An Ensemble-of-Experts Framework for Rehearsal-free Continual Relation Extraction

Shen Zhou[1], **Yongqi Li**[1], **Xin Miao**[1], **Tieyun Qian**[1,2*]
[1] School of Computer Science, Wuhan University, China
[2] Intellectual Computing Laboratory for Cultural Heritage, Wuhan University, China
{shenzhou,liyongqi,miaoxin,qty}@whu.edu.cn

Abstract

Continual relation extraction (CRE) aims to continuously learn relations in new tasks without forgetting old relations in previous tasks. Current CRE methods are all rehearsal-based, which need to store samples and thus may encounter privacy and security issues. *This paper targets rehearsal-free continual relation extraction for the first time* and decomposes it into task identification and within-task prediction sub-problems. Existing rehearsal-free methods focus on training a model (expert) for within-task prediction yet neglect to enhance the models' capability of task identification.

In this paper, we propose an **E**nsemble-**o**f-**E**xperts (EoE) framework for rehearsal-free continual relation extraction. Specifically, we first discriminatively train each expert by augmenting analogous relations across tasks to enhance the expert's task identification ability. We then propose a cascade voting mechanism to form an ensemble of experts for effectively aggregating their abilities. Extensive experiments show that our method outperforms current rehearsal-free methods and is even better than rehearsal-based CRE methods. [1]

1 Introduction

Relation extraction (RE) aims to identify the relation between two entities mentioned in the sentence, which is essential for many tasks like knowledge graph construction (Peng et al., 2020). Traditional RE methods based on pre-defined relation sets have achieved remarkable performance but cannot handle emerging relations in the real world (Han et al., 2020). Hence, continual relation extraction (CRE) is proposed (Wang et al., 2019) to deal with such a problem, with the goal of continuously learning new relations while not forgetting previously learned ones.

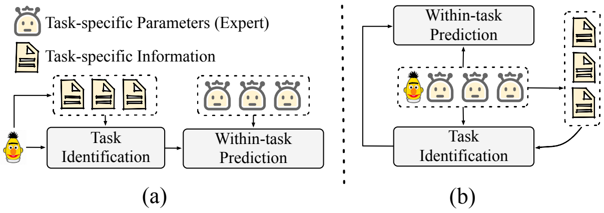


Figure 1: Comparison between (a) existing rehearsal-free methods for other tasks and (b) our rehearsal-free method for the RE task.

Existing CRE methods need to store a few typical samples for each learned relation and then utilize memory replay to avoid the catastrophic forgetting problem in continual learning. However, due to privacy and security issues, storing training data is not allowed in sensitive areas such as finance and biology (Wang et al., 2023d). In light of this, we consider the rehearsal-free continual relation extraction (RFCRE) problem for the first time, for which we expect to incrementally learn new relations without storing any training data.

Following the pioneering work in continual learning (Kim et al., 2022), we decompose RFCRE into two sub-problems: task identification and within-task prediction. Benefiting from parameter-efficient tuning (Li and Liang, 2021; Hu et al., 2021; Ding et al., 2022), the within-task prediction problem has been well addressed by saving the task-specific parameters of each task (named as an expert). Then, the core challenge lies in the task identification problem, i.e., how to identify the task-id for each test sample.

Along this line, Wang et al. (2023d) employs a pre-trained language model (PLM) (Devlin et al., 2019) based method for task identification in continual text classification. However, when applying this method to the RE task, we find its performance is not very satisfying. The main reason is that only one PLM is involved in task identification, as shown in Fig. 1(a). The knowledge in PLM

[*] Corresponding author.
[1] Our code and data are available at https://github.com/NLPWM-WHU/EoE-CRE.

Findings of the Association for Computational Linguistics: ACL 2024, pages 1410–1423
August 11-16, 2024 ©2024 Association for Computational Linguistics

might be enough for the text classification task but is insufficient for the RE task, whose performance relies heavily on task-related knowledge.

To address this problem, we propose an **E**nsemble-**o**f-**E**xperts (EoE) framework for rehearsal-free continual relation extraction. Our goal is to aggregate the experts of different tasks from different stages during the continual learning who already possess rich task-related knowledge for both within-task prediction and task identification, as shown in Fig. 1 (b). Despite the appealing target, we now face two new research questions.

[RQ1]: How to increase each expert's task identification ability besides the original within-task prediction ability?

[RQ2]: How to effectively aggregate multiple experts' abilities by assigning the right experts for the given task?

For [RQ1], we find the challenge in task identification mainly comes from the analogous relations across tasks, e.g., `father` and `sibling`, and it is pretty hard for an expert to discern such relations. In view of this, we introduce a discriminative training (Zhan et al., 2021) approach to increase each expert's task identification ability. Specifically, we develop a novel relation augmentation method to generate two types of analogous relations. One is to reverse the positions of two entities in an old relation (Han et al., 2021; Wang et al., 2022a). The other is to remove the context between two entities in an old relation. Since these augmented relations have not been seen by the expert of the current task, they will force the expert to distinguish old relations from the augmented ones during the discriminative training process.

For [RQ2], we find that directly aggregating the votes from all experts will bring about a transboundary risk. This is because the scope of task identification varies among different experts due to the temporal nature of continuous learning. For example, the t^{th} expert has not seen the $(t-1)^{th}$ task, so the $(t-1)^{th}$ task is out of the scope of the t^{th} expert. To handle this, we propose a cascade voting mechanism consisting of a two-phase voting procedure. Specifically, in the first phase, we use the first expert and the PLM to make an initial decision. Then, in the second phase, we dynamically aggregate votes from qualified experts for the final decision.

Our main contributions are as follows:

- To the best of our knowledge, we are the first to propose the rehearsal-free continual relation extraction (RFCRE) problem.

- We present an ensemble-of-experts framework for the RFCRE task, which aggregates experts' knowledge from different stages in continual learning for both task identification and within-task prediction.

- Extensive experimental results on two datasets show that our method not only outperforms the state-of-the-art rehearsal-free methods but also achieves better performance than rehearsal-based CRE methods.

2 Related Work

Continual Learning (CL) The goal of continual learning is to extend knowledge from a continuous data stream without catastrophic forgetting (Wang et al., 2023a). Existing CL methods can be roughly categorized into three groups: (a) *Regularization-based methods* mitigate catastrophic forgetting by introducing parameter regularization terms (Kirkpatrick et al., 2017) or knowledge distillation loss (Li and Hoiem, 2017), (b) *Rehearsal-based methods* store a small set of old training samples to assist the model in replaying learned knowledge. Current CRE methods (Wang et al., 2019; Han et al., 2020; Cui et al., 2021; Zhao et al., 2022; Hu et al., 2022; Wang et al., 2022a; Xia et al., 2023; Zhao et al., 2023; Xiong et al., 2023; Song et al., 2023; Nguyen et al., 2023) mainly focus on retaining or recovering the old relation knowledge during the memory replay stage, such as prototype learning (Cui et al., 2021), curriculum learning (Wu et al., 2021), knowledge distillation (Zhao et al., 2022), contrastive learning (Song et al., 2023). (c) *Network-based methods* dynamically expand or split the network parameters to learn new tasks. Thanks to the development of parameter-efficient tuning (PET) methods (Li and Liang, 2021; Liu et al., 2022; Hu et al., 2021), several studies (Wang et al., 2023d,c, 2022c) try to assign task-specific parameters to each task and then query the task-id for the given test sample.

In contrast to existing rehearsal-based CRE methods, we target rehearsal-free CRE, which is the first attempt in this field.

Task decomposition in CL Based on the parameter-efficient tuning technique, several CL

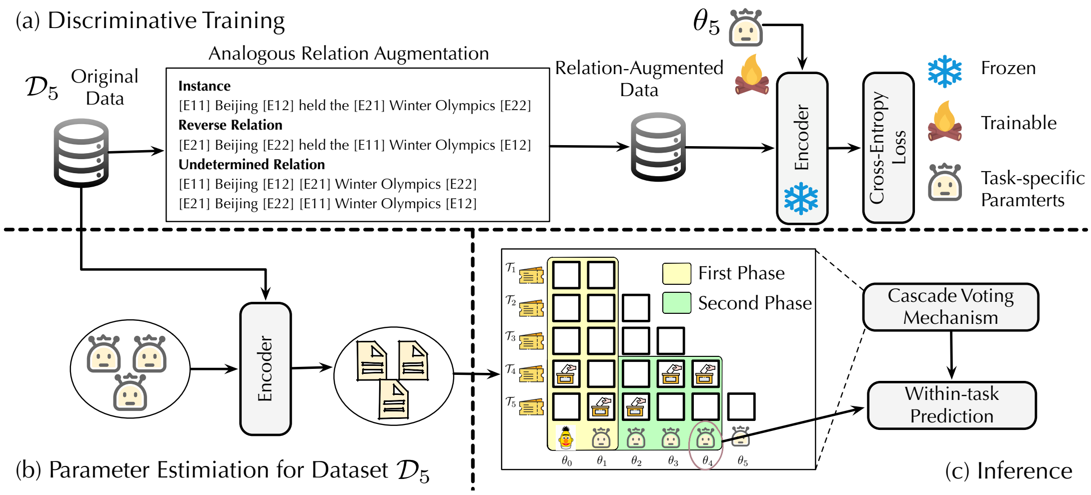

(a) Discriminative Training
Analogous Relation Augmentation
$\mathcal{D}_5$ Original Data
Instance
[E11] Beijing [E12] held the [E21] Winter Olympics [E22]
Reverse Relation
[E21] Beijing [E22] held the [E11] Winter Olympics [E12]
Undetermined Relation
[E11] Beijing [E12] [E21] Winter Olympics [E22]
[E21] Beijing [E22] [E11] Winter Olympics [E12]
Relation-Augmented Data
θ_5
Encoder
Cross-Entropy Loss
Frozen
Trainable
Task-specific Paramterts
Encoder
(b) Parameter Estimiation for Dataset $\mathcal{D}_5$
$\mathcal{T}_1$
$\mathcal{T}_2$
$\mathcal{T}_3$
$\mathcal{T}_4$
$\mathcal{T}_5$
First Phase
Second Phase
θ_0 θ_1 θ_2 θ_3 θ_4 θ_5
Cascade Voting Mechanism
Within-task Prediction
(c) Inference

Figure 2: Overall framework of the proposed method. We take task $\mathcal{T}_k$ (k=5) for illustration.

methods decompose the CL problem into task identification and within-task prediction (Kim et al., 2022). For example, L2P (Wang et al., 2022d) uses the PLM to retrieve a few prompts from a fixed prompt pool. EPI (Wang et al., 2023d) uses Mahalanobis distance for task identification, and ESN (Wang et al., 2023c) uses energy scores to conduct identification. Several studies directly summate (Wang et al., 2023b) and concatenate (Razdaibiedina et al., 2023) parameters for task identification.

Different from the above methods, we effectively aggregate experts from different stages for task identification via the discriminative training and cascade voting mechanism tailored for this purpose.

3 Problem Formulation

Relation extraction aims to determine whether the relation y holds for the head and tail entity in the given instance x. Continual relation extraction (CRE) aims to train the model from a series of class-incremental relation extraction tasks $\{\mathcal{T}_1, \mathcal{T}_2, \cdots, \mathcal{T}_N\}$, where N is the number of tasks, each task $\mathcal{T}_k$ contains its training set $\mathcal{D}_k = \{(x_i, y_i)\}_{i=1}^{N_k}$ and relation set $\mathcal{R}_k$. Specifically, x_i is an instance of relation $y_i \in \mathcal{R}_k$, N_k is the number of $\mathcal{D}_k$ and $\mathcal{R}_k$ is the label space of $\mathcal{T}_k$, where $\mathcal{R}_i \cap \mathcal{R}_j = \emptyset$ for $i \neq j$. The CRE methods aim to learn new relations from $\mathcal{D}_k$ and not forget previous relations. In other words, the trained model needs to classify all observed relations $\hat{\mathcal{R}}_k = \cup_{i=1}^{k} \mathcal{R}_i$ correctly after learning task $\mathcal{T}_k$. Unlike existing CRE methods that store a small set of representative samples to avoid catastrophic forgetting via memory replay, we focus on *rehearsal-free* setting without saving any original training samples to prevent potential privacy safety issues.

4 Methodology

As shown in Fig. 2, when the new task $\mathcal{T}_k$ comes, the overall framework is divided into three steps: (1) *Discriminative Training*: This step strengthens experts' task identification ability when encountering inter-task analogous relations. (2) *Parameter Estimation for Dataset $\mathcal{D}_k$*: This step stores some statistics about the current training dataset via experts $\{\theta_i\}_{i=0}^{k}$ for later task identification. (3) *Inference*: This step aims to aggregate the identification ability of different experts and then makes within-task predictions.

4.1 Discriminative Training

The core reason for the stability-plasticity dilemma in continual learning is the interference between old and new tasks (Wang et al., 2023a). Intuitively, the simplest way to eliminate it is to train separately with independent models, which may result in huge resource consumption. Benefiting from the PET methods, e.g., Prefix Tuning (Li and Liang, 2021; Liu et al., 2022), LoRA (Hu et al., 2021), we can insert a few additional parameters θ_k into the model and then only tune the inserted parameters to achieve the performance close to the full fine-tuning. In this work, we use LoRA (A brief introduction can be found in Appendix B) to obtain θ_k for the current task.

However, previous work (Wang et al., 2023d) only optimizes the θ_k based on the dataset $\mathcal{D}_k$. Although it is sufficient for within-task prediction, it does not perform satisfactorily on task identification when encountering inter-task analogous relations. The reason is that the lack of supervision for unknown analogous relations causes the expert to learn shortcuts for classifying relations $\mathcal{R}_k$ based on the dataset $\mathcal{D}_k$. Hence, we introduce analogous relation augmentation to generate new relations based on the known relations to avoid expert learning from shortcuts. Then, we formalize $|\hat{\mathcal{R}}_k|$-way ($|\hat{\mathcal{R}}_k| > |\mathcal{R}_k|$) classification task via merging generated new relations and known relations to obtain the expert θ_k.

① **Analogous Relation Augmentation** The key to relation data augmentation lies in how to construct challenging negative samples. Peng et al. (2020) have pointed out that relation extraction models tend to rely on shallow features leaked from entity mentions for relation classification. Thus, we apply two strategies to encourage the expert to have an in-depth understanding of entity and context information, including the reverse relation augmentation (Wang et al., 2022a) and undetermined relation augmentation.

Following Wang et al. (2022a), given an instance x from dataset $\mathcal{D}_k$, we first determine whether its relation is symmetric [2]. If the relation is not symmetric, we can construct a new relation by swapping the markers of the head and tail entities. For example, for the sentence "`[E11]` Beijing `[E12]` held the `[E21]` Winter Olympics `[E22]`", the relation changed from *host* to *place of hosting* when we switched the markers of the head and tail entities. After that, we denote the dataset augmented by this strategy as $\mathcal{D}_k^r$.

Wang et al. (2022b) constructs counterfactual examples by removing the context to eliminate entity bias. This implies a hypothesis: *If an instance only retains entity information and removes the context information, there should be no relation between the entities.* For example, it is hard for us to determine the relation between *Beijing* and *Winter Olympics* without context information; the relation may be *host*, *participate*, or *apply*. In view of this, we build a new relation by removing the context information of the instances named `undetermined_relation`. Note that this strategy is applied to both $\mathcal{D}_k$ and $\mathcal{D}_k^r$, and we denote the dataset with `undetermined_relation` as $\mathcal{D}_k^u$.

② **Training** After constructing the datasets $\mathcal{D}_k^r$ and $\mathcal{D}_k^u$, we insert them into the original dataset. For convenience, we denote relation-augmented dataset as $\hat{\mathcal{D}}_k = \mathcal{D}_k \cup \mathcal{D}_k^r \cup \mathcal{D}_k^u$ and merged relation set as $\hat{\mathcal{R}}_k$, where the maximum size of $\hat{\mathcal{R}}_k$ is $2 * |\mathcal{R}_k| + 1$. Given an input instance $x = \{w_1, w_2, \cdots, w_n\} \in \hat{\mathcal{D}}_k$ with the corresponding entity pair, we insert two pairs of special tokens `[E11]`/`[E12]` and `[E21]`/`[E22]` at the beginning and end of the head entity and the tail entity, respectively. Except for the first task, we inject additional parameters θ_k into BERT (Devlin et al., 2019) as backbone to encode the input to obtain the contextual representations $f_{\theta_k}(x) = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_n\}$. Then, we directly concatenate the representation of `[E11]` and `[E21]` as the relational representation of x_i, which is defined as

$$\boldsymbol{h}_i = \boldsymbol{w}_{\texttt{[E11]}} \oplus \boldsymbol{w}_{\texttt{[E21]}}. \tag{1}$$

Then, we initialize the relation embeddings $\phi_k \in \mathbb{R}^{|\hat{\mathcal{R}}_k| \times d}$ as $|\hat{\mathcal{R}}_k|$-way classifier weight to classify both old relations and new relations. Therefore, we can optimize θ_k and ϕ_k using cross-entropy loss, with the training objective as follows:

$$\mathcal{L}_{\text{ce}}(\theta_k, \phi_k) = -\frac{1}{|\hat{\mathcal{D}}_k|} \sum_{i=1}^{|\hat{\mathcal{D}}_k|} \sum_{j=1}^{|\hat{\mathcal{R}}_k|} \mathbb{I}(y_i = r_j) \log P_{\theta_k}(r_j | x_i), \tag{2}$$

where $P_{\theta_k}(x_i) = softmax(\boldsymbol{h}_i \cdot \phi_k)$ and y_i denotes the gold relation [3]. Note that we use full fine-tuning for the first task to provide a good initialization for subsequent tasks.

4.2 Parameter Estimation for Dataset $\mathcal{D}_k$

After training the k-th task, we need to save task-specific information for subsequent task identification due to the unavailability of the previous task data. Inspired by Wang et al. (2023d), we assume that the class-conditional distribution follows the multivariate Gaussian distribution based on Gaussian discriminant analysis. Then, with the task dataset $\mathcal{D}_k$ and θ_j, we can utilize the maximum likelihood estimator to estimate $|\mathcal{R}_k|$ class-conditional Gaussian distributions $\mathcal{N}(\mu_{k,c}^j, \Sigma_k^j)$ with a shared covariance Σ_k^j:

$$\mu_{k,c}^j = \frac{1}{|\mathcal{D}_{k,c}|} \sum_{y_i = c} \boldsymbol{h}_i, c \in \mathcal{R}_k, \tag{3}$$

$$\Sigma_k^j = \frac{1}{|\mathcal{D}_k|} \sum_{c} \sum_{y_i = c} (\boldsymbol{h}_i - \mu_{k,c}^j)(\boldsymbol{h}_i - \mu_{k,c}^j)^\top, \tag{4}$$

[2] Symmetric relations imply that swapping the order of the head and tail entities does not change the relation, such as `per:siblings`.

[3] After training the k-th task, we discard the relation embeddings of augmented relations.

where $\boldsymbol{h}_i = f_{\theta_j}(x_i)$ and $\mathcal{D}_{k,c}$ denotes samples with a relation c in D_k. In practice, we further share the covariance of seen tasks by θ_j to avoid numeric deviation:

$$\Sigma^j = \frac{1}{k-j+1} \sum_{i=j}^{k} \Sigma_i^j \tag{5}$$

4.3 Inference

Currently, we obtain experts $\{\theta\}_{i=1}^k$ for each task, but we cannot access the task-id of each test instance at the testing stage. Thus, when given a test instance, we must decide which expert to use for within-task prediction. Wang et al. (2023d) rely on the class-conditional Gaussian distribution estimated by the original PLM and uses Mahalanobis distance (Lee et al., 2018) for task identification. However, the knowledge in PLM is insufficient for RE, and it also ignores the potential identification capability of experts with task-related knowledge from different stages for task identification.

Limited by the sequential temporal nature of continuous learning, directly aggregating votes from different experts suffers from transboundary risk. For example, the 2^{nd} expert has not seen the first task, so the first task is out of the scope of θ_2. Hence, we proposed a cascade voting mechanism to overcome it. For convenience, we sequentially explain the voting protocol of θ_k, cascade voting mechanism, and within-task prediction.

① **Voting Protocol of θ_k** Firstly, we map the test instance x into the representation $\boldsymbol{h}$ by the expert model $\boldsymbol{h} = f_{\theta_k}(x)$. Then, with the above induced class-conditional Gaussian distributions $\{\mathcal{N}(\mu_{i,c}^k, \boldsymbol{\Sigma}^k)\}_{c=1}^{|\mathcal{R}_i|}$ of task $\mathcal{T}_i$ by expert θ_k, we can define the confidence score through computing the Mahalanobis distance between $\boldsymbol{h}$ and $\{\mathcal{N}(\mu_{i,c}^k, \boldsymbol{\Sigma}^k)\}_{c=1}^{|\mathcal{R}_i|}$:

$$\mathcal{C}_k^{\mathcal{T}_i}(x) = \min_{c} \left(\boldsymbol{h} - \mu_{i,c}^k\right)^\top {\boldsymbol{\Sigma}^{k}}^{-1} \left(\boldsymbol{h} - \mu_{i,c}^k\right), i \geq k, \tag{6}$$

where $\mathcal{C}_k^{\mathcal{T}_i}(x)$ denotes the confidence score of task $\mathcal{T}_i$ by θ_k. Then, the voting result $\mathcal{V}_k^x$ of expert k can be defined as:

$$\mathcal{V}_k^x = \arg\min_i \mathcal{C}_k^{\mathcal{T}_i}(x), i \geq k. \tag{7}$$

.

② **Cascade Voting Mechanism** As shown in Fig. 2, if we directly allow experts at each stage to participate in voting, the available voting range for each expert is inconsistent. To this end, we propose a cascade voting mechanism to solve the transboundary risk. Specifically, we first introduce the original PLM θ_0 as an additional expert[4] to form a fair voting stage with θ_1 in the first phase. Then, based on the voting results of the first phase, we dynamically aggregate votes from qualified experts from the final decision in the second phase. The detailed procedure of the cascade voting mechanism for the task $\mathcal{T}_k$ is shown in Algorithm 1, which mainly contains two phases:

Algorithm 1 Cascade Voting Mechanism

```
Require: Voting results $\mathcal{V}_0^x$ and $\mathcal{V}_1^x$ of expert $\theta_0$ and $\theta_1$, Maximum allowed experts $m$, Confidence scores $\{\mathcal{C}_j^{\mathcal{T}_i}(x) | 2 \leq j \leq k-1, j \leq i \leq k\}$
Output: Voting result $\mathcal{V}^x$
 1: if $\mathcal{V}_0^x = \mathcal{V}_1^x$ then
 2:     return $\mathcal{V}_0^x$;
 3: end if
 4: $\mathrm{start} = 0$;
 5: $\mathrm{end} = \min(\min(\mathcal{V}_0^x, \mathcal{V}_1^x), m)$;
 6: $\mathrm{result} = \mathrm{dict}()$;
 7: for $e \in [\mathrm{start}, \mathrm{end}]$ do
 8:     $\mathcal{V}_e^x = \arg\min_i \mathcal{C}_e^{\mathcal{T}_i}(x), i \geq \mathrm{end}$;
 9:     $\mathrm{result}[\mathcal{V}_e^x]$ += 1;
10: end for
11: return $\arg\max_c \mathrm{result}[c]$;
```

First Phase (line 1 $\sim$ 5): If the voting results of θ_0 and θ_1 are consistent, we directly return the results. If inconsistent, the minimum value of the voting result $\min(\mathcal{V}_0^x, \mathcal{V}_1^x)$ is taken as the maximum range of allowed voting experts. Noted that to avoid the surge in the number of voting experts as tasks increase, we introduce an additional threshold m to limit the maximum number of experts selected.

Second Phase (line 6 $\sim$ 11): For each expert $\theta_j, j \in [\mathrm{start}, \mathrm{end}]$ is allowed to vote, we frame its voting range at $[\mathrm{end}, k]$ to ensure the fairness, and calculate the task with the highest votes as the voting result.

③ **Within-task Prediction** After obtaining the voting result $\mathcal{V}^x$, we can use the expert $\theta_{\mathcal{V}^x}$ with the corresponding relation embeddings $\phi_{\mathcal{V}^x}$ to predict the relation of the given instance x.

5 Experiments

5.1 Experimental Setup

Datasets We employ two widely used relation extraction datasets, i.e., TACRED (Han et al., 2018) and FewRel (Han et al., 2018), to evaluate the performance of our proposed method. To ensure a fair comparison, we follow the order and the division of tasks in Cui et al. (2021) by dividing

[4] To adapt the PLM to the relation extraction, we have tried several different representation extraction methods and selected Entity Avg. finally. Please see Appendix A for details.

FewRel	Mem.	$\mathcal{T}_1$	$\mathcal{T}_2$	$\mathcal{T}_3$	$\mathcal{T}_4$	$\mathcal{T}_5$	$\mathcal{T}_6$	$\mathcal{T}_7$	$\mathcal{T}_8$	$\mathcal{T}_9$	$\mathcal{T}_{10}$
CEAR‡	$10 * \lvert\hat{\mathcal{R}}_k\rvert$	98.1	95.8	93.6	91.9	91.1	89.4	88.1	86.9	85.6	84.2
CDec‡		98.2	94.9	93.2	91.9	91.3	89.6	88.3	87.1	86.0	84.6
RationaleCL‡		98.6	95.7	93.4	92.3	91.3	89.7	88.2	87.3	86.3	85.1
CFDR ‡		98.3	94.7	93.1	91.4	90.6	89.4	87.9	86.9	85.4	84.3
InfoCL‡		98.3	95.2	93.4	92.1	91.3	89.7	88.5	87.7	86.8	85.4
FT	0	**98.4**	94.1	90.4	84.8	82.6	79.3	76.6	72.9	68.4	64.6
EWC		98.4	94.4	90.2	85.5	82.7	79.6	77.9	74.1	70.5	66.7
LwF		98.4	93.5	88.3	80.6	73.8	68.0	61.1	54.0	48.1	43.8
L2P		97.4	90.8	83.6	76.5	68.9	64.1	61.0	57.4	50.1	44.6
EPI (Prefix)		97.5	94.7	92.5	91.3	90.0	88.1	86.6	85.0	83.7	81.9
EPI (LoRA)		97.3	94.9	92.7	91.4	90.2	88.3	86.8	85.1	83.8	82.1
EoE	0	97.8	**95.0**	**93.6**	**92.5**†	**91.6**†	**90.0**†	**88.9**†	**87.9**†	**86.9**†	**85.5**†

TACRED	Mem.	$\mathcal{T}_1$	$\mathcal{T}_2$	$\mathcal{T}_3$	$\mathcal{T}_4$	$\mathcal{T}_5$	$\mathcal{T}_6$	$\mathcal{T}_7$	$\mathcal{T}_8$	$\mathcal{T}_9$	$\mathcal{T}_{10}$
CEAR‡	$10 * \lvert\hat{\mathcal{R}}_k\rvert$	97.7	94.3	92.3	88.4	86.6	84.5	82.2	81.1	80.1	79.1
CDec‡		97.9	93.1	90.1	85.8	84.7	82.6	81.0	79.6	79.5	78.6
RationaleCL‡		98.6	94.4	91.5	88.1	86.5	84.9	84.5	82.5	81.6	80.8
CFDR ‡		98.1	93.8	89.8	85.8	84.4	83.4	81.6	79.9	79.7	79.1
InfoCL‡		96.3	92.4	88.9	87.3	83.9	82.4	82.0	79.7	78.4	78.2
FT	0	98.5	90.6	78.1	73.1	71.2	65.6	61.0	58.1	55.0	50.5
EWC		98.5	90.1	79.0	73.3	69.8	65.8	62.6	59.2	55.7	50.2
LwF		98.5	88.7	71.5	64.5	60.4	53.3	48.9	45.4	42.9	37.4
L2P		96.9	88.2	73.8	68.6	66.3	63.1	60.4	59.1	56.8	54.8
EPI (Prefix)		98.0	94.5	89.4	86.4	85.7	84.5	82.9	82.0	81.7	80.4
EPI (LoRA)		97.8	**94.7**	89.0	85.8	85.4	84.3	82.7	81.6	81.4	80.0
EoE	0	**98.7**	**94.7**	**90.6**	**87.8***	**87.2***	**85.9***	**84.3**	**83.2***	**82.7***	**81.5***

Table 1: Classification accuracy (%) on all observed relations after learning each task. The baseline results with "‡" are retrieved from Zhao et al. (2023); Song et al. (2023) while other results are reproduced using their released code. The best and the second best accuracy scores under the rehearsal-free setting are in **bold** and underlined, respectively. The best accuracy score under rehearsal-based settings is in wave. † and $*$ denote the statistically significant improvements with $p < 0.01$ and $p < 0.05$ over the results by the best rehearsal-free baseline EPI.

each dataset into 10 sub-datasets according to relations, each representing a task. Please refer to the Appendix C.1 for the details of the datasets.

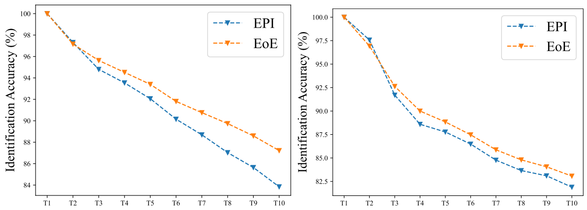
Metrics We use two metrics for evaluating the performance. **Identification Accuracy (IA)** denotes the accuracy of the task identification; **Classification Accuracy (CA)** denotes the relation classification accuracy on observed relations, which serves as the main metric.

Baselines We first compare our proposed EoE method with rehearsal-free continual learning methods. Since there is no existing RFCRE method, we migrate four rehearsal-free CL methods from other fields to the relation extraction task, and divide them into two groups: (1) **Regularization-based methods**: EWC (Kirkpatrick et al., 2017) and LwF (Li and Hoiem, 2017). (2) **Network-based methods**: L2P (Wang et al., 2022d), and EPI (Wang et al., 2023d). Both network-based methods are implemented with prefix-tuning. For a more fair comparison, we extend EPI with a LoRA implementation.

We then compare our EoE method with five latest **rehearsal-based continual relation extraction (CRE)**, including CEAR (Zhao et al., 2023), CDec (Xia et al., 2023), RationaleCL (Xiong et al., 2023), CFDR (Nguyen et al., 2023), and InfoCL (Song et al., 2023). [5]

Implementation Details We implement the framework based on Pytorch (Paszke et al., 2019) and Huggingface (Wolf et al., 2020) and use `bert-base-uncased` (Devlin et al., 2019) as the backbone. For the first task, we use full fine-tuning, with the learning rates of the backbone and classifier set to 1e-5 and 1e-3, respectively. the number of epochs is set to 10, and the Adamw optimizer is employed. For the subsequent

[5] Please refer to Appendix C.2for the details of baselines.



(a) FewRel (b) TACRED

Figure 3: Task identification accuracy (%) on all observed tasks after learning each task.

tasks, we freeze the backbone's parameters and use LoRA (Hu et al., 2021) to assign a small number of task-specific parameters to each task, with its learning rate set to 5e-4, rank set to 8, the alpha parameter for Lora scaling set to 16, and dropout set to 0.1. The number of epochs is set to 5, and the learning rate of the classifier is set to 3e-2. All experiments were conducted on an NVIDIA RTX 3090 GPU, and all results were averaged by taking 5 different task sequences.

For a more comprehensive comparison experiment, we migrated four rehearsal-free methods to relation extraction. Since the task boundaries are known during training, we mask non-current task relations at the training for FT, EWC, LwF, and L2P. For EPI, which has the closest performance, we additionally implement LoRA (Hu et al., 2021) to guarantee a fairer comparison.

5.2 Main Results

Table 1 presents the comparison results of our proposed EoE method and baselines. Based on the results, we have the following findings.

(1) Our EoE method achieves state-of-the-art performance in almost all cases. It surpasses the latest rehearsal-based methods focusing on relation extraction and the latest rehearsal-free methods adapted from other tasks. Moreover, on the last $\mathcal{T}_{10}$ on FewRel, even the performance of the best rehearsal-free method EPI is significantly worse than all rehearsal-based methods. In contrast, our EoE still wins in this case without the help of stored samples. In summary, the improvements of our method over both the rehearsal-free and rehearsal-based methods clearly demonstrate the effectiveness of our ensemble of expert approaches.

(2) As the task proceeds, there is a performance decrease for all methods. For rehearsal-based methods, the reason lies in the interference between old and new tasks. For network-based methods like EPI and ours, the reason is mainly due to the increase in the number of tasks, which consequently brings about the increase in the difficulty of task identification. To have a close look, we compare the task identification accuracy between our method and EPI in Fig. 3. It can be seen that our method is significantly better than EPI which only utilizes BERT for task identification. We can conclude that the significant improvements in classification accuracy of our method over EPI can be primarily attributed to the experts' ability of task identification.

	CA	$\mathcal{T}_6$	$\mathcal{T}_7$	$\mathcal{T}_8$	$\mathcal{T}_9$	$\mathcal{T}_{10}$
FewRel	EoE	90.0	88.9	87.9	86.9	85.5
	w/o DT	90.1	88.8	87.7	86.5	85.0
	w/o UR	90.1	89.0	88.1	87.0	85.5
	w/o RR	90.0	88.9	87.8	86.7	85.3
	w/o CV	89.4	88.1	87.1	85.9	84.3
	w/o All	89.2	87.8	86.7	85.5	83.9
TACRED	EoE	85.9	84.3	83.2	82.7	81.5
	w/o DT	84.1	82.7	81.4	80.9	79.8
	w/o UR	84.8	83.5	82.5	81.8	80.8
	w/o RR	84.3	83.0	82.0	81.3	79.9
	w/o CV	84.0	82.5	81.1	80.5	79.3
	w/o All	82.6	81.2	79.9	79.5	78.2
	IA	$\mathcal{T}_6$	$\mathcal{T}_7$	$\mathcal{T}_8$	$\mathcal{T}_9$	$\mathcal{T}_{10}$
FewRel	EoE	91.8	90.8	89.7	88.6	87.2
	w/o DT	91.7	90.5	89.5	88.1	86.6
	w/o UR	91.9	90.8	89.8	88.6	87.2
	w/o RR	91.8	90.7	89.6	88.3	86.9
	w/o CV	91.1	89.9	88.9	87.6	86.1
	w/o All	90.8	89.6	88.4	87.1	85.5
TACRED	EoE	87.5	85.9	84.8	84.1	83.1
	w/o DT	86.1	84.6	83.2	82.5	81.6
	w/o UR	86.8	85.4	84.5	83.5	82.6
	w/o RR	86.4	84.9	83.9	82.9	81.7
	w/o CV	85.5	84.1	82.7	81.9	80.9
	w/o All	84.5	83.1	81.7	81.0	79.9

Table 2: Results in terms of CA (upper part) and IA (lower part) for ablation study.

6 Analysis

6.1 Ablation Study

To validate the effectiveness of each component in our proposed framework, we conduct an ablation study. Specifically, (1) w/o DT denotes we remove the discriminative training, (2) w/o UR denotes

Group	Model	FewRel	TACRED
G1	EoE w/o DT	77.0	72.6
	EoE	78.4 (+1.4)	74.3 (+1.7)
G2	EoE w/o DT	86.5	76.2
	EoE	87.0 (+0.5)	77.9 (+1.7)
G3	EoE w/o DT	95.6	87.5
	EoE	95.5 (-0.1)	89.0 (+1.5)

Table 3: Analysis results for discriminative training. After training the last task, we divide all relations into three groups according to the similarity between the relation and the task.

we remove undetermined relation augmentation in DT, (3) w/o RR denotes we remove reverse relation augmentation in DT, (4) w/o CV denotes we remove the cascade voting mechanism and only use the first model for task identification, (5) w/o All denotes we remove DT and CV simultaneously, which means we only optimize the expert with standard cross-entropy loss on each dataset and directly use the first expert for task identification.

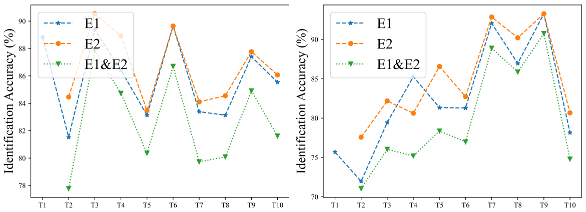
We present the ablation results in terms of classification accuracy (CA) and task identification accuracy (IA) in Table 2. From these results, we can observe that (1) Both DT and CV are effective, and the impact of the cascade voting mechanism is more significant. (2) Adding either reverse relations (RR) or undetermined relations (UR) can significantly enhance the experts' task identification capability. (3) Adding both RR and UR can further improve the task identification capability of the experts on TACRED, but the results on FewRel do not meet our expectations. We believe the main reason lies in the characteristics of the datasets. FewRel is collected from Wikipedia, where many samples involve commonsense relations between entities and do not rely too much on the textual context. (3) The removal of DT and CV results in the biggest drop in performance. This proves that the two components are inextricably linked and play an important role.

6.2 Analysis on Discriminative Training

To further exploit why the discriminative training can improve the task identification capability of the expert when encountering analogous relations, we employ a relation-task similarity metric to divide all observed relations into three nearly equal-size groups, where the similarity between the i-th relation in a task t and the j task is defined as:

$$\mathrm{sim}(i,j) = \mathcal{C}_0^{\mathcal{T}_j}(\mu_{t,i}^0), j \neq t, \tag{8}$$

where $\mathcal{C}_0^{\mathcal{T}_j}(\mu_{t,i}^0)$ is the Mahalanobis distance between the i-th relation and the j-th task computed by BERT. Then, we sort all relations based on $\min_j \mathrm{sim}(i,j), i \notin \mathcal{R}_j$ and divide them into three groups, and present the task identification results in Tab. 3.



(a) FewRel

(b) TACRED

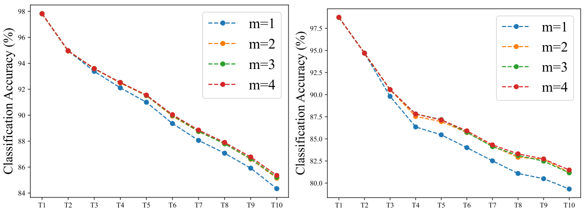
Figure 4: Task identification accuracy on each task after training all tasks. E1, E2, and E1&E2 denote results by the expert θ_1, θ_2, and θ_1 and θ_2 (intersection).

From these results, we can conclude that the analogous relations between tasks are the key reason for the decline in identification accuracy. Note that the relation-task similarity decreases from G1 to G3, while the increase of EoE with the discriminative training becomes more significant, e.g., a 1.4, 0.5, -0.1 gain on G1, G2, G3, respectively. The effect of DT is more obvious on the FewRel dataset than that on TACRED. The reason might be that the amount of each relation is the same for all tasks on FewRel, and thus the trend is much clearer.

6.3 Analysis on Cascade Voting

Overall, the generalization ability of an ensemble of multiple experts is greater than that of a single expert. Still, two points need to be verified. (1) The identification accuracy of experts should not be too bad and should be similar. (2) The identification results among different experts must have certain differences. Hence, to further explore why the cascade voting mechanism can bring performance gains, we report the identification accuracy only with θ_0 or θ_1 on each task after all tasks are finished.

From the results in Fig. 4, we can observe that: (1) There is no significant performance difference in identification accuracy between θ_1 and θ_2. We can also find that the expert θ_k significantly outperforms $\{\theta_i\}_{i=1}^{k-1}$ on the task $\mathcal{T}_k$. For example, θ_2 outperforms θ_1 on the $\mathcal{T}_2$ task by a large margin.



(a) FewRel

(b) TACRED

Figure 5: Impact of the threshold m.

(2) There is a large variability in the prediction results of θ_1 and θ_2, which is the source of the improvement in identification brought about by the subsequent cascade voting.

6.4 Analysis on Computational Cost

The inference in our RFCRE task is divided into two steps: task identification and within-task prediction. The main difference between the best rehearsal-free baseline (EPI) and our EoE is that we aggregate experts from different stages for task identification. However, our EoE needs to use all previous experts to estimate the distribution parameters of the dataset after training at each stage, which may bring about extra computational cost. To reduce the overhead, we introduce a threshold m to limit the number of models that can participate in estimation. Fig. 5 shows the changes in classification accuracy under different thresholds. $m = 1$ denotes we only use θ_1 for task identification [6]. $m = 2$ denotes we incorporate $\{\theta_0, \theta_1, \theta_2\}$.

From Fig. 5, we can find that: (1) Using only one expert for identification has a significant drop compared to multi-expert identification since experts trained on different tasks have their own biases. (2) Simply adding one additional expert like the expert θ_2 to select the results of the first phase in cascade voting can effectively improve the performance. (3) As more experts participate in the voting, the performance increases, but it is limited in magnitude. Thus, the number of experts can be selected according to the real world scenario.

The cascade voting mechanism consists of two phases, and only tasks with conflicting results in the first phase proceed to the second phase. To analyze the efficiency of the cascade voting mechanism, we calculate the percentage of each task that requires a two-phase voting procedure, as shown in Tab. 4. We can find that: (1) The proportion of two-phase voting rises as the number of tasks increases. This is because the increase in the number of tasks leading to an increase in the difficulty of task identification. (2) Even up to the 10-th task, the proportion of two-phase voting is about 12-14%, so nearly 85% of test samples only need the first phase to complete the task identification.

Dataset	$\mathcal{T}_2$	$\mathcal{T}_3$	$\mathcal{T}_4$	$\mathcal{T}_5$	$\mathcal{T}_6$	$\mathcal{T}_7$	$\mathcal{T}_8$	$\mathcal{T}_9$	$\mathcal{T}_{10}$
FewRel	3.01	4.80	5.79	6.88	8.31	9.35	10.64	11.55	12.72
TACRED	3.15	8.19	10.17	10.54	11.72	12.54	13.24	13.82	13.96

Table 4: The proportion (%) of test samples that need a two-stage voting during cascade voting.

7 Conclusion

In this work, we make the first attempt on the rehearsal-free continual relation extraction problem. We propose an ensemble-of-experts framework consisting of discriminative training and cascade voting. Specifically, we first introduce discriminative training to enhance the identification ability of experts when facing inter-task analogous relations. We then propose a cascade voting mechanism to aggregate experts' abilities from different stages. Extensive experimental results show that our method significantly outperforms existing rehearsal-free and rehearsal-based continual relation extraction methods.

Limitations

The main limitation of our proposed framework lies in the extra time and space cost. Though we introduce a threshold to limit the number of experts participating in task identification, additional overhead is unavoidable as long as multiple experts are involved. Besides, except for the first task, we do not consider inter-task knowledge transfer of the subsequent tasks' training while only assigning independent task-specific parameters for each task via PET methods.

Ethics Statement

Our work complies with the ACL Ethics Policy. We have not identified any significant ethical considerations associated with our work.

Acknowledgments

This work was supported by a grant from the National Natural Science Foundation of China (NSFC) project (No. 62276193).

[6] The reason θ_0 is not needed here is that a cascade voting mechanism is not possible with only two experts.

References

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. 2021. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7319–7328.

Li Cui, Deqing Yang, Jiaxin Yu, Chengwei Hu, Jiayang Cheng, Jingjie Yi, and Yanghua Xiao. 2021. Refining sample embeddings with relation prototypes to enhance continual relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 232–243.

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.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. 2022. Delta tuning: A comprehensive study of parameter efficient methods for pre-trained language models. *arXiv preprint arXiv:2203.06904*.

Pierre-Yves Genest, Pierre-Edouard Portier, Elöd Egyed-Zsigmond, and Laurent-Walter Goix. 2022. Promptore-a novel approach towards fully unsupervised relation extraction. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 561–571.

Xu Han, Yi Dai, Tianyu Gao, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2020. Continual relation learning via episodic memory activation and reconsolidation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6429–6440.

Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. 2021. Ptr: Prompt tuning with rules for text classification. *arXiv preprint arXiv:2105.11259*.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809.

Chengwei Hu, Deqing Yang, Haoliang Jin, Zhen Chen, and Yanghua Xiao. 2022. Improving continual relation extraction through prototypical contrastive learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1885–1895.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Gyuhak Kim, Changnan Xiao, Tatsuya Konishi, Zixuan Ke, and Bing Liu. 2022. A theoretical study on solving continual learning. *Advances in Neural Information Processing Systems*, 35:5065–5079.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597.

Zhizhong Li and Derek Hoiem. 2017. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland. Association for Computational Linguistics.

Huy Nguyen, Chien Nguyen, Linh Ngo, Anh Luu, and Thien Nguyen. 2023. A spectral viewpoint on continual relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9621–9629, Singapore. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Hao Peng, Tianyu Gao, Xu Han, Yankai Lin, Peng Li, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2020. Learning from Context or Names? An Empirical Study on Neural Relation Extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3661–3672, Online. Association for Computational Linguistics.

Anastasia Razdaibiedina, Yuning Mao, Rui Hou, Madian Khabsa, Mike Lewis, and Amjad Almahairi. 2023. Progressive prompts: Continual learning for language models. In *International Conference on Learning Representations*.

Yifan Song, Peiyi Wang, Weimin Xiong, Dawei Zhu, Tianyu Liu, Zhifang Sui, and Sujian Li. 2023. InfoCL: Alleviating catastrophic forgetting in continual text classification from an information theoretic perspective. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14557–14570, Singapore. Association for Computational Linguistics.

Hong Wang, Wenhan Xiong, Mo Yu, Xiaoxiao Guo, Shiyu Chang, and William Yang Wang. 2019. Sentence embedding alignment for lifelong relation extraction. 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 796–806.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. 2023a. A comprehensive survey of continual learning: Theory, method and application. *arXiv preprint arXiv:2302.00487*.

Peiyi Wang, Yifan Song, Tianyu Liu, Binghuai Lin, Yunbo Cao, Sujian Li, and Zhifang Sui. 2022a. Learning robust representations for continual relation extraction via adversarial class augmentation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6264–6278.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuan-Jing Huang. 2023b. Orthogonal subspace learning for language model continual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10658–10671.

Yabin Wang, Zhiheng Ma, Zhiwu Huang, Yaowei Wang, Zhou Su, and Xiaopeng Hong. 2023c. Isolation and impartial aggregation: A paradigm of incremental learning without interference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10209–10217.

Yiwei Wang, Muhao Chen, Wenxuan Zhou, Yujun Cai, Yuxuan Liang, Dayiheng Liu, Baosong Yang, Juncheng Liu, and Bryan Hooi. 2022b. Should we rely on entity mentions for relation extraction? debiasing relation extraction with counterfactual analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3071–3081, Seattle, United States. Association for Computational Linguistics.

Zhicheng Wang, Yufang Liu, Tao Ji, Xiaoling Wang, Yuanbin Wu, Congcong Jiang, Ye Chao, Zhencong Han, Ling Wang, Xu Shao, et al. 2023d. Rehearsal-free continual language learning via efficient parameter isolation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10933–10946.

Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. 2022c. Dual-prompt: Complementary prompting for rehearsal-free continual learning. In *European Conference on Computer Vision*, pages 631–648. Springer.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. 2022d. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 139–149.

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. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Tongtong Wu, Xuekai Li, Yuan-Fang Li, Gholamreza Haffari, Guilin Qi, Yujin Zhu, and Guoqiang Xu. 2021. Curriculum-meta learning for order-robust continual relation extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10363–10369.

Heming Xia, Peiyi Wang, Tianyu Liu, Binghuai Lin, Yunbo Cao, and Zhifang Sui. 2023. Enhancing continual relation extraction via classifier decomposition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10053–10062, Toronto, Canada. Association for Computational Linguistics.

Weimin Xiong, Yifan Song, Peiyi Wang, and Sujian Li. 2023. Rationale-enhanced language models are better continual relation learners. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15489–15497, Singapore. Association for Computational Linguistics.

Li-Ming Zhan, Haowen Liang, Bo Liu, Lu Fan, Xiao-Ming Wu, and Albert YS Lam. 2021. Out-of-scope intent detection with self-supervision and discriminative training. In *Proceedings of the 59th Annual*

Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3521–3532.

Kang Zhao, Hua Xu, Jiangong Yang, and Kai Gao. 2022. Consistent representation learning for continual relation extraction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3402–3411.

Wenzheng Zhao, Yuanning Cui, and Wei Hu. 2023. Improving continual relation extraction by distinguishing analogous semantics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1162–1175, Toronto, Canada. Association for Computational Linguistics.

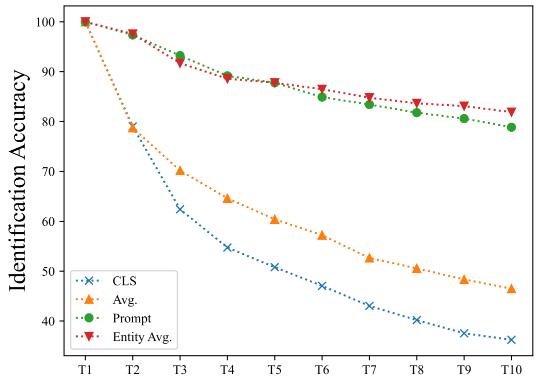


Figure 6: Changes in task identification accuracy (%) with different representation extraction methods for TACRED dataset.

A Impact of Different Representation Extraction Methods

EPI (Wang et al., 2023d) applies the original BERT (Devlin et al., 2019) to model the representation space of each relation as a multivariate Gaussian distribution. However, it uses the average representation of the last layer for each instance for coarse-grained text classification tasks, which is insufficient for relation extraction. For a fair comparison, we utilize several different methods to test the original BERT's ability of task identification in continual relation extraction, including:

- **CLS**: Referring to the pre-training objective, we directly use the `[CLS]` vector as the representation of the instance.

- **Avg.**: Following Wang et al. (2023d), we use the average representation of the last layer as the representation of the instance.

- **Prompt**: Following Genest et al. (2022), we convert the input instance with the prompt template "`[CLS]` x h `[MASK]` t `[SEP]`" where x is the input text, h is the head entity and t is the tail entity. Then, we use the `[MASK]` vector as the representation of the instance.

- **Entity Avg.**: We concatenate the average representation of the head entity h and tail entity t as the representation of the instance.

As shown in Figure 6, the task identification accuracy of CLS and Avg. decreases rapidly as the number of tasks increases because they ignore the task-related knowledge of relation extraction. In addition, we can find that Entity Avg. shows better identification performance than Prompt, so we use Entity Avg. as the default extraction method of the original BERT.

B Brief introduction of LoRA

Pre-trained language models have a low intrinsic dimension during the adaption of the downstream task (Aghajanyan et al., 2021). The essence of LoRA lies in introducing additional parameters $\mathbf{A} \in \mathbb{R}^{d\times r}$ and $\mathbf{B} \in \mathbb{R}^{r\times k}$ to supplement the pre-trained weight matrix $\mathbf{W_0} \in \mathbb{R}^{d\times k}$, where rank $r \ll min(d, k)$. At the training stage, LoRA freeze the update of $\mathbf{W_0}$ and utilize a low-rank decomposition $\mathbf{W}_0 + \Delta\mathbf{W} = \mathbf{W}_0 + \mathbf{AB}$ to represent its update. This means that the forward pass $\boldsymbol{h} = \mathbf{W}\boldsymbol{x}$, where $\mathbf{x}$ is the input vector, is modified as follows:

$$\boldsymbol{h} = \mathbf{W}\boldsymbol{x} + \Delta\mathbf{W}\boldsymbol{x} = \mathbf{W}\boldsymbol{x} + \mathbf{AB}\boldsymbol{x} \qquad (9)$$

C Details of Experimental Setups

C.1 Datasets

FewRel (Han et al., 2018): is a large-scale and balanced relation extraction dataset that contains 80 relations and 700 instances for each relation. Following Cui et al. (2021), we partitioned the dataset into 10 sub-datasets, each containing 8 relations. For each relation, we sampled 420 samples as the training set and 140 as the test set.

TACRED (Han et al., 2018): is a widely used RE dataset containing 42 relations (including `no_relation`). Following Cui et al. (2021), we remove `no_relation` in our experiments and partition the dataset into 10 sub-datasets, each containing 4 relations. The number of training samples of each relation is limited to 320, while the number of test samples of each relation is limited to 40.

C.2 Baselines

In this work, we compare five rehearsal-free methods as follows:

- **FT**: directly fine-tune the model without any strategy to prevent catastrophic forgetting, which can be viewed as the lower bound of continual learning.

- **EWC** (Kirkpatrick et al., 2017): maintains an importance matrix with the same scale as the model first and then uses L_2 loss to constrain the update of important parameters.

- **LwF** (Li and Hoiem, 2017): utilizes knowledge distillation to force the predicted probability of the old relations between the old model and the current model to be the same.

- **L2P** (Wang et al., 2022d): introduces a prompt-based continual learning framework that freezes the pre-trained encoder and then adapts to different tasks via a shared prompt pool.

- **EPI** (Wang et al., 2023d): leverages the pre-trained language model to estimate the input distribution for each task with a Gaussian and then uses the Mahalanobis distance for task identification.

Besides we also compare five rehearsal-based continual learning methods focus on relation extraction as follows:

- **CEAR** (Zhao et al., 2023): proposes memory-insensitive relation prototypes and memory augmentation during rehearsal replay to alleviate catastrophic forgetting.

- **CDec** (Xia et al., 2023): proposes a classifier decomposition framework with empirical initialization and adversarial training to alleviate classifier bias and representation bias.

- **RationaleCL** (Xiong et al., 2023): proposes a rationale-enhanced framework to improve the model's robustness in the face of future analogous relations via multi-task rationale tuning and contrastive rationale replay strategy.

- **CFDR** (Nguyen et al., 2023): proposes a class-wise feature decorrelation regularization to boost eigenvalues.

- **InfoCL** (Song et al., 2023): exploits fast-slow contrastive learning during new task training and current-past contrastive learning during rehearsal replay to learn more comprehensive representations.