Jaden He

11-775 HW 3

Mar 21th, 2022

Pipeline

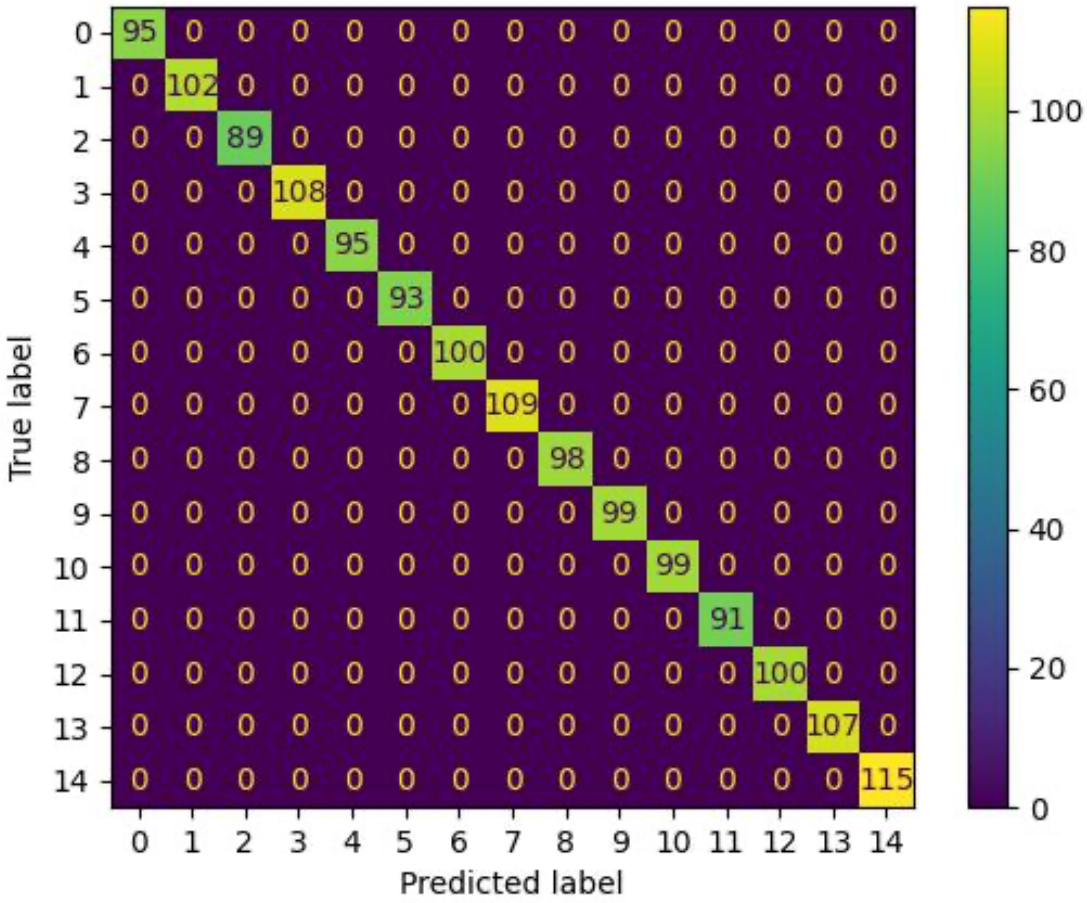
1. Get the audios part in .mp3 and .wav format from the dataset using ffmpeg.
2. Extract PANNs features from .mp3, and PASST from .wav.
3. Extract CNN features from the dataset using ResNet101 and ResNet152.
4. Implement early fusion, late fusion, and double fusion schemes.
5. Train the classification models on MLP (hidden_layer_sizes=(100,), activation="relu", solver="adam", max_iter=max_iter, lr=0.001) from above mentioned features, using three schemes.
6. Evaluate the performance of all models by Top-1 accuracy metric, 80% of the set to train and 20% to validate.
7. Generate confusion matrix for the models using tools/confution_matrix.py.

Table of results

Model	Features	Validation Accuracy	Test Accuracy	Running Time (s)
MLP – Early Fusion	PaSST + ResNet152	0.91000	0.93142	103.98
MLP – Early Fusion	PaSST + ResNet101	0.91067	0.92761	90.46

MLP – Early Fusion	PANNs + ResNet152	0.90867	0.92000	60.63
MLP – Early Fusion	PANNs + ResNet101	0.89933		61.09
MLP – Late Fusion	PaSST + ResNet152	1.0	0.86666	62.59
MLP - Late Fusion	PaSST + ResNet101	1.0		61.73
MLP - Late Fusion	PANNs + ResNet152	1.0	0.87047	89.26
MLP - Late Fusion	PANNs + ResNet101	1.0		82.62
MLP - Double Fusion	PaSST + ResNet152	1.0	0.94857	108.49
MLP - Double Fusion	PANNs + ResNet152	1.0	0.92952	153.11

Confusion matrix and best model



The model with the highest test accuracy on Kaggle is the MLP (hidden_layer_sizes=(100,), activation="relu", solver="adam", max_iter=max_iter, lr=0.001) trained from double fusion using PaSST audio features and ResNet152 features, which achieve 100% validation accuracy, and 94.857% test accuracy. Overall, this model did a perfect job on our 15-class classification. All videos were classified correctly.

Insights and Error analysis

In order to compare all schemes clearly, I use two features for all of them. In early fusion, late fusion, and double fusion three schemes, early fusion was the worst one in validation accuracy. Both late fusion and double fusion achieved 100% accuracy in validation. However, in the test accuracy part, double fusion was the best, a little bit better than early fusion. It was a surprise that double fusion was worse than early fusion on test accuracy. So it shows that concatenated feature vectors are more effective than separate supervised classifiers. The reason here is late fusion focuses on the individual strength of modalities, however, our videos features are significantly stronger than audio features. It is worth noting that three audio features were missing (LTQ5ODI3NjU5MTQ3OTQ4NTAwOQ==, NTkxNzA4MjE4OTM1ODg4NTYxOA==, LTgxOTM5Mzg2MTMwNzM4NjQzNzg=) since there are three videos without sounds. However, our video features were strong enough to overcome this missing information, especially for early and double fusion schemes.

Running time

1. Extracting .wav and .mp3 audios from videos took about half-hour respectively on my M1 Mac.
2. Extracting PANNs and PaSST scene features took two hours respectively on my M1 Mac.
3. Extracting resnet101 features took 2441 seconds on g4dn.4xlarge instance.
4. Extracting resnet152 features took 3381 seconds on g4dn.4xlarge instance.
5. MLP Running time on all schemes are on the accuracy table.

AWS credits

I have all my $50 AWS credits from HW1. I spent $35.302 on HW2. I spent $7.632 on HW3. So I have $107.066 on my account.