Contrastive Continual Learning with Importance Sampling and Prototype-Instance Relation Distillation

Jiyong Li[1, 2], Dilshod Azizov[3], Yang Li[4, 5], Shangsong Liang[1, 2, 3,*]
[1]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[2]Guangdong Key Laboratory of Big Data Analysis and Processing, Guangzhou, China
[3]Department of Machine Learning, Mohamed bin Zayed University of Artificial Intelligence, United Arab Emirates
[4]AI Thrust, Information Hub, The Hong Kong University of Science and Technology (Guangzhou), China
[5]Department of CSE, The Hong Kong University of Science and Technology, China
lijy373@mail2.sysu.edu.cn, dilshod.azizov@mbzuai.ac.ae, yli803@connect.hkust-gz.edu.cn,
liangshangsong@gmail.com

Abstract

Recently, because of the high-quality representations of contrastive learning methods, rehearsal-based contrastive continual learning has been proposed to explore how to continually learn transferable representation embeddings to avoid the catastrophic forgetting issue in traditional continual settings. Based on this framework, we propose ***C**ontrastive **C**ontinual **L**earning via **I**mportance **S**ampling (CCLIS)* to preserve knowledge by recovering previous data distributions with a new strategy for ***R**eplay **B**uffer **S**election (RBS)*, which minimize estimated variance to save hard negative samples for representation learning with high quality. Furthermore, we present the ***P**rototype-instance **R**elation **D**istillation (PRD)* loss, a technique designed to maintain the relationship between prototypes and sample representations using a self-distillation process. Experiments on standard continual learning benchmarks reveal that our method notably outperforms existing baselines in terms of knowledge preservation and thereby effectively counteracts catastrophic forgetting in online contexts. The code is available at https://github.com/lijy373/CCLIS.

Introduction

Deep neural networks are widely recognized to commonly face challenges in preserving performance in learned tasks after being trained on a new one, which is called Catastrophic Forgetting (McCloskey and Cohen 1989). Continual learning has been proposed to overcome Catastrophic Forgetting and transfer knowledge forward and backward during training for a stream of data, with only a small portion of the samples available at once. Recent research has integrated this scheme into continual learning due to the capability of contrastive learning to secure high-quality representations (Chen et al. 2020). A prime example of this integration is contrastive continual learning (Cha, Lee, and Shin 2021), which leverages the advantages of contrastive representations. Furthermore, we have identified two primary areas that remain underexplored in contrastive continual learning. First, the settings of continual learning inherently mean that only limited samples from previous tasks are retained and used alongside current tasks in contrastive continual models. This introduces a bias between contrastive representations trained in on-line and off-line settings, significantly contributing to Catastrophic Forgetting. Second, while the importance of hard negative samples (i.e., data points that are difficult to distinguish from anchors of different classes) for contrastive representation is recognized (Robinson et al. 2020), there has been limited investigation into the optimal selection and preservation of these samples in the context of contrastive continual learning, as observed in previous research (Cha, Lee, and Shin 2021). In this paper, we address two questions: (1) *How can we more effectively recapture the data distributions of earlier tasks using buffered samples?* and (2) *How can we improve the selection and retention of important samples?*

Consequently, we propose *CCLIS*, a contrastive continual learning algorithm that can alleviate Catastrophic Forgetting and select hard negative samples. We set out to make theoretical and empirical contributions as follows:

1. Based on a prototype-based supervised contrastive continual learning framework, we estimate previous task distributions via importance sampling with a weighted buffer during training, which can eliminate the bias between contrastive representations trained online and offline to overcome Catastrophic Forgetting.
2. Based on (1), we introduce a methodology to compute importance weights within a contrastive continual learning framework accompanied by robust theoretical justifications. In addition, we examine the interplay between importance sampling and hard negative selection. In particular, we observe that samples with higher scores, as determined by our method, are more likely to be retained as hard negative samples.
3. We propose the *PRD* loss to preserve the learned relationship between prototype and instance, which helps to use the importance sampling method to recover the distributions of tasks and maintain knowledge of previous tasks.
4. Finally, experiments demonstrate that our method outperforms most of the state-of-the-art non-stationary task distributions from three benchmark datasets. Empirically, our algorithms can recover the data distributions of pre-

*Corresponding author.

vious tasks as much as possible and store hard negative samples to enhance the performance of contrastive continual learning, which helps mitigate Catastrophic Forgetting.

Related Work

Contrastive Continual learning. Continual learning aims to extend the previous knowledge to new task adaptation and overcome Catastrophic Forgetting with restricted buffer and computing resources for the data stream, which has been widely used in Computer Vision, Reinforcement Learning etc (Wang et al. 2023). Continual learning algorithms are divided into three categories (De Lange et al. 2021): Replay methods (Riemer et al. 2018; Lopez-Paz and Ranzato 2017; Aljundi et al. 2019), Regularization-based methods (Li and Hoiem 2017; Zhang et al. 2021) and Parameter isolation methods (Rusu et al. 2016). Recently, replay methods have shown impressive performance, where the idea of extending contrastive learning (Jaiswal et al. 2020; Khosla et al. 2020) schemes to continual learning settings is noteworthy (Cha, Lee, and Shin 2021; Liang 2019, 2018; Liang et al. 2021; De Lange et al. 2021). However, due to the lack of previous task samples caused by continual learning, there is a gap in representations between continual contrastive learning and offline authentic contrastive learning. To fill this gap, recent work proposes continual contrastive learning (Fini et al. 2022) and contrastive continual learning (Cha, Lee, and Shin 2021). The first aims to explore the performance of contrastive learning in the continual learning setting, which is a different research goal from this paper. The latter adopts contrastive learning as a method to avoid suffering from Catastrophic Forgetting, which is the topic our work mainly focuses on. In this study, with the help of contrastive continual learning, we use the importance sampling method to recover the data distribution of previous tasks to alleviate Catastrophic Forgetting to some extent.

Coreset selection for Replay-based Continual Learning. Our algorithm focuses primarily on the replay method. Coreset selection for replay-based continual learning has been studied, and some buffer selection criteria have been proposed, e.g., herding-based exemplar selection (Rebuffi et al. 2017), gradient-based method (Aljundi et al. 2019), active learning (Ayub and Fendley 2022), minibatch similarity, sample diversity, and coreset affinity (Yoon et al. 2021) etc. Recent work (Tiwari et al. 2022) proposes weighted buffer selection, which can increase the capacity to preserve previous knowledge. Although coreset selection in continual learning and contrastive learning (Robinson et al. 2020) has been proposed, how to select and preserve important samples to help contrastive continual learning has not yet been studied. With contrastive continual learning, we apply the importance sampling method to weighted buffer, which assists mine hard negatives, estimate data distributions of previous tasks and overcome Catastrophic Forgetting issue.

Knowledge Distillation. As a vital method to transfer knowledge from teacher models to student models, knowledge distillation (Hinton, Vinyals, and Dean 2015) has been widely applied to continual learning in order to alleviate Catastrophic Forgetting by distilling past features into current models. Inspired by (Cha, Lee, and Shin 2021; Fang et al. 2021; Asadi et al. 2023), we propose to use knowledge distillation to keep the relation between prototypes and instances stable, which helps to apply the importance sampling strategy to better recover previous task distributions.

Background

Problem Setup: Continual Learning

We aim at supervised continual learning: given sequentially arriving tasks with timesteps $t \in \{1, 2, \ldots, T\}$, which are assumed to be sampled from a non-stationary distribution. In each task t, there are N_t input-label pairs drawn from data distributions of the specific task, i.e., $(x_i, y_i)_{i=1}^{N_t} \sim \mathcal{D}_t$, where x_i denotes the input and y_i denotes the responding class label. At each time step t, the model receives current task samples D_t separate from past task data $D_{1:t-1}$ due to the non-stationary distribution from which they are sampled.

We focus on two primary tasks in the setting of continual learning (Van de Ven and Tolias 2019): *Class-incremental learning (Class-IL) and Task-incremental learning (Task-IL).* For the first, given test samples, the model predicts the responding labels without task information during evaluation, that is, the goal is to obtain the model $\phi_\theta(x)$ by optimizing the following objective:

$$\min \sum_{t=1}^{T} \mathbb{E}_{(x,y)\sim D_t}[l(y, \phi_\theta(x))], \tag{1}$$

where l denotes the loss function such as the softmax function and θ denotes the model parameters. For the latter, the model is trained on sequentially arriving tasks with clear task boundaries and evaluated with task information:

$$\min \sum_{t=1}^{T} \mathbb{E}_{(x,y)\sim D_t}[l(y, \phi_\theta(x, t))]. \tag{2}$$

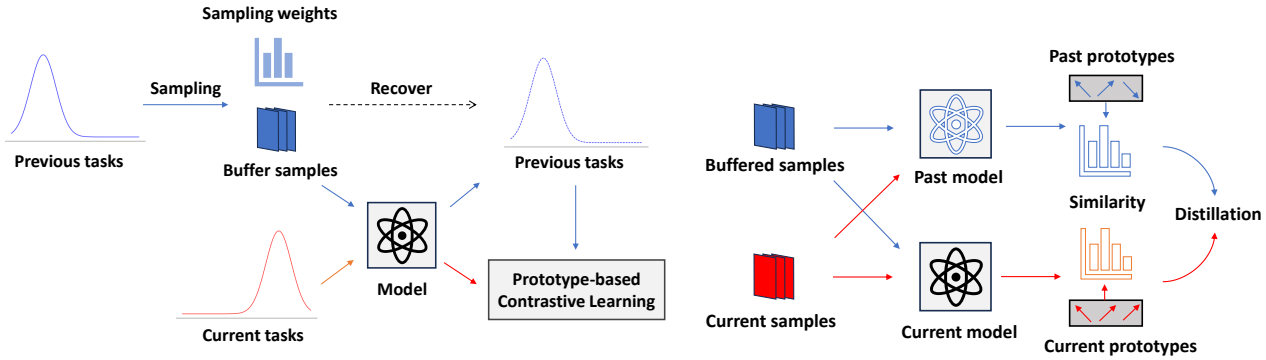
Preliminaries: Contrastive Learning

Given a training dataset $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$ of N samples, contrastive learning aims to obtain an embedding function f_θ mapping $\mathcal{X}$ to representation embeddings $\mathcal{Z} = \{z_1, z_2, \ldots, z_N\}$, i.e., $z_i = f_\theta(x_i)$, which can be applied to downstream tasks. Instance-wise contrastive learning (He et al. 2020) augments training samples into two views to have $2N$ inputs and gains high-quality embedding space by minimizing a contrastive loss function such as InfoNCE (Oord, Li, and Vinyals 2018):

$$L_{\text{InfoNCE}}(\theta) := \sum_{i=1}^{2N} -\log \frac{\exp(z_i \cdot z_i'/\tau)}{\sum_{j\neq i} \exp(z_i \cdot z_j/\tau)}, \tag{3}$$

where z_i and z_i' are positive embedding pairs of instance x_i with regards to the two views, while z_j includes one positive embedding z_i' and $2N-2$ negative embeddings of other instances, and τ denotes temperature.

To better apply our proposed method, we replace z_i with the prototype c of the responding class in equation (3) without augmenting samples into two views, which is different from the previous prototype-based contrastive learning framework (Li et al. 2020; Caron et al. 2020).



(a) Contrastive Learning via Importance Sampling (b) Prototype-instance Relation Distillation Loss

Figure 1: Illustration of Contrastive Learning via Importance Sampling and PRD Loss. (a) When new tasks are introduced, buffer samples are drawn with specific sampling weights. By using the Importance Sampling technique, we approximately recover the data distributions of previous tasks and apply prototype-based contrastive learning on previous and current data to have high-quality features. (b) Given samples of a minibatch, the PRD loss is designed to distill the relation between prototypes and instances from the previous model to the current one. We minimize the cross-entropy of prototype-instance similarity from the current and previous models with frozen parameters, which are computed with dot products of normalized embeddings.

Preliminaries: Importance Sampling

As a significant method in machine learning, importance sampling (Kong, Liu, and Wong 1994) has been proposed to approximate expectations without drawing samples from target distributions. Formally, given the data distribution π of sample z and the proposal distribution q, we can estimate the expectation of function $f(z)$ on π using the classical Monte Carlo Importance Sampling method:

$$\mathbb{E}_{z\sim\pi} f(z) = \mathbb{E}_{z\sim q} \frac{\pi(z)}{q(z)} f(z) \approx \frac{1}{L} \sum_{l=1}^{L} \frac{\pi(z_l)}{q(z_l)} f(z_l), \quad (4)$$

where L is the number of drawn samples. However, it is often the case that assuming $\pi = \hat{\pi}/Z_\pi$, the distribution π can only be evaluated up to an incremental normalization constant Z_π. By estimating Z_π, the biased importance sampling strategy provides the following approximation of the expectation:

$$\mathbb{E}_{z\sim\pi} f(z) \approx \frac{1}{L} \sum_{l=1}^{L} \frac{\hat{\pi}(z_l)/q(z_l)}{\hat{Z}_\pi} f(z_l), \quad (5)$$

where $\hat{Z}_\pi := \frac{1}{L}\sum_{l=1}^{L} \hat{\pi}(z_l)/q(z_l)$ is the estimator of Z_π.

Contrastive Continual Learning via Importance Sampling

Overview of Our Model

Based on previous contrastive continual learning (Cha, Lee, and Shin 2021), we propose *CCLIS* in order to overcome Catastrophic Forgetting. We propose a model structured in three distinct steps: First, we build the supervised version of prototype-based InfoNCE via importance sampling in the setting of the continual learning framework. Second, based on our modified version of contrastive loss, we propose a sampling method for *RBS* with sufficient theoretical guarantees. Finally, we introduce a distillation loss in our model. By doing so, our model boasts several significant advantages: (1) With importance sampling under the prototype-based InfoNCE, we can recover distributions of previous tasks to overcome Catastrophic Forgetting. (2) With our *RBS*, we can draw hard negative samples to preserve and eliminate the estimated variation in importance sampling. (3) By using the *PRD* loss, we can maintain the relationship between prototypes and instances, improving the performance of the importance sampling method. In addition, it helps to distill and preserve knowledge from previous tasks. The overview of the proposed model is presented in Algorithm 1.

Prototype-based InfoNCE Loss via Importance Sampling

In this section, our aim is to adopt a supervised version of the prototype-based InfoNCE to learn high-quality contrastive representation in a continual learning setting. However, due to the gap between offline and online scenarios, only a few samples of previous tasks are preserved in the replay buffer M, making it infeasible to optimize the loss function. To fill this gap, we aim to keep the performance of our model close to that trained on data drawn from the current task D_t and the samples available at the moment $t-1$, i.e., $R_{t-1} := M_{t-1} \cup D_{t-1}$, since it is difficult to achieve performance learning samples from all tasks $D_{1:t}$. Formally, for each task t, $\hat{C}_t$ and Y_t are defined as the class set of R_t and D_t respectively. Assuming that dataset $\mathcal{D}$ has N samples with K classes, we denote prototypes $\{c_1, \ldots, c_K\}$ as clustering centers of classes that are randomly initialized. By denoting $s_{ij} := c_i \cdot z_j/\tau$ for convenience, we focus on the following objective:

$$L := \sum_{i\sim\hat{C}_{t-1}\cup Y_t} \sum_{j\sim S_i} -\log \frac{\exp(s_{ij})}{\sum_{k\sim R_{t-1}\cup D_t} \exp(s_{ik})} \quad (6)$$

where $S_i := \{j | y_j = i \text{ and } (x_j, y_j) \in R_{t-1} \cup D_t\}$ denotes samples from specific class i in currently available data. To analyze the contrastive loss, we deviate from it with respect to the model parameters and obtain the gradients of this loss function. For a specific prototype c_i of the previous task, we denote $p_{ij} = \frac{\exp(s_{ij})}{\sum_{k \sim R_{t-1} \cup D_t} \exp(s_{ik})}$, then the gradient of the loss function (6) has the following form if we focus on the sample j drawn from S_i:

$$\nabla_\theta L_{i,j} = -\nabla_\theta s_{ij} + \sum_{k \sim R_{t-1} \cup D_t} p_{ik} \nabla_\theta s_{ik} \tag{7}$$

where $L_{i,j} = -\log \frac{\exp(s_{ij})}{\sum_{k \sim R_{t-1} \cup D_t} \exp(s_{ik})}$. Notice that the gradient $\nabla_\theta L_{i,j}$ can be decomposed into two parts, including $-\nabla_\theta s_{ij}$ and $\sum_{k \sim R_{t-1} \cup D_t} p_{ik} \nabla_\theta s_{ik}$, while the latter part can be written as the expectation of the gradients $\mathbb{E}_{k \sim p_i} \nabla_\theta s_{ik}$. However, it is difficult to apply the *classical Monte-Carlo* method to estimate expectations on various target distributions p_i. It is also difficult to directly adopt the *classical Importance Sampling* method, as the partition function $\sum_{k \sim R_{t-1} \cup D_t} \exp(s_{ik})$ is intractable due to previous lost samples. To address the issues, we use the biased importance sampling method to approximate the intractable normalizing function and give a biased estimation of the gradients, while positive samples of specific class i and samples from task t cannot be estimated. Unlike the traditional sampling strategy, each instance (x_j, y_j) is sampled from the discrete distribution of the specific class instead of the entire data distribution, avoiding the mixing of positive and negative samples. Formally, we denote $g^{(m)} = [g_j^{(m)}]_{j \sim S_m}$ as the proposal distribution of the specific class m, and J_m represents the samples drawn from $g^{(m)}$. Accordingly, we get the biased estimate of the stochastic gradient as below using biased importance sampling and Monte-Carlo techniques:

$$\begin{aligned} \nabla_\theta L_{i,j} \approx & -\nabla_\theta s_{ij} + \sum_{k \sim J_i \cup D_t} \frac{p_{ik}}{W_i} \nabla_\theta s_{ik} \\ & + \sum_{m \sim \hat{C}_{t-1} \setminus i} \frac{1}{|J_m|} \sum_{k \sim J_m} \frac{\omega_{ik}^{(m)}}{W_i} \nabla_\theta s_{ik}, \end{aligned} \tag{8}$$

where $\omega_{ik}^{(m)} := p_{ik}^{(m)} / g_k^{(m)}$ and $W_i := \sum_{k \sim J_i \cup D_t} p_{ik} + \sum_{m \sim \hat{C}_{t-1} \setminus i} \frac{1}{|J_m|} \sum_{k \sim J_m} \omega_{ik}^{(m)}$. Finally, our modified version of contrastive loss can be obtained as the sum of the antiderivative of the gradient in the specific task t:

$$L_{\text{Sample-NCE}}(\theta; t) = \sum_{i \sim Y_t} \sum_{j \sim S_i} L_{i,j} + \sum_{i \sim \hat{C}_t} \sum_{j \sim J_i} \hat{L}_{i,j}, \tag{9}$$

where:

$$\hat{L}_{i,j} = -\log \frac{\exp(s_{ij})}{\sum\limits_{k \sim J_i \cup D_t} \exp(s_{ik}) + \sum\limits_{m \sim \hat{C}_{t-1} \setminus i} \sum\limits_{k \sim J_m} \frac{\exp(s_{ik})}{g_k^{(m)} |J_m|}}. \tag{10}$$

The loss function can be applied directly to samples drawn from minibatches. Now, we propose a method to recover data distributions from previous tasks using weighted samples in the replay buffer. For more in-depth technical details, please refer to Appendix A.

Algorithm 1: ***C**ontrastive **C**ontinual **L**earning via **I**mportance **S**ampling (CCLIS).*

Require: Dataset $\{D_t\}_{t=1}^T$, model f_θ, buffer $M \leftarrow \{\}$, hyper-parameter λ, learning rate η,

```
Initialize model parameters θ,
for t=1,...,T do
    for batch B_t ~ D_t do
        B_M ~ M
        B ← B_t ∪ B_M
        L ← L_{Sample-NCE}(θ; B) with Eq.(9)
        if t > 1 then
            L ← L + λ · L_{PRD}(θ; θ_{prev}, B) with Eq.(14)
        end if
        θ ← θ − η∇_θ L
    end for
    Calculate the proposal distribution g with Eq.(12)
    M ← SELECT(M ∪ D_t, g) by weighted sampling without replacement
    θ_{prev} ← θ
end for
```

Replay Buffer Selection for Estimated Variance Minimization

While using the modified contrastive learning objective can help recover the previous data distributions, one may still benefit from selecting a suitable proposal distribution, which is crucial to the effectiveness of the importance sampling method. The closer the proposal distribution to the target distribution, the better the model performance, and the gap between the proposal distribution and the target distribution would cause the estimated variance, leading to a drop in performance. However, it is not trivial to select a proposal distribution $g^{(m)}$ of a specific class m in $\hat{C}_t$, because there are $n_t := |\hat{C}_t| - 1$ target distributions, except the distribution p_m to which the proposal distribution must be close. To overcome this, we propose ***RBS*** based on the estimated minimization of variance, which can better recover the previous data distributions.

For the set of a specific class m sampled from R_t, the target distribution with the fixed prototype c_i is p_i. Although samples from D_t should be considered, we can approximately minimize the estimated variance caused by the gap between the proposal distribution $g^{(m)}$ and the target distributions $\hat{p}_i^{(m)} := [\frac{\exp s_{ij}}{\sum_{j \sim R_{t-1}} \exp s_{ij}}]_{j \sim S_m}$ for all $i \in \hat{C}_t \setminus m$. We prove in Appendix A that we can minimize an upper bound on the estimated variance by optimizing the mean of Kullback–Leibler (KL) divergences of the proposal distribution from the target distributions:

$$\min \frac{1}{n_t} \sum_{i=1}^{n_t} KL(\hat{p}_i^{(m)} || g^{(m)}). \tag{11}$$

It is noticed that the above objective can be solved directly. The function achieves the minimum when the proposal distribution is equal to the mean of the target distributions:

$$g^{(m)} = \frac{1}{n_t} \sum_{i=1}^{n_t} \hat{p}_i^{(m)}, \tag{12}$$

As inferred from Equation (12), in order to reduce the estimated variance, we need to select negative samples that are close to the prototypes of all different categories on average, that is, hard negative samples that are difficult to distinguish by a classifier. More details can be seen in Appendix A.

PRD for Contrastive Continual Learning

Although our method readily provides a more transferable representation space, it is necessary to preserve the learned knowledge of previous tasks for stable representation features and prototype-instance relationships, which is crucial to limit the gap between the proposal distribution and the target distributions. To achieve this aim, we propose a *PRD* to regulate changes in the features of current and previous models through self-distillation (Fang et al. 2021), which can preserve learned knowledge to overcome Catastrophic Forgetting. Assuming that the samples are drawn from the minibatch B at the time step t and $Y_{1:t}$ is the class set of $D_{1:t}$, the prototype-instance score vector is the normalized similarity of the sample x_i to the prototypes that appear, which is formally defined as $q(x_j;\theta) = [q_{ij}]_{i\sim Y_{1:t}}$, where q_{ij} denotes the normalized prototype-instance similarity between prototype i and sample j:

$$q_{ij} = \frac{\exp(s_{ij})}{\sum_{i=1}^{|Y_{1:t}|}\exp(s_{ij})}. \tag{13}$$

Our proposed *PRD* loss measures the relation of prototype-instance similarity between the previous and current representation spaces as a self-distillation method. Formally, we denote the parameters of previous and current models as θ_{prev} and θ_{cur} respectively, and the *PRD* loss is defined as:

$$L_{\mathrm{PRD}}(\theta_{cur};\theta_{prev},B) = \sum_{x_j\sim B} -q(x_j;\theta_{prev})\log q(x_j;\theta_{cur}). \tag{14}$$

By using the frozen model with previous representations, *PRD* distill the learned relationship between prototypes and sample features to the current model, helping to keep the target distributions stable to better apply the importance sampling method and preserve previous knowledge better.

Objective Function

Similar to past literature(Cha, Lee, and Shin 2021), our loss function is composed of contrastive loss $L_{\mathrm{Sample-NCE}}$ and distillation loss L_{PRD} with the trade-off parameter λ as following:

$$L(\theta) = L_{\mathrm{Sample-NCE}}(\theta) + \lambda \cdot L_{\mathrm{PRD}}(\theta), \tag{15}$$

and all the trainable parameters, including prototypes and model parameters, can be updated through optimizing the objective function.

Experimental Setup

Research Questions. The remainder of the paper is guided by subsequent research questions: (**RQ1**) Can our proposed method outperform other baselines in various datasets in the continual learning setting? (**RQ2**) What is the impact of each component, i.e., importance sampling and *PRD*, on the performance of our method? (**RQ3**) What is the connection between the components in our method?

Baselines. We consider two continua learning settings: Class-Incremental Learning (Class-IL) and Task-Incremental Learning (Task-IL). For evaluation purposes, we take the following state-of-the-art baselines: (1) ER (Experience Replay) (Riemer et al. 2018) is a rehearsal-based method with random sampling in memory retrieval and reservoir sampling in memory updates. (2) iCarL (Rebuffi et al. 2017), representation learning with an incremental buffer proposes learning representation embeddings and classifiers in the Class-IL setting. (3) GEM (Lopez-Paz and Ranzato 2017) uses episodic memory to minimize negative knowledge transfer and Catastrophic Forgetting. (4) GSS (Aljundi et al. 2019), a gradient-based sample selection method that maximizes the variation of the gradients of the replay buffer samples. (5) DER (Dark Experience Replay) and DER++ (Buzzega et al. 2020), which promotes the consistency of the logit of the current task with the previous one. (6) Co2L (Contrastive Continual Learning) (Cha, Lee, and Shin 2021), a rehearsal-based continual learning algorithm with instance-wise contrastive loss and self-distillation. (7) GCR (Gradient Coreset-based Replay) (Tiwari et al. 2022), a new gradient-based strategy for *RBS* in continual learning.

Datasets. We verify the effectiveness of our method in Class-IL and Task-IL on the following three datasets: Seq-cifar-10 (Krizhevsky, Hinton et al. 2009), Seq-cifar-100 (Krizhevsky, Hinton et al. 2009), and Seq-tiny-imagenet (Le and Yang 2015), and all of them are commonly used as benchmarks in previous work(Tiwari et al. 2022). Seq-cifar-10 is the set of splits of Cifar10. Following general settings, we divide it into 5 tasks with two classes per task. Seq-cifar-100 is constructed from Cifar100 and divided into five tasks with 20 classes per task, similar to (Tiwari et al. 2022). Seq-tiny-imagenet splits Tiny-imagenet (Russakovsky et al. 2015) into ten tasks of 20 classes each. Both Class-IL and Task-IL share the same dataset splitting settings.

Settings. We train on three datasets using ResNet-18 (He et al. 2016) as the backbone and report our results with memory sizes 200 and 500. All model parameters, including prototypes and backbone parameters, are trained through backpropagation. To be consistent with previous contrastive continual learning, we freeze model parameters and train a linear classifier to evaluate on samples R_T. We adopt the same strategy of training the linear classifier as in (Cha, Lee, and Shin 2021) to avoid suffering from the class-imbalanced issue. *Accuracy* and *Average Forgetting* will be adopted to evaluate all the methods in our experiments following the CL literature (Chaudhry et al. 2018; Tiwari et al. 2022). Data preparation, model architecture, hyper-parameter selection, and training details, etc., can be referred to Appendix B.

Results and Discussions

In this section, we conduct sufficient experiments to answer the research questions. Additional comparisons, ablations, and analysis are shown in Appendix C.

Buffer	Dataset	Seq-Cifar-10		Seq-Cifar-100		Seq-Tiny-ImageNet	
	Scenario	Class-IL	Task-IL	Class-IL	Task-IL	Class-IL	Task-IL
200	ER	49.16±2.08	91.92±1.01	21.78±0.48	60.19±1.01	8.65±0.16	38.83±1.15
	iCaRL	32.44±0.93	74.59±1.24	28.0±0.91	51.43±1.47	5.5±0.52	22.89±1.83
	GEM	29.99±3.92	88.67±1.76	20.75±0.66	58.84±1.00	-	-
	GSS	38.62±3.59	90.0±1.58	19.42±0.29	55.38±1.34	8.57±0.13	31.77±1.34
	DER	63.69±2.35	91.91±0.51	31.23±1.38	63.09±1.09	13.22±0.92	42.27±0.90
	Co2L	65.57±1.37	93.43±0.78	27.73±0.54	54.33±0.36	13.88±0.40	42.37±0.74
	GCR	64.84±1.63	90.8±1.05	33.69±1.40	64.24±0.83	13.05±0.91	42.11±1.01
	CCLIS(Ours)	**74.95±0.61**	**96.20±0.26**	**42.39±0.37**	**72.93±0.46**	**16.13±0.19**	**48.29±0.78**
500	ER	62.03±1.70	93.82±0.41	27.66±0.61	66.23±1.52	10.05±0.28	47.86±0.87
	iCaRL	34.95±1.23	75.63±1.42	33.25±1.25	58.16±1.76	11.0±0.55	35.86±1.07
	GEM	29.45±5.64	92.33±0.80	25.54±0.65	66.31±0.86	-	-
	GSS	48.97±3.25	48.97±3.25	21.92±0.34	60.28±1.18	9.63±0.14	36.52±0.91
	DER	72.15±1.31	93.96±0.37	41.36±1.76	71.73±0.74	19.05±1.32	53.32±0.92
	Co2L	74.26±0.77	95.90±0.26	36.39±0.31	61.97±0.42	20.12±0.42	53.04±0.69
	GCR	74.69±0.80	94.44±0.32	45.91±1.30	71.64±2.10	19.66±0.68	52.99±0.89
	CCLIS(Ours)	**78.57±0.25**	**96.18±0.43**	**46.08±0.67**	**74.51±0.38**	**22.88±0.40**	**57.04±0.43**

Table 1: Class-IL and Task-IL Continual Learning. We report our performance and the results of rehearsal-based baselines on Seq-Cifar-10, Seq-Cifar-100 and Seq-Tiny-ImageNet with memory sizes 200 and 500, all of which are averaged across ten independent trails.

Buffer	Dataset	Seq-Cifar-10		Seq-Cifar-100		Seq-Tiny-ImageNet	
	Scenario	Class-IL	Task-IL	Class-IL	Task-IL	Class-IL	Task-IL
200	DER	35.79±2.59	6.08±0.70	62.72±2.69	25.98±1.55	64.83±1.48	40.43±1.05
	Co2L	36.35±1.16	6.71±0.35	67.06±0.01	37.61±0.11	73.25±0.21	47.11±1.04
	GCR	32.75±2.67	7.38±1.02	57.65±2.48	24.12±1.17	65.29±1.73	40.36±1.08
	CCLIS(Ours)	**22.59±0.18**	**2.08±0.27**	**46.89±0.59**	**14.17±0.20**	**62.21±0.34**	**33.20±0.75**
500	DER	24.02±1.63	3.72±0.55	49.07±2.54	25.98±1.55	59.95±2.31	28.21±0.97
	Co2L	25.33±0.99	3.41±0.8	51.96±0.80	26.89±0.45	65.15±0.26	39.22±0.69
	GCR	19.27±1.48	3.14±0.36	**39.20±2.84**	15.07±1.88	56.40±1.08	27.88±1.19
	CCLIS(Ours)	**18.93±0.61**	**1.69±0.12**	42.53±0.64	**12.68±1.33**	**50.15±0.20**	**23.46±0.93**

Table 2: Average Forgetting (lower is better) in Continual Learning, all averaged across five independent trails. For simplicity, we only compare our method with recent baselines.

RQ1:Performance on Squentially Arriving Tasks

To be consistent with most contrastive continual learning algorithms, we explore whether our method can learn high-quality embeddings to overcome Catastrophic Forgetting. To achieve this, we freeze the backbone parameters and train a linear classifier from scratch to verify the effectiveness of our algorithm. The results in Table 1 show that our method outperforms most state-of-the-art baselines on various datasets in non-stationary task distributions, and the average forgetting results in Table 2 show that our method can effectively alleviate the forgetting issue.

RQ2: Ablation Study

Effectiveness of importance sampling. To validate the effectiveness of the importance sampling method, we introduce three variants of *CCLIS:* Without importance sampling (*IS*) and *PRD*, where we train only with our prototype-based contrastive loss with the random sampling strategy; with importance sampling (*IS*) only, where we optimize the Sample-NCE without self-distill loss; With *PRD* only, where we preserve samples randomly and only use distillation loss to preserve previous knowledge. We compare our methods with the three variants in Seq-Cifar-10 with 200 buffered samples and show the results in Table 3 in the Class-IL scenario. It is noticed that without *PRD* loss, importance sampling brings about an improvement of 3.4% while we gain a 1.4% improvement with *PRD* loss on the average.

Furthermore, to illustrate whether *RBS* or the entire importance sampling method leads to improvement, we perform ablation experiments on Seq-Cifar-10 under the Class-IL and Task-IL scenarios. Table 4 illustrates that there is a gap between the data distributions of previous tasks and the preserved samples only with *RBS*, which leads to performance degradation; our method via importance sampling can recover the distributions of previous tasks by eliminating bias caused by sample selection, resulting in better performance than random sampling.

Effectiveness of *PRD* loss. Similar to the importance sampling ablation study, we verify the effectiveness of *PRD* loss in Table 3. According to the table, the improvement brought by *PRD* loss is 31.0% with random sampling, while there is a 28.4% relative improvement related to *PRD* loss with importance sampling. The increased performance growth shows that the *PRD* loss can preserve knowledge from previous tasks to alleviate Catastrophic Forgetting.

	Sampling	PRD	200
w/o IS and PRD	Random	×	56.43±1.54
w/ IS only	IS	×	58.37± 1.17
w/ PRD only	Random	✓	73.95±1.12
CCLIS(ours)	IS	✓	**74.95±0.61**

Table 3: Ablation study of importance sampling and *PRD*. We train our model on the Seq-CIFAR-10 dataset with 200 buffered samples under a Class-IL scenario to explore the effectiveness of importance sampling and *PRD*.

		200	500
Class-IL	RBS only	54.18±0.88	67.50±0.50
	IS	**57.24±1.46**	**68.50±0.76**
Task-IL	RBS only	85.00±1.20	91.19±0.30
	IS	**86.28±0.74**	**91.34±0.20**

Table 4: Ablation study of the importance sampling method. One experiment trains only with *RBS*, while the other trains with importance sampling. All results are given in Seq-CIFAR-10 under Class-IL and Task-IL scenarios. The improvements indicate that importance sampling can recover previous tasks data distributions, eliminating the bias caused by sample selection.

RQ3: Analysis of the Relationship between Components

Connections between importance sampling and *PRD* loss. As shown in experiments, the results validate that importance sampling and *PRD* loss can preserve the knowledge of previous tasks. We also find that *PRD* brought a large performance boost to the importance sampling method in the ablation study. The huge performance improvement indicates that *PRD* narrows the gap between the proposal distribution and the target distributions by keeping the relation between prototypes and instances stable, reducing the estimated variance in the sampling of importance.

To further analyze the joint impact of importance sampling and *PRD* loss on performance, we train the model by gradually tuning the trade-off parameter λ. As in Figure 2, the test accuracy gradually increases and stays steady after reaching the highest point with increasing parameters, indicating that *PRD* complements the importance sampling based contrastive loss by stabilizing the relationship between prototypes and instances in online settings.

Connections between importance sampling and hard negative mining. By re-examining *RBS*, we find that minimizing the estimated variance of importance sampling is equivalent to preserving hard negatives relative to prototypes of different classes. We have made an intriguing discovery and visualized the embedding space on Seq-Cifar-10 to compare importance sampling with random sampling. The visualization in Figure 3 indicates that our method draws the samples distributed at the edges of the clusters and learns better high-quality representations than random sampling.

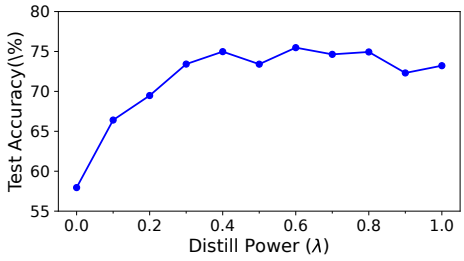


Figure 2: Performance variant with the distill power λ on Seq-CIFAR-10 under Class-IL scenario. *PRD* effectively enhances the performance of the importance sampling based contrastive learning by successfully maintaining the prototype-instance relationship.

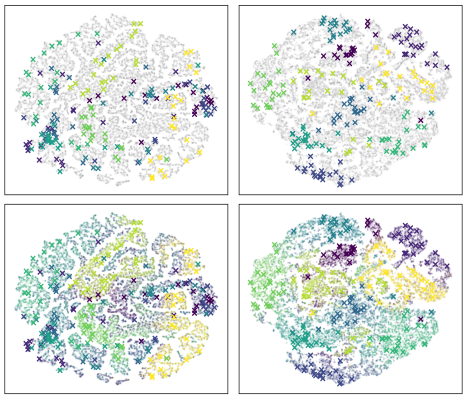


Figure 3: Top: t-SNE visualization of feature embeddings from replay buffer (colored) and all (gray) training samples of Seq-Cifar-10. Bottom: Similar to Top, but all samples are colored to distinguish different clusters clearly. Left: buffer features drawn by Co2L, a contrastive continual learning algorithm with random sampling, are spread uniformly in clusters. Right: Buffer features sampled by *CCLIS* are mainly distributed at the edge of clusters. These can be viewed as hard negatives of other classes to help the model learn high-quality contrastive representations.

Conclusion, Limitations and Future Work

In this paper, based on the previous contrastive continual learning method, we introduce an improved contrastive loss through importance sampling, yielding improved contrastive representations online. Furthermore, our *RBS* method, informed by importance weight, retains hard negative samples for future learning. In addition, we propose *PRD* to maintain the prototype-instance relation. Experiments show that our method can better preserve previous task knowledge to overcome Catastrophic Forgetting.

However, a limitation comes from the recovery of data distributions since we only recover the distributions of the last visible samples, which is still far from recovering the entire data distribution. We plan to develop algorithms to address this challenge as our future work.

References

Aljundi, R.; Lin, M.; Goujaud, B.; and Bengio, Y. 2019. Gradient based sample selection for online continual learning. *Advances in neural information processing systems*, 32.

Asadi, N.; Davari, M.; Mudur, S.; Aljundi, R.; and Belilovsky, E. 2023. Prototype-sample relation distillation: towards replay-free continual learning. In *International Conference on Machine Learning*, 1093–1106. PMLR.

Ayub, A.; and Fendley, C. 2022. Few-Shot Continual Active Learning by a Robot. *Advances in Neural Information Processing Systems*, 35: 30612–30624.

Buzzega, P.; Boschini, M.; Porrello, A.; Abati, D.; and Calderara, S. 2020. Dark experience for general continual learning: a strong, simple baseline. *Advances in neural information processing systems*, 33: 15920–15930.

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33: 9912–9924.

Cha, H.; Lee, J.; and Shin, J. 2021. Co2l: Contrastive continual learning. In *Proceedings of the IEEE/CVF International conference on computer vision*, 9516–9525.

Chaudhry, A.; Dokania, P. K.; Ajanthan, T.; and Torr, P. H. 2018. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European conference on computer vision (ECCV)*, 532–547.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

De Lange, M.; Aljundi, R.; Masana, M.; Parisot, S.; Jia, X.; Leonardis, A.; Slabaugh, G.; and Tuytelaars, T. 2021. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7): 3366–3385.

Fang, Z.; Wang, J.; Wang, L.; Zhang, L.; Yang, Y.; and Liu, Z. 2021. Seed: Self-supervised distillation for visual representation. *arXiv preprint arXiv:2101.04731*.

Fini, E.; Da Costa, V. G. T.; Alameda-Pineda, X.; Ricci, E.; Alahari, K.; and Mairal, J. 2022. Self-supervised models are continual learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9621–9630.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Jaiswal, A.; Babu, A. R.; Zadeh, M. Z.; Banerjee, D.; and Makedon, F. 2020. A survey on contrastive self-supervised learning. *Technologies*, 9(1): 2.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33: 18661–18673.

Kong, A.; Liu, J. S.; and Wong, W. H. 1994. Sequential imputations and Bayesian missing data problems. *Journal of the American statistical association*, 89(425): 278–288.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Le, Y.; and Yang, X. 2015. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7): 3.

Li, J.; Zhou, P.; Xiong, C.; and Hoi, S. C. 2020. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*.

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

Liang, S. 2018. Dynamic user profiling for streams of short texts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Liang, S. 2019. Collaborative, dynamic and diversified user profiling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 4269–4276.

Liang, S.; Tang, S.; Meng, Z.; and Zhang, Q. 2021. Cross-temporal snapshot alignment for dynamic networks. *IEEE Transactions on Knowledge and Data Engineering*.

Lopez-Paz, D.; and Ranzato, M. 2017. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30.

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

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

Rebuffi, S.-A.; Kolesnikov, A.; Sperl, G.; and Lampert, C. H. 2017. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2001–2010.

Riemer, M.; Cases, I.; Ajemian, R.; Liu, M.; Rish, I.; Tu, Y.; and Tesauro, G. 2018. Learning to learn without forgetting by maximizing transfer and minimizing interference. *arXiv preprint arXiv:1810.11910*.

Robinson, J.; Chuang, C.-Y.; Sra, S.; and Jegelka, S. 2020. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252.

Rusu, A. A.; Rabinowitz, N. C.; Desjardins, G.; Soyer, H.; Kirkpatrick, J.; Kavukcuoglu, K.; Pascanu, R.; and Hadsell, R. 2016. Progressive neural networks. *arXiv preprint arXiv:1606.04671*.

Tiwari, R.; Killamsetty, K.; Iyer, R.; and Shenoy, P. 2022. Gcr: Gradient coreset based replay buffer selection for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 99–108.

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

Wang, L.; Zhang, X.; Su, H.; and Zhu, J. 2023. A comprehensive survey of continual learning: Theory, method and application. *arXiv preprint arXiv:2302.00487*.

Yoon, J.; Madaan, D.; Yang, E.; and Hwang, S. J. 2021. Online coreset selection for rehearsal-based continual learning. *arXiv preprint arXiv:2106.01085*.

Zhang, Q.; Fang, J.; Meng, Z.; Liang, S.; and Yilmaz, E. 2021. Variational continual Bayesian meta-learning. *Advances in Neural Information Processing Systems*, 34: 24556–24568.