HW1: Building MED pipeline.

Email : nyarrabe@andrew.cmu.edu

1. **Data**:
   Given 8000 videos of short duration with ten event categories. TrainVal set is split into Train set and Validation set using 75-25 split ratio. Stratified split is used to ensure that the class label distribution remains the same for both the train set and validation set.

2. **Feature Extraction:**

   1. **MFCC feature extraction:**

      I modified opensmile configuration file to obtain mean normalized features.

      Bag-of-Words Representation :
      To reduce the high dimensional feature space to low dimensional feature, K means clustering could be used to obtain representative features from the full feature set.

      To determine the number of clusters, shiloute score is used to measure the goodness of the clustering. Inter-cluster distance and intra cluster points distance are used to fix the number of K. K=100 is set based on these experiments.

   2. **SoundNet Feature Extraction :**
      SoundNet is an unsupervised learning mechanism to extract features from videos. It uses convolutional neural network layer to encode video features and audio features for each class. Learning happens by bringing the two distributions closer to each other since both audio and video would have same class conditional distributions. I used mean pooling to perform the global pooling of the features among all frames.

      Since the lower layers of the network capture fine grained information about the features, they may not have significant discriminatory power for classifying the categories. I extracted features from layer 10 to 17 which are 1042 dimensional features.

3. **Fusion of both Soundnet features and MFCC features.**
   I experimented with fusing both SoundNet and MFCC features. Since each feature set may capture different nuances of the task, fusion would help in pooling diverse features about the task and will lead to enhances performance.

   To bring down the feature set size, I performed **PCA** on individual feature set to extract 50 features projected along the top 50 principal components from both SoundNet and MFCC features. These 100 features are used to train a MLP layer.

Modelling :

Support Vector Machine Classifier (SVM):

It is a discriminatory classifier used to learn the decision boundary for 1vs all class for each of the classesI also experimented with diff regularization weights;.

Kernel : It specifies the feature space for the transformation of features and a corresponding boundary is learnt in that space. . I experimented with various kernels like linear kernel, Poly Kernel, Gaussian Kernel. Since I used 100 dimensional feature space, polynomial kernel performed best among all the kernels.

Regularization Parameter : Specifies the cost imposed on the loss function for not maintaining the margin for a minor set of points.

Multi Layer Perceptron:

I also experimented with MLP learner. It uses a deep neural network architecture for learning the classes.

Layer Depth : Increasing the number of layers increases the model capacity and can learn any complex function. I have used 2 hidden layers for this model. Increasing it too much can overfit on the training data.

Layer Size: Used (200,100). Layer size indicates the number of functions learnt at each layer.

Activation Function : Experimented with tanh and Relu as activation functions.

Experiments :

Feature Set	Learning Method	Test Accuracy	Comments
MFCC (k=100)	SVM - 1 vs All	38.2	Used Polynomial kernel. BOF with 100 clusters for features.
MFCC	MLP	44.47	Used a 2 layer NN with (100,50)
Soundnet	MLP	56.49	
MFCC+ SOUNDET with PCA	MLP	**60.35**	Fused both Soundnet and MFCC features and used PCA individually for each of them to remove redundancies among features