Towards Class-wise Robustness Analysis

Tejaswini Medi[1], Julia Grabinski[1,2], and Margret Keuper[1,3]

[1] Visual Computing, University of Siegen
[2] Fraunhofer ITWM, Kaiserslautern and IMLA, University of Offenburg
[3] Max Planck Institute for Informatics, Saarland Informatics Campus

Abstract

While being very successful in solving many downstream tasks, the application of deep neural networks is limited in real-life scenarios because of their susceptibility to domain shifts such as common corruptions, and adversarial attacks. The existence of adversarial examples and data corruption significantly reduces the performance of deep classification models. Researchers have made strides in developing robust neural architectures to bolster decisions of deep classifiers. However, most of these works rely on effective adversarial training methods, and predominantly focus on overall model robustness, disregarding class-wise differences in robustness, which are critical. Exploiting weakly robust classes is a potential avenue for attackers to fool the image recognition models. Therefore, this study investigates class-to-class biases across adversarially trained robust classification models to understand their latent space structures and analyze their strong and weak class-wise properties. We further assess the robustness of classes against common corruptions and adversarial attacks, recognizing that class vulnerability extends beyond the number of correct classifications for a specific class. We find that the number of false positives of classes as specific target classes significantly impacts their vulnerability to attacks. Through our analysis on the Class False Positive Score, we assess a fair evaluation of how susceptible each class is to misclassification.

1. Introduction

Convolutional neural networks (CNNs) have achieved widespread success in various vision applications including image classification [8, 13, 27], image segmentation[16], and object detection [21]. Nonetheless, the existence of adversarial examples [6, 15, 5] and common corruptions [10] like blurring, zooming, or Gaussian noise, poses challenges in their real-world deployment. Extensive efforts have been devoted to defending against these adversarial attacks and enhancing model generalization [26]. Adversarial training has emerged as a prominent defense technique to improve the robustness of classification models [1, 6]. Prior works have analyzed adversarial training from different perspectives including robust optimization [23], robust generalization [20], and training strategy [28, 19, 24]. Howbeit, all these previous works have concentrated on improving the overall model robustness, neglecting the discrepancies in the robustness of individual classes. This imbalance in class-wise robustness can be exploited by attackers, who may target less robust classes. Therefore, a comprehensive understanding of adversarial training on class-wise robustness is crucial for improving the robustness of classification models in a meaningful way.

Recently a few studies have emphasized class-wise robustness disparity in adversarial training [22, 2]. However, their focus has been limited to comparing class-wise robust accuracy deviations to identify the vulnerable classes. While this is important, it is equally crucial to analyze class-to-class biases to gain insights into the latent space of robust models. Specifically, understanding which class labels are assigned erroneously or which classes are predominantly confused is essential. Therefore, we conduct a comprehensive study of class-wise robust accuracies with particular emphasis on false positives in class-wise misclassifications, to improve the understanding of class-wise biases.

2. Background: Network Evaluation Methods

When introducing new models for image classification tasks, one usually considers the network's performance in terms of accuracy[8, 27, 13]. Such evaluation is most important, as we do not need a network that classifies randomly and wrongly. Additionally, the evaluation of the network's robustness received more popularity in the last few years. Hence, a variety of robustness measures has been proposed [12]. On the one hand, common corruptions [10]

have been introduced, which incorporate natural and system noise that can lead to misclassifications in the classification systems. On the other hand, adversarial attacks gained a lot of popularity to evaluate the network's vulnerabilities. In consequence, a variety of attacks, *e.g.* [6, 15, 5] along with their defenses have been proposed [6]. The *robust accuracy* of a model is thereby usually defined as the model's accuracy under a specific adversarial attack or corruption. Thus, most of these studies focus on improving the overall robust accuracy of models under attacks or when facing corruptions. A few recent works further investigated the class biases in model accuracy and robustness, arguing for a fair training process that allows classifying all classes about equally well [25, 22]. These works also showed that adversarial training seems to amplify class-wise biases in model accuracy. Yet, only little effort has been devoted to studying which classes pre-dominantly attract incorrectly classified samples. To identify such classes, we study in this work the Class False Positive Score. We further argue that this perspective provides interesting insights into the model behavior and potentially allows to improve our understanding of the model's latent space and vulnerability.

2.1. Evaluatiuon Metrics

In this study, we calculate the **Class False Positive Score (CFPS)** to assess the vulnerability of each class c_j towards misclassifications with $j \in \{1, \dots, C\}$ in the classification model. To calculate the CFPS for a specific class, we calculate the number of misclassifications where samples from other classes, *i.e.* samples x_i from the test set $\{x_i\}_{i=1}^{N}$ of size N with labels $y_i \in \{c_j\}_{j=1}^{C}$ are incorrectly classified by model f_θ as this particular class, *i.e.* the cardinality of $\{x_i | f_\theta(x_i) = c_j, y_i \neq c_j\}$. We then divide this count by the total number of misclassifications across all classes,

$$\mathrm{CFPS}(c_j) = \frac{|\{x_i | f_\theta(x_i) = c_j, y_i \neq c_j\}|}{|\{x_i | f_\theta(x_i) \neq y_i\}|}. \quad (1)$$

A higher CFPS for a class indicates that it is more susceptible to being mistakenly assigned to samples from other classes by the model. The classes that are most likely mistaken as other classes have a high chance of manipulation by attackers, which impacts the overall reliability and security of our classification model. This enables us to focus on improving the robustness of these vulnerable classes.

The CFPS is complementary to the class-wise accuracy (CWA), which has been predominantly used in previous works such as [22, 2], which is defined as

$$\mathrm{CWA}(c_j) = \begin{aligned} &|\{x_i | f_\theta(x_i) = c_j, y_i = c_j\}|/N \\ &+|\{x_i | f_\theta(x_i) \neq c_j, y_i \neq c_j\}|/N. \end{aligned} \quad (2)$$

When evaluated under attack or corruption, we refer to these metrics as *robust accuracy* and *robust CFPS*, respectively.

3. Experiments

Class-Wise Accuracy Analysis. To carry out our experiments, we utilize the CIFAR-10 dataset, a simple and widely used benchmark for image classification tasks [14]. For robustness evaluations, we employ adversarially trained robust models from a standardized adversarial robustness benchmark [7, 4]. For our analysis, we select standard classification models like ResNet-18 and ResNet-50 from the ResNet family [8], DenseNet-169 [13], PreActResNet-18[9], WideResNet-70-16 [27] and a recent foundation model, DINOv2[18]. The ten classes of CIFAR-10, namely 'airplane', 'automobile', 'bird', 'cat', 'deer', 'dog', 'frog', 'horse', 'ship', and 'truck', are denoted as C1 to C10 respectively in the following sections of the work.

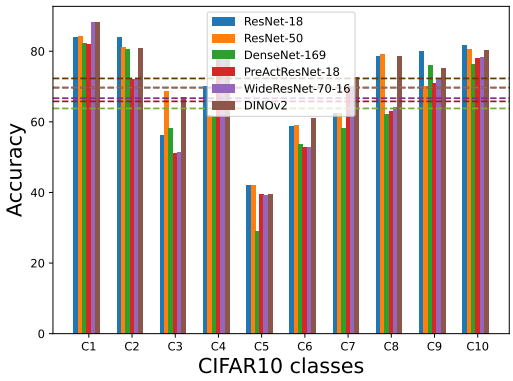

Figure 1: Class-wise accuracies of CIFAR10 across different robust model architectures. The horizontal lines in the figure depict the average overall accuracy of respective adversarially trained robust models.

Figure 1 illustrates the class-wise accuracy of the aforementioned architectures when evaluated on clean validation samples from the CIFAR10 dataset. The overall accuracy of each model allows us to categorize classes into two groups: strong classes and weak classes. This categorization is based on whether a class exhibits a class-wise accuracy above or below the average overall accuracy of the model. Notably, we observe that classes C3, C5, and C6, corresponding to 'bird', 'deer', and 'dog' respectively, fall into the category of weak classes. This determination is made due to their relatively lower accuracy compared to the other classes across different robust models. It is important to emphasize that this pattern remains consistent regardless of the specific architecture employed for training the robust models. Nevertheless, it is also not possible to conclude from this type of class-wise accuracy evaluation that these weak classes are more susceptible to attacks than strong classes when targeting the attack on specific classes.

In the context of class-wise robust analysis, previous research has commonly identified the weak robust classes

[22, 2]. These determinations were often made by calculating the class-wise accuracy deviations with respect to the overall model accuracy or strong class accuracy but they failed to see a common pattern of weak classes under the influence of common corruptions and attacks. However, such approaches may introduce potential biases, as the weakness of a class could be influenced by the overall model accuracy or the highest strong class accuracy. To ensure fairness and impartiality in our evaluation, we evaluate a metric called the Class False Positive Score, shortly CFPS. This metric focuses on model misclassifications among the classes independently of overall accuracy, enabling a comprehensive analysis of class-to-class biases exhibited by the models.

Class False Positive Score. Our evaluation of the CFPS for all classes of CIFAR-10 across different neural architectures is presented in figure 2. The results clearly demonstrate that the CFPS for classes C1, and C4 are comparatively higher, indicating these classes are highly susceptible to misclassifications. Conversely, the CFPS for the previously discussed weak classes C3, C5, and C6 is lower even though their class-wise accuracies are the least, suggesting relatively few samples are misclassified into these classes than into other classes. This finding underscores the importance of utilizing the CFPS metric, as it provides a more comprehensive and informative assessment of class-wise vulnerability to misclassifications and helps advance our understanding of the class-wise behavior of models in image classification tasks.

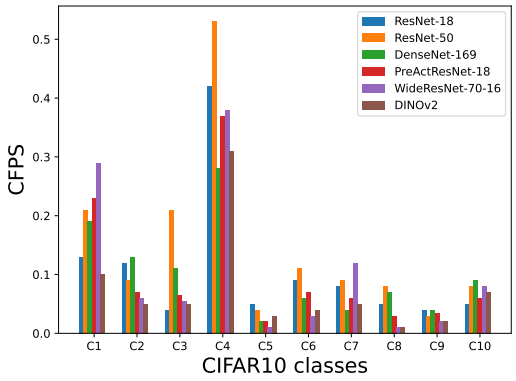

Figure 2: Class-wise CFPS of CIFAR10 across different robust model architectures.

Class-Wise Robustness Analysis Under Common Corruptions. We assess the consistency of class-wise robust classification accuracies and CFPSs with commonly corrupted sample types on the CIFAR10C dataset [11]. Figure 3 presents class-wise robust accuracies and CFPSs of the adversarially trained aforementioned models across various corruption types. Interestingly, weak classes still consistently exhibit the lowest robust classification accuracy even with the inclusion of common corruptions, while C4 ('cat') still maintains the highest CFPS. i.e., the class vulnerabilities to misclassifications remained constant after the addition of common corruptions but the magnitude of vulnerability varies.

Which classes are more vulnerable to adversarial attacks, weak or highly misclassified? Experiments have revealed that the two indicators of the class-wise performance of the model (accuracy and CFPS) point to the distinct properties of the classes. A crucial question here is whether weak classes based on the least class-wise accuracy or those that are mostly misclassified as others are more susceptible to adversarial attacks. Therefore, we further investigate the influence of adversarial attacks [17, 6] on class-wise robustness and also evaluate the most likely targetable class under the influence of attacks. We consider PGD attack [17] using ResNet-50 [3] for this experimentation.

Figure 4 displays the confusion matrix depicting ground truth classes (vertical axis) versus the average of predicted classes over aforementioned models (horizontal axis) after subjecting to PGD attack with $\epsilon = 8/255$ and 20 attack steps. Following the heatmap color, the diagonal elements with the brightest blue shade indicate the lowest number of correct classifications per class and red color indicates the highest. Notably, the class C5 (deer) exhibits the lowest number of correct classifications, implying its vulnerability after subjecting the models to an adversarial attack. Furthermore, an observable pattern is the brightest vertical line aligned with the class C4 (cat) indicating that a significant portion of other classes is being misclassified as this class.

We further evaluate the success rate of PGD-targeted attacks using ResNet-50 [3] with $\epsilon = 2/255$ and 20 attack steps on the CIFAR10 dataset. Figure 5 shows the success rate evaluations of all target classes. The success rate is generally defined as the percentage of misclassifications tricked by the classification model under a targeted attack in the desired target class. We achieve a higher success rate of attack for the target "cat" than that of "deer". This illustrates that the "cat" class is more vulnerable to targeted attacks than "deer".

4. Discussion

The evaluations demonstrate that assessing the class-wise properties of a classification model requires considering both class-wise robust accuracy and CFPS. While the robust accuracy provides insights into a class vulnerability during adversarial attacks and corruptions, it does not necessarily reflect its class-wise susceptibility to misclassifications. It is crucial to identify the class C5 (deer) with the lowest robust accuracy, as it signifies the most vulnerable target for attacks. However, this vulnerability may not translate similarly when facing targeted adversarial scenarios. By examining the CFPS, we gain valuable information

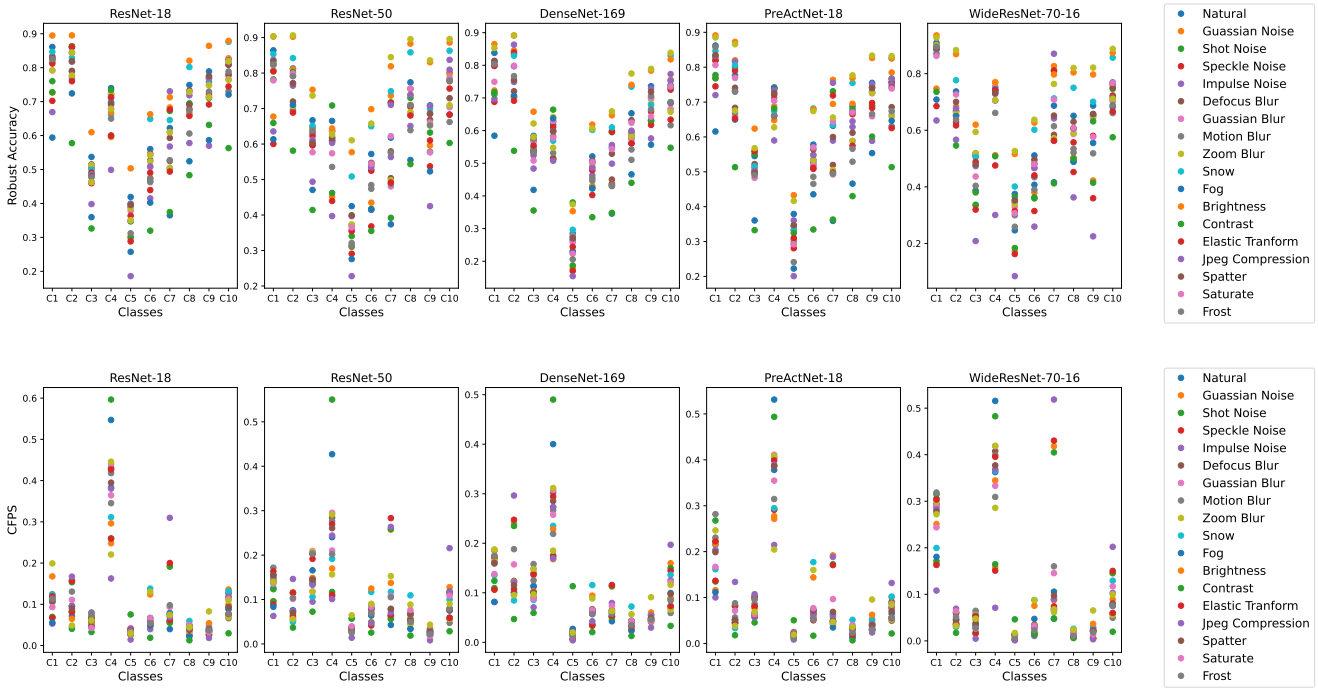

Figure 3: Class-wise robust accuracies(top) and robust CFPSs (bottom) across different model architectures under corruptions. Robust accuracies are presented in fractions. Some classes with reasonably high robust accuracies tend to easily attract false positives and are thus overall more vulnerable than expected.

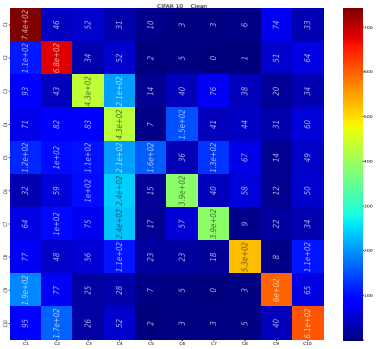

Figure 4: Confusion Matrix defining ground truth (vertical axis) versus predictions(horizontal axis) under PGD attack.

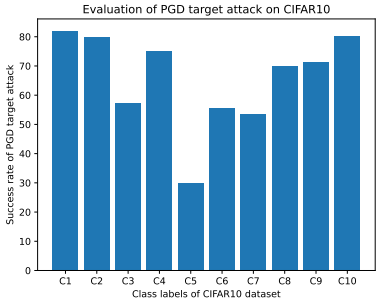

Figure 5: Evaluation of PGD target attack for all classes of CIFAR10 dataset using success rate.

on this, for example the class ”cat” as the most likely to be misclassified into (for CIFAR10). A potential reason is that class C4 (cat) is usually considered a rather difficult class because of the large intra-class variance in cat images. As a result, the label ”cat” might tend to form a rather complex decision space, such that decision boundaries to this label can be easily reached from almost anywhere in the latent space. While this is a specific example on a specific dataset, we assume that similar biases exist across different datasets and models. Models are trained to reach high classification accuracies on potentially difficult classes while it is particularly easy for attackers to fool these models to misclassify other (potentially easy) samples into these classes.

5. Conclusion

In summary, this work studies both class-wise accuracy and class-wise false positives of classes to gain a comprehensive understanding of class-wise vulnerabilities and class-biases present in robust models, empowering us to develop more resilient defenses against potential attacks and corruptions or, at least, to better understand the behavior of our models under domain shifts.

References

[1] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 274–283. PMLR, 10–15 Jul 2018.

[2] 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 *Preregister@NeurIPS*, 2020.

[3] Tianlong Chen, Sijia Liu, Shiyu Chang, Yu Cheng, Lisa Amini, and Zhangyang Wang. Adversarial robustness: From self-supervised pre-training to fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 699–708, 2020.

[4] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*, 2020.

[5] Francesco Croce and Matthias Hein. Mind the box: l_1-apgd for sparse adversarial attacks on image classifiers. In *ICML*, 2021.

[6] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015.

[7] Julia Grabinski, Paul Gavrikov, Janis Keuper, and Margret Keuper. Robust models are less over-confident. *Advances in Neural Information Processing Systems*, 35:39059–39075, 2022.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 630–645. Springer, 2016.

[10] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

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

[12] Kalun Ho, Franz-Josef Pfreundt, Janis Keuper, and Margret Keuper. Estimating the robustness of classification models by the structure of the learned feature-space. *arXiv preprint arXiv:2106.12303*, 2021.

[13] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

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

[15] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale, 2017.

[16] Zheng Lin, Zhao Zhang, Lin-Zhuo Chen, Ming-Ming Cheng, and Shao-Ping Lu. Interactive image segmentation with first click attention. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13336–13345, 2020.

[17] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2019.

[18] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[19] Tianyu Pang, Xiao Yang, Yinpeng Dong, Kun Xu, Jun Zhu, and Hang Su. Boosting adversarial training with hypersphere embedding, 2020.

[20] Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C. Duchi, and Percy Liang. Adversarial training can hurt generalization, 2019.

[21] Christian Szegedy, Alexander Toshev, and Dumitru Erhan. Deep neural networks for object detection. *Advances in neural information processing systems*, 26, 2013.

[22] Qi Tian, Kun Kuang, Kelu Jiang, Fei Wu, and Yisen Wang. Analysis and applications of class-wise robustness in adversarial training. In Feida Zhu, Beng Chin Ooi, and Chunyan Miao, editors, *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, pages 1561–1570. ACM, 2021.

[23] Yisen Wang, Xingjun Ma, James Bailey, Jinfeng Yi, Bowen Zhou, and Quanquan Gu. On the convergence and robustness of adversarial training. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6586–6595. PMLR, 09–15 Jun 2019.

[24] 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*, 2020.

[25] Zeming Wei, Yifei Wang, Yiwen Guo, and Yisen Wang. Cfa: Class-wise calibrated fair adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8193–8201, 2023.

[26] Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE Transactions on Neural Networks and Learning Systems*, 30:2805–2824, 2017.

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

[28] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*,

volume 97 of *Proceedings of Machine Learning Research*, pages 7472–7482. PMLR, 09–15 Jun 2019.