ADVERSARIAL TRAINING FOR FREE!

Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry Davis, Gavin Taylor, Tom Goldstein

2019

SUPERVISED MACHINE LEARNING

some training data (image, labels)

Pandas

or

Pumpkins

Training Algorithm

Classifier

ADVERSARIAL EXAMPLES

$$f(x) \rightarrow y$$

$$f(x+\delta) \neq f(x)$$

s.t.

$x+\delta$ $\boxed{\text{looks like}}$ x

$$\|\delta\|_p \leq \epsilon$$

“Ox” 85%

ℓ_∞

“Traffic light” 96%

ℓ_0

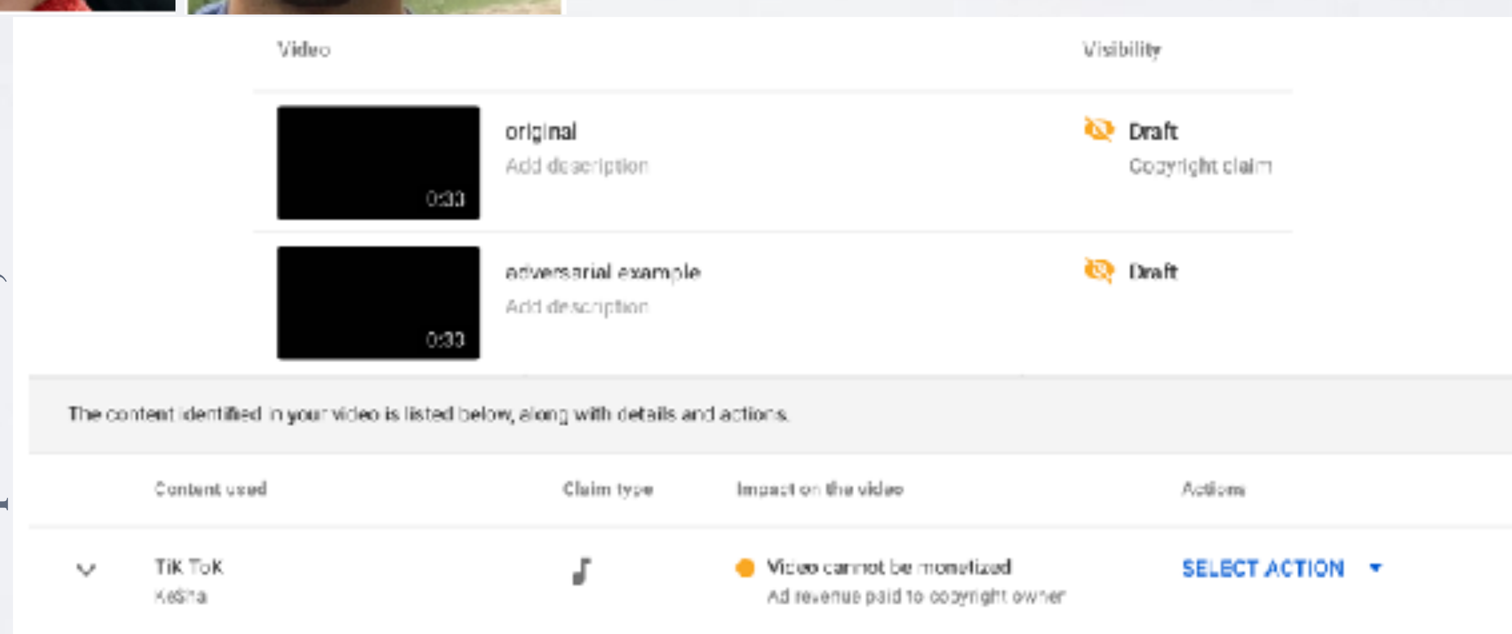
only 3% of pixels

REALISTIC ATTACKS

Sharif et al., 2016

Eykholt et al., 2018

Saadatpanah et al., 2019



CCS'16 Sharif, Bhagavatula, Bauer, Reiter "Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition"
CVPR 18 Eykholt, Evtimov, Fernandes, Li, Rahmati, Xiao, Prakash, Kohno, Song "Robust Physical-World Attacks on Deep Learning Visual Classification"
ArXiv 19 Saadatpanah, Shafahi, Goldstein "Adversarial Attacks on Copyright Detection"

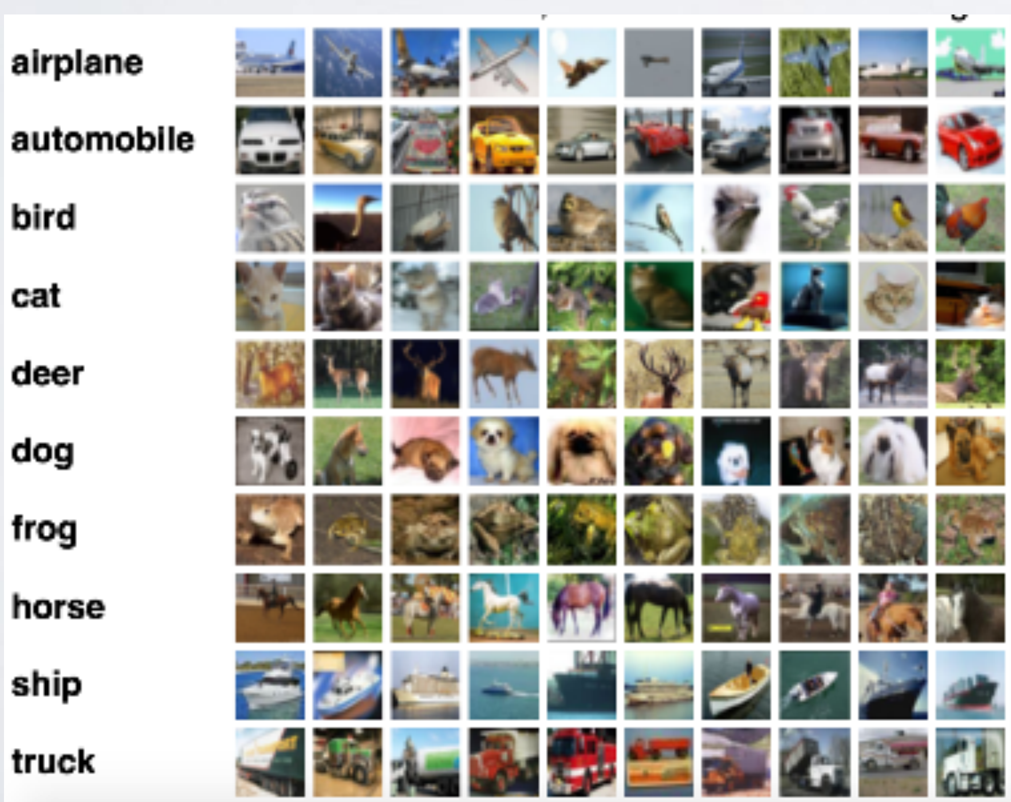
ROBUSTNESS AGAINST PER-INSTANCE PERTURBATIONS

Defending against non-targeted per-instance attacks is difficult...

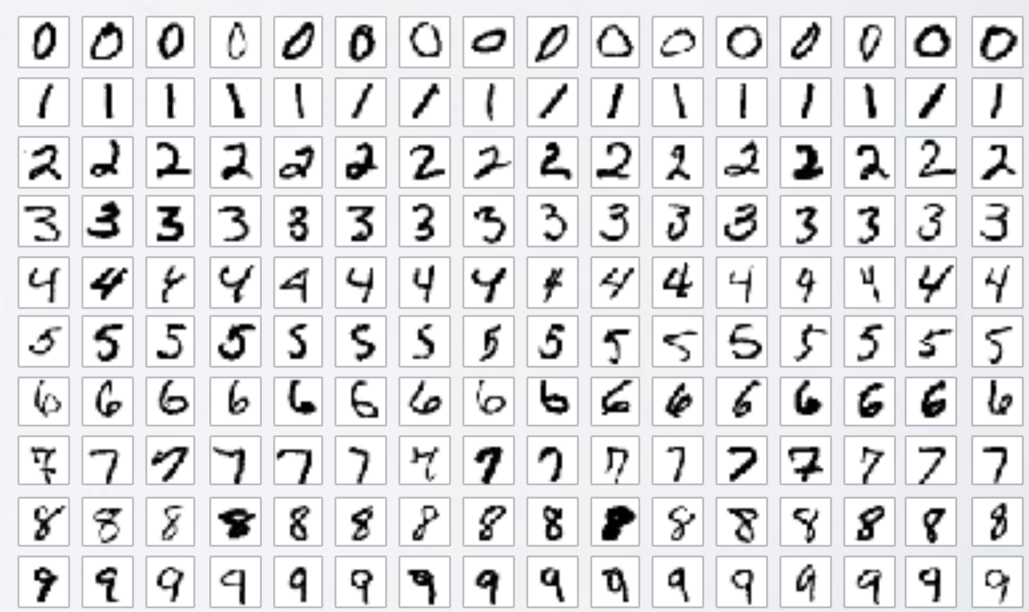
Small ϵ is used for p-norm bounded attacks

For larger datasets (ImageNet) defenses focused on random targets

Most studies focus on smaller datasets (CIFAR & MNIST)



CIFAR-10



MNIST

DEFENDING IS TOUGH

Defense	Defense type	Under which attack	Dataset	Distance	$\mathcal{A}_{\text{nat}}(f)$	$\mathcal{A}_{\text{rob}}(f)$
[BRRG18]	gradient mask	[ACW18]	CIFAR10	0.031 (ℓ_∞)	-	0%
[MLW+18]	gradient mask	[ACW18]	CIFAR10	0.031 (ℓ_∞)	-	5%
[DAL+18]	gradient mask	[ACW18]	CIFAR10	0.031 (ℓ_∞)	-	0%
[SKN+18]	gradient mask	[ACW18]	CIFAR10	0.031 (ℓ_∞)	-	9%
[NKM17]	gradient mask	[ACW18]	CIFAR10	0.015 (ℓ_∞)	-	15%
[WSMK18]	robust opt.	FGSM^{20} (PGD)	CIFAR10	0.031 (ℓ_∞)	27.07%	23.54%
[MMS+18]	robust opt.	FGSM^{20} (PGD)	CIFAR10	0.031 (ℓ_∞)	87.30%	**47.04%**
[ZSLG16]	regularization	FGSM^{20} (PGD)	CIFAR10	0.031 (ℓ_∞)	94.64%	0.15%
[KGB17]	regularization	FGSM^{20} (PGD)	CIFAR10	0.031 (ℓ_∞)	85.25%	45.89%
[RDV17]	regularization	FGSM^{20} (PGD)	CIFAR10	0.031 (ℓ_∞)	95.34%	0%

source: Zhang et. al, 2019

Defense	Dataset	Distance	Accuracy
Buckman et al. (2018)	CIFAR	0.031 (ℓ_∞)	0%*
Ma et al. (2018)	CIFAR	0.031 (ℓ_∞)	5%
Guo et al. (2018)	ImageNet	0.005 (ℓ_2)	0%*
Dhillon et al. (2018)	CIFAR	0.031 (ℓ_∞)	0%
Xie et al. (2018)	ImageNet	0.031 (ℓ_∞)	0%*
Song et al. (2018)	CIFAR	0.031 (ℓ_∞)	9%*
Samangouei et al. (2018)	MNIST	0.005 (ℓ_2)	55%**
Madry et al. (2018)	CIFAR	0.031 (ℓ_∞)	47%
Na et al. (2018)	CIFAR	0.015 (ℓ_∞)	15%

source: Athalye et. al, 2018

PGD Adversarial training!

ICML 18 Athalye, Carlini, Wagner “Obfuscated gradients give a false sense of security”
ICML 19 Zhang, Yu, Jiao, Xing, El Ghaoui, Jordan “Theoretically principled trade-off between robustness and accuracy”

ADVERSARIAL TRAINING

$$\min_{w} \quad \max_{\delta_i} \quad \frac{1}{N}\sum_{i=1}^{N} J(w, x_i + \delta_i)$$

s.t. $\quad \|\delta_i\|_p \leq \epsilon \quad \forall i \in \{1..N\}$

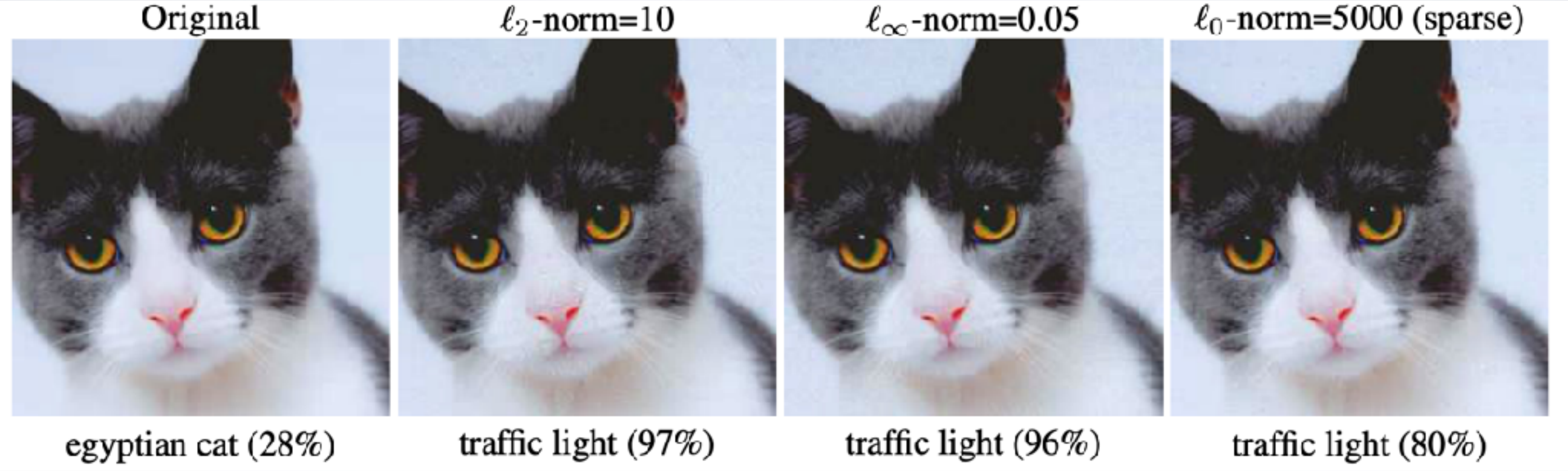


image source: Shafahi et. al, 2019

ICLR 19 Shafahi, Huang, Studer, Feizi, Goldstein "Are adversarial examples inevitable?"

PGD ADVERSARIAL TRAINING

Algorithm 1 Standard Adversarial Training (K-PGD)

Require: Training samples X, perturbation bound ϵ, step size ϵ_s, maximization iterations per minimization step K, and minimization learning rate τ

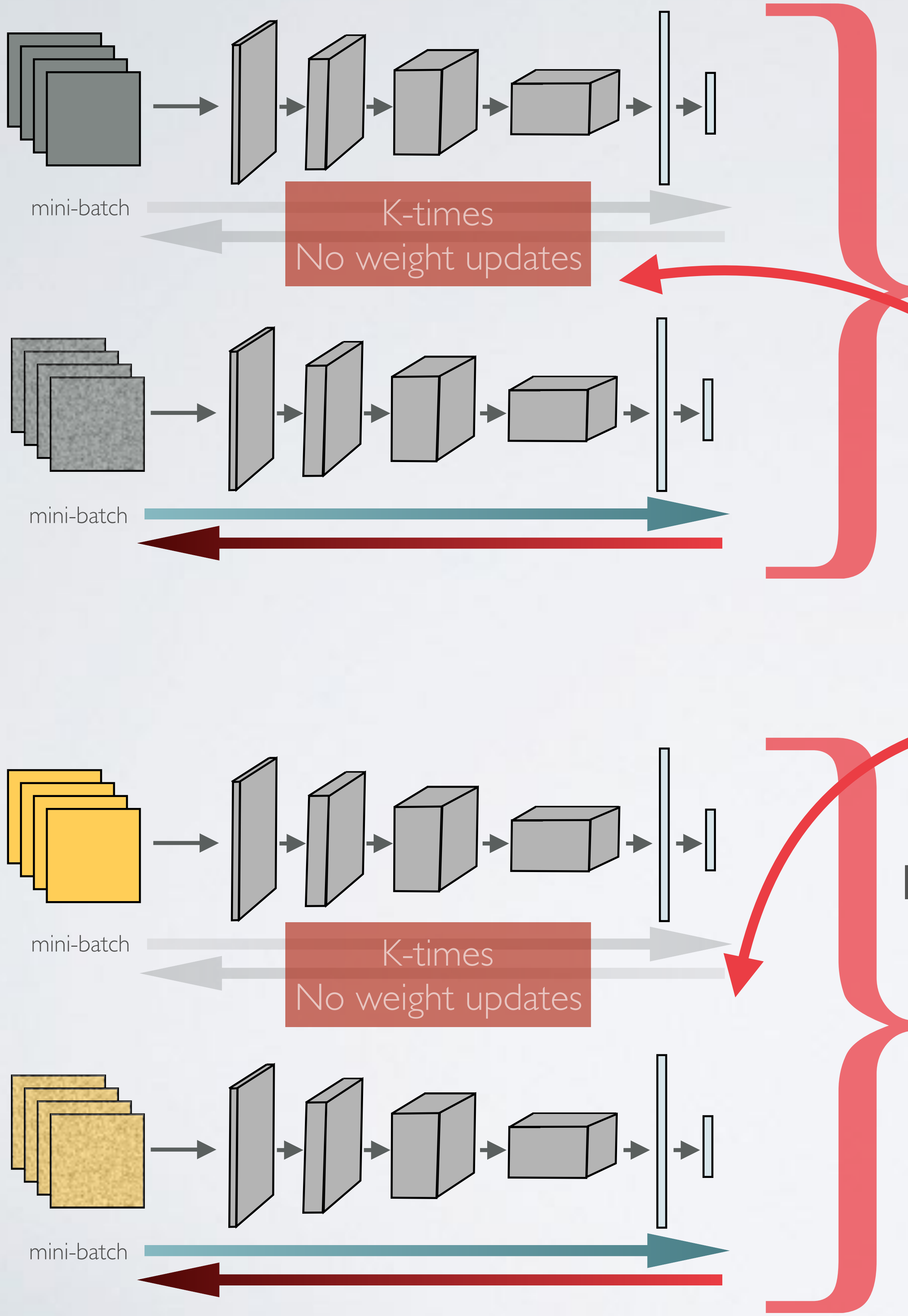
```
Initialize θ
for epoch = 1 ... N_ep do
    for minibatch B ⊂ X do
        Build x_adv for x ∈ B with PGD:
            Assign a random perturbation
                r ← U(−ε, ε)
                x_adv ← x + r
            for k = 1 ... K do
                g_adv ← ∇_x l(x_adv, y, θ)
                x_adv ← x_adv + ε_s · sign(g_adv)
                x_adv ← clip(x_adv, x − ε, x + ε)
            end for
        Update θ with stochastic gradient descent:
            g_θ ← E_{(x,y)∈B}[∇_θ l(x_adv, y, θ)]
            θ ← θ − τ g_θ
    end for
end for
```

Adversarial example generation

Madry et. al, 2018

ICLR 18 Madry, Makelov, Schmidt, Tsipras, Vladu “Towards deep learning models resistant to adversarial attacks”

PGD ADV. TRAINING



K-PGD adversarial training

$$\min_{w} \max_{\delta_i} \frac{1}{N}\sum_{i=1}^{N} J(w, x_i + \delta_i)$$

First weight update

Adds a K-factor Overhead:
We perform an additional K Forward and Backward passes without updating the network parameters

Second weight update

ADVERSARIAL TRAINING WITH PGD REQUIRES MANY FWD/BWD PASSES

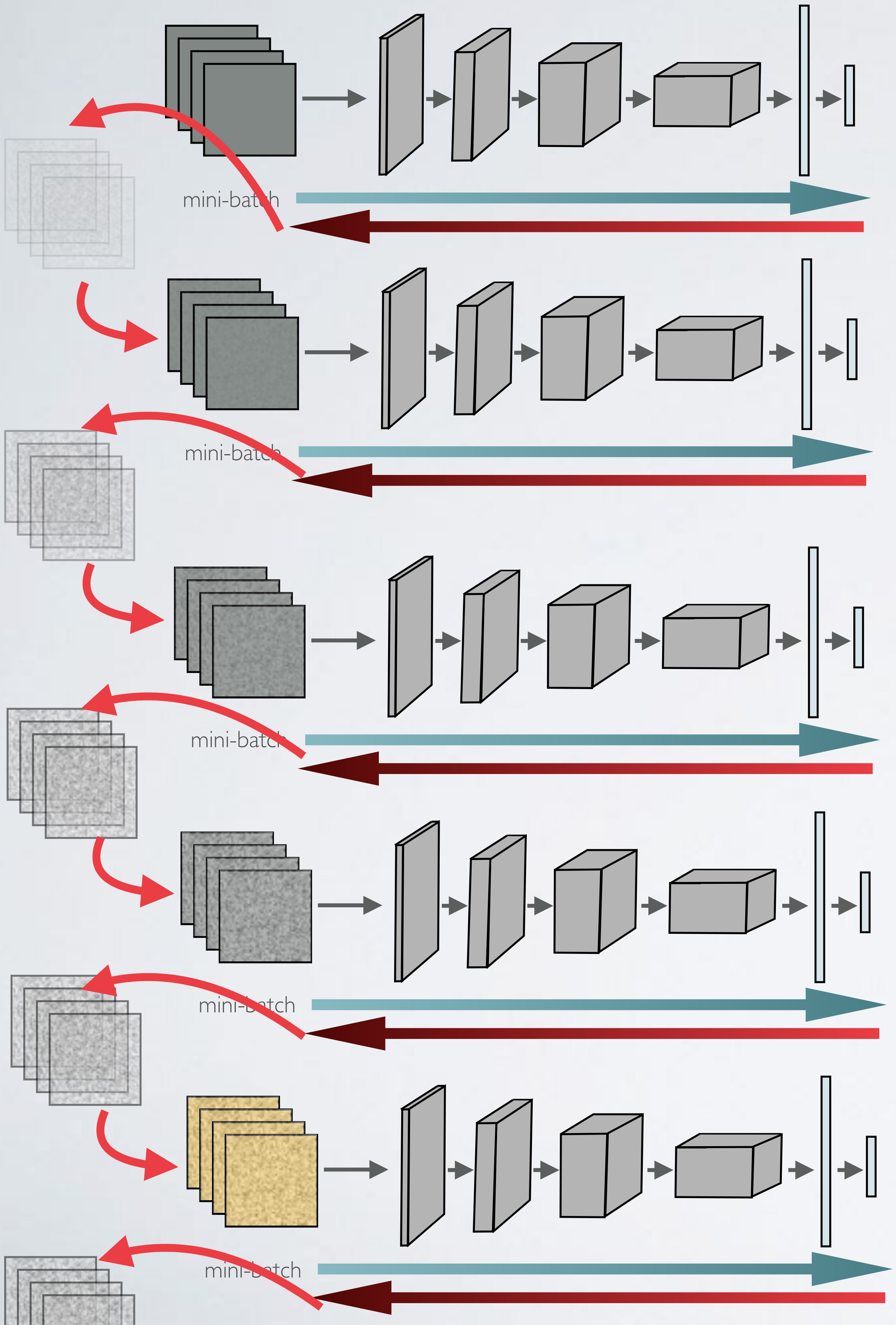
Impractical for ImageNet?

Not if you have a lot of compute…

Kannan et al., 2018	**53 P100s**
Xie et al., 2019	**128 V100s**
Qin et al., 2019	**128 TPUv3**

ArXiv 18 Kannan, Kurakin, Goodfellow "Adversarial Logit Pairing"
CVPR 19 Xie, Wu, Maaten, Yuille, He "Feature denoising for improving adversarial robustness"
NeurIPS 19 Qin, Martens, Gowal, Krishnan, Fawzi, De, Stanforth, Kohli "Adversarial Robustness Through Local Linearization"

ADV. TRAINING FOR FREE!



Free-m adversarial training

First weight update

Second weight update

m[=4] replays of the same mini-batch

Third weight update

Fourth weight update

Unlike K-PGD training, we update the network parameters every time we do a back-ward pass

Fifth weight update

ADVERSARIAL TRAINING FOR FREE!

Algorithm 2 "Free" Adversarial Training (Free-m)

Require: Training samples X, perturbation bound ϵ, learning rate τ, hop steps m

Initialize θ
$\delta \leftarrow 0$
for epoch $= 1 \dots N_{ep}/m$ **do**
    for minibatch $B \subset X$ **do**
        for i $= 1 \dots m$ **do**
            Update θ with stochastic gradient descent
                $g_\theta \leftarrow \mathbb{E}_{(x,y)\in B}[\nabla_\theta\, l(x+\delta, y, \theta)]$
                $g_{adv} \leftarrow \nabla_x\, l(x+\delta, y, \theta)]$
                $\theta \leftarrow \theta - \tau g_\theta$
            Use gradients calculated for the minimization step to update δ
                $\delta \leftarrow \delta + \epsilon \cdot \mathrm{sign}(g_{adv})$
                $\delta \leftarrow \mathrm{clip}(\delta, -\epsilon, \epsilon)$
        end for
    end for
end for

- Update both perturbation and network parameters in one pass
- Replay every mini-batch *m* times to simulate PGD training

NeurIPS 19 Shafahi, Najibi, Ghiasi, Xu, Dickerson, Studer, Davis, Taylor, Goldstein "Adversarial Training for Free!"

ADVERSARIAL TRAINING FOR FREE!

Table 1: Validation accuracy and robustness of CIFAR-10 models trained with various methods.

Training	Evaluated Against: Natural Images	PGD-20	PGD-100	CW-100	10 restart PGD-20	Training Time (minutes)
Natural	**95.01%**	0.00%	0.00%	0.00%	0.00%	**780**
Free $m = 2$	91.45%	33.92%	33.20%	34.57%	33.41%	816
Free $m = 4$	87.83%	41.15%	40.35%	41.96%	40.73%	800
Free $m = 8$	85.96%	**46.82%**	**46.19%**	**46.60%**	**46.33%**	785
Free $m = 10$	83.94%	46.31%	45.79%	45.86%	45.94%	785
Madry *et al.* (7-PGD trained)	87.25%	45.84%	45.29%	46.52%	45.53%	5418

Architecture: WRN 32-10
No. of iterations = 80k
batchsize=128
epsilon=8

Robust against many attacks

Fast

NeurIPS 19 Shafahi, Najibi, Ghiasi, Xu, Dickerson, Studer, Davis, Taylor, Goldstein "Adversarial Training for Free!"

ADV. TRAINING FOR FREE!

Free-m also maintains important valuable properties of PGD adversarially trained models

natural

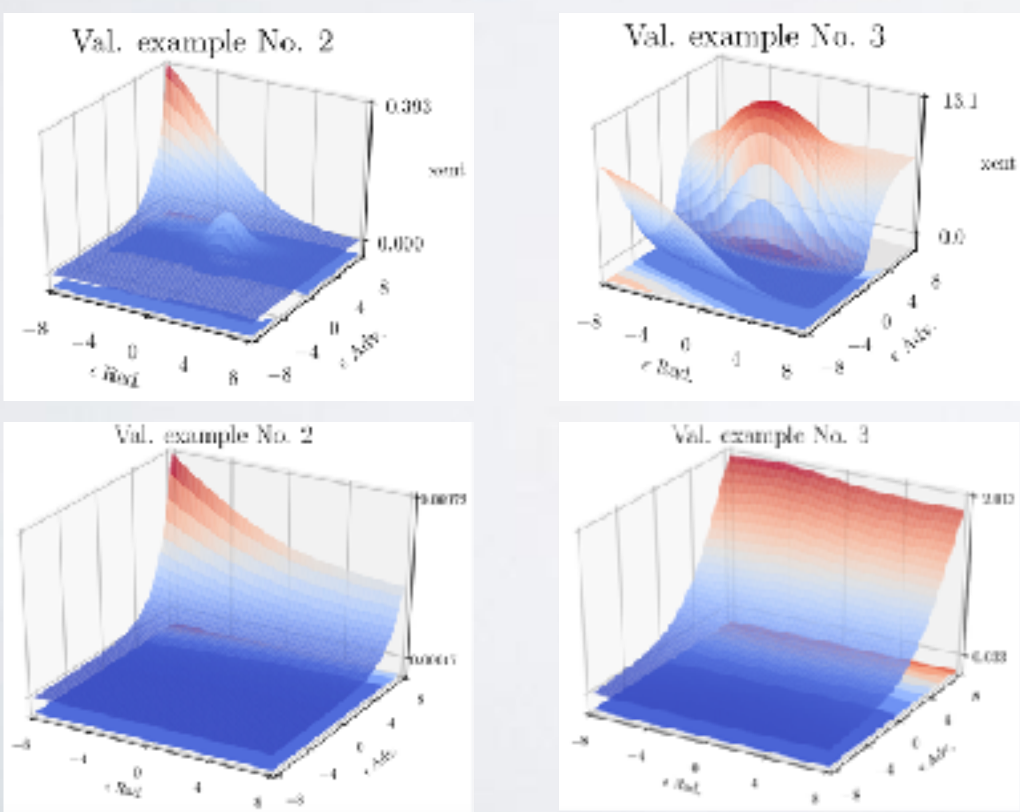
PGD-7

Free-8

Interpretable gradients

natural

Free-8

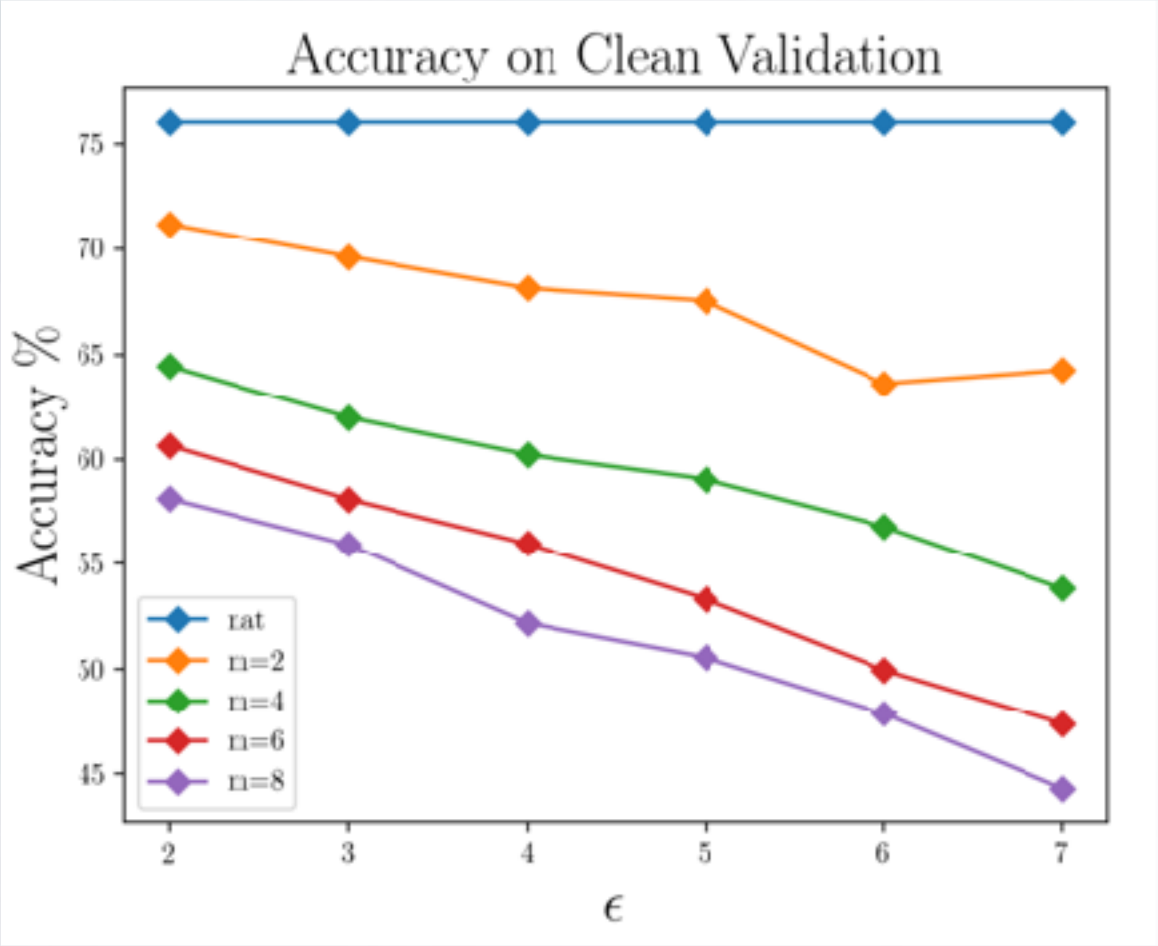


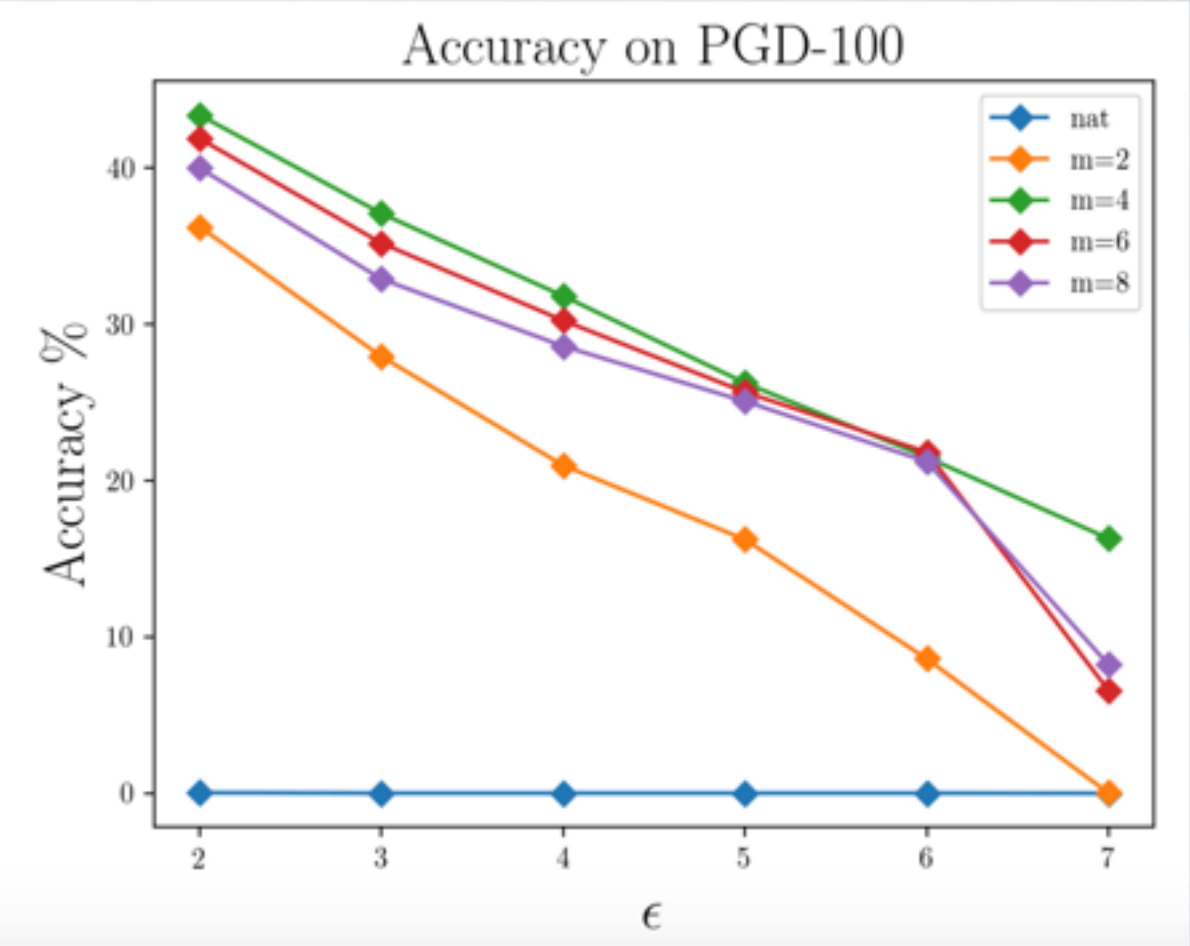
Smooth and flattened loss surface
compared to naturally trained models

NeurIPS 19 Shafahi, Najibi, Ghiasi, Xu, Dickerson, Studer, Davis, Taylor, Goldstein "Adversarial Training for Free!"

ADV. TRAINING FOR FREE!

ImageNet (ResNet-50)





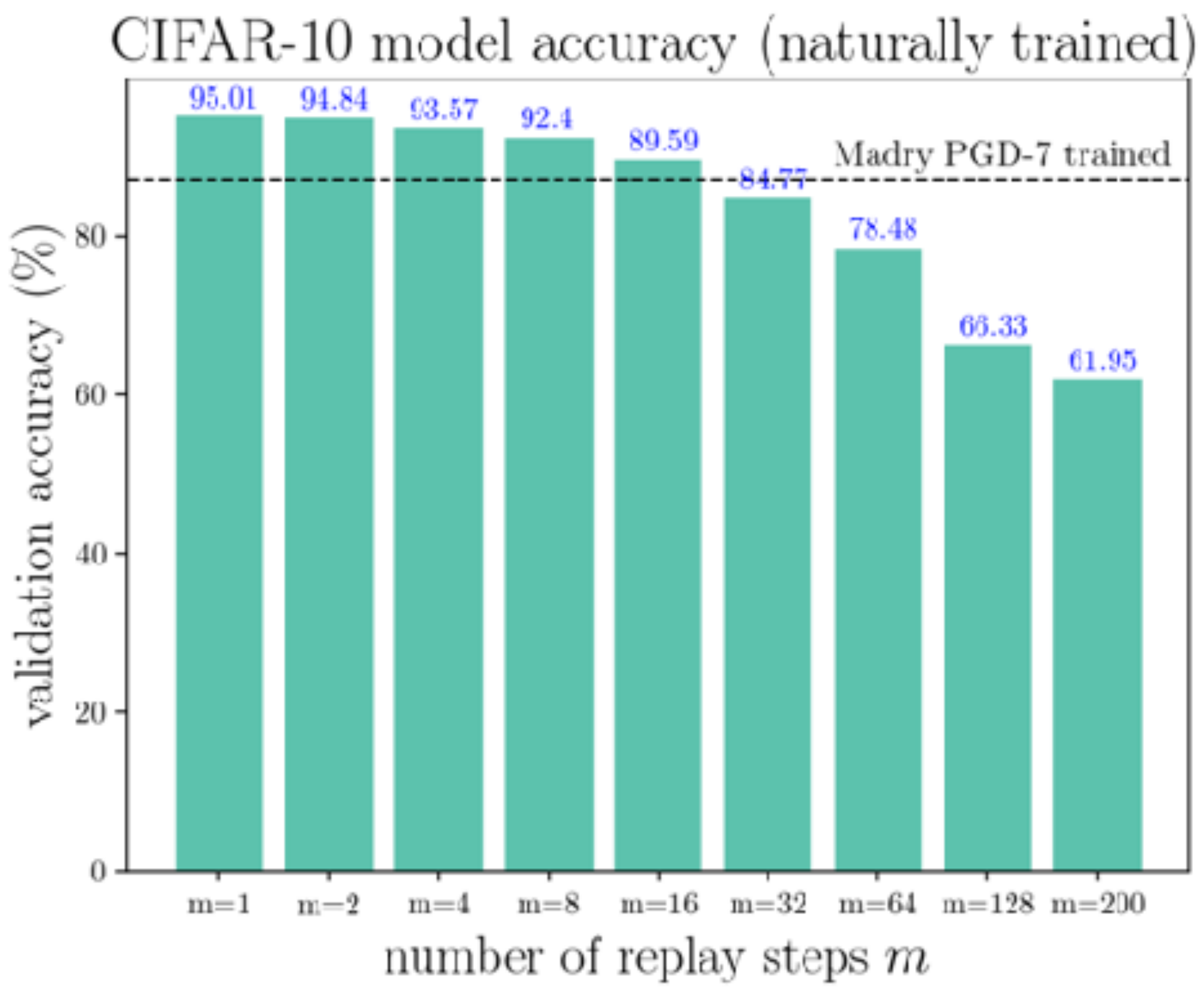
Free-4
(epsilon=4)
batchsize=256

Architecture	Evaluated Against			
	Natural Images	PGD-10	PGD-50	PGD-100
ResNet-50	60.206%	32.768%	31.878%	31.816%
ResNet-101	63.340%	35.388%	34.402%	34.328%
ResNet-152	**64.446%**	**36.992%**	**36.044%**	**35.994%**

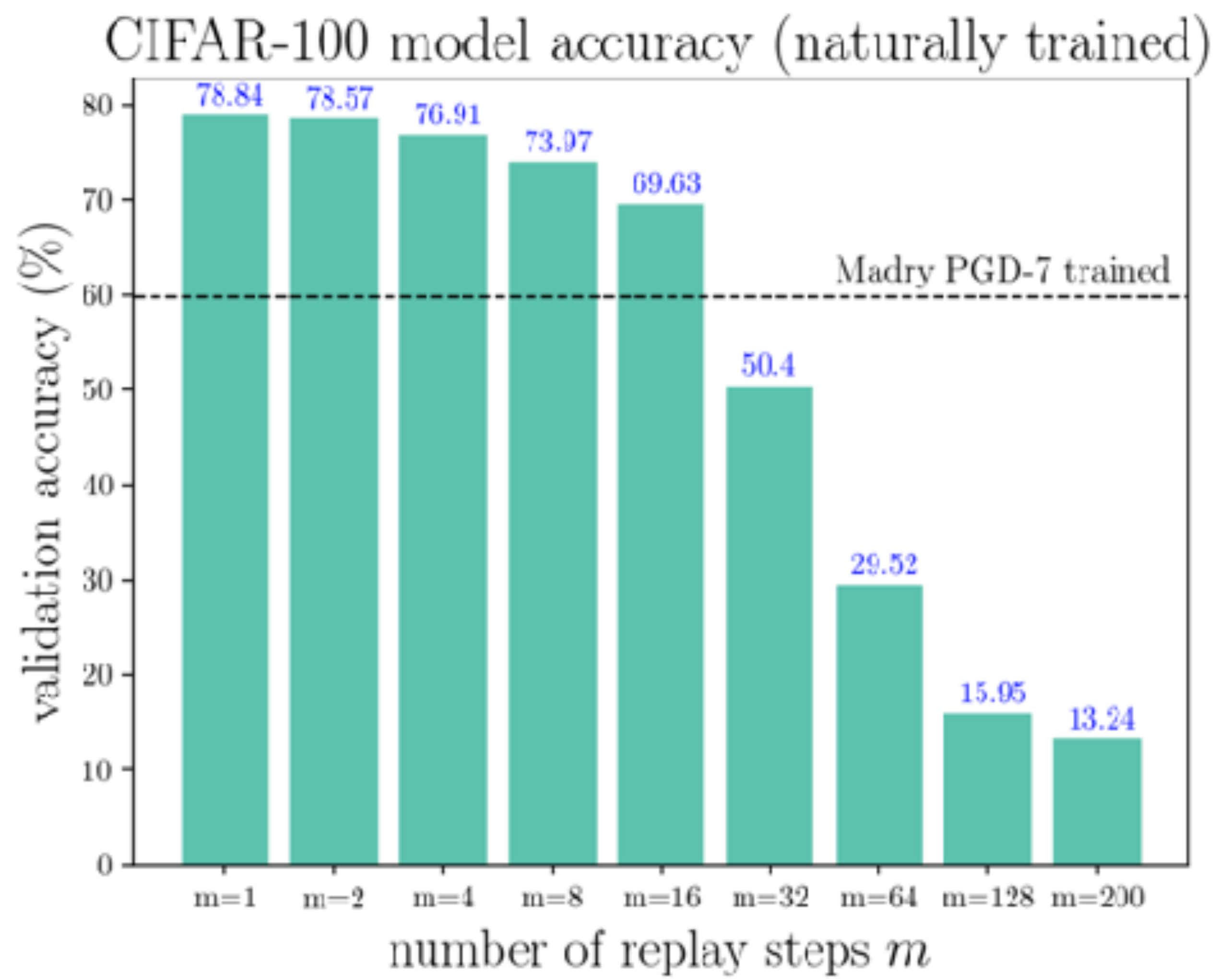
NeurIPS 19 Shafahi, Najibi, Ghiasi, Xu, Dickerson, Studer, Davis, Taylor, Goldstein "Adversarial Training for Free!"

ADVERSARIAL TRAINING FOR FREE!

How much replaying a mini-batch hurts…



(a) CIFAR-10 sensitivity to m



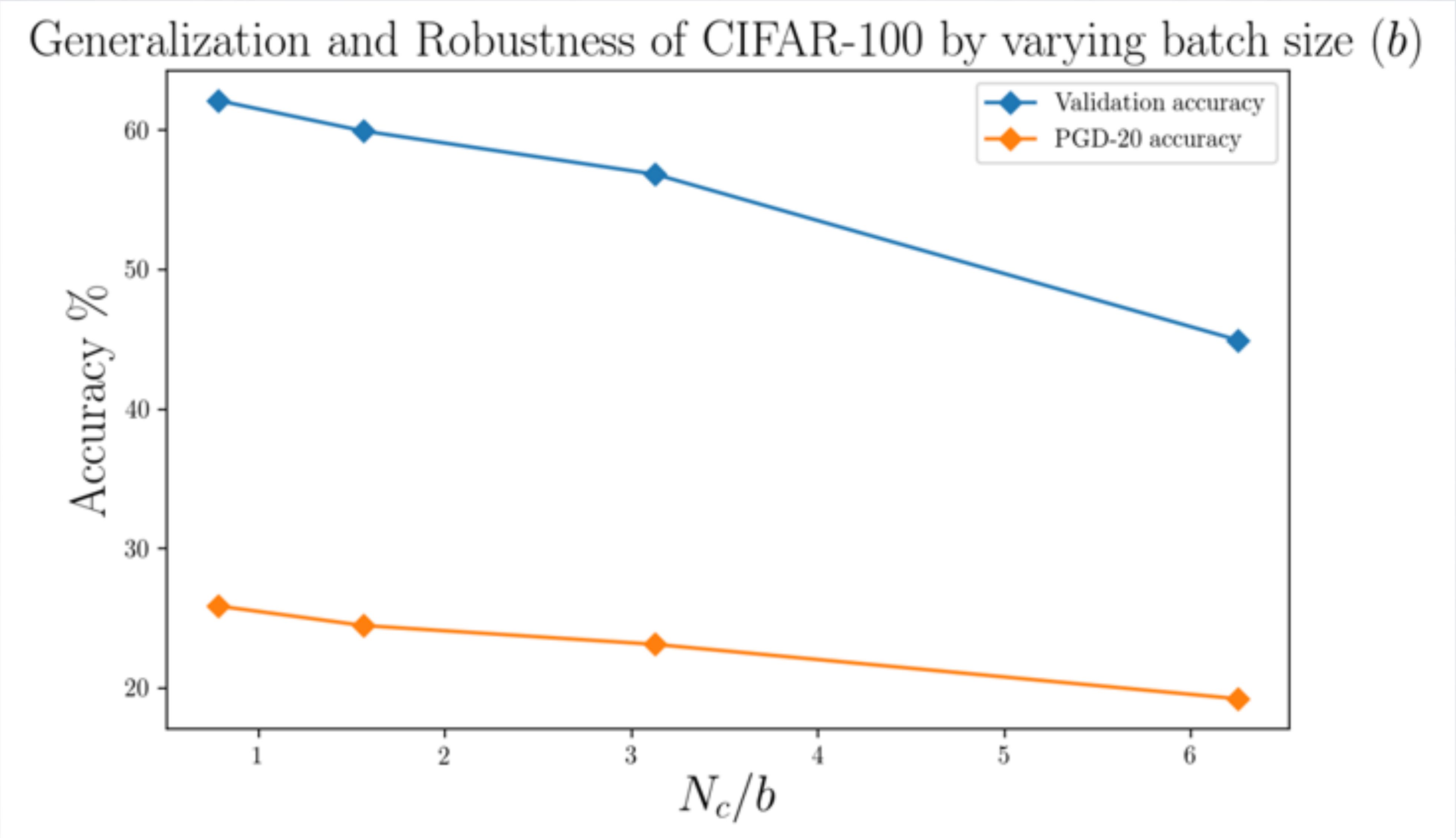
(b) CIFAR-100 sensitivity to m

Architecture: WRN 32-10
No. of iterations = 80k
batchsize=128

NeurIPS 19 Shafahi, Najibi, Ghiasi, Xu, Dickerson, Studer, Davis, Taylor, Goldstein "Adversarial Training for Free!"

ADVERSARIAL TRAINING FOR FREE!

Could compensate some of the negative effects of replay by increasing batch-size



Our ImageNet result were with a batch-size of 256 … $N_c/b = \frac{1000}{256} = 3.91$

FREE CODE!

Pytorch

ImageNet

Tensorflow

CIFAR

THANK YOU!