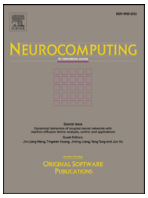
Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Automatic ICD-9 coding via deep transfer learning

Min Zeng[a], Min Li[a,*], Zhihui Fei[a], Ying Yu[a], Yi Pan[b], Jianxin Wang[a]

[a] *School of Information Science and Engineering, Central South University, Changsha 410083, PR China*
[b] *Department of Computer Science, Georgia State University, Atlanta, GA 30302, USA*

ARTICLE INFO

Article history:
Received 29 December 2017
Revised 10 March 2018
Accepted 4 April 2018
Available online 21 May 2018

Keywords:
Automatic ICD-9 coding
Deep learning
Transfer learning
MeSH

ABSTRACT

ICD-9 codes have been widely used to describe a patient's diagnosis. Accurate automatic ICD-9 coding is important because manual coding is expensive, time-consuming. Inspired by the recent successes of deep transfer learning, in this study, we propose a deep transfer learning framework for automatic ICD-9 coding. Our proposed method makes use of transferring MeSH domain knowledge to improve automatic ICD-9 coding. We demonstrate its effectiveness by achieving state-of-the-art performance with a value of 0.420 for Micro-average *F*-measure on MIMIC-III dataset, which indicates that our method outperforms hierarchy-based SVM and flat-SVM. Furthermore, we analyze the deep neural network structure to discover the vital elements in the success of our proposed method. Our experimental results indicate that transfer learning is the key component to improve the performance of automatic ICD-9 coding and deep learning approach is the foundation in the success of our proposed model. In addition, to explore the best network architecture, we also compare the performance of multi-scale and sequential network architectures and find that using multi-scale network is better. Finally, we investigate the effects of transferring different percentage of samples on transfer learning and the results show that the best performance of target domain task can be obtained when 100% number samples are transferred.

1. Introduction

The Ninth Revision of International Classification of Diseases (ICD-9) codes have been widely used to describe a patient's diagnosis including symptoms, statistical analysis of mortality rate and medical reimbursement [1]. ICD-9 codes mean that each disease has a unique code and are used in the electronic health records as a billing mechanism. In most cases, ICD-9 codes are undertaken by coders of the hospital's Medical Record Department, who assign an ICD-9 code to medical record according to a doctor's clinical diagnosis record [2]. To fulfill the task, the coders have to master knowledge in the field of medicine, coding rules and medical terminologies and thus manual coding is expensive and time-consuming. Taking into consideration of these constraints, there is an urgent need to develop an accurate and effective computational method for automatic ICD-9 coding.

Over the past two decades many scientists have explored how to automatically assign ICD-9 codes based on clinical records. From a computational perspective, automatic ICD-9 coding can be considered as a multi-label classification problem, where each ICD-9 code is a class label and each patient has multiple ICD-9 codes. To address this multi-label classification problem, researchers applied machine learning methods such as support vector machine (SVM) [3–6], Naive Bayes [7,8], k-nearest neighbors [9,10], topic model [11,12], and so on [13–15] to automatically assign ICD-9 codes. Adler et al. [6] presented novel evaluation metrics, which help them get a better sense of the usefulness of the hierarchical approach. SVM classifier was used in their experiment for classification. Chen et al. [13] proposed a semantic analytic technique based on dependency parsing to automatically assign clinical ICD-9 codes to complex medical patient records. They evaluated their technique with a real-world corpus and the results showed that their technique is indeed effective in relating strong similar document pairs. Pereira et al. [15] presented three different approaches (search engine, boosting algorithm and rule-based model) to predict the ICD-9 codes of radiology reports. Their experimental results showed that semantic information plays a key role in determining ICD-9 codes.

Although many machine learning methods have been developed for automatic ICD-9 coding, there is a room to improve the classification accuracy. To further improve machine learning methods for automatic ICD-9 coding, we borrow ideas from very recent breakthrough in transfer learning [16]. Transfer learning can learn from one related task and apply that knowledge to a target task. It has been proved a very effective method for classification and thus has been widely applied in bioinformatics field [17–19]. In this

[*] Corresponding author.
E-mail address: limin@mail.csu.edu.cn (M. Li).

https://doi.org/10.1016/j.neucom.2018.04.081
0925-2312/© 2018 Elsevier B.V. All rights reserved.

study we would like to utilize transfer learning for improving the performance of automatic ICD-9 coding because low level features learned from a related task should be helpful for learning target task.

We choose automatic MeSH indexing as an extra auxiliary task to help improve the performance of automatic ICD-9 coding. MeSH is a big medical literature data source which has tens of millions of samples [20,21]. From a computational perspective, these two tasks (MeSH Indexing and ICD-9 coding) have similar inputs (the medical text). Automatic MeSH indexing also can be viewed as a multi-label classification problem. The main difference is that MeSH dataset has a much larger sample size than ICD-9 codes dataset. Because of MeSH has such characteristics, we want to take knowledge learned from automatic MeSH indexing and transfer it to automatic ICD-9 coding. Having learned to automatic MeSH indexing, it might have learned some useful knowledge about word, phrase, sentence, that knowledge could help automatic ICD-9 coding network learn well with less data.

This paper proposes an end-to-end deep transfer learning method to transfer knowledge learned from MeSH source domain into ICD-9 codes domain. We first pre-train our deep learning model using a large number of MeSH dataset and then fine tune this neural network architecture on ICD-9 code dataset. Our experimental results show that our method greatly outperforms the state-of-the-art methods. Furthermore, we investigate the vital elements in the success of our proposed deep transfer learning method. The results demonstrate that transfer learning is the key elements to improve the performance of automatic ICD-9 coding model. In addition, to explore the best network architecture, we evaluated the performance of multi-scale and sequential network architectures and the results suggest that using multi-scale network has resulted in higher Micro-average F-measure. Finally, we investigate the effects of transferring different percentage of samples on transfer learning and the results show that the best performance of target domain task can be obtained when 100% of samples are transferred.

2. Materials

We make use of transfer learning to help improve accuracy and performance of automatic ICD-9 coding. In this study, we choose Multi-parameter Intelligent Monitoring in Intensive Care-III (MIMIC-III) dataset as ICD-9 code dataset and BioASQ3 dataset as MeSH dataset [22]. In the following, we briefly introduce these two datasets.

2.1. MIMIC-III dataset

Multi-parameter Intelligent Monitoring in Intensive Care (MIMIC) is a large, publicly available database, which has been developed by the MIT Lab. It is the most widely-used dataset for healthcare academic and industrial research. MIMIC-III is the third version of the MIMIC Intensive Care Unit (ICU) database and contains data associated with 58,929 distinct hospital admissions for adult patients between 2001 and 2012. Data includes discharge summaries, medications, laboratory measurements, procedure codes, diagnostic codes, survival data etc. In this work, we use only discharge summaries to study automatic ICD-9 coding.

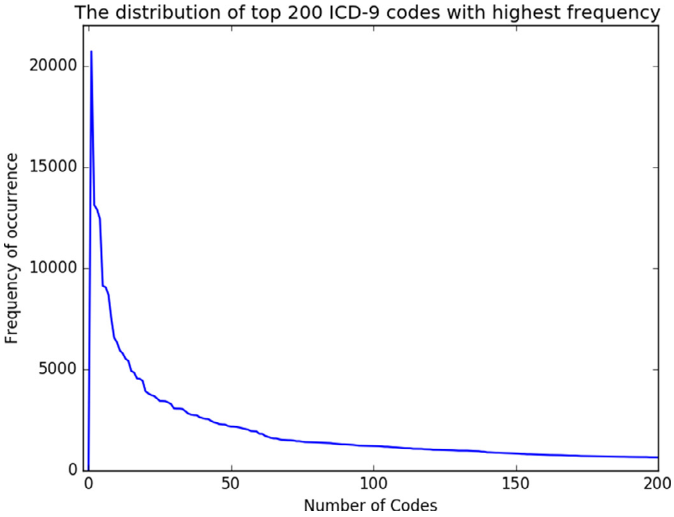


Fig. 1. The distribution of top 200 ICD-9 codes with highest frequency in MIMIC-III dataset.

Using MIMIC-III dataset for automatic ICD-9 coding has several difficulties. First, the distribution of ICD-9 code is highly biased. In Fig. 1 we show the distribution of top 200 codes with highest frequency. The number of all code categories is 6984, and the number of discharge summaries of 105 codes with the highest frequency makes up 50% of the entire samples. The bias of the distribution can be understood more quantitatively from Table 1, which shows four ICD-9 codes, ranked as first, 10th, 100th and 1,000th in terms of their frequencies in all 58,929 medical records from our dataset. The most frequent ICD-9 code, 401.9 (hypertension), appears in 35.1% medical records, while the 1,000th frequent ICD-9 code, 999.8 (other and unspecified transfusion reaction not elsewhere classified), appears in 0.14% medical records only. The bias of distribution usually leads to a poor performance of classification. Second, there is a large variation in the number of ICD-9 codes for each patient. For example, one patient may have 39 associated codes, while another may have only one code. Thirdly, average number of discharge summaries of a code is small. MIMIC-III has 6,984 ICD-9 codes and 58,929 discharge summaries. Average number of discharge summaries of a code is only about 8.43. Such a small average number means that many codes lack of enough samples for training, which leads to a difficulty of classification. Finally, there is a large variation in the length of discharge summary. The length of the longest discharge summary has 4314 words and the shortest one has only a few words.

The characteristics of the MIMIC-III dataset discussed above give rise to difficulties of automatic ICD-9 coding. To tackle these difficulties, we employ transfer learning to improve the performance of automatic ICD-9 coding.

2.2. BioASQ3 dataset

MeSH is the largest medical literature database and is developed by National Library of Medicine. MeSH has been widely used in many natural language processing task such as document searching, document clustering and query expansion. Thus accurate MeSH indexing of medical documents is very important for mining knowledge from this database. MeSH indexing is mostly undertaken by high-quality staff, which is expensive and time-consuming. In view of these limitations, developing an effective automatic MeSH indexing algorithm is very urgently needed. Automatic MeSH indexing also can be viewed as a multi-label classification problem, where each MeSH is a class label and each sample has some MeSHs.

We download MeSH data from Large Scale Biomedical Semantic Indexing Competition (BioASQ3) challenge. 12,208,342 indexed citations with both abstracts and titles are locally stored. There are 27,301 MeSH main headings (MHs) with average samples of a label being 477.12. Thus the total number of samples, total number of labels, and average samples of a label in MeSH dataset are much larger than those of MIMIC-III dataset. The larger number of samples in MeSH dataset gives a great advantage to transfer learning,

Table 1
The first, 10th, 100th and 1000th ICD-9 codes in terms of the frequencies of appearances in 58,929 medical records from MIMIC-III dataset.

ICD-9 Code	Correspond name	Frequency rank	Frequency of patients with code
401.9	Hypertension	1	0.3513
530.81	Esophageal reflux	10	0.1073
V10.46	Personal history of malignant neoplasm of prostate	100	0.0205
999.8	Other and unspecified transfusion reaction not elsewhere classified	1000	0.0014

Table 2
Basic statistics of the two datasets (BioASQ3 and MIMIC-III) used in our experiment.

Dataset	Task	Total number of samples	Total number of labels	Average samples of a label
BioASQ3	Multi-label classification	12,208,342	27,301	447.12
MIMIC-III	Multi-label classification	58,929	6984	8.43

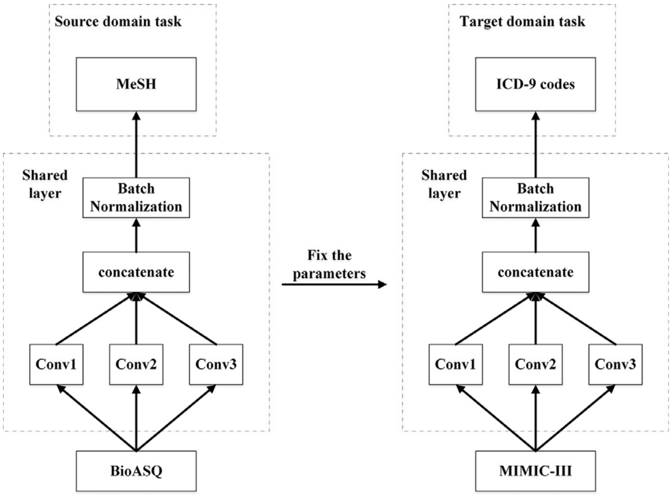


Fig. 2. Overview of our proposed model for automatic ICD-9 coding.

i.e. we have enough data to learn low level features. Detailed comparison between two datasets is summarized in Table 2.

3. Methods

In this section, we first introduce the overview of our proposed deep transfer learning framework in Section 3.1 and then give the details of convolutional neural network, transfer learning, evaluation metrics and baseline models in Sections 3.2, 3.3, 3.4 and 3.5, respectively.

3.1. Overview of deep transfer learning framework

As shown in Fig. 2, our deep learning framework consists of two parts. The first part is to train a neural network for automatic MeSH indexing using BioASQ3 dataset. The second one is to fix the shared network architecture parameters and then to retrain the weights of the output layer for automatic ICD-9 coding using MIMIC-III dataset. The shared network architecture is composed of multi-scale convolutional neural network (CNN) and batch normalization. Multi-scale CNNs are used to detect patterns and extract different scale features for input medical text. Batch normalization after multi-scale CNN layers is used to control the distributions of the output vector, which reduce internal covariate shift. After the shared network architecture, a fully connected layer with sigmoid activation function to predict the probability of label is utilized to classify the medical text. In the first part, we utilize all samples in MeSH dataset to train the shared network architecture parameters to learn high quality low level shared features. In the second part, the MIMIC-III dataset is divided into two parts of training set and testing set. We used shared features in the earlier layers in neural network to retrain the parameters of output layer by using the training set and then we used the testing set to evaluate the performance of our proposed model. The whole process can be considered as another kind of fine-tune operation using a source domain task architecture. The purpose is to use the shared features learned on automatic MeSH indexing to fine tune parameters on smaller target dataset during training and improve the performance of automatic ICD-9 coding.

3.2. Transfer learning

A common assumption in many machine learning methods is that the training and testing data are drawn from the same feature space with the same distribution [16]. However, in many real-world applications, it is difficult to collect sufficient training data to train an effective machine learning model. For example, we are sometimes interested in one domain which does not have sufficient training data. Nevertheless, we have sufficient training data in another relative domain. In such a case, transfer learning can learn from another task in relative domain and then apply that knowledge to target task in domain of interests, which can improve the performance of target task. Relative domain is called source domain and domain of interests is called target domain. In this study, we investigate automatic ICD-9 coding domain and we hope to transfer knowledge learned from automatic MeSH indexing domain to improve performance of target task. We formally defined target domain task and source domain task as follows:

Target domain task: automatic ICD-9 coding by using MIMIC-III dataset;

Source domain task: automatic MeSH indexing by using BioASQ3 dataset;

Source domain task as the auxiliary task can help improve our main task (target domain task).

3.3. Multi-scale CNN

CNN is a class of deep, feed-forward artificial neural network which has successfully been applied to various machine learning tasks [23–27]. Recently, CNNs have been employed for natural language processing (NLP) task such as semantic parsing, sentence modeling, sentence classification, search query retrieval and so on. CNNs utilize layers with kernel filters that are applied to extract local features. It means that CNN layers can automatically learn low-level features from input data. Based on the ability of CNNs, we do not need to handle engineered features.

Multi-scale learning has been proved to be an efficient method to combine different features for classification [28,29]. It is used to obtain multiple local contextual feature maps. As in most convolutional models, we used CNNs to extract features from medical free text. In order to extract better low-level features, we used multi-scale CNN layers with different kernel sizes (see Fig. 2). With the extracted multi-scale low-level features, we concatenate them into a vector as local context feature. Batch normalization after multi-scale CNN layers is used to control the distributions of the concatenated vectors, which reduce internal covariate shift.

3.4. Evaluation metrics

We denote M as the number of all ICD-9 codes, and N as the number of samples. Let y_i and $\hat{y}_i \in \{0, 1\}^M$ be the true and predicted label for sample i. Micro-average F-measure (MiF) is used to evaluate the performance of our proposed model [30,31]. Micro-average F-measure is the harmonic mean of Micro-average Precision (MiP) and Micro-average Recall (MiR).

$$MiF = \frac{2 \cdot MiP \cdot MiR}{MiP + MiR} \tag{1}$$

where

$$MiP = \frac{\sum_{m=1}^{M} \sum_{i=1}^{N} y_i^M \cdot \hat{y}_i^M}{\sum_{m=1}^{M} \sum_{i=1}^{N} \hat{y}_i^M} \tag{2}$$

$$MiR = \frac{\sum_{m=1}^{M} \sum_{i=1}^{N} y_i^M \cdot \hat{y}_i^M}{\sum_{m=1}^{M} \sum_{i=1}^{N} y_i^M} \tag{3}$$

3.5. Baselines to compare

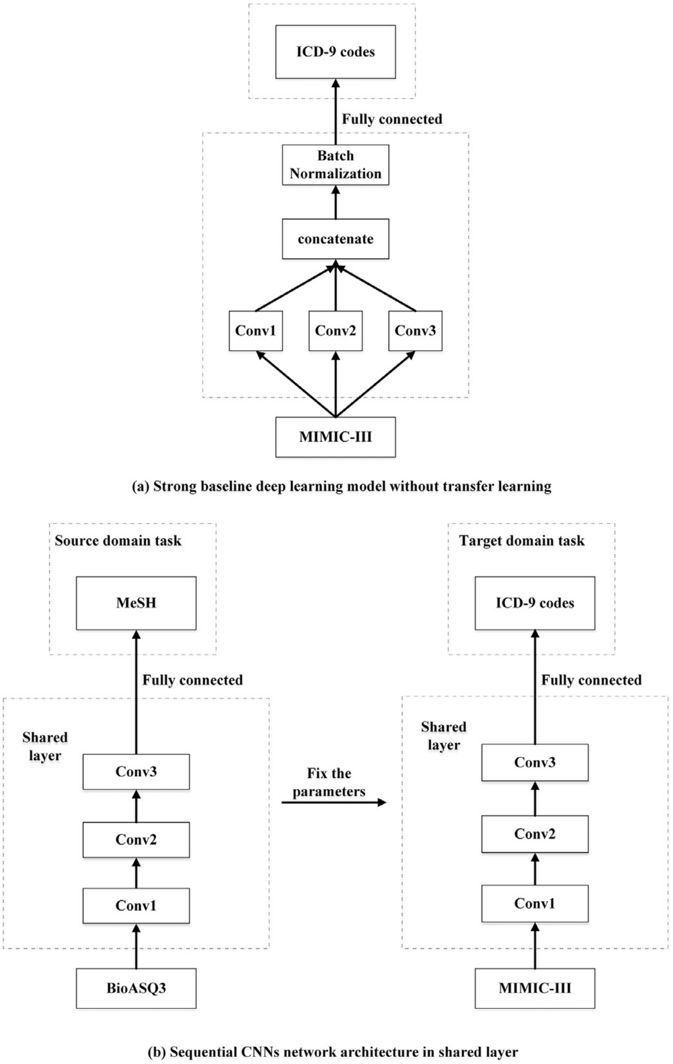
Adler et al. [6] demonstrated that flat hierarchy-based SVM obtained the state-of-the-art performance for automatic ICD-9 coding on MIMIC-II dataset. In their experiment, they used flat SVM as a baseline machine learning model. Thus it is worth to have a comparison with the results of flat SVM and hierarchy-based SVM. To the best of our knowledge, there is no previous work formally reported deep transfer learning framework for automatic ICD-9 coding. In addition to the non-deep learning methods of flat SVM and hierarchy-based SVM, we further developed a strong baseline deep learning model without transfer learning to compare with our deep transfer learning method to explore the validity of transfer learning of our proposed method (see Fig. 3(a)). Furthermore, in order to explore the best network architecture in shared layer, we also compare multi-scale CNNs with sequential network architecture (see Fig. 3(b)).

4. Results

In this section, we first introduce the experimental implementation details in Section 4.1. After that we compare the performance of the proposed approach with state-of-the-art and other base-line model in Section 4.2, and then analyze the architecture of the proposed method in Section 4.3. Furthermore, we compare different network structure as shared layer in Section 4.4. Finally we explore the effects of transferring different percentage of number of samples on transfer learning.

4.1. Implementation details

In this study, the discharge summaries of patients as the training dataset are used for training our model. After removing the discharge summaries with the number of words less than ten, we extracted 52,962 samples with their discharge summaries from MIMIC-III, and the total number of ICD-9 code is 6984. We randomly split the total discharge summaries into training set which consists of 47,665 documents and testing set which has 5297 documents. We use all MeSH data for training parameters of shared network structure. Our code is implemented in Tensorflow, a publicly available deep learning framework developed by Google [32].



(a) Strong baseline deep learning model without transfer learning

(b) Sequential CNNs network architecture in shared layer

Fig. 3. Model architectures: (a) strong baseline deep learning model without transfer learning (b) sequential CNNs network architecture in shared layer.

To obtain the best performance of our proposed method, we have tried a set of different parameters of network architectures to find best parameters for automatic ICD-9 coding. The detailed network structure is as follows. The word embedding size of each word in input medical free text is 100 and CNN layer has 700 hidden units. Each multi-scale layer contains 64 convolutional kernels of size 2, 3, and 4. Rectified Linear Units (ReLu) activation function is used in multi-scale layer. We use a dropout rate of 0.8 on each layer in the network to avoid overfitting. Then we apply batch normalization to reduce internal covariate shift with the batch size being set to 128. Sigmoid activation function is applied to fully connected layer to classification. We train all parameters in our deep network using the Adam optimizer [33]. The loss function we used in our experiments is cross-entropy loss function, which was widely used in classification tasks in deep learning field.

4.2. Comparison with results of baseline models

To evaluate the performance of our proposed method, we have compared our experimental results with those of the following baseline models: flat SVM and the state-of-the-art hierarchy-based SVM [6]. We used all BioASQ3 samples for training parameters in shared layer and computed three evaluation metrics for comparison, which are Micro-average F-measure, Micro-average

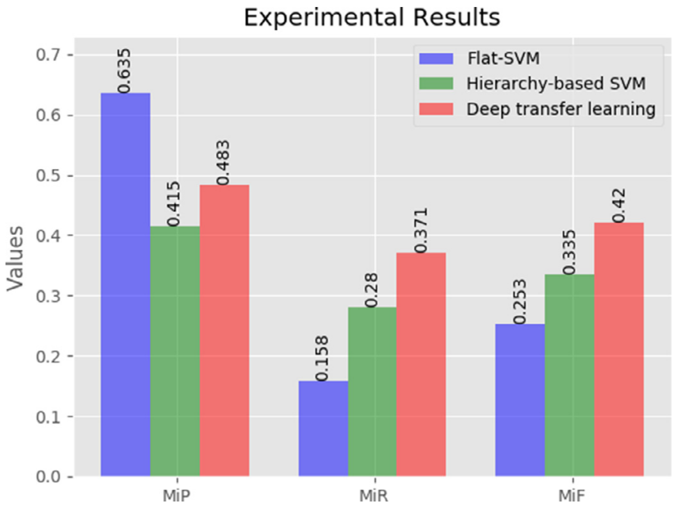


Fig. 4. The performances (Micro-average *F*-measure, Micro-average Precision and Micro-average Recall) of our proposed method and the compared baseline methods.

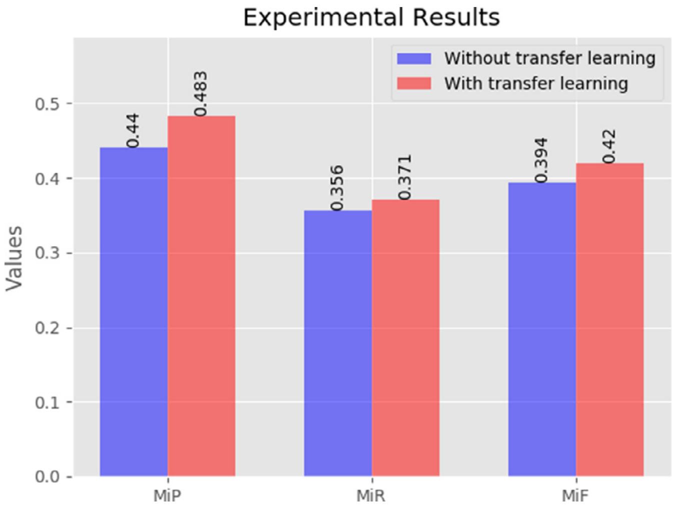


Fig. 5. The performances (Micro-average *F*-measure, Micro-average Precision and Micro-average Recall) of our proposed transfer learning method and the strong baseline deep learning model without transfer learning.

Precision and Micro-average Recall. Fig. 4 shows the performances of our proposed method and the other baseline models on MIMIC-III dataset. It is obvious that Micro-average F-measure predicted by our method significantly outperforms flat-SVM and hierarchy-based SVM. Our model obtained the values of Micro-average F-measure, Micro-average Precision and Micro-average Recall being 0.420, 0.483 and 0.371, respectively, which are better than flat-SVM (0.253, 0.635 and 0.158, respectively) and hierarchy-based SVM (0.335, 0.415 and 0.280, respectively). Although our value of 0.483 on Micro-average Precision is smaller than that of flat-SVM (0.635), the most important evaluation metrics is Micro-average F-measure, which reflects the whole classification performance of the classifier. Even though Micro-average Precision of flat-SVM is higher than that of our proposed method, our proposed method is a better classifier. Experimental results showed that deep transfer learning model is powerful and outperforms the traditional machine learning methods.

4.3. Comparing the performance of deep learning method without transfer learning

The previous experimental comparison used only traditional machine learning methods. In order to explore whether the utilization of transfer learning and deep learning improved the performances of automatic ICD-9 coding, we conduct a study by removing transfer learning component in our network. Specifically, we developed a strong baseline deep learning model without transfer learning to compare with our deep transfer learning method. Compared with the network architecture of our proposed model, the baseline deep learning model just removes transfer learning component. In the comparison, we used all BioASQ3 samples for training parameters in shared layer. Fig. 5 shows the performances (Micro-average *F*-measure, Micro-average Precision and Micro-average Recall) of deep transfer learning method and that of the strong baseline deep learning model without transfer learning. Two definite conclusions can be drawn from Fig. 5. First, deep transfer learning model achieves the state-of-the-art results, which indicates that transfer learning is the key component in improving the performance of automatic ICD-9 coding. Without transfer learning component, Micro-average *F*-measure, Micro-average Precision and Micro-average Recall drop from 0.415, 0.480 and 0.365 (deep transfer learning model) to 0.394, 0.440 and 0.356 (baseline deep learning model), respectively. Second, deep learning approach is the foundation in the success of our proposed model. The strong baseline deep learning model without transfer learning outperforms the non-deep learning baseline model. Without deep learning approach, Micro-average F-measure, Micro-average Precision and Micro-average Recall drop from 0.394, 0.440 and 0.356 (baseline deep learning model) to 0.330, 0.414 and 0.280 (hierarchy-based SVM), respectively. In summary, transfer learning and deep learning components of our model really improves the performances of automatic ICD-9 coding.

4.4. Comparing the performance of multi-scale and sequential network structure

Previous experimental results have shown that transfer learning and deep learning indeed improved the performances of automatic ICD-9 coding. In order to explore the best network architecture in shared layer, we have also compared the performance of multi-scale and sequential CNNS network architectures. Multi-scale and sequential CNNS network architectures have been widely applied to extracting image local features and proven to be effective for a lot of image processing tasks. Inspired by their success in image processing, we compare the performances of the different network architectures in text multi-label classification. In this comparison, we used all BioASQ3 samples for training parameters in shared layers. Fig. 6 shows the performances (Micro-average *F*-measure, Micro-average Precision and Micro-average Recall) of multi-scale CNNs network structure and that of sequential network structure. From Fig. 6, we can observe that multi-scale CNNs structure performs better than sequential network structure. Micro-average F-measure, Micro-average Precision and Micro-average Recall of multi-scale CNNs structure are 0.420, 0.483, and 0.371, respectively. These evaluation metrics of sequential network structure are 0.389, 0.433, and 0.354, respectively. Micro-average F-measure of multi-scale CNNs structure is 8.0% higher than that of sequential network structure. Micro-average Precision of multi-scale CNNs structure is 11.5% higher than that of sequential network structure. Micro-average Recall of multi-scale CNNs structure is 4.8% higher than that of sequential network structure, which indicated that the main improvement is Micro-average Precision. The experimental results showed that multi-scale CNNs structure is better than sequential CNNs structure in text multi-label classification. Therefore, multi-scale CNNs structure may provide more scales and more abundant context features of words that could be used

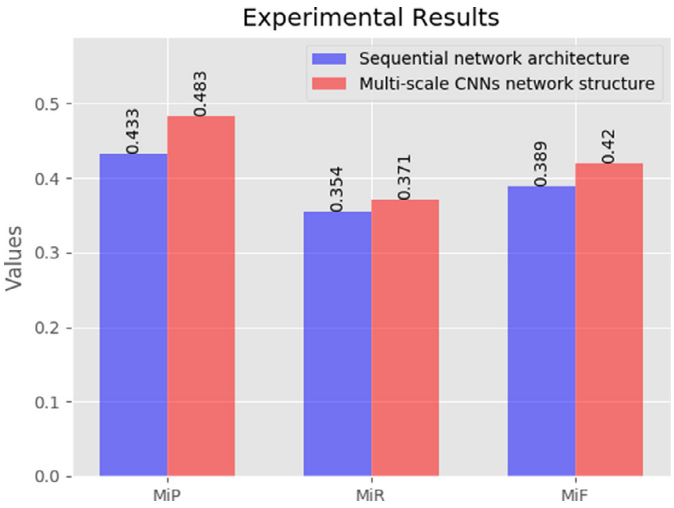


Fig. 6. Performances (Micro-average *F*-measure, Micro-average Precision and Micro-average Recall) of multi-scale CNNs network and sequential network architecture.

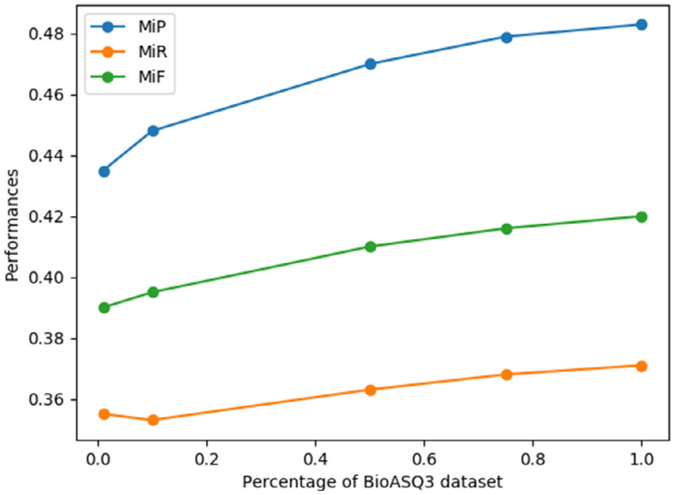


Fig. 7. Performances (Micro-average *F*-measure, Micro-average Precision and Micro-average Recall) of transferring 1%, 10%, 50%, 75% and 100% of samples of BioASQ3 dataset.

for automatics ICD-9 prediction. To some extent, it is similar to the mixture of language N-grams model, thus it can obtain a better performance.

4.5. Effects of different percentage of transferred number of samples on transfer learning

Section 4.3 has shown that transfer learning indeed improves the performance of automatic ICD-9 coding. BioASQ3 dataset is a big medical literature dataset and it has more than ten million samples. In previous experiments, we used all BioASQ3 samples for training parameters in shared layer. Training with such large number of samples takes a lot of time. In order to explore whether utilization of a small amount of samples of BioASQ3 dataset can achieves a good performance, we randomly selected different percentage of samples (1%, 10%, 50%, 75%, 100%) and used for transferring. Fig. 7 shows our experimental results of the performances (Micro-average *F*-measure, Micro-average Precision and Micro-average Recall) of automatic ICD-9 coding by transferring 1%, 10%, 50%, 75% and 100% of samples of BioASQ3 dataset. From Fig. 7, we can see the best performance achieves when transferring 100% samples of BioASQ3 dataset. At the percentage of 100%, Micro-average *F*-measure, Micro-average Precision and Micro-average Recall are 0.420, 0.483 and 0.371, respectively, which are better than those of 75% (0.410, 0.470 and 0.363, respectively), those of 50% (0.410, 0.470 and 0.363, respectively), those of 10% (0.395, 0.448 and 0.353,respectively) and of 1% (0.390, 0.435 and 0.355, respectively). Obviously when transferring 100% samples, our model learned high quality low level shared features such as word-level feature, sentence-level feature, which is helpful to automatic ICD-9 coding. At the percentage of 75%, the evaluation metrics are a little smaller than that of transferring 100% samples of BioASQ3 dataset. At the percentage of 50%, Micro-average *F*-measure is higher than that of using only the strong baseline deep learning model without transfer learning (0.410 versus 0.395), it showed that our proposed model have learned some knowledge that help for classification. When transferring 10% of the dataset, we have noticed that Micro-average *F*-measure is equivalent to that of using the strong baseline deep learning model without transfer learning. This means that 10% of the dataset is not enough to learn something useful features to help improve automatic ICD-9 coding. At the percentage of 1%, Micro-average *F*-measure is worse than that of using only the strong baseline deep learning model without transfer learning (0.390 versus 0.394). This shows that only transferring 1% of the data even decreases the performances of our target domain task. These experimental results suggest that the best performance of automaticICD-9 coding can be obtained at the percentage of 100%.

Table 3
Training time of transferring 1%, 10%, 50%, 75% and 100% of samples of BioASQ3 dataset.

Different percentage of samples used for training (%)	Training time
1	About 17 min
10	About 2.5 h
50	About 12 h
75	About 18 h
100	About 24 h

The entire deep network is trained on a single NVIDIA TITAN X GPU with 12 GB memory. In Table 3 we give the training time of transferring 1%, 10%, 50%, 75% and 100% of samples of BioASQ3 dataset. It takes about 24 h to train our deep network with the original dataset. The training time of transferring 1%, 10%, 50%, 75% is about 17 min, 2.5 h, 12 h, and 18 h, respectively. From the results presented in Table 3, we find that the training time increases significantly with the increase of training samples. The training time is roughly a linear relationship with the number of training samples.

5. Conclusion

In this study, we have proposed a deep transfer learning framework for automatic ICD-9 coding. By making use of a large number of MeSH domain knowledge, our model can significantly improve the performance of automatic ICD-9 coding. Experimental results suggest that our deep transfer learning model achieves state-of-the-art performance, outperforming hierarchy-based SVM and flat-SVM, which shows that our model is very powerful and effective. To understand why our deep transfer learning model works well on automatic ICD-9 coding, we have conducted a study by removing transfer learning component in our network. In particular, we have created a strong baseline deep learning model without transfer learning for comparison. Our experimental results indicate that transfer learning is the key component to improve the performance of automatic ICD-9 coding and deep learning approach is the foundation in the success of our proposed model. Furthermore, we have also compared the performance of multi-scale and sequential network architecture to explore the best network structure in shared layer. The experimental results showed that the better performance obtained by multi-scale CNNs. It has indicated that using multi-scale CNNs can capture text contextual features. We have also studied the effects of transferring different number of samples on transfer learning. Specifically, we use 1%, 10%, 50%, 75% and 100% samples of BioASQ3 dataset to explore the effect of transferring different number of samples. It turns out that transferring 100% samples of BioASQ3 dataset can achieve the best performance while transferring 1% samples even decrease the performance of our target domain task.

Transfer learning is an effective learning technique in target task through the transfer of knowledge from a related task. We conclude that transfer learning not only enhances the power of deep learning approaches, but also breaks the obstacle of insufficient data samples for training target domain task. Therefore, we believe that transfer learning can be generalized to other text classification tasks.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China under Grant nos. 61622213, 61732009 and 61728211.

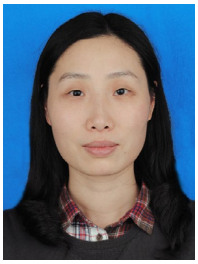
References

[1] P.B. Jensen, L.J. Jensen, S. Brunak, Mining electronic health records: towards better research applications and clinical care, Nat. Rev. Genet. 13 (2012) 395.
[2] Y. Chen, H. Lu, L. Li, Automatic ICD-10 coding algorithm using an improved longest common subsequence based on semantic similarity, Plos ONE 12 (2017) e0173410.
[3] F. Janela, H.M.G. Martins, Using structured EHR data and SVM to support ICD-9-CM coding, in: Proceedings of the IEEE International Conference on Healthcare Informatics, 2013, pp. 511–516.
[4] S. Wang, X. Li, X. Chang, L. Yao, Q.Z. Sheng, G. Long, Learning multiple diagnosis codes for ICU patients with local disease correlation mining, ACM Trans. Knowl. Discov. Data 11 (2017) 31.
[5] Y. Yan, G. Fung, J.G. Dy, R. Rosales, Medical coding classification by leveraging inter-code relationships, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2010, pp. 193–202.
[6] P. Adler, P. Rimma, N. Karthik, W. Nicole, W. Frank, E. Noémie, Diagnosis code assignment: models and evaluation metrics, J. Am. Med. Inf. Assoc. JAMIA 21 (2014) 231.
[7] 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, J. Am. Med. Inf. Assoc. 13 (2006) 516–525.
[8] J. Medori, Machine learning and features selection for semi-automatic ICD-9-CM encoding, in: Proceedings of the NAACL Hlt Second Louhi Workshop on Text and Data Mining of Health Documents, 2010, pp. 84–89.
[9] P. Ruch, J. Gobeilla, I. Tbahritia, A. Geissbühlera, From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding, AMIA Ann. Symp. Proc. 2008 (2008) 636–640.
[10] M. Erraguntla, B. Gopal, S. Ramachandran, R. Mayer, Inference of missing ICD 9 codes using text mining and nearest neighbor techniques, 1060–1069 (2012).
[11] M. Dermouche, J. Velcin, R. Flicoteaux, S. Chevret, N. Taright, Supervised topic models for diagnosis code assignment to discharge summaries, in: Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics, 2016.
[12] A. Perotte, N. Bartlett, F. Wood, Hierarchically supervised latent Dirichlet allocation, in: Proceedings of the International Conference on Neural Information Processing Systems, 2011, pp. 2609–2617.
[13] P. Chen, A. Barrera, C. Rhodes, Semantic analysis of free text and its application on automatically assigning ICD-9-CM codes to patient records, in: Proceedings of the IEEE International Conference on Cognitive Informatics, 2010, pp. 68–74.
[14] I. Goldstein, A. Arzrumtsyan, O Uzuner, Three approaches to automatic assignment of ICD-9-CM codes to radiology reports, in: Proceedings of the AMIA Symposium, 2007, p. 279. AMIA ... Annual Symposium Proceedings.
[15] L. Pereira, R. Rui, C. Silva, M. Agostinho, ICD9-based Text mining approach to children epilepsy classification ☆, Proc. Technol. 9 (2013) 1351–1360.
[16] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (2010) 1345–1359.
[17] R. Petegrosso, S. Park, T.H. Hwang, R Kuang, Transfer learning across ontologies for phenome-genome association prediction, Bioinformatics 33 (2017).
[18] F. Jiang, H. Liu, S. Yu, Y. Xie, Breast mass lesion classification in mammograms by transfer learning, in: Proceedings of the International Conference on Bioinformatics and Computational Biology, 2017, pp. 59–62.
[19] L. Chen, C. Cai, V. Chen, X Lu, Trans-species learning of cellular signaling systems with bimodal deep belief networks, Bioinformatics 31 (2015) 3008–3015.
[20] K. Liu, S. Peng, J. Wu, C. Zhai, H. Mamitsuka, S. Zhu, MeSHLabeler: improving the accuracy of large-scale MeSH indexing by integrating diverse evidence, Bioinformatics 31 (2015) i339.
[21] S. Peng, R. You, H. Wang, C. Zhai, H. Mamitsuka, S. Zhu, DeepMeSH: deep semantic representation for improving large-scale MeSH indexing, Bioinformatics 32 (2016) i70.
[22] A.E.W. Johnson, T.J. Pollard, L. Shen, L.H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L.A. Celi, R.G. Mark, MIMIC-III, a freely accessible critical care database, Sci. Data 3 (2016) 160035.
[23] Y. Kim, Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882, (2014).
[24] N. Kalchbrenner, E. Grefenstette, P. Blunsom, A convolutional neural network for modelling sentences. arXiv preprint arXiv:1404.2188, (2014).
[25] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, E. Hovy, Hierarchical attention networks for document classification, in: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2017, pp. 1480–1489.
[26] S. Lai, L. Xu, K. Liu, J. Zhao, Recurrent convolutional neural networks for text classification, AAAI 333 (2015) 2267–2273.
[27] W. Lotter, G. Sorensen, D. Cox, A multi-scale CNN and curriculum learning strategy for mammogram classification, Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer, Cham, 2017, pp. 169–177.
[28] J. Liu, Y. Pan, M. Li, Z. Chen, L. Tang, C. Lu, J. Wang, Applications of deep learning to MRI images: a survey, Big Data Min. Anal. 1 (1) (2018) 1–18.
[29] A. Roy, S. Todorovic, A multi-scale CNN for affordance segmentation in RGB images, in: European Conference on Computer Vision, Springer, Cham, 2016, pp. 186–201.
[30] E. Gaussier, E. Gaussier, E. Gaussier, G. Paliouras, I. Androutsopoulos, Evaluation measures for hierarchical classification: a unified view and novel approaches, Data Min. Knowl. Discov. 29 (2015) 820–865.
[31] R. Kavuluru, A. Rios, Y. Lu, An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records, Artif. Intell. Med. 65 (2015) 155.
[32] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, TensorFlow: large-scale machine learning on heterogeneous distributed systems, arXiv preprint arXiv:1603.04467, (2016).
[33] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980, (2014).

Min Zeng received the B.S. degree from Lanzhou University in 2013, and the M.S. degree from Central South University in 2016. He is currently working toward the Ph.D. degree in the School of Information Science and Engineering, Central South University, China. His research interests include bioinformatics, machine learning and deep learning.

Min Li received the Ph.D. degree in computer science from Central South University, China, in 2008. She is currently a professor in the School of Information Science and Engineering, Central South University, Changsha, Hunan, China. Her main research interests include bioinformatics and systems biology.

Zhihui Fei received his B.Sc. degrees in Hubei Normal University, China in 2015. He is currently a postgraduate student in Bioinformatics at Central South University. His currently research interests include bioinformatics, medical data mining and deep learning.



Ying Yu received her B.S. degree and M.S. degree from University of South China in 2002 and 2009 separately. She is studying for a Ph.D. in the School of Information Science and Engineering, Central South University, China. Her research focuses on machine learning, deep learning and analysis of healthcare big data.

Yi Pan is a Regents' Professor of Computer Science and an Interim Associate Dean and Chair of Biology at Georgia State University, USA. Dr. Pan joined Georgia State University in 2000 and was promoted to full professor in 2004, named a Distinguished University Professor in 2013 and designated a Regents' Professor (the highest recognition given to a faculty member by the University System of Georgia) in 2015. He served as the Chair of Computer Science Department from 2005 to 2013. He is also a visiting Changjiang Chair Professor at Central South University, China. Dr. Pan received his B.Eng. and M.Eng. degrees in computer engineering from Tsinghua University, China, in 1982 and 1984, respectively, and his Ph.D. degree in computer science from the University of Pittsburgh, USA, in 1991. His profile has been featured as a distinguished alumnus in both Tsinghua Alumni Newsletter and University of Pittsburgh CS Alumni Newsletter. Dr. Pan's research interests include parallel and cloud computing, wireless networks, and bioinformatics. Dr. Pan has published more than330 papers including over 180 SCI journal papers and 60 IEEE/ACM Transactions papers. In addition, he has edited/authored 40 books. His work has been cited more than 6500 times. Dr. Pan has served as an editor-in-chief or editorial board member for 15 journals including 7 IEEE Transactions. He is the recipient of many awards including IEEE Transactions Best Paper Award, 4 other international conference or journal Best Paper Awards, 4 IBM Faculty Awards, 2 JSPS Senior Invitation Fellowships, IEEE BIBE Outstanding Achievement Award, NSF Research Opportunity Award, and AFOSR Summer Faculty Research Fellowship. He has organized many international conferences and delivered keynote speeches at over 50 international conferences around the world.

Jianxin Wang received the B.Eng. and M.Eng. degrees in computer engineering from Central South University, China, in 1992 and 1996, respectively, and the Ph.D. degree in computer science from Central South University, China, in2001. He is the vice dean and a professor in School of Information Science and Engineering, Central South University, Changsha, Hunan, PR China. His current research interests include algorithm analysis and optimization, parameterized algorithm, bioinformatics and computer network. He has published more than 150 papers in various International journals and refereed conferences. He is a senior member of the IEEE.