Comparing performance between RNN & typical NN on detecting human anger by pupils

Chen Yang
Research School of Computer Science
Australian National University
Acton ACT 2601 Australia
U6688648@anu.edu.au

Abstract.

According to Manuel Oliva & Andrey Anikin [1], Pupil response is an autonomous response caused by stimuli rather than a response to cognitive emotion processing the human eye will convey many messages, such as emotions, when the human itself cannot detect. Human body responds differently to different mental conditions, and such reactions often express emotions more realistically.

This article mainly explains the use of pupillary detail information to detect a person's emotions. This report also describe the process of creating and improving RNN model, and compare the performance between the RNN model and typical NN model for this case.

Keywords.

anger recognition, pupillary response, pupil diameter, RNN, Neural network, LSTM.

1 Introduction

Based on the experience I have gained in daily life, our pupils will respond accordingly to our conscious or unconscious anger. M. E. Kret [2] argues that in many occasions, anger can make people's pupils dilate more than other emotions, which means there is a possible method to detect whether a person is angry through pupil diameter and the dilation rate of pupils.

This study trained a two-layer neural network classification model to classify whether the person is angry through multiple iterative learning and analyzed the relationship between different features in pupil data using explaining grade and extracting meaning method [4&5]. Data set of the NN model is collected from a research done by Chen, L., Gedeon, T., Hossain, M. Z., & Caldwell, S. [3]. This research also aims to generate a three-layer RNN to detect humans' anger and compare the performance of both model.

2 Method & Technology

2.1 Dataset

Dataset of RNN was collected from 19 participants' pupillary responses to 20 angry stimuli videos, genuine anger and post anger each account for half of these videos. In order to increase the versatility of the model, the participants are composed of people from different races and different countries. The data set collected for each video all participants'

left and right eye pupil diameters according to the time axis. The max length of the timeline is 186 and the minimal length is 36. The dataset of ordinary NN (mentioned as NN dataset below) contains 400 rows of pupil data which is extracted from 20 different videos played in 20 different sequence. Each pupil data has 8 features including video no, sequence no, mean pupil diameter, Standard deviation pupil diameter, etc.

2.2 Data Preprocessing

NN dataset was preprocessed by removing first two columns (video no & sequence no) which I think is not relevant to our research, and change the class label column to numeric (Genuine -> 1, Posed -> 0). All the features are normalized to range 0 to 1. RNN dataset is much more complex than the former one, because the different length of time observed by each participant leads to inconsistent length of pupil diameter data obtained from different participants, and there are much missing data among the timeline. First of all, the entire dataset was grouped by video no and participants. For each participant, the maximum and minimum values of all pupil diameters of him / her are extracted for data normalization, and the missing data is filled with the average of the two pupil diameters closest to the missing part on the timeline. As people from different race and different counties have different pupil size, So for each participant's pupil data, I only normalize it according to the maximum and minimum values of him / her recorded in all videos, which should be more reasonable. Padding time step by adding 0 to the end of each features to make sure they are all equal in length. Both study shuffle the processed dataset and randomly split data into training set (75%) and testing set (25%), and split both sets into input and target.

2.3 RNN&NN

The structure of the NN model: Six input neurons for six features (Mean, Std, Diff1, Diff2, PCAd1, PCAd2).Single hidden layer, 30 hidden neurons (30 performs best among [10,15,20,25,30,35,40]). Two output neurons respectively indicate the probability of being judged as G / P .The number of epoch: 600. Learning rate: 0.001, Activation function: ReLu. Loss function: CrossEntropyLoss and Optimizer: Adam.

For this research, I choose the LSTM (Long short-term memory) model. LSTM is a special RNN model. Compared to the typical RNN, The main purpose of LSTM is to solve the problem of gradient disappearance and gradient explosion during long sequence training, which means ,in a word , LSTM can perform better in longer sequences than ordinary RNN. For this case, each feature is processed to length 186, the long length time step is the reason to choose LSTM. During the training, I set the number of hidden layer to 2 and the hidden size to 15, because of the more complex dataset. Set Input size to 2 for each time input both left & right pupil diameter at one time step. Two output neurons respectively indicate the probability of being judged as G / P. Reduced the number of epoch to 200, because the training time became longer due to the complexity of the LSTM structure, and there was no significant improvement on the accuracy after 300 iterations. Batch size: 32. Loss function: CrossEntropyLoss and Optimizer: Adam, which are the same as the NN model. Learning rate : I have also tried 0.005 0.01 & 0.001, 0.005 & 0.01 performs bad which will reach a peak of accuracy during few epoch and quickly decrease, the training loss is unstable. At the beginning of the training process, the learning rate is set to 0.001 as the default learning rate of the Adam optimizer. To avoid overfitting, the learning rate will drop to 20% per 50 epoch.

Below are the plot of training process with 150 or 200 epoch using different learning rate. Four figures represent learning rate of 0.001, 0.005, 0.01, 0.001. Figure 1 means using a constant 0.001 learning rate throughout the training process. Auto-decay means during the training process the learning rate will drop to 20% per 50 epoch.

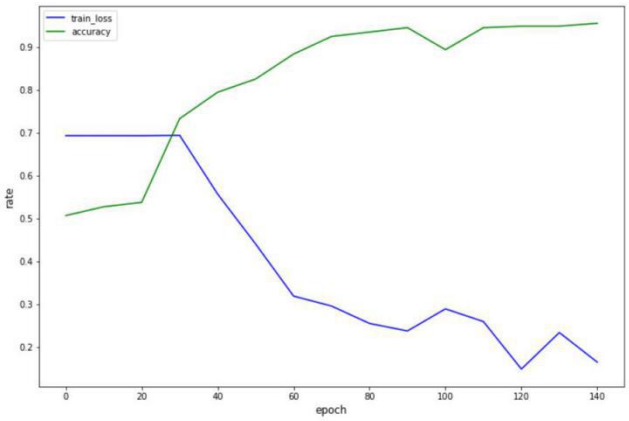

Figure 2.3.1 Plot of Learning rate = 0.001

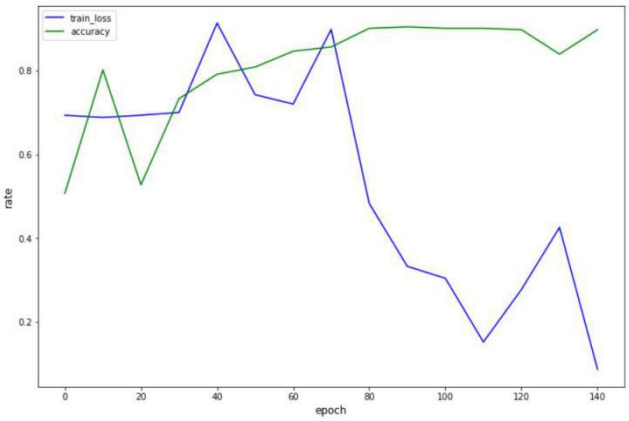

Figure 2.3.2 Plot of Learning rate = 0.005

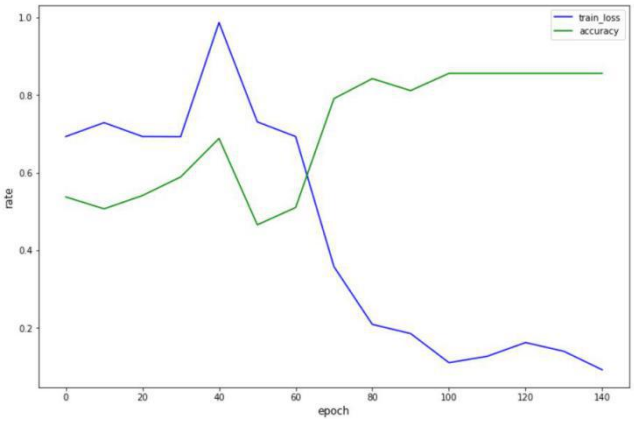

Figure 2.3.3 Plot of Learning rate = 0.01

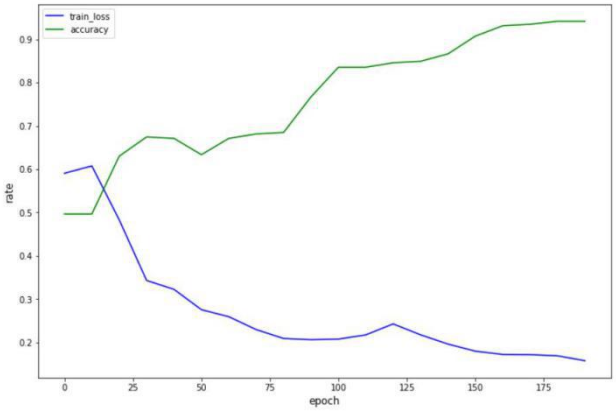

Figure 2.3.4 Plot of Learning rate = 0.001 with auto-decay

It can be discovered from these four charts that the model with a learning rate of 0.005 & 0.01 is obviously less stable than 0.001 in terms of train_loss, the fluctuation range is particularly large, and the final stable accuracy is also less than that with 0.001. The reason is that the learning rate is too large, which makes it difficult to converge. Compared with Figure 1, the one with auto-decay is more stable in terms of train_loss, but final accuracy little less than Figure 1and final loss is greater, which means the fourth model is not converged. So I increase the number of epoch to 200 and choose the fourth hypeparameters for the model.

3 Result

3.1 Results & Analysis

The confusion matrix of the NN:

Actual \ Predicted	Genuine	Posed
Genuine	120	46
Posed	45	129

Table 3.1.1 Confusion Matrix of Training

Actual \ Predicted	Genuine	Posed
Genuine	27	7
Posed	6	20

Table 3.1.2 Confusion Matrix of Testing

	Accuracy	Recall	Precision	F1 score
Training data	73.2%	72.7%	72.3%	72.5%
Testing data	78.3%	81.8%	79.4%	80.6%

The confusion matrix of the LSTM:

Actual \ Predicted	Genuine	Posed
Genuine	136	8
Posed	3	145

Table 3.1.3 Confusion Matrix of Training

Actual \ Predicted	Genuine	Posed
Genuine	47	2
Posed	3	46

Table 3.1.4 Confusion Matrix of Testing

	Accuracy	Recall	Precision	F1 score
Training data	96.2%	94.4%	97.8%	96.1%
Testing data	94.9%	95.9%	94%	94.9%

The performance of the LSTM is incredible. The accuracy reaches 95% which is the same as Chen[3]. The performance of the other aspects , Recall, Precision & F1score are very close, all around 95% . The three-layer LSTM model can accurately detect human angry expressions. I think that both the recall rate and the precision rate are very important in detecting angry expressions. Although the F1 score has become the main standard for measuring the quality of the model, which means this model has a certain ability to accurately identify anger through the F1 score of the testing data. Detecting someone is not angry is the same important as detecting anger, and comparing to the human verbal response, which has only 60% accuracy [3] my method perform much better than human being in detecting anger area. To compare the results between NN & LSTM, no matter which aspect, based on the results, the performance of LSTM is better than the ordinary NN model. Below is a histogram comparing the two results. I set the ylim from 60 to 100, 60 is the accuracy of participant.

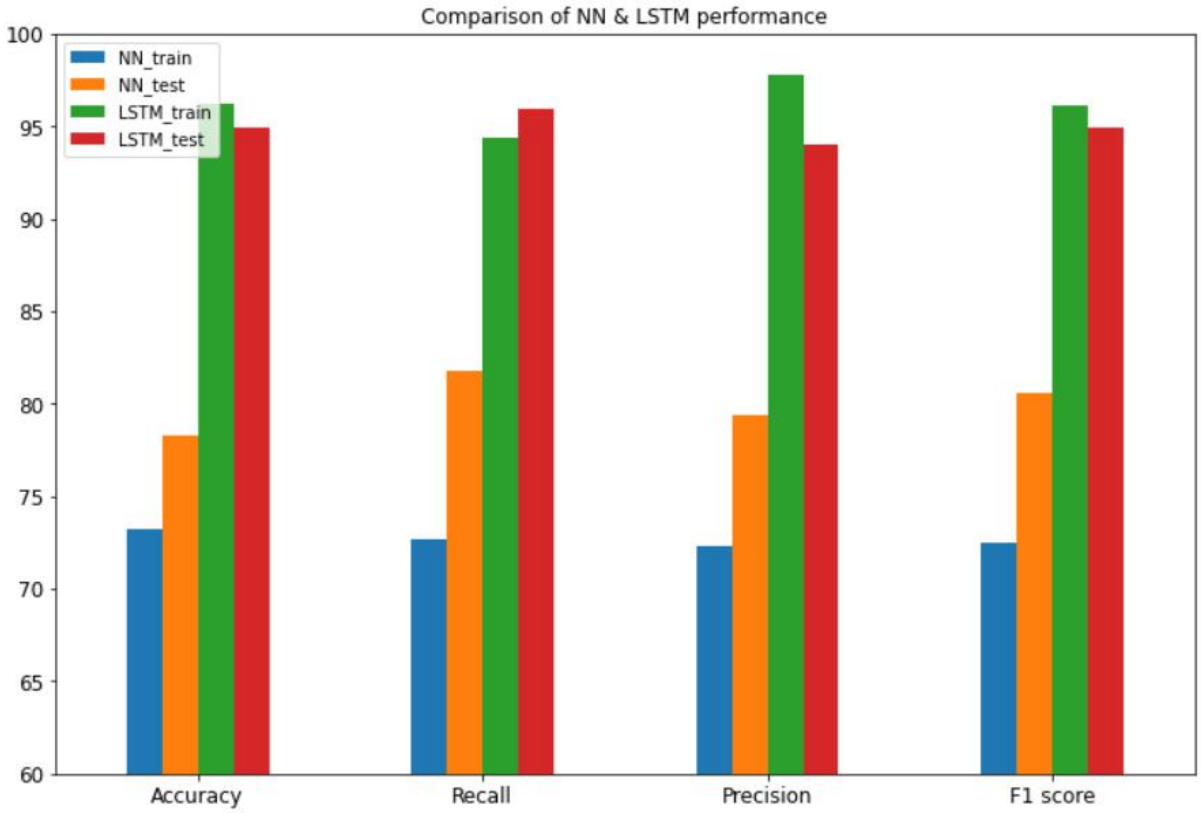

Figure 3.1.1 Comparison of NN & LSTM performance

4 Discussion

Raphael Féraud & Fabrice Clérot [6] argues that the neural network has excellent modeling capabilities. Both model have a great performance on detecting anger through pupil data, especially LSTM. Firstly, compare the RNN and ordinary NN models from a structural perspective. Different from NN in RNN, the output of hidden neurons can act on itself at the next timestamp, that is, the input of the i-th neuron at time m, in addition to the output of the (i-1) neuron at that time Including its own output at (m-1) time. If a person's pupil changes with time under anger, there is a certain correlation, then the RNN model can accurately identify anger. In addition, The structure of LSTM is more complicated. In order to remember the long-term state, LSTM adds an input and an output on the basis of RNN. The added path is the cell state [7]. So in theory LSTM should performs better than ordinary NN for this case, as the data is based on the timeline. Secondly, according to the actual results, the ordinary NN model can detect human anger with an accuracy rate of 70% to 80%, while the LSTM can reach to 95%. LSTM has more accurate detection function than ordinary NN, but at the same time it requires higher hardware. Faced with a same size dataset and the number of epochs is only one third of NN, LSTM has spent several times training time. It is because that LSTM has a more complicated structure.

5 Conclusion & Future work

In summary, my study showed that machine can detect human anger in a high accuracy by analyzing their pupil information. In addition, my research find the different performance between ordinary NN and LSTM for this case. The result charts shows that LSTM has much better detection accuracy, which means that LSTM has a good performance in predicting data based on time sequence. For future work, I assume that since LSTM has a good performance on data based on chronological order, whether it will also perform well on data based on logical order. So I may still do related

research about the performance of LSTM on data which based on logical order, in the meantime, compare the results between other models and LSTM.

Reference

[1] Oliva, M., & Anikin, A. (2018). Pupil dilation reflects the time course of emotion recognition in human vocalizations. *Scientific Reports, 8(1)*, 4871-10. doi:10.1038/s41598-018-23265-x

[2] Kret, M. E., Roelofs, K., Stekelenburg, J. J., & de Gelder, B. (2013). Emotional signals from faces, bodies and scenes influence observers' face expressions, fixations and pupil-size. *Frontiers in Human Neuroscience*, 7, 810. doi:10.3389/fnhum.2013.00810

[3] Chen, L., Gedeon, T., Hossain, M., & Caldwell, S. (2017). Are you really angry?: Detecting emotion veracity as a proposed tool for interaction. Paper presented at the 412-416. doi:10.1145/3152771.3156147

[4]Gedeon, T. D., & Turner, S. (1993). Explaining student grades predicted by a neural network. Paper presented at the *, 1* 609-612 vol.1. doi:10.1109/IJCNN.1993.713989

[5] Turner, H., & Gedeon, T. D. (1993). Extracting Meaning from Neural Networks. In *Proceedings 13th International Conference on AI* (Vol. 1, pp. 243-252).

[6] Féraud, R., & Clérot, F. (2002). A methodology to explain neural network classification. Neural Networks, 15(2), 237-246. doi:10.1016/S0893-6080(01)00127-7

[7] M. Sundermeyer, H. Ney and R. Schlüter, "From feedforward to recurrent LSTM neural networks for language modeling," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, (3), pp. 517-529, 2015.