INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Cascaded Multilingual Audio-Visual Learning from Videos

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, USA
[2]UT Austin, USA
[3]IBM Research AI, USA
[4]Columbia University, USA
[5]NYU, USA

roudi@mit.edu

Abstract

In this paper, we explore 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.

Index Terms: multilingual, audio-visual, videos, self-supervised, low-resource

1. Introduction

While technologies like Automatic Speech Recognition (ASR) and Machine Translation (MT) enable us to interact better with computers and each other, they are currently only available for less than 2% of the world's languages, in part due to the large amount of manually labelled data required for each language [1]. Recently, researchers have proposed models that can instead learn to recognize words from raw audio by associating them to semantically related images [2–8]. The first models were applied to English spoken audio captions, but further work applied the models to Hindi [9] and Japanese [10, 11] 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 leverage the recently proposed Audio-Video Language Network [12]. 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 [13], a large-scale dataset of 1.2M instructional videos, and achieved strong video retrieval performance on the YouCook2 [14] dataset of English cooking videos.

It would be challenging to collect large-scale instructional video datasets in other languages to train AVLnet given the significant engineering effort required to download and process them. Furthermore, there are currently fewer instructional videos available for other languages, especially low-resource languages. We address these limitations by proposing 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 [11] and Hindi [4], there are no instructional video datasets similar in size to YouCook2 in other languages. Therefore, we collected a dataset of instructional cooking videos in Japanese, named YouCook-Japanese. 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.

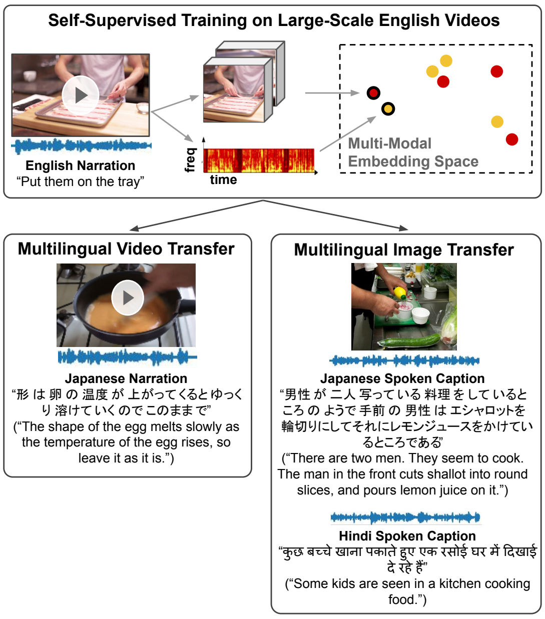
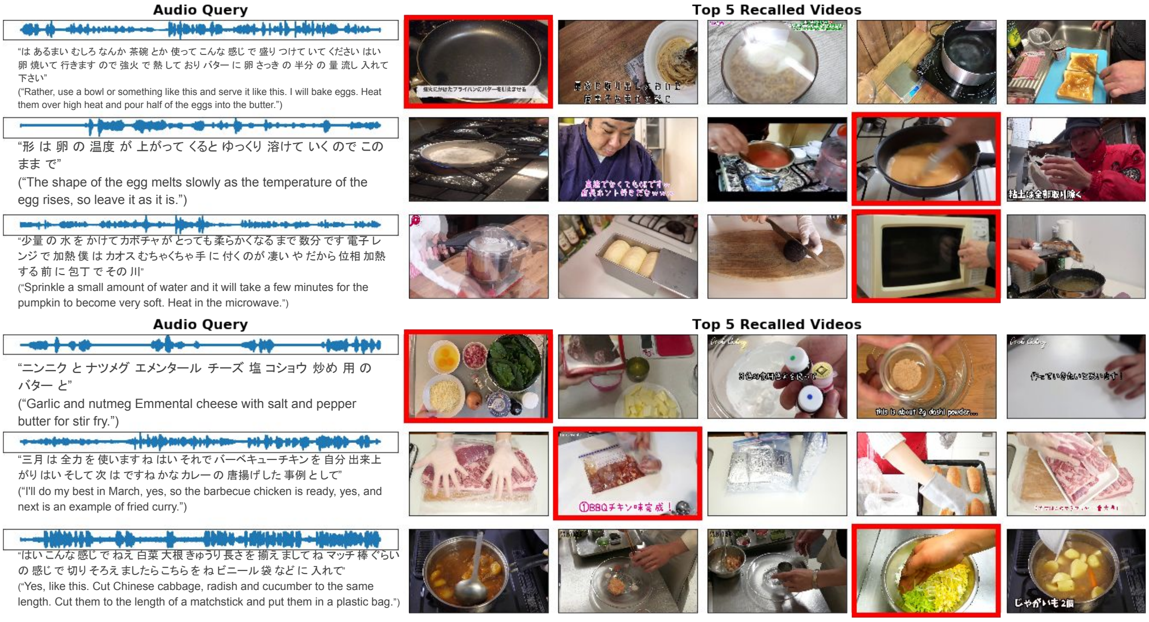

Self-Supervised Training on Large-Scale English Videos
English Narration
"Put them on the tray"
freq
time
Multi-Modal
Embedding Space
Multilingual Video Transfer
Japanese Narration
"形 は 卵 の 温度 が 上がってくるとゆっくり 溶けていくので このままで"
("The shape of the egg melts slowly as the temperature of the egg rises, so leave it as it is.")
Multilingual Image Transfer
Japanese Spoken Caption
"男性 が 二人 写っている 料理 をしている ところ の ようで 手前 の 男性 は エシャロットを 輪切りにしてそれにレモンジュースをかけているところである"
("There are two men. They seem to cook. The man in the front cuts shallot into round slices, and pours lemon juice on it.")
Hindi Spoken Caption
"कुछ बच्चे खाना पकाते हुए एक रसोई घर में दिखाई दे रहे हैं"
("Some kids are seen in a kitchen cooking food.")

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. Finally, we provide an analysis of the impact of the amount of English training videos on retrieval performance. We will release our code, trained models, and data at http://avlnet.csail.mit.edu

http://dx.doi.org/10.21437/Interspeech.2021-1352



“は あるまい むしろ なんか 茶碗 とか 使って こんな 感じ で 盛り つけて いて ください はい 卵 焼いて 行きます ので 強火 で 熱 して おり バター に 卵 さっき の 半分 の 量 流し 入れて 下さい”
(“Rather, use a bowl or something like this and serve it like this. I will bake eggs. Heat them over high heat and pour half of the eggs into the butter.”)

“形 は 卵 の 温度 が 上がって くると ゆっくり 溶けて いく ので この まま で”
(“The shape of the egg melts slowly as the temperature of the egg rises, so leave it as it is.”)

“少量 の 水 を かけて カボチャ が とっても 柔らかくなる まで 数分 です 電子レンジ で 加熱 僕 は カオス むちゃくちゃ 手 に 付くのが 凄い や だから 位相 加熱 する 前 に 包丁 で その 川”
(“Sprinkle a small amount of water and it will take a few minutes for the pumpkin to become very soft. Heat in the microwave.”)

“ニンニク と ナツメグ エメンタール チーズ 塩 コショウ 炒め 用 の バター と”
(“Garlic and nutmeg Emmental cheese with salt and pepper butter for stir fry.”)

“三月 は 全力 を 使います ね はい それで バーベキューチキン を 自分 出来上がり はい そして 次 は ですね かな カレー の 唐揚げした 事例 として”
(“I'll do my best in March, yes, so the barbecue chicken is ready, yes, and next is an example of fried curry.”)

“はい こんな 感じ で ねえ 白菜 大根 きゅうり 長さを 揃え ましてね マッチ 棒 ぐらい の 感じ で 切り そろえ ましたらこちらを ね ビニール 袋 など に 入れで”
(“Yes, like this. Cut Chinese cabbage, radish and cucumber to the same length. Cut them to the length of a matchstick and put them in a plastic bag.”)

Figure 2: *YouCook-Japanese video retrieval results with AVLnet. Top: retrieval after training on HowTo100M and without fine-tuning on YouCook-Japanese (zero-shot). Bottom: retrieval after training on HowTo100M and fine-tuning on YouCook-Japanese. Japanese ASR transcripts and English translations are shown, but AVLnet only uses audio as input. The correct clip match is in red.*

2. Related Work

Image and Spoken Caption Models. Several works [2, 3, 15] demonstrate the ability to learn semantic relationships between objects in images and the spoken words describing them using only the pairing between images and spoken captions as supervision. Using this framework, researchers have proposed improved image encoders, audio encoders, and loss functions [4–8, 16–20]. Harwath et al. [3, 4, 21] collected 400k spoken audio captions of images in the Places205 [22] dataset in English, which is one of the largest spoken caption datasets. For a recent survey of visually grounded models of spoken language, see Chrupała [23]. We instead use videos already available on the internet in English and other languages as our main source of training data. While we use the spoken narration naturally present in instructional videos, researchers have recently collected spoken captions for videos [24, 25].

Audio-Video Models. Instructional videos serve as a rich source of training data for learning the relationships between spoken words and visual content. While several models [26–29] have been proposed to learn from large-scale datasets such as How2 [30] and HowTo100M [13], they rely on having text transcripts. Our work instead builds from models that learn from the raw visual and audio channels. Boggust et al. [31] applied an image-caption [4] model to videos, using a single image frame from entire video clips to perform video to audio retrieval. Rouditchenko et al. [12] proposed the AVLnet model that pools visual information over entire clips using 2D and 3D visual CNNs. We use AVLnet trained on English HowTo100M videos and apply it to cooking videos in Japanese and images and spoken captions in Japanese and Hindi.

Multilingual Speech and Video Processing. The image and spoken caption models have been explored in the multilingual setting. Harwath et al. [9] collected 100k Hindi captions of Places images and proposed a bilingual audio-visual model. Building from this, Ohishi et al. [11] collected 100k Japanese captions and proposed a trilingual model. Other work has proposed bilingual models with synthetic spoken captions [10] and image text taggers [32], clustering for bilingual image-audio dictionaries [33], and pair expansion methods for learning from multilingual captions of disjoint images [34]. Instead of learning from multiple languages simultaneously, our approach is to learn from them one at a time in a cascade.

Several multilingual video datasets have been introduced, such as How2 [30] and VATEX [35] which contain parallel translations of English video captions in Portuguese and Chinese. Instead of collecting parallel translations, Sigurdsson et al. proposed versions of HowTo100M in Japanese, French, and Korean [36]. Thus far, all of the methods proposed on these datasets rely on text captions. Instead, we use AVLnet to learn from videos using speech audio and without requiring text.

Finally, multilingual ASR methods include simultaneous training on multiple languages [37–40] and cascaded approaches in which representations learned from one language are used as initialization for other languages [41, 42]. Our approach is similar to the cascaded methods, but it only requires audio-visual data without transcripts.

3. Technical Approach

3.1. Videos

Our goal is to learn audio-visual representations for videos in languages other than English using AVLnet [12]. AVLnet 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

Table 1: *Video retrieval on YouCook2 Videos (YC-EN) and YouCook-Japanese videos (YC-JP). HT100M=HowTo100M.*

(a) *English YouCook2 Videos (YC-EN)*

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	0.15	0.3	0.03	0.15	0.3
YC-EN	0.7	2.3	3.9	0.8	3.0	4.9
HT100M	27.4	51.6	61.5	27.3	51.2	60.8
HT100M + YC-EN	**30.7**	**57.7**	**67.4**	**33.0**	**58.9**	**68.4**
HT100M + YC-JP	19.4	40.4	51.3	19.8	43.5	53.7

(b) *YouCook-Japanese Videos (YC-JP)*

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
YC-JP	0.7	2.4	3.8	0.5	1.8	3.0
HT100M	4.6	12.1	18.2	5.6	14.6	21.3
HT100M + YC-EN	5.1	13.2	18.9	5.6	14.5	20.7
HT100M + YC-JP	**7.0**	**20.4**	**29.3**	**7.6**	**20.9**	**29.7**

only on the target videos in the new language. However, we find that hundreds of thousands of videos are necessary to learn strong representations, and there is simply not enough videos in 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. [36] 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 [43] 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. We report results on the held-out evaluation set and encourage other researchers to tune parameters only on the validation set.

3.2. 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.

Table 2: *Image retrieval on the Places Audio Caption 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. [10]	18.2	48.5	62.2	15.3	41.4	57.6
Ohishi et al. [34]	20.1	49.7	63.9	16.7	44.3	57.8
Ohishi et al. [11]	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. [9]	8.0	25.0	35.6	7.4	23.5	35.4
Havard et al. [10]	9.6	28.2	40.7	8.0	27.6	37.1
Ohishi et al. [34]	9.4	29.8	41.8	9.3	29.5	38.2
Ohishi et al. [11]	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**

Table 3: *Comparison of frozen versus trainable image encoder for fine-tuning on the Places Audio Caption dataset.*

Language	**Frozen Img. CNN**	**Audio to Image**			**Image to Audio**		
		R@1	R@5	R@10	R@1	R@5	R@10
Japanese	**Yes**	**23.5**	**57.3**	**70.4**	**24.3**	**56.6**	**70.0**
	No	20.8	50.9	64.9	20.9	49.5	63.5
Hindi	**Yes**	**15.2**	**38.9**	**51.1**	**17.0**	**39.8**	**51.5**
	No	12.1	30.9	44.1	11.9	30.8	41.7

4. Experiments

4.1. Datasets

Videos. We use the following instructional video datasets: HowTo100M [13] (1.2M videos), YouCook2 [14] and YouCook-Japanese. For YouCook2, we use 9,586 train clips and 3,350 validation clips as in [12]. We evaluate performance on audio to video clip retrieval and video clip to audio retrieval using the standard recall metrics R@1, R@5, R@10.

Images and Spoken Captions. We fine-tune and evaluate our model on the Places Audio Caption dataset [4], which contains 100k images from the Places205 dataset [22] each with a spoken caption in Japanese [11] and Hindi [9]. We evaluate the performance on audio to image and image to audio retrieval using the standard recall metrics R@1, R@5, R@10. We follow the prior work [9, 11] and report results on the validation sets of 1k images and spoken captions.

4.2. Implementation Details

For training AVLnet on HowTo100M, we follow the details in [12]. Each batch contains 128 videos with $M = 32$ clips per video, where each clip is $t = 10$ seconds long, for an effective batch size of $4,096$ clips per batch. The 2D visual feature extractor is a pre-trained ResNet-152 [44] model and the 3D visual feature extractor is a pre-trained ResNeXt-101 [45] model. Both are kept frozen during training. The trainable audio encoder is the ResDAVEnet model [5] which operates on log Mel filterbank spectrograms. For fine-tuning on video clips from YouCook2 and YouCook-Japanese, we use a batch size of 256 clips and a learning rate of 1e$-$4. We pad the audio or crop it up to 50 seconds in length. For fine-tuning AVLnet on images and spoken captions in Places, we either keep the ResNet-152 model frozen or fine-tune it. We use a learning rate of 1e$-$3 for

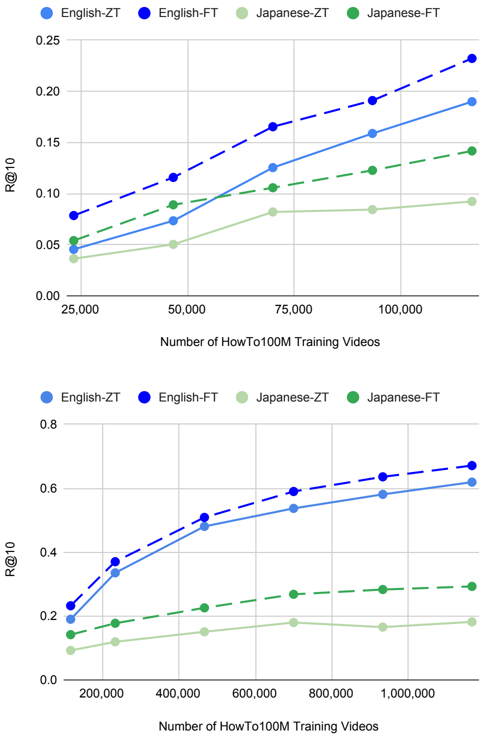


Figure 3: *Video retrieval performance when varying the % of HowTo100M videos. Top: {2,4,6,8,10}%. Bottom: {10,20,40,60,80,100}%. ZT=Zero-Shot, FT=Fine-tune.*

the frozen setting and a learning rate of 1e−4 for the trainable setting. Models were trained with the MMS loss [6].

4.3. Video Retrieval

YouCook2. Table 1a shows the video retrieval results on English YouCook2 videos. We note that some of the YouCook2 results have already been presented in [12], and we re-print them here for comparison with the results on YouCook-Japanese. Training on HowTo100M significantly improves performance compared with training only on YouCook2. In the zero-shot setting, ie. without fine-tuning on any YouCook2 videos, the model achieves strong retrieval performance, likely due to the similar instructional domain of HowTo100M and YouCook2 and shared language (English). Performance further improves after fine-tuning on YouCook2 videos. The final row of Table 1a shows that fine-tuning the model on YouCook-Japanese videos reduces the performance on YouCook2, indicating that the model is sensitive to the language present in the videos.

YouCook-Japanese. Table 1b shows the video retrieval results on YouCook-Japanese videos. AVLnet's performance when trained only on YouCook-Japanese is similar to AVLnet's performance on YouCook2 when trained only on YouCook2 videos, 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. In the zero-shot setting, ie. 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 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 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.

Qualitative results. Figure 2 shows qualitative YouCook-Japanese video retrieval results. In the zero-shot setting, without fine-tuning on Japanese videos, the model seems to perform retrieval using salient natural sounds, for example, sizzling sounds or microwave beeps. After fine-tuning the model on YouCook-Japanese, the model can handle more complex queries and retrieve video clips with specific ingredients mentioned in the audio queries.

Varying the % of HowTo100M videos. In Figure 3, we show the video retrieval performance when training AVLnet with a smaller percentage of HowTo100M videos. The plots show that performance generally increases with the number of HowTo100M videos. The gap between performance on English and Japanese videos is lower with 10% or less of the HowTo100M videos.

4.4. 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 [9, 11] 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.

Table 3 shows that retrieval results were higher for Japanese and Hindi with a frozen visual encoder (ResNet-152) than with a trainable encoder. We hypothesize that 100k images in the Japanese and Hindi training set is not enough to fine-tune the ResNet-152, and therefore it is better to leave it frozen. Furthermore, given that the visual encoders are also frozen during fine-tuning on videos, these results suggest that the visual branch is more language independent than the audio branch, and that the audio branch needs to be adapted to handle unseen languages.

5. 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. Our cascaded improves performance on the Japanese videos by nearly 10x compared to training on the Japanese videos solely. Our approach could plausibly work for instructional videos in any language. One direction that we plan to explore in the future is cross-lingual retrieval in videos, for example, retrieving Japanese audio from English audio through related videos.

6. Acknowledgements

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

7. References

[1] M. Prasad, D. van Esch, S. Ritchie, and J. F. Mortensen, "Building large-vocabulary asr systems for languages without any audio training data." in *INTERSPEECH*, 2019.

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

[3] D. Harwath, A. Torralba, and J. Glass, "Unsupervised learning of spoken language with visual context," in *NeurIPS*, 2016.

[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.

[5] ——, "Jointly discovering visual objects and spoken words from raw sensory input," *IJCV*, 2020.

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

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

[8] D. Merkx, S. L. Frank, and M. Ernestus, "Language learning using speech to image retrieval," in *INTERSPEECH*, 2019.

[9] D. Harwath, G. Chuang, and J. Glass, "Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech," in *ICASSP*, 2018.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] L. Zhou, C. Xu, and J. J. Corso, "Towards automatic learning of procedures from web instructional videos," in *AAAI*, 2018.

[15] D. Harwath and J. Glass, "Deep multimodal semantic embeddings for speech and images," in *ASRU*, 2015.

[16] D. Harwath, W.-N. Hsu, and J. Glass, "Learning hierarchical discrete linguistic units from visually-grounded speech," in *ICLR*, 2020.

[17] D. Suris, A. Recasens, D. Bau, D. Harwath, J. Glass, and A. Torralba, "Learning words by drawing images," in *CVPR*, 2019.

[18] M. S. Mortazavi, "Speech-image semantic alignment does not depend on any prior classification tasks," *INTERSPEECH*, 2020.

[19] R. Sanabria, A. Waters, and J. Baldridge, "Talk, don't write: A study of direct speech-based image retrieval," *arXiv preprint arXiv:2104.01894*, 2021.

[20] L. Wang, X. Wang, M. Hasegawa-Johnson, O. Scharenborg, and N. Dehak, "Align or attend? toward more efficient and accurate spoken word discovery using speech-to-image retrieval," in *ICASSP*, 2021.

[21] D. Harwath and J. Glass, "Learning word-like units from joint audio-visual analysis," in *ACL*, 2017.

[22] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *NeurIPS*, 2014.

[23] G. Chrupała, "Visually grounded models of spoken language: A survey of datasets, architectures and evaluation techniques," *arXiv preprint arXiv:2104.13225*, 2021.

[24] M. Monfort, S. Jin, A. Liu, D. Harwath, R. Feris, J. Glass, and A. Oliva, "Spoken moments: Learning joint audio-visual representations from video descriptions," in *CVPR*, 2021.

[25] A.-M. Oncescu, J. F. Henriques, Y. Liu, A. Zisserman, and S. Albanie, "Queryd: A video dataset with high-quality textual and audio narrations," *arXiv preprint arXiv:2011.11071*, 2020.

[26] N. Holzenberger, S. Palaskar, P. Madhyastha, F. Metze, and R. Arora, "Learning from multiview correlations in open-domain videos," in *ICASSP*, 2019.

[27] M. Wray, D. Larlus, G. Csurka, and D. Damen, "Fine-grained action retrieval through multiple parts-of-speech embeddings," in *ICCV*, 2019.

[28] J.-B. Alayrac, A. Recasens, R. Schneider, R. Arandjelović, J. Ramapuram, J. De Fauw, L. Smaira, S. Dieleman, and A. Zisserman, "Self-supervised multimodal versatile networks," *NeurIPS*, vol. 33, 2020.

[29] B. Chen, A. Rouditchenko, K. Duarte, H. Kuehne, S. Thomas, A. Boggust, R. Panda, B. Kingsbury, R. Feris, D. Harwath *et al.*, "Multimodal clustering networks for self-supervised learning from unlabeled videos," *arXiv preprint arXiv:2104.12671*, 2021.

[30] R. Sanabria, O. Caglayan, S. Palaskar, D. Elliott, L. Barrault, L. Specia, and F. Metze, "How2: a large-scale dataset for multimodal language understanding," in *Workshop on Visually Grounded Interaction and Language (ViGIL)*. NeurIPS, 2018.

[31] A. Boggust, K. Audhkhasi, D. Joshi, D. Harwath, S. Thomas, R. Feris, D. Gutfreund, Y. Zhang, A. Torralba, M. Picheny, and J. Glass, "Grounding spoken words in unlabeled video," in *CVPR Sight and Sound Workshop*, 2019.

[32] H. Kamper and M. Roth, "Visually grounded cross-lingual keyword spotting in speech," *arXiv preprint arXiv:1806.05030*, 2018.

[33] E. Azuh, D. Harwath, and J. R. Glass, "Towards bilingual lexicon discovery from visually grounded speech audio." in *INTERSPEECH*, 2019.

[34] 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.

[35] X. Wang, J. Wu, J. Chen, L. Li, Y.-F. Wang, and W. Y. Wang, "Vatex: A large-scale, high-quality multilingual dataset for video-and-language research," in *ICCV*, 2019.

[36] 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.

[37] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *ICASSP*, 2013.

[38] M. Karáfidt, M. K. Baskar, K. Veselỳ, F. Grézl, L. Burget, and J. Černockỳ, "Analysis of multilingual blstm acoustic model on low and high resource languages," in *ICASSP*, 2018.

[39] J. Cho, M. K. Baskar, R. Li, M. Wiesner, S. H. Mallidi, N. Yalta, M. Karafiat, S. Watanabe, and T. Hori, "Multilingual sequence-to-sequence speech recognition: architecture, transfer learning, and language modeling," in *SLT*, 2018.

[40] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[41] S. Thomas, S. Ganapathy, and H. Hermansky, "Multilingual mlp features for low-resource lvcsr systems," in *ICASSP*, 2012.

[42] A. Ghoshal, P. Swietojanski, and S. Renals, "Multilingual training of deep neural networks," in *ICASSP*, 2013.

[43] 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.

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[45] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" in *CVPR*, 2018.