YOUR CLIP MODEL MIGHT BE UNDERTRAINED

Anonymous authors
Paper under double-blind review

A EXPERIMENTAL SETUP

Pretraining datasets. In our experiments, we train our CLIP models on three datasets of increasing size, namely CC3M (Sharma et al., 2018), CC12M (Changpinyo et al., 2021), and LAION-400M (Schuhmann et al., 2022). Each of these dataset contains image-caption pairs of datapoints which are using to train CLIP model via contrastive learning.

Models. We consider three different models in our experiments: ResNet-50, ViT-B-32, and ViT-B-16. We first pretrain a version of these models from scratch on the above datasets (except for LAION-400M), matching the results of publicly available models on OpenCLIP[1]. For LAION-400M, we use the checkpoint available on OpenCLIP.

Validation datasets. To evaluate the performance of our models, we use several datasets, including ImageNet variations such as ImageNet-V2, ImageNet-A, ImageNet-R, ImageNet-S, and ObjectNet (Recht et al., 2019; Hendrycks et al., 2019; Barbu et al., 2019; Wang et al., 2019; Hendrycks et al., 2020), as well as suite of transfer learning datasets used in (Kornblith et al., 2019; Salman et al., 2020). We utilize the CLIP benchmarks[2] repository to evaluate all of our models.

Hyperparameters. When training our CLIP models on CC3M (Sharma et al., 2018) and CC12M (Changpinyo et al., 2021), we use hyperparameters similar to the ones employed in (Ilharco et al., 2021). Specifically, we use train our models for a total of 75 epochs using a global batch size of 2,560 (256 samples per GPU), a learning rate of 10^-3, and a weight decay of 0.5.

[1]OpenCLIP repository can be found here https://github.com/mlfoundations/open_clip.
[2]CLIP benchmarks can be found here https://github.com/LAION-AI/CLIP_benchmark.

B ADDITIONAL RESULTS

B.1 HOW MANY EXTRA EPOCHS?

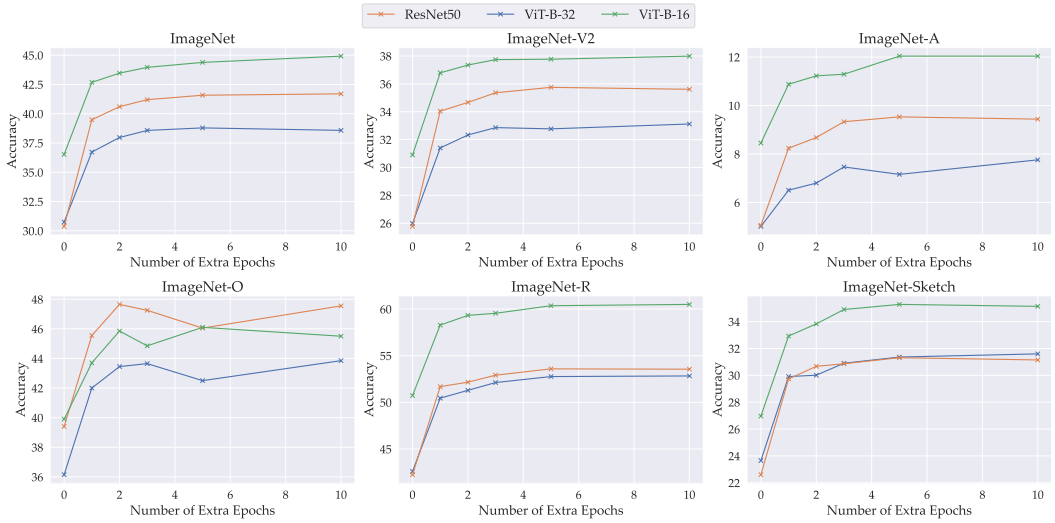

Figure 1: Applying the additional training procedure for few extra epochs is enough to improve performance. The zero-shot accuracy of several CLIP models (y-axis) increases as we apply the additional training procedure for more epochs (x-axis). Note that the performance improvement saturates after applying the procedure for only three additional epochs.

B.2 WHEN TO APPLY OUR STRATEGY?

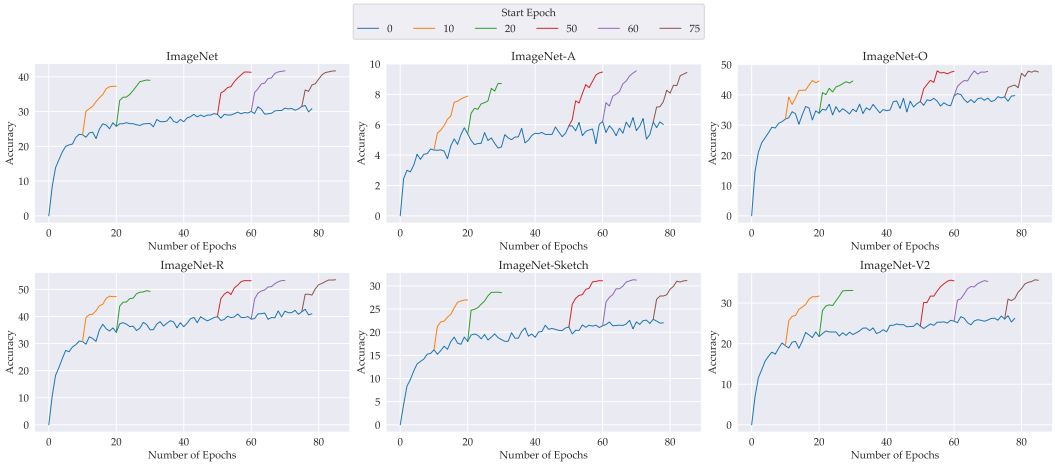

Figure 2: Applying our strategy early during training already improves performance. The blue curve corresponds to the accuracy of the original CLIP model. Each non-blue curve represents the zero-shot accuracy of the CLIP model after applying our strategy with different starting points. For example, the orange curve corresponds to applying our strategy on the CLIP model after it has been trained for 10 epochs (out of 75 epochs in total). Note that applying our strategy earlier during training leads to a performance improvement beyond the final accuracy reached by the model trained for 75 epochs.

CYCLIC LR SCHEDULER

This is a supplementary figure to Figure **??** which shows how cyclic learning rate schedule, instead of a cosine one, can lead to better zero-shot performance for CLIP models trained from scratch.

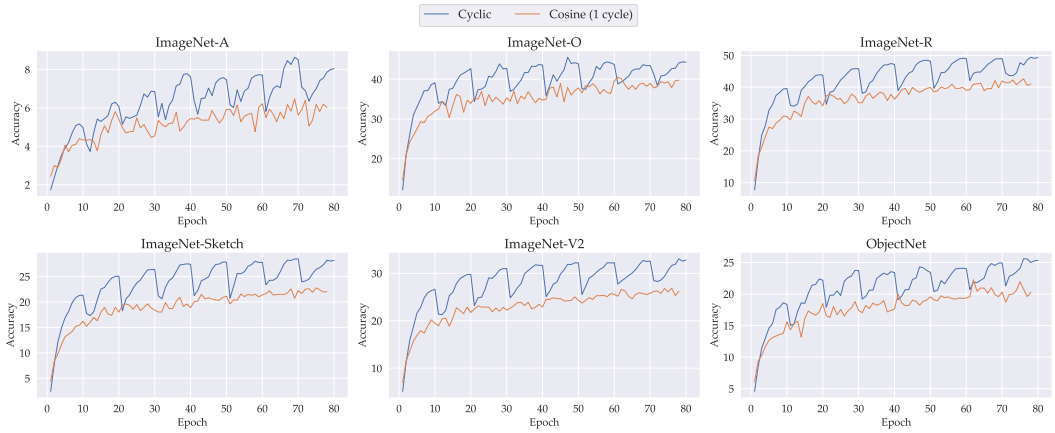

Figure 3: Applying a cyclic learning rate schedule improves performance. Each curve represents the zero-shot accuracy of a ResNet-50 CLIP model as a function of the number of training epochs. The orange curve corresponds to the standard training strategy using a cosine LR scheduler, while the blue curve corresponds to training the CLIP model with a cyclic LR Scheduler. Note that applying a cyclic LR improves performance.

B.3 ADDITIONAL ZERO-SHOT RESULTS ON DOWNSTREAM TASKS

Here we show the performance improvements of our models on a suite of transfer learning tasks, and a range of tasks from the CLIP benchmarks repository[3].

Model	Caltech101	Cars	CIFAR10	CIFAR100	DTD	FGVC Aircraft
ResNet-50	76.4 (+6.03%)	26.2 (+13.3%)	49.4 (+18.5%)	27.5 (+13.0%)	22.1 (+1.86%)	2.67 (+1.05%)
ViT-B-32	77.3 (+3.39%)	19.2 (+7.26%)	81.3 (+9.69%)	43.0 (+2.15%)	21.4 (-0.80%)	2.31 (+0.15%)
ViT-B-16	79.1 (+3.69%)	26.8 (+9.03%)	80.0 (+2.06%)	48.2 (+4.09%)	23.1 (-0.60%)	2.52 (+0.00%)

Table 4: Our simple training procedure consistently improves the performance of CLIP models trained on CC12M. This table shows the zero-shot accuracy of several CLIP models on different downstream tasks after applying our simple strategy. The numbers in parentheses represent the absolute change in zero-shot accuracy on the corresponding downstream classification task.

Model	Flowers	Pets	STL10	SUN397	SVHN
ResNet-50	34.6 (+11.1%)	62.0 (+13.0%)	89.6 (+3.20%)	47.5 (+2.93%)	13.6 (+6.93%)
ViT-B-32	34.0 (+10.6%)	57.8 (+2.80%)	92.0 (+2.47%)	47.3 (+2.47%)	22.6 (+5.42%)
ViT-B-16	37.8 (+14.1%)	64.7 (+12.2%)	93.9 (-0.20%)	48.6 (-0.60%)	19.7 (+2.45%)

Table 5: Our simple training procedure consistently improves the performance of CLIP models trained on CC12M. This table shows the zero-shot accuracy of several CLIP models on different downstream tasks after applying our simple strategy. The numbers in parentheses represent the absolute change in zero-shot accuracy on the corresponding downstream classification task.

[3]CLIP benchmarks can be found here https://github.com/LAION-AI/CLIP_benchmark

REFERENCES

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Neural Information Processing Systems (NeurIPS)*, 2019.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Computer Vision and Pattern Recognition*, 2021.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *arXiv preprint arXiv:1907.07174*, 2019.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization, 2020.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021.

Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *computer vision and pattern recognition (CVPR)*, 2019.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning (ICML)*, 2019.

Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *arXiv preprint arXiv:2210.08402*, 2022.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Association for Computational Linguistics*, 2018.

Haohan Wang, Songwei Ge, Eric P Xing, and Zachary C Lipton. Learning robust global representations by penalizing local predictive power. *Neural Information Processing Systems (NeurIPS)*, 2019.