Audio Caption in a Car Setting with a Sentence-Level Loss

Xuenan Xu, Heinrich Dinkel, Mengyue Wu, Kai Yu

MoE Key Lab of Artificial Intelligence
SpeechLab, Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China
`{wsntxxn, richman, mengyuewu, kai.yu}@sjtu.edu.cn`

Abstract

Captioning has attracted much attention in image and video understanding while a small amount of work examines audio captioning. This paper contributes a Mandarin-annotated dataset for audio captioning within a car scene. A sentence-level loss is proposed to be used in tandem with a GRU encoder-decoder model to generate captions with higher semantic similarity to human annotations. We evaluate the model on the newly-proposed *Car* dataset, a previously published Mandarin *Hospital* dataset and the *Joint* dataset, indicating its generalization capability across different scenes. An improvement in all metrics can be observed, including classical natural language generation (NLG) metrics, sentence richness and human evaluation ratings. However, though detailed audio captions can now be automatically generated, human annotations still outperform model captions on many aspects.

Index Terms: Audio Caption, Audio Caption Datasets, Sentence-level Loss, Natural Language Generation

1. Introduction

Automatic captioning is a challenging task that involves joint learning of different modalities. For example, image captioning requires extracting features from an image and combining them with a language model to generate reasonable sentences to describe the image. Similarly, video captioning learns features from a temporal sequence of images as well as audio to generate captions. However, since audio captioning is a relatively new field, it does not attract much attention like image-and video captioning.

One well-known task within audio processing, which is commonly associated with audio captioning, is Automatic Speech Recognition (ASR). There are two main characteristics of audio captioning compared with ASR: 1) audio captioning focuses on all sound events in an audio while ASR only focuses on speech (speech does not necessarily appear in the input to an audio captioning model) 2) audio captioning is an automatic summarization of the audio sound events while ASR directly outputs transcriptions of human speech in the audio. A comparison of the two tasks' goals is shown in Figure 1.

Though the success of an audio captioning task in the recent DCASE2020 challenge has prompted a plethora of novel approaches and papers [1, 2, 3, 4], limited attention is paid to audio captioning within Chinese language processing. A Mandarin-annotated 10 hour audio dataset within a hospital scene in conjunction with a baseline encoder-decoder model to generate natural language captions has recently been published [5].
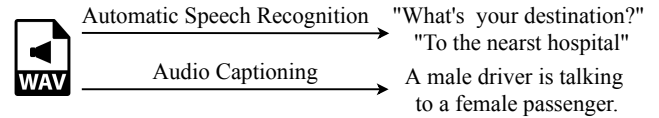
Figure 1: *Illustration of the difference between goals of ASR and audio captioning.*

Although the model performance evaluated by BLEU score is particularly high, human evaluation tells a different story. Most machine-generated Mandarin captions are monotonous and repetitive, while by contrast human annotations are much more specific in content and vivid in expression. Therefore, a captioning model should endeavor to generate various sentences that not only describe detailed audio content but also contain richer vocabulary and diverse sentence structures. For example, for a sound of a car crash in an audio clip, a well-performing model is expected to generate a caption like "The car went into a crash with others" or "A traffic accident happened", instead of repetitive "There is a sound of a car crash" or even "There are car sounds".

To achieve the goal of generating specific captions with various expressions, we first publish a dataset on car scene with five annotations per audio. Followed by that, we address the variety lacking problem by incorporating an additional sentence-level loss during training. Similar sequence-level loss has been proved effective in previous work [6, 7]. The sentence-level loss is based on context-aware sentence embeddings of diverse, vivid human annotations. Since there is a supervision signal from the overall sentence-level similarity with human annotations, the model is expected to generate more diverse sentences with similar meaning. In addition to classical natural language generation (NLG) metrics, we also evaluate our model by output richness, represented by the ratio between the number of unique sentences and total predicted sentences. Human evaluation is further performed to subjectively rate the output quality. Finally, with the newly proposed dataset on car scene, we evaluate our model's generalization capabilities.

2. Related Work

Audio Captioning Datasets To date, a few datasets including Audiocaps [8] and Clotho [9] for audio captioning have been published. Most of the current audio captioning datasets are in English. The only existing Mandarin-annotated dataset is the previously mentioned *Hospital Scene* dataset [5], a detailed comparison with the proposed *Car Scene* dataset will be provided in Section 3.

Captioning Model Image and video captioning have witnessed promising improvements recently. The development of
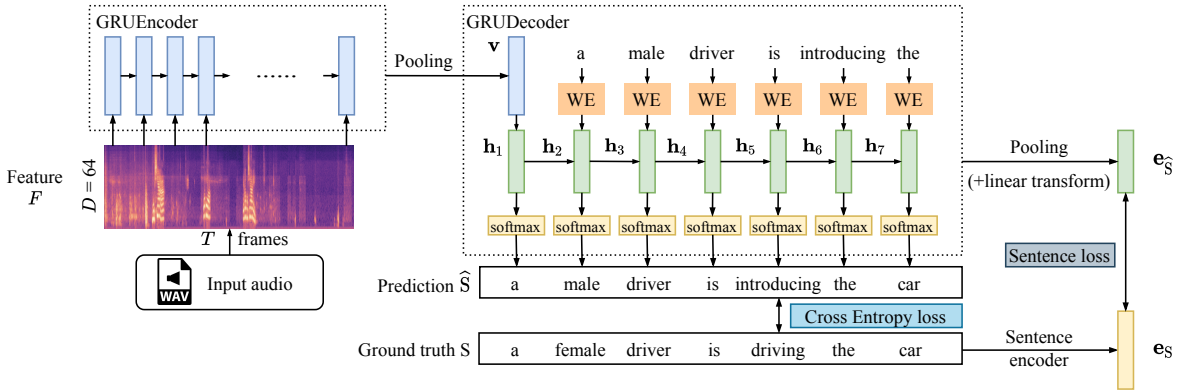
Figure 2: *Our proposed encoder-decoder model with sentence-level loss. Both the encoder and the decoder are single-layer GRUs. The encoder outputs a fixed-sized audio embedding $\mathbf{v}$ from the $T \times D$ input feature $\mathbf{F}$. Then the decoder predicts the sentence with $\mathbf{v}$ as the input to the first timestep. In addition to the standard cross entropy loss between the prediction $\hat{S}$ and the ground truth S, we propose a sentence-level loss, which is the dissimilarity between the prediction embedding $\mathbf{e}_{\hat{S}}$ and the ground truth embedding $\mathbf{e}_S$.*

sequence-to-sequence models enables well-performing video captioning models by simply using temporal image information [10]. Later, the attention mechanism is utilized to fuse audio with video information and assign different importance to time frames [11, 12, 13]. Shen *et al.* [14] generates multiple captions in different detail levels and temporal attention.

Sentence Embedding Early works like GloVe [15] and Word2Vec [16] in natural language processing (NLP) focus on context-free embedding of words. Recently, models like Cove [17], ELMo [18] and GPT [19] make use of the self-attention mechanism and transformers to build context-sensitive word representations. An unsupervised, C-BOW-like method to embed sentence to fixed-length vector [20] is later proposed. In this paper, our work is based on the state-of-art sentence embedding technique from BERT [21]. It contains large bidirectional transformers trained on a huge corpus, thus embeddings extracted from the pretrained BERT model perform well in many tasks.

Evaluation Metrics In previous captioning work, evaluation metrics are mainly borrowed from NLG tasks like machine translation and summarization: BLEU@1-4, METEOR, CIDEr and ROUGE-L scores. Most of these metrics are based on N-Gram overlaps between model predictions and human references [22, 23, 11, 13]. In addition, more effective metrics have also been explored. [24] and [25] treat image captioning as a sentence ranking task and use recall@k and median@r as their metric. Chuang *et al.* [13] embeds sentences to fixed length vectors, based on which, a cosine similarity between model predictions and human annotations is involved as a semantic evaluation. Our sentence-level loss function is thus inspired to focus on semantic similarity rather than individual wording choices.

3. An audio caption dataset in a car scene

This work publishes a 10 hours' Mandarin-annotated dataset on car scene that enables audio captioning. English translations using Baidu translator are also provided for broader accessibility. The proposed Car dataset contains 3602 car-scene related audio clips, each lasting for 10s. Each audio clip is annotated by five native Mandarin speakers with a concise labeling method: only natural sentence annotations are included while other metadata are generated from the annotations, e.g., sound events, subjects, etc.

Table 1: *A comparison of existing audio captioning datasets.*

Dataset	Language	Scene	# Audios	# Captions
AudioCaps	English	General	39,597	45,513
Clotho	English	General	4,981	24,905
Hospital	Mandarin	Specific	3,709	11,121
Car	Mandarin	Specific	3,602	18,010

This dataset exhibits a handful of discrepancies from the previously published datasets (see Table 1): 1) We provide scene-specific datasets for precise caption generation, in comparison with general purpose datasets like AudioCaps [8] and Clotho [9]; 2) The proposed Car dataset includes large quantities of real-life recordings which are suitable for real applications while the *Hospital* dataset [5] consists of more video clips from TV shows due to limited surveillance access in hospital.

Table 2: *Most Frequent Sound Events.*

Rank	Sound Event	# of events
1	Engine Sound	1442
2	Noise	872
3	Clicking Sound	812
4	Music	798
5	Speech	563

Table 2 shows the top 5 sound events of the proposed Car dataset, indicating that the sound events are quite scene-specific. The Car dataset is split into a development set and an evaluation set, which encompasses 3241 and 361 audio clips respectively. High sentence diversity is observed in both sets: only 6.7% annotations in the development set and 1.9% in the evaluation set are repeated. From the distribution of the top 5 tokens in Table 3 it can be seen that the development-evaluation split exhibits a similar token distribution.

4. Model Description

Since the previously utilized GRU encoder-decoder model [5] can generate audio relevant and grammatically correct sentences, we continue to incorporate a similar architecture with certain modifications for further performance enhancement.

Table 3: *Token Distribution in the proposed Car dataset.*

Rank	Token	Dev %	Eval %
1	is/are 在	6.01	6.01
2	driving 行驶	5.37	5.55
3	automobile 汽车	5.01	5.11
4	's 的	4.01	4.58
5	driver 司机	3.35	3.45
mean # of tokens		14.21	14.03

Our architecture consists of an audio embedder (encoder) and a text generator (decoder).

Encoder For each audio clip, our GRU encoder reads a log mel spectrogram (LMS) feature F and encodes it into a fixed-length feature vector $\mathbf{v}$.

Decoder The decoder takes the audio embedding $\mathbf{v}$ as its input to generate natural language captions. However, our decoder network works differently during training and evaluation. During training, when annotated captions are available, teacher forcing is used to accelerate the training process. $\mathbf{v}$ is thus concatenated with the ground truth annotation embeddings as the input to the decoder. Without annotation access during evaluation, $\mathbf{v}$ is directly fed to the decoder. For every timestep, the decoder generates a single token until the "<EOS>" (end of sentence) token is generated (see Figure 2).

In this paper, we utilize two loss functions during training: 1) standard Cross Entropy (CE) loss 2) the newly proposed sentence-level loss. Using a word and sentence loss combination, our model is expected to not only focus on wording selection but also sentence-level semantic similarity, which eventually leads to captions with human-like content while being diversified in sentence structure.

CE loss Standard cross entropy is used as the word-level loss Equation (1), which is defined as the negative log likelihood of the expected word S_t given the input audio feature F and the model parameters θ at time t.

$$\ell_{\text{CE}}(\theta; S, F) = -\sum_{t=1}^{T} \log p(S_t | \theta, F) \quad (1)$$

Sentence-level loss In addition to the standard CE loss at word-level, we propose a novel sentence-level loss to capture semantic similarity better. Since the decoder outputs a hidden state ($\mathbf{h}_t$) at each timestep t, we first pool the hidden states of all timesteps to get a single representation of the prediction. As Equation (2) shows, we use mean pooling on all $\mathbf{h}_t$ to obtain the representation $\mathbf{e}_{\hat{S}}$.

$$\mathbf{e}_{\hat{S}}(\theta, F) = \frac{1}{T} \sum_{t=1}^{T} \mathbf{h}_t(\theta, F) \quad (2)$$

In order to minimize the embedding difference between $\mathbf{e}_{\hat{S}}$ and annotated sentences ($\mathbf{e}_S$), we develop a sentence loss function opposed to cosine similarity (see Equation (3), where ϵ is a small number ensuring numerical stability). In this way, a small sentence loss indicates a high semantic similarity. In cases where $|\mathbf{e}_{\hat{S}}|$ differs from $|\mathbf{e}_S|$, a linear transformation layer is added after the mean pooling operation to ensure $\mathbf{e}_{\hat{S}}$ and $\mathbf{e}_S$ are of equal dimension.

$$\ell_{\text{sentence}}(\theta; \mathbf{e}_S, F) = 1 - \frac{\mathbf{e}_S \cdot \mathbf{e}_{\hat{S}}(\theta, F)}{\max(\|\mathbf{e}_S\|_2 \cdot \|\mathbf{e}_{\hat{S}}(\theta, F)\|_2, \epsilon)} \quad (3)$$

Accordingly, the training objective (Equation (4)) minimizes the weighted sum of the word (Equation (1)) and sentence (Equation (3)) loss, where α is a fixed hyperparameter.

$$\ell_{\text{combined}}(\theta; S, \mathbf{e}_S, F) = \ell_{\text{CE}}(\theta; S, F) + \alpha \cdot \ell_{\text{sentence}}(\theta; \mathbf{e}_S, F) \quad (4)$$

5. Experiments

5.1. Datasets

We first validate our proposed *Car* dataset for its effectiveness in audio captioning. To investigate the generalization capabilities of our model and the proposed sentence-level loss, in particular under cross-scene circumstances, we further experiment on another two datasets. One is the *Hospital* dataset, including 3709 audio clips with three human annotations for each audio clip; the other is the creation of a *Joint* dataset that merges the *Car* and *Hospital* datasets. It should be noted that the *Joint* dataset is domain balanced since the number of audio clips within the two datasets are similar (Car: 3602; Hospital: 3709).

5.2. Data preprocessing

Standard 64 dimensional LMS features from a 40 ms window are extracted every 20 ms. During training we apply global standardization (mean and variance) on each feature. Since the annotations are in Mandarin Chinese, a language that does not separate words by space in sentences, the annotations need to be tokenized. Here, Stanford core NLP tools [26] are used for parsing. We also use the public simplified and traditional Chinese BERT model[1] to obtain the fixed-length annotation embedding $\mathbf{e}_S$ for sentence-level loss training.

5.3. Training Details

Both the encoder and the decoder are composed of a single-layer GRU with a hidden size of 512. The dimension of $\mathbf{v}$ is 256. BERT encodes S into a 768 dimensional $\mathbf{e}_S$. The development set is further split into a training subset and a validation subset with a ratio of $9 : 1$. Training is done using the Adam optimization algorithm [27] with an initial learning rate $4e{-}4$, batch size of 32 and default beta values given by the pytorch framework [28]. α is set to 10 in combined loss training and the model is trained for a fixed amount of 25 epochs. Following previous work on image caption [29], we calculate the CIDEr score on the validation set after each epoch and choose the model with the highest CIDEr score for evaluation.

5.4. Results

Results are analyzed from two aspects: 1) the model performance evaluated by different metrics; 2) the model generalization capabilities on different datasets.

5.4.1. Evaluation Metrics

The presentation of our results is split into 1) Objective metrics, including BLEU@1-4, ROUGE and CIDEr; 2) Human Evalu-

[1] https://storage.googleapis.com/bert_models/2018_11_03/chinese_L-12_H-768_A-12.zip

Table 4: *Results on the Car evaluation set. ℓ_{CE} and $\ell_{combined}$ correspond to different loss functions (see Equation (1) and Equation (4)).*

Metric	Training loss	
	ℓ_{CE}	$\ell_{combined}$
BLEU$_1$	**0.720**	0.706
BLEU$_2$	0.549	**0.553**
BLEU$_3$	0.415	**0.433**
BLEU$_4$	0.322	**0.348**
ROUGE$_L$	0.491	**0.518**
CIDEr	0.364	**0.447**
Richness	0.116	**0.213**

ation, involving eight native speakers' ratings on machine- and human-generated captions.

Objective Metrics Classical NLG evaluation metrics include BLEU@1-4, ROUGE$_L$, CIDEr, METEOR and SPICE. However, METEOR and SPICE depend on a paraphrasing library named WordNet while there is no such library for Mandarin. Therefore, we only present BLEU@1-4, ROUGE$_L$ and CIDEr here. In addition to these scores, we also directly count the unique caption number in all predictions, indicating the output richness. Table 4 illustrates the objective results. Except BLEU$_1$, the model trained with $\ell_{combined}$ achieves a performance improvement on all evaluation metrics compared with ℓ_{CE} trained model. The most significant improvement lies in CIDEr, achieving a 22.8% relative gain. The improvement in objective metrics indicates that sentence-level loss is helpful in training the model to output content-correlated sentences. In addition, richness of the model predictions also increases from 0.116 to 0.213. Sentences generated by $\ell_{combined}$ trained model may have similar semantic meanings but are diverse in expression.

An example is provided here:

Human Annotation: When the car is moving, a man is talking with music on the car and the sound of braking.

CE Loss Prediction: When the car is moving, the male driver is talking with the female passenger accompanied by the engine sound.

Combined Loss Prediction: The car is moving with music playing. The male driver is talking and suddenly the car hits another one.

Human Evaluation Eight native Mandarin speakers are invited to evaluate the model predictions. We randomly pick five audios for evaluation. Human annotations, ℓ_{CE} predictions and $\ell_{combined}$ predictions are evaluated. Raters score each caption on a five-point scale, where 1 stands for the least and 5 signifies the most useful. Results show that human annotations averaged 4.05, followed by $\ell_{combined}$ (scored 3.63), with ℓ_{CE} predictions being the least useful (scored 3.18). Although $\ell_{combined}$ predictions show significant advantage against ℓ_{CE} predictions in terms of human evaluation scores, there is still a gap between our model predictions and human annotations. Examples are provided at the end of this section.

5.4.2. Generalization

In order to verify the generalization capabilities of our model and the combined loss function, we train the model on the other two datasets: *Hospital* dataset and *Joint* dataset. Results evaluated by different metrics can be seen in Table 5. The advantage in description accuracy and diversity are verified. Firstly, the GRU encoder-decoder model with our proposed sentence-level loss can be generalized to other datasets. There is an improvement on all metrics for both datasets, comparing $\ell_{combined}$ with the ℓ_{CE} baseline. Specifically, for the *Joint* dataset, annotations on different scenes are mixed while the improvement is still significant, indicating that the proposed sentence-level loss is not only effective on a specific scene. Secondly, the current model is capable of generating richer sentences. On both datasets, there is an about 30% relative increase in richness compared with the baseline model trained with only CE loss.

Table 5: *Results on Hospital and Joint datasets, trained by different loss functions.*

Datasets		Hospital		Joint	
Training loss		ℓ_{CE}	$\ell_{combined}$	ℓ_{CE}	$\ell_{combined}$
Metric	BLEU$_1$	0.526	**0.543**	0.614	0.614
	BLEU$_2$	0.430	**0.432**	0.235	**0.243**
	BLEU$_3$	0.205	**0.229**	0.311	**0.317**
	BLEU$_4$	0.144	**0.166**	0.235	**0.243**
	ROUGE$_L$	0.389	**0.392**	0.429	**0.442**
	CIDEr	0.326	**0.366**	0.435	**0.512**
	Richness	0.429	**0.566**	0.347	**0.437**

Hyp Score 5: Accurate, comprehensive and vivid description
Hyp: 汽车在行驶中男司机在和女乘客聊天伴随着发动机声
The male driver is chatting with the female passenger while the car is moving.
Ref 1: 行车过程中司机和后排乘客说话
The driver and the passenger on the back are talking during driving. (Score 4)
Ref 2: 车在行驶中男司机找女乘客搭讪女乘客小声应答
The driver strikes up a conversation with a female passenger while the car is moving. (Score 5)

Hyp Score 3: Generally correct, with some missing or redundant description
Hyp: 汽车停在路边男司机在介绍汽车
The car parks at the roadside and the male driver is introducing the car.
Ref 1: 汽车停靠在马路边司机讲解汽车性能有风噪声
The car parks at the roadside. The driver introduces the car performance along with wind noise. (Score 4)
Ref 2: 汽车停在路边男司机对女乘客讲解相关内容有车噪声
The car parks at the roadside. The male driver introduces it to the female passengers along with car noise. (Score 5)

Hyp Score 1 Not suitable at all
Hyp: 汽车在行驶中男司机和女乘客在聊天
The male driver and the female passenger are chatting while the car is driving
Ref 1: 车辆在高速行驶车里在放音乐
The car is running fast with music playing. (Score 3)
Ref 2: 汽车行驶中车内放着音乐外面传来物体落在车上的声音汽车停住了
When the car is running with music in it, there is sound outside the car. Then the car stops. (Score 5)

6. Conclusion

In this paper, we propose a 10 hour long Car scene corpus. Further, a sentence-level loss to provide a supervision signal from the sentence semantics is proposed. Metrics including classical NLG metrics and output richness show that our approach now generates more content-related captions with higher diversity. Human evaluation results also validate the advantage of the added sentence-level loss. Validation of the proposed approach is done on three Mandarin audiocaption datasets (Hospital, Car, Joint), verifying its generalization capability. Despite the effectiveness of our proposed GRU encoder-decoder model with a sentence-level loss, there is still a significant gap between model predictions and human annotations.

7. References

[1] Y. Koizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "The NTT DCASE2020 challenge task 6 system: Automated audio captioning with keywords and sentence length estimation," DCASE2020 Challenge, Tech. Rep., June 2020.

[2] Y. Wu, K. Chen, Z. Wang, X. Zhang, F. Nian, S. Li, and X. Shao, "Audio captioning based on transformer and pre-training for 2020 DCASE audio captioning challenge," DCASE2020 Challenge, Tech. Rep., June 2020.

[3] H. Wang, B. Yang, Y. Zou, and D. Chong, "Automated audio captioning with temporal attention," DCASE2020 Challenge, Tech. Rep., June 2020.

[4] X. Xu, H. Dinkel, M. Wu, and K. Yu, "The SJTU submission for DCASE2020 task 6: A CRNN-GRU based reinforcement learning approach to audiocaption," DCASE2020 Challenge, Tech. Rep., June 2020.

[5] M. Wu, H. Dinkel, and K. Yu, "Audio Caption: Listen and Tell," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2019-May. Institute of Electrical and Electronics Engineers Inc., may 2019, pp. 830–834.

[6] R. Prabhavalkar, T. N. Sainath, Y. Wu, P. Nguyen, Z. Chen, C.-C. Chiu, and A. Kannan, "Minimum word error rate training for attention-based sequence-to-sequence models," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4839–4843.

[7] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," *arXiv preprint arXiv:1511.06732*, 2015.

[8] C. D. Kim, B. Kim, H. Lee, and G. Kim, "AudioCaps: Generating captions for audios in the wild," in *Proc. Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 119–132. [Online]. Available: https://www.aclweb.org/anthology/N19-1011

[9] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020.

[10] R. Pasunuru and M. Bansal, "Multi-task video captioning with video and entailment generation," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017, pp. 1273–1283.

[11] C. Hori, T. Hori, T.-Y. Lee, Z. Zhang, B. Harsham, J. R. Hershey, T. K. Marks, and K. Sumi, "Attention-based multimodal fusion for video description," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 4193–4202.

[12] C. Hori, T. Hori, G. Wichern, J. Wang, T.-y. Lee, A. Cherian, and T. K. Marks, "Multimodal attention for fusion of audio and spatiotemporal features for video description," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, pp. 2528–2531.

[13] S.-P. Chuang, C.-H. Wan, P.-C. Huang, C.-Y. Yang, and H.-Y. Lee, "Seeing and hearing too: Audio representation for video captioning," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2017, pp. 381–388.

[14] Z. Shen, J. Li, Z. Su, M. Li, Y. Chen, Y.-G. Jiang, and X. Xue, "Weakly supervised dense video captioning," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1916–1924.

[15] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.

[16] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. International Conference on Machine Learning (ICML)*, 2014, pp. 1188–1196.

[17] B. McCann, J. Bradbury, C. Xiong, and R. Socher, "Learned in translation: Contextualized word vectors," in *Proc. Conference on Neural Information Processing Systems (NIPS)*, 2017, pp. 6294–6305.

[18] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proc. Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018, pp. 2227–2237.

[19] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018.

[20] M. Pagliardini, P. Gupta, and M. Jaggi, "Unsupervised learning of sentence embeddings using compositional n-gram features," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018, pp. 528–540.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019, pp. 4171–4186.

[22] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3156–3164.

[23] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. International Conference on Machine Learning (ICML)*, 2015, pp. 2048–2057.

[24] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3128–3137.

[25] 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.

[26] P. Qi, T. Dozat, Y. Zhang, and C. D. Manning, "Universal dependency parsing from scratch," in *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Brussels, Belgium: Association for Computational Linguistics, October 2018, pp. 160–170. [Online]. Available: https://nlp.stanford.edu/pubs/qi2018universal.pdf

[27] D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *The International Conference on Learning Representations (ICLR)*, pp. 1–13, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[28] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Proc. Conference on Neural Information Processing Systems (NIPS)*. Curran Associates, Inc., 2019, pp. 8026–8037. [Online]. Available: http://arxiv.org/abs/1912.01703

[29] S. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1179–1195.