Adversarial Examples and Adversarial Training

Ian Goodfellow, OpenAI Research Scientist
Presentation at San Francisco AI Meetup, 2016-08-18
In this presentation

• “Intriguing Properties of Neural Networks” Szegedy et al, 2013

• “Explaining and Harnessing Adversarial Examples” Goodfellow et al 2014

• “Adversarial Perturbations of Deep Neural Networks” Warde-Farley and Goodfellow, 2016
In this presentation

• “Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples” Papernot et al 2016

• “Practical Black-Box Attacks against Deep Learning Systems using Adversarial Examples” Papernot et al 2016

• “Adversarial Perturbations Against Deep Neural Networks for Malware Classification” Grosse et al 2016 (not my own work)
In this presentation

• “Distributional Smoothing with Virtual Adversarial Training” Miyato et al 2015 (not my own work)

• “Virtual Adversarial Training for Semi-Supervised Text Classification” Miyato et al 2016

Overview

• What are adversarial examples?

• Why do they happen?

• How can they be used to compromise machine learning systems?

• What are the defenses?

• How to use adversarial examples to improve machine learning, even when there is no adversary
Adversarial Examples

Timeline:
“Adversarial Classification” Dalvi et al 2004: fool spam filter
“Evasion Attacks Against Machine Learning at Test Time” Biggio 2013: fool neural nets
Szegedy et al 2013: fool ImageNet classifiers imperceptibly
Goodfellow et al 2014: cheap, closed form attack
Turning Objects into “Airplanes”
Attacking a Linear Model
Not just for neural nets

• Linear models
  • Logistic regression
  • Softmax regression
  • SVMs
• Decision trees
• Nearest neighbors
Adversarial Examples from Overfitting
Adversarial Examples from Excessive Linearity
Modern deep nets are very piecewise linear

- Rectified linear unit
- Carefully tuned sigmoid
- Maxout
- LSTM

(Goodfellow 2016)
Nearly Linear Responses in Practice

![Graph showing log ratio of class probabilities against interpolation coefficient. The curve decreases sharply as the interpolation coefficient increases.]
Maps of Adversarial and Random Cross-Sections

(collaboration with David Warde-Farley and Nicolas Papernot)
Maps of Adversarial Cross-Sections
Maps of Random Cross-Sections

Adversarial examples are not noise

(collaboration with David Warde-Farley and Nicolas Papernot)
Clever Hans

(“Clever Hans, Clever Algorithms,” Bob Sturm)
Small inter-class distances

Perturbation changes the true class

Random perturbation does not change the class

Perturbation changes the input to “rubbish class”

All three perturbations have L2 norm 3.96 This is actually small. We typically use 7!
The Fast Gradient Sign Method

\[ J(\tilde{x}, \theta) \approx J(x, \theta) + (\tilde{x} - x)^\top \nabla_x J(x). \]

Maximize

\[ J(x, \theta) + (\tilde{x} - x)^\top \nabla_x J(x) \]

subject to

\[ ||\tilde{x} - x||_\infty \leq \epsilon \]

\[ \Rightarrow \tilde{x} = x + \epsilon \text{sign} (\nabla_x J(x)). \]
Wrong almost everywhere
Cross-model, cross-dataset generalization
Cross-technique transferability

(Papernot 2016)

(Goodfellow 2016)
Transferability Attack

Target model with unknown weights, machine learning algorithm, training set; maybe non-differentiable

Train your own model

Substitute model mimicking target model with known, differentiable function

Deploy adversarial examples against the target; transferability property results in them succeeding

Adversarial crafting against substitute

Adversarial examples

(Goodfellow 2016)
Adversarial Examples in the Human Brain

These are concentric circles, not intertwined spirals.

(Pinna and Gregory, 2002)
Practical Attacks

• Fool real classifiers trained by remotely hosted API (MetaMind, Amazon, Google)

• Fool malware detector networks

• Display adversarial examples in the physical world and fool machine learning systems that perceive them through a camera
Adversarial Examples in the Physical World

(a) Printout

(b) Photo of printout

(c) Cropped image
Failed defenses

Generative pretraining
Adding noise at test time
Confidence-reducing perturbation at test time
Weight decay
Various non-linear units

Removing perturbation with an autoencoder
Ensembles
Error correcting codes
Multiple glimpses
Double backprop
Dropout
Adding noise at train time
Training on Adversarial Examples

![Graph showing validation set error over training time for clean examples and adversarial examples. The graph indicates that adversarial training results in lower error compared to standard training.](Goodfellow 2016)
Adversarial Training

Labeled as bird

Decrease probability of bird class

Still has same label (bird)

(Goodfellow 2016)
Virtual Adversarial Training

Unlabeled; model guesses it’s probably a bird, maybe a plane

New guess should match old guess (probably bird, maybe plane)

Adversarial perturbation intended to change the guess

(Goodfellow 2016)
Text Classification with VAT

RCV1 Misclassification Rate

6.00 - 8.00

[Graph showing misclassification rates for different models:
- Earlier SOTA: 7.70
- SOTA: 7.20
- Our baseline: 7.40
- Adversarial: 7.12
- Virtual Adversarial: 7.05
- Both: 6.97
- Both + bidirectional model: 6.68]

Zoomed in for legibility
Conclusion

• Attacking is easy

• Defending is difficult

• Benchmarking vulnerability is training

• Adversarial training provides regularization and semi-supervised learning