Journal of Artificial Intelligence Research 60 (2017) 1003-1030 Submitted 8/17; published 12/17

Learning Neural Audio Embeddings for Grounding Semantics in Auditory Perception

Douwe Kiela DKIELA@FB.COM
Facebook Artificial Intelligence Research
770 Broadway, New York, NY 10003, USA

Stephen Clark STEPHEN.CLARK@CL.CAM.AC.UK
Computer Laboratory, University of Cambridge
15 JJ Thomson Avenue, Cambridge CB3 0FD, UK

Abstract

Multi-modal semantics, which aims to ground semantic representations in perception, has relied on feature norms or raw image data for perceptual input. In this paper we examine grounding semantic representations in raw *auditory* data, using standard evaluations for multi-modal semantics. After having shown the quality of such auditorily grounded representations, we show how they can be applied to tasks where auditory perception is relevant, including two unsupervised categorization experiments, and provide further analysis. We find that features transfered from deep neural networks outperform bag of audio words approaches. To our knowledge, this is the first work to construct multi-modal models from a combination of textual information and auditory information extracted from deep neural networks, and the first work to evaluate the performance of tri-modal (textual, visual and auditory) semantic models.

1. Introduction

Distributional models (Turney & Pantel, 2010; Clark, 2015) have proved useful for a variety of core artificial intelligence tasks that revolve around natural language understanding. The fact that such models represent the meaning of a word as a distribution over other words, however, implies that they suffer from the *grounding problem* (Harnad, 1990); i.e. they do not account for the fact that human semantic knowledge is grounded in the perceptual system (Louwerse, 2008). Motivated by human concept acquisition, the field of multi-modal semantics enhances linguistic or textual representations with extra-linguistic perceptual input. Vision-based multi-modal semantic models have become increasingly popular over recent years, and have been shown to outperform language-only models on a range of tasks, including modeling semantic similarity and relatedness (Silberer & Lapata, 2012; Bruni, Tran, & Baroni, 2014), lexical entailment (Kiela, Rimell, Vulić, & Clark, 2015a), predicting compositionality (Roller & Schulte im Walde, 2013), bilingual lexicon induction (Kiela, Vulić, & Clark, 2015b) and metaphor identification (Shutova, Kiela, & Maillard, 2016). In fact, although surrogates of human semantic knowledge (i.e., feature norms elicited from human subjects) have also been used, raw image data has become the *de facto* perceptual modality in which to ground multi-modal models. See the review of Baroni (2016) for an excellent overview of visually grounded multi-modal models.

If the objective is to ground semantic representations in perceptual information, though, why stop at image data? The meaning of *violin* is surely not only grounded in its visual properties, such as shape, color and texture, but also in its sound, pitch and timbre. To understand how perceptual input leads to conceptual representation, we should cover as many perceptual modalities as possible. Recent preliminary studies have found that it is possible to derive semantic representations from sound data and that these representations can effectively be used in multi-modal models (Lopopolo & van Miltenburg, 2015; Kiela & Clark, 2015). Inspired by the "bag of visual words" (BoVW) (Sivic & Zisserman, 2003) approach in vision-based multi-modal semantics, these works make use of the so-called "bag of audio words" (BoAW) algorithm to obtain auditory-grounded representations.

In this work, we extend these preliminary results in various new directions. First, we introduce a deep convolutional neural network model for learning auditorily-grounded representations, called "neural audio embeddings" (NAE), and compare it to BoAW. We explore the transferability of NAE representations when trained on different types of data—either data from a narrow set of categories, such as musical instruments, or a broader set, such as naturally occurring environmental sounds (Section 5.1). We also provide a qualitative analysis (Section 5.2) and examine whether pre-training the network architecture on a much larger dataset improves performance (Section 5.3). Second, we show that the learned representations can be fused with other modalities, and examine two well-known deterministic multi-modal fusion methods (Section 5.4). We show that, in both cases, auditorily-grounded representations outperform text-only representations on the well-known MEN similarity and relatedness benchmark. In addition, we show that a tri-modal model, that incorporates textual, visual and auditory information, works even better. Finally, we show that the learned representations are valuable in downstream tasks that rely on auditory information, in this case unsupervised categorization (Section 5.5).

To our knowledge, this is the first work to construct multi-modal models from a combination of textual information and auditory information extracted from deep neural networks, and the first work to evaluate the performance of tri-modal (textual, visual and auditory) semantic models.

2. Related Work

Information processing in the brain can be roughly described to occur on three levels: perceptual input, conceptual representation and symbolic reasoning (Gazzaniga, 1995). Modeling the latter has a long history in AI and sprang from its "good old fashioned" roots (Haugeland, 1985), while the former has been advanced greatly through the application of pattern recognition to perceptual input (e.g. LeCun, Bengio, & Hinton, 2015). Understanding the middle level is arguably more of an open problem (Bengio, Courville, & Vincent, 2013): how is it that perceptual input leads to conceptual representations that can be processed and reasoned with? A key observation is that conceptual representations are, through perception, *grounded* in physical reality and sensorimotor experience (Harnad, 1990; Louwerse, 2008). There has been a surge of recent work on perceptually grounded semantic models that try to account for this fact, which have outperformed state-of-the-art text-based methods on a variety of natural language processing tasks.

2.1 Perceptual Grounding

Perceptually grounded models learn semantic representations from both textual and perceptual input. One method for obtaining perceptual representations is to rely on direct human semantic knowledge, in the shape of feature or association norms (e.g. Nelson, McEvoy, , & Schreiber, 2004; McRae, Cree, Seidenberg, & McNorgan, 2005), which have been used successfully in a range of multi-modal models (Silberer & Lapata, 2012; Roller & Schulte im Walde, 2013; Hill & Korhonen, 2014; Bulat, Kiela, & Clark, 2016). However, norms are elicited from human annotators and as a consequence are limited in coverage and relatively expensive to obtain. An alternative approach, that does not suffer from these limitations, is to make use of raw data as the source of perceptual information (Feng & Lapata, 2010; Leong & Mihalcea, 2011; Bruni et al., 2014). Raw data, for instance in the form of images, is cheap, plentiful, easy to obtain and has much better coverage (Baroni, 2016).

A popular approach has been to collect images associated with a concept, and then lay out each image as a set of keypoints on a dense grid, where each keypoint is represented by a robust local feature descriptor such as SIFT (Lowe, 2004). These local descriptors are subsequently clustered into a set of "visual words" using a standard clustering algorithm such as k-means and then quantized into vector representations by comparing the descriptors with the centroids. Kiela and Bottou (2014) introduced a more sophisticated approach that obtains much better visual representations (i.e. they perform much better on similarity and relatedness datasets) by transfering features from convolutional neural networks that were pre-trained on object recognition and aggregating these into a single concept representation. Various simple ways of aggregating image representations into visual representations for a concept have been proposed, such as taking the mean or the elementwise maximum of the individual image representations.

Ideally, one would jointly learn multi-modal representations from parallel multi-modal data, such as text containing images (Feng & Lapata, 2010) or images described with speech (Synnaeve, Versteegh, & Dupoux, 2014), but such data is hard to obtain, has limited coverage and can be noisy. Hence, image representations are often learned independently. Aggregated visual representations are subsequently combined with a traditional linguistic distributional space to form a multi-modal model. Mixing can be done in a variety of ways, ranging from simple concatenation to more sophisticated fusion methods (Bruni et al., 2014).

2.2 Auditory Representations

As this work intends to show, the source of perceptual input need not to be limited to the visual domain. Recent work in multi-modal semantics has started to go beyond vision as the single source of raw perceptual input, with preliminary investigations of auditory and even olfactory representations (Lopopolo & van Miltenburg, 2015; Kiela & Clark, 2015; Kiela, Bulat, & Clark, 2015). Auditory grounding was achieved by dividing sound files into frames, clustering these as "audio words" and subsequently quantizing them into representations by comparing frame descriptors with the centroids. More recently, Vijayakumar, Vedantam, and Parikh (2017) proposed an embedding scheme that learns specialized word embeddings grounded in sounds, using a variety of audio features. These techniques were found to work well for modeling human similarity and relatedness judgments and related

Dataset	MEN	AMEN
Textual	3000	258
Auditory	2590	233

Table 1: Number of concept pairs for which representations are available in each modality.

experiments. Here, we build on that work, using deep learning models that lead to auditory representations of a higher quality.

Various methods for general-purpose (i.e., as opposed to feature-engineered) auditory represention learning have been proposed in the literature. In a similar fashion to bag-of-words in computational linguistics and bag-of-visual-words in computer vision, bag-of-audio-words has successfully been applied in a variety of tasks, including e.g. event classification (Pancoast & Akbacak, 2012), audio document retrieval (Chechik, Ie, Rehn, Bengio, & Lyon, 2008), copy detection (Uchida, Sakazawa, Agrawal, & Akbacak, 2010) and emotion in speech classification (Schmitt, Ringeval, & Schuller, 2016). Recently there has been more interest in applying deep learning methods to auditory signal processing (Dahl, Yu, Deng, & Acero, 2012), both in end-to-end systems (Trigeorgis, Ringeval, Brueckner, Marchi, Nicolaou, Zafeiriou, et al., 2016) and with an explicit focus on representation learning (Hamel & Eck, 2010). Such methods have been successful in a variety of tasks, including music recommendation (Van den Oord, Dieleman, & Schrauwen, 2013), classification (Dieleman, Brakel, & Schrauwen, 2011), annotation (Hamel, Lemieux, Bengio, & Eck, 2011) and retrieval (Weston, Bengio, & Hamel, 2011). In work that is most related to this article, auditory feature extraction using deep learning was successfully applied to musical audio analysis (Dieleman, 2016). Here, we compare BoAW representations with convolutional neural network-derived features on tasks in semantics and show how NAE-grounded representations work better than the alternatives.

3. Evaluation

We evaluate on a standard similarity and relatedness dataset: the MEN test collection (Bruni et al., 2014). This dataset consists of concept pairs together with a human-annotated relatedness score. Relatedness here means that it assigns high scores to pairs such as *teacher-instructor* and *teacher-student* but low scores to unrelated pairs such as *jellyfish-bakery*. The human-assigned judgments were obtained by crowdsourcing using Amazon Mechanical Turk, only accepting English native speakers as annotators. The full dataset consists of 3,000 word pairs, randomly selected from words that occur at least 700 times in the Wackypedia corpora (Baroni, Bernardini, Ferraresi, & Zanchetta, 2009) and were used as tags for at least 50 times in the ESP game dataset (Von Ahn & Dabbish, 2004). Inter-annotator agreement was not calculated over all annotators, but was gauged separately by having two annotators annotate the full dataset and examining their agreement[1]: the Spearman correlation between their judgments was 0.68, while the correlation of their average ratings

1. https://staff.fnwi.uva.nl/e.bruni/MEN

MEN	human rating	relevant
automobile-car	1.00	✓
rain-storm	0.98	✓
sun-sunshine	0.94	
bird-eagle	0.88	✓
guitar-piano	0.86	✓
colour-red	0.82	
foliage-tulip	0.64	
dessert-orange	0.5	
hawk-insects	0.42	✓
frozen-shop	0.26	
monkeys-restaurant	0.1	✓
cheetah-phone	0.06	✓

Table 2: Illustrative examples of pairs in the datasets where auditory information is or is not relevant, together with their corresponding similarity rating as provided by human annotators.

with the MEN scores is at 0.84. These numbers serve as an indication of the upper bound on the dataset.

Evidence suggests that the inclusion of visual representations only improves performance for certain concepts, and that in some cases the introduction of visual information is detrimental to performance on similarity and relatedness tasks (Kiela, Hill, Korhonen, & Clark, 2014). The same is likely to be true for other perceptual modalities: in the case of comparisons such as *guitar-piano*, the auditory modality is certainly meaningful, whereas in the case of *democracy-anarchism* it is probably less so. This is even more likely to be the case for less dominant modalities such as auditory perception.

Therefore, we had two graduate students annotate the MEN dataset according to whether auditory perception is relevant to the pairwise comparison. The annotation criterion was as follows: if both concepts in a pairwise comparison have a distinctive associated sound, the modality is deemed relevant. Inter-annotator agreement was high, with $\kappa = 0.93$. Some examples of relevant pairs can be found in Table 2. Hence, we now have two evaluation datasets for conceptual similarity and relatedness: the MEN test collection **MEN**, and its auditory-relevant subset **AMEN**. Due to the nature of the auditory data sources, it is not possible to build auditory representations for all concepts in the test sets. Hence, we only evaluate on the covered subsets to ensure a fair comparison, that is, we only use the comparisons that have coverage in both the textual and auditory modalities. Table 1 shows how much of the test sets are covered for each modality.

4. Approach

One reason for using raw image data in multi-modal models is that there are many high quality resources available that contain tagged images, such as ImageNet (Deng, Dong, Socher, Li, Li, & Fei-Fei, 2009) and the ESP Game dataset (Von Ahn & Dabbish, 2004). Such resources do not exist for audio files, so instead we use the online search engine Freesound[2] (Font, Roma, & Serra, 2013) to obtain audio files. Freesound is a collaborative database released under Creative Commons licenses, in the form of snippets, samples and recordings, that is aimed at sound artists. The Freesound API allows users to easily search for audio files that have been tagged using certain keywords. For each of the concepts in the evaluation datasets, we used the Freesound API to obtain samples encoded in the standard open source OGG format[3]. The Freesound API allows for various degrees of keyword matching: we opted for the strictest keyword matching, in that the audio file needs to have been purposely tagged with the given word (the alternative includes searching the text description for matching keywords).

4.1 Auditory Representations

We experiment with two methods for obtaining auditory representations: bag-of-audio words and transfering a layer from a trained convolutional neural network. The former is a relatively simple approach that does not take into account any interdependencies between local feature descriptors, whereas the latter is more sophisticated and able to extract more elaborate patterns and interactions. While these methods vary significantly, in that they use different input features (local feature descriptors of frames versus spectrograms), their representations are constructed from the same sound files, allowing us to compare the two methods.

4.1.1 Bag of Audio Words (BoAW)

A common approach to obtaining acoustic features of audio files is the Mel-scale Frequency Cepstral Coefficient (MFCC) (O'Shaughnessy, 1987). MFCC features are abundant in a variety of applications in audio signal processing, ranging from audio information retrieval, to speech and speaker recognition, and music analysis (Eronen, 2003). Such features are derived from the mel-frequency cepstrum representation of an audio fragment (Stevens, Volkmann, & Newman, 1937). In MFCC, frequency bands are spaced along the mel scale, which has the advantage that it approximates human auditory perception more closely than e.g. linearly-spaced frequency bands. Hence, MFCC takes human perceptual sensitivity to audio frequencies into consideration, which makes it suitable for e.g. compression and recognition tasks, but also for our current objective of modeling auditory perception.

After having obtained MFCC descriptors, we cluster them using mini-batch k-means (Sculley, 2010) and quantize the descriptors into a "bag of audio words" (BoAW) (Foote, 1997) representation by comparing the MFCC descriptors to the cluster centroids. We set $k = 300$ — a number which has been found to work well for such representations (see, e.g. Kiela & Clark, 2015) — but do not apply any additional weighting. See Figure 1 for an

2. http://www.freesound.org.
3. http://www.vorbis.com.

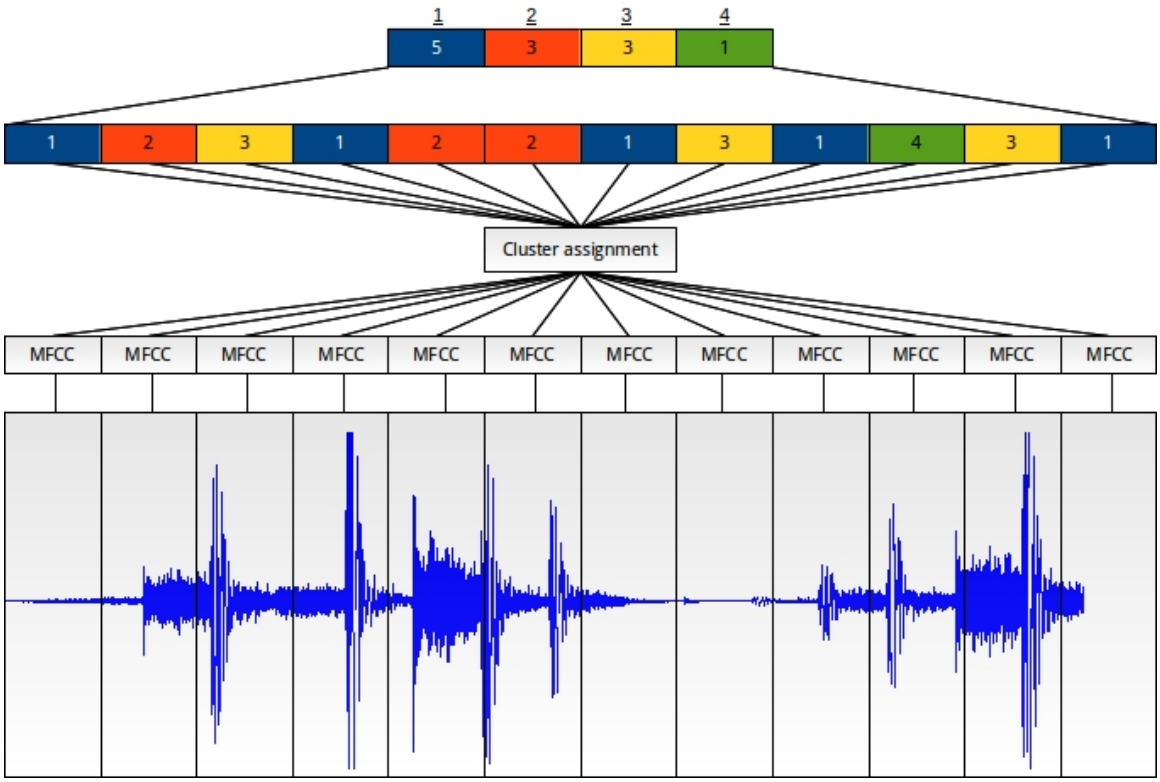

Figure 1: Illustration of the BoAW method. Each of the MFCC descriptors is assigned to a cluster. Assignments are subsequently quantized into a bag of audio words representation. In this illustration, $k = 4$ in k-means, which means there are four clusters and the value for each of the k clusters is the number of datapoints belonging to it. The colors in the diagram reflect the different clusters: for instance, cluster 1 (color-coded in blue here) occurs 5 times in this case.

illustration of the process for a single audio file. Auditory representations for a concept are obtained by taking the mean of the BoAW representations of the relevant audio files.

4.1.2 Neural Auditory Embeddings (NAE)

The work of Kiela and Bottou (2014) showed that it is possible to transfer and aggregate convolutional neural network layers in order to obtain a visual semantic representation. Their network was an adaptation of the well-known AlexNet (Krizhevsky, Sutskever, & Hinton, 2012), trained on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). Image representations were obtained by extracting the penultimate, pre-softmax layer from the network for each image, which were then aggregated into visual representations by taking the mean or pointwise maximum of the image representations. They found that such CNN-derived visual representations perform much better than traditional bag-of-visual-words-based ones, with substantial increases in correlation with human similarity and relatedness ratings.

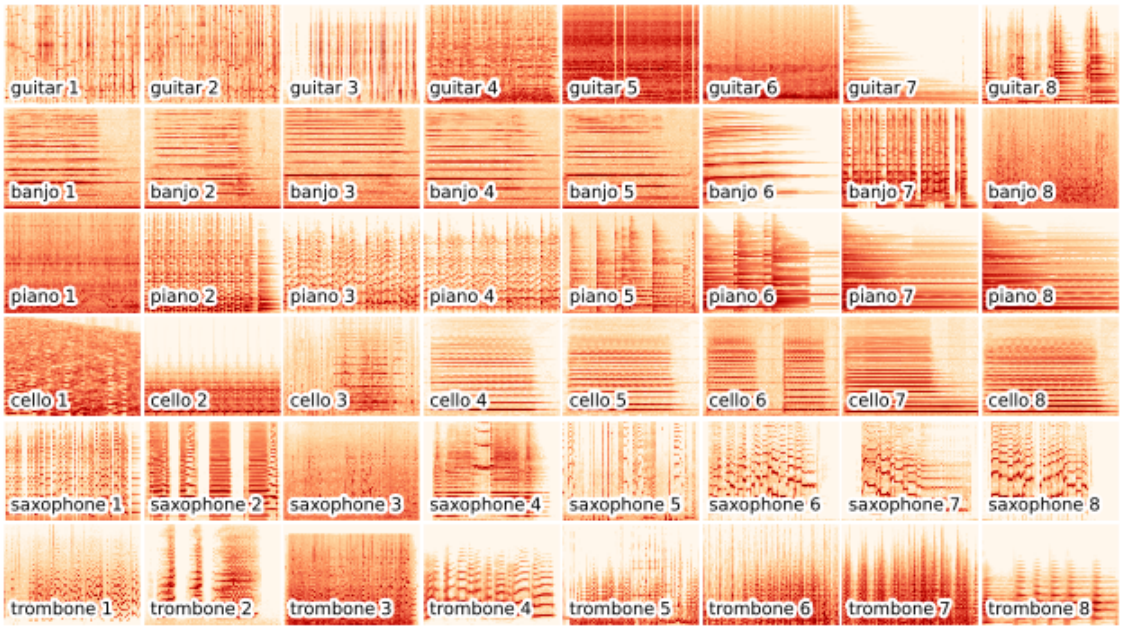

Figure 2: Examples of spectrograms, plotted for various musical instruments.

Here, we examine whether a similar methodology can be applied to auditory representations. The advantage of using a convolutional neural network instead of a recurrent one (RNN) is that it requires less memory and suffers less from the vanishing or exploding gradients problem (Pascanu, Mikolov, & Bengio, 2013). RNNs are especially susceptible to this problem for the current study, given that sound files can vary considerably in the number of frames (i.e., their duration in milliseconds), which means padding mini-batches is cumbersome and training is difficult. We obtain sound file representations by transfering the pre-softmax layer from a convolutional neural network trained on audio classification. We use a standard AlexNet architecture (Krizhevsky et al., 2012), without modifying the network to fit the auditory signal better (e.g. by adjusting the channels or color filtering), to make reproduction of our results as easy as possible. The neural auditory embedding approach can be summarized to comprise the following steps:

1. (train step) Train a neural network classifier $\mathcal{C}$ on the dataset $\{\langle f(s), L_s\rangle \mid s \in S_C\}$, where S_C is a set of audio files, f is a pre-processing function and L_s is the label for that file.

2. (transfer step) For each label L_x (where L_x is not necessarily also a label in S_C, but may be):

   (a) Retrieve a set of audio files S_x

   (b) For each file $s \in S_x$:

      i. Obtain the auditory representation $\mathbf{q}_s = g(f(s))$, where g is the neural network $\mathcal{C}$ up to the penultimate pre-softmax layer.

Figure 3: Illustration of the Neural Auditory Embedding method, using a convolutional neural network. The auditory signal is converted to a spectrogram which is fed to the neural network for classification. The pre-softmax layer, FC7, is transferred and taken as the neural audio embedding (NAE) for the given sound file.

(c) (aggregation) The overall representation for label L_x is then obtained by aggregating the per-file representations, that is, we take the mean of the relevant auditory representations, i.e., $\mathbf{r}_x = \frac{1}{|S_x|} \sum_{s_i \in S_x} \mathbf{q}_{s_i}$.

We use a Mel-scale spectrogram (Flanagan, 2013) of the sound file as the input to the network, i.e., the input sound file is converted (f in the algorithm above) into a three-dimensional representation of the spectrum of frequencies as they vary with time. A spectrogram can be interpreted as a visual rendering of an auditory signal, which means that we can apply a similar network architecture to deep neural networks used in computer vision, for classifying auditory patterns. Figure 2 shows how simple visual inspection already reveals some clear patterns for certain musical instruments, which convolutional networks are well-equipped to exploit.

Our architecture is as follows: the network consists of 5 convolutional layers, followed by two fully connected rectified linear unit (ReLU) layers that feed into a softmax for classification (Krizhevsky et al., 2012). The network learns through a multinomial logistic regression objective:

$$J(\theta) = -\sum_{i=1}^{D}\sum_{k=1}^{K} \mathbf{1}\{y^{(i)} = k\} \log \frac{\exp(\theta^{(k)\top} x^{(i)})}{\sum_{j=1}^{K} \exp(\theta^{(j)\top} x^{(i)})} \quad (1)$$

where $\mathbf{1}\{\cdot\}$ is the indicator function, $x^{(i)}$ is the input and D examples with K classes are used for training. We obtain audio embeddings by performing a forward pass with a given spectrogram and taking the 4096-dimensional fully connected layer that precedes the softmax (called FC7) as the representation of that sound file (see Figure 3).

We experiment with training the network on either a narrow dataset of musical instruments, or a broad dataset of naturally occurring environmental sounds; so one model has to be good at fine-grained distinctions between similar sounds (e.g., distinguishing between a mandolin, a ukelele and a banjo), while the other needs to be able to recognize general sound categories that can vary substantially in their audio signatures (e.g. distinguishing scissors from cows and airplanes). We use standard stochastic gradient descent (SGD) optimization, with an initial learning rate of 0.01. The learning rate was set to degrade in a stepwise fashion by a factor of 0.1 every 1000 iterations, until convergence.

Instruments Classifier We obtain up to 1000 sound files for a set of 54 musical instruments, yielding a total of 25324 sound files. We divide the data into a training and a

accordion	bagpipe	balalaika	banjo	baritone (sax.)
bass	bassoon	bell	bongo	bugle
carillon	castanets	celeste	cello	chimes
clarinet	claves	clavichord	clavier	conga
cornet	cowbell	cymbals	didgeridoo	drum
fiddle	flute	glockenspiel	gong	guitar
harmonica	harp	harpsichord	horn	keyboard
lute	lyre	mandolin	maracas	marimba
oboe	organ	piano	piccolo	saxophone
sitar	tambourine	trombone	trumpet	tuba
ukulele	violin	xylophone	zither	

Table 3: Labels for the musical instruments classifier.

validation set, sampling 75% for the former and taking the remainder for the latter. Using the training methodology described above, we obtain a classification accuracy of 92% on the validation set. The instruments are listed in Table 3. Embeddings transferred from this classifier are referred to as NAE-INST in what follows.

Environmental Sounds Classifier Gygi, Kidd, and Watson (2007) performed an extensive psychological study of auditory perception and its relation to environmental sound categories. We obtain up to 2000 sound files for the 50 classes used in their acoustic similarity and categorization experiments, which results in 31432 sound files. The classifier achieves 54% accuracy on the validation set. This number is substantially lower than the instruments classifier, which indicates that it is a significantly harder problem. The environmental labels (see Table 4) are much more varied and it is likely that FreeSound returns noisier sound files for these categories. Ultimately, we are less interested in the performance of the trained classifier, but more in the quality of the representations that can be extracted from that classifier, in order to use them for downstream tasks or applications. The set of labels has been specifically designed with the similarity and categorization of human auditory perception in mind, and hence it spans a wide range of sound categories and arguably reflects human auditory perception better than the instruments dataset. Embeddings transferred from this classifier are referred to as NAE-ENV.

4.1.3 Pre-training

One of the reasons behind the success of convolutional neural networks in computer vision is that they can be trained on millions of images. This allows for the lower layers of the network to become very good "edge detectors", and to become more specific to the final classification decision in higher layers, as shown by Zeiler and Fergus (2014). Since there are fewer sound files available, we additionally experiment with applying a transfer

airplane	axe	baby	basketball	bells
bird	bowling	bubbling	car accelerating	car start
cat	claps	clock	cough	cow
cymbals	dog	door	drums	footsteps
gallop	glass break	gun	harp	helicopter
honking	ice drop	keyboard	laugh	match
neigh	phone	ping pong	rain	rooster
saw	scissors	sheep	siren	sneeze
splash	thunder	toilet	train	typewriter
water	wave	whistle	wipers	zipper

Table 4: Labels for the environmental sound classifier, from Gygi et al. (2007).

learning technique, where we "finetune" a network that has already been trained on ILSVRC 2012. This means we can rely on the network to already perform well at recognizing visual patterns. In particular, we set the learning rate to a small number for the first five layers, and learn the fully connected weights that lead to the new softmax with different labels from scratch with a higher learning rate. This allows the use of edge-detectors that were trained on a massive dataset of images, but enables the fine-tuning of parameters for the particular task at hand, in this case the classification of auditory signals as represented by spectrograms. In this case, we set an initial learning rate of 0.01 for the fully connected layers and 0.001 for the earlier convolutional layers and learn for up to 4000 iterations using SGD. The learning rate was set to degrade in a stepwise fashion as above.

4.1.4 Duration and Number

The method for obtaining the auditory representations for the conceptual similarity and relatedness evaluations is as follows: For each word, we retrieve the first 100 sound samples from FreeSound with a maximum duration of 1 minute. The rationale behind this decision is that the duration of FreeSound samples varies significantly, with samples as short as one second and as long as half an hour. The same sound files are used as input in all models when extracting representations, to ensure direct comparability.

4.2 Textual Representations

We compare against textual representations, and combine auditory representations with textual representations to obtain multi-modal representations. For the textual representations we use the continuous vector representations from the log-linear skip-gram model of Mikolov, Chen, Corrado, and Dean (2013). Specifically, 300-dimensional vector representations were obtained by training on a dump of the English Wikipedia plus newswire

(8 billion words in total).[4] We train a skip-gram model for a sequence of training words $w_1, w_2, ..., w_T$ and a context size c by maximizing:

$$J = \frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0} \log \frac{\exp^{u_{w_{t+j}}^{\top} v_{w_t}}}{\sum_{w'\in W} \exp^{u_{w'}^{\top} v_{w_t}}}$$

where u_w and v_w are the context and target vector representations for the word w respectively, and W is the vocabulary. These types of representations have been found to yield very good performance on a variety of semantic similarity tasks (Baroni, Dinu, & Kruszewski, 2014).

4.3 Multi-modal Fusion Strategies

Since multi-modal semantics relies on two or more modalities, there are several ways of combining or *fusing* linguistic and perceptual cues (Bruni et al., 2014). When computing similarity scores, for instance, we can either jointly learn the representations (e.g. as in Lazaridou, Pham, & Baroni, 2015); learn them independently, combine (e.g. concatenate) them and compute similarity scores; or learn them independently, compute similarity scores independently and combine the scores. These possibilities have been called *early*, *middle* and *late* fusion, respectively. In this work, we restrict ourselves to middle and late fusion, since these options are relatively easy to compute, are less susceptible to noise (primarily because they do not force the model to learn to represent and to fuse at the same time) and have fewer hyperparameters than early fusion methods.

4.3.1 Middle Fusion

Whereas early fusion requires a joint training objective that takes into account both modalities, middle fusion allows for individual training objectives. Similarity between two multi-modal representations is calculated as follows:

$$sim(u, v) = g(f(u^l, u^a), f(v^l, v^a))$$

where g is some similarity function, u^l and v^l are textual representations, and u^a and v^a are the auditory representations. We call this model MM-MIDDLE.

A typical formulation in multi-modal semantics for $f(x, y)$ is $\alpha\, x \| (1-\alpha) y$, where $\|$ is concatenation (see, e.g. Bruni et al., 2014; Kiela & Bottou, 2014). The α parameter is a global parameter that governs how much of a given modality gets incorporated into the combined multi-modal representation. In what follows, we first keep α fixed, and then experiment with tuning it on a held-out development set.

4.3.2 Late Fusion

Late fusion can be seen as the converse of middle fusion, in that the similarity function is computed first before the similarity scores are combined:

$$sim(u, v) = h(g(u^l, v^l), g(u^a, v^a))$$

4. The demo-train-big-model-v1.sh script from https://code.google.com/archive/p/word2vec was used to obtain this corpus.

where g is some similarity function and h is a way of combining similarities, in our case a weighted arithmetic average: $h(x, y) = \alpha\, x + (1 - \alpha)\, y$; and we use $g = \frac{x \cdot y}{|x||y|}$ (cosine similarity). We call this model MM-LATE.

4.3.3 Tri-modal Fusion

Multi-modal fusion need not be limited to just two modalities. We also experiment with combining three modalities, i.e., combining textual, visual and auditory information. In that case we introduce an additional mixing parameter β that determines the contribution of the second modality. Applying that to the weighting functions f (for middle fusion) and h (for late fusion) above, we get $f(x, y, z) = \alpha\, x \parallel (1 - \alpha + \beta)\, y \parallel (1 - \alpha - \beta)\, z$ and $h(x, y, z) = \alpha\, x + (1 - \alpha + \beta)\, y + (1 - \alpha - \beta)\, z$, where $(\alpha + \beta \leq 1)$.

5. Results

We evaluate auditory representation quality by calculating the Spearman ρ_s correlation between the ranking of the concept pairs produced by the automatic similarity metric (cosine between the derived vectors) and that produced by the gold-standard similarity scores, the standard metric for these types of evaluations. The two main questions we aim to examine are: do NAEs outperform BoAW representations; and do auditorily-grounded multi-modal representations perform better than text-only. In addition, we examine auditorily-grounded representations qualitatively, experiment with pre-training, analyze fusion methods and evaluate on an unsupervised categorization task.

5.1 Representational Quality

The results are reported in Table 5, according to whether they are (a) uni-modal representations obtained from a single modality or (b) multi-modal representations that have undergone multi-modal fusion. Because cosine similarity is the normalized dot-product, and the uni-modal representations are themselves normalized, middle and late fusion are equivalent if we take the unweighted average (i.e., $\alpha = 0.5$, or $\alpha = \beta = \frac{1}{3}$ in the tri-modal case). Given this equivalence, we report one set of results in this experiment and omit whether they use middle or late fusion. An examination of the type of fusion and associated mixing parameters is provided in a later section.

In the uni-modal case, we compare auditory representations against those obtained from textual and visual[5] sources. We find that both visual and textual representations outperform auditory ones on the entire dataset (as we might have expected, given that audio is likely not to be the most dominant modality in this dataset). On the auditory-relevant subset AMEN, however, we observe that NAE-ENV representations outperform the visual model, though by a small margin. Importantly, NAE-ENV performs significantly better than the auditory BOAW alternative on both MEN ($z = 3.51, p < 0.001$) and AMEN ($z = 1.85, p < 0.05$). Embeddings extracted from the broad environmental sounds classifier (NAE-ENV)

5. Obtained by downloading 10 images from Google for each of the words, transferring the pre-softmax FC7 layer and taking the mean of the image representations to obtain an overall visual representation, as described in e.g. the work of Kiela and Bottou (2014).

(a): Uni-modal

Model	MEN	AMEN
RANDOM	0.02	0.16
TEXT	0.69	0.59
VISUAL	0.52	0.55
BOAW	0.23	0.43
NAE-INST	0.27	0.46
NAE-ENV	0.32	0.56

(b): Multi-modal

Model	MEN	AMEN
TEXT+VISUAL	0.72	0.64
TEXT+BOAW	0.62	0.64
TEXT+NAE-INST	0.61	0.64
TEXT+NAE-ENV	0.63	0.67
VISUAL+BOAW	0.50	0.63
VISUAL+NAE-INST	0.49	0.62
VISUAL+NAE-ENV	0.53	0.70
TEXT+VISUAL+BOAW	0.67	0.68
TEXT+VISUAL+NAE-INST	0.67	0.67
TEXT+VISUAL+NAE-ENV	0.68	0.70

Table 5: Spearman ρ_s correlation of (a) uni-modal representations and (b) multi-modal representations with untuned mixing parameters (in which case middle and late fusion are equivalent and modalities contribute equally). See Table 8 for results with tuned mixing parameters.

significantly outperform ($z = 1.45, p < 0.1$) embeddings extracted from the narrow musical instruments classifier (NAE-INST), which does only slightly better than BOAW.

Focusing on the multi-modal results, we see a large increase in performance over uni-modal representations, including over the textual model, for AMEN. The same effect is not observed for MEN, which is understandable given how few pairwise comparisons are auditory-relevant. In many cases we are still able to obtain sound files, but these tend to be of poor quality and lead to noisy representations (e.g., what does "sunlight" sound like?). As the results indicate, visual information is probably more useful in those cases. On the auditory-relevant subset, however, we see that the best performing auditorily-grounded multi-modal model TEXT+NAE-ENV performs significantly better ($z = 1.46, p < 0.1$) than TEXT. Although the AMEN dataset has been tagged with auditory relevance in mind, many of the selected comparisons (e.g. *cat-kittens* or *car-automobile*) are still dominated by visual or linguistic information, which means that the auditory representations must be of a high quality if they are able to mirror the human judgments, even in the uni-modal case. When we combine the modalities into a tri-modal model that incorporates textual, visual and auditory information, we see an even larger improvement over TEXT, with a highest Spearman correlation of 0.70 ($z = 2.09, p < 0.1$).

5.2 Qualitative Analysis

We performed a small qualitative analysis of the auditory representations for the words in the MEN dataset. As Table 6 shows, the nearest neighbors are remarkably semantically

TEXT							
engine	monster	children	dinner	splash	weather	birds	dawn
gasoline	zombie	kids	lunch	bucket	rain	mammals	dusk
vehicle	dragon	girls	wedding	skateboard	storm	animals	sunrise
airplane	creatures	women	breakfast	ink	fog	rodents	moon
aircraft	clown	people	cocktail	cocktail	cold	reptiles	night
motor	dog	boys	holiday	dripping	tropical	amphibians	misty
BOAW							
engine	monster	children	dinner	splash	weather	birds	dawn
motor	dead	female	eat	wet	storm	fabric	garden
car	zombie	cow	food	run	cold	summer	summer
storm	guitar	kids	tiles	lake	winter	forest	pond
drive	ship	animals	breakfast	wave	ford	village	parrot
automobile	dark	lady	floor	sea	building	food	birds
NAE-INST							
engine	monster	children	dinner	splash	weather	birds	dawn
cold	zombie	farm	school	wet	storm	forest	tropical
car	dead	sheep	kitchen	river	flag	summer	parrot
automobile	dark	animals	morning	lake	interior	nature	zoo
motor	ship	cow	home	run	car	morning	morning
vehicle	lion	party	coffee	dripping	aircraft	garden	birds
NAE-ENV							
engine	monster	children	dinner	splash	weather	birds	dawn
motor	zombie	protest	lunch	wet	storm	morning	tropical
automobile	dead	kids	coffee	lake	wind	summer	zoo
drive	guy	party	bar	dripping	alley	forest	birds
vehicle	lion	happy	mug	run	rain	tropical	morning
car	man	women	rusty	river	ocean	zoo	dusk

Table 6: Example nearest neighbors in MEN for textual representations and auditory BoAW and NAE representations.

coherent. For example, the auditory models group together sounds produced by cars and engines. Nearest neighbors for the textual model tend to be of a more abstract nature: where we find *wet* and *lake* as auditory neighbors for *splash*, the textual model gives us concepts like *bucket*, which can make splashes but do not sound like them. While auditory neighbors of *dawn* are related to sounds one might hear at that time of day (e.g. morning birdsong), the textual model knows that *dawns* relate to *night*, *moon* and *sunrise*. We observe that neighbors of *birds* in the textual model are all other types of animals—i.e., categorically related—while the auditory neighbors are related in a much more associative manner.

Model	Uni-modal	MM-*
NAE-IMAGENET	0.42	0.62
NAE-INST	0.46	0.64
NAE-INST-PRETRAINED	0.49	0.65
NAE-ENV	0.56	0.67
NAE-ENV-PRETRAINED	0.56	0.67

Table 7: Spearman ρ_s correlation on AMEN for either only training on ImageNet, pre-training on ImageNet and then on the auditory dataset, or only on the auditory dataset.

5.3 Pre-training Effects

Since our network architecture is essentially the same as a regular convolutional neural network used in computer vision tasks, except with spectrograms as inputs, it is natural to ask whether pre-training the network—using e.g. ImageNet to learn so-called "edge detectors"—improves results. It might be, for instance, that this results in finding different optima, since the network is already adept at basic pattern recognition from the start of training on the auditory data. Table 7 shows the results on AMEN for fine-tuning the network as described in Section 4.1.3. First, we can see that training the network specifically on auditory recognition yields benefits over simple visual recognition: in all cases, the networks improve over the NAE-IMAGENET baseline, which consists simply of a pre-trained network trained on ILSVRC-12. It is interesting to observe that that model does quite well already. This shows that, without any training in spectrogram recognition, we are still able to extract some relevant features due to the pattern recognition capabilities that were developed on image recognition.

Pre-training helps in the case of instruments, but does not have an effect on NAE-ENV representations. A possible explanation might be that the instruments domain is too narrow for representing the relatedness of very diverse concepts, which is what MEN and AMEN measure. In other words, we appear to get some additional generalization capability from pre-training on a set of more diverse concepts. This appears to indicate that the (number and variety of) classes used in the classifier are highly relevant for the transferability and quality of the network's representations. Since the MEN dataset comprises a wide variety of concepts, one could argue that this observation makes sense, but it is nonetheless a question that needs further investigation.

5.4 Fusion Strategies and Multi-modal Mixing

Even though auditorily-grounded multi-modal models performed reasonably well on the full MEN dataset, they did not yet match the textual model (TEXT). It is natural to ask whether it is possible to construct models such that including perceptual input is not detrimental to the quality of representations that have no auditory relevance, or where it might even

Model	Middle fusion	Late fusion
TEXT	.68±.02	.68±.02
TEXT+VISUAL	.72±.02 α=.54±.05	.72±.02 α=.62±.04
TEXT+BOAW	.69±.02 α=.74±.08	.69±.01 α=.84±.05
TEXT+NAE-INST	.69±.02 α=.76±.08	.69±.02 α=.86±.05
TEXT+NAE-ENV	.70±.01 α=.70±.00	.70±.01 α=.80±.00
VISUAL+BOAW	.55±.01 α=.36±.05	.56±.01 α=.22±.04
VISUAL+NAE-INST	.55±.01 α=.34±.05	.55±.00 α=.22±.04
VISUAL+NAE-ENV	.57±.01 α=.40±.00	.57±.01 α=.30±.00
TEXT+VISUAL+BOAW	.73±.02 α=.62±.04, β=.14±.08	.73±.01 α=.70±.00, β=.14±.05
TEXT+VISUAL+NAE-INST	.73±.02 α=.62±.04, β=.16±.08	.73±.02 α=.70±.00, β=.20±.06
TEXT+VISUAL+NAE-ENV	.74±.01 α=.60±.00, β=.10±.00	.73±.01 α=.68±.04, β=.12±.04

Table 8: Cross-validated performance of middle and late multi-modal fusion models on the MEN dataset, when varying the α mixing parameter.

improve performance. The mixing parameter α in the middle and late fusion models can be used to govern the influence of a given modality on the overall representation. We kept it fixed at 0.5 for the models in Table 5, but it is possible to use a development set to obtain a more optimal weighting.

Hence, we do a five-way cross-validated comparison where we tune the α parameter (and in the tri-modal case also the β) on a held-out validation set of 20% of the data and obtain the Spearman ρ_s correlation score for the other 80%. Since correlations cannot be averaged directly, we average the Fisher-transform and take its inverse to obtain the average correlation, i.e., $\rho_s = \tanh(\frac{1}{N}\sum_{i=0}^{N}\text{arctanh}(\rho_s^i))$ where N is the number of splits. Table 8 reports the results. First, note that there are only minor differences between the two fusion methods, which is likely a consequence of the tuning of the mixing parameter, allowing us to select the optimal contribution-weight for each modality. All text-based multi-modal models outperform the text-only representations. If the textual information is omitted, and we only include visual and auditory information, performance drops. The results show that the inclusion of auditory information is not detrimental to performance when we select the mixing parameters in a more intelligent way: in fact, the TEXT+NAE-ENV model significantly ($z = 1.37, p < 0.1$) outperforms the TEXT model, and the tri-modal models do even better, with a maximum improvement of 0.06 over the TEXT model ($z = 4.36, p < 0.001$).

The results also shed light on the question of how much input from a given modality is most useful for predicting human similarity and relatedness ratings. That is, since the mixing parameters are tuned, the results give insight into the relative contribution of each modality. The results clearly show that textual information is the most important. The better the auditory representation, the more we would want to include of it, which explains the lower α for TEXT+NAE-ENV compared to the other two for both types of fusion, and the lower α and β in the tri-modal model that incorporates that type of auditory embedding. In

(a): Clustering instruments

Model	Mean	Max
TEXT	0.30 ± 0.06	0.42
BOAW	0.20 ± 0.05	0.37
NAE-INST	0.25 ± 0.07	0.42
MM-BOAW	0.32 ± 0.07	0.50
MM-NAE-INST	0.37 ± 0.07	0.54

(b): Clustering environmental sounds

Model	Mean	Max
TEXT	0.15 ± 0.08	0.34
BOAW	0.24 ± 0.13	0.55
NAE-ENV	0.25 ± 0.09	0.45
MM-BOAW	0.25 ± 0.11	0.62
MM-NAE-ENV	0.26 ± 0.09	0.59

Table 9: V-measure performance for clustering (a) musical instruments and (b) environmental sounds. Mean is over 100 runs of k-means.

the audio-visual case, visual information appears to be more informative of human similarity and relatedness ratings than auditory information.

5.5 Unsupervised Categorization with Neural Auditory Embeddings

Categorization is a fundamental problem faced by the human cognitive system, and one of the main focal points of investigation in psychology (Fountain & Lapata, 2011). While the preceding experiments were applied to general semantic similarity and relatedness, in this section we focus on two audio-specific tasks: musical instrument categorization for NAE-INST and environmental sound categorization for NAE-ENV.

We explore categorization using an unsupervised clustering algorithm over the learned representations. These experiments provide two contributions. First, they allow for another test of whether NAEs outperform BoAW representations and whether auditorily-grounded representations outperform text-only ones. Second, they shed light on the question of whether the unsupervised categorization mirrors human categorization judgments. If this is the case, it serves as a further corroboration of the value of multi-modal representations.

5.5.1 Musical Instrument Categorization

The set of instruments in Table 3 was manually divided into 5 classes, based on how Wikipedia classified them: brass, percussion, piano-based, string and woodwind instruments. For each of the instruments, as many audio files as available[6] were obtained from FreeSound. We then performed k-means clustering over the aggregated auditory representations with five cluster centroids and compared results between textual, bag of audio words and NAE representations. We experiment with the pre-trained NAE-INST embeddings that were specialized for musical instrument identification.

This is an interesting problem because instrument classes are determined somewhat by convention (is a *saxophone* a brass or a woodwind instrument?). What is more, how instruments sound is rarely described in detail in text, so corpus-based linguistic representations cannot take this information into account. Table 9(a) shows the mean and standard

6. We did not restrict the set of retrieved audio files to exclude files used for training the classifier.

	TEXT
1	piccolo
2	flute, lute, harpsichord, marimba, zither, harp, clavichord, sitar, didgeridoo, carillon, lyre, keyboard
3	harmonica, mandolin, banjo, guitar, accordion, ukulele, fiddle, bass
4	xylophone, tambourine, glockenspiel, claves, maracas, castanets, cymbals, celeste, horn, balalaika, clavier, cowbell, bongo, bugle, drum, conga, chimes, bell, gong
5	clarinet, trombone, bassoon, cello, saxophone, piano, violin, oboe, tuba, trumpet, cornet, baritone

	NAE-INST
1	accordion, balalaika
2	trombone, piano, cello, violin, saxophone, flute, banjo, oboe, tuba, mandolin, clarinet, harmonica, guitar, harpsichord, bassoon, cornet, trumpet, marimba, sitar, harp, lute, ukulele, zither, didgeridoo, clavichord, fiddle, horn, bugle, baritone, bass
3	xylophone, glockenspiel, celeste, claves, carillon, clavier, chimes, cowbell, piccolo, keyboard, bongo, lyre, bell, conga
4	gong
5	tambourine, cymbals, drum, castanets, maracas

	BOAW
1	xylophone, glockenspiel, cowbell, tambourine, chimes, celeste, maracas, bell, conga
2	flute, piano, violin, clarinet, saxophone, mandolin, harmonica, harp, oboe, banjo, lute, trumpet, zither, harpsichord, sitar, marimba, accordion, cornet, ukulele, clavichord, fiddle, horn, cymbals, balalaika, claves, lyre, keyboard, castanets, bugle, drum
3	trombone, tuba, cello, guitar, bassoon, baritone, didgeridoo, bass, piccolo, carillon, bongo
4	gong
5	clavier

	MM-NAE-INST
1	glockenspiel, xylophone, celeste, zither
2	trombone, cello, violin, clarinet, flute, saxophone, oboe, tuba, cornet, bassoon, trumpet, harmonica, accordion, piccolo, fiddle, horn, balalaika, bugle
3	chimes, bell
4	piano, guitar, mandolin, banjo, lute, harpsichord, marimba, ukulele, harp, clavichord, sitar, didgeridoo, bongo, keyboard, bass, lyre, carillon, clavier, baritone, conga, cowbell, drum, gong
5	maracas, tambourine, castanets, claves, cymbals

Table 10: Musical instruments closest to cluster centroid for various models.

deviation of V-measure scores, a well-known clustering evaluation metric (Rosenberg & Hirschberg, 2007)[7], obtained by applying the clustering algorithm a total of 100 times in order to mitigate differences due to the random seeding phase in k-means. The V-measure is the harmonic mean between a clustering's homogeneity and completeness: the former reflects to what extent clusters contain only data points which are members of a single class, the latter reflects to what extent data points that are members of a given class are elements of the same cluster. The results clearly show that the multi-modal representation, which utilizes both linguistic information and auditory input, performs better on this task than the uni-modal representations.

It is interesting to observe that the textual representations perform better than the auditory ones: a possible explanation for this result is that audio files in FreeSound are in some cases samples of multiple instruments, so if a bass is often accompanied by a drum this might affect the overall representation. The clusters that were obtained by the maximally performing model are reported in Table 10: for the 5 clusters under the three uni-modal models, it shows the nearest instruments to the cluster centroids, qualitatively demonstrating the greater cluster coherence for the multi-modal models, in particular the one based on NAEs. Percussive instruments appear relatively easy to pick out using the auditory signal (e.g. cluster 5 for NAE-INST), except for some of the obvious ones (drums, bongos, gongs). Piano-based instruments (e.g. cluster 1 for BOAW and cluster 3 for NAE-INST) are also grouped together, but that cluster interestingly never includes piano.

7. We find the same patterns in the results with other clustering metrics such as purity and B-cubed.

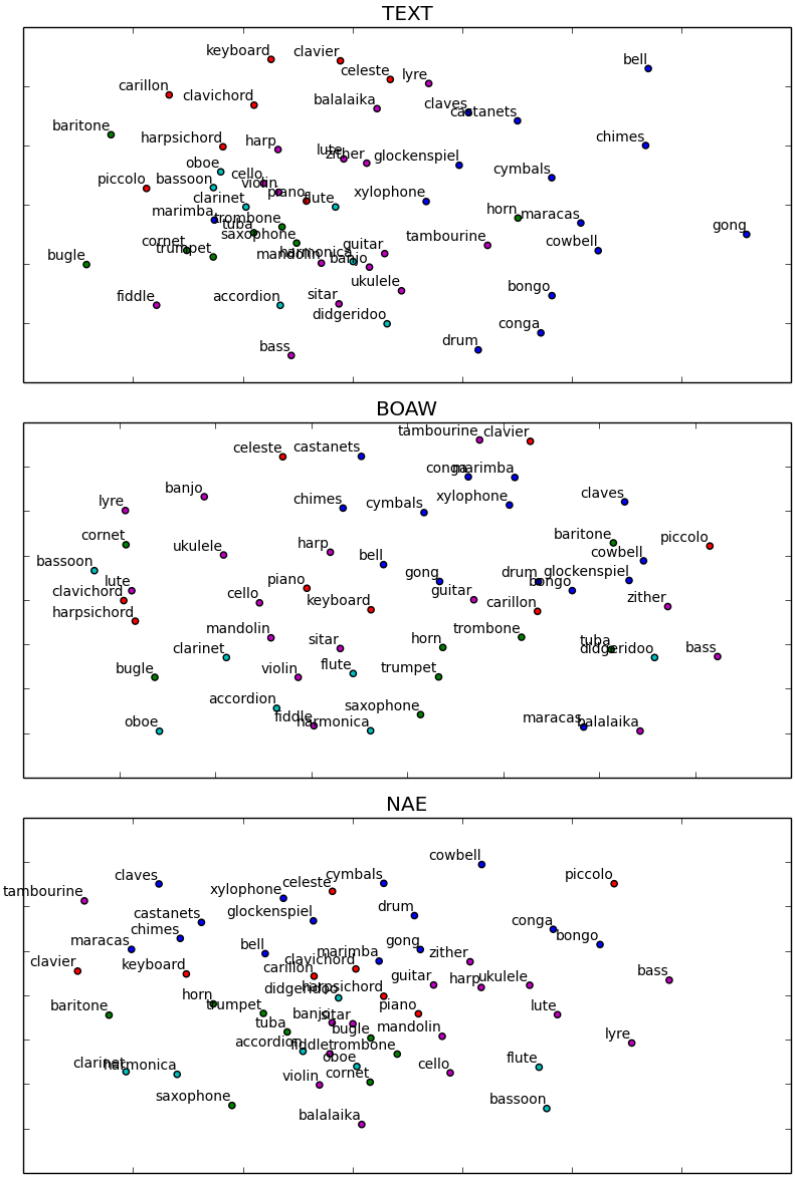

Figure 4: Multi-dimensional scaling of instrument representations.

The differences between the representations, together with the cluster assignments, can be visualized through multi-dimensional scaling (Hout, Papesh, & Goldinger, 2013): Figure 4 shows the instruments over the first two components, which shows that some of the instruments are clustered neatly by category.

5.5.2 Environmental Sounds Categorization

According to work by Gygi et al. (2007), a central challenge in the study of auditory perception and cognition is "to find the invariant acoustic information that specifies each

	TEXT
1	sneeze, cough, laugh, bubbling, splash
2	dog, cat, cow, toilet, rooster, baby, bird, airplane, typewriter, scissors, sheep, zipper, gun, helicopter, door, train, phone, wipers, water, clock, match, basketball, bowling
3	cymbals, thunder, claps, drums, bells, harp, whistle, honking, siren, neigh, keyboard, gallop, rain, wave, footsteps, saw

	NAE-ENV
1	harp, bells, keyboard, drums, cymbals, whistle, clock, siren, phone
2	splash, toilet, claps, laugh, zipper, water, rain, match, door, wipers, bowling, bubbling, sneeze, airplane, gun, cough, thunder, gallop, scissors, typewriter, helicopter, train, basketball, wave, saw
3	cat, dog, sheep, rooster, cow, bird, baby, neigh, honking, footsteps

	BOAW
1	whistle, siren, bird, laugh, bells, cat, baby, phone, sheep, harp, rooster, neigh, bubbling, keyboard, cow, dog
2	splash, rain, toilet, drums, saw, cymbals, water, scissors, match, gun, typewriter, zipper, wave, sneeze, claps, cough, clock, honking, footsteps, gallop, door, basketball
3	airplane, helicopter, train, wipers, thunder, bowling

	MM-NAE-ENV
1	harp, bells, keyboard, drums, cymbals, whistle, clock, siren, phone
2	splash, toilet, claps, laugh, zipper, water, rain, match, door, wipers, bowling, bubbling, sneeze, airplane, gun, cough, thunder, gallop, scissors, typewriter, helicopter, train, basketball, wave, saw
3	cat, dog, sheep, rooster, cow, bird, baby, neigh, honking, footsteps

Table 11: Environmental sound representations closest to cluster centroid for various models.

object or event", or alternatively "to determine how objects and events are identified in the absence of acoustic specificity". Here, we aim to show that learned auditory representations can be useful for examining these types of questions from a computational perspective: high-quality representations allow for clear differentiation between the acoustic information of objects.

To illustrate this, we compare our auditory representations with an experiment from Gygi et al. (2007). They collected similarity ratings from annotators for all classes in the dataset of 50 environmental sound categories. The similarity ratings were averaged to form a similarity matrix, after which they apply multi-dimensional scaling to examine categorizations of environmental sounds. They find some clearly defined groupings of environment sounds in their study: impacts, continuous sounds, and vocalizations and signals. We examine whether the same grouping are identifiable if we apply multi-dimensional scaling to representations learned from multi-modal data (rather than human similarity ratings).

Figure 5 plots the first two components in multi-dimensional scaling for both the textual and the NAE representations. The NAE-ENV classifier was trained on the same set of sound labels, so we can study how sounds are categorized with such representations and whether this matches the cognitive sound groups. It can be seen that the textual model clusters neatly in terms of relatedness—e.g., for instruments and animal sounds. Such a clustering, however, does not apply to auditory perception: a *gallop* and a *neigh* do not at all sound similar. The NAE representations are much more intuitive in that sense: a *sneeze* and a *cough* sound similarly, a *baby*'s sound is a *laugh* which (arguably) sounds quite similar to a *sheep*'s "baa".

Groupings similar to those found by Gygi et al. (2007) can be discerned in the figure: while impact sounds are not clustered for textual representations, they are clearly grouped in the bottom left corner for NAEs. In a similar fashion, signals and vocalizations (which

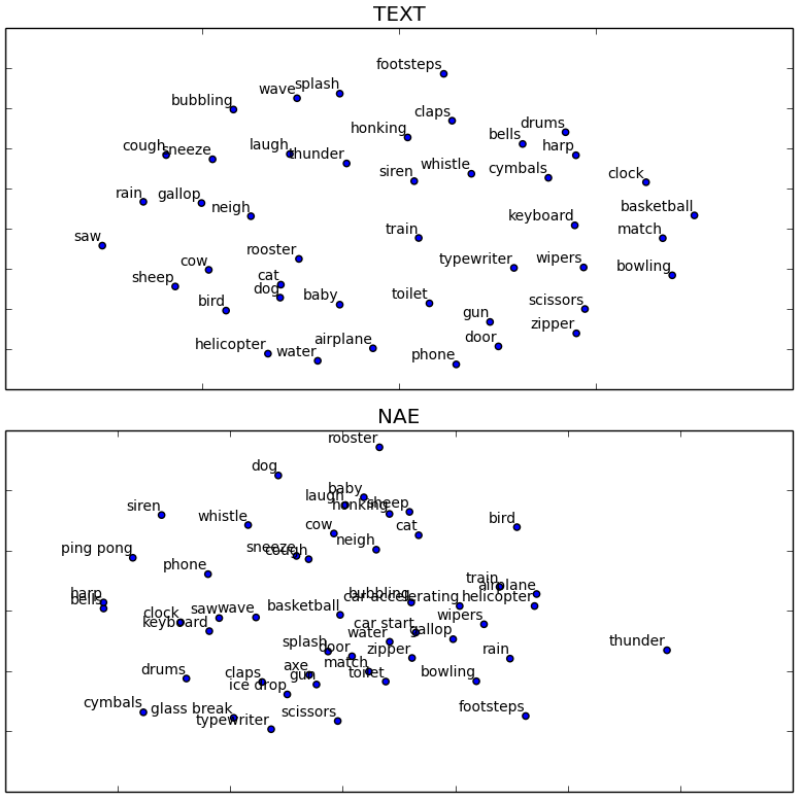

Figure 5: Multi-dimensional scaling of environmental sounds

include animal sounds) are grouped at the top. The set of continuous sounds is harder to identify, but there is a clear group of machine/tool sounds at the bottom of the center (tracing from trains and airplanes through to typewriters and scissors).

Gygi et al. (2007) do not provide an explicit categorization that directly allows for a quantitative evaluation. They do, however, list the categorization obtained from applying hierarchical clustering to the human similarity ratings[8]. They distinguish between three classes: continuous sounds, distinct sounds and harmonious sounds. Using this categorization, we can evaluate the clustering of neural auditory embeddings based on environmental sounds as well. Table 9(b) shows the mean and standard deviation of the V-measure scores when clustering the various models using k-means.

In short, the groupings in this preliminary multi-dimensional scaling analysis show remarkable similarity to Gygi et al.'s findings. We take this to indicate that auditory representations are not only useful for improving representations to be used in semantics tasks, but may also be useful for cognitive science experiments that involve auditory data, as was the case with the work of Gygi et al. (2007). A natural avenue for further exploration in this respect is examining the categorical knowledge transfer across auditory and visual domains, as recently demonstrated by Yildirim and Jacobs (2015).

8. See Figure 2 in Gygi et al. (2007).

6. Discussion

In these experiments we have relied on FreeSound as our source of sound files. Although the queries were restricted to a degree, we are ultimately dependent on FreeSound and its community efforts in uploading and tagging sound files, which may explain some of our findings. The fact that *gong* is in its own cluster for BOAW and NAE-INST representations seems to indicate that the audio samples on FreeSound already make it an outlier, as opposed to gongs having some special properties. In that respect, it is all the more interesting that such relatively noisy sound signals lead to improvements on semantic tasks, especially on auditory-relevant ones. A better, or more cleaned up, source of auditory data (e.g., with more stringent labelling or with outliers removed) might increase representational quality further.

The auditory representations learned here could be used in a variety of audio-related tasks that are not necessarily related to semantics, from musical preference prediction to identifying environmental background noise in video. We chose to evaluate on semantic relatedness here, because it shows how well the learned representations reflect human similarity and relatedness judgments. This type of intrinsic evaluation has been frequently used as an indicator of representation quality. However, such similarity and relatedness judgment datasets are not modality-specific, which means that they are susceptible to priming; i.e., if a previous comparison was clearly visual, e.g. *bright-light*, subjects might rely more on the visual modality for judging the next comparison. Furthermore, the dominance of vision in perceptually grounded cognition (Gazzaniga, 1995) probably biases similarity and relatedness judgments of concrete word pairs towards that modality. This might explain why visual grounding yields higher relative improvements than auditory grounding. Including auditory information, however, is not detrimental, as we have shown. In cases where auditory information is relevant, auditory grounding leads to large improvements, which merits further exploration of this particular modality.

The idea that learned representations can also shed light on cognitive questions goes back at least to the work of Landauer and Dumais (1997), and was reiterated in work by Lenci (2008) specifically for distributional semantics models. This is probably even more the case for grounded distributional models such as discussed here. In particular, multi-modal representations open up interesting possibilities for interdisciplinary studies between psychological, neurological and computational representation learning approaches (Kriegeskorte, Mur, & Bandettini, 2008).

7. Conclusions

We have studied grounding semantic representations in raw auditory perceptual information, using a bag-of-audio-words model and neural audio embeddings (NAEs) transferred from a convolutional neural network. NAEs were obtained by extracting the final layer from networks trained on audio recognition tasks, using spectrogram images. The auditory representations were compared to textual representations and combined with them using two standard fusion strategies. We evaluated on a well-known semantic similarity and relatedness benchmark and performed a detailed analysis of our findings. To show the applicability of auditory representations to auditory-relevant tasks, we examined musical

instrument clustering. To show how such auditory representations mirror findings in cognitive science studies, we performed a preliminary analysis comparing learned representations with psychological acoustic similarity experiments. We found that multi-modal representations perform much better than auditory or textual representations on musical instrument clustering, and that NAEs are useful for cognitive modeling of auditory perception, closely mirroring human categorizations of audio signals.

It may well be the case that the auditory modality is better suited for other evaluations or particularly useful in specific downstream tasks, but we have chosen to follow standard evaluations in multi-modal semantics to allow for a direct comparison. As indicated in the introduction, why stop at the visual modality? We hope to have shown that similar advances to those achieved by visually grounded models may be possible with non-visually grounded models as well. Our findings point toward fruitful applications of grounded representations in downstream tasks that relate to audio, as well as to the relatively unexplored area of linking grounded representations with cognitive studies. We hope that this will ultimately lead to perceptually grounded models in artificial intelligence that rely on data from all modalities, as a unified model that captures human semantic knowledge and experience.

Acknowledgments

This work was carried out while DK was a graduate student at the University of Cambridge. During part of this work, DK was supported by EPSRC grant EP/I037512/1. SC acknowledges ERC Starting Grant DisCoTex (306920) and EPSRC grant EP/I037512/1. We are grateful to Xavier Serra, Frederic Font Corbera, Alessandro Lopopolo, Emiel van Miltenburg, Ivan Vulić, Laura Rimell and the anonymous reviewers for useful suggestions and feedback.

References

Baroni, M. (2016). Grounding distributional semantics in the visual world. *Language and Linguistics Compass*, *10*(1), 3–13.

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, *43*(3), 209–226.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of ACL*, pp. 238–247, Baltimore, MA.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *35*(8), 1798–1828.

Bruni, E., Tran, N., & Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artifical Intelligence Research*, *49*, 1–47.

Bulat, L., Kiela, D., & Clark, S. (2016). Vision and Feature Norms: Improving automatic feature norm learning through cross-modal maps. In *Proceedings of NAACL-HLT 2016*, San Diego, CA.

Chechik, G., Ie, E., Rehn, M., Bengio, S., & Lyon, D. (2008). Large-scale content-based audio retrieval from text queries. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pp. 105–112. ACM.

Clark, S. (2015). Vector Space Models of Lexical Meaning. In Lappin, S., & Fox, C. (Eds.), *Handbook of Contemporary Semantic Theory*, chap. 16. Wiley-Blackwell, Oxford.

Dahl, G. E., Yu, D., Deng, L., & Acero, A. (2012). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(1), 30–42.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255.

Dieleman, S. (2016). *Learning feature hierarchies for musical audio signals*. Ph.D. thesis, University of Ghent.

Dieleman, S., Brakel, P., & Schrauwen, B. (2011). Audio-based music classification with a pretrained convolutional network. In *12th International Society for Music Information Retrieval Conference (ISMIR-2011)*, pp. 669–674.

Eronen, A. (2003). Musical instrument recognition using ICA-based transform of features and discriminatively trained HMMs. In *Proceedings of the Seventh International Symposium on Signal Processing and Its Applications*, Vol. 2, pp. 133–136.

Feng, Y., & Lapata, M. (2010). Visual information in semantic representation. In *Proceedings of NAACL*, pp. 91–99, Los Angeles, CA.

Flanagan, J. L. (2013). *Speech analysis synthesis and perception*, Vol. 3. Springer Science & Business Media.

Font, F., Roma, G., & Serra, X. (2013). Freesound technical demo. In *Proceedings of the 21st ACM international conference on multimedia*, pp. 411–412. ACM.

Foote, J. T. (1997). Content-based retrieval of music and audio. In *Voice, Video, and Data Communications*, pp. 138–147.

Fountain, T., & Lapata, M. (2011). Incremental models of natural language category acquisition. In Carlson, C., Hölscher, & Shipley, T. (Eds.), *Proceedings of the 33nd Annual Conference of the Cognitive Science Society*, Austin, TX: Cognitive Science Society.

Gazzaniga, M. S. (Ed.). (1995). *The Cognitive Neurosciences*. MIT Press, Cambridge, MA.

Gygi, B., Kidd, G. R., & Watson, C. S. (2007). Similarity and categorization of environmental sounds. *Perception & psychophysics*, *69*(6), 839–855.

Hamel, P., & Eck, D. (2010). Learning features from music audio with deep belief networks.. In *ISMIR*, pp. 339–344. Utrecht, The Netherlands.

Hamel, P., Lemieux, S., Bengio, Y., & Eck, D. (2011). Temporal pooling and multiscale learning for automatic annotation and ranking of music audio.. In *ISMIR*, pp. 729–734.

Harnad, S. (1990). The symbol grounding problem. *Physica D*, *42*, 335–346.

Haugeland, J. (1985). *Artificial Intelligence: The Very Idea*. Massachusetts Institute of Technology, Cambridge, MA, USA.

Hill, F., & Korhonen, A. (2014). Learning abstract concept embeddings from multi-modal data: Since you probably can't see what I mean. In *Proceedings of EMNLP*, pp. 255–265, Lisbon, Portugal.

Hout, M. C., Papesh, M. H., & Goldinger, S. D. (2013). Multidimensional scaling. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(1), 93–103.

Kiela, D., & Bottou, L. (2014). Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Proceedings of EMNLP*, pp. 36–45, Doha, Qatar.

Kiela, D., Bulat, L., & Clark, S. (2015). Grounding semantics in olfactory perception. In *Proceedings of ACL*, pp. 231–236, Beijing, China.

Kiela, D., & Clark, S. (2015). Multi- and cross-modal semantics beyond vision: Grounding in auditory perception. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 2461–2470, Lisbon, Portugal.

Kiela, D., Hill, F., Korhonen, A., & Clark, S. (2014). Improving multi-modal representations using image dispersion: Why less is sometimes more. In *Proceedings of ACL*, pp. 835–841, Baltimore, MA.

Kiela, D., Rimell, L., Vulić, I., & Clark, S. (2015a). Exploiting image generality for lexical entailment detection. In *Proceedings of ACL*, pp. 119–124, Beijing, China.

Kiela, D., Vulić, I., & Clark, S. (2015b). Visual bilingual lexicon induction with transferred convnet features. In *Proceedings of EMNLP*, pp. 148–158, Lisbon, Portugal.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, *2*, 1–28.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Proceedings of NIPS*, pp. 1106–1114.

Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge.. *Psychological review*, *104*(2), 211.

Lazaridou, A., Pham, N. T., & Baroni, M. (2015). Combining language and vision with a multimodal skipgram model. In *Proceedings of NAACL*, Denver, CO.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436—444.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics*, *20*(1), 1–31.

Leong, C. W., & Mihalcea, R. (2011). Going beyond text: A hybrid image-text approach for measuring word relatedness. In *Proceedings of IJCNLP*, pp. 1403–1407, Chiang Mai, Thailand.

Lopopolo, A., & van Miltenburg, E. (2015). Sound-based distributional models. In *Proceedings of IWCS*, London, UK.

Louwerse, M. M. (2008). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, *59*(1), 617–645.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91–110.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*(4), 547–559.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of ICLR*, Scottsdale, AZ.

Nelson, D. L., McEvoy, C. L., , & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods*, *36*(3), 402–407.

O'Shaughnessy, D. (1987). *Speech communication: human and machine*. Addison-Wesley series in electrical engineering: digital signal processing. Universities Press (India) Pvt. Limited.

Pancoast, S., & Akbacak, M. (2012). Bag-of-audio-words approach for multimedia event classification.. In *Proceedings of Interspeech*, pp. 2105–2108.

Pascanu, R., Mikolov, T., & Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In *Proceedings of ICML*, Atlanta, GA.

Roller, S., & Schulte im Walde, S. (2013). A multimodal LDA model integrating textual, cognitive and visual modalities. In *Proceedings of EMNLP*, pp. 1146–1157.

Rosenberg, A., & Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of EMNLP-CoNLL*, pp. 410–420.

Schmitt, M., Ringeval, F., & Schuller, B. (2016). At the Border of Acoustics and Linguistics: Bag-of-Audio-Words for the Recognition of Emotions in Speech. In *Proceedings of Interspeech*.

Sculley, D. (2010). Web-scale k-means clustering. In *Proceedings of WWW*, pp. 1177–1178. ACM.

Shutova, E., Kiela, D., & Maillard, J. (2016). Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of NAACL-HTL 2016*, San Diego, CA.

Silberer, C., & Lapata, M. (2012). Grounded models of semantic representation. In *Proceedings of EMNLP*, pp. 1423–1433, Jeju, South Korea.

Sivic, J., & Zisserman, A. (2003). Video google: A text retrieval approach to object matching in videos. In *Proceedings of ICCV*, pp. 1470–1477.

Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America*, *8*(3), 185—190.

Synnaeve, G., Versteegh, M., & Dupoux, E. (2014). Learning words from images and speech. In *NIPS Workshop on Learning Semantics*, Montreal, Canada.

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Zafeiriou, S., et al. (2016). Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5200–5204. IEEE.

Turney, P. D., & Pantel, P. (2010). From Frequency to Meaning: vector space models of semantics. *Journal of Artifical Intelligence Research*, *37*(1), 141–188.

Uchida, Y., Sakazawa, S., Agrawal, M., & Akbacak, M. (2010). Kddi labs and sri international at trecvid 2010: Content-based copy detection.. In *TRECVID*.

Van den Oord, A., Dieleman, S., & Schrauwen, B. (2013). Deep content-based music recommendation. In *Advances in Neural Information Processing Systems*, pp. 2643–2651.

Vijayakumar, A. K., Vedantam, R., & Parikh, D. (2017). Sound-word2vec: Learning word representations grounded in sounds. In *Proceedings of EMNLP*.

Von Ahn, L., & Dabbish, L. (2004). Labeling images with a computer game. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 319–326. ACM.

Weston, J., Bengio, S., & Hamel, P. (2011). Multi-tasking with joint semantic spaces for large-scale music annotation and retrieval. *Journal of New Music Research*, *40*(4), 337–348.

Yildirim, I., & Jacobs, R. A. (2015). Learning multisensory representations for auditory-visual transfer of sequence category knowledge: a probabilistic language of thought approach. *Psychonomic Bulletin & Review*, *22*(3), 673–686.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Proceedings of ECCV*, pp. 818–833. Springer.