Audio Captioning Transformer

Paper ID:68

Xinhao Mei[1], Xubo Liu[1], Qiushi Huang[2], Mark D. Plumbley[1] and Wenwu Wang[1]
[1]Centre for Vision, Speech and Signal Processing (CVSSP),
[2]Department of Computer Science,
University of Surrey, UK

Introduction

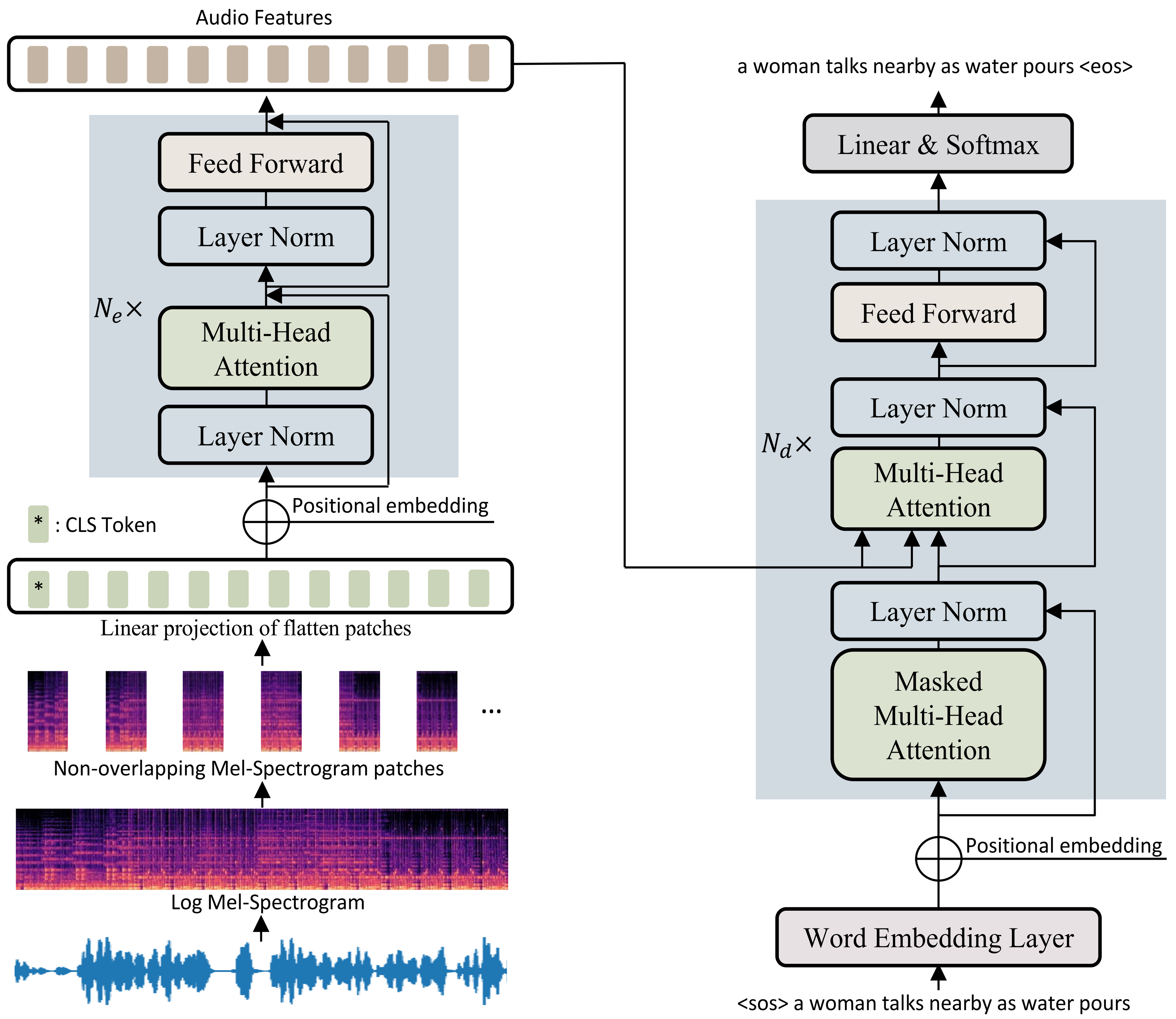
- Audio captioning requires detecting the audio events and their spatial-temporal relationships in an audio clip and describing these information using natural language.
- In this work, we propose a novel full Transformer network, **Audio Captioning Transformer** (ACT) which is based on self-attention and totally convolution free.
- ACT is a simple architecture, but shows competitive performance as compared to other state-of-the-art methods.

Issues

- Convolutional neural networks (CNNs) can be limited in modelling temporal relationships among the time frames in an audio signal.
- Recurrent neural networks (RNNs) can be limited in modelling the long-range dependencies among the time frames in an audio signal.

Methods

- ACT takes the log mel-spectrogram as input and split the mel-spectrogram into non-overlapping patches along the time axis.
- A class token designed to model the global information is appended at the start of the patch sequence.
- Audio features extracted by ACT encoder contains global information in CLS token and fine grid information in patch tokens.
- ACT decoder is a classical Transformer decoder and the whole model is trained with maximum likelihood estimation.



Encoder pre-training

- As Transformer usually requires more training data to achieve reasonable performance, the ACT encoder is adapted from a DeiT model and then pre-trained on AudioSet for an audio tagging task.
- ACT encoder achieves a mean average precision (mAP) of 0.43 on AudioSet.

Results

Model	$BLEU_1$	$BLEU_2$	$BLEU_3$	$BLEU_4$	$ROUGE_L$	METERO	CIDEr	SPICE	SPIDEr
ACT_m_DeiT_AudioSet	**0.653**	**0.495**	**0.363**	**0.259**	**0.471**	**0.222**	0.663	**0.163**	0.413
ACT_l_DeiT_AudioSet	0.647	0.488	0.356	0.252	0.468	0.222	0.679	0.160	0.420
CNN+Transformer	0.641	0.479	0.344	0.236	0.469	0.221	**0.693**	0.159	**0.426**