Robust Models are less Over-Confident

Julia Grabinski [1 2] **Paul Gavrikov** [3] **Janis Keuper** [3 2] **Margret Keuper** [1 4]

Abstract

Despite the success of convolutional neural networks (CNNs) in many academic benchmarks for computer vision tasks, their application in the real-world is still facing fundamental challenges. One of these open problems is the inherent lack of robustness, unveiled by the striking effectiveness of adversarial attacks. Adversarial training (AT) is often considered as a remedy to train more robust networks. In this paper, we empirically analyze a variety of adversarially trained models that achieve high robust accuracies when facing state-of-the-art attacks and we show that AT has an interesting side-effect: it leads to models that are significantly less overconfident with their decisions even on clean data than non-robust models. Further, our analysis shows that not only AT but also the models' building blocks (like activation functions and pooling) have a strong influence on the models' prediction confidences.

1. Introduction

Convolutional Neural Networks (CNNs) have been shown to successfully solve problems across various tasks and domains. However, distribution shifts in the input data can have a severe impact on the prediction performance. In real-world applications, these shifts may be caused by a multitude of reasons including corruption due to weather conditions, camera settings, noise, and maliciously crafted perturbations to the input data intended to fool the network (adversarial attacks). In recent years, a vast line of research (e.g. (Hendrycks & Dietterich, 2019; Goodfellow et al., 2015; Kurakin et al., 2017)) has been devoted to solving robustness issues, highlighting a multitude of causes for the limited generalization ability of networks and potential solutions to facilitate the training of better models.

[1]University of Siegen [2]CC-HPC, Fraunhofer ITWM, Kaiserslautern, [3]IMLA, Offenburg University [4]Max Planck Institute for Informatics, Saarland Informatics Campus, Germany. Correspondence to: Julia Grabinski <julia.grabinski1@itwm.fraunhofer.de>.

Adversarial Machine Learning Frontiers ICML2022.

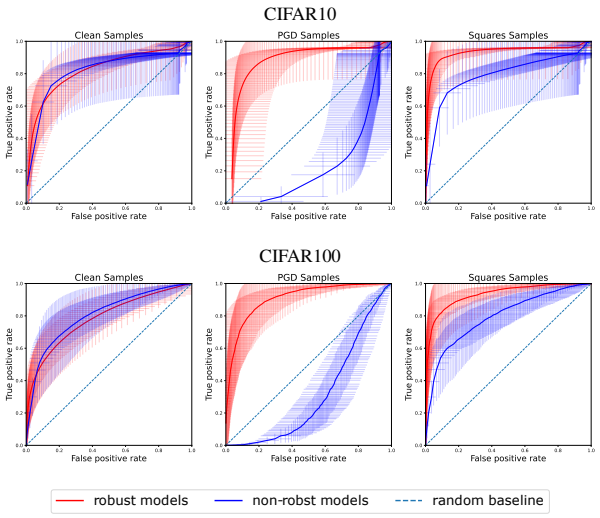

Figure 1. Average ROC curve for all robust and all non-robust models trained on CIFAR10 (top) and CIFAR100 (bottom). Standard deviation is marked by the error bars. The dashed line would mark a model which has the same confidence for each prediction. We observe that the models confidences can be an indicator for the correctness of the prediction. However, on PGD samples the non-robust models fail while the robust models can distinguish correct from incorrect predictions based on the prediction confidence.

A second, yet equally important issue that hampers the deployment of deep learning based models in practical applications is the lack of calibration concerning prediction confidences. In fact, most models are overly confident in their predictions, even if they are wrong (Lakshminarayanan et al., 2017; Guo et al., 2017; Nguyen et al., 2015). Specifically, most conventionally trained models are unaware of their own lack of expertise, i.e. they are trained to make confident predictions in any scenario, even if the test data is sampled from a previously unseen domain. Adversarial examples seem to leverage this weakness, as they are known not only to fool the network but also to cause very confident wrong predictions (Lee et al., 2018). In turn, adversarial training (AT) is known to improve the prediction accuracy under adversarial attacks (Goodfellow et al., 2015; Zhang et al., 2019b; Rony et al., 2019; Engstrom et al., 2019). Yet, a systematic synopsis of the two aspects, adversarial robustness and prediction confidence is still pending.

In this work, we provide an extensive empirical analysis of diverse adversarially robust models concerning their pre-

diction confidences. Therefore, we evaluate the prediction confidences of more than 70 adversarially robust models and their conventionally trained counterparts that show low robustness when exposed to adversarial examples. By measuring their predictive distributions on benign and adversarial examples for correct and erroneous predictions, we show that adversarially trained models have benefits beyond adversarial robustness and are less over-confident.

Our experiments on the datasets CIFAR10 (Krizhevsky, 2012), CIFAR100 and ImageNet (Deng et al., 2009) confirm that non-robust models are overconfident with their false predictions under adversarial attacks. This highlights the challenges for the usage in real-world applications. In contrast, we show that robust models are generally less confident in their predictions, and, especially CNNs which include improved building blocks (downsampling and activation) turn out to be better calibrated manifesting low confidence in wrong predictions and high confidence in their correct predictions.

Our contributions can be summarized as follows:

- We provide an extensive analysis of the prediction confidence of 71 adversarially trained models, and their conventionally trained counterparts. We observe that most non-robust models are exceedingly overconfident while robust models exhibit less confidence and especially are better calibrated. Thus achieving more reliable networks for real-world applications.
- We observe that specific layers, that are considered to improve model robustness impact the models' confidence. In detail, improved downsampling layers and activation functions can lead to an even better calibration of the learned model.

Our analysis provides a first synopsis of adversarial robustness and model calibration and aims to foster research that addresses both challenges jointly rather than considering them as two separate research fields.

2. Related Work

Adversarial Training and Robustness. Adversarial attacks intentionally add perturbations to the input samples, that are almost imperceptible to the human eye, yet lead to (high-confidence) false predictions of the attacked model (Goodfellow et al., 2015; Moosavi-Dezfooli et al., 2016; Szegedy et al., 2014b). In black-box attacks, the adversary has no knowledge of the model intrinsics (Andriushchenko et al., 2020; Chen et al., 2017; Ilyas et al., 2018; Tu et al., 2019). In white-box attacks, the adversary has access to the full model (Goodfellow et al., 2015; Kurakin et al., 2017) and can perform attacks using gradient ascent (Goodfellow et al., 2015; Kurakin et al., 2017; Moosavi-Dezfooli et al., 2016; Carlini & Wagner, 2017; Rony et al., 2019). *AutoAttack* (Croce & Hein, 2020) is a powerful ensemble of different attacks. It is used in the robustness benchmark RobustBench (Croce et al., 2020).

To improve robustness, adversarial training (AT) has proven to be quite successful on common robustness benchmarks. Some attacks can be simply defended by using their adversarial examples in the training set (Goodfellow et al., 2015; Rony et al., 2019) through an additional loss (Engstrom et al., 2019; Zhang et al., 2019b). Furthermore, the addition of more training data, by using external data, or data augmentation techniques such as generation of synthetic data, has been shown to be promising for more robust models (Rebuffi et al., 2021; Gowal et al., 2021a; Carmon et al., 2019; Sehwag et al., 2021; Gowal et al., 2021b; Wang et al., 2020). RobustBench (Croce et al., 2020) provides a leaderboard to study the improvements made by the aforementioned approaches on robustness in a comparable manner in terms of their robust accuracy. Yet, only very few but notable prior works such as (Lakshminarayanan et al., 2017) have investigated AT with respect to model calibration. Without providing a systematic overview, they show that AT can help to smooth the predictive distributions of CNN models. (Tomani & Buettner, 2021) introduce an adversarial calibration loss to reduce the calibration error. Complementary to (Croce et al., 2020), we provide an analysis of the predictive confidences of adversarially trained, robust models and release conventionally trained counterparts of the models from (Croce et al., 2020) to facilitate future research on the analysis of the impact of training schemes versus architectural choices.

Defense besides adversarial training, can be established by the detection and rejection of malicious input. Most such detectors use input sample statistics (Hendrycks & Gimpel, 2016; Li & Li, 2017; Harder et al., 2021; Feinman et al., 2017; Grosse et al., 2017), while others attempt to detect adversarial samples via inference on surrogate models.

Confidence Calibration. For many models that perform well with respect to standard benchmarks, it has been argued that the robust or regular model accuracy may be an insufficient metric (Amodei et al., 2016; DeGroot & Fienberg, 1983; Corbière et al., 2019; Varshney & Alemzadeh, 2017), in particular when real-world applications with potentially open-world scenarios are considered. In these settings, reliability must be established which can be quantified by the prediction confidence (Ovadia et al., 2019). Ideally, a reliable model would provide high confidence predictions on correct classifications, and low confidence predictions on false ones (Corbière et al., 2019; Nguyen et al., 2015). However, most networks are not able to provide a sufficient calibration instantly. Hence, confidence calibration is a vivid field of research and proposed methods are based on additional

loss functions (Lakshminarayanan et al., 2017; Gurau et al., 2018; Moon et al., 2020; Li & Hoiem, 2020; Hein et al., 2019), on adaptions of the training input by label smoothing (Szegedy et al., 2016; Reed et al., 2014; Müller et al., 2019; Qin et al., 2021) or on data augmentation (Zhang et al., 2017; DeVries & Taylor, 2017; Lakshminarayanan et al., 2017; Thulasidasan et al., 2019). Further, (Ovadia et al., 2019) present a benchmark on classification models regarding model accuracy and confidence under dataset shift. Further, different evaluation methods have been provided to distinguish between correct and incorrect predictions (Corbière et al., 2019; Naeini et al., 2015). Naeini et al. (2015) defined the networks *expected calibration error* (ECE)for a model f by with $0 \leq p \leq \infty$

$$\mathrm{ECE}_p = \mathbb{E}[|\hat{z} - \mathbb{E}[1_{\hat{y}=y}|\hat{z}]|^p]^{\frac{1}{p}} \tag{1}$$

where the model f predicts $\hat{y} = y$ with the confidence $\hat{z}$. This can be directly related to the over-confidence $o(f)$ and under-confidence $u(f)$ of a network as follows (Wenger et al., 2020):

$$|o(f)\mathbb{P}(\hat{y} \neq y) - u(f)\mathbb{P}(\hat{y} = y)| \leq \mathrm{ECE}_p, \tag{2}$$

where (Mund et al., 2015)

$$o(f) = \mathbb{E}[\hat{z}|\hat{y} \neq y] \quad u(f) = \mathbb{E}[1 - \hat{z}|\hat{y} = y], \tag{3}$$

i.e. over-confidence measures the expectation of $\hat{z}$ on wrong predictions, under-confidence measures the expectation of $1 - \hat{z}$ on correct predictions and ideally both are zero.

The ECE provides an upper bound for the difference between the probability of the prediction being wrong weighted by the networks over-confidence and the probability of the prediction being correctly weighted by the networks under-confidence and converges to this value for the parameter $p \to 0$ (in eq. 1]). We also recur to this metric as an aggregate measure to evaluate model confidence. Yet, it should be noted that the ECE is based on the assumption that networks make correct as well as incorrect predictions. A model that always makes incorrect predictions and is less confident in its few correct decisions than it is in its many erroneous decisions can end up with a comparably low ECE. Therefore, ECE values for models with an accuracy below 50% are hard to interpret.

Most common CNNs are over-confident (Lakshminarayanan et al., 2017; Guo et al., 2017; Nguyen et al., 2015). Moreover, the most dominantly used activation in modern CNNs (He et al., 2015; Szegedy et al., 2014a; Simonyan & Zisserman, 2015; Huang et al., 2017) is the ReLU function, while Hein et al. (2019) pointed out that ReLUs cause a general increase in the model's prediction confidences, regardless of the prediction validity. This is also the case for the vast majority of the adversarially trained models we consider, except for the model by (Dai et al., 2021) to which we devote particular attention.

3. Analysis

Experimental Setup We have collected 71 checkpoints of robust models listed on the ℓ_∞-*RobustBench* leaderboard (Croce et al., 2020) (see appendix D for a complete list of models). We compare each appearing architecture to a second model trained without AT or any specific robustness regularization, and without any external data (even if the robust counterpart relied on it). Training details can be found in appendix A.

Then we collect the predictions alongside their respective confidences of robust and non-robust models on clean validation samples, as well as on samples attacked by a white-box attack (PGD), and a black-box attack (Squares). PGD (and its adaptive variant APGD (Croce & Hein, 2020)) is the most widely used white-box attack and adversarial training schemes explicitly (when using PGD samples for training) or implicitly (when using the faster but strongly related FGSM attack samples for training) optimize for PGD robustness. In contrast, the *Squares* attack alters the data at random with an allowed budget until the label flips. Such samples are rather to be considered out-of-domain samples even for adversarially trained models and provide a proxy for a model's generalization ability.

3.1. Low Resolution Models

CIFAR10 (Krizhevsky, 2012) is a ten class dataset consisting of 50,000 training and 10,000 validation images with a resolution of 32×32. Since it is significantly cheaper to train on CIFAR10 than on e. g. ImageNet, this dataset became a welcome benchmark for robustness allowing to discount the additional cost of adversarial training, resulting in a number of RobustBench (Croce et al., 2020) entries.

Figure 2 shows an overview of all robust and non-robust models trained on CIFAR10 and CIFAR100 in terms of their accuracy as well as their confidence in their correct and incorrect predictions. The red star in the lower right corner indicates the optimal point where models are highly confident in correct predictions and have zero confidence in their incorrect predictions. Along the isolines, the ratio between confidence in correct and incorrect predictions is constant. The gray area indicates scenarios where models are even more confident in their incorrect predictions than in their correct predictions. Concentrating on the models' confidence, we can see that robust models (marked by a diamond) are in general less confident in their predictions, while non-robust models (marked by a circle) exhibit high confidence in all their predictions, both correct and incorrect. This indicates that non-robust models are not only more susceptible to (adversarial) distribution shifts but are also highly over-confident in their false predictions. Practically, such behaviour can lead to catastrophic consequences in safety-related, real-world applications. Robust models

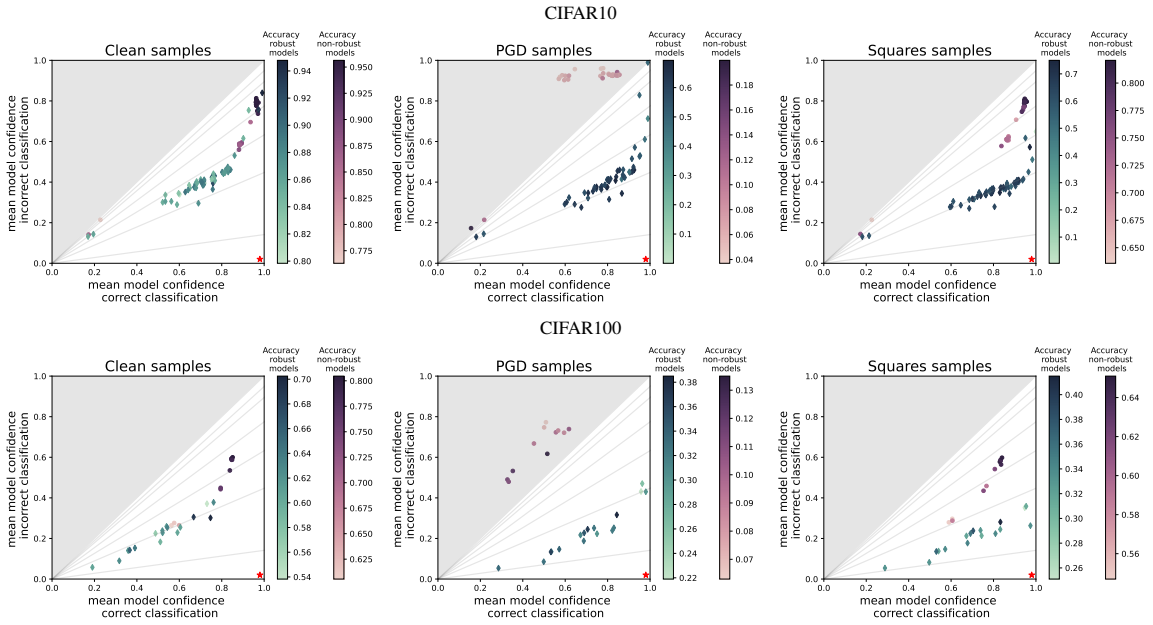

Figure 2. Mean model confidences on their correct (x-axis) and incorrect (y-axis) predictions over the full CIFAR10 dataset (top) and CIFAR100 dataset (bottom), clean (left) and perturbed with the attacks PGD (middle) and Squares (right). Each point represents a model. Circular points (purple color-map) represent non-robust models and diamond-shaped points (green color-map) represent robust models. The color of each point represents the models accuracy, darker signifies higher accuracy (better) on the given data samples. The star in the bottom right corner indicates the optimal model calibration and the gray area marks the area where the confidence distribution of the network is worse than random, i.e. more confident in incorrect predictions than in correct ones.

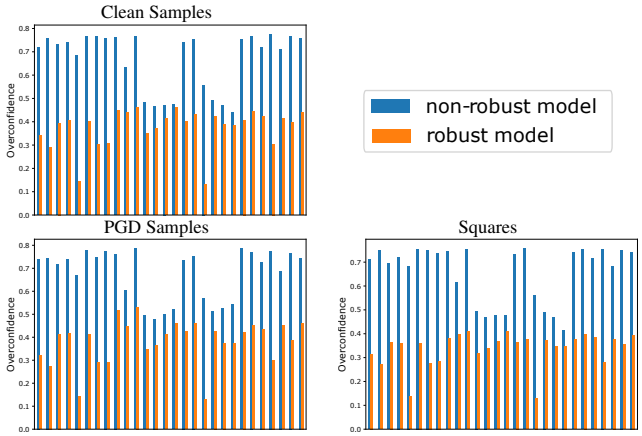

Figure 3. Overconfidence (lower is better) bar plots of robust models and their non-robust counterparts trained on CIFAR10. Non-robust models are highly overconfident, in contrast, their robust counterparts are less over-confident.

tend to have lower average confidence and a favorable confidence trade-off even on clean data (Figure 2, left). When adversarial samples using PGD are considered (Figure 2, middle), the non-robust models even fall into the gray area of the plot where more confident decisions are likely incorrect. As expected, adversarially trained models not only make fewer mistakes in this case but are also better adjusted in terms of their confidence. Black-box attacks (Figure 2, right) provide non-targeted out of domain samples. Adversarially trained models generalize overall well to this case, i.e. their mean confidences are hardly affected whereas non-robust models' confidences fluctuate heavily. Figure 12 further visualizes the significant decrease in over-confidence of robust models w.r.t. their non-robust counterparts. Robust models are better calibrated which results in a significantly lower overconfidence.

Model confidences can predict erroneous decisions. Next, we evaluate the prediction confidences in terms of their ability to predict whether a network prediction is correct or incorrect. We visualize the ROC curves for all models and compare the averages of robust and non-robust models in Figure 1 (top row for CIFAR10, bottom row for CIFAR100), which allows us to draw conclusions about the confidence behavior. While robust and non-robust models perform on average very similarly on clean data, robust model confidences can reliable predict erroneous classification results on adversarial examples where non-robust models fail. Also, for out-of-domain samples from the black-box attack *Squares* (right), robust models can reliably assess their prediction quality and can better predict whether their classification result is correct.

Robust model confidences can detect adversarial samples. Further, we evaluate the adversarial detection rate of the robust models based on their ROC curves (averaged

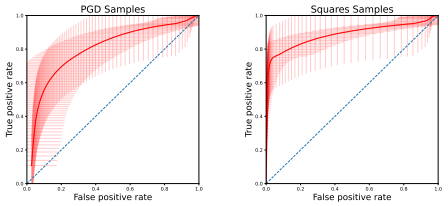

Figure 4. Average ROC curve (red) over all robust models on CIFAR10 of confidence on clean correctly classified samples and perturbed wrongly classified samples. The robust model confidences can be used as threshold for detection of adversarial attacks.

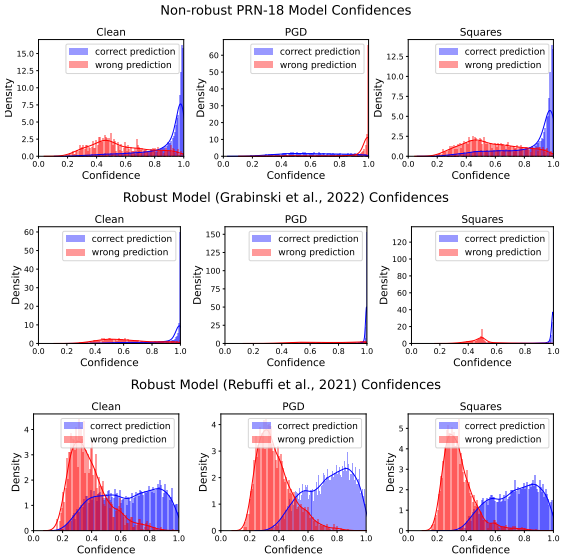

Figure 5. Confidence distribution on three different PRN-18. The first row shows a model without adversarial training and standard pooling, the second row the model by Grabinski et al. (2022b) which uses flc pooling instead of standard pooling and the third row shows the model by Rebuffi et al. (2021) adversarially trained and with standard pooling.

over all robust models) in Figure 4, comparing the confidence of correct predictions on clean samples and incorrect predictions caused by adversarial attacks. We observe that the confidences of robust models can be used to detect adversarial samples by simple thresholding.

Downsampling techniques. Most common CNNs apply downsampling to compress featuremaps with the intent to increase spatial invariance and overall higher sparsity. However, Grabinski et al. (2022b) stated that aliasing during the downsampling operation highly correlates with the lack of adversarial robustness, and provided a downsampling operation, called *frequency low cut pooling* (flc) (Grabinski et al., 2022a), which enables improved downsampling. Figure 5 compares the confidence distribution of three different networks. The top row shows a PRN-18 baseline without AT, the second row the approach by Grabinski et al. (2022a)

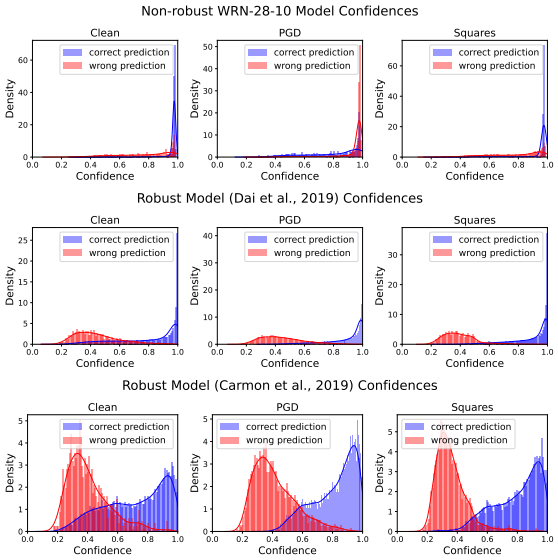

Figure 6. Confidence distribution on three different WRN-28-10. The first row shows a model without adversarial training and standard activation (ReLu), the second row the model by Dai et al. (2021) which uses learnable activation functions instead of fixed ones and the third row shows the model by Carmon et al. (2019) adversarially trained and with the standard activation (ReLu).

applied to the same architecture, and the third row a robust model trained by Rebuffi et al. (2021). The baseline model is highly susceptible to adversarial attacks, especially under white-box attacks, while the two robust counter-parts remain low-confident in false predictions, and show higher confidence in correct predictions. However, while the model of Rebuffi et al. (2021) shows a high variance amongst the predicted confidences, the approach by Grabinski et al. (2022a) significantly improves this by disentangling the confidences. Their model provides low-variance and high-confidence in correct predictions and reduced confidence in false predictions across all evaluated samples.

Activation functions. Next, we analyze the influence of activation functions. Only one RobustBench model utilizes an activation other than ReLU. Dai et al. (2021) introduce learnable activation functions to improve robustness. Figure 6 shows a WRN-28-10 model (top) without AT, the model by Dai et al. (2021) (center) and an adversarially trained model with the same architecture Carmon et al. (2019) (bottom). Although this is an arguably sparse basis for a thorough investigation, we observe that the model by (Dai et al., 2021) can retain high confidence in correct predictions for both clean and perturbed samples. Furthermore, the model is much less confident in its wrong predictions for the clean as well as the adversarial samples, which is in line with the findings of (Hein et al., 2019).

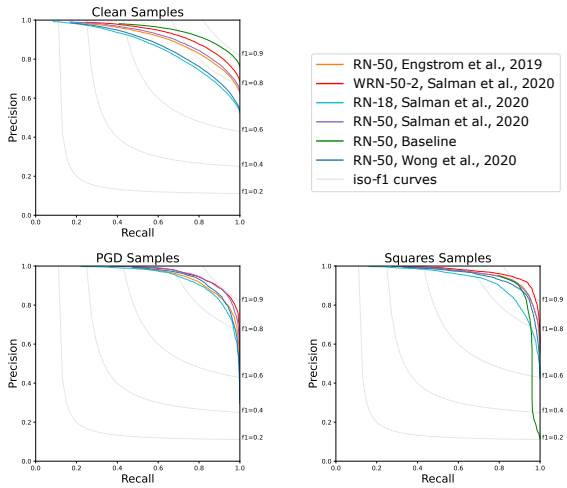

Figure 7. Precision Recall curves for robust and non-robust models trained on ImageNet provided by RobustBench (Croce et al., 2020) over 10000 samples. For the clean samples the non robust baseline can establish the best precision recall curve, followed by the WRN-50-2 by Salman et al. (2020). Similarly this robust WRN-50-2 performs best on the PGD and Squares samples.

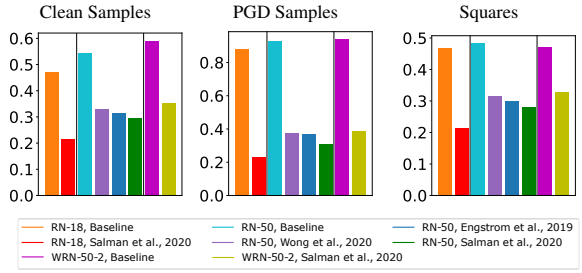

Figure 8. Overconfidence (lower is better) bar plots of the models trained on ImageNet provided by RobustBench (Croce et al., 2020) and their non-robust counterparts. The non-robust baselines exhibits the highest overconfidence. In contrast, the robust models are better calibrated.

3.2. High Resolution Models

As previously introduced, we rely on the models provided by RobustBench (Croce et al., 2020) for our ImageNet evaluation. We report the clean and robust accuracy against *PGD* and *Squares* in Table 3 in the appendix. The non-robust model, trained without AT, achieves the highest performance on clean samples but collapses under white- and black-box attacks. Further, the models trained with multistep adversaries by Engstrom et al. (2019) and Salman et al. (2020) achieve higher robust and clean accuracy than the model trained by Wong et al. (2020) which is trained with single-step adversaries. Moreover, the largest model, a WRN-50-2, yields the best robust performance. Still, the amount of robust networks on ImageNet is quite small, thus we can not make any generalized assumptions.

Figure 7 shows the precision-recall curve for our evaluated models. When evaluated on clean samples, the non-robust model without AT performs best. Under both attacks (*PGD* and *Squares*) the largest model (a WRN-50-2 by Salman et al. (2020)) performs best and the worst performer is the smallest model (RN-18). This may be suggesting that bigger models can not only achieve the better trade-off in clean and robust accuracy but also more successfully disentangle confidences between correct and incorrect predictions. Figure 8 confirms that the over-confidence is decreased in robust models and the ECE is lower than in non-robust models.

4. Discussion

Our experiments confirm that the prediction confidence of non-robust models is an insufficient indicator for the correctness of the prediction, especially under attacks. In contrast, robust models are better behaved and, thus, allow their prediction confidence to serve as a threshold to detect wrongly classified samples. Further, our results indicate that the selection of the activation functions as well as the downsampling are important factors for the models' performance and confidence. The method by Grabinski et al. (2022a), which improves the downsampling, as well as the method by Dai et al. (2021), which improves the activation function, exhibit the best calibration for the networks prediction. While further optimizing deep neural networks' architectures and training schemes, we should therefore consider the synopsis of model robustness and calibration instead of optimizing each of these aspects separately.

Limitations Our evaluation is based on the models provided on RobustBench (Croce et al., 2020). Thus the amount of networks on more complex datasets, like ImageNet, is rather small and therefore the evaluation not universally applicable. While the number of models for CIFAR10 and CIFAR100 is large, the proposed database can only be understood as a starting point for future research.

5. Conclusion

We provide an extensive study on the confidence of robust models and observe an overall trend: robust models tend to be less over-confident than non-robust models. Thus, while achieving a higher robust accuracy, adversarial training provides better calibrated models which are more suited for real-world application. Further, the prediction confidence of robust models can actually be used to reject wrongly classified samples on clean data and even adversarial examples.

Moreover, we see indications that exchanging simple building blocks like the activation function (Dai et al., 2021) or the downsampling method (Grabinski et al., 2022b) alters the properties of robust models with respect to confidence calibration. Our findings should nurture future research on jointly considering model calibration and robustness.

References

Addepalli, S., Jain, S., Sriramanan, G., Khare, S., and Radhakrishnan, V. B. Towards achieving adversarial robustness beyond perceptual limits. In *ICML 2021 Workshop on Adversarial Machine Learning*, 2021. URL https://openreview.net/forum?id=SHB_znlW5G7.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. Concrete problems in ai safety, 2016. URL https://arxiv.org/abs/1606.06565.

Andriushchenko, M. and Flammarion, N. Understanding and improving fast adversarial training. *Advances in Neural Information Processing Systems*, 33:16048–16059, 2020.

Andriushchenko, M., Croce, F., Flammarion, N., and Hein, M. Square attack: a query-efficient black-box adversarial attack via random search. In *European Conference on Computer Vision*, pp. 484–501. Springer, 2020.

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. IEEE, 2017.

Carmon, Y., Raghunathan, A., Schmidt, L., Duchi, J. C., and Liang, P. S. Unlabeled data improves adversarial robustness. *Advances in Neural Information Processing Systems*, 32, 2019.

Chen, E.-C. and Lee, C.-R. Ltd: Low temperature distillation for robust adversarial training, 2021.

Chen, J., Cheng, Y., Gan, Z., Gu, Q., and Liu, J. Efficient robust training via backward smoothing, 2021.

Chen, P.-Y., Zhang, H., Sharma, Y., Yi, J., and Hsieh, C.-J. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pp. 15–26, 2017.

Chen, T., Liu, S., Chang, S., Cheng, Y., Amini, L., and Wang, Z. Adversarial robustness: From self-supervised pre-training to fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 699–708, 2020.

Corbière, C., Thome, N., Bar-Hen, A., Cord, M., and Pérez, P. Addressing failure prediction by learning model confidence. *Advances in Neural Information Processing Systems*, 32, 2019.

Croce, F. and Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2020.

Croce, F., Andriushchenko, M., Sehwag, V., Flammarion, N., Chiang, M., Mittal, P., and Hein, M. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*, 2020.

Cui, J., Liu, S., Wang, L., and Jia, J. Learnable boundary guided adversarial training, 2021.

Dai, S., Mahloujifar, S., and Mittal, P. Parameterizing activation functions for adversarial robustness, 2021.

DeGroot, M. H. and Fienberg, S. E. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

DeVries, T. and Taylor, G. W. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Ding, G. W., Sharma, Y., Lui, K. Y. C., and Huang, R. Mma training: Direct input space margin maximization through adversarial training. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HkeryxBtPB.

Engstrom, L., Ilyas, A., Salman, H., Santurkar, S., and Tsipras, D. Robustness (python library), 2019. URL https://github.com/MadryLab/robustness.

Feinman, R., Curtin, R. R., Shintre, S., and Gardner, A. B. Detecting adversarial samples from artifacts. *arXiv preprint arXiv:1703.00410*, 2017.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples, 2015.

Gowal, S., Qin, C., Uesato, J., Mann, T., and Kohli, P. Uncovering the limits of adversarial training against norm-bounded adversarial examples, 2021a.

Gowal, S., Rebuffi, S.-A., Wiles, O., Stimberg, F., Calian, D. A., and Mann, T. A. Improving robustness using generated data. *Advances in Neural Information Processing Systems*, 34, 2021b.

Grabinski, J., Jung, S., Keuper, J., and Keuper, M. Frequencylowcut pooling–plug & play against catastrophic overfitting. *arXiv preprint arXiv:2204.00491*, 2022a.

Grabinski, J., Keuper, J., and Keuper, M. Aliasing coincides with CNNs vulnerability towards adversarial attacks. In *The AAAI-22 Workshop on Adversarial Machine Learning*

and Beyond, 2022b. URL https://openreview.net/forum?id=vKc1mLxBebP.

Grosse, K., Manoharan, P., Papernot, N., Backes, M., and McDaniel, P. On the (statistical) detection of adversarial examples. *arXiv preprint arXiv:1702.06280*, 2017.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Gurau, C., Bewley, A., and Posner, I. Dropout distillation for efficiently estimating model confidence. *arXiv preprint arXiv:1809.10562*, 2018.

Harder, P., Pfreundt, F.-J., Keuper, M., and Keuper, J. Spectraldefense: Detecting adversarial attacks on cnns in the fourier domain. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition, 2015.

Hein, M., Andriushchenko, M., and Bitterwolf, J. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem, 2019.

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

Hendrycks, D. and Gimpel, K. Early methods for detecting adversarial images. *arXiv preprint arXiv:1608.00530*, 2016.

Hendrycks, D., Lee, K., and Mazeika, M. Using pre-training can improve model robustness and uncertainty. In *International Conference on Machine Learning*, pp. 2712–2721. PMLR, 2019.

Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Huang, H., Wang, Y., Erfani, S. M., Gu, Q., Bailey, J., and Ma, X. Exploring architectural ingredients of adversarially robust deep neural networks, 2022.

Huang, L., Zhang, C., and Zhang, H. Self-adaptive training: beyond empirical risk minimization, 2020.

Ilyas, A., Engstrom, L., Athalye, A., and Lin, J. Black-box adversarial attacks with limited queries and information. In *International Conference on Machine Learning*, pp. 2137–2146. PMLR, 2018.

Krizhevsky, A. Learning multiple layers of features from tiny images. *University of Toronto*, 05 2012.

Kurakin, A., Goodfellow, I., and Bengio, S. Adversarial machine learning at scale, 2017.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Lee, K., Lee, K., Lee, H., and Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks, 2018. URL https://arxiv.org/abs/1807.03888.

Li, X. and Li, F. Adversarial examples detection in deep networks with convolutional filter statistics. In *Proceedings of the IEEE international conference on computer vision*, pp. 5764–5772, 2017.

Li, Z. and Hoiem, D. Improving confidence estimates for unfamiliar examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2686–2695, 2020.

Moon, J., Kim, J., Shin, Y., and Hwang, S. Confidence-aware learning for deep neural networks. In *international conference on machine learning*, pp. 7034–7044. PMLR, 2020.

Moosavi-Dezfooli, S.-M., Fawzi, A., and Frossard, P. 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.

Müller, R., Kornblith, S., and Hinton, G. E. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019.

Mund, D., Triebel, R., and Cremers, D. Active online confidence boosting for efficient object classification. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1367–1373, 2015. doi: 10.1109/ICRA.2015.7139368.

Naeini, M. P., Cooper, G., and Hauskrecht, M. Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Nguyen, A., Yosinski, J., and Clune, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 427–436, 2015.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., and Snoek, J. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

Pang, T., Yang, X., Dong, Y., Xu, K., Zhu, J., and Su, H. Boosting adversarial training with hypersphere embedding. *Advances in Neural Information Processing Systems*, 33:7779–7792, 2020.

Qin, Y., Wang, X., Beutel, A., and Chi, E. Improving calibration through the relationship with adversarial robustness. *Advances in Neural Information Processing Systems*, 34: 14358–14369, 2021.

Rade, R. and Moosavi-Dezfooli, S.-M. Helper-based adversarial training: Reducing excessive margin to achieve a better accuracy vs. robustness trade-off. In *ICML 2021 Workshop on Adversarial Machine Learning*, 2021. URL https://openreview.net/forum?id=BuD2LmNaU3a.

Rebuffi, S.-A., Gowal, S., Calian, D. A., Stimberg, F., Wiles, O., and Mann, T. Fixing data augmentation to improve adversarial robustness, 2021.

Reed, S., Lee, H., Anguelov, D., Szegedy, C., Erhan, D., and Rabinovich, A. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014.

Rice, L., Wong, E., and Kolter, Z. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pp. 8093–8104. PMLR, 2020.

Rony, J., Hafemann, L. G., Oliveira, L. S., Ayed, I. B., Sabourin, R., and Granger, E. Decoupling direction and norm for efficient gradient-based l2 adversarial attacks and defenses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4322–4330, 2019.

Salman, H., Ilyas, A., Engstrom, L., Kapoor, A., and Madry, A. Do adversarially robust imagenet models transfer better? *Advances in Neural Information Processing Systems*, 33:3533–3545, 2020.

Sehwag, V., Wang, S., Mittal, P., and Jana, S. Hydra: Pruning adversarially robust neural networks, 2020.

Sehwag, V., Mahloujifar, S., Handina, T., Dai, S., Xiang, C., Chiang, M., and Mittal, P. Robust learning meets generative models: Can proxy distributions improve adversarial robustness?, 2021.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition, 2015.

Sitawarin, C., Chakraborty, S., and Wagner, D. Sat: Improving adversarial training via curriculum-based loss smoothing, 2021.

Sridhar, K., Sokolsky, O., Lee, I., and Weimer, J. Improving neural network robustness via persistency of excitation, 2021.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions, 2014a.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014b. URL http://arxiv.org/abs/1312.6199.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Thulasidasan, S., Chennupati, G., Bilmes, J. A., Bhattacharya, T., and Michalak, S. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Tomani, C. and Buettner, F. Towards trustworthy predictions from deep neural networks with fast adversarial calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9886–9896, 2021.

Tu, C.-C., Ting, P., Chen, P.-Y., Liu, S., Zhang, H., Yi, J., Hsieh, C.-J., and Cheng, S.-M. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 742–749, 2019.

Varshney, K. R. and Alemzadeh, H. On the safety of machine learning: Cyber-physical systems, decision sciences, and data products. *Big data*, 5(3):246–255, 2017.

Wang, Y., Zou, D., Yi, J., Bailey, J., Ma, X., and Gu, Q. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rklOg6EFwS.

Wenger, J., Kjellström, H., and Triebel, R. Non-parametric calibration for classification. In *International Conference on Artificial Intelligence and Statistics*, pp. 178–190. PMLR, 2020.

Wightman, R. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

Wong, E., Rice, L., and Kolter, J. Z. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJx040EFvH.

Wu, D., Xia, S.-T., and Wang, Y. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems*, 33:2958–2969, 2020.

Zhang, D., Zhang, T., Lu, Y., Zhu, Z., and Dong, B. You only propagate once: Accelerating adversarial training via maximal principle, 2019a.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Zhang, H., Yu, Y., Jiao, J., Xing, E. P., Ghaoui, L. E., and Jordan, M. I. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, 2019b.

Zhang, J., Xu, X., Han, B., Niu, G., Cui, L., Sugiyama, M., and Kankanhalli, M. Attacks which do not kill training make adversarial learning stronger, 2020.

Zhang, J., Zhu, J., Niu, G., Han, B., Sugiyama, M., and Kankanhalli, M. Geometry-aware instance-reweighted adversarial training. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=iAX0l6Cz8ub.

A. Non-robust Model Training

For training, *CIFAR-10/100* data was zero-padded by 4 px along each dimension, and then transformed using 32×32 px random crops, and random horizontal flips. Channel-wise normalization was replicated as reported by the original dataset authors. Training hyper parameters have been set to an initial learning rate of 1e-2, a weight decay of 1e-2, a batch-size of 256 and a nesterov momentum of 0.9. We scheduled the SGD optimizer to decrease the learning rate every 30 epochs by a factor of $\gamma = 0.1$ and trained for a total of 125 epochs. The loss is determined using Categorical Cross Entropy and we used the model obtained at the epoch with the highest validation accuracy. Training *ImageNet1k* architectures with our hyperparameters resulted in a rather poor performance and we therefore rely on the baseline model without AT provided by *timm* (Wightman, 2019).

B. Additional Evaluation CIFAR10/100

Following we provided an overview over all robust and non-robust counterparts and their ECE on CIFAR10 and CIFAR100 are listed.

CIFAR100, includes 100 classes and can be seen as a more challenging classification task. This is reflected in the reduced model accuracy on the clean and adversarial samples (Figure 2 , bottom). On this data, robust models slightly closer to the optimal calibration point in the lower right corner even on clean data and perform significantly better on PGD samples where the confidences of non-robust models are again reversed (middle). The Squares attack again illustrates the stable behavior of robust models'. The models' full empirical confidence distributions are given in Figure 10. We also report the ECE values for CIFAR100 in Table 2. Please note that the accuracy of the CIFAR100 models is not very high (ranging between 56.87% and 70.25% even for clean samples), resulting in an unreliable calibration metric. Especially under PGD attacks, non-robust networks make mostly incorrect predictions such that the ECE collapses to being the expected confidence value of incorrect predictions (see eq. [1]), regardless of the confidences of the few correct predictions. In this case, ECE is not meaningful.

B.1. Confidence Distribution

The model confidences distributions are shown in Figure 9 and Figure 10. Each row contains the robust and non-robust counterpart and their confidence distributions on the clean samples and the perturbated samples by PGD and Squares.

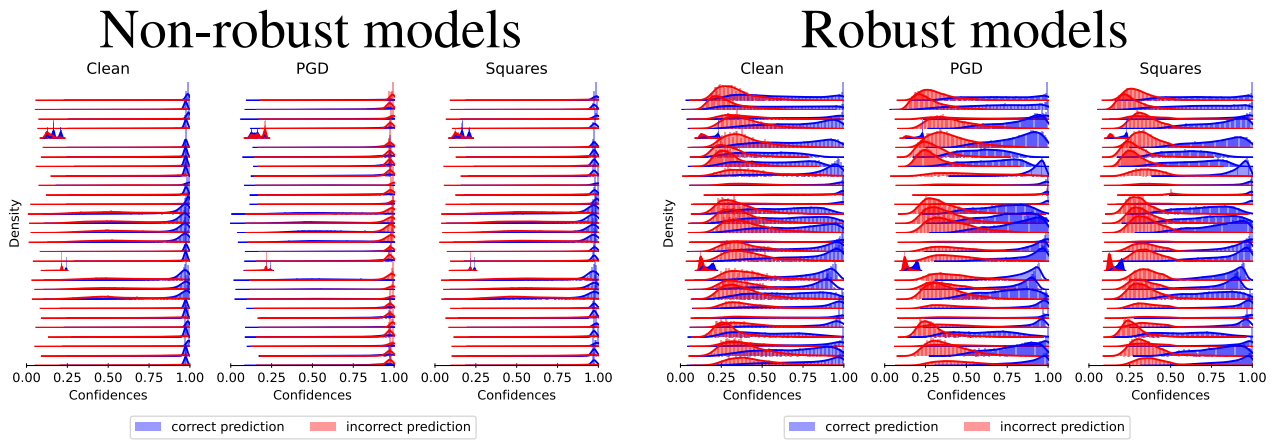

Figure 9. Density plots for robust and non-robust models on CIFAR10 over the models confidence on its correct and incorrect predictions. Each row contains the same model adversarially and standard trained. The non.robust models show high confidence on all of their predictions, however those might be wrong. Especially in the case of PGD samples the models are highly confidence in their false predictions. In contrast the robust models are better calibrated. The robust models are confident in their correct predictions and less confident in their false prediction.

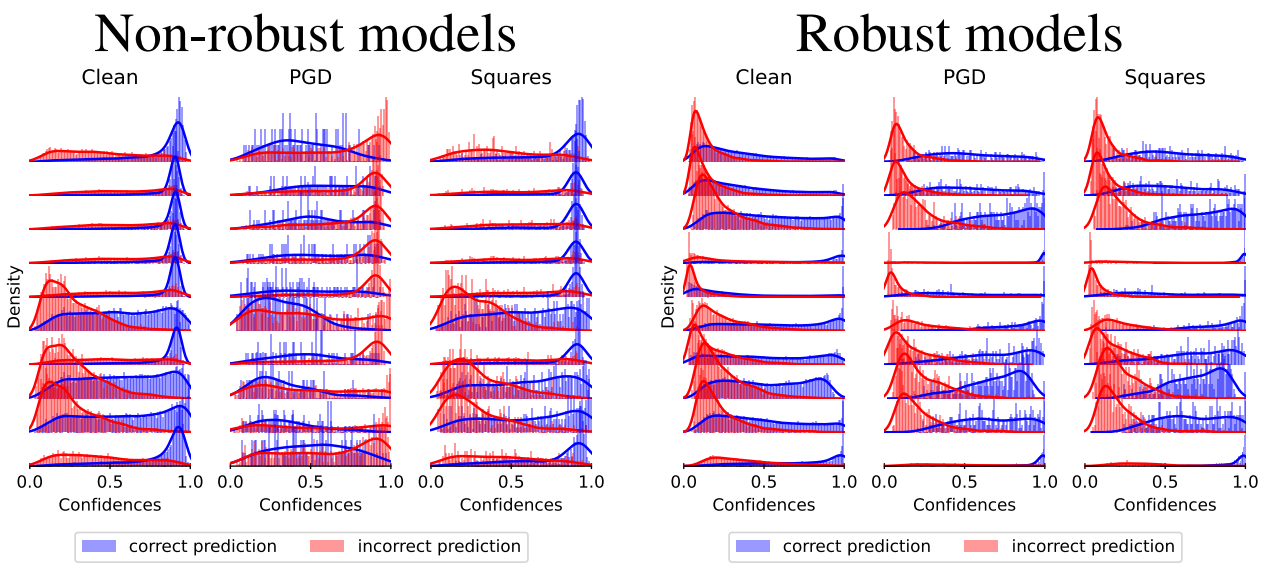

Figure 10. Density plots for robust and non-robust models on CIFAR100 over the models confidence on its correct and incorrect predictions. Each row contains the same model adversarially and standard trained. The non.robust models show high confidence on all of their predictions, however those might be wrong. Especially in the case of PGD samples the models are highly confidence in their false predictions. In contrast the robust models are better calibrated. The robust models are confident in their correct predictions and less confident in their false prediction.

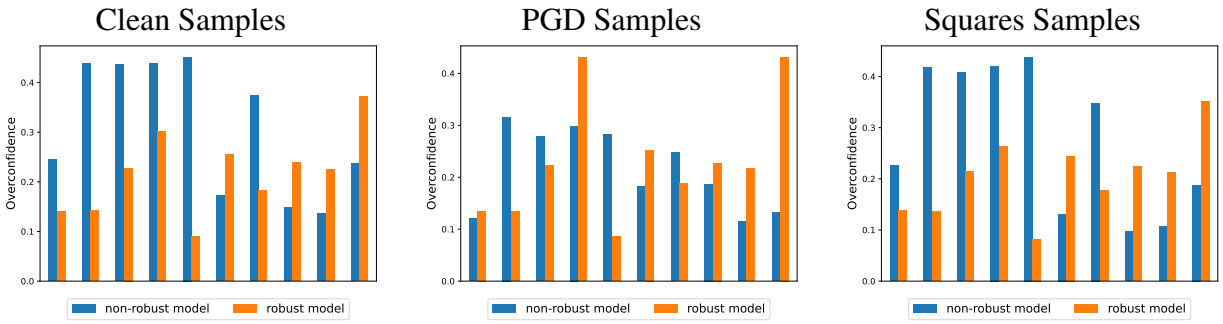

Figure 11. Overconfidence (lower is better) bar plots of robust models and their non-robust counterparts trained on CIFAR100.

B.2. Overconfidence and ECE

Table 1 reports the mean ECE over all robust models and their non-robust counterparts. Robust models are better calibrated which results in a significantly lower ECE. The models' full empirical confidence distributions are given in Figure 9.

Similar, the confidence distributions for the robust and non-robust counterparts on CIFAR100 are depicted in Figure 10.

Samples \ Robustness	Clean	PGD	Squares
non-robust models	0.3077 ± 0.1257	0.2159 ± 0.0738	0.2780 ± 0.1348
robust models	0.2962 ±0.1722	0.2307 ± 0.1494	0.2076 ± 0.1247

Table 2. Mean ECE (lower is better) and standard deviation over all non-robust model versus all their robust counterparts trained on CIFAR100. Robust model exhibit a significantly lower ECE on all samples.

B.3. Precision Recall

For completeness we included the Precision Recall curves on CIFAR10 and CIFAR100 as mean over all robust and non-robust models with marked standard deviation.

Samples \ Robustness	Clean	PGD	Squares
non-robust models	0.6736 ± 0.1208	0.6809 ± 0.1061	0.6635 ±0.1156
robust models	0.1894 ±0.1531	0.2688 ± 0.1733	0.2126 ± 0.1431

Table 1. Mean ECE (lower is better) and standard deviation over all non-robust models versus all their robust counterparts trained on CIFAR10. Robust models exhibit a significantly lower mean ECE.

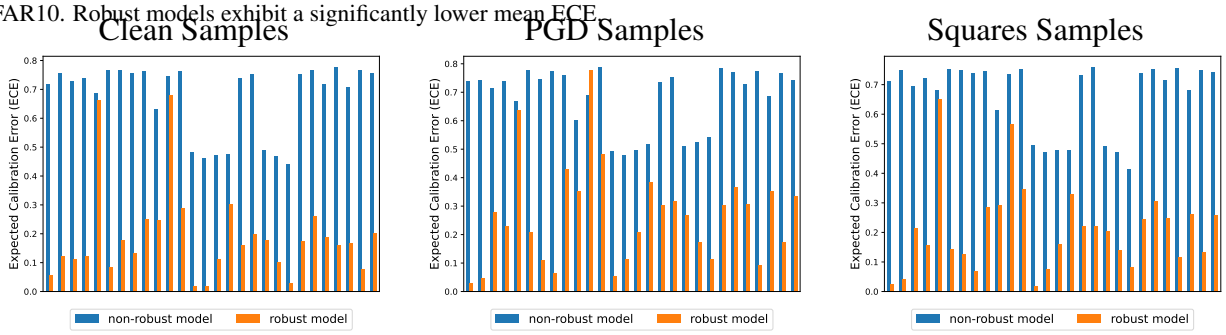

Figure 12. ECE (lower is better) bar plots of robust models and their non-robust counterparts trained on CIFAR10.

C. Additional Evaluation ImageNet

Following table 3 reports the accuracy evaluation of the robust models as well as the baseline on ImageNet. The accuracy is reported on the clean as well as on the perturbated samples by PGD and Squares with an ϵ of $4/255$.

For completeness we included the ROC curve in Figure on the clean as well as the perturbated samples for the robust models and the baseline on ImagNet in Figure 17 as well as the evaluation of the ECE in Figure 18.

Method	Architecture	Clean Acc ↑	PGD Acc ↑	Squares Acc ↑
Baseline	RN50	**76.13**	0.00	11.48
Engstrom et al. (2019)	RN50	62.41	35.47	54.93
Wong et al. (2020)	RN50	53.83	29.43	42.26
Salman et al. (2020)	RN50	63.87	42.23	56.58
Salman et al. (2020)	WRN50-2	68.41	**44.75**	**61.29**
Salman et al. (2020)	RN18	52.50	31.92	43.81

Table 3. Clean and robust accuracy on ImageNet against PGD and Squares (higher is better) over 10000 samples.

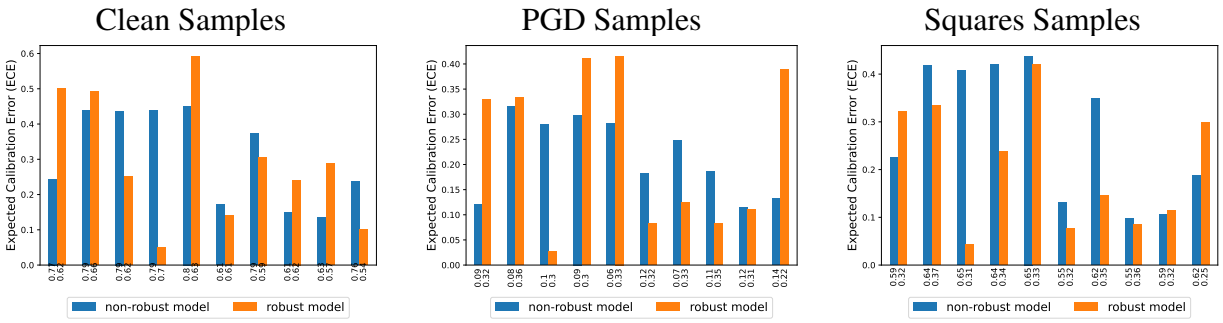

Figure 13. ECE (lower is better) bar plots of robust models and their non-robust counterparts trained on CIFAR100. The models accuracy are marked for the different samples for each bar.

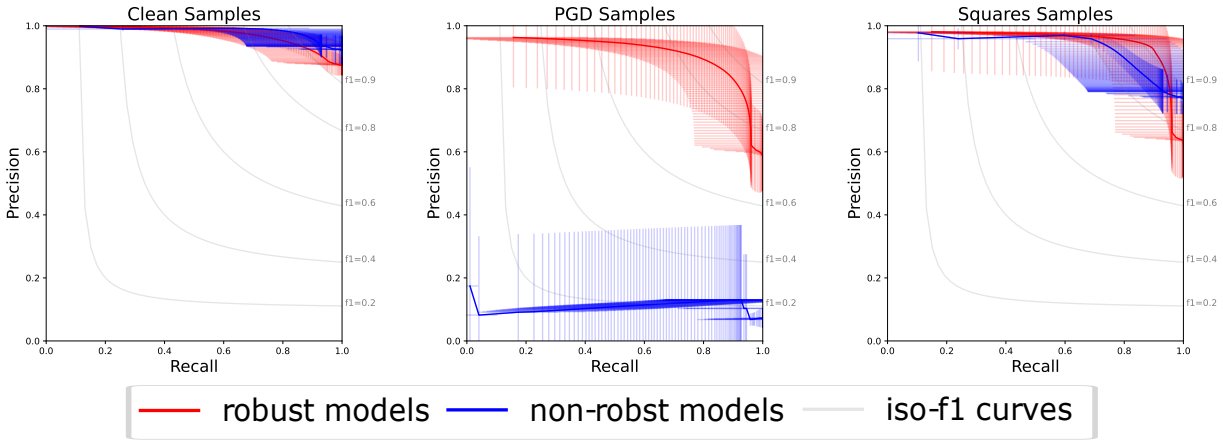

Figure 14. Average precision recall curve for all robust and all non-robust models trained on CIFAR10. Standard deviation is marked by the error bars. For the clean samples the non robust models can distinguish slightly better into correct and incorrect predictions based on the predictions confidence. The superior of the robust models is clearly visible on the samples created by PGD, the non.robust models are not able to distinguish. However, for the samples created by Squares the classification into correct and incorrect prediction based on the confidence is almost equally possible for robust and non-robust models.

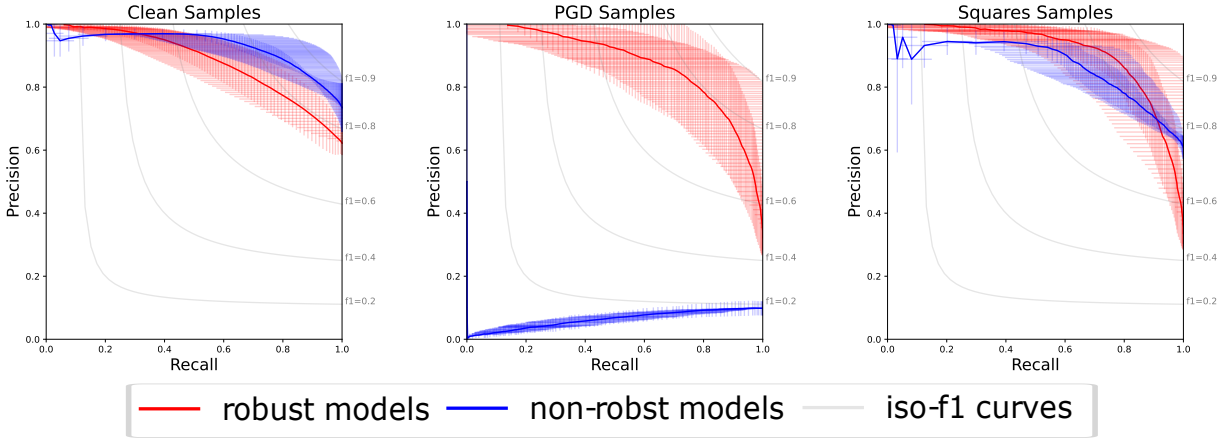

Figure 15. Average precision recall curve for all robust and all non-robust models trained on CIFAR100 for 1000 samples. Standard deviation is marked by the error bars. For the clean samples the non robust models can distinguish slightly better into correct and incorrect predictions based on the predictions confidence. The superior of the robust models is clearly visible on the samples created by PGD, the non.robust models are not able to distinguish. However, for the samples created by Squares the classification into correct and incorrect prediction based on the confidence is almost equally possible for robust and non-robust models.

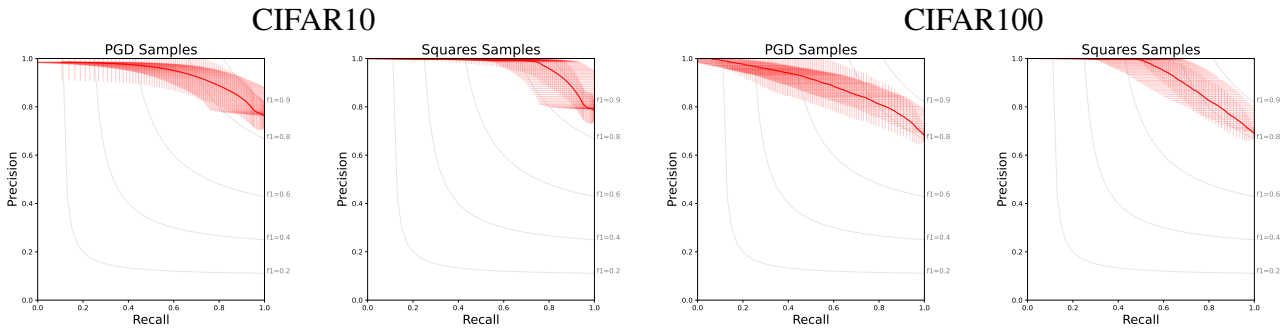

Figure 16. Precision Recall curve between confidence of clean correct samples and perturbated wrong samples on CIFAR10 and CIFAR100. The robust model confidences can be used as threshold for detection of adversarial attacks.

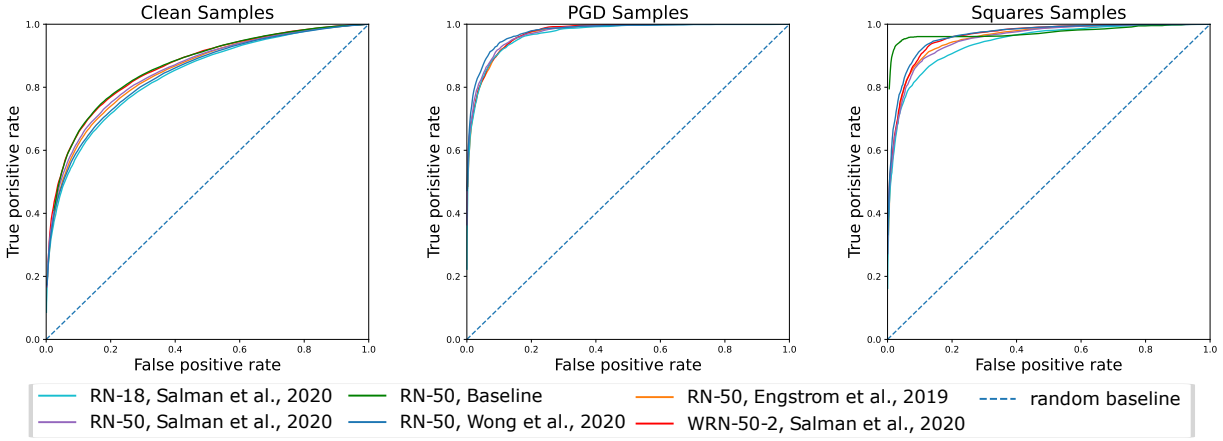

Figure 17. ROC curves for the robust models and the non-robust baseline trained on ImageNet provided on RobustBench (Croce et al., 2020).

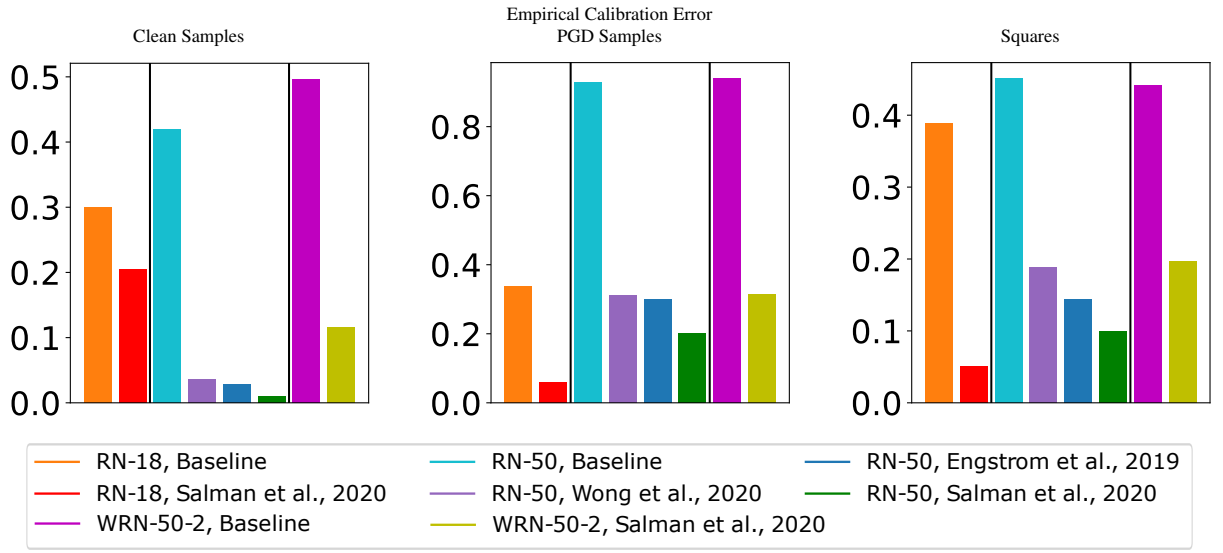

Figure 18. ECE (lower is better) bar plots of the models trained on ImageNet provided by RobustBench (Croce et al., 2020) and their non-robust counterparts. The non-robust baselines exhibits the highest ECE. In contrast, the robust models are better calibrated.

D. Model Overview

The robust checkpoints provided by *RobustBench* (Croce et al., 2020) are licensed under the MIT Licence. The clean models for ImageNet are provided by *timm* (Wightman, 2019) under the Apache 2.0 licence.

Paper	Dataset	Architecture	Adv. Trained Clean Acc.	Adv. Trained Robust Acc.	Norm. Trained Clean Acc.	Norm. Trained Robust Acc.
(Andriushchenko & Flammarion, 2020)	cifar10	PreActResNet-18	79.84	43.93	94.51	0.0
(Carmon et al., 2019)	cifar10	WideResNet-28-10	89.69	59.53	95.10	0.0
(Sehwag et al., 2020)	cifar10	WideResNet-28-10	88.98	57.14	95.10	0.0
(Wang et al., 2020)	cifar10	WideResNet-28-10	87.50	56.29	95.10	0.0
(Hendrycks et al., 2019)	cifar10	WideResNet-28-10	87.11	54.92	95.35	0.0
(Rice et al., 2020)	cifar10	WideResNet-34-20	85.34	53.42	95.46	0.0
(Zhang et al., 2019b)	cifar10	WideResNet-34-10	84.92	53.08	95.26	0.0
(Engstrom et al., 2019)	cifar10	ResNet-50	87.03	49.25	94.90	0.0
(Chen et al., 2020)	cifar10	ResNet-50	86.04	51.56	86.50	0.0
(Huang et al., 2020)	cifar10	WideResNet-34-10	83.48	53.34	95.26	0.0
(Pang et al., 2020)	cifar10	WideResNet-34-20	85.14	53.74	76.30	0.0
(Wong et al., 2020)	cifar10	PreActResNet-18	83.34	43.21	94.25	0.0
(Ding et al., 2020)	cifar10	WideResNet-28-4	84.36	41.44	94.33	0.0
(Zhang et al., 2019a)	cifar10	WideResNet-34-10	87.20	44.83	95.26	0.0
(Zhang et al., 2020)	cifar10	WideResNet-34-10	84.52	53.51	95.26	0.0
(Wu et al., 2020)	cifar10	WideResNet-28-10	88.25	60.04	95.10	0.0
(Wu et al., 2020)	cifar10	WideResNet-34-10	85.36	56.17	95.64	0.0
(Gowal et al., 2021a)	cifar10	WideResNet-70-16	85.29	57.20	87.91	0.0
(Gowal et al., 2021a)	cifar10	WideResNet-70-16	91.10	65.88	87.91	0.0
(Gowal et al., 2021a)	cifar10	WideResNet-34-20	85.64	56.86	88.33	0.0
(Gowal et al., 2021a)	cifar10	WideResNet-28-10	89.48	62.80	88.20	0.0
(Sehwag et al., 2021)	cifar10	WideResNet-34-10	85.85	59.09	95.64	0.0
(Sehwag et al., 2021)	cifar10	ResNet-18	84.38	54.43	94.87	0.0
(Sitawarin et al., 2021)	cifar10	WideResNet-34-10	86.84	50.72	95.26	0.0
(Chen et al., 2021)	cifar10	WideResNet-34-10	85.32	51.12	95.35	0.0
(Cui et al., 2021)	cifar10	WideResNet-34-20	88.70	53.57	95.44	0.0
(Cui et al., 2021)	cifar10	WideResNet-34-10	88.22	52.86	95.26	0.0
(Zhang et al., 2021)	cifar10	WideResNet-28-10	89.36	59.64	95.10	0.0
(Rebuffi et al., 2021)	cifar10	WideResNet-28-10	87.33	60.75	88.20	0.0
(Rebuffi et al., 2021)	cifar10	WideResNet-106-16	88.50	64.64	86.92	0.0
(Rebuffi et al., 2021)	cifar10	WideResNet-70-16	88.54	64.25	87.91	0.0
(Rebuffi et al., 2021)	cifar10	WideResNet-70-16	92.23	66.58	87.91	0.0
(Sridhar et al., 2021)	cifar10	WideResNet-28-10	89.46	59.66	95.10	0.0
(Sridhar et al., 2021)	cifar10	WideResNet-34-15	86.53	60.41	95.50	0.0
(Rebuffi et al., 2021)	cifar10	PreActResNet-18	83.53	56.66	89.01	0.0
(Rade & Moosavi-Dezfooli, 2021)	cifar10	PreActResNet-18	89.02	57.67	89.01	0.0
(Rade & Moosavi-Dezfooli, 2021)	cifar10	PreActResNet-18	86.86	57.09	89.01	0.0
(Rade & Moosavi-Dezfooli, 2021)	cifar10	WideResNet-34-10	91.47	62.83	88.67	0.0
(Rade & Moosavi-Dezfooli, 2021)	cifar10	WideResNet-28-10	88.16	60.97	88.20	0.0
(Huang et al., 2022)	cifar10	WideResNet-34-R	90.56	61.56	95.60	0.0
(Huang et al., 2022)	cifar10	WideResNet-34-R	91.23	62.54	95.60	0.0
(Addepalli et al., 2021)	cifar10	ResNet-18	80.24	51.06	94.87	0.0

Continued on next page

Model Confidences & Robustness

Paper	Dataset	Architecture	Adv. Trained Clean Acc.	Adv. Trained Robust Acc.	Norm. Trained Clean Acc.	Norm. Trained Robust Acc.
(Addepalli et al., 2021)	cifar10	WideResNet-34-10	85.32	58.04	95.26	0.0
(Gowal et al., 2021b)	cifar10	WideResNet-70-16	88.74	66.11	87.91	0.0
(Dai et al., 2021)	cifar10	WideResNet-28-10-PSSiLU	87.02	61.55	85.53	0.0
(Gowal et al., 2021b)	cifar10	WideResNet-28-10	87.50	63.44	88.20	0.0
(Gowal et al., 2021b)	cifar10	PreActResNet-18	87.35	58.63	89.01	0.0
(Chen & Lee, 2021)	cifar10	WideResNet-34-10	85.21	56.94	95.64	0.0
(Chen & Lee, 2021)	cifar10	WideResNet-34-20	86.03	57.71	95.29	0.0
(Gowal et al., 2021a)	cifar100	WideResNet-70-16	60.86	30.03	60.56	0.0
(Gowal et al., 2021a)	cifar100	WideResNet-70-16	69.15	36.88	60.56	0.0
(Cui et al., 2021)	cifar100	WideResNet-34-20	62.55	30.20	80.46	0.0
(Cui et al., 2021)	cifar100	WideResNet-34-10	70.25	27.16	79.11	0.0
(Cui et al., 2021)	cifar100	WideResNet-34-10	60.64	29.33	79.11	0.0
(Chen et al., 2021)	cifar100	WideResNet-34-10	62.15	26.94	78.75	0.0
(Wu et al., 2020)	cifar100	WideResNet-34-10	60.38	28.86	78.79	0.0
(Sitawarin et al., 2021)	cifar100	WideResNet-34-10	62.82	24.57	79.11	0.0
(Hendrycks et al., 2019)	cifar100	WideResNet-28-10	59.23	28.42	79.16	0.0
(Rice et al., 2020)	cifar100	PreActResNet-18	53.83	18.95	76.18	0.0
(Rebuffi et al., 2021)	cifar100	WideResNet-70-16	63.56	34.64	60.56	0.0
(Rebuffi et al., 2021)	cifar100	WideResNet-28-10	62.41	32.06	61.46	0.0
(Rade & Moosavi-Dezfooli, 2021)	cifar100	PreActResNet-18	56.87	28.50	63.45	0.0
(Rade & Moosavi-Dezfooli, 2021)	cifar100	PreActResNet-18	61.50	28.88	63.45	0.0
(Addepalli et al., 2021)	cifar100	PreActResNet-18	62.02	27.14	76.66	0.0
(Addepalli et al., 2021)	cifar100	WideResNet-34-10	65.73	30.35	79.11	0.0
(Chen & Lee, 2021)	cifar100	WideResNet-34-10	64.07	30.59	79.11	0.0
(Wong et al., 2020)	imagenet	ResNet-50	55.62	26.24	80.37	0.0
(Engstrom et al., 2019)	imagenet	ResNet-50	62.56	29.22	80.37	0.0
(Salman et al., 2020)	imagenet	ResNet-50	64.02	34.96	80.37	0.0
(Salman et al., 2020)	imagenet	ResNet-18	52.92	25.32	69.74	0.0
(Salman et al., 2020)	imagenet	WideResNet-50-2	68.46	38.14	81.45	0.0