Continual Learners are Incremental Model Generalizers

Jaehong Yoon [1 2] Sung Ju Hwang [1 3] Yue Cao [4]

Abstract

Motivated by the efficiency and rapid convergence of pre-trained models for solving downstream tasks, this paper extensively studies the impact of Continual Learning (CL) models as pre-trainers. In both supervised and unsupervised CL, we find that the transfer quality of the representation often increases gradually without noticeable degradation in fine-tuning performance. This is because CL models can learn improved task-general features when easily forgetting task-specific knowledge. Based on this observation, we suggest a new unsupervised CL framework with masked modeling, which aims to capture fluent task-generic representation during training. Furthermore, we propose a new fine-tuning scheme, *GLobal Attention Discretization (GLAD)*, that preserves rich task-generic representation during solving downstream tasks. The model fine-tuned with GLAD achieves competitive performance and can also be used as a good pre-trained model itself. We believe this paper breaks the barriers between pre-training and fine-tuning steps and leads to a sustainable learning framework in which the continual learner incrementally improves model generalization, yielding better transfer to unseen tasks.

1. Introduction

Unsupervised Representation Learning (URL) (Radford et al., 2015; Gidaris et al., 2018; Grill et al., 2020; Xie et al., 2021) is a pertinent branch of machine learning in which a model exploits data without human-generated signals to extract the generic representations. Although the standard URL scenario assumes that we have a complete unlabeled dataset before training, this setting is often unre-

alistic in the real world; as the world persistently changes, the model should cope with non-stationary data throughout its lifespan. It carries the lifelong learnability of the representation model. As motivated by the Continual Learning (CL) field (Thrun, 1995; Silver & Mercer, 2002; Kumar & Daume III, 2012; Li & Hoiem, 2016), Unsupervised Continual Learning (UCL) (Rao et al., 2019; Madaan et al., 2022; Fini et al., 2022) has recently been explored to address the limitations of the conventional representation learning setup and provides comprehensive analyses regarding the quality of learned representations along with their forgetting.

However, the recently proposed UCL frameworks have clear limitations in their interpretation of model transfer to downstream tasks. Suppose $r_{i,j}$ be the performance of a pre-defined supervised metric for task j using a sequentially pre-trained backbone model from the first to i^{th} task. They train on T sequential tasks $\{\mathcal{T}_t\}_{t=1}^T$ without labels, and measure the effectiveness of their representation model leveraging two supervised metrics: 1) the averaged performance gap measured *immediately after the task is learned* and *after all tasks are learned* $\sum_{t=1}^{T-1} r_t = r_{T,t} - r_{t,t}$ (backward transfer) (Lopez-Paz & Ranzato, 2017) and 2) the averaged performance of all tasks $\sum_{t=1}^T r_t = r_{T,t}$. Though maximizing the transferability of the learned representations on the target problem is essential for general-purpose models, prior UCL works have confined their validations and analyses to **linear evaluations** as updating the linear classifier with keeping fixed representation model backbones. This can be suitable for measuring direct differences in model drift (i.e., catastrophic forgetting) of continual learners. Yet, it cannot disclose the change in knowledge transferability during task sequential training, which is crucial in utilizing the pre-trained model in practice.

Beyond the limited understanding of knowledge transfer from prior works, we provide comprehensive transferability analyses with varying evaluation setups via supervised and unsupervised CL methods to explore their potential as a pre-trainer. In Figure 1, we perform a simple experiment to investigate the potential of incremental model generalization via a continual learning setup. A model sequentially learns the nine tasks from ImageNet1K-split (containing ten tasks in total) using supervised CL methods: *Base* (a CL model without any additional method), *SI* (Zenke et al., 2017), and *DER* (Buzzega et al., 2020). Then, we measure the change
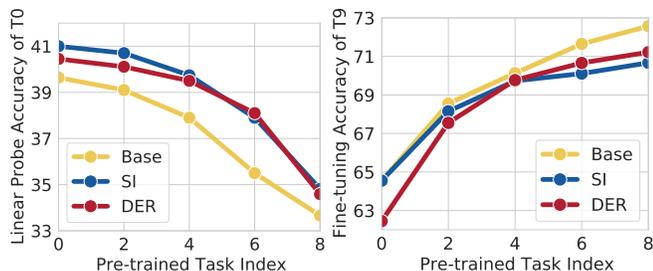

Figure 1: **Linear evaluation of the first task (T0) and fine-tuning of the unseen task (T9) during supervised CL** over a sequence of nine tasks from ImageNet1K-split (T0-T8). Sequential training on more tasks decreases the linear evaluation performance, but it increases fine-tuning performance on the unseen task.

of representation quality through linear evaluation for the first task, T0, and fine-tuning for the out-of-distribution task, T9. As shown in the increased performance in Figure 1 Right, we find that CL methods gain steady increases in their transferability to unseen tasks without the concern of performance degeneration from representational forgetting if the model trains on more tasks in a sequential manner. It is distinguishable from T0 linear evaluation results of the same methods, which suffer from performance degradation (Figure 1 Left).

We explain these phenomena because the transferability hinges upon rich task-generic features in pre-trained models (Xie et al., 2022a; Wei et al., 2022), while the model mostly loses task-specific features during CL, particularly severe when they train in a supervised/contrastive manner (please see Figures 2 and 5). Inspired by our above observations, we propose a new UCL framework based on Masked Image Modeling (MIM) (Xie et al., 2022b; He et al., 2022) that improves task-generic representation across all layers during training. Our framework retains fluent task-generic features over all layers by learning to predict masked regions of input images during unsupervised CL, conditioning other available areas. and outperforms existing supervised/unsupervised CL baselines in fine-tuning. Then we demonstrated that the suggested reconstruction-based CL framework achieves substantially higher fine-tuning performance on OOD tasks (Figure 6) than existing CL frameworks that aim to learn class-discriminative features, particularly at deeper layers.

Additionally, as we observe that continual learning models improve model generalization for downstream tasks, we raise the question of how a fine-tuned model can become a good pre-trained model itself since normal fine-tuning shifts generic features to be task-dependent. To this end, we additionally explore the potential of continual pre-training, reusing the fine-tuned model as a pre-trained model for other downstream tasks. The objective is to encourage the fine-tuning model to retain rich task-generic features during supervised fine-tuning. Leveraging our motivations for

building the MIM-based UCL framework, we suggest a new method, named **GL**obal **A**ttention **D**iscretization (**GLAD**). Our proposed method introduces a lightweight, trainable adapter in the multi-head attention module of the ViT backbone. For future reuse of fine-tuned networks, GLAD aims to keep improving global (task-generic) attention during fine-tuning through a constraint to encourage diversity of attention distance in the adaptor-free backbone while the model solves the downstream task by focusing on local (task-adaptive) features utilizing adaptor-guided attentions. We believe our observations and proposed approach lead to removing the barriers between the standard pretraining-finetuning scheme and continual learning towards improving model generalization via never-ending continual training while alleviating the threat of loss of generality of large pre-trained models during downstream task fine-tuning.

The main contributions of the paper are threefold:

- We unveil the behavior of representational transferability and forgetting of task-generic and task-specific features under multiple supervised/unsupervised continual learning frameworks at scale, with Vision Transformer backbones.

- We suggest a new learning/evaluation paradigm of the popular pretraining-finetuning scheme amalgamating to continual learning that aims to continuously increase the generalization of the pre-training backbone during the endless sequential fine-tuning phases.

- We further suggest a simple yet efficient remedy to increment task-generic feature expressiveness throughout continual pre-training, dubbed *GLAD*, which enables the model rapidly adapts to the target problem while preserving high transfer affinity to future tasks.

2. Related Work

Continual learning SI (Zenke et al., 2017) introduces an additional surrogate loss that reduces the weight shift during continual learning by maintaining the training trajectory according to the weight importance of previous tasks. DEN (Yoon et al., 2018) adaptively controls the network capacity by adding/pruning parameters when new tasks arrive. DER (Buzzega et al., 2020) stores a few training instances of previous tasks as well as their predicted logits and minimize the similarity to produce similar logit predictions on past tasks. BiC (Wu et al., 2019) adds a new layer at the top of the backbone to correct classification bias on new tasks. Similarly, WA (Zhao et al., 2020) corrects the prediction bias by rescaling the FC layer with averaged weights normalization on past tasks. DyToX (Douillard et al., 2022) adopts ViT and performs ensembled prediction with task-specific classifiers leveraging additional task-specific tokens. How-

ever, dominant research resorts to the sophisticated human annotation of inputs during training a sequence of tasks.

CURL (Rao et al., 2019) learns unsupervised representation on task sequences with a generative model adopting task-specific inference. However, the method is validated for only MNIST-scale datasets due to their limited scalability by design. Madaan et al. (2022) suggest a new unsupervised continual learning framework in a contrastive manner using Siamese structures. They demonstrate the scalability of the proposed framework through comprehensive analyses of learned representations. CaSSLe (Fini et al., 2022) utilizes a similar contrastive self-supervised framework for unsupervised continual learning, yet provides further extensive validations including diverse self-supervised learning backbones over ImageNet-100.

Very recently, Chen et al. (2023) provides intriguing discussions on forgetting and forward transfer of *FOMAML* (Finn et al., 2017) during supervised CL, primarily via quantitative evaluation with a few-shot linear probe. On the other hand, we extensively consider not only supervised CL but also siamese network-based contrastive and reconstruction-based unsupervised CL frameworks. In our work, we found that the continually learned representation behaves differently, depending on whether transferring via fine-tuning the entire model or linear evaluation with a scalable test set. And we deliver various discussion points on representation transferability through both quantitative and qualitative analyses, allowing the re-update of learned backbone weights for downstream tasks. Also, we further propose the GLAD module based on our findings, which preserves rich task-generic representation during solving downstream tasks. More discussions regarding meta-learning are provided in Appendix B.

Self-supervised learning Simsiam (Chen et al., 2020a) maximizes the similarity of input prediction upon two different augmentations using the Siamese network, learning input-invariant self-supervision. BarlowTwins (Zbontar et al., 2021) aims to remove cross-correlation across different feature vector embeddings from Siamese networks. DINO (Caron et al., 2021) distills teacher model predictions to the student by minimizing cross-entropy loss between their predictions, where the teacher model is updated through an exponential moving average from the student model. Unlike contrastive learning-based directions, Masked Image Modeling (MIM) has recently been developed inspired by masked language models for natural language understanding. SimMIM (Xie et al., 2022b) and MAE (He et al., 2022) adopt an encoder-decoder structure that zeroes out random spatial patches in each patchfied image and learns representations by predicting pixel values in masked patches. MSN (Assran et al., 2022) combines Siamese networks with masked modeling that maximize the

prediction similarity between patchfied masked inputs and the augmented target views.

3. Preliminaries

3.1. Pre-training and Fine-tuning

Given a neural network $f(\cdot; \boldsymbol{w})$ parameterized by weights $\boldsymbol{w}$, recent works have addressed the broad machine learning problems described to $\mathcal{D}_{target}$ by optimizing learnable weights with respect to complex objective functions. Beyond statistical initialization of network weights (Glorot & Bengio, 2010; He et al., 2015), pre-training, where leveraging learned weights from scaled benchmark datasets (e.g., ImageNet (Deng et al., 2009)) as the initialization of $\boldsymbol{w}$, has been widely adopted to promote a rapid and stable convergence curve during training. Self-supervised learning (Chen et al., 2020a; He et al., 2020; Caron et al., 2021; Bardes et al., 2021; Xie et al., 2022a) has recently become prevalent for pre-training, demonstrating superior generalization performance compared to supervised counterparts by capturing task-agnostic input features. While multiple different frameworks are considered for self-supervised learning, we exemplify the encoder-decoder framework in this paragraph. Let h and g be an encoder and a decoder parameterized by $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, respectively, the objective function is to minimize self-supervised loss given input data $\boldsymbol{d}$ without supervision:

$$\boldsymbol{\theta}^*, \ \boldsymbol{\phi}^* = \arg\min_{\boldsymbol{\theta}, \boldsymbol{\phi}} \ell\left(g_{\boldsymbol{\phi}} \circ h_{\boldsymbol{\theta}}\left(\boldsymbol{d}\right)\right), \quad (1)$$

where $\circ$ indicates function composition. The loss function is often designed in several formulations based on similarity, identity correlation, and contrastive loss. After the pre-training phase, the encoder transfers learned features to backbone neural networks for fine-tuning, $\boldsymbol{w} \leftarrow \boldsymbol{\theta}^*$.

3.2. Continual Learning Paradigms

Supervised Continual Learning (SCL) (Mallya & Lazebnik, 2018; Riemer et al., 2019; Aljundi et al., 2019; Chaudhry et al., 2019; 2020; Chrysakis & Moens, 2020; Titsias et al., 2020; Shen et al., 2020; Douillard et al., 2022; Yoon et al., 2020; 2022) is about a sustainable adaptation to unlimited task sequences while maintaining proficiency on previous tasks. Let us consider an intuitive example with the image-based problem: suppose $\mathcal{T} = \{\mathcal{T}_1, ..., \mathcal{T}_T\}$ be a sequence of T tasks, where the dataset $\mathcal{D}_t$ for the t-th task consists of n_t training instances $\mathcal{X}_t \in \mathbb{R}^{n_t \times C \times H \times W}$ and corresponding labels $\mathcal{Y}_t \in \mathbb{R}^{n_t}$. That is, C, H, and W denotes a channel, height, and width of images, respectively. A continual learner $f_{\boldsymbol{w}}$, parameterized by a set of weights $\boldsymbol{w}$, aims to predict classes by minimizing the optimization problem: $minimize_{\boldsymbol{w}} \sum_{t=1}^{T} L_{CE}\left(f\left(\mathcal{X}_t; \boldsymbol{w}\right), \mathcal{Y}_t\right)$, where L_{CE} is a cross-entropy loss. Yet, we assume that $f_{\boldsymbol{w}}$ can access each task in a specific timestep that loses the authorization to
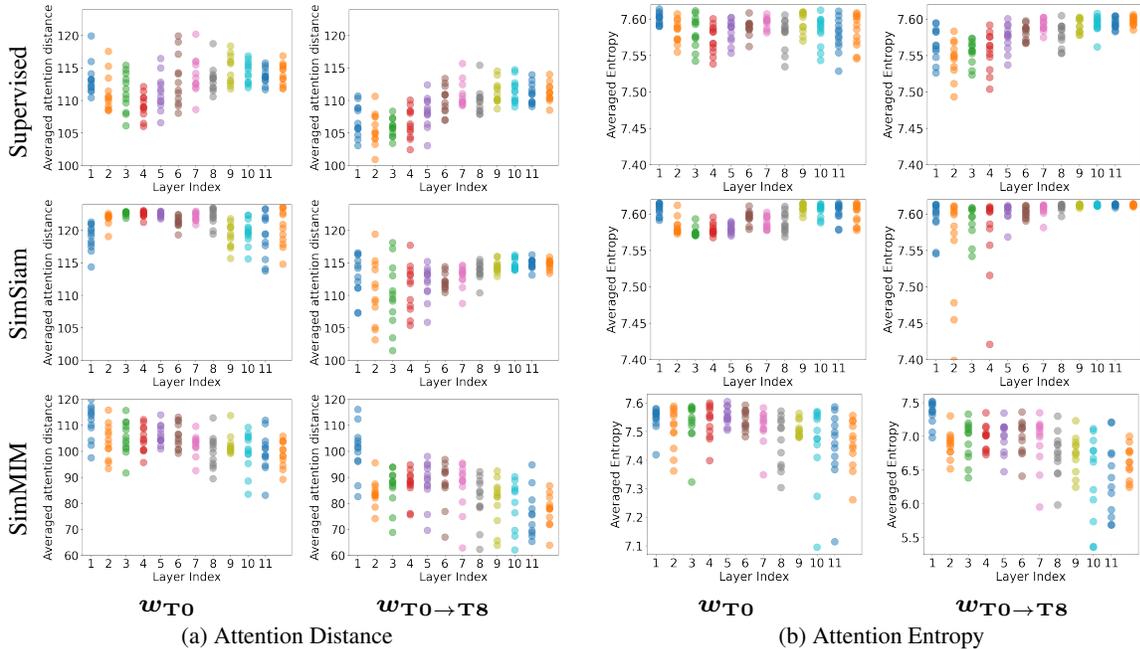
Figure 2: **(a): Visualization of the attention distance** of out-of-distribution task (T9) with respect to three continual learning frameworks right after the completion of the first (w_{T0}) and last task ($w_{T0 \to T8}$). **(b): Visualization of the entropy of each attention head's distribution** of unseen task T9. We use a ViT (Dosovitskiy et al., 2020) backbone for visualization.

revisit previous tasks' data instances when the next task arrives. That is, the model solves the following non-stationary problem at task t throughout task sequential training:

$$w^* = arg \min_{w} \ L_{CE} \left(f \left(\mathcal{X}_t; w \right), \mathcal{Y}_t \right)$$
$$\approx arg \min_{w} \ \sum_{i=1}^{t} L_{CE} \left(f \left(\mathcal{X}_i; w \right), \mathcal{Y}_i \right). \quad (2)$$

Obtained models directly evaluate the performance of each task, categorizing several incremental learning setups according to the accessibility to task oracle during inference.

Unsupervised Continual Learning (UCL) is formulated in representation learning frameworks on a sequence of unlabeled tasks $\{\mathcal{X}_t\}_{t=1}^{T}$, often referred to as continual self-supervised learning. A learner f_w aims to find the best solution that learns the informative representation of multiple tasks sequentially. At each timestep t, the model resorts to the accessible dataset $\mathcal{X}_t$ without any human-annotated labels to solve the problem:

$$w^*, \ w_{ext} = arg \min_{w} \ L \left(f \left(\mathcal{X}_t; w, w_{ext} \right) \right), \quad (3)$$

where L is an arbitrary loss function for representation learning (e.g., self-supervised losses (Chen et al., 2020b; Grill et al., 2020; Zbontar et al., 2021)) and w_{ext} is an (optional) extra weights for additional structures not included in backbone weights w, such as a decoder, a projection layer, and a predictor. Since a direct comparison of the quality of

representation models is intractable, recent representation learning literature validates obtained representation models by probing generic transferability on multiple downstream tasks. In a similar vein, prior UCL works (Madaan et al., 2022; Fini et al., 2022) adopt supervised prediction tools like the KNN classifier and linear evaluation while keeping the learned backbone fixed. However, we argue that such evaluation paradigms cannot appropriately measure the transferability of representation on unseen tasks.

4. Continual Learning for Incremental Model Generalization

4.1. The Role of Global and Local Attention during Continual Learning

Prior continual learning literature has demonstrated that a model with a standard CL setting suffers from forgetting due to loss of local features and attention to past tasks. But, we argue that they have barely discussed the generalization of continually learned representation, which can be a great source for deploying an improvable foundation model by fine-tuning an unlimited number of tasks. We throw a question mark at this point:

> "So, is the model generalization getting worse as it goes through training sequential tasks?"

Surprisingly, we found that this is not the case as the generalization (a.k.a., transfer quality) is consistently getting
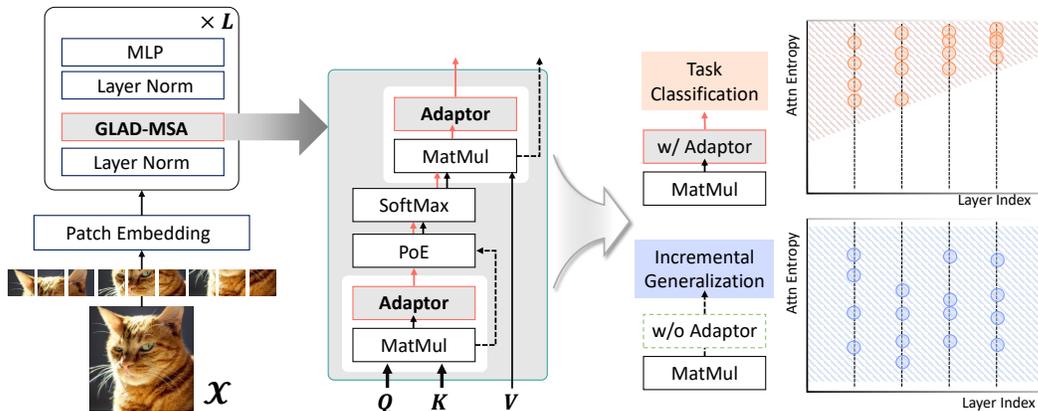
Figure 3: **Illustration of the proposed GLobal Attention Discretization (GLAD) for Continual Pre-training.** Our GLAD introduces a new Multi-head Self-attention operation named GLAD-MSA with a parametric adaptor. The pre-trained model with adaptors fine-tunes given problems with a constraint that encourages divergence of attention entropy in each layer, leading to incremental positive transfer of backbone parameters for future tasks.

improved during CL as shown in Figure 1 Right. To explicate behaviors of multi-head attention using transformer backbones, similar to the experiment in Figure 1, we train supervised and unsupervised CL frameworks on the ImageNet-1K split dataset without any specific CL methods. After that, we investigate changes in regional inductive bias by measuring the average distance of attention heads (defined by Equation (8)) in each layer, visualized as different points in Figure 2 (a) top and middle row. Supervised and contrastive self-supervised (*SimSiam*) continual learning models focus more on strong locality inductive bias at lower layers (decreased attention distance) and global attention at higher layers (increased attention distance) during continual learning as these frameworks are innately designed to cluster/classify input features in deeper layers. However, focusing on global attention to capture task-specific features is undesirable for transferring the representations to out-of-distribution tasks. Next, we visualize the entropy conditioned solely on the distribution of each attention head in Figure 2 (b) by computing $-\sum_i \boldsymbol{a}_i \log(\boldsymbol{a}_i)$ for each attention head $\boldsymbol{a}$. Supervised and contrastive unsupervised frameworks broadly focus on most attention heads during continual learning at deeper layers. This indicates that they already substantially adapt to pre-trained tasks while losing a degree of freedom to transfer downstream tasks.

To build a new UCL framework for a better generalizable representation model across all layers, we survey Masked Image modeling (MIM) (Pathak et al., 2016; He et al., 2022) that self-trains input representation by minimizing regression loss to predict RGB pixel values in randomly zeroed patches in patchified input images. MIM enjoys locality inductive bias with diverse attention across layers, allowing better transfer ability to unseen tasks.

4.2. Continual Self-supervised Learning with

Masked Modeling

We formulate a representational learner $f_{\boldsymbol{w}}$, composed of a neural encoder $h_{\boldsymbol{\theta}}$ and a decoder $g_{\boldsymbol{\phi}}$. We build backbones using Vision Transformer variants (Dosovitskiy et al., 2020; Liu et al., 2021) due to their powerful generality and remarkable performance on high-resolution visual tasks. They are flexible to transfer the obtained representations to downstream tasks requiring various input image sizes in demands when existing UCL frameworks (Madaan et al., 2022; Rao et al., 2019) allow the fixed image size for representation learning and fine-tuning since the architectures are basically composed of multi-layer perceptrons and convolutional neural networks. At training t-th task with a training instance $\boldsymbol{x}_t \in \mathbb{R}^{C \times H \times W} \in \mathcal{X}_t$, a model segments $\boldsymbol{x}_t$ into smaller image patches where the width and height are $s < H, W$, and randomly zeros a fraction of image patches out with a fixed ratio τ. An encoder tokenizes masked patches to the embedding space and fed into multiple self-attention blocks to capture latent representation features. A decoder reconstructs encoded features to approximate the input image. The objective is to minimize the following loss function for continual representation learning ($\| \cdot \|_{\mu}$ denotes any norm, often $\mu \in \{1, 2\}$ is used). Let $K = \left\lfloor \frac{H}{s} \right\rfloor \cdot \left\lfloor \frac{W}{s} \right\rfloor$ be the number of tokens at each image, we formulate the loss function as follows:

$$\ell\left(\boldsymbol{x}_t; \boldsymbol{w}\right) = \|f\left(\boldsymbol{m} *_s \boldsymbol{x}_t; \boldsymbol{w}\right) - \boldsymbol{m} *_s \boldsymbol{x}_t\|_{\mu}$$
$$= \|g\left(h\left(\boldsymbol{m} *_s \boldsymbol{x}_t; \boldsymbol{\theta}\right); \boldsymbol{\phi}\right) - \boldsymbol{m} *_s \boldsymbol{x}_t\|_{\mu}, \quad (4)$$
$$\text{where } \boldsymbol{m} = \{0, 1\}^K \sim B\left(K, \rho\right).$$

With patch size s, $*_s$ denotes a patch-wise multiplication operation between a training instance $\boldsymbol{x}$ and a generated mask vector $\boldsymbol{m}$ drawn by the binary distribution B with sparsity ratio ρ. $B(i, \rho)$ is a i independent binary sampling with a ratio ρ to pick 1. The model updates a set of weights

that predicts masked regions of input images, conditioning other available areas. We simply adopt ℓ_1 regularization to minimize the distance between predicted patches and the targets, followed by earlier reconstruction-based works (Xie et al., 2022b; He et al., 2022), and after completing a sequential training, the obtained encoder h_θ can be utilized for many different downstream tasks. And we find that our new framework outperforms supervised and contrastive benchmarks in model transferability during continual pre-training (Please see Figure 6). Interestingly, unsupervised continual learning with the masked autoencoder framework (*SimMIM* in Figure 2 bottom row) behaves very differently from the other two frameworks. Almost all layers have a diverse focus on locality, and this tendency becomes stronger as they continue to pre-train more tasks.

4.3. Continual Pre-training via Global Attention Discretization

Our aim is to utilize the backbone of the fine-tuned model as a pre-training checkpoint for another problem in a supervised manner. Given the target task dataset $\mathcal{D}_{target} = \{\mathcal{X}, \mathcal{Y}\}$ and a classifier $\delta(\cdot; \boldsymbol{u})$ parameterized by $\boldsymbol{u}$, we formulate the objective of continual pre-training as follows:

$$\underset{\boldsymbol{w}^{(t)}, \boldsymbol{u}^{(t)}}{\text{minimize}}\, \ell\left(\delta\left(f\left(\mathcal{X}; \boldsymbol{w}^{(t)}\right); \boldsymbol{u}^{(t)}\right), \mathcal{Y}\right), \quad (5)$$

We suppose each fine-tuning step independently introduces its own classifier. The formulation is aligned with the continual learning problem described in Section 3.2, but this setting is about never-ending model generalization to achieve a consistently improved adaptation to the out-of-distribution task in the future. That is, obtained representation model for fine-tuning task t, $\widehat{\boldsymbol{w}}^{(t)}$, will be reused for future task training ($\boldsymbol{w}^{(t+1)} = \widehat{\boldsymbol{w}}^{(t)}$).

However, fine-tuning often reduces the general transferability to adapt to different tasks, demonstrating a suboptimal model generalization of supervised CL compared to our MIM-based framework (Please see Figure 6 Left). Motivated by our findings in Section 4.1 and Section 4.2, we propose a new method for continual pre-training, named *GLobal Attention Discretization (GLAD)*. Our proposed method preserves diverse degrees of averaged distance at each attention head to preserve transferable backbone weights for future problems while capturing task-adaptive features guided by GLAD modules. As illustrated in Figure 3, a multi-head self-attention operation with adaptor weights $\boldsymbol{v}$. We transform the task-generic MSA features to input-dependent by propagating adaptor $\boldsymbol{v}$. Then, the model enables solving the current task problem while constraining the backbone weights to preserve abundant locality inductive bias. Let $\boldsymbol{a}^{l,i}$ be an averaged entropy of the attention passed over adaptor operation (*dark dashed arrow*) from i-th head at layer l, the objective function of our GLAD is

as follows:

$$\begin{aligned}
&\underset{\boldsymbol{w}, \boldsymbol{v}}{\text{minimize}} \sum_{n=1}^{N} \ell\left(f\left(\boldsymbol{x}^{(n)}; \boldsymbol{w}, \boldsymbol{v}\right), \boldsymbol{y}^{(n)}\right), \\
&+ \frac{1}{L}\sum_{l=1}^{L}\left(\left(\sqrt{E\left[\left(\boldsymbol{a}^{l,i} - \overline{\boldsymbol{a}}^l\right)^2\right]} + \epsilon\right)^{-1} + \lambda\left\|\overline{\boldsymbol{a}}^l\right\|_F^2\right),
\end{aligned} \quad (6)$$

where E indicates the expectation, $\overline{\boldsymbol{a}}^l = \frac{1}{H^l}\sum_i^{H^l}\boldsymbol{a}^{l,i}$ at layer l with the number of its attention head H^l, λ is a scaling factor. ϵ is a small constant value. We jointly minimize the task loss with an additional regularizer that constrains the entropy variance of attention heads to sufficiently diverge at each layer as an average of their inverse standard deviation. We add to minimize a Frobenius norm for $\overline{\boldsymbol{a}}$ to promote abundant locality inductive bias for backbone attention weights. Note that our proposed method is robust to utilize any kind of multi-head self-attention modules, we demonstrate the efficacy in vanilla Vision Transformer (Dosovitskiy et al., 2020) and Swin Transformer (Liu et al., 2021). The learned backbone weights $\boldsymbol{w}$ excepting classifier and GLAD-adaptors can be reused for fine-tuning future tasks. We describe the overall continual pre-training procedure with GLAD in Algorithm 1.

Algorithm 1 Continal Pre-training with GLAD

---

input A sequence of tasks $\{\mathcal{D}_1, \mathcal{D}_2, \cdots\}$, backbone network f, learning rate $\eta \in \mathbb{R}^+$, hyperparameter λ, small constant ϵ, initialization $\boldsymbol{w}_{\text{init}}, \boldsymbol{v}_{\text{init}}$

1: **for all** task $\mathcal{T}_t = \mathcal{T}_1, \mathcal{T}_2, \ldots$ **do**
2:      Build a model $f_{\boldsymbol{w}, \boldsymbol{v}}(\cdot)$ with GLAD-MSA   ▷ Figure 3
3:      Initialize $\boldsymbol{w} \leftarrow \boldsymbol{w}^*$ excluding classifier **if** $f_{\boldsymbol{w}^*}$ **exists, otherwise** $\boldsymbol{w} \leftarrow \boldsymbol{w}_{\text{init}}$
4:      Initialize $\boldsymbol{v} \leftarrow diag(\boldsymbol{v}_{\text{init}}) \coloneqq (1, \ldots, 1) \in \mathbb{R}^{d_{\text{out}}}$
5:      **for** batch $\boldsymbol{x}_n, \boldsymbol{y}_n \sim \mathcal{D}_t$ **do**
6:          $\mathcal{L} = \ell\left(f\left(\boldsymbol{x}_n; \boldsymbol{w}, \boldsymbol{v}\right), \boldsymbol{y}_n\right) +$
         $\frac{1}{L}\sum_{l=1}^{L}\left(\sqrt{E\left[\left(\boldsymbol{a}^{l,i} - \overline{\boldsymbol{a}}^l\right)^2\right]} + \epsilon\right)^{-1}$ ▷ Equation (6)
7:          $\boldsymbol{w} \leftarrow \nabla_{\boldsymbol{w}}\mathcal{L}, \boldsymbol{v} \leftarrow \nabla_{\boldsymbol{v}}\mathcal{L}$
8:      **end for**
9:      $\boldsymbol{w}^* \leftarrow \boldsymbol{w}$
10: **end for**

---

5. Experiments

We conduct the experiments on various continual learning frameworks with and without supervision using ImageNet-1K Split dataset against multiple strong baselines. for unsupervised continual learning that demonstrates the effectiveness of our proposed method on fine-tuning performance on downstream tasks.

5.1. Architectures and Baselines

We use ViT (Dosovitskiy et al., 2020) and Swin Transformer (Liu et al., 2021) as backbone architectures. We follow Siamese network structure by Madaan et al. (2022) and

Continual Learners are Incremental Model Generalizers

ImageNet 1K Split (T=10)	Supervised		Contrastive (Chen & He, 2021)		Masked Model (Xie et al., 2022b)	
	Final Acc	Neg. BWT	Final Acc	Neg. BWT	Final Acc	Neg. BWT
FT 1K Pretrained	87.48 / 98.08	–	–	–	–	–
22K Pretrained	87.76 / 98.48	–	–	–	–	–
Base Model	71.90 / 90.64	6.88 / 3.70	64.38 / 86.18	7.18 / 4.56	73.18 / 91.76	13.26 / 7.46
SI (Zenke et al., 2017)	70.00 / 90.38	7.40 / 4.52	61.46 / 84.82	4.15 / 2.42	71.54 / 90.76	11.92 / 6.58
DER (Buzzega et al., 2020)	70.57 / 90.12	8.94 / 4.86	62.37 / 85.46	9.28 / 6.08	70.10 / 90.10	19.55 / 12.24
LUMP (Madaan et al., 2022)	N/A	N/A	64.01 / 86.42	6.24 / 4.32	**75.11 / 92.38**	**21.28 / 12.42**
LP 1K Pretrained	87.48 / 97.98	–	–	–	–	–
22K Pretrained	86.53 / 98.06	–	–	–	–	–
Base Model	33.66 / 62.20	-5.98 / -4.30	17.10 / 40.60	**7.64 / 13.94**	17.46 / 40.60	4.76 / 6.36
SI (Zenke et al., 2017)	**34.82 / 63.18**	-6.18 / -5.56	15.24 / 36.76	-1.42 / -2.38	14.92 / 37.80	4.82 / 8.26
DER (Buzzega et al., 2020)	34.59 / 62.29	-5.86 / -6.16	14.84 / 36.13	3.68 / 5.22	6.22 / 21.52	-0.84 / -1.04
LUMP (Madaan et al., 2022)	N/A	N/A	18.50 / 42.05	7.54 / 11.38	19.26 / 43.21	7.38 / 11.22

CIFAR-100 Split (T=10)	Supervised		Contrastive (Chen & He, 2021)		Masked Model (Xie et al., 2022b)	
	Final Acc	Neg. BWT	Final Acc	Neg. BWT	Final Acc	Neg. BWT
FT Base Model	79.3	3.8	49.3	12.6	**88.9**	18.3
SI (Zenke et al., 2017)	78.0	5.8	57.3	14.5	86.4	**20.0**
LUMP (Madaan et al., 2022)	N/A	N/A	83.3	16.5	**88.6**	18.7
LP Base Model	70.3	0.0	45.7	5.9	73.0	**11.1**
SI (Zenke et al., 2017)	69.0	0.9	49.3	3.0	68.4	9.1
LUMP (Madaan et al., 2022)	N/A	N/A	73.6	9.8	**77.1**	**11.5**

Table 1: **Fine-tuning and linear evaluation performance with their negative backward transfer** of the first task on ImageNet 1K and CIFAR-100 Split after supervised/unsupervised continual learning. We report the Top-1/Top-5 performance for all individual experiments on ImageNet and the Top-1 performance on CIFAR-100. Higher is better for both metrics, and the best results are highlighted in **bold**.
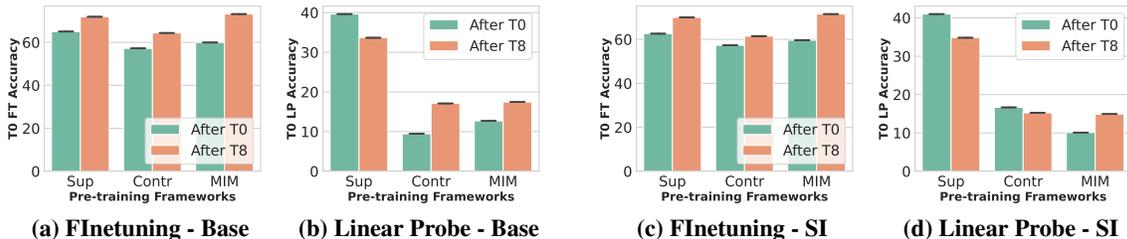

(a) FInetuning - Base **(b) Linear Probe - Base** **(c) FInetuning - SI** **(d) Linear Probe - SI**

Figure 4: **(a-b): Model transferability on ImageNet 1K Split.** We compare the fine-tuning accuracy on the OOD task (*T9*) after pre-training the first task (*After T0*) with the performance after pre-training on nine sequential tasks (*After T8*). We continually pre-train with Supervised learning (Sup), Contrastive Self-supervised learning (Contr), and Masked Image Modeling (MIM). **(c-d): Same visualization with a CL method**, *SI*, during pre-training.

implement a MIM-based continual self-supervised learning framework under SimMIM (Xie et al., 2022b) and MAE (He et al., 2022) for UCL. CURL (Rao et al., 2019) is one of the pioneer works on unsupervised continual learning literature, but it is not scalable for high-resolution visual images by design. We utilize several CL methods: SI (Zenke et al., 2017), DER (Buzzega et al., 2020), and LUMP (Madaan et al., 2022). We further describe details, including hyperparameter setups in Appendix A.

Datasets We use ImageNet-1K (Deng et al., 2009) and CIFAR-100 dataset (Krizhevsky et al., 2009) by splitting it into ten tasks where each task contains 100 and 10 classes, respectively. In Table 1, we construct a sequential dataset with nine earlier tasks and assign the last one as a downstream (validation) task with fine-tuning. Additionally, we

split CIFAR-100 into five tasks for the continual pre-training experiment.

5.2. Experimental Results

Evaluation performance during continual learning We validate the transfer quality of representation of continual learning models through the first task (T0)'s evaluation in Table 1. We measure the change in the linear evaluation and fine-tuning performance. The evaluation of T0 from full pre-trained models over the Imagenet-1K and -22K datasets obtains high validation accuracy on fine-tuning and linear evaluation as they train on entire datasets. Fine-tuning the base continual learning models, which perform a simple CL strategy without additional methods during training task sequences, achieves performance increases in T0 as they pre-trained longer task sequences, obtaining positive values

7

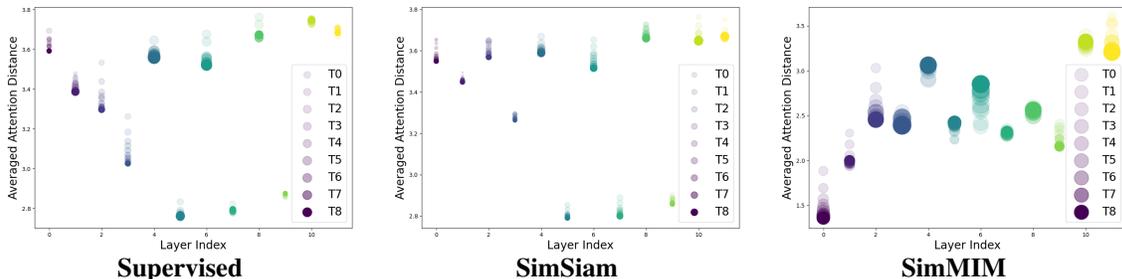**Supervised**   **SimSiam**   **SimMIM**

Figure 5: **Visualization of aggregated attention distance** on an OOD task (T9) at each layer at the end of each continual learning task phase (T0→T8). The radius of marker indicates the standard deviation over attention heads in the corresponding layer.
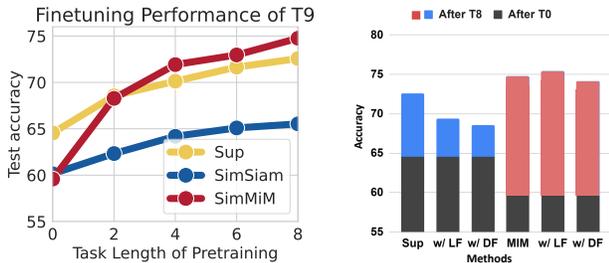

Figure 6: **Left: CL improves model generalization.** Sequential training on more tasks further increases fine-tuning performance on the unseen task (T9). **Right: The impact of freezing weights on a few lower/deeper layers during CL** after training the first task. We measure fine-tuning accuracy at the end of training the first task (T0) and last task (T8), and use supervised (*Sup*) and masked image modeling (*MIM*). *LF* and *DF* denote *freezing lower layers* and *freezing deeper layers*, respectively.

in backward transfer. In ImageNet 1K Split, the results are similar to all supervised and unsupervised continual learning frameworks, including Contrastive Self-supervised Learning (Madaan et al., 2022)- and Masked Image Modeling-based UCL (Please see Section 4.2). We also performed multiple continual learning methods, such as SI, DER, and LUMP. We found that they follow consistent tendencies according to the CL frameworks when LUMP with masked image modeling gains the highest accuracy on T0 with the strong backward transfer. On the other hand, the model degrades the linear evaluation performance in supervised continual learning, which had to do with catastrophic forgetting reported in conventional CL scenarios. SI and DER achieve increased final performance since the model mitigates the weight drift preserving the task-specific features on learned tasks. In CIFAR-100 Split, we similarly observed that fine-tuning the in-distribution task (i.e., the first task) of Base/SI/LUMP with reconstruction-based UCL framework achieves higher and positive BwT than the case of CL methods in a supervised manner, demonstrating that supervised CL suffers more severe forgetting of task-discriminative information at deeper layers, unlike reconstruction-based UCL focusing on task-generic features over all layers. Note that the linear evaluation on CIFAR-100 seems more robust to forgetting than ImageNet experiments. We expect that these

relative benefits in forgetting came from the shallower data distribution space and simpler visual features of CIFAR-100 compared to ImageNet.

Analyses for an Out-Of-Distribution (OOD) task   We denote the last task (T9) as an out-of-distribution problem, excluding it from the continual task sequence. In Figure 4, we visualize the top-1 validation accuracy on an OOD task over three continual learning frameworks w/ and w/o SI. Similar to in-distribution evaluation, the MIM-based UCL method achieves higher fine-tuning performance both on the base model and SI. The linear probe performance of Supervised CL surpasses unsupervised counterparts, and we expect that representation from supervised learning contains directly helpful features to classify the high-resolution and complex task problems even without the re-update of backbone weights. In contrast, MIM remarkably underperforms on linear evaluations due to its characteristic property; Masked Modeling focuses on capturing global attention rather than the local one, providing a better generalization ability to unseen tasks. But its linear evaluation without fine-tuning the weights is inadequate to solve the problem as they contain little task-discriminative information.

To understand how CL frameworks exhibit incremental model generalization and fine-tuning performance, we analyze the behavior of layer attention during continual learning using the Swin-T backbone. In Figure 5, we visualize the layer-by-layer changes in aggregated attention distance for T9 while the model trains ImageNet-1K-Split sequentially until the penultimate task (T0→T8). Interestingly, aggregated attention distance significantly decreases the scale and increases the diversity across attention heads. This demonstrates that the continual learner captures richer task-general (or low-level) features behaving with more local attention (i.e., lower attention distance), which retains localized information with strong local inductive bias, such as edges, patterns, and textures. In Supervised and Contrastive Continual Learning frameworks, lower layers tend to drastically change toward capturing task-generic attention, and it is also coincident with well-known observations that lower layers in neural networks are more concerned
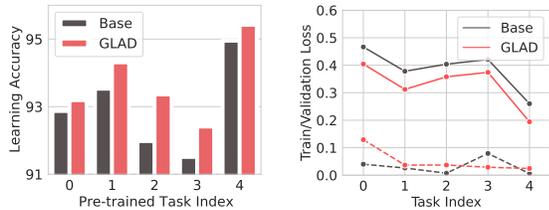
Figure 7: **Learning accuracy and loss of CIFAR-100 (T=5) at each continual pre-training step**. Right: Solid and dashed lines indicate validation and training loss, respectively.

PRE-TRAINED T0→T4	Continual learning over T5→T8			
	T5	T6	T7	T8
SUPERVISED	64.56	69.06	75.06	77.80
+ GLAD (OURS)	**65.78**	**69.84**	**76.12**	**79.04**
SIMMIM (Xie et al., 2022b)	**68.34**	72.22	77.38	80.12
+ GLAD (OURS)	68.24	**73.24**	**78.94**	**81.62**
MAE (He et al., 2022)	**42.19**	52.07	63.07	71.19
+ GLAD (OURS)	41.75	**54.74**	**68.01**	**74.30**

Table 2: **Fine-tuning accuracy of T9 at each continual pre-training step (T5 → T8)** over ImageNet1k-split. We initialized the model weights using continual (representation) learning over the earlier five sequential tasks (T0 → T4) to focus on continual pre-training performance on a few later tasks.

with low-level features. Also, Masked Image Modeling (SimMIM) results demonstrate the salient effectiveness of capturing task-generic attention compared with the other two frameworks.

Freezing the partial layer weights during continual learning We further analyze the effect of the layers for incremental generalization during continual learning in Figure 6 right. After supervised/unsupervised training of the first task (*After T0*), we freeze the two lowest or two deepest layer weights during the successive continual learning up to the final task (*After T8*). For the MIM-based UCL framework, we use MAE with a ViT-B backbone. In supervised learning, both partial gradient update policies reduce the degree of incremental generalization during continual learning. It significantly reduces the representation model's fine-tuning performance compared to the fully-trained model. However, interestingly, prohibiting the update of layer weights at both ends less affects MIM-based unsupervised continual learning. We expect that this property comes from its flexibility in learning diverse attention across all layers. For further analyses, please see Appendix C.4.

Improving model generalization via Global Attention Discretization We now validate our proposed method, *Global Attention Discretization (GLAD)*, which encourages incremental model generalization during supervised continual pre-training. As discussed earlier, supervised training tends to focus on task-specific attention at deeper layers.

That is, the model is hard to stray far from the weight space of the limited locality inductive bias, which is evident in the slower movement of averaged attention distance from supervised continual pre-training compared to the SimMIM-based (Please see Figure 5) and results in suboptimal adaptation to arriving tasks. In Table 2, we report the fine-tuning performance during continual pre-training. Note that to see the effect of MIM-based continual pre-training, we first perform continual pre-training over the earlier five tasks from ImageNet-1K Split under supervised and MIM-based unsupervised continual learning. Next, the pre-trained models fine-tune the remaining tasks in a sequential manner. We adopt SimMIM and Masked AutoEncoder (MAE) to understand general behaviors in MIM frameworks during unsupervised continual learning. Our proposed GLAD achieves significant gains in the performance of each task during sequential full-finetuning over different pre-trained initialization from supervised and unsupervised learning. In Figure 7, we visualize that our proposed GLAD consistently performs well in view of the generalization using the in-distribution task (CIFAR-100) during continual pre-training phase.

6. Conclusion

As powerful representation models have a great versatility to solve various downstream tasks, exploring incremental pre-training strategy on a number of sequential tasks can be a practical and important approach. This paper delves into how supervised and unsupervised continual learning affects model generalization from various perspectives. To our surprise, continual learning models preserve or even increment their transferability on in- and out-of-distribution tasks, increasing fine-tuning performance as pre-training more tasks. We scrutinize the behavior of representations in CL frameworks in the pre-training, including masked image modeling-based unsupervised continual learning, and find that the continual learner tends to forget class-discriminative features while progressively accumulating transferable features. Motivated by our observations, we propose a new method for continual pre-training to help backbone weights gain transferability during fine-tuning by introducing a new MSA module with parametric adaptors. We believe the exploration of continuous learnability of the representation model would contribute to developing eco-friendly and resource-efficient training regimes for broad research/industry fields.

Acknowledgements

References

Aljundi, R., Belilovsky, E., Tuytelaars, T., Charlin, L., Caccia, M., Lin, M., and Page-Caccia, L. Online continual learning with maximal interfered retrieval. In *Advances in Neural Information Processing Systems (NIPS)*, 2019.

Assran, M., Caron, M., Misra, I., Bojanowski, P., Bordes, F., Vincent, P., Joulin, A., Rabbat, M., and Ballas, N. Masked siamese networks for label-efficient learning. *arXiv preprint arXiv:2204.07141*, 2022.

Bardes, A., Ponce, J., and LeCun, Y. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Buzzega, P., Boschini, M., Porrello, A., Abati, D., and Calderara, S. Dark experience for general continual learning: a strong, simple baseline. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

Chaudhry, A., Rohrbach, M., Elhoseiny, M., Ajanthan, T., Dokania, P. K., Torr, P. H., and Ranzato, M. Continual learning with tiny episodic memories. *arXiv preprint arXiv:1902.10486*, 2019.

Chaudhry, A., Khan, N., Dokania, P. K., and Torr, P. H. Continual learning in low-rank orthogonal subspaces. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Chen, J., Nguyen, T., Gorur, D., and Chaudhry, A. Is forgetting less a good inductive bias for forward transfer? In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, 2020a.

Chen, X. and He, K. Exploring simple siamese representation learning. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Chen, X., Fan, H., Girshick, R., and He, K. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

Chrysakis, A. and Moens, M.-F. Online continual learning from imbalanced data. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Douillard, A., Ramé, A., Couairon, G., and Cord, M. Dytox: Transformers for continual learning with dynamic token expansion. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Fini, E., da Costa, V. G. T., Alameda-Pineda, X., Ricci, E., Alahari, K., and Mairal, J. Self-supervised models are continual learners. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

Gidaris, S., Singh, P., and Komodakis, N. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., Piot, B., kavukcuoglu, k., Munos, R., and Valko, M. Bootstrap your own latent - a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Hadsell, R., Rao, D., Rusu, A. A., and Pascanu, R. Embracing change: Continual learning in deep neural networks. *Trends in cognitive sciences*, 2020.

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2015.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Kumar, A. and Daume III, H. Learning task grouping and overlap in multi-task learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2012.

Li, Z. and Hoiem, D. Learning without forgetting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

Lomonaco, V., Maltoni, D., and Pellegrini, L. Rehearsal-free continual learning over small non-iid batches. In *CVPR Workshops*, 2020.

Lopez-Paz, D. and Ranzato, M. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Madaan, D., Yoon, J., Li, Y., Liu, Y., and Hwang, S. J. Rethinking the representational continuity: Towards unsupervised continual learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=9Hrka5PA7LW.

Mallya, A. and Lazebnik, S. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., and Efros, A. A. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Prabhu, A., Torr, P. H., and Dokania, P. K. Gdumb: A simple approach that questions our progress in continual learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Rao, D., Visin, F., Rusu, A., Pascanu, R., Teh, Y. W., and Hadsell, R. Continual unsupervised representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Riemer, M., Cases, I., Ajemian, R., Liu, M., Rish, I., Tu, Y., and Tesauro, G. Learning to learn without forgetting by maximizing transfer and minimizing interference. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

Shen, G., Zhang, S., Chen, X., and Deng, Z.-H. Generative feature replay with orthogonal weight modification for continual learning. *arXiv preprint arXiv:2005.03490*, 2020.

Silver, D. L. and Mercer, R. E. The task rehearsal method of life-long learning: Overcoming impoverished data. In *Conference of the Canadian Society for Computational Studies of Intelligence*. Springer, 2002.

Thrun, S. *A Lifelong Learning Perspective for Mobile Robot Control*. Elsevier, 1995.

Titsias, M. K., Schwarz, J., Matthews, A. G. d. G., Pascanu, R., and Teh, Y. W. Functional regularisation for continual learning with gaussian processes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

Wang, Z., Zhang, Z., Ebrahimi, S., Sun, R., Zhang, H., Lee, C.-Y., Ren, X., Su, G., Perot, V., Dy, J., et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

Wei, Y., Hu, H., Xie, Z., Zhang, Z., Cao, Y., Bao, J., Chen, D., and Guo, B. Contrastive learning rivals masked image modeling in fine-tuning via feature distillation. *arXiv preprint arXiv:2205.14141*, 2022.

Wu, Y., Chen, Y., Wang, L., Ye, Y., Liu, Z., Guo, Y., and Fu, Y. Large scale incremental learning. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Xie, Z., Lin, Y., Zhang, Z., Cao, Y., Lin, S., and Hu, H. Propagate yourself: Exploring pixel-level consistency for

unsupervised visual representation learning. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Xie, Z., Geng, Z., Hu, J., Zhang, Z., Hu, H., and Cao, Y. Revealing the dark secrets of masked image modeling. *arXiv preprint arXiv:2205.13543*, 2022a.

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., and Hu, H. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022b.

Yoon, J., Yang, E., Lee, J., and Hwang, S. J. Lifelong learning with dynamically expandable networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Sk7KsfW0-.

Yoon, J., Kim, S., Yang, E., and Hwang, S. J. Scalable and order-robust continual learning with additive parameter decomposition. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1gdj2EKPB.

Yoon, J., Madaan, D., Yang, E., and Hwang, S. J. Online coreset selection for rehearsal-based continual learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=f9D-5WNG4Nv.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

Zenke, F., Poole, B., and Ganguli, S. Continual learning through synaptic intelligence. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

Zhao, B., Xiao, X., Gan, G., Zhang, B., and Xia, S.-T. Maintaining discrimination and fairness in class incremental learning. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

A. Details for Problem Setups

Datasets We use ImageNet (Deng et al., 2009) dataset, containing 1000 classes of high-resolution object images with their corresponding labels. We split them into 10 tasks, where each task consists of 100 different classes. We use only 10% of training instances in each task for pre-training, and use the full set for the fine-tuning and linear probe. Accuracy is measured by the validation dataset for each task. Note that Table 2 also use 10% of the training set for pre-training earlier five sequential tasks from T0 to T4, and use the full training set for the sequential fine-tuning procedure from T5 to T8.

Architectures and baselines We follow (Madaan et al., 2022) for an unsupervised continual learning framework with contrastive self-supervised learning using Sim-Siam (Chen et al., 2020a). For masked image modeling, we follow the setting of SimMIM (Xie et al., 2022b) and MAE (He et al., 2022) using their official code repositories[1][2] where the masking ratio is 0.6 and 0.75, respectively. We use Vision Transformer (Dosovitskiy et al., 2020) (ViT-B) and Swin Transformer (Liu et al., 2021) (Swin-T) for backbone architectures. In ViT-B, the embedding dimension is 768, the layer depth is 12, the number of heads is 12, and the patch size is 16. In Swin-T, the embedding dimension is 96, the layer depth at each block is $[2, 2, 6, 2]$ (in total 12), the number of heads at each block is $[3, 6, 12, 24]$, the patch size is 4, and the sliding window size is 7. We set the input image size to 224 for all experiments but 192 for SimMIM pre-training. For continual learning methods, we use SI (Zenke et al., 2017), DER (Buzzega et al., 2020), and LUMP (Madaan et al., 2022). The implementation is built upon an official code of LUMP[3].

Training setups and hyperparameters We use AdamW optimizer (Loshchilov & Hutter, 2017) with cosine learning rate decay and the warmup for all experiments. For the pre-training phase at each task, we train the model 60 epochs on supervised learning and 100 epochs on unsupervised learning models as self-supervised learning methods without label supervision may require more iterations to converge. For fine-tuning, we basically perform 30 epochs training. For fine-tuning from the model pre-trained Imagenet 1K & 22K in Table 1, we set the number of training epochs to 10 as they rapidly converge within a few iterations. We set the hyperparameter for balancing the degree of regularization term $\lambda = 100$ for SI, $\lambda = 0.1$ for DER, and $\alpha = 0.1$ for LUMP. And the buffer size is 200 for rehearsal-based continual learning methods like DER and LUMP. We set the batch size to 64 for SimSiam pre-training, otherwise 128.

[1]https://github.com/microsoft/SimMIM
[2]https://github.com/facebookresearch/mae
[3]https://github.com/divyam3897/UCL

Table 3 summarizes the learning rate and training epochs for experiments, and we linearly scale the learning rate with $batch\_size/512$ in practice to reflect the input variance, followed by (Goyal et al., 2017).

Table 3: **Basic configurations** for three continual learning frameworks during pre-training and fine-tuning, where $\eta = 2e^{-4}$. We report the best combinations of the learning rate for pre-training and fine-tuning in the range of $[0.1, 0.2, 0.5, 1.0, 2.0, 5.0, 10.0] \times \eta$ and $[0.1, 0.5, 1.0, 5.0] \times \eta$, respectively.

	PRE-TRAINING	FINE-TUNING
SUPERVISED	lr: 1.0η, $epoch$: 60	lr: 0.5η, $epoch$: 30
CONTRASTIVE	lr: 0.2η, $epoch$: 100	lr: 1.0η, $epoch$: 30
MIM	lr: 5.0η, $epoch$: 100	lr: 5.0η, $epoch$: 30

Evaluation metrics We additionally introduce a linear classifier per task and measure fine-tuning and linear probe (or linear evaluation) accuracy by using the pre-trained backbone. For linear evaluation, we independently train a classifier over the training set of the corresponding task while freezing the backbone weights, and measure the accuracy on the validation set. For fine-tuning, we train a classifier as well as backbone weights over the training set of the target task. In this paper, we denote the term *backward transfer (BWT)*, acc_t^{bwt}, of task t as the fine-tuning accuracy (or linear probe accuracy) disparity between the model after training task t and the end of sequential training. Given T sequential tasks $\{\mathcal{T}_t\}_{t=1}^T$:

$$acc_t^{bwt} = acc_{t,T} - acc_{t,t}, \qquad (7)$$

where $acc_{i,j}$ is the measured accuracy for task i using sequentially pre-trained backbone model from the first task to j^{th} task. Followed by Xie et al. (2022a), we compute the attention distance at each attention head to analyze the change of attention during continual learning. Let $\boldsymbol{a}^{l,i}$ be an ith attention head output at layer l, the averaged attention distance $d_{\boldsymbol{a}^{l,i}}$ is measured by

$$d_{\boldsymbol{a}^{l,i}} = \sum_j c^* \bar{\boldsymbol{a}}_j^{l,i}, \qquad (8)$$

where c^* is the corresponding distance map and $\bar{\boldsymbol{a}}^{l,i}$ is a normalized attention matrix.

B. Further Discussions on Meta Learning

Chen et al. (2023) observes that the representation learned by *meta-learning* can be generic and useful for CL in terms of forgetting and forward transfer. Provided results and insights are interesting, therefore, here we discuss a few clear differences compared to Meta-learning (Finn et al.,

		SUPERVISED		Masked Model	
		FINAL ACC	NEG. BWT	FINAL ACC	NEG. BWT
FT	BASE	70.47 (± 7.91)	1.77 (± 1.56)	**86.17** (± 1.99)	20.90 (± 1.88)
	SI	70.93 (± 6.11)	4.80 (± 2.24)	83.50 (± 3.04)	16.80 (± 3.33)
	LUMP	N/A	N/A	85.80 (± 2.57)	18.30 (± 1.18)
LP	BASE	62.53 (± 6.58)	0.03 (± 0.05)	67.53 (± 4.05)	10.03 (± 1.17)
	SI	62.47 (± 5.12)	2.07 (± 2.16)	64.60 (± 2.72)	9.77 (± 0.87)
	LUMP	N/A	N/A	**68.97** (± 5.98)	12.60 (± 0.99)

Table 4: **Averaged fine-tuning and linear probe performance** on CIFAR-100 Split. We report the mean and standard deviation for three independent runs with randomly shuffled task orders.

2017). First, we do not need an external buffer for the re-hearsal, while FOMAML (Finn et al., 2017) requires keeping a (sub)set of past task data for inner loop update. Their reliance on rehearsal buffers has been criticized (Hadsell et al., 2020; Lomonaco et al., 2020; Wang et al., 2022) in the CL field. The performance of the rehearsal-based methods is sensitive to the size of the buffer (Prabhu et al., 2020), and they cannot be used in applications with concerns about data privacy. The MAML-based method requires exhaustive computational cost due to iterative and sequential inner loop updates for sampled past tasks at each iteration of current task training. On the other hand, our GLAD requires a marginal additional computation over the original training. Our GLAD gains benefit from adaptor-based modulation. In GLAD, the backbone focuses on capturing task-generic information, and a lightweight adaptor transforms them to task-adaptive attention for downstream task adaptation. Therefore, we can anytime recover the past model from the current one without storing full weights, by simply re-attaching the GLAD adaptor learned on the past task. Moreover, we can remove the unnecessary past task-specific knowledge from the model, or avoid the threat of performance degradation from training on noisy tasks/datasets - since the module for task-adaptive transformations is physically separate from generic representation (which is similar to (Yoon et al., 2020)). These practical utilizations are unavailable for Meta-learning as they update the entire weights without careful consideration of the knowledge modulation.

C. Additional Analyses

C.1. Evaluation on Task Order Shuffling

We additionally performed the evaluation on three different task orders for further reliable analyses in our experiments. At first, we measure fine-tuning and linear evaluation performance with the backward transfer using CIFAR-100. Similar to our experimental setting using ImageNet, we split CIFAR100 into ten tasks, containing ten disjoint classes per task. Then we train on the earlier nine tasks sequentially and measure the accuracy of the OOD task (10th), which is not seen during continual learning, as well as the forgetting

of the first task with the backward transfer. Regardless of the three orders of the task sequence, we observed that our proposed masked modeling-based UCL framework consistently surpasses supervised CL in terms of the fine-tuning performance on OOD task (T9) over all continual learning methods (Base/SI/LUMP), as similar to ImageNet results in Table 1. This is because masked modeling-based UCL continuously trains on more improved generic representations, evident in Figure 5 that continual learners gradually capture richer task-general (or low-level) features behaving with more local attention (i.e., lower attention distance), which retains localized information with strong local inductive bias, such as edges, patterns, and textures.

C.2. Attention Distance and Entropy from Different Transformer Backbones

We plot the attention distance and distribution of attention heads per layer for ViT-B in Figure 9 and Figure 10, respectively. And also, plot the attention distance and entropy of the distribution of attention heads per layer for Swin-T in Figure 11 and Figure 12. Both self-attention-based architectures similarly behave according to the learning frameworks, i.e., *Supervised*, *SimSiam*, and *SimMIM*. Note that two consecutive layers in Swin-T repeat relatively high and low values for both metrics since two successive swin transformer blocks (*S-MSA* and *SW-MSA* in their original paper) aggregate locality in different ranges.

C.3. Change of Aggregated Attention Distance and Entropy during Continual Learning

sup:subsec:analyses-entr We additionally visualize the movement of aggregated entropy of the distribution from each attention head in Figure 8. We visualize the plot for attention distance in Figure 5 again for a better comparison between them. Similar to observations in attention distance visualization, aggregated attention entropy gradually decreases as proceeding to pre-train more tasks, encouraging incremental model generalization. And aggregated attention entropy for supervised and contrastive learning-based continual learning frameworks suffer from a small diversity with a high average amount of information for all attention heads, compared to SimMIM.

C.4. Aggregated Attention Distance and Entropy while Freezing Partial Layers

We further visualize the movement of aggregated attention distance and entropy when freezing the two lowest and deepest layers in Figure 13 and Figure 14, respectively. These experiments are exactly from Table 2. Interestingly, if Sim-MIM freezes a few layers during continual learning, the remaining trainable layers tend to decrease their attention distance and entropy more actively. We believe that this
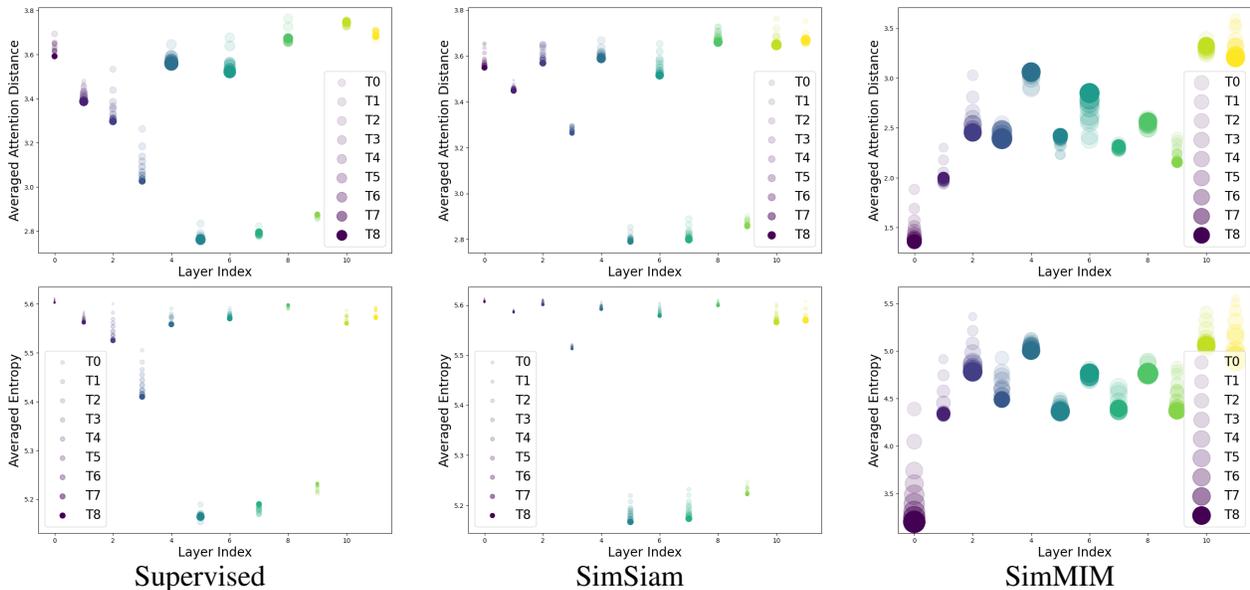
Figure 8: **Top row: Visualization of aggregated attention distance** on an OOD task (T9) at each layer at the end of each continual pre-training task phase (T0→T8). The radius of marker indicates the standard deviation over attention heads in the corresponding layer. **Bottom row: Aggregated attention entropy** on T9.

is a reason that the SimMIM does not find noticeable performance degeneration in fine-tuning, even freezing a few layers during continual learning. However, we didn't observe a significant change in supervised continual learning. This is because supervised learning is prone to rigidly focus on different features according to the layer depths (i.e., global to local), and therefore cannot flexibly cope with capturing locality inductive bias.

D. Limitations and Societal Impact

Limitations   Although we have shown promising results and findings in multiple CL frameworks, our proposed GLAD module requires extra memory and computation for the regularization term. Moreover, We design the experiment of continual pre-training with less than ten tasks, which is insufficient to evaluate lifelong learning. Reducing the memory cost for the additional adaptor and extending our framework to a larger number of continual pre-training tasks will be important for future work.

Negative societal impact   Our work doesn't store past task data in the buffer so that we can avoid the negative societal impact raised by data privacy issues in the community.
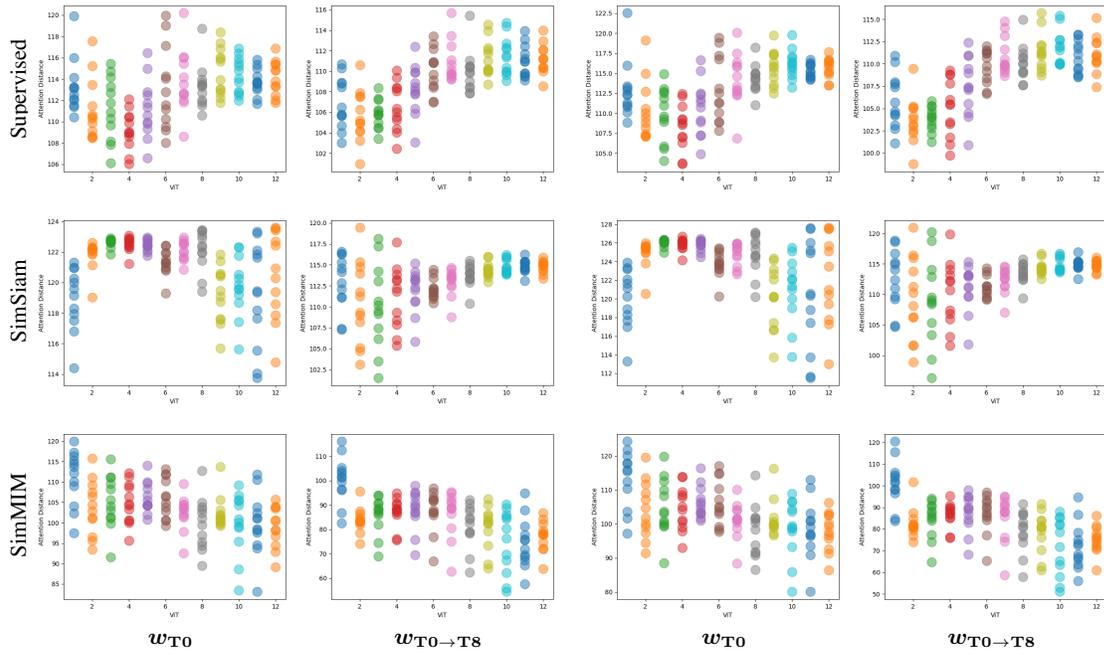
Figure 9: **ViT-B Attention distance** of the first (T0) and the last task (T8) in a task sequence with respect to three continual pre-trained frameworks right after the completion of the first (w_{T0}) and last task ($w_{T0 \rightarrow T8}$).
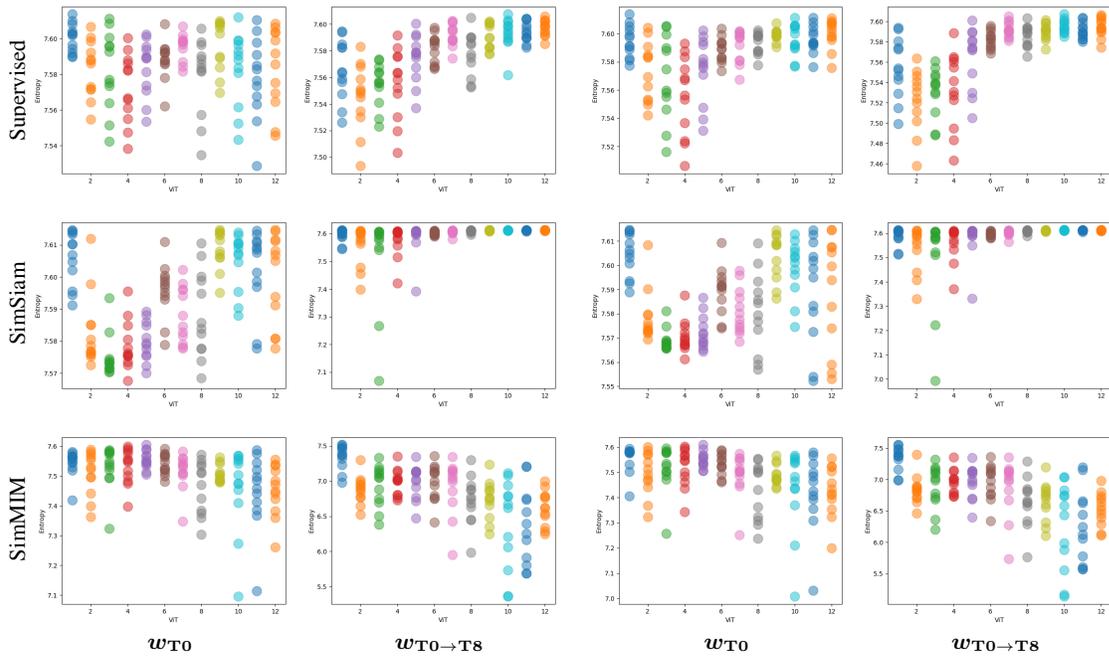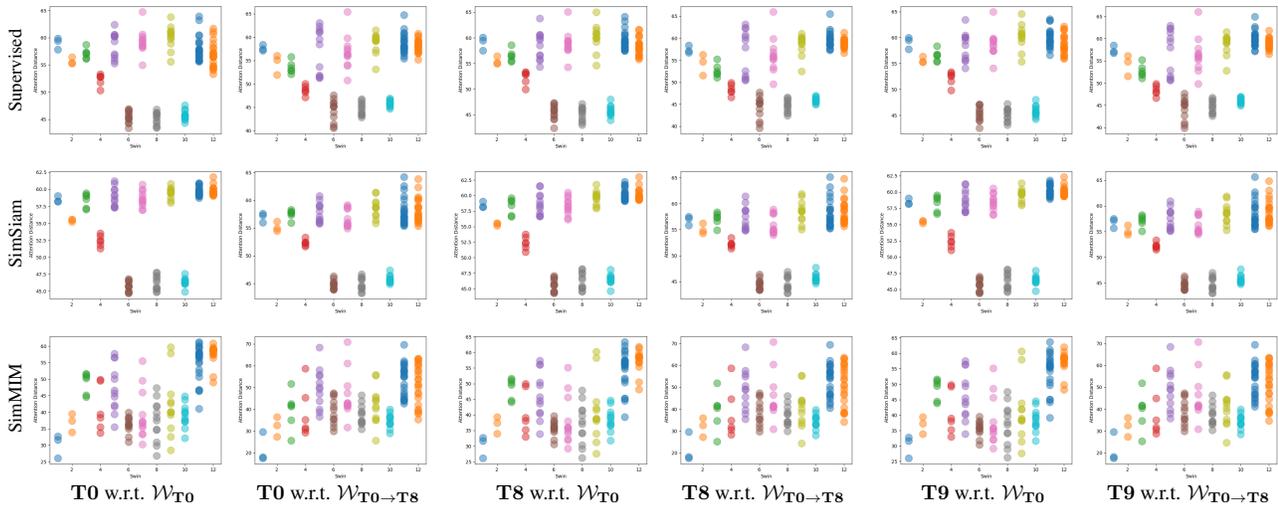

Figure 10: **ViT-B Attention entrtopy** of the first (T0) and the last task (T8) in a task sequence with respect to three continual pre-trained frameworks right after the completion of the first (w_{T0}) and last task ($w_{T0 \rightarrow T8}$).

Figure 11: **Swin-T attention distance** of an in-distribution (T0) and ood task (T9) with respect to three continual pre-trained frameworks right after the completion of the first ($\boldsymbol{w}_{\mathbf{T0}}$) and last task ($\boldsymbol{w}_{\mathbf{T0 \rightarrow T8}}$).
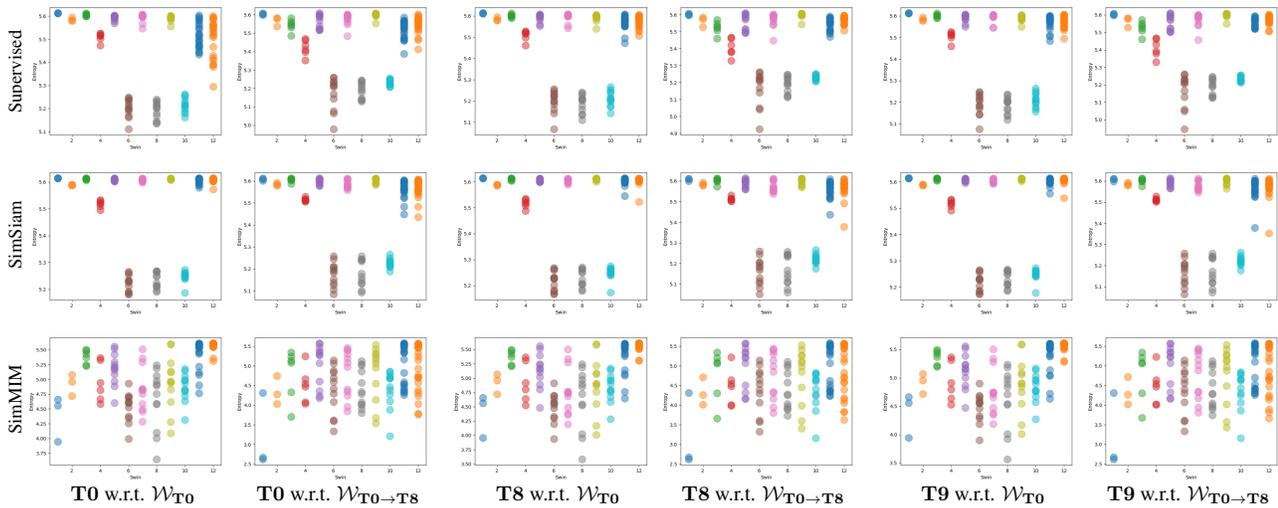

Figure 12: **Swin-T attention entropy** of an in-distribution (T0) and ood task (T9) with respect to three continual pre-trained frameworks right after the completion of the first ($\boldsymbol{w}_{\mathbf{T0}}$) and last task ($\boldsymbol{w}_{\mathbf{T0 \rightarrow T8}}$).
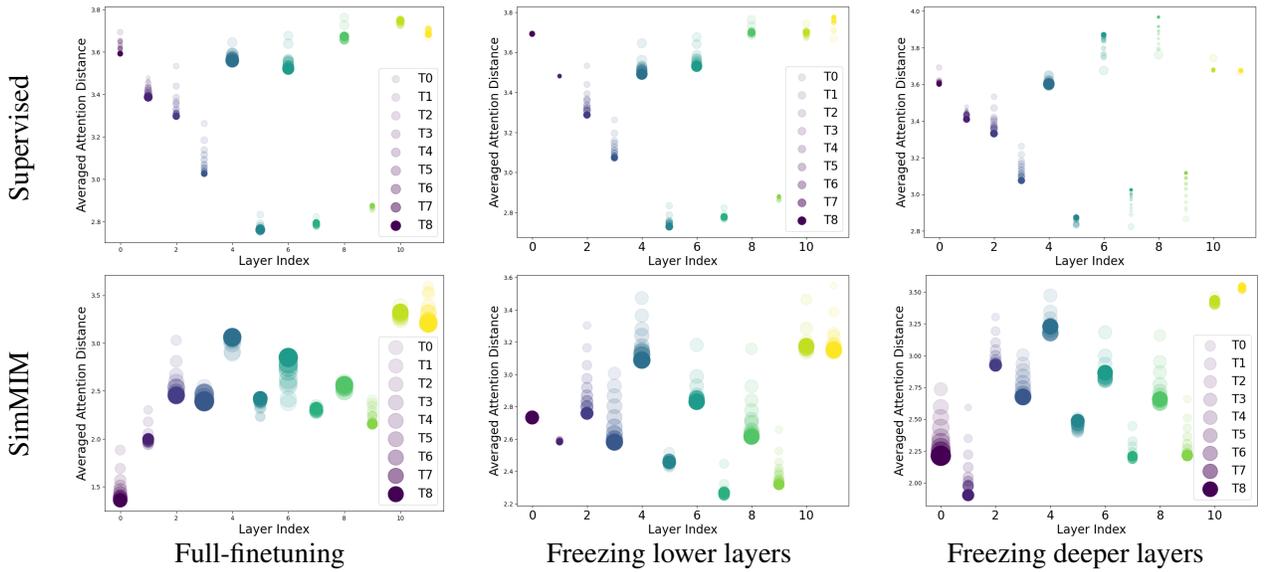
Figure 13: **Visualization of aggregated attention distance** on an OOD task (T9) at each layer at the end of each continual pre-training task phase (T0→T8). We freeze the two lowest or deepest layers after pre-training the first task (T0). The radius of the marker indicates the standard deviation over attention heads in the corresponding layer.
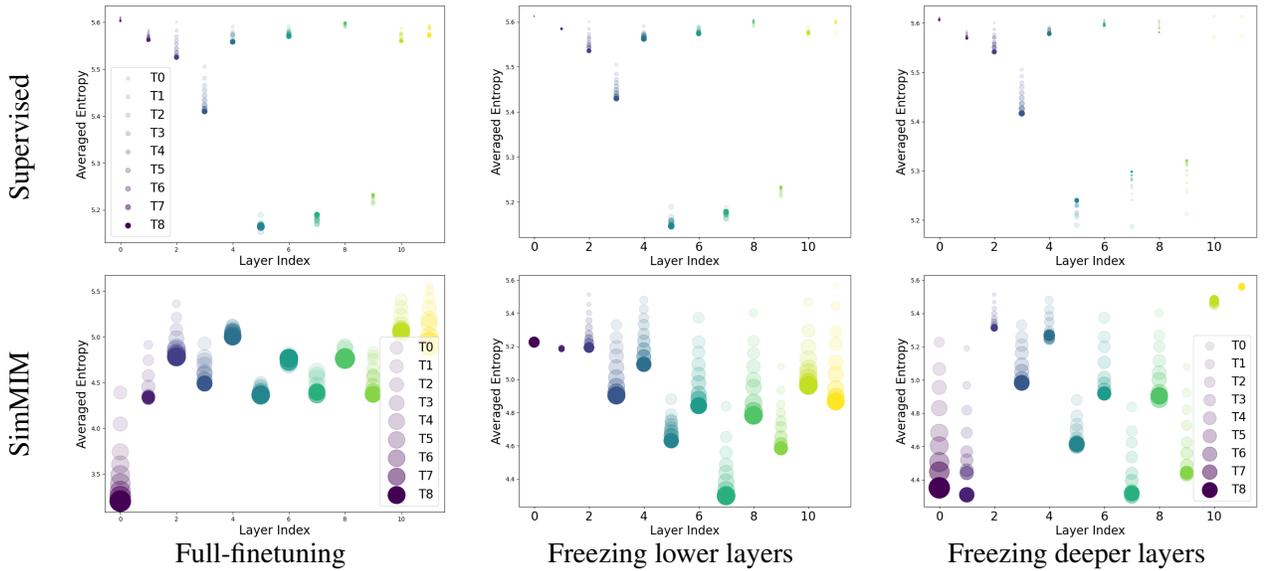

Figure 14: **Visualization of aggregated attention entropy** on an OOD task (T9) at each layer at the end of each continual pre-training task phase (T0→T8). We freeze the two lowest or deepest layers after pre-training the first task (T0). The radius of the marker indicates the standard deviation over attention heads in the corresponding layer.