Continual Relation Extraction via Sequential Multi-Task Learning

Thanh-Thien Le[1]*, Manh Nguyen[2]*, Tung Thanh Nguyen[3]*, Linh Ngo Van[2]†, Thien Huu Nguyen[4]

[1]VinAI Research, Vietnam
[2]Hanoi University of Science and Technology, Vietnam
[3]University of Michigan, USA
[4]University of Oregon, USA
v.thienlt3@vinai.io, manhnv195088@sis.hust.edu.vn, shawdsai@umich.edu,
linhnv@soict.hust.edu.vn, thienn@uoregon.edu

Abstract

To build continual relation extraction (CRE) models, those can adapt to an ever-growing ontology of relations, is a cornerstone information extraction task that serves in various dynamic real-world domains. To mitigate catastrophic forgetting in CRE, existing state-of-the-art approaches have effectively utilized rehearsal techniques from continual learning and achieved remarkable success. However, managing multiple objectives associated with memory-based rehearsal remains underexplored, often relying on simple summation and overlooking complex trade-offs. In this paper, we propose **C**ontinual **R**elation **E**xtraction via **S**equential Multi-**t**ask Learning (CREST), a novel CRE approach built upon a tailored Multi-task Learning framework for continual learning. CREST takes into consideration the disparity in the magnitudes of gradient signals of different objectives, thereby effectively handling the inherent difference between multi-task learning and continual learning. Through extensive experiments on multiple datasets, CREST demonstrates significant improvements in CRE performance as well as superiority over other state-of-the-art Multi-task Learning frameworks, offering a promising solution to the challenges of continual learning in this domain.

Introduction

In Natural Language Processing, Relation Extraction (RE) (Baldini Soares et al. 2019; Man et al. 2022; Lai et al. 2022) is the task of classifying the semantic relationships between entities/events in text into predefined relation types. Nonetheless, conventional relation extraction (Nguyen and Grishman 2015; Baldini Soares et al. 2019; Veyseh et al. 2020a,b) encounters challenges in dynamic environments characterized by a continuously expanding set of relations. This realization has prompted the development of Continual Relation Extraction (CRE) models, which recognize the ever-changing nature of information in practical settings.

The fundamental problem of Continual Relation Extraction (CRE) is catastrophic forgetting. This refers to the phenomenon where the model's performance on previous tasks declines after learning previously unseen relation types as new data emerges. To mitigate this issue, previous CRE approaches (Cui et al. 2021; Han et al. 2020; Zhao, Cui, and Hu 2023; Nguyen et al. 2023) have achieved impressive results using *memory-based* techniques (Lopez-Paz and Ranzato 2017; Shin et al. 2017a; Chaudhry et al. 2019). These methods retain a fraction of learned data in a small memory buffer, allowing the model to reinforce its past knowledge while learning new relations. For instance, Hu et al. (2022) mitigated catastrophic forgetting by integrating a classification network and a prototypical contrastive network, contrasting every newly encountered instance with the prototype of each relation stored in the replay buffer. Another state-of-the-art approach is Zhao et al. (2022), which leveraged the memory buffer to enable knowledge distillation from older tasks combined with learning the newest task using supervised contrastive learning. Despite their success, state-of-the-art CRE methods persists *two* lingering issues:

The first issues is their **reliance on a memory buffer**. In light of the diverse potential applications of CRE, many of which might involve highly confidential data, there are significant concerns regarding storing data in the long term while maintaining stringent privacy standards. In an attempt to solve this problem, recently, prompt-based, rehearsal-free methods for Continual Learning (Wang et al. 2022c,b; Smith et al. 2023) have emerged and achieved remarkable success in computer vision. However our empirical studies (Table 3) show extremely subpar results from these methods when applying to CRE.

Another problem that arises from such methods is the task of handling **multiple objectives during replay**. Among numerous advanced CRE baselines, primarily utilizing memory-based methods, there often exists at least two loss functions, corresponding to acquiring new information and strengthening previously acquired knowledge, respectively. In the above examples, Hu et al.'s method involves InfoNCE (Oord, Li, and Vinyals 2018) contrastive loss and contrastive margin loss, while Zhao et al.'s approach involves supervised contrastive loss and distillation loss. These methods oversimplistically aggregate these losses by weighted summation, hence overlook the inherent, complicated trade-offs between the objectives.

*These authors contributed equally.
†Corresponding Author

To tackle the challenge of training models with multiple objectives, gradient-based Multi-objective Optimization frameworks, designed for Multi-task Learning (MTL), seeking for a *Pareto-optimal* set of parameters (Sener and Koltun 2018; Yu et al. 2020; Liu et al. 2021a,b; Navon et al. 2022; Phan et al. 2022a), have emerged as some of the most successful approaches. Nevertheless, achieving effective application of a gradient-based MOO framework in continual NLP requires meticulous design considerations. Our empirical experiments have shown that directly applying these frameworks for CRE can lead to sub-optimal performance, as evident in Table 2. The decline in performance can be attributed to the intrinsic distinction between Continual Learning and Multi-task Learning. In the realm of Continual Learning, where tasks do not appear simultaneously, the updating direction must extend beyond seeking for the Pareto front. It should additionally take into consideration of the contrast in gradient magnitudes between the objectives, to prioritize the acquisition of novel knowledge while minimize forgetting in previously learnt tasks.

Contributions: To address the aforementioned challenges, we propose **C**ontinual **R**elation **E**xtraction via **S**equential Multi-**T**ask Learning (CREST), a novel Continual Relation Extraction framework that effectively mitigates catastrophic forgetting without the need to finetune the backbone encoder. **(i)** CREST introduces a novel gradient-based Multi-objective Optimization framework designed specifically for Continual Learning. By preserving the original proportions between gradients, our method leverages the inherent nature of continual learning, where the solution already resides in an optimized region for the seen tasks. We craft a direction that recognizes the original intensity of gradients and prioritizes learning the new task, outperforming traditional methods that seek balanced improvements. This sets a new standard in multi-objective optimization for continual learning, paving the way for unprecedented advancements in the field. **(ii)** CREST also utilizes a generative model, eliminating the dependency on explicit memory buffers for replay. Additionally, freezing the backbone encoder enables us to learn the underlying distribution and generating continuous latent relation representations, which is much more feasible than generating natural language text.

Background

Continual Relation Extraction

Continual Relation Extraction (Hu et al. 2022; Zhang et al. 2022) involves training a model, sequentially, on a series of K tasks, each with its own training set $\mathcal{D}_k$ and corresponding relation set $\mathcal{R}_k$; and $\{R_k\}_1^K$ are non-overlapping. Each dataset sample $(\boldsymbol{x}_i^k, y_i^k)$ represents input data point, comprising a natural language context and an entity pair, along with its corresponding relation label $y_i^k \in \mathcal{R}_k$.

For ease of understanding, each k-th task can be perceived as a conventional relation extraction problem, whose conventional solution framework is described in the subsequent paragraphs. The aim of Continual Relation Extraction (CRE) is to develop a model capable of acquiring knowledge from new tasks while maintaining its competence in previously encountered tasks.

The most fundamental deep learning-based framework to tackle conventional relation extraction (Ji et al. 2020; Wang and Lu 2020) involves utilizing a pretrained language model, such as BERT (Devlin et al. 2019). To maintain conciseness and align with the empirical experiments conducted in this paper, we will refer to the backbone pretrained language model as BERT from this point forward.

Given an input sentence (i.e., input context), special tokens $[E_{11}]/[E_{12}]$ and $[E_{21}]/[E_{22}]$ are inserted into the context to indicate the starting and ending positions of the head and tail entities, respectively (Baldini Soares et al. 2019; Hu et al. 2022). For example, consider the input context, where $< \cdot >$ denotes the entities whose relationship needs to be extracted:

$$< \mathrm{X} > \text{was born in} < \mathrm{Y} >.$$

After insertion of the special tokens, it becomes:

$$[\mathrm{E11}] < \mathrm{X} > [\mathrm{E12}] \text{ was born in } [\mathrm{E21}] < \mathrm{Y} > [\mathrm{E22}].$$

Subsequently, through embedding, we have the input context $\boldsymbol{x} = \boldsymbol{w}_{1:L} \in \mathbb{R}^{L \times d}$, where L denotes the length of the special-token-inserted sentence and d denotes the number of embedding dimensions. BERT encodes the tokenized input sentence $\boldsymbol{w}_{1:L}$ to obtain the contextual representations $\boldsymbol{w'}_{1:L}$. Let e_{11} and e_{21} denote the positions of $[E11]$ and $[E21]$ in the input sentence, respectively; the contextual representations $\boldsymbol{w'}_{e_{11}}$ and $\boldsymbol{w'}_{e_{21}}$ of the $[E11]$ and $[E21]$ tokens are concatenated to obtain the input relation representation $\boldsymbol{z}$.

Then, this relation representation $\boldsymbol{z}$ is passed through a Multilayer Perceptron (MLP) classifier to derive a feature vector $\boldsymbol{h}$. This feature vector h is then fed into a linear layer followed by a softmax layer, resulting in a probability distribution p over the predefined relation types:

$$p = Softmax(Linear(\boldsymbol{h})),$$

where $\boldsymbol{h} = MLP(\boldsymbol{z})$ and $\boldsymbol{z} = [\boldsymbol{w'}_{e_{11}}, \boldsymbol{w'}_{e_{21}}]$. Let $\mathcal{D}$, $\mathcal{N}$ denote the training dataset and the number of instances, respectively. We have the training loss as the cross-entropy loss:

$$L_{re} = \frac{1}{\mathcal{N}} \sum_{\boldsymbol{x} \in \mathcal{D}} \log p. \tag{1}$$

In our proposed method, the embeddings layer and BERT are frozen throughout the learning process. The embeddings corresponding with the special tokens $[E_{11}, E_{12}, E_{21}, E_{22}]$ are also kept frozen after completing the first task. This is different from state-of-the-art CRE baselines, which requires finetuning BERT.

Continual Learning

The continual settings of Continual Relation Extraction (CRE) fall into the category of class-incremental learning (Hu et al. 2022), which constitutes one of the three popular scenarios in Continual Learning (Ke and Liu 2022; Van de Ven and Tolias 2019). CRE can be perceived as a class-incremental learning (CIL) problem because the model needs to learn to adapt and classify new relations while avoiding catastrophic forgetting on previously learned ones. CIL entails the training of a learning agent on a singular

prediction problem; hence, the mention of different "tasks" should be interpreted as different training phases wherein novel classes are encountered, rather than denoting distinct prediction tasks. During testing, the model is anticipated to predict using the cumulative set of encountered labels, without explicit task identity. CIL is often considered the most difficult configuration among the three scenarios, as the unavailability of task identity imposes various constraints on the selection of methodologies.

Several state-of-the-art (SOTA) methods have emerged to address the challenge of catastrophic forgetting using three approaches: *Regularization-based approaches* (Jung et al. 2020; Phan et al. 2022b; Linh et al. 2022; Hai et al. 2023), *Architecture-based approaches* (Hung et al. 2019; Liu, Schiele, and Sun 2021), and *Replay-based approaches* (Farajtabar et al. 2020; Hou et al. 2019; Shin et al. 2017b). These three approaches represent prominent strategies employed to tackle catastrophic forgetting in continual learning, and memory-based techniques have been proven to be the most effective in the field of Natural Language Processing (de Masson d'Autume et al. 2019).

Gradient-Based Multi-Objective Optimization

Effective management of multiple objectives becomes crucial when training a Continual Learning model, especially when incorporating a replay process to reinforce previously learned knowledge. A Multi-objective Optimization problem is formulated as follows:

Let $\boldsymbol{\theta}$ denote the model parameters within a feasible set $\boldsymbol{\Theta}$, L_i as the i-th objective (i.e., loss), and K as the total number of objectives. We aim to minimize, simultaneously, all K losses:

$$\min_{\boldsymbol{\theta}} [L_1(\boldsymbol{\theta}), L_2(\boldsymbol{\theta}), ..., L_K(\boldsymbol{\theta})].$$

Given $\boldsymbol{\theta}^1$ and $\boldsymbol{\theta}^2$ as two feasible solutions to the problem above, we state that $\boldsymbol{\theta}^1$ **dominates** $\boldsymbol{\theta}^2$ if and only if $\boldsymbol{\theta}^1$ can enhance at least one objective without negatively impacting any other objectives, as compared to $\boldsymbol{\theta}^2$. A feasible solution is termed as **Pareto-optimal** if it is not dominated by any other solutions. The set of Pareto-optimal solutions is referred to as the **Pareto front**.

A straightforward method, which is popular among state-of-the-art CRE baselines, to solve the Multi-objective Optimization problem is to optimize the weighted sum of the K original objectives, $L_{total} = \sum_{i=1}^{K} \lambda_i L_i$, and then search for the optimal set of λ. However, this simple approach may encounter challenges when dealing with a non-convex Pareto front or when certain objectives have conflicting gradients (Sener and Koltun 2018). Moreover, it also requires searching for the best set of hyperparameters $\{\lambda_i\}_{i=1}^{K}$, which is both time- and data-consuming. Consequently, there needs to be more sophisticated approaches to solve such problems; currently, gradient-based MOO frameworks, designed for Multi-task Learning (MTL), such as **PCGrad** (Yu et al. 2020), **CAGrad** (Liu et al. 2021a), **IMTL** (Liu et al. 2021b), and **NashMTL** (Navon et al. 2022) are some of the most notable methods. They share a common idea of modeling the updating direction as a linear combination of individual gradients, i.e., $\Delta\boldsymbol{\theta} = \sum_1^K \alpha_i \boldsymbol{g}_i$; their differences lie in their strategies of choosing α. α can be perceived as a dynamic version, which changes at each descending step, of the coefficients λ in the weighted loss approach. However, even with those methods, given the difference between the nature of MTL and Continual Learning, directly applying these methods to our problems might yield deteriorated results as we have mentioned earlier.

Methodology

Reinforce Continual Relation Extraction via Representation Generation

A notable obstacle in replay-based CRE, and also in replay-based Continual Learning in general, arises from the limited size of the replay buffer in contrast to the continuous accumulation of data. This situation, apart from generating concerns about compromising privacy, introduces the risk of the model overfitting to the small memory buffer, thereby weakening the efficiency of replaying.

To address these issues and diversify the memory buffer, generative models, such as Variational Autoencoder (VAE) (Kingma and Welling 2013), Conditional Variational Autoencoder (cVAE) (Sohn, Lee, and Yan 2015), or Diffusion Models (Nichol and Dhariwal 2021), prove effective by synthesizing representations for each relation type. Notably, the choice of a generative model for continual learning should balance efficacy with cost, prompting the use of economical models like Gaussian Mixture Models (GMMs).

It is noteworthy that, since we keep the BERT encoder frozen during training and the embedding layer fixed after the first task, we can directly fit the generative model to the relation representations $\boldsymbol{z}$ of all the data, which is much more practical and feasible than fitting the model to the original embedding matrices of the text instances. As we have mentioned above, the relation representation $\boldsymbol{z}$ is what we obtained by concatenating the BERT-encoded representations at the position of the $[E11]$ and $[E21]$ tokens:

$$\boldsymbol{z} = \Big[f_b(\boldsymbol{x})[e_{11}, :],\ f_b(\boldsymbol{x})[e_{21}, :] \Big],$$

where f_b denotes the mapping function corresponding to *BERT*, and e_{11} and e_{21} denote the positions of $[E11]$ and $[E21]$ in the input sentence, respectively. This approach is only possible thanks to the freezing of all the blocks prior to $\boldsymbol{z}$, which eliminates any changes to the representations' distributions after each updating step of the model.

After training the model on task $k-1$, for each relation type $r \in \mathcal{R}_{k-1}$, we use a Gaussian Mixture Model (GMM) to learn the underlying data distribution of the relation representations $\boldsymbol{z}$ corresponding to the data from that specific label and store this distribution for future sampling. In the next task (k), for each relation type $r \in \mathcal{R}_{k-1}$, we use its corresponding learnt distribution to sample $\tilde{n}$ synthetic relation representations $\tilde{\boldsymbol{z}}_n \sim \sum_{i=1}^{K} \pi_i^r \mathcal{N}(\mu_i^r, \Sigma_i^r), n = 1, 2, ...\tilde{n}$ where K is the number of GMM components; π_i, μ_i and Σ_i

are the mixing coefficient, mean and diagonal covariance of i^{th} Gaussian distribution, respectively.

We denote the generated set as $\tilde{M}$. $\tilde{M}$ will facilitate the model in reinforcing its previous knowledge via knowledge distillation; the distillation loss in the context of our method is written as follows:

$$L_d = \frac{-1}{|\tilde{M}|} \sum_{\tilde{\boldsymbol{z}} \in \tilde{M}} p_{\tilde{\boldsymbol{z}}}^{k-1} \log p_{\tilde{\boldsymbol{z}}}^{k}, \tag{2}$$

where $|\tilde{M}|$ denotes the cardinality of $\tilde{M}$; $p_{\tilde{\boldsymbol{z}}}^{k-1}$ and $p_{\tilde{\boldsymbol{z}}}^{k}$ denote the probability distributions over learned relation types obtained from forwarding $\tilde{\boldsymbol{z}}$ through the old and current models, respectively. The distillation loss facilitates continual learning by transferring knowledge from previous tasks to the current task, enabling the model to retain previously learned information and avoid catastrophic forgetting.

Continual Relation Extraction With Sequential Multi-Task Learning

The training process of our model is a Multi-objective Optimization (MOO) problem, where we have to minimize two objectives simultaneously: L_{re} and L_d, as represented by equations (1) and (2), respectively. Multi-task Learning (MTL) frameworks, such as those discussed earlier, can be employed to address this problem.

Nevertheless, it is essential to recognize that the current MTL frameworks were not originally developed for continual learning, and there are fundamental differences between the two paradigms. When applying these MTL methods to Continual Relation Extraction, they often fail to consider the varying priority between different objectives. Specifically, at the beginning of training a new task, the objectives associated with maintaining performance on previously learned tasks are already in a better state than the objective related to learning new knowledge; however, state-of-the-art MTL approaches do not have mechanisms to leverage this information since they were not designed to handle such unique challenges encountered in continual learning settings.

Assume that after completing task k, we have obtained a good solution which works well with tasks from 1 to k. Therefore, when moving on to task $k+1$, the solution already lies within the proximity of the local optimal region of the loss associated with previous knowledge, namely the distillation loss (L_d). On the other hand, the model is completely untrained on task $k+1$. Based on this observation, we propose a novel gradient-based MOO algorithm, **A**daptive **U**nified **G**radient **D**escent, which allows the learning process of the model to recognize the difference in magnitudes of different gradient signals, thereby prioritize acquiring new knowledge.

Adaptive Unified Gradient Descent

In order to achieve the aforementioned learning paradigm, we can take advantage of the fact that the gradient signal corresponding to the objective of learning new knowledge are likely much stronger than that of re old knowledge.

Let $\boldsymbol{\theta}$ represents the learnable model parameters. Through backpropagation, we derive T gradients $\{\boldsymbol{g}_t = \nabla_{\boldsymbol{\theta}} L_t\}_{t=1}^{T}$ from the raw losses $\{L_t\}_{t=1}^{T}$. These gradients signify the optimal update directions for each respective objective. By modeling the final updating gradient as a linear combination of the individual gradients, our objective is to determine scalar coefficients α that enable the updating gradient $\boldsymbol{g} = \sum_{t=1}^{T} \alpha_t \boldsymbol{g}_t, \sum \alpha_t = 1$ to prioritize gradients based on their strengths, by ensuring that the projections of $\boldsymbol{g}$ onto $\boldsymbol{g}_t$ are proportionate to their original magnitudes.

Algorithm 1: Adaptive Unified Gradient Descent for CRE

Input: Model parameters $\boldsymbol{\theta}$ and differentiable loss functions L_d and L_{re}
Parameter: Learning rate η
Output: Updated parameter $\boldsymbol{\theta}^*$

1: **for** each $t \in [d, re]$ **do**
2: Compute gradient $\boldsymbol{g}_t := \nabla_{\boldsymbol{\theta}} L_t(\boldsymbol{\theta})$
3: Compute gradient unit vector $\boldsymbol{u}_t := \boldsymbol{g}_t / ||\boldsymbol{g}_t||$
4: **end for**
5: Calculate gradient differences $\boldsymbol{D}^\top := \left[\boldsymbol{g}_d^\top - \boldsymbol{g}_{re}^\top\right]$.
6: Calculate magnitude-scaled gradient unit differences:

$$\boldsymbol{U}^\top := \left[||\boldsymbol{g}_{re}||\boldsymbol{u}_d^\top - ||\boldsymbol{g}_d||\boldsymbol{u}_{re}^\top\right].$$

7: Calculate scalar coefficients for the objectives:

$$[\alpha_{re}] = \boldsymbol{g}_d \boldsymbol{U}^\top (\boldsymbol{D}\boldsymbol{U}^\top)^{-1},$$
$$\alpha_d = 1 - \alpha_{re}.$$

8: Update model parameter:

$$\boldsymbol{\theta}^* = \boldsymbol{\theta} - \eta \sum_{i \in \{d, re\}} \alpha_i \boldsymbol{g}_i$$

Let $\boldsymbol{u}_t = \boldsymbol{g}_t / ||\boldsymbol{g}_t||$ denote the unit-norm vector of $\boldsymbol{g}_t$, we want to achieve:

$$\frac{\boldsymbol{g}\boldsymbol{u}_1^\top}{||\boldsymbol{g_1}||} = \frac{\boldsymbol{g}\boldsymbol{u}_t^\top}{||\boldsymbol{g_t}||} \Leftrightarrow \boldsymbol{g}(||\boldsymbol{g}_t||\boldsymbol{u}_1 - ||\boldsymbol{g}_1||\boldsymbol{u}_t)^\top = 0,\ \forall 2 \le t \le T, \tag{3}$$

Similar to IMTL (Liu et al. 2021b), we will use the following notations in the upcoming equations:

$$\boldsymbol{\alpha} := [\alpha_2, \alpha_3, \ldots, \alpha_T] \Rightarrow \alpha_1 = 1 - \mathbf{1}\boldsymbol{\alpha}^\top,$$
$$\boldsymbol{G}^\top := \left[\boldsymbol{g}_2^\top, \boldsymbol{g}_3^\top, \ldots, \boldsymbol{g}_T^\top\right],$$
$$\boldsymbol{U}^\top := \big[||\boldsymbol{g}_2||\boldsymbol{u}_1^\top - ||\boldsymbol{g}_1||\boldsymbol{u}_2^\top,\ ||\boldsymbol{g}_3||\boldsymbol{u}_1^\top - ||\boldsymbol{g}_1||\boldsymbol{u}_3^\top,$$
$$\ldots,\ ||\boldsymbol{g}_T||\boldsymbol{u}_1^\top - ||\boldsymbol{g}_1||\boldsymbol{u}_T^\top\big],$$

where $\mathbf{1}$ denotes the all-one row vector. Denote $\mathbf{0}$ as the all-zero row vector, substitute $\boldsymbol{g}$ with $\sum_{t=1}^{T} \alpha_t \boldsymbol{g}_t$, and combine with the above notations, we can rewrite Equation (3) as:

$$\left[1 - \mathbf{1}\boldsymbol{\alpha}^\top, \boldsymbol{\alpha}\right] \begin{bmatrix} \boldsymbol{g}_1 \\ \boldsymbol{G} \end{bmatrix} \boldsymbol{U}^\top = \mathbf{0}. \tag{4}$$

From there, we can solve Equation 4 as in Liu et al.'s paper:

$$(4) \Leftrightarrow \boldsymbol{\alpha}\big(\mathbf{1}^\top \boldsymbol{g}_1 - \boldsymbol{G}\big)\boldsymbol{U}^\top = \boldsymbol{g}_1 \boldsymbol{U}^\top$$

Define $\boldsymbol{D}^\top := \boldsymbol{g}_1^\top \mathbf{1} - \boldsymbol{G}^\top = \big[\boldsymbol{g}_1^\top - \boldsymbol{g}_2^\top,\ \boldsymbol{g}_1^\top - \boldsymbol{g}_3^\top,\ \ldots,\ \boldsymbol{g}_1^\top - \boldsymbol{g}_T^\top\big]$, we have:

$$\boldsymbol{\alpha}\boldsymbol{D}\boldsymbol{U}^\top = \boldsymbol{g}_1 \boldsymbol{U}^\top \Leftrightarrow \boldsymbol{\alpha} = \boldsymbol{g}_1 \boldsymbol{U}^\top \big(\boldsymbol{D}\boldsymbol{U}^\top\big)^{-1}$$

From there we have the final scalar coefficients:

$$[\alpha_2, \ldots, \alpha_T] = \boldsymbol{\alpha} = \boldsymbol{g}_1 \boldsymbol{U}^\top \left(\boldsymbol{D}\boldsymbol{U}^\top\right)^{-1},$$
$$\alpha_1 = 1 - \sum_{t=2}^{T} \alpha_t, \tag{5}$$

and we can achieve the updating gradient $\boldsymbol{g} = \sum_{t=1}^{T} \alpha_t \boldsymbol{g}_t$. In summary, the essential steps of our proposed MOO framework, AUGD, is outlined in Algorithm 1.

As we talk about this method, it is necessary to revise the fact that our backbone BERT is frozen during training. This is absolutely crucial due to the fact that, when applying AUGD, or any other gradient-based MOO methods mentioned earlier, backpropagation needs to be executed T times at every descending step to obtain the gradient of each task loss with respect to the model parameters. This would result in an explosion of training time if we have to calculate backpropagation through the gigantic backbone LLM. By freezing BERT, we can significantly reduce the training time and make using gradient-base MOO methods possible.

Experimental Results

Datasets & Settings

We evaluate our proposed method and all baselines on two English datasets:

FewRel (Han et al. 2018) dataset comprises 80 relation types and contains a total of 56,000 samples. To make it suitable for experiments in a continual settings, the dataset is split into 10 non-overlapping groups, simulating the sequential arrival of data for 10 tasks. In line with Wang et al.'s paper (2019), this paper adopts the same configurations and utilizes the original training set and validation set as the foundation for conducting experiments. The FewRel dataset is a widely-used benchmark in the field of relation extraction, providing a diverse range of relations and samples for evaluating CRE models.

TACRED (Zhang et al. 2017) dataset presents an imbalanced scenario for relation extraction (RE) with 42 relations, including the "$no\_relation$" class, and a total of 106,264 samples. To maintain consistency with prior works, our work follows the experimental settings as in Cui et al.'s paper.

Additional details of our experiments can be found below:

- Batch size: 16, similar to CRL (Zhao et al. 2022)
- [1]Learning rate: $\{10^{-5}, \mathbf{2 \times 10^{-5}}, 10^{-4}\}$
- [1]Number of embeddings training epoch: $\{10, 20, \mathbf{50}\}$
- [1]Number of classifier training epoch: $\{100, 300, \mathbf{500}\}$
- [1]Number of GMM components: $\{\mathbf{1}, 3, 5\}$
- [1]Number of GMM samples: $\{64, 128, \mathbf{256}, 512\}$
- Computing infrastructure: Single NVIDIA A100 40GB. PyTorch 2.0.0+cu117 and Huggingface Transformer 4.33.0 are used to implement the models.
- Evaluation metric: mean overall accuracy after each task using 5 different random seeds (Zhao et al. 2022).

[1]Search range; **bold** indicates best value.

Baselines

We evaluate our proposed techniques by contrasting them against a range of established benchmarks in the context of Continual Relation Extraction (CRE):

EA-EMR by Wang et al. combines memory replay and embedding alignment to tackle the problem of catastrophic forgetting. **CML** (Wu et al. 2021) adopts a curriculum-meta learning strategy to effectively handle order-sensitivity and the issue of catastrophic forgetting in CRE. Han et al. put forth **EMAR + BERT**, which relies on memory activation and reconsolidation to preserve past knowledge. **RP-CRE** by Cui et al. utilizes a memory network to refine sample embeddings using relation prototypes, aiming to counteract the problem of catastrophic forgetting. Hu et al.'s **CRECL** merges a classification network with a prototypical contrastive network to mitigate the challenges associated with catastrophic forgetting. By adopting a contrastive replay mechanism and knowledge distillation, Zhao et al. present **CRL** in as a means to sustain the acquired knowledge. Building upon CRL, Wang et al. enhance it through the integration of a data augmentation mechanism, leading to their model **CRL+ACA** which bolsters the model's robustness. In the case of **CRE-DAS** (Zhao, Cui, and Hu 2023), they leverage memory-insensitive relation prototypes and memory augmentation to overcome overfitting, while also introducing integrated training and focal knowledge distillation to enhance performance on analogous relations. Most recently, Xia et al. propose **CDec+ACA**, a classifier decomposition framework aimed at addressing representation biases through robust representation learning while simultaneously retaining prior knowledge.

Main Results

The results in Table 1 demonstrate how our method compares to the current most successful CRE baselines. Even without retaining training data or directly fine-tuning BERT, CREST still produced nearly equivalent results to the current state-of-the-art (SOTA) baselines on both datasets. On TACRED, CREST even achieved a SOTA accuracy of 79.4% after training 10 tasks, 0.3% higher than the best result from the baselines. Similarly, on FewRel, the proposed method achieved a similar accuracy after 10 tasks in comparison to the previous SOTA result. This observation highlights the importance and effectiveness of our contributions. By utilizing generative modeling combined with a MOO method dedicated for Continual Learning, we allow the model to reinforce its previous knowledge while learning new relation types and find the suitable updating direction when working with multiple objectives associated with past and current knowledge. These contributions enable a finetuning-free CRE model like CREST to produce comparable, and in some cases, better performance compared to SOTA rehearsal-based, finetuning-included methods.

Effects of Choice of Gradient-Based MOO Method

To examine the effectiveness of our novel gradient-based MOO method designed for Continual Learning (CL), we have benchmarked our method in comparison against other

FewRel										
Model	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
EA-EMR	89.0	69.0	59.1	54.2	47.8	46.1	43.1	40.7	38.6	35.2
CML	91.2	74.8	68.2	58.2	53.7	50.4	47.8	44.4	43.1	39.7
EMAR+BERT	**98.8**	89.1	89.5	85.7	83.6	84.8	79.3	80.0	77.1	73.8
RP-CRE	97.9	92.7	91.6	89.2	88.4	86.8	85.1	84.1	82.2	81.5
CRECL	98.0	94.7	92.4	90.7	89.4	87.1	85.9	85.0	84.0	82.1
CRL	98.2	94.6	92.5	90.5	89.4	87.9	86.9	85.6	84.5	83.1
CRL+ACA	98.3	95.0	92.6	91.3	90.4	89.2	87.6	87.0	86.3	84.7
CRE-DAS	98.1	**95.8**	93.6	91.9	**91.1**	89.4	88.1	86.9	85.6	84.2
CDec+ACA	98.4	95.4	93.2	92.1	91.0	89.7	**88.3**	**87.4**	**86.4**	**84.8**
CREST (Ours)	98.7	93.6	**93.8**	**92.3**	91.0	**89.9**	87.6	86.7	86.0	**84.8**
TACRED										
Model	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
EA-EMR	47.5	40.1	38.3	29.9	24	27.3	26.9	25.8	22.9	19.8
CML	57.2	51.4	41.3	39.3	35.9	28.9	27.3	26.9	24.8	23.4
EMAR+BERT	96.6	85.7	81.0	78.6	73.9	72.3	71.7	72.2	72.6	71.0
RP-CRE	97.6	90.6	86.1	82.4	79.8	77.2	75.1	73.7	72.4	72.4
CRECL	97.3	93.6	90.5	86.1	84.6	82.1	79.4	77.6	77.9	77.4
CRL	97.7	93.2	89.8	84.7	84.1	81.3	80.2	79.1	79.0	78.0
CRL+ACA	**98.0**	92.1	90.6	85.5	84.4	82.2	80.0	78.6	78.8	78.1
CRE-DAS	97.7	**94.3**	**92.3**	**88.4**	**86.6**	**84.5**	**82.2**	**81.1**	**80.1**	79.1
CDec+ACA	97.7	92.8	91.0	86.7	85.2	82.9	80.8	80.2	78.8	78.6
CREST (Ours)	97.3	91.4	82.3	82.5	79.2	75.8	78.8	77.4	78.6	**79.4**

Table 1: Performance of CREST (%) on all observed relations at each stage of learning, in comparison with SOTA CRE baselines. The results of the baselines are directly taken from (Xia et al. 2023) and (Zhao, Cui, and Hu 2023).

SOTA MOO methods, namely PCGrad (Yu et al. 2020), CAGrad (Liu et al. 2021a), IMTL (Liu et al. 2021b), and Nash-MTL (Navon et al. 2022), in the context of CRE. The empirical results are presented in Table 2.

As mentioned earlier, due to the fact that Multi-task Learning (MTL) is inherently different from Continual Learning, directly applying gradient-based MOO methods built for MTL into our problem of CRE might yields sub-optimal results. The results in table 2 concurs with this statement. On TACRED, the best accuracy after 10 tasks achieved with a previous SOTA MOO method is 77.4% from PCGrad, approximately 2% lower than using the proposed method AUGD. Besides PCGrad, every other SOTA MOO method achieved a worse accuracy after 10 tasks than completely not using any MOO methods: IMTL achieved the next-best accuracy of 74.3% after 10 tasks, which is not only 5.1% lower than the accuracy of AUGD, but also 2.6% worse than not using an MOO method at all.

We can observe similar outcomes from experiments conducted on FewRel. On FewRel, although CAGrad yielded a 0.5% higher accuracy after 10 tasks compared to not using any MOO methods, our proposed AUGD still outperformed it by 1%. On the other hand, PCGrad, IMTL, and Nash-MTL all resulted in a decline of terminal accuracy after 10 tasks in comparison to not using any MOO methods. These results are concrete evidence, showcasing AUGD's effectiveness in enhancing the performance of CREST and highlighting the importance of designing an MOO framework dedicated to Continual Learning.

Comparison to MOO Rehearsal-Free Continual Learning Methods

Recently, prompt-based methods for Continual Learning (CL) have emerged as rehearsal-free and efficient-finetuning CL approaches, gaining attention due to their remarkable success in computer vision tasks, even surpassing SOTA memory-based methods (Wang et al. 2022c,b; Smith et al. 2023). Rather than directly finetuning the pretrained Transformer-based encoder, these approaches focus on tuning auxiliary embeddings known as prompts. These prompts can be dynamically inserted into the training process, adapted to individual instance features (Smith et al. 2023), or tailored to task-specific features (Wang et al. 2022c,b).

In similarity to CREST, these prompt-based methods neither require full finetuning of the backbone encoder nor rely on an explicit memory buffer. As such, it becomes essential to conduct a comparison between CREST and these prompt-based continual learning methods. Specifically, we will evaluate the proposed model against the state-of-the-art rehearsal-free continual learning baselines, namely L2P (Wang et al. 2022c), DualPrompt (Wang et al. 2022b), and CODA-Prompt (Smith et al. 2023). The assessment will be conducted in two distinct scenarios: one in which these techniques operate completely rehearsal-free, and the other involving their utilization of a small memory buffer for replay purposes. Given that these baselines were originally built for computer vision tasks, we re-implement them, to suit the domain of Continual Relation Extraction (CRE), employing BERT (Devlin et al. 2019) as the backbone encoder.

Table 3 illustrates the performance of current state-of-the-art prompt-based, rehearsal-free continual learning meth-

FewRel										
Model	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
CREST w/o MOO	98.7	93.5	93.1	92.2	90.9	89.5	**87.9**	86.1	85.4	83.3
CREST w/ **PCGrad**	98.7	92.0	92.3	90.8	89.0	87.9	84.7	83.2	82.8	81.2
CREST w/ **CAGrad**	98.7	92.8	93.2	91.9	90.5	89.6	87.4	86.4	84.9	83.8
CREST w/ **IMTL**	98.7	89.1	92.2	91.8	90.5	89.0	87.4	86.4	83.6	81.8
CREST w/ **Nash-MTL**	98.7	92.1	92.0	89.8	87.2	86.9	80.7	82.5	82.4	80.0
CREST	98.7	**93.6**	**93.8**	**92.3**	**91.0**	**89.9**	87.6	**86.7**	**86.0**	**84.8**
TACRED										
Model	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
CREST w/o MOO	97.3	**91.8**	81.8	79.4	77.1	**76.2**	77.9	74.9	77.5	76.9
CREST w/ **PCGrad**	97.3	91.4	80.0	82.3	76.1	74.8	77.0	76.7	78.2	77.4
CREST w/ **CAGrad**	97.3	91.4	81.8	**82.5**	77.1	76.1	74.2	75.0	69.6	71.4
CREST w/ **IMTL**	97.3	89.4	81.8	80.8	75.0	75.6	75.1	67.2	76.7	74.3
CREST w/ **Nash-MTL**	97.3	91.5	82.8	81.9	77.2	76.0	75.0	73.4	77.0	75.8
CREST	97.3	91.4	**82.3**	**82.5**	**79.2**	75.8	**78.8**	**77.4**	**78.6**	**79.4**

Table 2: Results of ablation studies on different MOO methods when applied to CREST.

FewRel										
Model	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
L2P	98.6	47.9	36.5	25.7	21.7	19.2	14.1	11.9	14.4	11.6
L2P w/ buffer	98.6	90.3	81.8	78.3	73.5	71.2	68.3	66.9	65.6	62.3
Dual-P	98.8	49.2	41.4	27.7	20.9	20.2	14.1	12.2	14.0	11.5
Dual-P w/ buffer	98.8	94.4	91.5	89.8	88.3	85.9	83.5	81.2	79.0	75.3
CODA-P	98.8	52.4	42.4	28.6	24.4	21.6	14.9	12.1	15.1	11.8
CODA-P w/ buffer	**98.8**	**94.5**	92.0	90.5	89.4	87.8	86.5	85.1	84.0	82.6
CREST	98.7	93.6	**93.8**	**92.3**	**91.0**	**89.9**	**87.6**	**86.7**	**86.0**	**84.8**
TACRED										
Model	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
L2P	96.6	40.8	32.4	24.1	19.7	15.4	13.6	9.1	10.8	9.7
L2P w/ buffer	96.6	91.4	86.9	82.0	79.3	74.7	72.7	69.7	68.8	66.6
Dual-P	96.6	40.0	33.4	23.8	19.9	15.2	13.8	9.6	11.1	11.4
Dual-P w/ buffer	96.6	92.1	86.1	81.9	79.5	76.0	74.4	72.1	71.6	70.0
CODA-P	95.9	40.8	34.2	24.7	19.9	15.1	14.1	12.1	12.3	12.2
CODA-P w/ buffer	95.9	**92.5**	**87.6**	83.7	**81.7**	**79.5**	77.4	76.4	75.5	73.9
CREST	**97.3**	91.4	82.3	**82.5**	79.2	75.8	**78.8**	**77.4**	**78.6**	**79.4**

Table 3: Comparison of our method's performance (%) with state-of-the-art rehearsal-free Continual Learning baselines on all observed relations at each learning stage. The results are obtained from our own implementations. For baselines that utilize a buffer, the buffer size consists of 10 samples from each relation type. Dual-P is short for DualPrompt (Wang et al. 2022b); CODA-P is short for CODA-Prompt (Smith et al. 2023).

ods in CRE in comparison to our method CREST. As we can see, they yield notably inferior results in the context of CRE. This observation strongly suggests that these approaches currently encounter challenges in effectively mitigating catastrophic forgetting across diverse domains. Upon integrating a memory buffer to bolster knowledge retention, their performance demonstrates substantial enhancements. Nevertheless, due to the absence of targeted strategies tailored specifically for CRE, their efficacy remains inferior to that of state-of-the-art CRE methods outlined in Table 1. In contrast, our proposed approach significantly outperforms all rehearsal-free baselines for continual learning in the CRE domain by substantial margins. CREST attains final accuracies of 84.8% and 79.4% on FewRel and TACRED, respectively. These results surpass the best outcomes achieved among the baselines, when they are supported by a small memory buffer, by 2.2% and 5.5% respectively.

Conclusion

In this paper, we have presented CREST, a novel method for Continual Relation Extraction (CRE) that effectively addresses the challenges of catastrophic forgetting and efficient knowledge acquisition. Acknowledging limitations of the current methodologies, we propose freezing the backbone encoder (BERT), employing a data generation technique using Gaussian Mixture Models (GMM), and proposing a novel gradient-based MOO framework which prioritizes new knowledge acquisition. CREST has achieved remarkable performance on the FewRel and TACRED datasets, producing extremely competitive results in comparison to SOTA CRE baselines and outperforming prompt-based rehearsal-free baselines for continual learning in the realm of Continual Relation Extraction.

Acknowledgements

This research has been supported by the NSF grant CNS-1747798 to the IUCRC Center for Big Learning and the NSF grant # 2239570.

References

Baldini Soares, L.; FitzGerald, N.; Ling, J.; and Kwiatkowski, T. 2019. Matching the Blanks: Distributional Similarity for Relation Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy.

Chaudhry, A.; Ranzato, M.; Rohrbach, M.; and Elhoseiny, M. 2019. Efficient Lifelong Learning with A-GEM. In *International Conference on Learning Representations*. New Orleans, LA, USA.

Cui, L.; Yang, D.; Yu, J.; Hu, C.; Cheng, J.; Yi, J.; and Xiao, Y. 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)*. Online.

de Masson d'Autume, C.; Ruder, S.; Kong, L.; and Yogatama, D. 2019. Episodic Memory in Lifelong Language Learning. In *Advances in Neural Information Processing Systems*. Vancouver, BC, Canada.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

Farajtabar, M.; Azizan, N.; Mott, A.; and Li, A. 2020. Orthogonal gradient descent for continual learning. In *International Conference on Artificial Intelligence and Statistics*. virtual.

Hai, N. L.; Nguyen, T.; Van, L. N.; Nguyen, T. H.; and Than, K. 2023. Continual variational dropout: a view of auxiliary local variables in continual learning. *Machine Learning*, 1–43.

Han, X.; Dai, Y.; Gao, T.; Lin, Y.; Liu, Z.; Li, P.; Sun, M.; and Zhou, J. 2020. Continual Relation Learning via Episodic Memory Activation and Reconsolidation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online.

Han, X.; Zhu, H.; Yu, P.; Wang, Z.; Yao, Y.; Liu, Z.; and Sun, M. 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*. Brussels, Belgium.

Hou, S.; Pan, X.; Loy, C. C.; Wang, Z.; and Lin, D. 2019. Learning a Unified Classifier Incrementally via Rebalancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA.

Hu, C.; Yang, D.; Jin, H.; Chen, Z.; and Xiao, Y. 2022. Improving Continual Relation Extraction through Prototypical Contrastive Learning. In *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea.

Hung, C.-Y.; Tu, C.-H.; Wu, C.-E.; Chen, C.-H.; Chan, Y.-M.; and Chen, C.-S. 2019. Compacting, Picking and Growing for Unforgetting Continual Learning. In *Advances in Neural Information Processing Systems*. Vancouver, BC, Canada.

Ji, B.; Yu, J.; Li, S.; Ma, J.; Wu, Q.; Tan, Y.; and Liu, H. 2020. Span-based Joint Entity and Relation Extraction with Attention-based Span-specific and Contextual Semantic Representations. In *IJCAI*.

Jung, S.; Ahn, H.; Cha, S.; and Moon, T. 2020. Continual Learning with Node-Importance based Adaptive Group Sparse Regularization. In *Advances in Neural Information Processing Systems*. virtual.

Ke, Z.; and Liu, B. 2022. Continual Learning of Natural Language Processing Tasks: A Survey. *arXiv preprint arXiv:2211.12701*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Lai, V.; Man, H.; Ngo, L.; Dernoncourt, F.; and Nguyen, T. 2022. Multilingual SubEvent Relation Extraction: A Novel Dataset and Structure Induction Method. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 5559–5570.

Linh, N. V.; Hai, N. L.; Pham, H.; and Than, K. 2022. Auxiliary Local Variables for Improving Regularization/Prior Approach in Continual Learning. In *Advances in Knowledge Discovery and Data Mining - Pacific-Asia Conference, PAKDD, Proceedings, Part I*. Chengdu, China.

Liu, B.; Liu, X.; Jin, X.; Stone, P.; and Liu, Q. 2021a. Conflict-Averse Gradient Descent for Multi-task learning. In *Advances in Neural Information Processing Systems*. virtual.

Liu, L.; Li, Y.; Kuang, Z.; Xue, J.-H.; Chen, Y.; Yang, W.; Liao, Q.; and Zhang, W. 2021b. Towards Impartial Multi-task Learning. In *International Conference on Learning Representations*. virtual.

Liu, Y.; Schiele, B.; and Sun, Q. 2021. Adaptive aggregation networks for class-incremental learning. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*. virtual.

Lopez-Paz, D.; and Ranzato, M. A. 2017. Gradient Episodic Memory for Continual Learning. In *Advances in Neural Information Processing Systems*. Long Beach, CA, USA.

Man, H.; Ngo, N. T.; Van, L. N.; and Nguyen, T. H. 2022. Selecting optimal context sentences for event-event relation extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10), 11058–11066.

Navon, A.; Shamsian, A.; Achituve, I.; Maron, H.; Kawaguchi, K.; Chechik, G.; and Fetaya, E. 2022. Multi-Task Learning as a Bargaining Game. *arXiv preprint arXiv:2202.01017*.

Nguyen, H.; Nguyen, C.; Ngo, L.; Luu, A.; and Nguyen, T. 2023. A Spectral Viewpoint on Continual Relation Extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 9621–9629.

Nguyen, T. H.; and Grishman, R. 2015. Relation Extraction: Perspective from Convolutional Neural Networks. In *Proceedings of the Workshop on Vector Space Modeling for NLP (VSM) at NAACL-HLT 2015*.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved Denoising Diffusion Probabilistic Models. In *Proceedings of the 38th International Conference on Machine Learning*. virtual.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Phan, H.; Tran, L.; Tran, N. N.; Ho, N.; Phung, D.; and Le, T. 2022a. Improving Multi-task Learning via Seeking Task-based Flat Regions. *arXiv preprint arXiv:2211.13723*.

Phan, H.; Tuan, A. P.; Nguyen, S.; Linh, N. V.; and Than, K. 2022b. Reducing Catastrophic Forgetting in Neural Networks via Gaussian Mixture Approximation. In *Advances in Knowledge Discovery and Data Mining - Pacific-Asia Conference, PAKDD, Proceedings, Part I*. Chengdu, China.

Sener, O.; and Koltun, V. 2018. Multi-Task Learning as Multi-Objective Optimization. In *Advances in Neural Information Processing Systems*. Montréal, Canada.

Shin, H.; Lee, J. K.; Kim, J.; and Kim, J. 2017a. Continual Learning with Deep Generative Replay. In *Advances in Neural Information Processing Systems*. Long Beach, CA, USA.

Shin, H.; Lee, J. K.; Kim, J.; and Kim, J. 2017b. Continual Learning with Deep Generative Replay. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach, California, USA.

Smith, J. S.; Karlinsky, L.; Gutta, V.; Cascante-Bonilla, P.; Kim, D.; Arbelle, A.; Panda, R.; Feris, R.; and Kira, Z. 2023. CODA-Prompt: COntinual Decomposed Attention-based Prompting for Rehearsal-Free Continual Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Vancouver, BC, Canada.

Sohn, K.; Lee, H.; and Yan, X. 2015. Learning Structured Output Representation using Deep Conditional Generative Models. In *Advances in Neural Information Processing Systems*. Montreal, Quebec, Canada.

Van de Ven, G. M.; and Tolias, A. S. 2019. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*.

Veyseh, A. P. B.; Dernoncourt, F.; Dou, D.; and Nguyen, T. H. 2020a. Exploiting the Syntax-Model Consistency for Neural Relation Extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Veyseh, A. P. B.; Dernoncourt, F.; Dou, D.; and Nguyen, T. H. 2020b. A Joint Model for Definition Extraction with Syntactic Connection and Semantic Consistency. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)*.

Wang, H.; Xiong, W.; Yu, M.; Guo, X.; Chang, S.; and Wang, W. Y. 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)*. Minneapolis, Minnesota.

Wang, J.; and Lu, W. 2020. Two are Better than One: Joint Entity and Relation Extraction with Table-Sequence Encoders. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online.

Wang, P.; Song, Y.; Liu, T.; Lin, B.; Cao, Y.; Li, S.; and Sui, Z. 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*. Abu Dhabi, United Arab Emirates.

Wang, Z.; Zhang, Z.; Ebrahimi, S.; Sun, R.; Zhang, H.; Lee, C.-Y.; Ren, X.; Su, G.; Perot, V.; Dy, J.; et al. 2022b. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European Conference on Computer Vision*. Tel Aviv, Israel.

Wang, Z.; Zhang, Z.; Lee, C.-Y.; Zhang, H.; Sun, R.; Ren, X.; Su, G.; Perot, V.; Dy, J.; and Pfister, T. 2022c. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans, LA, USA.

Wu, T.; Li, X.; Li, Y.-F.; Haffari, G.; Qi, G.; Zhu, Y.; and Xu, G. 2021. Curriculum-meta learning for order-robust continual relation extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*. virtual.

Xia, H.; Wang, P.; Liu, T.; Lin, B.; Cao, Y.; and Sui, Z. 2023. Enhancing Continual Relation Extraction via Classifier Decomposition. In *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada.

Yu, T.; Kumar, S.; Gupta, A.; Levine, S.; Hausman, K.; and Finn, C. 2020. Gradient Surgery for Multi-Task Learning. In *Advances in Neural Information Processing Systems*. virtual.

Zhang, H.; Liang, B.; Yang, M.; Wang, H.; and Xu, R. 2022. Prompt-Based Prototypical Framework for Continual Relation Extraction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Zhang, Y.; Zhong, V.; Chen, D.; Angeli, G.; and Manning, C. D. 2017. Position-aware Attention and Supervised Data Improve Slot Filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*. Copenhagen, Denmark.

Zhao, K.; Xu, H.; Yang, J.; and Gao, K. 2022. Consistent Representation Learning for Continual Relation Extraction. In *Findings of the Association for Computational Linguistics: ACL 2022*. Dublin, Ireland.

Zhao, W.; Cui, Y.; and Hu, W. 2023. Improving Continual Relation Extraction by Distinguishing Analogous Semantics. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics)*. Toronto, Canada.