Carnegie Mellon University
Software Engineering Institute

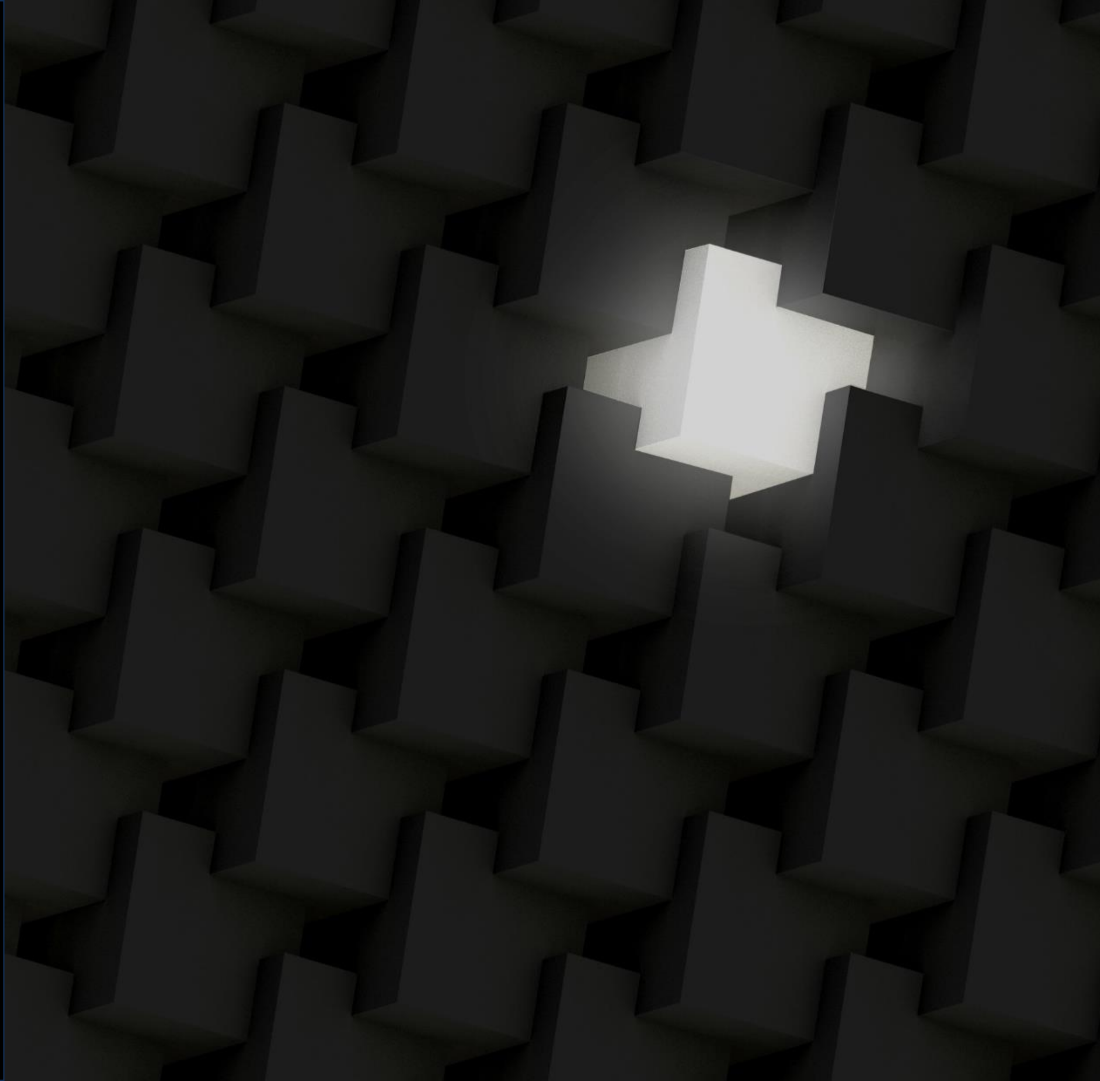
RESEARCH REVIEW 2020

Train, but Verify: Towards Practical AI Robustness

Presenter: Dr. Nathan VanHoudnos (van-HOD-ness)

SEI Team Members: Matt Churilla, Jon Helland, Grace Lewis, Nathan VanHoudnos, and Oren Wright

Carnegie Mellon University Team Members: Lujo Bauer, Matt Fredrickson, Aymeric Fromherz, Klas Leino, and Bryan Parno



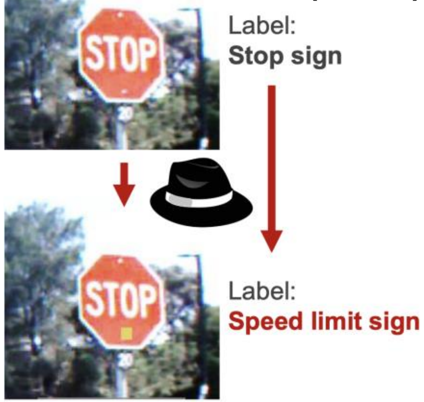
This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

DM20-0932

Beieler (2018): An attacker Can Make an ML System…

Learn the Wrong Thing

Gu et al. (2017)



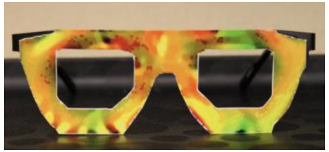
Do the Wrong Thing

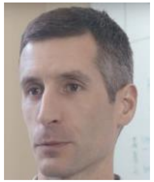
Sharif et al. (2016)

…

Carson Milla

Design Glasses





??? Milla

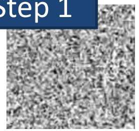
Reveal the Wrong Thing

Fredrickson et al. (2016)

…

Person A Person Z

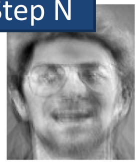
Step 1



P(A) = 0.03
P(B) = 0.04
…
P(Z) = 0.02

… …

Step N



P(A) = 0.01
P(B) = 0.00
…
P(Z) = 0.97

Train, but Verify

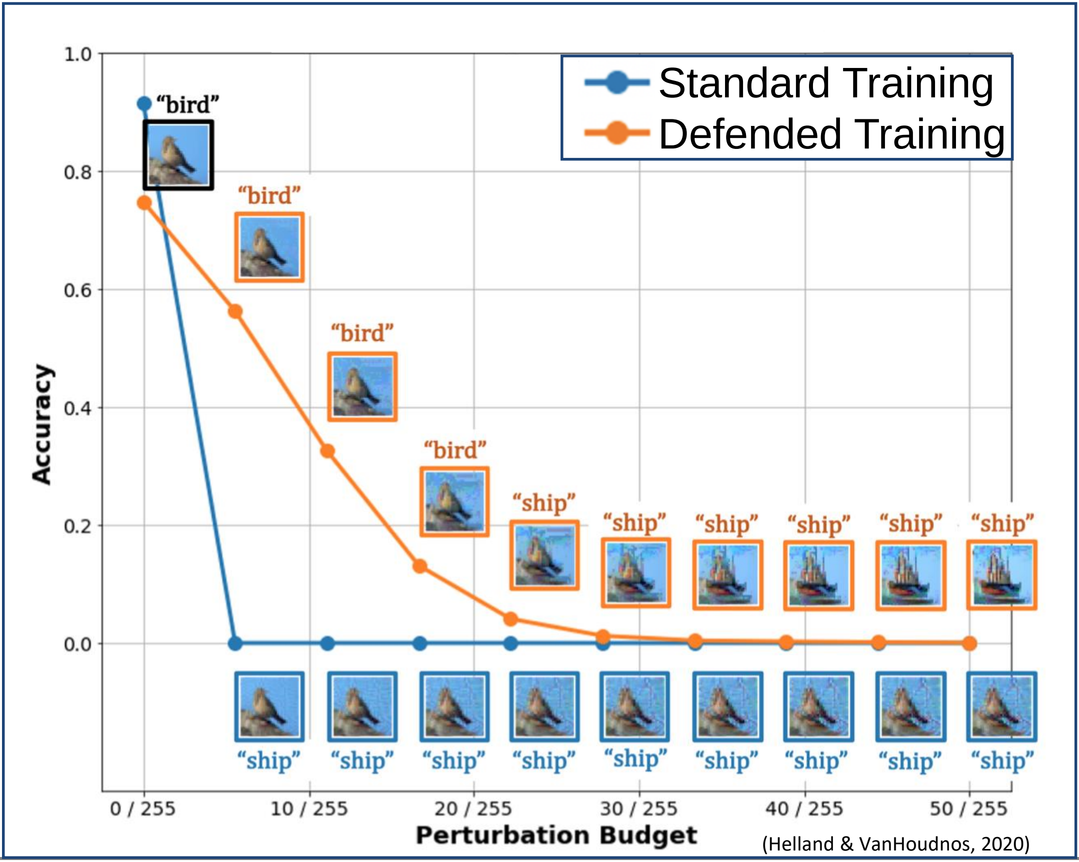
Train \ Verify	Verify "Learn" Policy	Verify "Do" Policy	Verify "Reveal" Policy
Train to enforce "learn" policy	IARPA TrojAI DARPA GARD		
Train to enforce "do" policy		DARPA GARD	?
Train to enforce "reveal" policy			NGA GURU

Problem

- AI promises capability for the DoD, but today is untrustworthy.
- Most defensive work focuses on one security policy, but the DoD has wider concerns.
  - What if a system makes high stakes decisions (do policy) and is trained on sensitive data (reveal policy)?

Defenses for Do Policies Reveal Information about the Data

Seed Image



Defended Example

Standard Example

First described by Tsipras et al. (2017).

Why does this happen?

Defenses for Do Policies Reveal Information about the Data

Consider a model that

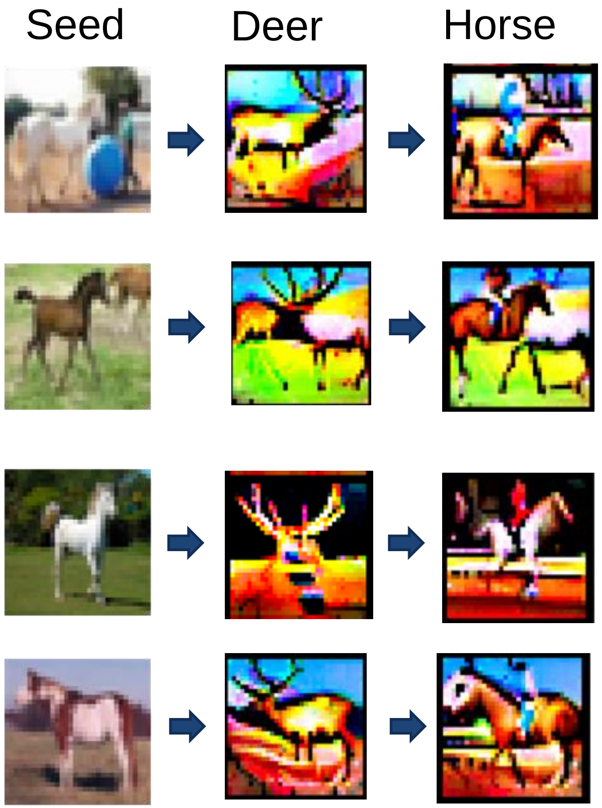
- has high stakes decisions (do)
- uses sensitive data (reveal)

The attacker’s goal is to reveal

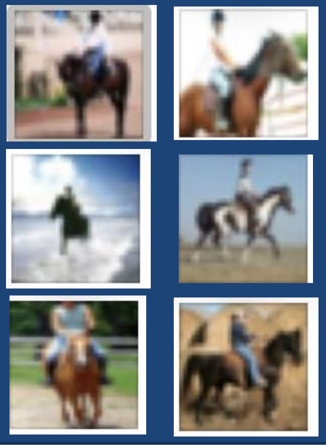
- How were the horse examples collected for CIFAR-10?

A novel use of a known attack:

- Generate adversarial examples against a defended model.



Recovers the presence of riders in the CIFAR 10 horse class (about 20% of examples)



Train, but Verify

Train \ Verify	Verify "Learn" Policy	Verify "Do" Policy	Verify "Reveal" Policy
Train to enforce "learn" policy	IARPA TrojAI DARPA GARD		
Train to enforce "do" policy		DARPA GARD	Helland & VanHoudnos (2020)
Train to enforce "reveal" policy			NGA GURU

Objectives of Train, but Verify

- Train secure AI systems by training ML models to enforce at least two security policies.
- Verify the security of AI systems by testing against declarative, realistic threat models.

This Talk

- will walk through of Helland & VanHoudnos (2020) and its implications for DoD.
- will ask: "What are the most interesting off diagonals to this community?"

Outline

What is a sufficient condition for training a convolutional neural network (CNN) image classifier such that adversarial examples against that model are recognizable to humans?

Comparison of Defensive Methods

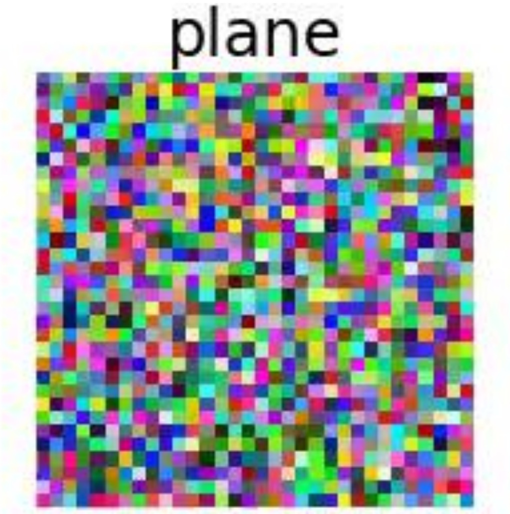
- Madry et al. (2017) + approximate methods
- TRADES (Zhang et al., 2019) + approximate methods
- Lemma: Defensive regularization drives down Lipschitz constant

Experimental Results

- Defensive regularization is sufficient for recognizability

Privacy

- Revealing characteristics of data collection



Adversarial walk for a CIFAR10 ResNet50 model trained via Madry PGD with ℓ_∞, ϵ=8/255

Defenses for Do: Comparison of Methods

Standard (undefended) training minimized expected loss across the training data:

$$\underset{f \in \mathcal{F}}{\text{minimize}}\ \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\left[\mathcal{L}(\boldsymbol{e}_y, f(\boldsymbol{x}))\right]$$

Madry Adversarial Training (Madry et al., 2017) trains on an internal adversary:

$$\underset{f \in \mathcal{F}}{\text{minimize}}\ \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\left[\max_{\boldsymbol{\delta} \in B_\epsilon} \mathcal{L}(\boldsymbol{e}_y, f(\boldsymbol{x}+\boldsymbol{\delta}))\right]$$

TRADES (Zhang et al., 2019) *trades* between expected loss and an internal adversary:

$$\underset{f \in \mathcal{F}}{\text{minimize}}\ \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\left[\mathcal{L}(\boldsymbol{e}_y,\ f(\boldsymbol{x})) + \max_{\boldsymbol{\delta} \in B_\epsilon} \beta\, \mathcal{L}(f(\boldsymbol{x}), f(\boldsymbol{x}+\boldsymbol{\delta}))\right]$$

Madry Adversarial Training Can Recover Other Methods

First order Taylor expansion of Madry connects to approximate first order methods:

Etmann et al. (2019), Finlay and Oberman (2019), and Ross and Doshi-Velez (2017)

$$\underset{f \in \mathcal{F}}{\text{minimize}}\ \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\left[\max_{\boldsymbol{\delta}\in B_\epsilon} \mathcal{L}(\boldsymbol{e}_y, f(\boldsymbol{x}+\boldsymbol{\delta}))\right]$$

$$\mathcal{L}(\boldsymbol{e}_y, f(\boldsymbol{x}+\boldsymbol{\delta})) = \mathcal{L}(\boldsymbol{e}_y, f(\boldsymbol{x})) + \boldsymbol{\delta}^\top \nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{e}_y, f(\boldsymbol{x})) + \mathcal{O}(\|\boldsymbol{\delta}\|_2^2)$$

$$\begin{aligned}\max_{\|\boldsymbol{\delta}\|_p \leq \epsilon} &\ \boldsymbol{\delta}^\top \nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{e}_y, f(\boldsymbol{x})) \\ &= \epsilon \left\|\nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{e}_y, f(\boldsymbol{x}))\right\|_q \\ &= \frac{\epsilon}{f(\boldsymbol{x})[y]} \left\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x})[y]\right\|_q\end{aligned}$$

$$\approx \underset{f \in \mathcal{F}}{\text{minimize}}\ \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\left[\underbrace{\mathcal{L}(\boldsymbol{e}_y, f(\boldsymbol{x}))}_{\text{Accuracy}} + \underbrace{\beta\epsilon \left\|\nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{e}_y, f(\boldsymbol{x}))\right\|_q}_{\text{Regularization}}\right]$$

TRADES Can Recover Other Methods, Step 1

TRADES $$\underset{f \in \mathcal{F}}{\text{minimize}}\ \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\left[L(\boldsymbol{e}_y, f(\boldsymbol{x})) + \max_{\boldsymbol{x}' \in \mathbb{B}(\boldsymbol{x},\varepsilon)} \beta\, L(f(\boldsymbol{x}), f(\boldsymbol{x}'))\right]$$

Virtual adversarial training (Miyato et al., 2018)

- Recall cross entropy loss: $L(\boldsymbol{p}, \boldsymbol{q}) = H(\boldsymbol{p}) + D_{KL}(\boldsymbol{p}||\boldsymbol{q})$
- Expand out boundary term:

$$\underset{f \in \mathcal{F}}{\text{minimize}}\ \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\left[L\left(\boldsymbol{e}_y, f(\boldsymbol{x})\right) + H\left(f(\boldsymbol{x})\right) + \max_{\boldsymbol{\delta} \in \mathbb{B}_\varepsilon} \beta\, D_{KL}\left(f(\boldsymbol{x})\, ||\, f(\boldsymbol{x}+\boldsymbol{\delta})\right)\right]$$

- Choose ℓ_2 ball to recover virtual adversarial training.

TRADES Can Recover Other Methods, Step 2

$$\underset{f \in \mathcal{F}}{\operatorname{minimize}} \, \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[L\left(\boldsymbol{e}_y, f(\boldsymbol{x})\right) + H\left(f(\boldsymbol{x})\right) + \max_{\boldsymbol{\delta} \in \mathbb{B}_\varepsilon} \beta \, D_{KL}\left(f(\boldsymbol{x}) \,||\, f(\boldsymbol{x}+\boldsymbol{\delta})\right) \right]$$

Expand the KL divergence to second order:

$$\underset{f \in \mathcal{F}}{\operatorname{minimize}} \, \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[L\left(\boldsymbol{e}_y, f(\boldsymbol{x})\right) + H\left(f(\boldsymbol{x})\right) + \frac{\beta}{2} \max_{\boldsymbol{\delta} \in \mathbb{B}_\varepsilon} \boldsymbol{\delta}^\top \boldsymbol{F}_{\boldsymbol{x}} \boldsymbol{\delta} \right]$$

Solve:

$$\max_{\boldsymbol{\delta} \in \mathbb{B}_\varepsilon} \boldsymbol{\delta}^\top \boldsymbol{F}_{\boldsymbol{x}} \boldsymbol{\delta} = \frac{\beta\varepsilon^2}{2} \lambda_{\max}\left(\boldsymbol{F}_{\boldsymbol{x}}\right)$$

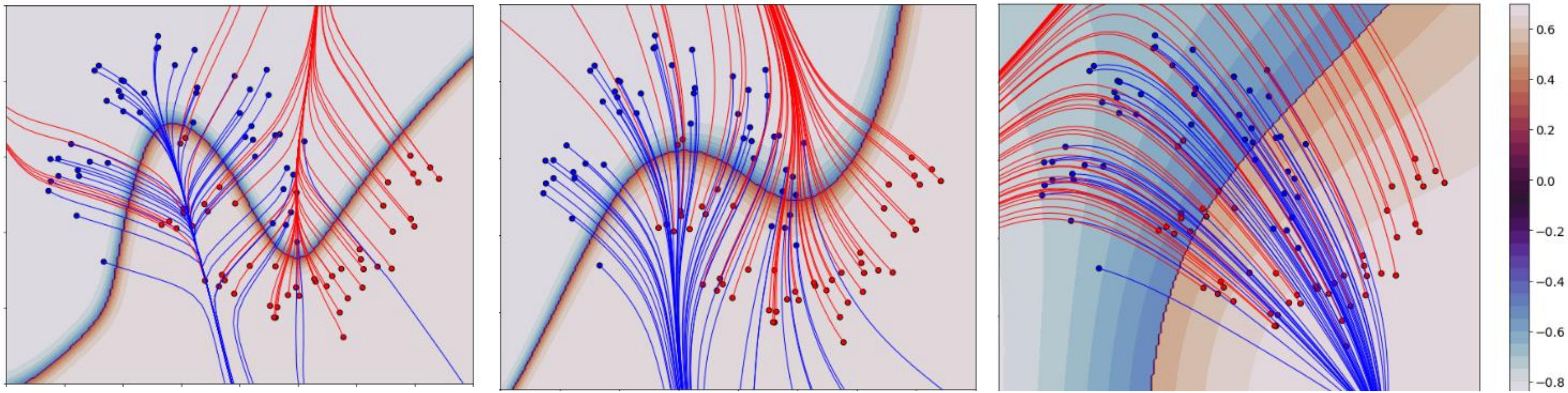
Various Fisher Information Matrix (FIM) methods fall out based on strategy for λ_{max}:

- Miyato et al. (2018) and Moosavi-Dezfooli et al. (2019) use finite-difference approximations.
- Zhao et al. (2019) uses power iteration.
- Shen et al. (2019) gives an upper bound on the full spectrum.

Madry, TRADES, and Approximate Methods Are Smooth

Adversarial walks on the half moon dataset. Levels are values of the loss.

- **Lemma**: Regularization of $\lambda_{max}(\boldsymbol{F_x})$ drives down the local Lipschitz constant.



$\beta = 0$ $\beta = 1$ $\beta = 100$

$$\underset{f \in \mathcal{F}}{\text{minimize}}\ \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\left[L(\boldsymbol{e}_y, f(\boldsymbol{x})) + \max_{\boldsymbol{x}' \in \mathbb{B}(\boldsymbol{x},\varepsilon)} \beta\, L(f(\boldsymbol{x}), f(\boldsymbol{x}'))\right]$$

Outline

What is a sufficient condition for training a convolutional neural network (CNN) image classifier such that adversarial examples against that model are recognizable to humans?

Comparison of defensive methods

- Madry et al. (2017) + approximate methods
- TRADES (Zhang et al., 2019) + approximate methods
- Lemma: Defensive regularization drives down Lipschitz constant.

Experimental Results

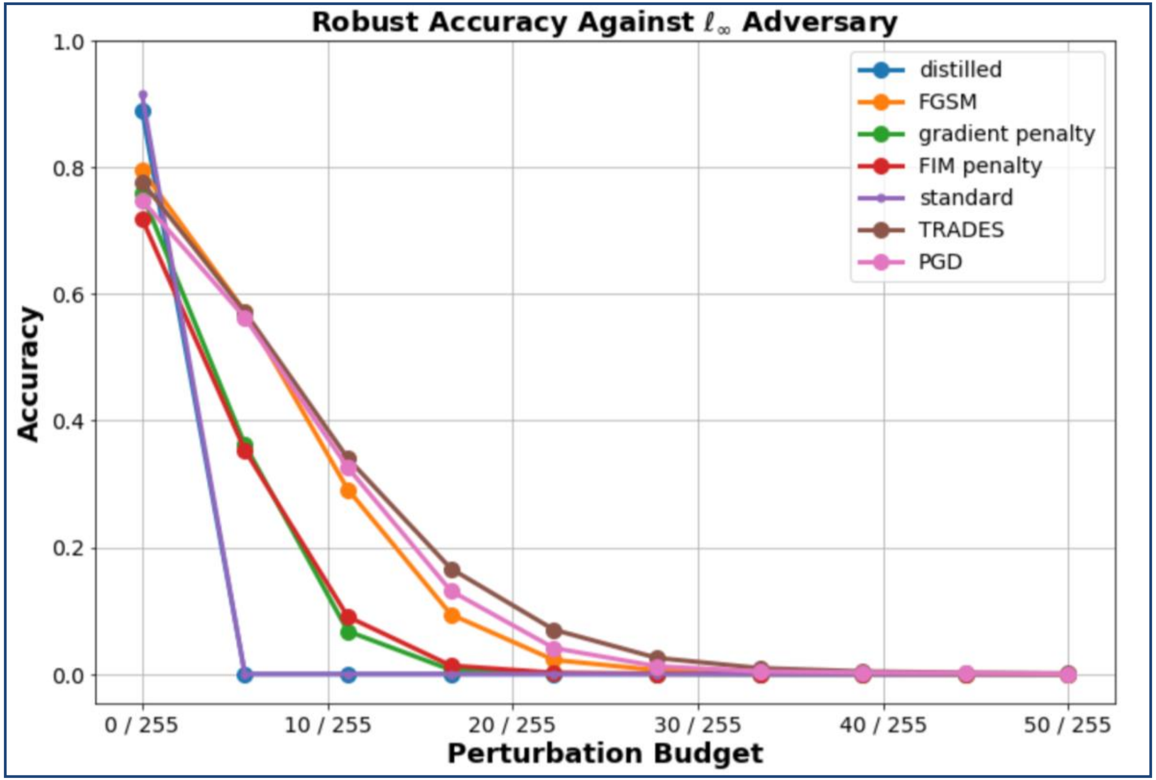
- Defensive regularization is sufficient for recognizability.

Privacy

- Revealing characteristics of data collection

Adversarial walk for a CIFAR10 ResNet50 model trained via Madry PGD with ℓ_∞, ϵ=8/255

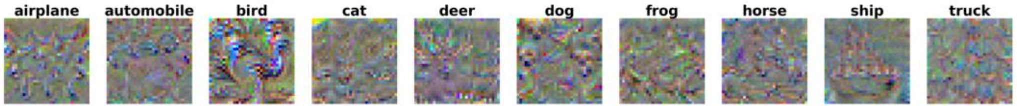
Experimental Results: Evaluation on Do Policy



- Standard (undefended) is not robust.
- Distillation (historical) is not robust.
- Gradient Penalty + FIM Penalty (approximate methods) are moderately robust.
- TRADES, PGD (Madry adversarial training) and FGSM (Madry with on iteration) are robust.

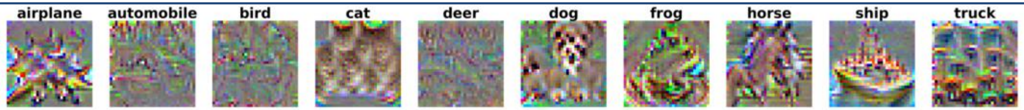
Experimental Results: Evaluation on Reveal Policy

Standard



Standard (undefended) is not recognizable.

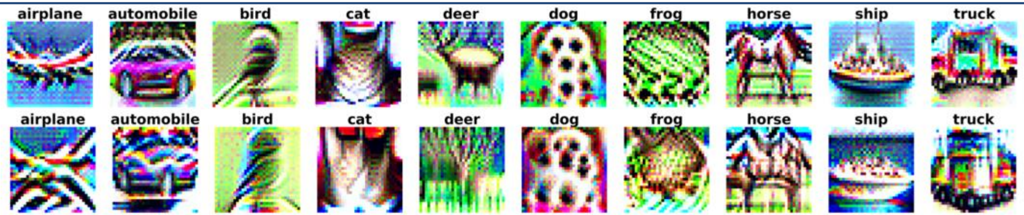
Distillation



Distillation (historical) is less recognizable.

Gradient penalty

FIM penalty

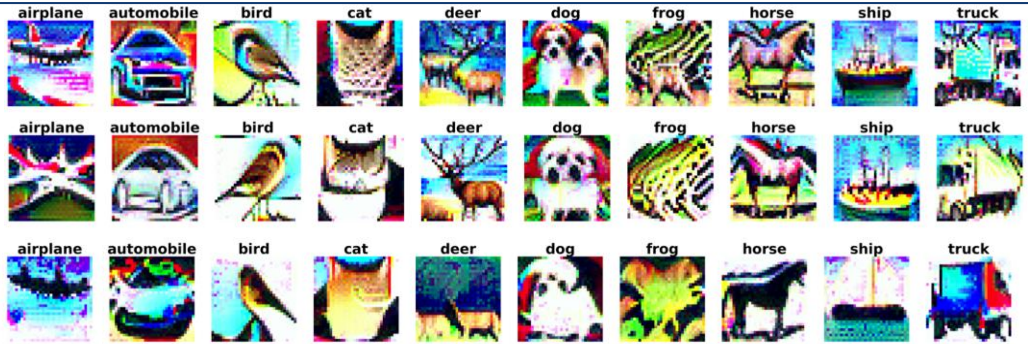


Approximate methods are moderately recognizable.

FGSM

PGD

TRADES



Full defenses are recognizable.

Outline

What is a sufficient condition for training a convolutional neural network (CNN) image classifier such that adversarial examples against that model are recognizable to humans?

Comparison of defensive methods

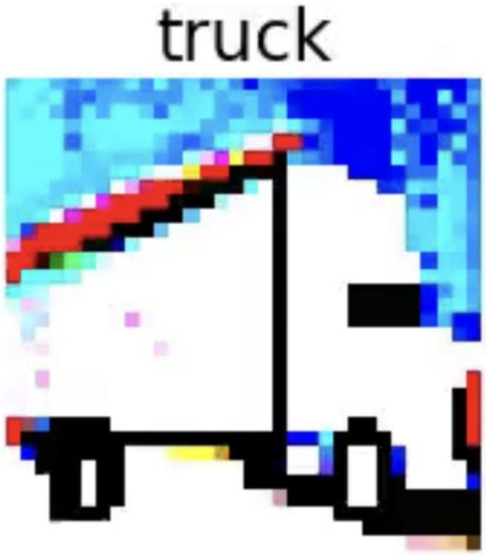
- Madry et al. (2017) + approximate methods
- TRADES (Zhang et al., 2019) + approximate methods
- Lemma: Defensive regularization drives down Lipschitz constant.

Experimental Results

- Defensive regularization is sufficient for recognizability.

Privacy

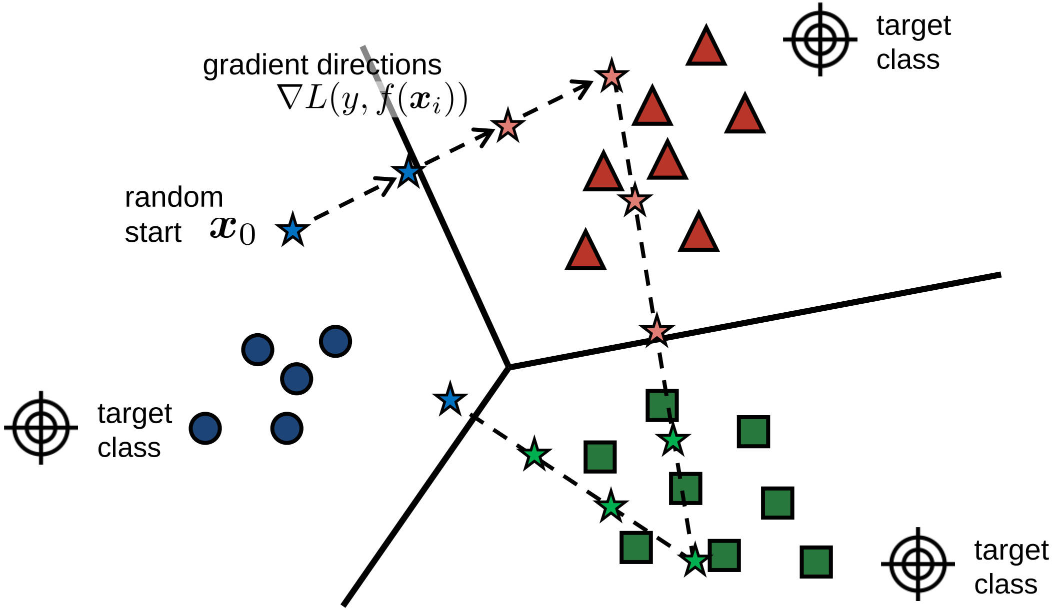
- Revealing characteristics of data collection



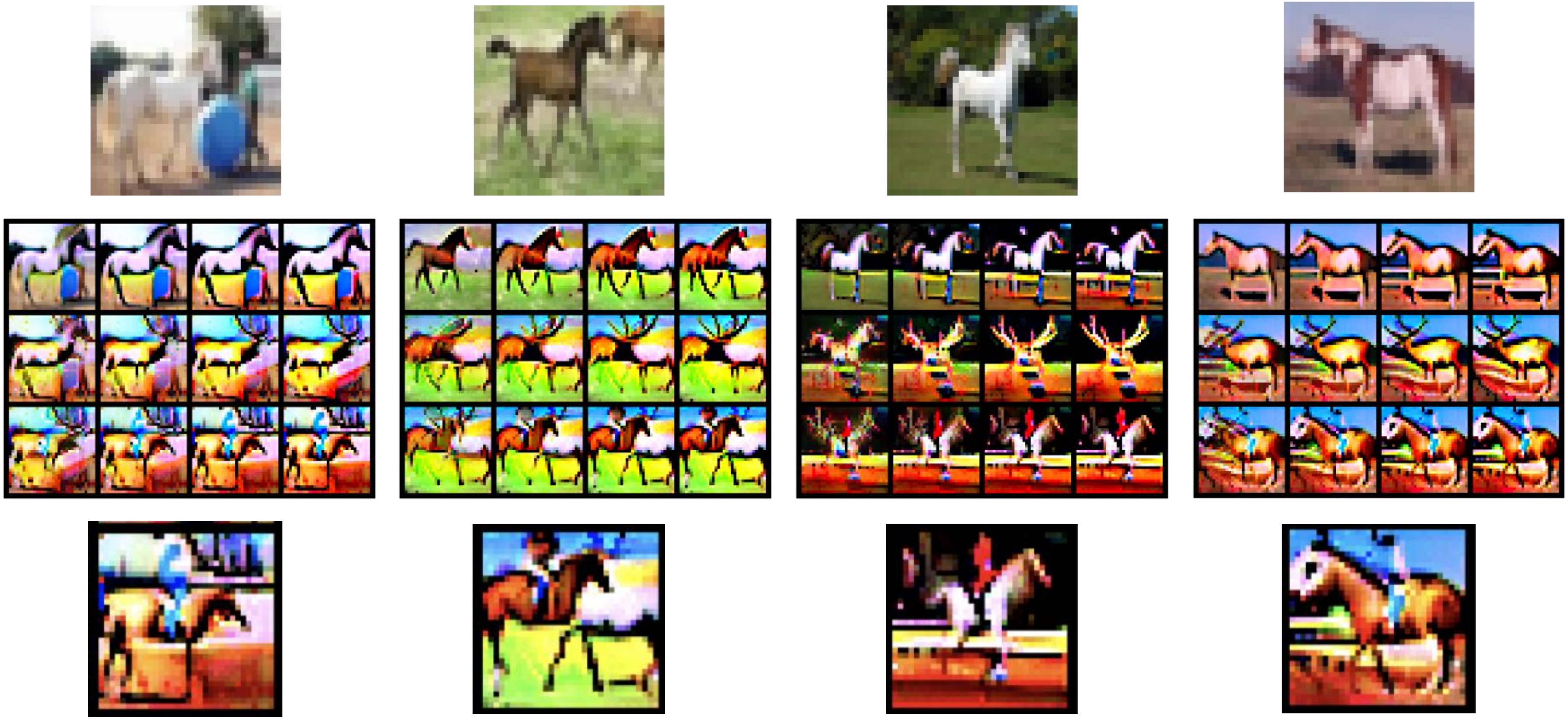
Adversarial walk for a CIFAR10 ResNet50 model trained via Madry PGD with ℓ_∞, ϵ=8/255

Adversarial Walks: Sequence of Adversarial Examples

Idea: do **unconstrained**, targeted **adversarial perturbation** towards each class in sequence.

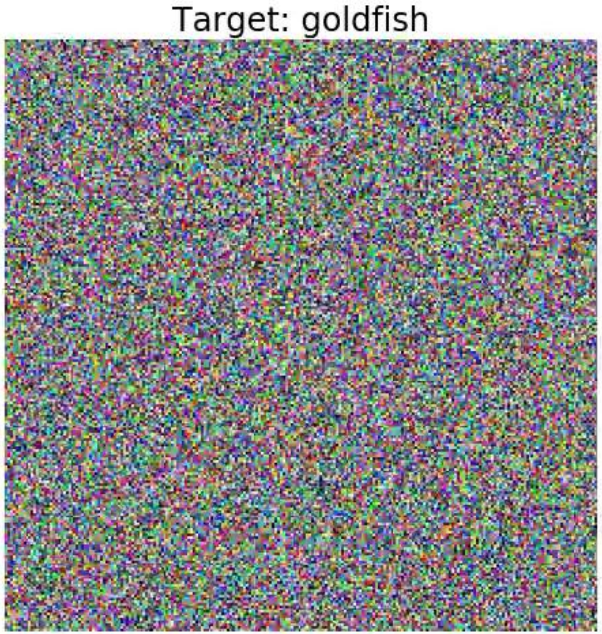


Privacy: Revealing Characteristics of Data (Model Access)

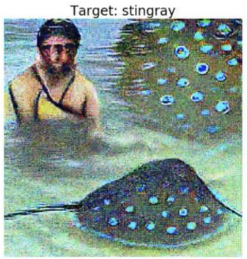


Revealing Characteristics of Data without Data Access

Fine Art: Adversarial Walk on ImageNet



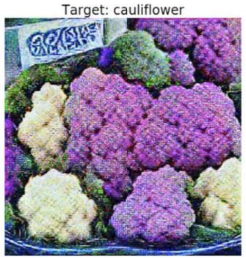
Attack: Characteristics of Training Data



Adversary with model access, but no data

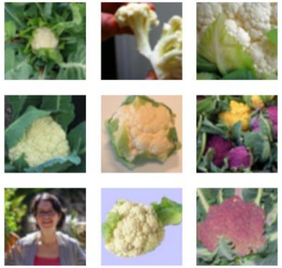
Some stingray images have swimmers in the water.

First 9 examples of synset n01498041 (stingray)



Adversary with model access, but no data

Cauliflower can be purple?



First 9 examples of synset n07715103 (cauliflower)

Summary & Roadmap

Train \ Verify	Learn	Do	Reveal
Learn			
Do			
Reveal			
Do & Reveal			

Train, but Verify

Summary

- State-of-the-art methods to enforce do policies are vulnerable to reveal attacks.
- Enforcing do and reveal will require new methods.

Roadmap

FY 2021:

- Quantify attacks to reveal policies.
- Develop new methods for do defenses and do attacks.
- Develop new methods to verify do policies (early version submitted ICLR ‘21).

FY 2022:

- Develop training methods for **do & reveal** that either
  - enforce both
  - trade between them

Team
SEI: Matt Churilla, Jon Helland, Grace Lewis, Nathan VanHoudnos, and Oren Wright
Carnegie Mellon University: Lujo Bauer, Matt Fredrickson, Aymeric Fromherz, Klas Leino, and Bryan Parno

info@sei.cmu.edu

References

M. Fredrikson, S. Jha, and T. Ristenpart. “Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures.” In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security - CCS ’15.* Pages 1322–1333. Denver, Colorado. 2015. doi: 10.1145/2810103.2813677.

I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and Harnessing Adversarial Examples.” *arXiv:1412.6572 [cs, stat].* Dec. 2014. Accessed: May 28, 2019. [Online]. Available: http://arxiv.org/abs/1412.6572.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. “Towards deep learning models resistant to adversarial attacks.” *arXiv preprint arXiv:1706.06083.* 2017.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. “Virtual adversarial training: a regularization method for supervised and semi-supervised learning.” *IEEE transactions on pattern analysis and machine intelligence.* Volume 41. Number 8. Pages1979–1993. 2018.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Jonathan Uesato, and Pascal Frossard. “Robustness via curvature regularization, and vice versa.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* Pages 9078–9086. 2019.

Chaomin Shen, Yaxin Peng, Guixu Zhang, and Jinsong Fan. “Defending against adversarial attacks by suppressing the largest eigenvalue of fisher information matrix.” *arXiv preprint arXiv:1909.06137.* 2019.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P Xing, Laurent El Ghaoui, and Michael I Jordan. “Theoretically principled trade-off between robustness and accuracy.” *arXiv preprint arXiv:1901.08573.* 2019.

Chenxiao Zhao, P Thomas Fletcher, Mixue Yu, Yaxin Peng, Guixu Zhang, and Chaomin Shen. “The adversarial attack and detection under the fisher information metric.” In *Proceedings of the AAAI Conference on Artificial Intelligence.* Volume 33. Pages 5869–5876. 2019.

SEI Team Members



Matt Churilla

Jon Helland

Grace Lewis


Nathan VanHoudnos

Oren Wright