Regularized Attentive Capsule Network for Overlapped Relation Extraction

Tianyi Liu[1]†, **Xiangyu Lin**[2]†, **Weijia Jia**[3,1]*, **Mingliang Zhou**[4], **Wei Zhao**[5]
[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2]Department of Computer and Information Science, University of Macau
[3]BNU-UIC Institute of AI and Future Networks, Beijing Normal University (Zhuhai)
[4]School of Computer Science, Chongqing University
[5]American University of Sharjah, Sharjah United Arab Emirates

Abstract

Distantly supervised relation extraction has been widely applied in knowledge base construction due to its less requirement of human efforts. However, the automatically established training datasets in distant supervision contain low-quality instances with noisy words and overlapped relations, introducing great challenges to the accurate extraction of relations. To address this problem, we propose a novel Regularized Attentive Capsule Network (RA-CapNet) to better identify highly overlapped relations in each informal sentence. To discover multiple relation features in an instance, we embed multi-head attention into the capsule network as the low-level capsules, where the subtraction of two entities acts as a new form of relation query to select salient features regardless of their positions. To further discriminate overlapped relation features, we devise disagreement regularization to explicitly encourage the diversity among both multiple attention heads and low-level capsules. Extensive experiments conducted on widely used datasets show that our model achieves significant improvements in relation extraction.

1 Introduction

Relation extraction aims to extract relations between entities in text, where distant supervision proposed by (Mintz et al., 2009) automatically establishes training datasets by assigning relation labels to instances that mention entities within knowledge bases. However, the wrong labeling problem can occur and various multi-instance learning methods (Riedel et al., 2010; Hoffmann et al., 2011; Surdeanu et al., 2012) have been proposed to address it. Despite the wrong labeling problem, each instance in distant supervision is crawled from web pages, which is informal with many noisy words and can express multiple similar relations. This problem is not well-handled by previous approaches and severely hampers the performance of conventional neural relation extractors. To handle this problem, we have to address two challenges: (1) Identifying and gathering spotted relation information from low-quality instances; (2) Distinguishing multiple overlapped relation features from each instance.

First, a few significant relation words are distributed dispersedly in the sentence, as shown in Figure 1, where words marked in red brackets represent entities, and italic words are key to expressing the relations. For instance, the clause "*evan_bayh son of birch_bayh*" in *S1* is sufficient to express the relation */people/person/children* of *evan_bayh* and *birch_bayh*. Salient relation words are few in number and dispersedly in *S1*, while others excluded from the clause can be regarded as noise. Traditional neural models have difficulty gathering spotted relation features at different positions along the sequence because they use Convolutional Neural Network (CNN) or Recurrent Neural Network (RNN) as basic relation encoders (Zeng et al., 2015; Liu et al., 2018; Ye and Ling, 2019), which model each sequence word by word and lose rich non-local information for modeling the dependencies of semantic salience. Thus, a well-behaved relation extractor is needed to extract scattered relation features from informal instances.

Second, each instance can express multiple similar relations of two entities. As shown in Figure 1, *Changsha* and *Hunan* possess the relations */location/location/contains* and */location/province/capital* in *S2*, which have similar semantics, introducing great challenges for neural extractors in discriminating

†make equal contributions
*Corresponding author: jiawj@sjtu.edu.cn

ID	Instances	Relations
S1	senator ***[evan_bayh]***, ***son of*** former senator ***[birch_bayh]*** of indiana, is organizing and testing the waters for a possible presidential bid in 2008.	/people/person/children
S2	that is one reason that ***[hunan]***'s fast-growing ***provincial capital***, ***[changsha]***, is beginning to siphon some workers.	/location/location/contains /location/province/capital
S3	the land is near calgary; while that is ***one of [alberta]***'s ***largest cities***, the ***capital*** is ***[edmonton]***.	/location/location/contains /location/province/capital /location/country/capital
…	…	…

Figure 1: Example of instances from the New York Times (NYT).

them clearly. Conventional neural methods are not effective at extracting overlapped relation features, because they mix different relation semantics into a single vector by max-pooling (Zeng et al., 2014) or self-attention (Lin et al., 2016). Although (Zhang et al., 2019) first propose an attentive capsule network for multi-labeled relation extraction, it treats the CNN/RNN as low-level capsules without the diversity encouragement, which poses the difficulty of distinguishing different and overlapped relation features from a single type of semantic capsule. Therefore, a well-behaved relation extractor is needed to discriminate diverse overlapped relation features from different semantic spaces.

To address the above problem, we propose a novel Regularized Attentive Capsule Network (RA-CapNet) to identify highly overlapped relations in the low-quality distant supervision corpus. First, we propose to embed multi-head attention into the capsule network, where attention vectors from each head are encapsulated as a low-level capsule, discovering relation features in an unique semantic space. Then, to improve multi-head attention in extracting spotted relation features, we devise relation query multi-head attention, which selects salient relation words regardless of their positions. This mechanism assigns proper attention scores to salient relation words by calculating the logit similarity of each relation representation and word representation. Furthermore, we apply disagreement regularization to multi-head attention and low-level capsules, which encourages each head or capsule to discriminate different relation features from different semantic spaces. Finally, the dynamic routing algorithm and sliding-margin loss are employed to gather diverse relation features and predict multiple specific relations. We evaluate RA-CapNet using two benchmarks. The experimental results show that our model achieves satisfactory performance over the baselines. Our contributions are summarized as follows:

- We first propose to embed multi-head attention as low-level capsules into the capsule network for distantly supervised relation extraction.
- To improve the ability of multi-head attention in extracting scattered relation features, we design relation query multi-head attention.
- To discriminate overlapped relation features, we devise disagreement regularization on multi-head attention and low-level capsules.
- RA-CapNet achieves significant improvements for distantly supervised relation extraction.

2 Related Work

Distantly supervised relation extraction has been essential for knowledge base construction since (Mintz et al., 2009) propose it. To address the wrong labeling problem in distant supervision, multi-instance and multi-label approaches are proposed (Riedel et al., 2010; Hoffmann et al., 2011; Surdeanu et al., 2012).

With the renaissance of neural networks, increasing researches in distant supervision have been proposed to extract precise relation features. Piecewise CNNs with various attention mechanisms are proposed (Zeng et al., 2015; Lin et al., 2016; Ji et al., 2017). Reinforcement learning and adversarial training

are proposed to select valid instances to train relation extractors (Feng et al., 2018; Qin et al., 2018b; Qin et al., 2018a). Recently, multi-level noise reduction is designed by (Ye and Ling, 2019; Jia et al., 2019).

Nevertheless, the above approaches ignore the effect of noisy words and overlapped relation features in each instance. To reduce the impact of noisy words, tree-based methods attempt to obtain the relevant sub-structure of an instance for relation extraction (Xu et al., 2015; Miwa and Bansal, 2016; Liu et al., 2018). To discriminate overlapped relation features, (Zhang et al., 2019) apply the capsule network (Sabour et al., 2017) for multi-labeled relation extraction. Inspired by the ability of multi-head attention in modeling the long-term dependency (Vaswani et al., 2017), (Zhang et al., 2020) attempt to reduce multi-granularity noise via multi-head attention in relation extraction.

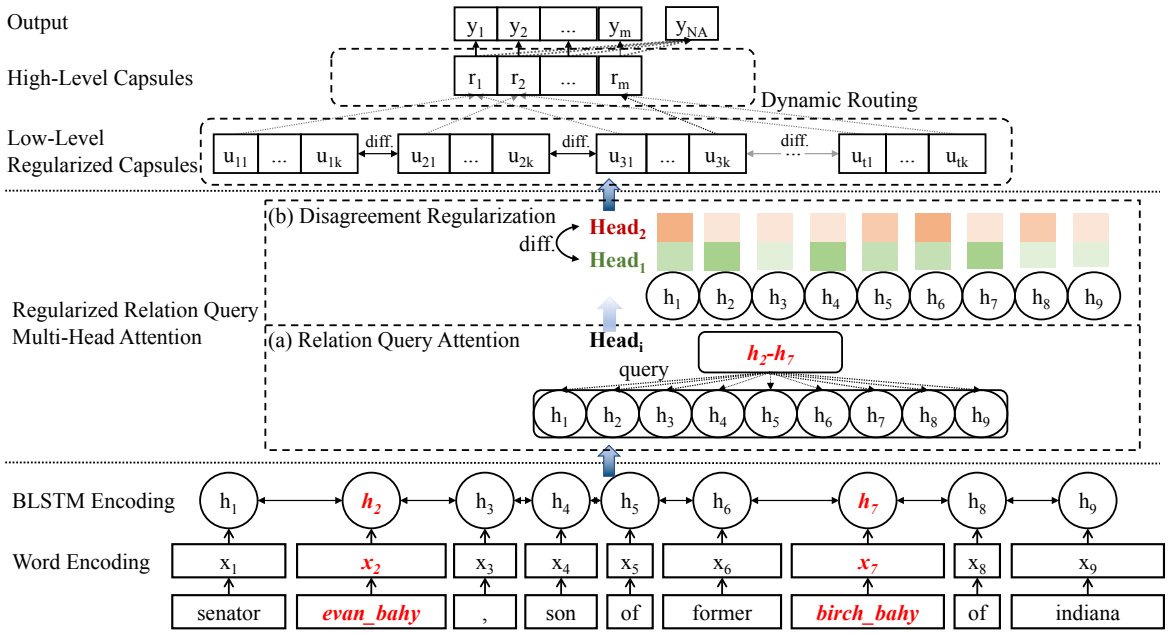

Figure 2: Overall architecture of RA-CapNet, expressing the process of handling an instance.

3 Methodology

As shown in Figure 2, we will introduce the three-layer RA-CapNet: (1) **The Feature Encoding Layer** primarily contains the word encoding layer and BLSTM encoding layer. (2) **The Feature Extracting Layer** chiefly includes relation query multi-head attention and disagreement regularization. (3) **The Relation Gathering Layer** mainly consists of a regularized capsule network and dynamic routing.

3.1 Feature Encoding Layer

Each instance is first input into the encoding layer to be transformed to the distributed representations for the convenience of calculation and extraction by neural networks.

Word Encoding Layer

As mentioned in (Zeng et al., 2014), the inputs of the relation extractor are word and position tokens, which are encoded by word embeddings and position embeddings at first. Then, the j_{th} input word x_{ij} in the i_{th} instance, is concatenated by one word embedded vector $x_{ij}^w \in R^k$ and two position embedded vectors x_{ij}^{p1} and $x_{ij}^{p2} \in R^p$, $x_{ij} = [x_{ij}^w; x_{ij}^{p1}; x_{ij}^{p2}]$, where k and p represent the dimensions of word vectors and position vectors respectively, and ; denotes the vertical concatenating operation. To simplify the mathematical expression, we denote x_{ij} as x_j.

BLSTM Encoding Layer

To further encode relation features inside the context, we adopt the Bidirectional Long-Short Term network (BLSTM) (Graves, 2013) as our basic relation encoder, which can access the future context as well

as the past. The encoding feature vector h_i of the i_{th} word is calculated as follows:

$$\overrightarrow{h_i} = \overrightarrow{LSTM}(x_i, \overrightarrow{h_{i-1}}) \tag{1}$$

$$\overleftarrow{h_i} = \overleftarrow{LSTM}(x_i, \overleftarrow{h_{i+1}}) \tag{2}$$

$$h_i = \overrightarrow{h_i} + \overleftarrow{h_i} \tag{3}$$

where $\overrightarrow{h_i}$ and $\overleftarrow{h_i} \in R^d$ are hidden state vectors of the LSTM. Finally, we obtain the sentence encoding vector $H = [h_1, h_2, \cdots, h_l]$, where l represents the instance length.

3.2 Feature Extracting Layer

First, relation query multi-head attention is devised to emphasize spotted relation features from different semantic spaces. Then, disagreement regularization is applied to encouraging the diversity of relation features that each head discovers.

Relation Query Multi-Head Attention

Multi-head attention is useful for modeling the long-term dependency of salient information in the context (Vaswani et al., 2017). Based on this mechanism, we propose relation query multi-head attention to improve the ability of multi-head attention in extracting spotted and salient relation features regardless of their irregular positions in the instance.

Formally, given an encoding instance H, we use the subtraction of two entities' states h_{en1} and h_{en2} as the relation representation, as inspired by (Bordes et al., 2013). The relation representation acts as a query vector as follows:

$$Q^{rel} = (h_{en1} - h_{en2})W^Q \tag{4}$$

where $W^Q \in R^{d \times d}$ is a weight matrix. The corresponding key K and value V vectors are defined:

$$K = HW^K \quad V = HW^V \tag{5}$$

where W^K and $W^V \in R^{d \times d}$ are weight matrices. Afterward, we calculate the logit similarity of the relation query vector and word representation vectors as attention scores:

$$energy = \frac{Q^{rel}K^T}{\sqrt{d}} \tag{6}$$

where the $energy$ can measure the importance of each word to relation extraction, which is leveraged to select salient and spotted relation features along the sequence:

$$ATT = softmax(energy)V \tag{7}$$

To extract diverse relation features, we employ relation query attention into multi-head attention:

$$head_i = ATT(Q_i^{rel}, K_i, V_i) \tag{8}$$

$$E^m = [head_1; head_2; \cdots; head_n]W^o \tag{9}$$

where $W^o \in R^{d \times d}$ is the weight matrix. Multiple heads can capture various semantic features.

After we acquire the output E^m of multi-head attention, a Feed-Forward Network (FFN) is applied:

$$H^r = max(0, E^m W_1^f + b_1^f)W_2^f + b_2^f \tag{10}$$

where $W_1^f \in R^{d \times d}$, $W_2^f \in R^{d \times d'}$, $b_1^f \in R^d$ and $b_2^f \in R^{d'}$ are parameters.

Disagreement Regularization on Multi-Head Attention

To further discriminate overlapped relation features from different heads in multi-head attention, we introduce the disagreement regularization based on (Yang et al., 2018).

Formally, given n heads $Head = [head_1, head_2, \cdots, head_n]$ as calculated in Eq. (8), we calculate the cosine similarity $cos(.)$ between the vector pair $head_i$ and $head_j$ in different value subspaces:

$$D_{ij}^{sub} = cos(head_i, head_j) = \frac{head_i \cdot head_j}{\|head_i\| \|head_j\|} \tag{11}$$

where $\| * \|$ represents the $L2$ norm of vectors. The average cosine distance among all heads is obtained:

$$D^{sub} = \frac{\sum_{ij} D_{ij}^{sub}}{n^2} \tag{12}$$

Our goal is to minimize D^{sub}, which encourages the heads to be different from each other, improving the diversity of subspaces among multiple heads. Accordingly, each head can discriminate overlapped relation features more clearly.

3.3 Relation Gathering Layer

To form relation-specific features, the relation gathering layer gathers scattered relation features from diverse low-level capsules using a dynamic routing algorithm.

Low-Level Capsules with Disagreement Regularization

The capsule network has been proven effective in discriminating overlapped features (Sabour et al., 2017; Zhang et al., 2019). In our application, a capsule is a group of neural vectors within one-head attention and regularized by a disagreement term. Thus, each capsule can capture relation features in an unique semantic space. In detail, the orientation of the attention vector inside one head indicates one certain factor of a specific relation, while its length means the probability that this relational factor exists.

We reorganize each attention head of H^r to form a low-level capsule denoted as $u \in R^{d_u}$, where each capsule captures information in a specific semantic space. Formally, the above process is as follows:

$$H^r = [h_1^r; ...; h_t^r] \tag{13}$$

$$u_k = g(h_k^r) = \frac{\|h_k^r\|^2}{1 + \|h_k^r\|^2} \frac{h_k^r}{\|h_k^r\|} \tag{14}$$

where t is the number of low-level capsules, which equals the quantity of heads. Eq. (14) is a squash function, shrinking the length of vectors from 0 to 1 to express the probability.

To encourage the diversity of these capsules, disagreement regularization is applied to them:

$$D_{ij}^{cap} = \frac{u_i \cdot u_j}{\|u_i\| \|u_j\|} \tag{15}$$

$$D^{cap} = \frac{\sum_{ij} D_{ij}^{cap}}{t^2} \tag{16}$$

To minimize D^{cap}, we can encourage the capsules to be different from each other, improving the diversity of subspaces among multiple capsules and discriminating overlapped relation features more clearly.

The final disagreement regularization term is the average of multi-head and capsule disagreement:

$$D = \frac{D^{sub} + D^{cap}}{2} \tag{17}$$

where D is the final disagreement regularization term which only works for the training process.

High-Level Capsules with Dynamic Routing

After the low-level capsules capturing the different aspects of semantic information, the high-level capsules $r \in R^{d_r}$ are produced from them to gather scattered information and form specific relation features, which are calculated as follows:

$$r_j = g(\sum c_{ij} W_j^h u_i) \tag{18}$$

where $W_j^h \in R^{d_u \times d_r}$ are parameters for high-level capsules and c_{ij} are coupling coefficients that are determined by the dynamic routing process described in (Sabour et al., 2017).

Loss Function

The sliding-margin loss function used in the capsule network enables the prediction of multiple overlapped relations, which sums up the loss for both the relations present and absent from the instances. This margin loss function is integrated into our model as follows:

$$\begin{aligned} L_j =& Y_j max(0, (S+\gamma) - \|r_j\|)^2 + \\ & \lambda(1-Y_j) max(0, \|r_j\| - (S-\gamma))^2 \end{aligned} \tag{19}$$

where γ is the width of the margin, S is a learnable threshold for "no relation" (NA), and λ is the down-weighting of the loss for absent relations. Y_j = 1 if the relation corresponding to r_j is present in the sentence and Y_j = 0 otherwise.

Afterward, the final loss is defined as follows:

$$loss = \sum_j L_j + \beta D + \beta' \|\theta\|^2 \tag{20}$$

where β and β' are hyperparameters used to restrict the disagreement regularization and $L2$ regularization of all parameters θ. In this paper, we use the Adam (Kingma and Ba, 2014) to minimize the final loss.

4 Experiments

Our experiments are devised to demonstrate that RA-CapNet can identify highly overlapped relations of informal instances in distant supervision. In this section, we first introduce the dataset and experimental setup. Then, we evaluate the overall performance of RA-CapNet and the effects of different parts of RA-CapNet. Finally, we present the case study.

4.1 Dataset and Experimental Setup

Dataset. To evaluate the effects of RA-CapNet, we conduct experiments on two datasets. **NYT-10** is a standard dataset constructed by (Riedel et al., 2010), which aligns relational tuples in Freebase (Bollacker et al., 2008) with the corpus of New York Times. Sentences from 2005-2006 are used as the training set, while sentences from 2007 are used for testing. **NYT-18** is a larger dataset constructed by (Zhang et al., 2020) with the same creation method as NYT-10, which crawls 2008-2017 contexts from the NYT. All the sentences are divided into five parts with the same relation distribution for five-fold cross-validation. The details of the datasets are illustrated in Table 1.

Datasets	Training (k)		Testing (k)		Rel.
	Sen.	Ent.	Sen.	Ent.	
NYT-10	523	281	172	97	53
NYT-18	2446	1234	611	394	503

Table 1: The dataset information. **Sen.**, **Ent.** and **Rel.** indicate numbers of sentences, entity pairs and relations (including NA).

Evaluation Metric. As mentioned in (Mintz et al., 2009), we use the held-out metrics to evaluate RA-CapNet. The held-out evaluation offers an automatic way to assess models with the PR curve and

Precision at top 100 or 10k predictions (P@100 on NYT-10 or P@10k on NYT-18) at all numbers of instances under each entity pair, which indicates that all instances under the entity pair are used to represent the relation.

Parameter. In our work, we use the *Skip-Gram* (Mikolov et al., 2013) to pretrain our word embedding matrices. The words of an entity are concatenated when it has multiple words. The grid search and cross-validation are used to adjust important hyperparameters of the networks. Our final parameter settings are illustrated in Table 2.

Parameter	NYT-10	NYT-18
Batch size b	50	50
Word embedding size k	50	360
Position embedding size p	5	5
Sentence length l	100	100
LSTM hidden size d	256	256
Multi-head number n	16	16
FFN hidden size d'	512	512
Capsule dimensions $[d^u, d^r]$	[16,16]	[16,16]
Low-level capsule number t	16	16
Valid relation class m	52	502
Sliding margin γ	0.4	0.4
Down-weighting λ	1.0	1.0
Learning rate lr	0.0001	0.0001
Dropout probability p	0.5	0.5
Weight of disagreement β	0.001	0.001
$L2$ penalty β'	1e-08	0.0

Table 2: Parameter settings.

4.2 Overall Performance

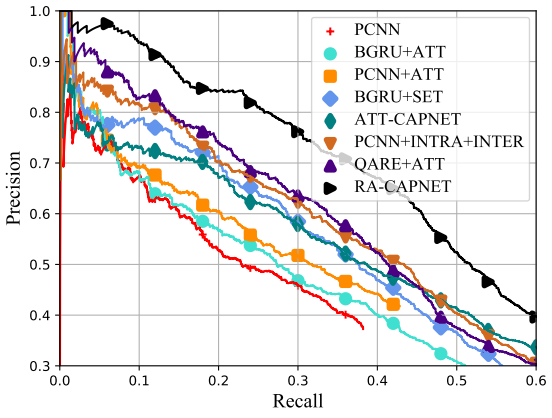

Figure 4: Precision-recall curves on NYT-10.

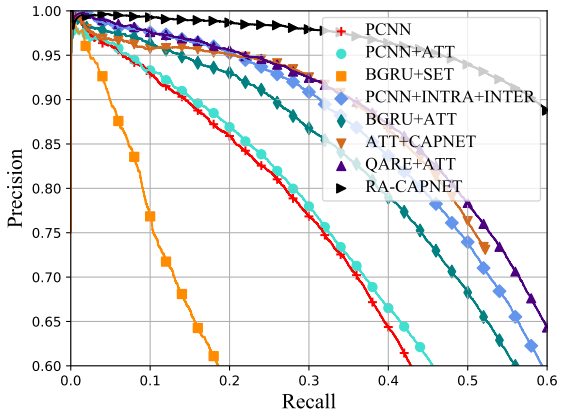

Figure 5: Precision-recall curves on NYT-18.

To evaluate our model, we select the following methods for comparison:

PCNN (Zeng et al., 2015) present a piecewise CNN for relation extraction.

PCNN+ATT (Lin et al., 2016) propose the selective attention mechanism with PCNN.

BGRU+ATT (Zhou et al., 2016) present a BGRU-based model with word-level attention.

BGRU+SET (Liu et al., 2018) propose a BGRU-based approach to reduce inner-sentence noise.
PCNN+INTRA+INTER (Ye and Ling, 2019) propose to emphasize true labeled sentences and bags.
ATT+CAPNET (Zhang et al., 2019) put forward an attentive capsule network for relation extraction.
QARE+ATT (Zhang et al., 2020) propose improved multi-head attention with transfer learning.

We compare our method with baselines on two datasets. For both datasets, the PR curves on NYT-10 and NYT-18 are shown in Figure 4 and Figure 5. We find that: (1) BGRU+SET performs well on NYT-10 but poorly on NYT-18. This demonstrates that BGRU+SET is not well-handled on highly informal instances because the complex instances in NYT-18 are difficult to be parsed precisely by the conventional parser. (2) RA-CapNet achieves the best PR curve among all baselines on both datasets, which improves the PR curve significantly. This verifies that our model is effective in capturing overlapped and scattered relation features. (3) RA-CapNet outperforms ATT+CAPNET, which indicates that the relation query multi-head attention and disagreement regularization are useful for overlapped relation extraction.

Model	PR curve area	
	NYT-10	NYT-18
BGRU+ATT	0.337	0.596
PCNN+ATT	0.356	0.511
BGRU+SET	0.392	0.290
ATT-CAPNET	0.415	0.647
PCNN+INTRA+INTER	0.423	0.617
QARE+ATT	0.428	0.645
RA-CAPNET	**0.526**	**0.780**

Table 3: Precision-recall curve areas.

Model	P@100	P@10k
PCNN	72.3	81.0
PCNN+ATT	82.0	82.2
BGRU+ATT	74.0	88.1
BGRU+SET	87.0	67.4
PCNN+INTRA+INTER	91.8	-
ATT+CAPNET	84.0	-
QARE+ATT	93.0	91.6
RA-CAPNET	**98.0**	**96.6**

Table 4: P@100 and P@10k.

A detailed comparison of all approaches, the areas of the PR curves, P@100 and P@10k on NYT-10 or NYT-18, are illustrated in Table 3 and Table 4. From the tables, we find that: (1) RA-CapNet is the first method to increase the PR curve area over 0.5 on NYT-10 while improving it on NYT-18 to 0.7. In P@100 and P@10k, our model also achieves superior performance. This result further demonstrates the effectiveness of RA-CapNet with multi-instance learning on overlapped relation extraction. (2) Capnet-based models achieve better performance on the highly complex NYT-18 dataset, which results from their capability of handling overlapped relations and complex sentences.

4.3 Ablation Study

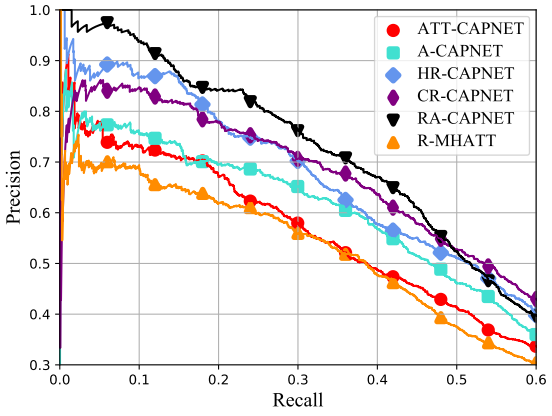

Figure 6: PR curves of our models.

Model	PR curve area
R-MHATT	0.383
ATT-CAPNET	0.415
A-CAPNET	0.449
HR-CAPNET	0.493
CR-CAPNET	0.501
RA-CAPNET	**0.526**

Figure 7: PR curve areas of our models.

To further evaluate the impacts of different parts on RA-CapNet, we compare the performance on the

NYT-10 dataset of RA-CapNet with five settings:

R-MHATT: Two multi-head attention layers with relation query attention.

ATT-CAPNET: The same as above.

A-CAPNET: RA-CapNet without disagreement regularization.

HR-CAPNET: RA-CapNet without capsule disagreement regularization.

CR-CAPNET: RA-CapNet without multi-head disagreement regularization.

RA-CAPNET: Our model.

In Figure 6 and 7, the result indicates that: (1) ATT-CAPNET improves the performance of R-MHATT by incorporating the capsule network for handling the multiple relations. (2) Compared with ATT-CAPNET, A-CAPNET improves the PR curve area from 0.415 to 0.449. This proves that relation query multi-head attention helps the capsule network extract salient relation features from different representations at different positions. (3) HR-CAPNET further increases the PR curve area to 0.493, which proves the effectiveness of our disagreement regularization on multiple heads in discriminating the diverse overlapped relation features. (4) Compared with A-CAPNET, CR-CAPNET achieves 0.501 of the PR curve area. This demonstrates that disagreement regularization on the capsules helps models distinguish multiple relation features more clearly. (5) Our complete model, RA-CAPNET, achieves the best performance, showing that the relation query multi-head attention and disagreement regularization term are both effective for relation extraction.

4.4 Case Study

ID	Instances					
S1	that is one reason that **[*hunan*]**'s fast-growing ***provincial capital***, **[*changsha*]**, is beginning to siphon some workers.					
	Model	PCNN+ATT	BGRU+ATT	ATT+CAPNET	QARE+ATT	RA-CAPNET
	LLC	✗	✗	✓	✗	✓
	LPC	✓	✓	✓	✓	✓
S2	the land is near calgary; while that is one of ***[alberta]***'s largest cities, the capital is ***[edmonton]***.					
	Model	PCNN+ATT	BGRU+ATT	ATT+CAPNET	QARE+ATT	RA-CAPNET
	LLC	✗	✗	✗	✗	✓
	LPC	✗	✗	✓	✗	✓
	LCC	✓	✓	✓	✓	✓

Figure 7: Prediction results of different models on some samples. "LLC", "LPC" and "LCC" represent relation labels of "/location/location/contains", "/location/province/capital" and "/location/country/capital".

In Figure 7, we randomly select two samples from NYT-10 to analyze the prediction performance of different models, where the entities are labeled in the red and bold brackets. From the figure, we find the following: (1) In *S1* and *S2*, compared with CNN/RNN/Attention-based methods, the capsule-based approaches can predict multiple similar relations. (2) In *S2*, only RA-CapNet predicts the correct relation of "/location/location/contains". This result demonstrates that by incorporating relation query multi-head attention and disagreement regularization in the capsule network, RA-CapNet makes further progress in discriminating overlapped relations.

5 Conclusion and Future Work

In this paper, we propose a novel regularized attentive capsule network for overlapped relation extraction. RA-CapNet embeds relation query multi-head attention into the capsule network and uses a novel disagreement regularization term to encourage the diversity among heads and capsules, making it capa-

ble of gathering salient information from diverse semantic spaces. Our model is resistant to the noise of distant supervision and achieves significant improvements on both standard and complex datasets.

In the future, we will experiment with different forms of regularization terms and their application to other components of our model.

Acknowledgements

This work is partially supported by Chinese National Research Fund (NSFC) Key Project No. 61532013 and No. 61872239; BNU-UIC Institute of Artificial Intelligence and Future Networks funded by Beijing Normal University (Zhuhai) and AI and data Science Hub, United International College (UIC), Zhuhai, Guangdong, China.

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 ACM SIGMOD*.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Proceedings of the NeurIPS*.

Jun Feng, Minlie Huang, Li Zhao, Yang Yang, and Xiaoyan Zhu. 2018. Reinforcement learning for relation classification from noisy data. In *Proceedings of the AAAI*.

Alex Graves. 2013. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the ACL*.

Guoliang Ji, Kang Liu, Shizhu He, and Jun Zhao. 2017. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In *Proceedings of the AAAI*.

Wei Jia, Dai Dai, Xinyan Xiao, and Hua Wu. 2019. Arnor: Attention regularization based noise reduction for distant supervision relation classification. In *Proceedings of the ACL*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the ACL*.

Tianyi Liu, Xinsong Zhang, Wanhao Zhou, and Weijia Jia. 2018. Neural relation extraction via inner-sentence noise reduction and transfer learning. In *Proceedings of the EMNLP*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

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 ACL and the AFNLP*.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using lstms on sequences and tree structures. In *Proceedings of the ACL*.

Pengda Qin, XU Weiran, and William Yang Wang. 2018a. Dsgan: Generative adversarial training for distant supervision relation extraction. In *Proceedings of the ACL*.

Pengda Qin, XU Weiran, and William Yang Wang. 2018b. Robust distant supervision relation extraction via deep reinforcement learning. In *Proceedings of the ACL*.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Proceedings of the ECML and PKDD*.

Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. 2017. Dynamic routing between capsules. In *Proceedings of the NeurIPS*.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 Joint Conference on EMNLP and CoNLL*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the NeurIPS*.

Yan Xu, Lili Mou, Ge Li, Yunchuan Chen, Hao Peng, and Zhi Jin. 2015. Classifying relations via long short term memory networks along shortest dependency paths. In *Proceedings of the EMNLP*.

Baosong Yang, Zhaopeng Tu, Derek F Wong, Fandong Meng, Lidia S Chao, and Tong Zhang. 2018. Modeling localness for self-attention networks. In *Proceedings of the EMNLP*.

Zhi-Xiu Ye and Zhen-Hua Ling. 2019. Distant supervision relation extraction with intra-bag and inter-bag attentions. In *Proceedings of the ACL*.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING*.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the EMNLP*.

Xinsong Zhang, Pengshuai Li, Weijia Jia, and Hai Zhao. 2019. Multi-labeled relation extraction with attentive capsule network. In *Proceedings of the AAAI*.

X. Zhang, T. Liu, P. Li, W. Jia, and H. Zhao. 2020. Robust neural relation extraction via multi-granularity noises reduction. *IEEE Transactions on Knowledge and Data Engineering*.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the ACL*.