Document-Level Relation Extraction with Adaptive Focal Loss and Knowledge Distillation

Qingyu Tan[*1,2] **Ruidan He**[†1] **Lidong Bing**[1] **Hwee Tou Ng**[2]
[1]DAMO Academy, Alibaba Group
[2]Department of Computer Science, National University of Singapore
{qingyu.tan,ruidan.he,l.bing}@alibaba-inc.com
{qtan6,nght}@comp.nus.edu.sg

Abstract

Document-level Relation Extraction (DocRE) is a more challenging task compared to its sentence-level counterpart. It aims to extract relations from multiple sentences at once. In this paper, we propose a semi-supervised framework for DocRE with three novel components. Firstly, we use an axial attention module for learning the interdependency among entity-pairs, which improves the performance on two-hop relations. Secondly, we propose an adaptive focal loss to tackle the class imbalance problem of DocRE. Lastly, we use knowledge distillation to overcome the differences between human annotated data and distantly supervised data. We conducted experiments on two DocRE datasets. Our model consistently outperforms strong baselines and its performance exceeds the previous SOTA by 1.36 F1 and 1.46 Ign_F1 score on the DocRED leaderboard.[1]

1 Introduction

The problem of document-level relation extraction[2] (DocRE) is highly important for information extraction and NLP research. The DocRE task aims to extract relations among multiple entities within a document. The DocRE task is more challenging than its sentence-level counterpart in the following aspects: (1) The complexity of DocRE increases quadratically with the number of entities. If a document contains n entities, classification decisions must be made on $n(n-1)$ entity pairs and most of them do not contain any relation. (2) Aside from the imbalance of positive and negative examples, the distribution of relation types for the positive entity pairs is also highly imbalanced. Considering the DocRED (Yao et al., 2019) dataset as an example, there are 96 relation types in total, where the top 10 relations take up 59.4% of all the relation labels. This imbalance significantly increases the difficulty of the document-level RE task.

Most existing approaches of DocRE leverage dependency information to construct a document-level graph (Zeng et al., 2021; Zeng et al., 2020), and then use graph neural networks for reasoning. Another popular strand of this field uses transformer-only (Vaswani et al., 2017) architecture (Zhou et al., 2021; Xu et al., 2021; Zhang et al., 2021). Such models are able to achieve state-of-the-art performance without explicit graph reasoning, showing that pre-trained language models (PrLMs) are able to implicitly capture long-distance relationships. However, there are three limitations of the existing DocRE methods. Firstly, existing methods mainly focus on the syntactic features from PrLMs while neglecting the interactions between entity pairs. Zhang et al. (2021) and Li et al. (2021) have used CNN structure to encode the interaction between entity pairs, but CNN structure cannot capture all the elements within the two-hop reasoning paths. Secondly, there is no prior work that explicitly tackles the class-imbalance problem for DocRE. Existing works (Zhou et al., 2021; Zhang et al., 2021; Zeng et al., 2020) only focus on threshold learning for balancing the positive and negative examples, but the class-imbalance problem within positive examples is not addressed. Lastly, there are very few works discussing the method of adapting distantly supervised data for the DocRE task. Xu et al. (2021) has shown that distantly supervised data is able to improve the performance of document-level relation extraction. However, it only uses the distantly supervised data to pre-train the RE model in a naive manner.

To overcome the limitations of existing works, we propose a semi-supervised learning framework for document-level relation extraction. Firstly, to

[*] Qingyu Tan is under the Joint PhD Program between Alibaba and National University of Singapore.
[†] Corresponding author
[1]Our code and data are available at https://github.com/tonytan48/KD-DocRE
[2]In this work, the task of relation extraction presumes that entities are given.

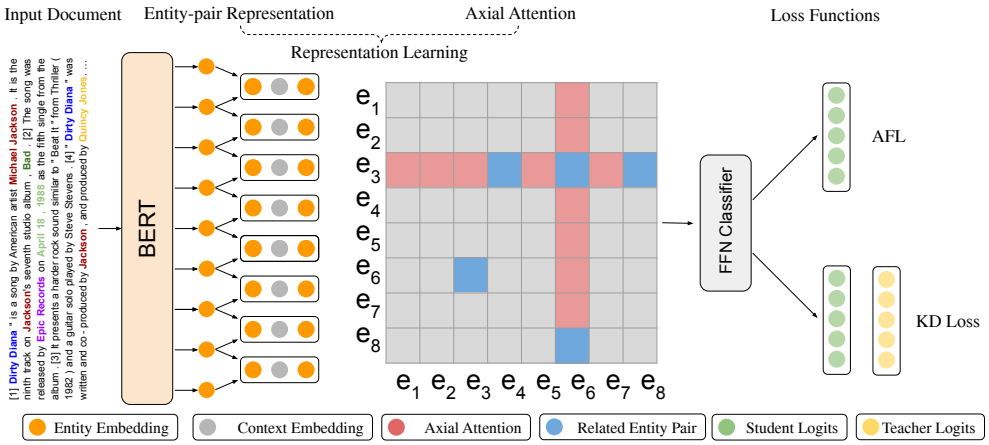

Figure 1: Model architecture of our DocRE system. We show the axial attention region for the entity pair (e_3, e_6).

improve the reasoning for two-hop relations, we propose to use an axial attention module as feature extractor. This module enables us to attend to elements that are within two-hop logical paths and capture the interdependency among the relation triplets. Secondly, we propose Adaptive Focal Loss to address the imbalanced label distribution problem. The proposed loss function encourages the long-tail classes to contribute more to the overall loss. Lastly, we use knowledge distillation to overcome the differences between the annotated data and the distantly supervised data. Specifically, we first train a teacher model with a small amount of human annotated data. The teacher model will then be used to generate predictions on a large amount of distantly supervised data. The generated predictions are used as soft labels for pre-training our student model. Finally, the pre-trained student model is further fine-tuned on the human annotated data.

We conducted experiments on two datasets – the DocRED (Yao et al., 2019) dataset and the HacRED (Cheng et al., 2021) dataset. Experimental results show that our model consistently outperforms competitive baselines. Moreover, our model significantly outperforms the existing state-of-the-art **SSAN-Adapt** (Xu et al., 2021) on the DocRED leaderboard by 1.36 in F1 score and 1.46 in Ign_F1 score.[3] Besides, we provide a thorough ablation study and error analysis to identify the bottleneck of our method.

[3]Refer to https://competitions.codalab.org/competitions/20717, where our model is named *KD-Roberta*.

2 Methodology

2.1 Problem Formulation

In this section, we describe the task formulation of document-level relation classification. Given a document D that contains a set of entities $\{e_i\}_{i=1}^{n}$, the document-level relation extraction task is to predict the relation types between entity pairs $(e_s, e_o)_{s,o\in\{1...n\},s\neq o}$, where the subscripts of e_s and e_o refer to subject and object. The set of relations is defined as $\mathbf{R} \cup \{\mathbf{NR}\}$, where **NR** stands for *no relation*. An entity may occur multiple times in a document, thus for each entity e_i, there can be multiple mentions $\{m_j^i\}_{j=1}^{N_{e_i}}$. If no relation exists between the entities in the pair (e_s, e_o), it will be labeled as **NR**. During test time, the relation labels for all entity pairs $(e_s, e_o)_{s,o\in\{1...n\},s\neq o}$ will be predicted. Essentially, this is a multi-label classification problem, as there can be multiple relations between e_s and e_o.

2.2 Model Architecture

As shown in Figure 1, our semi-supervised learning framework mainly consists of three parts: (1) representation learning; (2) adaptive focal loss; and (3) knowledge distillation for distant supervision pretraining. For representation learning, we first extract the contextual representation for each entity-pair by a pre-trained language model. The entity pair representations will be further enhanced by the axial attention module, which will encode the inter-dependent information between entity pairs. We then use a feedforward neural network (FFN) classifier to obtain the logits and compute their losses. We use our proposed adaptive focal loss to better learn from long-tail classes. Finally, we use

knowledge distillation to overcome the differences between human annotated data and distantly supervised data. Specifically, we train a teacher model with the annotated data and use its output as soft labels. We then pre-train a student model based on the soft labels and the distant labels. The pre-trained student model will be fine-tuned again with the annotated data. We will describe the details for each part in the following sections.

2.2.1 Representation Learning

Entity Representation We use a pretrained language model as the encoder. For a document D of length l, we have $D = [x_t]_{t=1}^{l}$, where x_t is the word at location t. Following prior works for relation classification, we use special token markers to represent entities. The entity mentions will be marked by a special token "*" at the start and end position. We then use a pre-trained language model (PrLM) to obtain the contextualized embeddings $\mathbf{H}$ of this document.

$$\mathbf{H} = PrLM([x_1, ..., x_l]) = [h_1, ..., h_l]) \quad (1)$$

where $\mathbf{H} \in \mathbb{R}^{l \times d}$ and d is the hidden dimension of the PrLM. If the document length exceeds the maximum position of the PrLM, the document will be encoded as multiple overlapping chunks, and the contextualized embeddings of the overlapping chunks will be averaged. We take the embedding of the special token "*" at the start of the mention as its embedding, which is denoted as h_{m_j}. Then, for each entity e_i with mentions $\{m_j^i\}_{j=1}^{N_{e_i}}$, where N_{e_i} is the number of mentions for entity e_i, its global representation is obtained by logsumexp pooling:

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

where $h_{e_i} \in \mathbb{R}^d$ is the aggregated feature of e_i.

Context-enhanced Entity Representation As prior works (Xu et al., 2021; Peng et al., 2020) have shown that contextual information is crucial for the relation classification task, our model also adapts contextual pooling method from Zhou et al. (2021). For each entity e_i, we first aggregate the attention output for its mentions by mean pooling $A_{e_i} = \sum_{j=1}^{N_{e_i}}(a_{m_j})$, where $a_{m_j} \in \mathbb{R}^{H \times l}$ is the the self-attention weight at the position of mention m_j, H is the number of attention heads, and l is the document length. Then the context query is calculated as:

$$q^{(s,o)} = \sum_{i=1}^{H}(A_{e_s}^i \circ A_{e_o}^i) \quad (3)$$

$$c^{(s,o)} = \mathbf{H}^\intercal q^{(s,o)} \quad (4)$$

where $A_{e_s} \in \mathbb{R}^{H \times l}$ is the aggregated attention output for entity e_s, likewise for e_o. $q^{(s,o)} \in \mathbb{R}^l$ is the mean-pooled attention weight for entity pair (e_s, e_o) and $\mathbf{H} \in \mathbb{R}^{l \times d}$ is the contextual embedding of the whole document. Then the context vector $c^{(s,o)} \in \mathbb{R}^d$ is fused with the entity representations.

$$z_s = tanh(\mathbf{W}_s h_{e_s} + \mathbf{W}_c c^{(s,o)}) \quad (5)$$

where $z_s \in \mathbb{R}^d$ is the context-enhanced representation of subject s for entity pair (e_s, e_o). We obtain the object representation z_o in the same manner.

Entity Pair Representation Following Zhou et al. (2021), we use a grouped bilinear function for feature combination. The entity embedding z_s will first will be split into k equal-sized groups, such that $z_s = [z_s^1, z_s^2, ..., z_s^k]$. We perform the same splitting for z_o. The value $g_i^{(s,o)}$ at each dimension of our entity pair representation is obtained by:

$$\begin{aligned} g_i^{(s,o)} &= \sum_{j=1}^{k}(z_s^{j\intercal} W_{g_i}^j z_o^j) + b_i \\ g^{(s,o)} &= [g_1^{(s,o)}, g_2^{(s,o)}, ..., g_d^{(s,o)}] \end{aligned} \quad (6)$$

where $W_{g_i}^j \in \mathbb{R}^{d/k \times d/k}$, for $i = 1, ..., d$, $j = 1, ..., k$, is the weight matrix for dimension i. b_i is a scalar bias of dimension i. $g^{(s,o)} \in \mathbb{R}^d$ is our final entity pair representation.

For a given document D with n entities, we need to classify $n(n-1)$ number of entity pair permutations. To help us encode all the entity pairs and their positions, we used an $\mathbb{R}^{n \times n \times d}$ matrix $\mathbf{G}$ to represent all the entity pairs of document D, and the diagonal of the $n \times n$ index is neglected during training and inference.

Axial Attention-Enhanced Entity Pair Representation Instead of using only head and tail embedding for relation classification, we propose to use two-hop attention to encode the axial neighboring information of each entity pair (e_s, e_o) representation. Although there are prior works that use Convolution Neural Networks (CNNs) to encode the neighbor information for relation classification (Zhang et al., 2021), we believe that attending to the axial elements is more effective and

intuitive. Given an $n \times n$ entity table, for entity pair (e_s, e_o), attending to its axial elements corresponds to attending to elements that are either (e_s, e_i) or (e_i, e_o). That is, if a two-hop relation (e_s, e_o) can be dissected into a path (e_s, e_i) and (e_i, e_o), then the most informative neighbors for classifying (e_s, e_o) are the one-hop candidates that share e_s or e_o with this entity pair. The axial attention is simply computed by self-attention along the height axis and the width axis, and each computation along the axes is followed by a residual connection. For the cell (e_s, e_o), we have:

$$\begin{aligned} r_w^{(s,o)} &= r_h^{(s,o)} + \sum_{p \in 1..n} softmax_p(q_{(s,o)}^T k_{(s,p)}) v_{(s,p)} \\ r_h^{(s,o)} &= g^{(s,o)} + \sum_{p \in 1..n} softmax_p(q_{(s,o)}^T k_{(p,o)}) v_{(p,o)} \end{aligned} \tag{7}$$

where we denote query $q_{(i,j)} = W_Q g^{(i,j)}$, key $k_{(i,j)} = W_K g^{(i,j)}$, and value $v_{(i,j)} = W_V g^{(i,j)}$, which are all linear projections of the entity pair representation g at position (i, j). $W_Q \in \mathbb{R}^{d \times d}$, $W_K \in \mathbb{R}^{d \times d}$, and $W_V \in \mathbb{R}^{d \times d}$ are all learnable weight matrices. The output of the axial attention module is $r_w^{(s,o)} \in \mathbb{R}^d$. The $softmax_p$ function denotes a softmax function that applies to all possible $p = (i, j)$ positions. The formulation of this mechanism resembles Wang et al. (2020). However, our motivation is different, as Wang et al. (2020) aim to use this mechanism to reduce the computational complexity of semantic segmentation, whereas our motivation is to attend to the one-hop neighbors for the two-hop relation triplets.

2.2.2 Adaptive Focal Loss

Finally, we have a linear layer for predicting relations:

$$l^{(s,o)} = \mathbf{W}_l r_w^{(s,o)} + b_l \tag{8}$$

where $l^{(s,o)} \in \mathbb{R}^c$ denotes the output logits for all relations, $\mathbf{W}_l \in \mathbb{R}^{d \times c}$ is the weight matrix that maps the relation embedding to the logit of each class and c is the number of classes.

Our relation extraction problem is essentially a multi-label classification problem. Traditionally, binary cross-entropy (BCE) loss is used to tackle this problem. However, this method relies on a global probability threshold for inference. Recently **Adaptive Thresholding Loss** (**ATL**, Zhou et al., 2021) has been proposed for multi-label classification. Instead of using a global probability threshold for all examples, ATL introduced a special class TH as the adaptive threshold value for each example. For each entity pair (e_s, e_o), the classes whose logits are larger than the TH class logit will be predicted as positive classes, and the rest will be predicted as negative classes.

We propose **Adaptive Focal Loss** (**AFL**) as an enhancement to **ATL** for long-tail classes. Our loss consists of two parts, the first part is for positive classes and the second part is for negative classes. During training, the label space is divided into two subsets: positive class subset $\mathcal{P}_T$ and negative class subset $\mathcal{N}_T$. The positive class subset $\mathcal{P}_T$ contains the relations that exist in entity pair (e_s, e_o), and if there is no relation between (e_s, e_o), $\mathcal{P}_T$ is empty ($\mathcal{P}_T = \varnothing$). The negative subset $\mathcal{N}_T$, on the other hand, contains the relation classes that do not belong to the positive classes, $\mathcal{N}_T = \mathcal{R} \setminus \mathcal{P}_T$. The probability of each positive class is computed as:

$$P(r_i | e_s, e_o) = \frac{exp(l_{r_i}^{(s,o)})}{exp(l_{r_i}^{(s,o)}) + exp(l_{TH}^{(s,o)})} \tag{9}$$

where the logit of r_i is ranked with the logit of threshold class TH individually. This is different from the original ATL, where all positive logits are ranked together with a softmax function. For simplicity, $P(r_i | e_s, e_o)$ is denoted as $P(r_i)$ in this section, because we are only discussing (e_s, e_o). For the negative classes, we use their logits to compute the probability of the TH class:

$$P(r_{TH} | e_s, e_o) = \frac{exp(l_{r_{TH}}^{(s,o)})}{\sum_{r_j \in \mathcal{N}_T \cup \{TH\}} exp(l_{r_j}^{(s,o)})} \tag{10}$$

Similarly, $P(r_{TH} | e_s, e_o)$ is referred to as $P(r_{TH})$ in the remainder of this section. Since the distribution of the positive labels is highly imbalanced, we leverage the idea of focal loss (Lin et al., 2017) for balancing the logits of the positive classes. We have our loss function as:

$$\mathcal{L}_{RE} = \sum_{r_i \in \mathcal{P}_T} (1 - P(r_i))^\gamma log(P(r_i)) + log(P(r_{TH})) \tag{11}$$

where γ is a hyper-parameter. Our loss is designed to focus more on the low-confidence classes. If $P(r_i)$ is low, the loss contribution from the relevant class will be higher, which enables a better optimization for long-tail classes.

2.2.3 Knowledge Distillation for Distant Supervison

In this section, we describe how we utilize the distantly supervised data in a more effective manner. The distantly supervised data included in the DocRed dataset (Yao et al., 2019) was obtained by

performing entity linking on the Wikidata Knowledge Base (Vrandečić and Krötzsch, 2014) and the Wikipedia data dump. It is shown that pre-training from the distantly supervised data is beneficial for document-level relation extraction (Xu et al., 2021). However, prior work only adapts the distantly supervised data in a naive manner. The key challenge for the distant supervision adaptation is to overcome the differences between probability distributions of the distantly supervised data and the human annotated data. We compare two strategies for adapting the distantly supervised data.

Naive Adaptation Adopting from (Xu et al., 2021), this method first pretrains the model with the distantly supervised data with the relation extraction loss $\mathcal{L}_{RE}$ (Eqn. 11), and then the model is fine-tuned on the human-annotated data with the same objective. We denote this method as **Naive Adaptation (NA)**.

Knowledge Distillation To further utilize the annotated data, we use a relation classification model trained on the human-annotated data (#Train in Table 1) as the teacher model. The teacher model is used to generate soft labels on the distantly supervised data. Specifically, the distantly supervised data is fed into the teacher model and the predicted logits will be the soft labels used for training the student model. The student model has the same configuration as the teacher model, but is trained with two signals simultaneously. The first signal is the supervision from the hard labels of the distantly supervised data and the second is from the predicted soft labels. We denote the loss computed on the hard labels as $\mathcal{L}_{RE}$ and the knowledge distillation loss computed on the soft labels as $\mathcal{L}_{KD}$. We use mean squared error (MSE) as the knowledge distillation loss function:

$$\mathcal{L}_{KD} = MSE(l_S^{(s,o)}, l_T^{(s,o)}) \quad (12)$$

where $l_S^{(s,o)}$ denotes the predicted logits of the student model and $l_T^{(s,o)}$ is the prediction of the teacher model. The student model is further fine-tuned with human-annotated data (#Train in Table 1) after it has been pre-trained on the distantly supervised data. The overall loss of pre-training with distantly supervised data is computed as:

$$\mathcal{L} = \mathcal{L}_{KD} + \mathcal{L}_{RE} \quad (13)$$

We denote this method as **KD** in our main experimental results section. Besides the MSE loss, we also compare different adaptation methods, such as KL-Divergence, in section 3.6.

3 Experiments

Statistics	DocRED	HacRED
# distant docs	101,873	–
# training docs	3,053	6,231
# dev docs	1,000	1,500
# test docs	1,000	1,500
# relations	97	27
Avg # entities per doc	19.5	10.8
Avg # mentions per entity	1.4	1.2
Avg # relations per doc	12.5	7.4

Table 1: Dataset statistics of the DocRED and HacRED datasets.

3.1 Dataset Statistics

We evaluated our model on two document-level relation extraction datasets – the DocRED (Yao et al., 2019) benchmark and the HacRED dataset (Cheng et al., 2021). DocRED is a crowd-sourced large-scale document-level relation extraction dataset. It contains 3,053/1,000/1,000 instances for training, validation, and test, respectively. HacRED is a Chinese relation extraction dataset that focuses on the hard cases of relation extraction. It contains 27 hard relations and is split into 6,231/1,500/1,500 instances for training, validation, and test. However, the test set of HacRED is not released yet. In this paper, we only provide the results on its dev set.

3.2 Implementation Details

We implemented our model with the PyTorch version of the Huggingface Transformers (Wolf et al., 2020). For experiments on DocRED, we experimented with Roberta-large (Liu et al., 2019) and Bert-base (Devlin et al., 2019) as our document encoder respectively. For experiments on HacRED, we use XLM-R base (Conneau et al., 2020) as the document encoder. AdamW (Loshchilov and Hutter, 2019) is used as the optimizer. At the knowledge distillation stage, we trained the model with the learning rate set to 1e-5 for 2 epochs. Warmup is applied on the initial 6% steps. The dropout rates between transformer layers are set to 0.1 and the maximum gradient norm is clipped at 1.0. During the fine-tuning stage, the learning rate is set to 1e-6 and we train the model for 10 epochs. We performed grid search for $\gamma \in [0, 0.5, 1.0, 1.5, 2.0]$ and set it to 0.5. Our model is trained on a single

	Dev		Test	
w/o Distant Supervision	Ign_F1	F1	Ign_F1	F1
Two-stage-B-b	56.67	58.83	56.47	58.69
ATLOP-B-b	59.22±0.15	61.09±0.16	59.31	61.30
SIRE-B-b	59.82	61.60	**60.18**	62.05
DocuNet-B-b	59.86±0.13	61.83±0.19	59.93	61.86
Ours-B-b	**60.08**±0.11	**62.03**±0.18	60.04	**62.08**
Coref-Rb-l	57.35	59.43	57.9	60.25
SSAN-Rb-l	59.40	61.42	60.25	62.08
GAIN-B-l	60.87	63.09	60.31	62.76
ATLOP-Rb-l	61.32±0.14	63.18±0.19	61.39	63.40
DocuNet-Rb-l	**62.23**±0.12	64.12±0.14	62.39	**64.55**
DocuNet-Rb-l*	61.56±0.14	63.58±0.17	61.79	63.73
Ours-Rb-l	62.16±0.10	**64.19**±0.16	**62.57**	64.28
with Distant Supervision	Ign_F1	F1	Ign_F1	F1
ATLOP-NA-Rb-l*	63.41±0.15	65.33±0.18	63.54	65.47
DocuNet-NA-Rb-l*	63.26±0.17	65.21±0.19	63.29	65.44
SSAN-NA-Rb-l	63.76	65.69	63.78	65.92
Ours-NA-B-b	62.18±0.12	64.17±0.16	61.77	64.12
Ours-KD-B-b	62.62±0.16	64.81±0.13	62.56	64.76
Ours-NA-Rb-l	63.38±0.11	65.64±0.17	63.63	65.71
Ours-KD-Rb-l	**65.27**±0.09	**67.12**±0.14	**65.24**	**67.28**

Table 2: Experimental results for the DocRED dataset. The reported metrics are F1 score and Ign_F1. We report the average of five random runs for the development set and the best checkpoint is used for the leaderboard submission for the test results. Results with * are obtained by our reproduction.

NVIDIA V100 GPU with 32 GB memory. The main evaluation metrics are Ign_F1 and F1 score following Yao et al. (2019), where Ign_F1 refers to the F1 score that ignores the triples that appear in the annotated training data.

3.3 Compared Methods

We denote Bert-base and Bert-large encoders as **B-b** and **B-l**. The Roberta-large model is denoted as **Rb-l**. We compare our model with the state-of-the-art systems on the DocRED leaderboard as well as strong baselines by our own implementation. They are the following models: Wang et al. (2019) has proposed to fine-tune BERT for document-level RE with a two-step process (**Two-stage-B-b**). The Bert model needs to classify whether the two entities have relation and then classify their relation if the first step is positive. The **Coref-Rb-l** (Ye et al., 2020) uses a co-reference module to aggregate the mention representations of the same entity. The **SSAN** (Xu et al., 2021) model utilizes co-occurrence information between entity mentions, leverages distantly supervised data for pretraining, and achieves the state of the art on the DocRED leaderboard. Since their best model **SSAN-Adapt** is equivalent to naive adaptation in our work, we denote it as **SSAN-NA-Rb-l** in our experiments. The **GAIN** (Zeng et al., 2020) model adds a graph neural network on top of a pre-trained language model, constructs a document-level graph for each example, and uses the graphical structure to extract relations. **SIRE** (Zeng et al., 2021) uses two encoders for different types of relation — a sentence-level encoder to extract intra-sentence relations and a document encoder to extract inter-sentence relations. **ATLOP** (Zhou et al., 2021) is purely based on the transformer architecture and a novel adaptive thresholding loss to deal with the multi-label problem for DocRE. Besides, it also fuses the contextual information with the aggregated attention weights for each entity. The **DocuNet** (Zhang et al., 2021) model treats the relation extraction task in a similar way as semantic segmentation in computer vision. We also conducted an experiment that pretrained the **ATLOP-Rb-l** model with distantly supervised data, as this model is the best model by our reproduction.

3.4 Main Results

Our main results for the DocRED dataset are shown in Table 2. Knowledge distillation is able to significantly improve the performance of our model. **Ours-KD-Rb-l** achieves the best single-model performance of 67.28 test F1. Our best model significantly ourperforms the previous state of the art **SSAN-NA-Rb-l** by 1.36 on test F1 and 1.46 on test Ign_F1. As of 11th Nov 2021, our best model achieves the highest scores on the DocRED leaderboard.

	P	R	F1
GAIN*	73.38	**80.07**	76.09
ATLOP*	76.97	78.29	77.63
Ours	**78.53**	78.96	**78.75**

Table 3: Experimental results on HacRED dev set. Results with * are implemented by us. All experiments used XLM-R-base as the encoder.

The experiment results for the HacRED dataset are shown in Table 3. The main difference of our method with the ATLOP baseline is the Adaptive Focal Loss and the Axial Attention Module. Our proposed method is able to exceed the ATLOP baseline by 1.12 F1. Besides the performance of the models, it is worth noting that for each method, the absolute performance of HacRED is significantly higher than its performance on DocRED. This is counter-intuitive as HacRED focuses on the hard relations whereas DocRED is more general. This can be caused by the following: 1) The human annotated training instances of the HacRED dataset are significantly more than DocRED, leading to better generalization performance. 2) Even though HacRED claims it focuses on the hard cases for relation extraction, it only has 27 classes, and the relation type distribution within the HacRED dataset is more balanced.

3.5 Ablation Study

We first separate our label space into two subsets. The first subset consists of the 10 most frequent labels, accounting for 59.4% of the positive relations in the training data. The second subset is denoted as the long-tail labels, which includes the rest of the 86 relations (the total label space is 97 and there is one TH class). Since our Adaptive Focal loss function is mainly designed for improving the performance on the less frequent classes, we show the ablation study by frequent and long-tail classes in Table 4. When we change the AFL loss to conventional Adaptive Thresholding Loss (Zhou et al., 2021), the overall performance with KD drops by 0.89 F1, and the F1 score for the frequent labels only drops by 0.65. Meanwhile, the long-tail labels' F1 drops by 1.78, which is significantly higher than the drop in overall performance and frequent performance. This indicates that our Adaptive Focal Loss is able to balance the weight of the frequent classes and infrequent classes. The axial attention module is also more beneficial for the long-tail classes than the frequent classes, which shows that our model's performance on the frequent classes is saturated.

	Frequent F1	Long-tail F1	Overall F1
ATLOP-Rb-l	70.93	50.01	63.12
Ours-Rb-l	**71.26**	**51.97**	**64.19**
w/o **Axial**	70.86	50.77	63.56
w/o **AFL** *w* **ATL**	70.94	50.86	63.67
With Distant Supervision			
ATLOP-NA-Rb-l	73.26	52.39	65.33
Ours-KD-Rb-l	**74.15**	**56.51**	**67.12**
w/o **Axial**	73.52	54.96	66.36
w/o **AFL** *w* **ATL**	73.50	54.73	66.23

Table 4: Experiment results for frequent and long-tail type relations. Frequent types refer to the most popular 10 relation types, and long-tail relations refer to the rest of the 86 relations.

	P	R	Infer-F1
GAIN-B-b	38.71	59.45	46.89
Ours-Rb-l	**42.15**	**61.56**	**50.04**
w/o **Axial**	40.26	60.60	48.37

Table 5: Ablation study for the Infer-F1 relation triples on the development set of DocRED.

We also provide an ablation study on the multi-hop relations in Table 5. We use the same evaluation method for multi-hop relations as Zeng et al. (2020). This evaluation method ignores all the one-hop relation triples. Our axial attention module effectively improves Infer-F1 by 1.67, while its improvement for overall performance is only 0.63.

3.6 Comparison of Adaptation Methods

In this section, we directly compare the knowledge adaptation methods on the development set of DocRED (Table 6). We mainly compare three methods for adaptation: 1) Naive Adaptation (**NA**), 2) $\mathbf{KD}_{KL}$ knowledge distillation with the KL divergence loss and 3) $\mathbf{KD}_{MSE}$ with mean squared error loss. The adaptation performance on the development set is positively correlated with the per-

Distant Adaptation	Ign_F1	F1
NA	52.29	54.67
KD$_{KL}$	53.89	56.97
KD$_{MSE}$	**55.28**	**57.74**
Continue-trained	Ign_F1	F1
NA	63.38	65.64
KD$_{KL}$	64.42	66.24
KD$_{MSE}$	**65.27**	**67.12**

Table 6: Development set performance of different knowledge adaptation methods for DocRED.

formance of downstream fine-tuning. In the distant adaptation setting, our best method **KD**$_{MSE}$ is able to outperform **NA** by 3.07 F1 and **KD**$_{KL}$ by 0.77 F1. Similar performance differences are observed in the continue-trained setting.

4 Error Analysis

Even though our final model significantly outperforms the previous state of the art on the DocRED leaderboard, the absolute performance of our model still does not match human performance. In this section, we provide a detailed error analysis of our model on the development set of DocRED.

		Ground Truth	
Predictions		$r \in$ **R**	**NR**
	$r \in$ **R**	**C**: 8,273 (51.4%)	**MR**: 3,814 (23.7%)
		W: 242 (1.5%)	
	NR	**MS**: 3,761 (23.4%)	380,703

Table 7: Statistics of our error distribution. The final evaluation score is evaluated on $r \in$ **R** triples, hence the correct predictions of **NR** are ignored when calculating the final scores.

We first construct the union of our model's predictions and the ground truth triples (without **NR** label). Then, we categorize the union into four categories: (1) **Correct (C)**, where prediction triples are in the ground truth. (2) **Wrong (W)**, where the predicted head entity and tail entity are in the ground truth but the predicted relation is wrong. (3) **Missed (MS)**, where the model predicts no relation for a pair of head entity and tail entity with some relation in the ground truth. (4) **More (MR)**, where the model predicts an extraneous relation for a pair of head entity and tail entity not related in the ground truth. From Table 7, we observe that the error percentage of the **W** category is very small. This indicates that for a pair of head entity and tail entity with some relation in the ground truth, and when our model predicts that there is a relation between these two entities, it is able to predict the correct relation rather accurately. However, we observe that most of our errors are under the **MR** and **MS** categories, and their counts are about the same. To better understand the performance bottleneck of the document-level RE task, we evaluate our model on a simplified subtask (Table 8). This subtask is binary classification, i.e., to determine whether two entities are related or not, and it is denoted as **Binary Labels**. In this subtask, we only care about predicting correctly that there is some relation between a head entity and a tail entity, but not what the exact relation is among the 97 relation classes. The performance on this simplified task is 68.64 F1 score, which is only marginally higher than the original F1 score of 67.12. This may be due to incomplete annotation of the two document-level relation extraction datasets, and we will illustrate this hypothesis in Figure 2.

	P	R	F1
Binary Labels	68.51	68.78	68.64
Original Labels	67.10	67.13	67.12

Table 8: Performance breakdown on the DocRED dev set.

"**Eivind Bolle** (**13 October 1923 – 10 June 2012**) was a **Norwegian** politician for the **Labour Party**. He was born in **Hol**. He was elected to the **Norwegian Parliament** from **Nordland** in 1973. ... On the local level he was a member of **Hol** municipality council from **1959** to **1963** , and later in Hol 's successor municipality **Vestvågøy**. He served as mayor from **1971** to **1973** , during which term he was also a member of **Nordland** county council ..."

More: (**Nordland**, *country*, **Norwegian**), (**Vestvågøy**, *country*, **Norwegian**),...

Correct: (**Labour Party**, *country*, **Norwegian**), (**Hol**, *country*, **Norwegian**),...

Figure 2: Example output of our model on the DocRED dev set.

In Figure 2, we show an example document from the dev set of DocRED and its predictions. We observe that many triples in the **MR** category are factually correct. That is, some of the pairs of entities are truly related but are labeled as **NR** throughout the dataset. For instance, from the ground truth, we can see that **Labour Party** and **Hol** are all entities from *country* **Norway**. Similarly, **Nordland** and **Vestvågøy** are all in **Norway**, but their relations with **Norway** are not present in the ground truth triples. Therefore, when our model predicts these triples, its performance would be unfairly penal-

ized during evaluation. This observation indicates that there are some incomplete annotations in the DocRED dataset. However, this is not the focus of this paper and we would like to leave this as future work.

5 Related Work

Early works on relation extraction mainly focused on sentence-level RE (Zhang et al., 2017; Baldini Soares et al., 2019; Peng et al., 2020). However, prior works have shown that a large number of relations can only be extracted from multiple sentences (Verga et al., 2018; Yao et al., 2019; Cheng et al., 2021). Various methods have been proposed to tackle document-level relation extraction (DocRE). Graph neural networks (GNNs; Scarselli et al., 2008) have been widely used for the DocRE task. Quirk and Poon (2017) used words as nodes and dependency information as edges to construct document-level graphs. This graph will be used to extract features for each entity pair. Later works extended this idea by applying different GNN architectures (Peng et al., 2017; Verga et al., 2018; Christopoulou et al., 2019; Nan et al., 2020; Zhang et al., 2018; Zeng et al., 2020). In particular, Nan et al. (2020) proposed the latent stucture refinement (LSR) model, which used structured attention to induce the document-level graph. Zeng et al. (2020) constructed the document-level graph by entity-mention nodes and sentence edges. Besides the graph-based methods, transformer-only architectures have also proven to be highly effective for the DocRE task (Tang et al., 2020; Zhou et al., 2021). Specifically, Zhou et al. (2021) proposed adaptive thresholding loss to tackle the multi-label classification problem in DocRE.

On the other hand, learning from distant supervision is another important problem for relation extraction. Qin et al. (2018) used generative adversarial training for selecting informative examples and Feng et al. (2018) used reinforcement learning to achieve the same goal. However, there are no existing works that jointly learn from annotated data and distant data. To this end, this paper is the first to overcome the differences between the human annotated and distantly supervised data. Moreover, this paper also tackles the under-explored class imbalance problem and the two-hop logical reasoning problem with novel solutions to the shortcomings of existing approaches.

6 Conclusions

In this paper, we have proposed a novel framework for document-level relation extraction, based on knowledge distillation, axial attention, and adaptive focal loss. Our proposed method is able to significantly outperform the previous state of the art on the DocRED leaderboard. Besides, we also conducted a thorough ablation study and error analysis to identify the bottleneck of the document-level relation extraction task.

7 Acknowledgements

We would like to thank the anonymous reviewers for their insightful feedback and comments.

References

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of ACL*.

Qiao Cheng, Juntao Liu, Xiaoye Qu, Jin Zhao, Jiaqing Liang, Zhefeng Wang, Baoxing Huai, Nicholas Jing Yuan, and Yanghua Xiao. 2021. HacRED: A large-scale relation extraction dataset toward hard cases in practical applications. In *Findings of ACL*.

Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Connecting the dots: Document-level neural relation extraction with edge-oriented graphs. In *Proceedings of EMNLP*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*.

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

Jingye Li, Kang Xu, Fei Li, Hao Fei, Yafeng Ren, and Donghong Ji. 2021. Mrn: A locally and globally mention-based reasoning network for document-level relation extraction. In *Findings of ACL*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of ICCV*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of ICLR*.

Guoshun Nan, Zhijiang Guo, Ivan Sekulic, and Wei Lu. 2020. Reasoning with latent structure refinement for document-level relation extraction. In *Proceedings of ACL*.

Hao Peng, Tianyu Gao, Xu Han, Yankai Lin, Peng Li, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2020. Learning from context or names? An empirical study on neural relation extraction. In *Proceedings of EMNLP*.

Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. 2017. Cross-sentence n-ary relation extraction with graph lstms. *Transactions of the Association for Computational Linguistics*.

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

Chris Quirk and Hoifung Poon. 2017. Distant supervision for relation extraction beyond the sentence boundary. In *Proceedings of EACL*.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. *IEEE Transactions on Neural Networks*.

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. In *Proceedings of KDD*.

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 NIPS*.

Patrick Verga, Emma Strubell, and Andrew McCallum. 2018. Simultaneously self-attending to all mentions for full-abstract biological relation extraction. In *Proceedings of NAACL*.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*.

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*.

Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. 2020. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In *Proceedings of ECCV*.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of EMNLP: System Demonstrations*.

Benfeng Xu, Quan Wang, Yajuan Lyu, Yong Zhu, and Zhendong Mao. 2021. Entity structure within and throughout: Modeling mention dependencies for document-level relation extraction. In *Proceedings of AAAI*.

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. In *Proceedings of ACL*.

Deming Ye, Yankai Lin, Jiaju Du, Zhenghao Liu, Peng Li, Maosong Sun, and Zhiyuan Liu. 2020. Coreferential reasoning learning for language representation. In *Proceedings of EMNLP*.

Shuang Zeng, Yuting Wu, and Baobao Chang. 2021. SIRE: Separate intra- and inter-sentential reasoning for document-level relation extraction. In *Findings of ACL*.

Shuang Zeng, Runxin Xu, Baobao Chang, and Lei Li. 2020. Double graph based reasoning for document-level relation extraction. In *Proceedings of EMNLP*.

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. In *Proceedings of IJCAI*.

Yuhao Zhang, Peng Qi, and Christopher D. Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of EMNLP*.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of EMNLP*.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Proceedings of AAAI*.