Human Genuine or Acted Anger Recognition through Neural Network and LSTM and Results Evaluation Through Casual Index and Characteristic Input Pattern

Jiaqi Zhang

Research School of Computer Science
Australian National University,
Acton ACT 2601
u6089193@anu.edu.au

Abstract. Facial emotion plays an important role in human communication. However, human facial expressions are sometimes deceptive, and it is very difficult to accurately identify the authenticity of facial expressions. In this paper, Neural Network (NN) and long short-term memory (LSTM) were employed to determine the authenticity of human anger by analysing their pupil size changes. Then the casual index and characteristic input pattern were engaged to explain the result of NN and LSTM. The final accuracy result for NN is 84% and LSTM is 95%. The result demonstrates that LSTM shows a significantly advantage over NN for time series data.

Keywords: Emotional Veracity, Anger Recognition, LSTM, Artificial Neural Network, Casual Index, Characteristic Input.

1 Introduction

Albert Mehrabian had inferred 7-38-55 Rule of communication: 7% verbal, 38% vocal and 55% facial [1]. More than half of communication depends on the facial liking. However, people may fake their facial expressions deliberately to cover up their true emotions and intentions. To accurately analyze people's true emotions through machine learning techniques, it is crucial to determine how to distinguish between real and acted facial expressions. This paper attempts to find out whether people's anger is intentionally manipulated or unconscious feelings through the changes in pupillary responses.

The raw data came from 22 participants' verbal responses and their pupillary changes when they watched anger videos. Among these videos, 10 were acted anger and 10 were genuine anger. To control for potential confounding factors, all videos were trimmed with similar settings so that participants would not be influenced by the environment light change [2]. The dataset contains three parts: participants' left, right and average pupillary changes when they watched each video. The most important information mainly came from pupil size change rather than its magnitude. This is because even the same person may have different sizes of pupils [2]. These bio signals may be far more reliable than human consciousness. Recent studies show that the accuracy of human verbal responses is only about 60%. However, predicting the authenticity anger using biomarker through building machine learning models provide 95% accuracy rate [2]. In this paper, two machine learning techniques, the neural network (NN) and long short-term memory (LSTM), would be introduced to predict whether participants can really be distinguished between genuine anger and acted anger.

The goal of this experiment is to analyze whether the anger of the participants is real or performed while they were watching videos. This is a binary classification problem. The paper is organized as follows. At first, this paper introduces a three-layer neural network method. The traditional neural network provides ability to learn and model non-linear and complex relationship, with advantages of high running efficiency and high resistance to noise. However, the neural network's "black box" feature directly causes the non-interpretability of the result. For example, when an image of a car is the input of a neural network, the output is a cat. It is hard to understand why NN yields the result with two non-related items. Despite of the regularization of neural network avoiding some potential overfitting problems, it would also decrease the prediction accuracy and reduce the robustness of neurons [3]. To explore the influence of features on the neural network, casual index and characteristic input pattern would be introduced. The method concentrates on the question that how the rate of change of the input neurons related to the output, which sheds lights on the mechanisms of how neural network generates results [3].

Despite neural network provides an acceptable result, there is a need to explore all characteristics in the dataset. Firstly, the dataset contains hidden time series because the data comes from participants watching the video on an ongoing basis. Secondly, LSTM can be used to deal with long distance dependencies (LDD) problems. Although Recurrent neural network (RNN) is also used for series data to allow network pass through information by several time stamps, some researchers have demonstrated that RNN do not work well for long-term dependency and exist potential gradient vanishing and exploration problems [4]. The reason why LSTM is more suitable for analysing human emotion problems in this paper is because of the unique structure of LSTM. It contains a cell and three kinds of gates, which are forget gate, input gate and output gate. Those cells will remember values or characteristics over an arbitrary times interval. The forget gate decides what information to discard from the cell, the input gate will decide which values to remember and update, and the output gate selects the most appropriate information and pass then to the next neural network. In this fashion, the previous important information from different time points contributes to the final prediction.

2 Method

2.1 Preprocessing data

The dataset for neural network contains 401 observations and 9 variables. The first step was to exclude the irrelevant variables. The first and second variables (ID of participants and name of Videos) and the first row (data label) were removed. Then data cleansing was conducted to minimize the influence of abnormal values: raw data was checked to correct invalid values. Besides, converting the data to the most appropriate type to ensure the data type was acceptable for building up the model, such as one-hot-encoding on categorical variables. Since this is a binary classification problem, the target variable was encoded to 0 and 1. Specifically, the representation transform method was used to concerted all 'Genuine' and 'Posed' for the variable 'Label', which initially a string type, to integer type '0' and '1'. All features were standardized to avoid gradient explosion or gradient vanishing problems and to ensure the fast convergence of gradient descent in backpropagation of the algorithm. After all these steps, the data has 400 observations and 7 variables. Then, the data was randomly split into training set and testing set with 80%: 20% proportion. Stratified train-test split method was used to preserve the representations of each class.

The dataset for LSTM includes information of left and right pupil size of participants when they were watching videos. Initially, variables with 80% missingness or more were discarded. For variables with missingness rate less than 80%, one-dimensional interpolation was used to impute missing values. Pupil data changes are attributed to one factor rather than multiple factors. Thus, one-dimensional interpolation is more appropriate than multidimensional interpolation. Besides, Besides, cubic spline interpolation was used to get more accurate revisions for missing values. Compared to linear, zero, slinear and quadratic methods, cubic method provides the smoothest filling for missing values and yields more reliable results. The comparison of different interpolation methods is shown in Figure 1.

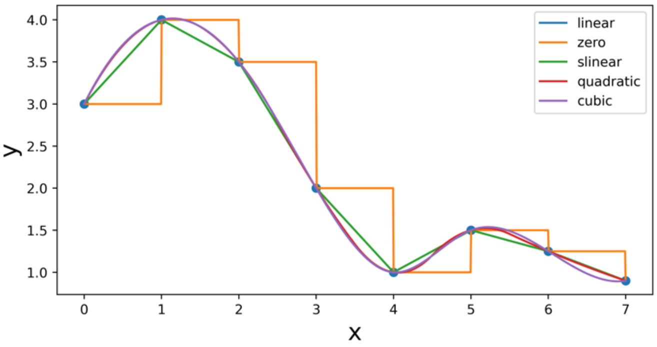


Fig. 1. Comparison of interpolation methods

However, cubic spline interpolation has potential Runge phenomenon, which will lead to high deviation problem. To overcome such limitation, I used the piecewise method was used: the cubic spline interpolation was used for different target classes separately. Then the dataset for left and right pupil datasets were merged. Features were normalised to ensure the speediness of gradient descent optimization. The sheet name in the dataset was also converted from string type, such as, 'T1', 'T2', and 'F2', to integer values. The sheet names include 'T', which represents genuine anger, these are converted to integer 0; All other names, representing fake anger, were converted to 1. The target is also corresponded to neural network dataset. Finally, irrelevant variable (ID of participants) was removed. The data was then split it to 80% training set and 20% testing set.

2.2 Three-layer neural network

The three-layer neural network consists of an input layer, two hidden layers and an output layer. All these layers are fully connected. To improve the accuracy of the neural network, choosing the most appropriate hyper parameters is crucial. Firstly, the number of input and output neurons is clear, which corresponds to 6 features and 2 classes. To find the most appropriate number of neurons for hidden layers, the rule of thumb and the equation in Figure 2 were combined [5], which yielding a result of 13 neurons for both hidden layers.

$$N_h = \frac{N_s}{(\alpha \cdot (N_i + N_0))}, where \left\{ \begin{matrix} N_s : Number\ of\ samples\ in\ training\ data\ set \\ N_i : Number\ of\ input\ neurons \\ N_0 : Number\ of\ output\ neurons \\ \alpha : An\ arbitrary\ scaling\ factor\ range\ [2, 10] \end{matrix} \right\}$$

Fig. 2. Hidden neuron equation

The activation function used for hidden layers is ReLU. Compare with sigmoid and tanh, ReLU has several advantages: (1) ReLU speeds up the backpropagation process; (2) ReLU is computational cheaper than sigmoid and tanh that involve exponential calculation; (3) ReLU's non-saturated property [6] can avoid the gradient vanishing problem which enable fitting a deeper network. As Bridle provided in [7], SoftMax is used in multiclass classification problem and is not necessarily used when the problem is binary classification. Instead, sigmoid is used in the output layer to get the activation of the classification. Learning rate and number of epochs are decided by the rule of thumb which based on grid search to find the optimized combination of hyperparameters. The loss function for classification problem in NN is cross entropy. Adam is used as the optimizer since it can automatically adjust the learning rate for each weight in every layer. With an optimizer of Adam in complex networks, the convergence speed is way faster, and the performance is better [8]. Finally, using L2 regularization is used to avoid large fluctuation and overfitting during model training. Additional techniques were also used to increase the performance of the neural network, such as dropout and early stopping. The main idea of dropout method is randomly dropping out neurons to fit a wider and deeper network and, at the same time, avoid overfitting. As for early stopping technique, it specifies that stopping adding more layers when the performance score of the evaluation set stops increasing. Considering the neural network result would be analyzed using casual index and characteristic input pattern, all technologies used in neural networks are designed to increase NN robustness while minimizing the significant impact change on input and output neurons

To summarize, the hyper parameters chosen by grid search are as follows. Neurons of hidden layers:13; learning rate: 0.01; weight decay: 0.001, and epochs: 200.

2.3 Casual index and Characteristic input pattern

Before the explanation of the output for the neural network, the casual index was calculated first. According to the algorithm provided in [3], the casual index is the rate of change of output neuron with respective to the input neuron. The Figure 3 shows all features rate of change. Since the neural network includes 6 different features, the casual index for each feature should be calculated separately. The characteristic input pattern was constructed based on the arithmetic mean of the vector components. The patterns can be named as: "ON Genuine" and "ON Posed". However, it is important to note that, even the problem in this study is a binary classification, the pattern cannot simply be defined as "ON Genuine" and all other not "ON Genuine" pattern as "Posed". The reason is that the characteristic input pattern and casual index are mainly used to explain the neural network results, rather than giving a precise prediction result. Besides, both "genuine" and "posed" may depend on same or different input features characteristic. Therefore, performing separate analyses on them can ensure the integrated rules to be complete and reliable.

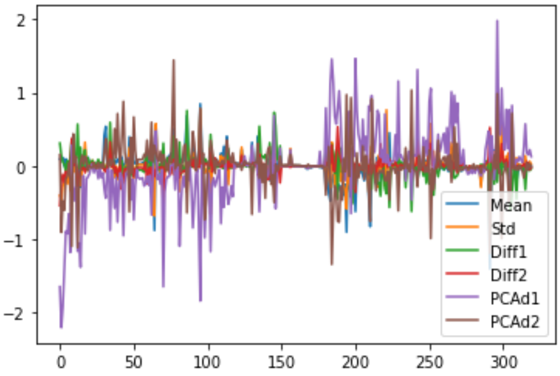


Fig. 3. Neuron rate of change

2.4 LSTM

Since the LSTM handles time series data, it mainly contains three important features: sequence length, time step and input size. The sequence length in the current dataset is 186. Thus, for time step and input size can choose from following combinations: (1, 186), (3, 62), (6, 31). After several comparison and testing, in balancing accuracy and computation costs the final choice of them is 6 and 31, separately. The LSTM also uses Adam optimizer and cross entropy loss function. Despite the advantages previously described, choosing these hyper parameters allows a better analytical comparison of the results of casual index and characteristic input pattern on f NN and LSTM. The batch size is 32 and learning rate is 0.001 which come from the performance of average of ten tests on the data. The result was shown in Figure 4.

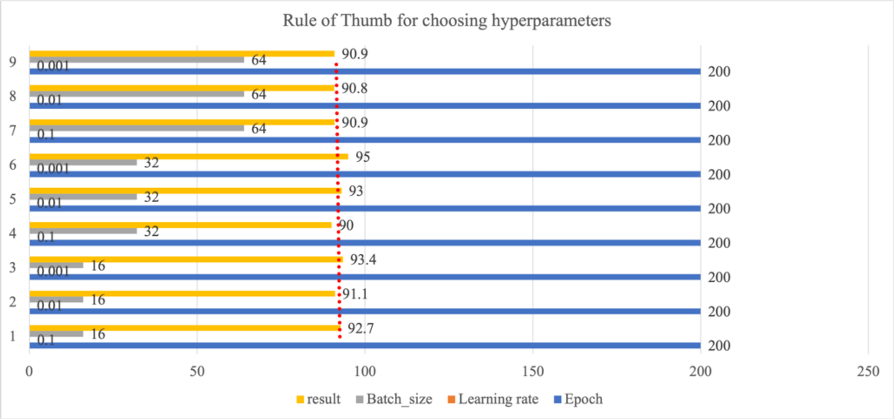


Fig. 4. Comparison of batch size and learning rate result

To sum up, the hyper parameters for LSTM are as follows. Input sequence: time step is 6 and input size is 31. Input layer neurons: 31. Hidden layer neurons: 62. Output later neurons: 2. Learning rate: 0.001; Batch size: 32. Epochs: 200.

3 *Result and Discussion*

3.1 *Experiment environment*

Operating system: macOS Big Sur 11.2.3
PyCharm Version: 2020.2.1
Python version: 3.8.5
PyTorch version: 1.8.1
CPU: Intel Core i9 2.3GHz 8 Cores
RAM: 32GB

3.2 *Neural Network*

Theoretically, 400 observations with 6 features are too small for a neural network to perform adequate training. The lack of data may cause the following problems: (1) the model cannot converge very well since the data does not provide with enough information about the underlying pattern; (2) potential underfitting problem. These may lead to a low accuracy score on both training and testing set. The result of the neural network also corresponds to my prediction. The accuracy result in training set is 94% and in testing set is 84%. However, the loss function in Figure 5 shows that even in 200 epochs the model still does not got converged. By increasing the number of epochs, I found the model converged in around 300 epochs, but in this case, the overfitting problem raised. The dilemma here is increasing epochs would cause overfitting but decreasing epochs will cause model convergence and underfitting issues. The reason here are: (1) features and data are insufficient; (2) noises in the data influenced model training. The inference for reason 1 is that, even the data includes 6 features with 400 observations, actually these data only came from 22 participants. And the target is to determine participants' anger is genuine or acted, so that the 'core' training data is only 154 (22 participants * 7) rows. The logic behind reason 2 is that, even all videos were trimmed to prevent the influence of environment light, if participants did not watch the video in a dark room, the raw data still had inevitable noise. In other words, the dataset did not provide enough confounding variables that can be used for the accurate prediction, which leads to the neural network picks up spurious relationships to make the prediction instead of the truly underlying patterns that we are interested in [9].

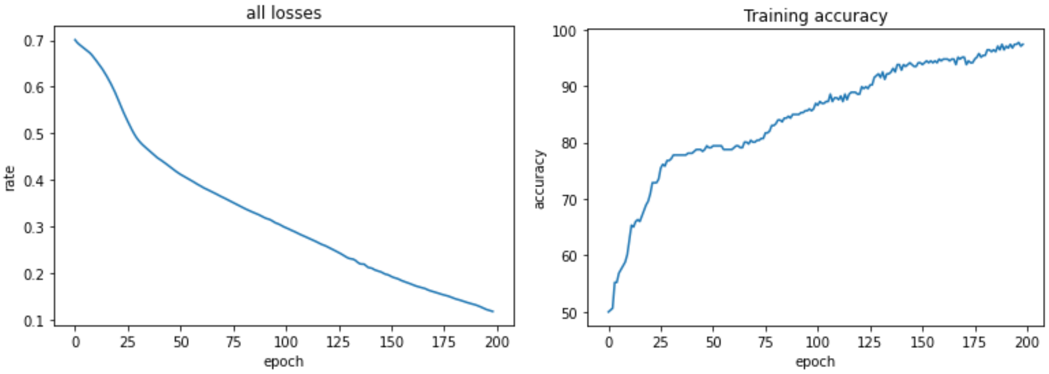
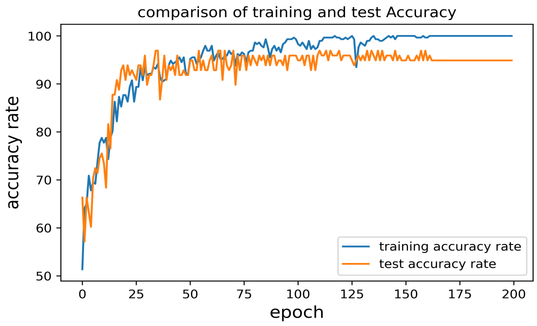


Fig. 5. Training losses and accuracy in neural network

3.3 LSTM

The LSTM results are consistent with my initial hypothesis. The accuracy result in training set is 99% and in testing set is 95%, as shown in Figure 6. The loss function in Figure 7 shows that the model converged very well. Therefore, for time series data, LSTM has a much better performance than neural network. In contrast to traditional neural networks LSTM has no fully connectivity between neurons. The LSTM provides a unique cell and three gates, forget gate, input gate and output gate. As information enters the LSTM model, the cell will make a judgement on the information and those that match the rules would be remained and those that doesn't, would be forgotten. This is also more in line with the thinking pattern of the human brain. Within the nervous system, memory is stored by the formation and disappearance of synapses in a number of brain regions. Most problems faced by humans contain complex temporal correlations and cause-effect relationships, while the traditional neural network lack the high dimension of information. But again, this explains why casual index and characteristic input pattern are more suitable for interpreting the results of neural network rather than LSTM. The long-dependency memory of the LSTM makes the rate of change of input and output neurons less obvious, which directly leads to the difficulty in computing casual index weights.



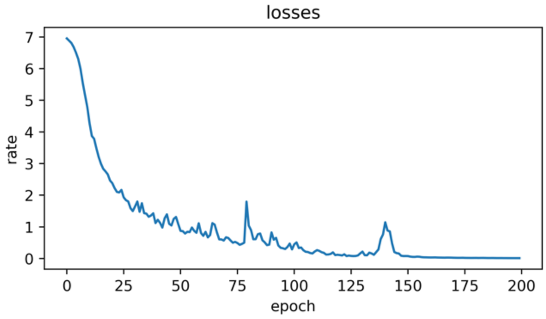


Fig. 6. LSTM accuracy result

Fig. 7 LSTM training losses

Human self-perception is only 60% accurate, which is much lower than NN and LSTM. Deep learning provides results that are much higher than human recognition [2]. This technique has a wide range of applications, for example, lie detection in psychology and automatic classification of the feelings expressed in videos in social media. The results lead to the hypothesis that the human subconscious unconsciously masks the expression of their true feelings, possibly as a self-protective mechanism. However, physiological expressions such as pupil contraction and sweaty palms are not hidden, cannot 'tell the lie'.

4 Conclusion and Future Work

This paper concludes the result of using the neural network and LSTM in predicting human anger authenticity. Also, we used the casual index and characteristic pattern on both neural network and LSTM to explain the result. On the one hand, these comparisons proves that neural network result is more appropriate to evaluate by casual index and characteristic pattern. On the other hand, casual index and characteristic pattern has potential limitations in explaining LSTM results. Complex cell and gates mechanisms can lead to insignificant rate of changes in input and output neurons. Besides, for the emotional veracity prediction of time series data, the LSTM has much better performance than traditional neural network.

The future exploration will include three parts. First, for neural network, we can use some more advanced techniques, such as feature extraction, feature fusion and data augmentation to minimize the effect of small sample size. Second, for LSTM, we can update the LSTM with an attention model. The attention model that stores more memory and selects the information that is useful for the current decision from a large amount of input information [10]. Thirdly, we will try to employ the casual index and characteristic pattern for LSTM context. The applicability of the casual index and characteristic pattern method can be further improved.

References

[1] Lapakko, D. (2009). Three cheers for language: A closer examination of a widely cited study of nonverbal communication. Communication Education, pp. 63-67.

[2] Lu Chen, T. G. (2017). Are you really angry? Detecting emotion veracity as a proposed tool for interaction. Australian Conference on Human-Computer Interaction (OzCHI '17), (p. 5). Brisbane, QLD, Australia.

[3] T. D. Gedeon, H. S. (1993). Explaining student grades predicted by a neural network. Proceedings of 1993 International Joint Conference on Neural Networks. P.O. Box 1, Kensington.

[4] Rafal Jozefowicz, W. Z., I. S. (2015). An Empirical Exploration of Recurrent Network Architectures. 32nd International Conference on Machine Learning, pp. 2342-2350

[5] Ning Liu, A. I. (2015). Diketopyrrolopyrrole (DPP) functionalized tetrathienothiophene (TTA) small molecules for organic thin film transistors and photovoltaic cells. Journal of Materials Chemistry C.

[6] A Krizhevsky, I. S. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Advances in neural information processing systems, pp. 1097-1105.

[7] Bridle, J. S. (n.d.). Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition. Neurocomputing (pp. 227-236). Springer, Berlin, Heidelberg.

[8] Rafi, U. L. (2016). An Efficient Convolutional Network for Human Pose Estimation. BMVC , 2.

[9] Hawkins, D. M. (2004). The Problem of Overfitting. Chemical Information and Computer Sciences, pp. 1-12.

[10] Zhou Xinjie, X. W., J. X. (2016). Attention-based LSTM network for cross-lingual sentiment classification. Proceedings of the 2016 conference on empirical methods in natural language processing, pp. 247-257