Medical Diagnosis System Based on Fast-weights Scheme

TIAN Shining, LU Jihua, GU Boyu, WANG Huan

(*School of Information and Electronics*, *Beijing Institute of Technology*, *Beijing* 100081, *China*)

Abstract: Clinical examination data often have the features of carrying vague information, missing data and incomplete examination records, which lead to higher probabilities of misdiagnosis. A variational recursive-discriminant joint model with fast weights (FWs) scheme is proposed. MIMIC-III data sets are trained and tested, and the results are used to diagnosing. Variational recurrent neural network (VRNN) with FWs can better obtain the temporal features with partly missing data, and discriminant neural network (DNN) is for decision. Moreover, layer regularization (LN) avoids the overflow of loss function and stabilize the dynamic parameters of each layer. For the simulations, 10 laboratory tests were selected to predict 10 diseases, 1600 samples and 400 samples were used for training and testing, respectively. The test accuracy of disease diagnosis without FWs is 72.55%, and that with FWs is 85.80%. Simulations reveal that the FWs mechanism can effectively optimize the system model, abstracting the features for diagnose, and significantly improve the accuracy of decision-making.

Key words: Fast Weights Scheme, Discriminant Neural Network, Variational Recurrent Neural Network, Diagnosis Accuracy

1 Introduction

In recent years, Diagnose decision by Artificial Intelligent (AI)-aided diagnosing has become increasingly widespread in the medical field[1]. Traditional diagnosing mainly depends on clinical data and personal experience, whose subjective observation and experiences have a great influence on the diagnosis results. Due to the complexity of pathology and human fatigue potential, machine-assisted diagnosing has entered the field of machine learning which has promoted the rapid development of disease decision[2]. Deep learning does not require on-site diagnosis of patients, only a small amount of data processing. Moreover, if it is necessary, the components of the data that provide diagnostic information can be discovered through self-learning[3-4]. Therefore, diagnosing is changing from professional to machine-based processing. The unprecedented achievements of machine learning can attribute to the following factors: the rapid progress of CPU and GPU technology; the development of big data; the progress of machine learning algorithms[5-8].

Recent years, because of the superiority of recurrent neural network (RNN) in processing sequential data tasks, researchers have gradually introduced RNN into medical diagnosing. In 2008, Che and Miotto's team applied denoising auto-encoders (DAE)[9] to learn the hidden features and use these features to diagnose[10]. Aczon and Ledbetter have developed an intelligent diagnosis system based on treatment records of more than 12,000 patients in Intense Care Unit (ICU). They applied RNN to integrating the newly generated information sequence and making accurate diagnosing decision according to the clinical data of patients detected in recent period. Then, Lipton team put forward a RNN model based on long-term and short-term memory (LSTM), and its performance is better than many excellent baseline models[12]. In this network, there is a problem of missing data, which is partly solved by heuristic forward filling and back-tracking[12]. Their diagnostic work is based on LSTM, in which a missing value indicator is introduced as part of the input. It is observed that dealing with missing data is more reasonable to better improve the accuracies. Choi and et al. first proposed a method of using past medical records to predict future diseases, and developed a system of "Artificial Intelligence Doctors" to predict diseases and the suitable drugs based on electronic health records (EHR), past medical conditions and drug use[13]. They also put forward another important mechanism of machine learning and proposed the neural attention[14]. Then, Che et. al. proposed a diagnostic network based on stacked denoising Auto-encoder (SAE) and LSTM. Moreover, they proposed a gradient enhancement tree to lower the complexity of feature learning[14]. An improved structure of gated Recurrent unit (GRU) is proposed to deal with incomplete input[15]. The models in [13]-[15] are all discriminative ones, which cannot solve the problem of missing data fundamentally. To solve this problem, a generation model should be adopted. For example, the Gaussian mixture model is a typical generation model. The missing data problem can be easily solved by expectation maximization (EM) algorithm in it. Marlin's team used GMMs to predict mortality through clinical data training[16]. However, these typical generation models are shallow, linear and Gaussian. In recent years, two novel diagnosis approach based on attention scheme and RNN have been proposed in [17] and [18]. Researchers have proposed several deep generation models, such as VAE and VRNN[19]. Compared with the traditional generation models, joint model with VAE and VRNN shows more rules with complex conditional distribution, and the accuracies are also improved[20]. Our model induces the FWs scheme in the end-to-end architecture of [20], and achieved better accuracy performance.

The paper is outlined as follows. In Section II, the proposed model with FWs scheme are detailed. The MIMIC-III data set is preprocessed in Section III. It is used to trainingand testing the proposed joint model. The training and testing simulation results obtained by the proposed model of VRNN-NN with FWs scheme are compared in detail in Section IV. The effects of the selection of iteration times, attenuation rate and batch size of fast weight. The performances of the simulation results are concluded in Section V.

2 System Model

The model structure of Variational Recurrent Neural Networks with Fast Weights is shown in Fig.1. In Fig.1, the proposed model consists of joint model and FWs scheme. The joint model contains a Variational Auto-Encoder (VAE) and a Gated Recurrent Unit (GRU) network.

2.1 Analysis of the Joint Model

In VAE model, Latent variables follows the distribution of Gaussian, whose mean and standard are conditioned on the previous state $h_{t-1}$and input variable x_t.

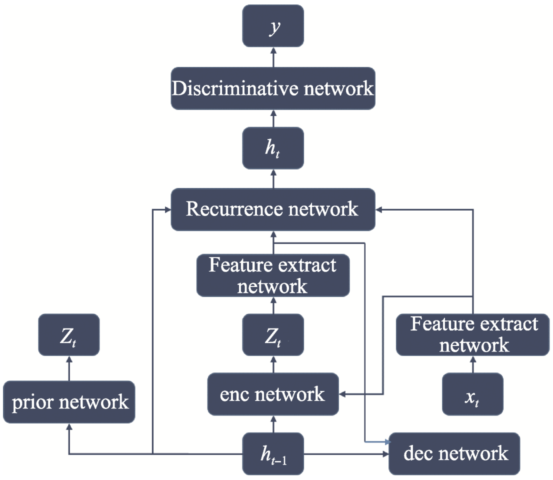

Fig.1 Diagram of Joint Model with Fast Weights

$$z_t = N(\mu_{z,t}, \sigma_{z,t}), where[\mu_{z,t}, \sigma_{z,t}] = \varphi^{enc}(x_t^{'}, h_{t-1}) \quad (1)$$

Here, $\mu_{z,t}$ and $\sigma_{z,t}$ represents the mean and the standard of the distribution, respectively. And $x_t^{'}$’ is feature of x_t, i.e. $x_t^{'} = \varphi_x(x_t)$. Then, define $z_t^{'}$ as the feature of z_t, i.e. $z_t^{'} = \varphi_z(z_t)$. Decoded X_t is defined as a Gaussian distribution variable.

$$X_t = N(\mu_{x,t}, \sigma_{x,t}), where[\mu_{x,t}, \sigma_{x,t}] = \varphi^{dec}(z_t, h_{t-1}) \quad (2)$$

where $\mu_{x,t}, \sigma_{x,t}$ and $\mu_{x,t}, \sigma_{x,t}$ represents the mean and the standard of the distribution, respectively. We define Z_t as the prior of latent variable.

$$Z_t = N(\mu_{0,t}, \sigma_{0,t}), where[\mu_{0,t}, \sigma_{0,t}] = \varphi^{prior}(h_{t-1}) \quad (3)$$

where $\mu_{0,t}, \sigma_{0,t}$ is the mean and the standard of Z_t. In the above process, φ^{dec},φ^{enc},φ^{prior},φ_z and φ_x are all linear neural networks. Then, combine z'_t and x'_t into u_t which is delivered to the next module as input, i.e. $u_t = [z'_t, x'_t]$.

The second module is a Gated Recurrent Unit (GRU) network, taking y_t as the input variable of GRU, and h_{t-1} as the previous state, recorded as H_{t-1}. The complete process of GRU is as follows.

$$r_t = Sigmoid(\varphi_{ir}(u_t) + \varphi_{hr}(H_{t-1})) \quad (4)$$

$$m_t = Sigmoid(\varphi_{im}(u_t) + \varphi_{hm}(H_{t-1})) \quad (5)$$

$$n_t = tanh(\varphi_{in}(u_t) + r_t\varphi_{hn}(H_{t-1})) \quad (6)$$

$$H_t = (1 - m_t)n_t + m_t H_{t-1} \quad (7)$$

r_t,m_t,n_t represents reset gate, update gate and new gate respectively, φ_{ir}, φ_{hr}, φ_{im}, φ_{hm}, φ_{in}, $\varphi_{hn}$are neural networks, H_t is the current state, which is delivered to the next module as state variable, $Sigmoid$ is the sigmoid function, written by $Sigmoid(x) = \frac{1}{1+e^{-x}}$.

2.2 Analysis of Fast Weights Scheme

The last module is Fast Weight. In this module, we introduce Fast Weight to describe the short-term correlation of data H_t, it is used as the preliminary vector, H_{t-1} as the previous state vector, y_t as the input. Fast weight update process is shown as follows.

$$A_t = lrA_{t-1} + dr[H_t{}^T * H_t] \quad (8)$$

where A_t is fast weight in time t, lr, dr is the learning rate and delay rate in fast weight updating. The state variable H_t^s will be updated for S times. The process of each update is as follows.

$$H_t^s = f\left(LN(u_t W_y + b_y + H_{t-1}W_h + H_t^{s-1}A_t)\right) \quad (9)$$

where W_y and W_h are slow weight matrices, A_t is fast weight matrix, $LN(.)$represents layer normalization, $f(.)$ represents the ReLU function, i.e. $f(x) = max(0, x)$. After 3 cycles, we get the final output vector H_t^S as the state variable at time t, recorded as h_t. We take the average ofh_t ,i.e. $\tilde{h} = average(h_1, h_2, \cdots, h_T)$, as the input of a Neural Networkφ_d. The output of NN is the prediction results of the proposed joint model, which is:

$$y = \varphi_d(\tilde{h}) \quad (10)$$

2.3 Calculation of Cross-entropy

The loss function of model consists of two parts, generative loss in VAE module and discriminative loss in NN, in which Generative loss is the Kullback–Leibler divergence between Z_t and z_t, and the Negative Log Likelihood between X_t and x_t, and discriminative loss is the Cross_Entropy between Y_n and y_n, where Y_n is the labels of data.

$$L_g = \sum_n^N \sum_t^T \left(Z_t^n \, log\left(\frac{Z_t^n}{z_t^n}\right) - (x_t^n \, log \, X_t^n + X_t^n \, log \, x_t^n) \right) \quad (11)$$

$$L_d = \textstyle\sum_n^N - log \frac{e^{Y_n}}{\sum_{j=1}^N e^{y_n(j)}} \quad (12)$$

where N is the number of classification. The overall loss is written as follow.

$$L = \sum_n^N \sum_t^T \left(Z_t^n \, log\left(\frac{Z_t^n}{z_t^n}\right) - (x_t^n \, log \, X_t^n + X_t^n \, log \, x_t^n) \right) + \eta * \textstyle\sum_n^N - log \frac{e^{Y_n}}{\sum_{j=1}^N e^{y_n(j)}} \quad (13)$$

3 Preprocessing of MIMIC-III Data

In this section, the VRNN+NN_FW model is trained and tested with MIMIC-III dataset. The laboratory tests data of patients during hospitalization are taken as inputs,

Step 1) Screening samples;

Specifically, M diseases and N inspection items with the largest number of samples were screened from MIMIC data sets, and data not belonging to these N diseases in "LABEVENT.csv" and data not belonging to this M disease in "DIAGNOSES_IDC.csv" were removed. The screened data and labels were saved into files "M_ITEMID.csv" and "N_ICD9CODE.csv" respectively.

Among them, the number of samples corresponds to the number of patients, referring to the number of patients related to the same related diseases and examination items.

Step 2) Data label matching;

Unify the data and labels of patients, that is, screen out the common items in "M_ITEMID.csv" and "N_ICD9CODE.csv", and remove the data of other patients, so that the data and labels correspond one by one, and save the results into files.

Step 3) Screening time nodes;

Delete the time nodes whose number of checks is less than or equal to K at a certain time and save them as files.

Step 4) Screening patients;

First set up the time step and then screen out the items of the patients' number larger than the time step and saved into files. By step 4, the examination items are screened in the process of screening patients.

Step 5) Replenish the missing data;

Replenish the data of the time nodes whose number of checks is greater than K and less than N by calculating the average value of N checks separately. Then, fill the missing data in the average value of corresponding checks. Finally, the results are saved into files.

Step 6) Data truncation;

Truncate the first TIME_STEP replenished data for each patient and intercept them as input.

Step 7) Screening labels;

By Step 4), the training data and the training labels are screened according to the examination items. After screening, the training labels are one-dimensional data with x lines of training labels. The results of the inspection items of the samples are kept as the training data.

Step 8) Separate the data;

Specifically, the training data is divided into training labels and test labels in the same proportion. X% is saved as training data for training, Y% is saved as test data for testing.

Among them, X%+Y%=100%;

Step 9) Data standardization;

That is, normalization of one-dimensional data in training data that has become line x*y*z.

Then the trained model is tested to get the predicted accuracy.

4 Simulations of Joint Model with FWs Scheme

The testing accuracies of three models are plotted in Fig.2, where VRNN, VRNN+NN and VRNN+NN_FW models are trained and tested respectively.

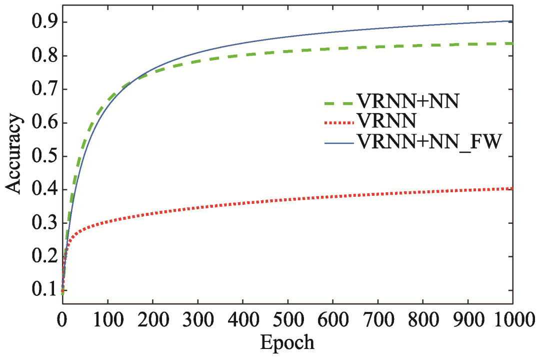

Fig.2 Accuracy versus Epoch for three models of VRNN, VRNN+NN, and VRNN+NN with FWs scheme, respectively

We observe from Fig.2 that the proposed joint model of VRNN+NN with FWs is superior to VRNN+NN and VRNN in terms of accuracy. This is reasonable, because the introduction of fast weight is a better way to express the relationship between short-term temporal data. VRNN_NN model performs better than VRNN in accuracy. This shows that the joint training considering classification targets performs better in feature learning.

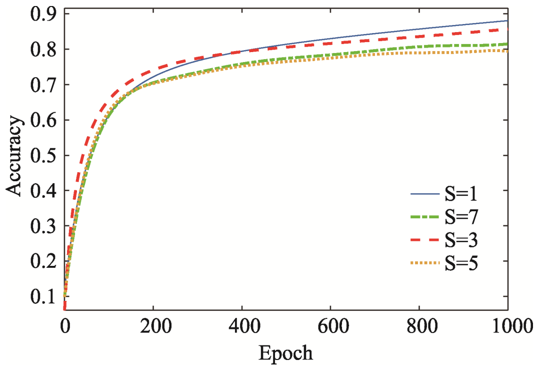

Fig.3 Testing accuracy versus epoch for FWs step = 1, 3, 5 and 7, respectively

From Fig.3, it shows that the training speed is better than others when fast weight iteration number equals to 3, and the accuracy performance achieves the best when fast weight iteration number equals to 1. It does not show the potential rule that the larger S is, the higher the accuracy is, or the faster the training speed is.

The testing accuracy versus epoch for decay rates of 0.3, 0.5, 0.7 and 0.95 are simulated in Fig.4.

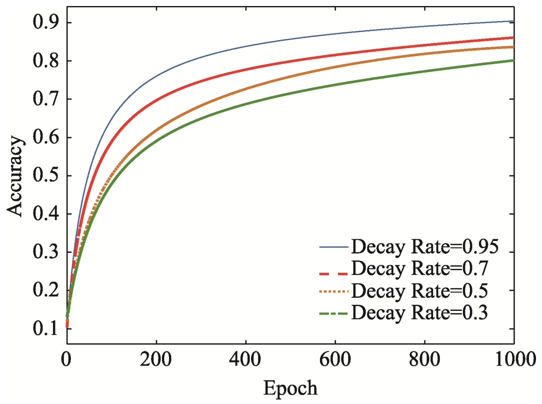

Fig.4 Testing accuracy versus epoch for decay rate = 0.3, 0.5, 0.7 and 0.95, respectively

From Fig.4, it can be seen clearly that the decay rate is positively correlated with the accuracy, that is, the higher the decay rate, the higher the accuracy.

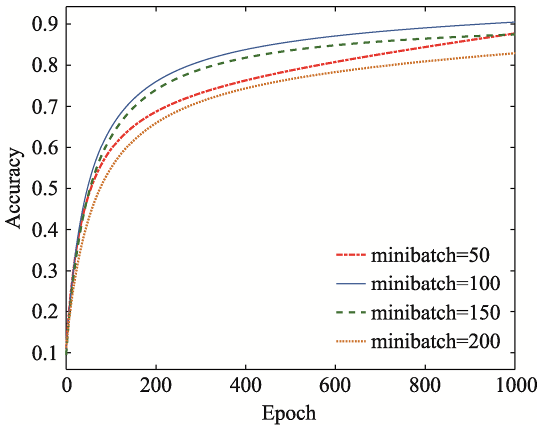

Fig.5 Testing accuracy versus epochs for batch size=50, 100, 150 and 200, respectively

From Fig.5, the accuracy reaches the maximum when the batch size is 100, and the accuracy becomes smaller as the batch size increases when it is more than 100. The accuracy becomes smaller as the batch size decreases when it is less than 100.

5 Conclusions

Improving the accuracy of medical decision is an important research direction in the field of AI-aided diagnosing. The proposed algorithm has achieved quite high accuracy, which is larger than 70%. However, to better express the temporal memorizing feature between the input data, and further improve the decision accuracy. We proposed a joint VRNN-NN model with FWs scheme. By training and testing on the pre-processed MIMIC-III dataset, the accuracies are greatly improved. In the training and testing process, the relationship between a series of important parameters such as, number of iterations, batch size, decay rate are simulated and compared. The results show that the higher the decay rate, the higher the training and testing accuracy. However, there is no trend consistency between the number of iterations, batch size and accuracy. When the batch size is 100, the accuracy of disease diagnosis is increased from 72.55% without FWs to 85.80% with FWs.

Acknowledgements

This work is supported by National Natural Science Foundation of China “Research on non- orthogonal multiple access technology for unauthorized transmission” (No. 61771051) and “Research on a new emergency positioning system for the integration of visible-light communication and MEMS inertial navigation” (No. 61675025).

References

[1] Shen, D., G. Wu, H. Suk. (2017). Deep Learning in Medical Image Analysis. *Annual Review of Biomedical Engineering.*19(1): p. 221-248.

[2] Schmidhuber J. (2015). Deep learning in neural networks: *An overview. Neural Networks.* 61: 85-117.

[3] Bengio Y. (2009). Learning deep architectures for ai. *Foundations and Trends in Machine Learning.* 2: 1-127.

[4] LeCun Y, Bengio Y, Hinton G. (2015). Deep learning. *Nature*. 521: 436-444.

[5] Hinton GE, Salakhutdinov RR. (2006). Reducing the dimensionality of data with neural networks. *Science*. 313: 504-507.

[6] Vincent P, Larochelle H, Lajoie I, et al. (2010). Stacked Denoising auto- encoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*. 11:3371-3408.

[7] Nair V, Hinton GE. (2010). Rectified linear units improve restricted Boltzmann machines. *Proceedings of International Conference on Machine Learning* (ICML).

[8] Srivastava N, Hinton G, Krizhevsky A, et al. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*.15:1929-1958.

[9] P Vincent. (2008). Extracting and composing robust features with denoising autoencoders. *In Proceedings of 25th International Conference on Machine Learning*.1096-1103.

[10] Z Che, D Kale. (2015). Deep Computational Phenotyping. *In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.507-516.

[11] M Aczon, D Ledbetter. (2017). Dynamic Mortality Risk Predictions in Pediatric Critical Care Using Recurrent Neural Networks. arXiv: 1701.06675.

[12] ZC Lipton. (2015). Learning to Diagnose with LSTM Recurrent Neural Networks. arXiv:1511.03677.

[13] Edward Choi, Doctor AI. (2016). Predicting Clinical Events via Recurrent Neural Networks[A]. *In Proceedings of the 1st Machine Learning for Healthcare Conference*. PMLR 56:301-318.

[14] Z Che. (2015). Distilling Knowledge from Deep Networks with Applications to Healthcare Domain. *Annales De Chirurgie*, 40 (8):529-32.

[15] Z Che. (2018). Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Scientific Reports*.8:6085.

[16] BM Marlin. (2012).Unsupervised pattern discovery in electronic health care data using probabilistic clustering models. *In Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium.* 389-398.

[17] Hojin Lee; Hyeyun Jeong; Gyogwon Koo; et al. 2020. Attention RNN Based Severity Estimation Method for Interturn Short-Circuit Fault in PMSMs. *IEEE Transactions on Industrial Electronics.*

[18] Wai-Kim Leung ; Xunying Liu ; Helen Meng. (2019). CNN-RNN-CTC Based End-to-end Mispronunciation Detection and Diagnosis.*ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

[19] Chung, J., et al. (2015). A Recurrent Latent Variable Model for Sequential Data.

[20] Shiyue Zhang, Pengtao Xie, Dong Wang, et al. (2017). Medical Diagnosis From Laboratory Tests by combining Generative and Discriminative Learning. Arxiv: 1711.04329v2.

Author biographies

TIAN Shining, received the B.Sc. degree in Telecommunications engineering from the Beijing Institute of Technology, China, in 2018. He is currently pursuing the M.S. degree in Communication and Information System from Beijing Institute of Technology, Beijing, China. His research interests include digital predistortion of nonlinear power amplifiers and the wideband receiver technology.

Email: 1120141155@bit.edu.cn

LU Jihua, (M’2012), corresponding author, received her B.S. degree in Electronics and Engineering from Shandong University, Shandong, China in 2000. Then, she received the M.S. and Ph. D. degrees in Communication and Information System from Beijing Institute of Technology, Beijing, China, in 2003 and 2012, respectively. She is currently an assistant professor in School of Information and

Electronics at Beijing Institute of Technology, Beijing, China. From 2013 to 2015, she was a visiting scholar of the Institute of Georgia Technology, Atlanta, USA. Her research interests include the secrecy capacity analysis, wireless channel modelling, synchronization & detection, performance analysis for wireless system, and machine learning.

Email: lujihua@bit.edu.cn.

GU Boyu, received the B.Sc. degree in Telecommunications engineering from the Beijing Institute of Technology, China, in 2020. He now works at China Mobile. His research interests include computer vision and machine learning.

Email: co52195144@163.com.

WANG Huan, received the B.Sc. degree in Automation major from three gorges university, China, in 2010. Then, she received the M.S. degrees in Integrated circuit engineering from Beijing Institute of Technology, Beijing, China, in 2020. She now works at HORIBA Precision Instruments (Beijing) Co., Ltd.. Her research interests include computer vision and machine learning.

Email: 349382752@163.com.