Boosting Continual Learning of Vision-Language Models via Mixture-of-Experts Adapters

Jiazuo Yu[1], Yunzhi Zhuge[1], Lu Zhang[1,*], Ping Hu[2], Dong Wang[1], Huchuan Lu[1] and You He[3]
[1] Dalian University of Technology, China
[2] University of Electronic Science and Technology of China
[3] Tsinghua University, China
yujiazuo@mail.dlut.edu.cn, zhangluu@dlut.edu.cn

Abstract

Continual learning can empower vision-language models to continuously acquire new knowledge, without the need for access to the entire historical dataset. However, mitigating the performance degradation in large-scale models is non-trivial due to (i) parameter shifts throughout lifelong learning and (ii) significant computational burdens associated with full-model tuning. In this work, we present a parameter-efficient continual learning framework to alleviate long-term forgetting in incremental learning with vision-language models. Our approach involves the dynamic expansion of a pre-trained CLIP model, through the integration of Mixture-of-Experts (MoE) adapters in response to new tasks. To preserve the zero-shot recognition capability of vision-language models, we further introduce a Distribution Discriminative Auto-Selector (DDAS) that automatically routes in-distribution and out-of-distribution inputs to the MoE Adapter and the original CLIP, respectively. Through extensive experiments across various settings, our proposed method consistently outperforms previous state-of-the-art approaches while concurrently reducing parameter training burdens by 60%. Our code locates at https://github.com/JiazuoYu/MoE-Adapters4CL

1. Introduction

Artificial Intelligence (AI), particularly in the realm of large-scale foundation models, has made significant strides in understanding the open world, as evidenced by recent advancements [44, 45, 56, 61, 67]. An ideal AI, akin to human cognition, should be able to continuously assimilate new knowledge from the dynamic environment. Traditional fully-supervised training paradigms can't adapt to this scenario due to the high computational costs of integrating new data with historical datasets. In contrast, Continual Learning (CL), offering an efficient incremental training strategy, emerges as a solution by focusing on new data at each training stage. However, CL faces the significant hurdle of "catastrophic forgetting" where a model loses previously acquired knowledge upon learning new tasks [24, 52].

To remedy this issue, one of the popular solutions in current CL methods [1, 16, 23, 43] is to develop dynamic expansion frameworks by incrementally adding task-specific components to a shared base model (see Figure 1 (a)). Although these methods show promise in memorization and scalability, they cannot distinguish unseen data and thus overlook zero-shot transfer capability. Recent advancements like ZSCL [79] have brought the zero-shot transfer ability into continual learning by leveraging a pretrained Vision Language Model (VLM). As illustrated in Figure 1 (b), this method relies on knowledge distillation to integrate zero-shot generalization ability from the frozen CLIP and uses parameter regularization to prevent knowledge degradation in continual learning. However, these designs often entail large computational burdens and exhibit limitations in long-term memorization. It's then natural to ask whether we can combine the merits of the pretrained foundation model and dynamic expansion strategy to form an effective system with robust memorization and zero-shot transfer abilities.

Recently, Parameter-Efficient Fine-Tune (PEFT) methods [22, 28, 30, 66, 74, 77] have demonstrated that large-scale models can quickly adapt to downstream tasks via only fine-tuning less-parameterized adapters. This inspires us to build a dynamic expansion framework on VLM with task-specific adapters to relieve the parameter burdens in long-term CL. Nevertheless, the intuitive approach of stacking adapters during incremental learning introduces a dependency on task identity. This poses challenges in practical scenarios such as class incremental learning where task identity may be unavailable. Furthermore, the use of independent adapters neglects the potential for inter-task knowl-

[*] Corresponding author

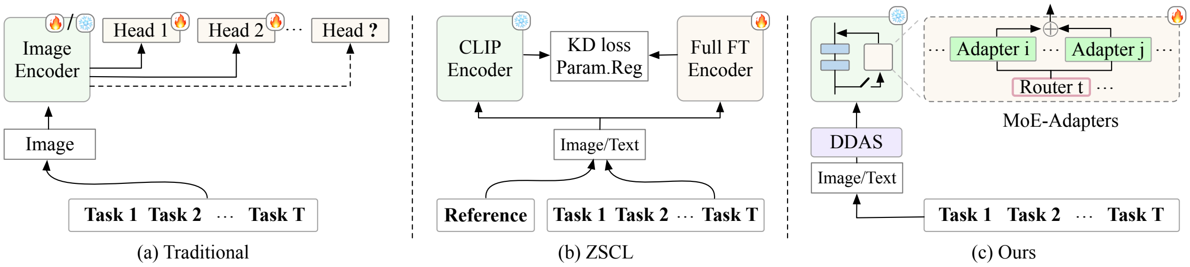


Figure 1. Comparison of various popular architectures to address CL. (a) Traditional dynamic expansion-based CL cannot distinguish unseen data. (b) Zero-shot CL [79] suffers from significant computational burdens. (c) The proposed MoE-Adapters and DDAS collaborate to form a parameter-efficient, zero-shot CL.

edge sharing and cooperation, resulting in a limited representation capability and efficacy.

To overcome the outlined challenges, we propose a parameter-efficient continual learning framework by leveraging the recent advance in the field of multi-task learning, Mixture-of-Experts (MoE) [32, 64]. We build a dynamic expansion architecture on a frozen CLIP model [61], dubbed as incremental MoE-Adapters, in which we take adapters as sparse experts and utilize incrementally incorporated task-specific routers to select the corresponding experts. In the continual learning process, we further apply a novel activate-freeze strategy to help the experts learn intra-task knowledge and encourage inter-task collaboration. Additionally, a Distribution Discriminative Auto-Selector (DDAS) is proposed to automatically allocate the testing data to MoE-Adapters or the pretrained CLIP, enabling effective predictions for seen data and zero-shot transfer for unseen data within a unified framework.

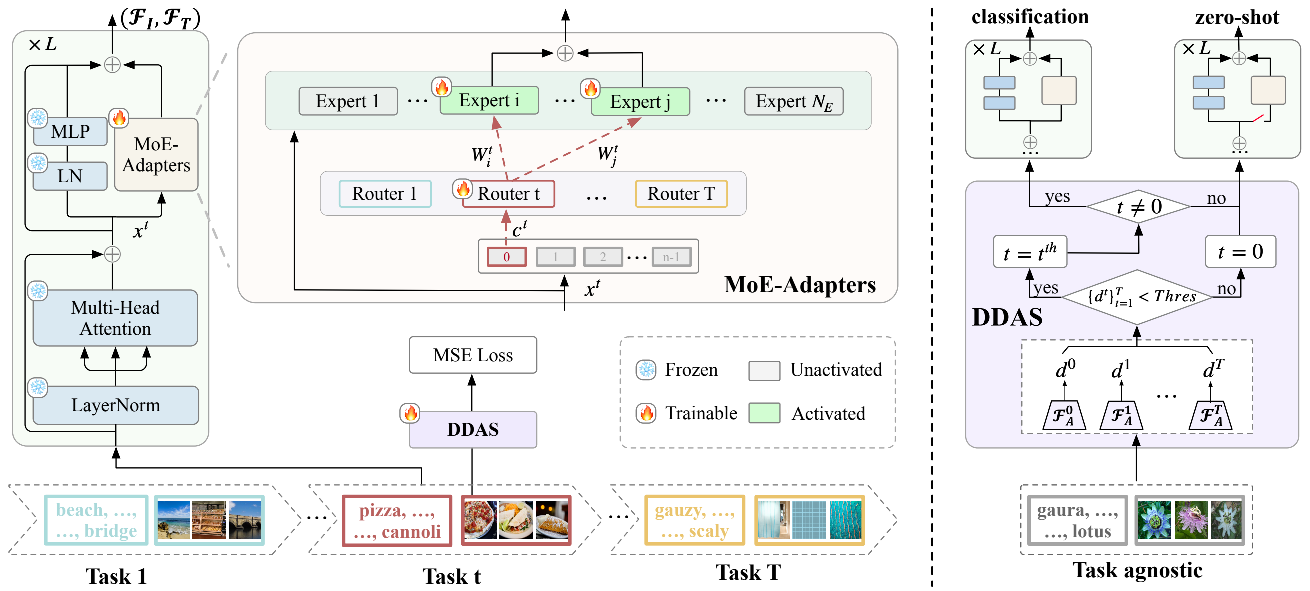
Our extensive experiments across various settings demonstrate the proposed method's effectiveness in addressing the catastrophic forgetting issue, significantly reducing the 60% parameter burdens and memory requirements during training. Furthermore, when applied to few-shot continual learning, the proposed model shows exceptional resistance to forgetting and outperforms the previous arts by 3.6%, 7.0% and 4.2% in a 5-shot setting. Our contributions can be summarized as follows:

- We introduce a parameter-efficient training framework for vision-language models in continual learning, employing a MoE-Adapters based dynamic expansion architecture for enhanced adaptability and efficiency.
- We develop an incremental activate-freeze strategy in the MoE framework, enabling experts to simultaneously acquire intra-task knowledge and engage in inter-task collaboration.
- We design a Distribution Discriminative Auto-Selector (DDAS) for automated substream assignment, effectively merging anti-forgetting and zero-shot transfer capabilities within a unified model.

2. Related Works

Continual Learning. Depending on the domain variations of incremental data, existing continual learning methods mainly focus on addressing *i.e.,* Class Incremental Learning (CIL) [3, 12, 34, 46, 71] and Task Incremental Learning (TIL) [51, 57, 79]. Existing efforts in this area have been made by developing various architectures [11], including memory-based, regularization-based and dynamic-based models. Memory-based methods [31, 41, 48, 58, 60, 62, 65] retain the historical knowledge by storing them in a memory bank, which will be accessed and updated in incremental learning. However, the continuously increasing learned data usually poses a burden on the memory bank, resulting in limited lifelong learning ability. Regularization-based methods add explicit regularization terms on weights [2, 37, 42, 75] or data [14, 18, 26, 43] to balance between the older and new tasks. They are usually used as an auxiliary trick in memory-based or dynamic models to alleviate the forgetting issue. Dynamic methods [1, 19, 29, 71–73] address continual learning by incrementally adding new parameters on the baseline, such as neurons, branches or prediction heads. Dynamic methods usually perform favorably against the other two pipelines. However, like memory-based methods, the dynamic architecture often incurs large-scale model sizes, limiting the models' efficiency. Despite the promising performance of the approaches aforementioned, addressing the crucial capability of AI agents, namely zero-shot transfer to unseen knowledge, remains challenging and complex to integrate into existing popular pipelines. In this paper, we propose incorporating the dynamic architecture on vision-language models to boost their memorization of historical knowledge and alleviate the degradation of zero-shot transfer ability. The highly related work is ZSCL [79], which uses parameter regularization in the continual learning of large-scale models. In contrast to the fully finetuning strategy in ZSCL, our method proposed an incremental MoE adapter to decrease the tuned parameters and enhance the collaboration of historically learned adapters and ongoing ones.

Parameter Efficient Fine-Tuning. In the realm of Natural



(a) The training stage

(b) The inference stage

Figure 2. Overall framework of the proposed method. (a) At the training stage, CLIP's image and text encoders $(\mathcal{F}_I, \mathcal{F}_T)$ take input samples from **Task *t***. In each of transformer blocks, there is a MoE-Adapters, whose input is the tokens $\boldsymbol{x}^t$ from MHSA. The router takes the task-specific `[CLS]` token $\boldsymbol{c}^t$ as input and produces experts' weights W_i^t and W_j^t to combine the expert's output. DDAS is trained using only images via the MSE loss defined by Eq. 3. (b) At the inference stage, the proposed DDAS determines the data distribution by comparing the distribution $\{d^t\}_{t=1}^T$ in each autoencoder of the **task-agnostic** images. It can automatically assign the testing data into MoE-Adapters or original CLIP to predict with either seen or unseen data.

Language Processing (NLP), fine-tuning large-scale models (*e.g.,* 175B GPT-3 [5]) imposes significant burdens in both parameter complexity and time consumption. Thus, several parameter-efficient fine-tuning methods [27, 28, 33, 36, 76] have been explored, which only set a few trainable parameters and fine-tune them for efficiency. Among these methods, LoRA [28] and Compacter [36] reduce the number of trainable parameters by attaching low-rank hyper-complex adapter layers or sharing adapter parameters across layers, respectively. The success of efficient tuning strategies in NLP promotes their applications on vision-language models [22, 35, 66, 77, 80] like CLIP [22]. Recently, Liu *et al.* [46] introduce efficient tuning strategies in the continual learning of CLIP, which uses trainable adapters and a parameter retention strategy for downstream task adaptation and historical knowledge memorization, respectively. However, this method is only applied to CIL and ignores the zero-shot ability of the original CLIP. In this paper, we propose a novel parameter-efficient tuning approach on CLIP to boost both the anti-forgetting and zero-shot abilities in continual learning. Our model can flexibly adapt to CIL and TIL and achieve promising performance even trained with few data (namely few-shot continual learning).

Mixture of Experts. The MoE [32] contains multiple experts and a routing network. It aggregates the expert outputs via a weighted strategy by the routing network. Based on the sparse architecture of MoE [64], some methods [10, 20, 54, 63] are proposed to decrease computational costs and improve model capacity. This technique is also introduced to continual learning to mitigate the forgetting issue. For example, Aljundi *et al.* [1] propose to train multiple backbones as experts and automatically feed the test samples to a relevant expert. Chen *et al.* [7] utilize the pre-trained experts and gates to store previous knowledge. These methods have demonstrated MoE's promising performance in continual learning. We propose an incremental MoE-Adapters for continual learning with CLIP. We use adapters as experts to increase the adaption speed and introduce an incremental expert interaction strategy to facilitate the collaboration of experts during continual learning.

3. Methodology

3.1. Continual Learning

Given a set of T tasks $\{\mathcal{T}^t\}_{t=1}^T$, continual learning works by sequentially accessing and learning on each task $\mathcal{T}^t = \{\mathcal{D}^t, \mathcal{C}^t\}$. Here, $\mathcal{D}^t = \{I_i^t, y_i^t\}_{i=1}^{N^t}$ represents the data of t^{th} task $\mathcal{T}^t$, where I_i^t is the input image, $y_i^t \in \mathcal{C}^t$ is the corresponding class label, and N^t is the size of data. The category set $\mathcal{C}^t = \{c_j^t\}_{j=1}^{M^t}$ encompasses the class names within $\mathcal{T}^t$, with a total of M^t classes. Continual learning aims to achieve good performance across all tasks and can be broadly categorized into Task Incremental Learning (TIL) and Class Incremental Learning (CIL). In TIL, the model generates predictions within a task-specific set $\mathcal{C}^t$, which is determined by the current task identity t. Meanwhile, in CIL, the challenge involves distinguishing between all the previously encountered classes $\cup_{i=1}^{t}\mathcal{C}^i$.

3.2. Framework Overview

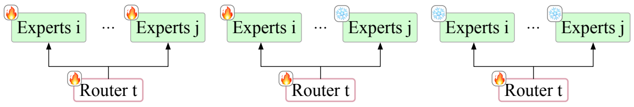
In this paper, we present a parameter-efficient framework designed to empower the continual learning capabilities of vision-language models, achieving robust historical knowledge memorization without sacrificing the zero-shot generalization abilities. Our method is built upon the CLIP [22] model, which contains parallel encoders $(\mathcal{F}_I, \mathcal{F}_T)$ to extract features of input images and texts, respectively. By following CLIP [22], we make predictions based on the cosine similarity between the final image embedding $\mathcal{F}_I(I_i^t)$ and text embedding $\mathcal{F}_T(c_j^t)$. The input image is then assigned to the class with the highest similarity.

The overall framework of our method is shown in Figure 2. We introduce MoE structure onto a frozen CLIP to consolidate all the downstream tasks within a unified model, in which the task-dependent routers are sequentially added to modulate the experts for each task. Adapter modules, such as LoRA [28], function as the experts in the MoE setup, enhancing adaptation speed during training. To relieve MoE's reliance on task identities, we further propose a Distribution Discriminative Auto-Selector (DDAS). The DDAS automatically infers the task context by analyzing the variations of the target image distributions. As a result, in-distribution data will be allocated to the corresponding routers within MoE, while the out-of-distribution inputs will be identified and directed to the original CLIP to perform zero-shot recognition.

3.3. Incremental Mixture-of-Experts Adapters

We leverage MoE [64] to build an expansible architecture to alleviate the "catastrophic forgetting" issue in the continual learning of CLIP. The MoE is composed of several experts $\{\mathcal{E}_i\}_{i=1}^{N_E}$ and routers, where N_E is the number of predefined experts. For current task $\mathcal{T}^t$, only a task-dependent router $\mathcal{R}^t, t \in [1, T]$ is added to the system, which integrates the experts' outputs via gated average.

Adapters as Experts. MoE in vision-language models usually incorporate the experts inside networks, which can be MLPs or attention heads [8, 64, 78]. However, inserting the MoE inside VLM might bring in significant computational burdens due to the full-parameter tuning. Some methods [22, 46] have demonstrated that adapters with few parameters can increase the adaption speed of VLM on downstream tasks and enable their applications in continual learning. Inspired by this, we use the effective adapter LoRA [28], which works by decoupling the original heavy and frozen parameters into low-rank trainable space, as the experts in MoE to speed up continual learning with CLIP.



(a) Both trainable (b) Mixture (c) Both frozen

Figure 3. The three distinct combinations among activated experts (a) both trained, (b) trainable and frozen, (c) both experts are frozen, and only the router is trainable.

Our MoE-Adapters are implemented in all the Transformer blocks of the parallel encoders $(\mathcal{F}_I, \mathcal{F}_T)$, as shown in Figure 2. To be more specific, in each Transformer block, the feature tokens $x^t \in \mathbb{R}^{n \times d}$ after the multi-head attention output are passed to all the experts in the MoE. Then, the task-specific router $\mathcal{R}^t$ is applied to fuse the experts' outputs via gated summation. Note that we implement the same MoE adapters in both the image encoder and the textual encoder, without sharing the parameters.

Incremental Mixture of Experts. In our MoE framework, task-specific routers $\mathcal{R}^t$ determine the activation of experts $\mathcal{E}_i$ to produce outcomes tailored to each task t. The combined output for a task, $\boldsymbol{y}^t$, is computed as:

$$\boldsymbol{y}^t = \sum_{i=1}^{N_E} W_i^t \mathcal{E}_i(\boldsymbol{x}^t), \tag{1}$$

where $W^t = \{W_i^t\}_{i=1}^{N_E}$ represents the gating weights assigned by $\mathcal{R}^t$, dictating each expert's contribution. $\boldsymbol{x}^t$ denotes the tokens processed for task t, and $\boldsymbol{y}^t$ is the corresponding output from the MoE-Adapters, matching the shape of $\boldsymbol{x}^t$. We refine the MoE-Adapters for continual learning with two key modifications. Unlike previous methods [10, 63] that input patch or image tokens into the router, we utilize the initial token, known as the `[CLS]` token ($\boldsymbol{c}^t \in \mathbb{R}^{1 \times d}$), to enhance processing efficiency. The gating weights are then computed as follows:

$$W^t = Softmax(Topk(\mathcal{R}^t(\boldsymbol{c}^t))), \tag{2}$$

where $\mathcal{R}^t$ projects $\boldsymbol{c}^t$ to a 1-D vector indicating each expert's likelihood of activation. The $Topk(\cdot)$ function selects the k most relevant experts, while setting the rest to be $-\infty$. The $Softmax(\cdot)$ function normalizes these weights to emphasize the selected experts' contribution.

Training MoE-Adapters. We train the MoE-Adapters through simple back-propagation, orchestrated by an incremental activate-freeze strategy. The objective is to augment experts with intra-task knowledge and inter-task collaboration. Specifically, after training on an older task, we count the distribution of its router's outputs. The Top-k most activated experts are then kept frozen during subsequent task training to preserve task-specific knowledge. In this manner, when faced with a new task, the respective router is able

	Method	Aircraft [50]	Caltech101 [21]	CIFAR100 [39]	DTD [9]	EuroSAT [25]	Flowers [55]	Food [4]	MNIST [13]	OxfordPet [59]	Cars [38]	SUN397 [70]	*Average*
CLIP	Zero-shot	24.3	88.4	68.2	44.6	54.9	71.0	88.5	59.4	89.0	64.7	65.2	65.3
	Full Fine-tune	62.0	95.1	89.6	79.5	98.9	97.5	92.7	99.6	94.7	89.6	81.8	89.2
	Fine-tune Adapter	56.8	92.6	89.4	79.0	98.4	97.0	92.9	99.2	94.1	89.1	82.7	88.3
Transfer	Continual-FT		67.1	46.0	32.1	35.6	35.0	57.7	44.1	60.8	20.5	46.6	44.6
	LwF [43]		74.5	56.9	39.1	**51.1**	52.6	72.8	60.6	75.1	30.3	55.9	58.9
	iCaRL [62]		56.6	44.6	32.7	39.3	46.6	68.0	46.0	77.4	31.9	60.5	50.4
	LwF-VR [15]		77.1	61.0	40.5	45.3	54.4	74.6	47.9	76.7	36.3	58.6	57.2
	WiSE-FT [68]		73.5	55.6	35.6	41.5	47.0	68.3	53.9	69.3	26.8	51.9	52.3
	ZSCL [79]		86.0	67.4	**45.4**	50.4	69.1	87.6	**61.8**	86.8	60.1	**66.8**	68.1
	Ours†		**87.9**	**68.2**	42.2	41.4	68.7	**88.7**	59.4	**89.1**	**64.5**	64.0	67.4**(-0.7)**
	Ours		**87.9**	**68.2**	44.4	49.9	**70.7**	**88.7**	59.7	**89.1**	**64.5**	65.5	**68.9(+0.8)**
Average	Continual-FT	25.5	81.5	59.1	53.2	64.7	51.8	63.2	64.3	69.7	31.8	49.7	55.9
	LwF [43]	36.3	86.9	72.0	59.0	73.7	60.0	73.6	74.8	80.0	37.3	58.1	64.7
	iCaRL [62]	35.5	89.2	72.2	60.6	68.8	70.0	78.2	62.3	81.8	41.2	62.5	65.7
	LwF-VR [15]	29.6	87.7	74.4	59.5	72.4	63.6	77.0	66.7	81.2	43.7	60.7	65.1
	WiSE-FT [68]	26.7	86.5	64.3	57.1	65.7	58.7	71.1	70.5	75.8	36.9	54.6	60.7
	ZSCL [79]	45.1	**92.0**	80.1	64.3	**79.5**	81.6	**89.6**	**75.2**	88.9	64.7	**68.0**	75.4
	Ours†	**54.3**	91.1	**85.1**	**69.7**	77.5	**84.5**	89.1	73.8	89.2	**69.0**	65.8	**77.2(+1.8)**
	Ours	50.2	91.9	83.1	69.4	78.9	84.0	89.1	73.7	**89.3**	67.7	66.9	76.7**(+1.3)**
Last	Continual-FT	31.0	89.3	65.8	67.3	88.9	71.1	85.6	**99.6**	92.9	77.3	81.1	77.3
	LwF [43]	26.3	87.5	71.9	66.6	79.9	66.9	83.8	**99.6**	92.1	66.1	80.4	74.6
	iCaRL [62]	35.8	**93.0**	77.0	70.2	83.3	88.5	90.4	86.7	93.2	81.2	81.9	80.1
	LwF-VR [15]	20.5	89.8	72.3	67.6	85.5	73.8	85.7	**99.6**	93.1	73.3	80.9	76.6
	WiSE-FT [68]	27.2	90.8	68.0	68.9	86.9	74.0	87.6	**99.6**	92.6	77.8	81.3	77.7
	ZSCL [79]	40.6	92.2	81.3	70.5	94.8	90.5	**91.9**	98.7	**93.9**	85.3	80.2	83.6
	Ours†	**54.3**	90.8	**88.8**	**80.3**	**98.1**	**97.5**	89.6	99.1	89.5	**89.2**	**83.8**	**87.4(+3.8)**
	Ours	49.8	92.2	86.1	78.1	95.7	94.3	89.5	98.1	89.9	81.6	80.0	85.0**(+1.4)**

Table 1. Comparison with state-of-the-art methods on MTIL benchmark in terms of "Transfer", "Average", and "Last" scores (%). "Ours†" and "Ours" indicate our method trained on 3k and 1k iterations, respectively. We label the best and second methods with **bold** and underline styles. The top block indicates the upper-bound solutions to adapt the CLIP on each task.

to access frozen experts for leveraging shareable knowledge from historical tasks, and optimize unfrozen experts to acquire specific information for the new task. As illustrated in Figure 3, during training the router can activate (a) only the untapped experts, (b) both untapped and previously learned experts, and (c) only the learned experts from previous tasks. As a result, this strategy allows experts to consolidate their knowledge collaboratively, resembling the human brain's mechanism of reinforcing and connecting new information with existing memories.

3.4. Distribution Discriminative Auto-Selector

The task-specific nature of the routers in our MoE-Adapters necessitates manual task identity to activate the appropriate router. Such manual intervention is not aligned with the automated and practical nature of Task Incremental Learning (TIL) and Class Incremental Learning (CIL), and restricts the inherent zero-shot generalization capability of the CLIP model. To address this limitation, we develop the Distribution Discriminative Auto-Selector (DDAS), which automatically selects the proper router with the task context inferred by analyzing the variation in the distribution of input.

DDAS extends the incremental MoE framework by introducing a series of task-specific autoencoders, $\{\mathcal{F}_A^t\}_{t=1}^T$, which are trained to independently capture the distribution characteristics for the tasks $\{\mathcal{T}^t\}_{t=1}^T$. The loss function employed for training the autoencoders is the Mean Squared Error (MSE), defined as:

$$d^t = ||\boldsymbol{f}_i^t - \boldsymbol{f}_o^t||^2, \tag{3}$$

where $\boldsymbol{f}_i^t$ is the intermediate feature extracted from the input image, and $\boldsymbol{f}_o^t = \mathcal{F}_A^t(\boldsymbol{f}_i^t)$ is the reconstructed feature representation by the autoencoder of task t. Since the autoencoder $\mathcal{F}_A^t$ is individually learned on the data of task t, the resulting reconstruction score d^t reflects the likelihood that an input image pertains to the task, with a lower score suggesting a higher probability.

Moreover, to preserve CLIP's zero-shot transfer ability during continual learning, we include an additional autoencoder, $\mathcal{F}_A^0$, trained on a reference dataset to identify out-of-distribution data. Upon completion of the learning process, DDAS computes a set of distribution scores $\{d^t\}_{t=1}^T$ for each input image. Should all scores surpass a specific threshold, $Thres$, the system classifies the input as "unseen

	Method	Aircraft [50]	Caltech101 [21]	CIFAR100 [39]	DTD [9]	EuroSAT [25]	Flowers [55]	Food [4]	MNIST [13]	OxfordPet [59]	Cars [38]	SUN397 [70]	*Average*
CLIP	Zero-shot	24.3	88.4	68.2	44.6	54.9	71.0	88.5	59.4	89.0	64.7	65.2	65.3
	5-shot Full Fine-tune	30.6	93.5	76.8	65.1	91.7	92.9	83.3	96.6	84.9	65.4	71.3	77.5
	5-shot Fine-tune Adapter	29.7	90.0	75.3	63.9	81.1	94.2	87.8	90.4	89.0	68.2	72.5	76.6
Transfer	Continual-FT		72.8	53.0	36.4	35.4	43.3	68.4	47.4	72.6	30.0	52.7	51.2
	LwF [43]		72.1	49.2	35.9	44.5	41.1	66.6	50.5	69.0	19.0	51.7	50.0
	LwF-VR [15]		82.2	62.5	40.1	40.1	56.3	80.0	60.9	77.6	40.5	60.8	60.1
	WiSE-FT [68]		77.6	60.0	41.3	39.4	53.0	76.6	58.1	75.5	37.3	58.2	57.7
	ZSCL [79]		84.0	68.1	**44.8**	46.8	63.6	84.9	61.4	81.4	55.5	62.2	65.3
	Ours		**87.9**	**68.2**	44.1	**48.1**	**64.7**	**88.8**	**69.0**	**89.1**	**64.5**	**65.1**	**68.9(+3.6)**
Average	Continual-FT	28.1	86.4	59.1	52.8	55.8	62.0	70.2	64.7	75.5	35.0	54.0	58.5
	LwF [43]	23.5	77.4	43.5	41.7	43.5	52.2	54.6	63.4	68.0	21.3	52.6	49.2
	LwF-VR [15]	24.9	89.1	64.2	53.4	54.3	70.8	79.2	66.5	79.2	44.1	61.6	62.5
	WiSE-FT [68]	**32.0**	87.7	61.0	55.8	68.1	69.3	76.8	71.5	77.6	42.0	59.3	63.7
	ZSCL [79]	28.2	88.6	66.5	53.5	56.3	73.4	83.1	56.4	82.4	57.5	62.9	64.4
	Ours	30.0	**89.6**	**73.9**	**58.7**	**69.3**	**79.3**	**88.1**	**76.5**	**89.1**	**65.3**	**65.8**	**71.4(+7.0)**
Last	Continual-FT	27.8	86.9	60.1	58.4	56.6	75.7	73.8	93.1	82.5	57.0	66.8	67.1
	LwF [43]	22.1	58.2	17.9	32.1	28.1	66.7	46.0	84.3	64.1	31.5	60.1	46.5
	LwF-VR [15]	22.9	**89.8**	59.3	57.1	57.6	79.2	78.3	77.7	83.6	60.1	69.8	66.9
	WiSE-FT [68]	**30.8**	88.9	59.6	60.3	80.9	81.7	77.1	**94.9**	83.2	62.8	70.0	71.9
	ZSCL [79]	26.8	88.5	63.7	55.7	60.2	82.1	82.6	58.6	85.9	66.7	70.4	67.4
	Ours	30.1	89.3	**74.9**	**64.0**	**82.3**	**89.4**	**87.1**	89.0	**89.1**	**69.5**	**72.5**	**76.1(+4.2)**

Table 2. Comparison with state-of-the-art methods on few-shot MTIL benchmark in terms of “Transfer”, “Average”, and “Last” scores (%). Ours converges in 500 iterations on few-shot. We label the best and second methods with **bold** and underline styles. The top block indicates the upper-bound solutions to adapt the CLIP on each task.

data” and redirects it to the frozen CLIP for zero-shot transfer. Conversely, inputs below this threshold are routed to the corresponding router with the lowest distribution score, ensuring efficient and accurate task identification.

4. Experiments

4.1. Experimental Setting

Datasets. We evaluate our method across two tasks: Multi-domain TIL (MTIL) and CIL. For MTIL, we follow the two-order training protocol proposed in [79]. For CIL, we follow [19] to conduct experiments on CIFAR100 [19] and TinyImageNet [71]. The 100 classes of CIFAR100 are divided into $\{10, 20, 50\}$ subsets, and the 100 classes from TinyImageNet are divided into $\{5, 10, 20\}$ subsets to evaluate class distribution adaptability.

Metrics. To evaluate our method on the MTIL, we utilize metrics proposed by [79], namely “Transfer”, “Average”, and “Last”. The “Transfer” metric assesses the model’s zero-shot transfer capability on unseen data. “Last” evaluates the model’s memorization ability on historical knowledge. “Average” is a composite metric measuring the mean performance across “Transfer” and “Last”. In CIL, following [19], we calculate the average accuracy over all subsets (“Average”) and specifically for the last subset (“Last”).

Implementation Details. As in [79], we use the CLIP model with ViT-B/16 [17] as our backbone for all the experiments. We adopt LoRA [28] as experts and set the total number $N_E = 22$. The router is a single MLP that mixes the experts with *top*-2 gating scores. In DDAS, the reference data is TinyImageNet [71], and the threshold is 0.065 and 0.06 for full-shot and few-shot. The autoencoder is built upon a pretrained AlexNet [40] with MLP and a non-linear layer. We use AdamW [49] optimizer and a label smoothing [53] technique for a better result. For TIL, we train 1k iterations on full-shot and 500 iterations on few-shot for each task. For DDAS, we train 1k and 300 iterations for reference datasets and incremental tasks, respectively. Except for the reference dataset, the MoE-Adapters and DDAS are jointly trained during continual learning.

4.2. Comparison with State-of-the-art Methods

Multi-domain Task Incremental Learning. Table 1 showcases a comparison between our proposed method and alternative approaches in the MTIL task. Our approach utilizes the predefined Order-I from [79], where datasets are trained and tested sequentially in a left-to-right order, as displayed in Table 1. Additional results for Order-II are provided in the supplementary material. The uppermost section of Table 1 displays the outcomes of applying CLIP independently on each task through zero-shot inference, full parameter fine-tuning, and parameter-efficient fine-tuning. The “zero-shot” represents the optimal results of CLIP’s zero-shot transfer on each task, while the other two rows

Method	10 step		20 step		50 step	
	Avg.	Last	Avg.	Last	Avg.	Last
UCIR [26]	58.66	43.39	58.17	40.63	56.86	37.09
Bic[69]	68.80	53.54	66.48	47.02	62.09	41.04
PODNet[18]	58.03	41.05	53.97	35.02	51.19	32.99
DER [71]	74.64	64.35	73.98	62.55	72.05	59.76
DyTox+[19]	74.10	62.34	71.62	57.43	68.90	51.09
DNE [29]	74.86	70.04	-	-	-	-
CLIP Zero-shot	74.47	65.92	75.20	65.74	75.67	65.94
Fine-tune	65.46	53.23	59.69	43.13	39.23	18.89
LwF [43]	65.86	48.04	60.64	40.56	47.69	32.90
iCaRL [62]	79.35	70.97	73.32	64.55	71.28	59.07
LwF-VR [15]	78.81	70.75	74.54	63.54	71.02	59.45
ZSCL [79]	82.15	73.65	80.39	69.58	79.92	67.36
Ours	**85.21**	**77.52**	**83.72**	**76.20**	**83.60**	**75.24**

Table 3. Comparison of different methods on CIFAR100 in class-incremental setting. We label the best and second-best methods with **bold** and underline styles.

Method	5 step		10 step		20 step	
	Avg.	Last	Avg.	Last	Avg.	Last
EWC [37]	19.01	6.00	15.82	3.79	12.35	4.73
EEIL [6]	47.17	35.12	45.03	34.64	40.41	29.72
UCIR [26]	50.30	39.42	48.58	37.29	42.84	30.85
MUC [47]	32.23	19.20	26.67	15.33	21.89	10.32
PASS [81]	49.54	41.64	47.19	39.27	42.01	32.93
DyTox [19]	55.58	47.23	52.26	42.79	46.18	36.21
CLIP Zero-shot	69.62	65.30	69.55	65.59	69.49	65.30
Fine-tune	61.54	46.66	57.05	41.54	54.62	44.55
LwF [43]	60.97	48.77	57.60	44.00	54.79	42.26
iCaRL [62]	77.02	70.39	73.48	65.97	69.65	64.68
LwF-VR [15]	77.56	70.89	74.12	67.05	69.94	63.89
ZSCL [79]	80.27	73.57	78.61	71.62	77.18	68.30
Ours	**81.12**	**76.81**	**80.23**	**76.35**	**79.96**	**75.77**

Table 4. Comparison of different methods on TinyImageNet dataset in class-incremental settings with 100 base classes. We label the best and second methods with **bold** and underline styles.

indicate the highest possible outcomes achieved by fine-tuning CLIP in each respective task. Our proposed method, labeled as "Ours", outperforms the second-best approach on most tasks, resulting in an overall improvement of by 0.8%, 1.3%, and 1.4% in terms of "Transfer", "Average" and "Last", respectively. Additionally, by increasing the training iterations from 1k to 3k, our method (labelled as "Ours†") achieves a further improvement of 1.8% and 3.8% on "Average" and "Last" while dropping 0.7% on "Transfer" in comparison to ZSCL [79]. Furthermore, our model achieves less degradation than ZSCL when compared with the upper bound methods, demonstrating favorable performance in anti-forgetting and zero-shot transfer.

Few-shot Multi-Domain Task Incremental Learning. we also ran experiments on few-shot MTIL, limiting the CLIP model to access only a few samples per task. In a 5-shot setting, the comparison results are shown in Table 2 using the same metrics as Table 1. Our method outperforms most state-of-the-art approaches on most datasets, surpassing the second-best method by 3.6%, 7.0%, and 4.2% in terms of "Transfer," "Average," and "Last". These results demonstrate the effectiveness of the proposed incremental MoE-Adapters in addressing the forgetting issue in long-term continual learning, even with limited samples. Furthermore, our proposed DDAS effectively learns data distribution discrimination with fewer training samples.

Method	Train Params ↓	GPU ↓	Times ↓
LWF [43]	149.6M	32172MiB	1.54s/it
LWF-VR [15]	149.6M	32236MiB	1.51s/it
ZSCL [79]	149.6M	26290MiB	3.94s/it
MoE-Adapters	51.1M	19898MiB	1.37s/it
DDAS	8.7M	2461MiB	0.21s/it
Ours	59.8M	22358MiB	1.58s/it
Δ	**-60.03%**	**-14.95%**	**-59.90%**

Table 5. Comparison of computational cost during training between our method and others in terms of training parameters, GPU burdens and training times of each iteration. And the Δ is the improvement relative to the SOTA ZSCL [79].

Class Incremental Learning. We conduct experiments on class incremental learning to verify our method's performance on single-domain CL. Unlike MTIL, the task id of the input image is unknown in CL. To this end, our MoE-Adapters use only one router with two experts to adapt to all the subsets. The comparison results between our method and state-of-the-art approaches on CIFAR100 and TinyImageNet are shown in Table 3 and 4, respectively. As we can see, the proposed method consistently outperforms the other competitors, including dynamic expansion and CLIP-based approaches, demonstrating the effectiveness and scalability of our MoE-Adapters in addressing single-domain CL.

Computational Cost. The experiments above have demonstrated the promising performance of our method in both MTIL and CIL. We further compare the computational cost of our method with others to prove its parameter and time efficiency during training. Table 5 shows that our method is superior to the SOTA method ZSCL, with a reduction of approximately 60%, 15%, and 60% in terms of training parameter (M), GPU burdens (MiB), and iteration time, respectively. Additionally, we analyze the efficiency of our two proposed components, MoE-Adapters and DDAS, demonstrating that they effectively enhance continual learning for CLIP while reducing significant computation burdens during training.

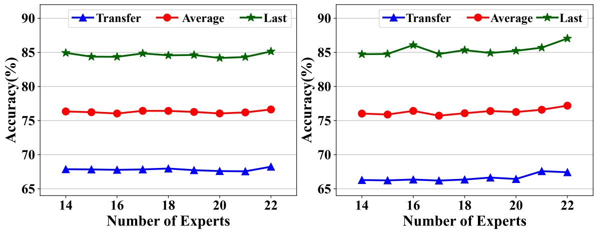
4.3. Ablation Study

In this section, we mainly analyze the efficacy of the proposed incremental MoE-Adapters and DDAS. All the experiments are conducted in MTIL setting. More analysis can be found in the supplementary material.

Analysis of MoE-Adapters. We conduct detailed ablation studies of different settings on MoE-Adapters, as shown in Table 6. Compared with the zero-shot CLIP and the fine-

Method	Transfer	Δ	Avg.	Δ	Last	Δ
CLIP Zero-shot	69.4	+0.5	65.3	-11.4	65.3	-19.7
+Adapter	45.0	-23.9	57.0	-19.7	71.5	-13.5
+2E/1R	45.1	-23.8	56.3	-20.4	71.1	-13.9
+2E/11R	68.1	-0.8	72.6	-4.1	77.9	-7.1
+22E/1R	44.1	-24.8	56.0	-20.7	66.2	-18.8
+22E/11R w/o F	68.6	-0.3	75.1	-1.6	82.0	-3.0
Ours	68.9	0.0	76.7	0.0	85.0	0.0

Table 6. Ablation studies on incremental MoE-Adapters. "mE/nR" indicates MoE with m experts and n routers, respectively. "F" represents the incremental activate-freeze strategy.

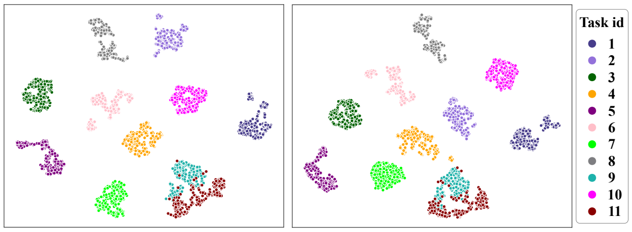


(a) 1k iterations **(b) 3k iterations**

Figure 4. Analysis of expert's number in different training iterations. The results can be referred to "Ours" and "Ours†" in Table 1.

tuned version with one adapter, our MoE-Adapters effectively mitigate the "catastrophic forgetting" issue and retain the zero-shot transfer ability on the unseen date. In the proposed MoE-Adapters, we use T task-specific routers to adaptively activate the $Topk$ experts from the predefined expert pool. Table 6 illustrates several different experts and routers combinations. As we can see, compared with using more experts, the task-specific routers contribute more to improve anti-forgetting and zero-shot transfer abilities. In the training of MoE, we propose an incremental activate-freeze strategy, enabling the collaboration of previously learned experts and inactivated ones for more accurate prediction. The comparison between "Ours" and "+22E/11R w/o F" demonstrates the effectiveness of this strategy.

Analysis of Expert Number. Figure 4 presents the ablation study on the number of experts used in MoE-Adapters. The experiments are conducted on the models trained by 1k / 3k iterations with T task-specific routers. The smoothness of the curves in the figure indicates our method's robustness for changing the number of experts. We can observe that the three metrics remain relatively stable as the number of experts changes. As shown in Figure 4 (a) and (b), it is more stable in the 1k iteration setting than in 3k iterations. This phenomenon arises due to the escalating number of iterations, leading to an increase in the overall frequency of expert selection. When the number of experts is small, our activate-freeze strategy may activate more untapped experts and train them a few times, leading to the router mistakenly activating incompletely trained experts during the inference phase. The fluctuation is within an acceptable range and does not significantly affect the final performance.



(a) Full-shot **(b) Few-shot**

Figure 5. t-SNE on DDAS's output of each task on full-shot and few-shot MTIL. The corresponding task names from $id = 1 - 11$ are matches with the datasets listed from left to right in Table 1.

Analysis of DDAS. We use DDAS to automatically distinguish the input images from seen or unseen data by learning the variations in data distribution with task-specific autoencoders. To verify the effectiveness of DDAS, we analyze the distribution discrimination in feature space, whose results are illustrated in Figure 5. We employ the reconstructed features $\boldsymbol{f}_o^t$ and plot the figure when the continual learning is finished. As we can see, the proposed DDAS is effective at learning the discriminative distribution of each learned task in full-shot and few-shot MTIL. As shown in Figure 5, the feature distributions of some samples from Task 9 overlap with that of Task 11. It is because these samples are misclassified by DDAS as out-of-distribution data and perform feature extraction with reference autoencoder. Although the inevitable misclassifications occur, our method still outperforms the state-of-the-art approaches in various metrics.

5. Discussion

We propose a parameter-efficient training framework to boost the continual learning of vision language models. We employ MoE-Adapters to help the CLIP model to adapt efficiently and generalize well on all tasks. Moreover, we introduce a Distribution Discriminative Auto-Selector (DDAS) to assign inference data automatically to either MoE-Adapters or the frozen CLIP. Extensive experimental results in various settings demonstrate the superiority of our method over previous arts in terms of classification accuracy and training efficiency.

One limitation of our framework is that the proposed DDAS requires a predefined threshold to determine downstream branches for all tasks. With the growth of task numbers, adapting all tasks with a single threshold would bring errors. Besides, incorporating the learned knowledge to improve the zero-shot transfer ability of the original CLIP is a future research direction.

Acknowledgement. This work was supported by National Key R&D Program of China under Grant 2021YFA0715202, National Natural Science Foundation of China under Grant 62206039 and 62293544, the Fundamental Research Funds for the Central Universities, and CCF-Baidu Open Fund (no. CCF-Baidu202314).

References

[1] Rahaf Aljundi, Punarjay Chakravarty, and Tinne Tuytelaars. Expert gate: Lifelong learning with a network of experts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3366–3375, 2017. 1, 2, 3

[2] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European conference on computer vision (ECCV)*, pages 139–154, 2018. 2

[3] Eden Belouadah and Adrian Popescu. Il2m: Class incremental learning with dual memory. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 583–592, 2019. 2

[4] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Proceedings of the European conference on computer vision (ECCV)*, pages 446–461, 2014. 5, 6

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 3

[6] Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 233–248, 2018. 7

[7] Wuyang Chen, Yanqi Zhou, Nan Du, Yanping Huang, James Laudon, Zhifeng Chen, and Claire Cui. Lifelong language pretraining with distribution-specialized experts. In *International Conference on Machine Learning*, pages 5383–5395. PMLR, 2023. 3

[8] Zitian Chen, Yikang Shen, Mingyu Ding, Zhenfang Chen, Hengshuang Zhao, Erik G Learned-Miller, and Chuang Gan. Mod-squad: Designing mixtures of experts as modular multi-task learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11828–11837, 2023. 4

[9] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 5, 6

[10] Erik Daxberger, Floris Weers, Bowen Zhang, Tom Gunter, Ruoming Pang, Marcin Eichner, Michael Emmersberger, Yinfei Yang, Alexander Toshev, and Xianzhi Du. Mobile v-moes: Scaling down vision transformers via sparse mixture-of-experts. *arXiv preprint arXiv:2309.04354*, 2023. 3, 4

[11] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3366–3385, 2021. 2

[12] Thomas De Min, Massimiliano Mancini, Karteek Alahari, Xavier Alameda-Pineda, and Elisa Ricci. On the effectiveness of layernorm tuning for continual learning in vision transformers. *arXiv preprint arXiv:2308.09610*, 2023. 2

[13] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012. 5, 6

[14] Prithviraj Dhar, Rajat Vikram Singh, Kuan-Chuan Peng, Ziyan Wu, and Rama Chellappa. Learning without memorizing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5138–5146, 2019. 2

[15] Yuxuan Ding, Lingqiao Liu, Chunna Tian, Jingyuan Yang, and Haoxuan Ding. Don't stop learning: Towards continual learning for the clip model. *arXiv preprint arXiv:2207.09248*, 2022. 5, 6, 7

[16] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655. PMLR, 2014. 1

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6

[18] Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 86–102, 2020. 2, 7

[19] 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, 6, 7

[20] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research*, 23(1):5232–5270, 2022. 3

[21] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. 5, 6

[22] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, pages 1–15, 2023. 1, 3, 4

[23] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 1

[24] Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013. 1

[25] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning

benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 5, 6

[26] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 831–839, 2019. 2, 7

[27] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019. 3

[28] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 1, 3, 4, 6

[29] Zhiyuan Hu, Yunsheng Li, Jiancheng Lyu, Dashan Gao, and Nuno Vasconcelos. Dense network expansion for class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11858–11867, 2023. 2, 7

[30] Zi-Yuan Hu, Yanyang Li, Michael R Lyu, and Liwei Wang. Vl-pet: Vision-and-language parameter-efficient tuning via granularity control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3010–3020, 2023. 1

[31] David Isele and Akansel Cosgun. Selective experience replay for lifelong learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 2

[32] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991. 2, 3

[33] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 3

[34] Quentin Jodelet, Xin Liu, Yin Jun Phua, and Tsuyoshi Murata. Class-incremental learning using diffusion model for distillation and replay. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3425–3433, 2023. 2

[35] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *European Conference on Computer Vision*, pages 105–124. Springer, 2022. 3

[36] Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. Compacter: Efficient low-rank hypercomplex adapter layers. *Advances in Neural Information Processing Systems*, 34:1022–1035, 2021. 3

[37] 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, 7

[38] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 5, 6

[39] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5, 6

[40] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 6

[41] Frantzeska Lavda, Jason Ramapuram, Magda Gregorova, and Alexandros Kalousis. Continual classification learning using generative models. *arXiv preprint arXiv:1810.10612*, 2018. 2

[42] Sang-Woo Lee, Jin-Hwa Kim, Jaehyun Jun, Jung-Woo Ha, and Byoung-Tak Zhang. Overcoming catastrophic forgetting by incremental moment matching. *Advances in neural information processing systems*, 30, 2017. 2

[43] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. 1, 2, 5, 6, 7

[44] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 1

[45] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 1

[46] Xialei Liu, Xusheng Cao, Haori Lu, Jia wen Xiao, Andrew D. Bagdanov, and Ming-Ming Cheng. Class incremental learning with pre-trained vision-language models, 2023. 2, 3, 4

[47] Yu Liu, Sarah Parisot, Gregory Slabaugh, Xu Jia, Ales Leonardis, and Tinne Tuytelaars. More classifiers, less forgetting: A generic multi-classifier paradigm for incremental learning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 699–716, 2020. 7

[48] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017. 2

[49] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[50] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 5, 6

[51] Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7765–7773, 2018. 2

[52] 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. 1

[53] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019. 6

[54] Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal contrastive learning with limoe: the language-image mixture of experts. *Advances in Neural Information Processing Systems*, 35:9564–9576, 2022. 3

[55] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008. 5, 6

[56] OpenAI OpenAI. Gpt-4 technical report. 2023. 1

[57] Guy Oren and Lior Wolf. In defense of the learning without forgetting for task incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2209–2218, 2021. 2

[58] Oleksiy Ostapenko, Timothee Lesort, Pau Rodriguez, Md Rifat Arefin, Arthur Douillard, Irina Rish, and Laurent Charlin. Continual learning with foundation models: An empirical study of latent replay. In *Conference on Lifelong Learning Agents*, pages 60–91. PMLR, 2022. 2

[59] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 5, 6

[60] Ameya Prabhu, Philip HS Torr, and Puneet K Dokania. Gdumb: A simple approach that questions our progress in continual learning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 524–540, 2020. 2

[61] 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. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2

[62] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017. 2, 5, 7

[63] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595, 2021. 3, 4

[64] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017. 2, 3, 4

[65] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30, 2017. 2

[66] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5227–5237, 2022. 1, 3

[67] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 1

[68] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7959–7971, 2022. 5, 6

[69] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 374–382, 2019. 7

[70] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 5, 6

[71] Shipeng Yan, Jiangwei Xie, and Xuming He. Der: Dynamically expandable representation for class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3014–3023, 2021. 2, 6, 7

[72] Fei Ye and Adrian G Bors. Self-evolved dynamic expansion model for task-free continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22102–22112, 2023.

[73] Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. *arXiv preprint arXiv:1708.01547*, 2017. 2

[74] Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*, 2021. 1

[75] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International conference on machine learning*, pages 3987–3995. PMLR, 2017. 2

[76] Jeffrey O Zhang, Alexander Sax, Amir Zamir, Leonidas Guibas, and Jitendra Malik. Side-tuning: a baseline for network adaptation via additive side networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 698–714, 2020. 3

[77] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021. 1, 3

[78] Xiaofeng Zhang, Yikang Shen, Zeyu Huang, Jie Zhou, Wenge Rong, and Zhang Xiong. Mixture of attention heads: Selecting attention heads per token. *arXiv preprint arXiv:2210.05144*, 2022. 4

[79] Zangwei Zheng, Mingyuan Ma, Kai Wang, Ziheng Qin, Xiangyu Yue, and Yang You. Preventing zero-shot transfer degradation in continual learning of vision-language models. *arXiv preprint arXiv:2303.06628*, 2023. 1, 2, 5, 6, 7

[80] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 3

[81] Fei Zhu, Xu-Yao Zhang, Chuang Wang, Fei Yin, and Cheng-Lin Liu. Prototype augmentation and self-supervision for incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5871–5880, 2021. 7