Multimodal Analysis of Mental Health

Team Members

- Rojin Bakhti
- Sukavanan Nanjundan

Problem Definition

- ❏ Detecting depression among patients using audio and text modalities.
- ❏ Creating unimodal (text-only or speech-only) models
- ❏ Creating multimodal model (text + speech)
- ❏ Assess the performance improvement going from unimodal to bimodal

Data

- Dataset used - Extended Distress Analysis Interview Corpus (E-DAIC)
- Contains semi-clinical interviews, designed to diagnose mental health disorders such as, anxiety, depression and PTSD.
- Includes audio and video recordings of the interview along with the automatically generated transcript of the interview.

Partition	# Subjects	Duration [h:min:s]
Training	163	43:30:20
Development	56	14:47:31
Test	56	14:52:42
All	**275**	**73:10:33**

Avec 2019 Paper, Ringeva et al.

- Depressed patients - 66, Non-depressed patients - 207.
- Labels: PHQ-8 Score and Binary label (0/1)

Method - Audio Modality

Features:

- Low Level Descriptors:
  - eGeMaps
  - MFCC*
- Deep Spectrum Features:
  - DS-VGG*

Model:

- Model built with eGeMaps and MFCC
- 2 layer 256-d GRU,
- Dropout 0.2
- 128-d fully connected layer followed by 64-d fully connected layer
- Finally, single output node
- Sequence length of 20 minutes (120,000 ms), downsampled to number of sentences uttered
- Batch size 2 - to fit in memory

*In progress

Result - Audio Modality

- Fine-tuning regression model improved the results from previous baseline for the eGeMAPS
- The CCC score is still low compared to the text modality (discussed in the next slides)

```
RMSE Score:  5.450313
CCC:  0.15015449159820934
```

Method - Text Modality

- ➢ Get the turn-wise sentences of the patient, (i.e.) collect every other row in the transcript, which is a sentence spoken by the patient.
- ➢ Use sentence transformer BERT to encode each of these sentences into sentence vectors.
- ➢ Use these vectors as input to the recurrent network model wherein each sentence is considered as one time-step.

Model Architecture - Text Modality

- Classification Model:
  -

```
sbert_GRU(
  (gru): GRU(768, 1024, batch_first=True)
  (dropout): Dropout(p=0.33, inplace=False)
  (linear): Linear(in_features=1024, out_features=512, bias=True)
  (relu): ReLU()
  (classifier): Linear(in_features=512, out_features=2, bias=True)
)
```

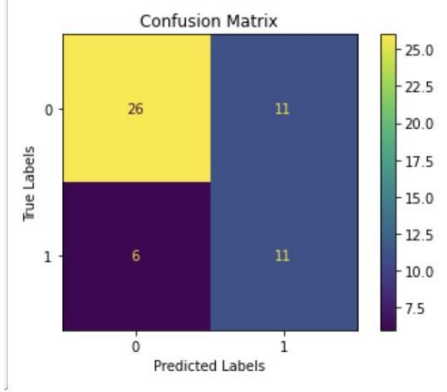
- Regression Model:
  -

```
sbert_GRU_regression(
  (gru): GRU(768, 1024, batch_first=True)
  (dropout): Dropout(p=0, inplace=False)
  (linear_1): Linear(in_features=1024, out_features=512, bias=True)
  (output): Linear(in_features=512, out_features=1, bias=True)
)
```

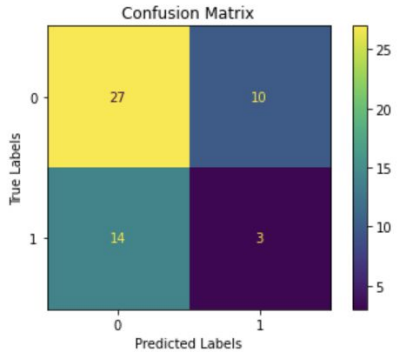
Results - Text Modality

- Classification Task Results

F1: 0.6588628762541806
Accuracy: 0.6851851851851852



F1: 0.4461538461538462
Accuracy: 0.5555555555555556



- Regression Task Results

RMSE Score: 5.0476055
CCC: 0.5827077995159888

RMSE Score: 6.646599
CCC: 0.010858672256921397

Method - Late Fusion

- Performing weighted late fusion with audio and text modality models.
- Weights are calculated as follows:

```
audio_ccc = 0.15
text_ccc = 0.583
total = audio_ccc + text_ccc
audio_ratio = audio_ccc / total
text_ratio = text_ccc / total
```

```
audio_ratio
```

```
0.20463847203274216
```

```
text_ratio
```

```
0.7953615279672578
```

Results - Late Fusion

```
RMSE Score:  4.8656716
CCC:  0.45623185773347164
```

- RMSE Score is better than that of both the models.
- But, the CCC score is lesser than what was achieved by the text modality model, indicating that the performance of the audio model is bringing down the whole performance of the late fusion model.
- For comparison, here are the results of equally weighted late fusion model:

```
RMSE Score:  7.8887587
CCC:  0.26125197515339016
```

Method - Early Fusion

- Performing early fusion with audio and text modality models.
- Model Architecture:

```
early_fusion_GRU_regression(
  (gru): GRU(791, 1024, batch_first=True)
  (dropout): Dropout(p=0, inplace=False)
  (linear_1): Linear(in_features=1024, out_features=512, bias=True)
  (output): Linear(in_features=512, out_features=1, bias=True)
  (relu): ReLU()
)
```

Results - Early Fusion

```
RMSE Score:  5.367013
CCC:  0.3544518095281107
```

- RMSE Score is similar to both the models.
- But, the CCC score is lesser than what was achieved by the text modality model, indicating that the performance of the audio model is bringing down the whole performance of the late fusion model.
- But, the CCC score is better than that of the late fusion model, which proves that early fusion is better than late fusion for this particular problem.
- In the future, with a better audio modality model/features, it would be more prudent to explore and improve early fusion model than late fusion models.

Conclusion

- Speech data provides less prediction power for this problem than the text
- Text modality is more reliable when predicting depression score
- RMSE improves when going from unimodal to late fusion bimodal learning for the depression detection task, but the CCC decreases
- Early fusion is better than late fusion for this particular problem.
- Early fusion in this type of problem can be computationally expensive
- Computation power + memory should be assessed carefully when dealing with large amount of data (i.e in the case of speech)

List of References

1. Ringeval, Fabien, et al. "AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition." *Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop*. 2019.
2. Pennington, Jeffrey & Socher, Richard & Manning, Christopher. (2014). Glove: Global Vectors for Word Representation. EMNLP. 14. 1532-1543. 10.3115/v1/D14-1162.
3. Lin, L I. “A concordance correlation coefficient to evaluate reproducibility.” *Biometrics* 45 1 (1989): 255-68.