Towards Systematic Robustness for Scalable Visual Recognition

Mohamed Omran [1 2] **Bernt Schiele** [1 2]

1. Introduction

There is widespread interest in developing *robust* recognition models, that can recognise a variety of object classes despite the presence of unlikely object, scene, or image properties. This is reflected by the proliferation of challenging benchmarks that focus on various kinds of robust recognition behaviour, e.g. robustness to image corruptions (Hendrycks & Dietterich, 2019), to spatial transformations (Engstrom et al., 2019), to harsh weather conditions (Sakaridis et al., 2018), to imperceptible perturbations (Gilmer et al., 2018), and to abstract depictions of objects (Rusak et al., 2021). While it is hard to formulate a compact definition of robustness that covers these diverse requirements and more, we can nonetheless impose a useful meta-requirement that we should pursue in general: For robust behaviour to be *scalable*, it should transfer flexibly across familiar object classes, and not be separately learned for every class of interest. We refer to this problem setting as **systematic robustness**. We argue in ongoing work that current training and evaluation protocols – especially for large-scale image classification – should target this requirement. We also demonstrate with extensive experiments that this remains a multi-faceted challenge for existing classification models and domain generalisation methods. This extended abstract describes the experimental setting and specific instantations thereof that we examine in our work.

While there is significant work on related problems, e.g. systematic generalisation and spurious correlations (Bahdanau et al., 2019; Ruis et al., 2020; Geirhos et al., 2020; Montero et al., 2021; Schott et al., 2022) experiments are typically conducted on small datasets: Either synthetic ones where we have full control over the data generation process and can correspondingly work with tightly-controlled train/test splits, or naturalistic datasets with limited variability but which nonetheless admit some form of control. Here we aim to bridge the gap between work on controlled systematic generalisation and work on large-scale recognition models, in which models are trained on large naturalistic datasets with an "almost anything goes" approach. We also depart from related work on domain generalisation (Gulrajani & Lopez-Paz, 2021; Koh et al., 2021), by acknowledging the need to learn specific robust behaviours from data: We are e.g. unlikely to be able to learn how to handle blurry images by only looking at clean and noisy ones. Noise and blur are perturbations with very different frequency characteristics, and it's not obvious that this should work without the appropriate evaluation. Our experiments also complement work on domain adaptation under label shift (Johansson et al., 2019; Zhao et al., 2019; Ben-David et al., 2010).

We formulate a variety of challenges probing systematic robustness w.r.t. (1) photometric transforms, (2) geometric transforms, (3) rendition styles. Our proposed benchmarks are a compromise between a tightly-controlled but overly simplistic setups and large-scale datasets without meaningful experimental controls, and they build on top of existing data (Russakovsky et al., 2015; Hendrycks & Dietterich, 2019; Peng et al., 2019). See Figure 1 for an example.

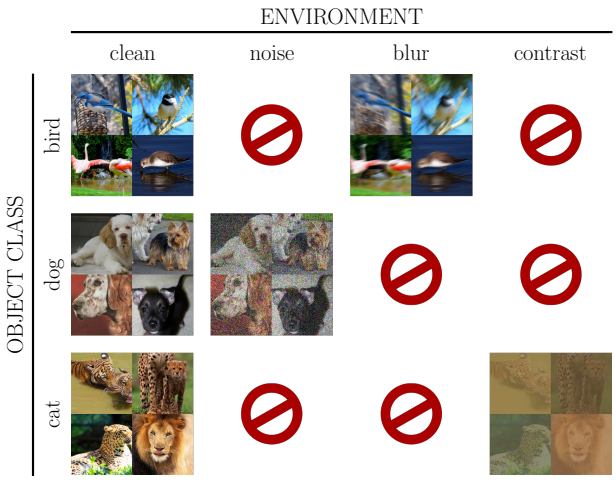

Figure 1. **Systematic Robustness:** We study the task of image classification in the presence of unlikely image properties, e.g. image corruptions, but under strict experimental controls: We systematically exclude certain combinations of class and property from the training data while ensuring that each is individually represented. At test time, we consider all possible combinations, seen and unseen. The goal is to encourage flexible robust behaviour that transfers in a non-trivial manner.

[1]Max Planck Institute for Informatics [2]Saarland Informatics Campus, Germany. Correspondence to: Mohamed Omran <mohomran@mpi-inf.mpg.de>.

Preprint.

2. Systematic Robustness

We now introduce the problem setting of systematic robustness, describe a simple metric for measuring the latter, and describe different problem instantiations that we examine.

Preliminaries In this work, we focus on the task of image classification. Our goal is to train a classifier to recognise object classes from some closed set $\mathrm{C} = \{\mathrm{c}_i\}_{i=1}^{N}$ given a training set of images annotated with class labels. The training data is drawn from a set of "environments" or "domains" $\mathcal{E} = \{\mathrm{E}_0, \ldots, \mathrm{E}_M\}$. Each environment $\mathrm{E}_k = \mathrm{T}_k \times \mathrm{C}_k$ can be characterised by a set of classes C_k and some property T_k that applies only to images from that environment, e.g. "contain Gaussian noise", " are rotated by 90° ", and "objects are depicted with line drawings". E_0 is the default environment in which all classes C are present. The remaining training environments are restricted to disjoint subsets of classes $\{\mathrm{C}_1, \ldots, \mathrm{C}_M\}$, s.t.: (1) $\bigcup_{k=1}^{M} \mathrm{C}_k \subseteq \mathrm{C}$, and (2) $\forall (i,j) \in \{1, \ldots, M\} : i \neq j \Rightarrow \mathrm{C}_i \cap \mathrm{C}_j = \emptyset$.

At test time, we consider all possible environments: $\{\mathrm{T}_0 \times \mathrm{C}\} \cup \{\mathrm{T}_i \times \mathrm{C}_j \mid i,j \in \{1, \ldots, M\}\}$. Thus for any valid i, j, if $i \neq j$ then $\mathrm{T}_i \times \mathrm{C}_j$ has not been seen during training. This allows us to contrast performance in seen and unseen environments, and determine how well robust behaviour learned for a subset of available classes transfers to the rest. This also imposes strict restrictions on training settings, and it is vital that methods not violate the letter and spirit of the exercise by resorting to augmentations that fill in the intentional gaps of the data distribution.

Measuring Systematic Robustness To measure systematic robustness of a model f w.r.t. some scene factor T_k, we first need to measure classification accuracy $\mathrm{acc}^f(\cdot)$ in two environments: (1) the control environment $\mathrm{T}_k \times \mathrm{C}_k$, i.e. the set of classes C_k for which robustness to T_k was learned during training, and (2) the experimental environment $\mathrm{T}_k \times \mathrm{C}_{\overline{k}}$, i.e. the environment in which the property T_k holds for classes $\mathrm{C}_{\overline{k}} = \mathrm{C} \setminus \mathrm{C}_k$. We can then simply normalise the accuracy in the experimental environment $\mathrm{acc}^f(\mathrm{T}_k \times \mathrm{C}_{\overline{k}})$ by the accuracy in the control environment $\mathrm{acc}^f(\mathrm{T}_k \times \mathrm{C}_k)$. This represents the degree to which robust behaviour learned in a seen environment transfers to an unseen one. Sometimes it will make sense to compensate for some baseline robustness, e.g. if we are fine-tuning a pre-trained model f_0 that already shows some degree of robust accuracy. We simply subtract average robust accuracy of the baseline model $\mathrm{acc}^{f_0}(\mathrm{T}_k \times \mathrm{C})$ from the accuracy measured in both environments. We refer to the resulting quantity as ρ^k, i.e. systematic robustness w.r.t. T_k, and compute it as follows while limiting the range to $[0, 1]$:

$$\rho^k = \min(1,\ \frac{\max(0,\ \mathrm{acc}^f(\mathrm{T}_k \times \mathrm{C}_{\overline{k}}) - \mathrm{acc}^{f_0}(\mathrm{T}_k \times \mathrm{C}))}{\min(1,\ \mathrm{acc}^f(\mathrm{T}_k \times \mathrm{C}_k) - \mathrm{acc}^{f_0}(\mathrm{T}_k \times \mathrm{C}))}) \tag{1}$$

It should be noted that a value of 1 is trivial to achieve if f is not much more robust than the reference classifier f_0. Thus on its own, this measure of robustness is insufficient and has to be paired with the other metrics that measure absolute robust accuracy. A related metric is that of *effective robustness* (Taori et al., 2020) which also relates robust accuracy to i.i.d. accuracy.

Problem Instantiations We opt to study three instantiations of this problem: systematic robustness w.r.t. (1) image corruptions (Hendrycks & Dietterich, 2019), (2) in-plane image rotations (Engstrom et al., 2019), and (3) rendition styles (e.g. natural images, line drawings) (Hendrycks et al., 2021; Rusak et al., 2021). These represent basic challenges in recognition that robust models should arguably handle in a flexible manner. They also differ qualitatively: The first are photometric perturbations that can drastically affect image statistics, whereas recognising objects from different viewpoints can require reasoning of a more geometric nature. While the way an object is depicted can also heavily affect the statistics of the resulting image, successfully managing this challenge requires non-trivial abstract reasoning about the essence of an object's texture and shape. We study these problems in the large-scale setting building on prior efforts to collect the relevant data. We study systematic robustness w.r.t. corruptions and rotations by applying selective data augmentation to *ImageNet* (Russakovsky et al., 2015; Hendrycks & Dietterich, 2019), and w.r.t. rendition styles by sampling images from *DomainNet* (Peng et al., 2019).

For image corruptions, we select four out of the 19 suggested for the *ImageNet-C* benchmark, each representing one of four categories: Gaussian noise ($\mathcal{N}$), Gaussian blur ($\mathcal{B}$), contrast ($\mathcal{D}$), and snow ($\mathcal{W}$). Thus we have up to five environments including the default "clean" one. To generate "corrupted" images, we sample one of five severity levels and apply the corresponding transformation. For image rotations, we have just one additional environment ($\mathcal{R}$) where images are randomly rotated by an angle sampled from a discrete set. When reporting performance for image corruptions, we measure average accuracy or systematic robustness over the available severity levels. For rotations, we restrict evaluation to right angles $\{90°, 180°, 270°\}$ as these do not introduce a confounding effect via pixel interpolation.

With image renditions, we use natural (or "real") images for the default environment, and then sample images corresponding to the remaining five rendition styles ("clipart", "quickdraw", "infograph", "sketch", and "painting") while restricting each style to a subset of the available classes.

Conclusion We believe that our proposed setting represents a step towards putting large-scale classification methods on a more solid experimental footing, by targeting a more nuanced evaluation of model behaviour. Our extensive experimental analysis will be the subject of a separate publication.

References

Bahdanau, D., Murty, S., Noukhovitch, M., Nguyen, T. H., de Vries, H., and Courville, A. C. Systematic generalization: What is required and can it be learned? In *International Conference on Learning Representations, ICLR*, 2019.

Ben-David, S., Lu, T., Luu, T., and Pál, D. Impossibility theorems for domain adaptation. In Teh, Y. W. and Titterington, D. M. (eds.), *International Conference on Artificial Intelligence and Statistics, AISTATS*, 2010.

Engstrom, L., Tran, B., Tsipras, D., Schmidt, L., and Madry, A. Exploring the landscape of spatial robustness. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *International Conference on Machine Learning, ICML*, 2019.

Geirhos, R., Jacobsen, J., Michaelis, C., Zemel, R. S., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, Nov 2020.

Gilmer, J., Adams, R. P., Goodfellow, I. J., Andersen, D. G., and Dahl, G. E. Motivating the rules of the game for adversarial example research. *CoRR*, abs/1807.06732, 2018.

Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. In *International Conference on Learning Representations, ICLR*, 2021.

Hendrycks, D. and Dietterich, T. G. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations, ICLR*, 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. In *IEEE/CVF International Conference on Computer Vision, ICCV*, 2021.

Johansson, F. D., Sontag, D. A., and Ranganath, R. Support and invertibility in domain-invariant representations. In Chaudhuri, K. and Sugiyama, M. (eds.), *International Conference on Artificial Intelligence and Statistics, AISTATS*, 2019.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B., Haque, I., Beery, S. M., Leskovec, J., Kundaje, A., Pierson, E., Levine, S., Finn, C., and Liang, P. WILDS: A benchmark of in-the-wild distribution shifts. In Meila, M. and Zhang, T. (eds.), *International Conference on Machine Learning, ICML*, 2021.

Montero, M. L., Ludwig, C. J., Costa, R. P., Malhotra, G., and Bowers, J. The role of disentanglement in generalisation. In *International Conference on Learning Representations, ICLR*, 2021.

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

Ruis, L., Andreas, J., Baroni, M., Bouchacourt, D., and Lake, B. M. A benchmark for systematic generalization in grounded language understanding. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Annual Conference on Neural Information Processing Systems NeurIPS*, 2020.

Rusak, E., Schneider, S., Gehler, P. V., Bringmann, O., Brendel, W., and Bethge, M. Adapting imagenet-scale models to complex distribution shifts with self-learning. In *IEEE/CVF International Conference on Computer Vision, ICCV*, 2021.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. S., Berg, A. C., and Fei-Fei, L. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3): 211–252, 2015.

Sakaridis, C., Dai, D., and Gool, L. V. Semantic foggy scene understanding with synthetic data. *Int. J. Comput. Vis.*, 126(9):973–992, 2018.

Schott, L., von Kügelgen, J., Träuble, F., Gehler, P. V., Russell, C., Bethge, M., Schölkopf, B., Locatello, F., and Brendel, W. Visual representation learning does not generalize strongly within the same domain. In *International Conference on Learning Representations, ICLR*, 2022.

Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., and Schmidt, L. Measuring robustness to natural distribution shifts in image classification. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Annual Conference on Neural Information Processing Systems, NeurIPS*, 2020.

Zhao, H., des Combes, R. T., Zhang, K., and Gordon, G. J. On learning invariant representations for domain adaptation. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *International Conference on Machine Learning, ICML*, 2019.