Following slides are from Ludwig Schmidt

Empirical science of ML (2019-2022)

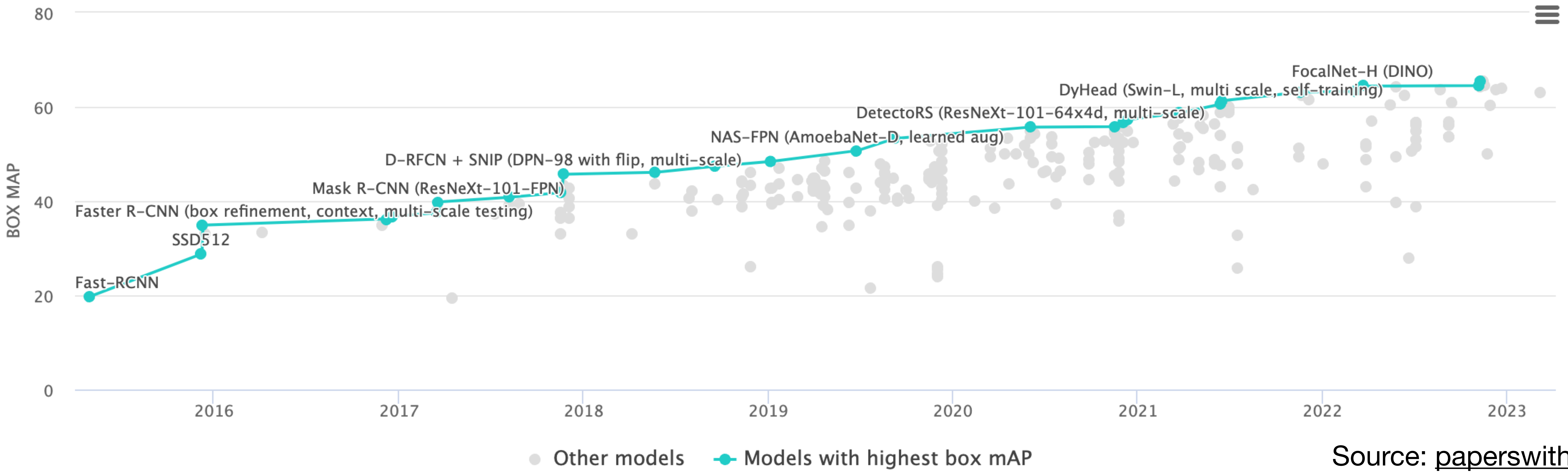
What is the path to reliable generalization?

ML = **algorithms** + **data**

- Optimization procedures
- Model architectures
- Loss functions
- … (thousands of papers)

?

Dominant paradigm in ML research: data fixed, improve models



Source: paperswithcode.com

Few papers experiment with improving the **training data**.

"Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI

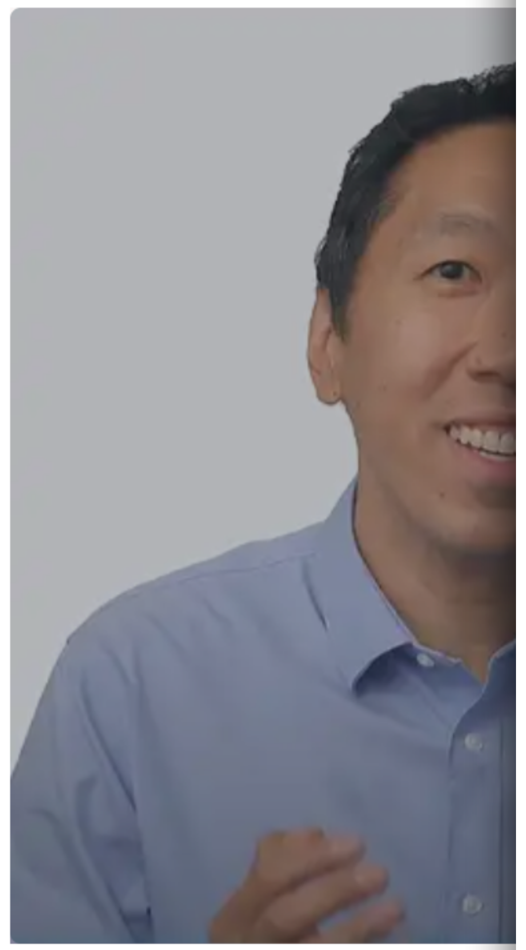
Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, Lora

DATA-CENTRIC AI

Topics Contribute NeurIPS 2021

Data-centric AI Resource Hub

Find the latest developments and best practices compiled here, so you can begin your Data-centric AI journey!



Neural Information Processing Systems Conference

Apr 7, 2021 · 4 min read · Listen

Announcing the NeurIPS 2021 Datasets and Benchmarks Track

Joaquin Vanschoren and Serena Yeung

There are no good models without good data (Sambasivan et al. 2021). The vast majority of the NeurIPS community focuses on algorithm design, but often can't easily find good datasets to evaluate their algorithms in a way that is maximally useful for the community and/or practitioners. Hence, many researchers resort to data that are conveniently available, but not representative of real applications. For instance, many algorithms are only evaluated on toy problems, or data that is plagued with bias, which could lead to biased models or misleading results, and subsequent public criticism of the field (Paullada et al. 2020).

🏆 Best paper a

[Deng, Dong, Socher, Li, Li, Fei-Fei'09]
[Russakovsky, Deng, Su, Krause, Satheesh, Ma, Huang, Karpathy, Khosla, Bernstein, Berg, Fei-Fei'15]

ImageNet

Large **image classification** dataset: 1.2M training images, 1,000 image classes.

Golden retriever

Great white shark

Minibus

[Deng, Dong, Socher, Li, Li, Fei-Fei'09]

[Russakovsky, Deng, Su, Krause, Satheesh, Ma, Huang, Karpathy, Khosla, Bernstein, Berg, Fei-Fei'15]

Robustness on ImageNet

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

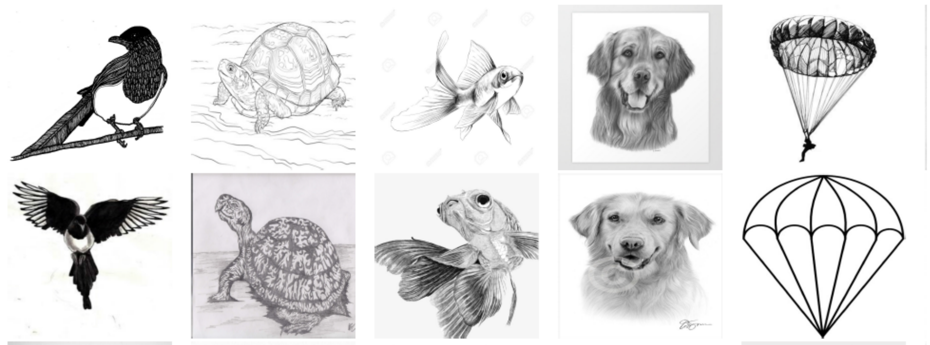
Evaluation: **new test sets**

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]

Measuring Robustness to Natural Distribution Shifts in Image Classification

Rohan Taori
UC Berkeley

Achal Dave
CMU

Vaishaal Shankar
UC Berkeley

Nicholas Carlini
Google Brain

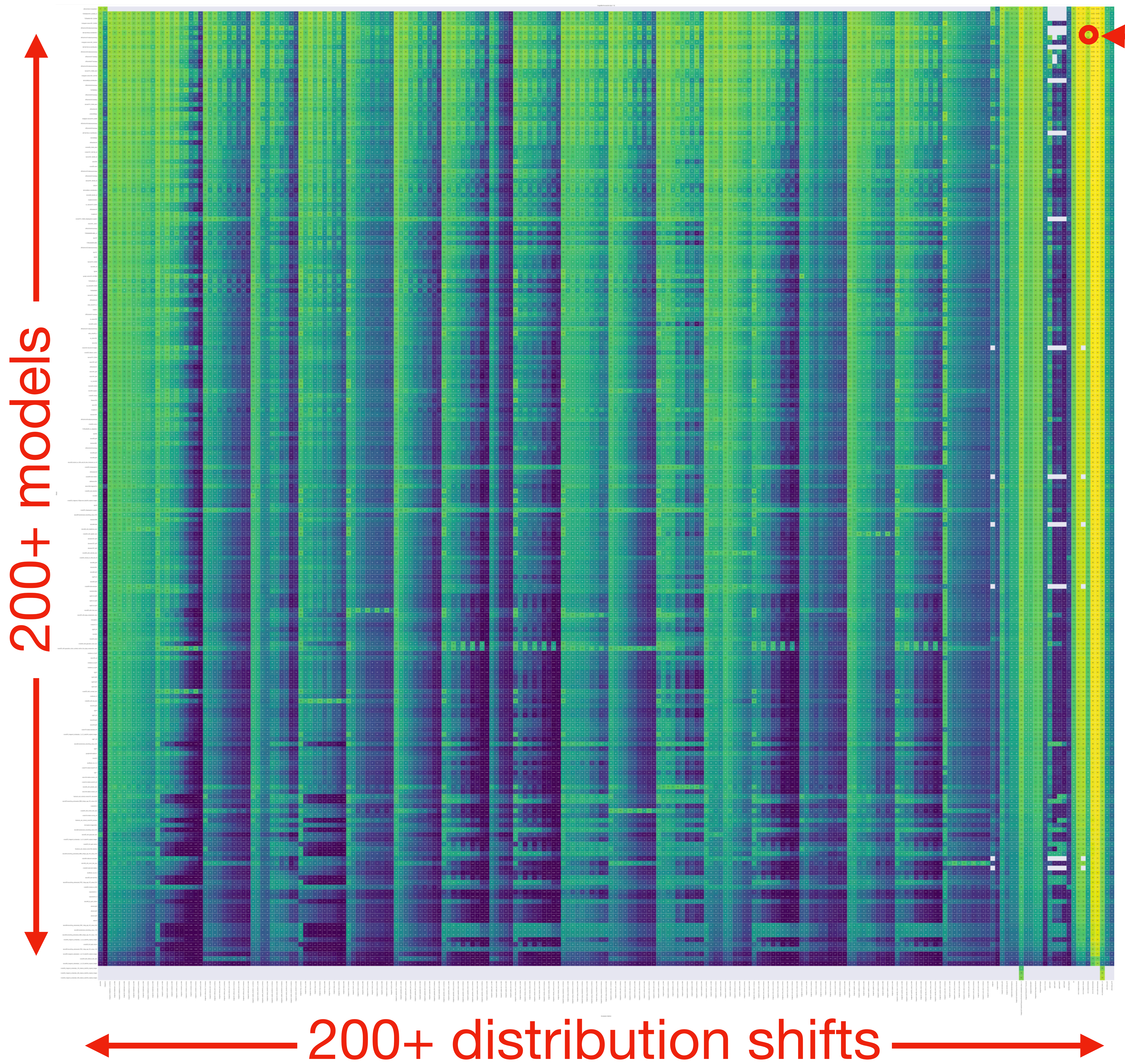
Benjamin Recht
UC Berkeley

Ludwig Schmidt
UC Berkeley

Abstract

We study how robust current ImageNet models are to distribution shifts arising from natural variations in datasets. Most research on robustness focuses on synthetic image perturbations (noise, simulated weather artifacts, adversarial examples, etc.), which leaves open how robustness on synthetic distribution shift relates to distribution shift arising in real data. Informed by an evaluation of 204 ImageNet models in 213 different test conditions, we find that there is often little to no transfer of robustness from current synthetic to natural distribution shift. Moreover, most current techniques provide no robustness to the natural distribution shifts in our testbed. The main exception is training on larger and more diverse datasets, which in multiple cases increases robustness, but is still far from closing the performance gaps. Our results indicate that distribution shifts arising in real data are currently an open research problem. We provide our testbed and data as a resource for future work at https://modestyachts.github.io/imagenet-testbed/.

Our approach: evaluate everything



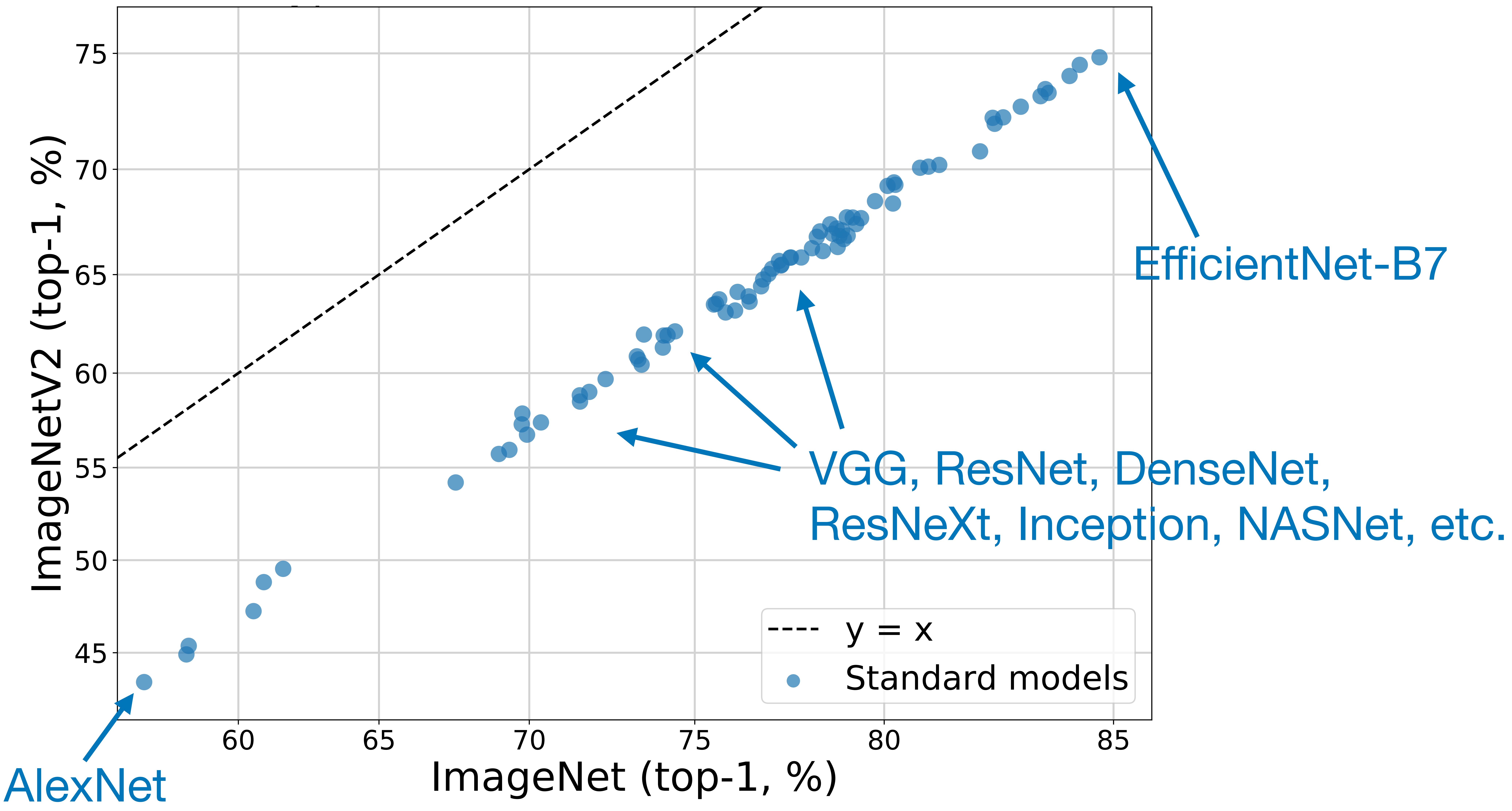
1 cell = 1 model evaluation on 1 dataset (total 10^9 image evaluations).

Models:

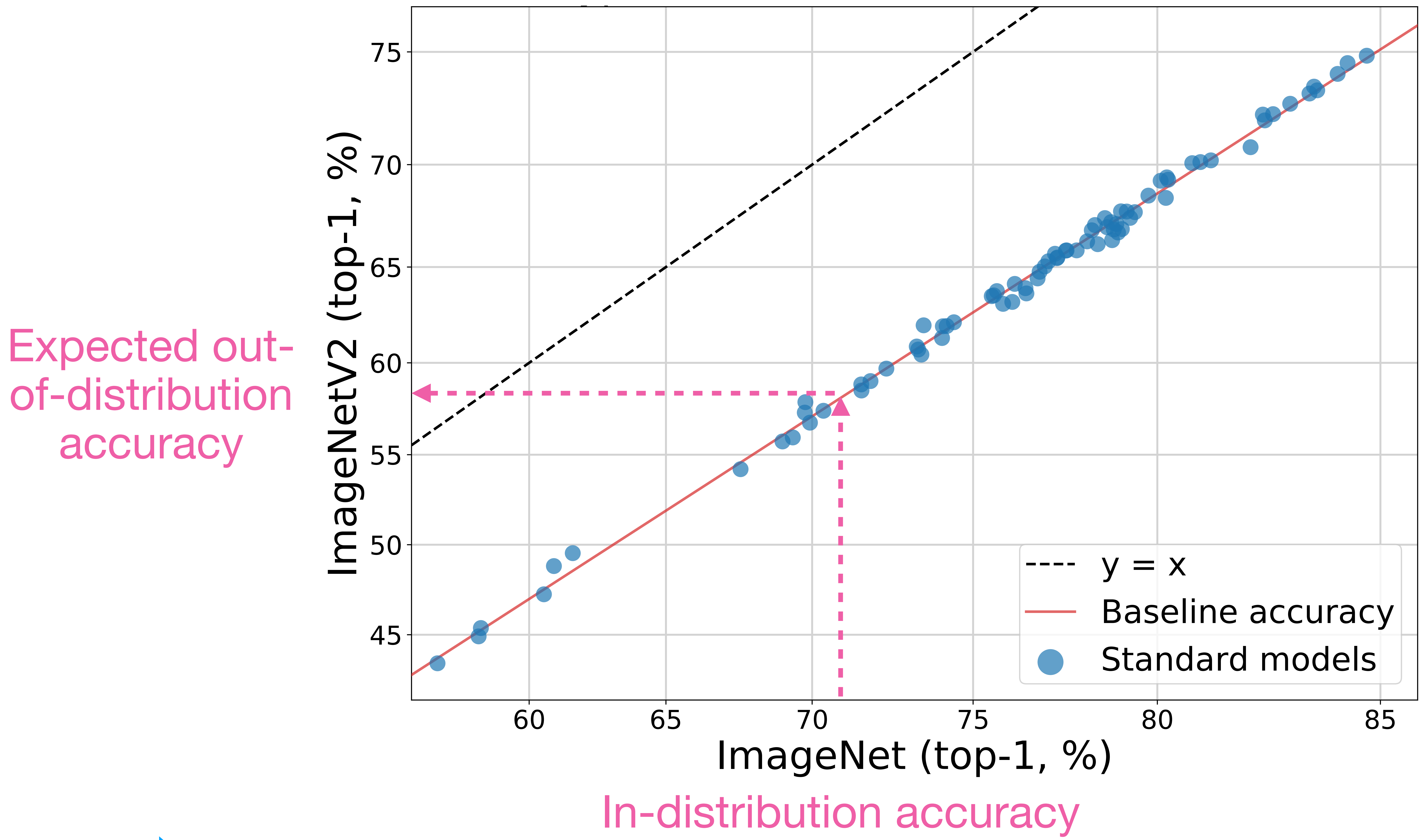
- **"Standard"** models (focus on ImgNet acc.)
- **Robust** models (adversarially robust models, models with special data augmentation, etc.)
- Models trained on **more data**

Distribution shifts

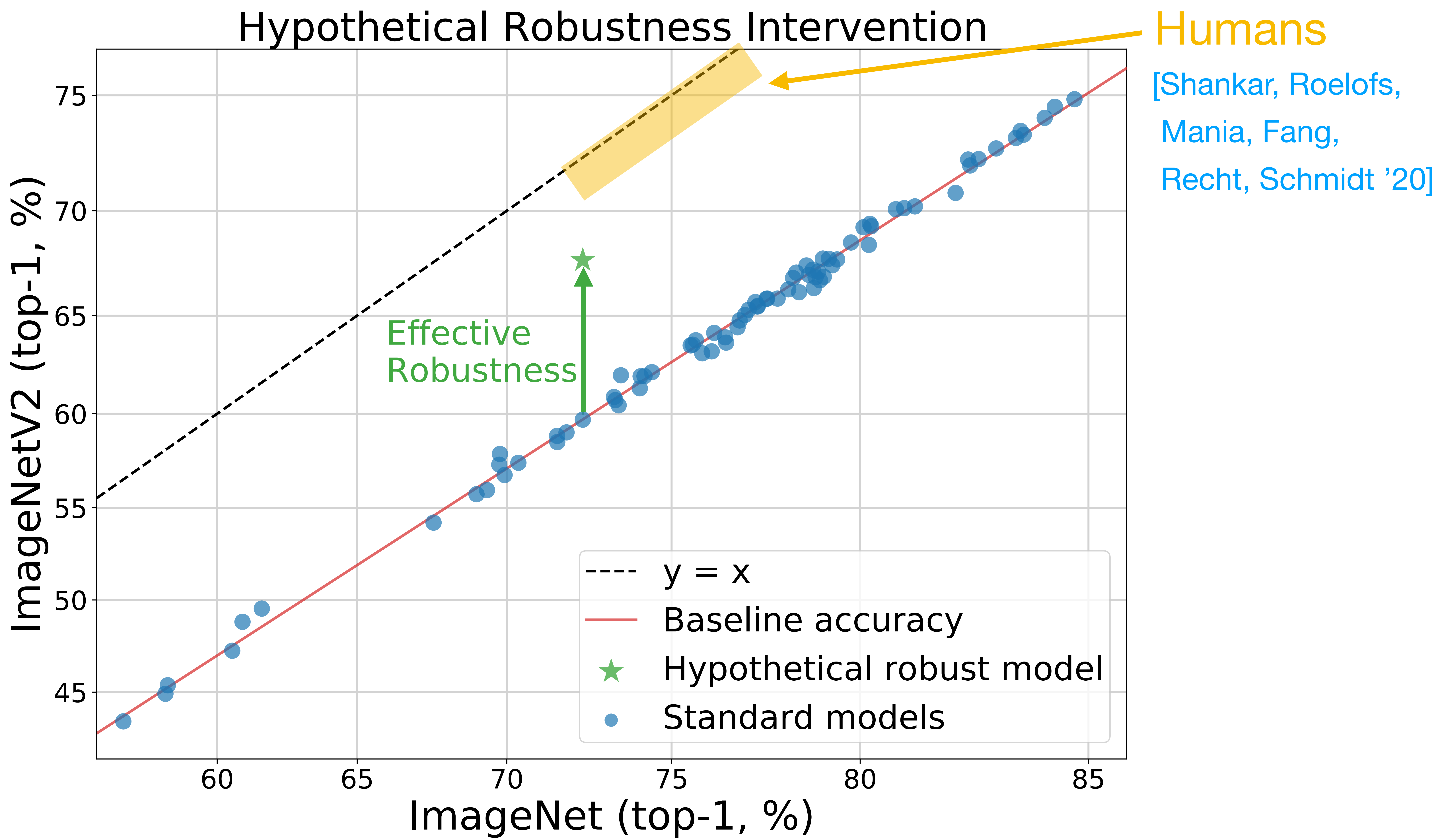
- ImageNet-V2
- ObjectNet
- ImageNet-R
- ImageNet-Sketch
- ImageNet-A
- ImageNetVid-Robust
- Adversarial attacks (L_p-norms)
- Image corruptions
- …



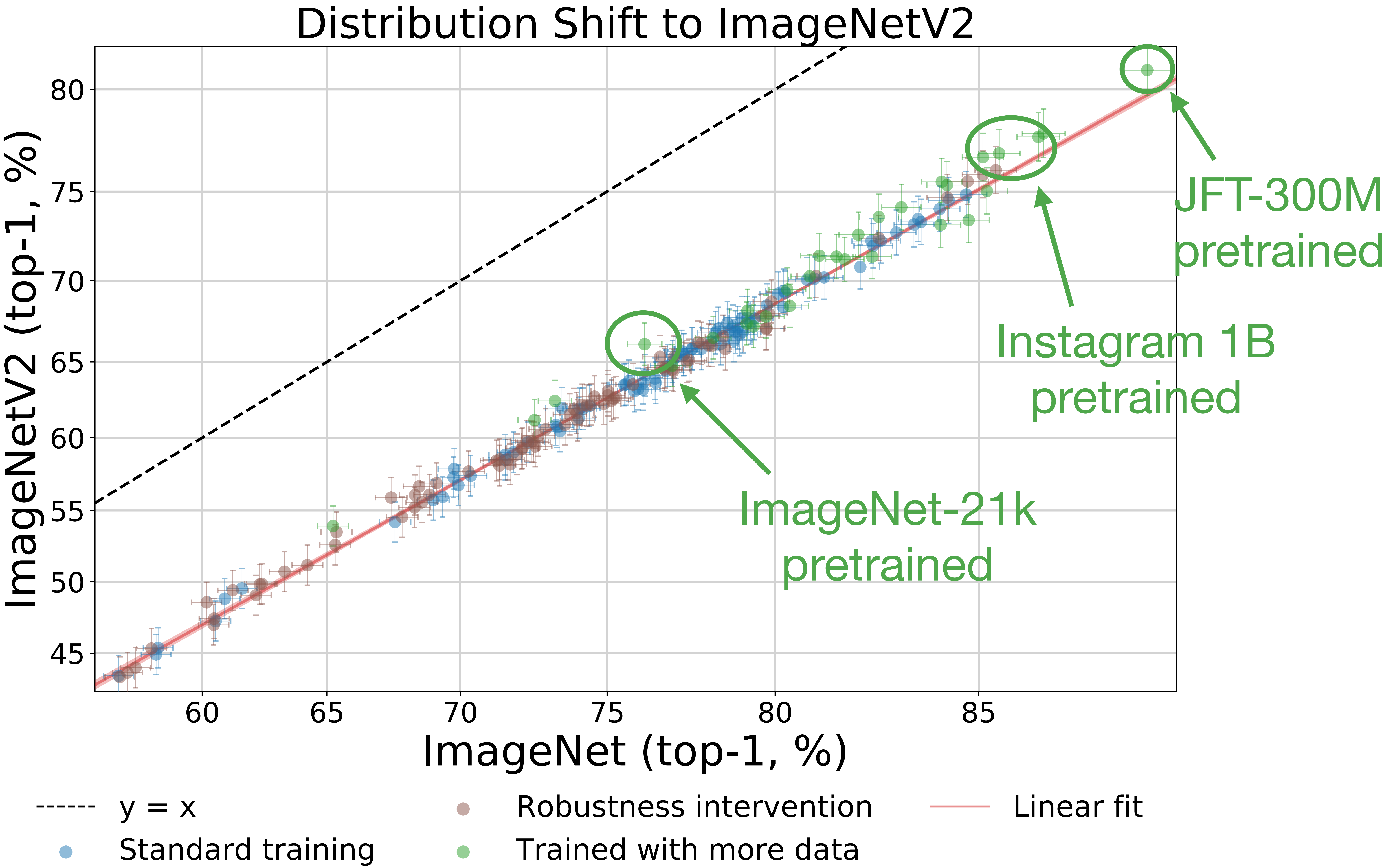
[Taori, Dave, Shankar, Carlini, Recht, Schmidt '20]



Baseline **out-of-distribution accuracy** from **in-distribution accuracy.**

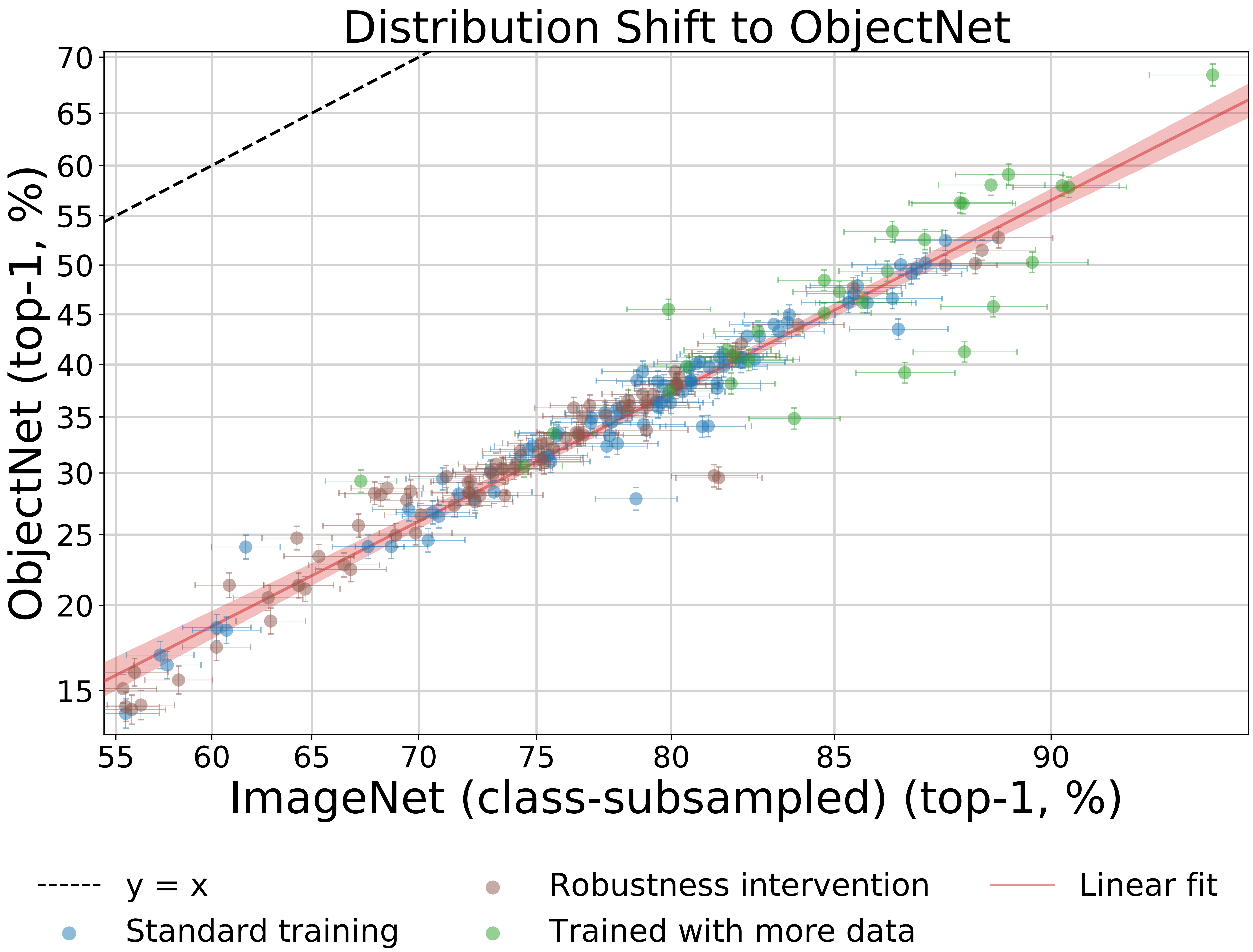


Do current robustness interventions achieve effective robustness?



➡ No current **robustness technique** achieves non-trivial effective robustness.

➡ Only training on (a lot) **more data** gives a small amount of effective robustness.



[Barbu, Mayo, Alverio, Luo, Wang, Gutfreund, Tenenbaum, Katz '19]

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

[Wang, Ge, Lipton, Xing '19]

[Hendrycks, Basart, Mu, Kadavath, Wang, Dorundo, Desai, Zhu, Parajuli, Guo, Song, Steinhardt, Gilmer '20]

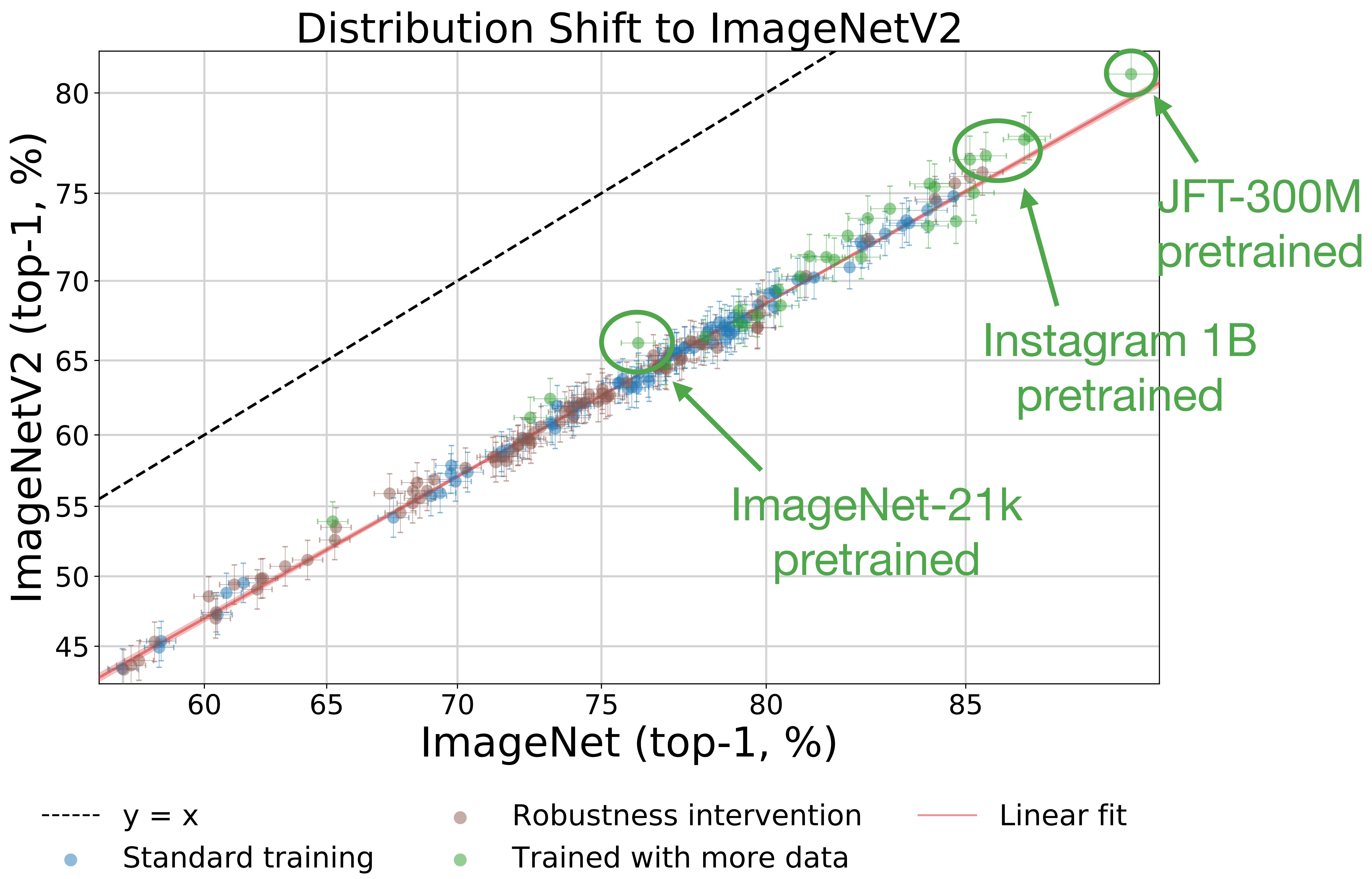
Some gains from **adv. training** and data augmentation. **More data** models still best.

Accuracy on the Line: On the Strong Correlation Between Out-of-Distribution and In-Distribution Generalization

John Miller[*] Rohan Taori[†] Aditi Raghunathan[†]
Shiori Sagawa[†] Pang Wei Koh[†] Vaishaal Shankar[*] Percy Liang[†]
Yair Carmon[‡] Ludwig Schmidt[§]

Abstract

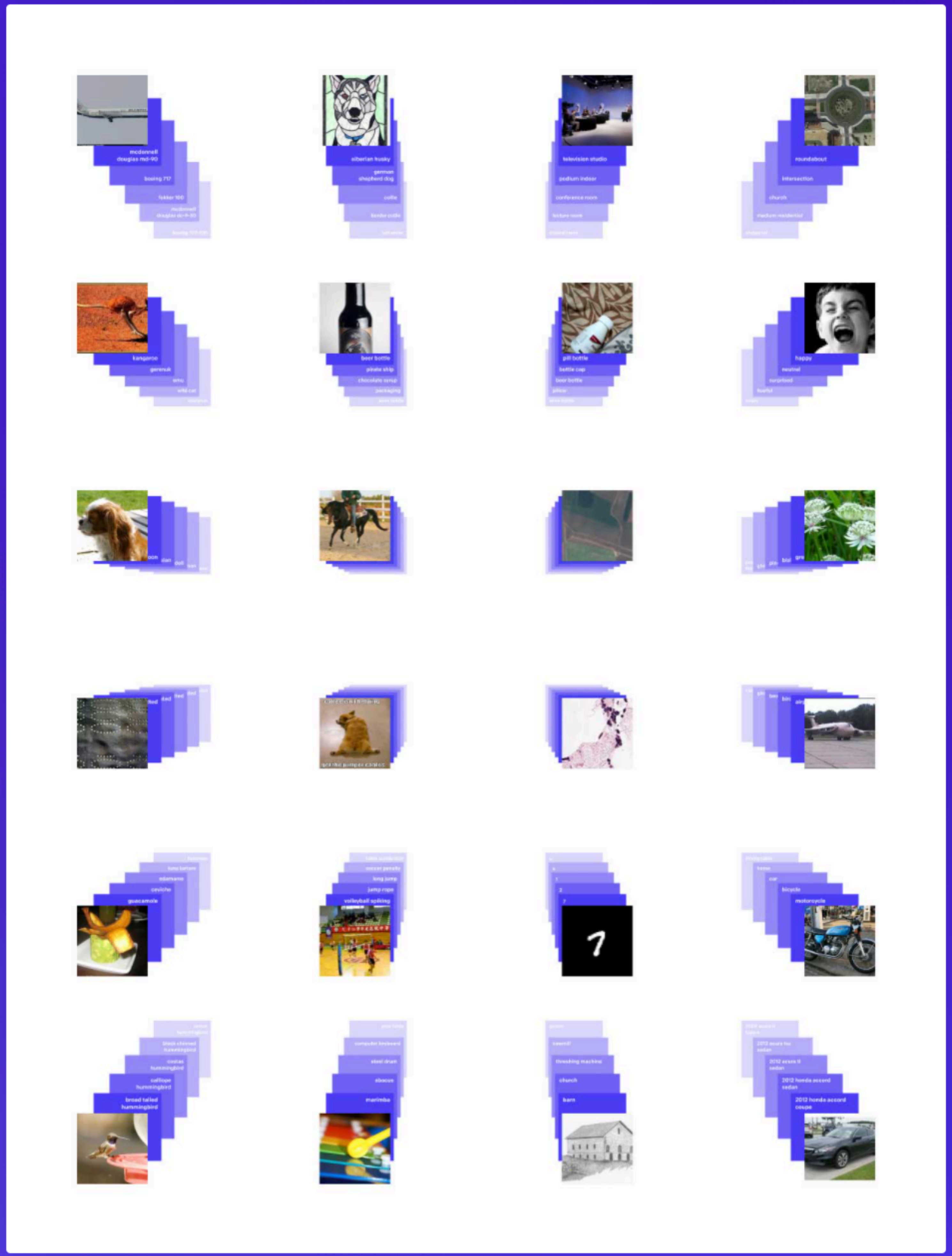
For machine learning systems to be reliable, we must understand their performance in unseen, out-of-distribution environments. In this paper, we empirically show that out-of-distribution performance is strongly correlated with in-distribution performance for a wide range of models and distribution shifts. Specifically, we demonstrate strong correlations between in-distribution and out-of-distribution performance on variants of CIFAR-10 & ImageNet, a synthetic pose estimation task derived from YCB objects, satellite imagery classification in FMoW-WILDS, and wildlife classification in iWildCam-WILDS. The strong correlations hold across model architectures, hyperparameters, training set size, and training duration, and are more precise than what is expected from existing domain adaptation theory. To complete the picture, we also investigate cases where the correlation is weaker, for instance some synthetic distribution shifts from CIFAR-10-C and the tissue classification dataset Camelyon17-WILDS. Finally, we provide a candidate theory based on a Gaussian data model that shows how changes in the data covariance arising from distribution shift can affect the observed correlations.



Training on (a lot) more data gives a **small** amount of effective robustness.

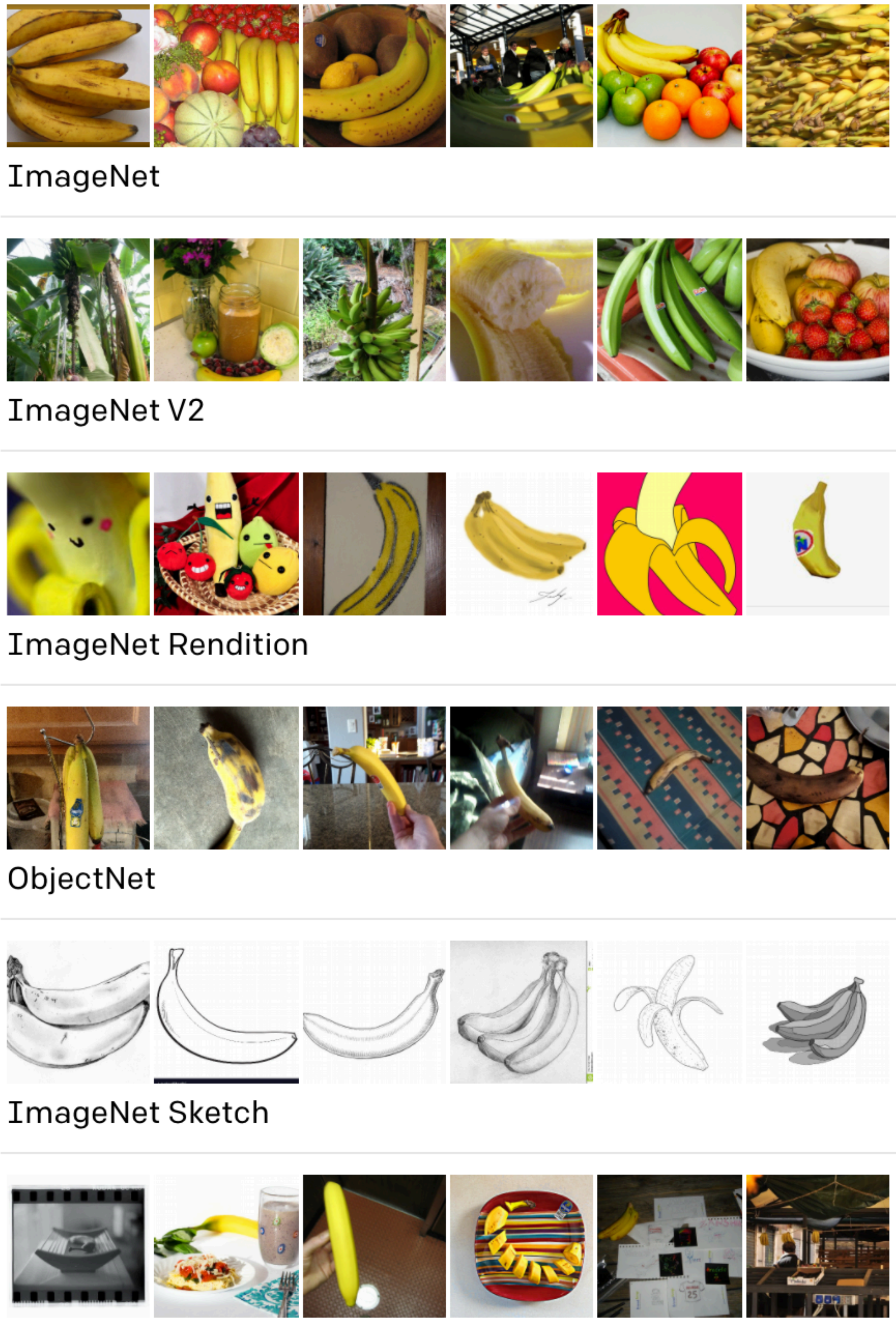
CLIP: Connecting Text and Images

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.



January 5, 2021
15 minute read

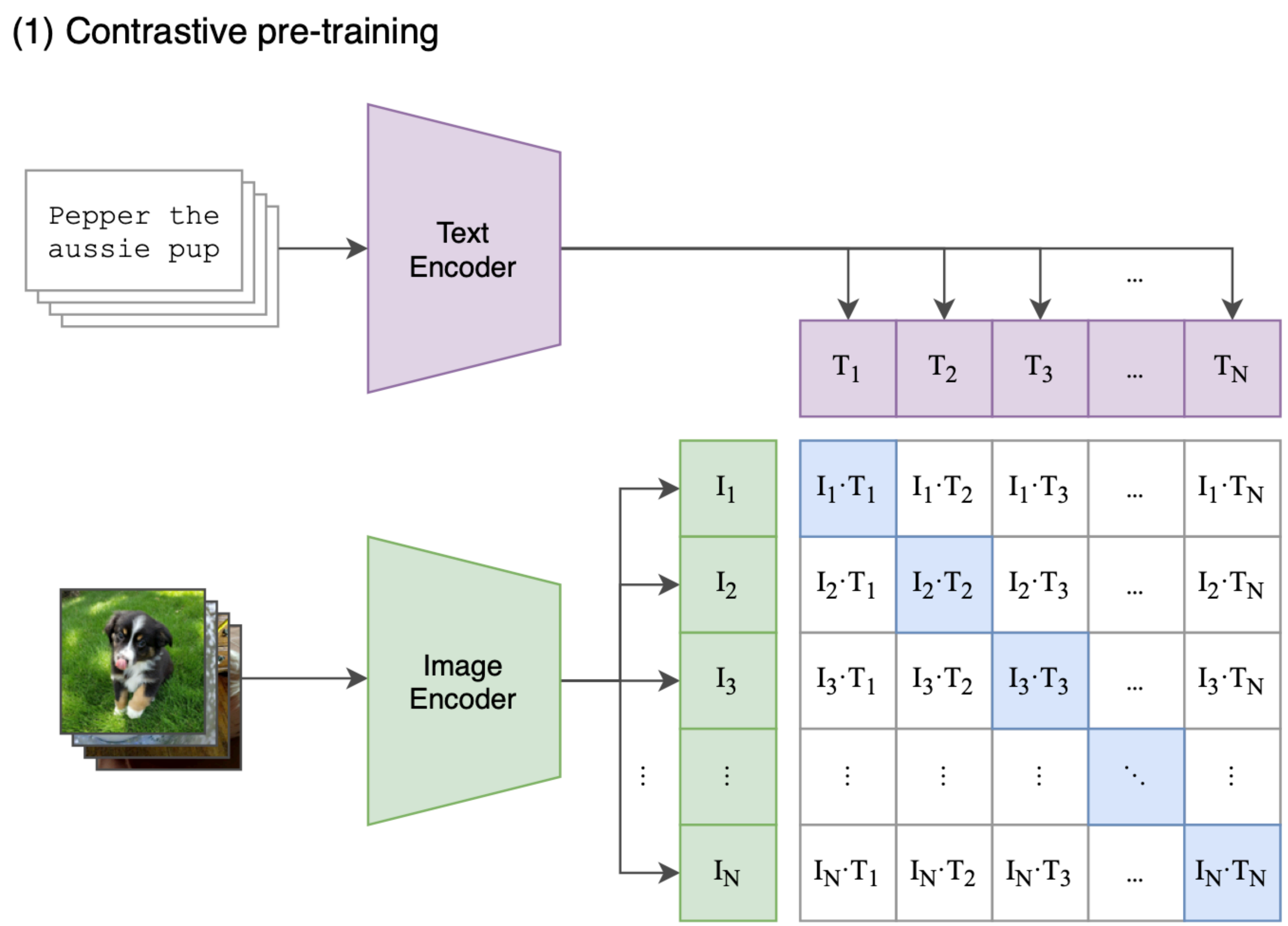
[Radford, Kim, Hallacy, Ramesh, Goh, Agarwal, Sastry, Askell, Mishkin, Clark, Krueger, Sutskever '21]



DATASET	IMAGENET RESNET101	CLIP VIT-L	Effective robustness
ImageNet	76.2%	76.2%	
ImageNet V2	64.3%	70.1%	+6%
ImageNet Rendition	37.7%	88.9%	+51%
ObjectNet	32.6%	72.3%	+40%
ImageNet Sketch	25.2%	60.2%	+35%
ImageNet Adversarial	2.7%	77.1%	+74%

→ **Very large** improvements in out-of-distribution robustness.

CLIP is not (explicitly) designed for robustness



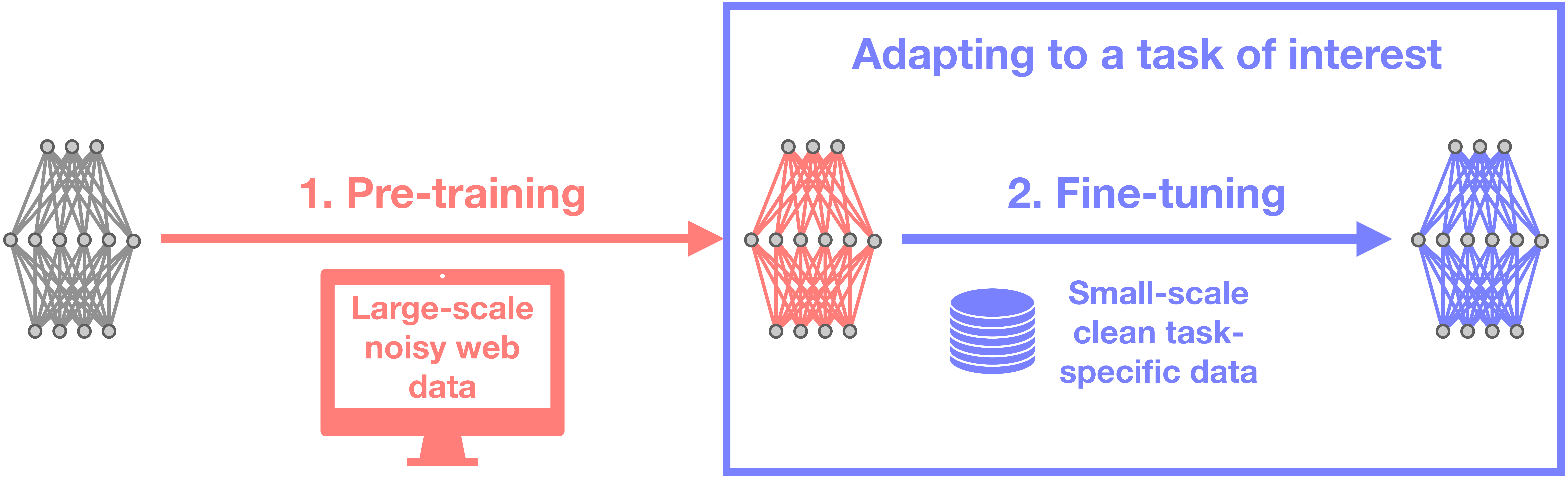
Training data: 400 million images collected from the web (dataset internal to OpenAI).

Compute: Trained on 250 - 600 GPUs for up to 18 days.

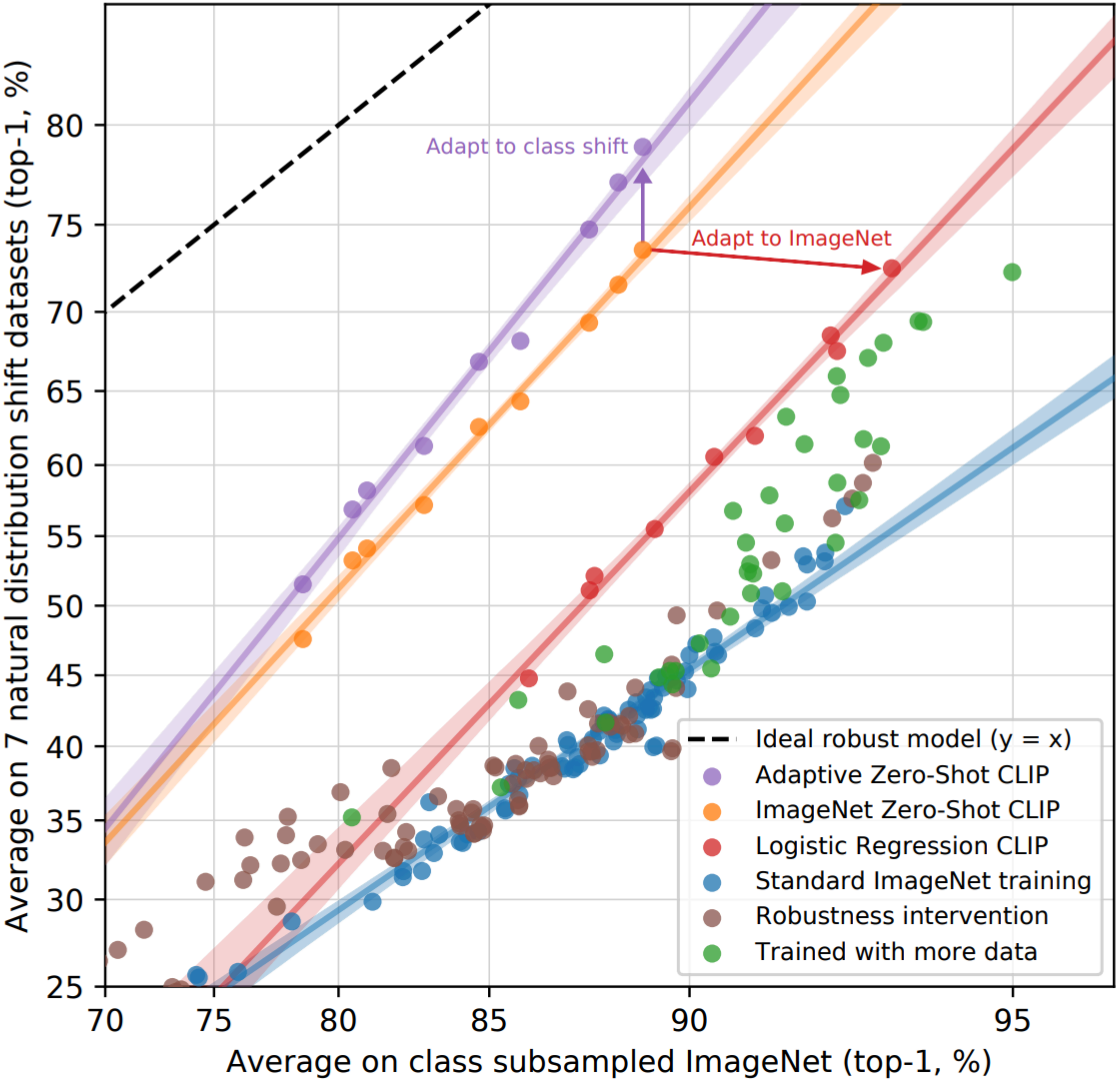
Model: ResNets and ViTs with up to 300M parameters.

Fine-tuning vs. zero-shot inference

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



CLIP skips fine-tuning: directly applies to task of interest via zero-shot inference.



[Radford, Kim, Hallacy, Ramesh, Goh, Agarwal, Sastry, Askell, Mishkin, Clark, Krueger, Sutskever '21]

Large robustness gains

➡ 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.

1 Introduction

Large pre-trained language-image models such as CLIP [27], ALIGN [21], and BASIC [26] have recently demonstrated unprecedented robustness on a variety of natural distribution shifts. In contrast to prior models that are trained on images with class annotations, CLIP and relatives[1] are directly trained on images and their corresponding unstructured text from the web. The resulting models achieve large robustness even on challenging distribution shifts such as ImageNetV2 [28] and ObjectNet [2]. No prior algorithmic techniques had enhanced robustness on these datasets even after multiple years of intensive research in reliable machine learning [13, 35]. As CLIP also improves robustness on a wide range of other distribution shifts, an important question emerges: *What causes CLIP's unprecedented robustness?*

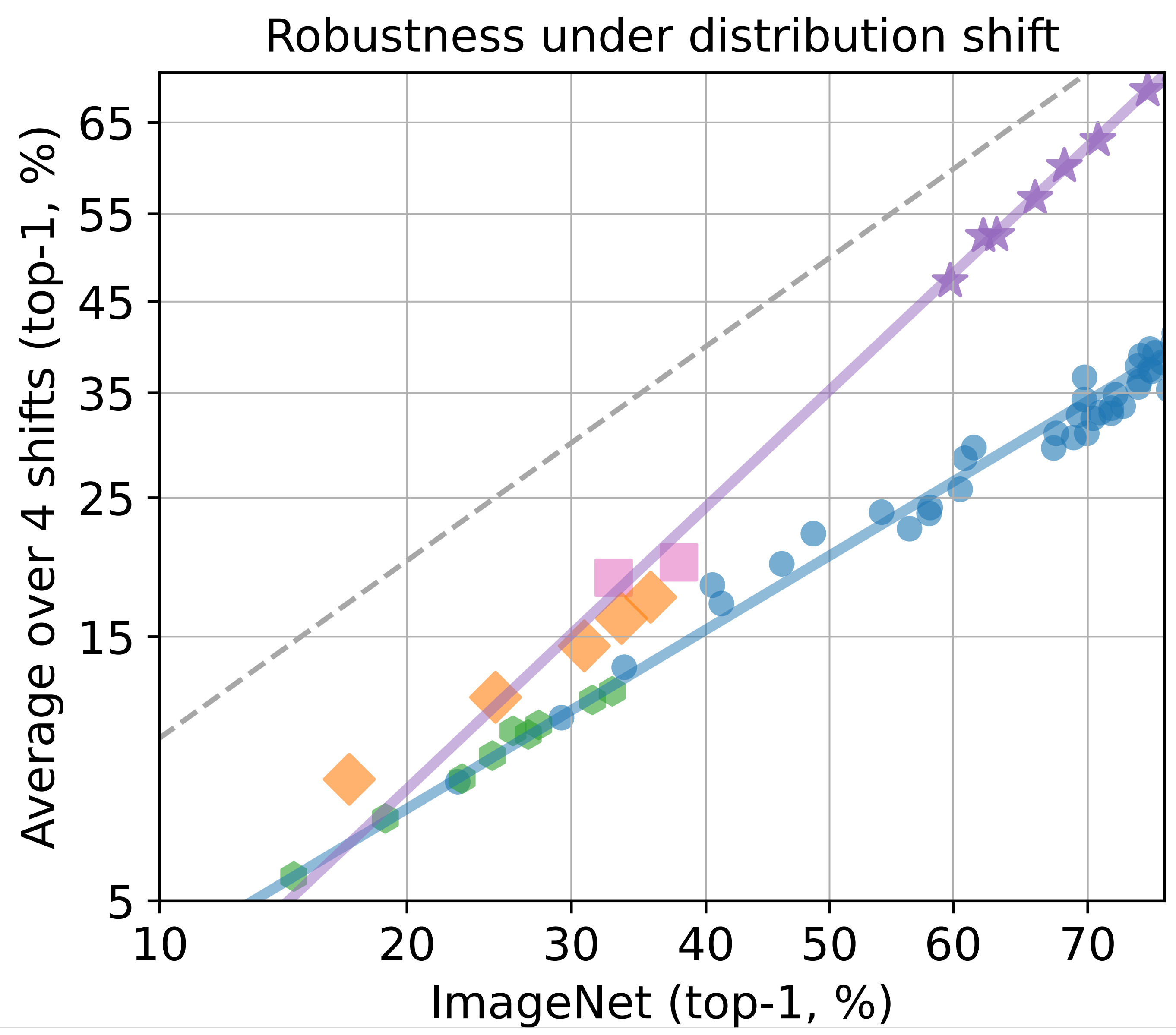
Hypotheses for CLIP's robustness

	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

	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~~

Conclusions



CLIP led to large robustness gains in image classification.

Image distribution is the main reason for CLIP's robustness.

- Not only scale but also **"diversity".**
- Language supervision helps with robustness indirectly: makes it **easier to collect training data.**

Open questions:

- How do we construct training sets that yield broadly reliable models?
- What about reasoning tasks (as opposed to recognition)?

github.com/mlfoundations/open_clip robustness.imagenetv2.org