Continual Learning via Manifold Expansion Replay

Zihao Xu[1], Xuan Tang[2], Yufei Shi[3], Jianfeng Zhang[1], Jian Yang[4], Mingsong Chen[1], Xian Wei[1*]

[1] Software Engineering Institute, East China Normal University, Shanghai, China

[2] School of Communication & Electronic Engineering, East China Normal University, Shanghai, China

[3] Department of Medical Informatics and Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou, China

[4] School of Geospatial Information, Information Engineering University, Zhengzhou, China

xian.wei@tum.de

Abstract—In continual learning, the learner learns multiple tasks in sequence, with data being acquired only once for each task. Catastrophic forgetting is a major challenge to continual learning. To reduce forgetting, some existing rehearsal-based methods use episodic memory to replay samples of previous tasks. However, in the process of knowledge integration when learning a new task, this strategy also suffers from catastrophic forgetting due to an imbalance between old and new knowledge. To address this problem, we propose a novel replay strategy called Manifold Expansion Replay (MaER). We argue that expanding the implicit manifold of the knowledge representation in the episodic memory helps to improve the robustness and expressiveness of the model. To this end, we propose a greedy strategy to keep increasing the diameter of the implicit manifold represented by the knowledge in the buffer during memory management. In addition, we introduce Wasserstein distance instead of cross entropy as distillation loss to preserve previous knowledge. With extensive experimental validation on MNIST, CIFAR10, CIFAR100, and TinyImageNet, we show that the proposed method significantly improves the accuracy in continual learning setup, outperforming the state of the arts.

Index Terms—continual learning, catastrophic forgetting, manifold diameter, Wasserstein distance

I. INTRODUCTION

Continual learning, also known as incremental learning [1]–[3], refers to the process of sequentially learning multiple tasks without forgetting previous knowledge. In this setting, catastrophic forgetting [4], [5] is a major challenge for continual learning, where previously learned knowledge is lost when learning new tasks.

Continual learning has recently gained increasing attention in the field of artificial intelligence. Various strategies have been proposed to overcome catastrophic forgetting [6], including rehearsal-based strategies [7]–[9], regularization-based strategies [10], [11], and parameter isolation-based strategies [12], [13]. These strategies are mutually orthogonal and can be combined in a specific scenario. Among

these strategies, rehearsal-based methods have proven to be a simple yet effective approach that uses episodic memory to replay training samples. Despite its encouraging success, there are still challenges that need to be addressed, including the issue of overfitting and biased knowledge representation due to knowledge imbalance in episodic memory. A naive but effective solution is increasing the memory size when new samples come. However, this approach increases the memory requirement and violates the setting of limited memory resource requirements in continual learning.

To address this issue, we propose a novel replay strategy called Manifold Expansion Replay (MaER). MaER investigates two factors to improve neural network performance in continual learning settings. Firstly, MaER views the process of continual learning as a fusion of implicit manifolds represented by knowledge. When the diameters of manifolds are imbalanced, the larger one will receive more bias while the smaller one will experience forgetting. Inspired by this, MaER adopts a greedy sampling strategy to manage memory, helping the neural network to learn unbiased presentation of all data. Secondly, MaER introduces the Wasserstein distance as distillation loss. The Wasserstein distance between two distributions is defined as the minimum cost required to match one distribution with another. Unlike traditional distance metrics such as Euclidean distance, Wasserstein distance considers the underlying structure of the compared distributions, which can help the neural network better fuse knowledge manifolds.

We mainly evaluate MaER on Permuted MNIST, Rotated MNIST, Split CIFAR10, Split CIFAR100, and Split TinyImageNet datasets. The extensive ablation studies and experimental results show that MaER gains significant performance improvement, outperforming state-of-the-art in accuracy.

Our contributions are summarized as follows:
- We propose a greedy sampling strategy to balance knowledge by expanding the diameter of the knowl-

* Corresponding author.

edge manifold in episodic memory.

- We propose to distill knowledge using Wasserstein distance, which helps neural networks effectively fuse knowledge in continual learning.

II. RELATED WORK

Here we briefly review previous research works. Existing works can be divided into three categories, i.e., rehearsal-based, regularization-based, and parameter isolation-based.

a) Rehearsal-based Strategy: The rehearsal-based strategy can be viewed as a review strategy that uses a capacity-limited buffer called episodic memory to replay a portion of the samples from the previous task at each training session. Despite its simplicity, this rehearsal strategy has been shown to be effective and work well when large memories are available. Typical approaches include Incremental Classifier and Representation Learning (iCaRL) [14], Experience Replay (ER) [7], Selective Experience Replay (SER) [15], Continual Prototype Evolution (CoPE) [16] and Tiny Experience Replay (TEM) [17].

b) Regularization-based Strategy: In contrast to the rehearsal strategy, the regularization strategy adopts a more strict continual learning setup. This strategy aims to reduce forgetting without accessing prior task data. Existing methods can be further divided into two categories, i.e., Data-focused and prior-focused [16]. The main idea of data-focused methods is to transfer knowledge from a teacher model to a student model using the knowledge distillation technique, where the teacher model has been trained on previous tasks. The idea of using knowledge distillation to improve performance on new tasks in continual learning was first proposed by [18]. Subsequent research has proven that knowledge distillation can also reduce forgetting. Typical methods include Learning without Forgetting (LwF) [19], Learning from Less (LFL) [20], and Dark Knowledge distillation with Memory Consolidation (DMC) [21]. The basic idea of prior-focused methods is to estimate the importance of parameters to previous tasks and then penalize changes to important parameters during training to prevent catastrophic forgetting. This strategy has been shown to be effective. Typical methods include Elastic Weight Consolidation (EWC) [11], Variational Continual Learning (VCL) [22], Incremental Moment Matching (IMM) [23], Synaptic Intelligence (SI) [10], and Riemannian Walk (RW) [2].

c) Paramters Isolation-based Strategy: This strategy overcomes catastrophic forgetting by selecting parameters from a fixed network or dynamically modifying the network structure. Typical methods, selecting a subnetwork for each task, include Hard Attention (HAT) [24], PackNet [25], and PathNet [26]. HAT uses a hard attention mask to selectively prune network parameters, retaining important features relevant to the current task and reducing forgetting. PackNet uses the network pruning technique to make the network adapt to multiple tasks. PathNet divides the network into sub-networks, with each sub-network responsible for each task. Dynamically expanding network structure and establishing new neural connections for new tasks has also been proven to be an effective strategy. Typical methods include Progressive Neural Networks (PNN) [12], Deep Adaptive Network (DAN) [27], and Reinforced Continual Learning (RCL) [28]. Both PNN and DAN adopt a hierarchical structure, where new network layers are established for new tasks, and each layer is responsible for a specific task. RCL, on the other hand, employs reinforcement learning techniques to adjust the network's learning strategy by rewarding and punishing its learning process.

III. CONTINUAL LEARNING SETUP

In continual learning, a learner needs to sequentially learn T tasks $\{(\mathcal{X}_1, \mathcal{Y}_1, ..., (\mathcal{X}_T, \mathcal{Y}_T)\}$. $(\mathcal{X}_t, \mathcal{Y}_t)$ represents a dataset $\mathcal{D}_t = \{(x_t^1, y_t^1), ..., (x_t^{n_t}, y_t^{n_t})\}$ from task t, randomly sampled from an unknown distribution $\mathcal{P}_t$, where x_t represents the sample, y_t represents the corresponding ground truth label, and n_t represents the number of samples in the dataset. We assume that the learner can use a capacity-limited buffer $\mathcal{M}$ to store a small number of samples during learning. Our goal is to train a predictor $f = (w \circ \Phi) : \mathcal{X} \to \mathcal{Y}$, composed of a feature extractor Φ and a classifier w, that minimizes the risk over all the data it has seen while only having access to a limited number of samples from previous tasks that are stored in $\mathcal{M}$:

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{(x,y) \sim P_t} [\ell(f(x; \theta), y)], \tag{1}$$

where θ denotes the parameters of model and ℓ denotes the loss function.

Evaluation Metrics. Following [2], [17], [29], we use average classification accuracy (ACC) and backward transfer (BWT) to evaluate performance. Formally, ACC is defined as:

$$\text{ACC} = \frac{1}{T} \sum_{j=1}^{T} a_{T,j}, \tag{2}$$

where $a_{i,j}$ denotes the test accuracy of the model on task j after learning task i. BWT is defined as the average change in accuracy of old tasks after learning a new task:

$$\text{BWT} = \frac{1}{T-1} \sum_{j=1}^{T-1} \max_{l \in \{1,...,T-1\}} (a_{l,j} - a_{T,j}). \tag{3}$$

A positive value of BWT means that learning a new task benefits the old tasks, while a negative value indicates that learning a new task interferes with the old tasks.

IV. MANIFOLD EXPANSION REPLAY

Recent research has shown that the replay strategy is a simple and effective approach to continual learning. However, to develop more robust methods based on the replay strategy, two issues need to be considered: (1) how to replay during the training phase and (2) how to sample and manage

TABLE I
CLASSIFICATION RESULTS FOR SPLIT CIFAR10, SPLIT TINYIMAGENET, PERMUTED MNIST AND ROTATED MNIST.

Buffer	Method	Split CIFAR10	Split TinyImageNet	P-MNIST	R-MNIST
-	JOINT	98.31 ± 0.12	82.04 ± 0.10	94.33 ± 0.17	95.76 ± 0.04
	SGD	61.02 ± 3.33	18.31 ± 0.68	40.70 ± 2.33	6.77 ± 8.53
-	oEWC	68.29 ± 3.92	19.20 ± 0.31	$\mathbf{75.79 \pm 2.25}$	$\mathbf{77.35 \pm 5.77}$
	SI	68.05 ± 5.91	36.32 ± 0.13	65.86 ± 1.57	71.91 ± 5.83
	LwF	63.29 ± 2.35	19.20 ± 0.31	-	-
	PNN	$\mathbf{95.13 \pm 0.72}$	$\mathbf{67.84 \pm 0.29}$	-	-
200	ER	91.19 ± 0.94	38.17 ± 2.00	72.37 ± 0.87	85.01 ± 1.90
	A-GEM	83.88 ± 1.49	22.77 ± 0.03	66.42 ± 4.00	81.91 ± 0.76
	iCaRL	88.99 ± 2.13	28.19 ± 1.47	-	-
	HAL	82.51 ± 3.20	-	74.16 ± 1.65	84.02 ± 0.98
	DER	91.40 ± 0.92	40.22 ± 0.67	81.74 ± 1.07	90.04 ± 2.61
	SNCL	$\mathbf{92.91 \pm 0.81}$	43.01 ± 1.67	86.23 ± 0.20	91.54 ± 2.58
	MaER (ours)	92.56 ± 0.49	$\mathbf{46.34 \pm 0.79}$	$\mathbf{90.04 \pm 0.28}$	$\mathbf{91.88 \pm 1.96}$
500	ER	93.61 ± 0.27	48.64 ± 0.46	80.60 ± 0.86	88.91 ± 1.44
	A-GEM	89.48 ± 1.45	25.33 ± 0.49	67.56 ± 1.28	80.31 ± 6.29
	iCaRL	88.22 ± 2.62	31.15 ± 3.27	-	-
	HAL	84.54 ± 2.36	-	80.13 ± 0.49	85.00 ± 0.96
	DER	93.40 ± 0.39	51.78 ± 0.88	87.29 ± 0.46	92.24 ± 1.12
	SNCL	$\mathbf{94.02 \pm 0.43}$	52.85 ± 0.67	88.53 ± 0.41	$\mathbf{93.05 \pm 1.02}$
	MaER (ours)	93.29 ± 0.42	$\mathbf{54.65 \pm 0.77}$	$\mathbf{92.34 \pm 0.54}$	92.55 ± 1.08

TABLE II
AVERAGE ACCURACY (ACC) AND FORGETTING (BWT) RESULTS ON SPLIT CIFAR100.

Method	Episodic Memory					
	ACC			BWT		
	100	300	500	100	300	500
A-GEM	54.9 ± 2.92	56.9 ± 3.45	$59.9 \pm 2.64]$	0.14 ± 0.03	0.13 ± 0.03	0.10 ± 0.02
ER	49.7 ± 2.97	57.7 ± 2.59	60.6 ± 2.09	0.19 ± 0.03	0.11 ± 0.01	0.09 ± 0.02
ER-RING	56.2 ± 1.93	60.9 ± 1.44	62.6 ± 1.77	0.13 ± 0.01	0.09 ± 0.01	0.08 ± 0.02
ER-RESERVOIR	53.1 ± 2.66	59.7 ± 3.87	65.5 ± 1.99	0.19 ± 0.02	0.12 ± 0.03	0.09 ± 0.02
MaER(ours)	$\mathbf{57.46 \pm 0.95}$	$\mathbf{62.61 \pm 1.59}$	$\mathbf{66.4 \pm 1.56}$	0.19 ± 0.01	0.13 ± 0.01	0.09 ± 0.01
FINETUNE	40.6 ± 3.83	-	-	-	-	-
EWC	41.2 ± 2.67	-	-	-	-	-

TABLE III
ABLATION STUDY OF THE DIFFERENT COMPONENTS IN PROPOSED METHOD. $\mathcal{L}_{WD}$ DENOTES THE WASSERSTEIN DISTANCE.

Method	Dataset	ACC (%) Memory size		
		100	300	500
$\mathcal{L}_{CE}$	P-MNIST	77.64	82.62	85.38
	Split CIFAR100	56.43	59.64	64.05
$\mathcal{L}_{CE} + \mathcal{L}_{WD}$	P-MNIST	81.79	87.72	90.14
	Split CIFAR100	57.49	62.46	66.37
MaER	P-MNIST	84.47	89.52	92.07
	Split CIFAR100	57.99	63.95	67.86

memory after each task training. Our approach, MaER, designs strategies for these two stages from a geometric perspective.

A. How to Replay

In the replay strategy, the number of samples collected for replay is limited. This poses a challenge in recalling the knowledge of the entire task from these samples. If a classification loss is used when training replay samples, the model can only learn to classify these samples rather than previous tasks. As a result, the model may still suffer from catastrophic forgetting as the number of tasks increases. For a specific task, we assume that each sample represents a piece of meta-knowledge. The model learns the entire knowledge manifold by learning from these samples. We now consider the first problem encountered: how to integrate this meta-knowledge when learning new tasks to form a more comprehensive knowledge manifold. To do this, we need to measure the distance between two knowledge manifolds. Intuitively, the distance between the new and old knowledge manifolds should be small because previously acquired meta-knowledge does not change. If we view the knowledge manifold as a distribution of meta-knowledge, one possible choice for a distance metric is KL divergence, which is commonly used to measure the distance between two distributions. However, KL divergence cannot be used as a strict distance function because it is asymmetric and cannot provide distance information when two distributions

do not overlap. Our method, MaER, uses Wasserstein distance to measure the distance between two knowledge manifolds.

In mathematics, the Wasserstein distance is a distance function defined between probability distributions on a given metric space (M, ρ), where $\rho(x, y)$ is a distance function for two instances x and y in the set M. Formally, the p-th Wasserstein distance between two probability measures μ and ν on M with p-moment is defined as:

$$W_p(\mu, \nu) := \left(\inf_{\gamma \in \Gamma(\mu, \nu)} \int_{M \times M} d(x, y)^p \, \mathrm{d}\gamma(x, y) \right)^{1/p}, \tag{4}$$

where $\Gamma(\mu, \nu)$ represents the set of all coupling of μ and ν. Wasserstein distance has some good properties. In contrast to KL divergence, Wasserstein distance is a symmetric metric and provides information even if the distributions do not overlap.

Now we describe how MaER uses the Wasserstein distance to facilitate meta-knowledge fusion in continual learning. When learning task i, we have a teacher model f_t that has learned the previous $i - 1$ tasks and a student model f_s responsible for learning task i. For a sample x, we use $\Phi_s(x; \theta)$ to denote the knowledge representation of f_s for that sample, where Φ_s denotes the feature extractor for f_s. Our objective is for the student model f_s to effectively learn task i. To achieve this, we train f_s on samples from task i using the cross-entropy loss function for classification. Concurrently, it is imperative that f_s retains the knowledge acquired from previous tasks. During training, we replay samples from the memory buffer $\mathcal{M}$, and in addition to learning to classify these samples accurately, we aim to minimize the Wasserstein distance between the knowledge representation learned by f_s and that of the teacher model f_t. Taking these considerations into account, the new loss function is defined as:

$$\mathcal{L} = \mathbb{E}_{(x,y) \sim D_t}[\ell(f_s(x; \theta), y)]$$
$$+ \mathbb{E}_{(x,y) \sim \mathcal{M}}[\ell(f_s(x; \theta), y) + W_p(\Phi_s(x; \theta), \Phi_t(x; \theta))], \tag{5}$$

where ℓ is the cross-entropy function in classification tasks. For ease of computation, we use the Wasserstein distance with p = 2 and assume that meta-knowledge is equally important. In MaER, the calculation of W_2 can be simplified to the following form:

$$W_2(P, Q) = (\frac{1}{n} \sum_i^n d(x, y)^2)^{1/2}, \tag{6}$$

where d is the distance function in metric space M, which can be Euclidean distance or geodesic distance.

B. How to Manage Memory

We now turn to the second problem: how to sample such that a small number of meta-knowledge can represent the knowledge manifold of the entire task as much as possible. We assume that the meta-knowledge is uniformly distributed over this knowledge manifold. We can only

---

Algorithm 1 Manifold Expansion Sampling

Input: $\Phi_s, D_t, \mathcal{M}, mem\_size, n$
Output: $\mathcal{M}$
$j \leftarrow 0$
for (x, y) in $\mathcal{D}_t$ **do**
  if $|\mathcal{M}| < mem\_size$ **then**
    $\mathcal{M}.\text{append}(x, y)$
  else
    $feature\_\mathcal{M} \leftarrow \Phi_s(\mathcal{M})$
    $\mathcal{C}, diameter \leftarrow \text{CentroidAndDiameter}(feature\_\mathcal{M})$
    $feature\_x \leftarrow \Phi_s(x)$
    if $\text{distance}(\mathcal{C}, feature\_x) > diameter$ **then**
      $i \leftarrow \text{randint}(0, |\mathcal{M}|)$
      $\mathcal{M}[i] \leftarrow (x, y)$
    else
      $i \leftarrow \text{randint}(0, n + j)$
      if $i < mem\_size$ **then**
        $\mathcal{M}[i] \leftarrow (x, y)$
      end if
    end if
  end if
  $j \leftarrow j + 1$
end for
$n \leftarrow n + j$
Return $\mathcal{M}$

---

sample a small fraction of meta-knowledge to represent the whole manifold. Intuitively, our sampling method needs to account for two aspects: (1) sample as uniformly as possible to maintain the geometric properties, (2) the sampled meta-knowledge should span the entire knowledge manifold as much as possible, avoiding bias towards new knowledge and preventing forgetting during learning. To address this problem, MaER employs a greedy strategy that incrementally enlarges the diameter of the manifold during the sampling process. First, the sampling process is stochastic. When a new sample arrives, if $\mathcal{M}$ is already full, then each sample in $\mathcal{M}$ has an equal chance of being replaced. Stochastic sampling preserves the consistency of the data distribution in $\mathcal{M}$ and the original distribution. Second, MaER's sampling strategy guarantees that samples that can augment the diameter of the manifold are always collected. For other samples, MaER will collect them with a certain probability. The criterion for determining whether a sample can augment the diameter of the manifold is essential for understanding MaER's sampling strategy. We begin by introducing the concepts of the centroid and diameter of a manifold. The Fréchet mean is a natural generalization of the centroid and can be applied to any manifold. For a metric space $\mathcal{X} = (X, d, \mu)$, its Fréchet mean is defined as:

$$\arg\min_{x \in X} \int_X d^2(x, y) d\mu(y). \tag{7}$$

With the manifold centroid $\mathcal{C}$, we can estimate the diameter of the manifold in a simple and efficient way. We define the

diameter as the largest distance from the centroid $\mathcal{C}$ to the sample x. Having defined these concepts, we can describe the sampling process for MaER, as shown in Algorithm 1.

V. Experiments

We apply MaER to different sequential tasks for continual learning and compare it with state-of-the-art replay methods, and then we empirically analyze the proposed algorithm.

A. Experimental Setting

Here, we begin by describing the continual learning benchmarks, implementation details, and compared methods.

Benchmarks. We conducted experiments on several continual learning datasets: Permuted MNIST, Rotated MNIST, Split CIFAR10, Split CIFAR100 and Split TinyImageNet. Permuted MNIST is derived from applying a random permutation to the pixels. Rotated MNIST is derived from rotating the image at a random angle. Split CIFAR10, Split CIFAR100, and Split TinyImageNet are derived from splitting CIFAR10, CIFAR100, and miniImageNet, respectively, such that the classes in the different tasks are disjoint. Split CIFAR10 consists of 5 tasks, while the other datasets consist of 20 tasks each.

Implementation details. In the network architecture, we utilized a three-layer MLP with 256 neurons in the hidden layer for MNIST and a standard resnet18 for CIFAR10, CIFAR100 and TinyImageNet. We optimize the parameters during training using SGD. The learning rate is set to 0.01 for MNIST and 0.003 for others. The batch size is set to 16 for all experiments. We trained the model for 10 epochs on MNIST, 5 epochs on Split CIFAR100 and 20 epochs on others.

Compared methods. We compare MaER with several baseline methods. The rehearsal-based baselines include iCaRL [14], ER [30], ER-RING [31], ER-RESERVOIR [31], A-GEM [32], HAL [33], DER [34], and SNCL [35], with DER and SNCL being the strongest baselines. We also compare with other strategy-based methods, including SI [10], EWC [11], oEWC [36], LwF [19], and PNN [12]. Additionally, we compare with two non-continual learning methods, JOINT and SGD, as upper and lower bounds.

B. Main Results

Table I shows the results of our method using a tiny buffer on Split CIFAR10, Split TinyImagNet, Permuted MNIST, and Rotated MNIST. On Split CIFAR10 and Rotated MNIST, our method achieved competitive results, differing from state-of-the-art replay methods by $0.5\% \sim 0.73\%$. On Permuted MNIST and Split TinyImageNet, our method achieves state-of-the-art performance, significantly outperforming baseline methods. On Split TinyImageNet, MaER surpassed SNCL by $1.80\% \sim 3.33\%$ in average accuracy when taking different buffer sizes. On Permuted P-MNIST, this gap was even more pronounced, with our

method leading SNCL by 3.81% in average accuracy when taking different buffer sizes. On Split CIFAR10 and Rotated MNIST, most baselines achieved considerable results. However, on Permuted MNIST and Split TinyImageNet, the gap between different methods became more pronounced. These results indicate that Permuted MNIST and Split TinyImageNet are more challenging when using a tiny buffer. MaER can work well in challenging scenarios and achieves state-of-the-art in average accuracy. On Split CIFAR100, we also compared MaER with more replay strategies. In Table II, our method surpassed all baselines when taking different buffer sizes, leading the baseline by $0.9\% \sim 7.64\%$ in average accuracy. As the buffer size increased, MaER was able to perform correspondingly better. Table III shows the ablation study results of different components of MaER. Here, $\mathcal{L}_{WD}$. represents the Wasserstein distance loss. Compared to the naive replay, both the replay strategy and memory management in MaER achieve certain performance improvements.

VI. Conclusion

In this paper, we propose a new replay strategy called MaER. MaER employs knowledge distillation techniques and introduces the Wasserstein distance between the features of the teacher and student models as a distillation loss to integrate old and new knowledge better. Intuitively, when integrating two imbalanced knowledge manifolds, the larger manifold will receive more bias, leading to catastrophic forgetting. MaER addresses this issue through manifold expansion sampling. Samples that can expand the manifold diameter are deterministically sampled, while those within the diameter range are randomly sampled. Our extensive experiments demonstrate that MaER performs well with tiny buffers and achieves state-of-the-art performance.

Limitation. To calculate the Wasserstein distance, MaER must compute the features of both the teacher and student models. This requires performing inference on the data twice, adding a certain computational burden. In memory management, manifold expansion sampling relies on calculating manifold diameters. Although we simplify the calculation of diameters, the cost of these calculations increases as the number of samples grows. When using large buffers, these computational overheads can make MaER slower to train than other methods.

References

[1] R. Aljundi, F. Babiloni, M. Elhoseiny, M. Rohrbach, and T. Tuytelaars, "Memory aware synapses: Learning what (not) to forget," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 139–154, 2018.

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

[3] R. Aljundi, P. Chakravarty, and T. Tuytelaars, "Expert gate: Lifelong learning with a network of experts," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3366–3375, 2017.

[4] R. M. French, "Catastrophic forgetting in connectionist networks," *Trends in cognitive sciences*, vol. 3, no. 4, pp. 128–135, 1999.

[5] R. Ratcliff, "Connectionist models of recognition memory: constraints imposed by learning and forgetting functions.," *Psychological review*, vol. 97, no. 2, p. 285, 1990.

[6] M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars, "Continual learning: A comparative study on how to defy forgetting in classification tasks," *arXiv preprint arXiv:1909.08383*, vol. 2, no. 6, p. 2, 2019.

[7] D. Rolnick, A. Ahuja, J. Schwarz, T. Lillicrap, and G. Wayne, "Experience replay for continual learning," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[8] A. Ayub and A. R. Wagner, "Storing encoded episodes as concepts for continual learning," *arXiv preprint arXiv:2007.06637*, 2020.

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

[10] F. Zenke, B. Poole, and S. Ganguli, "Continual learning through synaptic intelligence," in *International conference on machine learning*, pp. 3987–3995, PMLR, 2017.

[11] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

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

[13] X. Li, Y. Zhou, T. Wu, R. Socher, and C. Xiong, "Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting," in *International Conference on Machine Learning*, pp. 3925–3934, PMLR, 2019.

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

[15] D. Isele and A. Cosgun, "Selective experience replay for lifelong learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.

[16] M. De Lange and T. Tuytelaars, "Continual prototype evolution: Learning online from non-stationary data streams," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8250–8259, 2021.

[17] A. Chaudhry, M. Rohrbach, M. Elhoseiny, T. Ajanthan, P. K. Dokania, P. H. Torr, and M. Ranzato, "Continual learning with tiny episodic memories," 2019.

[18] D. L. Silver and R. E. Mercer, "The task rehearsal method of life-long learning: Overcoming impoverished data," in *Advances in Artificial Intelligence: 15th Conference of the Canadian Society for Computational Studies of Intelligence, AI 2002 Calgary, Canada, May 27–29, 2002 Proceedings 15*, pp. 90–101, Springer, 2002.

[19] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2935–2947, 2017.

[20] H. Jung, J. Ju, M. Jung, and J. Kim, "Less-forgetting learning in deep neural networks," *arXiv preprint arXiv:1607.00122*, 2016.

[21] J. Zhang, J. Zhang, S. Ghosh, D. Li, S. Tasci, L. Heck, H. Zhang, and C.-C. J. Kuo, "Class-incremental learning via deep model consolidation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1131–1140, 2020.

[22] C. V. Nguyen, Y. Li, T. D. Bui, and R. E. Turner, "Variational continual learning," *arXiv preprint arXiv:1710.10628*, 2017.

[23] S.-W. Lee, J.-H. Kim, J. Jun, J.-W. Ha, and B.-T. Zhang, "Overcoming catastrophic forgetting by incremental moment matching," *Advances in neural information processing systems*, vol. 30, 2017.

[24] J. Serra, D. Suris, M. Miron, and A. Karatzoglou, "Overcoming catastrophic forgetting with hard attention to the task," in *International Conference on Machine Learning*, pp. 4548–4557, PMLR, 2018.

[25] A. Mallya and S. Lazebnik, "Packnet: Adding multiple tasks to a single network by iterative pruning," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 7765–7773, 2018.

[26] C. Fernando, D. Banarse, C. Blundell, Y. Zwols, D. Ha, A. A. Rusu, A. Pritzel, and D. Wierstra, "Pathnet: Evolution channels gradient descent in super neural networks," *arXiv preprint arXiv:1701.08734*, 2017.

[27] A. Rosenfeld and J. K. Tsotsos, "Incremental learning through deep adaptation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 3, pp. 651–663, 2018.

[28] J. Xu and Z. Zhu, "Reinforced continual learning," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

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

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

[31] A. Chaudhry, M. Rohrbach, M. Elhoseiny, T. Ajanthan, P. K. Dokania, P. H. Torr, and M. Ranzato, "On tiny episodic memories in continual learning," *arXiv preprint arXiv:1902.10486*, 2019.

[32] A. Chaudhry, M. Ranzato, M. Rohrbach, and M. Elhoseiny, "Efficient lifelong learning with a-gem," *arXiv preprint arXiv:1812.00420*, 2018.

[33] A. Chaudhry, A. Gordo, P. Dokania, P. Torr, and D. Lopez-Paz, "Using hindsight to anchor past knowledge in continual learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, pp. 6993–7001, 2021.

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

[35] Q. Yan, D. Gong, Y. Liu, A. van den Hengel, and J. Q. Shi, "Learning bayesian sparse networks with full experience replay for continual learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 109–118, 2022.

[36] J. Schwarz, W. Czarnecki, J. Luketina, A. Grabska-Barwinska, Y. W. Teh, R. Pascanu, and R. Hadsell, "Progress & compress: A scalable framework for continual learning," in *International conference on machine learning*, pp. 4528–4537, PMLR, 2018.