International Journal of Computational Intelligence Systems (2021) 14:204
https://doi.org/10.1007/s44196-021-00053-6

RESEARCH ARTICLE

A Classification Model of Legal Consulting Questions Based on Multi-Attention Prototypical Networks

Jianzhou Feng[1] · Jinman Cui[1] · Qikai Wei[1] · Zhengji Zhou[2] · Yuxiong Wang[3]

Received: 22 May 2021 / Accepted: 29 November 2021

Abstract
Text classification is a research hotspot in the field of natural language processing. Existing text classification models based on supervised learning, especially deep learning models, have made great progress on public datasets. But most of these methods rely on a large amount of training data, and these datasets coverage is limited. In the legal intelligent question-answering system, accurate classification of legal consulting questions is a necessary prerequisite for the realization of intelligent question answering. However, due to lack of sufficient annotation data and the cost of labeling is high, which lead to the poor effect of traditional supervised learning methods under sparse labeling. In response to the above problems, we construct a few-shot legal consulting questions dataset, and propose a prototypical networks model based on multi-attention. For the same category of instances, this model first highlights the key features in the instances as much as possible through instance-dimension level attention. Then it realizes the classification of legal consulting questions by prototypical networks. Experimental results show that our model achieves state-of-the-art results compared with baseline models. The code and dataset are released on https://github.com/cjm0824/MAPN.

Keywords Legal consulting questions classification · Few-shot learning · Prototypical networks · Instance-dimension level attention

1 Introduction

Text classification is a common task in the field of Natural Language Processing (NLP), which is widely used in news classification, information retrieval and machine reading comprehension, etc. In the field of legal intelligent question-answering, accurately judging the legal category based on the facts described by users is also a text classification task, and a necessary prerequisite to realize intelligent question answering.

With the development of deep learning, researchers begin to apply deep learning models to achieve text classification. Kim [1] proposed to use Convolutional Neural Networks (CNN) to realize text classification, and achieved results that surpass previous supervision models. Later, Du et al. [2] and Huang et al. [3] used Recurrent Neural Network (RNN) and Graph Neural Networks (GNN) for text classification tasks, respectively. Deep learning models have gradually become the mainstream methods of text classification because they do not require manual feature extraction and the classification effect are significant. However, deep learning models need the support of large-scale labeled datasets. When there is less annotated data, they are easy to show the phenomenon of overfit [4]. Especially in specific application fields, such as medical and financial, annotators are required to have a high professional level, which leads to a high cost of annotation. So the development of deep learning is limited in these fields. In the case of sparse annotation, it has become a problem for researchers to achieve efficient text classification.

✉ Jinman Cui
cuijinman@163.com

Jianzhou Feng
fjzwxh@ysu.edu.cn

Qikai Wei
mrweiqk@163.com

[1] School of information Science and Engineering, Yanshan University, Qinhuangdao 066004, China

[2] Shengming Jizhi (Beijing) Technology Co., Ltd, Beijing 100000, China

[3] Chongqing NewGo AI Co., Ltd, Chongqing 401121, China

Published online: 01 January 2022

Currently, researchers have begun to study methods based on Few-Shot Learning [5] (FSL), such as model-agnostic meta-learning [6], relation networks [7], matching networks [8], etc. But these methods are mostly applied in computer vision, and are rarely found in NLP. Recently, Gao et al. [9] proposed hybrid attention-based prototypical networks for few-shot relation classification, and successfully applied FSL in text classification. However, in the legal intelligent question-answering system, due to the short length of legal consulting texts, serious oral language, the lack of applicable datasets and so on, the classification encounters great obstacles. For these reasons, we construct a few-shot dataset for lagal consulting questions classification, and propose a classification model based on multi-attention prototypical networks. Our contributions are summarized as follows:

- We construct a few-shot dataset for legal consulting questions. This dataset contains 46 categories and each category contains 30 to 300 instances.
- In view of the particularity of the dataset constructed in this paper, we propose a multi-attention prototypical networks model based on instance-dimension level attention.
- Experimental results on our dataset show that our model achieves state-of-the-art effects. We also test our model on other few-shot datasets. The experimental results show that our model has generalization ability.

2 Related Works

Currently, with the development of deep learning, researchers began to apply it to text classification and achieved good results. Kim [1] proposed to use CNN for text classification. First, he utilized pre-trained word vectors to represent text information as input, and then used the labeled data to train to achieve text classification. The result of classification proves that CNN is not only suitable for image classification tasks, but also for text classification tasks. Due to the sequential nature of the text information and the problem of long-term dependence, Zhou et al. [10] proposed to apply Long Shot-Term Memory (LSTM) to text classification, which solved the problem that CNN cannot model contextual information. Whether it is CNN or LSTM, the sentence is used as the input of model, the sentence features are extracted through the deep learning models, and the classifier is used to achieve classification. But these models ignore the information relevance between sentences. Yang et al. [11] proposed a hierarchical attention mechanism model for document classification, including word-level and sentence-level attention, so that the model has the ability to assign different attention to each word and each sentence in one document.

In the application field where annotated data is scarce, the existing deep learning models cannot play a very good role. In response to this problem, the few-shot learning methods have gradually attracted researchers' attention and began to be applied in the field of computer vision [4, 8, 12–14]. Gao et al. [9] believed that text is different from images, and there are problems such as information diversity and large noise. Therefore, they proposed a prototypical networks model based on the hybrid attention mechanism to achieve relation classification. This model designed instance-level and feature-level attention to alleviate the influence of noisy data and sparse features. Sun et al. [15] proposed a hierarchical attention prototypical networks for few-shot text classification. They designed the feature level, word level, and instance-level multi cross attention for the model to enhance the expressive ability of semantic space. Geng et al. [16] proposed an induction network for few-shot text classification. They used the dynamic routing algorithm in meta-learning to learn the generalized class-wise representation so that the model can be better generalized and promoted. Subsequently, Geng et al. [17] proposed dynamic memory induction networks for few-shot text classification. The model utilizes dynamic routing to provide more flexibility to memory-based FSL in order to better adapt the support sets, which is a critical capacity of few-shot classification models. Bao et al. [18] proposed a meta-learning model for few-shot text classification. They used an attention generator and a ridge regressor to make the model perform well in cross-category transfer.

3 Prototypical Networks Based on Multi-attention

3.1 Basic Concepts for Few-Shot Learning

The purpose of FSL is to generalize and analogize limited prior knowledge, and extend it to new tasks. It is used to solve the problem of how to improve the performance of the model in the case of less training data. FSL contains the following concepts:

Support set: N categories are randomly selected from the dataset at each iteration, and each category contains K instances, thus forming the support set containing $N \times K$ instances. Generally speaking, at each iteration, the model will be trained once on support set to learn features.

Query set: the query set is similar to support set. Q instances from each of the remaining instances of the N classes will be selected to form query set. During training, the model will calculate the loss on query set and update parameters after once training on support set.

N-way K-shot: N-way refers to N categories, and K-shot means that each category contains K instances. N-way

K-shot task refers to the model learns features from support set, so that it knows how to distinguish these N categories.

Training set and test set: the training set and test set in FSL task are consistent in data composition. But there is no category overlap between two datasets, and they both contain support set and query set. In the FSL task, to adapt to new tasks quickly, the model first trains parameters on training set, and then the support set in test is used to adjust the parameters. Finally, the model tests the performances on query set.

3.2 Model Framework

In this paper, we propose a Legal consulting questions Classification model based on Multi-Attention Prototypical Networks (MAPN-LC). This model uses prototypical networks as a basic framework, and realizes classification by calculating the distance between consulting questions and each class prototype. In the classification process, we introduce the instance-dimension level attention mechanism. Firstly, we assign different weights to each instance in one category, and then assign different weights to each dimension of weighted instances' feature vectors, thereby improving the contribution of key instances and key features.

We define the legal consulting questions classification as predicting the category of query instance q through a given support set S. Among them, the support set S is defined as follows:

$$S = \left\{ \left(s_1^1, s_1^2, \ldots, s_1^K\right), \ldots, \left(s_N^1, s_N^2, \ldots, s_N^K\right) \right\}, \tag{1}$$

where s_i^j denotes the j-th instance of the i-th category, and S is expressed in the form of N-way K-shot.

The framework of the model is shown in Fig. 1. For each instance in support set S, each word in the instance is expressed as a word vector through the word embedding module, and then the feature vector of instance is obtained through the feature extraction module. Then, the instance-dimension level attention is used on instances' feature vectors to obtain the weight vector β_i in the dimension. For query instance q, feature vector x_q is obtained through the word embedding module and feature extraction module. Finally, the class prototype c_i of each class is calculated through prototypical networks, and the distance function is generated by combining the weight vector β_i to predict the legal category of query instance q.

3.3 Instance Encoder

3.3.1 Embedding Layer

Since the text cannot be directly involved in the calculation, we use the pre-training word vectors to map each word in the instance into the form of word vector. Each word w_i in the instance $s = (w_1, w_2, \ldots, w_n)$ is expressed as a d_k dimensional vector e_i. Then the embedding vector $s' = (e_1, e_2, \ldots, e_n)$ of instance s is obtained, where $s' \in \mathbb{R}^{n \times d_k}$, and n denotes the maximum length of instance.

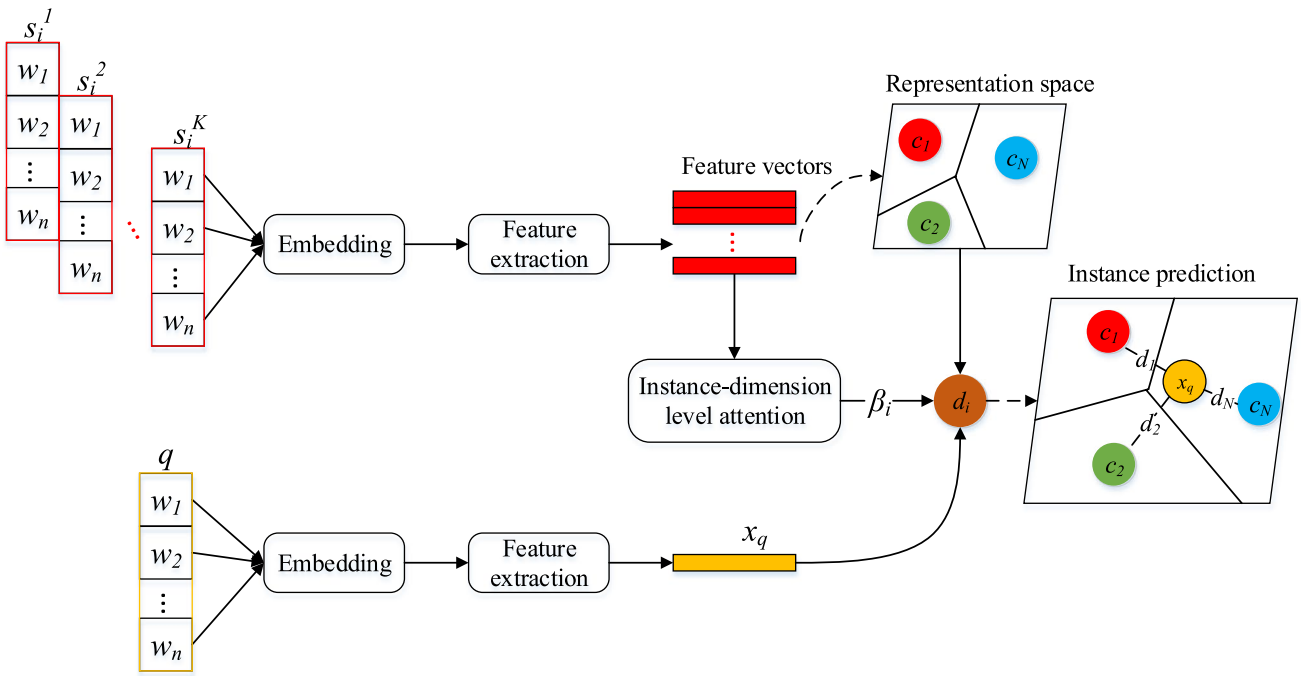

Fig. 1 Prototypical networks based on multi-attention

3.3.2 Feature Extraction Layer

Since the input of the model is a sentence, there is a strong correlation between each word in the sentence, so we can generate sentence feature vectors by extracting the semantic information of all words, and better express the complete semantic information of the sentence. In this paper, we perform CNN on the dimension of the word embeddings to extract the completed semantic information to obtain the corresponding sentence embedding, so as to better represent the sentence features. The feature extraction module is shown in Fig. 2. First, we use h convolutional kernels with $t \times d_k$ $(t < n)$ dimentional to perform on $n \times d_k$ dimensional embedding vector s'. And h local feature vectors with dimensions $m \times 1$ can be obtained. Then we use the max-pooling operations with the window size of $m \times 1$ to extract the maximum feature of each convolution kernel. Finally, we can obtain a $1 \times h$ dimensional feature vector, and instance s' is represented as a feature vector $x \in \mathbb{R}^{1 \times h}$.

3.4 Prototypical Networks

The main idea of prototypical networks is to use the class prototype vectors to represent the instances of each category. We represent the class prototype of each category by calculating the average value of its instances' feature vectors. The class prototype vector is calculated by Eq. (2):

$$c_i = \frac{1}{K} \sum_{j=1}^{K} x_i^j, \tag{2}$$

where x_i^j denotes the j-th instance's feature vector of the i-th category, K denotes instances' number of the i-th category in support set, and c_i denotes the class prototype of the i-th category.

3.4.1 Instance-Dimension Level Attention

Our dataset has a certain particularity, the length of the instances in our dataset is shorter and the instances of the same category have certain similarities. It can be seen from Table 1, although the three instances have different

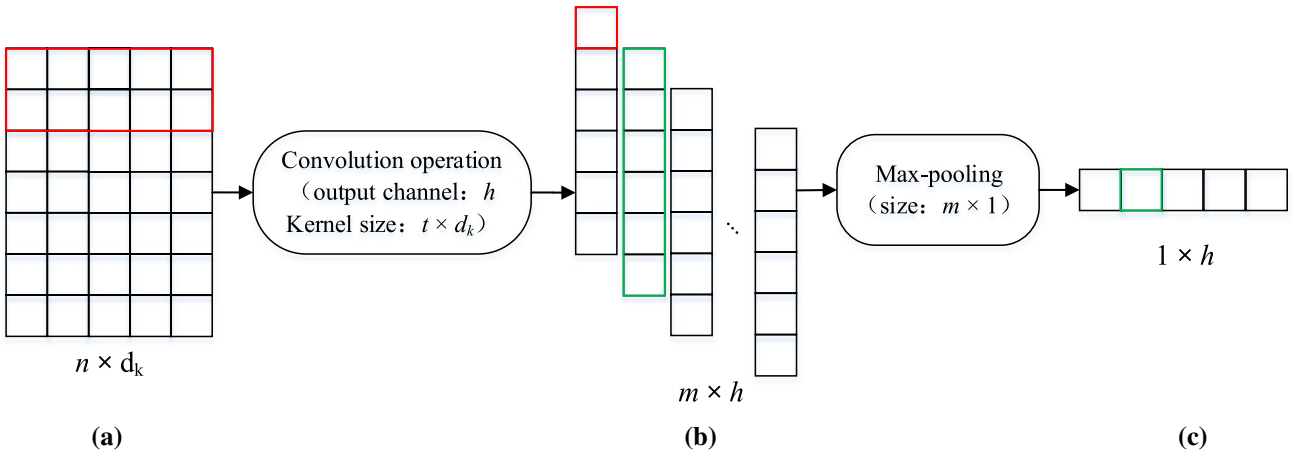

Fig. 2 Feature extraction module. The red rectangle in (**a**) represents a convolution kernel, and the red rectangle in (**b**) represents the result of a convolution operation performed by the convolution kernel on the sentence embedding vector; the green rectangle in (**b**) represents the filter size of the max-pooling operation, and the green rectangle in (**c**) represents the result of performing the max-pooling operation on a local feature vector

Table 1 Three instances in "*Joining rights protection*" randomly selected from the legal consultation question dataset

Number	Instance	Text length
1	我加盟了奶茶店, 但是还没开业, 可不可以退款? (I joined a milk tea shop, but it hasn't opened yet. Can I get a refund?)	13
2	你好, 我这边加盟了餐饮店, 现在没开成, 能退吗? (Hello, I joined a restaurant, but it hasn't opened now. Can I get a refund?)	16
3	交了加盟费, 不想开店了, 能退回来吗? (I have paid the franchise fee and don't want to open a shop. Can I get it back?)	10

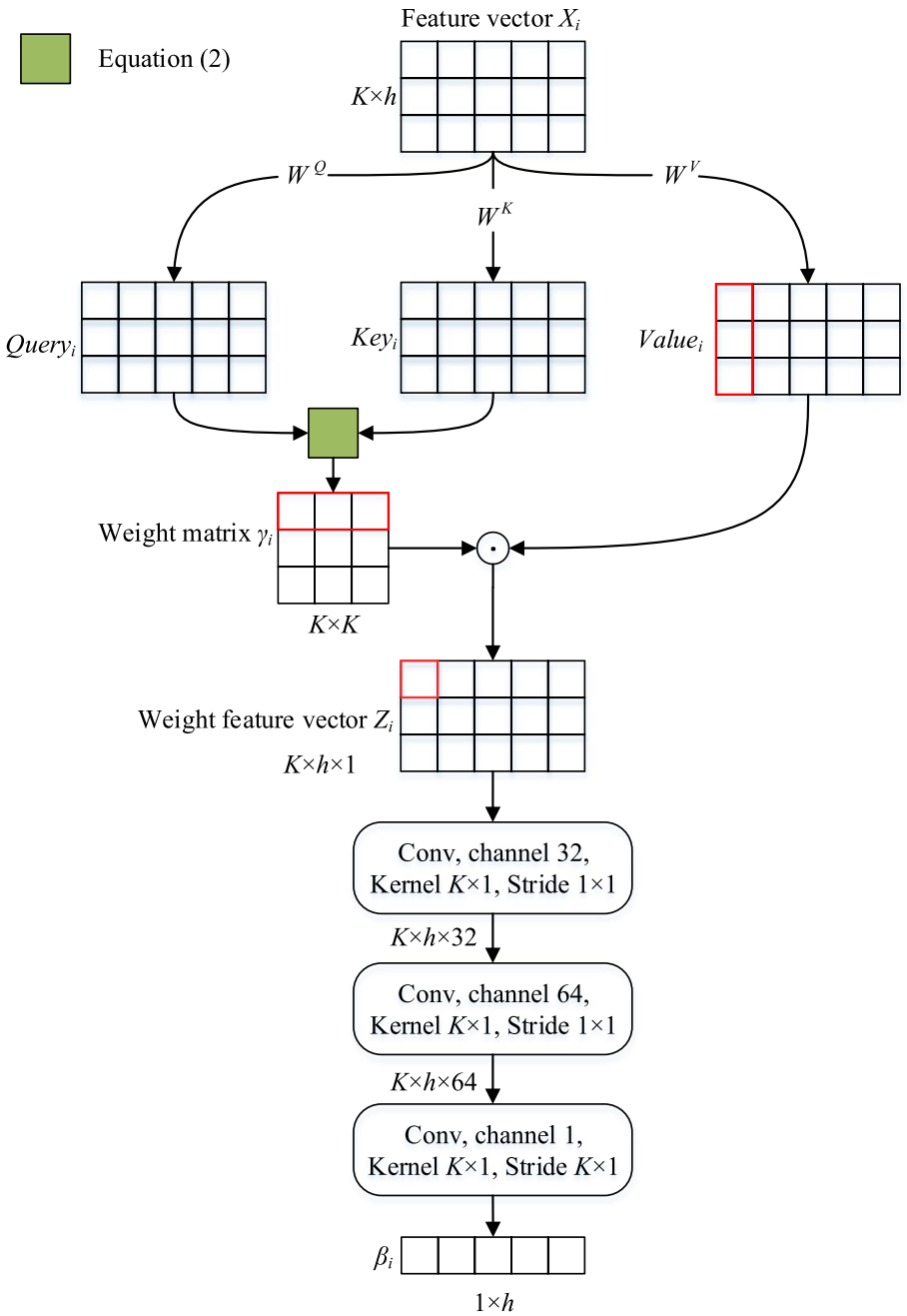

Fig. 3 Instance-dimension level attention module

expressions, they have similar semantics. Besides, there are fewer instances in the support set in FSL, and the features extracted from the support set have a datasparse problem. We introduce the instance-dimension level attention composed of instance-level and dimensional-level attention. It is shown in Fig. 3.

The instance-level attention draws on the idea of the self-attention mechanism [19], and performs cross-attention between K instances in the support set to assign different weights to each instance. If one instance has a higher similarity with other instances, it means that this instance is more representative of its category and should be given a higher weight. Thus, a certain interdependence relationship is established between K instances in the same category.

The instance-level attention module first multiplies the instances' feature vector set X_i of a certain category in support set with its projection matrices W^Q, W^K and W^V, respectively. Then through the linear transformation we can obtain three vectors, Query_i, Key_i and Value_i. Finally, we can get the self-attention score of each instance. That is, one instance can more represent its category, the higher corresponding score. And the weight matrix γ_i composed of similarity scores between each instance and other instances is obtained. Where $X_i = \{x_i^1, x_i^2, \ldots, x_i^K\} \in \mathbb{R}^{K\times h}$, i denotes

the i-th category, and x_i^j denotes the j-th instance's feature vector of the i-th category. $\text{Query}_i, \text{Key}_i, \text{Value}_i \in \mathbb{R}^{K\times h}$, and the weight matrix $\gamma_i \in \mathbb{R}^{K\times K}$ is shown as follows:

$$\gamma_i = \text{softmax}\left(\frac{\text{Query}_i \cdot \text{Key}_i^T}{\sqrt{h}}\right). \tag{3}$$

Then, as shown in Eq. (4), the weight matrix γ_i and the vector Value_i are element-multiplied to obtain the weighted feature vector set $Z_i = \{z_i^1, z_i^2, \ldots, z_i^K\} \in \mathbb{R}^{K\times h}$, where z_i^j denotes the j-th instance's weighted feature vector of the i-th category.

$$Z_i = \gamma_i \cdot \text{Value}_i. \tag{4}$$

When classifying special categories in the feature space, certain dimensions have stronger distinguishing abilities. Therefore after obtaining weighted feature vectors with a certain interdependence relationship, we introduce dimension-level attention to increase the weight of these dimensions. We apply a CNN-based feature attention mechanism according to Gao et al., to achieve dimension-level attention. Finally, through the convolution operations on each dimension of the weighted feature vector set Z_i, the weight vector β_i on the feature dimension is obtained.

3.4.2 Instance Prediction

To predict the category of query instance q, we calculate the distance from x_q to each class prototype c_i. The most commonly used distance function is Euclidean distance. For prototypical networks based on multi-attention proposed in this paper, the dimension-level weighted feature vector β_i of the support set can be obtained through the instance-dimension level attention. We use the distance function constructed based on the weight vector β_i to realize the label prediction of the instances [9]. It can make the model better adapt to the given categories and instances, and better alleviate the impact of feature sparse on model performance. The distance function is shown in Eq. (5):

$$d_i = \beta_i\left(c_i - x_q\right)^2. \tag{5}$$

Finally, we use the cross-entropy loss function to evaluate the gap between the predicted category and actual category of instance q, and use the stochastic gradient descent algorithm to adjust the parameters.

4 Experiments

To verify that the model proposed in this paper can better realize the legal consulting questions classification, we construct a few-shot consulting questions classification dataset in the legal field, and compare it with other FSL models.

4.1 Datasets

We evaluate our approach on legal consulting questions classification dataset, amazon product dataset, HuffPost headlines dataset, and FewRel.

Legal consulting questions dataset are obtained from real legal intelligent question-answering websites. This dataset contains 46 categories, such as *medical disputes*, *insurance claims* and *bond mortgages*, with a total of 10,402 legal consulting questions. Each category contains 30–300 instances with different text lengths, and the maximum text length is 40. According to the definition of FSL, we divide the dataset into a training set and a test set, and there is no overlap between the categories of two sets. The statistical information of the dataset is shown in Table 2.

Amazon product dataset contains customer reviews from 24 product categories. Since the original dataset is too large, we generate a subset by sampling 1000 reviews from each category and split it according to Bao et al. [18]

HuffPost headlines dataset [20] consists of news headlines published on HuffPost between 2012 and 2018. These headlines split among 41 classes.

FewRel [21] is a relation classification dataset developed for FSL. Each instance is a sentence, annotated with a head entity, a tail entity and their relation. We need to classify the relation between the head and tail entities based on the semantic information of the sentence.

4.2 Baselines

In this section, the baseline models in our experiments are introduced as follows:

- SNAIL [13] is a meta-learning model, which uses temporal convolutional neural networks and attention modules

Table 2 Dataset of legal consulting questions

Name	Traing set	Test set
Number of categories	36	10
Number of instances	8047	2355
Number of instances per category	30~300	30~300

to integrate information from past experience to achieve rapid learning.

- GNN [14] is proposed to embed instances of support set and correspond label information into graph nodes, and spread information between nodes. The query set instances receive information from support set, so as to realize few-shot classification.
- Siamese neural networks [12] is proposed to accomplish image classification. During training, this model combines different samples into pairs and inputs them into siamese network to extract the features, and then calculates the distance between the features to realize classification.
- Prototypical networks [4] is applied to a few-shot image classification task. This model maps the samples to a metric space. For each class of samples, it calculates the mean of feature vectors to represent the class prototype. When predicting an unknown sample, the Euclidean distance is used to calculate the distance between the sample and each class prototype to predict the label of the sample.
- Proto-HATT [9] is a hybrid attention-based prototypical networks for noisy few-shot relation classification.

4.3 Parameter Settings

In this experiment, Google's open-source toolkit word2vec is used to obtain word embedding vectors by training the dataset constructed in this paper. The hyper-parameters are determined by grid search algorithm, including CNN encoding window size $t \in \{2, 3, 4, 5\}$, batch size batch $\in \{3, 4, 5, 6\}$, and learning rate size lr$\in \{0.001, 0.01, 0.1\}$. In addition, we performed 30,000 iterations to train the model. The parameters used by the model are shown in Table 3.

4.4 Experimental Results

4.4.1 Overall Performance

To verify the superiority of the model proposed in this paper, we have compared it with the current mainstream FSL models.

Table 3 Parameter settings

Parameter	Value
Word vector's dimension, d_k	300
Maximum sentence length, n	40
Sentence's feature vectors dimension, d_h	230
CNN encoding window size, t	3
Learning rate, lr	0.001
Batch size, batch	4

As shown in Table 4: (1) In the four cases, the accuracy of MAPN-LC is higher than that of the baseline models, and it is 2–3% higher than Proto-HATT which works best among the comparison models. (2) From the experimental results of MAPN-LC in the four cases, the accuracy is the highest in 5-way 10-shot, and the accuracy is the lowest in 10-way 5-shot. It explains that with the same number of categories, as the number of training instances increases, the model performance will also improve. In the case of the same number of instances, as the number of categories increases, the difficulty also increases. (3) MAPN-LC is better than Proto-HATT. The reason is that although Proto-HATT uses feature-level attention, the difference is that we consider the impact of instance similarity on model performance when extracting instances' features. MAPN-LC increases the weight of instances with higher similarity to other instances when extracting features. Based on the above points, MAPN-LC proposed in this paper can better realize the classification of legal consulting questions.

To verify the attention purposed in the paper, the instance-level and dimension-level attention, which are applied on the basis of prototypical networks are helpful to improve the performance of model, we conduct an ablation experiment on the multi-attention prototypical networks. The results are shown in Table 5. Among them, Prototypical networks represent the basic prototypical networks model, DAPN-LC represents prototypical networks model

Table 4 Accuracy comparison between different models on our dataset (%)

Model	5-way		10-way	
	5-shot	10-shot	5-shot	10-shot
SNAIL	47.66	51.03	27.88	30.16
GNN	48.90	49.92	30.70	33.00
Siamese neural networks	57.37	63.44	43.41	49.01
Prototypical networks	62.83	67.94	49.11	54.97
Proto-HATT	66.31	73.64	53.00	60.36
MAPN-LC	**69.94**	**74.60**	**54.91**	**63.19**

Bold values indicate the model with the highest accuracy

Table 5 Ablation study of prototypical Networks based on multi-attention on our dataset (%)

Model	5-way		10-way	
	5-shot	10-shot	5-shot	10-shot
Prototypical networks	62.83	67.94	49.11	53.48
DAPN-LC	67.70	73.60	53.94	62.29
MAPN-LC	**69.94**	**74.60**	**54.91**	**63.19**

Bold values indicate the model with the highest accuracy

Table 6 Accuracy comparison of few-shot text classification on HuffPost (%)

Model	5-way		10-way	
	5-shot	10-shot	5-shot	10-shot
SNAIL	45.02	49.78	29.93	32.37
GNN	42.93	47.84	27.28	30.42
Siamese neural networks	46.43	48.82	31.66	33.87
Prototypical networks	**54.08**	58.36	**39.57**	43.82
Proto-HATT	52.64	57.90	38.52	44.22
MAPN-LC	52.46	**58.40**	38.06	**44.46**

Bold values indicate the model with the highest accuracy

Table 7 Accuracy comparison of relation classification on FewRel validation set (%)

Model	5-way		10-way	
	5-shot	10-shot	5-shot	10-shot
SNAIL	80.83	83.55	70.05	71.97
GNN	77.56	81.67	64.88	69.61
Siamese neural networks	82.04	85.59	70.49	71.48
Prototypical networks	85.64	87.88	74.62	78.05
Proto-HATT	85.72	88.35	75.06	78.73
MAPN-LC	**85.80**	**88.51**	**75.64**	**78.80**

Bold values indicate the model with the highest accuracy

that only applies dimension-level attention, and MAPN-LC represents prototypical networks model that integrates instance-dimension level attention.

It can be seen from Table 5: (1) In the four cases, DAPN-LC are 5% higher than the basic prototypical networks model on average, which proves that extracting instances features to express its category characteristics is helpful to improve the model performance. (2) MAPN-LC that applies instance-level attention on the basis of DAPN-LC is 1–2% higher than DAPN-LC, which proves that the dimensional features extracted after weighting the instances are more representative. (3) Based on the above experimental results, it can be seen that adding instance-dimension level attention on the basis of the prototypical networks can improve the performance of the model.

4.4.2 Generalization Ability Verification

To verify that the model proposed in this paper is also applicable to other public datasets, we experimented on HuffPost, FewRel and Amazon. The accuracy comparison results are shown in Tables 6, 7 and 8. It can be seen that the performance of MAPN-LC is higher than that of Proto-HATT in most cases, which proves that the model proposed in this paper is not only suitable for legal consulting questions dataset, but also for other general few-shot datasets. However, compared with Proto-HATT, the performance improvement is not obvious, indicating that this model is more suitable for the legal consulting questions dataset.

Table 8 Accuracy comparison of few-shot text classification on Amazon (%)

Model	5-way	
	5-shot	10-shot
SNAIL	46.47	48.51
GNN	42.55	47.59
Siamese neural networks	50.29	52.31
Prototypical networks	51.38	55.75
Proto-HATT	56.81	62.50
MAPN-LC	**57.08**	**63.09**

Bold values indicate the model with the highest accuracy

5 Conclusion and Future Work

In this paper, we propose a classification model based on multi-attention prototypical networks to solve the few-shot classification task in the field of legal intelligent question-answering. First, a few-shot consulting questions classification dataset in the legal field is constructed and the classification is realized by integrating multi-attention on the basis of the prototypical networks. Among them, multi-attention refers to the instance-dimension level attention. The instance-level attention is mainly used to weight each category instances so that the local features extracted by the dimension-level attention can better represent the characteristics of its category. The dimension-level attention is used to capture the semantic information of instances and alleviate the problem of feature sparseness. In the experiment, we compare with the current mainstream FSL methods, the results show that the model proposed in this paper is better than baseline models.

In the future, we will refine the classification dataset of few-shot consulting problem in the legal field which constructed in this paper, and explore more suitable models for the classification of legal consulting questions, so as to improve the performance of classification.

Acknowledgements This work is supported by sub-project of the National Key Research and Development Program (2020YFC0833404), Scientific and technological research projects of colleges and universities in Hebei Province (QN2018074), and the Nature Scientist Foundation of Hebei Province (F2019203157).

Declarations

Conflict of interest The authors declare that they have no conflicts of interest.

References

1. Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1746–1751, Doha, Qatar, Association for Computational Linguistics (2014)
2. Changshun, D., Huang, L.: Text classification research with attention-based recurrent neural networks. Int. J. Comput. Commun. Control **13**(1), 50–61 (2018)
3. Huang, L., Ma, D., Li, S., Zhang, X., Wang, H.: Text level graph neural network for text classification. In: 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), pp. 3444–3450, Hong Kong, China, Association for Computational Linguistics (2019)
4. Snell, J., Swersky, K., Zemel, R.S.: Prototypical networks for few-shot learning. arXiv:1703.05175 (2017)
5. Wang, Y., Yao, Q., Kwok, J.T., Ni, L.M.: Generalizing from a few examples: a survey on few-shot learning. ACM Comput. Surv. (CSUR) **53**(3), 1–34 (2020)
6. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of the 34th international conference on machine learning—vol. 70, ICML'17, pp. 1126–1135. JMLR.org (2017)
7. Sung, F., Yang, Y., Zhang, L., Xiang, T., HS Torr, P., Hospedales, T.M.: Learning to compare: relation network for few-shot learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1199–1208 (2018)
8. Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., Wierstra, D.: Matching networks for one shot learning. arXiv:1606.04080 (2016)
9. Tianyu Gao, X., Han, Z.L., Sun, M.: Hybrid attention-based prototypical networks for noisy few-shot relation classification. Proce. AAAI Conf. Artif. Intell. **33**, 6407–6414 (2019)
10. Zhou, C., Sun, C., Liu, Z., Lau, F.: A c-lstm neural network for text classification. arXiv:1511.08630 (2015)
11. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: human language technologies, pp. 1480–1489, San Diego, California, Association for Computational Linguistics (2016)
12. Koch, G., Zemel, R., Salakhutdinov, R.: Siamese neural networks for one-shot image recognition. In: ICML deep learning workshop, vol 2. Lille (2015)
13. Mishra, N., Rohaninejad, M., Chen, X., Abbeel, P.: A simple neural attentive meta-learner. In: International conference on learning representations (2018)
14. Satorras, V.G., Estrach, J.B.: Few-shot learning with graph neural networks. In: International conference on learning representations (2018)
15. Sun, S., Sun, Q., Zhou, K., Lv, T.: Hierarchical attention prototypical networks for few-shot text classification. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), pp. 476–485, Hong Kong, China, Association for Computational Linguistics (2019)
16. Geng, R., Li, B., Ye, Y., Jian, P., Sun, J.: Induction networks for few-shot text classification. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), pp. 3895–3904 (2019)
17. Geng, R., Li, B., Li, Y., Sun, J., Zhu, X.: Dynamic memory induction networks for few-shot text classification. In: Dan, J., Joyce, C., Natalie, S., Joel, R.T. (eds.) Proceedings of the 58th annual meeting of the association for computational linguistics, ACL 2020, Online, pp. 1087–1094. Association for Computational Linguistics (2020)
18. Bao, Y., Wu, M., Chang, S., Barzila, R.: Few-shot text classification with distributional signatures. In: International conference on learning representations (2020)
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, U., Polosukhin, I.: Attention is all you need. In: Proceedings of the 31st international conference on neural information processing systems, NIPS'17, pp. 6000-6010, Red Hook, Curran Associates Inc (2017)
20. Misra, R.: News category dataset (2018)
21. Han, X., Zhu, H., Yu, P., Wang, Z., Yao, Y., Liu, Z., Sun, M.: Fewrel: a large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In: Proceedings of the 2018 conference on empirical methods in natural language processing, pp. 4803–4809, Brussels, Belgium, Association for Computational Linguistics (2018)