KAICD: A knowledge attention-based deep learning framework for automatic ICD coding

Yifan Wu[1], Min Zeng[1], Zhihui Fei[1], Ying Yu[1], Fang-Xiang Wu[2], Min Li[1*]

[1]School of Computer Science and Engineering, Central South University, Changsha, 410083, P.R. China

[2]Division of Biomedical Engineering and Department of Mechanical Engineering, University of Saskatchewan, Saskatoon, SKS7N5A9, Canada.

E-mail addresses: limin@mail.csu.edu.cn

Abstract: Automatic International Classification of Diseases (ICD) coding is an important task in the future of artificial intelligence healthcare. In recent years, a lot of traditional machine learning-based methods have been proposed, and they achieved good results on this task. However, these traditional machine learning-based methods for automatic ICD coding only focus on the semantic features of clinical notes and ignore the feature extraction of ICD titles that are the descriptions of ICD codes. In this paper, we propose a knowledge attention-based deep learning framework called KAICD for automatic ICD coding. KAICD makes full use of the clinic notes and the ICD titles. The semantic features of clinic notes are extracted by a multi-scale convolutional neural network. For ICD titles, we use attention-based Bidirectional Gated Recurrent Unit (Bi-GRU) to build a knowledge database, which can offer additional information. Depending on input clinic notes, we can use the attention mechanism to obtain different knowledge vectors from the knowledge database where some ICD titles are more relevant to the input clinic notes. Last, we concatenate the knowledge vectors and the semantic features of clinic notes, and use them for the final prediction. KAICD is tested on a public dataset Medical Information Mart for Intensive Care III (MIMIC III); it achieves micro-precision of 0.502, micro-recall of 0.428, and micro-f1 of 0.462, which outperforms other competing methods. Furthermore, the results of the ablation study show that the knowledge database of ICD titles learned by the attention-based Bi-GRU enhances the feature expression and improves the prediction performance.

Key words: automatic ICD coding, clinic notes, ICD titles, Bidirectional Gated Recurrent Unit, attention, knowledge database

1 Introduction

In the past decades, large-scale biomedical datasets have been collected, analyzed and applied in precision medicine. Various computational methods and learning techniques have been used to extract symbolic and ontologic information to pave the way for better healthcare. In a medical text, the International Classification of Diseases (ICD) is essential and important. ICD is developed by the World Health Organization [1] for systematically classifying diseases worldwide. According to etiology, symptoms, and pathology, diseases are encoded into different codes which are generally composed of several letters and numbers. ICD coding plays an important role in hospital statistics, academic exchanges between different regions, and medical insurance reimbursement. In addition, it can facilitate the analysis and application of electronic medical records (EMR) to achieve future artificial intelligence health. Currently, it is mainly based on artificial ICD coding in practice. ICD coders need to memorize the contents of ICD codes and then label the clinic notes. This is an extremely time-consuming, laborious and expensive task. In addition, ICD coders are easy to make mistakes under long-term work time. It is estimated that the annual financial expenditure for ICD coding in USA is about $25 billion [2, 3]. Therefore, the implementation of automatic ICD coding is a very hot research topic.

Researchers applied traditional machine learning methods for automatic ICD coding, such as Support Vector Machine (SVM) [4], Bayes classifier [5], K-Nearest Neighbor (KNN) [6]. Those traditional machine learning algorithms have done a lot of valuable attempts for automatic ICD coding. However, almost these machine learning methods require a large number of feature engineering. With the development of deep learning [7], these problems have been gradually resolved. Deep learning greatly reduces the workload of feature engineering through deep network structure and a large number of learnable parameters. Deep learning techniques have improved a lot of valuable research in many tasks in the fields of bioinformatics and biomedicine, such as protein-protein interaction site prediction [8], protein function prediction [9], protein subcellular localization [10], essential protein prediction [11, 12], biomedical imaging

[13, 14], EMR [15, 16], omics [17], etc. These successful applications of deep learning in the fields of bioinformatics and biomedicine have also contributed to the development of automatic ICD coding. Li et al. proposed DeepLabeler [18] which combines the global semantic features extracted by Doc2Vec and the local semantic features extracted by a multi-scale convolutional neural network (CNN). Zeng et al. [19] pre-trained the word vector in biomedical texts of MEDLINE/PubMed database to improve the expression of the word representation. They used the structures of multi-scale CNN for automatic ICD coding on Medical Information Mart for Intensive Care III (MIMIC III) dataset. Guo et al. [20] used the MetaMap tool to extract the patient symptoms directly from clinic notes and used the Bidirectional Long Short-Term Memory (Bi-LSTM) and Term Frequency-Inverse Document Frequency (TF-IDF) to perform the prediction of the first three characters of ICD codes. Yu et al. [21] proposed the multilayer attention bidirectional LSTM for Chinese EMR and they extracted the semantic representations with the character vector to improve the performance. Cao et al. [22] used three Convolutional Attention for Multi-Label (CAML) structure to predict the 3-digit, 4-digit, and 6-digit of ICD codes. The output of the CAML for 3-digit is used as the feature input for the 4-digit CAML, and the output of the CAML for 4-digit is used as the feature input for the 6-digit CAML.

These studies have made a great contribution to automatic ICD coding and inspire our research [18, 19, 21, 23, 24]. Although the previous studies have reached great results, there is still much room for improvement. In the previous studies, researchers only focus on extracting features of clinic notes and ignore ICD titles. Actually, the information of ICD titles is critical for automatic ICD coding. ICD titles are language descriptions of ICD codes, and play a significant role in manually ICD coding. Thus, the lack of information of ICD titles can decrease the performance of existing methods.

To take advantage of the information in ICD titles, we design a knowledge attention-based deep learning framework called KAICD. The main idea is that we construct a knowledge database of all ICD titles by using an attention-based Bidirectional Gated

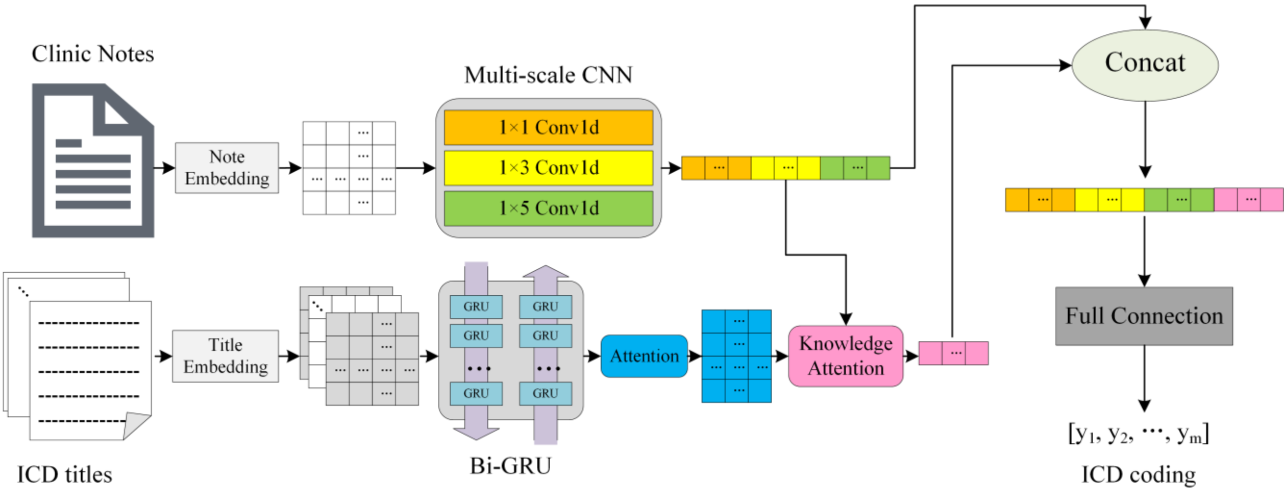

Figure 1. The architecture of KAICD. KAICD has two inputs: a clinic note and all ICD titles. For a clinic note, KAICD uses a multi-scale CNN to extract its features. For all ICD titles, KAICD uses Attention-based Bidirectional Gated Recurrent Unit (Bi-GRU) to build a knowledge database. Depending on input clinic notes, the attention mechanism is applied to obtain knowledge vectors from the knowledge database. Finally, the features of clinic notes and the knowledge vectors are concatenated to predict the probability of ICD codes.

Recurrent Unit (Bi-GRU). Then according to the different input clinic notes, we find some unique knowledge vectors which are more relevant to input clinic notes to offer more information about ICD codes. Lastly, we concatenate the knowledge vectors and the semantic features of clinic notes, and use them for the final prediction.

In this paper, we propose a knowledge attention-based deep learning framework called KAICD which combines the semantic features of clinic notes and the external knowledge database of all ICD titles. A multi-scale CNN and an attention-based Bi-GRU network are used to extract the features of clinic notes and ICD titles, respectively. Extensive experiments are conducted on MIMIC III and the results show that KAICD outperforms other competing methods.

2 Methods

In this part, the overview of KAICD is described in section 2.1. Then, the methods of data preprocessing and the feature extraction are described in sections 2.2, 2.3-2.5, respectively. Finally, other details are introduced in section 2.6.

2.1 Overview

The overview of KAICD is shown in Fig.1. KAICD has two kinds of inputs: clinic notes and ICD titles. Firstly, for a clinic note, KAICD encodes each of its words, and the multi-scale CNN is used to extract the features of clinic notes; for all ICD titles, the attention-based Bi-GRU is used to create the knowledge database. Depending on the input clinic note, KAICD uses the attention mechanism to obtain different knowledge vectors from the knowledge database where some ICD titles are more relevant to input clinic notes. Then the features of clinic notes and the knowledge vectors are concatenated as the input of the classifier. Finally, the probabilities of each ICD coding are obtained through the fully connected layer with the sigmoid activation.

2.2 Preprocessing raw clinic notes

There is a lot of noise in raw clinic notes and thus the first step is preprocessing them. Automatic ICD coding is an unbalanced and sparse multi-label classification task. For a clinic note, it usually only corresponds to rare ICD codes which are much smaller than 1% of the total number of ICD codes. Besides, in an EMR dataset, many low-frequency ICD codes only appear less than 10 times, while high-frequency ICD codes can cover more than 50% of the entire dataset. Therefore, the noise in clinic notes should be minimized to obtain high-quality features in the training process.

Many less important words are included in clinic notes, which can affect the performance of the model. In addition, some clinic notes are too long which brings lots of computational costs. Therefore, we need to refine clinic notes.

Term Frequency-Inverse Document Frequency (TF-IDF) is an effective measure of the importance of words in a text [25]. For the word i in text j, TF-IDF is defined as:

$$\text{TF-IDF}_{ij} = \text{TF}_{ij} \times \text{IDF}_i \tag{1}$$

where TF_{ij} is the term frequency of word i in document j, IDF_i is the inverse document frequency of the word i:

$$\mathrm{TF}_{ij}=\frac{n_{ij}}{\sum_k n_{kj}} \tag{2}$$

$$\mathrm{IDF}_i=\log(\frac{|D|}{1+|D_i|}) \tag{3}$$

where n_{ij} represents the number of occurrences of word i in document j, $|D|$ represents the total number of documents in the dataset, $|D_i|$ represents the total number of documents containing word i in the dataset.

We use TF-IDF as the measure of word effectiveness to refine clinic notes. Only the top k most effective words are retained for each clinic note. The results are all text with the same length (For those short clinic notes, zeros will be used to pad to the same length). For the title of ICD, we concatenate the short title and the long title as the final ICD title.

2.3 Feature extraction from clinic notes

The second step is representing a clinic note and extracting its features. Word2vec [26] is a powerful technique to represent words. At present, the word vector training models mainly include CBOW, Skip-Gram, GloVe, Bert. Among them, the Skip-Gram model trains word vectors by using the central words to predict the contexts, which has better mapping effectiveness on uncommon words. For the clinic notes and ICD titles, many words are low in frequency but important. Therefore, Skip-Gram is chosen in our model.

So far, we have obtained the embedding of clinic notes. The question then arises: how can we obtain useful features from the embedding? Since LeNet-5 [27] was proposed, CNN has become well-known and has achieved great success in the computer vision. Text data can also be viewed as a 1D image; therefore CNN is suitable for processing it. KIM YOON [28] applied CNN to text classification, and the multi-scale convolution kernels were used to extract features. The clinic note is the basis for ICD coding which contains almost all the information. Inspired by previous studies, we use a multi-scale CNN to extract the semantic features of different granularities in the clinic note.

Clinic notes are first encoded by the note embedding as sequences of word vectors. Their n-gram features are extracted by the multi-scale 1D CNN and the max-pooling layer. For each clinic note, the output of all CNNs are then concatenated together to obtain the final clinic notes' feature matrix $v \in \mathbb{R}^{1 \times \Sigma d_{\mathrm{n}}^{i}}$. d_{n}^{i} is defined as the kernel number of CNN i.

$$v = [v^1; v^2; v^3] \tag{4}$$

where $v^i \in \mathbb{R}^{1 \times d_{\mathrm{n}}{}^{i}}$ represents the feature matrixes extracted by the multi-scale CNN.

2.4 Creating a knowledge database of all ICD titles

The third step is creating a knowledge database of all ICD titles. RNN is a type of neural networks, which is well suited to deal with sequences. Since Long Short-Term Memory (LSTM) [29] and GRU [30] were proposed, the gradient disappearance and gradient explosion problems of traditional RNN have been greatly improved. Compared to CNN, RNN can handle the order between words, but its computational cost is greater. For ICD titles, it includes lots of semantic information about ICD codes. There are some semantic relationships between clinic notes and ICD title i if clinic notes have ICD code i. Therefore, the features of all ICD titles are extracted as the knowledge database by Bi-GRU. We define d_{h} as the hidden size of GRU, d_{h} as the feature size of an intermediate variable, k_{t} as the length of an ICD title， m as the number of ICD codes.

ICD titles are first encoded by the title embedding and then through an Attention-based Bi-GRU, the feature matrix $A = [A_1, A_2, ..., A_m]^{\mathrm{T}}$ is obtained.

$$A_i = h_i^{\mathrm{T}} \cdot \beta_i \tag{5}$$

where $A_i \in \mathbb{R}^{d_{\mathrm{h}} \times 1}$ is the feature matrix of ICD title i, m is the number of ICD codes, $h_i \in \mathbb{R}^{k_{\mathrm{t}} \times d_h}$ is the output of GRU hidden states for ICD title i, $\beta_i \in \mathbb{R}^{k_{\mathrm{t}} \times 1}$ is the attention weight vector of ICD title i at each position and it is computed by

$$h_i = \mathrm{f}_1(h_i) \tag{6}$$

$$\beta_i = \mathrm{softmax}(h_i \cdot W_{\mathrm{h}}) \tag{7}$$

where f_1 is the fully connected layer with tanh as the activation function, $h_i \in \mathbb{R}^{k_{\mathrm{t}} \times d_{\mathrm{h}}}$ is the output of f_1, $W_{\mathrm{h}} \in \mathbb{R}^{d_{\tilde{\mathrm{h}}} \times 1}$ is the weight matrix.

2.5 Knowledge attention

After obtaining the knowledge database matrix A, we need to calculate the weight of each ICD title according to the semantic features v of clinic notes. We use the attention mechanism [31] to implement it. Attention is a mechanism that simulates the human eye's focusing behavior. Bahdanau et al. [32] applied attention mechanism to natural language processing (NLP) field for the first time. The introduction of attention mechanism has greatly improved the effectiveness of many models, and we use it to construct the features of knowledge attention about clinic notes and ICD titles.

The knowledge attention is designed by referring to the attention structure of Bahdanau et al. [32] and Luong et al. [33]. The knowledge vector $u \in \mathbb{R}^{1 \times d_{\mathrm{h}}}$ is calculated as below and we use it to enhance the expression of clinic notes.

$$u = \alpha \cdot A \tag{8}$$

where $A \in \mathbb{R}^{m \times d_{\mathrm{h}}}$ is the feature matrix of ICD titles extracted by Bi-GRU, $\alpha \in \mathbb{R}^{1 \times m}$ is the attention weight of each ICD title and it is computed by

$$\tilde{v} = \mathrm{f}_2(v) \tag{9}$$

$$\alpha = \mathrm{softmax}(\tilde{v} \cdot A^{\mathrm{T}}) \tag{10}$$

where v is the clinic note feature matrix, f_2 is the fully connected layer with tanh as the activation function, $\tilde{v} \in \mathbb{R}^{1 \times d_{\mathrm{h}}}$ is the output of f_2.

2.6 Other details

The semantic feature matrix v of clinic notes and the knowledge vector u are concatenated, and a fully connected layer with the sigmoid activation is used to obtain the probability of occurrence of each ICD code. Finally, a multi-label one-versus-all loss based on the max-entropy and Adam [34] optimizer is used in the training process.

$$y = \text{sigmoid}([v;u] \cdot W_o + b_o) \tag{11}$$

where W_o is the weight matrix, b_o is the bias vector.

3 Results and discussion

In this part, we first introduce the public dataset MIMIC III and some processing details in section 3.1. Then we introduce the evaluation metrics in our experiments in section 3.2. The results and ablation studies are described in sections 3.3-3.4.

3.1 Dataset source

MIMIC III is a large, single-center database comprising information related to patients admitted to critical care units at a large tertiary care hospital [35]. We mainly use table ADMISSIONS, table NOTEEVENTS, table DIAGNOSES_ICD and table D_ICD_DIAGNOSES in this study. All clinic notes in table NOTEEVENTS are used to train the word vectors of clinic notes. All short and long titles in table D_ICD_DIAGNOSES are used to train the word vectors of ICD titles. All clinic notes with CATEGORY “Discharge summary” and DESCRIPTION “Report” are selected as the input of our model.

We drop low-frequency words, stop words and the words with too short length firstly: the minimum frequency and minimum length of words are set to 10 and 3, respectively. Then some special symbols and useless content are also removed: only the letters, numbers and punctuation marks are kept; the obvious useless parts in clinic notes like “admission date”, “dictated by …”, etc. are removed. After that, the statistics of clinic notes and its ICD codes are summarized in Table 1. It can be seen from Table 1 that the length of clinic notes is between 7 and 6808, with the average length is 1293 and the standard deviation of the length is 668. Therefore, if the short text is directly padded by 0 to achieve text alignment, too much useless data will be introduced. In addition, of the 6984 ICD codes included in MIMIC III, the maximum number of codes for a clinic note is only 39, which shows the sparsity of classification.

Table 1. Overview of clinic notes and its ICD codes.

Statistics	MIMIC III
Total number of notes	55095
Maximum number of words per note	6808
Minimum number of words per note	7
Average number of words per note	1293
Standard deviation of the number of words per note	668
Maximum number of words per ICD title	41
Minimum number of words per ICD title	1
Total number of codes	6984
Maximum number of codes per note	39
Minimum number of codes per note	1
Average number of codes per note	11
Standard deviation of the number of codes per note	6

We also analyze the frequencies of occurrence of each ICD codes, and plot Fig.2. MIMIC III is a critical care database, which means most ICD codes that appear in the dataset are related to the critical care. The most frequent 1000 of 6984 ICD codes account for nearly 90% of the total frequency and nearly 60% of ICD codes appear less than 10 times.

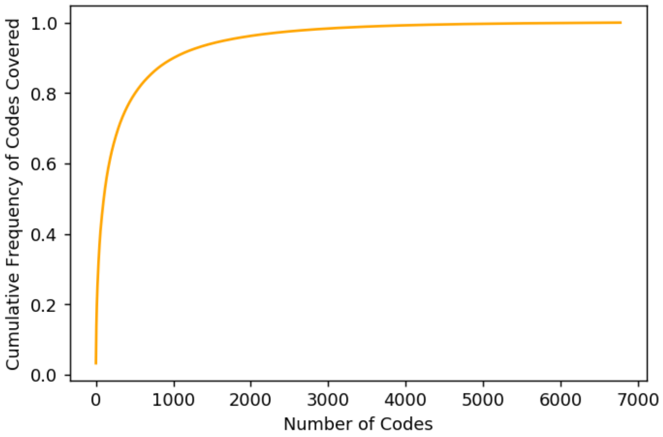

Figure 2. Cumulative frequency of codes covered. The horizontal axis represents the number of ICD codes, which are ranked according to the frequency of occurrence. The vertical axis represents the coverage of all codes on all clinic notes.

3.2 Evaluation metrics

The above analysis of the dataset demonstrates that the automatic ICD coding in MIMIC III is an unbalanced and sparse multi-label classification task. However, in practice, we usually give the priority to the accurate prediction of high-frequency categories than pay too much attention to low-frequency categories. Therefore, in this paper, the micro-averaging measurements: precisions (MiP), recall (MiR), f1-score (MiF) are used for model evaluation. Among them, f1-score is the harmonic mean of the other two and are used as the evaluation metrics for model evaluation.

$$\mathrm{MiP} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij} y_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}} \tag{12}$$

$$\mathrm{MiR} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij} y_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}} \tag{13}$$

$$\mathrm{MiF} = \frac{2 \times \mathrm{MiP} \times \mathrm{MiR}}{\mathrm{MiP} + \mathrm{MiR}} \tag{14}$$

where y_{ij} and y_{ij} are the true and predicted label for sample i in ICD j.

3.3 Results

The dataset is split into 80% for training and 20% for testing. All of the experiments are performed on a server with 1 NVIDIA TITAN XP GPUs. The batch size is set to 256 and most trainings are done within 30 epochs.

We fine-tune the hyper-parameters of our model for achieving the best performance of automatic ICD coding. The curves of different evaluation metrics on some hyper-parameters are shown in Fig.3. From Fig.3, we find that the number of keywords retained per note has a great impact on the performances and the best length is around 768. Excessively long or short notes can result in a decrease in MiF. As for the vector size of a word in note embedding, different evaluation metrics tend to be saturated when it exceeds 300. In addition, properly increasing the filter size in CNN can improve performance. Fig.3 (d) shows that the optimal threshold is around 0.2.

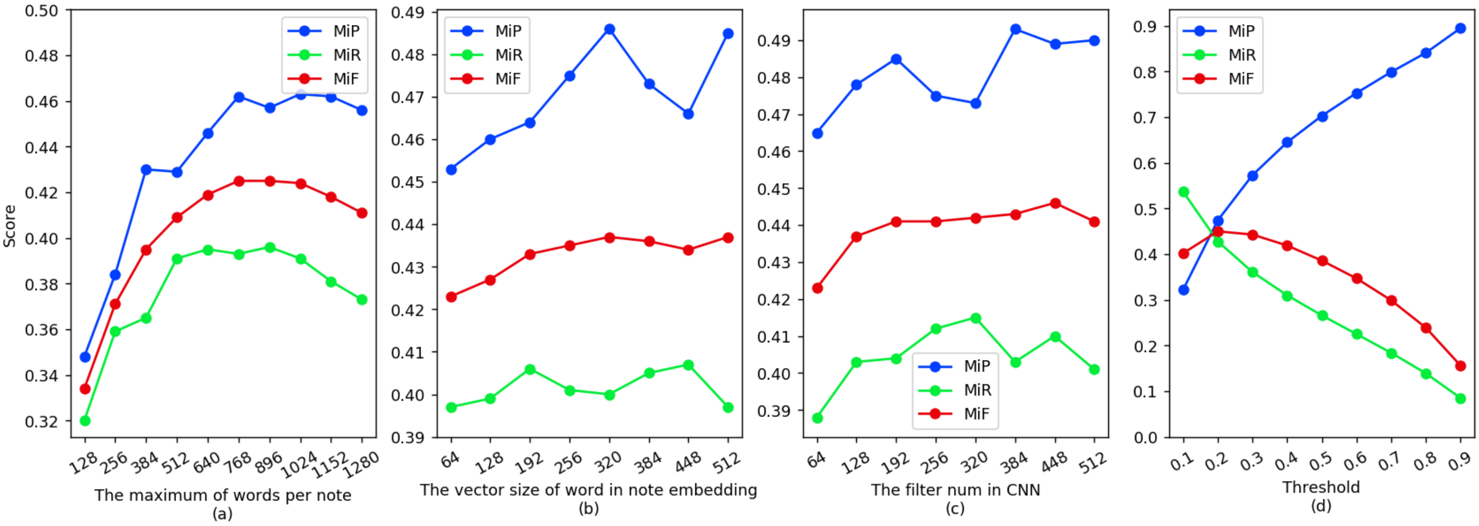

Figure 3. Performances of our model in different parameters.

After fine-tuning hyper-parameters, the optimal hyper-parameters of our model are shown in Table 2.

Table 2. The optimal parameters of our model.

Parameter	Value
Note maximum length	768
Vector size of word in note embedding	320
Vector size of word in title embedding	192
Kernel size in CNN	1,3,5
Kernel number in CNN	448
Hidden size in Bi-GRU	128
Dropout rate	0.5
Threshold	0.2

In order to verify the effectiveness of our proposed model, we select several previous models as baseline methods. The hierarchy-based SVM [4] is a classic machine learning model for automatic ICD coding. We implement and apply it on MIMIC III, which obtained the MiF of 0.335. DeepLabeler [18], TransferLabeler [19], MA-BiRNN [21] were proposed in our previous studies, which obtained the MiF of 0.408, 0.420, 0.420 on ICD-9 automatic coding of MIMIC III, respectively. The performances of different models are shown in Table 3.

Table 3. Performances of different models on MIMIC III.

Model	Micro Precision	Micro Recall	Micro F1
Hierarchy-based SVM [4]	0.415	0.280	0.335
DeepLabeler [18]	0.486	0.351	0.408
TransferLabeler [19]	0.483	0.371	0.420
MA-BiRNN [21]	0.432	0.408	0.420
KAICD	**0.502**	**0.428**	**0.462**

Two conclusions can be drawn from the experimental results. The first one is that the deep learning model (DeepLabeler, TransferLabeler, MA-BiRNN, KAICD) is generally superior to the traditional machine learning model (Hierarchy-based SVM) on this task. The MiF score of hierarchy-based SVM is lower than the deep learning model about 21%, which indicates that the deep learning models have better performance. Second, by introducing the knowledge of ICD titles, KAICD has learned more useful features on clinic notes and ICD titles, and has reached 0.502 in MiP, 0.428 in MiR, 0.462 in MiF score, which exceeds the hierarchy-based SVM [4] by 37.9%, the DeepLabeler [18] by 13.2%, the TransferLabeler [19] by 10.0%, and the MA-BiRNN [21] by 10.0% in MiF, respectively.

We also applied KAICD on our local private Chinese dataset, which comes from Xiangya Big Data Platform of Healthcare. Xiangya dataset is very similar to MIMIC III, and it is gathered Electronic Medical Records in recent years from three affiliated hospitals of Central South University. We used character embedding to encode the Chinese clinic notes on Xiangya dataset to avoid the errors caused by Chinese word segmentation. Table 4 shows the results of KAICD and MA-BiRNN. From Table 4, we can see that KAICD achieves decent results that exceed the MA-BiRNN by 5.2% in MiF.

Table 4. Performances of different models on Xiangya dataset

Model	Micro Precision	Micro Recall	Micro F1
MA-BiRNN [21]	0.704	0.586	0.639
KAICD	**0.747**	**0.610**	**0.672**

3.4 Ablation studies

KAICD combines the features of clinic notes extracted by the multi-scale CNN and the features of ICD titles extracted by the attention-based Bi-GRU. The combination of the two parts makes the model achieve a better performance than other methods. To show the effects of the multi-scale CNN and the attention-based Bi-GRU, we conduct some model ablation studies. The first step is to compare effects of the knowledge database which is created by the attention-based Bi-GRU. We use our model to compare another model without the attention-based Bi-GRU (only the CNN part). The experiments are performed on MIMIC III with the same parameters and the results are shown in Fig.4.

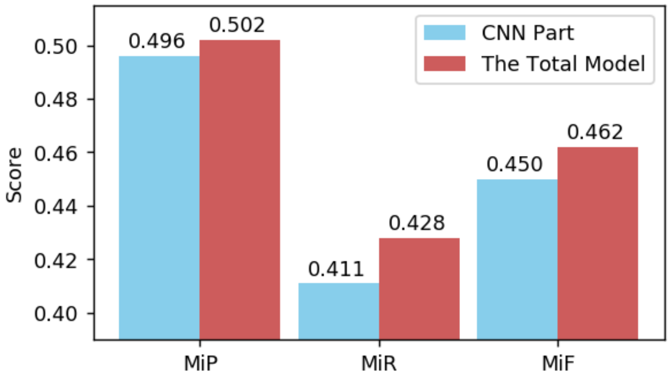

Figure 4. Performances of the CNN part and the total model.

According to the results from Fig.4, we find that the knowledge database which is created by the attention-based Bi-GRU is helpful in our model. Without the knowledge database, MiP, MiR, and MiF drop from 0.502, 0.428, and 0.462 to 0.496, 0.411, and 0.450, respectively. The results show the power of introducing the knowledge database of the ICD titles.

In addition, the word representation is the basis of all NLP tasks, and the quality of word representation can greatly affect the performance of the model. In this paper, Skip-Gram pre-trained word vectors are used in our previous experiments. To test which model is the most suitable, we also test different pre-trained word vectors (CBOW, GloVe and Skip-Gram) in our model. The Bert model is not compared due to its difficulty of fine-tuning the word embedding and its difficulty in migrating to the medical text.

Table 5. Performances of our model with different methods of pre-trained word embedding.

Methods of pre-trained word embedding	Micro Precision	Micro Recall	Micro F1
CBOW [26]	0.497	0.413	0.451
GloVe [33]	0.500	0.414	0.453
Skip-Gram [26]	**0.502**	**0.428**	**0.462**

Table 5 shows the performances on MIMIC III of our proposed model with different methods of pre-trained word embedding. It shows that the Skip-Gram method has achieved the best performance with MiP 0.502, MiR 0.428 and MiF 0.462. Compared with CBOW which uses the context word vector to predict the central word, Skip-Gram model trains word vectors by predicting the context based on the central word, and thus it has more advantages in dealing with uncommon words. This is the main reason why Skip-Gram can improve the performance of the model. As for GloVe [36] which is based on the co-occurrence matrix of words to achieve the training of word vectors, its performance is not as good as Skip-Gram. We think that it is caused by the effect of the small-scale corpus and the important low-frequency words.

As mentioned above, automatic ICD coding is an unbalanced multi-label classification. The frequency of codes has an impact on the results; we want to investigate the difference between the most categories and the least categories. Therefore, the scores of our model for 50 most common ICD codes and 3000 least common ICD codes were calculated, respectively, as shown in Table 6. From Table 6, the prediction accuracy of high-frequency ICD codes and low-frequency ICD codes varies greatly and the bottleneck of the model is in predicting the uncommon ICD codes correctly.

Table 6. Performances of our model for ICD codes with different frequencies.

Frequencies of ICD Codes	Micro Precision	Micro Recall	Micro F1
For 3000 least common ICD codes	0.024	0.015	0.019
For 50 most common ICD codes	0.639	0.610	0.624

In summary, the multi-scale CNN can extract rich features from clinic notes. The addition of knowledge attention provides the ICD title semantic features and improves the model effectiveness. The methods of pre-trained word embedding also have influences on the performances of KAICD, and Skip-Gram is the most suitable pre-training method on this task. In addition, KAICD has a lot of room for improvement in the prediction of low-frequency ICD codes.

4. Conclusion

In this paper, a knowledge-attention deep learning framework called KAICD for automatic ICD coding has been proposed. By introducing knowledge attention-based on ICD titles, the accuracy of automatic ICD coding has been improved, which outperforms other methods. The results of the ablation experiment show that the features of clinical notes extracted by a multi-scale CNN play a key role in our model, while the features of ICD titles learned by attention-based Bi-GRU enhance the feature expression and improve the performance. In addition, Skip-Gram is the most appropriate method for pre-training the word embedding, whose performance is better than CBOW and GloVe.

However, automatic ICD coding is an unbalanced and sparse multi-label classification problem. There are still some problems to be solved, especially the accurate prediction of low-frequency tags. In addition, how to design a powerful model to extract the useful features of clinic notes is a main challenge for researchers.

5. Acknowledgments

This work is supported in part by the NSFC-Zhejiang Joint Fund for the Integration of Industrialization and Informatization under Grant No. U1909208.

References

[1] W.H. Organization, International classification of diseases:[9th] ninth revision, basic tabulation list with alphabetic index, (1978).

[2] D. Lang, Consultant report-natural language processing in the health care industry, Cincinnati Children's Hospital Medical Center, Winter, 6 (2007).

[3] R. Farkas, G. Szarvas, Automatic construction of rule-based ICD-9-CM coding systems, BMC bioinformatics, (BioMed Central2008), pp. S10.

[4] A. Perotte, R. Pivovarov, K. Natarajan, N. Weiskopf, F. Wood, N. Elhadad, Diagnosis code assignment: models and evaluation metrics, Journal of the American Medical Informatics Association, 21 (2013) 231-237.

[5] S.V. Pakhomov, J.D. Buntrock, C.G. Chute, Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques, Journal of the American Medical Informatics Association, 13 (2006) 516-525.

[6] P. Ruch, J. Gobeill, I. Tbahriti, A. Geissbühler, From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding, AMIA Annual Symposium Proceedings, (American Medical Informatics Association2008), pp. 636.

[7] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, nature, 521 (2015) 436.

[8] M. Zeng, F. Zhang, F.-X. Wu, Y. Li, J. Wang, M. Li, Protein–protein interaction site prediction through combining local and global features with deep neural networks, Bioinformatics, (2019).

[9] F. Zhang, H. Song, M. Zeng, Y. Li, L. Kurgan, M. Li, DeepFunc: A Deep Learning Framework for Accurate Prediction of Protein Functions from Protein Sequences and Interactions, Proteomics, (2019) 1900019.

[10] L. Wei, Y. Ding, R. Su, J. Tang, Q. Zou, Prediction of human protein subcellular localization using deep learning, Journal of Parallel and Distributed Computing, 117 (2018) 212-217.

[11] M. Zeng, M. Li, Z. Fei, F. Wu, Y. Li, Y. Pan, J. Wang, A deep learning framework for identifying essential proteins by integrating multiple types of biological information, IEEE/ACM transactions on computational biology and bioinformatics, (2019).

[12] M. Zeng, M. Li, F.-X. Wu, Y. Li, Y. Pan, DeepEP: a deep learning framework for identifying essential proteins, BMC Bioinformatics, 20 (2019) 506.

[13] W. Shen, M. Zhou, F. Yang, C. Yang, J. Tian, Multi-scale convolutional neural networks for lung nodule classification, International Conference on Information Processing in Medical Imaging, (Springer2015), pp. 588-599.

[14] W. Zhu, X. Xiang, T.D. Tran, G.D. Hager, X. Xie, Adversarial deep structured nets for mass segmentation from mammograms, 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), (IEEE2018), pp. 847-850.

[15] A. Rajkomar, E. Oren, K. Chen, H.N. Dai AM, P. Liu, Scalable and accurate deep learning for electronic health records. npj Digit Med. 2018, (January).

[16] R. Miotto, L. Li, B.A. Kidd, J.T. Dudley, Deep patient: an unsupervised representation to predict the future of patients from the electronic health records, Scientific reports, 6 (2016) 26094.

[17] Z. Zhang, Y. Zhao, X. Liao, W. Shi, K. Li, Q. Zou, S. Peng, Deep learning in omics: a survey and guideline, Briefings in functional genomics, 18 (2019) 41-57.

[18] M. Li, Z. Fei, M. Zeng, F. Wu, Y. Li, Y. Pan, J. Wang, Automated ICD-9 coding via a deep

learning approach, IEEE/ACM transactions on computational biology and bioinformatics, (2018).

[19] M. Zeng, M. Li, Z. Fei, Y. Yu, Y. Pan, J. Wang, Automatic ICD-9 coding via deep transfer learning, Neurocomputing, 324 (2019) 43-50.

[20] D. Guo, G. Duan, Y. Yu, Y. Li, F.-X. Wu, M. Li, A disease inference method based on symptom extraction and bidirectional Long Short Term Memory networks, Methods, 173 (2020) 75-82.

[21] Y. Yu, M. Li, L. Liu, Z. Fei, F.-X. Wu, J. Wang, Automatic ICD code assignment of Chinese clinical notes based on multilayer attention BiRNN, Journal of biomedical informatics, 91 (2019) 103114.

[22] L. Cao, D. Gu, Y. Ni, G. Xie, Automatic ICD Code Assignment based on ICD's Hierarchy Structure for Chinese Electronic Medical Records, AMIA Summits on Translational Science Proceedings, 2019 (2019) 417.

[23] D. Guo, M. Li, Y. Yu, Y. Li, G. Duan, F.-X. Wu, J. Wang, Disease Inference with Symptom Extraction and Bidirectional Recurrent Neural Network, 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), (IEEE2018), pp. 864-868.

[24] Y. Yu, M. Li, L. Liu, Y. Li, J. Wang, Clinical big data and deep learning: Applications, challenges, and future outlooks, Big Data Mining and Analytics, 2 (2019) 288-305.

[25] K.S. Jones, A statistical interpretation of term specificity and its application in retrieval, Journal of documentation, (2004).

[26] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781, (2013).

[27] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE, 86 (1998) 2278-2324.

[28] Y. Kim, Convolutional neural networks for sentence classification, arXiv preprint arXiv:1408.5882, (2014).

[29] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation, 9 (1997) 1735-1780.

[30] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, arXiv preprint arXiv:1406.1078, (2014).

[31] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, International conference on machine learning2015), pp. 2048-2057.

[32] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473, (2014).

[33] M.-T. Luong, H. Pham, C.D. Manning, Effective approaches to attention-based neural machine translation, arXiv preprint arXiv:1508.04025, (2015).

[34] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980, (2014).

[35] A.E. Johnson, T.J. Pollard, L. Shen, H.L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L.A. Celi, R.G. Mark, MIMIC-III, a freely accessible critical care database, Scientific data, 3 (2016) 160035.

[36] J. Pennington, R. Socher, C. Manning, Glove: Global vectors for word representation,

Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)2014), pp. 1532-1543.