Large Scale Multimedia Analysis
Homework 3 Report

Ji In Kwak
Institute of Software Research
Carneige Mellon University
jiink@andrew.cmu.edu

1 Introduction

The purpose of the third homework is to use the features extracted in homework 1 and 2 and complete the multimedia event detection (MED) pipelines. By combining the feature extracted from different modalities, we can expect that the classification performance become better than the unimodal models.

2 Experiments

In the fusion experiments, I conducted two approaches for early fusion and one for late fusion methods. I utilized two feature extraction models for audio data and video data, respectively. For audio classification, the SoundNet model [1] and PaSST[2] model were used which both process the sequential frequency data. In video classification, since the video is constructed with the multiple frame images, the Resnet3D and R(2+1)D network [3] were used.

2.1 Early Fusion

The early fusion is one of the basic methods to deal with the multi-modal data. By concatenating the extracted features from two different modality, we can generate a new feature and use as an input of the MLP classifier. The MLP classifier is trained using the *sklearn.classifier* package. Since every extracted features have different dimensions, each MLP classifier also has different shape of weight parameters. As shown the Kaggle result in Table 1, the PaSST features has better result in the fusion classifier than the SoundNet model. Also, we can see that if the number of features increases from 2 to 3, the test accuracy also increased. This result can be interpreted that there exist more enough information in three embedded features than two of them although R(2+1)D and Resnet3D features comes from the same data modality.

2.2 Late Fusion

Different with the early fusion, late fusion approach focuses on the strength of individual modalities. After training different MLP classifiers for SoundNet and R(2+1)D features, I combined the softmax probability values of two. After computing the mean value of the probability output, the class with maximum probability is selected as the predicted class. The result of late fusion approach shows in the last row of Table 1 which shows better performance than early fusion with same settings but lower than other approaches using PaSST.

Preprint. Under review.

Table 1: The test set accuracy result for different fusion approaches

Fusion	# of features	Feature Extractors	Test Accuracy (Kaggle)
Early	2	SoundNet / R(2+1)D	0.87700
	2	PaSST / R(2+1)D	0.94652
	3	PaSST / R(2+1)D / Resnet3D	0.95187
Late	2	SoundNet / R(2+1)D	0.92780

References

[1] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*, 2016.

[2] Khaled Koutini, Jan Schlüter, Hamid Eghbal-zadeh, and Gerhard Widmer. Efficient training of audio transformers with patchout. *arXiv preprint arXiv:2110.05069*, 2021.

[3] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.