Large-scale representation learning from visually grounded untranscribed speech

Gabriel Ilharco[†*] **Yuan Zhang**[‡] **Jason Baldridge**[‡]
[†]Paul G. Allen School of Computer Science & Engineering,
University of Washington, Seattle, WA, USA
[‡]Google Research, Mountain View, CA, USA
gamaga@cs.washington.edu, {zhangyua,jridge}@google.com

Abstract

Systems that can associate images with their spoken audio captions are an important step towards visually grounded language learning. We describe a scalable method to automatically generate diverse audio for image captioning datasets. This supports pretraining deep networks for encoding both audio and images, which we do via a dual encoder that learns to align latent representations from both modalities. We show that a masked margin softmax loss for such models is superior to the standard triplet loss. We fine-tune these models on the Flickr8k Audio Captions Corpus and obtain state-of-the-art results—improving recall in the top 10 from 29.6% to 49.5%. We also obtain human ratings on retrieval outputs to better assess the impact of incidentally matching image-caption pairs that were not associated in the data, finding that automatic evaluation substantially underestimates the quality of the retrieved results.

1 Introduction

Natural language learning in people starts with speech, not text. Text is tidy: it comes in convenient symbolic units that vary little from one writer to another. Speech is continuous and messy: the sounds used to convey a given word are modified by those of surrounding words, and the rate of speech, its pitch, and more vary across speakers and even for the same speaker in different contexts. As such, problems involving speech provide distinct challenges and opportunities for learning language representations that text-based work—which represents the vast majority—gets a free pass on.

Recent work has explored various means to transform raw speech into symbolic forms with little or no supervision (Park and Glass, 2007; Varadarajan et al., 2008; Ondel et al., 2016; Kamper et al., 2017a; Bhati et al., 2018). However, learning natural language starts with grounded, contextualized speech. While infants as young as 8-months-old can segment word-like units without non-linguistic information (Jusczyk and Aslin, 1995) and adults can learn to segment words in artificial languages (Saffran et al., 1996), a learner must ultimately ground their representations of linguistic sequences (Harnad, 1990) to effectively use them to refer to objects, events and more. Furthermore, learning from rich perceptual data and interactions can be more efficient as it provides additional cues to the identities of words and their meaning in context.

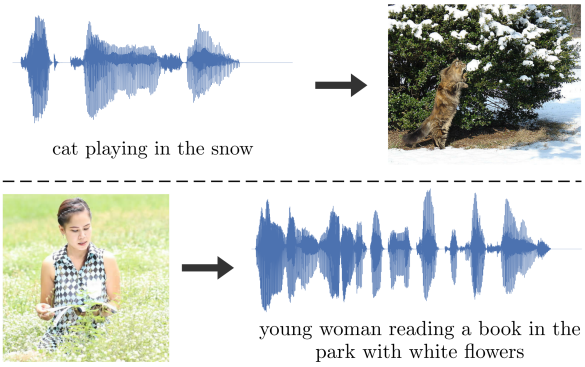

Figure 1: Models that encode speech segments and images into a shared latent space enable images to be retrieved using their audio descriptions (top) and to associate images with spoken captions (bottom). Text captions are provided for clarity; only speech and images are used by the models.

We address the problem of relating images to audio captions that describe them (Figure 1), building on previous research into learning from visually grounded, untranscribed speech (Harwath and Glass, 2015; Sun et al., 2016; Harwath et al., 2016; Chrupała et al., 2017; Kamper et al., 2017b; Chrupała, 2019; Harwath and Glass, 2019). Such problem settings provide opportunities both to improve our theoretical understanding of language

[*] Work done as a member of the Google AI Residency Program.

Proceedings of the 23rd Conference on Computational Natural Language Learning, pages 55–65
Hong Kong, China, November 3-4, 2019.

as well as to realize gains on practical problems—including voice interaction with virtual assistants, image retrieval based on speech, and generally better supporting people with visual impairments.

Our contribution is to improve performance on bidirectional speech/image retrieval through better data and better models for learning fixed dimensional latent representations of both modalities. We construct a synthetic speech caption dataset for pretraining by applying text-to-speech (TTS) on Conceptual Captions (Sharma et al., 2018), a dataset with 3.3 million diverse image-caption pairs. Unlike Chrupała et al. (2017), who similarly applied TTS to MS-COCO (Chen et al., 2015), we inject diversity by varying the voice, speech rate, pitch and volume gain on every synthetically produced audio caption. We refer to the resulting dataset as Conceptual Spoken Captions (CSC). CSC's scale allows us to train deeper models than previous work. We use Inception-ResNet-v2 (Szegedy et al., 2017) to encode both the audio and visual modalities in a dual encoder model, pretraining on CSC and then fine-tuning and evaluating on human speech in the smaller Flickr Audio Caption Corpus (FACC) (Harwath and Glass, 2015). Using an adapted batch loss function rather than the triplet loss used in previous work, we substantially improve on the previous state-of-the-art for the standard FACC retrieval tasks.

Image captioning datasets contain positively paired items—but that does not imply that a random image and caption cannot also be a valid match. For instance, in FACC there are many spoken captions about beaches and sunsets and plenty of images that match these captions; two different images with descriptions *"A surfer is riding a wave."* and *"A man surfs the wave"* are likely compatible. It is of course not feasible to exhaustively annotate all pairwise associations, so we have human raters judge the top five retrieved results for two models to assess the impact of this aspect of the data on automatic retrieval metrics used thus far. Unsurprisingly, models retrieve many compatible results that are unpaired in FACC: with the human evaluations, we find consistent increases in recall.

2 Data

Larger training datasets support better performance and generalization (Banko and Brill, 2001; Halevy et al., 2009; Sun et al., 2017), especially for deep models. Collecting labels from people has become easier via crowd computing (Buhrmester et al., 2011), but is still expensive and remains a bottleneck for creating broad and representative datasets. This motivates the case for exploiting incidental annotation (Roth, 2017) and automating some aspects of dataset creation. The current trend of using machine translation systems to produce augmented datasets for machine translation itself (Sennrich et al., 2016) and for monolingual tasks like classification (Yu et al., 2018) and paraphrasing (Wieting and Gimpel, 2018) is a good example of this.

For speech image captioning, Chrupała et al. (2017) used a Text-to-Speech (TTS) system to create audio from the textual captions given in the MS-COCO dataset, resulting in 300k unique images with 5 spoken captions each. We scale this idea to the larger and more diverse textual Conceptual Captions dataset with 3.3 million unique image and captions, additionally modifying the produced speech by using multiple voices and random perturbations to the rate, pitch and audio. Our goal is to make the resulting data more effective for pretraining models so they can learn more efficiently on smaller amounts of human speech.

2.1 Conceptual Captions

Image captioning datasets have ignited a great deal of research at the intersection of the computer vision and natural language processing communities (Lin et al., 2014; Vinyals et al., 2015; Bernardi et al., 2016; Anderson et al., 2018). Getting annotators to provide captions works well with crowd computing, but Sharma et al. (2018) exploit incidental supervision for this task to obtain greater scale with their Conceptual Captions dataset. It contains 3.3 million pairs of image and textual captions, where pairs are extracted from HTML web pages using the *alt-text* field of images as a starting point for their descriptions.

The textual captions are processed in a hypernymization stage. Named entities and syntactic dependency annotations are obtained using Google Cloud Natural Language APIs, which are matched to hypernym terms using the Google Knowledge Graph Search API. Proper nouns, numbers, units, dates, durations and locations are removed; identified named-entities are substituted with their hypernym, merging together analogous terms when possible. For example, the original alt-text (1) is converted to the conceptual caption (2).

(1) **alt-text**: *Musician Justin Timberlake per-*

forms at the 2017 Pilgrimage Music & Cultural Festival on September 23, 2017 in Franklin, Tennessee.

(2) **conceptual caption**: *pop artist performs at the festival in a city.*

There are many sequential filtering steps for improving the quality of the captions—see Sharma et al. (2018) for a thorough description. As quality control, a random sample of 4K conceptual captions were rated by human annotators, and 90.3% were judged "good" by at least 2 out of 3 raters.

2.2 Conceptual Spoken Captions

We use TTS to generate a high-fidelity spoken sentence for each of the 3.3 million textual captions in the Conceptual Captions dataset.[1] We use the Google Cloud Speech API[2] for TTS. Internally, the service uses a WaveNet model (Van Den Oord et al., 2016) to generate audio. For diversity, the speech is synthesized using parameter variations, as follows:

- *Voice*, which is sampled uniformly from a set of 6 different voices generated using a WaveNet model for American English.
- *Speaking rate* controls the speed of the synthesized audio. A speaking rate of 1.0 means the normal speed of a given voice, while a speaking rate of 2.0 means twice as fast. When synthesizing the data, we draw this parameter from a Gaussian distribution $\sim \mathcal{N}(1.0, 0.1^2)$.
- *Pitch* controls how high/deep the voice is. For example, if set to 1, this means the voice will be synthesized 1 semitones above the original pitch. This parameter is drawn from a Gaussian distribution $\sim \mathcal{N}(0.0, 1.0^2)$.
- *Volume gain* controls a gain in dB with respect to the normal native signal amplitude. If set to 0, the voice is synthesized without alterations in volume. This parameter is drawn from a Gaussian distribution $\sim \mathcal{N}(0.0, 2.0^2)$.

To avoid degenerate cases, we clip the values sampled from the Gaussian distributions described above such that they are never more than 2 times the standard deviation from the mean. All spoken captions are generated in 16000 Hz.

[1]The alt-text does not come with the dataset and cannot be redistributed, so we focus on the conceptual captions for ease of experimentation and reproducibility.

[2]https://cloud.google.com/text-to-speech/

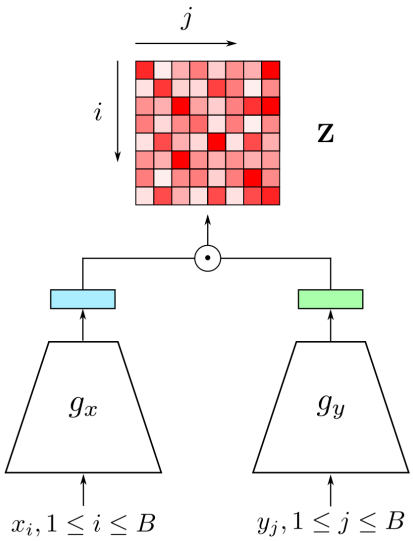

Figure 2: Dual-encoder model architecture.

2.3 Flickr Audio Caption Corpus

The Flickr Audio Caption Corpus (FACC) (Harwath and Glass, 2015) consists of 40,000 pairs of images and spoken captions, with 8000 unique images, of which 1000 are held for validation and 1000 for testing. The spoken captions are generated from humans reading the textual captions from the Flickr8k dataset (Hodosh et al., 2013), originally crowd-sourced and based on images from Flickr.

We use FACC for evaluation, both when pretraining on Conceptual Spoken Captions and when training on FACC from scratch. Like previous work, the core evaluation considered is retrieval of the known paired image given an audio caption within some top-k set of retrieved items (e.g. R@1 for whether the first item retrieved is the paired one and R@10 for whether it is in the top ten results). We also conduct human evaluations on retrieval outputs to detect the presence of unpaired but matching image-caption pairs identified by the models and thereby better assess their impact on performance.

3 Model

Dual encoders are used in a wide range of applications, including signature verification (Bromley et al., 1994), object tracking (Bertinetto et al., 2016), sentence similarity (Mueller and Thyagarajan, 2016), improving neural machine translation (Yang et al., 2019) and many others. The core of this set of architectures is a simple two-tower model illustrated in Figure 2, where inputs $x \in \mathcal{X}$ are processed by an encoder g_x and inputs $y \in \mathcal{Y}$ by a second encoder g_y. The inputs may come

from the same distribution—or they may be from entirely different sources or modalities. The towers may share the same architecture and weights—or they can be completely unlike and disconnected.

These models are standard in audiovisual image captioning (Harwath and Glass, 2015; Chrupała, 2019; Harwath et al., 2018). In this setting, the dual encoder model, is composed by a visual tower, g_{vis}, processing the images, and an audio tower, g_{aud}, processing the spoken captions. The model is trained to map both modalities into a joint latent space. Here, we extend previous work to consider a batched margin loss, which we show to be superior for learning dense representations for retrieval.

Notation. The inputs are processed in batches of size B. For each input x_k and y_k in the batch, $1 \leq k \leq B$, let $g_x(x_k)$ and $g_y(y_k)$ be their latent representations extracted by the corresponding tower. We define the $B \times B$ matrix $\mathbf{Z}$ as the similarity between the latent representations for each pair of elements in the batch. A natural choice for that similarity is the dot product between the latent representations:

$$\mathbf{Z}_{ij} = g_x(x_i) \cdot g_y(y_j) \tag{1}$$

As shown in Figure 2, $\mathbf{Z}$ encodes all pairwise associations in the batch. However, an additional aspect of some datasets must be taken into account: often times the same input x can match multiple inputs y or vice-versa—for instance, both Flickr8k and MS-COCO have multiple captions for the each image. To respect these pairs when they land in the same batch—and thus not penalize models for (correctly) associating them—we define a $B \times B$ masking matrix $\mathbf{M}$:

$$\mathbf{M}_{ij} = \begin{cases} 0, & \text{if } x_i \text{ matches } y_j \\ 1, & \text{otherwise} \end{cases} \tag{2}$$

All pairs (x_k, y_k) match and this equivalence is transitive, so $\mathbf{M}$ is symmetric and all diagonal elements $\mathbf{M}_{kk}$, $1 \leq k \leq B$ are zero.

Triplet Loss. Both Chrupała (2019) and Harwath et al. (2018) (and their previous work) employ the triplet loss function given in Equation 3.

$$\mathcal{L}_{\text{T}} = \sum_{k=1}^{B} \big(\max(0, \mathbf{Z}_{km} - \mathbf{Z}_{kk} + \delta) + \max(0, \mathbf{Z}_{nk} - \mathbf{Z}_{kk} + \delta) \big) \tag{3}$$

For each value k, m is randomly drawn from a uniform distribution over indices j such that $M_{kj} = 1$, and n over indices i such that $M_{ik} = 1$.

Masked Margin Softmax Loss. The triplet loss (3) used previously misses opportunities to learn against a wider set of negative examples, namely all those in the batch that are not known to be positively associated (i.e., $\mathbf{M}_{ij} = 1$). To exploit these additional negatives, we minimize the Masked Margin Softmax (MMS) loss function, inspired by Henderson et al. (2017) and Yang et al. (2019). MMS simulates x-to-y and y-to-x retrievals inside the batch. It is defined at a high level as:

$$\mathcal{L}_{\text{MMS}} = \mathcal{L}_{xy} + \mathcal{L}_{yx} \tag{4}$$

$\mathcal{L}_{\text{MMS}}$ is the sum of losses defined over x-to-y (Eq. 5) and y-to-x (Eq. 6) in-batch retrievals.

$$\mathcal{L}_{xy} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{e^{\mathbf{Z}_{ii} - \delta}}{e^{\mathbf{Z}_{ii} - \delta} + \sum_{j=1}^{B} \mathbf{M}_{ij} e^{\mathbf{Z}_{ij}}} \tag{5}$$

$$\mathcal{L}_{yx} = -\frac{1}{B} \sum_{j=1}^{B} \log \frac{e^{\mathbf{Z}_{jj} - \delta}}{e^{\mathbf{Z}_{jj} - \delta} + \sum_{i=1}^{B} \mathbf{M}_{ij} e^{\mathbf{Z}_{ij}}} \tag{6}$$

These are equivalent to a cross-entropy loss after a column-wise or row-wise softmax on the matrix $\mathbf{Z}$, subject to the masking constraints in $\mathbf{M}$ and margin δ.

The margin hyperparameter δ is gradually increased as training progresses. Empirically, we found that, with a fixed δ, large values lead to unstable performance in early training, while small values lead to negligible results in final performance. Starting with a small δ and increasing it does not hurt early training and forces the model to learn from a harder task later on. There many ways to increase δ along training—e.g. linearly, quadratically, and exponentially. The latter is used in this work.

Contrasting Equations 3 and 4, the former chooses a negative sample randomly, while the latter takes advantage of all negative pairs in the batch and thus improves sample efficiency. $\mathcal{L}_{\text{MMS}}$ has three main differences with Yang et al. (2019): (1) a masking term that accounts for the fact that there might be multiple positive choices in the batch for a given input; (2) a varying margin term δ, which is increased during training; (3) a log term that makes MMS more closely resemble a cross-entropy loss.

		Speech to Image					Image to Speech				
Loss	Batch Size	R@1	R@5	R@10	R@50	R@100	R@1	R@5	R@10	R@50	R@100
$\mathcal{L}_{\text{T}}$	48	.037	.109	.165	.367	.474	.031	.101	.155	.346	.455
$\mathcal{L}_{\text{MMS}}$	12	.025	.083	.129	.311	.432	.024	.083	.132	.315	.433
	24	.054	.143	.206	.418	.533	.046	.137	.197	.411	.520
	48	**.078**	**.204**	**.282**	**.499**	**.604**	**.074**	**.194**	**.269**	**.485**	**.587**

Table 1: Performance on the validation set of Conceptual Spoken Captions, comparing different loss functions and batch sizes.

4 Experiments

4.1 Experimental settings

Image preprocessing. During training, data augmentation is performed by randomly distorting the brightness and saturation of images. Each image is also randomly cropped or padded such that at least 67% of the area of the original image is covered, and re-scaled if necessary to 299×299. During evaluation, we do not perform color distortions, and we crop/pad the central portion of the images.

Audio preprocessing. We extract 128 Mel-Frequency Cepstral Coefficients (MFCCs) from the raw audio signals using a window size of 20ms. The audio signals have a sampling rate of 16000Hz. We compute features every 10ms, such that each window has a 50% overlap with its neighbors. During training, we randomly crop/pad the MFCCs in the temporal dimension, and perform data augmentation as in Park et al. (2019), using one mask with a frequency mask parameter of 20 and a time mask parameter of 40. We do not perform time warping.

Encoders. Both audio and image encoders are Inception-ResNet-v2 networks (Szegedy et al., 2017), allowing the model to reap the benefits of relatively low computational cost, fast training and and strong performance when combining the Inception architecture with residual connections.[3] Related to our setting for audio processing, Li et al. (2019) also uses residual convolutional neural networks for state of the art results on LibriSpeech dataset (Panayotov et al., 2015). For the audio tower, we stack 3 replicas of the MFCCs and treat them as images. For each modality, a 1536-dimensional latent space representation is extracted. Despite using the same architecture for both encoders, their weights are not shared. Unless specified otherwise, the models are *not* pretrained.

[3] See Bianco et al. (2018) for an extensive benchmark analysis of popular convolutional neural network architectures.

Optimization. Models are trained using Adam (Kingma and Ba, 2014), with an initial learning rate of 0.001 and an exponential decay of 0.999 every 1000 training steps, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1\text{e}{-8}$. We use a weight decay of $4\text{e}{-5}$, and train on 32 GPUs until convergence. Unless specified otherwise, the optimization objective is minimizing the loss $\mathcal{L}_{\text{MMS}}$ (Eq. 4) with a margin term initially set to $\delta = 0.001$ exponentially and increased by a factor of 1.002 every 1000 steps.

4.2 Retrieval: Conceptual Spoken Captions

Our primary aim with CSC is to use it for pretraining for later fine-tuning and evaluation on datasets with human speech instead of TTS. Nevertheless, we can compare different loss functions and different batch sizes on the CSC validation set to better understand the impact of these parameters.

We train models on CSC for 3 million steps, cropping/padding spoken captions to a duration of 3.5 seconds and using the loss functions $\mathcal{L}_{\text{T}}$ (Eq. 3) and $\mathcal{L}_{\text{MMS}}$ (Eq. 4). We find continuing improvements as batch size increases from 12 to 24 to 48. Furthermore, with the same batch size of 48, models optimized for minimizing $\mathcal{L}_{\text{MMS}}$ perform substantially better than those using $\mathcal{L}_{\text{T}}$, as summarized in Table 1. Of particular note is that R@1 scores for $\mathcal{L}_{\text{MMS}}$ (batch size 48) are more than double those of $\mathcal{L}_{\text{T}}$ in both directions.

4.3 Retrieval: Flickr Audio Caption Corpus

Table 2 compares previous results on the FACC dataset with those obtained by variations of our model. As a pre-processing step, spoken captions are cropped/padded to a duration of 8 seconds. After pretraining the model in CSC, we explore all possible combinations of using or not the pretrained weights for each of the branches g_{aud} and g_{vis} as a warm-starting point, training until convergence on FACC. Warm-starting each of the branches in the dual-encoder leads to substantial improvements

	Model	Caption to Image					Image to Caption				
		R@1	R@5	R@10	R@50	R@100	R@1	R@5	R@10	R@50	R@100
Text	Socher et al. 2014	-	-	.286	-	-	-	-	.2r90	-	-
	Karpathy et al. 2014	-	-	.425	-	-	-	-	.440	-	-
	Harwath and Glass 2015	-	-	.490	-	-	-	-	**.567**	-	-
	Chrupała et al. 2017	**.127**	**.364**	**.494**	-	-	-	-	-	-	-
Speech	Harwath and Glass 2015	-	-	.179	-	-	-	-	.243	-	-
	Chrupała et al. 2017	.055	0.163	.253	-	-	-	-	-	-	-
	Chrupała 2019	-	-	.296	-	-	-	-	-	-	-
	Ours (from scratch)	.018	.063	.101	.288	.428	.024	.072	.124	.332	.458
	Ours (warm-starting g_{aud})	.041	.138	.211	.467	.613	.550	.166	.241	.522	.654
	Ours (warm-starting g_{vis})	.062	.190	.279	.560	.703	.081	.242	.352	.664	.782
	Ours (warm-starting all)	**.139**	**.368**	**.495**	**.781**	**.875**	**.182**	**.435**	**.558**	**.842**	**.910**

Table 2: Retrieval scores on the test set of FACC.

over the baseline, and combining both branches leads to the best overall performance.

In particular, we improve R@10 for caption-to-image from the .296 obtained by Chrupała (2019) by 20% absolute to .495, without using multitask training or pretraining g_{vis} on ImageNet (Deng et al., 2009). The multitask training approach of Chrupała (2019) is complementary to our improvements, so further gains might be obtained by combining these strategies. Furthermore, very deep, residual convolutional neural networks over characters have been shown to perform well for text-based tasks (Conneau et al., 2017). We expect that our strategy of using the same basic architecture across different input types (speech, text and image) can be fruitfully extended to that setting. A related observation: while our results exceed previous results reported on text/image retrieval settings for FACC, we expect that recent advances in text encoding could easily beat those reported numbers.

We also explore very low-data regimes using our pretrained model (see Fig. 3). Using small training subsets randomly drawn from FACC, we report performance as a function of how much data the model sees. With as little as 10% of the original training data (3000 image/spoken caption pairs), the warm-started model performs competitively with a model trained on all training data.

Qualitative evaluation. Once a model is trained, any input (image or spoken caption) can be be used to query the corpus of images and spoken captions for nearest neighbors in the latent space. Figure 4 shows some examples of retrieved nearest neighbors in FACC's test set. Given a spoken caption or

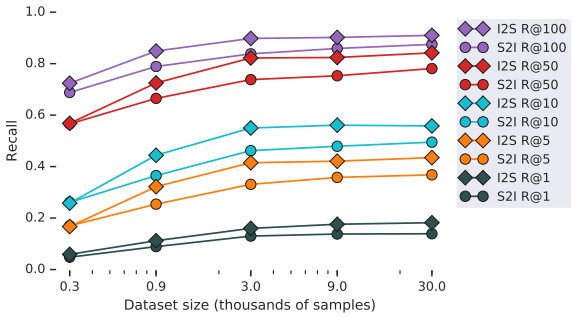

Figure 3: Ablations on low-data regime on FACC: chart shows recall scores for image-to-speech (I2S) and speech-to-image (S2I) retrieval, as a function of the amount of training data used for fine-tuning.

an image we show the five nearest image neighbors and five nearest caption neighbors. From these, it is clear that the representations capture many semantically salient attributes of the inputs. The retrieved items correctly share many thematic elements and many are clearly good matches even though the particular image-caption pairs are not associated in the data. This serves to reinforce our observation that R@k evaluations using only the known paired items is likely to underestimate the actual performance of the models—which we show to be the case with human evaluations in Section 4.4.

Only some items are substantially incompatible: e.g. an image of a car for a caption about a woman in a river (they share water spraying), a picture of three adults for a caption about children raising their hands, and a caption about a boy climbing a wall for an image of children playing leapfrog). That said, many details are poor matches, such as the count of objects (one ball versus many), colors

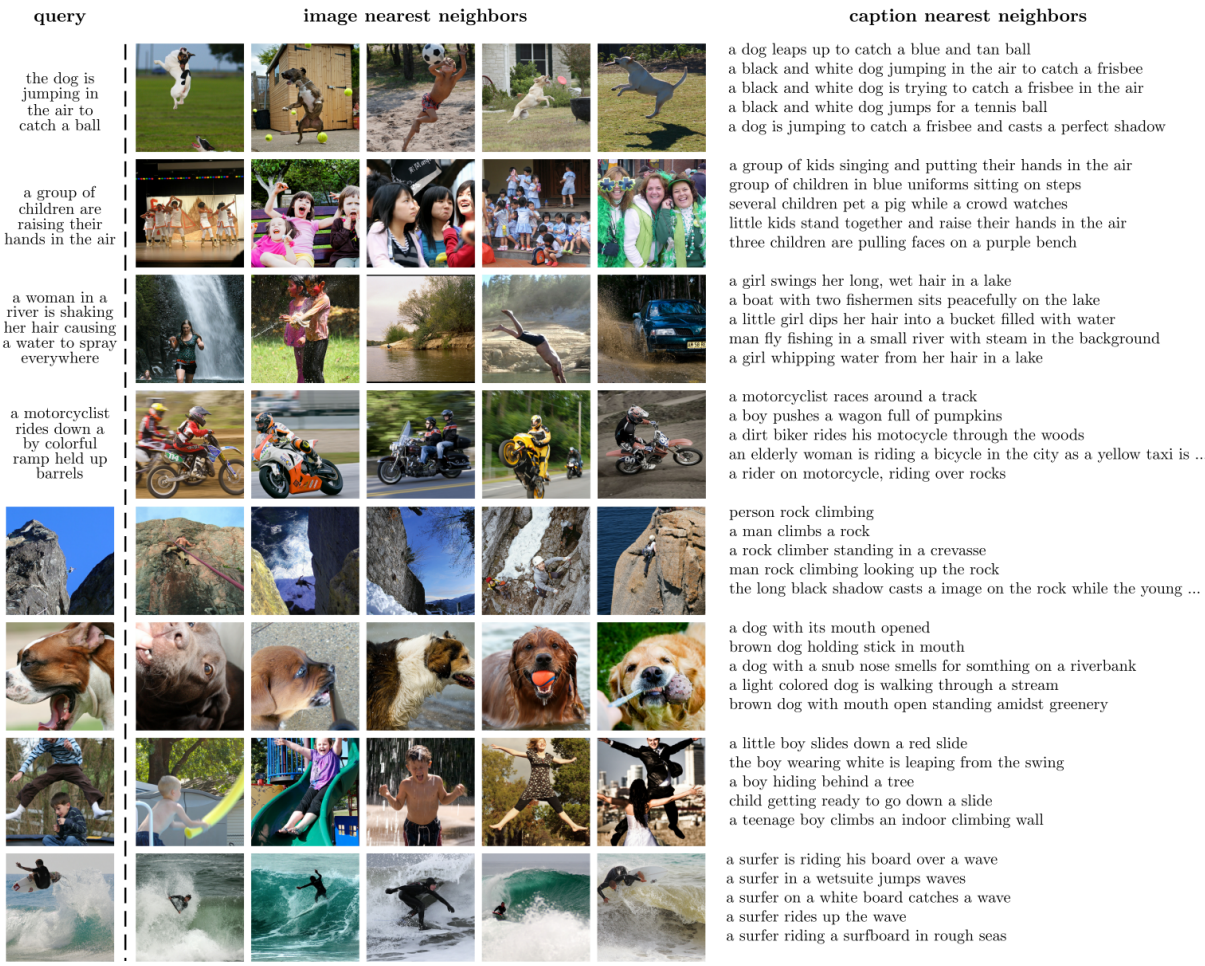

Figure 4: Nearest neighbors in the joint visual and acoustic latent space, best viewed with zoom: using 4 spoken captions and 4 images as queries, we extract from FACC's test set the closest 5 images and 5 spoken captions in the latent space for each of them. For simplicity, we show the text associated with each spoken caption.

(brown dogs versus multicolored ones), people descriptions (elderly woman versus male dirt biker), object identification (e.g. a yellow pool noodle viewed as similar to slides), processes (jumping versus sliding) and perspective (man looking up versus viewed from behind and climbing). As such, there is clearly significant headroom for better, more fine-grained modeling of both captions and images. Additionally, cross-modal attention mechanisms (Xu et al., 2015) and other explainability techniques (Ribeiro et al., 2016) could help better inspect and understand a model's predictions.

Furthermore, as noted by Chrupała et al. (2017), text-based retrieval models often handle misspellings poorly. In contrast, speech-based models are unlikely to suffer from similar problems because they inherently must deal with variation in the expression of words and utterances. For instance, the caption *"a dirt biker rides his **motocycle** through the woods"* (fourth row of Figure 4) is highly correlated with the correctly spelled sentences.

4.4 Human evaluation

We ran human evaluations to answer two questions: (1) how much does cropping limit model performance? and (2) how much do retrieval evaluations based only on positive associations underestimate model performance? Hints about both questions can be seen in the qualitative evaluation (Fig. 4).

To answer the first question, Table 3 shows the ratings for ground truth image/caption pairs in the FACC test set. The *uncropped* row shows that overall the captions are high quality and do match the full images. However, human ratings on images *cropped* at the center (which is what is provided to the models) show that there is considerable loss from cropping—only 62.5% of cropped images are rated as good matches by all five raters. Inspection makes it clear why cropping hurts: for example an

	"good" ratings (out of 5)				
	1+	2+	3+	4+	5
Cropped	.949	.918	.874	.800	.625
Uncropped	.995	.994	.989	.971	.891

Table 3: Human evaluation results on ground truth pairs on the test set of FACC, using either center cropped (which the models receive) or uncropped versions of the images.

image of a snowboarder in the air next to another on a ski lift is cropped such that the snowboarder is missing, and thus a poor match to captions mentioning the snowboarder. This clearly indicates that standard cropping (which we follow) inherently limits model performance and that strategies that use the full image should be explored.

Standard retrieval evaluations are blind to pairs that match but are not associated in the data. To address this and answer the second question posed above, we present the top-5 retrieved captions for each image and the top-5 retrieved images for each caption in FACC's test set to human raters. To increase speed and decrease costs, we show raters the original Flickr8k textual captions instead of the spoken ones. Each pair is judged by five raters as "good" or not. This gives a soft measure of the compatibility of each pair based on fast binary judgments from each rater. For retrieval evaluations of a model, we compute recall based on the majority of human raters approving each image-caption pair: R@1 is the percentage of top-1 results and R@5 the percentage of top-5 results that are evaluated as a match by at least 3 of the 5 raters.

Table 4 shows these metrics computed on retrieval outputs from two settings: FACC training from scratch and FACC fine-tuning after CSC pretraining. It also shows the corresponding automatic evaluations from Table 2 for easy comparison. These results make it clear that evaluation based only on positive associations is too rigid: speech-to-image retrieval based on human evaluations shows that a good matching item is returned in 52.2% of cases rather than just the 36.8% indicated by strict corpus matches. For image-to-speech retrieval the 55.8% strict measure goes up to 63.8%. That said, the results also show that the strict measure is nevertheless a useful indicator for comparing relative model performance: the model pretrained on CSC beats the corresponding one trained on FACC from scratch, on both human and automatic evaluations.

		S2I		I2S	
Eval	Pretrain	R@1	R@5	R@1	R@5
Auto		.018	.063	.024	.072
Auto	✓	.139	.368	.182	.558
Humans		.056	.154	.070	.196
Humans	✓	.229	.522	.306	.638

Table 4: Comparison of human rater scores (majority agreement) versus using only corpus-known pairs on all metrics for speech-to-image (S2I) and image-to-speech (I2S) retrieval. Rows with *Auto* evaluation correspond to *Ours (from scratch)* and *Ours (warm-starting all)* scores in Table 2.

5 Conclusion

Large-scale datasets are essential for training deep networks from scratch. In this paper, we present a scalable method for generating an audio caption dataset taking advantage of TTS systems to create millions of data pairs. Using the MMS loss, we demonstrate that pretraining on this dataset greatly improves performance on a human-generated audio caption dataset. As TTS models continue to improve and be developed for more languages, this data augmentation strategy will only become more robust and helpful over time. Finally, using human evaluations, we show evidence that corpus-based retrieval scores underestimate actual performance.

This present work is focused on the here and now since captions describe a snapshot in time and focus on the visual entities and events involved in them. We thus have little hope to learn representations for words like *visit*, *career* and *justice*, for example. Videos can help with process oriented words like *visit* and could get significant components of words like *career* (such as the visual contexts, but not the overall path with intermediate goals involved in careers). They are likely to be hopeless for abstract words like *justice*. To address problems of this sort, there are likely many opportunities to combine ideas from unsupervised term discovery (Kamper et al., 2016; Bansal et al., 2017) with audiovisual word learning (Harwath et al., 2018) and models of visual grounding that have been applied to text (Kiros et al., 2018). Being able to learn effective representations from raw audio associated with images could provide new possibilities for work that learns from videos and text (transcribed speech) (Chen et al., 2018), and in particular open up such techniques to new languages and domains.

Acknowledgements

The authors would like to thank Radu Soricut, Austin Waters, Alex Ku and Jeffrey Ling for the helpful comments that assisted the development of this work.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 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*, pages 6077–6086.

Michele Banko and Eric Brill. 2001. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th annual meeting of the Association for Computational Linguistics*, pages 26–33. Association for Computational Linguistics.

Sameer Bansal, Herman Kamper, Sharon Goldwater, and Adam Lopez. 2017. Weakly supervised spoken term discovery using cross-lingual side information. In *Proceedings of the 42nd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2017)*.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55:409–442.

Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. 2016. Fully-convolutional siamese networks for object tracking. In *European Conference on Computer Vision*, pages 850–865. Springer.

Saurabhch Bhati, Herman Kamper, and K Sri Rama Murty. 2018. Phoneme based embedded segmental k-means for unsupervised term discovery. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5169–5173. IEEE.

Simone Bianco, Remi Cadene, Luigi Celona, and Paolo Napoletano. 2018. Benchmark analysis of representative deep neural network architectures. *IEEE Access*, 6:64270–64277.

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1994. Signature verification using a "siamese" time delay neural network. In *Advances in Neural Information Processing Systems*, pages 737–744.

Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. 2011. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(1):3–5.

Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. 2018. Temporally grounding natural sentence in video. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 162–171, Brussels, Belgium. Association for Computational Linguistics.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Grzegorz Chrupała. 2019. Symbolic inductive bias for visually grounded learning of spoken language. In *To appear in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Grzegorz Chrupała, Lieke Gelderloos, and Afra Alishahi. 2017. Representations of language in a model of visually grounded speech signal. *arXiv preprint arXiv:1702.01991*.

Alexis Conneau, Holger Schwenk, Loïc Barrault, and Yann Lecun. 2017. Very deep convolutional networks for text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1107–1116, Valencia, Spain. Association for Computational Linguistics.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee.

Alon Halevy, Peter Norvig, and Fernando Pereira. 2009. The unreasonable effectiveness of data.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.

David Harwath and James Glass. 2015. Deep multimodal semantic embeddings for speech and images. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 237–244. IEEE.

David Harwath and James Glass. 2019. Towards visually grounded sub-word speech unit discovery. *arXiv preprint arXiv:1902.08213*.

David Harwath, Adria Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. 2018. Jointly discovering visual objects and spoken words from raw sensory input. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 649–665.

David Harwath, Antonio Torralba, and James Glass. 2016. Unsupervised learning of spoken language with visual context. In *Advances in Neural Information Processing Systems*, pages 1858–1866.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yunhsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652*.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

PW Jusczyk and RN Aslin. 1995. Infants detection of the sound patterns of words in fluent speech. *Cognitive psychology*, 29:1–23.

Herman Kamper, Aren Jansen, and Sharon Goldwater. 2016. Unsupervised word segmentation and lexicon discovery using acoustic word embeddings. *IEEE Transactions on Audio, Speech and Language Processing*, 24(4):669679.

Herman Kamper, Aren Jansen, and Sharon Goldwater. 2017a. A segmental framework for fully-unsupervised large-vocabulary speech recognition. *Computer Speech & Language*, 46:154–174.

Herman Kamper, Shane Settle, Gregory Shakhnarovich, and Karen Livescu. 2017b. Visually grounded learning of keyword prediction from untranscribed speech. *arXiv preprint arXiv:1703.08136*.

Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems*, pages 1889–1897.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Jamie Kiros, William Chan, and Geoffrey Hinton. 2018. Illustrative language understanding: Large-scale visual grounding with image search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 922–933, Melbourne, Australia. Association for Computational Linguistics.

Jason Li, Vitaly Lavrukhin, Boris Ginsburg, Ryan Leary, Oleksii Kuchaiev, Jonathan M Cohen, Huyen Nguyen, and Ravi Teja Gadde. 2019. Jasper: An end-to-end convolutional neural acoustic model. *arXiv preprint arXiv:1904.03288*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer.

Jonas Mueller and Aditya Thyagarajan. 2016. Siamese recurrent architectures for learning sentence similarity. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Lucas Ondel, Lukáš Burget, and Jan Černockỳ. 2016. Variational inference for acoustic unit discovery. *Procedia Computer Science*, 81:80–86.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. IEEE.

Alex S Park and James R Glass. 2007. Unsupervised pattern discovery in speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):186–197.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM.

Dan Roth. 2017. Incidental supervision: Moving beyond supervised learning. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Jenny R. Saffran, Elissa L. Newport, and Richard N. Aslin. 1996. Word segmentation: The role of distributional cues. *Journal of Memory and Language*, 35:4:606–621.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 843–852.

Felix Sun, David Harwath, and James Glass. 2016. Look, listen, and decode: Multimodal speech recognition with images. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 573–578. IEEE.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Aäron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *SSW*, 125.

Balakrishnan Varadarajan, Sanjeev Khudanpur, and Emmanuel Dupoux. 2008. Unsupervised learning of acoustic sub-word units. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 165–168. Association for Computational Linguistics.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164.

John Wieting and Kevin Gimpel. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Melbourne, Australia. Association for Computational Linguistics.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*.

Yinfei Yang, Gustavo Hernandez Abrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. *arXiv preprint arXiv:1902.08564*.

Adams Wei Yu, David Dohan, Quoc Le, Thang Luong, Rui Zhao, and Kai Chen. 2018. Fast and accurate reading comprehension by combining self-attention and convolution. In *International Conference on Learning Representations (ICLR)*.