Multi-Channel Convolutional Neural Networks with Adversarial Training for Few-Shot Relation Classification (Student Abstract)

Yuxiang Xie,[1,2] Hua Xu,[1*] Congcong Yang,[1,2] Kai Gao[2*]
[1]State Key Laboratory of Intelligent Technology and Systems, Department of Computer Science and Technology,
Tsinghua University, Beijing 100084, China
[2]School of Information Science and Engineering, Hebei University of Science and Technology,
Shijiazhuang 050018, China
{1021871473xyx, yang957421025}@gmail.com, xuhua@tsinghua.edu.cn, gaokai@hebust.edu.cn

Abstract

The distant supervised (DS) method has improved the performance of relation classification (RC) by means of extending the dataset. However, DS also brings the problem of wrong labeling. Contrary to DS, the few-shot method relies on few supervised data to predict the unseen classes. In this paper, we use word embedding and position embedding to construct multi-channel vector representation and use the multi-channel convolutional method to extract features of sentences. Moreover, in order to alleviate few-shot learning to be sensitive to overfitting, we introduce adversarial learning for training a robust model. Experiments on the FewRel dataset show that our model achieves significant and consistent improvements on few-shot RC as compared with baselines.

Introduction

RC is an important task of natural language processing (NLP). Previous RC works rely too heavily on supervised annotation, which is quite expensive. In order to obtain scaled RC dataset, Mintz et al. (2009) propose a distant supervised method (DS) to generate data automatically. DS method alleviates the issue of lack of supervised data, while at the same time, it brings about wrong labeling. How to extract relation on the basis of few clean and supervised data remains a challenging task.

Han et al. (2018) introduce a few-shot learning framework for RC, which can predict the unseen classes with few labeled data. Different from the DS method which improves model performance by training on automatic annotated DS dataset, the few-shot method tries to build a model with few labeled data or even without data. Snell, Swersky, and Zemel (2017) propose prototypical networks for the problem of few-shot classification based on the assumption that there exists a prototype for each class. Gao et al. (2019) design hybrid attention schemes based on prototypical networks for few-shot RC.

In this paper, we build our model on the few-shot RC dataset (FewRel) proposed by (Han et al. 2018). We built word embedding and position embedding into a multi-channel form similar to image representation and use multi-channel convolution neural networks for feature extraction. Besides, we attack the problem of overfitting in few-shot learning by introducing adversarial learning in our model training. We generate an adversarial example by adding small and continuous perturbations to the raw multi-channel vector representation for word. Empirically, we evaluate on the FewRel dataset and demonstrate significant and consistent improvements in RC as compared with the advanced methods.

Model

Problem Definition. In the training episode of few-shot classification, we define a function $\mathbf{T}_k : (\mathbf{S}, \mathbf{Q}) \rightarrow y$, choosing some examples from each class to build up support set $\mathbf{S}$ and the remainder to serve as query points $\mathbf{Q}$. The support set has k classes and each class includes j sentences $\mathbf{S} = \{(s_1^1, \cdots, s_1^j), (s_2^1, \cdots, s_2^j), \cdots, (s_k^1, \cdots, s_k^j)\}$, and the query set has k classes and each class includes h sentences $\mathbf{Q} = \{(q_1^1, \cdots, q_1^h), (q_2^1, \cdots, q_2^h), \cdots, (q_k^1, \cdots, q_k^h)\}$. The goal of our model is training a function $\mathbf{T}$ to predict the corresponding label y of query point q. Following the recent few-shot learning setting, we adopt K way J shot for our task as follows, $K = k, J = j$.

Vector Representation. Given a sentence s consisting of n words $s = \{w_1, w_2, \cdots, w_n\}$, e_1, e_2 are the two corresponding entities. Each word w_i has relative distances of w_i to e_1 and e_2. We build the word embedding v_i and the position embedding p_i^1, p_i^2 as a multi-channel vector representation of $r_i^m = [v_i : p_i^1 : p_i^2]$.

Multi-Channel Convolution. We use a multi-channel convolution filter $F^m = [F_w : F_{p_1} : F_{p_2}]$ to extract the local features from different spatial channels of multi-channel vectors. Next, we sum the extracted local features on each channel. The convolution operate of $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{m \times n}$ is defined as: $A \otimes B = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} b_{ij}$. We can obtain a word feature vector c_i by equation: $c_i = f(v_{i \sim i+j} \otimes F_w + b_w)$. We also can obtain a position feature vector q_i by equation: $q_i = f(p_{i \sim i+j} \otimes F_p + b_p)$. Here b_w, b_p are bias terms and $f(\phi)$ is a non-linear function. Finally, we can obtain the sentence vector representation by add up

[*]Corresponding author

the word feature and the position feature $d_i = c_i + q_i^1 + q_i^2$.

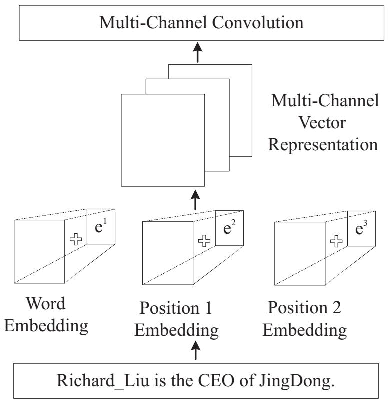
Multi-Channel Convolution
Multi-Channel Vector Representation
e1
e2
e3
Word Embedding
Position 1 Embedding
Position 2 Embedding
Richard_Liu is the CEO of JingDong.

Figure 1: The adversarial training for multi-channel word embedding. The multi-channel adversarial perturbation is represented by $e^m = [e^1 : e^2 : e^3]$.

Adversarial Training for Prototypical Network. There is a hypothesis in the prototypical network (Snell, Swersky, and Zemel 2017) that each class has a class prototype. The probability p_ϕ over classes for a query point is defined as a softmax over distances to the prototypes. The optimization is processed by minimizing the loss $L(\theta) = -\log p_\phi$ of the true class via SGD. Previous works get the prototype of a class by taking the mean of the vector representation of support examples belonging to its class. However, examples of the support set are randomly selected from the training set, and the mean of these samples does not fully represent the centre of the class. To remedy the error between the mean of the embedded support examples and the class prototype, we introduce adversarial training (AT) in prototypical networks. Different from Wu, Bamman, and Russell (2017), we generate a small multi-channel adversarial perturbation e^m to the multi-channel raw data r^m (as shown in Figure 1) that maximizes the loss function:

$$e^m = arg \max_{\|e\| \leq \epsilon} L(r^m + e; \hat{\theta}) \qquad (1)$$

where $\hat{\theta}$ denotes a fixed copy of the current model parameters and ϵ is a small bounder norm. In order to simplify the computation process of Eq.(1), Eq.(1) is replaced by: $e^m = \epsilon g / \|g\|$, with $g = \nabla_{r^m} L(r^m; \hat{\theta})$.

Experiments

Datasets. We evaluate our model on the few-shot RC dataset (FewRel), which is developed by (Han et al. 2018). There are 100 relationships, and each has 700 instances. The FewRel dataset uses separate sets of classes for training and testing, which uses 64, 16, and 20 relations for training, validation, and testing respectively.

Baseline Methods. Table 1 illustrates the performance of our model in comparison with Prototypical Networks (Proto-CNN) (Snell, Swersky, and Zemel 2017) and Hybrid Attention-Based Prototypical Networks (Proto-HATT) (Gao et al. 2019).

Model	5 way 5 shot	10 way 5 shot
Proto-CNN	84.79±0.16	75.55±0.19
Proto-HATT	90.12±0.04	83.05±0.05
Our (Proto-Mul)	86.90±0.16	77.30±0.24
Our (Proto-Mul+AT)	**90.70± 0.10**	**83.70±0.19**

Table 1: Accuracies(%) of different models on FewRel test set.

Result Analysis. From Table 1, we observed the following results: (1) The multi-channel convolutional method (Proto-Mul) has improved by $2\% \sim 3\%$ compared with CNN on the prototypical network. It indicates that the multi-channel vector representation can meaningfully fusion word embedding and position embedding, and the multi-channel convolution method can effectively extract sentence features from raw data. (2) With introducing adversarial training in our model (Proto-Mul+AT), the results under two few-shot experimental settings have improved by $6\% \sim 7\%$ compared with the strong baselines (Proto-CNN). Moreover, Proto-Mul+AT performs better compared to Proto-HATT, and they use an advanced hybrid attention mechanism more than our model. This shows that the few-shot method has overfitting on FewRel dataset. It is effective to add small and continuous multi-channel perturbation to the multi-channel vector representation.

Conclusion

In this paper, we propose a multi-channel convolutional method with adversarial training for few-shot relation classification. The experiment results demonstrate that our method significantly improves the performance and robustness of the model on few-shot RC task.

Acknowledgments: This paper is sponsored by the National Science Foundation of China (61673235, 61772075) and National Science Foundation of Hebei Province (F2017208012).

References

Gao, T.; Han, X.; Liu, Z.; and Sun, M. 2019. Hybrid attention-based prototypical networks for noisy few-shot relation classification. In *Proceedings of AAAI*, 6407–6414.

Han, X.; Zhu, H.; Yu, P.; Wang, Z.; Yao, Y.; Liu, Z.; and Sun, M. 2018. Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of EMNLP*, 4803–4809.

Mintz, M.; Bills, S.; Snow, R.; and Jurafsky, D. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of ACL-IJCNLP*, 1003–1011.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. In *Proceedings of NIPS*, 4077–4087.

Wu, Y.; Bamman, D.; and Russell, S. 2017. Adversarial training for relation extraction. In *Proceedings of EMNLP*, 1778–1783.