June 27, 2024

Angelos Potiriadis

Deep Learning - Multi Modal Dialogue Emotion Recognition

1. Dataset

For this semester's assignment for Deep Learning, I chose to use the Multimodal EmotionLines Dataset (MELD) dataset from Soujanya Poria et al. (2019) [1]. The dataset was chosen mainly for it's multiple modalities and thorough documentation. It is designed for models classifying emotion and sentiment in the context of conversations and contains approximately 13,000 utterances from 1,433 dialogues sourced from the TV series "Friends." The video files and text for each utterance is also provided along with annotations for both sentiment and emotion labels.

The dataset is split on a train - test - validation parts. Emotion labels are categorized into seven classes: anger, disgust, fear, joy, neutral, sadness, and surprise. Sentiment analysis is performed using three classes: positive, negative, and neutral.

Statistics	Train	Dev	Test
# of modality	{a,v,t}	{a,v,t}	{a,v,t}
# of unique words	10,643	2,384	4,361
Avg. utterance length	8.03	7.99	8.28
Max. utterance length	69	37	45
Avg. # of emotions per dialogue	3.30	3.35	3.24
# of dialogues	1039	114	280
# of utterances	9989	1109	2610
# of speakers	260	47	100
# of emotion shift	4003	427	1003
Avg. duration of an utterance	3.59s	3.59s	3.58s

(a) Dataset statistics

	Train	Dev	Test
Anger	1109	153	345
Disgust	271	22	68
Fear	268	40	50
Joy	1743	163	402
Neutral	4710	470	1256
Sadness	683	111	208
Surprise	1205	150	281

(b) Emotion label distribution.

Figure 1: Combined dataset statistics.

1.1 Modalities

The MELD dataset provides us with 3 different modalities, speech, text transcription, and video. Out of these, I utilized the available embeddings for text and audio data, which were extracted by the researchers using the following methods. To extract textual features, each token is initialized with pre-trained 300-dimensional GloVe embeddings. These embeddings are subsequently processed using a 1D-CNN to generate 100-dimensional textual feature representations. For audio features, the openSMILE toolkit was used to extract a wide set of features. Due to the high dimensionality of the audio features, L2-based feature selection

with sparse estimators, such as SVMs, is used to derive a dense representation of the audio segments.

There are multiple ways to deal with multimodal deep learning datasets. They can be fused using early fusion, which combines raw features or embeddings from different modalities at the input level (Summaira et al. 2021) [2], late fusion, which integrates high-level features from separate modality-specific models, or other methods more complicated methods which often combine both approaches . The but the goal of this project is to experiment with different modalities and determine whether combining them through early fusion will offer significant improvements over using a single modality.

2 Models

The models I've focused on are the Long Short Term Memory Networks (LSTM), and in total, three different models were tested accross all modalities. The most prominent ones were the standard LSTM and a bidirectional LSTM (bcLSTM).

Key Architecture components:

- **LSTM Layer**: This is the core of the model, consisting of one or more LSTM layers, defined in our hyperparameters.
- **Dropout Layer**: Applied after the LSTM to introduce regularization by randomly setting a fraction of the elements to zero.
- **Fully Connected (Linear) Layer**: A linear transformation applied to the output of the LSTM's final hidden state, mapping it to the desired output dimension.

The key difference between the LSTM and bcLSTM model, is that the latter processes sequences in both forward and backward directions. To account for the bidirectional outputs, in the fully connected layer, the hidden dimension is multiplied by 2.

2.1 Testing Environment

The models were tested using the following hyperparameters:

- **Input Dimension**: Variable, based on the input feature size
- **Hidden Dimension**: 128
- **Output Dimension**: Variable, based on the number of target classes
- **Number of Layers**: 3 LSTM layers were chosen for all experiments
- **Dropout Rate**: 0.25 was chosen as the dropout rate after the output layer
- **Learning Rate**: 10^{-4} was chosen as the learning rate for all models
- **Epochs**: For the results section, 50 epochs were chosen, because of time constraints. Recommended: 100 epochs

- **Patience**: Early stopping will be called if the model's F1 score during training doesn't improve over 15 epochs. For unimodal models, better results were observed with higher patience, around 20-25 epochs.

The results were cross-validated using 5 folds, with the same hyperparameters applied consistently across all experiments. The best model is saved in the results directory of the project, however the default behavior of the script is to re-train the model for ease of testing new hyperparameters.

3 Results

3.1 Best Model: bcLSTM

The metric I will be looking at for loss computation is weighted F1, in order to be able to compare my results with the state of the art models which are using the same metric. The F1 was evaluated in the end of every epoch, against the validation set. The improvement of F1 across epochs can be found in the following figure. It shows a sharp improvement in the first 15 epochs which gradually declines and finally stagnates around 40 epochs. Early stopping was not called, and there is room for improvement if we increase the epochs.

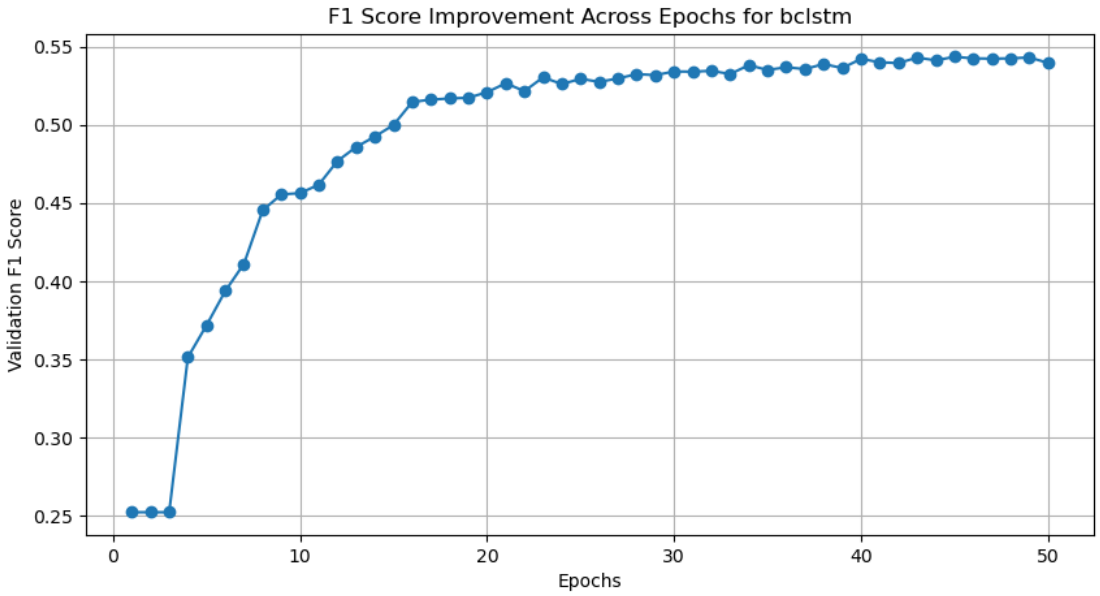

Figure 2: F1 improvement over 50 epochs.

The confusion matrix reveals that the bcLSTM model performs reasonably well for the "neutral" and "joy" classes, with high true positive rates. However, it struggles significantly with the "fear" and "disgust" classes, indicating a potential area for model improvement. The model also shows a tendency to confuse certain emotions, such as neutral and anger, which suggests that these emotions share similar features that the model finds difficult to distinguish. The dataset was also significantly imbalanced with much higher counts of the neutral class than other classes. As a future direction, this dataset imbalance should be accounted for with some sort of balancing, perhaps through resampling.

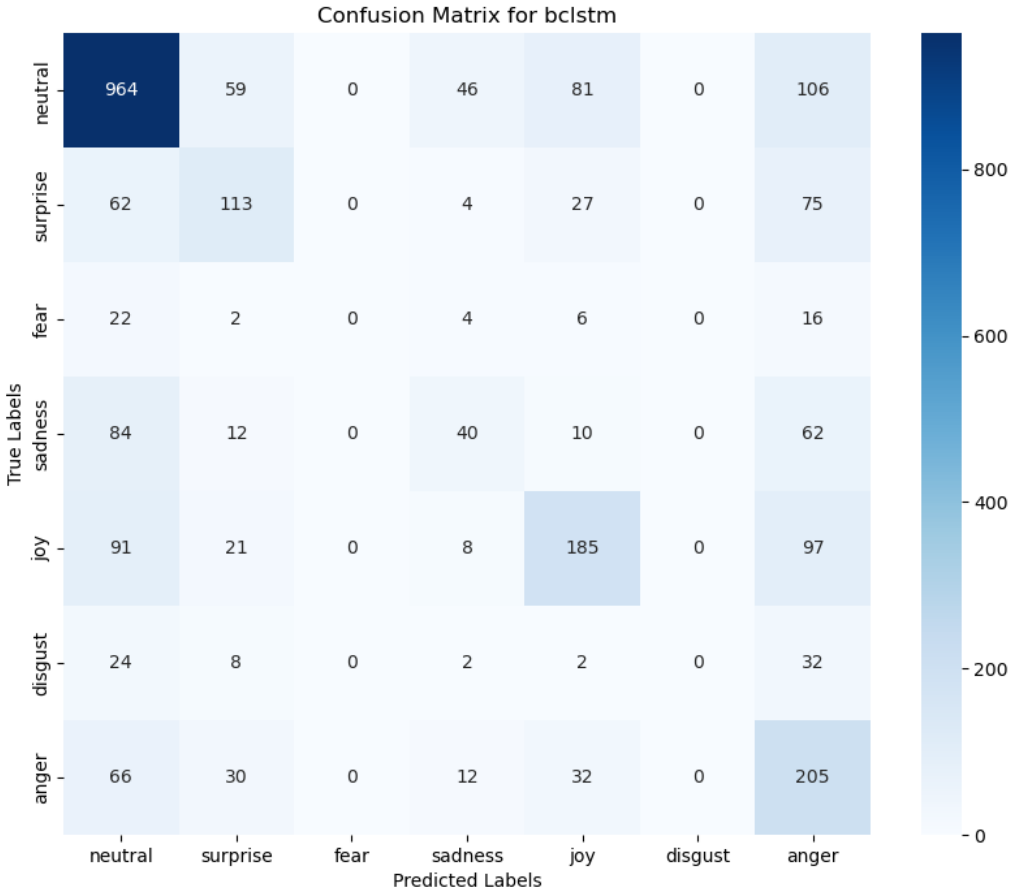

Figure 3: Confusion matrix.

The Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) are plotted in the following figure. The values indicate that the model is better than random chance in classifying all classes. Note that the value for our two problematic classes is deceiving, as the AUC of these classes is still above 0.50 because of the lack of missclasifications for these classes.

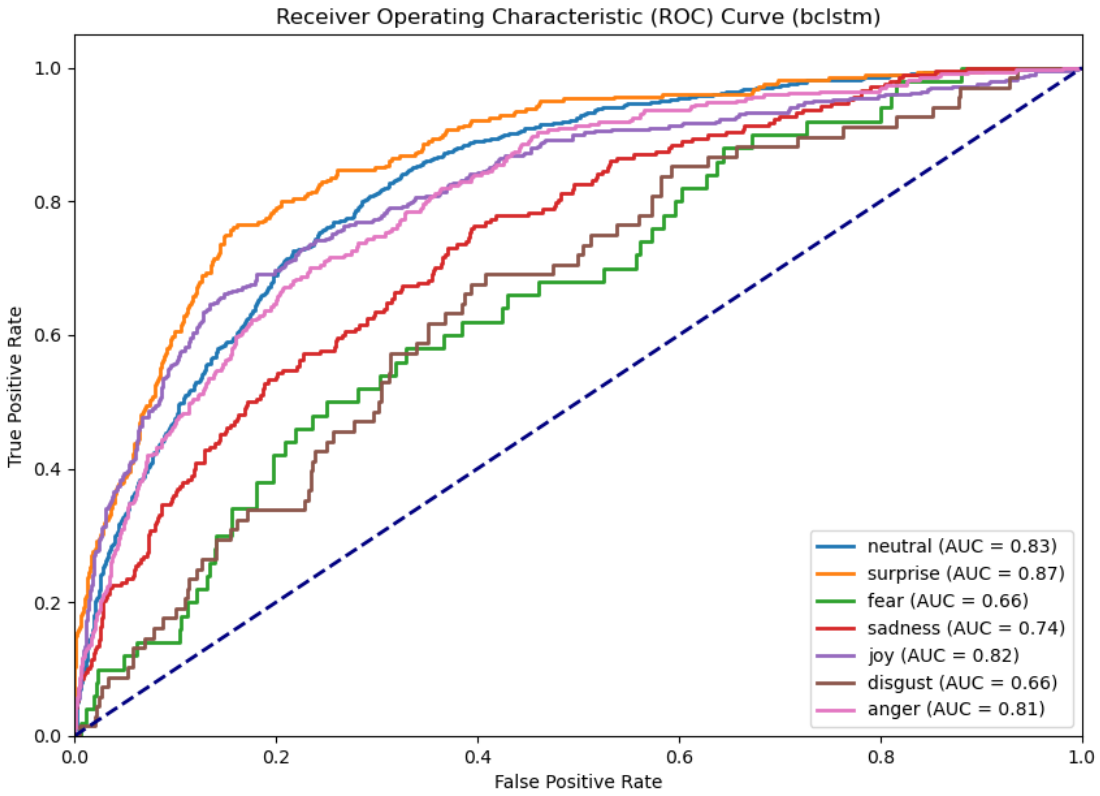

Figure 4: ROC Curve.

The Precision-Recall curve provides insight into the model's performance in terms of precision (positive predictive value) and recall (sensitivity). This curve is particularly useful for imbalanced datasets, where the ROC curve might present an overly optimistic view. Here we can see that our problematic classes are clearly the outliers with very low scores.

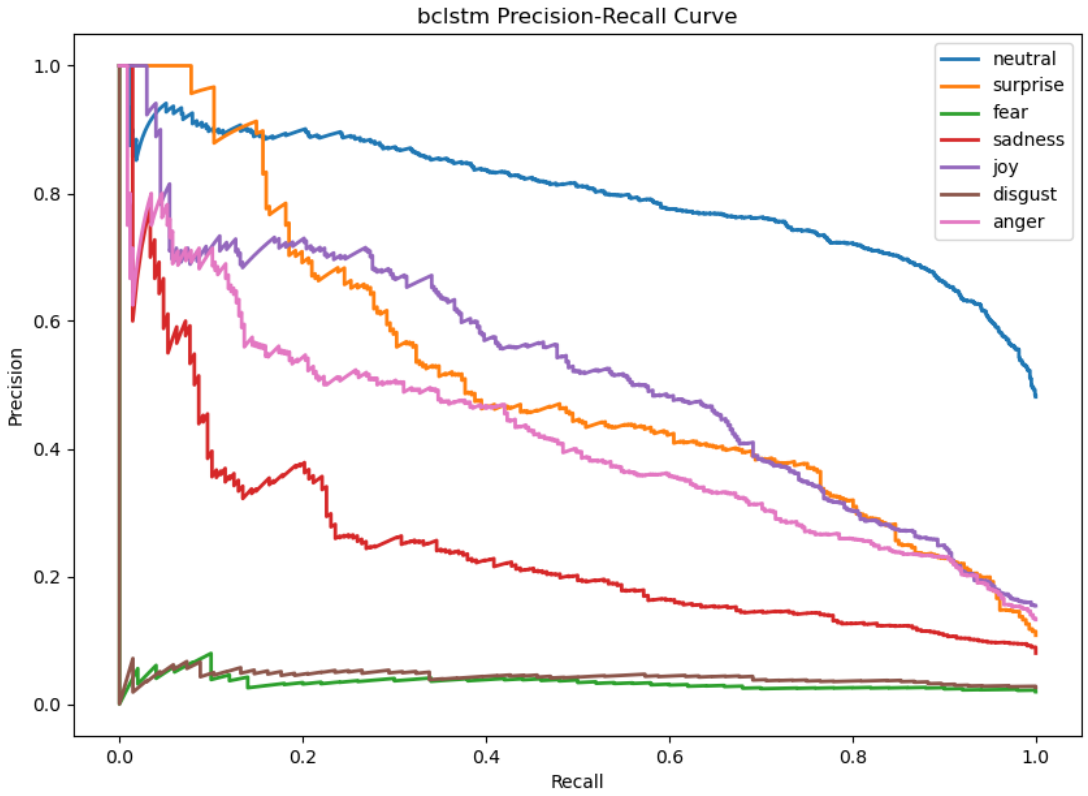

Figure 5: Precision Recall curve.

3.2 Fusion results and comparison with state of the art

For this section, I trained the bcLSTM model to classify Emotion and Sentiment for all modalities, using the previous hyperparameters mentioned in chapter 2.1. We can observe an sharp increase in performance when using both modalities fused together. The unimodal models seem to perform very bad for classifying emotion by themselves, but the advantage in using more than one modality is clear.

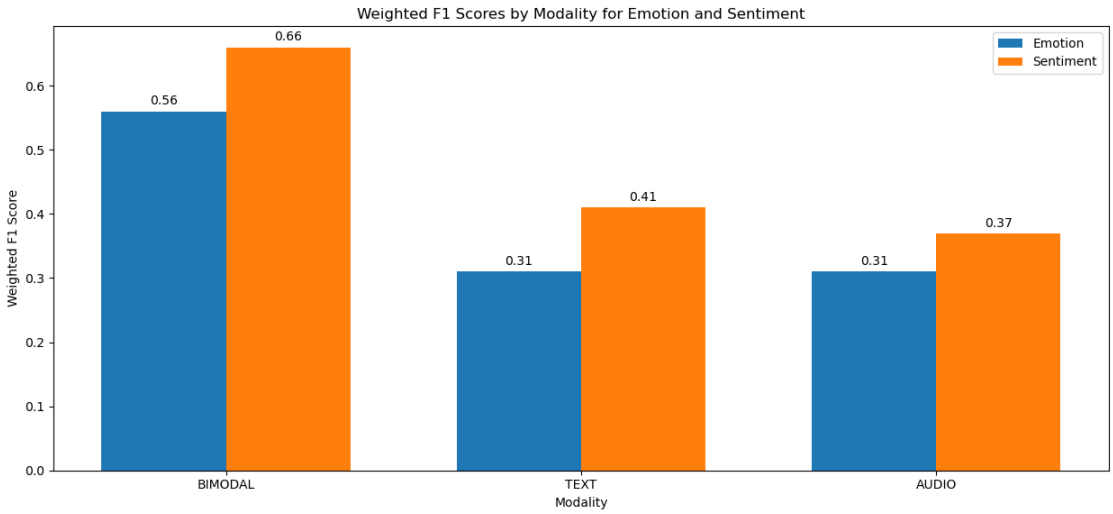

Figure 6: Comparing the results for all modalities.

Finally, in the following figure we can see how the state of the art models perform on the task of emotion recognition. The bcLSTM model of the researchers performs very close to my model, which indicates the model I designed was somewhat successful, with cross validation performance of 55%, a difference of only 1% with the original researchers' bcLSTM model. We can see that recent advancements have pushed the performance of these models to much higher than 55%, but these models use complicated methods and many of them utilize Large Language Models.

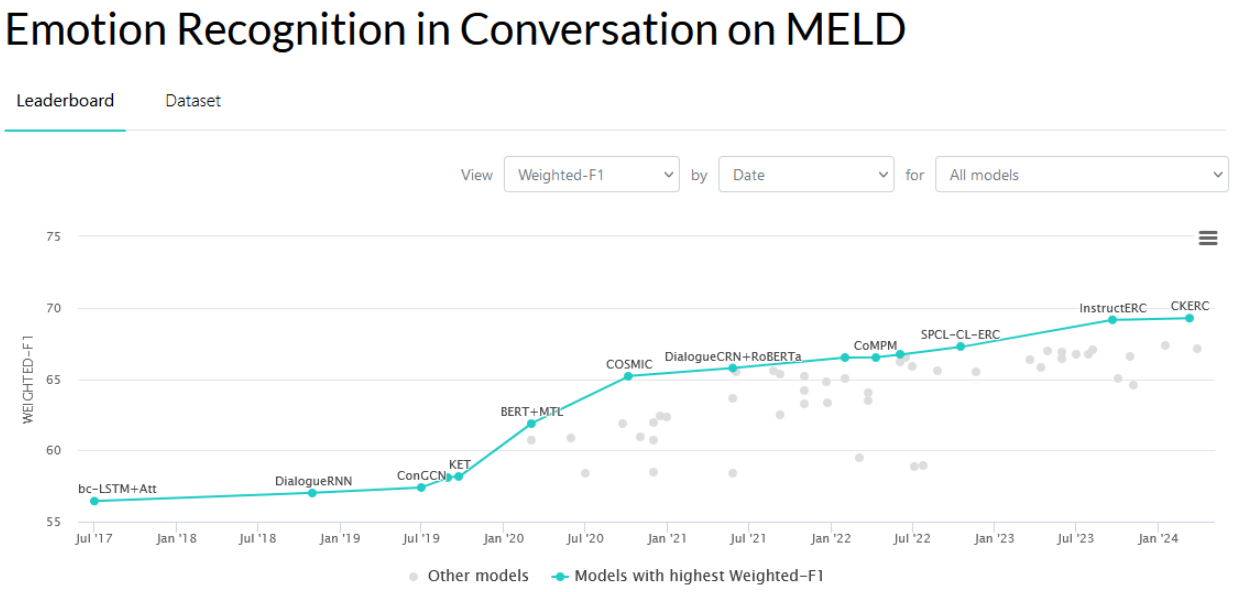

Figure 7: Comparing the results for all modalities.

References

[1] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations, 2019.

[2] Jabeen Summaira, Xi Li, Amin Muhammad Shoib, Songyuan Li, and Jabbar Abdul. Recent advances and trends in multimodal deep learning: A review, 2021.