@NicolasPapernot

Adversarial Examples in Machine Learning

Nicolas Papernot

Google PhD Fellow in Security, Pennsylvania State University

Talk at USENIX Enigma - February 1, 2017

Advised by Patrick McDaniel

Presentation prepared with Ian Goodfellow and Úlfar Erlingsson

Successes of machine learning

Google

NVIDIA

PayPlug

PayPal

invincea

deepinstinct

FeatureSmith

amazon web services

Google Cloud Platform

Autonomous driving	Financial fraud detection	Malware / APT detection	Machine Learning as a Service

Failures of machine learning: *Dave's talk*

An adversary forces a PDF detector trained with ML to make wrong predictions.

The scenario:

- Availability of the model for **intensive querying**
- Access to the model **label and score**
- **Binary** classifier (two outputs: malware/benign)

Failures of machine learning: *this talk*

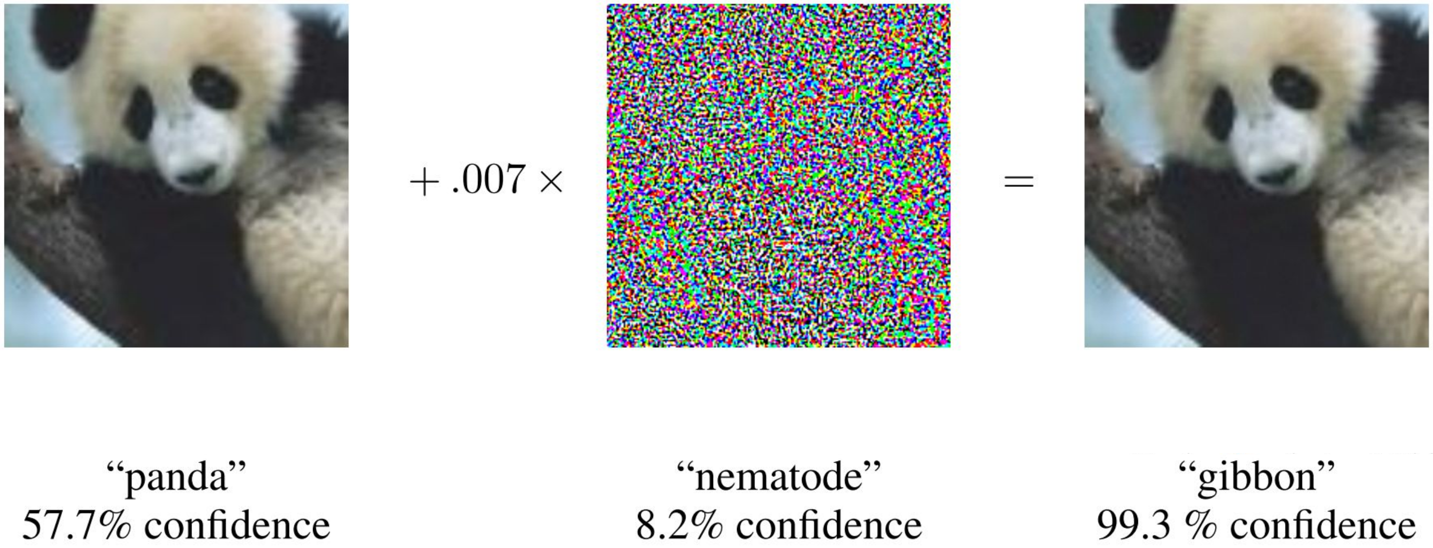
An adversary forces a computer vision model trained with ML to make wrong predictions.

The scenario:

- **Remote** availability of the model through an API
- Access to the model **label only**
- A **multi-class** classifier (up to 1000 outputs)

Adversarial examples represent *worst-case* domain shifts

Adversarial examples



[GSS15] Goodfellow et al. Explaining and Harnessing Adversarial Examples

Crafting adversarial examples: *fast gradient sign method*

During training, the classifier uses a loss function to **minimize** model prediction errors

After training, **attacker** uses loss function to **maximize** model prediction error

1. Compute its gradient with respect to the input of the model

$$\nabla_x J(\theta, x, y)$$

2. Take the sign of the gradient and multiply it by a threshold

$$x + \varepsilon \cdot sgn(\nabla_x J(\theta, x, y))$$

[GSS15] Goodfellow et al. Explaining and Harnessing Adversarial Examples

Black-box attacks and transferability

Threat model of a black-box attack

Adversarial capabilities

~~Training data~~
~~Model architecture~~
~~Model parameters~~
~~Model scores~~

(limited) oracle access: *labels*

Adversarial goal

Force a ML model remotely accessible through an API to misclassify

Example

amazon web services™

Google Cloud Platform

Our approach to black-box attacks

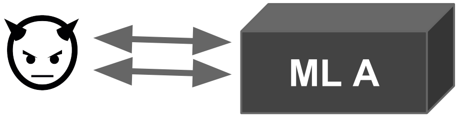
Alleviate lack of knowledge about model

Alleviate lack of training data

Adversarial example transferability

Adversarial examples have a **transferability** property:

samples crafted to mislead a model A are likely to mislead a model B

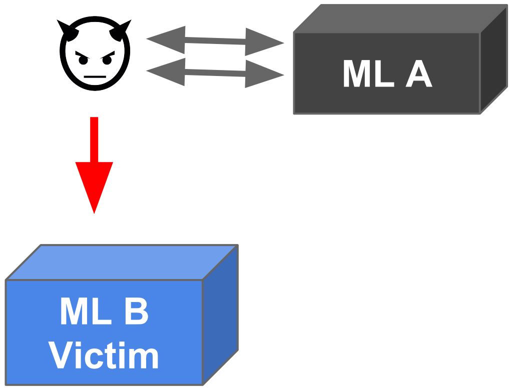


[SZS14] Szegedy et al. Intriguing properties of neural networks

Adversarial example transferability

Adversarial examples have a **transferability** property:

samples crafted to mislead a model A are likely to mislead a model B

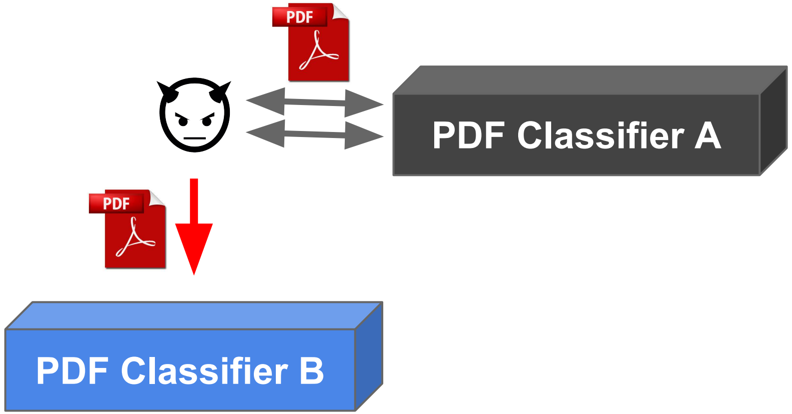


[SZS14] Szegedy et al. Intriguing properties of neural networks

Adversarial example transferability

Adversarial examples have a **transferability** property:

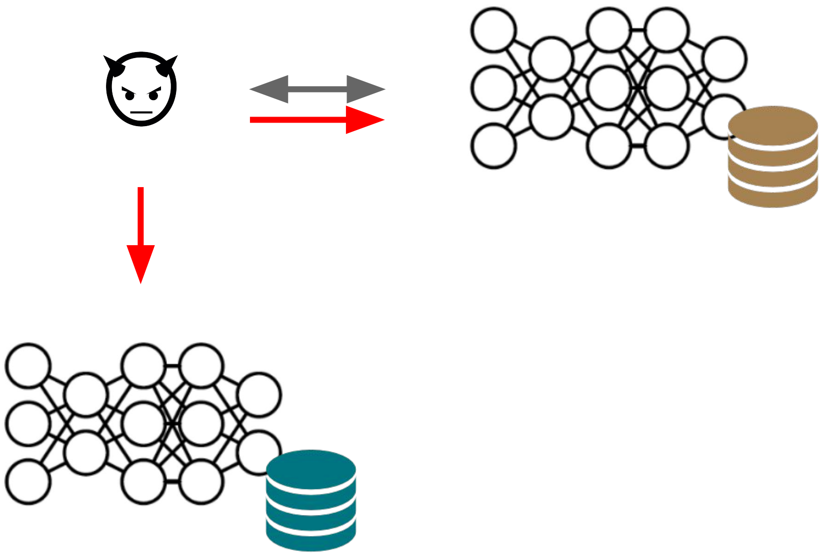
samples crafted to mislead a model A are likely to mislead a model B



Adversarial example transferability

Adversarial examples have a **transferability** property:

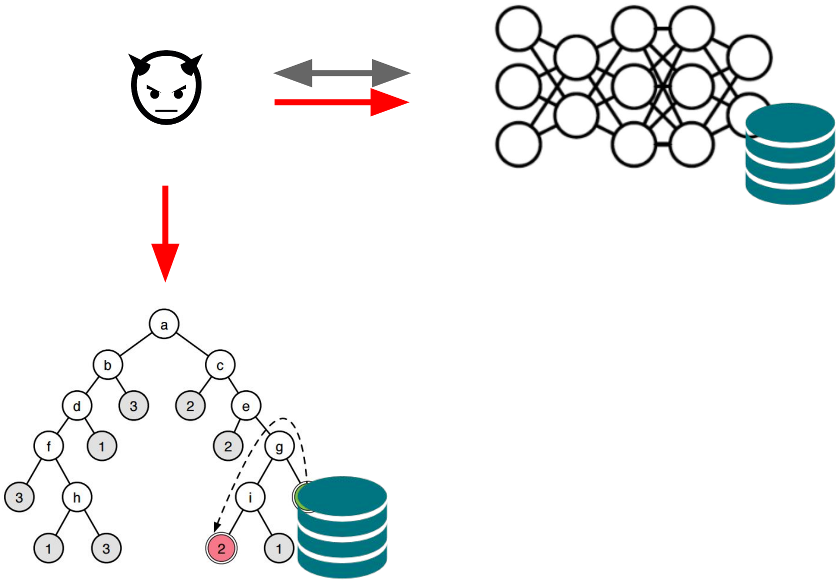
samples crafted to mislead a model A are likely to mislead a model B



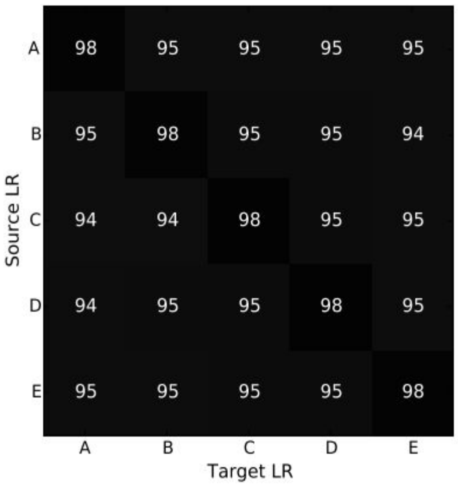
Adversarial example transferability

Adversarial examples have a **transferability** property:

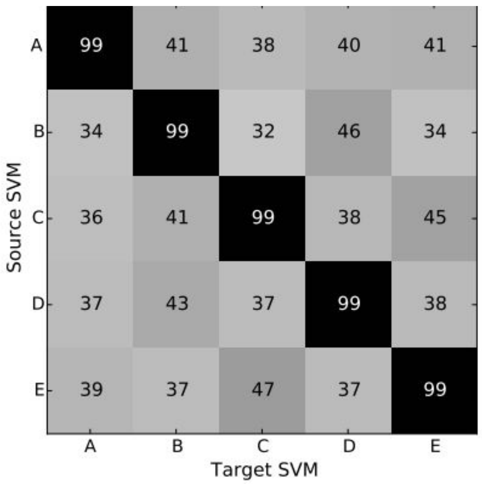
samples crafted to mislead a model A are likely to mislead a model B



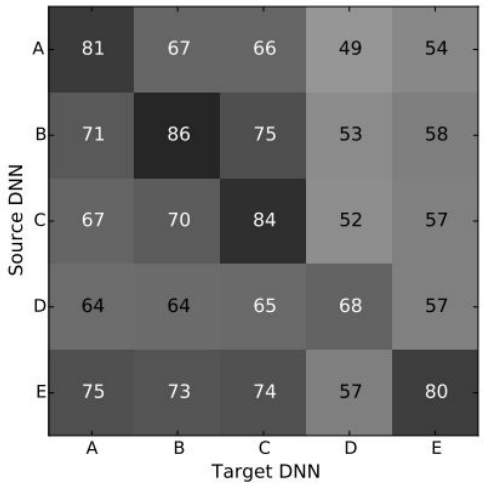
Intra-technique transferability: cross training data



Strong



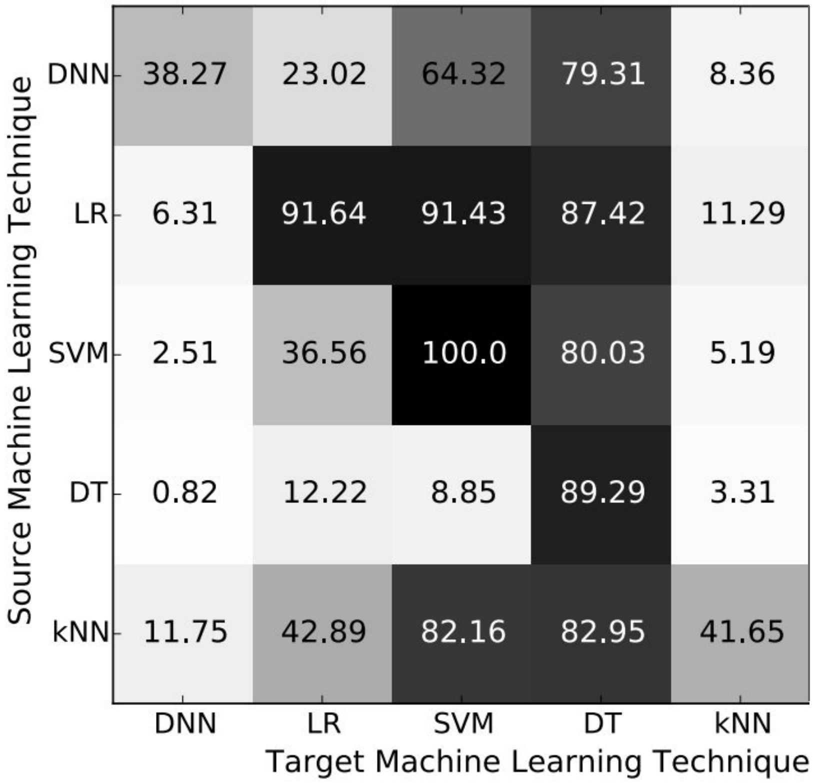
Weak



Intermediate

[PMG16b] Papernot et al. Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples

Cross-technique transferability



[PMG16b] Papernot et al. Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples

Cross-technique transferability

Source Machine Learning Technique \ Target Machine Learning Technique	DNN	LR	SVM	DT	kNN	Ens.
DNN	38.27	23.02	64.32	79.31	8.36	20.72
LR	6.31	91.64	91.43	87.42	11.29	44.14
SVM	2.51	36.56	100.0	80.03	5.19	15.67
DT	0.82	12.22	8.85	89.29	3.31	5.11
kNN	11.75	42.89	82.16	82.95	41.65	31.92

[PMG16b] Papernot et al. Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples

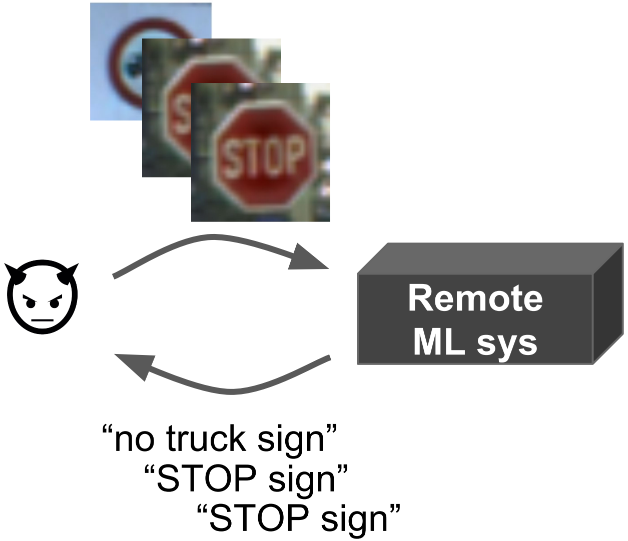
Our approach to black-box attacks

Alleviate lack of knowledge about model

Adversarial example transferability from a substitute model to target model

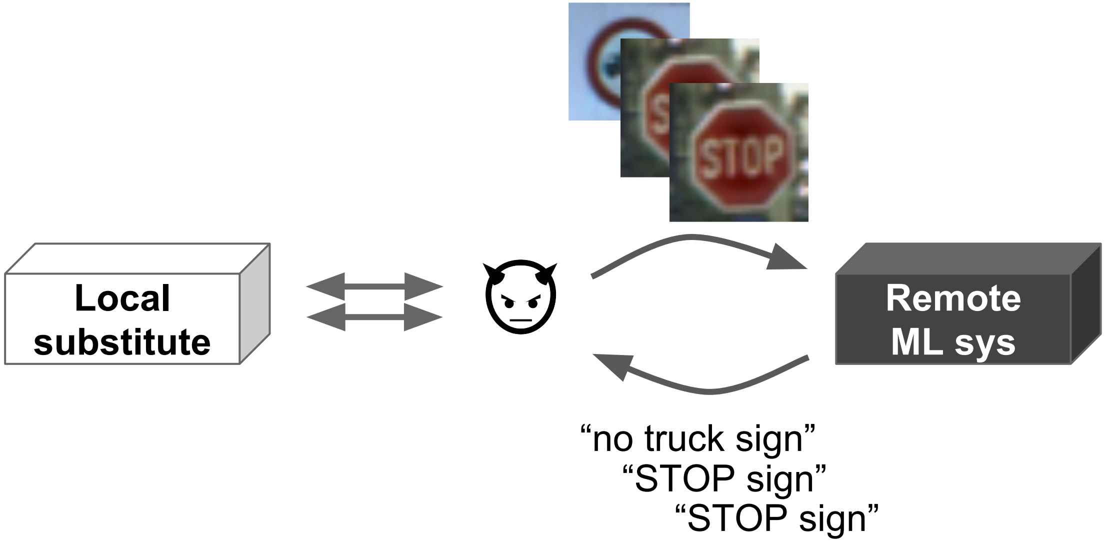
Alleviate lack of training data

Attacking remotely hosted black-box models



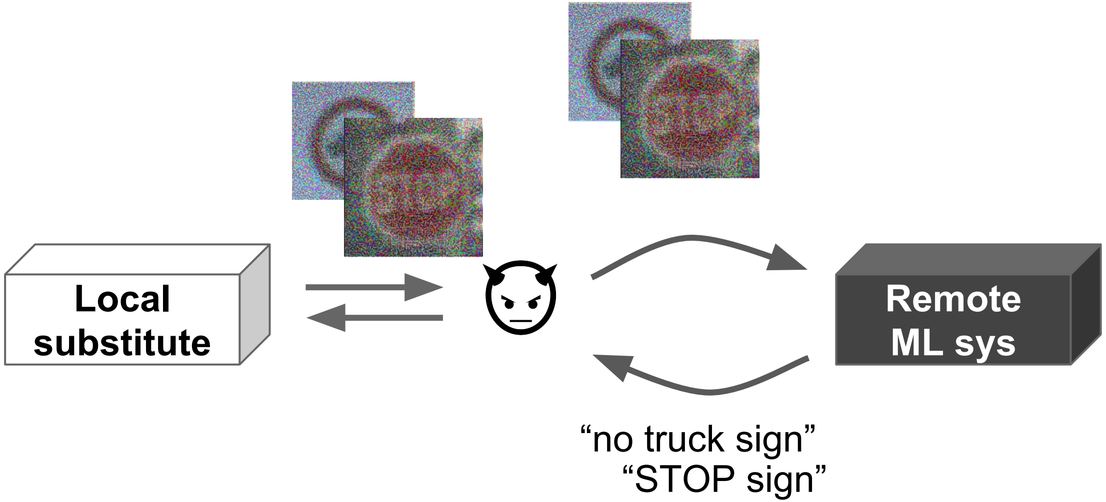
(1) The adversary queries remote ML system for labels on inputs of its choice.

Attacking remotely hosted black-box models



(2) The adversary uses this labeled data to train a local substitute for the remote system.

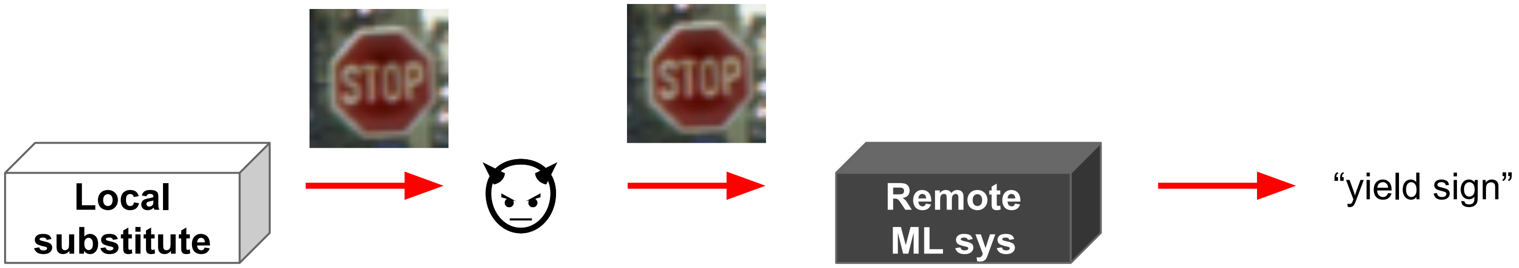
Attacking remotely hosted black-box models



$$S_{\rho+1} = \{\vec{x} + \lambda_{\rho+1} \cdot \mathrm{sgn}(J_F[\tilde{O}(\vec{x})]) : \vec{x} \in S_\rho\} \cup S_\rho$$

(3) The adversary selects new synthetic inputs for queries to the remote ML system based on the local substitute's output surface sensitivity to input variations.

Attacking remotely hosted black-box models



(4) The adversary then uses the local substitute to craft adversarial examples, which are misclassified by the remote ML system because of transferability.

Our approach to black-box attacks

Alleviate lack of knowledge about model

Adversarial example transferability from a substitute model to target model

+

Alleviate lack of training data

Synthetic data generation

Results on real-world remote systems

Remote Platform	ML technique	Number of queries	Adversarial examples misclassified (after querying)
MetaMind	Deep Learning	6,400	84.24%
amazon web services™	Logistic Regression	800	96.19%
Google Cloud Platform	Unknown	2,000	97.72%

All remote classifiers are trained on the MNIST dataset (10 classes, 60,000 training samples)

[PMG16a] Papernot et al. Practical Black-Box Attacks against Deep Learning Systems using Adversarial Examples

Defending and benchmarking machine learning models

Adversarial training

Intuition: **injecting** adversarial example during training with **correct** labels

Goal: improve model **generalization** outside of training manifold

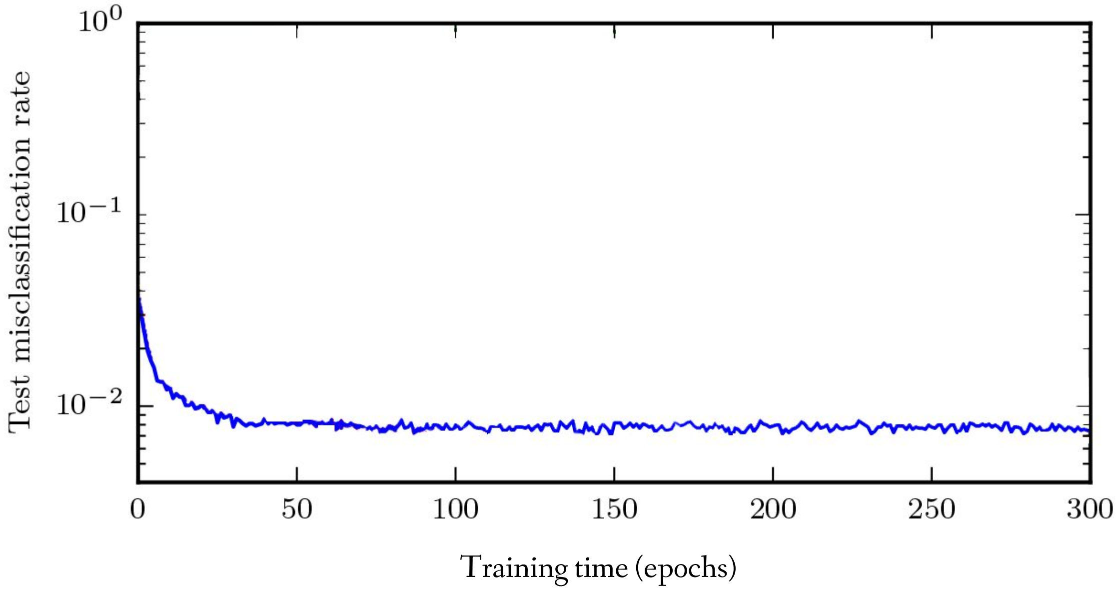


Figure by Ian Goodfellow

Adversarial training

Intuition: **injecting** adversarial example during training with **correct** labels

Goal: improve model **generalization** outside of training manifold

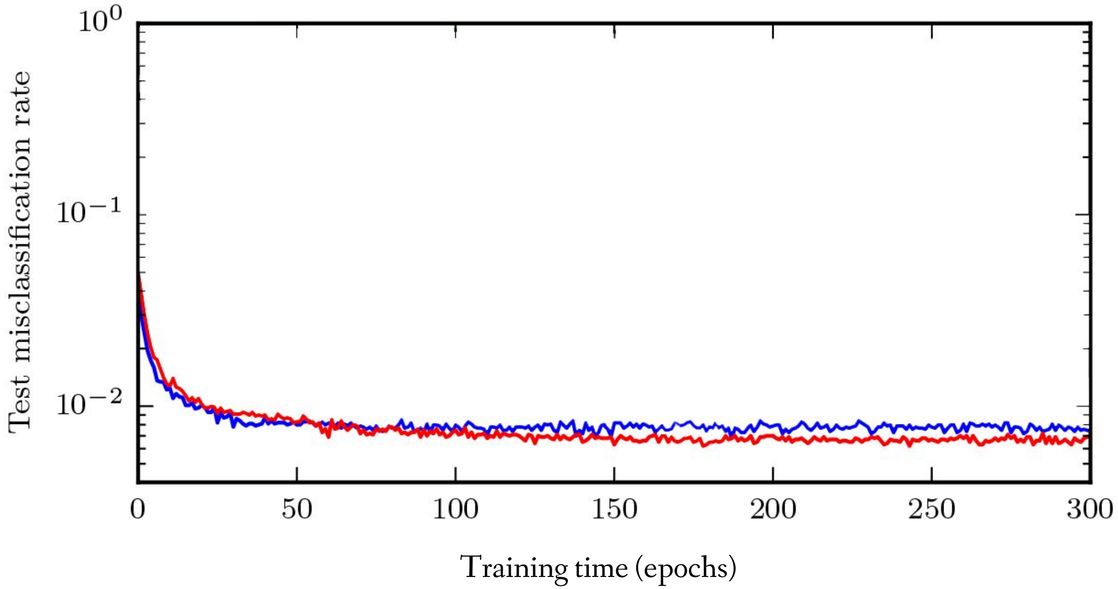


Figure by Ian Goodfellow

Adversarial training

Intuition: **injecting** adversarial example during training with **correct** labels

Goal: improve model **generalization** outside of training manifold

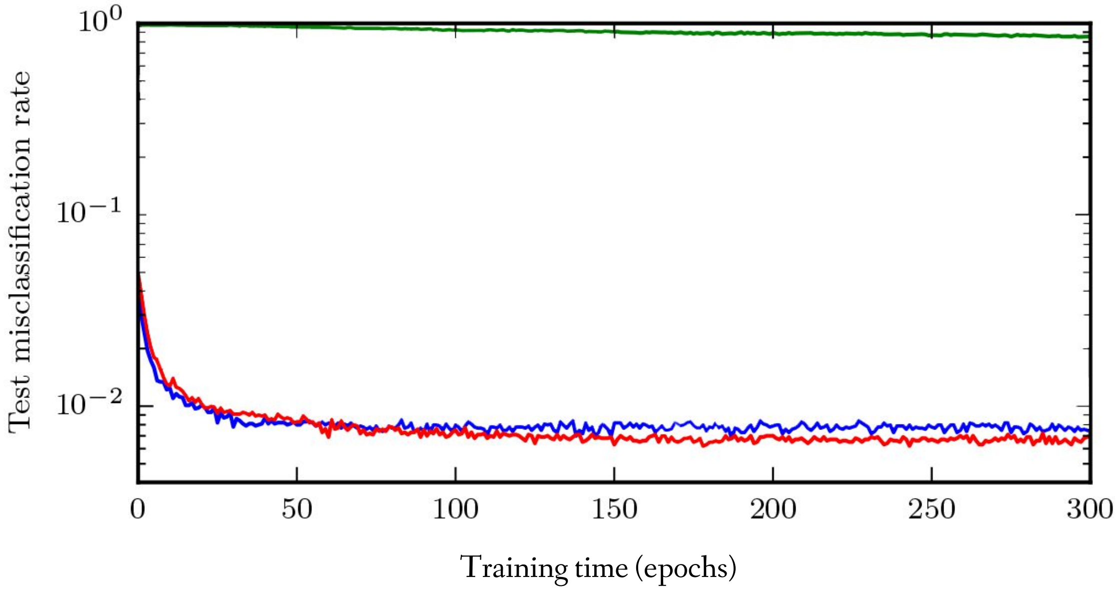


Figure by Ian Goodfellow

Adversarial training

Intuition: **injecting** adversarial example during training with **correct** labels

Goal: improve model **generalization** outside of training manifold

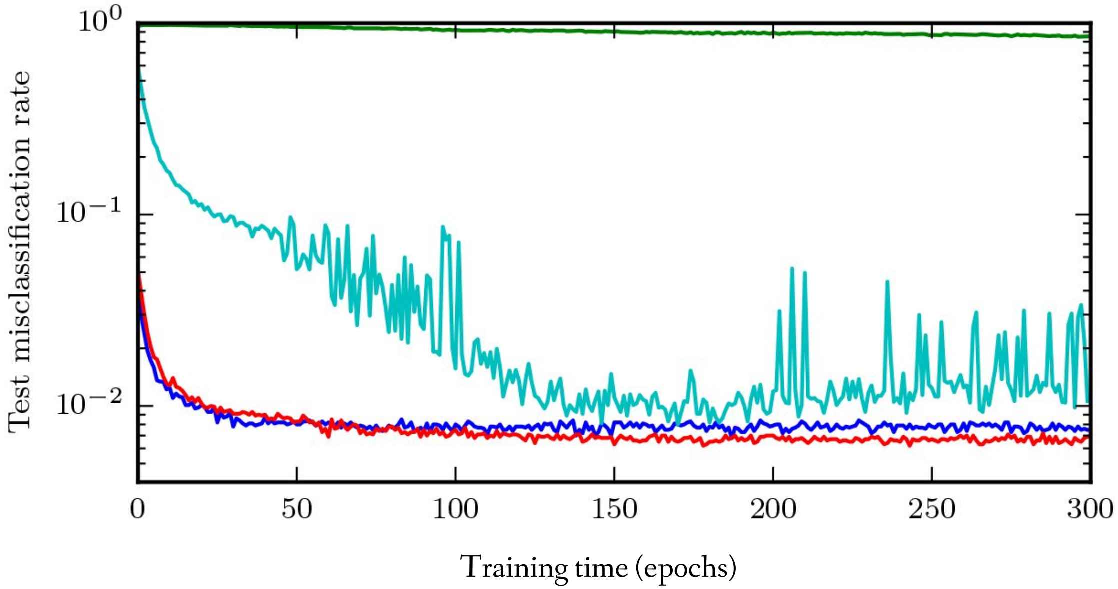


Figure by Ian Goodfellow

Adversarial training: *some limitations*

Works well here because **same attack** used by adversary and classifier

Harder to generalize model robustness to **adaptive attacks**

Classifier needs to be aware of **all** attacker strategies

The *cleverhans* library (and blog)

Code at: `github.com/openai/cleverhans`

Blog at: `cleverhans.io`

Benchmark models against adversarial example attacks

Increase model robustness with *adversarial training*

Contributions welcomed!

Hands-on tutorial with the MNIST dataset

```
# Define input TF placeholder
x = tf.placeholder(tf.float32, shape=(None, 1, 28, 28))
y = tf.placeholder(tf.float32, shape=(None, FLAGS.nb_classes))

# Define TF model graph
model = model_mnist()
predictions = model(x)

# Craft adversarial examples using Fast Gradient Sign Method (FGSM)
adv_x = fgsm(x, predictions, eps=0.3)
X_test_adv, = batch_eval(sess, [x], [adv_x], [X_test])

# Evaluate the accuracy of the MNIST model on adversarial examples
accuracy = tf_model_eval(sess, x, y, predictions, X_test_adv, Y_test)
print('Test accuracy on adversarial examples: ' + str(accuracy))
```

Implications of security research to the fields of ML & AI

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

Accurate extrapolation outside of training data (i.e., resilience to adversarial examples) is a prerequisite for model-based optimization

Image source: http://www.pullman-wa.gov/images/stories/Police/drugs.jpg

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

?

Thank you for listening!

Thank you to my sponsors: ARL Google