Quantified Task Misalignment to Inform PEFT: An Exploration of Domain Generalization and Catastrophic Forgetting in CLIP

Laura Niss[*], **Kevin Vogt-Lowell**[*], **Theodoros Tsiligkaridis**
Artificial Intelligence Technology Group, MIT Lincoln Laboratory
{laura.niss, kevin.vogt-lowell, ttsili}@ll.mit.edu

Abstract

Foundations models are presented as generalists that often perform well over a myriad of tasks. Fine-tuning these models, even on limited data, provides an additional boost in task-specific performance but often at the cost of their wider generalization, an effect termed catastrophic forgetting. In this paper, we analyze the relation between zero-shot text and image embedding alignment in the CLIP model and the performance of several simple parameter-efficient fine-tuning methods through the lens of domain generalization and catastrophic forgetting. We provide evidence that the silhouette score of the zero-shot image and text embeddings is a better measure of improvement gain from fine-tuning than the average cosine similarity of correct image/label embeddings, and discuss empirical relationships between zero-shot embedding alignments, fine-tuning method, domain generalization, and catastrophic forgetting. Additionally, the averaged results across tasks and performance measures demonstrate that a simplified method that trains only a subset of attention weights, which we call A-CLIP, provides a good balance between domain generalization and catastrophic forgetting.

1 Introduction

Motivation Contrastive Language-Image Pretraining (CLIP) (Radford et al., 2021) has demonstrated high zero-shot performance on average over many disparate tasks, yet many applications still predominantly use it for specific downstream tasks that require further fine-tuning to obtain optimal performance. The fine-tuning training method also has an impact on domain generalization performance (Vogt-Lowell et al., 2023). When fine-tuning the original model weights, varying degrees of catastrophic forgetting often occur, partially due to changes in the alignment between the text and image embedding spaces. This misalignment causes a decrease in model performance (Ding et al., 2022; Ni et al., 2023). Several parameter-efficient fine-tuning (PEFT) methods have been shown to improve either in-domain (ID) or out-of-domain (OOD) cross-dataset performance (Touvron et al., 2022; Liao et al., 2023). With regard to multi-modal models, we speculate that there exists a relation between the alignment of the zero-shot embeddings, catastrophic forgetting, and domain generalization when considering such PEFT methods of fine-tuning. Methods that attenuate catastrophic forgetting could result in poor generalization if the method is too 'surgical' in its learning, whereas methods that produce strong generalization on a specific task may distort other parts of the embedding space and trigger catastrophic forgetting. Motivated by this hypothesis, we consider five training datasets with two testing schema through which the performance and embedding space shifts of four PEFT methods were analyzed. To measure shifts in the embedding space and quantify zero-shot alignment, we consider both the average cosine similarity between image and correct caption embeddings and the silhouette score, a measure of overlap between two clusters, between all image and text embeddings.

Contributions To the best of our knowledge, we are the first to provide evidence that, with respect to CLIP: 1) the silhouette score of the zero-shot text and image embeddings can be used as a measure of available performance improvement; less initial overlap correlates with larger gains in performance

[*] Equal contribution

after fine-tuning. 2) BitFit is susceptible to catastrophic forgetting when the target task has a large zero-shot silhouette score; and 3) under fitting LoRA by using a low rank can significantly minimize catastrophic forgetting by restricting most changes in the embedding space to the area associated with the target task, though this precision comes at the cost of performance gains relative to other methods in terms of domain generalization and in-domain accuracy.

Additionally, we observe that our attention-based fine-tuning method, A-CLIP (a further refinement of Touvron et al. (2022)), yields the best balance of attenuating catastrophic forgetting while maintaining or improving performance in terms of domain generalization.

2 Related Work

Catastrophic forgetting in neural networks describes the significant drop in performance on previous tasks that accompanies increasing amounts of training. While this loss of knowledge has been proven to significantly affect traditional supervised learning techniques (McCloskey & Cohen, 1989; Kirkpatrick et al., 2017), recent research has shown that self-supervised, uni-modal models do not suffer from such severe forgetting (Ni et al., 2021; Hu et al., 2022a). Yet, with multi-modal foundation models, such as CLIP, continuous training can cause misalignment of the image and text embeddings, an outcome highlighted and explored in several works (Ding et al., 2022; Lin et al., 2023; Ni et al., 2023; Srinivasan et al., 2022; Fan et al., 2022).

PEFT methods could be considered among the parameter-isolation techniques explored in some catastrophic forgetting literature, and as such have the potential to attenuate catastrophic forgetting (Aljundi et al., 2018; Laborieux et al., 2021; Kirkpatrick et al., 2017). Motivated by the lack of research exploring the interplay between the misalignment problem and the possible attenuating effects of PEFT on catastrophic forgetting, we investigate the relationships between these fine-tuning methods, zero-shot task alignment, domain generalization, and catastrophic forgetting.

3 Methodology

Model Selection Our simple PEFT method, A-CLIP, freezes all of the weights in CLIP except for the in-projection weights of the attention layers, reducing the number of trainable weights even further than Touvron et al. (2022). We chose this method based on observations that with some learning parameters, the in-projection weights of CLIP's attention layers changed significantly more than the out-projection layers during fine-tuning. Further discussion can be found in Appendix A.4.

In addition to A-CLIP, we selected three SOTA PEFT methods for comparison, including CLIP-Adapter (Gao et al., 2023), LoRA (Hu et al., 2022b), and BitFit (bias-tuning) (Ben Zaken et al., 2022). All methods use CLIP ViT-B/32 provided by OpenCLIP (Cherti et al., 2023) as the model backbone and were also compared to a standard CLIP ViT-B/32 model fine-tuned end-to-end (full) using cross-entropy loss.

Alignment Measures We consider two measures of alignment: the average cosine similarity (labeled as cosine in figures) of image embeddings and their correct caption embeddings, and the silhouette score (ss) of the image and text embeddings. The silhouette score represents an average measure of the distance among points in their own cluster compared to that of the next nearest neighbor. A value of 0 implies greater cluster overlap and a value of 1 implies greater cluster separation. Formulas for both measures are found in Appendix A.3.

4 Results

Datasets To investigate zero-shot task alignment and model performance across a variety of domain shift magnitudes, we train our models on five diverse datasets divided into two cross-dataset evaluation schema: one with presumed small domain shifts to evaluate domain generalization (DG) and another with presumed large domain shifts to evaluate catastrophic forgetting (CF). For domain generalization, we use two training sets: 1) 16-shot ImageNet (Deng et al., 2009) evaluated on ImageNet-V2, ImageNet-Sketch, ImageNet-Rendition, and ImageNet-Adversarial, and 2) the in-distribution subset of the Functional Map of the World (FMoW-ID) satellite dataset (Christie et al.,

2018) from WILDS (Koh et al., 2021) evaluated on the satellite imagery benchmarks FMoW-OOD, RESISC45, and EuroSAT. For catastrophic forgetting, we selected three datasets on which zero-shot CLIP has exhibited extreme performances: 1) Stanford Cars (Krause et al., 2013), on which zero-shot CLIP outperforms ResNet with a linear probe, 2) GTSRB (Stallkamp et al., 2012) which underperforms ResNet with a linear probe, and 3) SVHN (Netzer et al., 2011) on which zero-shot CLIP also performs poorly. These datasets are evaluated with additional common vision benchmarks, namely CIFAR100 (Krizhevsky et al., 2009), DTD (Cimpoi et al., 2014), MNIST (LeCun et al., 1998), STL10 (Coates et al., 2011), and SUN397 (Xiao et al., 2010).

Experimental Setting Training for each method involved 40 epochs of learning using mini-batch stochastic gradient descent. The model checkpoint with the highest validation accuracy during training was used as the final model, and reported accuracies were calculated using held-out test data. Given that our interest lies in comparing models across task settings and evaluating their general performance rather than comparing state-of-the-art performances between PEFT methods, we opted to borrow several hyper-parameter selections made in Liao et al. (2023). Training details for each model can be found in Appendix A.1.

Model Performance to Evaluation Task In Figure 1, we have in-domain and out-of-domain results with respect to the catastrophic forgetting and generalizations schemas. All plots show changes in accuracy relative to zero-shot CLIP. The top plots giving in-domain (ID) accuracy show that across training and testing schema, full fine-tuning, A-CLIP, and BitFit achieved the highest accuracies, LoRA the lowest, and CLIP-Adapter varying in between. The bottom plots in Figure 1 highlight that in the catastrophic forgetting schema, BitFit has the worst out-of-domain performance and A-CLIP appears to have an advantage over full fine-tuning, especially for tasks with large zero-shot silhouette scores. LoRA performs the best with respect to catastrophic forgetting, though this is most likely an artifact of parameter choice resulting in poor in-domain performance gains. Further analysis of alignment changes suggests LoRA models with better in-domain performance would also induce catastrophic forgetting (Appendix A.6). Looking at the right hand plot for evaluating cross-dataset domain generalization, the best accuracy is achieved with full fine-tuning followed by A-CLIP.

From these results we highlight that with respect to catastrophic forgetting, A-CLIP balances in-domain performance gains with attenuating catastrophic forgetting, and with respect to domain generalization achieves performance near full fine-tuning while updating only a fraction of the parameters.

Additionally, Figure 1 gives empirical evidence for the positive correlation between the zero-shot silhouette score and the increase in in-domain performance with fine-tuning.

Silhouette Score, Accuracy, and Improvement To further test whether the zero-shot silhouette score holds empirically as an approximate measure of accuracy gain, we compare the zero-shot silhouette score and average cosine similarity to the change in accuracy in Figure 2. These results are from for all models and tasks tested on in-domain data, excluding results trained on 16-shot ImageNet or with LoRA. These exclusions were made because few-shot learning is an inherently different setting from the other tasks, and because our limited hyperparameter search resulted in LoRA training poorly on many of the ID tasks. We included six additional fully fine-tuned models from six datasets for more power: CIFAR100, DTD, EuroSAT, MNIST, RESISC45, and SUN397. The regression on the silhouette results in an R^2 value of 0.5, and is significant at $\alpha = 0.0003$. For the average cosine similarity R^2 is 0.07 and is significant at level $\alpha = 0.21$. These empirical results give evidence that the silhouette score is much more strongly correlated with in-domain performance gains with fine-tuning than the average cosine similarity.

In Figure 3, we explore the relationship between the change in measure and the change in accuracy between the zero-shot and fine-tuned models. Both the change in silhouette score and average cosine similarity have significant linear relationships to the change in in-domain accuracy, though the silhouette score has a stronger correlation than the average cosine similarity in in-domain, $R^2 = 0.67$ vs 0.21. This relationship is reversed for the out-of-domain measures ($R^2 = 0.2$ for ss and 0.42 for cosine). We include results and a discussion in Appendix A.5 giving evidence that neither measure is a good proxy for accuracy.

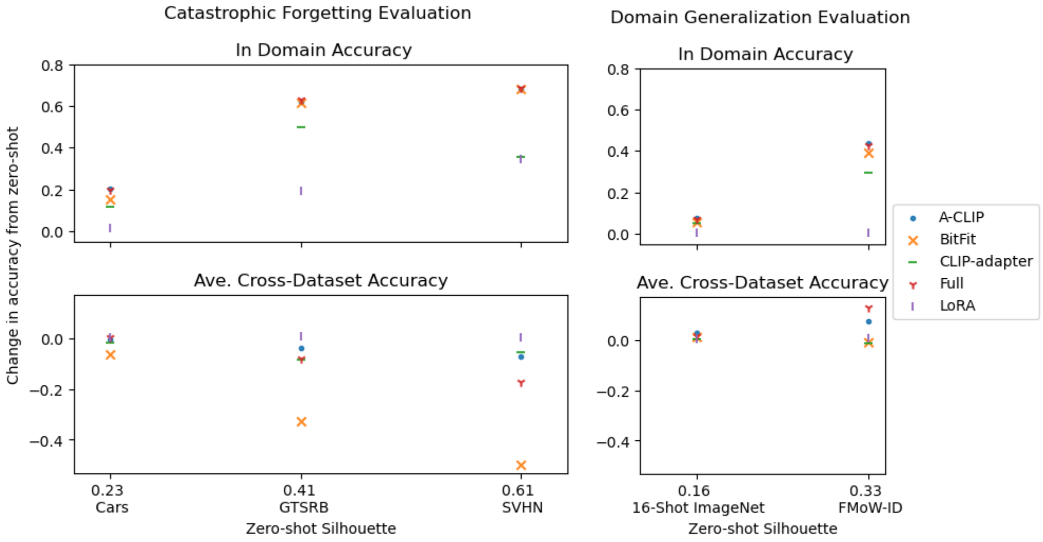


Figure 1: Silhouette score and average change in accuracy with fine-tuning; top plots show in-domain change in accuracy and bottom plots show out-of-domain cross-datasets change in accuracy. The left plots reflect the in-domain and out-of-domain performance within our catastrophic forgetting schema. The right plots reflect this performance with respect to our domain generalization schema. Cross-dataset accuracy for FMoW-ID averages over 3 other satellite datasets, 16-Shot ImageNet average over 4 ImageNet variants, and Cars, GTSRB, and SVHN average over 7 unrelated datasets.

5 Conclusion

In this paper, we gave evidence using CLIP that the silhouette score of the zero-shot text and image embeddings can be used as an approximation of available improvement gain from fine-tuning. We also provided evidence that several simple and well-known fine-tuning methods have different desirable and undesirable traits with respect to generalization and catastrophic forgetting, and that there is a positive correlation with the zero-shot silhouette score and forgetting. In particular, we highlight BitFit's susceptibility to catastrophic forgetting and A-CLIP's balanced performance for both in-domain and with respect to generalization. We believe further work can be done to quantify the alignment of zero-shot CLIP and other multi-modal models to inform fine-tuning strategies based on desired outcomes as well as similarities between tasks.

Acknowledgements

The authors acknowledge the MIT Lincoln Laboratory Supercomputing Center for providing the high performance computing resources that have contributed to the research results reported within this paper.

DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited.

This material is based upon work supported by the Under Secretary of Defense for Research and Engineering under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Under Secretary of Defense for Research and Engineering.

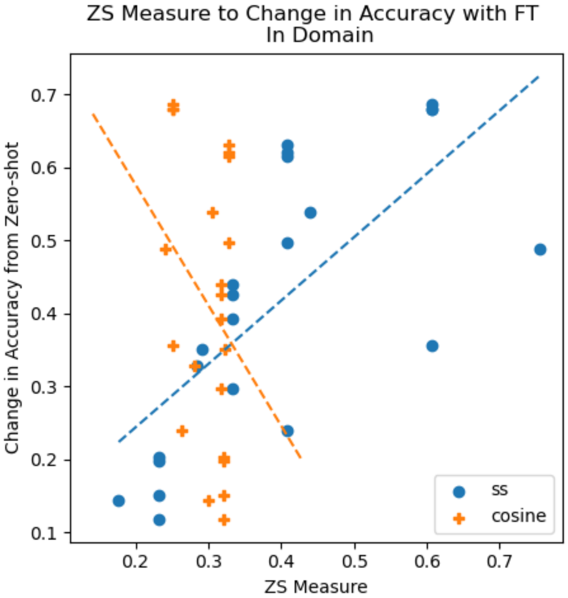


Figure 2: Zero-shot measure to change in accuracy with fine-tuning. Includes all data evaluated on a model trained on the same task. This excludes results from trained with 16-shot ImageNet as few shot learning is inherently different from training with full data, and results using LoRA as our limited hyperparameter search resulted in poor in-domain performance.

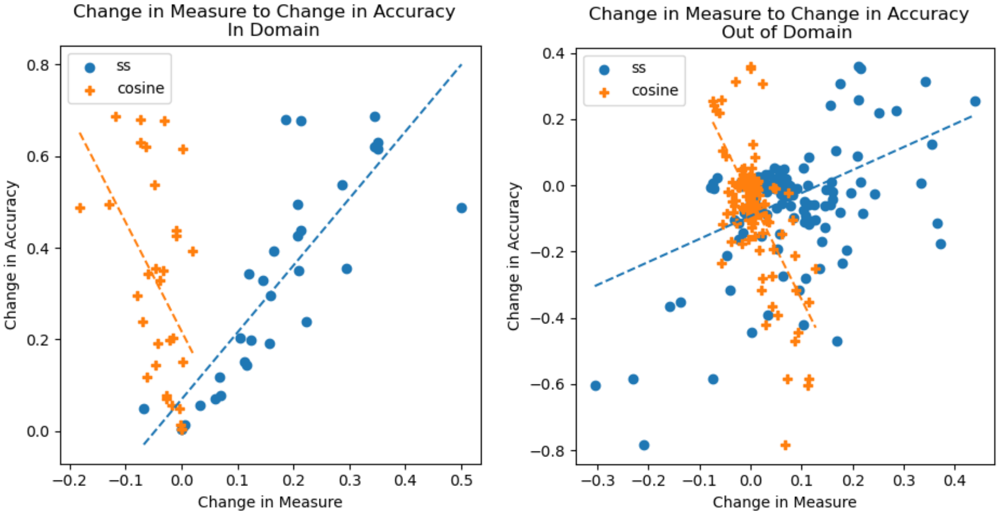


Figure 3: Left plot: change in alignment measure, fine-tuned - zero-shot, to change in accuracy. In-domain includes all data evaluated on a model trained on the same task. Right plot: change in alignment measure to change in accuracy for all data evaluated on a model fine-tuned on a different task.

References

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

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. BitFit: Simple Parameter-efficient Fine-tuning for Transformer-based Masked Language-models. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1–9, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.1.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829, June 2023.

Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional Map of the World. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6172–6180, Salt Lake City, UT, June 2018. IEEE. ISBN 978-1-5386-6420-9. doi: 10.1109/CVPR.2018.00646. URL https://ieeexplore.ieee.org/document/8578744/.

M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Yuxuan Ding, Lingqiao Liu, Chunna Tian, Jingyuan Yang, and Haoxuan Ding. Don't Stop Learning: Towards Continual Learning for the CLIP Model, July 2022. URL http://arxiv.org/abs/2207.09248.

Zhihao Fan, Zhongyu Wei, Jingjing Chen, Siyuan Wang, Zejun Li, Jiarong Xu, and Xuanjing Huang. A unified continuous learning framework for multi-modal knowledge discovery and pre-training, 2022. URL https://arxiv.org/pdf/2206.05555.pdf.

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*, 132(2):581–595, February 2023. ISSN 0920-5691, 1573-1405. doi: 10.1007/s11263-023-01891-x.

Dapeng Hu, Shipeng Yan, Qizhengqiu Lu, Lanqing Hong, Hailin Hu, Yifan Zhang, Zhenguo Li, Xinchao Wang, and Jiashi Feng. How well does self-supervised pre-training perform with streaming data?, 2022a. URL https://arxiv.org/abs/2104.12081.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022b. URL https://openreview.net/forum?id=nZeVKeeFYf9.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, March 2017. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1611835114.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, pp. 554–561, 2013. doi: 10.1109/ICCVW.2013.77.

Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009.

Axel Laborieux, Maxence Ernoult, Tifenn Hirtzlin, and Damien Querlioz. Synaptic metaplasticity in binarized neural networks. *Nature Communications*, 12(1):2549, May 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-22768-y.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Christopher Liao, Theodoros Tsiligkaridis, and Brian Kulis. Descriptor and word soups: Overcoming the parameter efficiency accuracy tradeoff for out-of-distribution few-shot learning, 2023. URL https://arxiv.org/pdf/2311.13612.pdf.

Yong Lin, Lu Tan, Hangyu Lin, Zeming Zheng, Renjie Pi, Jipeng Zhang, Shizhe Diao, Haoxiang Wang, Han Zhao, Yuan Yao, and Tong Zhang. Speciality vs Generality: An Empirical Study on Catastrophic Forgetting in Fine-tuning Foundation Models, October 2023. URL https://arxiv.org/abs/2309.06256.

Michael McCloskey and Neal J. Cohen. Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. In *Psychology of Learning and Motivation*, volume 24, pp. 109–165. Elsevier, 1989. ISBN 978-0-12-543324-2. doi: 10.1016/S0079-7421(08)60536-8.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading Digits in Natural Images with Unsupervised Feature Learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011. URL http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf.

Zixuan Ni, Siliang Tang, and Yueting Zhuang. Self-supervised class incremental learning, 2021. URL https://arxiv.org/abs/2111.11208.

Zixuan Ni, Longhui Wei, Siliang Tang, Yueting Zhuang, and Qi Tian. Continual Vision-Language Representation Learning with Off-Diagonal Information, June 2023. URL https://arxiv.org/abs/2305.07437.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763. PMLR, July 2021.

Tejas Srinivasan, Ting-Yun Chang, Leticia Pinto Alva, Georgios Chochlakis, Mohammad Rostami, and Jesse Thomason. CLiMB: A continual learning benchmark for vision-and-language tasks. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 29440–29453. Curran Associates, Inc., 2022.

J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks*, 32:323–332, 2012. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2012.02.016. URL https://www.sciencedirect.com/science/article/pii/S0893608012000457. Selected Papers from IJCNN 2011.

Hugo Touvron, Matthieu Cord, Alaaeldin El-Nouby, Jakob Verbeek, and Hervé Jégou. Three things everyone should know about vision transformers. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, pp. 497–515, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-20052-6. doi: 10.1007/978-3-031-20053-3_29. URL https://doi.org/10.1007/978-3-031-20053-3_29.

Kevin Vogt-Lowell, Noah Lee, Theodoros Tsiligkaridis, and Marc Vaillant. Robust fine-tuning of vision-language models for domain generalization. In *IEEE High Performance Extreme Computing Conference (HPEC)*, 2023.

J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, June 2010. doi: 10.1109/CVPR.2010.5539970.

A Appendix

A.1 Method details

Training for each method involved 40 epochs of learning using mini-batch stochastic gradient descent with a momentum of 0.9 and a batch size of 128. A weight decay of 1e-5 was used for training on GTSRB, SVHN, and 16-shot ImageNet (and with the additional Full models used in Figure 5), and a weight decay of 1e-4 was used for training on FMoW-ID. The learning rates employed per fine-tuning method can be seen in Table 1, with the only exception being that training on FMoW-ID via cross entropy used a learning rate of 1e-5 instead of 2e-5. No learning rate decay was applied.

Table 1: Learning rates per fine-tuning method

Method	Learning Rate
Full	2e-5
CLIP-Adapter	6e-3
LoRA	1e-5
BitFit	1e-3
A-CLIP	1e-5

For CLIP-Adapter, we used a reduction of 4. LoRA used rank 4 for FMoW-ID and 16-Shot ImageNet and rank 16 for SVHN, GTSRB, and Cars.

For 16-shot ImageNet, SVHN, and GTSRB, validation sets were generated by randomly selecting class-balanced samples from the training data. The validation set for FMoW-ID was provided by the WILDS collection.

All experiments were run non-distributed on a single NVIDIA V100 GPU using an Anaconda 2023b environment with CUDA 11.8.

A.2 Results

We include the raw results of the evaluations for reference. Though we bold the best performing model for each test, we remind the reader that we did not conduct a rigorous hyper-parameter search, therefore minuscule differences between model performances do not necessarily indicate superiority. Rather, we intend to highlight the large differences observed across testing schema.

Table 2 shows the in-domain evaluation results, Table 3 shows evaluation results under the domain generalization schema, and Table 4 shows the results for the catastrophic forgetting schema.

Table 2: In-domain test results.

Data	ID Model Performance					
	zero-shot	CLIP-Adapter	A-CLIP	BitFit	Full	LoRA
Cars	58.87	70.60	**79.12**	73.98	78.66	60.30
FMoW-ID	14.78	44.43	**58.67**	53.98	57.30	15.24
GTSRB	33.65	83.28	95.68	95.19	**96.72**	52.87
16-Shot ImageNet	59.24	64.15	**66.96**	64.93	66.34	59.56
SVHN	27.27	62.78	95.16	95.15	**96.00**	61.70

Table 3: Domain generalization test results.

Data	DG Model Performance					
	zero-shot	CLIP-Adapter	A-CLIP	BitFit	Full	LoRA
FMoW-ID	39.01	37.51	46.25	38.10	**51.89**	39.61
16-Shot ImageNet	48.06	48.53	**50.74**	49.35	49.69	48.37

Table 4: Catastrophic forgetting test results.

Data	CF Model Performance					
	zero-shot	CLIP-Adapter	A-CLIP	BitFit	Full	LoRA
Cars	52.24	50.62	51.64	45.49	**52.66**	52.45
SVHN	56.19	52.62	53.03	16.61	44.27	**57.25**
GTSRB	55.4	47.6	54.75	28.35	51.23	**57.37**

A.3 Alignment Measures

The following are the definitions of the two alignment measure used in this paper:

Average cosine similarity

For image embeddings $\boldsymbol{i} \in \mathbb{I}$ and correct label embedding $\boldsymbol{t}(\boldsymbol{i}) \in \mathbb{T}$, the average cosine similarity is

$$\text{average cosine similarity}(\mathbb{I}, \mathbb{T}) = \frac{1}{|\mathbb{I}|} \sum_{\boldsymbol{i} \in \mathbb{I}} \frac{\boldsymbol{i} \cdot \boldsymbol{t}(\boldsymbol{i})}{||\boldsymbol{i}|| \cdot ||\boldsymbol{t}(i)||}$$

Silhouette score

For a set of clusters $\{\mathbb{C}_1, \mathbb{C}_2 ..., \mathbb{C}_K\}$, vector $\boldsymbol{i} \in \mathbb{C}_I$, and a distance metric d, let

$$a(\boldsymbol{i}) = \frac{1}{|\mathbb{C}_I| - 1} \sum_{\boldsymbol{j} \in \mathbb{C}_I, \boldsymbol{j} \neq \boldsymbol{i}} d(\boldsymbol{i}, \boldsymbol{j})$$

$$b(\boldsymbol{i}) = \min_{J \neq I} \frac{1}{|\mathbb{C}_J|} \sum_{\boldsymbol{j} \in \mathbb{C}_J} d(\boldsymbol{i}, \boldsymbol{j})$$

$$s(\boldsymbol{i}) = \frac{b(\boldsymbol{i}) - a(\boldsymbol{i})}{\max\{b(\boldsymbol{i}), a(\boldsymbol{i})\}}$$

The silhouette score is defined as the mean $s(\boldsymbol{i})$ over all points,

$$ss = \frac{1}{\sum_{I=1}^{K} |\mathbb{C}_I|} \sum_{\boldsymbol{i} \in \cup_{I=1}^{K} \mathbb{C}_I} s(\boldsymbol{i})$$

We use the L2 norm as the distance metric in this work.

A.4 A-CLIP

We trained six additional models with full fine-tuning on Cars, EuroSAT, SUN397, CIFAR100, SVHN, and GTSRB with a learning rate of 1e-5 and a weight decay of 0.05. This larger weight decay revealed a disparity in the magnitude of change in the attention in-projection and out-projection layers not seen with smaller weight decays, leading us to explore CLIP's performance when limiting training to only those weights in the attention layers. We coined this PEFT method A-CLIP. Figure 4 shows the disparity of the magnitude changes of the two attention layers for the vision and text encoders.

Though we refrain from a detailed discussion in this paper, we note that further empirical results on these datasets suggest that fine-tuning only the attention in-projection layers of the first few transformer blocks can achieve in-domain performance on par with full fine-tuning, and we suspect that such fine-tuning would maintain the domain generalization and catastrophic forgetting performance associated with training all of the attention in-projection layers as we have done with A-CLIP in this paper.

A.5 Alignment Measures and Accuracy

We additionally explore if either measure is a useful proxy for accuracy. Looking at measures on in-domain data after fine-tuning, left plot of Figure 5, neither measure has a significant linear relationship with accuracy (p-values > 0.07) and the correlation is weak. With respect to out-of-domain data, right plot of Figure 5 (for which we also include the zero-shot measure to zero-shot accuracy), the silhouette score is significant at $\alpha > 0.0005$, and the average cosine similarity at $\alpha > 0.05$, though both have $R^2 < .08$ showing the relationships are very weak. We conclude that while the silhouette score shows evidence of being a good predictor of improvement gains with fine-tuning, neither the silhouette score nor the average cosine similarity are good proxies for accuracy.

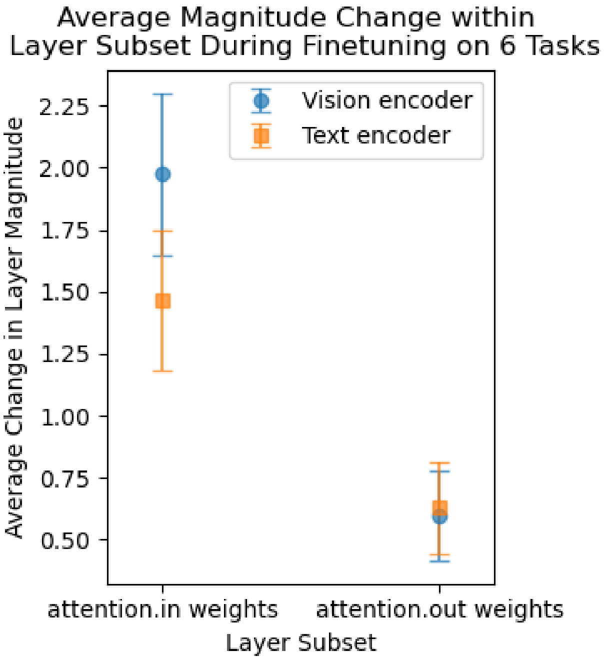


Figure 4: Average difference of layer subset magnitude with 95% confidence intervals.

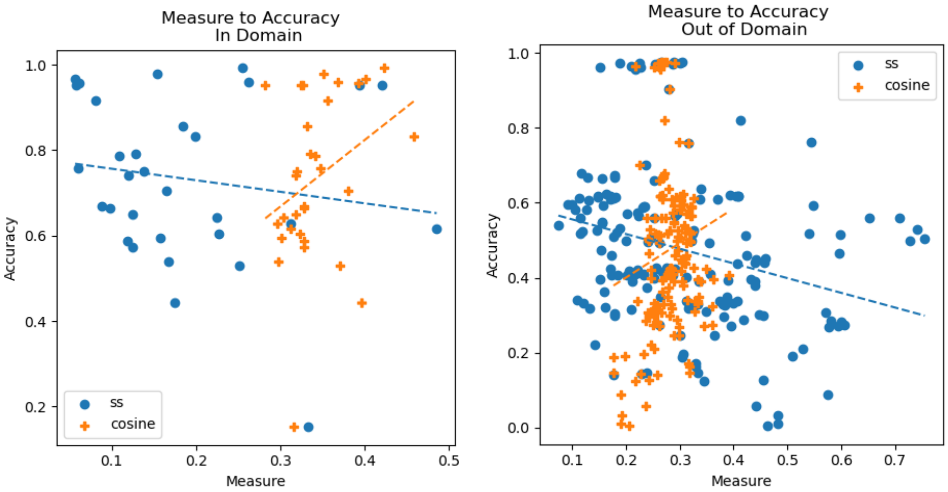


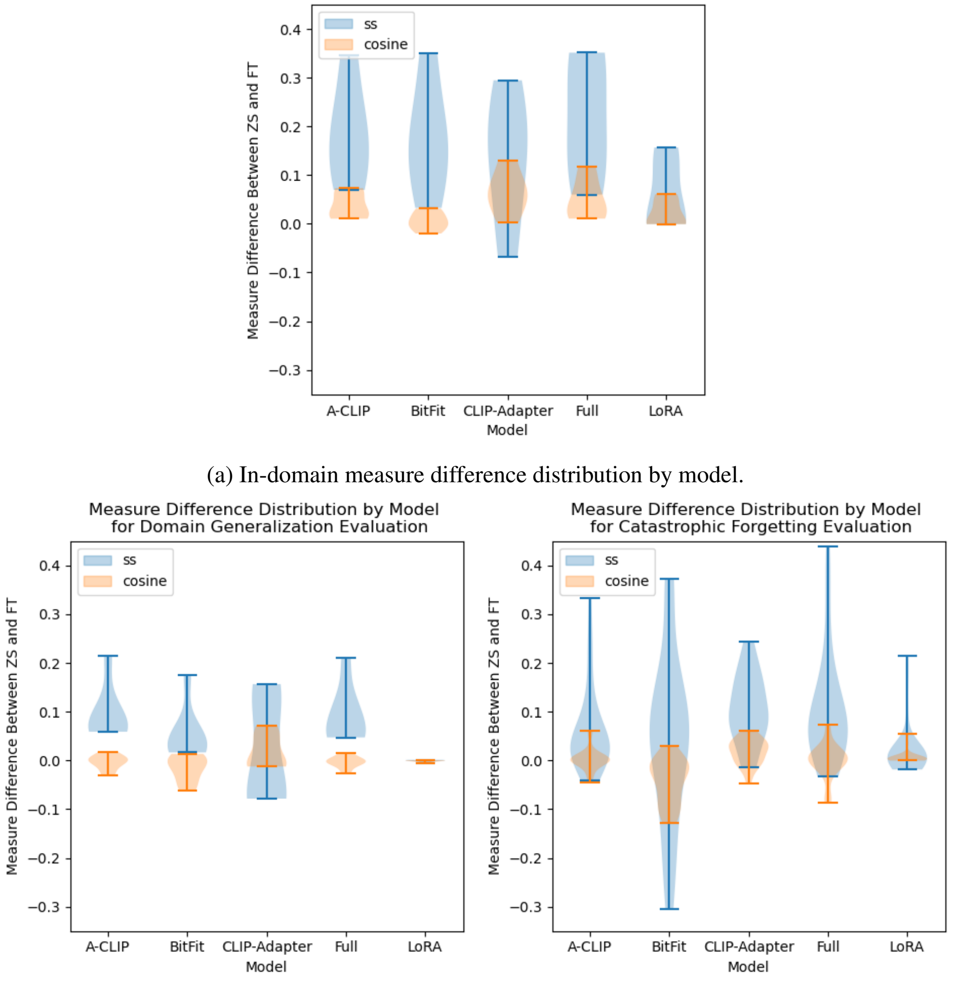
Figure 5: Left plot: alignment measure after fine-tuning fine-tuned accuracy. In-domain includes all data evaluated on on a model trained on the same task. Right plot: alignment measure to accuracy for all data evaluated on a model fine-tuned on a different task.

A.6 Alignment Measure Changes by Model

Looking at both the average cosine similarity and silhouette score, we can compare the alignment changes imposed by each fine-tuning method. In Figure 6, we compare the alignment changes in-domain, with respect to domain generalization, and with respect to catastrophic forgetting. Looking first at in-domain alignment in Figure 6a, A-CLIP, BitFit, and Full have similar measure distributions, with Full having more weight for slightly stronger alignments. CLIP-Adapter and LoRA have measurably smaller changes in alignment, with LoRA being particularly and expectedly small based on its in-domain performance.

In Figure 6b, there exist clear differences in the magnitude of measure change between models in both the domain generalization and catastrophic forgetting evaluation schema. In the domain generalization schema, LoRA has almost no change in alignment, yet it has clear changes in the catastrophic forgetting schema (though less relative to the other methods). These results suggest that hyper-parameter selections which produce better in-domain performance for LoRA will also likely exacerbate catastrophic forgetting.

The other results align with our observations about evaluation performance discussed in section 4. We see BitFit inducing less change in alignment in the domain generalization schema and significant changes in the catastrophic forgetting schema. A-CLIP has similar changes to Full in the domain generalization schema and less extreme changes than Full in the catastrophic forgetting schema, though with more weight towards no change.



(a) In-domain measure difference distribution by model.

(b) Measure difference by model for domain generalization and catastrophic forgetting.

Figure 6: Distributions of the difference in measure from the zero-shot to FT model for in-domain, domain generalization, and catastrophic forgetting. Here, ‘ss’ is the silhouette score of the zero-shot text and image embeddings minus the silhouette score of the FT text and image embeddings. A positive value thus means the clusters of image and text embeddings moved closer together. Similarly, ’cosine’ is the cosine similarity score of the FT text and image embeddings minus the cosine similarity of the zero-shot text and image embeddings. Again, a positive value means the image and appropriate label embeddings moved closer to each other.