Modeling Dense Cross-Modal Interactions for Joint Entity-Relation Extraction

Shan Zhao[1] , **Minghao Hu**[2] , **Zhiping Cai**[1*] and **Fang Liu**[3†]
[1]College of Computer, National University of Defense Technology, Changsha, China
[2]PLA Academy of Military Science, Beijing, China
[3]School of Design, Hunan University, Changsha, Hunan
{zhaoshan18, zpcai}@nudt.edu.cn, {huminghao16, liufang06}@gmail.com

Abstract

Joint extraction of entities and their relations benefits from the close interaction between named entities and their relation information. Therefore, how to effectively model such cross-modal interactions is critical for the final performance. Previous works have used simple methods such as label-feature concatenation to perform coarse-grained semantic fusion among cross-modal instances, but fail to capture fine-grained correlations over token and label spaces, resulting in insufficient interactions. In this paper, we propose a deep Cross-Modal Attention Network (CMAN) for joint entity and relation extraction. The network is carefully constructed by stacking multiple attention units in depth to fully model dense interactions over token-label spaces, in which two basic attention units are proposed to explicitly capture fine-grained correlations across different modalities (e.g., token-to-token and label-to-token). Experiment results on CoNLL04 dataset show that our model obtains state-of-the-art results by achieving 90.62% F1 on entity recognition and 72.97% F1 on relation classification. In ADE dataset, our model surpasses existing approaches by more than 1.9% F1 on relation classification. Extensive analyses further confirm the effectiveness of our approach.

1 Introduction

Extraction of entities and their relations from unstructured raw texts has attracted increasing attention due to its important application on knowledge base population, information retrieval, and question answering [Guo *et al.*, 2019]. Given a sentence, the task aims to find the location and type of mentioned entities, and further detect semantic relations among those entities. For example, in Figure 1, "*Tanya*" is a person entity (Peop), while "*Shabds Hospital*" and "*Gainesville*" are two location entities (Loc) connected by a "*Located In*" relation.

Traditionally, the task of extracting semantic relations between entities is decoupled into a pipeline of two separated subtasks, namely named entity recognition (NER) [Nadeau and Sekine, 2007] and relation extraction (RE) [Bach and Badaskar, 2007]. Since named entities interact closely with their relation information (two location entities are usually linked with a "*Located In*" relation), joint models that simultaneously learn NER and RE have been proposed and have achieved promising results [Miwa and Bansal, 2016; Adel and Schütze, 2017; Bekoulis *et al.*, 2018; Li *et al.*, 2019]. However, joint models only capture such *cross-modal* interaction by learning shared representations via multi-task training, but fail to take label information into account, which turns out to be a significant limitation. For example, if the model knows that "*Shabds Hospital*" and "*Gainesville*" are location entities beforehand, it can easily infer there may exist a "*Located In*" relation between them.

To overcome the problem of insufficient cross-modal interactions, some works [Miwa and Bansal, 2016; Bekoulis *et al.*, 2018] propose to enhance downstream RE performance by leveraging label information extracted from upstream NER process. These approaches adopt simple feature concatenation to fuse label information into contextualized representations, which results in promising performance improvement. However, such naive methods can only learn coarse-grained interactions of cross-modal instances via token-level semantic fusion, but cannot effectively infer the correlation between each token and each tagging label (e.g., it is beneficial that "*Shabds Hospital*" is aware of "*Gainesville*" being assigned with a "*B-LOC*" tag). Moreover, token-level self-correlation is also important for both NER and RE, which has been ignored by previous RNN or CNN based models [Wang *et al.*, 2016; Katiyar and Cardie, 2017]. For example, the fact that "*Shabds Hospital*" is highly relevant to "*Gainesville*" but less related with "*Tanya*" is helpful for entity recognition as well as relation classification.

To address the above issues, we propose a deep Cross-Modal Attention Network (CMAN) for joint entity and relation extraction. Inspired by multimodal learning in computer vision [Yu *et al.*, 2019], we view token and label spaces as two different modalities, and attempt to model dense cross-modal interactions over these two spaces. To achieve this,we design two basic attention units: a BiLSTM-enhanced self-attention (BSA) unit that aims to model intramodal interactions across different tokens (token-to-token); and a BiLSTM-enhanced label-attention (BLA) unit that is capable of modeling inter-

*Corresponding Author
†Corresponding Author

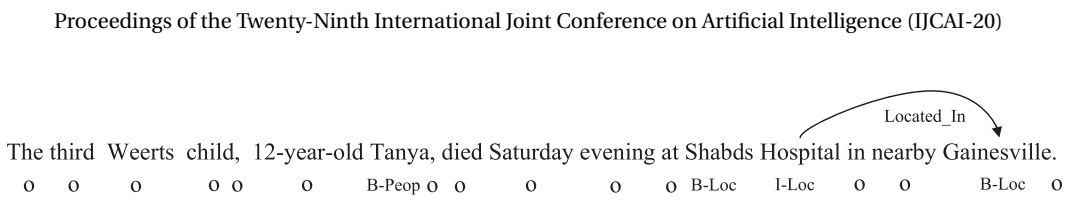

Figure 1: An example from CoNLL04 dataset, where the goal is to identify mentioned entities and corresponding relationships in the sentence.

modal interactions (label-to-token). BSA is able to build direct connections between two arbitrary tokens in a sentence despite of their distances, while BLA explicitly leverages label-space information to enhance contextualized token representations. Next, we construct the entire model by carefully stacking multiple attention units to form a deep network architecture for fully capturing cross-modal interactions, where gold label information is available only during training and is predicted during inference. Finally, we conducted extensive experiments on CoNLL04 and ADE datasets to evaluate the proposed model. In CoNLL04, our model obtains state-of-the-art results by achieving 90.62% and 72.97% F1 on entity recognition and relation classification respectively. Moreover, our model surpasses existing approaches by more than 1.9% F1 score on relation classification in ADE dataset.

2 Related Work

Joint entity-relation extraction. Due to the existence of close interactions between entity recognition and relation classification, joint models that simultaneously learn NER and RE have outperformed pipelined methods [Miwa *et al.*, 2009] by a large margin. Specifically, Miwa and Bansal [2016] employ bidirectional tree-structured RNNs, which extract relationships between entities based on word order information and dependent tree structure information. Wang et al. [2016] extract relations using multi-level attention CNNs. Then, a novel tagging scheme is proposed to convert the joint extraction problem into a sequence labeling problem [Zheng *et al.*, 2017], which is usually solved by RNNs-based decoding strategies. Yet, this tagging scheme is difficult to handle multiple relationships, which are relatively rare in many datasets. Therefore, Bekoulis et al [2018] propose a multi-head mechanism to support the prediction of multiple relationships. Compared to these approaches that adopt either RNNs or CNNs-based architecture, our model consists of cascaded attention units that combine bidirectional LSTM with multi-head attention [Vaswani *et al.*, 2017] to better capture correlations between any two modal instances despite of their relative distance.

Label-space information. Recently, label information has been applied to NLP tasks and achieves ideal results. Specifically, label knowledge has been exploited in the text classification task [Wang *et al.*, 2018]. Moreover, Zhang [2019] introduce label embeddings to the NER task. However, label-space information has not been carefully studied in joint entity and relation extraction. Prior approaches[Miwa and Bansal, 2016; Bekoulis *et al.*, 2018] exploit a naive way such as feature concatenation to utilize coarse-grained label. In contrast, we aim to model dense cross-modal interactions over token-label spaces, which delivers significantly better performance.

Multimodal learning. Multimodal learning is widely explored in computer vision and natural language processing. A typical task is visual question answering (VQA) [Antol *et al.*, 2015], which requires the model to perform fine-grained semantic understanding of both the image and the question. For example, Yu et al. [2019] propose modular attention mechanism to capture the interactions of multimodal instances (image and question). Inspired by recent advancements in this field, we regard token and label spaces as two different modalities and attempt to capture cross-modal interactions between them.

3 The Proposed Model

In this section, we introduce the deep Cross-Modal Attention Network (CMAN) in details, which is shown in Figure 2. We first obtain fixed-dimensional representations of token and label from different perspectives (§3.1). Then, we design a BiLSTM-enhanced self-attention unit and a BiLSTM-enhanced label-attention unit (§3.2). These two units are built to explicitly leverage token-label spaces information for modeling cross-modal interactions (e.g., token-to-token and label-to-token). After that, a deep network architecture based on these two units is carefully designed to utilize gold label information during training and predicted labels at inference time (§3.3). Finally, we apply a conditional random filed (CRF) [Lafferty *et al.*, 2001] and a multi-head mechanism [Bekoulis *et al.*, 2018] to perform the decoding for NER and RE (§3.4).

3.1 Representations in Token-Label Spaces

As mentioned above, sequence tokens and tagging labels are viewed as two different modalities, and therefore can be represented with different distributed representations. Below we will present how to construct these representations.

Token Representations

Word embeddings are used to map discrete words into continuous input vectors. Given a sentence containing n words, we map each token in the sentence to a real-valued embedding to express its semantic and syntactic meaning. Besides, we also utilize character embeddings, which is obtained by encoding character sequences with a bidirectional LSTM. Then, the input of each token is a concatenation of character embeddings, word embeddings, and ELMo embeddings [Peters *et al.*, 2018]. In this way, a sequence of input vectors $X \in \mathbb{R}^{n*d_w}$ can be obtained, where $d_w$is the token embedding dimension.

Label Representations

We adopt the BIO (Beginning, Inside, Outside) encoding scheme for NER, as illustrated in Figure 1. Motivated by [Miwa and Bansal, 2016], tagging labels are represented with randomly initialized vectors that are fine-tuned during training,

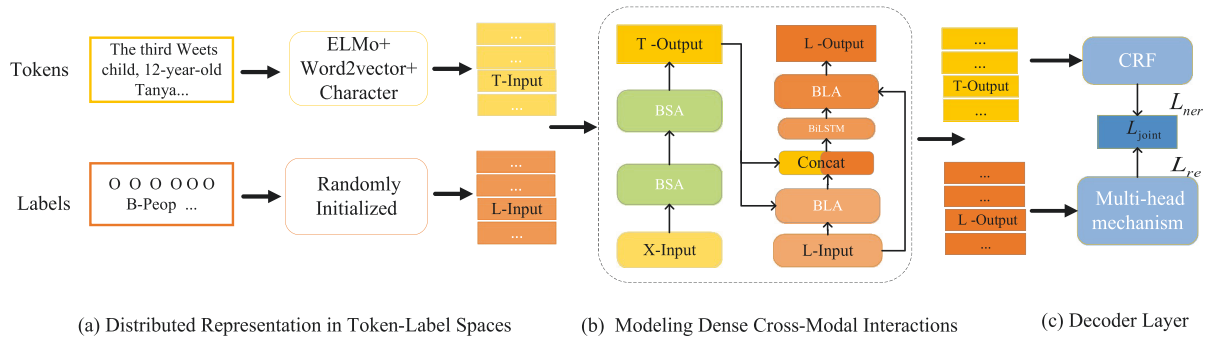

Figure 2: Overall flowchart of CMAN. Tokens and labels are first represented as distributed representations from multiple perspectives. A deep network architecture based on two attention units is then designed to utilize gold label information during training and predicted labels at inference time. Finally, a conditional random filed (CRF) [Lafferty *et al.*, 2001] and a multi-head mechanism [Bekoulis *et al.*, 2018] is used to perform decoding for NER and RE. T and L denote token and label, respectively.

thus yielding a sequence of label vectors $L \in \mathbb{R}^{n*d_l}$, where d_l is the label embedding dimension. Notice that ground-truth labels are used only during training, while predicted labels are utilized at inference time (see more details in §3.5).

3.2 Two Basic Attention Units

We first present a general architecture that contains BiLSTM and multi-head attention for encoding and attending any arbitrary sequence. Then we build two attention units based on this architecture to capture dense correlations among token-label spaces, namely a BiLSTM-enhanced self-attention (BSA) unit and a BiLSTM-enhanced label-attention (BLA) unit.

General architecture. Bidirectional LSTM (BiLSTM) is superior in build contextualized representations for various NLP tasks, as shown in [Katiyar and Cardie, 2017]. Hence, we utilize BiLSTM as the basic encoding component. Given a sequence of input vectors $X = [x_1, ..., x_n]$, a BiLSTM can be used to output hidden representations $H \in \mathbb{R}^{n*2d}$ as:

$$H = \text{BiLSTM}(X) \tag{1}$$

Multi-head attention [Vaswani *et al.*, 2017] has proven to be effective for capturing long-range dependencies by explicitly attending to all positions. Therefore, we apply the multihead attention as the basic attending component for capturing arbitrary correlations. Specifically, we project hidden representations $H = [h_1, ..., h_n]$ into three different representations, namely query, key, and value. Then z parallel heads are employed to capture correlations in different parts of channels:

$$\text{head}_i = \text{softmax}(\frac{(QW_i^Q)(KW_i^K)^T}{\sqrt{2d/z}})(VW_i^V) \tag{2}$$

$$T = \text{Concat}(\text{head}_1, ..., \text{head}_z)W^o \tag{3}$$

where $W_i^Q \in \mathbb{R}^{2d*2d/z}, W_i^K \in \mathbb{R}^{2d*2d/z}, W_i^V \in \mathbb{R}^{2d*2d/z}$, and $W_i^o \in \mathbb{R}^{2d*2d}$ are trainable parameter matrices. Finally, a residual connection [He *et al.*, 2016] along with layer normalization [Ba *et al.*, 2016] is applied on H and T to produce the final output features.

BiLSTM-enhanced self-attention. The BSA unit (see Figure 3a) is designed to model token-to-token self-correlations. Taking one group of input token features $X = [x_1, ..., x_n]$, the BiLSTM is first used to capture rich contextual information over token-space. Next, the multi-head attention receives the encoded hidden representations $H = [h_1, ..., h_n]$, and further learns the pairwise relationship between the paired sample $< h_i, h_j >$ within H and finally outputs attended output features by weighted summation across all instances. In summary, the computation of BSA unit can be defined as $\hat{X} = \text{BSA}(X)$.

BiLSTM-enhanced label-attention. The BLA unit (see Figure 3b) is capable of modeling intermodal interactions from label space to token space. It takes two groups of features $X \in \mathbb{R}^{n*d_w}$ and $L \in \mathbb{R}^{n*d_l}$ as inputs. The BiLSTM component is first used to encode label features as $\tilde{L} = \text{BiLSTM}(L)$. Next, the BLA unit models the pairwise relationship between each paired sample $< x_i, \tilde{l}_i >$ within X and $\tilde{L}$. Notice that we set token features X as query, and set encoded label features $\tilde{L}$ as key and value, so that each token can be fused with relevant label information. The calculation of BLA unit can be summarized as $\tilde{X} = \text{BLA}(X, L)$.

3.3 Dense Cross-Modal Interaction Learning

Taking the aforementioned token features X and label features Y as inputs, we perform dense cross-modal interaction learning by feeding input features into a deep network that contains carefully-designed cascaded attention units. Since there is no label information during entity recognition, we first pass token features X into several BSA units in a recursive manner.

$$X^i = \text{BSA}^i(X^{i-1}), \forall i \in [1, m] \tag{4}$$

where X^0 is set as X, and m is number of BSA units.

After multiple rounds of encoding, token features X^m contains rich information about self-correlations among input tokens. Thus, it can be directly used for entity recognition. As for relation classification, we further take self-aware token features X^m and label features L as inputs, and utilize a BLA unit to obtain initial label-aware token representations as:

$$\tilde{X}^1 = \text{BLA}^1(X^m, L) \tag{5}$$

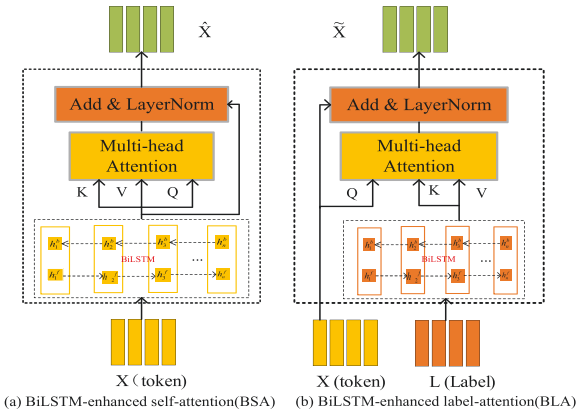

Figure 3: Two basic attention units. BSA is composed of a BiLSTM layer and a self-attention layer, which aims to model intramodal interactions across different tokens. BLA contains a BiLSTM layer and a label-attention layer, which is used to model intermodal interactions from label space to token space.

Next, we apply a concatenation-style residual connection [He *et al.*, 2016] on previous input and output token features, and further use another BiLSTM to fuse their semantic meanings:

$$\tilde{X}^2 = \text{BiLSTM}([X^m; \tilde{X}^1]) \tag{6}$$

Finally, taking $\tilde{X}^2$ and L as inputs, we apply another BLA unit to capture deep cross-modal correlations to form the final label-aware token features as:

$$\tilde{X}^3 = \text{BLA}^2(\tilde{X}^2, L) \tag{7}$$

Now $\tilde{X}^3$ is capable of capturing rich cross-modal interactions, and is suitable for the task of relation classification.

3.4 Decoder Layer

The decoder layer is responsible for predicting two subtasks, namely named entity recognition (NER) and relation extraction (RE), which is introduced below.

NER. A standard CRF layer is used to predict NER taggings, which takes self-aware token features $X^m = [x_1^m, ..., x_n^m]$ as inputs, and outputs a sequence of predicted taggings $Y = [y_1, ..., y_n]$. Let $Y^{'}$ denote the set of tagging labels (i.e., BIO scheme), the probability of the tagging sequence can then be calculated as follows:

$$Pr(Y|X^m) = \frac{\prod_{i=1}^{n} \varphi(y_{i-1}, y_i, X^m)}{\sum_{y' \in Y'} \prod_{i=1}^{n} \varphi(y'_{i-1}, y'_i, X^m)} \tag{8}$$

where $\varphi(y_{i-1}, y_i, X^m)$ is the transition score from y_{i-1} to y_i calculated by $\exp(W_\varphi X^m + b_\varphi)$, and W_φ and b_φ are trainable weight and bias.

RE. We utilize the multi-head mechanism for predicting RE, of which details can be found from [Bekoulis *et al.*, 2018]. Suppose that label-aware token features $\tilde{X}^3 = [\tilde{x}_1^3, ..., \tilde{x}_n^3]$ are given as inputs, and C is a set of relation labels. The idea of this mechanism is to predict a score for each tuple (w_i, w_j, c_k), where w_i is the head token, w_j is the tail token, and c_k denotes the k-th relation between them. Note that each pair of tokens $< w_i, w_j >$ can have multiple heads, where each head computes a score for one relation. We calculate the score between w_i and w_j given a relation c_k as follows:

$$s(\tilde{x}_i^3, \tilde{x}_j^3, c_k) = V_k \tanh(U_k \tilde{x}_i^3 + W_k \tilde{x}_j^3 + b_k) \tag{9}$$

where $V_k \in \mathbb{R}^{\tilde{d}}$, $W_k \in \mathbb{R}^{\tilde{d} * 2d}$, $U_k \in \mathbb{R}^{\tilde{d} * 2d}$, $b_k \in \mathbb{R}^{\tilde{d}}$ are parameters for the k-th relation, and $\tilde{d}$ is intermediate hidden size. Next, the probability of token w_i selected as the head of token w_j with the relation c_k is calculated as:

$$\begin{aligned} &Pr(\text{head} = w_i, \text{relation} = c_k | w_j) \\ &= \sigma(s(\tilde{x}_i^3, \tilde{x}_j^3, c_k)) \end{aligned} \tag{10}$$

where σ stands for the sigmoid function.

3.5 Training and Inference

During training, we optimize the parameters of the model by minimizing the following conditional likelihood for NER:

$$\mathcal{L}_{ner} = -logPr(Y|X^m) \tag{11}$$

As for RE, the cross-entropy loss is applied for training:

$$\begin{aligned} \mathcal{L}_{re} = \sum_{j=1}^{n} \sum_{i=1}^{n} \sum_{k=1}^{o} -logPr(&\text{head} = w_i, \\ &\text{relation} = c_k | w_j) \end{aligned} \tag{12}$$

where o is the number of relations. For the joint entity and relation extraction task, we calculate the objective as:

$$\mathcal{L}_{joint}(w; \theta) = \mathcal{L}_{ner} + \mathcal{L}_{re} \tag{13}$$

where w refers to tokens, and θ denotes model parameters.

Since gold NER tagging information is only available during training, we therefore use pseudo labels predicted by CRF at RE inference time. This method is feasible because that the NER task does not involve label space information, and token-label interactions are only modeled during RE.

To directly compare with previous works, we also apply Adversarial Training (AT) [Bekoulis *et al.*, 2018], which can be used to improve the robustness of neural models by adding small perturbations to training data:

$$\mathcal{L}_{final} = \mathcal{L}_{joint}(w; \theta) + \mathcal{L}_{joint}(w + \eta_{adv}; \theta) \tag{14}$$

where η_{adv} is the worst-case perturbation.

4 Experiments

4.1 Dataset

To evaluate the performance of our model, we conduct experiments on two datasets. The first one is the CoNLL04 dataset [Roth and Yih, 2004], which contains sentences with annotated named entities and relations extracted from news articles. There are four entity types in the dataset ("*Location*", "*Organization*", "*Person*", and "*Other*") and five relation types ("*Kill*", "*Live in*", "*Located in*", "*OrgBased in*" and "*Work for*"). The second one is the ADE dataset [Gurulingappa *et al.*, 2012]. This dataset aims to extract two kinds of entities ("*Drugs*" and "*Diseases*") and relations about which drug is associated with which disease. To directly compare with previous works,

Models	Entity	Relation
Table Representation[1]	80.70	61.00
Multi-head + AT [2]	83.61	61.95
Relation-Metric with AT [3]	84.57	62.28
Multi-turn QA [4]∗	87.80	68.20
SpERT [5]∗	88.94	71.47
CMAN (ours)	**90.62**	**72.97**

Table 1: Comparison of our method with other competing approaches in terms of F1 score on the CoNLL04 dataset. Miwa and Sasaki[2014][1], Bekoulis et al.[2018][2], Tran and Kavuluru[2019][3], Li et al.[2019][4], Eberts and Ulges[2019][5]. Results with ∗ indicate that the study apply BERT as their core model.

Models	Entity	Relation
Joint Model[1]	79.50	63.40
Neural joint model[2]	84.60	71.40
Multi-head + AT [3]	86.73	75.52
Relation-Metric with AT [4]	87.02	77.19
SpERT [5]∗	89.25	79.24
CMAN (ours)	**89.40**	**81.14**

Table 2: The performance of our method and other competing approaches in terms of F1 score on the ADE dataset. Li et al.[2016][1], Li et al.[2017][2], Bekoulis et al.[2018][3], Tran and Kavuluru[2019][4], Eberts and Ulges[2019][5]. Results with ∗ indicate that the study appy BERT as their core model.

we evaluate our model using 10-fold cross-validation similar to prior approaches on the ADE dataset [Li *et al.*, 2017; Bekoulis *et al.*, 2018].

We adopt standard Precision (Prec), Recall (Rec) and F1 score to evaluate the model. We use the strict evaluation: the boundary and type of extracted entities should be both correct for NER; named entities and the type of their relationships should be both correct for RE.

Implementation Details

We regularize our network using dropout with a rate tuned on the development set (the dropout rate is 0.2 for embeddings, 0.1 and 0.3 for BiLSTM on two datasets respectively). We utilize 2 BSA units in our network (m=2) and set the dimensionality of hidden size d as 128. We choose 25 as the dimensionality of label embeddings d_l. The size of character embeddings is 128, while the dimensionality of ELMo [Peters *et al.*, 2018] is 1024. Adam optimizer with a learning rate of 0.0005 is used to optimize parameters. The training takes 180 epochs for convergence.

4.2 Quantitative Results

In this section, we present the performance of different models on two datasets. For the CoNLL04 dataset, we compare the proposed model with several competing approaches, and show the results in Table 1. It can be seen that our model achieves state-of-the-art performance on entity recognition and relation classification by obtaining 90.62 and 72.97 F1 respectively. Compared with prior competing SpERT method [Eberts and Ulges, 2019] that relies on pre-trained language model (BERT) [Devlin *et al.*, 2018], our approach gets absolute F1 improvements of 1.68% and 1.50% on NER and RE respectively. We find even stronger performance increases with respect to NER (+7.01%) and RE (+11.02%) when compared to the Multi-head + AT baseline [Bekoulis *et al.*, 2018] , which uses feature concatenation for capturing interactions in token-label spaces and applies multi-head mechanism for RE decoding. The above results indicate the effectiveness of our method and suggest that CMAN is able to model dense cross-modal interactions for joint entity and relation extraction.

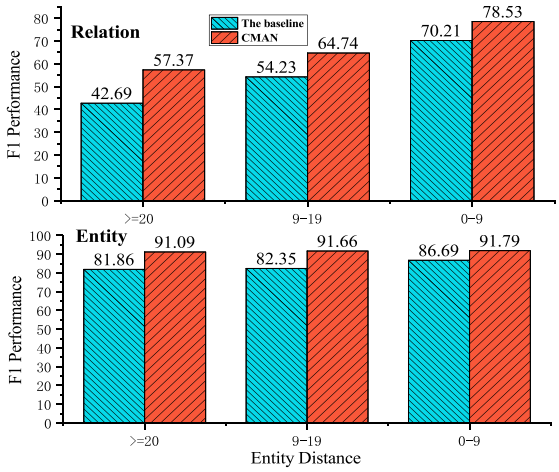

Figure 4: Comparison of the baseline and CMAN under different entity distances on the CoNLL04 development set. We use Multi-head + AT as the baseline, and and measure entity distance by computing the absolute character offset between the last character of the first occurring entity and last character of the second entity.

Table 2 presents the performance comparison between our approach and other competitive methods on the ADE dataset. Compared to the latest SpERT model, our approach only has a slight improvement (+0.15%) on NER. However, it can be found that our proposed model significantly outperforms SpERT by 1.90% F1 on RE. We think the reason may be that the ADE dataset contains less relations than CoNLL04, which is relatively easy for RE.

4.3 Performance against Entity Distance

Figure 4 shows F1 scores of the baseline model and CMAN under different entity distances on the CoNLL04 development set. Since Multi-head + AT [Bekoulis *et al.*, 2018] adopts CRF and multi-head mechanism for NER&RE decoding, we therefore set it as the baseline model. The CoNLL04 development set is split into three parts according to the metric of entity distance. We measure distance by computing the absolute character offset between the last character of the first occurring entity and the last character of the second mentioned entity, which is henceforth simply referred to as entity distance. The results indicate that CMAN significantly outperforms the baseline across different entity distances. In particular, the F1 score of CMAN is nearly 14.67% greater than that of the baseline for RE when the entity distance is more than 20 characters. It demonstrates that CMAN has a much greater advantage than

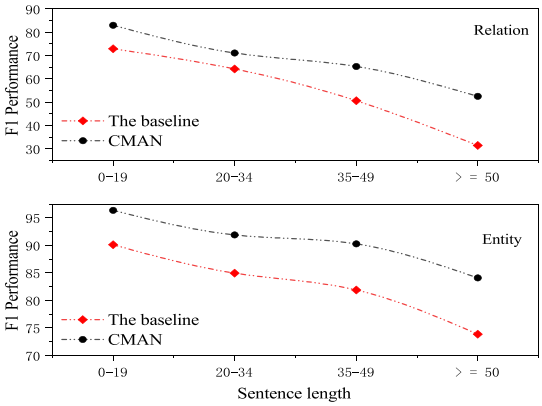

Figure 5: Comparison of the baseline and CMAN under different sentence lengths on the CoNLL04 development set, where Multi-head + AT is used as baseline.

Model	Entity	Relation
Predicted label	90.62	72.97
Golden label	90.44	72.80

Table 3: Performance of CMAN under different label sources on the CoNLL04 test set. At inference time, we evaluate the performance on either predicted labels or golden labels.

the baseline in dealing with entities that are far apart from each other. The reason is that CMAN can detect token-level self correlations by modeling dense intramodal interactions among tokens via the proposed BSA unit. Besides, we can notice that the effect of entity distance on RE is significantly higher than the impact on NER, likely due to that RE relies more on finding relevant distant entities.

4.4 Performance against Sentence Length

To investigate the influence of sentence length, we analyze the performance of baseline model and CMAN under grouped sentence lengths on the CoNLL04 development set, which is shown in Figure 5. Similarly, the Multi-head + AT model is used as the baseline. We partition the sentence length into four groups ([0-19], [20-34], [35-49], [$\geq$50]). We can observe that CMAN performs way better than the baseline under different sentence lengths. Moreover, the improvement achieved by CMAN is further enhanced when the sentence length consistently increases. In particular, CMAN outperforms the baseline by 10.26% and 21.42% F1 score for NER and RE respectively when the sentence length is large than 50. These results demonstrate that CMAN is more effective in terms of long sentences. It also verifies that our model can capture global dependencies of the whole sentence.

4.5 Performance against Label Source

In order to analyze the influence of different label sources, we evaluate our model with either golden labels or predicted labels at inference time on the CoNLL04 test set. To make the comparison fair, we set all hyperparameters unchanged but only feed the model with different NER tagging labels. The result is shown in Table 3. As can be seen from the Table, replacing predicted label with golden label barely changes the performance: the F1 score on NER slightly decreases by 0.18% and F1 on RE decreases by 0.17%. The reason may be that as the F1 score of entity recognition exceeds 90, the predicted label is very much the same as golden one. Therefore, predicted label sources can well carry label-space information, and the model can thus learn correct cross-modal interactions, which leads to unbiased performance.

Model	Entity	Relation
CMAN	90.62	72.97
- self-attention in BSA	90.16	70.18
- BLA unit	90.57	70.84
- Both units	89.34	68.95
replace BiLSTM with MLP	89.82	70.34

Table 4: Ablations on the CoNLL04 dataset.

4.6 Ablation Study

We conduct an ablation study to investigate the effectiveness of our attention units and network architecture in Table 4. Firstly, since the BiLSTM layer in BSA is a necessary component to encode tokens, we only remove the self-attention module to perform the ablation. We can observe that the F1 score drops by 0.46% and 2.79% for NER and RE tasks respectively, indicating self-attention is critical for capturing self-correlations among tokens. Secondly, we ablate the BLA unit and use self-aware token features for both tasks, and find that the performance slightly decreases, showing the beneficial effect of incorporating label-space information. Deleting both of attention units leads to further worse results on NER (-1.28%) and RE (-4.02%), which suggests that modeling dense cross-modal interactions plays a vital role in joint learning. Finally, to test the network architecture, we replace BiLSTM with Multi-Layer Perceptron (MLP) and find that the performance significantly drops to 89.82 and 70.34 F1 scores, implying the importance of building contextualized representations.

5 Conclusion

In this paper, we propose a deep Cross-Modal Attention Network (CMAN) for the task of joint entity-relation extraction. The network aims to capture dense cross-modal interactions by leveraging NER label information, where two basic attention units are proposed to model token-to-token and label-to-token correlations synergistically. We evaluate the proposed method on CoNLL04 and ADE datasets. The results show that CMAN achieves new state-of-the-art performance compared to other competing approaches.

Acknowledgments

This work is supported by The National Key Research and Development Program of China (2018YFB0204301, 2018YFB1800202, 2016YFB1000302, SQ2019ZD090149).

References

[Adel and Schütze, 2017] Heike Adel and Hinrich Schütze. Global normalization of convolutional neural networks for joint entity and relation classification. *arXiv preprint arXiv:1707.07719*, 2017.

[Antol *et al.*, 2015] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of ICCV*, pages 2425–2433, 2015.

[Ba *et al.*, 2016] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[Bach and Badaskar, 2007] Nguyen Bach and Sameer Badaskar. A review of relation extraction. *Literature review for Language and Statistics II*, 2:1–15, 2007.

[Bekoulis *et al.*, 2018] Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. Adversarial training for multi-context joint entity and relation extraction. *arXiv preprint arXiv:1808.06876*, 2018.

[Cui and Zhang, 2019] Leyang Cui and Yue Zhang. Hierarchically-refined label attention network for sequence labeling. *arXiv preprint arXiv:1908.08676*, 2019.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Eberts and Ulges, 2019] Markus Eberts and Adrian Ulges. Span-based joint entity and relation extraction with transformer pre-training. *arXiv preprint arXiv:1909.07755*, 2019.

[Guo *et al.*, 2019] Zhijiang Guo, Yan Zhang, and Wei Lu. Attention guided graph convolutional networks for relation extraction. *In Proceedings of ACL*, 2019.

[Gurulingappa *et al.*, 2012] Harsha Gurulingappa, Abdul Mateen Rajput, and Roberts. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of biomedical informatics*, 45(5):885–892, 2012.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of CVPR,2016*, pages 770–778, 2016.

[Katiyar and Cardie, 2017] Arzoo Katiyar and Claire Cardie. Going out on a limb: Joint extraction of entity mentions and relations without dependency trees. In *Proceedings of ACL,2017*, pages 917–928, 2017.

[Lafferty *et al.*, 2001] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[Li *et al.*, 2016] Fei Li, Yue Zhang, Meishan Zhang, and Donghong Ji. Joint models for extracting adverse drug events from biomedical text. In *IJCAI*, volume 2016, pages 2838–2844, 2016.

[Li *et al.*, 2017] Fei Li, Meishan Zhang, Guohong Fu, and Donghong Ji. A neural joint model for entity and relation extraction from biomedical text. *BMC bioinformatics*, 18(1):198, 2017.

[Li *et al.*, 2019] Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. Entity-relation extraction as multi-turn question answering. *arXiv preprint arXiv:1905.05529*, 2019.

[Miwa and Bansal, 2016] Makoto Miwa and Mohit Bansal. End-to-end relation extraction using lstms on sequences and tree structures. *arXiv preprint arXiv:1601.00770*, 2016.

[Miwa and Sasaki, 2014] Makoto Miwa and Yutaka Sasaki. Modeling joint entity and relation extraction with table representation. In *Proceedings of EMNLP, 2014.*, pages 1858–1869, 2014.

[Miwa *et al.*, 2009] Makoto Miwa, Rune Sætre, Yusuke Miyao, and Jun'ichi Tsujii. A rich feature vector for protein-protein interaction extraction from multiple corpora. In *Proceedings of EMNLP*, pages 121–130, 2009.

[Nadeau and Sekine, 2007] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[Peters *et al.*, 2018] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[Roth and Yih, 2004] Dan Roth and Wen-tau Yih. A linear programming formulation for global inference in natural language tasks. In *Proceedings of CoNLL-2004 at HLT-NAACL 2004*, pages 1–8, 2004.

[Tran and Kavuluru, 2019] Tung Tran and Ramakanth Kavuluru. Neural metric learning for fast end-to-end relation extraction. *arXiv preprint arXiv:1905.07458*, 2019.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[Wang *et al.*, 2016] Linlin Wang, Zhu Cao, Gerard De Melo, and Zhiyuan Liu. Relation classification via multi-level attention cnns. 2016.

[Wang *et al.*, 2018] Guoyin Wang, Chunyuan Li, Wenlin Wang, Yizhe Zhang, Dinghan Shen, and Zhang. Joint embedding of words and labels for text classification. *arXiv preprint arXiv:1805.04174*, 2018.

[Yu *et al.*, 2019] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the CVPR,2019*, pages 6281–6290, 2019.

[Zheng *et al.*, 2017] Suncong Zheng, Feng Wang, Hongyun Bao, Yuexing Hao, Peng Zhou, and Bo Xu. Joint extraction of entities and relations based on a novel tagging scheme. *arXiv preprint arXiv:1706.05075*, 2017.