TRILINGUAL SEMANTIC EMBEDDINGS OF VISUALLY GROUNDED SPEECH WITH SELF-ATTENTION MECHANISMS

Yasunori Ohishi†, *Akisato Kimura*†, *Takahito Kawanishi*†, *Kunio Kashino*†,
David Harwath††, *James Glass*††

†NTT Communication Science Laboratories, NTT Corporation, Atsugi, Kanagawa, Japan
††MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, USA

ABSTRACT

We propose a trilingual semantic embedding model that associates visual objects in images with segments of speech signals corresponding to spoken words in an unsupervised manner. Unlike the existing models, our model incorporates three different languages, namely, English, Hindi, and Japanese. To build the model, we used the existing English and Hindi datasets and collected a new corpus of Japanese speech captions. These spoken captions are spontaneous descriptions by individual speakers, rather than readings based on prepared transcripts. Therefore, we introduce a self-attention mechanism into the model to better map the spoken captions associated with the same image into the embedding space. We hope that the self-attention mechanism efficiently captures relationships between widely separated word-like segments. Experimental results show that the introduction of a third language improves the average performance in terms of cross-modal and cross-lingual retrieval accuracy, and that the self-attention mechanism added to the model works effectively.

Index Terms— Vision and spoken language, semantic embedding space, self-attention, and cross-lingual retrieval

1. INTRODUCTION

As the accuracy of visual object recognition improves, its application expands to more fields. However, recognition systems depend heavily on the learning datasets, and objects whose class labels are not included in the dataset are not correctly recognized. This increases the cost of dataset construction and calls for a complete definition of the classes to be recognized.

To cope with this problem, various unsupervised learning methods have been studied [1–7]. Among them, we focus on knowledge acquisition using co-occurrences between the visual information and spoken-language information, without human labeling. Harwath et al. [8–11] proposed a crossmodal information embedding model to associate visual objects with spoken words. The model comprises image and speech encoders that map corresponding signals, image or speech audio signals, to vectors in a shared embedding space. To train

Fig. 1. Spoken audio captions associated with the same image

the encoders, many pairs comprising a picture and its spoken audio caption were used. They showed the model's effectiveness through crossmodal information retrieval and co-localization tasks. They also demonstrated that training bilingual, namely, English and Hindi encoders for a common image dataset allows to obtain some kind of translation knowledge [12, 13].

In this paper, we extend their models to incorporate three different languages: English, Hindi, and a new corpus of Japanese-spoken captions. A natural question in this extension is whether the additional information provided by the third language improves the performance in terms of cross-modal information retrieval accuracy.

The problem arising with this extension is that the audio captions associated with the same image are not necessarily parallel data. In our setup, a pair comprising an image and an audio caption is intended to be an artificial co-occurrence of that information, so the audio caption is a spontaneous description by an individual speaker, rather than a reading from a prepared transcript. As shown in Fig. 1, considerable variation exists in content and duration due to differences in the culture and vocabulary behind languages. If such non-parallel data is directly mapped into the shared embedding space, the

Authorized licensed use limited to: University of Texas at Austin. Downloaded on September 03,2020 at 14:50:43 UTC from IEEE Xplore. Restrictions apply.

Table 1. Average duration and number of words per caption for each language

	Average duration	Average word count
English [11]	9.5 seconds	19.3 words
Hindi [12]	11.4 seconds	20.4 words
Japanese	19.7 seconds	44.6 words

appropriate semantics would not be represented in the embedding space. Against this background, we introduce a self-attention mechanism into the speech encoders. Our hope is that the self-attention mechanism efficiently captures relationships between widely separated word-like segments, and that the spoken captions associated with the same image are better mapped into the embedding space. Our experiments show that incorporating the third language improves the accuracy of the retrieval task, and that the self-attention mechanism works effectively. We also show that the proposed model acquires more accurate word translation knowledge than the model without the self-attention mechanism does.

The rest of this paper is organized as follows. Section 2 describes related work. Section 3 explains our dataset. Section 4 describes our model, and Section 5 shows experimental results. Finally, Section 6 concludes the paper.

2. RELATED WORK

Datasets for multiple modalities are being actively constructed. For spoken audio captions, Harwath et al. [11, 12] collected 400,000 English captions and 100,000 Hindi captions for a common image dataset. Chrupała et al. [14], Havard et al. [15], and Ilharco et al. [16] synthesized English and Japanese spoken captions from the text captions of the MS-COCO [17], STAIR [18], and Conceptual Captions datasets [19] and used them for training. In this paper, we investigate the effectiveness of a *trilingual* semantic embedding space using a dataset comprising spoken captions in three different languages.

Regarding model architectures, the dual encoder model has been used to train embeddings for images and audio captions. A pre-trained R-CNN [20], VGG16 [21], and Inception-ResNet-v2 [22] have been used as image encoders, and CNN-based DAVEnet [9] and ResNet-based ResDAVEnet [11] were proposed as speech encoders. The attention mechanism has been effectively utilized to model the temporal nature of spoken captions within a multi-layer recurrent highway network and a gated recurrent unit [14, 15].

Triplet loss has been introduced to learn the audio-visual embedding space [9, 14, 15]. Harwath et al. [11] also demonstrated that combining semi-hard negative training with the standard triplet loss worked well. More recently, Ilharco et al. [16] reported that a masked margin softmax loss has better characteristics than the standard triplet loss.

Other related works include studies of the association of

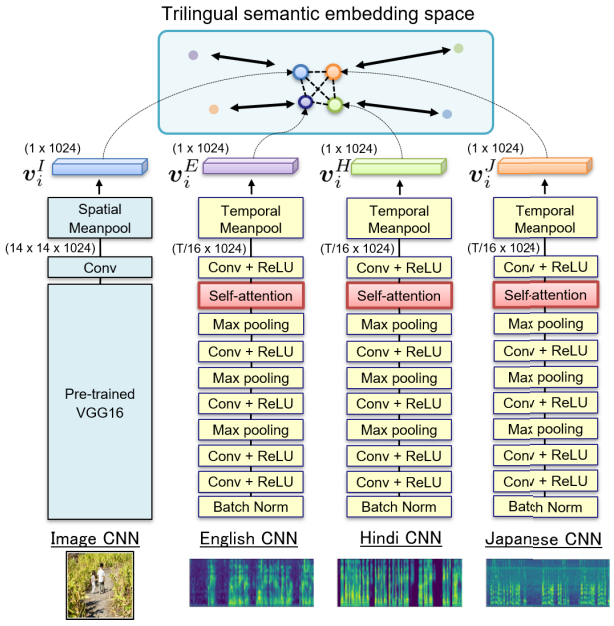


Fig. 2. Architecture of our neural networks

hand-written digits and spoken numbers [23, 24], visually grounded keyword spotting [25, 26], generation of audio descriptions of images [27, 28], an application to speech recognition [29], and audio-visual representation learning [30, 31].

3. JAPANESE SPOKEN CAPTIONS DATASET

We collected a new corpus of Japanese spoken captions for the Places205 dataset [32]. We used a subset of 100,000 images from the Places data that have both English [11] and Hindi [12] captions. To collect the Japanese corpus, we recorded a spontaneous, free-form spoken caption for each image using a crowdsourcing service in Japan.

To collect high-quality audio captions, all the recorded signals were checked by human listeners. If a signal contained too much noise or distortions, or was too short, it was excluded from the dataset, and the speaker was asked to re-record. Through this process, we collected 98,555 captions from 303 unique speakers (182 females and 121 males) in about three months. Table 1 shows the average duration and the number of words per caption for each language. The number of words was counted from the speech recognition results. As shown in the table, the listening verification process resulted in longer and richer spoken captions. We plan to make our dataset publicly available for academic purposes only.

4. MODELS

Our data takes the form of a collection of N *quadruples* $(I_i, A_i^E, A_i^H, A_i^J)$, where I_i is the i^{th} image, and A_i^E, A_i^H, and A_i^J are the speech audio signals of the English, Hindi,

Authorized licensed use limited to: University of Texas at Austin. Downloaded on September 03,2020 at 14:50:43 UTC from IEEE Xplore. Restrictions apply.

and Japanese captions describing the same image. Our model is based on the architecture proposed by Harwath et al. [12], in which a pair of convolutional neural networks (CNNs) are used to encode an image and a spoken caption to vectors in a shared embedding space. As shown in Fig. 2, we utilize four networks: one for the image, one for the English caption, one for the Hindi caption, and one for the Japanese caption. We hope this model can learn the visual-linguistic semantics from the co-occurrences across the different modalities.

We use an image encoder that takes all layers up through `conv5` from a pre-trained VGG16 network [21]. To map the VGG16 output into the embedding space, we apply 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 $\boldsymbol{v}^I$ of dimension d.

Our speech encoder is based on DAVEnet [11], but we added a self-attention layer [33] in the latter half of the network. The inputs are 40 log Mel filterbank energies per 25-ms frame of the caption at 10-ms shifts, and each speech encoder outputs an embedding vector of dimension d obtained by temporal mean pooling. We utilize truncation and zero-padding of each spectrogram to a fixed length of T frames ($T = 3072$, or approximately 30 seconds in our experiments), and then truncate the output features of each caption to remove the frames corresponding to zero-padding. Our data pre-processing follows the one in [12].

In the self-attention layer, the audio features from the third max pooling layer $\boldsymbol{x} \in \mathbb{R}^{C \times K}$ are first transformed into two feature spaces, where $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{W_f x}$, $\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{W_g x}$. Here, C is the number of channels, and K is the number of the audio feature locations. $\boldsymbol{W_f} \in \mathbb{R}^{\bar{C} \times C}$, $\boldsymbol{W_g} \in \mathbb{R}^{\bar{C} \times C}$ are the learned weight matrices, which are implemented as 1×1 pointwise convolutions. For memory efficiency, we choose $C = 512$ and $\bar{C} = C/8$ in all our experiments. The $\boldsymbol{x}_k$ is the k^{th} audio feature, and the attention map $\boldsymbol{\beta}$ is then given by

$$\beta_{k,l} = \frac{\exp(s_{k,l})}{\sum_{k=1}^{K} \exp(s_{k,l})}, \text{ where } s_{k,l} = \boldsymbol{f}(\boldsymbol{x}_k)^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{x}_l), \quad (1)$$

where $\beta_{k,l}$ indicates the extent to which this layer attends to the k^{th} location when synthesizing the l^{th} location. The output of the attention layer is $\boldsymbol{o} = \boldsymbol{x}\boldsymbol{\beta}^{\mathrm{T}} \in \mathbb{R}^{C \times K}$. The final output is given by $\boldsymbol{y} = \boldsymbol{x} + \gamma \boldsymbol{o}$, where γ is a trainable parameter. Our hope is that the attention layer enables the DAVEnet to efficiently model relationships between widely separated word-like segments in an audio caption.

Most previous studies employed triplet loss [34] to train the dual encoder models [12, 15, 31]. The triplet loss function is normally trained on a series of triplets $(\boldsymbol{a}, \boldsymbol{p}, \boldsymbol{i})$, where $\boldsymbol{a}$ is the anchor vector, $\boldsymbol{p}$ is a vector paired with $\boldsymbol{a}$, and $\boldsymbol{i}$ is an imposter vector. This function is designed to keep $\boldsymbol{a}$ closer to $\boldsymbol{p}$ than to $\boldsymbol{i}$, and it is widely used in many areas. On the other hand, inspired by [35, 36], Ilharco et al. [16] proposed masked margin softmax loss. In contrast to triplet loss, which chooses a negative sample randomly, masked margin softmax loss takes advantage of all negative pairs in the batch and thus improves the sample efficiency. More recently, Harwath et al. [11] found that semi-hard negative training worked much better when combined with the sampling-based triplet loss. Therefore, we compare these loss functions in our experiments. Given the quadruples, we apply these loss functions to neural embedding vectors $\boldsymbol{v}_i^I, \boldsymbol{v}_i^E, \boldsymbol{v}_i^H$, and $\boldsymbol{v}_i^J$ in 12 different ways, so that images and captions that belong to the quadruples are more similar in the embedding space than the mismatched image/caption or caption/caption pairs.

Table 2. Audio-visual retrieval recall scores for monolingual and trilingual models. "English caption" is abbreviated as E, "Hindi caption" as H, and "Japanese caption" as J.

Monolingual models

	Audio to Image			Image to Audio		
w/o self-attention layer	R@1	R@5	R@10	R@1	R@5	R@10
E	.105	.313	.437	.075	.261	.403
H	.090	.258	.365	.083	.252	.335
J	.190	.488	.626	.164	.431	.582
avg.	.128	.353	.476	.107	.315	.440
w/ self-attention layer						
E	**.145**	.355	.477	**.119**	.317	.452
H	.094	.274	.390	.089	**.249**	.373
J	**.204**	**.515**	**.651**	.214	**.489**	.635
avg.	**.148**	.381	.506	.141	**.352**	.487
w/ self-attention layer and masked margin softmax loss						
E	.141	**.364**	**.481**	.118	**.322**	**.457**
H	**.098**	**.280**	**.401**	**.099**	.247	**.380**
J	.200	.510	.642	**.218**	.475	**.637**
avg.	.146	**.385**	**.508**	**.145**	.348	**.490**

Trilingual models

w/o self-attention layer						
E	**.143**	.382	.518	.103	.342	.484
H	.103	.299	.429	**.110**	.295	.399
J	**.210**	.515	.667	.158	.435	.604
avg.	**.152**	.399	.538	.124	.357	.496
w/ self-attention layer and masked margin softmax loss						
E	.139	**.395**	**.529**	**.116**	**.358**	**.508**
H	**.112**	**.315**	**.445**	.108	**.313**	**.419**
J	.203	**.520**	**.667**	**.200**	**.468**	**.623**
avg.	.151	**.410**	**.547**	**.141**	**.380**	**.517**

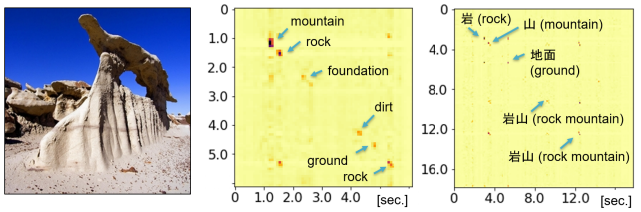
5. EXPERIMENTS

We evaluated the proposed model in terms of crossmodal and cross-lingual retrieval. We divided the dataset into a training set of 97,555 image/captions quadruples and a validation set

Authorized licensed use limited to: University of Texas at Austin. Downloaded on September 03,2020 at 14:50:43 UTC from IEEE Xplore. Restrictions apply.

Table 3. Cross-lingual retrieval for trilingual models

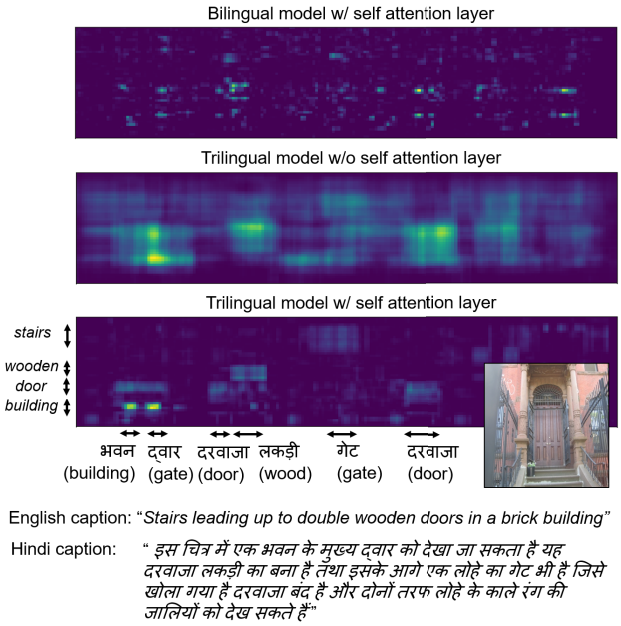
	w/o self-attention layer					
	Audio to Audio (→)			Audio to Audio (←)		
	R@1	R@5	R@10	R@1	R@5	R@10
E↔H	.055	.188	.278	.055	.176	.261
H↔J	.051	.193	.294	.063	.210	.286
J↔E	.072	.234	.328	.069	.221	.328
w/ self-attention layer and masked margin softmax loss						
E↔H	**.076**	**.225**	**.313**	**.076**	**.225**	**.313**
H↔J	**.085**	**.248**	**.334**	**.104**	**.246**	**.350**
J↔E	**.106**	**.317**	**.441**	**.105**	**.312**	**.437**



English caption: "*A **mountain rock foundation** with a lot of **dirt** on the **ground** and **rock***"

Japanese caption: "*真っ青な空の下に白い岩山がある。地面も真っ白で、波のような形をした大きな岩山が連なっていて、さらに大きな岩山が上のほうに乗り出していて複雑な形になっている*"

Fig. 3. Self-attention maps and speech recognition results



English caption: "*Stairs leading up to double wooden doors in a brick building*"

Hindi caption: "*इस चित्र में एक भवन के मुख्य द्वार को देखा जा सकता है यह दरवाजा लकड़ी का बना है तथा इसके आगे एक लोहे का गेट भी है जिसे खोला गया है दरवाजा बंद है और दोनों तरफ लोहे के काले रंग की जालियों को देख सकते हैं*"

Fig. 4. Similarity matrices between unpooled embeddings of English and Hindi captions

of 1,000 quadruples. We set the mini-batch size B and dimension d to 100 and 1024, respectively, and used a constant momentum of 0.9 and an initial learning rate of 0.001 which was decreased by a factor of 40 every ten epochs. Our model generally converged in less than 100 epochs.

The upper half of Table 2 lists the recall scores for the monolingual models trained from images and monolingual spoken captions. With the monolingual model, it was shown that the audio-visual retrieval accuracy for the Japanese-spoken captions exceeded that of other languages. This is probably because the Japanese captions were recorded longer and with more words, on average, than those of English and Hindi captions. It was also shown that the performance in these three languages was further improved by the self-attention layer inserted into the speech encoder. We found that the masked margin softmax loss function works slightly better than the combination of the sampling-based triplet loss and semi-hard negative training. We use the masked margin softmax loss function in all subsequent experiments.

The lower half of Table 2 shows the recall scores for the trilingual models. Comparing them with monolingual models, one can see that the retrieval performance is improved by learning three languages simultaneously. We believe that the shared embedding space is better learned by adding rich Japanese-spoken captions. Table 3 shows the cross-lingual retrieval results which indicate that incorporating the self-attention mechanism was also effective.

Figure 3 demonstrates examples of the self-attention map and speech recognition results. These results indicate that the characteristic words describing the image and the co-occurrences between the words are emphasized, and that the self-attention layers work well. We confirmed that attention in our model mainly focuses on nouns, as shown in [15]. Figure 4 compares similarity matrices between English and Hindi captions associated with the same image. Regions of high similarity correspond to translations of the underlying words. It can be seen that the use of the self-attention mechanism in the trilingual model results in less noise in the similarity matrix and clearer alignment. As future work, we are planning to learn an audio-visual picture dictionary, as shown in [13], from our trilingual models.

6. CONCLUSION

We proposed a trilingual semantic embedding model for visually grounded speech. In addition to the existing English and Hindi captions, we used a new corpus of Japanese-spoken captions. The experiments showed that the third language improves the performance in terms of crossmodal and cross-lingual information retrieval accuracy in most cases. Introducing the self-attention mechanism was also shown to be effective. Future work includes investigating ways to avoid combinatorial increases in the number of terms in the loss function and analyzing the model's performance when the quality or quantity of the spoken captions is not balanced among the languages. One of our challenges is to find audio-visual associations regarding not only still objects but also events, actions, or situations that would correspond to verbs, adverbs, or adjectives as well as nouns.

Authorized licensed use limited to: University of Texas at Austin. Downloaded on September 03,2020 at 14:50:43 UTC from IEEE Xplore. Restrictions apply.

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

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

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

[11] 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.

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

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

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

[15] 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 *Proc. ICASSP*, 2019.

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

[17] X. Chen, H. Fang, T-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick, "Microsoft COCO captions: Data collection and evaluation server," arXiv preprint arXiv:1504.00325, 2015.

[18] Y. Yoshikawa, Y. Shigeto, and A. Takeuchi, "STAIR Captions: Constructing a large-scale japanese image caption dataset," in *Proc. ACL*, 2017.

[19] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual Captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proc. ACL*, 2018.

[20] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. CVPR*, 2013.

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

[22] C. Szegedy, S. Ioffe, V. Vanhoucke, and A.A. Alemi, "Inception-v4,inception-resnet and the impact of residual connections on learning," in *Proc. AAAI*, 2017.

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

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

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

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

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

[28] 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.

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

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

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

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

[33] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. ICML*, 2019.

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

[35] M. Henderson, R. Al-Rfou, B. Strope, Y. Sung, L. Lukacs, R. Guo, S. Kumar, B. Miklos, and R. Kurzweil, "Efficient natural language response suggestion for smartreply," arXiv preprint arXiv:1705.00652, 2017.

[36] Y. Yang, G.H. Abrego, S. Yuan, M. Guo, Q. Shen, D. Cer, Y h. Sung, B. Strope, and R. Kurzweil, "Improving multilingual sentence embedding using bidirectional dual encoder with additive margin softmax," in *Proc. IJCAI*, 2019.

Authorized licensed use limited to: University of Texas at Austin. Downloaded on September 03,2020 at 14:50:43 UTC from IEEE Xplore. Restrictions apply.