AFC Coding Assignment 3

Vishaal Udandarao

April 27, 2020

1 Pre-processing, Feature Extraction and Baselines

The given dataset contains 24 speakers in total. Each speaker has 120 video samples each. The distribution of class samples is shown in the table below.

Class	Number of samples per speaker
Neutral	8
Calm	16
Happy	16
Sad	16
Angry	16
Fearful	16
Disgust	16
Surprise	16
Total	120

However, on further analysis of the dataset, each speaker's data can be separated into two subsets – video-only data and audio-visual data. The video data of both the subsets are exactly the same and hence the video-only data can be discarded without any loss of data. Therefore, our final dataset contains 60 samples (4 neutral samples and 7*8 emotional samples). Hence, totally there are 24*60=1440 samples in the dataset.

1.1 Literature Review for emotion detection

Most work on recent emotion detection in the wild (IEMOCAP[1]) and in regulated settings (RAVDESS[2]) datasets use unimodal approaches. Venkataramanan et al.[3] conduct a large scale study on the RAVDESS dataset for the task of emotion detection from speech. They do not consider the neutral emotion and have 14 classes in total – 7 (emotions) * 2 (genders). Their extracted features included Log-Mel Spectrogram, MFCCs, and pitch and energy. The models they applied were LSTMs, CNNs, Hidden Markov Models (HMMs) and deep MLPs. Their best model performed at around an accuracy of 68% using a CNN.

Damodar et al.[5] used a similar strategy by first extracting MFCC features and then employ a DT-CNN hybrid to classify emotions. For the RAVDESS dataset, they achieved a top accuracy of around 71%. Kwoon et al.[4] made use of a hybrid feature model to classify emotions using speech. They used a concatenated feature set mel-spectrograms and MFCC which were then passed into a CNN for classification. They achieved an unweighted accuracy of 80%.

There have also been attempts at incorporating multiple modalities to perform the task of emotion classification. Huang et al.[7] leveraged textual and auditory cues using an early fusion technique to pass through simple ML classifiers like SVM and RBMs. Further, Jannet et al.[6] used a simple early fusion concatenation strategy on deep features. They used an Inception v3 CNN to extract features from visual data and simple plotted waveforms as speech features. They achieved an accuracy of 96% on an inhouse emotional multi-modal dataset. Ghaleb et al.[8] use joint multi-modal embeddings for temporal feature extraction from the videos for the same task of emotion classification. They use frame sequences from the video and the waveforms from the audio as temporal data with similar granularities, and apply a metric learning based loss to learn a discriminative joint embedding space. On the 8-class classification, they achieve a best accuracy of 74%. Further, Beard et al.[9] use an attention based sequential fusion model for learning multi-modal embeddings. They modify existing LSTM blocks to take inputs from three different modalities, and then perform a fusion operation within the LSTM block itself. Their best accuracies are topped at 65%.

1.2 Feature Extraction

By following the approaches mentioned above, I experimented with different feature extraction techniques. The individual modalities are explained in detail further:

1.2.1 Audio Features Extraction

Firstly, I extracted the raw audios from the individual videos by using librosa's load function. Since each of the videos have different lengths, I thresholded all the videos to be at a maximum length of 3 seconds. The sampling rate for the audio signals was taken to 48000 Hz. Hence, every audio signal was represented as an array of length 144000. Then, for each speech sample, I extracted MFCC features by setting different hyperparameters (window size and stride). These MFCC features were then flattened and normalized (min-max norm) into a 1D feature array to be passed to the downstream classification layers.

1.2.2 Visual Features Extraction

Visual features were extracted using two different pre-trained convolutional networks. Firstly, I used a 2D ResNet-152 model trained on ImageNet as a feature extractor. For sampling frames, a sampling rate of 1 frame per second was chosen with an additional frame being captured at the very beginning of the video. Hence, per frame I could easily extract the final feature map of 2048 dimensionality. Since each video had 4 frames, the final dimensionality of the feature matrix was 4*2048, which on flattening returned a 1D array of 8192 dimensions. For the second feature extraction entity, i used a 3D ResNexT-101 pre-trained on the Kinetics dataset. Similar to the sampling and processing technique followed for the 2D feature extraction, the 3D model extracts a final feature vector of 8192 dimensionality. Finally for both the feature vectors extracted, min-max normalization is applied so that the learning algorithm has an easier time optimizing the classification problem.

The entire feature extraction pipeline is shown in the figure below:

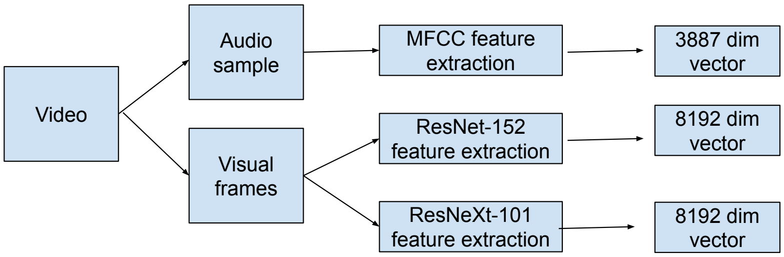

Figure 1: Feature extraction pipeline.

1.3 Training and Test splits

As mentioned in the assignment directive, I used the first 20 speakers for training and tested the performance on the held-out set of 4 speakers. Since we do a redundancy avoidance preprocessing for the dataset, we have a total of 1440 samples in the dataset. Among these, we have 1200 training samples (first 20 speakers) and 240 test samples (last 4 speakers).

2 Results and Analysis

All the models that I have implemented (both uni-modal and multi-modal) use the same base features sets (Audio-MFCC, Visual-2D ResNet-152, Visual-3D ResNeXt-101). Inspired by the uni-modal approaches cited in section 1.1, I take the following models to be my baseline approaches:

1. Uni-modal speech features with multiple downstream ML classifiers
2. Uni-modal visual features (2D and 3D separately) with multiple downstream ML classifiers
3. Simple early fusion model with downstream MLP and softmax

For the improvements, I tried two different model architectures:

1. Hybrid fusion model with modality-specific outputs (using downstream softmax layers)
2. Multi-task learning model with fused task-decisions (using a single downstream softmax layer)

The model architectures of both the improved models are shown below:

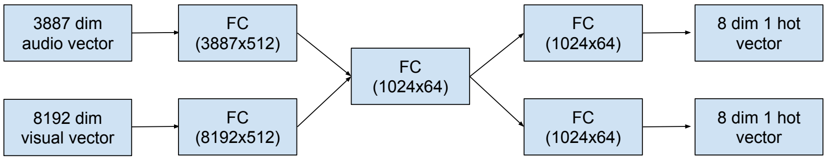

Figure 2: Improved architecture 1: Hybrid fusion with modality-specific outputs.

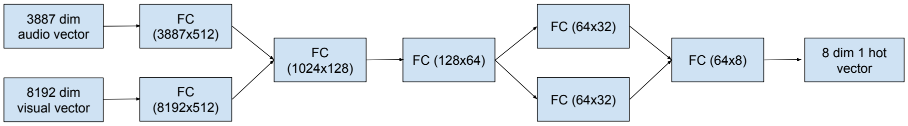

Figure 3: Improved architecture 2: MTL model with fused task outputs.

2.1 Baseline Results

The different ML classifiers used include SVM (RBF kernel), Logistic Regression, Decision Tree, Random Forest and Multi-Layer Perceptron. The baseline results are shown below. All accuracies are reported on final held-out test set.

Modality	Model	Accuracy
Uni-modal Audio	SVM	0.841
	Logistic Regression	0.633
	Decision Tree	0.329
	Random Forest	0.488
	MLP	0.747
Uni-modal Image 2D	SVM	0.12
	Logistic Regression	0.1
	Decision Tree	0.112
	Random Forest	0.171
	MLP	0.443
Uni-modal Image 3D	SVM	0.16
	Logistic Regression	0.13
	Decision Tree	0.12
	Random Forest	0.194
	MLP	0.468
Multi-modal	Early fusion	0.86

Based on the baseline results, we see that implicitly, the audio features are more discriminative than the visual features. For some of the visual features, the performance of the trained classifier is very similar to a random classifier. However, we also see that the early fusion model performs much better than the individual audio and visual MLPs. This empirically gives us evidence that innovative ways of fusing the modality-specific data might lead to better results. However, there is one caveat here – The plots below (Fig 5) show the training loss and accuracy curves for both the uni-modal MLP models. On analysing the curves, we see a strong case of overfitting in the model. Hence, the hypothesis is that by jointly fusion modalities in a hybrid fashion rather than at an early level, the inter-modal dynamics can be captured better.

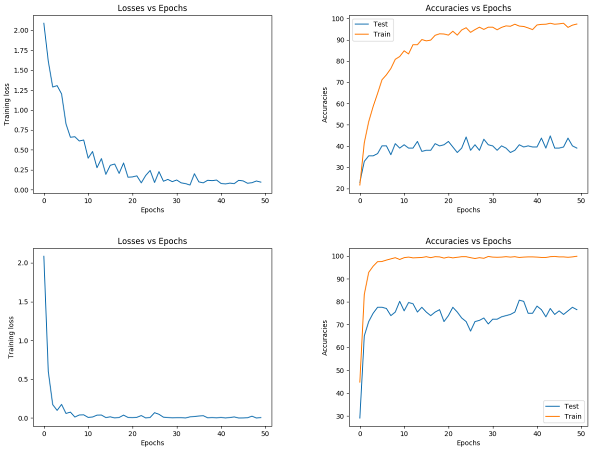

Figure 4: Left Top: Image training loss vs epochs, Right Top: Image accuracies vs epochs, Left bottom: Audio training loss vs epochs, Right bottom: Audio accuracies vs epochs

2.2 Improved Models Results

The results of both the improved models (Figs 2 and 3) are shown in the table 2.2 below. All accuracies are reported on final held-out test set.

Model	**Accuracy**
Hybrid Fusion (Image 2D features)	0.931
Hybrid Fusion (Image 3D features)	0.942
MTL model (Image 2D features)	0.987
MTL model (Image 3D features)	0.991

The training loss and accuracy plots also confirm the convergence and validity of all 4 models:

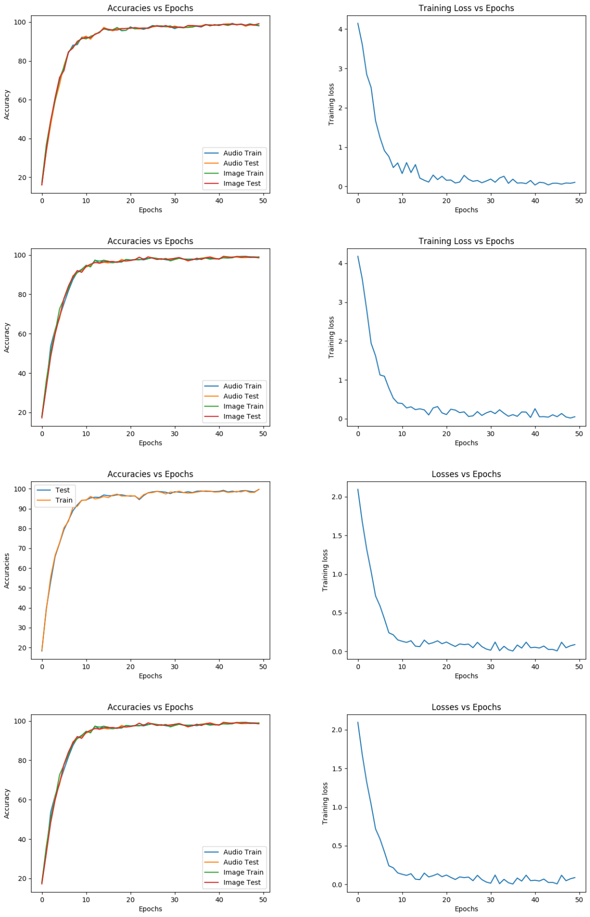

Figure 5: Training loss and accuracy plots of all 4 improved models

On analysing the results of the improved models, we easily see that training joint shared or multi-tasked networks for the task of emotion classification is a good approach and produce superior results to all the baseline models. This is possibly due to two major reasons:

- The joint models are able to aptly capture the significant traits of both early and late fusion. Due to this both the inter and intra-modal dynamics are preserved within the joint embedding space. This helps the model generalize better in a modality-invariant fashion.
- The embedding spaces produced by the models are well discriminated between the classes due to the interactions between samples from both modalities. Hence, this can be thought of as a contrastive algorithm which learns to discriminate class samples solely based on labels and not on their individual modalities.

References

[1] Busso et al, *IEMOCAP: interactive emotional dyadic motion capture database*, Language Resources and Evaluation, 2008

[2] Livingstone et al, *The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English*, PLOS ONE, 2012

[3] Venkataramanan et al, *Emotion Recognition from Speech*, arxiv.org, 2019

[4] Kwoon et al, *A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition*, Sensors, 2019

[5] Damodar et al, *Voice Emotion Recognition using CNN and Decision Tree*, IJITEE, 2019

[6] Jannat et al, *Ubiquitous Emotion Recognition using Audio and Video Data*, International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, 2018

[7] Huang et al, *Human Vocal Sentiment Analysis*, arxiv.org, 2019

[8] Ghaleb et al, *Multimodal and Temporal Perception of Audio-visual Cues for Emotion Recognition*, International Conference on Affective Computing and Intelligent Interaction (ACII), 2019

[9] Beard et al, *Multi-modal Sequence Fusion via Recursive Attention for Emotion Recognition*, Conference on Computational Natural Language Learning (CoNLL), 2018