Journal of Artificial Intelligence Research 71 (2021) 1183 - 1317 Submitted 09/2019; published 08/2021

Trends in Integration of Vision and Language Research: A Survey of Tasks, Datasets, and Methods

Aditya Mogadala AMOGADALA@LSV.UNI-SAARLAND.DE
Marimuthu Kalimuthu MKALIMUTHU@LSV.UNI-SAARLAND.DE
Dietrich Klakow DIETRICH.KLAKOW@LSV.UNI-SAARLAND.DE
Spoken Language Systems (LSV)
Saarland Informatics Campus
Saarland University
66123 Saarbrücken, Germany

Abstract

Interest in Artificial Intelligence (AI) and its applications has seen unprecedented growth in the last few years. This success can be partly attributed to the advancements made in the sub-fields of AI such as machine learning, computer vision, and natural language processing. Much of the growth in these fields has been made possible with deep learning, a sub-area of machine learning that uses artificial neural networks. This has created significant interest in the integration of vision and language. In this survey, we focus on ten prominent tasks that integrate language and vision by discussing their problem formulation, methods, existing datasets, evaluation measures, and compare the results obtained with corresponding state-of-the-art methods. Our efforts go beyond earlier surveys which are either task-specific or concentrate only on one type of visual content, i.e., image or video. Furthermore, we also provide some potential future directions in this field of research with an anticipation that this survey stimulates innovative thoughts and ideas to address the existing challenges and build new applications.

1. Introduction

Recent wave of unprecedented progress in deep learning methods has advanced the fields of Computer Vision (CV) and Natural Language Processing (NLP) to an extent that they are now making significant progress across several challenging tasks. Independent of NLP, computer vision has achieved prominent improvements in tasks such as visual content classification (He et al., 2016), object detection (Redmon & Farhadi, 2017), semantic segmentation (He et al., 2017), etc., using large annotated datasets or by employing self-supervision (Jing & Tian, 2019) on large-scale unlabeled data. Similarly, independent from computer vision, NLP has seen a surge of interest in solving multiple tasks at once with unsupervised pretraining of language models (Devlin et al., 2019; Radford et al., 2019; Conneau & Lample, 2019; Brown et al., 2020b) using large unlabeled corpora. However, there is a growing interest in solving challenges that combine linguistic and visual information from these traditionally independent fields. The methods which address the challenge of integration should provide a complete understanding of visual and/or textual content, and are expected to (1) generate comprehensible but concise and grammatically well-formed descriptions of the visual content, or vice versa by generating the visual content for a given textual description in a natural language of choice, (2) identify objects in the visual content

and infer their relationships to reason about, or answer arbitrary questions about them, (3) navigate through an environment by leveraging input from both vision and natural language instructions, (4) translate textual content from one language to another while leveraging the visual content for sense disambiguation, (5) generate stories about the visual content, and so on. Designing methods which can process and relate information from multiple modalities (i.e., linguistic and visual information) is usually considered to be a sub-part of multimodal learning models (Mogadala, 2015).

Efficiently solving the above mentioned and other related challenges can result in many potential real-world applications. For example, visually impaired individuals can be assisted by visual scene understanding, where they can get information about a scene from generated descriptions and by being able to ask questions about it. Other applications include automatic surveillance (Baumann et al., 2008), autonomous driving (Kim et al., 2018), human-computer interaction (Rickert et al., 2007), city navigation (de Vries et al., 2018), and so on. Also, solving such challenges can serve as an excellent test bed for computer vision and NLP systems, one that is much more intelligent and comprehensive than independent computer vision and NLP evaluations.

Given such a broad scope for fundamental and applied research, there has been several surveys in recent years aiming to provide a comprehensive overview of the integration of vision and language tasks. These surveys, however, have restricted themselves on covering specific vision and language integration tasks such as image description (Bernardi et al., 2016; Bai & An, 2018; Hossain et al., 2019) or video description generation (Aafaq et al., 2020), visual question answering (Kafle & Kanan, 2017; Wu et al., 2017), action recognition (Gella & Keller, 2017) and visual semantics (Liu et al., 2019). The surveys which went beyond these specific tasks have summarized dataset statistics (Ferraro et al., 2015), provided a comprehensive overview of only NLP tasks such as natural language generation (NLG) (Gatt & Krahmer, 2018; Garbacea & Mei, 2020) and commonsense reasoning (Storks et al., 2019). However, there was also an attempt to cover multiple modalities (including sound) (Baltrušaitis et al., 2019), but it was structured in a bottom-up manner giving more importance to underlying fusion technologies than the task itself. Also, there was some interest in understanding the limitations of the integration of vision and language research (Kafle et al., 2019). However, it is limited to the task of language-grounded image understanding. Furthermore, there were ideas to develop theories on the complementarity of language and visual data in the human-machine communication from a theoretical point of view (Moens et al., 2019).

With our efforts in this survey, we go beyond these and present a comprehensive overview of ten different tasks that are prominent in the current integration of vision and language research. We first begin with a background on the traditional tasks in computer vision and NLP separately, and then show how they facilitate in designing the prominent ten tasks for the integration of vision and language modalities in Section 2. Following that, we provide an in-depth exploration of each of the ten tasks and present details about the datasets, methods, results, and open challenges in separate sections beginning from Section 3 and ending at Section 9. In Section 10, we provide details about the joint pretraining of vision and language, which is gaining momentum in recent years, that aims to solve multiple tasks at once using learned representations. It is then followed in Section 11 by potential future research directions. Finally, in Section 12, we conclude our survey and offer some insights.

2. Background

In this section, we first briefly introduce some of the standard tasks that are studied in computer vision and NLP separately. We then present how the tasks are modified such that they facilitate in designing ten prominent tasks for the integration of vision and language.

2.1 Computer Vision (CV) Tasks

An array of different tasks are studied in computer vision. Keeping in mind the underlying goal of computer vision is to describe and explain visual information, we divide these tasks based on where the visual data arises. In this survey, we mainly focus on images and videos as the visual information, although RGB-D and point cloud data are becoming prevalent.

2.1.1 Image as Visual Information

We describe two different aspects of the use of images in computer vision: (1) the tasks where images are used as input, and (2) the representation of images. In the following, we discuss various computer vision tasks that use images as input and present the recent progress and improvements made for representing image data.

Tasks. The following tasks use images as input: (1) Image Classification (2) Object Localization (3) Object Detection (4) Object Segmentation (5) Object Identification (6) Instance Segmentation and (7) Panoptic Segmentation.

The fundamental difference between aforementioned tasks is that majority of them focus on carving out the exact position of visual object in an image, while rest of them provide a predefined class label for an image. There are also advanced tasks that use images as visual information and assist in the integration of computer vision and NLP. These tasks include (1) Image Style Transfer (2) Image Colorization and (3) Image Synthesis.

Representation. The advent of deep learning (LeCun et al., 2015; Bengio et al., 2021) has tremendously changed the field of computer vision. Convolutional Neural Networks (CNNs) (LeCun et al., 1995) have become the de facto standard for generating representations of images using end-to-end trainable models.

There are several variations of CNNs that learn image features with supervised or self-supervised techniques (Jing & Tian, 2019). Most of these techniques are designed to learn transferable general image features by leveraging tasks presented earlier.

Commonly, transferable global image representations are learned with deep CNN architectures such as AlexNet (Krizhevsky et al., 2012), VGGNet (Simonyan & Zisserman, 2015), GoogLeNet (Szegedy et al., 2015), Inception-v3 (Szegedy et al., 2016), Residual Networks (ResNets) (He et al., 2016), DenseNets (Huang et al., 2017), and Efficient Net (Tan & Le, 2019) using large datasets, viz. ImageNet[1] (Deng et al., 2009), MSCOCO[2] (Lin et al., 2014), and Visual Genome[3] (Krishna et al., 2017b). However, for some vision and language integration tasks, it is preferred to learn global image features during task-specific training as opposed to independently learning generic, pretrained representations.

[1] https://www.image-net.org
[2] http://cocodataset.org/#home
[3] https://visualgenome.org

For learning local features of objects that are typically represented with bounding boxes in images, the preferred choice is to utilize some region specific CNN architectures such as Region-based CNNs (R-CNN) (Ren et al., 2015b). More recently, there is an interest in using self-attention based approaches, namely Transformers (Vaswani et al., 2017) for achieving end-to-end object detection (Carion et al., 2020).

2.1.2 Video as Visual Information

Similar to images, when a video is used as visual data, we need to consider two crucial aspects: (1) knowing the tasks where videos are used as inputs, and (2) the representation of a video. In the following, we discuss different tasks in computer vision that use video as input and further present the recent progress made in video representations.

Tasks. Recently, the tasks on videos are also gaining importance, such as (1) Object Tracking (2) Action Classification (3) Emotion Detection (4) Scene Detection and (5) Automated Editing. The core difference between earlier tasks is that majority of them focus on tracking a visual object present in a scene of a video, while rest of them identify the task happening in a video such a action etc.

Representation. To account for the temporal nature of videos, RGB images are stacked as frames to form a 4D representation (i.e., video). Usually, visual data observed in videos is extracted in the form of screenshots that are amenable to the same techniques for image local and global representation. However, in addition, spatio-temporal features are also developed with general video analysis such as C3D (Tran et al., 2014), or from action recognition datasets i.e., Kinetics action recognition (Kay et al., 2017) to build RGB-D or Inflated 3D ConvNet (I3D) features (Carreira & Zisserman, 2017) using different CNN architectures.

2.2 NLP Tasks

Like in Section 2.1, the fundamental goals of most NLP tasks are to comprehend or generate language. In this section, we describe a few of the popular tasks that drive NLP research. We also discuss current approaches used to represent language.

Tasks. The aim of NLP tasks is two-fold: i) understanding language, ii) generating language. Some of the classical NLP tasks, that are used to comprehend language, are shallow parsing, syntax parsing, semantic role labeling, named entity recognition, entity linking, co-reference resolution, etc. Tasks to generate language in a conditional or unconditional manner are machine translation, text summarization, etc.

Representation. In deep learning based approaches, language is usually represented either as a bag-of-words or as distributed representations. For words in a sentence, initializations are commonly done with pretrained word embeddings (Mikolov et al., 2013; Pennington et al., 2014). Additionally, to represent variable-length text inputs, sequence learning techniques such as recurrent neural network variations like unidirectional Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997), or bidirectional LSTM (BiLSTM) and unidirectional Gated Recurrent Units (GRUs) (Chung et al., 2014), or bidirectional GRUs (BiGRUs) are applied. Recently, to provide parallelization in sequential

training, self-attention based approaches, viz. Transformers (Vaswani et al., 2017), have been employed to build architectures such as BERT (Devlin et al., 2019) and its variations.

2.3 CV and NLP Integration Tasks

Over the past few years, significant progress has been made in the research concerning the integration of language and vision. Several tasks exist which combine language, observed at different levels (i.e., words, phrases, sentences, paragraphs, and documents), with visual information, typically represented as images or videos. Initially, most works concentrated on combining low-level linguistic units, such as words with images or videos for building visual-semantic embeddings (Barnard et al., 2003; Frome et al., 2013; Kiros et al., 2014b; Liu et al., 2015; Cao et al., 2016; Tsai et al., 2017; Guo et al., 2018; Mogadala et al., 2018b; Wang et al., 2019; Kim et al., 2020), which are beneficial for downstream applications, as well as understanding adversarial attacks (Wu et al., 2019) to improve model robustness.

However, it will be appealing to look into those tasks that go beyond words and consider variable-length texts larger than words as language input. Most of these tasks can be seen as an extension to either CV, NLP, or both. Figure 1 provides an illustration of different tasks and their groupings.

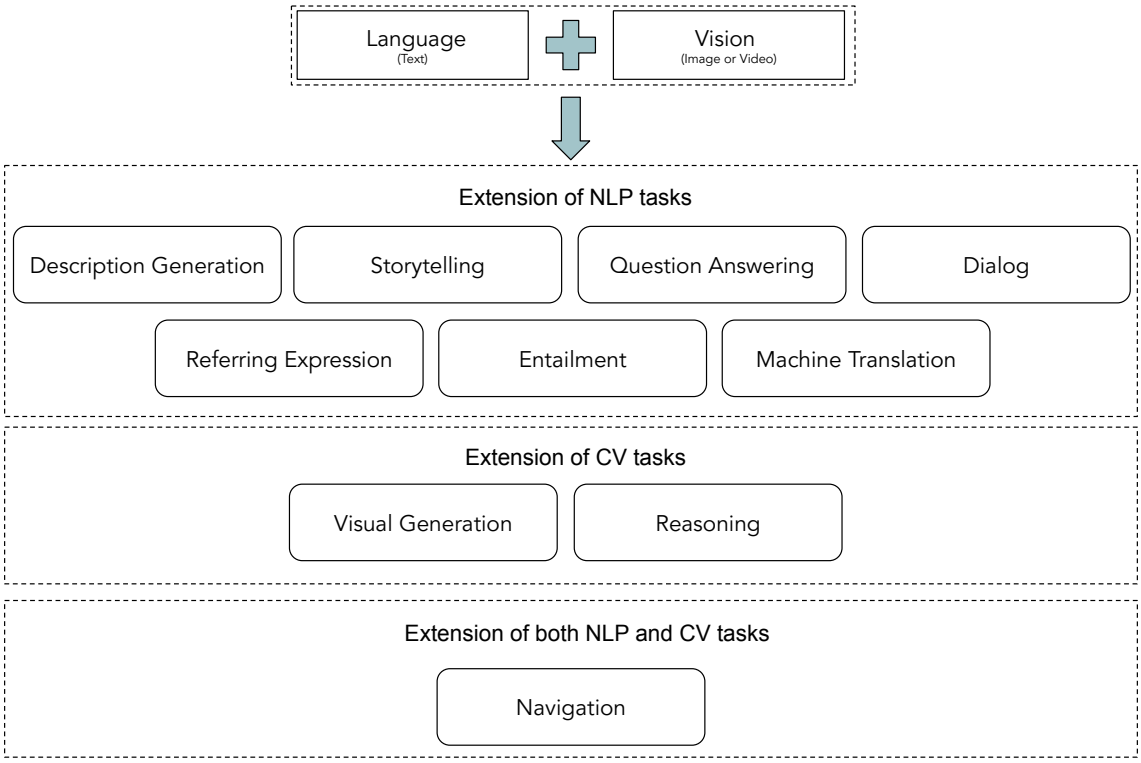

Figure 1: Ten different Language and Vision integration tasks.

To get a grasp on how these tasks are perceived as a natural extension of tasks in computer vision, NLP, or both, we briefly describe their relation with similar tasks addressed in their research.

Extension of NLP Tasks

- **Visual Description Generation** is closely related to conditional language modeling (De Mulder et al., 2015) or the Natural Language Generation (NLG) (Reiter & Dale, 2000) tasks in NLP. Given non-linguistic information (e.g., image or video), the goal is to generate a human-readable text snippet that describes the input.

- The task of **Visual Storytelling** solves a similar problem to visual description generation. However, instead of dealing with a single visual input, a sequence of visual inputs is used to generate a narrative summary based on the text aligned with them. It can be seen that the task is closely aligned to text summarization (Nallapati et al., 2016; Liu et al., 2018), mostly generating abstractive summaries.

- **Visual Question Answering** draws its inspiration from text-based question answering (Harabagiu et al., 2000; Strzalkowski & Harabagiu, 2006), which is one of the long-standing topics in NLP research. Here, answering questions about a visual input is perceived as its natural extension.

- The task of **Visual Dialog** aims at creating a meaningful dialog in a natural and conversational language about a visual content. It is perceived as the visual analogue of the text-based dialog and conversation system (Weizenbaum, 1966; Dodge et al., 2016; Li et al., 2016) that has been explored in NLP over many years.

- **Visual Referring Expression** is an extension of referring expression (Krahmer & Van Deemter, 2012) in natural language generation systems. Also, the sub-problem in visual referring expression (i.e., comprehension) is perceived as an analogy of pragmatics in linguistics (Thomas, 2014) due to its usage of context.

- **Visual Entailment** is an inference task for predicting whether the image semantically entails the text. The task has been proposed as a natural extension to natural language inference (Condoravdi et al., 2003; Bowman et al., 2015), where the premise is text, instead of a visual content.

- The goal in **Multimodal Machine Translation** is to achieve translation from source language(s) to target language(s) by leveraging the visual information as auxiliary modality along with the natural language text in source language(s). It is influenced by the well-known NLP task of machine translation that aims to automatically translate textual contents between any two natural languages (Brown et al., 1990; Bahdanau et al., 2015).

Extension of CV Tasks

- **Visual Generation** deals with the generation of visual content by conditioning on input text from a chosen natural language. It can be perceived as a multimodal extension of the popular computer vision tasks of image-to-image translation (Isola et al., 2017) and neural style transfer (Gatys et al., 2016).

- The task of **Visual Reasoning** is a direct extension of visual perception where standard computer vision tasks such as image classification (Krizhevsky et al., 2012),

object detection (Ren et al., 2015c), or semantic segmentation (Long et al., 2015) are performed. Instead of providing only class labels (in case of classification), bounding boxes (in case of detection), or segments (in case of segmentation), visual reasoning is expected to output a relationship between detected objects by generating an entire visual scene graph. Furthermore, the scene graph is leveraged to reason and answer questions about visual content. It can also be used to reason about whether a natural language statement is true or not regarding a visual input (Suhr et al., 2017).

Extension of both NLP and CV Tasks

- **Vision-and-Language Navigation** is one task that can be seen as a transition from standard vision-based navigation using only visual input (Sinopoli et al., 2001; Blösch et al., 2010) or natural language instruction based navigation (MacMahon et al., 2006; Vogel & Jurafsky, 2010). The expectation here is that the natural language navigation instruction should be interpreted based on visual input. Hence, it combines both vision and language.

Representation. In earlier sections, we discussed different architectures used to represent both vision and language separately. Combining representations of language and vision is essential to address vision and language integration tasks in an effective manner. There are various models that have been proposed for each task to build representations that integrate vision and language. We discuss these in greater detail in forthcoming sections where each of the tasks are introduced.

2.4 Summary

In background section, we have reviewed a variety of tasks that integrate vision and NLP. Additionally, we explored diverse methods that are used for the *representation* of vision and language modalities. Furthermore, we understood the training procedure used by different methods that use supervised learning. For example, models built using those methods leverage first-order optimization algorithms such as Stochastic Gradient Descent (SGD) (Bottou, 2010), ADAM (Kingma & Ba, 2015) or RMSProp (Tieleman & Hinton, 2012). While, some methods also utilize Reinforcement Learning (RL) (Sutton et al., 1998) in contrast with only supervised learning.

We will see that many of the models developed for these tasks use similar architectures for the representation of vision and language modalities and depend on standard gradient-based optimization algorithms for training. This shows that, although the aims of each task are different, the underlying principles to extract meaning from unstructured data remain the same.

3. Visual Description Generation and Storytelling

In this section, we explore two different tasks, *Visual Description Generation* and *Visual Storytelling*. Although the goals of these tasks do not perfectly line up, they share the common intention of generating a textual description when conditioned on visual input. In the following, we present more details about each of these tasks separately.

3.1 Visual Description Generation

The aim in description generation is to generate either a global description or dense captions for a given visual input. Depending on the type of visual input, i.e., either an image or a video, there are various ways to explore the problem.

3.1.1 Image Description Generation - Introduction

There are many subareas of image description generation where the underlying goal of generating global or dense descriptions remains the same, but the way those descriptions appear is different. In the following section, we explore some of the popular categories observed in image description generation.

Standard Image Description Generation. The goal in standard image description generation is to generate a sentence-level description of the scene in a given image. Here, methods leverage only vocabulary of the dataset to generate the best description that depicts the scene in the image. Figure 2 provides a conceptual representation of the task.

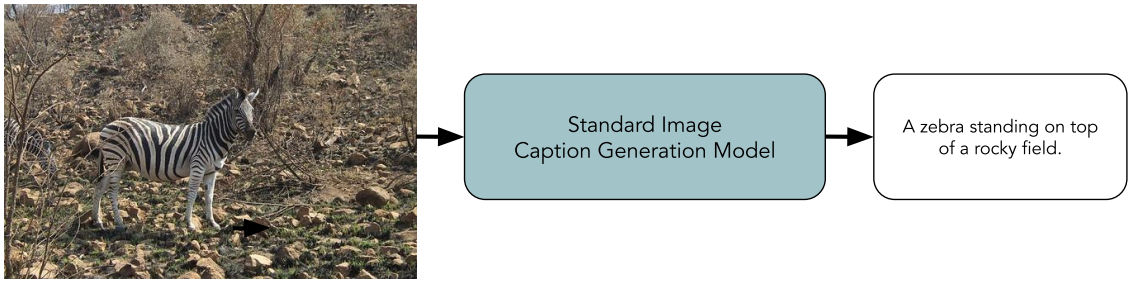

Figure 2: Given an *image*, the Standard Image Caption Generation Model generates a single global textual description of the scene.

Initially, several methods were developed based on templates, n-grams, and dependency parsing (Farhadi et al., 2010; Yang et al., 2011; Li et al., 2011; Mitchell et al., 2012; Kulkarni et al., 2013; Elliott & Keller, 2013; Fang et al., 2015). Recently, however, image description generation models based on the encoder-decoder a.k.a. Sequence-to-Sequence (Seq2Seq) frameworks (Cho et al., 2014; Vinyals et al., 2015) have become popular. Moreover, the above said frameworks have been extended with *attention* mechanisms (Bahdanau et al., 2015) to support the selection of local image features that are useful for the generation of words at each time step (Xu et al., 2015a). Table 1 summarizes different setups for generating image descriptions using neural network based non-attention, attention, and reinforcement learning techniques. Other variations include cross-lingual image captioning (Miyazaki & Shimizu, 2016) and multi-language image description generation (Elliott et al., 2015).

In the following, we explore some of the related ideas that expand the scope of image description generation.

Dense Image Description Generation. Dense image description generation task aims to create descriptions at the local object-level in a given image. It is referred to as dense captions since the commonly used image datasets have images containing multiple objects.

Approach	Attention	RL
MLBL (Kiros et al., 2014a)	✗	✗
m-RNN (Mao et al., 2015)	✗	✗
Minds Eye (Chen & Lawrence Zitnick, 2015)	✗	✗
BRNN (Karpathy & Fei-Fei, 2015)	✗	✗
NIC (Vinyals et al., 2015)	✗	✗
LRCN (Donahue et al., 2015)	✗	✗
Guided LSTM (Jia et al., 2015)	✗	✗
Deep Bidirectional LSTM (Wang et al., 2016)	✗	✗
Regional Visual Attributes (Wu et al., 2018)	✗	✗
Language CNN (Gu et al., 2017)	✗	✗
ConceptNet-NIC (Zhou et al., 2019)	✗	✗
Visual Attention (Xu et al., 2015a)	✓	✗
Region-based Attention (Jin et al., 2015)	✓	✗
Attribute Attention (You et al., 2016)	✓	✗
Review Attention (Yang et al., 2016)	✓	✗
Adaptive Attention (Lu et al., 2017b)	✓	✗
Areas of Attention (Pedersoli et al., 2017)	✓	✗
Contrastive Adaptive Attention (Dai & Lin, 2017)	✓	✗
Neural Baby Talk w/ Attention (Lu et al., 2018)	✓	✗
Convolutional Attention (Aneja et al., 2018)	✓	✗
Reflective Decoding Network (Ke et al., 2019a)	✓	✗
Self-Critical Attention (Rennie et al., 2017)	✓	✓
Policy Gradient (Liu et al., 2017)	✓	✓
Up-Down (Anderson et al., 2018a)	✓	✓
Multi-task Captioning (Zhao et al., 2018)	✓	✓
Stack Captioning (Gu et al., 2018)	✓	✓
Attention on Attention (Huang et al., 2019)	✓	✓
Meshed-Memory Transformer (Cornia et al., 2020)	✓	✓

Table 1: Summary of methods for generating a global description of an image. Approaches are segregated based on their usage of no-attention, attention, and RL techniques.

Several approaches (Plummer et al., 2017b; Johnson et al., 2016; Rohrbach et al., 2016a; Hu et al., 2017a) have been proposed to generate dense captions in images. Usually, they use representations of phrases and their relationships to generate descriptions (Kim et al., 2019).

Image Paragraph Generation. The aim in image paragraph generation is to create paragraphs instead of generating a single simple description, or dense descriptions for an image. Generated paragraphs are expected to be coherent and contain fine-grained natural language descriptions (Krause et al., 2017; Liang et al., 2017; Chatterjee & Schwing, 2018).

Spoken Language Image Description Generation. Spoken language image description generation expands the description generation task to work with spoken language, instead of limiting to only the written forms of language. Investigations such as visually

grounded speech signals (Chrupała et al., 2017) address the standard image description generation task from the perspective of a spoken language.

Stylistic Image Description Generation. Stylistic image description generation adds styles to the standard image description generation, where the generated descriptions adhere to a specific style. For example, Mathews et al. (2016) generated captions which capture the *sentiments* of an image, while Gan et al. (2017) attempted at generating humorous and romantic captions. In addition, this task has been extended by leveraging unpaired textual corpora (Mathews et al., 2018) to generate story-like captions. Furthermore, to make the generated captions more human-like, personality traits have been used to generate captions (Shuster et al., 2019). Recently, multi-style image description generation (Guo et al., 2019) has been explored, in which a single model using unpaired data is built to generate different stylized captions.

Unseen Objects Image Description Generation. Unseen objects image description generation leverages images which lack paired descriptions. Most of the paired image-description datasets have few visual objects to represent. Hence, methods such as Deep Compositional Captioning (DCC) (Hendricks et al., 2016), Novel Object Captioner (NOC) (Venugopalan et al., 2017), Constrained Beam Search (CBS) (Anderson et al., 2017), and LSTM-C (Yao et al., 2017) address the challenge of generating descriptions for these images. They generate descriptions for visual object categories that are previously unseen in image-description corpora, either by transferring information between seen and unseen objects before inference (i.e., before test time), or by keeping constraints on the generation of description words during inference (i.e., during test time). A few approaches (Mogadala et al., 2018a; Lu et al., 2018) have transferred information both before and during inference. Recently, pointing LSTM was designed to point to the novel objects (Li et al., 2019a) by balancing generation and copying of words. Nevertheless, earlier approaches work only with a limited set of objects. To address this issue, a large-scale nocaps dataset (Agrawal et al., 2019) was created.

Diverse Image Description Generation. Diverse image description generation task aims to incorporate variety and diversity in the generated captions. A few approaches (Dai et al., 2017; Shetty et al., 2017) have leveraged adversarial training, while Vijayakumar et al. (2016) used diverse beam search to decode diverse image captions in English. Approaches have also been proposed to describe cross-domain images (Chen et al., 2017).

Controllable Image Description Generation. Controllable image description generation task focuses on selecting specific objects in an image, defined by a control signal, to generate descriptions. Initially, Yin and Ordonez (2017) generated layouts from images, while Wang et al. (2018) counted image objects to produce multiple captions for a given image. Additionally, a control signal has been used to make the image captioning process more controllable, and also to generate diverse captions. Cornia et al. (2019) used either a sequence or a set of image regions. Also, chunks of the generated sentences were explicitly grounded on regions. Moreover, instead of making captions only diverse, there were also attempts to make the generated descriptions more accurate (Deshpande et al., 2018).

Image Caption Emendation as Generation. Caption emendation task is a variant of caption generation where the aim is to build a model to emend (a.k.a. edit or correct) both

syntactic and semantic errors in the captions. There has been a lot of interest in recent years on this emerging topic of research. Guo et al. (2020) proposed *Show, Tell, and Polish* framework to better mimic humans in sentence constructions. That is, coming up with a first version and then keep *polishing* it until it feels right. The core idea of this architecture is to perform a two-pass decoding, instead of the typical single-pass decoding. Thus, the model contains two decoder modules, viz., *base decoder* and *ruminant decoder*, whereby the base decoder generates a first version of caption which then feeds into the ruminant decoder for refinement (a.k.a. polishing). Along the same lines, Kalimuthu et al. (2020) introduced *fusion models for caption emendation*, which is a generic fusion model framework containing a standard encoder-decoder format image captioning model, a pretrained auxiliary language model (AuxLM - BERT MLM), and a fusion module component that fuses language-only representations of AuxLM and visual-linguistic representations of decoder using different fusion techniques. The intuition behind introducing an external language model trained on a large-scale language corpora is to capture world knowledge and rich linguistic features, which are both scarce in annotated captions data, in an attempt to generate fluent and accurate descriptions. In both of the above approaches, emendation is achieved by generating a caption while utilizing the baseline caption as a reference. That is, the model is trained to correct any errors and incongruencies in the baseline caption. Likewise, Sammani and Melas-Kyriazi (2020) propose *Show, Edit, and Tell* framework as an iterative adaptive refinement approach that utilizes attention LSTMs and denoising autoencoders for correcting captions.

3.1.2 Image Description Generation - Datasets

A wide range of datasets are available for conducting research in integration of vision and language. In fact, they are one of the main driving forces behind recent accelerated advancements that we are witnessing in this field (Bengio et al., 2021). Visual information associated with textual content in these datasets differ from each other in many aspects such as size, quality, and the way in which they are collected. In our survey, we summarize the characteristics of these datasets and provide basic statistics about them. However, we do not furnish a deeper analysis of them, as this was already done by Ferraro et al. (2015).

An array of diverse datasets, both of small and large-scale, were created and made available publicly in the past decade to address the challenge of image description generation. Some of the early large-scale datasets focus on image captions, while the others are only of small- or medium-scale. In the following sections, we cover only those datasets that are extensively used in the image captioning literature.

SBU Captioned Photo Dataset (SBU1M). SBU1M[4] (Ordonez et al., 2011) is an automatically collected image description dataset that uses query terms to retrieve images and associated text from Flickr[5]. This web-scale dataset is distributed as a single plain text file containing 1 million URLs of Flickr images and their corresponding captions. Although one of the older datasets in image description research, it has been rarely used in recent years. Table 2 provides basic statistics about this dataset.

[4] `http://vision.cs.stonybrook.edu/~vicente/sbucaptions`

[5] `https://www.flickr.com`

Total Images	Captions per Image	Total Captions	Object Categories
1,000,000	1	1,000,000	89

Table 2: Basic statistics of the SBU1M image description dataset.

Flickr8k. As with SBU1M, images in the Flickr8k[6] (Hodosh et al., 2013) dataset are also retrieved from Flickr[5]. However, unlike the automated way of collection of SBU1M, the images in Flickr8k are selected through user queries for specific objects and actions using the Amazon Mechanical Turk (AMT) platform. The images are then captioned by annotators on AMT such that each image contains five captions that are independently created. Table 3 presents the so-called *karpathy split*[7] of the dataset.

Split	Images	Captions per Image	Total Captions
Training	6,000	5	30,000
Validation	1,000	5	5,000
Test	1,000	5	5,000
Total	8,000	5	40,000

Table 3: Splits of the Flickr8k image description dataset.

Flickr30k. Flickr30k[8] (Young et al., 2014) is an extended version of the previously published Flickr8k dataset, containing images collected from Flickr[5] and captions obtained via crowdsourcing using AMT platform, following the same strategies employed in Flickr8k. Table 4 presents the previously-mentioned *karpathy split*[7] of the dataset.

Split	Images	Captions per Image	Total Captions
Training	29,000	5	145,000
Validation	1,014	5	5,070
Test	1,000	5	5,000
Total	31,014	5	155,070

Table 4: Splits of the Flickr30k image description dataset.

Flickr30k-Entities. Flickr30k-Entities[9] (Plummer et al., 2017b) extends Flickr30k with manually annotated bounding boxes for images and entity mentions in the captions in order to accomplish the task of language grounding in images, viz. *phrase localization*, while performing captioning. Specifically, there are 275,775 bounding boxes for the images of Flickr30k and 513,644 entity mentions in the 158k captions of Flickr30k. One peculiarity of this dataset is that it comes with 244k co-reference chains, in which each chain is a link between the mentions of the same entities across the five different captions of a given image. Some statistics and *karpathy split*[7] of this dataset is presented in Table 5.

[6] http://hockenmaier.cs.illinois.edu/8k-pictures.html
[7] https://cs.stanford.edu/people/karpathy/deepimagesent
[8] http://hockenmaier.cs.illinois.edu/Denotation.html
[9] http://bryanplummer.com/Flickr30kEntities

Split	Num. of Images	Object Categories	Objects per Category	Objects per Image	Captions per Image	Total Captions
Training	29,783	-	-	-	5	148,915
Validation	1,000	-	-	-	5	5,000
Test	1,000	-	-	-	5	5,000
Total	31,783	44,518	6.2	8.7	5	158,915

Table 5: Splits and statistics of the Flickr30k-Entities image description dataset.

MSCOCO. MSCOCO[2] (Lin et al., 2014) is a widely-used and considerably larger-scale dataset than the image captioning datasets discussed so far. It contains natural images that are collected from Flickr[5]. The AMT platform is then used to curate and collect descriptions for the images. This dataset does not have an official split, hence the *karpathy split*[7] from the above datasets is commonly used in the vision and language research community. The statistics and splits of the dataset can be found in Table 6.

Split	Images	Captions per Image	Total Captions	Object Categories
Training	113,287	5	566,435	-
Validation	5,000	5	25,000	-
Test	5,000	5	25,000	-
Total	123,287	5	616,435	80

Table 6: Splits of the MSCOCO image description dataset.

MSCOCO-Entities. MSCOCO-Entities[10] (Cornia et al., 2019) is a recently-introduced dataset based on the original MSCOCO (Lin et al., 2014) dataset, with the goal of achieving the twin challenges of grounding and controllability in generated image captions. Unlike Flickr30k-Entities, the grounding annotations in this dataset are obtained in a semi-automated way. Table 7 presents some statistics about the dataset as well as its split.

Split	Images	Total Captions	Noun chunks	Noun chunks per caption	Unique Classes
Training	113,287	545,202	1,518,667	2.79	1,330
Validation	5,000	7,818	20,787	2.66	725
Test	5,000	7,797	20,596	2.64	730

Table 7: Splits and statistics of the MSCOCO-Entities image description dataset.

STAIR Captions. STAIR Captions[11] (Yoshikawa et al., 2017) is a large-scale Japanese image captioning dataset that provides Japanese language descriptions for the 164,062 images of MSCOCO, while retaining the same dataset splits, viz. *karpathy split*[7], as with MSCOCO (see Table 6). The annotation of captions is done manually using crowdsourcing. Original statistics from the authors of the dataset is provided in Table 8.

[10] `https://github.com/aimagelab/show-control-and-tell`
[11] `http://captions.stair.center`

Total Num. of Images	Captions per Image	Total Num. of Captions	Vocabulary Size	Avg. Number of Chars
164,062 (123,287)	5	820,310 (616,435)	35,642 (31,938)	23.79 (23.80)

Table 8: Statistics of the STAIR Captions image description dataset (Japanese). Details on the public part of the dataset is indicated in brackets.

Multi30k-CLID. The Multi30k-CLID[12] (Elliott et al., 2016) dataset was designed for the task of Cross-Lingual Image Description (CLID) generation with an ultimate goal of pushing existing vision and language research towards *multilingual multimodal language processing*. In the first edition of the task in 2016, the Flickr30k-Entities[9] dataset (Plummer et al., 2017b) was extended to the German language by crowdsourcing the descriptions independently from their English language counterparts with the help of professional translators. As with original Flickr30k, each image comes with five descriptions in German. Hence, the English-German pairs are considered as comparable, though not parallel, corpora. The splits of this dataset for English and German languages can be found in Table 9.

Split	Images	Language of the Captions	
		English	German
Training	29,000	145,000	145,000
Validation	1,014	5,070	5,070
Testing	1,000	5,000	5,000

Table 9: Splits and statistics of the Multi30k-CLID (2016) dataset.

In the second version[13] of the task in 2017, the Flickr30k-Entities[9] dataset was further extended to support French language captions (Elliott et al., 2017). The annotations were again obtained via crowdsourcing following the same principles as with the previous version. Table 10 presents the number of instances in each language and the splits of the dataset.

Split	Images	Language of the Captions		
		English	French	German
Training	29,000	145,000	145,000	145,000
Validation	1,014	5,070	5,070	5,070
Testing	1,000	5,000	5,000	5,000

Table 10: Splits and statistics of the Multi30k-CLID (2017) dataset.

Similar to the earlier editions of the task, in the 2018 version[14] Czech language translations of the captions were added (Barrault et al., 2018). Following the same strategy of the prior versions of this dataset for obtaining annotations, human translators were employed to

[12] https://www.statmt.org/wmt16/multimodal-task.html
[13] https://www.statmt.org/wmt17/multimodal-task.html
[14] http://www.statmt.org/wmt18/multimodal-task.html

produce Czech translations for the captions of Flickr30k-Entities[9]. Table 11 presents splits and statistics of all four languages of the dataset.

Split	Images	Language of the Captions			
		Czech	English	French	German
Training	29,000	145,000	145,000	145,000	145,000
Validation	1,014	5,070	5,070	5,070	5,070
Testing	1,071	5,355	5,355	5,355	5,355

Table 11: Splits and statistics of the Multi30k-CLID (2018) dataset.

Conceptual Captions (CC). Conceptual Captions[15] (Sharma et al., 2018) is a recently introduced web-scale dataset containing more than 3.3M images paired with English language captions. The dataset was harvested from the web in an automatic manner in which the captions were extracted from the alt text of retrieved HTML webpages. As a consequence, contrary to other curated image captioning datasets in which each image is paired with five captions, the images in CC have only one description, a fact that is evident in Table 12 which also presents the dataset splits.

Split	Images	Captions
Training	3,318,333	3,318,333
Validation	15,840	15,840
Test	22,530	22,530

Table 12: Splits of the Conceptual Captions dataset.

Although it is a large-scale dataset with a wide variety and style in captions, continued availability of the dataset for downloading by future users is a major issue, primarily due to the fact that the dataset has been distributed as a CSV file containing URLs of images. Thus, it inherently suffers from the problem of URLs becoming stale (for instance due to contents being removed, unresponsive HTTP requests, etc.), which puts the dataset at a disadvantage.

Personality Captions (PC). Personality Captions[16] (Shuster et al., 2019) is a large scale image caption dataset that comes with so-called *personality traits* that are useful for controllable and style-based image captioning. Thus, the samples in the PC dataset are provided as triplets (*image*, *personality trait*, *caption*). Basic statistics such as vocabulary size, including the dataset splits, is provided in Table 13.

[15] `https://ai.google.com/research/ConceptualCaptions/download`
[16] `https://parl.ai/projects/personality_captions`

Split	Num. of Images	Captions per Image	Num. of Captions	Personality Types	Vocabulary Size	Avg. Tokens per Caption
Training	186,858	1	186,858	215	33,641	11.2
Validation	5,000	1	5,000	215	5,460	10.9
Test	10,000	5	50,000	215	16,655	11.1

Table 13: Splits and statistics of the Personality Captions dataset.

3.1.3 Image Description Generation - Evaluation Measures, Models, and Results

In this section, we describe only the evaluation measures which are used for the task of *Image Description Generation*, as **Models**, **Results**, and some **Discussion** have been broadly presented in recent surveys (Hossain et al., 2019).

Evaluation Measures. We divide the evaluation measures into three different categories. The first set of measures is "Language Metrics", the second category is about "Retrieval Metrics", and the third category denotes "Human Evaluation".

"Language Metrics" evaluate machine-generated text based on reference text by computing similarity scores using simple n-gram statistics and word overlaps.

- **Bilingual Evaluation Understudy (BLEU)** (Papineni et al., 2002) was originally developed for machine translation to compare machine generated output with human Ground Truth (GT). BLEU calculates the overlap between predicted unigrams (BLEU-1 (B-1)), or, more generally, n-grams (BLEU-2 (B-2) with bigrams, BLEU-3 (B-3) with trigrams, BLEU-4 (B-4) with quadrigrams, and so on.) from the set of candidate and reference sentences. To achieve a high BLEU score, generated descriptions should match the human GT words as well as their order. Maximum achievable BLEU score is 1.0 (or sometimes, equivalently 100), which is obtained when an exact match occurs between generated and reference sentence.

- **Metric for Evaluation of Translation with Explicit Ordering**, popularly known as **METEOR** (Banerjee & Lavie, 2005) has overcome some issues of BLEU, such as the need for exact word matching. Instead of a literal token matching, METEOR rather performs semantic matching by leveraging WordNet to match words at various levels, using synonymy and paraphrase matching. The METEOR score is then computed using the alignment between the machine generated output and the corresponding reference sentences. To be more specific, initially, the set of unigrams from the generated and reference sentences is used to perform an alignment. If multiple options are available for alignments between the generated and reference sentence, the alignment setting with least comparisons is preferred. After finalizing the alignment process, the METEOR score is calculated.

- **Recall Oriented Understudy for Gisting Evaluation (ROUGE)** (Lin, 2004) was designed to evaluate textual summaries. As opposed to BLEU, which concentrates on n-gram precision, ROUGE instead calculates the recall score of the generated sentences

corresponding to the reference sentences. The most prominent ROUGE variant used is ROUGE-L, which is based on the longest common subsequence. Other variants include ROUGE-W (Weighted Longest Common Sub-sequence) and ROUGE-S (Skip-Bigram Co-Occurrences Statistics). One advantage of ROUGE-L over BLEU and METEOR is that it checks for subsequences within a sentence. Moreover, specifying the n-gram length (as required in BLEU) is not necessary as it is automatically incorporated.

- **Consensus-based Image Description Evaluation (CIDEr)** (Vedantam et al., 2015) evaluates the consensus between a generated sentence and a set of reference sentences by performing different language pruning techniques, such as stemming and building a set of n-grams. N-grams that are common among the reference sentences of all visual data are given lower weight, as they are less informative about the visual content, and biased towards the textual content of the sentences. The weight for each n-gram is computed using Term Frequency (TF) - Inverse Document Frequency (IDF) (TF-IDF), where TF puts higher weight on frequently occurring n-grams in the reference sentence of the visual content, whereas IDF puts lower weight on commonly appearing n-grams across the whole dataset. To remove the mismatch between human evaluation and CIDEr scores, a variant of CIDEr, CIDEr-D, is used. It adds small variations, such as not performing stemming and ensuring that the words with high confidence are not repeated in a sentence by introducing a Gaussian penalty over length differences between the generated and reference sentences. As in the case of vanilla CIDEr, it produces high scores even if the sentences do not make sense.

- **Semantic Propositional Image Captioning Evaluation (SPICE)** (Anderson et al., 2016) measures the similarity between the scene graph tuples parsed from generated sentences and human created GT sentences. The scene graph encodes objects and their relationships through dependency parsing. Hence, it makes SPICE heavily dependent on parsing, which can be prone to errors. Similar to METEOR, SPICE uses WordNet to find and treat synonyms as positive matches when computing the F1 score between the tuples of generated sentences and the ground truth.

"Retrieval Metrics" evaluate the machine generated text based on standard information retrieval measures (Manning et al., 2010) and are presented in the following paragraphs.

- **Recall@k (R@k)**'s goal is to evaluate the number of relevant ground truth sentences retrieved in the Top-k (e.g., Top-1, Top-5 etc.) candidates. A higher R@k indicates better performance.

- **Median Rank (MedRank)** finds the median rank value of the retrieved ground truth. A lower MedRank value indicates better performance.

- **Mean Reciprocal Rank (MRR)** is a binary measure, where the rank of the highest ranking relevant document for a query is used to calculate the reciprocal rank averaged over all queries. A higher MRR indicates better performance.

- **Mean Rank (Mean)** refers to the mean rank achieved in retrieving the relevant sentence. A lower Mean value is better.

- **Normalized Discounted Cumulative Gain (NDCG)** is a variant of Discounted Cumulative Gain (DCG) (Järvelin & Kekäläinen, 2000). NDCG is a cumulative, multilevel measure of ranking quality that is usually truncated at a particular rank level.

"Human Evaluation" employs crowd-workers to evaluate the quality of the generated content and is described in the following paragraph.

- **Human Evaluation** The earlier discussed metrics provide only quantitative measures for evaluating different tasks. Due to the lack of high correlation between machine-generated textual or visual data with the human provided GT, most of the tasks, however, require human evaluations to judge the quality of the generated content. Therefore, based on the task, various kinds of instructions are given to humans who act as an evaluator in the evaluation study. In most tasks, we are interested only in finding relevance of the output to input.

3.1.4 Video Description Generation - Introduction

Going beyond images, the goal in video captioning is to comprehend the spatio-temporal information in a video for the purpose of generating either a single or multiple textual descriptions. As with image description generation (Section 3.1.1), we explore some of the popular types and categories of video description generation tasks in the following.

Global Video Description Generation. Global video description generation approaches (Motwani & Mooney, 2012; Regneri et al., 2013) initially started by grounding sentences that describe actions in the visual information extracted from videos. It was further expanded into generating global natural language descriptions for videos with various approaches, for example, leveraging latent topics (Das et al., 2013), corpora knowledge (Krishnamoorthy et al., 2013), graphical models (Rohrbach et al., 2013), and sequence-to-sequence learning (Venugopalan et al., 2015b, 2015a; Donahue et al., 2015; Srivastava et al., 2015; Xu et al., 2016; Ramanishka et al., 2016; Jin et al., 2016). Figure 3 depicts the description generation task for a complete video. The aforementioned approaches leverage only those

Figure 3: Given a *video* (represented as sequence of frames), the Video Caption Generation Model generates a single global description.

training datasets with a limited set of visual objects. However, the recognition and description of entities and activities in real-world videos is more difficult. Nevertheless, generating natural language descriptions for such videos is addressed with a factor graph by combining visual detection with language statistics (Thomason et al., 2014).

Additionally, sequence-to-sequence (seq2seq) based approaches have been improved with external corpora (Venugopalan et al., 2016) and also using *attention* with various techniques such as soft-attention (Yao et al., 2015), multimodal fusion (Hori et al., 2017), temporal

attention (Song et al., 2017), semantic consistency (Gao et al., 2017), and residual connections (Li et al., 2019). Apart from attention-based methods, novel architectures have also been explored, such as incorporation of semantic attributes learned from videos (Pan et al., 2017), ensemble-based description generator networks (Shetty et al., 2018) and encoder-decoder-reconstructors which leverage both the forward and backward flows, i.e., video-to-description and description-to-video, for video captioning (Wang et al., 2018). Multi-faceted attention has also been used to select the most salient visual features or semantic attributes, with which an overall sentence is generated (Long et al., 2018).

Apart from architecture improvements, different machine learning approaches have also been explored. Video captioning has been tackled using a multi-task learning scenario by sharing knowledge between two related tasks (such as temporal- and context-aware video) combined with entailment generation task (Pasunuru & Bansal, 2017a). Other approaches have leveraged reinforcement learning, either by providing entailment rewards (Pasunuru & Bansal, 2017b) , or to address the description generation for multiple fine-grained actions (Wang et al., 2018b). Further, Mazaheri and Shah (2018) proposed a deep network designed to detect inaccuracies in a sentence, and fix them by replacing the inaccurate word(s) with the help of a Visual Text Correction system. Recently, Zhang et al. Zhang et al. (2020) introduced an object relational graph (ORG) based encoder which encapsulates the relation among visual objects to build richer representation and a decoder the integrates the external language model to capture abundant linguistic knowledge for efficient video description generation.

In the following, we discuss some related ideas which expand the scope of video description generation.

Dense Video Description Generation. The aim of dense video description generation is to achieve fine-grained video understanding by addressing two sub-problems: (1) localizing events in a video, and (2) generating captions for these localized events (Zhou et al., 2018b; Xu et al., 2019). Further, extending earlier research, some approaches (Zhou et al., 2019) have explicitly linked the sentence to a corresponding bounding box in one of the frames of a video by annotating each of the noun phrases observed in the sentence. Incorporating background knowledge for video description generation is also another line of research (Whitehead et al., 2018). However, the core challenge, namely the automatic evaluation of video captioning, is still unsolved. It is currently being studied from the perspective of direct assessment with the help of human assessors (Graham et al., 2018).

Movie Description Generation. Movie description generation perceives the video description generation task from a different perspective, in which movie clips are used as inputs. Initially, aligning books to movies (Tapaswi et al., 2015; Zhu et al., 2015) was used to generate story-like explanations. Later, movie descriptions (Rohrbach et al., 2015) were directly created by transcribing audio descriptions by concentrating on precisely describing what is shown in the movie scenes.

3.1.5 Video Description Generation - Datasets

Similar to the *image* description generation task, several datasets have been created to address the task of *video* description generation. In the following, we cover those datasets

that are popular and extensively used. For the sake of brevity, we denote *hours* → *h*, *minutes* → *m*, and *seconds* → *s*.

Microsoft Video Description (MSVD). MSVD[17] (Chen & Dolan, 2011) is an open domain dataset collected from YouTube clips and annotated using AMT. The dataset is multilingual and contains human generated descriptions in languages such as German, English, Chinese, etc. On average, there are forty-one single sentence descriptions per clip. More statistics about the dataset are presented in Table 14 whereas Table 15 presents its split.

Total Videos	Total Classes	Total Length	Avg. Length	Total Clips	Total Sentences	Total Words	Vocabulary Size
1,970	218	5.3 h	10 s	1,970	70,028	607,339	13,010

Table 14: Statistics of the MSVD dataset.

Split	Frames	Videos
Training	33,682	1,200
Validation	3,275	100
Test	20,528	670
Total	57,485	1970

Table 15: Splits of the MSVD dataset.

MPII Cooking Activities. The MPII Cooking[18] (Rohrbach et al., 2012) dataset consists of 65 different cooking activities such as "wash hands", "put in bowl", etc., when participants are preparing one of 14 dishes such as *fruit salad*, *casserole*, etc. The dish preparation time ranges between 3 and 41 minutes. The videos are recorded in high resolution (1624x1224), following which the *activity* annotations are manually created by 6 people. Table 16 presents more statistics about the dataset whereas the splits of it can be found in Table 17.

Num. of Subjects	Total Clips	Total Videos	Total Frames	Video Length	Total Length	Num. of Activities	Total Dishes	Activity Annotations
12	5,609	44	881,755	3 to 41 m	8.0 h	65	14	5,609

Table 16: Statistics of the MPII Cooking Activities dataset.

YouCook. YouCook[19] (Das et al., 2013) is a more complex real-world cooking dataset when compared to MPII Cooking in which the complexity arises because of dynamic scene and camera changes. The videos are all downloaded from YouTube and are broadly categorized into 6 different cooking styles, viz. baking, grilling, etc. Video descriptions are

[17] https://www.cs.utexas.edu/users/ml/clamp/videoDescription

[18] https://www.mpi-inf.mpg.de/departments/computer-vision-and-machine-learning/research/human-activity-recognition/mpii-cooking-activities-dataset

[19] http://web.eecs.umich.edu/~jjcorso/r/youcook

Split	Frames	Subjects
Training	1,071	10
Validation	-	-
Test	1,277	7

Table 17: Splits of the MPII Cooking dataset.

obtained via crowdsourcing using AMT. On average, eight descriptions are collected per video. Frames are annotated with objects belonging to *categories* (such as bowls, utensils, etc.) and *actions*. More details and splits of the dataset can be found in Table 18 and Table 19 respectively.

Cooking Styles	Object Classes	Total Videos	Total Length	Num. of Sentences	Num. of Words	Vocabulary Size
6	10	88	2.3 h	2,688	42,457	2,711

Table 18: Statistics of the YouCook dataset.

Split	Videos
Training	49
Validation	-
Test	39

Table 19: Splits of the YouCook dataset.

YouCook II. Similar to the YouCook dataset, YouCook II[20] (Zhou et al., 2018a) also consists of instructional cooking videos that are all collected from YouTube. The videos include 89 cooking recipes from four regions: South Asia, East Asia, Europe/Middle East, and America. One unique aspect of this dataset when compared to previously discussed video description datasets is that that the videos are annotated with *procedure segments* that contain rich semantic information. Table 20 presents the statistics about the dataset.

Cooking Recipes	Total Videos	Total Video Length	Avg. Video Length	Procedure Seg. per Video	Total Clips	Num. of Sentences	Vocab. Size
89	2,000	175.6 h	316 s	3-16	15,400	15,400	2,600

Table 20: Statistics of the YouCook II dataset.

For each recipe, the videos are randomly split into training, validation, and testing in ratios of 67%, 23%, and 10% respectively. The actual numbers are presented in Table 21.

Textually Annotated Cooking Scenes (TACoS). The TACoS[21] (Regneri et al., 2013) dataset is an extended version of a subset of MPII Composites (Rohrbach et al., 2012)

[20] http://youcook2.eecs.umich.edu

[21] https://www.coli.uni-saarland.de/projects/smile/page.php?id=tacos

Split	Videos
Training	1,340
Validation	460
Test	200

Table 21: Splits of the YouCook II dataset.

which contains cooking videos that are each annotated with multiple textual descriptions. It contains only those videos that include activities such as manipulation of cooking ingredients. Around 26 cooking activities are collected with 127 videos. More statistics on the dataset is presented in Table 22 and Table 23. For building and evaluating models, the dataset is split into 50% for training, 25% for validation, and 25% for testing.

Total Videos	Total Clips	Descriptions per Video	Annotation Assignments	Annotations after filtering	Cooking Tasks/Dishes	Action Descriptions
127	7,206	20	2,540	2,206	26	17,334 (tokens)

Table 22: The TACoS dataset statistics - I.

Sentence Types	Total Words	Content Words (viz. nouns, verbs, adjectives)	Num. of Verbs (tokens)	Num. of Verbs (lemmas)
11,796	146,771	75,210	28,292	435

Table 23: The TACoS dataset statistics - II.

TACoS-MultiLevel. The above discussed TACoS dataset was extended into TACoS-MultiLevel[22] (Rohrbach et al., 2014) by collecting three levels of descriptions constituting (i) 15 detailed descriptions per video, (ii) 3-5 short descriptions, and (iii) a single sentence description, using AMT platform. Overall, the dataset comes with 2,600 triplets of descriptions. Further statistics on the dataset can be found in Table 24.

Total Videos	Total Clips	Total Video Length	Avg. Length	Number of Sentences	Total Words
185	14,105	27.1 h	360 s	52,593	2,000

Table 24: Statistics of the TACoS-MultiLevel dataset.

MPII Movie Description (MPII-MD). MPII-MD[23] (Rohrbach et al., 2015) dataset contains clips extracted from Hollywood movies and their transcribed audio descriptions. In addition, each clip is paired with a single sentence that is extracted from the script of

[22] https://www.mpi-inf.mpg.de/departments/computer-vision-and-machine-learning/research/vision-and-language/tacos-multi-level-corpus

[23] https://www.mpi-inf.mpg.de/departments/computer-vision-and-machine-learning/research/vision-and-language/mpii-movie-description-dataset

the movie. Furthermore, transcribed audio is associated with spoken sentences by using timestamps. Misalignment between the audio and visual content is handled by leveraging manual annotation. Additional statistics on the dataset is presented in Table 25.

	Unique Movies	Before alignment Words	After alignment Words	Sentences	Clips	Avg. Length	Total
Audio Desc.	55	346,557	332,846	37,272	37,266	4.1 s	42.5 h
Movie script	50	398,072	320,621	31,103	31,071	3.6 s	31.1 h
Total	94	744,629	653,467	68,375	68,337	3.9 s	73.6 h

Table 25: Statistics of the MPII-MD dataset.

For the task of video description, the MPII-MD dataset is split as follows: 11 movies with associated scripts and audio descriptions (in total 22 alignments, 2 per movie) are used as validation (8) and test sets (14). The remaining 83 movies are used for training purposes.

Montreal Video Annotation Dataset (M-VAD). M-VAD[24] (Torabi et al., 2015) is a large Descriptive Video Service (DVS)-derived video dataset that is created using 92 Movies, covering a wide variety of genres. It is collected in a semi-automatic manner with minimal human intervention. The words in the descriptions are annotated with Part-Of-Speech (POS) tags using the Stanford POS tagger. Around 500 proper names are removed from the corpus, since learning proper names is not interesting for a video description model.

Type	Movies	Words	Paragraphs	Sentences	Avg. Length	Total
Un-filtered	92	531,778	52,683	59,415	6.3 s	91 h
Filtered	92	510,933	48,986	55,904	6.2 s	84.6 h

Table 26: Statistics of the M-VAD dataset.

Table 26 presents some statistics about the dataset, while Table 27 presents the official dataset split that balances the genre within each split.

Split	Video Clips
Training	38,949
Validation	4,888
Test	5,149

Table 27: Splits of the M-VAD dataset.

MSR Video to Text (MSR-VTT). MSR-VTT[25] (Xu et al., 2016), also known as MSR-VTT-10k, is a large-scale video dataset containing automatically crawled videos belonging to 20 categories for the task of video description generation. The sentence annotations are obtained via crowdsourcing using AMT. In addition to the video content, the dataset also contains audio information. Table 28 presents more statistics about the dataset.

[24] https://mila.quebec/en/publications-archive/public-datasets/m-vad/

[25] http://ms-multimedia-challenge.com/2017/dataset

Categories	Videos	Clips	Sentences per Clip	Sentences	Words	Vocab.	Duration
20	7,180	10,000	20	200,000	1,856,523	29,316	41.2 h

Table 28: Statistics of the MSR-VTT dataset.

Out of 7.2k videos, 30k video clips have been created. However, only a random subset of 10k clips has been released. The dataset is split in the ratio of 65%:30%:5% for training, validation, and testing. Specific numbers are presented in Table 29.

Split	Video Clips
Training	6,513
Validation	497
Test	2,990

Table 29: Splits of the MSR-VTT dataset.

Videos Titles in the Wild (VTW). VTW[26] (Zeng et al., 2016) is a large-scale dataset of automatically crawled user-generated YouTube videos paired with titles and descriptions. The video clips are on average 90 seconds in duration and are described with one sentence per clip to enable video title generation. It also comes with augmented sentences that contain information that may not be present in the video clip. More statistics of the dataset can be found in Table 30.

Dataset	Sentences	Vocab.	Sentences/Word	Nouns	Verbs	Adjective	Adverb
VTW-title	18,100	8,874	2.0	5,850	2,187	1,187	224
VTW-full	44,603	23,059	1.9	13,606	6,223	3,967	846

Table 30: Statistics of the VTW dataset.

Similar to M-VAD, the dataset is randomly split into 80% for training and 10% each for validation and testing. Specific numbers are presented in Table 31.

Split	Videos	Sentences/Titles
Training	14,100	14,100
Validation	2,000	2,000
Test	2,000	2,000

Table 31: Splits of the VTW dataset.

ActivityNet Captions (ANetCap). ANetCap[27] (Krishna et al., 2017a) is a large-scale video dataset[28] that extends a subset of videos from ActivityNet with dense descriptions.

[26] http://aliensunmin.github.io/project/video-language/index.html#VTW
[27] http://activity-net.org/challenges/2017/captioning.html
[28] https://cs.stanford.edu/people/ranjaykrishna/densevid

There are multiple descriptions for every video and the videos contain multiple events occurring at the same time. Another notable aspect of this dataset is that the descriptions focus more on actions happening in videos. As a result, this dataset falls under the category of being more action-centric than object-centric.

Videos	Total Video Hours	Avg. Video Length	Sentences	Avg. Sentence Length
20,000	849	180 s	100,000	13.48 (words)

Table 32: Statistics of the ANetCap dataset.

Table 32 presents more statistics on the dataset, while Table 33 presents its split.

Split	Videos
Training	10,024
Validation	4,926
Test	5,044

Table 33: Splits of the ANetCap dataset.

ActivityNet Entities (ANetEntities). The ANetEntities[29] (Zhou et al., 2019) dataset augments ANetCap (Krishna et al., 2017a) with manually annotated bounding boxes, and was created for the task of grounding language in videos while generating descriptions. It adds around 158k bounding box annotations on ANetCap, each grounded to a Noun Phrase (NP) in the sentence description. More statistics and the dataset splits can be found in Table 34.

Split	Videos	Sentences	Objects	Bounding Boxes
Training	10,000	35,000	432	105,000
Validation	2,500	8,600	427	26,500
Test	2,500	8,500	421	26,100
Total	15,000	52,100	432	157,600

Table 34: Statistics and splits of the ANetEntities dataset.

COmprehensive INstructional video analysis (COIN). COIN[30] (Tang et al., 2019) is a large-scale dataset of instructional YouTube videos from 12 domains such as vehicles, gadgets, sports, etc., that are common in our daily lives. It is aimed at overcoming two limitations of current instructional video datasets, namely *diversity* and *scale*. It covers over 180 tasks in 12k videos.

One unique aspect of this dataset is that it introduces a three-level hierarchy, viz. *domain*, *task*, and *step*, for organizing videos. Table 35 shows some statistics of the dataset whereas Table 36 presents training and validation splits of COIN.

[29] https://github.com/facebookresearch/ActivityNet-Entities

[30] https://coin-dataset.github.io

Num. of Domains	Num. of Tasks	Total Videos	Total Segments	Total Duration	Avg. Video Length	Avg. Segment Length
12	180	11,827	46,354	476 h, 38 m	2.36 m	14.91 s

Table 35: Statistics of the COIN dataset.

Split	Videos
Training	9,030
Validation	-
Test	2,797

Table 36: Splits of the COIN dataset.

HowTo100M. HowTo100M[31] (Miech et al., 2019) is a large-scale dataset of narrated videos with emphasis on instructional YouTube videos where the video creators teach complex tasks with an explicit intention of explaining the visual content on screen. The dataset includes a wide variety of 23k activities from the domains such as gardening, personal care, fitness, hand crafting, cooking, etc. and is three orders of magnitude than the previously discussed video description datasets. Table 37 presents more statistics about the dataset.

Num. of Domains	Num. of Tasks	Total Videos	Total Clips	Total Duration	Total Captions	Avg. Video Length	Avg. Clip-Caption Pairs per Video
12	23,611	1.221M	136M	134,472 h	136M	6.5 m	110

Table 37: Statistics of the HowTo100M dataset.

This dataset has not yet been used for the task of video description generation. Hence, an official dataset split is not available for evaluation purposes.

3.1.6 Video Description Generation - Evaluation Measures, Models, and Results

In this section, we describe only the evaluation measures which are used for the task of *Image Description Generation* as **Models**, **Results**, and some **Discussion** have been broadly discussed in recent surveys (Aafaq et al., 2020).

Evaluation Measures. The measures used for *Video Description Generation* are the same as the Language metrics and Retrieval metrics used in *Image Description Generation* and are presented in the Section 3.1.3.

3.2 Visual Storytelling

The task of visual storytelling aims to encode a sequence of images or frames (in the video) to generate a paragraph which is story-like. This is usually considered more beneficial than generating a paragraph from a single image or video.

[31] `https://www.di.ens.fr/willow/research/howto100m`

3.2.1 Image Storytelling - Introduction

The aim of image storytelling is to generate stories from a sequence of images. Although sequence of images can be perceived as a video, consecutive images in the streams can have sharp changes of visual content, which can cause an abrupt discontinuity between consecutive sentences (Park & Kim, 2015). Hence, it is seen as a sequential vision-to-language task (Huang et al., 2016) where images are not considered in isolation. Figure 4 shows a schematic representation of image storytelling where a story in a sequence is generated.

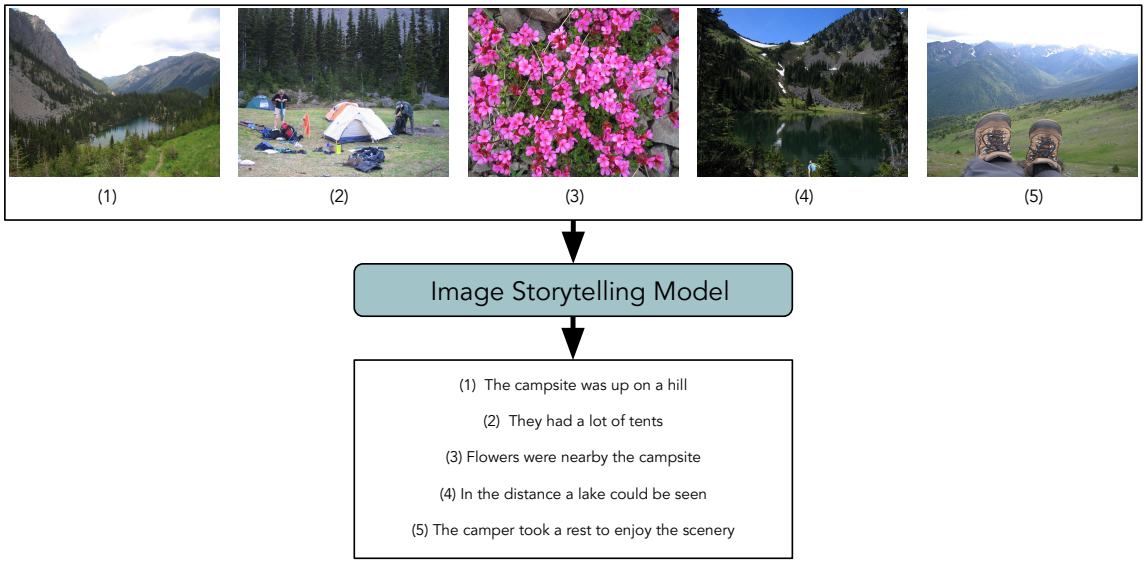

Figure 4: Given a *sequence of images*, an Image Storytelling Model generates a textual story in sequence.

Initially, semantic coherence in a photo stream is captured by reducing the visual variance. Further, the semantic space is acquired by jointly embedding each photo with its corresponding contextual sentence such that their correlations are discovered (Liu et al., 2017). It was then improved by exploiting hierarchical architecture (Yu et al., 2017) and further optimized by incorporating reinforcement learning with rewards (Wang et al., 2018) for generating relevant and expressive narrative paragraphs. Instead of *flat deep reinforcement learning*, a hierarchically structured reinforced training has also been studied (Huang et al., 2019) and has been shown to achieve significantly better performance than with a flat structure. Similarly, Wang et al. (2018a) used adversarial reward learning to learn an implicit reward function from human demonstrations to optimize policy search with the learned reward function.

Nevertheless, the standard form of narration suffers from repetitiveness, with the same objects or events serving to undermine a good story structure. Hence, inter-sentence diversity was explored with diverse beam search to generate more expressive stories (Hsu et al., 2018). The task has also been approached from a different perspective, in which, given a jumbled set of aligned image-description pairs that belong to a story, the task is to sort them such that the output sequence forms a coherent story (Agrawal et al., 2016).

While earlier research addresses only natural images, some approaches (Li et al., 2019) also incorporated medical domain knowledge to generate realistic and accurate descriptions for medical images.

3.2.2 Image Storytelling - Datasets

There are not many datasets created to address the creative task of image storytelling. In the following, we cover all datasets that have been used to advance this artistically interesting and challenging problem.

New York City Storytelling (NYC-Storytelling). The NYC-Storytelling[32] (Park & Kim, 2015) dataset was created from blogs in which users post their travelogues. The dataset is collected in a semi-automatic manner: automatic crawling followed by manual selection of travelogues and finally preprocessing using the NLTK[33] library. For evaluation purposes, the dataset is split in a ratio of 8:1:1 for training, validation, and testing respectively. Table 38 presents minimal statistics of the dataset.

Images	Blog posts
78,467	11,863

Table 38: Statistics of the NYC-Storytelling dataset.

Disneyland Storytelling. Similar to NYC-Storytelling, Disneyland Storytelling is also based on blogs documenting travelogues but specifically about *Disneyland Park*. This dataset was originally created by (Kim et al., 2015) but has been reused for visual storytelling tasks. The same ratio of data splits as with the NYC-Storytelling dataset is used for evaluation purposes. The minimal statistics of the dataset can be found in Table 39.

Images	Blog posts
60,545	7,717

Table 39: Statistics of the Disneyland-Storytelling dataset.

Sequential Image Narrative Dataset (SIND). SIND (Huang et al., 2016) is the first large-scale dataset created for the task of image storytelling. Natural language descriptions of the dataset are divided into three types: (i) Descriptions of Images-in-Isolation (DII), (ii) Descriptions of Images-in-Sequence (DIS), and (iii) Stories for Images-in-Sequence (SIS). The stories are collected via crowdsourcing using AMT. Similar to other image storytelling datasets, this dataset is split into 80%, 10%, and 10% for training, validation, and testing purposes respectively. Table 40 presents the statistics of the dataset.

Visual Storytelling Dataset (VIST). VIST[34] is the second version (v.2) of SIND (see Section 3.2.2) and is aimed at modeling the social language of humans for evolving AI to

[32] https://github.com/cesc-park/CRCN
[33] https://www.nltk.org
[34] http://visionandlanguage.net/VIST

	Images	Flickr Albums	(Text, Image)	Vocab
DII	-	-	151,800	13,800
DIS	-	-	151,800	5,000
SIS	-	-	252,900	18,200
Total	210,819	10,117	-	-

Table 40: Statistics of the SIND dataset.

be more human-like in understanding. Basic statistics of the dataset are shown in Table 41 while the splits of it can be found in Table 42.

Images	Text Sequences
81,743	10,117

Table 41: Statistics of the VIST (SIND v.2) dataset.

Split	Stories	Sentences
Training	40,155	200,775
Validation	4,990	24,950
Test	5,055	25,275

Table 42: Splits of the VIST dataset.

3.2.3 Image Storytelling - Evaluation Measures, Models, and Results

In this section, we review the measures used to evaluate different *Image Storytelling* models and the results obtained by them.

Evaluation Measures. To evaluate *Image Storytelling* models, the Language metrics and Retrieval metrics presented in Section 3.1.3 are used.

Models. Many models have been created in attempts to solve the *Image Storytelling* task. In Table 43, we present some exemplar architectures (refer to *Combined* column) created to address the task by integrating both image and language inputs. We also include a column that showcases the optimization techniques used to train those models.

Results. In Table 44, Table 45, Table 46, and Table 47 we present the results obtained with a subset of models which use the datasets presented earlier in Section 3.2.2.

3.2.4 Image Storytelling - Discussion

We observe that for *Image Storytelling*, the adversarial approach, i.e., Adversarial REward Learning (AREL) proposed by Wang et al. (2018a), achieves best results on both retrieval and language metrics for different datasets. This attests to AREL's ability to clone expert behaviors while still generating more human-like stories.

Approach	Image	Language	Combined	Optimizer	RL
(Kiros et al., 2014a)	AlexNet	LM	MLBL	-	✗
(Karpathy & Fei-Fei, 2015)	VGG	RNN	NeuralTalk	RMSprop	✗
(Vinyals et al., 2015)	GoogLeNet	LSTM	NIC	SGD	✗
(Park & Kim, 2015)	VGG	RNN	CRCN	RMSprop	✗
(Huang et al., 2016)	VGG	GRU	Story-Flat	-	✗
(Krause et al., 2017)	VGG	LSTM	HierarchicalRNN	ADAM	✗
(Liu et al., 2017)	VGG	LSTM	BARNN	-	✗
(Wang et al., 2018)	VGG	LSTM	GAN	ADAM	✓
(Wang et al., 2018a)	ResNet-152	GRU	AREL	ADAM	✓

Table 43: Exemplar *Image Storytelling* architectures.

Model	B-4	CIDEr	METEOR	R@1	R@5	MedRank
MLBL (Kiros et al., 2014a)	0.01	2.6	5.29	1.19	4.52	100.5
NeuralTalk (Karpathy & Fei-Fei, 2015)	0.00	0.5	1.34	0.48	2.86	120.5
NIC (Vinyals et al., 2015)	0.10	9.1	5.73	0.95	7.38	88.5
CRCN (Park & Kim, 2015)	**2.08**	30.9	7.69	11.67	31.19	14.00
Story-Flat (Huang et al., 2016)	-	-	7.37	-	-	-
HierarchialRNN (Krause et al., 2017)	-	-	6.07	-	-	-
BARNN (Liu et al., 2017)	-	**41.6**	-	**29.37**	**45.43**	**8**
AREL (Wang et al., 2018)	-	-	**8.39**	-	-	-

Table 44: Results of different models on the NYC-Storytelling dataset.

Model	B-4	CIDEr	METEOR	R@1	R@5	MedRank
MLBL (Kiros et al., 2014a)	0.01	3.4	4.99	1.02	4.08	62
NeuralTalk (Karpathy & Fei-Fei, 2015)	0.00	0.4	1.34	1.02	3.40	88
NIC (Vinyals et al., 2015)	0.07	10.0	4.51	2.83	10.38	61.5
CRCN (Park & Kim, 2015)	3.49	52.7	8.78	14.29	31.29	16
Story-Flat (Huang et al., 2016)	-	-	7.61	-	-	-
HierarchialRNN (Krause et al., 2017)	-	-	7.72	-	-	-
BARNN (Liu et al., 2017)	-	**54.1**	-	**35.01**	**49.07**	**6**
AREL (Wang et al., 2018)	-	-	**9.90**	-	-	-

Table 45: Results of various models on the Disneyland-Storytelling dataset.

Model	B-4	CIDEr	METEOR	R@1	R@5	MedRank
CRCN (Park & Kim, 2015)	-	-	-	9.87	28.74	21
Story-Flat (Huang et al., 2016)	3.50	6.84	10.25	-	-	-
HierarchialRNN (Krause et al., 2017)	3.7	6.51	9.97	-	-	-
AREL (Wang et al., 2018)	**5.16**	**11.35**	**12.32**	-	-	-

Table 46: Results of different models on the SIND dataset.

3.2.5 Video Storytelling - Introduction

In comparison to image storytelling, which only deals with a small sequence of images, the aim of video storytelling is to generate coherent and succinct stories for long videos.

Model	B-4	CIDEr	METEOR	R@1	R@5	MedRank
enc-attn-dec (Xu et al., 2015a)	-	4.96	32.98	-	-	-
h-attn-rank (Yu et al., 2017)	-	7.38	33.94	-	-	-
BARNN (Liu et al., 2017)	-	-	33.32	**24.07**	**44.29**	**9**
AREL-t-100 (Wang et al., 2018a)	14.1	**9.4**	**35.0**	-	-	-

Table 47: Results of various models on the VIST dataset.

However, video storytelling is less explored. The video storytelling task was pioneered by Li et al. (2020) to address challenges such as diversity in the story and the inherent complexity of video. They introduced residual Bidirectional RNNs (BiRNNs) for leveraging context and a narrator model with reinforcement learning. Further, Gella et al. (2018) created a multi-sentence video description dataset (VideoStory) to resemble stories from social media videos. The goal of social media-specific video description generation was to offer support to people with visual disabilities or other technical issues such as internet bandwidth limitations. Figure 5 illustrates the task of video storytelling where a story in a sequence is generated based on a video as the sole input.

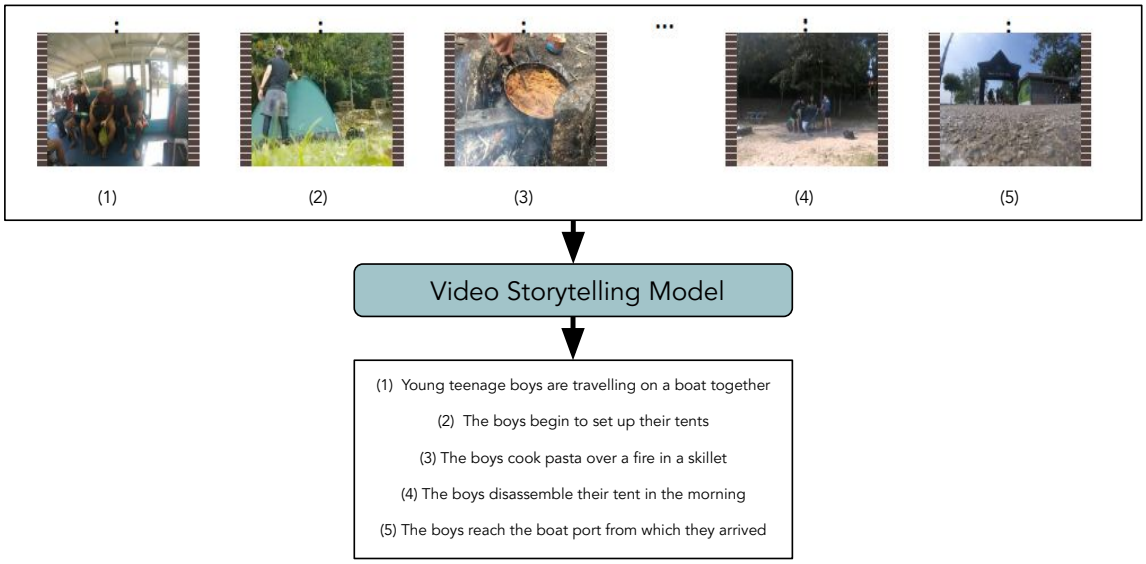

Figure 5: Given *video frames* (adopted from (Li et al., 2020)) as input, a Video Storytelling Model generates a textual story in sequence.

It is worth noting that this task bears close resemblance to the well-researched area of video summarization using only videos (Ma et al., 2002).

3.2.6 Video Storytelling - Datasets

Similar to *image storytelling* datasets, currently two different datasets are available to address the task of video storytelling. In the following, we elaborate on these two datasets.

VideoStory. VideoStory (Gella et al., 2018) is a multi-sentence description dataset created from social media videos that are selected to be highly diverse and engaging. Table 48 shows more statistics on the dataset.

Total Videos	Total Length	Total Clips	Avg. Video Duration	Total Sentences	Sentences per Video
20,000	396 h	123,000	70s	123,000	4.67

Table 48: Statistics of the VideoStory dataset.

Models can be evaluated locally on the earmarked *test* set whereas *test (blind)* is reserved for online evaluation purposes. However, the dataset including annotations has not been made public yet. Table 49 presents actual number of videos, clips, and sentence annotations for each of the splits.

Split	Videos	Clips	Paragraphs/video	Paragraphs	Words/paragraph
Training	17,098	80,598	1	17,098	61.76
Validation	999	13,796	3	2,997	59.88
Testing	1,011	14,093	3	3,033	59.77
Test (Blind)	1,039	14,139	3	3,117	69.45
Total	20,147	122,626	-	26,245	62.23

Table 49: Splits of the VideoStory dataset.

VideoStory-NUS. The VideoStory-NUS[35] (Li et al., 2020) dataset contains social event videos that were collected from YouTube by querying for common and complex events, namely *Birthday*, *Camping*, *Christmas*, and *Wedding*. Specifically, it comes with 105 manually chosen videos with sufficient inter-event and intra-event variations which are annotated with descriptive stories obtained through AMT. Each video is annotated by at least 5 different AMT workers, thus resulting in 529 stories in total. More statistics of the dataset can be found in Table 50.

Domain	Videos	Avg. Video Length	Avg. Story Length	Avg. Sentence Length	Vocab. Size
Open	105	12 m 35 s	162.6	12.1	4,045

Table 50: Statistics of the VideoStory-NUS dataset.

For experimental purposes, the dataset is randomly split in a ratio of 14:3:3 for training, validation, and testing respectively. Actual numbers are presented in Table 51.

3.2.7 Video Storytelling - Evaluation Measures, Models, and Results

In this section, we review the measures used to evaluate different *Video Storytelling* models and the results obtained by them.

[35] `https://zenodo.org/record/2383739`

Split	Percentage (%)	Videos
Training	70	73
Validation	15	16
Test	15	16

Table 51: Splits of the VideoStory-NUS dataset.

Evaluation Measures. To evaluate *Video Storytelling* models, the Language metrics and Retrieval metrics presented in Section 3.1.6 are used.

Models. There are a number of different models available for the task of *Video Storytelling*. These models combine representations of video and language in an efficient manner to address the task. In Table 52, we present some exemplar architectures (refer to *Combined* column) created to accomplish the task by integrating both video and language inputs. To understand the optimization techniques used, we also include a column that showcases the optimization method used to train the models.

Approach	Video	Frame	Language	Combined	Optimizer	RL
(Yu et al., 2016)	C3D	VGG	GRU	H-RNN	RMSProp	✗
(Gella et al., 2018)	R3D	ResNet-101	GRU	seq-seq+context	ADAM	✗
(Li et al., 2020)	-	ResNet-101	GRU	ResBRNN	ADAM	✓

Table 52: Exemplar *Video Storytelling* architectures.

Results. The *Video Storytelling* results showcases the efficacy of the proposed models. In Table 53 and Table 54 we present results obtained with a subset of models built using the datasets presented earlier in Section 3.2.6.

Model	B-4	CIDEr	METEOR	R@1	R@5	MedRank
seq-seq+context (Gella et al., 2018)	1.20	9.37	33.88	-	-	-

Table 53: Results obtained with different models on the VideoStory dataset.

Model	B-4	CIDEr	METEOR	R@1	R@5	MedRank
mRNN (Mao et al., 2015)	11.8	81.3	18.0	5.34	21.23	29
Deep Video-Text (Xu et al., 2015b)	11.5	79.5	17.7	4.72	19.85	31
H-RNN (Yu et al., 2016)	**16.1**	64.6	15.5	-	-	-
ResBRNN (Li et al., 2020)	14.7	94.3	19.6	**7.44**	**25.77**	**22**
ResBRNN-kNN (Li et al., 2020)	15.6	**103.6**	**20.1**	-	-	-

Table 54: Results obtained with different models on the VideoStory-NUS dataset.

3.2.8 Video Storytelling - Discussion

For *Video Storytelling*, a different set of methods are used for comparing two datasets. In Table 53, we observe that only one method utilizing the sequence-to-sequence paradigm

with contextual information (i.e., seq2seq+context) is evaluated on the "VideoStory" dataset. Nevertheless, another set of methods used for comparison for the "VideoStory-NUS" dataset is in Table 54. It shows that the approach proposed by Li et al. (2020) using Residual BRNN with k-Nearest Neighbours (i.e., ResBRNN-kNN) outperforms most of the baseline methods.

4. Visual Referring Expression Comprehension and Generation

In this section, we explore the task of *Visual Referring Expression Comprehension and Generation*. The objective of the task is to ground a natural language expression (e.g. a noun phrase or a longer piece of text) to objects in a visual input.

4.1 Image Referring Expression Comprehension and Generation

In the following, we provide a detailed description of the *Visual Referring Expression Comprehension and Generation* by using an image as the visual input.

4.1.1 Image Referring Expression Comprehension and Generation - Intro

In a natural environment, people use referring expressions to unambiguously identify, indicate, or point to particular objects. This is usually done with a simple phrase or within a larger context (e.g. a sentence). Having a larger context provides better scope for avoiding ambiguity and allows the referential expression to easily map to the target object. However, there can also be other possibilities in which people are asked to describe a target object based on its surrounding objects.

This provides us with two different possibilities for the visual referring expression task. In the first scenario, referring expressions deal with *generation*, in which an algorithm generates a referring expression for a given target object that is present in a visual scene. In the second scenario, the referring expression is used to perform comprehension, in which an algorithm locates in an image the object described by a given referring expression. Figure 6 shows an example for the task of referring expression comprehension.

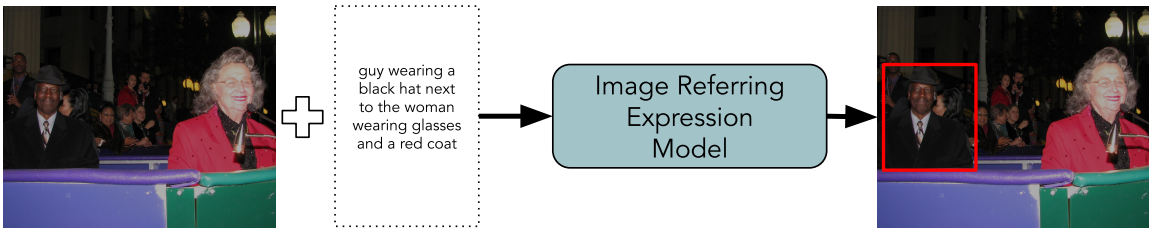

Figure 6: Given an *image* and a *referring expression*, an Image Referring Expression Comprehension Model identifies it in the image using bounding boxes.

Given these tasks, different approaches have been proposed for referring expression generation (Golland et al., 2010; Mitchell et al., 2013), comprehension (Kazemzadeh et al., 2014), and both combined (Mao et al., 2016; Yu et al., 2016). Note that there is a difference between referring expression tasks and grounding of free-form textual phrases (Rohrbach et al., 2016b) in an image.

Image Referring Expression Generation. An initial approach (FitzGerald et al., 2013) tackled the problem from the perspective of density estimation, in which the goal was to learn distributions over logical expressions identifying sets of objects in the world. Other research designed a comprehension-guided referring expression generator (Luo & Shakhnarovich, 2017) by using a comprehension module trained on human-generated expressions to generate referring expressions.

Image Referring Expression Comprehension. Nagaraja et al. (2016) investigated referring expression comprehension to integrate contexts between objects. Later on, techniques such as Multiple Instance Learning (MIL) were used to explore context regions and max-margin based MIL objective functions for training. Further, Hu et al. (2016) leveraged a natural language query of the object to localize a target object using a Spatial Context Recurrent Convnet (SCRC) model. It operates as a scoring function on candidate boxes for object retrieval, integrating spatial configurations and global scene-level contextual information. This explicit modeling of the referent and context region pairs has proven useful. Approaches such as compositional modular networks (Hu et al., 2017b) analyzed referential expressions by identifying entities and relationships mentioned in the input expression and grounding them all in the scene. Such an approach has been shown to effectively inspect local regions and pairwise interactions between them. A modular approach was also explored where three modular components related to subject appearance, location, and relationship to other objects was used to model with Modular Attention Network (Yu et al., 2018). It has proven effective at focusing on the subjects and their relationships. Approaches such has GroundNet (Cirik et al., 2018a) have leveraged syntactic analysis of the input referring expression to build a dynamic computation graph of neural modules that definesan architecture for performing localization. Variational models have also been used for referential expression comprehension where variational Bayesian methods called *variational context* (Zhang et al., 2018) were used to solve the problem of complex context modeling. These methods have proven capable of exploiting the relation between the referent and context, thereby reducing the search space of context. Furthermore, an accumulated attention mechanism (Deng et al., 2018) has been proposed to accumulate the attention for useful information in image, query, and objects. It has demonstrated the ability to reduce the redundancy and noise issues that were in other approaches.

Recently, a Cross-Modal Relationship Extractor (CMRE) and a Gated Graph Convolutional Network (GGCN) were combined into a cross-modal relationship inference network (Yang et al., 2019). CMRE has been shown to highlight objects and relationships which have connections with a given referring expression, while GGCN computes multimodal semantic contexts by fusing information from different modes and propagating multimodal information through the structured relation graph. Coming from a perspective of natural language understanding, a Recursive Grounding Tree (Hong et al., 2019) sought to automatically compose a binary tree structure by parsing the referring expression, in order to perform visual reasoning along the tree in a bottom-up fashion. It has been shown to allow gradients from continuous score functions with a discrete tree construction. There has also been interest in combining visual reasoning with referential expressions through the creation of new dataset (Liu et al., 2019). Most of the above approaches use bounding box

localization, but additionally object segmentation (Liu et al., 2017) has also been explored for referring expression comprehension.

Image Referring Expression Generation and Comprehension. Few approaches have performed both generation and comprehension tasks. Visual context (Mao et al., 2016; Yu et al., 2016) was initially used in referring expression models to find visual comparison to other objects within an image. It has shown significant improvements. Further, a unified framework (Yu et al., 2017a) was designed using a speaker, a listener, and a reinforcer. The *speaker* generates referring expressions, the *listener* comprehends referring expressions, and the *reinforcer* introduces a reward function to guide sampling of more discriminative expressions. Feedback from the discriminative reinforcer has proven capable of benefiting the tasks. The role of attributes (Liu et al., 2017) was also studied to show that they help in disambiguation when referring to a particular object.

4.1.2 Image Referring Expression Comprehension and Generation - Datasets

For the task of image referring expression, both *real* and *synthetic* image datasets have been designed. In the following, we present the details of the datasets in separate sections.

Real Images. In the real and natural images category, the ImageCLEF[36] and MSCOCO[2] (see Section 3.1.2) datasets are commonly used for creating referring expression annotations. From a subset of ImageCLEF's IAPR dataset[36], referring expressions are collected in a game-based setting, namely ReferItGame[38] (Kazemzadeh et al., 2014). The resulting dataset is called as RefCLEF[37] and its statistics can be found in Table 55.

Real Images	Distinct Objects	Referring Expressions	Train/Test Splits
19,894	96,654	130,525	Per-Image split

Table 55: Statistics of the RefCLEF dataset.

The RefCOCO[37], RefCOCO+[37] (Yu et al., 2016), and RefCOCOg (Mao et al., 2016) datasets were all created using MSCOCO images. For RefCOCO and RefCOCO+, the "People vs. Object" split evaluates images containing multiple people (Test A) and images containing multiple instances of all other objects (Test B). Both RefCOCO and RefCOCO+ were collected in the same interactive setting as above, ReferItGame[38] (Kazemzadeh et al., 2014). Table 56 presents the statistics of the RefCOCO dataset whereas Table 57 shows the statistics of the RefCOCO+ dataset.

Images	Total Objects	Referring Expressions	Train/Test Splits
19,994	50,000	142,209	People vs. Object

Table 56: Statistics of the RefCOCO dataset.

[36] https://www.imageclef.org/SIAPRdata
[37] https://github.com/lichengunc/refer
[38] http://tamaraberg.com/referitgame

One important distinction between the RefCOCO and RefCOCO+ datasets is that the latter was collected in a comparatively restrictive setting when compared to the former. Specifically, the usage of location words was not permitted in the referring expressions in case of RefCOCO+ whereas there was no such restriction on the language for RefCOCO.

Images	Total Objects	Referring Expressions	Train/Test Splits
19,992	49,856	141,564	People vs. Object

Table 57: Statistics of the RefCOCO+ dataset.

To overcome some of the limitations of RefCLEF, a dataset based on based on MSCOCO[2] was created. This dataset, known as RefCOCOg[39] (Mao et al., 2016), contains much longer sentences and was collected in a non-interactive setting using AMT, in contrast to the interactive setting used with RefCLEF, RefCOCO, and RefCOCO+. The statistics of this dataset is presented in Table 58.

Images	Total Objects	Referring Expressions	Train/Test Splits
26,711	54,822	85,474	Per-Object

Table 58: Statistics of the RefCOCOg dataset.

Earlier mentioned referring expression datasets use single sentences for image referring expression. In contrast, the GuessWhat[40] (de Vries et al., 2017) dataset was created with a cooperative two-player guessing game, the goal of which was to locate an unknown object in an image (collected from MSCOCO) by asking a sequence of questions. Hence, it creates multiple sentences (i.e., a dialog) for a given image in order to perform referring expression. Another notable aspect of this dataset is that only images containing a number of objects in the range of 3 to 20 are chosen from MSCOCO. The dialogue collection was achieved via crowdsourcing using AMT. For evaluation, the dataset is randomly split into 70% for training, 15% for validation, and 15% for testing. Table 59 presents more details about the dataset.

Dataset Type	Images	Objects	Dialogues	Questions	Words	Vocab. Size
Full	66,537	134,073	155,280	821,889	3,986,192	11,465
Finished	65,112	125,349	144,434	732,081	3,540,497	10,985
Success	62,954	114,271	131,394	648,493	3,125,219	10,469

Table 59: Statistics of "GuessWhat" dataset. The row 'Full' means all the dialogues are included, 'Finished' means all finished dialogues (successful and unsuccessful) are included, and 'Success' means only successful dialogues are included.

[39] https://github.com/mjhucla/Google_Refexp_toolbox
[40] https://github.com/GuessWhatGame/guesswhat

Synthetic Images. In the synthetic category, the CLEVR-Ref+[41] (Liu et al., 2019) dataset was introduced to address issues such as bias in datasets with real images, since it has been recently been shown that referring expression models suffer from unintended biases (Cirik et al., 2018b). CLEVR-Ref+ reuses the images from the CLEVR dataset (see Section 5.2.2), while replacing the questions in CLEVR with referring expressions and answers with referred objects. The main purpose of CLEVR-Ref+ is to diagnose image reasoning with referring expressions by exercising the desired control over the nature of samples. Table 60 present splits of the dataset.

Split	Images	Referring Expressions
Training	70,000	700,000
Validation	15,000	150,000
Test	15,000	150,000

Table 60: Splits of the CLEVR-Ref+ dataset.

4.1.3 Image Referring Expression Comprehension and Generation - Evaluation Measures, Models, and Results

In this section, we review the measures used to evaluate different *Image Referring Expression* models and the results achieved by them.

Evaluation Measures. The measure that is usually used for the evaluation of *Image Referring Expression* models is Precision@1, i.e., precision calculated with the Intersection over Union (IoU) ratio between the true and predicted bounding box.

Models. The models designed to approach the task of *Image Referring Expression* provide an effective way to optimize the Precision@1 measure by identifying the right object in a visual input which matches the textual phrase. In Table 61, we present some exemplar architectures (refer to *Combined* column) created to address the task by integrating both image and language inputs. We also include a column that showcases the optimization techniques used to train those models.

Results. Several models and datasets have been created to address the task of *Image Referring Expression*. These datasets provide variety in the content so that they enhance the generalization ability of the models. In this section, we cover the results obtained by the models on some representative datasets. Table 62 and Table 63 presents results obtained with a subset of models built using the datasets such as RefCOCO, RefCOCO+, and RefCOCOg presented in Section 4.1.2.

4.1.4 Image Referring Expression Comprehension and Generation - Discussion

For *Image Referring Expression*, on all MSCOCO based datasets (i.e., RefCOCO, RefCOCO+, and RefCOCOg) the technique proposed by Yang et al. (2019) outperforms existing baselines. This approach builds a Cross-Modal Relationship Extractor (CMRE) to

[41] https://cs.jhu.edu/~cxliu/2019/clevr-ref+.html

Approach	Image	Language	Combined	Optimizer	RL
(Mao et al., 2016)	VGG	LSTM	MMI	SGD	✗
(Nagaraja et al., 2016)	VGG	LSTM	Neg. Bag	SGD	✗
(Yu et al., 2016)	VGG	LSTM	Context	-	✗
(Luo & Shakhnarovich, 2017)	VGG	BiLSTM	CG	ADAM	✗
(Liu et al., 2017)	VGG	LSTM	Combined	ADAM	✗
(Hu et al., 2017b)	VGG	LSTM	CMN	-	✗
(Yu et al., 2017a)	VGG	LSTM	Reinforcer	ADAM	✓
(Zhang et al., 2018)	VGG	BiLSTM	VarContext	SGD	✓
(Deng et al., 2018)	VGG	LSTM	AccumulateAtt	SGD	✗
(Zhuang et al., 2018)	VGG	LSTM	ParallelAtt	ADAM	✗
(Yu et al., 2018)	ResNet-101	BiLSTM	MAttNet	-	✗
(Hong et al., 2019)	ResNet-101	BiLSTM	RVG-Tree	ADAM	✗
(Yang et al., 2019)	ResNet-101	BiLSTM	CMRIN	ADAM	✗

Table 61: Exemplar *Image Referring Expression and Comprehension* architectures.

Model	RefCOCO		
	val	testA	testB
MMI (Mao et al., 2016)	-	63.15	64.21
Neg. Bag (Nagaraja et al., 2016)	76.90	75.60	78.00
Context (Yu et al., 2016)	76.18	74.39	77.30
CG (Luo & Shakhnarovich, 2017)	-	74.04	73.43
Attributes (Liu et al., 2017)	-	78.85	78.07
CMN (Hu et al., 2017b)	-	75.94	79.57
Reinforcer (Yu et al., 2017a)	79.56	78.95	80.22
VarContext (Zhang et al., 2018)	-	78.98	82.39
AccumulateAtt (Deng et al., 2018)	81.27	81.17	80.01
ParallelAtt (Zhuang et al., 2018)	81.67	80.81	81.32
MAttNet+ResNet-101 (Yu et al., 2018)	85.65	85.26	84.57
RVG-Tree+ResNet-101 (Hong et al., 2019)	83.48	82.52	82.90
CMRIN+ResNet-101 (Yang et al., 2019)	**86.99**	**87.63**	**84.73**

Table 62: Comparison of Precision@1 (%) scores of different methods on RefCOCO.

highlight objects and their relationships. Furthermore, a Gated Graph Convolutional Network (GGCN) is used to compute multimodal semantic contexts by fusing information from different modes and propagating multimodal information. This Cross-Modal Relationship Inference Network (CMRIN) along with ResNet-101 visual features have been shown to achieve the best results.

4.2 Video Referring Expression Comprehension and Generation

In the following, we describe the setting of *Visual Referring Expression Comprehension and Generation* task when a video is used as the visual input.

Model	RefCOCO+			RefCOCOg	
	val	testA	testB	val	test
MMI (Mao et al., 2016)	-	48.73	42.13	-	-
Neg Bag (Nagaraja et al., 2016)	-	-	-	-	68.40
Context (Yu et al., 2016)	58.94	61.29	56.24	-	-
CG (Luo & Shakhnarovich, 2017)	-	60.26	55.03	-	-
Attributes (Liu et al., 2017)	-	61.47	57.22	-	-
CMN (Hu et al., 2017b)	-	59.29	59.34	-	-
Reinforcer (Yu et al., 2017a)	62.26	64.60	59.62	71.65	71.92
VariationalContext (Zhang et al., 2018)	-	62.56	62.90	-	-
AccumulateAttn (Deng et al., 2018)	65.56	68.76	60.63	-	-
ParallelAttn (Zhuang et al., 2018)	64.18	66.31	61.46	-	-
MAttNet+ResNet-101 (Yu et al., 2018)	71.01	75.13	66.17	78.10	78.12
RVG-Tree+ResNet-101 (Hong et al., 2019)	68.86	70.21	65.49	76.82	75.20
CMRIN+ResNet-101 (Yang et al., 2019)	**75.52**	**80.93**	**68.99**	**80.45**	**80.66**

Table 63: Comparison of Precision@1 (%) scores of different methods on the RefCOCO+ and RefCOCOg datasets.

4.2.1 Video Referring Expression Comprehension and Generation - Intro

When compared to image referring expression comprehension and generation, the task of video referring expression comprehension and generation is less explored at the time of publication of this survey. Although, there has been a surge in interest in tackling the spatio-temporal contexts and motion features that are inherent to videos, most of the works thus far, however, have concentrated on only one variant of image referring expression, namely comprehension. Vasudevan et al. (2018) used stereo videos to exploit richer and more realistic temporal-spatial contextual information along with gaze cues for referring expression comprehension. Figure 7 shows an example of the video referring expression comprehension. Another approach by Khoreva et al. (2018) explored Language Referring Expressions to

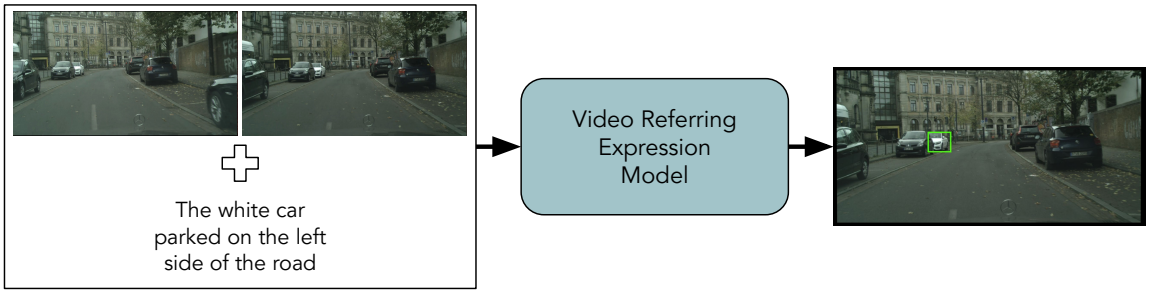

Figure 7: Given a *video* (represented as a sequence of frames from Vasudevan et al. (2018)) and a *referring expression*, a Referring Expression Comprehension Model identifies it in the video using bounding boxes.

point to the objects in the video to achieve object segmentation. Slightly different from

the described task, Wang et al. (2020) proposed an end-to-end boundary-aware model for video grounding. The model uses a lightweight branch to predict semantic boundaries corresponding to the given linguistic information. It aggregates contextual information by explicitly modeling the relationship between the current element and its neighbours.

4.2.2 Video Referring Expression Comprehension and Generation - Datasets

In this section, we present the datasets used to evaluate the task of *Video Referring Expression Comprehension.*

Object Referring in videos with Gaze (ORGaze). For performing Video Referring Expression, the Cityscapes[42] dataset containing a diverse set of stereo video sequences recorded in street scenes has been modified to have gaze information. Therefore, the ORGaze[43] (Vasudevan et al., 2018) dataset contains object referring in videos with language and human gaze. More details of the dataset is presented in Table 64.

Videos	Objects	Condition	Lighting	Annotations
5,000	30,000	Urban	Daytime	Bounding Boxes Gaze Recordings Language Expression

Table 64: Statistics of the ORGaze dataset.

The authors split the cities in the training set of Cityscapes for training and validation while using all the cities in validation set of Cityscapes for testing purposes. More concretely, the validation set is constructed by selecting one city (e.g., Zürich) from Cityscapes training set while leaving the rest of the cities as part of the training set. For constructing the test set, the videos from all the cities in Cityscapes validation set (e.g., Frankfurt, Lindau, Münster) of Cityscapes are used. Of the total 30,000 annotated objects, 80% has been used for *training* and the remaining 20% was reserved for model evaluation of the task.

4.2.3 Video Referring Expression Comprehension and Generation - Evaluation Measures, Models, and Results

In this section, we review the evaluation measures used to benchmark different *Video Referring Expression Comprehension* models and the results achieved by them.

Evaluation Measures. The measure that is used for the evaluation of *Video Referring Expression Comprehension* model is "Top-1 Accuracy" and also object proposal accuracy referred with Language-based Object Proposals (LOP), Faster R-CNN (FRCNN), and Edge-Box (Zitnick & Dollár, 2014).

Models. Many models have been created to solve the task of *Video Referring Expression Comprehension.* In Table 65, we present some exemplar architectures (refer to *Combined* column) created to address the task by integrating both video and language. We also include a column that showcases the optimization techniques used to train those models.

[42] https://www.cityscapes-dataset.com

[43] https://people.ee.ethz.ch/~arunv/ORGaze.html

Approach	Video	Frame	Language	Combined	Optimizer	RL
(Vasudevan et al., 2018)	-	VGG	LSTM	WithGaze	-	✗

Table 65: Exemplar *Video Referring Expression and Comprehension* architectures.

Results. As discussed earlier, several models have been created to approach the task of *Video Referring Expression Comprehension*. In Table 66 we present results obtained with a subset of models built using the ORGaze dataset presented earlier in Section 4.2.2.

Methods	Edgebox	FRCNN (↑)	LOP (↑)
MNLM (Kiros et al., 2014b)	-	23.954	32.418
VSEM (Liu et al., 2015)	-	24.833	32.961
MCB (Fukui et al., 2016)	-	26.445	33.366
SimModel (Plummer et al., 2017a)	4.5	18.431	35.556
WithGaze (Vasudevan et al., 2018)	-	**47.256**	**47.012**

Table 66: Comparison of Top-1 Accuracy (%) of different methods on the ORGaze dataset.

4.2.4 Video Referring Expression Comprehension and Generation - Discussion

The *Video Referring Expression Comprehension* task is benchmarked using a single dataset. Evaluated using different task-specific metrics, the approach proposed by Vasudevan et al. (2018), which uses the gaze information, produces the best results.

5. Visual Question Answering, Reasoning, and Entailment

In this section, we explore three different tasks, namely, *Visual Question Answering*, *Visual Reasoning*, and *Visual Entailment*. The goal of each of these tasks are different. However, they share the common intention of answering questions when conditioned on a visual input. In the following sections, we elaborate on each of these three tasks separately.

5.1 Visual Question Answering

The goal of *Visual Question Answering* (VQA) is to learn a model that comprehends visual content at both the global and local level for finding an association with pairs of questions and answers in the natural language form. The visual information for VQA includes both images and videos.

5.1.1 Image Question Answering - Introduction

The aim of *Image Question Answering* (Image Q&A) is to answer natural language questions about the contents of images. Earlier research efforts have focused on designing different algorithms and constructing datasets to address this challenge. The first approaches (Malinowski & Fritz, 2014; Malinowski et al., 2015; Geman et al., 2015) considered Image Q&A as a Visual Turing Test, where the expectation was to incorporate human-level abilities for

semantically accessing the visual information to answer different questions. These were then improved as fill-in-the-blank tasks (Yu et al., 2015), where the goal of the system was focused on multiple-choice question-answering for images. Also, it was expanded to address both multilingual (Gao et al., 2015) and automatic question generation, in which descriptions of sentences are converted into questions (Ren et al., 2015a). However, it lacked natural language questioning ability of humans. Hence, a broader task was proposed with an aim of addressing open-ended Image Q&A (Antol et al., 2015; Agrawal et al., 2017), where the challenge was to ask a free-form natural language question about an image and make the system to answer the question. Figure 8 provides a schematic representation of the task where a free-form question about the contents of an image is asked to obtain an answer.

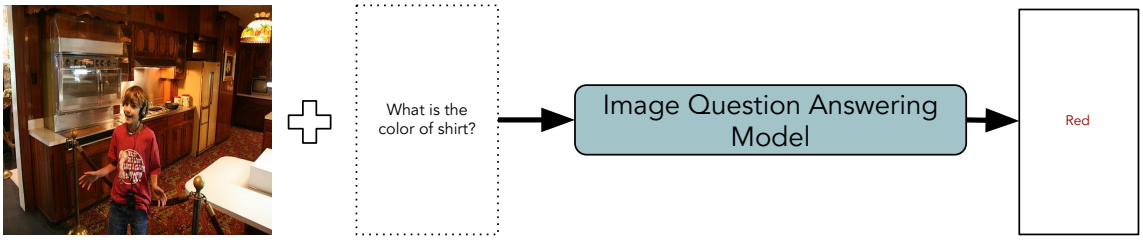

Figure 8: Given an *image* and a *question* about the image, an Image Question Answering model produces an answer to it.

However, designing such a system can contain several other challenges, such as coming up with strong baselines (Jabri et al., 2016). To address these, binary image Q&A (Zhang et al., 2016) was explored by providing complementary images for abstract scenes. These complementary images were used to provide visual verification of concepts contained in the questions. Some of the questions were understood as a loose, global association between Q&A sentences and images. Hence, more confined and dedicated tasks were created for relating local regions in the images (Zhu et al., 2016) by addressing object-level grounding. Some approaches (Zhang et al., 2018) concentrated only on counting objects in natural images. There are many methods that are proposed to address the challenging image Q&A task. The details about different methods are already covered in earlier surveys (Kafle & Kanan, 2017; Wu et al., 2017). Therefore, we briefly present new methods that were introduced after the publication of these surveys.

Recent works aim at interpretability or explainability by overcoming priors (Agrawal et al., 2018), concentrating better on the image to extract relevant information (Goyal et al., 2019), generating human-interpretable rules that provide better insights (Manjunatha et al., 2019), and cycle-consistency (Shah et al., 2019a), while other works try to understand the text inside an image to answer and reason about it (Singh et al., 2019). More recent works sought to incorporate outside knowledge (Marino et al., 2019) in the image Q&A framework to support real-world knowledge-aware question answering (Shah et al., 2019b).

There are different kinds of learning approaches used for image Q&A, such as Multi-task learning and Federated learning. A multi-task learning approach (Nguyen & Okatani, 2019) is used to learn a vision-language representation that is shared by many tasks from their diverse datasets to address image Q&A. In contrast, federated learning is used with the aimNet (Liu et al., 2020) and is validated on federated learning settings that include both

horizontal and vertical federated learning. To focus on language priors, a modular language attention mechanism is used by Jing et al. (2020) to parse a question into three phrase representations, namely type representation, object representation, and concept representation. It has prevented language priors from dominating the answering process.

5.1.2 Image Question Answering - Datasets

Several datasets were created in the past decade to address the challenge of image question answering. In the following, we cover the datasets that are extensively used for this Human-Computer Interaction (HCI) themed task.

VQA v1.0. VQA v1.0[44] (Antol et al., 2015) contains open-ended questions about images. These questions target different areas of an image, including background details and the underlying contexts. The answers are also open-ended and contain either a few words or a closed set of answers that can be provided in a multiple-choice format. Table 67 and Table 68 present the dataset splits of images with *real* and *abstract* scenes observed in the dataset respectively.

Dataset Split	Real Scenes	Questions per Image	Answers per Question	**Textual Annotations** Questions	Answers
Training	82,783	3	10	248,349	2,483,490
Validation	40,504	3	10	121,512	1,215,120
Test	81,434	3	10	244,302	2,443,020

Table 67: Splits of the VQA v1.0 dataset with *real* scenes.

Dataset Split	Abstract Scenes	Questions per Image	Answers per Question	**Textual Annotations** Questions	Answers
Training	20,000	3	10	60,000	600,000
Validation	10,000	3	10	30,000	300,000
Test	20,000	3	10	60,000	600,000

Table 68: Splits of the VQA v1.0 dataset with *abstract* scenes.

VQA v2.0. VQA v2.0 extends VQA v1.0 and has three parts: *Balanced Real Images*, *Balanced Binary Abstract Scenes*, and *Abstract Scenes*. Table 69 and Table 70 presents the dataset splits of the images with balanced real and binary abstract scenes observed in the dataset respectively. However, abstract scenes in VQA v2.0 are same as that of VQA v1.0.

The term *complementary pairs* in Table 69 means that a given question is associated with a pair of similar images such that the answer is different depending on the image (i.e. two different answers)

Outside Knowledge VQA (OK-VQA). OK-VQA[45] (Marino et al., 2019) uses a subset of MSCOCO (see Section 3.1.2) and is constructed with additional annotations such as

[44] https://visualqa.org
[45] https://okvqa.allenai.org

Dataset Split	Real Images	Answers per Question	Textual Annotations		
			Questions	Answers	Complementary Pairs
Training	82,783	10	443,757	4,437,570	200,394
Validation	40,504	10	214,354	2,143,540	95,144
Test	81,434	10	447,793	4,477,930	-

Table 69: Splits of the VQA v2.0 dataset with balanced real images.

Dataset Split	Binary Abstract Scenes	Answers per Question	Textual Annotations	
			Questions	Answers
Training	20,629	10	22,055	220,550
Validation	10,696	10	11,328	113,280

Table 70: Splits of VQA v2.0 with balanced binary abstract scenes.

questions, answers, knowledge category, etc. Table 71 presents more details about the dataset, while the Table 72 shows the splits of it.

Total Images	Total Questions	Answers per Question	Unique Questions	Unique Answers	Unique Ques. Words	Total Categories	Average Ans. Length
14,031	14,055	5	12,591	14,454	7,178	10 + 1	1.3

Table 71: Statistics of the OK-VQA dataset.

Split	Percent (%)	Questions
Training	64	9,009
Test	36	5,046
Total	100	14,055

Table 72: Splits of the OK-VQA dataset.

Knowledge-aware VQA (KVQA). The KVQA[46] (Shah et al., 2019b) dataset was designed to emphasize questions that require access to external knowledge. Table 73 presents more details about the dataset, while Table 72 shows the splits of it. In order to get a mean score, the KVQA dataset provides five such splits.

Total Images	Q&A Pairs	Unique Named Entities	Unique Answers	Avg. Ques. Len	Avg. Ans. Len	Avg. number of Questions per Image
24,602	183,007	18,880	19,571	10.14	1.64	7.44

Table 73: Statistics of the KVQA dataset.

[46] http://malllabiisc.github.io/resources/kvqa

Split	Percent (%)	Images	Q&A pairs
Training	70	17k	130k
Validation	20	5k	34k
Test	10	2k	19k

Table 74: Splits of the KVQA dataset.

5.1.3 Image Question Answering - Evaluation Measures, Models, and Results

In this section we describe only the evaluation measures used for *Image Question Answering* as **Models**, **Results**, and some **Discussion** are extensively presented in the recent surveys (Wu et al., 2017).

Evaluation Measures Image Q&A models are evaluated based on the Accuracy measure.

5.1.4 Video Question Answering - Introduction

The goal of *Video Question Answering* (Video Q&A) is to answer natural language questions about videos. Unlike Image Q&A, Video Q&A is less explored. Nevertheless, there are a few works which have explored this spatio-temporal domain. One of the early attempts in this domain was jointly parsing the videos with corresponding text to answer queries (Tu et al., 2014). Further, an open-ended Movie Q&A (Tapaswi et al., 2016) with multiple-choice question pairs was designed to solve challenging questions that require semantic reasoning over a long temporal domain. Additionally, to limit the involvement of crowdworkers, the task was modified using fill-in-the-blank questions (Zhu et al., 2017; Mazaheri et al., 2017) and were automatically generated from different manually created video description datasets (Section 3.1.5). Other works (Zeng et al., 2017) modified this dataset to support answering free-form natural language questions. Beyond this, open-ended video question answering is also addressed with methods such as spatio-temporal attentional encoder-decoder learning framework (Zhao et al., 2017). There has been interest shown in jointly addressing multiple tasks that handle video and language. High-level concept words (Yu et al., 2017b) are detected in order to be integrated with any video and language models addressing fill-in-the blank and multiple-choice test. Spatio-temporal reasoning from videos to answer questions has also been addressed by designing a spatial and temporal attention mechanism (Jang et al., 2017).

Recently, due to large interest in Video Q&A, similar to Movie Q&A, six popular TV shows were used to create a dataset, where questions are compositional (Lei et al., 2018). The TV Q&A dataset made the proposed multi-stream models to jointly localize relevant moments within a clip, comprehend subtitle-based dialogue, and then recognize relevant visual concepts. Furthermore, spatio-temporal grounding (Lei et al., 2020) is employed to link depicted objects to visual concepts in questions and answers. Figure 9 shows an example of this task, in which the model is given a video and a question and is asked to choose an answer from multiple choices.

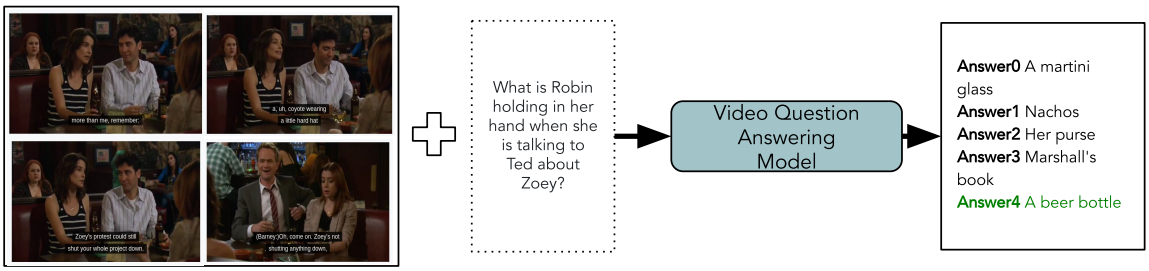

Figure 9: Given a *video* (represented as sequence of frames from TV Q&A dataset) and *question*, a Video Question Answering model finds the right answer from Multiple Options.

5.1.5 Video Question Answering - Datasets

Similar to image question answering, several datasets were created to address the challenge of video question answering. In the following, we cover those datasets that are popular and extensively used.

MovieQA. The MovieQA[47] (Tapaswi et al., 2016) dataset is used to evaluate story comprehension of both video and text in an automatic manner. The dataset consists of almost 15,000 multiple choice questions and answers obtained from over 400 movies having high diversity. Table 75 reports the statistics and splits of the dataset.

	Training	Validation	Test	Total
Movies with Plots and Subtitles				
Movies	269	56	83	408
QA pairs	9848	1958	3138	14944
Q words	9.3	9.3	9.5	9.3 ± 3.5
CA. words	5.7	5.4	5.4	5.6 ± 4.1
Movies with Video Clips				
Movies	93	21	26	140
QA pairs	4318	886	1258	6462
Video clips	4385	1098	1288	6771
Mean clip Length	201.0 s	198.5 s	211.4s	202.7 ± 216.2 s
Mean QA shots	45.6	49.0	46.6	46.3 ± 57.1

Table 75: Statistics & Splits of the MovieQA dataset. The column 'Total' represents mean counts with standard deviations.

TVQA. The TVQA[48] (Lei et al., 2018) dataset was created from videos of six different English TV shows, viz. *Friends*, *The Big Bang Theory*, *How I Met Your Mother*, *House M.D.*, *Grey's Anatomy*, and *Castle*. It consists of 460 hours of video and the questions are designed to be compositional, expecting the models to comprehend subtitles-based dialogue and to recognize relevant visual concepts. Table 76 presents the statistics of the dataset, while Table 77 shows the splits.

[47] http://movieqa.cs.toronto.edu/home
[48] http://tvqa.cs.unc.edu

Video Clips	Video Clip Length	Q&A Pairs	Total Duration	Questions per Video Clip	Answers per Video Clip
21,793	60 to 90 s	152,545	460 h	7	5

Table 76: Statistics of the TVQA dataset.

The testing data of TVQA is further split into two subsets named "test-public" containing 7,623 Q&A pairs and "test-reserved" consisting of 7,630 Q&A pairs. The *test-public* set is available for the TVQA leaderboard[49] whereas *test-reserved* is preserved for future use.

Split	Percent (%)	Q&A pairs
Training	80	122,039
Validation	10	15,253
Test	10	15,253

Table 77: Splits of the TVQA dataset.

The TVQA+[50] (Lei et al., 2020) is an augmented subset of the original TVQA dataset where the augmentation comes in the form of bounding boxes linking depicted objects to visual concepts in both questions and answers. Table 78 presents the splits of TVQA+ dataset.

Split	Q&As	Clips	Avg. Span Length (s)	Avg. Video Length (s)	Annotated Images	Bound. Boxes	Categories
Training	23,545	3,364	7.20	61.49	118,930	249,236	2,281
Validation	3,017	431	7.26	61.48	15,350	32,682	769
Test	2,821	403	7.18	61.48	14,188	28,908	680
Total	29,383	4,198	7.20	61.49	148,468	310,826	2,527

Table 78: Splits of the TVQA+ dataset.

5.1.6 Video Question Answering - Evaluation Measures, Models and Results

In this section, we present the evaluation measures, models, and results achieved with various architectures of Video Q&A.

Evaluation Measures. Video Q&A models are evaluated based on Accuracy. In addition, other measures such as Temporal mean Intersection-over-Union (Temp. mIoU) (Hendricks et al., 2017), Answer-Span joint Accuracy (ASA), that jointly evaluates both answer prediction and span prediction, and object grounding performance calculated with mean Average Precision (Grd. mAP) (Lei et al., 2020) are used.

Models. The models which are created to address the task of *Video Question Answering* aim to provide an overall understanding of the visual and the aligned textual content such as subtitles. In Table 79, we present some exemplar architectures (refer to *Combined* column)

[49] http://tvqa.cs.unc.edu/leaderboard.html
[50] http://tvqa.cs.unc.edu/download_tvqa_plus.html

created to address the task by integrating both video and language. We also include a column that showcases the optimization techniques used to train those models.

Approach	Video	Frame	Language	Combined	Optimizer	RL
(Jang et al., 2017)	C3D	ResNet-152	LSTM	ST-VQA	ADAM	✗
(Lei et al., 2018)	-	R-CNN+ResNet-101	BiLSTM	Two-stream	-	✗
(Lei et al., 2020)	-	R-CNN+ResNet-101	BERT	STAGE	ADAM	✗

Table 79: Exemplar *Video Question Answering* architectures.

Results. Several models have been created to approach the task of *Video Question Answering*. At the same time, many datasets have been created to provide diversity in the content so that they boost the generalization ability of the models. In this section, we cover the results achieved by the models on some representative datasets. Table 80 and Table 81 presents results obtained with a subset of models built using the TVQA and TVQA+ datasets presented in Section 5.1.5. Results for TVQA[51] and TVQA+[52] can also be found on the respective leaderboards.

Model	Accuracy (↑)
Random	20.00
Retrieval-SkipThought	24.77
Longest Answer	30.22
NNS-SkipThought (Subtitle)	38.29
NNS-TFIDF (Subtitle)	50.79
Two-stream (Subtitle+Videos) (Lei et al., 2018)	66.36
Three-stream (Subtitle+Videos+Questions) (Lei et al., 2018)	**68.48**

Table 80: Accuracy attained on TVQA test (public) set. All models use timestamp annotation without which the scores achieved by them are lower.

Model	Accuracy	Grd. mAP (↑)	Temp. mIOU (↑)	ASA (↑)
ST-VQA (Jang et al., 2017)	48.28	-	-	-
Two-stream (Lei et al., 2018)	68.13	-	-	-
STAGE-LXMERT (Lei et al., 2020)	71.46	21.01	26.31	18.04
STAGE (Lei et al., 2020)	**74.83**	**27.34**	**32.49**	**22.23**
Human (Lei et al., 2020)	90.46	-	-	-

Table 81: Results obtained on TVQA+ test set.

5.1.7 Video Question Answering - Discussion

It has been observed from STAGE (Lei et al., 2020) that aligned fusion is essential for improving Video Q&A performance. STAGE uses all of the existing information such as

[51] http://tvqa.cs.unc.edu/leaderboard.html

[52] https://competitions.codalab.org/competitions/22705#results

Subtitles, Video, and Questions to build an efficient model. It has also proven to be effective if the models have access to the timestamp information as shown in Table 80.

5.2 Visual Reasoning

The goal in visual reasoning is to learn a model that comprehends the visual content by reasoning about it. Both images and videos are used as visual inputs for visual reasoning. In the following, we provide a detailed description of this complex and challenging task.

5.2.1 Image Reasoning - Introduction

The goal of image reasoning is to answer sophisticated queries by reasoning about the visual world. Initial efforts (Johnson et al., 2017a) aimed at designing diagnostic tests going beyond benchmarks such as VQA. They reduced the biases by having detailed annotations describing the kind of reasoning each question requires. It has also been observed that VQA models struggle when comparing the attributes of objects, or when novel attribute combinations needs to be recognized (such as in compositional reasoning). A novel approach (Johnson et al., 2017b) used a program generator to construct an explicit representation of the reasoning process, and an execution engine to execute the resulting program, producing an answer. Then, end-to-end module networks (Hu et al., 2017) were proposed which learn to reason by directly predicting instance-specific network layouts without the aid of a parser as used in neural module networks. Santoro et al. (2017) went beyond and proposed Relation Networks (RNs) as a simple plug-and-play module to solve the problem of visual reasoning. RNs are further used to learn relation-aware visual features for content based image retrieval (Messina et al., 2018) and also Multi-Relational Networks (Chang et al., 2018). Furthermore, global context reasoning (Cao et al., 2018) is explored for better aligning image and language domains in diverse and unrestricted cases.

A recent approach (Perez et al., 2018) introduced a general-purpose conditioning method called Feature-wise Linear Modulation (FiLM) layers which influence neural network computation via a simple, feature-wise affine transformation based on conditioning information. FiLM was modified by Strub et al. (2018) to generate parameters of FiLM layers going up the hierarchy of a convolutional network in a multi-hop fashion rather than all at once. Cascaded Mutual Modulation (CMM) (Yao et al., 2018) is an end-to-end visual reasoning model that also uses the FiLM technique to enable the textual/visual pipeline to mutually control each other. Another approach modified neural modular networks (Hu et al., 2018) such that it performs compositional reasoning by automatically inducing a desired sub-task decomposition without relying on strong supervision. Mascharka et al. (2018) proposed a set of visual-reasoning primitives which, when composed, manifest as a model capable of performing complex reasoning tasks in an explicitly interpretable manner. Also, in the context of interpretable learning frameworks, Learning-By-Asking (LBA) (Misra et al., 2018b) attempted to closely mimic natural learning with the goal to make it more data efficient than the traditional VQA setting. Further, compositional attention networks (Hudson & Manning, 2018) were designed as fully differentiable neural network architectures to facilitate explicit and expressive reasoning. The goal of this architecture is to provide a strong prior for iterative reasoning, allowing it to support structured learning, as well as to generalize from a modest amount of data.

Recently, neural-symbolic visual question answering (Yi et al., 2018) attempted to combine deep representation learning with symbolic program execution. It first recovers structural scene representation from the image and a program trace from the question. This was extended with a Neuro-Symbolic Concept Learner (NS-CL) (Mao et al., 2019) that learns visual concepts, words, and semantic parsing of sentences without explicit supervision. It learns by simply looking at images and reading paired questions and answers. Further, a multimodal relational network (MuRel) (Cadène et al., 2019) was proposed to learn end-to-end reasoning over real images. Additionally, Aditya et al. (2019) used spatial knowledge to aid visual reasoning. Their framework combined knowledge distillation, relational reasoning, and probabilistic logical languages. Existing diagnostic tests have been further modified with referring expressions to handle bias (Liu et al., 2019) and with structural, relational, and analogical reasoning in a hierarchical representation (Zhang et al., 2019). Explainable and explicit neural modules (Shi et al., 2019) have also been explored with scene graphs. Objects as nodes and pairwise relationships as edges were used for explainable and explicit reasoning with structured knowledge.

Further expanding the scope of inquiry on this subject, Andreas et al. (2016a, 2016b) exploit the compositional linguistic structure of complex questions by forming neural module networks which query about the abstract shapes observed in an image. Improvement is further seen in how images are interpreted. For example, compositional question answering (Hudson & Manning, 2019) was addressed with scene graph structures on real-world images going beyond abstract shapes. Figure 10 demonstrates the task of reasoning about real-world images.

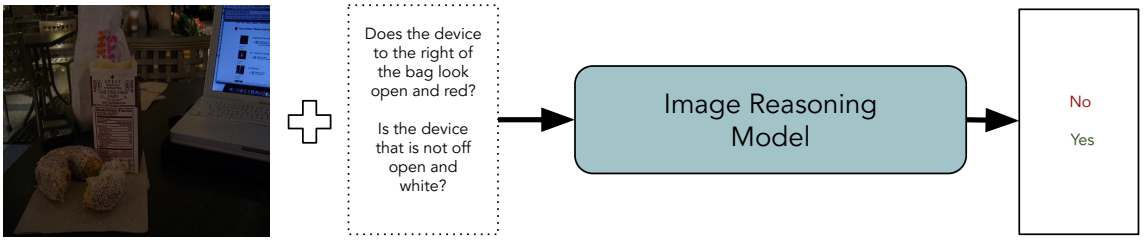

Figure 10: Given a *real-world image* and a *question*, an Image Reasoning Model reasons about the question to produce an answer.

Zellers et al. (2019) introduced a reasoning task that requires commonsense knowledge, while the goal of NLVR (Suhr et al., 2017) and NLVR2 (Suhr et al., 2019) tasks is to determine whether a sentence is true about a visual input or not.

5.2.2 Image Reasoning - Datasets

For image reasoning, both *real* and *synthetic* image datasets have been developed. In the following, we present the datasets belonging to both of these categories.

Compositional Language and Elementary Visual Reasoning (CLEVR). CLEVR[53] (Johnson et al., 2017a) is a diagnostic dataset created using a 3D computer graphics toolkit

[53] https://cs.stanford.edu/people/jcjohns/clevr

known as Blender[54]. It consists of synthetic images of simple 3D objects that vary in their attributes, viz. size, color, shape, and material. Images contain three to ten different combinations of these objects and attributes and are arranged in different spatial positions. Such complex configurations require good visual reasoning capabilities from VQA models to produce correct answers. Table 82 presents the splits of dataset.

Split	Images	Questions	Unique Questions	Overlap with train
Training	70,000	699,989	608,607	-
Validation	15,000	149,991	140,448	17,338
Test	15,000	149,988	140,352	17,335
Total	100,000	999,968	853,554	-

Table 82: Splits of the CLEVR dataset.

Natural Language Visual Reasoning (NLVR). The Cornell Natural Language for Visual Reasoning dubbed as NLVR[55] (Suhr et al., 2017) is a multimodal dataset that comes with natural language sentences grounded in synthetic images. The images are rendered and encapsulate different objects such as triangles, circles, and squares. These objects come in various sizes and are placed at different positions within images. The descriptions of the images were manually written by crowdworkers. Table 83 presents the official splits of the dataset for evaluation purposes.

Split	Unique Sentences	Examples
Training	3,163	74,460
Validation	267	5,940
Test-P	266	5,934
Test-U	266	5,910
Total	3,962	92,244

Table 83: Splits of the NLVR dataset. *Test-P* and *Test-U* means Test set (public) and Test set (unreleased) respectively.

Natural Language Visual Reasoning *for Real* (NLVR2). The limitations such as limited expressivity and semantic diversity that arose due to the synthetic nature of the NLVR dataset, has been addressed in the next incarnation of NLVR named as Natural Language for Visual Reasoning for Real, NLVR2[55] (Suhr et al., 2019). Similar to NLVR, the images in NLVR2 also come as a pair along with a grounded natural language description. Table 84 presents the official splits of the dataset.

CLEVR-CoGenT. A modified version of CLEVR is the Compositional Generalization Test (CLEVR-CoGenT)[53] (Johnson et al., 2017a) dataset. It is used to test models' ability to find novel combinations of attributes at test-time. There are two types of *conditions* in this dataset, viz. Condition A and Condition B, where based on the condition, the color

[54] `https://www.blender.org`
[55] `http://lil.nlp.cornell.edu/nlvr`

Split	Unique Sentences	Examples
Training	23,671	86,373
Validation	2,018	6,982
Test-P	1,995	6,967
Test-U	1,996	6,970
Total	29,680	107,292

Table 84: Splits of the NLVR2 dataset. *Test-P* denotes Test set Public, whereas *Test-U* means Test set Unreleased.

of the geometrical shape can vary as show in Table 85. Based on these *conditions*, the CLEVR-CoGenT dataset is divided for evaluation purposes as shown in Table 86.

Geometrical Shape	Condition	Colors of Geometrical Shape
Cubes	A	gray, blue, brown, yellow
	B	red, green, purple, cyan
Cylinders	A	red, green, purple, cyan
	B	gray, blue, brown, yellow
Spheres	A	any color
	B	any color

Table 85: *Conditions* in the CLEVR-CoGenT dataset.

Split	Condition	Images	Questions
Training	A	70,000	699,960
Validation	A	15,000	150,000
	B	15,000	149,991
Test	B	15,000	149,980
	B	15,000	149,992

Table 86: Splits of the CLEVR-CoGenT dataset.

GQA. The GQA[56] (Hudson & Manning, 2019) dataset was created to address the shortcomings in earlier VQA datasets. GQA consists of compositional questions over real-world images. Each image is associated with a scene graph of the image's objects, attributes, and relations. Also, each question is associated with a structured representation of its semantics. Table 87 presents the statistics and splits of the dataset.

Images	Questions	Vocabulary Size	Training	Validation	Testing	Challenge
113,018	22,669,678	3,097	70%	10%	10%	10%

Table 87: Statistics & splits of the GQA dataset.

[56] `https://cs.stanford.edu/people/dorarad/gqa`

Relational and Analogical Visual rEasoNing (RAVEN). The RAVEN[57] (Zhang et al., 2019) dataset was designed to perform relational and analogical visual reasoning. It is built by keeping in mind Raven's Progressive Matrices (RPM) (Burke, 1958). Furthermore, it associates vision with structural, relational, and analogical reasoning in a hierarchical representation. The dataset is split into training, validation, and testing in the ratio 6:2:2 respectively. Table 88 presents the statistics of the dataset.

Images	RPM Problems	Tree-structure per problem	Structural Labels	Rule Annotations	Avg. rules per problem
1,120,000	70,000	16	1,120,000	440, 000	6.29

Table 88: Statistics of the RAVEN dataset.

Visual Commonsense Reasoning (VCR). VCR[58] (Zellers et al., 2019) is a large-scale dataset for achieving cognition-level visual understanding. It contains about 110k images, 290k multiple choice questions and correspondingly 290k correct answers and rationales. This dataset is very diverse and, consequently, it is challenging. Table 89 presents the official splits and some high-level statistics of the dataset.

Dataset Characteristic	Train	Validation	Test
Number of questions	212,923	26,534	25,263
Number of answers per question	4	4	4
Number of rationales per question	4	4	4
Number of images	80,418	9,929	9,557
Number of movies covered	1,945	244	189
Average question length	6.61	6.63	6.58
Average answer length	7.54	7.65	7.55
Average rationale length	16.16	16.19	16.07
Average num. of objects mentioned	1.84	1.85	1.82

Table 89: High-level statistics of the VCR dataset. One fold in the dataset was held-out for blind evaluation at a later date. Hence, the statistics of that fold are not shown here.

Visual COMmonsense rEasoning in Time (Visual COMET). Visual COMET[59] (Park et al., 2020) is a large-scale dataset of Visual Commonsense Graphs for reasoning about the dynamic context of static images in order to achieve cognitive visual scene understanding. VisualCOMET contains images with person grounding (i.e., multimodal co-reference chains) and the images are connected with inference sentences. Table 90 presents the official splits and more statistics about the dataset.

[57] http://wellyzhang.github.io/project/raven.html
[58] https://visualcommonsense.com
[59] https://visualcomet.xyz

Split	Images/ Places	Events at Present	Inferences on Events Before	Inferences on Intents at Present	Inferences on Events After	Total Inferences
Train	47,595	111,796	467,025	237,608	469,430	1,174,063
Dev	5,973	13,768	58,773	28,904	58,665	146,332
Test	5,968	13,813	58,413	28,568	58,323	145,309
Total	59,356	139,377	584,211	295,080	586,418	1,465,704

Table 90: Statistics and splits of the Visual Commonsense Graph dataset.

5.2.3 Image Reasoning - Evaluation Measures, Models, and Results

In this section, we review the measures used to evaluate different models of *Image Reasoning* and the results obtained by them.

Evaluation Measures. The standard evaluation measures such as Accuracy are used for benchmarking purposes. However, there are evaluation measures that are explicitly used for *Image Reasoning* (e.g., CLEVR), viz. **Querying Attribute (QA)** that uses questions to ask about an attribute of a particular object, **Compare Attribute (CA)** which uses comparison questions for asking whether two objects have the same value for some attribute, **Compare Numbers (CN)** which uses comparison questions to ask which of two object sets is larger, **Count** which asks counting questions to find the number of objects fulfilling some conditions, and **Exist** which asks existence questions to check whether a certain type of object is present or not.

Models. The models that are designed to approach the task of *Image Reasoning* are built such that they provide an effective way of reasoning about vision with language as additional input. In Table 91, we present some exemplar architectures (refer to *Combined* column) created to address the task by integrating both image and language. We also include a column that showcases the optimization techniques used to train the *Image Reasoning* models.

Approach	Image	Language	Combined	Optimizer	RL
(Johnson et al., 2017a)	ResNet-101	LSTM	SA+MLP	ADAM	✗
(Hu et al., 2017)	VGG	LSTM	N2NMN	ADAM	✓
(Johnson et al., 2017b)	ResNet-101	LSTM	PGEE	ADAM	✓
(Santoro et al., 2017)	Custom	LSTM	RN	ADAM	✗
(Cao et al., 2018)	ResNet-101	BiLSTM	ACMN	ADAM	✗
(Perez et al., 2018)	ResNet-101	GRU	FiLM	ADAM	✗
(Hudson & Manning, 2018)	ResNet-101	BiLSTM	MAC	ADAM	✗
(Mascharka et al., 2018)	ResNet-101	-	TbD	ADAM	✗
(Haurilet et al., 2019)	ResNet-152	LSTM	FinalDestGraph	ADAM	✗
(Hu et al., 2019)	ResNet-101	LSTM	LCGN	ADAM	✗
(Mao et al., 2019)	ResNet-34	BiGRU	NS-CL	-	✓

Table 91: Exemplar *Image Reasoning* architectures. "Custom" - Own CNN architecture.

Results. The models designed on different *Image Reasoning* datasets aim to achieve generalization. In this section, we cover the results achieved by the models from some representative datasets. Table 92, Table 93, Table 94, and Table 95 presents results obtained with a subset of models built using the datasets such as CLEVR, GQA, VCR, and RAVEN that were presented in Section 5.2.2. Results for the NLVR and NLVR2 tasks can be found on the respective leaderboards[60].

Model	Count	Exist	CN	QA	CA	Overall
CNN+LSTM+SA+MLP (Johnson et al., 2017a)	59.7	77.9	75.1	80.9	70.8	73.2
N2NMN+700KProgLabel (Hu et al., 2017)	68.5	85.7	84.9	90.0	88.7	83.7
PGEE+700KProgLabel (Johnson et al., 2017b)	92.7	97.1	98.7	98.1	98.9	96.9
CNN+LSTM+RN (Santoro et al., 2017)	90.1	97.8	93.6	97.9	97.1	95.5
ACMN (Cao et al., 2018)	94.2	81.3	81.6	90.5	97.1	89.3
CNN+GRU+FiLM (Perez et al., 2018)	94.3	99.1	96.8	99.1	99.1	97.7
MAC (Hudson & Manning, 2018)	97.2	**99.5**	99.4	99.3	99.5	98.9
TbD+700KProgLabel (Mascharka et al., 2018)	97.6	99.2	99.4	**99.5**	99.6	**99.1**
FinalDestGraph (Haurilet et al., 2019)	91.3	98.6	**99.6**	**99.5**	**99.8**	97.5
LCGN+single-hop (Hu et al., 2019)	-	-	-	-	-	97.9
NS-CL (Mao et al., 2019)	**98.2**	98.8	99.0	99.3	99.1	98.9

Table 92: Comparison of different models on the CLEVR dataset.

Model	val	test-dev	test
CNN+LSTM (Hudson & Manning, 2019)	49.2	-	46.6
Bottom-up (Anderson et al., 2018b)	52.2	-	49.7
MAC (Hudson & Manning, 2018)	57.5	-	54.1
LCGN+single-hop (Hu et al., 2019)	**63.8**	**55.6**	**56.0**

Table 93: Comparison of accuracy (%) scores of different methods on the validation (val), test-dev, and test splits of the GQA dataset.

Model	(Q→ A)		(QA→ R)		(Q→ AR)	
	val	test	val	test	val	test
R2C (Zellers et al., 2019)	63.8	65.1	67.2	67.3	43.1	44.0
ViLBERT (Lu et al., 2019)	72.4	73.3	74.5	74.6	54.0	54.8
B2T2 (Alberti et al., 2019)	71.9	72.6	76.0	75.7	54.9	55.0
VL-BERT (Su et al., 2020)	73.7	74.0	74.5	74.8	55.0	55.5
Unicoder-VL (Li et al., 2020)	72.6	73.4	74.5	74.4	54.5	54.9

Table 94: Comparison of accuracy (%) scores of different models on the validation (val) and test splits of the VCR dataset.

[60] http://lil.nlp.cornell.edu/nlvr

Model	Acc	2x2 Grid	3x3 Grid	L-R	U-D	O-IC	O-IG
WReNDRT (Santoro et al., 2018)	15.02	23.26	29.51	6.99	8.43	8.93	12.35
ResNetDRT (Zhang et al., 2019)	**59.56**	**46.53**	**50.40**	**65.82**	**67.11**	**69.09**	**60.11**
Human (Zhang et al., 2019)	84.41	81.82	79.55	86.36	81.81	86.36	81.81
PerfectSolver	100	100	100	100	100	100	100

Table 95: Comparison of accuracy (%) scores of different models on the RAVEN dataset.

5.2.4 Image Reasoning - Discussion

The task of *Image Reasoning* has been studied using different types of datasets. Initially, a synthetic dataset, viz. CLEVR, was used. Later, real-world datasets like GQA were created for developing more complex vision and language integration models. Table 92 shows the results for the CLEVR dataset. Recently introduced Neuro-Symbolic Concept Learner (NS-CL) (Mao et al., 2019) reaches state-of-the-art results without explicit supervision on visual concepts, words, and semantic parsing of sentences. However, for the real-world image datasets like GQA, the approach by Hu et al. (2019) that creates Language-Conditioned Graph Networks (LGCN) providing different hops to effectively support relational reasoning achieve best results. Most of the works that outperform on the VCR task are pretrained and fine-tuned as shown in Table 94.

The RAVEN dataset differs from both CLEVR and GQA as it depends only on the image input. We can observe from Table 95 that a perfect solver achieves 100% accuracy, while the approach introduced by Zhang et al. (2019) achieves reasonable system performance.

5.2.5 Video Reasoning - Introduction

When compared to image reasoning, the video reasoning task is in its nascent stages and hence there is no clearly defined goal. However, for video reasoning, there exists a task of configurable visual question and answer (COG) designed by Yang et al. (2018). The goal of COG is to address problems related to visual and logical reasoning and memory. To be more concrete, the task is aimed at deducing the correct answer by pointing to the right object while taking into account the changes of the scene i.e., from both spatial and temporal perspective. Figure 11 demonstrates the task of temporal reasoning about synthetic 2D scenes resembling video input.

Further, Haurilet et al. (2019) addressed both image and video reasoning by introducing the concept of a question-based visual guide to constrain the potential solution space by learning an optimal traversal scheme. In their approach, the final destination nodes alone are used to produce the answers.

5.2.6 Video Reasoning - Datasets

There are not many datasets for video reasoning. One of the few examples is listed below.

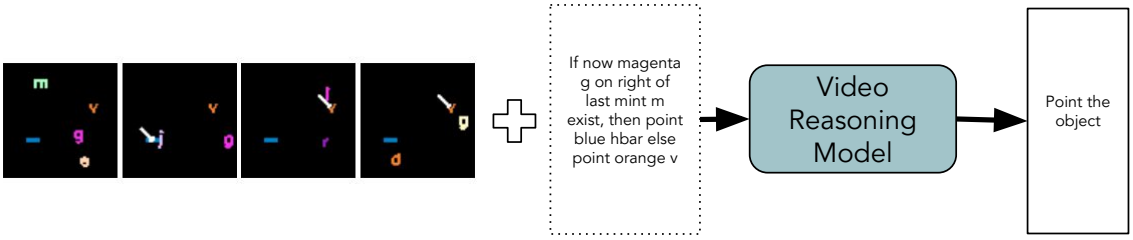

Figure 11: Given a *video* (represented as a sequence of synthetic 2D images (Yang et al., 2018)) and a *question*, a Video Reasoning Model reasons about the video to perform the task presented to it in the question.

Configurable Visual Question and Answer (COG). COG[61] (Yang et al., 2018) was created to parallel experiments in humans and animals. Table 96 presents splits of the dataset.

Split	Total Examples	Examples per Task Family
Training	10,000,320	227,280
Validation	500,016	11,364
Test	500,016	11,364

Table 96: Splits of the COG dataset.

5.2.7 Video Reasoning - Evaluation Measures, Models, and Results

In this section, we review the measures used to evaluate different models of *Video Reasoning* and the results obtained by them.

Evaluation Measures. For *Video Reasoning* (e.g., COG) the evaluation measures used are based on changes of the scene in three different query types.

- **Pointing (Point)** queries ask the system to point to a certain object.
- **Yes/No** seeks a binary decision, while **Conditional (Condit)** is composed of questions based on objects that needs to fulfill certain conditions.
- **Attribute-related (Atts)** which is composed of questions about certain attributes.

Models. Many models have been created to approach the task of *Video Reasoning*. In Table 97, we present some exemplar architectures (refer to *Combined* column) created to address the task by integrating video and language.

Results. As discussed earlier, several models have been created to approach the task of *Video Reasoning*. In Table 98, we present the results obtained with a subset of models built using the COG dataset presented in Section 5.2.6.

[61] https://github.com/google/cog#datasets

Approach	Video	Frame	Language	Combined	RL
(Yang et al., 2018)	-	Custom	LSTM	WorkMemory	✗
(Haurilet et al., 2019)	-	ResNet-152	LSTM	FinalDestGraph	✗

Table 97: Exemplar *Video Reasoning* architectures.

Model	Atts	Condit	Point	Yes/No	All
WorkMemory (Yang et al., 2018)	-	-	-	-	93.7
QuestionNodes (Haurilet et al., 2019)	73.7	63.5	92.5	57.9	63.3
FinalDestGraph (Haurilet et al., 2019)	**99.2**	**98.4**	**100.0**	**95.0**	**97.2**

Table 98: Comparison of measures using different methods on the COG dataset.

5.2.8 Video Reasoning - Discussion

The results presented in Table 98 show that the recently proposed approach by Haurilet et al. (2019) achieves the best result on different task-specific measures. This approach proposes a question-based visual guide, which constrains the potential solution space by learning an optimal traversal scheme of a graph.

5.3 Visual Entailment

Goal of the *Visual Entailment* task is to learn a model that predicts whether the visual content entails the augmented text along with hypothesis. Both images and videos are used as visual inputs. In the following, we describe the task, datasets used, and the approaches that have been proposed to tackle the problem.

5.3.1 Image Entailment - Introduction

To address the perceived drawbacks of VQA and visual reasoning, i.e. that they deal with similar objects and sentence structures, Vu et al. (2018) initially proposed a visually-grounded version of the Textual Entailment task where an image is augmented with textual premise and hypothesis. However, this task was refined by Xie et al. (2019) to predict whether the image semantically entails the text, given image-sentence pairs, where the premise is defined by an image instead of a natural language sentence. Figure 12 illustrates the task, where the image as a *premise* and a piece of text as *hypothesis* are used by the Image Entailment model to predict whether the hypothesis is an *entailment*, *contradiction*, or *neutral*.

5.3.2 Image Entailment - Datasets

The image entailment task is achieved using two different datasets. One dataset extends Natural Language Inference with Visually-grounded Natural Language Inference (V-SNLI) (Vu et al., 2018) while the other extends the Flickr30K dataset (see Section 3.1.2) into a visual entailment dataset (SNLI-VE)[62] (Xie et al., 2019). Table 99 and Table 100 presents the statistics and splits of these two datasets respectively.

[62] https://github.com/necla-ml/SNLI-VE

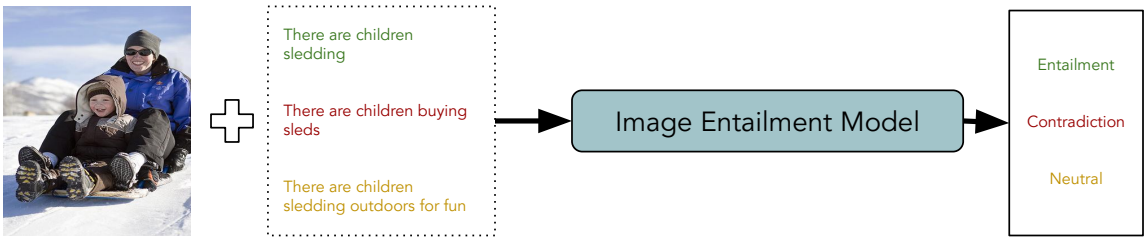

Figure 12: Given an image as a *premise* and a natural language text as a *hypothesis*, an Image Entailment Model predicts whether the hypothesis is an entailment, contradiction, or neutral by understanding the evidence(s) present in the image.

Split	Entailment	Neutral	Contradiction
Training	182,167	181,515	181,938
Validation	3,329	3,235	3,278
Test	3,368	3,219	3,237
V-SNLI$_{\text{hard}}$ Test	1,058	1,068	1,135

Table 99: Splits of the V-SNLI dataset.

Split	Images	Entailment	Neutral	Contradiction	Vocab
Training	29,783	176,932	176,045	176,550	29,550
Validation	1000	5,959	5,960	5,939	6,576
Test	1000	5,973	5,964	5,964	6,592

Table 100: Splits of the SNLI-VE dataset.

5.3.3 Image Entailment - Evaluation Measures, Models, and Results

In this section, we review the measures used to evaluate different models of *Image Entailment* and the results obtained by them.

Evaluation Measures. *Image Entailment* task is evaluated using the Accuracy measure.

Models. Two different models are created to approach the task of *Image Entailment*. In Table 101, we present some exemplar architectures (refer to *Combined* column) created to address the task. We also include a column that showcases the optimization techniques used to train those models.

Approach	Image	Language	Combined	Optimizer	RL
(Vu et al., 2018)	VGG	BiLSTM	V-BiMPM	ADAM	✗
(Xie et al., 2019)	ResNet-101	GRU	EVE-Image	ADAM	✗

Table 101: Exemplar *Image Entailment* architectures.

Results. The *Image Entailment* models leverage both image and textual input representations to build an entailment pipeline. In Table 102, Table 103, and Table 104 we present

results obtained with a subset of models that were built using the datasets presented in Section 5.3.2.

Model	Contradiction	Neutral	Entailment	Overall
Relation Network (Santoro et al., 2017)	67.29	68.86	66.50	67.55
Bottom-up (Anderson et al., 2018a)	70.52	**70.96**	65.23	68.90
Top-Down (Anderson et al., 2018a)	69.72	69.33	71.86	70.3
Hypothesis Only (Gururangan et al., 2018)	67.60	67.71	64.83	66.71
EVE-ROI (Xie et al., 2019)	67.69	69.45	**74.25**	70.47
EVE-Image (Xie et al., 2019)	**71.56**	70.52	71.39	**71.16**

Table 102: Comparison of accuracies (%) of different models on the SNLI-VE dataset.

Model	Contradiction	Neutral	Entailment	Overall
Hypothesis Only (Bowman et al., 2015)	66.29	66.36	72.65	68.49
LSTM (blind) (Bowman et al., 2015)	79.7	76.79	87.71	81.49
V-LSTM (Anderson et al., 2018a)	71.39	68.06	87.14	75.70
BiMPM (Wang et al., 2017)	86.25	82.79	90.03	86.41
V-BiMPM (Vu et al., 2018)	**87.53**	**82.91**	**90.38**	**86.99**

Table 103: Comparison of accuracies (%) of different models on the V-SNLI dataset.

Model	Contradiction	Neutral	Entailment	Overall
Hypothesis Only (Bowman et al., 2015)	25.29	20.22	31.28	25.57
LSTM (blind) (Bowman et al., 2015)	60.79	50.19	72.12	60.99
V-LSTM (Anderson et al., 2018a)	46.34	32.02	69.09	49.03
BiMPM (Wang et al., 2017)	**77.62**	59.36	80.43	72.55
V-BiMPM (Vu et al., 2018)	76.12	**63.67**	**81.38**	**73.75**

Table 104: Comparison of accuracy (%) scores of various models on V-SNLI$_{\text{hard}}$.

5.3.4 Image Entailment - Discussion

The task of *Image Entailment* was evaluated using two different datasets. Table 103 and Table 104 shows results obtained from V-SNLI in different settings. The approach proposed by Vu et al. (2018) that creates a visually grounded Bilateral Multi-Perspective Matching (BiMPM) model achieves the best result for the entailment task.

Similarly, evaluations conducted with SNLI-VE dataset (cf. Table 102) show that the Explainable Visual Entailment (EVE) approach proposed by Xie et al. (2019) achieves the best overall result.

5.3.5 Video Entailment - Introduction

Video entailment (Liu et al., 2020) aims to infer whether the natural language hypothesis is entailed or contradicted when given a video clip aligned with the subtitles information. The video contains diverse temporal dynamics, event shifts, and social interactions. Figure 13

illustrates the task: given a video clip with aligned subtitles as premise and a natural language hypothesis based on the video content, a video entailment model needs to infer whether the hypothesis is entailed or contradicted by the given video clip.

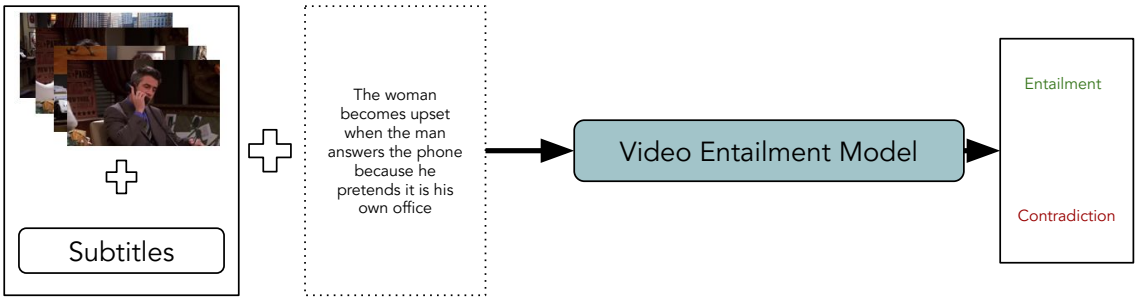

Figure 13: Given a video along with aligned subtitles as *premise* and a paired natural language text as *hypothesis*, the goal of a Video Entailment Model is to predict whether the hypothesis is an entailment or contradiction, by understanding the evidence(s) observed in the video. Example modified from Liu et al. (2020).

5.3.6 Video Entailment - Datasets

The Video Entailment task is proposed by Liu et al. (2020), with the introduction of a large-scale dataset called as VIdeO-and-Language INference (VIOLIN)[63]. Detailed statistics of the dataset is presented in Table 105.

Video Source (TV Show/Movie Clips)	Num. of Episodes	Num. of Clips	Avg. Clip Len	Avg. Pos. Stmnt Len	Avg. Neg. Stmnt Len	Avg. Sub-Title Len
Friends	234	2,676	32.89s	17.94	17.85	72.80
Desperate Housewives	180	3,466	32.56s	17.79	17.81	69.19
How I Met Your Mother	207	1,944	31.64s	18.08	18.06	76.78
Modern Family	210	1,917	32.04s	18.52	18.20	98.50
MovieClips	5,885	5,885	40.00s	17.79	17.81	69.20
All	6,716	15,887	35.20s	18.10	18.04	76.40

Table 105: Statistics of different video sources in the VIOLIN dataset.

For training and model evaluation purposes, the VIOLIN dataset is split into training, validation, and test splits in the ratio of 8:1:1. The exact number of triplet instances in each of the splits is shown in Table 106.

5.3.7 Video Entailment - Evaluation Measures, Models, and Results

In this section, we present the evaluation measures, models, and results achieved with various architectures introduced for solving the *Video Entailment* task.

Evaluation Measures. The *Video Entailment* models are evaluated using Accuracy.

[63] https://github.com/jimmy646/violin

Split	Number of Videos (V)	Number of Hypotheses (H)	Number of Triplets (V, S, H)
Training	12,687	76,122	76,122
Validation	1,600	9,600	9,600
Testing	1,600	9,600	9,600
Total	15,887	95,322	95,322

Table 106: Splits of the VIOLIN dataset.
(V: Video, S: Subtitle, H: Hypothesis)

Models. Very few models have been created to approach the task of *Video Entailment*. The variation of the *Video Entailment* models include the usage of different type of textual content such as subtitles, statements, etc. In Table 107, we present some exemplar architectures (refer to *Combined* column) created to address the task by integrating both video and language inputs. We also include a column that showcases the optimization techniques used to train those models.

Approach	Video	Frame	Language	Combined	Optimizer	RL
(Liu et al., 2020)	-	Detection Feat	BERT	SSV	ADAM	✗

Table 107: Exemplar *Video Entailment* architectures. SSV - Statement+Subtitles+Visual.

Results. Few models which have been designed to approach the task of *Video Entailment* use different types of textual content aligned with video. In Table 108 we present results obtained with subset of models built using the VIOLIN dataset presented in Section 5.3.6. For building textual or visual representations, models such as **SSV** has used pretrained vision and language integration models such as LXMERT (Tan & Bansal, 2019).

Model	Visual	Text	Accuracy
Statement (Liu et al., 2020)	-	BERT	54.20
Statement+Visual (Liu et al., 2020)	Detection Feat	BERT	59.45
Statement+Subtitles (Liu et al., 2020)	-	BERT	66.05
SSV (Tan & Bansal, 2019)	LXMERT	LXMERT	66.25
SSV (Liu et al., 2020)	Detection Feat	BERT	**67.84**

Table 108: Comparison of accuracies (%) of different methods on the VIOLIN dataset.

5.3.8 Video Entailment - Discussion

The task of *Video Entailment* was evaluated using the VIOLIN dataset and the recently proposed method by Liu et al. (2020) has shown that using multi-source information arising from different types of data such as Statements, Subtitles, and Visual features are useful for building a robust model. In addition, textual features generated using contextualized word embedding models are effective as well.

6. Visual Dialog

In this section, we explore the task of *Visual Dialog*. The objective of visual dialog is different from the previously discussed tasks and involves a complex interaction between a human and an artificial agent.

6.1 Image Dialog

In the following, we describe the setting of *Visual Dialog* where an image is used as the visual input.

6.1.1 Image Dialog - Introduction

The goal of the *image dialog* task is to create AI agents that can hold dialog with humans in a natural language of choice about a visual content (Das et al., 2017a), represented by an image. To be more specific, given an image, a history of dialogs, and a question about the image, the goal of an AI agent is to ground the question in the image, infer the context from the history, and then answer the question accurately. However, this problem can also be construed as a task where the goal of the AI system is to locate an unknown object in the image by asking a sequence of questions (de Vries et al., 2017) or to hold natural-sounding conversations about a shared image (Mostafazadeh et al., 2017). In Figure 14, we provide a visual depiction to illustrate the said task.

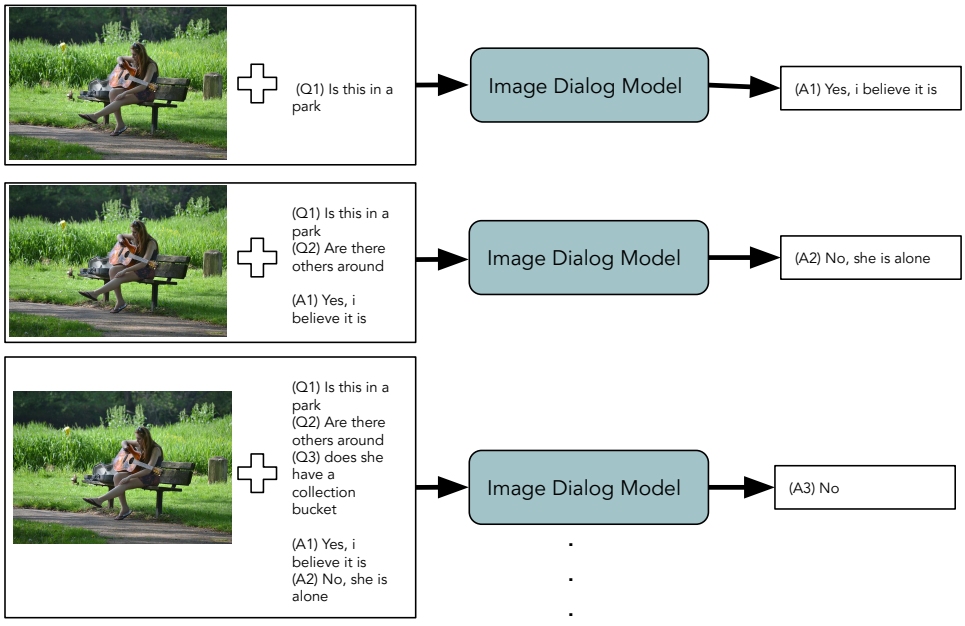

Figure 14: Given an *image*, *question* and the *dialog history*, an Image Dialog Model generates an answer based on these multimodal information.

Further, a standard agent can be extended to have a question and answer bot cooperating with each other for guessing images (Das et al., 2017b). To counter generic responses

in dialog generation, knowledge transfer from dialog generation was explored with a discriminative dialog module trained to rank a list of candidate human responses (Lu et al., 2017a). However, other approaches constrained themselves to specific domains and proposed end-to-end optimization schemes (Strub et al., 2017). Seo et al. (2017) introduced attentive memory that exploits visual attention in the past to resolve the current reference. Recently, reinforcement learning and Generative Adversarial Networks (GANs) were also used to generate more human-like responses to questions in the image-based dialog (Wu et al., 2018). Dialog can also be seen from the perspective of a system which asks questions, and demonstrates how a visual dialog can be generated from discriminative question generation and answering (Jain et al., 2018). Furthermore, co-reference resolution was also investigated (Kottur et al., 2018) to bridge the gap between nouns and pronouns with the usage of modules that form explicit and grounded co-reference resolution at word-level.

Recently, a novel attention mechanism called recursive visual attention (Niu et al., 2019) was proposed to resolve visual co-reference for visual dialog by browsing the dialog history. Another approach (Zheng et al., 2019) formalized the task as inference in a graphical model with partially observed nodes and unknown graph structures, i.e., relations in dialog. Further, Guo et al. (2019) extended one-stage solution to a two-stage solution by building an image-question-answer synergistic network to value the role of the answer for precise visual dialog. Other novel approaches (Shekhar et al., 2019) were also designed where a visually-grounded encoder was employed to synergize between guessing and asking questions. Further, a cooperative learning regime was followed to improve the accuracy.

6.1.2 Image Dialog - Datasets

For addressing the task of image dialog several datasets have been created. In the following, we elaborate each of them separately.

VisDial. For Image Dialog, there exists two versions of this dataset, VisDial v0.9 and VisDial 1.0[64] (Das et al., 2017a). VisDial was created using the MSCOCO dataset. For VisDial v0.9, splits are divided only into the training and validation set. Table 109 and Table 110 present details about the splits of VisDial v0.9 and VisDial v1.0 respectively.

Split	Images	Questions	Answers	Dialog Turns
Training	82,783	827,830	827,830	10
Validation	40,504	405,040	405,040	10
Test	-	-	-	-

Table 109: Splits of the VisDial v0.9 dataset.

CLEVR-Dialog. The CLEVR-Dialog[65] (Kottur et al., 2019) dataset was developed for studying multi-round reasoning in visual dialog. The dialog grammar is grounded in the scene graphs of the CLEVR dataset (Section 5.2.2), originally developed for reasoning about images. Table 111 provides statistics of the dataset, while Table 112 shows the dataset splits.

[64] https://visualdialog.org/data

[65] https://github.com/satwikkottur/clevr-dialog

Split	Images	Questions	Answers	Dialog Turns
Training	123,287	1,232,870	1,232,870	10
Validation	2,064	20,640	20,640	10
Test	8,000	80,000	80,000	1

Table 110: Splits of the VisDial v1.0 dataset.

CLEVR Images	Total Dialogs	Total Questions	Unique Questions	Unique Answers	Vocabulary Size	Dialog Turns	Mean Ques. Length
85k	425k	4.25M	73k	29	125	10	10.6

Table 111: Statistics of the CLEVR-Dialog dataset.

Split	Images	Q&A Pairs	Instances	Dialog Rounds
Training	70,000	3.5M	5	10
Validation	15,000	0.75M	5	10
Test	-	-	-	-

Table 112: Splits of the CLEVR-Dialog dataset.

6.1.3 Image Dialog - Evaluation Measures, Models and Results

In this section, we review the measures used to evaluate different models of *Image Dialog* and the results achieved by these models.

Evaluation Measures. The *Image Dialog* models are evaluated using the *Retrieval* metrics that have been discussed in Section 3.1.3.

Models. The models created to approach the *Image Dialog* task continuously process a stream of images and textual dialog information. In Table 113, we present some exemplar architectures (refer to *Combined* column) designed to integrate image and textual dialog to address the task.

Approach	Image	Language	Combined	RL
(Das et al., 2017a)	VGG	LSTM	MemoryNetwork	✗
(Lu et al., 2017a)	VGG	LSTM	HCIAE-NP-ATT	✗
(Seo et al., 2017)	VGG	LSTM	AMEM	✗
(Jain et al., 2018)	VGG	LSTM	SF	✗
(Kottur et al., 2018)	ResNet-152	LSTM	CorefNMN	✗
(Wu et al., 2018)	VGG	LSTM	CoAtt-GAN	✓
(Niu et al., 2019)	ResNet-152	LSTM	RvA	✗
(Zheng et al., 2019)	VGG	LSTM	GNN	✗
(Guo et al., 2019)	ResNet-101	LSTM	Synergistic	✗

Table 113: Exemplar *Image Dialog* Architectures (Discriminative and Generative).

Results. Models that are created to solve the task of *Image Dialog* effectively comprehends the complexity of the task. Several approaches are used to build the models with different

versions of the same dataset. However, few approaches share some commonalities such as usage of Memory Networks (Sukhbaatar et al., 2015). Table 114 and Table 115 presents the results obtained with a subset of both **discriminative** and **generative** models built using the "VisDial0.9" dataset. While Table 116 presents the results obtained only with a subset of **generative** models built using the "VisDial1.0" dataset presented earlier in Section 6.1.2.

Model	MRR	R@1	R@5	R@10	Mean
LF (Das et al., 2017a)	0.5807	43.82	74.68	84.07	5.78
HRE (Das et al., 2017a)	0.5846	44.67	74.50	84.22	5.72
HREA (Das et al., 2017a)	0.5868	44.82	74.81	84.36	5.66
MN (Das et al., 2017a)	0.5965	45.55	76.22	85.37	5.46
HCIAE-NP-ATT (Lu et al., 2017a)	0.6222	48.48	78.75	87.59	4.81
AMEM (Seo et al., 2017)	0.6227	48.53	78.66	87.43	4.86
CoAtt (Wu et al., 2018)	0.6398	50.29	80.71	88.81	4.47
SF (Jain et al., 2018)	0.6242	48.55	78.96	87.75	4.70
SCA (Wu et al., 2018)	0.6398	50.29	80.71	88.81	4.47
CorefNMN (Kottur et al., 2018)	0.641	50.92	80.18	88.81	4.45
GNN (Zheng et al., 2019)	0.6285	48.95	79.65	88.36	4.57
RvA (Niu et al., 2019)	**0.6634**	**52.71**	**82.97**	**90.73**	**3.93**

Table 114: Results of different **discriminative models** on the validation split of the VisDial v0.9 dataset.

Model	MRR	R@1	R@5	R@10	Mean
LF (Das et al., 2017a)	0.5199	41.83	61.78	67.59	17.07
HRE (Das et al., 2017a)	0.5237	42.29	62.18	67.92	17.07
HREA (Das et al., 2017a)	0.5242	42.28	62.33	68.17	16.79
MN (Das et al., 2017a)	0.5259	42.29	62.85	68.88	17.06
HCIAE-NP-ATT (Lu et al., 2017a)	0.5386	44.06	63.55	69.24	16.01
CorefNMN (Kottur et al., 2018)	0.535	43.66	63.54	69.93	15.69
CoAtt (Wu et al., 2018)	0.5411	44.32	63.82	69.75	16.47
CoAtt-RL (Wu et al., 2018)	**0.5578**	**46.10**	**65.69**	71.74	14.43
RvA (Niu et al., 2019)	0.5543	45.37	65.27	**72.97**	**10.71**

Table 115: Results of different **generative models** on the validation split of the VisDial v0.9 dataset.

6.1.4 Image Dialog - Discussion

For the *Image Dialog* task, two versions of the same dataset were used for evaluation. Similar approaches were used for the evaluation of both datasets with retrieval metrics. Nevertheless, the methods that achieve state-of-the-art performance on both datasets differ. Among the generative and discriminative methods on VisDial v0.9 dataset, the Recursive Visual Attention (RvA) approach proposed by Niu et al. (2019) achieves the best result. RvA refines the visual attention recursively by browsing through the dialog history until

the agent has sufficient confidence in its visual co-reference resolution. This has also been shown to generate interpretable attention maps without additional annotations.

For the VisDial v1.0 dataset, the results presented in Table 116 show that Synergistic-ensemble by Guo et al. (2019) outperform RvA.

Model	MRR	R@1	R@5	R@10	Mean	NDCG
LF (Das et al., 2017a)	0.5542	40.95	72.45	82.83	5.95	0.4531
LF-att (Das et al., 2017a)	0.5707	42.08	74.83	85.05	5.59	0.4976
HRE (Das et al., 2017a)	0.5416	39.93	70.45	81.50	6.41	0.4546
MN (Das et al., 2017a)	0.5549	40.98	72.30	83.30	5.92	0.4750
MN-att (Das et al., 2017a)	0.5690	42.43	74.00	84.35	5.59	0.4958
CorefNMN (Kottur et al., 2018)	0.615	47.55	78.10	88.80	4.40	0.547
GNN (Zheng et al., 2019)	0.6137	47.33	77.98	87.83	4.57	0.5282
RvA (Niu et al., 2019)	0.6303	49.03	80.40	89.83	4.18	0.5559
Synergistic-ensemble (Guo et al., 2019)	**0.6342**	**49.30**	**80.77**	**90.68**	**3.97**	**0.5788**

Table 116: Results of different **discriminative models** on the test-standard split of the VisDial v1.0 dataset.

6.2 Video Dialog

In this part, we present details about the *Visual Dialog* task in which a video is used as the visual input and a conversational chat with humans about the visual content is expected.

6.2.1 Video Dialog - Introduction

The aim of video dialog is to leverage scene information containing both audio (which can be transcribed as subtitles) and visual frames to hold a dialog (i.e., an exchange) with humans in a natural language of choice about the multimedia content (Alamri et al., 2019b, 2019a). A successful system is expected to ground concepts from the question in the video while leveraging contextual cues from the dialog history. Figure 15 illustrates the video dialog task.

Several approaches have been proposed to address the task, where initially multimodal attention-based video description features were used to improve dialog (Hori et al., 2019). Further, a novel baseline (Schwartz et al., 2019) analyzed components such as data representation, extraction, attention, and answer generation in order to show that there can be relative improvements as compared to other approaches.

6.2.2 Video Dialog - Datasets

Audio Visual Scene-Aware Dialog (AVSD)[66] (Alamri et al., 2019b) was created for the Scene-Aware Dialog Challenge, in which the agent grounds its responses on the dynamic scene, the audio, and the history (previous rounds) of the dialog. Table 117 presents some statistics and the splits of the AVSD dataset.

[66] `https://video-dialog.com`

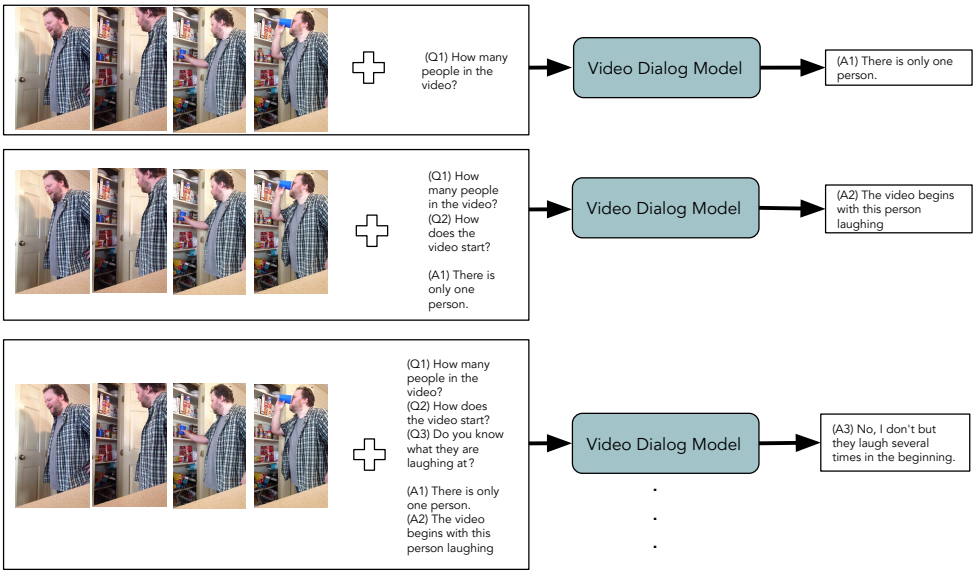

Figure 15: Given a *video* (represented as a sequence of frames), a *question*, and the *dialog history*, a Video Dialog Model generates answers based on these information.

Split	Dialogs	Turns	Words
Training	7,985	123,480	1,163,969
Validation	1,863	14,680	138,314
Test	1,968	14,660	138,790

Table 117: Splits of the AVSD dataset.

6.2.3 Video Dialog - Evaluation Measures, Models, and Results

In this section, we review the evaluation measures used to benchmark different models of *Video Dialog* and the results obtained by these models.

Evaluation Measures. The *Video Dialog* models are evaluated using the "Retrieval metrics" discussed in Section 3.1.3.

Models. Only a couple of models have been proposed so far to approach the task of *Video Dialog*. These models aim to capture the temporal aspect of a video and incorporate it in the textual dialog. In Table 118, we present some exemplar architectures (refer to *Combined* column) designed to address the task by integrating both video and language inputs. We also include a column that showcases the optimization techniques used to train those models.

Approach	Video	Frame	Language	Combined	Optimizer	RL
(Hori et al., 2019)	I3D	VGG	LSTM	MultimodalAtt	ADAM	✗
(Schwartz et al., 2019)	I3D	VGG	LSTM	i3d-rgb-spatial-10	ADAM	✗

Table 118: Exemplar *Video Dialog* architectures.

Results. As discussed earlier only few models have been created to address the task of *Video Dialog*. In Table 119 we present the results obtained with those models built using the "AVSD" dataset presented earlier in Section 6.2.2.

Model	B-1	B-2	B-3	B-4	METEOR	CIDEr
Att-base (Hori et al., 2019)	0.273	0.173	0.117	0.084	0.117	0.766
Att-weightshare (Schwartz et al., 2019)	**0.293**	**0.191**	0.133	0.097	0.127	0.923
i3d-rgb-spatial-10 (Schwartz et al., 2019)	0.290	0.190	**0.133**	**0.097**	0.127	0.928
Att-base-beam (Schwartz et al., 2019)	0.285	0.187	0.131	0.096	**0.128**	**0.941**

Table 119: Results of different models on the "AVSD" dataset.

6.2.4 Video Dialog - Discussion

The *Video Dialog* task is evaluated with the AVSD dataset. Different strategies have been explored to fuse the language and video features to create a strong baseline. In particular, the approach proposed by Schwartz et al. (2019), which uses beam search and the attention mechanism (i.e., Att-base-beam) over different modalities, outperforms other baseline methods.

7. Multimodal Machine Translation

In this section, we explore the task of *Multimodal Machine Translation* (MMT). The goal of this task is to translate natural language sentences that describe visual content (e.g. image) in a source language into a target language by taking the visual content as an additional input to the source language sentences.

7.1 Machine Translation with Image

In the following, we elaborate on the *Multimodal Machine Translation* task by considering image as the only visual input.

7.1.1 Machine Translation with Image - Introduction

The aim of MMT (Specia et al., 2016; Hitschler et al., 2016; Elliott et al., 2017; Barrault et al., 2018) is to translate sentences, that describe an image, in a source language into equivalent sentences in a target language. However, for any given image the description can be written in different source languages, resulting in multiple source language descriptions. This situation opens up the possibility to propose different variants of the MMT task. The first variant is a *single source translation* task, in which the image description in a single source language is translated to the target language with additional cues from the corresponding image. Figure 16 depicts this variant where an image is accompanied with its description in English and needs to be translated by the model into a description in German.

The second variant is a target language description generation task with additional source language cues, i.e., multiple source language descriptions of the same image termed as *multisource MMT*. Figure 17 illustrates this variant, where an image is accompanied with

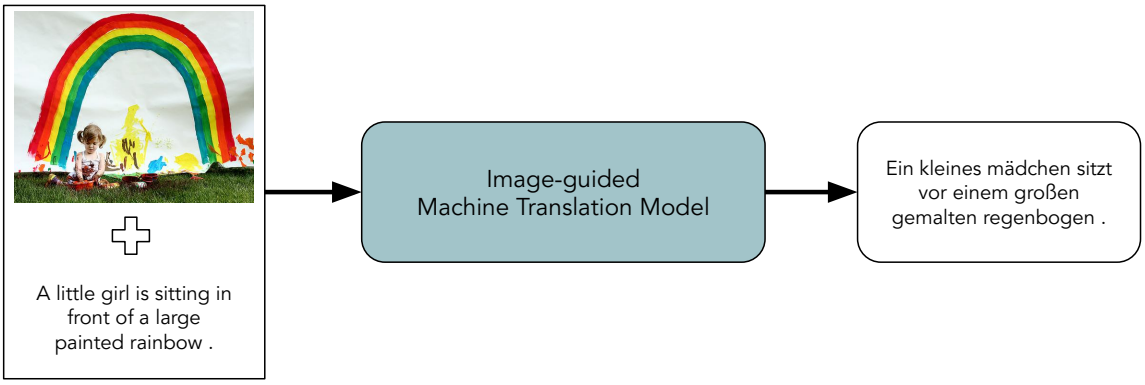

Figure 16: Given an *image* and its *description* in a source language (e.g. En), an Image-guided Machine Translation model produces a description in a target language (e.g. De).

its descriptions in English (en), French (fr), and Czech (cs), that are all used to generate the German (de) translation.

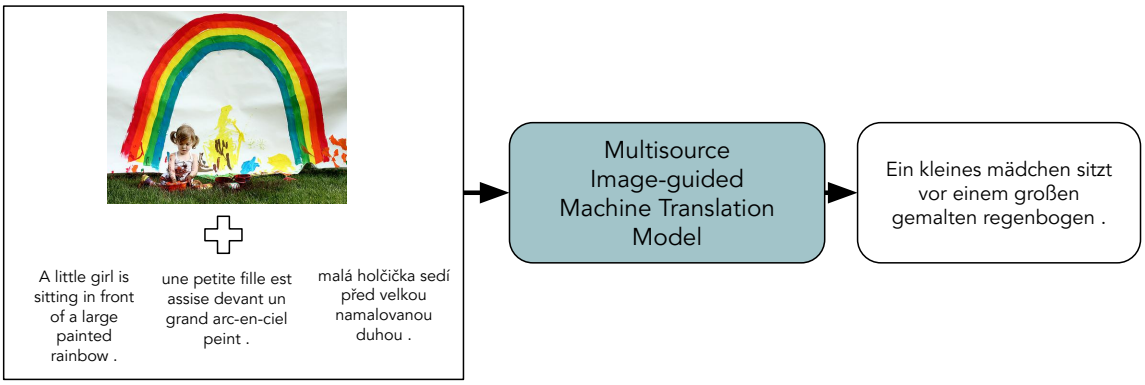

Figure 17: Given an *image* and its *description* in multiple source languages (e.g. en, fr, cs), a Multisource Image-guided Machine Translation model produces a description in a target language (e.g. de).

Different approaches have been proposed to handle single source MMT by associating visual and textual features with multimodal attention (Huang et al., 2016). Further, a novel approach where a doubly-attentive decoder incorporated visual features to bridge the gap between image description and translation was proposed (Calixto et al., 2017). In a similar vein, global visual features were incorporated in an attention-based multimodal NMT (Calixto & Liu, 2017). This is achieved by attending to source-language words and parts of an image independently by means of two separate attention mechanisms.

MMT task can also be solved using two sub-tasks: learning to translate, and learning visually grounded representations (Elliott & Kádár, 2017), both combined in a multi-task learning framework. Further, an advanced multimodal compact bilinear pooling method (Delbrouck & Dupont, 2017a, 2017b) has also been used for MMT in which the outer product of two vectors combines the attention features of the two modalities. Another model (Zhou

et al., 2018c) used a shared visual-language embedding and a translator for learning. This joint model leverages a visual attention grounding mechanism that links the visual semantics with the corresponding textual semantics. Due to the presence of large multimodal data on the web, noisy image captions have also been tried for MMT (Schamoni et al., 2018). A latent variable model (Calixto et al., 2019) has also been attempted in which the latent variable can be seen as a multimodal stochastic embedding of an image and its description in a foreign language.

MMT models have also been used in an adversarial setting. Elliott (2018) found that even in the presence of visual features from unrelated images there is no significant performance degradation. Due to the recent success of unsupervised machine translation (Lample et al., 2018), there is also a growing interest in extending it for unsupervised MMT (Su et al., 2019). Other studies (Caglayan et al., 2019) have reduced criticism of MMT by showing that under the limited textual context, MMT models are capable of leveraging the visual input to generate better translations. Regarding multisource models, Libovický and Helcl (2017) explored MMT using neural multi-source sequence-to-sequence learning.

7.1.2 Machine Translation with Image - Datasets

The main dataset used with the models above (Section 7.1.1) is the Multi30k-MMT[67] dataset (Barrault et al., 2018), extended using the Flickr30k dataset. Along with English, it contains human translated German, French, and Czech language sentences. The splits of this dataset can be found in Table 120.

Split	Images	Captions
Training	29,000	29,000
Validation	1,014	1,014
Test	1,000	1,000

Table 120: Splits of Multi30k-MMT for English, German, French, and Czech.

7.1.3 Machine Translation with Image - Evaluation Measures, Models, and Results

In this section, we review the evaluation measures used to benchmark different models of *Machine Translation with Image* and the results obtained by these models.

Evaluation Measures. To evaluate *Machine Translation with Image* models, the "Retrieval metrics" presented in the Section 3.1.3 are used.

Models. Several models have been created for the task of *Machine Translation with Image*. The aim of these models is to tackle translation using either a single or multiple language textual sources along with an image. In Table 121, we present some exemplar architectures (refer to *Combined* column) which integrate both image and language to address the task. We also include an "Optimizer" column that indicates the optimization techniques used to train those models.

[67] https://www.statmt.org/wmt18/multimodal-task.html

Approach	Image	Language	Combined	Optimizer	RL
(Calixto et al., 2017)	ResNet-50	BiGRU	DoubleAtt	Adadelta	✗
(Calixto & Liu, 2017)	VGG	BiGRU	GVF	Adadelta	✗
(Elliott & Kádár, 2017)	Inception-V3*	BiGRU	Imagination	ADAM	✗
(Caglayan et al., 2017)	ResNet-50	BiGRU	Lium-cvc-ensemble	ADAM	✗
(Calixto et al., 2019)	ResNet-50	BiGRU	$VMMT_F$	ADAM	✗
(Helcl et al., 2018)	ResNet-50	LSTM	CUNI-ensemble	ADAM	✗

Table 121: Exemplar *Machine Translation with Image* architectures. * - compares with ResNet-50 and VGG also.

Results. In Table 122 and Table 123 we present the results obtained with a subset of models built using the Multi30k-MMT dataset presented earlier in Section 7.1.2.

Results of Different Methods				
Model	Language	en → de	en → fr	en → cs
DoubleAtt (Calixto et al., 2017)	BLEU	36.5	-	-
	METEOR	55.0	-	-
GVF (Calixto & Liu, 2017)	BLEU	37.3	-	-
	METEOR	55.1	-	-
Imagination (Elliott & Kádár, 2017)	BLEU	36.8	-	-
	METEOR	55.8	-	-
Lium-cvc-ensemble (Caglayan et al., 2017)	BLEU	41.0	56.7	-
	METEOR	60.5	73.0	-
$VMMT_F$ (Calixto et al., 2019)	BLEU	37.6	-	-
	METEOR	56.0	-	-
CUNI-ensemble (Helcl et al., 2018)	BLEU	42.6	62.8	35.9
	METEOR	59.4	77.0	32.7

Translation

Table 122: Machine Translation with Image on the Multi30k test set [2016 (en → de), 2017 (en → fr), 2018 (en → cs)].

Results of Different Methods				
Model	Language	en → de	en → fr	en → cs
CUNI-single (Helcl et al., 2018)	BLEU	32.5	40.6	31.8
	METEOR	52.3	61.0	30.6
MeMAD (Grönroos et al., 2018)	BLEU	38.5	44.1	-
	METEOR	56.6	64.3	-

Table 123: Machine Translation with Image on Multi30k test set [2018 (en → de, en → fr, en → cs)].

7.1.4 Machine Translation with Image - Discussion

This task is evaluated using only one dataset, e.g., Multi30k-MMT, containing descriptions in three source languages and one target language. Results presented in Table 122 and Table 123 refer to the shared task proposed in different years. We can observe that based on different years of test set release, varied sets of approaches outperform the baseline methods.

7.2 Machine Translation with Video

In the following, we present more details about *Multimodal Machine Translation* by using the video as the visual input.

7.2.1 Machine Translation with Video - Introduction

The goal in video-guided machine translation (Wang et al., 2019b) is to translate a source language description into the target language equivalent using the video information as additional spatio-temporal context.

Figure 18 illustrates this task where the English language description accompanied by a video needs to be translated into the equivalent description in German.

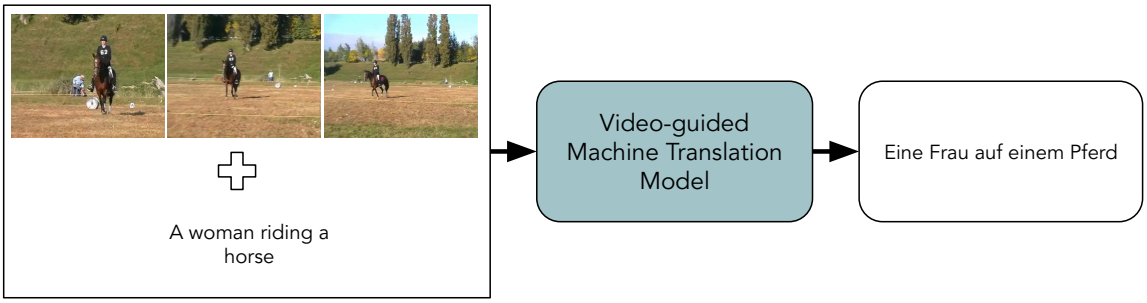

Figure 18: Schematic representation of Video-guided Machine Translation task.

7.2.2 Machine Translation with Video - Datasets

The VATEX[68] (Wang et al., 2019b) dataset was created for English and Chinese languages to perform machine translation with video and also for the task of generating multilingual video descriptions. Table 124 presents more details about the dataset.

Split	Videos	Action Label
Training	25,991	✓
Validation	3,000	✓
Public Test	6,000	-
Secret Test	6,278	-

Table 124: Splits of the VATEX dataset. *Secret Test* denotes human-annotated captions heldout for organizing challenges; Hence, this split is unavailable to the public.

[68] `http://vatex.org/main/index.html`

7.2.3 Machine Translation with Video - Evaluation Measures, Models, and Results

In this section, we review the measures used to evaluate different models of *Machine Translation with Video* and the results obtained by them.

Evaluation Measures. To evaluate the *Machine Translation with Video* models, the Language metrics discussed in Section 3.1.3 are used.

Models. Very few models have been created to investigate the task of *Machine Translation with Video*. The temporal aspect of a video is crucial for providing effective translations. In contrast to *Machine Translation with Image*, the task of *Machine Translation with Video* only has models which are built using single textual source. In Table 125, we present some exemplar architectures (refer to *Combined* column) which integrate both video and language inputs for addressing the task. We also include a column that showcases the optimization techniques used to train those models.

Approach	Video	Frame	Language	Combined	Optimizer	RL
(Wang et al., 2019b)	I3D	-	LSTM	NMT+LSTM VI	ADAM	✗

Table 125: Exemplar *Machine Translation with Video* architectures.

Results. The models that have been created to address the task of *Machine Translation with Video* is built using a single dataset, namely VATEX. In Table 126 we present results obtained with a subset of models built using the VATEX dataset presented earlier in Section 7.2.2.

Model	B-4	METEOR
NMT+LSTM VI (Wang et al., 2019b) [English → Chinese]	30.20	-
NMT+LSTM VI (Wang et al., 2019b) [Chinese → English]	27.18	-

Table 126: Comparison of different methods on the VATEX dataset.

7.2.4 Machine Translation with Video - Discussion

In Table 126, we observe that only one method utilizing LSTM with video features from the pretrained I3D model (i.e., NMT+LSTM VI) is evaluated using the language metrics on the challenging VATEX dataset for both English and Chinese.

8. Language-to-Vision Generation

In this section, we explore the task of *Language-to-Vision Generation*. The goal of this task is to generate visual content given their natural language descriptions. However, different variations of the task exist and will be discussed in the following.

8.1 Language-to-Image Generation

In the following, we describe the setting of *Language-to-Image Generation* where an image is desired from a piece of natural language text (e.g., a sentence) describing the scene.

8.1.1 Language-to-Image Generation - Introduction

A litany of different variations of the Language-to-Image Generation exists. For example, generation of an image can also be thought as a manipulation of an image. It allows for the generation of a new image using desired natural language description. We present some variations in the following.

Sentence-level Language-to-Image Generation. The goal is to generate images conditioned on the natural language descriptions. It is considered as a fundamental problem in many applications. The success of Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) has made possible the generation of interesting images of specific categories, such as room interiors, album covers, and faces (Radford et al., 2016). This has led to an interest in bridging the gap between natural language text and image modeling. Figure 19 shows the usage of natural language description for generating image with a Text-to-Image Generation Model.

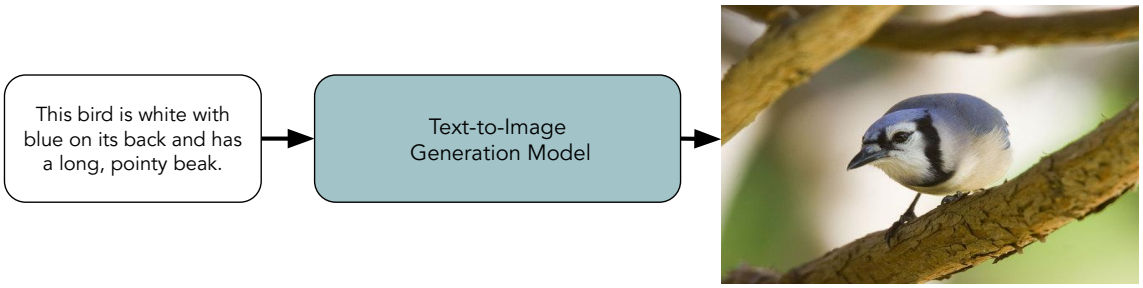

Figure 19: Given a *natural language description*, a Language-to-Image Model generates an image by conditioning on the provided description.

Initially, alignDRAW (Mansimov et al., 2016) was introduced to iteratively draw patches on a canvas, while attending to the relevant words in the description. Further, it was shown that visual concepts could be translated from characters to pixels (Reed et al., 2016b) with a conditional GAN. This was further improved by taking instructions about what content should be drawn in which location in order to achieve high-quality image generation (Reed et al., 2016a). Models which were developed to condition on classes for image generation (Nguyen et al., 2017) have also been used to generate images. However, the quality of images generated is much lower than when not conditioning on classes. Very close to this approach is Text-conditioned Auxiliary Classifier GAN (TAC-GAN) (Dash et al., 2017) which conditions images on both the sentence and class information, which has been shown to improve their structural coherence. To generate images with high resolution, several GANs were stacked together yielding stackGAN (Zhang et al., 2017, 2019) that used a global sentence representation. This helped generate images of different sizes. To overcome the bottleneck of global-level sentence representation, attention-based GAN like AttGAN (Xu

et al., 2018) was used to capture the fine-grained details at different sub-regions of the image. It pays attention to the relevant words in the natural language description.

In other research efforts, a hierarchical approach (Hong et al., 2018) was taken by inferring the semantic layout of the image. Instead of learning a direct description to an image mapping, the generation process is decomposed into multiple steps. First a semantic layout from the text is constructed by the layout generator. Then, the layout is converted to an image by the image generator. Other kinds of approaches such as HDGAN (Zhang et al., 2018) aim to accommodate hierarchical adversarial objectives inside the network to regularize mid-level representations and assist generator training in order to capture complex image information. This has been shown to generate images with high resolutions.

Later, instead of dealing with natural-language descriptions, Johnson et al. (2018) used image-specific scene graphs enabling explicitly reasoning about objects and their relationships. Further, for obtaining better high resolution images, coarse-resolution features were taken as input and Perceptual Pyramid Adversarial Network (PPAN) was introduced to directly synthesize multi-scale images conditioned on texts in an adversarial way (Gao et al., 2019). Another approach named MirrorGAN (Qiao et al., 2019) targets the main goal of visual realism and semantic consistency for generating images from text. It proposes global-local attention and semantics-preserving framework where the image generated from the text is further used to generate the text back. This has been shown to semantically align with the given text and generated description.

In the following, we explore some of the related ideas which expand the scope of language-to-image generation.

Image Manipulation. Image manipulation takes a different path from the earlier benchmark approaches about image generation, and so the TAGAN (Nam et al., 2018) was introduced to generate semantically manipulated images while preserving text-irrelevant contents. Here, the generator learns to generate images where only regions that correspond to the given text are modified. Another interesting approach is to have an interactive system that generates an image in an iterative manner. Recent approaches (Zhu et al., 2019) used attention in both the generator and the discriminator, while others (Li et al., 2020) have designed error correction modules to rectify mismatched attributes and complete the missing contents in the generated image. There are also other variations where the source image is manipulated via natural language dialogue (Cheng et al., 2018).

Fine-grained Image Generation. Fine-grained image generation uses a recurrent image generation model (El-Nouby et al., 2018) to take into account both the generated output up to the current step as well as all past instructions for generation. This has been shown to add new objects, apply simple transformations to existing objects, and correct previous mistakes. Earlier research never concentrated on fine-grained generation of images, i.e., localizing objects. Recently, control of the location of individual objects within an image was made possible (Hinz et al., 2019) by adding a pathway in an iterative manner and applying them at different locations specified by the bounding boxes to both the generator and the discriminator.

Sequential Image Generation. The sequential image generation approach StoryGAN (Li et al., 2019b), based on the sequential conditional GAN, concentrates on story by generating

a sequence of images, when given a multi-sentence paragraph. Termed as story visualization, it behaves exactly opposite to image storytelling and has been shown to generate images with high quality, while also achieving contextual consistency.

8.1.2 Language-to-Image Generation - Datasets

For image generation, existing image datasets have been modified to accommodate image descriptions. Initially, the Oxford-102[69] and Caltech-UCSD Birds (CUB)[70] datasets consisting of flower and bird images belonging to 102 and 200 classes respectively are expanded with image descriptions (Reed et al., 2016b). Table 127 and Table 128 presents splits of the datasets.

Split	Images	Captions per Image	Total Captions
Training	5,878	10	58,780
Validation	1,156	10	11,560
Test	1,155	10	11,550
Total	8,189	10	81,890

Table 127: Splits of the Oxford-102 dataset with image descriptions.

Split	Images	Captions per Image	Total Captions
Training	8,855	10	88,550
Validation	-	-	-
Test	2,933	10	29,330
Total	11,788	10	117,880

Table 128: Splits of the CUB dataset with image descriptions.

Similarly, the MSCOCO dataset (see Section 3.1.2) is also used for the reversed task of description generation, i.e., given a description, generate the image matching the description. We represent this dataset as MSCOCO-Gen. Table 129 presents the splits of the dataset.

Split	Images	Captions per Image	Total Captions
Training	82,783	5	413,915
Validation	-	-	-
Test	40,504	5	202,520
Total	123,287	5	616,435

Table 129: Splits of the MSCOCO-Gen dataset.

[69] http://www.robots.ox.ac.uk/~vgg/data/flowers/102
[70] http://www.vision.caltech.edu/visipedia/CUB-200-2011.html

8.1.3 Language-to-Image Generation - Evaluation Measures, Models, and Results

In this section, we review the measures used to evaluate different models of *Language-to-Image Generation* and the results obtained by them.

Evaluation Measures. There are different evaluation measures which are explicitly used for evaluating Language-to-Image generation models and are discussed below in detail.

- **Inception Score (IS)** (Salimans et al., 2016) was initially proposed to compare the quality of images generated by GAN models. A pretrained Inception-v3 model (Szegedy et al., 2016) is applied to the generated image to get the conditional label distribution with low entropy. A similar idea is applied for the generated images on the given text descriptions for automatic evaluation. Higher scores are better for IS.
- **Fréchet Inception Distance (FID)** (Heusel et al., 2017) is supposed to improve on IS by comparing the statistics of generated samples to original samples, instead of evaluating generated samples in an isolated manner. It also depends on the Inception-v3 model. In particular, the pool3 layer of the Inception-v3 is used for generating original samples for comparison. Lower FID is better as it corresponds to more similar generated and original samples.
- **R-precision** is inspired from the ranking retrieval results. It is used as a complementary evaluation metric for the language-to-image generation. Specifically, generated images are used to query their corresponding natural language descriptions to find how many relevant descriptions are retrieved.

Models. Many models have been created to approach the task of *Language-to-Image Generation*. In Table 130, we present some exemplar architectures (refer to *Combined* column) that integrate both image and language for addressing the task. We also include a column that showcases the optimization techniques used to train those models.

Approach	Image	Language	Combined	Optimizer	RL
(Reed et al., 2016b)	-	char-CNN-RNN	GAN-INT-CLS	ADAM	✗
(Reed et al., 2016a)	-	char-CNN-GRU	GAWWN	ADAM	✗
(Zhang et al., 2017)	-	-	StackGAN	ADAM	✗
(Xu et al., 2018)	Inception-v3	BiLSTM	AttGAN	-	✗
(Qiao et al., 2019)	-	BiLSTM	MirrorGAN	-	✗

Table 130: Exemplar *Language-to-Image Generation* architectures.

Results. In Table 131, Table 132, and Table 133 we present results obtained with a subset of models built using the CUB, Oxford-102, and COCO datasets presented earlier in Section 8.1.2.

8.1.4 Language-to-Image Generation - Discussion

The *Language-to-Image Generation* task has been evaluated using three different datasets. The CUB and Oxford-102 datasets contain only one visual object per image, while COCO

Model	Resolution	IS	FID	HR
GAN-INT-CLS (Reed et al., 2016b)	64x64	2.88 ± .04	68.79	2.76 ± .01
GAWWN (Reed et al., 2016a)	64x64	3.10 ± .03	53.51	-
	128x128	3.62 ± .07	72.65	1.95 ± .02
StackGAN (Zhang et al., 2017)	64x64	3.02 ± .03	35.11	-
	256x256	3.70 ± .04	51.89	1.29 ± .02
StackGAN++ (Zhang et al., 2019)	256x256	4.04 ± .05	15.30	1.19 ± .02
AttGAN (Xu et al., 2018)	256x256	4.36 ± .03	-	-
MirrorGAN (Qiao et al., 2019)	256x256	4.56 ± .05	-	-

Table 131: Comparison of different methods using generated images of different resolutions on the "CUB" dataset. R-precision (%) for 256x256 with AttGAN (53.31) and MirrorGAN (57.67). HR - Human Ranking.

Model	Resolution	IS	FID	HR
GAN-INT-CLS (Reed et al., 2016b)	64x64	2.66 ± .03	79.55	1.84 ± .02
StackGAN (Zhang et al., 2017)	64x64	2.73 ± .03	43.02	-
	256x256	3.20 ± .01	55.28	1.16 ± .02
StackGAN++ (Zhang et al., 2019)	256x256	3.26 ± .01	48.68	1.30 ± .03

Table 132: Comparison of different methods using generated images of different resolutions on the "Oxford-102" dataset.

Model	Resolution	IS	FID	HR
GAN-INT-CLS (Reed et al., 2016b)	64x64	7.88 ± .07	60.62	1.82 ± .03
StackGAN (Zhang et al., 2017)	64x64	8.35 ± .11	33.88	-
	256x256	8.45 ± .03	74.05	1.18 ± .03
StackGAN++ (Zhang et al., 2019)	256x256	8.30 ± .10	81.59	1.55 ± .05
PPGN (Nguyen et al., 2017)	256x256	9.58 ± .21	-	-
AttGAN (Xu et al., 2018)	256x256	25.89 ± .47	-	-
MirrorGAN (Qiao et al., 2019)	256x256	26.47 ± .41	-	-

Table 133: Comparison of different methods using generated images of different resolutions on the "COCO" dataset. R-precision (%) for 256x256 with AttGAN (72.13) and MirrorGAN (74.52).

has multiple objects. Several methods based on modified GAN objectives have been proposed for the generation of an image for a given textual description. From Table 131, Table 132, and Table 133 we observe the recent MirrorGAN (Qiao et al., 2019) achieves best results for different image resolution types using task-specific measures on CUB and COCO. It is built on the idea of back-translation of the image to text. However, for Oxford-102, StackGAN++ (Zhang et al., 2019) achieves the best result.

8.2 Language-to-Video Generation

In the following, we discuss the setting of *Language-to-Video Generation* where a video is desired as the visual output from a natural language text description of the scene in video.

8.2.1 Language-to-Video Generation - Introduction

The goal of Language-to-Video generation is to mimic language-to-image generation by considering the temporal aspect. However, language-to-video generation requires a stronger conditional generator than what is generally required for the language-to-image generation. This is because of the temporal nature of the videos. To address this challenge, a conditional generative model is trained (Li et al., 2018) to extract both static and dynamic information from text which combines variational autoencoders (VAE) (Kingma & Welling, 2014) with GANs. Figure 20 shows the usage of natural language description to generate a video with a text-to-video generation model.

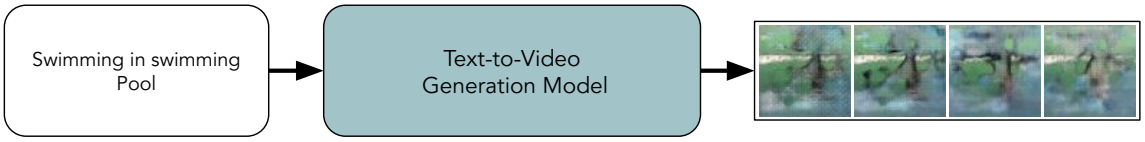

Figure 20: Given a *natural language description*, a Language-to-Video model generates a video (represented as sequence of frames from Li et al. (2018)) conditioned on the description.

Another novel approach is to generate video from script. The composition, retrieval, and fusion network (Craft) model (Gupta et al., 2018) is capable of learning knowledge from the video-description data and applying it in generating videos from novel captions. It has been shown that the Craft model performs better than the direct pixel generation approaches and generalizes well to unseen captions and to video databases with no text annotations.

8.2.2 Language-to-Video Generation - Datasets

For video generation there are no publicly available datasets. However, Li et al. (2018) have collected the Text2Video dataset belonging to ten different categories of YouTube videos, each ranging between 10-400 seconds for language-to-video generation. The categories of videos are biking in snow, playing hockey, jogging, playing soccer, playing football, kite surfing, playing golf, swimming, sailing and water skiing. For the purposes of model evaluation, the dataset is split into training, validation, and test sets in the ratio of 7:1:2 respectively, the details of which can be found in Table 134.

Split	Videos
Training	2800
Validation	400
Test	800

Table 134: Splits of Text2Video (Combines all categories).

8.2.3 Language-to-Video Generation - Evaluation Measures, Models, and Results

In this section, we review the measures used to evaluate different models of *Language-to-Video Generation* and the results obtained by them.

Evaluation Measures. The *Language-to-Video Generation* models are evaluated based on the Accuracy measure.

Models. Only a limited set of models have been created so far to handle the task of *Language-to-Video Generation*. In Table 135, we present an exemplar architecture (refer to *Combined* column) which integrates video and language to address the task. We also include a column that showcases the optimization technique used to train the model.

Approach	Video	Frame	Language	Combined	Optimizer	RL
(Li et al., 2018)	MotionFeatures	-	LSTM	T2V	ADAM	✗

Table 135: Exemplar *Language-to-Video Generation* architectures.

Results. In Table 136 we present results obtained with a subset of models built using the "TexttoVideo" dataset presented earlier in Section 8.2.2.

Model	Accuracy
DT2V-baseline (Li et al., 2018)	0.101
PT2V (Reed et al., 2016b)	0.134
GT2V (Li et al., 2018)	0.192
T2V (Li et al., 2018)	**0.426**

Table 136: Comparison of accuracy (%) scores of different models on Text2Video.

8.2.4 Language-to-Video Generation - Discussion

The task of *Language-to-Video Generation* is not as well-explored as the Language-to-Image generation task due to its complexity. Results presented in Table 136 show that the approach proposed by Li et al. (2018) achieves the best accuracy which is calculated using a simple video classifier which is a five-layer neural network model with 3D full convolutions and ReLU nonlinearities as activation functions.

9. Vision-and-Language Navigation

In this section, we explore the task of *Vision-and-Language Navigation*. The goal of this task is to carry out navigation in an environment by interpreting natural language instructions.

9.1 Image-and-Language Navigation

In the following, we provide a detailed description of the *Image-and-Language Navigation* task in which photorealistic images forming 3D environments are used as visual inputs.

9.1.1 Image-and-Language Navigation - Introduction

Most of the attempts at Vision-and-Language Navigation (VLN) use photorealistic images forming 3D environments. The goal of the Image-and-Language Navigation (ILN) task is to enable an autonomous agent (e.g., robot) to carry out navigation in an environment defined by the photo-realistic image views by means of interpreting natural language instructions (Anderson et al., 2018b). This requires the agent/robot to simultaneously process both vision and language inputs and navigate from a source to a target location. Figure 21 shows a visual depiction of the ILN task.

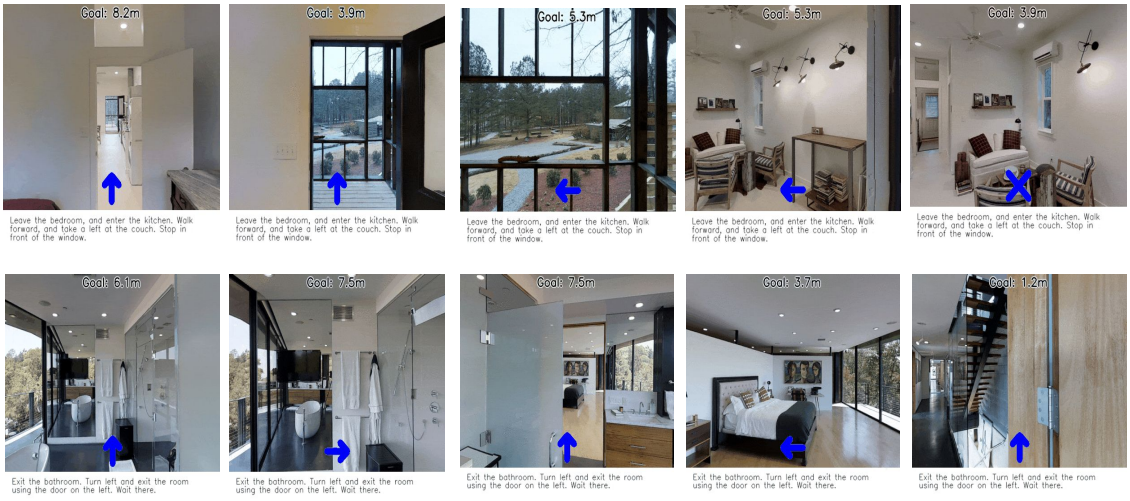

Figure 21: Given a *sequence of images* and few instructions in *textual* format (represented with a sequence of images from Anderson et al. (2018b)), an Image-and-Language Navigation model is expected to carry out the navigation of an agent in an environment (indicated by blue arrows in the pictures).

Initially, sequence-to-sequence models were proposed to address challenges in which the student-forcing approach achieved promising results in previously explored environments. One approach (Wang et al., 2018) integrated a module to combine model-based and model-free reinforcement learning techniques to better generalize to unseen environments. There is also the reinforced cross-modal matching approach (Wang et al., 2019a), which enforces both local and global cross-modal grounding via reinforcement learning.

ILN can also be viewed as a search on a navigation graph (Ma et al., 2019b) with a progress monitor as a learnable heuristic for search. It is improved by leveraging a visual-textual co-grounding attention mechanism to better align the instructions and visual scenes, and incorporates a progress monitor to estimate the agent's current progress towards the goal (Ma et al., 2019a). Another substantial improvement came from training an action space with an embedded speaker model (Fried et al., 2018). New instructions are synthesized for data augmentation and pragmatic reasoning was implemented for evaluating how well candidate action sequences explain an instruction. Improving over earlier approaches that make local action decisions or score entire trajectories using beam search, the novel approach of the FAST framework (Ke et al., 2019b) balances local and global signals when

exploring the environment allowing it to act greedily, but use global signals to backtrack when necessary. Also, Tan et al. (2019) explore a generalizable navigational agent by training it in two stages. In the first stage, mixed imitation and reinforcement learning is combined, while in the second stage, fine-tuning is performed via newly-introduced "unseen" triplets.

ILN can also be perceived as a form of visual question answering (see Section 5.1) that requires navigation to answer questions. Embodied Question Answering (Das et al., 2018a, 2018b) is explored with an agent that is spawned at a random location in a 3D environment and asked a question. For answering the question, the agent navigates through the 3D environment, for finding the information observed in the question. Other attempts used interactive question answering (Gordon et al., 2018) and grounded dialog (de Vries et al., 2018). Another set of approaches (Misra et al., 2018a) aims to map instructions to actions in 3D Environments with visual goal prediction. Recently, Chi et al. (2020) also made an interactive learning framework to endow the agent with the ability to ask for users' help in ambiguous situations.

9.1.2 Image-and-Language Navigation - Datasets

For the image-and-language navigation task, three different datasets have been designed so far. In the following, we present the details of these datasets in separate paragraphs.

Room-2-Room (R2R). The R2R[71] (Anderson et al., 2018b) dataset consists of real images of previously unseen building-scale 3D environments from Matterport3D (Chang et al., 2017). The navigation instructions have been collected with the help of humans using AMT. Table 137 presents splits of the dataset.

Split	Scenes	Navigation Instructions
Training	61	14,025
Validation (seen)	11	1,020
Validation (unseen)	11	2,349
Test	18	4,173

Table 137: Splits of the R2R dataset.

ASKNAV. Similar to R2R, the ASKNAV[72] (Nguyen et al., 2019) dataset is built on top of Matterport3D[73]. However, the objective differs in that the agent queries the advisor when in confusion and makes progress accordingly. It contains 10,800 panoramic views from 194,400 RGB-D images of 90 building-scale scenes. A data point in the dataset consists of a single starting viewpoint, but it has multiple goal viewpoints. Table 138 presents the splits of dataset.

TOUCHDOWN. Extending from building environments, the TOUCHDOWN[74] (Chen et al., 2019) dataset is designed for addressing tasks such as executing navigation instructions

[71] https://bringmeaspoon.org
[72] https://github.com/debadeepta/vnla
[73] https://niessner.github.io/Matterport
[74] https://github.com/lil-lab/touchdown

Split	Data points	Goals
Training	94,798	139,757
Validation (seen)	4,874	7,768
Validation (unseen)	5,005	8,245
Test (seen)	4,917	7,470
Test (unseen)	5,001	7,537

Table 138: Splits of the ASKNAV dataset.

(Navigation Only) and resolving spatial descriptions (SDR) in real-world environments. SDR is similar to the task of image referring expression (Section 4.1).

The *environment* includes 29,641 panoramas (360° Google Street View RGB images) and 61,319 edges from the New York City. Table 139 has more details about the dataset, while Table 140 presents its splits.

Dataset	Dataset Size	Vocab. Size	Mean Text Length
TOUCHDOWN (Complete task)	9,326	5,625	108.0
Navigation Only	9,326	4,999	89.6
SDR Only	25,575	3,419	29.7

Table 139: Statistics of the TOUCHDOWN dataset. *Vocabulary Size* and *Text Length* are computed by combining the training and validation sets.

Task	Split	Examples
Complete & Navigation Only	Training	6,526
	Validation	1,391
	Test	1,409
SDR Only	Training	17,880
	Validation	3,836
	Test	3,859

Table 140: Splits of the TOUCHDOWN dataset.

Cooperative Vision-and-Dialog Navigation (CVDN). CVDN[75] (Thomason et al., 2019) is a dataset[76] of embodied, human-human dialogs situated in a simulated, photorealistic home environment. Table 141 presents some statistics about the dataset.

Action Learning From Realistic Environments and Directives (ALFRED). ALFRED[77] (Shridhar et al., 2020) is a benchmark and interactive visual dataset for learning a mapping from natural language instructions and egocentric vision to sequences of actions for household tasks.

[75] https://cvdn.dev

[76] https://github.com/mmurray/cvdn/tree/master/tasks/CVDN/data

[77] https://askforalfred.com

Navigation Dialogs (Human-Human)	Navigation Trajectories	Total Scenes (MatterPort houses)
2,050	7,000	83

Table 141: Statistics of the CVDN dataset.

Data Split	Fold	Number of Scenes	Number of Annotations
Training	-	108	21,023
Validation	Seen	88	820
	Unseen	4	821
Testing	Seen	107	1,533
	Unseen	8	1,529

Table 142: Splits of the ALFRED dataset.

9.1.3 Image-and-Language Navigation - Evaluation Measures, Models, and Results

In this section, we present the evaluation measures, models, and results achieved with various architectures of *Image-and-Language Navigation*.

Evaluation Measures. The measures that are designed explicitly for the *Image-and-Language Navigation* system (e.g., R2R) are:

- **Path Length (PL):** PL is a trajectory length where it is the total length of the executed path.
- **Navigation Error (NE):** NE is based on the shortest path distance in the navigation graph, and is calculated by measuring the average distance between the end-location predicted by the follower agent and the true route's end-location.
- **Success Rate (SR):** SR is the percentage of predicted end-locations within 3 meters of the true location.
- **Oracle Success Rate (OSR):** OSR measures the success rate at the closest point to the goal that the agent has visited along the trajectory.
- **Success Path Length (SPL):** SPL is a trade-off between SR and PL, by weighting SR by inverse PL.

Models. Many models have been created to approach the task of *Image-and-Language Navigation*. In Table 143, we present some exemplar architectures (refer to *Combined* column) which integrate both image and language to address the task. We also include a column that showcases the optimization techniques used to train those models.

Results. As discussed earlier several models have been created to approach the task of *Image-and-Language Navigation*. Furthermore, many datasets have been created to provide variety in the content so that they improve the generalization ability of the models. In this

Approach	Image	Language	Combined	Optimizer	RL
(Anderson et al., 2018b)	ResNet-152	LSTM	Seq-to-Seq	ADAM	✗
(Wang et al., 2018)	ResNet-152	LSTM	RPA	-	✓
(Fried et al., 2018)	ResNet-152	LSTM	Speaker-Follower	-	✓
(Wang et al., 2019a)	ResNet-152	LSTM	RCM	ADAM	✓
(Ma et al., 2019a)	ResNet-152	LSTM	Self-Monitoring	ADAM	✗
(Tan et al., 2019)	ResNet-152	LSTM	BackTranslation	RMSprop	✓
(Ke et al., 2019b)	-	LSTM	FAST	-	✗

Table 143: Exemplar *Image-and-Language Navigation* architectures.

section, we cover the results obtained by the models from a representative dataset for this task. Table 144 – 146 present results obtained with a subset of models built and evaluated using the R2R dataset which was introduced in Section 9.1.2.

Model	PL	NE	OSR	SR	SPL
Random	9.89	9.79	18.3	13.2	12
Seq-to-Seq (Anderson et al., 2018b)	8.13	7.85	26.6	20.4	18
RPA (Wang et al., 2018)	9.15	7.53	32.5	25.3	23
Speaker-Follower (Fried et al., 2018)	14.82	6.62	44.0	35.0	28
Self-Monitoring (Ma et al., 2019a)	18.0	-	-	48.0	35
RCM (Wang et al., 2019a)	15.22	6.01	50.8	43.1	35
BackTranslation-Single (Tan et al., 2019)	11.7	-	-	**51.5**	**47**
TacticalRewind-Greedy (Ke et al., 2019b)	22.08	5.14	-	54	41
BackTranslation-PreExplore (Tan et al., 2019)	9.79	-	-	63.9	61
BackTranslation-Beam (Tan et al., 2019)	687	-	-	68.9	1
FAST-Beam (Ke et al., 2019b)	196.53	4.29	-	61.0	3

Table 144: Comparison of different methods on the R2R test set.

Model	PL	NE	OSR	SR	SPL
Speaker-Follower (Fried et al., 2018)	-	3.36	73.8	66.4	-
RCM+SIL (Wang et al., 2019a)	10.13	2.78	79.7	73.0	-
BackTranslation-Single (Tan et al., 2019)	11.0	3.99	-	62.1	59
TacticalRewind-Greedy (Ke et al., 2019b)	-	-	-	-	-
BackTranslation-PreExplore (Tan et al., 2019)	9.92	4.84	-	54.7	52
BackTranslation-Beam (Tan et al., 2019)	703	2.52	-	75.7	1
FAST-Beam (Ke et al., 2019b)	188.6	3.13	-	70.0	4

Table 145: Comparison of different methods on the *seen* validation set of R2R.

9.1.4 Image-and-Language Navigation - Discussion

Image-and-Language Navigation is evaluated with different splits of the R2R validation and test datasets. From Table 144, Table 145, and Table 146 we can observe that Frontier Aware Search with backTracking (FAST)-beam (Ke et al., 2019b) achieves the best result on the

Model	PL	NE	OSR	SR	SPL
Speaker-Follower (Fried et al., 2018)	-	3.36	73.8	66.4	-
RCM+SIL (Wang et al., 2019a)	10.13	2.78	79.7	73.0	-
BackTranslation-Single (Tan et al., 2019)	10.7	5.22	-	52.2	48
TacticalRewind-Greedy (Ke et al., 2019b)	21.17	4.97	-	56.0	43
BackTranslation-PreExplore (Tan et al., 2019)	9.57	3.78	-	64.5	61
BackTranslation-Beam (Tan et al., 2019)	663	3.08	-	69.0	1
FAST-Beam (Ke et al., 2019b)	224.42	4.03	-	63.0	2

Table 146: Comparison of different methods on the *unseen* validation set of R2R.

task-specific metrics. This approach balances local and global signals while exploring an unobserved environment. It also helps to act greedily but use global signals to backtrack whenever necessary.

10. Vision-and-Language Pretraining

Inspired by the works of pretraining only on vision (He et al., 2016) or solely on language data (Devlin et al., 2019; Radford et al., 2019; Brown et al., 2020a), the vision-and-language pretraining seeks to jointly learn representations using both visual and textual content for improving the efficiency of previously discussed vision and language integration tasks. Several methods will be discussed for vision-and-language pretraining, the architectures of which can be broadly divided into *Single-stream* and *Two-stream*. In the following, we provide more details on both types of architectures.

Single-stream Architectures. These neural architectures are based on BERT-like (Devlin et al., 2019) models where they incorporate an Image Embedder, a Text Embedder, and a multi-layer Transformer (Vaswani et al., 2017). The proposed models are pretrained on data which in general have parallel multimodal components i.e., videos or images along with captions. Further, the models are optimized with a combination of different objectives such as vision-based and text-based Masked Language Models (MLM), masked visual-feature modeling, and visual-linguistic matching. Learned representations are then used for different downstream tasks such as multimodal understanding or generation. For example, the VideoBERT (Sun et al., 2019) architecture has been designed to learn vision-language representations for a generative downstream task like video description generation (see Section 3.1.4). While there are several other approaches such as Bounding Boxes in Text Transformer (B2T2) (Alberti et al., 2019), Unicoder-VL (Li et al., 2020), VL-BERT (Su et al., 2020), UNITER (Chen et al., 2020) are all designed for multimodal understanding and facilitate downstream tasks. Works such as VLP (Zhou et al., 2020), OSCAR (Li et al., 2020) and also its extension VinVL (Zhang et al., 2021) have built unified models that can jointly understand and generate from cross-modal data. There is also an emergence of interest in probing vision-and-language pretrained models (Cao et al., 2020) to comprehend the contribution from each modality and also help in designing better model architectures and objectives.

Two-stream Architectures. In contrast to the single-stream architectures, two-stream architectures adopted two independent encoders for learning visual and text representations. ViLBERT (Lu et al., 2019) and LXMERT (Tan & Bansal, 2019) are examples of two-stream architectures which used self-attention principles to jointly learn representations from visual and textual data. ViLBERT builds a co-attentional transformer layer, while LXMERT uses a cross-modality encoder. Similar to single-stream, the two-stream architectures also optimize their models with pretraining tasks, such as MLM and vision-text matching. Sometimes they use additional text-only corpora for achieving better generalization on long and complex sentences.

In Table 147, we summarize both *Single-stream* and *Two-stream* architectures by enumerating the vision and language integration tasks they support. It has to be noted that these architectures only use *subsets of the datasets* from each task. Also, the type of tasks they select are limited and are mostly discriminative. Broadly, we denote with (✓) or (✗) to indicate whether they support the task in question or not.

Approach	VDG	VS	VRE	VQA	VR	VE	VDiag	MMT	LVG	VLN
Single-stream										
Unicoder-VL	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗
VL-BERT	✗	✗	✓	✓	✓	✗	✗	✗	✗	✗
VideoBERT	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
VLP	✓	✗	✗	✓	✗	✗	✗	✗	✗	✗
OSCAR	✓	✗	✗	✓	✓	✗	✗	✗	✗	✗
B2T2	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗
UNITER	✗	✗	✓	✓	✓	✓	✗	✗	✗	✗
VinVL	✓	✗	✗	✓	✓	✗	✗	✗	✗	✗
Two-stream										
ViLBERT	✗	✗	✓	✓	✓	✗	✗	✗	✗	✗
LXMERT	✗	✗	✗	✓	✓	✗	✗	✗	✗	✗

Table 147: Major Vision-and-Language Pretraining Architectures and their support for various Vision and Language Tasks. **VDG** - Visual Description Generation, **VS** - Visual Storytelling, **VRE** - Visual Referring Expression, **VQA** - Visual Question Answering, **VR** - Visual Reasoning, **VE** - Visual Entailment, **VDiag** - Visual Dialog, **MMT**- Multimodal Machine Translation, **LVG** - Language-to-Vision Generation, **VLN** - Vision-and-Language Navigation.

11. Future Directions

The integration of vision and language research has come a long way since the pioneering works, particularly after the adoption of deep learning techniques. Although the performance of current state-of-the-art models still needs to catch up with human abilities, the gap is diminishing at a steady rate. However, there is still ample room for theoretical and algorithmic improvements. Here, we enumerate several possible future directions that have the potential to advance the research overall.

Learning Common Sense and World Knowledge. There is a vast amount of out-of-domain data available which is unpaired with vision and language task-specific corpora. Leveraging such information as factual, hierarchical, or commonsense knowledge can significantly improve the intelligence of vision and language systems. Prior works have been shown to assist independent NLP tasks with pretrained language models such as commonsense reasoning (Rajani et al., 2019) and fact predictions (IV et al., 2019). It has also shown promise for image caption generation (Wu et al., 2018; Mogadala et al., 2018a) and question answering (Shah et al., 2019a; Marino et al., 2019). Extending such ideas to other tasks would be an interesting research direction to pursue. Another possibility could be to utilize images, videos, and text in a synchronous and synergistic manner as they encode different aspects of the world and implicitly. Here, an open question would be how to extract world and common sense knowledge from these sources.

Addressing Large-scale Data Limitations. Most approaches designed for tasks that integrate vision and language use large datasets for training. With this trend, it will soon become harder to design new tasks without having a dataset. To avoid such problems, future work will need to be adaptable to datasets of different sizes. Therefore, trade-off approaches are required where we know what amount of data is enough to master a certain task. This requires designing methods which might inspire from neuro-symbolic reasoning systems (Yi et al., 2018; Vedantam et al., 2019).

Combining Multiple Tasks. Some tasks are capable of sharing some ideas or representations of each other. For example, visual referring expression comprehension can be viewed as a visual dialog task (de Vries et al., 2017) where a sequence of questions is used to refer to an object in the image. Similarly, image caption generation can be viewed as the visual referring expression generation task (Mao et al., 2016).

Novel Neural Architectures for Representation. Up until late 2017, the de facto standard for learning language and visual feature representations were RNNs and CNNs respectively. However, over the last few years, with the introduction of novel ideas that address the limitations of aforementioned neural network types, either theoretically or computationally, there is a growing interest to adopt these new techniques. For instance, the Transformer (Vaswani et al., 2017) architecture that is used extensively for pure NLP tasks may see adoption for the integration of vision and language tasks. It has already shown its applicability for image caption generation (Sharma et al., 2018). In a similar manner, graph neural networks (Scarselli et al., 2008; Kipf & Welling, 2017; Battaglia et al., 2018) that were introduced to tackle graph-structured data, has already shown its promise in visual reasoning (Haurilet et al., 2019). Exploiting the compositionality of visual objects to describe an entire visual scene with neural modular networks is also an interesting direction to explore for many vision and language tasks.

Image vs Video. Most of the research into integrating vision and language concentrates on static images. This trend is clearly evident from the array of datasets and methods available for image and language integration tasks. Nevertheless, although a complex task, similar attention needs to be embraced for videos for which there is a scarcity of datasets. For instance, there is only one dataset available for tasks such as Video Dialog (Section 6.2), Video Referring Expression (Section 4.2), Language-to-Video Generation (Section 8.2), and

Machine Translation with Videos (Section 7.2), while tasks such as Vision-and-Language Navigation (Section 9) completely lack video-based datasets.

3D-Vision and Language. The world that we inhabit is inherently 3D. Thinking from this perspective, restricting vision and language research to just 2D, viz. images and videos, might be a hindrance for real world agents, e.g., humanoid robots, to fully understand the complexities of the 3D world and navigate with ease. To avoid such pitfalls, algorithms and techniques need to be developed for processing 3D inputs such as RGB-D, meshes, and point clouds in conjunction with language. Some pioneering works have already begun in this direction (Achlioptas et al., 2020; Chen et al., 2020; Liu et al., 2021; Roh et al., 2021) and we anticipate the trend[78] to shift more towards developing algorithms for understanding as well as the generation of 3D scenes (Briq et al., 2021), while utilizing language as a main or auxiliary modality.

Automatic Evaluation Measures. Automatic evaluation measures exist for several vision and language tasks. However, most of them are adaptations from standalone NLP tasks such as machine translation. For example, BLEU and METEOR metrics used for evaluating visual caption generation and storytelling models have been found not to correlate well with human judgements (Bernardi et al., 2016). The SPICE metric designed specifically for visual caption generation is dependent on parsing and is, therefore, not adaptable for other tasks such as storytelling. This kind of shortcoming shows us a promising research direction to pursue in developing evaluation measures applicable for several tasks. Recent attempts in developing BLEURT (Sellam et al., 2020) and BERTScore (Zhang et al., 2020) metrics show promising direction towards this goal. Analogously, language-to-vision generation, although having quantitative measures, is typically dependent on human evaluation. It needs to adopt novel techniques for effective quantitative evaluation. Other tasks such as vision-and-language navigation and visual reasoning have specific measures for evaluation which can be improved further.

12. Conclusion

In undertaking this survey, we provided an overview as well as elaborate details on the recent trends in integration of vision and language research. In the beginning, we started with a background on various tasks in computer vision and NLP. Then, we identified ten distinct prominent tasks that aim to integrate visual and language modalities. To draw connections from traditional research tasks to V&L integration tasks, we presented information about how each integration task is expanded from the standalone computer vision or NLP tasks on which they are originally based. Following that, we reviewed and analyzed each task separately by presenting a comprehensive introduction on how the tasks are designed in a bottom-up manner. Additionally, we presented different state-of-the-art methods used to address the tasks, along with exemplar architectures that are designed to integrate vision and language representations. We also provided a review on relevant datasets, evaluation measures, and the relative performance obtained by several state-of-the-art methods. Finally, in a separate section, we explored the various ways to pretrain generic models with large-scale multimodal data for supporting downstream vision and language integration tasks

[78] `https://language3dscenes.github.io`

with minimal fine-tuning efforts. Moreover, we outlined how much the existing pretraining approaches support the ten prominent integration tasks that we described in earlier sections.

When comparing the standalone research done individually in the fields of computer vision and NLP, the synergy of both, fuelled by advanced machine learning techniques, are expected to yield more intelligent and sustainable systems. Making them easily accessible can, therefore, have direct commercial and societal impact. However, despite the significant progress achieved so far in many integration tasks, large-scale evaluation of those systems show that they still fall behind human performance, by a large margin. This fact confirms that there is still a good deal of room for improvement. In particular, designing novel evaluation measures and architectures that can adequately deal with the complexity of vision and language integration problems has the potential to address some of the challenges. Towards this goal, we outlined a few possible future research directions in the final section.

We believe that our efforts in publishing this survey will help to systematize future research papers and also investigate the unsolved problems that are hindering the progress of effective integration of vision and language modalities.

Acknowledgments

This work was supported by the German Research Foundation (DFG) as a part of - Project-ID 232722074 - SFB1102. We extend our special thanks to Matthew Kuhn and Stephanie Lund for painstakingly proofing the whole manuscript. We also acknowledge the insightful comments of Marius Mosbach on the first version of the manuscript.

Appendix

ADAM ADAptive Moment estimation

AI Artificial Intelligence

AMEM Attention MEMory

AMT Amazon Mechanical Turk

ANetCap ActivityNet Captions

AREL Adversarial REward Learning

AVSD Audio Visual Scene-aware Dialog

BDD-X Berkeley Deep Drive eXplanation

BDD Berkeley Deep Drive

BERT Bidirectional Encoder Representations from Transformers

BiLSTM Bidirectional LSTM

BLEU BiLingual Evaluation Understudy

BRNN Bidirectional Recurrent Neural Network

CIDEr Consensus based Image Description Evaluation

CLEVR-CoGenT CLEVR Compositional Generalization Test

CLEVR Compositional Language and Elementary Visual Reasoning

CLID Cross-Lingual Image Description

CMM Cascaded Mutual Modulation

CMRE Cross-Modal Relationship Extractor

CNN Convolutional Neural Networks

CoAtt-GAN Co-Attention GAN

COG Configurable Visual Question and Answer

CorefNMN Coreference Neural Module Networks

CUB Caltech-UCSD Birds

DII Descriptions of Images-in-Isolation

DIS Descriptions of Images-in-Sequence

EVE Explainable Visual Entailment

FAST Frontier Aware Search with backTracking

FID Fréchet Inception Distance

FiLM Feature-wise Linear Modulation

GAN Generative Adversarial Network

GAWWN Generative Adversarial What-Where Network

GGCN Gated Graph Convolutional Network

GNN Graph Neural Network

GQA General Question Answering

GRU Gated Recurrent Unit

GVF Global Visual Features

HCIAE-NP-ATT History-Conditioned Image Attentive Encoder n-pair self-ATTentive

HREA Hierarchical Recurrent Encoder with Attention

HRE Hierarchical Recurrent Encoder

KVQA Knowledge-aware Visual Question Answering

LBA Learning-By-Asking

LF Late Fusion

LGCN Language-Conditioned Graph Networks

LSTM Long Short-Term Memory

LXMERT Learning Cross-Modality Encoder Representations from Transformers

M-VAD Montreal Video Annotation

MAC Memory, Attention, and Composition

MedRank Median Rank

METEOR Metric for Evaluation of Translation with Explicit Ordering

MIL Multiple Instance Learning

MMT Multimodal Machine Translation

MN Memory Network

MPII-MD MPII Movie Description

MPII Max Planck Institute for Informatics

MRR Mean Reciprocal Rank

MSCOCO Microsoft Common Objects in COntext

MSR-VTT Microsoft Research Video to Text

MSVD Microsoft Video Description

MuRel Multimodal Relational network

NDCG Normalized Discounted Cumulative Gain

NIC Neural Image Captioning

NMT Neural Machine Translation

NS-CL Neuro-Symbolic Concept Learner

NYC-Storytelling New York City Storytelling

OK-VQA Outside Knowledge Visual Question Answering

OSCAR Object-SemantiCs Aligned pRe-training

PPGN Plug & Play Generative Network

R-CNN Region-based CNN

R2R Room-2-Room

RAVEN Relational and Analogical Visual rEasoNing

RCM Reinforced Cross-modal Matching

RE Referring Expression

RGB-D Red, Green, Blue, Depth

RL Reinforcement Learning

RNs Relation Networks

ROUGE Recall Oriented Understudy for Gisting Evaluation

RvA Recursive Visual Attention

SCA Sequential Co-Attention

SCRC Spatial Context Recurrent Convnet

SDR Spatial Description Resolution

SF Similarity scoring + Fusion

SGD Stochastic Gradient Descent

SIL Self-supervised Imitation Learning

SIND Sequential Image Narrative Dataset

SIS Stories for Images-in-Sequence

SPICE Semantic Propositional Image Captioning Evaluation

TACoS Textually Annotated Cooking Scenes

UNITER UNiversal Image-TExt Representation Learning

V-SNLI Visually-grounded Natural Language Inference

VATEX Video And TEXt

VDG Visual Description Generation

VGG Visual Geometry Group

VIST Visual Storytelling

VLN Vision and Language Navigation

VLP Vision-Language Pre-training

VMMT$_F$ Variational MMT with Fixed Gaussian prior

VRE Visual Referring Expression

VTW Videos Titles in the Wild

References

Aafaq, N., Mian, A., Liu, W., Gilani, S. Z., & Shah, M. (2020). Video description: A survey of methods, datasets, and evaluation metrics. *ACM Comput. Surv.*, *52*(6), 115:1–115:37.

Achlioptas, P., Abdelreheem, A., Xia, F., Elhoseiny, M., & Guibas, L. (2020). Referit3d: Neural listeners for fine-grained 3D object identification in real-world scenes. In *16th European Conference on Computer Vision (ECCV), August 23-28, 2020*. Springer.

Aditya, S., Saha, R., Yang, Y., & Baral, C. (2019). Spatial knowledge distillation to aid visual reasoning. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 227–235. IEEE.

Agrawal, A., Batra, D., Parikh, D., & Kembhavi, A. (2018). Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4971–4980. IEEE Computer Society.

Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C. L., Parikh, D., & Batra, D. (2017). Vqa: Visual question answering. *International Journal of Computer Vision*, *123*(1), 4–31.

Agrawal, H., Anderson, P., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., & Lee, S. (2019). nocaps: novel object captioning at scale. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 8947–8956. IEEE.

Agrawal, H., Chandrasekaran, A., Batra, D., Parikh, D., & Bansal, M. (2016). Sort story: Sorting jumbled images and captions into stories. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 925–931.

Alamri, H., Cartillier, V., Das, A., Wang, J., Cherian, A., Essa, I., Batra, D., Marks, T. K., Hori, C., Anderson, P., Lee, S., & Parikh, D. (2019a). Audio visual scene-aware dialog. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 7558–7567. Computer Vision Foundation / IEEE.

Alamri, H., Hori, C., Marks, T. K., Batra, D., & Parikh, D. (2019b). Audio visual scene-aware dialog (avsd) track for natural language generation in dstc7. In *DSTC7 workshop at AAAI*.

Alberti, C., Ling, J., Collins, M., & Reitter, D. (2019). Fusion of detected objects in text for visual question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2131–2140.

Anderson, P., Fernando, B., Johnson, M., & Gould, S. (2016). SPICE: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pp. 382–398. Springer.

Anderson, P., Fernando, B., Johnson, M., & Gould, S. (2017). Guided open vocabulary image captioning with constrained beam search. In *EMNLP*.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018a). Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6077–6086.

Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., & van den Hengel, A. (2018b). Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3674–3683.

Andreas, J., Rohrbach, M., Darrell, T., & Klein, D. (2016a). Learning to compose neural networks for question answering. In Knight, K., Nenkova, A., & Rambow, O. (Eds.), *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pp. 1545–1554. The Association for Computational Linguistics.

Andreas, J., Rohrbach, M., Darrell, T., & Klein, D. (2016b). Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 39–48. IEEE Computer Society.

Aneja, J., Deshpande, A., & Schwing, A. G. (2018). Convolutional image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5561–5570.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., & Parikh, D. (2015). Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433.

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In Bengio, Y., & LeCun, Y. (Eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Bai, S., & An, S. (2018). A survey on automatic image caption generation. *Neurocomputing*, *311*, 291–304.

Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(2), 423–443.

Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72.

Barnard, K., Duygulu, P., Forsyth, D. A., de Freitas, N., Blei, D. M., & Jordan, M. I. (2003). Matching words and pictures. *Journal of Machine Learning Research*, *3*, 1107–1135.

Barrault, L., Bougares, F., Specia, L., Lala, C., Elliott, D., & Frank, S. (2018). Findings of the third shared task on multimodal machine translation. In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Jimeno-Yepes, A., Koehn, P., Monz, C., Negri, M., Névéol, A., Neves, M. L., Post, M., Specia, L., Turchi,

M., & Verspoor, K. (Eds.), *Proceedings of the Third Conference on Machine Translation: Shared Task Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pp. 304–323. Association for Computational Linguistics.

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., et al. (2018). Relational inductive biases, deep learning, and graph networks. *CoRR*, *abs/1806.01261*.

Baumann, A., Boltz, M., Ebling, J., Koenig, M., Loos, H., Merkel, M., Niem, W., Warzelhan, J., & Yu, J. (2008). A review and comparison of measures for automatic video surveillance systems. *EURASIP Journal on Image and Video Processing*, *2008*(1), 824726.

Bengio, Y., LeCun, Y., & Hinton, G. E. (2021). Deep learning for AI. *Commun. ACM*, *64*(7), 58–65.

Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., & Plank, B. (2016). Automatic description generation from images: A survey of models, datasets, and evaluation measures.. *Journal of Artificial Intelligence Research (JAIR)*, *55*, 409–442.

Blösch, M., Weiss, S., Scaramuzza, D., & Siegwart, R. (2010). Vision based mav navigation in unknown and unstructured environments. In *2010 IEEE International Conference on Robotics and Automation*, pp. 21–28. IEEE.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pp. 177–186. Springer.

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In Màrquez, L., Callison-Burch, C., Su, J., Pighin, D., & Marton, Y. (Eds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pp. 632–642. The Association for Computational Linguistics.

Briq, R., Kochar, P., & Gall, J. (2021). Towards better adversarial synthesis of human images from text. *CoRR*, *abs/2107.01869*.

Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J. D., Mercer, R. L., & Roossin, P. S. (1990). A statistical approach to machine translation. *Computational linguistics*, *16*(2).

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020a). Language models are few-shot learners. In *arXiv preprint arXiv:2005.14165*.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., et al. (2020b). Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., & Lin, H. (Eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Burke, H. R. (1958). Raven's progressive matrices: A review and critical evaluation. *The Journal of Genetic Psychology*, *93*(2), 199–228.

Cadène, R., Ben-younes, H., Cord, M., & Thome, N. (2019). MUREL: multimodal relational reasoning for visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 1989–1998. Computer Vision Foundation / IEEE.

Caglayan, O., Aransa, W., Bardet, A., García-Martínez, M., Bougares, F., Barrault, L., Masana, M., Herranz, L., & van de Weijer, J. (2017). LIUM-CVC submissions for WMT17 multimodal translation task. In Bojar, O., Buck, C., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno-Yepes, A., Koehn, P., & Kreutzer, J. (Eds.), *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017*, pp. 432–439. Association for Computational Linguistics.

Caglayan, O., Madhyastha, P., Specia, L., & Barrault, L. (2019). Probing the need for visual context in multimodal machine translation. In Burstein, J., Doran, C., & Solorio, T. (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4159–4170. Association for Computational Linguistics.

Calixto, I., & Liu, Q. (2017). Incorporating global visual features into attention-based neural machine translation. In Palmer, M., Hwa, R., & Riedel, S. (Eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pp. 992–1003. Association for Computational Linguistics.

Calixto, I., Liu, Q., & Campbell, N. (2017). Doubly-attentive decoder for multi-modal neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1, pp. 1913–1924.

Calixto, I., Rios, M., & Aziz, W. (2019). Latent variable model for multi-modal translation. In Korhonen, A., Traum, D. R., & Màrquez, L. (Eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 6392–6405. Association for Computational Linguistics.

Cao, J., Gan, Z., Cheng, Y., Yu, L., Chen, Y.-C., & Liu, J. (2020). Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *16th European Conference on Computer Vision (ECCV), August 23-28, 2020*. Springer.

Cao, Q., Liang, X., Li, B., Li, G., & Lin, L. (2018). Visual question reasoning on general dependency tree. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7249–7257.

Cao, Y., Long, M., Wang, J., Yang, Q., & Yu, P. S. (2016). Deep visual-semantic hashing for cross-modal retrieval. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1445–1454. ACM.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In *16th European Conference on Computer Vision (ECCV), August 23-28, 2020*. Springer.

Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308.

Chang, A. X., Dai, A., Funkhouser, T. A., Halber, M., Nießner, M., Savva, M., Song, S., Zeng, A., & Zhang, Y. (2017). Matterport3d: Learning from RGB-D data in indoor environments. In *2017 International Conference on 3D Vision, 3DV 2017, Qingdao, China, October 10-12, 2017*, pp. 667–676. IEEE Computer Society.

Chang, S., Yang, J., Park, S., & Kwak, N. (2018). Broadcasting convolutional network for visual relational reasoning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 754–769.

Chatterjee, M., & Schwing, A. G. (2018). Diverse and coherent paragraph generation from images. In Ferrari, V., Hebert, M., Sminchisescu, C., & Weiss, Y. (Eds.), *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part II*, Vol. 11206 of *Lecture Notes in Computer Science*, pp. 747–763. Springer.

Chen, D. Z., Chang, A. X., & Nießner, M. (2020). Scanrefer: 3d object localization in RGB-D scans using natural language. In *16th European Conference on Computer Vision (ECCV), August 23-28, 2020*. Springer.

Chen, D. L., & Dolan, W. B. (2011). Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pp. 190–200. Association for Computational Linguistics.

Chen, H., Suhr, A., Misra, D., Snavely, N., & Artzi, Y. (2019). Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12538–12547.

Chen, T., Liao, Y., Chuang, C., Hsu, W. T., Fu, J., & Sun, M. (2017). Show, adapt and tell: Adversarial training of cross-domain image captioner. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 521–530. IEEE Computer Society.

Chen, X., & Lawrence Zitnick, C. (2015). Mind's eye: A recurrent visual representation for image caption generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2422–2431.

Chen, Y.-C., Li, L., Yu, L., Kholy, A. E., Ahmed, F., Gan, Z., Cheng, Y., & Liu, J. (2020). Uniter: Universal image-text representation learning. In *16th European Conference on Computer Vision (ECCV), August 23-28, 2020*. Springer.

Cheng, Y., Gan, Z., Li, Y., Liu, J., & Gao, J. (2018). Sequential attention GAN for interactive image editing via dialogue. *CoRR*, *abs/1812.08352*.

Chi, T., Shen, M., Eric, M., Kim, S., & Hakkani-Tür, D. (2020). Just Ask: An interactive learning framework for vision and language navigation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI*

Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, pp. 2459–2466. AAAI Press.

Cho, K., van Merrienboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Moschitti, A., Pang, B., & Daelemans, W. (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 1724–1734. ACL.

Chrupała, G., Gelderloos, L., & Alishahi, A. (2017). Representations of language in a model of visually grounded speech signal. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1, pp. 613–622.

Chung, J., Gülçehre, Ç., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, *abs/1412.3555*.

Cirik, V., Berg-Kirkpatrick, T., & Morency, L.-P. (2018a). Using syntax to ground referring expressions in natural images. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Cirik, V., Morency, L., & Berg-Kirkpatrick, T. (2018b). Visual referring expression recognition: What do systems actually learn?. In Walker, M. A., Ji, H., & Stent, A. (Eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pp. 781–787. Association for Computational Linguistics.

Condoravdi, C., Crouch, D., De Paiva, V., Stolle, R., & Bobrow, D. G. (2003). Entailment, intensionality and text understanding. In *Proceedings of the HLT-NAACL 2003 workshop on Text meaning*.

Conneau, A., & Lample, G. (2019). Cross-lingual language model pretraining. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., & Garnett, R. (Eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pp. 7057–7067.

Cornia, M., Baraldi, L., & Cucchiara, R. (2019). Show, control and tell: A framework for generating controllable and grounded captions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 8307–8316. Computer Vision Foundation / IEEE.

Cornia, M., Stefanini, M., Baraldi, L., & Cucchiara, R. (2020). Meshed-memory transformer for image captioning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 10575–10584. IEEE.

Dai, B., Fidler, S., Urtasun, R., & Lin, D. (2017). Towards diverse and natural image descriptions via a conditional GAN. In *IEEE International Conference on Computer*

Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, pp. 2989–2998. IEEE Computer Society.

Dai, B., & Lin, D. (2017). Contrastive learning for image captioning. In *Advances in Neural Information Processing Systems*, pp. 898–907.

Das, A., Datta, S., Gkioxari, G., Lee, S., Parikh, D., & Batra, D. (2018a). Embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2054–2063.

Das, A., Gkioxari, G., Lee, S., Parikh, D., & Batra, D. (2018b). Neural modular control for embodied question answering. In *CoRL*, Vol. 87 of *Proceedings of Machine Learning Research*, pp. 53–62. PMLR.

Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J. M. F., Parikh, D., & Batra, D. (2017a). Visual dialog. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 1080–1089. IEEE Computer Society.

Das, A., Kottur, S., Moura, J. M. F., Lee, S., & Batra, D. (2017b). Learning cooperative visual dialog agents with deep reinforcement learning. In *ICCV*, pp. 2970–2979. IEEE Computer Society.

Das, P., Xu, C., Doell, R. F., & Corso, J. J. (2013). A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2634–2641. IEEE Computer Society.

Dash, A., Gamboa, J. C. B., Ahmed, S., Liwicki, M., & Afzal, M. Z. (2017). TAC-GAN - text conditioned auxiliary classifier generative adversarial network. *CoRR*, *abs/1703.06412*.

De Mulder, W., Bethard, S., & Moens, M.-F. (2015). A survey on the application of recurrent neural networks to statistical language modeling. *Computer Speech & Language*, *30*(1), 61–98.

de Vries, H., Shuster, K., Batra, D., Parikh, D., Weston, J., & Kiela, D. (2018). Talk the walk: Navigating new york city through grounded dialogue. *CoRR*, *abs/1807.03367*.

de Vries, H., Strub, F., Chandar, S., Pietquin, O., Larochelle, H., & Courville, A. C. (2017). Guesswhat?! visual object discovery through multi-modal dialogue. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 4466–4475. IEEE Computer Society.

Delbrouck, J.-B., & Dupont, S. (2017a). An empirical study on the effectiveness of images in multimodal neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 910–919.

Delbrouck, J., & Dupont, S. (2017b). Multimodal compact bilinear pooling for multimodal neural machine translation. *CoRR*, *abs/1703.08084*.

Deng, C., Wu, Q., Wu, Q., Hu, F., Lyu, F., & Tan, M. (2018). Visual grounding via accumulated attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7746–7755.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee.

Deshpande, A., Aneja, J., Wang, L., Schwing, A. G., & Forsyth, D. A. (2018). Diverse and controllable image captioning with part-of-speech guidance. *CoRR*, *abs/1805.12589*.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., & Solorio, T. (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics.

Dodge, J., Gane, A., Zhang, X., Bordes, A., Chopra, S., Miller, A. H., Szlam, A., & Weston, J. (2016). Evaluating prerequisite qualities for learning end-to-end dialog systems. In Bengio, Y., & LeCun, Y. (Eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2625–2634.

El-Nouby, A., Sharma, S., Schulz, H., Hjelm, D., El Asri, L., Ebrahimi Kahou, S., Bengio, Y., & Taylor, G. W. (2018). Keep drawing it: Iterative language-based image generation and editing. In *Neural Information Processing Systems (NeurIPS) Visually-Grounded Interaction and Language (ViGIL) Workshop*.

Elliott, D. (2018). Adversarial evaluation of multimodal machine translation. In Riloff, E., Chiang, D., Hockenmaier, J., & Tsujii, J. (Eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pp. 2974–2978. Association for Computational Linguistics.

Elliott, D., Frank, S., Barrault, L., Bougares, F., & Specia, L. (2017). Findings of the second shared task on multimodal machine translation and multilingual image description. In Bojar, O., Buck, C., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno-Yepes, A., Koehn, P., & Kreutzer, J. (Eds.), *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017*, pp. 215–233. Association for Computational Linguistics.

Elliott, D., Frank, S., & Hasler, E. (2015). Multi-language image description with neural sequence models. *CoRR*, *abs/1510.04709*.

Elliott, D., Frank, S., Sima'an, K., & Specia, L. (2016). Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language, hosted by the 54th Annual Meeting of the Association for Computational Linguistics, VL@ACL 2016, August 12, Berlin, Germany*. The Association for Computer Linguistics.

Elliott, D., & Kádár, À. (2017). Imagination improves multimodal translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Vol. 1, pp. 130–141.

Elliott, D., & Keller, F. (2013). Image description using visual dependency representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1292–1302.

Fang, H., Gupta, S., Iandola, F., Srivastava, R. K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J. C., et al. (2015). From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1473–1482.

Farhadi, A., Hejrati, S. M. M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., & Forsyth, D. A. (2010). Every picture tells a story: Generating sentences from images. In *ECCV (4)*, Vol. 6314 of *Lecture Notes in Computer Science*, pp. 15–29. Springer.

Ferraro, F., Mostafazadeh, N., Huang, T. K., Vanderwende, L., Devlin, J., Galley, M., & Mitchell, M. (2015). A survey of current datasets for vision and language research. In Màrquez, L., Callison-Burch, C., Su, J., Pighin, D., & Marton, Y. (Eds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pp. 207–213. The Association for Computational Linguistics.

FitzGerald, N., Artzi, Y., & Zettlemoyer, L. (2013). Learning distributions over logical forms for referring expression generation. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1914–1925.

Fried, D., Hu, R., Cirik, V., Rohrbach, A., Andreas, J., Morency, L.-P., Berg-Kirkpatrick, T., Saenko, K., Klein, D., & Darrell, T. (2018). Speaker-follower models for vision-and-language navigation. In *Advances in Neural Information Processing Systems*, pp. 3318–3329.

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Mikolov, T., et al. (2013). Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pp. 2121–2129.

Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., & Rohrbach, M. (2016). Multimodal compact bilinear pooling for visual question answering and visual grounding. In Su, J., Carreras, X., & Duh, K. (Eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pp. 457–468. The Association for Computational Linguistics.

Gan, C., Gan, Z., He, X., Gao, J., & Deng, L. (2017). Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3137–3146.

Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L., & Xu, W. (2015). Are you talking to a machine? dataset and methods for multilingual image question. In *Advances in neural information processing systems*, pp. 2296–2304.

Gao, L., Chen, D., Song, J., Xu, X., Zhang, D., & Shen, H. T. (2019). Perceptual pyramid adversarial networks for text-to-image synthesis. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pp. 8312–8319. AAAI Press.

Gao, L., Guo, Z., Zhang, H., Xu, X., & Shen, H. T. (2017). Video captioning with attention-based lstm and semantic consistency. *IEEE Transactions on Multimedia*, *19*(9), 2045–2055.

Garbacea, C., & Mei, Q. (2020). Neural language generation: Formulation, methods, and evaluation. *CoRR*, *abs/2007.15780*.

Gatt, A., & Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, *61*, 65–170.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2414–2423.

Gella, S., & Keller, F. (2017). An analysis of action recognition datasets for language and vision tasks. In Barzilay, R., & Kan, M. (Eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pp. 64–71. Association for Computational Linguistics.

Gella, S., Lewis, M., & Rohrbach, M. (2018). A dataset for telling the stories of social media videos. In Riloff, E., Chiang, D., Hockenmaier, J., & Tsujii, J. (Eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pp. 968–974. Association for Computational Linguistics.

Geman, D., Geman, S., Hallonquist, N., & Younes, L. (2015). Visual turing test for computer vision systems. *Proceedings of the National Academy of Sciences*, *112*(12), 3618–3623.

Golland, D., Liang, P., & Klein, D. (2010). A game-theoretic approach to generating spatial descriptions. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pp. 410–419. Association for Computational Linguistics.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680.

Gordon, D., Kembhavi, A., Rastegari, M., Redmon, J., Fox, D., & Farhadi, A. (2018). Iqa: Visual question answering in interactive environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4089–4098.

Goyal, Y., Khot, T., Agrawal, A., Summers-Stay, D., Batra, D., & Parikh, D. (2019). Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *International Journal of Computer Vision*, *127*(4), 398–414.

Graham, Y., Awad, G., & Smeaton, A. (2018). Evaluation of automatic video captioning using direct assessment. *PloS one*, *13*(9), e0202789.

Grönroos, S., Huet, B., Kurimo, M., Laaksonen, J., Mérialdo, B., Pham, P., Sjöberg, M., Sulubacak, U., Tiedemann, J., Troncy, R., & Vázquez, R. (2018). The memad submission to the WMT18 multimodal translation task. In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Jimeno-Yepes, A., Koehn, P., Monz, C., Negri, M., Névéol, A., Neves, M. L., Post, M., Specia, L., Turchi, M., & Verspoor, K. (Eds.), *Proceedings of the Third Conference on Machine Translation: Shared Task Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pp. 603–611. Association for Computational Linguistics.

Gu, J., Cai, J., Wang, G., & Chen, T. (2018). Stack-captioning: Coarse-to-fine learning for image captioning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Gu, J., Wang, G., Cai, J., & Chen, T. (2017). An empirical study of language cnn for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1222–1231.

Guo, D., Xu, C., & Tao, D. (2019). Image-question-answer synergistic network for visual dialog. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 10434–10443. Computer Vision Foundation / IEEE.

Guo, G., Zhai, S., Yuan, F., Liu, Y., & Wang, X. (2018). Vse-ens: Visual-semantic embeddings with efficient negative sampling. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Guo, L., Liu, J., Lu, S., & Lu, H. (2020). Show, tell, and polish: Ruminant decoding for image captioning. *IEEE Trans. Multim.*, *22*(8), 2149–2162.

Guo, L., Liu, J., Yao, P., Li, J., & Lu, H. (2019). Mscap: Multi-style image captioning with unpaired stylized text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4204–4213.

Gupta, T., Schwenk, D., Farhadi, A., Hoiem, D., & Kembhavi, A. (2018). Imagine this! scripts to compositions to videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 598–613.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., & Smith, N. A. (2018). Annotation artifacts in natural language inference data. In Walker, M. A., Ji, H., & Stent, A. (Eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pp. 107–112. Association for Computational Linguistics.

Harabagiu, S. M., Pasca, M. A., & Maiorano, S. J. (2000). Experiments with open-domain textual question answering. In *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*.

Haurilet, M., Roitberg, A., & Stiefelhagen, R. (2019). It is not about the journey; it is about the destination: Following soft paths under question-guidance for visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. B. (2017). Mask R-CNN. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 2980–2988. IEEE Computer Society.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Helcl, J., Libovický, J., & Varis, D. (2018). CUNI system for the WMT18 multimodal translation task. In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Jimeno-Yepes, A., Koehn, P., Monz, C., Negri, M., Névéol, A., Neves, M. L., Post, M., Specia, L., Turchi, M., & Verspoor, K. (Eds.), *Proceedings of the Third Conference on Machine Translation: Shared Task Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pp. 616–623. Association for Computational Linguistics.

Hendricks, L. A., Venugopalan, S., Rohrbach, M., Mooney, R., Saenko, K., & Darrell, T. (2016). Deep Compositional Captioning: Describing novel object categories without paired training data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–10.

Hendricks, L. A., Wang, O., Shechtman, E., Sivic, J., Darrell, T., & Russell, B. C. (2017). Localizing moments in video with natural language. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 5804–5813. IEEE Computer Society.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local nash equilibrium. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., & Garnett, R. (Eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pp. 6626–6637.

Hinz, T., Heinrich, S., & Wermter, S. (2019). Generating multiple objects at spatially distinct locations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Hitschler, J., Schamoni, S., & Riezler, S. (2016). Multimodal pivots for image caption translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural computation*, *9*(8), 1735–1780.

Hodosh, M., Young, P., & Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, *47*, 853–899.

Hong, R., Liu, D., Mo, X., He, X., & Zhang, H. (2019). Learning to compose and reason with language tree structures for visual grounding. *IEEE transactions on pattern analysis and machine intelligence*, *abs/1906.01784*.

Hong, S., Yang, D., Choi, J., & Lee, H. (2018). Inferring semantic layout for hierarchical text-to-image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7986–7994.

Hori, C., Alamri, H., Wang, J., Wichern, G., Hori, T., Cherian, A., Marks, T. K., Cartillier, V., Lopes, R. G., Das, A., et al. (2019). End-to-end audio visual scene-aware dialog using multimodal attention-based video features. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2352–2356. IEEE.

Hori, C., Hori, T., Lee, T.-Y., Zhang, Z., Harsham, B., Hershey, J. R., Marks, T. K., & Sumi, K. (2017). Attention-based multimodal fusion for video description. In *Proceedings of the IEEE international conference on computer vision*, pp. 4193–4202.

Hossain, M., Sohel, F., Shiratuddin, M. F., & Laga, H. (2019). A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CSUR)*, *51*(6), 118.

Hsu, C., Chen, S., Hsieh, M., & Ku, L. (2018). Using inter-sentence diverse beam search to reduce redundancy in visual storytelling. *CoRR*, *abs/1805.11867*.

Hu, R., Andreas, J., Darrell, T., & Saenko, K. (2018). Explainable neural computation via stack neural module networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 53–69.

Hu, R., Andreas, J., Rohrbach, M., Darrell, T., & Saenko, K. (2017). Learning to reason: End-to-end module networks for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 804–813.

Hu, R., Rohrbach, A., Darrell, T., & Saenko, K. (2019). Language-conditioned graph networks for relational reasoning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 10293–10302. IEEE.

Hu, R., Rohrbach, M., Andreas, J., Darrell, T., & Saenko, K. (2017a). Modeling relationships in referential expressions with compositional modular networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1115–1124.

Hu, R., Rohrbach, M., Andreas, J., Darrell, T., & Saenko, K. (2017b). Modeling relationships in referential expressions with compositional modular networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1115–1124.

Hu, R., Xu, H., Rohrbach, M., Feng, J., Saenko, K., & Darrell, T. (2016). Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4555–4564.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708.

Huang, L., Wang, W., Chen, J., & Wei, X. (2019). Attention on attention for image captioning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 4633–4642. IEEE.

Huang, P.-Y., Liu, F., Shiang, S.-R., Oh, J., & Dyer, C. (2016). Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, Vol. 2, pp. 639–645.

Huang, Q., Gan, Z., Çelikyilmaz, A., Wu, D. O., Wang, J., & He, X. (2019). Hierarchically structured reinforcement learning for topically coherent visual story generation. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pp. 8465–8472. AAAI Press.

Huang, T. K., Ferraro, F., Mostafazadeh, N., Misra, I., Agrawal, A., Devlin, J., Girshick, R. B., He, X., Kohli, P., Batra, D., Zitnick, C. L., Parikh, D., Vanderwende, L., Galley, M., & Mitchell, M. (2016). Visual storytelling. In Knight, K., Nenkova, A., & Rambow, O. (Eds.), *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pp. 1233–1239. The Association for Computational Linguistics.

Hudson, D. A., & Manning, C. D. (2018). Compositional attention networks for machine reasoning. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Hudson, D. A., & Manning, C. D. (2019). GQA: A new dataset for compositional question answering over real-world images. *CoRR*, *abs/1902.09506*.

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134.

IV, R. L. L., Liu, N. F., Peters, M. E., Gardner, M., & Singh, S. (2019). Barack's wife hillary: Using knowledge graphs for fact-aware language modeling. In Korhonen, A., Traum, D. R., & Màrquez, L. (Eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 5962–5971. Association for Computational Linguistics.

Jabri, A., Joulin, A., & Van Der Maaten, L. (2016). Revisiting visual question answering baselines. In *European conference on computer vision*, pp. 727–739. Springer.

Jain, U., Lazebnik, S., & Schwing, A. G. (2018). Two can play this game: Visual dialog with discriminative question generation and answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5754–5763. IEEE Computer Society.

Jang, Y., Song, Y., Yu, Y., Kim, Y., & Kim, G. (2017). TGIF-QA: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2758–2766.

Järvelin, K., & Kekäläinen, J. (2000). IR evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 41–48. ACM.

Jia, X., Gavves, E., Fernando, B., & Tuytelaars, T. (2015). Guiding the long-short term memory model for image caption generation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2407–2415.

Jin, J., Fu, K., Cui, R., Sha, F., & Zhang, C. (2015). Aligning where to see and what to tell: image caption with region-based attention and scene factorization. *CoRR*, *abs/1506.06272*.

Jin, Q., Chen, J., Chen, S., Xiong, Y., & Hauptmann, A. (2016). Describing videos using multi-modal fusion. In *Proceedings of the 24th ACM international conference on Multimedia*, pp. 1087–1091. ACM.

Jing, C., Wu, Y., Zhang, X., Jia, Y., & Wu, Q. (2020). Overcoming language priors in vqa via decomposed linguistic representations. In *Proc. Conf. AAAI*.

Jing, L., & Tian, Y. (2019). Self-supervised visual feature learning with deep neural networks: A survey. *CoRR*, *abs/1902.06162*.

Johnson, J., Gupta, A., & Fei-Fei, L. (2018). Image generation from scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1219–1228.

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., & Girshick, R. (2017a). CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2901–2910.

Johnson, J., Hariharan, B., van der Maaten, L., Hoffman, J., Fei-Fei, L., Lawrence Zitnick, C., & Girshick, R. (2017b). Inferring and executing programs for visual reasoning. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2989–2998.

Johnson, J., Karpathy, A., & Fei-Fei, L. (2016). Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4565–4574.

Kafle, K., & Kanan, C. (2017). Visual question answering: Datasets, algorithms, and future challenges. *Computer Vision and Image Understanding*, *163*, 3–20.

Kafle, K., Shrestha, R., & Kanan, C. (2019). Challenges and prospects in vision and language research. *Frontiers Artif. Intell.*, *2*, 28.

Kalimuthu, M., Mogadala, A., Mosbach, M., & Klakow, D. (2020). Fusion models for improved image captioning. In Bimbo, A. D., Cucchiara, R., Sclaroff, S., Farinella, G. M., Mei, T., Bertini, M., Escalante, H. J., & Vezzani, R. (Eds.), *Pattern Recognition. ICPR International Workshops and Challenges - Virtual Event, January 10-15, 2021, Proceedings, Part VI*, Vol. 12666 of *Lecture Notes in Computer Science*, pp. 381–395. Springer.

Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3128–3137.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., & Zisserman, A. (2017). The kinetics human action video dataset. *CoRR, abs/1705.06950*.

Kazemzadeh, S., Ordonez, V., Matten, M., & Berg, T. (2014). Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 787–798.

Ke, L., Pei, W., Li, R., Shen, X., & Tai, Y. (2019a). Reflective decoding network for image captioning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 8887–8896. IEEE.

Ke, L., Li, X., Bisk, Y., Holtzman, A., Gan, Z., Liu, J., Gao, J., Choi, Y., & Srinivasa, S. S. (2019b). Tactical rewind: Self-correction via backtracking in vision-and-language navigation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 6741–6749. Computer Vision Foundation / IEEE.

Khoreva, A., Rohrbach, A., & Schiele, B. (2018). Video object segmentation with language referring expressions. In Jawahar, C. V., Li, H., Mori, G., & Schindler, K. (Eds.), *Computer Vision - ACCV 2018 - 14th Asian Conference on Computer Vision, Perth, Australia, December 2-6, 2018, Revised Selected Papers, Part IV*, Vol. 11364 of *Lecture Notes in Computer Science*, pp. 123–141. Springer.

Kim, D., Choi, J., Oh, T., & Kweon, I. S. (2019). Dense relational captioning: Triple-stream networks for relationship-based captioning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 6271–6280. Computer Vision Foundation / IEEE.

Kim, D., Saito, K., Saenko, K., Sclaroff, S., & Plummer, B. A. (2020). MULE: multimodal universal language embedding. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 11254–11261. AAAI Press.

Kim, G., Moon, S., & Sigal, L. (2015). Ranking and retrieval of image sequences from multiple paragraph queries. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 1993–2001. IEEE Computer Society.

Kim, J., Rohrbach, A., Darrell, T., Canny, J., & Akata, Z. (2018). Textual explanations for self-driving vehicles. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 563–578.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In Bengio, Y., & LeCun, Y. (Eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. In Bengio, Y., & LeCun, Y. (Eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Kiros, R., Salakhutdinov, R., & Zemel, R. (2014a). Multimodal neural language models. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 595–603.

Kiros, R., Salakhutdinov, R., & Zemel, R. S. (2014b). Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, *abs/1411.2539*.

Kottur, S., Moura, J. M. F., Parikh, D., Batra, D., & Rohrbach, M. (2018). Visual coreference resolution in visual dialog using neural module networks. In *ECCV (15)*, Vol. 11219 of *Lecture Notes in Computer Science*, pp. 160–178. Springer.

Kottur, S., Moura, J. M., Parikh, D., Batra, D., & Rohrbach, M. (2019). CLEVR-dialog: A diagnostic dataset for multi-round reasoning in visual dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 582–595.

Krahmer, E., & Van Deemter, K. (2012). Computational generation of referring expressions: A survey. *Computational Linguistics*, *38*(1), 173–218.

Krause, J., Johnson, J., Krishna, R., & Fei-Fei, L. (2017). A hierarchical approach for generating descriptive image paragraphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 317–325.

Krishna, R., Hata, K., Ren, F., Fei-Fei, L., & Niebles, J. C. (2017a). Dense-captioning events in videos. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 706–715. IEEE Computer Society.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L., Shamma, D. A., Bernstein, M. S., & Fei-Fei, L. (2017b). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, *123*(1), 32–73.

Krishnamoorthy, N., Malkarnenkar, G., Mooney, R., Saenko, K., & Guadarrama, S. (2013). Generating natural-language video descriptions using text-mined knowledge. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105.

Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., Berg, A. C., & Berg, T. L. (2013). Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(12), 2891–2903.

Lample, G., Ott, M., Conneau, A., Denoyer, L., & Ranzato, M. (2018). Phrase-based & neural unsupervised machine translation. In Riloff, E., Chiang, D., Hockenmaier, J., & Tsujii, J. (Eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pp. 5039–5049. Association for Computational Linguistics.

LeCun, Y., Bengio, Y., et al. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, *3361*(10), 1995.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436.

Lei, J., Yu, L., Bansal, M., & Berg, T. L. (2018). TVQA: localized, compositional video question answering. In Riloff, E., Chiang, D., Hockenmaier, J., & Tsujii, J. (Eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pp. 1369–1379. Association for Computational Linguistics.

Lei, J., Yu, L., Berg, T. L., & Bansal, M. (2020). TVQA+: spatio-temporal grounding for video question answering. In Jurafsky, D., Chai, J., Schluter, N., & Tetreault, J. R. (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 8211–8225. Association for Computational Linguistics.

Li, B., Qi, X., Lukasiewicz, T., & Torr, P. H. (2020). ManiGAN: Text-guided image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, C. Y., Liang, X., Hu, Z., & Xing, E. P. (2019). Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pp. 6666–6673. AAAI Press.

Li, G., Duan, N., Fang, Y., Gong, M., & Jiang, D. (2020). Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 11336–11344. AAAI Press.

Li, J., Monroe, W., Ritter, A., Jurafsky, D., Galley, M., & Gao, J. (2016). Deep reinforcement learning for dialogue generation. In Su, J., Carreras, X., & Duh, K. (Eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pp. 1192–1202. The Association for Computational Linguistics.

Li, J., Wong, Y., Zhao, Q., & Kankanhalli, M. S. (2020). Video storytelling: Textual summaries for events. *IEEE Transactions on Multimedia*, *22*(2), 554–565.

Li, S., Kulkarni, G., Berg, T. L., Berg, A. C., & Choi, Y. (2011). Composing simple image descriptions using web-scale n-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pp. 220–228. Association for Computational Linguistics.

Li, X., Zhou, Z., Chen, L., & Gao, L. (2019). Residual attention-based LSTM for video captioning. *World Wide Web*, *22*(2), 621–636.

Li, X., Yin, X., Li, C., Hu, X., Zhang, P., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., Choi, Y., & Gao, J. (2020). Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*.

Li, Y., Yao, T., Pan, Y., Chao, H., & Mei, T. (2019a). Pointing novel objects in image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12497–12506.

Li, Y., Gan, Z., Shen, Y., Liu, J., Cheng, Y., Wu, Y., Carin, L., Carlson, D. E., & Gao, J. (2019b). Storygan: A sequential conditional GAN for story visualization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 6329–6338. Computer Vision Foundation / IEEE.

Li, Y., Min, M. R., Shen, D., Carlson, D. E., & Carin, L. (2018). Video generation from text. In McIlraith, S. A., & Weinberger, K. Q. (Eds.), *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 7065–7072. AAAI Press.

Liang, X., Hu, Z., Zhang, H., Gan, C., & Xing, E. P. (2017). Recurrent topic-transition gan for visual paragraph generation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3362–3371.

Libovickỳ, J., & Helcl, J. (2017). Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. 2, pp. 196–202.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer.

Liu, C., Lin, Z., Shen, X., Yang, J., Lu, X., & Yuille, A. (2017). Recurrent multimodal interaction for referring image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1271–1280.

Liu, D., Bober, M., & Kittler, J. (2019). Visual semantic information pursuit: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *abs/1903.05434*.

Liu, F., Flanigan, J., Thomson, S., Sadeh, N. M., & Smith, N. A. (2018). Toward abstractive summarization using semantic representations. *CoRR*, *abs/1805.10399*.

Liu, F., Wu, X., Ge, S., Fan, W., & Zou, Y. (2020). Federated learning for vision-and-language grounding problems. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 11572–11579. AAAI Press.

Liu, H., Lin, A., Han, X., Yang, L., Yu, Y., & Cui, S. (2021). Refer-it-in-rgbd: A bottom-up approach for 3d visual grounding in rgbd images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6032–6041.

Liu, J., Wang, L., & Yang, M. (2017). Referring expression generation and comprehension via attributes. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 4866–4874. IEEE Computer Society.

Liu, J., Chen, W., Cheng, Y., Gan, Z., Yu, L., Yang, Y., & Liu, J. (2020). VIOLIN: A large-scale dataset for video-and-language inference. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 10897–10907. IEEE.

Liu, R., Liu, C., Bai, Y., & Yuille, A. L. (2019). Clevr-ref+: Diagnosing visual reasoning with referring expressions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 4185–4194. Computer Vision Foundation / IEEE.

Liu, S., Zhu, Z., Ye, N., Guadarrama, S., & Murphy, K. (2017). Improved image captioning via policy gradient optimization of spider. In *Proceedings of the IEEE international conference on computer vision*, pp. 873–881.

Liu, W., Mei, T., Zhang, Y., Che, C., & Luo, J. (2015). Multi-task deep visual-semantic embedding for video thumbnail selection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3707–3715.

Liu, Y., Fu, J., Mei, T., & Chen, C. W. (2017). Let your photos talk: Generating narrative paragraph for photo stream via bidirectional attention recurrent neural networks. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440.

Long, X., Gan, C., & de Melo, G. (2018). Video captioning with multi-faceted attention. *Transactions of the Association of Computational Linguistics*, *6*, 173–184.

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., & Garnett, R. (Eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pp. 13–23.

Lu, J., Kannan, A., Yang, J., Parikh, D., & Batra, D. (2017a). Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model. In *NIPS*, pp. 313–323.

Lu, J., Xiong, C., Parikh, D., & Socher, R. (2017b). Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 3242–3250. IEEE Computer Society.

Lu, J., Yang, J., Batra, D., & Parikh, D. (2018). Neural baby talk. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7219–7228.

Luo, R., & Shakhnarovich, G. (2017). Comprehension-guided referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7102–7111.

Ma, C., Lu, J., Wu, Z., AlRegib, G., Kira, Z., Socher, R., & Xiong, C. (2019a). Self-monitoring navigation agent via auxiliary progress estimation. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Ma, C., Wu, Z., AlRegib, G., Xiong, C., & Kira, Z. (2019b). The regretful agent: Heuristic-aided navigation through progress estimation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 6732–6740. Computer Vision Foundation / IEEE.

Ma, Y.-F., Lu, L., Zhang, H.-J., & Li, M. (2002). A user attention model for video summarization. In *Proceedings of the tenth ACM international conference on Multimedia*, pp. 533–542. ACM.

MacMahon, M., Stankiewicz, B., & Kuipers, B. (2006). Walk the talk: Connecting language, knowledge, and action in route instructions. *Def*, *2*(6), 4.

Malinowski, M., & Fritz, M. (2014). A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in neural information processing systems*, pp. 1682–1690.

Malinowski, M., Rohrbach, M., & Fritz, M. (2015). Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE international conference on computer vision*, pp. 1–9.

Manjunatha, V., Saini, N., & Davis, L. S. (2019). Explicit bias discovery in visual question answering models. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 9562–9571. Computer Vision Foundation / IEEE.

Manning, C., Raghavan, P., & Schütze, H. (2010). Introduction to information retrieval. *Natural Language Engineering*, *16*(1), 100–103.

Mansimov, E., Parisotto, E., Ba, L. J., & Salakhutdinov, R. (2016). Generating images from captions with attention. In Bengio, Y., & LeCun, Y. (Eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Mao, J., Gan, C., Kohli, P., Tenenbaum, J. B., & Wu, J. (2019). The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A. L., & Murphy, K. (2016). Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 11–20.

Mao, J., Xu, W., Yang, Y., Wang, J., & Yuille, A. L. (2015). Deep captioning with multimodal recurrent neural networks (m-rnn). In Bengio, Y., & LeCun, Y. (Eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Marino, K., Rastegari, M., Farhadi, A., & Mottaghi, R. (2019). OK-VQA: A visual question answering benchmark requiring external knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 3195–3204. Computer Vision Foundation / IEEE.

Mascharka, D., Tran, P., Soklaski, R., & Majumdar, A. (2018). Transparency by design: Closing the gap between performance and interpretability in visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4942–4950.

Mathews, A., Xie, L., & He, X. (2018). Semstyle: Learning to generate stylised image captions using unaligned text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8591–8600.

Mathews, A. P., Xie, L., & He, X. (2016). Senticap: Generating image descriptions with sentiments. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Mazaheri, A., & Shah, M. (2018). Visual text correction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 155–171.

Mazaheri, A., Zhang, D., & Shah, M. (2017). Video fill in the blank using lstms with spatial-temporal attentions. In *ICCV*, pp. 1416–1425. IEEE Computer Society.

Messina, N., Amato, G., Carrara, F., Falchi, F., & Gennaro, C. (2018). Learning relationship-aware visual features. In *European Conference on Computer Vision*, pp. 486–501. Springer.

Miech, A., Zhukov, D., Alayrac, J., Tapaswi, M., Laptev, I., & Sivic, J. (2019). Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 2630–2640. IEEE.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119.

Misra, D., Bennett, A., Blukis, V., Niklasson, E., Shatkhin, M., & Artzi, Y. (2018a). Mapping instructions to actions in 3d environments with visual goal prediction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2667–2678.

Misra, I., Girshick, R., Fergus, R., Hebert, M., Gupta, A., & van der Maaten, L. (2018b). Learning by asking questions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11–20.

Mitchell, M., Han, X., Dodge, J., Mensch, A., Goyal, A., Berg, A., Yamaguchi, K., Berg, T., Stratos, K., & Daumé III, H. (2012). Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European*

Chapter of the Association for Computational Linguistics, pp. 747–756. Association for Computational Linguistics.

Mitchell, M., Van Deemter, K., & Reiter, E. (2013). Generating expressions that refer to visible objects. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics (ACL).

Miyazaki, T., & Shimizu, N. (2016). Cross-lingual image caption generation.. In *ACL (1)*.

Moens, M., Specia, L., & Tuytelaars, T. (2019). Joint processing of language and visual data for better automated understanding (dagstuhl seminar 19021). *Dagstuhl Reports*, *9*(1), 1–27.

Mogadala, A. (2015). Polylingual multimodal learning. In *ECML PKDD Doctoral Consortium*, p. 155. Citeseer.

Mogadala, A., Bista, U., Xie, L., & Rettinger, A. (2018a). Knowledge guided attention and inference for describing images containing unseen objects. In *European Semantic Web Conference*, pp. 415–429. Springer.

Mogadala, A., Kanuparthi, B., Rettinger, A., & Sure-Vetter, Y. (2018b). Discovering connotations as labels for weakly supervised image-sentence data. In *Companion Proceedings of the The Web Conference 2018*, pp. 379–386.

Mostafazadeh, N., Brockett, C., Dolan, B., Galley, M., Gao, J., Spithourakis, G. P., & Vanderwende, L. (2017). Image-grounded conversations: Multimodal context for natural question and response generation. In *IJCNLP(1)*, pp. 462–472. Asian Federation of Natural Language Processing.

Motwani, T. S., & Mooney, R. J. (2012). Improving video activity recognition using object recognition and text mining.. In *ECAI*, Vol. 1, p. 2.

Nagaraja, V. K., Morariu, V. I., & Davis, L. S. (2016). Modeling context between objects for referring expression understanding. In *European Conference on Computer Vision*, pp. 792–807. Springer.

Nallapati, R., Zhou, B., dos Santos, C. N., Gülçehre, Ç., & Xiang, B. (2016). Abstractive text summarization using sequence-to-sequence rnns and beyond. In Goldberg, Y., & Riezler, S. (Eds.), *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, pp. 280–290. ACL.

Nam, S., Kim, Y., & Kim, S. J. (2018). Text-adaptive generative adversarial networks: Manipulating images with natural language. In *Advances in Neural Information Processing Systems*, pp. 42–51.

Nguyen, A., Clune, J., Bengio, Y., Dosovitskiy, A., & Yosinski, J. (2017). Plug & play generative networks: Conditional iterative generation of images in latent space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4467–4477.

Nguyen, D.-K., & Okatani, T. (2019). Multi-task learning of hierarchical vision-language representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10492–10501.

Nguyen, K., Dey, D., Brockett, C., & Dolan, B. (2019). Vision-based navigation with language-based assistance via imitation learning with indirect intervention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12527–12537.

Niu, Y., Zhang, H., Zhang, M., Zhang, J., Lu, Z., & Wen, J.-R. (2019). Recursive visual attention in visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6679–6688.

Ordonez, V., Kulkarni, G., & Berg, T. L. (2011). Im2text: Describing images using 1 million captioned photographs. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F. C. N., & Weinberger, K. Q. (Eds.), *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pp. 1143–1151.

Pan, Y., Yao, T., Li, H., & Mei, T. (2017). Video captioning with transferred semantic attributes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6504–6512.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311–318. Association for Computational Linguistics.

Park, C. C., & Kim, G. (2015). Expressing an image stream with a sequence of natural sentences. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., & Garnett, R. (Eds.), *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 73–81.

Park, J. S., Bhagavatula, C., Mottaghi, R., Farhadi, A., & Choi, Y. (2020). VisualCOMET: Reasoning about the dynamic context of a still image. In *In Proceedings of the European Conference on Computer Vision (ECCV)*.

Pasunuru, R., & Bansal, M. (2017a). Multi-task video captioning with video and entailment generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1, pp. 1273–1283.

Pasunuru, R., & Bansal, M. (2017b). Reinforced video captioning with entailment rewards. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 979–985.

Pedersoli, M., Lucas, T., Schmid, C., & Verbeek, J. (2017). Areas of attention for image captioning. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 1251–1259. IEEE Computer Society.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.

Perez, E., Strub, F., De Vries, H., Dumoulin, V., & Courville, A. (2018). Film: Visual reasoning with a general conditioning layer. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Plummer, B. A., Mallya, A., Cervantes, C. M., Hockenmaier, J., & Lazebnik, S. (2017a). Phrase localization and visual relationship detection with comprehensive image-language cues. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1928–1937.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., & Lazebnik, S. (2017b). Flickr30k Entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *International Journal of Computer Vision*, *123*(1), 74–93.

Qiao, T., Zhang, J., Xu, D., & Tao, D. (2019). Mirrorgan: Learning text-to-image generation by redescription. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 1505–1514. Computer Vision Foundation / IEEE.

Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. In Bengio, Y., & LeCun, Y. (Eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8).

Rajani, N. F., McCann, B., Xiong, C., & Socher, R. (2019). Explain yourself! leveraging language models for commonsense reasoning. In Korhonen, A., Traum, D. R., & Màrquez, L. (Eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 4932–4942. Association for Computational Linguistics.

Ramanishka, V., Das, A., Park, D. H., Venugopalan, S., Hendricks, L. A., Rohrbach, M., & Saenko, K. (2016). Multimodal video description. In *Proceedings of the 24th ACM international conference on Multimedia*, pp. 1092–1096. ACM.

Redmon, J., & Farhadi, A. (2017). YOLO9000: better, faster, stronger. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 6517–6525. IEEE Computer Society.

Reed, S. E., Akata, Z., Mohan, S., Tenka, S., Schiele, B., & Lee, H. (2016a). Learning what and where to draw. In *Advances in Neural Information Processing Systems*, pp. 217–225.

Reed, S. E., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016b). Generative adversarial text to image synthesis. In Balcan, M., & Weinberger, K. Q. (Eds.), *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, Vol. 48 of *JMLR Workshop and Conference Proceedings*, pp. 1060–1069. JMLR.org.

Regneri, M., Rohrbach, M., Wetzel, D., Thater, S., Schiele, B., & Pinkal, M. (2013). Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, *1*, 25–36.

Reiter, E., & Dale, R. (2000). *Building natural language generation systems*. Cambridge university press.

Ren, M., Kiros, R., & Zemel, R. (2015a). Exploring models and data for image question answering. In *Advances in neural information processing systems*, pp. 2953–2961.

Ren, S., He, K., Girshick, R., & Sun, J. (2015b). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pp. 91–99.

Ren, S., He, K., Girshick, R., & Sun, J. (2015c). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pp. 91–99.

Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., & Goel, V. (2017). Self-critical sequence training for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 1179–1195. IEEE Computer Society.

Rickert, M., Foster, M. E., Giuliani, M., By, T., Panin, G., & Knoll, A. (2007). Integrating language, vision and action for human robot dialog systems. In *International Conference on Universal Access in Human-Computer Interaction*, pp. 987–995. Springer.

Roh, J., Desingh, K., Farhadi, A., & Fox, D. (2021). Languagerefer: Spatial-language model for 3d visual grounding. *CoRR*, *abs/2107.03438*.

Rohrbach, A., Rohrbach, M., Hu, R., Darrell, T., & Schiele, B. (2016a). Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, pp. 817–834. Springer.

Rohrbach, A., Rohrbach, M., Hu, R., Darrell, T., & Schiele, B. (2016b). Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, pp. 817–834. Springer.

Rohrbach, A., Rohrbach, M., Qiu, W., Friedrich, A., Pinkal, M., & Schiele, B. (2014). Coherent multi-sentence video description with variable level of detail. In Jiang, X., Hornegger, J., & Koch, R. (Eds.), *Pattern Recognition - 36th German Conference, GCPR 2014, Münster, Germany, September 2-5, 2014, Proceedings*, Vol. 8753 of *Lecture Notes in Computer Science*, pp. 184–195. Springer.

Rohrbach, A., Rohrbach, M., Tandon, N., & Schiele, B. (2015). A dataset for movie description. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 3202–3212. IEEE Computer Society.

Rohrbach, M., Amin, S., Andriluka, M., & Schiele, B. (2012). A database for fine grained activity detection of cooking activities. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1194–1201. IEEE.

Rohrbach, M., Qiu, W., Titov, I., Thater, S., Pinkal, M., & Schiele, B. (2013). Translating video content to natural language descriptions. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 433–440.

Rohrbach, M., Regneri, M., Andriluka, M., Amin, S., Pinkal, M., & Schiele, B. (2012). Script data for attribute-based recognition of composite activities. In Fitzgibbon, A. W., Lazebnik, S., Perona, P., Sato, Y., & Schmid, C. (Eds.), *Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October*

7-13, 2012, Proceedings, Part I, Vol. 7572 of *Lecture Notes in Computer Science*, pp. 144–157. Springer.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training GANs. In *Advances in neural information processing systems*, pp. 2234–2242.

Sammani, F., & Melas-Kyriazi, L. (2020). Show, edit and tell: A framework for editing image captions. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 4807–4815. IEEE.

Santoro, A., Hill, F., Barrett, D., Morcos, A., & Lillicrap, T. (2018). Measuring abstract reasoning in neural networks. In *International Conference on Machine Learning*, pp. 4477–4486.

Santoro, A., Raposo, D., Barrett, D. G., Malinowski, M., Pascanu, R., Battaglia, P., & Lillicrap, T. (2017). A simple neural network module for relational reasoning. In *Advances in neural information processing systems*, pp. 4967–4976.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2008). The graph neural network model. *IEEE Transactions on Neural Networks*, *20*(1), 61–80.

Schamoni, S., Hitschler, J., & Riezler, S. (2018). A dataset and reranking method for multimodal MT of user-generated image captions. In Cherry, C., & Neubig, G. (Eds.), *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas, AMTA 2018, Boston, MA, USA, March 17-21, 2018 - Volume 1: Research Papers*, pp. 140–153. Association for Machine Translation in the Americas.

Schwartz, I., Schwing, A. G., & Hazan, T. (2019). A simple baseline for audio-visual scene-aware dialog. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 12548–12558. Computer Vision Foundation / IEEE.

Sellam, T., Das, D., & Parikh, A. P. (2020). BLEURT: learning robust metrics for text generation. In Jurafsky, D., Chai, J., Schluter, N., & Tetreault, J. R. (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 7881–7892. Association for Computational Linguistics.

Seo, P. H., Lehrmann, A. M., Han, B., & Sigal, L. (2017). Visual reference resolution using attention memory for visual dialog. In *NIPS*, pp. 3722–3732.

Shah, M., Chen, X., Rohrbach, M., & Parikh, D. (2019a). Cycle-consistency for robust visual question answering. *CoRR*, *abs/1902.05660*.

Shah, S., Mishra, A., Yadati, N., & Talukdar, P. P. (2019b). KVQA: knowledge-aware visual question answering. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pp. 8876–8884. AAAI Press.

Sharma, P., Ding, N., Goodman, S., & Soricut, R. (2018). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Gurevych,

I., & Miyao, Y. (Eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pp. 2556–2565. Association for Computational Linguistics.

Shekhar, R., Venkatesh, A., Baumgärtner, T., Bruni, E., Plank, B., Bernardi, R., & Fernández, R. (2019). Beyond task success: A closer look at jointly learning to see, ask, and guesswhat. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2578–2587.

Shetty, R., Rohrbach, M., Hendricks, L. A., Fritz, M., & Schiele, B. (2017). Speaking the same language: Matching machine to human captions by adversarial training. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 4155–4164. IEEE Computer Society.

Shetty, R., Tavakoli, H. R., & Laaksonen, J. (2018). Image and video captioning with augmented neural architectures. *IEEE MultiMedia*, *25*(2), 34–46.

Shi, J., Zhang, H., & Li, J. (2019). Explainable and explicit visual reasoning over scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Shridhar, M., Thomason, J., Gordon, D., Bisk, Y., Han, W., Mottaghi, R., Zettlemoyer, L., & Fox, D. (2020). Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Computer Vision and Pattern Recognition (CVPR)*.

Shuster, K., Humeau, S., Hu, H., Bordes, A., & Weston, J. (2019). Engaging image captioning via personality. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 12516–12526. Computer Vision Foundation / IEEE.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In Bengio, Y., & LeCun, Y. (Eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., & Rohrbach, M. (2019). Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society.

Sinopoli, B., Micheli, M., Donato, G., & Koo, T.-J. (2001). Vision based navigation for an unmanned aerial vehicle. In *Proceedings 2001 ICRA. IEEE International Conference on Robotics and Automation (Cat. No. 01CH37164)*, Vol. 2, pp. 1757–1764. IEEE.

Song, J., Gao, L., Guo, Z., Liu, W., Zhang, D., & Shen, H. T. (2017). Hierarchical LSTM with adjusted temporal attention for video captioning. In Sierra, C. (Ed.), *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pp. 2737–2743. ijcai.org.

Specia, L., Frank, S., Sima'an, K., & Elliott, D. (2016). A shared task on multimodal machine translation and crosslingual image description.. In *WMT*, pp. 543–553.

Srivastava, N., Mansimov, E., & Salakhutdinov, R. (2015). Unsupervised learning of video representations using LSTMs. In Bach, F. R., & Blei, D. M. (Eds.), *Proceedings of*

the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015, Vol. 37 of *JMLR Workshop and Conference Proceedings*, pp. 843–852. JMLR.org.

Storks, S., Gao, Q., & Chai, J. Y. (2019). Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. *CoRR, abs/1904.01172*.

Strub, F., de Vries, H., Mary, J., Piot, B., Courville, A. C., & Pietquin, O. (2017). End-to-end optimization of goal-driven and visually grounded dialogue systems. In Sierra, C. (Ed.), *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pp. 2765–2771. ijcai.org.

Strub, F., Seurin, M., Perez, E., De Vries, H., Mary, J., Preux, P., & CourvilleOlivier Pietquin, A. (2018). Visual reasoning with multi-hop feature modulation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 784–800.

Strzalkowski, T., & Harabagiu, S. (2006). *Advances in open domain question answering*, Vol. 32. Springer Science & Business Media.

Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., & Dai, J. (2020). VL-BERT: pre-training of generic visual-linguistic representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Su, Y., Fan, K., Bach, N., Kuo, C. J., & Huang, F. (2019). Unsupervised multi-modal neural machine translation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 10482–10491. Computer Vision Foundation / IEEE.

Suhr, A., Lewis, M., Yeh, J., & Artzi, Y. (2017). A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pp. 217–223. Association for Computational Linguistics.

Suhr, A., Zhou, S., Zhang, A., Zhang, I., Bai, H., & Artzi, Y. (2019). A corpus for reasoning about natural language grounded in photographs. In Korhonen, A., Traum, D. R., & Màrquez, L. (Eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 6418–6428. Association for Computational Linguistics.

Sukhbaatar, S., Weston, J., Fergus, R., et al. (2015). End-to-end memory networks. In *Advances in neural information processing systems*, pp. 2440–2448.

Sun, C., Myers, A., Vondrick, C., Murphy, K., & Schmid, C. (2019). VideoBERT: A joint model for video and language representation learning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 7463–7472. IEEE.

Sutton, R. S., Barto, A. G., et al. (1998). *Introduction to reinforcement learning*, Vol. 135. MIT press Cambridge.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S. E., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *IEEE Conference*

on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, pp. 1–9. IEEE Computer Society.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 2818–2826. IEEE Computer Society.

Tan, H., & Bansal, M. (2019). Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5103–5114.

Tan, H., Yu, L., & Bansal, M. (2019). Learning to navigate unseen environments: Back translation with environmental dropout. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2610–2621.

Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pp. 6105–6114.

Tang, Y., Ding, D., Rao, Y., Zheng, Y., Zhang, D., Zhao, L., Lu, J., & Zhou, J. (2019). COIN: A large-scale dataset for comprehensive instructional video analysis. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 1207–1216. Computer Vision Foundation / IEEE.

Tapaswi, M., Bäuml, M., & Stiefelhagen, R. (2015). Book2movie: Aligning video scenes with book chapters. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1827–1835. IEEE Computer Society.

Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., & Fidler, S. (2016). MovieQA: Understanding Stories in Movies through Question-Answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Thomas, J. A. (2014). *Meaning in interaction: An introduction to pragmatics*. Routledge.

Thomason, J., Murray, M., Cakmak, M., & Zettlemoyer, L. (2019). Vision-and-dialog navigation. In Kaelbling, L. P., Kragic, D., & Sugiura, K. (Eds.), *3rd Annual Conference on Robot Learning, CoRL 2019, Osaka, Japan, October 30 - November 1, 2019, Proceedings*, Vol. 100 of *Proceedings of Machine Learning Research*, pp. 394–406. PMLR.

Thomason, J., Venugopalan, S., Guadarrama, S., Saenko, K., & Mooney, R. (2014). Integrating language and vision to generate natural language descriptions of videos in the wild. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 1218–1227.

Tieleman, T., & Hinton, G. (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, *4*(2), 26–31.

Torabi, A., Pal, C. J., Larochelle, H., & Courville, A. C. (2015). Using descriptive video services to create a large data source for video annotation research. *CoRR*, *abs/1503.01070*.

Tran, D., Bourdev, L. D., Fergus, R., Torresani, L., & Paluri, M. (2014). C3d: generic features for video analysis. *CoRR, abs/1412.0767*, *2*(7), 8.

Tsai, Y.-H. H., Huang, L.-K., & Salakhutdinov, R. (2017). Learning robust visual-semantic embeddings. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3591–3600. IEEE.

Tu, K., Meng, M., Lee, M. W., Choe, T. E., & Zhu, S. C. (2014). Joint video and text parsing for understanding events and answering queries. *IEEE MultiMedia*, *21*(2), 42–70.

Vasudevan, A. B., Dai, D., & Gool, L. V. (2018). Object referring in videos with language and human gaze. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 4129–4138. IEEE Computer Society.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., & Garnett, R. (Eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pp. 5998–6008.

Vedantam, R., Desai, K., Lee, S., Rohrbach, M., Batra, D., & Parikh, D. (2019). Probabilistic neural-symbolic models for interpretable visual question answering. *CoRR*, *abs/1902.07864*.

Vedantam, R., Zitnick, C. L., & Parikh, D. (2015). CIDEr: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 4566–4575. IEEE Computer Society.

Venugopalan, S., Hendricks, L. A., Mooney, R. J., & Saenko, K. (2016). Improving lstm-based video description with linguistic knowledge mined from text. In Su, J., Carreras, X., & Duh, K. (Eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pp. 1961–1966. The Association for Computational Linguistics.

Venugopalan, S., Hendricks, L. A., Rohrbach, M., Mooney, R., Darrell, T., & Saenko, K. (2017). Captioning images with diverse objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., & Saenko, K. (2015a). Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pp. 4534–4542.

Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R. J., & Saenko, K. (2015b). Translating videos to natural language using deep recurrent neural networks. In Mihalcea, R., Chai, J. Y., & Sarkar, A. (Eds.), *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pp. 1494–1504. The Association for Computational Linguistics.

Vijayakumar, A. K., Cogswell, M., Selvaraju, R. R., Sun, Q., Lee, S., Crandall, D. J., & Batra, D. (2016). Diverse beam search: Decoding diverse solutions from neural sequence models. *CoRR*, *abs/1610.02424*.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 3156–3164. IEEE Computer Society.

Vogel, A., & Jurafsky, D. (2010). Learning to follow navigational directions. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 806–814. Association for Computational Linguistics.

Vu, H., Greco, C., Erofeeva, A., Jafaritazehjan, S., Linders, G., Tanti, M., Testoni, A., Bernardi, R., & Gatt, A. (2018). Grounded textual entailment. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 2354–2368.

Wang, B., Ma, L., Zhang, W., & Liu, W. (2018). Reconstruction network for video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7622–7631.

Wang, C., Yang, H., Bartz, C., & Meinel, C. (2016). Image captioning with deep bidirectional lstms. In *Proceedings of the 2016 ACM on Multimedia Conference*, pp. 988–997. ACM.

Wang, J., Fu, J., Tang, J., Li, Z., & Mei, T. (2018). Show, reward and tell: Automatic generation of narrative paragraph from photo stream by adversarial training. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Wang, J., Ma, L., & Jiang, W. (2020). Temporally grounding language queries in videos by contextual boundary-aware prediction. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 12168–12175. AAAI Press.

Wang, J., Madhyastha, P. S., & Specia, L. (2018). Object counts! bringing explicit detections back into image captioning. In *NAACL-HLT*, pp. 2180–2193. Association for Computational Linguistics.

Wang, L., Li, Y., Huang, J., & Lazebnik, S. (2019). Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(2), 394–407.

Wang, X., Chen, W., Wang, Y.-F., & Wang, W. Y. (2018a). No metrics are perfect: Adversarial reward learning for visual storytelling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 899–909.

Wang, X., Chen, W., Wu, J., Wang, Y.-F., & Yang Wang, W. (2018b). Video captioning via hierarchical reinforcement learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4213–4222.

Wang, X., Huang, Q., Çelikyilmaz, A., Gao, J., Shen, D., Wang, Y., Wang, W. Y., & Zhang, L. (2019a). Reinforced cross-modal matching and self-supervised imitation learning

for vision-language navigation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 6629–6638. Computer Vision Foundation / IEEE.

Wang, X., Wu, J., Chen, J., Li, L., Wang, Y., & Wang, W. Y. (2019b). Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 4580–4590. IEEE.

Wang, X., Xiong, W., Wang, H., & Yang Wang, W. (2018). Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead vision-and-language navigation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 37–53.

Wang, Z., Hamza, W., & Florian, R. (2017). Bilateral multi-perspective matching for natural language sentences. In Sierra, C. (Ed.), *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pp. 4144–4150. ijcai.org.

Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45.

Whitehead, S., Ji, H., Bansal, M., Chang, S.-F., & Voss, C. (2018). Incorporating background knowledge into video description generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3992–4001.

Wu, H., Mao, J., Zhang, Y., Jiang, Y., Li, L., Sun, W., & Ma, W.-Y. (2019). Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6609–6618.

Wu, Q., Shen, C., Wang, P., Dick, A. R., & van den Hengel, A. (2018). Image captioning and visual question answering based on attributes and external knowledge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(6), 1367–1381.

Wu, Q., Teney, D., Wang, P., Shen, C., Dick, A. R., & van den Hengel, A. (2017). Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, *163*, 21–40.

Wu, Q., Wang, P., Shen, C., Reid, I. D., & van den Hengel, A. (2018). Are you talking to me? Reasoned visual dialog generation through adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6106–6115. IEEE Computer Society.

Xie, N., Lai, F., Doran, D., & Kadav, A. (2019). Visual entailment: A novel task for fine-grained image understanding. *CoRR*, *abs/1901.06706*.

Xu, H., Li, B., Ramanishka, V., Sigal, L., & Saenko, K. (2019). Joint event detection and description in continuous video streams. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 396–405. IEEE.

Xu, J., Mei, T., Yao, T., & Rui, Y. (2016). MSR-VTT: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and*

Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pp. 5288–5296. IEEE Computer Society.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A. C., Salakhutdinov, R., Zemel, R. S., & Bengio, Y. (2015a). Show, Attend and Tell: Neural image caption generation with visual attention. In Bach, F. R., & Blei, D. M. (Eds.), *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, Vol. 37 of *JMLR Workshop and Conference Proceedings*, pp. 2048–2057. JMLR.org.

Xu, R., Xiong, C., Chen, W., & Corso, J. J. (2015b). Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., & He, X. (2018). AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 1316–1324. IEEE Computer Society.

Yang, G. R., Ganichev, I., Wang, X., Shlens, J., & Sussillo, D. (2018). A dataset and architecture for visual reasoning with a working memory. In Ferrari, V., Hebert, M., Sminchisescu, C., & Weiss, Y. (Eds.), *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part X*, Vol. 11214 of *Lecture Notes in Computer Science*, pp. 729–745. Springer.

Yang, S., Li, G., & Yu, Y. (2019). Cross-modal relationship inference for grounding referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4145–4154.

Yang, Y., Teo, C. L., Daumé III, H., & Aloimonos, Y. (2011). Corpus-guided sentence generation of natural images. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 444–454. Association for Computational Linguistics.

Yang, Z., Yuan, Y., Wu, Y., Cohen, W. W., & Salakhutdinov, R. R. (2016). Review networks for caption generation. In *Advances in Neural Information Processing Systems*, pp. 2361–2369.

Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C. J., Larochelle, H., & Courville, A. C. (2015). Video description generation incorporating spatio-temporal features and a soft-attention mechanism. *CoRR*, *abs/1502.08029*.

Yao, T., Yingwei, P., Yehao, L., & Mei, T. (2017). Incorporating copying mechanism in image captioning for learning novel objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Yao, Y., Xu, J., Wang, F., & Xu, B. (2018). Cascaded mutual modulation for visual reasoning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 975–980.

Yi, K., Wu, J., Gan, C., Torralba, A., Kohli, P., & Tenenbaum, J. (2018). Neural-Symbolic VQA: Disentangling reasoning from vision and language understanding. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., & Garnett, R. (Eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural*

Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada, pp. 1039–1050.

Yin, X., & Ordonez, V. (2017). Obj2Text: Generating visually descriptive language from object layouts. In *EMNLP*, pp. 177–187. Association for Computational Linguistics.

Yoshikawa, Y., Shigeto, Y., & Takeuchi, A. (2017). STAIR Captions: Constructing a large-scale japanese image caption dataset. In Barzilay, R., & Kan, M. (Eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pp. 417–421. Association for Computational Linguistics.

You, Q., Jin, H., Wang, Z., Fang, C., & Luo, J. (2016). Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4651–4659.

Young, P., Lai, A., Hodosh, M., & Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, *2*, 67–78.

Yu, H., Wang, J., Huang, Z., Yang, Y., & Xu, W. (2016). Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4584–4593.

Yu, L., Bansal, M., & Berg, T. (2017). Hierarchically-attentive rnn for album summarization and storytelling. In *Empirical Methods in Natural Language Processing*.

Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., & Berg, T. L. (2018). MAttNet: Modular attention network for referring expression comprehension. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 1307–1315. IEEE Computer Society.

Yu, L., Park, E., Berg, A. C., & Berg, T. L. (2015). Visual Madlibs: Fill in the blank image generation and question answering. *CoRR*, *abs/1506.00278*.

Yu, L., Poirson, P., Yang, S., Berg, A. C., & Berg, T. L. (2016). Modeling context in referring expressions. In Leibe, B., Matas, J., Sebe, N., & Welling, M. (Eds.), *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*, Vol. 9906 of *Lecture Notes in Computer Science*, pp. 69–85. Springer.

Yu, L., Tan, H., Bansal, M., & Berg, T. L. (2017a). A joint speaker-listener-reinforcer model for referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7282–7290.

Yu, Y., Ko, H., Choi, J., & Kim, G. (2017b). End-to-end concept word detection for video captioning, retrieval, and question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3261–3269. IEEE Computer Society.

Zellers, R., Bisk, Y., Farhadi, A., & Choi, Y. (2019). From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6720–6731.

Zeng, K.-H., Chen, T.-H., Chuang, C.-Y., Liao, Y.-H., Niebles, J. C., & Sun, M. (2017). Leveraging video descriptions to learn video question answering. In *AAAI*, pp. 4334–4340. AAAI Press.

Zeng, K., Chen, T., Niebles, J. C., & Sun, M. (2016). Title generation for user generated videos. In Leibe, B., Matas, J., Sebe, N., & Welling, M. (Eds.), *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*, Vol. 9906 of *Lecture Notes in Computer Science*, pp. 609–625. Springer.

Zhang, C., Gao, F., Jia, B., Zhu, Y., & Zhu, S. (2019). RAVEN: A dataset for relational and analogical visual reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 5317–5327. Computer Vision Foundation / IEEE.

Zhang, H., Xu, T., & Li, H. (2017). StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 5908–5916. IEEE Computer Society.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., & Metaxas, D. N. (2019). StackGAN++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(8), 1947–1962.

Zhang, H., Niu, Y., & Chang, S.-F. (2018). Grounding referring expressions in images by variational context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4158–4166.

Zhang, P., Goyal, Y., Summers-Stay, D., Batra, D., & Parikh, D. (2016). Yin and yang: Balancing and answering binary visual questions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5014–5022.

Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., & Gao, J. (2021). Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5579–5588.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zhang, Y., Hare, J. S., & Prügel-Bennett, A. (2018). Learning to count objects in natural images for visual question answering. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Zhang, Z., Shi, Y., Yuan, C., Li, B., Wang, P., Hu, W., & Zha, Z. (2020). Object relational graph with teacher-recommended learning for video captioning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 13275–13285. IEEE.

Zhang, Z., Xie, Y., & Yang, L. (2018). Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6199–6208.

Zhao, W., Wang, B., Ye, J., Yang, M., Zhao, Z., Luo, R., & Qiao, Y. (2018). A multi-task learning approach for image captioning. In Lang, J. (Ed.), *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pp. 1205–1211. ijcai.org.

Zhao, Z., Yang, Q., Cai, D., He, X., & Zhuang, Y. (2017). Video question answering via hierarchical spatio-temporal attention networks. In Sierra, C. (Ed.), *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pp. 3518–3524. ijcai.org.

Zheng, Z., Wang, W., Qi, S., & Zhu, S.-C. (2019). Reasoning visual dialogs with structural and partial observations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6669–6678.

Zhou, L., Kalantidis, Y., Chen, X., Corso, J. J., & Rohrbach, M. (2019). Grounded video description. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 6578–6587. Computer Vision Foundation / IEEE.

Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J. J., & Gao, J. (2020). Unified vision-language pre-training for image captioning and VQA. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 13041–13049. AAAI Press.

Zhou, L., Xu, C., & Corso, J. J. (2018a). Towards automatic learning of procedures from web instructional videos. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Zhou, L., Zhou, Y., Corso, J. J., Socher, R., & Xiong, C. (2018b). End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8739–8748.

Zhou, M., Cheng, R., Lee, Y. J., & Yu, Z. (2018c). A visual attention grounding neural model for multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3643–3653.

Zhou, Y., Sun, Y., & Honavar, V. G. (2019). Improving image captioning by leveraging knowledge graphs. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2019, Waikoloa Village, HI, USA, January 7-11, 2019*, pp. 283–293. IEEE.

Zhu, D., Mogadala, A., & Klakow, D. (2019). Image manipulation with natural language using two-sided attentive conditional generative adversarial network. *CoRR*, *abs/1912.07478*.

Zhu, L., Xu, Z., Yang, Y., & Hauptmann, A. G. (2017). Uncovering the temporal context for video question answering. *International Journal of Computer Vision*, *124*(3), 409–421.

Zhu, Y., Groth, O., Bernstein, M. S., & Fei-Fei, L. (2016). Visual7w: Grounded question answering in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4995–5004. IEEE Computer Society.

Zhu, Y., Kiros, R., Zemel, R. S., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *ICCV*, pp. 19–27. IEEE Computer Society.

Zhuang, B., Wu, Q., Shen, C., Reid, I. D., & van den Hengel, A. (2018). Parallel Attention: A unified framework for visual object discovery through dialogs and queries. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 4252–4261. IEEE Computer Society.

Zitnick, C. L., & Dollár, P. (2014). Edge Boxes: Locating object proposals from edges. In Fleet, D. J., Pajdla, T., Schiele, B., & Tuytelaars, T. (Eds.), *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, Vol. 8693 of *Lecture Notes in Computer Science*, pp. 391–405. Springer.