WITCHcraft: Efficient PGD Attacks with Random Step Size

Ping-Yeh Chiang,[1] Jonas Geiping,[2]
Micah Goldblum,[1] Tom Goldstein[1] Renkun Ni,[1]
Steven Reich,[1] Ali Shafahi[1]

[1]University of Maryland, College Park, Maryland, USA
`tomg@cs.umd.edu`

[2]University of Siegen, Siegen, Germany
`jonas.geiping@uni-siegen.de`

April 17, 2020

Abstract

- Adversarial attacks use many steps and random restarts.
- Attacks saturate and explore image space inefficiently.
- Introduce adversarial attacks with coordinate-wise random step size.
- Better performance at a lower cost.

Adversarial Examples

- Adversarial attacks are small perturbations to inputs which cause pathological model behavior.
- Maximize loss w.r.t. inputs subject to constraints.

Original	ℓ_2-norm=10	ℓ_∞-norm=0.05	ℓ_0-norm=5000

egyptian cat (28%)	traffic light (97%)	traffic light (96%)	traffic light (80%)

Figure 1: Adversarial attacks against ResNet50 on ImageNet. ImageNet images have dimensions $224 \times 224 \times 3$ with pixel values between 0 and 1.

Adversarial Examples

- FGSM: $\delta = \epsilon\,\mathrm{sign}[\nabla_\delta \mathcal{L}(f_\theta(\mathbf{x}+\delta), y)]$ [GSS14]
- PGD attack: $\delta \leftarrow \pi_\epsilon[\delta + \alpha\,\mathrm{sign}[\nabla_\delta \mathcal{L}(f_\theta(\mathbf{x}+\delta), y)]]$, where π_ϵ denotes projection onto the ℓ_∞-ball of radius ϵ [Mad+18]
- Restarts from random initializations
- Targeted vs. untargeted
- Whitebox vs. blackbox

Injecting Randomness into Optimization

Adversarial attacks are a difficult nonconvex optimization, likely stuck in bad local minima.
Randomness is key to mitigate bad local minima:

- Stochastic optimization algorithms select data points at random.
- Stochastic preconditioners draw randomized preconditioning operators.
- Many iterative algorithms restart from random starting points.

WITCHcraft: Efficient Adversarial Attacks

- Combine the PGD attack with a randomly chosen coordinate-wise step size.
- Random step size is chosen independently for each entry in the gradient so that different pixels are perturbed different amounts with each iteration.
- WITCHcraft still incorporates a random initialization, which comes at no cost.
- Terminate the algorithm as soon as the attack fools the classifier.

WITCHcraft: Efficient Adversarial Attacks

Algorithm 1: The WITCHcraft attack algorithm.

Requires: Network f, input $\mathbf{x}$, label y, permissible perturbation set S, number of steps n, and expected step size parameter a.

Initialize perturbation δ with entries distributed independently according to distribution $\mathcal{U}(S)$.

for step = 1,..., n **do**

Sample τ with entries distributed independently according to distribution $\mathcal{U}(0, 2a)$.

$\delta \leftarrow \Pi_S[\delta + \tau \odot \text{sign}(\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x} + \delta, \text{class})$

If $\arg\max(f(\mathbf{x} + \delta)) \neq y$, return $\mathbf{x} + \delta$ and **break**.

Experimental Results

- Evaluate WITCHcraft on a canonical task: attacking the adversarially robust models introduced in [Mad+18] for CIFAR-10 and MNIST classification
- WideResNet(34-10) [ZK16] used for CIFAR-10
- CNN with 2 convolutional layers used for MNIST
- Both models adversarially trained using 7-step PGD
- Perturbations on CIFAR-10 and MNIST images are restricted to ℓ_∞-balls of radius 0.031 and 0.3, respectively.

Comparison to PGD Benchmarks

- Hyperparameters chosen to mirror those used for PGD attacks on the leaderboards [Mad19a] [Mad19b]
- On CIFAR-10, 20- and 100-step WITCHcraft beat equivalent PGD attacks (Table 1).
- On MNIST, 100-step WITCHcraft beat both 100- and 500-step PGD (Table 2).

Comparison to PGD Benchmarks

Attack	CIFAR-10 $\mathcal{A}_{adv}$
20-step PGD	47.04%
20-step WITCHcraft	**45.92%**
100-step PGD	45.29%
100-step WITCHcraft	**45.20%**
20-PGD w/ 10 restarts	45.21%

Table 1: Robust accuracy, $\mathcal{A}_{adv}$, of various adversarial attacks against the WideResNet(34-10) model trained on CIFAR-10, and released by the authors of [Mad+18]. Bolded entries indicate best attack results across fixed computational complexity. Randomized coordinate-wise learning rates (WITCHcraft) improve attack effectiveness with a fixed computational budget.

Comparison to PGD Benchmarks

Attack	MNIST $\mathcal{A}_{adv}$
100-step PGD	92.52%
100-step WITCHcraft	**91.68%**
500-step PGD	91.91%
500-step WITCHcraft	**91.00%**

Table 2: Robust accuracy, $\mathcal{A}_{adv}$, of various adversarial attacks against the two-layer CNN model trained on MNIST and released by the authors of [Mad+18]. Bolded entries indicate the best attack results across fixed computational complexity. Like we observed for the CIFAR-10 model, randomized coordinate-wise learning rates improve attack effectiveness with a fixed computational budget.

The Effect of Step Size

- How does expected step size affect WITCHcraft and PGD?
- Compare performance of both methods over a range of step sizes
- On CIFAR-10, our method is somewhat less sensitive to this parameter, and generally performs better than PGD (Figure 2).
- On MNIST, neither method appears very sensitive, but note that each accuracy result from our method beats every PGD result over this range (Figure 3).

The Effect of Step Size

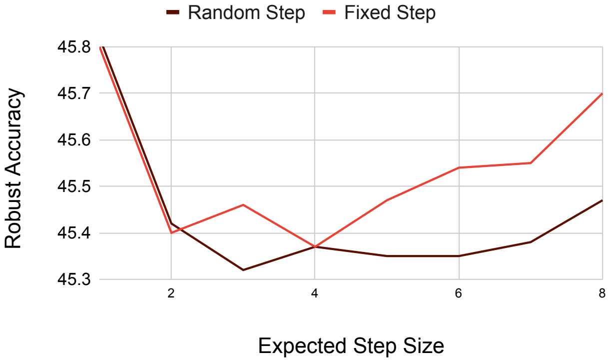

Figure 2: Sensitivity plot of a 40-step PGD attack compared with 40-step WITCHcraft for the CIFAR-10 challenge. We see that the randomized step size choice outperforms a deterministic step size choice, particularly when larger step sizes are used.

The Effect of Step Size

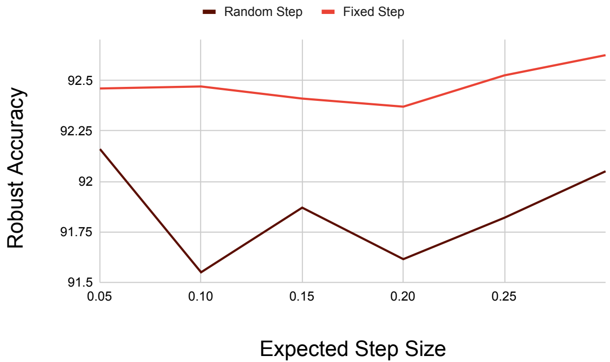

Figure 3: Sensitivity plot of a 40-step PGD attack compared with 40-step WITCHcraft. As we observed above for CIFAR-10, we see that randomized step sizes result in more effective attacks against robust MNIST classifiers.

Additional Attack Steps

- Examine how quickly the success rates of WITCHcraft and PGD saturate as the number of attack steps increases.
- For both tasks, WITCHcraft suffers less from diminishing returns (Figures 4, 5).
- We hypothesize that this is the result of randomness improving the exploratory power of the attack – the stochastic step size of WITCHcraft seems to better escape local minima.

Additional Attack Steps

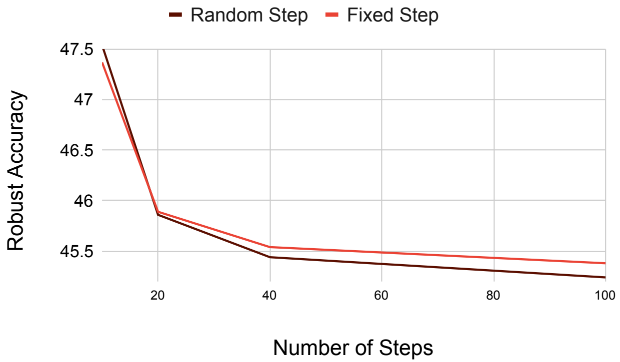

Figure 4: Comparison of robust accuracy as we increase the number of attack steps for WITCHcraft vs. PGD on CIFAR-10. Each reported robust accuracy is an average of 8 trials. As the number of steps increases, WITCHcraft outperforms PGD by a progressively wider margin.

Additional Attack Steps

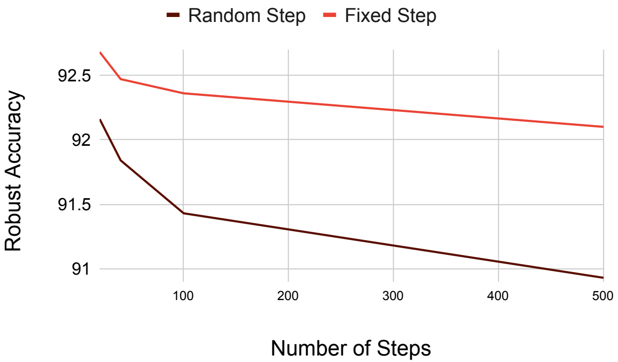

Figure 5: Comparison of robust accuracy as we increase the number of attack steps for WITCHcraft vs. PGD on MNIST. Each reported robust accuracy is an average of 6 trials. As the number of steps increases, WITCHcraft outperforms PGD by a progressively wider margin.

References I

[GSS14] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples". In: *arXiv preprint arXiv:1412.6572* (2014).

[Mad19a] Aleksander Madry. *CIFAR10 Adversarial Examples Challenge.* `https://github.com/MadryLab/cifar10_challenge`. 2019.

[Mad19b] Aleksander Madry. *MNIST Adversarial Examples Challenge.* `https://github.com/MadryLab/mnist_challenge`. 2019.

[Mad+18] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. "Towards Deep Learning Models Resistant to Adversarial Attacks". In: *International Conference on Learning Representations*. 2018. URL: `https://openreview.net/forum?id=rJzIBfZAb`.

[ZK16] Sergey Zagoruyko and Nikos Komodakis. "Wide residual networks". In: *arXiv preprint arXiv:1605.07146* (2016).