AUTOMATED AUDIO CAPTIONING USING PARAMETER EFFICIENT FINE-TUNING AND MERGING OF LLMS

Technical Report

Eungbeom Kim[1], *Jaeheon Sim*[1], *Jinwoo Lee*[2], *Kyogu Lee*[1,2]

[1] Interdisciplinary Program in Artificial Intelligence, Seoul National University,
[2] Department of Intelligence and Information, Seoul National University

ABSTRACT

This technical report introduces an audio captioning system, which is designed to tackle the task of Automated Audio Captioning (AAC) in the Detection and Classification of Acoustic Scenes and Events (DCASE) 2024 challenge. Our approach employs BEATs for robust audio representation learning and Llama 3 for high-quality text generation. To address the limitations of small datasets like Clotho, we fix the pre-trained weights of the BEATs and train a small linear model to map audio encoder dimensions to the LLM input. We further fine-tune the LLM using parameter-efficient fine-tuning method, LoRA, to train the model. We also explore the concatenation based LoRA merging method, achieving notable results on standard benchmarks.

Experimental results show that our proposed system achieves a FENSE [1] score of 0.5180 on the evaluation dataset.

Index Terms— Automated audio captioning, LLM, LoRA, fine-tuning

1. INTRODUCTION

Automated audio captioning aims at generating text descriptions for audio recordings. The task is similar to neural machine translation (NMT) in that the audio is translated to texts. For this reason, most of AAC systems employ an encoder-decoder architecture [2, 3, 4, 5]; audio encoder and text decoder. If the audio encoder outputs audio embedding the text decoder decodes caption text from the embedding. For the audio encoder, audio models that are pre-trained on large-scale audio data are used in general. For the text decoder, Transformers [6] or traditional recurrent neural networks are used.

Recently, however, pre-trained language models (LM) show great performance at generating sentences, some works adopt a pre-trained LM for text decoder [2, 3, 4, 5]. Moreover, as LMs scale to sizes of tons of parameters, large language models (LLM) have been developed, showing unimaginable performances. Thus, there have been attempts to utilize the power of LLMs. [3] uses ChatGPT [7] at data augmentation for more natural mixed-up sentences and [8] utilizes the decoder-only LLM by entering audio token and instruction tokens to LLM decoder as input.

However, our main dataset, Clotho [9], is too small to train LLMs with billions of parameters effectively. To address this problem, we keep the pre-trained weights of the LLM fixed and train a small linear model to map the dimensions of the audio encoder to the LLM's input dimensions. We then fine-tune the LLM using parameter-efficient methods.

2. SYSTEM DESCRIPTION

2.1. BEATs

BEATs [10] is a framework for audio representation learning that leverages self-supervised learning techniques. Unlike traditional models that rely on reconstruction loss, BEATs employs a self-distilled tokenizer to convert continuous audio signals into discrete labels, enabling mask and discrete label prediction pre-training. The framework iteratively refines an acoustic tokenizer and an audio SSL model through mutual optimization. Initially, random projection is used to train the SSL model, which then distills semantic knowledge to improve the tokenizer in subsequent iterations. BEATs achieves state-of-the-art performance on various benchmarks, including a 50.6% mean Average Precision (mAP) on AudioSet-2M and 98.1% accuracy on ESC-50, showing its effectiveness in generating robust and semantically rich audio embeddings.

2.2. Llama 3

The Llama [11] series are state-of-the-art LLMs developed for superior natural language understanding and generation. These models utilize advanced Transformer decoder-based architectures to achieve high performances across various NLP tasks, such as text generation, translation, and comprehension. Llama3, in particular, builds upon the strengths of its predecessors by incorporating more sophisticated algorithms and training techniques, enabling it to generate highly accurate and contextually relevant descriptive text based on prompts. This iteration of the Llama series excels in tasks requiring detailed and nuanced language generation, making it an exceptional tool for applications that demand high-quality natural language outputs.

Our approach involves fixing the BEATs representation and then adding a new 2-layer linear structure, followed by Global Mean Pooling to produce a representation for Llama tokens. For prompting, we utilize the Llama3-7B-instruct model with the format "Describe the following audio clip [audio]". The Llama3 model is trained using LoRA with the following parameters: rank (r) set to 16, dropout rate of 0.05, and training is performed by projecting only the query (q) and key (k) vectors. The training process involves initially training on the AudioCaps dataset and subsequently on the Clotho dataset. Both datasets utilize a learning rate that peaks

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2022-0-00320, Artificial intelligence research about cross-modal dialogue modeling for one-on-one multi-modal interactions)

	BLEU1	BLEU4	METEOR	ROUGE-L	CIDEr	SPICE	SPIDEr	FENSE
Submission1	0.5764	0.1553	0.1848	0.3800	0.4426	0.1336	0.2881	0.5060
Submission2	0.5485	0.1368	0.1868	0.3715	0.4122	0.1342	0.2732	0.5180

Table 1: *Results on evaluation split of Clotho dataset.*

at 0.0003, starting with a 10% warmup period, followed by a cosine decay schedule. Each dataset is trained for 50 epochs. Subsequently, based on the Clotho dataset, we fix the audio embedding models and produce multiple trained LoRA modules with the fixed audio embedding models. Then we merge the LoRA modules based on concatenation. This involves training the LoRA on the Clotho dataset four separate times, resulting in a total of five merged models.

3. RESULTS

Our results on the evaluation split of Clotho dataset is in Table 1.

4. CONCLUSION

In this technical report, we introduced a novel approach to automated audio captioning by combining BEATs for audio representation and Llama 3 for text generation. Our method addresses the challenge of small dataset sizes by fixing pre-trained LLM weights and using a small linear model for dimensional mapping, followed by fine-tuning with parameter-efficient techniques. Experimental results on the Clotho dataset demonstrate the effectiveness of our approach, with our system achieving competitive scores across multiple evaluation metrics. These findings highlight the potential of combining advanced audio and language models for enhanced audio captioning performance.

5. REFERENCES

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

[2] 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., May 2023.

[3] S.-L. Wu, X. Chang, G. Wichern, J.-w. Jung, F. Germain, J. L. Roux, and S. Watanabe, "Beats-based audio captioning model with instructor embedding supervision and chatgpt mix-up," DCASE2023 Challenge, Tech. Rep., May 2023.

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

[5] E. Labbé, T. Pellegrini, and J. Pinquier, "Irit-ups dcase 2023 audio captioning and retrieval system," DCASE2023 Challenge, Tech. Rep., May 2023.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[7] J. Schulman, B. Zoph, C. Kim, J. Hilton, J. Menick, J. Weng, J. F. C. Uribe, L. Fedus, L. Metz, M. Pokorny, *et al.*, "Introducing chatgpt," *OpenAI Blog*, 2022.

[8] S. Deshmukh, B. Elizalde, R. Singh, and H. Wang, "Pengi: An audio language model for audio tasks," *Advances in Neural Information Processing Systems*, vol. 36, pp. 18 090–18 108, 2023.

[9] 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.

[10] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, "Beats: Audio pre-training with acoustic tokenizers," *arXiv preprint arXiv:2212.09058*, 2022.

[11] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.