CHATGPT CAPTION PARAPHRASING AND FENSE-BASED CAPTION FILTERING FOR AUTOMATED AUDIO CAPTIONING

Technical Report

Inhan Choi, Hyeonuk Nam, Deokki Min, Seung-Deok Choi, Yong-Hwa Park

Korea Advanced Institute of Science and Technology
Department of Mechanical Engineering, 291, Daehak-ro
Yuseong-gu, Daejeon 34141, South Korea
{ds5amk, frednam, minducky, haroldchoi6, yhpark}@kaist.ac.kr

ABSTRACT

This paper presents an automated audio captioning (AAC) model developed for DCASE2024 Challenge Task 6. To address the scarcity of audio captioning datasets, we generated paraphrases of captions from the Clotho dataset as a data augmentation strategy utilizing ChatGPT. To ensure the selection of paraphrases with high semantic relevance, we filtered the captions with high FENSE scores, the metric adopted for this AAC task. By integrating ChatGPT paraphrasing and FENSE-based caption filtering to the AAC baseline model, our submitted model achieves a 0.521 FENSE score, outperforming the baseline with a FENSE score of 0.504.

Index Terms— Automated Audio Captioning, ChatGPT paraphrasing, FENSE-based caption filtering

1. INTRODUCTION

Automated audio captioning (AAC) is a task that aims to generate text describing input audio data. Machine learning systems require extensive high-quality datasets, but AAC field still faces scarcity of data. Unlike image captioning which leverages visual information, characteristics of audio captioning datasets vary significantly depending on the collection methods. For instance, AudioCaps enhanced caption accuracy by providing word hints and reference video clips to the annotators [1]. In contrast, annotators for Clotho had to solely rely on audio data [2]. In addition to such variations in data annotation methods, factors such as differences in audio data length, caption length, and writing style create differences between datasets and hinder effective model training. Consequently, many AAC models have employed a strategy of pre-training with several datasets, followed by fine-tuning on the target dataset [3, 4, 5]. Additionally, Labbé et al. proposed task embedding tokens to mitigate domain differences depending on the dataset [6].

To address performance limitation due to dataset scarcity and domain difference, we leveraged a large language model (LLM), ChatGPT [7], to augment the Clotho dataset by generating additional captions [8]. Additionally, the metric of the AAC task in DCASE2024 has been changed from SPIDEr-FL to FENSE [9, 10, 11], which utilizes Sentence-BERT for sentence similarity assessment [12]. Thus, we propose FENSE-based caption filtering, where we use FENSE to evaluate the paraphrases' similarity with the original captions and evaluate the original captions as well. Consequently, the paraphrases generated by ChatGPT showed high similarity with original captions in terms of FENSE, thereby preventing overfitting and enhancing model performance by increasing expression diversity.

2. PROPOSED METHOD

2.1. ChatGPT Paraphrasing

Current AAC datasets are challenging to be utilized effectively together due to domain differences [6]. AAC research suffers from a lack of data, necessitating the use of various data augmentation methods [13, 14]. To address the issue of insufficient high-quality captions, Cho et al. employed synonym substitution to replace specific words in captions with their synonyms and won second place in DCASE2023 Challenge Task 6a [15, 16]. Additionally, recent advancements in high-performance LLM have further facilitated data augmentation strategies. The WavCaps dataset generated over 400k audio captions using ChatGPT [17]. The first-place entry in DCASE2023 Task 6a utilized the LLM for caption mixup augmentation [8, 18].

To prevent model overfitting and enrich expression, we paraphrased captions for data augmentation purposes. We prompted ChatGPT to generate five paraphrases for each caption. This process generated a total of 19195 texts for 3839 audio files in the Clotho development set, in addition to the original texts.

2.2. FENSE-based Caption Filtering

To filter paraphrases with high semantic similarity, we employed FENSE, the metrics of AAC task. FENSE utilizes Sentence-BERT to generate semantically meaningful sentence embeddings and includes a fluency error detector [11, 12]. Therefore, we used FENSE to filter the captions to enhance the performance of ChatGPT paraphrasing. The average FENSE score for the paraphrases is 0.584, higher than the cross-referencing score of 0.574. We measured the cross-referencing score by using every reference

Table 1. Average FENSE score of paraphrase sets

Paraphrase set	**FENSE**
$\mathbb{P}_{sorted,1}$	0.659
$\mathbb{P}_{sorted,2}$	0.628
$\mathbb{P}_{sorted,3}$	0.597
$\mathbb{P}_{sorted,4}$	0.558
$\mathbb{P}_{sorted,5}$	0.476
mean	0.584

caption as a candidate for reference captions and averaging the results [6]. This indicates that the paraphrases generated by ChatGPT are of sufficiently high quality.

To investigate highly semantically similar paraphrase sets and the impact of paraphrase sets with varying FENSE averages on the AAC model, we compared the paraphrase set $\mathbb{P} = \cup_{i=1}^{N} \mathbb{P}^i$, $\mathbb{P}^i = \{p_1^i, p_2^i, p_3^i, p_4^i, p_5^i\}$ (p_m^i : m-th paraphrase in the i-th audio file, N: total number of audio files) for each of the audio files with the caption set $\mathbb{C} = \cup_{i=1}^{N} \mathbb{C}^i, \mathbb{C}^i = \{c_1^i, c_2^i, c_3^i, c_4^i, c_5^i\}$ (c_m^i : m-th caption in the i-th audio file), organizing the paraphrases in descending order of their FENSE scores as:

$$\mathbb{P}^i = \{p_1^i, p_2^i, p_3^i, p_4^i, p_5^i\} \quad (1)$$
$$s.t.\, f(p_1^i) \geq f(p_2^i) > \cdots > f(p_5^i)$$
$$where\ f(p_l^i) = FENSE(p_l^i | c_1^i, c_2^i, \ldots, c_5^i)$$

$\mathbb{P}$ are then grouped into five sets, $\mathbb{P}_{sorted,m} = \{p_m^i\}_{i=1}^{N}, m = 1,2,\ldots,5$. Data augmentation is performed by adding each sorted set to the original captions. The Average FENSE scores for $\mathbb{P}_{sorted,m}$ are presented in Table 1, showing that the scores for $m = 1,2$ and 3 exceed the cross-referencing score. We trained a model on the set $\mathbb{C}_j^+$, which is constructed by incrementally incorporating the sorted paraphrase sets $\mathbb{P}_{sorted,m}$ as follows:

$$\mathbb{C}_j^+ = \mathbb{C} \cup \{\cup_{m=1}^{j} \mathbb{P}_{sorted,m}\}, j = 1,2,\ldots,5 \quad (2)$$

Each $\mathbb{C}_j^+$ includes the initial set $\mathbb{C}$ and the union of the first j sets from the sorted collection $\mathbb{P}_{sorted,m}$. Therefore, the model trains $j + 5$ captions for each audio file on the $\mathbb{C}_j^+$

Furthermore, we created a modified caption set $\mathbb{C}'$ from the original set $\mathbb{C}$ for consistency with each audio file's captions. Clotho captions are generated exclusively using the corresponding audio files. As a result, numerous captions reflect varying interpretations of the same audio file by different annotators. To enhance the consistency of captions, we compared the caption with the lowest semantic similarity to the highest FENSE scoring paraphrase. The detailed process is as follows:

$$\mathbb{C}' = \cup_{i=1}^{N} \mathbb{C}_i' \quad (3)$$
$$s.t.\, \mathbb{C}_i' = \begin{cases} \mathbb{C}_i - \{c_{min}^i\} + \{p_1^i\} & if\ g(c_{min}^i) < f(p_1^i) \\ \mathbb{C}_i & otherwise \end{cases}$$
$$where\ c_{min}^i = \text{argmin}_{c^i \in \mathbb{C}^i} g(c^i)\ and\ g(c_m^i) = FENSE(c_m^i | \mathbb{C}^i)$$

Table 2. Comparison of original and modified caption set using FENSE (cross-referencing) and vocabulary size.

	FENSE (cross-referencing)	Vocabulary size
Original set($\mathbb{C}$)	0.574	4369
Modified set($\mathbb{C}'$)	0.668	5672

To give additional weight to the original captions, we included each caption itself when calculating the FENSE of c_m^i.The caption set $\mathbb{C}'$ has a cross-referencing score of 0.668 and a vocabulary size of 5672, indicating higher semantic consistency and a more diverse use of vocabulary compared to $\mathbb{C}$ as shown in Table 2. Additionally, we created and tested specialized dataset, $\mathbb{C}' + \mathbb{P}_{sorted,1}$.

3. EXPERIMENTAL SETTINGS

3.1. Model architecture

Our model builds upon a sequence-to-sequence baseline system [6]. The encoder employs ConvNeXt [19], a fully convolutional neural network pre-trained on AudioSet with frozen parameters. The decoder is a transformer decoder, as in the baseline. For the decoding algorithm, we utilized beam search with a beam size of three or four.

3.2. Data Augmentations

We implemented three data augmentation methods in our model: mixup with a parameter of 0.4, label smoothing [20] with a parameter of 0.2, and ChatGPT paraphrasing. We trained the model on various caption-augmented datasets $\mathbb{C}_1^+, \ldots, \mathbb{C}_5^+, \mathbb{C}'$, and $\mathbb{C}' + \mathbb{P}_{sorted,1}$ to find the optimized caption set for the AAC model.

3.3. Implementation Details

We implemented the model training with different hyperparameter settings. For the ablation study, we utilized the baseline set (BS) with a batch size of 64, epochs of 400, a beam size of 3, and validation loss as the monitor metric. Set1 used a batch size of 128, epochs of 500, a beam size of 4, and train loss as the monitor metric. Set2_num had a batch size of num, epochs of 600, a beam size of 3, and train loss as the monitor metric. And all other hyperparameters were consistent with the baseline.

4. RESULTS

To proceed with the ablation study of the $\mathbb{C}^+$ and $\mathbb{C}'$ sets, we first tested it using the same hyperparameter settings as the baseline [16]. The results, averaged over five experiments on the development-evaluation split of Clotho, are presented in Table 3. The findings indicate that incorporating paraphrases increased the FENSE scores without significant changes in other metrics. Adjusting the hyperparameter setting from BS to Set1, we observed a peak FENSE score of 0.515 and the SPIDEr-FL score of 0.313 with eight captions.

Table 3. Performance comparison of baseline, ablation study, and our submission on Clotho development-evaluation set.

	SPIDEr-FL	***FENSE***	***Vocab***
Baseline ($\mathbb{C}$)(BS)	0.296	0.504	551
Modified Set ($\mathbb{C}'$)(BS)	0.295	0.509	574.4
5C + 1P ($\mathbb{C}_1^+$)(BS)	0.293	0.505	553.6
5C + 2P ($\mathbb{C}_2^+$)(BS)	0.297	0.508	597
5C + 3P ($\mathbb{C}_3^+$)(BS)	0.302	0.513	577.2
5C + 4P ($\mathbb{C}_4^+$)(BS)	0.306	0.512	568
5C + 5P ($\mathbb{C}_5^+$)(BS)	0.301	0.509	565.6
5C + 3P ($\mathbb{C}_3^+$) (Set1)	**0.313**	0.515	585
$\mathbb{C}' + \mathbb{P}_{sorted,1}$ (Set2_256) (submission1)	0.299	**0.521**	623
$\mathbb{C}' + \mathbb{P}_{sorted,1}$ (Set2_128)	0.299	0.513	738
$\mathbb{C}' + \mathbb{P}_{sorted,1}$ (Set2_64) (submission2)	0.272	0.515	**829**

Additionally, we trained using $\mathbb{C}' + \mathbb{P}_{sorted,1}$. In submission 1, with Set2_256, FENSE score of 0.521 is achieved, which is the highest score. When we changed the batch size within the set, there were significant changes in the vocabulary size. In submission 2, when the batch size was adjusted to 64, the vocabulary size reached 829.

5. CONCLUSION

In this report, we evaluated the impact of ChatGPT-generated paraphrases on AAC model performance. We combined paraphrases with the original caption set for training and examined the effect of excluding captions with low semantic similarity to enhance consistency through FENSE-based filtering. Most AAC models trained on augmented-caption sets achieved higher FENSE scores than the baseline, with one model outperforming the baseline by 3.4% under optimized hyperparameters. In future research, we plan to utilize LLM to generate a larger variety of paraphrases, thereby creating caption sets with diverse qualities. We will also explore the correlation between caption quality and model performance.

6. REFERENCES

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

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

[3] T. Schaumlöffel, M. G. Vilas, and G. Roig, "PEACS: Prefix encoding for auditory caption synthesis", DCASE2023 Challenge, Tech. Rep., 2023.

[4] H. Sun, Z. Yan, Y. Wang, H. Dinkel, and J. Zhang, "Leveraging multi-task training and image retrieval with CLAP for audio captioning", DCASE2023 Challenge, Tech. Rep. 2023.

[5] Q. Han, W. Yuan, D. Liu, X. Li, and Z. Yang, "Automated Audio Captioning With Weakly Supervised Pre-training and Word Selection Methods," DCASE2021 Challenge, Tech. Rep., 2021.

[6] E. Labbé, T. Pellegrini, and J. Pinquier, "CoNeTTE: An efficient Audio Captioning system leveraging multiple datasets with Task Embedding," *arXiv preprint arXiv:2309.00454*, 2023.

[7] https://openai.com/index/chatgpt/.

[8] S.-L. Wu, X. Chang; G. Wichern et al. "Improving Audio Captioning Models with Fine-Grained Audio Features, Text Embedding Supervision, and LLM Mix-Up Augmentation" in *IEEE International Conf. on Acoustics, Speech and Signal Processing (ICASSP)*,2024.

[9] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, "Improved Image Captioning via Policy Gradient optimization of SPIDEr," *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 873–881.

[10] https://dcase.community/challenge2023/task-automated-audio-captioning#evaluation.

[11] Z. Zhou, Z. Zhang, X. Xu, Z. Xie, M. Wu, and K. Q. Zhu, "Can audio captions be evaluated with image caption metrics?" in *IEEE International Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 981–985.

[12] N. Reimers and I. Gurevych, "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.

[13] E. Kim et al., "Improving audio-language learning with mixgen and multi-level test-time augmentation," *arXiv preprint arXiv:2210.17143*, 2022.

[14] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Interspeech*, 2019, pp. 2613–2617.

[15] J.-H. Cho, Y.-A. Park, J. Kim, and J.-H. Chang, "HYU submission for the DCASE 2023 task 6a: automated audio captioning model using AL-MixGen and synonyms substitution," DCASE2023 Challenge, Tech. Rep., 2023.

[16] J. Wei and K. Zou, "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks," *arXiv preprint arXiv:1901.11196*, 2019.

[17] X. Mei, C. Meng, H. Liu et al., "WavCaps: A ChatGPT-Assisted Weakly-Labelled Audio Captioning Dataset for Audio-Language Multimodal Research," *arXiv preprint arXiv:2303.17395*, 2023.

[18] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations (ICLR)*, 2018.

[19] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11 966–11 976.

[20] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.