Momentum Iterative Gradient Sign Method Outperforms PGD Attacks

Sreenivasan Mohandas[1], Naresh Manwani[1] and Durga Prasad Dhulipudi[2]
[1]*Machine Learning Lab, International Institute of Information Technology, Hyderabad, India*
[2]*Lab for Spatial Informatics, International Institute of Information Technology, Hyderabad, India*

Keywords: Adversarial Attack, Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD) Method, Deep Learning, Training Time.

Abstract: Adversarial examples are machine learning model inputs that an attacker has purposefully constructed to cause the model to make a mistake. A recent line of work focused on making adversarial training computationally efficient for deep learning models. Projected Gradient Descent (PGD) and Fast Gradient Sign Method (FGSM) are popular current techniques for generating adversarial examples efficiently. There is a tradeoff between these two in terms of robustness or training time. Among the adversarial defense techniques, adversarial training with the PGD is considered one of the most effective ways to achieve moderate adversarial robustness. However, PGD requires too much training time since it takes multiple iterations to generate perturbations. On the other hand, adversarial training with the FGSM takes much less training time since it takes one step to generate perturbations but fails to increase adversarial robustness. Our algorithm achieves better robustness to PGD adversarial training on CIFAR-10/100 datasets and is faster than PGD string adversarial training methods.

1 INTRODUCTION

Adversarial examples are machine learning model inputs that an attacker has purposefully constructed to cause the model to make a mistake. A good adversarial example from MIT is the one that produced the 3D-printed turtle that, when viewed from almost any angle, modern ImageNet trained image classifiers misclassify as a rifle (Athalye et al., 2018) . Other research has found that making imperceptible changes to an image can fool a medical imaging system into correctly classifying a benign mole as malignant with 100 percent certainty (Finlayson et al., 2019) and that a few pieces of tape can fool a computer vision system into incorrectly classifying a stop sign as a speed limit sign (Eykholt et al., 2018). One of the few defenses against adversarial attacks that can resist powerful attacks is 'adversarial training', a network is explicitly trained on hostile samples. Unfortunately, traditional adversarial training on a large-scale dataset like ImageNet is impracticable due to the enormous expense of providing solid adversarial samples. A recent line of work focused on making adversarial training computationally efficient for deep learning models. Projected Gradient Descent (PGD) and Fast Gradient Sign Method (FGSM) are popular current techniques for generating adversarial examples efficiently. There is a tradeoff between these two in terms of robustness or training time.

Among these defense techniques, adversarial training with the PGD is considered one of the most effective ways to achieve moderate adversarial robustness. However, PGD requires too much training time since it takes multiple iterations to generate perturbations. On the other hand, adversarial training with the FGSM takes much less training time since it takes one step to generate perturbations but fails to increase adversarial robustness. Eric Wong (Wong et al., 2020) showed that adversarial training with the FGSM, when combined with random initialization, is as effective as PGD-based training with the lower computation time cost. This paper proposes our method, "Momentum Iterative gradient sign methods with Reuse of Perturbations (MIRP)." These perturbations introduced between epochs performs better than PGD, which was previously believed to be ineffective, rendering the method no more costly than standard training in practice. Our algorithm achieves better robustness to PGD adversarial training on CIFAR-10, CIFAR-100 datasets and is faster than PGD string adversarial training methods.

Deep neural networks (DNNs) have been recently achieving state-of-the-art performance on image classification datasets such as ImageNet. For example,

Mohandas, S., Manwani, N. and Dhulipudi, D.
Momentum Iterative Gradient Sign Method Outperforms PGD Attacks.
DOI: 10.5220/0010938400003116
In *Proceedings of the 14th International Conference on Agents and Artificial Intelligence (ICAART 2022) - Volume 3*, pages 913-916
ISBN: 978-989-758-547-0; ISSN: 2184-433X

He et al.(He et al., 2016) have achieved 96.43 % top-5 test accuracy with their ResNet architecture, which is difficult to exceed for humans. (Szegedy et al., 2014) discovered it is possible for a network to misclassify an image by adding a very small perturbation to it which humans will not notice the difference. These perturbed inputs are termed as adversarial examples. The existence of adversarial examples poses a threat to the security of systems incorporating machine learning models. For example, autonomous cars could be forced to crash, intruders could confound face recognition software or SPAM filters and other filters can be bypassed. Also, adversarial examples can often be generated not only with access to the model's architecture and parameters, but also without it. Therefore, it is important to understand the weakness of our models and come up with ways to defend them against potential adversaries. One popular approximation method successfully used in adversarial training is the PGD attack as it remains empirically robust to this day.

2 BACKGROUND

In this section, we provide the background knowledge as well as review the related works about adversarial attack and defense methods. Given a classifier f(x): x ε X $\rightarrow$ y ε Y that outputs a label y as the prediction for an input x, the goal of adversarial attacks is to seek an example x^* in the vicinity of x but is misclassified by the classifier. Specifically, there are two classes of adversarial examples - non-targeted and targeted ones. For a correctly satisfied input x with ground-truth label y such that f(x) = y, a non-targeted adversarial example x^* is crafted by adding small noise to x without changing the label, but misleads the classifier as $f(x^*)$!=y; and a targeted adversarial example aims to fool the classifier by outputting a specific label as $f(x^*) \neq y^*$, where y^* is the target label specified by the adversary, and $y^* \neq y$. In most cases, the Lp norm of the adversarial noise is required to be less than an allowed value as $\|x* - x\|_p \leq \varepsilon$, where p could be 0, 1, 2...Below text provides details of FGSM and PGD adversarial attacks.

Fast Gradient Sign Method (FGSM): FGSM perturbs clean examples x for one step by the amount of ε along the input gradient direction (Goodfellow et al., 2015):

$x_{adv} = x + \varepsilon \, . \, \text{sign}(\nabla_x l(h(x),y))$.

Project Gradient Descent (PGD): PGD (Goodfellow et al., 2015) perturbs a clean example x for a number of T steps with smaller step size. After each step of perturbation, PGD projects the adversarial example back onto the ε - ball of x, if it goes beyond:

$x^i = (x^{i-1} + \alpha \, . \, \text{sign}(\nabla_x l(h(x^{i-1}),y))$

where α is the step size, and x^i is the adversarial example at the i-th step (x^0 = x). The step size is usually set to $\varepsilon/T \leq \alpha < \varepsilon$ for overall T steps of perturbations.

Neural networks are trained with first-order gradient descent algorithms. There exist two families of gradient descent algorithms: accelerated stochastic gradient descent (SGD) and adaptive learning rate. SGD based algorithms include Momentum and Nesterov based methods. Adaptive learning rate include Adam, Adadelta, Adabelief. DNN's trained with SGD based algorithms has a strong generalization ability with low convergence rate which is vice-versa in adaptive family based algorithms. In general, solution based on these optimization families are considered for improving the transferability of adversarial examples.

Among many attempts (Metzen et al., 2017; Dong et al., 2017; Pang et al., 2017; Kurakin et al., 2017; Li and Gal, 2017; Tramèr et al., 2018; Papernot et al., 2016) , adversarial training is the most extensively investigated way to increase the robustness of DNNs (Kurakin et al., 2017; Tramèr et al., 2018; Goodfellow et al., 2015). However, running a strong PGD adversary within an inner loop of training is expensive, and some earlier work in this topic found that taking large but fewer steps did not always significantly change the resulting robustness of a network (Wang, 2019). Thus, an inherent tradeoff appears between computationally efficient approaches which aim at solving the optimum perturbation value in few iterations as possible and approaches which aim at solving the same problem more accurately but with more iterations. To be more specific, the PGD attack uses multiple steps of projected gradient descent (PGD), which is accurate but computationally expensive where as Fast Sign Gradient Method (FGSM) uses only one iteration of gradient descent which is computationally efficient.

To combat the increased computational overhead of the PGD defense, some recent work has looked at 'Fast and Free Adversarial training' where modified FGSM adversarial training achieves a accuracy closure to the model trained with PGD attacks. These include improvements include top performing training methods from DAWNBench competition (Coleman et al., 2017) are able to train CIFAR-10 and CIFAR-100 architectures to standard benchmark metrics in mere minutes, using only a modest amount of computational resources. Although some of the techniques can be quite problem specific for achieving bleeding edge performance, more general techniques such as cyclic learning rates (Smith and Topin, 2019)

and half-precision computations (Micikevicius et al., 2018) have been quite successful in the top ranking submissions, and can also be useful for adversarial training.

3 MIRP

With the Fast Adversarial training (Wong et al., 2020), updated FGSM training combined with random initialization is as effective as defense with PGD-based training. The key difference here is to use perturbation from the previous iteration as the initial starting point for the next iteration i.e., starting with the previous minibatch's perturbation or initializing from a uniformly random perturbation allow FGSM adversarial training to succeed at being robust to full-strength PGD adversarial attacks. FGSM with random initialization algorithm is shown below:

Algorithm 1: FGSM with random initialization adversarial training for R epochs, given some radius ε, step size α, and a dataset of size M for a network f_θ.

for $t = 1...R$ **do**
  for $i = 1...M$ **do**
    δ = Uniform (-δ,δ)
    $\delta = \delta + (\alpha * sign(\nabla_\delta l(f_\theta(x_i + \delta); y_i))$
    $\delta = max(min(\delta; \varepsilon); -\varepsilon)$
    $\theta = \theta - (\nabla_\delta l(f_\theta(x_i + \delta); y_i))$
  end for
end for

Also, with Boosting algorithm (Dong et al., 2018) (which doesn't use sign(.) method) achieves better performance than FGSM but with PGD attacks its accuracy drops to 0. On the above basis, we propose a new algorithm to use Momentum iterative gradient sign adversarial training with random initialization for the perturbation as shown below.

Algorithm 2: MIRP adversarial training for R iterations, T epochs, given some radius ε, N PGD steps, step size α, and a dataset of size M for a network f_θ.

for $t = 1...T$ **do**
  for $i = 1...M$ **do**
    for $j = 1...R$ **do**
      $m_{j+1} = \mu * m_j + \frac{(\nabla_\delta l(f_\theta(x_i+\delta); y_i))}{\|(\nabla_\delta l(f_\theta(x_i+\delta); y_i))\|_{L_1}}$
      $\delta = \delta + (\alpha * sign(m_{j+1}))$
      $\delta = max(min(\delta; \varepsilon); -\varepsilon)$
    end for
    $\theta = \theta - (\nabla_\delta l(f_\theta(x_i + \delta); y_i))$
  end for
end for

3.1 Tables

Below tables shows the test accuracy and the training time of the different models tested on different datasets.

Table 1: Standard and robust performances of various adversarial training methods on CIFAR-10 dataset

Method	PGD / Time
MIRP	50.59 / 47 min
Std Momentum Iterative	0 / 47.12 min
FGSM with random initialization	47.53 / 24.6 min
PGD	49.75 / 89.6 min

Table 2: Standard and robust performances of various adversarial training methods on CIFAR-100 dataset.

Method	PGD / Time
MIRP	28.22 / 44.55 min
Std Momentum Iterative	0 / 44.16 min
FGSM with random initialization	24.88 / 23.19 min
PGD	27.36 / 92.16 min

Here, Standard Momentum Iterative method is based on (Dong et al., 2018), FGSM with random initialization is based on (Wong et al., 2020). We used same configuration settings like step size, cyclic learning rate and mixed precision arithmetic as per Eric Wong (Wong et al., 2020). All experiments using MIRP are carried out with random starting points and step size alpha = 1.25*epsilon. All PGD adversaries used at evaluation are run with 10 random restarts for 50 iterations with ε = 8/255. Speedup with mixed-precision was incorporated with the Apex amp package at the O2 optimization level without loss scaling for CIFAR-10/100 experiments. All experiments are tested on Nvidia T4 machine.

Reuse between Epochs: We tested multiple combinations between momentum iterative gradient values and perturbation and below are the inferences drawn:

Table 3: Performance of MIRP on CIFAR-10 dataset with varying m and δ.

m , δ	PGD / Time
Non-zero, Non-zero	50.59 / 47 min
Non-zero, Zero	0 / 47.5 min
Zero , Non-zero	50.39 / 45.8 min
Zero , Zero	0 / 47.12 min

Decaying Factor (μ): Tested various values of Decaying factor from 0.1 to 0.6 and could see that model starts overfitting for any value beyond 0.1 as shown below:

Table 4: Performance of MIRP on CIFAR-10 dataset with varying decaying factor.

Decaying factor	PGD / Time
0.1	50.59 / 47 min
0.6	49.19 / 50.8 min

Here, Non-zero refers that the values are retained between batches and Zero refers to that the values are initialized to zero between batches.

4 CONCLUSIONS

Our findings show that MIRP adversarial training, when used with random initialization, can in fact be more effective as the more costly PGD adversarial training. As a result, we are able to learn adversarially robust classifiers for CIFAR10/100 in minutes. We believe that leveraging these significant reductions in time to train robust models will allow future work to iterate even faster, and accelerate research in learning models which are resistant to adversarial attacks.

REFERENCES

Athalye, A., Engstrom, L., Ilyas, A., and Kwok, K. (2018). Synthesizing robust adversarial examples.

Coleman, C. A., Narayanan, D., Kang, D., Zhao, T., Zhang, J., Nardi, L., Bailis, P., Olukotun, K., Ré, C., and Zaharia, M. A. (2017). Dawnbench : An end-to-end deep learning benchmark and competition.

Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., and Li, J. (2018). Boosting adversarial attacks with momentum. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9185–9193.

Dong, Y., Su, H., Zhu, J., and Bao, F. (2017). Towards interpretable deep neural networks by leveraging adversarial examples. *ArXiv*, abs/1708.05493.

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., and Song, D. (2018). Robust physical-world attacks on deep learning models.

Finlayson, S. G., Bowers, J., Ito, J., Zittrain, J., Beam, A., and Kohane, I. S. (2019). Adversarial attacks on medical machine learning. *Science*, 363:1287 – 1289.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Kurakin, A., Goodfellow, I. J., and Bengio, S. (2017). Adversarial machine learning at scale. *ArXiv*, abs/1611.01236.

Li, Y. and Gal, Y. (2017). Dropout inference in bayesian neural networks with alpha-divergences. In *ICML*.

Metzen, J. H., Genewein, T., Fischer, V., and Bischoff, B. (2017). On detecting adversarial perturbations. *ArXiv*, abs/1702.04267.

Micikevicius, P., Narang, S., Alben, J., Diamos, G. F., Elsen, E., García, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., and Wu, H. (2018). Mixed precision training. *ArXiv*, abs/1710.03740.

Pang, T., Du, C., and Zhu, J. (2017). Robust deep learning via reverse cross-entropy training and thresholding test. *ArXiv*, abs/1706.00633.

Papernot, N., Mcdaniel, P., Wu, X., Jha, S., and Swami, A. (2016). Distillation as a defense to adversarial perturbations against deep neural networks. *2016 IEEE Symposium on Security and Privacy (SP)*, pages 582–597.

Smith, L. N. and Topin, N. (2019). Super-convergence: very fast training of neural networks using large learning rates. In *Defense + Commercial Sensing*.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., and Fergus, R. (2014). Intriguing properties of neural networks. *CoRR*, abs/1312.6199.

Tramèr, F., Kurakin, A., Papernot, N., Boneh, D., and Mcdaniel, P. (2018). Ensemble adversarial training: Attacks and defenses. *ArXiv*, abs/1705.07204.

Wang, J. (2019). Bilateral adversarial training: Towards fast training of more robust models against adversarial attacks. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6628–6637.

Wong, E., Rice, L., and Kolter, J. Z. (2020). Fast is better than free: Revisiting adversarial training.