ImageNet-D: A new challenging robustness dataset inspired by domain adaptation

Evgenia Rusak [*1 2] Steffen Schneider [*1 2 3] George Pachitariu [1] Peter Gehler [4] Oliver Bringmann [1]
Matthias Bethge [1] Wieland Brendel [1]

Abstract

We propose a new challenging dataset to benchmark robustness of ImageNet-trained models with respect to domain shifts: ImageNet-D. ImageNet-D has six different domains (“Real”, “Painting”, “Clipart”, “Sketch”, “Infograph” and “Quickdraw”). We show that even state-of-the-art models struggle on this dataset and find that they make well-interpretable errors. For example, our best EfficientNet-L2 model experiences a large performance drop even on the “Real” domain from 11.6% on ImageNet clean to 29.2% on the “Real” domain.

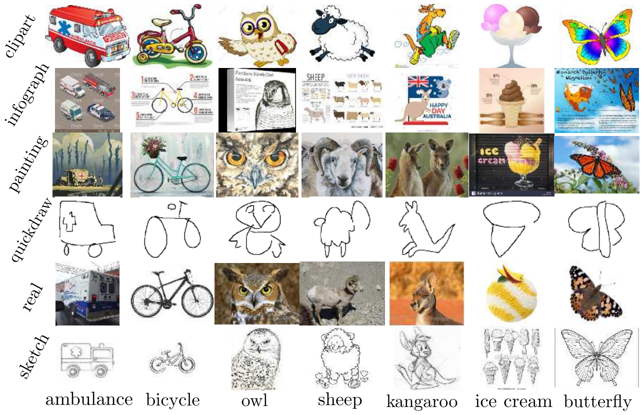

Figure 1. Example images from different domains of ImageNet-D.

Robustness datasets on ImageNet-scale (IN, Deng et al., 2009) have so far been limited to a few selected domains—image corruptions in ImageNet-C (IN-C, Hendrycks & Dietterich, 2019), image renditions in ImageNet-R (IN-R, Hendrycks et al., 2020a), difficult images for ResNet50 (He et al., 2016) classifiers in ImageNet-A (IN-A, Hendrycks et al., 2019) or unusual viewpoints in ObjectNet (Barbu et al., 2019). To enable researchers to benchmark their models on a wider range of complex distribution shifts, we re-purpose the dataset from the Visual Domain Adaptation Challenge 2019 (DomainNet, Saenko et al., 2019) as an additional robustness benchmark. This dataset comes with six image styles: Clipart, Real, Infograph, Painting, Quickdraw and Sketch. To benchmark robustness of IN-trained models out of the box, we filter out the classes that cannot be mapped to IN and refer to the smaller version of DomainNet as ImageNet-D (IN-D). We show example images from IN-D in Fig. 1. The benefit of IN-D over DomainNet is the re-mapping to ImageNet classes which allows robustness researchers to easily benchmark on this dataset, without the need of re-training a model (as is common in Unsupervised Domain Adaptation). ImageNet-D could also be used for studying the task of domain adaptation on ImageNet scale.

Related Work The most similar robustness dataset to IN-D is IN-R which contains renditions of IN classes, such as art, cartoons, deviantart, graffiti, embroidery, graphics and others. The benefit of IN-D over IN-R is that in IN-D, the images are separated according to the domain allowing for studying of systematic domain shifts, while in IN-R, the different domains are not distinguished. ImageNet-Sketch (Wang et al., 2019) is a dataset similar to the ”Sketch” domain of IN-D. We expect models to perform similarly on both datasets.

1. Creation of IN-D and the evaluation protocol

The original DomainNet dataset has 345 classes in total, out of which 164 overlap with IN. To create IN-D, we map these 164 DomainNet classes to 463 IN classes, e.g., for an image from the “bird” class in IN-D, we accept all 39 bird classes in IN as valid predictions. IN also has ambiguous classes, e.g., it has separate classes for “cellular telephone”

[*]Equal contribution [1]University of Tübingen, Germany [2]International Max-Planck Research School for Intelligent Systems (IMPRS-IS) [3]Work done during an internship at Amazon Tubingen [4]Amazon, Tübingen, Germany. Correspondence to: Evgenia Rusak <evgenia.rusak@uni-tuebingen.de>.

ICML 2022 Shift Happens Workshop: Crowdsourcing metrics and test datasets beyond ImageNet., Baltimore, Maryland, USA, 2022.

and “dial phone” or for “analog clock” and “digital clock” and “wall clock”, and others. For these cases, we accept all predictions as valid. In this sense, the mapping from IN-D to IN is a one-to-many mapping.

The mapping was first done by comparing the class labels in DomainNet and the synset labels on IN. Afterwards, the resulting label maps were cleaned manually, because simply comparing class label strings resulted in imperfect matches. For example, images of the class “hot dog” in DomainNet were mapped to the class “dog” in IN. Another issue is that IN synset labels of different animal species do not contain the animal name in the text label, e.g., the class “orangutan, orang, orangutang, Pongo pygmaeus” does not contain the word “monkey” and we had to add this class to the hierarchical class “monkey” manually. We verified the mappings by investigating the class-confusion matrix of the true DomainNet class and the predicted IN classes remapped to DomainNet on the “Real” domain, and checked that the predictions lay on the main diagonal, indicating that IN classes have not been forgotten.

The statistics for the mappings are shown in Table 1. Most IN-D classes (102) are mapped to one single IN class. A few IN-D classes are mapped to more than 20 IN classes: “monkey” and “snake” are mapped to 28 IN monkey and snake species classes, “bird” is mapped to 39 IN bird species classes, “dog” is mapped to 132 IN dog breed classes. The full mapping dictionary can be found in our code online.

The domains in IN-D differ in terms of their difficulty for the studied models. Therefore, to calculate an aggregate score over all six domains, we propose normalizing the error rates by the error achieved by AlexNet on the respective domains to calculate the mean error, following the approach in Hendrycks & Dietterich (2019) for IN-C. This way, we obtain the aggregate score mean Domain Error (mDE) by calculating the mean over different domains,

$$\mathrm{DE}_d^f = \frac{E_d^f}{E_d^{\mathrm{AlexNet}}}, \qquad \mathrm{mDE} = \frac{1}{D}\sum_{d=1}^{D} E_d^f, \qquad (1)$$

where E_d^f is the top-1 error of a classifier f on domain d. The top-1 errors achieved by AlexNet on the different IN-D domains are shown in Table 2.

1.1. Data access and evaluation code

Since IN-D is based on DomainNet, the first step in using IN-D is to download the images from `http://ai.bu.edu/M3SDA/`. We used both the train and test sets of DomainNet to create IN-D. We provide code to map the DomainNet classes to ImageNet classes. The mapping is done via the creation of symbolic links to a new directory containing ImageNet classes such that regular Pytorch (Paszke et al., 2017) dataloaders can be used.

2. Benchmarking state-of-the-art robust models on ImageNet-D

To benchmark models on IN-D, we evaluate the pre-trained and public checkpoints of SIN (Geirhos et al., 2019), ANT (Rusak et al., 2020), ANT+SIN (Rusak et al., 2020), AugMix (Hendrycks et al., 2020b), DeepAugment (Hendrycks et al., 2020a), DeepAug+Augmix (Hendrycks et al., 2020a) and EfficientNet-L2 Noisy Student (Xie et al., 2020). We show the results in Table 3 where we also show reference numbers on IN-C and IN-R.

More robust models perform better on IN-D. Comparing the performance of the vanilla ResNet50 model to its robust DeepAug+Augmix (Hendrycks et al., 2020a) variant which was trained with DeepAugment and AugMix data augmentations, we find that the DeepAug+Augmix model performs better on all domains, with the most significant gains on the “Clipart”, “Painting” and “Sketch” domains. We find that the best performing models on IN-D are also the strongest ones on IN-C and IN-R which indicates good generalization capabilities of the techniques combined for these models, given the large differences between the three considered datasets. However, even the best models perform 20 to 30 percentage points worse on IN-D compared to their performance on IN-C or IN-R, indicating that IN-D might be a more challenging benchmark.

All models struggle with some domains of IN-D. The EfficientNet-L2 Noisy Student model obtains the best results on most domains. However, we note that the overall error rates are surprisingly high compared to the model’s strong performance on the other considered datasets (IN-A: 14.8% top-1 error, IN-R: 17.4% top-1 error, IN-C: 22.0% mCE). Even on the “Real” domain closest to clean IN where the EfficientNet-L2 model has a top-1 error of 11.6%, the model only reaches a top-1 error of 29.2%.

Error analysis on IN-D. We investigate the errors a ResNet50 model makes on IN-D by analyzing the most frequently predicted classes for different domains to reveal systematic errors indicative of the encountered distribution shifts and show the results in Fig. 2. The colors of the bars indicate whether the predicted class is part of the IN-D dataset: “blue” indicates that the class appear in the IN-D dataset, while “orange” means that the class is not present in IN-D. We find most errors interpretable: the classifier assigns the label “comic book” to images from the “Clipart” or “Painting” domains, “website” to images from the “Infograph” domain, and “envelope” to images from the “Sketch” domain. Thus, the classifier predicts the domain rather than the class. We find no systematic errors on the “Real” domain which is expected since this domain should be similar to IN. We find the systematic errors on the “Clipart”, “Painting”, “Sketch” and “Infograph” domains to be consistent to the observation that neural networks tend to focus on object

Table 1. Statistics of one-to-many mappings from IN-D to IN.

Number of IN classes one IN-D class is mapped to	1	2	3	4	5	6	7	8	13	28	39	132
Frequency of these mappings	102	32	13	3	5	1	1	2	1	2	1	1

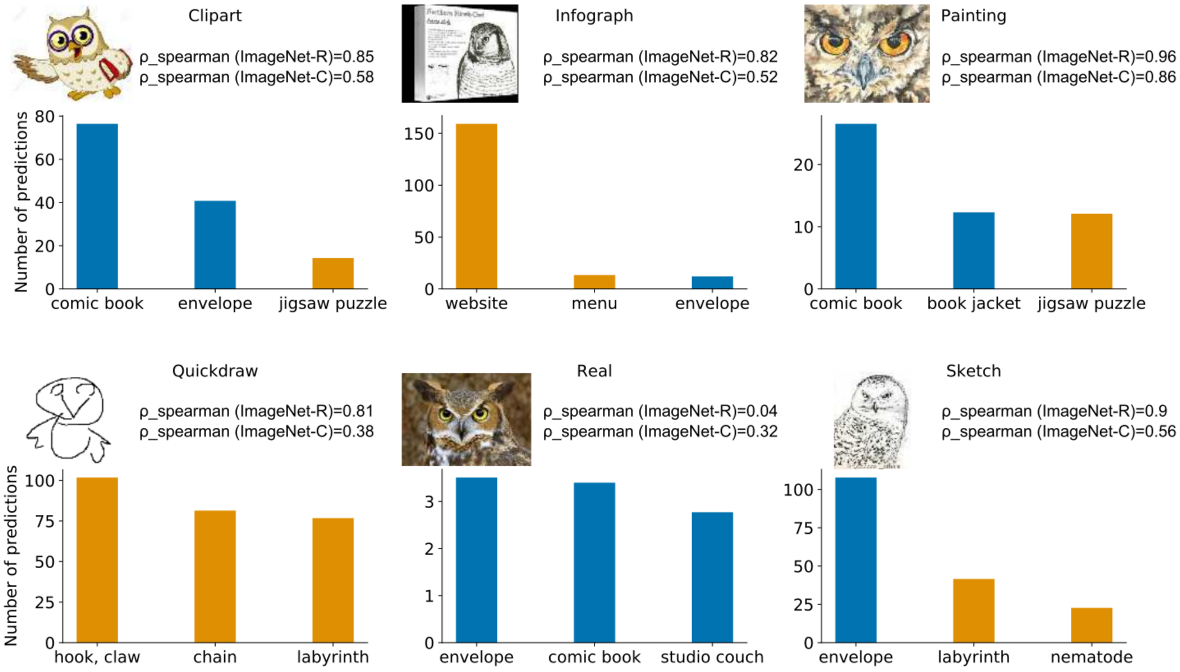

Figure 2. Systematic predictions of a vanilla ResNet50 on IN-D for different domains. The colors of the bars indicate whether the predicted class is part of the IN-D dataset: "blue" indicates that the class appear in the IN-D dataset, while "orange" means that the class is not present in IN-D.

Table 2. top-1 error on IN-D by AlexNet which was used for normalization.

Dataset	top-1 error in %
IN-D real	54.887
IN-D clipart	84.010
IN-D infograph	95.072
IN-D painting	79.080
IN-D quickdraw	99.745
IN-D sketch	91.189

textures rather than object shapes (Geirhos et al., 2019).

We also show the Spearman's rank correlation coefficients for errors on ImageNet-D correlated to errors on ImageNet-R and ImageNet-C for robust ResNet50 models. For this correlation analysis, we take the error numbers from Table 3. We find the correlation to be high between most domains in ImageNet-D and ImageNet-R which is expected since the distribution shift between ImageNet-R and ImageNet is similar to the distribution shift between ImageNet-D and ImageNet. The only domain where the Spearman's rank correlation coefficient is higher for ImageNet-C is the "Real" domain which can be explained with ImageNet-C being closer to real-world data than ImageNet-R. Thus, we find that the Spearman's rank correlation coefficient reflects the similarity between different datasets.

Filtering predictions on IN-D that cannot be mapped to ImageNet-D We perform a second analysis: For a vanilla ResNet50, we filter the predicted labels according to whether they can be mapped to IN-D and report the filtered top-1 errors as well as the percentage of filtered out inputs in Table 4. We note that for the domains "Infograph" and "Quickdraw", the ResNet50 predicts labels that cannot be mapped to IN-D in over 70% of all cases, highlighting the hardness of these two domains.

Filtering labels and predictions on IN that cannot be mapped to IN-D To test for possible class-bias effects, we test the performance of a ResNet50 model on IN classes that can be mapped to IN-D and report the results in Table 4.

Table 3. Top-1 error on IN-D in % as obtained by robust ResNet50 models. For reference, we also show the mCE on IN-C and the top-1 error on IN-R and clean IN. See main text for model references.

Model	Clipart	Infograph	Painting	Quickdraw	Real	Sketch	mDE	IN-C	IN-R	IN
vanilla ResNet50	76.0	89.6	65.1	99.2	40.1	82.0	88.2	76.7	63.9	23.9
SIN	71.3	88.6	62.6	97.5	40.6	77.0	85.6	69.3	58.5	25.4
ANT	73.4	88.9	63.3	99.2	39.9	80.8	86.9	62.4	61.0	23.9
ANT+SIN	68.4	88.6	60.6	95.5	40.8	70.3	83.1	60.7	53.7	25.9
AugMix	70.8	88.6	62.1	99.1	39.0	78.5	85.4	65.3	58.9	22.5
DeepAugment	72.0	88.8	61.4	98.9	39.4	78.5	85.6	60.4	57.8	23.3
DeepAug+Augmix	68.4	88.1	58.7	98.2	39.2	75.2	83.4	53.6	53.2	24.2
EfficientNet-L2 Noisy Student	45.0	77.9	42.7	98.4	29.2	56.4	67.2	16.5	23.5	11.6

Table 4. top-1 error on IN and different IN-D domains for different settings: left column: default evaluation, middle column: predicted labels that cannot be mapped to IN-D are filtered out, right column: percentage of filtered out labels.

Dataset	top-1 error	top-1 error on filtered labels	percentage of rejected inputs
IN val	12.1	13.4	52.7
IN-D real	40.2	17.2	27.6
IN-D clipart	76.1	59.0	59.0
IN-D infograph	89.7	59.3	74.6
IN-D painting	65.2	39.5	42.4
IN-D quickdraw	99.3	96.7	76.1
IN-D sketch	82.1	65.6	47.9

In addition, we map IN labels to IN-D to make the setting as similar as possible to our experiments on IN-D and report the top-1 error (12.1%). This error is significantly lower compared to the top-1 error a ResNet50 obtains following the standard evaluation protocol (23.9%). This can be explained by the simplification of the task: While in IN there are 39 bird classes, these are all mapped to the same hierarchical class in IN-D. Therefore, the classes in IN-D are more dissimilar from each other than in IN. Additionally, there are only 164 IN-D classes compared to the 1000 IN classes, raising the chance level prediction. If we further only accept predictions that can be mapped to IN-D, the top-1 error is slightly increased to 13.4%. In total, about 52.7% of all images in the IN validation set cannot be mapped to IN-D.

Conclusion

We proposed a new challenging dataset (ImageNet-D) to benchmark model robustness. While the error rates on IN-C, -R and -A are at a well-acceptable level for our largest EfficientNet-L2 model, IN-D performance is consistently worse (for all models). We propose to move from isolated benchmark settings like IN-R (single domain) to benchmarks more common in domain adaptation (like DomainNet) and make IN-D publicly available as an easy-to-use dataset for this purpose.

Acknowledgements

This work was supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting ER and StS. ER has been partly supported by the German Federal Ministry of Education and Research (BMBF) in the project Scale4Edge, FKZ: 16ME0129, and by the Ministry of Economic Affairs Baden-Württemberg in the project SEQUOIA under project number 036-840012. WB acknowledges financial support via an Emmy Noether Grant funded by the German Research Foundation (DFG) under grant no. BR 6382/1-1 and via the Open Philantropy Foundation funded by the Good Ventures Foundation. WB is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645. MB is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project No 390727645 and acknowledges support by the German Research Foundation (DFG): SFB 1233, Robust Vision: Inference Principles and Neural Mechanisms, TP 4, Project No: 276693517.

References

Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., and Katz, B. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 9448–9458, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/97af07a14cacba681feacf3012730892-Abstract.html.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Li, F. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer*

Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA, pp. 248–255. IEEE Computer Society, 2009. doi: 10.1109/CVPR.2009.5206848. URL https://doi.org/10.1109/CVPR.2009.5206848.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bygh9j09KX.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

Hendrycks, D. and Dietterich, T. G. Benchmarking neural network robustness to common corruptions and perturbations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. *ArXiv preprint*, abs/1907.07174, 2019. URL https://arxiv.org/abs/1907.07174.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ArXiv preprint*, abs/2006.16241, 2020a. URL https://arxiv.org/abs/2006.16241.

Hendrycks, D., Mu, N., Cubuk, E. D., Zoph, B., Gilmer, J., and Lakshminarayanan, B. Augmix: A simple data processing method to improve robustness and uncertainty. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020b. URL https://openreview.net/forum?id=S1gmrxHFvB.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*, 2017.

Rusak, E., Schott, L., Zimmermann, R., Bitterwolf, J., Bringmann, O., Bethge, M., and Brendel, W. Increasing the robustness of dnns against image corruptions by playing the game of noise. *ArXiv preprint*, abs/2001.06057, 2020. URL https://arxiv.org/abs/2001.06057.

Saenko, K., Peng, X., Usman, B., Saito, K., and Hu, P. *Visual Domain Adaptation Challenge (VisDA-2019)*, 2019. URL http://ai.bu.edu/visda-2019/.

Wang, H., Ge, S., Lipton, Z., and Xing, E. P. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pp. 10506–10518, 2019.

Xie, Q., Luong, M., Hovy, E. H., and Le, Q. V. Self-training with noisy student improves imagenet classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 10684–10695. IEEE, 2020. doi: 10.1109/CVPR42600.2020.01070. URL https://doi.org/10.1109/CVPR42600.2020.01070.