Carnegie Mellon University

Workshop on Detection and Classification of Acoustic Scenes and Events DCASE2021

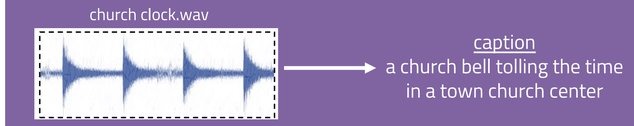
Leveraging State-of-the-art ASR techniques for Audio Captioning

Chaitanya Narisetty[1], Tomoki Hayashi[2], Ryunosuke Ishizaki[2], Shinji Watanabe[1], Kazuya Takeda[2]

[1] Carnegie Mellon University, USA, [2] Nagoya University, Japan

Automated Audio Captioning (AAC)

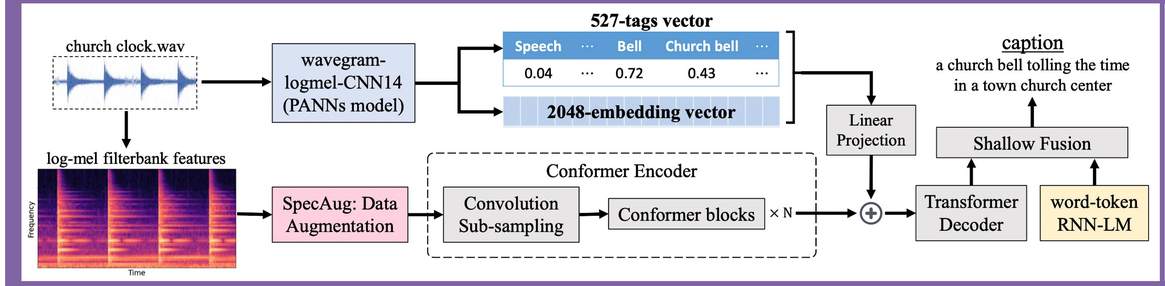
- Task: Generate descriptive captions for a given audio signal
- Intuition: Similar to automatic speech recognition (ASR), a seq-to-seq modeling task focusing on transcription
- Models: CNN, Transformer based encoder-decoder frameworks
- Popular datasets: Clotho and AudioCaps
- Challenge: Limited availability of captioned audio for training



Contributions

- Leveraged end-to-end ASR techniques and proposed a convolution augmented Transformer (Conformer) model, and performed shallow fusion with an RNN-language model (RNN-LM)
- Introduced pretrained audio neural networks (PANNs) to extract audio tags & embeddings, to be used as auxiliary inputs to our model
- Performed extensive evaluation on the DCASE2021 Task 6 dataset and showed significant improvement over the baseline model
- Expanded on our DCASE2021 challenge report with detailed methodology, key insights and ablation studies
- Released our captioning system for reproducible research
  https://github.com/chintu619/espnet/tree/aac_wordtokens/egs/clotho/aac_word

Methodology



Shallow Fusion

- Separately trained a word-token RNN language model using all captions
- Integrated with decoder output using shallow fusion during inference
- Combined the attention scores $\alpha_{\text{att}}(h)$ and look-ahead token scores $\alpha_{\text{lm}}(h)$ for a hypothesis h as:

$$\alpha(h) = \alpha_{\text{att}}(h) + \gamma \cdot \alpha_{\text{lm}}(h)$$

OVERVIEW

- Log-mel filterbank feature inputs, augmented with SpecAug
- Used PANNs to extract 527-tags & 2048-embedding vectors
- Both inputs are fed into our encoder-decoder framework
- Used a Conformer encoder to encode all the audio features
- Used a Transformer decoder to generate words in captions
- Integrated RNN-LM during inference using shallow fusion

Encoder-Decoder Framework

- Encoder: consists of convolution sub-sampling layer and several Conformer blocks. Each block consists of feed forward module (FFN), multi-head self-attention module (MHSA) and a second FNN, in sequence
- Decoder: consists of several Transformer blocks, where each block consists of MHSA, a linear layer, sandwiched between two normalization layers

PANNs

- Used CNN14, one of several PANNs models, trained on large scale AudioSet dataset with 527-tags for an audio classification task
- Used the output 527-tags prediction vector & 2048-embedding vector from last CNN layer
- Both vectors are concatenated, L2 normalized and projected to the same size as attention

Experiments and Results

- Downsampled the audio in Clotho dataset from 44.1kHz to 16kHz
- Trained on Clotho and AudioCaps datasets (with 46,000 samples)
- Used SpecAug based augmentation with time-warp, frequency and time masking parameters $W = 5$, $F_{\text{m}} = 30$ and $T_{\text{m}} = 40$
- Used DCASE2021 task 6 challenge baseline system for comparison
- Shallow fusion was performed with scaling parameter γ set to 0.2

Method	BLEU-1,2,3,4				ROUGE-L	METEOR	CIDEr	SPICE	SPIDEr
Baseline	0.389	0.136	0.055	0.015	0.262	0.074	0.084	0.033	0.054
Conformer	0.512	0.317	0.205	0.131	0.336	0.148	0.310	0.100	0.205
smaller enc-dec	0.500	0.311	0.203	0.129	0.336	0.144	0.299	0.099	0.199
smaller attention	0.490	0.307	0.199	0.127	0.332	0.143	0.310	0.096	0.203
+ larger-kernel	0.496	0.307	0.198	0.124	0.336	0.143	0.297	0.098	0.198
+ auxiliary features	**0.521**	**0.330**	**0.217**	**0.138**	**0.345**	**0.154**	**0.323**	**0.107**	**0.215**
+ dev-eval split	0.515	0.321	0.207	0.131	0.340	0.149	0.314	0.101	0.208
Ensemble	0.533	0.343	0.226	0.146	0.355	0.154	0.341	0.106	0.224

Table 1: Scores of evaluation metrics for the development-validation split.

Method	BLEU-1,2,3,4				ROUGE-L	METEOR	CIDEr	SPICE	SPIDEr
Baseline	0.378	0.119	0.050	0.017	0.078	0.263	0.075	0.028	0.051
Conformer	0.534	0.343	**0.233**	**0.158**	0.354	0.157	0.351	0.106	0.228
smaller enc-dec	0.524	0.331	0.219	0.144	0.356	0.153	0.329	0.103	0.216
smaller attention	0.506	0.320	0.212	0.140	0.349	0.152	0.337	0.102	0.219
+ larger-kernel	0.518	0.330	0.224	0.150	0.355	0.154	0.340	0.105	0.223
+ auxiliary features	0.536	0.341	0.225	0.146	0.357	0.160	0.346	0.108	0.227
+ dev-val split	**0.541**	**0.346**	0.231	0.152	**0.356**	**0.161**	**0.362**	**0.110**	**0.236**
Ensemble	0.546	0.356	0.243	0.165	0.369	0.163	0.381	0.110	0.246

Table 2: Scores of evaluation metrics for the development-evaluation split.

Method	CIDEr	SPICE	SPIDEr
Conformer + auxiliary input	0.323	**0.107**	**0.215**
- 527-tags	**0.325**	0.102	0.214
- 2048-embeddings	0.315	0.098	0.207
Conformer + auxiliary input	**0.346**	**0.109**	**0.227**
- 527-tags	**0.346**	0.104	0.225
- 2048-embeddings	0.342	0.106	0.224

Table 3: Evaluating contributions of PANNs tags and embeddings towards model performance on development-validation split (top) and development-evaluation split (bottom).

Method	CIDEr	SPICE	SPIDEr
Conformer	0.310	0.100	0.205
- RNN-LM	0.300	0.098	0.199
Conformer	0.351	0.106	0.228
- RNN-LM	0.344	0.105	0.225

Table 4: Evaluating contribution of RNN-LM towards model performance on development-validation split (top) and development-evaluation split (bottom).

Conclusion

Leveraged ASR techniques for automated audio captioning and opened potential research directions for joint modeling of ASR and AAC tasks