INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Pair Expansion for Learning Multilingual Semantic Embeddings using Disjoint Visually-grounded Speech Audio Datasets

Yasunori Ohishi[1], *Akisato Kimura*[1], *Takahito Kawanishi*[1], *Kunio Kashino*[1], *David Harwath*[2], *and James Glass*[2]

[1]NTT Corporation, Japan
[2]MIT Computer Science and Artificial Intelligence Laboratory, USA

yasunori.ooishi.uk@hco.ntt.co.jp

Abstract

We propose a data expansion method for learning a multilingual semantic embedding model using disjoint datasets containing images and their multilingual audio captions. Here, disjoint means that there are no shared images among the multiple language datasets, in contrast to existing works on multilingual semantic embedding based on visually-grounded speech audio, where it has been assumed that each image is associated with spoken captions of multiple languages. Although learning on disjoint datasets is more challenging, we consider it crucial in practical situations. Our main idea is to refer to another paired data when evaluating a loss value regarding an anchor image. We call this scheme "pair expansion". The motivation behind this idea is to utilize even disjoint pairs by finding similarities, or commonalities, that may exist in different images. Specifically, we examine two approaches for calculating similarities: one using image embedding vectors and the other using object recognition results. Our experiments show that expanded pairs improve crossmodal and cross-lingual retrieval accuracy compared with non-expanded cases. They also show that similarities measured by the image embedding vectors yield better accuracy than those based on object recognition results.

Index Terms: Vision and spoken language, multilingual semantic embeddings, disjoint datasets, pair expansion, cross-lingual retrieval

1. Introduction

As the accuracy of visual object recognition and speech recognition improves, these applications have spread across various fields. However, these recognition systems depend on learning datasets and require that the class labels to be recognized are defined in them. Preparing such a dataset is becoming recognized as a common bottleneck in terms of dataset construction costs and the difficulty in a priori definition of the classes to be recognized. In order to solve this problem, we aim to develope a robust, natural, and human-like learning scheme.

Many unsupervised learning methods have been studied so far [1–8]. Among them, we focus on knowledge acquisition based on co-occurrences of multiple modality inputs, such as visual information and multilingual speech information. In the literature, the so-called DAVEnet has been proposed that employs multiple encoders to convert information of different modalities to vectors in the common space [9–14]. These encoders were learned on the basis of metric learning using a large amount of paired data consisting of images and their speech captions. Multilingual models based on a dataset consisting of English, Hindi, and Japanese speech captions for a common image set has been shown to enable the acquisition of translation knowledge using the images as pivots [15–17].

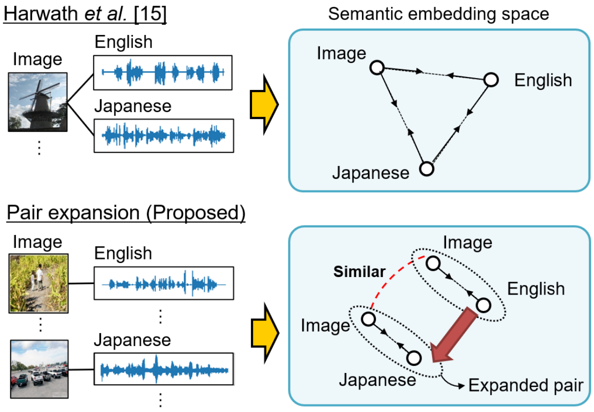

Figure 1: *Illustration of our method*

In these studies, as shown in the upper part of Figure 1, a set of aligned data, that is, multilingual audio captions describing a common image, were used for training. For example, given a set of images and their English/Japanese audio captions, the embedding space was constructed by metric learning in a pairwise manner. However, specific preparation is needed to build such a multilingual corpus, and in practice, a learning method that can be trained on disjoint datasets is desirable. Here, disjoint means there are no images shared by multiple language audio captions, as shown in the lower part of Fig. 1. This is because it is easier to collect disjoint datasets using e.g. TV shows or narrated videos on the Web [18–21], compared with the aligned datasets.

Thus, we propose a pair expansion method that can expand the image-audio caption pairs even for disjoint datasets. The method utilizes image similarities to find useful paired data for training. In this paper, we specifically examine two approaches for calculating the image similarities, one using image embedding vectors during training and the other using object recognition results for the images. Our experiments show that the expanded positive pairs improved crossmodal and cross-lingual retrieval accuracy compared with the original paired data. We also show that the use of image embedding vectors yields better results than the object recognition results.

Our paper is organized as follows. Section 2 introduces related work. Section 3 describes our method. Section 4 evaluates the effectiveness of our method in terms of crossmodal and cross-lingual retrieval tasks. Finally, Section 5 concludes this paper.

http://dx.doi.org/10.21437/Interspeech.2020-3078

2. Related work

Many methods have been proposed to construct an embedding space based on the correspondence between visual and natural language information and applied to crossmodal search, cross-lingual search, and machine translation [22–26]. In those methods, images and text captions in multiple languages were represented by vectors using corresponding encoders, and a latent space was constructed by metric learning based on paired data. Recently, Kádár *et al.* used disjoint image/text caption datasets for learning, which do not overlap images for each language [27, 28]. In [28], they demonstrated the effectiveness of creating pseudo pairs across datasets using the similarity between the embedding vectors of text captions and utilizing them in the metric learning process.

Models representing images and spoken languages, in the form of audio signals, in a common embedding space have also been proposed [13, 29, 30]. For example, an embedding space was constructed using English and Hindi spoken captions for a common image set [15,16]. Cross-lingual search and translation knowledge acquisition were also confirmed to be possible by using a trilingual model including English, Hindi, and Japanese speech captions [17]. Havard *et al.* built a monolingual model for each language of the English and Japanese speech caption datasets without common images, and achieved cross-lingual search using images as pivots [29]. Our approach is similar in motivation to those works, but novel in that it utilizes disjoint image-audio relationships. Unlike [28], for example, our method finds another image-audio pair in the learning process for a pivot image.

Other related works include studies dealing with the co-occurrence of visuals and speech, such as the association of handwritten digits and spoken numbers [31, 32], visually-grounded keyword spotting [33–35], image-based audio caption generation [36,37], applications to speech recognition [38], audio-visual representation learning [39, 40], and grounding spoken words in narrated videos [41].

3. Method

Given a triplet (I_i, A_i^X, A_i^Y) consisting of an image I and speech audio captions A in language X and Y for I, the parameters of the corresponding encoders are learned on the basis of the margin rank criterion [42], so that the d-dimensional embedding vectors $(\boldsymbol{I}_i, \boldsymbol{A}_i^X, \boldsymbol{A}_i^Y)$ obtained by inputting these into the image encoder and the audio encoders are placed close to each other [15]. The overall loss function can be written as:

$$\begin{aligned}\mathcal{L}_c = \sum_{i=1}^{B} \Big(& \mathrm{rank}(\boldsymbol{I}_i, \boldsymbol{A}_i^X, \boldsymbol{A}_{i_1}^X) + \mathrm{rank}(\boldsymbol{A}_i^X, \boldsymbol{I}_i, \boldsymbol{I}_{i_2}) \\ & + \mathrm{rank}(\boldsymbol{I}_i, \boldsymbol{A}_i^Y, \boldsymbol{A}_{i_3}^Y) + \mathrm{rank}(\boldsymbol{A}_i^Y, \boldsymbol{I}_i, \boldsymbol{I}_{i_4}) \\ & + \mathrm{rank}(\boldsymbol{A}_i^X, \boldsymbol{A}_i^Y, \boldsymbol{A}_{i_5}^Y) + \mathrm{rank}(\boldsymbol{A}_i^Y, \boldsymbol{A}_i^X, \boldsymbol{A}_{i_6}^X) \Big). \quad (1)\end{aligned}$$

The rank function for each term is:

$$\mathrm{rank}(\boldsymbol{a}, \boldsymbol{p}, \boldsymbol{n}) = \max(0, \eta - s(\boldsymbol{a}, \boldsymbol{p}) + s(\boldsymbol{a}, \boldsymbol{n})), \quad (2)$$

where $\boldsymbol{a}$ is an anchor vector, $\boldsymbol{p}$ is a positive vector paired with the anchor, $\boldsymbol{n}$ is a negative vector not paired with the anchor, $s(\boldsymbol{a}, \boldsymbol{p})$ is an inner product $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{p}$, and η is a hyperparameter representing a margin. The parameters of the encoders are learned so that the inner product of the anchor vector and the positive vector is larger than that of the negative vector. The negative vector indices $i_1, i_2, \dots, i_6$ are generally chosen at random from the mini-batch size B. The effectiveness of using semi-hard negative mining has been also reported in [13].

In this paper, we assume that the triplets (I_i, A_i^X, A_i^Y) cannot be obtained as input data. Our objective is to learn a multilingual semantic space under the condition that (I_i^X, A_i^X), a pair of an image and its audio caption in language X, and (I_j^Y, A_j^Y), a pair of an image and its audio caption in language Y, are respectively composed of the image sets that do not contain any images in common. In order to achieve this goal, we expand the image-audio caption pairs to the disjoint datasets on the basis of the image similarity. Here, we consider two approaches for calculating the image similarities as follows.

3.1. Embedded vector-based pair expansion

The first approach is to measure similarity on the basis of the embedded vectors in the embedding space. That is, this approach can be considered modality-independent. If the content of the images are similar, we assume that the audio captions of the images have similar intentions even if the languages are different. We refer to the pairs comprising such images and its audio captions as the *expanded* pairs. We use them in the computation of the cross-lingual terms of Eq. (1). First, given the images $\{I_i^X\}_{i=1}^B$, we sample N images $\{I_n^Y\}_{n=1}^N$ that are associated with the audio captions in language Y. Next, we calculate the similarity between those images using the inner product of their embedding vectors during training, to obtain a similarity matrix $\boldsymbol{S}$ of the size $B \times N$. Then, for image I_i^X, the most similar image $I_{l_i}^Y$ is selected among the N images as follows:

$$\boldsymbol{S}_{i,n} = s(\boldsymbol{I}_i^X, \boldsymbol{I}_n^Y), \quad l_i = \operatorname*{argmax}_n \boldsymbol{S}_{i,:}. \quad (3)$$

Finally, the audio caption $A_{l_i}^Y$ and the image $I_{l_i}^Y$ are considered as an expanded pair for the audio caption A_i^X of the image I_i^X. For the images $\{I_j^Y\}_{j=1}^B$, on the basis of the similarity matrix calculated in a similar manner, audio caption $A_{m_j}^X$ associated with $I_{m_j}^X$ is regarded as a positive counterpart to the image I_j^Y. Note that this method does not require additional pre-trained models except for the disjoint datasets, and it only incorporates new pairs as the learning proceeds.

On the basis of such expanded pairs, the overall loss function can be defined as follows:

$$\begin{aligned}\mathcal{L}_s = & \sum_{i=1}^{B} \Big(\mathrm{rank}(\boldsymbol{I}_i^X, \boldsymbol{A}_i^X, \boldsymbol{A}_{i_1}^X) + \mathrm{rank}(\boldsymbol{A}_i^X, \boldsymbol{I}_i^X, \boldsymbol{I}_{i_2}^X) \Big) \\ & + \sum_{j=1}^{B} \Big(\mathrm{rank}(\boldsymbol{I}_j^Y, \boldsymbol{A}_j^Y, \boldsymbol{A}_{j_1}^Y) + \mathrm{rank}(\boldsymbol{A}_j^Y, \boldsymbol{I}_j^Y, \boldsymbol{I}_{j_2}^Y) \Big) \\ & + \sum_{i=1}^{B} \mathrm{rank}(\boldsymbol{A}_i^X, \boldsymbol{A}_{l_i}^Y, \boldsymbol{A}_{l'}^Y) + \sum_{j=1}^{B} \mathrm{rank}(\boldsymbol{A}_j^Y, \boldsymbol{A}_{m_j}^X, \boldsymbol{A}_{m'}^X).\end{aligned} \quad (4)$$

In practice, we combined the sampling-based triplet loss and the semi-hard negative training with equal weights in accordance with [13]'s results. For both the randomly sampled and semi-hard negative mined loss functions, we selected the negative vectors from the same mini-batch, such that $i_1 \neq i, i_2 \neq i, j_1 \neq j, j_2 \neq j, l' \neq l_i$, and $m' \neq m_j$.

3.2. Object recognition-based pair expansion

The second approach is to calculate the image similarity on the basis of the object recognition results. This is a modality-specific approach. Among many image processing methods, we

Table 1: *Retrieval recall scores on the validation set in the disjoint setting for models trained on the English and Hindi spoken captions. [29]'s results are also reported for comparison.*

	I→E	**E→I**	**I→H**	**H→I**	**H→E**	**E→H**
Baseline						
R@10	.361	.397	.348	.370	.079	.069
R@5	.249	.270	.246	.279	.040	.048
R@1	.084	.080	.071	.085	.007	.016
VGG16 ($N = 10^2$)						
R@10	.389	.425	.354	.389	.090	.108
R@5	.287	.315	.278	.283	.056	.059
R@1	.095	.113	.094	.096	.013	.014
Embedding ($N = 10^2$)						
R@10	.412	.425	.376	.390	.125	.132
R@5	.284	.310	**.281**	.282	.078	.078
R@1	.100	.114	.087	.102	.019	.021
Embedding ($N = 10^3$)						
R@10	.407	.424	**.368**	.391	**.140**	.148
R@5	.284	.315	.271	.285	.084	.089
R@1	.096	.110	.093	**.105**	.017	.025
Embedding ($N = 10^4$)						
R@10	**.414**	**.430**	**.368**	**.414**	.139	**.156**
R@5	**.299**	**.321**	.266	**.286**	**.092**	**.091**
R@1	**.101**	**.121**	**.102**	.093	**.026**	**.029**
[29]						
R@10	.396	.425	.371	.407	.075	.077
R@5	.288	.313	.276	.282	.036	.042
R@1	.093	.094	.080	.096	.011	.007

Table 2: *Retrieval recall scores on the validation set in the aligned setting for the model trained on the English and Hindi spoken captions [15].*

	I→E	**E→I**	**I→H**	**H→I**	**H→E**	**E→H**
[15]						
R@10	.475	.501	.382	.418	.235	.223
R@5	.336	.367	.295	.298	.150	.156
R@1	.113	.137	.093	.094	.049	.040

Table 3: *Recall scores on the validation set in the disjoint setting for models trained on the English and Japanese spoken captions. [29]'s results are also reported for comparison.*

	I→E	**E→I**	**I→J**	**J→I**	**J→E**	**E→J**
Baseline						
R@10	.402	.433	.576	.607	.107	.080
R@5	.306	.330	.422	.485	.074	.055
R@1	.109	.116	.149	.182	.026	.019
VGG16 ($N = 10^2$)						
R@10	.415	.431	.553	.627	.142	.138
R@5	.296	.316	.432	.489	.087	.084
R@1	.105	.120	.150	.187	.023	.023
Embedding ($N = 10^2$)						
R@10	.425	.438	.578	.617	.170	.189
R@5	**.310**	.330	.423	.485	.109	.117
R@1	.103	.127	.154	.180	.031	.039
Embedding ($N = 10^3$)						
R@10	.405	.440	**.582**	**.629**	.203	.216
R@5	.299	.324	.425	**.489**	**.122**	.136
R@1	.101	.119	**.158**	**.190**	.031	.039
Embedding ($N = 10^4$)						
R@10	**.410**	**.442**	.574	.620	**.211**	**.233**
R@5	.295	**.340**	**.433**	.487	.121	**.144**
R@1	**.111**	**.126**	.156	.182	**.035**	**.047**
[29]						
R@10	.396	.425	.576	.622	.089	.149
R@5	.288	.313	.414	.485	.047	.077
R@1	.093	.094	.153	.182	.010	.010

Table 4: *Retrieval recall scores on the validation set in the aligned setting for the model trained on the English and Japanese spoken captions [15].*

	I→E	**E→I**	**I→J**	**J→I**	**J→E**	**E→J**
[15]						
R@10	.477	.516	.578	.639	.343	.352
R@5	.353	.402	.443	.497	.240	.234
R@1	.123	.138	.167	.201	.074	.073

employ the output of VGG16 [43], which consists of the posterior probabilities of 1,000 categories. Letting $\boldsymbol{O}_i^X$ and $\boldsymbol{O}_n^Y$ be the outputs for those images, we substitute the inner product of Eq. (3) with the Jensen-Shannon (JS) divergence [44] of $\boldsymbol{O}_i^X$ and $\boldsymbol{O}_n^Y$ as follows:

$$\boldsymbol{S}_{i,n} = -JS(\boldsymbol{O}_i^X, \boldsymbol{O}_n^Y), \quad l_i = \underset{n}{\operatorname{argmax}}\, \boldsymbol{S}_{i,:}. \tag{5}$$

The loss function is the same as Eq. (4). This method uses a pre-trained VGG16 model.

4. Experiments

We evaluated effectiveness of the proposed method in terms of crossmodal and cross-lingual retrieval tasks. For the experiments, we prepared two disjoint combinations of bilingual image-audio caption datasets: (1) English-Hindi and (2) English-Japanese, as follows. The Places205 [45] Hindi audio caption dataset [15] and Japanese audio caption dataset [17] contain common images with their audio captions in their respective languages. That is, Hindi-Japanese combination makes an aligned dataset. Therefore, we chose 97,555 image-audio pairs from both datasets, and then, we selected another subset of 97,555 image-caption pairs from the Places205 English audio caption dataset [13]. The English dataset originally contains 400,000 recordings, and we randomly sampled 97,555 images so that none of those images are contained in the abovementioned Hindi or Japanese datasets. These two sets of disjoint bilingual (English-Hindi and English-Japanese) image-audio caption datasets were used for training.

In the crossmodal retrieval task, an audio caption in one language was considered as a query and its associated image was the target, and vice versa. In the cross-lingual retrieval task, we considered that the captions in one language were the queries and those in another language associated with the same image were the targets. To do this, we chose 1,000 quadruplets as a validation set, each of which consisted of an image and its audio captions in Hindi, Japanese, and English.

We used an image encoder that takes all layers up through `conv5` from a pre-trained VGG16 network [43]. To map the VGG16 output into the embedding space, we applied a linear 3×3 convolution with d filters, followed by spatial mean pooling. For a 224×224 pixel RGB input image, the encoder outputs a vector of dimension d. Our speech encoder was based

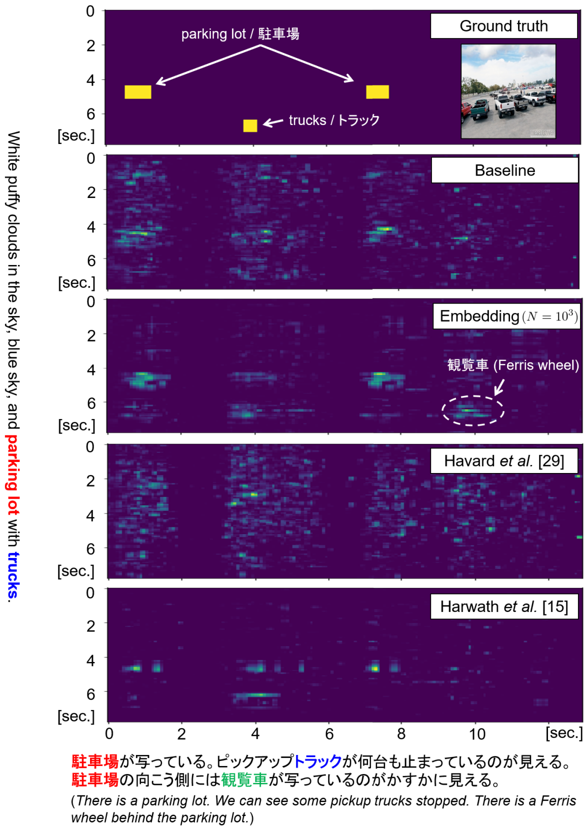

Figure 2: *Ground truth and similarity matrices between unpooled embeddings of English and Japanese captions. A text in each of the languages corresponding to the audio caption is displayed on each axis. A text obtained by translating the Japanese text into English is illustrated under the text in Japanese.*

on DAVEnet [13], for simplicity, while it is also possible to use ResDAVEnet [13] or other speech encoders for modeling the temporal nature of the audio captions within a gated recurrent unit and self-attention layer [17, 29, 40]. The audio inputs were converted to 40 log-Mel filterbank energies with a 25-ms frame with 10-ms shifts, and each speech encoder outputs an embedding vector of dimension d obtained by temporal mean pooling. We applied truncation and zero-padding of each spectrogram to a fixed length of T frames ($T = 2048$, or approximately 20 seconds in our experiments), and then truncate the output features of each caption to remove the frames corresponding to zero-padding. Such pre-processing followed the one in [15].

We set the mini-batch size B and dimension d to 100 and 1024, respectively, and used a constant momentum of 0.9. The initial learning rate was 0.001 and was decreased by a factor of 40 every ten epochs. Our model generally converged in less than 100 epochs.

Tables 1 and 3 show the crossmodal and cross-lingual retrieval recall scores in the disjoint settings at ranks 1, 5, and 10. We refer to the case of training on only the disjoint data as the baseline, where the loss function does not include the cross-lingual terms. Tables 2 and 4 list the recall scores in the aligned settings [15], where we used 97,555 image-caption quadruples for training that consist of the English, Hindi and Japanese audio captions associated with the same images. "English caption" is abbreviated as E, "Hindi caption" as H, "Japanese caption" as J, and "Image" as I. It is natural that the best scores came from training on the aligned dataset, but it is shown that our method achieved an accuracy closer to them without using the aligned data. It is also shown that using the image embedding vectors to expand positive pairs results in a better performance than using the image similarity based on VGG16 vectors, and the accuracy tends to be improved by increasing N. Tables 1 and 3 also list the results from the method proposed in [29], where we used 2,000 images as pivots to link the disjoint embedding spaces. In our experimentation, our method outperformed [29]. We consider this may indicate that the multilingual embedding model is more suitable than linking multiple individually-trained models after the training.

Table 5: *Cross-lingual word-to-word retrieval recall scores. "English word" is abbreviated as E and "Japanese word" as J.*

	J→E	**E→J**
Baseline	.424	.382
Embedding $(N = 10^3)$	**.473**	**.419**
Havard *et al.* [29]	.180	.169
Harwath *et al.* [15] (aligned)	.481	.432

We further examined whether word-level translation alignments were being learned using English and Japanese spoken caption pairs associated with the same images. First, 100 English and Japanese caption pairs were randomly selected from the validation set, and the bilingual word alignment was manually annotated as ground truths. Then, the matrix product of the outputs before the temporal mean pooling of the speech encoders for the English and Japanese captions was computed as a similarity matrix, where the regions of high similarity indicate the correspondence between the underlying words. Figure 2 compares the ground truth and the similarity matrices in the different settings. It can be seen that our method with $N = 10^3$ and [15] result in less noise and a clearer alignment in the similarity matrices. It is interesting that "観覧車 (Ferris wheel)" and "trucks", which are similar in a generic concept, are aligned to each other in our method. Next, we computed Recall@1 when searching for the labeled word segments in one language from those in another language in the similarity matrix. Table 5 shows that our method improved word-to-word retrieval performance compared with training on only the disjoint data.

5. Conclusion

We proposed a pair expansion method for crossmodal and cross-lingual semantic embeddings using multiple monolingual image and audio caption datasets where no images are shared across different languages. The method finds an image-audio caption pair and utilizes it in the learning process. Experiments using the crossmodal and cross-lingual retrieval tasks showed that the use of the expanded pairs clearly improves the accuracy. In particular, the method measuring the image similarity in the embedding space in the training process, called the embedded vector-based method, does not depend on any external data or models, such as object recognition, speech recognition, or machine translation. We hope this will enable us to fully utilize the monolingual resources in multilingual learning. Our future work will include examining the effectiveness of the proposed method with partially aligned datasets and disjoint datasets drawn from different image sources other than Places205 dataset.

6. References

[1] A. Jansen, K. Church, and H. Hermansky, "Toward spoken term discovery at scale with zero resources," in *Proc. INTERSPEECH*, 2010.

[2] A. Park and J. Glass, "Unsupervised pattern discovery in speech," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, pp. 186–197, 2008.

[3] H. Kamper, A. Jansen, and S. Goldwater, "Unsupervised word segmentation and lexicon discovery using acoustic word embeddings," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 24, pp. 669–679, 2016.

[4] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Proc. NIPS*, 2016.

[5] H. Nakayama and N. Nishida, "Zero-resource machine translation by multimodal encoder-decoder network with multimedia pivot," *Machine Translation*, vol. 32, pp. 49–64, 2017.

[6] A. Owens and A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proc. ECCV*, 2018.

[7] R. Arandjelović and A. Zisserman, "Objects that sound," in *Proc. ECCV*, 2018.

[8] A. Miech, J.-B. Alayrac, L. Smaira, I. Laptev, J. Sivic, and A. Zisserman, "End-to-end learning of visual representations from uncurated instructional videos," in *Proc. CVPR*, 2020.

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

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

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

[12] D. Harwath, W.-N. Hsu *et al.*, "Towards visually grounded subword speech unit discovery," in *Proc. ICASSP*, 2019.

[13] D. Harwath, A. Recasens, D. Surís, G. Chuang, A. Torralba, and J. Glass, "Jointly discovering visual objects and spoken words from raw sensory input," *International Journal of Computer Vision*, 2019.

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

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

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

[17] 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 *Proc. ICASSP*, 2020.

[18] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A large video description dataset for bridging video and language," in *Proc. CVPR*, 2016.

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

[20] D. Zhukov, J.-B. Alayrac, R.-G. Cinbis, D. Fouhey, I. Laptev, and J. Sivic, "Cross-task weakly supervised learning from instructional videos," in *Proc. CVPR*, 2019.

[21] 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 *Proc. ICCV*, 2019.

[22] S. Gella, R. Sennrich, F. Keller, and M. Lapata, "Image pivoting for learning multilingual multimodal representation," in *Proc. EMNLP*, 2017.

[23] I. Calixto and Q. Liu, "Sentence-level multilingual multi-modal embeddings for natural language processing," in *Proc. RANLP*, 2017.

[24] J. Wehrmann, D.-M. Souza, M.-A. Lopes, and R.-C. Barros, "Language-agnostic visual-semantic embeddings," in *Proc. ICCV*, 2019.

[25] S. Ruder, I. Vulic, and A. Søgaard, "A survey of cross-lingual word embedding models," *Journal of Artificial Intelligence Research*, vol. 65, pp. 569–630, 2019.

[26] A. Mohammadshahi, R. Lebret, and K. Aberer, "Aligning multilingual word embeddings for cross-modal retrieval task," in *Proc. IJCNLP*, 2019.

[27] Á. Kádár, D. Elliott, M.-A. Côté, G. Chrupała, A. Alishahi, and D. Elliott, "Lessons learned in multilingual grounded language learning," in *Proc. CoNLL*, 2018.

[28] Á. Kádár, G. Chrupała, A. Alishahi, and D. Elliott, "Bootstrapping disjoint datasets for multilingual multimodal representation learning," arXiv preprint arXiv:1911.03678, 2019.

[29] W. 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 *Proc. ICASSP*, 2019.

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

[31] K. Leidal, D. Harwath, and J. Glass, "Learning modality-invariant representations for speech and images," in *Proc. ASRU*, 2017.

[32] R. Eloff, H. Engelbrecht, and H. Kamper, "Multimodal one-shot learning of speech and images," in *Proc. ICASSP*, 2019.

[33] H. Kamper, S. Settle, G. Shakhnarovich, and K. Livescu, "Visually grounded learning of keyword prediction from untranscribed speech," in *Proc. Interspeech*, 2017.

[34] H. Kamper and M. Roth, "Visually grounded cross-lingual keyword spotting in speech," in *Proc. SLTU*, 2018.

[35] H. Kamper, A. Anastassiou, and K. Livescu, "Semantic query-by-example speech search using visual grounding," in *Proc. ICASSP*, 2019.

[36] M. Hasegawa-Johnson, A. Black, L. Ondel, O. Scharenborg, and F. Ciannella, "Image2speech: Automatically generating audio descriptions of images," in *Proc. ICNLSSP*, 2017.

[37] O. Scharenborg *et al.*, "Linguistic unit discovery from multimodal inputs in unwritten languages: Summary of the *speaking rosetta* JSALT 2017 workshop," in *Proc. ICASSP*, 2018.

[38] W.-N. Hsu, D. Harwath, and J. Glass, "Transfer learning from audio-visual grounding to speech recognition," in *Proc. ICASSP*, 2019.

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

[40] G. Chrupała, "Symbolic inductive bias for visually grounded learning of spoken language," in *Proc. ACL*, 2019.

[41] A. Boggust, K. Audhkhasi, D. Joshi, D. Harwath, S. Thomas, R. Feris, D. Gutfreund, Y. Zhang, A. Torralba, and M. Picheny, "Grounding spoken words in unlabeled video," in *Proc. CVPRW*, 2019.

[42] A. Karpathy, A. Joulin, and L. Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," in *Proc. NIPS*, 2014.

[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015.

[44] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NIPS*, 2014.

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