Measuring Robustness to Natural Distribution Shifts in Image Classification

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, Ludwig Schmidt

1 UC Berkeley, 2 CMU, 3 Google Brain

Main Result: Little to No Robustness

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?

<table>
<thead>
<tr>
<th></th>
<th>In-distribution (Source) Accuracy</th>
<th>Out-of-distribution (Target) Accuracy</th>
<th>Accuracy Drop</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model A</td>
<td>80%</td>
<td>75%</td>
<td>5%</td>
</tr>
<tr>
<td>Model B</td>
<td>90%</td>
<td>77%</td>
<td>13%</td>
</tr>
</tbody>
</table>

- 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.

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

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.

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:
Distribution Shift to ImageNetV2

Test set of objects:
Distribution Shift to ObjectNet

Video frames test set:
Distribution Shift to ImageNet-Vid-Robust

Adversarially filtered:
Distribution Shift to ImageNet-A

More analysis, code, and data at: tinyurl.com/imagenet-testbed