A Densely Connected Criss-Cross Attention Network for Document-level Relation Extraction

Liang Zhang[1] **Yidong Cheng** [1]
[1] Department of Artificial Intelligence, School of Informatics, Xiamen University
{lzhang,ydcheng}@stu.xmu.edu.cn

Abstract

Document-level relation extraction (RE) aims to identify relations between two entities in a given document. Compared with its sentence-level counterpart, document-level RE requires complex reasoning. Previous research normally completed reasoning through information propagation on the mention-level or entity-level document-graph, but rarely considered reasoning at the entity-pair-level. In this paper, we propose a novel model, called **Dense**ly Connected **C**riss-**C**ross Attention **Net**work (**Dense-CCNet**), for document-level RE, which can complete logical reasoning at the entity-pair-level. Specifically, the Dense-CCNet performs entity-pair-level logical reasoning through the Criss-Cross Attention (**CCA**), which can collect contextual information in horizontal and vertical directions on the entity-pair matrix to enhance the corresponding entity-pair representation. In addition, we densely connect multiple layers of the CCA to simultaneously capture the features of single-hop and multi-hop logical reasoning. We evaluate our Dense-CCNet model on three public document-level RE datasets, DocRED, CDR, and GDA. Experimental results demonstrate that our model achieves state-of-the-art performance on these three datasets.

以往的研究通常是通过在提及级或实体级文档图上的信息传播来完成推理的

但很少考虑在实体对级上的推理

Criss-Cross Attention (CCA)

在实体对矩阵上收集 水平和 垂直 方向的上下文信息，以增强相应的实体对表示

我们密集地连接CCA的多层，以同时捕获单跳和多跳逻辑推理的特征

1 Introduction

Relation extraction (RE) aims to identify relationships between two entities from raw texts. It is of great importance to many real-world applications such as knowledge base construction, question answering, and biomedical text analysis (Xu et al., 2021a). Most of the existing work focuses on sentence-level RE, which predicts the relationship between entities in a single sentence (Zhang et al., 2018; Soares et al., 2019). However, large amounts of relationships are expressed by multiple sentences in real life (Yao et al., 2019). According to the statistics of the DocRED (Yao et al., 2019) dataset which is obtained from Wikipedia documents, at least 40.7% of relations can only be extracted from multiple sentences. Therefore, researches on document-level RE models that can extract relational facts across sentences have gained increasing attention recently.

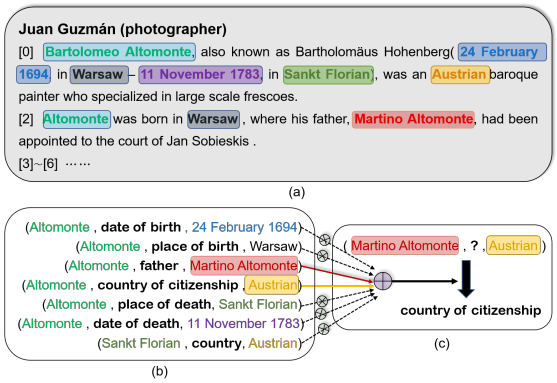

Figure 1: An example comes from the DocRED dataset, which shows that triples with overlapping entities provide important information for reasoning the complex inter-sentential relations. (a) is a document, in which different colors represent different entities. (b) lists some intra-sentential triplets, which can be easily identified. (c) shows a triple whose relationship requires logical reasoning techniques to be recognized. The arrows between (b) and (c) indicate the correlation among triples.

具有重叠实体的三元组为推理复杂的句子间关系提供了重要的信息。

Compared with sentence-level RE, the main challenge is that many relations in document-level RE could only be extracted through the technique of reasoning. Since these relationship facts are not explicitly expressed in the document, the model must captures the correlation between the relationships to infer these relationships. Therefore, capturing the relevance of the relationships is essential to improve the reasoning ability of document-level RE models. Figure 1 shows an example from the DocRED dataset. The Figure 1b lists the intra-sentential triplets, such as (*Altomonte, date of birth, 24 February 1694*), (*Altomonte, father, Martino Altomonte*), and (*Altomonte, country of citizen-*

捕捉这些关系的相关性 对RE很重要

两个相关的实体出现在同一个句子中的关系很好识别

但对于模型没有显示地表达的关系识别比较困难

ship, Austrian), which could be easily recognized since two related entities appear in the same sentence. However, it is non-trivial to predict the inter-sentential relations between *Martino Altomonte* and *Austrian* because the document does not explicitly express the relationship between them. In fact, the model needs to firstly capture the correlation among (*Altomonte, father, Martino Altomonte*), (*Altomonte, country of citizenship, Austrian*), and (*Martino Altomonte, country of citizenship, Austrian*) and use logical reasoning techniques to identify this complex relationship, as shown in Figure 1c.

这些方法通过token、提及或实体之间的信息传递来捕获关系之间的相关性，这可能是间接的和低效的

To extract these complex relationships, most current approaches constructed a document-level graph based on heuristics, structured attention, or dependency structures (Zeng et al., 2020; Nan et al., 2020; Christopoulou et al., 2019; Wang et al., 2020a), and then perform inference with graph convolutional network (GCN) (Guo et al., 2019; Kipf and Welling, 2016) on the graph. Meanwhile, considering the transformer architecture can implicitly model long-distance dependencies, some studies (Tang et al., 2020; Zhou et al., 2021) directly applied pre-trained models rather than explicit graph reasoning (Zhang et al., 2021). These methods captures the correlation between relationships through the information transfer between tokens, mentions or entities, which can be indirect and inefficient.

利用实体对之间的信息传递来更有效和直接地捕捉关系之间的相关性。

In this paper, we use the information transfer between the entity-pairs to capture the correlation between relationships more efficiently and directly. Moreover, as it can be seen in Figure 1, only (*Altomonte, father, Martino Altomonte*) and (*Altomonte, country of citizenship, Austrian*) triples, rather than the other triples, provide important information to infer the relations between *Martino Altomonte* and *Austrian*. Inspired by this phenomenon, we guess that the interaction between the triples with overlapping entities is a reasonable way of entity-pair-level reasoning.

推测具有重叠实体的三元组之间的交互是一种合理的实体对级推理的方法

Therefore, we propose a novel Dense-CCNet model by integrating the Criss-Cross Attention (**CCA**) (Huang et al., 2019) into the densely connected framework (Huang et al., 2017). The CCNet model (Huang et al., 2019) is an advanced semantic segmentation model recently proposed in the field of computer vision, which captures global context information from full-image through the CCA (as shown in Figure 2). The CCA applied to the entity-pair matrix can realize the interaction between entity-pairs with overlapping entities, which can complete the logical reasoning of the entity-pair-level. To fully capture the features of single-hop and multi-hop reasoning, we stack the multi-layer modules CCA modules by the densely connected framework. The lower layers in Dense-CCNet can capture local interdependence among entity-pairs and complete single-hop logical reasoning, while the upper layers can capture global interdependence among entity-pairs and complete multi-hop logical reasoning.

应用于实体对矩阵的CCA可以实现：具有重叠实体的实体对之间的交互，从而完成实体对级的逻辑推理

为了充分捕捉单跳和多跳推理的特点，我们将多层模块的CCA模块采用密集连接的框架进行叠加

低层捕获：实体对之间的局部相互依赖和完整的单跳逻辑推理

高层捕获：实体对之间的全局相互依赖性和完整的多跳逻辑推理

Since the CCA can only complete the reasoning mode of **A**$\rightarrow * \rightarrow$**B**, we expand the field (which single-layer CCA can pay attention to) to cover a wider range of reasoning modes, such as **A**$\rightarrow * \leftarrow$**B**, **A**$\leftarrow * \leftarrow$**B**, and **A**$\leftarrow * \rightarrow$**B**. In addition, we found that more than 90% of the entity pairs are irrelevant (that is, there is no relationship between two entities) in the document, and these entity pairs may limit the model's reasoning ability. To reduce the influence of unrelated entity-pairs, we use two techniques: **(1) Clustering loss**: The clustering loss separates the related entity-pairs (that is, there is relationship between two entities) from the unrelated entity-pairs in the representation space. **(2) Attention bias**: We add a bias term to a bias term to the attention score of the CCA, which makes the CCA pay more attention to related entity pairs.

In summary, our main contributions are as follows:

- We introduce the Dense-CCNet module that can more directly and effectively model the correlation between relationships through the entity-pair-level reasoning.
- We introduce four methods to further improve the reasoning ability of the CCA: Dense connection, Expanding the Field of Attention, Clustering loss, and Attention bias.
- Experimental results on three public document-level RE datasets shows that our Dense-CCNet model can achieve state-of-the-art performance.

2 Methodology

In this section, we elaborate on our Dense-CCNet mode. Our entire model (as shown in Figure 2) is mainly composed of three parts: Encoder module (Sec. 2.1), Dense-CCNet module (Sec. 2.2), and Classifier module (Sec. 2.3).

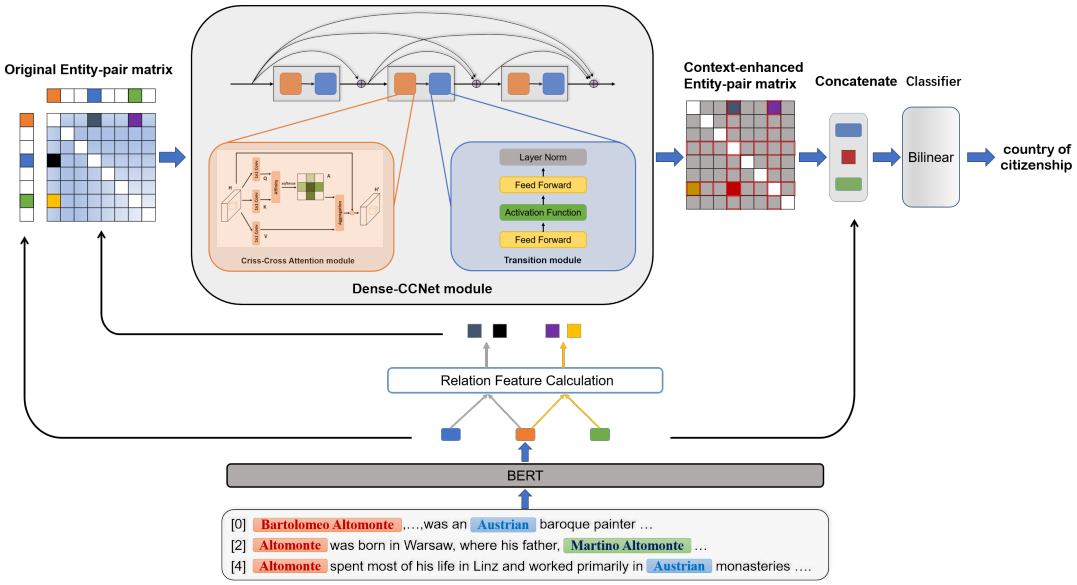

Figure 2: The overall architecture of our Dense-CCNet-based document-level RE model. Firstly, the BERT model encodes the input document to obtain the context embedding of each words, and then we obtains the representations of the entities (h_{e_s},h_{e_o}) through a pooling operation. Secondly, the relation features ($M_{s,o}/M_{o,s}$) of all entity pairs are calculated through the Relation Feature Calculation module, which is used to construct the original entity-pair matrix (M). Thirdly, the Dense-CCNet module transforms M into a context-enhanced entity-pair matrix (M'). Finally, the context-enhanced relation features ($M'_{s,o}$) of the entity pairs (e_s, e_o), the subject entity embedding (h_{e_s}), and object entity embedding (h_{e_o}) are concatenated and inputted to the classifier to predict the relationship.

2.1 Encoder Module

We first treat the document D as a sequence of words, i.e. $D=\{w_i\}_i^L$, where L is the total number of words in the document. Then, we insert special symbols $\langle E_i\rangle$ and $\langle\backslash E_i\rangle$ to mark the start and end positions of the mention respectively, where E_i is the entity id of the mention. It is an improved version of entity marker technology (Zhou et al., 2021; Zhang et al., 2021; Shi and Lin, 2019; Soares et al., 2019) by introducing entity id information which can help align the information of different mentions from the same entity. Finally, we leverage the pre-trained language model as an encoder to convert documents D into a sequence of contextual embeddings as follows:

$$H = [h_1, ..., h_L] = BERT([w_1, ..., w_L]). \quad (1)$$

We take the embedding of $\langle E_i\rangle$ at the start of mentions m_j as the mention embedding h_{m_j}. Then, we leverage logsumexp pooling (Jia et al., 2019), a smooth version of max pooling, to obtain the embedding h_{e_i} of entity e_i which contains p_i mentions $\{m_j^i\}_j^{p_i}$:

$$h_{e_i} = \log \sum_{j=1}^{p_i} exp(h_{m_j^i}). \quad (2)$$

After obtaining the embedding of all entities in the document, we construct an **Entity-Pair Matrix** $M \in R^{N_e \times N_e \times d}$ through the Relation Feature Calculation module, where N_e refers to the number of entities and d is the dimension of the relation feature vector. The $M_{s,o}$ item in M represents the relation feature vector between the entity e_s and the entity e_o, which is calculated as follows:

在正常ATLOP中的嵌入上将头尾实体对嵌入组合 通过FNN得到 实体对的嵌入

$$\begin{aligned} M_{s,o} &= FFNN([u_s, u_o]), \\ u_s &= \tanh(W_s[h_{e_s}, h_{doc}, c_{s,o}]), \\ u_o &= \tanh(W_o[h_{e_o}, h_{doc}, c_{s,o}), \end{aligned} \quad (3)$$

where h_{e_s} is subject entity embedding, h_{e_o} is object entity embedding, h_{doc} is document embedding, and $c_{s,o}$ is entity-pair-aware context feature, $FFN()$ refers to a feed-forward neural network, W_o, W_s is the learnable weight matrix.

We use the embedding of the document start token "[CLS]" as the document embedding h_{doc}, which can help aggregate cross-sentence information and provide document-aware representation.

The entity-pair-aware context feature $c_{s,o}$ represents the contextual information in the document that the entity e_s and the entity e_o pay attention to

together. The $c_{s,o}$ is formulated as follows:

与ATLOP中的上下文Cs，o类似

$$c_{s,o} = \sum_{i=1}^{L} A_{s,i} \cdot A_{o,i} \cdot h_i, \quad (4)$$

where $A_{s,i}$ is the attention score of the entity e_s paying attention to the i-th token h_i in the document.

2.2 Dense-CCNet Module

In this part, we introduce the Dense-CCNet module in detail. As shown in Figure 2, the Dense-CCNet module consists of densely connected N identical layers that are composed of two sub-modules: the **Criss-Cross Attention (CCA) module** and the **Transition module**.

CCNet model中CCA:

实体对矩阵中的每个 实体对(s,o) 都可以通过CCA模块关注其他实体对 在水平(s,i)和垂直(i,o)方向上 的关系特征

We followed the CCNet model (Huang et al., 2019) for the CCA module. Each entity pair in the entity-pair matrix can pay attention to the relation feature of other entity pairs in horizontal and vertical directions through the CCA module. The CCA module can be formulated as follows:

$$M_{s,o} = \sum_{i=1}^{N_e} \left(A_{(s,o)\to(s,i)} M_{s,i} + A_{(s,o)\to(i,o)} M_{i,o} \right)$$

where $A_{(s,o)\to(s,i)}$ is the attention score of the $M_{s,o}$ paying attention to the $M_{s,i}$. Therefore, the CCA module can complete entity-pair-level one-hop reasoning on the entity-pair matrix, and it is possible to complete multi-hop reasoning by stacking multiple layers of the CCA module.

CCA可以完成实体对级别的单条推理，堆叠多层CCA可以完成多跳推理

However, simply using Recurrent Criss-Cross Attention (**RCCA**) (Huang et al., 2019) to complete the logical reasoning of the entity-pair level may have several problems: **(1)** The RCCA only focuses on the high-level multi-hop inference feature and ignores the low-level single-hop inference feature which is also very important for document-level RE. **(2)** The CCA module can only model the reasoning mode of $\mathbf{A}\to * \to\mathbf{B}$, but cannot model the reasoning mode of $\mathbf{A}\to * \leftarrow\mathbf{B}$, $\mathbf{A}\leftarrow * \leftarrow\mathbf{B}$, and $\mathbf{A}\leftarrow * \to\mathbf{B}$. **(3)** Since most of the entity pairs are irrelevant in the document, the entity-pair matrix M contains a lot of noise which may affect the reasoning ability of the model. Therefore, distinguishing related entity-pairs from unrelated entity-pairs and strengthening the interaction of the relationship feature vectors of related entity-pairs is the key to improving The reasoning ability of the model. To solve these problems, we have introduced the following methods:

CCA的缺点：
1. 重点关注了高级多跳推理特性，而忽略了低级单跳推理特性
2. 建模能力有限
3. 由于文档中大多数实体对是无关的，实体对矩阵M包含大量的噪声，可能会影响模型的推理能力

区分相关的实体对和不相关的实体对

加强相关实体对的关系特征向量的相互作用

是提高模型推理能力的关键

Dense Connection: Since dense connections can reuse the features of low-level networks, we stack multiple layers of the CCA modules through the densely connected framework to solve the problem **(1)** . In addition, the dense connection can also reduce noise propagation to a certain extent.

CCA只关注多条推理

Dense Connection 可以重新利用低级的单跳推理

并在一定程度上降低了噪声的传播。

Expanding the Field of Attention: To allow the CCA module to model more inference modes, we modify the CCA module as follows:

CCA建模能力有限

Expanding the Field of Attention 可以覆盖更广泛的推理模式

$$M_{s,o} = \sum_{i=1}^{N_e} \Big(A_{(s,o)\to(s,i)} M_{s,i} + A_{(s,o)\to(i,o)} M_{i,o} + A_{(s,o)\to(s,i)} M_{s,i} + A_{(s,o)\to(i,o)} M_{i,o} \Big)$$

The modified CCA module can cover a wider range of reasoning modes including: $\mathbf{A}\to * \to\mathbf{B}$, $\mathbf{A}\to * \leftarrow\mathbf{B}$, $\mathbf{A}\leftarrow * \leftarrow\mathbf{B}$, and $\mathbf{A}\leftarrow * \to\mathbf{B}$.

Clustering Loss: We design a clustering loss function that separates the related entity pairs and the unrelated entity pairs in the feature space to reduce the influence of unrelated entity pairs on the inference process. Clustering Loss is formulated as follows:

区分相关的实体对和不相关的实体对加强相关实体对的关系特征向量的相互作用

要区分实体对之间是否相关

Clustering Loss区分相关实体对和不相关实体对

$$L_{dist} = (max\{0, (\mu + cos(v_0, v_1))\})^2,$$
$$L_{var1} = \sum_{i \in N_{pos}} (max\{0, (\lambda - cos(v_1, f_i))\})^2,$$
$$L_{var0} = \sum_{j \in N_{neg}} (max\{0, (2\lambda - cos(v_0, f_j))\})^2,$$
$$L_C = \alpha L_{dist} + \beta L_{var0} + \gamma L_{var1},$$

优化了嵌入

where N_{neg} is the set of the irrelevant entity pairs, N_{pos} is the set of the related entity pairs, v_0 is the average vector of the feature vectors of the entity pairs in the N_{neg}, v_1 is the average vector of the feature vectors of the entity pairs in the N_{pos}, f_i is the feature vector of the i-th entity pair.

Attention Bias: To make the CCA more focused on the related entity pairs, we added a bias to the attention score of the CCA:

使CCA更关注相关的实体对

在CCA的注意评分上增加了一个偏差

$$A_{(s,o)\to\star} = softmax(s_{(s,o)\to\star} + bias_\star), \quad (5)$$

where, $bias\star$ is a bias term of the entity pair $\star$, which reflects the confidence that the entity pair $\star$ is a related entity pair. $bias_\star$ is predicted and trained through a feed-forward neural network:

$$bias_* = FFNN(f_\star),$$
$$L_{bias} = BCE(bias_\star, label_{01}) \quad (6)$$

Where BCE is a cross-entropy loss function, and $label_{01}$ is the 0-1 label of the entity pair.

The Transition module controls the dimensions of the new features generated by each layer of the Dense-CCNet model, which reduces the computational complexity of the model.

2.3 Classification Module

We use the Dense-CCNet to convert the original entity-pair matrix M into a new context-enhanced entity-pair matrix M'. Given an entity pair (e_s, e_o), we first concatenate the two entity embedding (h_{e_s}, h_{e_o}) and new relation feature M'_{so}, then we obtain the distribution of relationship via a bilinear function. Formally, we have:

$$\begin{aligned} z_s &= \tanh(W'_s[h_{e_s}, M'_{so}]), \\ z_o &= \tanh(W'_o[h_{e_o}, M'_{so}]), \\ P(r|e_s, e_o) &= \sigma(z_s^T W_r Z_o + b_r). \end{aligned} \tag{7}$$

For the loss function, we use adaptive-thresholding loss (Zhou et al., 2021), which learns an adaptive threshold for each entity pair. The loss function is broken down into two parts as shown below:

$$L_1 = -\sum_{r \in P_D} \log\left(\frac{\exp(logit_r)}{\sum_{r' \in P_D \cup \{TH\}} exp(logit_{r'})}\right),$$

$$L_2 = -\log\left(\frac{\exp(logit_{TH})}{\sum_{r' \in N_D \cup \{TH\}} exp(logit_{r'})}\right),$$

$$L_{adap} = L_1 + L_2,$$

where TH is an introduced class to separate positive classes and negative classes: positive classes would have higher probabilities than TH, and negative classes would have lower probabilities than TH, P_D and N_D are the positive classes set and negative classes set in document D respectively.

Finally, our total loss function is defined as follows:

$$L = L_{adap} + L_{bias} + L_C, \tag{8}$$

3 Experiments

3.1 Datasets

We evaluate our Dense-CCNet model on three public document-level RE datasets. The statistics of the datasets could be found in Appendix A.

- **DocRED** (Yao et al., 2019): The DocRED is a large-scale crowdsourced dataset for document-level RE, which was constructed from Wikipedia and Wikidata. The DocRED contains 3053 documents for training, 1000 for validating, and 1000 for the test. It involves 97 types of target relations in total, and each document approximately contains 26 entities on average.
- **CDR** (Li et al., 2016): The CDR is a biomedical dataset is constructed by using the PubMed abstracts, which aims to predict the binary interactions between Chemical and Disease concepts. The CDR contains only one relationship and consists of 1500 human-annotated documents in total. The CDR are equally split into training, development, and test sets.
- **GDA** (Wu et al., 2019): Similar to the CDR, the GDA is also a dataset in the biomedical domain, but is constructed by distant supervision from the MEDLINE abstracts. The GDA contains 29192 documents as the training set and 1000 as the test set. Since there is no development set, we follow (Christopoulou et al., 2019) to divide the training set into two parts according to the ratio of 8:2 and use them as training set and development set respectively.

3.2 Experimental Settings

Our model was implemented based on PyTorch. We used cased BERT-base (Devlin et al., 2018) as the encoder on DocRED and SciBERT-base (Beltagy et al., 2019) on CDR and GDA. We set the number of layers of Dense-CCNet to 3. Our model is optimized with AdamW (Loshchilov and Hutter, 2017) with a linear warmup (Goyal et al., 2017) for the first 6% steps followed by a linear decay to 0. All hyper-parameters are tuned on the development set, some of which are listed in Appendix B.

3.3 Results on the DocRED Dataset

The experimental results of our model on DocRED are shown in Table 1. We followed (Yao et al., 2019) and used F_1 and IgnF_1 as the evaluation metrics to evaluate the overall performance of the model. IgnF_1 denotes the $F1$ score excluding the relational facts that are shared by the training and dev/test sets. We compare the Dense-CCNet model with the following two types of models on the DocRED dataset :

- **Graph-based Models**: these models first construct the document-graph from the document,

Model	Dev		Test	
	IgnF_1	F_1	IgnF_1	F_1
GEDA-BERT$_{\text{base}}$(Li et al., 2020)	54.52	56.16	53.71	55.74
LSR-BERT$_{\text{base}}$ (Nan et al., 2020)	52.43	59	56.97	59.05
GLRE-BERT$_{\text{base}}$(Wang et al., 2020a)	-	-	55.4	57.4
HeterGSAN-BERT$_{\text{base}}$(Xu et al., 2021b)	58.13	60.18	57.12	59.45
GAIN-BERT$_{\text{base}}$(Zeng et al., 2020)	59.14	61.22	59	61.24
BERT$_{\text{base}}$(Wang et al., 2019)	-	54.16	-	53.2
BERT-TS$_{\text{base}}$(Wang et al., 2019)	-	54.42	-	53.92
HIN-BERT$_{\text{base}}$(Tang et al., 2020)	54.29	56.31	53.7	55.6
CorefBERT$_{\text{base}}$(Ye et al., 2020)	55.32	57.51	54.54	56.96
ATLOP-BERT$_{\text{base}}$(Zhou et al., 2021)	59.22	61.09	59.31	61.3
DocuNet-BERT$_{\text{base}}$(Zhang et al., 2021)	59.86	61.83	59.93	61.86
Dense-CCNet-BERT$_{\text{base}}$	**60.72**($\pm$**0.12**)	**62.74**($\pm$**0.15**)	**60.46**	**62.55**

Table 1: Results (%) on the development and test set of the DocRED. We follow ATLOP (Zhou et al., 2021) and DocuNet (Zhang et al., 2021) for the scores of all baseline models. The results on the test set are obtained by submitting to the official Codalab.

and then perform inferences through GCN (Kipf and Welling, 2016) on the graph. We include GEDA (Li et al., 2020), LSR (Nan et al., 2020), GLRE (Wang et al., 2020a), GAIN (Zeng et al., 2020), and HeterGSAN (Xu et al., 2021b) for comparison.

- **Transformer-based Models**: These models directly use the pre-trained language models for document-level RE without graph structures. we compared BERT$_{\text{base}}$ (Wang et al., 2019), BERT-TS$_{\text{base}}$ (Wang et al., 2019), HIN-BERT$_{\text{base}}$ (Tang et al., 2020), CorefBERT$_{\text{base}}$ (Ye et al., 2020), CorefBERT$_{\text{base}}$(Ye et al., 2020), and ATLOP-BERT$_{\text{base}}$ (Zhou et al., 2021) with our model.

In addition, we also consider the DocuNet (Zhang et al., 2021) model in the comparison, which formulates document-level RE as a semantic segmentation problem.

As shown in Table 1, our Dense-CCNet model achieved **62.74**%F_1 and **62.55**% F_1 in the training set and test set, which outperforms the state-of-the-art model with **0.91**% F_1 and **0.69**% F_1 respectively. Compared with the GAIN model that is the state-of-the-art model of graph-based methods, our model exceeds it by **1.52**% F_1 on the dev set and **1.31**% F_1 on the test set. This proves that the logical reasoning on the entity-pairs level is more effective than previous methods on mentions or entities level.

3.4 Results on the Biomedical Datasets

On the CDR and GDA data sets, we compared BRAN (Verga et al., 2018), EoG (Christopoulou et al., 2019), LSR (Nan et al., 2020), DHG (Zhang et al., 2020c), GLRE (Wang et al., 2020a), ATLOP (Zhou et al., 2021), and DocuNet (Zhang et al., 2021) with our model. The experimental results on two biomedical datasets are shown in Table 2.

Our Dense-CCNet-SciBERT$_{\text{base}}$ model obtained **77.06**($\pm$**0.71**)% F_1 and **86.44**($\pm$**0.25**)% F_1 on two data sets respectively, which is also the new state-of-the-art result. The Dense-CCNet-SciBERT$_{\text{base}}$ improved the F_1 score by **0.76**% and **1.14**% on CDR and GDA compared with DocuNet-SciBERT$_{\text{base}}$. These results demonstrate the strong applicability and generality of our approach in the biomedical field.

3.5 Ablation Study

We conducted an ablation experiment to validate the effectiveness of different components of our Dense-CCNet model on the development set of the DocRED dataset. The results are listed in Table 3, where **w/o Dense Connection** replaces the Densely connected Criss-Cross Attention with the Recurrent Criss-Cross Attention (RCCA), **w/o Expanding Attention** uses standard Criss-Cross Attention and does not extend the field of attention, **w/o Clustering Loss** and **w/o Attention Bias** respectively removes the Clustering Loss and the Attention Bias from our model.

Model	CDR	GDA
BRAN (Verga et al., 2018)	62.1	-
EoG (Christopoulou et al., 2019)	63.6	81.5
LSR (Nan et al., 2020)	64.8	82.2
DHG (Zhang et al., 2020c)	65.9	83.1
GLRE (Wang et al., 2020a)	68.5	-
SciBERT$_{\text{base}}$ (Beltagy et al., 2019)	65.1	82.5
ATLOP-SciBERT$_{\text{base}}$(Zhou et al., 2021)	69.4	83.9
DocuNet-SciBERT$_{\text{base}}$(Zhang et al., 2021)	76.3	85.3
Dense-CCNet-SciBERT$_{\text{base}}$	**77.06**	**86.44**

Table 2: F1 scores (%) on test sets of the CDR and the GDA.

Model	Dev	
	IgnF_1	F_1
Dense-CCNet-BERT	**60.72**	**62.74**
w/o Dense Connection	59.23	61.12
w/o Expanding Attention	59.82	61.91
w/o Clustering Loss	59.97	62.02
w/o Attention Bias	60.65	61.60

Table 3: Ablation study of the Dense-CCNet on the development set of the DocRED. We turn off different components of the model one at a time.

Layer-number	Dev	
	IgnF_1	F_1
2-Layer	59.41	61.44
3-Layer	**60.72**	**62.74**
4-Layer	60.30	62.27

Table 4: Performance of the Dense-CCNet with the different numbers of layers on the development set of the DocRE.

From Table 3, we can observe that the w/o Dense leads to a drop of **1.62%** $F1$, which shows that the features of low-level inference are very helpful for relation extraction and the features of high-level inference may contain noises. The w/o Expanding Attention caused a performance drop of **0.83%** $F1$, which indicates that document-level RE may include multiple inference modes and our model can effectively expand the reasoning mode of the Criss-Cross Attention through the Expanding the Field of Attention technology.

低级推理的特征对关系的提取非常有帮助，而高级推理的特征可能包含噪声

文档级RE可能包含多种推理模式，我们的模型可以通过扩展注意领域技术有效地扩展交叉注意的推理模式。

The w/o Clustering Loss module and w/o Attention Bias module led to performance degradation of **0.72%** $F1$ and **1.14%** $F1$ points respectively, which reflects that reducing noise (irrelevant entity-pairs) may be the key to further improving entity-pair level inference. We guess that the most ideal entity-pair-level reasoning method may be to only propagate information between related entity pairs.

最理想的实体对级推理方法可能是只在相关的实体对之间传播信息。

In addition, we also introduced the ablation study of the number of layers of the Dense-CCNet, and the experimental results are shown in Table 4. When the number of layers increases from 2 to 3, our model can capture more multi-hop inference features, so the performance of the model is improved by **1.3%** $F1$. However, when the number of layers is increased to 4, the performance drops slightly by **0.37%** $F1$ points. The possible reasons is that the noise has a greater impact on the high-level feature or the model falls into over-fitting.

3.6 Case Study

We followed GAIN (Zeng et al., 2020) to select the same example and conduct a case study to further illustrate that our Dense-CCNet model can effectively capture the interdependence among entity-pairs and perform entity-pair-level logical reasoning compared with the baseline.

The experimental results are shown in Figure 3. Figure 3c demonstrates that our model has better logical reasoning ability than the baseline. Figure 3a shows that the entity pair (*"Without Me", May 26, 2002*) has more attention to the entity pairs (*"Without Me", The Eminem Show*) and (*The Eminem Show, May 26,2002*), which indicates that our model could capture the correlation these among entity-pairs.

4 Related Work

Sentence-level RE: Early research on RE focused on sentence-level RE, which predicts the relationship between two entities in a single sentence. Many approaches (Zeng et al., 2015; Feng et al., 2018; Zhang et al., 2020b,a; Wang et al., 2020b; Ye et al., 2020; Yu et al., 2020; Wu et al., 2021; Chen et al., 2021; Zheng et al., 2021) have been proven to effectively solve this problem. Since many relational facts in real applications can only be recognized across sentences, sentence-level RE

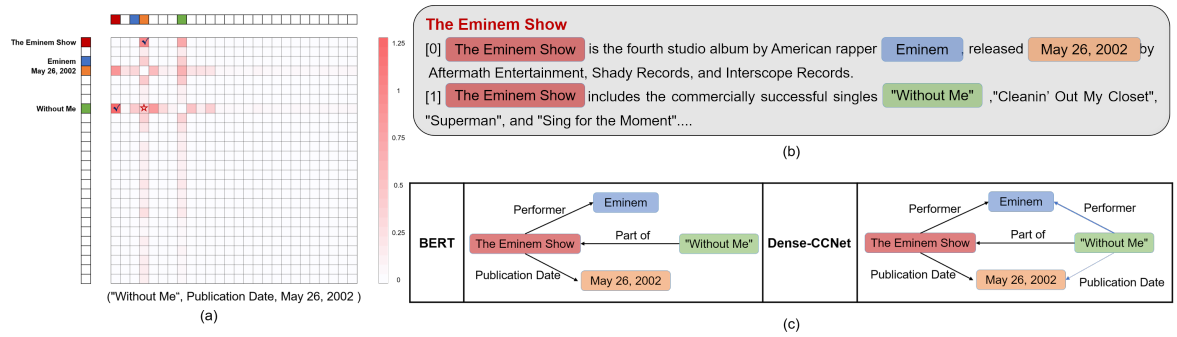

Figure 3: Case study of our Dense-CCNet mode and the baseline model. (c) shows that our model has better logical reasoning ability than the baseline. (a) visualize the attention scores of entity pairs (*"Without Me", May 26, 2002*) paying attention to other entity pairs, which shows that our model can effectively capture the correlation among entity-pairs.

face an inevitable restriction in practice.

Document-level RE: To solve the limitations of sentence-level RE in reality, a lot of recent work gradually shift their attention to document-level RE. Since graph neural network(GNN) can effectively model long-distance dependence and complete logical reasoning, Many methods based on document-graphs are widely used for document-level RE. Specifically, they first constructed a graph structure from the document, and then applied the GCN (Kipf and Welling, 2016; Huang et al., 2017) to the graph to complete logical reasoning. The graph-based method was first introduced by (Quirk and Poon, 2016) and has recently been extended by many works (Christopoulou et al., 2019; Li et al., 2020; Zhang et al., 2020c; Zhou et al., 2020; Wang et al., 2020a; Nan et al., 2020; Zeng et al., 2020; Xu et al., 2021b). (Li et al., 2020) proposed the Graph Enhanced Dual Attention network (GEDA) model and used it to characterize the complex interaction between sentences and potential relation instances. (Zeng et al., 2020) propose Graph Aggregation-and-Inference Network (GAIN) model. GAIN first constructs a heterogeneous mention-level graph (hMG) to model complex interaction among different mentions across the document and then constructs an entity-level graph (EG), finally uses the path reasoning mechanism to infer relations between entities on EG. (Nan et al., 2020) proposed a novel LSR model, which constructs a latent document-level graph and completes logical reasoning on the graph.

In addition, due to the pre-trained language model based on the transformer architecture can implicitly model long-distance dependence and complete logical reasoning, some studies (Tang et al., 2020; Zhou et al., 2021; Wang et al., 2019) directly apply pre-trained model without introducing document graphs. (Zhou et al., 2021) proposed an ATLOP model that consists of two parts: adaptive thresholding and localized context pooling, to solve the multi-label and multi-entity problems. Recently, the state-of-the-ar model, DocuNet (Zhang et al., 2021), formulates document-level RE as semantic segmentation task and capture global information among relational triples through the U-shaped segmentation module (Ronneberger et al., 2015).

However, none of the models completes the logical reasoning for document-level RE through the information propagation between the entity-pairs. Our Dense-CCNet model can capture the correlation among entity-pairs and complete the entity-pair-level reasoning by integrating the CCA (Huang et al., 2019) into the dense connection framework (Huang et al., 2017).

5 Conclusion and Future Work

In this work, we propose a novel Dense-CCNet model by integrating the Criss-Cross Attention into the densely connected framework. Dense-CCNet model can complete entity-pairs-level logical reasoning and model the correlation between entity pairs. Experiments on three public document-level RE datasets demonstrate that our Dense-CCNet model achieved better results than the existing state-of-the-art model. In the future, we will try to use our model for other inter-sentence or document-level tasks, such as cross-sentence collective event detection.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Xiang Chen, Xin Xie, Ningyu Zhang, Jiahuan Yan, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2021. Adaprompt: Adaptive prompt-based finetuning for relation extraction. *arXiv preprint arXiv:2104.07650*.

Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Connecting the dots: Document-level neural relation extraction with edge-oriented graphs. *arXiv preprint arXiv:1909.00228*.

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

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 conference on artificial intelligence*, volume 32.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. 2017. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*.

Zhijiang Guo, Yan Zhang, and Wei Lu. 2019. Attention guided graph convolutional networks for relation extraction. *arXiv preprint arXiv:1906.07510*.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. 2019. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 603–612.

Robin Jia, Cliff Wong, and Hoifung Poon. 2019. Document-level n-ary relation extraction with multiscale representation learning. *arXiv preprint arXiv:1904.02347*.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Bo Li, Wei Ye, Zhonghao Sheng, Rui Xie, Xiangyu Xi, and Shikun Zhang. 2020. Graph enhanced dual attention network for document-level relation extraction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1551–1560.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Guoshun Nan, Zhijiang Guo, Ivan Sekulić, and Wei Lu. 2020. Reasoning with latent structure refinement for document-level relation extraction. *arXiv preprint arXiv:2005.06312*.

Chris Quirk and Hoifung Poon. 2016. Distant supervision for relation extraction beyond the sentence boundary. *arXiv preprint arXiv:1609.04873*.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. *arXiv preprint arXiv:1906.03158*.

Hengzhu Tang, Yanan Cao, Zhenyu Zhang, Jiangxia Cao, Fang Fang, Shi Wang, and Pengfei Yin. 2020. Hin: Hierarchical inference network for document-level relation extraction. *Advances in Knowledge Discovery and Data Mining*, 12084:197.

Patrick Verga, Emma Strubell, and Andrew McCallum. 2018. Simultaneously self-attending to all mentions for full-abstract biological relation extraction. *arXiv preprint arXiv:1802.10569*.

Difeng Wang, Wei Hu, Ermei Cao, and Weijian Sun. 2020a. Global-to-local neural networks for document-level relation extraction. *arXiv preprint arXiv:2009.10359*.

Hong Wang, Christfried Focke, Rob Sylvester, Nilesh Mishra, and William Wang. 2019. Fine-tune bert for docred with two-step process. *arXiv preprint arXiv:1909.11898*.

Zifeng Wang, Rui Wen, Xi Chen, Shao-Lun Huang, Ningyu Zhang, and Yefeng Zheng. 2020b. Finding influential instances for distantly supervised relation extraction. *arXiv preprint arXiv:2009.09841*.

Tongtong Wu, Xuekai Li, Yuan-Fang Li, Reza Haffari, Guilin Qi, Yujin Zhu, and Guoqiang Xu. 2021. Curriculum-meta learning for order-robust continual relation extraction. *CoRR, abs/2101.01926*.

Ye Wu, Ruibang Luo, Henry CM Leung, Hing-Fung Ting, and Tak-Wah Lam. 2019. Renet: A deep learning approach for extracting gene-disease associations from literature. In *International Conference on Research in Computational Molecular Biology*, pages 272–284. Springer.

Benfeng Xu, Quan Wang, Yajuan Lyu, Yong Zhu, and Zhendong Mao. 2021a. Entity structure within and throughout: Modeling mention dependencies for document-level relation extraction. *arXiv preprint arXiv:2102.10249*.

Wang Xu, Kehai Chen, and Tiejun Zhao. 2021b. Document-level relation extraction with reconstruction. In *The 35th AAAI Conference on Artificial Intelligence (AAAI-21)*.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. Docred: A large-scale document-level relation extraction dataset. *arXiv preprint arXiv:1906.06127*.

Deming Ye, Yankai Lin, Jiaju Du, Zhenghao Liu, Peng Li, Maosong Sun, and Zhiyuan Liu. 2020. Coreferential reasoning learning for language representation. *arXiv preprint arXiv:2004.06870*.

Haiyang Yu, Ningyu Zhang, Shumin Deng, Hongbin Ye, Wei Zhang, and Huajun Chen. 2020. Bridging text and knowledge with multi-prototype embedding for few-shot relational triple extraction. *arXiv preprint arXiv:2010.16059*.

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

Shuang Zeng, Runxin Xu, Baobao Chang, and Lei Li. 2020. Double graph based reasoning for document-level relation extraction. *arXiv preprint arXiv:2009.13752*.

Ningyu Zhang, Xiang Chen, Xin Xie, Shumin Deng, Chuanqi Tan, Mosha Chen, Fei Huang, Luo Si, and Huajun Chen. 2021. Document-level relation extraction as semantic segmentation. *arXiv preprint arXiv:2106.03618*.

Ningyu Zhang, Shumin Deng, Zhen Bi, Haiyang Yu, Jiacheng Yang, Mosha Chen, Fei Huang, Wei Zhang, and Huajun Chen. 2020a. Openue: An open toolkit of universal extraction from text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 1–8.

Ningyu Zhang, Shumin Deng, Zhanlin Sun, Jiaoyan Chen, Wei Zhang, and Huajun Chen. 2020b. Relation adversarial network for low resource knowledge graph completion. In *Proceedings of The Web Conference 2020*, pages 1–12.

Yuhao Zhang, Peng Qi, and Christopher D Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. *arXiv preprint arXiv:1809.10185*.

Zhenyu Zhang, Bowen Yu, Xiaobo Shu, Tingwen Liu, Hengzhu Tang, Wang Yubin, and Li Guo. 2020c. Document-level relation extraction with dual-tier heterogeneous graph. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1630–1641.

Hengyi Zheng, Rui Wen, Xi Chen, Yifan Yang, Yunyan Zhang, Ziheng Zhang, Ningyu Zhang, Bin Qin, Ming Xu, and Yefeng Zheng. 2021. Prgc: Potential relation and global correspondence based joint relational triple extraction. *arXiv preprint arXiv:2106.09895*.

Huiwei Zhou, Yibin Xu, Weihong Yao, Zhe Liu, Chengkun Lang, and Haibin Jiang. 2020. Global context-enhanced graph convolutional networks for document-level relation extraction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5259–5270.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14612–14620.

Dataset	DocRED	CDR	GDA
Train	3053	500	23353
Dev	1000	500	5839
Test	1000	500	1000
Relations	97	2	2
Entities per Doc	19.5	7.6	5.4
Mentions per Doc	26.2	19.2	18.5
Entities per Sent	3.58	2.48	2.28

Table 5: Summary of DocRED, CDR and GDA datasets.

Hyperparam	DocRED BERT	CDR SciBERT	GDA SciBERT
Batch size	8	16	16
Epoch	100	20	5
lr for encoder	2e-5	1e-5	1e-5
lr for other parts	1e-4	5e-5	5e-5
$\{\mu, \lambda\}$	{1, 0.5}	{1, 0.5}	{1, 0.5}
$\{\alpha, \beta, \gamma\}$	{1, 1, 1}	{1, 1, 1}	{1, 1, 1}

Table 6: Hyper-parameters Setting.

A Datasets

Table 5 details the statistics of the three document-level relational extraction datasets, DocRED, CDR, and GDA. These statis-tics further demonstrate the complexity of entity structure in document-level relation extraction tasks.

B Hyper-parameters Setting

Table 6 details our hyper-parameters setting. All of our hyperparameters were tuned on the development set.