Spoken Moments: Learning Joint Audio-Visual Representations from Video Descriptions

Mathew Monfort*
MIT
mmonfort@mit.edu

SouYoung Jin*
MIT
souyoung@mit.edu

Alexander Liu
MIT
alexhliu@mit.edu

David Harwath
UT Austin
harwath@cs.utexas.edu

Rogerio Feris
IBM Research
rsferis@us.ibm.com

James Glass
MIT
glass@csail.mit.edu

Aude Oliva
MIT
oliva@mit.edu

Abstract

When people observe events, they are able to abstract key information and build concise summaries of what is happening. These summaries include contextual and semantic information describing the important high-level details (what, where, who and how) of the observed event and exclude background information that is deemed unimportant to the observer. With this in mind, the descriptions people generate for videos of different dynamic events can greatly improve our understanding of the key information of interest in each video. These descriptions can be captured in captions that provide expanded attributes for video labeling (e.g. actions/objects/scenes/sentiment/etc.) while allowing us to gain new insight into what people find important or necessary to summarize specific events. Existing caption datasets for video understanding are either small in scale or restricted to a specific domain. To address this, we present the Spoken Moments (S-MiT) dataset of 500k spoken captions each attributed to a unique short video depicting a broad range of different events. We collect our descriptions using audio recordings to ensure that they remain as natural and concise as possible while allowing us to scale the size of a large classification dataset. In order to utilize our proposed dataset, we present a novel Adaptive Mean Margin (AMM) approach to contrastive learning and evaluate our models on video/caption retrieval on multiple datasets. We show that our AMM approach consistently improves our results and that models trained on our Spoken Moments dataset generalize better than those trained on other video-caption datasets.

http://moments.csail.mit.edu/spoken.html

*equal contribution

1. Introduction

Video understanding has typically been focused on action recognition and object tracking as the temporal aspect of videos lends itself strongly to the task of representing motion, a key component of an action. Breaking down video analysis to simple tasks, such as action recognition, allows for efficient data annotation for building large datasets to train deep learning models [31, 45, 21] which has been extremely successful for images with object annotations [34]. A main difficulty is that, in contrast to an image, a video often captures an interaction between agents and objects that evolves over time. These interactions can be as simple as "a person picking up a glass of water", but even in this case three different objects ("person", "glass" and "water") are included in the interaction. Additionally, the video may also continue to depict the "person drinking from a glass" and the "person putting the glass back down on the table". These sequential events present additional challenges for video datasets where single annotations may not be sufficient to explain the events depicted. Multi-label approaches to video annotation have attempted to address this problem by labeling multiple actions in a video [46, 22, 72]. However, these methods focus on single domain annotations, such as actions or objects, and do not capture additional contextual information, such as "person *angrily* putting down the *dirty* glass on a *rusted* table", which can change the interpretation of an event and how it fits into a sequence of observations.

A solution for capturing more fully the content of video is to annotate multiple actions or objects in each video [22, 71, 46, 49]. However labels like "drinking", "glass", only provide a portion of the information needed to interpret the veracity of the event. Additional narratives may include intuitive descriptions and intentions, such as "an exhausted man picks up a dirty glass of water and drinks from

it before angrily putting it down on a table" which would dramatically change the event interpretation. The full lingual description combines these actions with adjectives and nouns (objects) that contextualize the events depicted leading to a better understanding of the video. This is our goal in providing a new large scale dataset for training models for full video understanding.

We introduce a large scale video caption dataset, Spoken Moments in Time (S-MiT), to allow large deep learning models for video understanding to learn contextual information. Most existing video description datasets [70, 59, 32, 20, 79] are limited in size when compared to the large datasets for action recognition [31, 45, 21]. A likely cause is the increased cost of collecting full text descriptions for videos compared to single label annotations. Recent work in image captioning [25] addressed this problem by collecting audio descriptions for a large set of images from the Places dataset [76]. Collecting spoken captions is faster and more efficient due to the low overhead of speaking compared to typing. In addition, recording of spontaneous speech rather than typed text can produce more natural descriptions of an event.An automatic speech recognition (ASR) system was then used to transcribe the spoken descriptions to text captions. In this work, both audio, text and video models were jointly trained via contrastive learning to learn joint cross-modal representations. We build on this approach and compare models that learn directly from the spoken captions to models that include a trained ASR model which feeds generated text transcriptions into an NLP language model. We then jointly train caption and visual models (based on concatenated video and image features) using a novel Adaptive Mean Margin (AMM) approach to contrastive learning to align the visual and caption representations. We evaluate our models on multiple datasets for video/caption retrieval and show that a model trained using AMM on S-MiT achieves the best general performance across four datasets.

Altogether, our novel contributions include:

1. The large-scale **Spoken Moments in Time dataset** (S-MiT) which includes 500k pairs of video clips and corresponding audio descriptions. This new dataset represents the largest video description dataset available and will serve as a new benchmark for the community.

2. **Benchmark models** with aligned spoken caption and video representations learned via contrastive learning. We compare approaches that learn directly from the spoken descriptions as well as approaches that include ASR transcriptions that feed into different language models to generate caption representations.

3. An **Adaptive Mean Margin** (AMM) approach to cross-entropy based contrastive learning.

2. Related work

2.1. Video Understanding

The field of video understanding has recently seen fast progress partly due to the availability of large scale video datasets including ActivityNet [6], Kinetics [31], Moments in Time [45, 46] and YouTube-8M [1]. These large datasets are used to pretrain models that are fine-tuned on smaller action recognition datasets such as UCF101 [61] and HMDB [35]. With the increased availability of large scale video datasets, many different models have been proposed to improve performance on a number of video understanding tasks. Two-stream convolutional neural networks (CNNs) combine optical flow with RGB frames to capture both temporal and spatial information [60]. I3D models [8] combine 3D CNNs [64], which use a 3D kernel to learn temporal information from a frame sequence, with optical flow to form a two-stream 3D network "inflated" from 2D filters pre-trained on ImageNet [16]. More recently a temporal shift module has been proposed to integrate temporal information into 2D models by shifting frame representations across the temporal dimension [39].

Recently multi-modal visual understanding methods have received significant attention [25, 63, 43, 66, 29, 4]. The DAVEnet model [25] has been proposed for jointly learning aligned representations between images and spoken captions, and has been extended to align frame-wise video representations with synchronized audio narration for cross-modal audio-visual concept learning [4]. Here, we build on the motivation from this paper and **learn aligned representations between videos and unsynchronized spoken descriptions** using the S-MiT Dataset.

2.2. Caption Datasets

There have been a number of different datasets released for providing language descriptions of visual information. Flickr8k [28] and Flickr30k [48] include 8k and 30k images respectively each sourced from Flickr. Each image is associated with 5 text captions describing what is in the image. An additional set of 5 audio captions per image in both sets was recently collected for learning joint embeddings between speech and images [25]. The Visual Genome dataset [33] includes captions for multiple regions of more than 180k images allowing for fine-grained descriptions of each image. The Places Audio Caption dataset [26] contains approximately 400k images from the Places 205 [77] image dataset with audio captions of people verbally describing each image. MS COCO [11] is a large image dataset for object recognition, segmentation, and captioning which includes roughly 1 million captions for 160k Flickr images. Conceptual Captions [58] contains 3.3M images with captions generated from HTML attributes associated with web based images. The Stock3M dataset [69] includes 3.2 mil-

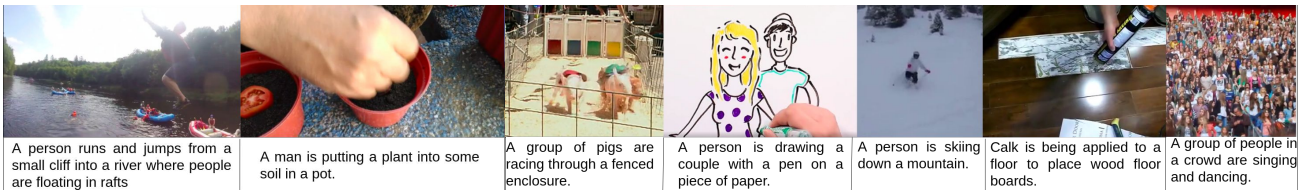

Figure 1: **Examples from the Spoken Moments Dataset:** The dataset is composed of videos and the corresponding spoken captions. We show some examples of the text transcriptions, automatically generated using the public Google ASR engine.

lion images each with a crowdsourced caption.

Beyond the numerous datasets available or image captioning [28, 48, 33, 11, 58, 69], including those that provide spoken descriptions [26, 25], there are a variety of different video caption datasets available. A number of these datasets are related to cooking [50, 51, 54, 13, 12] including YouCook [14] and YouCook II [79] which include 2k videos from YouTube each with multiple captions annotated at different segments of each video. MPII-Movie Description Corpus [52] contains transcribed audio descriptions from 94 Hollywood movies split into 68k clips where each clip is paired with a sentence from the movie scripts and an audio description of the visual content in each clip. Similarly, Large Scale Movie Description Challenge (LSMDC) dataset [53] contains 200 movies with 120K sentence descriptions. VideoStory [20] contains 20k social media videos where each video contains a paragraph length description. The ActivityNet Captions dataset [32] has 20k videos with 100k text descriptions. The Microsoft Video Description (MSVD) dataset [9] contains 2k YouTube clips with a 10-25 second duration and an average of 41 single sentence descriptions per clip. MSR-Video to Text (MSR-VTT) [70] contains 7k videos split into 10k clips with 20 captions per video.

HowTo100M [44] contains 136 million clips sourced from 1.22 million instructional videos with narrations generated from subtitles associated with each video. However, the subtitles are not human verified captions and the content is constrained to instructional videos. Since the text associated with the clips in the HowTo100M dataset are transcriptions of a narrator completing a task in the video, the short text phrases from the subtitles occasionally share noisy associations with the reference clip. In Section 5, and Table 2, we decided to compare our contributions using strict caption datasets as we are proposing a large-scale human annotated caption dataset with full human generated descriptions for each video.

VaTeX [68] contains 41k videos sourced from the Kinetics-600 dataset [31, 7] annotated with 10 English captions and 10 Chinese captions for multilingual captioning. VaTeX is the most similar to our proposed dataset in that it is sourced from an existing video dataset for action recognition and the captions are directly annotated.

In this work, we present a new dataset, *Spoken Moments in Time (S-MiT)*, which includes spoken audio captions for 500k unique three second clips each with different source videos from the Moments in Time dataset [45, 46]. In addition to vast increase in scale over other video-caption datasets, a major contribution is that we are using spoken descriptions rather than text. This allows us to train spoken caption models to directly align with video models. This is not possible with the other large video caption datasets and allows for spoken caption models to be analyzed with matching video information. We also show that models trained on our S-MiT dataset generalize much better in retrieval to the video-caption pairs in other datasets. This is due to the large coverage, diversity and scale of our proposed dataset.

2.3. Cross Modal Contrastive Learning

Cross modal learning has been used to jointly self-supervise audio-visual models [3, 47, 75] with synchronized information while NLP approaches have been leveraged to align joint representations for both visual and language modalities using spoken and text descriptions [2, 78]. This is typically done via Contrastive Learning where the alignment between positive pairs (language and visual input) is trained to be stronger than those of non-positive pairs [24]. For visual representations, a triplet based max-margin loss is commonly used to discriminate representations between positive and negative pairs [73, 74, 18]. Semi-hard negative mining [57] and a dot-product based similarity score have been used to jointly learn audio-visual embeddings between images and spoken captions [25] while batch-wise cross-entropy approaches to contrastive learning have been used to increase the amount of information utilized in learning by considering all negative examples in a mini-batch [65, 10]. Work on bidirectional speech/image retrieval using audio descriptions of images integrated ideas from max-margin contrastive learning and added a margin into the cross-entropy loss [29]. SimCLR [10] added a non-linear projection head that maps paired representations into a common space allowing for stronger representations.

A pretrained language model has recently been used to improve cross-modality learning with language and visual input pairs. VilBERT [42] added a pretrained BERT

[17] transformer to capture semantic language representations associated with object detection proposals from a pretrained faster RCNN network. VideoBERT [62] extended BERT to jointly learn the visual and linguistic domain by generating tokenized visual words. Inspired by this prior work, we propose adding a **pretrained language model that maps word predictions from a trained ASR model to semantic language features** in order to generate rich spoken caption representations. We then utilize an MLP to project these caption representations, and our video representations, to an aligned joint representation which can be used for video/caption retrieval (see Section 5).

2.3.1 Optimization Approaches

A common approach to optimization in contrastive learning settings is to use a similarity based loss function. We formulate the contrastive loss as, $\mathcal{L} = \mathcal{L}_{vc} + \mathcal{L}_{cv}$, where the goal is to maximize the discrimination between positive and negative paired captions c and videos v. The loss is split into two tasks where $\mathcal{L}_{vc}$ forms pairs from a fixed video and each caption in a sampled mini-batch, while $\mathcal{L}_{cv}$ fixes the caption and forms pairs with each video in the mini-batch. Below we discuss different approaches of $\mathcal{L}_{xy}$, where x and y are interchangeable with v and c.

Semi-hard negative mining (SHN) [57] has been used for learning aligned cross-modal embeddings using a triplet loss [25, 30]. This is an improvement over hard negative mining [19] since a sampled negative example is constrained to be less similar to the anchor than the positive sample while still being within the margin and thus contributing a loss at each step with the margin $M = 1$, $\mathcal{L}_{xy} = \max(\mathcal{S}(x_i, y_j) - \mathcal{S}(x_i, y_i) + M, 0)$, where $\mathcal{S}(x_i, y_j)$ is a similarity score for the representations of x_i and y_j, with x_i and y_i forming a positive pair.

Noise contrastive estimation (NCE) [23] has been applied to contrastive learning [10, 65] by using a log-likelihood based loss function that learns to discriminate between positive and negative pairs of feature embeddings,

$$\mathcal{L}_{xy} = -\frac{1}{B}\sum_{i=1}^{B} log \frac{e^{\mathcal{S}(x_i, y_i)}}{\sum_{j=1}^{B} I_{i \neq j} e^{\mathcal{S}(x_i, y_j)}}, \quad (1)$$

where $I_{i \neq j}$ is an indicator function that we only considers negative pairs in the denominator. This has been shown to improve feature alignment compared to SHN [10].

Masked Margin Softmax Loss (MMS) [29] and Large Margin Cosine Loss (LMCL) [67] incorporate a positive margin into the contrastive learning framework in order to improve feature discrimination among non-paired embeddings. MMS uses a monotonically increasing margin to allow for initial learning to begin to converge before a large alteration to the loss is added. LMCL proposes a theoretical limit on the maximum margin size of $1 - cos\frac{2\pi}{N}$ where N refers to the number of classes being discriminated. For aligning captions to visual information, the class size can be considered unbounded as each caption represents a slightly different representation that we want to discriminate leading to a max margin size of 1. Concretely, MMS proposes adding a margin to Equation 1,

$$\mathcal{L}_{xy} = -\frac{1}{B}\sum_{i=1}^{B} log \frac{e^{\mathcal{S}(x_i, y_i) - M}}{e^{\mathcal{S}(x_i, y_i) - M} + \sum_{j=1}^{B} I_{i \neq j} e^{\mathcal{S}(x_i, y_j)}}, \quad (2)$$

where the margin, M, starts as 0.001 and is exponentially increased by a factor of 1.002 every 1000 training steps.

We propose extending the idea of an increasing margin in MMS to an adaptive setting that does not require setting the initial value of the margin or the growth rate. We refer to this approach as an Adaptive Mean Margin (AMM) where the margin is set as the mean distance between the positive pair and the set of negative pairs in a batch. We describe AMM in more detail in Section 4.3.

3. The Spoken Moments Dataset

We begin with the Moments in Time dataset [45] as it includes over 1 million videos sourced from a number of different video hosting sites with strong inter & intra-varietal variation in terms of the number of events depicted in each video. Further, the videos are all cut to 3 seconds allowing for a concise description to effectively capture the localized information of each event. Here we refer to concise descriptions as those that focus on key events depicted in the video and does not imply partial descriptions. In data collection, annotators may watch a video as many times as desired. During recording, we block the annotators from seeing/hearing the video to encourage descriptions of important memorable events rather than every specific detail. This approach does not preclude the annotators from describing sequential or simultaneous events as shown in our qualitative examples (see Figure 1). We describe our annotation approach in more detail in the supplementary material.

3.1. Dataset Statistics

Our proposed Spoken Moments dataset contains 500k videos randomly chosen from the Multi-Moments in Time (M-MiT) training set and all of the 10k videos from the validation set. Each video in the training set contains at least one audio description. We transcribed each audio recording using the public Google Automatic Speech Recognition (ASR) engine to generate text captions for each video. When analyzing these transcriptions, we build a picture of the coverage and diversity of our captions. Table 2 (left) shows that our captions have an average length of 18 words with a unique vocabulary of 50,570 words consisting of 20,645 nouns, 12,523 adjectives and 7,436 verbs with a

Type	Total	Average	Unique
Words	5,618,064	18.01	50,570
Verbs	492,941	1.58	7,436
Nouns	1,365,305	4.37	20,645
Adjectives	386,039	1.24	12,523

Type	Dataset	Coverage
Objects	ImageNet	69.2%
	MS-COCO	100%
Actions	Kinetics	85.1%
	Moments in Time	96.2%
Scenes	Places365	47.4%

Dataset	Clips	Videos	Captions	Words	Vocab	Domain	Spoken
TACoS [50]	7,206	127	18,227	146,771	28,292	Cooking	
YouCook II [79]	15,400	2,000	15,400	121,418	2,583	Cooking	
MSVD [9]	1,970	1,970	70,028	607,339	13,010	General	
Charades [59]	10,000	10,000	27,800	645,636	32,804	General	
MPII-MD [52]	68,337	94	68,375	653,467	24,549	General	
MSR-VTT [70]	10,000	7,180	200,000	1,856,523	29,316	General	
ActivityNet Captions [32]	100,000	20,000	100,000	1,348,000	15,564	General	
VideoStory [20]	123,000	20,000	123,000	1,633,226	-	General	
Epic-Kitchens [13, 12]	76,885	633	76,885	227,974	1,737	Cooking	
Vatex-en [68]	41,300	41,300	413,000	4,994,768	44,103	General	
Spoken Moments	**515,912**	**459,742**	**515,912**	**5,618,064**	**50,570**	**General**	✓

Figure 2: **Dataset Statistics:** On the top-left we show the total and average number of words, verbs, nouns and adjectives in our captions as well as the number of unique examples of each. On the bottom-left we show the percentage of the class vocabulary from different datasets that occur in our captions. On the right we compare our proposed Spoken Moments dataset to existing video caption datasets. The word count and vocabulary for S-MiT are generated using ASR transcriptions.

total word count of 5.6 million. Table 2 (right) shows a comparison of our Spoken Moments dataset to other existing datasets for video captioning. Our dataset will be the largest public dataset in terms of video clips, source videos, total number of captions, total words in the captions and the vocabulary set of unique words occurring in the captions. The increase in vocabulary size is important as it shows that our increase in the number of videos over previous datasets does not simply include repeated events but covers a novel breadth of information. We can see the opposite effect of this in YouCook II [79] where the restricted domain of cooking videos results in a limited vocabulary used in the descriptions.

To understand how this vocabulary covers the class labels typically used for training computer vision models, we examined whether these labels exist in our vocabulary. Table 2 (right) shows that we have strong coverage of the two largest action recognition datasets for video understanding (Kinetics [31] and M-MiT [46]). We expected a large coverage of the events in M-MiT as we sourced our videos from this dataset and the action labels themselves are fairly general (e.g. "running" and "cooking"). For Kinetics, the labels are commonly tied to a noun preceded by a verb (e.g. "brushing hair"). For these labels we consider them to exist in our dataset if both the verb and noun are in the same caption. For example, "A boy is in a bathroom brushing his teeth" would cover the class "brushing teeth". With this approach we see a 85.1% coverage of the classes in Kinetics and a 96.2% coverage of the classes in M-MiT showing a strong level of event diversity. Similarly we see a strong overlap of the object classes in MS-COCO [40] (100%) and ImageNet [16] (69.2%) in our captions. ImageNet coverage is likely lower due to the specific labels used for many of its classes (e.g. "coucal"). Still, 69.2% coverage means 692 ImageNet classes appear in our captions. Similarly, Places [77] scene labels are very specific and don't necessarily match the language used in our descriptions. For example, an "abbey" will typically be described as a "church" or "monastery" in our captions. We did not account for all of the synonyms possible and are only considering direct matches in our captions. Even so we are able to find a 47.4% coverage of the scene labels in Places365 in our dataset.

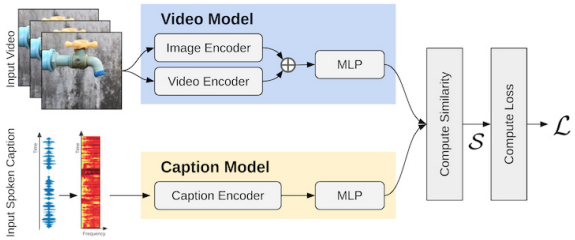

Figure 3: **Architecture:** Videos and captions are fed into the video/caption models where the outputs are used to compute a similarity matrix, $\mathcal{S}$, which is used to compute a loss, $\mathcal{L}$.

Here we provide information on some additional characteristics of our data that may be of interest. While we do not release demographic info of our annotators or captions, about 57% of the spoken captions were recorded by male voices and 43% female. For the audio streams of the videos, roughly 51% include natural sound, 5% have music as the audio and 44% have no audio. This is consistent with the M-MiT dataset [46] from which we source our videos. Additionally, we found that less than 3% of the videos contain captions that describe non-visible events (e.g. a car horn when no car is visible in the video frames). For this reason we have chosen to focus our approach on learning a strong visual model in Section 4.

4. Learning Audio-Visual Representations

In order to learn from the large set of spoken captions in the proposed S-MiT dataset, we adopt a cross-modal architecture used in prior work [44, 25, 55] which is composed of a video model and a caption model as depicted in Figure 3. Specifically, we take N video-caption pairs as input and encode each modality into a 4096-D feature vector. We do this by adding a multilayer perceptron (MLP) as a projection head on top of both the video and the caption model. This projection head is composed of two linear layers followed by gated linear units (GLU) [15]. We then compute the dot product between the video and caption representa-

tions to produce an NxN similarity matrix, $\mathcal{S}$, which is used to compute our contrastive loss for training. In Section 4.3, we describe our modified approach to margined contrastive learning which uses an Adaptive Mean Margin (AMM) which automatically adjusts itself during learning to improve the optimization signal during training.

4.1. Video Model

Following prior work [44], we use two encoders to represent input videos: image & video encoders. Specifically, we use a ResNet-152 [27] pretrained on ImageNet [34] and a temporal shift module (TSM) ResNet-50 model [39] pretrained on M-MiT [46]. Each encoder outputs a 2048-D feature vector after max-pooling over the temporal dimension (8 frames for the TSM ($\sim$3 fps) and 3 frames for the image model (1 fps)). We concatenate the two 2048-D vectors and feed the concatenated vector into an MLP projection head to get the final 4096-D visual representation. We examine the effect of using the image and video encoders as well as different pretrained models in the supplementary material.

4.2. Caption Model

4.2.1 Language Caption Model

Prior work in learning joint representations between audio captions and visual models has shown that utilizing ASR transcriptions greatly improves results [25]. We build on this idea and use the predicted words from a pretrained ASR model (e.g. Google's public ASR engine) to train our models. Concretely, we examine the effect of using different pretrained language models stacked on top of the ASR model predictions. We begin by comparing the results of using Fasttext [5], BART [36] and BERT [17] models to generate semantic and contextual word representations for our captions. During training, we randomly select 10 words from each caption to be included in training. In the case of the BART and BERT models, this selection happens after the full transformer model has been applied to avoid altering the results from the self-attention mechanisms. If less than 10 words occur in a caption then we allow words to be sampled multiple times in the random selection. This training augmentation allows different words in each caption to be represented differently at different training iterations. We examine the effect of this approach in the supplementary material. In test, we use the full transcription as input into the language model. We average the word representations from the output of the language model to generate a single representation for each caption which we align to the video representations described in the previous section.

4.2.2 Spoken Caption Model

We also train caption models with raw spoken captions instead of the corresponding transcription. For each caption, we randomly sample 10 seconds of speech for training and compute the 40-dimensional log Mel spectrogram to serve as the input of spoken caption model. The input is fed into a spoken caption model where we consider ResDavenet [25] (which is designed specifically for speech) and two ImageNet ResNet [27] models (ResNet-34, ResNet-50). For the ResNet models, we modify the first convolutional layer to take the 1-channel input so that spectrogram can be processed. In addition, the wav2vec [56] model, which takes raw waveform as the input, is also involved in our experiments. Spoken captions are first fed into the pre-trained wav2vec model, which produces 512-D vectors per 210 ms. We then feed them into a learnable ResStack, taken from ResDavenet, to learn representations of spoken captions.

4.3. Adaptive Mean Margin

We train our model using the contrastive loss with a similar setting to MMS (Equation 2). The only difference is that we replace the margin, M, with an adaptive margin based on the difference between the similarity of the positive pair and the set of negative pairs in each batch.

The challenge in using the MMS margin for mini-batch sampled contrastive learning is that the initial margin and growth schedule are difficult to tune for a specific dataset and similarity metric. Additionally, depending on the sampled pairs in a mini-batch, the margin calculated may be too weak if the positive pair is much more similar than the sampled negative pairs and too strong if it is very similar to the negative pairs. The approach to monotonically increase the margin during training is meant to address this as the positive and negative pairs will share similar alignment early in training and begin to diverge closer to convergence. However, variable rates of convergence of different models on different datasets make this growth rate difficult to tune and this approach does not account for differences in the negative samples that appear in different mini-batches. To address this, we propose an adaptive margin based on relative batch-wise similarity scores.

Class labels have been proposed to be used for generating adaptive margins based on class similarity between positive and negative pairs [37, 41]. Likewise, prior work explored a non-class dependant approach for an adaptive similarity-based margin for human pose estimation [38] where the mean joint error between a positive pose and a hard sampled negative pose was used as a margin with the triplet loss. This adaptively increases the margin when the sampled negative pair is dissimilar to the positive pair in order to maximize the learning signal on less aligned negative samples. We follow a similar intuition and simply replace M in Equation 2 with

$$M_{xy} = \alpha\big(\mathcal{S}(x_i, y_i) - \frac{1}{B-1}\sum_{j=1}^{B} I_{i\neq j}\mathcal{S}(x_i, y_j)\big), \qquad (3)$$

Language Caption Model	Caption to Video				Video to Caption				Mean			
	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP
Fasttext [5]	17.1±0.8	44.0±0.6	57.2±0.5	30.2±0.5	24.1±0.5	49.9±0.6	61.8±1.3	36.6±0.3	20.6±0.5	46.9±0.6	59.5±0.8	33.4±0.4
BERT [17]	25.9±0.6	55.5±1.2	67.0±1.1	39.7±0.7	33.3±1.4	62.1±1.0	72.0±0.6	46.5±1.2	29.6±0.8	58.8±1.0	69.5±0.8	43.1±0.8
BART [36]	**33.1**±0.9	**65.5**±1.5	**76.6**±1.3	**47.8**±1.1	**43.8**±0.7	**71.5**±1.2	**80.9**±1.6	**56.4**±0.7	**38.4**±0.4	**68.5**±1.3	**78.7**±1.4	**52.1**±0.8

Table 1: **Language Caption Model Comparison on Video/Caption Retrieval:** Here we compare the video/caption retrieval results on the test set of the Spoken Moments dataset using models trained with three different language models.

Dataset	Loss	Caption to Video				Video to Caption				Mean			
		R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP
Vatex [68]	NCE	43.6±1.4	77.4±1.4	86.5±1.4	58.4±1.2	39.4±1.3	74.3±1.0	84.7±0.8	54.7±1.0	41.5±1.2	75.8±1.1	85.6±1.1	56.5±1.0
	SHN	19.6±1.4	50.2±1.5	63.9±0.6	33.8±1.1	22.9±1.0	54.0±0.9	68.8±1.2	37.6±0.9	21.3±0.9	52.1±0.8	66.3±0.8	35.7±0.7
	MMS	46.2±1.5	79.7±0.8	88.1±0.8	60.7±1.0	42.0±0.7	**77.7**±0.7	**86.8**±0.3	57.5±0.6	44.1±1.1	78.7±0.7	87.4±0.5	59.1±0.7
	AMM	**48.7**±1.4	**82.0**±0.9	**89.3**±1.1	**63.0**±1.0	**43.0**±0.7	77.4±1.1	85.8±0.7	**58.3**±0.6	**45.9**±1.0	**79.7**±0.4	**87.5**±0.8	**60.7**±0.6
ActivityNet [32]	NCE	11.8±0.6	35.4±1.0	50.6±0.8	23.8±0.4	16.7±0.8	43.0±1.2	57.1±1.2	29.5±0.8	14.3±0.6	39.2±0.8	53.8±1.0	26.7±0.5
	SHN	9.9±0.9	31.2±1.3	45.2±0.9	20.9±0.9	13.7±1.1	38.5±0.9	53.4±0.9	25.9±1.0	11.8±0.9	34.9±0.8	49.3±0.7	23.4±0.9
	MMS	12.0±0.7	35.5±1.0	49.2±0.8	23.9±0.6	16.2±0.4	42.4±0.9	56.5±1.6	28.8±0.6	14.1±0.4	39.0±0.2	52.8±1.2	26.4±0.2
	AMM	**17.2**±1.1	**46.1**±1.4	**60.0**±0.8	**30.6**±0.6	**20.9**±1.1	**50.1**±1.3	**62.4**±0.8	**34.3**±0.6	**19.1**±1.0	**48.1**±1.2	**61.2**±0.6	**32.5**±0.6
MSR-VTT [70]	NCE	20.7±0.9	51.0±0.7	66.6±1.2	35.0±0.4	30.7±1.4	65.1±0.7	78.2±1.3	46.1±1.2	25.7±1.0	58.1±0.6	72.4±1.2	40.6±0.7
	SHN	11.3±0.2	32.0±1.0	44.9±1.4	21.9±0.3	22.1±0.9	54.5±1.6	68.9±1.4	37.0±1.1	16.7±0.5	43.3±0.5	56.9±0.9	29.5±0.5
	MMS	17.6±1.1	46.5±0.9	61.6±0.9	31.5±0.6	28.3±1.1	63.1±1.4	76.1±0.9	43.8±1.1	23.0±0.9	54.8±0.6	68.9±0.7	37.6±0.6
	AMM	**25.7**±0.8	**61.0**±0.8	**75.6**±0.7	**41.6**±0.6	**32.5**±1.5	**67.5**±1.7	**80.1**±1.4	**48.0**±1.2	**29.1**±0.8	**64.2**±1.0	**77.9**±1.0	**44.8**±0.8
S-MiT	NCE	**33.1**±0.9	**66.9**±1.9	**77.6**±1.2	**47.9**±0.7	43.0±0.8	**71.8**±0.9	80.7±1.2	55.8±0.7	38.0±0.5	**69.3**±1.4	**79.1**±1.1	51.8±0.6
	SHN	23.1±1.3	55.4±1.6	69.3±1.3	37.7±1.1	41.4±1.1	70.8±0.9	79.5±1.0	54.5±0.7	32.3±0.9	63.1±1.1	74.4±1.1	46.1±0.8
	MMS	26.5±1.3	58.3±1.4	72.0±0.9	41.1±1.1	43.3±1.3	71.2±1.4	79.9±0.8	55.8±1.2	34.9±1.2	64.8±1.2	76.0±0.8	48.5±1.1
	AMM	**33.1**±0.9	65.5±1.5	76.6±1.3	47.8±1.1	**43.8**±0.7	71.5±1.2	**80.9**±1.6	**56.4**±0.7	**38.4**±0.4	68.5±1.3	78.7±1.4	**52.1**±0.8

Table 2: **Loss Function Comparison for Video/Caption Retrieval:** Models trained on four datasets with different loss functions are compared. The proposed AMM loss function consistently achieves the best performance.

Spoken Caption Model	Loss	Caption to Video				Video to Caption				Mean			
		R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP
ResDavenet [25]	NCE	30.7±0.6	57.1±0.6	67.6±1.0	42.9±0.8	29.3±1.0	55.8±1.2	66.2±1.4	41.8±0.9	30.0±0.8	56.4±0.9	66.9±1.2	42.3±0.8
	SHN	30.2±1.1	56.9±0.8	66.8±0.5	42.6±1.0	31.0±1.2	57.2±0.8	67.1±0.9	43.2±1.0	30.6±1.1	57.0±0.8	67.0±0.7	42.9±1.0
	MMS	32.1±1.1	58.9±1.0	68.6±1.5	44.4±0.8	32.3±1.3	57.9±1.1	68.1±1.5	44.3±1.2	32.2±1.2	58.4±1.0	68.4±1.5	44.3±1.0
	AMM	**34.8**±1.1	**62.0**±1.1	**70.4**±1.2	**47.0**±1.1	**34.6**±1.5	**60.8**±1.6	**70.0**±0.9	**46.8**±1.2	**34.7**±1.2	**61.4**±1.4	**70.2**±1.1	**46.9**±1.1
Wav2Vec [56]	NCE	32.6±0.7	60.4±0.8	70.3±1.6	45.3±0.8	30.9±1.0	59.6±0.9	69.8±1.1	43.9±0.8	31.8±0.7	60.0±0.8	70.0±1.3	44.6±0.8
	SHN	27.8±1.0	54.2±1.7	64.9±1.8	40.1±1.0	28.4±0.7	53.7±1.6	64.2±1.7	40.4±0.8	28.1±0.8	53.9±1.6	64.6±1.7	40.2±0.9
	MMS	33.6±0.6	60.5±1.2	**71.4**±1.1	46.1±0.7	33.4±1.0	60.5±1.7	**70.3**±1.1	45.7±0.8	33.5±0.6	60.5±1.4	**70.8**±1.1	45.9±0.7
	AMM	**35.0**±0.4	**61.7**±0.9	71.0±0.9	**47.1**±0.6	**34.7**±1.5	**61.1**±0.9	70.2±0.9	**46.8**±1.2	**34.8**±0.9	**61.4**±0.9	70.6±0.8	**47.0**±0.9
ResNet-34	NCE	32.2±1.3	59.7±1.4	70.3±1.3	44.8±1.1	32.8±1.8	58.8±1.3	69.2±1.9	45.1±1.4	32.5±1.4	59.2±1.3	69.7±1.5	45.0±1.2
	SHN	32.7±1.1	60.3±1.3	71.0±1.1	45.5±1.0	33.1±1.0	60.1±1.5	70.1±1.3	45.6±0.9	32.9±1.0	60.2±1.4	70.6±1.2	45.6±0.9
	MMS	35.3±1.0	62.5±1.2	72.8±1.8	47.7±0.6	36.7±0.9	62.2±0.8	72.1±1.6	48.6±0.9	36.0±0.7	62.3±1.0	72.5±1.6	48.2±0.7
	AMM	**36.3**±0.5	**63.9**±1.7	**73.7**±1.6	**48.9**±0.8	**37.5**±1.7	**63.5**±1.9	**73.7**±1.6	**49.6**±1.5	**36.9**±1.1	**63.7**±1.7	**73.7**±1.5	**49.2**±1.2
ResNet-50	NCE	32.7±0.6	60.8±1.9	70.6±1.6	45.6±0.8	33.1±1.0	59.4±1.5	69.6±1.4	45.5±0.9	32.9±0.5	60.1±1.7	70.1±1.4	45.5±0.8
	SHN	33.9±0.6	60.1±1.4	70.9±1.3	45.8±0.7	34.0±1.2	60.6±1.8	70.1±1.4	46.0±1.1	34.0±0.8	60.3±1.5	70.5±1.3	45.9±0.8
	MMS	37.2±0.9	65.4±0.6	75.1±1.3	50.0±0.7	37.8±1.3	64.6±1.1	74.2±0.9	50.1±1.1	37.5±1.0	65.0±0.8	74.7±1.1	50.0±0.9
	AMM	**39.5**±1.3	**65.7**±1.5	**75.5**±1.3	**51.6**±1.1	**40.1**±0.7	**66.3**±1.1	**74.5**±1.2	**52.0**±0.7	**39.8**±0.9	**66.0**±1.2	**75.0**±1.1	**51.8**±0.8

Table 3: **Spoken Caption Model Comparison:** Models trained with different spoken caption architectures and different loss functions are compared for video/caption retrieval on the S-MiT test set. The proposed AMM loss function consistently achieves the highest performance while ResNet-50 is found to be significantly stronger than the other architectures.

Trained On	Evaluated On																			
	Vatex				ActivityNet				MSR-VTT				S-MiT				Mean			
	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP
Vatex	**45.9**	**79.7**	**87.5**	**60.7**	15.6	39.4	51.7	27.1	22.6	49.8	63.2	35.6	13.1	33.0	45.8	23.5	24.3	50.5	62.1	36.7
ActivityNet	25.0	56.0	68.4	39.1	**19.1**	**48.1**	**61.2**	**32.5**	15.1	37.1	50.4	26.4	9.8	28.7	40.6	19.7	17.3	42.5	55.2	29.4
MSR-VTT	21.0	51.3	64.8	35.1	9.9	28.3	39.7	19.6	29.1	64.2	**77.9**	44.8	14.6	39.3	53.4	26.9	18.7	45.8	59.0	31.6
S-MiT	42.7	75.4	84.2	57.1	17.6	41.6	53.8	29.2	**33.1**	**64.8**	77.4	**47.6**	**38.4**	**68.5**	**78.7**	**52.1**	**33.0**	**62.6**	**73.5**	**46.5**

Table 4: **Cross Dataset Evaluation on Video/Caption Retrieval:** Here we compare the generalization performance of models trained on four different datasets for video/caption retrieval. Each model is trained on a single dataset and we average the evaluation on five 1k video-caption samples from the test set of each other dataset. We additionally show the mean performance accross datasets. The S-MiT model shows it generalizes very strongly to the other datasets even beating the MSR-VTT model on its own test set.

where α is a dampening parameter to weight the strength of the margin. When M_{xy} in Equation 3 is applied to Equation 2 with $\alpha = 1$, the margin removes the positive pair similarity from the optimization. Ablation studies on different alpha values can be found in the supplementary material. In practice we use $\alpha = 0.5$ in our experiments.

This has the effect of increasing the margin as the difference between the true pair similarity and the similarity of the negative pairs increases. As the training progresses, and the learning approaches convergence, the margin generally increases with the increased separation between positive and negative pair-wise similarities. This also removes the need to tune the margin and growth rate which may have different optimal values for different similarity met-

rics, batch sizes and datasets.

We refer to this as an **Adaptive Mean Margin** (AMM) for contrastive learning and show in Section 5 the effect of applying this adaptive margin.

5. Results

5.1. Video/Caption Retrieval

In Tables 1, 2 and 3 we show results of R@k recall scores (for $k = 1, 5, 10$) and mean average precision (mAP) on both caption to video and video to caption retrieval. Results are averaged over five random sets of 1k video-caption pairs from the test set. Each model in Tables 1 and 2 uses the output of a pretrained ASR model, the Google Cloud ASR engine, as input into a trained language model to generate a feature representation for each caption. Alternatively, the spoken caption models align visual representations directly from the audio signal without pretrained modules.

Table 1 shows the result of using different language models to generate our caption representations from ASR text transcriptions. Each of these models was trained using the proposed AMM loss function described in Section 4.3. We evaluate the AMM loss in Table 2 where we compare the results on the NCE, SHN, MMS and AMM loss functions described in Sections 2.3.1 and 4.3 on four different datasets (the proposed Spoken Moments in Time dataset (S-MiT) as well as Vatex-en [68], MSR-VTT [70] and ActivityNet Captions[1] [32]). The proposed AMM loss function consistently achieves the best results across each dataset in Table 2 and the BART language model provides the strongest representations for the retrieval task in Table 1.

Table 2 shows a comparison of our AMM approach to other methods for cross-modal contrastive learning. We use the BART language model [36] to generate representations of words transcribed from the audio captions via a pretrained ASR model. Replacing the monotonically increasing margin used in MMS [29] with an adaptive margin that scales with the samples in a batch achieves the strongest results. We observed that as training continues and the margin in MMS continues to grow the training performance begins to degrade. This is likely due to the margin becoming too large for stable training as described in prior work [67].

In Table 3, we show a comparison of different spoken caption models with different loss functions. The proposed AMM approach beats the other loss functions consistently.

5.2. Cross Dataset Evaluation

To further examine the strength of our proposed Spoken Moments in Time (S-MiT) dataset, we compare the generalization performance of models trained on four different datasets (S-MiT as well as Vatex-en [68], MSR-VTT [70] and ActivityNet Captions [32]) for video/caption retrieval (see Table 2 (right) for comparisons of these datasets). We train each model on a single dataset using the approach described in Section 4.3 and evaluate on the test set from each other dataset. For example, a model trained on Vatex is evaluated on, in addition to its own, the test sets of ActivityNet Captions, MSR-VTT and S-MiT. We sample five sets of 1k video-caption pairs from each test set. This allows us to fairly compare results across test sets of different sizes (see supplementary material for full test set results). Each model in this evaluation was trained using the BART [36] language model and the proposed AMM loss function which was found to give the best results (see Tables 1, 2). We evaluate the models using the mean between the video-to-caption and caption-to-video retrieval tasks. We are not able to compare the spoken caption models from Table 3 here as the other datasets only include text captions.

In Table 4, we can see that the S-MiT model generalizes better than the other models in spite of the additional noise introduced by the ASR model. Additionally, the restriction to 3-second videos in S-MiT does not hinder it's ability to generalize to the much longer videos of the other datasets.

6. Conclusions

In this paper, we have introduced the Spoken Moments in Time dataset which includes 500k pairs of video clips and corresponding spoken descriptions. This new dataset represents the largest video caption dataset available and will serve as a new benchmark for the community. We compared various benchmark models for learning joint representations between captions and videos, and evaluated our approaches on multiple datasets to highlight the strength of the models as well as the ability of models trained on our proposed dataset to generalize to tasks in other datasets. With these results we are confident that the presented Spoken Moments dataset will have a positive impact on the fields of video understanding and cross-modal learning.

7. Acknowledgment

This work was supported by the MIT-IBM Watson AI Lab as well as the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/ Interior Business Center (DOI/IBC) contract number D17PC00341.

8. Disclaimer

The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/IBC, or the U.S. Government.

[1] We used the groundtruth timestamps to get corresponding video clips.

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Apostol (Paul) Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 2

[2] Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. Unsupervised learning from narrated instruction videos. In *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2016. 3

[3] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Adv. Neural Inform. Process. Syst.*, 2016. 3

[4] Angie Boggust, Kartik Audhkhasi, Dhiraj Joshi, David Harwath, Samuel Thomas, Rogerio Feris, Dan Gutfreund, Yang Zhang, Antonio Torralba, Michael Picheny, and James Glass. Grounding spoken words in unlabeled video. In *CVPR Sight and Sound Workshop*, 2019. 2

[5] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. 6, 7

[6] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015. 2

[7] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018. 3

[8] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Int. Conf. Comput. Vis.*, 2017. 2

[9] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 190–200, 2011. 3, 5

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 3, 4

[11] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 2, 3

[12] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, , Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision. *CoRR*, abs/2006.13256, 2020. 3, 5

[13] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *Eur. Conf. Comput. Vis.*, 2018. 3, 5

[14] Pradipto Das, Chenliang Xu, Richard F Doell, and Jason J Corso. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2634–2641, 2013. 3

[15] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International conference on machine learning*, pages 933–941, 2017. 5

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2009. 2, 5

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, pages 4171–4186, 2019. 4, 6, 7

[18] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Int. Conf. Comput. Vis.*, pages 1422–1430, 2015. 3

[19] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017. 4

[20] Spandana Gella, Mike Lewis, and Marcus Rohrbach. A dataset for telling the stories of social media videos. In *Empirical Methods in Natural Language Processing*, pages 968–974, Oct.-Nov. 2018. 2, 3, 5

[21] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. In *Int. Conf. Comput. Vis.*, Oct 2017. 1, 2

[22] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2018. 1

[23] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010. 4

[24] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *IEEE Conf. Comput. Vis. Pattern Recog.*, volume 2, pages 1735–1742, 2006. 3

[25] David Harwath, Adria Recasens, Didac Suris, Galen Chuang, Antonio Torralba, and James Glass. Jointly discovering visual objects and spoken words from raw sensory input. *Int. J. Comput. Vis.*, (128):620–641, 2020. 2, 3, 4, 5, 6, 7

[26] David Harwath, Antonio Torralba, and James Glass. Unsupervised learning of spoken language with visual context. In *Adv. Neural Inform. Process. Syst.*, 2016. 2, 3

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016. 6

[28] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013. 2, 3

[29] Gabriel Ilharco, Yuan Zhang, and Jason Baldridge. Large-scale representation learning from visually grounded untranscribed speech. In *Conference on Computational Natural Language Learning*, pages 55–65, Nov. 2019. 2, 3, 4, 8

[30] Aren Jansen, Manoj Plakal, Ratheet Pandya, Daniel PW Ellis, Shawn Hershey, Jiayang Liu, R Channing Moore, and Rif A Saurous. Unsupervised learning of semantic audio representations. In *IEEE international conference on acoustics, speech and signal processing*, pages 126–130, 2018. 4

[31] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1, 2, 3, 5

[32] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Int. Conf. Comput. Vis.*, 2017. 2, 3, 5, 7, 8

[33] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73, 2017. 2, 3

[34] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inform. Process. Syst.*, 25:1097–1105, 2012. 1, 6

[35] Hilde Kuehne, Hueihan Jhuang, Rainer Stiefelhagen, and Thomas Serre. Hmdb51: A large video database for human motion recognition. In *High Performance Computing in Science and Engineering '12*, pages 571–582, 2013. 2

[36] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019. 6, 7, 8

[37] Aoxue Li, Weiran Huang, Xu Lan, Jiashi Feng, Zhenguo Li, and Liwei Wang. Boosting few-shot learning with adaptive margin loss. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 6

[38] Sijin Li, Weichen Zhang, and Antoni B. Chan. Maximum-margin structured learning with deep networks for 3d human pose estimation. In *Int. Conf. Comput. Vis.*, December 2015. 6

[39] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Int. Conf. Comput. Vis.*, October 2019. 2, 6

[40] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Eur. Conf. Comput. Vis.*, pages 740–755. 2014. 5

[41] Hao Liu, Xiangyu Zhu, Zhen Lei, and Stan Z. Li. Adaptiveface: Adaptive margin and sampling for face recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2019. 6

[42] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Adv. Neural Inform. Process. Syst.*, pages 13–23, 2019. 3

[43] Danny Merkx, Stefan L. Frank, and Mirjam Ernestus. Language learning using speech to image retrieval. In *International Speech Communication Association*, September 2019. 2

[44] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Int. Conf. Comput. Vis.*, pages 2630–2640, 2019. 3, 5, 6

[45] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, and Aude Oliva. Moments in time dataset: one million videos for event understanding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(2):502–508, 2019. 1, 2, 3, 4

[46] Mathew Monfort, Kandan Ramakrishnan, Alex Andonian, Barry A McNamara, Alex Lascelles, Bowen Pan, Quanfu Fan, Dan Gutfreund, Rogerio Feris, and Aude Oliva. Multi-moments in time: Learning and interpreting models for multi-action video understanding. *arXiv preprint arXiv:1911.00232*, 2019. 1, 2, 3, 5, 6

[47] Andrew Owens, Jiajun Wu, Josh McDermott, William Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *Eur. Conf. Comput. Vis.*, 2016. 3

[48] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Int. Conf. Comput. Vis.*, December 2015. 2, 3

[49] Esteban Real, Jonathon Shlens, Stefano Mazzocchi, Xin Pan, and Vincent Vanhoucke. Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5296–5305, 2017. 1

[50] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 2013. 3, 5

[51] Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. Coherent multi-sentence video description with variable level of detail. In *German conference on pattern recognition*, pages 184–195, 2014. 3

[52] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3202–3212, 2015. 3, 5

[53] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. *Int. J. Comput. Vis.*, 123(1):94–120, 2017. 3

[54] Marcus Rohrbach, Anna Rohrbach, Michaela Regneri, Sikandar Amin, Mykhaylo Andriluka, Manfred Pinkal, and Bernt Schiele. Recognizing fine-grained and composite activities using hand-centric features and script data. *Int. J. Comput. Vis.*, 119(3):346–373, Sept. 2016. 3

[55] Andrew Rouditchenko, Angie Boggust, David Harwath, Dhiraj Joshi, Samuel Thomas, Kartik Audhkhasi, Rogerio Feris, Brian Kingsbury, Michael Picheny, Antonio Torralba, and James Glass. Avlnet: Learning audio-visual language representations from instructional videos. In *arXiv:2006.09199*, 2020. 5

[56] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. In *arXiv:1904.05862*, 2019. 6, 7

[57] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2015. 3, 4

[58] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Annual Meeting of the Association for Computational Linguistics*, pages 2556–2565, 2018. 2, 3

[59] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Eur. Conf. Comput. Vis.*, pages 510–526, 2016. 2, 5

[60] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Adv. Neural Inform. Process. Syst.*, pages 568–576, 2014. 2

[61] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2

[62] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Int. Conf. Comput. Vis.*, October 2019. 4

[63] Didac Suris, Adria Recasens, David Bau, David Harwath, James Glass, and Antonio Torralba. Learning words by drawing images. In *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2019. 2

[64] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015. 2

[65] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3, 4

[66] Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. Object referring in visual scene with spoken language. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1861–1870, 2018. 2

[67] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2018. 4, 8

[68] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Int. Conf. Comput. Vis.*, October 2019. 3, 5, 7, 8

[69] Yufei Wang, Zhe Lin, Xiaohui Shen, Scott Cohen, and Garrison W. Cottrell. Skeleton key: Image captioning by skeleton-attribute decomposition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, July 2017. 2, 3

[70] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. IEEE Conf. Comput. Vis. Pattern Recog., June 2016. 2, 3, 5, 7, 8

[71] Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *Int. J. Comput. Vis.*, 126(2-4):375–389, 2018. 1

[72] Junjie Zhang, Qi Wu, Chunhua Shen, Jian Zhang, and Jianfeng Lu. Multi-label image classification with regional latent semantic dependencies. *IEEE Trans. Multimedia*, 2018. 1

[73] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Eur. Conf. Comput. Vis.*, 2016. 3

[74] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 645–654, 2017. 3

[75] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Eur. Conf. Comput. Vis.*, September 2018. 3

[76] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017. 2

[77] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Adv. Neural Inform. Process. Syst.*, pages 487–495. 2014. 2, 5

[78] Luowei Zhou, Nathan Louis, and Jason J Corso. Weakly-supervised video object grounding from text by loss weighting and object interaction. In *Brit. Mach. Vis. Conf.*, 2018. 3

[79] Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018. 2, 3, 5