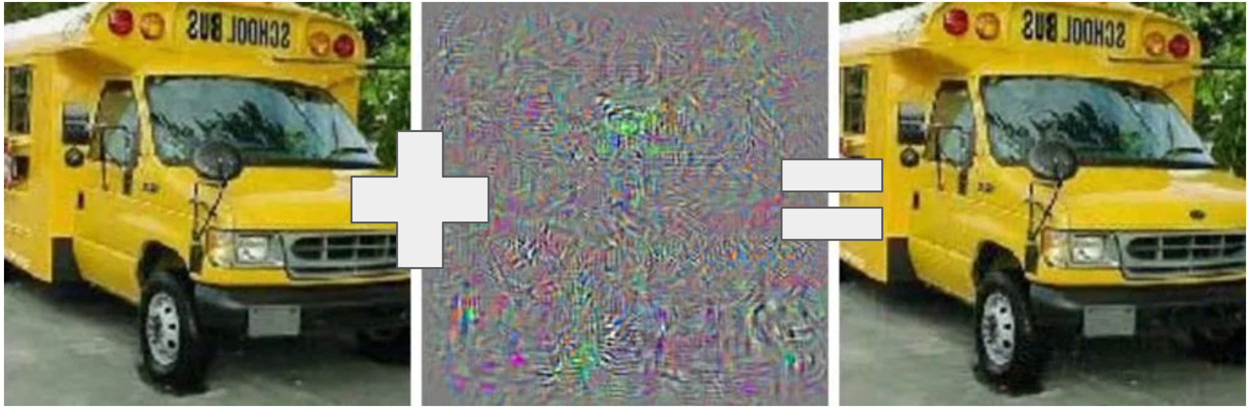
Deep Neural Networks are easily fooled

High Confidence Predictions for Unrecognizable Images

a paper of Anh Nguyen, Jason Yosinki, Jeff Clune

presented by Nils Wenzler

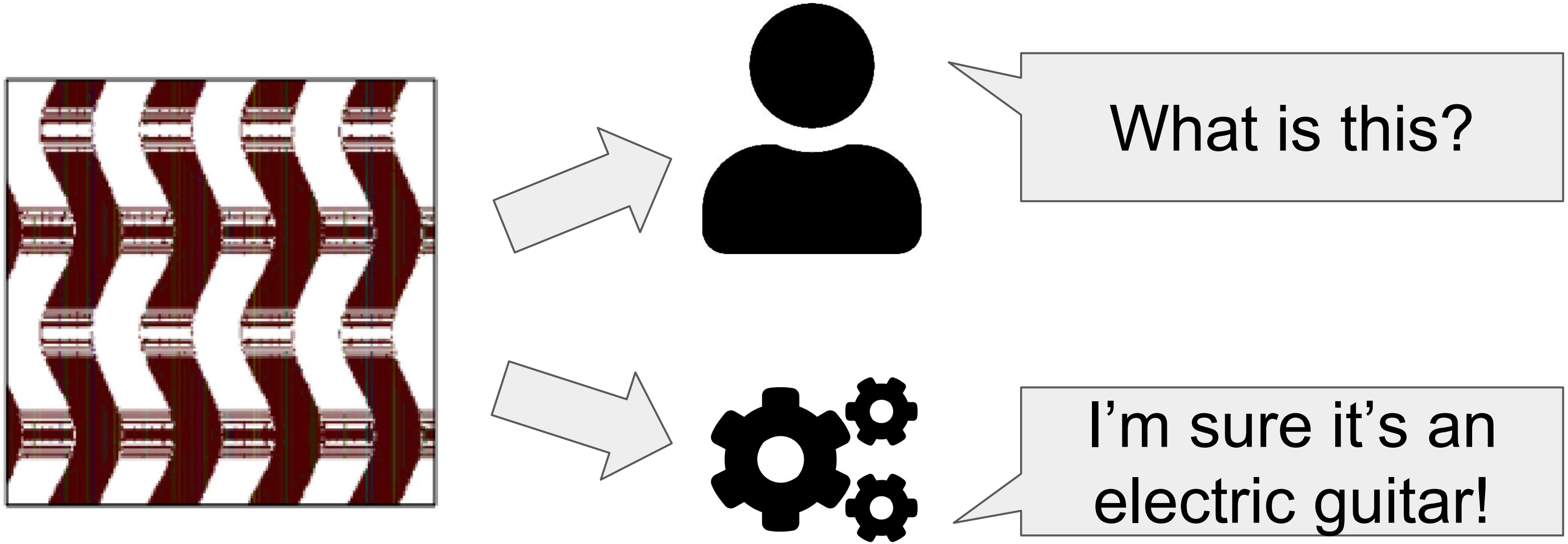
General idea: Fooling neural networks



school bus | imperceivable change | ostrich

from “Intriguing properties of neural networks” by Szegedy et al.

General idea:
Generate pictures that AI can “see” but humans not



But why?

Original goal:

- Visualize DNN perception
- Explain to what kind of features DNNs react to

Extended goal:

- Explain why DNNs are so easily fooled
- Research how to possibly improve resilience and robustness

Structure

Chapter 1: How to create fooling pictures?

Chapter 2: How do nowadays networks “see”?

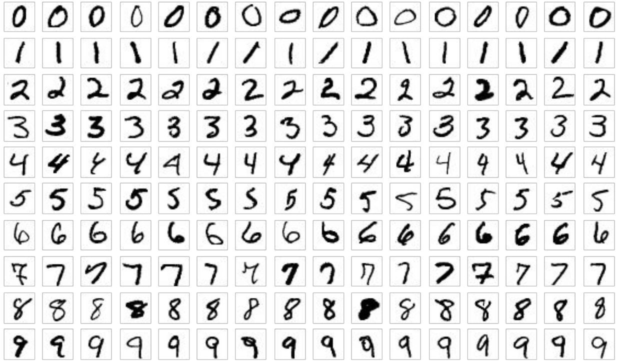
Chapter 3: How to defend against adversarial pictures?

Chapter 4: Lessons learned

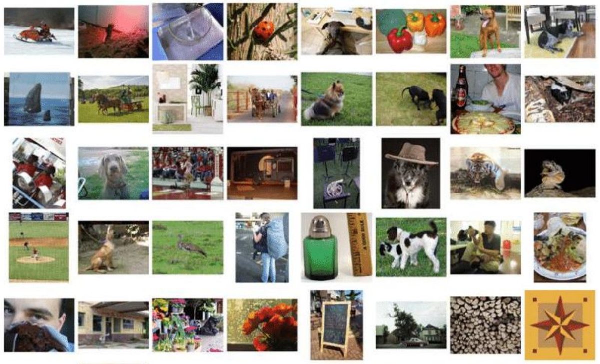
Chapter 1: How to create fooling pictures?

Data sets:

MNIST



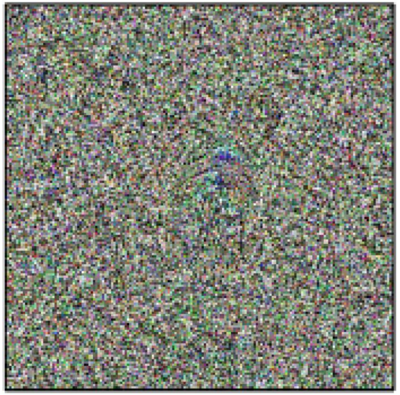
ImageNet



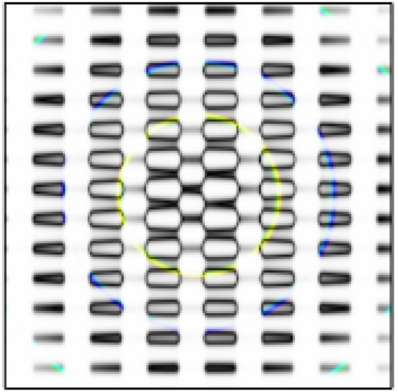
Chapter 1: How to create fooling pictures?

3 approaches:

direct model



indirect model



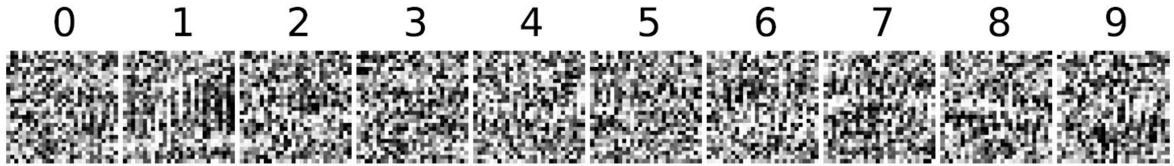
gradient ascend

Direct model

Learn all pixels with evolutionary algorithm

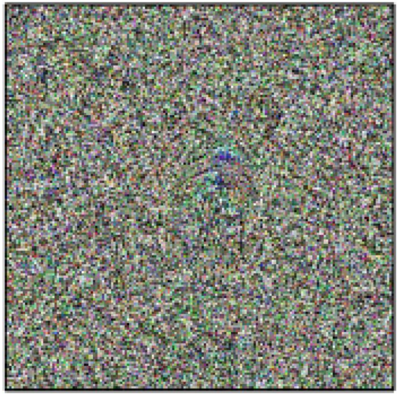
Results:

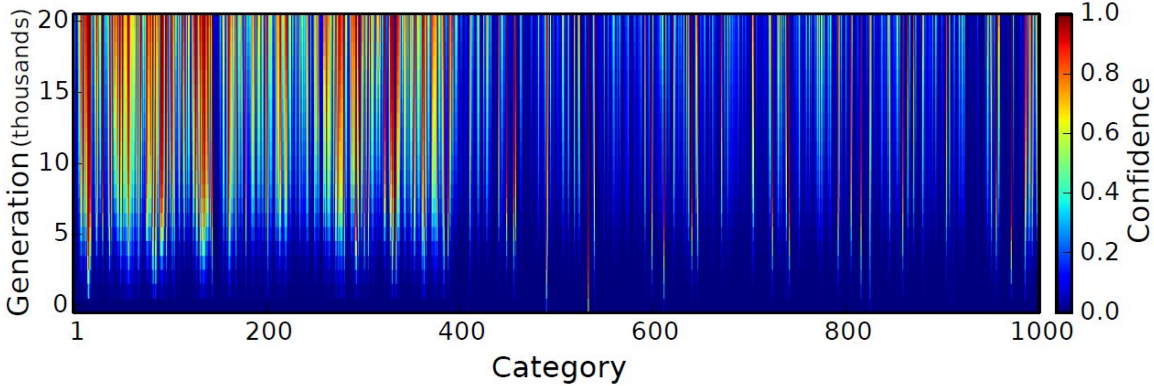
- 99,9% confidence that images are numbers (MNIST)



- Performance in ImageNet classification not very convincing

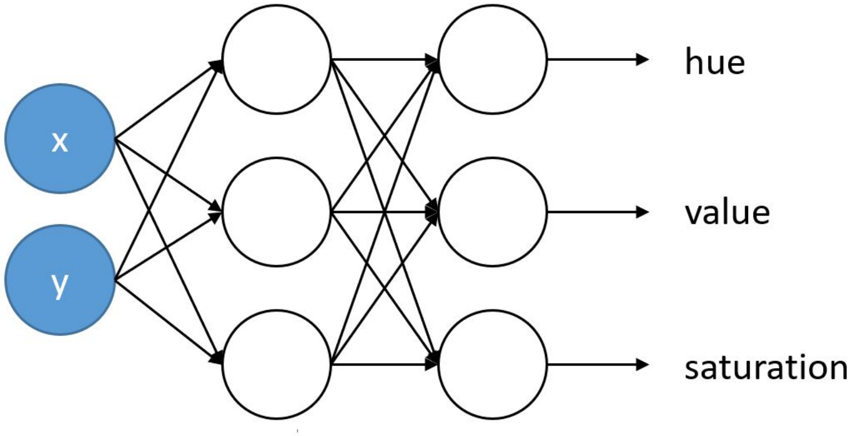
Direct model - ImageNet classification





Indirect model

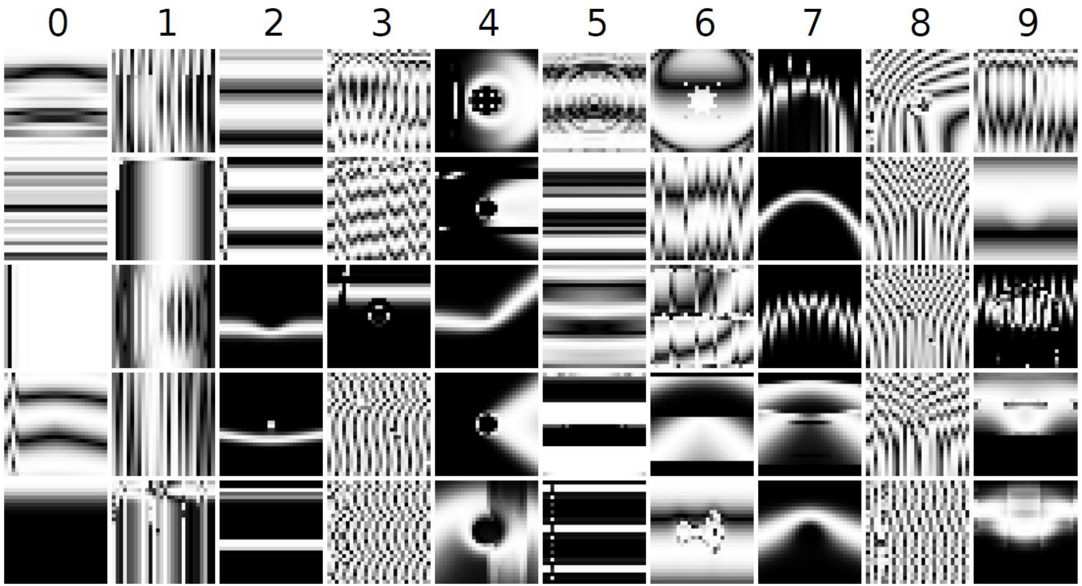
Evolve on compositional pattern-producing networks (CPPN)



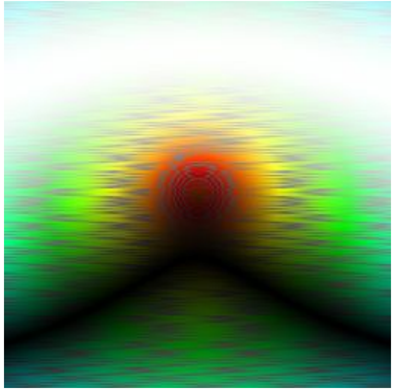
Indirect model

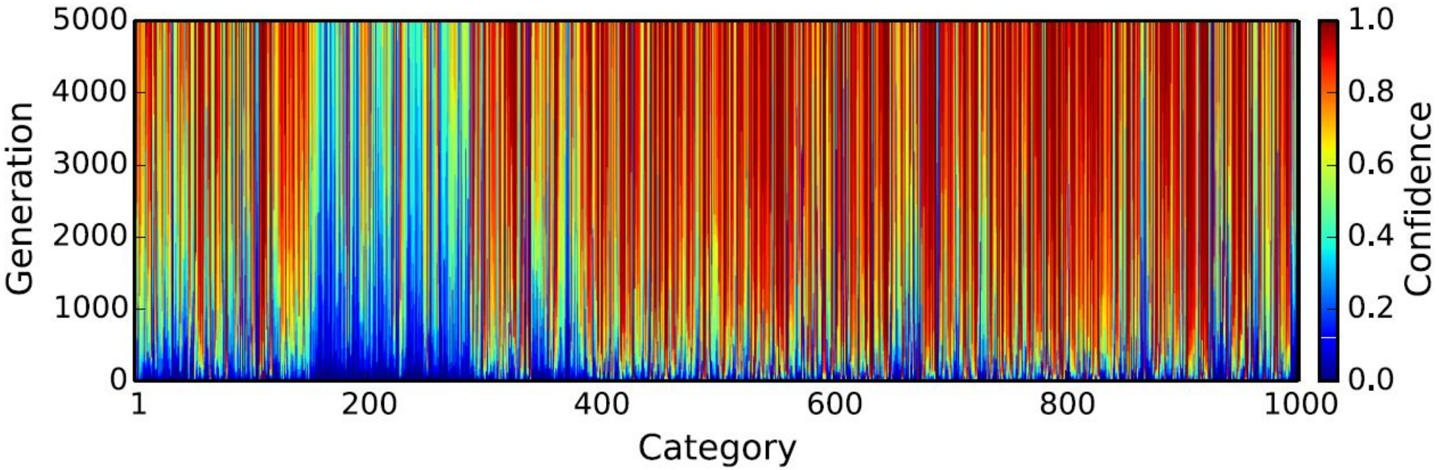
Results:

- 99,9% confidence that images are numbers (MNIST)



Indirect model - ImageNet classification





Gradient ascend

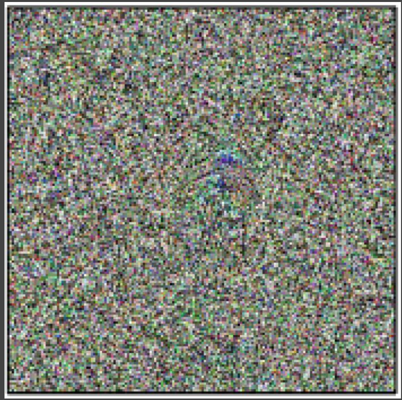
Whitebox approach which mathematically optimizes input image

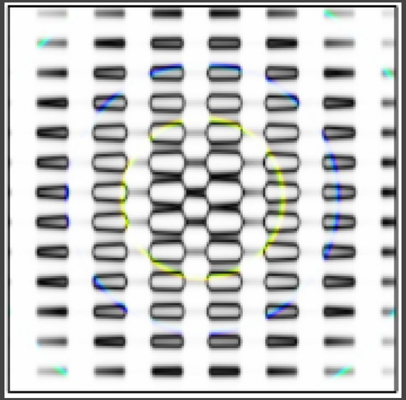
Works but not further considered because:

- Visualization is hard to understand
- Danger of being very network specific

Chapter 2: How do neural networks “see”?

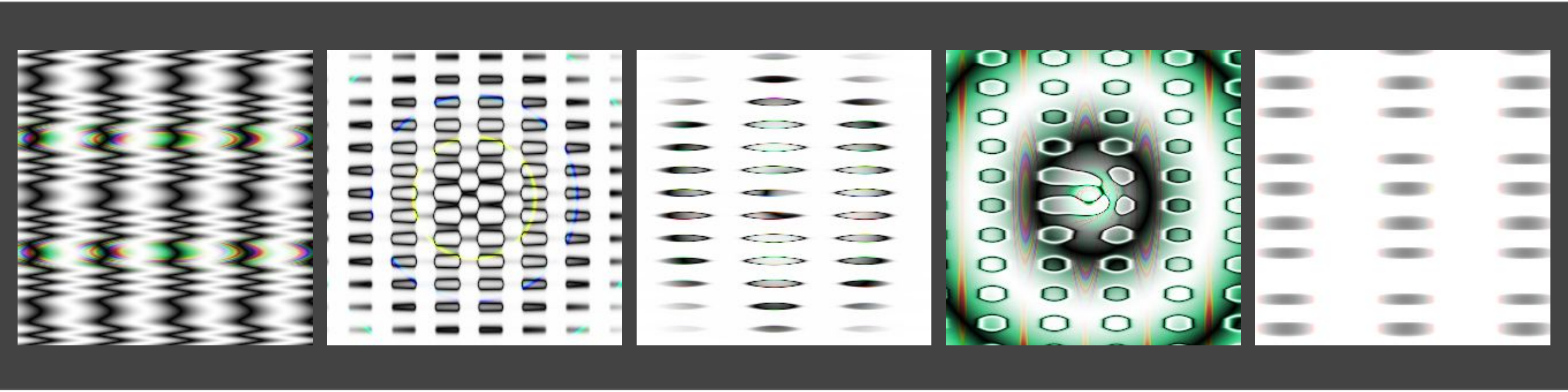
They “see” differently than human beings





Feature size

5 times a 99,9% confidence remote control:



Tend to react to medium sized features but not whole structures

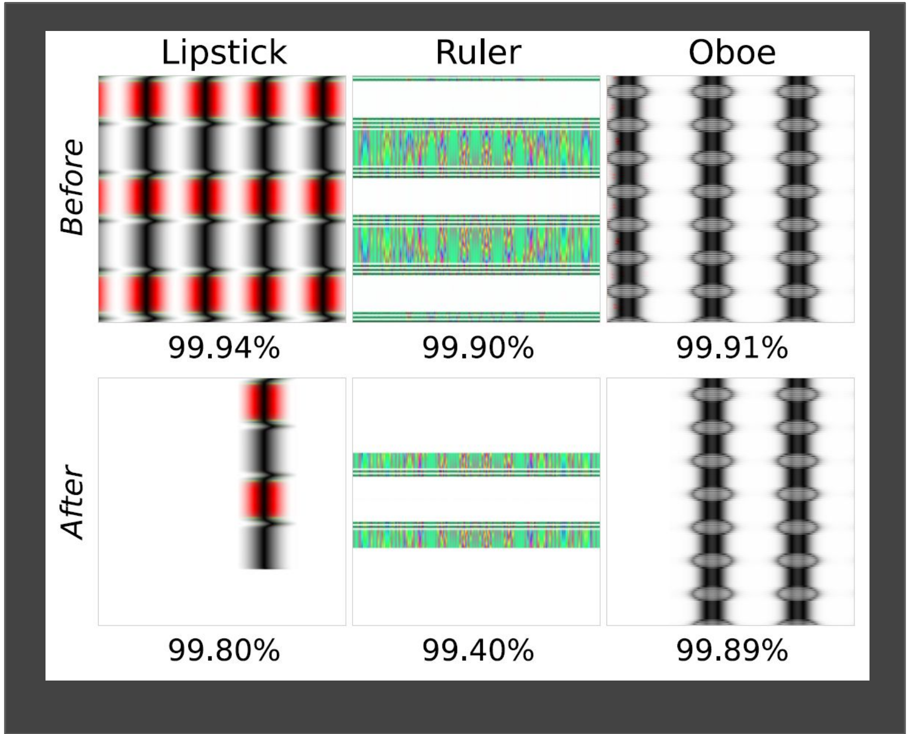
Adversarial images generalize to other networks

Some adversarial images trained for a DNN can fool another DNN as well.
Some images will only fool one of them.

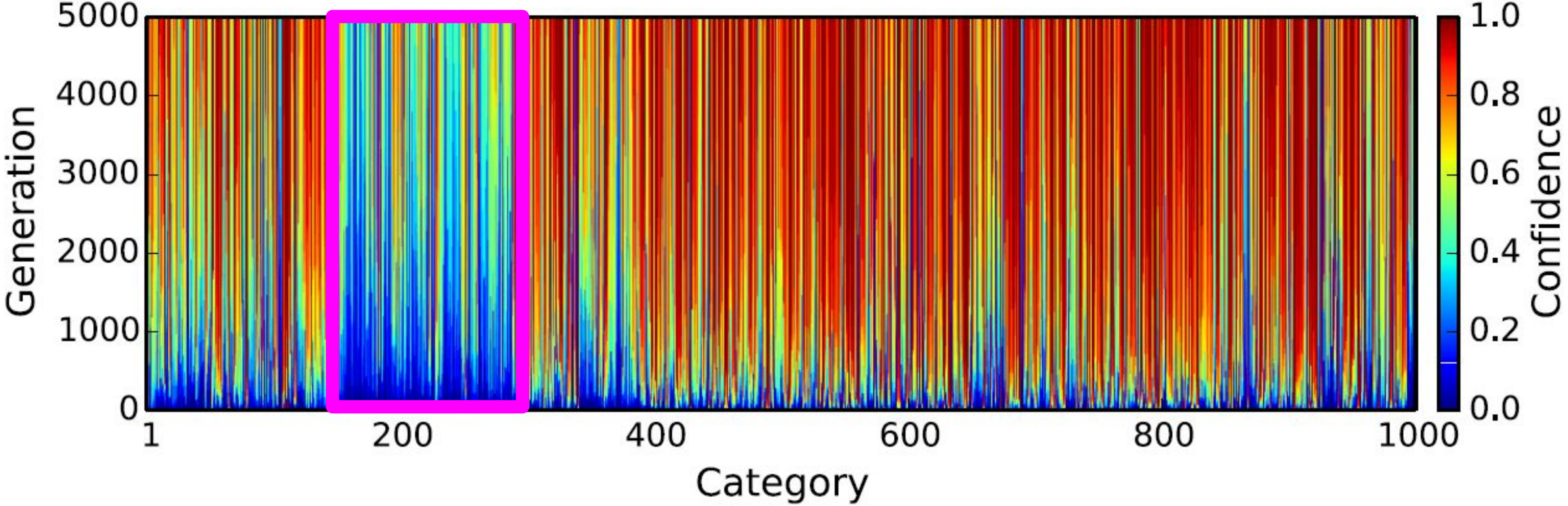
⇒ DNNs tend to observe similar features

Repeated instances

Several instances seem to improve confidence



Dogs are hard to differentiate



Chapter 3:
How do defend against adversarial pictures?

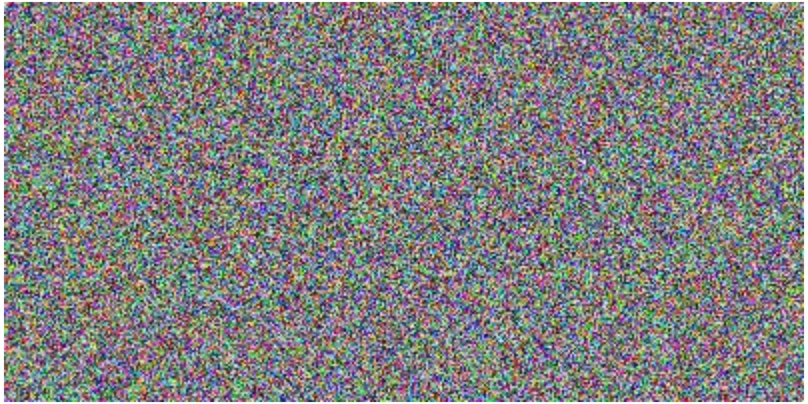
Just add adversarial pictures to training set with adversarial class?

- still easily fooled for MNIST digit recognition task
- learned to classify CPPN pictures for ImageNet (no exhaustive defense)

What’s the problem?

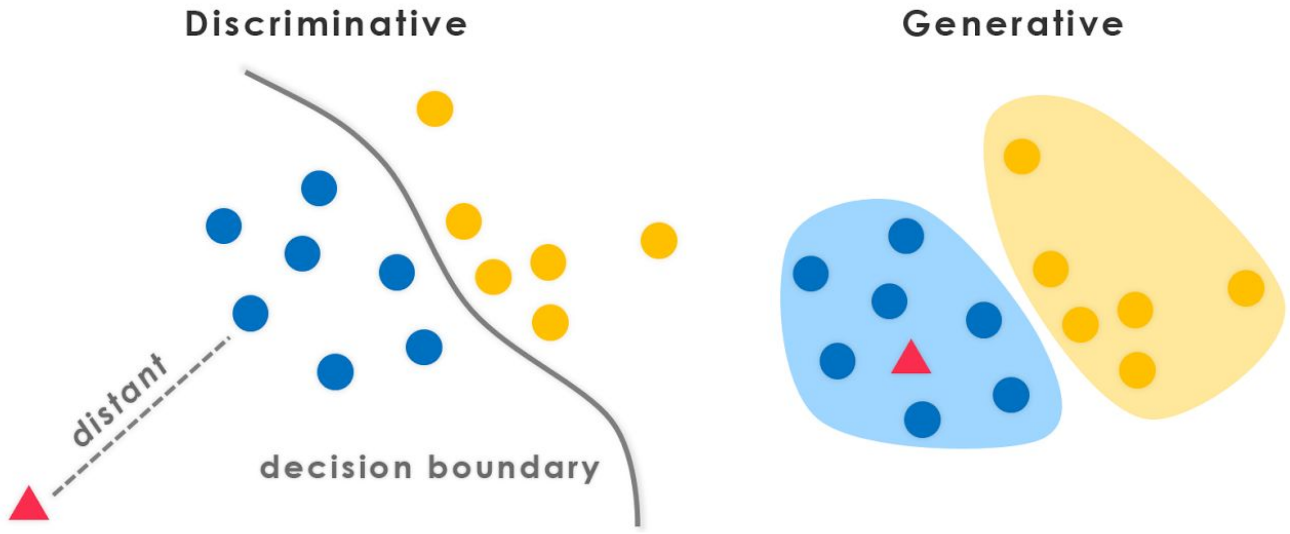
Image classification has a very high input space but significantly lower intrinsic dimensionality

If you create a random picture, how would it look like?



Discriminative versus Generative approaches

p(y|X) versus p(y, X)



Chapter 4: Lessons learned

Neural networks currently

- learn medium sized features
- can easily be fooled
- become more confident through feature repetition

Chapter 4: Lessons learned

Adversarial images

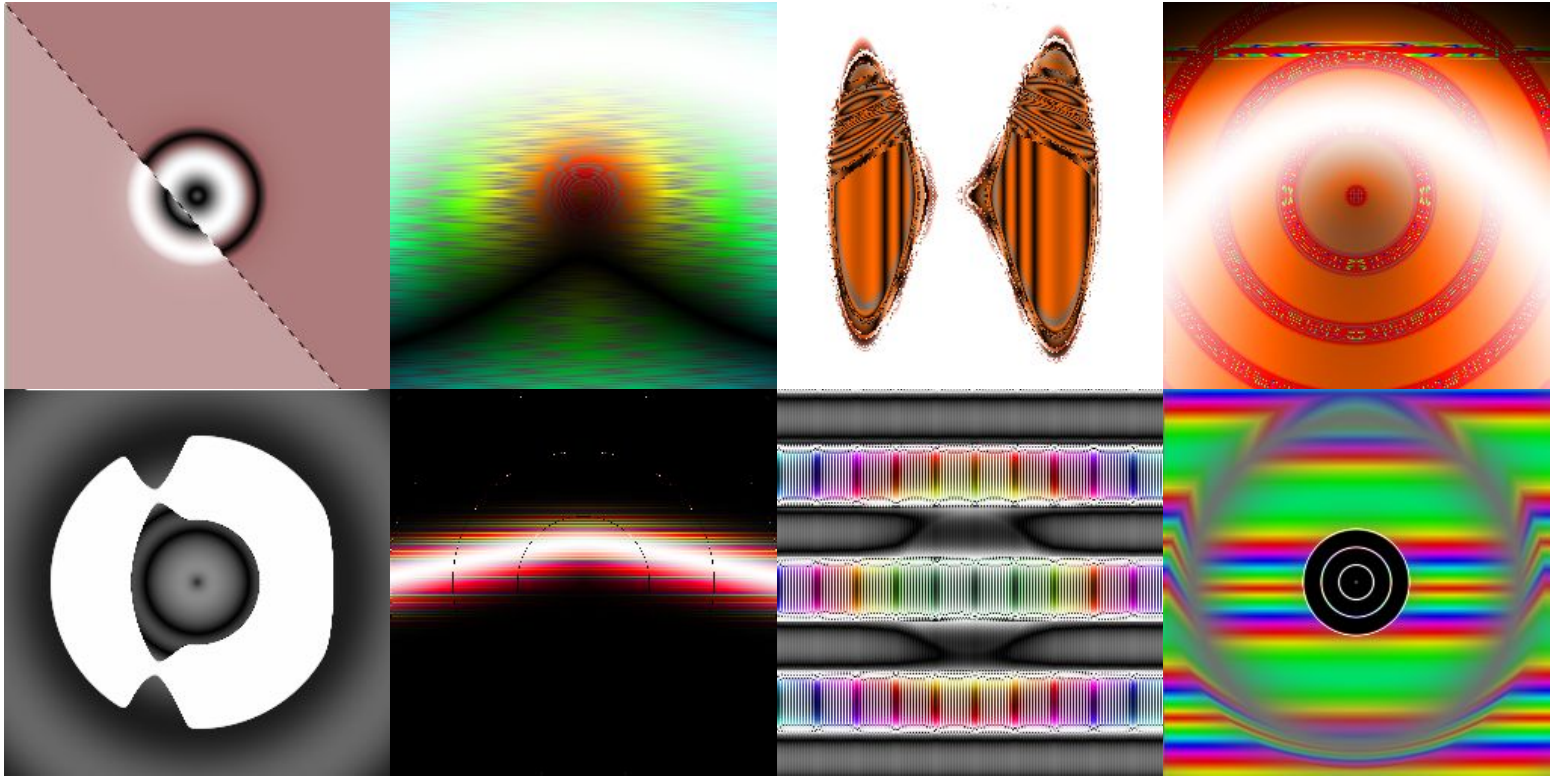
- can be hidden in perfectly normal pictures
- can be abstract images
- can generalize to other networks
- can not be dealt with by adding them to the training set

Chapter 4: Lessons learned

Furthermore,

- natural images have lower intrinsic dimensionality
- generative models may be more robust than discriminative models
- bigger and more diverse training sets make fooling harder

Neural networks as artists?



Thanks for your attention

Let's discuss the paper:

- natural images have lower intrinsic dimensionality
- generative models may be more robust than discriminative models
- bigger and more diverse training sets make fooling harder