Probing the Structure of Deep Neural Networks with Universal Adversarial Perturbations

Sanjit Bhat (Acton-Boxborough RHS), Mentor: Dimitris Tsipras (MIT)
PRIMES Conference, May 18, 2019

Acknowledgements

Thank you to:

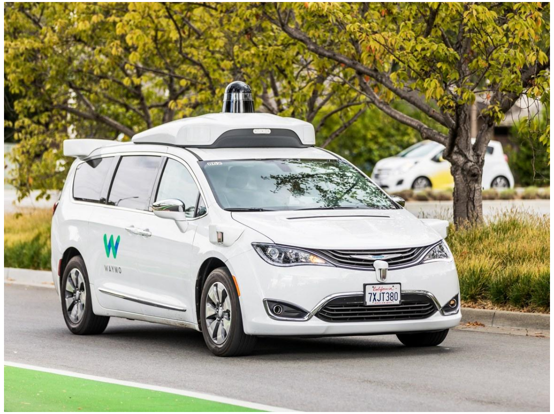
- My parents
- Dimitris Tsipras, for the useful discussions and guidance
- The Madry Lab, for making me feel at home at MIT
- Prof. Srini Devadas, for the PRIMES CS track
- Dr. Slava Gerovitch, for the PRIMES program

Introduction

Deep Learning (DL) can surpass humans

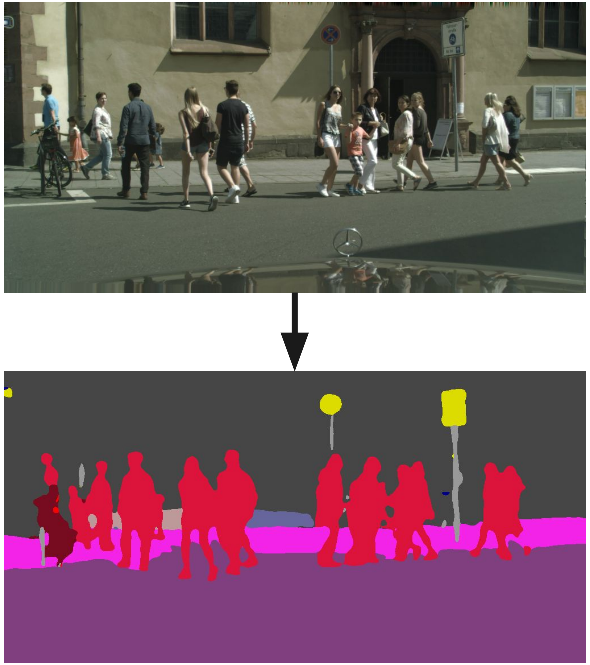
Input sentence:	*Translation (PBMT):*	*Translation (GNMT):*	*Translation (human):*
李克強此行將啟動中加總理年度對話機制，與加拿大總理杜魯多舉行兩國總理首次年度對話。	Li Keqiang premier added this line to start the annual dialogue mechanism with the Canadian Prime Minister Trudeau two prime ministers held its first annual session.	Li Keqiang will start the annual dialogue mechanism with Prime Minister Trudeau of Canada and hold the first annual dialogue between the two premiers.	Li Keqiang will initiate the annual dialogue mechanism between premiers of China and Canada during this visit, and hold the first annual dialogue with Premier Trudeau of Canada.

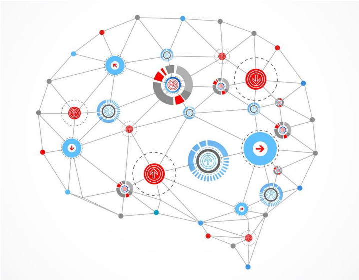
DL in security-critical applications



Is DL ready for this?

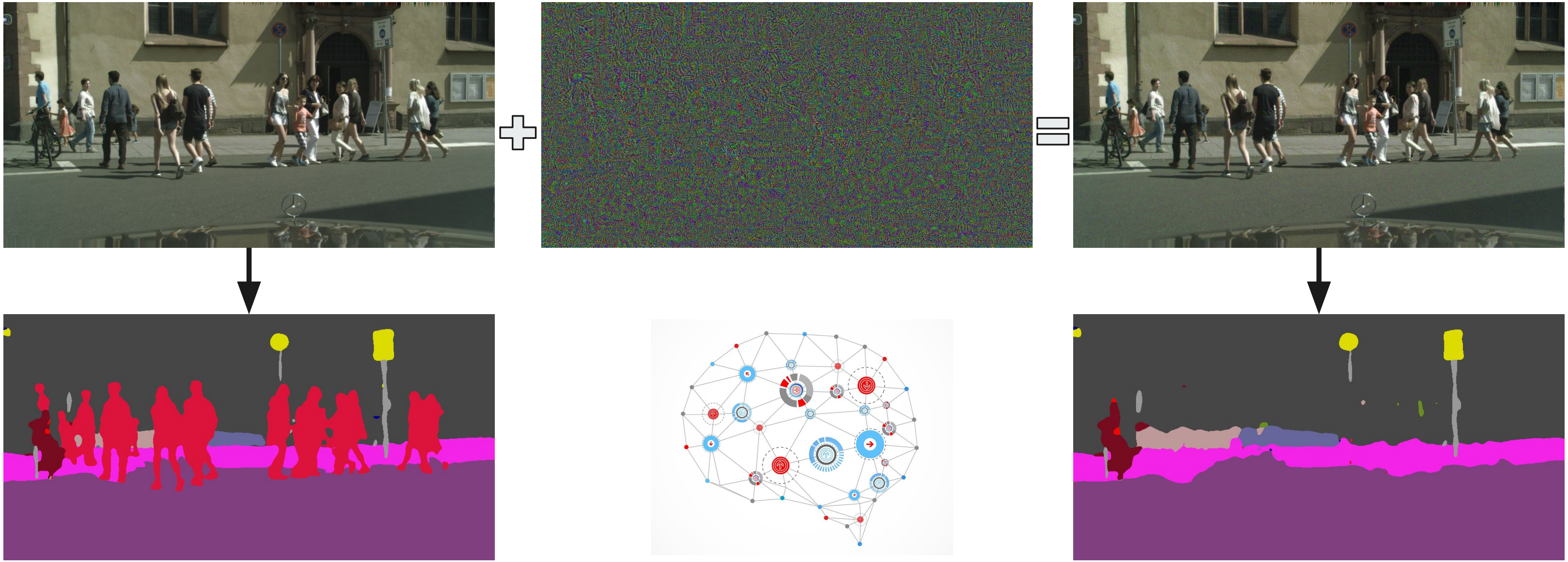
Deep Neural Network (DNN) - Natural Setting





V. Fischer, M. Kumar, J. Metzen, T. Brox
“Adversarial Examples for Semantic Image Segmentation”

DNN - Adversarial Setting



Why do we need robust DNNs?

Robustness to real-world perturbs

- Some natural perturbations (e.g., rain) can trick classifiers
- Train models that are more reliable in the natural world

Alignment with human intelligence

- Goal of ML: Make intelligent systems
- Most humans wouldn't get fooled, but these systems do

Background

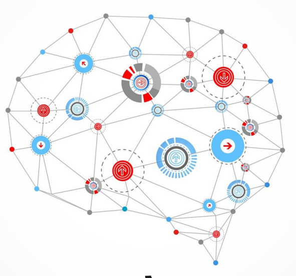
How do we train robust DNNs?

Adversarial Training - A robust training method

$$\min_{\theta}\left[\mathop{\mathbb{E}}_{(x,y)\sim\widehat{\mathcal{D}}}\mathcal{L}(x,y,\theta)\right]$$

Natural Training Set

$$\max_{\delta\in\mathcal{S}}\mathcal{L}(x+\delta,y,\theta)$$



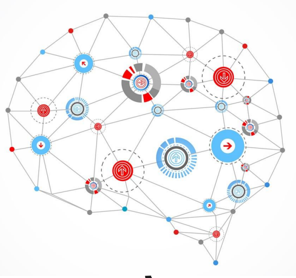
Model Parameters

Adversarial Training - A robust training method

$$\min_{\theta} \mathop{\mathbb{E}}_{(x,y)\sim\widehat{\mathcal{D}}} \left[\max_{\delta \in \mathcal{S}} \mathcal{L}(x+\delta, y, \theta) \right]$$

Adversarial Training Set

$$\max_{\delta \in \mathcal{S}} \mathcal{L}(x+\delta, y, \theta)$$



Model Parameters

Universal Adversarial Perturbations (UAPs)

Regular Adversarial Perturbations

$$\max_{\delta \in \mathcal{S}} \mathcal{L}(x + \delta, y, \theta)$$

- Image-specific (one perturbation per image)
- Stronger, more targeted

UAPs

$$\max_{\delta \in \mathcal{S}} \left[\sum_{i=1}^{n} \mathcal{L}(x_i + \delta, y, \theta) \right]$$

- Class-specific (one perturbation for all images in a particular class)
- More general
- Location-invariant

Goal: Use UAPs to study the general dynamics of Adversarial Training

Methodology

UAP Generation

Averaging

- Simplest, most obvious method

Singular Value Decomposition (SVD)

- Goal: Explain away variance
- Inputs: Data
- Outputs: Vectors that explain the most variance in data (eigenvectors) and their associated eigenvalues

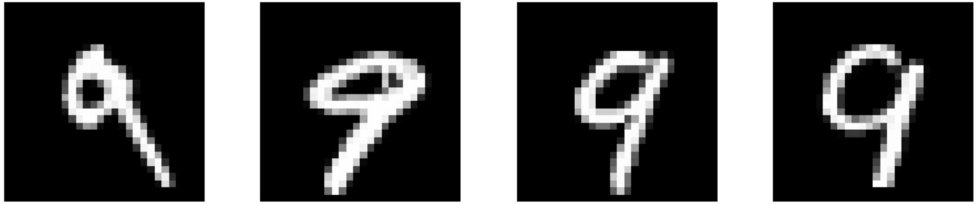
UAP Generation Cont.

- Pre-trained natural and adversarial models from Madry et al.
- UAPs generated and evaluated on MNIST (handwriting recognition) and CIFAR-10 (image recognition) test sets
- Focus on adversaries bounded in L2 norm - more interpretable perturbations

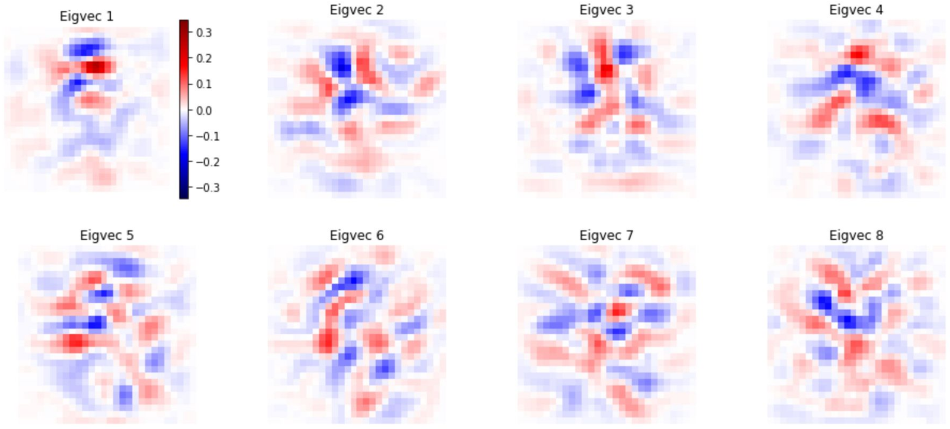
Experiments

Adversarial Training Induces More Human-Interpretable Features

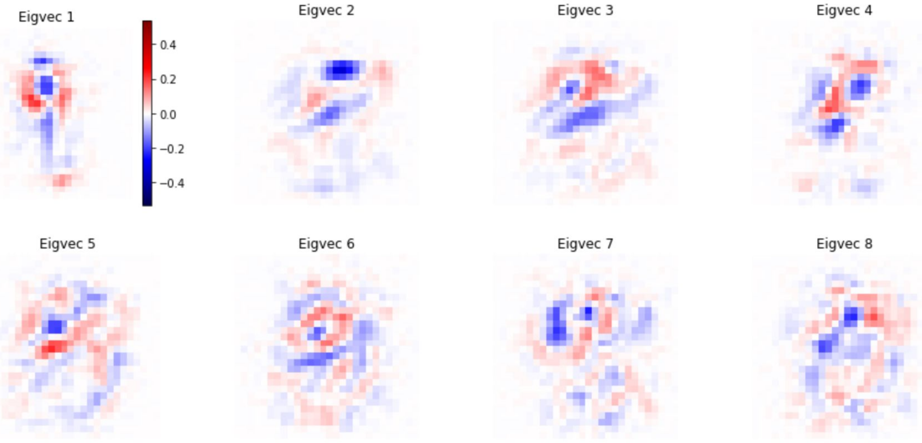
MNIST



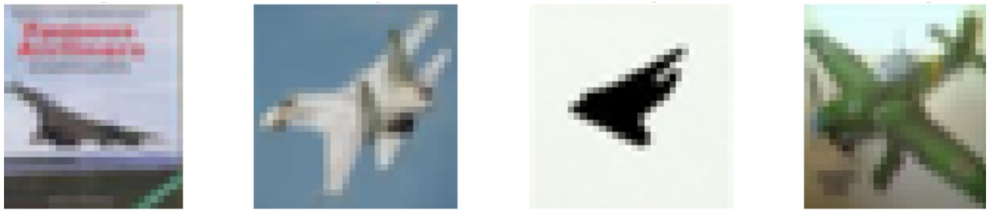
Naturally Trained



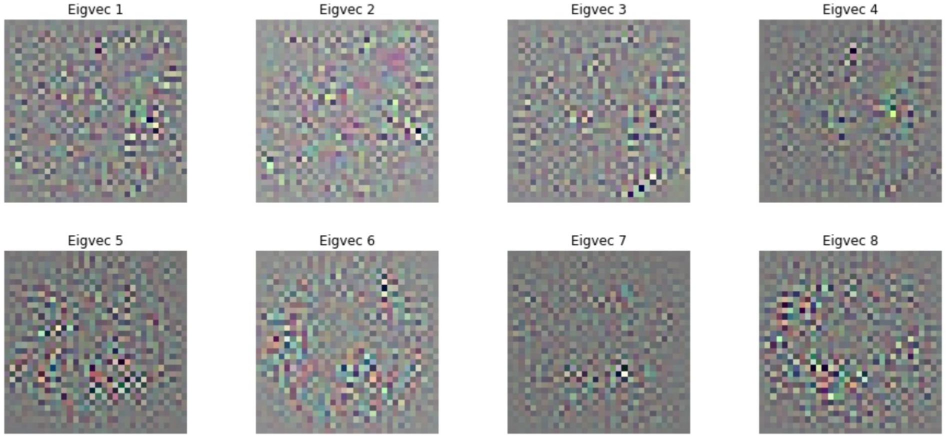
Adversarially Trained



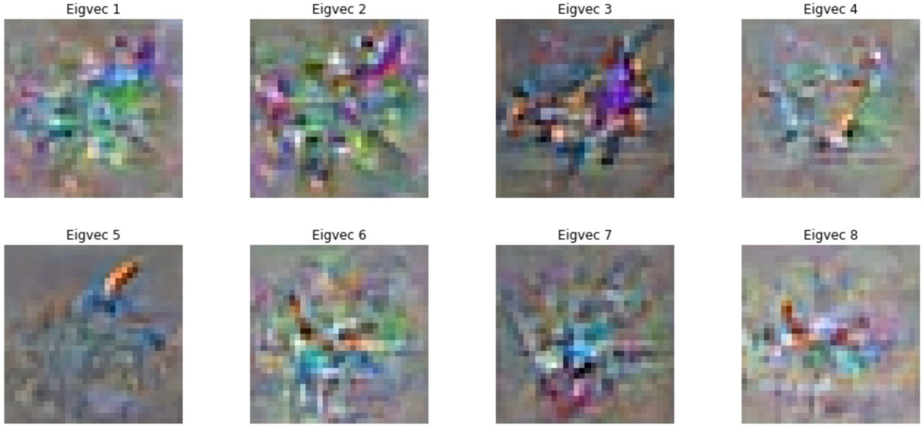
CIFAR-10



Naturally Trained



Adversarially Trained



Multiple UAP Directions Exist for MNIST

The Eigenvalue Spectra

MNIST

2.7	2.5	2.1	2.0	1.8

- No large drop
- Multiple universal directions
- Cause: Linear separability

CIFAR-10

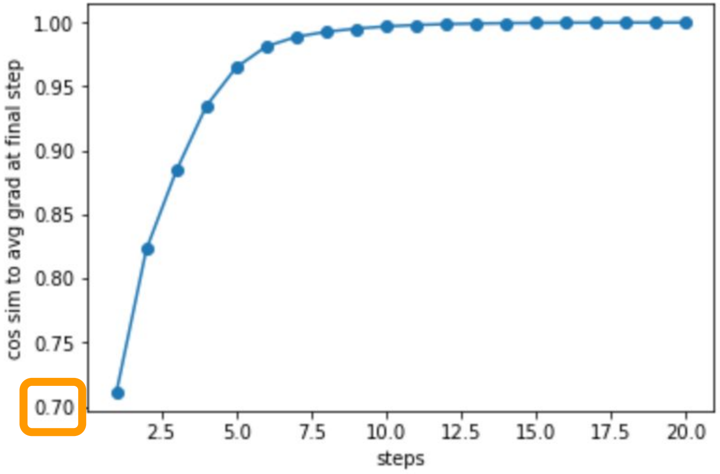
54.4	4.5	4.1	3.4	3.0

- Order of magnitude drop
- One main universal direction
- Cause: No linear separability, images mesh together

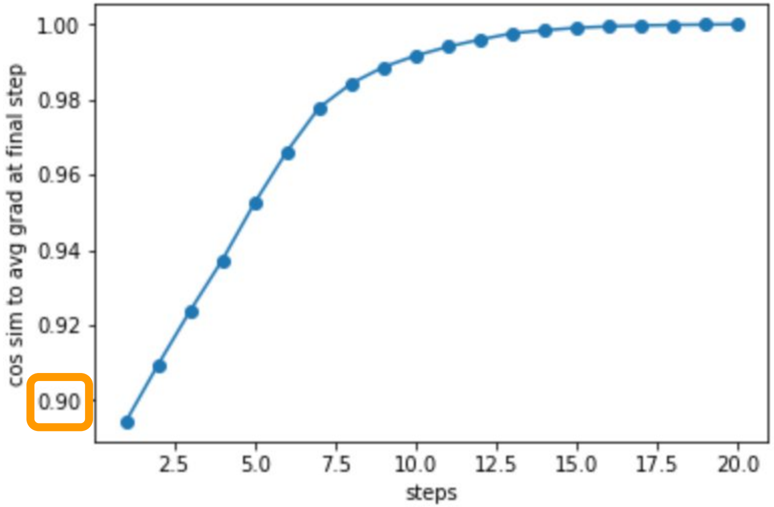
Adversarial Training Causes Local Loss Landscape Smoothening

Optimization Trajectories - MNIST

Naturally Trained

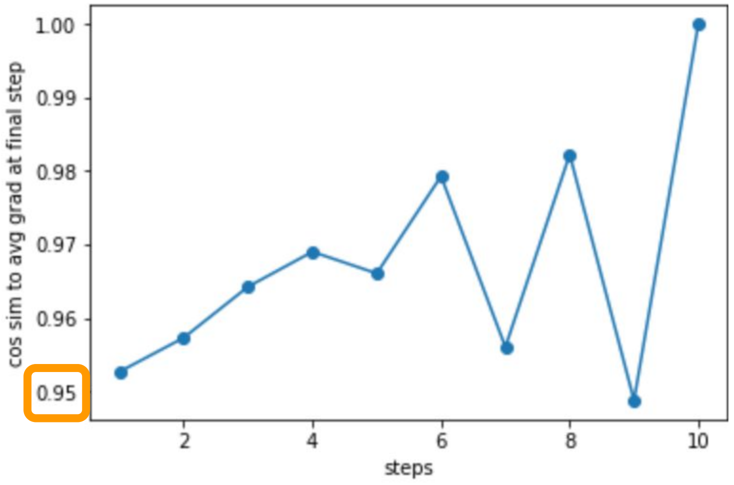


Adversarially Trained

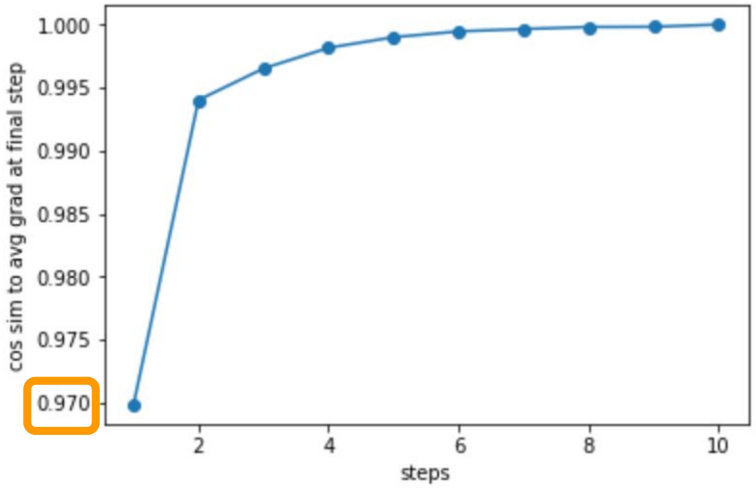


Optimization Trajectories - CIFAR-10

Naturally Trained

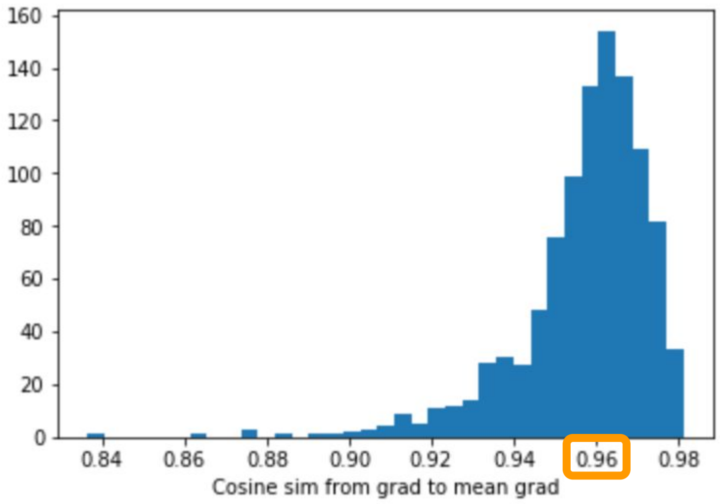


Adversarially Trained

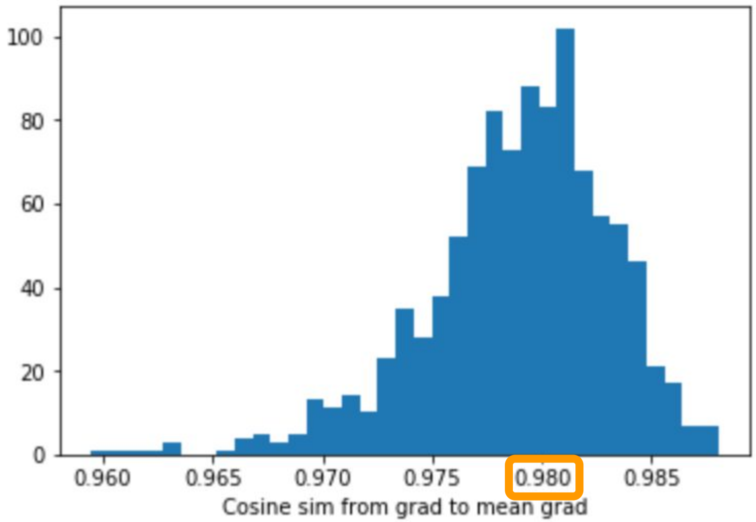


Distribution of Cos Similarities for Single Image

Naturally Trained



Adversarially Trained



Thank You!

Questions?