Distilling Causal Effect of Data in Continual Few-shot Relation Learning

Weihang Ye, Peng Zhang* , Jing Zhang, Hui Gao, Moyao Wang
College of Intelligence and Computing, Tianjin University
Tianjin, China
{whye, pzhang, zhang_jing, hui_gao, moyao}@tju.edu.cn

Abstract

Continual Few-Shot Relation Learning (CFRL) aims to learn an increasing number of new relational patterns from a data stream. However, due to the limited number of samples and the continual training mode, this method frequently encounters the catastrophic forgetting issues. The research on causal inference suggests that this issue is caused by the loss of causal effects from old data during the new training process. Inspired by the causal graph, we propose a unified causal framework for CFRL to restore the causal effects. Specifically, we establish two additional causal paths from old data to predictions by having the new data and memory data collide with old data separately in the old feature space. This augmentation allows us to preserve causal effects effectively and enhance the utilization of valuable information within memory data, thereby alleviating the phenomenon of catastrophic forgetting. Furthermore, we introduce a self-adaptive weight to achieve a delicate balance of causal effects between the new and old relation types. Extensive experiments demonstrate the superiority of our method over existing state-of-the-art approaches in CFRL task settings. Our codes are publicly available at: https://github.com/ywh140/CECF.

Keywords: Continual Few-Shot Relation Learning, Causal Effect, Catastrophic Forgetting

1. Introduction

Relation Extraction (RE) aims to extract relations between entities from unstructured text. For example, it involves detecting the relation *founder* in the sentence "*Jobs* is the founder of *Apple*." for the two entities *Jobs* and *Apple*. It holds substantial utility across various downstream applications, including knowledge augmentation, question answering, text generation, and summarization (Wang et al., 2022; Dong et al., 2015). Traditional RE methods rely on fixed pre-defined sets of relations (Han et al., 2018b; Gao et al., 2019; Xiong et al., 2017). Nevertheless, in practical scenarios, the new relation types emerge constantly, necessitating more flexible approaches to accommodate these evolving circumstances.

To address this issue, some researchers have explored formalizing RE as Continual Relation Learning (CRL) (Wang et al., 2019). As shown in Figure 1, the model continuously learns a series of tasks, each with its own set of relations. However, it is universally acknowledged that the major challenge in Continual Learning (CL) is catastrophic forgetting (McCloskey and Cohen, 1989). In other words, the model tends to forget previously learned relation tasks while learning new ones.

In the field of Natural Language Processing (NLP), memory-based methods have proven to be effective solutions to the catastrophic forgetting problem in CL (Wang et al., 2019; de Masson D'Autume et al., 2019). The memory-based methods select samples from previous tasks and stores them in memory. Subsequently, when learning a new task, the stored memory data is replayed to consolidate previously acquired knowledge. Prominent memory-based methods include EA-EMR (Wang et al., 2019), ARPER (Mi et al., 2020) and EMAR (Han et al., 2020).

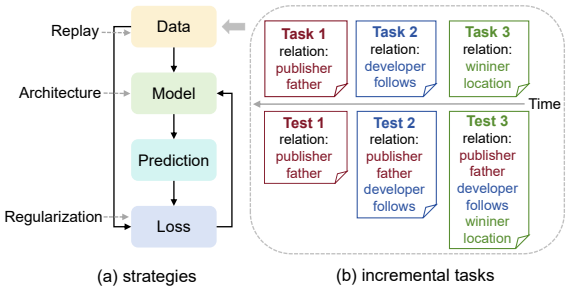


Figure 1: A conceptual framework of Continual Relation Learning (CRL). (a) Representative strategies have targeted various aspects of deep learning. (b) CRL requires adapting to incremental tasks with dynamic relation data.

Nevertheless, existing memory-based methods heavily rely on extensive training data to learn new relations. In real-world scenarios, acquiring large labeled datasets for nascent events can be challenging. In response to such circumstances, evolution has empowered human with strong adaptability to acquire knowledge from a limited number of samples. Naturally, we expect that CRL can also possess this ability. Unfortunately, some continual

* Corresponding author: Peng Zhang

LREC-COLING 2024, pages 5041–5051
20-25 May, 2024.

learning methods face issues of overfitting. Recently, Qin and Joty (2022) proposed a method for Continual Few-shot Relation Learning (CFRL).

Through causal inference, Hu et al. (2021) discovered that the forgetting is due to the absence of causal paths from old data to the predictions, resulting in the lack of causal effects from old data. In the general training process of CFRL, we first employ simple training to select memory samples and then proceed with anti-forgetting training. While the data replay method establishes a causal path to restore the causal effect in anti-forgetting training, there is no causal path in simple training to resist forgetting. Additionally, the memory data is not utilized at all in simple training. In the case of abundant data, simple training has a small impact on the final results because it constitutes a small proportion of the total training epochs. However, in the few-shot setting, as its proportion increases, its impact also grows.

Based on causal inference, this study presents a **C**ausal **E**ffect **C**ontinual **F**ew-shot Framework(**CECF**) to address these issues. Specifically, we propose to add a causal path between old data and predictions leveraging the colliding effect. In order to better leverage valuable information within the memory data, we establish an additional causal path by colliding memory data with old data in the old feature space. Because these added causal paths effectively preserve the causal effects of old data, the problem of catastrophic forgetting is mitigated. In addition, we introduce the self-adaptive weight mechanism to balance the causal effects of new and old relations. We conduct experiments in different settings on two benchmark datasets, and the results demonstrate that our approach outperforms existing state-of-the-art methods in CFRL. In summary, our main contributions are as follows:

- We are the first to understand CFRL from a causal graph perspective and propose a unified causal framework to effectively preserve the causal effects of old data.
- For mitigating forgetting, we are the first to distill causal effects from new data and memory data in simple training. We also introduce a self-adaptive weight to balance the causal effects of new and old relations.
- Through extensive experiments in various settings on two datasets, we observe consistent performance improvements. For example, CECF outperforms previous baselines by up to 3.04% on FewRel and 4.2% on TACRED.

2. Related Work

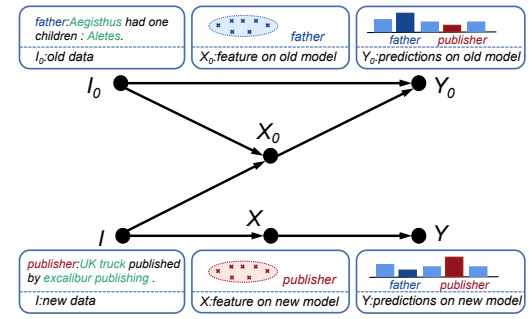
RE techniques are commonly classified into several logical categories: supervised (Zelenko et al., 2003; Zeng et al., 2014; Miwa and Bansal, 2016), semi-supervised (Chen et al., 2006; Sun et al., 2011; Hu et al., 2020) and distant supervision based techniques (Mintz et al., 2009; Zeng et al., 2015; Han et al., 2018a). These methods heavily rely on extensive pre-defined data. Yet in the real world, new relations emerge rapidly, and it is challenging to acquire a sufficient number of labeled data quickly at such times. Hence, some researchers propose **Continual Few-shot Relation Learning (CFRL)**, in order to learn relations without large pre-defined relation sets (Qin and Joty, 2022). The achievement of this goal involves two critical techniques: Continual Learning and Few-shot Learning.

Continual Learning (CL) focuses on learning from an infinite stream of data, with the goal of gradually extending the pool of acquired knowledge for future learning (De Lange et al., 2021). The main challenge is achieving CL without catastrophic forgetting (McCloskey and Cohen, 1989), where the knowledge of old tasks is lost when learning new relations. As depicted in Figure 1, there are currently three prevalent strategies in CL. First, data replay (memory-based) methods involve storing samples or generating synthetic samples through generative models. When learning new tasks, these samples from previous tasks are replayed to counteract the phenomenon of forgetting (Rebuffi et al., 2017; Lopez-Paz and Ranzato, 2017; Atkinson et al., 2018). Second, architecture-based methods dynamically modify the model's architecture to accommodate new information while preserving knowledge gleaned from prior tasks (Chen et al., 2015; Rusu et al., 2016; Fernando et al., 2017; Chaudhry et al., 2018). Finally, regularization-based methods introduce additional regularization terms into the loss function (Li and Hoiem, 2017; Kirkpatrick et al., 2017; Nguyen et al., 2017).

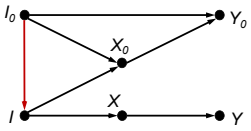
Few-shot Learning (FSL) refers to the problem of learning underlying patterns in data with only a few training samples. (Wang et al., 2020). FSL can be categorized into three directions: (1) data-based methods employ prior knowledge to enhance supervision experience (Santoro et al., 2016; Benaim and Wolf, 2018; Qin and Joty, 2022); (2) model-based methods reduce the size of hypothesis space with prior knowledge (Triantafillou et al., 2017; Hu et al., 2018); and (3) algorithm-based methods change the search method for optimal hypotheses in a given hypothesis space (Hoffman et al., 2013; Ravi and Larochelle, 2016; Finn et al., 2017).

Causal Inference (Pearl et al., 2016; Schölkopf, 2022) recently has been applied to aid NLP tasks, addressing issues such as spurious correlations (Gururangan et al., 2018; Veitch et al., 2021; Rosenfeld et al., 2020), biases in textual data (Hardt et al., 2016; Kilbertus et al., 2017), and model inter-

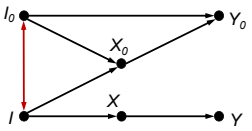
pretability (Feder et al., 2022; Karimi et al., 2021; Vig et al., 2020; Finlayson et al., 2021). Zheng et al. (2022) also leveraged causal inference to retain knowledge from samples of Other-Class for Named Entity Recognition (NER). Inspired by the pioneering work of Hu et al. (2021), which utilizes causal view to explain catastrophic forgetting, we aim to employ causal inference to confront long-standing challenges in CFRL.



(a) The Forgetting in CRL



(b) Data Replay



(c) Colliding Effect

Figure 2: The proposed causal graphs explaining the forgetting and anti-forgetting in CRL.

3. (Anti-) Forgetting in Causal Views

In this section, we provide a causal perspective on understanding the principles of forgetting and introduce how to create causal paths to mitigate forgetting.

3.1. Causal Graphs

Figure 2(a) illustrates a causal graph of the CRL. In this depiction, nodes serve as variables, while directed edges symbolize causal relations between pairs of nodes. Specifically, we define I_0 as the old data, I as the new data, X and X_0 as the extracted features from the new and old model, and Y and Y_0 as the predicted labels from the new and old model , respectively. The important links in this graph are as follows: (1) $I \rightarrow X$ signifies the utilization of the new model to extract features X from the input sentence I. (2) $X \rightarrow Y$ indicates using the extracted features X to predict the results Y. (3) $(I_0, I) \rightarrow X_0$ denotes that for a given new input sentence I, we can obtain the feature representation X_0 in the old feature space by using the old model trained on old data I_0. As observed in Figure 2(a), forgetting occurs when all the causal paths from I_0 to Y are blocked by colliders ($e.g. X_0$). Figure 2(b) represents the causal graph of the data replay method, where exists a causal path connecting I_0 and Y.

3.2. Colliding Effects

In addition to the commonly used anti-forgetting methods such as data replay and distillation, Hu et al. (2021) proposed a novel approach that leverages the colliding effect to establish causal paths between old data and predictions. As depicted in Figure 2(c), the colliding effect involves controlling the collider X_0 to establish mutual associations between the nodes I and I_0. A causal graph from real life can be used to aid understanding. For example, $Temperature \rightarrow Sensation \leftarrow Wind$ indicates that the perceived temperature is influenced by both the actual temperature and the wind. Typically, actual temperature and wind speed are considered as independent variables. However, when a person feels cold and the wind speed is low, it can be inferred that the actual temperature might be low, and vice versa.

4. Methodology

In this section, we present a unified causal framework for retrieving causal effects from both new data and memory data. Furthermore, a self-adaptive weight is introduced to effectively address the issue of data imbalance.

4.1. Problem Definition

In CFRL, the model M needs to learn a series of tasks $\mathbb{T} = (\mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_i, ..., \mathcal{T}_n)$. Each task $\mathcal{T}_i$ contains a set of relations R_i and corresponds to a specific dataset D_i. This dataset consists of samples $\{(x_k, y_k)\}_{k=1}^{|D|}$, where the relation label y_k comes from the relation set R_i. As discussed earlier, the few-shot task scenario is characterized by a limited number of labeled samples (e.g., 5), in all tasks except the initial one. Assuming that each few-shot task includes N relations, and each relation contains K samples, this setting is commonly referred to as N-way K-shot continual learning.

The target of CFRL is to learn the best RE model M_i to identify the relation set $\hat{R}_i = \bigcup_{k=1}^{i} R_k$. Achieving this goal necessitates addressing the challenge of catastrophic forgetting, which pertains to the model's ability to retain previously learned knowledge while acquiring new knowledge from limited data. To mitigate catastrophic forgetting, in the i-th step, the RE model M_{i-1}, previously trained on the dataset D_{i-1}, serves as the initialization point for the new model M_i. The teacher model M_{i-1} guides the learning process of the student model

M_i through knowledge distillation. Moreover, the memory data $\hat{\mathcal{Q}}_{i-1} = \{\mathcal{Q}_1, \cdots, \mathcal{Q}_{i-1}\}$ consists of the samples selected from each relation in previous tasks.

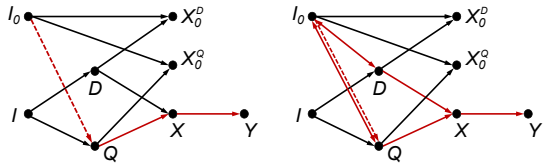


Figure 3: The causal graph for CFRL: (a) ERDA has only one causal path from old data to new predictions, which comes from data replay; (b) CECF builds additional two causal paths from old data to new predictions through new training data D and memory data Q.

4.2. Distilling Colliding Effect in CFRL

By analyzing the causal graph in Figure 3(a), we observe that the data replay methods establish a causal path. However, this path exists only in the anti-forgetting training. The simple training lacks intrinsic mechanisms to effectively mitigate forgetting. The utilization of the colliding effect emerges as a solution to introduce novel causal paths aimed at alleviating the phenomenon of forgetting.

To establish the new causal path using the colliding effect, we identify sentences in new data that have the similar feature representations X_0 to the input in the old feature space. Initially, we compute the old feature representations of the input and other samples. Following Hu et al. (2021), we employ the $\mathrm{K-Nearest-Neighbors}$ (KNN) strategy to search the K sentences whose features bear a greater similarity to the feature of input. Next, during the prediction phase for the input sentence, we leverage these matched sentences for joint prediction.

In simple training, the model does not utilize the valuable information within the memory data. We consider colliding the memory data with the old data in the old feature space to establish an additional causal path. The new causal paths augment the influence of old data on predictions, thereby ameliorating the model's propensity for forgetting during the training process.

Based on the distinctive attributes of CFRL, we extended the causal graph from Figure 3(a) to Figure 3(b). The key modification is that the new relation data D and the memory data Q collide separately with old data I_0 on nodes X_0^D and X_0^Q in the old feature space. In this way, we introduce two causal paths from old data I_0 to predictions Y. In the absence of colliding effects, the original class boundaries in the feature space are more likely to be disrupted during the training process, leading to forgetting. When the input sample collides with other sentences with similar representations, the original class boundaries in the feature space are more likely to be preserved, naturally alleviating forgetting.

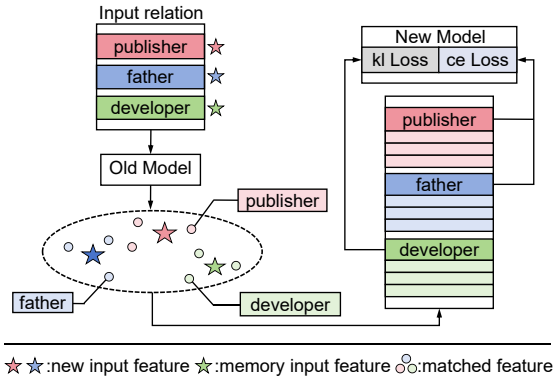


Figure 4: Our causal framework for CFRL.

4.3. Overall Framework

Our framework is illustrated in Figure 4, and the training process is depicted in Algorithm 1. The CFRL learning process typically encompasses three steps: learning with new data (simple training), selecting samples for memory, and alleviating forgetting through memory (anti-forgetting training). Our proposed improvements primarily concentrate on the first step. It is noteworthy that the settings for the subsequent steps are the same as Qin and Joty (2022).

In the i-th CFRL step, given the training data D_i for the task $\mathcal{T}_i$. If the task is the initial task, we have $\widetilde{D_i} = D_i$. For data augmentation, we use a fine-tuned BERT (Devlin et al., 2018) as the relational similarity model S to select unlabeled samples from the Wikipedia corpus $\mathcal{C}$ which have a high similarity score. Then, we merge these chosen samples with the training set D_i to create a new extended training set $\widetilde{D_i}$.

To distill causal effects from old data and enhance the utilization of memory data, we first initialize the model M_i using the pre-existing model M_{i-1} after the initial task. Then we use the model M_{i-1} to calculate the old feature representations $\mathbf{x}_k \in \mathbb{R}^d$ for the expanded training data $\widetilde{D_i}$ and $\mathbf{x}_l \in \mathbb{R}^d$ for the memory data $\hat{\mathcal{Q}}_{i-1}$. After that, we employ the KNN method to select K sentences that are most similar to the input sentence from $\widetilde{D_i}$ or $\hat{\mathcal{Q}}_{i-1}$ in the old feature space. Finally, we use the loss $\mathcal{L}$ to update the model M_i, and the total loss $\mathcal{L}$ is defined as:

Algorithm 1 training process at time step i

Require: The training set, D_i; The current task, $\mathcal{T}_i$; The current memory set $\hat{\mathcal{Q}}_{i-1}$ The relation set, R_i; The model M_i; The refer model M_{i-1}; The similarity model $\mathcal{S}$; The unlabeled text corpus $\mathcal{C}$.

if $i == 1$ **then** ▷ initial task
  $\widetilde{D_i} = D_i$
else ▷ few-shot task
  SELECT similar samples from $\mathcal{C}$ using $\mathcal{S}$ for every sample in D_i and store them in A
  $\widetilde{D_i} = A \bigcup D_i$
  $M_i \leftarrow M_{i-1}$ ▷ initialize new model
  $\mathcal{X}_0^d \leftarrow M_{i-1}(\widetilde{D_i})$ ▷ represent in old features
  $\mathcal{X}_0^q \leftarrow M_{i-1}(\hat{\mathcal{Q}}_{i-1})$
  $\mathcal{N}_d \leftarrow \text{KNN}(M_{i-1}(x_k), \mathcal{X}_0^d)$
  $\mathcal{N}_q \leftarrow \text{KNN}(M_{i-1}(x_l), \mathcal{X}_0^q)$
end if
INITIALIZE $\mathbf{r}_k$ for every relation $r_k \in R_i$
for $i = 1, ..., iter_1$ **do** ▷ train on new task
  UPDATE M_i with $\mathcal{L}$ on $\left\{\widetilde{D_i}, \hat{\mathcal{Q}}_{i-1}, \mathcal{N}_d, \mathcal{N}_q\right\}$
end for
SELECT key samples from D_i for every relation $r_k \in R_i$ to save in $\mathcal{Q}_i$
$\hat{R}_i = \hat{R}_{i-1} \bigcup R_i$
$\hat{\mathcal{Q}}_i = \hat{\mathcal{Q}}_{i-1} \bigcup \mathcal{Q}_i$ ▷ update memory
$\widetilde{H_i} = \widetilde{D_i} \bigcup \hat{\mathcal{Q}}_i$
for $i = 1, ..., iter_2$ **do**
  UPDATE M_i on $\widetilde{H_i}$
  UPDATE $\mathbf{r}_k$ for every relation $r_k \in \hat{R}_i$
end for

$$\mathcal{L} = \lambda_{ce}\mathcal{L}_{ce} + \lambda_{kl}\mathcal{L}_{kl} + \lambda_{mm}\mathcal{L}_{mm} + \lambda_{pm}\mathcal{L}_{pm} \quad (1)$$

where λ_{ce}, λ_{kl}, λ_{mm} and λ_{pm} are the weights corresponding to their respective losses. The detailed explanations for losses $\mathcal{L}_{ce}$ and $\mathcal{L}_{kl}$ can be found in sections 4.3.1 and 4.3.2 respectively. We also adopt a margin-based loss $\mathcal{L}_{mm}$ and a pairwise margin loss $\mathcal{L}_{pm}$ like Qin and Joty (2022) to improve the discriminative ability of the model.

4.3.1. Distilling Causal Effect of New Data

At time step i, we use the reference model M_{i-1} to compute the old feature representation $\mathbf{x}_k$ of each new sample $x_k \in \widetilde{D}_i$. $\mathbf{x}_k$ is stored in $\mathcal{X}_0^d$. In the old feature space, we select top K matched samples within the expanded dataset $\widetilde{D_i}$ that closely resemble to the input x_k to distill the colliding effect. The matched samples constitute the set $\mathcal{N}_d$. The distance calculation formula is as follows:

$$dist = \sqrt{\sum_{m=1}^{d}(\mathbf{p}_m - \mathbf{q}_m)^2} \quad (2)$$

where d is the number of dimensions, $\mathbf{p}_m$ and $\mathbf{q}_m$ are the m-th dimensions of the old feature representations $\mathbf{p}, \mathbf{q} \in \mathcal{X}_0^d$.

Then we calculate the loss $\mathcal{L}_{ce}$ by leveraging the matched samples $\mathcal{N}_d$ and the expanded data $\widetilde{D_i}$ together. The weight allocation during the joint training process is as follows:

$$\begin{aligned} \bar{Y_k} =& W_k Y_k + \textstyle\sum_{j=1}^{K} W_{kj} Y_{kj} \\ \text{s.t.} \quad & W_k \geq W_{k1} \geq W_{k2} \geq \cdots \geq W_{kK} \\ & W_k + \textstyle\sum_{j=1}^{K} W_{kj} = 1 \end{aligned} \quad (3)$$

Y_k,Y_{kj} are the predicted scores of the input x_k and its j-th matched sample. W_k,W_{kj} are their respective weights. K is the number of matched sentences. The weight constraints ensure that the sample closer to the input has a more significant impact. And as long as they satisfy the equation, their influence on the final result is not substantial (Hu et al., 2021). The specific cross entropy loss $\mathcal{L}_{ce}$ is used for relation classification:

$$\mathcal{L}_{ce} = \sum_{(x_k, y_k) \in \widetilde{D_i}} \sum_{n=1}^{|\hat{R}_i|} \delta_{y_k, r_n} \times log(\bar{Y_k}) \quad (4)$$

$$\bar{Y_k} = \frac{1}{2}\mathrm{S}(M_i(x_k)) + \frac{1}{2K}\sum_{j=1}^{K}\mathrm{S}(M_i(x_{kj})) \quad (5)$$

where $\hat{R}_i$ is the ensemble of all known relations at the present step, and x_{kj} is the j-th matched sample of the input x_k. δ_{y_k, r_n} signifies a symbolic function (0 or 1). It assumes 1 when the number of relation types in the sample y_k equals r_n, otherwise, it assumes 0. S represents the $\mathrm{Softmax}$ function.

4.3.2. Distilling Causal Effect of Memory Data

To combat forgetting and better leverage the effective information stored in the memory data, we establish an additional causal path. The model M_{i-1} encodes each memory sample $x_l \in \hat{\mathcal{Q}}_{i-1}$ into its representation $\mathbf{x}_l$ in the old feature space. These representations are stored in $\mathcal{X}_0^q$. We employ Equation 2 to search the K most similar samples for each input sample x_l, where $\mathbf{p}, \mathbf{q} \in \mathcal{X}_0^q$. The matched samples are saved in $\mathcal{N}_q$. Following this, we calculate the loss $\mathcal{L}_{kl}$ by jointly considering both the memory data $\hat{\mathcal{Q}}_{i-1}$ and the matched samples $\mathcal{N}_q$. The distillation loss $\mathcal{L}_{kl}$ is defined as:

$$\mathcal{L}_{kl} = \sum_{(x_l, y_l) \in \hat{\mathcal{Q}}_{i-1}} \sum_{n=1}^{|\hat{R}_{i-1}|} \delta_{y_l, r_n} \times log\frac{\bar{Y_s}}{Y_t} \quad (6)$$

$$\bar{Y_s} = \frac{1}{2}\mathrm{S}(\frac{M_i(x_l)}{T_s}) + \frac{1}{2K}\sum_{j=1}^{K}\mathrm{S}(\frac{M_i(x_{lj})}{T_s}) \quad (7)$$

$$Y_t = \mathrm{S}(\frac{M_{i-1}(x_l)}{T_t}) \quad (8)$$

where the temperature parameters T_t, T_s for the teacher model M_{i-1} and the student model M_i, respectively. x_{lj} represents the j-th matched sample of the memory sample x_l.

4.3.3. Balancing Causal Effects

The overall causal effect consists of learning new relations and reviewing old relations. As the steps increase, the number of relations to be learned for new tasks remains the same, while the number of relations to be reviewed for old tasks continues to grow. Therefore, there is a need to balance the learning ability between new and old relation types continuously. To achieve this, we introduce a self-adaptive weight:

$$\lambda_{kl} = \lambda\sqrt{\frac{R_o}{R_n}} \quad (9)$$

Where λ is an initial weight, R_o is the number of old relations, and R_n is the number of new relations. This way, as the number of old relations continues to increase, the weight can be automatically adjusted to balance the learning abilities.

5. Experiment

5.1. Settings

Benchmark We conduct experiments on two widely used datasets for CFRL. **FewRel** (Han et al., 2018b) comprises 80 relations, each with hundreds of samples. Following the CFRL benchmark (Qin and Joty, 2022), we divide these relations into 8 tasks, with each task consisting of 10 relations (10-way). The first task has 100 samples for each relation, while subsequent tasks have fewer samples. To demonstrate the effectiveness of our method, we perform 2-shot, 5-shot, and 10-shot experiments on the subsequent tasks.

Furthermore, we conduct experiments on **TACRED** (Zhang et al., 2017) to showcase the generalizability of our method. TACRED includes 41 available relations, which we divide into 8 tasks. The first task contains 6 relations, each with 100 samples, while the remaining tasks have 5 relations (5-way). Similar to Fewrel, we conduct experiments with 5-shot and 10-shot settings.

Metric At time step i, we evaluate the model's relation classification predictions by using the test set of the i-th task along with the test sets from all previous time steps. This evaluation metric effectively reflects whether the model has the ability to mitigate catastrophic forgetting in CFRL. Considering that the selection and order of relation sets may affect the model's performance, we conduct six experiments using different random seeds to eliminate randomness. The reported results are the average performance of these six experiments.

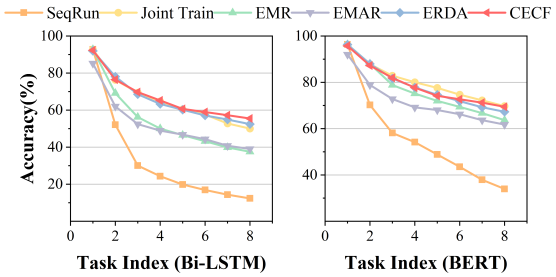


Figure 5: Comparison results of different methods with a Bi-LSTM encoder and a BERT encoder on **FewRel** benchmark. For both encoders, CECF is better than ERDA in the **10-way 5-shot** setting.

Baseline We compare our method with the following baselines:

- **SeqRun** fine-tunes the model directly on new task data would lead to severe catastrophic forgetting, serving as a lower bound.
- **Joint Training** combines the training data of the current task with the training data of all previous tasks and jointly trains. It serves as an upper bound.
- **EMR** (Wang et al., 2019) stores samples selected from all previous steps. In the subsequent step, these samples are added to the current training data for joint training.
- **EMAR** (Han et al., 2020) introduces Memory Activation and reconsolidation into CRL.
- **ERDA** (Qin and Joty, 2022) is the state-of-the-art in CFRL. It proposes a novel method based on embedding space regularization and data augmentation to avoid catastrophic forgetting.

Hyper-Parameter We set the number of matched samples $K = 3$. For the parameter of weight allocation, we set the weights $W_k = \frac{1}{2}$ and $W_{kj} = \frac{1}{2K}$. We initialize the self-adaptive weight $\lambda = 1.2$.

5.2. Main Results

We compare the performance of different methods using the same setting as ERDA (Qin and Joty, 2022), which uses a Bi-LSTM encoder and a BERT encoder.

Method	Task index							
	1	2	3	4	5	6	7	8
SeqRun	**92.78**	52.11	30.08	24.33	19.83	16.90	14.36	12.34
Joint Train	**92.78**	76.29	69.39	64.75	60.45	57.64	52.80	50.03
EMR	**92.78**	69.14	56.24	50.03	46.50	43.21	39.88	37.51
EMAR	85.20	62.02	52.45	48.95	46.77	44.33	40.75	39.04
ERDA	91.98	**78.09**	68.59	63.32	60.2	57.13	54.91	52.45
CECF	92.32	76.48	**69.73**	**65.24**	**60.65**	**58.99**	**57.26**	**55.49**

Table 1: Accuracy (%) of different methods at every time step on **FewRel** benchmark for **10-way 5-shot** CFRL. CECF is better than ERDA with a Bi-LSTM encoder.

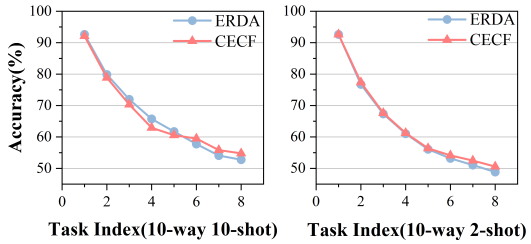


Figure 6: Comparison results at each time step on **FewRel** benchmark for **10-way 10-shot** and **2-shot** settings. For both settings, CECF is better than ERDA with a Bi-LSTM encoder.

FewRel Benchmark with Bi-LSTM We report the results of previous baselines and the proposed model on 10-way 5-shot in Table 1 and Figure 5. Additionally, Figure 6 displays the results on the 10-way 10-shot and 10-way 2-shot settings. We compare our experiments under the setting of FewRel 10-way 5-shot with all the baselines. For a clear comparison, other settings will be compared only with ERDA. From these results, we can observe that:

Our proposed method CECF consistently outperforms the baselines, highlighting the effectiveness of our approach. EMR and EMAR tend to overfit and suffer from catastrophic forgetting when confronted with few-shot tasks, because they rely on a substantial amount of training data. The performance of ERDA surpasses that of EMR and EMAR as ERDA is specifically designed for few-shot continual learning tasks. ERDA mitigates catastrophic forgetting by enforcing extra constraints on the relational embeddings and adding extra relevant data in a self-supervised manner. However, there are no anti-forgetting measures when the model learns new relations. Introducing new paths from old data to predictions through colliding effects can obtain causal effects from old data.

In CFRL, joint training is not always the upper bound due to the data imbalance in few-shot tasks. Since the ERDA model is designed for few-shot tasks, it outperforms joint training in the last two

Method	Task index							
	1	2	3	4	5	6	7	8
SeqRun	**96.35**	70.23	58.13	54.17	48.82	43.52	37.9	33.97
Joint Train	**96.35**	87.85	**82.87**	**80.05**	**77.62**	**74.69**	**72.23**	**69.74**
EMR	**96.35**	**88.02**	78.83	75.15	72	69.41	66.7	63.68
EMAR	92.03	78.87	72.81	69.19	68.05	66.23	63.68	61.77
ERDA	96.3	87.96	81.64	77.8	74.64	71.86	69.23	67.3
CECF	95.75	87.24	**81.94**	**77.71**	**74.18**	**72.69**	**71.21**	**69.44**

Table 2: Accuracy (%) of different methods at every time step on **FewRel** benchmark for **10-way 5-shot** CFRL. CECF is better than ERDA with a BERT encoder.

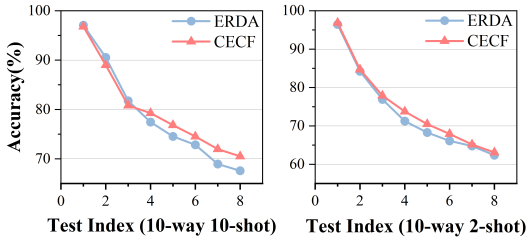


Figure 7: Comparison results of different methods on **FewRel** benchmark for **10-way 10-shot** and **2-shot** settings. CECF is better than ERDA with a BERT encoder for both settings.

tasks in the 10-way 5-shot setting. However, our method surpasses joint training to a greater extent. After the second task, CECF consistently surpasses joint training, and the performance gap continues to widen.

After learning all few-shot tasks, CECF outperforms ERDA by **1.73**%, **3.04**%, and **2**% in the 2-shot, 5-shot, and 10-shot settings, respectively. Moreover, we observe that the benefits of our method continue to improve as the number of tasks increases. This indicates that our approach is particularly suitable for longer CFRL tasks and effectively mitigates catastrophic forgetting.

FewRel Benchmark with BERT Table 2 and Figure 5 show the experiment results using a BERT encoder on Fewrel with the 10-way 5-shot setting. Figure 7 shows the results on Fewrel for 10-way 2-shot and 10-shot. It is evident that our method

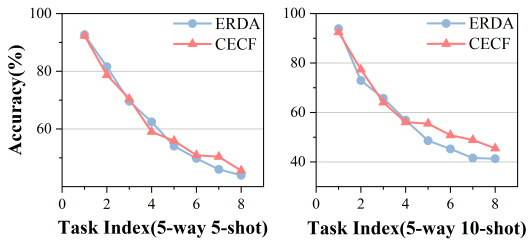


Figure 8: Comparison results of different methods on **TACRED** benchmark for **5-way 5-shot** and **10-shot** settings. CECF is better than ERDA with a BERT encoder for both settings.

Method	Task index							
	1	2	3	4	5	6	7	8
CECF	92.32	76.48	**69.73**	**65.24**	60.65	**58.99**	**57.26**	**55.49**
w.o. AW	92.4	77.49	69.08	65.03	60.89	58.42	56.33	54.16
$w.o.E_T$	**92.42**	**78.4**	69.68	64.32	**61.31**	58.23	56.86	53.84
$w.o.E_M$	92.25	77.77	68.93	64.04	59.91	57.34	55.84	54.42
$w.o.E_{M\&T}$	91.98	78.09	68.59	63.32	60.2	57.13	54.91	52.45

Table 3: Ablations on **FewRel** benchmark (**10-way 5-shot**). AW: adaptive weight; E_T: colliding effect in training data; E_M: colliding effect in memory data.

outperforms previous baselines with a BERT encoder by up to **2.92**% (10-way 10-shot), affirming the generalizability of our approach.

TACRED Benchmark Figure 8 depicts the 5-way 5-shot and 5-way 10-shot results on TACRED using a BERT encoder. It can be observed that our performance remains superior to the baseline by **1.63**% and **4.2**% on TACRED, demonstrating its strong generalization ability.

5.3. Ablation Study

We conduct several ablation experiments to analyze the contributions of different components of our method on FewRel in the 10-way 5-shot setting. It is important to note that our model is the same as ERDA when there is no causal effect. In each experiment iteration, we remove one component or combine the removal of two components: (a) the adaptive weight module AW, (b) the colliding effect E_T in training data, where we calculate the regular cross-entropy loss for classification, (c) the colliding effect E_M in memory data, (d) both the colliding effects $E_{M\&T}$ in training data and memory data. Combining the Equation 1 and Equation 9, we can observe that our AW module represents the weight λ_{kl} before the loss $\mathcal{L}_{kl}$. Therefore, in cases (c) and (d), removing the loss $\mathcal{L}_{kl}$ will inevitably result in the ineffectiveness of the AW module, where (d) is the same as the ERDA model.

From Table 3, we observe that all components improve the performance of our model. Specifically, the adaptive weight module contributes to a 1.33% accuracy boost by balancing the weights between old and new relations, particularly when there is a significant disparity in their quantities. The colliding effect in training data results in a 1.65% accuracy improvement, and it contributes more significantly over longer time steps. The colliding effect in memory data enhances performance by 1.07% accuracy, indicating improved utilization of information from the stored memory data. This demonstrates the effectiveness of establishing causal paths between old data and predictions.

K	1	2	3	5	10
Accuracy(%)	54.75	53.91	55.49	54.35	54.56

Table 4: Hyper-parameter K analysis on **FewRel** benchmark (**10-way 5-shot**).

λ	0.6	0.9	1.2	1.5	1.8
Accuracy(%)	53.97	53.64	55.49	53.61	53.90

Table 5: Hyper-parameter λ analysis on **FewRel** benchmark (**10-way 5-shot**).

Hyper-Parameter Analysis We provide a hyper-parameter analysis on FewRel in the 10-way 5-shot setting. We consider two hyper-parameters: the number of matched sentences K and the initial value of the adaptive weight λ. The results in Table 4 indicate that the best performance is achieved when $K = 3$. A larger K tends to perform better in scenarios with abundant data. However, in few-shot tasks, a larger K can lead to a higher likelihood of retrieving data from other incorrect classes, resulting in a decrease in performance. Additionally, Table 5 demonstrates that the model achieves the best accuracy with $\lambda = 1.2$. Please note that we did not perform an exhaustive search for the best hyper-parameters. Therefore, in specific cases, some carefully tuned hyper-parameters may lead to superior performance.

6. Conclusion

In Continual Few-shot Relation Learning (CFRL), causal inference can help us understand that forgetting occurs due to the loss of causal effects from old data. In order to address this issue, we propose a novel architecture called CECF from a causal graph perspective. CECF uses the colliding effect to establish two causal paths from old data to predictions, thereby enhancing the causal effects and effectively utilizing the valuable information in the memory data. Furthermore, we introduce a self-adaptive weight to balance the causal effects of new and old relation types. Comprehensive experimental results and analysis consistently demonstrate that CECF surpasses previous methods in terms of performance. In the future, we plan to further

address the issue of catastrophic forgetting from a causal perspective.

7. Limitations

Although our method partially alleviates catastrophic forgetting, it cannot guarantee an increasing gap from the baseline. Additionally, the improvements in simple training may affect the model's ability to learn new relations, thereby affecting the selection of memory data. We suspect that this may be the reason why our first two tasks did not perform well under some settings. Furthermore, since extra sentence matching needs to be computed, the training time inevitably increases.

8. Acknowledgements

This work is supported in part by the Natural Science Foundation of China (grant No.62276188), TJU-Wenge joint laboratory funding.

9. Bibliographical References

Craig Atkinson, Brendan McCane, Lech Szymanski, and Anthony Robins. 2018. Pseudo-recursal: Solving the catastrophic forgetting problem in deep neural networks. *arXiv preprint arXiv:1802.03875*.

Sagie Benaim and Lior Wolf. 2018. One-shot unsupervised cross domain translation. *advances in neural information processing systems*, 31.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. 2018. Efficient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420*.

Jinxiu Chen, Donghong Ji, Chew Lim Tan, and Zheng-Yu Niu. 2006. Relation extraction using label propagation based semi-supervised learning. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 129–136.

Tianqi Chen, Ian Goodfellow, and Jonathon Shlens. 2015. Net2net: Accelerating learning via knowledge transfer. *arXiv preprint arXiv:1511.05641*.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. 2021. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3366–3385.

Cyprien de Masson D'Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. 2019. Episodic memory in lifelong language learning. *Advances in Neural Information Processing Systems*, 32.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Li Dong, Furu Wei, Ming Zhou, and Ke Xu. 2015. Question answering over freebase with multi-column convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 260–269.

Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, et al. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

Chrisantha Fernando, Dylan Banarse, Charles Blundell, Yori Zwols, David Ha, Andrei A Rusu, Alexander Pritzel, and Daan Wierstra. 2017. Pathnet: Evolution channels gradient descent in super neural networks. *arXiv preprint arXiv:1701.08734*.

Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. 2021. Causal analysis of syntactic agreement mechanisms in neural language models. *arXiv preprint arXiv:2106.06087*.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR.

Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2019. Fewrel 2.0: Towards more challenging few-shot relation classification. *arXiv preprint arXiv:1910.07124*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*.

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, Pengfei Yu, Zhiyuan Liu, Maosong Sun, and Peng Li. 2018a. 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.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018b. Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. *arXiv preprint arXiv:1810.10147*.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Judy Hoffman, Eric Tzeng, Jeff Donahue, Yangqing Jia, Kate Saenko, and Trevor Darrell. 2013. One-shot adaptation of supervised deep convolutional models. *arXiv preprint arXiv:1312.6204*.

Xinting Hu, Kaihua Tang, Chunyan Miao, Xian-Sheng Hua, and Hanwang Zhang. 2021. Distilling causal effect of data in class-incremental learning. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 3957–3966.

Xuming Hu, Chenwei Zhang, Fukun Ma, Chenyao Liu, Lijie Wen, and Philip S Yu. 2020. Semi-supervised relation extraction via incremental meta self-training. *arXiv preprint arXiv:2010.16410*.

Zikun Hu, Xiang Li, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2018. Few-shot charge prediction with discriminative legal attributes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 487–498.

Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. 2021. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 353–362.

Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. *Advances in neural information processing systems*, 30.

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.

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

David Lopez-Paz and Marc'Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

Fei Mi, Liangwei Chen, Mengjie Zhao, Minlie Huang, and Boi Faltings. 2020. Continual learning for natural language generation in task-oriented dialog systems. *arXiv preprint arXiv:2010.00910*.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using lstms on sequences and tree structures. *arXiv preprint arXiv:1601.00770*.

Cuong V Nguyen, Yingzhen Li, Thang D Bui, and Richard E Turner. 2017. Variational continual learning. *arXiv preprint arXiv:1710.10628*.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Chengwei Qin and Shafiq Joty. 2022. Continual few-shot relation learning via embedding space regularization and data augmentation. *arXiv preprint arXiv:2203.02135*.

Sachin Ravi and Hugo Larochelle. 2016. Optimization as a model for few-shot learning. In *International conference on learning representations*.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. 2017. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. 2020. The risks of invariant risk minimization. *arXiv preprint arXiv:2010.05761*.

Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. 2016. Progressive neural networks. *arXiv preprint arXiv:1606.04671*.

Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. 2016. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pages 1842–1850. PMLR.

Bernhard Schölkopf. 2022. Causality for machine learning. In *Probabilistic and causal inference: The works of Judea Pearl*, pages 765–804.

Ang Sun, Ralph Grishman, and Satoshi Sekine. 2011. Semi-supervised relation extraction with large-scale word clustering. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 521–529.

Eleni Triantafillou, Richard Zemel, and Raquel Urtasun. 2017. Few-shot learning through an information retrieval lens. *Advances in neural information processing systems*, 30.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. 2021. Counterfactual invariance to spurious correlations: Why and how to pass stress tests. *arXiv preprint arXiv:2106.00545*.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401.

Hailin Wang, Ke Qin, Rufai Yusuf Zakari, Guoming Lu, and Jin Yin. 2022. Deep neural network-based relation extraction: an overview. *Neural Computing and Applications*, pages 1–21.

Hong Wang, Wenhan Xiong, Mo Yu, Xiaoxiao Guo, Shiyu Chang, and William Yang Wang. 2019. Sentence embedding alignment for lifelong relation extraction. *arXiv preprint arXiv:1903.02588*.

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

Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34.

Chenyan Xiong, Russell Power, and Jamie Callan. 2017. Explicit semantic ranking for academic search via knowledge graph embedding. In *Proceedings of the 26th international conference on world wide web*, pages 1271–1279.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel methods for relation extraction. *Journal of machine learning research*, 3(Feb):1083–1106.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1753–1762.

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.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Conference on Empirical Methods in Natural Language Processing*.

Junhao Zheng, Zhanxian Liang, Haibin Chen, and Qianli Ma. 2022. Distilling causal effect from miscellaneous other-class for continual named entity recognition. *arXiv preprint arXiv:2210.03980*.