11-775 Large Scale Multimedia Analysis
Spring 2019
HW 3

Sheetal Shalini
Andrew ID : sshalini

Collaboration Statement

1. Did you receive any help whatsoever from anyone in solving this assignment? Whiteboard discussion with Prashant Gupta, Shefali Garg, Karan Saxena, Ashwin Srinivasan
2. Did you give any help whatsoever to anyone in solving this assignment? Whiteboard discussion with Prashant Gupta, Shefali Garg, Karan Saxena, Ashwin Srinivasan
3. Did you find or come across code that implements any part of this assignment ? No

Problem Statement

The task is to perform multimedia event detection (MED) with a combination of audio features (MFCC, ASR, Soundnet) and video features (SURF, CNN, Resnet, Places).

Dataset

The dataset contains 2935 videos, with 3 positive events (P001: assembling shelter; P002: batting in run; P003: making cake) and 1 negative event class (NULL). There are 836 training instances and 400 validation instances. The validation set is used to tune hyper-parameters, conduct ablation studies and report the average precision scores. There are 1699 test videos, for which the results have been predicted by the designed model and their accuracies have been compared.

Features

Various combinations of the following features have been considered:

1. **Resnet** : Visual CNN features extracted from the keyframes using a pretrained Resnet50 model. Feature dimension = 2048

2. **Soundnet** : Audio features extracted using Soundnet. Feature dimension = 1401

3. **Places** : Visual CNN model trained on Places database for scene recognition. Feature dimension = 4096

4. **MFCC** : Audio features developed in HW1. Feature dimension = 400

5. **ASR** : Audio features developed in HW1. Feature dimension = 5609

6. **SURF** : Video features developed in HW2. Feature dimension = 450

Types of Fusions

The following types of fusions have been performed:

1. **Early Fusion** : The features have been concatenated and then trained using an MLP Classifier with 2 hidden layers of sizes 1024, 512.

2. **Late Fusion** : Individual features have been trained using an MLP Classifier with 2 hidden layers of sizes 1024, 512. The NULL, P001, P002, P003 scores obtained from these classifiers have been fused using an MLP meta-classifer with 1 hidden layer of size 100. I tried using an Extra Trees Classifier as the meta-classifier, but it led to a lesser performance, and hence continued thereafter with the MLP classifier.

3. **Double Fusion** : This is a combination of Early Fusion + Late Fusion, wherein individual features have been concatenated, trained and then fused with other combinations using the same meta-classifier mentioned above.

Experiments and Results

The results of Early Fusion, Late Fusion and Double Fusion with their Average Precision (AP) on the validation set have been reported in the tables below. All combinations of the above mentioned features have been tried out, but only the best few have been mentioned in the tables below.

Conclusion

Early Fusion

The best **MAP values of 0.987537, 0.818906, 0.950855, 0.598564** and a **test set accuracy of 0.84451** were obtained by the combination **Places + Resnet + MFCC**. Since they are a combination of both audio and visual features, they tend to perform well. Other combinations with the similar features also perform around the same, as shown in the table.

Features	AP-NULL	AP-P001	AP-P002	AP-P003
MFCC + Resnet	0.989117	0.76123	0.924856	0.657109
MFCC + Places	0.976097	0.75561	0.909373	0.436897
ASR + Resnet	0.987903	0.76141	0.942292	0.597872
ASR + Places	0.978681	0.794965	0.907876	0.458986
SURF + Resnet	0.989113	0.787833	0.930543	0.639593
SURF + Places	0.980401	0.799887	0.93198	0.470352
Resnet + MFCC	0.988956	0.768514	0.941094	0.658568
Resnet + SURF	0.989873	0.783585	0.934211	0.656323
Resnet + Soundnet	0.989807	0.78354	0.957102	0.560521
Resnet + Places	0.985736	0.791764	0.948148	0.540066
Places + Resnet	0.987702	0.834666	0.938588	0.57169
Places + MFCC	0.979668	0.808382	0.919698	0.434087
Places + SURF	0.977406	0.781897	0.934675	0.469657
MFCC + SURF + Resnet	0.989608	0.769013	0.934354	0.671847
MFCC + Resnet + SURF	0.989557	0.788532	0.943868	0.630263
MFCC + Resnet + Soundnet	0.989513	0.762435	0.95659	0.594457
MFCC + Resnet + Places	0.985818	0.83095	0.941647	0.581073
Resnet + SURF + Soundnet	0.989154	0.784721	0.946382	0.590192
Places + Resnet + MFCC	**0.987537**	**0.818906**	**0.950855**	**0.598564**
Places + Resnet + SURF	0.987056	0.819745	0.942289	0.641421
Resnet + SURF + MFCC	0.988842	0.78918	0.950214	0.649075
MFCC + Places + SURF + Resnet	0.986983	0.825942	0.950868	0.549951

Table 1: Early Fusion

Late Fusion

The best **MAP values of 0.985855, 0.815996, 0.898897, 0.573378** and a **test set accuracy of 0.81121** were obtained by the combination **Resnet + Places**. Other combinations with the similar features also perform around the same, as shown in the table.

Double Fusion

The best **MAP values of 0.987698, 0.84659, 0.900095, 0.626655** and a **test set accuracy of 0.82807** were obtained by the combination **{Resnet, Soundnet}+{Places, SURF}**. Other combinations with the similar features also perform around the same, as shown in the table.

Features	AP-NULL	AP-P001	AP-P002	AP-P003
MFCC + Resnet	0.974878	0.749503	0.896141	0.414402
MFCC + Places	0.967488	0.710601	0.826932	0.341601
SURF + Resnet	0.974944	0.72946	0.848029	0.435006
SURF + Places	0.960278	0.789373	0.872244	0.40424
Resnet + Soundnet	0.978425	0.769639	0.871001	0.515131
Resnet + Places	**0.985855**	**0.815996**	**0.898897**	**0.573378**
MFCC + SURF + Resnet	0.964274	0.783779	0.744785	0.414828
MFCC + Resnet + Soundnet	0.967087	0.718495	0.744714	0.450126
MFCC + Resnet + Places	0.976204	0.795592	0.895804	0.485875
Resnet + SURF + Soundnet	0.96732	0.773924	0.833583	0.440236
Places + Resnet + MFCC	0.975312	0.799298	0.895631	0.481057
Places + Resnet + SURF	0.974993	0.817406	0.881472	0.484084
Resnet + SURF + MFCC	0.963289	0.789181	0.795077	0.391004
MFCC + Places + SURF + Resnet	0.968552	0.81351	0.8951	0.444448

Table 2: Late Fusion

Features	AP-NULL	AP-P001	AP-P002	AP-P003
{Resnet, SURF} + {Places, MFCC}	0.98901	0.808628	0.900817	0.586162
{Resnet, MFCC} + {Places, SURF}	0.987059	0.822297	0.898872	0.61472
{Resnet, Soundnet} + {Places, SURF}	**0.987698**	**0.84659**	**0.900095**	**0.626655**
{Resnet, Soundnet} + {Places, MFCC}	0.988725	0.842974	0.900325	0.626852
{Resnet, SURF, Soundnet} + {Places}	0.986449	0.846018	0.899897	0.564745
{Resnet + {Places, Soundnet}	0.985225	0.805323	0.898914	0.573318

Table 3: Double Fusion

Overall

Even though double fusion combines the advantages of both early fusion and late fusion, it surprisingly performs slightly lesser than early fusion but better than late fusion. **Early Fusion gave the best MAP values on val set and best accuracy on test set.**

- CPU time taken to train classifiers : Almost instantaneous
- AWS credits not used

Future Work

To develop a custom fusion method for the above features.