Adversarial Learning

Nan Ye

School of Mathematics and Physics
The University of Queensland

Recall: Impressive Results

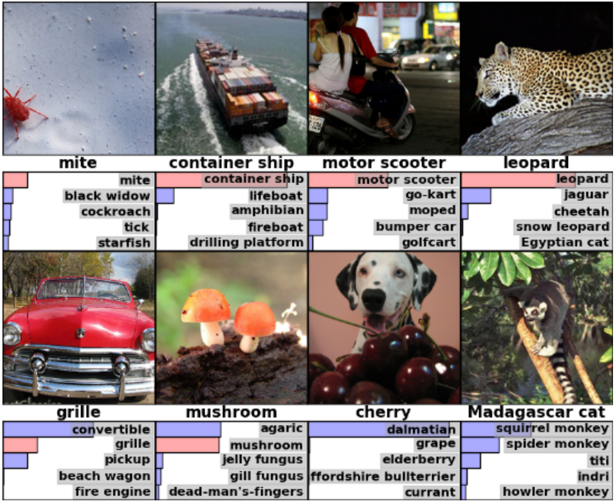
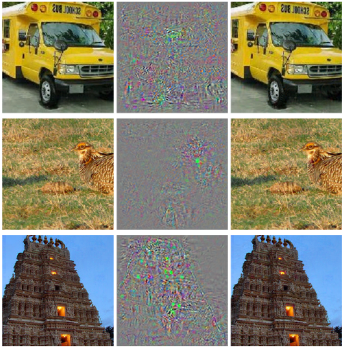
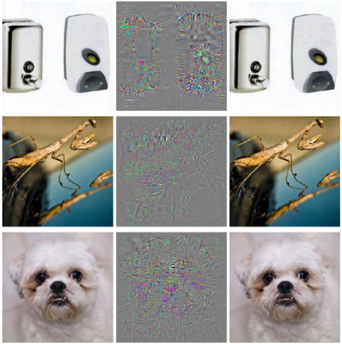


Image classification

Krizhevsky, Sutskever, and Hinton, Imagenet classification with deep convolutional neural networks, 2012

Adversarial Examples

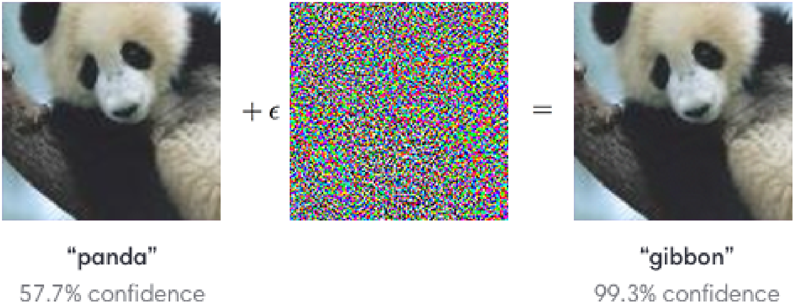




original image $+$ imperceptible noise $=$ ostrich

😱

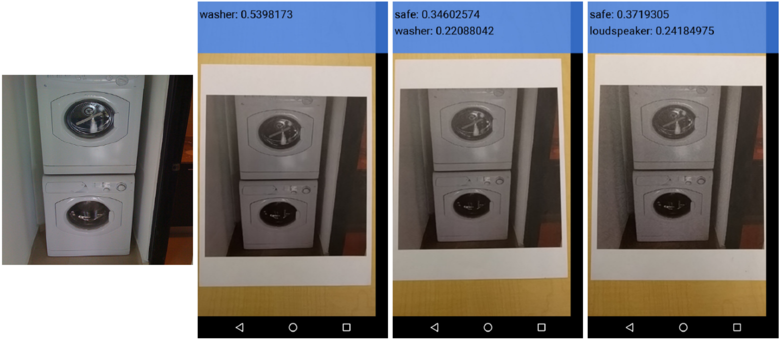
Szegedy, Zaremba, Sutskever, Bruna, Erhan, Goodfellow, and Fergus, Intriguing properties of neural networks, 2014



fast generation of adversarial examples

😱😱

Goodfellow, Shlens, and Szegedy, Explaining and harnessing adversarial examples, 2015



(a) Image from dataset (b) Clean image (c) Adv. image, $\epsilon = 4$ (d) Adv. image, $\epsilon = 8$

photoed adversarial images are still adversarial

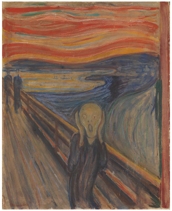
$\Rightarrow$ we can fool systems with camera sensors

😱😱😱

Kurakin, Goodfellow, and Bengio, Adversarial examples in the physical world, 2016

Adversarial Examples Are Universal

- Much attention has been paid on adversarial examples for neural nets
- But adversarial examples exist for many other models
  - Linear models: logistic regression, SVMs
  - Decision trees
  - Nearest neighbors
- and possibly for biological neural nets too...



The Dress

What's the color of the dress?

The Dress

What's the color of the dress?

what people see

- blue and black
- white and gold
- blue and brown
- others

RGB analysis

- dark yellow and light blue

The Dress

What's the color of the dress?

what people see

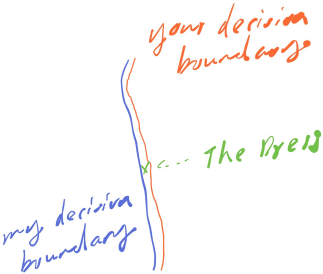
- blue and black
- white and gold
- blue and brown
- others

RGB analysis

- dark yellow and light blue

seller's photo

An explanation



Adversarial Examples are Transferable

- Adversarial examples generated for one model are often misclassified by another model — they are transferable.
- While a huge amount of effort has been spent to get good performance on many very large datasets, many cases still appear to be hard
  - e.g. we still can't get computer vision to work well enough for autonomous driving in unseen situations
  - scenes which appear to be essentially the same to us may behave like adversarial examples to the models
- Perhaps the set of hard cases (for ANNs) is much larger than the set of solvable cases?!

Implications of Adversarial Examples

- Machine learning algorithms do not learn smooth functions on natural inputs
  - What we consider as imperceptible perturbations are perceived as drastic changes by neural nets
- Machine learning algorithms do not generalise in the same way as human brains do
  - Machine learning algorithms are susceptible to adversarial attacks, while human brain does not.
  - Apparently, human brain seems to capture certain highly stable features on natural inputs.

- Machine learning systems are vulnerable to adversarial attacks
  - Someone wearing a mask can pretend to be you
  - A traffic sign can be imperceptibly changed to fool an autonomous vehicle to make dangerous moves

Explaining Adversarial Examples

- Adversarial examples can be present when a model overfits.
  - e.g. the decision boundary of 1NN is highly overfitting, and perturbation can easily change the class.
- Adversarial examples can be present when a model has excessive linearity.
  - Examples lying close to the linear boundary can be misclassified when perturbed.
  - Neural nets often try to work in the linear region!

Adversarial Attacks

Taxonomy of attacks

- A specific outcome desired?
  - Non-targeted: only an incorrect label required
  - Targeted: a specific outcome required
- Model known?
  - White box: full knowledge of the model
  - Black box with probing: no/limited knowledge of the model, but can probe or query the model
  - Black box without probing: no/limited knowledge of the model

White-box Attacks

Minimum perturbation method

- A white box targeted attack that searches for minimum perturbation needed to change prediction to a desired label
- Assume that the network computes $f : \mathbf{R}^m \to \mathbf{R}^k$
  - for an input $\mathbf{x} \in \mathbf{R}^m$, $f(\mathbf{x}) \in \mathbf{R}^k$ are the scores for the k classes
- Aim: classify perturbed example $\mathbf{x} + \delta$ to class y

$$\begin{aligned} \min_{\delta \in \mathbf{R}^m} \quad & \|\delta\|_2 \\ \text{s.t.} \quad & \arg\max_i f(\mathbf{x} + \delta)_i = y \text{ and } \mathbf{x} + \delta \in [0,1]^m. \end{aligned}$$

This is hard to solve, because it is hard to make sure δ satisfies the target class constraint.

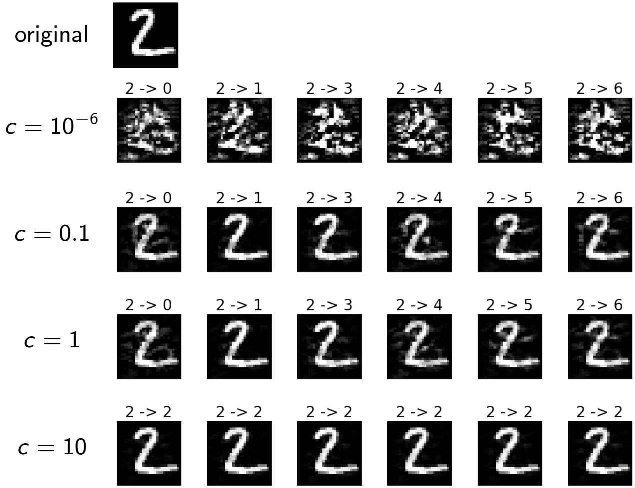
Szegedy, Zaremba, Sutskever, Bruna, Erhan, Goodfellow, and Fergus, Intriguing properties of neural networks, 2014

- Approximation

$$\begin{aligned} \min_{\delta \in \mathbf{R}^m} \quad & c\|\delta\|_2^2 + L(\mathbf{x} + \delta, y, f) \\ \text{s.t.} \quad & \mathbf{x} + \delta \in [0, 1]^m. \end{aligned}$$

- Intuitively, we want δ that
  - is a valid perturbation ($\Leftarrow$ box constraints),
  - is small ($\Leftarrow$ regularizer $c\|\delta\|_2^2$),
  - and the class of $x + \delta$ is *likely* to be y ($\Leftarrow$ minimizing $L(\mathbf{x} + \delta, y, f)$)
- The optimization problem is solvable using box-constrained L-BFGS

Can we perturb this image to fool a neural net to classify it as other digits like 1, 3, 4, 5, 6?



Min perturbation method examples for LeNet5 on MNIST

- The method involves solving a hard optimization problem, and is often slow.
- The perturbation is small, and sometimes easy to defend by reducing image quality.

Fast gradient sign method (FGSM)

- A white box non-targeted attack
- Aim: increase the loss $L(\mathbf{x}, y, f)$ for true class y

$$\begin{aligned} \max_{\delta \in \mathbf{R}^m} \quad & L(\mathbf{x} + \delta, y, f), \\ \text{s.t.} \quad & \|\delta\|_\infty \leq \epsilon. \end{aligned}$$

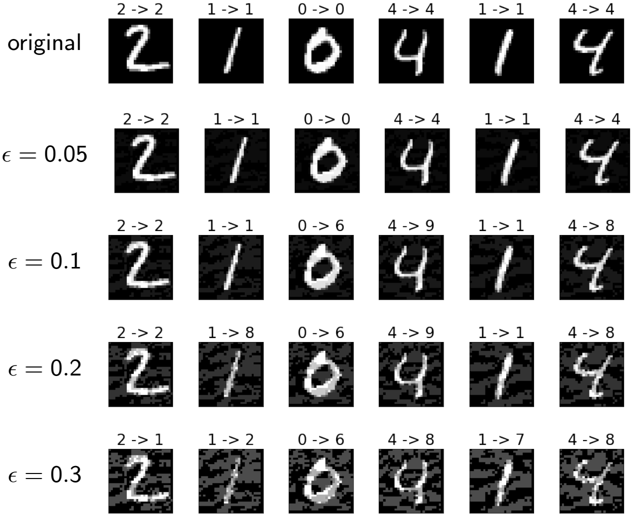
- This can be solved using gradient ascent methods.
- However, multiple function and gradient evaluations are needed, thus such methods are slow.

Goodfellow, Shlens, and Szegedy, Explaining and harnessing adversarial examples, 2015

- Approximation (linearize $L(\mathbf{x} + \delta, y, f)$)

$$\begin{aligned} \max_{\delta \in \mathbf{R}^m} \quad & L(\mathbf{x}, y, f) + \delta^\top \nabla_\mathbf{x} L(\mathbf{x}, y, f), \\ \text{s.t.} \quad & \|\delta\|_\infty \leq \epsilon. \end{aligned}$$

- Simple closed-form solution: $\delta = \epsilon \, \mathrm{sgn}(\nabla_\mathbf{x} L(\mathbf{x}, y, f))$.
- Only need to evaluate the gradient once!



FGSM examples for LeNet5 on MNIST

Black box attacks

- Black box attacks generally rely on the transferability of adversarial examples.
- In the complete black box scenario, an ensemble of models are used to increase transferability.
- When it is possible to query the target model, the attacker can use the responses to train a substitute model to increase transferability.

Defending Against Adversarial Examples

Defense methods

- Regularization: training with techniques like weight decay (i.e. ℓ_2 regularization), dropout
- Noise: add noise during training and/or testing
- Generative pre-training: learn a representation on a large unlabeled dataset using a generative model, then perform discriminative fine-tuning on a labeled dataset
- Ensembles: train on adversarial examples constructed for multiple models
- ...

- Attacking is easy
  - Various methods available, and transferability makes attacks easy even without much knowledge about the target model.
- Defending is difficult
  - No single method has been found to be very effective yet.

Adversarial Learning

Train with adversarial examples

- To make a model robust to adversarial examples, we can create many adversarial examples, and add them to the training set.
- This can be done iteratively by adding adversarial examples for the intermediate models.

Robust training objective

- Another approach is to explicitly modify the training objective to incorporate a term against adversarial examples
- A modified loss against FGSM

$$\tilde{L}(\mathbf{x}, y, f) = \alpha L(\mathbf{x}, y, f) + (1-\alpha) L(x + r\,\mathrm{sgn}(\nabla_{\mathbf{x}} L(\mathbf{x}, y, f)), y, f)$$

- The loss of the perturbed example can't be too much different from that or the original example.
- This prevents FGSM from significantly reducing the loss of the true label.
- Thus FGSM is less likely to change the class label.

What You Need to Know

- Adversarial examples: universality, transferability and implications
- Adversarial attacks: minimum perturbation method, fast gradient sign method.
- Defending against adversarial examples
- Adversarial learning: augment dataset with adversarial examples, robust training objectives.