On the Transferability of Parameter-Efficient Continual Learning for Vision Transformers

Leon Ackermann *
Faculty of Informatics
Università della Svizzera italiana
Lugano, Switzerland
ackerl@usi.ch

Van-Linh Nguyen
Department of Computer Science
National Chung Cheng University
Chiayi, Taiwan
nvlinh@ccu.edu.tw

Abstract

Continual Learning (CL) is the process of continually adapting a model to a new stream of data. Within CL, pre-trained transformer-based vision models such as the ogirinal Vision Transformer (ViT) have recently received increased attention. Various CL methods exist that adapt the base version of the Vision Transformer (ViT-Base) efficiently with outstanding results. However, ViT-Base underperforms several advanced transformer-based vision models on traditional image classification benchmarks. While in Natural Language Processing, state-of-the-art finetuning techniques are evaluated on the most up-to-date models, CL is missing such a comparison. Despite the existence of advanced transformer-based vision models in various sizes, state-of-the-art parameter-efficient CL methods fall back on ViT-Base for benchmarking. In this study, we address this gap by evaluating various sizes of ViT and multiple variants of DeiT3 and DinoV2, two of the best-performing vision transformers, on six state-of-the-art CL methods that are based on prompt tuning and adapter tuning. The experimental results show that the prompt-based techniques DualPrompt and L2P transfer more reliably to new model types and sizes compared to the adapter-based approaches. Furthermore, we show that model size is more important for prompt-based than adapter-based techniques. Finally, we select ViT-Large as the most performant and hence the model of choice. With these findings, we aim to further advance the understanding of the connection between model architecture and the continual learning approach.

1 Introduction

In Continual Learning (CL), a model is continuously trained on a stream of data that can belong to different distributions and possess different classes. A fundamental challenge in CL is catastrophic forgetting, where a model loses its knowledge of a previously learned task after being trained for the subsequent task (McCloskey and Cohen, 1989). Numerous algorithms have been proposed to address the issue of catastrophic forgetting (Rolnick et al., 2019; Kirkpatrick et al., 2017; Li and Hoiem, 2017). These methods generally fall into two categories for the domain of computer vision: (1) those that train a neural network from scratch, often utilizing a convolutional architecture, and (2) those that efficiently adapt pre-trained models (PTM), predominantly transformer-based image models (Zhou et al., 2024a). CL with transformer-based PTMs has recently received more attention due to the transformer's strong generalization capabilities (Raffel et al., 2020).

A range of different parameter-efficient (PEFT) CL methods for transformer-based PTMs exist and the original Vision Transformer (ViT) (Dosovitskiy et al., 2020) has become the benchmark

*Work done during an internship at National Chung Cheng University (CCU), Chiayi Taiwan.

38th Workshop on Fine-Tuning in Machine Learning (NeurIPS 2024).

for evaluation in the field. However, despite the existence of more advanced transformer-based image models such as DeiT3 (Touvron et al., 2022) or DinoV2 (Oquab et al., 2023), these have not yet been benchmarked with state-of-the-art CL methods. This lack of evaluation contrasts sharply with the Natural Language Processing (NLP) domain, where a diverse range of models, varying in architecture and pre-training strategies, are rigorously tested to assess the transferability of state-of-the-art fine-tuning approaches including parameter-efficient techniques (Hu et al., 2021; Lester et al., 2021). To the best of our knowledge, the field of CL lacks a similar comprehensive comparison of state-of-the-art parameter-efficient CL methods for various PTMs. This is a significant gap that needs to be addressed to fully understand how advanced CL methodologies transfer to novel model types. In this study, we evaluate six state-of-the-art parameter-efficient CL methods and traditional finetuning on Deit3 and DinoV2, two of the best-performing Vision Transformers, and include the original ViT as a baseline.

In particular, we select CL methods that build on the parameter-efficient finetuning techniques of prompt tuning (Lester et al., 2021) and adapter tuning (Pfeiffer et al., 2020). Research in NLP has shown that prompt-based approaches gain considerably from bigger models, resulting in greater performance (Lester et al., 2021). However, in the context of CL, only the base version of the original ViT has been evaluated, despite the fact that smaller and bigger ViT models exist. Together with the different model sizes of DeiT3 and DinoV2, the NLP studies demonstrate the possibility for scalability for CL with Vision Transformers.

In this study, we aim to contribute to the field of CL and fine-tuning techniques in the following:

1. We present the first evaluation of six state-of-the-art parameter-efficient CL methods and continual finetuning beyond the base model of the original Vision Transformer. Specifically, we include four sizes of ViT and DeiT3 and three sizes of DinoV2.
2. We show that the prompt-based approaches DualPrompt and L2P transfer best to new model types and sizes out of all methods. In addition, model size has a larger effect on prompt-based methods compared to adapter-based methods.
3. We identify ViT-Large as the model of choice as it performs the best for most of the CL methods and datasets.

2 Related Work

2.1 Class-Incremental Learning

Continual Learning is categorized into three different forms (Van de Ven et al., 2022). In **Task-Incremental Learning** (TIL), the algorithm must solve a series of tasks while knowing which task is being tested at any given time. In **Domain-Incremental Learning** (DIL), the algorithm learns different tasks over time without knowing the specific task domain during testing, but the labels remain consistent across tasks. In this work, we focus on the most challenging scenario of **Class-Incremental Learning** (CIL), where the algorithm faces tasks with different sets of classes and must learn to distinguish among all classes seen so far, without task-specific information during testing. CIL is a form of CL that can be formally defined as an algorithm f that must learn a function $f : \mathcal{X} \rightarrow \mathcal{Y}$ for a series of tasks $\{D_1, D_2, \ldots, D_n\}$. Here, $\mathcal{X}$ and $\mathcal{Y}$ represent the input and output spaces, respectively, which are defined for each task D_i. Specifically, in CIL, a model is trained to minimize $\sum_{(\mathbf{x}_j, y_j) \in \mathcal{D}_1 \cup \cdots \mathcal{D}_n} \ell\left(f\left(\mathbf{x}_j\right), y_j\right)$ where $\ell(\cdot, \cdot)$ is the loss function to compute the difference between the models prediction and the ground truth. A non-negative backward transfer and forward transfer are essential to achieve good performance on all tasks. This means that learning a new task should not degrade the performance on previous tasks (backward transfer) and should ideally improve performance on future tasks (forward transfer) (Lopez-Paz and Ranzato, 2017). The primary challenge in CL is to prevent significant negative backward transfer, commonly referred to as catastrophic forgetting (McCloskey and Cohen, 1989). In the field of Computer Vision, various methods have been developed to prevent catastrophic forgetting that is based on replay and regularization of parameter isolation (De Lange et al., 2021). Most methods train a convolutional architecture from scratch such as ER (Rolnick et al., 2019), EWC (Kirkpatrick et al., 2017) or LwF (Li and Hoiem, 2017). An exception is DyTox which trains a transformer-based architecture from scratch and expands the model

with a task-specific learned token (Douillard et al., 2022). Fine-tuning pre-trained transformer-based models is also popular within CL. While traditional methods like EWC or LwF adopted for the Vision Transformer (ViT) (Dosovitskiy et al., 2020) do not work well (Pelosin et al., 2022), most methods focus on using parameter-efficient prompt-based or adapter-based approaches to prevent catastrophic forgetting (Zhou et al., 2024a). State-of-the-art methods include DualPrompt (Wang et al., 2022a), L2P (Wang et al., 2022b) or CODA-Prompt (Smith et al., 2023)(prompt-based) and ADAM (Zhou et al., 2023), RanPAC (McDonnell et al., 2024) and EASE (Zhou et al., 2024b) (adapter-based).

2.2 Vision Transformers

The Vision Transformer (ViT) (Dosovitskiy et al., 2020) has become a benchmark model for CL in computer vision with pre-trained models. When first introduced, ViT demonstrated that a purely transformer-based architecture can achieve outstanding performance in image classification tasks as opposed to convolutional neural networks (CNN) that dominated computer vision previously. The ViT closely follows the structure of the original transformer architecture (Vaswani et al., 2017) with the difference of a patch embedding module to process images instead of text. Specifically, an image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ is transformed into 2D patches $\mathbf{x}_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$, which serve as input for the transformer model. Here, C denotes the number of channels, (P, P) represents the dimensions of each patch, and $N = \frac{H \cdot W}{P^2}$ indicates the total number of patches. Since the introduction of the original ViT, numerous transformer-based image models have been developed. Liu et al. (2023) proposed a taxonomy categorizing these models into six groups: (i) the original ViT, (ii) CNNs enhanced with transformer features, (iii) transformers augmented with CNN features or (iv) local attention mechanisms, (v) hierarchical transformers, (vi) Deep transformers, and (vii) transformers trained with self-supervision. Additionally, architectural variations include differences in patch size, image resolution used for pre-training, and the number of parameters. The survey authors demonstrate that improvements like the student-teacher self-supervised learning approach in DinoV2 (Oquab et al., 2023) or the CNN-enhanced training strategy in DeiT3 (Touvron et al., 2021) can substantially improve performance over the original ViT. Furthermore, the study indicates that among 40 distinct vision transformers, each accessible in varied sizes, most outperform the original ViT on ImageNet-1k (Deng et al., 2009). However, CL algorithms for pre-trained models (refer to section 3.2) have primarily been evaluated on the base version of the original ViT, leaving the potential of more advanced transformer-based vision models largely unexplored. One study has explored architectural differences primarily in CNNs, with a brief examination of the original ViT concerning continual learning (Mirzadeh et al., 2022). However, the authors restricted their analysis to simple continual fine-tuning, neglecting state-of-the-art CL methods. This approach led to suboptimal model performance, leaving a significant gap in the comprehensive evaluation of more advanced CL techniques across different transformer-based vision architectures.

3 Experiments

3.1 Evaluation Benchmarks

We use the recently introduced Deepfake Detection Benchmark for Continual Learning (CDDB) (Li et al., 2023). The benchmark is composed of three different sets of benchmarks with an EASY, HARD, and LONG part that all include Deepfakes from different generator models. We choose the HARD set of the benchmark that includes pairs of Deepfake and real images from 5 different generator models with a total of 26940 samples. In addition, we include CIFAR-100 (Krizhevsky et al., 2009) as another common benchmark used within CL to enable comparisons to other studies. We report the average test accuracy over all tasks defined as

$$\text{ACC} = \frac{1}{T} \sum_{i=1}^{T} R_{T,i}$$

and the backward transfer (forgetting) defined as

$$\text{BWT} = \frac{1}{T-1} \sum_{i=1}^{T-1} R_{T,i} - R_{i,i}$$

where $R \in \mathbb{R}^{T \times T}$ is a matrix where one entry $R_{i,j}$ is the test accuracy of the model of task t_j after being trained on task t_i (Lopez-Paz and Ranzato, 2017).

3.2 Continual Learning Methods

We assess the top-performing techniques from two categories: prompt-based and adapter-based approaches. All included methods perform CIL and do not store samples in the buffer during testing. Both prompt-based and adapter-based approaches fall under parameter-efficient finetuning techniques, as they modify less than 1% of the original parameters.
Prompt-based methods: We incorporate DualPrompt (Wang et al., 2022a), L2P (Wang et al., 2022b), CODA-Prompt (Smith et al., 2023), see Appendix Section B.1 for further details.
Adapter-based methods: We choose ADAM-Adapter (Zhou et al., 2023), RanPAC (McDonnell et al., 2024) and EASE (Zhou et al., 2024b), refer to Appendix Section B.2 for more details.
Baseline: Lastly, as a baseline we include the continual finetuning approach where all weights of the model are adjusted.

3.3 Vision Transformers backbones

We include four variants of DeiT3 (Touvron et al., 2022), three variants of DinoV2 (Oquab et al., 2023) and four variants ViT (Dosovitskiy et al., 2020). Each variant has a different parameter count, embedding dimension, and other architectural differences (see Table 1 and Appendix Section A for more details on the models). We run the experiments with the evaluation pipeline PILOT by Zhou et al. (2024a) and extend it to support the models DeiT3 and DinoV2.

Table 1: Model variants

Model	Variant	Input Size	# Layers	Hidden Size	MLP size	# Heads	# Parameters
ViT	Tiny	224×224	12	192	3072	3	2 M
	Small	224×224	12	384	3072	6	22 M
	Base	224×224	12	786	3072	12	86 M
	Large	224×224	22	1024	4096	16	307 M
DeiT3	Small	224×224	12	384	3072	6	5 M
	Medium	224×224	12	512	3072	6	38.8 M
	Base	224×224	12	768	3072	12	80 M
	Large	224×224	24	1024	4096	16	307 M
DinoV2	Small	518×518	12	384	3072	6	22 M
	Base	518×518	12	768	3072	12	86 M
	Large	518×518	24	1024	4096	16	307 M

4 Results

We report the average test accuracy ACC and the backward transfer (forgetting) BWT over all tasks in table 1a and table 1b for CDDB and CIFAR, respectively. Overall, the models achieve a higher accuracy on CIFAR than CDDB with a maximum of 95.72 and corresponding forgetting of 2.62 with ViT-Large and RanPAC. The same model-method configuration scores the highest performance on CDDB with an accuracy of 84.18 and 7.39 in forgetting. ViT-Large is most often the strongest-performing model out of all model types and sizes. For CIFAR, ViT-Large is the best model in all methods except for DualPrompt while in some cases other models like DeiT3-Large or DinoV2-Base perform best with certain methods on CDDB. When isolating model performance concerning the utilized method for CDDB, RanPAC works best for the variants of ViT and DeiT3 while ADAM-Adapter suits DinoV2 the best. This pattern partly continues as RanPAC works best for the ViT model family. However, the methods DualPrompt and L2P fit best for the model variants of DeiT3 and DinoV2. For both CDDB and CIFAR, the methods EASE, RanPAC, and CODA-Prompt perform significantly worse on the model variants of DeiT3 and DinoV2 when compared to the ViT model family. On CDDB, the combined average accuracies of the DeiT3 and DinoV2 models for CODA-Prompt, EASE, and RanPAC are 30.53, 37.44, and 58.04, respectively, while the average accuracies of the ViT variants are 50.50, 64.62, and 79.85. The same observation can be made for CIFAR. The other methods ADAM-Adapter, DualPrompt, and L2P perform smoothly with each variant of all model types except for DeiT3 on CDDB. Lastly, finetuning yields good results for ViT

Figure 1: Results. Values in blue correspond to the best-performing model for a single method. Values highlighted in yellow denote the best method for a single model. Values are colored in green if they have been identified as yellow and blue at the same time. **Bold** values highlight the best-performing model among all methods and model sizes of a specific model family.

(a) Resuts for CDDB

	L2P		DualPrompt		CODA-Prompt		ADAM-Adapter		EASE		RanPAC		Finetune	
Model variant	ACC	BWT	ACC	BWT	ACC	BWT	ACC	BWT	ACC	BWT	ACC	BWT	ACC	BWT
ViT-Tiny	29.54	28.62	35.94	9.78	43.03	23.81	56.81	7.06	59.44	15.28	71.70	4.21	41.47	52.32
ViT-Small	40.93	12.59	42.38	12.82	47.29	25.51	68.97	5.58	63.09	8.96	79.52	5.13	49.65	42.62
ViT-Base	52.91	13.76	41.18	19.16	55.17	12.52	70.27	4.82	66.66	6.63	84.00	2.76	48.22	59.62
ViT-Large	47.49	18.86	47.88	19.33	56.92	6.32	66.99	5.86	69.27	0.53	**84.18**	7.39	59.52	48.03
DeiT3-Small	34.15	28.09	46.45	18.06	27.29	33.69	60.36	6.56	41.60	18.67	56.80	7.19	43.82	51.22
DeiT3-Medium	44.76	19.20	47.57	5.73	33.27	7.19	57.11	4.34	33.22	31.72	59.74	6.04	44.75	24.69
DeiT3-Base	38.28	25.09	49.38	19.32	34.32	21.82	51.12	5.22	40.06	13.67	61.10	5.30	55.64	42.37
DeiT3-Large	45.38	11.55	51.42	7.36	21.40	19.38	48.48	8.97	40.23	12.13	50.09	8.97	**63.25**	17.55
DinoV2-Small	38.66	29.18	53.29	10.42	36.35	17.18	**77.11**	4.77	40.28	20.11	60.40	10.79	31.85	3.15
DinoV2-Base	50.21	12.83	56.57	10.60	27.87	32.79	75.74	5.48	33.20	27.09	61.35	8.08	33.13	13.36
DinoV2-Large	50.80	13.35	55.58	9.85	22.18	21.66	75.43	4.98	33.46	25.04	56.78	7.87	27.30	29.11

(b) Results for CIFAR

	L2P		DualPrompt		CODA-Prompt		ADAM-Adapter		EASE		RanPAC		Finetune	
Model variant	ACC	BWT	ACC	BWT	ACC	BWT	ACC	BWT	ACC	BWT	ACC	BWT	ACC	BWT
ViT-Tiny	67.64	11.23	77.16	10.58	75.92	15.88	77.42	9.9	77.92	13.13	86.83	6.73	53.16	65.32
ViT-Small	84.71	5.88	86.93	5.93	87.45	7.60	88.57	6.17	87.28	10.00	92.39	4.32	76.05	34.56
ViT-Base	87.58	5.10	88.52	8.54	91.33	5.07	90.91	5.23	90.61	8.37	94.33	3.64	79.10	30.47
ViT-Large	91.21	4.76	90.68	6.46	92.40	3.82	93.36	3.60	93.18	6.51	**95.72**	2.62	89.44	12.54
DeiT3-Small	71.22	11.43	76.34	6.91	16.27	9.92	56.11	9.77	11.94	13.04	36.15	12.50	65.01	54.63
DeiT3-Medium	78.88	10.08	80.35	9.48	13.63	8.56	53.62	8.92	11.06	11.64	36.99	13.09	68.58	50.09
DeiT3-Base	80.56	8.03	82.95	9.68	16.84	18.29	36.54	9.04	12.34	11.88	33.28	12.18	70.21	42.01
DeiT3-Large	84.31	10.48	**84.59**	8.67	14.91	13.63	30.75	8.57	13.26	10.74	28.58	10.64	82.16	34.30
DinoV2-Small	75.59	12.06	80.88	9.44	11.56	6.34	79.10	8.06	13.87	8.69	19.96	11.06	11.84	23.24
DinoV2-Base	85.51	9.09	88.21	7.54	14.34	5.26	82.85	5.99	10.99	10.68	19.92	18.88	17.40	46.90
DinoV2-Large	87.29	8.36	**91.56**	5.411	15.03	4.86	85.13	4.36	10.48	10.96	31.64	11.78	22.11	48.76

and DeiT3 while rarely producing model performances that compete with the outcomes of RanPAC, ADAM-Adapter, DualPrompt, or L2P. Moreover, models adapted with finetuning experience the strongest degree of forgetting across all methods.

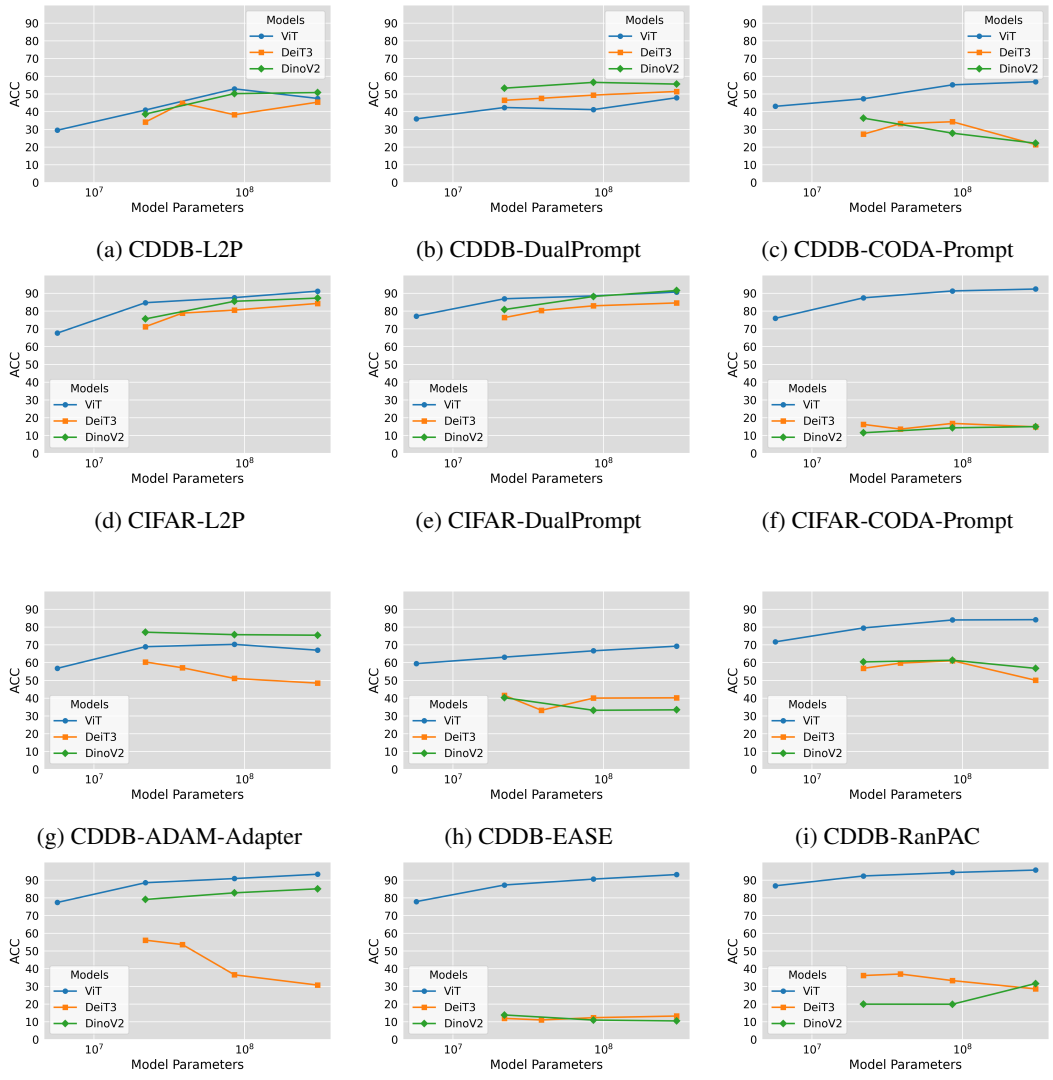
5 Discussion

We group our observations of the results into two categories. Firstly, we discuss how well the parameter-efficient CL methods transfer from the ViT to DeiT3 and DinoV2 (see Section 5.1). Secondly, in Section 5.2, we assess the importance of model size for the performance in CIL.

5.1 Transferability

As reported in Section 4, a pattern emerges where the methods DualPrompt, L2P, and ADAM-Adapter transfer well from DiT to DeiT3 and DinoV2 but RanPAC, EASE, and CODA-Prompt fail to do so. More generally, methods that are based on prompt tuning seem to transfer more easily to other models than methods based on adapter tuning. Figure 2 visualizes this observation. Similar conclusions were made by Su et al. (2022) in the domain of NLP showing that prompt tuning works for different models like Roberta (Liu, 2019) and T5 (Raffel et al., 2020). An exception to this insight is CODA-Prompt which fails to transfer from ViT to DeiT3 and DinoV2. While L2P and DualPrompt function similarly by training tunable prompts and selecting them during test time, CODA-Prompt adapts prompt components that combine to a final prompt during the evaluation. Further investigations could assess why this fundamental difference contributes to the poor transfer performance of CODA-Prompt. In the case of EASE and RanPAC, their unsatisfying transfer performance is surprising. A possible explanation could be that the methods are more sensitive to hyperparameter choice. However, we rule out this hypothesis since the performance differences between the ViT, DeiT3, and DinoV2 are too significant. The poor transfer performance of RanPAC and EASE is unexpected, as both

build on the ADAM-Adapter, which showed fewer issues when transferring to DeiT3 and DinoV2. RanPAC and EASE improve upon ADAM-adapter by adjusting modules for all tasks, whereas ADAM-Adapter only trains on the first task and generalizes to the remaining tasks during test time (Zhou et al., 2024b; McDonnell et al., 2024). A possible reason for RanPAC's and EASE's poor transfer performance could be that the different optimization strategy weakens the generalization of Deit3 and DinoV2, ultimately leading to more frequent forgetting. This question can be addressed in future work. Finally, ViT often outperforms DeiT3 and DinoV2, with a few outliers such as DualPrompt and ADAM-Adapter for CDDB. Future studies could investigate whether this remains true if more complex hyperparameter optimization techniques are used for each learning scenario (Semola et al., 2024).



(a) CDDB-L2P (b) CDDB-DualPrompt (c) CDDB-CODA-Prompt

(d) CIFAR-L2P (e) CIFAR-DualPrompt (f) CIFAR-CODA-Prompt

(g) CDDB-ADAM-Adapter (h) CDDB-EASE (i) CDDB-RanPAC

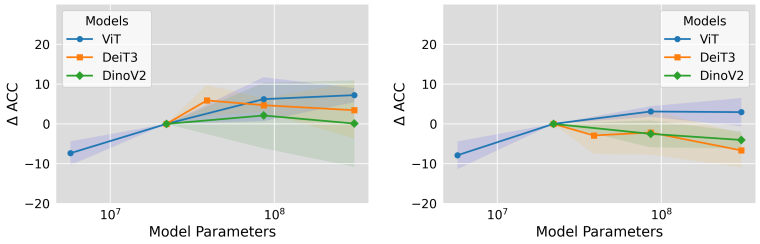
(j) CIFAR-ADAM-Adapter (k) CIFAR-EASE (l) CIFAR-RanPAC

Figure 2: Test accuracy of prompt-based approaches on CDDB (see figures 2a to 2c) and CIFAR (see figures 2d to 2f) and adapter-based approaches on CDDB (see figures 2g to 2i) and CIFAR (see figures 2j to 2l).

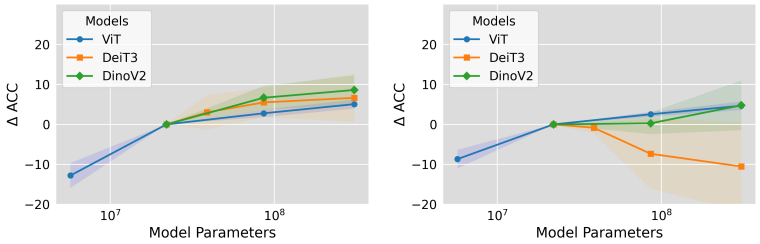
5.2 Impact of Model Size

Next to the transferability of the parameter-efficient CL methods to other model types, we discuss how well methods transfer to different model sizes. Figure 3 depicts the difference in test accuracy ACC

of all model variants compared to the small version of each model type. The differences are averaged across all methods belonging to the class of prompt-based (figures 3a and 3c) and adapter-based approaches (figures 3b and 3d). We observe that prompt-based methods scale well for each model type. For CDDB, the highest difference can be seen for ViT-Large that scores on average 7.23% higher than ViT-Small and $14,59\%$ better than ViT-Tiny. On CIFAR, scale has the largest impact on DinoV2 models as DinoV2-Large scores on average 8.61% higher compared to DinoV2-Small. The same relationship but less visible is true for ViT concerning the adapter-based approaches. ViT-Large scores on average 2.96% and 10.44% higher on CDDB and 4.67% and 13.36% higher on CIFAR compared to ViT-Small and ViT-Tiny respectively. However, for DeiT3 and DinoV2, the opposite is the case. Figures 3b and 3d show that scaling does not improve model performance but it even harms it. DeiT3 is the most affected by this phenomenon as DeiT3-Large scores 6.65% less on CDDB and 10.54% less on CIFAR compared to DeiT3-Small. This finding is consistent with the unsatisfactory transfer performance of adapter-based techniques for DeiT3 and DinoV2. The analysis of this insight is to be approached in future work.



(a) Prompt-based methods with CDDB (b) Adapter-based methods with CDDB



(c) Prompt-based methods with CIFAR (d) Adapter-based method with CIFAR

Figure 3: Average difference test accuracy of prompt-based and adapter-based approaches. The difference in test accuracy is measured relative to the small model of each model type. The shaded area represents the standard deviation.

6 Conclusion

In this paper, we investigated how six different state-of-the-art parameter-efficient Continual Learning (CL) methods transfer to different sizes of DeiT3 and DinoV2, two of the best-performing pre-trained transformer-based vision transformers, for the task of Deepfake Detection on CDDB. As a baseline, we included continual finetuning, multiple sizes of the original ViT, and the popular CL benchmark CIFAR100. Up to this date, the base version of the original Vision Transformer has been exclusively used as the benchmark model for CL with pre-trained transformer-based vision models. Hence, this paper addresses the lack of evaluation of more advanced transformer-based vision models within CL. Similar to other domains, we identify that model size is an important factor for model accuracy and forgetting. Overall, ViT-Large stands out as the model of choice performing best with most CL methods. Moreover, we show that the prompt-based methods DualPrompt and L2P transfer better to new model classes and sizes compared to their adapter-based counterparts. This insight can be used to further harness the capabilities of DeiT3 and DinoV2 to ultimately bypass the original ViT in parameter-efficient CL. With the findings, we hope to advance the understanding of how model performance can be increased within CL. We show that this is particularly important for the task of Deepfake Detection which urgently needs to profit from better methods that can adapt models continually.

References

M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3366–3385, 2021.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

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

A. Douillard, A. Ramé, G. Couairon, and M. 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.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

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

B. Lester, R. Al-Rfou, and N. Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

C. Li, Z. Huang, D. P. Paudel, Y. Wang, M. Shahbazi, X. Hong, and L. Van Gool. A continual deepfake detection benchmark: Dataset, methods, and essentials. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1339–1349, 2023.

Z. Li and D. Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

Y. Liu. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Y. Liu, Y. Zhang, Y. Wang, F. Hou, J. Yuan, J. Tian, Y. Zhang, Z. Shi, J. Fan, and Z. He. A survey of visual transformers. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

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

M. McCloskey and N. J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989.

M. D. McDonnell, D. Gong, A. Parvaneh, E. Abbasnejad, and A. van den Hengel. Ranpac: Random projections and pre-trained models for continual learning. *Advances in Neural Information Processing Systems*, 36, 2024.

S. I. Mirzadeh, A. Chaudhry, D. Yin, T. Nguyen, R. Pascanu, D. Gorur, and M. Farajtabar. Architecture matters in continual learning. *arXiv preprint arXiv:2202.00275*, 2022.

M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

F. Pelosin, S. Jha, A. Torsello, B. Raducanu, and J. van de Weijer. Towards exemplar-free continual learning in vision transformers: an account of attention, functional and weight regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3820–3829, 2022.

J. Pfeiffer, A. Kamath, A. Rücklé, K. Cho, and I. Gurevych. Adapterfusion: Non-destructive task composition for transfer learning. *arXiv preprint arXiv:2005.00247*, 2020.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

D. Rolnick, A. Ahuja, J. Schwarz, T. Lillicrap, and G. Wayne. Experience replay for continual learning. *Advances in neural information processing systems*, 32, 2019.

R. Semola, J. Hurtado, V. Lomonaco, and D. Bacciu. Adaptive hyperparameter optimization for continual learning scenarios. *arXiv preprint arXiv:2403.07015*, 2024.

J. S. Smith, L. Karlinsky, V. Gutta, P. Cascante-Bonilla, D. Kim, A. Arbelle, R. Panda, R. Feris, and Z. Kira. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11909–11919, 2023.

Y. Su, X. Wang, Y. Qin, C.-M. Chan, Y. Lin, H. Wang, K. Wen, Z. Liu, P. Li, J. Li, et al. On transferability of prompt tuning for natural language processing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3949–3969, 2022.

H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

H. Touvron, M. Cord, and H. Jégou. Deit iii: Revenge of the vit. In *European conference on computer vision*, pages 516–533. Springer, 2022.

G. M. Van de Ven, T. Tuytelaars, and A. S. Tolias. Three types of incremental learning. *Nature Machine Intelligence*, 4(12):1185–1197, 2022.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Z. Wang, Z. Zhang, S. Ebrahimi, R. Sun, H. Zhang, C.-Y. Lee, X. Ren, G. Su, V. Perot, J. Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European Conference on Computer Vision*, pages 631–648. Springer, 2022a.

Z. Wang, Z. Zhang, C.-Y. Lee, H. Zhang, R. Sun, X. Ren, G. Su, V. Perot, J. Dy, and T. Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 139–149, 2022b.

D.-W. Zhou, H.-J. Ye, D.-C. Zhan, and Z. Liu. Revisiting class-incremental learning with pre-trained models: Generalizability and adaptivity are all you need. *arXiv preprint arXiv:2303.07338*, 2023.

D.-W. Zhou, H.-L. Sun, J. Ning, H.-J. Ye, and D.-C. Zhan. Continual learning with pre-trained models: A survey. *arXiv preprint arXiv:2401.16386*, 2024a.

D.-W. Zhou, H.-L. Sun, H.-J. Ye, and D.-C. Zhan. Expandable subspace ensemble for pre-trained model-based class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23554–23564, 2024b.

A Vision Transformers

Original Vision Transformer As described in Section 2.2, the original ViT (Dosovitskiy et al., 2020) is based on the transformer architecture (Vaswani et al., 2017). To process images, it segments an image into non-overlapping patches and creates patch embeddings. The embeddings serve as input for the model with an additional class token that is used for classification at the final layer. The model is pre-trained on ILSVRC-2012 ImageNet dataset with 1k classes and 1.3M images (Deng et al., 2009).

DeiT3 An abbreviation for the data-efficient image transformer, the DeiT solves the dependence of the ViT on large-scale datasets for pretraining (Touvron et al., 2021). The model is trained by a CNN in a student-teacher distillation strategy and therefore DeiT belongs to the group of CNN-enhanced transformers. The CNN can transfer its inductive bias to the transformer, which leads to the student CNN outperforming its CNN teacher. Except for the smallest version, all three sizes of the model surpass ViT-Base on ImageNet-1K.

DinoV2 DINOv2 (Oquab et al., 2023) further improves on its predecessor DINO (Caron et al., 2021) which took inspiration from the student-teacher distillation approach of DeiT. By combining the DINO and iBOT loss and adding a regularizer, DINOv2 achieves excellent performance not only in image classification but also in image segmentation. Most importantly, it achieves the result in a self-supervised learning approach and thus solves the dependency on labeled image data. DINOv2 comes in four sizes and trains the original ViT architecture with the LVD-142M dataset.

B Continual Learning methods

B.1 Prompt-based methods

L2P Learning to Prompt (L2P) prepends tunable prompts to the input that share the inputs' embedding dimensions (Wang et al., 2022b). During training time, the models' parameters are frozen and the prompt weights are adjusted. L2P does not prepend the same prompt for samples of all tasks but designs a pool in which a prompt is adjusted for each task individually. In addition, during training, a learnable prompt selection strategy enables the model to select the prompt for the correct class during test time.

DualPrompt DualPrompt extends L2P by attaching prompts to multiple layers (Wang et al., 2022a). Moreover, DualPrompt introduces two sets of prompts: E(xpert)-Prompts for each distinct class and G(eneral)-Prompts to learn generalized features across classes. During testing, a query function selects the suitable E-Prompt for an input belonging to a certain class. The model then predicts the correct label of the selected class with the attached E-Prompts and G-Prompts.

CODA-Prompt DualPrompt and L2P both face the bottleneck of the optimal prompt selection process when it comes to performance improvement. CODA-Prompt addresses this challenge by learning a set of prompt components that are combined with an attention-based mechanism to a final prompt during test time (Smith et al., 2023). During training, all prompt components are adjusted contrary to DualPrompt and L2P where only the prompt that matches the task is tuned.

B.2 Adapter-based methods

ADAM-Adapter ADAM approaches the problem of CL with PTMs differently. It uses a parameter-efficient finetuning technique such as adapter tuning (ADAM-Adapter) to train the PTM on the first task and bridge the domain gap between the pre-trained model and task at hand (Zhou et al., 2023). ADAM therefore does not require training on all tasks of a continual learning scenario. During testing, it then combines the features of the PTM and adapted PTM to extract a prototype that is used as input for a classifier based on cosine similarity.

EASE EASE further extends ADAM by iteratively adapting PTMs to all tasks (Zhou et al., 2024b). To accommodate for the expanding feature space of the adapted PTMs, EASE designs a mapping between the old and new classifiers.

RanPAC Building on ADAM, RanPAC identifies that prototypes often exhibit correlations between classes (McDonnell et al., 2024). To address this, it recommends using an online LDA classifier to eliminate these class-wise correlations to improve separability. Additionally, to better fit a Gaussian distribution, RanPac introduces a random projection layer that projects features into a higher-dimensional space. This approach allows for more accurate computation of prototypes in the transformed space.