Adversarial Machine Learning - II

Heng Hsu

NIC Lab Group Meeting

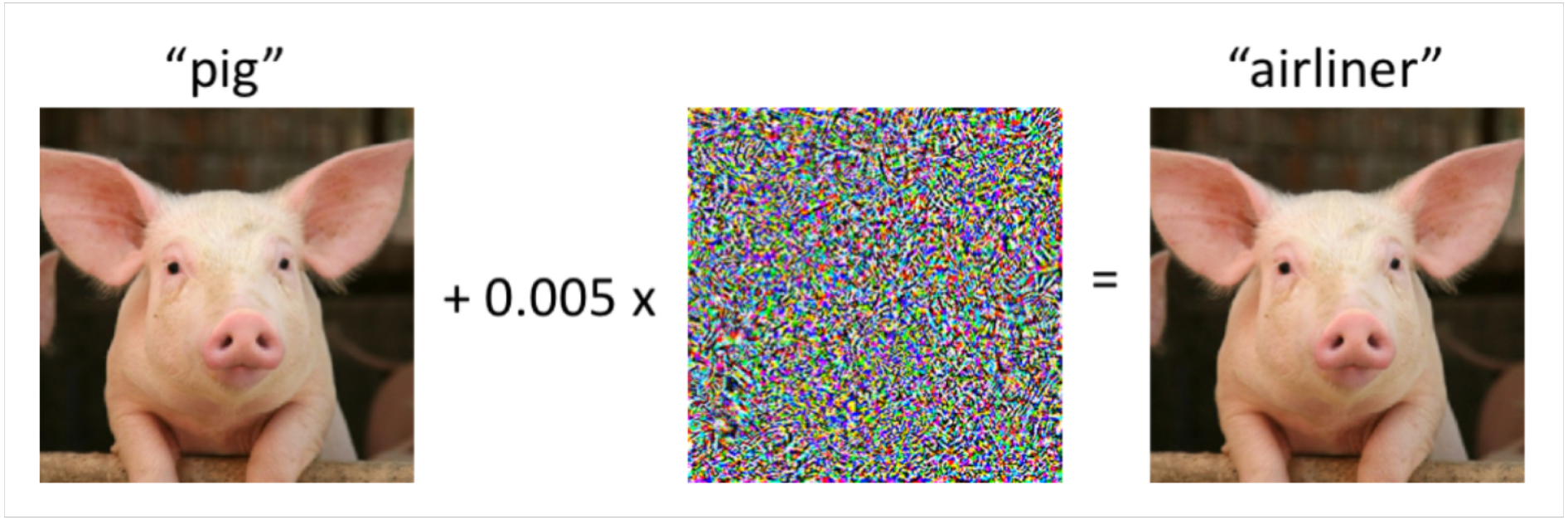
2018/10/30

Outline

- Quick review
- Two state-of-the-art methods
  - Ensemble Adversarial Training
  - Adversarial Logit Pairing
- Summary

Adversarial examples

- Adversarial examples are inputs to machine learning models that an attacker has **intentionally designed** to cause the model to make a mistake.[1]



[1] C. Szegedy et al., Intriguing properties of neural networks, ICLR 2014

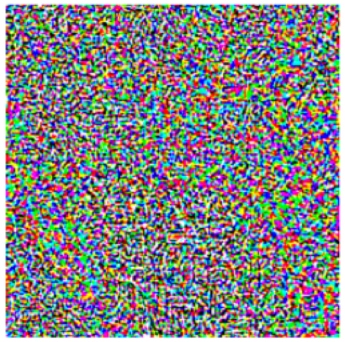
FGSM[2]

$$\boldsymbol{\eta} = \epsilon \text{sign}\left(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y)\right)$$

$\boldsymbol{x}$

"panda"
57.7% confidence

$+ .007 \times$



$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

$=$

$\boldsymbol{x} + \epsilon \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
99.3 % confidence

[2] I. Goodfellow et al., Explaining and Harnessing Adversarial Examples, ICLR 2015

Gradient Masking (Obfuscated Gradients)

- Some defenses look like they work because they break gradient-based white box attacks.
- The defense denies the attacker access to a useful gradient but these defenses are still vulnerable.
- Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples[3]

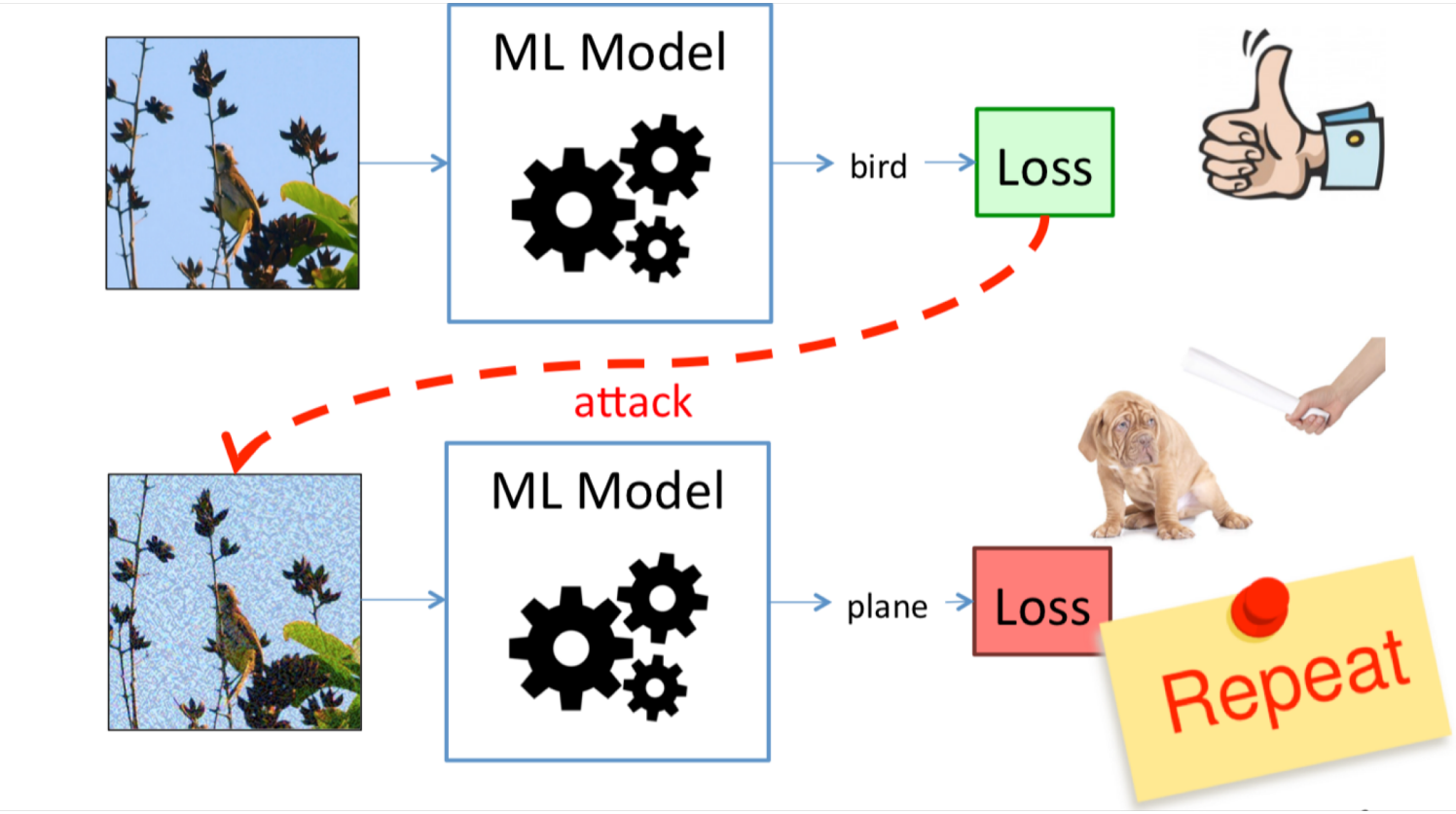
[3] A. Athalye et al., Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples, ICML 2018

Two state-of-the-art methods

- Ensemble Adversarial Training
- Adversarial Logit Pairing

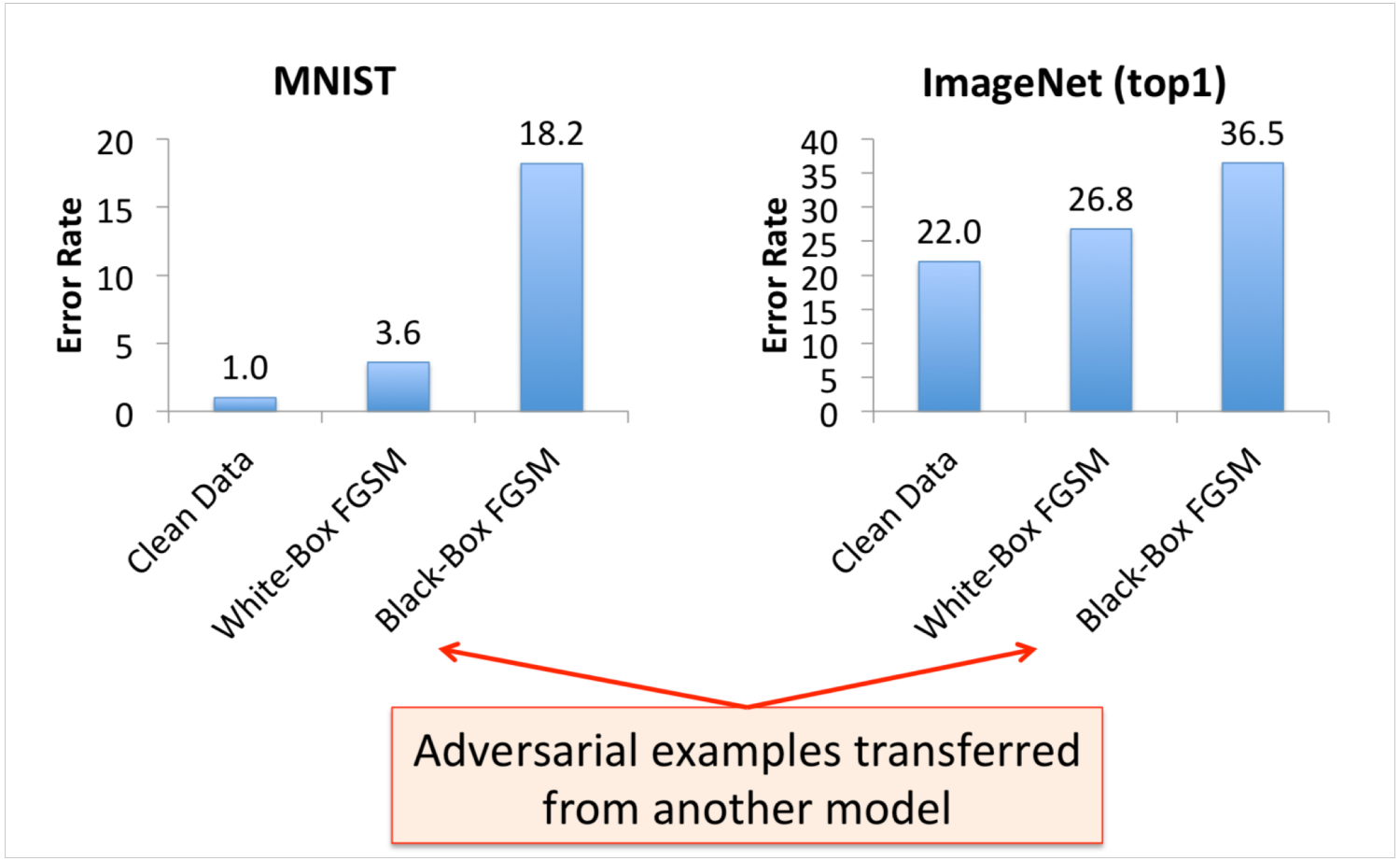
Ensemble Adversarial Training[4]

- Vanilla adversarial training:



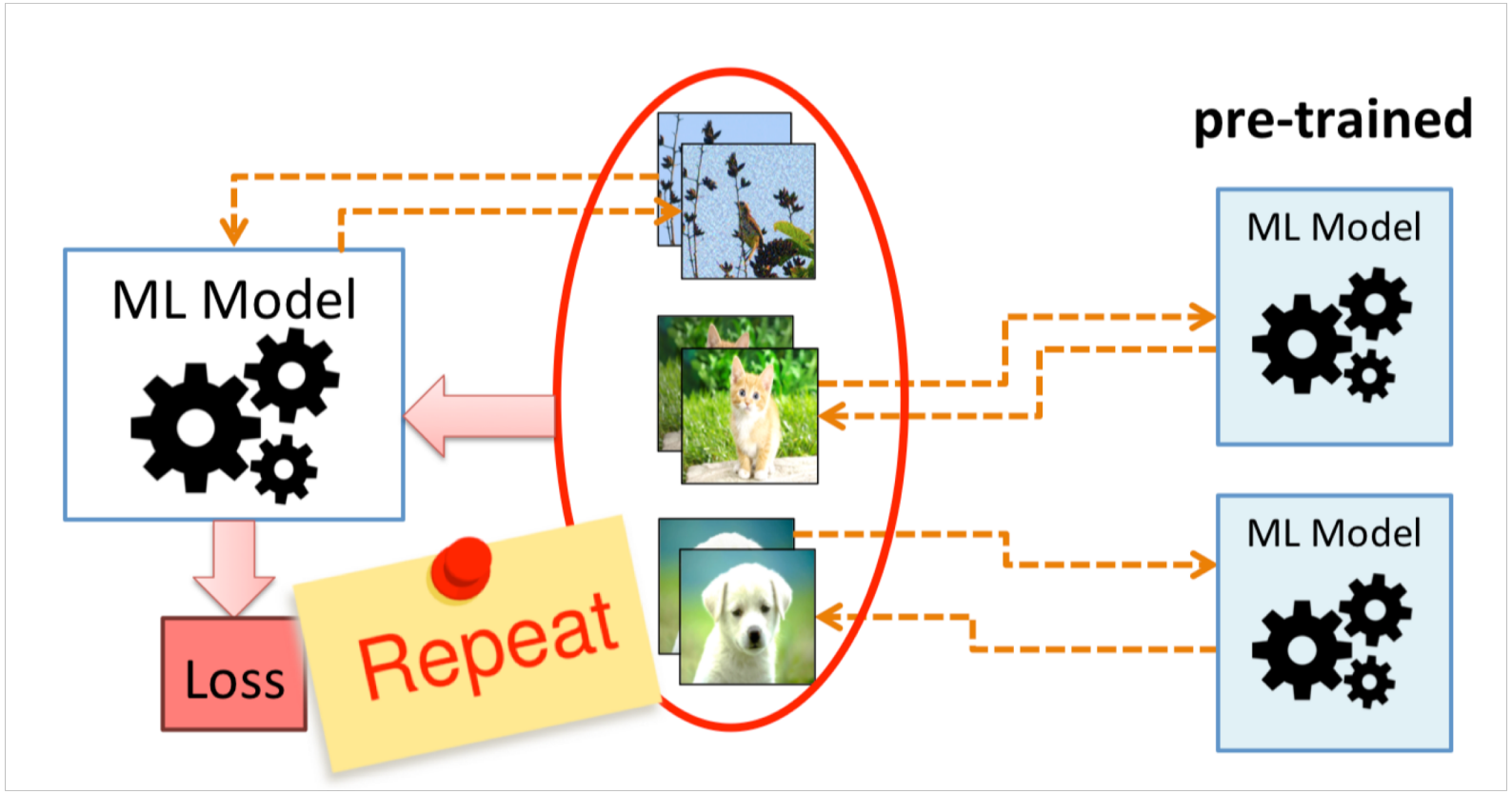
[4] F. Tramer et al., ENSEMBLE ADVERSARIAL TRAINING: ATTACKS AND DEFENSES, ICLR 2018

Vanilla Adversarial Training

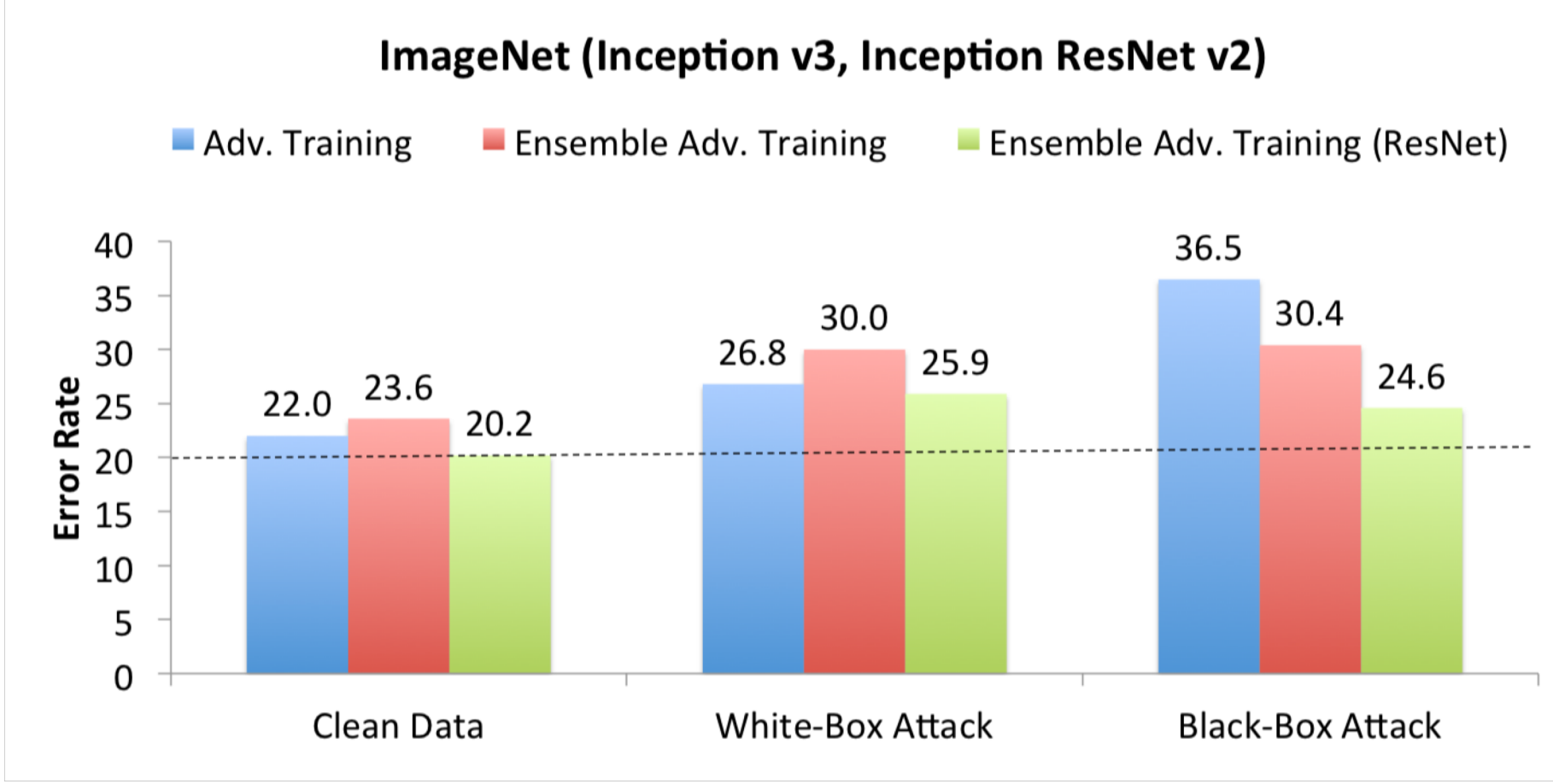


Defense – Ensemble Adversarial Training

- To *decouple* attack and defense

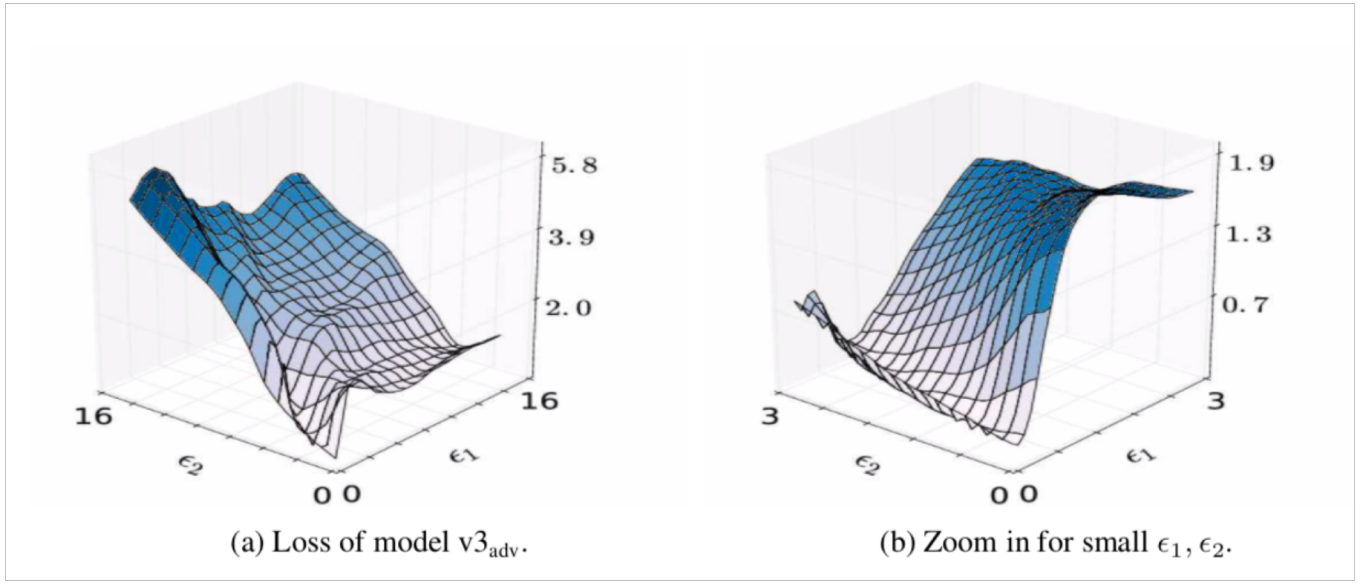


Results


ImageNet (Inception v3, Inception ResNet v2)
Adv. Training
Ensemble Adv. Training
Ensemble Adv. Training (ResNet)
Error Rate
40
35
30
25
20
15
10
5
0
22.0
23.6
20.2
26.8
30.0
25.9
36.5
30.4
24.6
Clean Data
White-Box Attack
Black-Box Attack

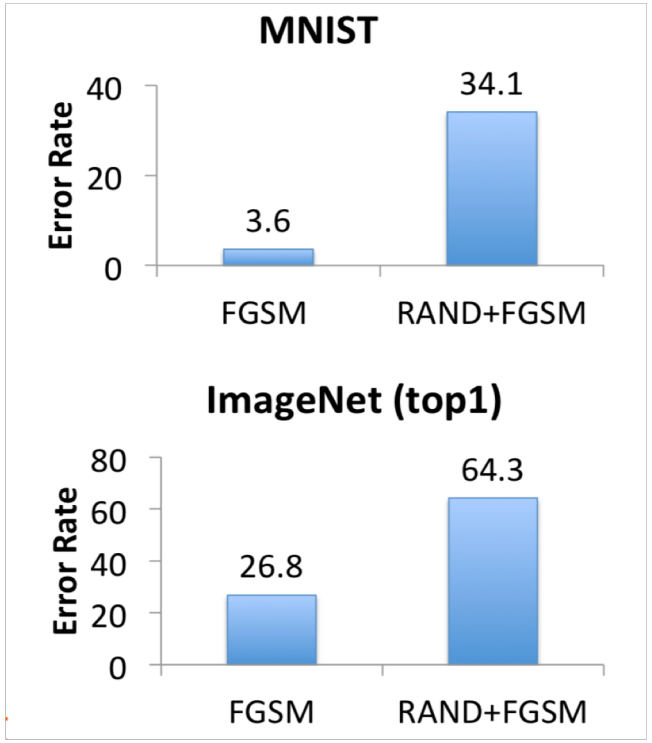
Attack – RAND + FGSM

- From previous slide, we find adversarial training can defend white-box attack but black-box attack.
- They proposed a white-box attack to beat adversarial training:
  1. Small random step
  2. Step in direction of gradient



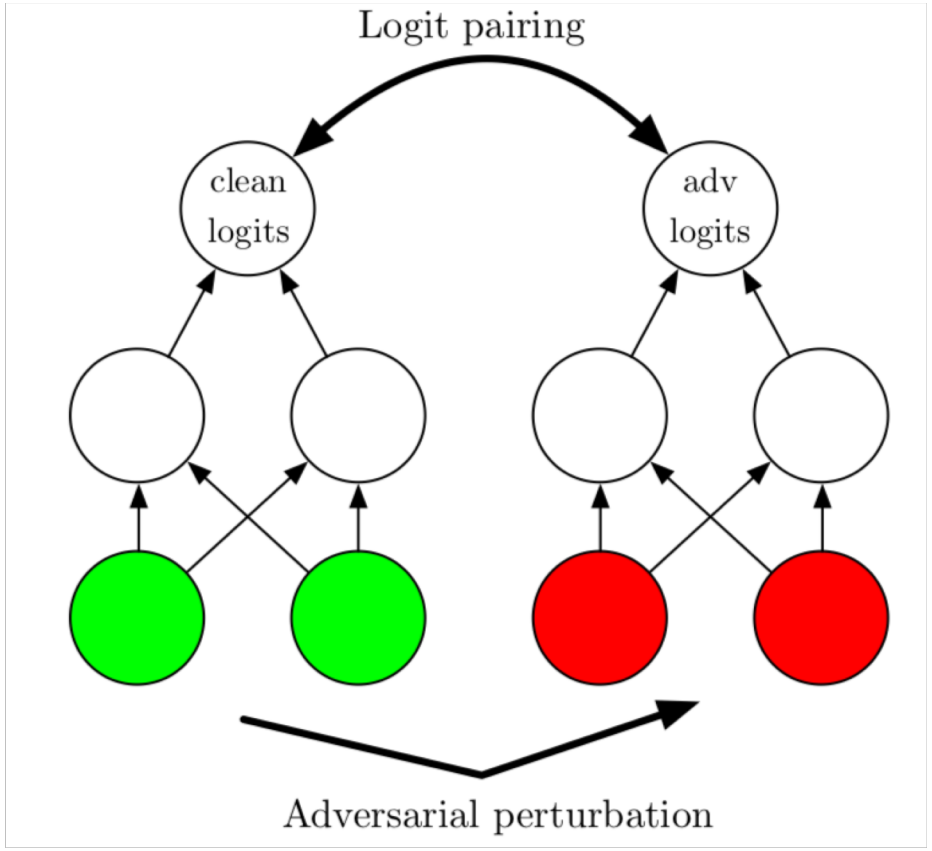
(a) Loss of model $v3_{adv}$.

(b) Zoom in for small ϵ_1, ϵ_2.



Adversarial Logit Pairing[5]

- Logit pairing:



[5] H. Kannan et al., Adversarial Logit Pairing, NIPS 2018

Adversarial Logit Pairing

- Vanilla adversarial training:

$$\theta^* = \arg\min_{\theta}[\mathcal{L}_{\text{cross entropy}}(x, t) + \mathcal{L}_{\text{cross entropy}}(x', t)]$$

- Adversarial logit pairing:

$$\theta^* = \arg\min_{\theta}[\mathcal{L}_{\text{cross entropy}}(x, t) + \mathcal{L}_{\text{cross entropy}}(x', t) + \lambda\mathcal{L}_{\text{logit pairing}}(f(x), f(x'))]$$

Results

- MNIST

Method	White Box	Black Box	Clean
M-PGD	93.2%	96.0%	98.5%
ALP	**96.4%**	**97.5%**	**98.8%**

- SVHN

Method	White Box	Black Box	Clean
M-PGD	44.4%	55.4%	**96.9%**
ALP	**46.9%**	**56.2%**	96.2%

Summary

- Today: enhanced adversarial training
- State-of-the-art but quite **heuristic**
- Proposed research direction:
  - Mathematical perspective explanation ?

THANK YOU