Transfer Learning Followed By Transformer For Automated Audio Captioning

Hyejin Won‡, Baekseung Kim‡ , Il-Youp Kwak *, Changwon Lim *
Department of Statistics, Chung-Ang University, Seoul, Republic of Korea

‡ These authors contributed equally * Corresponding author

Introduction

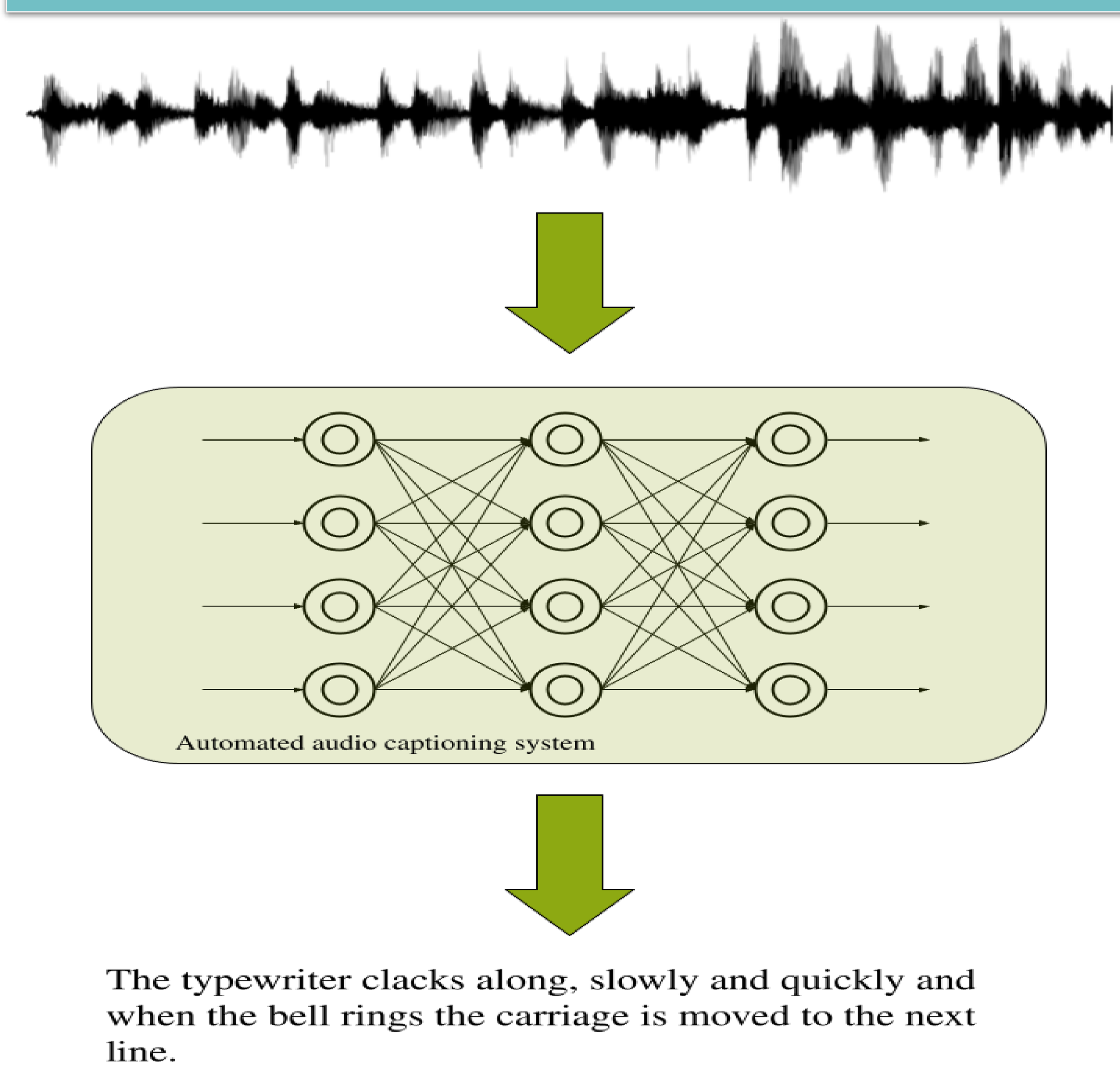

Figure 1: An example of an AAC system and process.

Automated Audio Captioning (AAC) automatically creates captions that can explain the given audio sound data using machine learning techniques.

Researchers investigate the solutions for AAC on DCASE 2021 audio captioning challenge(task 6). In the challenge, a model is required to generate natural language descriptions of a given audio signal.

- We propose transfer learning followed by transformer architecture. With the transfer learning, our proposed model takes two pre-trained networks, 14-layers CNN (CNN14) and 54- layers ResNet (ResNet54), trained on AudioSet as the encoder part.
- Audio feature preprocessing method is log-mel spectrogram.
- Data augmentation method is spec augmentation

http://dcase.community/challenge2021/task-automatic-audio-captioning

Proposed Method

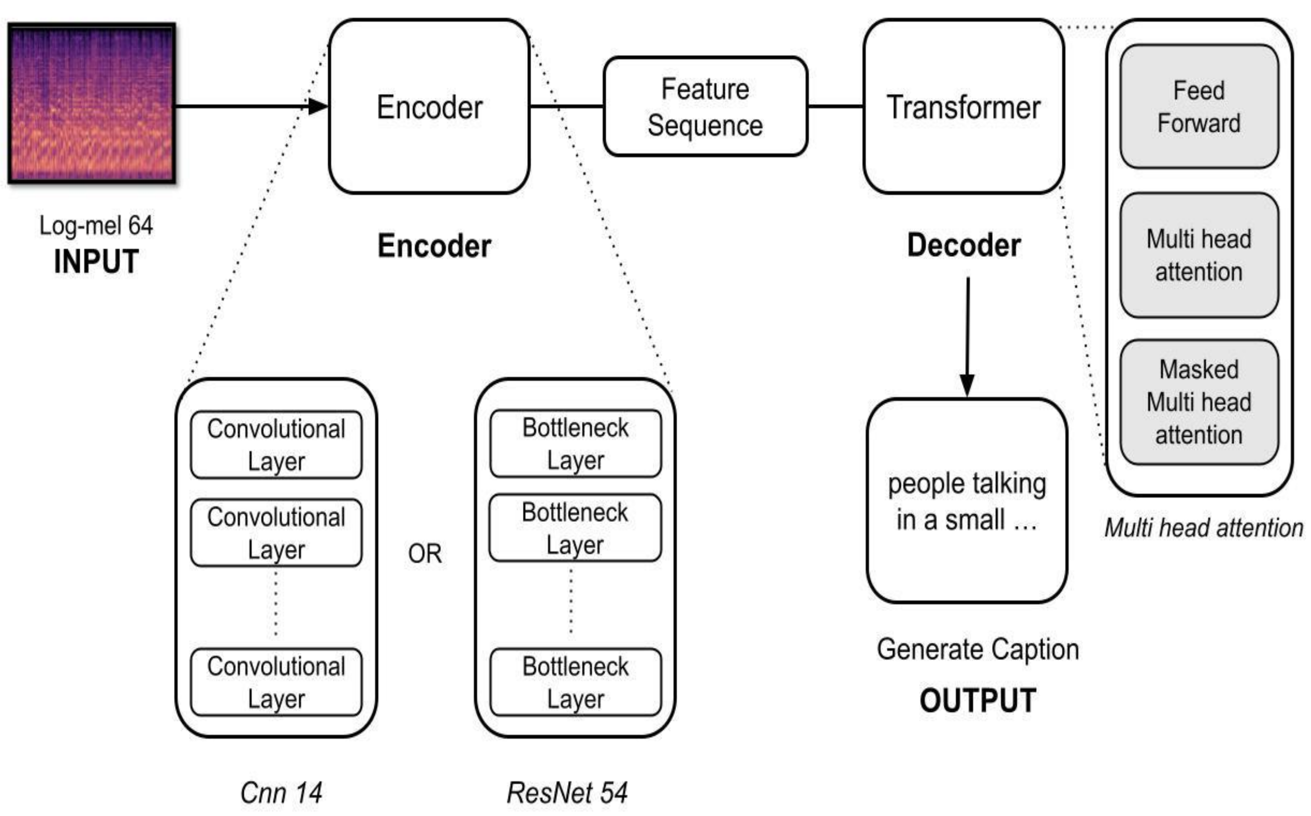

Figure 2. Proposed system architecture

CNN14
Log-mel spectrogram 64 mel bins
(3 × 3 @64,BN,ReLU)×2
Pooling 2 × 2
(3 × 3 @128,BN,ReLU)×2
Pooling 2 × 2
(3 × 3 @256,BN,ReLU)×2
Pooling 2 × 2
(3 × 3 @512,BN,ReLU)×2
Pooling 2 × 2
(3 × 3 @1024,BN,ReLU)×2
Pooling 2 × 2
(3 × 3 @2048,BN,ReLU)*2

Table 1: CNN14(Encoder)

ResNet54
Log-mel spectrogram 64 mel bins
(3 × 3 @512,BN,ReLU)×2
Pooling 2 × 2
(bottleneckB@64)×3
Pooling 2 × 2
(bottleneckB@128)×4
Pooling 2 × 2
(bottleneckB@256)×6
Pooling 2 × 2
(bottleneckB@512)×3
Pooling 2 × 2
(3 × 3 @512,BN,ReLU)×2

Table 2: ResNet54(Encoder)

Our model uses CNN14 or ResNet54 as an encoder and Transformer Decoder for natural language generation.

Encoder

- CNN14: The 14-layer CNN consists of four convolution blocks, each having two 3 × 3 convolution layers with ReLU activation function and batch normalization, with an 2 × 2 average pooling layer between the blocks.
- ResNet54:We had three bottleneck blocks with 64 filters, four bottleneck blocks with 128 filters, 6 bottleneck blocks with 256 filters, and 3 bottleneck blocks with 512 filters. Finally, two 3 x 3 convolutions are applied.

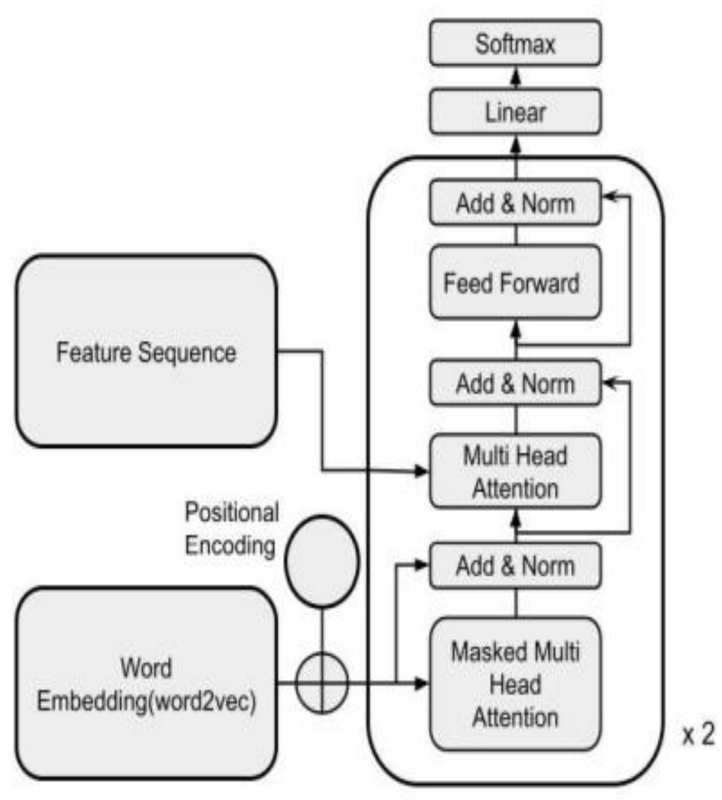

Figure 3: Multi head attention(Decoder)

Decoder

The decoder uses a standard transformer decoder consisting of multi-head self-attention.
The decoder uses a 2-layers transformer with a hidden dimension of 192 and 4 heads.

Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, “Panns: Large-scale pretrained audio neural networks for audio pattern recognition,” IEEE, 2020.
S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, “Improved image captioning via policy gradient optimization of spider,” IEEE, 2017.

Result

Model	$BLEU_1$	$BLEU_2$	$BLEU_3$	$BLEU_4$	$ROUGE_L$	METEOR	CIDEr	SPICE	**SPIDEr**
Baseline Model	0.378	0.119	0.050	0.017	0.263	0.078	0.075	0.028	0.051
CNN14 + Transformer (From Scratch)	0.466	0.262	0.156	0.092	0.309	0.137	0.208	0.087	0.148
ResNet54 + Transformer (From Scratch)	0.459	0.253	0.152	0.084	0.312	0.131	0.182	0.085	0.133
CNN14 + Transformer (Transfer-learning with fine-tuning)	0.552	0.364	0.244	0.159	0.378	0.168	0.395	0.118	0.257
ResNet54 + Transformer (Transfer-learning with fine-tuning)	0.546	0.358	0.239	0.156	0.373	0.166	0.379	0.113	0.246
CNN14 + Transformer (Transfer-learning)	0.564	0.376	0.254	0.163	0.388	0.177	0.441	0.128	0.285
ResNet54 + Transformer (Transfer-learning)	0.540	0.345	0.230	0.152	0.361	0.161	0.383	0.109	0.246

Table 3: Score for model performance on evaluation data

All scenarios of CNN14 Encoder+Transformer Decoder and ResNet54 Encoder+Transformer Decoder models have higher scores than the baseline model in all evaluation metrics.
Also, All Transfer learning scenarios have better performance than the training from the scratch scenario.
We also experiment with fine-tuning the last convolution block of encoder networks (CNN14 and ResNet54 models).
For CNN14 Encoder + Transformer Decoder, the transfer-learning model without fine-tuning works better than with the fine-tuning scenario in all evaluation metrics.

Conclusion

- In DCASE 2021, the use of external data is allowed. Thus, we propose transfer learning followed by a trans- former approach. We adopt CNN14 and ResNet54 pre-trained on AudioSet data because it achieves state-of-the-art performance on audio pattern recognition.
- The pre-trained CNN14 or ResNet54 models are taken as encoder networks for informative audio feature extraction. With the transferred encoder and a transformer decoder, our proposed systems outperform the baseline system with all evaluation metrics.
- Further, we experiment with three training scenarios, 1) from scratch, 2) transfer learning, and 3) transfer learning with fine-tuning.
- Among them, the transfer learning of CNN14 en- coder without fine-tuning works the best, achieving a SPIDEr score of 0.285. Our proposed system ranked 6th in this competition for DCASE 2021 task 6.

This work was supported by the National Research Foundation of Korea (NFR) grant funded by Ministry of Science and ICT (2021R1F1A1056516, 2020R1C1C1A01013020)