DCASE 2024 TASK6: AUTOMATED AUDIO CAPTIONING USING CONTRASTIVE LEARNING

Technical Report

Dan Epshtein, Yuval Amsalem, Alon Amar

Acoustics Research Center
Israel
Danep95@gmail.com

ABSTRACT

This technical report presents our proposed enhancements to improving the baseline results of the DCASE2024 challenge Task 6 on Automated Audio Captioning. We introduce an additional loss function for contrastive learning, incorporating the NTXent loss as proposed in [1][3] into the baseline platform.

Index Terms**—** Automated Audio Caption, Contrastive Learning, NTXent Loss

1. INTRODUCTION

Automated Audio Captioning (AAC) is the task of generating descriptive text for general audio content. Unlike speech-to-text systems, which transcribe spoken language, AAC is an inter-modal translation task where the input is an audio signal and the output is a textual description or caption of that signal.

AAC faces several primary obstacles. One significant challenge is the insufficient amount of available data. Unlike speech, non-speech sounds encompass a much larger variety of categories, making it difficult to gather comprehensive datasets. Additionally, different sources can produce similar sounds, complicating the task of distinguishing between them and generating accurate captions.

In this report, we propose enhancements to improve the baseline results of the DCASE2024 Challenge Task 6 on AAC. Our approach involves introducing an additional loss function for contrastive learning. Specifically, we incorporate the NTXent loss, as proposed in [1][3], into the existing baseline platform. This integration aims to enhance the model's ability to distinguish between different audio signals and generate more accurate and meaningful captions.

2. CONTRASTIVE LEARNING (CL)

As noted in [3], several Self-Supervised Learning (SSL) methods in AAC leverage Contrastive learning. One approach utilizes the audio-text representation from the decoder output to predict associations between audio and caption pairs. Another method focuses on constructing a proxy feature space where embeddings of captions from the same audio clip are encouraged to be closer together, while embeddings from different audio clips are pushed further apart.

SSL enhances performance by incorporating additional information about the matching or mismatching of audio-text pairs compared to training solely with captions as labels. This approach effectively leverages the contextual relationship between audio and text to improve overall performance.

Throughout this task, we explored both methods, ultimately opting to utilize the second due to its notably superior outcomes.

2.1. Integrate CL into the baseline system

As previously mentioned, we worked within the same framework as the baseline system, with the only change being the loss function. In the baseline framework, the loss function includes only the cross-entropy loss (L_{ce}) between the decoder output and the caption embeddings. We integrated the CL loss and added it to the original cross-entropy loss.

$$L = \mathrm{E}_{x,x^+,x^k}[-\log(\frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + \sum_{k=1}^{K} e^{f(x)^T f(x^k)}})] \quad (1)$$

Where x^+ is similar to x and f is an decoder.

(1) Describe NTXent Loss, also known as InfoNCE, mentioned in [2] and used as our CL loss (L_{cl}). The total loss is calculated as shown in (2)

$$L_{tot} = L_{ce} + 0.1 \cdot L_{cl} \quad (2)$$

Another approach we used was to initially train the model solely with cross-entropy loss for the first 100 epochs, and then incorporate the NTXent loss for 300 last epochs.

3. RESULTS

Metric	Baseline + CL
Meteor	0.1892
Cider	0.4732
Spice	0.1346
Spider	0.3039
Spider_fl	0.3016
fense	0.5035
vocabulary	578.000

Table 1: Development dataset result

4. CONCLUSION

We propose enhancements based on CL to improve the baseline results of the DCASE2024 Challenge Task 6. Compared to the baseline results on the development dataset, there is a minor improvement in some metrics.

5. REFERENCES

[1] Liu, X., Huang, Q., Mei, X., Ko, T., Tang, H. L., Plumbley, M. D., & Wang, W. (2021). CL4AC: A Contrastive Loss for Audio Captioning. *Detection and Classification of Acoustic Scenes and Events 2021, 15–19 November 2021, Online*. arXiv:2107.09990. Retrieved from https://arxiv.org/pdf/2107.09990

[2] Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., & Tang, J. (2021). Self-supervised Learning: Generative or Contrastive. *arXiv*. https://arxiv.org/pdf/2006.08218v5.pdf

[3] Xu, X., Xie, Z., Wu, M., & Yu, K. (2023). Beyond the Status Quo: A Contemporary Survey of Advances and Challenges in Audio Captioning. *arXiv*. https://arxiv.org/pdf/2205.05357v2.pdf