CL4AC: A Contrastive Loss for Audio Captioning

Paper ID: 65

Xubo Liu[1], Qiushi Huang[1,2*], Xinhao Mei[1], Tom Ko[2], H Lilian Tang[1], Mark D. Plumbley[1] and Wenwu Wang[1]*

[1]University of Surrey [2]Southern University of Science and Technology

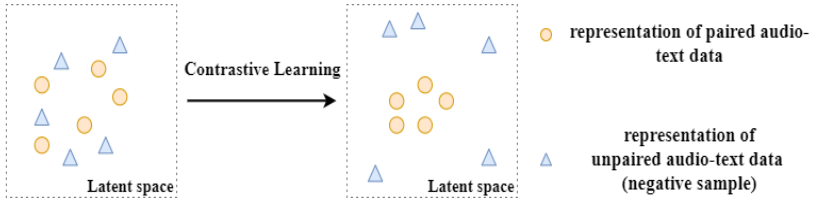
* The first two authors contributed equally to this work

Introduction

- Automated Audio Captioning (AAC) – using natural language to describe the content of an audio clip.
- Training of an AAC model often encounter the problem of data scarcity, which may lead to inaccurate representation between audio and texts.
- To address this issue, we propose a novel encoder-decoder framework: **C**ontrastive **L**oss for **A**udio **C**aptioning (CL4AC).
- In CL4AC, the self-supervision signals derived from the original audio-text paired data are used to exploit the audio-text correspondences by contrasting samples.
- CL4AC can improve the quality of latent representation and the alignment between audio and texts, while trained with limited data.

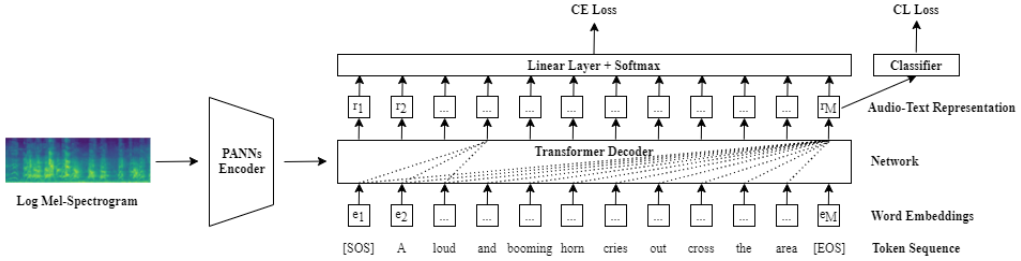
Contrastive Learning for AAC

- The audio-text representation of paired data are pulled together in the latent space while simultaneously pushing apart clusters of unpaired negative data by Contrastive Learning.



Proposed Framework: CL4AC

- PANNs encoder and Transformer decoder (state-of-the-art).
- Contrastive Loss (CL) is designed to maximize the difference between the audio-text representation of matched audio-caption pair derived from the negative pairs.
- Training objective: y=1 is negative pair, y=0 is matched pair.



$$\text{Loss}_{\text{Training}} = (1-y)\,\text{Loss}_{\text{CE}} + \text{Loss}_{\text{CL}}\,.$$

Experiment and Result

- Dataset: Clotho V2, same as that used for DCASE 2021 Challenge Task 6.
- Baseline: Our DCASE Challenge system without using reinforcement learning and transferring learning.
- Captions are generated using greedy search, evaluation metrics are same as those used for DCASE Challenge Task 6.

Model	BLEU_1	BLEU_2	BLEU_3	BLEU_4	ROUGE_L	METERO	CIDEr	SPICE	SPIDEr
Baseline	0.550	0.345	0.222	0.139	0.372	0.169	0.356	0.115	0.235
CL4AC	0.553	0.349	0.226	0.143	0.374	0.168	0.368	0.115	0.242

Conclusion

- We demonstrated the first attempt using contrastive learning for audio captioning – CL4AC.
- CL4AC can mitigate the data scarcity problem for AAC without introducing large-scale external data.
- Experimental results proved that CL4AC can improve the performance of a baseline that is already strong, while training with limited data.