An External Knowledge Enhanced Multi-label Charge Prediction Approach with Label Number Learning

Duan Wei and Li Lin

School of Computer Science and Technology, Wuhan University of Technology,
Wuhan 430070, China
314585797@qq.com, cathylilin@whut.edu.cn

Abstract. Multi-label charge prediction is a task to predict the corresponding accusations for legal cases, and recently becomes a hot topic. However, current studies use rough methods to deal with the label number. These methods manually set parameters to select label numbers, which has an effect in final prediction quality. We propose an external knowledge enhanced multi-label charge prediction approach that has two phases. One is charge label prediction phase with external knowledge from law provisions, the other one is number learning phase with a number learning network (NLN) designed. Our approach enhanced by external knowledge can automatically adjust the threshold to get label number of law cases. It combines the output probabilities of samples and their corresponding label numbers to get final prediction results. In experiments, our approach is connected to some state-of-the art deep learning models. By testing on the biggest published Chinese law dataset, we find that our approach has improvements on these models. We future conduct experiments on multi-label samples from the dataset. In items of macro-F1, the improvement of baselines with our approach is 3%-5%; In items of micro-F1, the significant improvement of our approach is 5%-15%. The experiment results show the effectiveness our approach for multi-label charge prediction.

Keywords: multi-label charge prediction · label number learning · external knowledge.

1 Introductionn

In recent years, NLP (Nature Language Processing) has a huge development in many research tasks, such as text classification, NER (Named Entity Recognition), semantic labeling, reading comprehending, etc. With the development of internet, the unit of global data total has crossed from GB to ZB. Among them, various valuable text data has greatly promoted the development of the NLP field, and help to alleviate defects of these works.

The legal charge prediction is a multi-label text classification task that learns from law case called fact to predict accusations (labels). This task makes use of the sample's complex text data to properly classify it into the appropriate labels (accusation types). A sample have one label or multiple labels, and there are subtle differences of contents of texts. For example, someone has violated two laws of theft and robbery because of a series of criminal activities. Therefore, we need to understand the description of a case called fact, extract the description of criminals breaking the law, and finally classify the sample into the corresponding accusations. In public Chinese law datasets, like Cail2018, the number of single-label samples is the largest. The number of training data is 154,592, but there is a huge difference between the number of single-label samples and multi-label samples. The number of multi-label samples accounts for a small portion of the total. The more labels the samples have, the less the size of them is in the total amount. For example, the number of 5 labels samples is 58 and the number of 9 labels sample is 1. Therefore multi-label legal charge prediction is a difficult multi-label text classification prediction task.

There are many models that can better solve text classification. For example, Kim purposed the TextCNN [12] that can synthesize local text content of different receptive field to capture the most important features of text; Shi purposed the CRNN [23] that can capture the local information, global information, and interrelated associations of texts; Johnson purposed the DPCNN [10] that can get more efficient and widely available global information while effectively obtaining local information; Vaswani purposed Multi-head Attention that can find out the degree of correlation between the text and the label [25]; And there are GRU and LSTM [4, 9] that can get better capture of longer distance text dependencies. We argue that for the legal charge prediction task, these models will be improved by the external knowledge from law provisions and a suitable way to decide label numbers.

Firstly, all models are only learning the logic between the content of the case and accusations, but they ignore the correlation between the content of the case and law provisions of corresponding accusations. If the model already has such a priori knowledgethe law provisions, it is easy to learn which case is more relevant to the accusations and assist itself learning to understand the content of the case, and finally make the prediction result more accurate.

Secondly, their models only get the output probability of belonging to the corresponding label. However, there is only two widely used strategies to map the probability to output label number as follow:

(1) Top-k strategy: For the output probability of each sample, it selects the first k labels that has the highest probability as the final prediction result;

(2) Threshold strategy: For each probability of the sample, it selects a value as the threshold, and all labels with the probability greater than the threshold are the final prediction result.

However, these two strategies have some disadvantages:

(1) We need to calculate the distribution, maximum value, median of the entire output probability. After getting understanding of the data, we can adjust the parameters to obtain the optimal result. However, it will take a lot of time and efforts to complete this goal;

(2) or most models or datasets, their output probabilities must be different. Therefore, we should select a threshold or k value respectively. It also takes time and efforts;

(3) These two widely used strategies will cause certain errors no matter how the parameters are selected, and the errors will greatly increase for the multi-label task.

Therefore, we propose an external knowledge enhanced multi-label charge prediction approach with automatic label number learning. Meanwhile, we purpose an efficient network structure called number learning network (NLN). For this legal text classification problem, we get external knowledge-corresponding to the law provision of each accusation, and assisting various deep learning models to understand legal texts. And the output probability can be mapped to probability of the label number through NLN. At last, combining them can get the better result than using popular deep learning models alone. Among it, the most important technology of the network is to construct a special embedding layer, that adaptively adjusts the threshold of each corresponding label probability through BP. After implementing this network any deep learning models, it is still an end-to-end model and can effectively solve the above problems..

In this paper, we have an overview of the legal text processing, multi-label text classification, memory mechanism and label number learning in Section 2. We propose an external knowledge enhanced end-to-end multi-label charge prediction approach with automatic label number learning and a number learning network (NLN). This network can be implemented with various deep learning model for the multi-label legal charge prediction. The details of approach will be elaborated in the Section 3. For the experiment results, we use the six deep learning methods with NLN on multi-label samples and our approach on the dataset. This approach has a huge improvement on these models, and the promotion is significant in the multi-label samples. The experiment results show that adding label number learning can also improve the model effect. The more details of experiment will be elaborated in the Section 4. Section 5 makes a conclusion of our work and discuss our future work.

2 Related Work

2.1 Legal Text

As the development of deep learning technology, there are many fields of NLP applications. Because of legal text form and the large amount of data, the law field of NLP becomes a hot topic. Luo adopted the attention based neural network with the supplement of relevant law articles [14,32,33]; Some researchers thought decisions of applicable law articles, charges, fines, and the terms of penalty have logical relationship [38]; Some researchers attempted a hybrid approach to summarize legal case reports [6]; And some researchers compared policy differences by the embedding of entire legal corpus of each institution [19]. More and more researchers paid attentions on the legal text, and exploited lots of deep learning models. [2, 24, 30]. However, for the classification of legal text, the current research is far away from real application.

2.2 Multi-Label Text Classification

In the direction of text classification, so many models have been proposed, and have achieved quite good results that makes an outstanding contribution to the development of this field. Kim purposed the attention-based classifier that can achieve multi-label emotion classification [11, 26]; Yang applied a sequence generation model with a novel decoder structure to solve it with correlations between labels [28, 29, 35]; [1, 15, 22, 25] created various attention mechanisms for NLP that applied into text classification by others; There was a new module that can sheep up training [7, 8]; Zhang achieved text classification with the correlation between different task data [37]; And [3] explored the influence of different semantic embedding to multi-label text classification. And various purposed embedding approaches for representation in document can be applied in multi-label text classification [5, 16, 20, 21, 27]. Most researchers put their energy into complex or novelty deep learning model to solve this problem, however, few researchers refer to the label number learning.

2.3 Memory Mechanism

The existing model has a weak storage capacity for memory, which cant store too much information, and it is easy to lose some semantic information, so the memory network memorizes information by external storage. Facebook AI proposed a memory mechanism to enhance memory of network, and later solved the problem of not being able to end-to-end train [31]. In the next study, it enhanced the size of memory and proposed the new dynamic memory network structure [17,18]. Therefore, we got the external knowledge through this memory mechanism to enhance the classification ability of the model.

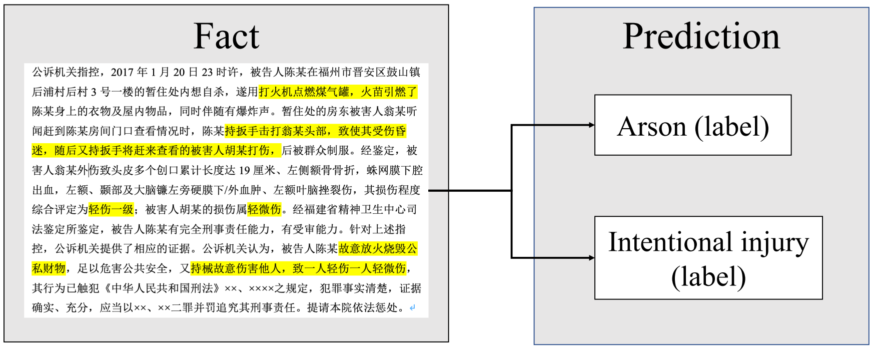
2.4 Label Number Learning

Label number learning is to learning the label number on the basis of predicted probability of deep learning model. There are two common strategies to solve this problem:1. Select an appropriate global threshold. If the predicted probability greater than it, the label will be true label; 2. Directly determine the number of labels of sample, such as top-k. However, there are respectively some defects. Yang purposed the approach that summarizes the whole process of multi-label classification. And she thought there is a need to learning to optimize the threshold over ranked listed on the per-instance basis [36]; Lenc refered to a simple neural network that can solve the problem by top-k [13]. Then these papers give the idea whether exist a neural network that can learn the label number. If it exists, then we can add this network after any deep learning model of the general prediction approach to enhance the performance of the model.

3 Our Approach

3.1 Problem Definition

The charge prediction is a task to predict accusations by analyzing the case description called fact in the Cail2018. In Figure 1 we see that a fact in law is a series of precise descriptions of the case and the presentation of criminal behavior. After segmentation of fact, we define the input as $T_i = \{t_{i1}, t_{i2}, t_{i3}, \ldots, t_{ij}\}$, , and each t_{ij} is a word, such as ability. In Figure 1, we analyze the facts through deep learning model, find out the text of the most relevant accusations (the text part on the yellow background), and then understand their association with each accusation, and finally predict that accusations of sample are respectively arson and intentional injury. We want to get the output is R_i, each R_i is a list, such as $R_i = [1, 55, 120]$. Each number in the list represents the index of the corresponding accusation, such as arson or intentional injury. And we purpose our own approach and NLN on the problem definition.



公诉机关指控，2017 年 1 月 20 日 23 时许，被告人陈某在福州市晋安区鼓山镇后浦村后村 3 号一楼的暂住处内想自杀，遂用打火机点燃煤气罐，火苗引燃了陈某身上的衣物及屋内物品，同时伴随有爆炸声。暂住处的房东被害人翁某听闻赶到陈某房间门口查看情况时，陈某持扳手击打翁某头部，致使其受伤昏迷，随后又持扳手将赶来查看的被害人胡某打伤，后被群众制服。经鉴定，被害人翁某外伤致头皮多个创口累计长度达 19 厘米、左侧额骨骨折，蛛网膜下腔出血，左额、颞部及大脑镰左旁硬膜下/外血肿、左额叶脑挫裂伤，其损伤程度综合评定为轻伤一级；被害人胡某的损伤属轻微伤。经福建省精神卫生中心司法鉴定所鉴定，被告人陈某有完全刑事责任能力，有受审能力。针对上述指控，公诉机关提供了相应的证据。公诉机关认为，被告人陈某故意放火烧毁公私财物，足以危害公共安全，又持械故意伤害他人，致一人轻伤一人轻微伤，其行为已触犯《中华人民共和国刑法》××、××××之规定，犯罪事实清楚，证据确实、充分，应当以××、××二罪并罚追究其刑事责任。提请本院依法惩处。↵

Fig. 1. Example of charge prediction

3.2 The Framework of Our Approach

For multi-label text classification tasks, our purposed prediction approach is: (1) use text preprocessing to obtain text vectors of input data and external knowledge; (2) use deep learning model and attention mechanism to obtain vectors, and combine them to get output results; (3) By machine learning training with output probability of the case sample, the probability of the label number of each sample is obtained; (4) The final result is obtained by label decision that combines the predicted label number of the case sample with output probability of each case sample. This new approach with automatic label number learning is also end-to-end. For the third step of the approach, we can think of it as an auxiliary system. And details of the framework of our approach are shown in Figure 2.

In this approach, we can use any text vectorization methods, whether it is one-hot, word-embedding or direct learning, etc. You can also use any kinds of deep learning models, whether it is TextCNN, Bi-GRU, etc. Therefore, it is easy to experiment for the charge prediction task. We make use of the framework of the memory network to obtain the correlation between text and legal provisions by attention mechanism in training, and combine their outputs with the joint training to get the results. Since both outputs are obtained at the same time, the loss function also needs to be adjusted. See 3.2 for details. Through this approach, adding a label number learning phase after the general prediction approach can effectively solve the multi-label number decision problem. In current research, for the output of the model prediction, both of top-k or threshold strategies need to observe the actual output, and try to manually determine the parameters. However, for the multi-label task, the number of labels is variable, and the probability distributions of different labels is different, so both strategies will produce a huge judgment error. Using this approach, the number of

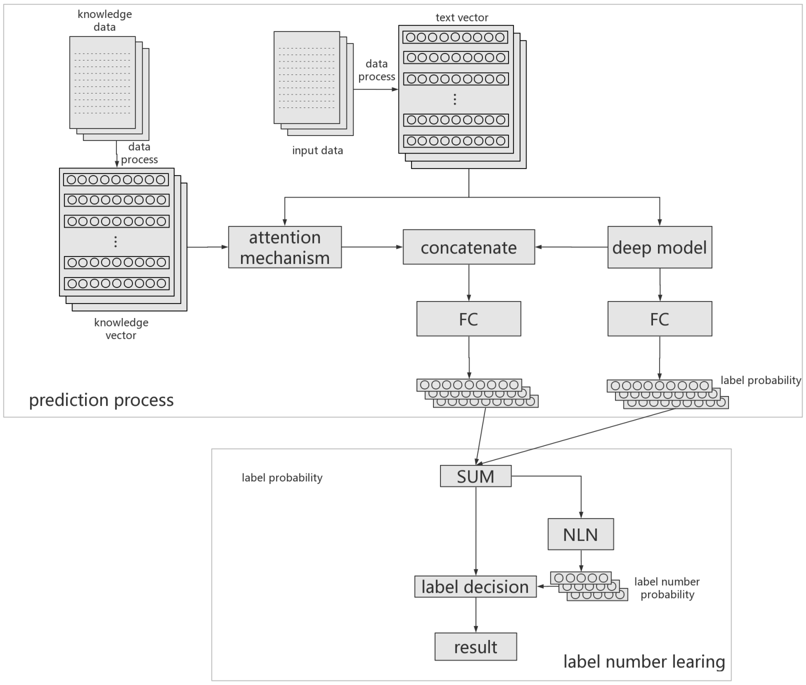


Fig. 2. External knowledge enhanced prediction with label number learning

multi-label charge prediction can be determined adaptively and the results are more accurate.

3.3 Loss Function

Since we hope that the network can learn to understand the logic of the text and it can also make decisions by matching the relevance of the text to the legal provisions, the network has multiple outputs. So we set the loss of these two outputs to $loss_1$ and $loss_2$respectively, and set the $Loss$ of the network to their weighted sum by the custom weight.

$$Loss = w_1 * loss_1 + w_2 * loss_2 \tag{1}$$

w_i is for them its respective weight.

3.4 Number Learning Network (NLN)

This paper adds a label number learning phase after general prediction process phase according to [34], and proposes a network that can better solve the label number problem of charge prediction called number learning network (NLN). Generally, the input of number learning network is $D = \{X_i, Y_i\}_{i=1}^{m}$. For the number learning network, X_i is the label output probability obtained from the previous phrase text prediction, Y_i is the corresponding one-hot encoding of number of labels. The details of X_i and Y_i are illustrated in equation 1,2.

$$X_i = \{x_{i1}, x_{i2}, x_{i3}, \ldots, x_{ij}\}, x_{ij} \in [0, 1] \tag{2}$$

$$Y_i = \{y_{i1}, y_{i2}, y_{i3}, \ldots, y_{ik}\}, x_{ik} = 0, 1 \tag{3}$$

The formula for neural networks is:

$$x_{i+1} = f(\sum_{i=0}^{m-1} w_{ij} x_i + b_j) \tag{4}$$

For the purpose of the experiment that it can adaptively adjust the threshold of the corresponding label, this paper modifies the network and purposes the number learning network (NLN). The most significant difference between the two networks is that the weight value of the first layer is fixed. For example, the first neuron of the first hidden layer is only connected to the first neuron of the second hidden layer. The specific expression is as follows:

$$w_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \tag{5}$$

Secondly, the number of hidden neurons fixed in the first hidden layer is the same as the number of neurons in the input layer. Through these two steps, this article creates a specific embedding process before the output probability is inputted into the FC network. For each x_{ij}, in the embedding process, it represents the probability that the sample belongs to jth accusation. Through this step of the embedding process, if the equality 5 is established, then x_{ij} can pass the filtering.

$$f(x_{ij} + b_j) > 0, \quad i = j \tag{6}$$

b_j is the threshold corresponding to all x_{ij}. In this paper, the back propagation (BP) process can adaptively adjust the b for each label as the corresponding threshold. There is a fine adjustment or screening work through the following several layers of fully connected (FC) network. After the specific embedding process, the FC network is used to map values of each label to the output probability. At the same time, the network can not be too complex to prevent gradient explosions or gradient dispersion. The details of network are illustrated in Figure 3.

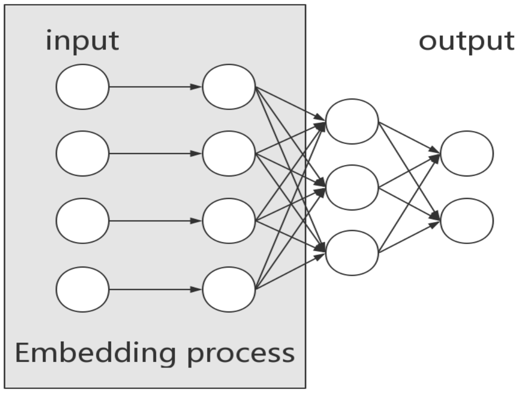


Fig. 3. Number learning network (NLN)

3.5 Label Decision

Finally, we get the label number probability for each sample, and choose the label number with the largest value as the final label number. Then we set the value of the label number is n, and select the top n of the largest value in the corresponding sample output probability as the final output R_i

4 Experiment

In this section, we evaluate our proposed approach on the dataset. We introduce the dataset, evaluation measure, experimental details, and all baseline models. Then, we compare our method with the baselines. Finally, we provide the analysis and discussions of experiment results.

4.1 Dataset and evaluation measure

All legal data are from Cail2018 [34] that contains more than 2.6 million criminal cases published by the Supreme People's Court of China. And it is divided small and big dataset. In this paper, we used the small dataset that consist of 154,592 samples for training, 32,508 samples for testing. Each of the samples contains complex legal text description and three label types. There is respectively accusation, relating article, and term of penalty. This paper only uses the accusation (label) that consists of 202 accusations. And each sample may have multiple labels. The details of small dataset are illustrated in Table 1. The ratio of the training dataset to the test dataset is about 5:1, and the number of labels is 202. Among them, the size of the single-label samples in the training dataset is about 4/5, and the size of multi-label samples is 1/5. And the number of samples that label number is 2 is the majority in multi-label samples.

Table 1. The details of Cail2018 small dataset

Dataset	Training	Valid	Test	Label	
Quantity	154592	17131	32508	202	
Label numer	1	2	3	4	Greater than 4
Quantity in training data	120475	30831	2914	288	96

Because this dataset is provided by Chinese 'fa yan bei' competition[1], we follow the evaluation measure of the competition. We respectively use micro-F1 and macro-F1 to evaluate the ability of different models that are employed on the dataset.

4.2 Baselines

We compare our proposed methods with the following baselines:

TextCNN [12]: It uses a variety of kernels to extract the key information of the text vectors, and then map the information to the low-dimensional space through the FC network to obtain the output probability.

CRNN [23]: It uses CNN before using RNN, and finally output the result through FC network.

[1] http://cail.cipsc.org.cn/

DPCNN [10]: It uses down sampling based on ResNet, get more efficient and widely available global information while effectively obtaining local information.

CNN + attention [25]: While understanding the text information, it finds out the degree of correlation between the text and the label, and maps it to the label space.

Bi-GRU/Bi-LSTM [4,9]: It better captures longer distance text dependencies.

4.3 Experiment Configuration

In experiment, we simply get the word2vec directly after the legal text segmentation, so that the text is mapped into a 512 dimensions vector and the number of words of input data and knowledge respectively is 400, 85. The convolution kernel is generally 3, and for TextCNN, the convolution is respectively [1, 2, 3, 4, 5]. The output dimension of the RNN is 512, and the parameters of the dropout phase are always set to 0.2. Finally, the first layer of the FC network is 1000 dimensions, and the number of dimensions of the second layers output is 202. Both loss functions of deep learning model and NLN are cross entropy while training.

For the multi-label charge prediction task, the number of the multi-label samples is generally less than single-label samples. if the label number of samples is too large, such as 8, the number of samples is less than 0.01% compared to the number of total samples. Therefore the samples are ignored directly by us. Because the number of samples with the number of labels greater than 4 is too small in table 1, we recognized the label number of these samples as 4 in the experiment.

4.4 Experiment on Multi-Label Samples

To prove the effect of our proposed approach, we extract all the samples that the numbers of labels are greater than 1 in the test dataset. We hope to prove the effect of our proposed approach with NLN. The details of multi-label samples are illustrated in Table 2 and the details of experiment are illustrated in Table 3. According to the Table 3, we can clearly see that scores of all the baselines with the number learning network (NLN) are higher in multi-label samples than the ones by threshold strategy. And the improvements of the scores of various baselines are very high.

Table 2. The details of multi-label samples

Label number	2	3	4	Greater than 4
Quantity	1913	135	6	2

With macro-F1 as the index, score of the best model TextCNN is 1.5% higher than the model with threshold strategy. And DPCNN has the most promotion in the baselines, the score of promotion is 5%. Meanwhile, we are surprised to find that with micro-F1 as the index, scores of all the baseline model have great improvements. Score of the best model TextCNN is 5% higher than the model with threshold strategy. The most promotion in the baseline model is DPCNN and the promotion of score is 17%. Therefore we can get a conclusion that by adding number learning network (NLN), we can make these kinds of deep learning models perform better on multi-label charge prediction task.

Table 3. The details of multi-label samples

	Add NLN	microF1(%)	macroF1(%)
TextCNN	with	**82.47**	**63.91**
	without	77.39	62.52
CRNN	with	**78.44**	**57.97**
	without	63.90	54.63
DPCNN	with	**80.20**	**60.80**
	without	63.90	54.63
Attention	with	**81.81**	**62.14**
	without	70.39	57.59
GRU	with	**76.50**	**53.62**
	without	60.48	50.54
LSTM	with	**76.62**	**51.41**
	without	63.14	48.49

4.5 Comparations of Baselines

In this part, we apply baselines to the whole dataset including single label samples and multi-label samples. The details of experiment results are illustrated in the Table 4. In Table 4, the single-label type shows that we only select the label with the highest probability as final prediction result, so we recognize it as single-label task; the multi-label type shows that we have selected all the labels that meet the conditions by threshold strategy as the output. Here we use the appropriate values as the threshold by many manual selections. We find that in all the baselines, using the RNN model to process word2vec vectors that the fact converted is the worst, and using TextCNN model with five different convolution kernels performs best. With micro-F1 as the evaluation index, the best model TextCNN scores 10% higher than the worst model Bi-GRU. With macro-F1 as the evaluation index, the best model TextCNN scores 15% higher than the worst model Bi-GRU.

Although the data is extremely imbalance, we can still find that after the selected threshold determining results, the F1 scores have improved, but the effect is not satisfactory. On this dataset, the difference between the multi-label results with all baselines and the single-label results is not large. Based on the evaluation measure, their improvement is less than 1% on average. Therefore we can make a judgment that the widely threshold strategy has defects in accuracy.

Table 4. Results of baselines

	type	microF1(%)	macroF1(%)
TextCNN	single-label	84.59	74.54
	multi-label	**85.95**	**76.18**
CRNN	single-label	77.67	61.19
	multi-label	78.04	61.58
DPCNN	single-label	80.55	65.03
	multi-label	80.9	65.49
Attention	single-label	82.69	71.36
	multi-label	83.44	72.10
GRU	single-label	75.48	58.16
	multi-label	75.76	58.44
LSTM	single-label	77.54	60.17
	multi-label	77.96	62.20

4.6 Results of Our Approach

In this part, we use our purposed approach with all the baseline models. The details of experiment are illustrated in Table 5. In Figure 5, we find that our approach has some improvements on various baselines. According to the Table 5, with micro-F1 as the index, we can observe that the results using our approach have a better performance. For all models with our approach the scores increase, and some models have 5%-8% improvements; With macro-F1 as the index, we can also observe that all the models have great improvements, and some models have 9%-13% improvements. Scores of most models reach above 70%. Because of the definition of macro-F1, we can also make a conclusion: we can correctly classify more multi-label samples with our approach. And results prove that our approach can solve this task better.

Table 5. Results of baselines with our approach

	Our Approach	microF1(%)	macroF1(%)
TextCNN	True	**86.23**	**77.50**
	Flase	85.95	76.18
CRNN	True	**84.58**	**71.83**
	Flase	78.04	61.58
DPCNN	True	**85.74**	**76.06**
	Flase	80.9	65.49
Attention	True	**83.50**	**74.52**
	Flase	83.44	72.10
GRU	True	**83.69**	**71.63**
	Flase	75.76	58.44
LSTM	True	**78.96**	**61.85**
	Flase	77.98	60.66

5 Conclusions and Future Work

n this paper, we propose an external knowledge enhanced end-to-end multi-label charge prediction approach with automatic label number learning and a number learning network (NLN). By using it, we prove that the approach can improve the performance of the deep learning model in the multi-label charge prediction task. This method has great improvement on all models. At the same time, we extract all the multi-label samples from the test dataset and test the approach on it. The experiment results show that our approach is better than the deep learning model with the threshold strategy. With macro-F1 as the index, the improvement of various deep learning models is 3%-5%; With micro-F1 as the index, it has increased 5%-15%.

In the future, we will connect deep learning models with number learning network (NLN) while training, rather than getting the output probability of sample first, and create a new loss function that is used in training. We hope that can improve accuracy of the charge prediction and reduce the training loss error. In particularly, we want to avoid predicting the single-label samples into multi-label category to increase the score of F1. At the same time, we will test the performance of NLN on other multi-label tasks, and hope that further study can improve its versatility.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. CoRR **abs/1409.0473** (2014)
2. Bajwa, I.S., Karim, F., Naeem, M.A., ul Amin, R.: A semi supervised approach for catchphrase classification in legal text documents. JCP **12**(5), 451–461 (2017)
3. Berger, M.J.: Large scale multi-label text classification with semantic word vectors. Tech. rep., Technical Report. Stanford University (2015)
4. Cho, K., van Merrienboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL. pp. 1724–1734 (2014)
5. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR **abs/1810.04805** (2018)
6. Galgani, F., Compton, P., Hoffmann, A.: Combining different summarization techniques for legal text. In: Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data. pp. 115–123. Association for Computational Linguistics (2012)
7. Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, Y.N.: Convolutional sequence to sequence learning. In: ICML 2017, Sydney, NSW, Australia, 6-11 August 2017. pp. 1243–1252 (2017)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 770–778 (2016)
9. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997)

10. Johnson, R., Zhang, T.: Deep pyramid convolutional neural networks for text categorization. In: ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers. pp. 562–570 (2017)
11. Kim, Y., Lee, H., Jung, K.: Attnconvnet at semeval-2018 task 1: Attention-based convolutional neural networks for multi-label emotion classification. In: NAACL-HLT, New Orleans, Louisiana, June 5-6, 2018. pp. 141–145 (2018)
12. Kim, Y.: Convolutional neural networks for sentence classification. In: EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL. pp. 1746–1751 (2014)
13. Lenc, L., Král, P.: Word embeddings for multi-label document classification. In: RANLP 2017, Varna, Bulgaria, September 2 - 8, 2017. pp. 431–437 (2017)
14. Luo, B., Feng, Y., Xu, J., Zhang, X., Zhao, D.: Learning to predict charges for criminal cases with legal basis. In: EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017. pp. 2727–2736 (2017)
15. Luong, T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. In: EMNLP 2015, Lisbon, Portugal, September 17-21, 2015. pp. 1412–1421 (2015)
16. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR **abs/1301.3781** (2013)
17. Miller, A.H., Fisch, A., Dodge, J., Karimi, A., Bordes, A., Weston, J.: Key-value memory networks for directly reading documents. In: EMNLP 2016, Austin, Texas, USA, November 1-4, 2016. pp. 1400–1409 (2016)
18. Miller, A.H., Fisch, A., Dodge, J., Karimi, A., Bordes, A., Weston, J.: Key-value memory networks for directly reading documents. In: EMNLP 2016, Austin, Texas, USA, November 1-4, 2016. pp. 1400–1409 (2016)
19. Nay, J.J.: Gov2vec: Learning distributed representations of institutions and their legal text. In: EMNLP 2016, Austin, TX, USA, November 5, 2016. pp. 49–54 (2016)
20. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL. pp. 1532–1543 (2014)
21. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers). pp. 2227–2237 (2018)
22. Shaw, P., Uszkoreit, J., Vaswani, A.: Self-attention with relative position representations. In: NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers). pp. 464–468 (2018)
23. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Trans. Pattern Anal. Mach. Intell. **39**(11), 2298–2304 (2017)
24. Sulea, O., Zampieri, M., Malmasi, S., Vela, M., Dinu, L.P., van Genabith, J.: Exploring the use of text classification in the legal domain. In: ICAIL 2017, London, UK, June 16, 2017. (2017)
25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA. pp. 6000–6010 (2017)
26. Wang, Y., Lin, X., Wu, L., Zhang, W., Zhang, Q., Huang, X.: Robust subspace clustering for multi-view data by exploiting correlation consensus. IEEE Trans. Image Processing **24**(11), 3939–3949 (2015)

27. Wang, Y., Wu, L.: Beyond low-rank representations: Orthogonal clustering basis reconstruction with optimized graph structure for multi-view spectral clustering. Neural Networks **103**, 1–8 (2018)
28. Wang, Y., Wu, L., Lin, X., Gao, J.: Multiview spectral clustering via structured low-rank matrix factorization. IEEE Trans. Neural Netw. Learning Syst. **29**(10), 4833–4843 (2018)
29. Wang, Y., Zhang, W., Wu, L., Lin, X., Fang, M., Pan, S.: Iterative views agreement: An iterative low-rank based structured optimization method to multi-view spectral clustering. In: IJCAI 2016, New York, NY, USA, 9-15 July 2016. pp. 2153–2159 (2016)
30. Wei, F., Qin, H., Ye, S., Zhao, H.: Empirical study of deep learning for text classification in legal document review. In: Big Data 2018, Seattle, WA, USA, December 10-13, 2018. pp. 3317–3320 (2018)
31. Weston, J., Chopra, S., Bordes, A.: Memory networks. In: ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015)
32. Wu, L., Wang, Y., Gao, J., Li, X.: Where-and-when to look: Deep siamese attention networks for video-based person re-identification. IEEE Trans. Multimedia **21**(6), 1412–1424 (2019)
33. Wu, L., Wang, Y., Li, X., Gao, J.: Deep attention-based spatially recursive networks for fine-grained visual recognition. IEEE Trans. Cybernetics **49**(5), 1791–1802 (2019)
34. Xiao, C., Zhong, H., Guo, Z., Tu, C., Liu, Z., Sun, M., Feng, Y., Han, X., Hu, Z., Wang, H., Xu, J.: CAIL2018: A large-scale legal dataset for judgment prediction. CoRR **abs/1807.02478** (2018)
35. Yang, P., Sun, X., Li, W., Ma, S., Wu, W., Wang, H.: SGM: sequence generation model for multi-label classification. In: COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018. pp. 3915–3926 (2018)
36. Yang, Y., Gopal, S.: Multilabel classification with meta-level features in a learning-to-rank framework. Machine Learning **88**(1-2), 47–68 (2012)
37. Zhang, H., Xiao, L., Wang, Y., Jin, Y.: A generalized recurrent neural architecture for text classification with multi-task learning. In: IJCAI 2017, Melbourne, Australia, August 19-25, 2017. pp. 3385–3391 (2017)
38. Zhong, H., Guo, Z., Tu, C., Xiao, C., Liu, Z., Sun, M.: Legal judgment prediction via topological learning. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018. pp. 3540–3549 (2018)