LSTM Neural Networks: Input Features Analysis on Anger Veracity Detection

Xue Ling Teh

Research School of Computer Science,
Australian National University, Canberra Australia
u6462117@anu.edu.au

Abstract. The popularity of using neural networks in performing data analysis and supervised learning predictions has led to the emergence of various analysis techniques, which continue to be refined based on previous findings. This research aims to investigate binary classification using a basic neural network and Long Short-Term Memory (LSTM) model on time series data, in particular the pupillary response of human participants based on the anger video dataset. The application of brute-force analysis suggests that the choice of summarized statistical input features (or lack thereof) appear to have little impact on the model prediction accuracy, yet the initial weights assigned from input to hidden neurons affects the output predictions. Given the statistical summary of eye-gaze input features, a basic neural network does not reliably classify the nature of anger portrayed on videos.

Keywords: Neural network, LSTM, analysis technique, magnitude measure, brute force, input feature contribution, time series data, anger recognition, eye gaze, pupillary response, physiological signals

1 Introduction

Neural networks are researched extensively in the computing field and used in a variety of situations in solving business problems, such as generating forecasts based on existing data. It is often challenging to decide on the optimal hyperparameters to train a neural network model. When training a neural network model, the choice of hyperparameters includes the initialisation of weights, input features, and the number of hidden neurons in the hidden layer. Existing weights are updated with the newly calculated weights based on the error values computed in the backpropagation step. In weight visualization curves (WV-curves), the average contribution of a neuron from one layer to another is calculated based on the weight values [6]. Early literature has discovered that by identifying hidden neurons with similar functionality, these somewhat redundant neurons can be pruned from the network in order to improve generalization [2].

This research paper explores the use of analysis techniques on time series data, particularly on human physiological signals, using a binary classification neural network model. Analysis techniques discussed in this report include brute force analysis, applying magnitude measures such as Q-contribution, a coined term used in this report which refers to the contribution of an input neuron to an output neuron, based on the Q formula in Gedeon's 1997 paper on input analysis on magnitude and functional measures [3]. The final results produced by the model are affected by the initial weights which are assigned from input to hidden neurons, which supports the notion that weight initialisation of inputs has a substantial impact on the network performance.

A study in 2017 by Hossain and Gedeon [4] used a "leave-one-video out process" in measuring model classification accuracy on an observers' ability to distinguish between elicited real and posed (acted) smiles in videos. Another study in 2017 by Chen, Gedeon, Hossain, and Caldwell [1] investigated a total of 22 participants' physiological signal, mainly the pupillary response, compared to their verbal response in detecting the veracity of anger expressed in videos. The obtained mean accuracy classification results in the anger recognition study were 95% for pupillary response compared to 60% from the verbal response of participants. The anger dataset was chosen for the research described in this report. The objective in applying the aforementioned analysis techniques is to find out how each input feature contributes to the final output obtained, which is useful to determine how effective the network is when provided by summarized statistical measures of a particular type of time series data. This research has been extended to include the analysis of a more complete version of the anger dataset, with the raw data of participants' left and right eye pupillary response. For the extended research, a supervised deep learning classifier, the Long Short-Term Memory (LSTM) [7] model was trained and tested on the updated dataset. The analysis approach using both types of models on the anger dataset is described in the next section.

2 Analysis Techniques and Implementation

The main goal of this research is to determine the importance of individual statistical measures in time series data, particularly in the chosen anger dataset, when applied to a binary classification neural network and LSTM model, based

on existing analysis techniques in literature. In order to achieve the objective, the following process has been undertaken: initial preprocessing of both versions of the anger dataset, defining a customised neural network structure and LSTM model, dividing the dataset into training and testing data, applying the analysis techniques when training and then testing the models with the respective train and test data, and finally evaluating the performance of the neural network model.

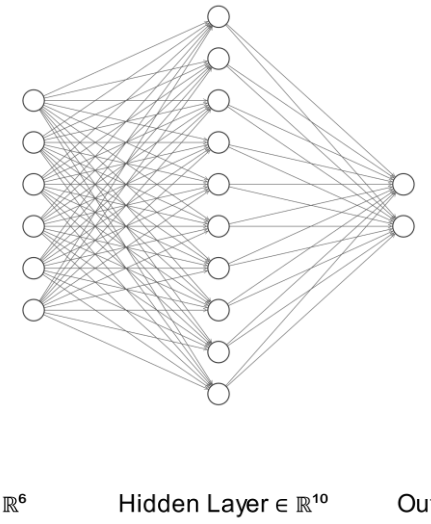

Fig. 1. Fully-connected neural network structure (*6 input neurons, 10 hidden neurons, and 2 output neurons*).

When pre-processing the initial version of the data, the desired input features of the dataset were selected and column features which were not meaningful (i.e. order of video appearance, video name) were removed. In the readme file of the first version of the anger dataset, it was stated that column 2 (video name) should be omitted when predicting for the class label, which is either 'Genuine' or 'Posed'. Version 1 consisted of the statistical measures of 22 participants' eye gaze (pupillary response) data, which were already normalised to the 0-1 range. The output class labels were converted into integer value representations: Genuine is 1 and Posed is 0. I have selected 6 input features: 'Mean', 'Std', 'Diff1', 'Diff2', 'PCAd1', 'PCAd2' and 2 output class labels (targets): 'Genuine' or 'Posed'. After a random shuffling to the order of the data, 80% were allocated as the training set, with the remaining 20% as the test set. The training set was fed into a custom-defined neural network as shown in Figure 1, with uniform distribution of initial weights, 10 hidden neurons, using the sigmoid activation function. The two-layer neural network is trained using error backpropagation and Stochastic Gradient Descent (SGD) as an optimiser, which holds the current state and updates the parameters based on the computed gradients. Cross-entropy loss is used to evaluate the network's performance on the training set.

An updated second version of the anger dataset was used for the deep learning approach. The second version, which consisted of raw data samples, had a different representation than the statistical summary in the initial dataset. The raw pupillary diameter (PD) for each participant is recorded separately in two respective Excel files for the left (PDLeft.xlsx) and right (PDRight.xlsx) eye. Each Excel file contains multiple tabs, each tab representing a different video. Each column in the Excel sheet tab represents the sequence of a participant's pupillary response over time, throughout the duration of watching that particular video. The timestamp of each row is approximately $1/60^{th}$ second after the previous row. The timestamp and sampled data for each participant watching each specific video in both files correspond with each other. An Excel file consisting of the mean PD values for left, right and both eyes combined, of all participants across all videos was also provided.

When pre-processing the raw values of left and right PD (in their respective files), the first step was to perform min-max normalization according to each participant's PD sequence. A sequence in this context is defined as the collective PD values (either right or left) of a participant throughout the duration of a specific video. The min-max normalisation formula, which transforms the current data values to a range between 0-1 is provided below in (1).

$$d_i' = \frac{d_i - min}{\max - min} \tag{1}$$

The 0 values in each PD sequence is then imputed with the mean of all PD values in that particular sequence for both left and right PDs. Empty columns which contained Nan values were removed, as not all participants watched every video. Video IDs which began with 'T' were videos portraying true anger, and labelled as 'Genuine' and converted to integer as class 1, and those starting with 'F' were videos of fake anger, which were then labelled 'Posed' and converted to class 0. A unique session ID was assigned for every unique combination of participant ID and video ID. The left and right PD sequences correspond to the assigned session ID for each participant watching a particular video.

The order of session ID was randomly shuffled. Similar to the previous vanilla neural network, 80% of the sequence data were allocated as the training set, with the remaining 20% as the test set. The training set was fed into a custom-defined LSTM, with random distribution of initial weights, 100 hidden neurons, using the sigmoid activation function. The LSTM is trained using error backpropagation and Stochastic Gradient Descent (SGD) as an optimiser, which holds the current state and updates the parameters based on the computed gradients. Cross-entropy loss is used to evaluate the network's performance on the training set.

2.1 Brute force analysis

Brute force analysis was performed on the pre-processed anger dataset by eliminating inputs and comparing the test set model predictions with the actual labels from the original dataset. According to [3], eliminating only 1 input led to inconsistent result. Hence, in my analysis I have decided to implement pair-wise elimination of inputs. Since there are 6 input features, the total number of input pair combinations is 15. Similar to [3], for each of the 15 possibilities, 4 networks with the same topology (6-10-2) were trained. During training, the weights for the 2 eliminated inputs are excluded.

2.2 Q-contribution

[6] first defined the measure P_{ij} to measure for an input's contribution to a neuron in a hidden layer. [3] proposed an extension of their technique, which was a measure P_{jk} to measure the contribution from hidden neuron to output neuron, leading to the Q-contribution formula, which is the sum of the cross-product of P_{ij} and P_{jk}. Q-contribution, the contribution of an input neuron to an output neuron in a neural network, is a coined term used in this report, based on the definition of the Q formula introduced in [3]. The formulae are listed below.

$$P_{ij} = \frac{|\, w_{ij} \,|}{\sum_{p=1}^{n\,i} |\, w_{ij} \,|} \tag{2}$$

$$P_{jk} = \frac{|\, w_{jk} \,|}{\sum_{r=1}^{n\,h} |\, w_{rk} \,|} \tag{3}$$

$$Q_{ik} = \sum_{r=1}^{n\,h} (P_{ir} \times P_{rk}) \tag{4}$$

3 Results and Discussion

Each run of the source code file produced some variation in results due to the randomness of the train-test data split and the neural network weight initialisations. For the purpose of discussion, below are the results of an example run.

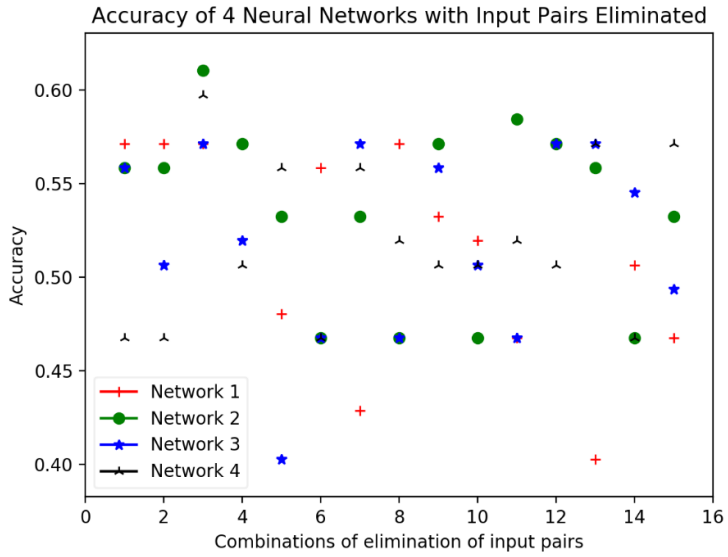

Fig. 2. Accuracy results of networks 1-4 (*data points with 4 different respective markers*) with 15 different combinations of input pair elimination.

Figure 2 shows a considerable difference in the range of accuracy results for four different networks trained with each input pair elimination combination, from the lowest (41.6%) to highest (62.6%). This shows that there is some effect (21% difference) on the accuracy of the model classification predictions for the test set when training with any combination of input pair elimination.

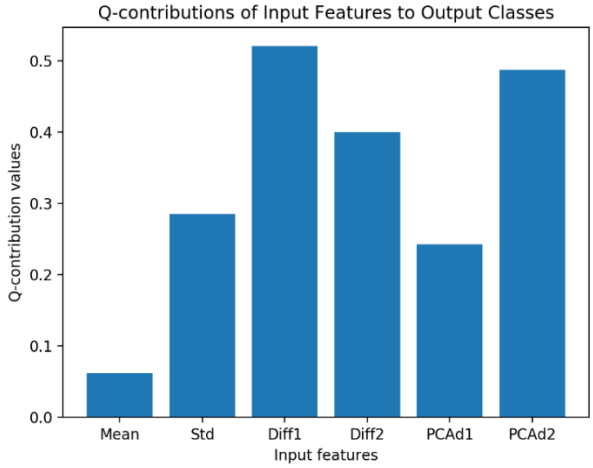

Fig. 3. Q-contribution values of input features 1-6 (*mean, std, diff1, diff2, PCAd1, PCAd2*) to output classes.

According to Figure 3, the bar graph of Q-contribution values for the 6 input features indicates that all features except 'mean' have a Q-contribution of greater than 0.2. The highest Q-contribution value is diff1, with more than 0.5, which suggests that it has the most impact or contribution in the network. The top 3 significant inputs to the network, in order, are diff1, PCAd2, and diff2. Table 1 below shows the ranking of input features (according to their column position number), from the most to least significant.

Table 1. Q-contribution of input features from most significant to least significant

Model	Input features (Most significant to Least significant)
Q-contribution	3,6,4,2,5,1

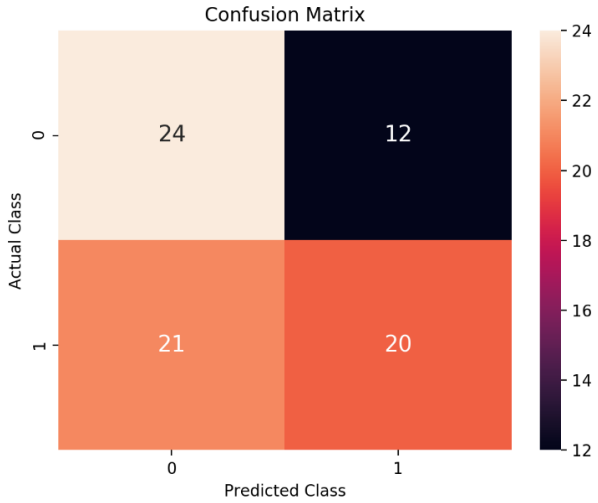

Fig. 4. Confusion matrix for evaluating the performance of a trained neural network, showing the number of predicted class labels against the actual class labels (*predicted class vs actual class*).

Confusion matrix is a good summary visualization of the neural network's performance and provides more insight to the accuracy of the predictions. In Figure 4, the class labels 0 and 1 for both predicted and actual are 'Posed' and 'Genuine', respectively. The x-axis (column values) and y-axis (row values) represent the predicted classes and actual classes, respectively. The top left and bottom right quadrants are the True Positives (TP) and True Negatives (TN), in other words, the total number predictions in which the model classified correctly [5]. There are 21 False Positives (FP), in which the model predicted as 'Posed' when the actual label is 'Genuine', and 12 False Negatives (FN), which is the opposite. The results show the number of unseen (test) data that the network predicts belong in each class and compares the predictions to the respective actual class labels. In the context of confusion matrix, the accuracy is calculated based on the sum of TP and TN divided by the sum of all of the quadrants (TP, FN, FP, TN).

4 Conclusion and Future Work

In the original study on anger recognition [1], with the focus on pupillary response, the mean classification accuracy was 95%, whereas the verbal response returned 60%. In the case of brute-force analysis, the accuracy of 4 neural network model predictions with 15 combinations of input pairs eliminated shows that there is a 21% difference in range, which can either be moderately higher or moderately lower by approximately 10%. For the Q-contribution analysis, the input which has the most impact in this particular run example is diff1. The mean pupil diameter of the 22 participants seems to have the least impact in the learning of the neural network. These results suggest that the application of brute-force analysis and Q-contribution techniques did not provide additional benefits to increase the performance of the binary classification model.

The selected statistical summary features of the anger dataset did not seem to be representative of the dataset or have distinct enough patterns for the model to learn from in order to make an accurate prediction that is consistently greater than chance (>50%). This implies that the summarized statistical input features of multiple participants' pupillary response do not lead to reliable neural network predictions of the nature of anger portrayed on videos. Therefore, based on the applied analysis techniques above, the choice of input features, including the removal of any number of the chosen 6 input features during training, appears to have little impact on the model predictions. Extended research on the chosen anger dataset to predict videos instead of the 'Genuine' or 'Posed' label would lead to further insights on using statistical features of the pupillary response. Other possible future work involves investigating the effect of applying the techniques described in this paper or in related literature on other types of physiological signal data, such as Electrodermal activity (EDA), heart rate variability (HRV) and skin temperature (ST).

References

1. 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.
2. Gedeon, T. D. (1995, November). Indicators of hidden neuron functionality: the weight matrix versus neuron behaviour. In *Artificial Neural Networks and Expert Systems, 1995. Proceedings, Second New Zealand International Two-Stream Conference on* (pp. 26-29). IEEE.
3. Gedeon, T. D. (1997). Data mining of inputs: analysing magnitude and functional measures. *International Journal of Neural Systems*, *8*(02), 209-218.
4. Hossain, M.Z., and Gedeon, T.: 'Classifying posed and real smiles from observers' peripheral physiology', in Editor (Eds.): 'Book Classifying posed and real smiles from observers' peripheral physiology' (2017, edn.), pp. 460-463
5. Raschka, S.: 'An overview of general performance metrics of binary classifier systems', arXiv preprint arXiv:1410.5330, 2014
6. Wong, P.M., Gedeon, T.D., and Taggart, I.J.: 'An improved technique in porosity prediction: a neural network approach', IEEE Transactions on Geoscience and Remote Sensing, 1995, 33, (4), pp. 971-980.
7. Zebin, T., Sperrin, M., Peek, N., and Casson, A.J.: 'Human activity recognition from inertial sensor time-series using batch normalized deep LSTM recurrent networks', in Editor (Eds.): 'Book Human activity recognition from inertial sensor time-series using batch normalized deep LSTM recurrent networks' (IEEE, 2018, edn.), pp. 1-4