DCASE2024 CHALLENGE - TASK 6
HYPERPARAMETER TUNING OF THE CONETTE AUDIO CAPTIONING SYSTEM

Technical Report

Jakob De Jesus Silva, Justus Tobias, Sebastian Sonderegger

Johannes Kepler University, Linz, Austria

ABSTRACT

In the course of this challenge, we explored various methods to achieve a state-of-the-art audio captioning model. Initially, we worked with the baseline provided by the challenge organizers, then we also constructed several models from scratch, using diverse architectures. The best outcome we could achieve, was by tuning the hyperparameters of the baseline model CoNeTTE[1]. Our systematic approach involved finding hyperparameters that had the most effect on performance and their best combination. Although our enhanced baseline model demonstrated some performance gains, it still did not achieve a significant breakthrough over the original baseline. This is a student project in course of the lecture "Machine-learning and Audio: A Challenge" at JKU.

***Index Terms*—** Automated Audio Captioning, CoNeTTE, ConvNext, Transformer, DCASE2024, Task 6, ML Challenge

1. INTRODUCTION

The DCASE Challenge is a yearly held challenge in the field of Machine Listening, consisting of 10 different tasks where researchers can compete in. We chose to compete in Task 6, Automated Audio Captioning (AAC), which is about generating a short descriptive sentence, called caption, from raw audio with the help of a machine learning model. Recent approaches, like the baseline architecture we used, employ encoder-decoder architectures for this. The raw audio is first converted to log-mel-spectrograms in the preprocessing, and these spectrograms then serve as inputs to Image-to-Text models. ConvNeXt[2] is a prominent Transformer[3] architecture used in image-to-text prediction, and has been adapted to audio in the baseline model of the challenge [4]. With our own models not quite reaching the performance of this baseline, we opted for hyperparameter-tuning instead. Results are summarized in Chapter 5.

2. BASELINE

The baseline system, CoNeTTE[1], is a state-of-the-art Audio Captioning model that utilizes a pre-trained and frozen ConvNeXt-tiny[2][4] architecture (29.6M parameters) as audio encoder, combined with a Transformer decoder (11.9M parameters). The ConvNeXt architecture, originally adapted from the vision domain, provides robust audio embeddings that enhance the caption generation process. Additionally, CoNeTTe leverages Task Embedding tokens, to handle training and inference on multiple datasets while mitigating their biases. This approach enabled the model to compete with models up to forty times it's size, by identifying the source dataset for each input sample, thereby reducing performance gaps and ensuring a more generalized and robust audio captioning capability. An overview of the Baseline architecture is presented in Figure1.

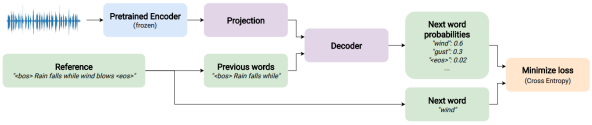

Figure 1: Overview of Baseline AAC training process [1]

3. DATA

Clotho [5] is a high-quality audio captioning dataset. It contains around 7K audio snippets from the Freesound[6] portal, and each snippet was manually labeled by five human annotators. Captions are between 8 and 20 words long, and each snippet is not longer than 30 seconds. Among other widely used AAC datasets like WavCaps[7] and AudioCaps [8] the upside of using Clotho for challenges is its reasonable size, making it accessible, even with smaller computational resources. Additionally it provides dedicated subset-splits for training, evaluation and analysis as well as a non-public testing-split, to evaluate challenge submissions. An example datapoint is illustrated in Table 1.

Filename	Captions
Clock Ticking.wav	A machine is making a loud clicking noise. A pendulum moves back and forth and a pendulum ticks. A pendulum ticks as it moves back and forth. The machine is making a loud clicking noise. the ticktocking of a loud clock in the foreground

Table 1: Example data from the Clotho Dataset

4. EVALUATION

FENSE [9], which stands for Fluency Enhanced Sentence-BERT Evaluation, is a novel metric designed to address the limitations of existing image captioning metrics when applied to audio captions. The metric combines the strengths of Sentence-BERT for capturing semantic similarity and an Error Detector to penalize fluency issues. This dual approach ensures that FENSE can effectively evaluate the quality of audio captions by not only measuring semantic relevance

but also addressing common errors like incomplete sentences, repeated events, and missing conjunctions.

5. EXPERIMENTS

We did a set of experiments to learn the effects of the most commonly tweaked hyperparameters individually and then narrowed down our selection to the three most impactful, namely learning rate, batch size and dimension of hidden layers in the decoder model. Then we did a small grid search to find their best combination. Using larger hidden-layers (512 instead of 256) required to train the model with a bigger batch size, which makes sense intuitively, as more complex structures can be learned, but also the model can be more prone to overfitting, while larger batch sizes can counter this. The increased size of the hidden layers increased the model size significantly, from 11.9M to 30.1M trainable parameters. With the found optimal hyperparameters we then not only trained the model on Clotho's development-train split (Baseline+_1), but also on the combined development-train and development-validation split (Baseline+_2). This resulted in another increase in the final FENSE-score, when evaluated on the Clotho development-evaluation split. (see Table 2).

	Training set	**lr**	**bs**	**h_dim**	**FENSE**
Baseline	cl-dev-train	5e-4	32	256	0.5040
Baseline+_1	cl-dev-train	4e-4	64	512	0.5059
Baseline+_2	cl-dev-train cl-dev-val	4e-4	64	512	**0.5079**

Table 2: Final evaluation scores of our best models

6. CONCLUSIONS

As the baseline system is already a very sophisticated architecture, our hyperparameter tuning resulted only in minor improvements. To achieve further enhancement, we were planning to incorporate additional data from AudioCaps [8], for which we had already generated enhanced captions in the style of Clotho. For this we used a GPT-3.5-turbo model [10], finetuned through the Open-AI API [11] with Clotho captions. Unfortunately, we have not yet succeed in integrating our own data into the complex data preparation framework of the baseline system.

7. ACKNOWLEDGMENTS

This project was realized in the course of the lecture "Machine Learning and Audio: A Challenge" at JKU Linz, thus we want to give special thanks to Professor Gerhard Widmer's team at the Institute of Computational Perception, in particular, we would like to mention Florian Schmid and Paul Primus and thank them for the inspiring lecture and guidance.

8. REFERENCES

[1] E. Labbé, T. Pellegrini, and J. Pinquier, "CoNeTTE: An efficient Audio Captioning system leveraging multiple datasets with Task Embedding," *CoRR*, vol. abs/2309.00454, 2023.

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

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 2017, pp. 5998–6008.

[4] T. Pellegrini, I. K. Hassani, E. Labbé, and T. Masquelier, "Adapting a ConvNeXt model to audio classification on AudioSet," *CoRR*, vol. abs/2306.00830, 2023.

[5] 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)*, 2020, pp. 736–740.

[6] Freesound, "Freesound: Collaborative Database of Creative Commons Licensed Sounds," 2024, accessed: 2024-06-10. [Online]. Available: https://freesound.org

[7] X. Mei, C. Meng, H. Liu, Q. Kong, T. Ko, C. Zhao, M. D. Plumbley, Y. Zou, and W. Wang, "Wavcaps: A Chatgpt-Assisted Weakly-Labelled Audio Captioning Dataset for Audio-Language Multimodal Research," *CoRR*, vol. abs/2303.17395, 2023.

[8] 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.

[9] 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)*, 2022, pp. 981–985.

[10] J. Ye, X. Chen, N. Xu, C. Zu, Z. Shao, S. Liu, Y. Cui, Z. Zhou, C. Gong, Y. Shen, J. Zhou, S. Chen, T. Gui, Q. Zhang, and X. Huang, "A Comprehensive Capability Analysis of GPT-3 and GPT-3.5 Series models," *CoRR*, vol. abs/2303.10420, 2023.

[11] OpenAI, "OpenAI API," 2024, accessed: 2024-06-10. [Online]. Available: https://www.openai.com/api/