Effective Robustness against Natural Distribution Shifts for Models with Different Training Data

Zhouxing Shi[1,*], Nicholas Carlini[2], Ananth Balashankar[2], Ludwig Schmidt[3], Cho-Jui Hsieh[1,2], Alex Beutel[4,*], Yao Qin[2,5]

[1]University of California, Los Angeles [2]Google [3]University of Washington
[4]OpenAI [5]University of California, Santa Barbara [*]Work done at Google

Effective Robustness in the Literature

We consider robustness to natural distribution shifts evaluated on benchmark test sets.

- **Effective robustness** measures the extra out-of-distribution (OOD) robustness **beyond** what can be predicted from the in-distribution (ID) performance.
- **Controlling for ID accuracy** distinguishes:
  An effectively improved robustness v.s. An expected outcome of a higher ID accuracy.
- ImageNet models usually have similar effective robustness.
- **Zero-shot Contrastive Language-Image Pre-trained (CLIP)** models apparently achieved significant effective robustness gains.

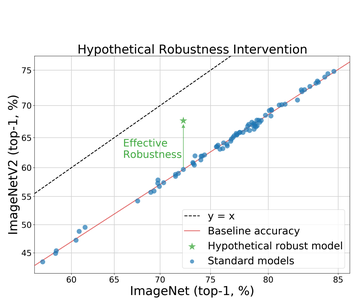

Figure 1. ImageNet models evaluated by Taori et al., 2020.

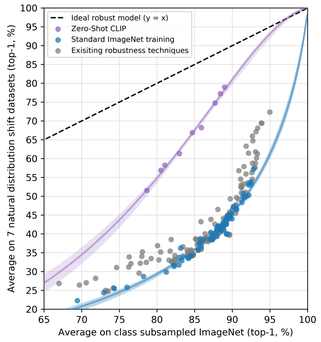

Figure 2. CLIP models v.s. ImageNet models evaluated by Radford et al., 2021.

Do zero-shot CLIP models truly have stronger effective robustness?

A subtle issue:

- ImageNet was regarded as an "in-distribution" data for all the models.
- **CLIP models here were NOT trained on ImageNet.**
- A mismatch between CLIP's training data and the ID data in the evaluation.

Limitations of the previous evaluation:

- It requires a single fixed ID test set.
- It can become **problematic when there are models trained on different data**.

Contradictory Results under Varying ID Test Sets

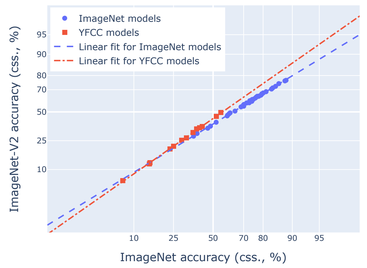

Figure 3. Using ImageNet as the ID test set.

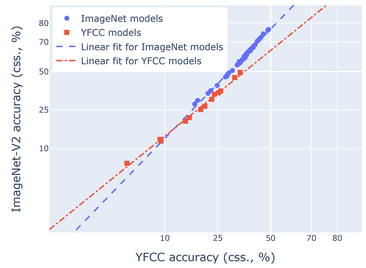

Figure 4. Using YFCC as the ID test set.

- The models appear to be more robust when there is a mismatch between their training data and the ID test set.

Our Generalized Evaluation

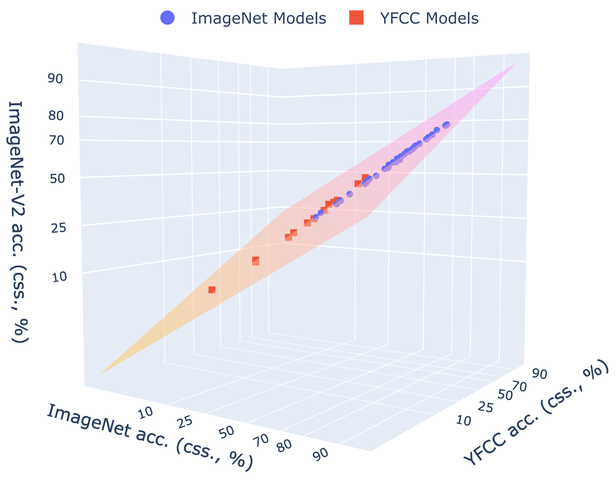

Figure 5. Accuracy on the plane.

- We control for the accuracy evaluated on **multiple ID test sets** to cover the training distributions of all the models.
- Baseline function: a fitting plane $\beta(x, y)$ from baseline models.
- **Multi-ID effective robustness**:

$$\rho(f) = \mathrm{acc}_{\mathrm{ood}}(f) - \beta(\mathrm{acc}_1(f), \mathrm{acc}_2(f)).$$

Accuracy on the Plane

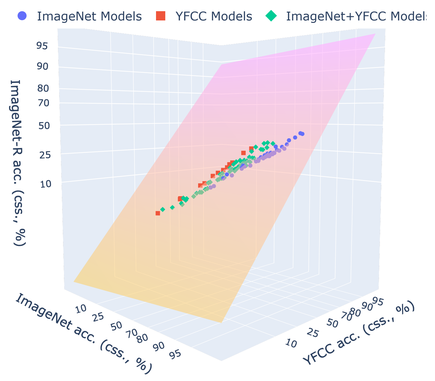

Figure 6. ImageNet models v.s. YFCC models on ImageNet-R.

Figure 7. Interactive visualization.

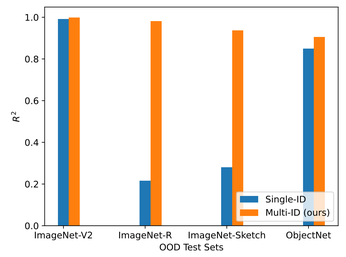

Figure 8. Fitting quality by R^2.

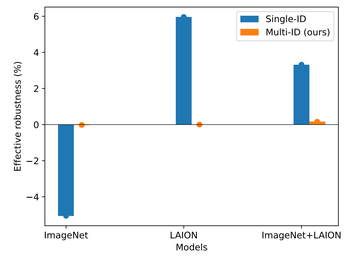

Figure 9. Effective robustness values.

- Our multi-ID evaluation improves the fitting quality of the baseline functions.
- No model has a clear effective robustness gain under our new evaluation.

Conclusion

- Although CLIP models pre-trained on some datasets can improve the accuracy on OOD test sets, **this improvement is not an effective robustness gain**.
- Our work provides a new effective robustness evaluation for models trained on different data, and we also provide a new understanding on the effective robustness gains of CLIP-like models observed in previous works.

NeurIPS 2023, New Orleans, LA
Contact: z.shi@ucla.edu