Benchmarking Low-Shot Robustness to Natural Distribution Shifts

Aaditya Singh*†, Kartik Sarangmath*, Prithvijit Chattopadhyay, Judy Hoffman
Georgia Institute of Technology

ICCV23 PARIS

1 Highlights

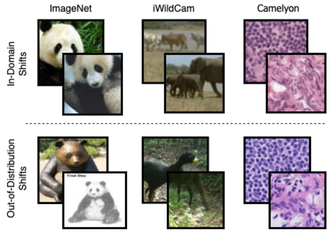
- **Low-Shot Robustness:** Study the effect of limited (~10^4 images) in-domain (ID) data on out-of-distribution (OOD) robustness
- ***No single model or initialization is consistently more robust*** across multiple datasets and natural shifts
- ***Existing robustness interventions often fail to improve robustness*** across different data regimes

2 Motivation

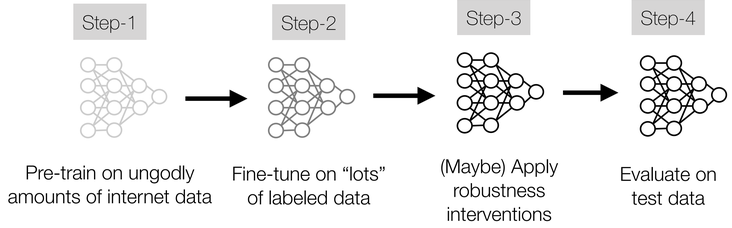
- **Task of interest:** Visual classification under natural distribution shifts

Train / Fine-tune model on in-domain data

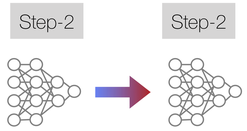
Test model out-of-the-box on OOD data



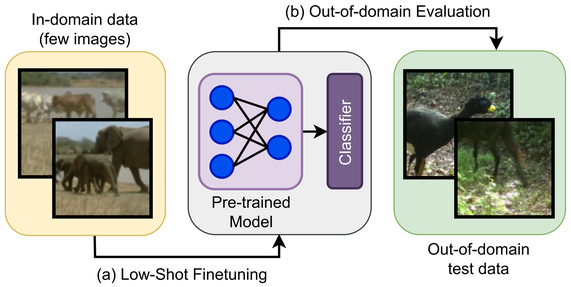
- **Standard paradigm:** Existing works for evaluating and improving robustness almost always assume access to a large amount (~10^6) of in-domain (ID) data



- ***The extent to which the current observations hold for a lesser amount of ID data remains unknown***
- **Low-Shot Robustness Setting:** Pre-trained model with a classifier head is trained on few ID images



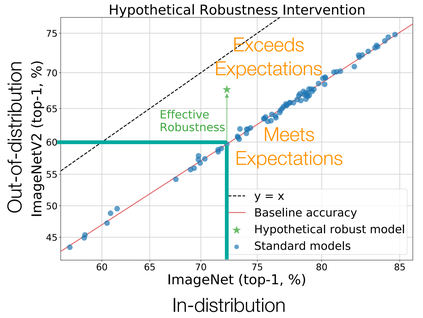
3 Low-Shot Robustness Setting



- We consider **three distinct natural distribution shifts**
- We consider **three low-shot regimes of increasing difficulty**

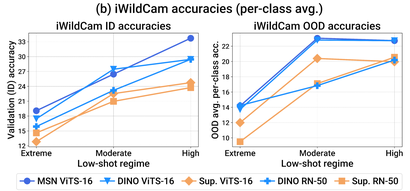
Dataset	Low-Shot Regimes (Imgs / Class)		
	Extreme	Moderate	High
1 ImageNet [5]	1	5	~ 13
2 iWildCam [18]	1-480	1-4802	1-9604
3 Camelyon [20]	1500	7500	15000

- **Models and Interventions to Improve Robustness**:
  - **ImageNet (IN1k) pre-trained models:** Supervised ResNet and ViT (DEIT), Self-supervised (SSL) ResNets (SwAV, DINO) and ViTs (MSN, DINO)
  - **Classifier heads:** Logistic Regression, Mean-Centroid Classifier, Baseline++
  - **Robustness interventions:** LP-FT, CLIP, WiSE-FT, Model Soups, RobustViT
- We use the **effective robustness framework** for considering the effect of ID accuracy while comparing OOD robustness
- Use a bank of standard models to **establish "baseline" expectations**
- A method that **exceeds expectations is effectively robust**
- We study this phenomena in various low and full-shot and regimes



4 Findings

- Amongst IN1k pre-trained models, ***SSL ViTs are often more robust but no single initialization or model size is the best*** across datasets



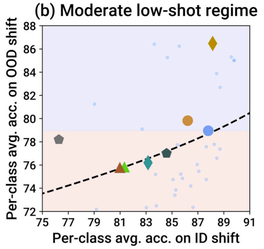
	ImageNet		iWildCam	
	ID	OOD	ID	OOD
1 MSN ViTS-16 [14]	58.99	21.51	26.41	**19.99**
2 DINO ViTS-16 [21]	53.78	19.09	24.78	19.75
3 MSN ViTB-16 [14]	**61.40**	**22.81**	24.78	19.65
4 DINO ViTB-16 [21]	56.72	21.98	**27.40**	19.82

- Without additional interventions, **CLIP is worse than IN1k ViTs on datasets other than ImageNet**

	ImageNet		iWildCam		Camelyon	
	ID	OOD	ID	OOD	ID	OOD
1 CLIP zero shot [4, 16]	**67.93**	**57.37**	9.67	16.82	50.48	51.55
2 CLIP [4]	50.8	27.50	23.75	19.10	84.9	77.3
3 MSN (IN1k) [14]	61.40	22.81	24.78	19.65	86.40	78.84
4 DINO (IN1k) [21]	56.72	21.98	**27.40**	**19.82**	**86.93**	**84.33**

- ***Robustness in full-shot regime ⇏ robustness in low-shot regimes***, as interventions fail to consistently ↑ robustness, ***except CLIP with WiSE-FT***

	ImageNet		iWildCam		Camelyon	
	$\rho\uparrow$	$\tau\uparrow$	$\rho\uparrow$	$\tau\uparrow$	$\rho\uparrow$	$\tau\uparrow$
Full-Shot Regime						
1 LP-FT [15]	5.16	-0.61	-1.41	-0.17	-0.45	7.48
2 + CLIP	19.60*	13.77*	-3.60	-6.09	0.37	11.28
3 WiSE-FT [16]	6.66	-0.86	-3.84	-5.87	6.22	12.66
4 + CLIP	22.24*	16.41*	3.98	4.78	2.85	14.18
5 Model Soups [16]	0.53	-10.58	-0.93	-0.14	-0.35	11.68
6 + CLIP	11.00†	4.29†	3.20	-4.84	5.93	9.50
7 RobustViT [68]	6.73	1.13	N/A	N/A	N/A	N/A
8 CLIP zero-shot [4, 16]	30.28	10.79	8.46	-23.167	-14.63	-28.54



5 Future Work

- Intervention performance is largely dataset and model dependent, but **weight-space ensembling [1] is promising for low-shot robustness**
- We hope to motivate researchers to focus on this practically important setting!

[1] Wortsman, et al. "Robust fine-tuning of zero-shot models." CVPR. 2022.