The Application of Long Short-Term Memory (LSTM) Network in Detecting Genuine/Posed Anger

Zhetao Zhong[1]

Australian National University, Acton ACT 2601, Australia
email: zhetao.zhong@anu.edu.au

Abstract. Neural Networks in one of the powerful tools in data science which is widely used in processing data. Among the variants of architectures and structures in different areas, Long Short-Term Memory (LSTM) is a kind of artificial Recurrent Neural Network (RNN) architectures widely used in deep learning area.

In this work, I constructed an artificial neural network using LSTM and linear neural network to genuine or posed anger. Data used in this work is the size of both left and right eye's pupil (all numeric data). This time series data is captured during the previous experiment. Packed in 2 different files, one for each side, the given data tells how participant's pupil sizes change among the time as well as whether the anger shown is genuine or fake.

Although the result of this implementation is not as impressive as expected and should have a long way to improve, the implementation itself shows a possible method to detect genuine or posed anger using the model proposed in this work. For wider usage, this architecture can be modified to work with other kinds of time series data.

Keywords: LSTM Network · Neural Networks · Anger Recognition · Machine Classification.

1 Introduction

Facial expressions play an important role in social interactions. As a central part of human communication [1], they provide significant help in understanding each other as well as the conversation situation because it usually reflects the internal mental state of the displayer of the emotion [2]. Sometimes understanding a facial expression wrongly can directly lead to a massive misunderstanding. In HCI area, understanding users' affects better often leads to better decisions.

One of the most common expressions, anger, is very easy to express and be understood. But here is a question, whether all the anger we found on someone's face is real? That is, a person can act or pretend to be angry and shows a fake anger expression on his/her face. We call this Posed Anger in this work. There are already some existing work in this area, capturing the difference between genuine and fake emotion even if they appear the same. Usually the genuine and fake ones have different physiological signals which can be used by machine to distinguish them [3][4].

Long Short-Term Memory (LSTM) can be used as a effective and scalable tool to train the model using sequential data [5]. It is widely used in analysing sentences (sequence of words) and capturing features from videos (continuous images). In this work, it is aiming to figure out how LSTM can be use in neural network for distinguishing fake anger. A classification model is constructed using LSTM and linear neural network and experiments are performed. The data used is participants' pupil size in time series format.

This work provides support for structuring and setting up neural networks to classify and analyse facial expressions, helpful in a variety of ways including HCI user feedback [6], handling user affects [7].

2 Related Work

A number of existing work has already been done on facial expression recognition and distinguishing real/fake expressions, using different network structure and algorithms.

[2] provided a work on detecting emotion veracity uses perceivers physiological signals to classify whether the anger is genuine or fake. Some part of the data used in [2] is highly similar to the one used in my work. In his experiment, 22 participants were required to view anger stimuli in two different types. During the viewing, experimenters recorded their physiological signals including skin conductance, blood volume pulse, heart rate, eye gaze and pupil size. They also designed several specific questions to ask after participants' viewing for testing how they recognised or were aware of the real/fake anger. In the end, it proved that machines can retrieve useful features from emotion perceiver's physiological signals, especially pupillary response, to achieve a much higher accuracy to differentiating genuine and acted anger than human beings.

Another high-accuracy system to distinguish real and fake smiles was carried out by Md Zakir and Tom[8]. It senses observers galvanic skin response (GSR, indicating electrical changes measured at the surface of human skin). They applied different combinations of feature-selecting algorithms and network structures including SVM, KNN and NN, comparing the performance. They succeeded in finishing up with the accuracy of 96.5% using simple NN network and selected features.

Likewise, [9] proposed a speech emotion classification based on LSTM architecture. They extracted speech features from the original waveform, replacing traditional statistical ones. The sequence of frames preserve the timing relations in the original speech. Then the classification is performed on the extracted frame data. Their result shows that after proper training, the LSTM-based model provides a good performance on time series data.

3 Method

3.1 Data Process

This work begins with raw data (participants' pupil size) coming from previous experiment. The data comes in 2 separate files, containing the left eye data and right eye data respectively. In each file, there are 20 worksheets, 10 of them are for positive samples and the rest are for negative ones. In every single worksheet, one column stands for one participant and the rows are time series. The time interval between two continuous rows is 1/60 second.

The first step is reading all data from the file. Only non-empty columns are used. That is, going through all the worksheets and looking for all the columns. If one of the left-eye or right-eye data is missing, then this sample is abandoned. At this moment, all data read is at the same length. However, some of the samples contains 0 which means at that time, the participant closed that eye. So the next step is removing all the zeros from the samples.

It is possible that at the same time, one of the eyes was closed but the other was not. To solve the problem, I first get the participants' left-eye and right-eye data separately then go through the time axis. A pair of data will only be used if both left and right eyes are open (non-zero). Otherwise, this single pair of data will be skipped. After this step, the data currently in storage is all non-zero.

The next problem is, pupil sizes of different persons differ. Some people have large pupils naturally when some others' are relatively small. To avoid the influence caused by nature and focus on the changing trend, all data pairs (left and right eye data) are standardised. Here a method similar to statistical standardisation [10] is preformed. All the data is calculated by (data - mean) / variance. The mean and variance are the mean value and variance value of that single list of data.

3.2 Network Setting

As stated in [5], Long Short-Term Memory is a model showing good effectiveness and scalability. Compared to normal Recurrent Neural Network, the key idea behind LSTM is a memory cell which can keep its states as the time goes. So it is more capable of processing data changing with time flows. Another key point of a LSTM cell is its gates. Normally, there are 3 gates which are input, output and forget gate.

The input gate evaluate the importance of a new input and update the cell in a selective way. That is, it will select important information to update. The forget gate 'forgets' in a selective way. It decides what should be forgot from the previous state and what should be kept. The output gate is in charge of deciding what in the memory cell is going to be the output. Because of these improvement, the LSTM should perform better in avoiding gradient vanishing and exploding.

My model starts with a 2-layer LSTM with hidden size of 2 after which is 2 full-connected linear layers whose sizes are 32-16-2. As we are doing binary classification, an output with width of 2 should be the chosen. The reason why the input size of linear layers is 32 but the hidden size of LSTM is only 2 will be discussed later. The activation function used between linear layers is ReLU. Loss function applied here is cross entropy loss which is a common usage in calculating probabilities.

The feeding of my training is one by one. That is, the time series data is fed in to the model one sentence after another. One sentence is just one piece of information which contain a sequential length list of data pairs. A pair of data is made up of the left and right pupil sizes captured at the same time (shown in the data part). For instance, a sequence of 120 pairs should be formatted into (120, 1, 2). So, the h_0 and c_0 are both in shape of (2, 1, 2), indicating that the dimension is 2, 1 sentence per time, the number of hidden layers is 2. From LSTM to linear part, a simple conversion is done. As the output shape is the same as that of input for LSTM (sequential data), I capture the last 16 pairs of output data, flatten the 16 * 2 into a 32 * 1 for formatting input to the linear layers. This is also explains why the input width is 32 for the top linear layer.

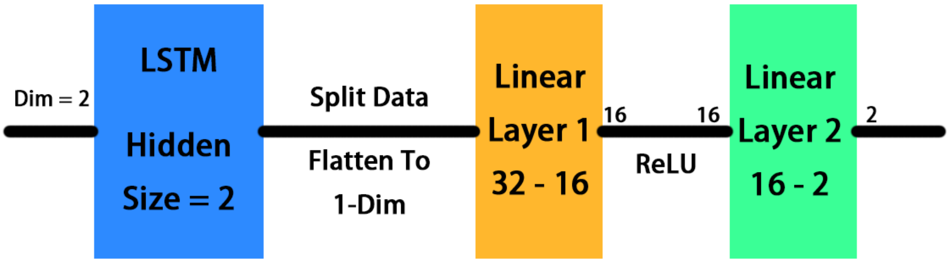

Fig. 1. The structure of LSTM classification model

In choosing the optimiser for network learning, Adam is finally used. Adam is a kind of adaptive optimising algorithm [11] for multi-layer feed-forward network. It aims to control the learning rate adaptively for different parameters, based on the statistics of gradient. This algorithm simplifies the settings and can result in faster convergence. Resilient propagation (RPROP) [12] was also considered in design but not used at last. The RPROP was also adaptive, designed to overcome pure gradient-descent's inherent disadvantages. The sign of gradient is the only thing it cares whose change is a critical for learning process control.

To increase the credibility, Cross-validation [13][14] which is one of the most widely used methods for evaluating model performance is a part of the application. Firstly, the dataset is cut into several pieces. Among them, a larger part (in this case) is used to fit into the model for training when the rest is used to measure how the model works after training. The prediction value is compared with the actual value in validation data to calculate the accuracy and to show how this model works on independent data. In this case, the model validates 5 times during the whole process. Each time, the dataset is newly splitted in random order.

4 Experiment and Comparison

Several experiments were performed with different choices of learning rate, optimiser and network setting up. Compared to the previous which was a cascade neural network built using techniques stated and recommended in [15] to perform the same classification, the result is not impressive at all. The previous model can achieve an accuracy around 85% while the current one only does 65%.

Firstly, learning rate. A number of values between 0.1 and 0.0001 have been tested. For rate near 0.1, the network updated too fast. The learning process was not stable and convergence was found really difficult. For value near 0.001, it updated too slow. It took too long to see an improvement in the model. Among all the values tested, 0.01 is the one finally used as it showed the most satisfying process and result.

Secondly, whether all sequences should have same length was tested. One option was feeding all original data sequences into the LSTM and they can be different in length. That is, the first sequence was 100-pair long and the second can be in length of 120. As LSTM is also widely used in language processing, it was thought unnecessary to control the sequence length. Another option is resampling the data and make every single sequence 100-pair long. The experiments showed that both options could converge but the one with fixed sequence length did better than variable one by 4%.

Additionally, there were 2 options of choosing optimiser as also mentioned in the method part. Using RPROP or Adam did not make a difference. Finally, Adam was used in the model.

Table 1. The result from experiments

N-th Fold Validation	Accuracy
1	58.14%
2	62.04%
3	63.07%
4	67.65%
5	59.69%
Avg	62.12%

Table 1 gives a summary of the result got from the model. It was set as: learning rate = 0.01, epoch = 300, loss function: Cross Entropy, optimiser: Adam, sequence length: 100 fixed. There's a huge gap between the performance of this model and the previous work.

Different network structures can cause the gap, but more importantly, the data used are totally not the same. Data in last work is pre-processed pupil size data, 20 entries for one video and 20 videos in total. That is 400 pieces of information. The total number of new data sequences is also near 400, but they are time series data. For a single sequence, there are more than 100 pieces of data to fit in. It needs to learn the 'pattern' rather than classification using just one line of numeric data.

One of the factors need to be improved is the pre-processing of data. It goes through some processing steps currently but not enough for the model to extract obvious patterns and find common/different trends between genuine/fake samples. These pieces of data should be more relevant to each other. Another point is enhance the network structure. An improvement to make it more powerful in extracting pattern features and capable of working on more generic data is expected.

5 Conclusion and Future Work

I introduced the application of LSTM-based neural network in detecting genuine or posed anger. Unlike the previous work, this one is fed with time series data for every single sample. Experiments were done and several different options of network setting were tested. Although the performance is not impressive, it still gives an example of using LSTM in this area.

As mentioned above, the future work will be in 2 parts. The first one is focusing more on data preprocessing, enhance the relevance between pieces of data in same format. Secondly, the network itself needs improvement, concentrating on extracting general patterns over a certain amount of data instead of something specific. Training with more sentences at the same time is also considered to be added in future.

Finally, this work is brand new to a great extent as the data is totally different as well as the training method. It is also expected to combine this one with the previous one sometime.

References

1. Mri, R.M. (2016), Cortical control of facial expression. *J. Comp. Neurol.*, 524: 1578-1585. https://doi.org/10.1002/cne.23908
2. Chen, L., Gedeon, T., Hossain, M. Z., Caldwell, S. (2017, November). Are you really angry?: detecting emotion veracity as a proposed tool for interaction. *In Proceedings of the 29th Australian Conference on Computer-Human Interaction* (pp. 412-416). ACM.
3. Masood Mehmood Khan, Robert D. Ward, and Michael Ingleby. 2009. Classifying pretended and evoked facial expressions of positive and negative affective states using infrared measurement of skin temperature. *ACM Trans. Appl. Percept.* 6, 1, Article 6 (February 2009), 22 pages. https://doi.org/https://dx.doi.org/10.1145/1462055.1462061
4. Chi Jung Kim, and Min-Hyuk Chang. 2015. Actual Emotion and False Emotion Classification by Physiological Signal. In *Signal Processing, Image Processing and Pattern Recognition (SIP '15)*, 21-24. IEEE. https://doi.org/http://doi.ieee.org/10.1109/SIP.2015.17
5. K. Greff, R. K. Srivastava, J. Koutnk, B. R. Steunebrink and J. Schmidhuber "LSTM: A Search Space Odyssey," in *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 10, pp. 2222-2232, Oct. 2017, https://doi.org/: 10.1109/TNNLS.2016.2582924.
6. Zhihong Zeng, Jilin Tu, Ming Liu, Tong Zhang, Nicholas Rizzolo, Zhenqiu Zhang, Thomas S. Huang, Dan Roth, and Stephen Levinson. 2004. Bimodal HCI-related affect recognition. *In Proceedings of the 6th international conference on Multimodal interfaces (ICMI 04).* Association for Computing Machinery, New York, NY, USA, 137143. https://doi.org/https://doi-org.virtual.anu.edu.au/10.1145/1027933.1027958
7. Eva Hudlicka. 2003. To feel or not to feel: the role of affect in human-computer interaction. *Int. J. Hum.-Comput. Stud.* 59, 12 (July 2003), 132. https://doi.org/https://doi.org/10.1016/S1071-5819(03)00047-8
8. Md Zakir Hossain and Tom Gedeon. Observers Galvanic Skin Response for Discriminating Real from Fake Smiles. *Australian Conference on Information Systems (ACIS '16)*, Wollongong, 1-8. Retrieved from: http://ro.uow.edu.au/acis2016/papers/1/33
9. Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou and B. Schuller, "Speech Emotion Classification Using Attention-Based LSTM," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1675-1685, Nov. 2019, https://doi.org/10.1109/TASLP.2019.2925934.
10. Shanker, M., Hu, M. Y., and Hung, M. S. (1996). Effect of data standardization on neural network training. *Omega*, 24(4), 385-397.
11. Z. Zhang, "Improved Adam Optimizer for Deep Neural Networks," *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*, Banff, AB, Canada, 2018, pp. 1-2, https://doi.org/10.1109/IWQoS.2018.8624183.
12. Riedmiller, M., Braun, H.: A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm. In: *IEEE Int. Conf. on Neural Networks*, pp. 586591 (1993)
13. Stone, M. (1974). Crossvalidatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2), 111-133.
14. Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American statistical Association*, 70(350), 320-328.
15. Khoo, Suisin and Gedeon, Tom. (2008). Generalisation Performance vs. Architecture Variations in *Constructive Cascade Networks*. 236-243. 10.1007/978-3-642-03040-6_29.