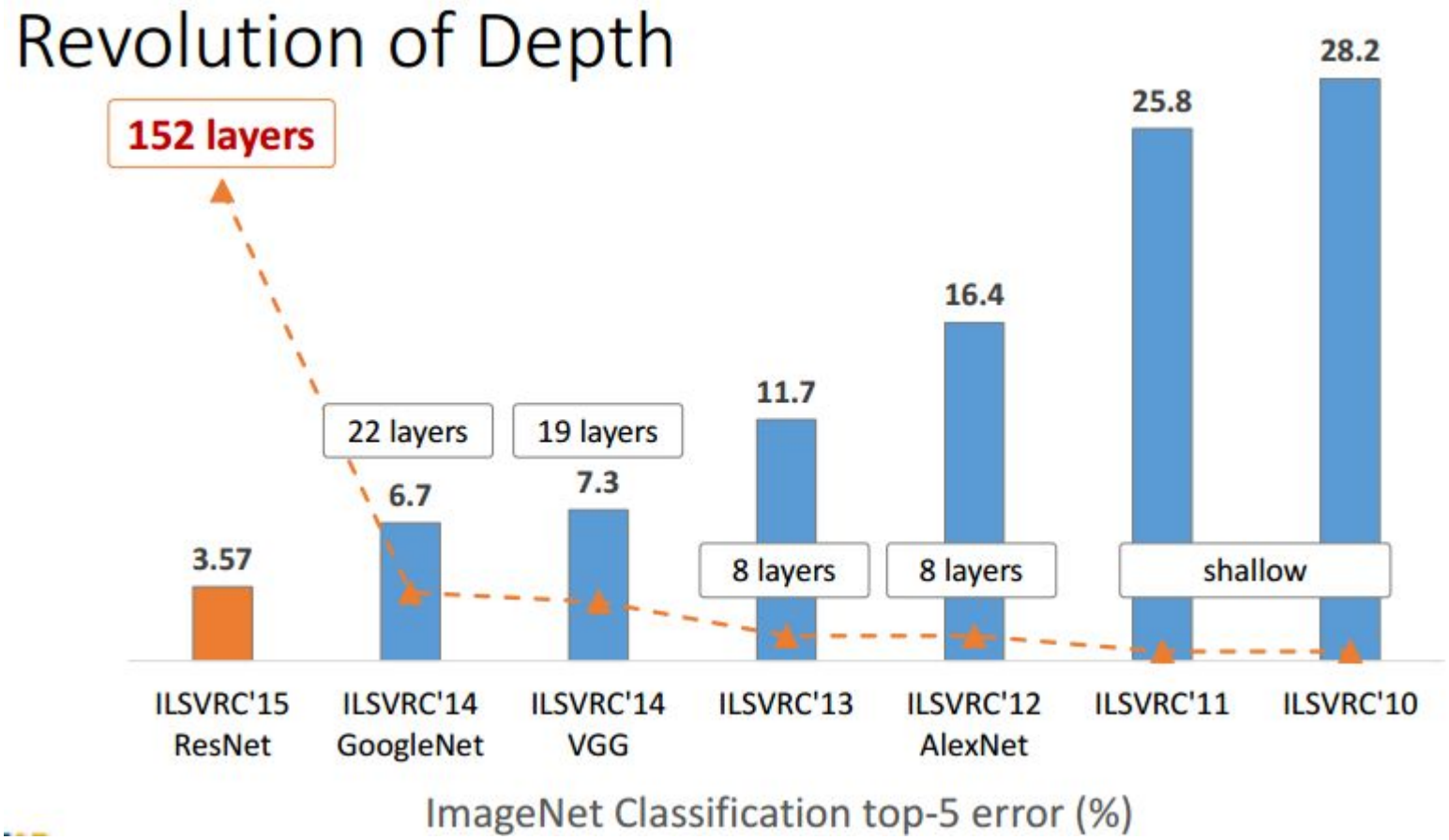
Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images

Anh Nguyen
University of Wyoming
anguyen8@uwyo.edu

Jason Yosinski
Cornell University
yosinski@cs.cornell.edu

Jeff Clune
University of Wyoming
jeffclune@uwyo.edu

Presenter: Zhenghao Fei
10/20/2016 CS289 Class



However!

Slide from Kaiming He's presentation
http://kaiminghe.com/ilsvrc15/ilsvrc2015_deep_residual_learning_kaiminghe.pdf

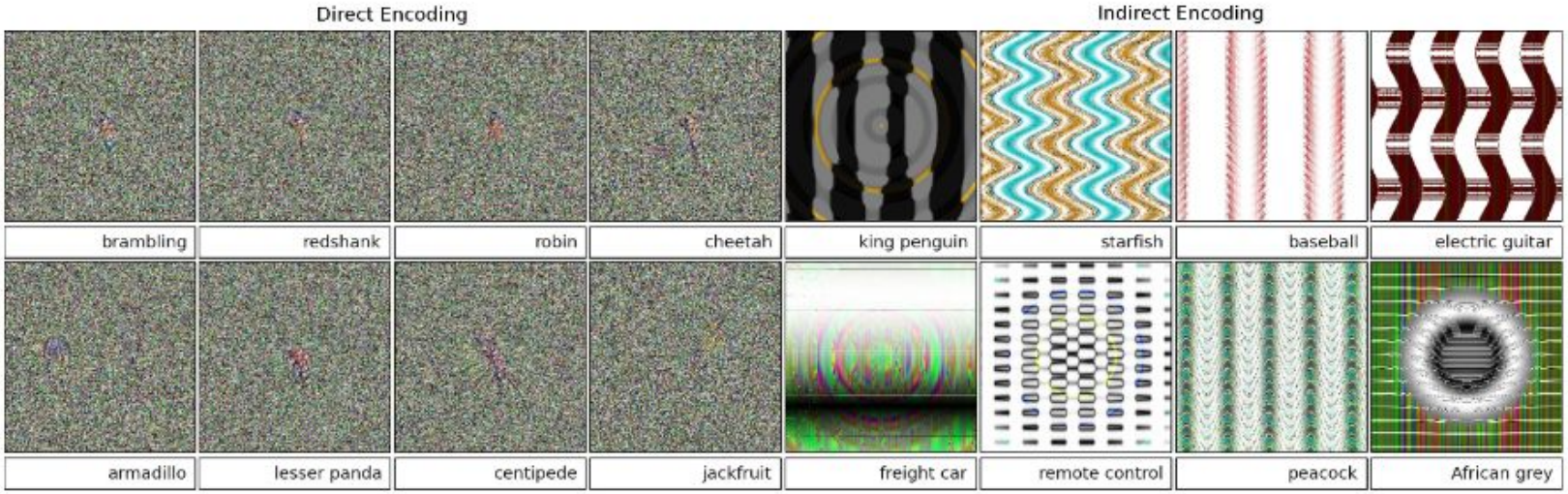
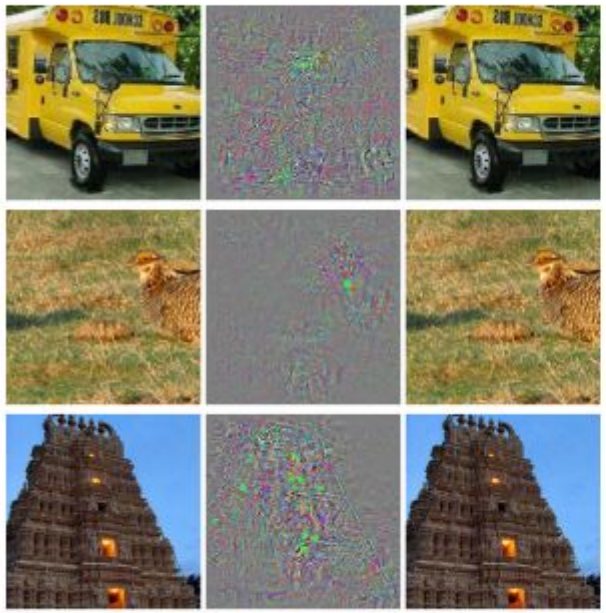


Figure 1: Evolved images that are unrecognizable to humans, but that state-of-the-art DNNs trained on ImageNet believe with >= 99.6% certainty to be a familiar object. This result highlights differences between how DNNs and humans recognize objects. Left: Directly encoded images. Right: Indirectly encoded images.

From http://www.evolvingai.org/fooling

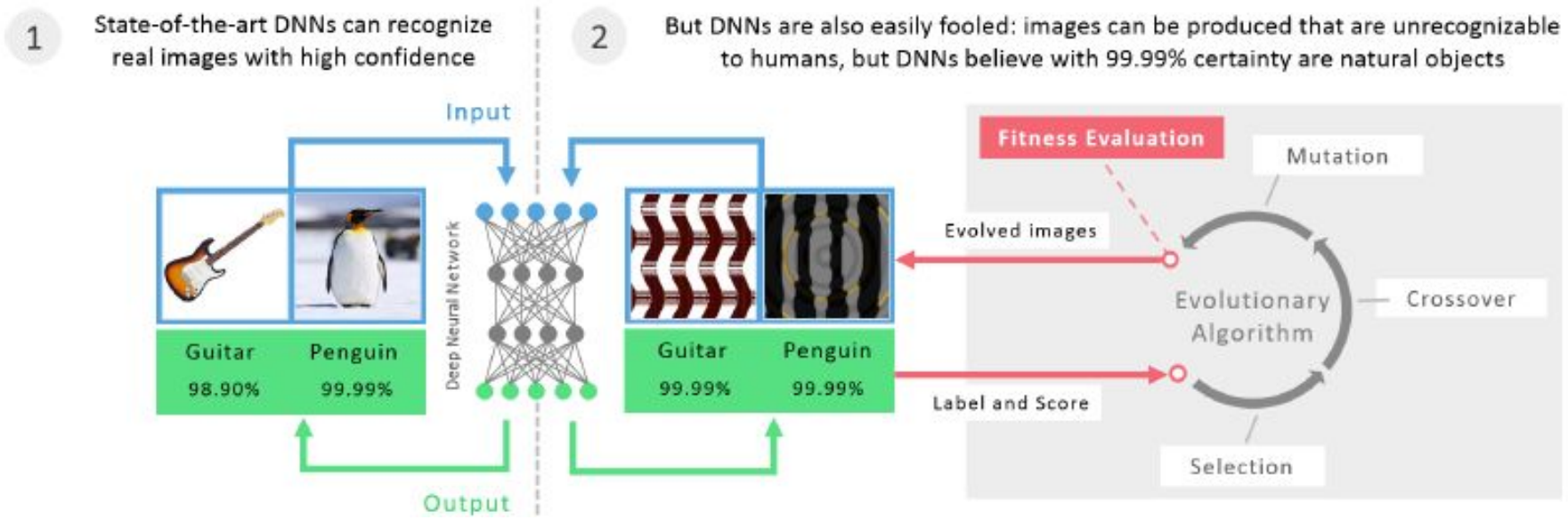
Bus? Not a bus?



(a)

(b)

Christian Szegedy et al Intriguing properties of neural networks
https://arxiv.org/pdf/1312.6199v4.pdf

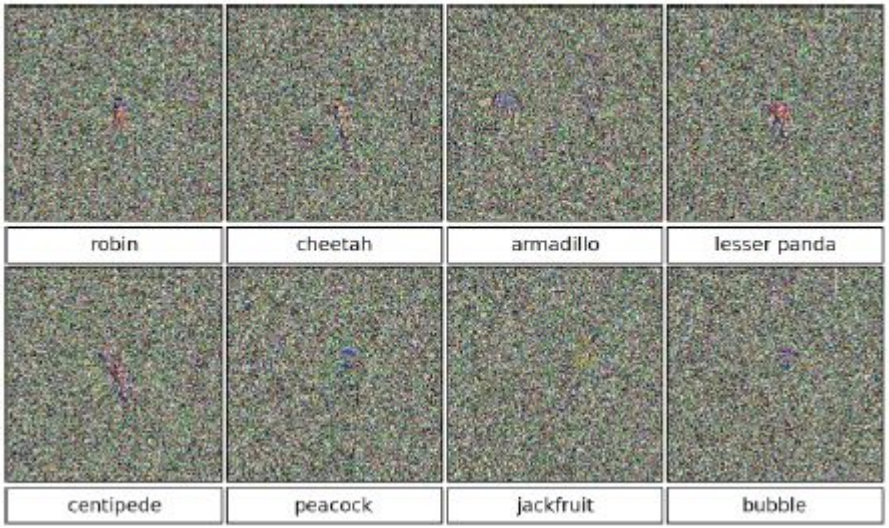


The CNN model they fooled:
AlexNet
Poor AlexNet

Using evolutionary algorithms or gradient ascent to generate images that are given high prediction scores by convolutional neural networks

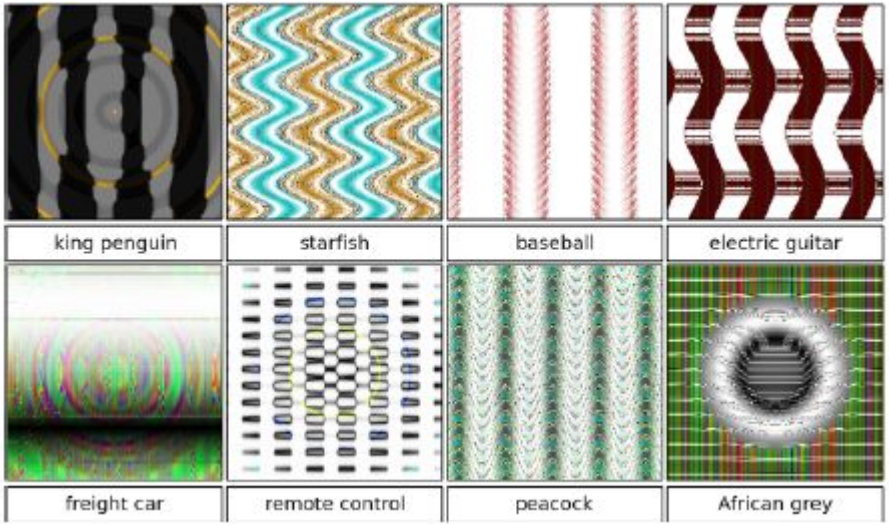
Evolutionary algorithms

Direct encoding & Indirect encoding



Direct encoding:

Each pixel value is initialized with uniform random noise within the range



Indirect encoding:

Compositional Pattern-Producing Network (CPPN)
More complex, regular image that resemble nature an man-made objects

Results

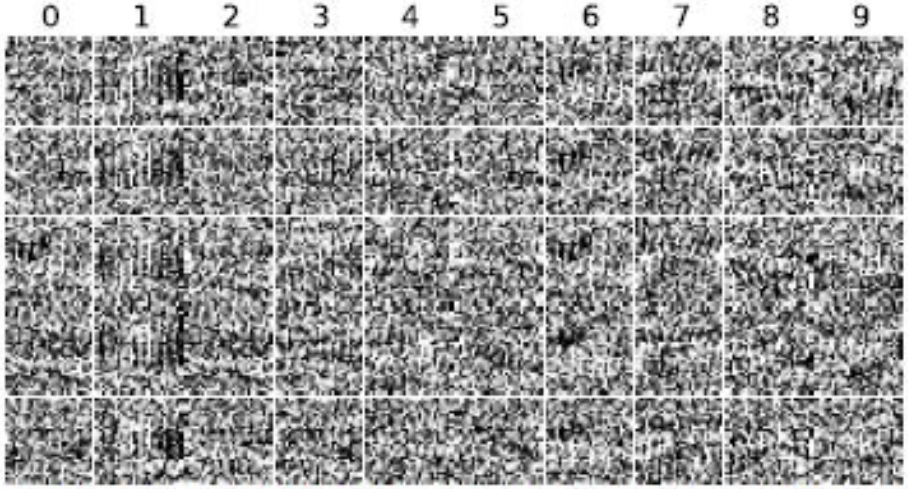


Figure 4. Directly encoded, thus irregular, images that MNIST DNNs believe with 99.99% confidence are digits 0-9. Each column is a digit class, and each row is the result after 200 generations of a randomly selected, independent run of evolution.

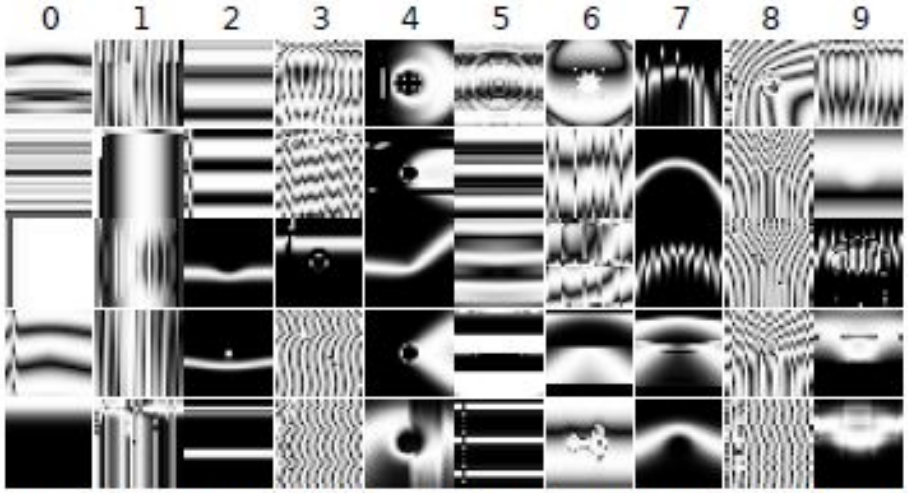


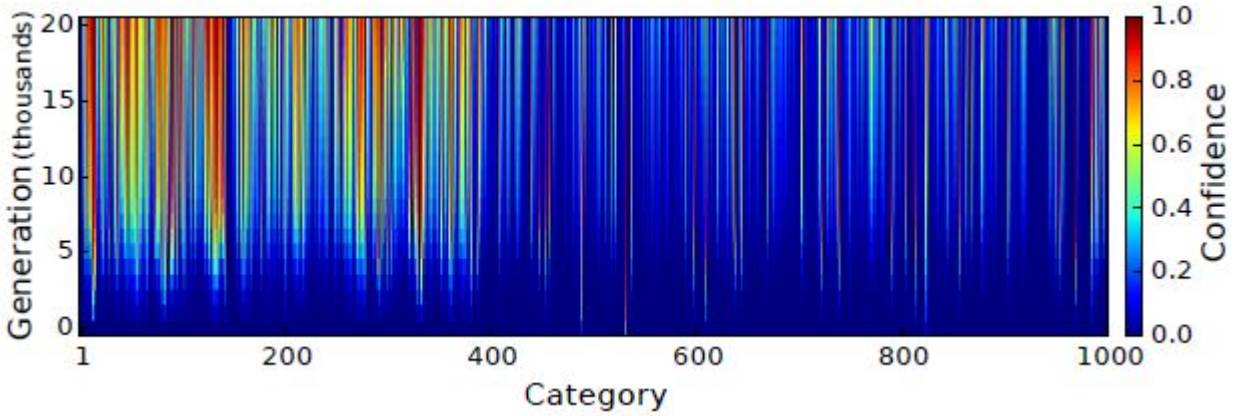
Figure 5. Indirectly encoded, thus regular, images that MNIST DNNs believe with 99.99% confidence are digits 0-9. The column and row descriptions are the same as for Fig. 4.

DataSet: MNIST

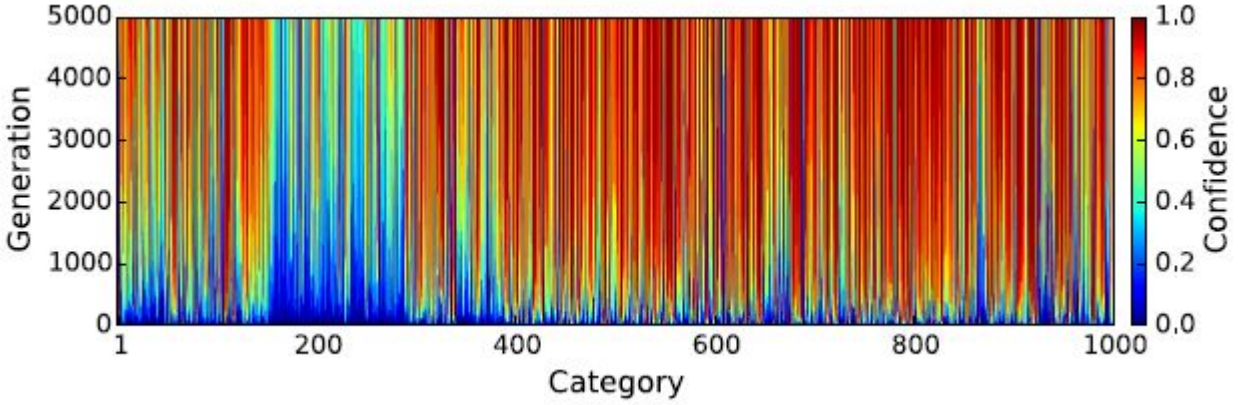
Results

45 classed with >= 99% confidence

Direct encoding: median score is 21.59%

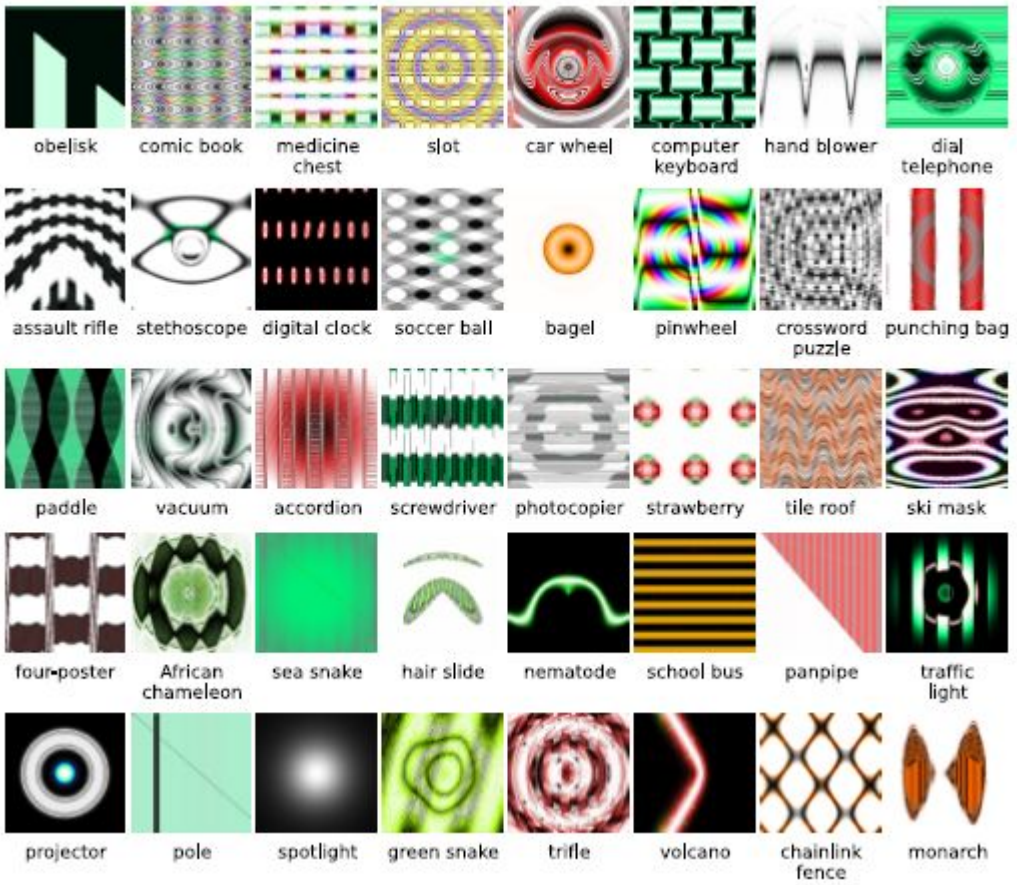


Indirect encoding: median score is 88.11%



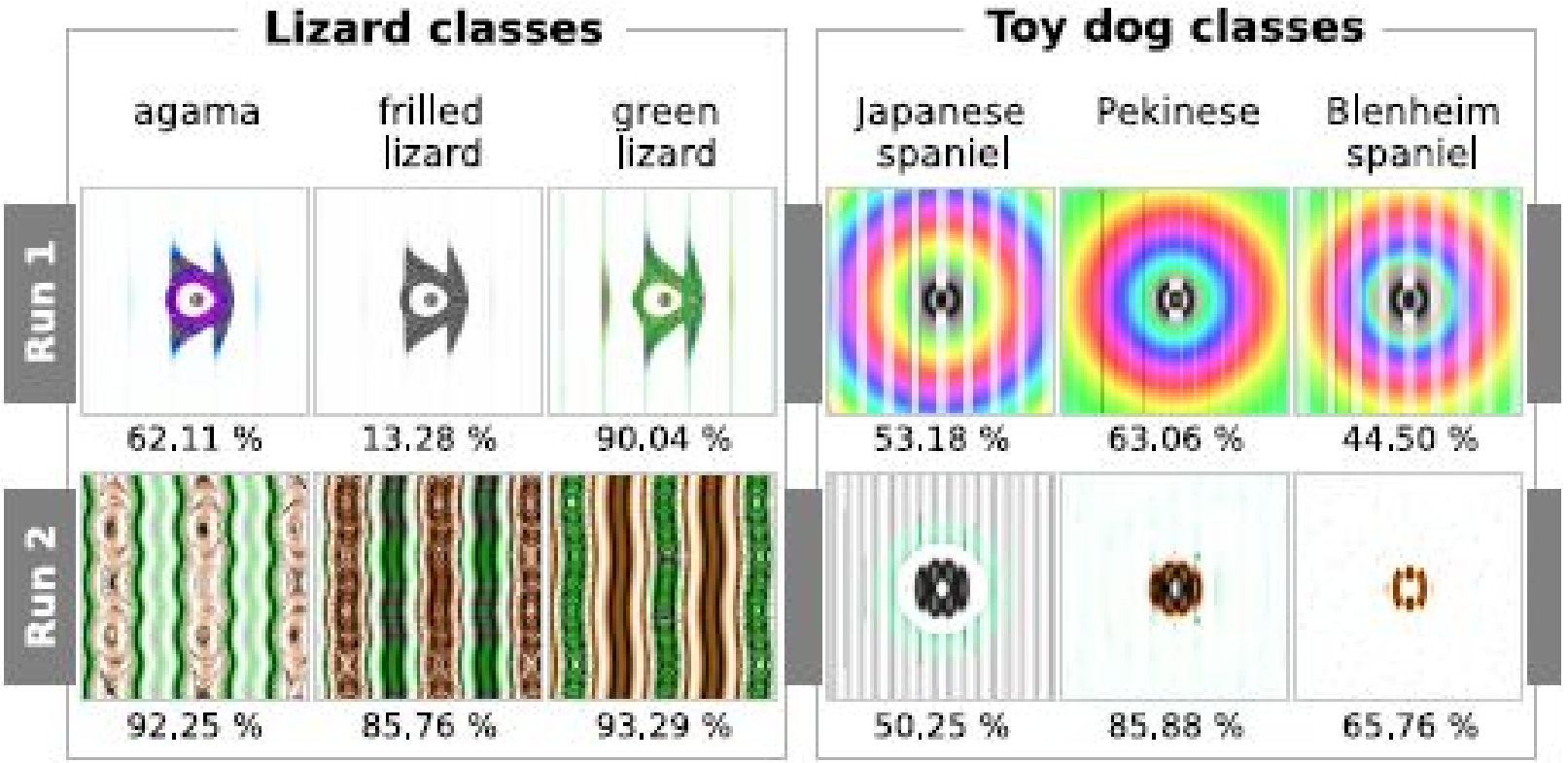
DataSet: ImageNet

Evolving images to match DNN classes produces a tremendous diversity of images



Only need to produce features that are unique to,
or discriminative for, a class,
rather than produce an image that contains all of the typical features of a class.

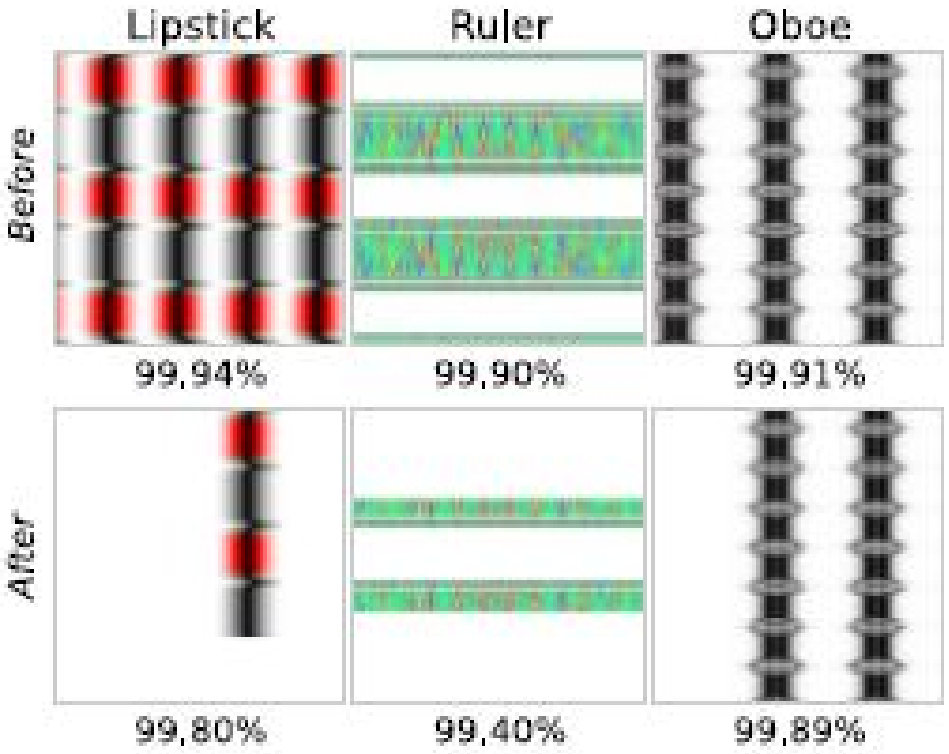
Different runs of evolution, however, produce different image types for these related categories, revealing that there are different discriminative features per class that evolution exploits.



Local? Global?

Extra copies make the DNN more confident that the image belongs to the target class.

These results suggest that DNNs tend to learn lowand middle-level features rather than the global structure of objects



Images that fool one DNN generalize to others

(1) DNNA and DNNB have identical architectures and training, differ only in their randomized initializations;

(2) DNNA and DNNB have different DNN architectures, but are trained on the same dataset.

Most fool image can fool both, while some can’t.

How about training networks to recognize fooling images ?

“fooling images” and can go in the n+1 category

1. **Training MNIST DNNs with fooling images**

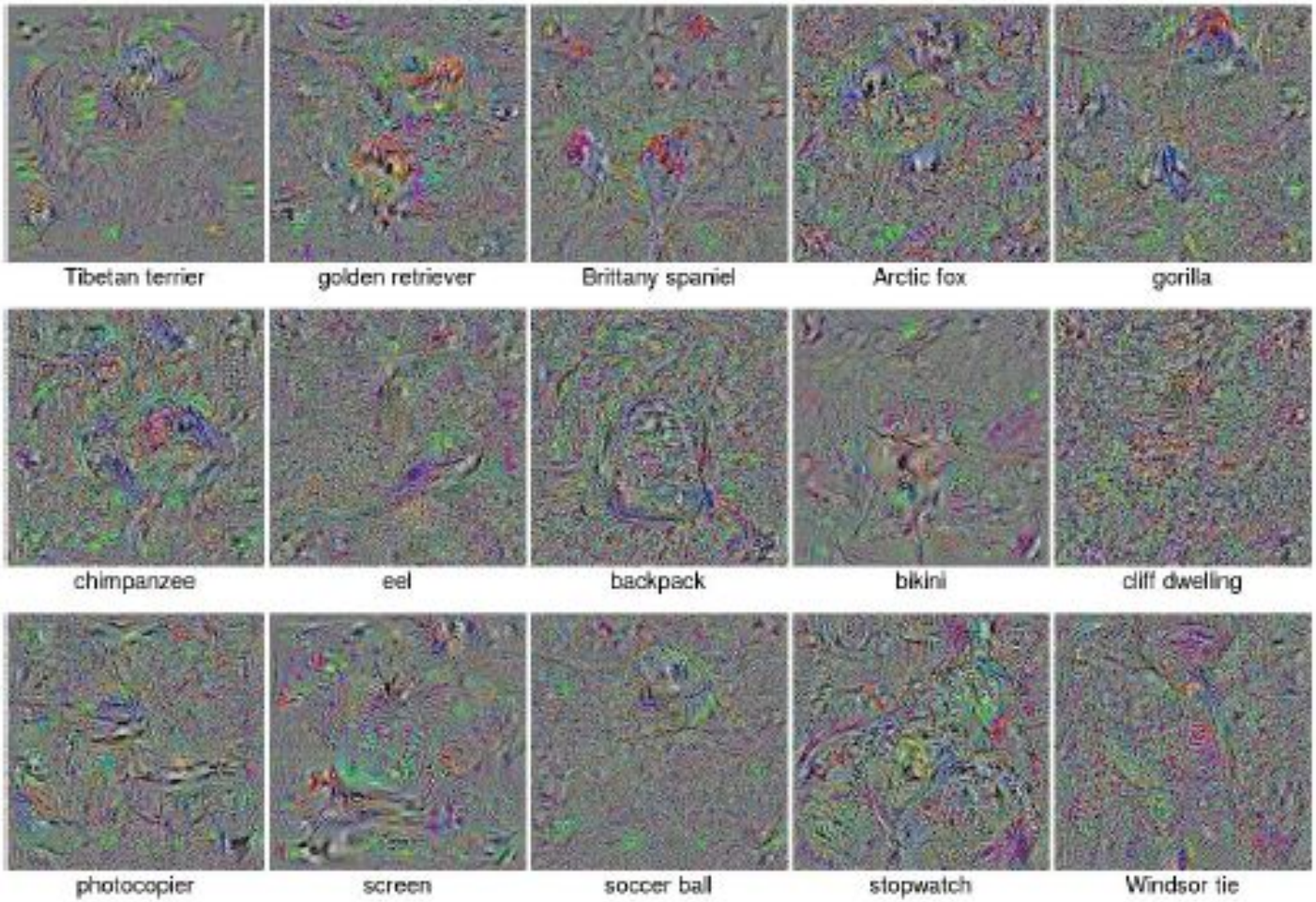
Evolution still produces many unrecognizable images for DNN2 with confidence scores of 99.99%.

2. **Training ImageNet DNNs with fooling images**

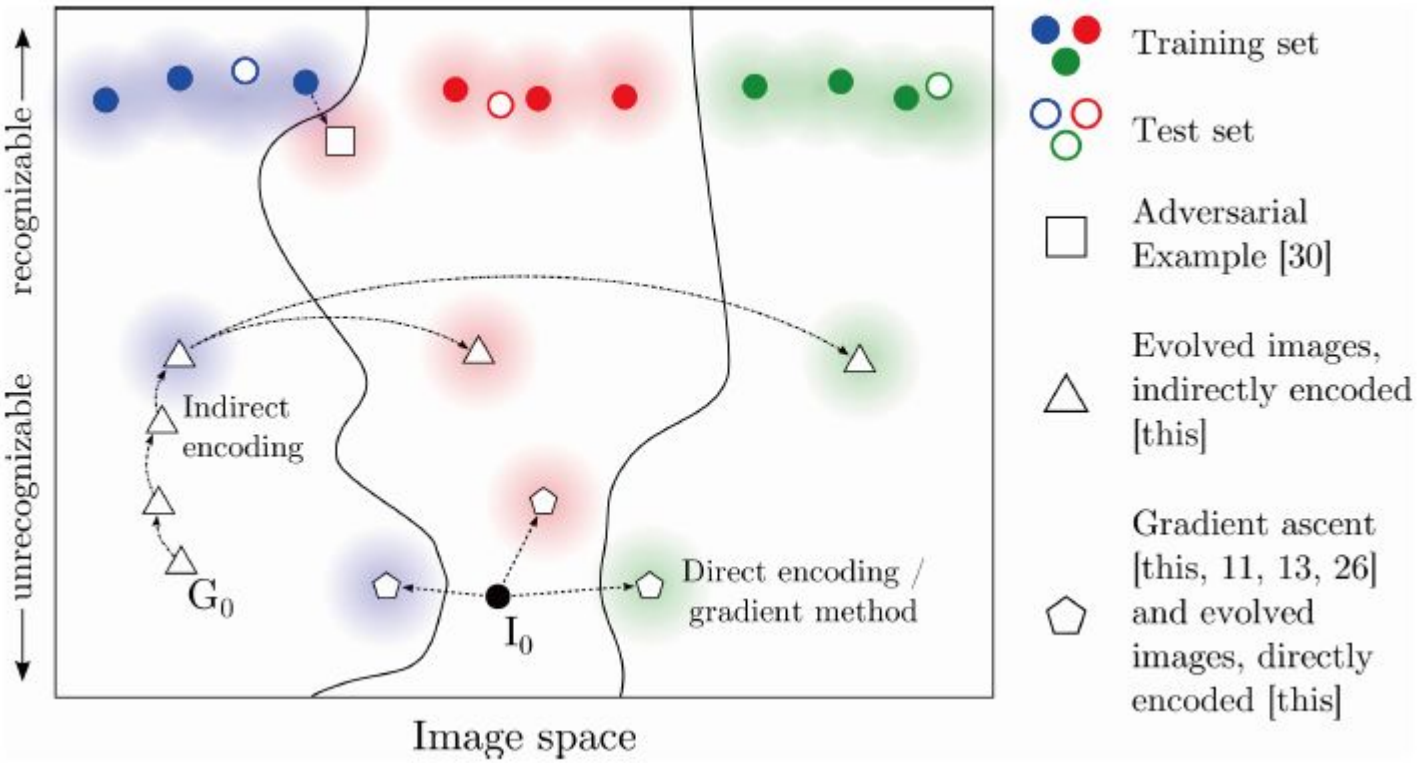
The median confidence score significantly decreased from 88.1% for DNN1 to 11.7% for DNN2

Producing fooling images via gradient ascent

with respect to the input image using backprop, and then we follow the gradient to increase a chosen unit's activation.



Evolution produced high-confidence unrecognizable images



Discriminative model: learn p(y|X)

Generative model: learn p(y, X)

Where y is a label vector and X is input example

Concerns:

A security camera that relies on face or voice recognition.

Image-based search engine rankings.

Safety-critical ones such as driverless cars.

From google image search

From Google Self-Driving Car Project