Integrative Solution for Catastrophic Forgetting in Data-Free Class Incremental Learning

Anonymous submission

Abstract

Data-Free Class Incremental Learning (DFCIL) scenario enables the model to continually learn without accessing old data, which prevents the training process from violating data privacy. However, we observe that previous DFCIL approaches still severely suffer from *catastrophic forgetting* caused by the class imbalance problem. To overcome catastrophic forgetting from class imbalance, we introduce Integrative Solution for Catastrophic Forgetting (ISCF), a novel DFCIL framework consisting of *Weight Equalizer* (WEQ) and the combined knowledge distillation approach. We firstly propose *Weight Equalizer* (WEQ) to solve the class imbalance issue by correcting biased weights toward new classes in the classification head. WEQ makes the weight norms of the current linear head follow the average of weight norms in the previous training step. Furthermore, we suggest exploiting *Similarity Preserving Knowledge Distillation* (SPKD) with *Logit Knowledge Distillation* (LKD) that distills the features of intermediate layers and representations of the previous linear head to gain more stability (*i.e.*, not to forget previous knowledge). Extensive experiments on CIFAR-100 and Tiny-ImageNet demonstrate significant accuracy improvement of our proposed method over previous works. We also show that our proposed method largely remains the previously learned knowledge compared to the existing methods.

Introduction

Conventional Deep Neural Networks (DNNs) typically train with an offline batch where all data are available at once. However, many real-world applications require the model to learn with continuously incoming new class data, which is called the Class Incremental Learning (CIL) paradigm (Silver, Yang, and Li 2013; Parisi et al. 2019). One of the challenging problems in CIL is catastrophic forgetting, that the model forgets the previously learned knowledge when gradually learning new information (McCloskey and Cohen 1989; De Lange et al. 2021). Various CIL works alleviate catastrophic forgetting via storing a small portion of previously observed data in a fixed memory budget (Rebuffi et al. 2017; Castro et al. 2018; Wu et al. 2019). Despite successful works, saving old samples is not trivial in real-world scenarios due to data privacy and constrained memory size. To address data privacy issue, several works emphasize the *Data-Free Class Incremental Learning* (DFCIL) approaches. (Li and Hoiem 2017; Shin et al. 2017; Cong et al. 2020; Xiang et al. 2019; Yin et al. 2020; Smith et al. 2021). Firstly, Li and Hoiem (2017) proposes the CIL strategy without remaining any old samples to mitigate the data privacy concern but this shows poor performance due to the absence of old data. Shin et al. (2017); Cong et al. (2020); Xiang et al. (2019) concurrently train the model and generator that generates observed classes before training the next task. However, generator-based methods potentially violate data privacy since the generator captures the sensitive information in the real data during training.

To protect data privacy, Yin et al. (2020); Smith et al. (2021) introduce model inversion-based approaches, which synthesize the images using the current model. Although the inversion-based works recover the old samples without any real data, mitigating catastrophic forgetting only with synthetic images is extremely challenging. As shown in Fig.1, all of the previous DFCIL approaches are severely prone to catastrophic forgetting. Especially, Fig.2 illustrates that their weights in the classification head are heavily biased toward new classes, indicating that they suffer from class imbalance (He, Wang, and Chen 2021). Since samples per old class are getting decreased compared to new classes while training, the model tends to be overconfident in newly seen classes. This phenomenon, which is the class imbalance issue, results in an accuracy drop for old classes and eventually ends up with catastrophic forgetting.

Inspired by our observation, we introduce *Integrative Solution for Catastrophic Forgetting* (ISCF), an effective strategy for preserving old knowledge in DFCIL scenario. ISCF consists of *Weight Equalizer* (WEQ) and the hybrid knowledge distillation approach. WEQ corrects the biased weights towards new classes in the classification head while training the model. To be specific, WEQ regularizes the weight norm scale of the current classification head to follow the average of weight norms in the classification head of the previous step. Notably, WEQ perfectly maintains equivalence between weight norm scales over all of the seen classes. In addition, to further preserve previously learned information, we exploit the *Similarity Preserving Knowledge Distillation* (SPKD) with the *Logit Knowledge Distillation* (LKD). While SPKD transfers the spatial attention of the prior model to the current model encouraging the current network to mimic similar semantic features of previously

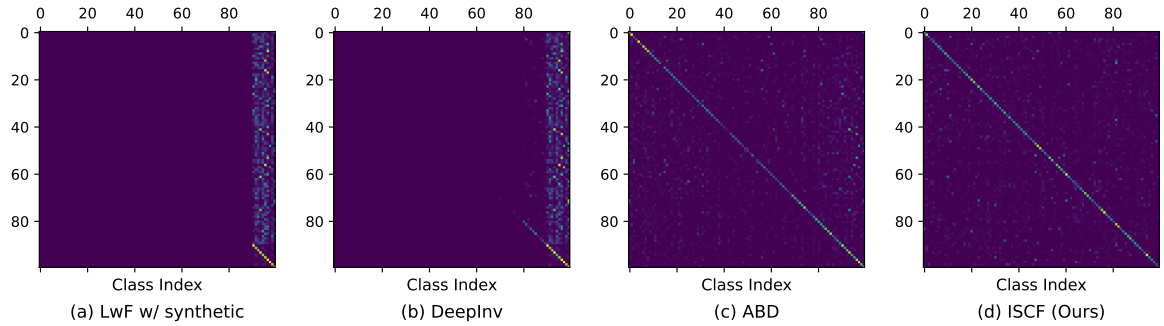

Figure 1: Confusion matrix of previous DFCIL methods and ours on CIFAR-100 in 10-task class incremental learning. For fair comparison, we feed synthetic old class images to LwF based on inversion while training new task. The (a-c) show poor performance when predicting old classes, indicating that they suffer from catastrophic forgetting. On the contrary, ISCF performs well on predicting old classes.

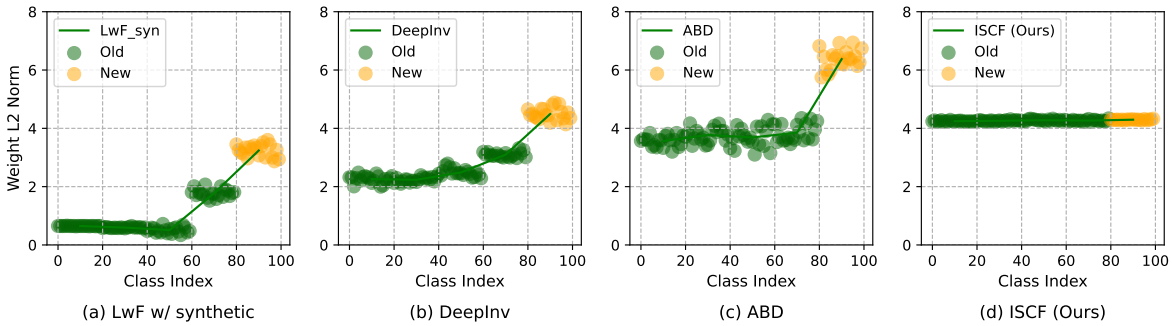

Figure 2: Visualization of weight norm in classification head on last step in CIFAR-100 5-task class incremental learning. (a), (b), and (c) display considerably larger weight norm scale of new classes than old. In contrast, ISCF balances weight norm scales of old and new perfectly via WEQ.

observed classes, LKD constraints output probability for remaining stability (i.e., the ability to remember old task information). The proposed knowledge distillation approach helps the model prevent catastrophic forgetting by penalizing the excessive change of the current model.

Experiments on CIFAR-100 and Tiny-ImageNet demonstrate the superiority of our proposed method over existing works. In addition, we verify that our proposed method achieves a significant improvement in the prediction performance of old classes without forgetting. To summarize, the contribution of our proposed methods are as follows:

- To mitigate the class imbalance problem, we introduce the *Weight Equalizer* (WEQ) that reduces the bias of the classification head toward new classes.
- We propose a hybrid knowledge distillation strategy exploiting *Similarity Preserving Knowledge Distillation* (SPKD) with *Logit Knowledge Distillation* (LKD), which achieves higher stability by transferring feature information as additional guidance and constraining output probability from the previous model.
- Our proposed method achieves outstanding performance in top-1 accuracy compared to previous Data-Free Class Incremental Learning (DFCIL) methods by alleviating catastrophic forgetting.

Background and Related works.

Data-Free Class Incremental Learning (DFCIL)

The difference between Class Increment Learning (CIL) and Data-Free Class Incremental Learning (DFCIL) is whether the model continually learns with or without storing any previously learned data. Li and Hoiem (2017) firstly proposed Learning Without Forgetting (LwF), which is the knowledge distillation-based incremental learning paradigm without any old samples. The main drawback of LwF is performing knowledge distillation with only current task data.

Shin et al. (2017); Wu et al. (2018); Cong et al. (2020) simultaneously train the generator and model for generating old samples. However, generative replay approaches also potentially violate data privacy because the generative model captures the sensitive information of real images during the training. In addition, since the distribution of generated images might be far from the real data distribution, the model could not create a decision boundary by the semantic gap, resulting in poor performance.

To mitigate data privacy, Yin et al. (2020); Smith et al. (2021) propose the model inversion-based approach to synthesize old class data. The model inversion-based approaches, which optimize the input space from noise to image, synthesize previous task images following the real data

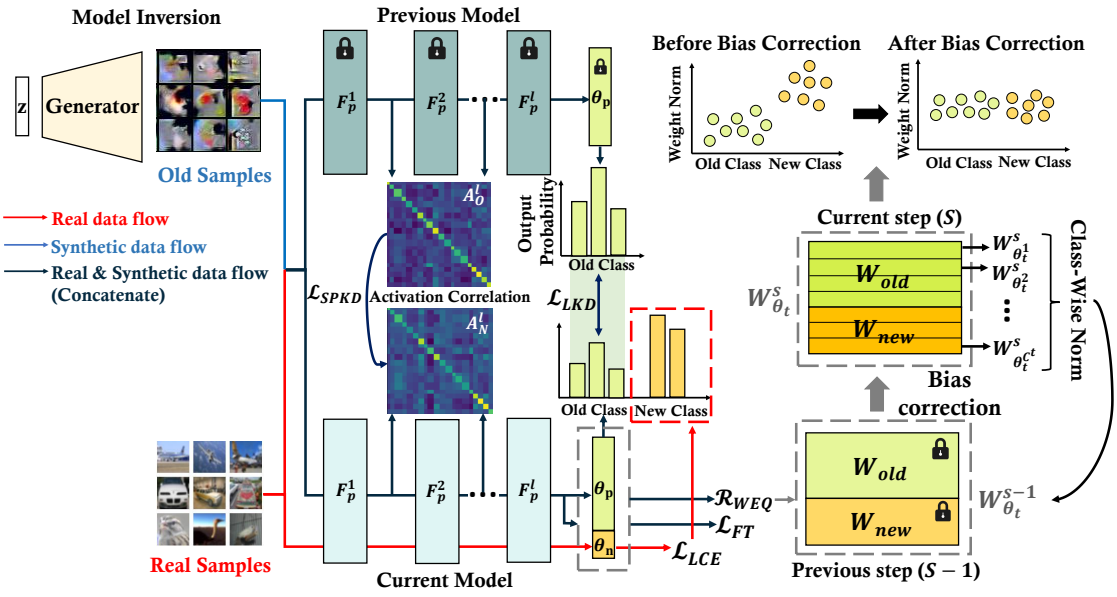

Figure 3: Overall diagram of ISCF. Our method consists of 5 techniques for addressing catastrophic forgetting. 1) Local cross-entropy loss $\mathcal{L}_{LCE}$ computed on only new classes for enhancing plasticity 2) *Similarity Preserving Knowledge Distillation* ($\mathcal{L}_{SPKD}$) for preserving the semantic information in each class. 3) *Logit Knowledge Distillation* ($\mathcal{L}_{LKD}$) to maintain the output probability in the past. 4) Fine-tuning loss ($\mathcal{L}_{FT}$) to find ideal decision boundary between old and new classes. 5) Weight Equalizer ($\mathcal{R}_{WEQ}$) for correcting the bias towards new class by equalizing the weight norm scale in classification head following previous step.

statistics stored in the model after finishing the current task training. Synthesized samples through model inversion effectively make the model create a decision boundary by the semantic gap, not the distribution gap between real and synthetic data. We adopt the model inversion approach that does not infringe on the data privacy problem and reduce the domain gap.

Catastrophic Forgetting

When a model continually learns new class data, the model forgets the previously learned information due to overwritten parameters upon learning new information. Several works (Prabhu, Torr, and Dokania 2020; Ahn et al. 2021; Zhou et al. 2019; Liu et al. 2020) propose to alleviate the catastrophic forgetting problem in Class Incremental Learning (CIL). In particular, we briefly discuss the (1) bias correction and (2) knowledge distillation in CIL.

Bias Correction. The rehearsal strategy in CIL stores the old historical data with a limited memory budget to prevent catastrophic forgetting (Van de Ven and Tolias 2019). However, as learning new classes, the stored data per old class gradually decreases, resulting in a class imbalance between real and old data (Zhao et al. 2020; He, Wang, and Chen 2021; Kim, Jeong, and Kim 2020). The class imbalance problem causes the model to be significantly biased towards newly observed classes over time, leading to the performance degradation of predicting old classes. Wu et al. (2019) proposes the bias correction layer to correct the biased output logits in new classes.

Zhao et al. (2020) empirically finds that the norms of the weights in the classification head are biased to the new categories. Regarding this, Zhao et al. (2020) aligns the weights of new classes in the classification head to have a similar weight norm scale of old classes. He, Wang, and Chen (2021) proposes the post-scaling method to correct the biased logits in the inference phase. The above works imply that the model suffers from catastrophic forgetting due to the class imbalance problem. Inspired by prior works, we introduce *Weight Equalizer* (WEQ) that completely balances the magnitude of weight norm in a classification head to tackle the class imbalance issue.

Knowledge Distillation (KD). Knowledge distillation was pioneered by Hinton et al. (2015), which transfers the knowledge from the pre-trained teacher to the student to improve the training outcomes under the guided information. Tung and Mori (2019) introduces Similarity Preserving Knowledge Distillation (SPKD) that encourages the student network to produce semantically similar attention by transferring the spatial attention maps of intermediate layers. The above works demonstrate that transferring a teacher's intermediate representation assists in training student models effectively. With the CIL setting, several works (Li and Hoiem 2017; Rebuffi et al. 2017; Castro et al. 2018) utilize KD to overcome catastrophic forgetting by preventing the deterioration of old knowledge while learning new task. In common, above studies conduct the knowledge distillation focused on only the final output. To further preserve the old knowledge stored in the previous model, Douillard et al.

(2020) proposes the integrated KD strategy transferring the pooled outputs from intermediate layers and final output embedding. Their experiments demonstrate outstanding performance compared with other CIL methods. Driven from the above approach, we suggest combining SPKD with LKD as a hybrid KD strategy.

Methodology

In this section, we describe preliminary and our method, which includes the hybrid knowledge distillation approach and the *Weight Equalizer* (WEQ).

Preliminary

The given network continually learns the series of tasks in Class Incremental Learning (CIL) denoted by $\mathcal{T} = \{\mathcal{T}_1, ..., \mathcal{T}_N\}$. When learning each t_{th} task, $\mathcal{T}_t$, the model observes a set of training sets $\mathcal{D}_t$ corresponding to the class groups, $\mathcal{C}_t$. Specifically, in the rehearsal-based CIL method, the stream of training sets $\mathcal{D}_t$ includes samples of new classes $\mathcal{C}_t^{new}$ and the previously observed classes $\mathcal{C}_t^{old}$ denoted by $\mathcal{D}_t = \{\mathbf{x}_t, \mathbf{y}_t\} = \{(\mathbf{x}_t^{new} \cup \mathbf{x}_t^{old}), (\mathbf{y}_t^{new} \cup \mathbf{y}_t^{old})\}$. The goal of CIL is that the model classifies all seen classes, $\mathcal{C}_{1:t} = \mathcal{C}_t$.

Our baseline CIL framework based on knowledge distillation utilizes the cross-entropy loss $\mathcal{L}_{CE}$ in conjunction with knowledge distillation loss $\mathcal{L}_{KD}$ given by:

$$\mathcal{L}_{CIL} = \mathcal{L}_{CE}(\mathbf{x}, \mathbf{y}) + \mathcal{L}_{KD}(\mathbf{x}). \tag{1}$$

Generally, the cross-entropy loss is simultaneously computed for all samples from new classes and old samples from the memory buffer. However, Smith et al. (2021) verifies that applying the cross-entropy loss for all seen classes leads to separating the decision boundary with not the semantic information but distribution gap between real and synthetic data. Therefore, We utilize the *local* cross-entropy loss that computes cross-entropy loss only with new classes below:

$$\mathcal{L}_{LCE} = \sum_{k=C_t^{old}+1}^{C_t^{new}} -\delta_{y=k} \log(p_k(\mathbf{x}^k)), \tag{2}$$

where the $\delta_{y=k}$ is the indicator function and p_k means the output probabilities about the k_{th} class in new classes. Generally, the distillation loss encourages the output of the current model to approximate the outputs of the previous model to maintain the model's stability. Our distillation loss is formulated as follows:

$$\mathcal{L}_{KD} = \sum_{k=1}^{C_t^{old}} -\hat{q_k}(\mathbf{x}^k) \log(q_k(\mathbf{x}^k)), \tag{3}$$

where $\hat{q_k}$=$\frac{e^{\hat{o}^k(\mathbf{x}^k)/T}}{\sum_{j=1}^{C_t^{old}} e^{\hat{o}^j(\mathbf{x}^j)/T}}$, and q_k=$\frac{e^{o^k(\mathbf{x}^k)/T}}{\sum_{j=1}^{C_t^{old}} e^{o^j(\mathbf{x}^j)/T}}$.

$\hat{\mathbf{o}}$ is output logits of previous model denoted by $\hat{\mathbf{o}} = \{\hat{o}_{1:t-1}^1(\mathbf{x}), ..., \hat{o}_{1:t-1}^{c^{old}}(\mathbf{x})\}$ and $\mathbf{o} = \{o_t^1(\mathbf{x}), ..., o_t^{c^{old}}(\mathbf{x})\}$, which means the output logits of current model. Note that the training set $\mathbf{x}$ includes new and synthetic samples. Finally, both feature extractor, $\mathcal{F}_t$ and the classification head, θ_t are updated by $\mathcal{L}_{CIL}$.

Hybrid Knowledge Distillation Strategy

In CIL, the Knowledge Distillation (KD) based methods are widely used to preserve knowledge of previous data by restricting the excessive change of the weights. Our intuition is that transferring additional guidance from the previous model helps to consistently conserve the historical weights. To maintain the important weights in the previous model, we propose a hybrid KD approach that incorporates *Similarity Preserving Knowledge Distillation* (SPKD) and *Logit Knowledge Distillation* (LKD).

More specifically, we guide the activation correlations induced in the previous model for concentrating similar parts when incoming data. Tung and Mori (2019) transfers the spatial attention extracted from the intermediate layers of the teacher network to the student network to mimic the semantically similar features of the teacher network. Inspired by the above works, we suggest applying the similarity-preserving distillation approach to our proposed DFCIL framework, distilling which spatial areas of the input the previous model focused on most for class prediction. For applying SPKD to our framework, we extract the activation maps at the down-sampling point with different spatial attention information. Given an input batch b, we can extract the activation map, $\mathbf{M}^{(l)} = \{\mathbf{M}_{old}^{(l)} \cup \mathbf{M}_{new}^{(l)}\}$ from each intermediate layer l, with $l \in L = \{l_1, l_2, ..., l_L\}$, of old and new models independently. We denote the $\mathbf{M}^{(l)} \in \mathbf{R}^{b \times c \times w \times h}$ with the number of output channels as c, width and height as w and h. The SPKD distills the L2-normalized outer product vector of activation maps from previous model below:

$$\mathcal{A}_{[i,:]}^{(l)} = \tilde{\mathcal{A}}_{[i,:]}^{(l)} / \|\tilde{\mathcal{A}}_{[i,:]}^{(l)}\|_2; \;\; \tilde{\mathcal{A}}^{(l)} = Q^{(l)} \cdot Q^{(l)\top}, \tag{4}$$

where the $Q^{(l)} \in \mathbf{R}^{b \times cwh}$ reshapes the $\mathbf{M}^{(l)}$ and the $\tilde{\mathcal{A}}^{(l)}$ reveals the activation similarity between i_{th} and j_{th} images within one mini-batch. Finally, we calculate the L2-normalized attention map from the previous and current model denoted as $\mathcal{A}^{(l)} = \{\mathcal{A}_{new}^{(l)} \cup \mathcal{A}_{old}^{(l)}\}$, where the $[i, :]$ means the i_{th} row in attention similarity matrix. The total loss of activation similarity-based SPKD is defined below:

$$\mathcal{L}_{SPKD}(\mathcal{A}_{new}, \mathcal{A}_{old}) = \frac{1}{b^2} \sum_{l \in L} \|\mathcal{A}_{new}^{(l)} - \mathcal{A}_{old}^{(l)}\|_2 \tag{5}$$

We minimize SPKD loss by forwarding real and synthetic data to maintain the model explanations. Forwarding real and synthetic data at once enables the model to separate decision boundary with semantic information.

To further keep the representation of the previous classification head, we utilize the LKD to strictly maintain the output probability, we directly distill the logits from the previous to the current classification head as:

$$\mathcal{L}_{LKD} = \|\theta_t(\mathcal{F}_t(\mathbf{x})) - \theta_{1:t-1}(\mathcal{F}_{1:t-1}(\mathbf{x}))\|_2, \tag{6}$$

where $\mathcal{F}_{1:t-1}$ and $\theta_{1:t-1}$ mean the previous models trained until the $\mathcal{T}_{t-1}$. We apply the LKD upon real and synthetic data to reinforce important parts of the past classification head. Finally, the proposed KD loss combines above two components:

$$\mathcal{L}_{KD} = \lambda_{SPKD}\mathcal{L}_{SPKD}(\mathbf{x}) + \lambda_{LKD}\mathcal{L}_{LKD}(\mathbf{x}), \tag{7}$$

where the λ_{SPKD} and λ_{LKD} are the hyperparameters for balancing the contribution of two losses. More detailed hyperparameters are described in the appendix. Our proposed KD strategy ensures the model preserves the final logits and the intermediate layer information to further alleviate the catastrophic forgetting problem.

Weight Equalizer

Unfortunately, as the model incrementally learns incoming data, the model is gradually biased to newly learned classes due to class imbalance. In our DFCIL frameworks, although the real and synthetic images are reasonably considered one batch at a time, the class imbalance problem still occurs because the model inversion approach synthesizes accumulated classes image over time. This phenomenon leads to the prediction bias, which is overconfident in newly seen classes, exacerbating catastrophic forgetting. To address this issue, we propose the effective weight regularizer, *Weight Equalizer* (WEQ) for bias correction in classification head. We aim to prevent the model from the class imbalance issue by balancing the overall parameter scale in the current classification head. Specifically, we consistently align the class-wise norm of the weight and bias vectors of the current classification head along with the average weight norm of the previous step while training. Note that the step indicates one gradient update within one epoch; thus one epoch consists of several training steps. Since the bias of the classification head can be defined in the same way as weight, we express only the weight to define the equation for convenience. We denote the weights of the classification head as follows:

$$\mathbf{W}^{s}_{\theta_t} = (\mathbf{w}^{s}_{\theta^1_t}, \mathbf{w}^{s}_{\theta^2_t}, ..., \mathbf{w}^{s}_{\theta^{C_t}_t}) \in \mathbf{R}^{d\times C_t},$$

$$\mathbf{W}^{s-1}_{\theta_t} = (\mathbf{w}^{s-1}_{\theta^1_t}, \mathbf{w}^{s-1}_{\theta^2_t}, ..., \mathbf{w}^{s-1}_{\theta^{C_t}_t}) \in \mathbf{R}^{d\times C_t},$$

where $\mathbf{W}^{s}_{\theta_t}$ intuitively indicates the weight vectors of classification head in current step, and $\mathbf{W}^{s-1}_{\theta_t}$ is those in previous step, $s-1$. We freeze the $\mathbf{W}^{s-1}_{\theta_t}$ and induce class-wise $\mathbf{W}^{s}_{\theta^c_t}$ to mimic the scale of these fixed weights. The average norm of weights, which is the representative value of weights in previous step, and class-wise weight norm in current step is written as:

$$\begin{aligned} \|\mathbf{W}^{s}_{\theta_t}\|_1 &= \{\|\mathbf{W}^{s}_{\theta^i_t}\|_1, ..., \|\mathbf{W}^{s}_{\theta^{C_t}_t}\|_1\}, \\ \|\mathbf{W}^{s}_{\theta^i_t}\|_1 &= \left|\mathbf{W}^{s}_{\theta^i_t}\right|, i \in C_t \\ \|\mathbf{W}^{s-1}_{\theta_t}\|_1 &= \frac{1}{C_t}\sum_{i=1}^{C_t}\left|\mathbf{W}^{s-1}_{\theta^i_t}\right|, \end{aligned} \tag{8}$$

Finally, the WEQ effectively adjusts the class-wise weight and bias norm in current step similar to the average norm of those in the previous step during training.

$$\begin{aligned} \mathcal{R}_{WEQ} = \lambda_{WEQ}\frac{1}{C^t}\sum_{i=1}^{C_t}(&|\|\mathbf{W}^{s}_{\theta^i_t}\|_1 - \|\mathbf{W}^{s-1}_{\theta_t}\|_1\|_2 \\ &+ |\|\mathbf{b}^{s}_{\theta^i_t}\|_1 - \|\mathbf{b}^{s-1}_{\theta_t}\|_1\|_2), \end{aligned} \tag{9}$$

$\mathcal{R}_{WEQ}$ penalizes the extreme change of average norm of weights and bias in current classification head over a time. It enables the model to preserve the previously learned knowledge by correcting bias toward new classes to reduce inaccurate predictions without forgetting. In addition, the variation of weight norm becomes small over all classes.

Overall Training Framework

Fig.3 describes our proposed DFCIL framework. Before starting a new incremental task, we recover the old data following the model inversion-based (Smith et al. 2021) strategy, synthesizing old samples given the model learned up to the previous task without accessing previously observed data. Synthesizing the same number old samples with real data let the model preserve old knowledge during learning new information. First, we adopt the local classification loss, $\mathcal{L}_{LCE}$, which is the cross-entropy loss computed on only the new classification head with real data. $\mathcal{L}_{LCE}$ helps the model classify newly observed classes by increasing the plasticity. Second, we simultaneously forward the combined dataset to the previous and current model to perform our proposed knowledge distillation strategy for improving the stability of model. Third, we fine-tune the classification head of the current model with all observed data after freezing the feature extraction layer. We formulate fine-tuning loss as:

$$\mathcal{L}_{FT} = \sum_{k=1}^{C^t} -\delta_{y=k}\log(p_k(\mathbf{x}^k)). \tag{10}$$

During fine-tuning, we only optimize the current classification head to obtain the ideal decision boundary between the real and old classes, using global cross-entropy loss that computes the loss on all seen data. Lastly, we correct the bias towards the new classes via WEQ to alleviate the class imbalance problem without forgetting knowledge. Our final objective function is summarized as:

$$\begin{aligned} \min_{\mathcal{F}_t,\theta_t} \mathcal{L}_{LCE}(\mathbf{x}^{new}_t, \mathbf{y}^{new}_t) + \mathcal{L}_{KD}(\mathbf{x}_t) \\ + \mathcal{L}_{FT}(\mathbf{x}_t, \mathbf{y}_t) + \mathcal{R}_{WEQ} \end{aligned} \tag{11}$$

Experiment

In this section, we conduct various experiments to validate the effectiveness of Integrative Solution for Catastrophic Forgetting (ISCF). Especially, we evaluate the final performance of our proposed framework on CIFAR-100 (Krizhevsky, Hinton et al. 2009) and Tiny-ImageNet (Le and Yang 2015). Our experiments are divided into two folds, (1) comparison to other methods on benchmark dataset and (2) analysis of our proposed methods. Our experiments settings can be found in appendix.

Proposed DFCIL Framework Performance

For comparison with other methods, we follow the protocol which continuously trains all 100 classes in several splits, N including 5, 10, and 20 incremental steps. Following prior works (Rebuffi et al. 2017; Hou et al. 2019), we train the

Data-Free	**Task (A_N, %)**			**Data-Access**	**Task (A_N, %)**		
Method	5	10	20	Method	5	10	20
U.B	69.9 ± 0.2			U.B	69.9 ± 0.2		
Base	16.4 ± 0.4	8.8 ± 0.1	4.4 ± 0.3	Rehearsal	34.0 ± 0.2	24.0 ± 1.0	14.9 ± 0.7
LwF	17.0 ± 0.1	9.2 ± 0.0	4.7 ± 0.1				
DGR	14.4 ± 0.4	8.1 ± 0.1	4.1 ± 0.3	LwF	39.4 ± 0.3	27.4 ± 0.8	16.6 ± 0.4
LwF w/syn	16.7 ± 0.1	8.9 ± 0.0	4.7 ± 0.0				
DI	18.8 ± 0.1	10.9 ± 0.6	5.7 ± 0.3	BiC	53.7 ± 0.4	45.9 ± 1.8	37.5 ± 3.2
ABD	43.9 ± 0.9	33.7 ± 1.2	20 ± 1.4				
ISCF (Ours)	$\mathbf{47.56 \pm 0.15}$	$\mathbf{39.01 \pm 0.08}$	$\mathbf{21.24 \pm 1.84}$	**ISCF (Ours)**	$\mathbf{47.56 \pm 0.15}$	$\mathbf{39.01 \pm 0.08}$	$\mathbf{21.24 \pm 1.84}$

Table 1: Top-1 accuracy result (%) for data-free (DFCIL) and Data-Access class-incremental learning (DACIL) on CIFAR-100 for 5, 10, and 20-task experiments. According to the replay data type, we divide the DFCIL methods into three groups. ('None' includes Base, LwF , 'Generator' includes DGR and 'Model Inversion' include other methods). DACIL methods store 2k real images as coreset. All the results are reported as an average of 3 runs.

Data-Free	A_N **(%)**	**Data-Access**	A_N **(%)**
Upper Bound	73.6	Upper Bound	73.6
Base	4.1	Naive Rehearsal	6.6
LwF	4.4	LwF	6.9
LwF w/ MC	8.8		
LwF w/ syn	4.75	EEIL	16.9
DI	5.1	BiC	17.4
ABD	12.1		
ISCF (Ours)	**16.68**	**ISCF (Ours)**	**16.68**

Table 2: Top-1 accuracy results (%) for DFCIL and DACIL on Tiny-ImageNet for 20-task (5 classes per each task). According to the replay data type, we divide the DFCIL methods into two groups ('None', and 'Model Inversion') using row line.

ResNet-32 (He et al. 2016) on benchmark dataset. To evaluate the model performance, we utilize the class incremental learning accuracy, A_t with full test set defined by:

$$\mathcal{A}_t = \frac{1}{\mathcal{D}_t^{test}} \sum_{(x,y)\in\mathcal{D}_t^{test}} \mathbf{1}(\hat{y} = y), \qquad \hat{y} = \arg\max \theta_t(\mathcal{F}_t(x)) \tag{12}$$

where the **1** means the indicator function. From above metric, we can evaluate methods by each task, t. We report the final task accuracy of each methods denoted as A_N in extensive experiments.

Evaluation of CIFAR-100. Table.1 summarizes the final task accuracy on CIFAR-100 reporting averages and standard deviations. The result validates that our proposed method achieves high performance over other methods by a considerable margin in different incremental steps. Specifically, our proposed method outperforms ABD, the state-of-the-art method for DFCIL, up to 3.66%, 5.31% and 1.24% on 5,10, and 20 task respectively. Moreover, our approach achieves comparable performance with Data-Access Class Incremental Learning (DACIL) methods. DACIL methods follow the rehearsal strategy of storing the 2k coreset real images of old classes, and then train the model by continuously accessing real old images in buffer. Despite not storing any real data, the performance of ISCF is even better than Naive Rehearsal, LwF, and has comparable performance compared to BiC (Wu et al. 2019), which has the best performance. These results indicate that our method contributes to significant accuracy improvement in existing DFCIL methods, even decreasing the performance gap with DACIL method.

Evaluation of Tiny-ImageNet. To further validate our proposed method, we compare DFCIL and DACIL methods with ISCF on Tiny-ImageNet, which is a more challenging dataset to train. We use the same settings with CIFAR-100 experiments but only perform the validation test under 20 tasks, learning five classes per task. We further compare the LwF.MC (Rebuffi et al. 2017) and EEIL (Castro et al. 2018) well-known for achieving higher performance on the large-scale dataset. The results are reported in Table.2. We can observe that our method consistently surpasses previous DFCIL methods by a large margin and achieves comparable performance with the DACIL approaches. In particular, our method outperforms ABD by 4.58% incremental accuracy, and the small gap with EEIL and BiC from ISCF, which is about 0.22% and 0.72% respectively. Through these results, we consistently prove that our method has verifies the superiority of our proposed methods.

Analysis of ISCF

We perform several studies to verify our approach alleviate the catastrophic forgetting. As part of our experiments, we conduct (a) analyzing the effect of ISCF (b) ablation study to show the effect of each crucial component in ISCF, and (c) comparison of incremental learning accuracy curve.

The Effect of ISCF Framework. Our goal is to alleviate the catastrophic forgetting problem by sufficiently transferring old knowledge and correcting the bias toward new categories. To validate our achievement, we conduct two experiments on CIFAR-100. Firstly, we calculate the weight norm of the classification head and plot them with other methods to analyze the effectiveness of *Weight Equalizer* (WEQ). As shown in Fig.2 (a-c), the norm of weight corresponding to new classes in prior works is even larger than those for old classes. This phenomenon causes prediction bias towards

Method	Loss components	5task (A_N)	10 task (A_N)
Full method (w/ WEQ)	$\mathcal{L}_{LCE} + \mathcal{L}_{SPKD} + \mathcal{L}_{LKD} + \mathcal{L}_{FT} + \mathcal{R}_{WEQ}$	**47.54**	**39.07**
Baseline w/ SPKD + LKD	$\mathcal{L}_{LCE} + \mathcal{L}_{SPKD} + \mathcal{L}_{LKD} + \mathcal{L}_{FT}$	46.94	37.02
Baseline w/ SPKD + FKD	$\mathcal{L}_{LCE} + \mathcal{L}_{FKD} + \mathcal{L}_{SPKD} + \mathcal{L}_{FT}$	44.61	34.31
Baseline w/ FKD	$\mathcal{L}_{LCE} + \mathcal{L}_{FKD} + \mathcal{L}_{FT}$	43.9	33.7

Table 3: Ablation Study of each component in ISCF on CIFAR-100 with ResNet-32. We report the Top-1 accuracy result (%) for each task (5 and 10).

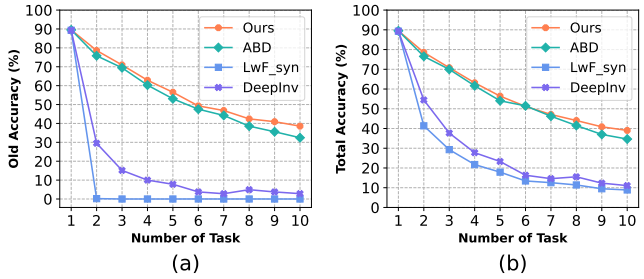

Figure 4: Top-1 Accuracy (%) curves on CIFAR-100 for 10 task compared to several DFCIL methods. The performance on only old classes (a) and all classes (b). Ours shows significant accuracy improvement for old classes over ABD, DeepInversion, LwF with synthetic images. 6.13%, 35.73%, 38.56% respectively.

new classes, which can harm the classification of old classes. In contrast, WEQ attempts to balance the parameter uniformly on overall classes following the previous step's scale. Fig.2 (d) shows that WEQ effectively corrects the bias in linear head. Also, our KD method help the model maintain the semantic information of both old and new classes. As a result, Fig.1 (d) exhibit superiority of ISCF that performs great on all classes compared to the previous approaches in Fig.1 (a-c), which show a strong bias towards the newest ten classes in confusion matrix.

Ablation Study. we conduct the ablation study to understand how each component affects our overall method by adding one component to the baseline model. We evaluate the final performance on the 5 and 10 incremental steps in CIFAR-100 independently for more specific analysis. We use the ABD (Smith et al. 2021) as a baseline model. They proposes $\mathcal{L}_{FKD}$, which is the important weighted-based feature distillation. They extract the output features of $\mathcal{F}_t(\mathbf{x})$ and $\mathcal{F}_{t-1}(\mathbf{x})$ respectively, then constrain $\theta_{t-1}(\mathcal{F}_t(\mathbf{x}))$ to keep $\theta_{t-1}(\mathcal{F}_{t-1}(\mathbf{x}))$ weighted by previous classification head for preserving important weight. We add the $\mathcal{L}_{SPKD}$ loss from the baseline model and then replace the $\mathcal{L}_{FKD}$ with our proposed KD strategy, both $\mathcal{L}_{SPKD}$ and $\mathcal{L}_{LKD}$. Lastly, we show the final performance of our proposed method by adding the $\mathcal{R}_{WEQ}$. Table.3 reports the result of top-1 accuracy on CIFAR-100 for 5 and 10-task. We find out that each component has different impacts on the final performance according to experimental settings. Since SPKD transfers the previously learned semantic feature to the current model, more significant knowledge can be extracted from the sufficiently learned model. Therefore, SPKD achieves a larger performance improvement in a 5-task experiment than 10-task. On the contrary, WEQ improves performance on 10-task incremental learning since the bias is more extreme due to overfitting caused by a small portion of the training set. To sum up, we typically show that the performance of the model is further improved with 3.64% and 5.37% from baseline in 5 and 10-task respectively when using all components.

Incremental Learning Accuracy Curve on Overall Tasks. To show how much knowledge of old classes in each incremental step affects the overall performance, we report the top-1 accuracy curve with old classes and all classes. We conduct 10-task experiments on CIFAR-100 with previous DFCIL methods. Fig.4 describes the top-1 accuracy on the old and all seen classes. For example, the accuracy in second task reports the performance on old classes (0-10) in Fig.4 (a) and total classes (0-19) in Fig.4 (b). We observe that the top-1 accuracy gap of old classes between ISCF and other methods increases as the task progresses. At the last task, the gap between ISCF and other methods (e.g., ABD, DeepInversion and LwF-synthetic) are about 6.13%, 35.73%, and 38.56%; thus, we can see that our proposed method handles catastrophic forgetting superior to the other methods. In addition, preserving previously learned knowledge also affects to the overall performance, thereby our approach outperforms other methods in total accuracy performance as well. The final task accuracy of each method is specified in the 10 task result on CIFAR-100 in table 1.

Conclusion

In this paper, we empirically observe that existing Data-Free Class Incremental Learning (DFCIL) works still suffer from catastrophic forgetting. Therefore, we suggest a practical framework that alleviates catastrophic forgetting. First, we propose *Weight Equalizer* that minimizes class imbalance, and corrects bias towards new classes in the classification head. Second, we propose the combined knowledge distillation strategy consisting of *Similarity Preserving Knowledge Distillation* and *Logit Knowledge Distillation* for preserving the old information by transferring the additional guidance and constraint the output probability. As a result, we achieve state-of-the-art performance over previous DFCIL methods on CIFAR-100 and TinyImageNet. We also demonstrate the effectiveness of our proposed method via experimental analysis. We believe this work gives a promising direction for overcoming catastrophic forgetting in DFCIL scenario.

References

Ahn, H.; Kwak, J.; Lim, S.; Bang, H.; Kim, H.; and Moon, T. 2021. Ss-il: Separated softmax for incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 844–853.

Castro, F. M.; Marín-Jiménez, M. J.; Guil, N.; Schmid, C.; and Alahari, K. 2018. End-to-end incremental learning. In *Proceedings of the European conference on computer vision (ECCV)*, 233–248.

Cong, Y.; Zhao, M.; Li, J.; Wang, S.; and Carin, L. 2020. Gan memory with no forgetting. *Advances in Neural Information Processing Systems*, 33: 16481–16494.

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.

Douillard, A.; Cord, M.; Ollion, C.; Robert, T.; and Valle, E. 2020. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *European Conference on Computer Vision*, 86–102. Springer.

He, C.; Wang, R.; and Chen, X. 2021. A tale of two cils: The connections between class incremental learning and class imbalanced learning, and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3559–3569.

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.; Dean, J.; et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7).

Hou, S.; Pan, X.; Loy, C. C.; Wang, Z.; and Lin, D. 2019. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 831–839.

Kim, C. D.; Jeong, J.; and Kim, G. 2020. Imbalanced continual learning with partitioning reservoir sampling. In *European Conference on Computer Vision*, 411–428. Springer.

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

Liu, Y.; Parisot, S.; Slabaugh, G.; Jia, X.; Leonardis, A.; and Tuytelaars, T. 2020. More classifiers, less forgetting: A generic multi-classifier paradigm for incremental learning. In *European Conference on Computer Vision*, 699–716. Springer.

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.

Parisi, G. I.; Kemker, R.; Part, J. L.; Kanan, C.; and Wermter, S. 2019. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113: 54–71.

Prabhu, A.; Torr, P. H.; and Dokania, P. K. 2020. Gdumb: A simple approach that questions our progress in continual learning. In *European conference on computer vision*, 524–540. Springer.

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.

Shin, H.; Lee, J. K.; Kim, J.; and Kim, J. 2017. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30.

Silver, D. L.; Yang, Q.; and Li, L. 2013. Lifelong machine learning systems: Beyond learning algorithms. In *2013 AAAI spring symposium series*.

Smith, J.; Hsu, Y.-C.; Balloch, J.; Shen, Y.; Jin, H.; and Kira, Z. 2021. Always be dreaming: A new approach for data-free class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9374–9384.

Tung, F.; and Mori, G. 2019. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1365–1374.

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

Wu, C.; Herranz, L.; Liu, X.; van de Weijer, J.; Raducanu, B.; et al. 2018. Memory replay gans: Learning to generate new categories without forgetting. *Advances in Neural Information Processing Systems*, 31.

Wu, Y.; Chen, Y.; Wang, L.; Ye, Y.; Liu, Z.; Guo, Y.; and Fu, Y. 2019. Large scale incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 374–382.

Xiang, Y.; Fu, Y.; Ji, P.; and Huang, H. 2019. Incremental learning using conditional adversarial networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6619–6628.

Yin, H.; Molchanov, P.; Alvarez, J. M.; Li, Z.; Mallya, A.; Hoiem, D.; Jha, N. K.; and Kautz, J. 2020. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8715–8724.

Zhao, B.; Xiao, X.; Gan, G.; Zhang, B.; and Xia, S.-T. 2020. Maintaining discrimination and fairness in class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13208–13217.

Zhou, P.; Mai, L.; Zhang, J.; Xu, N.; Wu, Z.; and Davis, L. S. 2019. M2kd: Multi-model and multi-level knowledge distillation for incremental learning. *arXiv preprint arXiv:1904.01769*.