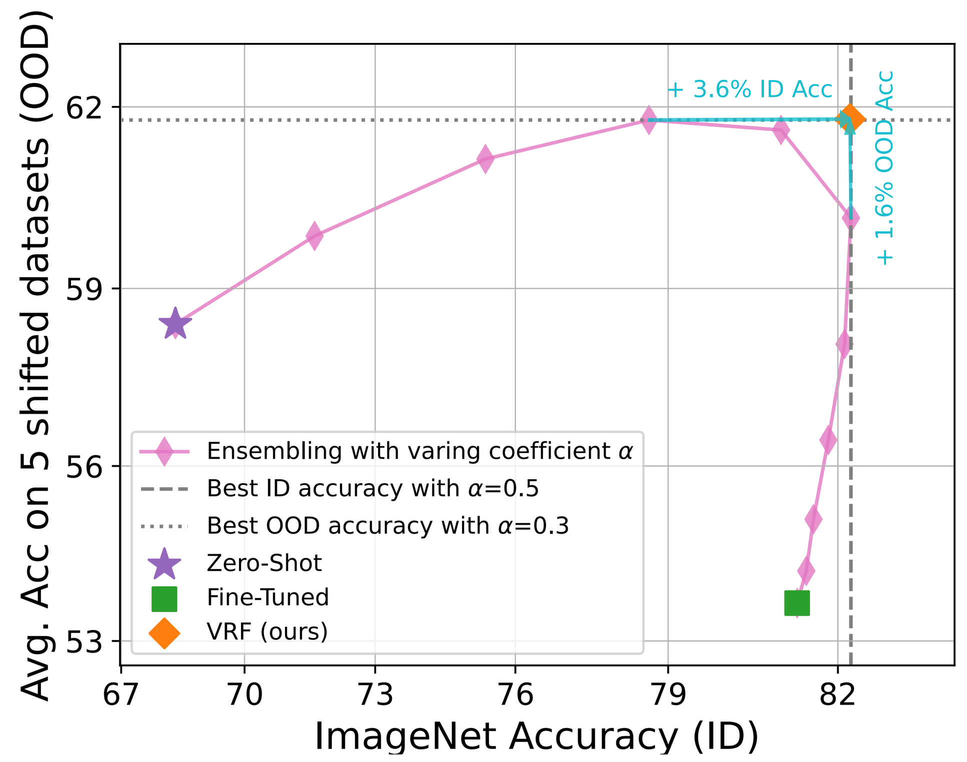
Robust Fine-tuning of Zero-shot Models via Variance Reduction

Beier Zhu, Jiequan Cui, Hanwang Zhang

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

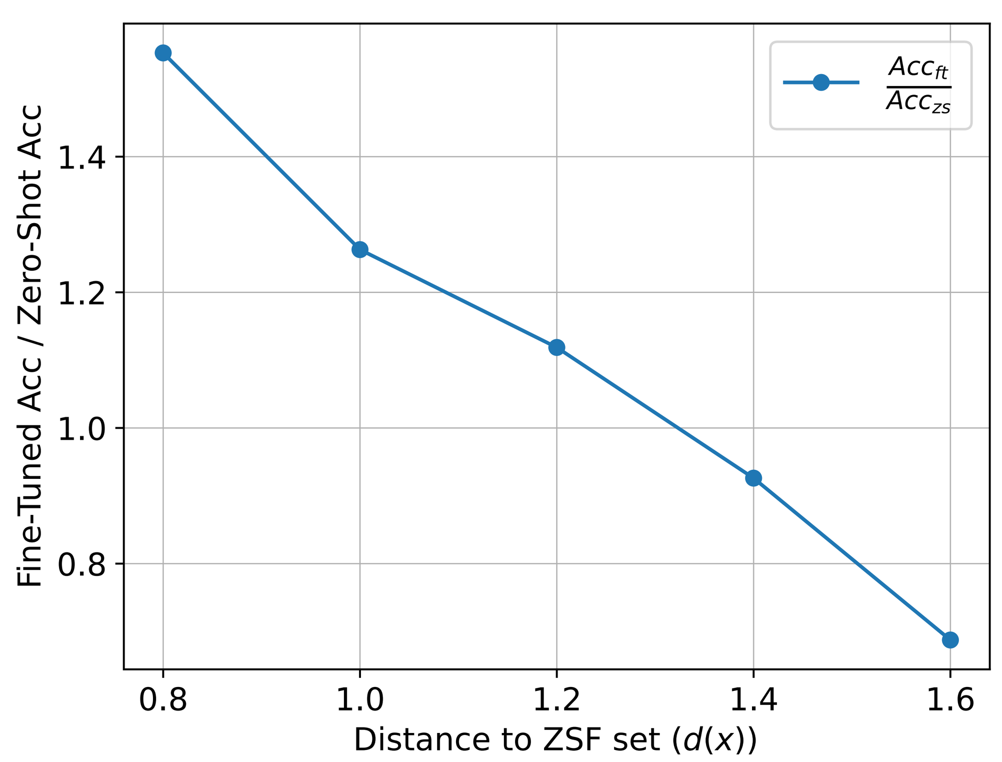
NEURAL INFORMATION PROCESSING SYSTEMS

ID-OOD Trade-offs



- Fine-tuning zero-shot models often compromises OOD performance. (■ excels in ID but lags in OOD compared to ★).
- Recently, ensemble-based models (ESMs) have shown great potential in addressing the ID-OOD dilemma.
- However, ESMs ♦ cannot fully solve the ID-OOD trade-offs: they achieve peak performance for ID and OOD accuracy at different mixing coefficients (best ID at $\alpha = 0.5$, best OOD at $\alpha = 0.3$.)

Intriguing Finding



- Zero-Shot Failure (ZSF) set: for each training sample, if the fine-tuned model correctly predicts the label while the zero-shot model fails, we collect its feature representation.
- We measure the distance of each test sample to the ZSF set. Based on this distance, test samples are grouped into bins, and we compute the ratio of fine-tuned accuracy to zero-shot accuracy $\frac{\mathrm{Acc}_{\mathrm{ft}}}{\mathrm{Acc}_{\mathrm{zs}}}$.
- **Finding**: the ratio monotonically decreases as distance increases.

Method

Core Idea: using the distance to assign weights in ensembling -- a smaller distance results in a higher weight for the fine-tuned model, and vice versa.

Given: Training dataset $\mathcal{D}$, a zero-shot model f_{zs}, and a fine-tuned model f_{ft}.

Step 1 (Identification). We build the zero-shot failure set as

$$\mathcal{V} = \{\mathbf{v}_i \text{ s.t. } y_i = \mathrm{pred}(f_{\mathrm{ft}}(\mathbf{x}_i)) \text{ and } y_i \neq \mathrm{pred}(f_{\mathrm{zs}}(\mathbf{x}_i))\}$$

where $\{\mathbf{x}_i, y_i\} \in \mathcal{D}$, $\mathbf{v}_i$ is the feature representation of $\mathbf{x}_i$.

Step 2 (Distance Calculation). The distance of a test sample to $\mathcal{V}$ is defined as the l_2 distance to the k-th nearest neighbor in $\mathcal{V}$

$$d(\mathbf{x}) = \|\mathbf{v} - \mathbf{v}_{(k)}\|_2$$

Step 3 (Sample-Wise Ensembling). We implement sample-wise output-space in the form:

$$\widehat{\mathbb{P}}_{\mathrm{vrf}}(y|\mathbf{x}) = \omega(\mathbf{x})\widehat{\mathbb{P}}_{\mathrm{ft}}(y|\mathbf{x}) + (1-\omega(\mathbf{x}))\widehat{\mathbb{P}}_{\mathrm{zs}}(y|\mathbf{x}),$$

where $\omega(\mathbf{x}) = \sigma(-(d(\mathbf{x}) - a)/b)$, $\sigma(\cdot)$ is the sigmoid function and a, b are two hyperparameters.

Justification

The probability output of a classifier parameterized by θ can be expressed as:

$$\widehat{\mathbb{P}}(y|\mathbf{x};\theta) = \mathbb{P}(y|\mathbf{x}) + \eta_y(\mathbf{x})$$

where $\mathbb{P}(y|\mathbf{x})$ denotes the true *a posterior and* $\eta_y(\mathbf{x})$ is the error term. The expected error of the estimated classifier is:

$$E = \frac{\mathbb{V}[\eta_y(\mathbf{x})]}{s},$$

where s is a constant factor related to the derivative of the true a posterior distribution and is independent of the trained model, and $\mathbb{V}[\eta_y(\mathbf{x})]$ is the variance.

Let $g_{\mathrm{zs}}(\cdot)$ and $g_{\mathrm{ft}}(\cdot)$ be two functions that produce weights for ensembling the models. Subject to the constraint that $g_{\mathrm{zs}}(\mathbf{x}) + g_{\mathrm{ft}}(\mathbf{x}) = 1$, the variance of our model can be expressed as:

$$\mathbb{V}[\eta_{\mathrm{vrf}}(\mathbf{x})] = g_{\mathrm{zs}}(\mathbf{x})^2\mathbb{V}[\eta_{\mathrm{zs}}(\mathbf{x})] + g_{\mathrm{ft}}(\mathbf{x})^2\mathbb{V}[\eta_{\mathrm{ft}}(\mathbf{x})].$$

To obtain the minimal variance, the optimal weight function should be

$$g_{\mathrm{ft}}(\mathbf{x}) = \frac{\mathbb{V}[\eta_{\mathrm{zs}}(\mathbf{x})]}{\mathbb{V}[\eta_{\mathrm{zs}}(\mathbf{x})] + \mathbb{V}[\eta_{\mathrm{ft}}(\mathbf{x})]} = \frac{E_{\mathrm{zs}}}{E_{\mathrm{zs}} + E_{\mathrm{ft}}} \propto \frac{\mathrm{Acc}_{\mathrm{ft}}}{\mathrm{Acc}_{\mathrm{zs}}}$$

Results

Table 1: Accuracy of various methods on ImageNet and derived distribution shifts for CLIP ViT-B/32

Method	IN	Distribution shifts					Avg
		IN-V2	IN-Sketch	IN-A	IN-R	ObjectNet	shifts
Zero-shot [20]	63.3	55.9	42.3	31.5	69.3	43.5	48.5
Linear classifier [20]	75.4	63.4	38.8	26.1	58.7	41.5	45.7
E2E-FT [28]	76.2	64.2	38.7	21.0	57.1	40.1	44.2
+ Weight-space ensemble [28]	77.9	67.2	45.1	28.8	66.4	45.1	50.5
+ Output-space ensemble	77.3	66.0	44.2	27.1	68.4	44.4	50.0
+ VRF (ours)	77.6	66.7	47.0	29.2	70.9	46.3	52.0
Δ	+0.3	+0.7	+2.8	+2.1	+2.5	+1.9	+2.0
LP-FT [15]	76.9	64.8	39.9	25.7	69.9	42.6	48.6
+ Weight-space Ensemble [28]	78.0	67.0	44.8	31.2	65.8	46.1	51.0
+ Output-space Ensemble	77.8	66.3	44.0	29.5	66.2	45.5	50.3
+ VRF (ours)	77.8	66.7	46.1	31.0	70.0	46.3	51.8
Δ	+0.0	+0.4	+2.1	+1.5	+3.8	+0.8	+1.5

We observe that our VRF boosts the accuracy of fine-tuned models, including ensembling baseline models, across five ImageNet distribution shifted datasets, while maintaining or improving the ImageNet in-distribution performance.