Few-Shot Relation Classification Research Based on Prototypical Network and Causal Intervention

ZHIMING LI, FEIFAN OUYANG, CHUNLONG ZHOU, YIHAO HE, AND LIMIN SHEN (Member, IEEE)
School of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China
The Key Laboratory for Computer Virtual Technology and System Integration of Hebei Province, Qinhuangdao 066004, China
The Key Laboratory for Software Engineering of Hebei Province, Qinhuangdao 066004, China
Corresponding authors: Zhiming Li (lizm@ysu.edu.cn) and Limin Shen (shenllmm@sina.com)

ABSTRACT To improve the accuracy of the few-shot relation classification task, the research, which weakens the influence of confounder on the performance of the model and enhances the semantic representation and feature extraction ability of the model, is carried out based on the prototypical network. And then, the weaken-confounders method based on causal intervention (WCCI) is proposed, and the RBERTI-Proto model is constructed. In WCCI, the pre-trained knowledge is stratified by the backdoor adjustment based on causal intervention, the optimal stratified number is determined by a stratified method, and the BN layer is introduced for the gradient disappearance problem. In the RBERTI-Proto model, the abilities of semantic representation and feature extraction of the model are enhanced by which the RoBERTa is used as the feature extractor of the model. Experimental results demonstrate the effectiveness of our proposed methods and the RoBERTa model as feature extractor of our model, and the ACC value of the RBERTI-Proto model achieve 93.38% on the 5-way 5-shot scenario of the FewRel dataset.

INDEX TERMS Few-shot learning, causal intervention, prototypical network, relation classification, RoBERTa, confounder.

I. INTRODUCTION

In recent years, with the support of massive annotated data, deep learning methods, especially deep Convolutional Neural Networks, have achieved excellent prediction performance in various machine learning tasks. However, the obtaining of massive annotated data is difficult and costly in practice, which restricts the development and application of deep learning. Inspired by human’s ability to recognize new things based on the pre-trained knowledge quickly, few-shot learning (FSL) can achieve a pleasing classification effect only by using pre-trained knowledge of similar tasks and the few annotated data related to the current task, which has attracted widespread attention of researchers.

A lot of research works on FSL in the field of computer vision (CV) are carried out, and several classification models with pleasing effect such as Meta-Learning model [1]–[4], siamese networks model [5], and prototypical networks model [6]. The basic principle of the prototypical networks model is as follows: firstly, the feature space based on pre-trained knowledge of similar tasks is constructed; then, the class prototypes are obtained based on the feature space and few annotated data related to the current task, and the classification task is completed according to the distance between the class prototype and the samples in the query set. Because of its simple structure and excellent classification effect, the prototypical networks model has gradually developed into the baseline model.

As a multi-classification problem of natural language processing (NLP), relation classification (RC) aims to determine the relation between two entities in a sentence. As for the RC task, researches are carried out from multiple perspectives, such as the kernel methods (Zelenko [7], Mooney [8]), the embedding method (Gormley [9]), the deep neural network method (Zeng [10]), and the distant supervision [11], [12]. The above RC task methods need many
annotated data as support and are not competent for RC tasks with few samples. Therefore, Han [13] proposed using FSL to solve RC tasks in few-shot scenarios, constructed a dataset named FewRel for the few-shot RC task research, and applied few-shot learning models in the CV field on this dataset for performance comparison experiments. Experimental result shows that the prototypical networks model has the best effect, and the accuracy rate reaches 84.79% in the 5-way 5-shot.

At present, most machine learning algorithms, such as deep learning and few-shot learning, which are essentially suitable methods of observed data, aim to find the correlation information between things. However, it is difficult to explain the causality behind the correlation between things. The causality model [14] is characterized by improving the predictive performance and interpretability of the model by exploring and using the causal relation behind the correlation between things, which is different from traditional machine learning models. The three levels of causality proposed by Pearl [14] are association, intervention, and counterfactual, which are known as the “causal ladder.” These three levels correspond to three cognitive abilities in human evolution. Association, which refers to finding rules among things through passive observation, corresponds to seeing ability. Intervention, more advanced than association, not only finds the rules between things through passive observation, but also actively applies the planned behavior to improve the result, corresponds to doing ability. Counterfactual, which refers to the negation and re-representation of the facts that have happened, and to construct a possible hypothesis, corresponds to imagining ability. Traditional machine learning models, no matter how large the data set or how complex the network structure, cannot jump from level 1 to level 2 as long as the data used is “passively” collected.

Through the above analysis, we find that the following challenges exist in the study of few-shot relationship classification.

(1) The difference between the CV field and the NLP field lies in the input mode, so most of the FSL models in the CV field are difficult to apply to the RC task in the NLP field directly. The image is composed of pixel points, which is the abstraction of the real world by the computer. However, natural language is a paragraph of words or a text, which is an abstraction of the natural world by humans, and there are various complicated situations such as polysemy. Therefore, the research is carried out for few-shot learning models from the CV field to make them more suitable for the RC task problem.

(2) Focusing on the text diversity and its noise, Proto-HATT [15] proposed by Gao improves the performance and robustness of the RC model in the noisy few-shot learning scenarios by highlighting key features and examples. The accuracy of Proto-HATT reaches 90.12% on the 5-way 5-shot scenario of the FewRel dataset. To solve the problem of the long-tailed distributions in few-shot RC, MLMAN [16], proposed by Zhi-Xiu Y based on the prototypical network, codes the state of the query set and class prototypes in an interactive way, measures the consistency among its support instances by auxiliary loss function designed. The accuracy of MLMAN reaches 92.66% on the 5-way 5-shot scenario of the FewRel dataset. However, the above models use CNN model as the feature extractor, due to the limitation of convolution kernel size, so CNN model cannot capture long-distance contextual semantic information.

(3) From the point of view of causal reasoning, it can see that Proto-HATT and MLMAN belong to the prediction models established at the association level, ignoring the influence of confounders [14] on the performance of prediction models. In terms of studies on weakening the influence of confounders to the performance of prediction models by causal intervention [17]–[20], the most representative method of these studies is a method which is proposed by Yue [20]. The method have realized causal intervention by backdoor adjustment to the pre-trained knowledge and applied to some standard meta-learning models [3], [21]–[24] and fine-tuning models (Linear, Cosine, k-NN) to improve the prediction performance of the models significantly. However, in this method, the concept of dividing the feature vectors into N equal parts is proposed, but the determination strategy for the optimal stratified number is not given.

To sum up, there are several problems with the few-shot relationship classification.

(1) Due to the difference between the CV field and the NLP field lies in the input mode, most of the FSL models in the CV field are difficult to apply to the RC task in the NLP field directly.

(2) Due to the limitation of convolution kernel size, CNN model cannot capture long-distance contextual semantic information, it will lead to CNN model unable to obtain accurate semantic representation.

(3) Most of the FSL models belong to the prediction models established at the association level, ignoring the influence of confounders on the performance of prediction models. Though the method of Yue [20] considers confounders, but the optimal stratified number is assigned manually, it will lead to the increase the experimental workload and cannot reuse the experience accumulated through a large number of experiments.

According to the above problems, to improve the accuracy of the few-shot RC task, the research, which weakens the influence of confounders on the performance of the model through the method of the causal intervention, is carried out based on the prototypical network and the RoBERTa [25] model. In summary, our contributions in this paper are three-fold.

(1) To further enhance the semantic representation and feature extraction capability of the model, RoBERTa is used to replace the CNN part of the prototypical network as a feature extractor, which can make the model more suitable for RC task scenarios and effectively improve the prediction performance of the model.
(2) WCCI is proposed. In WCCI, the pre-trained knowledge is stratified by the backdoor adjustment based on causal intervention, the gradient disappearance problem caused by the backdoor adjustment is eliminated with the BN layer introduced, and the optimal stratified number is determined by which the stratified number is taken as a network parameter and trained during the model training.

(3) Based on the above work, a few-shot RC model named RBERTI-Proto based on the prototypical network and causal intervention is constructed. Experimental results demonstrate the effectiveness of our proposed methods and the RoBERTa model as feature extractor of our model, and the ACC value of the RBERTI-Proto model achieve 93.38% on the 5-way 5-shot scenario of the FewRel dataset.

The remainder of the paper is organized as follows. Section II introduces the Few-shot RC task and related definitions and theorems. Section III proposes the weaken-confounders method based on causal intervention(WCCI), Section IV constructs the RBERTI-Proto model and Section V reports the experimental results and analysis. The paper concludes with Section VI.

II. TASK DEFINITION
A. THE FEW-SHOT RC TASK
The goal of the few-shot RC task is to learn a function \( f: (S, x, R) \rightarrow y, R = \{r_1, r_2, r_3, \ldots, r_n\} \) represents the set of relations in the data set, \( S \) represents the support set, \( Q \) represents the query set, and \( x \in Q \), the relation class \( y \in Q \) of unlabeled instance \( x \) in the query set \( Q \) is predicted by the information on the support set \( S \). Where the support set \( S \) are as follows:

\[
S = \{ (x_1, r_1), (x_2, r_1), (x_3, r_1), \ldots, (x_n, r_1) \}
\]

\[
\ldots
\]

\[
(1)
\]

N-way K-shot is generally used to represent different scenarios on the few-shot task, where N-way means that there are N relation classes in the support set \( S \), and K-shot implies that there are K samples under each class. Therefore, the support set \( S \) contains \( N \times K \) samples, in Equation(1),

B. RELATED DEFINITIONS AND THEOREMS
For the convenience of the following statement, the necessary definitions and theorems are defined as follows.

1) CLASS原型 [6]
In the prototypical network, the mean feature vector of the embedded support points belonging to its class is called the class prototype of its class:

\[
e_k = \frac{1}{|S_k|} \sum_{(x_i, y_i) \in S_k} F(x_i) \quad (2)
\]

where \( e_k \) represents the class prototype of class \( k \), \( S_k \) represents the set of examples labeled with class \( k \) in the support set, \( x_i \) is the feature vector of an example and \( y_i \) is the corresponding class, \( F(\cdot) \) is the embedding function used to extract the feature vector representation of \( x_i \).

2) BACKDOOR CRITERION
In the causal diagram shown in Figure 1, if the node in \( Z \) is not the descendant of \( X \), and each path between \( X \) and \( Y \) can be blocked under the condition of \( Z \), namely the backdoor path of \( X \), then the variable set \( Z \) conforms to the backdoor criterion relative to the ordered pair \( (X, Y) \) of variables.

![FIGURE 1. The causal diagram conforming to the backdoor criterion.](image)

Reference 14 also mentioned that if the variable set \( Z \) meets the backdoor criterion of \( (X, Y) \), the causal effect of \( X \) on \( Y \) (in short, the causal effect is that the variation of \( Y \) in the presence of \( X \) and in the absence of \( X \)) can be obtained by adjusting the variable set \( Z \), the adjustment is also known as the backdoor adjustment. The specific adjustment way is that the total probability of \( Z \) is based on the variable set \( Z \) and \( X \), and the backdoor adjustment formula is shown in Equation (3).

\[
p(Y = y | do(X = x)) = \sum_z p(Y = y | X = x, Z = z)p(Z = z) \quad (3)
\]

III. THE WEAKEN-CONFOUNDERS METHOD BASED ON CAUSAL INTERVENTION(WCCI)
A. THE METHOD OF CAUSAL INTERVENTION BASED ON BACKDOOR ADJUSTMENT
Our idea is inspired from the Pearl [14] and Yue [20], the whole prediction process of the RC model is abstracted into a causal diagram, as shown in Figure 2, where \( D \) represents pre-trained knowledge of the pre-training model, \( X \) represents the vector representation of the input, \( C \) represents the projection of \( X \) on the pre-training data manifold, and \( Y \) represents label to be predicted.

![FIGURE 2. The causal diagram of the prediction process of RC model.](image)
confounder when D will lead to a pseudo-correlation between X and Y, also known as bias, which will affect the prediction of true causality. The complex situation, which the pseudo-correlation is mixed with true causality caused by confounders, is called confounding bias. The purpose of weakening the effect of confounders is to eliminate the negative influence of bias on the prediction of true causality as much as possible. According to the backdoor criterion, X ← D → C → Y is the backdoor path of X, and D is not the descendant of X. Therefore, D conforms to the backdoor criterion of (X, Y), and the causal effect of X on Y can be obtained by backdoor adjustment of D, and the causal intervention can be realized in the predicted process of RC model, as shown in Figure 3.

The causal intervention on x is equivalent to stratify D, and the posterior probability \( p(Y = y|do(X = x)) \) in the prediction of the original model is equal to Equation (4).

\[
p(Y = y|do(X = x)) = \sum_D p(Y = y|X = x, C = c, D = d)p(D = d) \quad (4)
\]

It can be known from Equation (3), the posterior probability of the prediction model becomes the sum of the product of the marginal distribution and the conditional probability with different values of D after the backdoor adjustment, and the number of values of d is equivalent to the stratified number of pre-trained knowledge.

According to reference 20, in the metric layer of the prototypical network, the operations of backdoor adjustment are as follows: First, the feature vectors of class prototype and query set inputted into the metric layer are divided into N equal-size disjoint subsets according to their dimensions; Then, the corresponding N equal-size feature vectors of class prototype and query set are inputted into the corresponding classifier; Finally, the average of N classifier prediction probability is taken as the final prediction result. The classifier is composed of distance metrics and softmax. We take the four equal-size as an example, the network structure of backdoor adjustment in the metric layer as shown in Figure 4.

**B. THE PROCESSING METHOD OF GRADIENT DISPERSION**

The causal intervention based on backdoor adjustment though weakens the influence of confounders on the model, yet brings new problems. The operations of backdoor adjustment will generate multiple classifiers in the metric layer of the prototypical network, so that when updating parameters in backpropagation, the parameters of the metric layer which close to the output quickly reach the convergence state, while the parameters of the embedded layer which close to the input are updated slowly and are almost in the initial state. This phenomenon is named gradient disappearance. As a result, the model needs more epochs to offset the effect of gradient disappearance, which increases the model training time and reduces the model accuracy.

According to the gradient disappearance problem, the BN layer was introduced in behind of the embedding layer to normalize the feature vector and further solve the gradient disappearance problem.

**C. THE METHOD TO DETERMINE THE OPTIMAL STRATIFIED NUMBER OF PRE-TRAINED KNOWLEDGE**

The stratified number of pre-trained knowledge has a significant impact on the predicted effect of the model. Therefore, it is also a significant problem to determine the optimal stratified number of pre-trained knowledge. To determine the stratified number of pre-trained knowledge, the stratified number is set usually according to experience or cognition. This method will significantly increase the experimental workload. In addition, different data sets and network models will lead to different the optimal stratified number, so it is difficult to reuse the experience accumulated through a large number of experiments.

According to the above problems, a method to determine the optimal stratified number of pre-trained knowledge was proposed. Firstly, the stratified number was set as parameter T and initialized as 1. The learning rate and the maximum epoch M were determined; then, the parameter T was taken as a model parameter and updated during the model training. When the gradient tends to 0 (less than the preset threshold) or the epoch reaches the maximum, the training is over. The value of parameter T obtained by model training was the optimal stratified number of pre-trained knowledge. The flow of determining the optimal stratified number of pre-trained knowledge is shown in Figure 5.

**IV. THE RBERTI-PROTO MODEL**

To improve the performance of the prediction model, the prototypical networks were improved by WCCI. In addition, to enhance the semantic representation and feature extraction capability of the model, RoBERTa, the upgraded version of BERT [26], was used to replace the CNN part of the prototypical networks as a feature extractor. The overall architecture of the model is shown in Figure 6. To facilitate subsequent expression, the model shown in Fig. 6 was named RoBERTa-Intervention-Proto, and for short RBERTI-Proto.

Compared with CNN, Transformer [27] in RoBERTa and BERT use a multi-head attention mechanism to capture bi-directional relations in sentences more thoroughly. In addition, RoBERTa uses a more extensive training set and a larger batch size to acquire more pre-trained knowledge during
training and utilizes the dynamic mask to replace the static mask in BERT, which enhances the semantic representation and feature extraction ability of the model. Therefore, RoBERTa will make the RBERTI-Proto model more suitable for RC task scenarios and effectively improve the prediction performance.

To solve multi-classification problems, the cross-entropy loss function is selected as the loss function in the RBERTI-Proto model, as shown in Equation (5),

$$\min_{\theta} \sum_m - \log(p_m(Y = y_m | x_m^Q))$$  (5)

where $\theta$ represents all parameters that need to be trained in the network, $y_m^Q$ represents the real label of $x_m^Q$, and $p_m(Y = y_m^Q | x_m^Q)$ represents the prediction probability that $x_m^Q$ belongs to the $y_m$ class in the query set $Q$.

In the RBERTI-Proto model, the causal intervention method based on backdoor adjustment is used for the pre-trained knowledge stratification. As shown in Fig. 4, multiple classifiers need to be set up after feature vectors and class prototypes of the query set are equalized by dimension in the metric layer. Compared with the prototypical networks, the calculation of cross-entropy loss function in the RBERTI-Proto model is unique, as follows. Assuming that the number of classifiers is $n$, and the number of relational classification classes is $N$. When $y_m^Q = i$, the formula of $p_m(Y = y_m^Q | x_m^Q)$ in the $K$th classifier is shown in Equation (6), which is composed of distance measurement function and softmax function.

$$p_m^k(Y = i|x_m^Q) = \frac{\exp(-d(F_{\alpha}^k(x_m^Q), c_{ik}^k))}{\sum_{j=1}^N \exp(-d(F_{\alpha}^k(x_m^Q), c_{jk}^k))}$$  (6)

where $c_{ik}^k$ represents the class prototypes of class $i$ in the $K$th classifier, $c_{jk}^k$ represents the class prototypes of class $j$ in the $K$th classifier, $F_{\alpha}^k(\cdot)$ is the feature extraction function, $\alpha$ represents the required training parameters needed $F_{\alpha}^k(\cdot)$, $F_{\alpha}^k(x_m^Q)$ represents the feature vector of a query point $x_m^Q$ in the $K$th classifier, $d$ is the distance metric function, and $k \in [1, n], i, j \in [1, N]$.

As for the predicted classification probability of the sample that belongs to class $i$, the result of Equation (6) is only the predicted classification probability of one classifier. In the RBERTI-Proto model, the mean of the predicted classification probability of $N$ classifiers is taken as the final predicted classification probability of the sample that belongs to class $i$,
TABLE 1. Configuration of each model.

| Model    | Feature extractor | whether to add backdoor adjustment | whether to add BN layer |
|----------|------------------|------------------------------------|-------------------------|
| Proto_1  | CNN              | No                                 | No                      |
| Proto_2  | RoBERTa          | No                                 | No                      |
| Proto_3  | RoBERTa          | Yes                                | No                      |
| Proto_4  | RoBERTa          | Yes                                | Yes                     |

as shown in Equation (7).

\[
p_m(Y = i | x_m^Q) = \frac{\sum_{n=1}^{n} p_n^k(Y = i | x_n^Q)}{n} \tag{7}
\]

In the RBERTI-Proto model, the Bregman divergence function is used as the metric function \(d\), when \(f(x) = \|x\|^2\) in the Bregman divergence function, \(d\) is shown in Equation (8).

\[
d_f(x, y) = \|x\|^2 - (\|y\|^2 + (2y \cdot x - y)) = \|x - y\|^2 \tag{8}
\]

V. EXPERIMENT

A. DATASET

FewRel, which is suitable for few-shot RC problem studies, was used to validate the performance of our proposed model. This dataset was created by Han [13] based on Wikipedia corpus. It is the largest and highest-quality few-shot RC dataset which contains 100 relation classes, each with 700 sentence instances. In this paper, 80 relation classes that are opened in FewRel are used and divided into the training sets (48 types), the validation sets (16 types), and the test sets (16 types) without crossing.

B. SETTING AND EVALUATION METRIC

Comparison experiment and ablation experiment were used to verify the model’s performance, and accuracy (ACC), which is commonly used in RC field, was selected as the evaluation metric of the model’s performance.

The Siamese model, GNN model, Proto model, Proto-Hatt model, MLMAN model, and RBERTI-Proto model were selected to carry out the performance comparison experiments under the scenarios of 5-way 1-shot, 5-way 5-shot, 10-way 1-shot and 10-way 5-shot respectively to verify the superiority of our model compared with other few-shot learning models in the field of RC. Among them, Siamese(CNN), GNN(CNN), and Proto(CNN), belonging to the few-shot classical models for the image classification, are directly applied into the RC task, and Proto-Hatt (CNN) and MLMAN(CNN) are the improved models for the RC task.

The ablation experiments were carried out in 5-way 1-shot, 5-way 5-shot, 10-way 1-shot, and 10-way 5-shot scenarios respectively, to verify the effectiveness of improvement on each part of the Proto_1. The Proto_1 represents the prototypical network. The Proto_2, the Proto_3, and the Proto_4 models were constituted based on the Proto_1 by the selection of feature extractor, whether to add backdoor adjustment and whether to add BN layer, among which the Proto_4 model is the RBERTI-Proto model.

C. RESULTS AND ANALYSIS

To achieve objective comparison of model’s performance, CNN was taken as the feature extractor in the Siamese, GNN, Proto, Proto-HATT, and MLMAN models, and the 50-dimension Glove was used to obtain word embeddings. In addition, other basic parameters of each model was kept the same settings. All experiments were conducted on a computer with a 8 core 2.1G Intel Xeon E5-2620 and 16GB RAM. The main hyper-parameters of the RBERTI-Proto model are set through the validation set, and they are shown in Table 2. The results of comparative experiments and ablation experiments are shown in Table 3 and Table 4.

TABLE 2. Main hyper-parameter configuration.

| Hyper-parameter            | Value |
|----------------------------|-------|
| Word embedding dimension   | 768   |
| Batch size                 | 4     |
| Learning rate              | 2e-5  |
| Max length                 | 50    |
| Train iterator             | 20000 |
| Validation iterator        | 1000  |
| Test iterator              | 10000 |
| Optimization strategy      | SGD   |
TABLE 3. ACC(%) of different models on the partitioned FewRel test set.

| Model (Feature extractor) | 5-way 1-shot | 5-way 5-shot | 10-way 1-shot | 10-way 5-shot |
|---------------------------|--------------|--------------|---------------|---------------|
| Siamese (CNN)             | 70.74        | 79.67        | 56.11         | 66.63         |
| GNN (CNN)                 | 64.86        | 77.93        | 52.44         | 65.17         |
| Proto (CNN)               | 72.51        | 85.07        | 60.57         | 74.51         |
| Proto-HATT (CNN)          | 73.92        | 86.44        | 60.59         | 75.04         |
| MLMAN (CNN)               | 76.49        | 86.76        | 64.31         | 78.23         |
| RBERTI-Proto              | 82.87        | 93.38        | 72.94         | 87.74         |

TABLE 4. ACC(%) of the models in this paper under different configurations.

| Model | 5-way 1-shot | 5-way 5-shot | 10-way 1-shot | 10-way 5-shot |
|-------|--------------|--------------|---------------|---------------|
| Proto_1 | 72.51 | 85.07        | 60.57         | 74.51         |
| Proto_2 | 80.53 | 91.82        | 69.53         | 84.67         |
| Proto_3 | 81.32 | 91.87        | 70.57         | 85.04         |
| Proto_4 | 82.87 | 93.38        | 72.94         | 87.74         |

2) ABLEATION STUDY
From Table 1 and Table 4, we observe that based on the prototypical network, RoBERTa which replaces CNN was taken as the feature extractor in the Proto_2 model, and the ACC value of the Proto_2 model is greatly improved. This indicates that the enhancement of semantic representation and feature extraction ability has an essential influence on the performance of the few-shot RC model. It also verifies the correctness of the theoretical analysis that CNN in the prototypical network was replaced by RoBERTa mentioned above.

Based on the Proto_2 model, the backdoor adjustment was applied in the Proto_3 model, but the ACC value of the Proto_3 model is not obviously improved compared with the Proto_2 model, because of the gradient dispersion caused by backdoor adjustment operation. Based on the model Proto_3, the BN layer was added into the model Proto_4 to eliminate gradient dispersion. Compared with the model Proto_2, the ACC value of the model Proto_4 increases significantly, which fully shows the necessity of introducing the BN layer in WCCI.

VI. CONCLUSION
In this paper, to improve the accuracy of the few-shot RC task, the research, which weakens the influence of confounders on the performance of the model and enhances the semantic representation and feature extraction ability of the model, has been carried out based on the prototypical network. The following results are obtained:

1) The RoBERTa is selected as the feature extractor in the few-shot relation classification model, which significantly improves the prediction accuracy of the model.

2) WCCI has been proposed for few-shot relation classification. In WCCI, the pre-trained knowledge is stratified by the backdoor adjustment based on causal intervention, the optimal stratified number is determined by which the stratified number is taken as a network parameter and trained during the model training, and the gradient disappearance problem caused by the backdoor adjustment is eliminated with the BN layer introduced.

3) The RBERTI-Proto model has been constructed. Experimental results demonstrate the effectiveness of our proposed methods and the RoBERTa model as feature extractor of our model, which ACC value achieve 93.38% on the 5-way 5-shot scenario of the FewRel dataset.

Though we obtained good results, there are many questions that need further research. In the future, the following questions warrant a thorough theoretical analysis and more empirical studies: 1) How to further improves the prediction accuracy of our model by using better stratified strategies; 2) How to make our method more general by applying WCCI to other FSL models.

REFERENCES
[1] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, “Meta-learning with memory-augmented neural networks,” in Proc. Int. Conf. Mach. Learn., Jun. 2016, pp. 1842–1850.
[2] S. Ravi and H. Larochelle, “Optimization as a model for few-shot learning,” in Proc. ICLR, 2017, pp. 1–4.
[3] C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” in Proc. Int. Conf. Mach. Learn., Jul. 2017, pp. 1126–1135.
[4] T. Munkhdalai and H. Yu, “Meta networks,” in Proc. Int. Conf. Mach. Learn., Jul. 2017, pp. 2554–2563.
[5] G. Koch, R. Zemel, and R. Salakhutdinov, “Siamese neural networks for one-shot image recognition,” in Proc. ICML deep Learn. Workshop, vol. 2, Jul. 2017, pp. 1–30.
[6] J. Snell, K. Swersky, and R. S. Zemel, “Prototypical networks for few-shot learning,” 2017, arXiv:1703.05175.
[7] D. Zelenko, C. Aone, and A. Richardella, “Kernel methods for relation extraction,” J. Mach. Learn. Res., vol. 3, pp. 1083–1106, Feb. 2003.
[8] R. Bunescu and R. J. Mooney, “Subsequence kernels for relation extraction,” in Proc. NIPS, Dec. 2005, pp. 171–178.
[9] M. R. Gormley, M. Yu, and M. Dredze, “Improved relation extraction with feature-rich compositional embedding models,” 2015, arXiv:1505.02419.
[10] D. Zeng, K. Liu, S. Lai, G. Zhou, and J. Zhao, “Relation classification via convolutional deep neural network,” in Proc. 25th Int. Conf. Comput. Linguistics, Tech. Papers, Aug. 2014, pp. 2335–2344.
[11] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, “Distant supervision for relation extraction without labeled data,” in Proc. Joint Conf. 47th Annu. Meeting ACL, 2009, pp. 1003–1011.
[12] S. Riedel, L. Yao, and A. McCallum, “Modeling relations and their mentions without labeled text,” in Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases. Berlin, Germany: Springer, Sep. 2010, pp. 148–163.

[13] X. Han, H. Zhu, P. Yu, Z. Wang, Y. Yao, Z. Liu, and M. Sun, “FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation,” 2018, arXiv:1810.10147.

[14] J. Pearl and D. Mackenzie, The Book of Why: The New Science of Cause and Effect. New York, NY, USA: Basic books, 2018.

[15] T. Gao, X. Han, Z. Liu, and M. Sun, “Hybrid attention-based prototypical networks for noisy few-shot relation classification,” in Proc. AAAI Conf. Artif. Intell., Jul. 2019, vol. 33, no. 1, pp. 6407–6414.

[16] Z.-X. Ye and Z.-H. Ling, “Multi-level matching and aggregation network for few-shot relation classification,” 2019, arXiv:1906.06678.

[17] S. Yu, Y. Niu, S. Wang, J. Jiang, and Q. Sun, “Counterfactual variable control for robust and interpretable question answering,” 2020, arXiv:2010.05581.

[18] X. Yang, H. Zhang, and J. Cai, “Deconfounded image captioning: A causal retrospect,” IEEE Trans. Pattern Anal. Mach. Intell., early access, Oct. 21, 2021, doi: 10.1109/TPAMI.2021.3121705.

[19] Y. Niu, K. Tang, H. Zhang, Z. Lu, X.-S. Hu, and J.-R. Wen, “Counterfactual VQA: A cause-effect look at language bias,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 12700–12710.

[20] Z. Yue, H. Zhang, Q. Sun, and X.-S. Hu, “Interventional few-shot learning,” 2020, arXiv:2009.13000.

[21] A. A. Rusu, D. Rao, J. Sygnowski, O. Vinyals, R. Pascanu, S. Osindero, and R. Hadsell, “Meta-learning with latent embedding optimization,” 2018, arXiv:1807.05960.

[22] Q. Sun, Y. Liu, T.-S. Chua, and B. Schiele, “Meta-transfer learning for few-shot learning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 403–412.

[23] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, “Matching networks for one shot learning,” in Proc. Adv. Neural Inf. Process. Syst., vol. 29, 2016, pp. 3630–3638.

[24] S. Xu Hu, P. G. Moreno, Y. Xiao, X. Shen, G. Obozinski, N. D. Lawrence, and A. Damianou, “Empirical Bayes transductive meta-learning with synthetic gradients,” 2020, arXiv:2004.12696.

[25] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “RoBERTa: A robustly optimized BERT pretraining approach,” 2019, arXiv:1907.11692.

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” 2018, arXiv:1810.04805.

[27] A. Vaswani, N. Shazeer, and N. Parmar, “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 5998–6008.