Integrating Empirical Knowledge into Multi-View Feature Attention Network for Disease Diagnosis

Anonymous ACL submission

Abstract

As one of the currently significant problems in AI-enabled healthcare research, disease diagnosis based on the medical text has made substantial progress. However, the length of the diagnostic evidences is different, leading to the difficulty of capturing multi-scale features of each disease. And recent studies have discovered that structural knowledge from medical text is critical for disease diagnosis. This paper proposes integrating empirical knowledge of disease into a multi-view feature attention network to address these issues. The multi-view feature attention network employs multi encoders to capture segment information of diagnostic evidences of each illness. Meanwhile, we used an abductive causal graph constructed from medical text to extract the empirical knowledge representation of diseases by graph convolutional network. The evaluation conducted on the MIMIC-III-50 dataset and Chinese dataset demonstrates that the proposed method outperforms the structural knowledge-based state-of-the-art models.[1]

EMR Text:
Cough and expectoration for 1 month, worsening with chest tightness for 10 days. The patient had a cough with no obvious cause before 1 month, showing paroxysmal cough, coughing a small amount of white foamy sputum, no fever, no chills, no chills, chest tightness when coughing, no shortness of breath, chest pain, and no night sweats, with fatigue and appetite. Check the sputum to find AFB: acid-fast bacilli (+++), given infusion and oral medication. After the treatment, the patient's cough improved, but the patient's chest tightness increased in the past 10 days. The patient's chest CT examination in our hospital indicated: left lower lobe lesions, consideration of left lung Ca and obstructive atelectasis.
...
MRI of the lumbar spine showed that the disc degeneration of the lumbar 3-sacral segment 1, mild bulging of the lumbar 4-5 intervertebral disc, and bulging of the lumbar 5-sacral 1 interve rtebral disc. Please consider after the orthopedic surgery consultation: give "Celecoxib Capsule" 0.2g po qd, and externally apply Huoxuezhitong ointment.

Admission diagnosis:
Tuberculosis; Lumbar disc herniation

Figure 1: Each admission diagnosis corresponds to diagnostic evidences of different lengths. The segments highlighted are the diagnostic evidence of tuberculosis. The lumbar disc herniation is like it.

1 Introduction

With the rapid growth of the population, some common diseases occupy a large amount of public medical resources, which leads to the problem of uneven distribution (Bohmer et al., 2020). And due to the enormous work pressure, the misdiagnosis rate of doctors will also increase. Meanwhile, AI technology has been applied in many fields, such as face recognition (Adjabi et al., 2020; Wang et al., 2020b), machine translation (Bapna and Firat, 2019; Fan et al., 2021), etc. Therefore, it is particularly significant to employ AI technology to establish an auxiliary diagnosis system, which can improve the work efficiency of doctors, reduce the misdiagnosis rate, and alleviate the problem of lack of medical resources.

Most methods are based on the Electronic Medical Record (EMR) text for disease diagnosis, mainly including the chief complaint, history of present illness, past history, and test results information. Some existing methods treat it as a multi-label text classification task, such as CNN-based (Mullenbach et al., 2018; Li and Yu, 2020; Liu et al., 2021), RNN-based (Cho et al., 2014; Vu et al., 2020). These methods employ a sequence model and attention mechanism, which mainly focus on the information representation of the entire medical text. Since medical texts often contain professional knowledge and terminology, some studies incorporate additional medical knowledge into diagnostic models, e.g., the description of diseases (Xie et al., 2019; Wang et al., 2020a). Besides, the entity-level features and their relationships are also essential for disease diagnosis. Since GCN (Kipf and Welling, 2017) was proposed, some studies have tried to leverage structural knowledge graphs to diagnose disease, such as (Yuan et al., 2020; Xie et al., 2020; Chen et al., 2020; Sun et al., 2020; Chen et al., 2021a). Moreover, doctors will accumulate abundant empirical knowledge in clinical practice, which will assist them in diagnosing diseases more accurately. Therefore, empirical knowl-

[1]The code is available at https://github.com/FutureForMe/MVFAN-EK

edge of diseases is also essential (Quaranta, 2021).

Although the current methods have made significant progress in disease diagnosis, there are still the following challenges: 1) Since there are many types of diseases, and each disease corresponds to diagnostic evidence segments of different lengths, as shown in Figure 1. The first challenge is how to accurately extract the diagnostic evidence of each disease from the segment information. 2) The empirical knowledge that doctors gain from clinical experience is also essential in disease diagnosis. Therefore, the other is how to extract the empirical knowledge of diseases from medical texts reasonably and effectively.

To address these challenges, we propose integrating **E**mpirical **K**nowledge into the **M**ulti-**V**iew **F**eature **A**ttention **N**etwork (**MVFAN-EK**) model, which employs multiple CNNs combined with the label attention mechanism, which can extract diagnostic evidence segment information of each disease from the long medical text. Besides, we also propose a framework for knowledge fusion on the abductive causal graph to obtain empirical knowledge of diseases.

The main contributions of this paper are as follows:

- We propose a multi-view feature attention module that captures disease diagnosis segment information of different lengths corresponding to each disease.
- We first put forward an abductive causal graph constructed from electronic medical records. Through GCN fusion, we can obtain disease representations that incorporate empirical knowledge.
- The experiment results conducted on the real medical dataset demonstrate that our proposed method outperforms previous state-of-the-art methods, which validates the effectiveness of our proposed method.

2 Related Work

In this section, we will briefly introduce disease diagnosis models based on text classification and structural knowledge, and finally, discuss the importance of empirical knowledge of diseases.

Based on Text Classification Disease diagnosis has been a hot topic in the healthcare domain for more than 20 years (de Lima et al., 1998). Recent works utilized sequence models (Kim, 2014; Cho et al., 2014) and attention mechanisms for disease diagnosis. Some researchers employed CNN to extract n-gram features from the medical text (Yang et al., 2018; Mullenbach et al., 2018; Li and Yu, 2020; Liu et al., 2021). In addition, there is a growing interest in using RNN to capture long-range dependent information (Shi et al., 2017; Vu et al., 2020). Different from previous work, our work improves model performance by extracting segments of diagnostic evidence at different scales for each disease.

Based on Structural Knowledge Since GCN (Kipf and Welling, 2017) was proposed, it has attracted the attention of many researchers. To capture structural knowledge, some researchers have begun to construct knowledge graph from medical text to diagnose diseases (Xie et al., 2019; Cao et al., 2020; Yuan et al., 2020). As the best model at present,SHiDAN (Chen et al., 2021a) incorporates a subgraph convolutional network and hierarchical diagnostic attentive network to extract the layered structural features. The difference of our proposed method to SHiDAN (Chen et al., 2021a) is that our disease expression is empirical knowledge extracted from clinical experience, which can be applied to all patients rather than personalized.

Empirical Knowledge of Disease In clinical practice, the empirical knowledge of diseases can help doctors diagnose diseases and reduce the rate of misdiagnosis. (Quaranta, 2021) emphasized the importance of empirical knowledge in clinical practice in the present society. (Joto et al., 2021) constructed a knowledge base by the clinical empirical knowledge of neurosurgery to assist in disease diagnosis.

3 Method

In this section, we first describe our MVFAN-EK model (Figure 2), which consists of two major modules: multi-view feature attention module that employs multiple encoders to produce the representation of each disease from different perspectives, and empirical knowledge representation module that uses GCN to obtain the empirical knowledge of each disease on an abductive causal graph.

3.1 Problem Definition

We treat disease diagnosis as a multi-label classification task under the medical text. The input of the model is the EMRs text data W =

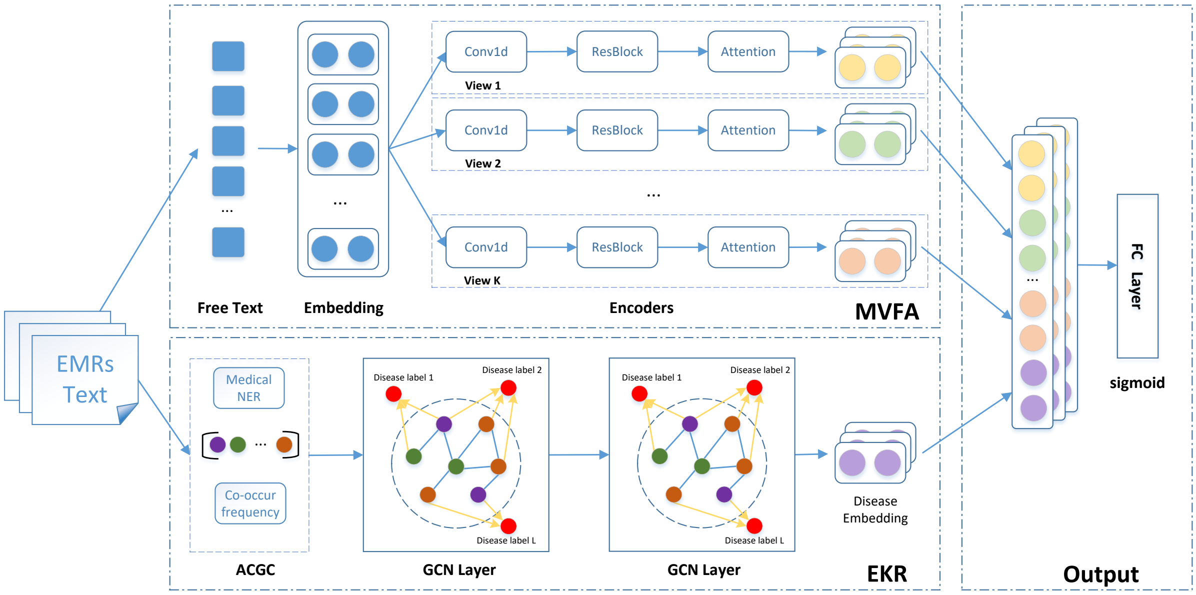


Figure 2: Architecture of our MVFAN-EK model which contains two main modules, MVFA and EKR. The MVFA module extracts the diagnostic evidence segment information for each disease from K views. The EKR module extracts the empirical knowledge of diseases by graph convolutional network in an abductive causal graph.

$[w_1, w_2, ..., w_N]$, where N denotes the length of medical tokens. The output is the prediction result $\hat{\mathbf{y}} = [y_1, y_2, ..., y_L]$, where L is the number of disease and $y_i \in \{0, 1\}$.

3.2 Multi-View Feature Attention (MVFA)

To capture multi-scale diagnostic evidence information of each disease, we define multi encoders including convolutional layer, residual block, and label attention. Each encoder can extract the segment information from one view for each disease.

Embedding Layer

We employ a tokenizer to obtain the word tokens for the input medical text data W, such as EMR. Then by the pre-trained model, like Word2Vec (Mikolov et al., 2013) and Bert (Devlin et al., 2019), we can acquire the word embedding $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N}]$, where $\mathbf{x_i} \in R^{d_w}$, d_w is the dimension of word embedding.

Multi View Convolutional

We apply multiple CNNs of different scales to extract the diagnostic evidence segment from different views. For example, a CNN with a convolution kernel size of 3 and a convolution kernel channel of 100 can capture a segment pattern of length 3. The convolutional procedure can be formalized as :

$$\mathbf{H_c} = BatchNorm(tanh(Conv1d(\mathbf{X}))) \quad (1)$$

where $Conv1d$ represents the 1-dimensional convolution. Here we forced the row number N of the output $\mathbf{H_c} \in R^{N \times d_f}$ to be same as that input $\mathbf{X}$. d_f indicates the out-channel size of the filter.

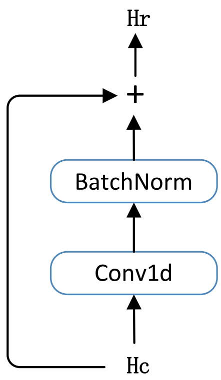


Figure 3: The architecture of a residual block. "+" represents the element-wise addition.

Residual Block

In order to reduce the gradient vanishing issue, we add a residual network (He et al., 2016) after the convolutional layer. The structure of the residual block in our model is shown in Figure 3. The output of multi-view convolutional layer $\mathbf{H_c}$ is input the residual block as:

$$\mathbf{H_r} = \mathbf{H_c} + BatchNorm(Conv1d(\mathbf{H_c})) \quad (2)$$

where $\mathbf{H_r} \in R^{N \times d_f}$ and the $Conv1d$ is the same as before.

Label Attention

To capture each label representation from different views, we employ a label attention mechanism (Lin et al., 2017) to transform $\mathbf{H_c}$ into label-specific vectors. First, we compute the label-specific weight as:

$$\mathbf{Z} = tanh(\mathbf{W}\mathbf{H_r}^\top) \tag{3}$$

$$\mathbf{A} = softmax(\mathbf{U}\mathbf{Z}) \tag{4}$$

Eq 3 is a non-liner projection, where $\mathbf{W} \in R^{d_p \times d_f}$ is a matrix. Then we use a matrix $\mathbf{U} \in R^{|L| \times d_p}$ to compute the label-specific weight matrix $\mathbf{A} \in R^{|L| \times N}$. The attention weight matrix $\mathbf{A}$ is used to produce the label-specific vectors as:

$$\mathbf{V} = \mathbf{H_r}\mathbf{A}^\top \tag{5}$$

Finally, the matrix $\mathbf{V} \in R^{d_f \times |L|}$ is the representation of diseases from a view.

Supposing that we set K encoders in this module, which can obtain the representations of each disease $\mathbf{V_1}, \mathbf{V_2}, ..., \mathbf{V_K}$ from K views.

3.3 Empirical Knowledge Representation (EKR)

To simulate the empirical knowledge representation of diseases obtained from clinical experience, we construct an abductive causal graph from all EMRs, and then use GCN for knowledge fusion to get the empirical knowledge representation of each disease.

Abductive Causal Graph Construction (ACGC)

The first step in constructing an abductive causal graph is Named Entity Recognition (NER) (Li et al., 2020; Chen et al., 2021b). As a sub-direction of NER, there are many mature models of medical NER (Wu et al., 2017; Wang et al., 2019), which can extract medical entities such as symptoms, test results, etc., from medical texts. We use the existing medical NER model to obtain the entity set $E = \{e_1, e_3, ..., e_M\}$, where M is the number of entities.

The abductive causal graph can be constructed using co-occurrence frequencies between entities, similar to the previous work (Chen et al., 2021a; Yuan et al., 2020). Firstly, we construct the co-occurrence relationship between symptoms, test results, and other entities by setting a threshold (e.g., 30). Secondly, we construct the reverse causal relationship between symptoms, test results, and labeled diseases. The construction process is similar to the former, but the relationship is directed from symptoms and test results to labeled diseases. This kind of directed edge can represent the process of inferring the disease from the symptoms and the test results. Through the above two steps, the final abductive causal graph $G = (E, R)$ can be constructed.

GCN Layer

The GCN (Kipf and Welling, 2017) has been widely used in the modeling of graph structure data. But the original GCN was designed for undirected graphs. For propagating the information of a node to its nearest neighbors on the directed graph, (Fu et al., 2019; Bian et al., 2020) improved the original GCN. Therefore, we use improved GCN to obtain the high-level representation of medical entities considering the graph structure among the entities.

After getting the abductive causal graph from all EMRs, we employ an embedding layer to produce their initial vector representation $\mathbf{H_g^{(0)}} \in R^{M \times d_e}$ of medical entities. The disease entities can fuse the information of their nearest medical entities through GCN. Empirical knowledge of diseases can be represented through multi-layer GCN fusion as:

$$\mathbf{H_g^{(l+1)}} = \sigma(\mathbf{\hat{D}^{-1}}\mathbf{\hat{A}}\mathbf{H_g^{(l)}}\mathbf{W^{(l)}}) \tag{6}$$

where $\mathbf{W^{(l)}}$ is a weight matrix for the $\mathbf{l}$-th neural network layer and $\sigma(\cdot)$ is a non-linear activate function like $ReLU$. With $\mathbf{\hat{A}} = \mathbf{A} + \mathbf{I}$, where $\mathbf{I}$ is the identity matrix and $\mathbf{\hat{D}}$ is the diagonal node degree matrix of $\mathbf{\hat{A}}$.

Provided that $\mathbf{H_g} \in R^{M \times d_e}$ is the output of last GCN layer, we extract the vector representation of each disease entity $\mathbf{H_d} \in R^{|L| \times d_e}$ from it.

3.4 Output Layer

We concat the empirical knowledge representation of diseases $\mathbf{H_d}$ with the patient feature representation from multiple view attention $\mathbf{V_1}, ..., \mathbf{V_M}$, and then input it into the fully connected layer (FC).

$$\mathbf{H} = concat[\mathbf{V_1}^\top, ..., \mathbf{V_K}^\top, \mathbf{H_d}] \tag{7}$$

$$\mathbf{\hat{y}} = sigmoid(\mathbf{W}\mathbf{H} + \mathbf{b}) \tag{8}$$

where $\mathbf{H} \in R^{|L| \times (K \times d_f + d_e)}$. The probability of disease $\mathbf{\hat{y}}$ can be predicted by the $sigmoid$ activation function. Here, the probability is used to

# views	kernel size	MIMIC-III-50		ChineseEMR	
		macro F1	micro F1	macro F1	micro F1
1	(3,)	63.8%	69.0%	71.0%	74.7%
	(5,)	65.6%	70.1%	72.0%	75.7%
2	(3,5)	65.7%	70.2%	72.7%	76.6%
	(3,9)	65.8%	70.3%	70.5%	74.3%
3	(3,5,9)	66.1%	70.5%	76.2%	79.2%
	(3,5,15)	65.2%	69.9%	72.4%	75.6%
4	(3,5,9,15)	66.0%	70.0%	75.4%	77.8%
	(3,5,9,19)	65.3%	69.9%	74.5%	77.9%
5	(3,5,9,15,19)	64.8%	69.4%	73.8%	76.1%

Table 1: Performance comparisons using different configurations on MIMIC-III-50 and ChineseEMR datasets.

predict the binary output $y_i \in \{0, 1\}$ using a predefined threshold, such as 0.5. The training objective is to minimize the binary cross-entropy loss between the prediction $\hat{\mathbf{y}}$ and the target $\mathbf{y}$ as:

$$L(W, G, \mathbf{y}, \theta) = - \sum_{j=1}^{L} y_j log\hat{y}_j + (1 - y_j) log(1 - \hat{y}_j) \quad (9)$$

where θ denotes all the trainable parameters, W denotes the input word sequence and G is the abductive causal graph.

4 Experiment

4.1 Datasets

In order to make the results more convincing and robust, we conducted experiments on the public English dataset (MIMIC-III-50) and the Chinese dataset (ChineseEMR) respectively.

MIMIC-III-50

Similar to the previous work (Xie et al., 2019; Li and Yu, 2020), we focus on the prediction of the final diagnosis based on the discharge summary of the patient. In particular, we did not use all diseases but selected the top 50 diseases with the highest frequency for experiments, including 8,067 discharged summaries for training, 1,574 for validation, and 1,730 for testing.

ChineseEMR

This dataset is the real EMR data of a tertiary A hospital in China, involving 38 common diseases and 4,864 EMRs, including 3,392 EMRs are used for training, 729 for validation, and 743 for testing.Each EMR contains the chief complaint, current medical history, past history, and related test results. More details of the datasets are in table 2.

Metrics	MIMIC-III-50	ChinsesEMR
# of EMRs	11,368	4,864
# of diagnosis codes	50	38
avg # of diagnosis per EMR	5.8	1
avg # of entities per EMR	81.4	52.9
avg length of EMRs	1,876	740

Table 2: The statistical results of the MIMIC-III–50 and ChineseEMR datasets. The " # " means number.

Preprocessing

For the MIMIC-III-50 dataset, we follow the previous work (Li and Yu, 2020). Due to the biomedical and clinical English model packages of Stanford (Zhang et al., 2021) having achieved the best results on the English open-source dataset 2010 i2b2/VA dataset (Uzuner et al., 2011), we use it to obtain the medical entities in the MIMIC-III-50 dataset. For the ChineseEMR dataset, we tokenize the text based on character. Since Bert has shown excellent performance in many natural language processing fields, we employ Bert as a pre-training model to acquire the initial word embedding of the Chinese dataset. The length of EMRs is from 182 to 1,569. Therefore, we truncate all EMRs to the maximum length of 1,024. Besides, as the classic NER model, we trained the Bi-LSTM-CRF (Lample et al., 2016) model in real EMRs manually marked by medical experts, whose F1 score is reported about 95.07% in the validation set.

4.2 Evaluation Metrics

In order to ensure the fairness of the model in comparison, the same evaluation metrics as the previous work (Li and Yu, 2020), macro-F1, macro-AUC, micro-F1, micro-AUC, and P@5 are applied in the MIMIC-III-50 dataset. In the ChineseEMR dataset, since only a few EMRs have multiple diagnostic re-

Model	Macro		Micro		P@5
	F1	AUC	F1	AUC	
Bi-GRU (Cho et al., 2014)	48.4%	82.8%	54.9%	86.8%	59.1%
CNN (Kim, 2014)	57.6%	87.6%	62.5%	90.7%	62.0%
CAML (Mullenbach et al., 2018)	53.2%	87.5%	61.4%	90.9%	60.9%
DR-CAML (Mullenbach et al., 2018)	57.6%	88.4%	63.3%	91.6%	61.8%
HyperCore (Cao et al., 2020)	60.9%	89.5%	66.3%	92.9%	63.2%
MultiResCNN (Li and Yu, 2020)	60.6%	89.9%	67.0%	92.8%	64.1%
MSATT-KG (Xie et al., 2019)	63.8%	**91.4%**	68.4%	**93.6%**	**64.4%**
GMAN (Yuan et al., 2020)	62.4%	–	66.0%	–	–
SHiDAN (Chen et al., 2021a)	**64.7%**	–	**69.2%**	–	–
MVFAN-EK (Our Model)	**66.1%**	**92.0%**	**70.5%**	**94.1%**	**65.9%**

Table 3: The experiment results on the MIMIC-III-50. Since our model and the baseline models use the same dataset and evaluation metrics, the results of baselines are directly cited from the origin papers.

model	Macro		Micro		P@1	R@1
	F1	AUC	F1	AUC		
Bi-GRU (Cho et al., 2014)	63.0%	**96.1%**	66.0%	**97.0%**	71.6%	65.1%
CNN (Kim, 2014)	66.8%	92.7%	68.2%	94.1%	65.2%	71.4%
CAML (Mullenbach et al., 2018)	69.7%	94.9%	71.1%	95.9%	74.8%	74.7%
MultiResCNN (Li and Yu, 2020)	**70.3%**	93.8%	**72.7%**	94.3%	**76.6%**	**76.5%**
MVFAN-EK (Our Model)	**76.2%**	**97.0%**	**79.2%**	**97.5%**	**79.8%**	**79.6%**

Table 4: The experiment results on the ChineseEMR. For the Chinese dataset, we select some baselines with source code for comparison.

sults, except the previous evaluation metrics macro-F1, macro-AUC, micro-F1, and micro-AUC, we have added P@1 and R@1 as evaluation metrics. As detailed in (Schütze et al., 2008).

4.3 Experiment Implementation

We implement out MVFAN-EK model using PyTorch (Paszke et al., 2019). During training model, we apply AdamW (Loshchilov and Hutter, 2019) as optimizer, and set its learning rate to 0.001. The number of epochs and batch size are set to 200 and 8. If there is no improvement of the micro-F1 score on the validation dataset in 10 continuous epochs, we will stop early. In addition, we also implement a dropout mechanism with dropout probability of 0.3. Note that to ensure the accuracy of the experiment, we ran our model three times with the same hyper-parameters using different random seeds and reported the scores averaged over three times.

To explore a better configuration for the number of views and the kernel sizes of each encoder, we follow the previous work (Li and Yu, 2020) to design some experiments. The experimental results are shown in Table 1. We choose the best configuration for experimentation, which is 3 encoders, and the kernel size is (3,5,9). More parameter sensitivity analysis in section 5.3.

4.4 Baselines

We compared the following deep learning models for disease diagnosis, which include text classification-based and structural knowledge-based models:

The baselines of text classification-based include CNN (Kim, 2014), Bi-GRU (Cho et al., 2014), CAML [CNN with label-wise attention] (Mullenbach et al., 2018), DR-CAML [CAML added text description] (Mullenbach et al., 2018) and MultiResCNN [Multi-filter CNN with ResNet] (Li and Yu, 2020). The baselines of structural knowledge-based include HyperCore [GCN with hyperbolic representation of disease] (Cao et al., 2020), MSATT-KG [Multi-scale CNN with attention integrated into the structural knowledge of disease] (Xie et al., 2019), GMAN [GCN with mutual attention] (Yuan et al., 2020) and current state-of-the-art model SHiDAN [subgraph convolutional network with hierarchical attentive network] (Chen et al., 2021a).

5 Results and Analysis

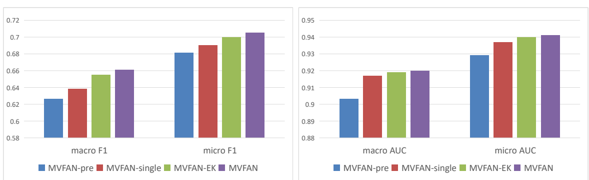
5.1 Results

Table 3 shows the comparative results of the evaluation across all quantitative metrics on the MIMIC-III-50 dataset. Compared with the text classification model, our proposed model gets better results on all metrics. Compared to the previous state-of-the-art model based on structural knowledge SHiDAN (Chen et al., 2021a), MVFAN-EK produces notable improvements of 1.4% and 1.3% in macro-F1 and micro-F1. Besides, our model has improved by 0.6%, 0.5%, and 1.5% in macro-AUC, micro-AUC, and P@5 compared to the previous model MSATT-KG (Xie et al., 2019).

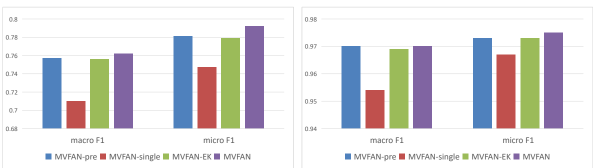
Table 4 shows the result on ChineseEMR dataset. MVFAN-EK outperforms all the baseline models across all the metrics. Compared to the previous model MultiResCNN (Li and Yu, 2020), MVFAN-EK produces notable improvements of 5.9%, 0.9%, 6.5%, 0.5%, 3.2% and 3.1% in macro-F1, macro-AUC, micro-F1, micro-AUC, P@1, and R@1.

From the results on MIMIC-III-50 and ChineseEMR, we come to a conclusion that the performance of the MVFAN-EK model in disease diagnosis can be remarkably improved by combining the diagnostic evidence segment information from different views and the disease empirical knowledge representations.

5.2 Ablation Experiment



(a) The results of ablation studies on MIMIC-III-50



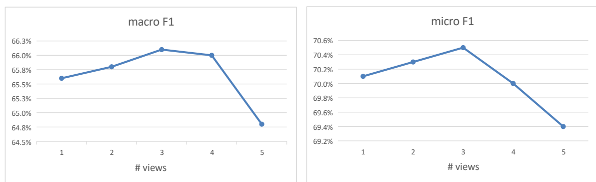
(b) The results of ablation studies on ChineseEMR

Figure 4: "MVFAN-X" means MVFAN without module X. "pre" means not applicable for pretrained model. "single" means use one encoder. "EK" means no structured knowledge.

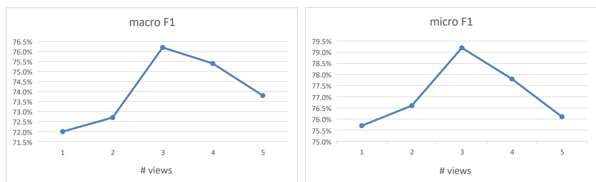
To study the contribution of each component in the MVFAN-EK, we remove each module with an ordinary replacement without changing other modules. Figure 4(a) and figure 4(b) illustrate the results of ablation studies on the MIMIC-III-50 dataset and the ChineseEMR dataset, which verify the effectiveness of each module on the proposed model.

The reduction amount of the MIMIC-III-50 dataset is more apparent than the Chinese dataset. However, when we use one encoder in the MVFA module, the results on the two datasets have a greater degree of decline. For comparative groups without empirical representation of the diseases model, we can see that the performance has slightly decreased. It is observed that removing each component results in reducing all metrics, showing the effectiveness of these three components. Among the three components, the MVFA module has a more significant impact on all datasets, which means that the multi-view feature attention module can indeed extract diagnostic evidence segment information of different lengths for each disease. It reveals that using a structural knowledge graph can capture empirical knowledge of the diseases, which helps the model better diagnose diseases.

5.3 Parameter Sensitivity



(a) The results of parameter sensitivity on MIMIC-III-50.



(b) The results of parameter sensitivity on ChineseEMR.

Figure 5: Parameter sensitivity of MVFAN-EK.

This section investigates the influence of the number of views K and the kernel sizes. As shown figure 5(a) and figure 5(b), we use macro-F1 and micro-F1 as primary metrics on MIMIC-III-50 and ChineseEMR datasets. We vary K from 1 to 5, and there are two different settings for each view, except when $K = 5$. For different views, we selected the best results for plotting. It can be observed from the results, with the increase of K, the

ICD: 96.71 [Continuous invasive mechanical ventilation for less than 96 consecutive hours]
... basos atyps metas myelos promyelo nuc rbcs 24pm hypochrom anisocyt poikilocy macrocyt microcyt normal polychrom burr stippled 24pm plt smr very low plt count 24pm pt ptt inr pt 24pm alt sgpt ast sgot ld ldh alk phos tot bili 44pm type art po2 pco2 ph total co2 base xs brief hospital course patient was transferred from hospital hospital after days of worsening abdominal pain severe hypotension and lactic acidosis he was admitted to hospital hospital on morning was intubated started on pressors and antibiotics and after notifying the transplant center he was transferred in the afternoon and admitted to the surgical icu of hospital1 patient was started on neo synephrine norepinephrine and vasopressin continued of broad spectrum antibiotics and attempted to correct his coagulopathy with blood products prior to perform a diagnostic paracentesis with hepatology this showed wbc and rbc but no microorganisms on the gram stain a right chest thoracentesis for a large right pleural was also performed by the sicu to improve his ventilatory settings and improve his oxygenation which drained l of fluid patient tolerated both procedures well initially but was never stable enough to bring him to ct scan at midnight he started with increasing pressure requirement and was maximized on neo synephrine levophed and vasopressin his profound lactic acidosis with a worsening lactate up to was attempted to be corrected with sodium bicarb with no improvement on his ph of his wife was name ni who decided to continue measures and after giving 5l of fluids including crystalloids colloids blood and at a maximum dose of pressures he was not able to hold his bp patient expired on at am after his the pastor of his church arrived to the sicu his wife doctor first name was doctor first name while she was on her way the admitting office was notified and the medical examiner waived the case his family consented for an autopsy which will be done at hospital1 medications on admission last name un clobetasol clotrimazole 10mg 5x day vit d units weeks lactulose 15mg q4hrs viread 300mg daily mag oxide 400mg hospital1 lasix 80mg hospital1 rifaxamin 550mg hospital1 spironolactome 200mg hospital1 discharge medications none discharge disposition expired discharge diagnosis cardiopulmonary arrest septic shock multiorgan failure renal liver neurologic cardiac end stage liver disease congenital hepatitis b discharge condition expired discharge instructions autopsy to be performed first name11 name pattern1 last name namepattern4 md md number completed by

ICD: 285.9 [Acidosis]
... basos atyps metas myelos promyelo nuc rbcs 24pm hypochrom anisocyt poikilocy macrocyt microcyt normal polychrom burr stippled 24pm plt smr very low plt count 24pm pt ptt inr pt 24pm alt sgpt ast sgot ld ldh alk phos tot bili 44pm type art po2 pco2 ph total co2 base xs brief hospital course patient was transferred from hospital hospital after days of worsening abdominal pain severe hypotension and lactic acidosis he was admitted to hospital hospital on morning was intubated started on pressors and antibiotics and after notifying the transplant center he was transferred in the afternoon and admitted to the surgical icu of hospital1 patient was started on neo synephrine norepinephrine and vasopressin continued of broad spectrum antibiotics and attempted to correct his coagulopathy with blood products prior to perform a diagnostic paracentesis with hepatology this showed wbc and rbc but no microorganisms on the gram stain a right chest thoracentesis for a large right pleural was also performed by the sicu to improve his ventilatory settings and improve his oxygenation which drained l of fluid patient tolerated both procedures well initially but was never stable enough to bring him to ct scan at midnight he started with increasing pressure requirement and was maximized on neo synephrine levophed and vasopressin his profound lactic acidosis with a worsening lactate up to was attempted to be corrected with sodium bicarb with no improvement on his ph of his wife was name ni who decided to continue measures and after giving 5l of fluids including crystalloids colloids blood and at a maximum dose of pressures he was not able to hold his bp patient expired on at am after his the pastor of his church arrived to the sicu his wife doctor first name was doctor first name while she was on her way the admitting office was notified and the medical examiner waived the case his family consented for an autopsy which will be done at hospital1 medications on admission last name un clobetasol clotrimazole 10mg 5x day vit d units weeks lactulose 15mg q4hrs viread 300mg daily mag oxide 400mg hospital1 lasix 80mg hospital1 rifaxamin 550mg hospital1 spironolactome 200mg hospital1 discharge medications none discharge disposition expired discharge diagnosis cardiopulmonary arrest septic shock multiorgan failure renal liver neurologic cardiac end stage liver disease congenital hepatitis b discharge condition expired discharge instructions autopsy to be performed first name11 name pattern1 last name namepattern4 md md number completed by

Figure 6: Visualization of the label attention score.

performance is boosted at first since more views mean more scale features but drops after $K = 3$ as the diagnostic evidence segment may not be too long. From the results, we proposed MVFAN-EK achieves the best performance when $K = 3$ and the kernel size = (3,5,9).

5.4 Interpretability Analysis

Figure 6 illustrates an example of a patient with two diseases, which is randomly selected from the testing set. We extract the attention value of the corresponding disease from the label attention in multiple encoders for visualization. The different colors indicate different diseases of the patient, and We highlight words according to their different weights. The higher the weight, the more obvious the highlight. The visualization concludes that the multi-view feature attention module can extract disease-related information from different views. Thus, we can interpret the diagnosis results through our proposed model to help doctors diagnose diseases.

5.5 Limitations

In our work, the Chinese dataset used has a limited amount of data and involves fewer types of diseases. The next step is to increase the amount of data and further improve the model. On the other hand, our method is currently only in the experimental stage. We have already cooperated with some large tertiary hospitals in China. Next, we will put our work into practical application.

6 Conclusions

In this paper, we propose integrating empirical knowledge into the multi-view feature attention network (MVFAN-EK) method, which consists of two parts. The first portion is the multi-view feature attention module which can capture diagnostic evidence segment information of different lengths for each disease by multi CNNs and label-wise attention mechanism. Besides, we employ GCN to extract the empirical knowledge representation of diseases from an abductive causal graph in the second portion. We mainly use the MVFAN-EK model containing the above two parts for disease diagnosis, and conduct experiments on two real-world EMR datasets. The experimental results prove that the effectiveness of the proposed model in disease diagnosis. We will further expand this study to diagnose more kinds of diseases by incorporating more structural external knowledge.

References

Insaf Adjabi, Abdeldjalil Ouahabi, Amir Benzaoui, and Abdelmalik Taleb-Ahmed. 2020. Past, present, and future of face recognition: A review. *Electronics*, 9(8):1188.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1538–1548.

Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. 2020. Rumor detection on social media with bi-directional graph convolutional networks. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 549–556.

R Bohmer, G Pisano, R Sadun, and T Tsai. 2020. How hospitals can manage supply shortages as demand surges. *Harvard Business Review*, 3.

Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu, and Weifeng Chong. 2020. HyperCore: Hyperbolic and co-graph representation for automatic ICD coding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3105–3114.

Jun Chen, Xiaoya Dai, Quan Yuan, Chao Lu, and Haifeng Huang. 2020. Towards interpretable clinical diagnosis with Bayesian network ensembles stacked on entity-aware CNNs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3143–3153.

Jun Chen, Quan Yuan, Chao Lu, and Haifeng Huang. 2021a. A novel sequence-to-subgraph framework for diagnosis classification. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3606–3612. Main Track.

Weile Chen, Huiqiang Jiang, Qianhui Wu, Börje Karlsson, and Yi Guan. 2021b. AdvPicker: Effectively Leveraging Unlabeled Data via Adversarial Discriminator for Cross-Lingual NER. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 743–753.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

Luciano RS de Lima, Alberto HF Laender, and Berthier A Ribeiro-Neto. 1998. A hierarchical approach to the automatic categorization of medical documents. In *Proceedings of the seventh international conference on Information and knowledge management*, pages 132–139.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Tsu-Jui Fu, Peng-Hsuan Li, and Wei-Yun Ma. 2019. GraphRel: Modeling text as relational graphs for joint entity and relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1409–1418.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778.

Ayuki Joto, Takahiro Fuchi, Hiroshi Noborio, Katsuhiko Onishi, Masahiro Nonaka, and Tsuneo Jozen. 2021. Construction of a knowledge base for empirical knowledge in neurosurgery. In *Human-Computer Interaction. Interaction Techniques and Novel Applications*, pages 521–537.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270.

Fei Li and Hong Yu. 2020. ICD coding from clinical text using multi-filter residual convolutional neural network. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8180–8187.

Xiaonan Li, Hang Yan, Xipeng Qiu, and Xuanjing Huang. 2020. FLAT: Chinese NER using flat-lattice transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6836–6842.

Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Yang Liu, Hua Cheng, Russell Klopfer, Matthew R Gormley, and Thomas Schaaf. 2021. Effective convolutional attention network for multi-label clinical document classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5941–5953.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035.

Alessandra Quaranta. 2021. The consilia by learned physicians pietro andrea mattioli and francesco partini: Dialectic relations between doctrine, empirical knowledge and use of the senses in sixteenth-century europe. *Social History of Medicine*.

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*.

Haoran Shi, Pengtao Xie, Zhiting Hu, Ming Zhang, and Eric P Xing. 2017. Towards automated icd coding using deep learning. *ArXiv preprint*.

Zhenchao Sun, Hongzhi Yin, Hongxu Chen, Tong Chen, Lizhen Cui, and Fan Yang. 2020. Disease prediction via graph neural networks. *IEEE Journal of Biomedical and Health Informatics*, (3):818–826.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, (5):552–556.

Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. 2020. A label attention model for ICD coding from clinical text. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3335–3341.

Ke Wang, Xuyan Chen, Ning Chen, and Ting Chen. 2020a. Automatic emergency diagnosis with knowledge-based tree decoding. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3407–3414.

Qi Wang, Yangming Zhou, Tong Ruan, Daqi Gao, Yuhang Xia, and Ping He. 2019. Incorporating dictionaries into deep neural networks for the chinese clinical named entity recognition. *Journal of biomedical informatics*, page 103133.

Xiaobo Wang, Shifeng Zhang, Shuo Wang, Tianyu Fu, Hailin Shi, and Tao Mei. 2020b. Mis-classified vector guided softmax loss for face recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12241–12248.

Yonghui Wu, Min Jiang, Jun Xu, Degui Zhi, and Hua Xu. 2017. Clinical named entity recognition using deep learning models. In *AMIA Annual Symposium Proceedings*, page 1812. American Medical Informatics Association.

Jing Xie, Jingchi Jiang, Yehan Wang, Yi Guan, and Xitong Guo. 2020. Learning an expandable emr-based medical knowledge network to enhance clinical diagnosis. *Artificial Intelligence in Medicine*, page 101927.

Xiancheng Xie, Yun Xiong, Philip S. Yu, and Yangyong Zhu. 2019. EHR coding with multi-scale feature attention and structured knowledge graph propagation. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, pages 649–658.

Zhongliang Yang, Yongfeng Huang, Yiran Jiang, Yuxi Sun, Yu-Jin Zhang, and Pengcheng Luo. 2018. Clinical assistant diagnosis for electronic medical record based on convolutional neural network. *Scientific reports*, 8(1):1–9.

Quan Yuan, Jun Chen, Chao Lu, and Haifeng Huang. 2020. The graph-based mutual attentive network for automatic diagnosis. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3393–3399.

Yuhao Zhang, Yuhui Zhang, Peng Qi, Christopher D Manning, and Curtis P Langlotz. 2021. Biomedical and clinical english model packages for the stanza python nlp library. *Journal of the American Medical Informatics Association*, (9):1892–1899.