CNN-BASED DUAL-STREAM NETWORK FOR AUDIO-VISUAL SCENE CLASSIFICATION

Technical Report

Yuanbo Hou[1*], *Yizhou Tan*[2], *Yue Chang*[3], *Tianyang Huang*[3], *Shengchen Li*[4], *Xi Shao*[3], *Dick Botteldooren*[1]

[1] Ghent University, Gent, Belgium. [2] Beijing University of Posts and Telecommunications, Beijing, China.
[3] Nanjing University of Posts and Telecommunications, Nanjing, China.
[4] Xi'an Jiaotong-Liverpool University, Suzhou, China.

ABSTRACT

This technical report presents the CNN-based dual-stream network for audio-visual scene classification in DCASE 2021 Challenge (Task 1 Subtask B). The proposed method in this report is only trained based on the development dataset in Task 1 Subtask B and does not use any external dataset. For the performance, the model proposed in this report gets 0.318 log-loss and 90.0% accuracy for scene classification on the development dataset, and the log-loss and accuracy in the baseline are 0.658 and 77.0%, respectively. Our results are reproducible, source code is available here: https://github.com/Yuanbo2020/DCASE2021-T1B.

Index Terms— Convolutional neural network (CNN), audio-visual scene classification, dual-stream network

1. MODEL DESCRIPTION

The proposed CNN-based dual-stream network in Figure 1 consists of the audio-based module, image-based module, and audio-visual fusion module. The audio-based module generates audio embeddings that can represent audio context information from the input acoustic features, and the image-based module extracts visual embeddings with spatial information from the input image sequences. Then the audio-visual fusion module fuses the dual-modal information based on self-attention.

1.1. The audio-based module

Motivated by the good performance of acoustic features based on OpenL3 [1] in the baseline system [2], acoustic features from OpenL3 are used as input in the audio branch. For details, acoustic features are calculated with a window length of 1 s and a hop length of 0.1 s, 256 mel filters, using the "environment" content type, resulting in an audio embedding vector of length 512.

Acoustic features will be input into the self-attention [3] layer, and self-attention is a global way of attention. Based on this global attention, the model can focus on information related to the task and ignore irrelevant noise. Then there are four convolutional layers with different parameters, the purpose of which is to focus on the local features of the audio and perform downsampling. The output dimension of the audio module is $(batch, frames, emb\_dim)$, where emb_dim represents the dimension of embeddings. In this report, it is 256 and $frames$ is 10. This means that we expect the model to re-represent the input acoustic features from the OpenL3 as 10 frames, and each frame is a 256-dimensional vector. For the specific parameters of the model, please see the source code.

*Email: Yuanbo.Hou@UGent.be .

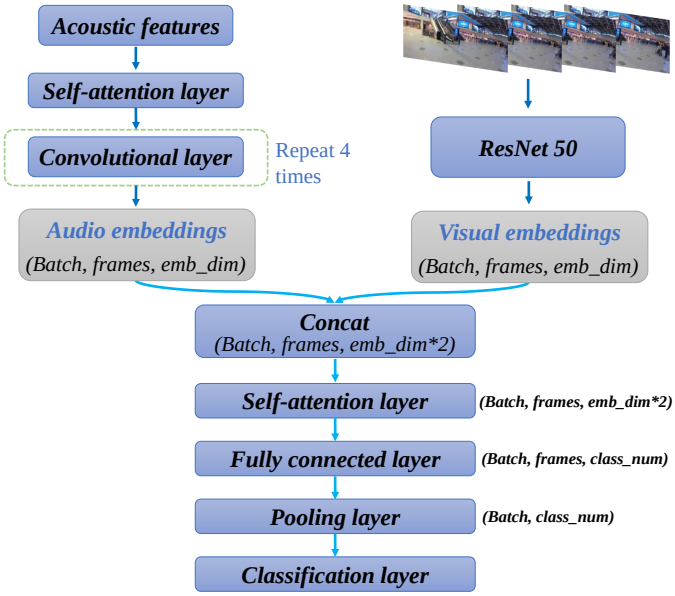

Figure 1: The proposed CNN-based dual-stream network.

1.2. The image-based module

In view of the ResNet [4] showing excellent performance in image processing, the pre-training model ResNet50 is used as a general visual feature extractor in the image module. Furthermore, to make the ResNet50 with strong represent ability more suitable for this task, we fine-tune the parameters of ResNet50 in the training. Consistent with the audio module, the output dimension of the image module is also $(batch, frames, emb\_dim)$, which means that we expect the input image sequence can be re-represented as a embedding with 10 frames, and each frame is a 256-dimensional vector.

1.3. The fusion module

After getting the embeddings of audio and image, in order to better integrate them instead of simply splicing them together, we again use the self-attention layer. Then the pooling layer is applied after the fully connected layer to summarize the high-level audio-visual representations into a fixed-length vector. To combine the advantages of maximum and average pooling [5], we sum the averaged and maximized vectors.

2. EXPERIMENT

2.1. Dataset

The model proposed in this report only uses the development dataset in the TAU Urban Audio Visual Scenes 2021 and does not use any external datasets. The development set contains audio and video data from 10 cities. The total amount of audio in the development set is 34 hours. Provided files have a length of 10 seconds. In order to comprehensively consider the audio-visual context information, the entire audio and video clips with a length of 10 seconds are used during the training phase. Instead of using audio and video clips with a length of 1 second like the processing way in the baseline.

2.2. Experimental setup

Since this task is a multiclass classification task, the activation function of the last classification layer is Softmax, and the loss function is cross-entropy loss. Adam optimizer [6] is used with a learning rate set to 0.0001 and weight decay of 0.1. All models are trained up to 50 epochs with a batch size of 16, data shuffling between epochs. For other parameters, please refer to our source code.

3. ACKNOWLEDGEMENTS

This research received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme.

4. REFERENCES

[1] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 609–617.

[2] S. Wang, T. Heittola, A. Mesaros, and T. Virtanen, "Audio-visual scene classification: analysis of dcase 2021 challenge submissions," 2021.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[5] J. Pons Puig, O. Nieto Caballero, M. Prockup, E. M. Schmidt, A. F. Ehmann, and X. Serra, "End-to-end learning for music audio tagging at scale," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2018.

[6] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of 3rd International Conference on Learning Representations (ICLR)*, 2015.