Robustness May Be at Odds with Fairness: An Empirical Study on Class-wise Accuracy

Philipp Benz*
pbenz@kaist.ac.kr

Chaoning Zhang*
chaoningzhang1990@gmail.com

Adil Karjauv
mikolez@gmail.com

In So Kweon
iskweon77@kaist.ac.kr

Abstract

Recently, convolutional neural networks (CNNs) have made significant advancement, however, they are widely known to be vulnerable to adversarial attacks. Adversarial training is the most widely used technique for improving adversarial robustness to strong white-box attacks. Prior works have been evaluating and improving the model average robustness without per-class evaluation. The average evaluation alone might provide a false sense of robustness. For example, the attacker can focus on attacking the vulnerable class, which can be dangerous, especially, when the vulnerable class is a critical one, such as "human" in autonomous driving. In this preregistration submission, we propose an empirical study on the class-wise accuracy and robustness of adversarially trained models. Given that the CIFAR10 training dataset has an equal number of samples for each class, interestingly, preliminary results on it with Resnet18 show that there exists inter-class discrepancy for accuracy and robustness on standard models, for instance, "cat" is more vulnerable than other classes. Moreover, adversarial training increases inter-class discrepancy. Our work aims to investigate the following questions: (a) is the phenomenon of inter-class discrepancy universal for other classification benchmark datasets on other seminal model architectures with various optimization hyper-parameters? (b) If so, what can be possible explanations for the inter-class discrepancy? (c) Can the techniques proposed in the long tail classification be readily extended to adversarial training for addressing the inter-class discrepancy?

1 Introduction

Convolutional neural networks (CNNs) [27] have achieved enormous success in a wide range of applications [49, 23, 24, 32, 48, 44, 25, 47]. However, they are still vulnerable to adversarial attacks. Numerous endeavors have been attempted to improve model adversarial robustness, and adversarial training, to our best knowledge, is the only one that has not been broken by strong white-box attack [14, 29, 7]. Prior works mainly report the model accuracy and robustness averaging on samples from all classes without per-class evaluation. This average performance alone might be misleading for giving a wrong sense of robustness. For example, in autonomous driving, a well-performing model with high accuracy and/or robustness averaging on all classes is particularly dangerous if a certain important class, such as *human*, is vulnerable.

Recognizing its practical relevance, we perform an empirical study to evaluate the per-class accuracy and robustness of adversarially trained models. Preliminary investigation is conducted for CIFAR10 on ResNet18 [17] with both standard training (see Figure 1) and adversarial training (see Figure 2).

*Equal contribution

Pre-registration workshop NeurIPS (2020), Vancouver, Canada.

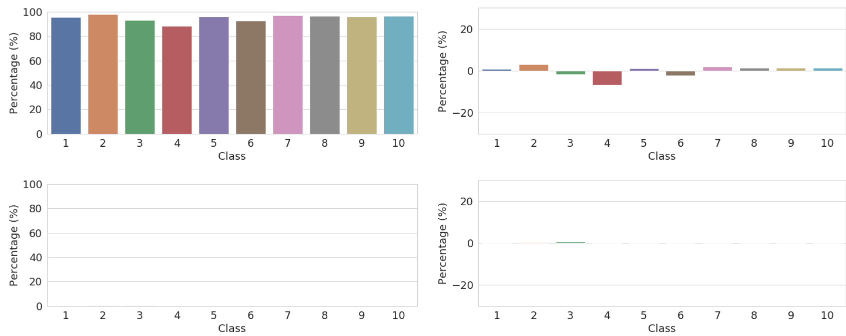

Figure 1: Inter-class discrepancy for standard model. First row: accuracy w/o (left) and w/ (right) mean subtracted. Second row: robustness w/o (left) and w/ (right) mean subtracted.

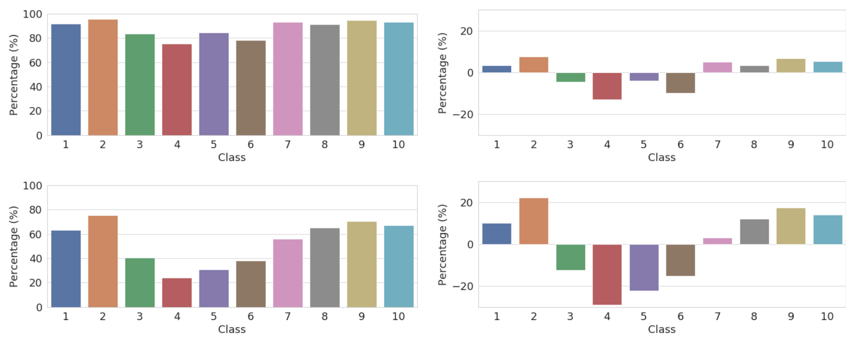

Figure 2: Inter-class discrepancy for adversarially trained model. First row: accuracy w/o (left) and w/ (right) mean subtracted. Second row: robustness w/o (left) and w/ (right) mean subtracted.

There are several intriguing observations. First, there is a non-trivial inter-class discrepancy, even though the long-tail issue does not exist, *i.e.* each class is balanced with the same number of training samples. Second, a similar trend can be observed for the adversarially trained model, more notably, the inter-class discrepancy is more significant than that of a standard model. Third, the imbalance is the most significant for the adversarially trained model. Overall, it suggests that there exists inter-class discrepancy under balanced training dataset and adversarial training increases inter-class discrepancy for both accuracy and robustness.

Our empirical analysis will address the following questions regarding class-wise accuracy and robustness:

- Is the phenomenon of inter-class discrepancy universal in other setups?
- What are possible explanations for this inter-class discrepancy in accuracy and robustness?
- Can the techniques proposed in the long-tail setup be readily extended to adversarial training for addressing the inter-class discrepancy?

2 Related work

2.1 Adversarial examples

CNNs are widely known to be vulnerable to adversarial examples [37, 14, 20, 4, 3, 1], which has inspired numerous investigations on both image-dependent attack [37, 14, 29] and universal attack [30, 45, 46, 2] and defense [33, 43]. Most of the defense techniques have been broken, and currently, adversarial training [14, 29] is the most widely adopted one, empirically proven effective. In the past few years, numerous adversarial training methods [50, 41, 34, 51, 42] have been proposed for improving either effectiveness or efficiency. Despite different motivations and implementations, they all fall into a min-max optimization problem [29], *i.e.* adversarial attack solving the loss maximization problem to generate adversarial examples and network training minimizing the loss to update the network weights. Adversarial training leads to some interesting findings. For example, adversarial training leads to higher robustness while at the cost of accuracy, inspiring to bridge the gap for the trade-off between robustness and accuracy [51]. Adversarial training leads to a model with more robust features [20], consequently reducing information while improving transferability [39]. Adversarially trained models have also been found to be more fit for down-stream transfer learning [40]. Complementary to their findings, we empirically find that adversarial training increases inter-class discrepancy for accuracy and robustness.

2.2 Inter-class discrepancy in long tail recognition

The inter-class discrepancy problem in long-tail recognition is a fundamental issue in machine learning. In the real world setting, the long-tail problem exists when data is inherently imbalanced [22], which undermines the performance of algorithms that do not take this problem into account [5, 21]. Due to its practical relevance, there is a body of work devoted to tackling this problem [16]. The core issue in the long tail recognition lies in low accuracy for the rare classes, and various techniques have been developed to improve the accuracy of those rare classes. Our preliminary results suggest that there exists an inter-class discrepancy for accuracy and robustness for a balanced dataset, especially, adversarial training leads to a more significant inter-class discrepancy. Conceptually, the vulnerable class in the adversarial training is similar to the rare class in the long tail recognition, since the accuracy of them is low and needs to be improved. Straightforwardly, the techniques proposed to solve the long tail recognition might also help mitigate the inter-class discrepancy for the accuracy and/or robustness of adversarially trained models. There are two common techniques in long-tail recognition for handling a class-imbalanced dataset. The first one is to modify the dataset to reduce the imbalance. One can either collect more data samples for the deficient classes [8, 15] or remove samples from the abundant classes to increase balance [11]. Alternatively, the sampling strategy can be designed to increase the sampling frequency for the rare classes during training. The second technique is called cost-sensitive learning, which modifies misclassification costs to account for the imbalance in the number of samples [38, 12, 18]. A recent method [9] introduces the weighting factors for each class to re-balance the loss function. These weights are inversely proportional to the effective number of samples for every class. [6] proposes a new re-balancing optimization procedure with a new loss function to encourage larger classification margins for deficient classes.

3 Methodology and experimental protocol

3.1 Is the phenomenon of inter-class discrepancy universal in other setups?

To test how universal the inter-class discrepancy for accuracy and robustness is, we mainly take into account three factors, *i.e*. datasets, model architectures and optimization methods. For dataset, we plan to test it on various benchmark datasets, including MNIST [28], SVHN [31], CIFAR10 [26], CIFAR100 [26], TinyImageNet, ImageNet [10]. For models, we plan to evaluate on the most seminal models ranging from networks stacking a few convolutional layers to very deep networks, specifically including LeNet [28], VGG family (VGG16 and VGG19) [35], ResNet family (ResNet18 and ResNet50) [17], DenseNet family (DenseNet121 and DenseNet169) [19], Inception family (GoogleNet, Inception-v3) [36]. For optimization factors, we mainly consider optimizer, and learning rate schedule and weight decay.

To check whether the same phenomenon can be observed on the above datasets, we will adopt ResNet18 to train a standard and adversarially trained model for each dataset and evaluate the per-class accuracy and robustness. For the ImageNet dataset, to avoid expensive computation, we will use the online available pre-trained ResNet50, for both standard and adversarially trained ones. For various models, we will test them only on CIFAR10 dataset for avoiding redundancy. For the optimization factors, we will test them with CIFAR10 on ResNet18, but with different optimizers, such as SGD and ADAM, different learning rate schedule (stepwise decrease and cyclic), and different weight decay factors.

Additionally, it would be interesting to see whether the vulnerable class changes during the training. Thus, we will also report the trend of per-class accuracy and robustness during the whole training stage.

3.2 What might be possible explanations for the inter-class discrepancy?

At this stage, we assume that the results from the above investigation would support the following conclusion: The phenomenon of inter-class discrepancy should exist universally for a wide range of datasets on various model architectures with different optimization hyper-parameters. In the balanced dataset setup, each class has the same number of training samples, which increases the chance that the model architectures and/or optimization strategies might influence this phenomenon. We aim to analyze the following results:

- Is the same trend of inter-class discrepancy observed for different networks trained on the same dataset, *i.e.* a robust/vulnerable class on *model A* is also robust/vulnerable on *model B*? If so, we can conclude that the inter-class discrepancy has little to do with model architectures. Otherwise, model architecture can be one factor that causes the phenomenon of the inter-class discrepancy.
- Similarly, we can check the trend regarding different optimization factors. If the same trend is observed for models trained with different optimization, such as standard training vs. adversarial training, or SGD vs. ADAM, we can conclude that this inter-class discrepancy has little to do with optimization factors. Otherwise, optimization factors can be one factor influencing the inter-class discrepancy.

Another important factor is the semantic features in a dataset. We can conduct experiments on CIFAR100, which has super-classes, under which multiple sub-classes share similar semantic features. We can visualize a matrix of ground-truth classes and predicted classes. For example, for samples from ground-truth *class x*, we can count the number of predicted classes and the majority of the samples are likely to be predicted as *x*. If *class x* is semantically similar to *class y*, we expect that there would be a non-trivial amount of samples misclassified to be *class y*. Vice versa, there would also be a nontrivial amount of samples from ground-truth *class y* misclassified to be *class x*. Cross-class feature similarity, represented by the cosine similarity between the output logit vector of two different ground-truth labels, might be a good metric to measure the similarity between classes. Specifically, we measure the average logit vector for samples from each ground-truth label and then perform cross-class feature cosine similarity. If a label has high cross-class feature similarity with other labels, it means the model perceives it to be close to other labels, which might lead to the samples from this label being vulnerable to misclassification.

Feature perspective. Recently, adversarial robustness has been attributed to non-robust features in the dataset [20]. It has been shown in [20] that both robust features and non-robust features are useful for classification. It would be interesting to distinguish whether robust or/and non-robust features lead to the inter-class discrepancy. Following [20], we will construct a dataset that has non-robust features, and train a new model on them and perform per-class accuracy evaluation. As a control study, we will also construct a robust dataset and repeat the above procedure.

3.3 Can the techniques from long tail be extended to adversarial training?

Since our setup has a balanced number of samples for each class, we mainly evaluate the cost-sensitive learning strategy, by giving higher weight on the vulnerable class(es). Since we do not know the performance yet, it is challenging to provide more concrete procedures. We take the following two scenarios into account. First, we assume each class is equally important for the model and the target is to decrease the inter-class discrepancy while minimizing the decrease of the overall average accuracy and/or robustness. Second, we assume that a certain class is critical and the target is to increase the accuracy and/or robustness for that class while minimizing the decrease of accuracy/robustness for other classes. Additionally, we consider including a regularizer term to decrease the inter-class cosine similarity. For adversarial training, we will experiment with a targeted attack by choosing the vulnerable or important class as the target class, which intuitively might make that class more robust. Depending on the performance, we will tailor the above strategies accordingly.

3.4 Additional explorations for robustness against natural corruptions

It has been shown in [13] that corruptions such as noise corruptions, fog or contrast influence standard and adversarially trained models differently. It would be insightful to see whether they might influence the per-class performance differently. Moreover, comparing and analyzing the behavior between adversarial perturbation and natural corruptions can provide insight into the phenomenon of the inter-class discrepancy.

References

[1] Philipp Benz, Chaoning Zhang, Tooba Imtiaz, and In-So Kweon. Data from model: Extracting data from non-robust and robust models. *arXiv preprint arXiv:2007.06196*, 2020.

[2] Philipp Benz, Chaoning Zhang, Tooba Imtiaz, and In So Kweon. Double targeted universal adversarial perturbations. In *ACCV*, 2020.

[3] Philipp Benz, Chaoning Zhang, Adil Karjauv, and In So Kweon. Revisiting batch normalization for improving corruption robustness. *WACV*, 2021.

[4] Philipp Benz, Chaoning Zhang, and In So Kweon. Batch normalization increases adversarial vulnerability: Disentangling usefulness and robustness of model features. *arXiv preprint arXiv:2010.03316*, 2020.

[5] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 2018.

[6] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *NeurIPS*, 2019.

[7] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Symposium on Security and Privacy (SP)*, 2017.

[8] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 2002.

[9] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

[11] Chris Drummond, Robert C Holte, et al. Class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on learning from imbalanced datasets II*, 2003.

[12] Charles Elkan. The foundations of cost-sensitive learning. In *IJCAI*, 2001.

[13] Justin Gilmer, Nicolas Ford, Nicholas Carlini, and Ekin Cubuk. Adversarial examples are a natural consequence of test error in noise. In *ICML*, 2019.

[14] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

[15] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *International joint conference on neural networks*, 2008.

[16] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *Transactions on knowledge and data engineering*, 2009.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[18] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *CVPR*, 2016.

[19] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.

[20] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *NeurIPS*, 2019.

[21] Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intelligent data analysis*, 2002.

[22] Maurice George Kendall et al. The advanced theory of statistics. *The advanced theory of statistics.*, (2nd Ed), 1946.

[23] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Recurrent temporal aggregation framework for deep video inpainting. *TPAMI*, 2019.

[24] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Video panoptic segmentation. In *CVPR*, 2020.

[25] Myungchul Kim, Sanghyun Woo, Dahun Kim, and In So Kweon. The devil is in the boundary: Exploiting boundary representation for basis-based instance segmentation. *WACV*, 2021.

[26] Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009.

[27] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 2015.

[28] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

[29] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

[30] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *CVPR*, 2017.

[31] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NeurIPS*, 2011.

[32] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *CVPR*, 2020.

[33] Shilin Qiu, Qihe Liu, Shijie Zhou, and Chunjiang Wu. Review of artificial intelligence adversarial attack and defense technologies. *Applied Sciences*, 2019.

[34] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! In *NeurIPS*, 2019.

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[36] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[37] 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.

[38] Yuchun Tang, Yan-Qing Zhang, Nitesh V Chawla, and Sven Krasser. Svms modeling for highly imbalanced classification. *Transactions on Systems, Man, and Cybernetics*, 2008.

[39] Matteo Terzi, Alessandro Achille, Marco Maggipinto, and Gian Antonio Susto. Adversarial training reduces information and improves transferability. *arXiv preprint arXiv:2007.11259*, 2020.

[40] Francisco Utrera, Evan Kravitz, N Benjamin Erichson, Rajiv Khanna, and Michael W Mahoney. Adversarially-trained deep nets transfer better. *arXiv preprint arXiv:2007.05869*, 2020.

[41] Jianyu Wang and Haichao Zhang. Bilateral adversarial training: Towards fast training of more robust models against adversarial attacks. In *ICCV*, 2019.

[42] Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. *ICLR*, 2020.

[43] 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*, 2019.

[44] Chaoning Zhang, Philipp Benz, Dawit Mureja Argaw, Seokju Lee, Junsik Kim, Francois Rameau, Jean-Charles Bazin, and In So Kweon. Resnet or densenet? introducing dense shortcuts to resnet. In *WACV*, 2021.

[45] Chaoning Zhang, Philipp Benz, Tooba Imtiaz, and In-So Kweon. Cd-uap: Class discriminative universal adversarial perturbation. In *AAAI*, 2020.

[46] Chaoning Zhang, Philipp Benz, Tooba Imtiaz, and In-So Kweon. Understanding adversarial examples from the mutual influence of images and perturbations. In *CVPR*, 2020.

[47] Chaoning Zhang, Philipp Benz, Adil Karjauv, Geng Sun, and In Kweon. Udh: Universal deep hiding for steganography, watermarking, and light field messaging. *NeurIPS*, 2020.

[48] Chaoning Zhang, Francois Rameau, Junsik Kim, Dawit Mureja Argaw, Jean-Charles Bazin, and In So Kweon. Deepptz: Deep self-calibration for ptz cameras. In *WACV*, 2020.

[49] Chaoning Zhang, Francois Rameau, Seokju Lee, Junsik Kim, Philipp Benz, Dawit Mureja Argaw, Jean-Charles Bazin, and In So Kweon. Revisiting residual networks with nonlinear shortcuts. In *BMVC*, 2019.

[50] Dinghuai Zhang, Tianyuan Zhang, Yiping Lu, Zhanxing Zhu, and Bin Dong. You only propagate once: Accelerating adversarial training via maximal principle. In *NeurIPS*, 2019.

[51] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. In *Proceedings of Machine Learning Research*, 2019.