Detecting emotion by LSTM

Haonan Zhang

Research School of Computer Science
Australian National University
Acton ACT 2601 Australia
u6885863@anu.edu.au

Abstract. In the past, we judged people's moods by their descriptions which are easily influenced, so the accuracy is relatively low. In this article, we introduce a new method to judge people's emotion through their pupillary responses. We have built a neural network to classify whether a person is angry which the LSTM(Long Short Term Memory networks) layers and FCN(full connected network) layers. Finally, we compare the result of different hidden layers and get the ideal result.

Keywords: Pupillary Response · Genuine Posed Anger · Long Short Term Memory Networks.

1 Introduction

We used to judge people's moods by their descriptions. However, the verbal response is influenced by many factors. The result showed that the accuracy of human verbal response is just 60%. We desires to find another effective method to detect emotion more accurately. Lu Chen [2] focused on pupillary responses and proved them effective. Comparing with the verbal response, the pupullary responses make a huge improvement on emotion classifications, with raise the accuracy from 60% to 95%. This research shows that machines can reach high accuracy in distinguishing acted and genuine anger by using physiological signals,.
In our experiment, we collect 20 participants' pupillary responses in viewing two types of anger stimuli. Each type of anger stimuli contains 10 kinds of videos, so the total dataset has 200 positive samples and 200 negative samples. This is a typical classification problem with only two labels. We used artificial neural network technique to deal with this problem and compare with different techniques to filter dataset.

2 Dataset description

The dataset, as the description above, contains 200 positive samples and 200 negative samples. Each sample contains each participant's pupillary responses on one type of anger stimuli. We used pupil size to quantify pupillary responses. Firstly, we set 145 recorded frame to record instantaneous pupil size at each recorded frame. Then we replaced those zero values with linear interpolation between the nearby nonzero values, because participants occasional close eyes resulting zero values.
In Figure 1,we can observe that these two curves have obvious distinction that the pupil sizes of acted anger stimuli are usually bigger than the pupil sizes of genuine anger stimuli. In our experiment, we gather 400 kinds of sequential data to deal with this classfication task. In Figure 2, we can see that the labels are balanced, so we do not change the ratio of the samples.

2.1 Data Processing

For each sequential data, we have two two dimensions of information: the left eye's size and the right eye's size. Because my data is not very complete, for example, we have many missing values which we might not measure and some zero values where the participants closed their eyes, we should pre-process the sequential data before. For those data that contain too many missing values of zero values, since they have little useful information, we have to abandon them. For those data that contain a few missing values of zero values, we should use interpolation to replace them. In practical, we replaced the missing values with the previous value of it.
However, there is another question we should solve. The videos have different durations, so our sequential data have different length. While the neural network require the data have the same length. Thus, we should re-arrange the data into the same length. In practical, we assume that our data is producted by a unknown function $f(t)$, where t

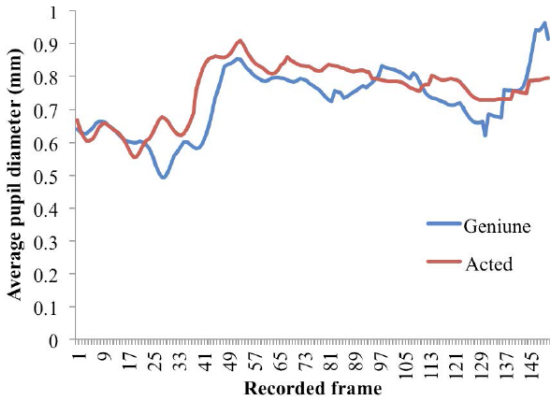

Fig. 1. Pupillary response to genuine versus acted anger stimuli

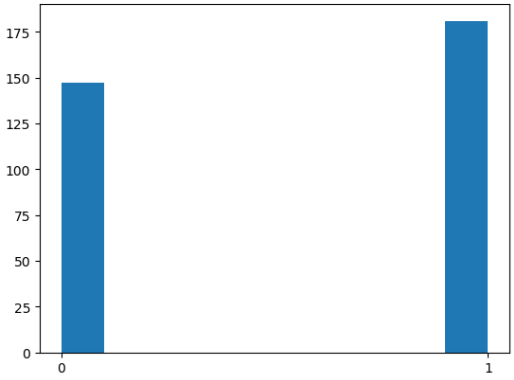

Fig. 2. Distribution of the labels

is a discrete independent variable like $0, 1, 2...$ and $f(t)$ is the obversed value at t time. After we fitted this function, we choose 100 evenly spaced points from 0 to $t_{\max}$, and then we would get a new sequential data with length of 100.

2.2 Analyze features

Before we contrust the neural network structure, we often qualitatively analyze each feature. Typically, we draw the curve of each feature with different labels. In Figure 3, we can see that there is an obvious difference between Genuine label and Posed label. Both eyes' size with Genuine label are significantly greater than those with Posed label. Otherwise, there are also subtle differences between left eyes and right eyes with the same label. The mean and variance of different eyes are distinctive. Thus, before we put the data into the neural network, we should normalize them. In practical, we let each data minus its mean and then divided by its standard deviation and then put these normalized data into the neural network.

3 Neural network structure

A neural network commonly contains an input layer, hidden layers and an output layer. Because the input data in our task is the sequential data and the size of our dataset is $400 \times 100 \times 2$. We connect the LSTM(Long Short Term Memory networks) layer [3] with the input layer and then connect a FC layer after the LSTM layer Otherwise, we are doing with a classification task with two labels, so the dimension of the output layer is one. The loss function we choose is cross entropy [6]. Our main task is to adjust parameters in the hidden layers and network structure.

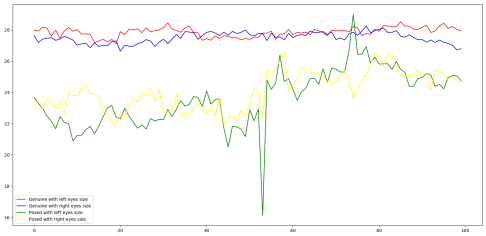

Fig. 3. Compare two eyes size with different labels

For the LSTM layer which has the shown in Figure 4. We have three gates in this structures: the forget gate, the

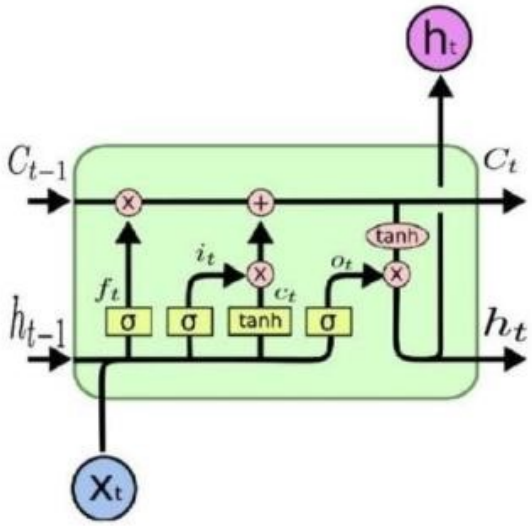

Fig. 4. Neural network structure

input gate, and the output gate. I will discuss them in detail following.
The first step in LSTM is to determine what information the cell state needs to throw away.This is handled through a sigmoid unit called the forget gate. It sends a value between 0 and 1 due to h_{t-1} and x_t to control how much information C_{t-1} should abandon like Equation 1.

$$f_t = \sigma(X_f[h_{t-1}, x_t] + b_f) \tag{1}$$

The next step is to decide what new information to add to the cell state. First, h_{t-1} and x_t are used to determine which information to update through an operation called an input gate. The new candidate cell information $\tilde{C}_t$ was obtained through a tanh layer using h_{t-1} and x_t, which may be updated to the cell information like Equation 2 and Equation 3.

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{2}$$

$$\tilde{C}_t = \tanh(W_C[h_{t-1}, x_t] + b_C) \tag{3}$$

Next, the old cell information C_{t-1} will be updated to the new cell information C_t.The updated rule is to select a part of the old cell information to be forgotten by the forget gate, and a part of the candidate cell information to be added by the input gate to get the new cell information C_t like Equation 4

$$C_t = f_t C_{t-1} + i_t \tilde{C}_t \tag{4}$$

After the cell state is updated, you need to determine the status characteristics of the output cell, which needs to be evaluated by a sigmoid layer called the output gate. Then the cell state is obtained by the tanh layer which is the final output.

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{5}$$

$$h_t = o_t \tanh(C_t) \tag{6}$$

After the LSTM layer, we build a FCN(full connected network) layer, where we compute a linear combination of the previous data and then use an activation function to do the unlinear mapping. Put the result as the input data in the next layers and repeat above steps again and again. In our experiment, we choose ReLU function as the activation function in the all hidden layers.
Otherwise we should decide how wide and deep our neural network is. Typically, we ask the dimension of our network to decrease, as it become deeper, which means that the number parameters of the current hidden layer should be less than the previous one. Although we have chosen the unsaturated activation function 'ReLU', gradient vanishing will also occur when the network become deeper. Thus, we add a Batch Norm layer [5] after each hidden layer to prevent from gradient vanishing.

$$a_i \leftarrow \frac{x_i - \mu}{\sigma} \tag{7}$$

In Equation 7, a_i is denoted by the output of the Batch Norm. x_i is denoted by the input of the previous hidden layer. μ is denoted by the mean of the previous hidden layer. σ is denoted by the standard deviation of the previous hidden layer. This layer can prevent each input for the next hidden layer is a standard data of mean equal to zero and variance equal to one. So The final network structurce sketch is shown in Figure 5.

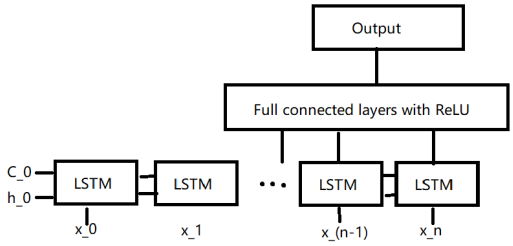

Fig. 5. Neural network structure

4 Experiment and Comparison

As we all know, the LSTM layer is not the only method to deal with the sequential data. RNN layers [1] are often used in this kind of task. In this section, we will compare the tendency and effect with different hidden layers. Although, these two kind of loss functions will finally all converge, we can see that the loss function with LSTM layer is more stable. Thus although we don't adjust parameters, we will also get a ideal result. The 5-fold cross validation has been shown in Table 1.The LSTM has a tiny advantage with the RNN. The final accuracy is around 93%, which is tiny less than the accuracy in [2].

Table 1. Comparing accuracy with LSTM and RNN in 5-fold cross validation

Name	1	2	3	4	5	average
RNN	0.83	0.94	0.89	0.95	0.86	0.89
LSTM	0.89	0.92	1.00	0.85	0.87	0.90

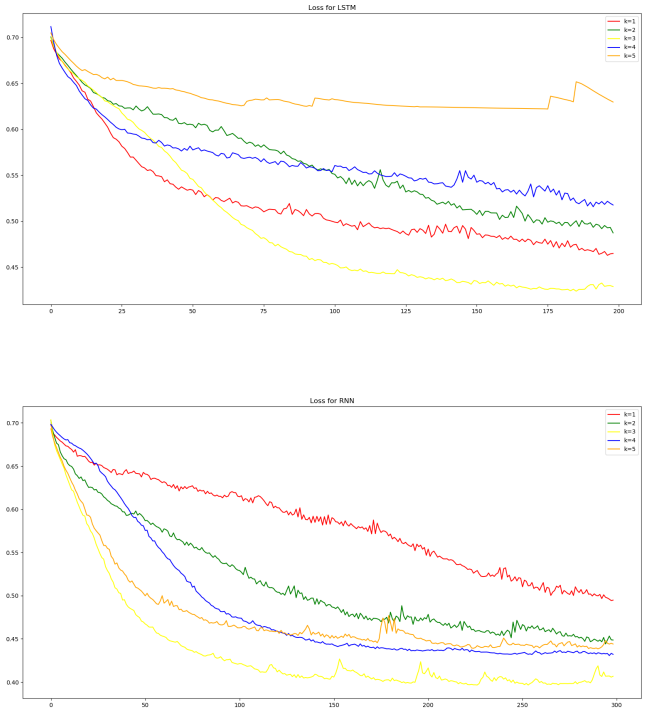

Fig. 6. Comparing loss with LSTM and RNN in 5-fold cross validation

5 Conclusion and Future Work

In our experiment, we trend to judge people's emotion through their pupillary responses. We used the neural network to deal with the classfication task, whether a person is angry. Then, we explored how to how to get useful information from the sequential data and it seems to be effective. Finally, the accuracy of our model is 93% which is similar with Lu Chen's [2]. it is far higher than the accuracy of the verbal responses. It shows that pupillary responses is a good way to judge people's emotion.

In the future, we will improve this model in two main parts. Firstly, We will recruit more volunteers. The more samples, the more robustness the model has. Thus, we want to enrich our train data as soon as possible. In another part, we will improve more complex network structure. Recently, many new network structures have been raised like self-attention layers [4] for sequential data. We want to explore a better method to improve the performance.

References

1. Buber, E., Diri, B.: Web page classification using rnn. Procedia Computer Science **154**, 62–72 (2019)
2. Chen, l., Gedeon, T., Hossain, M., Caldwell, S.: Are you really angry?: detecting emotion veracity as a proposed tool for interaction. pp. 412–416 (11 2017). https://doi.org/10.1145/3152771.3156147
3. Ordóñez, F.J., Roggen, D.: Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. Sensors **16**(1), 115 (2016)
4. Wang, Y., Liu, Y., Ma, Z.M.: The scale-invariant space for attention layer in neural network. Neurocomputing **392**, 1–10 (2020)
5. Wu, S., Li, G., Deng, L., Liu, L., Xie, Y., Shi, L.: L1-norm batch normalization for efficient training of deep neural networks. IEEE Transactions on Neural Networks and Learning Systems **PP** (02 2018). https://doi.org/10.1109/TNNLS.2018.2876179
6. Xie, Z., Huang, Y., Zhu, Y., Jin, L., Liu, Y., Xie, L.: Aggregation cross-entropy for sequence recognition. pp. 6531–6540 (06 2019). https://doi.org/10.1109/CVPR.2019.00670