Robust Fine-tuning of Zero-shot Models

Ludwig Schmidt

(Samir Gadre filling in)
Fine-tuning vs. zero-shot inference

State-of-the-art ML models often come from a **two-step process**.

1. **Pre-training**
   - Large-scale noisy web data

2. **Fine-tuning**
   - Small-scale clean task-specific data

**Adapting to a task of interest**

What is the best way to fine-tune a large pre-trained model?
Focus today: out-of-distribution robustness

Transportation

Health care

Robotics

Chat assistants

Need reliable machine learning
Robustness on ImageNet

Lots of progress on ImageNet over the past 10 years, but models are still not robust.

Evaluation: new test sets

- ImageNetV2
  - [Recht, Roelofs, Schmidt, Shankar ’19]
- ObjectNet
  - [Barbu, Mayo, Alverio, Luo, Wang, Gutfreund, Tenenbaum, Katz ’19]
- ImageNet-Sketch
  - [Wang, Ge, Lipton, Xing ’19]
- ImageNet-R
  - [Hendrycks, Basart, Mu, Kadavath, Wang, Dorundo, Desai, Zhu, Parajuli, Guo, Song, Steinhardt, Gilmer ’20]
What robustness interventions help?

[Image showing a scatter plot with the following markers:
- EfficientNet-B7
- VGG, ResNet, DenseNet, ResNeXt, Inception, NASNet, etc.
- AlexNet

Equation: \( y = x \)

Standard models

[Taori, Dave, Shankar, Carlini, Recht, Schmidt ’20]
What robustness interventions help?

Expected out-of-distribution accuracy

Baseline out-of-distribution accuracy from in-distribution accuracy.
What robustness interventions help?

Do current robustness interventions achieve **effective robustness**?
No current robustness technique achieves non-trivial effective robustness.

Only training on (a lot) more data gives a small amount of effective robustness.
Distribution Shift to ObjectNet

Same trend: only **more data** gives effective robustness.

[Barbu, Mayo, Alverio, Luo, Wang, Gutfreund, Tenenbaum, Katz ’19]
CLIP: Connecting Text and Images

We’re introducing a neural network called CLIP which efficiently learns visual concepts from natural language supervision. CLIP can be applied to any visual classification benchmark by simply providing the names of the visual categories to be recognized, similar to the “zero-shot” capabilities of GPT-2 and GPT-3.
| DATASET          | IMAGENET RESNET101 | CLIP VIT-L |
|------------------|--------------------|------------|
| ImageNet         | 76.2%              | 76.2%      |
| ImageNet V2      | 64.3%              | 70.1%      |
| ImageNet Rendition| 37.7%              | 88.9%      |
| ObjectNet        | 32.6%              | 72.3%      |
| ImageNet Sketch  | 25.2%              | 60.2%      |
| ImageNet A       | 2.7%               | 77.1%      |

Effective robustness

+6%
+51%
+40%
+35%
+74%

Very large improvements in out-of-distribution robustness.
Large robustness gains

What makes CLIP robust?

But: fine-tuning reduces robustness

Can we get both high in-distribution and out-of-distribution accuracy?
What makes CLIP robust?

Data Determines Distributional Robustness in Contrastive Language Image Pre-training (CLIP)

Alex Fang† Gabriel Ilharco† Mitchell Wortsman† Yuhao Wan†

Vaishaal Shankar○ Achal Dave○ Ludwig Schmidt†○

Abstract
Contrastively trained image-text models such as CLIP, ALIGN, and BASIC have demonstrated unprecedented robustness to multiple challenging natural distribution shifts. Since these image-text models differ from previous training approaches in several ways, an important question is what causes the large robustness gains. We answer this question via a systematic experimental investigation. Concretely, we study five different possible causes for the robustness gains: (i) the training set size, (ii) the training distribution, (iii) language supervision at training time, (iv) language supervision at test time, and (v) the contrastive loss function. Our experiments show that the more diverse training distribution is the main cause for the robustness gains, with the other factors contributing little to no robustness. Beyond our experimental results, we also introduce ImageNet-Captions, a version of ImageNet with original text annotations from Flickr, to enable further controlled experiments of language-image training.
## Hypotheses for CLIP’s robustness

| Feature                        | CLIP          | Standard ImageNet supervised learning |
|-------------------------------|---------------|---------------------------------------|
| Language supervision         | Yes           | No                                    |
| Training distribution        | ???           | ImageNet                              |
| Training set size            | 400M          | 1.2M                                  |
| Loss function                | Contrastive   | Supervised                            |
| Test-time prompting          | Yes           | No                                    |
| Model architecture           | ViTs          | CNNs                                  |
Hypotheses for CLIP’s robustness

| Hypothesis                        | CLIP       | Standard ImageNet supervised learning |
|-----------------------------------|------------|---------------------------------------|
| Language supervision              | Yes        | No                                    |
| Training distribution             | ??         | ImageNet                              |
| Training set size                 | 400M       | 1.2M                                  |
| Loss function                     | Contrastive| Supervised                            |
| Test-time prompting               | Yes        | No                                    |
| Model architecture                | ViTs       | CNNs                                  |
One takeaway: datasets are a key for improving models

DATAComp:
In search of the next generation of multimodal datasets

Samir Yitzhak Gadre*2 Gabriel Ilharco*1 Alex Pang*1 Jonathan Hayase1 Georgios Smyrnis5
Thao Nguyen1 Ryan Marten7,9 Mitchell Wortsman1 Dhruva Ghosh1 Jieyu Zhang1
Eyal Orgad3 Rahim Entezari10 Giannis Daras5 Sarah Pratt1 Vivek Ramanujan1
Yonatan Bitton11 Kalyani Marathe1 Stephen Mussmann1 Richard Vencu6
Mehdi Cherti6,8 Ranjay Krishna1 Pang Wei Koh1,12 Olga Sorkh10 Alexander Ratner1,13
Shuran Song2 Hannaneh Hajishiriz1,7 Ali Farhadi1 Romain Beaumont6
Sewoong Oh1 Alexandros G. Dimakis5 Jenia Itsev6,8
Yair Carmon3 Vaishaal Shankar1 Ludwig Schmidt1,6,7

Abstract
Multimodal datasets are a critical component in recent breakthroughs such as Stable Diffusion and GPT-4, yet their design does not receive the same research attention as model architectures or training algorithms. To address this shortcoming in the ML ecosystem, we introduce DATAComp, a testbed for dataset experiments centered around a new candidate pool of 12.8 billion image-text pairs from Common Crawl. Participants in our benchmark design new filtering techniques or extract new data sources and then evaluate their new dataset by running our standardized CHP

Workshop tomorrow at ICCV!
Can we fine-tune CLIP without losing robustness?

Robust fine-tuning of zero-shot models

Mitchell Wortsman*† Gabriel Ilharco*† Jong Wook Kim§ Mike Li‡

Simon Kornblith○ Rebecca Roelofs○ Raphael Gontijo-Lopes○

Hannaneh Hajishirzi†○ Ali Farhadi*† Hongseok Namkoong*‡ Ludwig Schmidt†∆

Abstract

Large pre-trained models such as CLIP or ALIGN offer consistent accuracy across a range of data distributions when performing zero-shot inference (i.e., without fine-tuning on a specific dataset). Although existing fine-tuning methods substantially improve accuracy on a given target distribution, they often reduce robustness to distribution shifts. We address this tension by introducing a simple and effective method for improving robustness while fine-tuning: ensembling the weights of the zero-shot and fine-tuned models (WiSE-FT). Compared to standard fine-tuning, WiSE-FT provides large accuracy improvements under distribution shift, while preserving high accuracy on the target distribution. On ImageNet and five derived distribution shifts, WiSE-FT improves accuracy under distribution shift by 4 to 6 percentage points (pp) over prior work while increasing ImageNet accuracy by 1.6 pp. WiSE-FT achieves similarly large robustness gains (2 to 23 pp) on a diverse set of six further distribution shifts, and accuracy gains of 0.8 to 3.3 pp compared to standard fine-tuning on seven commonly used transfer learning datasets. These improvements come at no additional computational cost during fine-tuning or inference.
The problem with fine-tuning

Raised as an open problem by researchers from OpenAI, Stanford, Google, etc.
A simple but effective solution

Weight-space ensembles for fine-tuning (WiSE-FT)

Task accuracy
Robustness

Building on [Nagarajan, Kolter ’19], [Frankle, Dziugaite, Roy, Carbin ’20], [Neyshabur, Sedghi, Zhang ’20].
Training from scratch

Linearly interpolating the weights of two models trained from scratch encounters a high error barrier (Frankle et al., 2020).
Accuracy remains high when linearly interpolating the weights of two networks fine-tuned from a shared initialization (Neyshabur et al., 2020).
Key difference between fine-tuning and training from scratch

- **From schematic to experiment**: fine-tuned models often appear to lie in a single, low-error region.
In-distribution (ID) accuracy
Out-of-distribution (OOD) accuracy

Models trained on in-distribution train set
Zero-shot CLIP models
• Weight-space ensemble for $\alpha \in [0, 1]$: $\theta_\alpha = (1 - \alpha) \cdot \theta_{\text{zero-shot}} + \alpha \cdot \theta_{\text{fine-tuned}}$

Schematic: previous methods
(Higher ID but lower OOD accuracy after fine-tuning)

Schematic: our method, WiSE-FT
(Better OOD accuracy without decreasing ID accuracy)

Accuracy on $D$

Accuracy on $D'$

Models trained on in-distribution train set
CLIP zero-shot
Linear fit (CLIP zero-shot)
CLIP fine-tuned end-to-end
CLIP fine-tuned with a linear classifier
Weight-space ensemble (end-to-end)
Weight-space ensemble (linear classifier)
Best OOD without reducing ID
Standard ImageNet models
Linear fit (standard ImageNet models)
In-distribution (ID) accuracy
Out-of-distribution (OOD) accuracy

Models trained on
in-distribution train set

Zero-shot CLIP models

Effective robustness
Fine-tuned CLIP

Schematic: previous methods
(Higher ID but lower OOD accuracy after fine-tuning)

Schematic: our method, WiSE-FT
(Better OOD accuracy without decreasing ID accuracy)

Avg. accuracy on 5 distribution shifts

Real data: our method

ImageNet (top-1, %)

CLIP zero-shot
Linear fit (CLIP zero-shot)
CLIP fine-tuned end-to-end
Weight-space ensemble (end-to-end)
Best OOD without reducing ID
Standard ImageNet models
Linear fit (standard ImageNet models)

\( y = x \)
| Train | Test (OOD) |
|---|---|
| $d = \text{Location 1}$ | $d = \text{Location 2}$ | $d = \text{Location 245}$ | $d = \text{Location 246}$ |
| Vulturine Guineafowl | African Bush Elephant | ... | unknown |
| Cow | Cow | Southern Pig-Tailed Macaque | Great Curassow |

| Test (ID) |
|---|
| $d = \text{Location 1}$ | $d = \text{Location 2}$ | $d = \text{Location 245}$ |
| Giraffe | Impala | Sun Bear |

**FMoW**

| Year / Region | Train | Test |
|---|---|---|
| 2002 / Americas | | |
| 2009 / Africa | | |
| 2012 / Europe | | |
| 2016 / Americas | | |
| 2017 / Africa | | |

| Building / Land Type | Train | Test |
|---|---|---|
| shopping mall | | |
| multi-unit residential | | |
| road bridge | | |
| recreational facility | | |
| educational institution | | |

- **+3.0pp OOD**
  - Predicted: domestic cat
- **+2.2pp OOD**
  - Predicted: monkey

Christie et al., 2018

Koh et al., 2021

Beery et al., 2018

CIFAR-10.1.
Recht et al., 2019

CIFAR-10.2.
Lu et al., 2020

ImageNet-Vid-Robust
Shankar et al., 2019

YTBBRobust

+8.3pp OOD

+14.7pp OOD

🥈 Best paper finalist, CVPR 2022
Robustness gains invariant as compute scale increases

- In-distribution (ID) accuracy
- Out-of-distribution (OOD) accuracy

Models trained on in-distribution train set

- Zero-shot CLIP models
- Effective robustness

Schematic: previous methods
(Higher ID but lower OOD accuracy after fine-tuning)

Schematic: our method, WiSE-FT
(Better OOD accuracy without decreasing ID accuracy)

- Weight-space ensemble for $\mathcal{E} \in [0 \rightarrow 1]$
- $\mu_{\mathcal{E}} = (1 \rightarrow \mathcal{E}) \leq \mu_{\text{zero-shot}} + \mathcal{E} \leq \mu_{\text{fine-tuned}}$

Real data: our method

Final result (high accuracy models)

Reliable extrapolation via “Accuracy on the line”

Where most experiments happened (low accuracy models)
- cheaper → faster iteration

ImageNet (top-1, %)

Avg. accuracy on 5 distribution shifts
All experiments measured effective robustness
Robustness gains invariant as compute scale increases

Experiment with the full-scale model worked on first try

ID-OOD trends are a reliable scaling law for model design

Real data: our method

Final result (high accuracy models)

Reliable extrapolation via “Accuracy on the line”

Where most experiments happened (low accuracy models)

→ cheaper → faster iteration

ImageNet (top-1, %)

Avg. accuracy on 5 distribution shifts

CLIP zero-shot

Linear fit (CLIP zero-shot)

CLIP fine-tuned end-to-end

Weight-space ensemble (end-to-end)

Best OOD without reducing ID

Standard ImageNet models

Linear fit (standard ImageNet models)

y = x

Final result (high accuracy models)

Where most experiments happened (low accuracy models)

→ cheaper → faster iteration

Experiment with the full-scale model worked on first try

ID-OOD trends are a reliable scaling law for model design
Finetune like you pretrain: Improved finetuning of zero-shot vision models

Sachin Goyal¹, Ananya Kumar², Sankalp Garg¹, Zico Kolter¹,³, and Aditi Raghunathan¹

¹Carnegie Mellon University
²Stanford University
³Bosch 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).
Why stop at averaging two models?

Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time
Conventional procedure for maximizing accuracy while fine-tuning

1. Fine-tune with various hyper-parameters.
Conventional procedure for maximizing accuracy while fine-tuning

1. Fine-tune with various hyper-parameters.

2. Choose the model with the best accuracy on the held-out validation set.

- LR=1e-5, aug=minimal: 81.3%
- LR=2e-5, mixup=0.8: 79.6%
- LR=3e-5, decay=0.1: 79.6%
- LR=2e-5, mixup=0.8: 82.1%
Downsides of the conventional fine-tuning recipe

Choosing the best individual model on the held-out validation set

Lower accuracy

Ensemble

Higher inference cost
Model soups

\[ \frac{1}{n} \sum_{i=1}^{n} \theta_i \]

Best of both worlds:
- Same **high accuracy** as the ensemble
- Same **fast inference** time as an individual model

**Results**
- ImageNet SotA 🏆
- Gains on many more datasets
- Widely used for multimodal models
Can we fine-tune a model while preserving its zero-shot abilities?

Patching open-vocabulary models by interpolating weights

Gabriel Ilharco* University of Washington ganaga@cs.washington.edu
Mitchell Wortsman* University of Washington mmitch@cs.washington.edu
Samir Yitzhak Gadre* Columbia University sy@cs.columbia.edu
Shuran Song Columbia University shurans@cs.columbia.edu
Hannaneh Hajishirzi University of Washington hannah@cs.washington.edu
Simon Kornblith Google Research, Brain Team skornblith@google.com
Ali Farhadi University of Washington ali@cs.washington.edu
Ludwig Schmidt University of Washington schmidt@cs.washington.edu

Abstract

Open-vocabulary models like CLIP achieve high accuracy across many image classification tasks. However, there are still settings where their zero-shot performance is far from optimal. We study model patching, where the goal is to improve accuracy on specific tasks without degrading accuracy on tasks where performance is already adequate. Towards this goal, we introduce PAINT, a patching method that uses interpolations between the weights of a model before fine-tuning and the weights after fine-tuning on a task to be patched. On nine tasks where zero-
Conclusions

Pre-trained models often can be improved by fine-tuning on task-specific data.

Both in **vision** and in **NLP** (instruction tuning, RLHF, etc.)

“Standard” fine-tuning can **negatively affect the capabilities** of the pre-trained model.

Interpolating between the pre-trained and fine-tuned models can preserve robustness while improving task performance.

**Open questions**

- Simple weight interpolation seems naive $\rightarrow$ are there better fine-tuning methods?
- Can we remove fine-tuning entirely and improve pre-training instead?