Enhancing Contrastive Learning with Noise-Guided Attack: Towards Continual Relation Extraction in the Wild

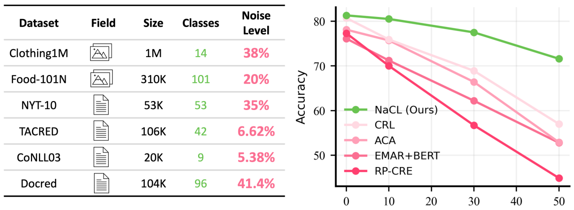
Ting Wu[1]*, Jingyi Liu[1]*, Rui Zheng[1], Qi Zhang[1,3], Tao Gui[2], Xuanjing Huang[1,3]
[1]School of Computer Science, Fudan University
[2]Institute of Modern Languages and Linguistics, Fudan University
[3]Shanghai Collaborative Innovation Center of Intelligent Visual Computing
`{tingwu21, liujingyi21}@m.fudan.edu.cn`

Abstract

The principle of continual relation extraction (CRE) involves adapting to emerging novel relations while preserving old knowledge. Existing CRE approaches excel in preserving old knowledge but falter when confronted with contaminated data streams, likely due to an artificial assumption of no annotation errors. Recognizing the prevalence of noisy labels in real-world datasets, we introduce a more practical learning scenario, termed as *noisy-CRE*. In response to this challenge, we propose a noise-resistant contrastive framework called Noise-guided Attack in Contrastive Learning (NaCL), aimed at learning incremental corrupted relations. Diverging from conventional approaches like sample discarding or relabeling in the presence of noisy labels, NaCL takes a transformative route by modifying the feature space through targeted attack. This attack aims to align the feature space with the provided, albeit inaccurate, labels, thereby enhancing contrastive representations. Extensive empirical validations demonstrate the consistent performance improvement of NaCL with increasing noise rates, surpassing state-of-the-art methods [1].

1 Introduction

Alongside the predictive wins of relation extraction (RE) on various benchmarks (Trisedya et al., 2019; Ye et al., 2022), the need for the ability to acquire sequential experience in dynamic environments stands out the significance. Catering to the real-world learning requirement, a new RE formulation, namely continual relation extraction (CRE), has been proposed (Wang et al., 2019).



Dataset	Field	Size	Classes	Noise Level
Clothing1M	🖼	1M	14	38%
Food-101N	🖼	310K	101	20%
NYT-10	📄	53K	53	35%
TACRED	📄	106K	42	6.62%
CoNLL03	📄	20K	9	5.38%
Docred	📄	104K	96	41.4%

Figure 1: Left Table: Noisy labels exist widely in well-annotated benchmarks. Right Plot: Performance of the state-of-the-art CRE methods drop significantly on TACRED with noise ratio ranging from 0% to 50%.

Under this topic, catastrophic forgetting (McCloskey and Cohen, 1989) where previous knowledge is overwritten as new concepts are learned, remains a key challenge. To prevent forgetting, a variety of sophisticated methods are developed by memory replay (Rebuffi et al., 2017; Sun et al., 2020), weight regularization (Kirkpatrick et al., 2017) or architecture expansion (Hung et al., 2019). Wang et al. (2019) explicitly store past experiences into a limited memory and replay them to complement new tasks learning. In comparison to exemplars storage, Dong et al. (2021) impose constraints on the update of the important network weights for old knowledge consolidation. As for architecture-based method, it dynamically changes model architectures to acquire new information while remembering previous knowledge (Ehret et al., 2021).

Despite the effectiveness, all of these methods implicitly assume the correctness of the labels for the streaming data. In practice, such an assumption is rather artificial even impossible to satisfy since label shifts are inevitable in real-world scenarios. Worse still, official statistics in the table of Figure 1 reveal that the widely used benchmarks with elaborate human annotations, likewise, contain a certain proportion of noisy labels. Due to the ignorance of noisy labels over data streams, it is clear to see in Figure 1 that state-of-the-art CRE models fail to defend against label inconsistency, resulting in significant performance drops.

To break the impractical structure of current CRE setup and to enhance the noise-resistant capacity of models, in this paper, we present a more

* Equal Contribution.

[1] Our code and data are available at `https://github.com/CuteyThyme/Noisy-CRE.git`.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2227–2239
August 11-16, 2024 ©2024 Association for Computational Linguistics

generalized learning setting coined as *noisy-CRE*. In this challenging scenario, there is a potential for mislabeled samples to contaminate the sequential stream in every incremental task. We assume that models trained under the noisy-CRE setting can reflect their ability to adapt to new relations in the real world.

In the face of the great challenge, in this paper, we propose a robust contrastive framework as **N**oise-guided **a**ttack **C**ontrative **L**earning (NaCL) for noisy-CRE. Generally, handling noisy labels can be relaxed to a subsequent process of clean sample selection and noisy sample correction. In NaCL, we introduce an auxiliary model to play the two roles. **First**, at each new task, the auxiliary model will be re-initialized to train for new relations learning. Intriguingly, we term it as *reboot*, which can make the model escape the interference of prior knowledge so that its logit outputs can be a measure of clean sample selection for current task. **Second**, this model will translate a novel sight into feature space for correction by performing *noise-guided attack*. This attack can actively drive the feature distribution of noisy negatives more aligned with their given labels.

To demonstrate the effectiveness of NaCL, we design two benchmarks based on FewRel and TACRED. Empirical results and in-depth analyses show that our NaCL can achieve consistent improvements when noise rates vary from light to heavy, and it outperforms all state-of-art baselines far ahead. In summary, the contributions of this work are three-fold:

- We define a practical noisy-CRE setting and construct well-designed benchmarks. To the best of our knowledge, this is the first work to improve the robustness of CRE models against noisy labels.
- We propose NaCL, a noise-resistant contrastive framework that can jointly prevent catastrophic forgetting and learn with noisy labels.
- We provide empirical results and extensive assessments to verify the effectiveness of NaCL, outperforming other state-of-the-art baselines adapted from CRE methods by a large margin.

2 Noisy-CRE Setting Formulation

Continual relation extraction is defined as training models on non-stationary data from sequential tasks. In the setup of noisy-CRE, we first define a sequence of tasks $\mathbb{T} = (\mathcal{T}^1, \cdots, \mathcal{T}^n)$. For the k-th task $\mathcal{T}^k$, its training dataset is denoted as $\mathcal{D}^k_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{N_k}$ containing tuples of the input sample $x_i \in \mathcal{X}$ and corresponding relation label $y_i \in \mathcal{Y}$, where $\mathcal{Y}$ has a probability of rate to be corrupted. Our goal is to train a single model $f_\theta : \mathcal{X} \to \mathcal{Y}$ parameterized by θ, such that it predicts the label $y = f_\theta(\mathbf{x}) \in \mathcal{Y}$ given an unseen test sample x from arbitrary learned tasks.

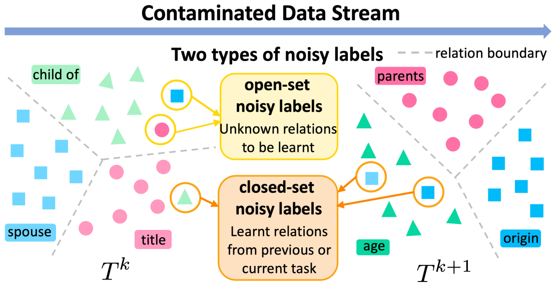


Figure 2: The generalized setting of noisy-CRE with two types of noisy labels existing in the contaminated data stream.

Protocols for Label Corruption. In an ideal CRE mode, each task has independent relation space $\mathcal{Y}$. However, for noisy-CRE, due to the inevitable label corruption, this assumption does not hold in the training set. As shown in Figure 2, the relation space $\mathcal{Y}^k$ of the k-th task can be contaminated arbitrarily by samples from label space $\mathcal{Y}^i$ with $i \in \{1, \cdots, k-1, k+1, \cdots, n\}$, thus leading to two kinds of noisy labels. When $i \leq k$, we term these noisy labels as *closed-set* ones, since their gold relations are embedded in the model knowledge and can be recovered. In contrast, when $i > k$, the gold relations of the noisy ones are unreachable and formed as *open-set* noise.

3 NaCL: Towards Noise-resistant CRE

In this section, we present NaCL, our noise-resistant contrastive learning framework designed to simultaneously handle closed-set and open-set noisy labels in the noisy-CRE scenario.

3.1 Overall Framework

Building upon noisy-CRE setting, the learning process of each task contains two components: new relations learning with noisy labels and memory replay for old knowledge consolidation, as presented in the overall framework depicted in Figure 5.

New Relations Learning. When learning a new task $\mathcal{T}^k$, the presence of noisy labels can lead to the introduction of false contrastive pairs in vanilla contrastive learning framework. To mitigate this issue, NaCL employs two procedures. First, a rebooted selection process is executed to identify clean positive samples, as described in Section 3.2. Second, a noise-guided attack is performed on noisy samples to generate hard negatives, which is discussed in Section 3.3.

Old Knowledge Replay. Once new relations are well-learned at the completion of each task, clean and representative samples stored in the memory buffer will be replayed for old relations prevention.

3.2 Rebooted Selection for Clean Positives

To handle the noisy labels, a broadly applied criterion is to select samples with small losses and treat them as clean data. It is inspired by empirical observations that deep learning models tend to learn simple patterns first before overfitting on the noisy labels (Arpit et al., 2017; Zhang et al., 2017a).

As shown in Figure 3, we can observe the model quickly converges to a small loss for the first task. However, as the task progresses, an obvious loss threshold between clean and noisy samples gradually disappears. We recognize this **failure of small-loss-based selection** is attributed to the old knowledge of prior tasks embedded in model parameters, which prevents the model from learning incremental tasks from scratch.

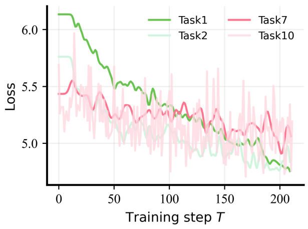


Figure 3: Training loss at different tasks on FewRel with 30% noise ratio.

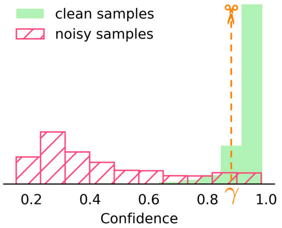


Figure 4: Confidence distribution of clean and noisy samples at Task 10.

For the sake of overcoming the problem originating from knowledge intervention, we propose to introduce an auxiliary model $f_A(\cdot, \theta^*)$ and *reboot* it to help select clean samples at each incremental task. With the decomposition into $f_A = \mathcal{F}_A \circ \mathcal{E}_A$, $\mathcal{E}_A$ being the feature extractor and $\mathcal{F}_A$ the classifier, we train f_A with the following classification loss:

$$J(\mathbf{x}, \mathbf{y}) = -\log p(\mathbf{y}|\mathbf{x}) \quad (1)$$

In light of the fact that $f_A(\cdot, \theta^*)$ is re-initialized at each new task, it can avoid being intervened by previous knowledge. With a classifier introduced in the auxiliary model f_A, we can use the logit probability $p(\mathbf{x})$ as a measure of confidence to differentiate between clean and noisy samples. As shown in Figure 4, for the tenth task trained on FewRel with a 30% noise ratio, a high confidence threshold γ successfully identifies almost all clean samples. Consequently, we can predict pseudo clean and noisy set for $\mathcal{T}^k$ as follows:

$$\mathcal{D}^k_{train} = \begin{cases} \widetilde{D}_{\text{clean}}(\mathbf{x}), & p(\mathbf{x}) \geq \gamma, \\ \widetilde{D}_{\text{noisy}}(\mathbf{x}), & p(\mathbf{x}) < \gamma, \end{cases} \quad (2)$$

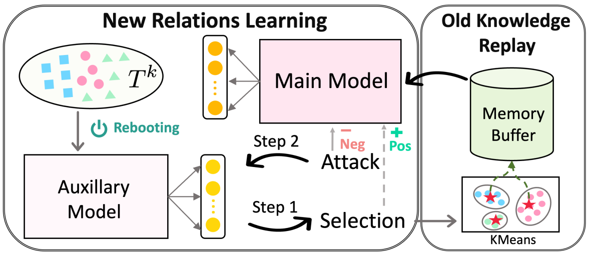


Figure 5: Main framework of NaCL and the training pipeline for $\mathcal{T}^k$ learning.

3.3 Noise-guided Attack for Hard Negatives

Since errors are costly but abstention is manageable, selecting clean samples first and then discarding the noisy ones is a natural approach in the context of learning with noisy labels (Jiang et al., 2018; Xia et al., 2022). Nonetheless, over the contaminated data stream, training samples for each task are limited, and thus direct discarding can lead to a loss of abundant context information. Furthermore, the reduction of negative samples will impair contrastive representation learning (Chen et al., 2020). Account of the two reasons, making use of noisy samples becomes essential.

Noise Correction in Feature Sapce. One typical way to utilize the noisy samples is to relabel them for correction (Li et al., 2020a; Zhou et al., 2021). Faced with the challenge of the co-existence of open-set and closed-set noise, it is impossible for NaCL to apply off-the-shelf techniques to relabel as some noisy labels are unreachable up to current task learning. This inaccessible to label space drives NaCL to translate a novel sight into feature space for noise correction, performed by a variant of targeted attack as *noise-guided attack*.

Noise-guided attack intends to modify the feature to let them match the noisy labels, compared with relabeling that modifies labels to match the given sample features. Within the framework of

NaCL, we re-utilize the auxiliary model f_A to implement the attack. As shown in Figure 5, at each new task $\mathcal{T}^k$, after training for clean sample selection, f_A will act as the proxy to generate adversarial perturbation on the input embeddings of noisy samples. Assuming the noisy labels $\mathbf{y}$ as the attack targets $\mathbf{y}^{tgt}$, the adversarial loss of f_A is essentially to maximize the probability of classification into $\mathbf{y}^{tgt}$ as follows:

$$\mathbf{x}' \leftarrow \Pi_\epsilon\big(\mathbf{x} - \epsilon \text{sign}(\nabla_{\mathbf{x}'}(J(\mathbf{x}', \mathbf{y}^{tgt})))\big) \tag{3}$$

To further help in generating targeted adversarial examples to match the noisy labels actively, we encourage every adversarial sample to move far away from its starting point in the feature space. To achieve this goal, we add a regularization term to the training objective of Equation 3:

$$\begin{aligned}\mathbf{x}' \leftarrow \Pi_\epsilon\big(\mathbf{x} - \epsilon \text{sign}(\nabla_{\mathbf{x}'}(J(\mathbf{x}', \mathbf{y}^{tgt}) \\ + \lambda \text{KL}(f_A(\mathbf{x};\theta^*)||f_A(\mathbf{x}';\theta^*))))\big)\end{aligned} \tag{4}$$

where KL is the Kullback–Leibler divergence, we name this KL regularization as the feature-disruption term, and λ is the fixed hyper-parameter to weigh the contribution of this feature disruption.

Attack as Hard Negative Mining. From the perspective of contrastive representation learning, under the noise-guided attack, noisy samples serving as the negatives all move towards the same direction of the feature space where their noisy label lies. To this extent, it can be viewed as hard negative mining which generates more informative negative samples. What's more, given the fixed attack steps s, some noisy samples originally closer to the positive region can be successfully pushed into this region for positives diversified. Specifically, denoting the relation-wise centroid as c_r by calculating the mean of the hidden representations for each relation from $\widetilde{\mathcal{D}}_{\text{clean}}$, we can obtain $d_{\max}$ that measures the maximum euclidean distance of the clean sample to its centroid c_r. If the distance between the attacked sample $\mathbf{x}'$ and its corresponding relation centroid c_r is smaller than $d_{\max}$, we can recognize this noisy sample is attacked successfully. Consequently, the **attack success rate** (ASR) can be calculated as follows:

$$ASR = \frac{\sum \mathbb{1}\left[||\mathcal{E}_M(\mathbf{x}') - c_r\|_2 <= d_{\max}\right]}{|\widetilde{\mathcal{D}}_{\text{noisy}}|} \tag{5}$$

New Contrastive Pool. We add the successfully attack samples from $\widetilde{\mathcal{D}}_{\text{noisy}}$ into the positive set as $\mathcal{D}_{\text{att-pos}}$. To this end, we can obtain following contrastive samples pool for current task learning:

$$A = \underbrace{\widetilde{\mathcal{D}}_{\text{clean}} \cup \mathcal{D}_{\text{att-pos}}}_{\text{Positive Set } P(\mathbf{x})} \cup \mathcal{D}_{\text{neg}} \tag{6}$$

Final Learning Objective. Hence, we come to the training objective of NaCL for new relations learning:

$$\mathcal{L}_{\text{NaCL}} = -\frac{1}{|P(\mathbf{x})|}\sum_{j\in P(\mathbf{x})} \log \frac{\exp\left(\mathbf{z}_i \cdot \mathbf{z}_j/\tau\right)}{\sum_{k\in A}\exp(\mathbf{z}_i \cdot \mathbf{z}_k/\tau)} \tag{7}$$

where $\mathbf{z}_\ell = \text{Proj}(\mathcal{E}_M(\mathbf{x}))$, $\tau \in \mathbb{R}^+$ is a scalar temperature parameter.

3.4 Memory Replay and Inference

After the stage of k-th task training for new relations, NaCL will select representative samples from $\mathcal{D}^k_{\text{train}}$ to store in the memory buffer $\mathcal{B}$. The buffer size is the number of memory samples needed for each relation, i.e., 20 in our experiments. Like previous rehearsal-based methods for CRE (Han et al., 2020; Cui et al., 2021), we apply K-Means in the representation space produced by $\mathcal{E}_M$ for exemplar selection, which is only carried out in $\widetilde{\mathcal{D}}_{\text{clean}}$. As for each cluster, the sample closest to the cluster center will be selected to store in the buffer $\mathcal{B}$. When the memory buffer is updated with all the seen relations stored, we train f_M with these exemplars of following supervised contrastive loss:

$$\mathcal{L}_{\text{SCL}} = -\frac{1}{|P'(\mathbf{x})|}\sum_{j\in P'(\mathbf{x})} \log \frac{\exp\left(\mathbf{z}_i \cdot \mathbf{z}_j/\tau\right)}{\sum_{k\in \mathcal{B}}\exp(\mathbf{z}_i \cdot \mathbf{z}_k/\tau)} \tag{8}$$

Relation inference. Given a test sample x_i, nearest class mean (NCM) is utilized to obtain the relation predicted by f_M. Concretely, after the training pipeline of $\mathcal{T}^k$, we can obtain the prototype for each seen relation as p_r by calculating the mean of the features from its corresponding exemplars in the buffer $\mathcal{B}$. To be noted, the calculation of the features is in the space after the projector of the main model f_M. Then, we compare the projected representation of x_i with all the prototypes of seen relations and assign the relation label with the closest prototype:

$$\widetilde{y} = \underset{r=1,\dots,C}{\arg\min}\, \|\text{Proj}(\mathcal{E}_M(\mathbf{x})) - p_r\| \tag{9}$$

4 Experiments

4.1 Benchmark Construction

Datasets. We carry out our experiments on widely-used **FewRel** (Han et al., 2018b) and **TACRED** (Zhang et al., 2017b). FewRel is an RE dataset that contains 80 relations, each with 700 instances, and TACRED contains 42 relations and 106,264 samples in total. To be noted, previous works for CRE employ two different task partitioning methods to construct the continual benchmarks, one is the imbalanced division based on clustering of relation embeddings (Wang et al., 2019; Han et al., 2020; Wu et al., 2021) and the other is a random partition with balanced relations for each task (Cui et al., 2021; Zhao et al., 2022). This diversion in task construction makes the baselines incomparable, and we unify them into the same second policy that we split FewRel and TACRED into 10 clusters of relations, leading to 10 tasks and each relation just belongs to only one task.

Noise generation. We design four levels of random noisy labels to accommodate varying noise rates in real-world data, including clean data, 10% noisy data, 30% noisy data, and 50% noisy data for $\mathcal{D}^k_{\text{train}}$ at each task $\mathcal{T}^k$. To generate synthetic noises that contain both close-set and open-set noisy labels, we first randomly flip the relation labels across the whole dataset according to the noise ratio. Then, we partition the dataset based on the flipped relations and cluster them into ten sequential tasks.

4.2 Baselines

We adapt the following state-of-the-art CRE baselines to the proposed noisy-CRE setting and make a comparison with our NaCL model.

EA-EMR (Wang et al., 2019) employs memory replay and embedding alignment to tackle the problem of embedding space distortion when training on new tasks.

EMAR (Han et al., 2020) applies episodic memory activation and reconsolidation mechanism to maintain learned knowledge.

CML (Wu et al., 2021) adopts meta learning and curriculum learning to cope with the challenges of catastrophic forgetting and order-sensitivity in continual relation extraction.

RP-CER (Cui et al., 2021) refines sample embeddings with an attention-based memory network fed with relation prototypes to alleviate catastrophic forgetting.

CRL (Zhao et al., 2022) proposes a consistent representation learning that maintains the stability of the relation by adopting contrastive learning and knowledge distillation when replaying memory.

ACA (Wang et al., 2022) points out catastrophic forgetting problem of previous CRE models mainly lies in shortcuts learning and applies a simple yet effective adversarial class augmentation mechanism to learn more robust representations.

Joint-training corresponds to training a model from scratch during each incremental task with the total dataset containing all data about new and past classes. We treat the performance of joint-training model on clean dataset as *upper bound*.

Finetuning in the other hand represents the *lower bound* of performance, as it is a simple training setup that fine-tunes the model at each incremental task with no replay, regularization or model expansion.

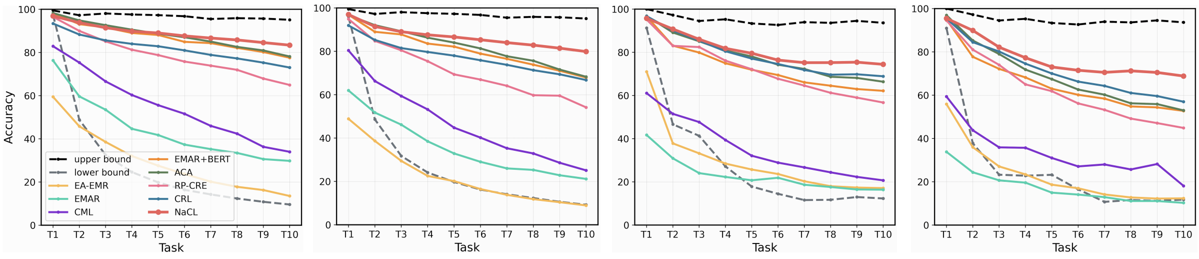
4.3 Training Details and Evaluation Metrics

Implementation Details. The main model f_M is composed of a feature extractor $\mathcal{E}_M$ implemented by BERT-base (Devlin et al., 2019) and a projector of 2-layer MLP. For the auxiliary model f_A, its feature extractor is implemented by another BERT-base, and the output dimension of the classifier $\mathcal{F}_A$ is the relation numbers of each incremental task, *i.e.*, 8-dim for FewRel and 4-dim for TACRED. At each session k, we will re-initialized $f_A(;\theta^*)$ and train it for 3 epochs to help select the clean samples. Following the baseline methods (Cui et al., 2021; Zhao et al., 2022), we adopt Adam as the optimizer with the learning rate of 1e-5 on FewRel and 2e-5 on TACRED for both main model and auxiliary model. Considering that baselines all leverage memory replay to help attenuate catastrophic learning, we set a fixed memory size of 20 for relation-wise storage when re-implementing all methods for the sake of a fair comparison.

Evaluation Metrics. As the main performance metric, we adopt **last test accuracy**, where after all tasks are learned, testing on the test sets of all tasks. We report the average accuracy over 5 random runs. Additionally, we introduce a **normalized forgetting** metric to quantify the severity of catastrophic learning. As a self-relative metric on the performance drop of the first task, the forgetting measure from previous works (Liu et al., 2020) applied to a noisy setting could be misleading since even if a model performs poorly, small forgetting metric

Models	FewRel						TACRED					
	Acc (%) ↑			*Forget* (%) ↓			*Acc* (%) ↑			*Forget* (%) ↓		
	◷	◔	◑	◷	◔	◑	◷	◔	◑	◷	◔	◑
Joint-training	88.1	73.7	56.4	–	–	–	87.3	70.2	50.4	–	–	–
Finetuning	10.0	9.6	9.3	100.0	100.0	100.0	12.6	12.3	11.7	100.0	100.0	100.0
EA-EMR (Wang et al., 2019)	22.3	13.5	8.9	84.3	93.9	96.1	23.6	17.1	12.3	89.5	95.7	95.9
EMAR (Han et al., 2020)	37.2	29.8	21.2	64.7	72.2	78.2	19.7	16.4	10.3	78.8	76.2	88.5
CML (Wu et al., 2021)	37.1	34.0	25.1	68.2	85.3	89.4	22.4	20.7	18.1	70.1	79.2	81.3
EMAR+BERT	83.0	77.6	67.9	22.1	33.0	42.1	71.2	62.2	52.8	27.7	37.5	47.7
RP-CRE (Cui et al., 2021)	77.1	65.0	54.2	30.2	42.7	56.7	70.0	56.7	44.9	37.4	52.5	64.7
CRL (Zhao et al., 2022)	77.7	73.0	66.8	13.7	17.3	19.9	75.9	68.9	57.0	21.1	27.4	41.9
ACA (Wang et al., 2022) †	84.1	78.1	68.3	18.9	27.3	38.9	75.7	66.4	52.9	25.8	38.2	54.6
NaCL	**84.1**	**83.7**	**80.5**	**11.4**	**16.0**	**16.8**	**80.5**	**77.5**	**71.6**	**13.1**	**16.8**	**24.6**

Table 1: **Last test accuracy and forgetting** on FewRel and TACRED with noise ratio of {◷10%, ◔30%, ◑50%}. We re-implement all the baselines with equal task division and evaluation for a fair comparison. † indicates EMAR+ACA since ACA is implemented based on the backbone of EMAR and RP-CRE, and it achieves better accuracy.



(a) FewRel 30% noise (b) FewRel 50% noise (c) Tacred 30% noise (d) Tacred 50% noise

Figure 6: **Accuracy** (%) **on all seen relations** at the stage of learning current tasks with varying noise rates on FewRel and TACRED.

values will be observed due to its little information to forget from the beginning. Therefore, we normalize this forgetting on the accuracy of the first task.

$$Forget = \frac{|\mathcal{A}^n_{\mathcal{T}=1} - \mathcal{A}^1_{\mathcal{T}=1}|}{\mathcal{A}^1_{\mathcal{T}=1}} \tag{10}$$

where $\mathcal{A}^k_{\mathcal{T}=1}$ denotes the accuracy on the first task at the session k. For *accuracy*, the larger is better, while for *forget*, the smaller will be better.

4.4 Main Results

We compare the proposed NaCL with nine baselines on FewRel and TACRED with varying label noise and summarize the results in Table 1.

Overall Performance. Table 1 clearly demonstrates that NaCL achieves consistent performance improvements with noise rate from light to heavy, and outperforms all the baselines by a large margin. Furthermore, we can observe that: **(i)** Apart from our NaCL, all the baselines suffer from the vulnerability of label flips in the continual stream, indicating current CRE models are not resistant to noisy labels. It is apparent to see as the noise rate increases, their last test accuracy declines sharply and the forget rate remains high. **(ii)** Comparison among the baselines validates that BERT-like pre-trained language models are better continual learners since EA-EMR, EMAR, and CML that leverage LSTM as main feature extractor attain worse performances. **(iii)** There is a close connection between model learning accuracy and the ability to defend against catastrophic forgetting. As shown in Figure 6, test accuracy over ten incremental tasks depicts a vivid trend that if a model achieves high accuracy at each incremental task, its final forget rate tends to retain at a low level.

Purity of Memory Buffer. As rehearsal-based methods served for old knowledge consolidation,

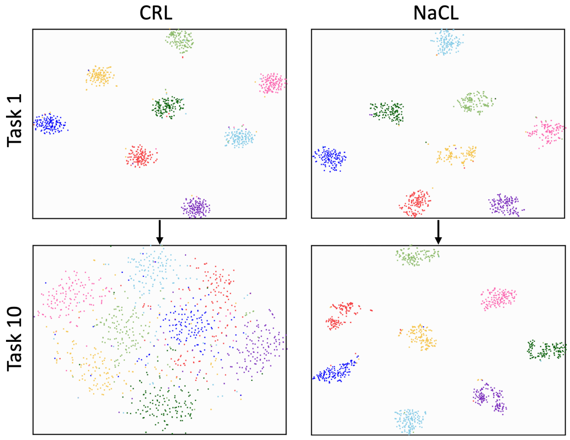


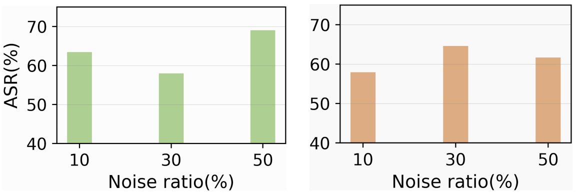
Figure 7: **t-SNE visualization** of relation representation learned from Task 1 and tested by CRL and NaCL at the last task, with a noise rate of 50% on FewRel. Colors stand for different relations.

the purity of the memory buffer is vital. Therefore, we compare the ratio of clean samples in the memory between NaCL and the high-performing baselines. As shown in Table 2, we observe that EMAR-BERT, RP-CRE and CRL all experience a significant decrease in the purity of the memory buffer as the noise rate increases. In contrast, NaCL is able to maintain comparative purification even with the noise rate increasing.

	FewRel			TACRED		
noise rate(%)	10	30	50	10	30	50
EMAR-BERT	80.2	58.9	40.7	76.1	60.0	46.1
RP-CRE	88.1	76.4	63.8	79.1	63.1	50.9
CRL	68.3	47.2	36.3	71.4	53.6	41.2
NaCL	**98.6**	**96.4**	**80.3**	**94.8**	**82.4**	**71.5**

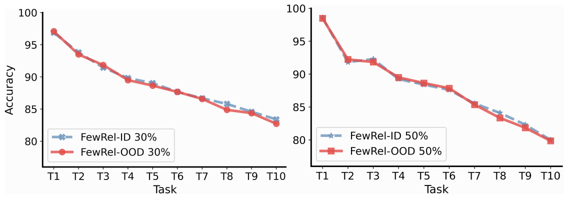
Table 2: **Purity** of the memory buffer.

Preserve of Cluster Relative Positions. We further demonstreate the t-SNE visualization of the representations learned at the first task and tested at the subsequent tasks in Figure 7. As we can observe, compared to CRL, NaCL can achieve more compact clustering of the representations in the feature space and better preserve the relative positions of each relation cluster. It is worth noting that when approaching the last task, relations learned with CRL become indistinguishable, while NaCL maintains their structures, revealing that NaCL has a better capacity to prevent catastrophic forgetting.



(a) FewRel (b) TACRED

Figure 8: **Attack success rate** with noise ratio of $\{10\%, 30\%, 50\%\}$.



(a) 30% Noise Ratio (b) 50% Noise Ratio

Figure 9: **Accuracy** (%) **on all seen relations** at the stage of learning current tasks with varying noise rates on FewRel ID set and OOD set (TACRED).

5 Analysis and Discussion

5.1 Effectiveness of Adversarial Attack

From the results in Table 3, we can conclude that compared with discarding the expected noisy samples directly, employing targeted adversarial attack can de facto make better use of the noisy ones, thus leading to performance improvements. To better investigate the influence of attack, we calculate attack success rate by Equation 5 on FewRel and TACRED with different noise rates. As shown in Figure 8, by imposing a small perturbation on the input embedding, noise-guided attack can successfully force a great number of samples to the direction of their noisy labels in the feature space.

		FewRel			TACRED		
		Acc (%) ↑			*Acc* (%) ↑		
Noise	Attack	10	30	50	10	30	50
Discarding		81.1	80.7	76.9	77.8	72.4	68.5
✓		83.0	82.1	78.0	78.6	75.5	70.5
✓	✓	**84.1**	**83.7**	**80.5**	**80.5**	**77.5**	**71.6**

Table 3: **Ablation studies** on the noise-guided attack, compared with noisy samples discarding.

5.2 Globally Open-set Label Noise

In real-world applications, we expect a robust continual learner to be able to adapt well to noisy data streams, even with out-of-distribution (OOD) samples. Empirical results have demonstrated that NaCL can successfully handle both closed-set label flips and open-set outliers. However, the meaning of *open-set* we introduced before is only from a local perspective relative to the task progression. To explore the potential for noisy label learning from a global OOD set, as for FewRel, we further construct the label noise completely from TACRED. As the experimental results in Figure 9 show, NaCL achieves consistent performance when transferring from FewRel-ID to FewRel-OOD with varying noise rates, which demonstrates the superiority of NaCL for the strong noise resistance.

6 Related Work

6.1 Continual Learning

Prevalent methods for continual learning to tackle catastrophic forgetting problem can be categorized into three macro-types: *rehearsal-based*, *regularization-based*, and *architecture-based* ones. Specifically, rehearsal-based methods construct a data buffer to save samples from older tasks to train with data at the current task (Rebuffi et al., 2017). When the buffer storage is limited, exemplar selection techniques (Aljundi et al., 2019) or generative modeling (Sun et al., 2020) are developed to help approximate the old data distribution. Viewed as exemplar-free methods without storing old task data, regularization-based ones consolidate old knowledge by limiting the learning rate on important parameters for previous tasks (Kirkpatrick et al., 2017). Differently, architecture-based methods aim at having separate components for each task, and these task-specific components can be identified by expanding the network (Loo et al., 2021) or attending to task-specific sub-networks (Gurbuz and Dovrolis, 2022).

Among them, rehearsal-based methods are substantiated to be the most effective paradigm in consolidating old knowledge (Wang et al., 2019; Sun et al., 2020). In this work, we consider combining NaCL with memory replay to help handle the severe forgetting problem.

6.2 Learning with Noisy Labels

Deep neural networks are validated to easily overfit noisy labels resulting in poor generalization performance (Arpit et al., 2017). To improve model generalization with noisy labels, numerous approaches have been developed from various perspectives, *e.g.*, loss correction (Hendrycks et al., 2018), robust loss functions with provable noise tolerance (Ma et al., 2020), sample-reweighting (Ren et al., 2018), curriculum learning (Zhou et al., 2021) and model co-teaching (Han et al., 2018a; Yu et al., 2019). The principle idea shared among these methods is to detect clean labels while discarding, down-weighting or relabeling the wrong labels.

Up to now, none of the works has focused on continual learning with noisy labels. Although strategies above seem to be well-handled for noisy labels, they are confined to *closed-set* label flips and hence cannot be applied to our noisy-CRE setting. To be more generalized, our NaCL undertakes noise correction in the feature space to resolve both closed-set and open-set label noise.

6.3 Contrastive Representation Learning

As a dominant paradigm for representation learning, unsupervised contrastive learning (UCL) has achieved comparable performance. The core idea behind UCL is to pull the anchor and the positive sample close to each other while pushing apart the anchor and the negative sample in embedding space (He et al., 2020). Usually, the positives are produced from data augmentation while the negatives are random samples from the batch or the whole dataset. Concerned with the negative sampling distribution, recent works (Robinson et al., 2021; Ge et al., 2021) further validate that using *hard negative samples*, i.e., the negative samples that are difficult to distinguish from the anchor can improve performance. Concurrently, supervised contrastive learning (SCL) has developed to extend the unsupervised batch contrastive approach to a *fully-supervised* setting that can leverage label information to select the positive and negative samples (Khosla et al., 2020; Gunel et al., 2021).

Motivated by the hard-negative sampling strategies in UCL and the value of label information in SCL, our proposed NaCL utilizes both label information to retain the clean positives and attack the noisy samples to move closer to the decision boundary as a kind of hard negative mining.

7 Conclusion

Building on the recent wave of learning without forgetting, in this paper, we demonstrate current continual learners are vulnerable under natural label shifts. Hence, we propose a novel noise-resistant contrastive learning framework NaCL to correct the false contrastive pairs brought by the co-existence of closed-set and open-set label noise. Comprehensive experiments and analyses validate that our method can achieve the *triple wins* that boost old knowledge, new task learning and noisy label robustness in one integrated algorithm.

Limitations

The problem of natural shifts in label space over streaming data exists in various domains and datasets. To validate the effectiveness of our method for a better comparison, we conduct comprehensive experiments on relation extraction. Therefore, it is intriguing to generalize our noise-resistant contrastive learning framework to other applications for more robust continual learners. On the other hand, our method directly lineages the step of memory replay from previous work for its certified performance. However, from the perspective of efficiency and online learning, to maintain the plasticity-stability trade-off without replaying is worth further refinement.

Ethics Statement

There is an ongoing trend of developing continual learners to adapt the streaming data without forgetting previously learned knowledge. We hope our work can encourage the community to consider a more generalized setting of continual learning for better robustness. Moreover, our noise-resistant contrastive learning framework provides insight into dealing with false contrastive pairs with better views of positives and hard negatives mining.

Acknowledgements

The authors wish to thank the anonymous reviewers for their helpful comments. This work was partially funded by National Natural Science Foundation of China (No.62376061,62206057,62076069), Shanghai Rising-Star Program (23QA1400200), Natural Science Foundation of Shanghai (23ZR1403500).

References

Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. 2019. Gradient based sample selection for online continual learning. In *NeurIPS*.

Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. 2017. A closer look at memorization in deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 233–242. PMLR.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

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

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

Songlin Dong, Xiaopeng Hong, Xiaoyu Tao, Xinyuan Chang, Xing Wei, and Yihong Gong. 2021. Few-shot class-incremental learning via relation knowledge distillation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(2):1255–1263.

Benjamin Ehret, Christian Henning, Maria Cervera, Alexander Meulemans, Johannes Von Oswald, and Benjamin F Grewe. 2021. Continual learning in recurrent neural networks. In *International Conference on Learning Representations*.

Songwei Ge, Shlok Mishra, Chun-Liang Li, Haohan Wang, and David Jacobs. 2021. Robust contrastive learning using negative samples with diminished semantics. In *Advances in Neural Information Processing Systems*, volume 34, pages 27356–27368. Curran Associates, Inc.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. In *International Conference on Learning Representations*.

Mustafa B Gurbuz and Constantine Dovrolis. 2022. NISPA: Neuro-inspired stability-plasticity adaptation for continual learning in sparse networks. In *Proceedings of the 39th International Conference on Machine*

Learning, volume 162 of *Proceedings of Machine Learning Research*, pages 8157–8174. PMLR.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. 2018a. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, pages 8535–8545.

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

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

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. Momentum contrast for unsupervised visual representation learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735.

Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. 2018. Using trusted data to train deep networks on labels corrupted by severe noise. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Ching-Yi Hung, Cheng-Hao Tu, Cheng-En Wu, Chien-Hung Chen, Yi-Ming Chan, and Chu-Song Chen. 2019. Compacting, picking and growing for unforgetting continual learning. In *Advances in Neural Information Processing Systems*, pages 13647–13657.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. 2018. MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2304–2313. PMLR.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Junnan Li, Richard Socher, and Steven C.H. Hoi. 2020a. Dividemix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*.

Yang Li, Guodong Long, Tao Shen, Tianyi Zhou, Lina Yao, Huan Huo, and Jing Jiang. 2020b. Self-attention enhanced selective gate with entity-aware embedding for distantly supervised relation extraction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8269–8276.

Yaoyao Liu, Anan Liu, Yuting Su, Bernt Schiele, and Qianru Sun. 2020. Mnemonics training: Multi-class incremental learning without forgetting. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12242–12251.

Noel Loo, Siddharth Swaroop, and Richard E Turner. 2021. Generalized variational continual learning. In *International Conference on Learning Representations*.

Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. 2020. Normalized loss functions for deep learning with noisy labels. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6543–6553. PMLR.

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

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. 2017. iCaRL: incremental classifier and representation learning. In *CVPR*.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. 2018. Learning to reweight examples for robust deep learning. In *ICML*.

Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*.

Fan-Keng Sun, Cheng-Hao Ho, and Hung-Yi Lee. 2020. {LAMAL}: {LA}nguage modeling is all you need for lifelong language learning. In *International Conference on Learning Representations*.

Bayu Distiawan Trisedya, Gerhard Weikum, Jianzhong Qi, and Rui Zhang. 2019. Neural relation extraction for knowledge base enrichment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 229–240, Florence, Italy. Association for Computational Linguistics.

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

Peiyi Wang, Yifan Song, Tianyu Liu, Binghuai Lin, Yunbo Cao, Sujian Li, and Zhifang Sui. 2022. Learning robust representations for continual relation extraction via adversarial class augmentation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. 2021. Learning with noisy labels revisited: A study using real-world human annotations. *Learning*.

Tongtong Wu, Xuekai Li, Yuan-Fang Li, Gholamreza Haffari, Guilin Qi, Yujin Zhu, and Guoqiang Xu. 2021. Curriculum-meta learning for order-robust continual relation extraction. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 10363–10369. AAAI Press.

Xiaobo Xia, Tongliang Liu, Bo Han, Mingming Gong, Jun Yu, Gang Niu, and Masashi Sugiyama. 2022. Sample selection with uncertainty of losses for learning with noisy labels. In *International Conference on Learning Representations*.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. Docred: A large-scale document-level relation extraction dataset. *arXiv preprint arXiv:1906.06127*.

Deming Ye, Yankai Lin, Peng Li, and Maosong Sun. 2022. Packed levitated marker for entity and relation extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4904–4917, Dublin, Ireland. Association for Computational Linguistics.

Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. 2019. How does disagreement help generalization against label corruption? In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7164–7173. PMLR.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017a. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017b. 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.

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

Tianyi Zhou, Shengjie Wang, and Jeff Bilmes. 2021. Robust curriculum learning: from clean label detection to noisy label self-correction. In *International Conference on Learning Representations*.

Wenxuan Zhou and Muhao Chen. 2021. Learning from noisy labels for entity-centric information extraction. *arXiv preprint arXiv:2104.08656*.

A Supplementary Explanation

A.1 Real-world Noise

Dataset	Noise Level
Clothing1M	38% (Wei et al., 2021)
Food-101N	20% (Wei et al., 2021)
NYT-10	35% (Li et al., 2020b)
TACRED	6.62% (Zhou and Chen, 2021)
CoNLL03	5.38% (Zhou and Chen, 2021)
Docred	41.4% (Yao et al., 2019)

Table 4: References for the noise level in Figure 1.

Notation	Meaning
f_M	Main Model
$\mathcal{E}_M$	Main Feature Extractor
Proj	Projector in Main Model
f_A	Auxiliary Model
$\mathcal{E}_A$	Auxiliary Feature Extractor
$\mathcal{F}_A$	Classifier in Auxiliary Model

Table 5: Model Components Notation.

B Training Algorithm

We present the whole training procedure for $\mathcal{T}^k$ in Algorithm 1.

Algorithm 1 Training procedure for $\mathcal{T}^k$

```
Receives: $\mathcal{D}^k_{\text{train}}$: contaminated training set of the k-th task, $f_M(\cdot, \theta)$: main model, $f_A(\cdot, \theta^*)$: auxiliary model, $\mathcal{B}$: memory buffer with exemplars stored
Require: learning rate $\eta$ for $f_M$ and $f_A$, batch size $m_s$, training epochs $E_1, E_2$, perturbation radius $\epsilon$, noise-guided attack step $s$
for epoch= $1, \cdots, E_2$ do                                   ▷ Selection
    Sample a batch $\{(x_i, y_i)\}_{i=1}^{m_s}$ from $\mathcal{D}^k_{\text{train}}$
    Training $f_A$ by Equation 1
end for
Obtain $\widetilde{\mathcal{D}}_{\text{clean}}$ and $\widetilde{\mathcal{D}}_{\text{noisy}}$ by Equation 2
for $(x_i, y_i) \in \widetilde{\mathcal{D}}_{\text{noisy}}$ do                          ▷ Attack
    $x'_i \leftarrow x_i + \delta$, where $\delta \sim$ Uniform$(-\epsilon, \epsilon)$
    for fixed step $s = 1, \cdots, S$ do
        Perform noise-guided attack by Equation 4
    end for
    Group $(x_i, y_i)$ with success attack to $\mathcal{D}_{\text{att-pos}}$ and $\mathcal{D}_{\text{neg}}$ otherwise
end for
for epoch= $1, \cdots, E_1$ do                                ▷ $\mathcal{T}^k$ Training
    Sample a batch $\{(x_i, y_i)\}_{i=1}^{m_s}$ from $\widetilde{\mathcal{D}}_{\text{clean}}$
    Contrastive training of $f_M$ by Equation 7
end for
if $\mathcal{T}^k$ is not the first task then                        ▷ Replay
    Update memory buffer $\mathcal{B}$ with exemplars selected from $\widetilde{\mathcal{D}}_{\text{clean}}$
    for epoch= $1, \cdots, E_1$ do
        Sample a batch $\{(x_i, y_i)\}_{i=1}^{m_s}$ from $\mathcal{B}$
        Training $f_M$ by Equation 8
    end for
end if
```

C Hyper-parameter Setup

All the hyper-parameters in our experiments for reproduction are shown in Table 6.

Parameter	Meaning	FewRel	TACRED
γ	selection threshold (Equation 2)	0.8,0.6,0.5 for $\{10\%, 30\%, 50\%\}$	0.9,0.75,0.6 for $\{10\%, 30\%, 50\%\}$
λ	trade-off for attack (Equation 4)	0.1	0.1
ϵ	perturbation size (Equation 4)	0.1	0.1
s	attack steps (Equation 4)	5	5
τ	temperature (Equation 7)	0.1,0.05,0.2 for $\{10\%, 30\%, 50\%\}$	
n	total task numbers	10	10
$\mathcal{C}$	classes of each incremental task	8	4
η	learning rate for f_M and f_A	1e-5	2e-5
m_s	training batch size	16	16
dim	projection dimension	64	64
E_1	training epoch of f_M for new relations	1	1
E_2	training epoch of f_A for selection	3	3

Table 6: List of hyper-parameters for our approach to reproduce the results in Table 1.