Ca*DRL*: Document-level Relation Extraction via Context-aware Differentiable Rule Learning

Kunli Zhang[‡], Pengcheng Wu[‡], Bohan Yu[‡], Kejun Wu[‡], Aoze Zheng[‡], Xiyang Huang[‡]
Chenkang Zhu[‡], Min Peng[♭], Hongying Zan[‡], Yu Song[‡*]
[‡]School of Computer Science and Artificial Intelligence, Zhengzhou University
[♭]School of Computer Science, Wuhan University
{ieysong, ieklzhang}@zzu.edu.cn, {pcwu2022, alexyu}@gs.zzu.edu.cn

Abstract

Document-level Relation Extraction (DocRE) aims to extract relations from documents. Compared with sentence-level relation extraction, it is necessary to extract long-distance dependencies. Existing methods enhance the output of trained DocRE models either by learning logical rules or by extracting rules from annotated data and then injecting them into the model. However, these approaches can result in suboptimal performance due to incorrect rule set constraints. To mitigate this issue, we propose **C**ontext-**a**ware **D**ifferentiable **R**ule **L**earning or **Ca*DRL*** for short, a novel differentiable rule-based framework that learns the doc-specific logical rule to avoid generating suboptimal constraints. Specifically, we utilize Transformer-based relation attention to encode document and relation information, thereby learning the contextual information of the relation. We employ a sequence-generated differentiable rule decoder to generate relational probabilistic logic rules at each reasoning step. We also introduce a parameter sharing training mechanism in Ca*DRL* to reconcile the DocRE model and the rule learning module. Extensive experimental results on three DocRE datasets demonstrate that Ca*DRL* outperforms existing rule-based frameworks, significantly improving DocRE performance and making predictions more interpretable and logical.

1 Introduction

In recent years, document-level relation extraction (DocRE) has garnered significant attention from researchers. Unlike sentence-level relation extraction (RE) (Zeng et al., 2014; Zhang et al., 2017; Han et al., 2018; Wang et al., 2021), DocRE presents unique challenges: 1) capturing the complex remote dependencies between entity pairs in documents. 2) The absence of logical frameworks makes it susceptible to logical reasoning errors.

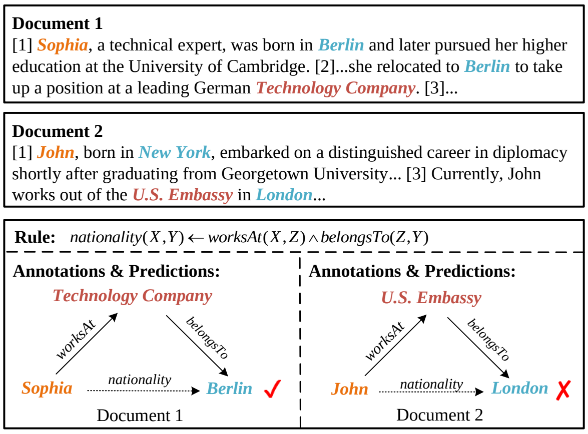

Figure 1: Example of a global rule in the DocRED dataset that fails to apply correctly. While the rule holds in Document 1, it produces an incorrect result in Document 2, highlighting the need for domain-specific rules for Document 2.

In response to the challenges, existing research can primarily be divided into three categories: the *sequence-based model*, the *graph-based model*, and the *rule constraints model*. In sequence-based and graph-based models, the focus is learning more powerful implicit representations (Devlin et al., 2019; Liu et al., 2019; Zeng et al., 2020; Zhou et al., 2020). However, these methods lack logic and transparency. Logical rules can effectively address these issues. Rule learning is now widely applied not only in knowledge graphs but also in relation triples (Xu et al., 2023). The structured nature of logical rules gives them an irreplaceable advantage in mining implicit relations, circumventing the difficulties of capturing long-term dependencies and using inherent correlations to explain results. If models could automatically learn and utilize these rules for prediction, we would achieve better RE performance and greater interpretability.

The existing rule-constrained DocRE models include LogiRE (Ru et al., 2021) and MILR (Fan et al., 2022). LogiRE first learns logical rules based

*Corresponding authors.

Proceedings of the 31st International Conference on Computational Linguistics, pages 8272–8284
January 19–24, 2025.

on the output logits of a trained neural model. In contrast, MILR first learns logical rules from annotated data and then trains a model penalized by an auxiliary loss to account for rule violations. However, these approaches may result in the unintended consequence of erroneous rule set constraints. Using a static prior rule set, we predicted the relations between the training and test sets on DWIE, treating cases that adhered to the rules as successes. The failure rates were 12.9% in the training set and 28.6% in the test set, highlighting the problem of incorrect constraints due to rule solidification. In Figure 1, the rule `nationality(X,Y)` $\leftarrow$ `worksAt(X,Z)` $\wedge$ `belongsTo(Z,Y)` reflects specific logic in the prediction of relations in Document 1, but this a priori rule does not apply to Document 2, leading to incorrect results.

In this paper, we propose **C**ontext-**a**ware **D**ifferentiable **R**ule **L**earning or **Ca*DRL*** for short, a novel differentiable rule-based framework that learns the *doc-specific logical rule* to avoid generating suboptimal constraints. Specifically, we leverage Transformer-based relational attention to encode both document and relation information, enabling the model to capture the context of relations. We employ a sequence-generated, differentiable rule decoder to produce relation probabilistic logic rules at each step. In training, we introduce a parameter-sharing training mechanism in **Ca*DRL*** to integrate the DocRE model with the rule learning module, facilitating more efficient collaboration between the two components, to further improve the performance. We experimented **Ca*DRL*** with four enhanced DocRE models, including BiLSTM (Yao et al., 2019), GAIN (Zeng et al., 2020), ATLOP (Zhou et al., 2021) and DREEAM (Ma et al., 2023a). We further evaluated it on large language models (LLMs), including ChatGPT and GPT-4. Experimental results on three DocRE datasets DWIE (Zaporojets et al., 2021), DocRED (Yao et al., 2019), HacRED (Cheng et al., 2021) demonstrate that the proposed **Ca*DRL*** framework is superior to the rule-based framework for DocRE. Our main contributions are as follows:

- We propose a DocRE constraint framework called **Ca*DRL***, which constrains the DocRE model using a differentiable rule learning module and employs parameter sharing for joint training. As far as we know, this is the first differentiable rule learning approach imposes logical rules on DocRE models.
- We introduced an encoder based on relational attention and a differentiable rule decoder based on sequence generation. We utilize the TensorLog mechanism to obtain high-quality RE results.
- Extensive experiments on three DocRE datasets demonstrate that **Ca*DRL*** consistently achieves improvements across various backbones and further enhances RE performance. The improvement on the test set is greater than that on the validation set, which shows the superiority of **Ca*DRL*** in the dynamic learning of rules.

2 Related Work

2.1 Document-level Relation Extraction

Previous research on DocRE has primarily focused on improving representation learning. Advanced neural network architectures, including attention mechanisms (Yao et al., 2019; Zhou et al., 2021), graph neural networks (Zhang et al., 2017; Zeng et al., 2020), and pre-trained language models ((Jia et al., 2019; Tang et al., 2020; Xu et al., 2021)), have been employed as encoders to generate representations of entity pairs. Several studies (Tan et al., 2022; Ma et al., 2023a) adopt knowledge distillation, where evidence information serves as a supervisory signal to guide the attention module in assigning higher weights to relevant evidence. In addition, other researchers (Zhu et al., 2024; Xue et al., 2024; Li et al., 2024) have explored the use of prompt learning to enhance the generative capabilities of LLMs, applying this approach to improve performance in DocRE tasks.

2.2 Differentiable Rule Learning

Differentiable rule learning approaches based on TensorLog (Cohen, 2016), are introduced to address the limitations of symbolic-based methods that mine rules discretely. Neural-LP (Yang et al., 2017), a pioneering method, focuses on learning probabilistic closed-path rules and simultaneously optimizes both the parameters and structure of these rules. Subsequent developments like DRUM (Sadeghian et al., 2019) enhance the architectural framework of Neural-LP, achieving superior performance. Neural-Num-LP (Wang et al., 2020) extends this concept to include numerical rules, providing significant insights into potential reasoning patterns. Ruleformer (Xu et al., 2022) prioritizes the selection of the most appropriate rule among various candidates. To broaden the scope of rule diversity, Neural Logic Inductive Learning (NLIL) (Yang and Song, 2019) addresses non-closed path rules by integrating elementary statements.

2.3 Rule Constraint in the DocRE Model

To capture more complex interdependencies between entity pairs and enhance interpretability, some studies have added logical reasoning frameworks to GNN-based and attention-based methods for constraints (Ru et al., 2021; Fan et al., 2022; Liu et al., 2023; Qi et al., 2024). LogiRE (Ru et al., 2021) treats logic rules as latent variables and introduces them into the neural network via a rule generator and a relation extractor, explicitly capturing remote dependencies and obtaining better explanations. MILR (Fan et al., 2022) statically mines logic rules from the training set based on confidence and then trains a DocRE model constrained by a training loss function. BCBR (Liu et al., 2023) improves on the rule mining strategy of MILR by modeling rules through beta distributions and constructing bi-directional logical constraint loss to regulate the output of the DocRE model. JMRL (Qi et al., 2024) uses an end-to-end model to constrain the neural model, which introduces residual connections and auxiliary loss to unify the DocRE model with the logical reasoning module.

However, these approaches can result in the problem of erroneous rule set constraints leading to suboptimal results. In contrast, our proposed **Ca*DRL*** framework employs a dynamic, differentiable rule approach to learn the logical rules specific to each document. This method achieves targeted and independent rule constraints, thereby reducing the occurrence of suboptimal outcomes.

3 Preliminaries

3.1 Problem Formulation

Given a document $\mathcal{D}$ containing a set of named entities $\mathcal{E}_{\mathcal{D}} = \{e_i\}_{i=1}^{n_d}$, the task of DocRE involves predicting the relations r between entity pairs $(e_h, e_t)_{h,t\in 1,\ldots,n,h\neq t}$, where $r \in \mathbb{R}$ and $\mathbb{R} = \mathcal{R} \cup \mathcal{NA}$. Here, $\mathcal{R}$ denotes a pre-defined set of relation types, and $\mathcal{NA}$ stands for "no relation", respectively. An entity e_i can be mentioned multiple times in $\mathcal{D}$ as $\{m_j^i\}_{i=1}^{N_{e_i}}$, where the existence of $r_{e_h}^{e_t}$ between e_h and e_t is determined by the corresponding mentions.

3.2 Atoms and Logical Rules

The atom (e_h, r, e_t) or $r\,(e_h, e_t)$ is a binary variable that indicates whether the relation $r \in \mathbb{R}$ exists between e_h and e_t. If r exists, $r\,(e_h, e_t) = 1$. Otherwise $r\,(e_h, e_t) = 0$.

First-order logic (FOL) is formed from *constants*, *variables* and *predicates* with propositional connectives $\wedge$, $\vee$, $\neg$ and quantifiers. We focus on learning the conjunctive paradigm $\wedge$. A clause can be written in the form of a rule: $H \leftarrow B_1 \wedge \ldots \wedge B_k$, where H is called rule head and $B_1 \ldots B_k$ is the rule body. A FOL is referred to as a FOL-L if it contains L body atoms. We define the set of FOLs as $\widetilde{\mathcal{R}} \in \mathcal{R} \cup \{r^{-1} \mid r \in \mathcal{R}\}$.

4 The Ca*DRL* Framework

To adopt doc-specific logical rule constraints for the DocRE model, we propose a new rule-constrained framework, named Context-aware Differentiable Rule Learning, or **Ca*DRL*** for short, as illustrated in Figure 2. **Ca*DRL*** initially employs a Transformer-based relational attention mechanism to encode document and relation information. Subsequently, the document encoding is provided to the DocRE model, and the relation encoding is supplied to a sequence generation-based differentiable relation decoder. We also introduce a parameter-sharing method to jointly train the DocRE model and the rule constraint module, achieving the objectives by minimizing the original DocRE's relational classification loss and rule constraint loss.

4.1 Relational Attention Encoder

Ca*DRL* adopts a Transformer-based relational attention encoder to aggregate contextual information about relations. This encoder jointly encodes document information and relational information, feeding the relational encoding into a rule decoder to learn relational context, thus generating higher-quality rules. To enhance the results, relational information is also integrated into the document encoding, thereby optimizing the output of the DocRE model.

After embedding, $\mathcal{D}$ can be obtained as $\mathcal{T}_{\mathcal{D}} = [t_1, t_2, \ldots, t_n, \ldots, blank]$, where t_i represents the embedding of each token, including entity and mentioned embeddings. n is the number of tokens. $blank$ is a special embedding. Attention calculation is performed on the embedding of relation $r^{(e_h,e_t,\mathcal{D})}$ and the position i in the document, which is shown in:

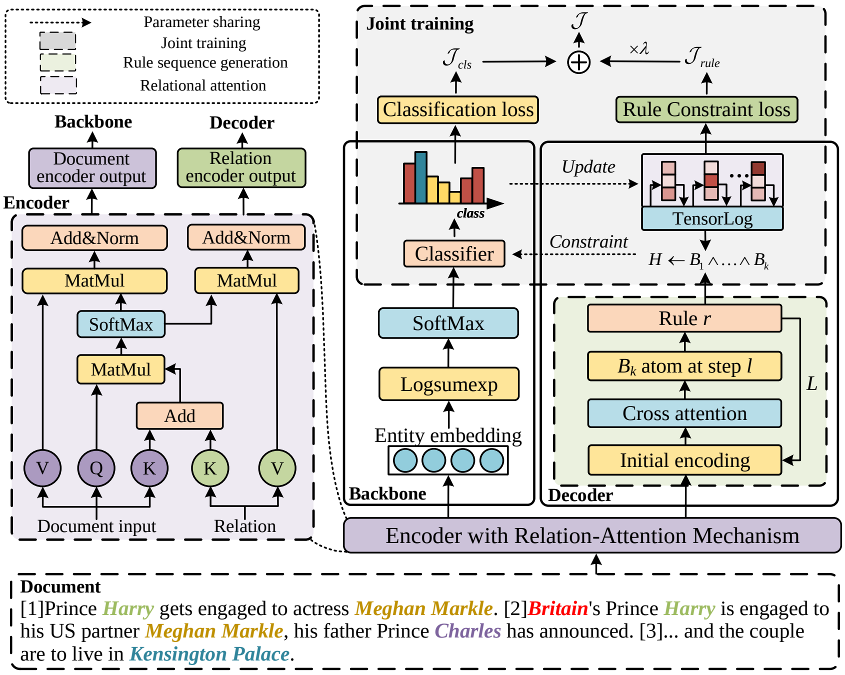

Figure 2: The overview of the proposed **Ca*DRL*** framework.

$$\phi^{\mathcal{D}}_{(r,i)} = \frac{\left(r^{(e_h,e_t,\mathcal{D})}\, \mathcal{Q}_r^{(e_h,e_t,\mathcal{D})}\right)\left(t_i \mathcal{K}_i^{\mathcal{D}}\right)}{\sqrt{d_k+\epsilon}}, \tag{1}$$

where $\mathcal{Q}_r^{(e_h,e_t,\mathcal{D})}$ is the query matrix for $r^{(e_h,e_t,\mathcal{D})}$, $\mathcal{K}_i^{\mathcal{D}}$ is the key matrix at i, d_k is the dimension of $\mathcal{K}_i^{\mathcal{D}}$, ϵ is the smoothing factor. Then, the attention $\phi^{\mathcal{D}}_{(i,j)}$ between the embeddings at positions i and j in $\mathcal{D}$ is calculated:

$$\phi^{\mathcal{D}}_{(i,j)} = \frac{\left(t_i \mathcal{Q}_i^{\mathcal{D}}\right)\left(t_j \mathcal{K}_j^{\mathcal{D}} + \sum_{r=1}^{|\mathbb{R}|}\left(r^{(e_h,e_t,\mathcal{D})} \mathcal{K}_r^{(e_h,e_t,\mathcal{D})}\right)\right)}{\sqrt{d_k+\epsilon}}, \tag{2}$$

where $|\mathbb{R}|$ is the size of relation set. $\phi^{\mathcal{D}}_{(r,i)}$ and $\phi^{\mathcal{D}}_{(i,j)}$ are subjected to $SoftMax$ calculation.

The document encodings $\psi_i^{\mathcal{D}}$ and relation encodings $\psi_r^{\mathcal{D}}$ are obtained by normalizing the products of $\phi^{\mathcal{D}}_{(r,i)}$ and $\phi^{\mathcal{D}}_{(i,j)}$ with their respective value matrices. The calculation process is identical for both $\psi_i^{\mathcal{D}}$ and $\psi_r^{\mathcal{D}}$, although only the derivation of $\psi_r^{\mathcal{D}}$ is presented here. The corresponding formulas are as follows:

$$\psi_i^{\mathcal{D}} = \sum_{j=1}^{n} \phi^{\mathcal{D}}_{(i,j)} \left(t_j \mathcal{V}_j^{\mathcal{D}} + \sum_{r=1}^{|\mathbb{R}|} r^{(e_h,e_t,\mathcal{D})} \mathcal{V}_r^{(e_h,e_t,\mathcal{D})} \right), \tag{3}$$

where $\mathcal{V}_r^{(e_h,e_t,\mathcal{D})}$, $\mathcal{V}_j^{\mathcal{D}}$ are the value matrices for r and positions j, respectively. Finally, $\psi_r^{\mathcal{D}}$ is used as the output of the encoder and input into the decoder to generate rules, while $\psi^{\mathcal{D}} = \left[\psi_1^{\mathcal{D}}, \psi_2^{\mathcal{D}}, \ldots, \psi_n^{\mathcal{D}}\right]$ is used as the output of the document encoding and input into backbone for RE.

4.2 Sequence Generation Rule Decoder

Existing studies use a priori rule sets, which may yield suboptimal outcomes due to inappropriate constraints. Therefore, **Ca*DRL*** introduces a rule decoder based on sequence generation and utilizes TensorLog to render rule extraction a differentiable process. TensorLog facilitates complex logical reasoning efficiently through matrix operations.

Let L be the maximum number of atoms in each rule and $\mathcal{R}_+ = \mathcal{R} \cup \mathcal{R}^- \cup \mathcal{NA}$. $\mathcal{R}^-$ represents the inverse relation. Suppose $\mathcal{R} = \{r_i\}_{1\leq i\leq n}$, then $\mathcal{R}^- = \{r_i\}_{n+1\leq i\leq 2n}$. The decoder uses $\mathcal{NA}$ as r_h in FOL-L and generates the relation with the highest probability at each step until the sequence output achieves the predetermined rule length L. Following this, cross-attention computation with $\psi_r^{\mathcal{D}}$ yields an intermediate vector. To determine $S_{r,L}^{(e_h,e_t,\mathcal{D})}$, MLP (Taud and Mas, 2018) is employed to calculate the probability ω_l^r of $S_{r,l}^{(e_h,e_t,\mathcal{D})}$ in step l. $S_{r,l}^{(e_h,e_t,\mathcal{D})}$ with the highest ω_l^r is selected and incorporated into next step. If $r_{l+1}^{(e_h,e_t,\mathcal{D})}$ is a maximum probability relation, then $S_{r,l+1}^{(e_h,e_t,\mathcal{D})} = \left[S_{r,l}^{(e_h,e_t,\mathcal{D})}, r_{l+1}^{(e_h,e_t,\mathcal{D})}\right]$. After repeating L times, the rule is obtained. The rule with lengths less than L is populated using the $\mathcal{NA}$ in the relation set.

The quality of generation should be evaluated using the probability of r corresponding to B_i. We multiply the probabilities of B_i in their ω_l to obtain the score:

$$\Theta_r^{(e_h,e_t,\mathcal{D})} = \prod_{l=1}^{L} \omega_l^{max}. \tag{4}$$

However, unlike the sequence generation task in machine translation, **Ca*DRL*** lacks labels to verify whether the generated relations are the optimal choice. We draw on the idea of **TensorLog** to obtain predictions of relation labels. Specifically, we represent e_i as a one-hot vector $\mathbf{v}^{e_i} \in \{0,1\}^{|\mathcal{E}|}$, $|\mathcal{E}|$ being the size of the entity set, and represent the extracted relation as an adjacency matrix $\mathbf{M}^{r_k} \in \{0,1\}^{|\mathcal{E}|\times|\mathcal{E}|}$. If $\mathbf{M}_{ij}^{r_k} = \mathbf{1}$, it means that e_i and e_j have relation r_k, $k = 1,\dots,n$, otherwise $\mathbf{M}_{ij}^{r_k} = 0$. The tail entities e_j can be obtained as follows:

$$\mathbf{v}_{(e_i,r_k)}^{e_j} = \mathbf{v}^{e_i}\mathbf{M}^{r_k}, \tag{5}$$

where $\mathbf{v}_{(e_i,r_k)}^{e_j}$ is a one-hot vector containing information about multiple entities. The triples obtained at each step constitute B_i, allowing the current B_i to be organized into an adjacency matrix $\mathbf{M}^{r_k}$. This rule obtains the tail entity through multiplication with the L step adjacency matrix $\mathbf{M}^{r_i}$. See the Appendix B for the detailed derivation process.

Sequence generation can be formalized into a differentiable process for training. For the triple (e_h, r, e_t), r is used as H to verify B_i and to construct the loss function. B_i atoms at step l are derived as follows:

$$\xi_l^{(e_i,r_k,\mathcal{D})} = \xi_{l-1}^{(e_i,r_k,\mathcal{D})} \times \sum_{n=1}^{|\mathbb{R}|} \omega_l^n \mathbf{M}^{r_n}, \tag{6}$$

where $\xi_{l-1}, \xi_l \in \mathbb{R}^{|\mathcal{E}|\times 1}$ are the representations of the entities in steps $l-1$ and l, respectively. The result $\xi_L^{(e_i,r_k,\mathcal{D})}$ is obtained after L steps of reasoning. The reasoning score is derived from the similarity between $\xi_L^{(e_i,r_k,\mathcal{D})}$ and the target entity's one-hot vector $\mathbf{v}$:

$$\Psi(e_t \mid e_h, r) = \mathbf{v}\cdot\log\left[\xi_L^{(e_i,r_k,\mathcal{D})}, \gamma\right]_+, \tag{7}$$

where $\left[\xi_L^{(e_i,r_k,\mathcal{D})}, \gamma\right]_+$ denotes the maximum value of each element in $\xi_L^{(e_i,r_k,\mathcal{D})}$ for γ, γ is a stabilizing constant. The loss function of the differentiable rules is as follows:

$$\mathcal{J}_{rule}^{(e_h,e_t,\mathcal{D})} = \sum_{(e_h,r,e_t)\in\mathcal{D}} (1 - \Psi(e_t \mid e_h, r)). \tag{8}$$

4.3 Joint Training

The DocRE model $\mathcal{F}$ calculates a logit $\mathcal{F}(e_h, e_t, \mathcal{D}) \in \mathbb{R}^{n+1}$ for each entity pair $(e_h, r, e_t)_{h,t\in\{1,\dots,n\},h\neq t,r\in\mathcal{R}}$, where $n = |\mathcal{R}|$, $[\mathcal{F}(e_h, e_t, \mathcal{D})]_i$ denotes the logit for a normal relation for all $1 \leq i \leq n$ and $[\mathcal{F}(e_h, e_t, \mathcal{D})]_{n+1}$ denotes the logit for $\mathcal{NA}$.

The input of the DocRE model is the document information encoding that contains relation, $\psi^{\mathcal{D}} = \left[\psi_1^{\mathcal{D}}, \psi_2^{\mathcal{D}}, \dots, \psi_n^{\mathcal{D}}\right]$. A DocRE model utilizes the logsumexp pooling method (Jia et al., 2019) to compute the embedding of e_i. The model is typically trained by minimizing either the binary cross-entropy (BCE) loss (Yao et al., 2019; Zeng et al., 2020) or the adaptive thresholding loss (ATLoss) (Zhou et al., 2021). During inference, the set of predicted facts $\{(e_h, r, e_t) \mid [\sigma(\mathcal{F}(e_h, e_t, \mathcal{D}))]_r > \delta\}$ is derived by applying a threshold to the predicted probabilities for each entity pair, where δ represents the given threshold and σ is an activation function such as the sigmoid or softmax function. $\varphi_r^{(e_h,e_t,\mathcal{D})}$ = $[\mathcal{F}(e_h, e_t, \mathcal{D})]_r$ is the final predicted logit. We use *ATLoss* as the DocRE model's loss:

$$\mathcal{J}_{cls}^{(e_h,e_t,\mathcal{D})} = -\sum_{r\in\mathcal{R}_p^{\mathcal{D}}} \log \frac{\exp(\varphi_r^{(e_h,e_t,\mathcal{D})})}{\sum_{r'\in\mathcal{R}_p^{\mathcal{D}}\cup\{\mathcal{NA}\}} \exp(\varphi_{r'}^{(e_h,e_t,\mathcal{D})})} - \log \frac{\exp(\varphi_r^{(e_h,e_t,\mathcal{D})})}{\sum_{r'\in\mathcal{R}_n^{\mathcal{D}}\cup\{\mathcal{NA}\}} \exp(\phi_{r'}^{(e_h,e_t,\mathcal{D})})}, \tag{9}$$

where $\mathcal{R}_p^{\mathcal{D}} = r \mid (e_h, r, e_t) \in \mathcal{D}, r \in \mathcal{R}$ and $\mathcal{R}_p^{\mathcal{D}} = r \mid (e_h, r, e_t) \notin \mathcal{D}, r \in \mathcal{R}$. $\mathcal{R}_p^{\mathcal{D}}$ and $\mathcal{R}_n^{\mathcal{D}}$ are positive and negative examples respectively.

Parameter Sharing. To enhance the mutual promotion between the DocRE model and the rule learning module, we adopted the concept of multi-task learning (Zhang and Yang, 2021), allowing the two modules to share parameters. This enables the rules to guide the DocRE model better while minimizing suboptimal results caused by erroneous constraints. **Ca*DRL*** forms $\mathcal{D}$'s corresponding knowledge graph by predicting triples and implements parameter sharing by updating the ruleset through TensorLog. After continuous updates, it returns the RE results and a high-quality ruleset. **Ca*DRL*** is trained by minimizing the total loss:

$$\mathcal{J} = \sum_{i\in D} \sum_{(e_h,e_t)\in\mathcal{C}_i, e_h\neq e_t} \mathcal{J}_{cls}^{(e_h,e_t,\mathcal{D})} + \lambda\cdot\mathcal{J}_{rule}^{(e_h,e_t,\mathcal{D})}, \tag{10}$$

where λ is a hyper-parameter to trade off the two losses.

Method	Dev			Test		
	Ign F1	F1	Logic	Ign F1	F1	Logic
ChatGPT (zero-shot)	-	-	-	-	20.00	-
ChatGPT (5-shot)	-	-	-	-	26.72	-
BiLSTM (Yao et al., 2019)	32.14	39.66	52.24	33.88	43.54	60.53
+LogiRE (Ru et al., 2021)	32.39(+0.25)	40.32(+0.66)	69.24(+17.00)	34.21(+0.33)	43.95(+0.45)	73.13(+12.60)
+MILR (Fan et al., 2022)	34.05(+1.91)	41.22(+1.56)	74.62(+22.38)	35.09(+1.21)	44.65(+1.11)	73.92(+13.39)
+BCBR (Liu et al., 2023)	36.15(+4.01)	42.10(+2.44)	76.47(+24.23)	39.85(+5.97)	47.90(+4.36)	74.65(+14.12)
+**Ca*DRL*** (our work)	38.26(+6.16)	44.02(+4.36)	78.35(+26.11)	42.77(+8.89)	51.43(+7.89)	75.98(+15.45)
GAIN (Zeng et al., 2020)	58.89	63.81	85.25	61.36	67.45	86.85
+LogiRE (Ru et al., 2021)	58.98(+0.09)	64.90(+1.09)	91.25(+6.00)	61.58(+0.22)	68.71(+1.26)	91.71(+4.86)
+MILR (Fan et al., 2022)	61.25(+2.36)	65.85(+2.04)	93.77(+8.52)	62.76(+1.40)	69.23(+1.78)	91.92(+5.07)
+BCBR (Liu et al., 2023)	62.35(+3.46)	65.20(+1.39)	91.50(+6.25)	63.40(+2.04)	69.70(+2.25)	92.15(+5.30)
+**Ca*DRL*** (our work)	63.51(+4.62)	66.49(+2.68)	**96.27(+11.02)**	66.63(+5.27)	70.22(+2.77)	**94.74(+7.89)**
ATLOP (Zhou et al., 2021)	63.37	69.87	86.14	67.29	75.13	88.62
+LogiRE (Ru et al., 2021)	64.54(+1.17)	70.66(+0.79)	90.33(+4.19)	68.13(+0.84)	75.67(+0.54)	91.42(+2.80)
+MILR (Fan et al., 2022)	67.18(+3.81)	72.05(+2.97)	94.85(+8.71)	69.84(+2.55)	76.51(+1.38)	93.16(+4.54)
+BCBR (Liu et al., 2023)	67.42(+4.05)	72.28(+2.41)	93.72(+7.58)	70.02(+2.73)	76.64(+1.51)	93.27(+4.65)
+**Ca*DRL*** (our work)	68.32(+4.95)	74.02(+4.15)	95.03(+8.89)	71.52(+4.23)	78.36(+3.23)	93.82(+5.20)
DREEAM (Ma et al., 2023a)	64.06	70.63	87.18	68.41	77.15	90.17
+LogiRE (Ru et al., 2021)	64.95(+0.89)	71.22(+0.59)	87.69(+0.51)	68.94(+0.53)	77.86(+0.71)	90.87(+0.70)
+MILR (Fan et al., 2022)	67.81(+3.75)	72.67(+2.04)	93.28(+6.10)	69.55(+1.14)	77.56(+0.41)	94.09(+3.92)
+BCBR (Liu et al., 2023)	68.02(+3.96)	72.85(+2.22)	93.91(+6.73)	70.10(+1.69)	77.90(+0.75)	94.15(+3.98)
+**Ca*DRL*** (our work)	**69.03(+4.97)**	**74.52(+3.89)**	95.66(+8.48)	**72.07(+3.66)**	**78.83(+1.68)**	94.29(+4.12)

Table 1: Main results on DWIE (%). (The underlined statistics pass a t-test for significance with p-value < 0.01.)

5 Experiments

5.1 Experimental Setups

We evaluate our approach on three datasets, DWIE (Zaporojets et al., 2021), DocRED (Yao et al., 2019), and HacRED (Cheng et al., 2021). We provide detailed datasets in Appendix A.1. We evaluate our method using three metrics: F1, Ign F1, and Logic. The Ign F1 score excludes triplets that are involved with either the train set or the dev set, thus preventing information leakage from the test set. Logic is used to assess the adherence of our predictions to the golden rule. The detailed description of the baseline model we used is provided in Appendix A.2. We also compared **Ca*DRL*** with LLMs such as ChatGPT (Han et al., 2023), GPT-4 (Peng et al., 2023), and FLAN-UL2 (Peng et al., 2023). We provide detailed hyperparameter settings in Appendix A.3, with all parameters tuned to maximize the Ign F1 score on the development set. We utilized public repositories of backbone models, including BiLSTM[1],GAIN[2], ATLOP[3], and DREEAM[4], to conduct our experiments. The hyperparameter λ for loss balance was set to 1e-5 in all experiments.

5.2 Results & Discussions

We conducted experiments on three datasets, primarily comparing the results on the DWIE dataset with logical labels. The following is an analysis of the results. We denote the enhanced models using +**Ca*DRL*** (resp. +LogiRE, +MILR or +BCBR).

5.2.1 Results on DWIE.

Table 1 displays the results on DWIE, showing that when integrated with four different backbones, **Ca*DRL*** consistently surpasses both LogiRE and MILR in terms of RE and logical consistency. This underscores **Ca*DRL***'s generalizability and compatibility across various backbones. The differentiable rule learning employed by **Ca*DRL*** enhances the specificity of the rule sets, which in turn improves its performance on the test set compared to LogiRE and MILR. Notably, on the DREEAM dataset, **Ca*DRL*** achieves the best-recorded performance to date, enhancing the Ign F1 score by 3.66%, the F1 score by 1.68%, and the Logic score by 4.12%.

Furthermore, we evaluated **Ca*DRL*** against ChatGPT, which employs zero-shot and 2-shot contextual learning (Han et al., 2023). Operating without specific training, ChatGPT depends solely on its general language capabilities and achieves lower in-

[1]https://github.com/thunlp/DocRED
[2]https://github.com/DreamInvoker/GAIN
[3]https://github.com/wzhouad/ATLOP
[4]https://github.com/YoumiMa/dreeam

Method	PLM	DocRED		HacRED	
		Ign F1	F1	Ign F1	F1
ChatGPT (2-shot)	ChatGPT	-	12.40	-	9.79
ChatGPT (5-shot)	ChatGPT	-	32.21	-	26.15
GPT-4 (2-shot)	GPT4	-	27.90	-	20.56
FLAN-UL2 (2-shot)	FLAN-UL2(20B)	-	1.90	-	-
FLAN-UL2 (fine-tuned)	FLAN-UL2(20B)	-	54.50	-	-
ATLOP (Zhou et al., 2021)	$\text{BERT}_{\text{base}}$	57.93	60.53	75.22	76.84
+LogiRE (Ru et al., 2021)	$\text{BERT}_{\text{base}}$	58.62(+0.69)	60.71(+0.18)	75.63(+0.41)	77.39(+0.55)
+MILR (Fan et al., 2022)	$\text{BERT}_{\text{base}}$	59.06(+1.13)	61.23(+0.70)	75.92(+0.70)	77.67(+0.83)
+BCBR (Liu et al., 2023)	$\text{BERT}_{\text{base}}$	60.14(+2.21)	62.08(+1.55)	76.47(+1.25)	78.29(+1.45)
+**Ca*DRL*** (our work)	$\text{BERT}_{\text{base}}$	61.42(+3.49)	62.97(+2.44)	77.03(+1.81)	80.47(+3.63)
DREEAM (Ma et al., 2023a)	$\text{BERT}_{\text{base}}$	59.12	60.91	75.53	77.28
+LogiRE (Ru et al., 2021)	$\text{BERT}_{\text{base}}$	59.85(+0.73)	61.52(+0.61)	75.81(+0.28)	78.02(+0.74)
+MILR (Fan et al., 2022)	$\text{BERT}_{\text{base}}$	60.07(+0.95)	61.79(+0.88)	76.42(+0.89)	78.35(+1.07)
+BCBR (Liu et al., 2023)	$\text{BERT}_{\text{base}}$	61.03(+1.91)	62.35(+1.44)	77.18(+1.65)	79.05(+1.77)
+**Ca*DRL*** (our work)	$\text{BERT}_{\text{base}}$	**62.78(+3.66)**	**64.02(+3.11)**	**79.63(+4.10)**	**81.07(+3.79)**

Table 2: Main results on DocRED and HacRED (%). (The underlined statistics pass a t-test for significance with p-value < 0.01.)

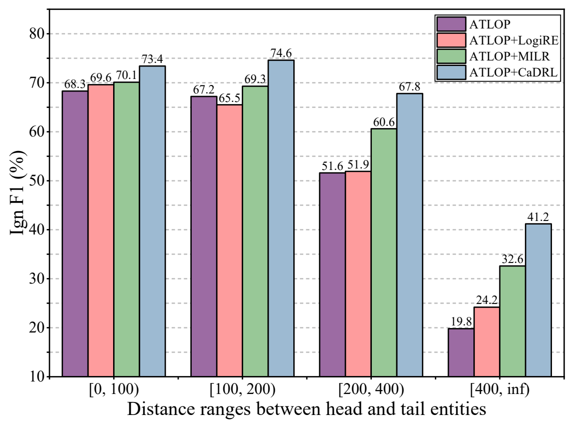

Figure 3: Comparison results for different distances.

ference scores. The intricacy of the DocRE task surpasses ChatGPT's capacity when untrained, highlighting the specialized demands of such tasks.

5.2.2 Results on DocRED and HacRED.

Table 2 illustrates the performance of models under various logical constraints on the DocRED and HacRED test sets. Notably, **Ca*DRL*** has enhanced the F1 score on the DocRED test set by 3.11% over the previously leading MILR framework, which was combined with the DREEAM model, and has increased the Ign F1 score by 3.66%. In contrast, LogiRE did not demonstrate significant improvements on the DocRED dataset, largely due to the high incidence of false negative labels. Previous methods relied on rule sets derived from training set labels, which often proved suboptimal for test and development sets. **Ca*DRL*** tackles this challenge by generating unique rule sets for specific documents where universal rules fall short, ensuring that the rules applied in RE on the test sets are tailored to the current document context. This approach is clearly effective, as evidenced by **Ca*DRL***'s significantly larger gains in the Ign F1 score, which discounts the influence of the training set, compared to the F1 score.

The accurate annotations in HacRED lead to very few false negative labels, thereby minimizing noise introduced by rule constraints. However, due to the limited number of relation categories in HacRED, the potential benefits of rule specificity are not fully realized, with generic rules predominantly used. Moreover, the paucity of relation types means that **Ca*DRL*** does not significantly enhance the Ign F1 score, performing comparably to the F1 score.

Furthermore, we conducted a comparative analysis of **Ca*DRL*** with LLMs such as ChatGPT, GPT-4, and FLAN-UL2 (fine-tuning) (Han et al., 2023; Peng et al., 2023). The results, detailed in Table 2, show that even when fine-tuned, LLMs exhibit suboptimal performance on the DocRED and HacRED datasets. Additionally, the generative nature of LLMs renders them less suitable for DocRE tasks, which are fundamentally classification tasks. See the Appendix C for more discussion.

5.2.3 Results on Long-range Dependencies.

Logical rules offer shortcuts to comprehension. To determine if introducing logical rules aids in capturing long-range dependencies between entity men-

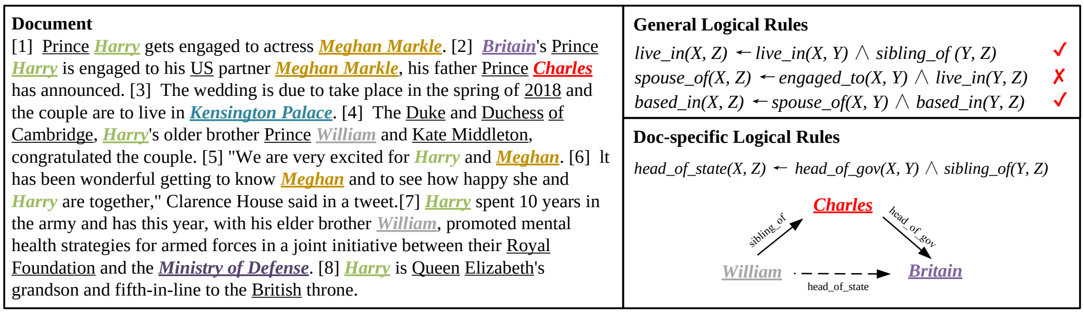

Figure 4: Case Study of ATLOP+**Ca*DRL*** on DWIE. A check mark (✓) denotes the availability of a rule, while a cross (✗) indicates that the rule is not applicable.

Dataset	DWIE		DocRED	
	IgnF1	F1	IgnF1	F1
$L = 1$	70.15	77.94	60.82	62.33
$L = 2$	71.52	78.36	61.42	62.97

Table 3: Comparison on hyper-parameters L.

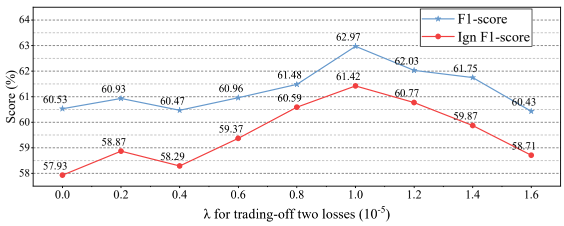

Figure 5: Analysis on the hyper-parameter λ.

tions, we categorized entity pairs into four groups based on the distance between them, defined by the minimum number of tokens separating their mentions within the document. Figure 3 illustrates the comparative results on the DWIE dataset, where ATLOP+**Ca*DRL*** consistently surpasses all baseline models across these groups. Redundant information complicates semantic mapping and limits the potential of representation-based methods. Our approach addresses this by focusing on local logical units, ignoring background noise, and integrating higher-level conceptual connections to derive answers.

5.2.4 Analysis on the impacts of L.

We performed an analysis to assess the impact of the hyperparameters L on the performance of DocRE. Specifically, we created several variants of ATLOP+**Ca*DRL*** with different values for L, and evaluated their performance on the DWIE and DocRED datasets. The results of these comparisons are presented in Table 3.

Dataset	Model	Ign F1	F1
DWIE	**Ca*DRL***	**72.07**	**78.83**
	-encoder	70.81(-1.26)	77.34(-1.49)
	-decoder	69.57(-2.50)	75.79(-3.04)
	-joint training	71.64(-0.43)	78.17(-0.66)
DocRED	**Ca*DRL***	**62.78**	**64.02**
	-encoder	61.23(-1.55)	62.82(-1.20)
	-decoder	60.37(-2.41)	61.28(-2.74)
	-joint training	61.95(-0.83)	63.57(-0.45)

Table 4: Ablation study on the DocRED and DWIE datasets (%).

5.2.5 Results on the Hyper-parameter λ.

We analyzed the hyperparameter λ used for balancing the loss and conducted experiments on the DocRED dataset based on ATLOP+**Ca*DRL***. Figure 5 shows the comparison results. It can be observed that within the λ range of 0 to 6e-5, both the F1-score and the Ign F1-score fluctuate with changes in λ, reaching their maximum values at λ = 1e-5. Therefore, we set λ = 1e-5 in all our experiments.

5.3 Ablation study

We conducted ablation studies on the DWIE and DocRED datasets using DREEAM as the DocRE model, with results presented in Table 4. In this table, "-encoder" indicates a substitution of the current encoder module with that of a standard Transformer, "-decoder" signifies the removal of the decoder, and "-joint training" denotes the elimination of the joint training mechanism. The results demonstrate that our method consistently outperforms the baseline, even when a component is omitted, underscoring the robustness of these elements. The effectiveness of these components is further evidenced by the critical roles played by the quality of the rules and the logical constraints.

5.4 Case Study

We conduct a sample study on DWIE, as shown in Figure 4. It can be seen that for this document, the general rules `spouseOf(X,Y)` $\leftarrow$ `engagedTo(X,Z)` $\wedge$ `liveIn(Z,Y)` will produce incorrect reasoning results in this document. It is necessary to extract exclusive rules for the document and learn doc-specific logical rules.

6 Conclusion and Future work

In this paper, we propose a context-aware differentiable rule learning framework named **Ca*DRL***, aimed at enhancing the inference capabilities of existing DocRE models. Notably, in **Ca*DRL***, we introduce a new encoder and decoder module to simulate the inference of logical rules and adopt a parameter-sharing approach to jointly train the rule constraint module with the DocRE model, thereby learning doc-specific logical rules. Moreover, the effectiveness of **Ca*DRL*** is validated through experimental results on three benchmark datasets. Future work will employ logical rule constraints on LLMs to enhance the capabilities of the rule-learning module and extract more accurate rules.

Limitations

Ca*DRL* may have a major limitation. Since **Ca*DRL*** needs to learn doc-specific rules, it will incur a large time complexity. Therefore, **Ca*DRL*** needs a golden rule set for the training set to reduce its time loss by filtering out the documents with incorrect constraints. We will make up for the above shortcomings in future work to better learn logical rules and constrain the DocRE model.

Ethics Statement

Ca*DRL* is a rule-constrained, interpretable scheme for DocRE tasks. However, employing **Ca*DRL*** in DocRE tasks could potentially expose personal privacy. To mitigate this risk, we restrict our evaluations to public benchmark datasets, which do not contain personally identifiable information.

Acknowledgements

We thank the anonymous reviewers for their feedback on this work. This paper is supported by the Key Program of the National Natural Science Foundation of China (No.U23A20316), Henan Provincial Science and Technology Department's scientific and technological research project (NO.232102211039), the Key Research and Development Program of Henan Province (No.241111212700), the Key Lab of Information Network Security, Ministry of Public Security (No.C23600-04).

References

Qiao Cheng, Juntao Liu, Xiaoye Qu, Jin Zhao, Jiaqing Liang, Zhefeng Wang, Baoxing Huai, Nicholas Jing Yuan, and Yanghua Xiao. 2021. Hacred: A large-scale relation extraction dataset toward hard cases in practical applications. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2819–2831.

William W Cohen. 2016. Tensorlog: A differentiable deductive database. *arXiv preprint arXiv:1605.06523*.

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

Shengda Fan, Shasha Mo, and Jianwei Niu. 2022. Boosting Document-Level Relation Extraction by Mining and Injecting Logical Rules. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10311–10323.

Ridong Han, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan. 2023. Is information extraction solved by chatgpt? an analysis of performance, evaluation criteria, robustness and errors. *arXiv preprint arXiv:2305.14450*.

Xu Han, Pengfei Yu, Zhiyuan Liu, Maosong Sun, and Peng Li. 2018. Hierarchical Relation Extraction with Coarse-to-Fine Grained Attention. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2236–2245, Brussels, Belgium. Association for Computational Linguistics.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging. *arXiv preprint*. ArXiv:1508.01991 [cs].

Robin Jia, Cliff Wong, and Hoifung Poon. 2019. Document-Level N-ary Relation Extraction with Multiscale Representation Learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3693–3704.

Xingzuo Li, Kehai Chen, Yunfei Long, and Min Zhang. 2024. Llm with relation classifier for document-level relation extraction. *arXiv preprint arXiv:2408.13889*.

Yichun Liu, Zizhong Zhu, Xiaowang Zhang, Zhiyong Feng, Daoqi Chen, and Yaxin Li. 2023. Document-level Relationship Extraction by Bidirectional Constraints of Beta Rules. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2256–2266.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint*. ArXiv:1907.11692 [cs].

Youmi Ma, An Wang, and Naoaki Okazaki. 2023a. DREEAM: Guiding Attention with Evidence for Improving Document-Level Relation Extraction. *arXiv preprint*. ArXiv:2302.08675 [cs].

Yubo Ma, Yixin Cao, Yong Hong, and Aixin Sun. 2023b. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10572–10601.

Hao Peng, Xiaozhi Wang, Jianhui Chen, Weikai Li, Yunjia Qi, Zimu Wang, Zhili Wu, Kaisheng Zeng, Bin Xu, Lei Hou, et al. 2023. When does in-context learning fall short and why? a study on specification-heavy tasks. *arXiv preprint arXiv:2311.08993*.

Kunxun Qi, Jianfeng Du, and Hai Wan. 2024. End-to-end learning of logical rules for enhancing document-level relation extraction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7247–7263.

Dongyu Ru, Changzhi Sun, Jiangtao Feng, Lin Qiu, Hao Zhou, Weinan Zhang, Yong Yu, and Lei Li. 2021. Learning Logic Rules for Document-level Relation Extraction. *arXiv preprint*. ArXiv:2111.05407 [cs].

Ali Sadeghian, Mohammadreza Armandpour, Patrick Ding, and Daisy Zhe Wang. 2019. Drum: End-to-end differentiable rule mining on knowledge graphs. *Advances in Neural Information Processing Systems*, 32.

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

Hengzhu Tang, Yanan Cao, Zhenyu Zhang, Jiangxia Cao, Fang Fang, Shi Wang, and Pengfei Yin. 2020. HIN: Hierarchical Inference Network for Document-Level Relation Extraction. In Hady W. Lauw, Raymond Chi-Wing Wong, Alexandros Ntoulas, Ee-Peng Lim, See-Kiong Ng, and Sinno Jialin Pan, editors, *Advances in Knowledge Discovery and Data Mining*, volume 12084, pages 197–209. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Hind Taud and Jean-Franccois Mas. 2018. Multilayer perceptron (mlp). *Geomatic approaches for modeling land change scenarios*, pages 451–455.

Po-Wei Wang, Daria Stepanova, Csaba Domokos, and J Zico Kolter. 2020. Differentiable learning of numerical rules in knowledge graphs. In *ICLR*.

Yijun Wang, Changzhi Sun, Yuanbin Wu, Hao Zhou, Lei Li, and Junchi Yan. 2021. ENPAR:Enhancing Entity and Entity Pair Representations for Joint Entity Relation Extraction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2877–2887, Online. Association for Computational Linguistics.

Benfeng Xu, Quan Wang, Yajuan Lyu, Yong Zhu, and Zhendong Mao. 2021. Entity structure within and throughout: Modeling mention dependencies for document-level relation extraction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14149–14157. Issue: 16.

Zezhong Xu, Peng Ye, Hui Chen, Meng Zhao, Huajun Chen, and Wen Zhang. 2022. Ruleformer: Context-aware rule mining over knowledge graph. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2551–2560.

Zezhong Xu, Peng Ye, Juan Li, Huajun Chen, and Wen Zhang. 2023. Differentiable learning of rules with constants in knowledge graph. *Knowledge-Based Systems*, 275:110686. Publisher: Elsevier.

Lilong Xue, Dan Zhang, Yuxiao Dong, and Jie Tang. 2024. Autore: Document-level relation extraction with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 211–220.

Fan Yang, Zhilin Yang, and William W Cohen. 2017. Differentiable learning of logical rules for knowledge base reasoning. *Advances in neural information processing systems*, 30.

Yuan Yang and Le Song. 2019. Learn to explain efficiently via neural logic inductive learning. *arXiv preprint arXiv:1910.02481*.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A Large-Scale Document-Level Relation Extraction Dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777.

Klim Zaporojets, Johannes Deleu, Chris Develder, and Thomas Demeester. 2021. DWIE: An entity-centric dataset for multi-task document-level information extraction. *Information Processing & Management*, 58(4):102563. Publisher: Elsevier.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation Classification via Convolutional Deep Neural Network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Shuang Zeng, Runxin Xu, Baobao Chang, and Lei Li. 2020. Double Graph Based Reasoning for Document-level Relation Extraction. *arXiv preprint*. ArXiv:2009.13752 [cs].

Yu Zhang and Qiang Yang. 2021. A survey on multi-task learning. *IEEE transactions on knowledge and data engineering*, 34(12):5586–5609.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware Attention and Supervised Data Improve Slot Filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.

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

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14612–14620. Issue: 16.

Weiran Zhu, Xinzhi Wang, Xue Chen, and Xiangfeng Luo. 2024. Refining chatgpt for document-level relation extraction: A multi-dimensional prompting approach. In *International Conference on Intelligent Computing*, pages 190–201. Springer.

A Experimental Setups

A.1 Datasets

To validate the effectiveness and generalization of the DocRE model, the researchers constructed datasets from various domains. Initially, the datasets used to evaluate the DocRE model mainly include TACRED, MUC6, and C77, etc. These datasets have limitations in terms of scale, application domain, and applicability, and they are more limited in terms of the type and number of relations, which makes it difficult to adapt to the needs of DocRE in complex application scenarios. To further promote the research progress of DocRE, it is necessary to establish large, high-quality benchmark datasets for more effective training and evaluation of DocRE models. Currently widely used datasets include DocRED, CDR, GDA, CHR, SCIREX, HacRED, and DWIE. The dataset statistics are shown in Table 5. A summary of the DocRED, DWIE, and HacRED datasets used in this paper is given below:

Statistics	**DWIE**	**DocRED**	**HacRED**
#Train	602	3053	6231
#Dev	98	1000	1500
#Test	99	1000	1500
#Relations	65	97	27
Avg.# entities per Doc.	27.4	19.5	10.8
Avg.# relations per Doc.	24.4	12.5	7.4

Table 5: Statistics of the datasets in experiments.

A.1.1 DWIE

The DWIE (Deutsche Welle Information Extraction Corpus) is a newly developed document-level multitasking information extraction dataset that incorporates four key subtasks: named entity recognition, co-reference resolution, relation extraction, and entity linking (Zaporojets et al., 2021). This study utilizes the dataset solely for DocRE experiments. DWIE is an entity-centered dataset designed to explore entity interactions at the document level, presenting a departure from the prevalent mention-driven approach that typically focuses on detecting and categorizing named entity mentions within individual sentences. The DWIE dataset, sourced randomly from Deutsche Welle's English-language online content, employs annotation schemes that closely mirror the real content, offering a more realistic setting compared to datasets with predetermined annotation schemes and annotations adjusted by non-uniform sampling. Additionally, DWIE includes rule labels that are essential for evaluating the logic of DocRE methods that operate under rule constraints. For our experiments, we exclusively used the dataset, which comprises 802 documents with 23,130 entities, allocating 702 documents for training and 100 for testing.

A.1.2 DocRED

To address the limitation of single-domain focus in DocRE datasets, (Yao et al., 2019) developed a generalized domain dataset based on Wikipedia and Wikidata. This dataset integrates manual annotations with remote supervision, deriving its content from Wikipedia text and Wikidata. The DocRED dataset comprises 101,873 documents obtained through remote supervision and 5,053 documents acquired via manual annotation, with 3,053 designated for training, 1,000 for validation, and 1,000 for testing. Moreover, DocRED encompasses a diverse array of 96 relation types (excluding "$\mathcal{NA}$"), spanning fields such as geography, art, and science.

A.1.3 HacRED

While some existing relation extraction methods perform well on experimental datasets, their results are often less satisfactory in real-world applications. In response to these challenges, the Data Science Laboratory at Fudan University introduced HacRED (Cheng et al., 2021). The HacRED dataset utilizes a case-oriented construction framework specifically designed to create challenging relation extraction datasets. Comprising 9,231 documents that encapsulate 65,225 relational facts across various fields, HacRED is one of the largest DocRE datasets in Chinese and has achieved an F1 score of 96% in terms of data quality.

A.2 Baselines

To assess the generalizability of our method as a plugin model for DocRE, we select four backbone models: BiLSTM (Huang et al., 2015), GAIN (Zeng et al., 2020), ATLOP (Zhou et al., 2021), and DREEAM (Ma et al., 2023a). For fairness, we utilize BERT-base-cased (Devlin et al., 2019) as the pretraining model for GAIN, ATLOP, and DREEAM. We compared our model with other logic constraint DocRE models, namely LogiRE (Ru et al., 2021), MILR (Fan et al., 2022) and BCBR (Liu et al., 2023), simultaneously.

A.3 Hyper-parameter Details

To facilitate the reproduction of our results, we have detailed the hyperparameter settings used in our experiments in Table 6. This table outlines the specific settings for various baseline models and datasets, optimized to maximize the Ign F1 scores on the development set.

A.4 Experimental environment

To make the experiment reproducible, we list the experimental environment in Table 7.

B Proof of TensorLog

Given a document $\mathcal{D}$ containing a set of named entities $\mathcal{E}_{\mathcal{D}} = \{e_i\}_{i=1}^{n_d}$, the task of DocRE involves predicting the relations r between entity pairs $(e_h, e_t)_{h,t\in\{1,\ldots,n_d\},h\neq t}$, where $r \in \mathbb{R}$, and $\mathbb{R}$ represents the set of possible relations, which includes $\mathcal{R}$ (the set of known relations) and $\mathcal{NA}$ (representing "No Relation").

Each entity e_i is represented as a one-hot encoded vector $\mathbf{v}_i \in \{0,1\}^{|\mathcal{E}|}$, where $\mathcal{E}$ is the set of all entities, and the i-th entry is 1, with all other entries being 0. TensorLog introduces an operator $\mathbf{M}^{R_k}$ for each relation R_k, where $\mathbf{M}^{R_k} \in \{0,1\}^{|\mathcal{E}|\times|\mathcal{E}|}$ is a matrix that encodes the presence of a relation $r_{\mathcal{D}}^{(e_i,e_j)}$ between entities e_i and e_j. Specifically, the (e_i, e_j) entry of $\mathbf{M}^{R_k}$ is set to 1 if and only if the relation $r_{\mathcal{D}}^{(e_i,e_j)}$ exists and belongs to $\mathcal{R}$.

We now establish a connection between TensorLog and logical rule inference. Consider a rule of the form:

$$H(x,y) \leftarrow B_1(x,z_1) \wedge B_2(z_1,z_2) \wedge \ldots \wedge B_k(z_{k-1},y), \tag{11}$$

where $B_1, B_2, \ldots, B_k$ are the body predicates. Using the operators described above, we can approximate the rule inference by performing matrix multiplications:

$$\mathcal{S} = \sum_{l=1}^{L}\left(\alpha_l\left(\prod_{k\in\beta_l}\mathbf{M}^{R_k}\mathbf{v}_x\right)\right). \tag{12}$$

In this formulation, the sum is over all possible rules indexed by l, where α_l represents the confidence associated with rule l, and β_l is the ordered list of relations in that rule. The product $\prod_{k\in\beta_l}\mathbf{M}^{R_k}\mathbf{v}_x$ computes the result of applying the relations in the body of the rule to the head entity x. The score vector $\mathcal{S}$ is the weighted sum of these results, with weights α_l corresponding to the confidence of each rule.

Finally, the score for each entity e_j is computed as:

$$\Gamma_{e_j}^{\mathcal{D}} = \mathbf{v}_{e_j}^{T}\cdot\mathcal{S}, \tag{13}$$

where $\mathbf{v}_{e_j}$ is the one-hot encoded vector of entity e_j. The score $\Gamma_{e_j}^{\mathcal{D}}$ indicates how likely entity e_j is to be the correct tail entity for the given head entity e_h, based on the rule-based inference process.

C Discussion on LLMs

In this section, we discuss in detail the comparison of **Ca*DRL*** with current mainstream LLMs, including ChatGPT, GPT-4, and FLAN-UL2. The comparison results on the DWIE and DocRED datasets are presented in Table 1 and Table 2, where the results for LLMs come from (Han et al., 2023;

Hyper-parameter	DWIE				DocRED		HacRED	
	BiLSTM	**GAIN**	**ATLOP**	**DREEAM**	**ATLOP**	**DREEAM**	**ATLOP**	**DREEAM**
Maximum length L	2	2	2	2	2	2	2	2
Optimizer for training	Adam	AdamW	AdamW	AdamW	AdamW	AdamW	AdamW	AdamW
Maximum training epoch	100	100	100	100	100	100	150	150
Learning rate	1e-4	1e-4	1e-4	1e-4	1e-4	1e-4	2e-5	2e-5
Batch size for training	4	4	4	4	4	4	4	4
Dropout rate	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.1
λ for trading-off losses	1e-5	1e-5	1e-5	1e-5	1e-5	1e-5	1e-5	1e-5

Table 6: Comparison of hyper-parameters across different models.

Option	Setting
OS	Ubuntu 20.04
CUDA	11.8
GPU	RTX 4090
GPU driver	525.85.12
System memory	256GB@40 cores
Language	python 3.9
Deep learning framework	PyTorch 2.0.1

Table 7: Experimental environment.

Peng et al., 2023). The results indicate that there is a certain performance gap between LLMs and DREEAM+**Ca*DRL*** on both datasets. It is also observed that the performance of FLAN-UL2 significantly improves after fine-tuning on DocRED, suggesting that LLMs with limited-sample in-context learning (ICL) struggle to leverage the full domain knowledge in the training data. Furthermore, even after fine-tuning FLAN-UL2 on the training data, our method remains significantly superior to FLAN-UL2. This is because, compared to the **Ca*DRL***-enhanced DocRE model as a relation classification model, FLAN-UL2 cannot adapt to the classification task of DocRE. There is a significant gap between the generative training objectives and discriminative training objectives of classification tasks. Additionally, LLMs themselves have issues with hallucinations, which may lead to unexpected relations being predicted as the final outcome. This problem currently cannot be fully resolved through fine-tuning.

However, combining **Ca*DRL*** with LLMs is a prospective method for further improving performance. LLMs' few-shot ICL does not generalize well in information extraction tasks (Ma et al., 2023b), but LLMs can solve some difficult cases. Therefore, we can use existing DocRE models to handle most simple cases and use LLMs for difficult cases that DocRE models cannot handle, allowing a two-stage relation extraction process to help **Ca*DRL*** adapt to knowledge-intensive scenarios. Additionally, LLMs can generate logical inference rules by using relational paths as input. Thus, the rules generated by LLMs can serve as a golden rule set to initialize the rule constraint module. This can help **Ca*DRL*** learn more logical reasoning rules, thereby achieving better convergence and performance. To cope with dynamically changing environments in practical scenarios, LLMs' generative capabilities can be used to update their knowledge bases and serve as the rule set for the DocRE model in real-time. This allows for continuous learning from new data, adapting to environmental changes, and improving decision-making accuracy and adaptability.

D Discussion on More Applications

Ca*DRL* is a differentiable rule learning framework for jointly training specific neural models and logical rules. Therefore, we believe that **Ca*DRL*** can be used in more application scenarios that use logical rules. For example, **Ca*DRL*** can be applied to other information extraction tasks such as document-level event extraction, document-level aspect-level sentiment analysis, and document-level event causal relationship identification. The exploration of **Ca*DRL*** in these applications is also part of our future work.