Classification of Human's Deceptive Response Using Neural Network and Long Short-term Memory Network

Tianai Qiu

Research School of Computer Science
Australian National University, Canberra Australia
u6744700@anu.edu.au

Abstract. There has been an increasing amount of research on thermal imaging (a.k.a. thermography) and deception detection recently. Based on minor physiological changes recorded by thermal cameras, thermal imaging can be adopted as an efficient lie detector in law enforcement and other fields. This paper introduces a neural network and an LSTM model, respectively, to classify deception responses using two versions of thermal datasets. A magnitude measure is implemented in Neural Network for feature pruning. The results show 59.33% testing accuracy in the two-layer fully connected neural network, and 77.07% testing accuracy in LSTM model, which turns out that the LSTM model performs relatively better than neural network trained in this paper. The low performance of the NN model after implementing the feature pruning technique indicates that the 1st thermal dataset used in this paper is not suitable to use this kind of functionality-based feature selecting technique. Furthermore, the high performance of the LSTM model indicates that the sequence data organized in 2nd version of the thermal dataset can effectively record deception pattern of subjects and is somewhat more suitable for classification.

Keywords: Neural network, facial thermal imaging, deception detection, input analysis, long short-term memory, deep learning

1 Introduction

Deception detection is usually applied in law enforcement-related areas as a reference tool in investigations [1]. There are many techniques like polygraphs are utilized for detecting deceptive response of subjects. In addition to typical polygraphs, using thermal imaging (a.k.a. thermography) recorded by an infrared thermal camera to measure facial skin temperature as a cue to deception is also a way of deception detection [1]. Although this technology has not yet been used in law enforcement areas, thermal image analysis for polygraph testing has already gained a US patent [2] and is empirically supported by previous studies with results suggesting that it has the potential to detect deception with considerable accuracy [3].

Several experiments and discussion were centred on the relationship between emotional stimulations and temperature changes. Puri [4] suggested in his study that the temperature in the forehead area can be an indicator of stress level. In addition to forehead, other facial areas are also found to have different blood movement patterns. Derakhshan et al [5]'s study shows that the blood movement is faster in the periorbital area during threatening situations but facial area like the nose is lower because it usually requires less blood. Based on these findings, thermal imaging has been suggested as a deception detection tool because it can track the immediate partial change of facial temperature [6].

This paper focuses on the relationship between human's stress response and the blood flow in different facial areas, and try to classify subjects' deceptive response. The task for this paper is to predict the label (deceive or truth) of subjects based on the 2 thermal datasets. The 1st thermal dataset is composed of 31 subjects' granger causality from one facial region of interest to another. A two-layer fully connected neural network is trained to make classification by using this dataset. Considering the feature pruning in order to eliminate the effects of irrelevant features, a magnitude measure of contributions is adopted to select features to see whether this technique can boost the performance of the model. The final result of the neural network model will be also compared with that of Derakhshan et al [5]'s study. The 2nd dataset is composed of sequence data that indicates 31 subjects' the max and min value of blood flow in 5 face regions over up to 20 seconds in response to a question. LSTM is usually used to solve classification problems based on individual time steps of the sequence data [7]. So as for 2nd dataset, an LSTM will be trained to make classification of deceptive response.

2 Method

2.1 Datasets

Both of two datasets used in this paper consist of thermal imaging records of 31 samples collected by Derakhshan et al [5]. In Derakhshan et al [5]'s study, experimental data is collected from an experiment by using a mock crime protocol. 31 subjects are divided into two groups, one group of subjects did crime by "stealing" a necklace are labelled as

"deceptive". Another group of subjects didn't perform any "criminal" act. Then all subjects are interview and answer 8 questions categorized as "Relevant, Irrelevant and Neutral" questions. During the interview, the subjects' minor physiological changes were monitored using facial thermal imaging. The 1st thermal dataset contains thermal data only for question 6 "Did you steal the necklace" which is most relevant to the crime while the 2nd thermal dataset contains data for all 8 questions.

2.1.1 The 1st Thermal Dataset

This dataset contains 31 participants and 20 features which indicates the Granger causality from one facial region to another as shown in Table 1. For example, feature 1 represents the effect of the periorbital region on the forehead region.

Face Region	Periorbital	Forehead	Cheek	Perinasal	Chin
Periorbital	/	1	2	3	4
Forehead	5	/	6	7	8
Cheek	9	10	/	11	12
Perinasal	13	14	15	/	16
Chin	17	19	19	20	/

Table 1. Feature Index Table

The data in this dataset was scaled from 0 and 1, and the number of examples for each label is about 15, which is quite balanced. So the original dataset has been preprocessed in a good manner. Using the z-score method to normalize the dataset turns out that the performance of the model did not become better but even worse, so there is no particular normalization technique applied to preprocess the data except shuffling the order of the dataset's rows since all rows which have label "1" are on top, followed by the rows with label "2". It can on some extent stabilize the performance in the test set.

2.1.2 The 2st Thermal Dataset

The second dataset has 31 tabs for 31 subjects. Each tab contains max and min value of blood flow in 5 face regions when answering the 8 questions, so the original dataset has 80 features (8 questions* 5 face regions * 2 = 80 features). To classify the deceive response, this paper only considers question 6 for classification as it is the most relevant question like what Derakhshan et al [5] did in their study. In the preprocessing step, only 10 features related to question 6 are selected, which are min and max value of thermal data in 5 face regions (5 face region * 2 = 10 features). Among these 31 samples, 27 samples consisting of 199 rows, 3 samples consist of 178 rows, and 1 sample consists of 149 lines of row. To ensure the shape of the whole dataset remains consistent, 4 samples in which the number of rows is not equal to 199 are then deleted from the dataset so that the new dataset now consists of 27 samples. Normalization is performed to scale each feature to the range (0, 1) in order to make the training less sensitive to the scale of features [8]. It turns out the performance of the neural network model after normalization increased by scaling features in the given range.

2.2 Neural Network

The neural network applied in this paper is a two-layer fully connected, feedforward neural network. As shown in Figure 1, the network topology was 20-16-2, being twenty inputs, sixteen hidden neurons, and two output neurons. In this network, the link between each unit is a simple weighted link, and the basic sigmoid function, y=1+ e-x-1 , is chosen as the activation function in output layer for binary classification. Adaptive Moment Estimation (Adam) was adapted as the optimizer, the cross-entropy loss is adopted as identification loss function and the network was trained using error-backpropagation. The number of hidden neurons is chosen by trying different alternatives and compare the performances of these models. Model performances are evaluated by K-fold cross-validation. Since the dataset is quite small with only 31 samples, K=10 is chosen which turns out to be relatively the most suitable choice for this small dataset. When K=10, 4 samples among the total 31 samples would be chosen to combine as a test set, while the remaining 27 samples combined as a training set in every fold.

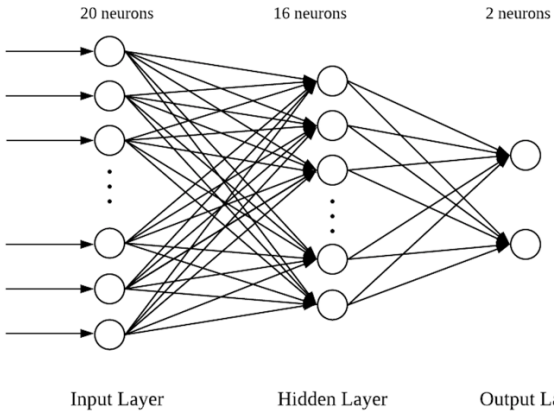

Fig 1. 20-16-2 Two-Layer Fully Connected Neural Network

Other hyperparameters are also chosen based on higher performance. The learning rate is set as 0.01 after trying alternatives 0.1, 0.01, 0.001 and 0.0001. The number of epochs is set as 70 after trying alternatives 20, 40, 60, 70, 80, 100, 150, 300. Because of this small dataset, the number of epochs is tended to set relatively small to avoid overfitting in some way. The following section further describes the technique applied to analysis input.

2.3 Input Analysis Technique

The technique implemented in this paper is derived from the distinctiveness analysis technique [9] which can be used to examine "the functionality of hidden neurons based on weight matrix" [10]. Based on that, to examine the functional difference between inputs, the formulae were shown as below [11].

$$angle(i,j) = tan^{-1}\left(\sqrt{\frac{\sum_p^{pats} sact(p,\mathrm{i})^2 * \sum_p^{pats} sact(p,\mathrm{j})^2}{\sum_p^{pats}(sact(p,\mathrm{i}) * sact(p,j))^2}} - 1\right)$$

where $sact(p,\mathrm{h})$ = norm (weight(h)) – 0.5

In this formula, the significance of features is determined by the effect of inputs on output. Neural network weights are supposed to be extracted to perform the calculation for input analysis. The least two significant inputs (features) then will be deleted which is inspired from Gedeon [11]'s removing method. The topology of the network, in turn, was adjusted as 18-16-2 after feature selection. The pruned dataset then will be used for model training to see whether this technique improves the performance of the model.

2.4 Deep Neural Network

As for the 2nd thermal dataset, the data which indicates the minor physiological changes are organized as a type of sequence data. In this case, the model trained to perform the classification is supposed to learn the pattern of minor physiological changes over time. Adopting the neural network model as 1st dataset is not appropriate because neural networks are memoryless [12]. Long Short-Term Memory (LSTM) is one of the most widely used recursive structures in sequence modelling. It controls the flow of information in recursive computations by using gate circuits. Because of this mechanism, LSTM network has a good performance in maintaining long-time domain memory [12].

Therefore, a LSTM network is trained to perform classification of deceptive response based on this dataset. The structure of the LSTM network is shown in Figure 2, which is a many to one model. As for this kind of model, the input is the sequence data, which in this case, is the thermal data over time, while the output is the prediction of fight or flight response.

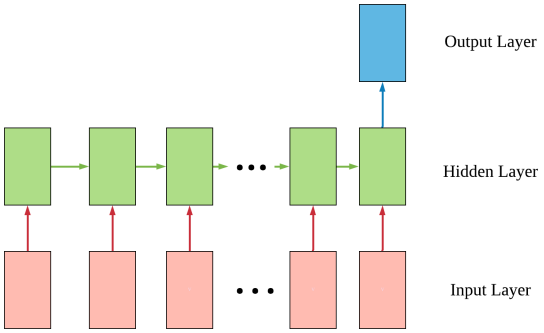

Figure 2. Structure of Many to One LSTM Model

In this LSTM network, Stochastic gradient descent (SGD) is adapted as the optimizer with the learning rate to be 0.1. As the purpose of the model is to classify the deceptive response, the cross-entropy loss is adopted as an identification loss function. The number of hidden neurons is set as 100, which is chosen by trying different alternatives and compare the performances of these models. The number of epochs is set as 1000 to avoid overfitting. Model performances are evaluated by K-fold cross-validation with K= 5. The choice of k is made by considering the bias can be reduced and the time takes to compute. It turns out that K = 5 makes a relatively better trade-off between the efficiency and the accuracy of error prediction. In this case, for each fold, 5 samples among the total 27 samples would be chosen to combine as a test set, while the remaining 22 samples combined as a training set.

3 Results and Discussion

3.1 Results and Discussion for the 1st Thermal Dataset

As for the neural network trained by using the 1st version of the thermal dataset, the results are shown in Figure 3. The average testing accuracy of the model before adopting the feature selection technique is 59.33% while the average testing accuracy of the model decreases to 42.79% after implementing the technique. The performance of the model decreases about 27.88% after adopting the technique, which indicates that the data analysis technique used didn't boost the performance of the neural network, instead, it even decreases the model performance. The purpose of using this magnitude measure for feature selection is to reduce the noise of irrelevant features [5]. Based on this, the reason for the low performance after adopting this technique might be that the 1st dataset used in this paper is quite balanced that contains few irrelevant and redundant input fields. Likewise, this result indicates that the magnitude measure technique used in this paper would be more suitable for the dataset which contains a lot of redundant information.

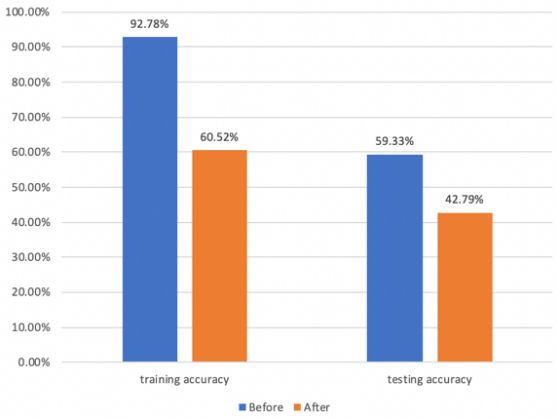

Figure 3. Average accuracy before and after using the technique

Besides, by comparison, there are some differences between the results get from this paper and the results reported in Derakhshan [5]'s research. "The preliminary results with all features show a successful classification rate of 67.1%. We further improved our classification rate to 87.1% by employing feature selection procedures" Derakhshan et al [5]. The preliminary accuracy rate of classification in this paper is 59.33% which is lower than what Derakhshan et al [5] got. The accuracy rate of classification with selected features is only 42.79% in this paper, which is much lower than that of Derakhshan [5]'s research (87.1%).

3.2 Results and Discussion for the 2nd Thermal Dataset

The LSTM model fed by the 2nd thermal dataset performed well with a testing accuracy of 77.07%. As shown in figure 4 and figure 5, the training loss of the neural network decreases smoothly as the number of epoch increase, which indicates that the network is learning from the training set and the network is not overfitting until the testing loss began to increase. In figure 5, the testing loss began to increase after about 1000 epochs, so eventually, the number of the epoch is set as 1000 to avoid overfitting. With this number of the epoch, the training accuracy of the model is 94.79%, while the testing accuracy reaches to 77.07%. Generally, the LSTM trained in this paper can classify the deceptive response with relatively high accuracy by compared with the performance of the neural network trained in this paper.

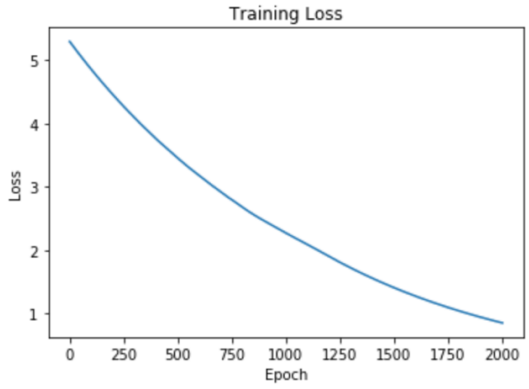

Figure 4. Training loss with 2000 epochs

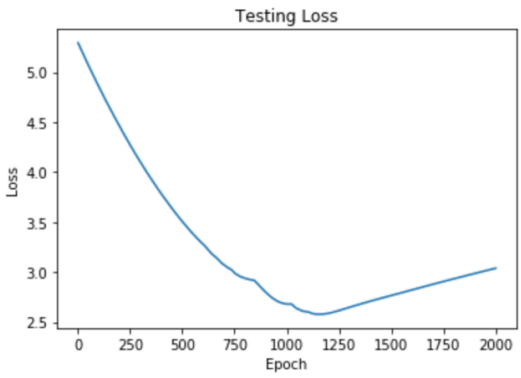

Figure 5. Testing loss with 2000 epochs

On the other hand, this result also suggests that the data of the 2nd database is organized in a better way for the model to make classification, in which, the network may learn the patterns of deceptive response and honest response more effectively.

5 Conclusion and Future Work

The main goal of this study is to classify human's deception response based on the blood flow change data obtained from 31 subjects. Two models (neural network and LSTM network) are trained to make classification respectively. Results show that the neural network trained in this paper can predict the label which is "deceive" or "truth" in an accuracy rate of 59%, while the testing accuracy of LSTM network reaches 77.07%. These two models are fed by two different versions of thermal datasets. Different performances of two models may indicate that 1) LSTM network can work more effectively in this case to make classification of deceptive response. 2) The 2nd thermal dataset in which the data is organized into sequence data could be regarded as a better way of data organization in deception detection problem.

By applying the input analysis technique in neural network based on weight matrix [11], the performance of model decreases about 27.88%, which indicates that the thermal dataset used in this paper is not suitable to apply this input analysis technique to select features. This input analysis technique is based on measuring the functionality of neurons which produced a better indicator of the significance of inputs in Gedeon et al [11]'s study. So the further work can be as applying other measure techniques mentioned in Gedeon et al [11]'s study (like sensitivity analysis or magnitude-based technique) to see whether can help increase the performance of the model.

Further work could try more techniques to increase the performance of the two models. Based on these two datasets, some regularization techniques like dropout and early stopping can be applied when trying to reduce overfitting. Randomly dropped the neurons out of the network in dropout technique may result in a less sensitive network which is less likely to overfit the training data. A relatively big number of epochs can easily lead to overfitting in these thermal datasets, especially in the 1st thermal dataset, so early stopping can also be considered as a way to reduce overfitting.

References

[1] Park, K. K., Suk, H. W., Hwang, H., & Lee, J.: A functional analysis of deception detection of a mock crime using infrared thermal imaging and the concealed information test. Frontiers in Human Neuroscience, doi:http://dx.doi.org.virtual.anu.edu.au/10.3389/fnhum.2013.00070 (2013)

[2] Pavlidis, I, U.S. Patent No. 6, 854, 879B2, Morristown, NJ: U.S. (2005).

[3] Pollina, D. A., Dollins, A. B., Senter, S. M., Brown, T. E., Pavlidis, I., Levine, J. A., et al.: Facial skin surface temperature changes during a "concealed information" test. Annals of Biomedical Engineering, 34(7), 1182-9. doi:http://dx.doi.org.virtual.anu.edu.au/10.1007/s10439-006-9143-3 (2006).

[4] Puri L et al. : Stresscam: Non-Contact Measurement of Users' Emotional States Through Thermal Imaging (Portland, OR: ACM) pp 1725–172. (2005)

[5] Derakhshan, A., Mikaeili, M., Nasrabadi, A. M., & Gedeon, T.: Network physiology of "fight or flight" response in facial superficial blood vessels. Physiological measurement, vol. 40, no. 1, p. 014002 (2018).

[6] Warmelink, L., Vrij, A., Mann, S., Leal, S., Forrester, D., & Fisher, R. P.: Thermal imaging as a lie detection tool at airports. Law and Human Behavior, 35(1), 40-8. doi:http://dx.doi.org.virtual.anu.edu.au/10.1007/s10979-010-9251-3 (2011).

[7] Zhang, Y., Xu, S., Ma, S., Yang, Y., & Ren, X.: Does higher order LSTM have better accuracy for segmenting and labeling sequence data?. Ithaca: Cornell University Library, arXiv.org. Retrieved from https://search-proquest-com.virtual.anu.edu.au/docview/2073900563?accountid=8330 (2018).

[8] Defilippi, R: Standardize or Normalize? — Examples in Python, Medium, Available: https://medium.com/@rrfd/standardize-or-normalize-examples-in-python-e3f174b65dfc (2018).

[9] Gedeon, T.D. and Harris, D.: Network Reduction Techniques. Proceedings International Conference on Neural Networks Methodologies and Applications, AMSE, San Diego, vol. 1: 119-126 (1991).

[10] Gedeon, T.D.: Indicators of Hidden Neuron Functionality: Static versus Dynamic Assessment. Australasian Journal of Intelligent Information Systems. vol. 3 (2): 1-9 (1996).

[11] Gedeon, T. D.: Data Mining of Inputs: Analysing Magnitude and Functional Measures. International Journal of Neural Systems 08, 209–218 (1997).

[12] Laddad, A.: Basic understanding of LSTM. from https://blog.goodaudience.com/basic-understanding-of-lstm-539f3b013f1e (2019).