Journal of Artificial Intelligence Research 68 (2020) 661-689 Submitted 03/2020; published 08/2020

Image Captioning using Facial Expression and Attention

Omid Mohamad Nezami OMID.MOHAMAD-NEZAMI@HDR.MQ.EDU.AU
Macquarie University, Sydney, NSW, Australia
CSIRO's Data61, Sydney, NSW, Australia

Mark Dras MARK.DRAS@MQ.EDU.AU
Macquarie University, Sydney, NSW, Australia

Stephen Wan STEPHEN.WAN@DATA61.CSIRO.AU
CSIRO's Data61, Sydney, NSW, Australia

Cécile Paris CECILE.PARIS@DATA61.CSIRO.AU
Macquarie University, Sydney, NSW, Australia
CSIRO's Data61, Sydney, NSW, Australia

Abstract

Benefiting from advances in machine vision and natural language processing techniques, current image captioning systems are able to generate detailed visual descriptions. For the most part, these descriptions represent an objective characterisation of the image, although some models do incorporate subjective aspects related to the observer's view of the image, such as sentiment; current models, however, usually do not consider the emotional content of images during the caption generation process. This paper addresses this issue by proposing novel image captioning models which use facial expression features to generate image captions. The models generate image captions using long short-term memory networks applying facial features in addition to other visual features at different time steps. We compare a comprehensive collection of image captioning models with and without facial features using all standard evaluation metrics. The evaluation metrics indicate that applying facial features with an attention mechanism achieves the best performance, showing more expressive and more correlated image captions, on an image caption dataset extracted from the standard Flickr 30K dataset, consisting of around 11K images containing faces. An analysis of the generated captions finds that, perhaps unexpectedly, the improvement in caption quality appears to come not from the addition of adjectives linked to emotional aspects of the images, but from more variety in the actions described in the captions.

1. Introduction

Image captioning systems aim to describe the content of an image using Computer Vision and Natural Language Processing approaches which have led to important and practical applications such as helping visually impaired individuals (Vinyals et al., 2015). This is a challenging task because we have to capture not only the objects but also their relationships, and the activities displayed in the image in order to generate a meaningful description. The impressive progress of deep neural networks and large image captioning datasets has resulted in a considerable improvement in generating automatic image captions (Vinyals et al., 2015; Xu et al., 2015; Johnson et al., 2016; You et al., 2016a; Rennie et al., 2017; Chen et al., 2017; Lu et al., 2017; Anderson et al., 2018; Tian et al., 2019).

However, current image captioning methods often overlook the emotional aspects of the image, which play an important role in generating captions that are more semantically correlated with the

Figure 1: The examples of Flickr 30K dataset (Young et al., 2014) with emotional content. The green color indicates words with strong emotional values.

visual content. For example, Figure 1 shows three images with their corresponding human-generated captions including emotional content. The first image at left has the caption of "a dad smiling and laughing with his child" using "smiling" and "laughing" to describe the emotional content of the image. In a similar fashion, 'angry" and "happy" are applied in the second and the third images, respectively. These examples demonstrate how image captioning systems that recognize emotions and apply them can generate richer, more expressive and more human-like captions; this idea of incorporating emotional content is in fact one that is typical to intelligent systems, which researchers like Lisetti (1998) have identified as necessary to generate more effective and adaptive outcomes. Although detecting emotions from visual data has been an active area of research in recent years (Fasel & Luettin, 2003; Sariyanidi et al., 2015), designing an effective image captioning system to employ emotions in describing an image is still an open and challenging problem.

A few models have incorporated sentiment or other non-factual information into image captions (Gan et al., 2017; Mathews et al., 2016; Chen et al., 2018); they typically require the collection of a supplementary dataset, from which a sentiment vocabulary is derived, drawing on work in Natural Language Processing (Pang & Lee, 2008) where sentiment is usually characterized as one of positive, neutral or negative. Mathews et al. (2016), for instance, constructed a sentiment image-caption dataset via crowdsourcing, where annotators were asked to include either positive sentiment (e.g. *a cuddly cat*) or negative sentiment (e.g. *a sinister cat*) using a fixed vocabulary; their model was trained on both this and a standard set of factual captions. These kinds of approaches typically embody descriptions of an image that represent an *observer's* view towards the image (e.g. *a cuddly cat* for a positive view of an image, versus *a sinister cat* for a negative one); they do not aim to capture the emotional content of the image, as in Figure 1.

To capture the emotional content of the image, we propose two groups of models: FACE-CAP[1] and FACE-ATTEND. FACE-CAP feeds in a fixed one-hot encoding vector similar to Hu et al. (2017) and You et al. (2018). In comparison, we represent the aggregate facial expressions of the input image at different time steps of our caption generator, which employs a long short-term memory (LSTM) architecture. To construct the vector, we train a state-of-the-art facial expression recognition (FER) model which automatically recognizes facial expressions (e.g. happiness, sadness, fear, and so on). However, the recognized facial expressions are not always reliable because the FER model

1. An earlier version of FACE-CAP has already been published (Nezami et al., 2018a).

is not 100% accurate. This can result in an image captioning architecture that propagates errors. Hence, we propose an alternative representation that uses more fine-grained facial expression features (e.g. convolutional features) which could potentially be more useful than the one-hot encoding representation. The FACE-ATTEND architecture realises this representation through a visual attention mechanism and dual LSTMs for facial features and general visual content.

The main contributions of the paper are as follows:

- We propose FACE-CAP and FACE-ATTEND models to effectively employ facial expression features with general visual content to generate image captions. To the authors' knowledge, this is the first study to apply facial expression analyses in image captioning tasks.
- Our generated captions using the models are evaluated by all standard image captioning metrics. The results show the effectiveness of the models comparing to a comprehensive list of image captioning models using the FlickrFace11K dataset,[2] the subset of images from the Flickr 30K dataset (Young et al., 2014) that include human faces.
- We further assess the quality of the generated captions in terms of the characteristics of the language used, such as variety of expression. Our analysis suggests that the generated captions by our models improve over other image captioning models by better describing the actions performed in the image.

2. Previous Work

In the following sections, we review image captioning and facial expression recognition models as they are the key parts of our work.

2.1 Image Captioning

There are three main types of image captioning systems: template-based models, retrieval-based models and deep-learning based models (Bernardi et al., 2016; Hossain et al., 2019). Template-based ones first detect visual objects, their attributes and relations and then fill a pre-defined template's blank slots (Farhadi et al., 2010). Retrieval-based ones generate captions using the available captions corresponding to similar images in their corresponding datasets (Hodosh et al., 2013). These classical image captioning models have some limitations. For example, template-based ones cannot generate a wide variety of captions with different lengths, and retrieval-based ones are not able to generate specifically-designed captions for different images. Moreover, classical models do not incorporate the detection and generation steps using an end-to-end training approach. Because of these limitations, modern image captioning models using deep learning are currently the most popular.

Modern image captioning models usually use an encoder-decoder paradigm (Kiros et al., 2014; Vinyals et al., 2015; Xu et al., 2015). They apply a top-down approach where a Convolutional Neural Network (CNN) model learns the image content (encoding), followed by a Long Short-Term Memory (LSTM) generating the image caption (decoding). This follows the paradigm employed in machine translation tasks, using deep neural networks (Sutskever et al., 2014), to translate an image into a caption. This top-down mechanism directly converts the extracted visual features into image captions (Chen & Lawrence Zitnick, 2015; Donahue et al., 2015; Johnson et al., 2016; Karpathy & Fei-Fei, 2015; Mao et al., 2014). However, attending to fine-grained and important fragments of visual data

2. Our dataset splits and labels are publicly available: `https://github.com/omidmnezami/Face-Cap`

in order to provide a better image description is usually difficult using a top-down paradigm. To solve this problem, a combination of a top-down approach and a bottom-up approach, inspired from the classical image captioning models, is proposed by You et al. (2016a). The bottom-up approach overcomes this limitation by generating the relevant words and phrases, which can be detected from visual data with any image resolution, and combining them to form an image caption (Elliott & Keller, 2013; Farhadi et al., 2010; Kulkarni et al., 2013; Kuznetsova et al., 2012).

To attend to fine-grained fragments, attention-based image captioning models have been recently proposed (Xu et al., 2015). These kinds of approaches usually analyze different regions of an image in different time steps of a caption generation process, in comparison to the initial encoder-decoder image captioning systems which consider only the whole image (Vinyals & Le, 2015) as an initial state for generating image captions. They can also take the spatial information of an image into account to generate the relevant words and phrases in the image caption. The current state-of-the-art models in image captioning are attention-based systems (Anderson et al., 2018; Rennie et al., 2017; Xu et al., 2015; You et al., 2016a), explained in the next section, similar to our attention-based image captioning systems.

2.1.1 IMAGE CAPTIONING WITH ATTENTION

Visual attention is an important aspect of the visual processing system of humans (Koch & Ullman, 1987; Corbetta & Shulman, 2002; Spratling & Johnson, 2004; Rensink, 2000), dynamically attending to salient spatial locations in an image with special properties or attributes which are relevant to particular objects. The first image captioning model with attention was proposed by Xu et al. (2015), drawing on earlier use in, for example, machine translation (Bahdanau, Cho, & Bengio, 2014). The model uses visual content extracted from the convolutional layers of CNNs, referred to as spatial features, as the input of a *spatial* attention mechanism to selectively attend to different parts of an image at every time step in generating an image caption. Image captioning with attention differs from previous encoder-decoder image captioning models by concentrating on the salient parts of an input image to generate its equivalent words or phrases simultaneously. Xu et al. (2015) proposed two types of attention including a hard (stochastic) mechanism and a soft (deterministic) mechanism. In the soft attention mechanism, a weighted matrix is calculated to weight a particular part of an image as the input to the decoder (interpreted as a probability value for considering the particular part of the image). The hard attention mechanism, in contrast, picks a sampled annotation vector corresponding to a particular part of an image at each time step as the input to the decoder.

Rennie et al. (2017) extended the work of Xu et al. (2015) by using the CIDEr metric (Vedantam et al., 2015), a standard performance metric for image captioning, to optimize their caption generator instead of optimizing maximum likelihood estimation loss. Their approach was inspired by a Reinforcement Learning approach (Williams, 1992; Sutton & Barto, 1998) called self-critical sequence training, which involves normalizing the reward signals calculated using the CIDEr metric at test time.

Yu et al. (2017) and You et al. (2016a) applied a notion of *semantic* attention to detected visual attributes, learned in an end-to-end fashion, where bottom-up approaches were combined with top-down ones to take advantage of both paradigms. For instance, they acquired a list of semantic concepts or attributes, regarded as a bottom-up mechanism, and used the list with visual features, as an instance of top-down information, to generate an image caption. Semantic attention is used to attend to semantic concepts detected from various parts of a given image. Here, the visual content

was only used in the initial time step. In other time steps, semantic attention was used to select the extracted semantic concepts. That is, semantic attention differs from spatial attention, which attends to spatial features in every time step, and does not preserve the spatial information of the detected concepts.

To preserve spatial information, salient regions can be localized using spatial transformer networks (Jaderberg et al., 2015), which get the spatial features as inputs. This is similar to Faster R-CNN's generation of bounding boxes (Ren et al., 2017), but it is trained in an end-to-end fashion using bilinear interpolation instead of a Region of Interest pooling mechanism as proposed by Johnson et al. (2016). Drawing on this idea, Anderson et al. (2018) applied spatial features to image captioning by using a pre-trained Faster R-CNN and an attention mechanism to discriminate among different visual-based regions regarding the spatial features. Specifically, they combined bottom-up and top-down approaches where a pre-trained Faster R-CNN is used to extract the salient regions from images, instead of using the detected objects as high-level semantic concepts in the work of You et al. (2016a); and an attention mechanism is used to generate spatial attention weights over the convolutional feature maps representing the regions.

In our image captioning systems, we use an attention mechanism weighting visual features as a top-down approach. We also use another attention mechanism to attend to facial expression features as a bottom-up approach. This combination allows our image captioning models to generate captions which are highly correlated with visual content and facial features. To do so, we train a state-of-the-art facial expression recognition model to extract the features. Then, we use the features, via the attention mechanism at each time step, to enrich image captions with emotional content.

2.1.2 IMAGE CAPTIONING WITH STYLE

Most image captioning systems concentrate on describing objective visual content without adding any extra information, giving rise to factual linguistic descriptions. However, there are also stylistic aspects of language which play an essential role in enriching written communication and engaging users during interactions. Style helps in clearly conveying visual content (Mathews et al., 2018), and making the content more attractive (Gan et al., 2017; Chen et al., 2018). It also conveys personality-based (Pennebaker & King, 1999) and emotion-based attributes which can impact on decision making (Mathews et al., 2016).

There are a few models that have incorporated style or other non-factual characteristics into the generated captions (Mathews et al., 2016; Gan et al., 2017; Nezami et al., 2018c, 2019a). In addition to describing the visual content, these models learn to generate different forms or styles of captions. For instance, Mathews et al. (2016) proposed the Senti-Cap system to generate sentiment-bearing captions. Here, the notion of sentiment is drawn from Natural Language Processing (Pang & Lee, 2008), with sentiment either *negative* or *positive*. The Senti-Cap system of Mathews et al. (2016) is a full switching architecture incorporating both factual and sentiment caption paths. In comparison, the work of Gan et al. (2017) consists of a Factored-LSTM learning the stylistic information in addition to the factual information of the input captions. Chen et al. (2018) subsequently applied a mechanism to weight the stylistic and the factual information using Factored-LSTM. All these approaches need two-stage training: training on factual image captions and training on sentiment-bearing image captions. Therefore, they do not support end-to-end training.

To address this issue, You et al. (2018) designed two new schemes, Direct Inject and Sentiment Flow, to better employ sentiment in generating image captions. For Direct Inject, an additional

dimension was added to the input of a recurrent neural network (RNN) to express sentiment, and the sentiment unit is injected at every time step of the generation process. The Sentiment Flow approach of You et al. (2018) injects the sentiment unit only at the initial time step of a designated sentiment cell trained in a similar learning fashion to the memory cell in LSTMs.

All of the above work is focused on subjective descriptions of images using a given sentiment vocabulary, rather than representing the emotional content of the image, as we do in this work. In order to target content-based emotions using visual data, we propose FACE-CAP and FACE-ATTEND models employing attention mechanisms to attend to visual features. We aim to apply the emotional content, recognized using a facial expression analysis, of images themselves during a caption generation process. We use the emotional content to generate image captions without any extra style-based or sentiment-bearing vocabulary: our goal is to see whether, given the existing vocabulary, incorporating the emotional content can produce better captions.

2.2 Facial Expression Recognition

Facial expression is a form of non-verbal communication conveying attitudes, affects, and intentions of individuals. It happens as the result of changes over time in facial features and muscles (Fasel & Luettin, 2003). It is also one of the most important communication means for showing emotions and transferring attitudes in human interactions. Indeed, research on facial expressions started more than a century ago when Darwin published his book titled, "The expression of the emotions in man and animals" (Ekman, 2006). Since then a large body of work has emerged on recognizing facial expressions, usually using a purportedly universal framework of a small number of standard emotions (*happiness*, *sadness*, *fear*, *surprise*, *anger*, and *disgust*) or this set including a *neutral* expression (Field et al., 1982; Kanade et al., 2000; Fasel & Luettin, 2003; Yin et al., 2006; Fridlund, 2014; Sariyanidi et al., 2015; Nezami et al., 2019b) or more fine-grained facial features such as facial action units, defined as the deformations of facial muscles (Tian et al., 2001). Recently, recognizing facial expressions has been paid special attention because of its practical applications in different domains such as education (Nezami et al., 2017, 2018b), health-care and virtual reality (Zeng et al., 2008; Fasel & Luettin, 2003). It is worth mentioning that the automatic recognition of facial expressions is a difficult task because different people express their attitudes in different ways and there are close similarities among various types of facial expressions (Zeng et al., 2018) as shown in Figure 2.

To find effective representations, deep learning based methods have been recently successful in this domain. Due to their complex architectures including multiple layers, they can capture hierarchical structures from low- to high-level representations of facial expression data. Tang (2013), the winner of the 2013 Facial Expression Recognition (FER) challenge (Goodfellow et al., 2013), trained a Convolutional Neural Network (CNN) with a linear support vector machine (SVM) to detect facial expressions. He replaced the softmax layer, used to generate a probability distribution across multiple classes, with a linear SVM and showed a consistent improvement compared to the previous work. Instead of cross-entropy loss, his approach optimizes a margin-based loss to maximize margins among data points belonging to diverse classes.

CNNs are also used for feature extraction and transfer learning in this domain. For example, Kahou et al. (2016) applied a CNN model to recognize facial expressions. Their approach uses a combination of deep neural networks to learn from diverse data modalities including video frames, audio data and spatio-temporal information. The CNN model, as the best model in this work, aims to

Figure 2: Examples from the Facial Expression Recognition 2013 dataset (Goodfellow et al., 2013) including seven standard facial expressions.

recognize emotions from static video frames. Then the recognized emotions are combined across a video clip by a frame aggregation technique and classified using an SVM with a radial basis function kernel. As another example, Zhang et al. (2015) proposed a CNN-based method to recognize social relation traits (e.g. friendly, competitive and dominant) from detected faces in an image. The method includes a CNN model to recognize facial expressions projected into a shared representation space. The space combines the extracted features from two detected faces in an image and generates the predictions of social traits.

The models mentioned above usually use conventional CNN architectures to report the performance on different facial expression recognition datasets including the FER-2013 dataset (Goodfellow et al., 2013), which is a publicly available dataset with a large number of human faces captured in real-life settings. Pramerdorfer and Kampel (2016) instead used an ensemble of very deep architectures of CNNs such as VGGnet, Inception and ResNet by identifying the bottlenecks of the previous state-of-the-art facial expression recognition models on the FER-2013 dataset and achieving a new state-of-the-art result on the dataset. The quality of these recent models is high: it is at least as good as human performance (Goodfellow et al., 2013). The idea of applying VGGnet in facial expression recognition tasks motivates our work to make a facial expression recognition module reproducing the state-of-the-art result on FER-2013 dataset. We use the module to extract facial features from human faces to apply in our image captioning models.

3. Approach

In this section, we describe FACE-CAP and FACE-ATTEND, our proposed models for generating image captions using facial expression analyses. The models are inspired by two popular image captioning models, specifically Show-Attend-Tell (Xu et al., 2015) and Up-Down-Captioner (Anderson et al., 2018).

Show-Attend-Tell is a well-known and widely used image captioning system that incorporates an attention mechanism to attend to spatial visual features. It demonstrates a significant improvement over earlier image captaining models that do not have an attention mechanism. From this starting

point, we propose the FACE-CAP model which similarly attends to visual features and additionally uses facial expression analyses in generating image captions. FACE-CAP incorporates a one-hot encoding vector as a representation of the facial expression analysis, similar to the representations used for sentiment by Hu et al. (2017) and You et al. (2018).

Up-Down-Captioner is a current state-of-the-art image captioning model, defining a new architecture to incorporate attended visual features in generating image captions. In this model, the features directly relate to the objects in the image and two LSTMs (one for generating attention weights and another one for a language model) are used to generate image captions. We propose FACE-ATTEND based on this kind of architecture, as we can apply more fine-grained facial expression features and use two LSTMs, one to attend to those features and the second to general visual features. Because Up-Down-Captioner already incorporates attention over objects in the image, our models derived from this allow us to examine the effectiveness of the facial expression features beyond just recognition of the face as an object.

In what follows, we describe our datasets and our facial expression recognition model that are used by FACE-CAP and FACE-ATTEND. We then explain FACE-CAP in Section 3.3.1 and FACE-ATTEND in Section 3.3.2.

3.1 Datasets

In this work, we use two datasets described below to train our facial expression recognition and image captioning models, respectively.

Facial Expression Recognition To train our facial expression recognition model, we use the facial expression recognition 2013 (FER-2013) dataset (Goodfellow et al., 2013). It includes images labeled with standard facial expression categories (*happiness*, *sadness*, *fear*, *surprise*, *anger*, *disgust* and *neutral*). It consists of 35,887 examples (standard splits are 28,709 for training, 3589 for public and 3589 for private test), collected by means of the Google search API. The examples are in grayscale at the size of 48-by-48 pixels. For our purposes, we split the standard training set of FER-2013 into two sections after removing 11 completely black examples: 25,109 for training our models and 3589 for development and validation. Similar to other work in this domain (Kim et al., 2016; Pramerdorfer & Kampel, 2016; Yu & Zhang, 2015), we use the private test set of FER-2013 for the performance evaluation of the model after the training phase. To compare with the related work, we do not apply the public test set either for training or for validating the model.

Image Captioning To train FACE-CAP and FACE-ATTEND, we have extracted a subset of the Flickr 30K dataset with image captions (Young et al., 2014) that we name FlickrFace11K. It contains 11,696 images including human faces detected using a convolutional neural network-based face detector (King, 2009). (The latest version of Dlib library is applied.) Each image has five ground-truth captions. We observe that the Flickr 30K dataset is a good source for our dataset, because it has a larger portion of images that include human faces, in comparison with other image caption datasets such as the MSCOCO dataset (Lin et al., 2014). We split the FlickrFace11K samples into 8696 for training, 2000 for validation and 1000 for testing. Since we aim to train a facial expression recognition model on FER-2013 and use it as a facial expression feature extractor on the samples of FlickrFace11K, we need to make the samples consistent with the FER-2013 data. To this end, the face detector is used to pre-process the faces of FlickrFace11K. The faces are cropped from each

sample. Then, we transform each face to grayscale and resize it into 48-by-48 pixels, which is the same as in the FER-2013 data.

3.2 Facial Expression Recognition Model

For our core models, we train a facial expression recognition (FER) model using the VGG-B architecture (Simonyan & Zisserman, 2014), because of its strong performance in Pramerdorfer and Kampel (2016). We remove the last convolutional block, including two convolutional layers, and the last max pooling layer from the architecture. We use 3×3 kernel sizes for all remaining convolutional layers. We use a batch normalization layer (Ioffe & Szegedy, 2015) after every remaining convolutional block. Our FER model gives a similar performance to the state-of-the-art under a similar experimental setting, as described in Pramerdorfer and Kampel (2016); this is higher than reported human performance (Goodfellow et al., 2013).

From the FER model, we extract two classes of facial expression features to use in our image captioning models. The first class of features is the output of the final softmax layer of our FER model, $a_i = (a_{i,1}, \ldots, a_{i,7})$, representing the probability distribution of the facial expression classes for the ith face in the image. For the image as a whole, we construct a vector of facial expression features $s = \{s_1, \ldots, s_7\}$ used in our image captioning model as in Equation 1.

$$s_k = \begin{cases} 1 & \text{for } k = \arg\max \sum_{1 \leq i \leq n} a_{i,j}, \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where n is the number of faces in the image. That is, s is a one-hot encoding, which we refer to as the facial encoding vector, of the aggregate facial expressions of the image.

The second class of features consist of convolutional features extracted from the FER model, giving a more fine-grained representation of the faces in the image. For each face in an image, we extract the last convolutional layer of the model, giving $6 \times 6 \times 512$ features. We convert these into a 36×512 representation for each face. We restrict ourselves to a maximum of three faces: in our FlickrFace11K dataset, 96.5% of the images have at most three faces. If one image has more than three faces, we select the three faces with the biggest bounding box sizes. We then concatenate the features of the three faces leading to 108×512 dimensions, $f = \{f_1, ..., f_{K^\star}\}, f_i \in \mathbb{R}^D$, where $K^\star$ is 108 and D is 512; we refer to these as facial features. If a sample includes fewer than three faces, we fill in dimensions with zero values.

In addition to these VGG-based representations, for comparison we trained two FER models using the high-performing ResNet and Inception architectures (Szegedy et al., 2015; He et al., 2016). The performance of these two additional models is similar to the VGG architecture. Similar to the VGG-based model, we can use these ResNet and Inception-based models to extract FER features. We use these primarily for comparison within our FACE-ATTEND models to assess the effect of different fine-grained facial representations.

3.3 Image Captioning Models

Our image captioning models aim to generate an image caption, $x = \{x_1, \ldots, x_T\}$, where x_i is a word and T is the length of the caption, using facial expression analyses. As a representation of the image, all our models use the last convolutional layer of the VGG-E architecture (Simonyan & Zisserman, 2014). In addition to our proposed facial features, the VGG-E network trained on

ImageNet (Russakovsky et al., 2015) produces a $14 \times 14 \times 512$ feature map. We convert this into a 196×512 representation, $c = \{c_1, ..., c_K\}, c_i \in \mathbb{R}^D$, where K is 196 and D is 512; we refer to this as the visual features. The specifics of the image captioning models are explained below.

3.3.1 FACE-CAP

These models essentially extend the Show-Attend-Tell architecture of Xu et al. (2015). Like these models, we use a long short-term memory (LSTM) network as our caption generator. The LSTM incorporates the emotional content of the image in the form of the facial encoding vector defined in Equation 1. We propose two variants, FACE-CAP-REPEAT and FACE-CAP-MEMORY, that differ in terms of how the facial encoding vector is incorporated.

FACE-CAP-REPEAT In FACE-CAP-REPEAT, in each time step (t), the LSTM uses the previous word embedded in M dimensions ($w_{t-1} \in \mathbb{R}^M$ selected from an embedding matrix learned without pre-training from random initial values), the previous hidden state (h_{t-1}), the attention-based features ($\hat{c}_t$), and the facial encoding vector (s) to calculate input gate (i_t), forget gate (f_t), output gate (o_t), input modulation gate (g_t), memory cell (c_t), and hidden state (h_t).

$$\begin{aligned}
i_t &= \sigma(W_i w_{t-1} + U_i h_{t-1} + C_i \hat{c}_t + S_i s + b_i) \\
f_t &= \sigma(W_f w_{t-1} + U_f h_{t-1} + C_f \hat{c}_t + S_f s + b_f) \\
o_t &= \sigma(W_o w_{t-1} + U_o h_{t-1} + C_o \hat{c}_t + S_o s + b_o) \\
g_t &= \tanh(W_g w_{t-1} + U_g h_{t-1} + C_g \hat{c}_t + S_g s + b_g) \\
c_t &= f_t c_{t-1} + i_t g_t \\
h_t &= o_t \tanh(c_t)
\end{aligned} \tag{2}$$

where W, U, C, S, and b are learned weights and biases and σ is the logistic sigmoid activation function. From now on, we show this LSTM equation using the shorthand of Equation 3.

$$h_t = \text{LSTM}(h_{t-1}, [\hat{c}_t, w_{t-1}, s]) \tag{3}$$

To calculate $\hat{c}_t$, for each time step t, FACE-CAP-REPEAT weights visual features (c) using a soft attention mechanism as in Equation 4 and 5.

$$\begin{aligned}
e_{i,t} &= W_e^T \tanh(W_c c_i + W_h h_{t-1}) \\
e'_t &= \text{softmax}(e_t)
\end{aligned} \tag{4}$$

where $e_{i,t}$ are unnormalized weights for the visual features (c_i) and e'_t are the normalized weights using a softmax layer at time step t. Our trained weights are represented by W_x. Finally, our attention-based features ($\hat{c}_t$) are calculated using:

$$\hat{c}_t = \sum_{1 \le i \le K} e'_{i,t} c_i \tag{5}$$

To initialize the LSTM's hidden state (h_0), we feed the facial features through a standard multilayer perceptron, shown in Equation 6.

$$h_0 = \text{MLP}_{init}(s) \tag{6}$$

We use the current hidden state (h_t) to calculate the negative log-likelihood of s in each time step (Equation 7); we call this the face objective function.

$$L_f = -\sum_{1 \leq i \leq 7} s_i \log(p_e(i|h_t)) \tag{7}$$

where a multilayer perceptron generates $p_e(i|h_t)$, which is the categorical probability distribution of the current hidden state across the facial expression classes. We adapt this from You et al. (2018), who use this objective function for injecting ternary-valued sentiment (positive, neutral, negative) into captions. This loss is estimated and averaged, over all steps, during the training phase.

The general objective function of FACE-CAP-REPEAT is defined as:

$$L_{g1} = -\sum_{1 \leq t \leq T} \log(p_x(x_t \,|\, \hat{c}_t, h_t)) + \sum_{1 \leq k \leq K} (1 - \sum_{1 \leq t \leq T} c_t)^2 \tag{8}$$

A multilayer perceptron and a softmax layer is used to calculate p_x, the probability of the next generated word:

$$p_x(x_t \,|\, \hat{c}_t, h_t) = \text{softmax}(W'_c \hat{c}_t + W'_h h_t + b') \tag{9}$$

where the learned weights and bias are given by W' and b'. The last term in Equation 8 is to encourage FACE-CAP-REPEAT to equally pay attention to different sets of c when a caption generation process is finished.

FACE-CAP-MEMORY The above FACE-CAP-REPEAT model feeds in the facial encoding vector at the initial step (Equation 6) and at each time step (Equation 3), shown in Figure 3 (top). The LSTM uses the vector for generating every word because the vector is fed at each time step. Since not all words in the ground truth captions will be related to the vector — for example in Figure 1, where the majority of words are not directly related to the facial expressions — this mechanism could lead to an overemphasis on these features.

Our second variant of the model, FACE-CAP-MEMORY, is as above except that the s term is removed from Equation 3: we do not apply the facial encoding vector at each time step (Figure 3 (bottom)) and rely only on Equation 7 to memorize this facial expression information. Using this mechanism, the LSTM can effectively take the information in generating image captions and ignore the information when it is irrelevant. To handle an analogous issue for sentiment, You et al. (2018) implemented a sentiment cell, working similarly to the memory cell in the LSTM, initialized by the ternary sentiment. They then fed the visual features to initialize the memory cell and hidden state of the LSTM. Similarly, FACE-CAP-MEMORY uses the facial features to initialize the memory cell and hidden state. Using the attention mechanism, our model applies the visual features in generating every caption word.

3.3.2 FACE-ATTEND

Here, we apply two LSTMs to attend to our more fine-grained facial features (f) explained in Section 3.2, in addition to our visual features (c). We propose two variant architectures for combining these features, DUAL-FACE-ATT and JOINT-FACE-ATT, explained below.

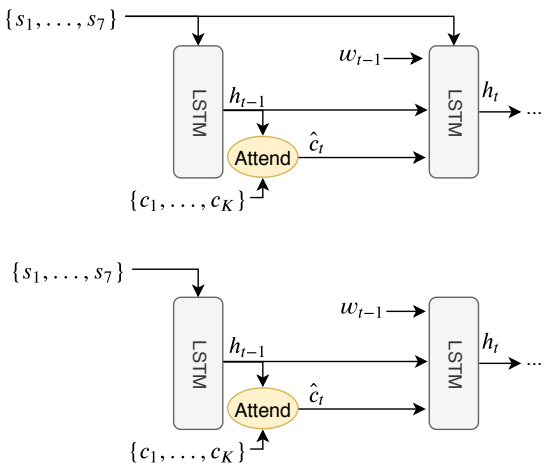

Figure 3: The frameworks of FACE-CAP-REPEAT (top), and FACE-CAP-MEMORY (bottom). Attend is our attention mechanism attending to the visual features, $\{c_1, \ldots, c_K\}$.

DUAL-FACE-ATT The framework of DUAL-FACE-ATT is shown in Figure 4. To generate image captions, DUAL-FACE-ATT includes two LSTMs: one, called F-LSTM, to attend to facial features and another one, called C-LSTM, to attend to visual content. Both LSTMs are defined as in Equation 10, but with separate training parameters.

$$h_{t,z} = \text{LSTM}(h_{t,z-1}, [\hat{z}_t, w_{t-1}]) \tag{10}$$

In both LSTMs, to calculate $\hat{z}_t$ at each time step (t), features z (the facial features (f) for F-LSTM and the visual features (c) for C-LSTM) are weighted using a soft attention mechanism, but with separately learned parameters.

$$\begin{aligned} e_{i,t,z} &= W_{e,z}^T \tanh(W_z z_i + W_{h,z} h_{t,z-1}) \\ e'_{t,z} &= \text{softmax}(e_{t,z}) \end{aligned} \tag{11}$$

where $e_{i,t,z}$ and $e'_{t,z}$ are unnormalized weights for features z_i, and normalized weights using a softmax layer, respectively. Our trained weights are W_z. Finally, our attention-based features ($\hat{z}_t$) are calculated using:

$$\hat{z}_t = \sum_{1 \le i \le K_z} e'_{i,t,z} z_i \tag{12}$$

K_z is $K^\star$ for F-LSTM and K for C-LSTM. The initial LSTM's hidden state ($h_{0,z}$) is computed using a standard multilayer perceptron:

$$h_{0,z} = \text{MLP}_{init,z}(\frac{1}{K_z} \sum_{1 \le i \le K_z} z_i) \tag{13}$$

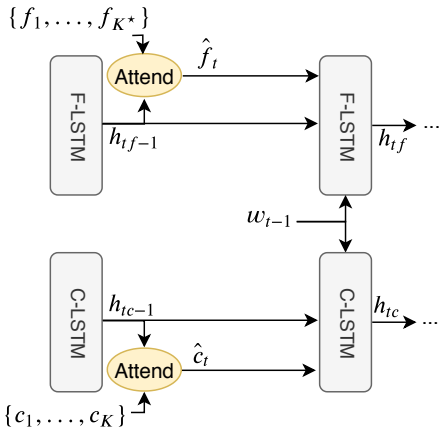

Figure 4: DUAL-FACE-ATT model enables generating image captions with both facial features $\{f_1, \ldots, f_{K^\star}\}$ and visual content $\{c_1, \ldots, c_K\}$.

The objective function of DUAL-FACE-ATT is defined using Equation (14).

$$L_{g2} = -\lambda[\sum_{1\leq t\leq T} \log(p_{x,c}(x_t \mid \hat{c}_t, h_{t,c})) + \sum_{1\leq k\leq K} (1 - \sum_{1\leq t\leq T} c_{t,k})^2] - (1-\lambda)[\sum_{1\leq t\leq T} \log(p_{x,f}(x_t \mid \hat{f}_t, h_{t,f})) + \beta_1 \sum_{1\leq k\leq K^*} (1 - \sum_{1\leq t\leq T} f_{t,k})^2] \quad (14)$$

where a multilayer perceptron and a softmax layer, for each LSTM, are used to calculate $p_{x,f}$ and $p_{x,c}$ (the probabilities of the next generated word on the basis of facial expression features and visual features, respectively):

$$\begin{aligned} p_{x,f}(x_t \mid \hat{f}_t, h_{t,f}) &= \text{softmax}(W_f \hat{f}_t + W_{h,f} h_{t,f} + b_f) \\ p_{x,c}(x_t \mid \hat{c}_t, h_{t,c}) &= \text{softmax}(W_c \hat{c}_t + W_{h,c} h_{t,c} + b_c) \end{aligned} \quad (15)$$

λ and β_1 are regularization constants. The ultimate probability of the next generated word is:

$$p_x(x_t \mid \hat{f}_t, h_{t,f}, \hat{c}_t, h_{t,c}) = \lambda p_{x,f}(x_t \mid \hat{f}_t, h_{t,f}) + (1-\lambda) p_{x,c}(x_t \mid \hat{c}_t, h_{t,c}) \quad (16)$$

JOINT-FACE-ATT The above DUAL-FACE-ATT model uses two LSTMs: one for attending to visual features and another one for attending to facial features. In the model, both LSTMs also play the role of language models (Equation 16) and directly impact on the prediction of the next generated word. However, the recent state-of-the-art image captioning model of Anderson et al. (2018) achieved better performance by using two LSTMs with differentiated roles: one for attending only to visual features and a second one purely as a language model. Inspired by this, we define our JOINT-FACE-ATT variant to use one LSTM, which we call A-LSTM, to attend to image-based

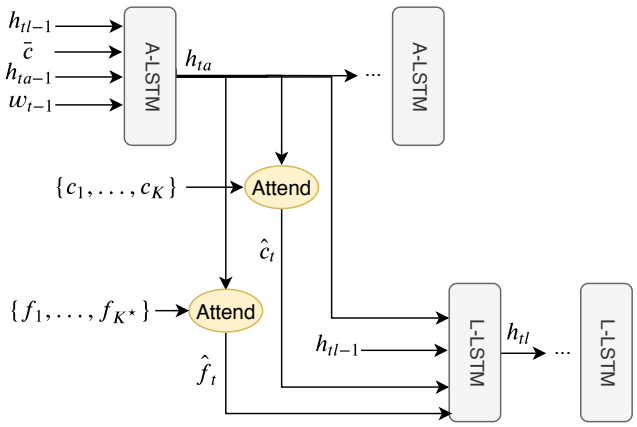

Figure 5: JOINT-FACE-ATT model enables generating image captions with two LSTMs for learning attention weights and generating captions, separately.

features, both facial and visual; and a second one, which we call L-LSTM, to generate language (Figure 5). Here, we calculate the hidden state of A-LSTM using:

$$h_{t,a} = \text{LSTM}(h_{t,a-1}, [\bar{c}, h_{t,l-1}, w_{t-1}]) \tag{17}$$

where $\bar{c} = \frac{1}{K}\sum_{1\leq i\leq K} c_i$ is the mean-pooled visual features and $h_{t,l-1}$ is the previous hidden state of L-LSTM. We also calculate the hidden state of L-LSTM using:

$$h_{t,l} = \text{LSTM}(h_{t,l-1}, [\hat{f}_t, \hat{c}_t, h_{t,a}]) \tag{18}$$

where $\hat{f}_t$ and $\hat{c}_t$ are the attended facial features and visual features, respectively. They are defined analogously to Equation 11 and 12, but $h_{t,z-1} = h_{t,a}$ with different sets of trainable parameters. h_a and h_l are similarly initialized as follows using two standard multilayer perceptrons:

$$\begin{aligned} h_{0,l} &= \text{MLP}_{init,l}(\frac{1}{K}\sum_{1\leq i\leq K} c_i) \\ h_{0,a} &= \text{MLP}_{init,a}(\frac{1}{K}\sum_{1\leq i\leq K} c_i) \end{aligned} \tag{19}$$

The objective function of JOINT-FACE-ATT is:

$$L_{g3} = -[\sum_{1\leq t\leq T} \log(p_x(x_t \,|\, \hat{c}_t, \hat{f}_t, h_{t,l})) + \sum_{1\leq k\leq K}(1 - \sum_{1\leq t\leq T} c_{t,k})^2 + \beta_2 \sum_{1\leq k\leq K^\star}(1 - \sum_{1\leq t\leq T} f_{t,k})^2] \tag{20}$$

where β_2 is a regularization constant and p_x is the probability of the next generated word calculated as follows:

$$p_x(x_t \,|\, \hat{c}_t, \hat{f}_t, h_{t,l}) = \text{softmax}(W_{c,l}\hat{c}_t + W_{f,l}\hat{f}_t + W_{h,l}h_{t,l} + b_l) \tag{21}$$

where $W_{x,l}$ and b_l are trainable weights and bias, respectively.

4. Experiments

In the following sections, we describe the evaluation setup and discuss the experimental results. At the end, we analyse the failure cases.

4.1 Evaluation Metrics

Following previous work, we evaluate our image captioning model using standard evaluation metrics including BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Denkowski & Lavie, 2014), CIDEr (Vedantam et al., 2015), and SPICE (Anderson et al., 2016). Larger values are better results for all metrics. BLEU calculates a weighted average for n-grams with different sizes as a precision metric. ROUGE is a recall-oriented metric that calculates F-measures using the matched n-grams between the generated captions and their corresponding reference summaries. METEOR uses a weighted F-measure matching synonyms and stems in addition to standard n-gram matching. CIDEr uses n-gram matching, calculated using cosine similarity, between the generated captions and the consensus of the reference captions. Finally, SPICE calculates F-score for semantic tuples derived from scene graphs.

4.2 Systems for Comparison

The core architectures for our FACE-CAP and FACE-ATTEND models come from Show-Attend-Tell (Xu et al., 2015) and Up-Down-Captioner (Anderson et al., 2018), respectively. We therefore use these models, trained on the FlickrFace11K dataset, as baselines, in order to provide an ablative assessment of the effect of adding facial expression information. We call these baseline models SHOW-ATT-TELL and UP-DOWN. (Moreover, Anderson et al. (2018) has the state-of-the-art results for image captioning.)

We further look at two additional models to investigate the impact of the face loss function in using the facial encoding in different schemes. We train the FACE-CAP-REPEAT model, which uses the facial encoding in every time step, without calculating the face loss function (Equation (7)); we refer to this (following the terminology of Hu et al. (2017) and You et al. (2018)) as the STEP-INJECT model. The FACE-CAP-MEMORY model, which applies the facial encoding in the initial time step, is also modified in the same way; we refer to this as the INIT-FLOW model.

4.3 Implementation Details

The size of the word embedding layer, initialized via a uniform distribution, is set to 300 except for UP-DOWN and JOINT-FACE-ATT which is set to 512. We fixed 512 dimensions for the memory cell and the hidden state in this work. We use the mini-batch size of 100 and the initial learning rate of 0.001 to train each image captioning model except UP-DOWN and JOINT-FACE-ATT where we set the mini-batch size to 64 and the initial learning rate to 0.005. We used different parameters for UP-DOWN and JOINT-FACE-ATT in comparison with other models because using similar parameters led to worse results for all models. The Adam optimization algorithm (Kingma & Ba, 2014) is used for optimizing all models. During the training phase, if the model does not have an improvement in METEOR score on the validation set in two successive epochs, we divide the learning rate by two (the minimum learning rate is set to 0.0001) and the previous trained model with the best METEOR is reloaded. This method of learning rate decay is inspired by Wilson et al. (2017), who advocated tuning the learning rate decay for Adam. In addition to learning rate decay, METEOR is applied to

select the best model on the validation set because of a reasonable correlation between METEOR and human judgments (Anderson et al., 2016). Although SPICE can have higher correlations with human judgements, METEOR is quicker to calculate than SPICE, which requires dependency parsing, and so more suitable for a training criterion. The epoch limit is set to 30. We use the same vocabulary size and visual features for all models. λ and β_1 in Equation 14 are empirically set to 0.8 and 0.2, respectively. β_2 in Equation 20 is also set to 0.4. Multilayer perceptrons in Equation 6, 13 and 19 use $\tanh$ as an activation function.

4.4 Experimental Results

In the following sections, quantitative and qualitative results are explained in detail.

4.4.1 QUANTITATIVE ANALYSES

Performance Metrics The FlickrFace11K splits are used for training and evaluating all image captioning models in this paper. Table 1 summarizes the results on the FlickrFace11K test set. DUAL-FACE-ATT and JOINT-FACE-ATT outperform other image captioning models using all the evaluation metrics. For example, DUAL-FACE-ATT achieves 17.6 for BLEU-4 which is 1.9 and 0.4 points better that SHOW-ATT-TELL (the first baseline model) and FACE-CAP-MEMORY (the best of the FACE-CAP models), respectively. JOINT-FACE-ATT also achieves a BLEU-4 score of 17.7 which is 0.4 better than UP-DOWN, the baseline model it builds on, and 0.5 better than FACE-CAP-MEMORY. DUAL-FACE-ATT and JOINT-FACE-ATT show very close results, with DUAL-FACE-ATT demonstrating a couple of larger gaps in performance, in the BLEU-1 and ROUGE-L metrics. Among the FACE-CAP models, FACE-CAP-MEMORY is clearly the best.

Table 2 compares DUAL-FACE-ATT-VGG with FER features derived from the VGG architecture (DUAL-FACE-ATT in Table 1 which is our core version in the paper) against DUAL-FACE-ATT-RES using the ResNet architecture and DUAL-FACE-ATT-INC using the Inception architecture (see Section 3.2). This comparison is to investigate the variability of FER features derived from different architectures on the image captioning task; we choose DUAL-FACE-ATT for this as the highest-performing model from Table 1. All three DUAL-FACE-ATT in the table perform similarly, and outperform the SHOW-ATT-TELL model, using all the image captioning metrics. This confirms the broadly similar effectiveness of the FER features from different architectures.

Entropy, Top$_4$ and Ranking of Generated Verbs To analyze what it is about the captions themselves that differs under the various models, with respect to our aim of injecting information about emotional states of the faces in images, we first extracted all generated adjectives, which are tagged using the Stanford part-of-speech tagger software (Toutanova et al., 2003). Perhaps surprisingly, emotions do not manifest themselves in the adjectives in our models: the adjectives used by all systems are essentially the same.

To investigate this further, we took the NRC emotion lexicon (Mohammad & Turney, 2013) and examined the occurrence of words in the captions that also appeared in the lexicon. This widely-used lexicon is characterised as "a list of English words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust)" whose labels have been manually annotated through crowd-sourcing. The labels are based on word associations — annotators were asked "which emotions are associated with a target term" — rather than whether the word *embodies* an emotion; the lexicon thus contains a much larger set of words than is useful for our purposes. (For example, the most frequent word overall in the reference captions that appears in the lexicon is

Table 1: The results of different image captioning models (%) on FlickrFace11K test split. B-N is the BLEU-N metric. The best performances are bold.

Model	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDEr	SPICE
SHOW-ATT-TELL	56.0	35.4	23.1	15.7	17.0	43.7	21.9	9.3
UP-DOWN	57.9	37.3	25.0	17.3	17.5	45.1	24.4	10.1
STEP-INJECT	58.4	37.6	24.8	17.0	17.5	45.0	22.8	9.9
INIT-FLOW	56.6	36.5	24.3	16.9	17.2	44.8	23.1	9.8
FACE-CAP-REPEAT	57.1	36.5	24.1	16.5	17.2	44.8	23.0	9.7
FACE-CAP-MEMORY	58.9	37.9	25.1	17.2	17.4	45.5	24.7	10.0
DUAL-FACE-ATT	**59.4**	**38.2**	25.4	17.6	**17.6**	**45.8**	**24.9**	10.1
JOINT-FACE-ATT	58.6	38.1	**25.6**	**17.7**	**17.6**	45.5	24.8	**10.2**

Table 2: DUAL-FACE-ATT with different sets of FER features, extracted by our FER models using high-performing CNN architectures including VGG, ResNet (RES) and Inception (INC).

Model	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDEr	SPICE
DUAL-FACE-ATT- VGG	59.4	38.2	25.4	17.6	17.6	45.8	24.9	10.1
DUAL-FACE-ATT- RES	58.8	38.1	25.4	17.6	17.3	45.3	23.4	9.8
DUAL-FACE-ATT- INC	58.7	38.0	25.4	17.7	17.3	45.3	23.5	9.7

young, which presumably has some positive emotional associations.) In addition, the set of emotions used in lexicon labels does not exactly correspond to our set. We therefore do not propose to use this lexicon purely automatically, but instead to help in understanding the use of emotion-related words.

Among the reference captions, as noted above the most frequent word from the emotion lexicon was *young*, followed by *white*, *blue* and *black*; all of these presumably have some emotional association, but do not generally embody an emotion. The first word embodying the expression of an emotion is the verb *smiling*, at rank 8, with other similar verbs following closely (e.g. *laughing*, *enjoying*). The highest ranked emotion-embodying adjective is *happy* at rank 26, with a frequency of around 15% of that of *smiling*; other adjectives were much further behind. It is clear that verbs form a more significant expression of emotion in this particular dataset than do adjectives.

To come up with an overall quantification of the different linguistic properties of the generated captions under the models, we therefore focused our investigation on the differences in distributions of the generated verbs. To do this, we calculated three measures. The first is entropy (in the information-theoretic sense), which can indicate which distributions are closer to deterministic and which are more spread out (with a higher score indicating more spread out): in our context, it will indicate the amount of variety in selecting verbs. We calculated entropy using the standard Equation (22).

$$\text{Entropy} = -\sum_{1 \leq i \leq V} p(v_i) \times \log_2(p(v_i)) \tag{22}$$

where V indicates the number of the unique generated verbs and $p(v_i)$ is the probability of each generated verb (v_i), estimated as the Maximum Likelihood Estimate from the sample.

Table 3: The Entropies of all generated verbs and the probability mass of the Top_4 generated verbs (*is*, *are*, *sitting*, and *standing*). Reference means the ground-truth captions.

Model	**Entropy**	**Top$_4$**
Reference	6.9963	32.63%
SHOW-ATT-TELL	2.7864	77.05%
UP-DOWN	2.7092	79.24%
STEP-INJECT	2.9059	74.80%
INIT-FLOW	2.6792	78.78%
FACE-CAP-REPEAT	2.7592	77.68%
FACE-CAP-MEMORY	2.9306	73.65%
DUAL-FACE-ATT	3.0154	71.14%
JOINT-FACE-ATT	2.8074	77.69%

As a second measure, we looked at the four most frequent verbs (Top_4), which are the same for all models (*is*, *sitting*, *are*, *standing*) — these are verbs with relatively little semantic content, and for the most part act as syntactic props for the content words of the sentence. The amount of probability mass left beyond those four verbs is another indicator of variety in verb expression.

Table 3 shows that DUAL-FACE-ATT can generate the most diverse distribution of the verbs compared to other models because it has the highest Entropy. It also shows that DUAL-FACE-ATT has the lowest (best) proportion of the probability mass taken up by Top_4, leaving more for other verbs. In contrast to the results of the standard image captioning metrics shown in Table 1, DUAL-FACE-ATT and JOINT-FACE-ATT show very different behaviour: DUAL-FACE-ATT is clearly superior. Among the FACE-CAP models, as for the overall metrics, FACE-CAP-MEMORY is the best, and is in fact better than JOINT-FACE-ATT. (As a comparison, we also show Entropy and Top_4 for all reference captions (5 human-generated captions per image): human-generated captions are still much more diverse than the best models.)

The two measures above are concerned only with variety of verb choice and not with verbs linked specifically to emotions or facial expressions. For a third measure, therefore, we look at selected individual verbs linked to actions that relate to facial emotion expression, either direct or indirect. Our measure is the rank of the selected verb among all those chosen by a model; higher (i.e. lower-numbered) ranked verbs mean that the model more strongly prefers this verb. Our selected verbs are among those that ranked highly in the reference captions and also appeared in the emotion lexicon.

Table 4 shows a sample of those verbs such as *singing*, *reading* and *laughing*. The baseline SHOW-ATT-TELL model ranks all of those relatively low, where our other baseline UP-DOWN and our models incorporating facial expressions do better. Only FACE-CAP-MEMORY (the best of our FACE-CAP models by overall metrics) and our FACE-ATTEND models manage to use verbs like *laughing* and *reading*.

SAT: Two women and a man are posing for a picture.

UD: A group of people are posing for a picture.

SI: Two men and a woman are smiling.

IF: Two men and a woman are smiling at the camera.

FR: Two men and a woman are smiling.

FM: Two men and a woman are smiling at a camera.

DFA: Three women are smiling and laughing.

JFA: A group of people are posing for a picture.

SAT: A woman with a black shirt and black pants is standing in front of a microphone.

UD: A man in a black shirt and a woman in a black shirt and a woman in a black shirt.

SI: A man with a beard and a beard is playing a guitar.

IF: A man in a black shirt and a black hat is playing a guitar.

FR: A woman with a black shirt and a black hat is holding a microphone.

FM: A woman in a black shirt is holding a microphone.

DFA: A woman in a black dress is singing into a microphone.

JFA: A woman in a black shirt is singing into a microphone.

SAT: A man in a white shirt is sitting at a table with a computer.

UD: A man in a yellow shirt is sitting at a table with a book in his lap.

SI: A man in a yellow shirt is working on a computer.

IF: A woman in a yellow shirt is sitting at a table with a computer.

FR: A man in a yellow shirt is sitting at a table with a computer.

FM: A woman in a yellow shirt is working on a computer.

DFA: A woman in a yellow shirt is reading a book.

JFA: A man in a yellow shirt is working on a computer.

SAT: Two young girls are sitting in a chair.

UD: A woman in a striped shirt is holding a small child in a striped shirt.

SI: A woman with a brown shirt and a blond woman in a blue shirt are smiling.

IF: A woman with a white shirt and a young girl in a blue shirt are sitting in a chair.

FR: A woman and a woman are smiling at the camera.

FM: A woman and a young girl are smiling.

DFA: A man and a woman are smiling at the camera.

JFA: A woman in a striped shirt is smiling at the camera.

Figure 6: Example generated captions using SAT (SHOW-ATT-TELL), UD (UP-DOWN) SI (STEP-INJECT), IF (INIT-FLOW), FR (FACE-CAP-REPEAT), FM (FACE-CAP-MEMORY), DFA (DUAL-FACE-ATT) and JFA (JOINT-FACE-ATT) models.

Table 4: Comparison of different image captioning models in ranking example generated verbs. Higher ranks mean better results.

Model	Smiling	Looking	Singing	Reading	Eating	Laughing
Reference	11	10	27	35	24	40
SHOW-ATT-TELL	19	n/a	15	n/a	24	n/a
UP-DOWN	14	13	9	n/a	15	n/a
STEP-INJECT	11	18	10	n/a	15	n/a
INIT-FLOW	10	21	12	n/a	14	n/a
FACE-CAP-REPEAT	12	20	9	n/a	14	n/a
FACE-CAP-MEMORY	9	18	15	22	13	27
DUAL-FACE-ATT	14	16	9	19	19	25
JOINT-FACE-ATT	15	13	8	15	17	23

4.4.2 QUALITATIVE ANALYSES

In Figure 6, we compare some generated captions by different image captioning models using four representative images. The first one shows that DUAL-FACE-ATT correctly uses *smiling* and *laughing* to capture the emotional content of the image. STEP-INJECT, INIT-FLOW, FACE-CAP-REPEAT and FACE-CAP-MEMORY are also successful in generating *smiling* for the image. For the second sample, DUAL-FACE-ATT and JOINT-FACE-ATT use the relevant verb *singing* to describe the image, while other models cannot generate the verb. Similarly, DUAL-FACE-ATT generates the verb *reading* for the third image. Moreover, most models can correctly generate *smiling* for the forth image except SHOW-ATT-TELL and UP-DOWN which do not use the facial information. INIT-FLOW also cannot generate *smiling* because it uses the facial information only at initial step which provides a weak emotional signal for the model. Here, DUAL-FACE-ATT can generate the most accurate caption ("A man and a woman are smiling at the camera") for the image, while other models generate some errors. For example, FACE-CAP-MEMORY generates "A woman and a young girl are smiling", which does not describe the man in the image.

4.5 Failure Analyses

We also carried out an analysis on examples where our image captioning models fail to generate better captions than the baseline models. We first look quantitatively at these examples via image captioning metrics, focussing on SPICE, and then show a few of these examples.

4.5.1 SEMANTIC SUBCATEGORIES

For our failure analyses, we use the SPICE metric to compare generated captions by different models: SPICE is specifically designed for fine-grained analyses, as described in Anderson et al. (2016), as it can break down scoring into semantic proposition subcategories including object, relation, and attribute; it can also break down attributes further into color, count and size, for example.

Table 5: SPICE and F-scores of the semantic subcategories for all captions generated using different models.

Model	SPICE	Object	Relation	Attribute	Color	Count	Size
SHOW-ATT-TELL	9.3	19.4	3.0	3.7	8.0	2.3	2.9
FACE-CAP-MEMORY	10.0	20.5	3.1	4.5	10.1	2.4	1.3
DUAL-FACE-ATT	10.1	20.1	3.2	4.5	10.1	4.1	2.3

Table 6: SPICE and F-scores of the semantic subcategories where our models generate captions with lower scores compared to the baseline model.

Model	SPICE	Object	Relation	Attribute	Color	Count	Size
SHOW-ATT-TELL	15.6	30.7	6.4	6.9	13.9	6.6	6.5
FACE-CAP-MEMORY	8.7	18.5	2.0	4.0	8.7	3.5	0.8
DUAL-FACE-ATT	9.3	18.8	3.3	4.4	9.0	6.4	2.4

To identify examples where our image captioning models perform worse, we first calculate SPICE scores on individual examples. As image captioning metrics are designed to be applied to a set of captions rather than individual ones, this only gives a rough idea of the quality of an individual caption; we therefore set a threshold on the difference between our models and the baseline (0.05) so as not to include ones where scores are very close and therefore may not be a reliable indicator that the caption is actually worse.

Our analysis uses SHOW-ATT-TELL as the baseline model without FER features, and two of our models: FACE-CAP-MEMORY (our best version using the FER one-hot encoding), and DUAL-FACE-ATT (our best version using the FER convolutional features).

We first show the SPICE F-scores for subcategories over *all* captions, in Table 5. We observe in general that although the overall SPICE scores for our models are better (as in Table 1), they are lower for the size attribute, showing that adding facial expression features can reduce the focus on this attribute in describing visual content. This is particularly the case for FACE-CAP-MEMORY which uses the one-hot encoding version of the features. FACE-CAP-MEMORY is similar to SHOW-ATT-TELL but worse than DUAL-FACE-ATT in terms of the count attribute, perhaps because the one-hot encoding here presents the aggregate facial expressions of the input image (Section 3.2) and ignores the number of individuals in the image.

In terms of the selected subset of captions where our models perform worse, Table 6 shows the average SPICE F-scores for these subcategories. The key difference here is that SHOW-ATT-TELL performs a lot better on this subset in terms of overall SPICE score than it does on all captions (15.6 vs 9.3) while our two models perform just slightly worse on these than on all captions (FACE-CAP-MEMORY: 8.7 vs 10.0; DUAL-FACE-ATT: 9.3 vs 10.1). This relationship also holds for the object subcategory, and for SHOW-ATT-TELL and FACE-CAP-MEMORY for the relation category. This may be because the FER features encourage the models to generate relevant verbs (e.g., *smiling*, *looking*) and nouns (e.g. *camera*, *microphone*) as shown in Table 4 and Figure 6, which are sometimes less

Figure 7: Example images where FACE-CAP and DUAL-FACE-ATT fail to generate better results than SHOW-ATT-TELL.

relevant. Overall, the trade-off that our models appear to make is that performance degrades slightly on images that they are less well-suited to, while boosting performance overall.

4.5.2 EXAMPLES

Figure 7 shows some examples where our two models produce substantially worse captions than the baseline according to the SPICE metric. In the topmost one the baseline SHOW-ATT-TELL generates "a man with a beard and a woman in a black shirt are playing a guitar" while FACE-CAP-MEMORY generates "a woman is playing a guitar and singing into a microphone" and DUAL-FACE-ATT generates "two men are playing a guitar and singing". Notwithstanding some gender confusions, the baseline is scoring higher because of the mention of the clothing, which appears in two of the human

Figure 8: An example image where the SHOW-ATT-TELL model expresses the size attribute in its generated caption.

reference captions, while our models have incorrectly guessed the people are singing (perhaps not unreasonably, given the guitar); FACE-CAP-MEMORY also postulates the existence of a microphone.

In the middle image, SHOW-ATT-TELL generates "two women in a white dress and a man in a white shirt are standing in a crowd" while FACE-CAP-MEMORY generates "a group of people are dancing together" and DUAL-FACE-ATT generates "a man in a white shirt and a woman in a white shirt are standing in front of a microphone". This is another instance where a face-focussed model posits the existence of a microphone, as an object that is commonly near a face.

The final image appears to be one of a number of instances where SPICE is likely to be an inaccurate reflection of human judgement of the relative quality of the models. SHOW-ATT-TELL has "a woman in a red shirt is sitting on a bench with a large large crowd on the side of", while FACE-CAP-MEMORY has "a man in a red shirt is riding a bike" and DUAL-FACE-ATT has "a group of people are riding bikes on a street". None of the human reference captions use the word "bike" even though that is a prominent aspect of the image (there is "moped" and (sic) "mo pad"), while the less salient "crowd" is mentioned in one reference caption, boosting the score of SHOW-ATT-TELL.

As we were also curious about the unexpected advantage that the baseline SHOW-ATT-TELL has in terms of the size attribute, and noting the repeated "large large" generated by SHOW-ATT-TELL in this last example, while conducting our failure analysis we also looked at other examples where this model expressed size. We found that a large number of them looked like the one in Figure 8, where SHOW-ATT-TELL generated "a man in a blue shirt is standing in a room with a large large large large large large large". This was in contrast to FACE-CAP-MEMORY's "a man in a blue jacket is standing in front of a yellow wall" and DUAL-FACE-ATT's "a man in a blue shirt is standing in front of a green car", which are less problematic even though the SHOW-ATT-TELL model scored better than DUAL-FACE-ATT. This repeated word problem is known from neural machine translation (Mi et al., 2016) and is common to neural models in general. Exploring this issue is beyond the scope of this paper, but we do note based just on our observations that our models seemed less prone to this problem of repeated size attributes, even though the SPICE size attribute scores suggest the baseline SHOW-ATT-TELL is evaluated at being better at describing sizes.

5. Conclusion

In this work, we have presented several image captioning models incorporating emotion-related information from facial features. All of our models produce better captions on images including

faces than strong baseline systems, as measured by standard metrics on the FlickrFace11K dataset. In investigating these models, we made the following findings:

- Our models that use a distributed representation of facial emotion (FACE-ATTEND) outperformed those that use a one-hot encoding (FACE-CAP).
- For FACE-CAP models, injecting facial expression information only once at the start outperformed injecting at all steps in caption generation, suggesting that the models shouldn't encourage too strongly the incorporation of this facial expression information. For FACE-ATTEND models, our two different methods for separating information — separate LSTMs for visual and facial features (DUAL-FACE-ATT) versus separate LSTMs for visual and language functions (JOINT-FACE-ATT) — performed fairly similarly in terms of overall metrics.
- A linguistic analysis of the generated captions showed that much of the improvement in our models was manifested through verbs. In particular, under measures of diversity of caption generation, DUAL-FACE-ATT was substantially better than all other models.
- An ablative study on the distributed facial emotion representations in FACE-ATTEND showed similar performance regardless of which of three high-performing facial emotion recognition systems was used.
- A failure analysis showed only minor degradation of performance in those cases where the baseline outperformed out new models.

In terms of improvements to our models, the failure analysis suggests the addition of some mechanism that prevents the models from too strongly encouraging the caption generator to incorporate objects that are associated with faces; the findings regarding the location for incorporating facial expressions in the architecture (in FACE-CAP-MEMORY versus FACE-CAP-REPEAT) could be explored in our other models too.

In terms of broader application of the ideas of this work, there is other recent work that explore other aspects of emotional content in images; we note specifically the dataset of You et al. (2016b). In future work, we are interested in exploring this broader emotional content of images, which is reflected in the NRC Emotion Lexicon we used in our linguistic analysis of captions.

Acknowledgments

The first author was partially supported by scholarships from Macquarie University and CSIRO's Data61, and the second author by a Collaborative Research Project with CSIRO's Data61.

References

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

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). 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 (CVPR)*, pp. 6077–6086.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

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*, *55*, 409–442.

Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., & Chua, T.-S. (2017). Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6298–6306. IEEE.

Chen, T., Zhang, Z., You, Q., Fang, C., Wang, Z., Jin, H., & Luo, J. (2018). factual or emotional: Stylized image captioning with adaptive learning and attention. *arXiv preprint arXiv:1807.03871*.

Chen, X., & Lawrence Zitnick, C. (2015). Mind's eye: A recurrent visual representation for image caption generation. In *CVPR*, pp. 2422–2431. IEEE.

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews neuroscience*, *3*(3), 201.

Denkowski, M., & Lavie, A. (2014). METEOR universal: Language specific translation evaluation for any target language. In *WMT*, pp. 376–380.

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 *CVPR*, pp. 2625–2634. IEEE.

Ekman, P. (2006). *Darwin and facial expression: A century of research in review*. Ishk.

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.

Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., & Forsyth, D. (2010). Every picture tells a story: Generating sentences from images. In *ECCV*, pp. 15–29. Springer.

Fasel, B., & Luettin, J. (2003). Automatic facial expression analysis: a survey. *Pattern recognition*, *36*(1), 259–275.

Field, T. M., Woodson, R., Greenberg, R., & Cohen, D. (1982). Discrimination and imitation of facial expression by neonates. *Science*, *218*(4568), 179–181.

Fridlund, A. J. (2014). *Human facial expression: An evolutionary view*. Academic Press.

Gan, C., Gan, Z., He, X., Gao, J., & Deng, L. (2017). Stylenet: Generating attractive visual captions with styles. In *CVPR*. IEEE.

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., & Lee, D.-H. (2013). Challenges in representation learning: A report on three machine learning contests. In *ICONIP*, pp. 117–124. Springer.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778.

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.

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.

Hu, Z., Yang, Z., Liang, X., Salakhutdinov, R., & Xing, E. P. (2017). Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1587–1596. JMLR. org.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Jaderberg, M., Simonyan, K., & Zisserman, A. (2015). Spatial transformer networks. In *Advances in neural information processing systems*, pp. 2017–2025.

Johnson, J., Karpathy, A., & Fei-Fei, L. (2016). Densecap: Fully convolutional localization networks for dense captioning. In *CVPR*, pp. 4565–4574. IEEE.

Kahou, S. E., Bouthillier, X., Lamblin, P., Gulcehre, C., Michalski, V., Konda, K., Jean, S., Froumenty, P., Dauphin, Y., & Boulanger-Lewandowski, N. (2016). Emonets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces*, *10*(2), 99–111.

Kanade, T., Cohn, J. F., & Tian, Y. (2000). Comprehensive database for facial expression analysis. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, pp. 46–53. IEEE.

Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pp. 3128–3137. IEEE.

Kim, B.-K., Dong, S.-Y., Roh, J., Kim, G., & Lee, S.-Y. (2016). Fusing aligned and non-aligned face information for automatic affect recognition in the wild: A deep learning approach. In *CVPR Workshops*, pp. 48–57. IEEE.

King, D. E. (2009). Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, *10*(Jul), 1755–1758.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kiros, R., Salakhutdinov, R., & Zemel, R. S. (2014). Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*.

Koch, C., & Ullman, S. (1987). Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of intelligence*, pp. 115–141. Springer.

Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., Berg, A. C., & Berg, T. L. (2013). Baby talk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(12), 2891–2903.

Kuznetsova, P., Ordonez, V., Berg, A. C., Berg, T. L., & Choi, Y. (2012). Collective generation of natural image descriptions. In *ACL*, pp. 359–368. Association for Computational Linguistics.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer.

Lisetti, C. (1998). Affective computing..

Lu, J., Xiong, C., Parikh, D., & Socher, R. (2017). Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*, Vol. 6, p. 2.

Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., & Yuille, A. (2014). Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*.

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 *AAAI*, pp. 3574–3580.

Mi, H., Sankaran, B., Wang, Z., & Ittycheriah, A. (2016). Coverage Embedding Models for Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 955–960, Austin, Texas. Association for Computational Linguistics.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon.. *29*(3), 436–465.

Nezami, O. M., Dras, M., Anderson, P., & Hamey, L. (2018a). Face-cap: Image captioning using facial expression analysis. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 226–240. Springer.

Nezami, O. M., Dras, M., Hamey, L., Richards, D., Wan, S., & Paris, C. (2018b). Automatic recognition of student engagement using deep learning and facial expression. *arXiv preprint arXiv:1808.02324*.

Nezami, O. M., Dras, M., Wan, S., & Paris, C. (2018c). Senti-attend: Image captioning using sentiment and attention. *arXiv preprint arXiv:1811.09789*.

Nezami, O. M., Dras, M., Wan, S., Paris, C., & Hamey, L. (2019a). Towards generating stylized image captions via adversarial training. In *Pacific Rim International Conference on Artificial Intelligence*, pp. 270–284. Springer.

Nezami, O. M., Lou, P. J., & Karami, M. (2019b). Shemo: a large-scale validated database for persian speech emotion detection. *Language Resources and Evaluation*, *53*(1), 1–16.

Nezami, O. M., Richards, D., & Hamey, L. (2017). Semi-supervised detection of student engagement.. In *PACIS*, p. 157.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, *2*(1-2), 1–135.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *ACL*, pp. 311–318. Association for Computational Linguistics.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference.. *Journal of personality and social psychology*, *77*(6), 1296.

Pramerdorfer, C., & Kampel, M. (2016). Facial expression recognition using convolutional neural networks: State of the art. *arXiv preprint arXiv:1612.02903*.

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, pp. 1137–1149.

Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., & Goel, V. (2017). Self-critical sequence training for image captioning. In *CVPR*, Vol. 1, p. 3.

Rensink, R. A. (2000). The dynamic representation of scenes. *Visual cognition*, *7*(1-3), 17–42.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., & Bernstein, M. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, *115*(3), 211–252.

Sariyanidi, E., Gunes, H., & Cavallaro, A. (2015). Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE transactions on pattern analysis and machine intelligence*, *37*(6), 1113–1133.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Spratling, M. W., & Johnson, M. H. (2004). A feedback model of visual attention. *Journal of cognitive neuroscience*, *16*(2), 219–237.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *NIPS*, pp. 3104–3112.

Sutton, R. S., & Barto, A. G. (1998). *Introduction to reinforcement learning*, Vol. 135. MIT press Cambridge.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9.

Tang, Y. (2013). Deep learning using linear support vector machines. *arXiv preprint arXiv:1306.0239*.

Tian, Y.-I., Kanade, T., & Cohn, J. F. (2001). Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence*, *23*(2), 97–115.

Tian, Y., Wang, X., Wu, J., Wang, R., & Yang, B. (2019). Multi-scale hierarchical residual network for dense captioning. *Journal of Artificial Intelligence Research*, *64*, 181–196.

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *NAACL HLT*, pp. 173–180. Association for Computational Linguistics.

Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). CIDEr: Consensus-based image description evaluation. In *CVPR*, pp. 4566–4575. IEEE.

Vinyals, O., & Le, Q. (2015). A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *CVPR*, pp. 3156–3164. IEEE.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, *8*(3-4), 229–256.

Wilson, A. C., Roelofs, R., Stern, M., Srebro, N., & Recht, B. (2017). The marginal value of adaptive gradient methods in machine learning. In *NIPS*, pp. 4151–4161.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pp. 2048–2057.

Yin, L., Wei, X., Sun, Y., Wang, J., & Rosato, M. J. (2006). A 3d facial expression database for facial behavior research. In *7th international conference on automatic face and gesture recognition (FGR06)*, pp. 211–216. IEEE.

You, Q., Jin, H., & Luo, J. (2018). Image captioning at will: A versatile scheme for effectively injecting sentiments into image descriptions. *arXiv preprint arXiv:1801.10121*.

You, Q., Jin, H., Wang, Z., Fang, C., & Luo, J. (2016a). Image captioning with semantic attention. In *CVPR*, pp. 4651–4659. IEEE.

You, Q., Luo, J., Jin, H., & Yang, J. (2016b). Building a Large Scale Dataset for Image Emotion Recognition: The Fine Print and The Benchmark. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pp. 308–314.

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, Y., Ko, H., Choi, J., & Kim, G. (2017). End-to-end concept word detection for video captioning, retrieval, and question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3261–3269. IEEE.

Yu, Z., & Zhang, C. (2015). Image based static facial expression recognition with multiple deep network learning. In *ICMI*, pp. 435–442. ACM.

Zeng, N., Zhang, H., Song, B., Liu, W., Li, Y., & Dobaie, A. M. (2018). Facial expression recognition via learning deep sparse autoencoders. *Neurocomputing*, *273*, 643–649.

Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2008). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, *31*(1), 39–58.

Zhang, Z., Luo, P., Loy, C.-C., & Tang, X. (2015). Learning social relation traits from face images. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3631–3639.