What does a Car-ssette tape tell?
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 little work examines audio captioning. This paper contributes a manually-annotated dataset on car scene, in extension to a previously published hospital audio captioning dataset. An encoder-decoder model with pretrained word embeddings and additional sentence loss is proposed. This current model can accelerate the training process and generate semantically correct but unseen unique sentences. We test the model on the current car dataset, previous Hospital Dataset and the Joint Dataset, indicating its generalization capability across different scenes. Further, we make an effort to provide a better objective evaluation metric, namely the BERT similarity score. It compares the semantic-level similarity and compensates for drawbacks of N-gram based metrics like BLEU, namely high scores for word-similar sentences. This new metric demonstrates higher correlation with human evaluation. However, though detailed audio captions can now be automatically generated, human annotations still outperform model captions in many aspects.

Index Terms: Audio Caption, Audio Dataset, Evaluation Metric, 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 those features 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, audio captioning does not attract much attention [1], unlike in the image and video fields.

A human-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 [2]. Although the model performance evaluated by BLEU score is particularly high, human evaluation tells a different story. Most machine-generated sentences are monotonous and repetitive, while by contrast human annotations are much more specific in content and vivid in expression. Therefore, our model should endeavor to generate more unique sentences that not only describe detailed audio content but also contain richer vocabulary and diverse sentence structures. For example, for a sound of car crash in an audio clip, a nice model is expected to generate a caption like “The car went into a crash with other cars” or “Traffic accident happened”, instead of repetitive “There is a sound of car crash” or even “There is a sound of cars”.

To achieve the goal of generating specific captions with various expressions, we first publish a dataset on car scene with an updated labelling strategy. Followed by that, we address the variety lacking problem by making use of context-aware sentence embedding with variable length. In addition, we endeavor to provide a more reliable way to evaluate the machine generated captions. As pointed out previously, current objective metrics could not evaluate the machine-generated sentences as expected [2]. Metrics based on N-Grams mostly consider contextual word co-occurrence, which work well in examining the word/sentence structure similarity rather than semantic relevance. Since we aim to generate semantically human-like but form-diversified sentences, a new robust objective metric is in urgent need.

2. Related Work
Captioning Model Image and video captioning has witnessed promising improvement recently. The development of sequence-to-sequence models enables well-performing video captioning models by simply using temporal image information [3]. Later, the attention mechanism was utilized to fuse audio with video information and assign different importance to time frames [4].[5]. [6]. [7] and [8] generated multiple captions in different detail levels and temporal attention.

Sentence Embedding Early works like GloVe [8] and Word2Vec [9] in natural language processing (NLP) focused on context-free embedding of words. Recently, models like Cove [10], ELMo [11] and GPT [12] made use of the self-attention mechanism and transformer to build context-sensitive word representations. An unsupervised, C-BOW like method to embed sentence to fixed length vector [13] was later proposed. In this paper, our work is based on the state-of-art sentence embedding technique from Google named BERT [14]. It contains large bidirectional transformers trained on huge corpus, thus embeddings extracted from pretrained BERT model can perform well in many tasks with a little fine-tuning work.

Evaluation Metric In previous captioning work evaluation metrics were mainly borrowed from machine translation: BLEU@1-4, METEOR, CIDEr and ROUGE-L scores were calculated. All these metrics are based on N-Gram overlaps between hypothesis and reference [15][16][17][18]. [18] treated image captioning as a sentence ranking task and used recall@k and median@r as their metric. Chuang et al. [6] used the Sent2Vec model from [13] to embed sentences to fixed length vectors. In addition to a BLEU score, a sentence embedding cosine similarity between model outputs and human transcriptions was involved as a semantic evaluation. In our current work, we were inspired by such a sentence-level evaluation.
3. A Cassette on Car Scene: Car-ssette

This work publishes a 10h Mandarin-annotated dataset on car scene that enables audio captioning, in extension to the aforementioned Hospital Dataset [2]. English translations by using Baidu translator are also provided for broader accessibility. We refer to this current car dataset as ‘car-ssette’ since it contains 3602 car-scene related audio clips (‘cassette tapes’), each lasting for 10s. This dataset exhibits a handful of discrepancies from the Hospital Dataset:

• This current car dataset includes large quantities of real-life recordings while the Hospital Dataset consists of more video clips from TV shows. Table 1 shows the top 5 sound events. “Engine sound” appears in over $\frac{1}{3}$ of all audio clips.

Table 1: 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         |

• We adjust the number of annotations for each clip and labelling strategy: the number of native Mandarin annotators is updated from 3 in previous hospital scene to the current 5; in contrast to the previously comprehensive labelling strategy, this current dataset utilizes a concise method by only including natural sentence annotations, and generating other metadata e.g. sound events, subjects, etc. directly from the natural descriptions.

Albeit the successful generation of grammatically-correct and scene-related captions in the Hospital Dataset, the lack of specificity and variety is to be resolved. More interestingly, the commonly embedded objective metrics could not evaluate our results effectively. The current adjusted dataset prompts a shift in focus onto improving the machine-generated sentence quality and searching for a better evaluation metric.

The car dataset is split into a training set and a development set, which encompasses 3241 and 361 audio clips respectively. High sentence diversity is observed in both sets: only 6.7% transcriptions in the training set and 1.9% in the development set are repeated. From the distribution of the top 5 tokens in Table 2 it can be seen that the train-development set exhibits a similar token distribution.

Table 2: Token Distribution in DataSet

| Rank | Token | Train % | Dev % |
|------|-------|---------|-------|
| 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  |

To investigate our model generalization capabilities, in particular under cross-scene circumstances, we further experimented on two other datasets. One is the Hospital Dataset [2], including 3707 audio clips with 3 human annotations for each; the other is the creation of a Joint Dataset that merges the car and hospital datasets. It should be noted that this is a balanced dataset since the number of audio clips within the 2 datasets are similar (car: 3602; hospital: 3707).

4. Model Description

Since the encoder-decoder model [2] can generate audio relevant and grammatically correct sentences, we continue to incorporate a similar architecture with certain modifications for performance enhancement.

For each audio clip, the GRU encoder reads a filterbank (Fbank) feature and encodes it into a fixed length feature vector. During training, teacher forcing is used to accelerate the training process. The fixed length feature vector is thus concatenated with the ground truth transcription embedding and then fed to the decoder. During evaluation and testing, no transcriptions are available, thus an Fbank feature is directly fed to the decoder in order to generate natural language captions. For every timestep, the decoder generates a single token until the “<EOS>” token is generated (see Figure 1).

$$\ell_{CE}(S, S') = -\sum_{t=1}^{T} S'(t) \log S(t)$$

Standard cross entropy is used as the word-level loss eq. (1), which is defined as the negative log likelihood of the expected word $S(t)$ given transcription $S'(t)$ at time $t$. In addition, we proposed a sentence-level loss function to compare the semantic similarity on the basis of an added mean pooling layer.

Pooling Layer Since the decoder outputs a vector equal to the size of our sentence representation (e.g., BERT) at each timestep, pooling is required in order to compare semantic sim-
ilarity between human transcription and model prediction. In this work mean pooling is exclusively utilized.

On the basis of such a pooling layer, the semantic similarity could be calculated by a cosine similarity between the two sentence embeddings. In order to minimize the embedding difference between predicted ($S_e$) and transcribed sentences ($S_{e'}$), we developed a sentence loss function opposed to cosine similarity (see Equation (2), where $\epsilon$ is a small number ensuring numerical stability). In this way, the smaller the sentence loss is, the higher the cosine similarity can be.

$$\ell_S(S_e, S_{e'}) = 1 - \frac{S_e \cdot S_{e'}}{\max(||S_e||_2, ||S_{e'}||_2, \epsilon)}$$  \hspace{1cm} (2)

Accordingly, the training objective (Equation (1)) minimizes the sum of the word (Equation (1)) and sentence (Equation (2)) loss.

$$\ell(S, S', S_{e}, S_{e'}) = \ell_{CE}(S, S') + \ell_S(S_e, S_{e'})$$  \hspace{1cm} (3)

Using a word and sentence loss combination, our model is expected to generate captions with human-like content while being diversified in sentence structure.

**Pretrained Word Embeddings** Due to dataset limitations, the model in [2] trains the embedding layer from scratch, which is likely to bias embeddings towards the most commonly seen words. This potentially leads the model to generate repeated captions. In order to partially circumvent this problem and generate sentences with richer vocabulary, we utilized pretrained Word2Vec and BERT embeddings. Word2Vec, a context-insensitive embedding method, could be pretrained from human annotations in our dataset. By contrast, BERT sentence embeddings are pretrained on large, scene-free corpus and possess a strong sensitivity to context.

5. Experiments

5.1. Data preprocessing

Typical 64 dimensional FBank features from a 25ms window were extracted every 10ms. During training we applied 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, we need to tokenize transcriptions. Here, Stanford core NLP tools[19] were used for parsing. Word and sentence embeddings based on Word2Vec[9] were trained. We also utilized BERT[14] on published simplified and traditional Chinese models to initialize fixed-length embeddings.

5.2. Training Process

It is worth mentioning that during the training stage, cross validation was not utilized. We justify this practice by reiterating that the focus of this work is to generate unique, previously unseen sentences, rather than reproducing them. A word-level cross validation could only provide information on how well the model can generate same-looking captions, being of little help for the variety purpose.

Training was done using the Adam optimization algorithm with learning rate $4e^{-4}$, batch size of 32 and default beta values given by the pytorch framework[20]. We experimented on three embedding layer initialization methods: no pretrained embedding, pretrained Word2Vec, and pretrained BERT. The training loss curve for each epoch (see Fig[2]) is plotted to visualize the effect of different embeddings with and without pretraining. It revealed that a pretrained embedding speeds up training (lower initial loss and faster convergence). Furthermore, by using BERT embeddings to estimate the sentence cost $\ell_S$, a noticeably lower final loss was obtained. This may be in part due to the context-sensitive nature of BERT sentences with similar semantic meaning are embedded in close proximity to each other, whereas Word2Vec sentences are strictly embedded on word-level.

5.3. Results

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

5.3.1. Evaluation Metrics

In addition to the classic metric BLEU score, an N-Gram based evaluation method, and human evaluation, we proposed a semantic similarity metric: BERT similarity score. The presentation of our metric results is thus split into 1) Objective metrics, including BLEU and BERT scores, with comparison to the number of unique sentences generated, which stands for caption richness; 2) Human Evaluation, involving 20 native speakers’ ratings on machine- and human-generated captions.

**Objective Metrics** For every input utterance, we have $R$ independent transcriptions, and $D$ is the number of input utterances in the evaluation set. Objective scores per utterance are calculated by picking 1 transcription as hypothesis and treating the other $R-1$ as references. Accordingly, every utterance should have $R-1$ scores. The maximum value was chosen as the representative score. Regarding BLEU score, this work exclusively used the 4-gram BLEU scores, the calculation method of which is shown in Equation (4), which signifies the N-gram similarity. The BERT similarity score, a metric for semantic similarity, is calculated as the opposite of $\ell_S$ (see Equation (5)).

$$S(\text{BLEU}) = \frac{1}{D} \sum_{d=1}^{D} \max_{r=1}^{R} \text{BLEU}_4(\text{Ref}_r, \text{Hyp}_d)$$  \hspace{1cm} (4)

$$S(\text{BERT}) = \frac{1}{D} \sum_{d=1}^{D} \max_{r=1}^{R} [1 - \ell_S(\text{Ref}_r, \text{Hyp}_d)]$$  \hspace{1cm} (5)

Table[3] illustrates the results evaluated by BLEU$_4$, BERT similarity score, and output richness (number of unique captions generated). Interestingly, such a richness evaluation accords with neither BLEU$_4$ nor the BERT score, indicating that simply counting on unique sentence generation could not ensure its similarity with human annotations. It should be noted that our goal is to generate unique and semantically human-like captions.
By comparing the results of BLEU4 and the BERT similarity score with the output richness, it can be summarised that BLEU4 overlooked how unique its outputs are, while the BERT similarity score worked more effectively by evaluating semantic similarity without sacrificing output richness. For instance, Word2Vec achieved the highest and near-human BLEU4 score but its outputs were far from satisfaction; the caption “The male driver is chatting with the female passenger while the car is moving” was repeatedly generated for nearly 1000 times over the all 3602 audio clips. Without any doubt, the number of unique sentences it generated was the lowest. However, these sentences largely overlapped with human transcriptions on an N-Gram basis, consequently leading to a high BLEU4. This again echoes the finding[2] that BLEU is less efficient in evaluating captioning tasks.

When evaluating by the BERT similarity score, BERT+ℓs embedding scored the highest and could generate moderately rich captions. Again, though Baseline could generate the most unique sentences but its content-correctness was relatively low.

By comparing the results of BLEU4 and the BERT similarity score on Joint Dataset are relatively good, meaning that the baseline model can distinguish different scenes. Due to the fact that every audio in hospital scene only had 3 annotations, our model showed its preference towards the car scene: it mistakenly generated hospital-related captions for only 12 car inputs and 308 car-related captions for hospital inputs.

Table 3: Evaluation results on the car development set. ℓs(S,S′) is the additional sentence embedding loss

| Embedding       | BLEU4 | BERT | Unique # |
|-----------------|-------|------|----------|
| Baseline        | 0.220 | 0.919| 710      |
| Word2Vec        | 0.261 | 0.922| 269      |
| Word2Vec + ℓs(S,S′) | 0.235 | 0.917| 273      |
| BERT            | 0.245 | 0.924| 562      |
| BERT + ℓs(S,S′) | 0.246 | 0.925| 429      |
| Human           | 0.266 | 0.935| 3503     |

6. Conclusion

Table 4: Results of the baseline model trained on 3 datasets.

| BLEU4 | BERT |
|-------|------|
| Model | Human| Model | Human|
| Hospital | 0.127 | 0.137 | 0.937 | 0.942 |
| Car    | 0.220 | 0.266 | 0.919 | 0.935 |
| Joint  | 0.157 | 0.185 | 0.925 | 0.954 |

Previous work on audio captioning was limited in scenery (hospital) richness, the question of expanding towards a larger corpus thus comes to mind. We provide a 10h long car scene corpus in order to cope with data limitations. The results show that by fine-tuning a model using pretrained BERT embeddings as well as our proposed BERT sentence loss, rich sentences with content-related captions can now be generated. Results on the Joint Dataset show that most outputs from the baseline model are content-related, verifying its generalization capability. This paper puts forth another major problems for audio captioning, that an effective and reliable objective metric is lacking. To ascertain the usefulness of a caption is mainly reliant on costly, manual scoring. Along with the investigation on different pre-trained embeddings, we draw a final conclusion that BERT embeddings with added sentence loss could generate most useful captions, which is the same evaluation result by our proposed BERT similarity score and human evaluation.

Hyp Score 4: Accurate, comprehensive and useful
Hyp: 汽车停在路边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: 行驶过程中司机和乘客聊天
The driver and the passenger on the back are talking during driving. (Score 3)
Ref 2: 汽车停在路边男司机在介绍汽车
The driver strikes up a conversation with a female passenger while the car is moving. (Score 4)

Hyp Score 2: General description, not specific enough
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: 汽车停在路边司机在介绍汽车
The car parks at the roadside. The driver introduces the car performance along with wind noise. (Score 3)
Ref 2: 汽车停在路边司机在介绍汽车
The car parks at the roadside. The male driver introduces it to female passengers along with car noise. (Score 4)

Hyp Score 0 Not suitable at all
Hyp: 在行驶中男司机和女乘客在聊天
The male driver and the female passenger chatting while the car is driving
Ref 1: 车辆在高速行驶里怠速
The car is running fast with music playing. (Score 2)
Ref 2: 汽车行驶中车内乘客声音外面由单位体体验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 3)

5.3.2. Generalization

In order to verify our model’s generalization capabilities, we trained the baseline model on all 3 datasets. Results evaluated by different metrics can be seen in Table 3. Firstly, our model is capable of being generalized to other datasets, in particular the cross-scene dataset: both the BLEU4 and the BERT similarity score on Joint Dataset are relatively good, meaning that the baseline model can distinguish different scenes.

An example could be found here: The car parks at the roadside. The male driver introduces it to the female passenger while the car is moving. (Score 4)

Hyp Score 4: Accurate, comprehensive and useful
Hyp: 汽车在行驶中男司机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: 行驶过程中司机和乘客聊天
The driver and the passenger on the back are talking during driving. (Score 3)
Ref 2: 汽车停在路边男司机在介绍汽车
The driver strikes up a conversation with a female passenger while the car is moving. (Score 4)

Hyp Score 2: General description, not specific enough
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: 汽车停在路边司机在介绍汽车
The car parks at the roadside. The driver introduces the car performance along with wind noise. (Score 3)
Ref 2: 汽车停在路边司机在介绍汽车
The car parks at the roadside. The male driver introduces it to female passengers along with car noise. (Score 4)

Hyp Score 0 Not suitable at all
Hyp: 在行驶中男司机和女乘客在聊天
The male driver and the female passenger chatting while the car is driving
Ref 1: 车辆在高速行驶里怠速
The car is running fast with music playing. (Score 2)
Ref 2: 汽车行驶中车内乘客声音外面由单位体体验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 3)
7. References

[1] K. Drossos, S. Adavanne, and T. Virtanen, “Automated audio captioning with recurrent neural networks,” in 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), Oct 2017, pp. 374–378.

[2] M. Wu, H. Dinkel, and K. Yu, “Audio caption: Listen and tell,” arXiv preprint arXiv:1902.09254, 2019.

[3] R. Pasunuru and M. Bansal, “Multi-task video captioning with video and entailment generation,” arXiv preprint arXiv:1704.07489, 2017.

[4] 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, 2017, pp. 4193–4202.

[5] 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 Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 2528–2531.

[6] 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 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2017, pp. 381–388.

[7] 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 Conference on Computer Vision and Pattern Recognition, 2017, pp. 1916–1924.

[8] J. Pennington, R. Socher, and C. Manning, “Glove: Global vectors for word representation,” in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[9] Q. Le and T. Mikolov, “Distributed representations of sentences and documents,” in International conference on machine learning, 2014, pp. 1188–1196.

[10] B. McCann, J. Bradbury, C. Xiong, and R. Socher, “Learned in translation: Contextualized word vectors,” in Advances in Neural Information Processing Systems, 2017, pp. 6294–6305.

[11] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, “Deep contextualized word representations,” arXiv preprint arXiv:1802.05365, 2018.

[12] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving language understanding by generative pre-training,” URL https://s3-us-west-2.amazonaws.com/ai2-s2.0/pdf/2018/03/24/language_understanding_language_understanding.pdf, 2018.

[13] M. Pagliardini, P. Gupta, and M. Jaggi, “Unsupervised learning of sentence embeddings using compositional n-gram features,” arXiv preprint arXiv:1703.02507, 2017.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018.

[15] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3156–3164.

[16] 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 International conference on machine learning, 2015, pp. 2048–2056.

[17] A. Karpathy and L. Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3128–3137.

[18] 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.

[19] 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/qa2018universal.pdf

[20] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in pytorch,” 2017.