The Graph-based Mutual Attentive Network for Automatic Diagnosis

Quan Yuan, **Jun Chen***, **Chao Lu** and **Haifeng Huang**
Baidu Inc, Beijing, China
{yuanquan02, chenjun22, luchao, huanghaifeng}@baidu.com

Abstract

The automatic diagnosis has been suffering from the problem of inadequate reliable corpus to train a trustworthy predictive model. Besides, most of the previous deep learning based diagnosis models adopt the sequence learning techniques (CNN or RNN), which is difficult to extract the complex structural information, e.g. graph structure, between the critical medical entities. In this paper, we propose to build the diagnosis model based on the high-standard EMR documents from real hospitals to improve the accuracy and the credibility of the resulting model. Meanwhile, we introduce the Graph Convolutional Network into the model that alleviates the sparse feature problem and facilitates the extraction of structural information for diagnosis. Moreover, we propose the mutual attentive network to enhance the representation of inputs towards the better model performance. The evaluation conducted on the real EMR documents demonstrates that the proposed model is more accurate compared to the previous sequence learning based diagnosis models. The proposed model has been integrated into the information systems in over hundreds of primary health care facilities in China to assist physicians in the diagnostic process.

Sections	**Texts**
CC	Intermittent shoulder and back pain for 3 years. Feeling short of breath and having syncope for 2 days. (间歇性肩背部疼痛3年, 气短伴晕厥2天)
HPI	The patient suffers from shoulder and back pain without obvious causes, and feels chest tightness and sweating from 3 years ago. (3年前无明显诱因出现肩背部疼痛,伴胸闷,汗出)
PE	T:37.8℃, P:86 BPM, BP:120/80mmHg. (体温:37.8℃, 心率:86 BPM, 血压:120/80mmHg)
SE	Echocardiography: Segmental wall motion abnormality, moderate mitral, tricuspid and aortic insufficiency, moderate pulmonary hypertension, left ventricular dysfunction. (心脏超声:节段性室壁运动异常,二尖瓣,三尖瓣,主动脉瓣中度关闭不全,中度肺动脉高压,左心功能减低)
Diagnosis	Miocardial infarction (心肌梗死)
Findings	shoulder pain (肩痛), back pain (背痛), short of breath (气短), syncope (晕厥), chest tightness (胸闷), moderate aortic valve insufficiency (主动脉瓣中度关闭不全).

Table 1: The example of a real EMR document. **CC**: chief complaint. **HPI**: history of present illness. **PE**: physical examination. **SE**: supplementary examination (e.g. imaging reports or lab test results). Findings are extracted with NER tools from EMR.

1 Introduction

The automatic diagnosis has been very popular in recent years due to the advancement of Artificial Intelligence [Anandan *et al.*, 2019]. The AI-enabled decision-making has been successfully applied in many enterprise diagnostic systems such as Babylon Health and Ping An Good Doctor, to assist physicians and patients through diagnostic process.

Besides the studies of diagnosis on the Web [Xia *et al.*, 2020; Chen *et al.*, 2019], there are extensive studies of automatic diagnosis based on EMR (Electronic Medical Record) documents for clinical use [Yang *et al.*, 2018; Girardi *et al.*, 2018; Mullenbach *et al.*, 2018; Liang *et al.*, 2019]. An EMR document consists of multiple text sections written by certificated physicians with high standards. The text sections describe a patient's illness such as chief complaint, history of present illness, physical examination and so forth. Table 1 shows an example of a real EMR document from a hospital in China. Each section has one or more paragraphs of pure texts where there are critical medical entities like symptoms and signs called *findings* which can be extracted by the Named Entity Recognition (NER) [Dai *et al.*, 2019].

Most of the previous deep learning based diagnosis models consider it as a sequence learning problem, and predict diagnosis with Convolutional Neural Network (CNN) [Girardi *et al.*, 2018; Mullenbach *et al.*, 2018; Yang *et al.*, 2018] or Recurrent Neural Network (RNN) [Sha and Wang, 2017] models. Unfortunately, there still exist three major issues: (1) There are complex relations between findings and diseases (diagnosis), e.g. one disease may cause multiple findings to occur and one finding can be caused by multiple diseases. But the sequence learning models on pure texts can-

[*] Jun Chen is the corresponding author.

not effectively extract the structural information (e.g. graph) between the findings and the diseases in the original texts. (2) The mentions of the same finding in the original texts vary a lot due to different writing styles of the physicians, which leads to the *sparse feature problem* because the feature is diluted into different relations. For example, *hemorrhage of brain stem* and *brainstem hemorrhage* are exactly the same but may be recognized as different findings. Besides, the minor difference in the expression also adds to the sparse feature problem, e.g. formation of softening lesions in the *left* basal ganglia (左基底节软化灶形成) and formation of softening lesions in the *right* basal ganglia (右基底节软化灶形成) are both kinds of formation of softening lesions in the basal ganglia (基底节软化灶形成). (3) The medical entities in the EMR are very important for the physician to make diagnosis, but the previous self-attentive models [Girardi *et al.*, 2018; Sha and Wang, 2017] cannot take advantage of the extra information from other data sources to enhance the importance of the critical words in the input.

To tackle the above issues, we propose the **G**raph-based **M**utual **A**ttentive **N**etwork (**GMAN**) in this paper. It improves the effectiveness of sequence learning models by incorporating the graph convolutional network and the mutual attentive network.

We summarize the major contributions of this paper as:

- We introduce the Graph Convolutional Network (GCN) to faciliate the representation learning of findings and diseases in automatic diagnosis based on the disease hierarchy and the causal graph of diseases we construct. GCN enables the proposed model to effectively extract the structural information between the critical medical entities and alleviate the sparse feature problem.
- We bring forward the novel mutual attentive network between pure texts and medical entities in the diagnosis model. It firstly enhances the representation of medical entities with pure texts, and then enhances the representation of pure texts with the generated features of medical entities. The proposed mutual attentive network emphasizes the critical information in both the original texts and the medical entities towards the better performance of diagnosis.
- The experiments conducted on both the real Chinese EMR documents and the benchmarking English clinical dataset show that the proposed GMAN model outperforms the previous methods in the automatic diagnosis. Besides, we have collaborated with the Regional Healthcare Committee in several major cities in China to integrate GMAN into the information systems in over hundreds of primary health care facilities to assist the physicians throughout the diagnostic process.

2 Related Work

We briefly introduce the related work on deep learning based diagnosis, GCN and attentive networks.

2.1 Deep Learning based Diagnosis

There are many studies of automatic diagnosis based on deep learning. Yang [2018] proposed a multi-layer convolutional network for high level semantic understanding, which is used in automatic diagnosis. Mullenbach [2018] employs a label-wise attentive mechanism, which allows the model to learn distinct document representations for each label. Based on that, Rios [2018] learns to predict the few-shot and the zero-shot labels by matching the discharge summaries in EMR documents to feature vectors for each label by exploiting the structured label space with GCN. RETAIN [Choi *et al.*, 2016b] is an interpretable predictive model, which employed the reverse time attentive mechanism in an RNN for binary prediction. DoctorAI [Choi *et al.*, 2016c] is a straightforward approach with simple RNN for sequential patient data modeling. MNN [Qiao *et al.*, 2019] incorporates clinical text data and medical codes for diagnosis prediction. Meanwhile, Girardi [2019] detects the warning symptoms based on the deep attentive neural network.

2.2 Graph Convolutional Networks

The Graph Convolutional Networks have attracted the growing attention recently and have been widely used in many tasks like recommender system [Zhang *et al.*, 2019], relation extraction [Guo *et al.*, 2019] and reading comprehension [Ding *et al.*, 2019]. The most similar application of GCN to automatic diagnosis is the text classification that models long documents as graphs. Peng [2018] proposes to convert a document into a word co-occurrence graph, which is used as the input to the GCN layers. Yao [2019] models words and documents into a unified graph where the edges between words are computed with the point-wise mutual information (PMI) and the edges connecting words and documents are calculated with TF-IDF features. Liu [2018] proposes a siamese GCN model in the text matching problem by modeling two documents in an interaction graph. Zhou [2018] adopts a similar strategy but uses GCN to match the article with a short query.

2.3 Attentive Networks

Attention is generally used to attend to the most critical part of texts, images or other types of data. It has been successfully applied in machine translation [Vaswani *et al.*, 2017] and question answering [Chen *et al.*, 2019]. In automatic diagnosis, the attentive mechanism is mostly combined with the convolutional networks or recurrent networks to obtain the interpretable prediction results [Sha and Wang, 2017; Girardi *et al.*, 2018; Choi *et al.*, 2016b; Qiao *et al.*, 2019]. Choi [2017] improves disease representation learning by incorporating attention from the preceding nodes on disease hierarchy. The attentive pooling network [Santos *et al.*, 2016] introduces the bidirectional attention in question answering, which is similar to our mutual attentive network. The major difference is that we apply the attention in two consecutive steps rather than the parallel manner in [Santos *et al.*, 2016].

3 The GMAN Model

Fig. 1 shows the architecture of the proposed **G**raph-based **M**utual **A**ttentive **N**etwork (GMAN) model. It consists of three components: medical graph construction, GCN encoding and mutual attentive network.

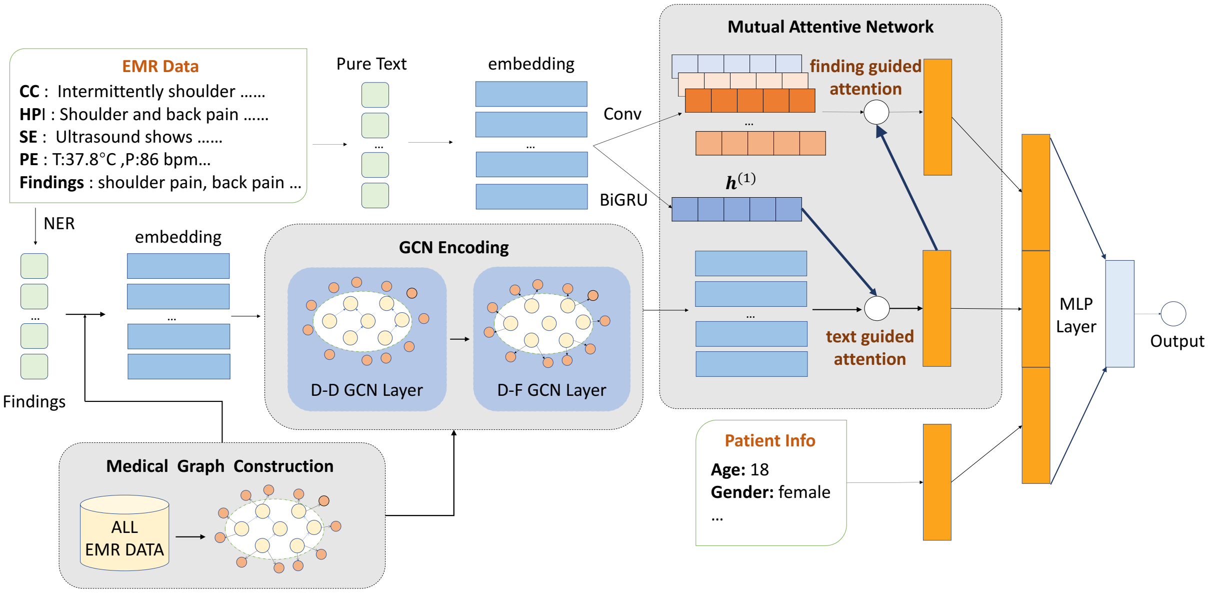


Figure 1: The Architecture of the GMAN model.

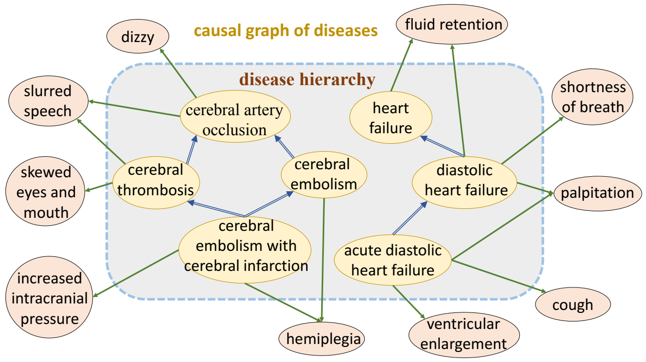


Figure 2: The illustration of the medical graphs. The disease hierarchy surrounded by dashed lines connects between diseases. The causal graph of diseases connects between diseases and findings.

3.1 Medical Graph Construction

As shown in Fig. 2, the GMAN model is built on top of two medical graphs: the disease hierarchy and the causal graph of diseases. The disease hierarchy shows the affiliation between diseases. It consists of disease nodes and the *isA* relationship ($\Rightarrow$) between them. The directed edge $d_i \Rightarrow d_j$ means that disease d_i *is a* kind of disease d_j, e.g. bacterial pneumonia $\Rightarrow$ pneumonia. One disease may have multiple parent diseases or multiple children diseases in the disease hierarchy. We obtain the disease hierarchy from the International Classification of Diseases, 10th Revision (ICD-10) [1].

The causal graph of diseases is mined from the EMR documents, which connects disease d and the findings in the EMR documents where d is the main diagnosis. Let $G = (V, E)$ denote the causal graph of diseases where V and E are the sets of nodes and edges, respectively. V mainly consists of two types of nodes: *finding* and *disease* where *finding* can be the evidence (e.g. *symptom* or *sign*) found in the EMR that supports the diagnosis, and *disease* is the main diagnosis of the EMR. E contains the directed edge $d \rightarrow f$ which means disease d causes finding f to occur. In this study, we propose to construct the causal graph G with the following steps:

Firstly, the EMR document undergoes the named entity recognition (NER) [Dai *et al.*, 2019] to extract the medical entities in the EMR like symptoms, signs and diseases together with the polarity of entities (positive, negative or unknown). The F1 score of the NER is 91% in a separate evaluation conducted on 1000 deduplicated sentences from EMRs annotated by physicians. The original causal graph is obtained by adding a directed edge from the main diagnosis (disease) to each of the positive findings in the same EMR.

Secondly, the original causal graph is pruned based on the causal weight matrix $A \in \mathcal{R}^{|V_d| \times |V_f|}$, which is defined as:

$$A_{i,j} = n(f_j|d_i) * \log \frac{N}{1 + n(d_i)}, \tag{1}$$

where V_d and V_f are the set of disease nodes and that of finding nodes, respectively, and $V_d \cup V_f = V$. $n(f_j|d_i)$ is the frequency of finding f_j in the EMR documents where d_i is the main diagnosis. $n(d_i)$ is the number of EMR documents of disease d_i and N is the total number of EMR documents. $A_{i,j}$ measures the likelihood that disease d_i causes finding f_j. The causal weight is normalized by diseases, and the final causal weight matrix is:

$$\widetilde{A}_{i,j} = \frac{A_{i,j}}{\sum_j A_{i,j}}. \tag{2}$$

The causal weight $\widetilde{A}_{i,j}$ has a long-tail distribution by diseases, and most of the low-weight edges are noise. Thus, we preserve the Top-k edges in each row so that the most important signals on each disease are captured in the result graph. The rest edges are removed and their weights in the graph are set to zero. Therefore, each disease is connected to at most k neighbors in the causal graph. We empirically set $k = 5$ in our experiments.

[1] https://www.cdc.gov/nchs/icd/icd10cm.htm

3.2 GCN Encoding

Most of the previous deep learning based automatic diagnosis models conduct sequence learning on the pure texts of EMR documents [Yang *et al.*, 2018; Mullenbach *et al.*, 2018; Girardi *et al.*, 2018]. Although the sequence learning models can extract important information from the plain texts, yet they are not designed to model the complex inherent relationship between the medical entities, e.g. graph structure. In contrast, the Graph Convolutional Network (GCN) has been widely used in the modeling of graph structure data [Fu *et al.*, 2019]. Thus, we propose to use GCN to obtain the high-level representation of medical entities considering the graph structure among the entities.

Let $\mathbf{d}_i \in \mathcal{R}^m$ and $\mathbf{f}_j \in \mathcal{R}^m$ denote the embedding of disease d_i and that of finding f_j, respectively. Inspired by [Fu *et al.*, 2019], GCN convolves the features of the neighbors to update the embedding of the target node. We propose the update rules as:

$$\hat{\mathbf{d}}_i = ReLU(\mathbf{W}^{(1)}\mathbf{d}_i + \sum_{u\in N_p(i)} \frac{\mathbf{W}^{(2)}\mathbf{d}_u}{|N_p(i)|} + \sum_{v\in N_c(i)} \frac{\mathbf{W}^{(3)}\mathbf{d}_v}{|N_c(i)|} + \mathbf{b}^{(1)}), \quad (3)$$

$$\hat{\mathbf{f}}_j = ReLU(\mathbf{W}^{(4)}\mathbf{f}_j + \frac{1}{|N_g(j)|} \sum_{i\in N_g(j)} \widetilde{A}_{i,j}\mathbf{W}^{(5)}\hat{\mathbf{d}}_i + \mathbf{b}^{(2)}), \quad (4)$$

where $\mathbf{W}^{(1)}$, $\mathbf{W}^{(2)}$, $\mathbf{W}^{(3)}$, $\mathbf{W}^{(4)}$, $\mathbf{W}^{(5)} \in \mathcal{R}^{m\times m}$, $\mathbf{b}^{(1)}$ and $\mathbf{b}^{(2)} \in \mathcal{R}^m$ are the trainable parameters of the GCN encoder. $N_p(i)$ and $N_c(i)$ are the set of parent nodes and that of children nodes of disease d_i on the disease hierarchy, respectively. $N_g(j)$ is the set of neighbors (diseases) of finding f_j on the causal graph. In Eq. (3), we update the embedding of disease d_i based on the features of its parent diseases, children diseases and the feature of itself. In Eq. (4), we update the embeding of finding f_j based on the features of the neighboring disease nodes and the feature of f_j itself. Thus, the structural information of disease hierarchy and causal graph is encoded in the embeddings of diseases and findings with the GCN encoder. Eq. (3) and (4) can be considered as: a *Disease-Disease* (*D–D*) GCN layer followed by a *Disease-Finding* (*D–F*) GCN layer. We update disease embeddings before updating finding embeddings because disease causes findings to occur instead of the other way around. There is an ordinal causal relationship in between. In the GMAN model, the GCN encoder is used to generate the high-level representations of the findings from the given EMR document.

3.3 Mutual Attentive Network

As shown in Fig. 1, the pure texts and the findings extracted from the given EMR are fed into a convolutional neural network with attention. In this study, we propose a mutual attentive network to jointly model pure texts and findings (i.e. medical entities) in automatic diagnosis. Specifically, we use the pure texts to enhance the representation of findings with attention, and vice versa.

Firstly, we concatenate the texts of chief complaint, history of present illness, physical examination and supplementary examination (e.g. lab test result or imaging report) together, and perform Chinese word segmentation with Jieba [2].

Secondly, on one hand, the word sequence is passed through a convolutional layer to get the n-gram features where n is the size of the convolutional kernel, e.g. $n = 3, 4, 5$. On the other hand, the word sequence is fed into a Bidirectional Gated Recurrent Unit (Bi-GRU) layer to get the hidden feature $\mathbf{h}^{(1)}$. Let $\{w_1, ..., w_N\}$ denote the input word sequence. The features of the hidden GRU layer are:

$$\begin{aligned} \overrightarrow{\mathbf{h}^i} &= \overrightarrow{GRU}(w_i) \qquad \overrightarrow{\mathbf{h}^i} \in \mathcal{R}^r, \\ \overleftarrow{\mathbf{h}^i} &= \overleftarrow{GRU}(w_i) \qquad \overleftarrow{\mathbf{h}^i} \in \mathcal{R}^r, \end{aligned} \quad (5)$$

where r denotes the number of recurrent units per direction. We use the step-wise average pooling as the hidden feature:

$$\mathbf{h}^{(1)} = \frac{1}{N}\sum_{i=1}^{N} \begin{bmatrix} \overrightarrow{\mathbf{h}^i} \\ \overleftarrow{\mathbf{h}^i} \end{bmatrix} \qquad \mathbf{h}^{(1)} \in \mathcal{R}^{2r}. \quad (6)$$

$\mathbf{h}^{(1)}$ is used to compute the attention weights α of the embeddings (Eq. (4)) w.r.t each finding:

$$\mathbf{u}_j = tanh(\mathbf{W}^{(6)} \begin{bmatrix} \mathbf{f}_j \\ \mathbf{h}^{(1)} \end{bmatrix} + \mathbf{b}^{(3)}), \quad (7)$$

$$\alpha_j = \frac{exp(\mathbf{v}^{(1)} \cdot \mathbf{u}_j)}{\sum_l exp(\mathbf{v}^{(1)} \cdot \mathbf{u}_l)}, \quad (8)$$

where $\mathbf{W}^{(6)}$, $\mathbf{b}^{(3)}$ and $\mathbf{v}^{(1)}$ are the parameters. Thereafter, we compute the compound representation of all findings as the attentive weighted sum:

$$\mathbf{h}^f = \sum_j \alpha_j \mathbf{u}_j. \quad (9)$$

We call the process (Eq. (5)-(9)) the *text-guided attention*.

Next, we use the findings to enhance the representation of pure texts with attention mechanism because there are key information in the word sequence of pure texts that are not included in the extracted findings due to the missed recall of NER and the other critical but non-medical keywords. For example, the duration of symptoms like "one hour ago" and "for ten years" in the pure texts are critical to tell whether it is acute disease or not.

We use the multi-channel CNN model [Kim, 2014] to process the input pure texts with 3-gram, 4-gram and 5-gram kernels. Each channel has l (e.g. $l = 100$) kernels and the sequence is padded with zeros so that the size of the resulting feature maps is the same. Let $\mathbf{X} \in \mathcal{R}^{N\times k}$ denote the word embeddings of the input pure texts where N is the length of the input word sequence and k is the number of dimensions of word embedding. The output of the convolutional layer is:

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}^{(1)} \\ \mathbf{Y}^{(2)} \\ \mathbf{Y}^{(3)} \end{bmatrix} = MultiChannelCNN(\mathbf{X}), \qquad \mathbf{Y} \in \mathcal{R}^{3N\times l}, \quad (10)$$

[2] Jieba (https://github.com/fxsjy/jieba) is an open-sourced Chinese word segmentation package, widely used in Chinese NLP studies. It generates the DAG structure of all possible segmentations based on Trie Tree, and uses dynamic programming to find the most probable segmentation based on word frequency.

Metrics	**Neurology**	**Cardiology**	**MIMIC-III-50**
# of training samples	20,545	16,130	8,067
# of testing samples	1,425	1,108	1,574
# of unique findings	35,111	23,158	77,949
# of unique diseases	154	110	50
avg. length of text	1,691	2,208	1,530
avg. # of findings per EMR	12	15	25

Table 2: The statistics of the datasets. # means the *the number*.

where $\mathbf{Y}^{(1)}$, $\mathbf{Y}^{(2)}$, $\mathbf{Y}^{(3)} \in \mathcal{R}^{N\times l}$ are the output feature maps of the CNN channels w.r.t. 3-gram, 4-gram and 5-gram, respectively. Due to page limits, we refer the readers to [Kim, 2014] for the detail of multi-channel CNN model.

Similarly, let $\mathbf{Y}_i$ denote the i-th row of $\mathbf{Y}$. We compute the hidden feature of w.r.t. the i-th row as:

$$\mathbf{z}_i = tanh(\mathbf{W}^{(7)} \begin{bmatrix} \mathbf{Y}_i^\top \\ \mathbf{h}^{(f)} \end{bmatrix} + \mathbf{b}^{(4)}), \tag{11}$$

where $\mathbf{h}^{(f)}$ is the *text-guided attention* in Eq. (9). $\mathbf{W}^{(7)}$ and $\mathbf{b}^{(4)}$ are the parameters. The attentive weight of gram is:

$$\alpha_i = \frac{exp(\mathbf{v}^{(2)} \cdot \mathbf{z}_i)}{\sum_k exp(\mathbf{v}^{(2)} \cdot \mathbf{z}_k)}. \tag{12}$$

Finally, the compound representation of all pure texts is computed as the attentive weighted sum:

$$\mathbf{h}^t = \sum_i \alpha_i \mathbf{z}_i. \tag{13}$$

We call Eq. (10)-(13) the *finding-guided attention*.

To this end, the findings and the pure texts are mutually used to enhance the representation of each other. We name this process the *mutual attentive network*. Different from [Santos *et al.*, 2016], the two attentions are performed consecutively in GMAN to introduce more information exchange between free-texts and entities during representation learning. That is, GMAN computes $\mathbf{h}^f$ first, and the computation of $\mathbf{h}^t$ is based on $\mathbf{h}^f$ unlike the parallel attention mechanism used in question answering in [Santos *et al.*, 2016] where the update is separately performed based on the snapshot features of the counterpart's previous state. Finally, the attentive features $\mathbf{h}^f$ and $\mathbf{h}^t$ are concatenated together with the patient's basic information like gender and age before feeding to the last fully connected layer for classification.

4 Experiments

In this section, we evaluate GMAN on the real-world EMR data and compare them with the baseline models.

4.1 Experimental Settings

We have collaborated with many top hospitals in China to conduct research on automatic diagnosis. In this way, we collected the real EMR documents for experiments. We select two medical departments, *Neurology* and *Cardiology*, in the evaluation with the following reasons: (1) It is difficult to distinguish between diseases in the same department since they usually share symptoms and signs. (2) Most of the diseases in Neurology and Cardiology are highly risky, e.g. heart failure and brainstem hemorrhage. It is of highly clinical value to automatically diagnose the diseases in the two departments. The main diagnosis by the certificated doctor in each EMR is selected as its ground-truth label to predict.

For the reproducibility concerns, we choose MIMIC-III-50 [Mullenbach *et al.*, 2018] as the English dataset in the evaluation besides the Chinese datasets. Table 2 shows the statistics of the datasets in the evaluation[3]. The training and the testing sets in the above three datasets are disjoint. For MIMIC-III-50, we use the same training and testing sets from the original study[4]. The public English NER for clinical notes, CliNER[5], is used to process MIMIC-III-50, which reports 83.8% F1 score in the original paper [Boag *et al.*, 2018].

We compare the proposed model with the following deep learning based automatic diagnosis methods:

CNN [Yang *et al.*, 2018]: It proposes a convolutional neural network to extract features from EMR documents for automatic diagnosis.

BiGRU [Choi *et al.*, 2016a]: It introduces the bidirectional GRU upon the word sequence to get the hidden features before aggregating to a compound representation by average pooling for diagnosis prediction.

ACNN [Girardi *et al.*, 2018]: It combines the CNN model with the gram-level attention to predict the diagnosis and detect the warning symptoms.

CAML [Mullenbach *et al.*, 2018]: It proposes a label-wise attention on top of a convolutional neural network to predict diagnosis from clinical texts.

In order to validate the effectiveness of the proposed method in automatic diagnosis, we use **GCN**, **MAN** and **GMAN** to denote the proposed models with GCN encoding alone, mutual attention alone and both, respectively. **GPAP** is the method that replaces the mutual attentive network of **GMAN** with Parallel Attentive Pooling [Santos *et al.*, 2016].

4.2 Results

We use recall at K (R@K) and precision at K (P@K) as the metrics to measure the performance. Since there is only one ground-truth label in each testing sample in the Chinese datasets Neurology and Cardiology, the R@1 and P@1 are reported. However, since there are multiple ground-truth labels for each sample in MIMIC-III-50, we report R@5 and P@5 to be consistent with [Mullenbach *et al.*, 2018]. Since the original paper [Mullenbach *et al.*, 2018] only reports the P@5 result, we reproduce CAML on the same training and testing datasets, and report both P@5 and R@5 in our evaluation. The reproduced P@5 is 61.50%, very similar to that (61.80%) in the original paper.

Table 3 shows the evaluation results. As we can see, the proposed models outperform all baseline models on three datasets under all metrics including the benchmarking English dataset. Compared to the best of the baselines, there are about 6%-10% absolute improvement of recall and precision on Neurology, while there are about 1.6%-4% abso-

[3] We do not have the permission from hospitals to publish the Chinese EMR data since they are legally protected by the laws. Please focus on the contributions of the proposed diagnosis models.

[4] https://github.com/jamesmullenbach/caml-mimic

[5] https://github.com/text-machine-lab/CliNER

Models	Neurology		Cardiology		MIMIC-III-50	
	R@1	P@1	R@1	P@1	R@5	P@5
CNN	62.39%	68.02%	48.37%	48.89%	59.66%	61.67%
BiGRU	64.00%	67.59%	51.32%	60.22%	57.51%	59.10%
ACNN	64.58%	69.31%	52.62%	58.01%	59.77%	61.85%
CAML	62.53%	68.91%	50.75%	56.84%	59.23%	61.50%
GCN	70.31%	77.32%	53.73%	61.19%	61.22%	62.53%
MAN	66.45%	72.36%	53.26%	60.43%	60.05%	62.19%
GPAP	71.87%	78.63%	55.79%	61.94%	61.57%	62.83%
GMAN	**73.59%**	**79.81%**	**57.31%**	**62.62%**	**62.13%**	**63.46%**

Table 3: Prediction results for the different models

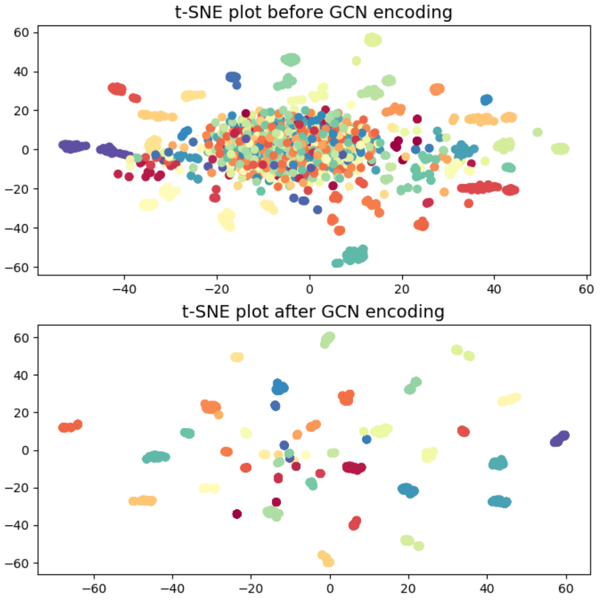


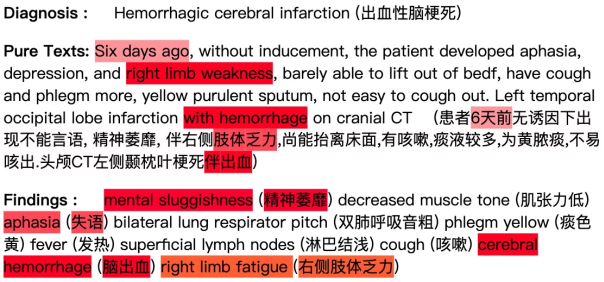
Figure 3: The t-SNE plots of the findings before (upper) and after (bottom) GCN encoding. The findings most close to the same disease are in the same color.

lute improvement of recall and precision on Cardiology and MIMIC-III-50, which prove the effectiveness of the proposed models in automatic diagnosis.

Among the proposed methods, we can see that both GCN and MAN improve the performance of automatic diagnosis. It means that both the proposed GCN encoding and the mutual attentive network bring benefits to model patient's illness. By comparing GMAN and GPAP, we can see that the proposed mutual attentive network outperforms parallel attentive pooling in this task. Meanwhile, the experiments show that when applying both methods together, the performance improvement is the best.

We attribute the improvement of performance of the proposed models to two aspects:

Firstly, the incorporation of GCN encoding can effectively extract the structural information between findings and diseases, and it also alleviates the sparse feature problem. Fig. 3 illustrates the t-SNE plots of the findings before and after GCN encoding. In order to show the validity, we use the same color to draw the findings if they are the most close to the same disease. We randomly select up to 20 findings for each disease from the Cardiology department where the findings are the most close to the selected disease based on the cosine distance between their embeddings. Apparently, the findings w.r.t. the same disease are much closer after GCN encoding, which means GCN is capable of encoding complex structural information in the embeddings.



Diagnosis : Hemorrhagic cerebral infarction (出血性脑梗死)

Pure Texts: Six days ago, without inducement, the patient developed aphasia, depression, and right limb weakness, barely able to lift out of bedf, have cough and phlegm more, yellow purulent sputum, not easy to cough out. Left temporal occipital lobe infarction with hemorrhage on cranial CT （患者6天前无诱因下出现不能言语, 精神萎靡, 伴右侧肢体乏力,尚能抬离床面,有咳嗽,痰液较多,为黄脓痰,不易咳出.头颅CT左侧颞枕叶梗死伴出血)

Findings : mental sluggishness (精神萎靡) decreased muscle tone (肌张力低) aphasia (失语) bilateral lung respirator pitch (双肺呼吸音粗) phlegm yellow (痰色黄) fever (发热) superficial lymph nodes (淋巴结浅) cough (咳嗽) cerebral hemorrhage (脑出血) right limb fatigue (右侧肢体乏力)

Figure 4: Visualization of mutual attention weights

Secondly, the proposed mutual attentive network can correctly enhance the weights of both the critical findings and the important original words from the input towards getting more accurate diagnosis. Fig. 4 illustrates an example of the mutual attention weights. The example is randomly selected from the testing set. The proposed mutual attentive network learns to impose higher weights upon the critical findings as well as the important words. In Fig. 4, the words with high attention weights are highlighted. The higher the weight is, the darker the background color is. As we can see, the highlighted findings (e.g. *mental sluggishness*, *decreased muscle tone*, *cerebral hemorrhage* and *limb fatigue*) are all the critical symptoms and signs of the diagnosis *hemorrhagic cerebral infarction*. Similarly, the highlighted words (e.g. *right limb weakness* and *with hemorrhage*) in the original texts are also the keywords highly relevant to the diagnosis. Besides, the attention weights can be further used to *explain* the diagnosis prediction because they are considered critical to make the diagnosis by the model. Thus, the mutual attentive network sheds some light on the interpretability of the automatic diagnosis models.

In all, the evaluation results prove that the proposed models are effective in automatic diagnosis on both the Chinese and the English datasets. The improvement results from the better representation learning of medical entities by GCN and the interpretability brought by mutual attentive network.

5 Conclusion

In this paper, we propose GMAN, an automatic diagnosis model which is built upon the reliable corpus of EMR documents from hospitals. GMAN consists of medical graph construction, GCN encoding and mutual attentive network. We construct the disease hierarchy and the causal graph of diseases based on the international standards as well as the clinical EMR data. The proposed GCN encoding extracts the complex structural information between the critical findings found on the patient. The mutual attentive network enhances the feature representations of the critical words and findings by imposing higher attention weights on them. The experimental results show that the proposed models outperform the previous deep learning based diagnosis methods on the real-world clincial EMR data.

References

[Anandan *et al.*, 2019] Padmanabhan Anandan, Yan Huang, Kazumi Nishikawa, BBorie Park, Eric S. Sullivan, Jingyu Wang, and Xu Shan. AI in health care: Capacity, capability, and a future of active health in Asia. *MIT Technology Review Insights*, pages 1–25, October 2019.

[Boag *et al.*, 2018] Willie Boag, Elena Sergeeva, Saurabh Kulshreshtha, Peter Szolovits, Anna Rumshisky, and Tristan Naumann. CliNER 2.0: Accessible and accurate clinical concept extraction. In *arXiv:1803.02245*, 2018.

[Chen *et al.*, 2019] Jun Chen, Jingbo Zhou, Zhenhui Shi, Bin Fan, and Chengliang Luo. Knowledge abstraction matching for medical question answering. In *IEEE BIBM*, pages 342–347, 2019.

[Choi *et al.*, 2016a] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor AI: Predicting clinical events via recurrent neural networks. In *MLHC*, pages 301–318, 2016.

[Choi *et al.*, 2016b] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. In *NeurIPS*, pages 3504–3512, 2016.

[Choi *et al.*, 2016c] Edward Choi, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Using recurrent neural network models for early detection of heart failure onset. *JAMIA*, 24(2):361–370, 2016.

[Choi *et al.*, 2017] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F. Stewart, and Jimeng Sun. GRAM: Graph-based attention model for healthcare representation learning. In *KDD*, pages 787–795, 2017.

[Dai *et al.*, 2019] Dai Dai, Xinyan Xiao, Yajuan Lyu, Shan Dou, Qiaoqiao She, and Haifeng Wang. Joint extraction of entities and overlapping relations using position-attentive sequence labeling. In *AAAI*, 2019.

[Ding *et al.*, 2019] Ming Ding, Chang Zhou, Qibin Chen, Hongxia Yang, and Jie Tang. Cognitive graph for multihop reading comprehension at scale. In *ACL*, pages 2694–2703, 2019.

[Fu *et al.*, 2019] Tsu-Jui Fu, Peng-Hsuan Li, and Wei-Yun Ma. GraphRel: Modeling text as relational graphs for joint entity and relation extraction. In *ACL*, pages 1409–1418, 2019.

[Girardi *et al.*, 2018] Ivan Girardi, Pengfei Ji, An phi Nguyen, Nora Hollenstein, Adam Ivankay, Lorenz Kuhn, Chiara Marchiori, and Ce Zhang. Patient risk assessment and warning symptom detection using deep attention-based neural networks. In *EMNLP Workshop*, pages 139–148, 2018.

[Guo *et al.*, 2019] Zhijiang Guo, Yan Zhang, and Wei Lu. Attention guided graph convolutional networks for relation extraction. In *ACL*, pages 241–251, 2019.

[Kim, 2014] Yoon Kim. Convolutional neural networks for sentence classification. In *EMNLP*, pages 1746—-1751, 2014.

[Liang *et al.*, 2019] Huiying Liang, Brian Y. Tsui, Hao Ni, Carolina C. S. Valentim, Sally L. Baxter, and et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nature Medicine*, 25:433–438, 2019.

[Liu *et al.*, 2018] Bang Liu, Ting Zhang, Di Niu, Jinghong Lin, Kunfeng Lai, and Yu Xu. Matching long text documents via graph convolutional networks. *arXiv preprint arXiv:1802.07459*, 2018.

[Mullenbach *et al.*, 2018] James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. Explainable prediction of medical codes from clinical text. In *NAACL*, pages 1101–1111, 2018.

[Peng *et al.*, 2018] Hao Peng, Jianxin Li, Yu He, Yaopeng Liu, Mengjiao Bao, Lihong Wang, Yangqiu Song, and Qiang Yang. Large-scale hierarchical text classification with recursively regularized deep Graph-CNN. In *WWW*, pages 1063–1072, 2018.

[Qiao *et al.*, 2019] Zhi Qiao, Xian Wu, Shen Ge, and Wei Fan. MNN: Multimodal attentional neural networks for diagnosis prediction. In *IJCAI*, pages 5937–5943, 2019.

[Rios and Kavuluru, 2018] Anthony Rios and Ramakanth Kavuluru. Few-shot and zero-shot multi-label learning for structured label spaces. In *EMNLP*, pages 3132–3142, 2018.

[Santos *et al.*, 2016] Cicero dos Santos, Ming Tan, Bing Xiang, and Bowen Zhou. Attentive pooling networks. *arXiv preprint arXiv:1602.03609*, 2016.

[Sha and Wang, 2017] Ying Sha and May D. Wang. Interpretable predictions of clinical outcomes with an attention-based recurrent neural network. In *ACM BCB*, pages 233–240, 2017.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.

[Xia *et al.*, 2020] Yuan Xia, Jingbo Zhou, Zhenhui Shi, Chao Lu, and Haifeng Huang. Generative adversarial regularized mutual information policy gradient framework for automatic diagnosis. In *AAAI*, 2020.

[Yang *et al.*, 2018] Zhongliang Yang, Yongfeng Huang, Yiran Jiang, Yuxi Sun, Yu-Jin Zhang, and Pengcheng Luo. Clinical assistant diagnosis for electronic medical record based on convolutional neural network. *Scientific Reports*, 8(1), April 2018.

[Yao *et al.*, 2019] Liang Yao, Chengsheng Mao, and Yuan Luo. Graph convolutional networks for text classification. In *AAAI*, pages 7370–7377, 2019.

[Zhang *et al.*, 2019] Jiani Zhang, Xingjian Shi, Shenglin Zhao, and Irwin King. STAR-GCN: Stacked and reconstructed graph convolutional networks for recommender systems. In *IJCAI*, pages 4264–4270, 2019.

[Zhou *et al.*, 2018] Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, and Maosong Sun. Graph neural networks: A review of methods and applications. *arXiv preprint arXiv:1812.08434*, 2018.