Paper ID: 67

An Encoder-Decoder Based Audio Captioning System with Transfer and Reinforcement Learning

Xinhao Mei[1], Qiushi Huang[1,4], Xubo Liu[1], Gengyun Chen[2], Jingqian Wu[3], Yusong Wu[3], Jinzheng Zhao[1], Shengchen Li[3], Tom Ko[4], H Lilian Tang[1], Xi Shao[2], Mark D. Plumbley[1], Wenwu Wang[1]

[1]University of Surrey, Guildford, United Kingdom, [2]Nanjing University of Posts and Telecommunications, Nanjing, China

[3]Xi'an Jiaotong-Liverpool University, Suzhou, China, [4]Southern University of Science and Technology, Shenzhen, China

Introduction:

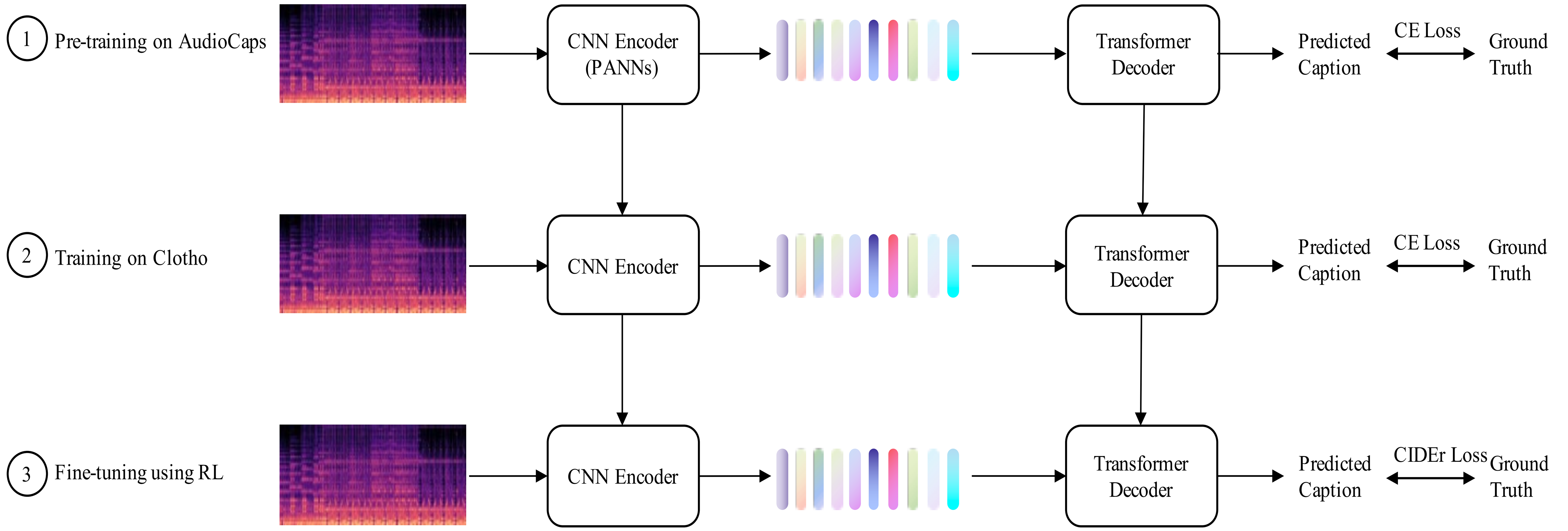
- Automated audio captioning aims to use natural language to describe the content of audio data.
- Existing audio captioning systems almost follow an encoder-decoder architecture, where the decoder predicts words based on audio features extracted by the encoder.
- In this work, we propose to use transfer learning and reinforcement learning to improve an audio captioning system.
- The resulting system was ranked 3rd in DCASE 2021 Task 6, and it was the best system without using ensemble technique.
- Reinforcement learning may impact adversely on the quality of the generated captions.

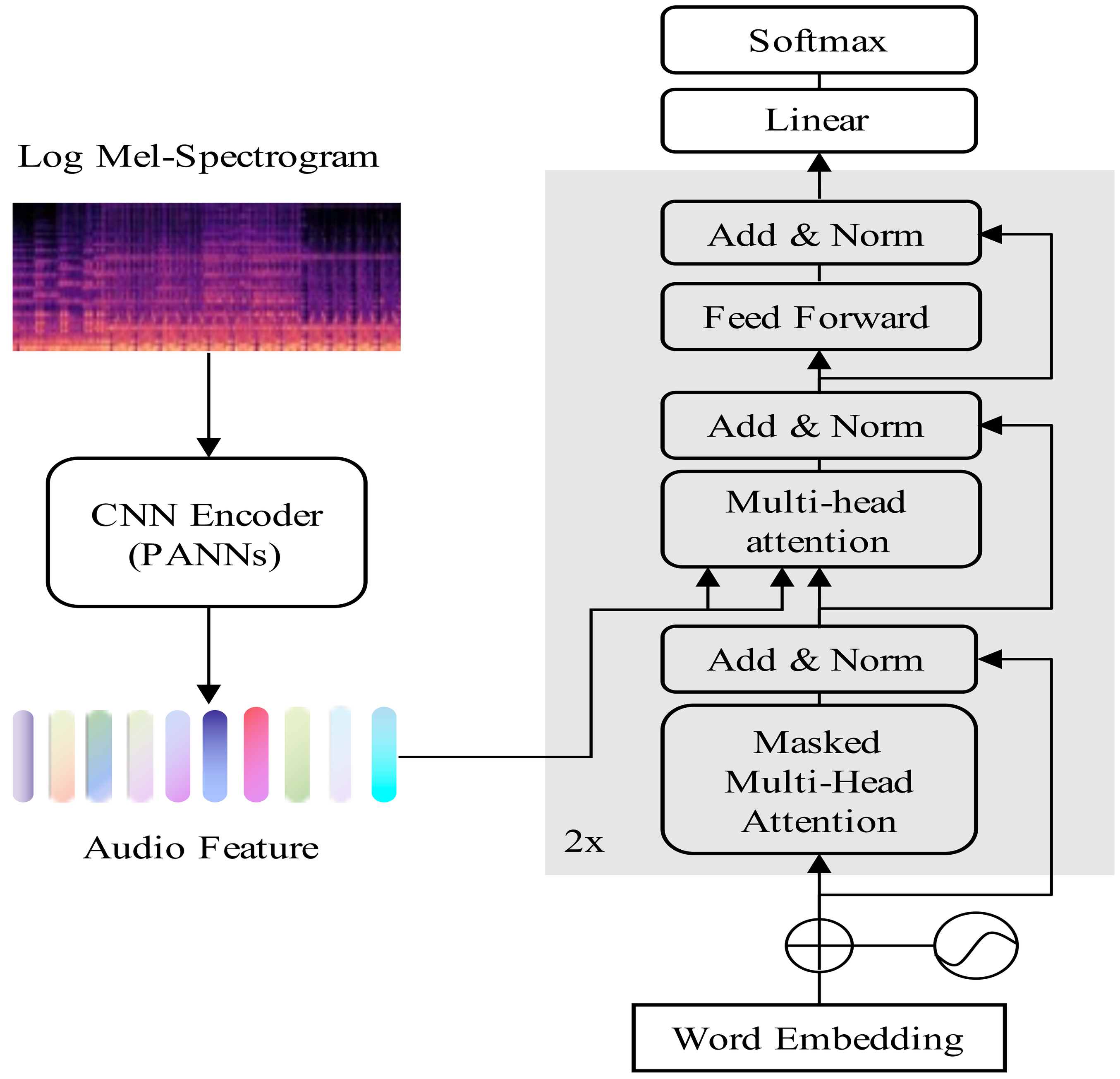
Issues:

- The official dataset Clotho is limited, which just contains 5929 audio clips.
- Maximum likelihood training introduces 'exposure bias'.
- Training objective mismatches with the evaluation metrics.

Model and methods:

- CNN-Transformer model is used as our baseline.
- Transfer learning is used to solve the data scarcity problem.
  - Transferring from upstream task (PANNs)
  - Transferring from a large in-domain dataset (AudioCaps)
- Reinforcement learning is used to address the 'exposure bias' problem and directly optimize the evaluation metric CIDEr.





Results:

Model	$BLEU_1$	$BLEU_2$	$BLEU_3$	$BLEU_4$	$ROUGE_L$	METERO	CIDEr	SPICE	SPIDEr
Baseline	0.525	0.344	0.237	0.163	0.359	0.154	0.352	0.100	0.226
B+PANNs	0.564	0.375	0.255	0.171	0.383	0.172	0.421	0.120	0.270
B+PANNs+AC	0.561	0.374	0.257	0.174	0.379	0.171	0.426	0.124	0.275
B+PANNs+RL	0.639	0.415	0.276	0.174	0.401	0.186	0.452	0.131	0.292
B+PANNs+AC+RL	0.634	**0.423**	**0.288**	**0.185**	**0.410**	**0.187**	**0.476**	**0.134**	**0.305**

Conclusion:

- Transfer learning and reinforcement learning both significantly improve the score of the evaluation metrics.
- Reinforcement learning may impact adversely on the quality of the generated captions.
- Conventional evaluation metrics may not correlate well with human judgements.