Mutual-pairing Data Augmentation for Fewshot Continual Relation Extraction

Nguyen Hoang Anh[1*], Quyen Tran[2*], Thanh Xuan Nguyen[1*], Diep Thi-Ngoc Nguyen[4], Linh Ngo Van[3†], Thien Huu Nguyen[5], Trung Le[6]

[1]Oraichain Labs Inc., US, [2]VinAI Research [3]Hanoi University of Science and Technology, [4]VNU University of Engineering and Technology, [5]University of Oregon, [6]Monash University

Abstract

Data scarcity is a major challenge in Few-shot Continual Relation Extraction (FCRE), where models must learn new relations from limited data while retaining past knowledge. Current methods, restricted by minimal data streams, struggle with catastrophic forgetting and overfitting. To overcome this, we introduce a novel *data augmentation strategy* that transforms single input sentences into complex texts by integrating both old and new data. Our approach sharpens model focus, enabling precise identification of word relationships based on specified relation types. By embedding adversarial training effects and leveraging new training perspectives through special objective functions, our method enhances model performance significantly. Additionally, we explore Sharpness-Aware Minimization (SAM) in Few-shot Continual Learning. Our extensive experiments uncover fascinating behaviors of SAM across tasks and offer valuable insights for future research in this dynamic field.

1 Introduction

Relation extraction (RE) is a problem in Information Extraction that seeks to extract semantic relationships between pairs of entities in a sentence. For example, given the sentence *"Kamala Harris and Donald Trump were political opponents in the US presidential election"*, the relation to be extracted between "Kamala Harris" and "Donald Trump" is "political opponents". In more in-depth research on real-world scenarios for RE, *Few-shot Continual Relation Extraction* (FCRE) (Qin and Joty, 2022c; Chen et al., 2023; Nguyen et al., 2025) is a challenging setting that has recently attracted a lot of attention. In this setting, the models need to continuously capture semantic information of new emerging relations from *a small and limited amount data*, while avoiding forgetting knowledge of previously learned ones. Therefore, the two main concerns in this scenario involve dealing with *catastrophic forgetting* (Thrun and Mitchell, 1995; Le et al., 2024a; Hai et al., 2024; Van et al., 2022; Phan et al., 2022; Le et al., 2025) and *overfitting* of FCRE models.

Recent work (Wang et al., 2023b; Qin and Joty, 2022c; Chen et al., 2023) tackles these problems by leveraging memory-based approaches inspired by conventional Continual learning methods (Nguyen et al., 2023; Le et al., 2024b; Dao et al., 2024; Le et al., 2024c). These methods intentionally save a small amount of data samples from old tasks and propose various strategies to enhance the models' abilities to distinguish relation representations. However, a significant issue is that previous models are often fine-tuned with only one sample from each old class and a few samples from new classes, which leads to forgetting and overfitting in the challenging FCRE scenario.

To address these challenges, this work explores a new approach to generate more diverse and meaningful learning samples for FCRE. By leveraging strong guidance, our method relies solely on the available datasets, avoiding the need for external resources. Specifically, we introduce *a data augmentation method* that combines both old and new data to create new training samples, transforming *single input sentences into complex texts* and weaving together knowledge from all tasks so far. Starting with an original sentence containing a pair of entities, we append an arbitrary sentence, either before or after it. This augmentation does not change the relation between the two entities in each original sentence. Thus, our approach not only expands the training data but also guides models to focus on identifying word relationships based on the specified relation type, rather than on grammatical or other semantic associations. More specifically, our method encourages the attention mechanism to pri-

*Equally contributed.
†Corresponding author: linhnv@soict.hust.edu.vn

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 4057–4075
April 29 - May 4, 2025

oritize words in the original sentence related to the entity pair, while minimizing attention to the added noisy sentences. Especially, we propose special objective functions to inherently deliver adversarial training effects and novel perspectives during training, significantly boosting model performance—particularly in preventing forgetting and reducing overfitting.

Furthermore, in the effort to explore solutions for the overfitting problem in Few-shot Continual learning models, we conducted extensive experiments on the application of Sharpness-Aware Minimization (SAM) (Foret et al., 2020). Specifically, we assessed the effectiveness of SAM in improving the latest state-of-the-art methods for both Few-Shot Continual Relation Extraction (FCRE) and Few-Shot Continual Event Detection (FCED) (Zhang et al., 2024; Cao et al., 2020; Yu et al., 2021a; Liu et al., 2022). The results yielded surprising and intriguing insights, showing that SAM is not always a suitable solution for few-shot continual scenarios, which we believe will be valuable for future research.

In summary, the key contributions of this work are as follows:

- We introduce a novel data augmentation approach to enrich the limited datasets in the Few-Shot Continual Relation Extraction (FCRE) scenario. This strategy not only enables the model to benefit from adversarial training but also provides diverse perspectives during training, thereby significantly enhancing its ability to mitigate overfitting, prevent forgetting, and ultimately improve overall performance.

- Through extensive experiments on applying Sharpness-Aware Minimization (SAM) across various tasks and datasets within the Few-Shot Continual Learning domain, we offer valuable insights for the community.

2 Background

2.1 Problem formulation

In the setting of Few-Shot Continual Relation Extraction (FCRE) (see more details in A), we consider a model that incrementally acquires knowledge through a sequence of tasks. For each task $\mathcal{T}^t$, the model is trained on dataset $D^t = (x_i^t, y_i^t)_{i=1}^{N\times K}$, which includes N new relations of relation set R^t, each consisting of K data samples. Particularly, each instance (x_i^t, y_i^t) consists of a sentence x_i containing a pair of entities (e_h, e_t) and its corresponding relation label $y_i \in R^t$. This paradigm is commonly called *"N-way-K-shot"* learning.

Upon completion of task $\mathcal{T}^t$, the dataset D^t becomes unavailable for subsequent learning phases. Then the model performance is evaluated on the testing dataset of all tasks so far, to identify relations in the expanded set $\tilde{R}^t = \bigcup_{i=1}^{t} R^i$.

In alignment with prior research (Han et al., 2020a; Qin and Joty, 2022a; Wang et al., 2023b), we employ a memory buffer $M = \{M^1, M^2, \ldots, M^t\}$, which retains m representative samples for each relation from all previous tasks. During the training of $\mathcal{T}^t$, the model can access the memory $\widehat{M}^{t-1} = \bigcup_{i=1}^{t-1} M^i$. For conciseness, we denote $\mathcal{D}^t = D^t \cup \widehat{M}^{t-1}$ as the complete training data used in t. In the few-shot setting, we store only a single sample per relation ($m = 1$).

2.2 The Base Methods

In this work, we build upon "Making Pre-trained Language Models Better Continual Few-Shot Relation Extractors (CPL)" (Ma et al., 2024), a recent state-of-the-art approach of FCRE problems, as the base of our method. Please refer to Appendix B.1 for more details. Additionally, we extend our experiments to the Few-shot Continual Event Detection (FCED) task to rigorously evaluate the effectiveness of our proposed method. Details regarding FCED are provided in Appendix B.2.

3 Proposed Method

In this section, we introduce our novel data augmentation technique and two special objective functions in Section 3.1. We then discuss interesting findings related to applying Sharpness Aware Minimization (SAM) - a method for enhancing model generalizability, in the challenging scenario of FCRE, where models are prone to overfitting (Section 3.2).

3.1 Mitigating Data Scarcity and Consolidating Feature Extraction

FCRE is inherently challenged by data scarcity due to the limited number of labeled samples for each relation. This restriction impedes model performance, particularly in extracting nuanced relationships between entities. The problem is compounded in Continual learning settings, where the model must continually adapt to emerging relations while retaining previous knowledge. To address

this problem, we propose a dual strategy, including (I) a *Mutual-pairing augmentation* technique to enrich the provided training dataset, and (II) accompanying *advanced objective functions* to optimize this data usage. This approach aims to mitigate the negative effects of data scarcity and enhance representation in latent space, thereby improving overall FCRE performance.

3.1.1 Mutual-Pairing Augmentation.

When training task t, given the training set $\mathcal{D}^t$, we randomly select data instances and then pair them to get at most $L = C^2_{|\mathcal{D}^t|} = \frac{|\mathcal{D}^t|(|\mathcal{D}^t|+1)}{2}$ pairs of data samples: $\{[(x_i, y_i), (x_j, y_j)]_k\}_{k=1}^{L}$, where $i, j \in I$, for simplicity, we ignore the superscript of taskID. Combining the original samples within each pair, we obtain the corresponding augmented dataset:

$$E^t = \{(x_k, [y_k^1; y_k^2])\}_{k=1}^{L} = \{(x_k, [y_k^1; y_k^2])\}_{k \in I^E} \tag{1}$$

where $x_k := \textit{concat}(x_i, x_j)$; y_k^1 and y_k^2 stand for y_i and y_j, respectively; $I^E = \{1, ...L\}$ is the index set of augmented samples. For convenience, we refer to *"the original single-label space"* as the space corresponding to the labels of the original dataset, where each data sample has a unique label. In addition, we refer to *"the dual-label space"* as the space containing pairs of labels corresponding to each augmented data sample.

In this way, without any external resources or support, we can significantly increase the volume of the training set by $L = C^2_{|\mathcal{D}^t|}$, with diverse "dual-label samples". Furthermore, during the training process, the attention mechanism in Language models' architecture allows [MASK] embeddings of the original samples in the respective augmented ones to interact mutually. Since the input sentences in the dataset are often unrelated or only weakly related, pairing them will make these embeddings to be perturbed to some extent. This is similar to the perturbation strategy in Adversarial training (Goodfellow et al., 2015; Zhang et al., 2019; Wong et al., 2020), where the use of challenging samples during training makes the model more robust. This also implies that increasing the number of samples in our strategy does not merely involve duplicating existing data, but rather makes the data more diverse and nuanced, thereby effectively reducing overfit for models.

3.1.2 Advanced accompanying loss function

To fully leverage the new augmented data, we propose advanced objective functions, which enable models to have multiple perspectives on the data during training

thereby improving generalization and reducing the risk of overfitting.

Margin Loss for Original Single-label Space Building upon the augmented dataset, we first employ this loss function to ensure the behavior of models w.r.t the original label space. In particular, while the relation pair in each augmented sample mutually interacts to create the permuted embeddings, this loss function helps models recognize the relations in individual sentences and separate their respective permuted representations. Given an augmented sample $(x_k, [y_k^1; y_k^2])$, let r_k^1 and r_k^2 be the representations corresponding to the pair of relations, the loss function can be formulated as follows:

$$\mathcal{L}_{CR} = \frac{\sum_{k \in I^E} \mathbb{I}_{y_k^1 \neq y_k^2} \max(0, m - sim(r_k^1, r_k^2))}{\sum_{k \in I^E} \mathbb{I}_{y_k^1 \neq y_k^2}} \tag{2}$$

where m is a hyperparameter of margin and $sim(\cdot, \cdot)$ returns similarity between 2 vectors, $\mathbb{I}_{y_k^1 \neq y_k^2} = 1$ if $y_k^1 \neq y_k^2$ else 0.

By seamlessly integrating this loss function with our data augmentation strategy, our method guides models to identifying word relationships based on the specified relation type, steering clear of distractions from grammatical or semantic nuances. It actively encourages the attention mechanism to prioritize words connected to the entity pair in the original sentence, while effectively filtering out added noise.

Contrastive Loss for Dual-label Space. To further take advantage of our pairing data strategy, this loss function encourages models to exploit a new perspective on the data, via the constraint between augmented samples. Intuitively, the representation vectors of paired samples with the same original set of labels are expected to be close together, while the pairs with at least one different component label are pushed apart.

Particularly, let $\bar{r}_k = \frac{1}{2}(r_k^1 + r_k^2)$ be the representative embedding of k^{th} augmented sample. We define the index set of positive samples w.r.t this augmented sample as $P_k^E = \{p \in I^E : \{y_k^1, y_k^2\} = \{y_p^1, y_p^2\}\}$, and the corresponding negative set as

$N_k^E = I^E \setminus P_k^E$. Then the loss function can be formulated as follows:

$$\mathcal{L}_{DL} = -\sum_{k \in I^E} \frac{1}{|P_k^E|} \sum_{l \in P_k^E} \log \frac{u(\bar{r}_k, \bar{r}_l)}{\sum_{l' \in N_k^E} u(\bar{r}_k, \bar{r}_{l'})} \tag{3}$$

where $u(\bar{r}_k, \bar{r}_l) = \exp(sim(\bar{r}_k, \bar{r}_l)/\tau')$, and τ' is a temperature parameter. In this way, models have an opportunity to consider the representations of augmented samples corresponding to "the dual-label space" (i.e., space containing $[y_k^1, y_k^2], k \in I^E$) in a new point of view, thereby further consolidating the original representations of each relation in the original latent space as well as reducing variance caused by conventional training on the limited training data, and finally improving the generalization ability of models.

Finally, the objective function used for training can be summarized as:

$$\mathcal{L}_{=} \mathcal{L}_0 + \beta_1 \mathcal{L}_{CR} + \beta_2 \mathcal{L}_{DL} \tag{4}$$

where $\mathcal{L}_0$ is the base objective function of the original method (i.e., $\mathcal{L}_{MCL}$ of CPL), β_1 and β_2 are hyperparameters that control the respective contributions of our proposed loss functions. It's worth noting that our method not only improves CPL but can also be flexibly integrated into any existing FCRE methods to enhance model performance.

3.2 Study about the application of Sharpness Aware Minimization in FCRE problem

Sharpness Aware Minimization (SAM) (Foret et al., 2020) is known as an effective solution for improving generalization and reducing overfitting in deep learning models. However, there are no studies on SAM for Few-shot Continual learning scenarios in the problem of Information Extraction in general, as well as FCRE in particular. Therefore, in this section, we present our key findings, aiming to provide useful insights for future work.

Failure of SAM in Few-shot Continual Learning scenarios. To begin with, we conducted experiments applying SAM to the training process of each task t of the latest FCRE and FCED methods, as described by the following formula:

$$\min_{\phi} \max_{\|\delta\| \leq \sigma} \mathcal{L}(\widehat{M}^{t-1} \cup D^t; \phi + \delta) \tag{5}$$

where ϕ denotes the model parameters, $\mathcal{L}$ represents the objective function, and $\sigma > 0$ is the perturbation threshold.

The results in Table 2 indicate that, for FCRE, while SAM improves efficiency in some cases of CPL when using BERT as the backbone, it significantly reduces performance by more than 6% when LLM2Vec is used as the backbone. Furthermore, with CPL+MI, the current strongest method, SAM not only fails to enhance performance but also leads to 2% drop in final accuracy on TACRED. For FCED, the results mostly show that SAM, negatively impacts model performance.

Applying SAM for only current task training data? Based on the above surprising experimental results, we thoroughly conducted extra experiments to verify whether the scarcity of FCRE data caused SAM not to perform as expected. Specifically, we supposed that a sample per learned relation may sometimes lack enough information to determine a good update direction for SAM, and might even return a bad guidance that potentially diminishes the model's generalization capability. Therefore, when training task t, instead of applying SAM to the entire $\mathcal{D}^t$, we only apply it to the current task's data D^t and ignore the data from the memory buffer $\widehat{M}^{t-1}$. Particularly, we conduct experiments following the optimization process as below:

$$\min_{\phi} \mathcal{L}(\widehat{M}^{t-1}; \phi) + \max_{\|\delta\| \leq \sigma} \mathcal{L}(D^t; \phi + \delta) \tag{6}$$

Consequently, the experimental results in Figure 1 and 2 demonstrate that excluding SAM from the memory buffer consistently enhances model performance when using SAM. Therefore, it can be said that data scarcity in Few-shot Continual learning scenario is the main weakness of SAM, and applying SAM only to the current task, due to the large and representative enough data, could be a temporary but intriguing solution to preserve the positive impact of SAM.

How our data augmentation strategy can help ensure the effectiveness of SAM? Motivated by the results above, we find that if the weakness of SAM in FCRE is the lack of data, our method can help when significantly increasing the amount of training data. Suppose at the time of training task t, the memory buffer contains n_m data samples of the corresponding n_m learned relations, and the current task has n_c data samples of new relations. By cross-matching all available data, we obtain the following results:

FewRel *(10-way 5-shot)*

Method	$\mathcal{T}^1$	$\mathcal{T}^2$	$\mathcal{T}^3$	$\mathcal{T}^4$	$\mathcal{T}^5$	$\mathcal{T}^6$	$\mathcal{T}^7$	$\mathcal{T}^8$	$\Delta\downarrow$
ERDA	92.43	64.52	50.31	44.92	39.75	36.36	34.34	31.96	60.47
CRECL	93.93	82.55	74.13	69.33	66.51	64.60	62.97	59.99	33.94
ConPL**	**95.18**	79.63	74.54	71.27	68.35	63.86	64.74	62.46	32.72
SCKD	94.77	82.83	76.21	72.19	70.61	67.15	64.86	62.98	31.79
SCKD+MI	94.75	83.88	76.71	72.34	70.78	67.36	65.08	63.95	30.80
CPL	94.87	85.14	78.80	75.10	72.57	69.57	66.85	64.50	30.37
CPL+MI	94.69	**85.58**	80.12	75.71	73.90	70.72	68.42	66.27	28.42
SCKD+augment	94.67	84.04	77.57	74.23	71.79	68.13	65.36	63.83	30.84
SCKD+MI+augment	95.10	84.88	78.22	74.04	71.81	68.72	66.36	64.35	30.75
CPL+augment	95.02	84.88	**80.85**	76.39	75.20	72.09	69.86	67.82	27.20
CPL+MI+augment	94.76	85.48	80.24	**77.69**	**75.6**	**72.94**	**70.74**	**68.36**	**26.40**

TACRED *(5-way 5-shot)*

Method	$\mathcal{T}^1$	$\mathcal{T}^2$	$\mathcal{T}^3$	$\mathcal{T}^4$	$\mathcal{T}^5$	$\mathcal{T}^6$	$\mathcal{T}^7$	$\mathcal{T}^8$	$\Delta\downarrow$
ERDA	81.88	53.68	40.36	36.17	30.14	22.61	22.29	19.42	62.46
CRECL	87.09	78.09	61.93	55.60	53.42	51.91	47.55	45.53	41.56
ConPL**	**88.77**	69.64	57.50	52.15	58.19	55.01	52.88	50.97	37.80
SCKD	88.42	79.35	70.61	66.78	60.47	58.05	54.41	52.11	36.31
SCKD+MI	87.55	79.39	70.7	66.68	61.94	59.81	55.1	53.63	33.92
CPL	86.27	81.55	73.52	68.96	63.96	62.66	59.96	57.39	28.88
CPL+MI	85.67	**82.54**	75.12	70.65	66.79	65.17	61.25	59.48	26.19
SCKD+augment	88.10	81.70	71.79	66.60	61.10	59.97	55.81	54.53	33.57
SCKD+MI+augment	87.74	80.16	72.46	68.67	62.89	61.28	58.03	54.67	33.07
CPL+augment	86.68	81.99	75.27	70.41	66.30	**65.71**	62.16	60.26	26.42
CPL+MI+augment	86.33	82.31	**76.35**	**70.93**	**68.28**	65.04	**62.60**	**61.97**	**24.36**

Table 1: Accuracy (%) of different BERT-based methods after training for each task on TACRED and FewRel in *5-shot* settings. We highlight the rows corresponding to our method. The best result in each group is in **bold**, and the corresponding runner-up is underlined. **Results of ConPL are reproduced

FCRE	FewRel		TACRED	
	CPL	+ SAM	CPL	+ SAM
BERT	64.50	**67.80**	57.39	**60.75**
LLM2Vec	**69.49**	68.54	**71.35**	65.16
BERT (+MI)	66.27	**67.01**	**59.48**	57.26

FCED	ACE		MAVEN	
	HANet	+ SAM	HANet	+ SAM
2way-5shot	**57.85**	57.18	**53.62**	52.86
2way-10shot	**61.02**	60.20	56.13	**56.46**

Table 2: Performance comparison of existing FCRE and FCED SOTA methods and those when using SAM (i.e., +SAM).

- For the memory buffer, cross-matching between old samples yields $C^2_{n_m}$ additional samples. Besides, cross-matching with the current task samples adds $n_m \times n_c$ more samples, representing the old task's information. Compared to the existing method (CPL), the data representing the old task in our strategy has increased by $\frac{n_m+1}{2} + n_c$ times.
- Similarly, for the current task, the amount of data representing this task increases by $\frac{n_c+1}{2} + n_m$ times.

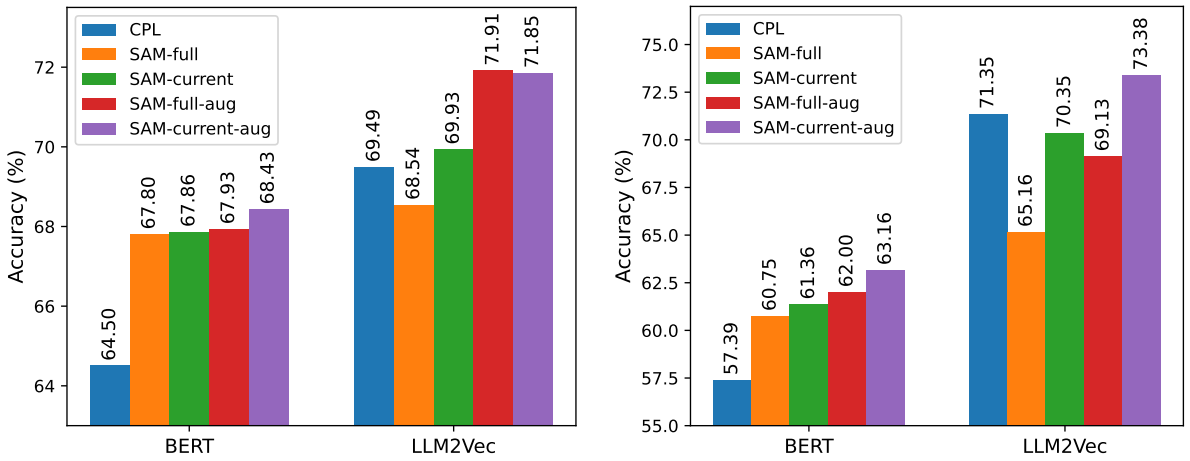
The results in Figure 1 and 2 also show that, when combined with our augmentation strategy, SAM effectively helps the model improve significantly, both when applied to the current task data and to the entire training dataset, including data from the poor memory buffer.

4 Experiments

In this section, we present experimental results demonstrating the effectiveness of our proposed data augmentation strategy.

4.1 Experimental Setup

In our main experiments, we use current state-of-the-art FCRE methods as baselines, includ-



(a) FewRel (b) TACRED

Figure 1: Applications of SAM in FCRE models. *"SAM-full"* indicates the case we apply SAM on all data $\mathcal{D}^t$ from the current task and memory buffer, *"SAM-current"* is when we apply SAM only on the data of the current task.

ing: ERDA (Qin and Joty, 2022b), CRECL (Hu et al., 2022), SCKD (Wang et al., 2023b), ConPL (Chen et al., 2023), CPL (Ma et al., 2024) and CPL+MI (Tran et al., 2024). Besides, the models are evaluated using pre-trained models consisting of BERT (Devlin et al., 2018), and LLM2Vec (BehnamGhader et al., 2024), on two benchmark datasets: FewRel (Han et al., 2018) and TACRED (Zhang et al., 2017). Besides, to further demonstrate the efficiency and flexibility of our proposed method, we conducted additional experiments on the setting of FCED in two datasets: MAVEN (Wang et al., 2020) and ACE (Walker et al., 2005).

We note that we have reproduced the results of ConPL (Chen et al., 2023) under the same setting as SCKD and CPL. The reason is that the evaluation strategy in this paper is impractical for continual learning scenarios. Please refer to Appendix C for more details.

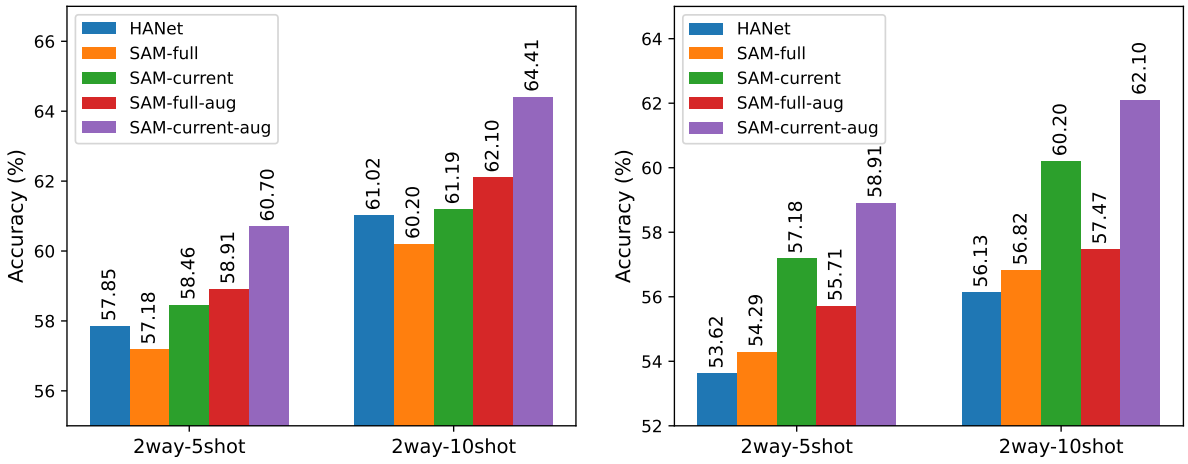
4.2 Main Results

Performance comparison Table 1 compares our method and FCRE baselines, on TACRED and FewRel datasets. Our approach, which integrates data augmentation and special objective functions, demonstrates consistent improvements regarding both final average accuracy $\mathcal{A}_8$ and forgetting rate $\Delta = \mathcal{A}_1 - \mathcal{A}_8$, which is the discrepancy between accuracy after learning the first task ($\mathcal{A}_1$) and after learning the final one ($\mathcal{A}_8$).

On the FewRel dataset, our method achieves a notable performance gain of up to 2.09% compared to the strongest baseline. Similarly, on the TACRED dataset, we observe an even more substantial improvement, with a performance gap of 2.49%. These results underscore the effectiveness of our approach in mitigating catastrophic forgetting, as evidenced by the significantly reduced accuracy drops. Specifically, our model outperforms CPL with accuracy drops of only 26.40% on FewRel and 24.36% on TACRED, indicating enhanced retention of knowledge from previous tasks.

Ablation study To gain deeper insights into the contributions of our proposed components, we conduct an ablation study using CPL as the baseline on both TACRED and FewRel datasets. Table 3 illustrates the impact of the Margin Loss for Original Single-label Space ($\mathcal{L}_{CR}$) and the Contrastive Loss for Dual-label Space ($\mathcal{L}_{DL}$) on model performance.

The results demonstrate that both loss functions play crucial roles in the overall performance of our method. Removing $\mathcal{L}_{CR}$ leads to a noticeable decrease in accuracy across most tasks, particularly in the later stages of continual learning. This suggests that $\mathcal{L}_{CR}$ is essential for maintaining model behavior with respect to the original label space. Similarly, the absence of $\mathcal{L}_{DL}$ results in performance degradation, especially in the middle and later tasks. This observation highlights the importance of exploiting the new perspective on data provided by our pairing strategy, which encourages the model to learn more robust representations.



(a) ACE (b) MAVEN

Figure 2: Applications of SAM in FCED models. *"SAM-full"* indicates the case we apply SAM on all data $\mathcal{D}^t$ from the current task and memory buffer, *"SAM-current"* is when we apply SAM only on the data of the current task.

TACRED *(5-way 5-shot)*

Method	$\mathcal{T}^1$	$\mathcal{T}^2$	$\mathcal{T}^3$	$\mathcal{T}^4$	$\mathcal{T}^5$	$\mathcal{T}^6$	$\mathcal{T}^7$	$\mathcal{T}^8$
CPL + augment	**95.02**	84.88	**80.85**	76.39	**75.2**	**72.09**	69.86	**67.82**
w.o $\mathcal{L}_{CR}$	94.20	84.22	79.83	**76.67**	73.86	71.59	**70.88**	66.49
w.o $\mathcal{L}_{DL}$	94.56	**86.02**	77.4	75.16	75.16	70.0	69.66	63.66

FewRel *(10-way 5-shot)*

Method	$\mathcal{T}^1$	$\mathcal{T}^2$	$\mathcal{T}^3$	$\mathcal{T}^4$	$\mathcal{T}^5$	$\mathcal{T}^6$	$\mathcal{T}^7$	$\mathcal{T}^8$
CPL + augment	**86.68**	**81.99**	75.27	70.41	**66.30**	**65.71**	**62.16**	**60.26**
w.o $\mathcal{L}_{CR}$	86.46	81.37	75.59	69.96	64.61	65.53	60.13	57.36
w.o $\mathcal{L}_{DL}$	85.64	81.62	**75.88**	**71.72**	66.06	64.82	62.10	59.59

Table 3: Ablation study - Our special objective function

4.3 Improved SAM performance in Few-shot Continual Learning

Table 4 presents the results when applying our data augmentation method to ensure the effectiveness of Sharpness Aware Minimization (SAM), across various baselines and datasets. For each setting, we compare the performance of the original baseline models, models with SAM applied only to the current task's data (SAM$_{current}$), and our proposed method combining SAM$_{current}$ with data augmentation.

On the TACRED dataset (5-way 5-shot), our method combined with SAM, consistently outperforms both the original baselines and the corresponding SAM-based versions (+SAM$_{current}$) across all tasks and model configurations (CPL$_{BERT}$, CPL$_{BERT}$+MI, CPL$_{LLM2Vec}$). Notably, compared with CPL$_{BERT}$+MI, our approach can achieve the accuracy of 65.11% after the final task, outperforming the baseline and its SAM-based version by a gap up to 10%.

On the FewRel dataset (10-way 5-shot), our method also demonstrates the improvements from the baselines' versions, especially in later tasks. For CPL$_{BERT}$, we achieve the final accuracy of 68.43%, outperforming both the baseline (64.50%) and its SAM-based (66.84%) versions by the gap up to 4%.

In addition, extended experiments in the setting of FCED, on ACE (2-way 5-shot) and MAVEN (2-way 10-shot) datasets, further confirm the effectiveness of our approach in ensuring the positive effect of SAM in the challenging setting of Fewshot Continual Learning (FCL).

These comprehensive results across multiple tasks and datasets confirm that our data augmentation strategy effectively mitigates the primary limitation of SAM in FCL scenarios, as outlined in 3.2.

TACRED *(5-way 5-shot)*

Method	$\mathcal{T}^1$	$\mathcal{T}^2$	$\mathcal{T}^3$	$\mathcal{T}^4$	$\mathcal{T}^5$	$\mathcal{T}^6$	$\mathcal{T}^7$	$\mathcal{T}^8$
CPL$_{BERT}$	86.27	81.55	73.52	68.96	63.96	62.66	59.96	57.39
CPL$_{BERT}$+SAM$_{current}$	86.33	81.84	74.53	71.9	65.78	66.87	64.47	61.36
CPL$_{BERT}$+SAM$_{current}$+augment	**86.58**	**82.05**	**75.35**	**73.31**	**69.20**	**68.55**	**64.68**	**62.62**
CPL$_{BERT}$+MI	85.67	82.54	75.12	70.65	66.79	65.17	61.25	59.48
CPL$_{BERT}$+MI+SAM$_{current}$	**86.77**	81.48	73.13	70.12	64.02	62.13	57.71	55.00
CPL$_{BERT}$+MI+SAM$_{current}$+augment	86.55	**82.75**	**76.16**	**73.93**	**71.14**	**70.37**	**67.36**	**65.11**
CPL$_{LLM2Vec}$	**89.12**	**82.93**	78.24	75.20	74.37	74.23	71.55	71.35
CPL$_{LLM2Vec}$+SAM$_{current}$	88.74	80.87	79.57	75.97	75.98	74.62	73.08	70.35
CPL$_{LLM2Vec}$+SAM$_{current}$+augment	88.45	82.85	**80.32**	**76.79**	**77.13**	**77.23**	**74.16**	**73.38**

FewRel *(10-way 5-shot)*

Method	$\mathcal{T}^1$	$\mathcal{T}^2$	$\mathcal{T}^3$	$\mathcal{T}^4$	$\mathcal{T}^5$	$\mathcal{T}^6$	$\mathcal{T}^7$	$\mathcal{T}^8$
CPL$_{BERT}$	**94.87**	**85.14**	78.80	75.10	72.57	69.57	66.85	64.50
CPL$_{BERT}$+SAM$_{current}$	94.68	85.06	80.55	76.99	75.11	71.94	69.37	66.84
CPL$_{BERT}$+SAM$_{current}$+augment	94.48	84.34	**80.80**	**77.29**	**75.75**	**73.07**	**70.45**	**68.43**
CPL$_{BERT}$+MI	**94.69**	85.58	80.12	75.71	73.90	70.72	68.42	66.27
CPL$_{BERT}$+MI+SAM$_{current}$	94.53	84.97	80.12	76.38	74.48	71.62	69.57	67.41
CPL$_{BERT}$+MI+SAM$_{current}$+augment	94.59	**85.62**	**80.79**	**77.40**	**75.73**	**72.61**	**70.04**	**68.57**
CPL$_{LLM2Vec}$	**96.38**	87.22	82.67	79.20	77.00	74.63	72.22	69.49
CPL$_{LLM2Vec}$+SAM$_{current}$	96.12	87.65	81.33	78.56	77.21	74.04	71.16	69.93
CPL$_{LLM2Vec}$+SAM$_{current}$+augment	95.98	**88.11**	**83.88**	**80.34**	**78.54**	**75.60**	**73.61**	**71.85**

ACE *(2-way 5-shot)*

Method	$\mathcal{T}^1$	$\mathcal{T}^2$	$\mathcal{T}^3$	$\mathcal{T}^4$	$\mathcal{T}^5$
HANet$_{BERT}$	61.16	63.07	**57.50**	53.21	54.31
HANet$_{BERT}$+SAM$_{current}$	61.34	64.71	55.79	54.53	55.97
HANet$_{BERT}$+SAM$_{current}$+augment	**63.51**	**64.99**	56.46	**61.45**	**57.11**

MAVEN *(2-way 10-shot)*

Method	$\mathcal{T}^1$	$\mathcal{T}^2$	$\mathcal{T}^3$	$\mathcal{T}^4$	$\mathcal{T}^5$
HANet$_{BERT}$**	57.64	53.28	58.67	56.23	54.85
HANet$_{BERT}$+SAM$_{current}$	57.18	54.33	58.90	56.77	**56.91**
HANet$_{BERT}$+SAM$_{current}$+augment	**58.31**	**56.59**	**58.98**	**56.93**	56.53

Table 4: Effectiveness of Our Method in Improving SAM Performance Across Various Datasets and Baselines. **Results for MAVEN dataset under *2-way 10-shot* settings were reproduced using the provided source code.

By significantly increasing the amount of training data through cross-matching, SAM is better able to identify flat regions where the learned model demonstrates improved performance and generalization across all tested configurations.

5 Conclusion

This work addresses the challenges of catastrophic forgetting and overfitting in few-shot continual relation extraction under conditions of limited data availability. First, we introduce a novel data augmentation technique that generates additional training samples by combining both old and new data, transforming them into more complex textual structures. Our method not only expands the training data but also enhances the model's ability to capture word relationships based on the specified relation type, rather than relying solely on grammatical or other general semantic associations. Furthermore, we propose specialized objective functions designed to inherently induce adversarial training effects and increase discriminative representation among relation types. Finally, this work represents the first step toward investigating the application of Sharpness-Aware Minimization (SAM) in few-shot continual information extraction. The findings pro-

vide valuable insights for the research community.

Limitations

Our data augmentation method, while effective, may introduce training imbalances in Few-shot Continual Learning scenarios. The augmentation process amplifies the existing disparity between current and previous task data. Consequently, with limited batch sizes, many minibatches may contain samples from only a few specific labels, primarily from the current task. This imbalance could potentially impact the model's ability to maintain equal representation across all tasks, affecting its performance in mitigating catastrophic forgetting and potentially biasing predictions towards the current task. Future work could explore balanced augmentation techniques or adaptive sampling methods to address this limitation.

Acknowledgements

This research was partially supported by NSF Grant #2239570. This research is also supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract 2022-22072200003. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government.

Trung Le was supported by ARC DP23 grant DP230101176 and by the Air Force Office of Scientific Research under award number FA2386-23-1-4044.

References

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2Vec: Large Language Models Are Secretly Powerful Text Encoders. *arXiv preprint arXiv:2404.05961*.

Debora Caldarola, Barbara Caputo, and Marco Ciccone. 2022. Improving generalization in federated learning by seeking flat minima. In *European Conference on Computer Vision*, pages 654–672. Springer.

Pengfei Cao, Yubo Chen, Jun Zhao, and Taifeng Wang. 2020. Incremental event detection via knowledge consolidation networks.

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. 2021. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34:22405–22418.

Xiudi Chen, Hui Wu, and Xiaodong Shi. 2023. Consistent prototype learning for few-shot continual relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 7409–7422. Association for Computational Linguistics.

Viet Dao, Van-Cuong Pham, Quyen Tran, Thanh-Thien Le, Linh Ngo, and Thien Nguyen. 2024. Lifelong event detection via optimal transport. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12610–12621.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Gintare Karolina Dziugaite and Daniel M. Roy. 2017. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *UAI*. AUAI Press.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. 2020. Sharpness-aware minimization for efficiently improving generalization.

Robert M. French. 1993. Catastrophic interference in connectionist networks: Can it be predicted, can it be prevented? In *Advances in Neural Information Processing Systems 6, [7th NIPS Conference, Denver, Colorado, USA, 1993]*, pages 1176–1177. Morgan Kaufmann.

Robert M. French and Nick Chater. 2002. Using noise to compute error surfaces in connectionist networks: A novel means of reducing catastrophic forgetting. *Neural Comput.*, 14(7):1755–1769.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*.

Nam Le Hai, Trang Nguyen, Linh Ngo Van, Thien Huu Nguyen, and Khoat Than. 2024. Continual variational dropout: a view of auxiliary local variables in continual learning. *Machine Learning*, 113(1):281–323.

Xu Han, Yi Dai, Tianyu Gao, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2020a. Continual relation learning via episodic memory activation and reconsolidation.

Xu Han, Yi Dai, Tianyu Gao, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2020b. Continual relation learning via episodic memory activation and reconsolidation. In *Proceedings of the*

58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pages 6429–6440. Association for Computational Linguistics.

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

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

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. 2020. Fantastic generalization measures and where to find them. In *ICLR*. OpenReview.net.

Minh Le, Tien Ngoc Luu, An Nguyen The, Thanh-Thien Le, Trang Nguyen, Tung Thanh Nguyen, Linh Ngo Van, and Thien Huu Nguyen. 2025. Adaptive prompting for continual relation extraction: A within-task variance perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Minh Le, An Nguyen, Huy Nguyen, Trang Nguyen, Trang Pham, Linh Van Ngo, and Nhat Ho. 2024a. Mixture of experts meets prompt-based continual learning. In *Advances in Neural Information Processing Systems*.

Thanh-Thien Le, Viet Dao, Linh Nguyen, Thi-Nhung Nguyen, Linh Ngo, and Thien Nguyen. 2024b. Sharpseq: Empowering continual event detection through sharpness-aware sequential-task learning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3632–3644.

Thanh-Thien Le, Manh Nguyen, Tung Thanh Nguyen, Linh Ngo Van, and Thien Huu Nguyen. 2024c. Continual relation extraction via sequential multi-task learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18444–18452.

Minqian Liu, Shiyu Chang, and Lifu Huang. 2022. Incremental prompting: Episodic memory prompt for lifelong event detection.

Shengkun Ma, Jiale Han, Yi Liang, and Bo Cheng. 2024. Making pre-trained language models better continual few-shot relation extractors. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 10970–10983. ELRA and ICCL.

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

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks.

Xuan Thanh Nguyen, Duc Le Anh, Tran Quyen, Le Thanh-Thien, Linh Ngo Van, and Thien Huu Nguyen. 2025. Few-shot, no problem: Descriptive continual relation extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Henning Petzka, Michael Kamp, Linara Adilova, Cristian Sminchisescu, and Mario Boley. 2021. Relative flatness and generalization. In *NeurIPS*, pages 18420–18432.

Hoang Phan, Anh Phan Tuan, Son Nguyen, Ngo Van Linh, and Khoat Than. 2022. Reducing catastrophic forgetting in neural networks via gaussian mixture approximation. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 106–117. Springer.

Chengwei Qin and Shafiq Joty. 2022a. Continual few-shot relation learning via embedding space regularization and data augmentation.

Chengwei Qin and Shafiq Joty. 2022b. Continual few-shot relation learning via embedding space regularization and data augmentation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2776–2789, Dublin, Ireland. Association for Computational Linguistics.

Chengwei Qin and Shafiq R. Joty. 2022c. Continual few-shot relation learning via embedding space regularization and data augmentation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2776–2789. Association for Computational Linguistics.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. 2017. icarl: Incremental classifier and representation learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5533–5542. IEEE Computer Society.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. 2017. Continual learning with deep generative replay. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 2990–2999.

Sebastian Thrun and Tom M. Mitchell. 1995. Lifelong robot learning. *Robotics and Autonomous Systems*,

15(1):25–46. The Biology and Technology of Intelligent Autonomous Agents.

Quyen Tran, Nguyen Xuan Thanh, Nguyen Hoang Anh, Nam Le Hai, Trung Le, Linh Van Ngo, and Thien Huu Nguyen. 2024. Preserving generalization of language models in few-shot continual relation extraction.

Linh Ngo Van, Nam Le Hai, Hoang Pham, and Khoat Than. 2022. Auxiliary local variables for improving regularization/prior approach in continual learning. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 16–28. Springer.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2005. Ace 2005 multilingual training corpus.

Hong Wang, Wenhan Xiong, Mo Yu, Xiaoxiao Guo, Shiyu Chang, and William Yang Wang. 2019. Sentence embedding alignment for lifelong relation extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 796–806. Association for Computational Linguistics.

Pengfei Wang, Zhaoxiang Zhang, Zhen Lei, and Lei Zhang. 2023a. Sharpness-aware gradient matching for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3769–3778.

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. Maven: A massive general domain event detection dataset.

Xinyi Wang, Zitao Wang, and Wei Hu. 2023b. Serial contrastive knowledge distillation for continual few-shot relation extraction. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 12693–12706. Association for Computational Linguistics.

Eric Wong, Leslie Rice, and J Zico Kolter. 2020. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations (ICLR)*.

Pengfei Yu, Heng Ji, and Prem Natarajan. 2021a. Lifelong event detection with knowledge transfer.

Pengfei Yu, Heng Ji, and Prem Natarajan. 2021b. Lifelong event detection with knowledge transfer.

Chenlong Zhang, Pengfei Cao, Yubo Chen, Kang Liu, Zhiqiang Zhang, and Mengshu Sunand Jun Zhao. 2024. Continual few-shot event detection via hierarchical augmentation networks.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. 2019. Theoretically principled trade-off between robustness and accuracy. In *Proceedings of ICML*, pages 7472–7482. PMLR.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.

Appendices

A Related work

Continual Learning (CL) is a learning scenario that challenges models to continuously acquire new knowledge from a sequence of tasks over time. A major problem in CL is *catastrophic forgetting* (French, 1993; Thrun and Mitchell, 1995; French and Chater, 2002), where the model significantly loses its ability to perform previous tasks. To address this issue, one effective approach is memory-based techniques (Rebuffi et al., 2017; Shin et al., 2017; Wang et al., 2019; Han et al., 2020b), which proposed storing a few key samples from the current task in a memory buffer and revisiting them when learning new tasks to reinforce past knowledge.

Fewshot Continual Relation Extraction (FCRE) is a challenging scenario, which was introduced by (Qin and Joty, 2022c) for Relation Extraction problems. To deal with catastrophic forgetting and overfitting phenomenon caused by the extremely limited availability of data for each new task, recent work like Wang et al. (2023b); Chen et al. (2023); Ma et al. (2024) propose memory-based approaches, which suggest imposing regularization functions during training. Specifically, Wang et al. (2023b) proposed using serial objective functions based on contrastive and distillation, Qin and Joty (2022c) proposed leveraging extra training data from external unlabeled text, and Chen et al. (2023) proposes a prototype-based learning strategy to help the model enhance the ability to distinguish between different relation representations. Recently, (Tran et al., 2024) introduced a novel approach where the often discarded component - pretrained LM heads are employed as a regularization strategy, which helps reduce overfitting as well as forgetting significantly.

We find that the root cause of FCRE is the limited training data. Therefore, unlike existing works, we propose a novel data augmentation strategy where models can have opportunities to gain fresh perspectives on new data samples, incorporating both old and new task knowledge. This creates a robust training strategy and achieves superior testing results.

Sharpness Aware Minimization Flat minimizers have been shown to be more robust to the shifts between training and test losses, thereby enhancing the generalization ability of neural networks (Jiang et al., 2020; Petzka et al., 2021; Dziugaite and Roy, 2017). Among the flat minimizers, Sharpness-Aware Minimization (SAM), introduced by Foret et al. (2020), has gained significant attention due to its effectiveness and scalability. SAM's versatility has been leveraged across a wide range of tasks and domains, including domain generalization (Cha et al., 2021; Wang et al., 2023a), federated learning (Caldarola et al., 2022), etc.

However, the potential of SAM in Few-shot Continual Information Extraction tasks, particularly in FCRE, remains underexplored. To address this gap, we conducted extensive experiments and offer valuable insights that contribute to advancing research in this area.

B Background

B.1 The base FCRE method (CPL)

This method proposed techniques including prompt design, representation learning loss, and memory augmentation strategy to address the challenges of catastrophic forgetting and overfitting in FCRE:

(a) The prompt designing CPL employs a soft prompting technique, where learnable tokens guide the behavior of models instead of explicit, human-understandable words. The template is defined as:

$$\begin{aligned} T(x) = x.[v_{0:n_0-1}]e_h[v_{n_0:n_1-1}][MASK] \\ [v_{n_1:n_2-1}]e_t[v_{n_2:n_3-1}]. \end{aligned} \tag{7}$$

where x is the input sentence, e_h and e_t are the head and tail entities, $[v_i]$ are learnable continuous tokens, n_j determines the number of tokens in each phrase, and $[MASK]$ represents the relation between entities.

(b) For representation learning, CPL leverages a margin-based contrastive learning (MCL) objective function, where the loss for each sample $\boldsymbol{x_i}$ is formulated as follows:

$$L_{MCL}(i) = -\sum_{p \in P(i)} \log \frac{\exp(\alpha_{i,p} \cdot s_{i,p}/\tau)}{\sum_{a \in I} \exp(\alpha_{i,a} \cdot s_{i,a})/\tau} \tag{8}$$

where $P(i) = \{p \in I : y_p = y_i\}$ is the index set of positive samples w.r.t sample $\boldsymbol{x_i}$, $I = \{1, 2, ...n\}$ is index set of all n training samples; $s_{i,p} = \frac{\boldsymbol{z}_i \cdot \boldsymbol{z}_p}{\|\boldsymbol{z}_i\| \cdot \|\boldsymbol{z}_p\|}$ is the similarity between the representations $\boldsymbol{z}_i$ and $\boldsymbol{z}_p$ of samples $\boldsymbol{x_i}$ and $\boldsymbol{x_p}$, respectively; $\alpha_{i,p}$ is a relaxation factor, and τ is a temperature parameter.

(c) Finally, *the memory augmentation component* utilizes ChatGPT to generate diverse samples guided by well-crafted prompts, aiming to reduce overfitting in this low-resource scenario. These augmented samples are combined with the original ones to form a new training set for memory replay.

The training process consists of two primary steps: (I) current task training, where the model is trained on new relation samples using the MCL loss, and (II) memory replaying, where the model revisits augmented samples from previous tasks to consolidate knowledge and prevent forgetting.

For relation prediction, CPL employs a Nearest-Class-Mean (NCM) classifier, defined as:

$$\hat{y}_{\boldsymbol{x}} = \underset{r \in \tilde{R}^t}{\operatorname{argmin}} \|\boldsymbol{z}_{\boldsymbol{x}} - p_r\|_2, \quad p_r = \frac{1}{L} \sum_{i=0}^{L} \boldsymbol{z}_{\hat{x}_i^r}, \tag{9}$$

where p_r is the prototype of relation r, $\hat{\boldsymbol{x}}_i^r$ denotes the sample with label $y_{\hat{\boldsymbol{x}}_i^r} = r$ in the memory buffer.

B.2 Few-Shot Continual Event Detection

To further demonstrate the efficacy and versatility of our proposed methodology, we conducted additional experiments on the Few-Shot Continuous Event Detection (FCED) problem.

FCED aims to detect emerging events with limited sample data. Given a set of tasks $T = \{T_1, T_2, \ldots, T_n\}$, each task T_i comprises individual training, validation, and testing sets: $T_i = D_i^{\text{train}}, D_i^{\text{dev}}, D_i^{\text{test}}$. Each set $D_i = \{(X_i^j, Y_i^j)\}_{j=1}^m$ consists of samples X and their corresponding labels Y, where m denotes the number of event types in each task. The initial task T_1 serves as the base task T_{base}, containing a substantial number of training samples. Subsequent tasks are defined as few-shot incremental tasks $T_{\text{inc}} = T_2, T_3, \ldots, T_n$, each containing only a limited number of samples (e.g., 5 or 10) for each new event type. It is important to note that for any two tasks T_i and T_j, their event types are mutually exclusive: $T_i \cap T_j = \emptyset$. At time step t, for FCED task C_t, the training set is defined as $C_t^{\text{train}} = D_t^{\text{train}}$, while the validation/testing set is $C_t^{\text{test}} = D_t^{\text{test}} \cup C_{t-1}^{\text{test}}$. This formulation requires the CFED system to maintain consistent performance on all previously observed labels $L_t = \bigcup_{i=1}^t \{Y_i^j\}_{j=1}^m$ using only the currently available training samples in task T_t.

For comparative analysis, we consider the Continual Few-shot Event Detection via Hierarchical Augmentation Networks (HANet) (Zhang et al., 2024) as a baseline. HANet proposes a memory-based framework incorporating two key components: prototypical augmentation and contrastive augmentation. The HANet model employs a BERT-based event detector for trigger extraction and classification. Given an input sentence $S = x_1, x_2, \ldots [e_s, \ldots, e_e], \ldots, x_n$ containing event triggers $E = [e_s, \ldots, e_e]$, the model generates a hidden representation $H \in \mathbf{R}^{n \times d}$. The hidden states of a trigger H_e are constructed by concatenating their start and end representations. The probability of an event type $y_i \in L_t$ at stage t is computed as:

$$p(y_i|h_e) = \frac{\exp(W_i^T h_e + b_i)}{\sum_j \exp(W_j^T h_e + b_j)}$$

where W_i and b_i are learnable parameters. To mitigate catastrophic forgetting, HANet implements prototypical augmentation in the memory set. For each event type, an exemplar is selected, and its feature space is reconstructed using a Gaussian distribution:

$$\hat{H}_e^j = \{\hat{h}_{e,1}^j, \ldots, \hat{h}_{e,n}^j\} \sim \mathcal{N}(\mu_j, \sigma_j^2)$$

where μ_j represents the exemplar's representation and σ_j^2 denotes the variance calculated during exemplar selection. To address overfitting in few-shot scenarios, HANet introduces contrastive augmentation at the token level. This involves constructing positive and negative pairs from augmented tokens and applying contrastive losses for sentence and trigger representations.

To integrate our method into HANet, we consider the representation of the trigger as a proxy for the relation representation.

C Experimental setting

C.1 Datasets

C.1.1 Few-shot Continual Relation Extraction (FCRE)

Our experiments for the FCRE scenario utilize two benchmark datasets:

- **FewRel** (Han et al., 2018): This dataset comprises 100 relations with 70,000 samples. Following Qin and Joty (2022c), we employ a configuration of 80 relations, partitioned into 8 tasks, each containing 10 relations (10-way). The initial task, $\mathcal{T}^1$, includes 100 samples per relation, while subsequent tasks are structured as few-shot tasks under 5-shot settings.

- **TACRED** (Zhang et al., 2017): This dataset encompasses 42 relations with 106,264 samples extracted from Newswire and Web documents. Consistent with (Qin and Joty, 2022c), we exclude instances labeled as "no_relation" and distribute the remaining 41 relations across 8 tasks. The first task, $\mathcal{T}^1$, comprises 6 relations with 100 samples each, while subsequent tasks involve 5 relations (5-way) in 5-shot configurations.

C.1.2 Few-shot Continual Event Detection (FCED)

For the FCED scenario, we construct benchmarks based on two publicly available datasets, following the approach of (Zhang et al., 2024):

- **MAVEN** (Wang et al., 2020): This dataset originally contains 168 event types, representing a comprehensive general domain event detection corpus. We adopt the training/validation/testing split methodology of Yu et al. (2021b), constructing the test set from the initial development set and randomly selecting samples from the original training set to form a new development set. For incremental task splits, we select the most frequent event types to construct FCED tasks, randomly sampling 100 instances for each type in the base task, and 5 or 10 instances for each type in the incremental tasks.

- **ACE 2005** (Walker et al., 2005): This dataset consists of 33 event types. We utilize the training/validation/testing split as established in previous works (Nguyen et al., 2016). The incremental task split methodology is identical to that applied to the MAVEN dataset for constructing FCED tasks.

Our experimental design incorporates 5 sub-tasks. We define an m-way k-shot CFED task as one containing m event types per subtask and k training samples per type. We select the 10 most frequent types to conduct 2-way 5-shot and 2-way 10-shot tasks. For the base task T_{base}, we randomly sample 100 instances per type, while for incremental tasks T_{inc}, we sample 5 and 10 instances per type, respectively.

C.2 Baselines

This study evaluates our approach against state-of-the-art methods in FCRE and FCED. The selected baselines are as follows:

C.2.1 FCRE Baselines

- **CRECL** (Hu et al., 2022): extends beyond conventional few-shot learning by imposing additional constraints on training data. It accomplishes this by integrating information regarding support instances to augment instance representations. Furthermore, it advocates for open-source task

enrichment to facilitate cross-domain knowledge aggregation and introduces the TinyRel-CM dataset tailored specifically for few-shot relation classification with restricted training data. Experimental results illustrate its efficacy in enhancing performance under conditions of limited data availability.

- **ERDA** (Qin and Joty, 2022b): This study introduces Continual Few-Shot Relation Learning (CFRL) as a novel challenge, recognizing the constraints of current methodologies that demand substantial labeled data for new tasks. CFRL endeavors to acquire knowledge of novel relations with minimal data while averting catastrophic forgetting. Addressing this challenge, ERDA presents a methodology grounded in embedding space regularization and data augmentation. This strategy imposes constraints on relational embeddings and integrates supplementary relevant data through self-supervision. Extensive experimentation showcases ERDA's substantial performance enhancements over prior state-of-the-art approaches in CFRL scenarios.

- **SCKD** (Wang et al., 2023b) SCKD implements a systematic knowledge distillation strategy to preserve knowledge from previous tasks. The method integrates contrastive learning techniques with pseudo samples to enhance the discriminative power of relation representations.

- **ConPL** (Chen et al., 2023) introduces a method comprising three core components: a prototype-based classification module, a memory-enhanced module, and a consistent learning module designed to maintain distribution consistency and mitigate forgetting. Furthermore, ConPL employs prompt learning to enhance representation learning and integrates focal loss to reduce confusion among closely related classes.

- **CPL** (Ma et al., 2024) CPL introduces a framework that employs prompts to generalize across categories and utilizes margin-based contrastive learning to address challenging samples. This approach aims to mitigate catastrophic forgetting and overfitting. Additionally, CPL incorporates a memory augmentation strategy, leveraging ChatGPT to generate diverse samples, further addressing overfitting in low-resource FCRE scenarios.

  In this paper, to conduct the ablation study in Table ...

- **CPL+MI** (Tran et al., 2024) (Mutual Information Maximization) is designed to complement and enhance existing baseline methods. It utilizes the often-neglected language model heads to preserve prior knowledge from pre-trained backbones and improve representation learning. This is achieved by maximizing the mutual information between the latent representations of the language model head branch and the main classifier branch.

C.2.2 FCED Baselines

To date, the work of Zhang et al. (2024) represents the sole comprehensive study addressing FCED (elaborated in Section B.2). Consequently, we adopt their methodology and comparative baselines as the foundation for our FCED experiments. The following approaches serve as our benchmarks:

- **KCN** (Cao et al., 2020) A prominent continual event detection method that employs a memory replay-knowledge distillation paradigm.

- **KT** (Yu et al., 2021a) This approach primarily adheres to the memory-based paradigm, incorporating a novel initialization technique for effective knowledge transfer.

- **EMP** (Liu et al., 2022) In addition to memory replay, this method integrates prompt learning for each event type to facilitate the retrieval of knowledge from previous types.

C.3 Training Configurations

This section delineates the optimal hyperparameter values employed across various experimental settings. Tables 5, 6, 7, and 8 present the specific configurations for each model variant.

Hyperparameter	Value
Current-task training epochs	8
Memory training epochs	6
Learning rate	1×10^{-5}
Encoder output dimension	768
BERT input maximum sequence length	256
Margin (m) for Margin Loss	1.0
β_1	0.25
β_2	0.25
σ	0.1

Table 5: Hyperparameter configuration for CPL$_{BERT}$+SAM$_{current}$+augment

Hyperparameter	Value
Current-task training epochs	10
Memory training epochs	5
Learning rate	1×10^{-5}
Encoder output dimension	768
BERT input maximum sequence length	256
Margin (m) for Margin Loss	1.0
β_1	0.25
β_2	0.25
σ	0.05

Table 6: Hyperparameter configuration for CPL$_{BERT}$+MI+SAM$_{current}$+augment

D Additional experimental results

D.1 Our method when ensure the efficiency of SAM in Fewshot Continual Learning

The empirical evidence presented in Tables 9 and 10 provides compelling support for the efficacy of our proposed methodology when integrated with SAM, particularly in comparison to current state-of-the-art baselines. Notably, our approach demonstrates significant performance improvements, yielding increases in average accuracy of up to 3.34% on the TACRED dataset under the FCRE scenario, and 3.39% on the ACE dataset in the FCED scenario compared with the strongest baseline.

D.2 Our method helps avoid forgetting

To illustrate how our method helps avoid forgetting, we provide the accuracy of the learned model on each task overtime in Table 11.

D.3 Training time

While our method does increase the training data volume, it's important to note that in the FCRE scenario from task $\mathcal{T}^2$ onwards, the training data only includes a small number k samples per class (e.g., 5 samples in 5-shot setting). Therefore, the actual training overhead is not substantial. To quantify this, we conducted timing experiments on an A100 40GB GPU (Table 12).

D.4 Additional ablation study

In this part, we provide additional ablation experiments to demonstrate the behaviour of our method on the latest state-of-the-art (CPL+MI+augment) and obtained the notable results in Table 13. These results are consistent with our previous findings on CPL+augment and further validate the crucial role of both loss functions $\mathcal{L}_{CR}$ and $\mathcal{L}_{DL}$ in improving model performance.

Hyperparameter	Value
Current-task training epochs	8
Memory training epochs	6
Learning rate	1×10^{-5}
Encoder output dimension	4096
BERT input maximum sequence length	256
Margin (m) for Margin Loss	1.0
LoRA α	16
LoRA rank	8
LoRA dropout rate	0.05
β_1	0.25
β_2	0.25
σ	0.05

Table 7: Hyperparameter configuration for CPL$_{LLM2Vec}$+SAM$_{current}$+augment

Hyperparameter	Value
Training epochs	30
Learning rate	2×10^{-5}
Encoder output dimension	768
BERT input maximum sequence length	256
Margin (m) for Margin Loss	1.0
β_1	0.5
β_2	0.5
σ	0.05

Table 8: Hyperparameter configuration for HANet+SAM$_{current}$+augment

FewRel *(10-way 5-shot)*

Method	$\mathcal{T}^1$	$\mathcal{T}^2$	$\mathcal{T}^3$	$\mathcal{T}^4$	$\mathcal{T}^5$	$\mathcal{T}^6$	$\mathcal{T}^7$	$\mathcal{T}^8$	avg
ERDA	92.43	64.52	50.31	44.92	39.75	36.36	34.34	31.96	49.32
CRECL	93.93	82.55	74.13	69.33	66.51	64.60	62.97	59.99	71.75
ConPL**	**95.18**	79.63	74.54	71.27	68.35	63.86	64.74	62.46	72.50
SCKD	94.77	82.83	76.21	72.19	70.61	67.15	64.86	62.98	73.95
CPL	94.87	85.14	78.80	75.10	72.57	69.57	66.85	64.50	75.93
CPL+MI	94.69	85.58	80.12	75.71	73.90	70.72	68.42	66.27	76.93
CPL+augment+SAM$_{current}$	94.48	84.34	**80.80**	77.29	**75.75**	**73.07**	**70.45**	68.43	78.08
CPL+MI+augment+SAM$_{current}$	94.59	**85.62**	80.79	**77.4**	75.73	72.61	70.04	**68.57**	**78.17**

TACRED *(5-way 5-shot)*

Method	$\mathcal{T}^1$	$\mathcal{T}^2$	$\mathcal{T}^3$	$\mathcal{T}^4$	$\mathcal{T}^5$	$\mathcal{T}^6$	$\mathcal{T}^7$	$\mathcal{T}^8$	avg
ERDA	81.88	53.68	40.36	36.17	30.14	22.61	22.29	19.42	38.32
CRECL	87.09	78.09	61.93	55.60	53.42	51.91	47.55	45.53	60.14
ConPL**	**88.77**	69.64	57.50	52.15	58.19	55.01	52.88	50.97	60.64
SCKD	88.42	79.35	70.61	66.78	60.47	58.05	54.41	52.11	66.28
CPL	86.27	81.55	73.52	68.96	63.96	62.66	59.96	57.39	69.28
CPL+MI	85.67	82.54	75.12	70.65	66.79	65.17	61.25	59.48	70.83
CPL+augment+SAM$_{current}$	86.58	82.05	75.35	73.31	69.20	68.55	64.68	62.62	72.79
CPL+MI+augment+SAM$_{current}$	86.55	**82.75**	**76.16**	**73.93**	**71.14**	**70.37**	**67.36**	**65.11**	**74.17**

Table 9: Comparative analysis of accuracy (%) for various BERT-based methodologies evaluated on the TACRED and FewRel datasets under 5-shot settings in the FCRE scenario.

ACE *(2-way 5-shot)*

Method	$\mathcal{T}^1$	$\mathcal{T}^2$	$\mathcal{T}^3$	$\mathcal{T}^4$	$\mathcal{T}^5$	avg
KCN	60.86	56.38	47.56	38.62	37.05	48.09
KT	53.16	42.55	33.93	38.48	31.27	39.88
EMP	54.78	40.49	24.32	27.15	22.53	33.85
HANet	61.16	63.07	**57.50**	53.21	54.31	57.85
HANet+augment+SAM$_{current}$	**63.51**	**64.99**	56.46	**61.45**	**57.11**	**60.70**

ACE *(2-way 10-shot)*

Method	$\mathcal{T}^1$	$\mathcal{T}^2$	$\mathcal{T}^3$	$\mathcal{T}^4$	$\mathcal{T}^5$	avg
KCN	60.86	59.41	57.39	46.48	44.3	53.69
KT	53.16	59.12	50.02	49.02	28.54	47.97
EMP	54.78	37.28	19.6	34.69	24.19	34.11
HANet	61.16	**66.84**	64.68	58.02	54.37	61.02
HANet+augment+SAM$_{current}$	**63.60**	65.96	**67.06**	**64.67**	**60.75**	**64.41**

MAVEN *(2-way 5-shot)*

Method	$\mathcal{T}^1$	$\mathcal{T}^2$	$\mathcal{T}^3$	$\mathcal{T}^4$	$\mathcal{T}^5$	avg
HANet**	57.64	50.23	55.82	52.07	52.35	53.62
HANet+augment+SAM$_{current}$	**58.42**	**54.58**	**57.10**	**54.60**	**53.84**	**55.71**

MAVEN *(2-way 10-shot)*

Method	$\mathcal{T}^1$	$\mathcal{T}^2$	$\mathcal{T}^3$	$\mathcal{T}^4$	$\mathcal{T}^5$	avg
HANet**	57.64	53.28	58.67	56.23	54.85	56.13
HANet+augment+SAM$_{current}$	**58.31**	**56.59**	**58.98**	**56.93**	**56.53**	**57.47**

Table 10: Comparative analysis of HANet baseline and HANet integrated with our proposed methodology on ACE and MAVEN datasets. Results depict accuracy (%) under 5-shot and 10-shot settings in the FCED scenario.

CPL-MI on FewRel *(10way-5shot)*

Task	$\mathcal{T}^1$	$\mathcal{T}^2$	$\mathcal{T}^3$	$\mathcal{T}^4$	$\mathcal{T}^5$	$\mathcal{T}^6$	$\mathcal{T}^7$	$\mathcal{T}^8$
$\mathcal{T}^1$	94.87	–	–	–	–	–	–	–
$\mathcal{T}^2$	89.00	78.27	–	–	–	–	–	–
$\mathcal{T}^3$	84.10	75.35	76.03	–	–	–	–	–
$\mathcal{T}^4$	79.62	72.42	69.89	78.88	–	–	–	–
$\mathcal{T}^5$	75.22	69.03	65.56	76.35	76.84	–	–	–
$\mathcal{T}^6$	71.28	65.65	62.99	72.23	73.02	68.91	–	–
$\mathcal{T}^7$	70.12	62.35	59.33	70.82	70.27	65.72	73.96	–
$\mathcal{T}^8$	66.21	60.32	58.00	65.28	69.51	60.78	68.96	71.73
$\Delta \downarrow$	28.67	17.97	18.03	13.52	7.33	8.12	5.00	–

CPL-MI-aug on FewRel *(10way-5shot)*

Task	$\mathcal{T}^1$	$\mathcal{T}^2$	$\mathcal{T}^3$	$\mathcal{T}^4$	$\mathcal{T}^5$	$\mathcal{T}^6$	$\mathcal{T}^7$	$\mathcal{T}^8$
$\mathcal{T}^1$	94.45	–	–	–	–	–	–	–
$\mathcal{T}^2$	90.22	80.48	–	–	–	–	–	–
$\mathcal{T}^3$	87.43	77.23	78.22	–	–	–	–	–
$\mathcal{T}^4$	82.08	74.21	71.69	80.62	–	–	–	–
$\mathcal{T}^5$	78.66	71.18	68.36	78.64	78.51	–	–	–
$\mathcal{T}^6$	75.55	68.44	66.01	76.36	76.13	72.53	–	–
$\mathcal{T}^7$	73.14	66.06	63.09	73.75	73.77	68.34	75.65	–
$\mathcal{T}^8$	70.74	64.04	59.12	70.23	71.14	65.76	73.16	74.64
$\Delta \downarrow$	23.71	16.44	19.08	10.39	7.37	6.77	2.49	–

Table 11: Accuracies on each task overtime

Setting	**Model**	**Training Time (mins)**	**Overhead Increase**
5-way 5-shot TACRED	CPL-MI (Baseline)	97.30	-
5-way 5-shot TACRED	CPL-MI-SAM-aug	107.92	10.91%
10-shot FewRel	CPL-MI (Baseline)	141.80	-
10-way 5-shot FewRel	CPL-MI-SAM-aug	160.09	12.90%

Table 12: Training Time and Overhead Increase for Different Models

FewRel *(10way-5shot)*

Method	$\mathcal{T}^1$	$\mathcal{T}^2$	$\mathcal{T}^3$	$\mathcal{T}^4$	$\mathcal{T}^5$	$\mathcal{T}^6$	$\mathcal{T}^7$	$\mathcal{T}^8$
CPL+MI+augment	**94.76**	**85.48**	**80.24**	77.69	**75.60**	72.94	**70.74**	**68.36**
w.o. L_{CR}	94.01	84.84	79.70	**78.10**	74.23	**73.11**	68.92	67.03
w.o. L_{DL}	94.53	83.90	78.74	76.90	73.98	71.35	67.76	67.50

TACRED *(5way-5shot)*

Method	$\mathcal{T}^1$	$\mathcal{T}^2$	$\mathcal{T}^3$	$\mathcal{T}^4$	$\mathcal{T}^5$	$\mathcal{T}^6$	$\mathcal{T}^7$	$\mathcal{T}^8$
CPL+MI+augment	**86.33**	**82.31**	**76.35**	70.93	**68.28**	65.04	**62.60**	**61.97**
w.o. L_{CR}	85.43	81.23	76.29	**71.16**	65.20	**66.20**	61.18	59.47
w.o. L_{DL}	83.89	80.79	75.80	69.25	67.01	65.43	61.88	60.34

Table 13: Ablation study on CPL+MI