Learning from Students: Online Contrastive Distillation Network for General Continual Learning

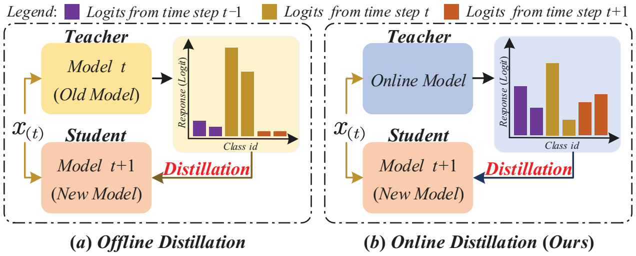
Jin Li[1] , **Zhong Ji**[1*] , **Gang Wang**[1,2] , **Qiang Wang**[1] , **Feng Gao**[2]
[1]School of Electrical and Information Engineering, Tianjin University
[2]CETC Key Laboratory of Aerospace Information Applications
{lijincm, jizhong, wanggg, qiangwang306}@tju.edu.cn

Abstract

The goal of General Continual Learning (GCL) is to preserve learned knowledge and learn new knowledge with constant memory from an infinite data stream where task boundaries are blurry. Distilling the model's response of reserved samples between the old and the new models is an effective way to achieve promising performance on GCL. However, it accumulates the inherent old model's response bias and is not robust to model changes. To this end, we propose an Online Contrastive Distillation Network (OCD-Net) to tackle these problems, which explores the merit of the student model in each time step to guide the training process of the teacher model. Concretely, the teacher model is devised to help the student model to consolidate the learned knowledge, which is trained online via integrating the parameters of the student model to accumulate the new knowledge. Moreover, our OCD-Net incorporates both relation and adaptive response to help the student model alleviate the catastrophic forgetting, which is also beneficial for the teacher model to preserve the learned knowledge. Extensive experiments on six benchmark datasets demonstrate that our OCD-Net significantly outperforms state-of-the-art approaches in $3.03\% \sim 8.71\%$ with various buffer sizes. Our code is available at https://github.com/lijincm/OCD-Net.

1 Introduction

Human beings have the gift of quickly learning new knowledge on the basis of learned knowledge without interfering the stability of learned knowledge, but it remains challenging in current machine learning technology. On the one hand, the new task will seriously interfere with the old task's performance when the neural network is trained directly on the new task, which is the notorious catastrophic forgetting problem [McCloskey and Cohen, 1989]. On the other hand, training a neural network on both old and new tasks obviously consumes more resources. Continual learning, also known as incremental learning and lifelong learning, has been developed for addressing these problems. It focuses on quickly learning new tasks from a non-stationary data stream and overcoming the catastrophic forgetting problem [Delange *et al.*, 2021].



(*a*) *Offline Distillation* **(*b*) *Online Distillation* (*Ours*)**

Figure 1: The illustration of utilizing knowledge distillation in GCL. (a) Diagram of the popular offline distillation approach. At time step $t + 1$, the teacher model's response is biased to the classes from the time step t when the reserved sample $x_{(t)}$ from time step t is input into the teacher model. (b) Diagram of the proposed online distillation approach. The teacher model is trained simultaneously by integrating the student model's parameters, which is beneficial for obtaining less biased responses to guide the training process of the student model. Note that there are responses for time step $t + 1$ since it is essential to initialize the model by providing ample class classification heads to avoid aligning outputs of different length.

Currently, lots of studies simplify the continual learning problem by providing the additional prior knowledge, such as assuming that tasks don't overlap each other and applying task boundaries during training stage [Tao *et al.*, 2020], [Liu *et al.*, 2021], [Zhao *et al.*, 2021], [Quang *et al.*, 2021]. However, these prior knowledge don't always exist in the real-world scenarios where tasks are complex and diverse. Therefore, General Continual Learning (GCL) has been proposed to meet the requirements of real-world scenarios. It emphasizes that the training and testing processes of continual learning don't rely on the task boundaries and the memory size is bounded. There are mainly two lines of approaches to tackle this issue: Sample strategy-based approaches [Rahaf *et al.*, 2019], [Ji *et al.*, 2021b] and knowledge transfer-based approaches [Buzzega *et al.*, 2020], [Cha *et al.*, 2021]. The former selects vital samples by designing sample strategies while the latter utilizes knowledge distillation to transfer the learned knowledge. Specifically, the knowledge transfer-based approaches usually leverage the old model as the teacher model and distill its response to help the

[*]Corresponding Author

new model alleviate the catastrophic forgetting.

However, the response bias towards classes of current step is also inherited, which limits the positive effect of knowledge distillation, as shown in Fig. 1. This response bias is caused by the imbalanced data between the old and the new classes since the old classes are reserved in the buffer with a small amount of data while the new ones are input currently with abundant data. It leads to the weights of the classifier (i.e. the last fully connected layer) towards classes of current step [Wu *et al.*, 2019], [Zhao *et al.*, 2020], [Ahn *et al.*, 2021]. The old model's response helps the student model to alleviate the catastrophic forgetting on some old classes as well as accumulating inherent response bias when distilling the output of the teacher model to the student model. Thus, taking the old model as the teacher model has severe limitation.

Apart from distilling the response of samples, maintaining the relation among classes is also crucial to preserve the learned knowledge. The response of a single sample is not robust to model changes, which may mislead the process of knowledge distillation. That is, it inevitably damages the model's ability to preserve the learned knowledge by only aligning the response of a single sample without considering the internal structure information among classes.

To address the above limitations, an Online Contrastive Distillation Network (OCD-Net) is proposed. It includes an Online Response Distillation Module (ORD-Module) and a Contrastive Relation Distillation module (CRD-Module), respectively. The ORD-Module is designed to align those less biased responses between the teacher and the student models. Particularly, the teacher model is trained by accumulating the student model's parameters since the student model at each time step is good at classifying different classes. The CRD-Module utilizes an embedding space to maintain the consistency of class similarities among classes, which is more robust to model changes. Moreover, an adaptive perception approach is devised to enforce the ORD-Module to fix attention on distilling the high-quality response. With the combination of ORD-Module and CRD-Module, the teacher model enables the student model to learn richer information to preserve the learned knowledge, the student model helps the teacher model accumulate new knowledge with less forgetting at the same time. The main framework of the proposed approach is illustrated in Fig. 2. The highlights are summarized as follows:

- We propose an Online Contrastive Distillation Networks (OCD-Net) for GCL, which leverages the merit of the student model by integrating their parameters to consolidate the learned knowledge and accumulate the new knowledge.
- To exploit internal structure information among classes, we further introduce a contrastive relation distillation objective function to maintain the consistency of class similarities between the teacher and the student models.
- We propose an adaptive distillation approach to perceive and strengthen the high-quality teacher model's response in online distillation, which is beneficial for the student model to consolidate the learned knowledge.

2 Related Work

To alleviate the catastrophic forgetting problem, a bounded replay buffer is usually utilized to store the old data in GCL. Typical works such as [David and Akansel, 2018], [Chaudhry *et al.*, 2019], [Rahaf *et al.*, 2019], and [Ji *et al.*, 2021b] can be classified as sample strategy-based approaches, which focus on devising effective criteria for selecting the appropriate replay samples. For example, [David and Akansel, 2018] utilized reservoir sample strategy [Vitter, 1985] to ensure each sample to be selected in the buffer with an equal probability. Based on this approach, [Chaudhry *et al.*, 2019] introduced the average episodic memory loss over the previous tasks to constrain the gradient of the model's update. [Rahaf *et al.*, 2019] further proposed an approximated greedy strategy to increase the sample diversity between the gradient of individual samples. [Ji *et al.*, 2021b] leveraged the training loss value as a metric to discard those less vital samples.

Another line of studies [Buzzega *et al.*, 2020], [Simon *et al.*, 2021], [Cha *et al.*, 2021] can be summarized as knowledge transfer based approaches, which exploits knowledge distillation [Zhang *et al.*, 2020], [Chen *et al.*, 2021], [Zhu *et al.*, 2021b] to mitigate the catastrophic forgetting via transfer the learned knowledge (e.g. logits and features) to the new model. For example, [Buzzega *et al.*, 2020] utilized knowledge distillation between the old and the new models' response. [Cha *et al.*, 2021] stored the old model and regulated the instance-wise changes in feature relation.

Our proposed OCD-Net falls into the category of the knowledge transfer-based approaches and is related to [Buzzega *et al.*, 2020] and [Cha *et al.*, 2021]. Different from them, we replace the offline teacher model with the online one, which effectively accumulates the new knowledge and consolidates the learned knowledge via integrating the student model's parameters.

3 Method

GCL aims at sequentially learning a model to classify all seen classes from a non-stationary data stream $\mathcal{D}$ without utilizing task boundaries. Specifically, a GCL model is trained by the labeled sample $\mathcal{D}_t = \{(x, y)\}$ at time step t. Additionally, we introduce a bounded replay buffer $\mathcal{M} = \{(x_i, y_i)_{i=1}^{\mathcal{B}}\}$ with fixed size $\mathcal{B}$ to store few previous samples.

The architecture of our OCD-Net mainly consists of an Online Response Distillation Module (ORD-Module), and a Contrastive Relation Distillation Module (CRD-Module), as shown in Fig. 2. In the ORD-Module, we leverage the online distillation scheme to transfer less biased response from the teacher model. Meanwhile, the CRD-Module aims at maintaining the consistency of class similarities between the teacher and the student models, which reduces the intra-class variance. Moreover, an adaptive perception idea is developed to ensure that the student model learns high-quality response by adjusting the weights of each sample.

3.1 Online Response Distillation Module

To alleviate the inherent response bias of the teacher model in the offline knowledge distillation, we propose an Online Response Distillation Module (ORD-Module), which explores

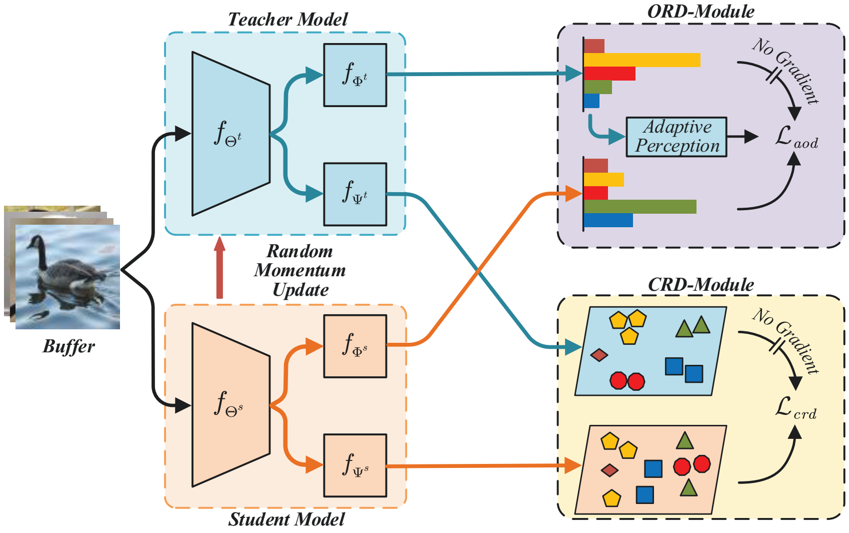


Figure 2: The main architecture of the proposed OCD-Net. We adopt the online knowledge distillation framework where the student model is trained to align the relation and response of the teacher model, and the teacher model is updated via a random momentum update of the student's parameters.

the merit of the student model to obtain a model with less forgetting on the learned knowledge. It aims at learning the reserved samples' response of the teacher model to preserve the learned knowledge. Obviously, the teacher model is crucial to preserve the learned knowledge in the online distillation scheme, in which a key challenge is how to train the teacher model. Inspired by the idea of momentum update [He *et al.*, 2020] in contrastive learning, we simply train the teacher model by random momentum update, which randomly integrates the student model's parameters to accumulate the new knowledge. In this way, the teacher model is capable of achieving knowledge accumulation.

Formally, we employ two feature extractors f_{Θ^t} and f_{Θ^s} to encode feature representations from sample x. Then, two classifiers f_{Φ^t} and f_{Φ^s} map the corresponding feature representations into the label space. The student model is trained by the online response distillation loss, which is written as:

$$\mathcal{L}_{ord} = \mathbb{E}_{x_i \sim \mathcal{M}}[\|f_{\Theta^t,\Phi^t}(x_i) - f_{\Theta^s,\Phi^s}(x_i)\|_2^2], \tag{1}$$

where the $\|\cdot\|_2$ operator refers to the ℓ_2 norm.

Meanwhile, the teacher model is trained by the random momentum update approach, which is formulated as:

$$\Theta^t \leftarrow m\Theta^t + (1-m)[(1-X)\Theta^t + X\Theta^s], \tag{2}$$

$$\Phi^t \leftarrow m\Phi^t + (1-m)[(1-X)\Phi^t + X\Phi^s], \tag{3}$$

where m is a momentum coefficient and X obeys Bernoulli distribution, which is denoted as:

$$P(X=k) = p^k(1-p)^{1-k}, k = \{0,1\}, \tag{4}$$

where Bernoulli probability p is in range $(0,1)$.

By doing so, we avoid adjusting the particularly sensitive momentum coefficient m. Besides, to help the teacher model better accumulate new knowledge, the value of momentum coefficient m is designed to increase gradually in the early stage:

$$m = min(n/(n+1), \eta), \tag{5}$$

where n is the number of model iterations and constant η is set to 0.999.

3.2 Contrastive Relation Distillation Module

Apart from distilling the response from the teacher model, maintaining the consistency of class similarities is also essential to alleviate the catastrophic forgetting. Since samples from the same class usually have similar embeddings, the embeddings from the student model should be consistent with those in the same class in the teacher model. To this end, we introduce a contrastive relation distillation loss to encourage the same class embeddings to be pulled closer and the other samples be pushed away between the teacher and the student models. Concretely, we introduce two learnable projectors f_{Ψ^t} and f_{Ψ^s} to map the samples' features to an embedding space where the contrastive relation distillation loss is applied. The projector f_{Ψ^t} is also updated by random momentum update:

$$\Psi^t \leftarrow m\Psi^t + (1-m)[(1-X)\Psi^t + X\Psi^s]. \tag{6}$$

Given a batch of data with samples x, the embedding can be expressed by $\mathbf{z}^t = f_{\Theta^t,\Psi^t}(x)$ and $\mathbf{z}^s = f_{\Theta^t,\Psi^s}(x)$. To maintain class similarities between the teacher and the student models within a batch, we fix the teacher's embedding $\mathbf{z}^t$ as anchor and enumerate the student's embedding $\mathbf{z}^s$ to approach the embeddings of the same class in the teacher model. The contrastive relation distillation loss is:

$$\mathcal{L}_{crd} = -\mathbb{E}_{\mathbf{z}_i^s \sim \mathbf{z}^s} \sum_{\mathbf{z}_j^t \sim \mathbf{z}^{t+}} \log \frac{h(\mathbf{z}_i^s, \mathbf{z}_j^t)}{\sum_{\mathbf{z}_k^t \sim \mathbf{z}^t} h(\mathbf{z}_i^s, \mathbf{z}_k^t)}, \tag{7}$$

where $\mathbf{z}^{\mathbf{t}+}$ represents the set of all teacher embeddings of the same label with $\mathbf{z}_i^s$, and critic function $h : \{\mathbf{z}_i, \mathbf{z}_j\} \to [0,1]$

indicates whether the embedding tuple $(\mathbf{z}_i, \mathbf{z}_j)$ is drawn from the joint distribution $p(\mathbf{z}_i, \mathbf{z}_j)$, which is defined as:

$$h(\mathbf{z}_i, \mathbf{z}_j) = \frac{\exp((\mathbf{z}_i/\|\mathbf{z}_i\|_2)^\top(\mathbf{z}_j/\|\mathbf{z}_j\|_2)/\tau)}{\exp(1/\tau)}, \tag{8}$$

where τ is a temperature hyper-parameter and the $(\cdot)^\top$ operator means transpose.

In addition to maintaining the internal structure information with the teacher model, the student model also encodes the data's inherent co-occurrence relationships for better generalization. Following CoCa [Ji *et al.*, 2021a], we also employ the idea of supervised contrastive learning [Khosla *et al.*, 2020] to raise the critic values from the same class and reduce them from different classes, which is formulated as:

$$\mathcal{L}_{scl} = -\mathbb{E}_{\mathbf{z}_i^s \sim \mathbf{z}^s} \sum_{\mathbf{z}_j^s \sim \mathbf{z}^{s+}} \log \frac{h(\mathbf{z}_i^s, \mathbf{z}_j^s)}{\sum_{\mathbf{z}_k^s \sim \mathbf{z}^s} h(\mathbf{z}_i^s, \mathbf{z}_k^s)}, \tag{9}$$

where $\mathbf{z}^{s+}$ represents the set of positive embeddings form $\mathbf{z}^s$.

Noticeably, the contrastive relation distillation loss $\mathcal{L}_{crd}$ and supervised contrastive learning loss $\mathcal{L}_{scl}$ are cooperative. The former is designed to match the marginal feature distribution $p(\mathbf{z}^s)$ with $p(\mathbf{z}^t)$ for consolidating the learned knowledge while the latter learns marginal distribution $p(\mathbf{z}^s)$ for better generalization. By minimizing the $\mathcal{L}_{crd}$ and $\mathcal{L}_{scl}$ simultaneously, the student model better consolidates the learned knowledge and learns new knowledge. Therefore, learning from the student model is capable of accumulating less biased knowledge.

3.3 Adaptive Perception

To further mitigate the catastrophic forgetting in our proposed OCD-Net, we deploy an adaptive perception approach to evaluate the quality of the teacher model's response, which enforces the student model to pay more attention to learn high-quality response. Specifically, the adaptive perception gets as input a sample response $r^t = f_{\Theta^t,\Phi^t}(x_i)$ with its label y_i, and yields as output a quality score $\omega(x_i)$:

$$\omega(x_i) = \frac{exp(r^t_{y_i}/\rho)}{\sum_{c'=1}^{C} exp(r^t_{c'}/\rho)}, \tag{10}$$

where ρ is the temperature parameter and C refers to the number of possible classes. By constructing the quality score, we dynamically adjust the contribution of each teacher model's response, in which a high score response is emphasized and a low score response is weakened in knowledge distillation. To this end, the online response distillation loss is substituted by the adaptive online distillation loss, which is denoted as:

$$\mathcal{L}_{aod} = \mathbb{E}_{x_i \sim \mathcal{M}}[\omega(x_i)\, \|f_{\Theta^t,\Phi^t}(x_i) - f_{\Theta^s,\Phi^s}(x_i)\|_2^2]. \tag{11}$$

3.4 Overall Objective

In summary, the total training objective for the student model in our proposed OCD-Net can be expressed by:

$$\mathcal{L} = \mathcal{L}_{ce} + \alpha_1\mathcal{L}_{scl} + \alpha_2\mathcal{L}_{aod} + \alpha_3\mathcal{L}_{crd}, \tag{12}$$

where α_1, α_2 and α_3 are weighting factors, and the classification loss $\mathcal{L}_{ce}$ is calculated by:

$$\mathcal{L}_{ce} = \mathbb{E}_{(x,y)\sim\mathcal{D}\cup\mathcal{M}}\ell(f_{\Theta^s,\Phi^s}(x), y). \tag{13}$$

Algorithm 1 The training algorithm of the OCD-Net

```
Input: Stream data $\mathcal{D}$, Buffer $\mathcal{M}$, Learning rate $\gamma$, Bernoulli Probability $p$, Temperature hyper-parameters $\tau$ and $\rho$, Weighting factors $\alpha_1$, $\alpha_2$ and $\alpha_3$
Parameter: Teacher parameters $\Theta^t$, $\Phi^t$ and $\Psi^t$, Student parameters $\Theta^s$, $\Phi^s$ and $\Psi^s$
Output: Learned teacher parameters $\Theta^t$ and $\Phi^t$
Initialize $\Theta^t = \Theta^s$,$\Phi^t = \Phi^s$,$\Psi^t = \Psi^s$, $\mathcal{M} \leftarrow \{\}$.
for $(X, Y)$ in $\mathcal{D}$ do
    $\mathcal{M} \leftarrow Reservoir(\mathcal{M}, (X, Y))$.
    $(X_m, Y_m) \leftarrow Sample(\mathcal{M})$.
    Calculate $\mathcal{L}_{aod}$ loss by Eq. (11).
    $\mathbf{z}^t \leftarrow f_{\Theta^t,\Psi^t}(X_m)$, $\mathbf{z}^s \leftarrow f_{\Theta^s,\Psi^s}(X_m)$.
    Calculate $\mathcal{L}_{crd}$ loss by Eq. (7).
    Calculate overall loss $\mathcal{L}$ by Eq. (12).
    Update $\Theta^s$, $\Phi^s$, $\Psi^s$ by gradient descent:
        $\vartheta \leftarrow \vartheta - \gamma\nabla_\vartheta\mathcal{L}, \vartheta \in \{\Theta^s, \Phi^s, \Psi^s\}$.
    Update $\Theta^t$, $\Phi^t$, $\Psi^t$ by Eq. (2), Eq. (3) and Eq. (6).
end for
return $\Theta^t$, $\Phi^t$.
```

We elaborate on the overall training procedure of our proposed OCD-Net in Alg. 1. When new data come, we first utilize the reservoir sample strategy [Vitter, 1985] to update the buffer. Then, we optimize the student model by the overall objective $\mathcal{L}$. Finally, we modify the teacher model by the random momentum update approach. At the test stage, we leverage the teacher model for testing. The reason lies in that the student models at different time steps are good at classifying different classes, the teacher model learned from the student models could accumulatively learn their merits. Therefore, the teacher model has a stronger ability in classifying all seen classes than a student model.

4 Experiments

4.1 Datasets

We evaluate our OCD-Net on both spilt and smooth benchmarks. **Split benchmarks** mean that the classes of each task are disjoint and the number of classes of each task is equal. Following [Buzzega *et al.*, 2020], [Chaudhry *et al.*, 2019], we adopt three spilt benchmarks in our experiments: Split CIFAR-10 (S-CIFAR-10) [Krizhevsky and Hinton, 2009], Split CIFAR-100 (S-CIFAR-100) [Krizhevsky and Hinton, 2009] and Split Tiny-ImageNet (S-Tiny-ImageNet) [Hadi and Saman, 2015]. They consist of 5, 20 and 10 tasks, each including 2, 5 and 20 classes respectively. **Smooth benchmarks** refer to the class of each task emerges irregularly and the distribution of class shifts gradually, which are closer to real-world scenarios. For this type of datasets, we choose MNIST-360 [Buzzega *et al.*, 2020] and generalized CIFAR-100 [Mi *et al.*, 2020] datasets. Specifically, MNIST-360 dataset offers a stream of data in which each task constitutes by two continuous rotating digits. Generalized

Method	**S-CIFAR-10**			**S-CIFAR-100**			**S-Tiny-ImageNet**		
JOINT	92.20			69.55			59.99		
SGD	19.62			4.33			7.92		
Buffer Size	200	500	5120	200	500	5120	200	500	5120
ER [David and Akansel, 2018]	44.79	57.74	82.47	9.84	14.64	44.79	8.49	9.99	27.40
GSS [Rahaf *et al.*, 2019]	39.07	49.73	67.27	6.35	7.44	9.71	8.55	9.63	14.16
A-GEM [Chaudhry *et al.*, 2019]	20.04	22.67	21.99	4.73	4.74	4.87	8.07	8.06	7.96
DER [Buzzega *et al.*, 2020]	61.93	70.51	83.81	15.22	24.11	44.8	11.87	17.75	36.73
DER++ [Buzzega *et al.*, 2020]	64.88	72.70	85.24	18.66	28.70	51.20	10.96	19.38	39.02
CER [Ji *et al.*, 2021b]	68.07	76.63	85.38	19.02	30.50	52.33	11.48	18.05	37.64
GeoDL[Simon *et al.*, 2021]	49.20	61.83	85.91	13.38	23.06	54.57	10.08	12.03	36.29
CO^{2}L [Cha *et al.*, 2021]	65.57	74.26	84.27	18.85	24.45	46.18	13.88	20.12	37.14
OCD-Net (Ours)	**72.61**	**81.28**	**89.67**	**27.24**	**36.80**	**59.74**	**21.54**	**28.42**	**47.03**

Table 1: Average classification accuracy (%) on split benchmarks. The best results are marked in bold, and the second best results are marked underlined.

Method	**MNIST-360**			**UG-CIFAR-100**			**LG-CIFAR-100**		
JOINT	82.98			60.19			54.72		
SGD	19.02			18.05			15.02		
Buffer Size	200	500	1000	200	500	1000	200	500	1000
ER[David and Akansel, 2018]	49.27	65.04	75.18	24.20	27.65	32.89	21.56	24.79	32.25
GSS [Rahaf *et al.*, 2019]	43.92	54.45	63.84	20.99	24.27	26.27	20.50	22.35	25.03
A-GEM[Chaudhry *et al.*, 2019]	28.34	28.13	29.21	18.13	18.87	19.91	15.53	16.69	17.49
DER[Buzzega *et al.*, 2020]	55.22	69.11	75.97	29.23	34.79	40.01	27.46	33.50	38.38
DER++[Buzzega *et al.*, 2020]	54.16	69.62	76.03	28.79	33.14	40.51	25.27	32.59	36.97
CER[Ji *et al.*, 2021b]	59.18	72.16	78.36	26.86	32.91	36.02	26.08	30.99	35.61
GeoDL[Simon *et al.*, 2021]	54.17	70.64	78.98	25.78	33.53	37.22	23.88	32.00	36.73
CO^{2}L[Cha *et al.*, 2021]	59.30	69.10	76.83	24.67	26.12	33.69	22.97	24.04	31.02
OCD-Net (Ours)	**67.27**	**76.80**	**82.01**	**37.94**	**38.98**	**43.77**	**33.19**	**38.50**	**42.00**

Table 2: Average classification accuracy (%) on smooth benchmarks. The best results are marked in bold, and the second best results are marked underlined.

CIFAR-100 has two variants: Uniform Generalized CIFAR-100 (UG-CIFAR-100) and Longtail Generalized CIFAR-100 (LG-CIFAR-100). Both of them have 20 tasks, the difference is whether the class distribution is uniform.

4.2 Implementation Details

Following [Buzzega *et al.*, 2020], [Ji *et al.*, 2021a], we adopt a fully-connected network with two layers for MNIST-360 dataset and ResNet-18 [He *et al.*, 2016] for the other datasets as feature extractors. A linear layer and an MLP with one hidden layer are instantiated as classifier and projector, respectively. The stochastic gradient descent optimizer is employed for optimization. The model is trained on each task with 50 epochs for all datasets except for MNIST-360 dataset, which is trained with a single epoch.

4.3 Comparison with the State-of-the-Art

We compare our OCD-Net against a series of GCL approaches, including ER [David and Akansel, 2018], GSS [Rahaf *et al.*, 2019], DER [Buzzega *et al.*, 2020], DER++ [Buzzega *et al.*, 2020], and CER [Ji *et al.*, 2021b]. Meanwhile, we also modify A-GEM [Chaudhry *et al.*, 2019], GeoDL [Simon *et al.*, 2021], and CO^{2}L [Cha *et al.*, 2021] in line with the GCL setting. Besides, we also provide approximate upper and lower bounds on all datasets, in which the former is trained on all shuffled samples (JOINT) and the latter trains the model without any strategy (SGD). For the same dataset, all approaches utilize the same architecture to ensure a fair comparison.

Comparison on Split Benchmarks. Table 1 summarizes the average classification accuracy of OCD-Net and competitors on three split benchmarks with three different buffer sizes. We could observe that OCD-Net clearly outperforms the competitors in all cases. Particularly, it obtains at least 4.29% improvements among these datasets with different buffer sizes. For example, OCD-Net achieves 81.28%, 36.80% and 28.42% accuracies on S-CIFAR-10, S-CIFAR-100 and S-Tiny-ImageNet datasets with 500 buffer sizes, outperforming the second-best approaches by 4.65%, 6.30% and 8.30%, respectively. Similar to other approaches, it could also be observed that increasing the buffer size is beneficial to improve the performance. Moreover, it can be observed that the performance on S-CIFAR-10 dataset is higher than S-CIFAR-100 and S-Tiny-ImageNet datasets since it has fewer classes

Method	S-CIFAR-10			UG-CIFAR-100		
OCD w/o BUF	10.00			1.42		
Buffer Size	200	500	5120	200	500	1000
OCD $w/$ ON	67.38	73.13	85.86	31.76	31.92	37.99
OCD w/o CRD	67.75	76.30	86.32	35.67	35.86	40.90
OCD w/o ORD	68.13	75.63	88.59	36.73	37.89	42.49
OCD w/o AP	72.35	80.36	89.44	37.51	38.33	42.81
OCD-Net	**72.61**	**81.28**	**89.67**	**37.94**	**38.98**	**43.05**

Table 3: Ablation Studies of OCD-Net on S-CIFAR-10 and UG-CIFAR-100 datasets.

and tasks than the other datasets.

Comparison on Smooth Benchmarks. Table 2 reports the results of the comparison with the same eight competitors on three smooth benchmarks. It demonstrates that our proposed OCD-Net improves steadily on all datasets under all buffer sizes. Concretely, our OCD-Net achieves an 8.71% improvement over the sub-optimal approach DER on UG-CIFAR-100 dataset with 200 buffer sizes, which is quite notable. It can also be observed that the results on LG-CIFAR-100 dataset perform worse than those on UG-CIFAR-100 dataset, which illustrates the imbalance of classes aggravate the catastrophic forgetting problem.

4.4 Ablation Studies

Table 3 shows the ablation studies for each component in OCD-Net on S-CIFAR-10 and UG-CIFAR-100 datasets. We first validate the impact of the buffer by removing the replay buffer (OCD w/o BUF). Then, we train OCD-Net by removing the CRD-Module and the ORD-Module (OCD $w/$ ON). Finally, we train OCD-Net by removing the CRD-Module (OCD w/o CRD), the ORD-Module (OCD w/o ORD) and the adaptive perception (OCD w/o AP), respectively. From the results, we have the following observations: (1) OCD w/o BUF brings a sharp decline on both datasets. The results of OCD w/o BUF and OCD $w/$ ON indicate that it is essential to help the student model to preserve the learned knowledge in our OCD-Net. (2) The proposed ORD-Module, CRD-Module and adaptive perception approach are all proved to be positive in OCD-Net. Moreover, they are mutually complementary in mitigating the catastrophic forgetting of the student model.

4.5 Quantitative Analysis

The Impact of Bernoulli Probability p. An experiment is conducted to further discuss the influence on the teacher and the student models of the Bernoulli probability p in Eq. (4), as shown in Fig. 3. This Bernoulli probability is designed to control the updating frequency of the teacher model. From the results, we could observe similar trends in both the teacher and the student models, indicating that the student model is crucial for the teacher model in online distillation schemes. Additionally, the teacher model achieves the apex when probability $p = 0.2$. We speculate that higher probability accumulates more inherent bias, while lower probability accumulates less learned knowledge.

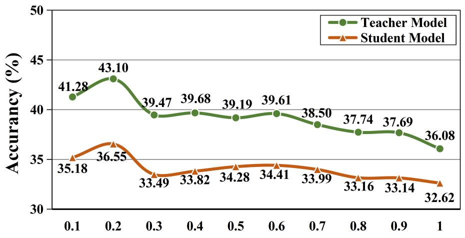
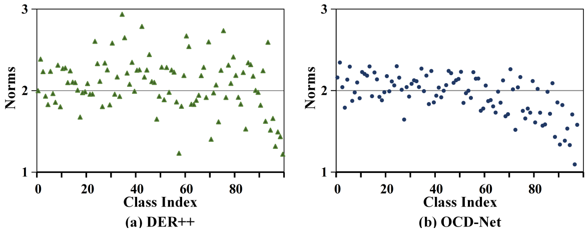


Figure 3: The impact of Bernoulli probability p on LG-CIFAR-100 dataset with buffer size of 500.



(a) DER++

(b) OCD-Net

Figure 4: Norms of the Weight Vectors in the Classifier on LG-CIFAR-100 dataset with buffer size of 500.

Norms of the Weight Vectors in the Classifier. Figure 4 shows the norms of the weight vectors in the classifier (i.e. the last fully connected layer) on LG-CIFAR-100 dataset. We take the popular offline distillation method DER++ [Buzzega *et al.*, 2020] for comparison. As presented in Fig. 4(a), the norms of the weight vectors show a large variance, which illustrates the classifier is biased towards some classes. By contrast, the weight learned by our proposed OCD-Net is obviously less biased, as shown in Fig. 4(b). It proves that the proposed online distillation scheme is capable of alleviating the inherent classifier bias widely exists in the offline distillation scheme.

5 Conclusion

In this paper, we have proposed the Online Contrastive Distillation Networks (OCD-Net) for GCL, which exploits a strong teacher model by accumulatively learning the merit of the student model via an online distillation scheme. The teacher model is trained via integrating the student model's parameters to accumulate the new knowledge while the student model is trained via maintaining the relation and adaptive response to consolidate the learned knowledge. Therefore, the teacher model integrated from the student models not only quickly learns new knowledge but also effectively alleviates the catastrophic forgetting. The experimental results on six benchmarks demonstrate that the proposed OCD-Net outperforms state-of-the-art approaches by a large margin. Future works include extensions of the OCD-Net on more challenging scenarios like semi-supervised GCL [Brahma *et al.*, 2021] and non-exempler GCL [Zhu *et al.*, 2021a].

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) under Grants 62176178.

References

[Ahn *et al.*, 2021] Hongjoon Ahn, Jihwan Kwak, Subin Lim, Hyeonsu Bang, Hyojun Kim, and Taesup Moon. Ss-il: Separated softmax for incremental learning. In *ICCV*, pages 844–853, 2021.

[Brahma *et al.*, 2021] Dhanajit Brahma, Vinay Kumar Verma, and Piyush Rai. Hypernetworks for continual semi-supervised learning. In *IJCAI Workshop on Continual Semi-Supervised Learning*, 2021.

[Buzzega *et al.*, 2020] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. In *NeurIPS*, pages 15920–15930, 2020.

[Cha *et al.*, 2021] Hyuntak Cha, Jaeho Lee, and Jinwoo Shin. Co2l: Contrastive continual learning. In *ICCV*, pages 9516–9525, 2021.

[Chaudhry *et al.*, 2019] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-GEM. In *ICLR*, 2019.

[Chen *et al.*, 2021] Liqun Chen, Dong Wang, Zhe Gan, Jingjing Liu, Ricardo Henao, and Lawrence Carin. Wasserstein contrastive representation distillation. In *CVPR*, pages 16296–16305, 2021.

[David and Akansel, 2018] Isele David and Cosgun Akansel. Selective experience replay for lifelong learning. In *AAAI*, pages 3302–3309, 2018.

[Delange *et al.*, 2021] Matthias Delange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Greg Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE TPAMI*, 2021.

[Hadi and Saman, 2015] Pouransari Hadi and Ghili Saman. Tiny imagenet visual recognition challenge. *CS231N course, Stanford University, Stanford, CA, USA*, 2015.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[He *et al.*, 2020] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9239–9738, 2020.

[Ji *et al.*, 2021a] Zhong Ji, Jin Li, Qiang Wang, and Zhongfei Zhang. Complementary calibration: Boosting general continual learning with collaborative distillation and self-supervision. *arXiv preprint arXiv: 2109.02426*, 2021.

[Ji *et al.*, 2021b] Zhong Ji, Jiayi Liu, Qiang Wang, and Zhongfei Zhang. Coordinating experience replay: A harmonious experience retention approach for continual learning. *Knowledge-Based Systems*, 234:107589, 2021.

[Khosla *et al.*, 2020] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *NeurIPS*, pages 18661–18673, 2020.

[Krizhevsky and Hinton, 2009] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Handbook of systemic autoimmune diseases*, 1(4), 2009.

[Liu *et al.*, 2021] Yaoyao Liu, Bernt Schiele, and Qianru Sun. Adaptive aggregation networks for class-incremental learning. In *CVPR*, pages 2544–2553, 2021.

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

[Mi *et al.*, 2020] Fei Mi, Lingjing Kong, Tao Lin, Kaicheng Yu, and Boi Faltings. Generalized class incremental learning. In *CVPR Workshop*, pages 970–974, 2020.

[Quang *et al.*, 2021] Pham Quang, Liu Chenghao, and Hoi Steven. Dualnet: Continual learning, fast and slow. In *NeurIPS*, pages 16131–16144, 2021.

[Rahaf *et al.*, 2019] Aljundi Rahaf, Lin Min, Goujaud Baptiste, and Bengio Yoshua. Gradient based sample selection for online continual learning. In *NeurIPS*, pages 11816–11825, 2019.

[Simon *et al.*, 2021] Christian Simon, Piotr Koniusz, and Mehrtash Harandi. On learning the geodesic path for incremental learning. In *CVPR*, pages 1591–1600, 2021.

[Tao *et al.*, 2020] Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong. Few-shot class-incremental learning. In *CVPR*, pages 12183–12192, 2020.

[Vitter, 1985] Jeffrey S. Vitter. Random sampling with a reservoir. *ACM Transactions on Mathematical Software*, 11(1):37–57, 1985.

[Wu *et al.*, 2019] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *CVPR*, pages 374–382, 2019.

[Zhang *et al.*, 2020] Youcai Zhang, Zhonghao Lan, Yuchen Dai, Fangao Zeng, Yan Bai, Jie Chang, and Yichen Wei. Prime-aware adaptive distillation. In *ECCV*, pages 658–674, 2020.

[Zhao *et al.*, 2020] Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, and Shu-Tao Xia. Maintaining discrimination and fairness in class incremental learning. In *CVPR*, pages 13205–13214, 2020.

[Zhao *et al.*, 2021] Hanbin Zhao, Yongjian Fu, Mintong Kang, Qi Tian, Fei Wu, and Xi Li. Mgsvf: Multi-grained slow vs. fast framework for few-shot class-incremental learning. *IEEE TPAMI*, 2021.

[Zhu *et al.*, 2021a] Fei Zhu, Xuyao Zhang, Chuang Wang, Fei Yin, and Chenglin Liu. Prototype augmentation and self-supervision for incremental learning. In *CVPR*, pages 5871–5880, 2021.

[Zhu *et al.*, 2021b] Jinguo Zhu, Shixiang Tang, Dapeng Chen, Shijie Yu, Yakun Liu, Mingzhe Rong, Aijun Yang, and Xiaohua Wang. Complementary relation contrastive distillation. In *CVPR*, pages 9260–9269, 2021.