None Class Ranking Loss for Document-Level Relation Extraction

这种正-负不平衡使得多标签损失更多地关注负样本，并降低了罕见阳性样本的贡献

Yang Zhou[1] , **Wee Sun Lee**[1]
[1]School of Computing, National University of Singapore
{zhouy, leews}@comp.nus.edu.sg

Abstract

Document-level relation extraction (RE) aims at extracting relations among entities expressed across multiple sentences, which can be viewed as a multi-label classification problem. In a typical document, most entity pairs do not express any pre-defined relation and are labeled as "none" or "no relation". For good document-level RE performance, it is crucial to distinguish such *none* class instances (entity pairs) from those of pre-defined classes (relations). However, most existing methods only estimate the probability of pre-defined relations independently without considering the probability of "no relation". This ignores the context of entity pairs and the label correlations between the none class and pre-defined classes, leading to sub-optimal predictions. To address this problem, we propose a new multi-label loss that encourages large *margins* of label confidence scores between each pre-defined class and the none class, which enables captured label correlations and context-dependent thresholding for label prediction. To gain further robustness against positive-negative imbalance and mislabeled data that could appear in real-world RE datasets, we propose a margin regularization and a margin shifting technique. Experimental results demonstrate that our method significantly outperforms existing multi-label losses for document-level RE and works well in other multi-label tasks such as emotion classification when none class instances are available for training.

为了获得良好的文档级RE性能，将这些无类实例（实体对）与那些预定义的类（关系）实例区分开来是至关重要的

然而，大多数现有的方法只独立估计预定义关系的概率，而没有考虑"无关系"的概率

这忽略了实体对的上下文以及无类和预定义类之间的标签相关性，导致了次优预测

1 Introduction

Identifying relationships among entities from unstructured text, namely relation extraction (RE), is a fundamental task in information extraction. Previous works have studied sentence-level RE that aims at extracting relations between two entities in a single sentence [Miwa and Bansal, 2016; Zhang *et al.*, 2018]. More recently, document-level RE, which aims to identify the relations of various entity pairs expressed across *multiple sentences*, has received increasing research attention [Yao *et al.*, 2019; Yu *et al.*, 2020].

Since one entity pair can express multiple relations in a document, document-level RE can be formulated as a *multi-label* classification problem. Different from conventional multi-label settings where each instance at least has one positive pre-defined label, most instances (entity pairs) in document-level RE express *none* of the pre-defined relations, where all the pre-defined labels are *known to be negative*. For good document-level RE performance, it is crucial to distinguish such *none* class instances (entity pairs) from those of pre-defined classes (relations).

However, most existing methods adopt binary cross entropy (BCE) loss for document-level RE. They only estimate the probability of pre-defined relations *independently* without considering the probability of "no relation". This ignores the fact that the probabilities of both the *none* class and pre-defined classes are important for multi-label prediction, as they are highly correlated (the increased probability of the none class implies the decreased probabilities of all the pre-defined classes). Without knowing the none class probability, whether an entity pair has certain relations needs to be determined via global thresholding (i.e., selecting labels whose confidence scores are higher than a single threshold as final predictions). This fails to take the context of each entity pair into account and may produce either too many or too few labels, leading to sub-optimal predictions.

它们只独立地估计预定义关系的概率，而不考虑"无关系"的概率

In this work, we propose a new multi-label performance measure and its surrogate loss, called None Class Ranking Loss (NCRL), to address the aforementioned disadvantages of BCE in document-level relation extraction. NCRL aims to maximize the *margin* of label confidence scores between the none class and each pre-defined class, such that positive labels are ranked above *the none class label* and negative labels are ranked below. In this way, the none class score becomes an instance-dependent threshold that treats labels being ranked above the none class as positive, and the label margin reflects the uncertainty of such prediction. By formulating the probability of each class as label margins, NCRL not only takes the context of entity pairs in multi-label prediction but also (as will be shown later) enables captured label correlations without directly comparing the scores of positive and negative labels. Moreover, we theoretically justify the effectiveness of NCRL by proving that NCRL satisfies *Bayes consistency* w.r.t. the targeted multi-label performance measure, which is a desirable property for any surrogate loss in

最大化无类和每个预定义类之间的标签置信度分数的边际

正标签排在无类标签上面，负标签排在下面。

NCRL不仅在多标签预测中考虑了实体对的上下文，而且能够捕获标签之间的相关性，而无需直接比较正负标签的得分

multi-label learning.

In a document, entity pairs generally have a few positive relations and many negative ones. This positive-negative imbalance makes multi-label losses focus more on negative samples and down-weight the contributions from the rare positive samples [Ben-Baruch *et al.*, 2021]. Another practical problem of document-level RE is that entity pairs are prone to being mislabeled as "no relation", which results in deteriorated classification performance. To alleviate the positive-negative imbalance problem, we propose a margin regularization technique, which penalizes negative samples whose confidence scores are far below that of the none class. For robustness against mislabeled samples, we propose a margin shifting approach to attenuate the impact of negative samples with a very large score, which are suspected as mislabeled. This helps to reduce the label noise in document-level RE datasets.

为了缓解正负样本不平衡的问题，我们提出了一种边界正则化技术，它惩罚那些负样本的置信度远低于“ none ”类的样本。

为了对错误标记样本的稳健性，我们提出了一种边际转移方法来衰减得分非常大的阴性样本的影响，这被怀疑是错误标记。

Our NCRL loss is general enough and can be applied to other multi-label tasks when none class instances are available for training. Experimental results on two document-level RE datasets, DocRED [Yao *et al.*, 2019] and DialogRE [Yu *et al.*, 2020], and one fine-grained emotion classification dataset, GoEmotions [Demszky *et al.*, 2020], show that NCRL achieves significantly better results than existing multi-label losses. The main contributions of this paper are summarized as follows:

- We propose a new multi-label performance measure along with its surrogate loss, NCRL, that prefers large label margins between the none class and each pre-defined class. Our loss captures the label correlations and enables context-dependent thresholds for multi-label prediction.
- We prove that NCRL satisfies multi-label consistency, which guarantees that minimizing NCRL can achieve the expected objective.
- We propose a margin regularization and a margin shifting technique to alleviate positive-negative sample imbalance and gain further robustness against mislabeled samples, respectively.

2 Related Work

Multi-label learning. Multi-label classification is a well-studied problem with rich literature. Here, we focus on loss functions for multi-label learning. BCE is the most popular multi-label loss, which reduces the multi-label problem into a number of independent binary (one-vs-all) classification tasks, one for each label. Another common multi-label loss function is pairwise ranking loss, which transforms multi-label learning into a ranking problem via pairwise (one-vs-one) comparison. Many attempts have been made to improve these two types of multi-label loss functions [Fürnkranz *et al.*, 2008; Yeh *et al.*, 2017; Li *et al.*, 2017; Durand *et al.*, 2019; Xu *et al.*, 2019; Wu *et al.*, 2019; Ben-Baruch *et al.*, 2021]. However, as far as we know, none of the existing losses as well as theoretical studies consider none class instances, which commonly appear in real-world multi-label applications.

Document-level relation extraction. Existing works on document-level RE mainly focus on learning better entity pair representations to improve classification performance. Some representative approaches include incorporating long-distance entity interactions into recurrent neural networks or transformer-based models [Christopoulou *et al.*, 2019; Nan *et al.*, 2020], aggregating local and global entity information with graph neural networks [Wang *et al.*, 2020; Zeng *et al.*, 2020; Xu *et al.*, 2021b; Dai *et al.*, 2021], and integrating latent word semantics with pre-trained language models [Zhou *et al.*, 2021; Xu *et al.*, 2021a].

All the aforementioned document-level RE methods, except [Zhou *et al.*, 2021], use BCE as the objective function for training and do not consider the none class. ATL [Zhou *et al.*, 2021] is probably the most related method to our NCRL loss function. Both ATL and NCRL treat the none class label as an adaptive threshold to separate positive relations from negative ones, leading to context-dependent multi-label prediction. However, the proposed NCRL can learn a more discriminative classifier with theoretical guarantees and consistently outperforms ATL in our experiments. For clarity, we defer the detailed comparisons between NCRL and ATL in Sec. 3.5.

同：
ATL和NCR都将无类标签视为一个自适应阈值，以区分正关系和负关系

异：
NCRL可以学习一个具有理论保证的更具有鉴别性的分类器，并且在我们的实验中始终优于ATL

3 None Class Ranking Loss

In this section, we first introduce None Class Ranking Error (NCRE), a new multi-label performance measure that takes into account the none class in multi-label learning. Then, we propose None Class Ranking Loss (NCRL), which is a surrogate loss of NCRE with Bayes consistency. Finally, we propose a margin regularization and a margin shifting strategy to address the positive-negative imbalance and mislabeling problems that commonly appear in real-world RE datasets.

3.1 Problem Statement

Document-level RE is essentially a multi-label classification problem, where each entity pair is an instance and the associated relations are label samples. Let $\mathcal{X}$ be an instance space and $\mathcal{Y} = \{0, 1\}^K$ be a label space, where K is the number of pre-defined classes. An instance $\boldsymbol{x} \in \mathcal{X}$ is associated with a subset of labels, identified by a binary vector $\boldsymbol{y} \in \mathcal{Y} = (y_1, \ldots, y_K)$, where $y_i = 1$ if the i-th label is positive for $\boldsymbol{x}$, and $y_i = 0$ otherwise. In particular, none class instances can be represented by setting all the labels to be negative ($y_i = 0, i = 1, \ldots, K$). To simplify the notation and facilitate deriving the proposed NCRL, we introduce the none class and assign it with the 0-th label y_0 without loss of generality. Then none class instances can be simply labeled with $y_0 = 1$.

3.2 Performance Measure

Different from document-level RE, conventional multi-label datasets often assume that each instance at least has one positive label, which means that none class instances are not available for training. For this reason, existing multi-label performance measures only focus on the predictions of pre-defined labels individually, and do not consider how well the none class is predicted. Nevertheless, the probabilities of the pre-defined classes are highly correlated with the probability of the none class. Failing to consider such label correlations

现有方法没有考虑无类的预测程度

in the performance measure could overlook bad predictions that should have been penalized during training, and thus increases the risk of misclassification, especially for document-level RE, where most entity pairs are from the none class.

To better measure the performance of document-level RE, we transform the multi-label prediction of each instance into a label ranking, where positive pre-defined labels should be ranked higher than the none class label, while negative ones should be ranked below. Formally, we propose NCRE, a new multi-label performance measure as follows:

None Class Ranking Error (NCRE)

$$\ell_{\mathrm{NA}}(\boldsymbol{y},\boldsymbol{f}) = \sum_{i=1}^{K}\Big([\![y_i>0]\!][\![f_i<f_0]\!] + [\![y_i\le 0]\!][\![f_i>f_0]\!] + \frac{1}{2}[\![f_i=f_0]\!]\Big), \quad (1)$$

where f_i is the confidence score of the i-th class, and $[\![\pi]\!]$ is the mapping that returns 1 if predicate π is true and 0 otherwise. NCRE treats the none class label as an adaptive threshold that separates positive labels from negative ones, and counts the number of reversely ordered label pairs between the none class and each pre-defined class, which considers errors in ranking the none class and pre-defined labels simultaneously.

3.3 NCRL Surrogate Loss

NCRE很难被最小化，所以我们最小化NCRE的替代损失NCRL

Although our goal is to find an optimal classifier w.r.t. NCRE, it is difficult to minimize NCRE directly. For practical multi-label learning, we need to minimize a surrogate loss of NCRE instead, which should be simple and easy for optimization. To this end, we propose None Class Ranking Loss (NCRL) as follows:

None Class Ranking Loss (NCRL)

$$L_{\mathrm{NA}}(\boldsymbol{y},\boldsymbol{f}) = -\sum_{i=1}^{K}\Big(y_i\cdot\log\sigma(m_i^+) + (1-y_i)\cdot\log\sigma(m_i^-)\Big), \quad (2)$$

where $m_i^+ = f_i - f_0$ and $m_i^- = f_0 - f_i$ represent *positive and negative label margins*, respectively, and $\sigma(x)$ is the sigmoid function that transforms label margins into probability outputs within $[0,1]$. Intuitively, minimizing NCRL encourages positive (negative) labels to have higher (lower) scores than that of the none class as much as possible. At the test time, we can return labels with higher scores than f_0 as predicted labels or return NA if such labels do not exist.

Captured correlations. BCE assumes pre-defined classes are independent and thus ignores label correlations and dependencies. Pairwise ranking loss is commonly used to capture label correlations in multi-label learning, which aims to maximize the margin between each pair of positive and negative labels. Compared with BCE and pairwise ranking loss, NCRL relaxes the independent assumption of BCE and captures label correlations by connecting the probability of each pre-defined class $\mathbb{P}(y_i = 1|\boldsymbol{x})$ to the label margins $m_i^+ = f_i - f_0$. In this way, maximizing $\mathbb{P}(y_i = 1|\boldsymbol{x})$ not only leads to increased f_i but also decreased f_0, which in turn affects the margins (probabilities) of all the other classes.

Moreover, thanks to the introduced none class, the margin between any pair of pre-defined labels can be induced from

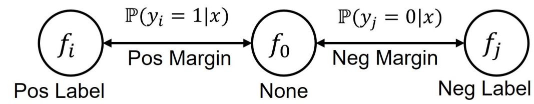

Figure 1: An illustration of NCRL. For $y_i = 1$, NCRL maximizes the positive margin $f_i - f_0 \propto \mathbb{P}(y_i = 1|\boldsymbol{x})$ (see Theorem 1). For $y_j = 0$, NCRL maximizes the negative margin $f_0 - f_j \propto p(y_j = 0|\boldsymbol{x})$. In this way, the margin between positive and negative labels $f_i - f_j = (f_i - f_0) + (f_0 - f_j)$ is also maximized.

the sum of m_i^+ and m_j^-, i.e., $f_i - f_j = (f_i - f_0) + (f_0 - f_j)$. Therefore, NCRL captures pairwise label correlations and *indirectly* maximizes the margin between every pair of positive and negative labels without expensive pairwise comparisons. Figure 1 illustrates how NCRL learns a proper label ranking by separating positive and negative labels via f_0.

NCRL捕获了成对的标签相关性，并间接地最大化了每对正标签和负标签之间的边际，而不进行昂贵的成对比较

Consistency analysis. Now we theoretically justify the effectiveness of NCRL by proving that NCRL (2) is *Bayes consistent* w.r.t. NCRE (1). Bayes consistency is an important property for a surrogate loss. It guarantees that the classifier obtained by minimizing the surrogate loss converges to the one with minimal multi-label risk, and thus is likely to achieve good classification performance w.r.t. the corresponding multi-label performance measure.

Given an instance $\boldsymbol{x}$, let $\Delta_i = \mathbb{P}(y_i = 1|\boldsymbol{x})$ be the marginal probability when the i-th label is positive. Then the conditional risk w.r.t. NCRE (1) can be written as follows:

$$R_{\ell_{\mathrm{NA}}}(\mathbb{P},\boldsymbol{f}) = \mathbb{E}[\ell_{\mathrm{NA}}(\mathbb{P},\boldsymbol{f})|\boldsymbol{x}] = \sum_{i=1}^{K}\Big(\Delta_i[\![f_i<f_0]\!] + (1-\Delta_i)[\![f_i>f_0]\!] + \frac{1}{2}[\![f_i=f_0]\!]\Big), \quad (3)$$

which corresponds to the *expected penalty* for an instance $\boldsymbol{x}$ when $\boldsymbol{f}(\boldsymbol{x})$ is used as the score function. From (3), the Bayes optimal score function $\boldsymbol{f}^*_{\ell_{\mathrm{NA}}}$ that minimizes the multi-label risk (3) is given by:

$$\boldsymbol{f}^*_{\ell_{\mathrm{NA}}} \in \{\boldsymbol{f} : f_i > f_0 \text{ if } \Delta_i > \frac{1}{2}, \text{ and } f_i < f_0 \text{ if } \Delta_i < \frac{1}{2}\}. \quad (4)$$

Theorem 1. *NCRL (2) is Bayes consistent w.r.t. NCRE (1).*

Proof sketch. We prove the Bayes consistency of NCRL by showing that given an instance, the optimal label margin $f_i^* - f_0^*$ obtained by minimizing NCRL *coincides with* the marginal label probability $\Delta_i = \mathbb{P}(y_i = 1|\boldsymbol{x})$, such that $f_i^* - f_0^* > 0 \Rightarrow \Delta_i > \frac{1}{2}$, which is an optimal solution with the minimal NCRE risk according to (4). The complete proof is provided in Appendix. □

3.4 Margin Regularization and Shifting

Margin regularization. In real-world document-level RE datasets, it is common that only a small number of entity pairs express pre-defined relations, implying that the number of positive samples per relation is much larger than that of negative ones. Such positive-negative sample imbalance makes NCRL push a large number of negative samples away from the threshold f_0 to achieve a low training loss. As a

Margin regularization —— label imbalance
这种正负样本不平衡使得NCRL将大量负样本远离阈值f0，以达到较低的训练损失，使用不平衡数据进行训练，往往会对具有罕见阳性样本的预定义标签做出负面预测。

result, training with imbalanced data tends to make negative predictions for pre-defined labels with rare positive samples.

To address this problem, we propose to control the average margin and design the following regularization term:

$$L_{\mathrm{NA}_0}(\boldsymbol{y}, \boldsymbol{f}) = -y_0 \cdot \log \sigma(m_0^+) - (1 - y_0) \cdot \log \sigma(m_0^-) \quad (5)$$

where $m_0^+ = f_0 - \frac{1}{K}\sum_{i=1}^{K} f_i$ and $m_0^- = \frac{1}{K}\sum_{i=1}^{K} f_i - f_0$ are the average positive and negative margins, respectively. For instances with $y_0 = 0$, the margin regularization $L_{\mathrm{NA}_0}(\boldsymbol{y}, \boldsymbol{f})$ penalizes the average score over all pre-defined labels if it is smaller than f_0, and thus avoids over-suppressing scores of negative samples. On the other hand, for none class instances with $y_0 = 1$, the threshold f_0 should be larger than the largest score among all the pre-defined labels. However, directly maximizing the margin $f_0 - \max_{i \neq 0} f_i$ can lead to unstable results, as the pre-defined label with $\max_{i \neq 0} f_i$ would keep changing during training. Instead, $L_{\mathrm{NA}_0}(\boldsymbol{y}, \boldsymbol{f})$ maximizes the average margin $f_0 - \frac{1}{K}\sum_{i=1}^{K} f_i$, which is an upper-bound of $f_0 - \max_{i \neq 0} f_i$, to encourage f_0 being at the top of the label ranking.

y0=0时，L会避免过度抑制负样本分数

y0=1时，阈值f0应该大于所有预定义标签中的最大分数

Margin shifting. Another practical issue in document-level RE is the mislabeling problem. As a document can describe various contents, it is very difficult for a human annotator to annotate all relations accurately and extensively for each entity pair. As a result, real-world document-level RE datasets often contain considerable noise in labeling. Among mislabeled data, false negative samples are particularly common in practice.

To gain more robustness against mislabeled samples, we propose to shift the transformed negative margin $\sigma(m_i^-)$ in NCRL (2) as follows:

$$p_i^- = \max(\sigma(m_i^-) + \gamma, 1), i = 0, 1, \ldots, K, \quad (6)$$

where $0 < \gamma < 1$ is a hyper-parameter. Such margin shifting approach increases the probability of all negative samples by a factor γ such that the impact of very hard negative samples with small $\sigma(m_i^-) = \mathbb{P}(y_i = 0|\boldsymbol{x})$, which are suspected as mislabeled, are attenuated during training. In addition, very easy samples with $\sigma(m_i^-) > 1 - \gamma$ get a zero loss and are ignored, which makes NCRL focus more on harder samples.

hard negative

σ(m~i_)<1-γ
目的:硬负样本影响变小,
m~i=f0-fi

Although similar shifting approaches have been proposed [Lin *et al.*, 2017; Ben-Baruch *et al.*, 2021; Wu *et al.*, 2020], they do not consider the none class label and shift the *score* f_i rather than the *margin* m_i^-. Since the none class score f_0 *aligns* the margins m_i^- of pre-defined labels in the same range, margin shifting is more suitable than score shifting in adjusting the prediction probability for different training samples. Armed with the margin regularization and shifting, our final loss is given as follows:

$$L_{\mathrm{NA}}(\boldsymbol{y}, \boldsymbol{f}) = -\sum_{i=0}^{K} \Big(y_i \cdot \log \sigma(m_i^+) + (1 - y_i) \cdot \log p_i^-\Big). \quad (7)$$

3.5 Comparison with ATL

There are a few related works that consider the none class in document-level RE. To the best of our knowledge, adaptive thresholding loss (ATL) [Zhou *et al.*, 2021] is the most related method to our NCRL. ATL also treats the none label class as a learnable threshold and encourages positive labels to be ranked above the none class label and negative labels to be ranked below for document-level RE. Specifically, ATL is given as follows:

$$L_{ATL}(\boldsymbol{y}, \boldsymbol{f}) = -\sum_{i: y_i = 1} \log \left(\frac{\exp(f_i)}{\exp(f_0) + \sum_{j: y_j = 1} \exp(f_j)} \right) - \log \left(\frac{\exp(f_0)}{\exp(f_0) + \sum_{j: y_j = 0} \exp(f_j)} \right),$$

where $i, j = 1, \ldots, K$.

Although ATL and NCRL share the same spirit of learning the none class label, they essentially learn distinct label rankings. ATL is constructed in two terms. The first term prefers the none class score f_0 being *ranked at the bottom* among f_0 and the scores of positive pre-defined labels, and the second term prefers f_0 being *ranked at the top* among f_0 and the scores of negative pre-defined labels. ATL only focuses on positioning f_0 between positive and negative sets of labels, whereas fails to maximize the margin between each pair of positive and negative labels, which is desirable for obtaining discriminative classifiers. Due to the softmax function in the first term of ATL, increasing the score of one positive labels f_i will penalize the scores of all the other positive labels, and could leads to large loss values, which is not reasonable in the multi-label setting. On the contrary, our NCRL separates positive labels from negative ones as much as possible by maximizing the label margins, and enjoys the theoretical guarantee of Bayes consistency. In addition, NCRL is equipped with the margin regularization and shifting to alleviate the positive-negative imbalance and mislabeling problems, which are not applicable for ATL.

ATL未能最大化每对正标签和负标签之间的边界，这是获得鉴别分类器的理想条件

ATL softmax增加一个积极标签分数fi会惩罚其他积极标签的分数

4 Experiments

In this section, we evaluate NCRL on two document-level RE datasets. To demonstrate the effectiveness of NCRL for other multi-label tasks, we also test NCRL for fine-grained emotion classification.

4.1 Experimental Setups

Datasets. *DocRED* [Yao *et al.*, 2019] is a large-scale document-level RE dataset, which is constructed from Wikipedia articles. DocRED consists of 5053 documents with 96 pre-defined relations, where 3053 documents are used for training, 1000 for development, and 1000 for testing. DocRED also provides a large distantly supervised dataset, which includes 101,873 documents with noisy labels. *DialogRE* [Yu *et al.*, 2020] is a dialogue-based RE dataset, which is constructed from the transcripts of the American TV situation comedy, Friends. DialogRE consists of 1788 dialogues with 36 pre-defined relations, where 60% of dialogues are used for training, 20% for development, and 20% for testing. *GoEmotions* [Demszky *et al.*, 2020] is an emotion classification dataset of 58009 Reddit comments labeled for 27 emotion categories or Neutral, where 80% data are used for training, 10% for development, and 10% for testing.

Model	Dev		Test	
	Ign F_1	F_1	Ign F_1	F_1
HIN-BERT$_{\text{Base}}$ [Tang *et al.*, 2020]	54.29	56.31	53.70	55.60
CFER-BERT$_{\text{Base}}$ [Dai *et al.*, 2021]	58.00	60.06	57.89	59.82
Coref-RoBERTa$_{\text{Large}}$ [Ye *et al.*, 2020]	57.35	59.43	57.90	60.25
SSAN-RoBERTa$_{\text{Large}}$ [Xu *et al.*, 2021a]	60.25	62.08	59.47	61.42
BCE + ATLOP-RoBERTa$_{\text{Large}}$	61.02 ± 0.21	63.19 ± 0.20	61.58	63.46
ATL + ATLOP-RoBERTa$_{\text{Large}}$ [Zhou *et al.*, 2021]	61.32 ± 0.14	63.18 ± 0.19	61.39	63.40
NCRL + ATLOP-RoBERTa$_{\text{Large}}$	62.21 ± 0.22	64.18 ± 0.20	61.94	64.14
BCE + ATLOP-DeBERTa$_{\text{Large}}$	61.92 ± 0.13	63.96 ± 0.15	61.83	63.92
ATL + ATLOP-DeBERTa$_{\text{Large}}$	62.16 ± 0.15	64.01 ± 0.12	62.12	64.08
NCRL + ATLOP-DeBERTa$_{\text{Large}}$	$\mathbf{62.98 \pm 0.18}$	$\mathbf{64.79 \pm 0.13}$	**63.03**	**64.96**
With pre-training on the distantly supervised dataset				
BCE + ATLOP-DeBERTa$_{\text{Large}}$	65.05 ± 0.24	66.71 ± 0.22	64.91	66.69
ATL + ATLOP-DeBERTa$_{\text{Large}}$	64.34 ± 0.12	66.18 ± 0.15	63.95	65.90
NCRL + ATLOP-DeBERTa$_{\text{Large}}$	$\mathbf{66.11 \pm 0.14}$	$\mathbf{67.92 \pm 0.14}$	**65.81**	**67.53**

Table 1: Classification results (%) on the DocRED dataset.

Implementation details. For each dataset, we adopt an encoding model upon pre-trained language models for representation learning. The classification results are then obtained by training the encoding model with different loss functions. We use Huggingface's Transformers [Wolf *et al.*, 2020] to implement all the models. We use AdamW [Loshchilov and Hutter, 2019] as the optimizer with learning rates $\in \{1\mathrm{e}{-5}, 2\mathrm{e}{-5}, \ldots 5\mathrm{e}{-5}\}$, and apply a linear warmup [Goyal *et al.*, 2017] at the first 10% steps followed by a linear decay to 0. The number of training epochs is selected from $\{5, 8, 10, 20, 30\}$. At the test time, the best checkpoint on the development set is used for the final prediction. For BCE, we follow the same settings with ATL [Zhou *et al.*, 2021] to use a global threshold for multi-label prediction and select the threshold from $\{0.1, 0.2, ..., 0.9\}$. For NCRL, the hyper-parameter γ in margin shifting (6) are selected from $\{0, 0.01, 0.05\}$. All hyper-parameters are tuned on the development set, where those with the highest F_1 scores are selected for testing. All the experiments are conducted with 1 GeForce RTX 3090 GPU. We provide detailed configurations of the encoding models and the hyper-parameter settings on each dataset in Appendix. Our code is available at https://github.com/yangzhou12/NCRL.

Competing methods. We compare our NCRL with existing multi-label loss functions including BCE and ATL [Zhou *et al.*, 2021]. We also test pairwise ranking loss and the log-sum-exp pairwise loss [Li *et al.*, 2017], but do not report their results for simplicity, as they are consistently worse than BCE in our experiments. For each method, we report the mean and standard deviation of performance measures based on 5 runs with different random seeds. The results with citations are taken from the corresponding papers.

4.2 Results and Discussions

Results on DocRED. On DocRED, we follow the same setting of ATL and use its encoding model called ATLOP [Zhou *et al.*, 2021] for fair comparisons, which is built on RoBERTa-large [Liu *et al.*, 2019]. To demonstrate our NCRL can still improve the performance on stronger encoding models, we also test a more recent pre-trained language model, DeBERTa-large [He *et al.*, 2021]. In addition, we consider an improved RE model, which is first trained on the distantly supervised DocRED dataset and then finetuned on the human-annotated dataset. For comprehensiveness, we also compare recent document-level RE methods including HIN [Tang *et al.*, 2020], Coref [Ye *et al.*, 2020], CFER [Dai *et al.*, 2021], and SSAN [Xu *et al.*, 2021a], which use BCE for training. Following [Yao *et al.*, 2019], we use F_1 and Ign F_1 in evaluation, where the Ign F_1 denotes the F_1 score excluding the relational facts that already appear in the training and dev/test sets.

Table 1 shows the document-level RE results on DocRED, where the test scores are obtained by submitting the predictions of the best checkpoint on the development set to the official Codalab evaluation system. Overall, BCE with a manually tuned threshold already achieves comparable or better results than most competing methods such as CFER and SSAN. By considering the none class in multi-label learning, ATL and our NCRL outperform BCE in most cases, and avoid the overhead of threshold tuning. This indicates that exploiting the correlations of the none class label can improve the performance of document-level RE. Moreover, our NCRL consistently outperforms ATL in all cases. This demonstrates the advantage of NCRL in learning a label ranking with maximized label margins. When the distantly supervised data are used for training, all the methods obtain much better results. In particular, NCRL achieves a state-of-the-art Ign F_1 score of 65.81% on the leaderboard[1]. This indicates that NCRL is robust against noisy labels and can better transfer the information from the distantly supervised dataset for document-level RE. On the other hand, ATL seems to be sensitive to label noise and gets much worse results than NCRL.

Results on DialogRE. On DialogRE, we adopt the official baseline, BERTs$_{\text{Base}}$ [Yu *et al.*, 2020], as the encoding model, where the dialogue inputs are augmented with entity type and coreference information. We follow [Yu *et al.*,

[1]https://competitions.codalab.org/competitions/20717

Model	Dev		Test	
	F_1	F_{1c}	F_1	F_{1c}
NN [Yu *et al.*, 2020]	46.1 ± 0.7	43.7 ± 0.5	48.0 ± 1.5	45.0 ± 1.4
LSTM [Yu *et al.*, 2020]	46.7 ± 1.1	44.2 ± 0.8	47.4 ± 0.6	44.9 ± 0.7
BiLSTM [Yu *et al.*, 2020]	48.1 ± 1.0	44.3 ± 1.3	48.6 ± 1.0	45.0 ± 1.3
$\text{BERTs}_{\text{Base}}$ [Yu *et al.*, 2020]	63.0 ± 1.5	57.3 ± 1.2	61.2 ± 0.9	55.4 ± 0.9
GDPNet [Xue *et al.*, 2021]	67.1 ± 1.0	61.5 ± 0.8	64.9 ± 1.1	60.1 ± 0.9
BCE + $\text{BERTs}_{\text{Base}}$	66.10 ± 0.84	61.78 ± 0.68	63.25 ± 0.45	59.26 ± 0.37
ATL + $\text{BERTs}_{\text{Base}}$	67.15 ± 0.75	62.66 ± 0.70	$\mathbf{66.06 \pm 1.09}$	61.62 ± 0.83
NCRL + $\text{BERTs}_{\text{Base}}$	$\mathbf{68.10 \pm 0.31}$	$\mathbf{64.26 \pm 0.37}$	66.00 ± 0.28	$\mathbf{62.45 \pm 0.22}$

Table 2: Classification results (%) on the DialogRE dataset.

2020] to use F_1 and F_{1c} in evaluation, where F_{1c} is the F_1 score computed by taking the first few turns instead of the entire dialogue as input. Table 2 shows the dialogue-based RE results on DialogRE. As can be seen, our implementation of $\text{BERTs}_{\text{Base}}$ with BCE obtains significantly better results than the reported ones and is slightly worse than the state-of-the-art method, GDPNet, which combines the graph neural networks and $\text{BERTs}_{\text{Base}}$ for better representation learning and uses BCE for training. ATL and our NCRL outperform both BCE and GDPNet by a large margin, showing the benefit of learning the none class for document-level RE. Moreover, NCRL achieves the best and most stable results in most cases. This is because NCRL is more robust against imbalanced and mislabeled samples by applying the margin regularization and shifting approaches. It is worth noting that our margin regularization and shifting techniques are *inapplicable* for BCE and ATL, as they do not formulate the probability of each class as label margins.

Results on GoEmotions. For the GoEmotions dataset, sentences labeled with Neutral are considered as none class instances. Originally, there are a small number *ambiguous* instances labeled with both Neutral and certain emotions. We remove these instances from the train/dev/test set for simplicity. We use $\text{BERT}_{\text{Base}}$ as the encoding model as in [Demszky *et al.*, 2020], and finetune it with the plain NCRL (2) without any regularization, which is enough to achieve good performance in our experiments. We use mean Average Precision (mAP) for evaluation, which summarizes a precision-recall curve at each threshold and measures the model ability in label ranking for each instance. Table 3 shows the classification results on GoEmotions. Again, NCRL obtains the best results, indicating that NCRL works well on multi-label tasks when none class instances are available for training. Specifically, our NCRL improves BCE and ATL by 1.49% and 1.25% in mAP, respectively. This demonstrates that NCRL produces more accurate label rankings by capturing the correlations between the none class and pre-defined classes via the label margins.

Ablation studies. Finally, we perform ablation studies of NCRL on DialogRE. Table 4 provides the results of NCRL when each component of NCRL is excluded at a time, where M-Reg and M-shift denote the margin regularization and the margin shifting, respectively. In addition, we also compare BCE with probability shifting (P-shift) that is used in [Ben-Baruch *et al.*, 2021; Wu *et al.*, 2020] for multi-label learning. We can observe that all components of NCRL contribute to multi-label learning. Both the margin regularization and shifting techniques are important for NCRL to achieve good performance. Even without these two approaches, the plain NCRL still outperforms BCE and its variant, which implies that learning the none class from none class instances is indeed helpful for document-level RE.

Model	Dev mAP	Test mAP
Encoding model	$\text{BERT}_{\text{Base}}$	
BCE	49.79 ± 0.62	48.28 ± 0.37
ATL	49.95 ± 0.59	48.52 ± 0.78
NCRL	$\mathbf{50.99 \pm 0.35}$	$\mathbf{49.77 \pm 0.19}$

Table 3: Results (%) on the GoEmotions dataset.

Model	Dev		Test	
	F_1	F_{1c}	F_1	F_{1c}
NCRL	68.10	64.26	66.00	62.45
- M-Reg	67.27	63.22	65.57	61.77
- M-Shift	67.52	63.59	65.61	62.01
- Both	67.56	63.04	65.86	61.65
BCE	66.10	61.78	63.25	59.25
+ P-Shift	66.69	62.50	63.79	59.71

Table 4: Ablation studies on the DialogRE dataset.

5 Conclusion

We have proposed NCRL, a new multi-label loss that considers the probability of "no relation" in document-level RE. NCRL maximizes the label margin between the none class and each pre-defined class, capturing the label correlations of the none class and enabling adaptive thresholding for label prediction. Through theoretical analysis, we justify the effectiveness of NCRL by proving that NCRL is Bayes consistent w.r.t. its targeted performance measure NCRE. In addition, a margin regularization and a margin shifting technique are proposed to gain robustness against imbalanced and mislabeled samples. Experiments on three benchmark datasets demonstrate that NCRL significantly outperforms existing multi-label loss functions for document-level RE and emotion classification.

A Appendix

A.1 Proof of Bayes Consistency

Loss and risk. In general, multi-label learning methods aim to find a classifier $\boldsymbol{h}$ that minimizes the *multi-label risk* $R_\ell(\boldsymbol{h}) = \mathbb{E}_{(x,y)}[\ell(\boldsymbol{y}, \boldsymbol{h}(\boldsymbol{x}))]$, i.e., the expected classification loss over the joint distribution $\mathbb{P}(X, Y)$, where $\ell(\boldsymbol{y}, \boldsymbol{h}(\boldsymbol{x}))$ is a *multi-label loss* that measures the prediction performance of $\boldsymbol{h}$. A multi-label classifier $\boldsymbol{h}(\boldsymbol{x}) = (h_0(\boldsymbol{x}), \ldots, h_K(\boldsymbol{x}))$ is often induced from a *score function* $\boldsymbol{f}(\boldsymbol{x}) = (f_0(\boldsymbol{x}), \ldots, f_K(\boldsymbol{x}))$ via *thresholding*. Typically, one can take $h_i(\boldsymbol{x}) = [\![f_i(\boldsymbol{x}) > t(\boldsymbol{x})]\!]$ with a thresholding function $t(\boldsymbol{x})$, where $[\![\pi]\!]$ is the mapping that returns 1 if predicate π is true and 0 otherwise. Classical multi-label performance measures include Hamming loss $\ell_H(\boldsymbol{y}, \boldsymbol{h})$ and ranking loss $\ell_{rnk}(\boldsymbol{y}, \boldsymbol{h})$, whose definitions are given as follows:

$$\ell_H(\boldsymbol{y}, \boldsymbol{h}) = \sum_{i=1}^{K} [\![y_i \neq h_i]\!],$$

$$\ell_{rnk}(\boldsymbol{y}, \boldsymbol{h}) = \sum_{(i,j): y_i > y_j} \Big([\![f_i < f_j]\!] + \frac{1}{2}[\![f_i = f_j]\!]\Big).$$

Intuitively, Hamming loss counts the number of misclassified labels, and ranking loss counts the number of reversely ordered label pairs.

Surrogate loss and consistency. Although our goal is to find an optimal classifier w.r.t. certain performance measure (loss), directly optimizing a performance measure such as Hamming loss and ranking loss is generally intractable. Because of this, one typically needs to minimize a *surrogate loss* instead. For example, BCE and pairwise ranking loss are the surrogate losses of Hamming loss and ranking loss, respectively. A proper surrogate loss should be easy for optimization and *consistent* with the corresponding performance measure. Formally, the multi-label consistency is defined as follows:

Definition 1 (Multi-label consistency [Gao and Zhou, 2013])**.** *The surrogate loss L is Bayes consistent w.r.t. the multi-label performance measure ℓ if and only if it holds for any sequence $\boldsymbol{f}_n$ that*

$$R_L(\boldsymbol{f}_n) \to R_\ell^* \Rightarrow R_\ell(\boldsymbol{f}_n) \to \min R_L^*, \tag{8}$$

where R_ℓ^* and R_L^* are the minimal risk w.r.t. ℓ and L, respectively.

Consistency is important for good classification performance, which guarantees that minimizing the surrogate loss can achieve the expected objective. Nevertheless, it is nontrivial to design a consistent surrogate loss. In fact, even some commonly used surrogate losses are inconsistent. For example, it is known that all convex surrogate loss functions including pairwise ranking loss is *inconsistent* with the ranking loss [Gao and Zhou, 2013]. In what follows, we prove the Bayes consistency of NCRL w.r.t. NCRE.

Theorem 1. *NCRL (2) is Bayes consistent w.r.t. NCRE (1).*

Proof. Let $\Delta_i = \mathbb{P}(y_i = 1|\boldsymbol{x})$ be the marginal probability when the i-th label is positive. The conditional risk of NCRE (1) is given by

$$R_{L_{\text{NA}}}(\mathbb{P}, \boldsymbol{f}) = \sum_{i=1}^{K} \Big(\Delta_i \cdot \log \sigma(f_i - f_0) + (1 - \Delta_i) \cdot \log \sigma(f_0 - f_i)\Big). \tag{9}$$

For $i = 1, \ldots, K$, the partial derivative can be computed by

$$\frac{\partial}{f_i}\mathbb{E}[L_{\text{NA}}(\mathbb{P}, \boldsymbol{f})|\boldsymbol{x}] = \Delta_i \frac{-\exp(f_0 - f_i)}{1 + \exp(f_0 - f_i)} + (1 - \Delta_i)\frac{\exp(f_i - f_0)}{1 + \exp(f_i - f_0)}, \tag{10}$$

Since NCRL (2) is convex and differentiable, we can obtain the optimal $\boldsymbol{f}^*$ by setting the partial derivatives to zero, which leads to

$$f_i^* - f_0^* = \log(\frac{\Delta_i}{1 - \Delta_i}), i = 1, \ldots, K. \tag{11}$$

This means that, for the optimal score function $\boldsymbol{f}^*$, $f_i^* > f_0^*$ if and only if $\Delta_i > \frac{1}{2}$, which minimizes the NCRE risk according to (4). Therefore, NCRL is Bayes consistent w.r.t. NCRE (1). □

A.2 Detailed Configurations of Encoding Models

We construct the encoding model by finetuning a pre-trained language model (PLM) for representation learning. Since the inputs of different datasets are different, we follow existing works to build the encoding model for each dataset as follows:

ATLOP on DocRED. We adopt ATLOP [Zhou *et al.*, 2021] as the encoding model on DocRED. Given a document $\mathcal{D} = [w_i]_{i=1}^{L}$ of length L, ATLOP inserts a special symbol "*" around each entity mentions in the document $\mathcal{D}$, and then feeds the document into a PLM to obtain d-dimensional contextual embeddings:

$$\mathbf{H} = [\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_L] = \text{PLM}([w_1, w_2, ..., w_L]). \tag{12}$$

In ATLOP, the embedding of "*" at the start of mentions is used as the mention embeddings. For an entity e_i with N_{e_i} mentions $\{e_i^j\}_{j=1}^{N_{e_i}}$, the entity embedding $\boldsymbol{h}_{e_i}$ is computed by applying the logsumexp pooling [Jia *et al.*, 2019] as follows:

$$\boldsymbol{h}_{e_i} = \log \sum_{j=1}^{N_{e_i}} \exp\Big(\boldsymbol{h}_{e_i^j}\Big). \tag{13}$$

Given the embeddings $(\boldsymbol{h}_{e_s}, \boldsymbol{h}_{e_o})$ of an entity pair (e_s, e_o), the latent representations $\boldsymbol{z}_s$ and $\boldsymbol{z}_o$ of e_s and e_o are obtained by fusing $\boldsymbol{h}_{e_s}$ and $\boldsymbol{h}_{e_o}$ with a localized context embedding followed by a non-linear projection (please refer to [Zhou *et al.*, 2021] for more details). Finally, the entity pair representation is computed by $\boldsymbol{x} \in \mathbb{R}^{d^2} = \boldsymbol{z}_o \otimes \boldsymbol{z}_s$, and the confidence scores $\boldsymbol{f} \in \mathbb{R}^{K+1}$ given $\boldsymbol{x}$ is obtained by

$$\boldsymbol{f} = \mathbf{W}\boldsymbol{x} + \boldsymbol{b}, \tag{14}$$

where $\otimes$ is the Kronecker product, $\boldsymbol{b} \in \mathbb{R}^{K+1}$ is a bias term, and $\mathbf{W} \in \mathbb{R}^{K+1 \times d^2}$ is the model parameters for the multi-label classifier.

Model	Dev		Test	
	F_1	F_{1c}	F_1	F_{1c}
BCE	66.10 ± 0.84	61.78 ± 0.68	63.25 ± 0.45	59.26 ± 0.37
BCE + Finer-Grained	66.40 ± 0.55	62.03 ± 0.44	63.37 ± 0.42	59.36 ± 0.32
BCE + Per-Label	67.65 ± 1.13	64.19 ± 0.58	62.91 ± 1.44	60.55 ± 0.91
ATL	67.15 ± 0.75	62.66 ± 0.70	$\mathbf{66.06 \pm 1.09}$	61.62 ± 0.83
NCRL	$\mathbf{68.10 \pm 0.31}$	$\mathbf{64.26 \pm 0.37}$	66.00 ± 0.28	$\mathbf{62.45 \pm 0.22}$

Table 5: Classification results (%) on the DialogRE dataset.

BERTs on DialogRE. We adopt BERTs [Yu *et al.*, 2020], an improved BERT baseline, as the encoding model on DialogRE. Given a dialogue $\mathcal{D}$ and an entity pair (e_s, e_o), BERTs augments the dialogue with the entity pair by [CLS] $\mathcal{D}$ [SEP] e_s [SEP] e_o [SEP], where [CLS] and [SEP] are the classification and separator tokens defined in BERT, respectively. If e_s or e_o indicates the name of a speaker, all the mentions of such speaker in the augmented dialogue are replaced by a special token [Sub] (or [Obj]) to locate the positions of relevant turns. After feeding the augmented dialogue into the BERT model, the embedding of [CLS] is used as the entity pair representation $\boldsymbol{x}$ for classification.

BERT on GoEmotions. We follow [Demszky *et al.*, 2020] and use the plain BERT as the encoding model on GoEmotions for emotion classification. By feeding a given sentence into BERT, the embedding of [CLS] is used as the instance representation $\boldsymbol{x}$ for classification.

A.3 Effect of Threshold Tuning

In this section, we test BCE with carefully tuned thresholding and compare it with the proposed NCRL. Please note that NCRL and ATL can learn the threshold automatically, and only BCE needs to select the best threshold for prediction. Specifically, we consider two threshold tuning strategies for BCE: (1) We select a *global* threshold from a finer-grained grid of $\{0.1, 0.11, ..., 0.9\}$, after confirming that the performance will not be further improved on the development set by searching more values. (2) We tune a *per-label* threshold from $\{0.1, 0.11, ..., 0.9\}$.

Table 5 shows the results of BCE with the threshold tuned by different strategies, where we also include the results reported in Table 2 for clarity. As can be seen, with the finer-grained global threshold, BCE obtains slightly better results than the originally reported ones. When the pre-label threshold is used, BCE obtains much better results on the development set, whereas the improvement is not that significant on the test test. In particular, BCE even gets lower F1 scores than the originally reported results, indicating that the pre-label threshold overfits the development set. Even with a carefully tuned per-label threshold, BCE is still worse than our NCRL in all the cases. This demonstrates that by exploiting the none class instances, NCRL not only learns a better threshold but also achieves a better model for multi-label learning.

A.4 Importance of None Class Instances

Although the proposed NCRL is a generic and can be used for multi-label tasks other than document-level RE, its effectiveness comes from the information contained in none class instances. When there is no none class instance for training, it is difficult for NCRL to estimate the confidence score of the none class and capture correct correlations between the none class and pre-defined classes via label ranking. Without positive samples of $y_0 = 1$, the score f_0 does not reflect the confidence for an instanced to be labeled as the none class anymore, as f_0 will always be ranked below certain pre-defined labels during training. Consequently, NCRL losses its ability in adjusting label rankings and performing adaptive thresholding.

Model	Dev F1	Test F1
BCE	75.21	**73.79**
ATL	73.93	72.25
ATL+	75.15	73.44
NCRL	73.65	71.85
NCRL+	**75.81**	73.13

Table 6: Results (%) on the AAPD dataset.

To study how none class instances affect the performance of NCRL, we conduct experiments on the AAPD dataset [Yang *et al.*, 2018] for multi-label text classification, which contains the abstract of 55840 papers from 54 subjects in the field of computer science. All the abstracts in the AAPD dataset have at least one subject (pre-defined label), and there is no none class instance. We use the BERT base pre-trained model for text encoding and then finetune it for classification.

Table 6 shows the classification results on AAPD. ATL+ and NCRL+ represent ATL and NCRL using a tuned threshold as in BCE rather than the none class label for multi-label prediction, respectively. As can be seen, when the learned threshold is used, ATL and NCRL are worse than BCE. On the other hand, when the tuned threshold is used, all the losses obtain similar results. This implies that the lack of none class instances prevents ATL and NCRL from learning the correct confidence scores of the none class for thresholding. Different from AAPD, document-level relation extraction datasets contain many none class instances, which can be utilized by NCRL for better classification performance.

A.5 Sensitivity to Margin Shifting

Figure 2 shows the performance of NCRL with different γ on the DocRED dataset. As can be seen, NCRL is not very sensitive to γ.

A.6 Hyper-Parameter Settings

Table 7 provides the detailed hyper-parameter settings on different datasets.

Hyper-param	DocRED			DialogRE	GoEmotions	AAPD
	BERT	RoBERTa	DeBERTa	BERT	BERT	BERT
Batch size	4	4	4	24	16	16
Gradient accumulation steps	1	2	2	6	1	1
# Epoch	8	8	8	20	10	30
lr for encoder	5e-5	3e-5	1e-5	3e-5	5e-5	2e-5
lr for classifier	1e-4	1e-4	1e-4	3e-5	5e-5	2e-5
γ for margin shifting	0.05	0.05	0.05	0.01	0	0.01

Table 7: Hyper-parameter settings on different datasets.

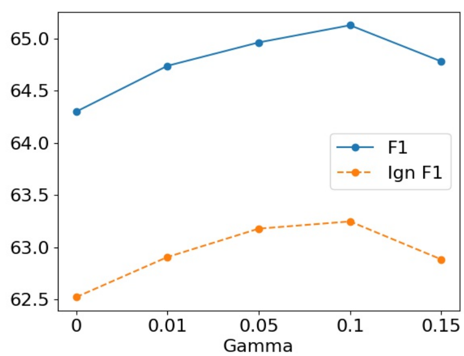

Figure 2: Classification results of NCRL with different γ on DocRED.

Acknowledgments

This research is supported in part by the National Research Foundation, Singapore under its AI Singapore Program (AISG Award No: AISG2-RP-2020-016).

References

[Ben-Baruch *et al.*, 2021] Emanuel Ben-Baruch, Tal Ridnik, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of ICCV*, 2021.

[Christopoulou *et al.*, 2019] Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. Connecting the dots: Document-level neural relation extraction with edge-oriented graphs. In *Proceedings of EMNLP-IJCNLP*, pages 4927–4938, 2019.

[Dai *et al.*, 2021] Damai Dai, Jing Ren, Shuang Zeng, Baobao Chang, and Zhifang Sui. Coarse-to-fine entity representations for document-level relation extraction. In *Proceedings of AAAI*, 2021.

[Demszky *et al.*, 2020] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of ACL*, pages 4040–4054, 2020.

[Durand *et al.*, 2019] Thibaut Durand, Nazanin Mehrasa, and Greg Mori. Learning a deep convnet for multi-label classification with partial labels. In *Proceedings of CVPR*, pages 647–657, 2019.

[Fürnkranz *et al.*, 2008] Johannes Fürnkranz, Eyke Hüllermeier, Eneldo Loza Mencía, and Klaus Brinker. Multilabel classification via calibrated label ranking. *Machine Learning*, 73(2):133–153, 2008.

[Gao and Zhou, 2013] Wei Gao and Zhi-Hua Zhou. On the consistency of multi-label learning. *Artificial Intelligence*, 199-200:22–44, 2013.

[Goyal *et al.*, 2017] Priya Goyal, P. Dollár, Ross B. Girshick, P. Noordhuis, L. Wesolowski, Aapo Kyrola, Andrew Tulloch, Y. Jia, and Kaiming He. Accurate, large minibatch SGD: Training imagenet in 1 hour. *ArXiv*, abs/1706.02677, 2017.

[He *et al.*, 2021] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced bert with disentangled attention. In *Proceedings of ICLR*, 2021.

[Jia *et al.*, 2019] Robin Jia, Cliff Wong, and Hoifung Poon. Document-level n-ary relation extraction with multiscale representation learning. In *NAACL-HLT*, 2019.

[Li *et al.*, 2017] Yuncheng Li, Yale Song, and Jiebo Luo. Improving pairwise ranking for multi-label image classification. In *Proceedings of CVPR*, pages 3617–3625, 2017.

[Lin *et al.*, 2017] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of ICCV*, pages 2980–2988, 2017.

[Liu *et al.*, 2019] Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019.

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

[Miwa and Bansal, 2016] Makoto Miwa and Mohit Bansal. End-to-end relation extraction using lstms on sequences and tree structures. In *Proceedings of ACL*, 2016.

[Nan *et al.*, 2020] Guoshun Nan, Zhijiang Guo, Ivan Sekulic, and Wei Lu. Reasoning with latent structure refinement for document-level relation extraction. In *Proceedings of ACL*, pages 1546–1557, 2020.

[Tang *et al.*, 2020] Hengzhu Tang, Yanan Cao, Zhenyu Zhang, Jiangxia Cao, Fang Fang, Shigang Wang, and Pengfei Yin. Hin: Hierarchical inference network for document-level relation extraction. In *Proceedings of PAKDD*, 2020.

[Wang *et al.*, 2020] Difeng Wang, Wei Hu, Ermei Cao, and Weijian Sun. Global-to-local neural networks for document-level relation extraction. In *Proceedings of EMNLP*, pages 3711–3721, 2020.

[Wolf *et al.*, 2020] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of EMNLP*, pages 38–45, 2020.

[Wu *et al.*, 2019] Jiawei Wu, Wenhan Xiong, and William Yang Wang. Learning to learn and predict: A meta-learning approach for multi-label classification. In *Proceedings of EMNLP-IJCNLP*, pages 4354–4364, 2019.

[Wu *et al.*, 2020] Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *Proceedings of ECCV*, pages 162–178, 2020.

[Xu *et al.*, 2019] Miao Xu, Yu-Feng Li, and Zhi-Hua Zhou. Robust multi-label learning with pro loss. *IEEE Transactions on Knowledge and Data Engineering*, 32(8):1610–1624, 2019.

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

[Xu *et al.*, 2021b] Wang Xu, Kehai Chen, and Tiejun Zhao. Document-level relation extraction with reconstruction. In *Proceedings of AAAI*, 2021.

[Xue *et al.*, 2021] Fuzhao Xue, Aixin Sun, Hao Zhang, and Eng Siong Chng. GDPNet: Refining latent multi-view graph for relation extraction. In *Proceedings of AAAI*, 2021.

[Yang *et al.*, 2018] Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. SGM: Sequence generation model for multi-label classification. In *Proceedings of ACL*, pages 3915–3926, 2018.

[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 ACL*, 2019.

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

[Yeh *et al.*, 2017] Chih-Kuan Yeh, Wei-Chieh Wu, Wei-Jen Ko, and Yu-Chiang Frank Wang. Learning deep latent space for multi-label classification. In *Proceedings of AAAI*, 2017.

[Yu *et al.*, 2020] Dian Yu, Kai Sun, Claire Cardie, and Dong Yu. Dialogue-based relation extraction. In *Proceedings of ACL*, pages 4927–4940, 2020.

[Zeng *et al.*, 2020] Shuang Zeng, Runxin Xu, Baobao Chang, and Lei Li. Double graph based reasoning for document-level relation extraction. In *Proceedings of EMNLP*, pages 1630–1640, 2020.

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

[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 AAAI*, 2021.