ACCURACY OF WHITE BOX AND BLACK BOX ADVERSARIAL ATTACKS ON A SIGN ACTIVATION 01 LOSS NEURAL NETWORK ENSEMBLE

Anonymous authors
Paper under double-blind review

ABSTRACT

In this work we ask the question: is an ensemble of single hidden layer sign activation 01 loss networks more robust to white box and black box adversarial attacks than an ensemble of its differentiable counterpart of cross-entropy loss with relu activations and an ensemble of the approximate differentiable counterpart of cross-entropy loss with sign activations? We consider a simple experimental setting of attacking models trained for binary classification on pairwise CIFAR10 datasets - altogether a total of 45 datasets. We study ensembles of **bcebp**: binary cross-entropy loss with relu activations trained with back-propagation, **bceban**: binary cross-entropy loss with sign activations trained with back-propagation with the straight through estimator gradient, **01scd**: 01-loss with sign activations trained with gradient-free stochastic coordinate descent, and **bcescd**: binary cross-entropy loss with relu activation trained with gradient-free stochastic coordinate descent (to isolate the effect of 01 loss from gradient-free training). Our four models have mean test accuracies of 87%, 85%, 83%, and 87% on pairwise CIFAR10 datasets but under PGD white-box attacks they each drop to 0.2%, 0.1%, 31%, and 0.2% - the 01 loss network ensemble has a drop of 4% in clean test accuracy but a gain of 31% in white box attack accuracy. Even training with the gradient-free stochastic coordinate descent can be attacked thus suggesting that the defense lies in 01 loss. In a transfer black-box attack we find adversaries produced from the bcebp model also lower the accuracy of bceban to 0.6% but 01scd to only 59% - we see similar accuracies if we transfer from bceban to bcebp and 01scd. However, if we transfer from 01scd to bcebp and bceban their accuracies drop to only 79% and 75%. While our results are from from those of multi-class and convolutional networks it suggests that 01 loss are hard to attack naturally without adversarial training. All models, data, and code to reproduce results here are available from `https://github.com/xyzacademic/mlp01example`.

1 INTRODUCTION

Adversarial attacks remain a security vulnerability in neural networks today since they were first discovered Szegedy et al. (2014). A recent paper evaluating the operational feasibility of adversarial attacks in military defense Zhang et al. (2022) found that patch attacks posed a minimal danger in practice but both white and black-box attacks can be significantly more effective - a white box attack reduces the target model's accuracy by 65% and a black-box attack reduces it by 55% to 63% depending upon knowledge and access to training data and target model architecture. Various defenses have been developed and broken Athalye et al. (2018). Adversarial training Kurakin et al. (2016), which is to train the model with clean and adversarial data, remains the most effective solution to date but it is computationally expensive and lowers clean test accuracy Raghunathan et al. (2019); Clarysse et al. (2022).

Based on previous work Xie et al. (2019); Xue et al. (2020b;a); Yang et al. (2020) and intuition we hypothesize that a 01-loss neural network model would be hard to attack. With a new stochastic coordinate descent algorithm Xie et al. (2019) we were able to train a single hidden layer 01-loss neural network with accuracy and runtimes comparable to its differentiable and cross-entropy loss counterparts (see Table 1 below).

Table 1: Accuracy of our models on clean test data averaged across all 45 pairs of classes in CIFAR10. Each model is an ensemble of 8 single ones each with a different random number generator seed.

	Accuracy (mean of all 45 pairs)				**Runtime to train one pair (seconds)**			
Data	bcebp	bceban	01scd	bcescd	bcebp	bceban	01scd	bcescd
Train	94.2	94.88	88.6	92.4	40	46	64	56
Test	87.2	85.42	83.2	86.8	N/A	N/A	N/A	N/A

2 MAIN RESULT

In a white box attack the adversary has full access to the target model complete with all architecture details and model weights. Both FGSM Goodfellow et al. (2015) and PGD Madry et al. (2017) are popular attack methods. In FGSM the adversary uses the target model gradient once ($x' = x + \epsilon sign(\nabla_x \mathcal{L}(\theta, x, y))$) whereas in PGD this is performed iteratively ($x^{t+1} = x^t + \alpha \text{sign}(\nabla_x \mathcal{L}(\theta, x^t, y))$). Thus PGD can be more effective but may also be detected due to multiple queries. We perform 20 steps of PGD with standard parameter settings $\alpha = 2/256$ and $\epsilon = 16/256$.

Since both FGSM and PGD need the target model gradient we approximate the 01 loss gradient with the cross-entropy loss gradient and use the straight through estimator Courbariaux et al. (2016) for sign activation gradient: $\frac{df}{dx} = 1$ if $|x| \leq \sum_{x' \in X} x'$ and 0 if $|x| > \sum_{x' \in X} x'$. This is the same estimator to train the cross-entropy loss network with sign activation (also known as binary neural network) with back propagation. For each image in each test dataset we produce an adversary with the target model's gradient (averaged across each model in the ensemble) and then evaluate its accuracy on the target model.

In Table 2 we see that both binary cross-entropy models bcebp and bceban trained with back propagation have adversarial accuracy almost 0%. Even bcescd, which is the same model as bcebp but trained with gradient free stochastic coordinate descent, can be attacked (see accuracies of 9.7% and 0.2% in FGSM and PGD attacks below). Thus we conjecture that the 01 loss is the main defense that makes 01scd harder to attack.

Table 2: FGSM and PGD white box accuracy of our models averaged across all 45 pairs of classes in CIFAR10. Each model is an ensemble of 8 single ones each with a different random number generator seed.

FGSM white box accuracy				**PGD white box accuracy**			
bcebp	bceban	01scd	bcescd	bcebp	bceban	01scd	bcescd
0.8%	1.4%	44.2%	9.7%	0.2%	0.1%	31.4%	0.2%

In our black box attacks we produce adversaries targeting a "source" model and evaluate them on target models. This simulates a scenario where the attacker does not have the target model complete with weights but has the same training dataset and same architecture as the target model. In Table 3 we see that adversaries sourced from bcebp and bceban are also effective on each other but much less so on 01scd.

Table 3: PGD black box accuracy of our models averaged across all 45 pairs of classes in CIFAR10

Source	**bcebp**		**bceban**		**01scd**	
Target	bceban	01scd	bcebp	01scd	bcebp	bceban
	0.6%	59.3%	0.9%	60.7%	78.9%	75%

URM STATEMENT

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2023 Tiny Papers Track.

REFERENCES

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

Jacob Clarysse, Julia Hörmann, and Fanny Yang. Why adversarial training can hurt robust accuracy. *arXiv preprint arXiv:2203.02006*, 2022.

Matthieu Courbariaux, Itay Hubara, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1. *arXiv preprint arXiv:1602.02830*, 2016.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. URL `http://arxiv.org/abs/1412.6572`.

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

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C Duchi, and Percy Liang. Adversarial training can hurt generalization. In *Identifying and Understanding Deep Learning Phenomena ICML Workshop*, 2019.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. URL `http://arxiv.org/abs/1312.6199`.

Meiyan Xie, Yunzhe Xue, and Usman Roshan. Stochastic coordinate descent for 0/1 loss and its sensitivity to adversarial attacks. In *Proceedings of 18th IEEE International Conference on Machine Learning and Applications - ICMLA 2019*, pp. to appear, 2019.

Yunzhe Xue, Meiyan Xie, and Usman Roshan. Towards adversarial robustness with 01 loss neural networks. In *IEEE International Conference on Machine Learning and Applications*, 2020a.

Yunzhe Xue, Meiyan Xie, and Usman Roshan. On the transferability of adversarial examples between convex and 01 loss models. In *IEEE International Conference on Machine Learning and Applications*, 2020b.

Zhibo Yang, Yanan Yang, Yunzhe Xue, Frank Y Shih, Justin Ady, and Usman Roshan. Accurate and adversarially robust classification of medical images and ecg time-series with gradient-free trained sign activation neural networks. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 2456–2460. IEEE, 2020.

Li Ang Zhang, Gavin S. Hartnett, Jair Aguirre, Andrew J. Lohn, Inez Khan, Marissa Herron, and Caolionn O'Connell. *Operational Feasibility of Adversarial Attacks Against Artificial Intelligence*. RAND Corporation, Santa Monica, CA, 2022. doi: 10.7249/RRA866-1.