Emphasis on Easy Samples for Distantly Supervised Relation Extraction

Anonymous ACL submission

Abstract

There are many wrongly-labeled samples and low-quality samples in automatically generated Distantly Supervised Relation Extraction datasets. Overfitting these samples leads to decline of generalization. To address this issue, the learning of high-quality samples should be prioritized. In this paper, we propose the Emphasis on Easy Samples (EES) mechanism to emphasize high-quality samples using weight distribution regularization at sentence level and priority weighting at bag level. Experiments on a widely used benchmark show that our approach achieves significant improvements.

1 Introduction

Distantly Supervised Relation Extraction (DSRE) (Mintz et al., 2009) is proposed for effective construction of knowledge bases. However, it also introduces sentences and sentence bags with wrong labels, which can be called **Noisy Samples**. In addition, due to the low quality of web-crawled corpus, some of the sentences are poorly structured or overly ambiguous. Sometimes all the sentences in a bag are of low quality. Such sentences and sentence bags can be viewed as **Hard Samples**. The remaining well-structured high-quality samples with correct labels are **Easy Samples**. For example, for entity pair david ben-gurion and israel with relation */people/person/nationality*:

- **Easy Sample**: he said israel's first leader, david ben-gurion, ...
- **Noisy Sample**: mr.bar-zohar, a noted biographer of david ben-gurion, first wrote this book ... and it was published in israel in december 2005.
- **Hard Sample**: mr.feldman was sent to meet quietly with israeli leaders, particularly david ben-gurion ..., about matters including ... and whether israel was building a nuclear weapon.

In the hard sample of the example, the pair entities are far from each other and have no direct connections, making it hard to fit during training.

Overfitting noisy and hard samples may hinder the generalization of the model. Therefore, many of previous methods focus on alleviating the impact of noisy and hard samples in the bag (Zeng et al., 2015; Lin et al., 2016) or superbag (Yuan et al., 2019b; Ye and Ling, 2019). However, there is little discussion about how to distinguish easy samples from the noisy and hard ones. Moreover, without explicitly emphasizing easy samples, overfitting of hard/noisy samples still occurs during the training of previous models (Zhang et al., 2017).

To address these issues, we leverage the Logit Margin (LM) (Huang et al., 2021) to capture easy samples and devise a two-level approach named **Emphasis on Easy Samples (EES)** to avoid overfitting on hard/noisy samples. At sentence level, we apply regularization on the weight distribution within the sentence bag to emphasize easy sentences. At bag level, we introduce a priority weight to prioritize the learning of easy bags while slowing the overfitting of hard/noisy bags.

Our contributions can be summarized as follows:

- We are the first one to address the overfitting of low-quality samples in DSRE. We utilize the logit matrix to measure the sample quality.
- We design the EES mechanism, which highlights high-quality sentences and sentence bags during training, to alleviate the overfitting problem. No extra parameters are needed in our approach.
- The experiments show that our method significantly improve the generalization of the model.

2 Related Work

Distantly Supervised Relation Extraction (Mintz et al., 2009) is proposed for automatic annotation in large-scale relation extraction. To alleviate the impact of noisy sentences introduced by the strong assumption of DSRE, multi-instance learning for DSRE is proposed (Riedel et al., 2010), followed by various noise-reduction methods. Some methods only select valuable sentences and drop the rest (Zeng et al., 2015; Qin et al., 2018; Feng et al., 2018). For better information utilization, sentence-level attention is applied by Lin et al. (2016) to dynamically reduce the weight of noisy sentences. Yuan et al. (2019a) down-weights the sentences with low similarity to the best sentence in the bag. As an attempt to alleviate noisy bag problem, Yuan et al. (2019b) and Ye and Ling (2019) employ bag-level attention under each superbag. There are also soft label methods (Liu et al., 2017; Wang et al., 2018) that avoid using noisy relation labels. However, explicitly distinguishing high-quality (easy) samples from low-quality (hard/noisy) ones remains a challenge for DSRE. Moreover, overfitting of hard/noisy samples during training is not discussed in previous work.

According to Pleiss et al. (2020) and Huang et al. (2021), the logit matrix can be utilized to distinguish easy samples from hard/noisy ones. Furthermore, we apply the Logit Margin (Huang et al., 2021) as the reference for sample quality to emphasize easy samples during training and avoid overfitting of hard/noisy samples.

3 Methodology

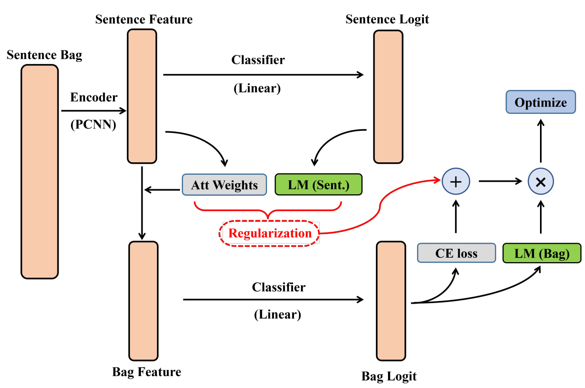

Figure 1: The overall procedure of our method.

As shown in Figure 1, our methods is a two-level approach. At sentence level, we calculate the LM score of the sentence and use it as the reference for dynamically-learned weight distribution. At bag level, we leverage the LM score of the bag as the priority weight for optimization. Further details will be discussed in this section.

3.1 Input Representations

The representation of each word in the sentence consists of two parts: word embedding and position embeddings. Each word is first mapped into a d_w-dimensional word embedding $v_j \in R^{d_w}$. To describe the relative distance to the two entities, the position embeddings $p_j^{e1}, p_j^{e2} \in R^{d_p}$, proposed by Zeng et al. (2014), are concatenated with the word embedding to form the representation of each word $w_j = [v_j; p_j^{e1}; p_j^{e2}]$ of $d_w + 2d_p$ dimensions.

3.2 Sentence Encoder

The sentence encoder in our model can be employed as a variety of neural encoders such as CNNs and RNNs. Since the Piecewise Convolution (PCNN) layer (Zeng et al., 2015) is widely used in previous work, we employ it as the default sentence encoder. The PCNN contains a convolution layer and a piecewise max-pooling layer. The input sentence is processed by a CNN with d_c filters and window size l. Then, piecewise max-pooling is adopted to extract features from the three segments of CNN outputs, which are segmented by the positions of the two entities. Finally, the sentence representation $s \in R^{3d_c}$ is obtained by concatenating the max-pooled outputs of the three segments.

3.3 Classifier

Our model follows Lin et al. (2016) and uses soft attention over the sentences (ATT) in multi-instance learning layer. The attention weight α_i for the i_{th} sentence is calculated using the bilinear form:

$$e_i = s_i A r \tag{1}$$

$$\alpha_i = \frac{exp(e_i)}{\sum_j exp(e_j)} \tag{2}$$

where A is a weighted diagonal matrix, and r is the query vector indicating the relation. For each entity pair, the logit score for the bag o is calculated from the bag representation x, which is the weighted sum of sentence representations:

$$x = \sum_i \alpha_i s_i \tag{3}$$

$$o = Mx + d \tag{4}$$

where M is the representation matrix of the relations and d is the bias vector.

3.4 Emphasis on Easy Samples (EES)

The goal of EES is to prioritize the learning of easy samples and avoid overfitting hard/noisy samples. The first step is to distinguish them. As observed in Pleiss et al. (2020), the model fits easy samples better than hard/noisy ones, especially in the early epochs. Such difference is reflected in the logit matrix, where easy samples have prominent values on the logit corresponding to the label relation. Therefore, we utilize the difference between the logit value of the label relation and the maximum logit of other relations, which is the Logit Margin (LM) (Huang et al., 2021), to distinguish easy samples from hard/noisy ones. The LM is calculated as follows:

$$LM = o_{j^*} - \max_{j \neq j^*} o_j \tag{5}$$

where j^* is the given DS label. The easy samples tend to have higher LM scores than hard samples. In contrast, The LM scores of noisy samples are more likely to be negative.

At sentence level, we hope that easy sentences have larger proportion in the weight distribution. Focusing on the easy sentences is also consistent with the at-least-one assumption (Riedel et al., 2010). Therefore, we design a regularization term to minimize the difference D between relative magnitude of LM and the distribution of attention weights:

$$\alpha_i^{LM} = \frac{exp(LM_i^{sen})}{\sum_j exp(LM_j^{sen})} \tag{6}$$

$$D = KL(\alpha^{LM}, \alpha) \tag{7}$$

where LM_i^{sen} represents the LM of i-th sentence in the bag. The difference D is calculate as the KL-divergence between α and α^{LM}, where α^{LM} is the target distribution.

At bag level, to prioritize the learning of easy bags, we introduce a priority weight based on the LM score of the bag. The calculation is simple:

$$W_i = exp(LM_i^{bag}) \tag{8}$$

where LM_i^{bag} is the LM value for the i-th bag. Since easy bags have larger LM scores, their priority weights are much larger than hard/noisy bags. Thus, the easy bags are prioritized in the optimization. In contrast, the priority weights of noisy bags are very small due to exp of negative values. The LM scores of hard bags are generally close to 0, so the magnitude of weight is dynamically controlled in a viable range (near 1).

3.5 Loss Function

Our model aims to maximize the conditional probability for the target relation given the sentence bag of the entity pair:

$$p(y_i|s,\theta) = \frac{o_i}{\sum_j exp(o_j)} \tag{9}$$

With Emphasis on Easy Sample, the loss function is implemented as cross entropy with priority weight W on sentence bag and regularization term D on weight distribution within the bag:

$$L(s_j,\theta) = W_j(-\sum_i logp(y_{ji}|s_j,\theta) + D_j) \tag{10}$$

$$L(\theta) = \sum_j L(s_j,\theta) + \beta||\theta||^2 \tag{11}$$

where β is a hyper-parameter to restrict the L_2 regularization.

4 Experiments

Experiments are conducted on widely used NYT-10 (Riedel et al., 2010) benchmark to test our approach. We first introduce the details of dataset and experiment settings before presenting our results.

4.1 Dataset and Settings

NYT-10 is a standard dataset constructed by aligning relation facts in Freebase (Bollacker et al., 2008) with the New York Times corpus. It has 281k training entity pairs , 97k testing entity pairs and 53 relation classes.

Parameter	Value
Batch size b	128
Word embedding size d_w	50
Position embedding size d_p	5
Sentence length l	70
Hidden size d_c	230
Window Size l	7
Learning rate lr	0.001
Dropout probability pr	0.3
L_2 penalty β	1e-04

Table 1: Parameter settings.

The hyper-parameters are shown in Table 1. In the experiments, we use Adam(Kingma and Ba, 2014) optimizer to optimize our model. We compare the models in terms of precision at top N predictions(P@N) and precision-recall curve.

Methods	One				Two				All			
	100	200	300	mean	100	200	300	mean	100	200	300	mean
(Lin et al., 2016)	73.3	69.2	60.8	67.8	77.2	71.6	66.1	71.6	76.2	73.1	67.4	72.2
PCNN+ATT (ours)	79.0	71.5	61.6	70.7	81.0	74.5	67.6	74.4	84.0	75.5	71.3	76.9
PCNN+ATT+D	79.0	71.0	64.0	71.3	83.0	75.5	70.0	76.1	84.0	76.0	72.3	77.4
PCNN+ATT+W	79.0	73.0	**67.0**	73.0	85.0	78.0	70.3	77.8	88.0	85.0	77.6	83.6
PCNN+ATT+EES	**84.0**	**78.0**	66.7	**76.2**	**87.0**	**80.5**	**75.3**	**80.9**	**92.0**	**86.0**	**78.7**	**85.6**

Table 2: P@N values of the models on NYT-10. **Bold** numbers indicate the best results. One/Two/All means randomly selecting one/two/all sentence(s) in each testing entity pair with more than one sentence.

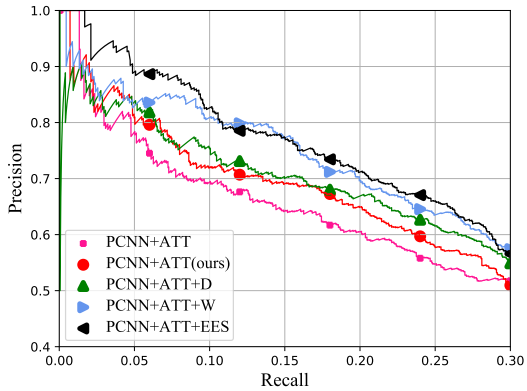

Figure 2: The precision-recall curve of the models.

4.2 Comparison with Previous Work

In the experiments, we select the widely used PCNN+ATT (Lin et al., 2016) model as the baseline. We implement PCNN+ATT using our own settings and achieve better performance than the original paper. We repeat the training multiple times and report the median. As shown in Table 2 and Figure 2, the PCNN+ATT+EES achieves significantly better results comparing with PCNN+ATT, indicating that the model trained with EES generalize better to test set. It is because that EES prevents overfitting on low-quality samples and improves the generalization of the model. Note that our implementations use the same set/amount of parameters, which means that the improvement comes solely from better training.

4.3 Ablation Study

To further explore the effects of the components, we conduct ablation study using two variants: PCNN+ATT+D and PCNN+ATT+W. D indicates the regularization on intra-bag weight distribution and W is the priority weighting on sentence bags. The result shows that both weight distribution regularization D and priority weighting W improve the overall performance. Although weight distribution regularization seems less effective, in practice, the training of model is much slower without it. The reason is that without explicitly emphasizing easy sentences, the model may make false prediction based on hard/noisy sentences in the bag. Therefore, combining both D and W is strongly recommended.

5 Conclusions and Future Work

In this paper, we propose a two-level Emphasis on Easy Samples mechanism to improve the generalization of DSRE model. At sentence level, the regularization term on intra-bag weight distribution is employed to emphasize high-quality sentences in the bag. At bag level, we apply the priority weight to promote the learning of high-quality sentence bags. The experimental results show that our approach significantly improves the generalization of the model on unseen data.

In the future, we will conduct more experiments using other existing frameworks. There are still some limitations, for example, the regularization on attention weight distribution is not applicable to non-attentive methods such as reinforcement learning. In addition, the start-up of our model is slower because the LM scores are generally low in the early stage of training.

References

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.

Jun Feng, Minlie Huang, Li Zhao, Yang Yang, and Xiaoyan Zhu. 2018. Reinforcement learning for relation classification from noisy data. *arXiv preprint arXiv:1808.08013*.

Xiusheng Huang, Yubo Chen, Shun Wu, Jun Zhao, Yuantao Xie, and Weijian Sun. 2021. Named entity recognition via noise aware training mechanism with data filter. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4791–4803.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv: Learning*.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2124–2133.

Tianyu Liu, Kexiang Wang, Baobao Chang, and Zhifang Sui. 2017. A soft-label method for noise-tolerant distantly supervised relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1790–1795.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011.

Geoff Pleiss, Tianyi Zhang, Ethan R Elenberg, and Kilian Q Weinberger. 2020. Identifying mislabeled data using the area under the margin ranking. *arXiv preprint arXiv:2001.10528*.

Pengda Qin, Weiran Xu, and William Yang Wang. 2018. Robust distant supervision relation extraction via deep reinforcement learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2137–2147.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 148–163. Springer.

Guanying Wang, Wen Zhang, Ruoxu Wang, Yalin Zhou, Xi Chen, Wei Zhang, Hai Zhu, and Huajun Chen. 2018. Label-free distant supervision for relation extraction via knowledge graph embedding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2246–2255.

Zhi-Xiu Ye and Zhen-Hua Ling. 2019. Distant supervision relation extraction with intra-bag and inter-bag attentions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2810–2819.

Changsen Yuan, Heyan Huang, Chong Feng, Xiao Liu, and Xiaochi Wei. 2019a. Distant supervision for relation extraction with linear attenuation simulation and non-iid relevance embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7418–7425.

Yujin Yuan, Liyuan Liu, Siliang Tang, Zhongfei Zhang, Yueting Zhuang, Shiliang Pu, Fei Wu, and Xiang Ren. 2019b. Cross-relation cross-bag attention for distantly-supervised relation extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 419–426.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1753–1762.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers*, pages 2335–2344.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. Understanding deep learning requires rethinking generalization.