ME-Net: Towards Effective Adversarial Robustness with Matrix Estimation

Yuzhe Yang, Guo Zhang, Dina Katabi and Zhi Xu

MIT Computer Science and Artificial Intelligence Laboratory

Adversarial Examples

- Adversarial noise is **highly structured**
- Such structure is designed to fool neural nets

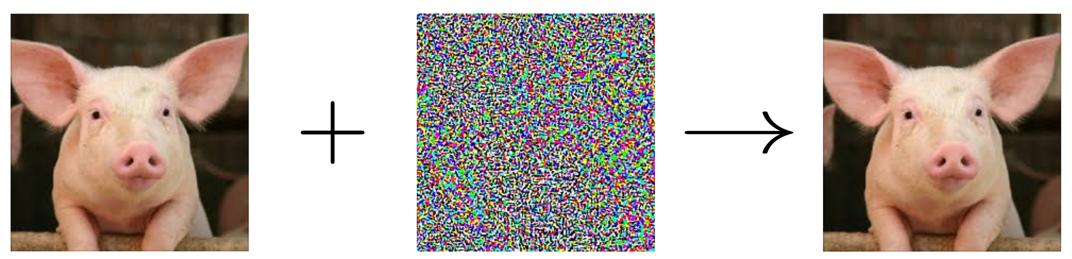

Figure 1: An illustration of adversarial examples.

Design Motivations

- Destroy the structure of adversarial noise
- Emphasize the **global structure** in the image

Idea: Images are approximately low-rank

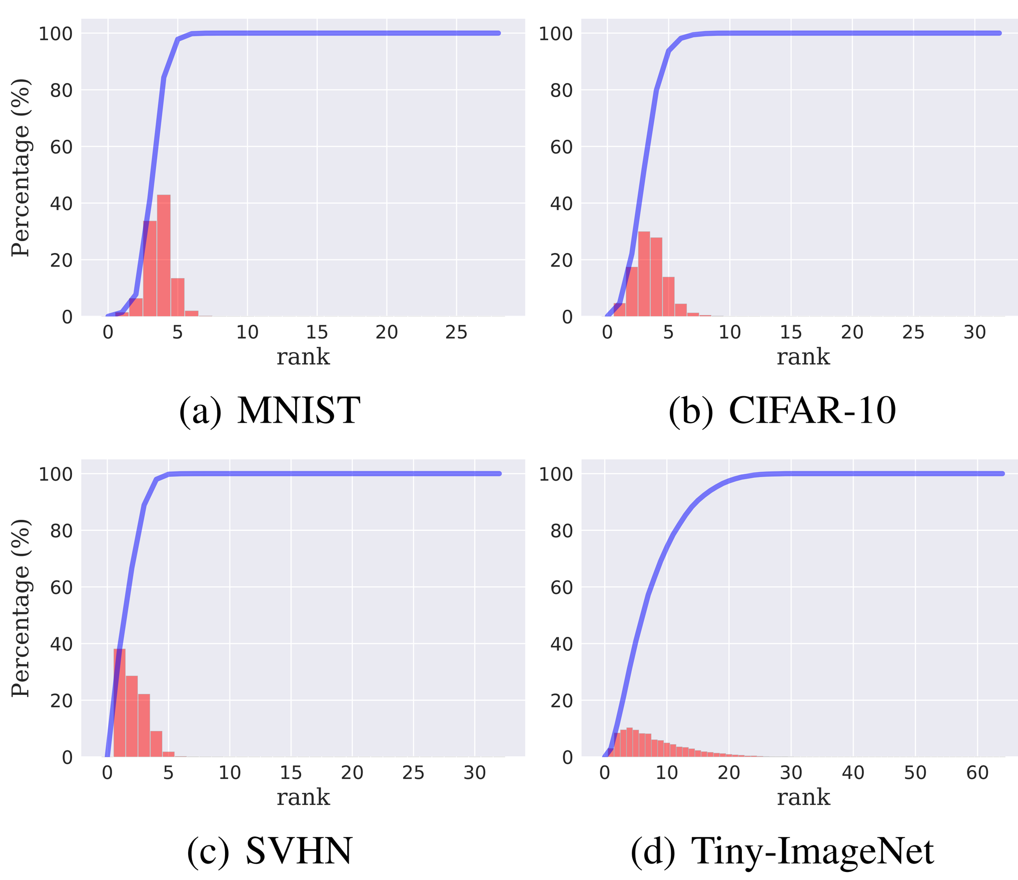

Figure 2: The approximate rank of different datasets.

Matrix Estimation

- Recover the underlying global structure from noisy and incomplete observations
- Theoretically guaranteed if true data matrix has some global structures (e.g., low rank)
- Algos: USVT, Soft-Impute, Nuclear norm . . .

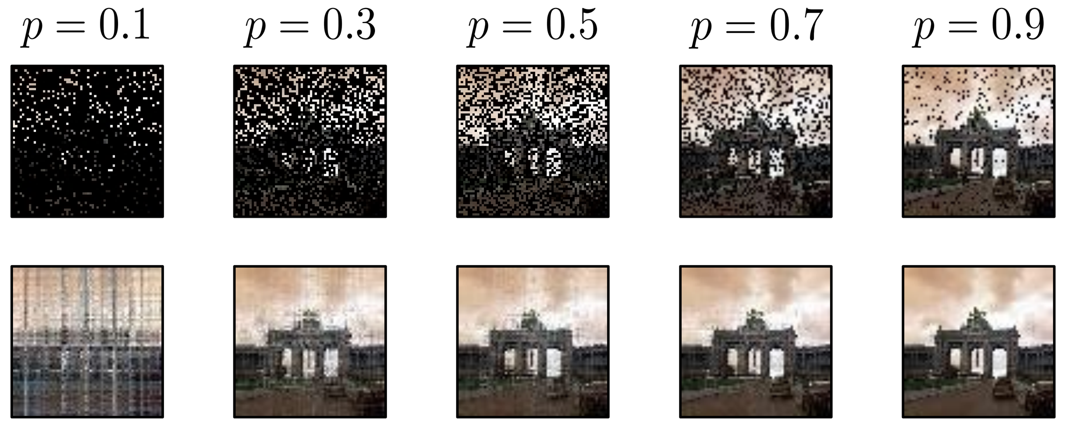

Figure 3: An example of how ME affects the input images.

ME-Net

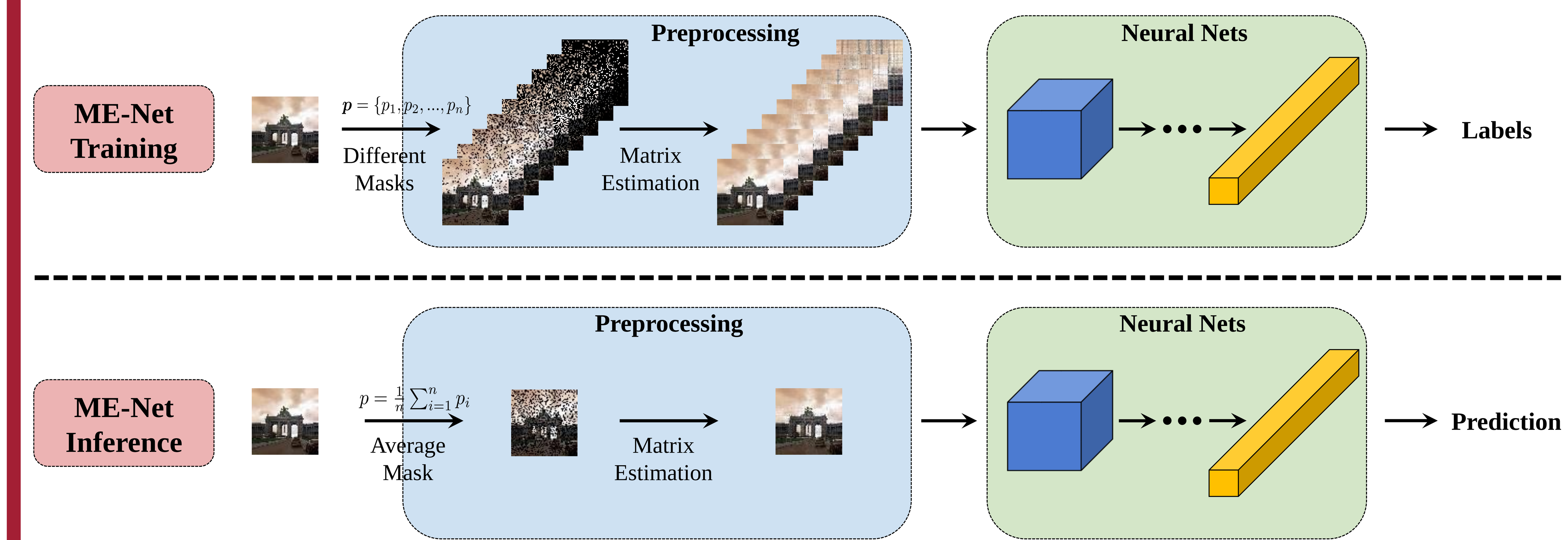

Figure 4: An illustration of ME-Net training and inference process.

- A new defense method that emphasizes the ***global structure*** in images using matrix estimation
- Creates more data for training by generating randomly subsampled versions for each example
- Can be combined with adversarial training, to further increase the robustness

Black-box Attacks

Threat model

- l_∞-bounded perturbation (8/255 for CIFAR)

Three types of black-box attacks

- **Transfer-based:** using FGSM, PGD, and CW
- **Decision-based:** Boundary attack
- **Score-based:** SPSA attack

Attack	Vanilla	Madry et al.	ME-Net
FGSM	24.8%	67.0%	**92.2%**
PGD	7.6%	64.2%	**91.8%**
CW	8.9%	78.7%	**93.6%**
Boundary	3.5%	61.9%	**87.4%**
SPSA	1.4%	47.0%	**93.0%**

Table 1: CIFAR-10 black-box attacks results.

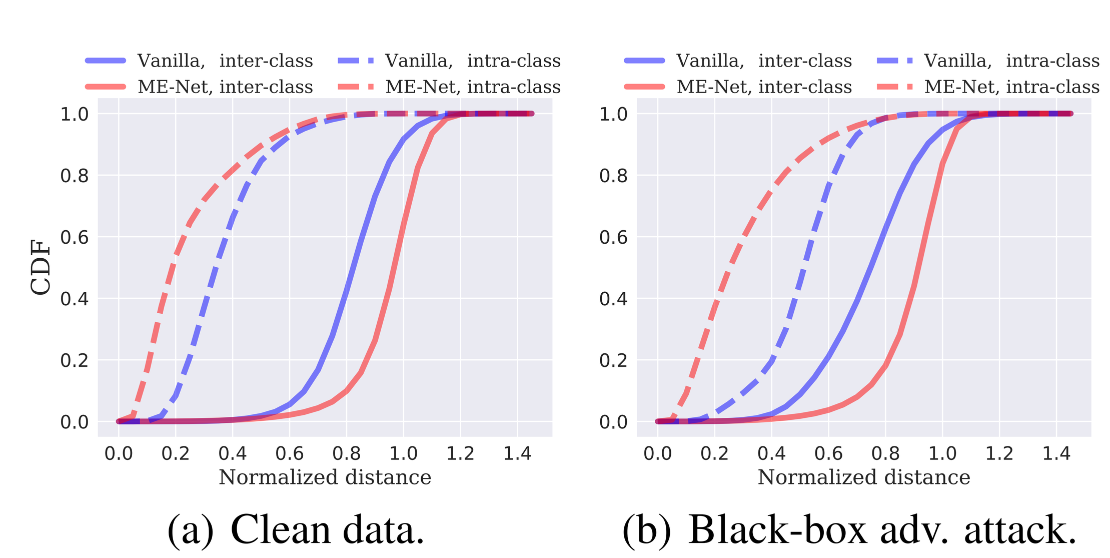

Figure 5: Empirical CDF of distance within and among classes.

White-box Attacks

- Compared with pure preprocessing methods

Method	Type	Steps	Accuracy
Thermometer	Prep.	40	0.0%
PixelDefend	Prep.	100	9.0%
TV Minimization	Prep.	100	0.4%
ME-Net	Prep.	1000	**40.8%**

Table 2: White-box attack against pure preprocessing schemes.

- Compared with SOTA adversarial training

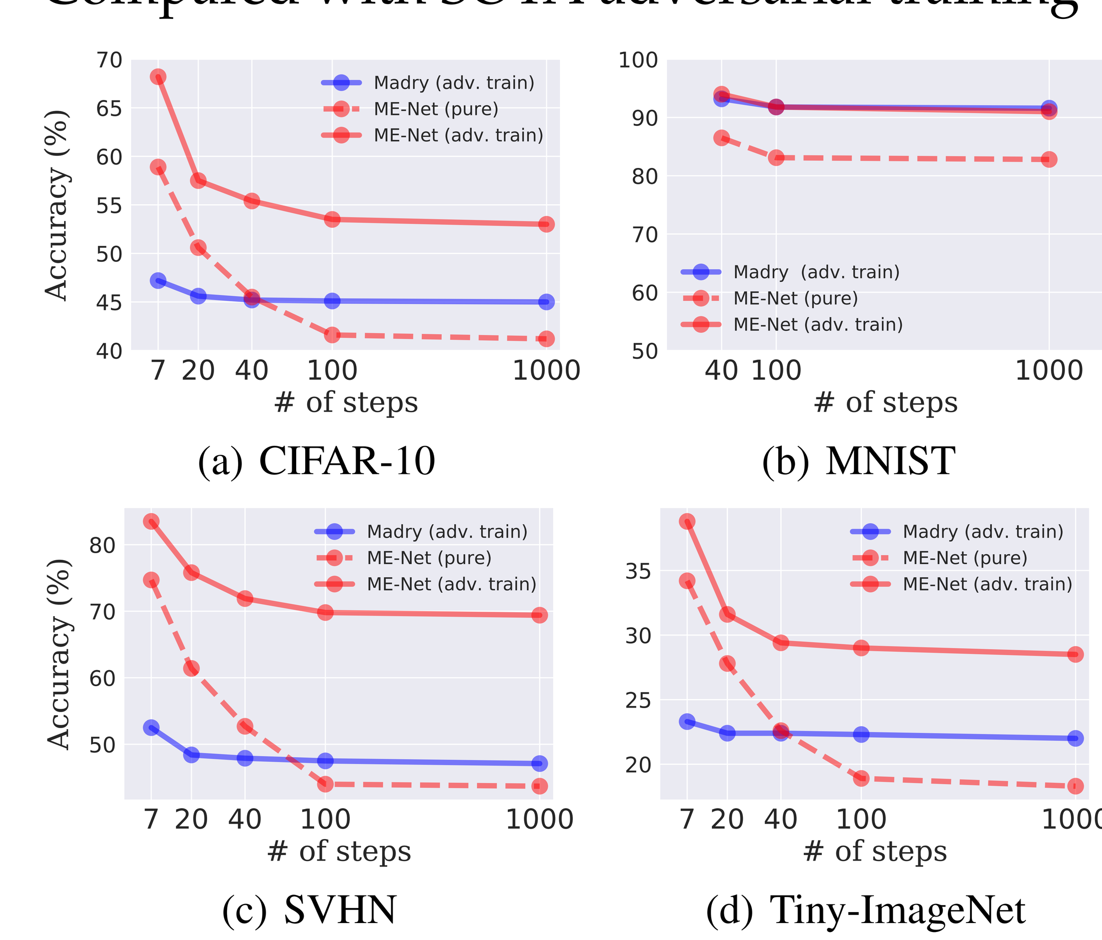

Figure 6: White-box attack results on different datasets.

Adaptive Attacks

- Constructed after a defense is specified
- Takes advantage of knowledge of the defense

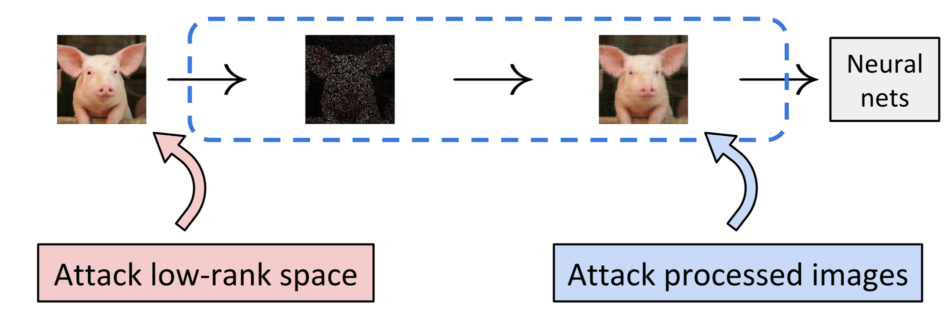

Figure 7: An illustration of two proposed adaptive attacks.

Approximate input attack

- uses exact preprocess to approximate inputs
- attacks the constructed inputs using BPDA

Projected BPDA attack

- attacks directly the main structural space
- projects grads to low-rank space iteratively

Training	Steps	Approx. Input	Projected BPDA
Pure	1000	**41.5%**	**64.9%**
Adversarial	1000	**62.5%**	**74.7%**

Table 3: Results of adaptive white-box attacks on CIFAR-10.

Additional Benefits

Improving generalization for both

- standard training (only with clean data)
- adversarial training (with adv. examples)

Method	Training	MNIST	CIFAR	SVHN	Tiny-ImageNet
Vanilla	Pure	98.8%	93.4%	95.0%	66.4%
ME-Net	Pure	**99.2%**	**94.9%**	**96.0%**	**67.7%**
Madry	Adv.	98.5%	79.4%	87.4%	45.6%
ME-Net	Adv.	**98.8%**	**85.5%**	**93.5%**	**57.0%**

Table 4: The generalization performance on clean data.

More Information

Source Code

Project Page