Predict Depression Level by Long Short-Term Memory Network and Explain Result with Attention

Jiang ping Gao[1]

Research School of Computer Science
The Australian National University
Canberra ACT 2601
u6987832@anu.edu.au

Abstract. This paper implemented a four-layer, long short-term memory network with attention mechanism to classify the depression grade of the physiological signal data set collected while watching depression videos. This paper suggests that when different physiological signal features are selected, the classification effect does vary as described in the data set. In selecting different physiological signals, this paper proves that the performance of the four-layer LSTM network with attention mechanism is poor, easy to underfit, and unable to complete the classification task. This paper uses attention mechanism to explain the rationality of the rules.

Keywords: LSTM · Attention · Neural network · Depression · Galvanic Skin Response · Skin Temperature · Pupil Dilation.

1 Introduction

1.1 Problem Description

Depression is a global public health problem, with significant and persistent low mood as the main clinical characteristics. Symptoms of depression usually include insomnia, lethargy, decreased appetite, anhedonia, anxiety, decreased motivation, decreased memory, feelings of worthlessness, and hopelessness.World Health Organization (WHO) states that depression is the leading cause of disability as measured by Years Lived with Disability (YLDs) and the fourth leading contributor to the global burden of disease [1].

As depression is the leading cause of disability worldwide, large-scale surveys have been conducted to establish the occurrence and risk factors of depression [5]. However, the multiple factors and inherent complexity of depression (such as physicians' subjective judgments and patients' inexpressiveness) leads to the challenges of objectively and accurately detecting mental disorders. In recent years, the development of machine learning (ML) frameworks for automatic diagnosis of depression has escalated to a next level of deep learning frameworks [2], demonstrating its potential to assess factors contributing to the prevalence and clinical presentation of depression.

People with depression express differently from normal individuals in many ways, such as eye contact and body movements, which are able to be captured and observed by others. In react to watch different levels of depression patients, observers are possible to produce different physical signals, which are expected to be utilised by deep learning algorithms to objectively detect the depression levels. This paper aims to examine the availability of using neural networks to predict depression by collecting the changes of physical signals of observers when they are watching videos of different levels of depressed people. In this project, a four-layered long short-term memory network is provided, with time series features at the input layer, 64 dimension set for the hidden layer and 4 outputs for the depression level classification. Min-Max Normalization, k-fold cross validation and stratified sampling methods are employed for data preprocessing to normalize data and prevent overfitting problems, Cross-entropy is performed as loss function while the Adam optimizer and L2 regularization are mentioned for the optimization step. Attention is used to explain the rationality of the rules generated by the neural network.

1.2 Dataset Description

The dataset considered in this project is selected from the data built in [3]. People in [3] picked out 16 over 300 webcam videos from the 2014 Audio-Visual Emotion Challenge (AVEC2014) dataset [4], which are at similar lengths and covers all the four depression categorises mentioned in the AVEC2014 dataset (no or minimal, mild, moderate and severe). To generate the original research data, fourteen participants aging from 18 to 27 were involved in the experiment, who were requested to watch those depression videos and whose actions were collected and measured. Participants' reactions responding to the videos are recorded by sensors and the collected physical signals are classified into three categories: Galvanic Skin Response (GSR), Skin Temperature (ST) and Pupillary Dilation (PD). The resulted dataset contains 192 complete responses with time series features, which are pre-labelled by the four depression levels and covered by the three physical signal categories.

In the following sections, Section 2 presents the methodologies that used in this project, where the procedure of data preprocessing is introduced, along with the description of the Long short-term memory network framework and techniques utilised in dropout mechanism, regularization, loss function, optimization, evaluation and attention. Result analyses, comparison and discussion are demonstrated in Section 3, involving the analyses of both result prediction and explanation. Section 4 concludes this paper and showcases a discussion of the future work.

2 Method

2.1 Data Preprocess

As mentioned in [3], physical signals of participants vary from individuals, which indicates that a similar signal may represent the reaction to different levels of depression videos among different observers. This trick problem heralds the obligation to normalise data to a same scale. In this paper, Min-Max Normalization is employed to scale down data to 0 to 1 for each observer, promising the signals of all participants are at the same level and eliminating the individual differences. According to different length of time series data, this paper cuts off all time series data to same length which is the shortest of each kind of data.

Unlike most real databases, statistics for this dataset show that data under each category is evenly distributed (48 of each)(see Fig.1), thus no requirement is needed to adjust the distribution of the data. However, the limited amount of data (192) compared with the large number of features (at least 144) suggest that approaches are required to avoid over-fitting issues. This paper accepts stratified sampling, with the premise of ensuring that the category distribution ratio of subsets does not change, all data is divided into 8 groups, with the training set and the test set separated in a 3:1 ratio.

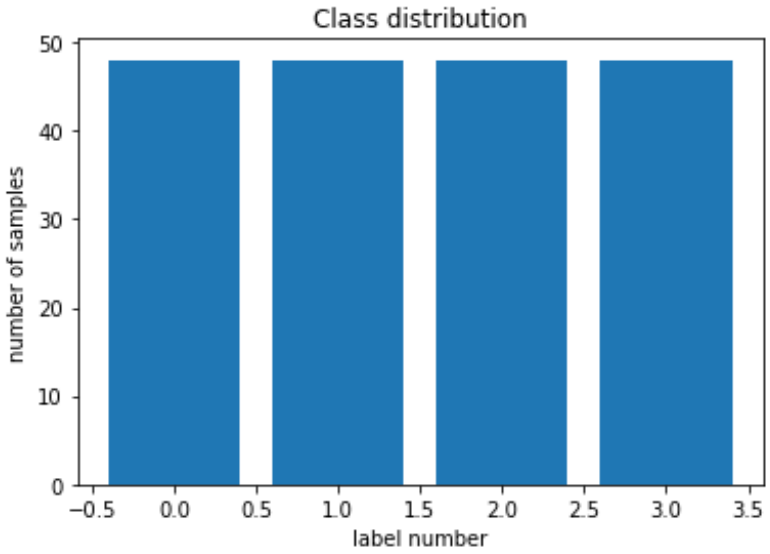


Fig. 1. Label Distribution in Original Dataset

2.2 Long Short-Term Memory Network

Long short-term memory network[8][9] is a variant of recurrent neural network, which can effectively solve the gradient explosion or vanishing problem of a simple recurrent neural network. LSTM mainly controls the internal information transfer by adding a gating mechanism and a new internal state. The internal state c_t of LSTM is specially used for linear circular information transmission and outputs information (non-linearly) to the external state of the hidden layer h_t at the same time.

$$\begin{aligned} c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t, \\ h_t &= o_t \odot \tanh\left(c_t\right), \\ \tilde{c}_t &= \tanh\left(W_c x_t + U_c h_{t-1} + b_c\right) \end{aligned} \tag{1}$$

In 1, f_t, o_t and i_t are three gates to control the path of information transmission. $\odot$ is the product of vector elements; c_{t-1} is the memory unit of the last moment; $\tilde{c}_t$, are candidate states obtained by nonlinear functions. At each moment t, the internal status of the LSTM network c_t is marked with a log of historical information up to the current moment

The gating mechanism is composed of an input gate, forgetting gate and output gate. Forgetting gate f_t controls the internal status of the last moment to switch 1 how much information needs to be forgotten. The input gate i_t controls how much information needs to be saved for the current moment of the candidate state, light. The output gate o_t controls how much information the internal status sign at the current moment needs to output to the external status gate. The gate in the LSTM has a value between 0 and 1, indicating that a certain percentage of information is allowed to pass through. The three gates functions are expressed as 2.

$$\begin{aligned} i_t &= \sigma\left(W_i x_t + U_i h_{t-1} + b_i\right) \\ f_t &= \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right) \\ o_t &= \sigma\left(W_o x_t + U_o h_{t-1} + b_o\right) \end{aligned} \tag{2}$$

Where $\sigma(\cdot)$ is a Logistic function with an output interval of (0, 1), x_t is the input of the current moment, and h_{t-1} is the external state of the previous moment.

When $f_t = 0$, $i_t = 1$, the internal status will empty history information and will be the candidate state vector. When $f_t = 1$, $i_t = 0$, the internal status will copy the contents of the previous moment and does not write new information.

This paper uses four-layers of long short-term memory networks to predict the depression level. The first layer is the input layer, and the dimension number of the input layer is the length of time series of each row of data. The second layer is the LSTM layer with layer number of four, which have 64 as the hidden dimensions. The third layer is an attention layer which helps explain the result and can improve the classified accuracy. The last but not least layer is a linear layer. It helps make output of attention layer to dimension which is the number of class. How do researchers determine that the number of hidden layer neurons has been a more complex problem? This article adopts the way of experiment many times which shows in section 3.2 to define the hyper parameters. On the one hand, this paper improves the complexity of the model to make it have better classification effect. In this paper, the dropout layer is added to deactivate part of neurons to improve the robustness of the neural network and avoid overfitting. Because it is a classification problem, cross-entropy is used as the loss function in this paper, and the Adam optimizer updates the parameters of the model.

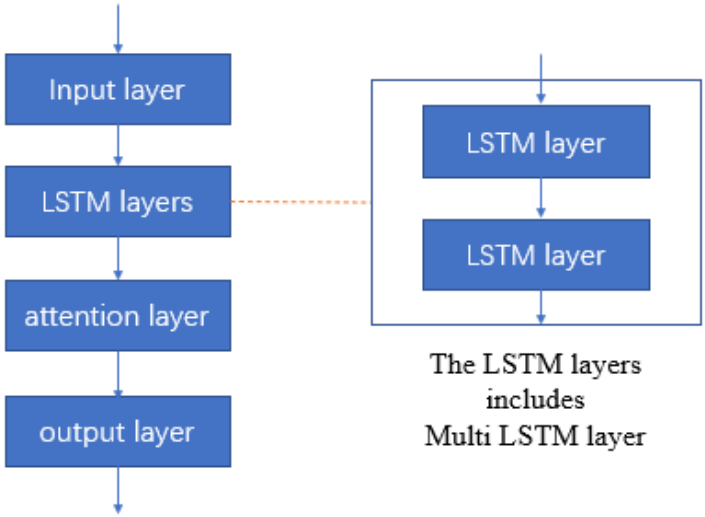


Fig. 2. Pipeline of model

2.3 Attention

In the case of limited computing power, the attention mechanism can also be called the attention model. As a resource allocation scheme, Attention Mechanism uses limited computing resources to deal with more critical information, which is the primary method to solve information overload. In order to save computational resources, it is not necessary to input all the information into the neural network but to select some task-related information from the input data. The calculation of the attention mechanism can be divided into two steps: one is to calculate the distribution of attention on all input information; the other is to calculate the weighted average of input information according to the distribution of attention.

To select information related to a particular class from input vectors of batch size, this paper need to introduce a class-related representation called a Query Vector and calculate the correlation between each input Vector and the Query Vector using a scoring function. Given a task-related query vector q, the learnable parameter of the query vector q which is the hidden layer state in this paper. This paper uses the attention variable z to represent the index position of the selected information. Therefore, z= n indicates that the zth input vector is selected. This paper uses a "soft" information selection mechanism. The probability α of choosing the nth input vector given q and x is calculated, where α is called attention distribution, and $S(x, q)$ is the scoring function for Attention. In this paper, the dot product method is used to calculate:

$$\begin{aligned} &\alpha_n = p(z = n \mid X, q) \\ &= softmax\left(s\left(x_n, q\right)\right) \\ &= \frac{\exp(s(x_n,q))}{\sum_{j=1}^{N}\exp(s(x_j,q))}, \\ &s(x, q) = x^\top q \end{aligned} \tag{3}$$

The attention distribution α can be interpreted as the degree to which the nth input vector receives Attention for a given category-related query q. This paper use a "soft" information selection mechanism to summarize the input information.

$$att(X, q) = \sum_{n=1}^{N} \alpha_n x_n, \ = E_{z\sim p(z|X,q)}\left[x_z\right] \tag{4}$$

2.4 Dropout

The dropout function is to cause the activation value of a neuron to stop working in accordance with a certain probability during the process of forwarding propagation.[6] Because the value of the node of the next layer is obtained through the interaction between the nodes of the upper layer, reducing this interaction can significantly enhance the generalization ability of the model so that it will not rely too much on some local features, to avoid the over-fitting problem of the model. In this paper, the value of dropout is 0.2, which means each neuron is not involved in forwarding transmission with 20% probability.

2.5 Regularization

To avoid model overfitting, this paper sets the parameter of weight attenuation of the Adam optimizer to 0.01. This optimizer is the equivalent of an L2 regular. In the fitting process, L2 regularization usually tends to make the weight as small as possible, and finally, construct a model with all the parameters relatively small. Because it is generally believed that the model with a small parameter value is relatively simple, can adapt to different data sets, and avoids overfitting to a certain extent.

2.6 Loss Function

This paper uses cross-entropy as the loss function. Cross entropy is used to measure the difference between actual and predicted probability distributions.The function is shown below. M means the number of labels, y_{ic} is 1, if the category is the same as sample i, otherwise, it is 0 and p_{ic} is the predicted probability for the observed sample i belonging to the category c. When cross-entropy is used as a loss function, its partial derivative depends on the error degree of the model. The larger the error degree is, the worse the model effect will be. However, the larger the value is, the larger the partial derivative value will be so that the model learning speed will be faster. The smaller the value of cross-entropy, the better the prediction effect of the model.

$$L = \frac{1}{N}\sum_i L_i = \frac{1}{N}\sum_i -\sum_{c=1}^{M} y_{ic}\log\left(p_{ic}\right) \tag{5}$$

2.7 Optimization

The Adam optimizer calculates the update step size by comprehensively considering the first-order Moment Estimation (i.e. the mean of the gradient) and second-moment Estimation (i.e. the uncentralized variance of the gradient) of the gradient. Because of its straightforward implementation, efficient calculation, less memory demand, parameter update, which is not affected by the scaling of the gradient and automatic adjustment of the learning rate, the Adam optimizer is adopted in this paper. In this paper, the learning rate of the Adam optimizer is set to 0.01, and the L2 regularization mentioned above is also set in the optimizer.

2.8 Evaluation

This paper practices accuracy and confusion matrix in evaluating the classification results. Accuracy is the number of correct classifications divided by the total number of samples. Confusion matrix is a method to evaluate the classification result of each class.

3 Results and Discussion

3.1 experiment environment

The experiment environment shows in Table 3.1.

—PC	Parts	Settings
Software	System	Windows 10
	Python	3.7.8
	Pytorch	1.4.0
Hardware	CPU	Intel Core i7-9700K 8 Cores 4.9 GHz
	GPU	Nvidia GeForce RTX 2060(6GB)
	Memory	Corsair DDR4 DRAM 32GB 3200MHz

Table 1. Experiment Environment Settings.

3.2 Hyper parameters chosen

In this paper, we chosen different candidate hyper parameters to test the LSTM with attention model. The result shows in Table 3 in the appendix. From the table result, this paper chose (hidden size,batch size,learning rate, dropout, number of epoch) to (64, 50, 0.01, 0.2, 500) as hyper parameters.

3.3 Predict Result

The model's performance in this paper is not good in either training or test set. The train and test accuracy are both close to 30%-40%. This is an underfitting phenomenon. It is proved that the model of the three-layer neural network designed in this paper does not perform well on the data set as input. According to the influence of different selected characteristics on the accuracy of the neural network, this paper separately analyzes and runs the data sets of the three physiological indicators, and the performance results are as Table 2. Here, we change epoch number to 1000 and learning rate to 0.001. The predict result have the same distribution as mentioned in [3].

Dataset	Train accuracy	Test accuracy
GSR	33.99%	33.33%
Pupil	24.84%	25.64%
Pupir	25.49%	23.08%
Skin temperature	33.33%	29.41%

Table 2. Result of Single Physiological Signal Feature

From confusion matrix, it is obvious that test cases have a high possibility to judge as a wrong class. The only correct class is label 1. And label 2 and label 3 have 1 or 2 correct prediction cases. This model also have no predict of label 0, this shows the weakness of model in judge label 0 data. All above mean the model have a weak ability to classify this dataset.

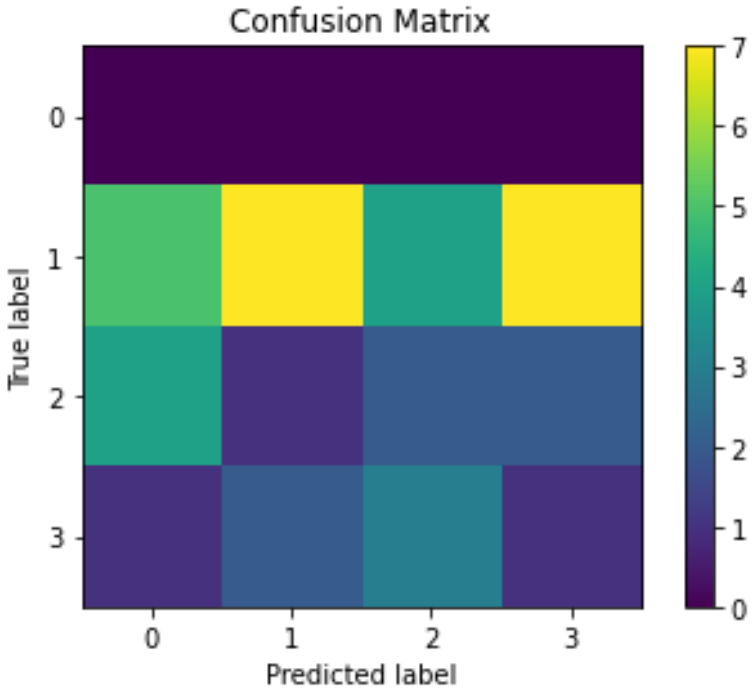


Fig. 3. Confusion Matrix

From Figure 4.b, it is obvious that different people have different physical signals to same video. For same person, it is hard to figure out the reaction to different level of videos in Figure 4.a. And compare Figure 4.c with Figure 4.d, there seems no evidence for good classify after normalize the dataset. Therefore, this dataset needs more preprocssing, the performance of LSTM network can be forgiven.

3.4 Explain Result with Attention method

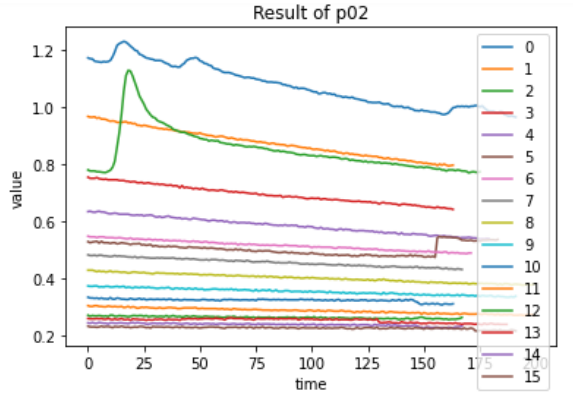
The interpretability of neural networks has always been criticized. Different from causal index and characteristic chosen technology [7], this paper tries to explain the neural networks with the help of attention. In this paper, the attentions of the LSTM network shown how important of each time stamp data influence to the final result. As Figure 5 shows below, different time stamps have different influences(it show as the degree of black) to the final prediction, The ideal result for each class should have close distribution of attentions. For example, sk data is smaller than 0.4 at fifth frame will increase the possibility of class 0.

4 Conclusion and Future Work

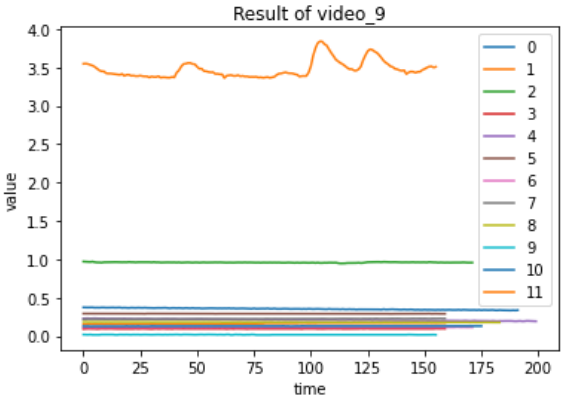
Although many technologies (such as regularization, dropout, Attention, etc.) were used in the experiment of predicting the level of depression videos based on the physiological signals of viewers to avoid over-fitting of the model, the effect of the final experiment was still unsatisfactory. The performance in the test set was weak, and the phenomenon of underfitting still occurred. All these problems make the model inexplicable to the solution of the problem.

Because the samples have too many features and the correlation of each feature is little, there may be highly characteristic noise. In this regard, future research can start from enhancing the data set or improving the rationality of feature selection. The data set may need to introduce more data information to enhance the model's performance or manual feature filtering to find beneficial feature information.

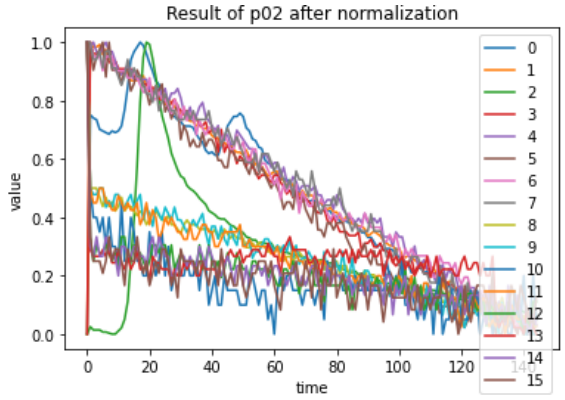
The model-fitting problem cause cannot effectively extract rules information from attention. This paper made the following conjecture: 1)If reduce some noise data or add the people label as feature, the model may have better performance and the explain rules will directly influence by kind of people and their behaviors. 2) use some statistic values such as mean value or variance to help improve the performance of model and generate expalin rules from attention directly.



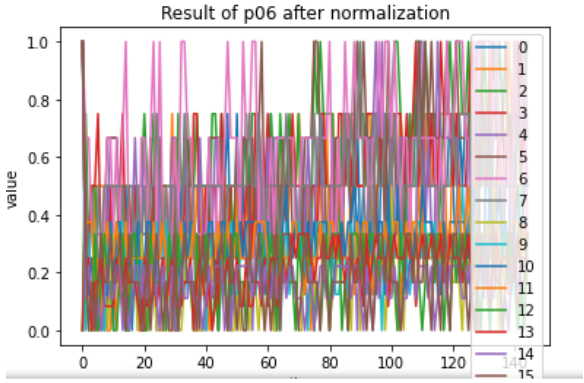
(a) gsr of same person to different videos



(b) gsr of same video to different people



(c) person02's gsr to different videos(normalization)



(d) person06's gsr to different videos(normalization)

Fig. 4. Data set analysis

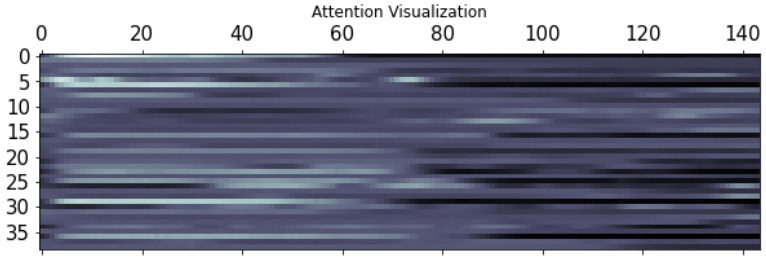


Fig. 5. Attention Matrix

References

1. Reddy, M. S. Depression: the disorder and the burden. SAGE Publications Sage India: New Delhi, India 1-2. (2010)
2. Mumtaz, Wajid, and Abdul Qayyum. A Deep Learning Framework for Automatic Diagnosis of Unipolar Depression." International Journal of medical Informatics 132 (2019): 103983.
3. X. Zhu, T. Gedeon, S. Caldwell and R. Jones, Detecting Emotional Reactions to Videos of Depression, In: IEEE 23rd International Conference on Intelligent Engineering Systems (INES), pp. 000147-000152 (2019) https://doi.org/10.1109/INES46365.2019.9109519
4. M. Valstar et al., Avec: 3d Dimensional Affect and Depression Recognition Challenge, In: Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, pp. 3-10, (2014).
5. Oh J, Yun K, Maoz U, et al. Identifying Depression in the National Health and Nutrition Examination Survey Data Using a Deep Learning Algorithm. Journal of Affective Disorders, (2019), 257: 623-631.
6. Hinton G E, Srivastava N, Krizhevsky A, et al. Improving Neural Networks by Preventing Co-Adaptation of Feature Detectors. arXiv preprint arXiv:1207.0580, (2012). MLA
7. Gedeon, T. D., and H. S. Turner. "Explaining Student Grades Predicted by a Neural Network." In: Proceedings of 1993 International Conference on Neural Networks (IJCNN-93-Nagoya, Japan). Vol. 1. IEEE, (1993).
8. Hochreiter S, Schmidhuber J; Long short-term memory[J]. Neural computation, 9(8):1735-1780 (1997).
9. Gers F A, Schmidhuber J, Cummins F; Learning to forget: Continual prediction with lstm[J].Neural Computation (2000).

5 Appendix

num_epoch	hidden size	lr	batch siz e	dropout	accuracy	num_epoch	hidden size	lr	batch siz e	dropout	accuracy
300	16	0.01	30	0.0	23.08 %	300	16	0.01	30	0.2	33.33 %
500	16	0.01	30	0.0	23.08 %	500	16	0.01	30	0.2	25.64 %
300	16	0.001	30	0.0	23.08 %	300	16	0.001	30	0.2	20.51 %
500	16	0.001	30	0.0	30.77 %	500	16	0.001	30	0.2	33.33 %
300	16	0.01	50	0.0	20.51 %	300	16	0.01	50	0.2	15.38 %
500	16	0.01	50	0.0	23.08 %	500	16	0.01	50	0.2	28.21 %
300	16	0.001	50	0.0	23.08 %	300	16	0.001	50	0.2	20.51 %
500	16	0.001	50	0.0	23.08 %	500	16	0.001	50	0.2	17.95 %
300	16	0.01	70	0.0	25.64 %	300	16	0.01	70	0.2	33.33 %
500	16	0.01	70	0.0	23.08 %	500	16	0.01	70	0.2	28.21 %
300	16	0.001	70	0.0	30.77 %	300	16	0.001	70	0.2	30.77 %
500	16	0.001	70	0.0	35.90 %	500	16	0.001	70	0.2	17.95 %
300	64	0.01	30	0.0	25.64 %	300	64	0.01	30	0.2	33.33 %
500	64	0.01	30	0.0	25.64 %	500	64	0.01	30	0.2	25.64 %
300	64	0.001	30	0.0	28.21 %	300	64	0.001	30	0.2	30.77 %
500	64	0.001	30	0.0	28.21 %	500	64	0.001	30	0.2	28.21 %
300	64	0.01	50	0.0	25.64 %	300	64	0.01	50	0.2	20.51 %
500	64	0.01	50	0.0	25.64 %	500	64	0.01	50	0.2	33.33 %
300	64	0.001	50	0.0	20.51 %	300	64	0.001	50	0.2	23.08 %
500	64	0.001	50	0.0	28.21 %	500	64	0.001	50	0.2	23.08 %
300	64	0.01	70	0.0	20.51 %	300	64	0.01	70	0.2	23.08 %
500	64	0.01	70	0.0	23.08 %	500	64	0.01	70	0.2	30.77 %
300	64	0.001	70	0.0	23.08 %	300	64	0.001	70	0.2	35.90 %
500	64	0.001	70	0.0	28.21 %	500	64	0.001	70	0.2	35.90 %
300	128	0.01	30	0.0	23.08 %	300	128	0.01	30	0.2	23.08 %
500	128	0.01	30	0.0	28.21 %	500	128	0.01	30	0.2	15.38 %
300	128	0.001	30	0.0	35.90 %	300	128	0.001	30	0.2	15.38 %
500	128	0.001	30	0.0	23.08 %	500	128	0.001	30	0.2	30.77 %
300	128	0.01	50	0.0	25.64 %	300	128	0.01	50	0.2	17.95 %
500	128	0.01	50	0.0	23.08 %	500	128	0.01	50	0.2	23.08 %
300	128	0.001	50	0.0	17.95 %	300	128	0.001	50	0.2	17.95 %
500	128	0.001	50	0.0	25.64 %	500	128	0.001	50	0.2	33.33 %
300	128	0.01	70	0.0	30.77 %	300	128	0.01	70	0.2	25.64 %
500	128	0.01	70	0.0	33.33 %	500	128	0.01	70	0.2	33.33 %
300	128	0.001	70	0.0	20.51 %	300	128	0.001	70	0.2	30.77 %
500	128	0.001	70	0.0	28.21 %	500	128	0.001	70	0.2	23.08 %
300	256	0.01	30	0.0	25.64 %	300	256	0.01	30	0.2	28.21 %
500	256	0.01	30	0.0	25.64 %	500	256	0.01	30	0.2	28.21 %
300	256	0.001	30	0.0	23.08 %	300	256	0.001	30	0.2	20.51 %
500	256	0.001	30	0.0	25.64 %	500	256	0.001	30	0.2	28.21 %
300	256	0.01	50	0.0	23.08 %	300	256	0.01	50	0.2	20.51 %
500	256	0.01	50	0.0	28.21 %	500	256	0.01	50	0.2	20.51 %
300	256	0.001	50	0.0	20.51 %	300	256	0.001	50	0.2	12.82 %
500	256	0.001	50	0.0	20.51 %	500	256	0.001	50	0.2	23.08 %
300	256	0.01	70	0.0	23.08 %	300	256	0.01	70	0.2	30.77 %
500	256	0.01	70	0.0	28.21 %	500	256	0.01	70	0.2	23.08 %
300	256	0.001	70	0.0	23.08 %	300	256	0.001	70	0.2	23.08 %
500	256	0.001	70	0.0	33.33 %	500	256	0.001	70	0.2	20.51 %

Table 3. Hyper Parameters Table