EMOTION RECOGNITION USING PHYSIOLOGICAL SIGNALS

Name of author

ABSTRACT

The ability to recognize and classify emotions using physiological signals can help solve problems in many fields such as medicine and education. This is made possible using a recurrent neural network, which has been designed to handle sequences of temporal data. After running several tests where the network hyper-parameters were adjusted, it was demonstrated that this type of neural network is capable of learning how to classify emotions from raw sensor data. Furthermore, it was found that certain data types within the dataset provided stronger representations of emotion for the network to learn from when compared to other data types. Therefore, the results from this experiment can serve as further evidence of the possibilities with neural networks in a temporal domain, as well as demonstrate that some physiological signals are more useful than others for a given task.

1. INTRODUCTION

A Recurrent neural network (RNN) is capable of dealing with temporal data. This is important in problems where the information between frames, or time steps, is more important than each individual step. For this experiment, a special kind of RNN called a Long Short-Term Memory Network (LSTM) was used. An important distinction between LSTMs and standard RNNS is that LSTMs are capable of learning long sequences, whereas RNNs are not. This is made possible due to the addition of a gated memory cell. The gates within an LSTM cell are able to dictate which information to remember or forget. In short, this solves the vanishing gradient problem, which RNNs suffer from, and allows the network to learn over longer sequences [1].

For training, the BP4D+ dataset was used [2]. This dataset is comprised of multimodal data collected from 140 subjects, of which 82 are female and 58 are male. Their ages range from 18 to 66 years old, and include a wide range of ethnic backgrounds. Each subject responded to ten different tasks, making up the ten classes that the neural network should classify. These include: happy, sad, surprise, pain, disgust, afraid, startled, skeptical, embarrassment, and fear. During each experiment, a set of instruments were used to record various physiological signals. The physiological signals considered in this experiment include: electrodermal activity, pulse rate, respiration, respiration rate, blood pressure, systolic blood pressure, diastolic blood pressure, mean blood pressure, and all signals combined.

2. RELATED WORKS

In recent years, a lot of research has been conducted on new ways to use RNNs for emotion recognition using signals. Alhagry et al. [3] have proposed a deep learning method to recognize emotion from raw Electroencephalogram (EEG) signals. Their method uses an LSTM network to learn features from the EEG signals and classify them by low/high arousal, valence, and liking. Subsequently, they were able to use the DEAP dataset to test and verify their method. The tests resulted in an average accuracy of 85.65%, 85.45%, and 87.9% for arousal, valence, and liking, respectively. They were able to conclude that these results introduced a higher accuracy method when compared to other methods.

Jinkyu Lee and Ivan Tashev [4] presented a RNN model for the task of speech emotion recognition. They utilized a bidirectional long short-term memory (BLSTM) model to accomplish this task. In order to overcome certain challenges with speech emotion recognition systems, their proposed system accounts for the long-range context effect and the uncertainty of emotional label expressions. Furthermore, all frames in the same utterance, or unit of speech, are mapped to their corresponding emotion label, where each label is a sequence of random variables. After training the sequences using the proposed algorithm, it was concluded that the accuracy of this emotion recognition system was an improvement by up to 12% over the previous extreme learning machine (ELM) based systems.

Shizhe Chen and Qin Jin [5] showcased an approach to emotion recognition with recurrent neural networks using a subset of the RECOLA dataset. From this, they experimented with different approaches to improve the performance of emotion recognition. For instance, they experimented with bidirectional LSTMs, which combine a forward LSTM and a backward LSTM. In this instance, they figured that the bidirectional LSTM did not show enough performance improvement to be worth the time consumption and overfitting issues that were present. After testing several combinations of various novel loss functions, modality features, network architectures, they came to several conclusions. Among these, a single direction LSTM was found to be good enough for dimensional emotion regression for the given task.

Liao et al. [6] demonstrated a convolutional recurrent neural network based method to solve the problem of low emotion recognition rates for single-mode physiological signals. Their method used a convolutional neural network to learn spatial representations of multi-channel EEG signals. Also, they used a Long Short-term Memory network to learn the temporal respresentations of peripheral physiological signals, such as electromyogram and electrooculogram. These two representations were combined to classify emotions. They conducted their experiments using the DEAP dataset. In the combined feature emotion classification, they were able to achieve 93.06% and 91.95% accuracy for the arousal and valence dimensions, respectively.

Kahou et al. [7] proposed a system to tackle the 2015 Recognition in the Wild (EmotiW) Challange. Their system combines previous convolutional neural network based solutions for dealing with emotion recognition in video, with a recurrent neural network framework. Furthermore, they were able to conclude that the fusion of representations from models that were trained using different modalities, specifically their feature-level fusion network, showed higher accuracy when compared to any of the single modality classifiers. Given some challenges with their dataset, they still found their hybrid CNN-RNN architecture yielded promising results for handling emotion recognition in video.

Zhang et al. [8] introduced a new framework called spatial-temporal recurrent neural network (STRNN). Their framework is used for combining the learning of two different signal sources into a single model. They accomplish this by combining two RNN layers, a multi-directional recurrent neural network layer with a bi-directional temporal RNN. The multi-directional RNN traverses the spatial region of each time slice from multiple angles. This allows their framework to capture long-rage contextual cues. Then, the bi-directional layer concatenates spatial features from each time slice produced by the spatial RNN layer. This allows the framework to learn discriminative temporal dependencies from sequences. They tested their STRNN framework on two datasets: the SJTU Emotion EEG dataset (SEED) and the CK+ dataset. In both instances, STRNN showed favorable results over comparable methods.

Yang et al. [9] recognize the challenge with automatic real-time emotion recognition, and address the challenge with a relatively simple pre-processing method to improve the accuracy of emotion recognition. Utilizing a hybrid CNN-RNN model, they were able to effectively train by learning spatial-temporal representations of raw EEG streams. An LSTM module is used to extract contextual information. Using the DEAP benchmarking dataset, they were able to conclude that their proposed pre-processing method is able to improve emotion recognition accuracy by approximately 32%. Furthermore, their hybrid CNN-RNN model achieved a mean accuracy of 90.80% and 91.03% on valence and accuracy, respectively.

Yao et al. [10] developed a framework that incorporated three classifiers: a deep neural network, a convolutional neural network, and a recurrent neural network. They passed three different signal types to the three networks separately and used a fusion strategy they developed to integrate the strengths of these classifiers. This created the models they used for their experiment: LLD-RNN, MS-CNN, and HSF-DNN. They conducted three experiments using the IEMOCAP dataset. The multi-task learning strategy they implemented with their models yielded a notable performance. After taking into account their proposed fusion system, they achieved a weighted accuracy of 57.1%. This result demonstrated the effectiveness of their proposed classifier fusion approach.

Dimitrios Kollias and Stefanos Zafeiriou [11] presented an approach to solving the One-Minute Gradual-Emotion Regognition (OMG-Emotion) Challenge. Their solution involved using a CNN-RNN neural network architecture, but extending it to allow a combination of multiple features generated by the CNN component to be explored by RNN subnets. Furthermore, they used a loss function based on the Concordance Correlation Coeffecient metric, which they claim has shown to provide better insight. After training their model, the valence and arousal prediction accuracies scored 49% and 31%, respectively; This was much higher than their given baseline scores for the challenge.

Ma et al.[12] showcased a multimodal residual LSTM for emotion recognition. The goal with their experiment was to address the lack of exploration in high-level temporal-feature learning with deeper LSTM networks. Their network architecture, MMResLSTM, is capable of learning the correlation between various physiological signals by sharing the weights across the modalities in each layer. They tested their network using the DEAP dataset. From their experiments, they were able to achieve an accuracy of 92.87% and 92.30% for arousal and valence, respectively.

3. METHOD

The network architecture consists of two LSTM layers and a dense output layer. The first LSTM layer contains 64 neurons and uses ReLU for an activation function. The second LSTM layer contains 32 neurons and uses sigmoid for an activation function. Finally, the dense output layer uses the sigmoid activation function. Following each LSTM layer is a dropout layer with a dropout rate of 0.2. The dropout layers are used to reduce the problem of overfitting. The final dense layer outputs 10 units which reflect the 10 emotion classes represented in the BP4D+ dataset. In addition, the Adam optimizer is used with the default learning rate (0.001). The network is trained over 50 epochs and yields an accuracy of 43.97% using specified data from the BP4D+ dataset, discussed in section 4.

The reasons for picking this network architecture is that it is simple and yields good results. There were attempts to add

additional LSTM layers, or use different activation functions, such as Tanh. However, none of these changes made much consistent impact across the datatypes when training.

4. EXPERIMENTS AND RESULTS

For the experiment, the first challenge was to prepare the data given in the BP4D+ dataset so it could be used with the neural network. To do this, the format of the data needed to be considered. The dataset consisted of several text files which were labeled with their corresponding datatypes, along with which class they belonged to. Each text file contained a sequence of sensor data over a period of time. All of the files were guaranteed to be under 1000 lines long, but not all of them were the same length. So each sequence of sensor readings was read from the file into an array. From here, each sequence in the array was padded to fill in the remaining space, if there was any. This ensured that each array was exactly of size 1000, which was needed for the network to be able to train.

During the training process, hyper-parameters were adjusted to determine the best performance with the network architecture that was used. Initially, the network was trained over 20 epochs using a batch size of 128. However, it did not converge with any of the data types in the dataset. Then 40 epochs was tested with the same batch size, and the model still did not converge. Finally, the network was trained over 50 epochs. Here it seemed to converge. Next, the batch size was lowered to 64. This showed a small but consistent performance improvement across the dataset, and was decided to be the final batch size of the network. The learning curves for this network over 50 epochs are shown in Figure 1 and Figure 2.

Metric	Macro Average (%)	Micro Average (%)
Precision	43.97	43.5
Recall	43.5	43.5
F1 Score	41.24	43.5

Table 1. Comparisons between the metrics considered when looking at the performance of the model. The similarities between macro and micro averaging indicate that the dataset does not contain any notable class imbalances.

In determining the performance of the model, there was a question of which metrics to use. During the testing phase of the experiment, several metrics were recorded and saved. These metrics include: precision, recall and F1 score; Micro and macro variants of each of these averages were considered as well. In general, macro averaging considers each class independently, before computing the average metric. Whereas, micro averaging will combine the contributions from all classes before computing the average metric. For the BP4D+ dataset, using the macro averaging made the most sense due to the classes being relatively balanced. This can be shown by looking at comparisons between the micro and macro averages of each of the metrics on a trained model, as seen in Table 1. Of the three metrics, F1 score seemed the most sensible to use as it considers both, recall and precision, in its calculation. Thus, macro F1 score is the metric that was used for determining the final performance of the model.

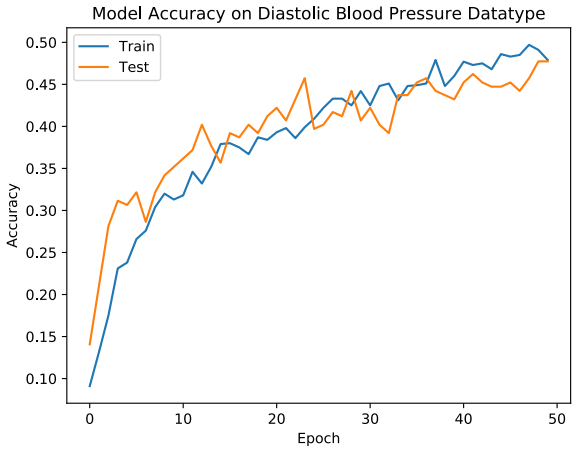

Fig. 1. A comparison between the learning curve of the neural network training and testing on the diastolic blood pressure data type from the BP4D+ dataset. The model converges at around 50 epochs.

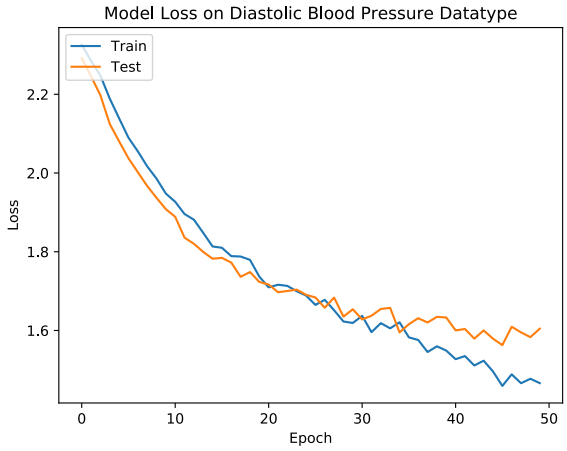

Fig. 2. A comparison between the loss over 50 epochs when training and testing the neural network on the diastolic blood pressure data type.

When looking at the performance across the different datatypes of the BP4D+ dataset, diastolic blood pressure showed to be the best data type for the model to train on. As seen in Table 2, diastolic blood pressure performed over 5%

better when compared to the next highest performing data type. It might be worth noting that there were some clearly worse performing data types, such as respiration. The respiration signal performed the worst out of all data types in the dataset, indicating that breathing patterns might not be affected as much by emotional response when compared to diastolic blood pressure. Furthermore, combining all data types into one set of signals does not yield favorable results. These results indicate that specificity with which modality is being used for determining emotion response is an important factor to consider.

BP4D+ Data Type	**Macro F1 Score (%)**
Electrodermal Activity	30.40
Mean Blood Pressure	35.78
Blood Pressure	37.95
Diastolic Blood Pressure	**43.97**
Systolic Blood Pressure	31.28
Pulse Rate	36.78
Respiration	30.35
Respiration Rate	35.15
All Signals	35.96

Table 2. Comparisons between the Macro F1 scores of the data types considered in the BP4D+ dataset.

5. CONCLUSION

The results of this experiment demonstrate the possibility of recognizing and classifying emotions from physiological signals through the use of a neural network. Using a Long Short-term Memory Network and some fine-tuning, it is possible to achieve a model that can reasonably learn to classify emotions. The multimodality of the BP4D+ dataset allows for some interesting comparisons between different physiological signals. Comparing the performance of the neural network against the physiological signals in the dataset showed that the diastolic blood pressure signal type allowed the network to learn better when compared to the other signals such as respiration activity, which performed the worst. From these comparisons, it can be concluded that certain physiological signals might prove more useful than others in recognizing emotions, and that it is important to distinguish between them when working on emotion recognition tasks.

6. REFERENCES

[1] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–80, 12 1997.

[2] Zheng Zhang, Jeffrey M. Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andrew Horowitz, Huiyuan Yang, Jeffrey F. Cohn, Qiang Ji, and Lijun Yin, "Multimodal spontaneous emotion corpus for human behavior analysis," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3438–3446.

[3] Salma Alhagry, Aly Aly Fahmy, and Reda A. El-Khoribi, "Emotion recognition based on eeg using lstm recurrent neural network," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 10, 2017.

[4] Jinkyu Lee and Ivan Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," 09 2015.

[5] Shizhe Chen and Qin Jin, "Multi-modal dimensional emotion recognition using recurrent neural networks," New York, NY, USA, 2015, AVEC '15, p. 49–56, Association for Computing Machinery.

[6] Jinxiang Liao, Qinghua Zhong, Yongsheng Zhu, and Dongli Cai, "Multimodal physiological signal emotion recognition based on convolutional recurrent neural network," *IOP Conference Series: Materials Science and Engineering*, vol. 782, no. 3, pp. 032005, mar 2020.

[7] Samira Ebrahimi Kahou, Vincent Michalski, Kishore Konda, Roland Memisevic, and Christopher Pal, "Recurrent neural networks for emotion recognition in video," New York, NY, USA, 2015, ICMI '15, p. 467–474, Association for Computing Machinery.

[8] Tong Zhang, Wenming Zheng, Zhen Cui, Yuan Zong, and Yang Li, "Spatial-temporal recurrent neural network for emotion recognition," *CoRR*, vol. abs/1705.04515, 2017.

[9] Yilong Yang, Qingfeng Wu, Ming Qiu, Yingdong Wang, and Xiaowei Chen, "Emotion recognition from multi-channel eeg through parallel convolutional recurrent neural network," in *2018 International Joint Conference on Neural Networks (IJCNN)*, 2018, pp. 1–7.

[10] Zengwei Yao, Zihao Wang, Weihuang Liu, Yaqian Liu, and Jiahui Pan, "Speech emotion recognition using fusion of three multi-task learning-based classifiers: Hsf-dnn, ms-cnn and lld-rnn," *Speech Communication*, vol. 120, 03 2020.

[11] Dimitrios Kollias and Stefanos Zafeiriou, "A multi-component CNN-RNN approach for dimensional emotion recognition in-the-wild," *CoRR*, vol. abs/1805.01452, 2018.

[12] Jiaxin Ma, Hao Tang, Wei-Long Zheng, and Bao-Liang Lu, "Emotion recognition using multimodal residual lstm network," New York, NY, USA, 2019, MM '19, p. 176–183, Association for Computing Machinery.