SCENE CLASSIFICATION SYSTEM WITH MULTI-MODAL FEATURE FUSION

Technical Report

Yujie Yang
Tsinghua University
HCSI Lab
yyj20@mails.tsinghua.edu.cn

Yanan Luo
Tencent
Youtu Lab
yananluo@tencent.com

Zhiyong Wu
Tsinghua University
HCSI Lab
zywu@sz.tsinghua.edu.cn

ABSTRACT

In this report, we provide a brief overview of our submission for the audio-visual scene classification task of the DCASE 2021 challenge. This report focuses on the joint use of audio and video features to improve the performance of scene classification. We propose a system that fuses multimodal data for scene classification, which incorporates both image and audio information. In order to extract audio features, we train a CNN model and a transformer model to classify the log-mel spectra. In order to extract video features, we use convolutional vision transformer to train an image classifier. We have trained a feature fusion network using mixed features to build an image audio feature classifier. As a result the best system achieved an accuracy of 93.9% and a logloss of 0.223 on the DCASE2020 challenge's test set.

Index Terms— Audio-visual scene classification, multi-model feature fusion, convolutional neural network, transformer

1. INTRODUCTION

This challenge is from the task 1b of Detection and Classification of Acoustic Scenes and Events (DCASE) [1]. Different from previous years, which focused on acoustic scene classification alone, this year's competition provides video data corresponding to audio. By fusing multimodal data, it is theoretically possible to improve the performance of the scene classifier. This task's goal is to classify 10 scenes (Airport, Shopping mall, Metro station, Park, Pedestrian street, Public square, Street traffic, Tram, Bus, Metro) in individual one second audio and video data.

Our solution works on three fronts, extracting audio features, extracting video features and using both features for classification. We used the model cnn14 [2], which performs well on AudioSet [3], to train the acoustic scene classifier. In addition to this, we trained a transformer model for acoustic scene classification, which was used to extract audio features. As visual transformers have surpassed traditional convolutional neural networks on image classification tasks, we use a convolutional vision transformer [4] which is pretrained on Imagenet dataset [5] to train the image scene classifier. After this, audio and image features are used to train a classifier with a mixed strategy. On the officially divided test set, the classifier fusing cnn audio features, transformer audio features and image features can achieve 93.9% classification accuracy and 0.223 logloss, which significantly outperform the baseline model.

2. SCENE CLASSIFICATION SYSTEM

Our system is organized into three sections, an acoustic feature extraction module, an image feature extraction module and a feature fusion classification module.

2.1. CNN acoustic feature extractor

The pretrained audio neural networks (PANNs) which trained on raw AudioSet recordings with a wide range of neural networks performs well on audio classification. We learned the structure of the cnn14 model but did not use the parameters pre-trained on AuioSet. Through experimental analysis, we did not use the log-mel spectrum and wavegram double branches mentioned in the paper, but used only the log-mel spectrum as input. The model consists of 12 convolutional layers followed by batch normalization relu and pooling and two fully connected layers. We train the model to do the acoustic scene classification and extract the output before the last fully connected layer as acoustic feature, and the acoustic features are 2048-dimensional vectors.

2.2. Transformer acoustic feature extractor

We also use the transformer model to extract acoustic features. The transformer structure is the same as the original transformer[6], and only the encoder is kept since we just focus on the classification task. All frames of the log-mel spectrum concatenating class token are sent to the encoder layer of the transformer after positional encoding. The output of class token is followed by a fully concatenated layer for classification. The dimension of the transformer encoder's hidden layer is 512, the number of attention heads is 8, and the encoder has 6 layers. We then extract the output of the class token as audio features, and the acoustic features are 512-dimensional vectors.

2.3. CvT image feature extractor

The analysis of the video data revealed that the video content changes very slightly in each second, so we decided to use image features for classification. We use convolutional vision transformer (CvT-13)[4] to extract image features. Compared with ViT[7], CvT has some excellent features of CNN, local perceptual field, shared convolutional weights, and spatial downsampling. Therefore the amount of data for training can be greatly reduced. The CvT-13 model is first pre-trained on imagenet, followed by video frame scene classification. We take the class token output as the image features, and the image features are 384-dimensional vectors.

Table 1: Average accracy and logloss for each system

Features	Accuracy	logloss
cnn14&cvt	92.6%	0.261
transformer&cvt	93.1%	0.230
cnn14&transformer&cvt	93.9%	0.223

2.4. Fusion Classifier

Borrowing from baseline's feature fusion experiments, we also choose not to directly concatenate different features, but to go through fully-connected layers before adding the two features. Unlike the approach of fixing the parameters of the first few layers in the baseline, we adopt the strategy of training from scratch. We adopt a training strategy where, instead of training all the features of the same sample, we randomly select a portion of the features to train the classifier, the features not selected are padding with zero. The mixed feature input allows the classifier to learn how to classify different features at the same time, thus making the classification of multimodal features more accurate.

3. EXPERIMENT

3.1. Data Preparation

We follow the data division in the the official website and baseline for the experiments. The training set contains 7907 samples, the validation set contains 739 samples, and the test set contains 3645 samples, and the samples with the same location id will only appear in one subset. For the video data, we sample a frame for each second of video. The training set contains 79070 images, the validation set contains 7390 images, and the test set contains 36450 images.

3.2. Augmentation

The data augmentation methods we use in audio classification include mixup[8] and SpecAugment[9]. The data augmentation methods we use in image classification include mixup[8], cutmix[10] and random flip.

4. RESULT

We present a total of three versions of experimental results with classification models using cnn14 acoustic features and cvt image features, transformer acoustic features and cvt image features and all three features, respectively.

We use the training set to train the audio visual scene classification system, and the performance on the test set is at table1.

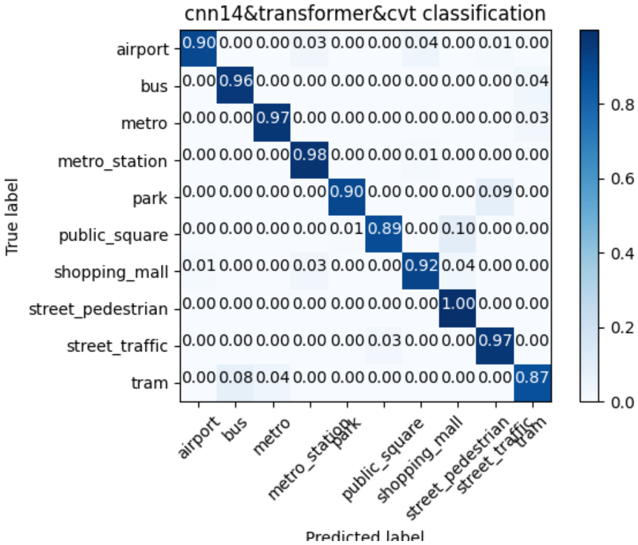

Figure 1: Normalized confusion matrix for the cnn14 transformer cvt features classification model.

5. REFERENCES

[1] http://dcase.community/challenge2021/.

[2] 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/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[3] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[4] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "Cvt: Introducing convolutions to vision transformers," *arXiv preprint arXiv:2103.15808*, 2021.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[8] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[9] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[10] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6023–6032.