Continual Learning of Foundation Models with Limited Labelled Data

Shuvendu Roy[1,2] Elham Dolatabadi[1,3], Arash Afkanpour[1], Ali Etemad[2]
[1]Vector Institute, [2]Queen's University, Canada, [3]York University, Canada
shuvendu.roy@queensu.ca, edolatab@yorku.ca,
arash.afkanpour@vectorinstitute.ai, ali.etemad@queensu.ca

Abstract

We explore a new paradigm of continual learning dubbed Few-Shot Class-Incremental Tuning (FSCIT), which facilitates continual tuning of vision foundation models to continuously learn new classes with few samples per class. Unlike traditional Few-Shot Class-Incremental Learning (FSCIL), FSCIT does not assume the availability of a large in-distribution base session to initially train the model in a fully supervised setting, prior to the few-shot class-incremental sessions. To this end, we propose Consistency-guided Asynchronous Contrastive Tuning (CoACT), a new approach to continually tune foundation models for new classes in few-shot settings. CoACT comprises three components: (*i*) asynchronous contrastive tuning, which learns new classes by including LoRA modules in the pre-trained encoder while enforcing consistency between two asynchronous encoders; (*ii*) controlled fine-tuning, which facilitates effective tuning of a subset of the foundation model; and (*iii*) consistency-guided incremental tuning, which enforces additional regularization during later sessions to reduce forgetting of the learned classes. We perform an extensive study on 16 diverse datasets where CoACT outperforms the best baseline method by 2.47% on average and with up to 12.52% on individual datasets. Additionally, CoACT shows reduced forgetting and robustness in low-shot experiments. As an added bonus, CoACT outperforms current SOTA on FSCIL.

1 Introduction

Large foundation models pre-trained on web-scale unlabeled data demonstrate robust generalization on downstream tasks when fine-tuned with a relatively small amount of labelled data [1, 2]. However, the sheer size of the pre-trained models poses significant challenges for fine-tuning such models, particularly when learning from limited labelled data [3]. Despite recent advancements such as parameter-efficient tuning [4, 5] and regularization [3], fine-tuning the model in a few-shot setting often leads to a decline in the generalization capability of the foundation model. Yet, real-world applications not only necessitate learning from a few samples but also demand continual learning of new classes without forgetting previously learned ones. To this end, we introduce a novel continual learning paradigm, **Few-Shot Class-Incremental Tuning (FSCIT)**, which enables tuning a vision foundation model to continuously learn new classes with a few samples per class.

In continual learning, Few-shot Class-Incremental Learning (FSCIL) is a closely related area (different from conventional class-incremental learning [6]) that first trains the model on a *large in-distribution* labelled base session in a fully supervised setting, followed by few-shot learning of new classes over incremental sessions [7, 8]. In practice, such an in-distribution base session with a large number of classes and a large number of samples per class is difficult to collect, undermining the motivation of few-shot continual learning. Unlike FSCIL, FSCIT focuses on tuning vision foundation models in a few-shot class-incremental setup, without assuming the availability of an initial in-distribution base session. Due to the unavailability of the in-distribution base session, existing methods from the FSCIL literature can not be directly adopted to FSCIT. Additionally, few-shot tuning methods [9, 3] for foundation models are also not applicable to FSCIT since they do not facilitate **continual** tuning.

To address this challenge, we first establish a few baselines by combining the prominent approaches from FSCIL and few-shot tuning literature to facilitate FSCIT. However, these baselines fall short of

NeurIPS 2024 Workshop on Scalable Continual Learning for Lifelong Foundation Models.

addressing the critical challenge of preserving the generalizability of the foundation model and preventing catastrophic forgetting of previously learned classes. Next, we propose Consistency-guided Asynchronous Contrastive Tuning (CoACT), a novel framework for class-incremental tuning of vision foundation models in few-shot settings. CoACT comprises three components: (**i**) asynchronous contrastive tuning, (**ii**) controlled fine-tuning, and (**iii**) consistency-guided incremental tuning. **Asynchronous contrastive tuning** learns from the first incremental session using a novel asynchronous contrastive loss that strikes a balance between adaptability to learn new classes and retaining generalizable knowledge of the pre-trained foundation model. To further enhance adaptability, we introduce **controlled fine-tuning**, which is a two-step training protocol for training the first incremental session. Finally, **consistency-guided incremental tuning** is a novel regularization technique that ensures effective learning of classes in the following incremental sessions while preventing forgetting of previously learned classes and preserving the generalization capabilities of the foundation model at the same time. To achieve this, we enforce consistency between the predictions of the learnable encoder in the incremental sessions and the frozen encoder from the first incremental session.

We conduct a comprehensive study on 16 diverse image recognition datasets to investigate the effectiveness of our method. Our comprehensive experiments demonstrate that CoACT achieves a 2.47% average improvement over the best baseline method, with up to 12.79% performance gain on individual datasets. More importantly, CoACT exhibits reduced forgetting of already learned classes as the number of classes increases. We provide detailed ablation studies showing the effectiveness of each component of our method. Additionally, CoACT outperforms current SOTA on FSCIL.

2 Related works

Class-incremental learning is a continual learning process that focuses on continuous learning of new classes while retaining the knowledge of already learned ones [6]. In practice, machine learning models often need to learn new classes from a few labelled samples per class [7], while having no access to samples from already learned classes. This scenario has given rise to a new learning task called few-shot class-incremental learning or FSCIL [10]. The existing literature on FSCIL can be broadly categorized into two main groups: methods that continuously train both the encoder and classifier over each incremental session [8, 11, 12, 13], and methods that keep the encoder frozen during the incremental learning sessions [14, 15, 16]. While methods in the first group generally offer greater *adaptability* to new classes compared to the second group, they require additional constraints to avoid overfitting to the new classes and thus catastrophic forgetting, and they are not generally feasible in the context of foundation models as encoders are quite large, resulting in strong overfitting in few-shot settings. While methods in the second group generally offer greater *stability* on already learned classes stability often comes at the cost of adaptability toward learning new classes, and these methods are generally not compatible with foundation models given lack of control over the base training of such off-the-shelf models. A number of techniques have recently been proposed to tune foundation models without the need to re-train them from scratch [17, 18, 19, 20, 21, 22, 23]. Nonetheless, such methods are not designed for *continuous* tuning of a foundation model since there are no mechanisms in them to prevent loss of generalization capabilities and catastrophic forgetting.

3 Method

Asynchronous contrastive tuning. To strike a balance between adaptability to new classes and retaining generalizable knowledge of the foundation model, we introduce asynchronous contrastive tuning as the first component in our framework. This involves fine-tuning the pre-trained model using our novel Asynchronous Contrastive Learning (ACL) approach while incorporating LoRA modules into the foumodel. Let $h_i = f_\theta^{(i)}(h_{i-1})$ be the output of i^{th} layer of the pre-trained encoder, h_{i-1} be the output of the $(i-1)^{\text{th}}$ hidden layer of the encoder, and $f_\theta^{(i)}$ be the i^{th} layer of the encoder. With the new learnable LoRA layers, the

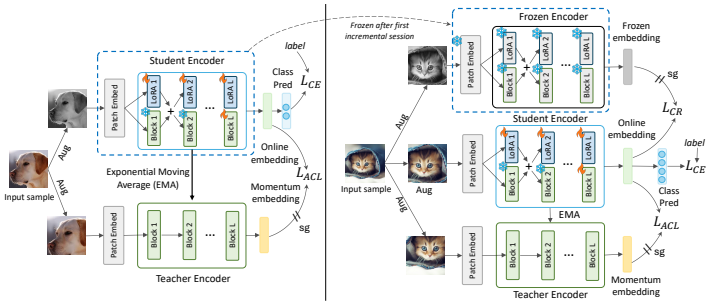

Figure 1: Illustration of CoACT. (Left) Training on the first incremental session with asynchronous contrastive tuning and controlled fine-tuning. (Right) Consistency-guided incremental tuning over continual sessions.

output of the i^{th} layer of the network can be represented as $h_i^{'} = f_\theta^{(i)}(h_{i-1}) + f_{\text{LoRA}}^{(i)}(h_{i-1})$, where, $f_{\text{LoRA}}^{(i)}$ is the i^{th} LoRA layer added to the pre-trained encoder. For brevity, we denote the encoder with learnable LoRA layers as $f_{\theta'}$. We can train $f_{\theta'}$ on $\mathcal{D}_{train}^1$ to learn the first incremental session as: $\mathcal{L}_{sup} = \mathcal{L}_{ce}(W^T f_{\theta'}(x), y)$. However, it has been shown in prior work that cross-entropy alone does not learn a well-separable embedding space [24] and has a higher over-fitting tendency, especially in a few-shot setting [3]. To reduce the possibility of overfitting and retain the generalization in the learnable encoder $f_{\theta'}$, we regularize its output distribution with a teacher encoder by maximizing their agreement in the embedding space. Here, the teacher encoder is identical to the pre-trained model (does not contain the LoRA modules) and learned through the Exponential Moving Average (EMA) of the student $f_{\theta'}$ as $\theta^{''} = m \cdot \theta^{''} + (1 - m) \cdot \theta$, where θ and $\theta^{''}$ are the parameters of the student (without LoRA) and teacher encoders, and m is the momentum parameter. Given that the teacher and student encoders differ in their architecture due to the addition of LoRA to the student, learning occurs asynchronously. This asynchronous encoder design and slow-moving update of the teacher through EMA ensures that the predictions from the teacher do not fluctuate. Since the teacher encoder is also initialized from the foundation model, consistency with the teacher effectively regularizes the student from overfitting. In practice, we maximize the agreement between the embeddings of the student and teacher encoders on all the samples from each class using a supervised contrastive loss:

$$\mathcal{L}_{ACL} = -\sum_i \frac{1}{|C_i|} \sum_{j \in C_i} \log \frac{\exp(\langle q_i, k_j \rangle / \tau)}{\sum_{l \neq i} \exp(\langle q_i, k_l \rangle / \tau)}, \tag{1}$$

where $C_i \stackrel{\text{def}}{=} \{j : y_j = y_i\}$, $\langle \cdot, \cdot \rangle$ denotes inner product, $q_i = f_{\theta'}(\mathcal{A}_1(x_i))$ and $k_j = f_{\theta''}(\mathcal{A}_2(x_j))$ are online embeddings and momentum embeddings of augmentations of x_i and x_j from the student and the teacher encoder respectively, and $\mathcal{A}_1$ and $\mathcal{A}_2$ are random augmentations. Finally, we train the model with the $\mathcal{L}_{ACL}$ and $\mathcal{L}_{sup}$ as: $\mathcal{L}_{ACL} + \lambda \cdot \mathcal{L}_{sup}$, where λ controls the impact of the ACL loss.

Controlled fine-tuning. To further enhance the adaptability of the model, we enable controlled fine-tuning of some of the pre-trained layers of the encoder. Since the newly added LoRA modules are randomly initialized, we begin by training only the LoRA modules of the student encoder with a higher learning rate for the initial E_c epochs of training. This is followed by a fine-tuning stage where the last C_l layers of the pre-trained encoder are fine-tuned with a reduced LR (scaled by a factor of C_f). We focus on fine-tuning only the last C_l layers, as the later layers of a pre-trained model are responsible for learning domain-specific fine-grained features, whereas the earlier layers are more general and transferable to a wide range of tasks [25].

Consistency-guided incremental tuning. While the first two modules facilitate tuning the foundation model ($f_{\theta'}$) during the first session, later sessions also require the retention of previously learned classes. To facilitate this, we propose consistency-guided incremental tuning, which prevents forgetting by regularizing the output distribution of the student $f_{\theta'}$ when training on the incremental sessions. More specifically, we enforce consistency between the predictions of the student encoder and a frozen encoder after the first session, effectively discouraging substantial changes in the learned representations of the student. Let f_{θ_β} be the frozen encoder after the first session, and the frozen embedding of this encoder be $p_i = f_{\theta_\beta}(\mathcal{A}(x_i))$. We define our consistency regularizer as:

$$\mathcal{L}_{CR} = -\sum_i \frac{1}{|C_i|} \sum_{j \in C_i} \log \frac{\exp(\langle q_i, p_j \rangle / \tau)}{\sum_{l \neq i} \exp(\langle q_i, p_l \rangle / \tau)}. \tag{2}$$

Finally, we train the model after the first session with: $\mathcal{L}_{CR} + \gamma \mathcal{L}_{ACL} + \lambda \mathcal{L}_{sup}$, where γ and λ controls the relative importance of the loss functions.

4 Experiments

We present the main results on CoACT on FSCIT in Table 1 on 16 datasets using three different encoder backbones. With ViT-B/16 as the backbone, while CoACT shows average improvements of 2.66%, 2.54%, and 2.47% improvement over prototype learning, linear tuning, and LoRA, respectively. Notably, it shows up to a 12.79% improvement on some individual datasets such as Resisc-45. In general, we find relatively higher improvements for more challenging datasets. For instance, on the 5 datasets with the least accuracy (Country211, FGVCAircraft, GTSRB, Resisc-45, and StanfordCars), CoACT shows an average improvement of 6.14% over the best baseline method (LoRA). We observe similar improvements in the performance with the other two encoders (ViT-B/32 and ViT-L/16), where CoACT outperforms the best baseline (LoRA) by average accuracies of 2.08% and 1.67%.

Next, we report the forgetting of learned classes for each method, measured as the drop in accuracy w.r.t. the first session, in Figure 2 (left). As we observe, CoACT has the least amount of forgetting compared to the baselines with approximately 1.5% less forgetting than both LoRA and linear tuning, and 3.2% than prototype learning. We also present a breakdown of accuracies into the first incremental session, the remaining incremental sessions, and all sessions in Figure 2 (right). Here, for all methods, the first session shows particularly higher accuracy than other sessions since there is no interference (or forgetting) of other classes. While LoRA and linear tuning have higher overall accuracy than prototype tuning, the improvement mainly comes from the higher accuracy in the first session only. All baselines, however, perform relatively similarly in the remaining sessions. In contrast, CoACT shows higher improvement in all sessions.

Table 1: Performance of CoACT on 16 datasets and its comparison to the baselines.

Method	*Encoder*	*Caltech101*	*CIFAR100*	*Country211*	*CUB200*	*DTD*	*EuroSat*	*FGVC*	*Food101*	*GTSRB*	*MiniIN*	*Flowers102*	*OxfordPets*	*Resisc-45*	*Cars*	*SUN397*	*VOC 2007*	*Average*
Pro. lear.	ViT-B/16	83.47	75.58	7.42	74.10	54.72	62.20	15.51	71.09	15.53	95.31	98.09	88.10	49.83	18.90	65.68	64.72	58.77
	ViT-B/32	83.90	74.29	6.79	69.32	52.83	68.91	13.50	63.57	19.84	92.76	97.09	84.57	45.30	16.85	61.29	66.64	57.34
	ViT-L/16	85.10	75.04	7.23	76.31	54.33	65.48	19.03	74.66	12.24	98.05	98.82	87.64	54.85	24.69	66.58	60.16	60.01
Lin. tun.	ViT-B/16	83.60	75.58	7.45	74.41	54.72	62.20	15.32	71.18	16.92	95.32	98.09	88.10	49.83	18.96	65.87	64.72	58.89
	ViT-B/32	83.87	74.35	6.75	69.84	53.22	68.89	11.29	63.51	26.28	92.75	97.64	85.06	51.17	16.57	61.36	66.58	58.07
	ViT-L/16	85.25	75.33	7.26	76.11	54.24	65.32	19.44	74.51	13.01	97.61	98.34	87.34	54.99	25.21	66.99	61.01	60.12
LoRA	ViT-B/16	83.87	75.81	7.45	74.65	54.70	62.20	15.32	71.24	17.23	95.32	98.09	88.10	49.82	18.94	65.90	64.72	58.96
	ViT-B/32	83.54	74.72	6.81	69.85	53.22	68.89	10.88	63.52	29.84	92.75	97.55	85.09	49.14	16.09	61.32	66.56	58.11
	ViT-L/16	85.15	75.34	7.29	76.26	54.52	65.68	19.13	74.75	12.28	97.88	98.71	87.89	54.85	24.99	66.78	60.56	60.13
CoACT	ViT-B/16	86.86	78.31	7.42	77.38	55.11	62.25	18.98	71.59	26.05	95.34	98.18	87.79	62.62	24.40	66.11	64.45	61.43
	ViT-B/32	85.59	77.26	6.85	72.21	54.78	67.71	14.56	68.41	29.99	94.81	98.01	86.21	55.34	20.18	64.49	66.71	60.19
	ViT-L/16	87.25	79.27	7.45	78.29	56.19	66.36	21.24	74.99	17.22	97.18	98.18	87.95	62.25	25.43	67.18	62.23	61.79

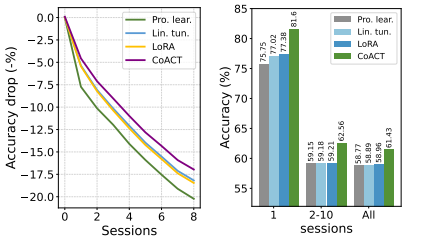

Figure 2: (left) Forgetting of learned classes. (right) Accuracy breakdown over sessions.

To further evaluate the effectiveness of our method, we also investigate its performance in the traditional training setup of FSCIL. In Table 2, we compare the performance of CoACT with prior works on the CIAFR-100 dataset, where we divide the existing methods into two groups: the first group, which trains randomly initialized models and the second group which use a pre-trained encoder. As we observe in the table, CoACT outperforms the previous SOTA without a pre-trained encoder (BOT [27]) by 25.88% and the SOTA with a pre-trained encoder (CPE-CLIP [29]) by 4.11%. For a fair comparison with existing methods, we evaluate SAVC [24] and BOT [27] with the same pre-trained encoder (ViT-B/16) as ours, where both methods show

Table 2: Comparison to prior works on FSCIL on CIFAR-100.

Method	**Pre-trained**	**Acc. in each session (%) ↑**								
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
SAVC [24]	✗	79.85	73.70	69.37	65.28	61.91	59.27	57.24	54.97	53.12
SoftNet [26]	✗	79.88	75.54	71.64	67.47	64.45	61.09	59.07	57.29	55.33
BOT [27]	✗	80.25	77.20	75.09	70.82	67.83	64.86	62.73	60.52	58.75
SAVC* [24]	ViT-B/16	82.98	75.35	74.01	71.51	70.64	69.78	68.98	67.84	66.24
BOT* [27]	ViT-B/16	83.75	78.14	76.85	73.23	72.95	72.03	71.56	70.66	69.72
SV-T [28]	SwinT	86.77	82.82	80.36	77.20	76.06	74.00	72.92	71.68	69.75
CPE-CLIP [29]	CLIP-B/16	87.83	85.86	84.93	82.85	82.64	82.42	82.27	81.44	80.52
CoACT (Ours)	ViT-B/16	90.46	88.46	88.11	86.94	86.98	86.52	86.39	86.0	84.63

Ablation study. We present an ablation study on the proposed components of CoACT in Table 3, where ACT, C.F., and C.I.T stand for asynchronous contrastive tuning, controlled fine-tuning, and consistency-guided incremental tuning, respectively. Given that the asynchronous contrastive tuning component of our method could not be removed as it contains the trainable parameters, we start this study by removing controlled fine-tuning and consistency-guided incremental tuning modules individually and simultaneously. Interestingly, we observe that while individual removal of these components does not show considerable drops in performance, their concurrent application within our framework results in a significant boost in performance of 1.17% across 16 datasets. Finally, with the ablation of all three components and only training a linear classifier, we observe a 2.66% drop in performance.

Table 3: Ablation study.

ACT	**C.F.**	**C.I.T**	**Acc**
✓	✓	✓	61.43
✓	✗	✓	61.31
✓	✓	✗	61.28
✓	✗	✗	60.26
✗	✗	✗	58.77

5 Conclusion

To enable few-shot class-incremental learning with pre-trained large vision models, we propose CoACT. Our method can effectively tune a foundation model to learn new classes without losing the generalization of the pre-training or forgetting previously learned classes. Extensive studies show the effectiveness of our method, achieving higher accuracy, lower forgetting, and robustness in low-shot settings. CoACT also outperforms prior SOTA in the standard FSCIL setup by a large margin. We present comprehensive experiments on different components of CoACT and make our code available to foster rapid developments in the area.

References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021. 1

[2] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 1

[3] Shuvendu Roy and Ali Etemad. Consistency-guided prompt learning for vision-language models. In *International Conference on Learning Representations*, 2024. 1, 3

[4] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. 1

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

[6] Marc Masana, Xialei Liu, Bartłomiej Twardowski, Mikel Menta, Andrew D Bagdanov, and Joost Van De Weijer. Class-incremental learning: survey and performance evaluation on image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5513–5533, 2022. 1, 2

[7] Da-Wei Zhou, Han-Jia Ye, Liang Ma, Di Xie, Shiliang Pu, and De-Chuan Zhan. Few-shot class-incremental learning by sampling multi-phase tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1, 2

[8] Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong. Few-shot class-incremental learning. In *TPAMI*, pages 12183–12192, 2020. 1, 2

[9] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 2022. 1

[10] Songsong Tian, Lusi Li, Weijun Li, Hang Ran, Xin Ning, and Prayag Tiwari. A survey on few-shot class-incremental learning. *Neural Networks*, 169:307–324, 2023. 2

[11] Ali Cheraghian, Shafin Rahman, Pengfei Fang, Soumava Kumar Roy, Lars Petersson, and Mehrtash Harandi. Semantic-aware knowledge distillation for few-shot class-incremental learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2534–2543, 2021. 2

[12] Songlin Dong, Xiaopeng Hong, Xiaoyu Tao, Xinyuan Chang, Xing Wei, and Yihong Gong. Few-shot class-incremental learning via relation knowledge distillation. In *Association for the Advancement of Artificial Intelligence*, pages 1255–1263, 2021. 2

[13] Hanbin Zhao, Yongjian Fu, Mintong Kang, Qi Tian, Fei Wu, and Xi Li. Mgsvf: Multi-grained slow vs. fast framework for few-shot class-incremental learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[14] Kai Zhu, Yang Cao, Wei Zhai, Jie Cheng, and Zheng-Jun Zha. Self-promoted prototype refinement for few-shot class-incremental learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6801–6810, 2021. 2

[15] Guangyuan Shi, Jiaxin Chen, Wenlong Zhang, Li-Ming Zhan, and Xiao-Ming Wu. Overcoming catastrophic forgetting in incremental few-shot learning by finding flat minima. In *Advances in Neural Information Processing Systems*, pages 6747–6761, 2021. 2

[16] Chi Zhang, Nan Song, Guosheng Lin, Yun Zheng, Pan Pan, and Yinghui Xu. Few-shot incremental learning with continually evolved classifiers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12455–12464, 2021. 2

[17] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Learning Representations*, pages 2790–2799. PMLR, 2019. 2

[18] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, 2021. 2

[19] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *International Joint Conference on Natural Language Processing*, pages 4582–4597, 2021. 2

[20] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021. 2

[21] Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. Compacter: Efficient low-rank hypercomplex adapter layers. *Advances in Neural Information Processing Systems*, 34:1022–1035, 2021. 2

[22] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727, 2022. 2

[23] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022. 2

[24] Zeyin Song, Yifan Zhao, Yujun Shi, Peixi Peng, Li Yuan, and Yonghong Tian. Learning with fantasy: Semantic-aware virtual contrastive constraint for few-shot class-incremental learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24183–24192, 2023. 3, 4

[25] Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? *Advances in Neural Information Processing Systems*, 33:512–523, 2020. 3

[26] Haeyong Kang, Jaehong Yoon, Sultan Rizky Hikmawan Madjid, Sung Ju Hwang, and Chang D Yoo. On the soft-subnetwork for few-shot class incremental learning. In *International Conference on Learning Representations*, 2023. 4

[27] Shuvendu Roy, Chunjong Park, Aldi Fahrezi, and Ali Etemad. A bag of tricks for few-shot class-incremental learning. *arXiv preprint arXiv:2403.14392*, 2024. 4

[28] Wenhao Qiu, Sichao Fu, Jingyi Zhang, Chengxiang Lei, and Qinmu Peng. Semantic-visual guided transformer for few-shot class-incremental learning. In *IEEE International Conference on Multimedia and Expo*, pages 2885–2890, 2023. 4

[29] Marco D'Alessandro, Alberto Alonso, Enrique Calabrés, and Mikel Galar. Multimodal parameter-efficient few-shot class incremental learning. In *IEEE/CVF International Conference on Computer Vision*, pages 3393–3403, 2023. 4