RobustAugMix: Joint Optimization of Natural and Adversarial Robustness

Josué Martínez-Martínez*
Department of Computer Science & Engineering
University of Connecticut
Storrs, CT
josue.martinez-martinez@uconn.edu

Olivia Brown
AI Technology Group
MIT Lincoln Laboratory
Lexington, MA
olivia.brown@ll.mit.edu

Abstract

Machine learning models often suffer performance degradation when faced with corrupted data. In this work, we explore a technique that combines a data augmentation strategy (AugMix) with adversarial training, in order to increase robustness to both natural and adversarial forms of data corruption.

1 Introduction

Traditionally, machine learning models are trained using Empirical Risk Minimization (ERM), where model parameters are optimized to minimize the average error on the training dataset [20]:

$$\min_{\theta} E_{(x,y)\sim D}[L(f_\theta(x), y)], \tag{1}$$

where D is the distribution over inputs x with corresponding labels y, L is a loss function (e.g., cross-entropy), and f_θ is the machine learning model parameterized by θ, such as a deep neural network. The foundation of ERM is the assumption that training data is sampled independently from the same data distribution as test (e.g., operational) data [1]. In reality, operational data often differs from training data, and this difference can degrade the model's performance. Differences between training and operational data may be due to natural or adversarial factors, such as changes in the data collection process [17, 10], changes in the environment or sensors [7], or adversarial examples [19, 2, 13].

When deploying machine learning models in safety- or mission-critical applications, it is imperative that these models are designed to be robust to various forms of data corruption. A common approach to build robustness is to train the model on augmented data, and various strategies exist for applying natural [6, 3, 24, 22, 9] and adversarial [5, 15] data augmentations during training. However, the research community largely has been exploring each type of robustness (e.g., natural vs. adversarial) in isolation. This leads to scenarios such as a model that is robust to adversarial examples, but not robust to common corruptions, and vice versa. Real systems necessitate models that are robust to a whole suite of possible data corruptions and shifts, thus, there is a need for methods to jointly optimize multiple dimensions of robustness.

In this paper, we propose a new approach to improve robustness against both common corruptions and adversarial examples. The approach involves combing the AugMix data augmentation technique from [9] with the adversarial training methodology from [15]. We call our combined approach RobustAugMix. We train models using our approach on the CIFAR-10 dataset [12], and demonstrate an improvement (over baselines considered) in both robustness to common corruptions on the CIFAR-10-C dataset [8] and robustness to adversarial examples.

*Work completed under summer research program at MIT Lincoln Laboratory

ML Safety Workshop, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

2 Related Work

Data Augmentation Data augmentation has become a standard approach to increase generalization performance, and is beginning to emerge as a useful strategy for enhancing robustness to common corruptions. For image classification tasks, random flips, rotations and crops are commonly used to increase the size and variation of the training set and hence increase performance [6]. More sophisticated techniques such as Cutout [3], which produces random occlusions, CutMix [22], which replaces parts of an image with another, and MixUp [24], which linearly interpolates between two images, have also shown improvements over standard data transformations.

To increase robustness to naturally-corrupted data, [9] presented an approach called AugMix, which utilizes a chain of simple augmentation operations in concert with a consistency loss. The augmentations are sampled stochastically and combined via a weighted sum to produce a wide diversity of augmented images. The developers enforce a consistent embedding by the classifier across diverse augmentations of the same input image through the use of Jensen-Shannon divergence as a consistency loss. The AugMix training objective is:

$$\min_{\theta} E_{(x,y)\sim D}\{L_{CE}(f_\theta(x), y) + \lambda * L_{JSD}[f_\theta(x), f_\theta(g_{Aug}(x)), f_\theta(g_{Aug}(x))]\}, \tag{2}$$

where, L_{CE} is the cross-entropy loss, L_{JSD} is the Jensen-Shannon divergence loss, $g_{Aug}(x)$ is the stochastic function for applying a chain of augmentations, and λ is a scalar to help balance the contributions of the two loss terms. For more details on each of these functions, refer to [9].

There have also been several proposed methods that build off of or outperform AugMix. For example, PixMix [11] is comprised of two main components: a set of structurally complex pictures ("Pix"), and a pipeline for augmenting clean training images with these pictures ("Mix"). In [4], NoisyMix combines data augmentations with stability training and noise injections to improve model robustness and in-domain accuracy. In our work, we fix AugMix as our baseline, and consider how it can be combined with adversarial training to increase robustness to adversarial examples, and reserve the consideration of these alternative data augmentations strategies for future work.

Adversarial Training Adversarial training was first introduced in [5] as a defense against adversarial examples [19]. Adversarial training can be thought of as a data augmentation strategy in which adversarial perturbations are applied to the training inputs. [15] formalized the adversarial training objective as a form of robust optimization:

$$\min_{\theta} E_{(x,y)\sim D}[\max_{||\delta||_p<\epsilon} L(f_\theta(x+\delta), y)], \tag{3}$$

where, δ is the adversarial perturbation for input x with true label y, $||\cdot||_p$ is an ℓ_p norm (e.g., $p = 2$ is the Euclidean norm), and ϵ is the constraint on the size of the perturbation given that norm. The inner maximization is typically approximated using projected gradient descent (PGD), which performs iterative updates using the gradient of the loss with respect to the perturbation δ, and projected onto the ϵ-ball.

There has been some inital work aimed at combining aspects of data augmentation with adversarial training. [16] attempts to tackle the robust overfitting problem by incorporating data augmentation into adversarial training. In [21], the authors develop an extension to AugMix that first randomly samples multiple augmentation operators, then learns an adversarial mixture of the selected operators. In both of these works, the authors are still primarily focused on a single robustness objective (e.g., improving adversarial robustness *or* robustness to corruptions). In our work, we aim to jointly optimize robustness to both natural and adversarial shifts in data.

3 RobustAugMix

On their own, AugMix is not robust to adversarial examples, and adversarial training will not guarantee robustness to non-adversarial perturbation types, such as common corruptions. In order to combine the benefits of AugMix and adversarial training, and to develop models that are robust to both natural and adversarial data corruptions, we propose to optimize the following objective:

$$\min_{\theta} E_{(x,y)\sim D}[L_{CE}(f_\theta(x), y) + \lambda * L_{JSD}[f_\theta(x), f_\theta(g_{Aug}(x)), f_\theta(x + \delta^*)]] \tag{4}$$

$$\text{where } \delta^* = \arg \max_{||\delta||_p < \epsilon} L_{CE}(f_\theta(x + \delta), y), \tag{5}$$

where the third input to the Jensen-Shannon loss term from AugMix is replaced with an adversarial example. By minimizing this loss, our hope is to enable the model to learn to perform well on clean data (due to cross-entropy term), and learn to produce outputs for images that have been corrupted by both natural and adversarial data augmentations that are close to the outputs for the corresponding clean images.

We compare training a model via the RobustAugMix approach in Equation 4 with various baselines, including standard training via Equation 1, vanilla AugMix via Equation 2, and robust (i.e., adversarial) training via Equation 3.

4 Experiments

Dataset The CIFAR-10 [12] dataset contains small 32 × 32 × 3 color natural images, with 50,000 training images and 10,000 testing images, and has 10 classification categories. In order to measure a model's resilience to natural data shift, we evaluate our models on the CIFAR-10-C [8]. This dataset is constructed by corrupting the original CIFAR-10 test set. There are a total of 19 corruption types, including noise, blur, weather, and digital corruptions, each appearing at 5 severity levels or intensities. Since the CIFAR-10-C corruptions are used to measure network behavior under data shift, the 19 corruptions are reserved for testing, and not introduced during the training procedure.

Training Setup The neural network architecture used for this study was a 50-2 Wide ResNet [23] and was trained for 100 epochs with batch size of 128. It was optimized with stochastic gradient descent using Nesterov momentum. The networks were trained using an initial learning rate of 0.1 which decays following a cosine learning rate [14]. The input images were pre-processed with standard random left-right flipping and cropping prior to any additional augmentations, and normalized with the mean of the standard deviation of the CIFAR-10 dataset. For AugMix, we used the same augmentation scheme as presented in [9], with 3 augmentation chains and $\lambda = 12$. The adversarial examples for training the robust models were solved for using PGD with 7 steps of gradient descent, a step size of $2.5 * \epsilon/7$, and $\epsilon = 1.0$ consrained by the Euclidean norm (e.g., $p = 2$).

5 Results

Figure 1 presents a sample of the results of each model when tested against a corruption from four of the different corruption categories. The full results over all 19 corruptions are presented in Appendix A. As presented in the plots, the proposed RobustAugMix approach is more robust compared to the baselines when tested against Gaussian noise, glass blur, and JPEG compression. However, it does not perform as well against Fog. One hypothesis is that adversarial training enhances robustness to high-frequency corruptions (e.g., noise) at the cost of reduced robustness to lower-frequency corruptions (e.g., Fog). While RobustAugMix does not always outperform vanilla AugMix on CIFAR10-C, RobustAugMix improves over standard training in 15 out of 19 of the common corruptions, particularly when tested against corruptions of higher severity levels. RobustAugMix also outperforms robust (i.e., adversarial) training for every corruption type.

Figure 2 presents the adversarial accuracy obtained by each model when tested against adversarial perturbations of increasing size (i.e., epsilon). We generate adversarial examples for testing using the CIFAR-10 test set, using 10 steps of PGD, a step size of $2.5 * \epsilon/7$, and ϵ constrained by the Euclidean norm. Note that RobustAugMix achieves very similar adversarial accuracy to Robust training, and significantly improves over Standard and AugMix. RobustAugMix, thus, achieves our goal of enhancing robustness to naturally corrupted images, while also remaining robust to adversarial attacks.

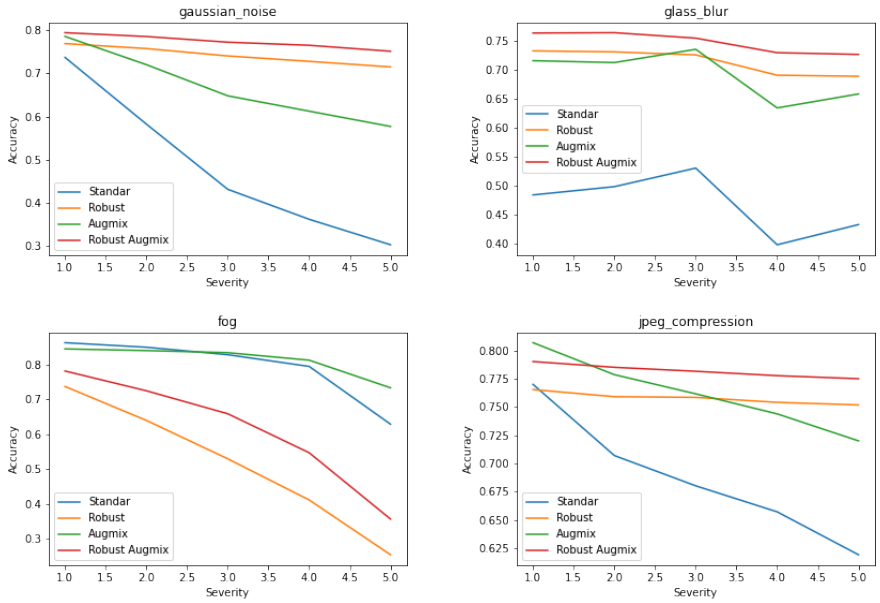

Figure 1: Accuracy of models trained using a standard approach, robust approach (i.e., adversarial training), AugMix, and RobustAugMix, when tested against 4 types of common corruptions (gaussian noise, glass blur, fog, and jpeg compression) of increasing severity.

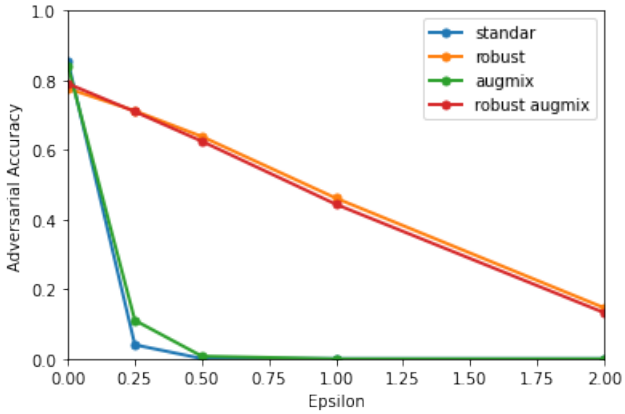

Figure 2: Accuracy of models trained using a standard approach, robust (i.e., adversarial) training, AugMix, and RobustAugMix, when tested against the CIFAR10 test set perturbed using ℓ_2 adversarial perturbations of increasing size (ϵ).

6 Conclusions and Future Work

This paper presents a new approach to achieve robustness against corrupted images and adversarial attacks. The AugMix method brings robustness against low-frequency domain corruptions and the robust training enhances robustness against high-frequency domain corruptions. Some potential limitations of RobustAugMix include a drop in clean accuracy compared to Standard and AugMix training, and an increased computational cost needed to train with adversarial perturbations.

Additional experiments are needed to better characterize the performance of RobustAugMix. Future work will include comparing to other baselines (e.g., an ensemble of models trained with AugMix and adversarial training, independently) and data augmentation techniques (e.g., AugMax [21]), adding additional augmentations into the AugMix strategy (e.g., Fourier perturbations [18]), and repeating these experiments on other datasets.

Acknowledgments and Disclosure of Funding

DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited.

This material is based upon work supported by the Under Secretary of Defense for Research and Engineering under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Under Secretary of Defense for Research and Engineering.

© 2022 Massachusetts Institute of Technology.

Delivered to the U.S. Government with Unlimited Rights, as defined in DFARS Part 252.227-7013 or 7014 (Feb 2014). Notwithstanding any copyright notice, U.S. Government rights in this work are defined by DFARS 252.227-7013 or DFARS 252.227-7014 as detailed above. Use of this work other than as specifically authorized by the U.S. Government may violate any copyrights that exist in this work.

A portion of this research was sponsored by the United States Air Force Research Laboratory and the United States Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the United States Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

[1] Dan A Calian, Florian Stimberg, Olivia Wiles, Sylvestre-Alvise Rebuffi, Andras Gyorgy, Timothy Mann, and Sven Gowal. Defending against image corruptions through adversarial augmentations. *arXiv preprint arXiv:2104.01086*, 2021.

[2] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 3–14, 2017.

[3] Terrance DeVries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout, 2017.

[4] N. Benjamin Erichson, Soon Hoe Lim, Winnie Xu, Francisco Utrera, Ziang Cao, and Michael W. Mahoney. Noisymix: Boosting model robustness to common corruptions, 2022.

[5] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[7] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

[8] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations, 2019.

[9] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019.

[10] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021.

[11] Dan Hendrycks, Andy Zou, Mantas Mazeika, Leonard Tang, Bo Li, Dawn Song, and Jacob Steinhardt. Pixmix: Dreamlike pictures comprehensively improve safety measures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16783–16792, June 2022.

[12] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[13] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018.

[14] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[15] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[16] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy A Mann. Data augmentation can improve robustness. *Advances in Neural Information Processing Systems*, 34:29935–29948, 2021.

[17] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019.

[18] Ryan Soklaski, Michael Yee, and Theodoros Tsiligkaridis. Fourier-based augmentations for improved robustness and uncertainty calibration. *arXiv preprint arXiv:2202.12412*, 2022.

[19] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[20] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

[21] Haotao Wang, Chaowei Xiao, Jean Kossaifi, Zhiding Yu, Anima Anandkumar, and Zhangyang Wang. Augmax: Adversarial composition of random augmentations for robust training. *Advances in neural information processing systems*, 34:237–250, 2021.

[22] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.

[23] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[24] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization, 2017.

A Appendix

We present results of our experiments on the full CIFAR10-C dataset, broken out by corruption type and severity level in Figures 3-6. Noise corruptions are shown in Figure 3, blur corruptions in Figure 4, weather corruptions in Figure 5, and digital corruptions in Figure 6.

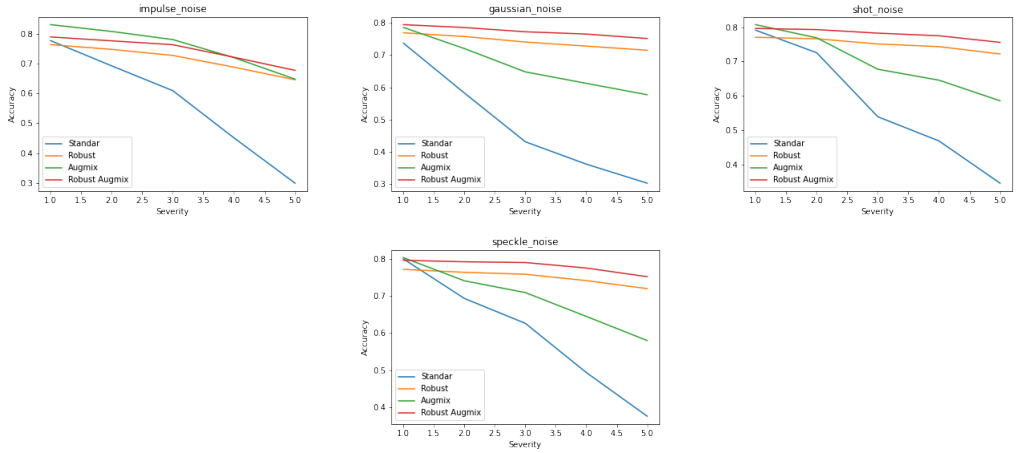

Figure 3: Noise Corruptions

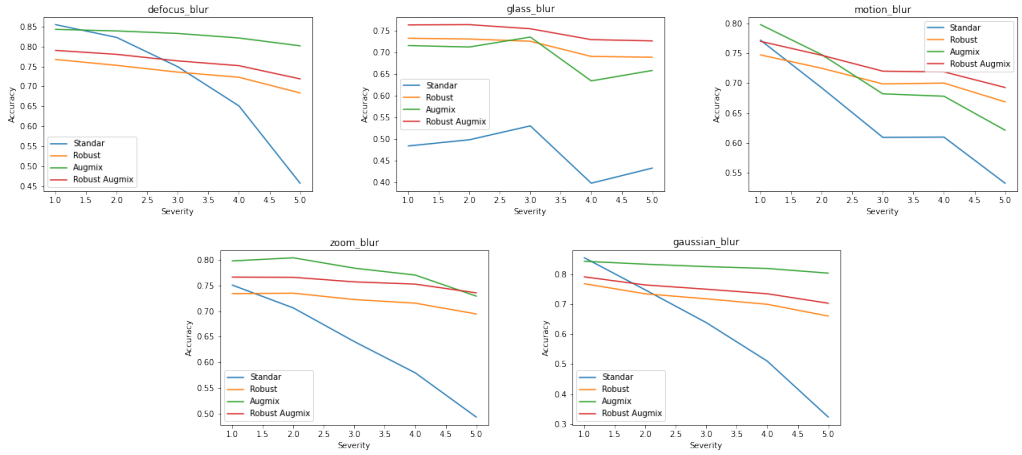

Figure 4: Blur Corruptions

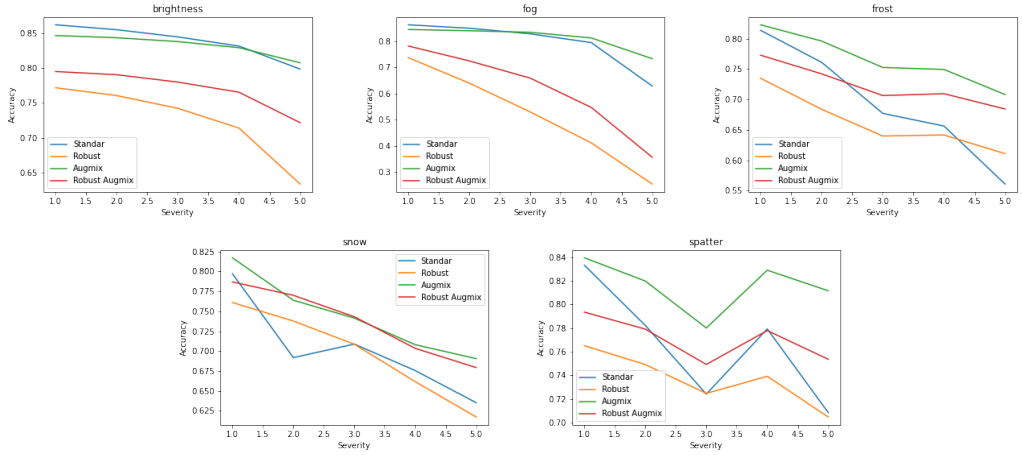

Figure 5: Weather Corruptions

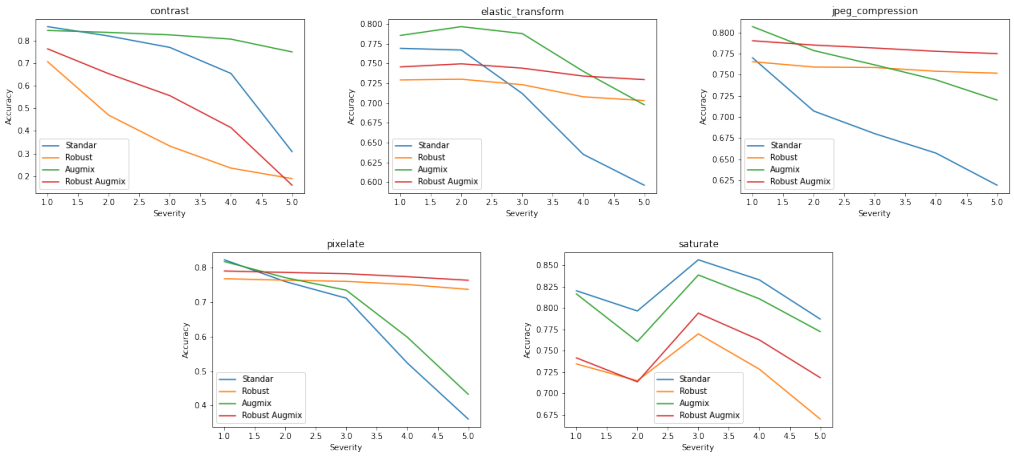

Figure 6: Digital Corruptions