IALP 2019, Shanghai, Nov 15-17, 2019

BERT with Enhanced Layer for Assistant Diagnosis Based on Chinese Obstetric EMRs

Kunli Zhang*†, Chuang Liu*, Xuemin Duan*†, Lijuan Zhou*, Yueshu Zhao‡† and Hongying Zan*†
*School of Information Engineering
Zhengzhou University
†Peng Cheng Laboratory
‡The Third Affiliated Hospital of Zhengzhou University
Email : ieklzhang@zzu.edu.cn ; 214674227@qq.com ; xueminduan@stu.zzu.edu.cn ;
ieljzhou@zzu.edu.cn ;zyswr@163.com ; iehyzan@zzu.edu.cn

***Abstract*—This paper proposes a novel method based on the language representation model called BERT (Bidirectional Encoder Representations from Transformers) for Obstetric assistant diagnosis on Chinese obstetric EMRs (Electronic Medical Records). To aggregate more information for final output, an enhanced layer is augmented to the BERT model. In particular, the enhanced layer in this paper is constructed based on strategy 1(A strategy) and/or strategy 2(A-AP strategy). The proposed method is evaluated on two datasets including Chinese Obstetric EMRs dataset and Arxiv Academic Paper Dataset (AAPD). The experimental results show that the proposed method based on BERT improves the F1 value by 19.58% and 2.71% over the state-of-the-art methods, and the proposed method based on BERT and the enhanced layer by strategy 2 improves the F1 value by 0.7% and 0.3% (strategy 1 improves the F1 value by 0.68% and 0.1%) over the method without adding enhanced layer respectively on Obstetric EMRs dataset and AAPD dataset.**

***Keywords*-EMRs; assistant diagnosis; BERT; the enhanced layer;**

I. Introduction

EMRs(Electronic Medical Records) are detailed records of medical activities by medical personnel. The most important form of EMRs is free text data. With the development of medical informatization, hospitals have accumulated massive amounts of EMRs. These EMRs contain a lot of medical knowledge and patients health information. It is one of the most important tasks in the medical field to use NLP(Natural Language Processing) technology for assistant diagnosis based on EMRs. Since family planning was issued as one of the fundamental state policies in China, the policy of late marriage and late childbearing has brought many benefits. However, it has also led to an increase in the proportion of older pregnant women over 35 years of age [1]. After the implementation of China's Universal Two-child Policy in 2016, the proportion of older pregnant women will has become greater. The incidence of dystocia, fetal malformations and complications among older pregnant women is higher than that of normal pregnant women, it will be a great challenge for obstetrics in medical institutions to solve this problem. EMRs not only records the patient's complaint, physical examination, auxiliary examination and other information, but also records the doctor's initial diagnosis, diagnosis based on differential diagnosis and treatment plan. Usually, ad-mission diagnosis includes normal diagnosis, pathological diagnosis and description of complications, rather than a single diagnosis. In this paper, we treat the obstetric diagnostic task of EMRs as text multi-label classification task. To solve this problem [2] proposed the BERT model, which not only improves greatly on multiple data sets of different tasks, but also adapts to different tasks only by fine-tuning the pre-training version of BERT. On the basis of real EMRs, through screening and processing the original medical records, we transform tasks into multi-label text classification tasks. The contribution of this paper are as follows.

- To the best of our knowledge, BERT was firstly applied to the auxiliary diagnosis of Chinese EMRs.
- An enhanced layer was augmented to the BERT model based on two strategies for further improvement of the diagnosis effect.
- The enhanced layer works equally well on other domains of dataset.

Experiments on EMRs datasets and public datasets of text multi-label classification show that the results of Bert with enhanced layer model have been improved, which demonstrates the effectiveness and generality of the enhanced layer.

II. Related Works

Traditional multi-label classification mainly transforms multi-label classification tasks into multi-classification problem. In the neural network learning, changing the deep learning model and the loss function in order to improve the effect of multi-label classification [3]–[10]. Pre-training technology has become the corner-stone of NLP task. Pre-training technology can effectively improve the performance of NLP task results. Among them, word embedding technology [11], [12] has also made considerable progress, and has become a standard technology in different tasks. However, word embedding technology also has some drawbacks, such as the inability to distinguish polysemy. For example, the inaccuracy of word segmentation and the inability to distinguish polysemy will affect the quality of word or word encoding. In recent studies, pre-trained language models can effectively address the above shortcomings, such as ELMO [13], OpenAI GPT [14], and BERT [2]. In addition to using word embedding technology, these language models also use different deep learning coders for context encoding, in which ELMO uses Bi-LSTM [15], GPT uses one-way Transformer [16]. Coder, BERT uses bidirectional Transformer encoder. The above language model has obvious improvement on different NLP tasks, and can be applied to different NLP tasks only by fine-tuning the output of different NLP tasks. In the diagnosis of Obstetric based on EMRs, [17] proposed a multi-label classification method to study the problem of assistant diagnosis based on Chinese obstetric EMR. [18] used vector stitching method to fuse the numerical characteristics of EMRs for experiments, which improved the effect of assistant diagnosis. The traditional multi-label classification model is used in both tasks, and the experimental dataset is about 10,000. On the basis of expanding the dataset, this paper will use the pre-training model to study the problem of assistant diagnosis based on Chinese obstetric EMRs.

III. Model and Method

This section presents the details of the proposed model. Firstly, we will give the overall structure of the model, which is divided into three parts: encoding layer, enhanced layer and output layer. Then the three parts are detailed in the following.

A. Overview

In this paper, we treat the assistant diagnosis task of obstetric EMRs as multi-label text classification task. Let $\chi = R^d$ be a d dimensional instance space, $\mathbf{y} = y_1, y_2, ..., y_q$ is a set of q categories.Given a training set , where each instance $T = (x_1, Y_1), (x_2, Y_2), ..., (x_m, Y_m)(x_i \in \chi, Y_i \subseteq y)$ is a d dimensional feature vector, the goal of multi-label learning is to learn a multi-label classifier which satisfies some evaluation criteria. The model consists of Encoding layer, Enhanced layer and Output layer. The encoding layer uses BERT to obtain all the hidden layer representations of input sequences and hidden representation [C] for classification tasks (represented by C shown in Figure1). The design goal of enhanced layer is to further enhance the [C] hidden layer representations to cover as much sequence information as possible. Detailed content is described in enhanced layer. The main structure of the model is shown in Figure1.

B. BERT

BERT is a encoding structure composed of bidirectional transformer model, in which the transformer model is the attention encoding model proposed by [18]. Transformer uses a multi-head attention mechanism and each head calculates the attention weight independently. Then the model splices the results of each head, so the multi-head attention mechanism can be represented at different levels of the sequence. Because location information cannot be obtained in simple attention calculation, Transformer uses special position vector encoding for sequence location information. Due to the performance advantages of transformer model, it has recently become one of the most important

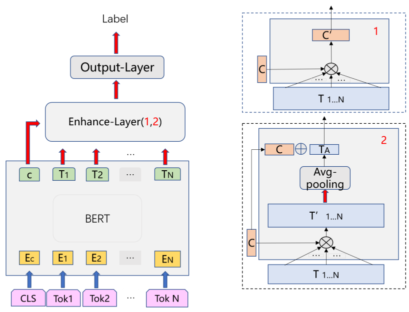

Figure 1. The left side of the figure is the framework of the proposed model. After sequence input, it first generates one additional output representation which is represented by C shown in Figure1 through BERT model, then generates a better hidden layer representation by enhanced layer. The right side of the figure illustrates two strategies of enhanced layer. Among them, red number 1 represents attention strategy (namely, A strategy) and red number 2 represents Attention-average pooling strategy (namely, A-AP strategy).

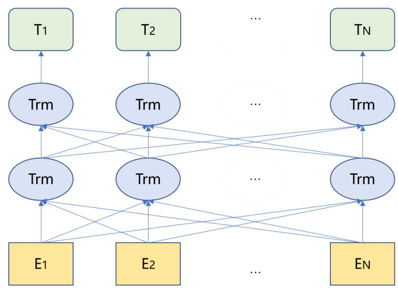

Figure 2. BERT uses a bidirectional transformer structure (represented by a blue elliptical Trm), and BERT encoding connects information in both directions of the context. Figure adapted from [2].

models in NLP field, and has been widely used in various sub-tasks, so the details of transformer model are not elaborated in this paper. In BERT, the input of the model can be a single sequence or a sequence pair. In this paper, we regard EMRs as single sequence to do experiments. In the Chinese task, a sequence consists of three parts: embedding representation of each independent Chinese character, location embedding and input marking. In the output of BERT, in addition to the classified representation embedding [C] mentioned in the previous section, all the representations of the sequence are also recorded as hi, where i is the position of each word in the input sequence x, all the sequence representations are defined as Hm, and m represents the number of words in the text. In order to obtain better classification representation for final classification, the self-attention mechanism is introduced to further enhance the output representation of encoding. Because BERT has been encoded by several transformer blocks, each position of each layer sequence is encoded by transformer blocks,we can see the construct of BERT in Figure2. But it is noteworthy that when the original BERT model is output, only one position encoding information is used as the input of the latter classifier, which is the position of [C] in Figure1. Although this position is a representation of the sequence, other positions in the final output still contain the information of the sequence which can also be used to enhance the representation of the output vector. So this paper designs enhanced layer after BERT to enhance the representation of output. Details of enhanced layer are discussed below.

C. Enhanced Layer

In this paper, we proposed two enhanced layer strategies. Two strategies are used to enhance the classification representation of the original BERT h[C]. Details of the two strategies are described below.

1) Strategy 1(namely, A strategy): In the strategy 1, the enhanced layer uses the information on the position of [C] after the last encoding to calculate the attention of all the position information in this layer sequence. In this case, the output vector of [C] position not only contains the whole sequence information of the upper layer, but also interacts with the information of all positions of this layer.As shown in the upper right of Figure1, the hidden layer representation h[C] and the whole sequence representation Tn are used to calculate the attention equation (1) (2). After using equation (3), the hidden layer representation containing all sequence information is obtained which is written as h[C], and then h[C] is used as the input of the classifier.

$$e = v_a^T \ \tan(W_a \ h_[c] + U_t T_n) \tag{1}$$

$$a = Softmax(e) \tag{2}$$

$$h'_{[c]} = aT_n \tag{3}$$

2) Strategy 2(namely, A-AP strategy): In the strategy 2, all the sequences will be generated. The information is computed with the information at [C] position, and the enhanced representation of all sequences is obtained. Inspired by [19] and [20], simple pooling operation extracts the effective information of the sequence. Therefore, this paper adopts the pooling operation to extract the information, and then output the information after splicing with the original [C]. This paper uses Concat Attention [21] mechanism to make an attention interaction between the representation h[C] of [C] and the representation of all sequences.As shown in the lower right of Figure1, we first exchange the calculation order of hidden layer representation h[C] and the representation of Tn of all sequences, then use the same attention equation as strategy 1 to calculate, and use equation (4) to get the new representation of Tn of all sequences. We use average pooling (5) operation to obtain a new representation of Tavg. The new representation Tavg and original classification representation h[C] are joined together (6) as input of classifier.

$$T'_n = \sum_{i=1}^{m} ah_{[c]} \tag{4}$$

$$T_{avg} = Avg - pooling\,(T'_n) \tag{5}$$

$$h'_{[c]} = Concat(h_{[c]}; T_{avg}) \tag{6}$$

Table I

SUMMARY OF DATASETS. TOTAL SAMPLES, LABEL SETS DENOTE THE TOTAL NUMBER OF SAMPLES AND LABELS, RESPECTIVELY. LABELS/SAMPLE IS THE AVERAGE NUMBER OF LABELS PER SAMPLE.

	Total Samples	Label Sets	Labels/Sample
Obstetrics EMRs	24,339	73	2.49
AAPD	55,840	54	2.41

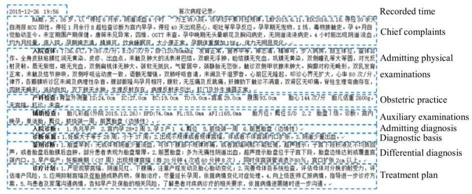

Figure 3. The example of the first course of disease record. Figure adapted from [12].

It should be noted that the classification representation h[C] obtained in Strategy 1 is a d-by-T matrix and T is the size of batch size. In strategy 2, the classification representation h[C] is a 2d-by-T matrix.

D. Output Layer

We optimize the entire model end-to-end, with the additional softmax classifier parameters $W \in R^{k \times H}$ where H is the dimension of the hidden state vectors and K is the number of classes. In this paper, K is also the number of diseases to be classified. The sigmod function is used to classify each label into 0 or 1, where the threshold is set to 0.5. And we minimize the BCEWithLogitsLoss for multi-label tasks.

IV. EXPERIMENT

A. Datasets

In this paper, the experiment was carried out on the Obstetrics EMRs and the open multi-label data set Arxiv Academic Paper Dataset (AAPD) [10]. The scale information of the data set is shown in Table I.

1) AAPD Dataset: Arxiv Academic Paper Dataset (AAPD) is a large multi-label text categorization dataset. Totally, it includes 55,480 abstracts of papers collected from computer science websites. An academic paper may contain a variety of topics. There are 54 topics in this dataset. The task of data sets is to predict the corresponding topics according to the abstracts of the papers.

2) Chinese Obstetric EMRs Dataset: The Chinese obstetric EMRs datasets used in this paper is randomly extracted from 15 hospitals. A total of 24,339 EMRs were pre-processed. The experiment in this paper mainly uses the information of chief complaint, admission examination, obstetric examination and auxiliary examination in the first course record to give a preliminary diagnosis. Therefore, the structure of the first course is mainly analyzed. The first course record sample is shown in Figure3.

Table II
EXPERIMENTAL RESULTS OF OBSTETRICS EMRS. A REPRESENT A STRATEGY, A-AP REPRESENT A-AP STRATEGY.

Model	F1(%)	Average Precision(%)	Hamming Loss	One Error
SGM	60.00	39.00	0.0200	0.0630
BERT	79.58	84.97	0.0132	0.0961
BERT+A	80.26	85.42	0.0129	0.0863
BERT+A-AP	80.28	85.74	0.0129	0.0891

Table III
EXPERIMENTAL RESULTS OF AAPD. A REPRESENT A STRATEGY, A-AP REPRESENT A-AP STRATEGY.

Model	F1(%)	Average Precision(%)	Hamming Loss	One Error
CNN	66.40	-	0.0256	-
CNN-RNN	66.40	-	0.0278	-
SGM	71.00	-	0.0245	-
BERT[24]	73.40	-	-	-
BERT	73.71	79.89	0.0227	0.022
BERT+A	73.81	79.51	0.0225	0.023
BERT+A-AP	74.01	79.74	0.0225	0.024

Table IV
COMPARISON OF OBSTETRICS EMRS OF DIFFERENT POOLING MECHANISMS IN STRATEGY 2(A-AP STRATEGY). A-AP REPRESENT A-AP STRATEGY.

Pooling	F1(%)	Average Precision(%)	Hamming Loss	One Error
Avg-pooling(A-AP)	80.28	85.74	0.0129	0.0891
Max-pooling	80.07	85.69	0.0130	0.0933
(Max+Avg)-pooling	80.16	85.55	0.0130	0.0941

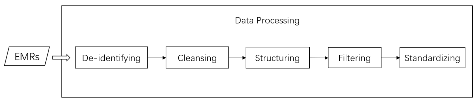

Figure 4. Anonymous identification process

B. Experimental Setup

The collected electronic medical records are pretreated by de-identifying [22], data cleaning, structuring, data filtering and standardization of diagnostic labels. The process is shown in Figure4. Data filtering is the filtering of repeated and similar information contained in the first course of disease, not the decisive factor of filtering diagnosis. It filters by calculating sentence similarity, and only retains personality information. The standardization of diagnostic labels uses different descriptions of the same diagnosis for different doctors, and standardizes the diagnostic results according to ICD10 under the guidance of doctors. In this paper, BERT-base-Chinese is chosen as the language model version of electronic medical record diagnosis in obstetrics and gynecology, and BERT-base is used as the model language model version of AAPD dataset. The parameters are set by default. The main parameters are: hidden_size 768, max_position_embedding 512, num_attention_heads 12, num_hidden_layers 12, maximum input length 512, optimizer Adam [23], learning rate 5e-5, batch_size 2, training epoch 20. We run the rest of the experiments on a GTX 1080 GPU.

C. Results

According to the distribution of diagnostic results, 21,905 of them were used as training set and 2,434 as test set. The experimental results of EMRs diagnosis in obstetrics and gynecology are shown in Table II. Four evaluation indexes [24] (hamming loss, F1-micro, One-error, Average precision) are used to test the effects of BERT and the model with enhanced layer strategy. In the experiment, the Sequence Generation Model (SGM) proposed by [10] is used as the experimental contrast model in this paper. The AAPD dataset used in this experiment is also from SGM paper [10]. BERT represents the result of using the BERT-base-Chinese version alone which is one of the BERT models by [2], and BERT + represents the author's introduction of a mechanism to enhance the original BERT model after the original BERT. The experimental results of AAPD dataset are shown in Table III. The evaluation indicators also use the above four indicators. The results of CNN, CNN-RNN and SGM are from [10], which are shown in the first to third rows of Table III respectively. The result of BERT are from [25], which are shown in the four row of Table III respectively. In this experiment, BERT represents the result of using the BERT-base version alone, and BERT + also represents the result of using the corresponding model layer to enhance BERT.

D. Analysis By Different Pooling Mechanisms in strategy 2

Strategy 1(strategy A) plays an important role in enhancing classification representation through sequence information. The experimental results further prove its effectiveness. But in strategy 2(A-AP strategy), why choose average pooling mechanism instead of other pooling mechanism? This paper then gives a more detailed experimental comparison on Chinese Obstetric EMRs, through the experimental results to analyze the reasons. In Table

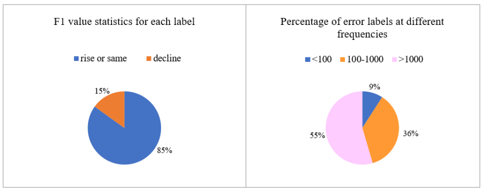

Figure 5. F1 value statistics of each diagnostic label for A strategy. The left chart shows that the F1 value of 85% tags increases or remains unchanged after adding A strategy, while only 15% of tags have a F1 value decline. The right chart shows that among all the wrong labels, labels with frequencies higher than 1000 account for 55% of the total, labels with frequencies between 100 and 1000 account for 36% of the total, while labels with frequencies lower than 100 account for only 9%.

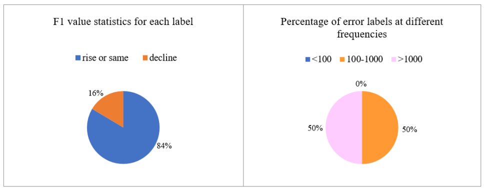

Figure 6. F1 value statistics of each diagnostic label for A-AP strategy. The left chart shows that the F1 value of 84% tags increases or remains unchanged after adding A-AP strategy, while only 16% of tags have a F1 value decline. The right chart shows that among all the wrong labels, labels with frequencies higher than 1000 account for 50% of the total, labels with frequencies between 100 and 1000 account for 50% of the total, while no labels with frequencies lower than 100.

IV, the author compares the pooling mechanism used in strategy 2(A-AP strategy) with other pooling mechanisms on Chinese Obstetric EMRs.

The experimental results show that the average pooling mechanism used in strategy 2(A-AP strategy) is the best based on all the indicators. Besides, the other indicators of maximum pooling are better than the stitching pooling except that the F1 value is lower. This is because the average pooling mechanism is used to get the representation that contains all sequence information. In the maximum pooling mechanism, only the most important representation is obtained. Result from the rich information in gynecological electronic medical records, there are more than one information conducive to diagnosis, and only using the maximum pooling mechanism will cause a certain degree of information loss. In stitching pooling, the use of both pooling methods will get more abundant information theoretically, but due to the existence of maximum pooling, it will still interfere with the final tag prediction to a certain extent. Therefore, the average pooling mechanism is adopted in strategy 2(A-AP strategy) rather than others.

E. Analysis By Different Strategy

In order to analyze the effects of the two enhanced layer strategies proposed in this paper more intuitively, we have made fine-grained statistics on the F1 values of each tag. The experiment found that after adding strategy 1(A strategy), the F1 values of 62 tags in all 73 tags were increased or unchanged, and only 11 Tags showed a decrease. The author further counted the wrong labels, and found that the F1 value of only one label decreased in labels with frequency less than 100. Figure5 is a visual representation of the results. With the addition of strategy 2(A-AP strategy), the F1 value of 34 Tags was increased in all 73 tags, the F1 value of 27 Tags remained unchanged, and the F1 value of only 12 Tags decreased. Among them, none of the labels with frequencies less than 100 has a F1 value decrease. Figure6 is a visual representation of the results. The above analysis shows that the strategy 1(A strategy) and strategy 2(A-AP strategy) in enhanced layer can increase the F1 value of labels by 84% and 85% respectively. According to the frequency of labels appearing, the experimental results show that the two strategies are more effective for low-frequency labels. This also shows that in the case of sufficient data size, BERT itself has been excellent enough, but in the data set with balanced label distribution, the effect of low-frequency label still has great room for improvement.

V. Conclusion

In this experiment, we regard the diagnosis of electronic medical records of gynecology and obstetrics as the task of text multi-label classification.We propose two strategies based on the current advanced language model (BERT) and verify its effectiveness.Of course, in the electronic medical records of obstetrics, besides the existence of information in the form of text, the inspection indicators are also important diagnostic basis.In the next step, the introduction of more domain knowledge and information of inspection indicators deserves attention too.On the other hand, even fine-tuning basic version of BERT-base model will bring a huge amount of computation cost, which limits the deployment and utilization of language models.Therefore, in future work, how to reduce the computation cost of the language models without affecting their performance is equally important.

References

[1] Y. Z. Yang Y.L., “Effect of older pregnancy on maternal and fetal outcomes,” *Chin J Obstet Emerg(Electronic Edition)*, pp. 129–135, 2016.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[3] M.-L. Zhang and Z.-H. Zhou, “Multilabel neural networks with applications to functional genomics and text categorization,” *IEEE transactions on Knowledge and Data Engineering*, vol. 18, no. 10, pp. 1338–1351, 2006.

[4] J. Nam, J. Kim, E. L. Mencía, I. Gurevych, and J. Fürnkranz, “Large-scale multi-label text classificationrevisiting neural networks,” in *Joint european conference on machine learning and knowledge discovery in databases*. Springer, 2014, pp. 437–452.

[5] F. Benites and E. Sapozhnikova, "Haram: a hierarchical aram neural network for large-scale text classification," in *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*. IEEE, 2015, pp. 847–854.

[6] G. Kurata, B. Xiang, and B. Zhou, "Improved neural network-based multi-label classification with better initialization leveraging label co-occurrence," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 521–526.

[7] G. Chen, D. Ye, Z. Xing, J. Chen, and E. Cambria, "Ensemble application of convolutional and recurrent neural networks for multi-label text categorization," in *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, pp. 2377–2383.

[8] S. Baker and A. Korhonen, "Initializing neural networks for hierarchical multi-label text classification," in *BioNLP 2017*, 2017, pp. 307–315.

[9] S. Ma, X. Sun, Y. Wang, and J. Lin, "Bag-of-words as target for neural machine translation," *arXiv preprint arXiv:1805.04871*, 2018.

[10] P. Yang, X. Sun, W. Li, S. Ma, W. Wu, and H. Wang, "Sgm: sequence generation model for multi-label classification," *arXiv preprint arXiv:1806.04822*, 2018.

[11] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[12] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[13] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *arXiv preprint arXiv:1802.05365*, 2018.

[14] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding with unsupervised learning," Technical report, OpenAI, Tech. Rep., 2018.

[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[17] K. Zhang, H. Ma, Y. Zhao, H. Zan, and L. Zhuang, "The comparative experimental study of multilabel classification for diagnosis assistant based on chinese obstetric emrs," *Journal of healthcare engineering*, vol. 2018, 2018.

[18] Y. Z. H. Z. L. Z. Hongchao Ma, Kunli Zhang, "The study of multi-label assistant diagnosis of obstetrics based on feature fusion," *Journal of Chinese Information Processing*, pp. 128–136, 2018.

[19] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," *arXiv preprint arXiv:1607.01759*, 2016.

[20] D. Shen, G. Wang, W. Wang, M. R. Min, Q. Su, Y. Zhang, C. Li, R. Henao, and L. Carin, "Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms," *arXiv preprint arXiv:1805.09843*, 2018.

[21] T. Rocktäschel, E. Grefenstette, K. M. Hermann, T. Kočiskỳ, and P. Blunsom, "Reasoning about entailment with neural attention," *arXiv preprint arXiv:1509.06664*, 2015.

[22] Y.-S. Zhao, K.-L. Zhang, H.-C. Ma, and K. Li, "Leveraging text skeleton for de-identification of electronic medical records," *BMC medical informatics and decision making*, vol. 18, no. 1, p. 18, 2018.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[24] M. L. Zhang and Z. H. Zhou, "Ml-knn: A lazy learning approach to multi-label learning," *Pattern Recognition*, vol. 40, no. 7, pp. 2038–2048, 2007.

[25] A. Adhikari, A. Ram, R. Tang, and J. Lin, "Docbert: Bert for document classification," *arXiv preprint arXiv:1904.08398*, 2019.