Improving robustness against common corruptions by covariate shift adaptation

ICML 2020 Workshop on Uncertainty & Robustness in Deep Learning

Steffen Schneider\textsuperscript{1,2\*}, Evgenia Rusak\textsuperscript{1,2\*}, Luisa Eck\textsuperscript{3}, Oliver Bringmann\textsuperscript{1\dagger}, Wieland Brendel\textsuperscript{1\dagger}, Matthias Bethge\textsuperscript{1\dagger}

July 17, 2020

\textsuperscript{1}University of Tübingen \textsuperscript{2}IMPRS-IS \textsuperscript{3}LMU Munich

Web: domainadaptation.org/batchnorm
Contact: steffen@bethgelab.org
Benchmarking corruption robustness
Benchmarking corruption robustness: ImageNet-C (Hendrycks et al., ‘19)

<table>
<thead>
<tr>
<th>Category</th>
<th>Test Corruptions</th>
<th>Holdout</th>
</tr>
</thead>
<tbody>
<tr>
<td>blur</td>
<td><img src="image1" alt="Blur Images" /></td>
<td><img src="image2" alt="Blur Holdout" /></td>
</tr>
<tr>
<td>digital</td>
<td><img src="image3" alt="Digital Images" /></td>
<td><img src="image4" alt="Digital Holdout" /></td>
</tr>
<tr>
<td>noise</td>
<td><img src="image5" alt="Noise Images" /></td>
<td><img src="image6" alt="Noise Holdout" /></td>
</tr>
<tr>
<td>weather</td>
<td><img src="image7" alt="Weather Images" /></td>
<td><img src="image8" alt="Weather Holdout" /></td>
</tr>
</tbody>
</table>

For $C = 15$ test corruptions and $S = 5$ severities.
Benchmarking corruption robustness: ImageNet-C (Hendrycks et al., ‘19)

<table>
<thead>
<tr>
<th>Category</th>
<th>Test Corruptions</th>
<th>Holdout</th>
</tr>
</thead>
<tbody>
<tr>
<td>blur</td>
<td><img src="blur.png" alt="Image" /></td>
<td><img src="blur.png" alt="Image" /></td>
</tr>
<tr>
<td>digital</td>
<td><img src="digital.png" alt="Image" /></td>
<td><img src="digital.png" alt="Image" /></td>
</tr>
<tr>
<td>noise</td>
<td><img src="noise.png" alt="Image" /></td>
<td><img src="noise.png" alt="Image" /></td>
</tr>
<tr>
<td>weather</td>
<td><img src="weather.png" alt="Image" /></td>
<td><img src="weather.png" alt="Image" /></td>
</tr>
</tbody>
</table>

Mean Corruption Error (lower is better):

\[
mCE(\text{model}) = \frac{1}{C} \sum_{c=1}^{C} \frac{\sum_{s=1}^{S} \text{err}_{c,s}^{\text{model}}}{\sum_{s=1}^{S} \text{err}_{c,s}^{\text{AlexNet}}} \]

For \( C = 15 \) test corruptions and \( S = 5 \) severities.
Adaptation of Batch Norm Statistics

Current practice

ImageNet

$\mu_s, \Sigma_s$

$\mu_s, \Sigma_s$

$\mu_s, \Sigma_s$

squirrel
Adaptation of Batch Norm Statistics

Current practice

ImageNet

ImageNet-C

squirrel

July 17, 2020 Schneider & Rusak et al. — Improving robustness against common corruptions by covariate shift adaptation
Adaptation of Batch Norm Statistics

Proposed (large sample)

ImageNet

\[ \mu_s, \Sigma_s \]

\[ \mu_s, \Sigma_s \]

\[ \mu_s, \Sigma_s \]

\[ \text{squirrel} \]

ImageNet-C

\[ \mu_t, \Sigma_t \]

\[ \mu_t, \Sigma_t \]

\[ \mu_t, \Sigma_t \]

\[ ? \]
Adaptation of Batch Norm Statistics

Proposed (small sample)

ImageNet

$\mu_s, \Sigma_s$

BN

$\mu_s, \Sigma_s$

BN

$\mu_s, \Sigma_s$

BN

$\mu_s, \Sigma_s$

BN

$\mu_s, \Sigma_s$

squirrel

ImageNet-C

$\bar{\mu}, \bar{\Sigma}$

BN

$\bar{\mu}, \bar{\Sigma}$

BN

$\bar{\mu}, \bar{\Sigma}$

BN

$\bar{\mu}, \bar{\Sigma}$

?
Issue 1:

- Robustness is benchmarked in an *ad hoc* setting, assuming access to one sample.
Motivation: Rethinking robustness evaluation

Issue 1:

- Robustness is benchmarked in an ad hoc setting, assuming access to one sample.
- In many practical problems (medical imaging, quality control, ...), it is a reasonable assumption that distributions only slowly drift — or abruptly change, but only from time to time.
Motivation: Rethinking robustness evaluation

Issue 1:

• Robustness is benchmarked in an *ad hoc* setting, assuming access to one sample.
• In many practical problems (medical imaging, quality control, ...), it is a reasonable assumption that distributions only slowly drift — or abruptly change, but only from time to time.
Motivation: Rethinking robustness evaluation

Issue 1:

- Robustness is benchmarked in an *ad hoc* setting, assuming access to one sample.
- In many practical problems (medical imaging, quality control, ...), it is a reasonable assumption that distributions only slowly drift — or abruptly change, but only from time to time.

Issue 2:

- Many computer vision models are trained using batch normalization.
Motivation: Rethinking robustness evaluation

Issue 1:

- Robustness is benchmarked in an *ad hoc* setting, assuming access to one sample.
- In many practical problems (medical imaging, quality control, ...), it is a reasonable assumption that distributions only slowly drift — or abruptly change, but only from time to time.

Issue 2:

- Many computer vision models are trained using batch normalization.
- Problem with BN in *o.o.d.* scenarios: Training stats are not optimal at test time.
Issue 1:

- Robustness is benchmarked in an ad hoc setting, assuming access to one sample.
- In many practical problems (medical imaging, quality control, ...), it is a reasonable assumption that distributions only slowly drift — or abruptly change, but only from time to time.

Issue 2:

- Many computer vision models are trained using batch normalization.
- Problem with BN in o.o.d. scenarios: Training stats are not optimal at test time.
Motivation: Rethinking robustness evaluation

Issue 1:

- Robustness is benchmarked in an *ad hoc* setting, assuming access to one sample.
- In many practical problems (medical imaging, quality control, ...), it is a reasonable assumption that distributions only slowly drift — or abruptly change, but only from time to time.

Issue 2:

- Many computer vision models are trained using batch normalization.
- Problem with BN in o.o.d. scenarios: Training stats are not optimal at test time.

**Hypothesis:** Current robustness results underestimate model performance.

We propose a simple baseline for IN-C evaluation beyond the ad hoc settings.
Adaptation boosts robustness of a vanilla trained ResNet-50 model.

\[
\bar{\mu} = \frac{N\mu_s + n\mu_t}{N + n}
\]

\[
\bar{\sigma}^2 = \frac{N\sigma_s^2 + n\sigma_t^2}{N + n}
\]

- \(n\): Target batch size
- \(N\): Pseudo batch size
Adaptation boosts robustness of a vanilla trained ResNet-50 model.

\[ \bar{\mu} = \frac{N \mu_s + n \mu_t}{N + n} \]

\[ \bar{\sigma}^2 = \frac{N \sigma_s^2 + n \sigma_t^2}{N + n} \]

$n$: Target batch size  

$N$: Pseudo batch size

---

**Graph:**

- **RN50 AM**
  - $N = 0$
  - $N$ best
  - $N = \infty$

- **AssembleNet (SoTA)**
  - $N = \infty$

**Axes:**
- **y-axis**: mCE (mean classification error)
- **x-axis**: Batch size

**Legend:**
- Green dashed line: $N = 0$
- Red dotted line: $N$ best
- Green solid line: $N = \infty$
Adaptation boosts robustness of a vanilla trained ResNet-50 model.

\[ \mu = \frac{N\mu_s + n\mu_t}{N + n} \]
\[ \sigma^2 = \frac{N\sigma_s^2 + n\sigma_t^2}{N + n} \]

\( n \): Target batch size

\( N \): Pseudo batch size
Adaptation yields new state of the art on ImageNet-C for robust models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean Corruption Error [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-50</td>
<td>77</td>
</tr>
<tr>
<td>SIN</td>
<td>69</td>
</tr>
<tr>
<td>ANT</td>
<td>63</td>
</tr>
<tr>
<td>ANT+SIN</td>
<td>61</td>
</tr>
<tr>
<td>AssembleNet</td>
<td>54</td>
</tr>
<tr>
<td>AugMix</td>
<td>65</td>
</tr>
</tbody>
</table>

Baseline
Adaptation yields new state of the art on ImageNet-C for robust models.

The graph shows the mean corruption error in percent for different models: ResNet-50, SIN, ANT, ANT+SIN, and AssembleNet. The Baseline is compared to the Adapted (n=8) versions. The error rates range from 54% to 77%.
Adaptation yields new state of the art on ImageNet-C for robust models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean Corruption Error [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>Adapted (full)</td>
</tr>
<tr>
<td>ResNet-50</td>
<td>77</td>
</tr>
<tr>
<td>SIN</td>
<td>69</td>
</tr>
<tr>
<td>ANT</td>
<td>63</td>
</tr>
<tr>
<td>ANT+SIN</td>
<td>61</td>
</tr>
<tr>
<td>AssembleNet</td>
<td>65</td>
</tr>
<tr>
<td>AugMix</td>
<td>52</td>
</tr>
</tbody>
</table>

Bar graph showing the mean corruption error for different models on ImageNet-C, comparing baseline and adapted (full) versions.
Adaptation yields new state of the art on ImageNet-C for robust models.

\[
\bar{\mu} = \frac{N\mu_s + n\mu_t}{N + n}
\]

\[
\bar{\sigma}^2 = \frac{N\sigma_s^2 + n\sigma_t^2}{N + n}
\]

\(n\): Target batch size

\(N\): Pseudo batch size
Adaptation yields new state of the art on ImageNet-C for robust models.

\[
\bar{\mu} = \frac{N\mu_s + n\mu_t}{N + n}
\]

\[
\bar{\sigma}^2 = \frac{N\sigma_s^2 + n\sigma_t^2}{N + n}
\]

- \(n\): Target batch size
- \(N\): Pseudo batch size

![Graph showing the effect of batch size on mCE for different models and batch sizes.]

**RN50 AM**
- \(N = 0\)
- \(N\) best
- \(N = \infty\)

**AssembleNet (SoTA)**
- \(N = \infty\)
Adaptation yields new state of the art on ImageNet-C for robust models.

\[
\bar{\mu} = \frac{N\mu_s + n\mu_t}{N + n}
\]

\[
\bar{\sigma}^2 = \frac{N\sigma_s^2 + n\sigma_t^2}{N + n}
\]

- \(n\): Target batch size
- \(N\): Pseudo batch size

### Graph

- **ASSEMBLENET (SoTA)**
  - **RN50 AM**
    - \(N = 0\)
    - \(N = \infty\)

### Table

<table>
<thead>
<tr>
<th>Batch size</th>
<th>mCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>140</td>
</tr>
<tr>
<td>8</td>
<td>120</td>
</tr>
<tr>
<td>64</td>
<td>100</td>
</tr>
<tr>
<td>512</td>
<td>80</td>
</tr>
<tr>
<td>50000</td>
<td>60</td>
</tr>
<tr>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>80</td>
<td>40</td>
</tr>
<tr>
<td>120</td>
<td>40</td>
</tr>
<tr>
<td>140</td>
<td>40</td>
</tr>
</tbody>
</table>

July 17, 2020

Schneider & Rusak et al. — Improving robustness against common corruptions by covariate shift adaptation
Adaptation (●) consistently improves corruption robustness over Baseline (○) across ImageNet trained models.

![Graph showing comparison between IN-C mCE and IN Top-1 Error across different models](image-url)
Severity of covariate shift correlates with performance degradation.

ImageNet statistics on ImageNet-C

<table>
<thead>
<tr>
<th>category</th>
<th>test corruptions</th>
<th>holdout</th>
</tr>
</thead>
<tbody>
<tr>
<td>blur</td>
<td>defocus</td>
<td>Gaussian</td>
</tr>
<tr>
<td>digital</td>
<td>contrast</td>
<td>saturate</td>
</tr>
<tr>
<td>noise</td>
<td>Gaussian</td>
<td>speckle</td>
</tr>
<tr>
<td>weather</td>
<td>snow</td>
<td></td>
</tr>
<tr>
<td>clean</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

avg $W$ (across layers)
Severity of covariate shift correlates with performance degradation.

ImageNet-C statistics on ImageNet-C

<table>
<thead>
<tr>
<th>Category</th>
<th>Test Corruptions</th>
<th>Holdout</th>
</tr>
</thead>
<tbody>
<tr>
<td>blur, defocus</td>
<td>glass, motion, zoom, Gaussian</td>
<td>x Gaussian, saturate, speckle</td>
</tr>
<tr>
<td>digital</td>
<td>contrast, elastic, pixelate, jpeg</td>
<td>x clean, snow, brightness, spatter</td>
</tr>
<tr>
<td>noise, Gaussian</td>
<td>shot, impulse, -</td>
<td>x clean, snow, brightness, spatter</td>
</tr>
<tr>
<td>weather</td>
<td>snow, clean, clean</td>
<td>x clean, snow, brightness, spatter</td>
</tr>
</tbody>
</table>

July 17, 2020 Schneider & Rusak et al. — Improving robustness against common corruptions by covariate shift adaptation — 27 —
Severity of covariate shift correlates with performance degradation.

ImageNet-C statistics on ImageNet

<table>
<thead>
<tr>
<th>Category</th>
<th>Test Corruptions</th>
<th>Holdout</th>
</tr>
</thead>
<tbody>
<tr>
<td>blur</td>
<td>defocus</td>
<td>Gaussian</td>
</tr>
<tr>
<td>digital</td>
<td>contrast</td>
<td>saturate</td>
</tr>
<tr>
<td>noise</td>
<td>Gaussian</td>
<td>speckle</td>
</tr>
<tr>
<td>weather</td>
<td>snow</td>
<td></td>
</tr>
<tr>
<td>clean</td>
<td>clean</td>
<td></td>
</tr>
</tbody>
</table>

avg $W$ (across layers)

Top-1 error
Large scale pre-training alleviates the need for adaptation.

<table>
<thead>
<tr>
<th>ResNeXt101</th>
<th>ImageNet-C mCE (↘)</th>
</tr>
</thead>
<tbody>
<tr>
<td>32x8d, IN</td>
<td>BN, non-adapt 66.6</td>
</tr>
<tr>
<td>32x8d, IG-3.5B</td>
<td>BN, adapted 56.7 (−9.9)</td>
</tr>
<tr>
<td>32x48d, IG-3.5B</td>
<td>45.7 (−0.1)</td>
</tr>
<tr>
<td>32x48d, IG-3.5B</td>
<td>47.3 (+1.6)</td>
</tr>
</tbody>
</table>
GroupNorm, Fixup better than BN non-adapt, worse than adaptation.

<table>
<thead>
<tr>
<th>Model</th>
<th>Fixup</th>
<th>GroupNorm</th>
<th>BN, non-adapt</th>
<th>BN, adapted</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-50</td>
<td>72.0</td>
<td>72.4</td>
<td>76.7</td>
<td>62.2</td>
</tr>
<tr>
<td>ResNet-101</td>
<td>68.2</td>
<td>67.6</td>
<td>69.0</td>
<td>59.1</td>
</tr>
<tr>
<td>ResNet-152</td>
<td>67.6</td>
<td>65.4</td>
<td>69.3</td>
<td>58.0</td>
</tr>
</tbody>
</table>
Control: Same performance on iid data

![ImageNet and ImageNet-V2 Top-1 Error vs Batch Size graphs](image.png)
Limitation: No gains on more difficult domain shifts (ObjectNet; Barbu et al. ‘19)
Conclusion

- We empirically showed that BN adaptation improves all commonly used models on IN-C, often by 10–15% points.
- Focussing on the ad-hoc scenario ($n = 1$) underestimates model performance.
- Instead, we suggest to report ad-hoc, small sample size ($n = 8$) and full adaptation scores.
- When evaluating robustness on systematic, well-defined corruptions like in ImageNet-C, batch normalization is a strong and very simple baseline. We regard this as the very minimal technique to try in future work. It can be quickly implemented with minimal changes to the source code.

Read our paper at domainadaptation.org/batchnorm
Special thanks to Julian Bitterwolf, Roland S. Zimmermann, Lukas Schott, Mackenzie W. Mathis, Alexander Mathis, Asim Iqbal, David Klindt, Robert Geirhos and other members of the Bethge and Mathis labs for helpful suggestions for improving our manuscript and providing ideas for additional experiments.

We thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting E.R. and St.S.; St.S. acknowledges his membership in the European Laboratory for Learning and Intelligent Systems (ELLIS) PhD program.

This work was supported by the German Federal Ministry of Education and Research (BMBF) through the Tübingen AI Center (FKZ: 01IS18039A), by the Deutsche Forschungsgemeinschaft (DFG) in the priority program 1835 under grant BR2321/5-2 and by SFB 1233, Robust Vision: Inference Principles and Neural Mechanisms (TP3), project number: 276693517.