Continual Learning with Lifelong Vision Transformer

Zhen Wang[1], Liu Liu[1], Yiqun Duan[3], Yajing Kong[1], Dacheng Tao[2,1]
[1]The University of Sydney, Australia, [2]JD Explore Academy, China, [3]University of Technology Sydney, Australia
{zwan4121,liu.liu1,ykon9947}@sydney.edu.au, yiqun.duan@student.uts.edu.au, dacheng.tao@gmail.com

Abstract

Continual learning methods aim at training a neural network from sequential data with streaming labels, relieving catastrophic forgetting. However, existing methods are based on and designed for convolutional neural networks (CNNs), which have not utilized the full potential of newly emerged powerful vision transformers. In this paper, we propose a novel attention-based framework Lifelong Vision Transformer (LVT), to achieve a better stability-plasticity trade-off for continual learning. Specifically, an inter-task attention mechanism is presented in LVT, which implicitly absorbs the previous tasks' information and slows down the drift of important attention between previous tasks and the current task. LVT designs a dual-classifier structure that independently injects new representation to avoid catastrophic interference and accumulates the new and previous knowledge in a balanced manner to improve the overall performance. Moreover, we develop a confidence-aware memory update strategy to deepen the impression of the previous tasks. The extensive experimental results show that our approach achieves state-of-the-art performance with even fewer parameters on continual learning benchmarks.

1. Introduction

Humans can continuously learn novel concepts throughout their lifetime and accumulate visual knowledge from past experiences [5, 69]. In contrast, artificial neural networks forget the information learned in the previous tasks while learning new ones, resulting in a drastic drop in performance on the previous tasks. This phenomenon, known as *catastrophic forgetting* or *catastrophic interference* [52, 59], stems from changes in the input data distribution that cause the new input information to interfere severely with the previously learned knowledge [8, 51]. To address this challenge, the field of continual learning (also called lifelong or incremental learning) [17, 61, 62, 72] studies the problem of learning from a non-stationary stream of data, with the goal of maintaining and extending the acquired knowledge over time.

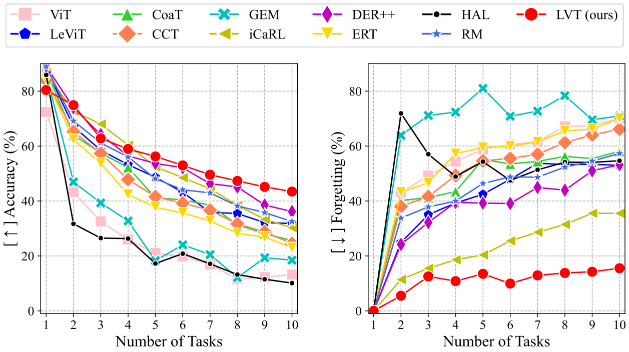


Figure 1. Incremental accuracy and forgetting evaluated on all tasks observed so far during continual learning. We compare our model with vision transformers (ViT [21], LeViT [25], CvT [83], and CCT [32]) and prior continual learning methods (GEM [47], iCaRL [61], DER++ [10], HAL [14], ERT [11], and RM [7]) on the experiment CIFAR100 of 10 splits with memory size 500. [↑] higher is better, [↓] lower is better.

Continual learning requires neural networks to be stable to prevent forgetting, but also plastic to learn new streaming labels, which is referred to as the *stability-plasticity* dilemma [27, 53]. Most of the early works in continual learning focus on the task-incremental learning (task-IL), where oracle knowledge of the task identity is available at inference time for selecting the corresponding classifier [2, 17, 44, 65, 68]. For example, regularization-based methods penalize the changes of important parameters during the learning process of new tasks and typically assign a separate output layer (classifier) for each task [13, 40, 64, 91]. Recently, various works have focused on the more difficult and realistic class-incremental learning (class-IL) [3, 9, 14, 20, 36, 61, 75, 78, 89, 94], where the network is evaluated on all classes observed during the training, without requiring the task identity. Among them, rehearsal-based methods [4, 7, 10] storing a small portion of observed data in a limited memory for replaying have shown promising results; besides, distillation-based methods [12, 61, 93] alleviate deterioration in later tasks by using knowledge distillation [34] to maintain the representation.

However, existing methods are based on and designed for convolutional neural networks (CNNs) [33], which have not taken the full utilization of the potential from newly emerged powerful vision transformers [31, 39]. Vision transformers, recently, have shown superiority on certain computer vision tasks based on the self-attention mechanism [16, 21, 25, 46, 55, 83, 90]. The merits of vision transformers bring a new perspective to the development of continual learning. Nevertheless, current vision transformers are not directly applicable to modeling a stream of tasks [31,39], since transformers lack the mechanism to prevent catastrophic forgetting on previous tasks. As shown in Figure 1, vision transformers [21, 25, 32, 83] with rehearsal strategy suffer from catastrophic forgetting and performance degradation on previous tasks. Thus, it is challenging to incorporate transformers well for further improving continual learning.

In this work, we propose a novel framework, Lifelong Vision Transformer (LVT), which plays the strengths of the attention mechanism in continual learning, achieving a better stability-plasticity trade-off. Unlike vanilla self-attention in vision transformers [21, 25, 32, 74] that derives the attention map by computing similarity between self-queries and self-keys, we propose an *inter-task attention* mechanism to obtain attention maps by computing the affinities between self-queries and a learnable external key with an attention bias, which implicitly absorbs the previous tasks information. Also, inter-task attention saves the number of parameters compared to self-attention. Besides, we consolidate the important attention weights by preventing them from changing in future tasks, thereby avoiding catastrophic forgetting of past tasks. Different from the existing rehearsal-based methods [7, 9, 10, 14, 17, 47, 64, 75, 78] that use the same classifier for learning new tasks and replaying previous data, LVT proposes to utilize two classifiers: an *injection classifier* is used to inject new task representation into the model, mitigating interference with previous tasks; and an *accumulation classifier* focuses on integrating the previous and new knowledge in a balanced manner to improve the overall performance.

Moreover, we propose a simple and effective confidence-based memory update strategy to store *impressive* exemplars in the limited memory. The impressive exemplars have the distinctive characteristics of their classes. Like memories in the brain [23, 26, 51], recalling these impressive exemplars is more beneficial for the model to consolidate previous knowledge and thus reduce forgetting. We systematically compare state-of-the-art and well-established methods for the continual learning problem in both the class-IL and task-IL settings. Experimental results show that the proposed framework significantly outperforms other methods in terms of accuracy and forgetting even with fewer parameters. Using various ablation experiments, we validate the components of our approach.

The main contributions of this paper are four-fold:

- We propose a novel attention-based framework Lifelong Vision Transformer (LVT), to achieve a better stability-plasticity trade-off for continual learning. LVT contains an inter-task attention mechanism that consolidates previous knowledge and alleviates the forgetting on previous tasks.
- LVT presents a novel dual-classifier structure to independently inject new task representation avoiding catastrophic interference, and accumulate the new and previous knowledge in a balanced manner.
- We develop a confidence-aware memory update strategy to deepen the impression of the previous tasks.
- The extensive experimental results show that our approach achieves state-of-the-art performance with even fewer parameters on continual learning benchmarks.

2. Related Work

2.1. Continual learning

Rehearsal-based methods prevent catastrophic forgetting by replaying a subset of exemplars of previous tasks stored in limited memory [3, 9, 22, 35, 36, 41, 47, 54, 58, 67, 79]. Experience Replay (ER) [60, 62, 63] jointly optimizes the network parameters by interleaving the previous task exemplars with current task data. ERT [11] further improves ER by a balanced sampling strategy and bias control. GSS [4] introduces a gradient-based sampling to store optimally selected exemplars in the memory buffer. HAL [14] complements experience replay with an additional objective, keeping intact the predictions on some *anchor* points of past tasks. GEM [47] and AGEM [15] utilize episodic memory to compute previous task gradients to constrain the current update step. iCaRL [61] trains a nearest-class-mean classifier while maintaining the representation in later tasks via a self-distillation loss term. DER++ [10] mixes rehearsal with distillation loss for retraining past experience and achieves state-of-the-art performance. RM [7] presents a sampling strategy by leveraging uncertainty and data augmentation.

Other Approaches. Regularization-based methods try to estimate the importance of each network parameter for prior tasks and penalize the changes of important parameters during the learning of new tasks [2, 13, 40, 64, 66, 91]. The difference between these works is the way to compute network parameter importance. Structure-based methods [1, 29, 43, 49, 50, 57, 68, 87] expand networks as new tasks arrive and keep the parameters of sub-networks related to previous tasks fixed. However, most structure-based methods need task identity during inference to allocate distinguished sets of parameters to distinct tasks. Label-based methods [28, 77, 80–82, 88] model streaming labels from sequential data with label relationships. The proposed method in this paper belongs to the rehearsal-based method.

2.2. Vision Transformers

Transformer is firstly proposed in [74] for machine translation tasks, and since then, transformer architectures have become the state-of-the-art models for natural language processing (NLP) tasks [19, 24, 48, 56]. The core component in a transformer is the attention module, which aggregates information from the entire input sequence. Recently, Vision Transformer (ViT) [21] makes a pure Transformer architecture scalable for image classification when the data is large enough. Following that, lots of efforts have been dedicated to improving Vision Transformers for data efficiency and model efficiency [31, 39, 90]. A popular research direction explores integrating explicit convolution or properties of convolution into the Transformer architecture [16, 73, 86, 92]. CoaT [85] designs a conv-attention module to realize relative position embeddings with convolutions. LeViT [25] replaces the uniform structure of a Transformer by a pyramid with pooling to learn convolutional-like features. CCT [32] eliminates the requirement for class token and positional embeddings through a sequence pooling strategy and the use of convolutions.

However, current vision transformers are not directly applicable to modeling a stream of tasks; existing continual learning algorithms designed for CNNs may not be optimal for vision transformers either. To this end, we propose the Lifelong Vision Transformer (LVT) with the inter-task attention designed for continual learning and achieve better performance than other transformers and CNN baselines.

3. Methodology

3.1. Problem Setup

Formally, a continual learning problem is split in a sequence of T supervised learning tasks $\mathcal{T}_t$, $t \in \{1, ..., T\}$. For task $\mathcal{T}_t$, input samples $x \in \mathcal{X}_t$ and the corresponding ground truth labels $y \in \mathcal{Y}_t$ are drawn from an i.i.d. distribution $\mathcal{D}_t$. The label space of the model is all observed classes $\cup_{i=1}^{t} \mathcal{Y}_i$ and the model is expected to predict well on the all classes. The model observes one task at a time in a sequential manner, and hence it is infeasible to optimize all observed classes jointly, but a small amount of data can be stored in a limited memory $\mathcal{M}$ for future rehearsing.

3.2. Lifelong Vision Transformer (LVT)

We propose the attention-based framework Lifelong Vision Transformer (LVT) to effectively alleviate catastrophic forgetting for continual learning. An overview of the framework is depicted in Figure 2. The major contributing components in LVT are introduced as follows:

1) *Inter-task attention* in the lifelong transformer block implicitly absorbs the previous tasks information into the attention maps and slows down the learning on attention maps based on the importance to previously observed tasks.

2) Dual-classifier: *Injection classifier* injects new task representation into the model, avoiding catastrophic interference; *Accumulation classifier* integrates past and new knowledge in a balanced manner, to improve the stability-plasticity trade-off.

3.2.1 Inter-task Attention Mechanism.

Unlike vanilla self-attention in vision transformers [21, 25, 32, 74] that derives the attention map by computing similarity between self-queries and self-keys, we propose inter-task attention mechanism to obtain attention map by computing the affinities between self-queries and a learnable external key K_W with an attention bias B, which implicitly injects previous task information in the attention mechanism. Moreover, inter-task attention can save the number of parameters compared to self-attention. When the task changes, the important weights of K_W and B are consolidated by preventing them from changing in the future tasks, thereby avoiding catastrophic forgetting of the past tasks.

Suppose the input tensor is X, we apply linear transformation with parameters W_q, W_v to obtain the vanilla self-query $Q_X = W_q X$ and self-value $V_X = W_v X$. We use a external key K_W [30] to replace input-depended self-key, and explicitly add a learnable attention bias B to the attention maps. Suppose there are H attention heads, these terms are uniformly split into H segments Q_X^h, K_W^h, V_X^h, and B^h. Then the inter-task attention mechanism computes the head-specific attention map A^h and concatenates the multi-head attention as follows:

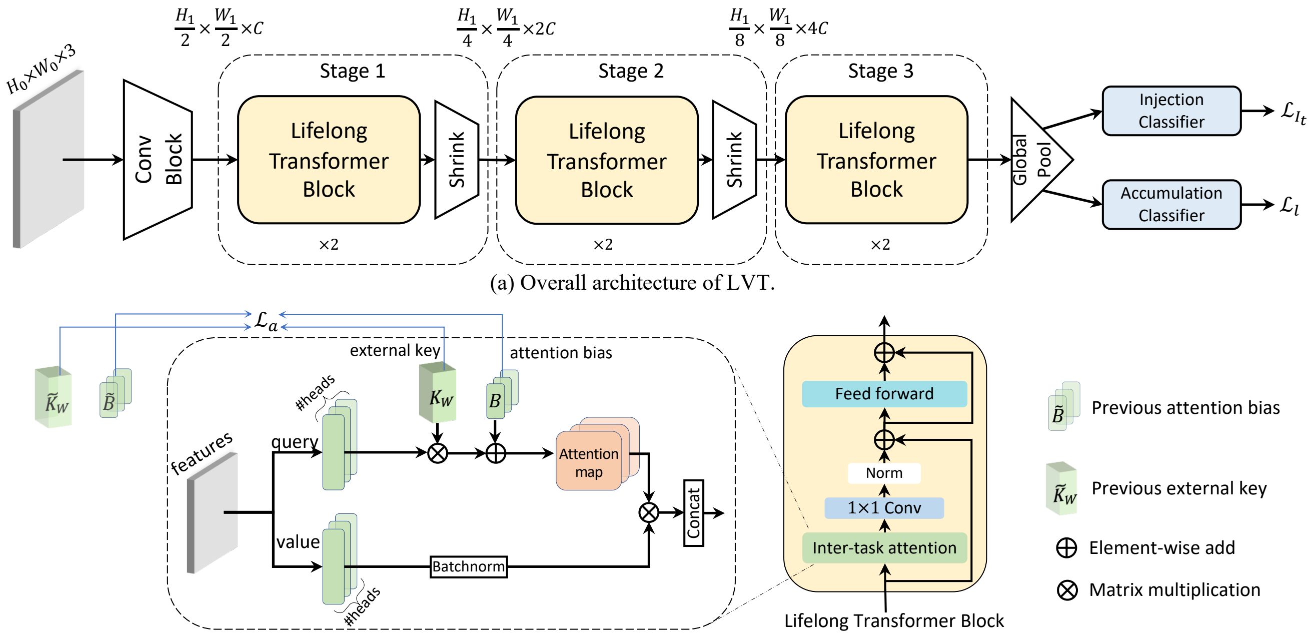
$$\begin{aligned} A^h &= \text{Softmax}\left(\frac{\text{Norm}(Q_X^h (K_W^h)^\top) + B^h}{\sqrt{d/H}}\right), \\ X_{out}^h &= A^h V_X^h, \quad h = 1, ..., H, \end{aligned} \tag{1}$$

where d is the dimension of the key and query; Norm() denotes batch normalization. The external key K_W and attention bias B do not rely on the input of current features and can be optimized by using an end-to-end way, which can capture previous tasks' information.

Furthermore, the learnable parameters external key K_W and attention bias B interact with the previous tasks by a regularization function to maintain the stability of the attention map and reduce forgetting. Specifically, we compute a weighted ℓ_1-norm between the current parameters (K_W and B) and the parameters corresponding to the last task, i.e., $\widetilde{K}_W$ and $\widetilde{B}$, given by:

$$\mathcal{L}_a = \left\| \nabla_{\widetilde{K}_W} \mathcal{L}_{I_t} \odot \left(K_W - \widetilde{K}_W \right) \right\|_1 + \left\| \nabla_{\widetilde{B}} \mathcal{L}_{I_t} \odot \left(B - \widetilde{B} \right) \right\|_1, \tag{2}$$

where $\odot$ is the Hadamard product; $\|\cdot\|_1$ denotes ℓ_1-norm; $\mathcal{L}_{I_t}$ is a cross-entropy loss defined in Eq. (3); $\nabla_{\widetilde{K}_W} \mathcal{L}_{I_t}$ and $\nabla_{\widetilde{B}} \mathcal{L}_{I_t}$ are the importance computed by the averaged gradients of loss on the last task with respect to the parameters of



(a) Overall architecture of LVT.

(b) Lifelong transformer block with inter-task attention mechanism.

Figure 2. The architecture of Lifelong Vision Transformer (LVT). (a) Overall architecture. LVT is composed of stacked lifelong transformer blocks after a simple convolutional block. Shrink module performs downsampling to reduce the resolution of the activation maps and increase their number of channels between LVT stages. After global average pooling, two classifiers serve for knowledge injection and accumulation, respectively. (b) Illustration of lifelong vision transformer block with an inter-task attention mechanism. Different from the vanilla self-attention, we apply external key and attention bias to compute attention map and interact with those of the previous task.

$\widetilde{K}_W$ and $\widetilde{B}$ respectively. During the learning of new tasks, the larger the gradient magnitude is, the greater the importance degree of the parameters will be. Thus greater penalties will be given to the more important parameters. We demonstrate that penalizing the changes in attention maps helps to retain information of the previous tasks, as new task arrives. What is worth mentioning is that this loss is similar to the Fisher information used in regularization-based methods [40, 64, 91], which prevents forgetting while allowing LVT to learn new task representations. Since K_W and B are linear learnable units, inter-task attention has less number of parameters compared with the original self-attention. The size of the two saved previous parameters $\widetilde{K}_W$ and $\widetilde{B}$ is negligible in relation to the overall model.

Normalization. Different from most vision transformers that use layer normalization (LN) [6] before each attention, we adopt batch normalization (BN) [38] after attention computing. We find that BN is more appropriate for vision transformers in continual learning than LN by using the distribution of the summed input to a neuron over a mini-batch composed of different task (non-i.i.d.) data.

3.2.2 Dual-classifier Structure

Most rehearsal-based methods [7, 9, 10, 14, 17, 47, 64, 75, 84] use the same classifier for learning the new task and replaying previous data in memory $\mathcal{M}$, which could cause catastrophic interference between new and previous tasks. To address this problem, LVT proposes to utilize a novel dual-classifier structure for independently injecting new representation without interference and accumulating the new and previous knowledge in a balanced manner.

Injection Classifier. First, we introduce the *injection classifier*. Let $g(x)$ be the feature of a sample x outputted from the backbone of LVT before the classifier. When the current task data arrives, we utilize the output from an independent injection classifier to compute a classification loss:

$$\mathcal{L}_{I_t} = \mathbb{E}_{(x,y)\sim\mathcal{D}_t}\left[\ell(y, f_I(g(x)))\right], \tag{3}$$

where f_I denotes the injection classifier; ℓ adopts a cross-entropy loss. Injection classifier is only trained on current task data and does not participate in the inference stage. The representation of current task is injected into the backbone of LVT from this classifier, to reduce the interference with previous tasks. Besides, with the benefits of the injection classifier focusing on the current task, $\mathcal{L}_{I_t}$ also serves for computing the importance weights in Eq. (2) and confidence in Eq. (8).

Accumulation Classifier. Then, we introduce the accumulation classifier as follows. Since the injection classifier mainly undertakes the representation learning on the current task, we employ a *accumulation classifier* to focus on improving the stability-plasticity trade-off by integrating previous and new knowledge in a balanced manner. The accumulation classifier is used at inference stage for outputting the prediction.

Rehearsing the limited memory data during learning new tasks is a crucial way to maintain previous knowledge. We replay the exemplars stored in the memory buffer with their ground truth labels by minimizing:

$$\mathcal{L}_r = \mathbb{E}_{(x',y')\sim\mathcal{M}}\big[\ell(y', f_A(g(x')))\big], \tag{4}$$

where f_A denotes the accumulation classifier. We approximate the expectation by computing gradients on batches sampled from the memory buffer.

Moreover, we retain the network's logits $z = h_A(g(x))$ while storing the exemplars x, where h_A is the accumulation classifier without the softmax operation, $f_A(g(x)) \triangleq \mathrm{softmax}(h_A(g(x)))$. The *dark knowledge* can be obtained by the distillation loss:

$$\mathcal{L}_d = \mathbb{E}_{(x',y',z')\sim\mathcal{M}}\big[D_{KL}\left(\mathrm{softmax}(z')\,||\,f_A(g(x'))\right)\big], \tag{5}$$

where D_{KL} denotes the KL divergence. We can set the temperature of softmax to produce suitable soft labels (targets).

Besides, the accumulation classifier also needs a supervised signal from current task data. Thanks to the injection classifier which helps learn the representation of current task, we can have the flexibility to adjust the weight of the current task in f_A with the goal of maintaining a balance between the old and new classes. Based on the above, we give the accumulation classifier loss:

$$\mathcal{L}_l = \alpha\,\mathcal{L}_r + \beta\,\mathcal{L}_d + r(t)\mathcal{L}_{A_t}, \tag{6}$$

where $\mathcal{L}_{A_t}$=$\mathbb{E}_{(x,y)\sim\mathcal{D}_t}\big[\ell(y, f_A(g(x)))\big]$; α and β are the coefficients balancing knowledge consolidation; $r(t)$ is a monotonically decreasing function with respect to t, the number of tasks observed so far, which aims to reduce the weight of the current task over time and pay more attention to fight against forgetting.

Overall, the total loss used in LVT is the sum of Eq. (2), Eq. (3), and Eq. (6):

$$\mathcal{L} = \mathcal{L}_l + \mathcal{L}_{I_t} + \gamma\mathcal{L}_a, \tag{7}$$

where γ is the coefficient balancing $\mathcal{L}_a$.

3.3. Confidence-aware Memory Update

A key question for rehearsal-based methods is how to update memory exemplars when new task arrives? Most of the methods employ the *Reservoir* sampling [76] or *Herding* sampling [61] to update memory, in which Reservoir randomly samples examples from the input stream with the same probability, and Herding stores the examples which are close to the mean of features for each class.

In this work, we design a confidence-aware sampling based on the injection classifier of LVT to store the *impressive* exemplars in the limited memory. We argue that the exemplars which are selected to be stored should have the distinctive characteristics of their classes, i.e., they can be accurately distinguished by the model. In analogy to memories in the brain [23, 26, 51], recalling these impressive exemplars can further consolidate previous knowledge for continual learning. In order to choose impressive exemplars, we propose a simple and effective sampling which stores the samples with the highest confidence scores of their classes.

Given the memory capacity M, we assign $K = \lfloor M/|\mathcal{C}| \rfloor$ exemplars for each class, where $\mathcal{C}$ is the set of classes observed so far. At the end of current task $\mathcal{T}_t$ with a set of classes $\mathcal{C}_t$, we put the samples x of each class into the model and get the logits z from the injection classifier. We can obtain the confidence score ρ of each sample by:

$$\rho(x) = \frac{e^{z^c}}{\sum_{i=1}^{|\mathcal{C}|} e^{z^i}},\ x \in \{\hat{x}|(\hat{x},\hat{y}) \in \mathcal{D}_t, \hat{y} = y_c\},\ c \in \mathcal{C}_t, \tag{8}$$

where z^i is the i-th element of z; y_c is the label corresponding to class c. We select the K samples with the highest confidence scores ρ for each class. These exemplars are not only representative for their corresponding class, but also discriminative to the other classes. We store the exemplars in a descending order based on the corresponding values of ρ, where exemplars come earlier in the order have a higher value of ρ. Memory update also includes removing exemplars of each previous class, we reduce the number of exemplars of each previous class to K in an ascending order.

4. Experiment

4.1. Experimental Setup and Implementation

We consider a strict evaluation setting [37, 72] for continual learning, which includes task incremental learning (**Task-IL**) and class incremental learning (**Class-IL**). Task-IL splits the training samples into partitions of tasks, which requires task identities to select corresponding classifiers at inference time. Class-IL sequentially increases the number of classes to be classified without requiring the task identities, as the hardest scenario [10].

Datasets. The CIFAR-100 dataset [42] contains 100 classes and each class has 500 train and 100 test color images. TinyImageNet [70] consists 200 classes that include 100,000 images for training and 10,000 images for validation. ImageNet100 [61] contains 100 classes randomly chosen from ILSVRC [18] with the average resolution 469×387. It includes about 120,000 images for training and 5,000 images for validation.

Baselines. We compare LVT with state-of-the-art and well-established methods, including eight rehearsal-based

Memory Buffer	**Method**	**#Paras**	**5 splits**		**10 splits**		**20 splits**	
			Class-IL	*Task-IL*	*Class-IL*	*Task-IL*	*Class-IL*	*Task-IL*
–	Joint	11.2	70.21±0.15	85.25±0.29	70.21±0.15	91.24±0.27	71.25±0.22	94.02±0.33
	SGD	11.2	17.27±0.14	42.24±0.33	8.62±0.09	34.40±0.53	4.73±0.06	40.83±0.46
200	ER [62]	11.2	21.94±0.83	62.41±0.93	14.23±0.12	67.57±0.68	9.90±1.67	70.82±0.74
	GEM [47]	11.2	19.73±0.34	57.13±0.94	13.20±0.21	62.96±0.67	8.29±0.18	66.28±1.49
	AGEM [15]	11.2	17.97±0.26	53.55±1.13	9.44±0.29	55.04±0.87	4.88±0.09	41.30±0.56
	iCaRL [61]	11.2	30.12±2.45	55.70±1.87	22.38±2.79	60.81±2.48	12.62±1.43	62.17±1.93
	FDR [9]	11.2	22.84±1.49	63.75±0.49	14.85±2.76	65.88±0.60	6.70±0.79	59.13±0.73
	GSS [4]	11.2	19.44±2.83	56.11±1.50	11.84±1.46	56.24±0.98	6.42±1.24	51.64±2.89
	DER++ [10]	11.2	27.46±1.16	62.55±2.31	21.76±0.78	59.54±0.77	15.16±1.53	61.98±0.91
	HAL [14]	22.4	13.21±1.24	35.61±2.95	9.67±1.67	37.49±2.16	5.67±0.91	53.06±2.87
	ERT [11]	11.2	21.61±0.87	54.75±1.32	12.91±1.46	58.49±3.12	10.14±1.96	62.90±2.72
	RM [7]	11.2	32.23±1.09	62.05±0.62	22.71±0.93	66.28±0.60	15.15±2.14	68.21±0.43
	LVT (ours)	**8.9**	**39.68**±1.36	**66.92**±0.40	**35.41**±1.28	**72.80**±0.49	**20.63**±1.14	**73.41**±0.67
500	ER [62]	11.2	27.97±0.33	68.21±0.29	21.54±0.29	74.97±0.41	15.36±1.15	74.97±1.44
	GEM [47]	11.2	25.44±0.72	67.49±0.91	18.48±1.34	72.68±0.46	12.58±2.15	**78.24**±0.61
	AGEM [15]	11.2	18.75±0.51	58.70±1.49	9.72±0.22	58.23±0.64	5.97±1.13	59.12±1.57
	iCaRL [61]	11.2	35.95±2.16	64.40±1.59	30.25±1.86	71.02±2.54	20.05±1.33	72.26±1.47
	FDR [9]	11.2	29.99±2.23	69.11±0.59	22.81±2.81	74.22±0.72	13.10±3.34	73.22±0.83
	GSS [4]	11.2	22.08±3.51	61.77±1.52	13.72±2.64	56.32±1.84	7.49±4.78	57.42±1.61
	DER++ [10]	11.2	38.39±1.57	70.74±0.56	36.15±1.10	73.31±0.78	21.65±1.44	70.55±0.87
	HAL [14]	22.4	16.74±3.51	39.70±2.53	11.12±3.80	41.75±2.17	9.71±2.91	55.60±1.83
	ERT [11]	11.2	28.82±1.83	62.85±0.28	23.00±0.58	68.26±0.83	18.42±1.92	73.50±0.82
	RM [7]	11.2	39.47±1.26	69.27±0.41	32.52±1.53	73.51±0.89	23.09±1.72	75.06±0.75
	LVT (ours)	**8.9**	**44.73**±1.19	**71.54**±0.93	**43.51**±1.06	**76.78**±0.71	**26.75**±1.29	78.15±0.42

Table 1. Results (overall accuracy %) on CIFAR100 benchmark which is averaged over five runs. #Paras means the number of parameters in the model, which is counted by million.

methods (ER [62], GEM [47], AGEM [15], GSS [4], FDR [9], HAL [14], ERT [11], and RM [7]), two methods leveraging Knowledge Distillation (iCaRL [61] and DER++ [10]). Besides, we also compare SOTA vision transformers (ViT [21], LeViT [25], CoaT [85], and CCT [32]) with rehearsal strategy for continual learning. We further provide an upper bound (JOINT) obtained by training all tasks jointly and a lower bound simply performing SGD without any countermeasure to forgetting.

Metrics. We evaluate continual learning methods in terms of accuracy and forgetting following [10, 13, 15]. The accuracy is defined by $\mathbf{A}_T{=}\frac{1}{T}\sum_{t=1}^{T} a_{T,t}$, and the forgetting is defined by $\mathbf{F}_T{=}\frac{1}{T-1}\sum_{t=1}^{T-1} \max_{i\in\{1,\dots,T-1\}}(a_{i,t}-a_{T,t})$, where $a_{T,t}$ is the testing accuracy on task $\mathcal{T}_t$ when the model completed learning task $\mathcal{T}_T$.

Implementation Details. To fairly compare each method, we train all networks using the stochastic gradient descent (SGD) optimizer. The training images are randomly cropped and flipped for all methods following [10, 11, 67]. We adopt 50 and 100 epoch with mini-batch size of 32, learning rate of 0.1 for CIFAR100 and TinyImageNet respectively, following [10, 11, 61, 91]. For ImageNet100, we resize the images to 224×224 and use batch-size of 128, step annealing learning-rate schedule ranged from 0.1 to 0.001, and the number of epochs of 100, which are used from [7, 61]. Continual learning baselines use ResNet18 [33] as backbone, and cross-entropy as classification loss, following [7, 10, 14, 15, 67, 71]. The implementation of transformer block is based on ViT [21] and LeViT [25]. LVT uses GELU activation and dropout in transformer blocks and applies a global average pooling to the last activation map. We set hyper-parameters by performing a grid-search on a validation set, which is obtained by sampling 10% from the training dataset. The details of the setup are in Appendix A.

4.2. Comparison to State-of-the-Art Methods

Evaluation on CIFAR100. We follow the protocol proposed in [61, 87], which trains all 100 classes in several splits including 5, 10, 20 incremental tasks. Table 1 summarizes the overall accuracy on CIFAR100 with 200 and 500 memory size. It is shown that LVT outperforms other methods by a considerable margin in different incremental splits, e.g., LVT can improve the accuracy of continual learning by more than **12%** in 10-split with 200 memory capacity. Especially in the case of small memory, the advantage of LVT is more obvious, which indicates LVT per-

Memory Buffer	Method	#Paras	TinyImageNet		#Paras	ImageNet100	
			Class-IL	*Task-IL*		*Class-IL*	*Task-IL*
–	Joint	11.2	59.36±0.19	81.95±0.15	11.2	73.82±0.23	81.58±0.31
	SGD	11.2	7.87±0.24	18.31±0.63	11.2	8.72±0.37	21.32±0.61
200	ER [62]	11.2	8.79±0.21	39.16 ±2.14	11.2	9.58±0.34	36.24±1.69
	AGEM [15]	11.2	8.28±0.15	23.79±0.11	11.2	9.27±0.08	25.20±0.35
	iCaRL [61]	11.2	8.64±0.78	28.41±1.53	11.2	12.59±0.68	33.75±1.81
	FDR [9]	11.2	8.77±0.82	40.15±0.67	11.2	10.08±0.36	37.80±0.91
	DER++ [10]	11.2	11.16±0.95	40.97±1.16	11.2	11.92±0.12	31.96±1.65
	ERT [11]	11.2	10.85±0.24	39.54±1.90	11.2	13.51±1.13	36.94±1.54
	RM [7]	11.2	13.58±1.07	41.96±1.28	11.2	16.76±0.84	35.18±1.43
	LVT (ours)	**9.0**	**17.34**±1.13**(+3.76)**	**46.15**±1.21**(+4.19)**	**9.4**	**19.46**±1.06**(+2.70)**	**41.78**±2.03**(+3.98)**
500	ER [62]	11.2	10.15±0.32	50.11±0.53	11.2	11.68±0.25	42.04±0.47
	AGEM [15]	11.2	9.67±0.18	26.79±0.81	11.2	10.92±0.16	34.22±0.68
	iCaRL [61]	11.2	10.69±1.53	35.89±2.47	11.2	16.44±1.35	36.89±0.72
	FDR [9]	11.2	10.58±0.22	49.91±0.78	11.2	11.78±0.40	42.60±0.64
	DER++ [10]	11.2	19.33±1.41	51.90±0.62	11.2	14.52±1.86	35.46±0.66
	ERT [11]	11.2	12.13±0.36	50.87±0.49	11.2	20.42±1.13	41.56±1.78
	RM [7]	11.2	18.96±1.34	52.08±0.84	11.2	14.56±2.64	38.66±2.47
	LVT (ours)	**9.0**	**23.97**±1.27**(+4.64)**	**57.39**±0.75**(+5.31)**	**9.4**	**26.32**±1.67**(+5.90)**	**47.84**±1.33**(+5.24)**

Table 2. Results (overall accuracy %) on TinyImageNet and ImagNet100, which are averaged over three runs. #Paras means the number of parameters in the model, which is counted by million. The **green numbers** represent gains.

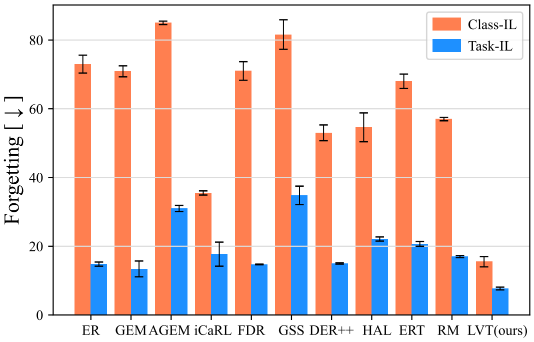


Figure 3. Forgetting results (%) on CIFAR100 (lower is better).

forms better in more realistic and challenging data-scarcity situations. It is worth noting that although LVT uses fewer parameters (8.9M) than other methods (11.2M∼22.4M), it can still achieve state-of-the-art performance. One reason is that LVT inherits the merits of transformers and designs the architecture for modeling the stream of tasks, and thus works well in continual learning without stacking a lot of parameters.

Evaluation on ImageNet datasets. Table 2 summarizes the experimental results for the TinyImageNet and ImageNet100 datasets with 10 splits. It is shown that LVT consistently surpasses other methods with a considerable margin for Class-IL and Task-IL on TinyImageNet and ImageNet100 datasets. Specifically, our method outperforms the state-of-the-art with about **5.9%** for the Class-IL accuracy on the ImageNet100 benchmark. For TinyImageNet benchmark, the Task-IL accuracy is improved from 52.08% to 57.39%(**+5.31%**). Moreover, LVT takes fewer parameters compared to other CNN-based methods.

Method	#Paras	Accuracy [↑]		Forgetting [↓]	
		Class-IL	*Task-IL*	*Class-IL*	*Task-IL*
ViT [21]	16.2	13.19	54.53	70.16	24.79
LeViT [25]	10.9	31.84	72.76	52.93	14.67
CoaT [85]	10.3	25.44	66.15	58.01	17.28
CCT [32]	**3.9**	24.50	71.37	66.17	20.20
ResNet18 [33]	11.2	36.98	73.23	47.43	15.36
LVT (ours)	8.9	**43.51**	**76.78**	**15.54**	**7.76**

Table 3. Comparison with vision transformer and CNN architectures for continual learning.

Forgetting. To compare the preventing forgetting capability, we assess the *average forgetting* [10, 13] that measures the performance degradation in subsequent tasks. Figure 3 shows that LVT suffers from less forgetting than all the other methods in both of Class-IL and Task-IL settings with memory size 500. This is because LVT constructs an inter-task attention architecture and leverages the injection and accumulation strategy, which improve the stability of the vision transformer network.

Comparison to Transformer and CNN Architectures. We compare LVT to SOTA Vision Transformers (ViT [21], LeViT [25], CvT [83], and CCT [32]) and the CNN benchmark ResNet18 [33] under the proposed rehearsal strategy in continual learning. The results from Table 3 and Figure 1 indicate that ViT is not up to the task of continual learning, since it is "data hungry" and only fits to i.i.d. large datasets. LeViT, CvT, and CCT contain CNN structures to obtain inductive biases, which improve the generalizability

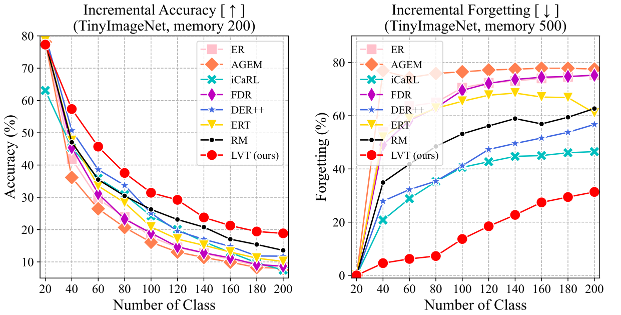


Figure 4. Incremental performance evaluated on all tasks observed so far. [↑] higher is better, [↓] lower is better.

Module				CIFAR100		TinyImageNet	
IT-*att*	f_I	f_A	ρ	*Class-IL*	*Task-IL*	*Class-IL*	*Task-IL*
	√	√	√	36.93	73.52	19.35	52.14
√		√	√	39.76	74.78	20.03	54.34
√	√		√	38.42	75.49	18.85	55.07
√	√	√		40.25	73.71	21.16	54.42
√	√	√	√	**43.51**	**76.78**	**23.97**	**57.39**

Table 4. Ablation study on each component of LVT. IT-*att* represents the transformer with inter-task attention mechanism; f_I and f_A denotes the injection classifier and accumulation classifier respectively; And ρ denotes the confidence-ware memory update.

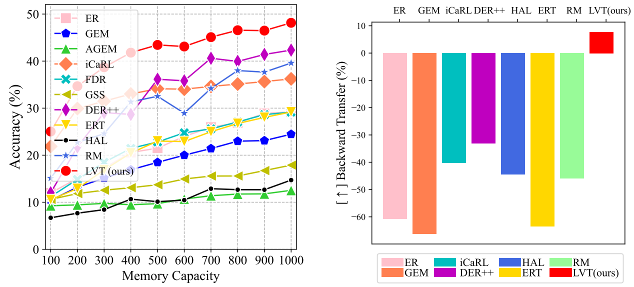
of transformer but still suffer from catastrophic forgetting in continual learning. Using Vision Transformer directly for continual learning is not even as good as ResNet performance. Only LVT exploits the strengths of transformer with even fewer parameters to achieve better performance for continual learning, which benefits from the inter-task attention mechanism and the dual-classifier structure.

Incremental Performance. We demonstrate the *average incremental performance* [10, 61] under the Class-IL setting, which is the result of evaluating on all the tasks observed so far after completing each task. As shown in Figures 1 and 4, the performance of most methods degrades rapidly as new tasks arrive, while our method consistently outperforms the state-of-the-art methods at every step in both of accuracy and forgetting.

4.3. Ablation Study and Analysis

Effect of Each Component. Table 4 shows the effect of each component of LVT on CIFAR100 and TinyImageNet with 500 memory. We can see that the average accuracy on CIFAR100 is improved significantly from 36.93% to 43.51% by the proposed transformer block with inter-task attention. The dual-classifier structure obtains 5.09% gain in Class-IL setting. The performance of the model is further improved with 2.97% gain using the confidence-ware memory update strategy on TinyImageNet.

Sensitive Analysis on Memory Size. We assess the effectiveness of the proposed method on various memory capacities. Figure 5a shows that LVT consistently performs better than other methods at various memory capacities on CIFAR100. We also notice that the improvement of LVT becomes more significant with small memory size, which illustrates that our method can be better adapted to real-world situations with limited sources.



(a) Sensitivity analysis regarding the memory capacity.

(b) Backward transfer (BWT) analysis under Class-IL setting.

Figure 5. Analyses for memory capacity and backward transfer.

Backward Transfer (BWT) Analysis. BWT [10, 15, 47] is the influence of learning a task on the performance of previous tasks, defined by BWT=$\frac{1}{T-1}\sum_{t=1}^{T-1}(a_{T,t}-a_{t,t})$, where $a_{T,t}$ is the testing accuracy on task $\mathcal{T}_t$ when the model completed learning task $\mathcal{T}_T$. We analyze BWT for different methods on CIFAR100 of 10 splits with memory 1000. As shown in Figure 5b, other methods have large negative BWT in the Class-IL setting, which means severe forgetting. In contrast, our method even achieves positive BWT, which means the learning of new tasks may help previous tasks' performance. This result further proves the superiority of our method. More detailed results in Appendix B.

5. Conclusion

To the best of our knowledge, this paper is the first in the literature to design a vision transformer for continual learning. The proposed Lifelong Vision Transformer (LVT) contains an inter-task attention mechanism and a dual-classifier structure, which can consolidate previous knowledge and alleviate the forgetting on previous tasks. Moreover, we develop a confidence-aware memory update strategy to deepen the impression of the previous tasks. Extensive experimental results show that our approach significantly outperforms current state-of-the-art methods with fewer parameters. Ablation analyses validate the effectiveness of the proposed components.

6. Acknowledgements

Dr Liu Liu, Mr Zhen Wang and Ms Yajing Kong are supported in part by ARC FL-170100117, DP-180103424, IC-190100031, and LE-200100049.

References

[1] Davide Abati, Jakub Tomczak, Tijmen Blankevoort, Simone Calderara, Rita Cucchiara, and Babak Ehteshami Bejnordi. Conditional channel gated networks for task-aware continual learning. In *CVPR*, pages 3931–3940, 2020. 2

[2] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *ECCV*, pages 139–154, 2018. 1, 2

[3] Rahaf Aljundi, Punarjay Chakravarty, and Tinne Tuytelaars. Expert gate: Lifelong learning with a network of experts. In *CVPR*, pages 3366–3375, 2017. 1, 2

[4] Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. In *NeurIPS*, 2019. 1, 2, 6

[5] P Alvarez and L R Squire. Memory consolidation and the medial temporal lobe: a simple network model. *Proceedings of the National Academy of Sciences*, 91(15):7041–7045, 1994. 1

[6] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4, 12

[7] Jihwan Bang, Heesu Kim, YoungJoon Yoo, Jung-Woo Ha, and Jonghyun Choi. Rainbow memory: Continual learning with a memory of diverse samples. In *CVPR*, pages 8218–8227, 2021. 1, 2, 4, 6, 7

[8] Yoshua Bengio, Ian Goodfellow, and Aaron Courville. *Deep learning*, volume 1. MIT press Massachusetts, USA:, 2017. 1

[9] Ari S Benjamin, David Rolnick, and Konrad P Kording. Measuring and regularizing networks in function space. *ICLR*, 2019. 1, 2, 4, 6, 7

[10] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark Experience for General Continual Learning: a Strong, Simple Baseline. In *NeurIPS*, 2020. 1, 2, 4, 5, 6, 7, 8, 12, 13

[11] Pietro Buzzega, Matteo Boschini, Angelo Porrello, and Simone Calderara. Rethinking experience replay: a bag of tricks for continual learning. In *International Conference on Pattern Recognition (ICPR)*, pages 2180–2187. IEEE, 2021. 1, 2, 6, 7

[12] Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *ECCV*, pages 233–248, 2018. 1

[13] Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *ECCV*, pages 532–547, 2018. 1, 2, 6, 7

[14] Arslan Chaudhry, Albert Gordo, Puneet K Dokania, Philip HS Torr, and David Lopez-Paz. Using hindsight to anchor past knowledge in continual learning. In *AAAI*, 2021. 1, 2, 4, 6

[15] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient Lifelong Learning with A-GEM. In *ICLR*, 2019. 2, 6, 7, 8, 13

[16] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. In *NeurIPS*, 2021. 2, 3

[17] M. Delange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 1, 2, 4

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 5

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[20] Prithviraj Dhar, Rajat Vikram Singh, Kuan-Chuan Peng, Ziyan Wu, and Rama Chellappa. Learning without memorizing. In *CVPR*, pages 5138–5146, 2019. 1

[21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1, 2, 3, 6, 7, 12

[22] Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *ECCV*, pages 86–102, 2020. 2

[23] Hagar Gelbard-Sagiv, Roy Mukamel, Michal Harel, Rafael Malach, and Itzhak Fried. Internally generated reactivation of single neurons in human hippocampus during free recall. *Science*, 322(5898):96–101, 2008. 2, 5

[24] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *CVPR*, pages 244–253, 2019. 3

[25] Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. Levit: A vision transformer in convnet's clothing for faster inference. In *ICCV*, pages 12259–12269, October 2021. 1, 2, 3, 6, 7, 12

[26] Matthew D. Grilli and Elizabeth L. Glisky. Imagining a better memory: Self-imagination in memory-impaired patients. *Clinical Psychological Science*, 1(1):93–99, 2013. 2, 5

[27] Stephen Grossberg. Adaptive resonance theory: How a brain learns to consciously attend, learn, and recognize a changing world. *Neural networks*, 37:1–47, 2013. 1

[28] Jiaxian Guo, Mingming Gong, Tongliang Liu, Kun Zhang, and Dacheng Tao. LTF: A label transformation framework for correcting label shift. In *ICML*, volume 119, pages 3843–3853, 2020. 2

[29] Jiaxian Guo, Mingming Gong, and Dacheng Tao. A relational intervention approach for unsupervised dynamics generalization in model-based reinforcement learning. In *ICLR*, 2022. 2

[30] Meng-Hao Guo, Zheng-Ning Liu, Tai-Jiang Mu, and Shi-Min Hu. Beyond self-attention: External attention using two linear layers for visual tasks. *arXiv preprint arXiv:2105.02358*, 2021. 3

[31] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, Zhaohui Yang, Yiman Zhang, and Dacheng Tao. A survey on visual transformer. *ArXiv*, abs/2012.12556, 2020. 2, 3, 13

[32] Ali Hassani, Steven Walton, Nikhil Shah, Abulikemu Abuduweili, Jiachen Li, and Humphrey Shi. Escaping the big data paradigm with compact transformers. *arXiv preprint arXiv:2104.05704*, 2021. 1, 2, 3, 6, 7, 13

[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2, 6, 7

[34] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. In *NeurIPS workshop*, 2014. 1, 13

[35] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Lifelong learning via progressive distillation and retrospection. In *ECCV*, pages 437–452, 2018. 2

[36] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *CVPR*, pages 831–839, 2019. 1, 2

[37] Yen-Chang Hsu, Yen-Cheng Liu, Anita Ramasamy, and Zsolt Kira. Re-evaluating Continual Learning Scenarios: A Categorization and Case for Strong Baselines. In *NeurIPS Continual learning Workshop*, 2018. 5

[38] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, pages 448–456. PMLR, 2015. 4, 12

[39] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *arXiv preprint arXiv:2101.01169*, 2021. 2, 3, 13

[40] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. 1, 2, 4

[41] Yajing Kong, Liu Liu, Jun Wang, and Dacheng Tao. Adaptive curriculum learning. In *ICCV*, pages 5067–5076, 2021. 2

[42] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5

[43] Xilai Li, Yingbo Zhou, Tianfu Wu, Richard Socher, and Caiming Xiong. Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting. In *ICML*, 2019. 2

[44] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(12), 2017. 1

[45] Xuan Liu, Xiaoguang Wang, and Stan Matwin. Improving the interpretability of deep neural networks with knowledge distillation. In *IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2018. 13

[46] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 2

[47] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *NeurIPS*, 2017. 1, 2, 4, 6, 8, 13

[48] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421. Association for Computational Linguistics, Sept. 2015. 3

[49] Arun Mallya, Dillon Davis, and Svetlana Lazebnik. Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In *ECCV*, 2018. 2

[50] Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *CVPR*, 2018. 2

[51] James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419, 1995. 1, 2, 5

[52] Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. volume 24 of *Psychology of Learning and Motivation*, pages 109–165. Academic Press, 1989. 1

[53] Martial Mermillod, Aurélia Bugaiska, and Patrick Bonin. The stability-plasticity dilemma: Investigating the continuum from catastrophic forgetting to age-limited learning effects. *Frontiers in psychology*, 4:504, 2013. 1

[54] Oleksiy Ostapenko, Mihai Puscas, Tassilo Klein, Patrick Jahnichen, and Moin Nabi. Learning to remember: A synaptic plasticity driven framework for continual learning. In *CVPR*, pages 11321–11329, 2019. 2

[55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 2

[56] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 3

[57] Jathushan Rajasegaran, Munawar Hayat, Salman H Khan, Fahad Shahbaz Khan, and Ling Shao. Random path selection for continual learning. In *NeurIPS*, 2019. 2

[58] Amal Rannen, Rahaf Aljundi, Matthew B Blaschko, and Tinne Tuytelaars. Encoder based lifelong learning. In *ICCV*, pages 1320–1328, 2017. 2

[59] Roger Ratcliff. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*, 97(2):285, 1990. 1

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

[61] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *CVPR*, 2017. 1, 2, 5, 6, 7, 8

[62] Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to Learn without Forgetting by Maximizing Transfer and Minimizing Interference. In *ICLR*, 2019. 1, 2, 6, 7

[63] Anthony Robins. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7(2):123–146, 1995. 2

[64] Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & Compress: A scalable framework for continual learning. In *ICML*, 2018. 1, 2, 4

[65] Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. In *ICML*, volume 80, pages 4548–4557, 2018. 1

[66] Yujun Shi, Li Yuan, Yunpeng Chen, and Jiashi Feng. Continual learning via bit-level information preserving. In *CVPR*, pages 16674–16683, 2021. 2

[67] Christian Simon, Piotr Koniusz, and Mehrtash Harandi. On learning the geodesic path for incremental learning. In *CVPR*, pages 1591–1600, 2021. 2, 6

[68] Pravendra Singh, Pratik Mazumder, Piyush Rai, and Vinay P Namboodiri. Rectification-based knowledge retention for continual learning. In *CVPR*, pages 15282–15291, 2021. 1, 2

[69] Paul Smolen, Douglas A Baxter, and John H Byrne. How can memories last for days, years, or a lifetime? proposed mechanisms for maintaining synaptic potentiation and memory. *Learning & Memory*, 26(5):133–150, 2019. 1

[70] Stanford. Tiny ImageNet Challenge (CS231n), 2015. http://tiny-imagenet.herokuapp.com/. 5

[71] Shixiang Tang, Dapeng Chen, Jinguo Zhu, Shijie Yu, and Wanli Ouyang. Layerwise optimization by gradient decomposition for continual learning. In *CVPR*, pages 9634–9643, 2021. 6

[72] Gido M van de Ven and Andreas S Tolias. Three continual learning scenarios. *NeurIPS Continual Learning Workshop*, 2018. 1, 5

[73] Ashish Vaswani, Prajit Ramachandran, Aravind Srinivas, Niki Parmar, Blake Hechtman, and Jonathon Shlens. Scaling local self-attention for parameter efficient visual backbones. In *CVPR*, pages 12894–12904, 2021. 3

[74] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 2, 3

[75] Eli Verwimp, Matthias De Lange, and Tinne Tuytelaars. Rehearsal revealed: The limits and merits of revisiting samples in continual learning. In *ICCV*, 2021. 1, 2, 4

[76] Jeffrey S Vitter. Random sampling with a reservoir. *ACM Transactions on Mathematical Software (TOMS)*, 11(1):37–57, 1985. 5

[77] Zhen Wang, Yiqun Duan, Liu Liu, and Dacheng Tao. Multi-label few-shot learning with semantic inference. In *AAAI*, volume 35, pages 15917–15918, 2021. 2

[78] Zhen Wang, Liu Liu, Yiqun Duan, and Dacheng Tao. Continual learning through retrieval and imagination. In *AAAI*, 2021. 1, 2

[79] Zhen Wang, Liu Liu, Yiqun Duan, and Dacheng Tao. Continual learning with embeddings: Algorithm and analysis. In *ICML 2021 Workshop on Theory and Foundation of Continual Learning*, 2021. 2

[80] Zhen Wang, Liu Liu, Yiqun Duan, and Dacheng Tao. Sin: Semantic inference network for few-shot streaming label learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 2

[81] Zhen Wang, Liu Liu, and Dacheng Tao. Deep streaming label learning. In *ICML*, volume 119, pages 9963–9972, 2020. 2

[82] Tong Wei, Jiang-Xin Shi, and Yu-Feng Li. Probabilistic label tree for streaming multi-label learning. In *SIGKDD*, pages 1801–1811, 2021. 2

[83] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *ICCV*, pages 22–31, October 2021. 1, 2, 7, 12, 13

[84] Haoning Xi, Liu He, Yi Zhang, and Zhen Wang. Differentiable road pricing for environment-oriented electric vehicle and gasoline vehicle users in the bi-objective transportation network. *Transportation Letters*, pages 1–15, 2021. 4

[85] Weijian Xu, Yifan Xu, Tyler Chang, and Zhuowen Tu. Co-scale conv-attentional image transformers. In *ICCV*, pages 9981–9990, October 2021. 3, 6, 7

[86] Yufei Xu, Qiming Zhang, Jing Zhang, and Dacheng Tao. Vitae: Vision transformer advanced by exploring intrinsic inductive bias. *Advances in Neural Information Processing Systems*, 34, 2021. 3

[87] Shipeng Yan, Jiangwei Xie, and Xuming He. Der: Dynamically expandable representation for class incremental learning. In *CVPR*, 2021. 2, 6, 12

[88] Shan You, Chang Xu, Yunhe Wang, Chao Xu, and Dacheng Tao. Streaming Label Learning for Modeling Labels on the Fly. *arXiv preprint arXiv:1604.05449*, 2016. 2

[89] Lu Yu, Bartlomiej Twardowski, Xialei Liu, Luis Herranz, Kai Wang, Yongmei Cheng, Shangling Jui, and Joost van de Weijer. Semantic drift compensation for class-incremental learning. In *CVPR*, pages 6982–6991, 2020. 1, 12

[90] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis E.H. Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *ICCV*, pages 558–567, October 2021. 2, 3, 12, 13

[91] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *ICML*, 2017. 1, 2, 4, 6

[92] Qiming Zhang, Yufei Xu, Jing Zhang, and Dacheng Tao. Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond. *arXiv preprint arXiv:2202.10108*, 2022. 3

[93] Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, and Shu-Tao Xia. Maintaining discrimination and fairness in class incremental learning. In *CVPR*, 2020. 1

[94] Junjie Zhu, Bingjun Luo, Sicheng Zhao, Shihui Ying, Xibin Zhao, and Yue Gao. Iexpressnet: Facial expression recognition with incremental classes. In *ACM International Conference on Multimedia*, pages 2899–2908, 10 2020. 1