Predicting the Performance of Foundation Models via Agreement-on-the-Line

Aman Mehra[1]*, Rahul Saxena[1]*, Taeyoun Kim[1]*, Christina Baek[1], Zico Kolter[1,2], Aditi Raghunathan[1]

[1]Carnegie Mellon University, [2] Bosch Center for AI

Problem Motivation

Foundation Models (FMs) are pre-trained on vast amounts of open world data, and then fine-tuned on task-specific data. Given a set of foundation models pre-trained on *potentially* different data

$$\mathscr{B} = \{B_1, B_2, \ldots, B_m\} \text{ s.t. } B_i : \mathbb{X} \to \mathbb{R}^d$$

We obtain a collection of models $\mathscr{H} = \{h_1, h_2, \ldots, h_n\}$ by fine-tuning models in $\mathscr{B}$ on in-distribution (ID) data

$$X_{train}, y_{train} \sim \mathscr{D}_{ID} \text{ s.t. } h_i : \mathbb{X} \to \mathbb{Y}$$

Say we are provided *labeled* ID validation data and *unlabeled* OOD test data

$$X_{val}, y_{val} \sim \mathscr{D}_{ID} \text{ and } X_{test} \sim \mathscr{D}_{OOD}$$

Can we predict the OOD performance of models in $\mathscr{H}$ without access to OOD labels?

Agreement-on-the-line (AGL)

For two models h_1, h_2, some distribution $\mathscr{D}$, and an appropriate

$$\text{Acc}(h_1) = \mathbb{E}_{x,y\sim\mathscr{D}}[l(h_1(x), y)]$$
$$\text{Agr}(h_1, h_2) = \mathbb{E}_{x\sim\mathscr{D}}[l(h_1(x), h_2(x)]$$

Empirically, in **CNN's trained from scratch**, when ID vs OOD accuracy is strongly linearly correlated [1], ID vs OOD agreement is also strongly linearly correlated with the same slope and bias [2].

$$\text{Acc}_{\text{OOD}}(h_1, h_2) = a \cdot \text{Acc}_{\text{ID}}(h_1, h_2) + b$$
$$\text{Agr}_{\text{OOD}}(h_1, h_2) = c \cdot \text{Agr}_{\text{ID}}(h_1, h_2) + d$$
$$\text{where} \quad a \approx c, b \approx d$$

ALine [2] linearly transforms ID accuracy using c, d to estimate OOD accuracy.

Using ALine methods for OOD estimation

In **diverse** ensembles, ALine is a good method to estimate OOD accuracy for many datasets and tasks.

Table 1. Mean Absolute Percentage Error of OOD performance estimates.

Dataset		ALine-D	ALine-S	Agr	ATC [3]	AC [4]	Doc [5]
CIFAR-10C		6.99	**6.92**	44.33	31.28	48.66	32.79
CIFAR-10.1		**2.42**	3.03	41.52	6.48	54.57	8.51
CIFAR-100C		**11.94**	12.67	46.13	18.69	80.81	37.36
ImageNetC		**10.91**	11.04	56.76	27.25	79.00	37.86
ImageNetV2		**4.96**	5.03	47.65	8.96	77.34	7.86
WILDS	fMoW	**2.59**	2.74	83.94	9.03	44.59	5.86
	iWildCam	**22.05**	25.29	46.42	37.25	57.31	69.58
OfficeHome		**10.01**	12.78	49.76	31.89	79.98	35.92
SQuAD Shifts	Reddit	**1.20**	2.6	20.21	12.74	49.25	6.09
	Amazon	**1.64**	3.10	20.40	15.35	51.06	7.39
	NYT	**0.82**	1.33	18.46	3.11	38.61	3.18
	NewWiki	**3.08**	3.18	18.87	5.46	41.26	1.50
MNLI-MM		0.49	0.51	6.19	0.55	**0.47**	0.59
SNLI		2.43	**1.88**	11.60	3.38	5.19	4.96

AGL when Fine-Tuning a Single Foundation Model

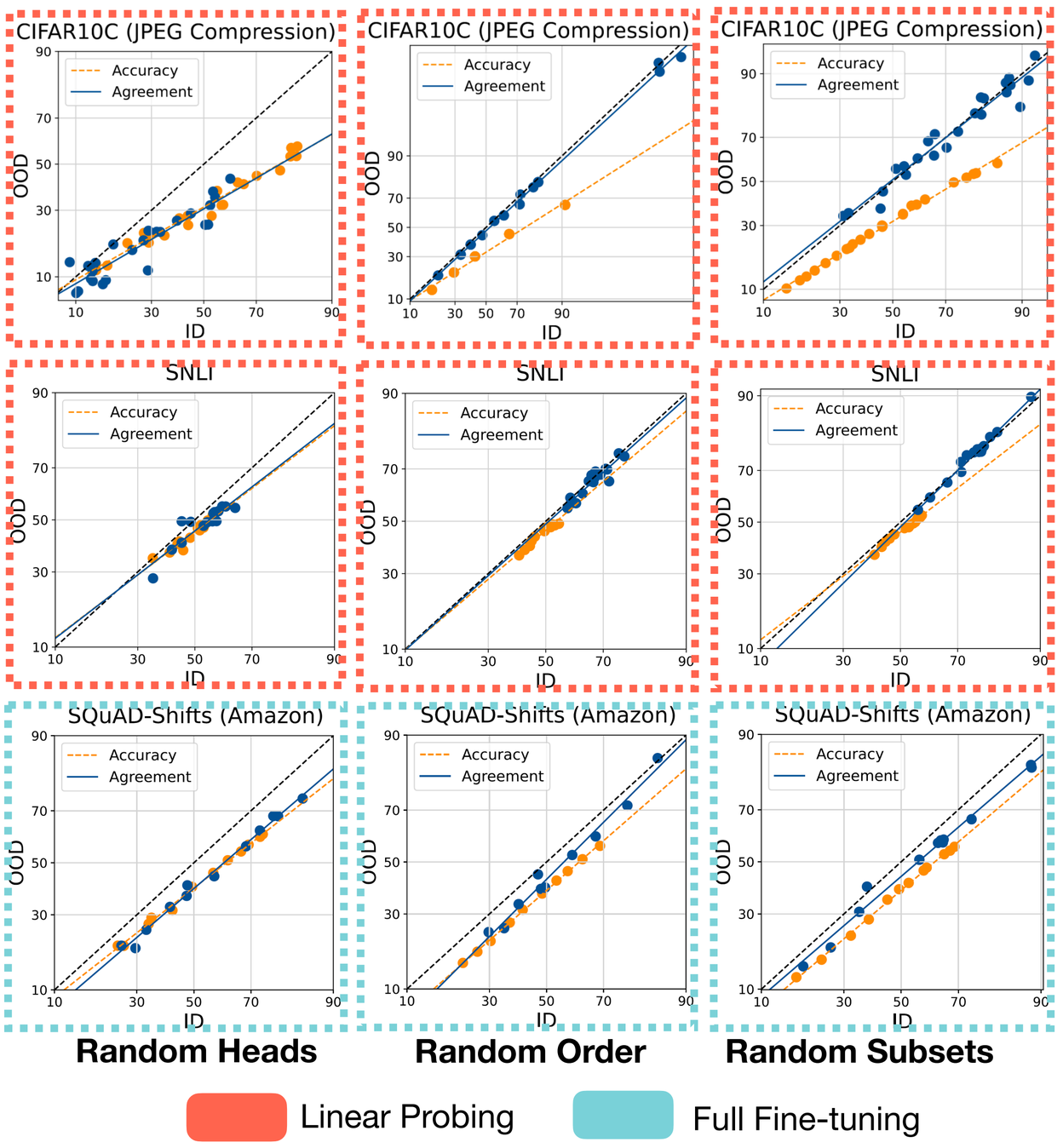
Do models obtained by *lightly* fine-tuning a single FM have diverse enough predictions to observe AGL?

- Light fine-tuning over the same foundation model may yield models that agree too highly ID and/or OOD as they're pre-trained on the same corpus
- Thus AGL might not hold due to lack of diversity in predictions

Given a single FM, we introduce this diversity by training an ensemble of fine-tuned models that are:

1. Initialized with **Random linear heads**
2. Presented with the training data in **Random order**
3. Trained on independently sampled **Random subsets** of the data

*Only FMs fine-tuned with random heads exhibit **enough diversity OOD** to observe AGL!*



AGL when Fine-Tuning Multiple Foundation Models

*Agreement between models fine-tuned from **different base FMs** have enough diversity to exhibit **AGL!***

References

1. John Miller, et. al. **Accuracy on the line: On the strong correlation between out-of-distribution and in-distribution generalization.** ICML, 2021.
2. Christina Baek, et al. **Agreement-on-the-line: Predicting the performance of neural networks under distribution shift,** Neurips 2022.
3. Benjamin Recht, et. al. **Do ImageNet classifiers generalize to ImageNet?,** ICML 2019.
4. Dan Hendrycks and Thomas G. Dietterich. **Benchmarking neural network robustness to common corruptions and perturbations.** ICLR, 2019.
5. Gordon Christie, et al. **Functional map of the world.** CVPR, 2018.