Cascaded Multilingual Audio-Visual Learning from Videos - Extended Abstract

Andrew Rouditchenko[1], Angie Boggust[1], David Harwath[2], Samuel Thomas[3], Hilde Kuehne[3], Brian Chen[4], Rameswar Panda[3], Rogerio Feris[3], Brian Kingsbury[3], Michael Picheny[5], James Glass[1]
[1]MIT CSAIL, [2]UT Austin, [3]IBM Research AI, [4]Columbia University, [5]NYU

roudi@mit.edu

Abstract

In this extended abstract, we describe our recent work on self-supervised audio-visual models that learn from instructional videos. Prior work has shown that these models can relate spoken words and sounds to visual content after training on a large-scale dataset of videos, but they were only trained and evaluated on videos in English. To learn multilingual audio-visual representations, we propose a cascaded approach that leverages a model trained on English videos and applies it to audio-visual data in other languages, such as Japanese videos. With our cascaded approach, we show an improvement in retrieval performance of nearly 10x compared to training on the Japanese videos solely. We also apply the model trained on English videos to Japanese and Hindi spoken captions of images, achieving state-of-the-art performance. We encourage readers to check our full paper, which has been accepted to Interspeech 2021 and will be available publicly soon, for the full details and more experiments.

1. Introduction

Recently, researchers have proposed models that can learn to recognize words from raw audio by associating them to semantically related images [1, 4–6, 8, 9, 15]. The first models were applied to English spoken audio captions, but further work applied the models to Hindi [3] and Japanese [7, 12] captions. We are also interested in learning multilingual representations from audio-visual data, but we aim to do this from instructional videos that are naturally present on the internet and do not require recorded spoken captions.

To learn multilingual representations, we use the recently proposed Audio-Video Language Network [13]. Compared to prior image and spoken audio caption models, it learns from entire video clips and raw audio from instructional videos. The model was trained on HowTo100M [10], a dataset of 1.2M instructional videos, and achieved strong video retrieval performance on the YouCook2 [16] dataset of English cooking videos. Here, we propose a cascaded approach that applies the AVLnet model trained on English videos to videos in Japanese. While spoken audio captions of images already exist for Japanese [12] and Hindi [4], there are no instructional video datasets similar in size to YouCook2 in other languages. Therefore, we introduce the YouCook-Japanese instructional video dataset. Applying our cascaded approach, we show an improvement in retrieval performance of nearly 10x on YouCook-Japanese compared to training on the Japanese videos solely.

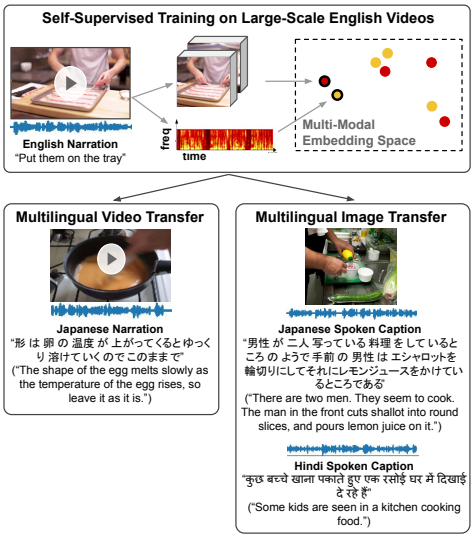

Figure 1: Given an audio-video model (AVLnet) trained on videos in English, we transfer the representations to videos in Japanese. We also transfer the representations to images and spoken captions in Japanese and Hindi.

We also show that our cascaded approach can work as a bridge between English instructional videos and the spoken audio captions of images in Japanese and Hindi. Given the AVLnet model trained on English videos, we fine-tune it on Japanese and Hindi spoken captions of images, achieving state-of-the-art performance. We will release our code, trained models, and data at avlnet.csail.mit.edu.

2. Technical Approach

2.1. Videos

AVLnet [13] is trained through a contrastive loss to discriminate between temporally aligned audio-video pairs and temporally mismatched pairs from both within the same video and from other videos. This results in an audio-video embedding space which colocates semantically similar audio and visual inputs. Since AVLnet does not require any

annotations besides the raw video data, we only assume that a set of videos in the target language is given, but without any additional annotation. One approach is to simply train AVLnet only on the target videos in the new language. However, we find that a large number of videos, typically hundreds of thousands, is necessary to learn strong representations from scratch, and there is simply not enough videos in downstream datasets such as YouCook2 to train the model from scratch. Therefore, our proposed approach is simple: given the AVLnet model trained on English HowTo100M videos, we apply it to videos in Japanese by directly fine-tuning it on the Japanese videos. This represents a cascade since the model only learns from videos in one language at a time (ie. first English, then Japanese).

YouCook-Japanese. There are currently no other instructional video datasets in other languages similar in size to YouCook2. Therefore, we collected a dataset of Japanese cooking videos, and call it YouCook-Japanese to indicate the similarity in content and size to YouCook2. As a starting point, Sigurdsson et al. [14] proposed a version of HowTo100M in Japanese with approximately 300k videos. We followed the steps to download Japanese instructional videos from YouTube, except we limited the search to cooking videos only. We used a CNN-based audio segmentation toolkit [2] to segment the videos into clips containing speech, and then filtered the clips to be at least 5s and at most 50s. To make the dataset similar in size to YouCook2, we selected 10k random clips for training, 3k clips for validation, and 3k clips for evaluation, with the constraint that each video can only appear in one set.

2.2. Images and Spoken Captions

Since instructional videos and spoken captions of images both contain descriptive audio of visual scenes, our cascaded approach is also applicable to images and spoken captions. Specifically, we use the AVLnet model trained on HowTo100M videos and fine-tune it on the spoken captions and images in the Places Audio Caption Dataset in Japanese and Hindi. For these experiments, we train AVLnet using only the 2D features in the visual branch so that the model can work on both videos and images.

3. Experiments

3.1. Video Retrieval

YouCook-Japanese. Table 1 shows the video retrieval results on YouCook-Japanese videos. AVLnet's performance when trained only on YouCook-Japanese (row a) is similar to AVLnet's performance on YouCook2 when trained only on YouCook2 videos [13], indicating that the two datasets are similar in difficulty. Using our cascaded approach, we apply the AVLnet model trained on HowTo100M to the Japanese videos which significantly improves performance.

Table 1: Video retrieval on the YouCook-Japanese (YC-JP) dataset.

AVLnet Train Data	Video Clip (A→V)			Language (V→A)		
	R@1	R@5	R@10	R@1	R@5	R@10
Random	0.03	0.17	0.33	0.03	0.17	0.33
(a) YC-JP	0.7	2.4	3.8	0.5	1.8	3.0
(b) HT100M	4.6	12.1	18.2	5.6	14.6	21.3
(c) HT100M + YC-EN	5.1	13.2	18.9	5.6	14.5	20.7
(d) HT100M + YC-JP	**7.0**	**20.4**	**29.3**	**7.6**	**20.9**	**29.7**

Table 2: Image retrieval on the Places Audio dataset.

(a) Places Audio Captions - Japanese

Method	Audio to Image			Image to Audio		
	R@1	R@5	R@10	R@1	R@5	R@10
Random	0.1	0.5	1.0	0.1	0.5	1.0
Havard et al. [7]	18.2	48.5	62.2	15.3	41.4	57.6
Ohishi et al. [11]	20.1	49.7	63.9	16.7	44.3	57.8
Ohishi et al. [12]	20.3	52.0	66.7	20.0	46.8	62.3
AVLnet	**23.5**	**57.3**	**70.4**	**24.3**	**56.6**	**70.0**

(b) Places Audio Captions - Hindi

Method	Audio to Image			Image to Audio		
	R@1	R@5	R@10	R@1	R@5	R@10
Random	0.1	0.5	1.0	0.1	0.5	1.0
Harwath et al. [3]	8.0	25.0	35.6	7.4	23.5	35.4
Havard et al. [7]	9.6	28.2	40.7	8.0	27.6	37.1
Ohishi et al. [11]	9.4	29.8	41.8	9.3	29.5	38.2
Ohishi et al. [12]	11.2	31.5	44.5	10.8	31.3	41.9
AVLnet	**15.2**	**38.9**	**51.1**	**17.0**	**39.8**	**51.5**

In the zero-shot setting (row b), i.e., without fine-tuning, the retrieval performance is nearly 5x the performance compared with training on YouCook-Japanese only. This is surprising considering that the model has only been trained on English videos. Fine-tuning the model on the Japanese videos (row d) further increases the performance to nearly 10x the performance compared with training on YouCook-Japanese only. We also note that fine-tuning the model on English YouCook2 videos (row c) instead of Japanese videos is comparable to the zero-shot performance, further indicating that the model is actually sensitive to the language present in the videos.

3.2. Image Retrieval

Table 2 shows the retrieval results on the Places Audio Caption dataset in Hindi and Japanese. For our cascaded approach, we fine-tune AVLnet trained on HowTo100M videos to each language in Places independently. We compare our approach to the state-of-the-art models for each dataset. While previous models are not trained on HowTo100M videos, some of them [3, 12] are trained on images with parallel spoken captions in multiple languages.

Our cascaded approach involves training on one language at a time, achieving large gains over prior baselines.

3.3. Conclusion

We propose a cascaded approach to learn multilingual audio-visual representations. Given the AVLnet model trained on English HowTo100M videos, we fine-tuned and evaluated it on YouCook-Japanese videos and the images and spoken captions in the Places Audio Caption dataset in Japanese and Hindi. The representations learned from HowTo100M serve as a strong initialization for fine-tuning on Japanese videos through our cascaded approach, which improves performance by nearly 10x compared to training on the Japanese videos solely.

Acknowledgements. This research was supported by the MIT-IBM Watson AI Lab. We also thank the IBM Japan team for help with Japanese ASR.

References

[1] G. Chrupała, L. Gelderloos, and A. Alishahi. Representations of language in a model of visually grounded speech signal. In *ACL*, 2017. 1

[2] D. Doukhan, J. Carrive, F. Vallet, A. Larcher, and S. Meignier. An open-source speaker gender detection framework for monitoring gender equality. In *ICASSP*, 2018. 2

[3] D. Harwath, G. Chuang, and J. Glass. Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech. In *ICASSP*, 2018. 1, 2

[4] D. Harwath, A. Recasens, D. Surís, G. Chuang, A. Torralba, and J. Glass. Jointly discovering visual objects and spoken words from raw sensory input. In *ECCV*, 2018. 1

[5] D. Harwath, A. Recasens, D. Surís, G. Chuang, A. Torralba, and J. Glass. Jointly discovering visual objects and spoken words from raw sensory input. *IJCV*, 2020. 1

[6] D. Harwath, A. Torralba, and J. Glass. Unsupervised learning of spoken language with visual context. In *NeurIPS*, 2016. 1

[7] W. N. Havard, J.-P. Chevrot, and L. Besacier. Models of visually grounded speech signal pay attention to nouns: A bilingual experiment on english and japanese. In *ICASSP*, 2019. 1, 2

[8] G. Ilharco, Y. Zhang, and J. Baldridge. Large-scale representation learning from visually grounded untranscribed speech. In *CoNLL*, 2019. 1

[9] D. Merkx, S. L. Frank, and M. Ernestus. Language learning using speech to image retrieval. In *INTERSPEECH*, 2019. 1

[10] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, 2019. 1

[11] Y. Ohishi, A. Kimura, T. Kawanishi, K. Kashino, D. Harwath, and J. Glass. Pair expansion for learning multilingual semantic embeddings using disjoint visually-grounded speech audio datasets. *INTERSPEECH*, 2020. 2

[12] Y. Ohishi, A. Kimura, T. Kawanishi, K. Kashino, D. Harwath, and J. Glass. Trilingual semantic embeddings of visually grounded speech with self-attention mechanisms. In *ICASSP*, 2020. 1, 2

[13] A. Rouditchenko, A. Boggust, D. Harwath, D. Joshi, S. Thomas, K. Audhkhasi, R. Feris, B. Kingsbury, M. Picheny, A. Torralba, et al. Avlnet: Learning audio-visual language representations from instructional videos. *arXiv preprint arXiv:2006.09199*, 2020. 1, 2

[14] G. A. Sigurdsson, J.-B. Alayrac, A. Nematzadeh, L. Smaira, M. Malinowski, J. Carreira, P. Blunsom, and A. Zisserman. Visual grounding in video for unsupervised word translation. In *CVPR*, 2020. 2

[15] G. Synnaeve, M. Versteegh, and E. Dupoux. Learning words from images and speech. *NeurIPS Workshop on Learning Semantics*, 2014. 1

[16] L. Zhou, C. Xu, and J. J. Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018. 1