Multi-level Attention Network using Text, Audio and Video for Depression Prediction

Anupama Ray
IBM Research, India

Siddharth Kumar, Rutvik Reddy, Prerana Mukherjee
Indian Institute of Information Technology, Sri City, India

Ritu Garg
Intel Corporation, India

Paper ID: 8

20-10-2019

Introduction

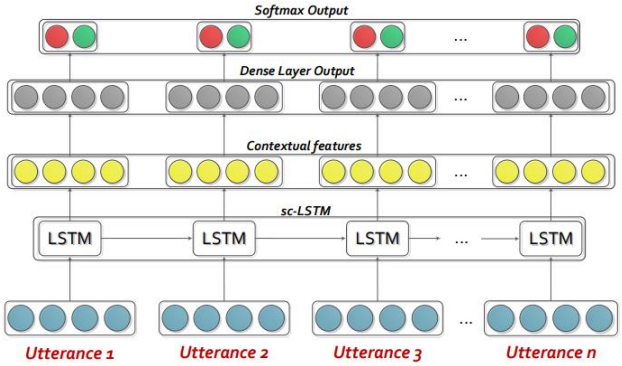
- Depression has been the leading cause of mental-health illness worldwide. Major depressive disorder (MDD), is a common mental health disorder that affects both psychologically as well as physically which could lead to loss of lives.

- Leading cause of mental disability, has tremendous psychological and pharmacological affects and leading to suicidal attempts.

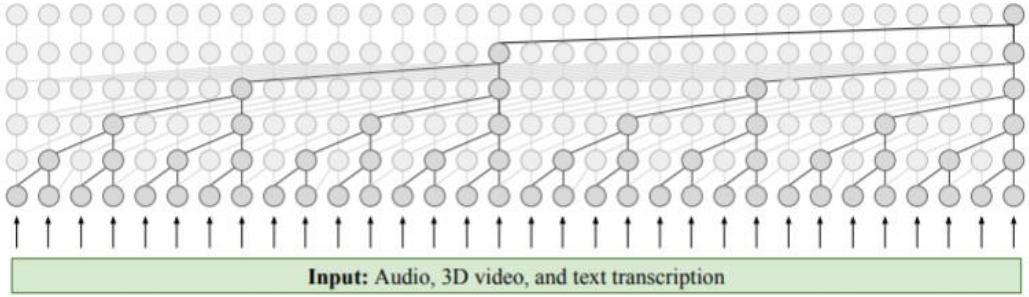
- Our work proposes a multi-level attention framework for feeding in multi-modal features from audio, text and video.

- Our model was able to outperform the current baseline by 17.52%.

Related Works



[1] Multimodal analysis using utterance level features



[2] Causal CNN to produce multimodal feature representation

[1] S.Poria et al, “Context-Dependent Sentiment Analysis in User-Generated Videos.” ACL 2017.
[2] A. Haque et. Al, “Measuring Depression Symptom Severity from Spoken Language and 3D Facial Expressions”.

Related Works

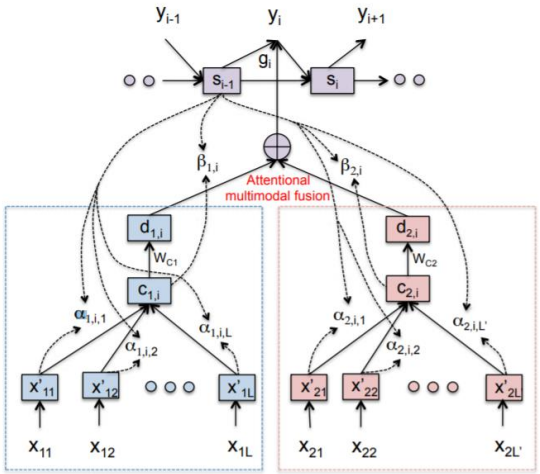


Figure 2. Our Attentional Fusion of multimodal features.

[3] Feature level attention for video and audio modalities

[3] C. Hori et al., “Multimodal Attention for Fusion of Audio and Spatiotemporal Features for Video Description.”, CVPRW, 2018

Gaps

- Previous works explore various methods of fusion, however, there is often very little insight the model provides about contributions of these modalities as they are fused in determining the final outcome.
- Primarily only a single feature of each modality is chosen in most multimodal works. We look to explore multiple sub modalities and look at their individual contributions by applying modality level attention over that. We can then make smart inferences about which sub modalities to finally consider for best performance.

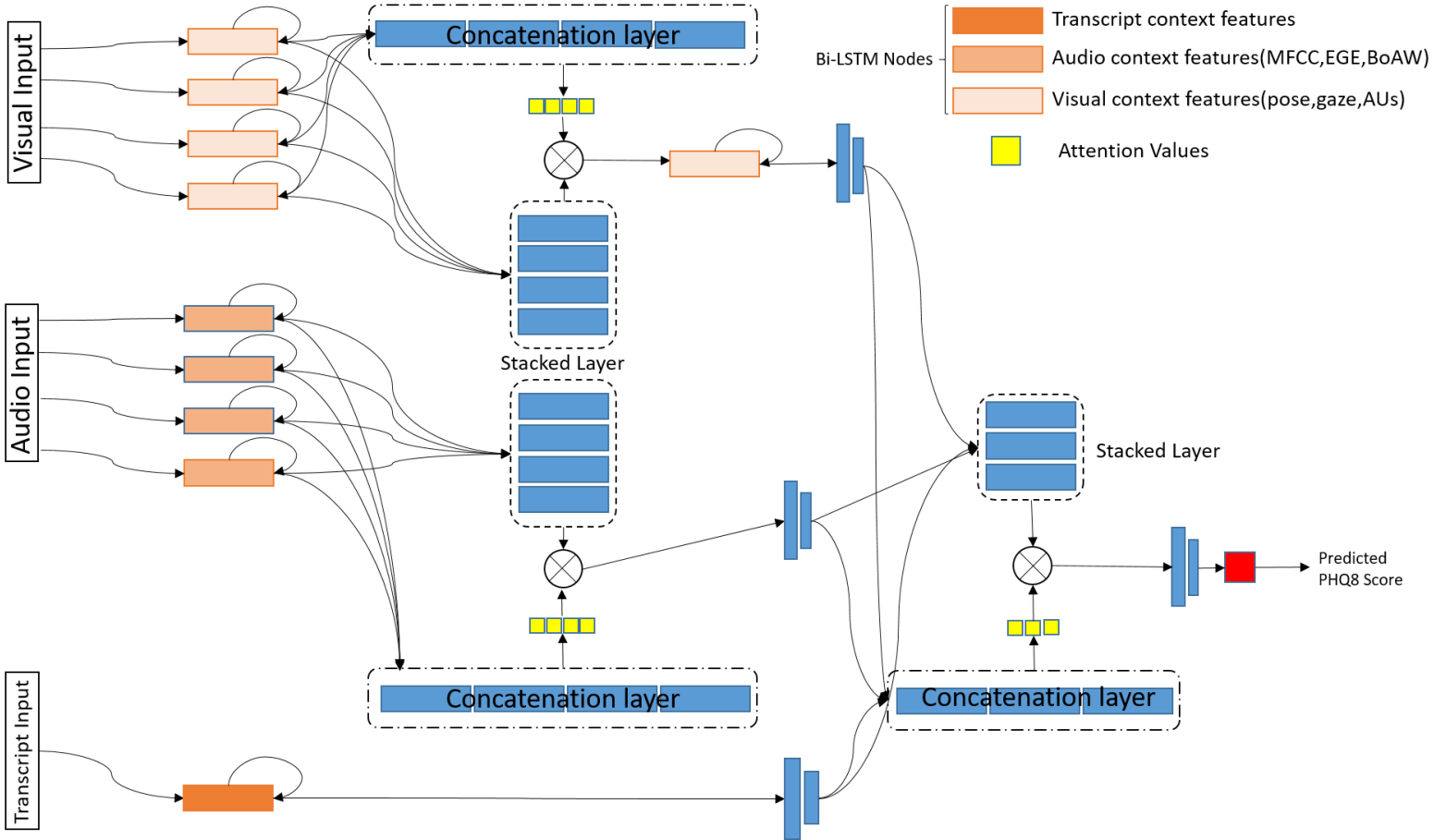
Proposed Framework

- Multi-level attention network that learns to predict depression by learning to attend to the important features within each modality and as well as importance of each modality for the decision.

Ablation studies: Experiments with each feature (of audio and video) as well as each modality separately and all possible combinations to understand role of each feature and modality.

Understanding *correlations between each input modality* and its content.

Proposed Framework



Methodology: Text Modality

- We use the speech-to-text output for the participants in the data.
- Preprocessing: we modified the utterances by replacing short forms of words with the original full word, to take care of out-of-vocabulary words for language modeling and neural network training.
- We used pretrained Universal Sentence Encoder to get sentence embeddings.
- To obtain constant size of the tensors, we zero pad shorter sentences and have a constant number of timesteps as 400 for each file.
- We used 2 layers of stacked BLSTM with sentence embeddings as input and PHQ scores as output to train a regression model.
- We take sum of output of each time step for each file and this is sent as an input to a feedforward layer for regression to get the PHQ8 scores.

Methodology: Audio Modality

- For the audio modality we created different models with each of the low-level descriptor feature and its functional as provided in the challenge.
- For each audio feature, the span of vectors were the participant has spoken was only considered in our experimentation.
- The functional audio and deep densenet features are sampled at 1Hz, whereas the Bag-of- AudioWords (BoAW)] is sampled at 10Hz and the low-level audio descriptors are sampled at 100Hz.
- Each audio model and one model based on fusion of all audio features are all stacked BLSTM network with two layers each having 200 hidden units. We take the last layer output and pass it to a multi-layer perceptron to regress PHQ8 scores.
- For the fusion of audio features, we pass the features through an attention layer before the LSTM layers to understand the impact of each audio feature in the final decision.

Methodology: Video Modality

- We created separate models for each low-level descriptor video feature as video BoW features. Each low-level descriptor features for Pose, Gaze and Facial Action Units (FAU) are sampled at 10Hz.
- We experimented with both low level features as well as functional features, and observed similar performances.
- Since the deep LSTM networks could also learn similar properties from the data (like functionals and more abstract information), we chose to use the low-level descriptors as it has more information than its mean and standard deviation.
- We also built a fusion network with all video features with 2 layers of BLSTM and 1 attention after each feature to learn to attend to the video feature .

Methodology(Fusion of Modalities)

- We propose a multi-layer attention based network that learns the importance of each feature and weighs them accordingly leading to better early fusion and prediction.
- Such an attention network gives us insights of which features in a modality are more influential in learning. It also gives an understanding of the ratio of contribution of each modality towards the prediction.
- Our fusion model extracts feature vectors from Audio, Video and Text.
- These feature vectors are extracted by passing individual modalities through their subnetworks and a uniform sized vector is obtained for each of them.
- Attention scores are then computed for each modality and the vector obtained is multiplied with its respective softmax attention scores.

Methodology(Fusion of Modalities)

- Each of the individual features is fed through a BiLSTM network and at each timestep
- The output of each of these features is concatenated after passing through an attention layer
- This concatenated output at each timestep is further fed to a BiLSTM network and we take the mean of the obtained output, then pass it through two fully connected layers, each with 200 neurons before finally regressing to obtain a final PHQ8 score.
- We experimented with various combinations such as sum of all outputs, mean of outputs and also by max-pooling as three alternatives, but max-pooling worked best, so we have utilized max-pooling over the LSTM outputs.

Observations from Fusion of Modalities

- On testing the model with individual modalities, we observed that both Video feature model and audio feature model have a much steeper descent than the ASR model, on fusion, the model often got stuck on the minima of the video and audio features which are both quite close.
- To mitigate this and nudge the model towards a minima which takes the path of the minima reached by ASR transcripts, we multiply the final outputs of the attention layer element-wise with a variable vector initialized with values in the reciprocal ratios of the rmse loss for each individual modality in order to prioritize the the text modality initially.
- This led to a stable decline of train and validation loss, more stable than the individual modality loss also, and the final attention scores are indicative of contributions of each individual modality.
- Upon convergence, the attention ratios were *[0.21262352, 0.21262285, 0.57475364]* for video, audio and text respectively

Results from ASR (submitted to challenge)

Text Sub-modality

Partition	**Text**
Dev-proposed	4.37
Dev-Baseline	-

Results

Audio Sub-modality

Partition	Funct MFCC	Funct eGeMAPS	BoAW-M	BoAW-e	DS-DNet
Dev-proposed	5.11	5.52	5.66	5.50	5.65
Dev-Baseline	7.28	7.78	6.32	6.43	8.09

Results

Video Sub-modality

Partition	Pose-LLD	Gaze-LLD	FAU-LLD	BoVW
Dev-proposed	5.85	6.13	5.96	5.70
Dev-Baseline	-	-	7.02	5.99

Results

Attentive Fusion Models

Partition	video-LLD-fused	video-BoVW-fused	Video-Text fused	Audio-Text fused	All-feature-fusion
Dev-proposed	5.55	5.38	4.64	4.37	4.28
Dev-Baseline	-	-	-	-	5.03

Take away notes

- Attention network gives us insights of which features in a modality, and which modality are more influential in learning. Attention gives an understanding of the ratio of contribution of each modality towards the prediction.
- LSTMs with attention prove to be ideal candidate for learning such sequential data.
- Future work: Due to limited availability of labelled training data, we are currently trying to build solutions with Meta Learning for better generalization capabilities.

Thank You!