NARROWING CLASS-WISE ROBUSTNESS GAPS IN ADVERSARIAL TRAINING

Fatemeh Amerehi, Patrick Healy
Computer Science and Information Systems
University of Limerick
Limerick, Ireland
Fatemeh.Amerehi@ul.ie

ABSTRACT

Efforts to address declining accuracy as a result of data shifts often involve various data-augmentation strategies. Adversarial training is one such method, designed to improve robustness to worst-case distribution shifts caused by adversarial examples. While this method can improve robustness, it may also hinder generalization to clean examples and exacerbate performance imbalances across different classes. This paper explores the impact of adversarial training on both overall and class-specific performance, as well as its spill-over effects. We observe that enhanced labeling during training boosts adversarial robustness by 53.50% and mitigates class imbalances by 5.73%, leading to improved accuracy in both clean and adversarial settings compared to standard adversarial training.

1 INTRODUCTION

Adversarial examples are inputs to machine learning models that have been intentionally crafted to cause incorrect model predictions. These examples are typically made by introducing slight perturbations that distort model outputs while maintaining a high degree of similarity Goodfellow et al. (2014); Papernot et al. (2016); Tramèr et al. (2017); Madry et al. (2017).

Formally, given an image x_i with its corresponding label y_i, and a model f that correctly classifies it, i.e., $f(x_i) = y_i$, an adversarial example x_i' is defined as an image that satisfies two conditions: first, it causes the classifier to misclassify the image, such that $f(x_i) \neq f(x_i')$; and second, it remains visually similar to the original image x_i Carlini et al. (2019); Engstrom et al. (2019); Wang et al. (2019). Typically, similarity between the two images is measured using an ℓ_p-norm, meaning that $x_i' = x_i + \delta$ is considered a valid adversarial example if and only if $\|\delta\|_p \leq \varepsilon$, where ε is a small constant, and $p \in [0, \infty]$. Under these similarity constraints, adversarial examples x_i' are often crafted as in Eq. 1, to maximize the loss of the model when processing the sample x_i.

$$\max_{\|\delta\|_p \leq \varepsilon} \mathcal{L}(f(x_i + \delta), y_i). \tag{1}$$

The set of adversarial data points generated by this maximization implicitly defines a distribution of adversarial examples Goodfellow (2019). The Fast Gradient Sign Method (FGSM) Goodfellow et al. (2014) approximates the above maximization, and generates adversarial examples by back-propagating the gradient of the loss to the input data to compute $\nabla_{x_i}\mathcal{L}(x_i, y_i)$. This gives direction in which the loss function increases the most with respect to small changes in the input data x_i. It then moves the data in this direction (i.e., $\text{sign}\nabla_{x_i}\mathcal{L}(x_i, y_i)$) that maximizes the loss for x_i, as shown in Eq. 2:

$$x_i' = x_i + \varepsilon \cdot \text{sign}(\nabla_{x_i}\mathcal{L}(x_i, y_i)), \tag{2}$$

where the added perturbation is scaled down by ε to maintain similarity distance. FGSM Goodfellow et al. (2014) is designed for fast generation rather than optimality. Basic Iterative Method Kurakin et al. (2016) and Projected Gradient Descent (PGD) Madry et al. (2017) are iterative extensions of

FGSM, designed to generate stronger adversarial examples. These examples typically viewed as the worst-case form of distributional shift, where even minor perturbations can lead to significant misclassifications Rice et al. (2021).

To address distribution shifts, data augmentation techniques are commonly used to improve model performance in image classification by increasing data variety and reducing the gap between training and test data distributions Hendrycks & Dietterich (2019); Xu et al. (2023). Likewise, to enhance model robustness against attacks, adversarial training is the most effective method, involving the augmentation of training data with adversarial examples Goodfellow et al. (2014); Kurakin et al. (2016); Moosavi-Dezfooli et al. (2016); Madry et al. (2018); Athalye et al. (2018).

While data augmentation enhances overall accuracy and robustness, its effects are often highly class-dependent. Techniques like random cropping, for instance, can introduce class imbalance—improving average test performance while significantly degrading accuracy for certain classes Kirichenko et al. (2024). Augmentations may also produce unintended spillover effects; for example, color jittering strengthens robustness to brightness and color shifts yet unexpectedly weakens robustness to pose Idrissi et al. (2022). Similarly, while adversarial training enhances robustness, it comes at a cost. It imposes trade-offs between robustness and accuracy Tsipras et al. (2018), as well as between in- and out-of-distribution generalization Zhang et al. (2019), while also amplifying disparities in performance across different classes. As a result, certain classes may be disproportionately disadvantaged, affecting model fairness Benz et al. (2021).

A data augmentation policy that fails to preserve label integrity can further disrupt class balance. When applied uniformly across all classes, augmentations may degrade label information unevenly, leading to imbalances even in originally well-balanced datasets Kim et al. (2020); Balestriero et al. (2022); Islam et al. (2024). In this paper, we aim to mitigate the class imbalance caused by adversarial training by adjusting the labels used during the adversarial training process.

2 Related works

Adversarial training is a form of data augmentation that incorporates adversarial examples into the training pipeline. Typically, augmentation techniques expand training data by randomly applying transformations to promote invariance, thus encouraging models to make consistent predictions across different views of each sample Geiping et al. (2023). In general, augmentation techniques can be classified into two categories: label-preserving and label-mixing techniques.

Label-preserving augmentation uses transformations on images that preserve their semantic content. While these methods have shown improvements in generalization for some factors, they can also negatively impact others Idrissi et al. (2022). Label-mixing approaches use convex combinations of pairs of examples and their labels to encourage linear behavior between training examples, which helps to regularize the model Zhang et al. (2018). Despite their effectiveness, these methods have been found to introduce label ambiguities due to random placements of images, resulting in misleading signals for supervision Kim et al. (2020); Islam et al. (2024). Similar to both types, Label Augmentation Amerehi & Healy (2024) aims to maintain invariance to the class identity of images while also encouraging separation between class identity and transformation. It does so by assigning one-hot labels z_j to each transformation operation o_j applied during augmentations. Rather than merely augmenting the transformed data $\tilde{x_i}$ during training, without distinguishing between labels for transformed and untransformed inputs, it augments the labels by concatenating the original input labels with the operation labels:

$$\tilde{y}_i = \text{Concat}[(1-\delta)y_i, \delta z_j], \tag{3}$$

where, where δ is a scaling factor that prevents excessive deviation of the model toward the augmented label. The training objective is thus expressed as:

$$\mathcal{L}_{LA}(\tilde{y}_i \tilde{p}_i) = -\sum_{k=1}^{K+M} \tilde{y}_{ik} \log \tilde{p}_{ik}, \tag{4}$$

Training	Clean	mCE	PGD-40
Std	**20.87**	**58.06**	94.06
Adv	32.18	71.41	70.32
Adv$^+$	26.70	66.43	**32.70**

Table 1: Error Rate of Adversarial Training With/Without Label Augmentation on ResNet-50

where, $\tilde{p}_i$ denotes the softmax of predictions for $\tilde{x}_i$, and $\tilde{y}_{ik}$ is a vector of length $K + M$, which merges the original K class labels and the M transformation labels. This method has been shown to improve both the clean and robust accuracy. In the following sections, we examine whether Label Augmentation can be incorporated into adversarial training to reduce its negative side effects.

3 Experimental Setup

We study how average and class-level performance change when concatenating labels during 10-step PGD Madry et al. (2017) adversarial training, with ℓ_∞ constraints and constraint budget $\varepsilon = 0.03$. Our evaluation focuses on robustness against common and adversarial perturbations, as well as the induced side effects of adversarial training.

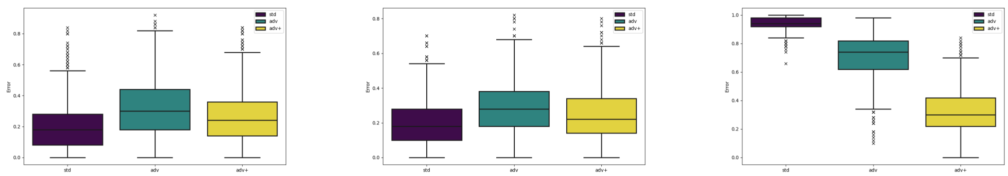
Architecture and Training Details. We run all experiments on an RTX-3080 GPU with CUDA Version 12.5 using PyTorch version 2.0.1. We fine-tune the ResNet50 model with default weights on the ImageNet (IN) for 10 epochs. The training starts with a learning rate of 0.01, which decays by a factor of 0.0001 according to a cosine annealing learning rate schedule (Loshchilov & Hutter, 2016). We optimize the models using stochastic gradient descent with a momentum of 0.9. The batch sizes for training and evaluation are set to 64. We set the scaling factor $\delta = 0.03$ to reflect the strength of the added perturbation. We evaluate both average error and per-class error using the original ImageNet and the corresponding robustness benchmark datasets: IN-C Hendrycks & Dietterich (2019), IN-X Idrissi et al. (2022), and IN-ReaL Beyer et al. (2020). Additionally, we assess their adversarial robustness using a 40-step PGD attack, with ℓ_∞ constraints and constraint budget $\varepsilon = 0.03$.

Evaluation metrics. The Clean error denotes the standard classification error on uncorrupted test data. For a given corruption c within IN-C, the error at severity s is denoted as $\mathrm{E}_{c,s}$. The Corruption Error (CE_c) is the average error over severities: $\mathrm{CE}_c = \frac{1}{5}\sum_{s=1}^{5} \mathrm{E}_{c,s}$. The mean Corruption Error (mCE) is then averaged across all 15 corruptions: $\mathrm{mCE} = \frac{1}{15}\sum_{c=1}^{15} \mathrm{CE}_c$. This single value enables comparisons against common corruptions Hendrycks & Dietterich (2019). To account for label noise, we evaluate on IN-ReaL Beyer et al. (2020), which provides re-assessed multi-label annotations for the ImageNet validation set. These metrics quantify a model's error in assigning incorrect labels. To explore spillover effects, we evaluate on IN-X Idrissi et al. (2022), which includes human annotations of failure modes across 16 variation factors, such as pose, size, color, and occlusions. We compute standard classification error as well as error ratio for each factor as $\mathrm{E}_f = \frac{1-\mathrm{accuracy(factor)}}{1-\mathrm{accuracy(model)}}$, which measures how much a model's errors increase for a specific variation factor relative to its overall performance Idrissi et al. (2022).

4 Results

Average Error. Table 1 presents error rates under three evaluation conditions. Compared to the standard model, adversarial training (Adv) improves adversarial robustness by 25.24% but compromises clean error and corruption robustness by 54.19% and 22.99%, respectively. Adv$^+$ further enhances adversarial robustness by 65.23% while weakening clean error by 27.93% and corruption robustness by 14.42%. Both methods involve trade-offs, but Adv$^+$ achieves a better balance, mitigating the losses in clean error by 17.03% and corruption robustness by 6.97% compared to Adv, while achieving a 53.50% more improvement in adversarial robustness.

Class-Wise Error. Figure 1 shows the class-wise error comparessions. Both Clean ImageNet 1a and ImageNet-ReaL 1b show that, compared to the standard model, adversarial methods increase



(a) Clean ImageNet. (b) ImageNet-ReaL. (c) PGD-40 on ImageNet.

Figure 1: Class-wise errors on different settings.

	Clean		ReaL		PGD-40	
Method	Mean	SD	Mean	SD	Mean	SD
std	**20.87**	**15.65**	**19.43**	**12.78**	94.06	**4.71**
adv	32.17	17.78	29.00	15.24	70.32	15.77
adv+	26.69	16.76	24.27	14.04	**32.70**	15.33

Table 2: Mean and Standard Deviation (SD) across Different Methods

the existing class-wise imbalance. However, Adv^{+} outperforms Adv by reducing both mean error and class-wise error variability—by 5.73% on Clean and 7.87% on IN-ReaL settings. Under a PGD-40 attack 1c, the standard model exhibits the highest error and low variability, suggesting uniform vulnerability. In contrast, adversarial models exhibit similar class-wise imbalances, with Adv^{+} reducing imbalance by 2.79% compared to Adv while also achieving greater adversarial robustness.

Error Rates Across Corruptions and Categories. Figure 2 presents error rates on ImageNet-C Hendrycks & Dietterich (2019) across various corruption types and severity levels. In most cases, the standard model performs better across different corruption types and severity levels. For adversarial models, Adv^{+} consistently outperforms Adv, except in JPEG and Pixelate corruptions. Figures 3 and 4 show the error rates and error ratios across different IN-X categories. While the standard model achieves lower overall error, all models exhibit similar types of mistakes. However, adversarial models demonstrate improved error ratios in style and texture variations.

5 CONCLUSION

While adversarial training helps mitigate distribution shifts from adversarial examples, it often results in reduced performance on clean samples and increased class-wise error disparities. Modifying labels during adversarial training is easy to implement, enhancing overall robustness while achieving a more favorable trade-off compared to standard adversarial training.

ACKNOWLEDGMENTS

This publication has emanated from research conducted with the financial support of Taighde Éireann – Research Ireland under Grant No. 18/CRT/6223.

REFERENCES

Fatemeh Amerehi and Patrick Healy. Label augmentation for neural networks robustness. In *Conference on Lifelong Learning Agents*, 2024. URL `arXivpreprintarXiv:2408.01977`.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pp. 274–283. PMLR, 2018.

Randall Balestriero, Leon Bottou, and Yann LeCun. The effects of regularization and data augmentation are class dependent. In S. Koyejo, S. Mohamed, A. Agarwal,

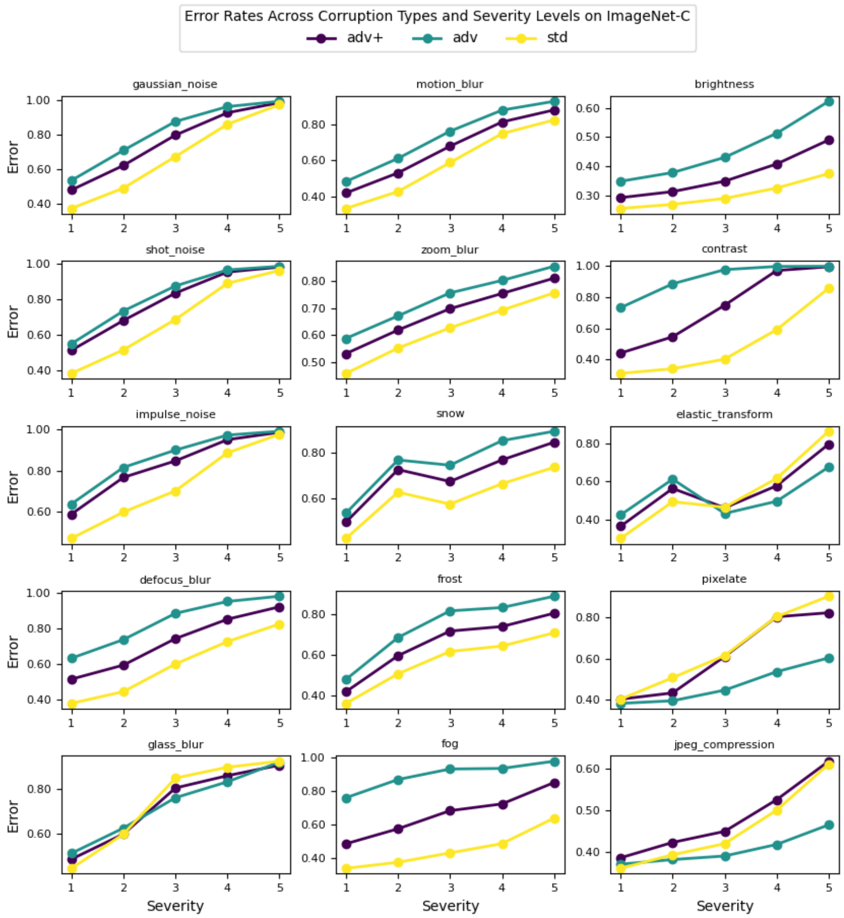


Figure 2: Error rates across corruption types and severity levels on ImageNet-C.

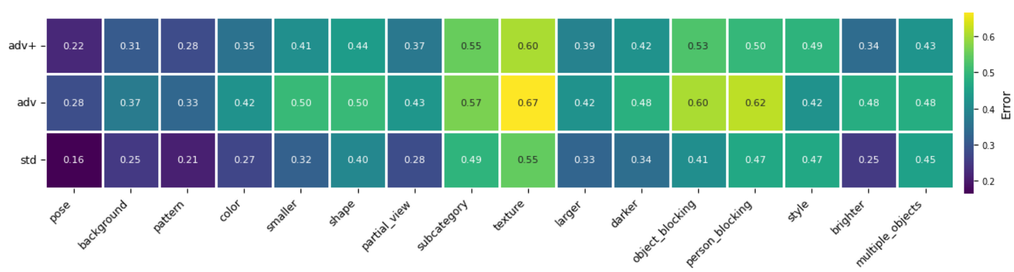


Figure 3: Error rate across ImageNet-X categories.

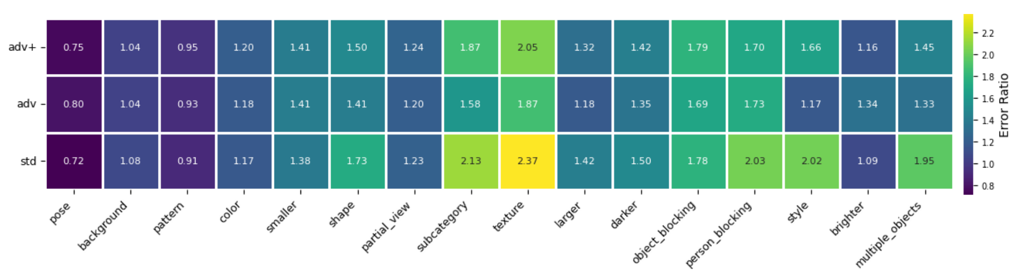


Figure 4: Error ratio across ImageNet-X categories.

D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 37878–37891. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/f73c04538a5e1cad40ba5586b4b517d3-Paper-Conference.pdf.

Philipp Benz, Chaoning Zhang, Adil Karjauv, and In So Kweon. Robustness may be at odds with fairness: An empirical study on class-wise accuracy. In *NeurIPS 2020 Workshop on pre-registration in machine learning*, pp. 325–342. PMLR, 2021.

Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020.

Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.

Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. In *International conference on machine learning*, pp. 1802–1811. PMLR, 2019.

Jonas Geiping, Micah Goldblum, Gowthami Somepalli, Ravid Shwartz-Ziv, Tom Goldstein, and Andrew Gordon Wilson. How much data are augmentations worth? an investigation into scaling laws, invariance, and implicit regularization. In *International Conference on Learning Representations (ICLR)*, March 2023. URL https://arxiv.org/pdf/2210.06441.

Ian Goodfellow. A research agenda: Dynamic models to defend against correlated attacks. *arXiv preprint arXiv:1903.06293*, 2019.

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

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Badr Youbi Idrissi, Diane Bouchacourt, Randall Balestriero, Ivan Evtimov, Caner Hazirbas, Nicolas Ballas, Pascal Vincent, Michal Drozdzal, David Lopez-Paz, and Mark Ibrahim. Imagenet-x: Understanding model mistakes with factor of variation annotations. *arXiv preprint arXiv:2211.01866*, 2022.

Khawar Islam, Muhammad Zaigham Zaheer, Arif Mahmood, and Karthik Nandakumar. Diffusemix: Label-preserving data augmentation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27621–27630, 2024.

Jang-Hyun Kim, Wonho Choo, and Hyun Oh Song. Puzzle mix: Exploiting saliency and local statistics for optimal mixup. In *International conference on machine learning*, pp. 5275–5285. PMLR, 2020.

Polina Kirichenko, Mark Ibrahim, Randall Balestriero, Diane Bouchacourt, Shanmukha Ramakrishna Vedantam, Hamed Firooz, and Andrew G Wilson. Understanding the detrimental class-level effects of data augmentation. *Advances in Neural Information Processing Systems*, 36, 2024.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

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

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.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2574–2582, 2016.

Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.

Leslie Rice, Anna Bair, Huan Zhang, and J Zico Kolter. Robustness between the worst and average case. *Advances in Neural Information Processing Systems*, 34:27840–27851, 2021.

Florian Tramèr, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. The space of transferable adversarial examples. *arXiv preprint arXiv:1704.03453*, 2017.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International conference on learning representations*, 2019.

Mingle Xu, Sook Yoon, Alvaro Fuentes, and Dong Sun Park. A comprehensive survey of image augmentation techniques for deep learning. *pattern recognition*, 137:109347, 2023.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pp. 7472–7482. PMLR, 2019.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1Ddp1-Rb.