DIVERSITY-CONTROLLABLE AND ACCURATE AUDIO CAPTIONING BASED ON NEURAL CONDITION

Xuenan Xu, Mengyue Wu†, Kai Yu†

MoE Key Lab of Artificial Intelligence
X-LANCE Lab, Department of Computer Science and Engineering
AI Institute, Shanghai Jiao Tong University, Shanghai, China

ABSTRACT

With the incorporation of pre-training, transfer learning and keyword input, notable improvement has been made in audio captioning on generating accurate audio event descriptions in recent years. However, current captioning models tend to generate repetitive and generic sentences which often contain the most frequent patterns in the training data. Some works in natural language generation make an effort to improve the diversity by attending to specific contents or increasing the generated caption number. However, these approaches often enhance the diversity with the sacrifice of description accuracy. In this work, we propose a novel neural conditional captioning model to balance the diversity and accuracy trade-off. Compared with the statistical condition, the neural condition is the posterior given by a neural discriminator. Given the reference condition, the captioning model is trained to generate captions with a similar posterior. The captioning model and the discriminator are trained in an adversarial way. We evaluate the proposed approach on Clotho and Audiocaps. The results show that compared with baselines, our approach can improve the output diversity with the least accuracy decline.

***Index Terms*—** Audio captioning, conditional generation, adversarial training, diverse caption generation

1. INTRODUCTION

Automatic audio captioning is a challenging task which requires recognizing and understanding audio contents then summarizing them with natural language. The summarization may include acoustic scenes, sound events, sound properties or even high-level abstraction [1]. It is more related to human processing than structured label outputs, suitable for automatic content description or intelligent human-machine interaction applications.

Audio captioning has attracted much attention in recent years. Researchers aim to enhance the description accuracy by incorporating techniques like pre-training [2, 3] and keyword indicator [4, 5]. However, like most natural language generation tasks, audio captioning also suffers from the diversity lacking problem. Systems trained by maximum likelihood estimation (MLE) tend to generate generic outputs [6], which are often the most common patterns in the training corpus. In contrast, human annotations may describe the same audio clip with different styles, i.e., sentence structure, wording choices.

Though some previous work has addressed the diversity problem, most of which exhibits higher diversity with declined accuracy. Some works focus on generating more descriptive, content-specific outputs [7]. For example, the system is encouraged to output "a knife" instead of "a metal object" for a knife-sharpening audio clip. In this way, when generating the same number of captions for one audio, the diversity of this audio-caption set ("set-diversity") is improved since the description is more detailed. Other works endeavor to improve diversity by generating more outputs for a single input [8, 9, 10, 11]. Multiple outputs given the same input result in a higher diversity for an input instance (we call it "instance-diversity") compared with the single-output system. For works promoting set-diversity, the captioning accuracy decrease brought by diversity improvement is in particular dramatic, for example GAN-based approaches [12, 13].

In audio captioning, Ikawa *et al.* [14] proposes a system to control the output specificity, i.e., how specific the generated caption content is. The sum of the inverse word frequency is used as the sentence specificity indicator. Multiple captions can be generated with different input specificities. Although such a statistical condition is straightforward, it only captures the word frequency characteristics, while patterns like phrase and sentence structures are not considered. The captioning model may be encouraged to generate word patterns which are not the most frequent, but still common in the training corpus ("sub-generic" patterns).

To circumvent the problem of statistical condition, we propose a new conditional audio captioning system where the condition is provided by a neural network, which uses a discriminator to tell whether a caption is generated from a human or a model. The discriminator and the captioning model are trained in an adversarial way to penalize generating sub-generic patterns. Compared with GAN-based methods, we incorporate MLE in training the captioning model to ensure accuracy. By alternating the condition, we control the description specificity, thus control the output set-diversity. Experiments are carried out on benchmark datasets, Clotho and Audiocaps. Compared with model-agnostic approaches and statistical condition approaches, our system achieves the best diversity-accuracy trade-off. The output set-diversity can be controlled by the condition with the least influence on the captioning accuracy.

2. NEURAL CONDITIONAL AUDIO CAPTIONING

In this section, we first give a brief overview of our proposed neural conditional audio captioning system. Then each part of the system is described, including an audio encoder, a neural discriminator and a text decoder. Finally, the adversarial training strategy is introduced.

2.1. System Overview

As shown in Figure 1, our proposed system generates descriptions with two inputs: an input audio sequence and a condition embed-

†Mengyue Wu and Kai Yu are the corresponding authors.

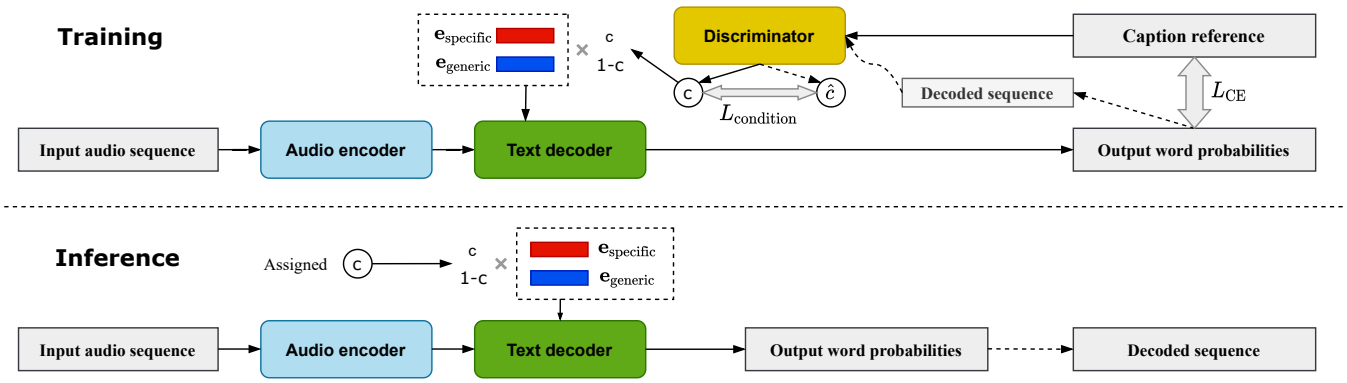

Fig. 1. Proposed neural conditional captioning framework. During **training**, the text decoder takes the reference condition and the encoded audio embedding as the input to give word probabilities. During **inference**, we manually assign the condition and feed to the text decoder to control the output set-diversity.

ding. The audio encoder transforms the input feature into an embedding sequence $\{\mathbf{e}_t^A\}_{t=1}^T$ (T is the embedding sequence length). The text decoder produces word probabilities based on both the audio embedding sequence and the condition embedding.

2.2. Audio Encoder

Since features extracted by pre-trained deep neural networks show better performance and effectiveness in training [5], we adopt PANNs [15] as the feature extractor in this work. Specifically, we use the pre-trained 14-layer convolutional neural network (CNN14). The feature map before the global pooling layer is taken as the input. Since mean-max pooling is done between convolution blocks in CNN14, each frame of the feature sequence represents a short segment of the original audio clip. We use a three-layer bidirectional gated recurrent unit (GRU) as the audio encoder to learn temporal dependencies between these segments. Taking the original audio signal as the input, the audio encoder comprises a fixed CNN14 feature extractor and a trainable GRU encoder.

2.3. Neural Discriminator

The neural discriminator takes an input audio caption and estimates its specificity c. The caption can be either a human annotation or a machine-generated one. We use a two-layer bidirectional long-short term memory (LSTM) to encode the input caption. The last timestep hidden is taken as the sequence representation and is transformed to a single value by a linear layer. Finally, c is obtained after the sigmoid activation.

2.4. Text Decoder

The input condition embedding $\mathbf{c}$ is a weighted sum of two trainable style embeddings, where the weight is a condition value $c \in [0, 1]$.

$$\mathbf{c} = c \cdot \mathbf{e}_{\text{specific}} + (1 - c) \cdot \mathbf{e}_{\text{generic}}$$

$c = 0$ denotes a generic output style while $c = 1$ encourages the system to generate content-specific descriptions. During training, the reference caption is fed into the neural discriminator to obtain c. During inference, c is manually assigned to generate descriptions with corresponding styles (generic or specific).

We utilize a unidirectional single layer GRU as the text decoder to estimate the word probabilities given the audio embedding sequence $\{\mathbf{e}_t^A\}_{t=1}^T$ and $\mathbf{c}$. Attention mechanism [16] is adopted to aggregate $\{\mathbf{e}_t^A\}_{t=1}^T$. At each timestep n, the hidden state is updated depending on the input word w_n, aggregated audio embedding $\tilde{\mathbf{e}}_A$ and $\mathbf{c}$:

$$\tilde{\mathbf{e}}_A = \text{Attention}\left(\mathbf{h}_{n-1}, \{\mathbf{e}_t^A\}_{t=1}^T\right)$$
$$\mathbf{h}_n = \text{GRU}\big([\tilde{\mathbf{e}}_A; \text{WE}(w_n); \mathbf{c}], \mathbf{h}_{n-1}\big)$$
$$\mathbf{o}_n = \text{Linear}(\mathbf{h}_n)$$

The word embedding layer WE transforms w_n into a continuous vector. Finally a linear layer outputs the probability vector $\mathbf{o}_n \in \mathbb{R}^{|\mathcal{V}|}$ of the current timestep, where $\mathcal{V}$ is the vocabulary.

2.5. Adversarial Training

Since we utilize a neural network to provide the condition c, labels are required for training the neural discriminator. Based on the observation that system outputs are often generic descriptions while human annotations are generally more diverse, we use such intuitive labels: human annotations are positive samples while system outputs are negative ones. Therefore, c of an input caption is the posterior given by the discriminator. It can be seen as a "human-like" score which measures the extent to which the caption resembles human annotations.

The encoder-decoder captioning network and the discriminator are trained in an adversarial way. During training, the following two stages are carried out alternatively. In the first stage, only the captioning model parameters are updated while the discriminator is fixed. The captioning model estimates the word probabilities at each timestep with the audio input $\mathcal{A}$ and the reference specificity c. Then the standard cross entropy (CE) loss is calculated as follows:

$$c = \text{Dis}\left(\{w_n\}_{n=1}^N\right)$$
$$\{\hat{p}_n\}_{n=1}^N = \text{Dec}\left(\text{Enc}(\mathcal{A}), c\right)$$
$$\mathcal{L}_{\text{CE}} = \sum_{n=1}^{N} -\log\left(\hat{p_n}(w_n)\right)$$

where Enc, Dec and Dis denote the encoder, decoder and discriminator, respectively. In addition to CE loss, an extra condition loss

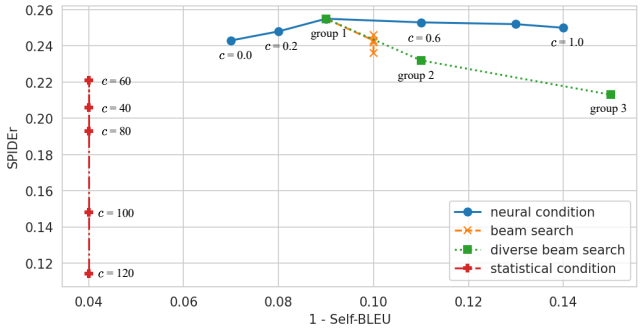

Fig. 2. The accuracy-diversity curve on Clotho.

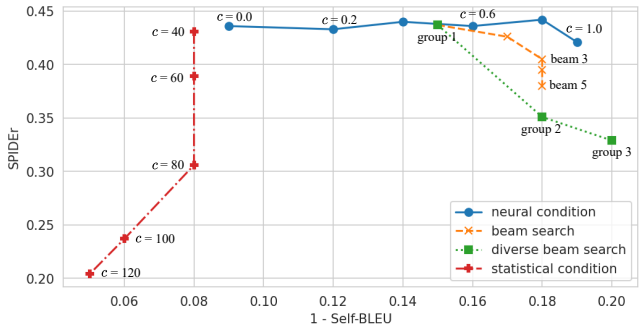

Fig. 3. The accuracy-diversity curve on Audiocaps.

$\mathcal{L}_{\text{condition}}$ is added to encourage the captioning model to generate descriptions with the corresponding specificity. It is the binary cross entropy loss between the decoded caption specificity $\hat{c}$ and c. The total captioning loss is the weighted sum of $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{condition}}$:

$$\hat{c} = \text{Dis}(\hat{\mathbf{s}}) \quad \hat{\mathbf{s}} = \underset{\mathbf{s}}{\text{argmax}}\, \hat{p}(\mathbf{s})$$
$$\mathcal{L}_{\text{condition}} = c \log(\hat{c}) + (1 - c) \log(1 - \hat{c})$$
$$\mathcal{L}_{\text{caption}} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{condition}}$$

where $\hat{\mathbf{s}} = \{\hat{w}_n\}_{n=1}^{N}$ is the model generated caption and λ is a hyperparameter. We use the reparameterization trick in the greedy decoding process to ensure the network can be trained by backpropagation.

In the second stage, the captioning network is fixed while the discriminator is trained on both human annotations and the captioning network outputs. For an input sentence $\mathbf{s}$, the discriminator training loss is calculated as:

$$c = \text{Dis}(\mathbf{s})$$
$$\mathcal{L}_{\text{discriminator}} = y \log(c) + (1 - y) \log(1 - c)$$

where the label $y = 1$ if $\mathbf{s}$ is a human annotation, otherwise $y = 0$.

3. EXPERIMENTAL SETUP

3.1. Datasets

Experiments are conducted on benchmark audio captioning datasets, Clotho [17] and Audiocaps [18]. The latest version 2.1 of Clotho is used, containing about 6k audio clips. Audiocaps contains about 50k audio clips. We use the official training, validation and testing splits for both datasets.

3.2. System Configuration

We follow the same configuration of CNN14 in PANNs [15] to extract audio features. During the captioning model training, special tokens <bos> and <eos> are added to the beginning and the end of each caption. A baseline sequence-to-sequence captioning model without the condition input is trained for 25 epochs with a maximum learning rate of 5×10^{-4}. The discriminator is also pre-trained on the mixture of the baseline model outputs and human annotations. Then the neural conditional model is initialized by the baseline model parameters and trained for 20 epochs with a maximum learning rate of 2×10^{-4}. In the first 15 epochs, only CE loss is used ($\lambda = 0$) and the discriminator is fixed. Then the whole model is trained and λ is set to 5×10^{-5} for the rest five epochs. During both the baseline model and the neural conditional model training, the learning rate linearly warms up to the maximum value in $\frac{1}{5}$ of the total iterations and then exponentially decayed to 5×10^{-7}. We use stochastic weight average [19] so that the model parameters from the last five epochs are averaged for evaluation.

3.3. Evaluation Metrics

The evaluation is conducted from two aspects: accuracy and diversity. Standard metrics including $\text{BLEU}_{1\text{-}4}$ [20], ROUGE_{L} [21], METEOR [22], CIDEr [23] and SPICE [24] are used for evaluating the captioning accuracy. SPIDEr, the mean of CIDEr and SPICE, is taken as an overall accuracy indicator since it relies on both n-gram and semantic similarities. Following previous works [7, 8], we evaluate the diversity by *distinct-1*, *distinct-2* and Self-BLEU [25]. The vocabulary size of the output captions is also calculated to measure diversity. For these diversity metrics, a higher value indicates better diversity except self-BLEU. Among them we choose Self-BLEU as the representative because it measures the {1-4}-gram overlaps between generated captions while *distinct-1* and *distinct-2* only consider unigrams and bigrams.

4. RESULTS AND ANALYSIS

4.1. Comparison with Baselines

We first compare our proposed neural condition approach with several baselines: a) model-agnostic approaches, including beam search and diverse beam search [9]; b) statistical condition [14]. Each compared method is independently evaluated by diversity and accuracy metrics. For a fair comparison, we use beam search with a beam size of 5 in all approaches. Results from one beam form an output set. In diverse beam search, the 1-best results of each group form an output set. For statistical and the proposed neural condition, each output set contains results produced by one input condition.

Results are shown in Figure 2 and Figure 3. We use (1 - Self-BLEU) as the diversity indicator in the curve so that higher values denote better diversity. Generated captions from different beams in beam search show little variance in diversity while the accuracy slightly degrades from the first beam to the last beam. Output captions from diverse beam search are much more diverse than other approaches. However, better diversity is achieved with the sacrifice of accuracy. When generating captions for the current group, it penalizes words which have appeared in previous groups. However, these words are often more accurate. This indicates that previous

Table 1. Detailed accuracy and diversity results of the proposed approach with different input c on Clotho. "# Vocabulary" denotes the output vocabulary size. For all metrics except Self-BLEU, higher is better.

c	Accuracy Metrics					Diversity Metrics			
	$BLEU_4$	$ROUGE_L$	METEOR	CIDEr	SPICE	# Vocabulary	*distinct-1*	*distinct-2*	Self-BLEU
0.0	16.5	38.2	17.1	37.0	11.6	276	.026	.07	.93
0.2	16.8	38.2	17.2	37.8	11.7	302	.028	.078	.92
0.4	16.8	38.4	17.2	39.0	12.0	335	.031	.084	.91
0.6	16.4	38.1	17.1	38.6	11.9	383	.035	.095	.89
0.8	15.9	37.7	17.2	38.4	12.1	437	.039	.11	.87
1.0	15.4	37.5	17.1	37.9	12.1	466	.041	.12	.86

Table 2. Examples of neural conditional captioning model generated captions with different input c.

filename	*Clatter.wav*	*Deutz-Tractor-Engine-1972.wav*
c	Generated Caption	
0.0	a hard object is being hit on a hard surface	a motor is running and then the engine revs
0.4	a person is hammering on a wooden door	a diesel engine is idling and then the engine revs
0.8	a wooden object is being hit against a hard surface	a large diesel engine is idling and then the gears
1.0	someone is walking on a wooden door with a windshield	a large diesel engine is idling and the engine is being revved up

works on promoting the output instance-diversity cannot generate accurate captions with the increase of the output set-diversity.

Captions generated by the statistical conditional model perform the worst in both accuracy and diversity. We find captions generated with high statistical conditions mostly contain repeated meaningless words (e.g., "and", "is", "a"), resulting in low accuracy and diversity. In contrast, our proposed neural conditional captioning model is able to control the output diversity without sacrificing accuracy. On Clotho, the diversity variance of neural conditional outputs is similar to that of diverse beam search while the captioning accuracy almost keeps unchanged. On Audiocaps, the captioning accuracy slightly drops when $c = 1.0$ but the degradation is still the least when achieving the same diversity score. Clotho contains five reference captions (possibly with different styles) for each training audio clip while Audiocaps contains one. This may lead to it being more difficult for models to learn the mapping from c to output styles on Audiocaps.

4.2. Accuracy-diversity Trade-off

To analyze the effect of different input c on the generated captions, we list the detailed accuracy and diversity results on Clotho with different input c in Table 1. All diversity metrics improve with higher c. Although the variance of the accuracy metrics is slight, we can find the trend that a higher c results in lower BLEU, $ROUGE_L$ and METEOR but higher SPICE. It indicates that captions generated by higher c have fewer n-gram overlaps with the reference but the semantic level accuracy (objects, attributes and relationships) is better. With a high input c, The model is encouraged to generate n-grams different from those in the reference while keeping the semantic contents the same. $c = 0.4$ achieves the highest CIDEr, indicating a moderate c results in the best TF-IDF-based similarity. Results on Audiocaps show a similar trend so they are not listed here.

To give an intuitive result, we show an example of generated captions with different input c on the same audio in Clotho evaluation split in Table 2. High c results in more detailed descriptions: "a wooden door" instead of "a hard surface" and the additional object "windshield". Therefore the output set-diversity improves. In contrast, low c leads to generic descriptions which are probable to be correct but contain less information.

5. CONCLUSION

This paper aims to control the audio captioning output set-diversity. A neural conditional captioning model is proposed to generate generic or specific outputs with different condition signal c. During training, the reference condition is provided by a neural discriminator, which is trained with the captioning model in an adversarial way. Experiments on Clotho and Audiocaps show that the proposed approach can generate captions with different set-diversity. Compared with beam search, diverse beam search and statistical condition, the neural conditional approach has the least influence on accuracy to generate outputs with the same diversity. Detailed accuracy metrics also show that captions with a higher input c have fewer n-gram overlaps with the reference but contain correct audio contents.

6. ACKNOWLEDGEMENTS

This work has been supported by National Natural Science Foundation of China (No.61901265), State Key Laboratory of Media Convergence Production Technology and Systems Project (No.SKLMCPTS2020003) and Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102). Experiments have been carried out on the PI supercomputer at Shanghai Jiao Tong University.

7. REFERENCES

[1] 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.

[2] K. Chen, Y. Wu, Z. Wang, X. Zhang, F. Nian, S. Li, and X. Shao, "Audio captioning based on transformer and pre-trained cnn," in *Proceedings of the Detection and Classifica-*

tion of Acoustic Scenes and Events Workshop (DCASE), Tokyo, Japan, November 2020, pp. 21–25.

[3] Y. Koizumi, Y. Ohishi, D. Niizumi, D. Takeuchi, and M. Yasuda, "Audio captioning using pre-trained large-scale language model guided by audio-based similar caption retrieval," *arXiv preprint arXiv:2012.07331*, 2020.

[4] Y. Koizumi, R. Masumura, K. Nishida, M. Yasuda, and S. Saito, "A Transformer-based Audio Captioning Model with Keyword Estimation," in *Proceedings of Conference of the International Speech Communication Association*, 2020, pp. 2–6.

[5] A. Ö. Eren and M. Sert, "Audio Captioning Based on Combined Audio and Semantic Embeddings," in *Proceedings of IEEE International Symposium on Multimedia (ISM)*, 2020.

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

[7] J. Li, M. Galley, C. Brockett, J. Gao, and W. B. Dolan, "A diversity-promoting objective function for neural conversation models," in *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2016, pp. 110–119.

[8] L. Wang, A. G. Schwing, and S. Lazebnik, "Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space," in *Proceedings of Conference on Neural Information Processing Systems (NIPS)*, 2017, pp. 5758–5768.

[9] 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.

[10] M. Cornia, L. Baraldi, and R. Cucchiara, "Show, control and tell: A framework for generating controllable and grounded captions," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8307–8316.

[11] K. Yu, Z. Zhao, X. Wu, H. Lin, and X. Liu, "Rich short text conversation using semantic-key-controlled sequence generation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 26, no. 8, pp. 1359–1368, 2018.

[12] B. Dai, S. Fidler, R. Urtasun, and D. Lin, "Towards diverse and natural image descriptions via a conditional gan," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2970–2979.

[13] R. Shetty, M. Rohrbach, L. Anne Hendricks, M. Fritz, and B. Schiele, "Speaking the same language: Matching machine to human captions by adversarial training," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 4135–4144.

[14] S. Ikawa and K. Kashino, "Neural audio captioning based on conditional sequence-to-sequence model," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, New York University, NY, USA, October 2019, pp. 99–103.

[15] 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.

[16] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

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

[18] C. D. Kim, B. Kim, H. Lee, and G. Kim, "Audiocaps: Generating captions for audios in the wild," in *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019, pp. 119–132.

[19] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. G. Wilson, "Averaging weights leads to wider optima and better generalization," *arXiv preprint arXiv:1803.05407*, 2018.

[20] P. Kishore, R. Salim, W. Todd, and Z. Wei-Jing, "BLEU: a Method for Automatic Evaluation of Machine Translation," in *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002, pp. 311–318.

[21] L. Chin-Yew, "Rouge: A package for automatic evaluation of summaries," in *Proceedings of the workshop on text summarization branches out*, no. 1, 2004, pp. 25–26.

[22] A. Lavie and A. Agarwal, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," in *Proceedings of the Second Workshop on Statistical Machine Translation*, no. June, 2007, pp. 228–23.

[23] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 07-12-June, 2015, pp. 4566–4575.

[24] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: Semantic propositional image caption evaluation," in *Proceedings of European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 382–398.

[25] Y. Zhu, S. Lu, L. Zheng, J. Guo, W. Zhang, J. Wang, and Y. Yu, "Texygen: A benchmarking platform for text generation models," in *International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018, pp. 1097–1100.