< 2022/04/18 >

MLAI

University of Seoul

Robust fine-tuning of zero-shot models

Mitchell, Wortsman, et al. CVPR ‘22

Hoyoon Byun

Contents

- Overview
- Problem
- Experimental setup
- Methods (Weight-space ensembles for fine-tuning)
- Results
- Discussion & Conclusion

($+\alpha$) Individual opinion

Source:

Overview

- A foundation goal of machine learning is to develop models that work reliably across a broad range of data distributions
- Recently, large pre-trained models such as CLIP, ALIGN and BASIC have demonstrated novel to these challenging distribution shifts.
- However, these robustness improvements are largest in the zero-shot setting.
- When a pre-trained zero-shot model is fine-tuned on target distribution, which often yields large performance gains on the target distribution.
- This paper propose a simple and effective methods for improving robustness while fine-tuning
  - Ensembling between the weights of the zero-shot and fine-tuned models

Source:

Problem

- The main question what this paper try to answer is
  - Can zero-shot models be fine-tuned without reducing accuracy under distribution shift?

Source:

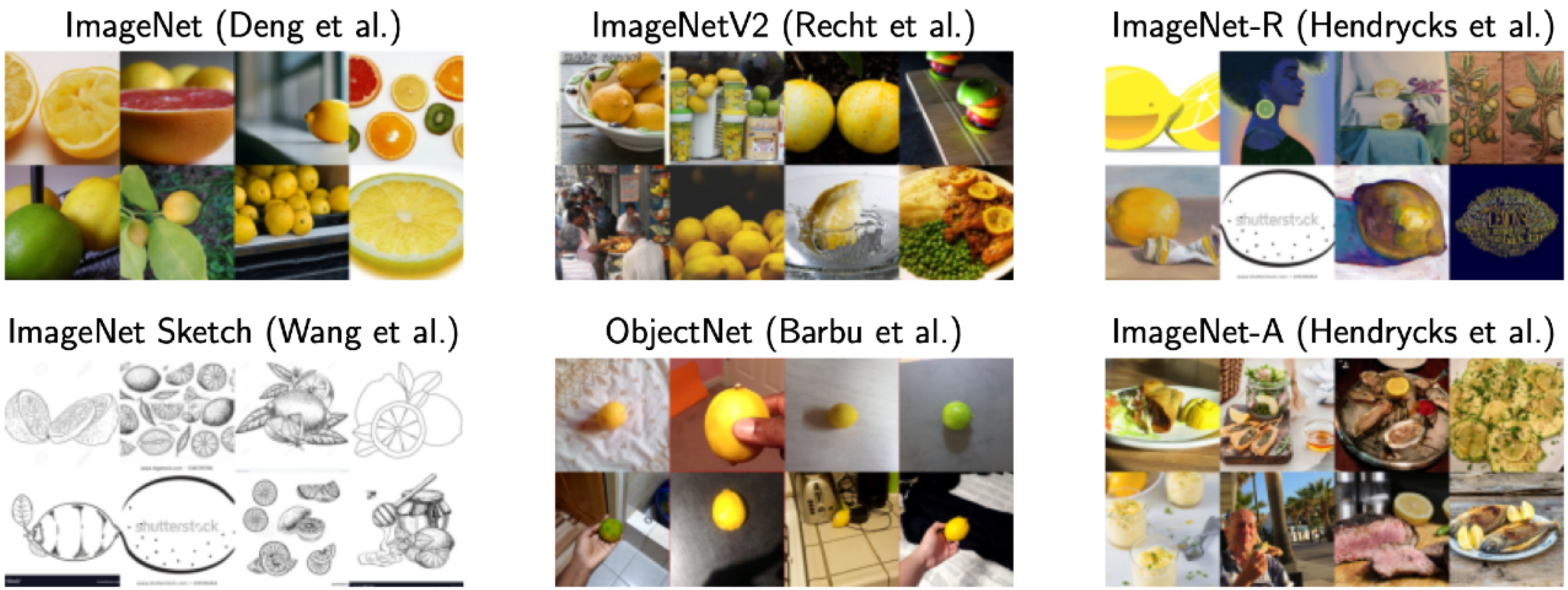
Experimental setup

- Distribution shifts
  - Taori et al. categorized distribution shifts into two broad categories
    - Synthetic : artificial change (contrast, brightness)
    - Natural : Naturally occurring variations in lighting, geographic location, crowdsourcing process, image styles etc.
- Effective robustness and scatter plots
  - The effective robustness framework introduced by Taori et al.
    - Effective robustness : Quantifies robustness as accuracy beyond a baseline trained on reference distribution.
  - Scatter plots display accuracy on the reference distribution on the x-axis and accuracy under distribution shift on the y-axis
    - Empirically, when applying logit axis scaling, models trained on the reference distribution approximately lie on linear trend.
- Zero-shot models and CLIP
  - Zero-shot models exhibit effective robustness and lie on a qualitatively different linear trend

Source:

Experimental setup

- Distribution shifts
  - Taori et al. categorized distribution shifts into two broad categories
    - Synthetic : artificial change (contrast, brightness)
    - Natural : Naturally occurring variations in lighting, geographic location, crowdsourcing process, image styles etc.



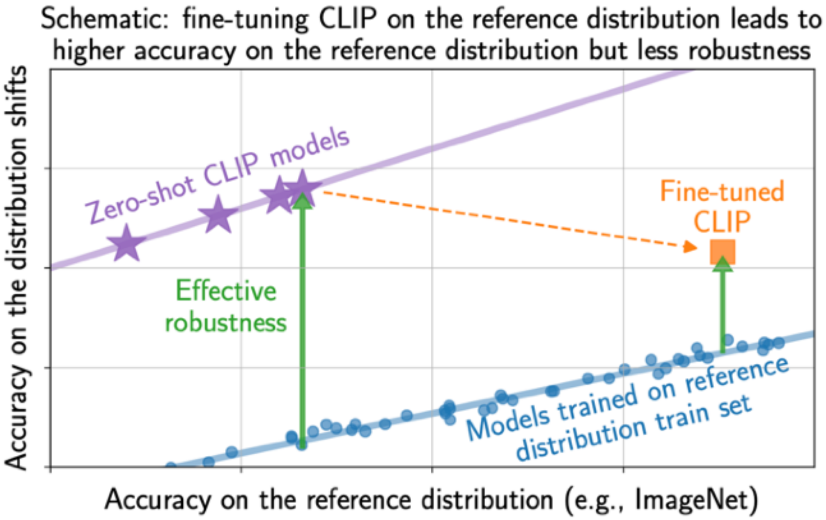
Source:

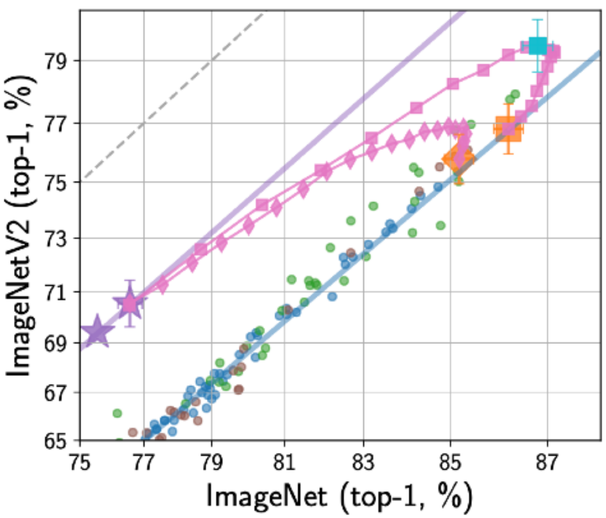
- Effective robustness and scatter plots
  - The effective robustness framework introduced by Taori et al.
    - Effective robustness

$$\rho(f) = \text{Acc}_{shift}(f) - \beta(\text{Acc}_{ref}(f))\,.$$

*When we say that a model is robust to distribution shift, we mean that effective robustness is positive.

  - Scatter plots

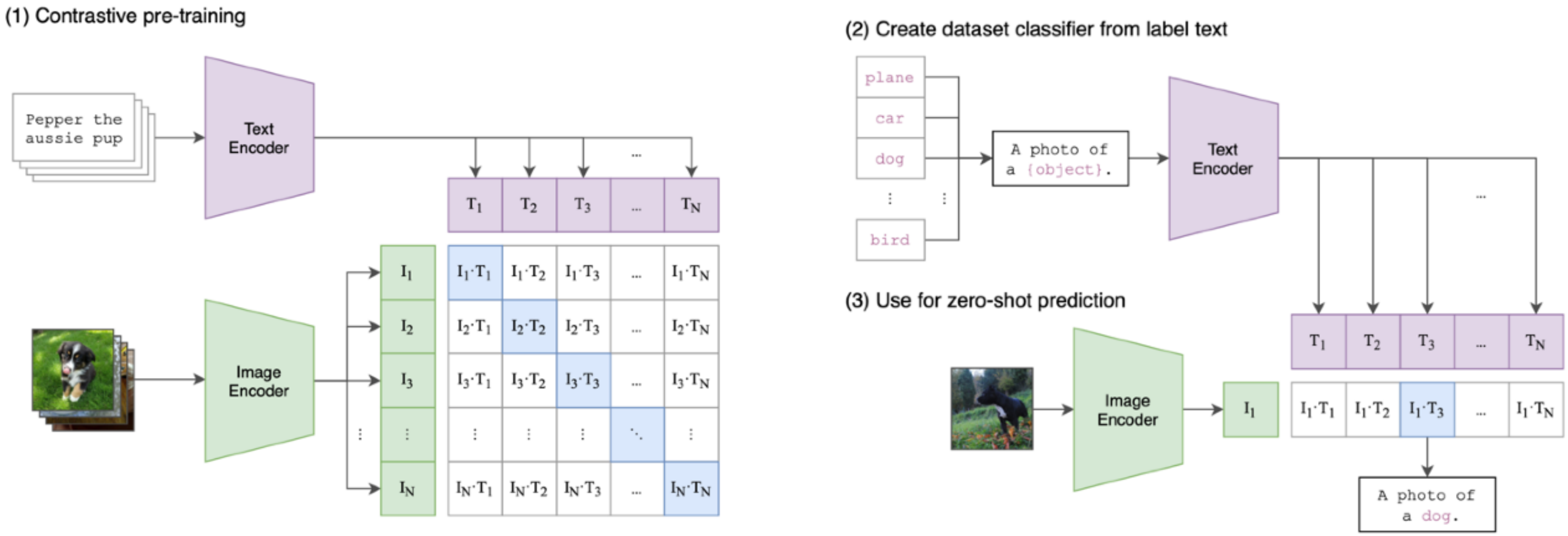




Source:

Figure 1 p2

- Zero-shot models and CLIP



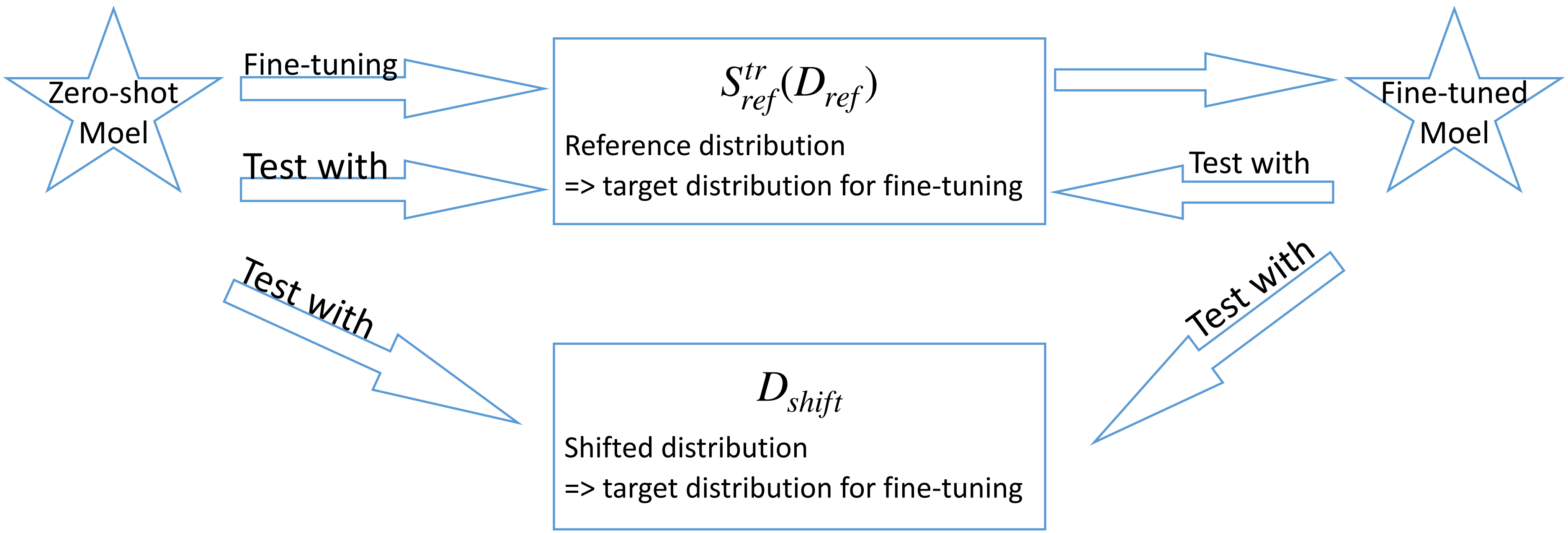
- CLIP-like models perform zero-shot k-way classification given an image x and class names C = $\{c_1, \ldots, c_k\}$ by matching x with potential captions.
- Using caption s_i = "a photo of a $\{c_i\}$" for each class I, the zero shot model predicts the class via argmax$_j$ $< g(x), h(s_j) >$

$$f(x) = g(x)^T \mathbf{W}_{zero-shot}$$

$$\mathbf{W}_{zero-shot} \in \mathbb{R}^{d \times k}$$

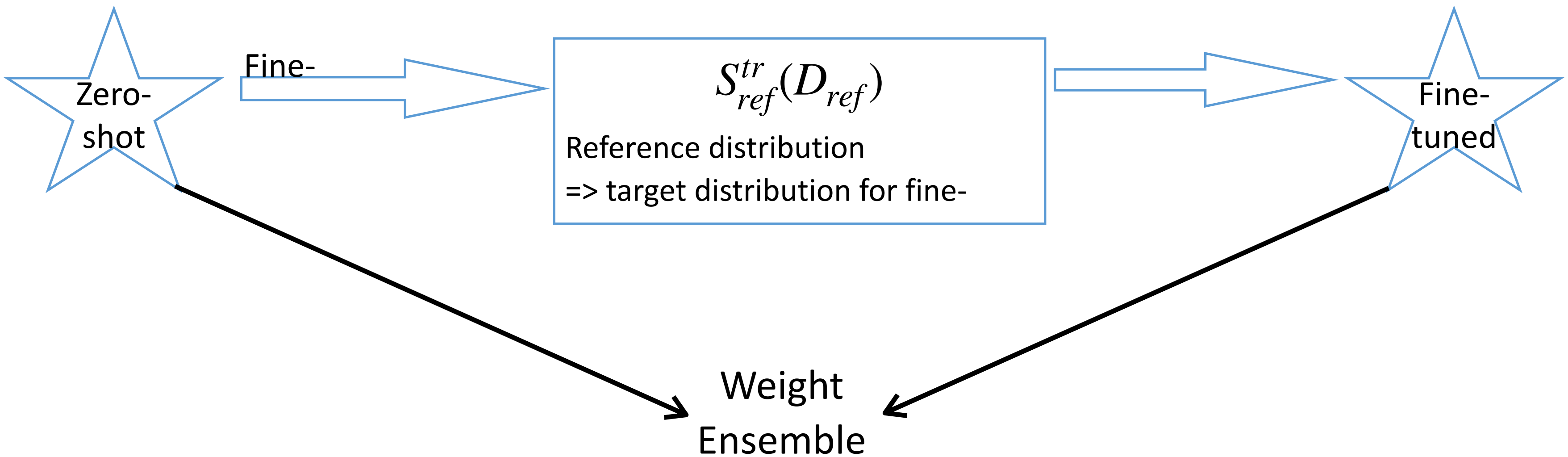
Source: [Learning Transferable Visual Models From Natural Language Supervision]

Experimental setup



Source:

Methods (Weight-space ensembles for fine-tuning)



- WiSE-FT consists of two simple steps
  - First, fine-tune the zero-shot model on application-specific data.
  - Second, combine the original zero-shot and fine-tuned models by "linearly interpolating between their weights"

Source:

Methods (Weight-space ensembles for fine-tuning)

- Step 1 : Standard fine-tuning
  - f(x, θ) = g(x, $\mathbf{V}_{enc}$)$\mathbf{W}_{classifier}$ where $\mathbf{W}_{classifier} \in \mathbb{R}^{d \times k}$
  - The parameters of f : $\theta = [\mathbf{V}_{enc}, \mathbf{W}_{classifier}]$
  - argmin$_\theta$\{ $\sum_{(x_i, y_i) \in S_{ref}^{tr}} l(f(x_i, \theta), y_i) + \lambda R(\theta)$\} , l : cross-entropy loss
  - Two most common variants of fine-tuning
    - ❖ end-to-end
    - ❖ fine-tuning only a linear classifier. => $\mathbf{V}_{\mathbf{enc}}$ is fixed
- Step 2 : Weight-space ensembling

$$wse(x, \alpha) = f(x, (1 - \alpha) \cdot \theta_0 + \alpha \cdot \theta_1)$$

α : mixing coefficient $\in [0,1]$

θ_0 : parameters of zero-shot model

θ_1 : parameters of fine-tuned model

Source:

Methods (Weight-space ensembles for fine-tuning)

- Weight-space ensembling

$$wse(x, \alpha) = f(x, (1 - \alpha) \cdot \theta_0 + \alpha \cdot \theta_1)$$

α : mixing coefficient $\in [0,1]$

θ_0 : parameters of zero-shot model

θ_1 : parameters of fine-tuned model

  - When fine-tuning only the linear classifier, weight-space ensembling is equivalent to the traditional output-space ensemble

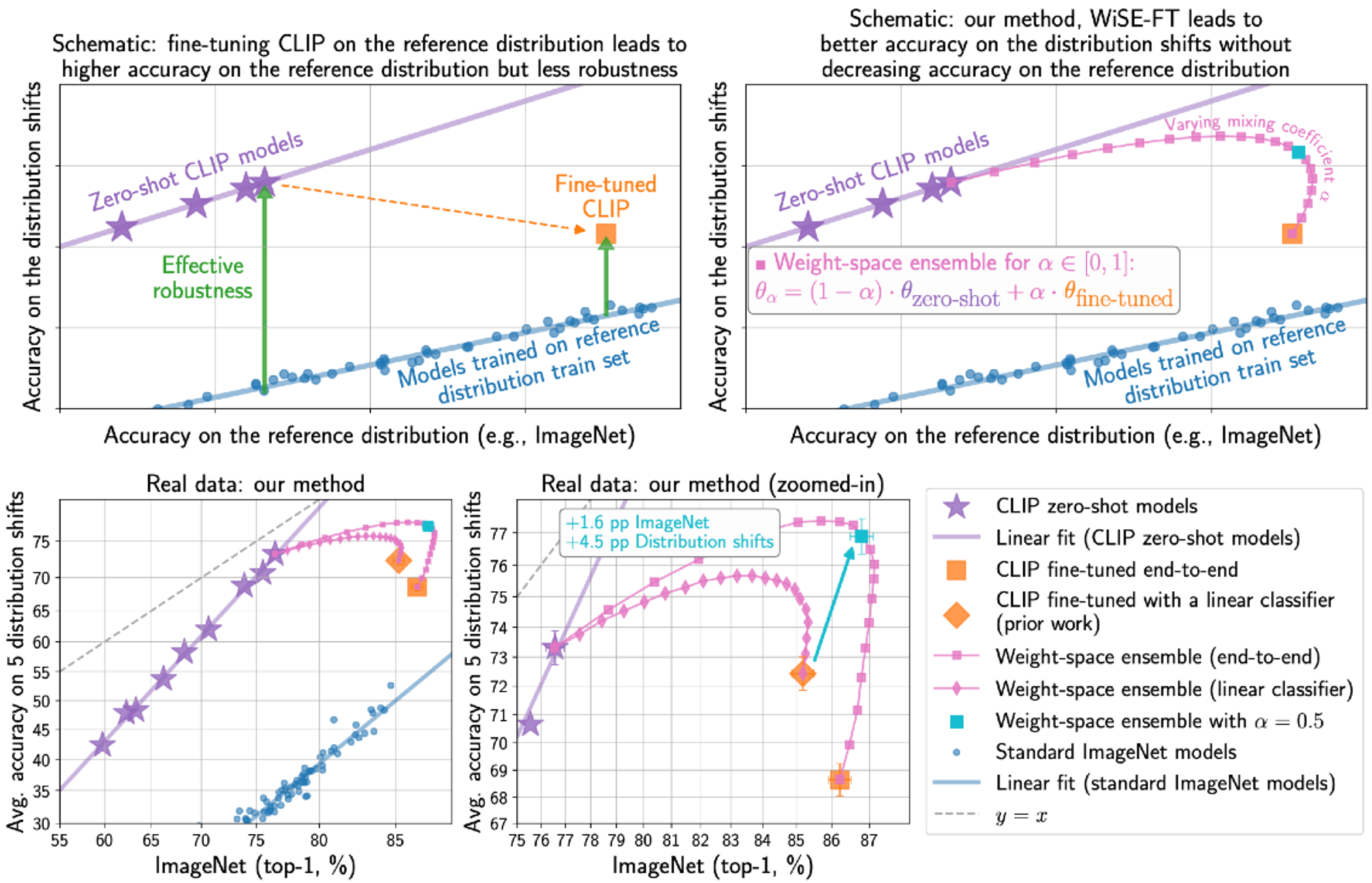
$$(1 - \alpha) \cdot f(x, \theta_0) + \alpha \cdot f(x, \theta_1)$$

$$(1 - \alpha) \cdot g(x, \mathbf{V}_{enc})^T \mathbf{W}_{zero-shot} + \alpha \cdot g(x, \mathbf{V}_{enc})^T \mathbf{W}_{classifier}$$

Source:

Results

Main results : ImageNet and associated distribution shifts

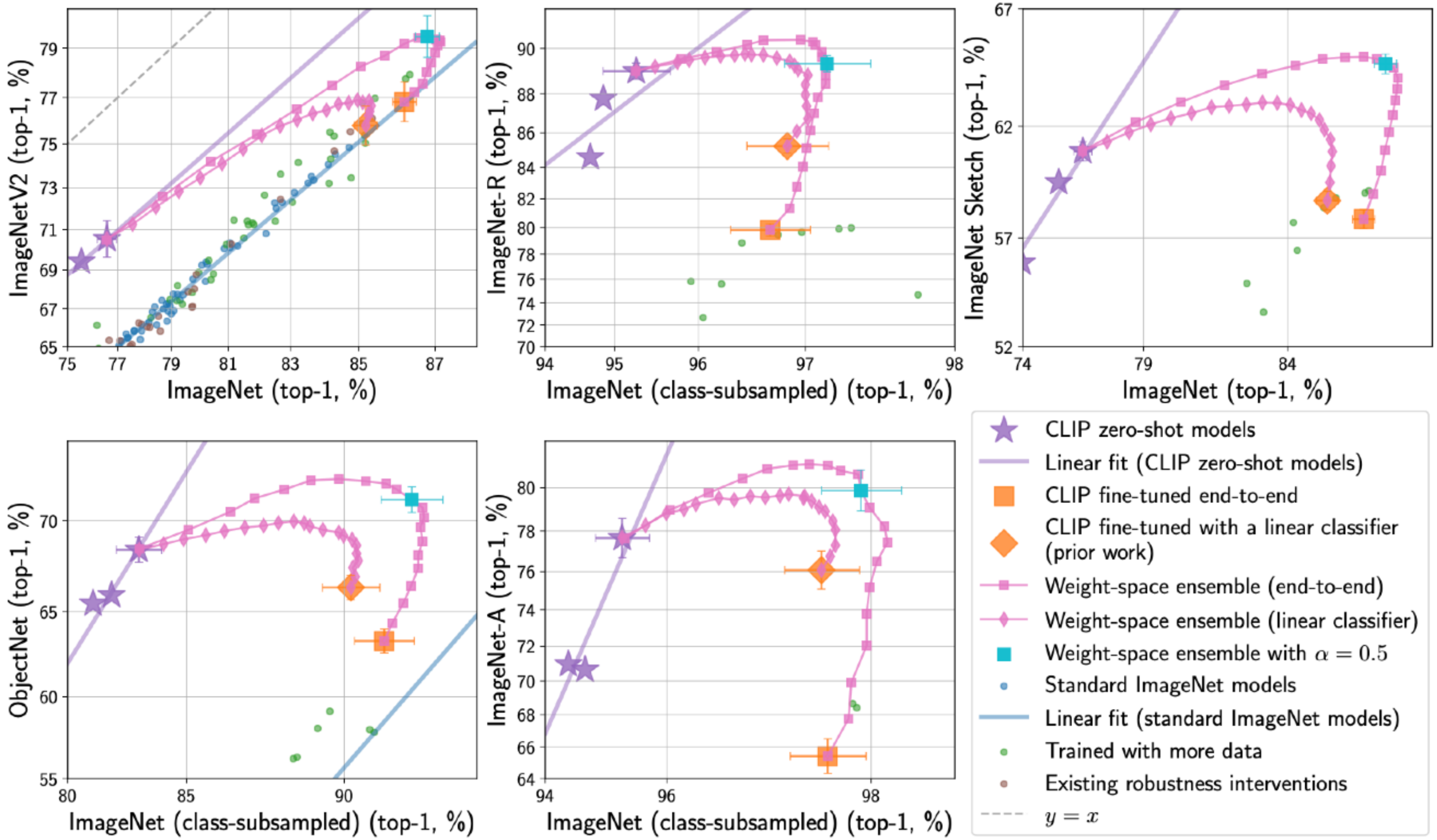


Source:

Figure 1 p2

Results

Main results : ImageNet and associated distribution shifts



Source:

Main results : ImageNet and associated distribution shifts

- Appendix B
  - You should find optimal α manually. Recommend using 0.5 when no domain knowledge is available.
  - There is no additional cost(train) when you find α.

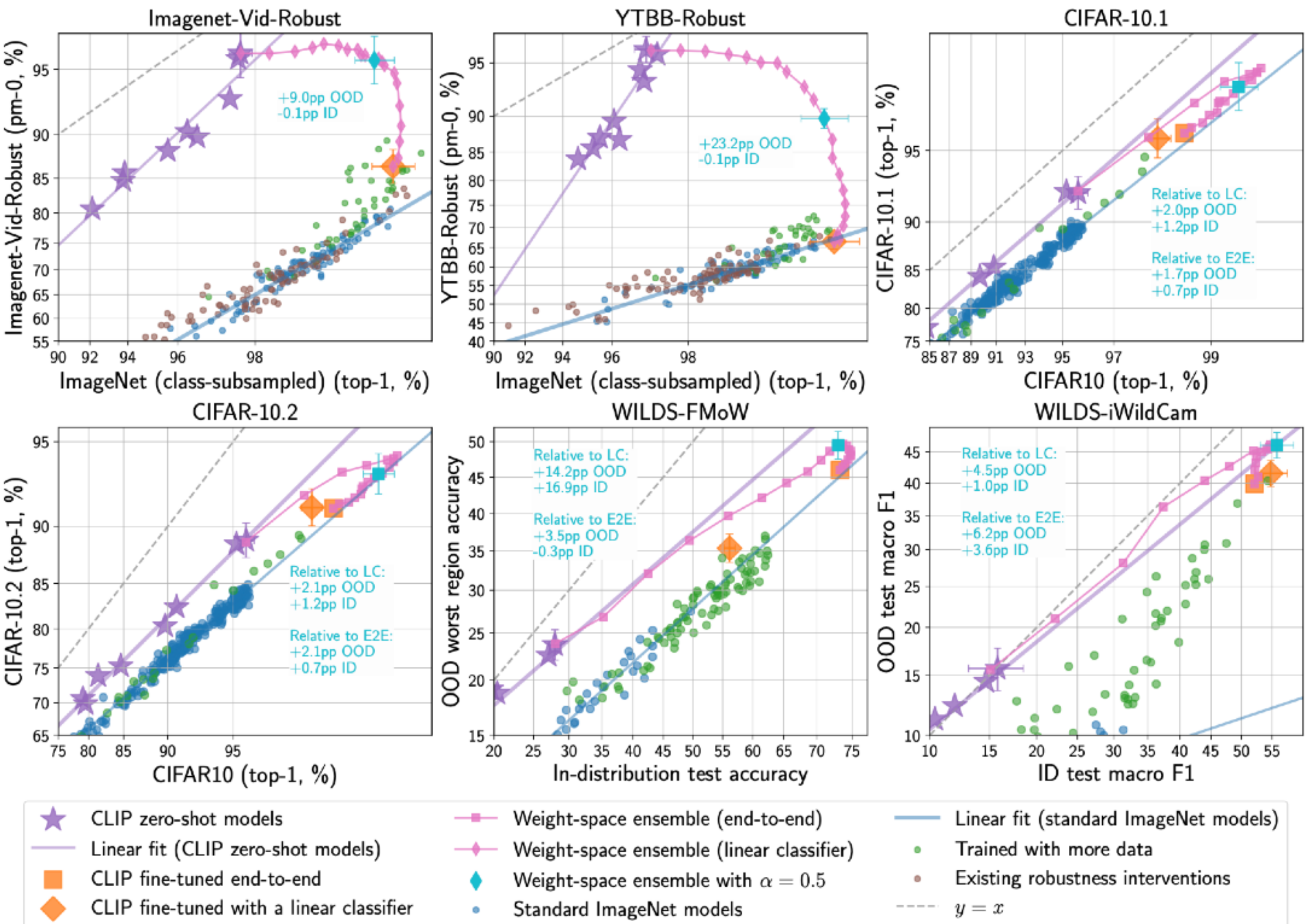
	IN (ref.)	Distribution shifts IN-V2	IN-R	IN-Sketch	ObjectNet	IN-A	Avg shifts	Avg ref., shifts
ViT-B/16, end-to-end	0.9	0.4	1.4	0.2	0.4	2.4	0.5	0.0
ViT-B/16, linear classifier	1.8	0.6	1.2	0.1	0.2	0.6	0.1	0.2
ViT-L/14@336, end-to-end	0.3	0.0	0.9	0.3	1.0	1.1	0.5	0.1
ViT-L/14@336, linear classifier	1.6	0.6	0.2	0.0	0.0	0.0	0.0	0.4

Table 3: Difference in performance (percentage points) between WiSE-FT using the optimal mixing coefficient and a fixed value of α=0.5 for CLIP ViT-B/16 and ViT-L/14@336. For each cell in the table, the optimal mixing coefficient α is chosen individually such that the corresponding metric is maximized. Results for all mixing coefficients are available in Tables 4 and 5. *Avg shifts* displays the mean performance among the five distribution shifts, while *Avg reference, shifts* shows the average of ImageNet (reference) and Avg shifts.

Source:

Results

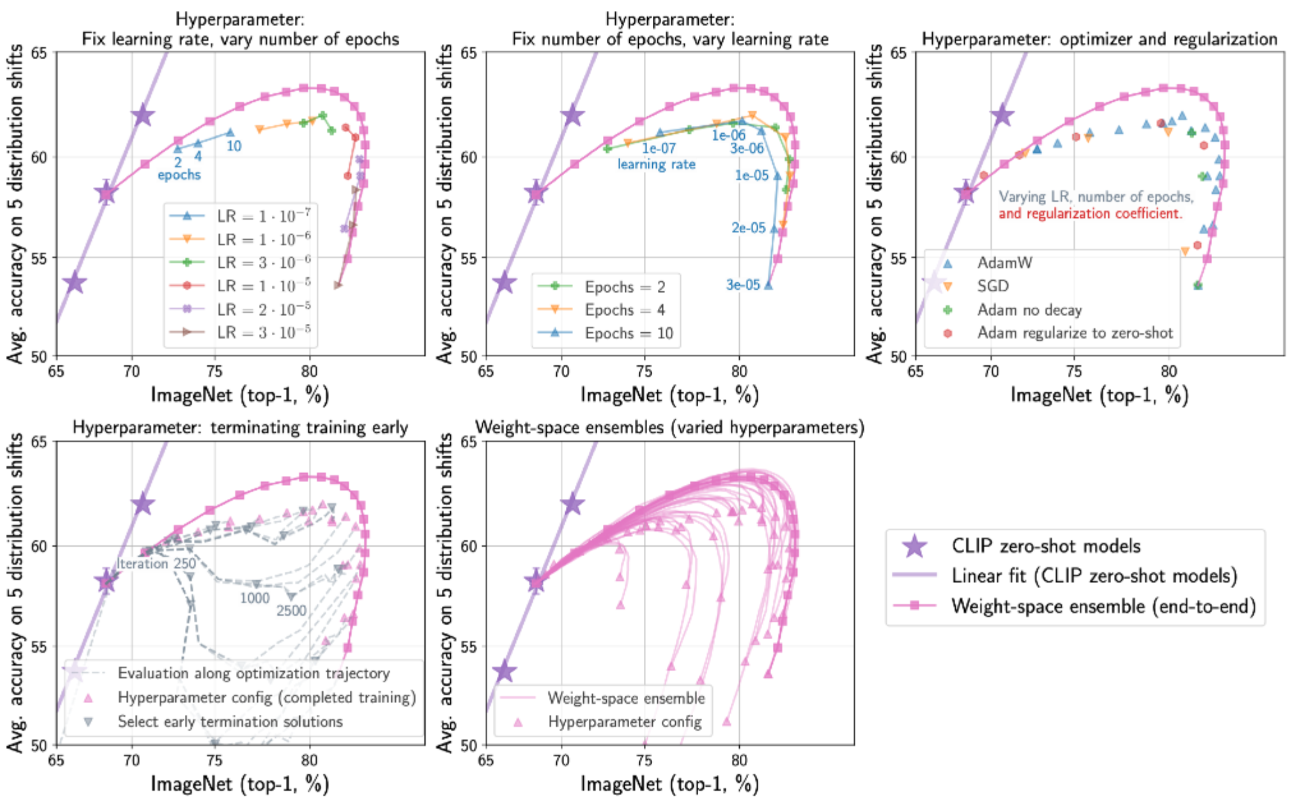
Robustness on additional distribution shifts



Source:

Hyperparameter variation and alternatives

- While there is a huge improvement of accuracy on distribution shift, it is not seen that much on reference distribution.
- Tuning the hyper-parameters on ImageNet dataset could deteriorate robustness.

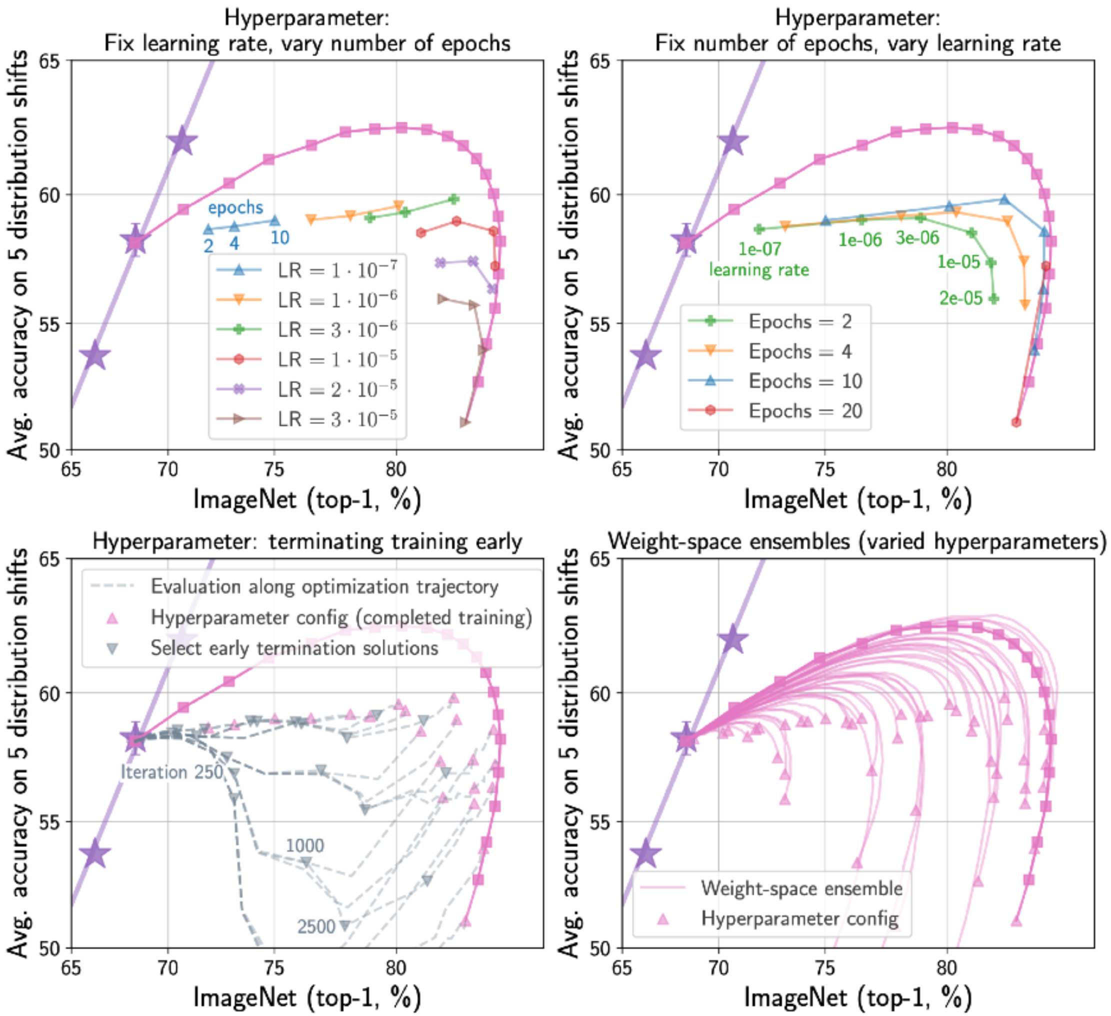


Source:

Figure 3 p8

Hyperparameter variation and alternatives

- C.4 Changes in data augmentation



Source:

Accuracy gains on reference distributions

- WiSE-FT has higher accuracy than fine-tuned model at the target distributions

	ImageNet	CIFAR10	CIFAR100	Cars	DTD	SUN397	Food101
Standard fine-tuning	86.2	98.6	92.2	91.6	81.9	80.7	94.4
WiSE-FT (α=0.5)	86.8 (+0.6)	99.3 (+0.7)	93.3 (+1.1)	93.3 (+1.7)	84.6 (+2.8)	83.2 (+2.5)	96.1 (+1.6)
WiSE-FT (opt. α)	87.1 (+0.9)	99.5 (+0.8)	93.4 (+1.2)	93.6 (+2.0)	85.2 (+3.3)	83.3 (+2.6)	96.2 (+1.8)

Table 2: Beyond robustness, WiSE-FT can improve accuracy after fine-tuning on several datasets.

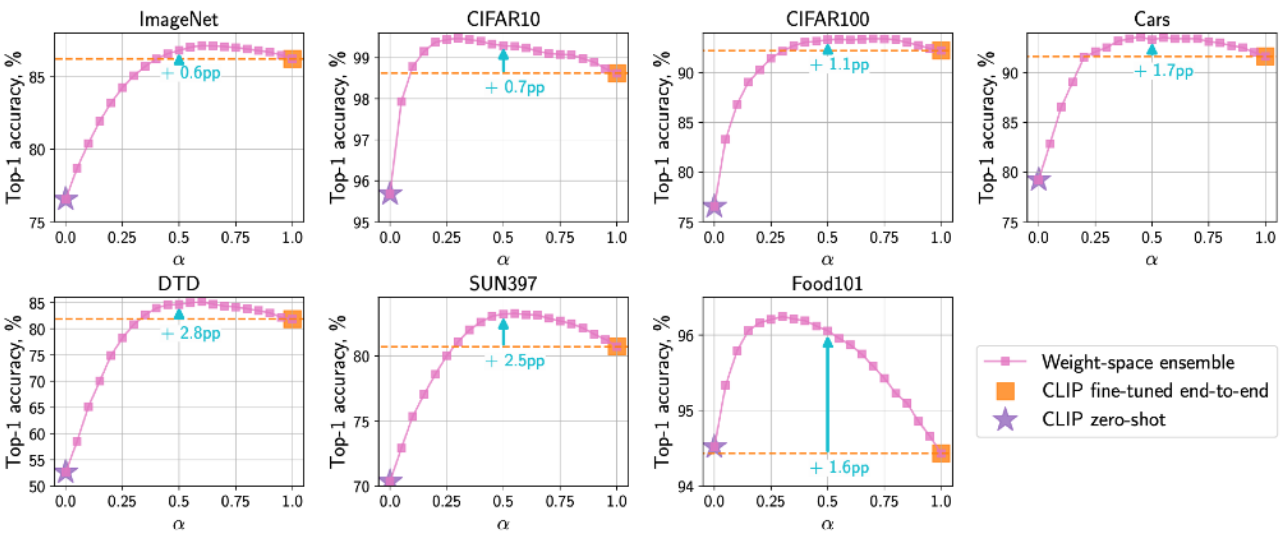


Figure 15: The accuracy of WiSE-FT (end-to-end) with mixing coefficient α on ImageNet and a number of datasets considered by Kornblith et al. [50]: CIFAR-10, CIFAR-100 [52], Describable Textures [14], Food-101 [10], SUN397 [101], and Stanford Cars [51].

Source:

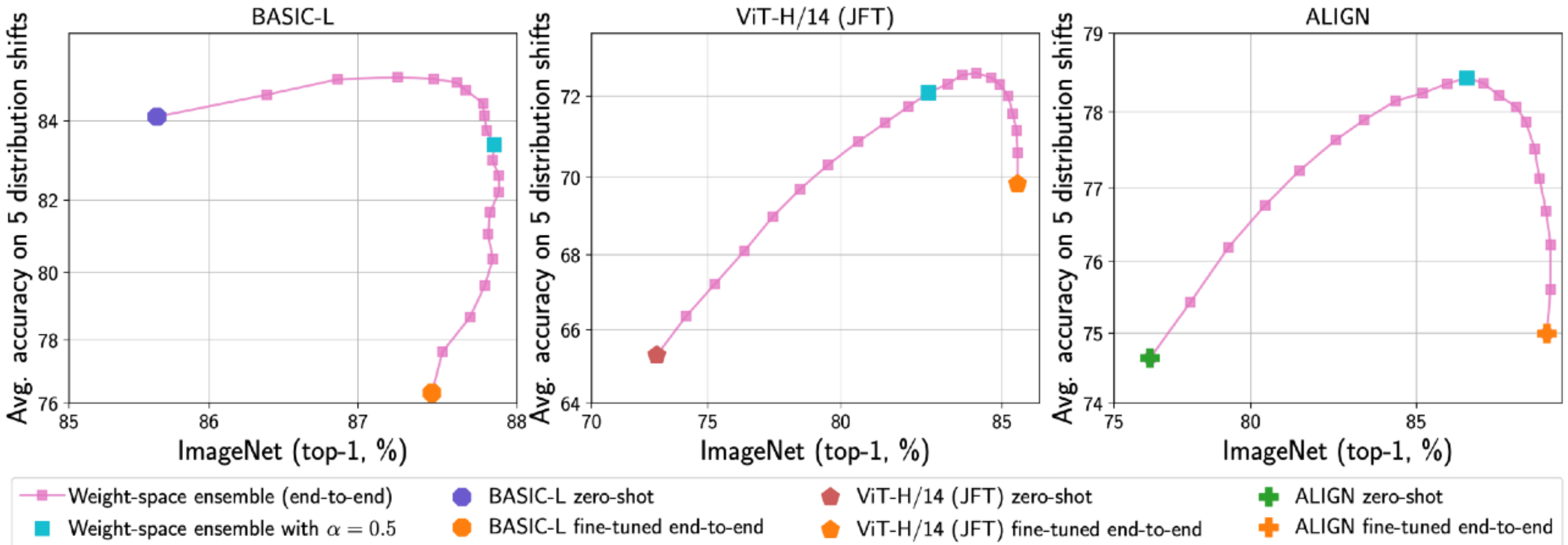


Figure 4: WiSE-FT applied to BASIC-L [75], a ViT-H/14 [21] model pre-trained on JFT-300M [91] and ALIGN [44].

Source:

Discussion & Conclusion

- Zero-shot and fine-tuned models are complementary
  - Zero-shot and fine-tuned models are diverse
  - Models are more confident where they excel
- An error landscape perspective
  - Observation 1
  - Observation 2

Source:

Zero-shot and fine-tuned models are complementary

- Zero-shot and fine-tuned models are diverse
  - Two measures of diversity
    - Prediction diversity

$$\mathrm{PD}(f, g, \mathcal{S}) = \frac{1}{N} \sum_{1 \leq i \leq N} \mathbb{1}\left[d_f \vee d_g\right],$$

$$d_f = \left(\hat{y}_f^{(i)} = y^{(i)} \wedge \hat{y}_g^{(i)} \neq y^{(i)}\right).$$

$$d_g = \left(\hat{y}_f^{(i)} \neq y^{(i)} \wedge \hat{y}_g^{(i)} = y^{(i)}\right).$$

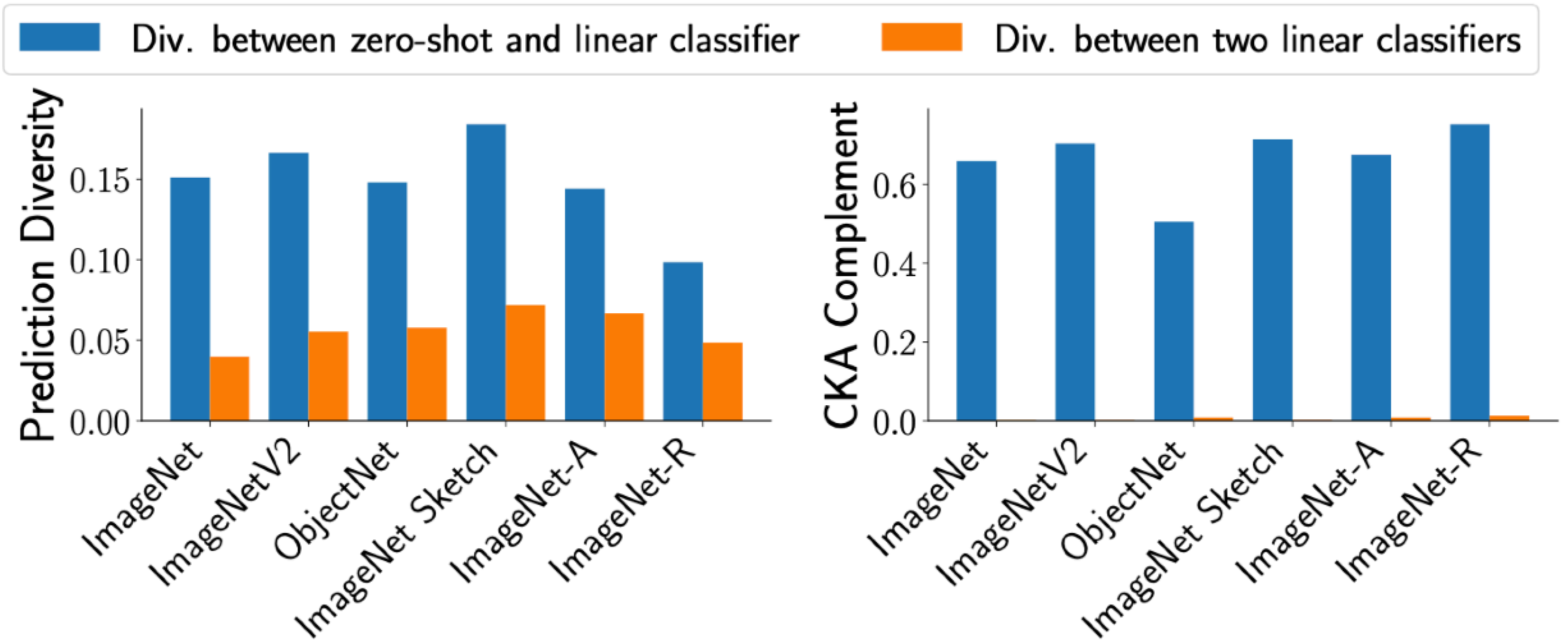
    - Centered Kernel Alignment Complement (CKA)

$$\mathrm{CKA}(f, g, \mathcal{S}) = \frac{||S_g^\top S_f||_F^2}{||S_f^\top S_f||_F ||S_g^\top S_g||_F},$$

Source:

Zero-shot and fine-tuned models are complementary

- Zero-shot and fine-tuned models are diverse

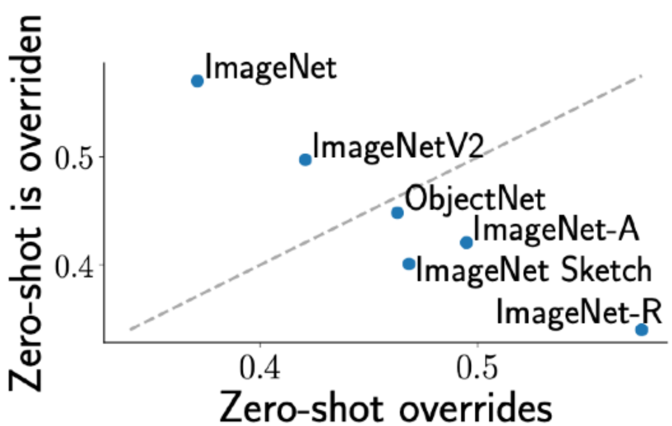


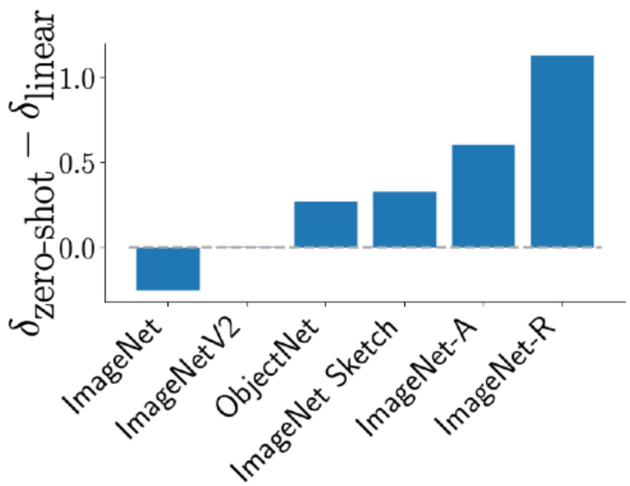
Source:

Figure 5 p10

Zero-shot and fine-tuned models are complementary

- Zero-shot and fine-tuned models are diverse
- Models are more confident where they excel
  - If zero-shot and fine-tuned model's prediction disagree and the zero-shot prediction matches with weight-space ensemble, we say the zero-shot model overrides.
  - When we say the zero-shot is overridden which means the opposite case to the above situation.
  - Measuring model confidence : the margin between the largest and second largest output of each classifier





Source: Figure 5 p10

Discussion & Conclusion

An error landscape perspective

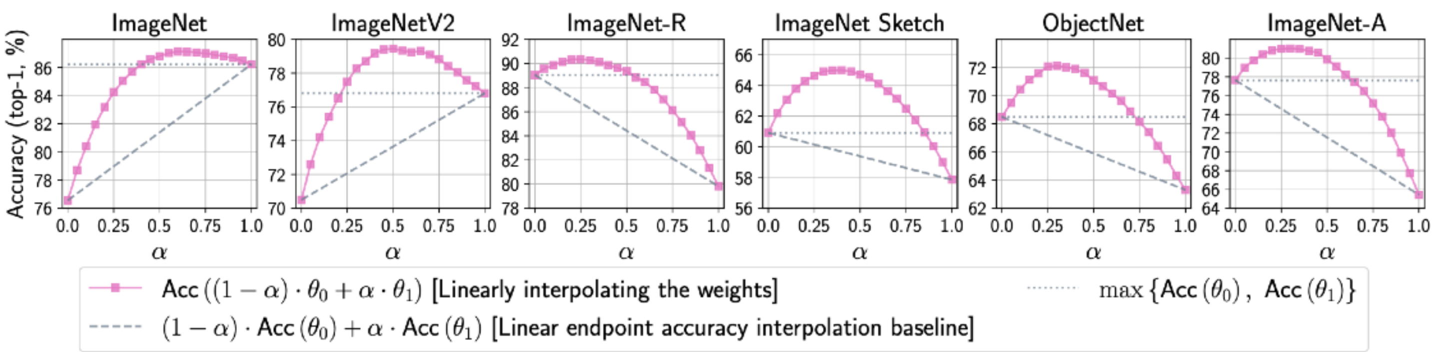
- Observation 1
  - Where the accuracy of both endpoints are similar, this equation is equivalent to the definition of Linear Mode Connectivity of Frankle et al.

$$\mathrm{Acc}_{\mathcal{D},f}((1-\alpha)\cdot\theta_0+\alpha\cdot\theta_1) \ \geq\ (1-\alpha)\cdot\mathrm{Acc}_{\mathcal{D},f}(\theta_0)+\alpha\cdot\mathrm{Acc}_{\mathcal{D},f}(\theta_1)$$

  - Linear mode connectivity has been observed, when
    - Part of the training trajectory is shared
    - Two models are fine-tuned with a shared initialization [Neyshabur et el.]

    => It may give us a clue about the reason why weight-space ensemble attain high accuracy.

- Observation 2
  - $\mathrm{Acc}_{\mathcal{D},f}\left((1-\alpha)\cdot\theta_0+\alpha\cdot\theta_1\right) \ \geq\ \max\left\{\mathrm{Acc}_{\mathcal{D},f}\left(\theta_0\right),\ \mathrm{Acc}_{\mathcal{D},f}\left(\theta_1\right)\right\}.$



Source: https://arxiv.org/pdf/1912.05671.pdf

Figure 6 p11

Discussion & Conclusion

An error landscape perspective

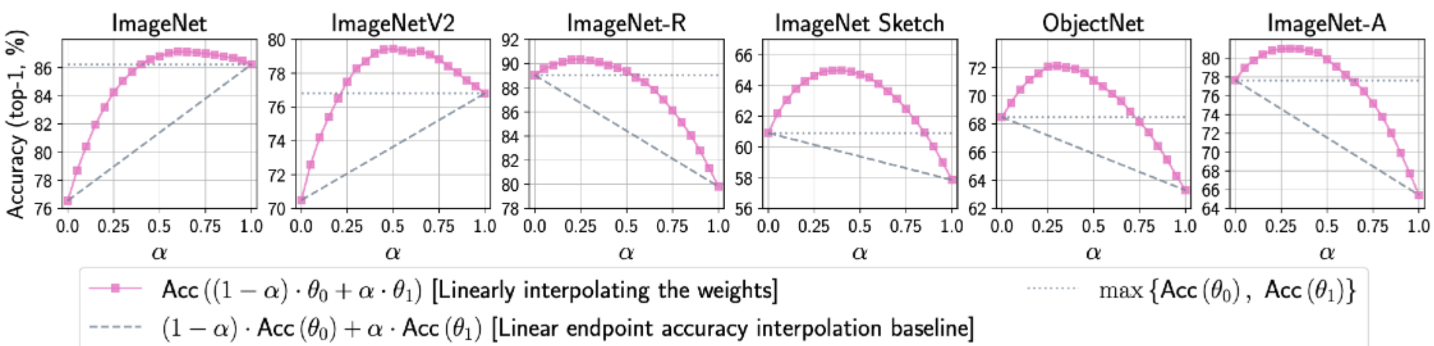
- Observation 1
  - ❖ Where the accuracy of both endpoints are similar, this equation is equivalent to the definition of Linear Mode Connectivity of Frankle et al.

$$\mathrm{Acc}_{\mathcal{D},f}((1-\alpha)\cdot\theta_0+\alpha\cdot\theta_1)\ \geq\ (1-\alpha)\cdot\mathrm{Acc}_{\mathcal{D},f}(\theta_0)+\alpha\cdot\mathrm{Acc}_{\mathcal{D},f}(\theta_1)$$

  - ❖ Linear mode connectivity has been observed, when
    - ➢ Part of the training trajectory is shared
    - ➢ Two models are fine-tuned with a shared initialization [Neyshabur et el.]

      => It may give us a clue about the reason why weight-space ensemble attain high accuracy.

- Observation 2
  - ❖ $\mathrm{Acc}_{\mathcal{D},f}\left((1-\alpha)\cdot\theta_0+\alpha\cdot\theta_1\right)\ \geq\ \max\left\{\mathrm{Acc}_{\mathcal{D},f}\left(\theta_0\right),\ \mathrm{Acc}_{\mathcal{D},f}\left(\theta_1\right)\right\}.$



Source: https://arxiv.org/pdf/1912.05671.pdf

(+α) Individual opinion

- What if our downstream task has a different number of class(or objective function/task) than the pre-training dataset?
- How much similar $\mathbf{W}_{zero-shot}$ (CLIP) and $\mathbf{W}_{classifier}$ (fine-tuned only linear classifier)

Source: