CVPR SEATTLE WASHINGTON
JUNE 16-18 2020

Benchmarking Adversarial Robustness on Image Classification

Yinpeng Dong, Qi-An Fu, Xiao Yang, Tianyu Pang, Zihao Xiao, Hang Su, Jun Zhu
Dept. of Comp. Sci. and Tech., BNRist Center, Institute for AI, THBI Lab,
Tsinghua University, Beijing, 100084, China
Contact: dyp17@mails.tsinghua.edu.cn; fqa19@mails.tsinghua.edu.cn

Adversarial Examples

An adversarial example is crafted by adding a small perturbation, which is visually indistinguishable from the corresponding normal one, but yet are misclassified by the target model.

There is an "arms race" between attacks and defenses, making it hard to understand their effects.

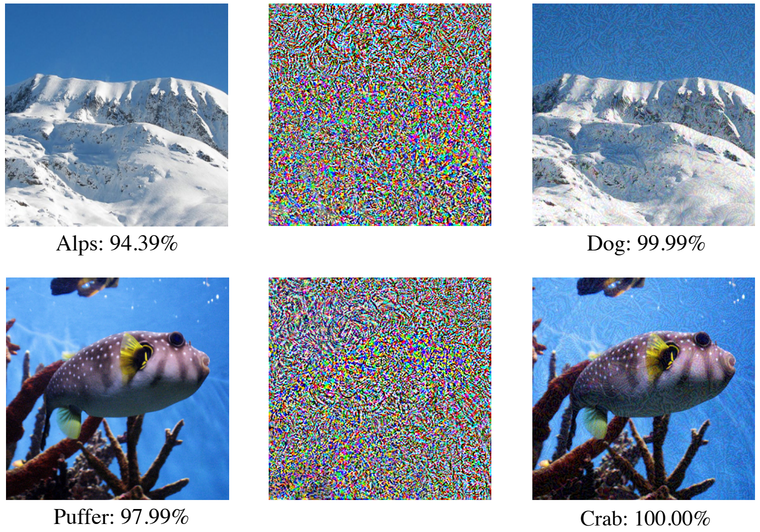


Figure from Dong et al. (2018).

Attacks

Adaptive attacks [Athalye et al., 2018]

Optimization-based attacks [Carlini and Wagner, 2017]

Iterative attacks[kurakin et al., 2016]

One-step attacks [Goodfellow et al., 2014]

Defenses

Randomization, denoising [Xie et al., 2018; Liao et al., 2018]

Defensive distillation [Papernot et al., 2016]

Adversarial training with FGSM [Kurakin et al., 2015]

Robustness Benchmark

- Threat Models: we define complete threat models
- Attacks: we adopt 15 attacks
- Defenses: we adopt 16 defenses on CIFAR-10 and ImageNet
- Evaluation Metrics:
  - **Accuracy (attack success rate) vs. perturbation budget** curves
  - **Accuracy (attack success rate) vs. attack strength** curves

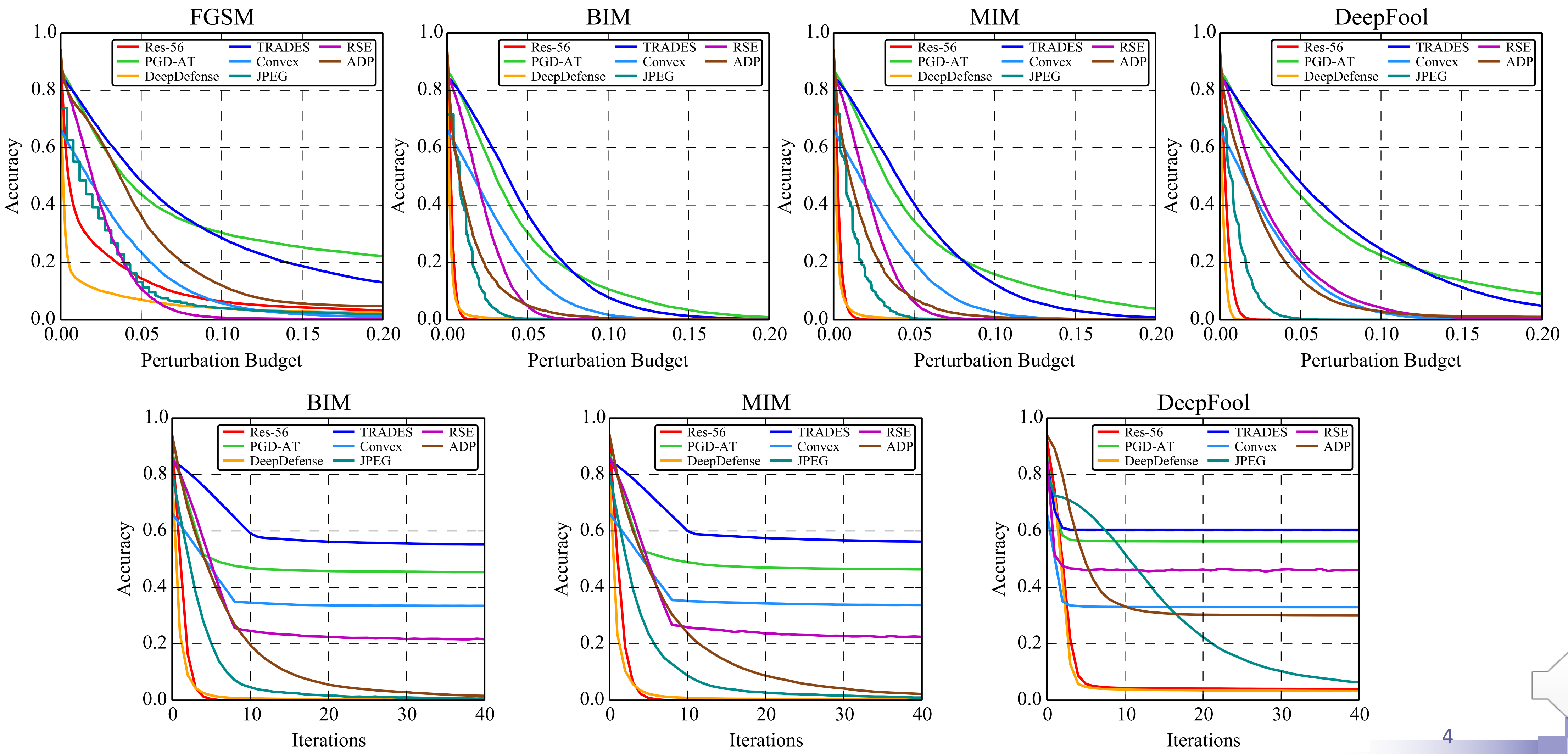
Attack Method	Knowledge	Goals	Capability	Distance
FGSM [17]	white & transfer	un. & tar.	constrained	ℓ_∞, ℓ_2
BIM [27]	white & transfer	un. & tar.	constrained	ℓ_∞, ℓ_2
MIM [13]	white & transfer	un. & tar.	constrained	ℓ_∞, ℓ_2
DeepFool [34]	white	un.	optimized	ℓ_∞, ℓ_2
C&W [7]	white	un. & tar.	optimized	ℓ_2
DIM [59]	transfer	un. & tar.	constrained	ℓ_∞, ℓ_2
ZOO [8]	score	un. & tar.	optimized	ℓ_2
NES [22]	score	un. & tar.	constrained	ℓ_∞, ℓ_2
SPSA [52]	score	un. & tar.	constrained	ℓ_∞, ℓ_2
$\mathcal{N}$ATTACK [29]	score	un. & tar.	constrained	ℓ_∞, ℓ_2
Boundary [3]	decision	un. & tar.	optimized	ℓ_2
Evolutionary [14]	decision	un. & tar.	optimized	ℓ_2

CIFAR-10 [25]				ImageNet [43]			
Defense Model	Category	Intended Threat	Acc.	Defense Model	Category	Intended Threat	Acc.
Res-56 [19]	natural training	-	92.6	Inc-v3 [49]	natural training	-	78.0
PGD-AT [33]	robust training	ℓ_∞ ($\epsilon = 8/255$)	87.3	Ens-AT [51]	robust training	ℓ_∞ ($\epsilon = 16/255$)	73.5
DeepDefense [61]	robust training	ℓ_2	79.7	ALP [23]	robust training	ℓ_∞ ($\epsilon = 16/255$)	49.0
TRADES [63]	robust training	ℓ_∞ ($\epsilon = 8/255$)	84.9	FD [58]	robust training	ℓ_∞ ($\epsilon = 16/255$)	64.3
Convex [54]	(certified) robust training	ℓ_∞ ($\epsilon = 2/255$)	66.3	JPEG [15]	input transformation	General	77.3
JPEG [15]	input transformation	General	80.9	Bit-Red [60]	input transformation	General	61.8
RSE [31]	rand. & ensemble	ℓ_2	86.1	R&P [57]	(input) rand.	General	77.0
ADP [35]	ensemble	General	94.1	RandMix [64]	(certified input) rand.	General	52.4

Evaluation Results on CIFAR-10

CVPR SEATTLE WASHINGTON JUNE 16-18 2020

ℓ_∞ norm; untargeted attacks; white-box; accuracy curves



Platform: RealSafe

- We developed a new platform for adversarial machine learning research called `RealSafe` focusing on benchmarking adversarial robustness on image classification correctly & efficiently.
- Available at https://github.com/thu-ml/realsafe (Scan the QR code for this URL).

Feature highlights:

- Modular implementation, which consists of attacks, models, defenses, datasets, and evaluations.
- Support `tensorflow` & `pytorch` models with the same interface.
- Support 11 attacks & many defenses benchmarked in this work.
- Provide ready-to-use pre-trained baseline models (8 on ImageNet & 8 on CIFAR10).
- Provide efficient & easy-to-use tools for benchmarking models with the 2 robustness curves.

Thanks