Rendezvous between Robustness and Dataset Bias: An empirical study

Prabhu Pradhan*
Max Planck Institute for Intelligent Systems (MPI-IS), Tübingen, Germany.
ppradhan@tue.mpg.de

Ruchit Rawal*
Netaji Subhas Institute of Technology (NSIT), New Delhi, India.
ruchitr.ec.17@nsit.net.in

Gopi Kishan
Indian Institute of Technology Roorkee, IIT-R, India.
gkishan@cs.iitr.ac.in

Abstract

We aim to shine a light on the effects of various techniques for perturbarial robustness on dataset-distribution bias (i.e. class imbalance). This relationship between data-skewness and such performance-enhancing measures remains largely unexplored. Deep learning models are seeing real-world deployment, hence it's crucial to gauge the reliability of neural networks since undetected (side)effects of robustness-enhancement techniques on dataset-bias could be catastrophic. Our focus will also be on efficient/compact models, since studies have shown them to have relatively inferior generalization capability. We will evaluate methods for model robustness (against common corruptions as well as adversarial perturbations) by their effects on dataset bias through a variety of specialized metrics.

1 Introduction/Motivation

Fueled by large-scale annotated data and inexpensive compute, Deep Learning has witnessed unprecedented growth across a plethora of domains. Although extremely proficient at classifying patterns in high-dimensional benchmark datasets, deep learning still fares poorly on real-world data (generally, long-tailed distribution where some (head) classes have many examples, and most (i.e. tail) classes have few). ones) (Wang, Ramanan, and Hebert, 2017). Feldman (2019) in their work demonstrated that learning algorithms tend to "memorize" examples in the training-set when operated on long-tailed data leading to degraded performance on the under-represented classes. This may prove detrimental in scenarios that demand generalization on all the classes or in high-stakes real-world settings such as self-driving cars and medical diagnostics where rare cases may be critical.

Machine learning algorithms are based on an implicit assumption that datapoints encountered at the test time are governed by the same distribution as the one's seen during training time. This assumption, however, doesn't hold merit in the wild. For example, a self-driving car's training data might not comprehend all possible weather and lighting conditions, bearing disastrous consequences when deployed in the real world. Also, a lack of robustness against adversarial examples have been observed (Goodfellow, Shlens, and Szegedy, 2015), making physical-world adversarial attacks on perception systems feasible. Moreover, Ilyas et al. (2019) in their work demonstrated that adversarial examples can be directly attributed to the presence of non-robust features i.e. features (derived from patterns in the data distribution) that are highly predictive, yet brittle and (thus) incomprehensible to humans, further validating an ever-pressing need to build an adversarially robust model. Consequently, model robustness to different forms of image perturbations (noise, blurs, adversarial examples, stimulated weather-artifacts, etc) has become a growing concern in machine learning research as well as AI deployment in real-world applications.

*Equal contribution.

Pre-registration Workshop @ 34th Neural Information Processing Systems (NeurIPS 2020), Virtual.

Most of the discussion around the utility of various robustness enhancement techniques is centered on the trade-off between robustness to image perturbations and basic metrics such as accuracy. However, this trade-off is futile in cases of long-tailed distribution (which are common in the real world) as poor-performance on under-represented features would only cause minimal changes to accuracy. Buolamwini and Gebru (2018) in their work observe that darker skin tones have a high misclassification rate due to under-representation in facial analysis datasets, indicating poor predictive performance on under-represented classes often corresponds to serious reliability and fairness considerations (Hooker et al., 2020). Thus, understanding the relationship between model robustness and dataset-bias is essential to the deployment of machine learning solutions.

In this work, we aim to tap into the effects (if any) of improving model robustness on the dataset-distribution bias, and since basic metrics (such as accuracy) fail to give deeper insights (Buda, Maki, and Mazurowski, 2018; Lipton and Steinhardt, 2019) we will use specialized metrics for imbalance (F-1 score, balanced accuracy). Given the surge of interest in image perturbations, we focus our study on robustness to common corruptions and adversarial-perturbation. Despite several papers on robustness to diverse perturbations, robustness-transfer, etc, the relational dynamics between these techniques and imbalance bias is currently absent.

The proposed contributions of this study are:

- Insights on the relational dynamics between robustness-enhancement methods and Imbalance-bias (either mitigation or amplification). Since counter-productivity (if any) of these methods on dataset-distribution bias remains unknown.
- To encompass a variety of contamination strategies, we use a variety of robustness schemes i.e. against common corruptions as well as adversarial samples (specifically, universal adversarial perturbations)
- To widen the scope and applicability of our work, we use specialized (bias-specific) metrics (such as F1-score, balanced-accuracy) which is uncommon in the literature related to robustness.
- We focus on compact networks such as SqueezeNet owing to their widespread adoption in computer-constraint environments.

In section 2, we present current research efforts in these parallel topical domains.

2 Related Work and Literature Trends

Mitigating adverse effects of imbalance data is a well-studied problem in the machine learning literature. Popular methods to deal with class-imbalance are re-sampling (Buda, Maki, and Mazurowski, 2018; Byrd and Lipton, 2019), (which includes oversampling the minority class or undersampling the majority class to equalize the class-count), and Re-weighting (Huang et al., 2016, 2019) (which assigns a distinct set of weights to different classes. More recently, the focus has been to deal with imbalance at the classifier's level instead of data, giving rise to numerous methodologies to deal with class-imbalance such as label-distribution sensitive loss functions (Cui et al., 2019; Cao et al., 2019), learning rate schedules (Smith, 2015) , training routines (Izmailov et al., 2018), etc. Rawal and Pradhan (2020a) in their work provide a comparative overview of these diverse methodologies.

As deep learning is increasingly applied to open-world perception problems in safety-critical domains such as robotics and autonomous driving, its robustness properties become of paramount importance. Researchers have tried to train robust Deep Neural Networks by training on corrupted images, however, recent work by Geirhos et al. (2018) and Vasiljevic, Chakrabarti, and Shakhnarovich (2016) has shown that Deep Neural Networks tend to memorize specific distortions shown during training and fail to generalize on unseen distortions. ImageNet-C benchmark (Hendrycks and Dietterich, 2019) proposed measuring performance on unseen corruptions as a yardstick for the model's generalization capability. Gaussian noise, Pepper noise, Glass blur, Zoom blur, Pixelation, JPEG compression are a few examples of the common corruptions present in ImageNet-C that manifest themselves in real-world settings via electronic noise, lighting conditions, bit errors, lossy-compression, etc.

In addition to these standard corruptions, we also have their worst-case scenarios- commonly known as Adversarial perturbations which when added to an input image, cause the network output to change drastically without making perceptible changes to the input image. While per-instance adversarial

perturbation varies for different samples in a dataset, there exist image-agnostic perturbations called universal adversarial perturbations (UAPs) as introduced by Moosavi-Dezfooli et al. (2017), that can fool state-of-the-art (SOTA) recognition models on most natural images with high probability. These UAPs are more interesting than input-specific ("per-image") ones as UAPs reveal systemic vulnerabilities that models are sensitive to regardless of the input and they will be the focus of our study. UAPs are more likely to transfer effectively between models that learn similar features (Carlini and Wagner, 2017; Fawzi, Moosavi-Dezfooli, and Frossard, 2016). An analysis of universal perturbation and their properties is provided by Dezfooli et al. (2018); Chaubey et al. (2020).

Taori et al. (2020) in their work observed that though much progress has been made in building models that are robust to synthetic distributions-shifts (common-corruptions, adversarial perturbations, etc), this robustness doesn't necessarily transfer to distribution shifts that arise naturally from real-data without any synthetic modification. They argue that since models are likely to encounter natural distribution shifts when deployed in the wild, new algorithmic ideas, as well as a better understanding of how training data affects the robustness, is the need of the hour. Independently, Hooker et al. (2020) in their recent work illustrated that several quantization and pruning techniques disproportionately affect performance on under-represented features in long-tailed scenarios.

3 Methodology and Experimental Protocol

To gauge the effects of robustness enhancing techniques on a model's ability to generalize in the presence of dataset bias (class imbalance), we evaluate various imbalance-focused metrics namely balanced accuracy, F1 score on a diverse set of deep neural networks.

We aim to perform a comparative study of various robustness-enhancing techniques (along with the corresponding baselines) for operating on skewed datasets that are suffering from class-imbalance problems. Extensive experiments are conducted on artificially induced long-tailed CIFAR100 (Krizhevsky, 2009), and Droughtwatch (Hobbs and Svetlichnaya, 2020) (a real-world imbalanced dataset). The experiments are repeated multiple times by gradually increasing the degree of imbalance in the datasets to better understand the effects of robustness-enhancing techniques on model's generalization capability in presence of data-skewness.

- **Base Networks**: We will perform our experiments on distinct base models. We begin with a vanilla Convolutional Neural Network and a standard ResNet-50 (He et al., 2016). Efficient and compact networks (arguably) have inferior generalization capabilities, (Brutzkus and Globerson, 2019; Yehudai and Shamir, 2019) but are gaining popularity due to deployability. A rigorous investigation of the effects of robustness enhancement on efficient networks in presence of skewed datasets is vital to deploying trustworthy AI solutions, thus we also will evaluate a widely popular compact networks *viz.* SqueezeNet (Iandola et al., 2016).
- **Datasets**: We perform comprehensive experiments on Droughtwatch (Hobbs and Svetlichnaya, 2020) and artificially induced long-tailed CIFAR100 (Krizhevsky, 2009) dataset. The

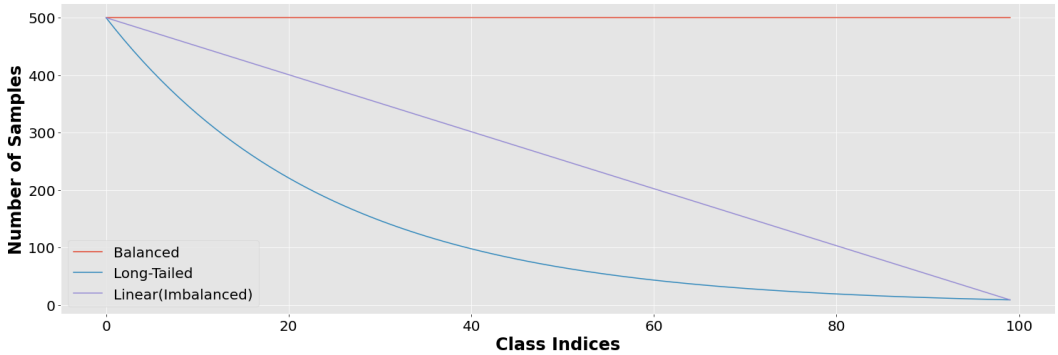

Figure 1: Proposed Levels of Distribution Bias in CIFAR-100; Training samples per class in artificially induced long-tailed CIFAR-100 with increasing degree of imbalance

dataset is highly imbalanced (roughly 60% of the data gathered is of class 0, classes 1 and 2 have 15% each, and the remaining 10% is class 3).

Cui et al. (2019) constructed artificially induced long-tailed CIFAR100 dataset by reducing training-samples from each class as per an exponential function $n = n_i\lambda^i$ where i denotes the class-index, n_i refers to original samples in the class and $\lambda \in (0, 1)$. To better understand the intricacies of robustness-enhancement and data-skewness we repeat experiments for three levels of distribution bias - Balanced, Linear Imbalance and Long-Tailed distribution (see Figure 1).

- **Robustness Enhancement Techniques**: Building models that are satisfactorily robust as well as generalize well on all the inherent classes is essential to the deployment of high stakes real-world machine learning solutions, however, the complex interactions between robustness and dataset biases such as class-imbalance have largely been left neglected.
  Rawal and Pradhan (2020b) in their work observed that Augmix (Hendrycks et al., 2020b), a state-of-the-art robustness enhancement technique on common corruptions, tends to deteriorate the model's performance on imbalance focused-metrics, however, the results were limited and carried out for a different evaluation. We are inspired to explore further in this area, and conduct a comprehensive study, by evaluating numerous robustness-enhancement techniques for multiple models on two datasets at varied imbalance levels. Additionally, we also seek to evaluate the effects of robustness to universal adversarial perturbations on class-imbalance metrics, which to the best of our knowledge is still unexplored.
  1. **Common Corruption**: We will evaluate various state-of-the-art robustness to common corruption techniques. Firstly we have Augmix (Hendrycks et al., 2020b), a highly modular data augmentation technique that employs chains of layered transformations, stochastically sampling from a predefined pool of augmentations. The final augmented image consists of a weighted sum across various augmentation chains and the original image. Secondly, we have DeepAugment+Augmix (Hendrycks et al., 2020a), it uses image-to-image neural networks for data augmentations and can be easily integrated with other augmentation techniques (such as Augmix in our case). Lastly, we have Adversarial Noise Training (ANT) (Rusak et al., 2020) which trains a classifier jointly with a noise generator neural network. The goal of the noise generator is to produce spatially uncorrelated noise that when added to inputs, can fool a given classifier. The ANT scheme ensures the classifier becomes robust to adversarial noise as well as common corruption.
  2. **Adversarial Corruption**: We will evaluate on two robustness enhancement methods to Universal adversarial perturbations(UAPs). Firstly, we will apply Universal Adversarial Training as proposed by Shafahi et al. (2020) which increases adversarial robustness by modelling the training process as a two-player min-max game where the minimization is over the target model parameters, and the maximization is over the universal adversarial perturbation. Secondly, we will use the idea of 'shared adversarial training' as proposed by Mummadi, Brox, and Metzen (2019) to handle the trade-off between enhanced robustness against UAPs vs. reduced performance on clean data samples. We will study these methods with a focus on imbalance bias of clean datasets.
- **Training Routine**: Our training setup consists of Adam Optimizer and Cross-Entropy loss. Standard data augmentation e.g. Random Horizontal-Flips, Random Vertical-Flips, and Random Rotation are also applied after normalizing the data. Additionally, we employ Cyclical Learning Rates (Smith, 2015), a learning rate routine that oscillates between a range of values (contrary to the conventional wisdom of step-wise decreasing learning rate as training progresses) to streamlines hyperparameter selection. Cyclical Learning Rates also ensures quick traversals over saddle points and sharp minima.
- **Result**: In this section, we present an outline of possible results (see *Table 1*). For a particular dataset, there will be two tables, each reporting a unique imbalance-focused metric namely Balanced accuracy & F-1 score. These experiments are repeated over a variety of imbalance-levels (Balanced, Linear, heavy-tailed versions) to meticulously examine the interactions between robustness-enhancement and performance on varying degree of distribution bias.
  In reference to *Table 1*- Different models in our experiment would correspond to distinct Deep Neural Network architecture families for eg. ResNet-50 and SqueezeNet, specific insights can then be drawn regarding effect of various robustness enhancement techniques (eg. AugMix, ANT) across a varied degree of imbalance (Balanced, Linear, Long-Tails).

Table 1: CIFAR-100 results with different models (I & II) and robustness-methods (a & b)

Bal. Acc.	Level of distribution bias			***F1 Score***	Level of distribution bias		
	Balanced	(Linear) Imbalance	Long-Tailed		Balanced	(Linear) Imbalance	Long-Tailed
Model-I	-	-	-	Model-I	-	-	-
Model I a	-	-	-	Model I a	-	-	-
Model I b	-	-	-	Model I b	-	-	-
Model II	-	-	-	Model II	-	-	-
Model II a	-	-	-	Model II a	-	-	-
Model II b	-	-	-	Model II b	-	-	-

Acknowledgments

We would like to sincerely thank Tatsunori Hashimoto for insightful discussions and Narendra Ahuja for initial feedback on the project. PP also thanks Krikamol Muandet for his support.

References

Brutzkus, A., and Globerson, A. 2019. Why do larger models generalize better? a theoretical perspective via the xor problem. In *International Conference on Machine Learning (ICML)*.

Buda, M.; Maki, A.; and Mazurowski, M. A. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks* 106:249–259.

Buolamwini, J., and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Friedler, S. A., and Wilson, C., eds., *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, 77–91. New York, NY, USA: PMLR.

Byrd, J., and Lipton, Z. C. 2019. What is the effect of importance weighting in deep learning? In *International Conference on Machine Learning (ICML)*.

Cao, K.; Wei, C.; Gaidon, A.; Aréchiga, N.; and Ma, T. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Carlini, N., and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, 39–57. IEEE.

Chaubey, A.; Agrawal, N.; Barnwal, K.; Guliani, K. K.; and Mehta, P. 2020. Universal adversarial perturbations: A survey. *arXiv preprint arXiv:2005.08087*.

Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; and Belongie, S. J. 2019. Class-balanced loss based on effective number of samples. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9260–9269.

Dezfooli, S. M.; Fawzi, A.; Omar, F.; Frossard, P.; and Soatto, S. 2018. Robustness of classifiers to universal perturbations: A geometric perspective. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Fawzi, A.; Moosavi-Dezfooli, S.-M.; and Frossard, P. 2016. Robustness of classifiers: from adversarial to random noise. In *Advances in Neural Information Processing Systems*, 1632–1640.

Feldman, V. 2019. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing (STOC) 2020*.

Geirhos, R.; Temme, C. R. M.; Rauber, J.; Schütt, H. H.; Bethge, M.; and Wichmann, F. A. 2018. Generalisation in humans and deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Goodfellow, I.; Shlens, J.; and Szegedy, C. 2015. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*.

He, K.; Zhang, X.; Ren, S.; and Sun., J. 2016. Deep residual learning for image recognition. In *Proc.. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

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

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. 2020a. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ArXiv* abs/2006.16241.

Hendrycks, D.; Mu, N.; Cubuk, E. D.; Zoph, B.; Gilmer, J.; and Lakshminarayanan, B. 2020b. AugMix: A simple data processing method to improve robustness and uncertainty. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Hobbs, A., and Svetlichnaya, S. 2020. Satellite-based prediction of forage conditions for livestock in northern kenya. ICLR 2020 Workshop on Computer Vision for Agriculture (CV4A).

Hooker, S.; Moorosi, N.; Clark, G.; Bengio, S.; and Denton, E. L. 2020. Characterising bias in compressed models. *ArXiv* abs/2010.03058.

Huang, C.; Li, Y.; Loy, C. C.; and Tang, X. 2016. Learning deep representation for imbalanced classification. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 5375–5384.

Huang, C.; Li, Y.; Loy, C. C.; and Tang, X. 2019. Deep imbalanced learning for face recognition and attribute prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

Iandola, F. N.; Han, S.; Moskewicz, M. W.; Ashraf, K.; Dally, W. J.; and Keutzer, K. 2016. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. *arXiv preprint arXiv:1602.07360*.

Ilyas, A.; Santurkar, S.; Tsipras, D.; Engstrom, L.; Tran, B.; and Madry, A. 2019. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems (NeurIPS)*, 125–136.

Izmailov, P.; Podoprikhin, D.; Garipov, T.; Vetrov, D. P.; and Wilson, A. G. 2018. Averaging weights leads to wider optima and better generalization. In *Conference on Uncertainty in Artificial Intelligence (UAI)*.

Krizhevsky, A. 2009. Learning multiple layers of features from tiny images. Technical report, University of Toronto. TR-2009.

Lipton, Z. C., and Steinhardt, J. 2019. Troubling trends in machine learning scholarship. *Queue* 17(1):80:45–80:77.

Moosavi-Dezfooli, S.-M.; Fawzi, A.; Fawzi, O.; and Frossard, P. 2017. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1765–1773.

Mummadi, C. K.; Brox, T.; and Metzen, J. H. 2019. Defending against universal perturbations with shared adversarial training. In *Proceedings of the IEEE International Conference on Computer Vision*, 4928–4937.

Rawal, R., and Pradhan, P. 2020a. Climate adaptation: Reliably predicting from imbalanced satellite data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. arXiv 2004.12344.

Rawal, R., and Pradhan, P. 2020b. Generalizing across the (in)visible spectrum. The ICML Workshop on Extreme Classification (XC): Theory and Applications.

Rusak, E.; Schott, L.; Zimmermann, R.; Bitterwolf, J.; Bringmann, O.; Bethge, M.; and Brendel, W. 2020. A simple way to make neural networks robust against diverse image corruptions. In *European Conference on Computer Vision (ECCV)*.

Shafahi, A.; Najibi, M.; Xu, Z.; Dickerson, J. P.; Davis, L. S.; and Goldstein, T. 2020. Universal adversarial training. In *AAAI*, 5636–5643.

Smith, L. N. 2015. Cyclical learning rates for training neural networks. *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)* 464–472.

Taori, R.; Dave, A.; Shankar, V.; Carlini, N.; Recht, B.; and Schmidt, L. 2020. Measuring robustness to natural distribution shifts in image classification. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Vasiljevic, I.; Chakrabarti, A.; and Shakhnarovich, G. 2016. Examining the impact of blur on recognition by convolutional networks. *ArXiv* abs/1611.05760.

Wang, Y.-X.; Ramanan, D.; and Hebert, M. 2017. Learning to model the tail. In *Advances in Neural Information Processing Systems*, 7029–7039.

Yehudai, G., and Shamir, O. 2019. On the power and limitations of random features for understanding neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.