Improving Adversarial Robustness via Adversarial Attacks

Sanghyun Hong
Oregon State University
sanghyun.hong@oregonstate.edu

Shubhankar Joshi
Oregon State University
joshishu@oregonstate.edu

ABSTRACT

The goals of this project were two-fold: first, to demonstrate the potential impact of adversarial attacks like FGSM, PGD, and AutoAttack that add small perturbations to inputs to fool neural network classifiers; and second, to implement and evaluate defenses like adversarial training and the novel pNML technique to improve model robustness. The novelty comes from the use of advanced ResNet and WideResNet architectures as base models, as well as the implementation of pNML defense specifically for image classification tasks. The technical approach taken was to first train models on the MNIST and CIFAR10 datasets, then apply attacks to quantify accuracy drops. Finally, defenses were implemented through adversarial training with PGD examples and the pNML method which perturbs inputs to target each possible label. Key results showed a complete accuracy reduction to 0% on undefended models, with adversarial training recovering 40-50% accuracy and pNML restoring 60-68% accuracy under PGD and FGSM attacks. The main findings confirm that adversarial attacks can easily undermine classifier reliability, but proper defenses can recover substantial performance. Although limitations exist in terms of model and dataset scope, this project clearly demonstrates the potency of attacks as well as viable defenses that together further understanding of adversarial machine learning.

1 INTRODUCTION

Adversarial attacks represent an emerging threat that can completely undermine neural network model performance in ways that are difficult to detect. By applying subtle perturbations to inputs that are imperceptible to humans, adversarial attacks can lead machine learning models to produce grossly incorrect predictions with extremely high confidence. Over recent years, the sophistication and effectiveness of attack techniques have rapidly evolved, with methods like the Fast Gradient Sign Attack (FGSM), iterative Projected Gradient Descent (PGD), and the impressively potent AutoAttack demonstrating the ability to destroy model accuracy entirely. Defending against such attacks remains an open challenge, however the research community has made significant progress in developing countermeasures to restore robustness. Adversarial training augments clean training data with adversarial examples to increase model resilience. More advanced proposals like the adversarial Predictive Normalized Maximum Likelihood (pNML) technique show promise in recovering non-trivial amounts of accuracy post-attack. The proposed Adversarial pNML has a simple methodology. For a given input, it generates adversarial examples targeting every label using fast gradient sign attack. It then collects the probability outputs from the pretrained model for each such refined adversarial input, with each targeted attack's output interpreted as support for that label hypothesis. These hypothesis probabilities are normalized to get a valid distribution, and the label with maximum normalized likelihood is predicted. The normalization methodology is based on the predictive NML framework which compares hypotheses to a reference model.

This project conducts an extensive evaluation of state-of-the-art attacks like FGSM, PGD, and AutoAttack in the context of image classification tasks using benchmark datasets such as MNIST and CIFAR10. Multiple model architectures including ResNets and WideResNets are attacked and then defended with adversarial training and the novel pNML defense. Testing under consistent conditions on established datasets provides an impartial perspective into quantifiable gaps that persist, despite remarkable efforts from both attacks and defenses continuing to push boundaries. The insights from this project aim to contribute empirical clarity between ever-increasing adversarial threats and rapid response through countermeasures intended to uphold model integrity.

RELATED WORK

This project builds on influential recent work that provides insight into adversarial examples, as well as explores defense methodologies against them. Ilyas et al. (2019) provided a vital perspective by arguing that adversarial examples exploit blind spots in models regarding meaningful features of the true data distribution, rather than just being artifacts of model bugs or errors. This motivated research into robustness enhancements through adversarial training that expose models to such blindspots during training. Madry et al. (2018) pioneered this by proposing a min-max adversarial training procedure that trains models to minimize loss under the worst case perturbations generated by multi-step, bounded projected gradient descent attacks. This approach called TRADES substantially improved robustness over prior empirical defense proposals. However, most adversarial training schemes rely on untargeted attacks that try to cause any misclassification.

The key reference paper this project focuses on (Pesso et al., 2022) hypothesizes that incorporating targeted attacks tailored to each class can further expose model vulnerabilities and signal useful patterns to improve robustness. They proposed Adversarial pNML which generates such a diverse set of per-class targeted attacks for each input, collects the probability outputs from the pre-trained model, interprets these as normalized likelihoods for each label hypothesis. It is based on principles from the predictive NML framework and predicts the label with maximum normalized likelihood. Extensive experiments on ImageNet, CIFAR10 and MNIST datasets demonstrate state-of-the-art robustness against white-box and black-box adversarial attacks relative to the adversarial training baseline. Thus, this paper provides a novel perspective and promising approach based on precisely engineered targeted adversarial attacks that can better reveal and address model vulnerabilities to advance robustness that this project seeks to implement, evaluate and extend.

2 APPROACH

The core methodology focused on evaluating adversarial attacks and defenses for image classification models on the CIFAR-10 and MNIST datasets. Models used were ResNet50, ResNet18 and WideResNet-28-10. On CIFAR-10 Resnet 18 was trained for 50 epochs using stochastic gradient descent with cross-entropy loss. Similar WideResNet was also trained. The initial accuracy of the models was above 86 % for the CIFAR 10 dataset. On the MNIST dataset, a pretrained ResNet 50 and Wide ResNet-28-10 were used. The models were trained for 20 epochs and accuracy of more than 98 % was secured for both models. The initial learning rate was 0.001 for all the models.

After obtaining converged models, various adversarial attacks were implemented including FGSM, PGD, and AutoAttack to reduce model accuracy. FGSM applies single-step fixed perturbations while PGD applies iterative perturbations over 10 steps. Larger epsilon values induce greater accuracy drops. . FGSM and PGD attacks apply fixed and iterative epsilons of 0.03 and 0.3 for CIFAR 10 and MNIST datasets respectively. AutoAttack integrates APGD-CE, APGD-DLR, FAB and Square attacks, applying perturbations with random restarts and early stopping on misclassifications. The same epsilon values were used for AutoAttack as well. In this project, we considered only APGD CE and APGD - DLR.

The first defense technique used was adversarial training. Here, PGD adversarial examples were generated by attacking the trained models and blending them with clean samples. This improves robustness to perturbation attacks seen during training. ResNet18 and WideResNet-28-10 were trained on the CIFAR-10 dataset from scratch for 110 and 60 epochs using stochastic gradient descent. The cross-entropy loss was optimized using learning rate scheduling. The initial learning rate for WideResNet was 0.1 and for ResNet 18 was 0.01. Models were trained on NVDIA T4 GPU with batch size 64. Training took 11 and 16 hours for ResNet 18 and WideResNet-28-10 respectively. The clean adversarial accuracy achieved for both models was above 80 % respectively. ResNet 50 and WideResNet-28-10 were trained for 40 epochs for the MNIST dataset. The training time was 2 hours and 5 hours for both models respectively. The accuracy achieved after training was more than 90 % and learning rate chosen was 0.1 and 0.01 for WideResNet 28-10 and ResNet 50 respectively.

The second defense was predictive normalized maximum likelihood (pNML). This works by perturbing every input towards each possible label using gradient sign perturbations scaled by a lambda hyperparameter. Well-tuned lambda values were used for both datasets respectively. For CIFAR10, different lambda values from 0.01 to 0.03 were implemented and for MNIST 0.1 to 0.3 were implemented. The perturbed inputs are classified by the model and the label with maximum predicted probability is chosen, providing prediction robustness. The pNML defense for each lambda value took around 1 hour for the CIFAR 10 dataset and 40 minutes MNIST dataset with clean examples. And when it was tested with PGD examples it took 6 and 4 hours for CIFAR 10 and MNIST datasets respectively.

The PGD attack creates adversarial examples by iteratively applying small perturbations to move the input away from the correct class. These PGD adversarial examples were then passed into the pNML defense to evaluate its accuracy under perturbation attacks. Similar actions were done with the FGSM attack as well.

3 EXPERIMENTAL RESULTS

Attack Evaluation

The work evaluates model robustness against FGSM, PGD, and AutoAttack. For the MNIST dataset, the ResNet50 model sees its accuracy drop from 99% on clean data to 0% under the AutoAttack perturbations, indicating complete failure. WideResNet-28-10 retains some robustness under AutoAttack with 12-13% accuracy. However, under the stronger PGD attack with ε=0.3 over 10 steps, the accuracy drops more significantly to 4.6-7.4%.

	Model	**CLEAN**	**APGD CE**	**APGD D**	**PGD**	**FGSM**
MNIST	ResNeT50	99 %	0 %	0 %	22.79 %	30 .56 %
	Wide ResNet 28-10	99.10 %	0 %	0 %	4.61 %	7.44 %
CIFAR10	ResNeT18	88.9%	0.0 %	0.0 %	0.0 %	3.89 %
	Wide ResNet 28-10	89.39 %	12.45 %	12.19 %	0.04 %	22.5 %

On the more complex CIFAR10 dataset, all models see essentially 0% accuracy under AutoAttack. This highlights a lack of intrinsic robustness generalizing beyond the MNIST digits dataset. Under the PGD and AutoAttack with ε=0.03, the accuracy of ResNet18 and WideResNet drop essentially to 0% robustness across architectures.

Defense Evaluation

Applying adversarial training improves resilience for both MNIST and CIFAR10 datasets. On CIFAR10 under the PGD ε=0.03 attack and niter = 7, accuracy increases from 0% to 40-45% for ResNet18 and WideResNet. This demonstrates adversarial training can recover significant robustness, although still far from clean accuracy. For MNIST, under the stronger PGD ε=0.3 attack, accuracy rises further to 68-95% across models. The ResNet50 architecture reaches over 90% accuracy, indicating adversarial training can confer substantial robustness even under such tailored attacks.

	Model	**CLEAN**	**APGD CE**	**APGD D**	**PGD (0.3), 10**	**FGSM**
MNIST	ResNeT50	98.80 %	71.05 %	68.10 %	93.25 %	95.32 %
	Wide ResNet 28-10	98.58 %	83.74 %	82. 66 %	94.24 %	96.17 %
CIFAR10	ResNeT18	80.5 %	41.70 %	39.84 %	43.5 %	51.37 %
	Wide ResNet 28-10	81.8 %	44.72 %	42.92 %	45.7 %	53.98 %

On CIFAR10 Clean data, pNML slightly reduces accuracy to 79-81% from 80%. But under PGD and FGSM attacks it increases robustness significantly, achieving 62-68% and 64-69% accuracy respectively with ε=0.03. This is comparable to adversarial training results. For MNIST, under the stronger ε=0.3 attacks, pNML confers even greater robustness with 95% accuracy on ResNet50, and 95-96% on WideResNet. This indicates pNML can be an effective defense method compared to adversarial training.

	Model	**CLEAN**	**PGD (0.3), 10**	**FGSM**
MNIST	ResNeT50	98.80 %	93.25 %	95.32 %
	Wide ResNet 28-10	98.58 %	94.24 %	96.17 %
CIFAR10	ResNeT18	79.8 %	62 %	64.3 %
	Wide ResNet 28-10	81.2 %	66.75 %	68.7 %

4 Discussion

Future work could expand these models to more complex datasets. Testing was limited to MNIST and CIFAR-10, lacking adversarial feature distributions and diversity to conclusively evaluate real-world viability. Analyzing performance on large-scale datasets like ImageNet or MS-COCO with additional perturbation types would provide meaningful stress testing

In terms of defenses, both adversarial training and pNML improved robustness but recovered accuracy only to 40-68% under key attacks. This leaves considerable room for enhancement. Exploring augmented training with more perturbation types could help, as could tuning lambda, training epochs and ratios in a broader hyperparameter search. Finally, applying an ensemble combining pNML, adversarial training and other defenses could provide multiplicative benefits in robustness.

5 Conclusions

This work demonstrates that standard image classifiers are highly vulnerable to adversarial attacks - with accuracy dropping to 0% under perturbations from FGSM, PGD and AutoAttack. Defenses like adversarial training and predictive normalized maximum likelihood (pNML) can recover some accuracy but significant gaps remain. The methodology provides a blueprint for holistically evaluating model security against various attack types. Core lessons learned are the tradeoffs between accuracy and robustness in vision models, the need for proactive security evaluations, and the promise of emerging defenses. These findings advance the understanding of reliability gaps and lay the groundwork for developing resilient machine-learning models.

References

1. Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., & Madry, A. (2019, August 12). Adversarial examples are not bugs, they are features. arXiv.org. https://arxiv.org/abs/1905.02175
2. Pesso, Uriya & Bibas, Koby & Feder, Meir. (2021). Utilizing Adversarial Targeted Attacks to Boost Adversarial Robustness.
3. Croce, F., & Hein, M. (2020). Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. *arXiv (Cornell University)*. http://arxiv.org/pdf/2003.01690.pdf