ImageNet-Cartoon and ImageNet-Drawing: Two domain shift datasets for ImageNet

Tiago Salvador [1 2] **Adam M. Oberman** [1 2]

Abstract

Benchmarking the robustness to distribution shifts traditionally relies on dataset collection which is typically laborious and expensive, in particular for datasets with a large number of classes like ImageNet. An exception to this procedure is ImageNet-C (Hendrycks & Dietterich, 2019), a dataset created by applying common real-world corruptions at different levels of intensity to the (clean) ImageNet images. Inspired by this work, we introduce ImageNet-Cartoon and ImageNet-Drawing, two datasets constructed by converting ImageNet images into cartoons and colored pencil drawings, using a GAN framework (Wang & Yu, 2020) and simple image processing (Lu et al., 2012), respectively. Code is available at `https://github.com/oberman-lab/imagenet-shift`.

1. Introduction

The main challenge in testing the robustness to domain shifts lies in data availability. Dataset collection is a laborious and expensive procedure. Previously proposed datasets to test the robustness of ImageNet models include ImageNet-v2 (Recht et al., 2019), ImageNet-Sketch (Wang et al., 2019), ImageNet-A (Hendrycks et al., 2021b), ImageNet-R (Hendrycks et al., 2021a), all of which required collecting new images. The sole exception to this is ImageNet-C (Hendrycks & Dietterich, 2019) that applies synthetic corruptions to the validation set in the original ImageNet dataset. The transformation preserves the image label and therefore there is no need to collect new data. The bulk of the work consisted in carefully defining the corruptions so they reflect real-world transformations. The corruptions proposed include brightness (variations in daylight intensity), gaussian noise (in low-lighting conditions) and defocus blur (when the image is out of focus), among others.

The two datasets we propose here, ImageNet-Cartoon and ImageNet-Drawing, follow the same principle of applying label-preserving transformations as in ImageNet-C. We choose cartoons and colored drawings for two main reasons: (i) for the original ImageNet dataset the annotators were instructed to collect "photos only, no painting, no drawings, etc." (Deng, 2012) and thus cartoons and drawings constitute a natural domain shift; (ii) cartoons and sketches[1] are part of the common domain shifts included in the datasets used in the domain adaptation literature: DomainNet (Peng et al., 2019) includes cliparts and sketches, Office-Home (Venkateswara et al., 2017) includes cliparts and PACS Li et al. (2017) includes cartoons and sketches.

The images in ImageNet-Cartoon are obtained using the GAN framework proposed by Wang & Yu (2020) which converts photo images into cartoons. To generate the images in ImageNet-Drawing we follow the work of Lu et al. (2012) that transforms real photos into colored pencil drawings using image processing alone. In Table 1, we display the accuracies of PyTorch pre-trained ImageNet models with different architectures. While it is reasonable to say that for humans the task difficulty remains essentially the same, there is on average an 18 and 45 percent points accuracy drop for the deep neural network models on ImageNet-Cartoon and ImageNet-Drawing, respectively, which highlights the how challenging these datasets are for current models.

2. Related Work

Deep neural networks and 3D graphics engines have been used to create new datasets as they allow us to generate new labeled images. Nakkiran et al. (2021) introduce CIFAR-5m, a dataset of 6 million synthetic CIFAR-10-like images which are generated with the Denoising Diffusion generative model of Ho et al. (2020). More recently, Li et al. (2022) leverage a BigGAN (Brock et al., 2019) and a VQGAN (Esser et al., 2021) to synthesize a new ImageNet bench-

[1] Department of Mathematics and Statistics, McGill University, Montreal, Canada [2] Mila, Montreal, Canada. Correspondence to: Tiago Salvador <tiago.saldanhasalvador@mcgill.ca>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

[1] In this work, we see sketches as the black and white simplified version of a colored drawings.

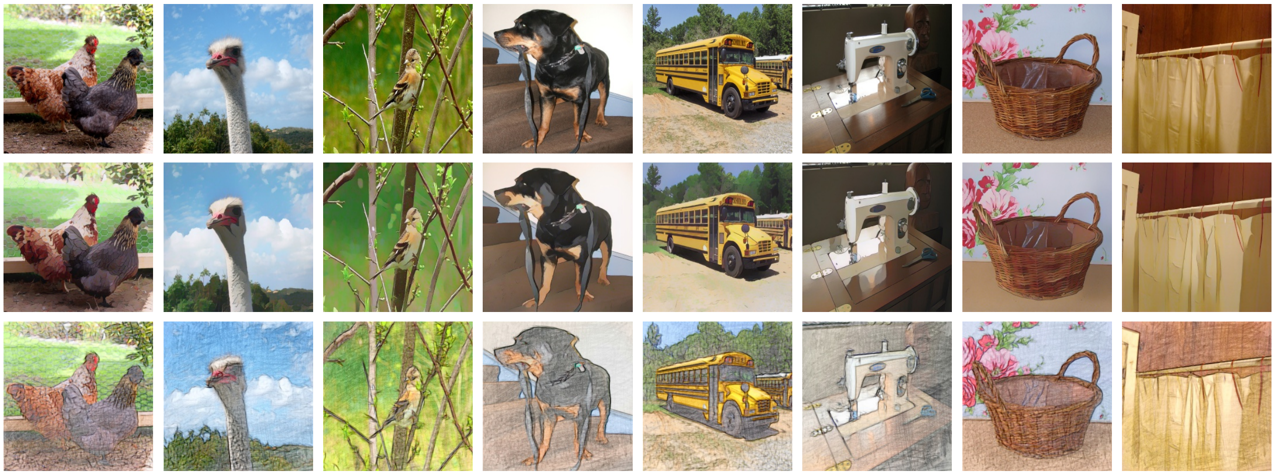

Figure 1: Several examples of ImageNet images (top) and their respective ImageNet-Cartoon (middle) and ImageNet-Drawing (bottom) versions. Additional examples in Figure 5 in the Appendix.

Table 1: Accuracy of pre-trained ImageNet models on the original ImageNet dataset, ImageNet-Cartoon and ImageNet-Drawing.

Architecture	ImageNet	Cartoon	Drawing
AlexNet	56.52	39.10	15.46
VGG-19-BN	74.22	49.63	20.86
ResNet-50	76.13	53.96	23.74
DenseNet121	74.43	57.09	27.59
ResNeXt-101-32x8d	79.31	62.38	31.78
Wide-ResNet-50-2	78.47	57.90	27.12
ViT-B/16	81.07	69.03	46.89
ConvNeXt-S	83.30	69.02	44.63

mark with pixel-wise labels. More pertinent to our work, is the VisDA-C dataset (Peng et al., 2017) whose training set consists of 3D renderings of different objects, while the validation and test sets consist in real-images from two different sources. However, it is limited to only 12 different classes and as pointed in (Li et al., 2022) "creating ImageNet-level of class and instance diversity via the graphics approach would require significant 3D content acquisition efforts".

The aforementioned ImageNet-C (Hendrycks & Dietterich, 2019), where different synthetic corruptions at different intensity levels are applied to ImageNet images, is perhaps the closest work to ours. Similarly, style transformations were considered in (Geirhos et al., 2019) leading to Stylized-ImageNet (SIN): the texture of each image is altered based on the style of a randomly selected painting through AdaIN style transfer (Huang & Belongie, 2017). By exposing the model to SIN images during training the goal was to reduce the texture bias of Convolutional Neural Networks and force the model to recognize objects based on their shape. However, the images in SIN, while partially retaining their shape content, have their overall appearance severely altered with a loss of semantic meaning and thus are not conducive to test the robustness to domain shift (see Figure 2 for an example).

3. Dataset synthesis

In this section we discuss the two methods used to transform the images in the validation set of the (original) ImageNet dataset into cartoons and drawings, forming respectively our proposed ImageNet-Cartoon and ImageNet-Drawing. Examples are shown in Figure 1 and additional ones in Figure 5 in the Appendix.

3.1. ImageNet-Cartoon

Wang & Yu (2020) propose a GAN framework to cartoonize a real photo into a cartoon that we leverage here to generate our dataset ImageNet-Cartoon. Their method relies on decomposing images into three representations: (i) surface representation capturing the smooth surface of the images, mimicking a first rough draft drawn by cartoonists that is later retouched and filled with details; (ii) structure representation that emulates flattened global content, sparse color blocks, and clear boundaries, thus capturing the sparse visual effects; (iii) texture representation, a gray-scale representation retaining the details and edges but independent of color and luminance. Three independent modules are used to extract the above representations, while the GAN framework itself contains a generator G, a fully-convolutional U-Net-like (Ronneberger et al., 2015) network, and two discriminators D_s and D_t, with PatchGAN (Isola et al., 2017) networks, that distinguish the surface and structure representations of the generator output and cartoon images. In addition, a pre-trained VGG network (Simonyan & Zis-

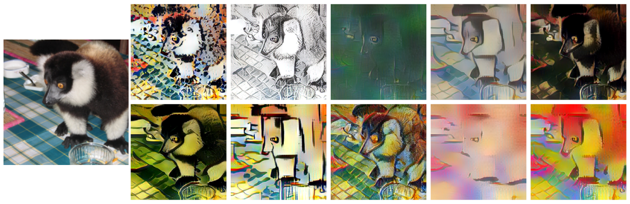

Figure 2: Visualisation of Stylized-ImageNet. Left: randomly selected ImageNet image of class `ring-tailed lemur`. Right: ten examples of images with content/shape of left image and style/texture from different paintings. Image taken from Geirhos et al. (2019).

serman, 2014) is used to extract high-level features and to impose spatial constrains on the global content. We refer the reader to Wang & Yu (2020) for additional details.

Here we make use of the pre-trained generator model G provided in the official Tensorflow implementation[2] to construct ImageNet-Cartoon. If $\boldsymbol{x}$ is an ImageNet image, then we simply store $G(\boldsymbol{x})$ in our dataset. Using a single Tesla P100 PCIe GPU with 16 GB, we are able to generate ImageNet-Cartoon in under 48 minutes. For comparison, downloading Imagenet-Cartoon at an average speed of 10Mbps would take 53 minutes. In the supplemental material, we provide the necessary code to produce ImageNet-Cartoon.

3.2. ImageNet-Drawing

Lu et al. (2012) propose a multi-stage procedure to create colored pencil drawings from real photos. First, a line drawing with strokes, $\boldsymbol{s}$, is produced based on the convolution of kernels representing lines in 8 possible different directions with the gradient norm of the gray-scale version of the input image. Then, tone adjustment is performed leveraging parametric histogram models learned based on statistics from a set of sketch examples producing the tone map $\boldsymbol{j}$. This is then combined with a predefined drawing pattern $\boldsymbol{h}$ to produce the pencil texture rendering $\boldsymbol{t}$. Finally, the pencil stroke $\boldsymbol{s}$ is combined with the tonal texture $\boldsymbol{t}$ through pointwise multiplication generating the grayscale pencil sketch $\boldsymbol{r} = \boldsymbol{s} \cdot \boldsymbol{t}$. The final color pencil drawing is obtained by taking sketch $\boldsymbol{r}$ as the brightness layer (the Y channel in the YUV color space) of the original image. We illustrate the entire process in Figure 4. For the implementation, we used (Daniel, 2018). Perhaps a bit surprising, the choice of the drawing pattern $\boldsymbol{h}$ has a significant impact in the accuracy of deep neural networks. In Table 2, we report the accuracies on ImageNet-Drawing datasets generated with 4 different drawing patterns. The latter are shown in Figure 3. We choose the pattern that causes the most drop in accuracy to form the ImageNet-Drawing dataset.

Table 2: Accuracy of pre-trained ImageNet models on different versions of ImageNet-Drawing created with different drawing patterns.

Architecture	Drawing-I	Drawing-II	Drawing-III	Drawing-IV
AlexNet	15.46	20.31	18.06	31.63
VGG-19-BN	20.86	27.32	25.14	39.75
ResNet-50	23.74	31.11	29.32	45.63
DenseNet121	27.59	34.68	32.29	48.89
ResNeXt-101-32x8d	31.78	38.82	36.83	53.57
Wide-ResNet-50-2	27.12	34.30	31.91	49.35
ViT-B/16	46.89	53.68	51.45	65.01
ConvNeXt-S	44.63	50.72	48.34	61.84

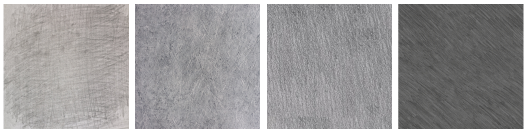

Figure 3: Drawing patterns used to generate different drawing domain shifts. The left most one was used to generate ImageNet-Drawing and they correspond as well to Drawing-I, Drawing-II, Drawing-III and Drawing-IV in Table 2.

4. Evaluation Protocol

ImageNet-Cartoon and ImageNet-Drawing constitute domain shifts with respect to the (original) ImageNet dataset. Therefore they can be used to test the robustness to domain shifts with the metric being the standard classification accuracy. Our experiments show that there is an average of 18 and 45 percent points drop in accuracy for ImageNet-Cartoon and ImageNet-Drawing, respectively (see Table 1). On the other hand, one would expect humans to maintain their accuracy. This highlights the usefulness of the proposed datasets.

In addition, calibration metrics can be considered as well. These are of particular importance as vision models are increasingly used in safety-critical applications and therefore it is important to have a precise estimation of the predictive uncertainty of the models. The most common metric is the Expected Calibration Error (ECE)

$$\mathrm{ECE} = \frac{1}{|B_m|}|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)| \tag{1}$$

Let $f(\boldsymbol{x})$ denote the output of the deep neural network for the input image $\boldsymbol{x}$ with associated label y, after applying the softmax layer. The predicted class $\hat{y}$ is given by the most likely output and the associated score is taken to be the confidence

$$\hat{y}(\boldsymbol{x}) = \underset{k}{\mathrm{argmax}} f(\boldsymbol{x})_k; \quad \hat{c}(\boldsymbol{x}) = \max_k f(\boldsymbol{x})_k \tag{2}$$

Then, given a set of samples $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}\}$, the ECE (Eq. 1) is calculated in two steps. First the confidence

[2] `https://github.com/SystemErrorWang/White-box-Cartoonization`

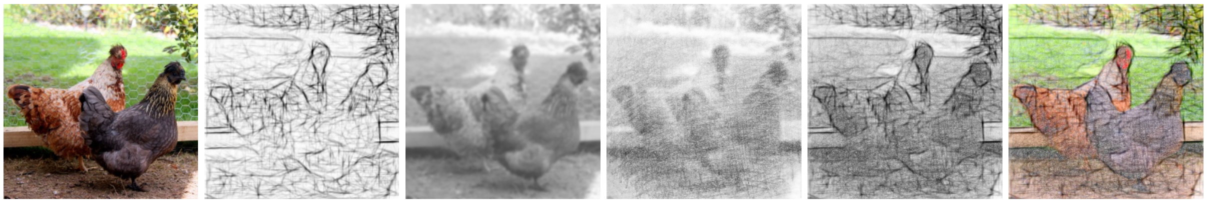

Figure 4: Visualization of the conversion of a real image to a colored pencil drawing. We display, in the following order, the original image, its pencil stroke representation $\boldsymbol{s}$, the tone map $\boldsymbol{j}$, the tonal texture $\boldsymbol{t}$, the grayscale pencil sketch $\boldsymbol{r}$ and the final image.

Table 3: ECE of pre-trained ImageNet models on the original ImageNet dataset, ImageNet-Cartoon and ImageNet-Drawing.

Architecture	ImageNet	Cartoon	Drawing
AlexNet	1.99	5.31	18.95
VGG-19-BN	3.75	9.27	20.11
ResNet-50	3.71	8.28	22.52
DenseNet121	2.52	5.70	16.57
ResNeXt-101-32x8d	8.06	13.71	26.69
Wide-ResNet-50-2	5.29	9.98	23.81
ViT-B/16	5.54	3.90	4.57
ConvNeXt-S	16.59	20.68	10.91

scores $\{\hat{c}^{(1)}, \ldots, \hat{c}^{(n)}\}$ of samples are partitioned into M bins $\{B_m\}_{m=1}^{M}$ of equal mass. Second, the weighted average of the differences between the average confidence $\mathrm{conf}(B_m) = \frac{1}{|Bm|}\sum_{i\in B_m}\hat{c}^{(i)}$ and the accuracy $\mathrm{acc}(Bm) = \frac{1}{|Bm|}\sum_{i\in B_m}\mathbf{1}_{y^{(i)}=\hat{y}^{(i)}}$ in each bin is computed as the ECE metric, where $|B_m|$ denotes the size of bin B_m, $\mathbf{1}$ is the indicator function. While it is common to use equally spaced bins, we emphasize that one should use equal mass bins as described above in order to mitigate the bias in the estimation (Roelofs et al., 2022). In addition, Classwise-ECE (Kull et al., 2019) and Brier score (Brier, 1950) which measure classwise calibration can also be considered. We show the results for the ECE error, as described above, using 15 bins in Table 3. In general, all models are poorly calibrated, some more than other. This is in agreement with Minderer et al. (2021) whose results suggest that the architecture plays a major role in determining calibration properties. In addition, we highlight the following: (i) the ViT-B/16 model has better calibration on ImageNet-Cartoon and Imagenet-Drawing than on ImageNet; (ii) the ConvNeXt-S model, despite achieving the best accuracy on ImageNet and ImageNet-Cartoon it has the worst calibration on those same datasets.

5. Conclusions

By leveraging existing work, we are able to create two new datasets, ImageNet-Cartoon and ImageNet-Drawing, that can be used to test the robustness of models to domain shift. We show that current pre-trained ImageNet models fail to generalize to these datasets exhibiting an 18 and 45 percent points decrease in accuracy on average for ImageNet-Cartoon and ImageNet-Drawing, respectively, in comparison to ImageNet.

Acknowledgements

Thanks to CIFAR for their support through the CIFAR AI Chairs program. TS thanks KC Tsiolis and Vikram Voleti for constructive feedback on the first draft of this manuscript.

References

Brier, G. W. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1 – 3, 1950. doi: 10.1175/1520-0493(1950)078⟨0001:VOFEIT⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/78/1/1520-0493_1950_078_0001_vofeit_2_0_co_2.xml.

Brock, A., Donahue, J., and Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1xsqj09Fm.

Daniel, T. Python implementation of the pencil drawing by sketch and tone algorithm. https://github.com/taldatech/image2pencil-drawing, 2018.

Deng, J. *Large scale visual recognition*. PhD thesis, Princeton University, 2012.

Esser, P., Rombach, R., and Ommer, B. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12873–12883, 2021.

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 *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bygh9j09KX.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., and Gilmer, J. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021a.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. *CVPR*, 2021b.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Huang, X. and Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 1501–1510, 2017.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

Kull, M., Perello Nieto, M., Kängsepp, M., Silva Filho, T., Song, H., and Flach, P. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. M. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Li, D., Ling, H., Kim, S. W., Kreis, K., Barriuso, A., Fidler, S., and Torralba, A. Bigdatasetgan: Synthesizing imagenet with pixel-wise annotations, 2022.

Lu, C., Xu, L., and Jia, J. Combining Sketch and Tone for Pencil Drawing Production. In Asente, P. and Grimm, C. (eds.), *International Symposium on Non-Photorealistic Animation and Rendering*. The Eurographics Association, 2012. ISBN 978-3-905673-90-6. doi: 10.2312/PE/NPAR/NPAR12/065-073.

Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., Tran, D., and Lucic, M. Revisiting the calibration of modern neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

Nakkiran, P., Neyshabur, B., and Sedghi, H. The deep bootstrap framework: Good online learners are good offline generalizers. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=guetrIHLFGI.

Peng, X., Usman, B., Kaushik, N., Hoffman, J., Wang, D., and Saenko, K. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017.

Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., and Wang, B. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do ImageNet classifiers generalize to ImageNet? In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5389–5400. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/recht19a.html.

Roelofs, R., Cain, N., Shlens, J., and Mozer, M. C. Mitigating bias in calibration error estimation. In *International Conference on Artificial Intelligence and Statistics*, pp. 4036–4054. PMLR, 2022.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Venkateswara, H., Eusebio, J., Chakraborty, S., and Panchanathan, S. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

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.

Wang, X. and Yu, J. Learning to cartoonize using white-box cartoon representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.