CAN AUDIO CAPTIONS BE EVALUATED WITH IMAGE CAPTION METRICS?

Zelin Zhou, Zhiling Zhang*, Xuenan Xu*, Zeyu Xie, Mengyue Wu†, Kenny Q. Zhu†*

Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China

ABSTRACT

Automated audio captioning aims at generating textual descriptions for an audio clip. To evaluate the quality of generated audio captions, previous works directly adopt image captioning metrics like SPICE and CIDEr, without justifying their suitability in this new domain, which may mislead the development of advanced models. This problem is still unstudied due to the lack of human judgment datasets on caption quality. Therefore, we first construct two evaluation benchmarks, *AudioCaps-Eval* and *Clotho-Eval*. They are established with pairwise comparison instead of absolute rating to achieve better inter-annotator agreement. Current metrics are found in poor correlation with human annotations on these datasets. To overcome their limitations, we propose a metric named *FENSE*, where we combine the strength of Sentence-BERT in capturing similarity, and a novel Error Detector to penalize erroneous sentences for robustness. On the newly established benchmarks, FENSE outperforms current metrics by 14-25% accuracy. [1]

***Index Terms*—** Audio captioning, image captioning, caption evaluation, pre-trained model

1. INTRODUCTION

Automated audio captioning [1] is the task of automatically generating human-like content description of an audio signal using free text. Recent progress has been focused on the development of caption datasets [2, 3, 4], in which novel algorithms [5, 6, 7] are cultivated and fostered rapidly. However, little attention has been addressed on the automatic evaluation metrics. Current evaluations for audio caption directly adopt the metrics from the image captioning literature, including those for general Natural Language Generation (NLG) (*e.g.* BLEU [8], ROUGE [9], METEOR [10]) or specifically for image captioning (*e.g.* CIDEr [11], SPICE [12]), without considering their generalizability to audio domain.

In this work, we ask: *"Can audio captions be evaluated with image caption metrics?"*, since a biased evaluation may hinder the improvement of algorithms or even lead the journey to a skewed direction. In Fig 1, we show that current metrics cannot properly evaluate audio captioning. On one hand, traditional metrics like BLEU may fail to capture semantic similarity beyond exact word matching [13, 14]. Here, Caption A roughly paraphrases reference with all the sound events aligned, while Caption B partially aligned to reference with a misprediction. However, all the N-gram overlap based metrics give much higher scores to Caption B since it has more words in common with reference.

Reference: *Paper crackling* with *female speaking* lightly in the background

Caption A: Young *woman speaking* with *crinkling noise*

Caption B: Wind blowing lightly followed by a *female speaking*

Caption C: Wind blowing lightly followed by a *female speaking* and in the

Metric	A	B	C
BLEU_1	0.10	0.33	0.45
ROUGE_L	0.12	0.23	0.40
CIDEr	0.11	0.41	0.64
SPICE	0	0.44	0.44
Sentence-BERT	0.78	0.62	0.62

Fig. 1. Column A, B and C shows the corresponding scores of Caption A, B and C given by different metrics. One cell is colored if the caption is favoured by a certain metric of this row.

On the other hand, the specificity in vision-focused captions may cause the failure of image caption specific metrics. In image captioning, attention has been drawn to the visible attributes of objects, along with their spatial relationships (*e.g. A women in blue blouse sits at a table, with crinkled paper in her hand*). Conversely in audio captioning, importance has been attached to auditory properties of events as well as their temporal connections (*e.g. Young women speaking with crinkling noise*). By this means, scene graph based metrics (*e.g.* SPICE [12]) are unlikely applicable to audio caption evaluation, in that object attributes and relations are seldom mentioned. It could be seen that SPICE even leaves a zero score on Caption A for its scene graph disjointness with reference, despite their semantic affinity.

Moreover, *fluency issues* like incomplete or inconsistent sentences are quite common in current generated captions. However, current metrics do not penalize, and sometimes even favour captions with these errors, so they can be easily gamed by systems exploiting the weakness. As shown in Fig 1, Caption C attaches three meaningless words to Caption B, which further increases its N-gram overlap with reference. Scores indicate that all N-gram overlap based metrics are cheated by this trick, while the others are indifferent, which is also incompetence.

Human evaluation is considered the golden standard in NLG tasks. However, by far there is no benchmark human judgments available for audio captioning. It has certainly caused difficulty in conducting fair comparisons on evaluation metrics. Therefore, we annotated the first two human evaluation benchmarks for audio captioning, based on the widely accepted AudioCaps [4] and Clotho [2] dataset. We experimented with two annotation protocols adopted from image captioning [11]: absolute rating and pairwise comparison. We find that absolute rating leads to poor inter-annotator agreement, and thus establish our final annotation based on pairwise comparison, resulting in the annotation of 3,421 pairs. We then bench-

*Equal contribution

† Corresponding Author. Kenny Q. Zhu was partly supported by SJTU-CMBCC Joint Research Scheme.

[1]Code, data, and web demo available at: https://github.com/blmoistawinde/fense

mark current metrics, and find that metrics like BLEU_1 and SPICE perform no better than random guess on deciding a better generated caption in a pair.

To perform better audio caption evaluation, we explore the use of pre-trained model based metrics, including BERTScore [15] and BLEURT [16], and witness significant advantage over current methods. We also repurpose Sentence-BERT [17], which produces sentence embedding for similarity calculation and is mainly utilized for information retrieval, to perform caption evaluation (Fig 1, Last Row), and it surprisingly achieved the best performance. However, even methods with enhanced capability for capturing semantic relatedness can still fail to penalize fluency issues properly. We thus further propose *Error Detector* to recognize erroneous sentences and penalize them accordingly. We refer to the combined metric of Sentence-BERT and fluency penalty as **F**luency **EN**hanced **S**entence-bert **E**valuation (FENSE), which significantly outperforms existing metrics.

In summary, our contributions are:

1. We establish the first two benchmark datasets, *AudioCaps-Eval* and *Clotho-Eval*, for the comparison of audio caption evaluation metrics.
2. We propose FENSE, a metric consists of Sentence-BERT for similarity calculation and *Error Detector* to penalize fluency issues for robustness.
3. Results on the new benchmarks show that FENSE can outperform previous metrics in pairwise comparison accuracy by 14-25 points, and ablations suggest a significant contribution of both Sentence-BERT and Error Detector.

2. BENCHMARK DATASET CONSTRUCTION

2.1. Audio Captioning Systems

To generate candidate captions for evaluation, we involve multi-faceted audio captioning systems: 1) Nearest neighbor (NN) retrieval system. For a test audio, we retrieve its most similar training audio and take its annotation as a prediction. The similarity between two audio clips are measured by the cosine similarity of their embeddings, extracted by a pre-trained CNN [18]. 2) Fully Connected (FC) input system. An architecture similar to [3], where the input audio is first transformed to a vector and then decoded into a caption. 3) Attention (ATT) input system. The same architecture as [5], where the input audio is first transformed into an embedding sequence and then decoded into a caption using temporal attention mechanism. 4) Reinforcement Learning (RL) system. The same architecture as ATT system while the model is further fine-tuned using reinforcement learning [5], achieving the second place in the recent DCASE challenge evaluated by CIDEr and SPICE.

FC, ATT and RL systems are all sequence-to-sequence encoder-decoder frameworks. They all use a 10-layer CNN as the encoder and a single-layer GRU as the decoder. The detail structure can be found in [19]. To increase the diversity of generated captions, for all systems except NN, we employ various decoding strategies, including greedy decoding, beam search, diverse beam search [20]. In beam search, different temperatures are utilized.

2.2. Data Annotation

Based on captions generated by the aforementioned systems, we build the audio captioning evaluation dataset by collecting human judgments on captions. We build our datasets based on two benchmark audio caption datasets: AudioCaps [4] and Clotho [2]. There are two protocols to collect human judgments in image captioning: absolute rating and pairwise comparison. We first make a preliminary exploration on absolute rating protocol. Given an audio clip and a caption, raters are asked to score the caption from 1 to 4 considering both its relevance to the audio and its fluency. We randomly sample 100 audio-caption pairs. Each audio-caption pair is rated by four different human raters. However, Fleiss Kappa score [21] for different raters is only 0.18 on Clotho and 0.23 on AudioCaps, indicating poor inter-annotator agreement. Annotators also report difficulty in rating the caption.

Therefore, the second protocol, pairwise comparison, is adopted. Given an audio clip and a pair of candidate captions, four raters are asked to choose which candidate describes the audio better in terms of description accuracy and fluency. Raters are allowed to choose "I'm not sure" if they cannot distinguish which candidate is better. Following [11], we generate four pair groups, namely human-human correct (HC), human-human incorrect (HI), human-machine (HM) and machine-machine (MM). HC contains two human annotations describing the same audio. HI also contains two human annotations but one describes another randomly-picked audio. HM is formed by a human annotation and a machine generated caption for the same audio. MM is composed of two machine generated captions describing the same audio. On both Clotho and AudioCaps, we randomly pick 250 audio clips and generate the four kinds of pairs. For each audio clip, we form one HC pair, one HI pair, one HM pair and several MM pairs since the evaluation metrics are usually applied to compare captions generated by different machine systems. The number of MM pairs for each audio clip range from one to four.

To avoid indistinguishable cases between similar pairs, we adopt a filtering strategy for pair generation. We first generate all candidate pairs and calculate the similarity of each pair using Sentence-BERT embedding. Then we exclude candidate pairs with a similarity higher than 0.9. We randomly sample pairs from the remaining captions to obtain the final generated pairs for comparison. For HC and HM, we randomly sample the pairs. For MM, we randomly sample four pairs if there are more than four candidate pairs, otherwise we use all the pairs left. In case of no pairs left after filtering, we select the pair with the lowest similarity. In this way, we finally obtain human judgments on 1,750 pairs on Clotho and 1,671 pairs on AudioCaps. Higher inter-annotator agreement is achieved on both datasets, indicated by a Fleiss Kappa score of 0.48 and 0.33 respectively. We will refer to them as *AudioCaps-Eval* and *Clotho-Eval*. To our best knowledge, this is the first dataset of human judgments on the quality of audio captions. We evaluate the effectiveness of different metrics using this dataset.

3. EVALUATION METRICS

In this section we briefly introduce the automatic evaluation metrics investigated in this work. Traditional NLG Metrics and Image Caption Metrics are the current adopted ones for audio caption. Pre-trained Model Based Metrics are the existent metrics utilized in other domains and proposed to use for audio caption. Error Detector is the newly customized metric for evaluating the vastly present fluency issues in generated captions.

Traditional NLG Metrics Traditional NLG metrics mainly rely on N-gram matching. **BLEU** [8] counts the exact N-gram matches between a candidate and its corresponding reference sentences and calculate the precision, while **ROUGE** [9] is the recall-based counterpart of BLEU. A clear limitation for these methods is that even a subtle difference in the wording for expressing a similar meaning

will be counted as an error. **METEOR** [10] is proposed to alleviate the problem by supporting word stems, synonyms and simple paraphrases. However, similar words beyond such changes are still unhandled, and none of them are capable of capturing similarity based on contextualized semantics.

Image Caption Metrics For the evaluation of Image captioning models, many methods have been proposed to further leverage the specific characteristics of image captions. Since image captions can be diverse even for the same picture, **CIDEr** [11] tries to capture the consensus among multiple annotators by rating a candidate by the mean TF-IDF similarity across reference captions. Moreover, since image captions usually describe the objects, attributes and relations depicted in the image, **SPICE** [12] proposes to utilize the scene graph representation parsed from caption to capture such key concepts, and rate a candidate caption by the F1-score over scene graph tuples between a candidate and its references.

These metrics have shown high correlation with human judgments for image captioning. However, some specific characteristics of image captions may not apply to audio captioning. For instance, the difficulty in identifying complex relations and attributes for an audio may pose challenges for the parse of scene graphs. They also suffer from the limitations of traditional metrics.

Pre-trained Model Based Metrics Recently, pre-trained language models like BERT [22] have been utilized for enhancing the evaluation of NLG. Their ability to produce contextualized word representations can overcome the limitation of N-gram matching, and the large pre-training corpus may further facilitate their domain generalizability, which inspired us to investigate their performance for the evaluation of audio captioning. **BERTScore** [15] leverages the contextualized word embeddings from the original BERT to calculate similarity on word-level, and aggregates them into sentence similarity with IDF weighting scheme. **BLEURT** [16] follows the original BERT to pair the sentences as input and uses a linear layer on top of the [CLS] embedding to predict the score. They further finetune the BERT backbone specifically for evaluation.

We also explore a novel use of **Sentence-BERT** [17]. It is a modification of BERT that uses siamese network structure to learn sentence embeddings so that their cosine similarity can reflect the semantic similarity. Although it was originally proposed for similarity search or clustering, we hypothesize that its capability of measuring semantic similarity may also benefit the evaluation of audio captioning. Therefore, we rate a candidate caption by its average cosine similarity with references according to Sentence-BERT embeddings.

Error Detector Fluency issues like repeated events and incomplete sentences are prevalent in current audio captioning systems. However, few of the current evaluation metrics are able to take them into consideration, resulting in their overestimation of system quality. To mitigate this problem, we propose to use a separate error detector to penalize the scores given by other evaluation metrics when fluency issues are detected. We investigated the outputs of the captioning systems (§2.1), and found several typical types of fluency issues. We summarize them and give examples in Table 1.

Type	Example
Incomplete Sentence	a woman is giving a speech and a *(...)*
Repeated Event	music plays followed by *music playing*
Repeated Adverb	sheep bleats nearby several times *nearby*
Missing Conjunction	people speaking *(and)* a train horn blows
Missing Verb	food sizzles and a pan *(verb)*

Table 1. Types and examples of fluency issues. We use parentheses to mark the missing information.

To train a model for detecting such errors, we produce a synthetic training set by using rules to corrupt the correct captions into erroneous ones with such issues. Specifically, we use the captions in Clotho and AudioCaps training set as correct captions. We then tailor modifications for each error type to apply on the correct captions and label the modified ones as having the specific error type. For example, to produce incomplete captions, we append phrases that frequently appears at the tail of observed problematic examples like "and" , "and a", "follow by", etc. We empirically produce one error per caption 90% of the time, and 2 errors otherwise. Then we mix the correct captions and corrupted captions into one set, and add an overall *Error* label to samples with at least one error. Finally, we train a BERT model on this set for a multi-label classification of each error type and an overall *Error* label.

To include the trained model for penalization, we take the model's predicted probability for *Error*. If it exceeds a predefined threshold (0.9 in this work), we will divide the original score by 10. We observe improvement when combining the penalty with any metric, and obtain the best performance with Sentence-BERT (§4.3). We thus propose FENSE as the combination of them.

4. EXPERIMENTS

4.1. Experimental Setup

We conducted experiments on the two established datasets to compare different metrics' effectiveness on deciding a better caption out of two candidate sentences. For each pair of candidates, we compute all the metrics for two sentences and consider the one with a higher score as better according to the metric. A decision is considered correct if it agrees with human judgment. We calculate per-category accuracy by dividing correct predictions with total predictions within the categories (HC, HI, HM and MM), and merge them as a total score with micro average. Note that we exclude samples with equal number of opposite human judgments (excluding "not sure") so that the chosen samples can be considered distinguishable.

Since we may also include reference sentences in comparison, we only choose 4 out of 5 sentences from original references in the calculation of metrics. For an HC pair, we exclude each used reference itself from the five references. For an HI pair or HM pair, we eliminate the first human correct sentence from original references as their common new references. For an MM pair, we form the C_5^4 combinations selecting 4 out of 5 references as their common new references and calculate metrics on each of the 5 derivatives, then the five scores are averaged counting for the final score.

To enable a fair comparison across the pre-trained model based metrics, we choose models of roughly equal size (usually small) and pre-trained on paraphrase tasks (if possible), which are BERTScore (*t5-paraphraser*), BLEURT (*BLEURT-tiny*) and Sentence-BERT (*paraphrase-TinyBERT-L6-v2*). Note here BERTScore (*t5-paraphraser*) is much larger than the other two models. BLEU_1, BLEU_4, ROUGE_L, METEOR, CIDEr and SPICE are implemented with tools provided by DCASE challenge. [2]

4.2. Standalone Evaluation of Error Detector

To evaluate the performance of the proposed *Error Detector*, we randomly sampled 723 machine generated captions, and annotated if the caption has a fluency issue. We then compare the model's prediction on the overall *Error* class (threshold=0.9) with our annotations. The model achieved a precision of 96.8, recall of 76.4, and F1 of 85.4.

[2] https://github.com/audio-captioning/caption-evaluation-tools

Metrics	AudioCaps-Eval					Clotho-Eval				
	HC	HI	HM	MM	Total	HC	HI	HM	MM	Total
$BLEU_1$	58.6	90.3	77.4	50.3	62.4	51.0	90.6	65.5	50.3	59.0
$BLEU_4$	54.7	85.8	78.7	50.6	61.6	52.9	88.9	65.1	53.2	60.5
METEOR	66.0	96.4	90.0	60.1	71.7	54.8	93.0	74.6	57.8	65.4
$ROUGE_L$	61.1	91.5	82.8	52.1	64.9	56.2	90.6	69.4	50.7	60.5
CIDEr	56.2	96.0	90.4	61.2	71.0	51.4	91.8	70.3	56.0	63.2
SPICE	50.2	83.8	77.8	49.1	59.7	44.3	84.4	65.5	48.9	56.3
BERTScore	60.6	97.6	**92.9**	65.0	74.3	57.1	**95.5**	70.3	61.3	67.5
BLEURT	**77.3**	93.9	88.7	72.4	79.3	59.0	93.9	75.4	67.4	71.6
Sentence-BERT	64.0	**99.2**	92.5	73.6	79.6	60.0	**95.5**	75.9	66.9	71.8
FENSE	64.5	98.4	91.6	**84.6**	**85.3**	**60.5**	94.7	**80.2**	**72.8**	**75.7**

Table 2. Benchmarking metrics for audio caption evaluation. Results are the correlation with human on pairwise comparisons.

	AudioCaps-Eval			Clotho-Eval		
	w/o	w	gain	w/o	w	gain
$Bleu_1$	62.4	75.3	12.9	59	68.2	9.2
SPICE	59.7	62.8	3.1	56.3	62.1	5.8
Sentence-BERT	79.6	**85.3**	5.7	71.8	**75.7**	3.9

Table 3. Ablation results of Error Detector in total accuracy.

4.3. Results on *AudioCaps-Eval* and *Clotho-Eval*

Table 2 shows our experimental results on AudioCaps-Eval and Clotho-Eval across different metrics. The results are consistent on two datasets for a common explanation. Intuitively, HC and MM pairs pose bigger challenges for all metrics, as they are either equally good or equally bad on describing the corresponding audio clip. Nonetheless, pre-trained model based metrics generally perform much better than other metrics, especially our proposed FENSE which achieves the best overall accuracy on both datasets. BERTScore falls 4-5 percent behind other pre-trained model based metrics despite its biggest model size. On AudioCaps-Eval, N-gram exact match metrics (BLEU, ROUGE) perform 14-18 points worse than Sentence-BERT, while $BLEU_1$ and $BLEU_4$ struggle to exceed a random guess accuracy on MM pairs, showing their shortage differentiating two machine generated audio captions. METEOR and CIDEr attain a moderate performance gain beyond N-gram exact match metrics, however, they are still far behind Sentence-BERT. Not surprisingly, SPICE performs the worst, even unable to obtain 50% accuracy on AudioCaps MM and Clotho-Eval HC & MM, which strongly confirms our hypothesis that image captioning specified metrics cannot be adopted in audio caption evaluation.

Table 3 shows the overall accuracy of three representative metrics with and without sentence penalty of our proposed *Error Detector* (The last row: FENSE). With *Error Detector*, all the three metrics achieve extra performance gain, even for Sentence-BERT. This suggests that metrics leveraging either N-gram matching or contextual semantic similarity will fail in detecting *fluency issues*, and *Error Detector* can help mitigate such issues, thus setting a new SOTA.

Integrating *Error Detector* into Sentence-BERT metric, we get our proposed metric FENSE. Since table 2 and 3 have shown that FENSE significantly outperforms all the above metrics on the two evaluation datasets, we can use it to evaluate existing audio captioning system mentioned in §2.1. Fig 2 illustrates the judgments over four audio captioning systems made by humans, FENSE and SPICE on Clotho-Eval. Here we choose ATT, FC and RL whose beam search temperature equals to 0.5. On the y-axis we show the fraction of times one system wins another (rated better than another systems under this metric). We can see that NN is voted as the best system by both humans and FENSE, while it's the worst under SPICE. Humans and FENSE ranked ATT, FC and RL accordingly after NN. However, SPICE fails to reach their consensus again. This indicates the superiority of FENSE, and further confirms our claim that image captioning specified metrics like SPICE are inappropriate to be transferred into audio caption evaluation.

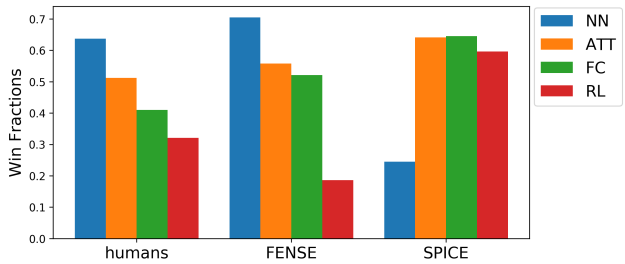

Fig. 2. Illustration of judgments made by humans, FENSE and SPICE on Clotho-Eval. y-axis shows the win fractions of 4 audio captioning systems.

5. CONCLUSION

In this work, we establish the first two human judgment datasets for audio caption evaluation, AudioCaps-Eval and Clotho-Eval, with pairwise comparison annotations. We benchmark commonly used metrics adopted from image captioning literature on the two datasets, and find their performance unsatisfying. We thus leverage Sentence-BERT for better estimation of semantic similarity and propose Error Detector to penalize sentences with fluency issues. The metric combining them two, named as FENSE, exhibits significant advantage.

Our findings suggest that the currently popular metrics for the research and competitions of audio captioning may not be a reliable measure of system quality, and better constitutes may be preferred. Although FENSE has achieved relatively good performance, we think that there is still sufficient room for improvement. Since semantic similarity does not equal to acoustic relevance [23], and only the former is covered in FENSE, better utilization of modality-specific knowledge may facilitate better evaluation as has been witnessed in image captioning [24].

6. 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)*. IEEE, 2017, pp. 374–378.

[2] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 736–740.

[3] M. Wu, H. Dinkel, and K. Yu, "Audio caption: Listen and tell," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 830–834.

[4] C. D. Kim, B. Kim, H. Lee, and G. Kim, "Audiocaps: Generating captions for audios in the wild," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 119–132.

[5] X. Xu, Z. Xie, M. Wu, and K. Yu, "The sjtu system for dcase2021 challenge task 6: Audio captioning based on encoder pre-training and reinforcement learning," DCASE2021 Challenge, Tech. Rep, Tech. Rep., 2021.

[6] X. Mei, Q. Huang, X. Liu, G. Chen, J. Wu, Y. Wu, J. Zhao, S. Li, T. Ko, H. L. Tang *et al.*, "An encoder-decoder based audio captioning system with transfer and reinforcement learning for dcase challenge 2021 task 6," DCASE2021 Challenge, Tech. Rep, Tech. Rep., 2021.

[7] X. Xu, H. Dinkel, M. Wu, and K. Yu, "A crnn-gru based reinforcement learning approach to audio captioning," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2020, pp. 225–229.

[8] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[9] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[10] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[11] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.

[12] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *European conference on computer vision*. Springer, 2016, pp. 382–398.

[13] J. Novikova, O. Dušek, A. C. Curry, and V. Rieser, "Why we need new evaluation metrics for nlg," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2241–2252.

[14] A. Chaganty, S. Mussmann, and P. Liang, "The price of debiasing automatic metrics in natural language evalaution," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 643–653.

[15] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations*, 2019.

[16] T. Sellam, D. Das, and A. Parikh, "Bleurt: Learning robust metrics for text generation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7881–7892.

[17] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3982–3992.

[18] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[19] X. Xu, H. Dinkel, M. Wu, Z. Xie, and K. Yu, "Investigating local and global information for automated audio captioning with transfer learning," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 905–909.

[20] A. K. Vijayakumar, M. Cogswell, R. R. Selvaraju, Q. Sun, S. Lee, D. Crandall, and D. Batra, "Diverse beam search for improved description of complex scenes," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, pp. 7371–7379.

[21] J. L. Fleiss, "Measuring nominal scale agreement among many raters." *Psychological bulletin*, vol. 76, no. 5, p. 378, 1971.

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[23] Z. Zhang, Z. Zhou, H. Tang, G. Li, M. Wu, and K. Q. Zhu, "Enriching ontology with temporal commonsense for low-resource audio tagging," *arXiv preprint arXiv:2110.01009*, 2021.

[24] M. Jiang, Q. Huang, L. Zhang, X. Wang, P. Zhang, Z. Gan, J. Diesner, and J. Gao, "Tiger: Text-to-image grounding for image caption evaluation," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 2141–2152.