ARPDL: Adaptive Relational Prior Distribution Loss as an Adapter for Document-Level Relation Extraction

Huangming Xu, **Fu Zhang**[*], **Jingwei Cheng** and **Xin Li**
School of Computer Science and Engineering, Northeastern University, China
xuhuangming@foxmail.com, zhangfu@mail.neu.edu.cn

Abstract

The goal of document-level relation extraction (DocRE) is to identify relations between entities from multiple sentences. As a multi-label classification task, a common approach is to determine whether there are relations for an entity pair by selecting a multi-label classification threshold, with scores of relations above the threshold predicted as positive and the rest as negative. However, we find that predicting multiple relations for entity pairs causes the decrease of predicted scores in positive classes. This could lead to many positive classes being incorrectly predicted as negative. Additionally, our analysis suggests that fitting the distribution of predicted relations to the prior distribution of relations can help improve prediction performance. However, previous studies have not explored or leveraged the prior distribution of relations. To address these issues and findings, we for the first time propose the idea of incorporating the relational prior distribution into the loss calculation in DocRE tasks. We innovatively propose an **A**daptive **R**elational **P**rior **D**istribution **L**oss (ARPDL), which can adaptively adjust relation prediction scores based on the relational prior distribution. Our designed relational prior distribution component can also be integrated as an adapter into other threshold-based losses to improve prediction performance. Experimental results demonstrate that ARPDL consistently improves the performance of existing DocRE models, achieving new state-of-the-art results. Furthermore, integrating our relational prior distribution adapter into other losses significantly enhances their performance in DocRE tasks, validating the effectiveness and generality of our approach. Code is available at https://github.com/xhm-code/ARPDL.

1 Introduction

Document-level relation extraction (DocRE), which is an information extraction task developed based on sentence-level relation extraction, aims to extract one or more relations for an entity pair from sentences in a document. This task has garnered increasing attention due to its alignment with real-world applications, where a large number of relational facts are expressed in multiple sentences [Yao *et al.*, 2019].

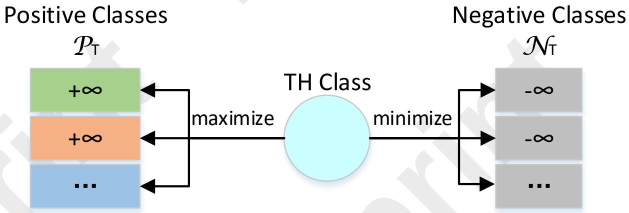


Figure 1: A threshold (TH) class is introduced to distinguish between positive and negative classes: during the training phase, the ATL is designed to ensure that the scores of positive classes are significantly higher than TH, while the scores of negative classes are significantly lower than TH.

As a multi-label classification task, the complexity of DocRE grows exponentially with the number of entities. When a document contains n entities, relations need to be extracted for $n(n-1)$ entity pairs. Moreover, DocRE exhibits a severe *class imbalance* [Yao *et al.*, 2019; Tan *et al.*, 2022b] between the positive and negative classes[1]. Additionally, even among positive classes, various relation types are highly imbalanced, forming a *long-tail* problem. To address these issues, an adaptive thresholding loss (ATL)-based selection approach [Zhou *et al.*, 2021] is commonly used. As illustrated in Figure 1, for each entity pair (e_s, e_o), classes with scores exceeding the threshold are predicted as positive, while the remaining classes are predicted as negative.

The subsequent losses such as adaptive focal loss (AFL) [Tan *et al.*, 2022a], none class ranking loss (NCRL) [Zhou and Lee, 2022], and adaptive hinge balance loss (HingeABL) [Wang *et al.*, 2023] are all improvements based on the ATL. NCRL and HingeABL reveal that too many negative classes

[*] Corresponding author

[1] Given a pre-defined set R of relations of interest, *Positive classes* $\mathcal{P}_T \subseteq R$ are the relations that exist for an entity pair, while *Negative classes* $\mathcal{N}_T \subseteq R$ are the relations that do not exist for an entity pair. That is, $R = \mathcal{P}_T \cup \mathcal{N}_T$. Moreover, 'NA' (no relation) denotes the absence of any relation between an entity pair and is considered a distinct category, separate from the Negative classes.

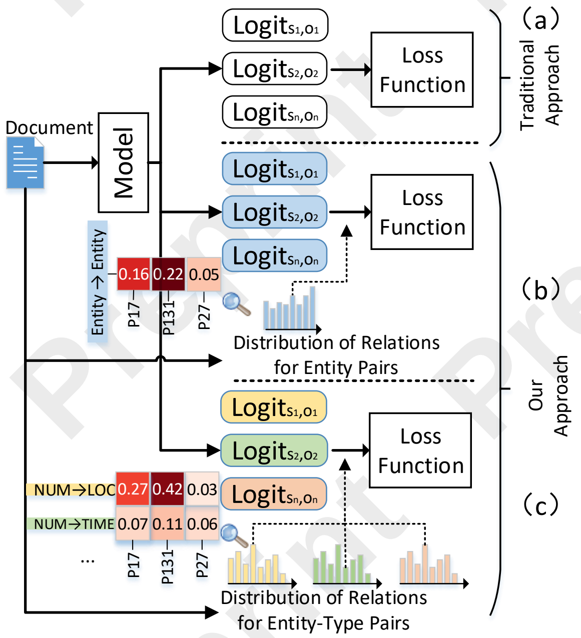


Figure 2: (a) Flow of the loss used in the traditional method. (b) Loss based on relational prior distribution of entity pairs. (c) Loss based on relational prior distribution of entity-type pairs. *Logit* denotes the score output by the model, s and o represent the subject and object of an entity pair. In (b), the same relational prior distribution is introduced for different entity pairs. In (c), for an entity pair, a relational prior distribution corresponding to the entity-type pair is introduced. Different colors indicate different entity-type pairs.

raise the learned threshold, causing this threshold to exceed the scores of positive classes. This may lead to numerous false-negative predictions, thus they improve the ATL by reducing the proportion of negative classes. AFL mitigates the impact of head classes (frequent relation types) in long-tail distributions, thereby allowing tail classes (infrequent relation types) to receive more attention. However, these strategies, which either reduce the proportion of negative classes or focus on tail classes, struggle to effectively address both class imbalance and long-tail issues simultaneously, thereby limiting their performance.

Further, our findings reveal *two new issues* in ATL and its improved losses, which are crucial to improving the performance of DocRE and may also help address the class imbalance and long-tail issues: (1) As the number of relations between entities increases, the prediction scores of these relations decrease and tend to be close. This could lead to many positive classes being incorrectly predicted as negative, as shown in our theoretical proof (Section 3.3) and experiments (Section 6.1). (2) Our analysis indicates that the prior distribution of relations may guide the models to adjust the prediction scores. And if the distribution of predicted relations fits the prior distribution of relations, it intuitively enhances the prediction performance. However, previous studies have not explored or leveraged the prior distribution of relations.

To address these issues, we propose a novel idea of leveraging and incorporating relational prior distribution into loss as shown in Figure 2. Firstly, we introduce relational prior distribution to dynamically adjust relation prediction scores during training, tackling the issue of decrease in the scores of positive classes when predicting multiple relations. Secondly, in order to address the false-negative problem caused by the class-imbalance, we for the first time propose to solve this problem by increasing the contribution of positive classes to loss. Additionally, introducing relational prior distribution can also expand the training knowledge for tail classes with few relation types in the long-tail problem, significantly enhancing the prediction performance of these tail classes.

Based on the above findings and analysis, we propose an **A**daptive **R**elational **P**rior **D**istribution **L**oss (ARPDL), which introduces relational prior distribution from two different granularities: One at the entity pair level, as illustrated in Figure 2(b), which treats each entity pair as having the same relational prior distribution; The other at the entity-type pair level, as shown in Figure 2(c), which more finely considers the relational prior distribution corresponding to the types of entity pairs. Our contributions are as follows:

- We, for the first time, propose the idea of incorporating relational prior distribution into loss in DocRE tasks. ARPDL can adaptively adjust and improve relation prediction scores based on the relational prior distribution.
- We propose a strategy in ARPDL to address the class-imbalance problem by increasing the contribution of positive classes to loss, while also effectively mitigating the long-tail problem. This is different from the previous methods of reducing the proportion of negative classes.
- Our relational prior distribution component can be integrated as an adapter into ATL-based losses, which can significantly enhance their performance in DocRE tasks.

Experimental results on DocRE datasets demonstrate that our loss ARPDL consistently improves the performance of different backbones, achieving new state-of-the-art (SOTA) results. In addition, integrating our relational prior distribution adapter into other losses significantly enhances their performance, validating the effectiveness and generality of our idea based on the relational prior distribution.

2 Related Work

2.1 Document-Level Relation Extraction

Existing DocRE methods are mainly divided into two categories: Transformer-based and graph-based models. The former [Zhou *et al.*, 2021; Xie *et al.*, 2022; Xiao *et al.*, 2022; Ma *et al.*, 2023; Lu *et al.*, 2023; Zhang *et al.*, 2024; Gao *et al.*, 2024] uses pre-trained language models to encode documents to obtain contextual information, which can capture relations between entities. The latter [Xu *et al.*, 2021; Peng *et al.*, 2022; Sun *et al.*, 2023; Jain *et al.*, 2024] uses graph neural networks to integrate and reason about relations between entities by constructing a graph structure of different text information (e.g., entities and sentences).

2.2 Loss Functions in DocRE

In DocRE multi-label classification tasks, the adaptive thresholding loss (ATL) is the most commonly used loss [Zhou *et* *al.*, 2021]. In ATL, a threshold class (TH class) is introduced. If there is a relation for an entity pair, the score of the relation is higher than the threshold, otherwise it is lower than the threshold. Subsequently, many losses for DocRE are improved based on ATL, such as AFL [Tan *et al.*, 2022a], adaptive margin loss (AML) [Wei and Li, 2022], NCRL [Zhou and Lee, 2022], PEMSCL [Guo *et al.*, 2023], separate adaptive thresholding (SAT) [Wang *et al.*, 2023], and HingeABL [Wang *et al.*, 2023]. These losses mainly solve the problem of imbalance between positive and negative classes or the long-tail problem in the positive class.

3 Methodology

In this section, we first give the definition of DocRE. Then, we analyze the classic loss ATL in DocRE. Finally, we propose the loss ARPDL we designed.

3.1 Problem Formulation

DocRE is a multi-label classification task. Given a document D that contains a known set of entities $\{e_i\}_{i=1}^{n}$ and their corresponding types, these entities form multiple entity pairs (e_s, e_o), where e_s is the subject and e_o is the object. The model will determine which of the pre-defined relation set $R \cup \{\mathrm{NA}\}$ each entity pair (e_s, e_o) belongs to. R represents a pre-defined set of relations of interest, and NA indicates that there is no relation for an entity pair.

3.2 Adaptive Thresholding Loss

Adaptive Thresholding Loss (ATL) [Zhou *et al.*, 2021] is a loss widely used in DocRE tasks. Compared to binary cross-entropy (BCE) loss, ATL can adaptively set thresholds for each entity pair (BCE uses a global threshold). As shown in Eq.(1), ATL is composed of two parts, $\mathcal{L}_1$ and $\mathcal{L}_2$. $\mathcal{L}_1$ includes positive classes $\mathcal{P}_T$ and the threshold (TH) class, which needs to be learned during training. $\mathcal{L}_2$ includes negative classes $\mathcal{N}_T$ and the TH class. The optimization goal of ATL is to ensure positive class scores are significantly higher than the TH class score, and negative class scores are significantly lower. In the testing phase, if an entity pair is predicted to have one or more relations, the scores for these relations must be above the threshold. When all scores are below the threshold, the entity pair is predicted as NA.

$$\begin{aligned}
\mathcal{L}_1 &= -\sum_{r\in\mathcal{P}_T} \log\left(\frac{\exp(logit_r)}{\sum_{r'\in\mathcal{P}_T\cup\{\mathrm{TH}\}}\exp\left(logit_{r'}\right)}\right) \\
\mathcal{L}_2 &= -\log\left(\frac{\exp(logit_{\mathrm{TH}})}{\sum_{r'\in\mathcal{N}_T\cup\{\mathrm{TH}\}}\exp\left(logit_{r'}\right)}\right) \\
\mathcal{L}_{ATL} &= \mathcal{L}_1 + \mathcal{L}_2
\end{aligned} \tag{1}$$

3.3 Our Empirical Analysis of ATL

For $\mathcal{L}_1$, let $Z = \sum_{r'\in\mathcal{P}_T\cup\{TH\}}\exp(logit_{r'})$. Then $\mathcal{L}_1$ can be decomposed into Eq.(2):

$$\begin{aligned}
\mathcal{L}_1 &= -\sum_{r\in\mathcal{P}_T}\log\left(\frac{\exp(logit_r)}{Z}\right) = -\sum_{r\in\mathcal{P}_T}\left(logit_r - \log(Z)\right) \\
&= -\sum_{r\in\mathcal{P}_T} logit_r + |\mathcal{P}_T|\log(Z)
\end{aligned} \tag{2}$$

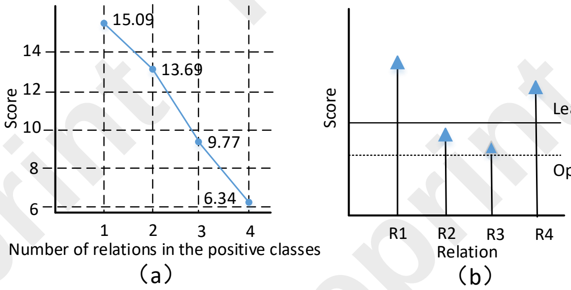


Figure 3: (*a*) The $\mathcal{L}_1$ problem in ATL: the output of ATLOP [Zhou *et al.*, 2021] on the Re-DocRED [Tan *et al.*, 2022b] dataset, utilizing BERT$_{base}$ as the encoder and employing ATL as the loss. The horizontal axis represents the number of relations in positive classes corresponding to entity pairs, and the vertical axis represents the average predicted score of positive classes. (*b*) The $\mathcal{L}_2$ problem in ATL. The threshold (TH) learned by ATL is much higher than the candidate score, resulting in an increase in the number of false-negative predictions.

Taking the derivative of $logit_r$ of $\mathcal{L}_1$, we get:

$$\frac{\partial\mathcal{L}_1}{\partial logit_r} = -1 + \frac{|\mathcal{P}_T|}{Z}\frac{\partial Z}{\partial logit_r} \tag{3}$$

$$\begin{aligned}
\frac{\partial Z}{\partial logit_r} &= \frac{\partial}{\partial logit_r}\left(\sum\nolimits_{r'\in\mathcal{P}_T\cup\{TH\}}\exp(logit_{r'})\right) \\
&= \exp(logit_r)
\end{aligned} \tag{4}$$

Further, substituting Eq.(4) into Eq.(3), we obtain Eq.(5):

$$\frac{\partial\mathcal{L}_1}{\partial logit_r} = -1 + \frac{|\mathcal{P}_T|}{Z}\exp(logit_r) \tag{5}$$

We set the derivative to zero and find the extreme point:

$$-1 + \frac{|\mathcal{P}_T|\exp(logit_r)}{Z} = 0 \tag{6}$$

$$\frac{\exp(logit_r)}{Z} = \frac{\exp(logit_r)}{\sum_{r'\in\mathcal{P}_T\cup\{TH\}}\exp(logit_{r'})} = \frac{1}{|\mathcal{P}_T|} \tag{7}$$

Finally, according to Eq.(7), we deduce that each $logit_r$ (where $r \in \mathcal{P}_T$) is close.

For $\mathcal{L}_2$, according to the deduction in [Wang *et al.*, 2023], we can obtain Eq.(8). If $\mathcal{L}_2 \to 0$, then $\sum_{r'\in\mathcal{N}_T}\exp(\mathrm{logit}_{r'} - \mathrm{logit}_{\mathrm{TH}}) \approx 0$, meaning $\mathrm{logit}_{r'} \ll \mathrm{logit}_{\mathrm{TH}}$. This indicates that the threshold is significantly higher than the predicted scores of relations, leading to numerous false-negative predictions.

$$\begin{aligned}
\mathcal{L}_2 &= -\log\left(\frac{\exp(logit_{\mathrm{TH}})}{\sum_{r'\in\mathcal{N}_T\cup\{\mathrm{TH}\}}\exp(logit_{r'})}\right) \\
&= -\log\left(\frac{1}{1+\sum_{r'\in\mathcal{N}_T}\exp(logit_{r'} - logit_{\mathrm{TH}})}\right)
\end{aligned} \tag{8}$$

Through the above deduction, we draw the following conclusions:

- When there are n relations for an entity pair ($n \geq 1$), according to Eq.(7), we deduce the scores of relations $r \in \mathcal{P}_T$ for the entity pair tend to be close. Therefore, in $\mathcal{L}_1$, as the number of relations existing for an entity pair increases, this may lead to the **decrease in the scores of positive classes**, as shown in Figure 3(*a*);
- For $\mathcal{L}_2$, since the number of relations in negative classes is greater than that of positive classes, the contribution of negative classes to the loss is greater than that of positive classes, which will **increase the threshold** (as shown in Figure 3(*b*)) and make the model more inclined to predict the NA label for entity pairs, **resulting in** a large number of **false-negative predictions**.

3.4 Adaptive Relational Prior Distribution Loss

To address the issues above, we propose a new adaptive relational prior distribution loss, ARPDL, by leveraging and incorporating relational prior distribution into loss. Considering that there are two types of relational prior distributions associated with entity pairs, namely, relational prior distributions of *entity pairs* and *entity-type pairs*. The latter refers to the relational prior distribution related to the types of entities in entity pairs. Therefore, we propose our loss from two granularities: entity pair level and entity-type pair level.

ARPDL based on Relational Prior Distribution of Entity Pairs

We introduce the relational prior distribution at the entity pair level into the loss, which treats each entity pair as having the same relational prior distribution, as illustrated in Figure 2(b). We first count the frequency of occurrence of each relation in the training set as the *prior probability* of this relation, and then form *prior distribution* of the relations. Then, we calculate the predicted probability of a relation r for a given entity pair as follows:

$$P(r) = \log \frac{\exp(logit_r)}{\sum\limits_{r' \in R \cup \{\text{NA}\}} \exp(logit_{r'})} - \frac{\exp(rd_r)}{\sum\limits_{i=1}^{|R \cup \{\text{NA}\}|} \exp(rd_i)} \tag{9}$$

where rd_i is the prior probability of the i-th relation.

Moreover, in order to address the issues of the decrease in the scores of positive classes and the class-imbalance, we propose the idea of *increasing the contribution of positive classes* to loss. This is different from the previous methods of reducing the proportion of negative classes. To achieve this, we propose two strategies: one is to calculate $P(r)$ only on positive classes, i.e., $r \in \mathcal{P}_T$, and the other is to set the total number of NA label occurrences to 1 before obtaining the relational prior distribution, thereby ignoring the contribution of a large number of NA labels to the loss. Finally, we obtain the loss that fuses the relational prior distribution of entity pairs as follows:

$$\mathcal{L}_3 = -\sum\nolimits_{r \in \mathcal{P}_T} P(r) \tag{10}$$

ARPDL based on Relational Prior Distribution of Entity-Type Pairs

The loss at the entity pair level above considers each entity pair as having the same relational prior distribution for calculation, without considering the potential connection between entity-type pairs and pre-defined relations. For example, the P175 (“performer”) and P569 (“date of birth”) relations will not appear in the entity pairs corresponding to the type ORG → NUM. Therefore, we consider computing the loss at the entity-type pair level, which incorporates the relational prior distribution corresponding to the entity-type pairs.

Dataset	Split	#Doc.	Avg. #Ents	Avg. #Facts
Re-DocRED	Train	3053	19.4	28.1
	Dev	500	19.4	34.6
	Test	500	19.6	34.9
DWIE	Train	602	27.4	23.9
	Dev	98	28.4	26.8
	Test	99	26.5	24.8

Table 1: Dataset statistics.

$$P(r) = \log \frac{\exp(logit_r)}{\sum\limits_{r' \in R \cup \{\text{NA}\}} \exp(logit_{r'})} - \frac{\exp(etd_{r,j})}{\sum\limits_{i=1}^{|R \cup \{\text{NA}\}|} \exp(etd_{i,j})} \tag{11}$$

where $j \in M$, M denotes the total number of entity-type pairs. $etd_{i,j}$ is the prior probability of the i-th relation for the j-th entity-type pair, and the distribution of each entity-type pair is different.

Our final loss is represented as Eq.(12). Here, α serves as a hyperparameter that adjusts the weighting of the relational prior distribution within the overall loss.

$$\mathcal{L}_{ARPDL} = \mathcal{L}_1 + \mathcal{L}_2 + \alpha\mathcal{L}_3 \tag{12}$$

4 Experimental Settings

Datasets. Re-DocRED [Tan *et al.*, 2022b] is a widely-adopted DocRE dataset. It is the scientifically revised version of the original DocRED [Yao *et al.*, 2019] dataset, aiming to address the problem of missing annotations in DocRED. Re-DocRED includes 4053 documents and 96 pre-defined relations. DWIE [Zaporojets *et al.*, 2021] dataset contains 799 documents and 65 relations. In Table 1, “Avg. Ents” and “Avg. Facts” represent the average number of entities and relation facts in each document, respectively.

Implementation Details. All experiments are implemented based on Transformers [Wolf *et al.*, 2020]. We use BERT$_{base}$ [Devlin *et al.*, 2019] and RoBERTa$_{large}$ [Liu *et al.*, 2019] as the encoder of DocRE backbones. We employ AdamW as optimizer with learning rates set to {1e-5, 2e-5, 3e-5, 4e-5, 5e-5} and {6, 8, 10, 20, 30} epochs. The optimal α is 1.2 for BERT encoder and 1.5 for RoBERTa. We use F1 and Ign-F1 metrics for evaluation. Ign-F1 measures F1 while disregarding triples that are present in training set. For each experiment, we run 5 times and report the average score. All experiments are conducted on a GeForce RTX 3090 GPU, and use the loss based on entity-type pair level.

With the exception of Table 2, **all other experiments** are conducted on the Re-DocRED test set, and **for a fair comparison** with other loss methods, we **use ATLOP as the representation module and BERT$_{base}$ as the encoder**.

Model	Dev				Test			
	F1	F1 with A$_L$	Ign-F1	Ign-F1 with A$_L$	F1	F1 with A$_L$	Ign-F1	Ign-F1 with A$_L$
Re-DocRED with RoBERTa$_{large}$								
ATLOP [Zhou *et al.*, 2021] *	77.63	**79.40**(+1.77)	76.88	**78.54**(+1.66)	77.73	**79.90**(+2.17)	76.94	**79.10**(+2.16)
DocuNet [Zhang *et al.*, 2021] *	78.16	**79.70**(+1.54)	77.53	**78.83**(+1.30)	77.92	**79.52**(+1.60)	77.27	**78.69**(+1.42)
KD-DocRE [Tan *et al.*, 2022a] *	78.65	**79.18**(+0.53)	77.92	**78.42**(+0.50)	78.35	**78.95**(+0.60)	77.63	**78.20**(+0.57)
Dreeam [Ma *et al.*, 2023] *	-	81.05	-	80.38	80.73	**81.28**(+0.55)	79.66	**80.66**(+1.00)
AA [Lu *et al.*, 2023] *	81.15	**81.48**(+0.33)	80.04	**80.83**(+0.79)	81.20	**81.66**(+0.46)	80.12	**81.06**(+0.94)
TTM-RE [Gao *et al.*, 2024] ◇	78.13	**80.98**(+2.85)	78.05	**80.27**(+2.22)	79.95	**80.96**(+1.01)	78.20	**80.27**(+2.07)
DWIE with BERT$_{base}$								
ATLOP [Zhou *et al.*, 2021] ‡	64.82	**65.30**(+0.48)	59.03	58.77(-0.26)	69.94	**71.20**(+1.26)	62.09	**63.27**(+1.18)
DocuNet [Zhang *et al.*, 2021] †	67.27	**68.41**(+1.14)	61.03	**61.40**(+0.37)	73.74	**73.98**(+0.24)	66.57	66.07(-0.50)
KD-DocRE [Tan *et al.*, 2022a] †	67.90	**68.47**(+0.57)	61.88	**62.23**(+0.35)	73.46	**73.97**(+0.51)	66.83	**67.03**(+0.20)
Dreeam [Ma *et al.*, 2023] †	67.86	**68.43**(+0.57)	62.00	61.96 (-0.04)	70.10	**71.47**(+1.37)	63.72	**64.89**(+1.17)
AA-NoFusion [Lu *et al.*, 2023] †	67.54	**68.32**(+0.78)	61.82	**62.00**(+0.18)	70.37	**70.90**(+0.53)	64.22	**64.27**(+0.05)
DWIE with RoBERTa$_{large}$								
ATLOP [Zhou *et al.*, 2021] †	76.65	**76.97**(+0.32)	72.47	**72.50**(+0.03)	81.39	**81.45**(+0.06)	76.83	76.51(-0.32)
DocuNet [Zhang *et al.*, 2021] †	76.46	**76.99**(+0.53)	72.69	**72.86**(+0.17)	81.32	**81.47**(+0.15)	77.20	76.75(-0.45)
KD-DocRE [Tan *et al.*, 2022a] †	76.55	**77.77**(+1.22)	72.01	**73.30**(+1.29)	80.92	**81.48**(+0.56)	75.67	**76.51**(+0.84)

Table 2: Performance of different DocRE models with ARPDL replacing their losses. A$_L$ denotes ARPDL. Results of F1 and Ign-F1 marked with † are from our reproduction, ‡ from [Ru *et al.*, 2021], * from [Lu *et al.*, 2023], and ◇ from the original paper.

Loss Function	F1	Ign-F1
ATL [Zhou *et al.*, 2021] *	73.29	72.46
Balanced-Softmax [Zhang *et al.*, 2021] *	73.68	72.85
AML [Wei and Li, 2022] *	72.60	71.78
AFL [Tan *et al.*, 2022a] *	74.15	73.20
NCRL [Zhou and Lee, 2022] †	73.87	72.79
PEMSCL [Guo *et al.*, 2023] †	73.98	73.06
HingeABL$_{SAT}$ [Wang *et al.*, 2023] *	73.46	72.61
HingeABL$_{MeanSAT}$ [Wang *et al.*, 2023] *	74.68	72.90
HingeABL [Wang *et al.*, 2023] *	75.15	73.84
ARPDL	**75.90**	**74.81**

Table 3: Comparison of losses on the Re-DocRED test set. Results with † are from our reproduction, and * from [Wang *et al.*, 2023].

5 Main Results and Analysis

5.1 Different Loss Methods

To compare performance with different losses, we conduct experiments to compare our method with the ATL [Zhou *et al.*, 2021] loss and its extended methods. The results are shown in **Table 3**. To ensure a fair comparison with other ATL-based extended losses, our ARPDL is also implemented based on the ATL [Zhou *et al.*, 2021]. The results show that our ARPDL *outperforms all ATL-based loss methods* and achieves the highest F1 and Ign-F1 of 75.90 and 74.81.

5.2 Different DocRE Models with ARPDL

To evaluate the effectiveness and generality of our loss on different DocRE models, we select several recently competitive DocRE models and *replace their native losses with our ARPDL*. As depicted in **Table 2**, the original methods of these models, ATLOP, Dreeam, and AA employ ATL loss; DocuNet employs Balanced-Softmax loss; KD-DocRE employs AFL loss; and TTM-RE employs a non-ATL positive unlabeled loss [Wang *et al.*, 2022].

Table 2 shows the performance of our loss on different backbones. The results show that using our proposed ARPDL as loss demonstrates significant improvements, particularly on the Re-DocRED dataset. On the ATLOP, using our ARPDL loss shows an improvement of **2.17** in F1 and **2.16** in Ign-F1 on the Re-DocRED test set. Similarly, on the dev set, it boosts the recent TTM-RE by **2.85** F1 and **2.22** Ign-F1. Results show that ARPDL *improves the performance of different baseline models*, demonstrating that our loss is effective and general.

5.3 Analyzing Generalizability of Our Relational Prior Distribution Adapter

To demonstrate the applicability of our proposed relational prior distribution adapter in the ATL-based family loss, we directly add the adapter (i.e., $\mathcal{L}_3$ in Eq.(12)) to different losses. The results are shown in **Table 4**.

After adding our adapter into the ATL loss, the F1 is increased by 2.61 and the Ign-F1 is increased by 2.35; for the AML loss, the F1 is increased by 1.85 and the Ign-F1 is increased by 1.69. The results indicate that *our adapter* is applicable to the ATL-based losses and *effectively improves the performance of these original losses*.

Loss	F1	F1 with RPD	Ign-F1	Ign-F1 with RPD
ATL*	73.29	**75.90**(+2.61)	72.46	**74.81**(+2.35)
AML*	72.60	**74.45**(+1.85)	71.78	**73.47**(+1.69)
AFL*	74.15	**75.30**(+1.15)	73.20	**74.12**(+0.92)
NCRL†	73.87	**74.80**(+0.93)	72.79	**73.81**(+1.02)
PEMSCL†	73.98	**75.10**(+1.12)	73.06	**73.92**(+0.86)
HingeABL$_{SAT}$*	73.46	**74.96**(+1.50)	72.61	**73.88**(+1.27)
HingeABL*	75.15	**75.77**(+0.62)	73.84	**74.42**(+0.58)

Table 4: Integrating relation prior distribution (RPD) adapter into ATL-based losses on Re-DocRED. Results of F1 and Ign-F1 marked with † are from our reproduction, and * from [Wang *et al.*, 2023].

6 Further Analysis

6.1 Analyzing the Decrease of Scores

To assess whether incorporating relational prior distribution into the loss function can address the new issue we discovered: the decrease in predicted scores for positive classes when predicting multiple relations. In **Figure 4**(a), we observe that the scores of both ATL and ARPDL decrease as the number of relations increases. However, the scores of ATL's positive classes gradually fall below the threshold, whereas our ARPDL's are mostly higher than it.

Moreover, since the scores are relative to different thresholds, we cannot directly compare the scores of ARPDL and ATL. Therefore, in order to eliminate the impact of different thresholds, we use relative difference to represent the distance between the score and the threshold, which we call the *relative score*, as shown in Eq.(13).

$$Relative\_Score = \frac{logit_r - TH}{(\frac{logit_r+TH}{2})} \times 100\% \qquad (13)$$

Figure 4(b) shows that the relative score of ARPDL is higher than ATL's. This indicates that the loss *after integrating the relational prior distribution* can make the boundary between the relation scores and the threshold more obvious, which *relatively improves the scores of relations* in positive classes, thus achieving better prediction performance.

6.2 Analysis of Class-Imbalance Problem

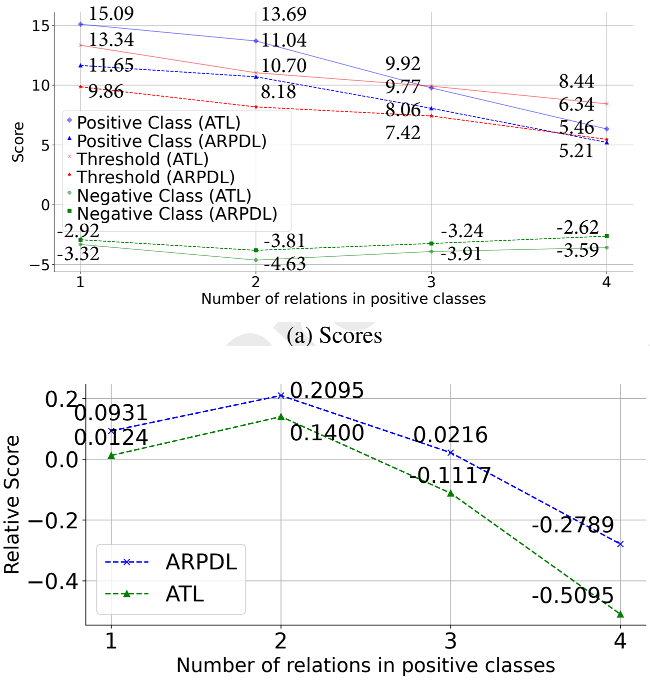
To illustrate the effectiveness of our relational prior distribution adapter in alleviating the false-negative problem caused by class-imbalance, we add the adapter to several different losses and count the number of four prediction patterns.

The results in **Table 5** show that the values of both FN and FN/(FP+FN) decrease after integrating our adapter, indicating that the false-negative problem *has been significantly mitigated*. Also, for the NA problem in false-negative, the values of both FN_NA and FN_NA/(FP+FN) also decrease, indicating that the problem of false-negative samples being predicted with NA labels is mitigated. In addition, we observe an increase in the number of false positive (FP) samples. We speculate that this is due to the adapter increasing the contribution of positive classes to the loss, thereby raising the predicted scores of relations and consequently increasing the number of false positive samples.

6.3 Analysis of Long-Tail Problem

In order to analyze the impact of relational prior distribution on long-tail problems, we first rank in descending order all pre-defined relations by the number of entity pairs that are labeled with them. Next, we classify them into four categories: head-10 relations (the top 10 relations, accounting for 64.02% of Re-DocRED's training data), mid-76 relations (the 10th to 86th relations, accounting for 35.47%), tail-20 relations (the bottom 20 relations, accounting for 1.93%), tail-10 relations (the bottom 10 relations, accounting for 0.51%).

The results in **Table 6** show that the relational prior distribution can *effectively alleviate the long-tail problem*, especially in the Tail-20 and Tail-10, where the improvement is particularly significant. For example, AML's performance



(a) Scores

(b) Relative Scores

Figure 4: Scores in positive classes using ATL and ARPDL. For all sub-graphs, the horizontal axis represents the number of relations in positive classes corresponding to entity pairs, and the vertical axis represents the average predicted score of the relations in positive classes. In Figure 4(a), the scores are the logit output of the model; in Figure 4(b), the scores are relative scores obtained by Eq.(13).

Loss	FP ↓	FN ↓	FN _NA↓	FN _Rel↓	FN_NA /(FP+FN)↓	FN /(FP+FN)↓
ATL	1887	6253	5498	755	67.54	76.82
with RPD	2781	5076	4426	650	56.33	64.60
AML	2032	6363	5515	848	65.69	75.80
with RPD	2272	5699	4919	780	61.71	71.50
AFL	2300	5744	4898	846	60.89	71.41
with RPD	2975	4933	4241	692	53.63	62.38
NCRL	2770	5603	4872	731	58.19	66.92
with RPD	2157	5666	5039	627	64.41	72.43
PEMSCL	2264	5746	4926	820	61.50	71.74
with RPD	3060	4978	4283	695	53.28	61.93
SAT	1749	6241	5363	878	67.12	78.11
with RPD	2647	5389	4604	785	57.29	67.06
HingeABL	2935	5083	4306	777	53.70	63.39
with RPD	3510	4590	3903	687	48.19	56.67

Table 5: Statistics of false prediction patterns before and after integrating the relational prior distribution (RPD) adapter. FP (False-Positive): predicts a negative example as a positive example. FN (False-Negative): predicts a positive example as a negative example. FN_NA: predicts a positive example as a negative example, and the predicted label is NA. FN_Rel: predicts a positive example as a negative example, and the predicted label is a label in negative classes.

Loss	Head-10↑	Mid-76↑	Tail-20↑	Tail-10↑
ATL	77.40	66.63	41.49	38.96
with RPD	79.11	68.54	47.87	41.99
AML	76.68	65.72	43.64	40.00
with RPD	78.52	68.47	51.09	47.40
AFL	78.22	68.33	46.15	42.17
with RPD	79.73	70.06	51.89	48.17
NCRL	77.61	67.72	44.18	37.74
with RPD	78.53	69.64	48.34	41.21
PEMSCL	78.16	68.82	48.46	42.78
with RPD	79.17	70.15	49.82	44.09
SAT	77.48	67.64	48.95	41.21
with RPD	78.79	69.02	49.91	46.93
HingeABL	79.04	69.92	52.56	43.93
with RPD	79.86	70.13	51.86	45.50

Table 6: F1 results on different proportions of long-tail distribution.

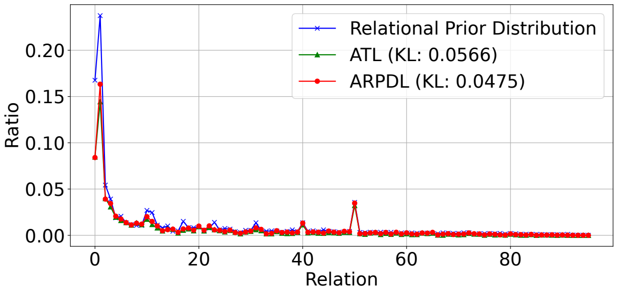


Figure 5: Results of distribution fitting. When calculating the distribution of predicted relations, we first perform softmax on the scores of 97 classes (including TH class) for each entity pair, and then obtain relation scores of all entity pairs. Finally, we average the scores for each class to obtain the distribution of predicted relations (excluding the TH class).

Method on Re-DocRED test set	F1	Ign-F1
ATLOP	73.29	72.46
ATLOP with ARPDL (Our original)	**75.90**	**74.81**
ATLOP with ARPDL (randomly generated distribution)	75.49	74.42

Table 7: Results under different relation prior distributions.

Model	Dev		Test	
	F1	Ign-F1	F1	Ign-F1
on Re-DocRED				
Entity-Type Pairs	75.78	74.65	75.90	74.81
Entity Pairs	75.78	74.65	75.90	74.81
on DWIE				
Entity-Type Pairs	65.28	58.26	71.24	63.27
Entity Pairs	65.55	58.79	71.48	63.46

Table 8: Results of ARPDL with different granularities.

on Tail-20 improves from 43.64% to 51.09%, and on Tail-10 improves from 40.00% to 47.40%. This trend is also significant on other losses, further verifying the effectiveness of integrating relational prior distribution in processing long-tail distribution data.

6.4 Analysis of the Fit Between Relation Prediction Distribution and Relation Prior Distribution

To verify whether the distribution of predicted relations fits the prior distribution of relations, we use Kullback-Leibler (KL) divergence to compare the fitting degree of two distributions. As shown in **Figure 5**, when using ARPDL, the KL divergence between the relation prediction distribution and the relation prior distribution is 4.75%, while the KL divergence when using ATL is 5.66%. This shows that the predicted relations after using ARPDL better fit the prior distribution.

In our experiments, *we cannot ensure that the training and test relational distributions are similar*. Thus, we further assess ARPDL's robustness under distribution shifts by comparing its performance using the train set's prior distribution and a randomly generated distribution. As shown in **Table 7**, ATLOP achieves an F1 score of 73.29, which improves to 75.90 with ARPDL. Even with a random distribution, ARPDL maintains F1 score of 75.49, demonstrating its adaptability to distribution variations. Our analysis suggests that the main reasons for this improvement are the incorporation of prior distribution and the optimization of model performance by ignoring the NA label and increasing the contribution of positive classes to loss.

6.5 Entity Pairs *vs* Entity-Type Pairs

To compare the performance of our loss ARPDL based on different granularities of entity pairs and entity-type pairs, we conduct experiments on Re-DocRED and DWIE datasets as shown in **Table 8**. Experimental results show that both granularities perform equally well on the Re-DocRED, with only slight differences observed on DWIE. We attribute this to the following main factors: first, the relational prior distribution of entity pairs in each of the two datasets is highly consistent with the distribution of entity-type pairs; second, the negative classes contain more relations than the positive classes, which are relatively sparse. Our prior distribution loss focuses exclusively on the positive classes (Eq.(9)-(11)), which may lead to negligible differences in the outcomes between the two granularities. This results in Table 8 also further implicitly demonstrate the stability of our relational prior distribution loss.

7 Conclusion

We propose ARPDL, a new multi-label classification loss that can guide the DocRE models to adaptively adjust relation prediction scores using prior distribution of relations. Moreover, we propose a novel strategy in ARPDL to address the class-imbalance and long-tail problems by increasing the contribution of positive classes to loss. In addition, our designed relational prior distribution component can also be applied as a plug-in adapter to the other multi-label threshold losses to further improve their performance on DocRE. Experiments show that ARPDL consistently improves DocRE models, achieving SOTA results and enhancing performance when integrated into other losses, validating our approach's effectiveness and generality. Since our method is independent of a specific model, it has the potential to be widely used in other multi-label classification scenarios.

Acknowledgements

The authors sincerely thank the reviewers for their valuable comments, which improved the paper. The work is supported by the National Natural Science Foundation of China (62276057), and Sponsored by CAAI-MindSpore Open Fund, developed on OpenI Community.

References

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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 (NAACL-HLT)*, pages 4171–4186, 2019.

[Gao *et al.*, 2024] Chufan Gao, Xuan Wang, and Jimeng Sun. TTM-RE: memory-augmented document-level relation extraction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

[Guo *et al.*, 2023] Jia Guo, Stanley Kok, and Lidong Bing. Towards integration of discriminability and robustness for document-level relation extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 2598–2609, 2023.

[Jain *et al.*, 2024] Monika Jain, Raghava Mutharaju, Ramakanth Kavuluru, and Kuldeep Singh. Revisiting document-level relation extraction with context-guided link prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pages 18327–18335, 2024.

[Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[Lu *et al.*, 2023] Chonggang Lu, Richong Zhang, Kai Sun, Jaein Kim, Cunwang Zhang, and Yongyi Mao. Anaphor assisted document-level relation extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 15453–15464, 2023.

[Ma *et al.*, 2023] Youmi Ma, An Wang, and Naoaki Okazaki. Dreeam: Guiding attention with evidence for improving document-level relation extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 1971–1983, 2023.

[Peng *et al.*, 2022] Xingyu Peng, Chong Zhang, and Ke Xu. Document-level relation extraction via subgraph reasoning. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4331–4337, 2022.

[Ru *et al.*, 2021] Dongyu Ru, Changzhi Sun, Jiangtao Feng, Lin Qiu, Hao Zhou, Weinan Zhang, Yong Yu, and Lei Li. Learning logic rules for document-level relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1239–1250, 2021.

[Sun *et al.*, 2023] Qi Sun, Kun Huang, Xiaocui Yang, Pengfei Hong, Kun Zhang, and Soujanya Poria. Uncertainty guided label denoising for document-level distant relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL)*, pages 15960–15973, 2023.

[Tan *et al.*, 2022a] Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. Document-level relation extraction with adaptive focal loss and knowledge distillation. In *Findings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1672–1681, 2022.

[Tan *et al.*, 2022b] Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. Revisiting docred-addressing the false negative problem in relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8472–8487, 2022.

[Wang *et al.*, 2022] Ye Wang, Xinxin Liu, Wenxin Hu, and Tao Zhang. A unified positive-unlabeled learning framework for document-level relation extraction with different levels of labeling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

[Wang *et al.*, 2023] Jize Wang, Xinyi Le, Xiaodi Peng, and Cailian Chen. Adaptive hinge balance loss for document-level relation extraction. In *Findings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3872–3878, 2023.

[Wei and Li, 2022] Ying Wei and Qi Li. Sagdre: Sequence-aware graph-based document-level relation extraction with adaptive margin loss. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 2000–2008, 2022.

[Wolf *et al.*, 2020] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP)*, pages 38–45, 2020.

[Xiao *et al.*, 2022] Yuxin Xiao, Zecheng Zhang, Yuning Mao, Carl Yang, and Jiawei Han. Sais: Supervising and augmenting intermediate steps for document-level relation extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 2395–2409, 2022.

[Xie *et al.*, 2022] Yiqing Xie, Jiaming Shen, Sha Li, Yuning Mao, and Jiawei Han. Eider: Empowering document-level relation extraction with efficient evidence extraction and inference-stage fusion. In *Findings of ACL*, pages 257–268, 2022.

[Xu *et al.*, 2021] Wang Xu, Kehai Chen, and Tiejun Zhao. Document-level relation extraction with reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pages 14167–14175, 2021.

[Yao *et al.*, 2019] Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. Docred: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 764–777, 2019.

[Zaporojets *et al.*, 2021] Klim Zaporojets, Johannes Deleu, Chris Develder, and Thomas Demeester. DWIE: an entity-centric dataset for multi-task document-level information extraction. *Inf. Process. Manag.*, 58(4):102563, 2021.

[Zhang *et al.*, 2021] Ningyu Zhang, Xiang Chen, Xin Xie, Shumin Deng, Chuanqi Tan, Mosha Chen, Fei Huang, Luo Si, and Huajun Chen. Document-level relation extraction as semantic segmentation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3999–4006, 2021.

[Zhang *et al.*, 2024] Fu Zhang, Qi Miao, Jingwei Cheng, Hongsen Yu, Yi Yan, Xin Li, and Yongxue Wu. SRF: enhancing document-level relation extraction with a novel secondary reasoning framework. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 15426–15439, 2024.

[Zhou and Lee, 2022] Yang Zhou and Wee Sun Lee. None class ranking loss for document-level relation extraction. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4538–4544, 2022.

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