Learning fine-grained semantics in spoken language using visual grounding

Wang, Xinsheng; Tian, Tian ; Zhu, Jihua; Scharenborg, Odette

DOI
10.1109/ISCAS51556.2021.9401232

Publication date
2021

Document Version
Accepted author manuscript

Published in
2021 IEEE International Symposium on Circuits and Systems (ISCAS)

Learning Fine-grained Semantics in Spoken Language Using Visual Grounding

Xinsheng Wang[†]
School of Software Engineering
Xi'an Jiaotong University
Xi'an, China
wangxinsheng@stu.xjtu.edu.cn

Tian Tian
Multimedia Computing Group
Delft University of Technology
Delft, The Netherlands
T.Tian-1@student.tudelft.nl

Jihua Zhu[*]
School of Software Engineering
Xi'an Jiaotong University
Xi'an, China
zhujh@xjtu.edu.cn

Odette Scharenborg
Multimedia Computing Group
Delft University of Technology
Delft, The Netherlands
o.e.scharenborg@tudelft.nl

***Abstract*—In the case of unwritten languages, acoustic models cannot be trained in the standard way, i.e., using speech and textual transcriptions. Recently, several methods have been proposed to learn speech representations using images, i.e., using visual grounding. Existing studies have focused on scene images. Here, we investigate whether fine-grained semantic information, reflecting the relationship between attributes and objects, can be learned from spoken language. To this end, a Fine-grained Semantic Embedding Network (FSEN) for learning semantic representations of spoken language grounded by fine-grained images is proposed. For training, we propose an efficient objective function, which includes a matching constraint, an adversarial objective, and a classification constraint. The learned speech representations are evaluated using two tasks, i.e., speech-image cross-modal retrieval and speech-to-image generation. On the retrieval task, FSEN outperforms other state-of-the-art methods on both a scene image dataset and two fine-grained datasets. The image generation task shows that the learned speech representations can be used to generate high-quality and semantic-consistent fine-grained images. Learning fine-grained semantics from spoken language via visual grounding is thus possible.**

***Index Terms*—Multimodal modelling, semantic retrieval, visual grounding, image generation, speech representation learning.**

I. Introduction

Standard automatic speech recognition (ASR) depends on large amounts of transcribed speech data for training the acoustic models. However, for around 98% of the world's languages not enough training material is available to train ASR systems [1]. Moreover, for unwritten languages, i.e., languages without a common writing system, the textual transcriptions are necessarily lacking.

Inspired by human infants' ability to learn to understand speech from exposure to spoken language and from watching objects and gestures (e.g., pointing), recently, several methods have been proposed to learn speech models using visual information for grounding [2]–[9]. In these works, various tasks were considered, such as cross-modal retrieval [2]–[4] and speech unit discovery [5], [6].

These methods train speech models using scene images that contain broad semantic classes, e.g., dog, beach, man. The combination of different objects plays a key role in existing tasks. For instance, in the scene "A boy is playing football", the "boy" and "football" are key factors for speech-image cross-modal retrieval and keyword discovery. Here, we go one step further and investigate whether fine-grained semantic information can be learned from speech, i.e., we explore whether a visually grounded speech model can learn the differences and relationships between attributes and objects rather than mainly focus on objects.

Although scene images can contain attribute-object pairs, such as "(a boy in a) black T-shirt", they typically do not contain those to the level that would allow us to investigate transfer to novel attribute-object instances similar to what humans are able to do. For instance, if we know "red apple" and "yellow banana", we would be able to understand "yellow apple". In contrast, the combinations of attribute-object pairs are the most important factor in recognizing fine-grained images. For instance, Fig. 1 shows two similar birds, which share the same attributes, i.e., black, yellow, and white, and the same objects, i.e., head and wings. The relationship between different attributes and objects is the key to differentiate between these two bird species. Speech representations should also have the ability to capture the relationships between attributes and objects.

Fig. 1. Examples from CUB dataset [10]. These two birds are two different species. The corresponding language descriptions for these two birds share most of the semantic information (in blue), and only differs for the head descriptions (in red).

To learn discriminative speech representations that can capture the relationships between attributes and objects, we propose a Fine-grained Semantic Embedding Network (FSEN). In FSEN, a pyramidal RNN followed by a self-attention module is used for the speech embedding branch to cope with the long sequence of audio signals and obtain more discriminative speech representations. A similar structure is also adopted in the image embedding branch to obtain discriminative representations of the images.

To evaluate the proposed model, two fine-grained datasets, i.e., CUB [10], which contains photos of birds, and Oxford-102 [11], containing photos of flowers, are adopted. During training, FSEN only sees a subset of the bird species, i.e., classes (in case of CUB) or of the flower species (in case of Oxford-102). During testing, the model sees test classes disjoint from the training classes. Thus, it can be regarded as a kind of zero-shot learning, and the cross-modal retrieval performance on the test set then reflects the model's ability to associate attributes and objects learned from known instances and generalize to new instances. A second challenging downstream task tests the learned speech representations on speech-to-image generation, in which the generated images give us a direct reflection of the semantic information carried by the learned speech representations.

[†]Xinsheng Wang was supported by the China Scholarship Council (CSC).
[*]Corresponding author.

II. Approach

A. Datasets

Two fine-grained image datasets, i.e., CUB [10] and Oxford-102 [11], are used both for the cross-modal retrieval and image generation tasks. CUB is a bird dataset that contains 11,788 bird images from 200 classes, and Oxford-102 is a flower dataset that contains 8,189 flower images from 102 classes. Following [12], we split them into class-disjoint training and test sets. Specifically, the CUB training set consists of 150 classes, and the test set of 50 classes non-overlapping with the training classes. Similarly, the Oxford-102 has 82 training classes and 20 test classes. Each image from both datasets has 10 descriptions. For our research, these descriptions are synthesized by a text-to-speech system[1] with tacotron2 [13] to obtain spoken descriptions according to the text descriptions collected by [12].

B. Model architecture

Dual encoders are popular in visually grounded speech learning tasks [3], [4], and other cross-modal learning tasks [14]–[16]. Here, we also take the dual-encoder structure for visually grounded speech learning. As shown in Fig. 2, this dual-encoder structure contains two encoders, i.e., a speech encoder and an image encoder. These two encoders embed the input images and speech into a common space, such that the image representations in this space can be used as supervision information to train the speech encoder, and vice versa.

The overall structure of this model is similar to that in [3]; however, we propose a new structure for both the speech encoder and the image encoder. Specifically, we discard the 1-D convolutional layer and replace the naive RNN with the pyramidal RNN [17] to deal with the long sequences of the speech signals. Moreover, in addition to the self-attention module in the speech encoder in [3], we propose an attention module along with an RNN to get important local features for the visual representations.

The speech encoder is modeled as a pyramidal Bidirectional Gated Recurrent Unit [18] (pBGRU) followed by a self-attention module. This pBGRU has three hidden layers, and the input of each layer is the concatenation of two consecutive state vectors from the former layer, which is similar to the pBLSTM in [17]. Thus, this three-layer pBGRU module reduces the time resolution 8 times. The hidden state vectors are 1024-d resulting from concatenating the bidirectional 512-d representations. The final speech representations $\boldsymbol{y}_i$ are calculated by the self-attention layer via a weighted sum over all hidden state vectors. The self-attention layer is similar to that in [3]. The speech for the input of the speech encoder consists of log Mel filter bank spectrograms, which are obtained using 40 Mel-spaced filter banks with 25 ms Hamming window and 10 ms shift.

The image encoder is implemented as a one-layer pBGRU with the same attention layer as that in the speech encoder. The input to the pBGRU is created by scanning the last convolutional layer in a raster-scan order, i.e., left-to-right, top-to-bottom. Specifically, we use the ResNet-101 [19] trained on ImageNet [20] as the pre-encoder. The last convolutional layer has 2048 channels, each has a size of 8×8, resulting in a 2048-d pseudo-temporal sequence with a sequence length of 64. The subsequent self-attention module is the same as that in the speech encoder, and is used to obtain the final representation $\boldsymbol{x}_i$ of an image.

C. Objective Function

As shown in Fig. 2, the loss function consists of three parts, i.e., a matching loss, a discriminative loss, and an adversarial loss.

[1]https://github.com/NVIDIA/tacotron2

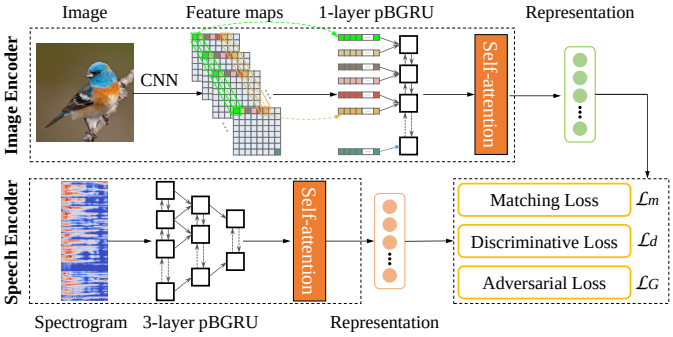


Fig. 2. Framework of the proposed method.

The matching loss is used to make the speech and its corresponding image closer in the embedding space. Similar to the DAMSMS loss in [21], given a batch of speech-image representation pairs $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_i^n$, with batch size of n, the matching loss function is defined as

$$\mathcal{L}_m = -\sum_{i=1}^{n} \log P\left(\boldsymbol{x}_i|\boldsymbol{y}_i\right) - \sum_{i=1}^{n} \log P\left(\boldsymbol{y}_i|\boldsymbol{x}_i\right), \tag{1}$$

where $P\left(\boldsymbol{x}_i|\boldsymbol{y}_i\right)$ is the possibility of $\boldsymbol{y}_i$ matching with $\boldsymbol{x}_i$:

$$P\left(\boldsymbol{x}_i|\boldsymbol{y}_i\right) = \frac{\exp\left(\beta S\left(\boldsymbol{y}_i, \boldsymbol{x}_i\right)\right)}{\sum_{j=1}^{n} M_{i,j} \exp\left(\beta S\left(\boldsymbol{y}_i, V_j\right)\right)}, \tag{2}$$

where β is a smoothing factor, and set to 10 following [21]. $S\left(\boldsymbol{y}_i, \boldsymbol{x}_i\right)$ is the cosine similarity of $\boldsymbol{y}_i$ and $\boldsymbol{x}_i$. We assume that only $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ are matched pairs in a batch, and $M_{i,j} \in \mathbb{R}^{n \times n}$ is used to deactivate the effect of other pairs from the same class

$$M_{ij} = \begin{cases} 0, & \text{if } \boldsymbol{y}_i \text{ matches } \boldsymbol{x}_j \ \& \ i \neq j, \\ 1, & \text{otherwise .} \end{cases} \tag{3}$$

In the same way, we compute $P\left(\boldsymbol{y}_i|\boldsymbol{x}_i\right)$, which is the possibility of $\boldsymbol{x}_i$ matching with $\boldsymbol{y}_i$.

The discriminative loss is used to learn class distinctive representations, i.e., representations that can distinguish different classes (e.g., bird species). Here, we take the classification objective function as the discriminative loss. Specifically, with $f(.)$ representing a perception layer that transfers representations of images and speech from the common embedding space to the label space, in which the vector of an image or a stretch of speech represents the probability distribution for each class label, the loss function is defined as

$$\mathcal{L}_d = -\frac{1}{2} \sum_{i=1}^{n} \Big(\log \hat{P}\left(c_i|f(\boldsymbol{y}_i)\right) + \log \hat{P}\left(c_i|f(\boldsymbol{x}_i)\right) \Big), \tag{4}$$

where $\hat{P}(c_i|f(\boldsymbol{y}_i))$ represents the softmax probability of $f(\boldsymbol{y}_i)$ belonging to its corresponding class c_i.

The adversarial loss is used to reduce the modality gap, and has shown to yield clear improvements in the text-image cross modal retrieval task [22] by reducing the modality gap. Generative Adversarial Networks (GANs) [23]–[25] are trained in a two-player mini-max game between a discriminator and a generator. Here, a modality classifier D works as the discriminator, which is trained with the loss function:

$$\mathcal{L}_D = -\sum_{i=1}^{n} \Big(\mathop{\mathbb{E}}_{\boldsymbol{y}_i \sim \boldsymbol{y}} \left[\log D\left(\boldsymbol{y}_i\right)\right] + \mathop{\mathbb{E}}_{\boldsymbol{x}_i \sim \boldsymbol{x}} \left[\log\left(1 - D\left(\boldsymbol{x}_i\right)\right)\right] \Big). \tag{5}$$

The encoders for speech and images work as the generator. They are trained to fool the discriminator D, such that the classifier D cannot

distinguish the modality differences between the speech and image representations. The loss function is given by

$$\mathcal{L}_G = -\sum_{i=1}^{n}\left(\mathop{\mathbb{E}}_{\boldsymbol{x}_i\sim\boldsymbol{x}}\left[\log D\left(\boldsymbol{x}_i\right)\right] + \mathop{\mathbb{E}}_{\boldsymbol{y}_i\sim\boldsymbol{y}}\left[\log\left(1-D\left(\boldsymbol{y}_i\right)\right)\right]\right). \quad (6)$$

The total loss for speech representation learning is given by

$$\mathcal{L} = \mathcal{L}_m + \mathcal{L}_d + \mathcal{L}_G. \quad (7)$$

This loss function is used to update the parameters in the dual encoder. Note that $\mathcal{L}_D$ is only used to train the modality discriminator. These two training processes are performed in an alternate way. Adam [26] is adopted and the initial learning rate is 0.0001 with a decay rate of half every 50 epochs.

III. Experiments

A. Cross-modal retrieval

The task of cross-modal retrieval is to retrieve the correct image given a spoken description, and vice versa. In order to properly evaluate our model, we first compare our model with the state-of-the-art method [3] on speech-image cross-modal retrieval on the Flickr8k [27] dataset, which is a commonly used scene image dataset in cross-modal retrieval. The spoken descriptions of Flick8k are collected via Amazon Mechanical Turk and are natural speech [28] . We followed the training/test split in [3]. Note, Flickr8k does not have class information, so each image is treated as a single class to perform the discriminative loss.

Since this is the first work on fine-grained image based cross-modal retrieval using spoken input, we compare our model with two state-of-the-art cross-modal speech-based retrieval methods for scene images [2], [3], and three state-of-the-art text-image cross-modal retrieval methods [16], [22], [29] for datasets with class information. To implement the methods originally designed for text-image cross-modal retrieval tasks, the text encoder was replaced by our speech encoder. For a fair comparison, the input speech for all methods was represented by log Mel filter bank spectrograms as in our method. In order to compare architectures and objective function rather than training schemes, the training tricks used in [3], i.e., cyclic learning rate and snapshot ensembling, which have been shown to be beneficial were replaced with a learning rate with an initial value of 0.0001 with a decay rate of half every 50 epochs.

The evaluation metrics for the cross-modal retrieval task are R(ank)@K and mAP@50. R@K indicates the percentage of the queries for which at least one ground-truth, i.e., images or speech from the correct class, are retrieved among the top-K results. mAP@50 is the mean Average Precision (mAP), but only the top 50 retrieved results are taken into consideration.

B. Speech-to-image generation

We test the semantic information carried by the learned speech representations in a speech-to-image generation task. Since text-to-image generation was made possible by Reed *et al*. [30], many efforts have been carried out [21], [31], [32] to improve the performance of text-to-image generation. Here, we use the structure of StackGAN-v2 [31], which showed outstanding performance on the text-to-image generation task, to perform the spoken description-to-image generation task. In our experiment, we replace the original text representations with our learned spoken language representations, and compare the performances of StackGAN-v2 with the original text representations [31] and our spoken language representations.

The evaluation metrics for the image generation task follow those in [31]. Specifically, we use inception score (IS) [33] and Fréchet inception distance (FID) [34] to evaluate the diversity and quality of generated images, respectively. Higher IS means better diversity and lower FID means a smaller distance between the real and generated image distributions, indicating better performance on image generation. Additionally, in order to evaluate the semantic consistency between the generated images and the ground-truth images, we conduct a retrieval task using ground-truth images to retrieve synthesized images. The retrieval performance is evaluated with mean Average Precision (mAP). Larger mAP means better semantic consistency between the generated images and corresponding speech descriptions.

TABLE I
Cross-modal retrieval performance on CUB and Oxford datasets. The best results are shown in bold.

Dataset	CUB (Bird)				Oxford102 (Flower)			
Method	Speech-to-image		Image-to-speech		Speech-to-image		Image-to-speech	
	R@1	mAP@50	R@1	mAP@50	R@1	mAP@50	R@1	mAP@50
[2]	6.9	5.6	13.6	10.1	17.1	17.2	22.8	19.6
[3]	9.6	8.4	12.3	10.3	10.6	10.1	17.9	14.6
[16]	13.9	11.2	23.3	19.1	25.7	19.9	30.0	25.7
[29]	18.8	16.0	28.8	24.9	34.9	30.7	45.2	40.5
[22]	21.0	17.7	31.9	26.1	36.1	31.8	46.1	41.2
FSEN	**33.5**	**27.9**	**50.2**	**41.0**	**48.1**	**42.5**	**63.2**	**53.5**

IV. Results

A. Cross-modal retrieval

The comparison of our model's results with those of [3] on Flickr8k [27] showed that our method outperforms [3] on both speech-to-image and image-to-speech cross-modal retrieval tasks. Specifically, on the speech-to-image retrieval task, we achieve 10.1%, 28.8%, and 40.7% accuracy for R@1, R@5, and R@10 respectively, outperforming the corresponding accuracy 8.4%, 25.7%, and 37.6% achieved by [3]. On the image-to-speech retrieval task, our model's performances are 13.7%, 36.1%, and 49.3%, while performances of [3] are 12.2%, 31.9%, and 45.2%. These results show the state-of-the-art performance of our method on learning semantic representations of spoken language grounded by visual information.

The Fine-grained image based cross-modal retrieval performance is shown in Table I. As shown, on both CUB and Oxford datasets, our method outperforms all the compared methods with a substantial margin on all evaluation metrics. Specifically, on the CUB dataset, our method achieves 12.5% and 18.3% improvements on R@1 over the second-best method for the speech-to-image and image-to-speech retrieval tasks, respectively. These results indicate that our method is effective in matching previously unseen speech descriptions and images by associating known attributes and objects.

B. Speech-to-image generation

Subjective results are shown in Fig. 3. As can be seen, conditioned on our learned speech representations, the StackGAN-v2 is able to synthesize high-quality photo-realistic images. The synthesized images show good semantic consistency with the corresponding spoken descriptions, which indicates that FSEN has the ability to learn speech representations with fine-grained semantics.

Objective results are shown in Table II. As can be seen, conditioned on our learned speech representations, StackGAN-v2 shows better performance on the image generation task than when it is conditioned on the original text embedding, indicating the good performance of our method on learning semantic representations for spoken descriptions.

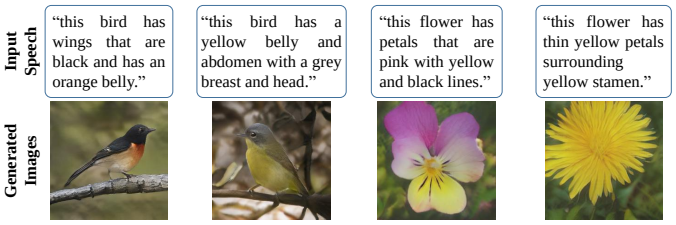


Fig. 3. Examples of generated images based on spoken descriptions from CUB (birds) and Oxford102 (flowers).

TABLE II
PERFORMANCE ON THE IMAGE GENERATION TASK. THE BEST RESULTS ARE SHOWN IN BOLD.

	CUB (Bird)			Oxford-102 (Flower)		
Input	mAP	FID	IS	mAP	FID	IS
text [31]	7.01	20.94	4.02±0.03	9.88	**50.38**	3.35±0.07
speech	**7.83**	**18.63**	**4.13±0.03**	**13.02**	54.53	**3.75±0.09**

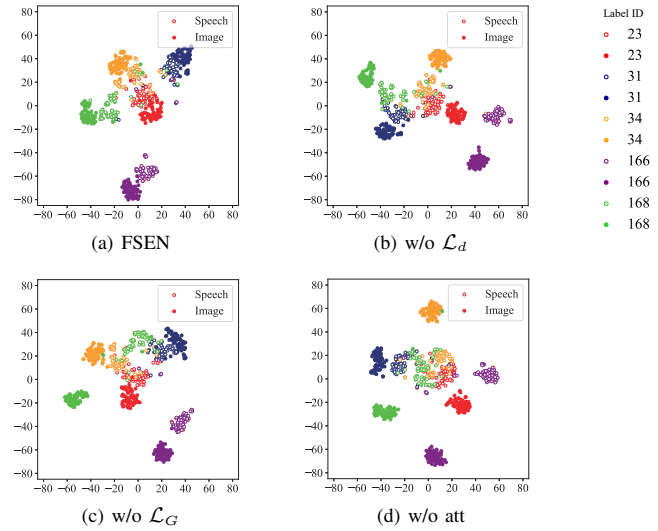
C. Component analysis

To investigate the effectiveness of the different components in FSEN, we evaluated various variants of FSEN by removing each component respectively. Table III shows the cross-modal retrieval performance of those variant FSENs on the CUB and Oxford-102 datasets. The removal of the attention modules was achieved by replacing the attention modules with averaging. As can be seen, removing any component of FSEN except for $\mathcal{L}_d$ brings a significant decline in the performance of cross-modal retrieval. Although removing $\mathcal{L}_d$ leads to a slight improvement on R@1 for speech-to-image retrieval, it still gives an overall decrease. Moreover, discarding any attention module in image encoder or speech encoder brings obvious negative effects. These results show that each part of the model contributes to the success of the model.

Ideally, a good speech-image dual encoder should cluster speech representations and image representations from the same class together and separate them from representations belonging to other classes. To show the effect of each component on learning the representations in an intuitive way, we visualized the speech and image representation distributions produced by four FSEN variants using t-SNE [35], see Fig. 4. For ease of inspection, the presented data are from 5 randomly selected classes from the CUB test dataset, and only the first speech description of each image is taken. Representations of speech and images that are from the same class are plotted using the same color. As shown, the removal of $\mathcal{L}_d$ leads to a trend that the speech representations mix with one another, demonstrating the effectiveness of $\mathcal{L}_d$ on learning speech representations that are discriminative among different classes. Training the model without $\mathcal{L}_G$ leads to a separation of the speech representations from their corresponding image representations, which indicates that $\mathcal{L}_G$ plays an important role in reducing the modality gap between speech and image. The attention modules are the backbone in the dual encoder: when removed, both separation of the representations from the two modalities and mixing of the speech representations from different classes occur, even when $\mathcal{L}_d$ and $\mathcal{L}_G$ are used.

TABLE III
COMPONENT ANALYSIS OF FSEN; W/O MEANS WITHOUT; ATT MEANS ALL ATTENTION MODULES IN THE FSEN; ATT-I MEANS THE ATTENTION MODULE IN THE IMAGE BRANCH AND ATT-S MEANS THE ATTENTION MODULE IN THE SPEECH BRANCH. THE BEST RESULTS ARE SHOWN IN BOLD.

Dataset	CUB (Bird)				Oxford102 (Flower)			
Method	Speech-to-image		Image-to-speech		Speech-to-image		Image-to-speech	
	R@1	mAP@50	R@1	mAP@50	R@1	mAP@50	R@1	mAP@50
w/o $\mathcal{L}_d$	**33.9**	27.8	49.5	40.4	**48.7**	42.0	57.3	51.0
w/o $\mathcal{L}_G$	33.1	27.4	48.2	39.8	47.2	40.9	60.2	50.5
w/o att	29.1	24.4	41.4	34.6	44.5	38.4	55.0	47.7
w/o att-I	32.6	27.4	45.1	38.6	46.8	41.1	58.8	51.1
w/o att-S	30.6	25.5	45.1	36.1	44.1	39.3	57.9	50.7
FSEN	33.5	**27.9**	**50.2**	**41.0**	48.1	**42.5**	**63.2**	**53.5**



(a) FSEN (b) w/o $\mathcal{L}_d$ (c) w/o $\mathcal{L}_G$ (d) w/o att

Fig. 4. Distribution visualization of speech and image representations learned by variants of FSEN. w/o means without. att means all attention modules.

V. DISCUSSION AND CONCLUSION

In this paper, a fine-grained semantic embedding network FSEN was proposed to learn fine-grained semantic representations of spoken descriptions. Evaluated on two fine-grained image datasets, the proposed FSEN shows good performance both on a speech-image cross-modal retrieval task and a speech-to-image generation task, indicating that it is feasible to learn fine-grained semantic representations of spoken languages which capture the relationship between attributes and objects, via visual grounding. Moreover, the component analysis demonstrated the effectiveness of each component of the proposed method.

In the current work, the model was evaluated on synthesized speech (but note its state-of-the-art performance on the Flickr8k dataset which contains natural speech). In the future, we will investigate the model's ability to learn fine-grained semantic speech representations from natural speech.

REFERENCES

[1] G. Adda, S. Stüker, M. Adda-Decker, O. Ambouroue, L. Besacier, D. Blachon, H. Bonneau-Maynard, P. Godard, F. Hamlaoui, D. Idiatov *et al.*, "Breaking the unwritten language barrier: The bulb project," *Procedia Computer Science*, vol. 81, pp. 8–14, 2016.

[2] D. Harwath, A. Torralba, and J. Glass, "Unsupervised learning of spoken language with visual context," in *NeurIPS*, 2016, pp. 1858–1866.

[3] D. Merkx, S. Frank, and M. Ernestus, "Language learning using speech to image retrieval," *Proceedings of Interspeech 2019.*, pp. 1841–1845, 2019.

[4] G. Ilharco, Y. Zhang, and J. Baldridge, "Large-scale representation learning from visually grounded untranscribed speech," in *CoNLL*, 2019, pp. 55–65.

[5] H. Kamper, S. Settle, G. Shakhnarovich, and K. Livescu, "Visually grounded learning of keyword prediction from untranscribed speech," *Proc. Interspeech 2017*, pp. 3677–3681, 2017.

[6] H. Kamper, G. Shakhnarovich, and K. Livescu, "Semantic speech retrieval with a visually grounded model of untranscribed speech." *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 27, no. 1, pp. 89–98, 2019.

[7] 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 *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 649–665.

[8] 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*, pp. 1–22, 2019.

[9] D. Harwath, W.-N. Hsu, and J. Glass, "Learning hierarchical discrete linguistic units from visually-grounded speech," *arXiv preprint arXiv:1911.09602*, 2019.

[10] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

[11] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*. IEEE, 2008, pp. 722–729.

[12] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 49–58.

[13] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *ICASSP*. IEEE, 2018, pp. 4779–4783.

[14] A. Senocak, T.-H. Oh, J. Kim, M.-H. Yang, and I. So Kweon, "Learning to localize sound source in visual scenes," in *CVPR*, 2018, pp. 4358–4366.

[15] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 154–162.

[16] L. Zhen, P. Hu, X. Wang, and D. Peng, "Deep supervised cross-modal retrieval," in *CVPR*, 2019, pp. 10 394–10 403.

[17] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4960–4964.

[18] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[20] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[21] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "Attngan: Fine-grained text to image generation with attentional generative adversarial networks," in *CVPR*, 2018, pp. 1316–1324.

[22] N. Sarafianos, X. Xu, and I. A. Kakadiaris, "Adversarial representation learning for text-to-image matching," in *CVPR*, 2019, pp. 5814–5824.

[23] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NeurIPS*, 2014, pp. 2672–2680.

[24] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," *arXiv preprint arXiv:1701.07875*, 2017.

[25] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *CVPR*, 2017, pp. 2223–2232.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[27] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, 2013.

[28] D. Harwath and J. Glass, "Deep multimodal semantic embeddings for speech and images," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 237–244.

[29] Y. Zhang and H. Lu, "Deep cross-modal projection learning for image-text matching," in *ECCV*, 2018, pp. 686–701.

[30] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *33rd International Conference on Machine Learning*, 2016, pp. 1060–1069.

[31] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan++: Realistic image synthesis with stacked generative adversarial networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1947–1962, 2018.

[32] T. Qiao, J. Zhang, D. Xu, and D. Tao, "Mirrorgan: Learning text-to-image generation by redescription," in *CVPR*, 2019, pp. 1505–1514.

[33] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *NeurIPS*, 2016.

[34] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *NeurIPS*, 2017.

[35] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, 2008.