Measuring Robustness to Natural Distribution Shifts in Image Classification

Rohan Taori[1], Achal Dave[2], Vaishaal Shankar[1], Nicholas Carlini[3], Benjamin Recht[1], Ludwig Schmidt[1]

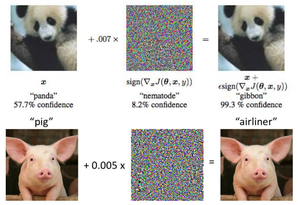
[1]UC Berkeley, [2]CMU, [3]Google Brain

Synthetic vs. Natural Distribution Shifts

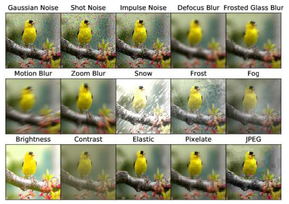
Synthetic Distribution Shifts

Created by modifying existing images according to a defined transformation.

Adversarial Examples



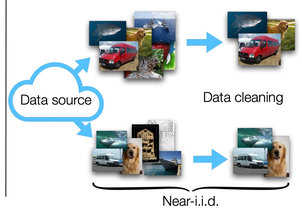
Artificial Corruptions



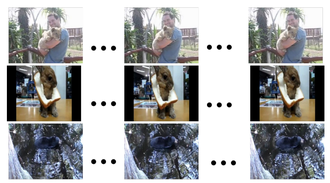
Natural Distribution Shifts

Created by modifying the underlying procedure used to sample the distribution.

ImageNetV2



ImageNet-Vid-Robust



- Real-world distribution shifts are likely hard to predict or characterize.
  - They shift the image generation process rather than modify specific pixels.
- There has been much work in the community creating synthetically robust models.
- **Main Question: Are vision models robust to natural distribution shift?**
  - We construct a large testbed of 10^9 model evaluations to answer this.

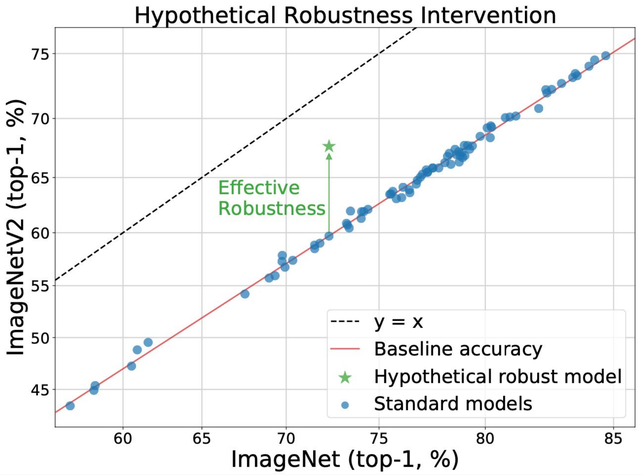
Measuring Robustness with Effective Robustness

Hypothetical Robust Model Example:

- Model B has higher target accuracy.
- Model A has a smaller accuracy drop.
- Which is more robust?

	In-distribution (Source) Accuracy	Out-of-distribution (Target) Accuracy	Accuracy Drop
Model A	80%	75%	5%
Model B	90%	77%	13%

Hypothetical Robustness Intervention



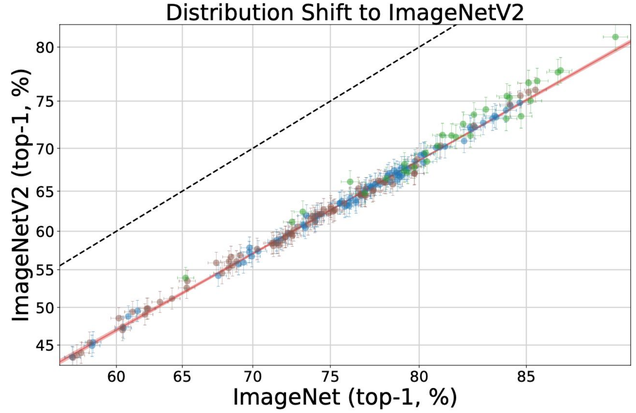
- Measuring robustness is difficult as standard accuracy acts as a confounder.
- Want to know: Does model B have target accuracy beyond what's expected from having a higher source accuracy?
- **Effective Robustness** is our notion of robustness beyond baseline accuracy.
- The log-linear fit is straightforward to compute from our testbed as models display a clear trend under shift.

Main Result: Little to No Robustness

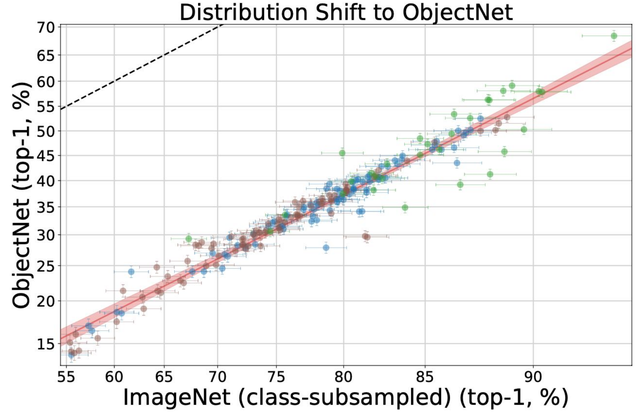
High-level takeaways:

- Most models & training strategies provide little to no effective robustness.
- Main outlier to the above is models trained on more data (but the effect isn't uniform).
- **Recommendations**: 1) Measure effective robustness, and 2) Evaluate on natural shifts.

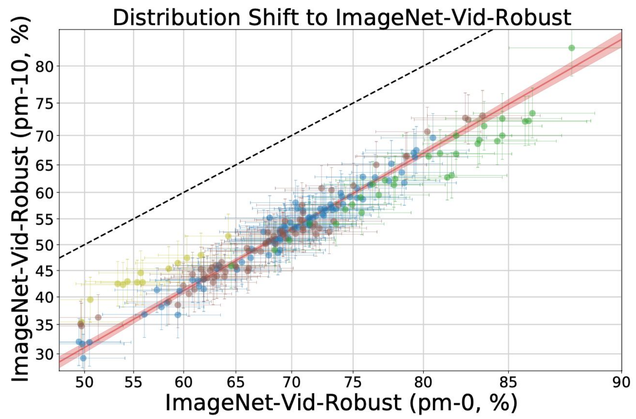
Near-i.i.d. test set:



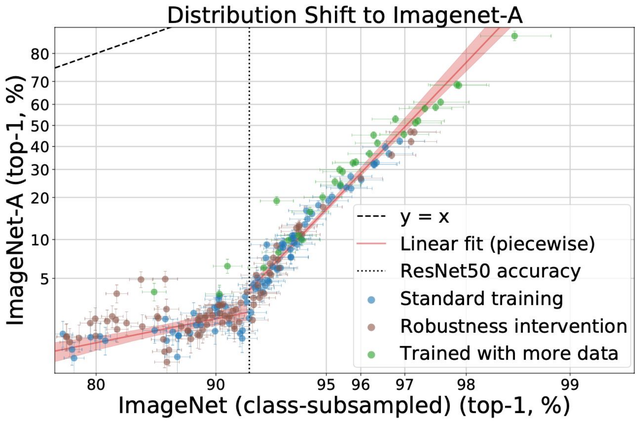
Test set of objects:



Video frames test set:



Adversarially filtered:



Models and Datasets in our Testbed

200+ models:

- Standard models - architectures from AlexNet to EfficientNet.
- Robust models - adversarially robust models & models with data-aug (cutout, augmix, etc.).
- Models trained on more data - Instagram-1B, JFT-300M, YFCC-100M, & other datasets.

200+ distribution shifts:

- Most current natural distribution shifts:
  - ImageNetV2, ObjectNet, ImageNet-Vid-Robust, YTBB-Robust, ImageNet-A.
- Synthetic distribution shifts - Lp attacks & image corruptions.