Improving Long Tailed Document-Level Relation Extraction via Easy Relation Augmentation and Contrastive Learning

Yangkai Du[1], Tengfei Ma[2], Lingfei Wu[3], Yiming Wu[1], Xuhong Zhang[1]
Bo Long[3], Shouling Ji[1]
[1]Zhejiang University; [2]IBM Research; [3]JD.COM
{yangkaidu,wuyiming,zhangxuhong,sji}@zju.edu.cn
tengfei.ma1@ibm.com, {lingfei.wu,bo.long}@jd.com

Abstract

Towards real-world information extraction scenario, research of relation extraction is advancing to document-level relation extraction(DocRE). Existing approaches for DocRE aim to extract relation by encoding various information sources in the long context by novel model architectures. However, the inherent long-tailed distribution problem of DocRE is overlooked by prior work. We argue that mitigating the long-tailed distribution problem is crucial for DocRE in the real-world scenario. Motivated by the long-tailed distribution problem, we propose an Easy Relation Augmentation(ERA) method for improving DocRE by enhancing the performance of tailed relations. In addition, we further propose a novel contrastive learning framework based on our ERA, i.e., ERACL, which can further improve the model performance on tailed relations and achieve competitive overall DocRE performance compared to the state-of-arts.

1 Introduction

Relation extraction plays an essential role in information extraction, which aims to predict relations of entities in texts. Early work on relation extraction mainly focuses on sentence-level relation extraction, i.e., predicting relation from a single sentence, and has achieved promising results. Recently, the research of relation extraction has advanced to document-level relation extraction, a scenario more practical than sentence-level relation extraction and more challenging.

The relation pattern between entity pairs across different sentences is often more complex, and the distance of these entity pairs is relatively long. Therefore, DocRE requires models to figure out the relevant context and conduct reasoning across sentences instead of memorizing the simple relation pattern in a single sentence. Moreover, multiple entity pairs co-exist in one document, and each entity may have more than one mention appearing across sentences. Thus, DocRE also requires the model to extract relations of multiple entity pairs from a single document at once. In other words, DocRE is a one-example-multi-instances task while sentence-level RE is a one-example-one-instance task.

Another unique challenge of DocRE that cannot be overlooked is long-tailed distribution. Long-tailed distribution is a common phenomenon in real-world data. In DocRE, we also observe the long-tailed distribution. Figure 1 presents the relation distribution of DocRED(Yao et al., 2019), a widely-used DocRE dataset: 7 most frequent relations from 96 relations takes up 55.12% of total relation triples; while the frequencies of 60 relations are only less than 200. Vanilla training on long-tailed data will cause the model to achieve overwhelming performance on head relations but underfitting on tailed relations. Although the overall DocRE performance is largely dependent on performance on head relations since they are the majority, model failure on tailed relations is a big concern in real-world DocRE scenarios.

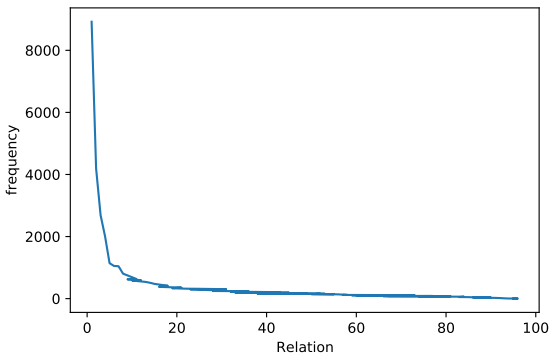

Figure 1: Relation Distribution of DocRED Train set. Relation index are sorted by frequency count from high to low.

Data augmentation is a commonly used strategy for addressing the long-tailed problem. Nonetheless, applying data augmentation efficiently on

DocRE is non-trivial. Ordinary data augmentation operation on the document, including text random-dropping or replacing(Wei and Zou, 2019) would require the DocRE model for extra encoding process of the entire document, which is computation in-efficient on DocRE since the document may contain numerous sentences. Besides, DocRE is a one-example-multi-instances task, so tailed relations and head relations presumably co-exist in one document. As a result, the head relations would also be augmented if we augment the tailed relations by aforementioned trivial augmentation methods on text, which is unexpected and may lead to over-fitting on head relations.

In this paper, we propose a novel data augmentation mechanism for DocRE, named ERA, for improving the document-level relation extraction by mitigating the long-tailed problem. The proposed ERA method applies augmentation on relation representations rather than texts, so it can augment tail relations without another encoding operation of the long document, which is computation-efficient and also effective for improving performance on tailed relations.

In addition, we propose a contrastive learning framework based on our ERA method, i.e., ERACL, for pre-training on the distantly-supervised data. The proposed ERACL framework can further enhance the model performance on tailed relations and achieve comparable overall DocRE performance compared to the state-of-art methods on DocRED.

2 Background and Related Works

2.1 Problem Formulation

Given a document $\mathcal{D} = \{w_1, w_2, ..., w_l\}$ with l words, a set of n entities $\mathcal{E} = \{e_i\}_{i=1}^{n}$ are identified by human annotation or external tools. For each entity e_i, m mentions of e_i denoted as $\{m_{ij}\}_{j=1}^{m}$ are also annotated by providing the start position and end position in $\mathcal{D}$. In addition, the relation scheme $\mathcal{R}$ is also defined.

The objective of DocRE is to extract the relation triple set $\{(e_h, r, e_t)|e_h \in \mathcal{E}, r \in \mathcal{R}, e_t \in \mathcal{E}\} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$ from all possible relation triples, where each realtion triple (e_h, r, e_t) extracted by the model can be interpreted as relation $r \in \mathcal{R}$ holds between head entity $e_h \in \mathcal{E}$ and tail entity $e_t \in \mathcal{E}$. For future simplicity, we denote tail relations as $\mathcal{R}^t \subset \mathcal{R}$ and head relations as $\mathcal{R}^h \subset \mathcal{R}$.

2.2 Document-Level Relation Extraction

To address the prior challenges in DocRE, one main branch of DocRE works use **Graph-based Methods**(Sahu et al., 2019; Christopoulou et al., 2019; Wang et al., 2020; Zeng et al., 2020; Nan et al., 2020; Li et al., 2020; Xu et al., 2021b). The general idea of graph-based methods is to conduct multi-hop reasoning across entities, mentions and sentences in a document by graph neural networks. First a document is converted to a document graph by human designed heuristics, attention mechanism or dependency parser. Then the document graph is encoded by graph neural networks(Kipf and Welling, 2017; Chang et al., 2018; Wu et al., 2021) to conduct multi-hop reasoning across graph nodes and edges. Another branch of DocRE methods adopt **Transformer-based Methods**(Wang et al., 2019; Ye et al., 2020; Xu et al., 2021a; Zhou et al., 2021; Zhang et al., 2021). Transformer-based methods rely on the strong long-context representation capability of pre-trained transformers(Devlin et al., 2019; Liu et al., 2019). In addition, self-attention mechanism in transformer architecture can implicitly model the dependency between entities, mentions and contexts, which can be utilized for relation reasoning(Zhou et al., 2021).

Different from previous works, in this paper we focus more on addressing the challenges of long-tailed distribution in DocRE.

2.3 Contrastive Learning

Contrastive learning is proved to be a promising self-supervised pre-training task for image recognition(Chen et al., 2020; He et al., 2020). The principle of contrastive learning is to increase the representation similarity of anchor example x and positive examples x^+ while decreasing the representation similarity of anchor example x and negative examples x^- by INFONCE loss(van den Oord et al., 2018).

Under the self-supervised setting, positive samples x^+ are constructed by data augmentation operation, including image cropping, resizing on anchor samples. The motivation of creating x^+ via data augmentation is that augmented samples are still similar or even the same in semantic space, then it can provide training signals for self-supervised pre-training. Therefore, models pre-trained by self-supervised contrastive learning can learn task-agnostic and robust representation for down-streaming tasks, which also can capture the

semantic information of input samples.

The general contrastive learning framework has been applied in language tasks and achieved competitive performance. Fang et al. (2020) adapted the contrastive learning framework for self-supervised pre-training on transformers and achieved superior performance compared to BERT(Devlin et al., 2019). Gunel et al. (2021) proposed to use supervised contrastive learning for more robust fine-tuning on pre-trained transformers. In relation extraction, Peng et al. (2020) and Qin et al. (2021) adopted contrastive learning as one pre-training task to improve the relation understanding capability of BERT. It has been demonstrated that proper adaptation of contrastive learning framework can encourage the model to learn more robust task-agnostic or task-related representations, especially when the training data is limited. The problem of long-tail relation distribution is essentially lack of training samples of certain relation types. Considering this, we proposed a new contrastive learning framework based on a new data augmentation method in DocRE, called Easy Relation Augmentation(ERA), which can learn more robust relation representations for DocRE, especially for the tailed relations.

3 Easy Relation Augmentation

3.1 Overview

We summarize the main components of ERA framework as follow: ERA takes a document $\mathcal{D}$ as input, then the Document Encoding and Relation Encoding modules will encode each entity pair $(e_h, e_t) \in \mathcal{E} \times \mathcal{E}$ from two aspects: contextualized entity representation and pooled context representation via self-attention mechanism of Pre-trained Transformers(Zhou et al., 2021). Afterwards, we proposed a novel Easy Relation Augmentation(ERA) mechanism to enhance the entity pair representation by applying a random mask on pooled context representation. The proposed ERA mechanism can augment the tail relations $r \in \mathcal{R}^t$ without another Relation Encoding and Document Encoding, which is computation-efficient and also effective. Finally, we train the relation prediction module on the augmented relation representations.

3.2 Document Encoding

In light of the promising capability of Pre-trained Transformers(Devlin et al., 2019; Liu et al., 2019) for modeling the long-range text dependency, We resort to pre-trained transformers for document encoding. We add a special entity marker "*"(Zhang et al., 2017) at the start and end position of each mention m_{ij}, and "*" can be replaced with other special tokens. Entity markers can spotlight the mention words and also provide entity positional information for Pre-trained Transformers, which proves to be effective in DocRE(Zhou et al., 2021). Feeding the document $\mathcal{D}$ to the pre-trained transformers, we can get the contextualized representation $\mathbf{H}$ of all words and vanilla multi-head self-attention $\mathbf{A}$ from the last block of Pre-trained Transformers($\mathbf{Ptr}$).

$$\mathbf{H}, \mathbf{A} = \mathbf{Ptr}(\mathcal{D} = \{w_1, w_2, ..., w_l\}) \quad (1)$$

Where $\mathbf{H} \in \mathbb{R}^{l \times d}$, $\mathbf{A} \in \mathbb{R}^{l \times l \times h}$. d is the model dimension of the Pre-trained Transformers and h is the number of self-attention heads of Pre-trained transformers.

3.3 Relation Encoding

Given the contextualized representation $\mathbf{H}$ and self-attention $\mathbf{A}$ of the document $\mathcal{D}$, the goal of Relation Encoding module is to encode each entity pair $(e_h, e_t) \in \mathcal{E} \times \mathcal{E}$ by aggregating the contextualized entity representation and pooled context representation, which are crucial for relation understanding and reasoning across the long document.

Contextualized entity representation can provide the contextualized entity naming and entity typing information for relation inference. For entity $e_h \in \mathcal{E}$, we obtain the contextualized mention representation by collecting the Pre-trained transformer last layer output of "*" marker at the start of mention m_{ij}, denoted as $\mathbf{m_{hj}}$. Subsequently, we can get the final contextualized entity representation $\mathbf{e_h}$ by logsumexp pooling (Jia et al., 2019), which can achieve better results compared to max pooling and average pooling on DocRE(Zhou et al., 2021).

$$\mathbf{e_h} = log \sum_{j=1}^{m} exp(\mathbf{m_{hj}}) \quad (2)$$

As mentioned in Section 2.2, DocRE requires the model to capture the dependencies among entities, mentions, and context words, and also filter out the unnecessary context information from the long document. We named the aforementioned information as pooled context information. The self-attention matrix $\mathbf{A} \in \mathbb{R}^{l \times l \times h}$ obtained from Pre-trained transformers have already implicitly modeled the

dependency among entities, mentions, and context words, which can be utilized for getting meaningful pooled context representation(Zhou et al., 2021). We follow Zhou et al. (2021) to obtain the pooled context information by utilizing the self-attention matrix $\mathbf{A}$.

Given a entity pair $(e_h, e_t) \in \mathcal{E} \times \mathcal{E}$, one can get the pooled context representation $\mathbf{c_{h,t}}$ by Equation 3 and 4.

$$\mathbf{c_{h,t}} = \mathbf{H^T} \cdot \frac{\mathbf{A_{h,t}}}{\mathbf{1^T} \cdot \mathbf{A_{h,t}}} \tag{3}$$

$$\mathbf{A_{h,t}} = \mathbf{A_h} * \mathbf{A_t} \tag{4}$$

Where $\mathbf{A_h} \in \mathbb{R}^{l \times 1}$,$\mathbf{A_t} \in \mathbb{R}^{l \times 1}$ and $\mathbf{1} \in \mathbb{R}^{l \times l}$. $\mathbf{A_h}$ is the attention score of entity e_h to all words in $\mathcal{D}$, which is obtained by averaging the attention score of all entity mentions m_{hj}, denoted as $\mathbf{A_{m_{hj}}}$. Similar to contextualized mention representation $\mathbf{m_{hj}}$, we obtain the mention attention score $\mathbf{A_{m_{hj}}}$ by indexing the vanila self-attention matrix $\mathbf{A}$ with position of starting "*" marker. In addition, note that the vanila self-attention matrix is first averaged over all attention heads before performing the indexing. $\mathbf{A_t}$ is also calculated following the same procedure.

In the end, for the entity pair (e_h, e_t), we can form a triple represention $\mathcal{T}_{\mathbf{h,t}} = (\mathbf{e_h}, \mathbf{c_{h,t}}, \mathbf{e_t})$. $\mathcal{T}_{\mathbf{h,t}}$ contains all the information for relation prediction and form the basis for our Easy Relation Augmentation and Contrastive Leaning framework.

3.4 Relation Representation Augmentation

To address the long-tailed problem residing in the DocRE, we propose a novel Easy Relation Augmentation(ERA) mechanism to increase the frequency of tailed relations and enhance the entity pair representation.

Denote the set of triple representation of all entity pairs as $\mathcal{T}_{orig} = \{(\mathbf{e_h}, \mathbf{c_{h,t}}, \mathbf{e_t}) | e_h \in \mathcal{E}, e_t \in \mathcal{E}\}$. In addition, we can manually select the set of relations need to be augmented, i.e., $\mathcal{R}^{aug} \subseteq \mathcal{R}$.

Given a entity pair (e_h, e_t) whose relation $r \in \mathcal{R}^{aug}$, we first retrieve the original triple representation $(\mathbf{e_h}, \mathbf{c_{h,t}}, \mathbf{e_t})$ from $\mathcal{T}_{orig}$. Recall that the pooled context representation $\mathbf{c_{h,t}}$ encodes the unique context information for relation inference, and a slight perturbation on the context should not affect the relation prediction. Established on this intuition, we add a small perturbation on $\mathbf{c_{h,t}}$.

We first apply a random mask on $\mathbf{A}_{h,t}$ described in Equation 3 by multiplying $\mathbf{A}_{h,t}$ with a randomly generated mask vector $\mathbf{p} \in \mathbb{R}^{l \times 1}$. Each dimension of $\mathbf{p}$ is in $\{0, 1\}$ and generated by a Bernoulli distribution with parameter p.

$$\mathbf{A'_{h,t}} = \mathbf{p} * \mathbf{A_{h,t}} \tag{5}$$

Applying the random mask on attention score $\mathbf{A_{h,t}} \in \mathbb{R}^{l \times 1}$ can be interpreted as randomly filter out some context information since the attention score for them are set to 0. In addition, the degree of perturbation can be controlled by setting proper p. Then we can get the perturbed pooled context representation $c'_{h,t}$ in Equation 6.

$$\mathbf{c'_{h,t}} = \mathbf{H^T} \cdot \frac{\mathbf{A'_{h,t}}}{\mathbf{1^T} \cdot \mathbf{A_{h,t}}} \tag{6}$$

For all the entity pairs (e_h, e_t) whose relation r in $\mathcal{R}^{aug}$, we apply the prior steps to get α distinct pertubed context representations $\{\mathbf{c'_{i,h,t}}\}_{i=1}^{|\alpha|}$ by using α random mask, where α is a hyperparameter for controlling the number of ERA operations. Eventually, we can get the augmented triple representation set $\mathcal{T}_{aug}$, which can be formulated in Equation 7.

$$\mathcal{T}_{aug} = \{(\mathbf{e_h}, \mathbf{c'_{i,h,t}}, \mathbf{e_t}) | e_h \in \mathcal{E}, r \in \mathcal{R}^{aug}, e_t \in \mathcal{E}\} \tag{7}$$

Combining the original triple representation set $\mathcal{T}_{orig}$ and $\mathcal{T}_{aug}$, we can get the total tripe representation set $\mathcal{T}$ for relation prediction and our Contrastive Learning framework.

$$\mathcal{T} = \mathcal{T}_{orig} \cup \mathcal{T}_{aug} \tag{8}$$

3.5 Relation Prediction

Based on the triple representation of all entity pairs, the relation prediction module finally predict the relations hold between each pair. For a triple representation $(\mathbf{e_h}, \mathbf{c_{ht}}, \mathbf{e_t}) \in \mathcal{T}$, we first apply two linear transformations with Tanh activation to fuse the pooled context representation $\mathbf{c}_{ht}$ with $\mathbf{e}_h$ and $\mathbf{e}_t$.

$$\mathbf{h} = tanh(\mathbf{W}_h \cdot \mathbf{e}_h + \mathbf{W}_{c1} \cdot \mathbf{c}_{h,t}) \tag{9}$$

$$\mathbf{t} = tanh(\mathbf{W}_t \cdot \mathbf{e}_t + \mathbf{W}_{c2} \cdot \mathbf{c}_{h,t}) \tag{10}$$

Where $\mathbf{W}_h, \mathbf{W}_t, \mathbf{W}_{c1}, \mathbf{W}_{c2} \in \mathbb{R}^{d \times d}$, which are trainable parameters of the model. Following Zhou et al. (2021), then we used a grouped bi-linear layer to calculate a score for relation r, which splits the

vector representation to k groups and performs bilinear within group.

$$score_r = \sum_{i=1}^{k} \mathbf{h}^{iT} \mathbf{W}_r^i \mathbf{t}^i \tag{11}$$

Where $\mathbf{W}_r^i \in \mathbb{R}^{d/k \times d/k}$ is the bilinear parameter of group i. During training stage, we apply the adaptive thresholding loss (Zhou et al., 2021) to dynamically learn a threshold $\theta_{h,t}$ for each entity pair by introducing a threshold class TH.

$$\mathcal{L}_{h,t} = -\sum_{r \in \mathcal{P}_{h,t}} log \left[\frac{exp(score_r)}{\sum_{r' \in \mathcal{P}_{h,t} \cup \{TH\}} exp(score_{r'})} \right] - log \left[\frac{exp(score_{TH})}{\sum_{r' \in \mathcal{N}_{h,t} \cup \{TH\}} exp(score_{r'})} \right] \tag{12}$$

$\mathcal{P}_{h,t} \subset \mathcal{R}$ is the set of all valid relations that hold between entity pair (e_h, e_t), and it is empty when no relation hold between the pair. In addition, $\mathcal{N}_{h,t} = \mathcal{R} - \mathcal{P}_{h,t}$. In the inference stage, the threshold θ for valid relation scores is set to $score_{TH}$.

4 Contrastive Learning for relation pre-training

4.1 Overview

We propose a contrastive learning(CL) framework for unifying the augmented relation representations and improving the robustness of learned relation representations, especially for tailed relations. Specifically, we use the CL framework for pre-training on the distantly-supervised DocRE dataset(Yao et al., 2019), which is annotated by querying the knowledge graph but is noised. Considering that the model will be fine-tuned on the human-annotated dataset after the representation learning stage, the noise in the distantly supervised dataset is acceptable and correctable.

Under DocRE setting, we claim that the semantically-similar samples should be the entity pairs that have the same relation r, including both of the original pairs and augmented pairs by ERA. However, only a few entity pairs have the same relation within one document, especially for the tailed relation r.

Increasing the mini-batch size can partially mitigate the problem, but it requires large GPU memory for training which may not be accessible. Thus, we adapted the MOCO framework (He et al., 2020) to the DocRE setting, named MoCo-DocRE. The

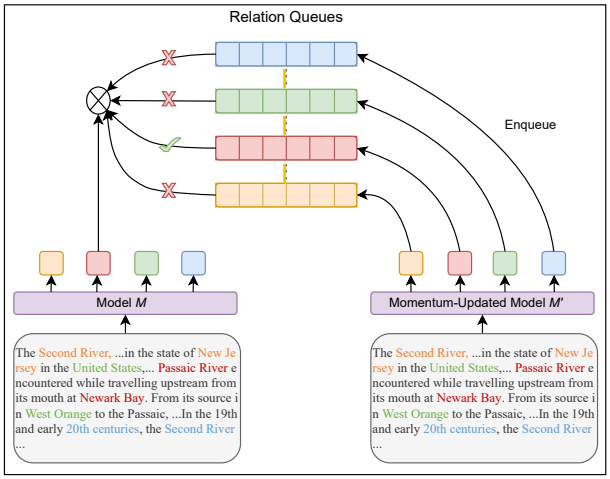

Figure 2: Overview of the proposed MoCo-DocRE framwork.

moCo-DocRE framework can conduct the CL without using large batch-size by keeping a relation representation queue Q_r holding q relation representations from the previous mini-batch for each relation $r \in \mathcal{R}$. This allows us to reuse the encoded positive and negative relation representations in prior mini-batches. We summarize our contrastive learning framework in Figure 2.

4.2 Anchor relation encoding

For document $\mathcal{D}$ in pre-training dataset, we first conduct the aforementioned document encoding, relation encoding and Easy Relation Augmentation(ERA) and obtain the triple representation set $\mathcal{T}$ of all entity pairs. For a triple representation $(\mathbf{e_h}, \mathbf{c_{h,t}}, \mathbf{e_t}) \in \mathcal{T}$, we use two linear transformations to fuse the triple representation, which are same as Equation 9 and 10. Next, we use a MLP layer with ReLU activation for final relation representation:

$$\mathbf{x} = relu(\mathbf{W}_2(\mathbf{W}_1[\mathbf{h} : \mathbf{t}] + \mathbf{b}_1) + \mathbf{b}_2) \tag{13}$$

Where $[:]$ denotes the vector concatenation operation, $\mathbf{W}_1 \in \mathbb{R}^{2d \times d}$ and $\mathbf{W}_2 \in \mathbb{R}^{d \times d_r}$ are trainable model parameters in pre-training stage, and d_r is the dimension of final relation representation $\mathbf{x}_{h,t}$. After contrastive pre-training, the MLP layer will not be used for relation prediction in fine-tuning.

4.3 MoCo-DocRE

To keep the consistency of relation representation in Q_r, we also use a momentum updated model to encode the positive and negative samples in contrastive learning(He et al., 2020). The original model $\mathcal{M}$ is updated via back-propagation, and

the momentum-updated model $\mathcal{M}'$ is updated by Equation 14.

$$\mathcal{M}' = m \cdot \mathcal{M}' + (1 - m) \cdot \mathcal{M} \quad (14)$$

Where m is the momentum hyper-parameter, which can control the evolving speed of M'. Next we feed the document $\mathcal{D}$ to $\mathcal{M}'$ for $\mathbf{x}'$ by following the same procedure as getting anchor relation representation $\mathbf{x}$. Then we push $\{\mathbf{x}'|(\mathbf{e}_h, \mathbf{c}_{h,t}, \mathbf{e}_t) \in \mathcal{T}'\}$ to $|\mathcal{R}|$ relation representation queues according to their relation labels. If relation r holds between (e_h, e_t), then x' will be pushed to Q_r. Eventually, we can get the set of positive and negative relation representations of $\mathbf{x}$ from queues, i.e, $\mathcal{P} = \cup_{r \in \mathcal{P}_{h,t}} Q_r$ and $\mathcal{N} = \cup_{r \in \mathcal{N}_{h,t}} Q_r$.

For anchor relation representation $\mathbf{x}$, now we can formalize the INFONCE loss (van den Oord et al., 2018) under our MoCo-DocRE in Equation 15.

$$\mathcal{L} = -\sum_{x^+ \in \mathcal{P}} log \left[\frac{e^{\mathbf{x}^T \mathbf{x}^+ / \tau}}{e^{\mathbf{x}^T \mathbf{x}^+ / \tau} + \sum_{x^- \in \mathcal{N}} e^{\mathbf{x}^T \mathbf{x}^- / \tau}} \right] \quad (15)$$

Where τ is the temperature hyperparameter. In addition $\mathbf{x}, \mathbf{x}^+, \mathbf{x}^-$ in Equation 15 are *l2*-normalized.

5 Experiment

5.1 Experiment Setup and Dataset

Dataset: We evaluate the proposed ERA and contrastive learning framework on two popular DocRE datasets, DocRED(Yao et al., 2019) and HacRED(Cheng et al., 2021). DocRED contains 5053 English documents extracted from Wikipedia and 96 relations, which are human-annotated. Besides, DocRED also provide a distantly-supervised dataset with 101873 documents, and the relation of entitity pairs are annotated by querying Wikidata. HacRED is a human annotated Chinese dataset with 26 relations. Statistics of Datasets are listied on Table 2.

Implementation Details: We use the pre-trained BERT-base-cased(Devlin et al., 2019) as our backbone for DocRED dataset. All the hyperparameters are tuned on the development set. Specifically, we set the random mask probability p to 0.1 and the number of augmentation α to 2. In addition, the number of grouped bilinear k is set to 64. The temperature parameter τ is set to 0.5 and the size of Q_r, i.e., q is set to 500, and the momentum m is set to 0.99. The learning rate is set to $1e-5$ for pre-training on our CL framework. In the fine-tune on human-annotated data, we set the learning rate to $5e-5$ for parameters of BERT and $1e-4$ for other parameters. We use AdamW(Loshchilov and Hutter, 2019) for optimization of all parameters and a linear-decayed scheduler with a warmup ratio 0.06. Gradients whose norm is larger than 1 are clipped. For HacRED dataset, we use XLM-Roberta-base(Conneau et al., 2020) as backbone. Under HacRED scenario, we set the random mask probability p to 0.05 and the number of augmentation α to 3. All the other parameters are same as the DocRED scenario.

5.2 Evaluation Metric

DocRED benchmark provide two evaluation metrics, i.e. F_1 Ign F_1. F_1 is the minor F_1 value for all predicted relations in test/development dataset, which can reflect the overall performance of DocRE. Compared to F_1, Ign F_1 excludes the entity pairs which appear both on training and test/dev data. To demonstrate how ERA and contrastive learning can improve the performance of tailed relations, we propose to use the following evaluation metrics:

Macro: it computes the F_1 value by first calculating F_1 for each relation separately and then getting the average of all relation classes. Compared to minor F_1, macro F_1 treat all relation classes equally, F_1 of tailed relations will have equal impact compared to head relations.

Macro@500,Macro@200,Macro@100: Those metrics target at tailed relations whose frequency count in train dataset is less than 500,200,100 respectively. Values are computed by averaging the F1 value of the targeted relations.

5.3 Main Results

We compare the proposed ERA and ERACL methods to graph-based and transformer-based models on the DocRED benchmark by using F_1 and Ign F_1 metrics on the dev/test dataset. Results are reported in Table 1. The proposed ERACL method, which first conducts contrastive learning under our MoCo-DocRE framework on the distantly supervised dataset and then conducts ERA fine-tuning on the training set, can achieve competing F_1 and Ign F_1 value, compared to state-of-art graph-based methods and transformer-based methods. Besides, compared to ATLOP(Zhou et al., 2021) which is the baseline of ERACL, ERACL can improve the minor F_1 on the development set by 0.71 and ERA can improve the minor F_1 by

Model	Dev		Test	
	Ign F_1	F_1	Ign F_1	F_1
Graph-based Methods				
GLRE(Wang et al., 2020)	–	–	55.40	57.40
LSR(Nan et al., 2020)	52.43	59.00	56.96	59.05
GAIN(Zeng et al., 2020)	59.14	61.22	59.00	61.24
DRN(Xu et al., 2021b)	59.33	61.39	59.15	61.37
Transformer-based Methods				
SSAN(Xu et al., 2021a)	57.04	59.19	56.06	58.41
ATLOP(Zhou et al., 2021)	59.22	61.09	59.31	61.30
DocuNet(Zhang et al., 2021)	59.86	61.83	59.93	61.86
AFLKD(Tan et al., 2022)	60.08	62.03	60.04	62.08
Our Methods				
ERA	59.30 ± 0.09	61.30 ± 0.08	58.71	60.97
ERACL	59.72 ± 0.19	61.80 ± 0.20	59.08	61.36

Table 1: Overall DocRE performance evaluated on DocRED benchmark. We report the mean and standard deviation of 3 runs with different random seeds on the development set. The official test results are reported by the best checkpoint on the development set. Results of all other models are reported in their original paper and use BERT-base-cased as backbone encoder.

Statistics	DocRED	HacRED
# Train	3053	6231
# Dev	1000	1500
# Test	1000	1500
# Relations	96	26

Table 2: Dataset Statistics of DocRED and HacRED

0.29, which demonstrate the effectiveness of the proposed ERA method and contrastive learning pretraining.

5.4 Results on tailed relation

To demonstrate the effectiveness of our ERA and ERACL on improving model performance on tailed relations, we evaluated ERA and ERACL using Macro, Macro@500, Macro@200, and Macro@100 metrics. Besides, we compare ERA and ERACL with three baseline methods which are used for addressing long-tailed distribution on DocRED. **Text Random Deletion/Mask** are commonly used data augmentation techniques for NLP tasks. We apply the Text Random Deletion and masking as data augmentation for documents which contain tailed relations. **Adaptive Focal Loss** proposed by Tan et al. (2022) is a adaptation of Focal Loss(Lin et al., 2017) on DocRE scenario. AFL proved to be effective on tail relations. We implement those three methods based on ATLOP. Results are listed on Table 3. The proposed ERA and ERACL method can outperform those three baselines on tailed relations, which demonstrates the effectiveness and necessities of our ERA and ERACL for addressing the long-tailed distribution on DocRE scenario.

Moreover, the proposed ERA method can improve the Macro over ATLOP by 1.01, 1.01 on Macro@500, 1.69 on Macro@200, and 1.74 on Macro@100. We observe that the improvements are more significant on relations that appear less frequently. In addition, the proposed ERACL method can further gain improvements over ERA: 0.79 on Macro, 0.92 on Macro@500, 0.92 on Macro@200, 1.81 on Macro@100, which also show similar trends as ERA over ATLOP.

To better illustrate the performance gain on the tailed relations, we sort 96 relations according to their frequency count in the DocRED train set from high to low, then slice 96 relations to 10 relation clusters equally for more clear visualization. For each cluster, we calculate the cluster F1 by averaging the F1 of relation within the cluster. The results are demonstrated in Figure 3. We observe that the proposed ERA method gain improvements compared to ATLOP on relation clusters 4-10, which correspond to the tailed relation in DocRED, and also achieve competing performance on clusters

Methods	Macro	Macro@500	Macro@200	Macro@100
DocuNet	40.69	36.54	28.95	**22.37**
ATLOP	39.54	35.20	26.82	18.76
–Deletion	39.22	34.88	26.62	18.78
–Mask	38.31	33.80	25.30	17.35
–AFL	40.04	35.78	27.78	20.52
ERA	40.55	36.21	28.51	20.50
ERACL	**41.34**	**37.13**	**29.43**	22.31

Table 3: Evaluation on tailed relations. All the results are averaged on 3 runs with different random seeds on development set. Relation labels of test set are not accessible, so the results on test set cannot be reported.

Methods	F1	Macro	Macro@500	Macro@200	Macro@100
ERACL	61.80	41.34	37.13	29.43	22.31
– ERA	61.36	40.49	36.22	28.61	21.52
– CL	61.30	40.55	36.21	28.51	20.50
– both	60.97	39.54	35.20	26.82	18.76

Table 4: Ablation Study on development set.

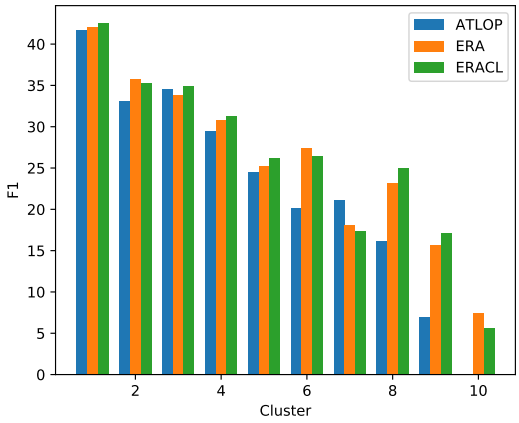

Figure 3: Comparison of F1 across 10 relation clusters. All results are averaged by 3 runs with different random seeds.

1-3, which correspond to the head relations. Those findings show that our ERA methods are effective for improving the DocRE performance on tailed relations while keeping the performance on head relations. In addition, similar performance gain is also achieved by the proposed ERACL method, and ERACL can further improve the tailed relations compared to ERA and achieve competing performance on head relations.

In addition, we conduct another set of experiments by manually reducing the percentage of training data in order to explore the performance of the proposed ERA methods and ERACL methods under a limited-data scenario. The results are listed in Table 5. Compared to the setting that uses all of the train data, we observe that the performance gain of the proposed ERA and ERACL under 10% and 5% settings are more significant, which also indicate that the proposed ERA and ERACL can improve the DocRE performance by mitigating the long-tailed problem and are especially effective when training data is limited.

Methods	F1	Macro	Macro@200
10%			
ATLOP	51.92	23.68	7.97
ERA	52.46	24.80	11.20
ERACL	**53.73**	**27.88**	**14.89**
5%			
ATLOP	45.17	16.17	5.59
ERA	47.18	17.27	6.22
ERACL	**49.40**	**23.47**	**11.84**

Table 5: Results on the development set under the limited-data setting. 10% refers to only using 10% of training data, and 5% refers to only using 5% of training data. All results are reported by averaging 3 runs with different random seeds.

Besides, we also conduct experiments on HacRED to investigate whether our ERA can generalize well on other long-tailed DocRE datasets. Results are shown on Table 6. We observe that ERA

can still outperform the ATLOP on tailed relations.

Methods	F1	Macro	Macro@500
ATLOP	77.84	70.99	55.11
ERA	78.27	71.73	57.13

Table 6: Results on HacRED. Since HacRED do not have distant-labeled data, we can only evaluate ERA on HacRED. ATLOP results are implemented by us. All experiments use XLM-Roberta-base as the backbone encoder.

5.5 Ablation Study

To evaluate the contribution of the ERA and contrastive learning(CL) framework separately, we conduct an ablation study on the development set by reducing one component at a time. The results are shown in Table 4. All of the results are tuned on the development set for best performance.

Note that reducing ERA refers to turning off the relation representation augmentation operation described in Section 3.4 and only keeping the original relation representations. In addition, reducing CL means without conducting contrastive learning on distantly supervised data. We observe that the ERA component and contrastive learning(CL) framework are almost equally important, which lead to 0.44 and 0.50 performance drop on F1 metric, 0.85 and 0.79 performance drop on Macro F1.

6 Conclusion

We propose a novel Easy Relation Augmentation(ERA) method for the Document-level Relation Extraction task, which improves the DocRE performance by addressing the long-tailed problem residing in DocRE by augmentation on relation representations. In addition, we propose a novel contrastive learning framework based on ERA, i.e., MoCo-DocRE, for unifying the augmented relation representations and improving the robustness of learned relation representations, especially for tailed relations. Experiments on the DocRED dataset demonstrate that the proposed ERA and ERACL can achieve competing performance compared to state-of-arts models, and we demonstrate that the performance gain of ERA and ERACL are mainly from the tailed relations.

Nonetheless, addressing the long-tailed problem is still challenging for DocRE. One limitation of our method is it still relies on large amount of annotated data to achieve overwhelming performance. We hope it can be mitigated in future research.

References

Jianlong Chang, Jie Gu, Lingfeng Wang, GAOFENG MENG, SHIMING XIANG, and Chunhong Pan. 2018. Structure-aware convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

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 the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2819–2831, Online. Association for Computational Linguistics.

Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Connecting the Dots: Document-level Neural Relation Extraction with Edge-oriented Graphs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4925–4936, Hong Kong, China. Association for Computational Linguistics.

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 the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. CERT: Contrastive Self-supervised Learning for Language Understanding. *arXiv:2005.12766 [cs, stat]*. ArXiv: 2005.12766.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738.

Robin Jia, Cliff Wong, and Hoifung Poon. 2019. Document-level n-ary relation extraction with multiscale representation learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3693–3704, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

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, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (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. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Guoshun Nan, Zhijiang Guo, Ivan Sekulic, and Wei Lu. 2020. Reasoning with Latent Structure Refinement for Document-Level Relation Extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1546–1557, Online. Association for Computational Linguistics.

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 the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3661–3672, Online. Association for Computational Linguistics.

Yujia Qin, Yankai Lin, Ryuichi Takanobu, Zhiyuan Liu, Peng Li, Heng Ji, Minlie Huang, Maosong Sun, and Jie Zhou. 2021. ERICA: Improving entity and relation understanding for pre-trained language models via contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3350–3363, Online. Association for Computational Linguistics.

Sunil Kumar Sahu, Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Inter-sentence Relation Extraction with Document-level Graph Convolutional Neural Network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4309–4316, Florence, Italy. Association for Computational Linguistics.

Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. 2022. Document-level relation extraction with adaptive focal loss and knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1672–1681, Dublin, Ireland. Association for Computational Linguistics.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

Difeng Wang, Wei Hu, Ermei Cao, and Weijian Sun. 2020. Global-to-Local Neural Networks for Document-Level Relation Extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3711–3721, Online. Association for Computational Linguistics.

Hong Wang, Christfried Focke, Rob Sylvester, Nilesh Mishra, and William Wang. 2019. Fine-tune Bert for DocRED with Two-step Process. *arXiv:1909.11898 [cs]*. ArXiv: 1909.11898.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Lingfei Wu, Yu Chen, Kai Shen, Xiaojie Guo, Hanning Gao, Shucheng Li, Jian Pei, and Bo Long. 2021. Graph neural networks for natural language processing: A survey. *arXiv preprint arXiv:2106.06090*.

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. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14149–14157. AAAI Press.

Wang Xu, Kehai Chen, and Tiejun Zhao. 2021b. Discriminative Reasoning for Document-level Relation Extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1653–1663, Online. Association for Computational Linguistics.

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 the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.

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 the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7170–7186, Online. Association for Computational Linguistics.

Shuang Zeng, Runxin Xu, Baobao Chang, and Lei Li. 2020. Double Graph Based Reasoning for Document-level Relation Extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1630–1640, Online. Association for Computational Linguistics.

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 the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 3999–4006. ijcai.org.

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 the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14612–14620. AAAI Press.