Grounding Spoken Words in Unlabeled Video

Angie Boggust[1], Kartik Audhkhasi[2], Dhiraj Joshi[2], David Harwath[1], Samuel Thomas[2]
Rogerio Feris[2], Dan Gutfreund[2], Yang Zhang[2], Antonio Torralba[1], Michael Picheny[2], James Glass[1]

[1] Massachusetts Institute of Technology [2] IBM Research AI

Abstract

In this paper, we explore deep learning models that learn joint multi-modal embeddings in videos where the audio and visual streams are loosely synchronized. Specifically, we consider cooking show videos from the YouCook2 dataset and a subset of the YouTube-8M dataset. We introduce varying levels of supervision into the learning process to guide the sampling of audio-visual pairs for training the models. This includes (1) a fully-unsupervised approach that samples audio-visual segments uniformly from an entire video, and (2) sampling audio-visual segments using weak supervision from off-the-shelf automatic speech and visual recognition systems. Although these models are preliminary, even with no supervision they are capable of learning cross-modal correlations, and with weak supervision we see significant amounts of cross-modal learning.

1. Introduction

Babies learn to perceive the world around them by understanding speech and recognizing objects with extremely weak supervision, aided only by observation, repetition, and multi-modal context. An obvious existence proof that this occurs is the feedback a young child gets when initially learning language. For example, take the case when a child sees a round object and hears the audio segment corresponding to the word "ball". After seeing these two events co-occur some number of times, the child learns that a round object is called a "ball" and that the term "ball" refers to a round object. In this paper, we explore this type of loosely synchronous, multi-stream, audio-visual learning in a machine learning context. Specifically, we develop deep learning models that demonstrate cross-modal learning capabilities trained using unannotated videos. By leveraging naturally occurring cross-modal correspondences, we demonstrate that these models can learn relationships between what they see and what they hear in both general audio and speech.

There is a growing body of research which studies the problem of joint audio-visual modeling. Correlating visual objects with the sounds they produce has been used as a signal for *self-supervised* learning of both auditory and visual features [2, 4, 16, 17]. This idea has been further developed for visually-guided audio source separation [8, 18, 19], generating sounds from silent videos [17, 21], localizing sounds in video frames [3, 19], learning association of faces and voices [13, 15], and video-based audio spatialization [7, 14]. Most of the recent work targets synchronous audio and visual signals (such as sight and sound of waterfalls, musical instruments, etc.). In contrast, our work focuses on the speech audio and visual pairing which is loosely synchronous. More closely related to our work, there have been limited attempts to combine visual and speech information in the *zero resource* (unsupervised) setting [6, 9, 10, 12]. Our work builds upon recent research results that demonstrate an ability to uncover concepts from images paired with spoken descriptions [9, 10] by learning a joint audio-visual latent space which reflects the underlying semantics of both modalities. We extend the aforementioned work by learning from videos containing descriptive speech. In this initial work, we focus on cooking shows as they exhibit tight relationships between the audio and video components. The key challenge taken up in our work is the asynchronous nature of spoken audio and visual descriptions in such videos. We study the effect of introducing varying levels of supervision into the learning process and create soft performance upper bounds that a completely unsupervised learning system can achieve in such scenarios. We plan to release the machine and human labeled audio-visual pairs which formed the basis of our evaluation in this paper in an attempt to create a collaborative research ecosystem around joint speech-visual pair modeling.

2. DAVEnet Modeling

We make use of the Deep Audio-Visual Embedding network (DAVEnet) architecture [10] to model the grounding between image frames and audio segments. The DAVEnet model is comprised of two convolutional branches: an image branch f which takes as input an RGB image and an

audio branch g which accepts a log-Mel frequency spectrogram. Both f and g output D-dimensional feature maps which attempt to capture the semantics of both modalities. We follow [10] and use the triplet loss function as a training objective [5] with a margin hyperparameter of 1 and the dot product similarity between average-pooled feature maps. We blend two variants of the triplet loss with equal weight: in the first case, impostor samples are randomly drawn from other examples within a minibatch, as in [10]. The second loss term employs semi-hard negative mining [11].

3. Video Processing

The DAVEnet model takes still images paired with variable length audio segments as input. Applying the DAVEnet model to continuous video requires us to process each video in the dataset into discrete pairs of video frames and corresponding audio segments. Discretizing a video into frame-audio pairs presents the challenge of identifying all conceptually-relevant pairs in the video while minimizing the number of redundant pairs. For example, sampling too frequently may lead to adjacent pairs that contain redundant information, while sampling too few pairs may miss important concepts from the video. In our approach, we uniformly extract video frames at rate of 1 frame per second (fps). Each extracted frame is paired with the T seconds of audio centered around it. T is a tunable parameter that we explored and found $T = 2$ seconds to perform the best.

3.1. Off-the-shelf Systems for Labeling

We weakly label our dataset by passing the extracted video frames through the IBM food detector (built as part of IBM Watson Visual Recognition service[1]). The detector uses a multi-branch network, consisting of a backbone ResNet50 CNN to classify an image into a set of 2200 food classes, and a parallel sub-network branch to perform food versus non-food classification on the image. The food concept detector produces posteriors for food/non-food and all 2200 food classes for each input frame.

We transcribe the audio channel of each video using the US English broadband model of IBM's Watson Speech-To-Text (STT) service.[2] The acoustic and language models used in the STT system were both trained on a large set of data from various corpora. The STT system predicts a sequence of word hypotheses and word boundaries given the input audio. We did not adapt the food concept detector or the STT system on the YouCook2 or YouTube-8M datasets.

We use these automatic labels to construct the conceptually-relevant audio-video frame pairs for our weakly-supervised experiments. The details of this data preparation are in Section 6.

[1] https://www.ibm.com/watson/services/visual-recognition/
[2] https://www.ibm.com/watson/services/speech-to-text/

4. Datasets

In real world descriptive video datasets, the audio and visual channels are often disjoint (e.g. a chef speaking about a childhood memory while chopping vegetables). In order to increase the number of semantically relevant frame-audio pairs used to train our model, we combine the YouCook2 dataset [20] and the YouTube-8M dataset [1], which are summarized in Table 1.

YouCook2: The YouCook2 dataset consists of approximately 2000 cooking videos from YouTube filmed from the third person point of view. The videos were randomly separated into a 67-23-10 training, validation, and testing split and categorized into one of 89 recipe types. Each recipe type contains 22 videos on average.

YouTube-8M: The YouTube-8M dataset is a large scale video dataset consisting of 6.1 million YouTube videos. Videos in the training and validation sets are assigned to one or more of over 3800 categories. To maintain consistency with the YouCook2 dataset, we only use videos from the *baking*, *cooking*, *cooking show*, *cuisine*, *dish*, *food*, and *recipe* categories. The resulting corpus contains videos with speech from many different languages, as well as videos that do not contain any speech. To address these inconsistencies, we select approximately 3500 English videos from the YouTube-8M dataset using their YouTube metadata audio language tag.

		Video Length (sec)		
Subset	**N Videos**	**Min**	**Max**	**Mean**
YC2 Train	1262	46.0	1106.1	317.7
YC2 Val	439	44.3	829.4	308.9
YC2 Test	205	37.8	722.8	318.1
YT8 English	3535	119.0	499.8	275.5

Table 1: A summary of the YouCook2 dataset and the English cooking subset of the YouTube-8M dataset.

Food Word Analysis: To better understand the semantic content of the videos in our dataset, we analyze the number of food-related words spoken per video. We manually select 350 food nouns from the entire set of words generated by the automatic speech recognizer on the YouCook2 and metadata-tagged YouTube-8M videos (see Section 3.1). In our dataset, each video contains an average of 10.5 unique food words and 22 total food words. Together, these results suggest there are a significant number of food concepts spoken and repeated throughout the videos that can be used as a basis for learning.

Synchronization Analysis: We analyze the temporal synchronization between the audio and visual streams of our video data to better understand the nature of descriptive data and to motivate our video parsing scheme and weakly-supervised experiments. As described in Section 3, we parse the videos in our dataset into frames and pair them with a symmetric window of T seconds of audio. This procedure assumes that the chosen audio segment describes the center video frame.

In order to test the validity of this assumption, we utilize the automatically-derived food labels (see Section 3.1). We select all video frames from the YouCook2 validation set with a food probability greater than 0.5 as determined by the food concept detector. For each selected video frame, we select its five most likely food labels from the food concept detector. We then search for any word in these food labels within the STT transcript of the audio in a T second window centered around the frame. We find that in nearly 16% of the cases, at least one food label word occurred in the transcript within a 20 second window of audio around the video frame. This is a non-trivial number of frames considering that the automatically-derived audio/video labels may have errors, and we are searching for an exact match of the visual food label words in the audio transcription.

5. Unsupervised Semantic Learning

In this section, we evaluate the performance of an unsupervised approach to learning audio-visual objects from unannotated video. We first parse the YouCook2 and metadata-tagged YouTube-8M datasets by uniformly extracting frame-audio pairs, as described in Section 3. This results in 1,169,255 training pairs which are separated into minibatches of 100 pairs each for training DAVEnet (see Section 2) with stochastic gradient descent.

In the evaluation task, the model is given a set of frame-audio pairs, but not told the specific pairings. Given a frame, the model must rank the similarity of each audio segment to that frame. We evaluate this task using Recall@10, which we define as the fraction of examples for which the model returns the true audio pair in the top ten most similar audio segments. We also perform this task in the reverse, holding the audio segment constant and searching on the frames.

To perform the evaluation, we create two validation datasets from the YouCook2 validation set: the Val 1000 dataset and the Food Val 1000 dataset. The Val 1000 dataset consists of 1000 randomly chosen pairs from the YouCook2 validation set. The Food Val 1000 dataset contains 1000 frame-audio pairs from the YouCook2 validation set whose audio segment contains a food word as identified by the automatic speech recognition model (see Section 3.1). The best performance of our model on these two datasets is shown in Table 2.

We visualize the model's output to further evaluate its results and better interpret the learned features. For a subset of evaluation videos, we extract the frame-audio pairs at a rate of 24 fps. We feed each pair into the trained DAVEnet model and extract the similarity matchmap for the frame-audio relationship. Using the matchmap, we highlight the pixels in the frame that have a high similarity to the audio segment. These highlighted frames are then recombined with the original audio track, resulting in a video showing the semantically relevant parts of the image in real time.

Frame level examples are shown in Figure 1. Objectively, our model performs best on videos where the camera is focused on the food items. On videos focused primarily on the chef, our model tends to learn the relationship between the person's voice and the scene, due to the smaller number of semantically relevant pixels in each frame.

Validation Set	Audio R@10	Visual R@10
Val 1000	18.2%	18.3%
Food Val 1000	20.3%	19.4%

Table 2: Performance of the unsupervised model on held out validation data.

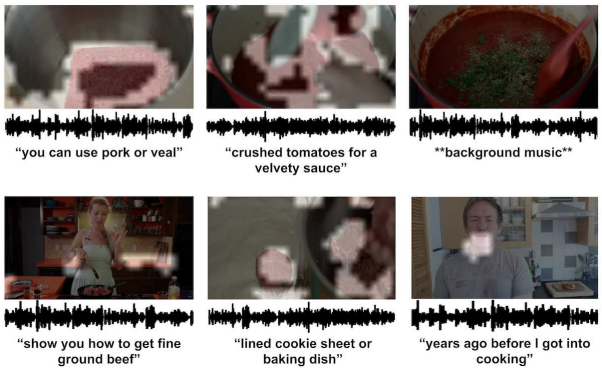

Figure 1: Examples of the unsupervised model's semantic correlation on videos from the YouCook2 dataset. The top row shows positive examples where the highlighted portions of the frame are related to the audio shown below. The bottom row shows examples in which our model fails to pick up on semantically relevant pixels or identifies pixels unrelated to the audio.

6. Weakly-Supervised Semantic Learning

To evaluate the validity of a weakly supervised learning approach, we utilize the food concept detector and STT systems described in Section 3.1 to weakly-label the YouCook2 dataset. We identify the set of frames from all video frames in YouCook2 that have been assigned to a food class by the food concept detector with a probability of at least 0.5. For each identified frame, we select its top-5 food classes and search for the exact food class name within the STT tran-

scription of the 20 second audio segment surrounding the frame. There are a total of 2860/1067/381 such food frame-audio pairs in the YouCook2 training/validation/testing sets, respectively. We further augment the training examples with an equal number of non-food frames as predicted by the food classifier with a confidence at least 0.95 and pair each one with the surrounding 1 second segment of audio.

Before conducting the weakly-supervised experiments, we performed manual checking of the 1067 YouCook2 validation set food frame-audio pairs. A total of 845 examples contained correct labels and were picked to evaluate the models. We note that this validation set is different from the one used in Table 2.

We first evaluate the recall performance of DAVEnet trained in an unsupervised fashion. This unsupervised DAVEnet obtains an audio Recall@10 of 21.9% and video Recall@10 of 22.1% on the YouCook2 validation set. In Table 3, we analyze the effect of using weak-supervision to fine-tune the unsupervised DAVEnet model as well as an untrained model and a model whose vision channel was pretrained on ImageNet. We find that weak supervision improves the recall of the DAVEnet model trained in fully unsupervised fashion. Weak supervision is also able to achieve audio/visual recalls of 39.1%/38.1% with only ImageNet-supervised pretraining of the visual sub-network.

		Recall@10	
Audio Init	**Visual Init**	**Audio**	**Visual**
Random	Random	13.4%	14.6%
Random	ImageNet	39.1%	38.1%
Unsupervised	Unsupervised	27.2%	23.3%

Table 3: Audio and visual Recall@10 for the labeled YouCook2 validation set with initialization settings. "Random" refers to random initialization, "ImageNet" refers to initialization of visual sub-network by a ImageNet-trained VGG network, and "Unsupervised" refers to unsupervised training on YouCook2 + YouTube-8M data sets as shown in Section 5.

7. Conclusion

We present a novel unsupervised approach to cross-modal learning of audio-visual concepts from unannotated instructional video. We eliminate the need for expensive and time consuming data collection and annotation by extending our model to learn from publicly available videos. This work establishes a benchmark and sets a basis for future work on unsupervised multi-modal learning of video and speech. We have begun to explore such future work including: unsupervised methods to guide video frame extraction (e.g. zero resource pattern discovery), tools for modelling audio and visual alignment (e.g. multiple instance learning), and application of these methods to a larger corpora of descriptive video (e.g. utilizing language detection techniques to mine English videos from YouTube-8M).

References

[1] S. Abu-El-Haija et al. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 2

[2] R. Arandjelovic et al. Look, listen, and learn. In *ICCV*, 2017. 1

[3] R. Arandjelović et al. Objects that sound. In *European Conference on Computer Vision*, 2018. 1

[4] Y. Aytar et al. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems 29*, pages 892–900. 2016. 1

[5] G. Chechik et al. Large scale online learning of image similarity through ranking. *J. Mach. Learn. Res.*, 11:1109–1135, Mar. 2010. 2

[6] G. Chrupala et al. Representations of language in a model of visually grounded speech signal. In *ACL*, 2017. 1

[7] R. Gao et al. 2.5 d visual sound. *arXiv preprint arXiv:1812.04204*, 2018. 1

[8] R. Gao et al. Learning to separate object sounds by watching unlabeled video. In *ECCV*, 2018. 1

[9] D. Harwath et al. Unsupervised learning of spoken language with visual context. In *NIPS*, 2016. 1

[10] D. Harwath et al. Jointly discovering visual objects and spoken words from raw sensory input. *CoRR*, abs/1804.01452, 2018. 1, 2

[11] A. Jansen et al. Unsupervised learning of semantic audio representations. In *ICASSP*, 2018. 2

[12] H. Kamper et al. Visually grounded learning of keyword prediction from untranscribed speech. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2017. 1

[13] C. Kim et al. On learning associations of faces and voices. *arXiv preprint arXiv:1805.05553*, 2018. 1

[14] P. Morgado et al. Self-supervised generation of spatial audio for 360 video. In *NeurIPS*, 2018. 1

[15] A. Nagrani et al. Seeing voices and hearing faces: Cross-modal biometric matching. In *CVPR*, 2018. 1

[16] A. Owens et al. *Ambient Sound Provides Supervision for Visual Learning*, pages 801–816. 2016. 1

[17] A. Owens et al. Visually indicated sounds. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2405–2413, 2016. 1

[18] A. Owens et al. Audio-visual scene analysis with self-supervised multisensory features. *CoRR*, abs/1804.03641, 2018. 1

[19] H. Zhao et al. The sound of pixels. In *ECCV*, September 2018. 1

[20] L. Zhou et al. Towards automatic learning of procedures from web instructional videos. In *AAAI Conference on Artificial Intelligence*, pages 7590–7598, 2018. 2

[21] Y. Zhou et al. Visual to sound: Generating natural sound for videos in the wild. In *CVPR*, 2018. 1