Adversarial Robust Memory-Based Continual Learner

Xiaoyue Mi[1,2], Fan Tang[1,2]*, Zonghan Yang[3,4], Danding Wang[1,2], Juan Cao[1,2], Peng Li[3], Yang Liu [3,4]
[1]Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]Institute for AI Industry Research (AIR), Tsinghua University
[4]Department of Computer Science & Technology, Tsinghua University

Abstract

Despite the remarkable advances that have been made in continual learning, the adversarial vulnerability of such methods has not been fully discussed. We delve into the adversarial robustness of memory-based continual learning algorithms and observe limited improvement in robustness by directly applying adversarial training techniques. Our preliminary studies reveal the twin challenges for building adversarial robust continual learners: ***accelerated forgetting*** *in continual learning and* ***gradient obfuscation*** *in adversarial robustness. In this study, we put forward a novel adversarial robust memory-based continual learner that adjusts data logits to mitigate the forgetting of pasts caused by adversarial samples. Furthermore, we devise a gradient-based data selection mechanism to alleviate the gradient obfuscation caused by limited stored data. The proposed approach can widely integrate with existing memory-based continual learning and adversarial training algorithms. Extensive experiments on Split-CIFAR10/100 and Split-TinyImageNet demonstrate the effectiveness of our approach, achieving a maximum forgetting reduction of* 34.17% *in adversarial data for ResNet, and* 20.10% *for ViT.*

1. Introduction

Continual learning [48, 62] enables intelligent models to continually acquire new knowledge and adapt to changing environments while maintaining previous capabilities. It has recently emerged as a promising direction to achieve ideal intelligent models, and some pioneer works have investigated the robustness of continual learning in various aspects [42, 53, 56]. Achieving robust continual learners represents a crucial advancement in enabling intelligent models to adapt to dynamic scenarios effectively.

However, when dealing with streaming non-*i.i.d.* data, continual learners are usually vulnerable to adversarial examples [23]. Chen *et al*. [12] first attempt at preserving adversarial robustness in continual learning [29] by leveraging adversarial training [52, 57] and a large amount of unlabeled data. Despite the impressive results of Chen *et al*. [12], the inherent challenges for conducting adversarial robust continual learners are still under disclosure.

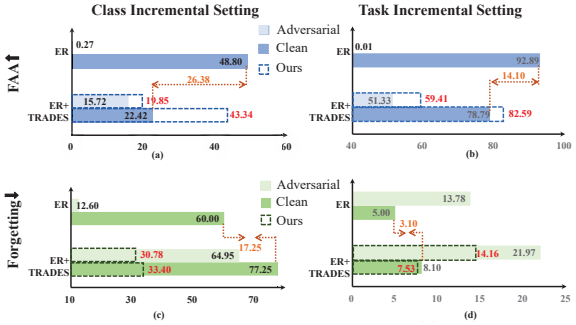

Figure 1. A preliminary analysis of the combination of continual learning and adversarial training on Split-CIFAR10 with buffer size 200. We consider the final average accuracy (FAA) and forgetting of adversarial and clean samples in both class incremental and task incremental settings. The red numbers indicate the performance after adding our approach.

In this study, we investigate building adversarial robust continual learners without abundant unlabeled data. We select two representative approaches for continual learning and adversarial defenses: memory-based continual learning and adversarial training. Figure 1 shows the direct combination of memory-based continual learning (ER [10]) with an adversarial training approach (TRADES [57]). The results for adversarial data reconfirm that continual learners risk adversarial attacks [12, 20, 23, 24], and demonstrate a critical trade-off: while adversarial training can strengthen continual learners against attacks [12, 20, 23, 24], it simultaneously increases forgetting on clean data, thus compromising the core objective of continual learning.

Further experimental analysis in Sec. 3 reveals two key

*Corresponding author

challenges when combining adversarial training with continual learning: (1) Adversarial data increase negative gradients toward past classes compared with clean data, accelerating catastrophic forgetting in continual learning; (2) Insufficient historical data in continual learning leads to gradient obfuscation [4], compromising the effectiveness of adversarial training compared to joint training. These challenges emerge naturally from the direct integration of the two approaches.

Based on the analysis, we propose an adversarial robust memory-based continual learner without a large amount of additional data, which consists of an anti-forgettable logit calibration (AFLC) module and a robustness-aware experience replay (RAER) strategy. The AFLC module designs different logit calibration strategies for past, current, and future classes, thus mitigating the negative gradients of adversarial samples for continual learning. By leveraging attacking difficulty, RAER selects data that are adversarial safety and maintains the diversity of data distribution, rather than storing data overfitting to the decision boundary [9]. Our approach can be combined with most memory-based continual learning as well as adversarial training algorithms, and we have demonstrated the effectiveness of our approach on the Split-CIFAR10/100 and more challenging dataset Split-Tiny-ImageNet across ResNet and ViT architectures.

Our contributions can be summarized as follows:

- We reveal the twin critical challenges to carrying out adversarial robust memory-based continual learning algorithms: accelerated forgetting for continual learning and gradient obfuscation for adversarial robustness.
- We propose anti-forgettable logit calibration to reduce negative influences from adversarial training through task-adaptive logit calibration, and robustness-aware experience replay by storing robust and distribution-representative past data to alleviate gradient obfuscation.
- Experiments on Split-CIFAR10/100 and Split-Tiny-ImageNet across ResNet and ViT demonstrate our effectiveness, in improving both clean accuracy and adversarial robustness of continual learning and mitigating forgetting in both class-incremental and task-incremental settings without additional data.

2. Related Works

Continual learning. Continual learning [38, 48, 62] enables models to adapt to evolving data distributions while preserving past knowledge. To alleviate catastrophic forgetting, existing approaches [34, 39] can be categorized into three types: memory-based [7, 8, 10], regularization-based [25, 29, 30], and dynamic architecture [17, 41, 50, 55, 61]. Continual learning can be categorized into class-incremental (class-IL) and task-incremental (task-IL) settings [33] based on the availability of task information during inference. In both settings, memory-based methods have demonstrated superior performance without requiring model expansion [7, 60]. Researchers have explored continual learning robustness under various constraints [19, 53, 56], and saft risks about backdoor attacks [1, 22, 47] and privacy concerns [21]. Recent studies have revealed both the vulnerability of continual models to adversarial attacks and the potential of adversarial samples for mitigating forgetting [20, 23, 24, 27, 49]. Chen *et al*. [12] combined LwF [29] with adversarial training using additional data, while our work uniquely analyzes the core challenges in achieving adversarial robustness for continual learning without such requirements.

Adversarial defense. The vulnerability of deep neural networks to adversarial examples [18, 46, 52, 59] has led to numerous defense methods [2, 5], with adversarial training emerging as the most effective approach [14, 15, 28, 31, 32, 36, 37, 57]. Recent research has extended to real-world scenarios, such as long-tailed distributions [54] and open-world settings [43]. Meanwhile some studies show the potential of continual algorithms in adapting to new attacks [13, 40]. Different from them, we specifically address adversarial robustness in class-incremental and task-incremental settings.

3. Baselines and Their Inherent Problems

3.1. Preliminaries

Memory-based continual learning. Following [6, 7], we consider continual learning across T distinct classification tasks. We pre-allocate classification heads for all tasks at initialization rather than expanding them incrementally. For the t-th task ($1 \leq t \leq T$), we denote its data distribution as D_t, with input samples x_t and corresponding labels y_t. A model $\mathrm{M}_\theta(\cdot)$ with parameters θ is optimized sequentially in the given task order. Let $h_\theta(x_t) \in \mathbb{R}^C$ denote the output logits where $h_\theta(x_t) = \mathrm{M}_\theta(x_t)$, and C is the total number of classes. The softmax output is defined as $f_\theta(x_t) \triangleq \mathrm{softmax}(h_\theta(x_t))$. Given a logit vector $h_\theta(x_t) = [h_\theta(x_t)_1, ..., h_\theta(x_t)_C]$ where $h_\theta(x_t)_i$ represents the logit for the i-th class, we partition the classes into three sets based on the current task t:

$$\begin{aligned} p =& \{1, 2, ..., (t-1)b\}, \\ c =& \{(t-1)b+1, (t-1)b+2, ..., tb\}, \\ u =& \{tb+1, tb+2, ..., C\}, \end{aligned} \tag{1}$$

where b is the number of classes per task. Accordingly, we divide $h_\theta(x_t)$ into three parts: $h_\theta(x_t)_p$, $h_\theta(x_t)_c$, and $h_\theta(x_t)_u$, representing logits for past, current, and future classes respectively.

Continual learner aims to correctly classify data of all

Buffer Size	Method	W/ AT	Class Incremental Setting				Task Incremental Setting				CRD↓	FRI↓	RRD↓
			Clean Data		Adversarial Data		Clean Data		Adversarial Data				
			FAA↑	Forgetting↓	FAA↑	Forgetting↓	FAA↑	Forgetting↓	FAA↑	Forgetting↓			
-	**Joint**	✗	$91.81_{\pm 0.07}$	-	$0.00_{\pm 0.00}$	-	$98.17_{\pm 0.12}$	-	$0.00_{\pm 0.00}$	-	7.92	-	-
	(upper bound)	✓	$79.33_{\pm 0.36}$	-	$50.93_{\pm 1.28}$	-	$94.80_{\pm 0.30}$	-	$74.63_{\pm 1.13}$	-			
-	**SGD**	✗	$19.66_{\pm 0.07}$	$96.39_{\pm 0.67}$	$5.17_{\pm 2.58}$	$33.31_{\pm 5.19}$	$66.84_{\pm 6.50}$	$37.42_{\pm 7.39}$	$4.44_{\pm 3.78}$	$33.34_{\pm 3.49}$	0.91	−3.73	41.10
	(lower bound)	✓	$19.34_{\pm 0.05}$	$91.93_{\pm 0.20}$	$15.65_{\pm 0.18}$	$69.71_{\pm 1.77}$	$65.33_{\pm 5.91}$	$34.43_{\pm 7.16}$	$27.71_{\pm 1.88}$	$54.56_{\pm 3.47}$			
200	**ER**	✗	$48.80_{\pm 1.84}$	$60.00_{\pm 2.25}$	$0.27_{\pm 0.18}$	$12.60_{\pm 2.13}$	$92.89_{\pm 0.26}$	$5.00_{\pm 0.26}$	$0.01_{\pm 0.01}$	$13.78_{\pm 0.41}$	14.51	12.91	31.70
		✓	$28.18_{\pm 0.69}$	$80.58_{\pm 1.05}$	$17.86_{\pm 0.29}$	$69.58_{\pm 1.15}$	$84.49_{\pm 0.61}$	$10.23_{\pm 0.98}$	$44.30_{\pm 1.05}$	$36.89_{\pm 1.11}$			
5120		✗	$83.34_{\pm 1.38}$	$15.35_{\pm 1.71}$	$0.08_{\pm 0.11}$	$13.65_{\pm 2.34}$	$96.84_{\pm 0.17}$	$0.71_{\pm 0.19}$	$0.06_{\pm 0.04}$	$21.83_{\pm 8.69}$	13.53	12.11	18.80
		✓	$61.88_{\pm 0.74}$	$37.72_{\pm 0.73}$	$27.28_{\pm 0.56}$	$41.66_{\pm 1.22}$	$91.24_{\pm 0.19}$	$2.56_{\pm 0.12}$	$56.59_{\pm 0.88}$	$19.34_{\pm 1.11}$			
200	**DER**	✗	$62.71_{\pm 2.76}$	$38.75_{\pm 4.11}$	$0.01_{\pm 0.00}$	$7.42_{\pm 2.13}$	$91.32_{\pm 0.83}$	$6.95_{\pm 1.44}$	$0.09_{\pm 0.05}$	$7.39_{\pm 2.18}$	17.72	15.73	34.80
		✓	$35.79_{\pm 2.85}$	$65.67_{\pm 4.16}$	$16.44_{\pm 0.21}$	$52.84_{\pm 3.71}$	$82.79_{\pm 0.21}$	$11.48_{\pm 0.66}$	$39.52_{\pm 0.37}$	$40.64_{\pm 1.38}$			
5120		✗	$83.99_{\pm 0.99}$	$10.57_{\pm 1.51}$	$0.07_{\pm 0.07}$	$7.05_{\pm 2.03}$	$95.67_{\pm 0.13}$	$1.63_{\pm 0.13}$	$0.05_{\pm 0.05}$	$7.39_{\pm 0.48}$	16.74	15.66	24.32
		✓	$56.58_{\pm 3.87}$	$40.40_{\pm 4.07}$	$23.26_{\pm 0.44}$	$45.06_{\pm 1.48}$	$89.62_{\pm 0.67}$	$3.12_{\pm 0.23}$	$53.66_{\pm 2.01}$	$22.95_{\pm 3.68}$			
200	**DER++**	✗	$65.55_{\pm 2.81}$	$32.17_{\pm 4.26}$	$0.24_{\pm 0.32}$	$7.38_{\pm 3.37}$	$92.11_{\pm 0.55}$	$8.85_{\pm 2.98}$	$0.22_{\pm 0.33}$	$10.45_{\pm 0.94}$	14.47	8.24	37.14
		✓	$45.43_{\pm 3.28}$	$46.67_{\pm 5.09}$	$14.46_{\pm 1.35}$	$35.09_{\pm 2.70}$	$83.29_{\pm 1.07}$	$10.83_{\pm 1.77}$	$36.82_{\pm 0.59}$	$34.09_{\pm 2.23}$			
5120		✗	$85.74_{\pm 0.65}$	$7.32_{\pm 0.88}$	$1.51_{\pm 1.27}$	$6.10_{\pm 0.53}$	$96.13_{\pm 0.53}$	$1.14_{\pm 0.35}$	$0.82_{\pm 0.52}$	$6.06_{\pm 0.49}$	10.63	8.21	22.18
		✓	$68.67_{\pm 1.11}$	$23.68_{\pm 2.66}$	$26.95_{\pm 0.29}$	$27.68_{\pm 2.48}$	$91.95_{\pm 0.33}$	$1.20_{\pm 0.90}$	$54.26_{\pm 0.41}$	$16.87_{\pm 1.72}$			
200	**X-DER**	✗	$60.96_{\pm 1.09}$	$12.29_{\pm 0.09}$	$0.00_{\pm 0.00}$	$3.84_{\pm 1.18}$	$94.33_{\pm 0.63}$	$2.49_{\pm 0.17}$	$0.13_{\pm 0.03}$	$2.59_{\pm 0.64}$	20.23	7.65	24.02
		✓	$34.04_{\pm 0.90}$	$25.13_{\pm 4.03}$	$16.82_{\pm 0.98}$	$27.84_{\pm 4.16}$	$80.8_{\pm 0.66}$	$4.96_{\pm 1.66}$	$60.83_{\pm 0.56}$	$10.99_{\pm 0.81}$			
5120		✗	$69.93_{\pm 0.63}$	$4.33_{\pm 0.26}$	$0.21_{\pm 0.12}$	$4.52_{\pm 1.39}$	$96.94_{\pm 0.02}$	$0.29_{\pm 0.04}$	$0.13_{\pm 0.02}$	$2.04_{\pm 0.43}$	24.96	18.52	21.27
		✓	$34.29_{\pm 0.65}$	$38.92_{\pm 0.03}$	$19.24_{\pm 0.10}$	$38.79_{\pm 0.41}$	$82.67_{\pm 1.68}$	$2.74_{\pm 0.83}$	$64.11_{\pm 0.12}$	$8.14_{\pm 0.98}$			

Table 1. We conduct a systematic experimental analysis combining adversarial training (specifically Vanilla AT [31]) with memory-based continual learning algorithms. See Sec. 3.2 for a detailed summary of observations. Performance is evaluated using several metrics (detailed in Sec. 5.1): Final Average Accuracy (FAA), Forgetting, Clean Relative Decrease (CRD), Forgetting Relative Increase on clean data (FRI), and Robust Relative Decrease (RRD) compared to joint training. All adversarial samples are generated using PGD-20. Note that the reported CRD, FRI, and RRD values are averaged across both task-incremental and class-incremental settings.

observed tasks, and its objective can be formalized as:

$$\arg\min_{\theta} \sum_{t=1}^{T} \mathcal{L}_t, \quad \text{where} \quad \mathcal{L}_t \triangleq \mathbb{E}_{(x_t,y_t)\sim D_t}[\ell(f_\theta(x_t), y_t)], \tag{2}$$

where ℓ is the task-specific loss. The unavailability of past task data poses great challenges for continual learning. However, direct sequential optimization can lead to severe catastrophic forgetting. To address the above challenge, memory-based continual learners usually store a few past task data and replay them during the current task training phase [7], and become leading-edge approaches. When training the t-th task, its loss function can be formalized as the following:

$$\mathcal{L}_t \triangleq \mathbb{E}_{(x,y)\sim D_t \cup \mathcal{M}}[\ell(f_\theta(x), y)], \tag{3}$$

where $\mathcal{M}$ is the episodic memory of the current task, which preserves a small subset of observed past data $\langle x_\mathcal{M}, y_\mathcal{M} \rangle$.

Adversarial training. Adversarial training is currently the most reliable way [5] to improve the adversarial robustness of a model. Following [31], it can be formally defined as a min-max optimization problem:

$$\arg\min_{\theta} \mathbb{E}_{(x,y)\sim D} \left(\max_{\delta \in S} \ell(f_\theta(x+\delta), y) \right), \tag{4}$$

where δ is an adversarial perturbation, D is data distribution, and S is a perturbation set. $S = \left\{ \delta | \left\| \delta \right\|_p \leq \epsilon \right\}$, ϵ is the perturbation size. The goal of adversarial training is to correctly classify all adversarial examples ($\widetilde{x} = x + \delta$).

Here we integrate memory-based continual learner and adversarial training as adversarial robust memory-based continual learner baselines, and the new optimization objective during the training stage of t-th task is:

$$\arg\min_{\theta} \mathbb{E}_{(x,y)\sim D_t \cup \mathcal{M}} \left(\max_{\delta \in S} \ell(f_\theta(x+\delta), y) \right). \tag{5}$$

3.2. Vanilla Solutions

To investigate the challenges in adversarial robust continual learning, we analyze several popular memory-based continual learning methods (ER [10], DER [7], DER++ [7], and X-DER [6]) on Split-CIFAR10 [26] under both class and task incremental settings. We employ Vanilla AT [31] for adversarial training due to its classic and strong performance [36]. Performance is evaluated against an upper bound ("Joint") and a lower bound ("SGD") using multiple metrics: Final Average Accuracy (FAA), forgetting, Clean Relative Decrease (CRD), Forgetting Relative Increase in clean data (FRI), and Robust Relative Decrease (RRD). Detailed metric definitions are provided in Sec. 5.1, and additional implementation details and hyper-parameters are provided in the supplementary material.

Results in Table 1 reveal that direct integration of adversarial training with continual learning is suboptimal. First, the results validate our observations from Sec. 1: while

adversarial training enhances robustness, it compromises clean performance (positive CRDs) and accelerates forgetting (positive FRIs). Larger memory sizes consistently improve performance across all metrics. Moreover, continual models show inferior robustness improvement compared to joint models (positive RRDs).

To investigate the underlying mechanisms, we select ER as our base model for further analysis. While DER/DER++ and X-DER demonstrate better stability with lower forgetting, they struggle with new task learning, resulting in poor robust FAA. In contrast, ER combined with adversarial training exhibits strong robustness.

3.3. The Inherent Problems

By analyzing gradients during training, we identify two critical challenges in robust continual learning: accelerated forgetting in continual learning and gradient obfuscation in adversarial training.

Accelerated forgetting in continual learning. Adversarial training hastens the forgetting of clean data in continual learners. According to [31], adversarial examples are designed to mislead model predictions by increasing probabilities for incorrect classes while suppressing those of the true class. In the continual learning setting, adversarial examples from the current task tend to increase probabilities for past classes while decreasing probabilities for their true class. Specifically, for an adversarial example $\widetilde{x}_t$ generated from a clean sample x_t of current task t:

$$f_\theta(\widetilde{x}_t)_p > f_\theta(x_t)_p, f_\theta(\widetilde{x}_t)_{\text{true class}} < f_\theta(x_t)_{\text{true class}}. \quad (6)$$

When using cross-entropy as the loss function, adversarial data amplify negative gradients for current data x_t for past tasks:

$$\left[\frac{\partial \ell}{\partial h_\theta(\widetilde{x}_t)}\right]_p = f_\theta(\widetilde{x}_t)_p > f_\theta(x_t)_p = \left[\frac{\partial \ell}{\partial h_\theta(x_t)}\right]_p. \quad (7)$$

Our empirical analysis reveals that these adversarial examples generate stronger negative gradients toward past tasks, especially during the initial training phases. In Figure 2, experiments with ER+AT on Split-CIFAR10 using a buffer size of 200 show adversarial data consistently produces larger gradient norms compared to clean data. The larger gradient norms indicate stronger positive gradients for the true class and stronger negative gradients for other classes, including past classes, thus accelerating forgetting of previously learned tasks. While memory-based methods using stored adversarial samples can help mitigate catastrophic forgetting, the limited buffer size (typically small in practical applications) means that adversarial training on the entire dataset still accelerates forgetting.

Gradient obfuscation in adversarial training. Gradient obfuscation [4] occurs when gradient information is obscured by non-differentiable operators or randomness, leading to inferior robustness. We observe a specific form of gradient obfuscation, termed "shattered gradients", in adversarial training under continual learning, resulting in less effective robustness compared to joint models.

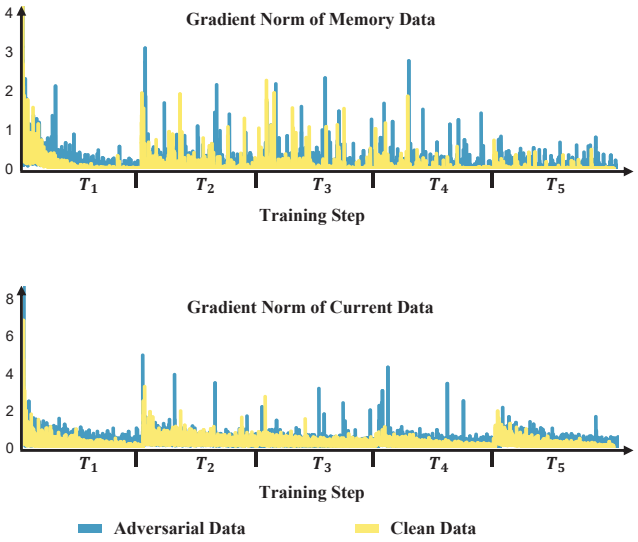

Figure 2. Gradient norms of adversarial data and clean data during the training stage. We experiment ER+AT with a 200 buffer size on the Split-CIFAR10 dataset.

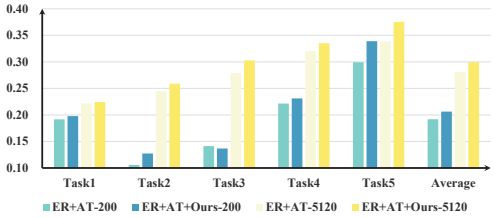

Figure 3. The gradient cosine similarities between robust joint model and robust continual model, which indicates a special gradient obfuscation only in continual learning. We experiment with buffer size=200/5,120 on the Split-CIFAR10 dataset. Gradient obfuscation weakens as the buffer size gets larger, and our method also mitigates gradient obfuscation.

In continual learning, gradient information is biased due to limited replay data. With few old task samples, the model tends to overfit, causing direction-biased gradients when generating adversarial examples. Our analysis of gradient directions between adversarial robust joint and continual models reveals significant differences, contributing to lower adversarial robustness in continual learners (see Figure 3).

4. Method

4.1. Overview

To address the identified challenges, we propose a novel framework consisting of two key components: anti-forgettable logit calibration (AFLC) to mitigate forgetting acceleration, and robustness-aware experience replay (RAER) to reduce gradient obfuscation. The full pipeline is illustrated in Figure 4: during training each task, the

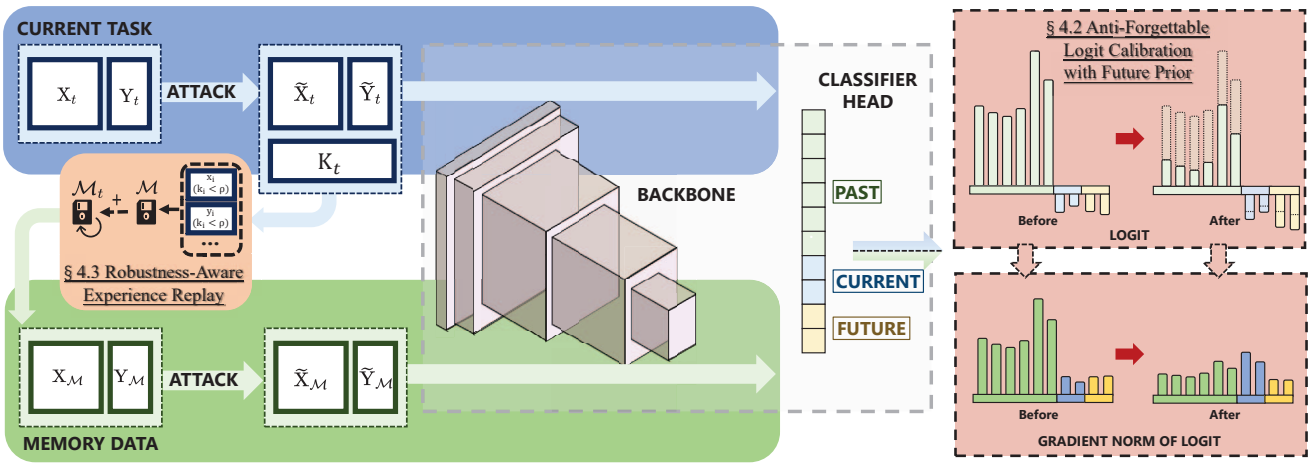

Figure 4. The pipeline of our adversarial robust memory-based continual learner. First, generate adversarial samples ($\widetilde{X}_t$, $\widetilde{X}_{\mathcal{M}}$) and robust difficulty factors (K_t) for both current task data (X_t, Y_t) and memory data ($X_{\mathcal{M}}, Y_{\mathcal{M}}$). The model then processes these samples through AFLC (Sec. 4.2), which calibrates logits based on task order to reduce negative gradients toward past tasks. Meanwhile, RAER (Sec. 4.3) leverages the robust difficulty factors K_t to select representative samples for memory storage, mitigating gradient obfuscation.

model processes both current task data and memory samples through the model. AFLC then calibrates logits based on task order (past/current/future) before loss computation. Simultaneously, RAER evaluates current task samples using a robustness difficulty factor k to select representative data for memory storage.

4.2. Anti-Forgettable Logit Calibration

We introduce AFLC to regulate the influence of adversarial examples on past task knowledge. The core idea is to calibrate logits differently for past and current tasks to control gradient flow.

Given an input x (either from current task x_t or memory buffer $x_{\mathcal{M}}$), AFLC applies a task-dependent calibration:

$$h_\theta^{\text{lc}}(x)_i = h_\theta(x)_i - v_i, v_p > v_c, \tag{8}$$

where v_i represents the calibration factor for class i:

$$v_i = -\log(\frac{n_i}{\sum_{j=1}^{C} n_j}) - \alpha_i, \tag{9}$$

where n_i denotes sample count for class i, and α_i is a hyperparameter. When all tasks share the classification head, we realize that making logit changes to past and current classes only is equivalent to negative gradient increases. So we make a **Further Prior (FP) adjustment** that essentially averages the adjustments applied to current and past classes and applies this value to future classes, helping to balance the gradients across all periods.

$$v_i = \frac{1}{t \cdot b} \sum_{j=1}^{t \cdot b} (v_j), \tag{10}$$

where t is the current task order and b is the number of classes per task.

The calibrated probability distribution is computed as:

$$f_\theta^{\text{lc}}(x)_i = \frac{\exp(h_\theta^{\text{lc}}(x)_i)}{\sum_{j=1}^{C} \exp(h_\theta^{\text{lc}}(x)_j)}, \tag{11}$$

and leads to two key properties: For current task samples,

$$\left[\frac{\partial \ell}{\partial h_\theta^{\text{lc}}(\widetilde{x})}\right]_p = f_\theta^{\text{lc}}(\widetilde{x}_t)_p < f_\theta(\widetilde{x}_t)_p, \tag{12}$$

and for memory buffer samples,

$$\left[\frac{\partial \ell}{\partial h_\theta^{\text{lc}}(\widetilde{x_{\mathcal{M}}})}\right]_c = f_\theta^{\text{lc}}(\widetilde{x}_{\mathcal{M}})_c > f_\theta(\widetilde{x}_{\mathcal{M}})_c. \tag{13}$$

In that way, AFLC can mitigate forgetting acceleration by reducing the negative gradient to past classes from current adversarial data and improving the negative gradient to current classes from memory adversarial data.

Comparisons with other logit calibration CL. X-DER [6] also adjusts the negative gradients of the new task data affecting the past tasks by updating the new task with the logit masking. Logit masking can be seen as a special case of our method at the softmax layer, which can be viewed as $\mathrm{v}_p = \mathrm{v}_c = 0$ for past data, and $\mathrm{v}_p = +\infty, \mathrm{v}_c = 0$ for current data. However, their extreme margin v inhibits the model's ability to learn new knowledge. X-DER further adds a safeguard threshold to limit the activations of past and future heads, which also can be combined with us.

4.3. Robustness-Aware Experience Replay

To address gradient obfuscation, we propose RAER, which selects samples based on their robustness characteristics. For each sample x, we measure its robustness difficulty factor k during PGD-10 attack, without additional forward or

backward calculation, where $k \in [0, 10] \cap \mathbb{N}$ indicates the number of successful attacks. Samples meeting the threshold condition $k < \rho$ are candidates for memory storage. The memory update process can be formalized as:

$$\mathcal{M} \leftarrow \text{Random Sample}(\{x | k_x < \rho\}, m), \tag{14}$$

where m is the memory size and ρ is set to 5 in practice.

Why RAER works. The robustness difficulty factor k shows how easily a sample can be attacked: a high k suggests vulnerability and proximity to decision boundaries, while a low k implies robustness and distance from boundaries. We selectively store samples with $k < \rho$ in memory, excluding boundary-overfitting outliers. Unlike prior continual learning [9] and adversarial training [58] methods, RAER emphasizes distribution-representative and robust learnable samples while maintaining model reliability.

The computational overhead of RAER is minimal, as k is calculated during the existing adversarial example generation process, without additional forward or backward calculation.

Comparisons with other data-selection CL. Existing works [3, 44] alleviate catastrophic forgetting using data selection and focus solely on clean samples. RAER uniquely targets gradient obfuscation in continual robust settings by: Selecting adversarially safe samples for better retention; excluding high-k boundary samples; and maintaining representative data distribution. As demonstrated in Sec. 5, RAER outperforms traditional data selection strategies combined with adversarial training, particularly in adversarial robustness.

4.4. Overall Objective

To sum up, we design AFLC to reduce negative gradients of the current adversarial sample to past tasks, and RAER to select data that alleviate gradient obfuscation to store. We set α of AFLC is 3.5 in practice and and $x_{\mathcal{M}}$ and $y_{\mathcal{M}}$ are selected by RAER with $\rho = 5$. When ER+AT combines with ours, the loss of task t is:

$$\mathcal{L}_t \triangleq \text{CE}(f_\theta^{lc}(\widetilde{x}_t), y_t) + \text{CE}(f_\theta^{lc}(\widetilde{x}_{\mathcal{M}}), y_{\mathcal{M}}). \tag{15}$$

5. Experiments

5.1. Experimental Settings

Datasets and training details. Following standard practices in adversarial training and continual learning [6, 37, 63], we evaluate our approach on Split-CIFAR10 and Split-CIFAR100 [26] and the model architecture is ResNet18, trained using SGD optimizer with learning rate 0.1. Split-CIFAR10 comprises ten classes ($5,000$ training, $1,000$ test samples per class) divided into five binary classification tasks. Split-CIFAR100 contains 100 classes (500 training, 100 test samples per class) split into ten-way classification tasks. **Evaluations on Split-Tiny-ImageNet [45] and ViT [16] in supplementary materials**.

For adversarial training, we use perturbations of $8/255$ and step size of $2/255$. Data augmentation includes random cropping (4-pixel padding) and horizontal flipping for all data. Unlike Chen *et al*. [12], we achieve robustness without additional unlabeled data[1]. All experiments are conducted with three different random seeds.

Evaluation metrics. We evaluate both adversarial robustness and continual learning performance using Final Average Accuracy (FAA) and forgetting metrics on clean and adversarial samples. For adversarial evaluation, we employ PGD-20 (white-box attack) and Auto Attack (AA) [14], which combines black-box and white-box attacks for reliable robustness assessment [36, 37, 51]. **Extended robustness verification using additional black-box (RayS [11]) and adaptive attacks in the supplementary material.** Let a_i^t denote the accuracy for task i after training on task t. The metrics are defined as:

$$\text{FAA} \triangleq \frac{1}{T} \sum_{i=1}^{T} a_i^T, \tag{16}$$

$$\text{forgetting} \triangleq \frac{1}{T-1} \sum_{j=1}^{T-1} f_j, \; s.t. \; f_j = \max_{l \in \{1,\dots,T-1\}} a_i^l - a_j^T. \tag{17}$$

Forgetting, ranging from [-100, 100], measures the average decrease in accuracy across tasks.

For analysis, we further define:

$$\text{CRD} \triangleq \text{FAA}_{\text{clean}} - \widetilde{\text{FAA}}_{\text{clean}}, \tag{18}$$

$$\text{FRI} \triangleq \widetilde{\text{Forgetting}}_{\text{clean}} - \text{Forgetting}_{\text{clean}}, \tag{19}$$

$$\text{RRD} \triangleq (\widetilde{\text{FAA}}_{\text{adv}}^{\text{Joint}} - \text{FAA}_{\text{adv}}^{\text{Joint}}) - (\widetilde{\text{FAA}}_{\text{adv}} - \text{FAA}_{\text{adv}}), \tag{20}$$

where "clean" and "adv" denote metrics on clean and adversarial data, respectively, tilde ($\sim$) indicates robust models, and "Joint" refers to a joint training model.

5.2. Main Results

We evaluate our method against various adversarial training and memory-based continual learning combinations on Split-CIFAR10/100. **Evaluations on Split-Tiny-ImageNet [45] and ViT [16] in supplementary materials**. Due to computational constraints, we focus on integrating our approach with ER+AT and ER+TRADES ($\beta = 6.0$), which showed superior adversarial FAA performance among baselines.

Buffer Size	Method	Class Incremental Setting					Task Incremental Setting				
		Clean Data		Adversarial Data			Clean Data		Adversarial Data		
		FAA↑	Forgetting↓	PGD-20↑	AA↑	Forgetting↓	FAA↑	Forgetting↓	PGD-20↑	AA↑	Forgetting↓
200	ER+AT	28.18	80.58	17.86	16.94	69.58	84.49	10.23	44.30	44.69	36.89
	ER+TRADES	22.42	77.25	15.72	15.53	64.95	78.79	8.10	51.33	51.50	21.97
	ER+FAT	33.61	69.21	15.14	14.81	49.04	83.40	10.35	43.69	43.96	28.56
	ER+LBGAT	25.68	84.47	16.65	16.56	70.50	78.19	18.85	40.69	40.73	40.83
	ER+SCORE	**48.65**	56.79	2.40	0.93	**18.96**	**88.90**	9.82	7.25	6.72	**8.70**
	ER+AT+Ours	35.68(↑ 7.50)	71.18(↓ 9.40)	18.40(↑ 0.54)	18.16(↑ 1.22)	67.85(↓ 1.73)	84.87(↑ 0.38)	9.93(↓ 0.30)	47.30(↑ 3.00)	47.61(↑ 2.92)	34.04(↓ 2.85)
	ER+TRADES+Ours	43.34(↑ 20.92)	**33.40**(↓ 43.85)	**19.85**(↑ 4.13)	**18.35**(↑ 2.82)	30.78(↓ 34.17)	82.59(↑ 3.80)	**7.53** (↓ 0.57)	**59.41** (↑ 8.08)	**59.59** (↑ 8.09)	14.16(↓ 7.81)
5,120	ER+AT	61.88	37.72	27.28	26.69	41.66	91.24	2.56	56.59	56.90	19.34
	ER+TRADES	20.36	85.14	16.30	16.18	72.85	88.48	1.59	64.36	64.52	12.47
	ER+FAT	54.55	43.18	19.68	18.91	42.15	91.50	2.12	56.72	56.87	14.79
	ER+LBGAT	62.45	37.73	27.42	26.66	47.83	91.10	3.25	56.57	56.31	19.38
	ER+SCORE	51.37	52.84	3.14	1.10	**20.29**	**95.63**	**1.58**	14.66	13.92	**2.53**
	ER+AT+Ours	**64.34**(↑ 2.46)	**23.64**(↓ 14.08)	**31.31**(↑ 4.03)	**30.49**(↑ 3.80)	20.46(↓ 21.20)	91.00(↓ 0.24)	3.51(↑ 0.95)	60.49(↑ 3.90)	60.61 (↑ 3.71)	13.27 (↓ 6.07)
	ER+TRADES+Ours	39.80(↑ 19.44)	44.08(↓ 41.06)	23.07(↑ 6.77)	21.98(↑ 5.80)	41.37(↓ 31.48)	86.48(↓ 2.00)	1.71(↑ 0.12)	**69.25**(↑ 4.89)	**69.38**(↑ 4.86)	**5.33**(↓ 7.14)

(a) Results on Split-CIFAR10. We chose two buffer sizes of 200 and 5, 120.

Buffer Size	Method	Class Incremental Setting					Task Incremental Setting				
		Clean Data		Adversarial Data			Clean Data		Adversarial Data		
		FAA↑	Forgetting↓	PGD-20↑	AA↑	Forgetting↓	FAA↑	Forgetting↓	PGD-20↑	AA↑	Forgetting↓
500	ER+AT	11.94	73.54	5.66	5.54	38.03	52.71	28.44	17.35	19.56	25.67
	ER+TRADES	7.59	68.13	5.43	5.11	44.13	50.75	20.22	26.89	27.69	20.34
	ER+FAT	11.48	73.99	5.35	5.23	37.26	56.10	24.71	20.01	22.99	22.31
	ER+LBGAT	11.77	75.10	6.16	5.82	39.77	42.02	26.45	13.99	14.43	23.14
	ER+SCORE	16.22	77.71	0.91	0.62	**6.13**	**68.87**	**11.21**	3.23	7.95	**4.94**
	ER+AT+Ours	**24.14**(↑ 12.20)	53.03(↓ 20.51)	7.13(↑ 1.47)	6.68(↑ 1.14)	25.81(↓ 12.22)	56.00(↑ 3.29)	26.19(↓ 2.25)	17.95(↑ 0.60)	20.26(↑ 0.70)	24.87(↓ 0.80)
	ER+TRADES+Ours	23.24(↑ 15.65)	**24.29**(↓ 43.84)	**9.93**(↑ 4.50)	**7.50**(↑ 2.39)	11.70(↓ 32.43)	56.68(↑ 5.93)	21.39(↑ 1.17)	**27.39**(↑ 0.50)	**28.50**(↑ 0.81)	16.76(↓ 3.58)
2,000	ER+AT	18.77	65.06	7.20	7.01	33.56	62.01	17.97	21.16	24.04	20.47
	ER+TRADES	9.50	70.49	5.35	5.01	42.08	60.63	13.78	26.68	29.19	18.60
	ER+FAT	17.09	66.91	6.12	5.93	33.62	63.88	15.99	23.48	26.75	17.12
	ER+LBGAT	20.58	63.31	7.93	7.05	30.09	55.77	25.11	17.95	18.88	19.18
	ER+SCORE	30.10	61.51	0.75	0.51	**4.10**	**76.90**	19.60	4.88	11.97	**5.25**
	ER+AT+Ours	**31.93**(↑ 13.16)	40.50(↓ 24.56)	9.61(↑ 2.41)	9.16(↑ 2.15)	17.78(↓ 15.78)	63.77(↑ 1.76)	17.16(↓ 0.81)	23.11(↑ 1.95)	25.59(↑ 1.55)	17.57(↓ 2.90)
	ER+TRADES+Ours	28.73(↑ 19.23)	**24.16**(↓ 46.33)	**12.75**(↑ 7.40)	**11.02**(↑ 6.01)	14.56(↓ 27.52)	62.01(↑ 1.38)	**13.16**(↓ 0.62)	**34.81**(↑ 8.13)	**35.36**(↑ 6.17)	11.60(↓ 7.00)

(b) Results on Split-CIFAR100. We choose two buffer sizes of 500 and 2, 000.

Table 2. Experiment results on Split-CIFAR10/100 datasets and model architecture is ResNet18. Here, PGD-20 and AA are adversarial data Final Average Accuracy (FAA) generated by PGD-20 and Auto Attack (AA), respectively. Forgetting of adversarial data is computed based on PGD-20. **Bold** represents the best experimental results for the same settings, underline indicates the second-best result, grean signifies relative improvement and red indicates relative degradation. By applying our methods, various model performances can be improved across the board.

Method	Class Incremental Setting				Task Incremental Setting			
	Clean Data		Adversarial Data		Clean Data		Adversarial Data	
	FAA↑	Forgetting↓	FAA↑	Forgetting↓	FAA↑	Forgetting↓	FAA↑	Forgetting↓
ER+AT	28.18	80.58	17.86	69.58	84.49	10.23	44.30	36.89
ER+AT+Ours	35.68(↑ 7.50)	71.18(↓ 9.40)	18.40(↑ 0.54)	67.85(↓ 1.73)	84.87(↑ 0.38)	9.93(↓ 0.30)	47.30(↑ 3.00)	34.04(↓ 2.85)
GSS+AT	27.59	80.78	16.67	68.53	84.41	9.83	44.25	34.79
GSS+AT+Ours	36.93(↑ 9.34)	67.72(↓ 13.06)	16.84(↑ 0.17)	60.04(↓ 8.49)	85.57(↑ 1.16)	8.86(↓ 0.97)	47.11(↑ 2.86)	31.83(↓ 2.97)
ASER+AT	18.85	87.78	14.06	65.57	73.87	19.01	30.65	44.85
ASER+AT+Ours	24.45(↑ 5.61)	81.73(↓ 6.05)	14.91(↑ 0.85)	62.74(↓ 2.83)	77.70(↑ 3.83)	15.21(↓ 3.80)	34.50(↑ 3.85)	38.55(↓ 6.30)
X-DER+AT	34.04	25.13	16.82	27.84	80.80	4.96	60.83	10.99
X-DER+AT+Ours	43.25(↑ 9.21)	20.77(↓ 4.36)	17.22(↑ 0.41)	18.56(↓ 9.28)	84.87(↑ 4.07)	1.74(↓ 3.21)	61.68(↑ 0.85)	7.03(↓ 3.97)

Table 3. Experiments with other data selection-based and logit masking-based continual learning methods.

Results on different adversarial training methods. Table 2 reveals several key findings. Buffer size significantly impacts performance, with larger buffers generally improving results, except for ER+TRADES on Split-CIFAR10 at buffer size 5, 120. Regarding robustness, ER+AT demonstrates superior performance in class incremental settings, while ER+TRADES achieves optimal results under task incremental learning, albeit with reduced clean sample accuracy. ER+LBGAT exhibits balanced performance in class incremental settings, while ER+FAT and ER+SCORE excel on clean samples but underperform on adversarial ones.

Our method significantly enhances both ER+TRADES

[1]The 80M-TinyImage dataset used in [12] has been withdrawn due to privacy concerns

AFLC	RAER	FP	Class Incremental Setting				Task Incremental Setting			
			Clean Data		Adversarial Data		Clean Data		Adversarial Data	
			FAA↑	Forgetting↓	FAA↑	Forgetting↓	FAA↑	Forgetting↓	FAA↑	Forgetting↓
			22.42	77.25	15.72	64.95	78.79	8.10	51.33	21.97
✓			37.88	21.76	18.58	21.77	81.66	10.86	53.64	17.60
✓	✓		37.45	**10.29**	19.81	**9.61**	78.13	15.39	55.71	**14.13**
✓	✓	✓	**43.34**	33.40	**19.85**	30.78	**82.59**	**7.53**	**59.41**	14.16

Table 4. Ablation experiments on the Split-CIFAR10 dataset with the buffer size of 200, using ER+TRADES as the baseline. **Bold** represents the best experimental results for the same settings. Experiments have demonstrated that AFLC mitigates clean-sample accelerated forgetting from adversarial samples, RAER mitigates gradient obfuscation (Adversarial FAA has a boost) and robust forgetting; adding FP improves the model's ability to learn new tasks (FAA has an overall increase).

and ER+AT across Split-CIFAR10/100 datasets, particularly in class incremental settings. For Split-CIFAR10, we achieve FAA improvements of $\Delta_{clean} = 20.92\%$ and $\Delta_{adv} = 8.09\%$, with forgetting reductions of $\Delta_{f_{clean}} = 43.85\%$ and $\Delta_{f_{adv}} = 34.17\%$. On Split-CIFAR100, improvements reach $\Delta_{clean} = 19.23\%$ and $\Delta_{adv} = 8.13\%$ for FAA, with forgetting reductions of $\Delta_{f_{clean}} = 46.33\%$ and $\Delta_{f_{adv}} = 32.43\%$.

Our approach demonstrates positive improvements in 93% of evaluation metrics (Table 2). Minor performance decreases in task incremental settings with buffer size $5,120$ reflect AFLC's diminishing returns and the known clean-robust accuracy trade-off [31, 35, 37, 57].

Results on different CL methods. We evaluate our method against state-of-the-art continual learning approaches on Split-CIFAR10 with buffer size 200, including gradient-based GSS, Shapley value-based ASER, and logit masking-based X-DER. These methods primarily address clean-data continual learning, whereas our approach specifically targets adversarial robustness.

Table 3 demonstrates that existing data selection methods (GSS, ASER) show limited effectiveness in adversarial settings, with robustness performance not exceeding ER+AT. Our method, when integrated with these approaches, enhances both robustness and clean accuracy across class and task incremental settings.

X-DER employs logit masking, a specialized case of our AFLC, for memory updates and future preparation. While effective at forgetting mitigation, its aggressive gradient suppression limits model learning capacity, resulting in lower robust FAA than ER+AT in class incremental settings. However, integration with our method yields significant improvements in both clean and adversarial metrics.

The consistent performance gains across all evaluation metrics validate our method's effectiveness in simultaneously addressing adversarial robustness and catastrophic forgetting. These empirical results demonstrate the broad applicability of our approach across different continual learning frameworks. **Accuracy curves during continual training can be found in the supplementary material.**

5.3. Ablation Experiments

In Table 4, we have performed ablation experiments based on ER+TRADES under the Split-CIFAR10 dataset. The results demonstrate that AFLC (Sec. 4.2) can effectively mitigate the increased forgetting caused by adversarial training under class incremental setting (55.49% for clean samples and 43.18% for adversarial samples forgetting, with corresponding FAA improvements of 15.46% and 2.86%, respectively). AFLC does not show significant improvement in the task incremental setting due to excessive suppression of future task classification heads and the use of the same calibration value for classes within the same task.

RAER (Sec. 4.3) can further improve the robust accuracy of AFLC by 1.23% for class incremental setting and 2.07% for task incremental setting and reduce the robust forgetting by 12.16% and 3.47% respectively. That proves the data selected by RAER describe the overall data distribution more accurately and effectively mitigate the gradient obfuscation.

When considering the future prior adjustment (FP in Table 4), we find that although the forgetting of the class incremental setting is higher, the FAA of both clean samples and adversarial samples has been significantly improved, and the forgetting of the task incremental setting has been further reduced, which proves that FP can reduce negative gradients to future classes and help learn new tasks. **We also provide sensitivity analysis of hyperparameters (α in AFLC and ρ in RAER) in supplementary materials.**

6. Conclusion and Future Work

This paper explores achieving adversarial robust continual learning without additional data. We analyze the challenges of directly combining continual learning and adversarial training - accelerated forgetting and gradient obfuscation. To address these, we propose: anti-forgettable logit calibration to reduce negative effects of past knowledge, and a robust difficulty-based data selection mechanism to mitigate gradient obfuscation. Experiments demonstrate the effectiveness of our approach. As AI systems increasingly need to learn continuously in dynamic environments, our work represents a further step toward more resilient and adaptable machine learning systems.

Acknowledgements

This work was supported in part by the Beijing Natural Science Foundation under No. L221013, Strategic Priority Research Program of the Chinese Academy of Sciences (No. XDB0680202), and the Innovation Funding of ICT, CAS under Grant No. E561160, E561090.

References

[1] Ali Abbasi, Parsa Nooralinejad, Hamed Pirsiavash, and Soheil Kolouri. Brainwash: A poisoning attack to forget in continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24057–24067, 2024. 2

[2] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018. 2

[3] Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. *Advances in neural information processing systems*, 32, 2019. 6

[4] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*, pages 274–283. PMLR, 2018. 2, 4

[5] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness. *arXiv preprint arXiv:2102.01356*, 2021. 2, 3

[6] Matteo Boschini, Lorenzo Bonicelli, Pietro Buzzega, Angelo Porrello, and Simone Calderara. Class-incremental continual learning into the extended der-verse. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2, 3, 5, 6

[7] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *Advances in neural information processing systems*, 33:15920–15930, 2020. 2, 3

[8] Lucas Caccia, Rahaf Aljundi, Nader Asadi, Tinne Tuytelaars, Joelle Pineau, and Eugene Belilovsky. New insights on reducing abrupt representation change in online continual learning. In *International Conference on Learning Representations*, 2022. 2

[9] Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *European Conference on Computer Vision*, pages 532–547. Springer, 2018. 2, 6

[10] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc'Aurelio Ranzato. Continual learning with tiny episodic memories. *arXiv preprint arXiv:1902.10486, 2019*, 2019. 1, 2, 3

[11] Jinghui Chen and Quanquan Gu. Rays: A ray searching method for hard-label adversarial attack. In *Proceedings of the 26rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2020. 6

[12] Tianlong Chen, Sijia Liu, Shiyu Chang, Lisa Amini, and Zhangyang Wang. Queried unlabeled data improves and robustifies class-incremental learning. *Transactions on Machine Learning and Data Mining*, 2022. 1, 2, 6, 7

[13] Ting-Chun Chou, Jhih-Yuan Huang, and Wei-Po Lee. Continual learning with adversarial training to enhance robustness of image recognition models. In *2022 International Conference on Cyberworlds*, pages 236–242. IEEE, 2022. 2

[14] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 2, 6

[15] Jiequan Cui, Shu Liu, Liwei Wang, and Jiaya Jia. Learnable boundary guided adversarial training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15721–15730, 2021. 2

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, G Heigold, S Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 6

[17] Arthur Douillard, Alexandre Ramé, Guillaume Couairon, and Matthieu Cord. Dytox: Transformers for continual learning with dynamic token expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9285–9295, 2022. 2

[18] Ranjie Duan, Yuefeng Chen, Dantong Niu, Yun Yang, A Kai Qin, and Yuan He. Advdrop: Adversarial attack to dnns by dropping information. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7506–7515, 2021. 2

[19] Sebastian Farquhar and Yarin Gal. Towards robust evaluations of continual learning. *arXiv preprint arXiv:1805.09733*, 2018. 2

[20] Yunhui Guo, Mingrui Liu, Yandong Li, Liqiang Wang, Tianbao Yang, and Tajana Rosing. Attacking lifelong learning models with gradient reversion. 2020. 1, 2

[21] Ahmad Hassanpour, Majid Moradikia, Bian Yang, Ahmed Abdelhadi, Christoph Busch, and Julian Fierrez. Differential privacy preservation in robust continual learning. *IEEE Access*, 10:24273–24287, 2022. 2

[22] Siteng Kang, Zhan Shi, and Xinhua Zhang. Poisoning generative replay in continual learning to promote forgetting. In *International Conference on Machine Learning*, pages 15769–15785. PMLR, 2023. 2

[23] Hikmat Khan, Nidhal Carla Bouaynaya, and Ghulam Rasool. Adversarially robust continual learning. In *2022 International Joint Conference on Neural Networks*, pages 1–8. IEEE, 2022. 1, 2

[24] Hikmat Khan, Pir Masoom Shah, Syed Farhan Alam Zaidi, et al. Susceptibility of continual learning against adversarial attacks. *arXiv preprint arXiv:2207.05225*, 2022. 1, 2

[25] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran

Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 2

[26] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009. 3, 6

[27] Lilly Kumari, Shengjie Wang, Tianyi Zhou, and Jeff Bilmes. Retrospective adversarial replay for continual learning. In *Advances in Neural Information Processing Systems*, 2022. 2

[28] Sungyoon Lee, Hoki Kim, and Jaewook Lee. Graddiv: Adversarial robustness of randomized neural networks via gradient diversity regularization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2645–2651, 2023. 2

[29] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. 1, 2

[30] Guoliang Lin, Hanlu Chu, and Hanjiang Lai. Towards better plasticity-stability trade-off in incremental learning: A simple linear connector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 89–98, 2022. 2

[31] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 2, 3, 4, 8

[32] Pratyush Maini, Eric Wong, and Zico Kolter. Adversarial robustness against the union of multiple perturbation models. In *International Conference on Machine Learning*, pages 6640–6650. PMLR, 2020. 2

[33] Marc Masana, Xialei Liu, Bartłomiej Twardowski, Mikel Menta, Andrew D. Bagdanov, and Joost van de Weijer. Class-incremental learning: Survey and performance evaluation on image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5513–5533, 2023. 2

[34] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, pages 109–165. Elsevier, 1989. 2

[35] Tianyu Pang, Kun Xu, Yinpeng Dong, Chao Du, Ning Chen, and Jun Zhu. Rethinking softmax cross-entropy loss for adversarial robustness. In *International Conference on Learning Representations*, 2020. 8

[36] Tianyu Pang, Xiao Yang, Yinpeng Dong, Hang Su, and Jun Zhu. Bag of tricks for adversarial training. In *International Conference on Learning Representations*, 2021. 2, 3, 6

[37] Tianyu Pang, Min Lin, Xiao Yang, Jun Zhu, and Shuicheng Yan. Robustness and accuracy could be reconcilable by (proper) definition. *International Conference on Machine Learning*, 2022. 2, 6, 8

[38] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019. 2

[39] Roger Ratcliff. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*, 97(2):285, 1990. 2

[40] Mohammad Rostami, Leonidas Spinoulas, Mohamed Hussein, Joe Mathai, and Wael Abd-Almageed. Detection and continual learning of novel face presentation attacks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14851–14860, 2021. 2

[41] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016. 2

[42] Fahad Sarfraz, Elahe Arani, and Bahram Zonooz. Error sensitivity modulation based experience replay: Mitigating abrupt representation drift in continual learning. In *The Eleventh International Conference on Learning Representations*, 2023. 1

[43] Rui Shao, Pramuditha Perera, Pong C Yuen, and Vishal M Patel. Open-set adversarial defense. In *European Conference on Computer Vision*, pages 682–698. Springer, 2020. 2

[44] Dongsub Shim, Zheda Mai, Jihwan Jeong, Scott Sanner, Hyunwoo Kim, and Jongseong Jang. Online class-incremental continual learning with adversarial shapley value. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9630–9638, 2021. 6

[45] Stanford. Tiny imagenet challenge (cs231n). `http://tiny-imagenet.herokuapp.com/`, 2015. Accessed: Feb 21, 2023. 6

[46] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 2

[47] Muhammad Umer, Glenn Dawson, and Robi Polikar. Targeted forgetting and false memory formation in continual learners through adversarial backdoor attacks. In *2020 International Joint Conference on Neural Networks*, pages 1–8. IEEE, 2020. 2

[48] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1, 2

[49] Zhenyi Wang, Li Shen, Le Fang, Qiuling Suo, Tiehang Duan, and Mingchen Gao. Improving task-free continual learning by distributionally robust memory evolution. In *International Conference on Machine Learning*, pages 22985–22998. PMLR, 2022. 2

[50] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 139–149, 2022. 2

[51] Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. Better diffusion models further improve adversarial training. In *International Conference on Machine Learning*, 2023. 6

[52] Zeyu Wang, Xianhang Li, Hongru Zhu, and Cihang Xie. Revisiting adversarial training at scale. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 24675–24685, 2024. 1, 2

[53] Tongtong Wu, Massimo Caccia, Zhuang Li, Yuan-Fang Li, Guilin Qi, and Gholamreza Haffari. Pretrained language model in continual learning: A comparative study. In *International Conference on Learning Representations*, 2021. 1, 2

[54] Tong Wu, Ziwei Liu, Qingqiu Huang, Yu Wang, and Dahua Lin. Adversarial robustness under long-tailed distribution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8659–8668, 2021. 2

[55] Ju Xu, Jin Ma, Xuesong Gao, and Zhanxing Zhu. Adaptive progressive continual learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6715–6728, 2022. 2

[56] Jaehong Yoon, Saehoon Kim, Eunho Yang, and Sung Ju Hwang. Scalable and order-robust continual learning with additive parameter decomposition. *arXiv preprint arXiv:1902.09432*, 2019. 1, 2

[57] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, pages 7472–7482. PMLR, 2019. 1, 2, 8

[58] Jingfeng Zhang, Jianing Zhu, Gang Niu, Bo Han, Masashi Sugiyama, and Mohan Kankanhalli. Geometry-aware instance-reweighted adversarial training. In *International Conference on Learning Representations*, 2021. 6

[59] Jie Zhang, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, Lei Zhang, and Chao Wu. Towards efficient data free black-box adversarial attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15115–15125, 2022. 2

[60] Da-Wei Zhou, Qi-Wei Wang, Zhi-Hong Qi, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Deep class-incremental learning: A survey. *arXiv preprint arXiv:2302.03648*, 2023. 2

[61] Da-Wei Zhou, Hai-Long Sun, Han-Jia Ye, and De-Chuan Zhan. Expandable subspace ensemble for pre-trained model-based class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23554–23564, 2024. 2

[62] Da-Wei Zhou, Qi-Wei Wang, Zhi-Hong Qi, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Class-incremental learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1, 2

[63] Kaijie Zhu, Jindong Wang, Xixu Hu, Xing Xie, and Ge Yang. Improving generalization of adversarial training via robust critical fine-tuning. *arXiv preprint arXiv:2308.02533*, 2023. 6