Visual Emotion & Depression Recognition using LSTM

Anonymous
Department of Computer Science
Golisano College of Computing and Information Sciences
Rochester Institute of Technology
Rochester, NY 14586

Abstract—Automatic emotion & depression recognition has been studied over a decade. Very work has striven to develop a system that automates the recognition process with minimum error. Better systems enable better human-computer interaction & better diagnostic tools for mental health. Our approach tackles emotion recognition & depression recognition as two separate problems using the Audio/Visual Emotion Challenge (AVEC) 2017 dataset. We have restricted our approach to using only visual features, wherein we use a temporal LSTM-RNN model for scoring emotional dimensions such as Arousal, Valence & Likability per frame of the subject video & for identifying a depression possibility per subject. Our trust in temporal-learning for improving the error for both problems has paid rich dividends. Our proposed method helped us fetch 0.125, 0.116 & 0.108 RMSE over Arousal, Valence & Likability dimension respectively, and 2.8310 RMSE for the depression problem.

Index Terms—Temporal Learning; Unimodal features; LSTM; Emotion Dimensions; Depression.

I. INTRODUCTION

BACKGROUND

An improved solution to the Emotion & Depression Recognition problem helps us better the quality of human-computer interactions. Human beings, being constantly driven by their thoughts, need a medium to have a mental health check-up, just like they would have a physical health check-up. Hence, a resolution to this problem would have us make leaps in mental diagnostics. Along with that, gauging the student responses to a computer tutoring session or audience responses to an auditorium seminar, that without any human intervention, provides us huge leaps in automating our analysis.

Emotion classification problem is a problem that deals with classifying emotions into a very fixed set of categories. These may include: happy, sad, angry, surprised, disgust, neutral & contempt. The solution for this problem has recorded a state-of-the-art accuracy of 71.2% [7]. However, the categories mentioned above do not reference to subtle & continuous emotions. Human emotions are mostly never singleton. They are a combination of multiple emotions and therefore complicated in nature. Dimensional emotion theory has proven to model these set of complicated emotions in the required subtle & continuous fashion [1]. Arousal, Valence & Likability are a subset of the total emotional dimensions.

On the other hand, depression cannot be counted as an emotional dimension. In AVEC 2017, depression was recognized using a metric called as Patient Health-care Questionnaire (PHQ). There were a total of 8 classes used to generate the final PHQ-8 score for the subject. These 8 classes were: NoInterest, Depressed, Sleep, Tired, Appetite, Failure, Concentrating, Moving. The net PHQ-8 score generated on addition of the scores per class was further used to gauge the depression level of the subject. A score greater than or equal to 10 would evaluate the subject as depressed else otherwise.

Our approach for both the problems aimed at coming up with a single model that provided a resolution for both. Since we operated on the AVEC 2017 dataset, we worked on raw videos for the Emotion problem & previously derived features for the Depression problem. For the emotion recognition, we generated features on a per frame basis using the VGG-CNN & fed those into our LSTM model for training. Having experimented with multiple CNN architectures, we concluded VGG-CNN performed better on colored frames. For the depression problem, unlike our reference which used Topic-based modeling, we performed temporal modeling. For training the model, we concatenated various Constrained Local Neural Field (CNLF) subject features such as Action Units (AUs), 2D feature points, 3D feature points, gaze & pose. Since the depressed subjects were under-represented, we oversampled features derived from such subjects.

Our proposed method for both problems use Root Mean Square Error (RMSE) as a performance metric. Due to the unavailability of test dataset labels from the AVEC 2017 authors, we were constrained to working on the combination of the train & validation dataset. We used a small subset of this combination to act as our test dataset. Therefore, it is to be noted that the numbers we claim in this paper cannot be directly compared to the ones in our main references due to the difference in the quantity & subjects of test dataset.

RELATED WORK

[1] considers multi-modal inputs such as audio, visual & textual for resolving emotion Recognition problem. The primary step that [1] performs is exploring different features per mode, fusing them together & readying them for training.
either a temporal or a non-temporal model. Further, it performs a comparison by training a non-temporal model SVR alongside a temporal LSTM model. And, finally, uses a customized loss function when training the LSTM model in order to take into the account the shared representations between the two emotional dimensions, Arousal & Valence. Its results (with Concordance Correlation Coefficient (CCC) as the metric) to challenge the CCC metric of baseline system for AVEC 2017 Emotion challenge and claim to outperform it with a CCC of 0.675 for Arousal, 0.756 for Valence & 0.509 for Likability.

[2] uses topic-based modeling to execute a context-aware analysis of the given subjects. It aimed at emulating the success of this method in text-mining for depression analysis. [2], too, utilized multi-modal inputs such as audio, visual & visual features for determining depression possibility in the subject. The experiments are primarily based on challenging the context-unaware methods for recognizing depression as well as outperforming the AVEC 2017 baseline system for all metrics (i.e. RMSE, MAE, CC and F1-score). Table 3 in [2] proves its proposed method to outperform the challenge baseline for all metrics & Context-unaware baseline for RMSE and MAE. Each of the above metrics have been reported individually across the Cross-Validation, Development & Test datasets from the AVEC 2017 Depression Challenge in [2].

II. METHOD

EMOTION RECOGNITION

As discussed in [1], temporal models have been a preferred choice when dealing with emotion recognition, especially LSTM model. We have followed suite with our approach by using a LSTM model. However, unlike [1], the features we generated from the given videos were using VGG-CNN. Before discussing the architecture of our system, we should look into the distribution of the AVEC 2017 Emotion Challenge dataset.

Emotion Dataset

The AVEC 2017 Emotion dataset consists of videos in the "avi" format. Since there are 64 subjects in total, we fetched 64 videos in the dataset with every video being of approximately 3 minutes. Each video has a 50 frames per second & every frame has scores for emotional dimensions arousal, valence & likability in the subject corresponding label "csv" files. The dataset consists of Male as well as Female subjects, with an age range of 18-50. Table I summarizes the dataset distribution. We trained our temporal model on a large subset of the combination of the Train & Development frames as we did not receive access to the Test labels from the authors of the AVEC 2017 Emotion dataset. Therefore, Table II summarizes our new train, development (or validation) & test datasets for the emotion problem.

<table>
<thead>
<tr>
<th>Phase</th>
<th>Total Frames</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>325,000</td>
</tr>
<tr>
<td>Development</td>
<td>100,000</td>
</tr>
<tr>
<td>Test</td>
<td>25,000</td>
</tr>
</tbody>
</table>

TABLE II
PHASE DISTRIBUTION (EMOTION)

The labels for the frames have the dimensions scored in the range [-1,1]. It is to be noted that the labels have been provided for frames in the multiple of 100ms. However, the distribution between the labels is not very uniform. Most of the frames have labels in the range 0.04 to 0.09, especially for arousal & valence emotion dimensions. We could try oversampling the frames vectors that have labels consisting of higher values for arousal & valence (i.e. in range [0.70, 1]). Apparently, we did not perform oversampling as we wanted to try the dataset by staying true to it.

In addition, the dataset also provides contextual features such as facial orientation per frame in terms of pitch, yaw & roll, pixel coordinates for eye points, pixel coordinates for facial landmarks, etc. However, we did not make use of those features as we were not experimenting with a non-temporal model for this project.

Feature Extraction

Since we used videos in the "avi" format as input for this problem, we needed to perform the feature extraction on a frame-by-frame basis. We experimented with two CNN architectures namely, VGG-CNN & DenseNet-CNN. Based on our experimental results, VGG-CNN performed better than DenseNet-CNN over colored frames in videos. We used the VGG19 model in Keras to perform the feature extraction. The model utilized pre-trained weights from the ImageNet dataset. Further, we fetched the features from the "block5_pool" layer of the model. Every feature vector was of the dimension (7, 7, 512). We later reshaped this vector to (25088,) for training purpose.

Emotion LSTM Model Description

Long Short Term Memory Models (LSTM) & Recurrent Neural Network (RNN) are widely used in applications such as language translation, image captioning, speech recognition, etc. The main characteristic of these deep learning models is their ability to persist information. RNNs blocks make use of loops to persist information but are only able to maintain a very short term persistence. For example, RNNs can be very effective in learning sequences as it only needs to persist the previous number in the sequence. However, LSTMs help
maintain a much longer persistence. LSTMs can be termed as special RNNs with the ability to add or remove information from its cell states through the use of internal gates. Hence, LSTMs prove to be ideal for applications such as image captioning or recognizing emotional dimensions in videos. Figure 1 displays the model we used for recognizing emotional dimensions. The model is a single LSTM model with every LSTM cell of dimension 32. The features derived from the VGG19 model were horizontally stacked together to be fed into the LSTM model for training. The training set consisted of 45,090 feature vectors with every vector of dimension (25088,1). The input was fed to the model in batches of 16 & time steps of 10. Every LSTM cell being of dimension 32, the output produced per cell was (16, 32). This was fed into the Dense layer to produce the final scores for arousal, valence & likability with a dimension of (16, 3) per batch. The Dropout layer was added to avoid the data from over-fitting. The Dense layer used a linear activation for generating scores for the 3 output classes. Also, it is to be noted that we utilized a stateful LSTM for this problem (i.e. we retained the cell states between batches while training over a single epoch). The cell states were reset only at the end of every training epoch. Mean square error (MSE) was chosen as the performance metric with “rmsprop” as the model optimizer.

**DEPRESSION RECOGNITION**

[2] claims context-aware topic-based modeling to effectively recognize depression in subjects. It compares its method to the context-unaware technique of recognizing depression in subjects. However, we aimed at designing a singular model that not only helps recognize emotional dimensions in videos but also gauge the depression possibility of the subject. Before describing the model for this problem, we should look into the distribution of the AVEC 2017 Depression Challenge dataset.

**Depression Dataset**

Similar to AVEC 2017 emotion dataset, we trained our temporal model on a large subset of the combination of the Train & Development feature vectors as we did not receive access to the Test labels from the authors of the AVEC 2017 depression dataset. Therefore, Table III summarizes our new train, development (or validation) & test datasets for the depression problem. It is to be noted that the dataset does not contain raw videos. Instead, it contains files containing feature vectors per frame such as Action Units (AUs), 2D & 3D feature points from subject’s faces, gaze, pose, etc. We perform concatenation of all these features across files to fetch the final set of feature vectors per frame. Also, the dataset was under-represented for the depressed subjects feature vectors. So, we performed oversampling on such feature vectors by replicating them to evenly balance the dataset. The number displayed in table III are after the ones obtained after oversampling.

<table>
<thead>
<tr>
<th>Subjects</th>
<th>Total Feature Vectors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>4,532,435</td>
</tr>
<tr>
<td>Development</td>
<td>1,529,838</td>
</tr>
<tr>
<td>Test</td>
<td>498,445</td>
</tr>
</tbody>
</table>

**TABLE III**

**PHASE DISTRIBUTION (DEPRESSION)**

Figure 2 displays the model we used for recognizing depression possibility in a subject. The model is a single layer LSTM model with every LSTM cell of dimension 32. The training set consisted of 4,532,435 feature vectors with every vector of dimension (379,1). Hence, the net dimension of the training set was (379, 4532435, 1). The input was fed to the model in batches of 16 & time steps of 10. Every LSTM cell being of dimension 32, the output produced per cell was (16, 32). The Dropout layer was added to avoid the data from over-fitting. The Dense layer used a linear activation for generating scores for the 8 output classes. Also, it is to be noted that we utilized a stateful LSTM for this problem (i.e. we retained the cell states between the batches while training over a single epoch). The cell states were reset only at the end of every training epoch. Mean square error (MSE) was chosen as the performance metric with “rmsprop” as the model optimizer.

**III. RESULTS**

**For Emotion Recognition**

Though we have used temporal learning via LSTM for both the problems, we trained two separate models on the individual datasets. We used Root Mean Square Error (RMSE) as our metric for checking the performance of each model. The comparison performed for deriving the RMSE was between the predicted values & ground truth. The table IV compares the RMSE values we derived for the emotional
unlike emotion recognition, depression recognition was a mix of a regression & classification problem. as discussed in the depression recognition section in section ii, we used regression to score the 8 phq classes necessary for gauging the possibility of depression in a subject. the table v compared our result with [2] on our validation dataset. similar to the emotion problem, we could not test on the test dataset of aavec 2017 depression challenge. hence, it is to be noted that our results have been obtained on validating on a subset of the combination of train & validation dataset. we had combined the train & validation dataset as our total dataset and further parted it into our train, validation & test dataset.

for the purpose of demoing the result to our project instructor, we summarized the scores obtained for the 8 classes per video. this way we obtained a single vector of size 8 containing to 5.5% & 6% to 6.5%.

for depression recognition

unlike emotion recognition, depression recognition was a mix of a regression & classification problem. as discussed in the depression recognition section in section ii, we used regression to score the 8 phq classes necessary for gauging the possibility of depression in a subject. the table v compared our result with [2] on our validation dataset. similar to the emotion problem, we could not test on the test dataset of aavec 2017 depression challenge. hence, it is to be noted that our results have been obtained on validating on a subset of the combination of train & validation dataset. we had combined the train & validation dataset as our total dataset and further parted it into our train, validation & test dataset.

for the purpose of demoing the result to our project instructor, we summarized the scores obtained for the 8 classes per video. this way we obtained a single vector of size 8 containing to 5.5% & 6% to 6.5%.

Fig. 3. Active Subject

Fig. 4. Mellow Subject

TABLE V
RESULTS FOR EMOTION RECOGNITION

<table>
<thead>
<tr>
<th>Set</th>
<th>Results (RMSE)</th>
<th>Results from [2] (RMSE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validation</td>
<td>2.83</td>
<td>3.54</td>
</tr>
</tbody>
</table>

TABLE IV
RESULTS FOR EMOTION RECOGNITION

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Results (RMSE)</th>
<th>Results from [1] (RMSE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arousal</td>
<td>0.125</td>
<td>0.086</td>
</tr>
<tr>
<td>Valence</td>
<td>0.116</td>
<td>0.091</td>
</tr>
<tr>
<td>Likability</td>
<td>0.108</td>
<td>0.113</td>
</tr>
</tbody>
</table>

For the purpose of demoing the result to our project instructor, we summarized the scores obtained for the 8 classes per video. This way we obtained a single vector of size 8 containing to 5.5% & 6% to 6.5%.

dimensions with the ones stated in [1]: as mentioned in the description of the emotion dataset, we could not test on the test dataset of aavec 2017 emotion challenge. hence, it is to be noted that our results have been obtained on testing on a subset of the combination of train & validation dataset. we had combined the train & validation dataset as our total dataset and further parted it into our train, validation & test dataset.

For the purpose of demoing the result to our project instructor, we plotted the scores for each of the three dimensions against the subject video in real time. The scatter plot per video was updated on a frame-by-frame basis. figure 3 & 4 show two subjects, the first being more active in its facial expressions than the later. It is quite evident from the population of the plots for both subjects that the more active subject was more consistent with his expressions leading to more even distribution in the 5% to 9.5% range for arousal & valence combined. On the other hand, the mellow subject did not display such consistency and hence had most of its scores for arousal & valence constrained to the range of 5%
the summarized scores per class. All the values in this vector were summed to obtain a single PHQ-8 score. A PHQ-8 score of greater than or equal to 10 indicated the subject to be depressed, else not. We summarized our results for 15 test subjects (from our customized test dataset) into the confusion matrix displayed in Figure 5. The figure shows that 12 out of the 15 test subjects were correctly identified to their respective categories, with 9 being correctly identified as depressed & 3 as not depressed.

![Depression Confusion Matrix](image)

**Fig. 5.** Depression Confusion Matrix for 15 test subjects

**IV. CONCLUSION**

In conclusion, we believe we have shown that both Emotion & Depression Recognition could be solved with temporal learning with a similar LSTM model. Though we have achieved better results than [1] and [2] for the likability dimension in the Emotion Recognition problem & validation RMSE for the Depression Recognition respectively, we cannot prove it unless we get to verify our model on the same test dataset as used in AVEC 2017 Challenge.

Further experiments could be conducted for both problems by using a multi-layer, stateful LSTM. It is usually observed that deeper LSTM model tend to perform better owing to the fact that it gets more number of cells to stores states and that, too, without facing the issue of vanishing or exploding gradients.

**ACKNOWLEDGMENT**

I would like to thank Dr. Ifeoma Nwogu, my advisor for this project, for thinking about tackling this problem in an ingenuous way. Without her passion for Computer Vision & Deep learning as well as her willingness to constantly solve new problems in the domain this project would not have been possible. I would also like to thank the Department for allotting us space & resources on CS Lab machines “Champ” & “Nessie”. Computational speed for our computer programs increased manifold on these CS machines & helped us save a lot of time during the project.

**REFERENCES**


