Disease Diagnosis Prediction of EMR Based on BiGRU-Att-CapsNetwork Model

Pin Ni*†, Yuming Li*†, Jiayi Zhu*, Junkun Peng*, Zhenjin Dai*, Gangmin Li*‡ and Xuming Bai§
*Research Lab for Knowledge and Wisdom, Xi'an Jiaotong-Liverpool University, China
†Department of Computer Science, University of Liverpool, UK
§Department of Interventional Radiology, Second Affiliated Hospital of Soochow University, China
Email: ‡Gangmin.Li@xjtlu.edu.cn, §2005baixuming@163.com

***Abstract*—Electronic Medical Records (EMR) carry a large number of diseases characteristics, history and other specific details of patients, which has great value for medical diagnosis. These data with diagnostic labels can help automated diagnostic assistant to predict disease diagnosis and provide a rapid diagnostic reference for doctors. In this study, we designed a BiGRU-Att-CapsNetwork model based on our proposed CMed-BERT Chinese medical domain pre-trained language model to predict disease diagnosis in Chinese EMR. In the wide-ranging comparative experiments involving a real EMR dataset (SAHSU) and an academic evaluation task dataset (CCKS 2019), our model obtained competitive performance.**

***Index Terms*—Disease Diagnosis Prediction, Capsule Network, BiGRU, Attention, EMR**

I. Introduction

EMR contains a large number of health information that is closely related to the patient's condition. Among them, chief complaint, history of present illness, initial diagnosis and other information in the EMR have a high reference value for subsequent diagnosis and treatment. These data are also an important resource for automated medical diagnostic assistants. Through modeling a large number of historical EMR text data containing the "Admitting Diagnosis" label to predict diseases with similar symptoms, can provide doctors with a rapid and accurate reference during the actual diagnosis process.

Therefore, this study uses the CMed-BERT language model through pre-training the text data contains 36681228 (36M) Chinese character and constructed BiGRU-Att-CapsNetwork to model the context of EMR of each disease, through the two directions (front to back and back to front) of the EMR text to capture semantics and finer-grained information, and encapsulate all detected feature states into a vector form through the Capsule Network [1] to achieve the modeling of the orientation and relative spatial relationship between the high-level features and the low-level features and finally converted to the vector containing the highest feature in the text by the flattening operation, and output by the fully connected dense layer with a Softmax to achieve preliminary disease diagnosis prediction of the content of the EMR text data. We trained and tested a dataset integrated a real-world EMR dataset SAHSU and an academic evaluation task dataset CCKS 2019 to predict 44 major diseases and compare 9 other previous state-of-the-art deep learning models. The final test results demonstrate the effectiveness of the constructed model structure. As far as we know, this is the first disease diagnosis prediction model structure to realize the number of multi-category and inter-disciplinary diseases in Chinese EMR. The experimental results show that our method has obtained competitive results in disease diagnosis prediction.

II. Literature Review

Meystre et al. [2] found out most systems mostly based on two different sets of approaches: pattern matching and machine learning. Li et al. [3] and Gao et al. [4] proved that Natural Language Processing is the best method to solve the task of electronic medical record processing. Demner et al. [5] and Nadkarni et al. [6] proposed that it is one of the most commonly used methods of unstructured medical data processing. Promoting the development of the medical field through artificial intelligence has become the common goal of computer science and medical field [7]–[9].

Capsule Network firstly proposed by Sabour et al. [1] for image classification task, after that, the theory has been applied to many fields such as text classification [10], relation extraction [11], intent detection [12], and hypernymy relationship detection [13]. Yu et al. [14] used Bi-GRU with attention mechanism for sentiment analysis at the aspect level. Khayibi et al. [15] proposed a student answer assessment model based on Bi-GRU and Capsule network, and got a higher accuracy (72.5%). Chen et al. [16] use Bi-GRU to capture the context semantics, then use Capsule Network and softmax as a classifier for question target classification.

III. Methodology

The core of our model consists of the following three parts:

A. BiGRU Layer

The two sub-networks (forward $\overrightarrow{f_W}$ and backwards $\overleftarrow{f_W}$) model the two directions of the context (back to the front and front to the back), respectively.

Cho et al. [17] proposed a Gated Recursive Unit (GRU) that allows each recursive unit to adaptively capture dependencies on different time scales. The Bi-GRU layer models the sequential sentence information and calculates as follows:

$$z_i = \sigma(W_z X_i + V_z h_{i-1} + b_z) \quad (1)$$

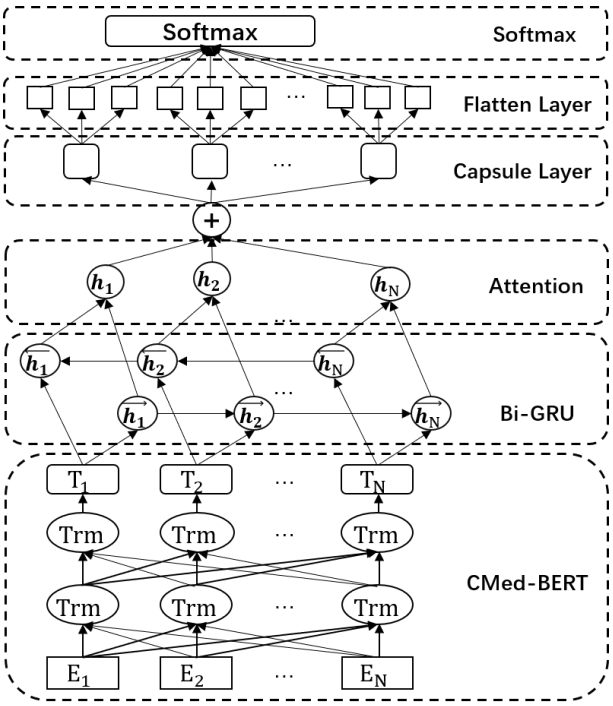

Fig. 1. The Structure of BiGRU-Att-CapsNetwork Model

$$r_i = \sigma(W_r X_i + V_r h_{i-1} + b_r) \tag{2}$$

$$\widetilde{h_i} = tanh(W X_i + V(r_i * h_{i-1}) + b) \tag{3}$$

$$h_i = z_i * h_{i-1} + (1 - z_i) * c_i \tag{4}$$

Where X_i is the i^{th} vector value output by the pre-training layer, h_{i-1} represents the previous hidden state, and z_i and r_i represent the update gate and the reset gate, respectively. The update gate is used to control the degree to which the status information of the previous moment is brought into the current state. The larger the value of the update gate is, the more the status information is brought in at the previous moment. The reset gate controls how much information is written to the current candidate set c_i in the previous state. The smaller the reset gate, the less the information of the previous state is written. σ represents the sigmoid function and θ represents the product of the matrix. And two fundamental components composite the GRU: an update gate z is used to decide whether to hide the state to update the new hidden state $\tilde{h}$ or not and reset gate r to decide the ignore the previous hidden state. Finally, merge the forward and backwards information and output the vector h_i for each i word.

B. Attention Layer

In this layer, we mainly use the Attention to enhance the semantic vector representation of the target word in context.

The constituent elements in the $Source$ can be considered to be composed of a series of $< Key, Value >$ data pairs. In this case, given a certain element $Query$ in the $Target$, by calculating the similarity or correlation between $Query$ and each Key, the weight coefficient of each Key corresponding to $Value$ can be obtained. And the weighted sum of $Value$ to obtain the final Attention $Value$. Therefore, the Attention mechanism actual is to weight the sum of the $Value$ of the elements in the $Source$, and $Query$ and Key are used to calculate the weight coefficients of the corresponding $Value$. The whole conception can be represented as the following equation:

$$\begin{aligned} &Attention(Query, Source) = \\ &\textstyle\sum_{i=1}^{L_x} similarity(Query, Key_i) \times Value_i \end{aligned} \tag{5}$$

Which, L_x represents the length of the $Source$. The role of Attention here can be to selectively screen out and focus on a small amount of important information from a large amount of information, ignoring most of the minor information. Therefore, we use Attention to enhance the semantic representation of feature words in context.

C. Capsule Layer

The prediction vector $u_{j|i}$ can be seen as each of the capsule neurons in the last layer outputs to a neuron in the next layer with different intensity connections. This is calculated by matrix multiplication of the BiGRU layer output vector v_i and the transformation matrix W'_{ij}, which encodes the important spatial relationship between the low-level and high-level features within text (Equation 6).

$$\hat{u}_{j|i} = W'_{ij} v_i \tag{6}$$

Perform the weighted sum of the input vectors, find the coupling coefficient c through the dynamic routing algorithm (Equation 7, 8).

$$c_{ij} = \frac{\exp(b_{ij})}{\sum_k \exp(b_{ik})} \tag{7}$$

$$s_j = \sum_i c_{ij} \hat{u}_{j|i} \tag{8}$$

And use the non-linear "Squash" activation function to compress the vector to the interval length between $0 - 1$ as the probability, while retaining its orientation, which the calculation of output V is as follows (Equation 9).

$$\mathbf{v}_j = \frac{\|\mathbf{s}_j\|^2}{1 + \|\mathbf{s}_j\|^2} \frac{\mathbf{s}_j}{\|\mathbf{s}_j\|} \tag{9}$$

IV. Experiment and Results

The experimental environment is based on GPU: Dual NVIDIA 1080Ti GPU (11 GB Memory/Card), CPU: Intel Xeon E5-2678 v3 Dual 2.50GHz, RAM: 64GB. And the experimental results of the diagnosis prediction of 44 diseases in real-world EMRs show that BiGRU-Att-CapsNetwork model has a greater improvement than other mainstream deep learning models (Fig. 2, 3). With 100 epochs (10 epochs early stopping) and 512 Sequence Length training settings, the F1-Score of our model reached 73.89% (Table I). At the same time, our CMed-BERT has also been proven by the experimental results, which can improve the performance of domain downstream tasks compared to the results of most of the original BERT-based downstream models.

TABLE I
PREDICTION RESULTS OF COMPARISON MODELS

Model	Original Model			Model	Pre-trained Model		
	Precision	Recall	F1 Score		Precision	Recall	F1 Score
BERT-CNN-LSTM	0.4287	0.3762	0.3823	CMed-BERT-CNN-LSTM	0.5731	0.5273	0.5285
BERT-CNN	0.5544	0.5300	0.5103	CMed-BERT-CNN	0.5921	0.5562	0.5448
BERT-BiGRU	0.6091	0.5918	0.5741	CMed-BERT-BiGRU	0.5971	0.5829	0.5631
BERT-CNN-GRU	0.6296	0.5835	0.5791	CMed-BERT-CNN-GRU	0.6073	0.5802	0.5745
BERT-BiLSTM	0.6006	0.5696	0.5657	CMed-BERT-BiLSTM	0.6359	0.6259	0.6136
BERT-Dropout-BiGRU	0.6561	0.6506	0.6310	CMed-BERT-Dropout-BiGRU	0.7063	0.6826	0.6730
BERT-Dropout-AVRNN	0.6862	0.6585	0.6417	CMed-BERT-Dropout-AVRNN	0.6996	0.6861	0.6745
BERT-AVRNN	0.6610	0.6701	0.6487	CMed-BERT-AVRNN	0.6998	0.6971	0.6800
BERT-AVCNN	0.6706	0.6605	0.6487	CMed-BERT-AVCNN	0.6976	0.7158	0.6911
BERT-BiGRU-Att-CapsNet.	**0.7463**	**0.7410**	**0.7295**	**CMed-BERT-BiGRU-Att-CapsNet.**	**0.7654**	**0.7590**	**0.7389**

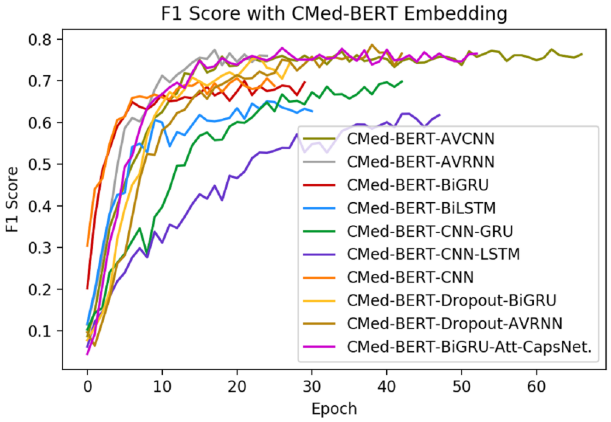

Fig. 2. The Training Process of Downstream Models with CMed-BERT Embedding

V. CONCLUSION

In this study, We designed BiGRU-Att-CapsNetwork to realize disease diagnosis prediction in EMR based on the CMedBERT language model. BiGRU-Att-CapsNetwork can use BiGRU to capture semantic dependence and finer-grained information from the front and back directions of the medical record; through the encoding of the Attention layer, it can be better identify important features in lower-level capsules; and Capsule Network can realize the modeling of the orientation and relative spatial relationship between the high-level features and the low-level features. By comparing 9 mainstream deep learning models, our model achieved 73.89% F1-Score in the real-world EMR dataset, which is higher than other models. This also provides support for more accurate Chinese EMR text diagnosis prediction in the clinical auxiliary diagnosis system.

VI. ACKNOWLEDGEMENT

This work is partially supported by the AI University Research Centre (AI-URC) through XJTLU Key Programme Special Fund (KSF-P-02) and KSF-A-17. And it is also partially supported by Suzhou Science and Technology Programme Key Industrial Technology Innovation programme with project code SYG201840. We appreciate their support and guidance.

REFERENCES

[1] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Advances in neural information processing systems*, 2017, pp. 3856–3866.
[2] S. M. Meystre, F. J. Friedlin, B. R. South, S. Shen, and M. H. Samore, "Automatic de-identification of textual documents in the electronic health record: a review of recent research," *BMC medical research methodology*, vol. 10, no. 1, p. 70, 2010.
[3] L. Li, R. Zhou, and D. Huang, "Two-phase biomedical named entity recognition using crfs," *Computational biology and chemistry*, vol. 33, no. 4, pp. 334–338, 2009.
[4] H. Gao, E. J. A. Bowles, D. Carrell, and D. S. Buist, "Using natural language processing to extract mammographic findings," *Journal of biomedical informatics*, vol. 54, pp. 77–84, 2015.
[5] D. Demner-Fushman, W. W. Chapman, and C. J. McDonald, "What can natural language processing do for clinical decision support?" *Journal of biomedical informatics*, vol. 42, no. 5, pp. 760–772, 2009.
[6] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Natural language processing: an introduction," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 544–551, 2011.
[7] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel, "Learning to diagnose with lstm recurrent neural networks," *arXiv preprint arXiv:1511.03677*, 2015.
[8] T. Cohen and M. Welling, "Group equivariant convolutional networks," in *International conference on machine learning*, 2016, pp. 2990–2999.
[9] N. Razavian, J. Marcus, and D. Sontag, "Multi-task prediction of disease onsets from longitudinal laboratory tests," in *Machine Learning for Healthcare Conference*, 2016, pp. 73–100.
[10] J. Kim, S. Jang, E. Park, and S. Choi, "Text classification using capsules," *Neurocomputing*, 2019.
[11] N. Zhang, S. Deng, Z. Sun, X. Chen, W. Zhang, and H. Chen, "Attention-based capsule networks with dynamic routing for relation extraction," *arXiv preprint arXiv:1812.11321*, 2018.
[12] C. Xia, C. Zhang, X. Yan, Y. Chang, and P. S. Yu, "Zero-shot user intent detection via capsule neural networks," *arXiv preprint arXiv:1809.00385*, 2018.
[13] Q. Wang, C. Xu, Y. Zhou, T. Ruan, D. Gao, and P. He, "An attention-based bi-gru-capsnet model for hypernymy detection between compound entities," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2018, pp. 1031–1035.
[14] C. Yu, S. Wang, and J. Guo, "Learning chinese word segmentation based on bidirectional gru-crf and cnn network model," *International Journal of Technology and Human Interaction (IJTHI)*, vol. 15, no. 3, pp. 47–62, 2019.
[15] N. A. Khayi and V. Rus, "Bi-gru capsule networks for student answers assessment."
[16] S. Chen, B. Zheng, and T. Hao, "Capsule-based bidirectional gated recurrent unit networks for question target classification," in *China Conference on Information Retrieval*. Springer, 2018, pp. 67–77.
[17] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.