CSCI 535 Assignment 2

Mehak Piplani

Question 1:

I have used mean pooling for reducing the temporal dimension of the audio and visual files while loading the files into a PyTorch data loader.

Question 2, Question 3:

I tried different types of networks for each modality and listed networks perform well (better than random).

a) Visual Modality

Model: 4 Linear layer model
Hyper-parameters: The number of output neurons for the 2nd (hp_1), and 3rd linear layer (hp_2) and learning rate of the optimizer (hp_3).
Range for Hyper-parameters:

hp_1: [1024, 512, 256]
hp_2: [256, 64, 32]
hp_3: [0.00001, 0.0001]

Results:

F1-Micro: 0.36556 for hyper parameters: 512,64,1e-05

b) Lexical Modality

Model: 4-layer model: Embedding (input), dropout, GRU and linear
Hyper-parameters: The number of neurons for the GRU layer (hp_1 and learning rate of the optimizer (hp_2).
Range for Hyper-parameters:

hp_1: [300, 100, 64, 32]
hp_2: [1e-06, 0.00001,0.0001, 0.001]

Results:

F1-Micro: 0.63509 for hyper parameters: 300,0.0001

c) Acoustic Modality

Model: 4-layer model: Embedding (input), dropout, GRU and linear
Hyper-parameters: The number of neurons for the GRU layer (hp_1) and learning rate of the optimizer (hp_2).
Range for Hyper-parameters:

hp_1: [100, 64, 32]
hp_2: [1e-06, 0.00001,0.0001, 0.001]

Results:

F1-Micro: 0.50499 for hyper parameters: 100,0.0001

Question 4:

The dataset provided has an issue of class imbalance. The statistics of the 4 label classes are as follows:

0: 328, 1: 308, 2: 180, 3: 520, this shows that samples corresponding to class 2 are highly skewed.

I know two techniques for handling the class imbalance issue:

1) Up sampling the minority class
2) Using a weighted loss function

In this assignment I have experimented with a weighted loss function, the weight of the class defined by the result of dividing the number of samples in the majority case by the number of samples in the class for which we must find the weight.

The confusion matrices for the different models:

a) Visual Uni-modal

[[318 188 30 279]
[155 300 42 269]
[131 109 57 166]
[366 322 91 603]]

b) Textual Uni-modal

[[229 17 9 73]
[24 192 23 69]
[21 27 64 68]
[65 72 33 350]]

c) Acoustic Uni-modal

[[214 20 7 87]
[15 194 5 99]
[40 39 58 47]
[79 108 8 325]]

d) Late-Fusion Multi-Modal

[[240 10 0 78]
[15 205 3 85]
[30 24 29 97]
[61 76 3 380]]

e) Early–Fusion Multi-Modal

[[228 18 14 68]
[20 213 21 54]
[22 22 71 65]
[60 67 42 351]]

Question 5:

Early Fusion:

Model: 6 layers: Linear layer, 2 Convolution layers, 3 linear layers
Hyper-parameters: learning rate of the optimizer (hp_1).
Range for Hyper-parameters:

hp_1: [1e-06, 1e-05,0.0001,0.001]

Results:

F1-Micro: 0.65416 for hyper parameters: 0.001

Late Fusion:

I added the probabilities of the output of the model from the three modalities and then predicted the label.

Results:

F1-Micro: 0.63959

Both the fusion methods have resulted in almost similar F1 scores but when we compare the confusion matrices, we observe that early fusion has done a better job in classifying the labels specially in the case of the minority label.

Reason:

In late fusion the result depends how well the three uni-modal networks were trained and hence when we look at the poor performance of visual modality and acoustic modality also being reflected in the fusion model's result, we could say that complementary information is present in the different modalities as discussed in class. Whereas in the case of late fusion, the feature concatenation kind of provides the network to look at all the features at once and is probably able to identify a better equation to model/predict the emotions.

Question 6:

Comparison within Uni-modal Models:

1) Lexical Model outperforms Acoustic and Visual because text features are BERT embeddings which capture context of what is being said and hence can predict the emotion better.
2) The acoustic features do not perform as good as lexical model may be due to ambiguity between tone, lack of context in short statements, etc.
3) Visual modality models lack context and are only face expressions are not indicative of emotion and hence perform the worst.

Comparison of the Uni-modal Models with Multi-modal models:

1) The fusion models outperform all the uni-modal models, but they are not much better since the improvement in F1 score only by a small margin. The fusion techniques are very simple and do not capture the interaction between the three modalities.
2) For all the models, samples corresponding to happiness are confused with Neutral emotion due to a smaller number of training samples belonging to the happiness category.
3) The Early fusion network performs best as the model is probably able to look at all the features at once and identify a relationship with the label making it efficient enough to predict happiness correctly.