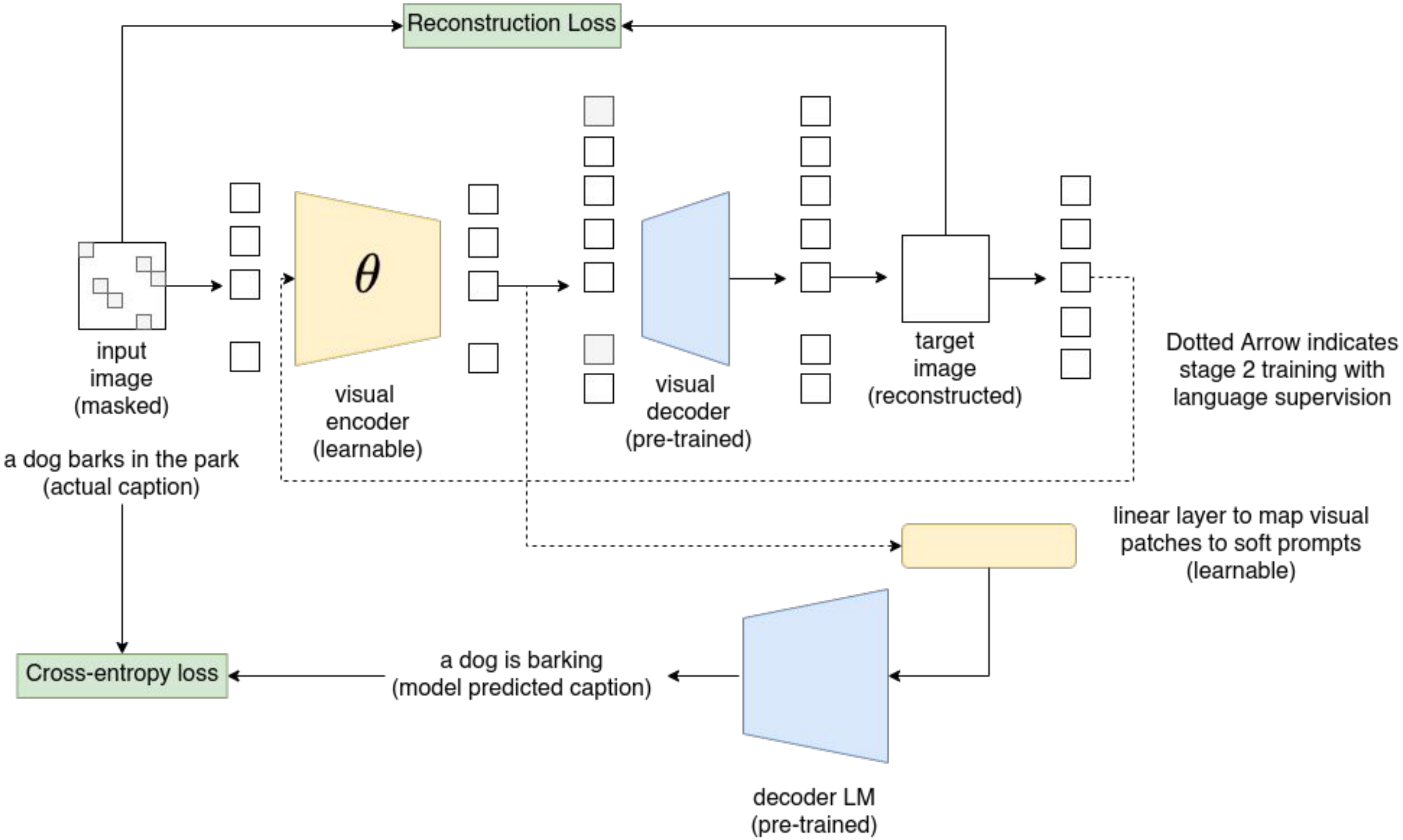
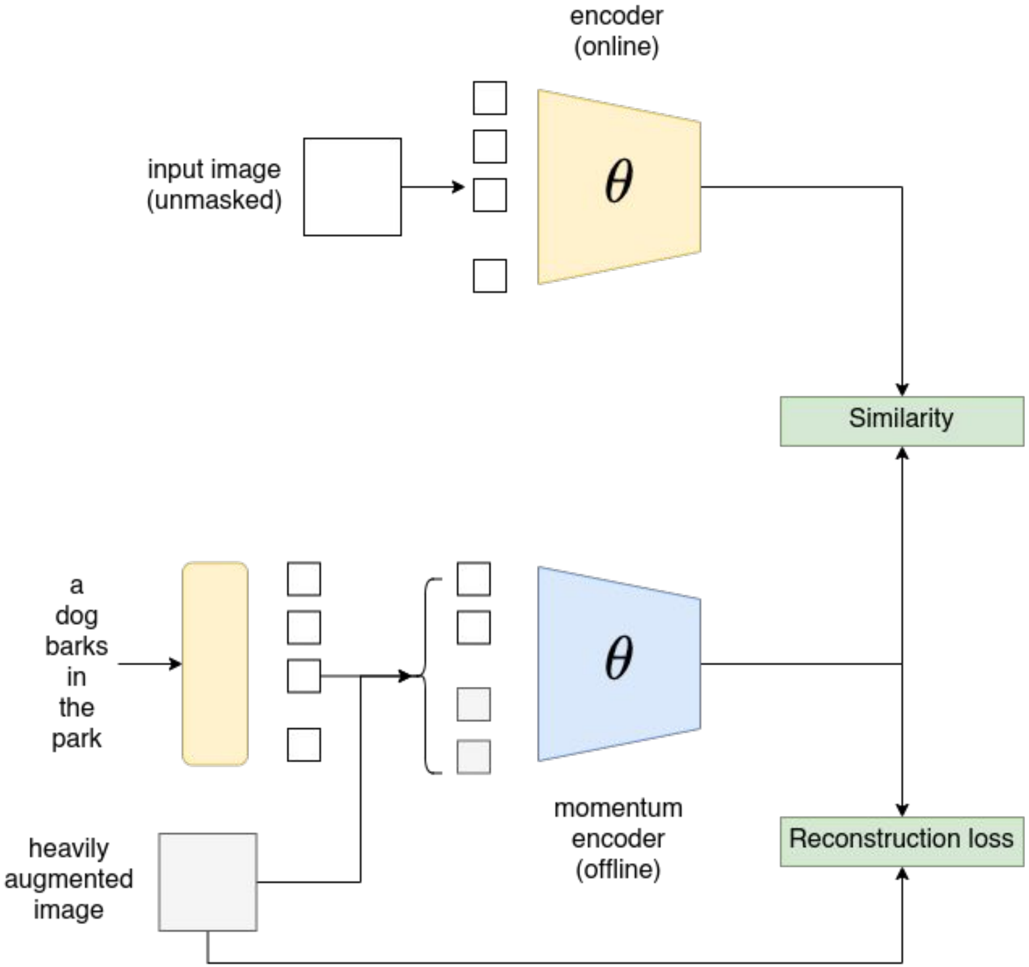
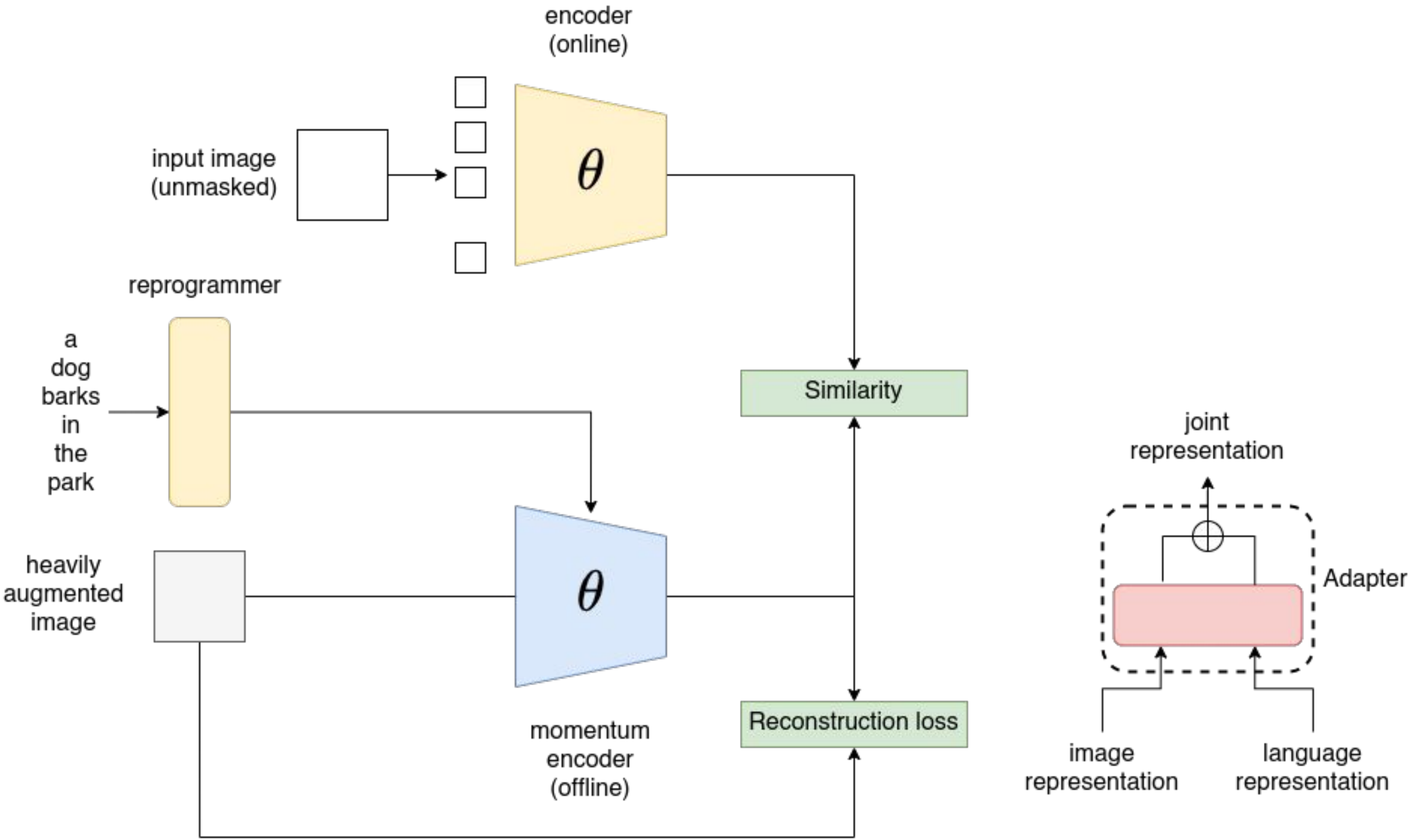
Weak language supervision fine-tuning of vision encoders

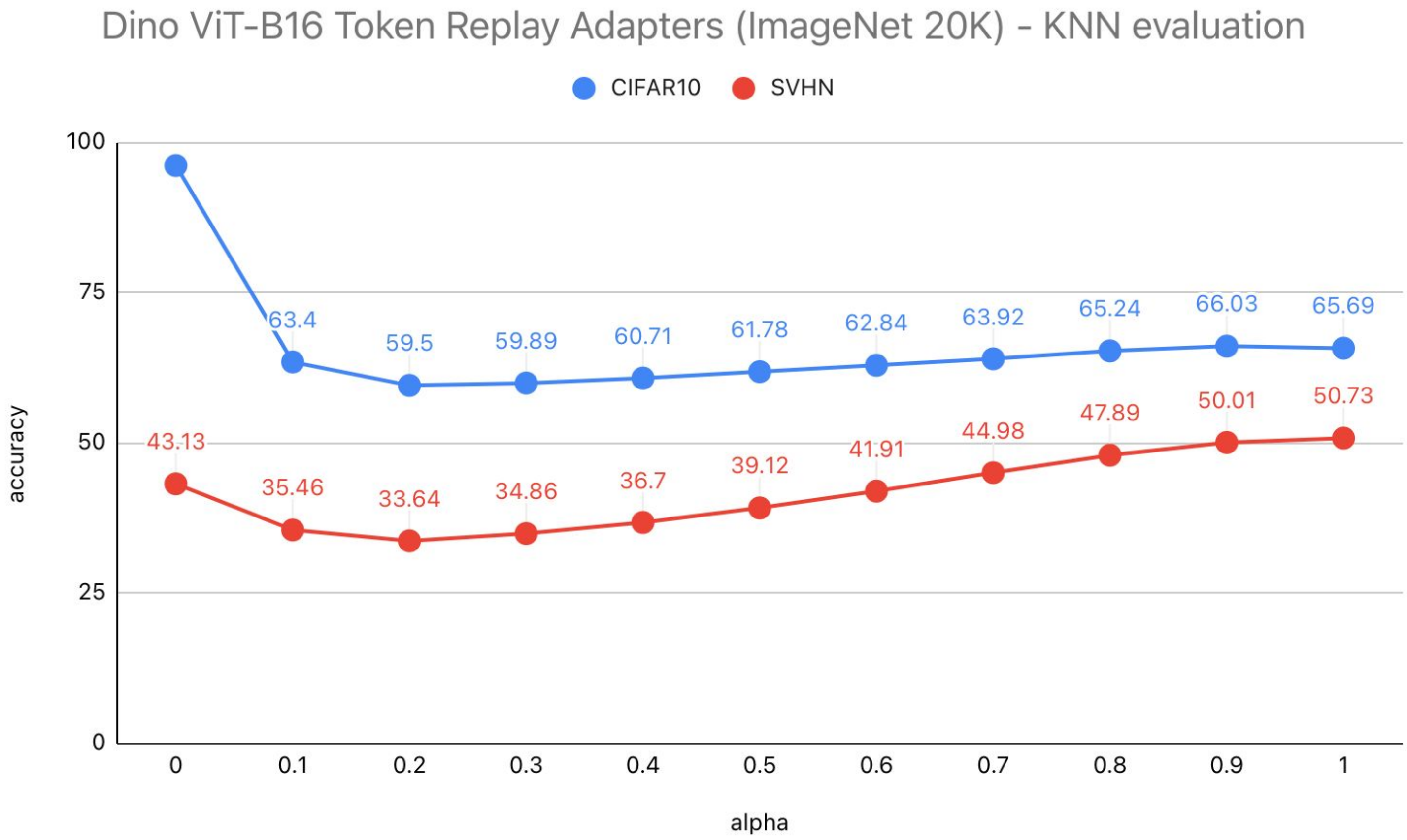
Diganta Misra


Reconstruction Loss
θ
input image (masked)
visual encoder (learnable)
visual decoder (pre-trained)
target image (reconstructed)
Dotted Arrow indicates stage 2 training with language supervision
a dog barks in the park (actual caption)
linear layer to map visual patches to soft prompts (learnable)
Cross-entropy loss
a dog is barking (model predicted caption)
decoder LM (pre-trained)

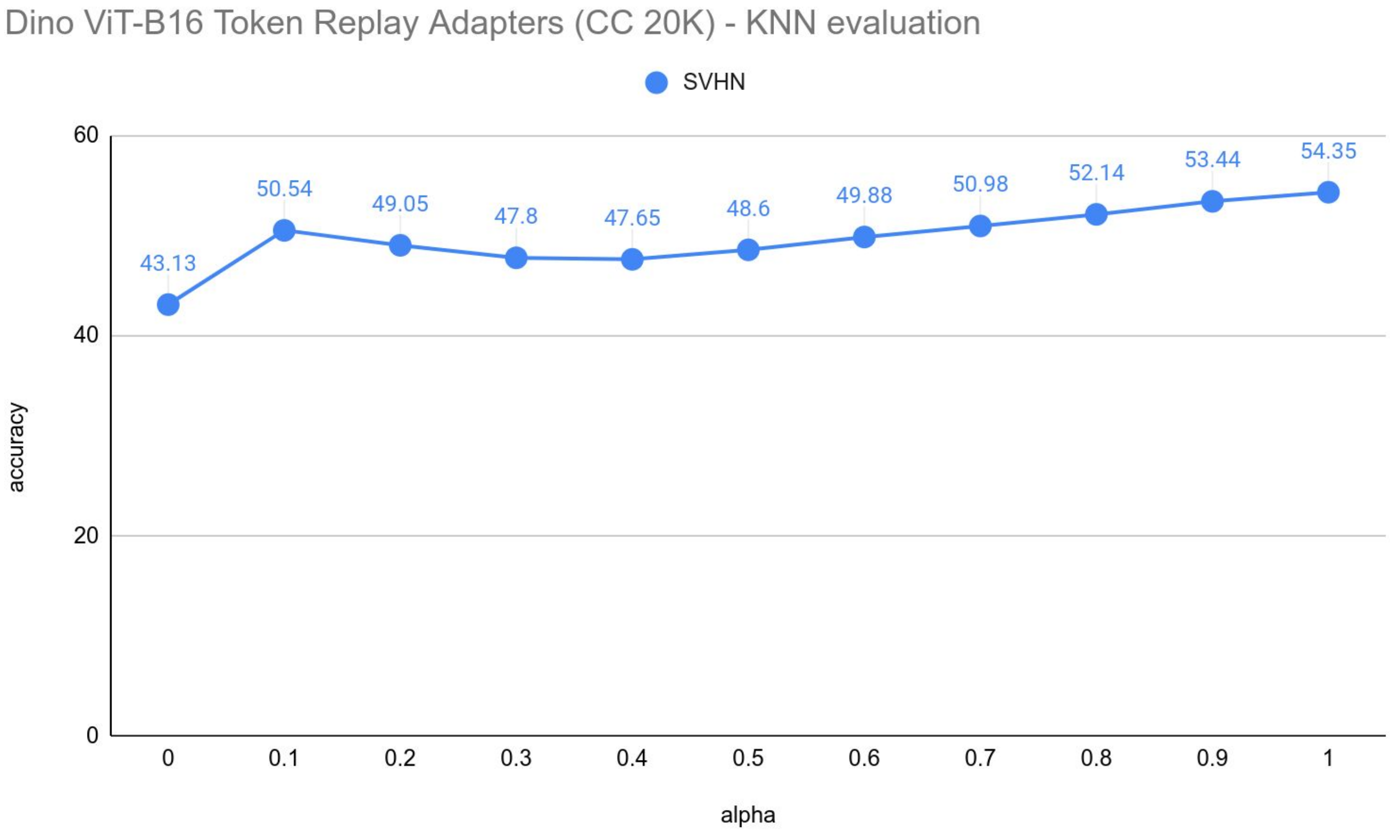

encoder (online)
input image (unmasked)
θ
Similarity
a dog barks in the park
θ
momentum encoder (offline)
Reconstruction loss
heavily augmented image


encoder
(online)
input image
(unmasked)
θ
reprogrammer
a
dog
barks
in
the
park
Similarity
joint
representation
heavily
augmented
image
θ
Adapter
Reconstruction loss
momentum
encoder
(offline)
image
representation
language
representation

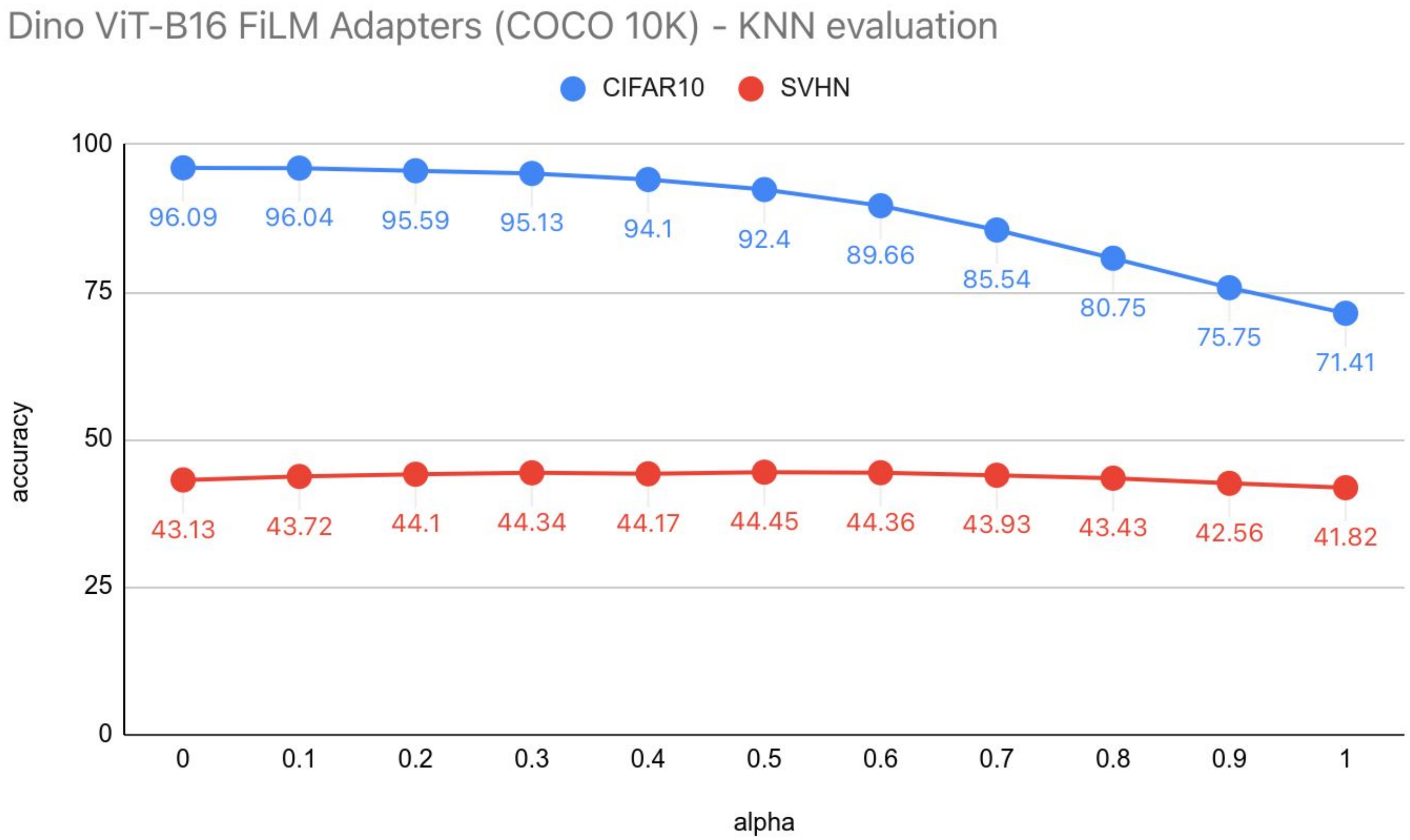
alpha x finetuned model + (1 - alpha) x pretrained model

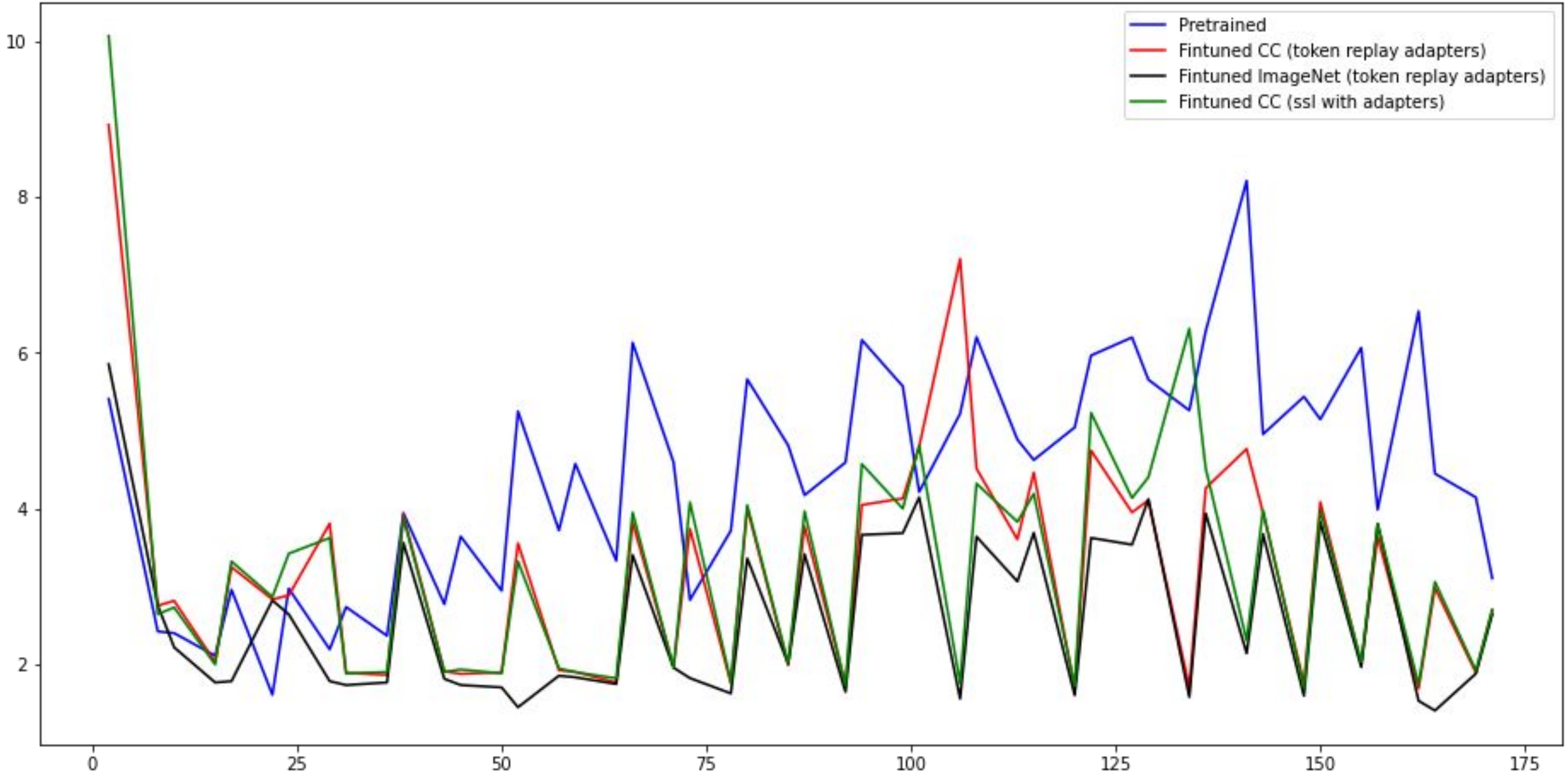


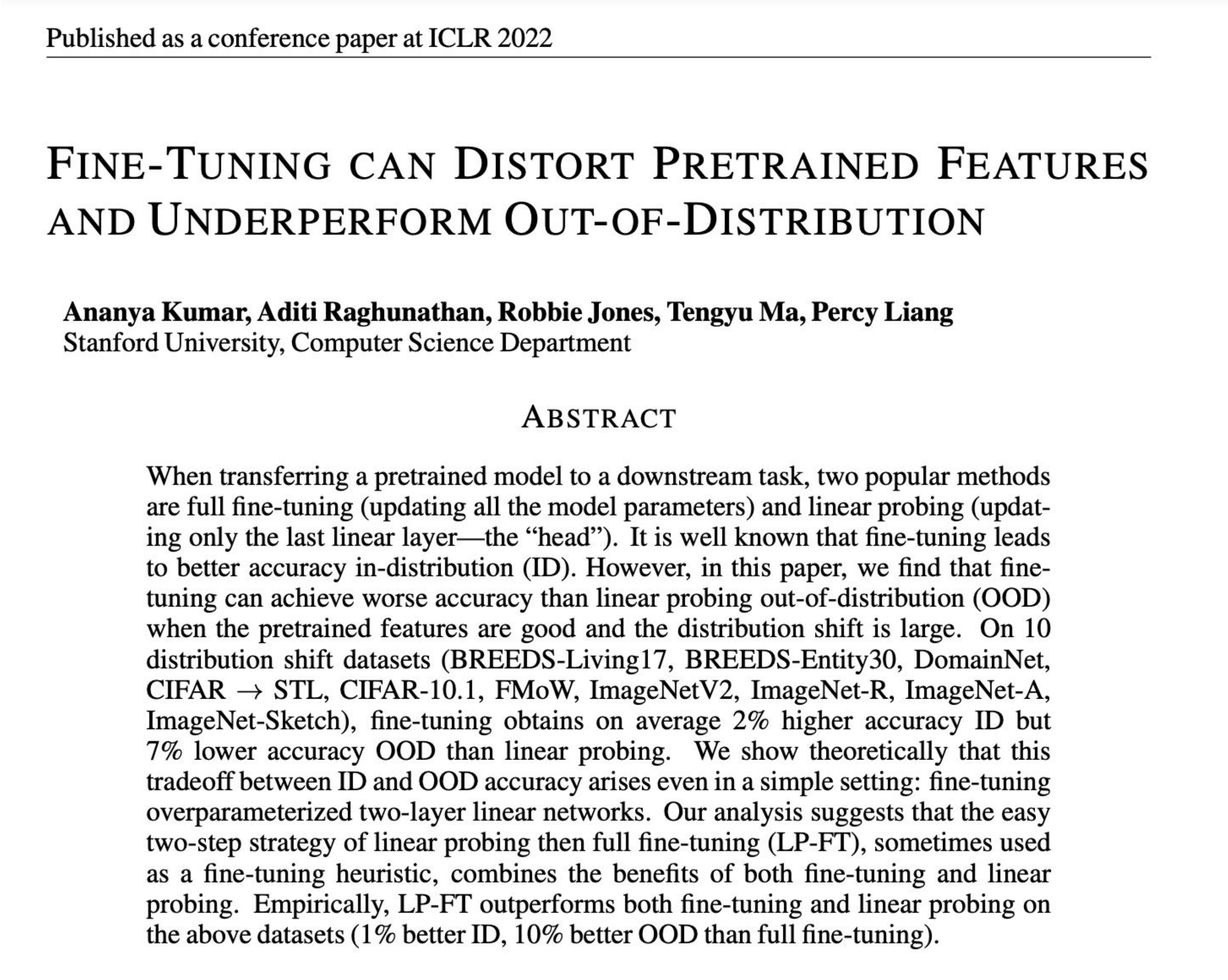
Dino ViT-B16 Token Replay Adapters (ImageNet 20K) - KNN evaluation



Dino ViT-B16 Token Replay Adapters (CC 20K) - KNN evaluation


Dino ViT-B16 FiLM Adapters (COCO 10K) - KNN evaluation
● CIFAR10
● SVHN
accuracy
100
75
50
25
0
alpha
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
96.09
96.04
95.59
95.13
94.1
92.4
89.66
85.54
80.75
75.75
71.41
43.13
43.72
44.1
44.34
44.17
44.45
44.36
43.93
43.43
42.56
41.82


Pretrained
Fintuned CC (token replay adapters)
Fintuned ImageNet (token replay adapters)
Fintuned CC (ssl with adapters)
10
8
6
4
2
0
25
50
75
100
125
150
175

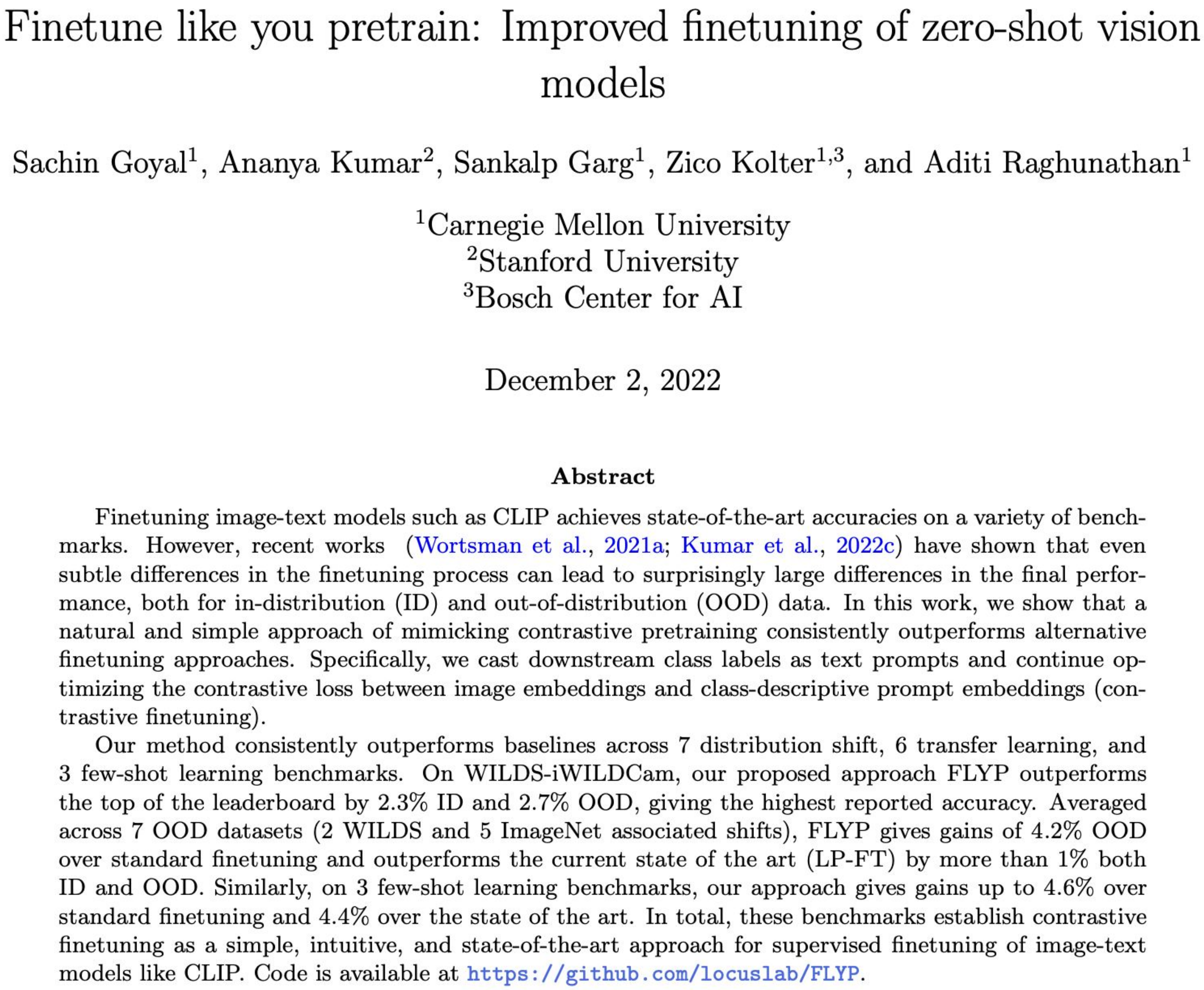


FINE-TUNING CAN DISTORT PRETRAINED FEATURES AND UNDERPERFORM OUT-OF-DISTRIBUTION

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, Percy Liang
Stanford University, Computer Science Department

ABSTRACT

When transferring a pretrained model to a downstream task, two popular methods are full fine-tuning (updating all the model parameters) and linear probing (updating only the last linear layer—the "head"). It is well known that fine-tuning leads to better accuracy in-distribution (ID). However, in this paper, we find that fine-tuning can achieve worse accuracy than linear probing out-of-distribution (OOD) when the pretrained features are good and the distribution shift is large. On 10 distribution shift datasets (BREEDS-Living17, BREEDS-Entity30, DomainNet, CIFAR → STL, CIFAR-10.1, FMoW, ImageNetV2, ImageNet-R, ImageNet-A, ImageNet-Sketch), fine-tuning obtains on average 2% higher accuracy ID but 7% lower accuracy OOD than linear probing. We show theoretically that this tradeoff between ID and OOD accuracy arises even in a simple setting: fine-tuning overparameterized two-layer linear networks. Our analysis suggests that the easy two-step strategy of linear probing then full fine-tuning (LP-FT), sometimes used as a fine-tuning heuristic, combines the benefits of both fine-tuning and linear probing. Empirically, LP-FT outperforms both fine-tuning and linear probing on the above datasets (1% better ID, 10% better OOD than full fine-tuning).



Finetune like you pretrain: Improved finetuning of zero-shot vision models

Sachin Goyal[1], Ananya Kumar[2], Sankalp Garg[1], Zico Kolter[1,3], and Aditi Raghunathan[1]

[1]Carnegie Mellon University
[2]Stanford University
[3]Bosch Center for AI

December 2, 2022

Abstract

Finetuning image-text models such as CLIP achieves state-of-the-art accuracies on a variety of benchmarks. However, recent works (Wortsman et al., 2021a; Kumar et al., 2022c) have shown that even subtle differences in the finetuning process can lead to surprisingly large differences in the final performance, both for in-distribution (ID) and out-of-distribution (OOD) data. In this work, we show that a natural and simple approach of mimicking contrastive pretraining consistently outperforms alternative finetuning approaches. Specifically, we cast downstream class labels as text prompts and continue optimizing the contrastive loss between image embeddings and class-descriptive prompt embeddings (contrastive finetuning).

Our method consistently outperforms baselines across 7 distribution shift, 6 transfer learning, and 3 few-shot learning benchmarks. On WILDS-iWILDCam, our proposed approach FLYP outperforms the top of the leaderboard by 2.3% ID and 2.7% OOD, giving the highest reported accuracy. Averaged across 7 OOD datasets (2 WILDS and 5 ImageNet associated shifts), FLYP gives gains of 4.2% OOD over standard finetuning and outperforms the current state of the art (LP-FT) by more than 1% both ID and OOD. Similarly, on 3 few-shot learning benchmarks, our approach gives gains up to 4.6% over standard finetuning and 4.4% over the state of the art. In total, these benchmarks establish contrastive finetuning as a simple, intuitive, and state-of-the-art approach for supervised finetuning of image-text models like CLIP. Code is available at https://github.com/locuslab/FLYP.

Questions?