Neural Networks 142 (2021) 213–220

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Self-selective attention using correlation between instances for distant supervision relation extraction

Yanru Zhou [1], Limin Pan [1], Chongyou Bai [1], Senlin Luo *, Zhouting Wu

Information System and Security & Countermeasures Experimental Center, Beijing Institute of Technology, Beijing 100081, China

ARTICLE INFO

Article history:
Received 2 November 2020
Received in revised form 18 March 2021
Accepted 23 April 2021
Available online 28 April 2021

Keywords:
Distant supervision relation extraction
Convolution neural network
Self-attention mechanism

ABSTRACT

Distant supervision relation extraction methods are widely used to extract relational facts in text. The traditional selective attention model regards instances in the bag as independent of each other, which makes insufficient use of correlation information between instances and supervision information of all correctly labeled instances, affecting the performance of relation extractor. Aiming at this problem, a distant supervision relation extraction method with self-selective attention is proposed. The method uses a layer of convolution and self-attention mechanism to encode instances to learn the better semantic vector representation of instances. The correlation between instances in the bag is used to assign a higher weight to all correctly labeled instances, and the weighted summation of instances in the bag is used to obtain a bag vector representation. Experiments on the NYT dataset show that the method can make full use of the information of all correctly labeled instances in the bag. The method can achieve better results as compared with baselines.

1. Introduction

Relation extraction is mainly to extract semantic relation between entities, which is a very basic task in the field of information extraction. It has important applications in knowledge base building (Heng & Ralph, 2011) and question answering system (Yu et al., 2017).

The supervision relation extraction system(Mooney & Bunescu, 2006) requires a large number of manually labeled datasets, and manual labeling of data requires a lot of time and effort. To reduce the reliance on manually labeled data, 2009, Mintz et al. (2009) use distant supervision to automatically label data. Distant supervision assumes that if there is a relation between two entities in the knowledge base, all sentences containing these two entities will express this relation; if there is no relation between two entities in the knowledge base, the sentence will be labeled as NA (Not A relation). Distant supervision method can automatically label large-scale training data. The training data labeled by this method is successfully applied to the joint extraction of entities and relations (Bai, Pan, et al., 2020). However, the method assumption is too strong and cause the problem of incorrect labeling, which affect the performance of the relation extractor.

On purpose of eliminating the interference of noise instances, a multi-instance learning framework is introduced into the task of distant supervision relation extraction. Under the multi-instance learning framework, some methods (He et al., 2017; Hoffmann et al., 2011; Riedel, Yao, & McCallum, 2010; Surdeanu et al., 2012; Zeng et al., 2015) use the "at-least-one" assumption to select the most confident instances in the bag as positive instances to train model, the method loses a lot of information of correctly labeled instances.

For similar reasons, for the scene where the training set is disturbed by noise, a sentence-level selective attention model (Alt, Hübner, & Hennig, 2019; Hu et al., 2019; Ji et al., 2017; Lin et al., 2016; Liu et al., 2018; Wu, Bamman, & Russell, 2017; Xing & Luo, 2019) is proposed. The method is used to dynamically reduce the weight of noise instances in the bag and uses correctly labeled instances to calculate the bag vector representation. The method still has two defects. First, the multi-instance learning method puts sentences with the same entity pair into the same bag. These instances with the same entity pair in the bag have a certain connection. This is important supervision information for instances in the bag. However, the sentence-level selective attention model assumes that instances in the bag are all independent and uniformly distributed, the correlation between instances in the bag is ignored, which leads to the loss of supervisory information. Second, the sentence-level attention model depends on the target relation to select effective instances in the bag, which makes the weight of some correctly labeled instances in the bag very small, resulting in the loss of supervision information of the correctly labeled instances.

* Corresponding author.
E-mail addresses: zhouyrbit@gmail.com (Y. Zhou), luosenlin2019@126.com (S. Luo).

[1] The authors contribute equally.

https://doi.org/10.1016/j.neunet.2021.04.032
0893-6080/

Table 1
The weight distribution of instances in the bag calculated by the traditional attention model.

Triplet		Instances	Noise?	Weight
⟨*/people/person/place_lived*, *chingy*, **st._louis**⟩	S1	it sounds a little like the **st._louis** rapper **chingy**'s mega-single "right thurr".	No	0.980
	S2	chingy after a string of pop hits, the **st._louis** rapper **chingy** split with his former label boss, ludacris.	No	0.003
	S3	it's a good time for **st._louis** thanks to the doors that nelly and **chingy** and j-kwon were able to open.	Yes	0.017

To illustrate the second defect, a bag is randomly selected, and the weight distribution of instances in this bag is calculated using the traditional selective attention model. The calculation results are shown in Table 1, where bold words indicate entities and italic words indicate relation types. There are three instances in this bag. According to the assumption of distant supervision, all instances in this bag can express the relation type of "/people/person/place_lived". The semantics of the third instance cannot indicate that the relation exists between these two entities, so this instance is a noise instance.

Aiming at reducing the interference of noise instances, attention mechanisms are used to dynamically decrease the weights of noise instances. It can be seen from Table 1 that the weight of noise instance calculated by the traditional attention model is 0.017, and the method can indeed reduce the influence of noise instances. The weight of the second correctly labeled instance S2 is only 0.003, which shows that the weight distribution calculated by the method is not accurate. The method fails to pay attention to all correctly labeled instances, resulting in the loss of supervision information of correctly labeled instances.

Aiming at above problems, a distant supervision relation extraction method with self-selective attention is proposed. Riedel et al. (2010) find that the precision of aligning the relations in Freebase to the New York Times corpus is about 70%. It is inferred that most of instances in the bag are correctly labeled. The correctly labeled instances in the same bag can express the same relation type, so the more similar their semantic vector representation. When most of instances in the bag are labeled correctly, instances that are similar to most of instance vectors in the bag are more likely to be correct instances. Therefore, our model uses the similarity between instance vectors to measure the correlation between instances in the bag, and assigns higher weight to instances similar to other instances in the bag. In addition, to better learn the semantic vector representation of sentences, convolution and self-attention mechanisms are used instead of PCNN and CNN as sentence encoder.

The contributions of this paper as follows:

(1) A distant supervision relation extraction method with self-selective attention is proposed. This method uses the correlation between instances in the bag to assign a higher weight to all correctly labeled instances in the bag, which can pay attention to all correctly labeled instances in the bag.

(2) Experiment on the widely used NYT dataset. Experimental results show that SelfCON + SATT model with self-selective attention module outperforms PCNN + ATTRA + BAGATT model (Ye & Ling, 2019) with 3.1% improvement.

2. Related work

Relation extraction is an important task in the field of natural language processing. Some early works regard relation extraction as a supervised learning task. Due to the lack of large-scale manual labeled data, a distant supervision relation extraction (Mintz et al., 2009) method is proposed. The method generates the relation labels of the entity pairs by automatically aligning the original text with the knowledge base. The method has the problem of noise label. To alleviate this problem, early distant supervision relation extraction methods use multi-instance learning (Riedel et al., 2010) and multi-instance multi-label learning (Hoffmann et al., 2011; Surdeanu et al., 2012) to model the assumption that "at least one sentence can correctly express the relation type for each bag". The performance of these methods relies heavily on the features of manual design.

With the development of deep neural networks, many neural network models are used for distant supervision relation extraction. Zeng et al. (2015) used PCNN to learn the vector representation of sentences under multi-instance learning framework, and select the vector representation of the highest confident instance of the bag as bag vector representation. Lin et al. (2016) used a sentence-level selective attention model to dynamically reduce the influence of noise instances. Ji et al. (2017) adopt a similar attention strategy and introduced the feature of entity description to calculate the weight of instance. Liu et al. (2018) used subtree parsing method and entity-wise attention mechanism to reduce noise in sentences; Xing and Luo (2019) used independent head-to-tail convolutional neural networks to encode sentences, and filter out instances with no actual relation to alleviate noisy data problems; Alt et al. (2019) used a pre-trained language model-Transformer to capture semantic features, syntactic features and a lot of "common sense" knowledge to identify more relation types. Liu et al. (2017) proposed a soft label method to reduce the influence of noise instances. Vashishth et al. (2018) used graph convolutional networks to encode grammatical information, and used additional information in the knowledge base to improve the performance of relation extraction. Bai, Jin, et al. (2020) proposed a method for integrating entity type information into a neural network based distantly supervised relation extraction model. Instead of using selective attention mechanism, Li et al. (2020) designed a pooling-equipped gate mechanism as an aggregator to generate bag-level representation, which performs stably and promisingly even if only one sentence appears in a bag and thus keeps the consistency across all training examples. Shang et al. (2020) propose a novel method for distant supervised relation extraction, which employs unsupervised deep clustering to generate reliable labels for noisy sentences, which will be transformed into useful training data and benefit the model's performance. Yu et al. (2020) formulated distantly supervised relation extraction as a hierarchical classification task and propose a novel hierarchical classification framework, which extracts the relation in a top-down manner. These methods all use the weighted sum of instance vectors in the bag to represent bag, and use bag vector representation to calculate the probability of classifying the bag into each relation during training.

In addition, some methods are proposed to solve the case where instances in the bag are all noise. Feng et al. (2018) proposed a sentence-level relation classification model based on reinforcement learning. The model can select high-quality sentences and filter out noisy sentences. The reward of the reward function in this model is calculated according to the predicted probability of relation classifier. Qin, Xu, and William (2018b) also tried to use a deep reinforcement learning framework to automatically identify noise instances. In this method, the reward of the reward function is calculated according to the performance change of relation classifier. Qin, Xu, and William (2018a)

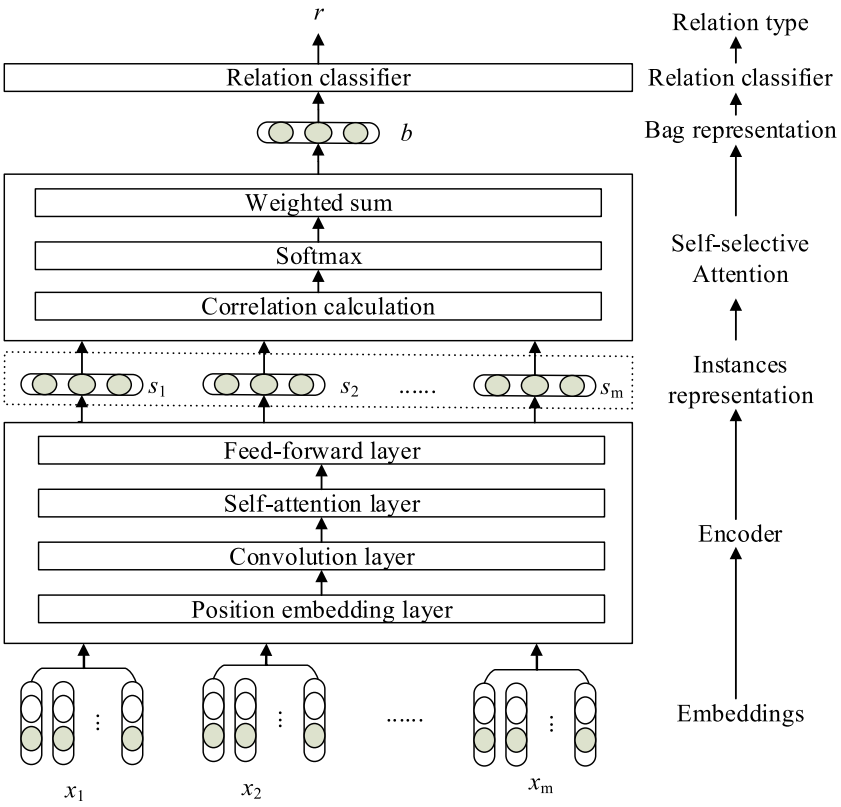

Fig. 1. The principle framework of self-selective attention using correlation between instances for distant supervision relation extraction.

used a generative adversarial network to automatically identify noise instances in each relation type without any supervision information. Ye and Ling (2019) proposed a model that combines intra-bag and inter-bag attention. The model considers both sentence-level noise and bag-level noise. Yuan et al. (2019) proposed a non-IID relevance attention module that weights of instances in the bag are calculated by the relevance of instances with a single instance, which best represents the relationship selected via multi-instance learning (MIL).

However, above methods all use the traditional selective attention model, which makes insufficient use of correlation information between instances and supervision information of all correctly labeled instances, affecting the performance of relation extractor. Different from these methods, a self-selective attention method is proposed, which uses the correlation between instances in the bag to assign a higher weight to the correctly labeled instances in the bag, so it can pay attention to all correctly labeled instances.

3. Method

A distant supervision relation extraction method with self-selective attention is proposed. The model is modeled under multi-instance learning framework, which mainly includes 4 modules: Embeddings, Encoder, Self-selective attention, and Relation classifier. The principle framework of the model is shown in Fig. 1.

Embeddings first obtain the word vector of each word in the input sentence by pre-training; then, to specify the position of entity pair in sentence, position vector needs to be added to the word vector; finally, the vector representation of all words in the input sentence is stacked into an embedding matrix x. The encoder first reads the embedding matrix x and two entities; then it learns the vector representation s of instance through convolution and self-attention mechanisms. The self-selective attention module first puts the vector representations of instances with the same entity pair into the same bag; then calculates the weights of instances in the bag based on the correlation between instances in the bag; and finally, bag vector representation is calculated by weighted summation. Unlike the traditional selective attention model, this method does not use target relations to select instances, and selects correctly labeled instances based on the correlation between instances in the bag. The relation classifier first reads bag vector representation; then predicts the relation type on bag vector representation.

3.1. Embeddings

The input of encoder is original word, and each word in the input needs to be converted into a low-dimensional real-valued vector representation. In addition, to specify the position of entity pair, it is also necessary to use position embedding for all words in the input.

For the ith instance $x_i = \{w_1, w_2, \ldots, w_n\}$ in the bag $B = \{x_1, x_2, \ldots, x_m\}$, where, m represents the number of instances in the bag, n represents the length of instance. w_j represents the vector representation of the jth word in instance, which consists of two parts: a word vector and a position vector. First, each word in instance is mapped to a vector representation with dimension d_w by looking up the embedded representation matrix trained by Word2vec tool; Then the position embedding method proposed by Zeng et al. (2014) is used to calculate the vector representation of the relative distance of each word in sentence from the head and tail entities, the dimension is d_p; finally, the word vector and position vector corresponding to each word in sentence are connected together to form a new vector representation w_j, dimension $d = d_w + 2d_p$.

3.2. Encoder

The encoder is used to learn the semantic vector representation $s_i \in R^{d_c}$ of instance x_i in the bag, as shown in Eq. (1), where d_c represents the dimension of the encoder.

$$s_i = \mathrm{f}(x_i) \tag{1}$$

where, f represents the functions of encoder.

The model does not use CNN or PCNN to encode instances, and uses self-attention mechanism and convolution (SelfCON) to learn better vector representations of instances. The encoder includes 4 layers, which are position embedding layer, convolution layer, self-attention layer, and feed-forward layer. There is a layer normalization between each layer. The dropout (Srivastava et al., 2014) is used between the convolution layer, the self-attention layer, and the feed-forward layer to avoid the model overfitting. The self-attention mechanism cannot learn the position information of the words in sequence, so a position embedding needs to be added at the beginning of the encoder. Position embedding uses the method proposed by Vaswani et al. (2017) to encode the position information of each word. The convolution layer uses a deep separable convolutional neural network with better memory efficiency and better generalization ability (Francois, 2017). The self-attention layer uses a multi-head attention mechanism (Vaswani et al., 2017). The feed-forward layer consists of a convolution operation and a max-pooling operation.

3.3. Self-selective attention

In distant supervision relation extraction, the traditional selective attention model uses target relation labeled by distant supervision method as "query" to select correctly labeled instances, without considering the correlation between instances in the bag. If two instances in the bag can express the same relation type, the more similar the semantic vector representation of two instances. Therefore, our model uses the similarity between instance vectors to measure the correlation between instances in the bag, and assigns higher weight to instances similar to other instances in the bag. This method does not depend on the target relation when selecting instances, and only depends on instances in the bag to select correctly labeled instances, so it is called self-selective attention (SATT).

Given a bag $B = \{x_1, x_2, \ldots, x_m\}$ containing m instances, these instances contain the same entities pair. After all instances in the bag have been encoded, the vector representation s_i of each instance can be obtained. The bag $B = \{s_1, s_2, \ldots, s_m\}$, where, $s_i \in R^{d_c}$. To make full use of the information of correctly labeled instances in the bag, the correlation between instances is used to dynamically assign weights to instances in the bag, and the bag vector representation b is obtained by the weighted sum method. The calculation process is shown in Eq. (2).

$$b = \sum_{i=1}^{m} \alpha_i s_i \tag{2}$$

where, $b \in R^{d_c}$, α_i represents the weight corresponding to the ith instance in the bag, and its calculation process is shown in Eq. (3).

$$\alpha_i = \frac{\exp(e_i)}{\sum_{i=1}^{m} \exp(e_i)} \tag{3}$$

where, e_i represents the sum of the similarity between the ith instance and other instances in the bag. A simple dot product operation is used to calculate the similarity between instances in the bag. The calculation process is shown in (4).

$$e_i = \sum_{j=1,2,\ldots,m, j\neq i} \bar{s_i}\bar{s_j}^T \tag{4}$$

where, instance vector representation s is normalized to $\bar{s}$, and the normalization process is shown in Eq. (5).

$$\bar{s} = s/\|s\|_2 \tag{5}$$

3.4. Relation classifier

After bag vector representation $b \in R^{d_c}$ is calculated, the relation classifier predicts relation at bag level.

Firstly, the dropout strategy is applied to bag representation b to prevent overfitting. Then, the similarity calculation is performed on the bag vector representation b and the relation matrix $r \in R^{h\times d_c}$ to obtain the confidence that bag is predicted as each relation type. The calculation process is shown in Eq. (6).

$$o = br^T + a \tag{6}$$

where, $a \in R^h$ represents the bias vector, h represents the number of relations.

Finally, the Softmax function is used to obtain the conditional probability of predicting bag B as the kth relation type. The calculation process is shown in Eq. (7).

$$p(k|B,\theta) = \frac{\exp(o_k)}{\sum_{i=1}^{h} \exp(o_i)} \tag{7}$$

where, θ represents all trainable parameters in the model.

The loss function is defined as the cross-entropy, and the calculation process is shown in Eq. (8).

$$L(\theta) = \sum_{i=1}^{D} \log(p(k|B_i,\theta)) \tag{8}$$

where D represents the number of bags in training set and θ represents all trainable parameters in the model. Stochastic gradient descent is used to update the parameters of the model.

3.5. Training and test

When training, the self-selective attention module is used to calculate the attention weight of instances in the bag, and the vector representation of instances in the bag is weighted and summed to obtain bag representation b.

When testing, as the method proposed by Lin et al. (2016), the vector representation of all relations is used as "query" to calculate the posterior probability of each relation, and relation with the highest probability value is selected as the classification result.

4. Experiments

4.1. Dataset and evaluation metrics

The New York Times (NYT) dataset published by Riedel et al. (2010) is widely used in the study of distant supervision relation extraction (Wu et al., 2017; Ye & Ling, 2019). This dataset was generated by aligning Freebase (Bollacker et al., 2008) with the NYT corpus automatically. This dataset uses the sentences in 2005–2006 article as training set, and the sentences in 2007 article as test set. This dataset contains 53 relation types, including a NA (Not A relation) relation, indicating that there is no relation between two entities. The detailed statistics of this dataset is shown in Table 2.

The held-out test set of NYT dataset is used to evaluate the model and compare the relations predicted by the model with the relations in Freebase. The evaluation metrics use Precision/Recall (PR) curve, Area Under Curve (AUC) values and Top-N Precision (P@N).

Table 2
Statistics of NYT dataset.

	Number of sentences	Number of entity pairs	Number of triplets
Training set	570,088	291,699	19,429
Test set	172,448	96,678	1,950

Table 3
Parameter settings.

Component	Parameter	Value
Word embedding	Dimension	50
Position embedding	Dimension	5
Self-attention	Head number	8
Dropout	Dropout rate	0.9
Optimization	Batch size	32
	Learning rate	0.01
CNN	Window size	7
	Filter number	400

4.2. Hyperparameters

All parameter settings of the model are shown in Table 3. The word embedding matrix is published by Lin et al. (2016).

4.3. Overall performance

Eight different models are selected and compared with SelfCON + SATT model proposed in this paper.

Mintz (Mintz et al., 2009) (2009), MultiR (Hoffmann et al., 2011) (2011) and MIMLRE (Surdeanu et al., 2012) (2012) are three feature-based methods, PCNN + ATT (Lin et al., 2016) (2016) and PCNN + ATT + soft-label (Liu et al., 2017) (2017) are two methods with selective attention mechanism,[2] PCNN + ATT + RL (Qin et al., 2018b) (2018), PCNN + ATT + GAN (Qin et al., 2018a) (2018) 和 PCNN + ATTRA + BAGATT (Ye & Ling, 2019) (2019) are three new methods in distant supervision relation extraction.

Where, PCNN means that encoder uses piecewise convolutional neural network, ATT means selective attention method proposed by Lin et al. (2016), SelfCON means that encoder uses self-attention mechanism and convolution, and SATT means self-selective attention.

4.3.1. PR curves

For visual comparison, the PR curve with only the top 2000 points of SelfCON + SATT and these 8 models is plotted, as shown in Fig. 2.

From the PR curve of Fig. 2, it can be seen that:

(1) Compared with other models, SelfCON + SATT model has better PR performance. It shows that SelfCON + SATT model uses the correlation between instances in the bag to assign weights to instances to make full use of the supervision information of all correctly labeled instances, thereby improving the performance of relation classifier.

(2) The overall performance of SelfCON + SATT model is more balanced. When Recall is small, PCNN + ATTRA + BAGATT model performs better, but after Recall is greater than 0.18, the Precision of this model drops faster than SelfCON + SATT model. It shows that SelfCON + SATT model has a more balanced overall performance.

4.3.2. AUC values

In addition, for quantitative comparison, the AUC values of PCNN + ATT + RL (Qin et al., 2018b), PCNN + ATT + GAN (Qin et al., 2018a), PCNN + ATTRA + BAGATT (Ye & Ling, 2019) and SelfCON + SATT model is compared, and the experimental results are shown in Table 4.

Table 4
AUC value of SelfCON + SATT and 3 comparison models.

Model	AUC
PCNN + ATT + RL (Qin et al., 2018b)	0.281
PCNN + ATT + GAN (Qin et al., 2018a)	0.281
PCNN + ATTRA + BAGATT (Ye & Ling, 2019)	0.312
SelfCON + SATT	**0.320**

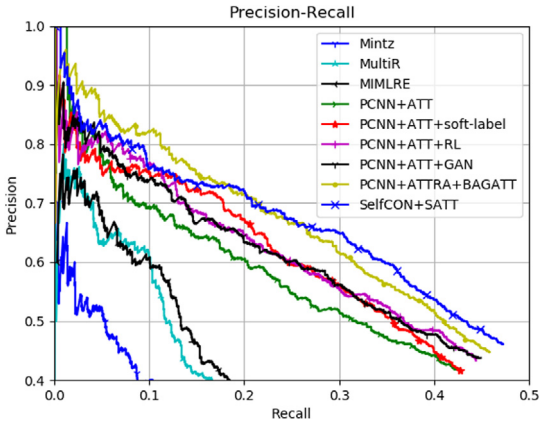

Fig. 2. PR curve of SelfCON + SATT and 8 comparison models.

From the experimental results in Table 4, SelfCON + SATT model achieves the highest AUC value of 0.320 (the AUC value of PR curve with the top 2000 points). Compared with the PCNN + ATT + RL model and the PCNN + ATT + GAN model, the AUC value increased by 3.9%. Compared with the PCNN + ATTRA + BAGATT model, it is improved by 0.8%. It shows that SelfCON + SATT model can make full use of the supervision information of all correctly labeled instances by using the correlation between instances in the bag, thereby improving the performance of relation classifier.

4.4. P@N

Like Lin et al. (2016), we also evaluate the proposed model on entity pairs with multiple instances. First three different test sets are constructed:

One: For each bag in the original test set, randomly select one instance for relation classification;

Two: For each bag in the original test set, randomly select two instances for relation classification;

All: For each bag in the original test set, all instances are used for relation classification.

The P@100, P@200, P@300 values and their mean values of the models on these three test sets are shown in Table 5.

From the experimental data in Table 5, it can be seen that:

(1) On the One and Two test sets, SelfCON + SATT obtains the best P@N value.

(2) When CNN and PCNN are used as encoders, the effect of SATT method is better than ATT method on these three test sets; When SelfCON is used as encoder, SATT method is better than ATT

[2] The PR curve data of these five methods come from https://github.com/tyliupku/soft-label-RE.

Table 5
P@N values of entity pairs with different number of instances.

# of instances	One				Two				All			
P@N (%)	100	200	300	Mean	100	200	300	Mean	100	200	300	Mean
CNN + ATT	76	71.5	65.7	71.1	80	78	72.3	76.8	81	76.5	74	77.2
CNN + SATT	79	72	67.3	72.7	86	80	75.6	80.5	86	81	76	80.9
PCNN + ATT	76	72	65.3	71.1	85	77	70.7	77.6	84	77.5	72.3	77.9
PCNN + SATT	**86**	74.5	68.6	76.3	89	83	**76.6**	82.8	90	**85**	78.3	84.4
SelfCON + ATT	80	73	65.7	72.9	84	82.5	73.7	80.1	**91**	84.5	**79**	**84.8**
SelfCON + SATT	84	**78.5**	**73.3**	**78.6**	**90**	**83.5**	75.7	**83.1**	88	83	**79**	83.3

Table 6
AUC values of whole PR curves of different models.

Model	AUC
CNN + ATT	0.384
CNN + SATT	0.405
PCNN + ATT	0.399
PCNN + SATT	0.424
PCNN + ATTRA + BAGATT (Ye & Ling, 2019)	0.422
SelfCON + ATT	0.426
SelfCON + SATT	**0.453**

method on the One and Two test sets. This is because the SATT method can make full use of the supervision information of all correctly labeled instances by using the correlation between instances in the bag, thereby improving the performance of relation classifier.

(3) Whether using ATT method or SATT method, SelfCON encoder has better experimental results than PCNN and CNN. It shows that using convolution and self-attention mechanisms as encoders can learn the better semantic features.

(4) With the increase in the number of instances, the performance of all models will be improved, which shows that more information is used.

4.5. Effects of self-selective attention

Seven models are implemented to verify the effectiveness of self-selective attention module. The names of these models are shown in Table 6. For visual comparison, the whole PR curves of these 7 models are plotted in Figs. 3–5, respectively. For quantitative comparison, the AUC values of the whole PR curves of these models are listed in Table 6. From the results of Figs. 3, 4, 5 and Table 6, it can be seen that:

(1) Compared with CNN + ATT model, the AUC value of CNN + SATT model has increased by 2.1%. Compared with PCNN + ATT model, the AUC value of PCNN + SATT model has increased by 2.5%. Compared with SelfCON + ATT model, the AUC value of SelfCON + SATT model has increased by 2.7%. Experimental results prove that SATT method is better than ATT method. SATT method uses the correlation between instances in the bag to calculate the attention weight, which can learn a higher weight for all correctly labeled instances and can pay attention to all correctly labeled instances in the bag.

(2) SelfCON + SATT model achieves the best AUC value of 0.453. Compared with PCNN + ATTRA + BAGATT, the AUC value increased by 3.1%.

4.6. Case study

To further verify the effectiveness of self-selective attention module. The self-selective attention module (SATT) and the traditional selective attention method (ATT) are used to calculate the weights of all instances in a randomly selected bag. The results are shown in Table 7.

In the first bag, these two entities corresponding to the bag are “chingy” and “st._louis”, and the relation type between these

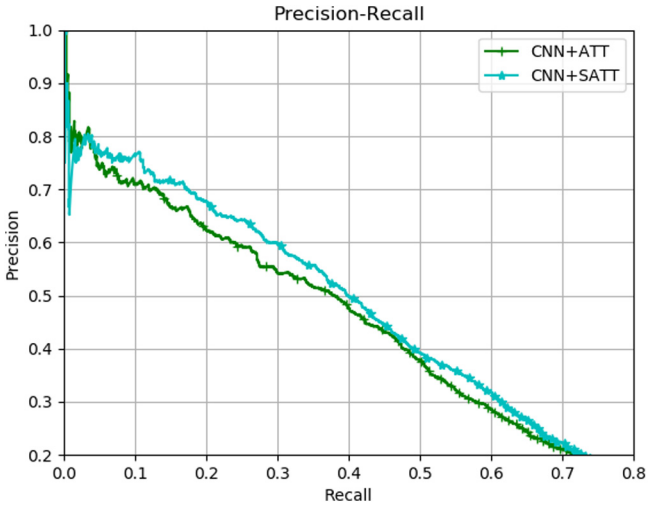

Fig. 3. PR curves of CNN + ATT and CNN + SATT models.

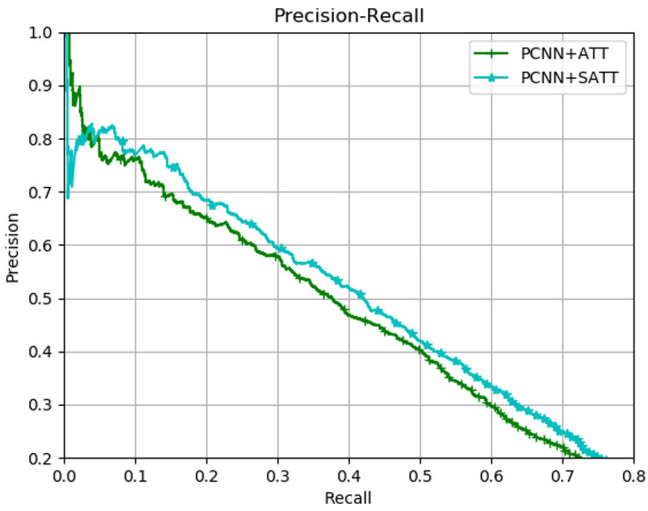

Fig. 4. PR curves of PCNN + ATT and PCNN + SATT models.

two entities is “/people/person/place_lived”. The instance S3 is a noise instance because S3 can’t express the relation “/people/person/place_lived” between “chingy” and “st._louis”. In the Second bag, these two entities corresponding to the bag are “jane_jacobs” and “toronto”, and the relation type between these two entities is “/people/deceased_person/place_of_death”. There are as many instances of noise as there are correct instances. While in the third bag, there is only one correct instance in the third package. These two entities corresponding to the bag are “hermine_braunsteiner” and “vienna”, and the relation type between these two entities is “/people/person/place_of_birth”.

As shown in Table 7, for the first case where there is only one instance of noise, the weight of noise instance S3 calculated by

Table 7
The weight distribution of instances in the bag calculated by SATT and ATT.

Triplet		Instances	Noise?	ATT	SATT
⟨*/people/person/place_lived*, chingy,st._louis⟩	S1	it sounds a little like the **st._louis** rapper **chingy**'s mega-single "right thurr".	No	0.9798	0.3556
	S2	chingy after a string of pop hits, the **st._louis** rapper **chingy** split with his former label boss, ludacris.	No	0.0032	0.3533
	S3	it's a good time for **st._louis** thanks to the doors that nelly and **chingy** and j-kwon were able to open.	Yes	0.0170	0.2911
⟨*/people/deceased_person/ place_of _death*, **jane_jacobs**, **toronto**⟩	S4	**jane_jacobs**, the writer and thinker who brought penetrating eyes and ingenious insight to the sidewalk ballet of her own greenwich village street and came up with a book that challenged and changed the way people view cities, died yesterday in **toronto**, where she moved in 1968.	No	0.9876	0.3758
	S5	alice sparberg alexiou, the author of the biography "**jane_jacobs**: urban visionary" lrbrutgers university press, will participate in a panel discussion based on the work of ms. jacobs, the urban planner who died in april in **toronto**.	No	0.0107	0.2507
	S6	dovercourt has a penchant for arriving at rock clubs and bars with books by the famed urban critic **jane_jacobs**, who has made **toronto** her home for nearly 40 years.	Yes	0.0011	0.1944
	S7	**jane_jacobs**, the activist who took him on, now lives in **toronto**.	Yes	0.0004	0.1789
⟨*/people/person/place_of _birth*, **hermine_braunsteiner**, **vienna**⟩	S8	the call was from russell ryan, the electrician who met **hermine_braunsteiner** in **vienna**, then took her to america as his wife, enabling her to become a citizen.	Yes	0.0001	0.2132
	S9	as a freshman general-assignment reporter in the newsroom of the times, my task that morning was to go to queens and check out a tip from simon wiesenthal, the renowned nazi hunter in **vienna**, that a notorious death-camp guard and convicted war criminal, **hermine_braunsteiner**, was now living there under the name mrs. ryan.	Yes	0.0191	0.2105
	S10	**hermine_braunsteiner** was born in vienna on july 16, 1919.	No	0.9808	0.5763

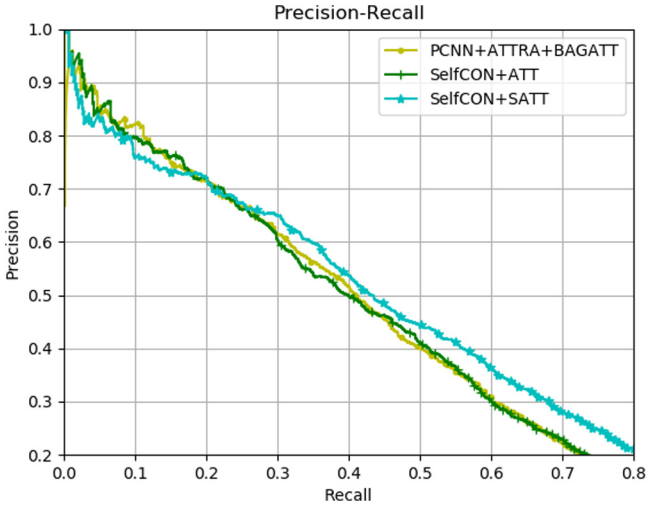

Fig. 5. PR curves of SelfCON + ATT, SelfCON + SATT and PCNN + ATTRA + BAGATT models.

ATT is 0.0170, and the weights of the correctly labeled instances S1 and S2 are 0.9798 and 0.0032, respectively. The weights of the correctly labeled instances S1 and S2 calculated by SATT are 0.3556 and 0.3533, respectively. The results show that ATT can only focus on one correctly labeled instance and ignore other correctly labeled instances, resulting in the loss of supervision information of other correctly labeled instances.

For the second case where there is half instance of noise, the weights of noise instances S6 and S7 calculated by ATT are 0.0011, 0.0004, and the weights of the correctly labeled instances S4 and S5 are 0.9876 and 0.0107, respectively. The weights of the correctly labeled instances S4 and S5 calculated by SATT are 0.3758 and 0.2507, respectively. It can be found that in the case of different noisy bags, our model (SATT) achieves appropriate weight allocation, and realizes the use of all correctly labeled instances.

For the last case where there is only one correctly labeled instance, the weights of noise instance S8 and S9 calculated as 0.0001, 0.0191 by ATT, and 0.2132, 0.2105 by SATT. The weight of the correctly labeled instance S10 is calculated as 0.9808 by ATT and 0.5763 by SATT. When there is only one correct label instance in the bag, and the others are all noise, our model can still achieve a relatively appropriate allocation.

Therefore, SATT can make full use of the supervision information of all correctly labeled instances, which helps to improve the performance of relation classifier.

5. Conclusion

In this paper, a distant supervision relation extraction method with self-selective attention is proposed to make full use of the correlation information between instances and the information of all correctly labeled instances in the bag. The method regards the distant supervision relation extraction task as a multi-instance learning task. First, a layer of convolution and self-attention mechanism are used as encoder to learn the better instances vector representation; then, self-selective attention module is used to dynamically assign weights to instances in the bag, and instance vectors in the bag are weighted and summed to obtain bag vector representation; finally, relation classification is performed on the bag vector representation. The experimental results on the NYT dataset show that our method can only use the correlation between instances in the bag to assign higher weight to all correctly labeled instances, which can focus on all correctly labeled instances, thereby improving the performance of relation classifier. However, our model cannot handle the case where all instances in a bag are noise. Therefore, future work will pay more attention to noise bags.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work supported by 242 National Projects (No. 2019A021).

References

Alt, C., Hübner, M., & Hennig, L. (2019). Fine-tuning pre-trained transformer language models to distantly supervision relation extraction. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 1388–1398). Stroudsburg, PA: Association for Computational Linguistics.

Bai, L., Jin, X., Zhuang, C., et al. (2020). Entity type enhanced neural model for distantly supervised relation extraction. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 13751–13752).

Bai, Chongyou, Pan, Limin, Luo, Senlin, et al. (2020). Joint extraction of entities and relations by a novel end-to-end model with a double-pointer module. *Neurocomputing*, *377*, 325–333.

Bollacker, K., Evans, C., Paritosh, P., et al. (2008). Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the ACM SIGMOD international conference on management of data* (pp. 1247–1250). New York, NY: Association for Computing and Machinery.

Feng, Jun, Huang, Minlie, Zhao, Li, et al. (2018). Reinforcement learning for relation classification from noisy data. In *Proceedings of the thirty-second AAAI conference on artificial intelligence* (pp. 5779–5786). Menlo Park, CA: Association for the Advancement of Artificial Intelligence.

Francois, Chollet (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1251–1258). Los Alamitos, CA/Washington/Brussels/Tokyo: IEEE Computer Society.

He, Dengchao, Zhang, Hongjun, Hao, Wenning, et al. (2017). Distant supervision relation extraction via long short term memory networks with sentence embedding. *Intelligent Data Analysis*, *21*(5), 1213–1231.

Heng, Ji, & Ralph, Grishman (2011). Base population: Successful approaches and challenges. In *The 49th annual meeting of the association for computational linguistics: human language technologies* (pp. 1148–1158). Stroudsburg, PA: Association for Computational Linguistics.

Hoffmann, R., Zhang, Congle, Ling, Xiao, et al. (2011). Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies-Volume 1* (pp. 541–550). Stroudsburg, PA: Association for Computational Linguistics.

Hu, Linmei, Zhang, Luhao, Shi, Chuan, et al. (2019). Improving distantly-supervision relation extraction with joint label embedding. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing* (pp. 3812–3820). Stroudsburg, PA: Association for Computational Linguistics.

Ji, Guoliang, Liu, Kang, He, Shizhu, et al. (2017). Distant supervision for relation extraction with sentence-level attention and entity descriptions. In *Thirty-first AAAI conference on artificial intelligence* (pp. 3060–3066). Menlo Park, CA: Association for the Advancement of Artificial Intelligence.

Li, Y., Long, G., Shen, T., et al. (2020). Self-attention enhanced selective gate with entity-aware embedding for distantly supervised relation extraction. In *Proceedings of the AAAI conference on artificial intelligence*. (pp. 8269–8276).

Lin, Yankai, Shen, Shiqi, Liu, Zhiyuan, et al. (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)* (pp. 2124–2133). Stroudsburg, PA: Association for Computational Linguistics.

Liu, Tianyu, Wang, Kexiang, Chang, Baobao, et al. (2017). A soft-label method for noise-tolerant distantly supervision relation extraction. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 1790–1795). Stroudsburg, PA: Association for Computational Linguistics.

Liu, Tianyi, Zhang, Xinsong, Zhou, Wanhao, et al. (2018). Neural relation extraction via inner-sentence noise reduction and transfer learning. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 2195–2204). Stroudsburg, PA: Association for Computational Linguistics.

Mintz, M., Bills, S., Snow, R., et al. (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: Volume 2-Volume 2* (pp. 1003–1011). Stroudsburg, PA: Association for Computational Linguistics.

Mooney, Raymond J., & Bunescu, Razvan C. (2006). Subsequence kernels for relation extraction. In *Advances in neural information processing systems* (pp. 171–178). Long Beach, CA: MIT Press.

Qin, Pengda, Xu, Weiran, & William, Yang Wang (2018a). Dsgan: Generative adversarial training for distant supervision relation extraction. In *Proceedings of the 56th annual meeting of the association for computational linguistics (Volume 1: Long papers)* (pp. 496–505). Stroudsburg, PA: Association for Computational Linguistics.

Qin, Pengda, Xu, Weiran, & William, Yang Wang (2018b). 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)* (pp. 2137–2147). Stroudsburg, PA: Association for Computational Linguistics.

Riedel, S., Yao, Limin, & McCallum, A. (2010). A modeling relations and their mentions without labeled text. In *Joint European conference on machine learning and knowledge discovery in databases* (pp. 148–163). Berlin, Heidelberg: Springer.

Shang, Y., Huang, H., Mao, X., et al. (2020). Are noisy sentences useless for distant supervised relation extraction? In *Proceedings of the AAAI conference on artificial intelligence* (pp. 8799–8806).

Srivastava, N., Hinton, G., Krizhevsky, A., et al. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(1), 1929–1958.

Surdeanu, M., Tibshirani, J., Nallapati, R., et al. (2012). Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 455–465). Stroudsburg, PA: Association for Computational Linguistics.

Vashishth, S., Joshi, R., Prayaga, S. S., et al. (2018). Reside: Improving distantly-supervised neural relation extraction using side information. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 1257–1266). Stroudsburg, PA: Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. In *Proceedings of the 31st conference on neural information processing systems* (pp. 5998–6008). Long Beach, CA: MIT Press.

Wu, Yi, Bamman, D., & Russell, S. (2017). Adversarial training for relation extraction. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 1778–1783). Stroudsburg, PA: Association for Computational Linguistics.

Xing, Rui, & Luo, Jie (2019). Distant supervision relation extraction with separate head-tail CNN. In *Proceedings of the 2019 EMNLP workshop W-NUT: The 5th workshop on noisy user-generated text* (pp. 249–258). Stroudsburg, PA: Association for Computational Linguistics.

Ye, Zhixiu, & Ling, Zhenhua (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)* (pp. 2810–2819). Stroudsburg, PA: Association for Computational Linguistics.

Yu, E., Han, W., Tian, Y., et al. (2020). ToHRE: A top-down classification strategy with hierarchical bag representation for distantly supervised relation extraction. In *Proceedings of the 28th international conference on computational linguistics* (pp. 1665–1676).

Yu, M., Yin, W., Hasan, K. S., et al. (2017). Improved neural relation detection for knowledge base question answering. In *Proceedings of the 55th annual meeting of the association for computational linguistics* (pp. 571–581). Stroudsburg, PA: Association for Computational Linguistics.

Yuan, C., Huang, H., Feng, C., et al. (2019). Distant supervision for relation extraction with linear attenuation simulation and non-IID relevance embedding. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 7418–7425).

Zeng, Daojian, Liu, Kang, Chen, Yubo, et al. (2015). Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1753–1762). Stroudsburg, PA: Association for Computational Linguistics.

Zeng, Daojian, Liu, Kang, Lai, Siwei, et al. (2014). Relation classification via convolutional deep neural network. In *Proceedings of the 25th international conference on computational linguistics international conference* (pp. 2335–2344). Stroudsburg, PA: Association for Computational Linguistics.