Gradient Masking in Machine Learning

Nicolas Papernot
Pennsylvania State University

ARO Workshop on Adversarial Machine Learning, Stanford
September 2017

Thank you to our collaborators

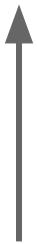
Pieter Abbeel (Berkeley)
Michael Backes (CISPA)
Dan Boneh **(Stanford)**
Z. Berkay Celik (Penn State)
Yan Duan (OpenAI)
Ian Goodfellow **(Google)**
Matt Fredrikson (CMU)
Kathrin Grosse (CISPA)

Sandy Huang (Berkeley)
Somesh Jha (U of Wisconsin)
Alexey Kurakin **(Google)**
Praveen Manoharan (CISPA)
Patrick McDaniel **(Penn State)**
Arunesh Sinha (U of Michigan)
Ananthram Swami (US ARL)
Florian Tramèr **(Stanford)**
Michael Wellman (U of Michigan)

Gradient Masking

Training

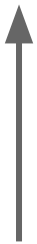
$$\text{loss}(x, y)$$



Small when prediction is correct on legitimate input

Adversarial training

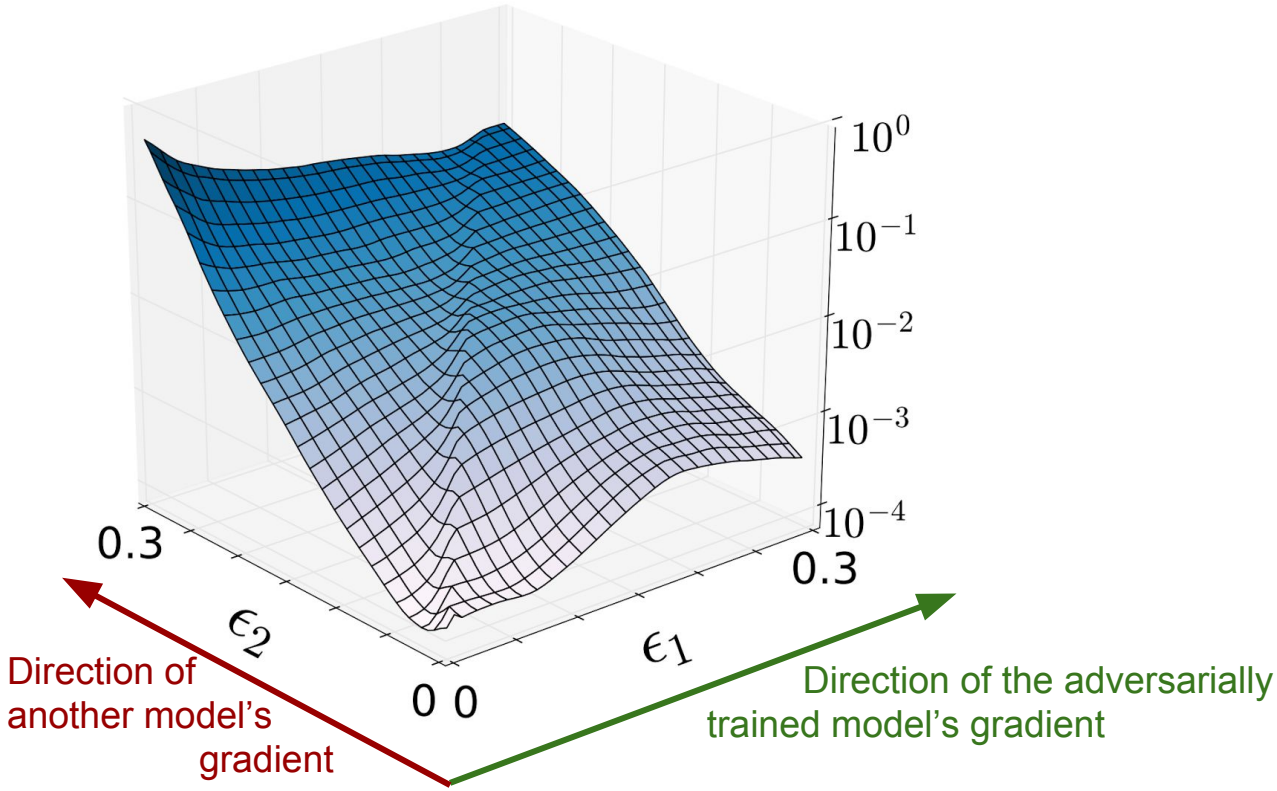
$$\text{loss}(x, y) + \text{loss}(x + \epsilon \cdot \texttt{sign}(\text{grad}), y)$$



Small when prediction is correct on legitimate input

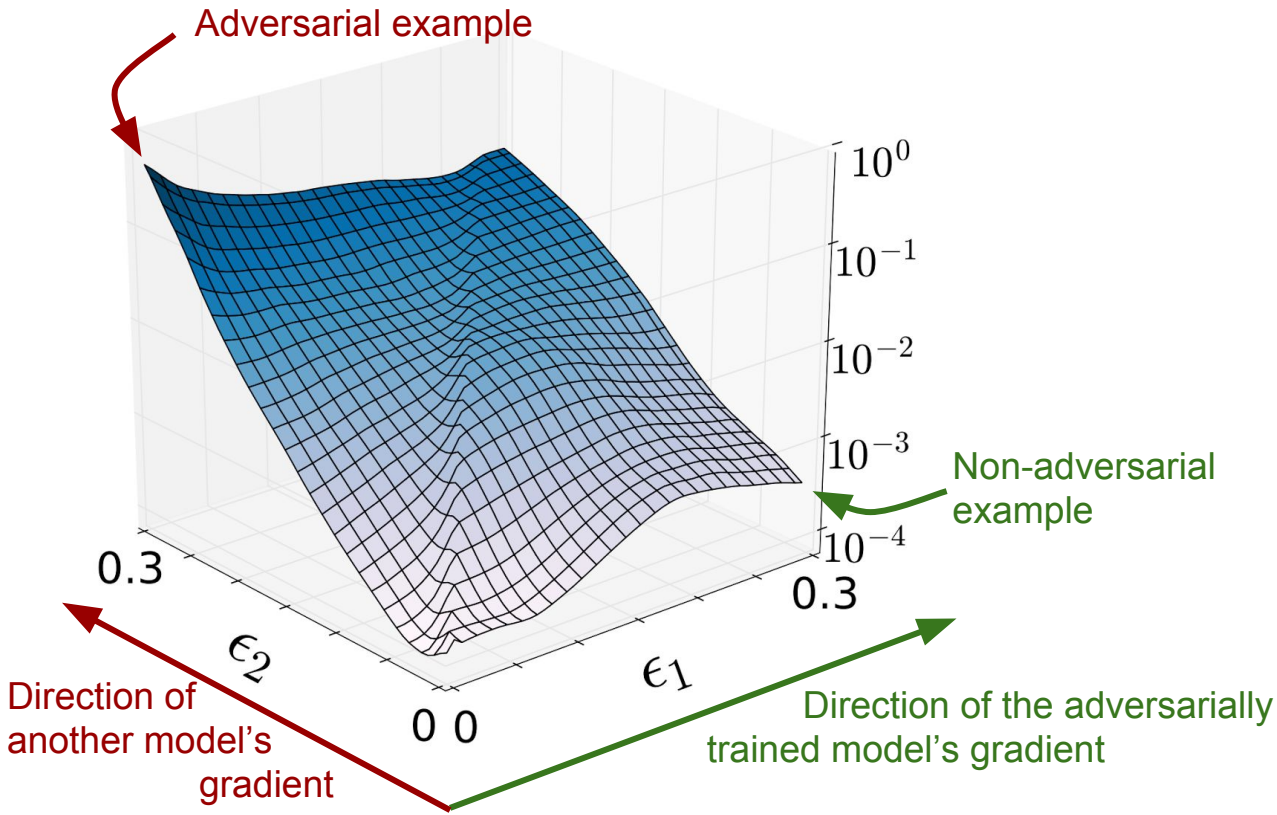
Small when prediction is correct on adversarial input

Gradient masking in adversarially trained models



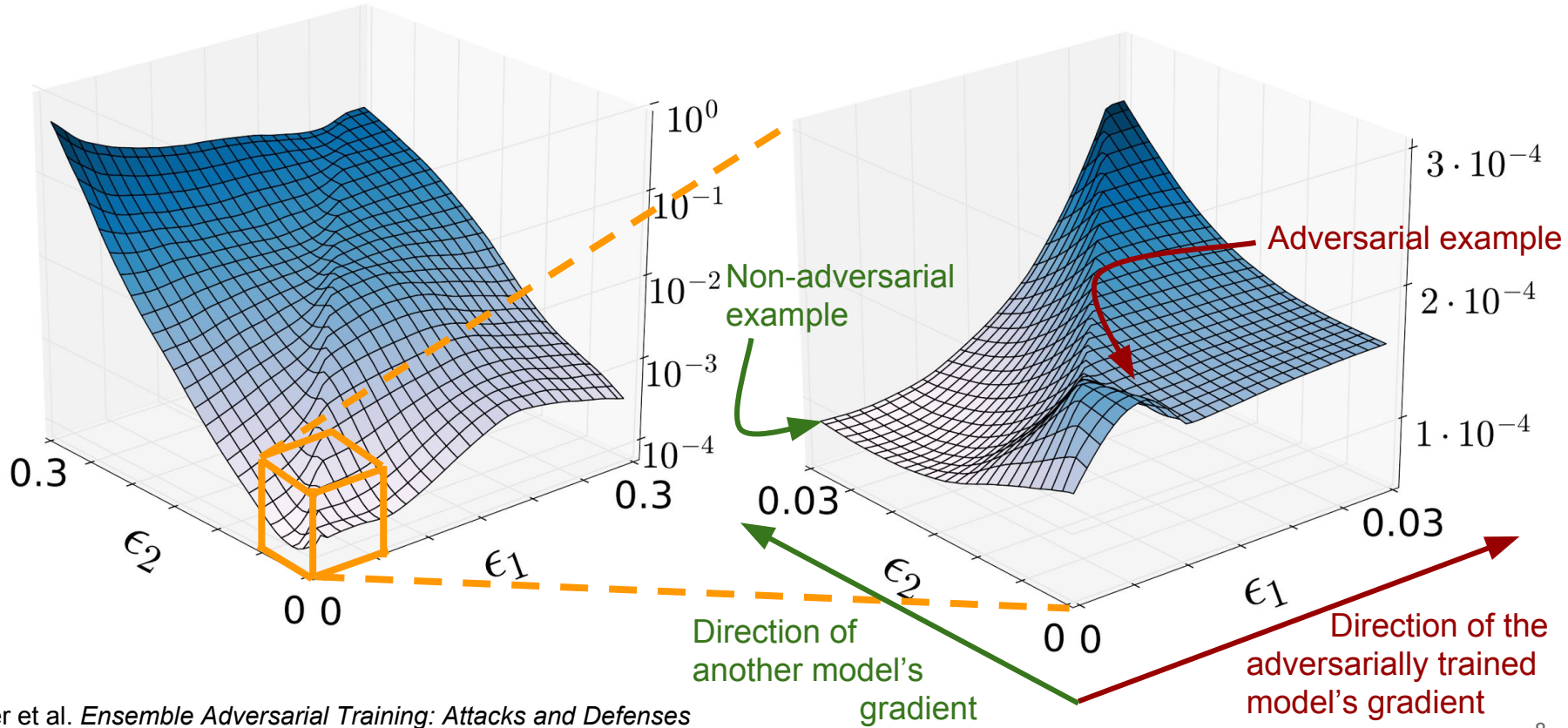
Tramèr et al. *Ensemble Adversarial Training: Attacks and Defenses*
Illustration adapted from slides by Florian Tramèr

Gradient masking in adversarially trained models



Tramèr et al. *Ensemble Adversarial Training: Attacks and Defenses*
Illustration adapted from slides by Florian Tramèr

Gradient masking in adversarially trained models



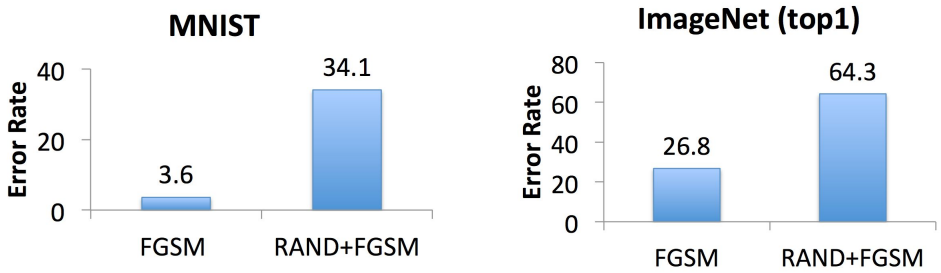
Tramèr et al. *Ensemble Adversarial Training: Attacks and Defenses*
Illustration adapted from slides by Florian Tramèr

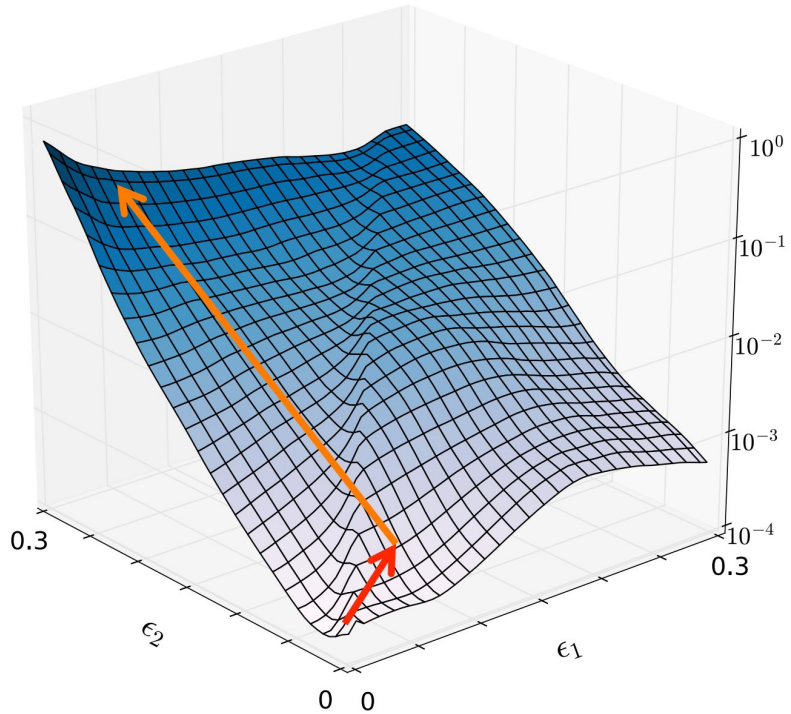
Evading gradient masking (1)

Threat model: white-box adversary

Attack:

(1) Random step (of norm alpha)
(2) FGSM step (of norm eps - alpha)





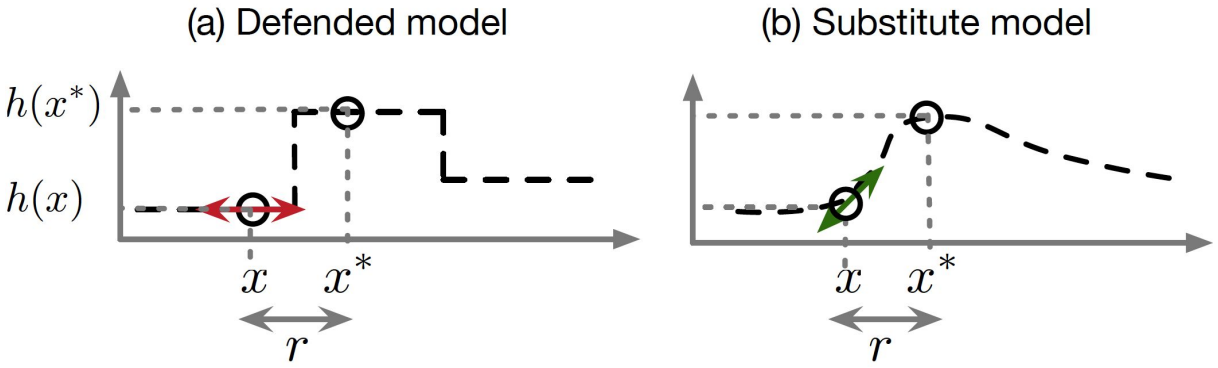
Tramèr et al. *Ensemble Adversarial Training: Attacks and Defenses*
Illustration adapted from slides by Florian Tramèr

Evading gradient masking (2)

Threat model: black-box attack

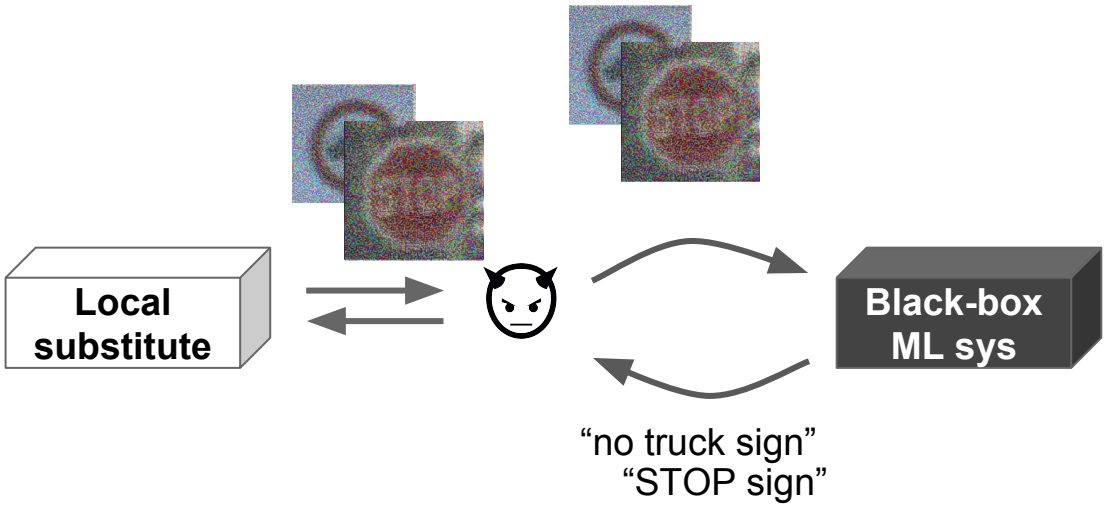
Attack:

(1) Learn substitute for defended model
(2) Find adversarial direction using substitute



Papernot et al. *Practical Black-Box Attacks against Machine Learning*
Papernot et al. *Towards the Science of Security and Privacy in Machine Learning*

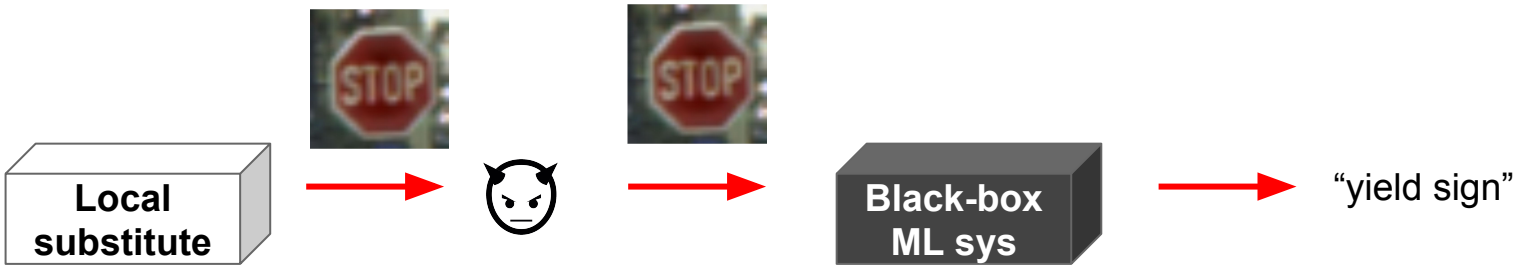
Attacking black-box models



(1) The adversary queries the remote ML system with synthetic inputs to learn a local substitute.

Papernot et al. *Practical Black-box Attacks Against Machine Learning*

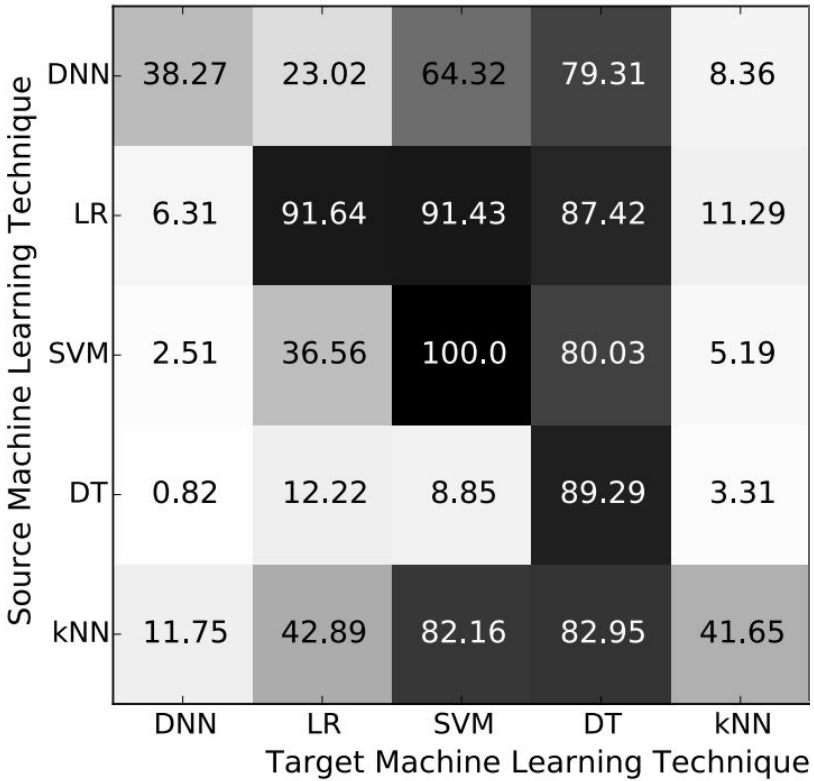
Attacking black-box models



(2) The adversary uses the local substitute to craft adversarial examples.

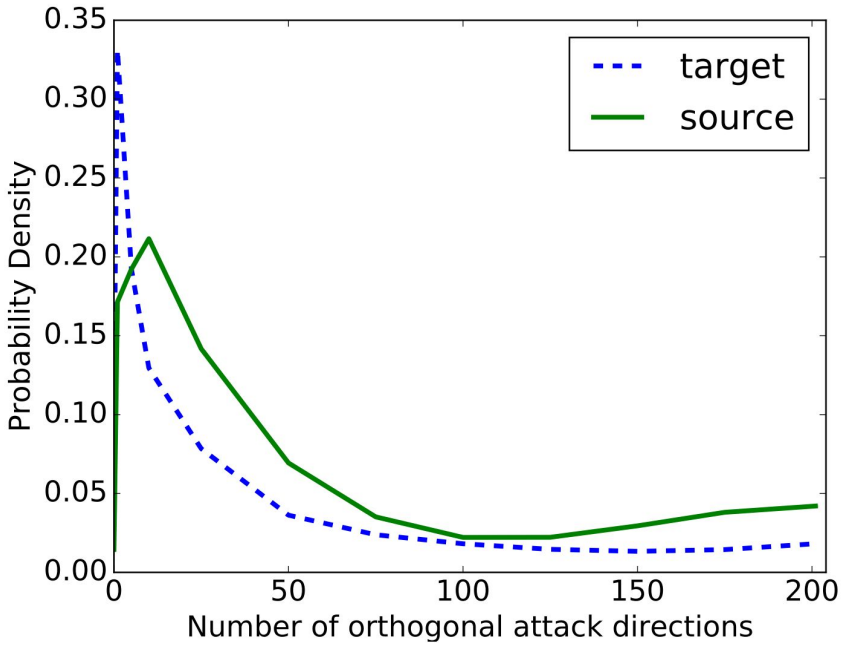
Papernot et al. *Practical Black-box Attacks Against Machine Learning*

Adversarial example transferability



Papernot et al. *Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples*

Large adversarial subspaces enable transferability

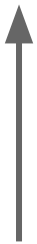


On average: 44 orthogonal directions -> 25 transfer

Tramèr et al. *The Space of Transferable Adversarial Examples*

Adversarial training

$$\text{loss}(x, y) + \text{loss}(x + \epsilon \cdot \texttt{sign}(\text{grad}), y)$$



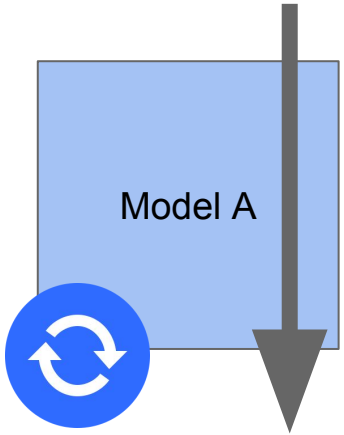
Small when prediction is correct on legitimate input

Small when ~~prediction is correct on adversarial input~~ **gradient is not adversarial**

Ensemble Adversarial Training

Ensemble adversarial training

Intuition: present adversarial gradients from multiple models during training

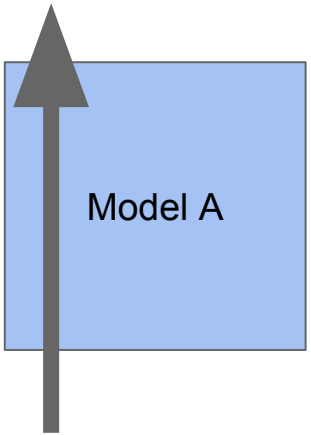


Adversarial gradient

$$\text{loss}(x, y) + \text{loss}(x + \epsilon \cdot \texttt{sign}(\text{grad}), y)$$

Ensemble adversarial training

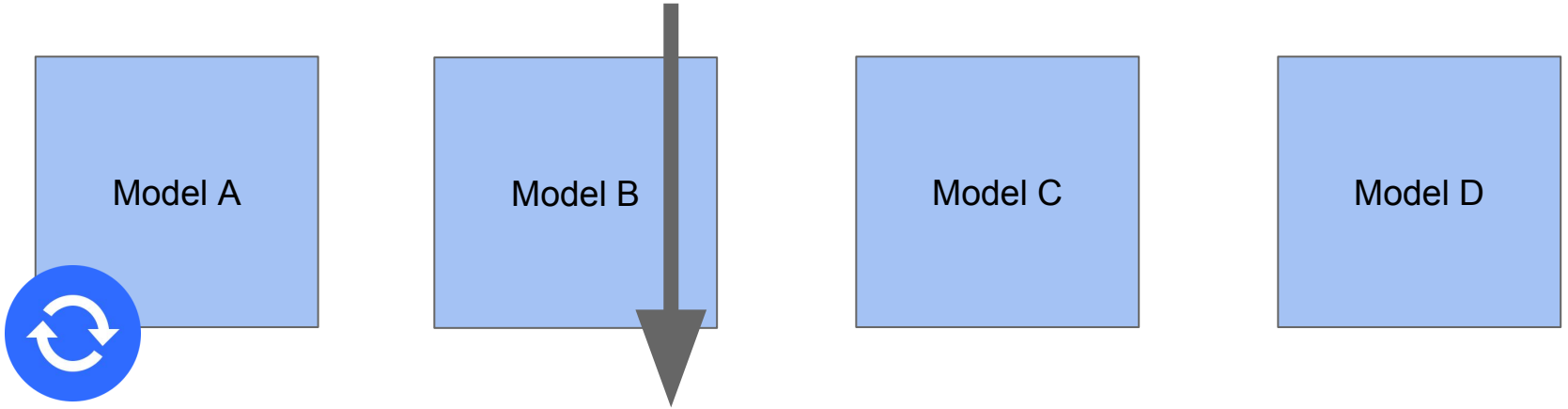
Intuition: present adversarial gradients from multiple models during training



Inference

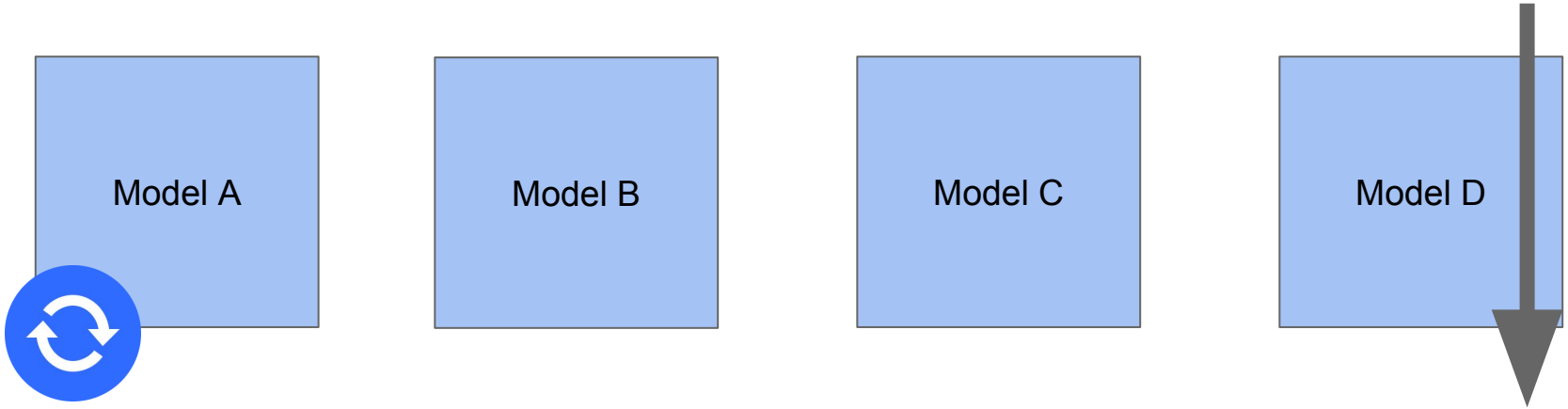
Ensemble adversarial training

Intuition: present adversarial gradients from multiple models during training



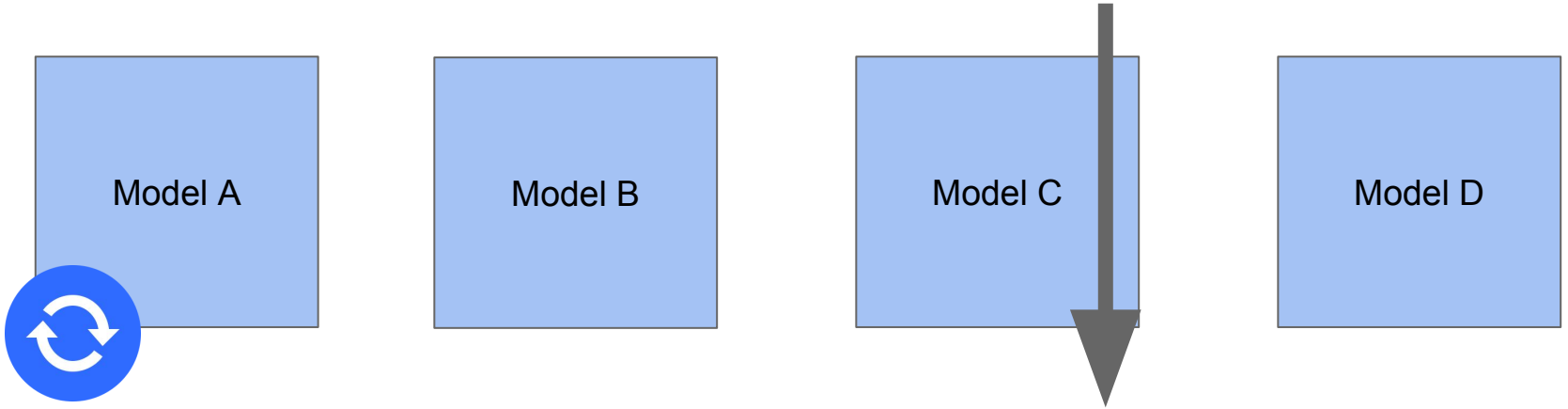
Ensemble adversarial training

Intuition: present adversarial gradients from multiple models during training



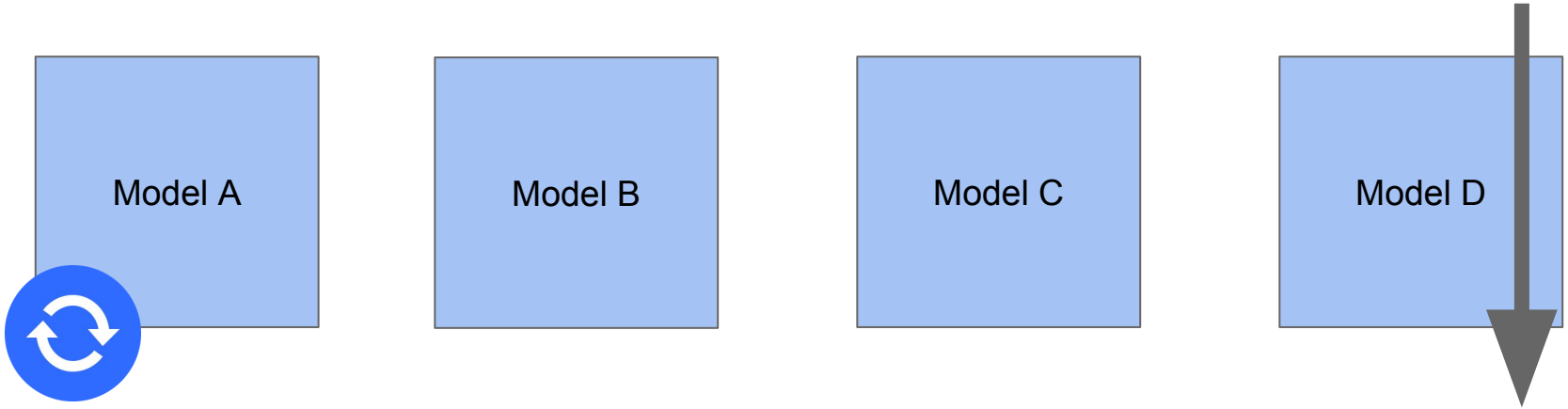
Ensemble adversarial training

Intuition: present adversarial gradients from multiple models during training



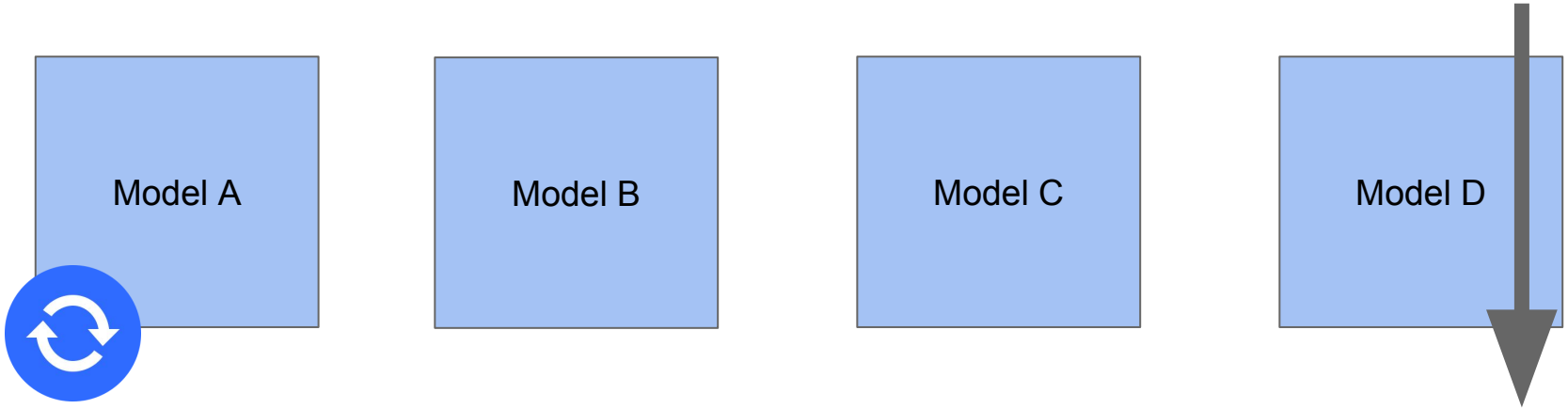
Ensemble adversarial training

Intuition: present adversarial gradients from multiple models during training



Ensemble adversarial training

Intuition: present adversarial gradients from multiple models during training

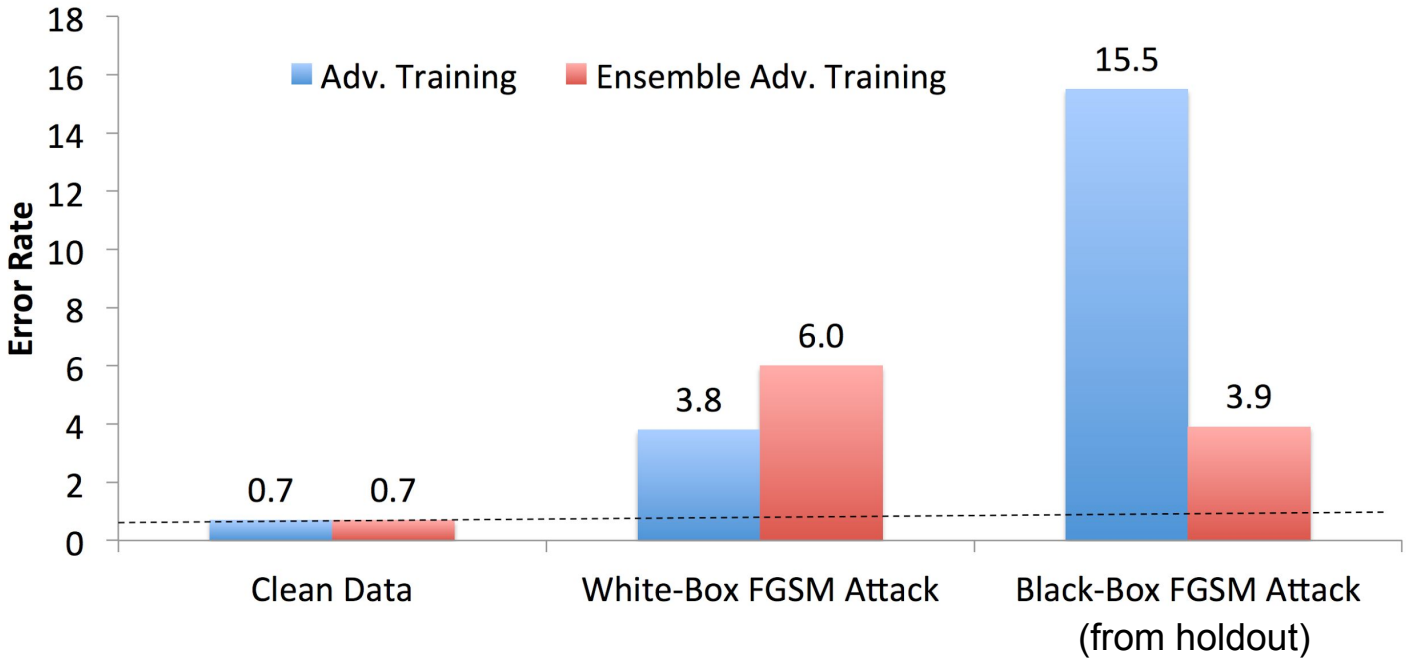


Ensemble adversarial training

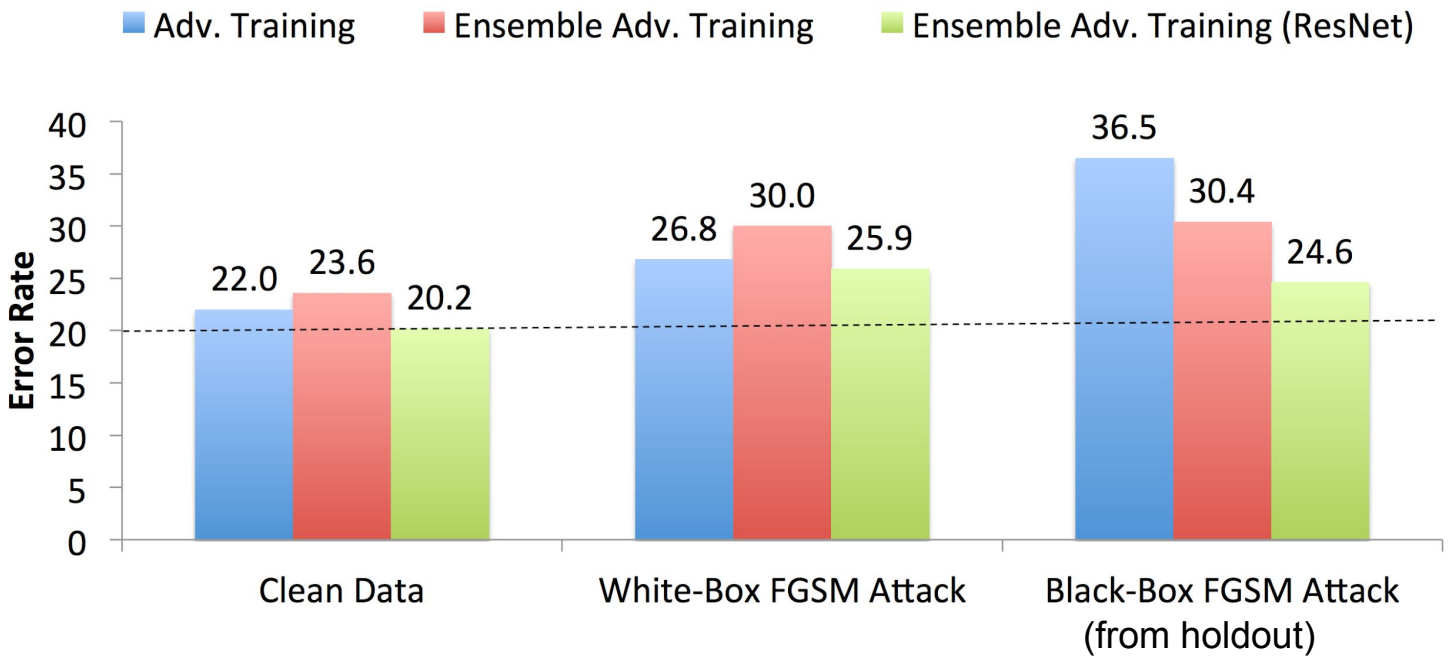
Intuition: present adversarial gradients from multiple models during training

Inference

Experimental results on MNIST



Experimental results on ImageNet



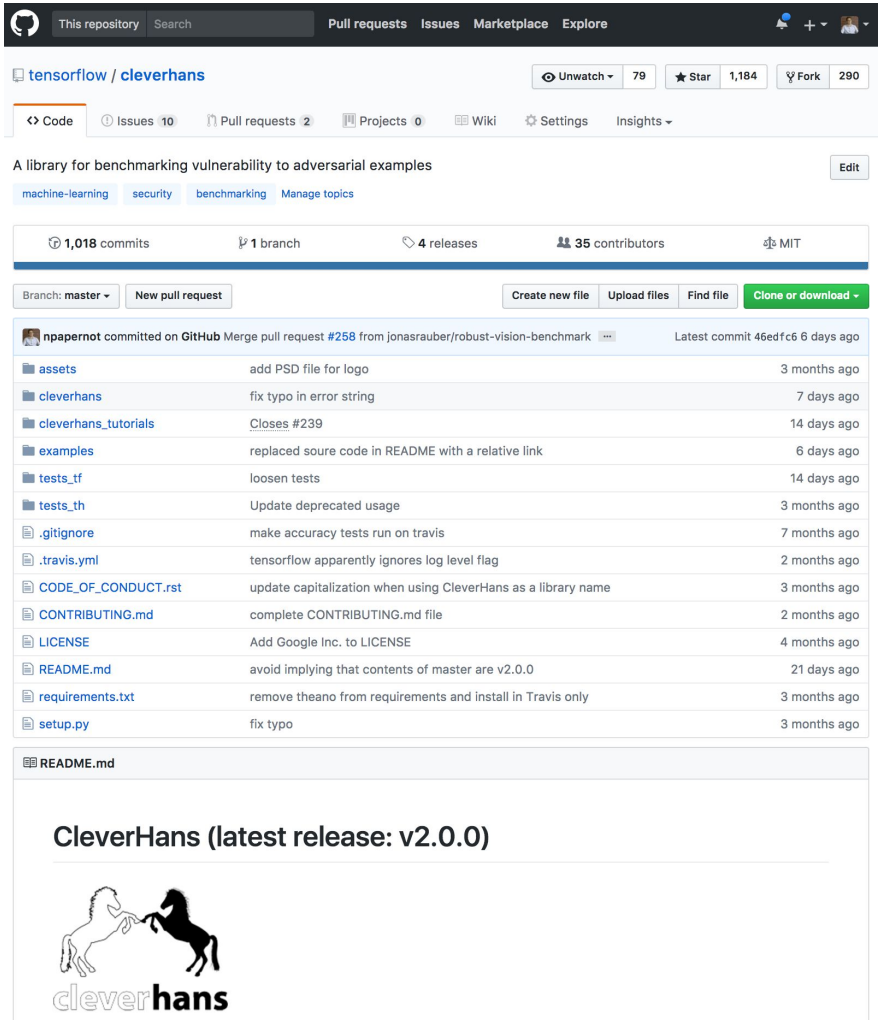
Reproducible Adversarial ML research with CleverHans

CleverHans library guiding principles

1. Benchmark reproducibility

2. Can be used with any TensorFlow model

3. Always include state-of-the-art attacks and defenses



Growing community

1.1K+ stars

290+ forks

35 contributors

Adversarial examples represent

worst-case distribution drifts

[DDS04] Dalvi et al. *Adversarial Classification* (KDD)

Adversarial examples are a *tangible* instance of hypothetical AI safety problems

Image source: http://www.nerdist.com/wp-content/uploads/2013/07/Space-Odyssey-4.jpg

nicolas@papernot.fr
www.cleverhans.io
@NicolasPapernot

Thank you for listening! Get involved at:

github.com/tensorflow/cleverhans

This research was funded by: ARL Google