Improving robustness against common corruptions by covariate shift adaptation

ICML 2020 Workshop on Uncertainty & Robustness in Deep Learning

Steffen Schneider[1 2*], Evgenia Rusak[1 2*], Luisa Eck[3],
Oliver Bringmann[1†], Wieland Brendel[1†], Matthias Bethge[1†]

July 17, 2020

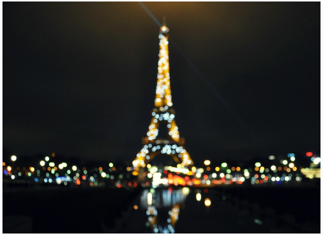
[1]University of Tübingen [2]IMPRS-IS [3]LMU Munich

Web: domainadaptation.org/batchnorm
Contact: steffen@bethgelab.org

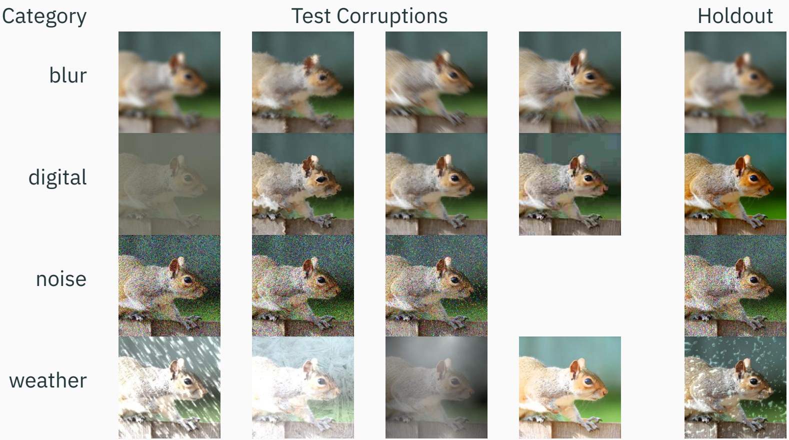





Benchmarking corruption robustness



Benchmarking corruption robustness: ImageNet-C (Hendrycks et al., ‘19)



Benchmarking corruption robustness: ImageNet-C (Hendrycks et al., ‘19)

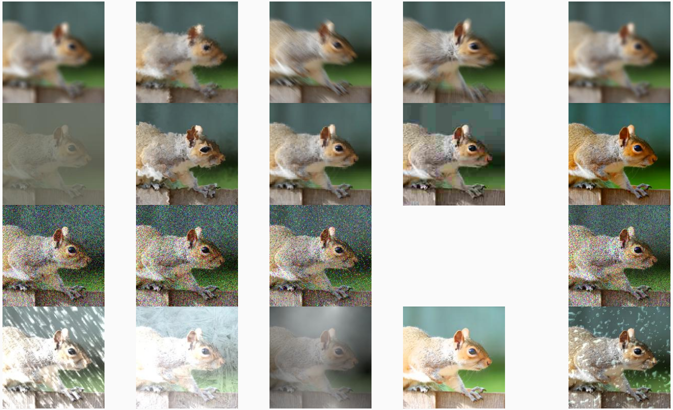
Category | Test Corruptions | Holdout

blur

digital

noise

weather



Mean Corruption Error (lower is better):

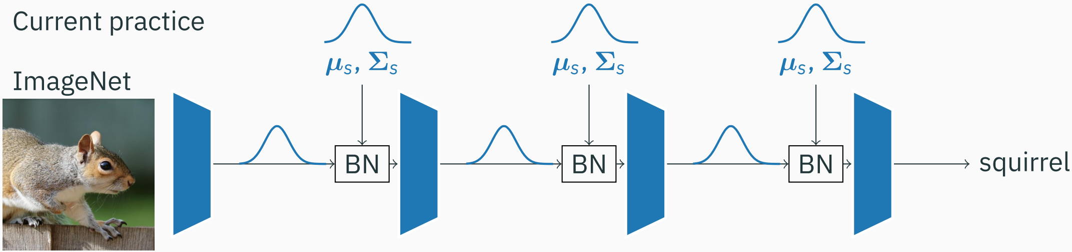
mCE(model)

$$= \frac{1}{C}\sum_{c=1}^{C}\frac{\sum_{s=1}^{S}\mathrm{err}_{c,s}^{\mathrm{model}}}{\sum_{s=1}^{S}\mathrm{err}_{c,s}^{\mathrm{AlexNet}}}$$

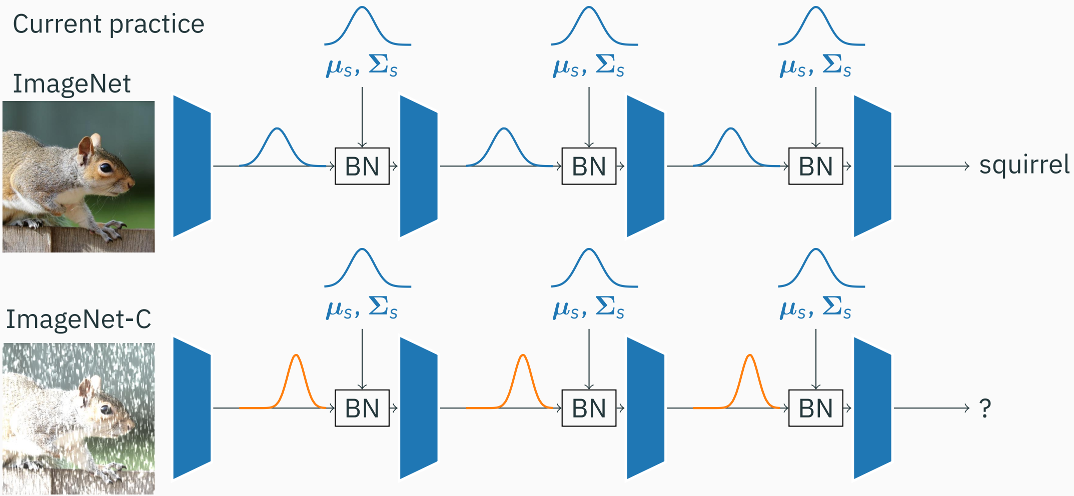
For $C = 15$ test corruptions and $S = 5$ severities.

Adaptation of Batch Norm Statistics

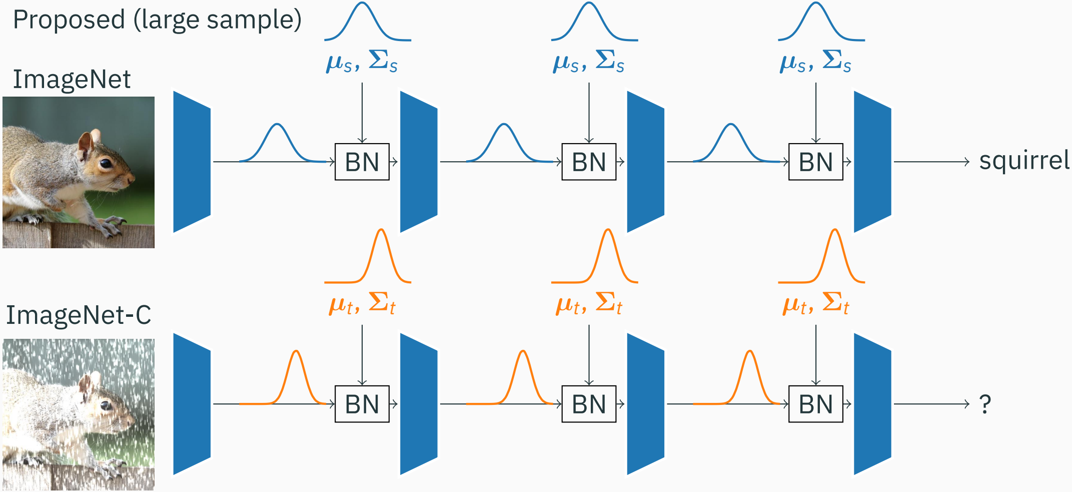
Current practice



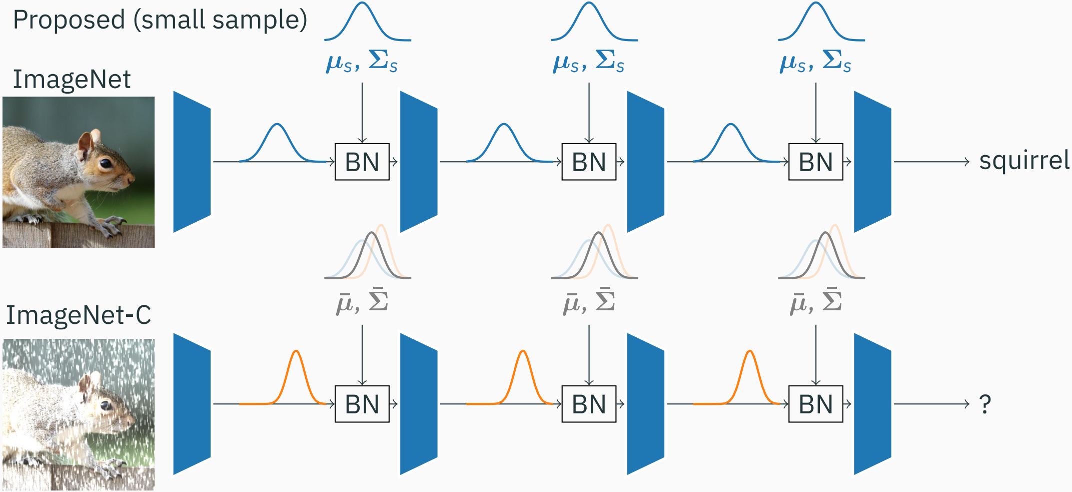
Adaptation of Batch Norm Statistics



Adaptation of Batch Norm Statistics



Adaptation of Batch Norm Statistics



Motivation: Rethinking robustness evaluation

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.

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.

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.

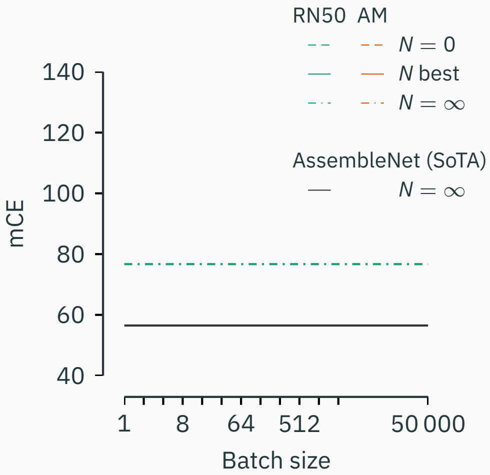
Adaptation boosts robustness of a vanilla trained ResNet-50 model.

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

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

n: Target batch size

N: Pseudo batch size



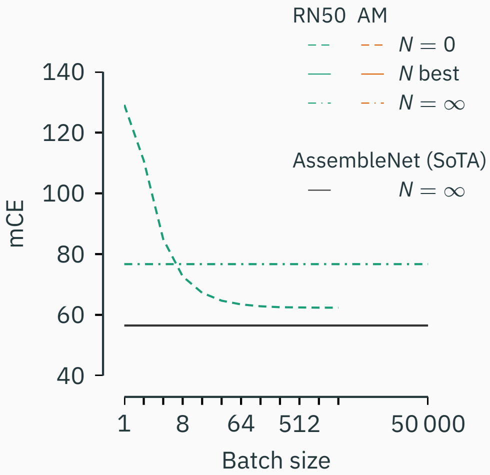
Adaptation boosts robustness of a vanilla trained ResNet-50 model.

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

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

n: Target batch size

N: Pseudo batch size



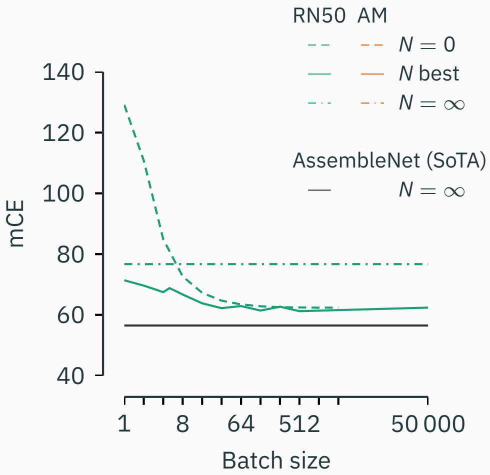
Adaptation boosts robustness of a vanilla trained ResNet-50 model.

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

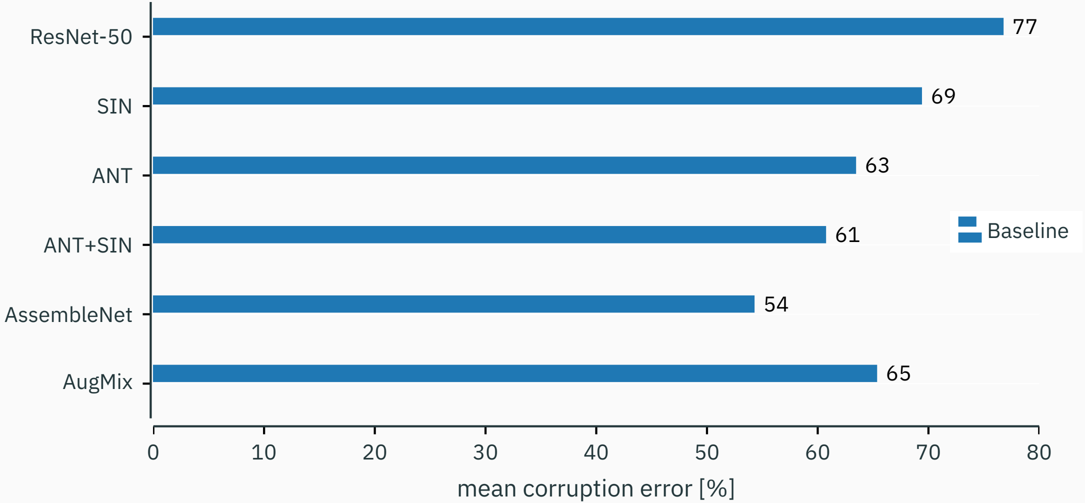
$$\bar{\sigma}^2 = \frac{N\sigma_s^2 + n\sigma_t^2}{N + n}$$

n: Target batch size

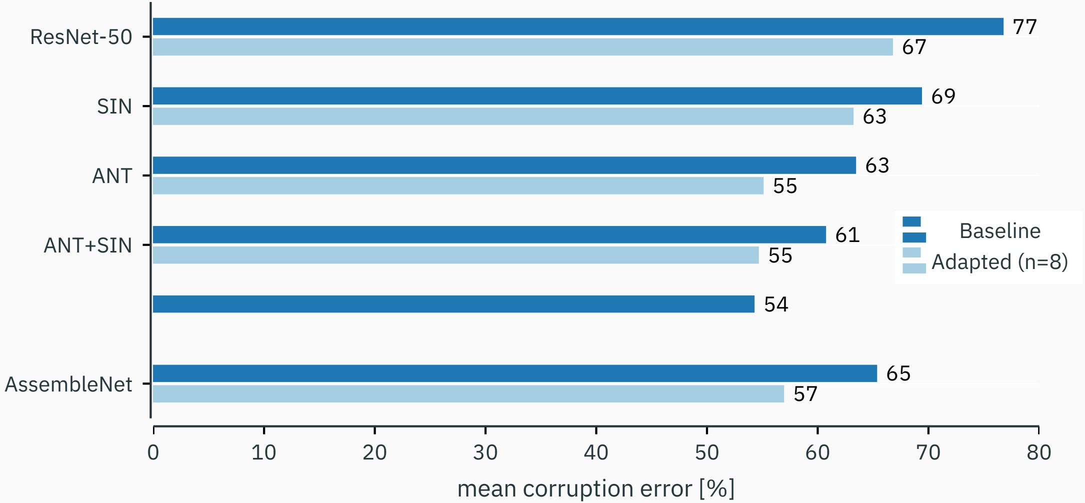
N: Pseudo batch size



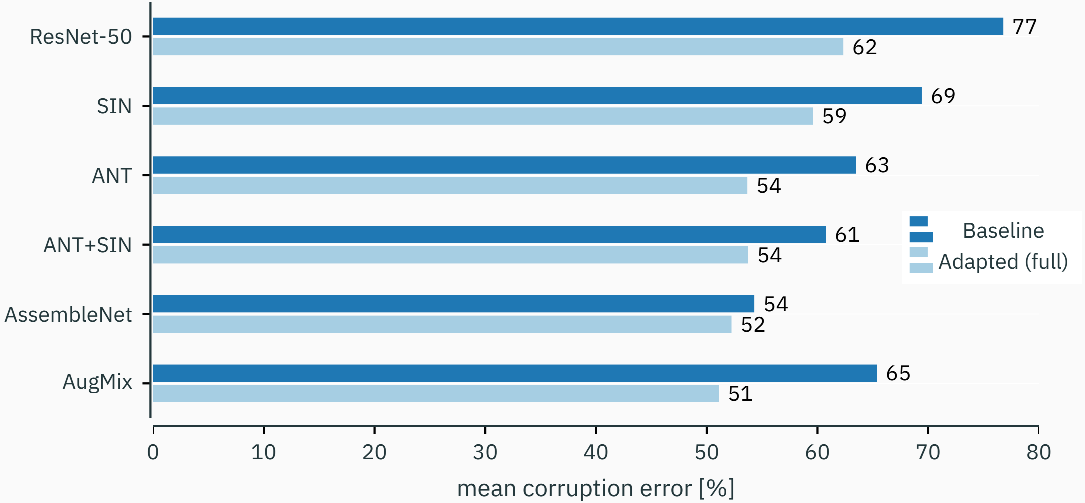
Adaptation yields new state of the art on ImageNet-C for robust models.



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



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



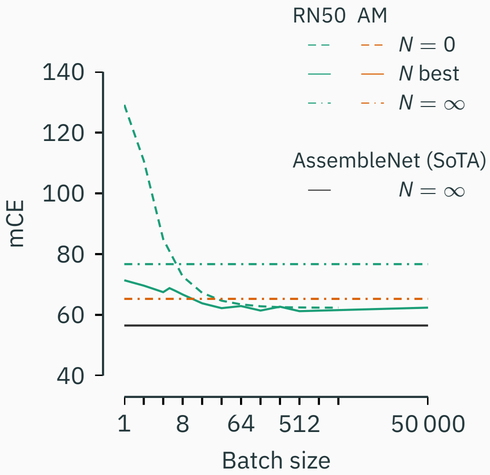
Adaptation yields new state of the art on ImageNet-C for robust models.

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

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

n: Target batch size

N: Pseudo batch size



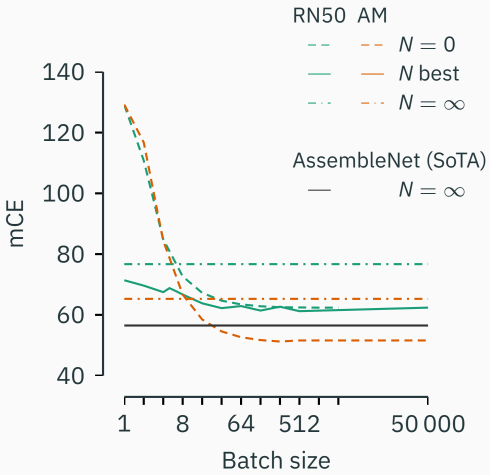
Adaptation yields new state of the art on ImageNet-C for robust models.

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

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

n: Target batch size

N: Pseudo batch size



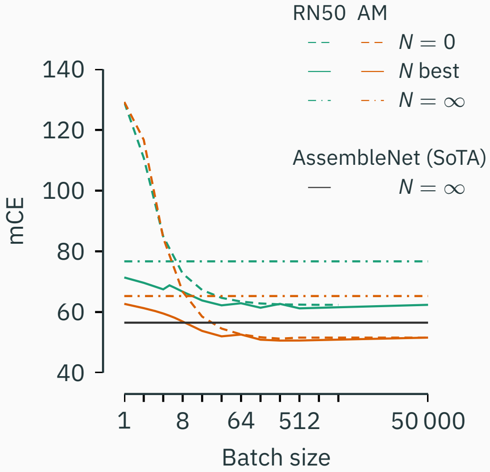
Adaptation yields new state of the art on ImageNet-C for robust models.

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

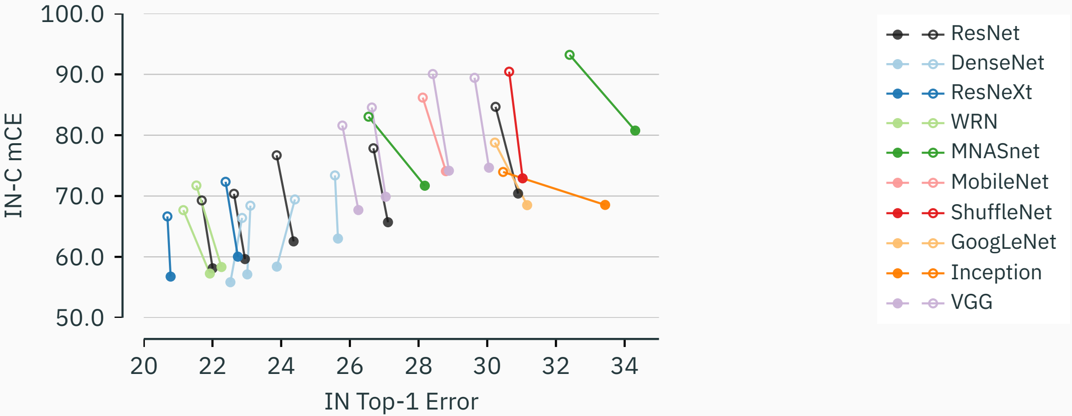
$$\bar{\sigma}^2 = \frac{N\sigma_s^2 + n\sigma_t^2}{N+n}$$

n: Target batch size

N: Pseudo batch size

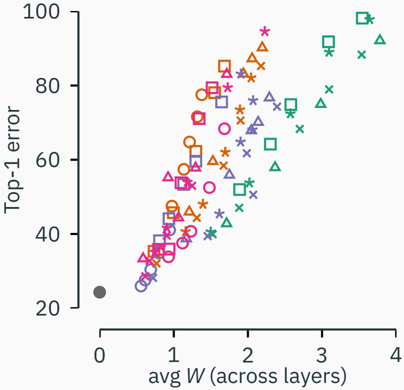


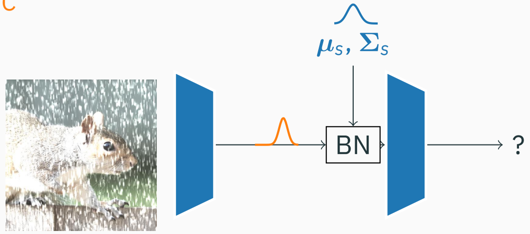
Adaptation [•] consistently improves corruption robustness over Baseline [○] across ImageNet trained models.



Severity of covariate shift correlates with performance degradation.

ImageNet statistics on ImageNet-C

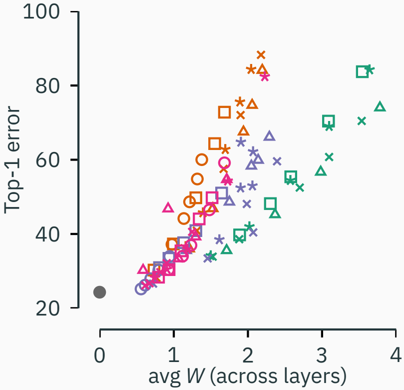


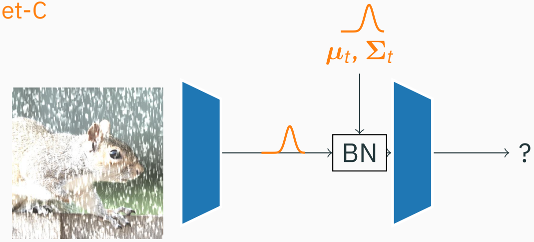


category	test corruptions				holdout
blur	⋆ defocus	△ glass	□ motion	○ zoom	× Gaussian
digital	⋆ contrast	△ elastic	□ pixelate	○ jpeg	× saturate
noise	⋆ Gaussian	△ shot	□ impulse	–	× speckle
weather	⋆ snow	△ frost	□ fog	○ brightness	× spatter
clean	● clean				

Severity of covariate shift correlates with performance degradation.

ImageNet-C statistics on ImageNet-C

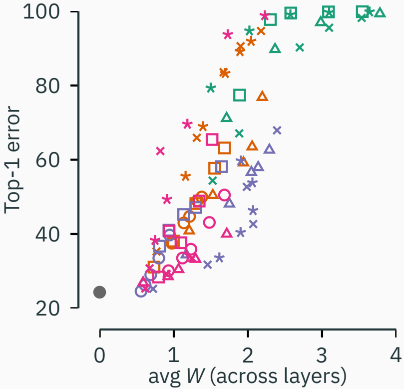


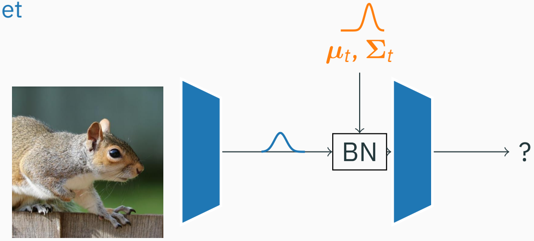


category	test corruptions				holdout
blur	⋆ defocus	△ glass	□ motion	○ zoom	× Gaussian
digital	⋆ contrast	△ elastic	□ pixelate	○ jpeg	× saturate
noise	⋆ Gaussian	△ shot	□ impulse	–	× speckle
weather	⋆ snow	△ frost	□ fog	○ brightness	× spatter
clean	● clean				

Severity of covariate shift correlates with performance degradation.

ImageNet-C statistics on ImageNet





category	test corruptions				holdout
blur	⋆ defocus	△ glass	□ motion	○ zoom	× Gaussian
digital	⋆ contrast	△ elastic	□ pixelate	○ jpeg	× saturate
noise	⋆ Gaussian	△ shot	□ impulse	–	× speckle
weather	⋆ snow	△ frost	□ fog	○ brightness	× spatter
clean	● clean				

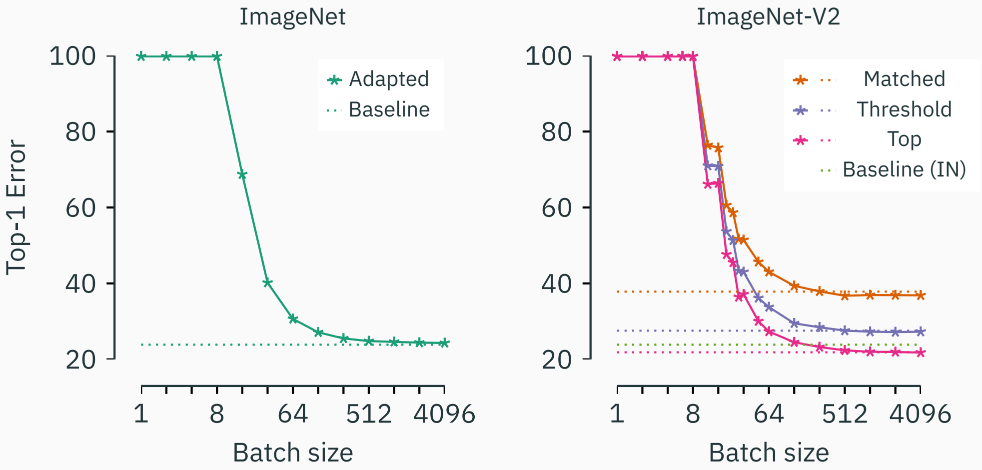
Large scale pre-training alleviates the need for adaptation.

ResNeXt101	ImageNet-C mCE ($\searrow$) BN, non-adapt	ImageNet-C mCE ($\searrow$) BN, adapted
32x8d, IN	66.6	56.7 (–9.9)
32x8d, IG-3.5B	51.7	51.6 (–0.1)
32x48d, IG-3.5B	**45.7**	**47.3** (+1.6)

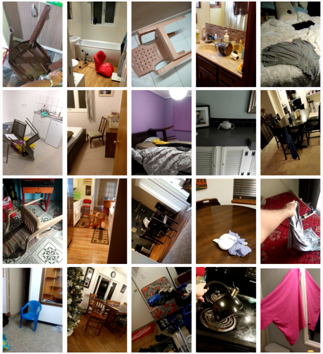
GroupNorm, Fixup better than BN non-adapt, worse than adaptation.

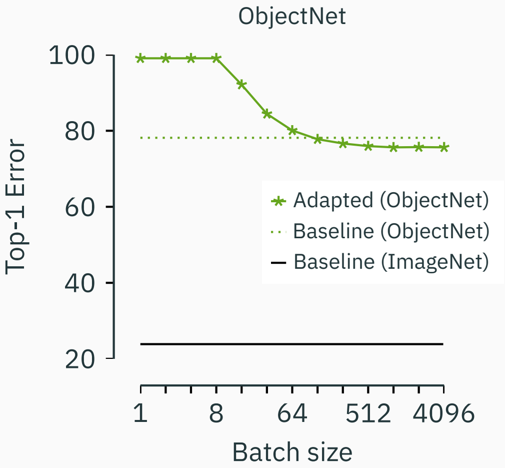
	ImageNet-C mCE (↘)			
Model	Fixup	GroupNorm	BN, non-adapt	BN, adapted
ResNet-50	72.0	72.4	76.7	**62.2**
ResNet-101	68.2	67.6	69.0	**59.1**
ResNet-152	67.6	65.4	69.3	**58.0**

Control: Same performance on iid data



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

Acknowledgements & Funding sources

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.

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

imprs-is