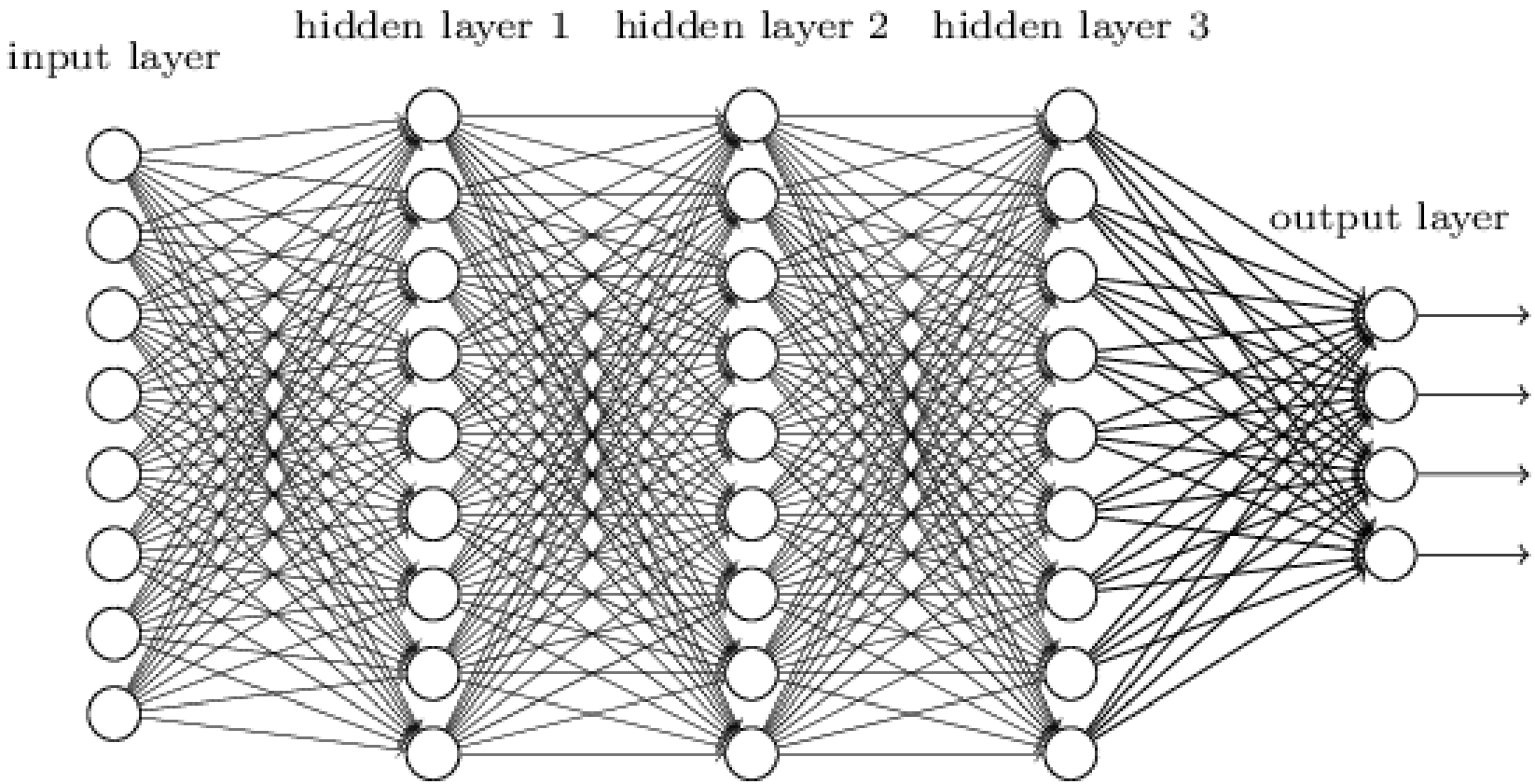
Deep Dreams

Toronto Deep Learning Meetup

Alex Yakubovich

July 30, 2015

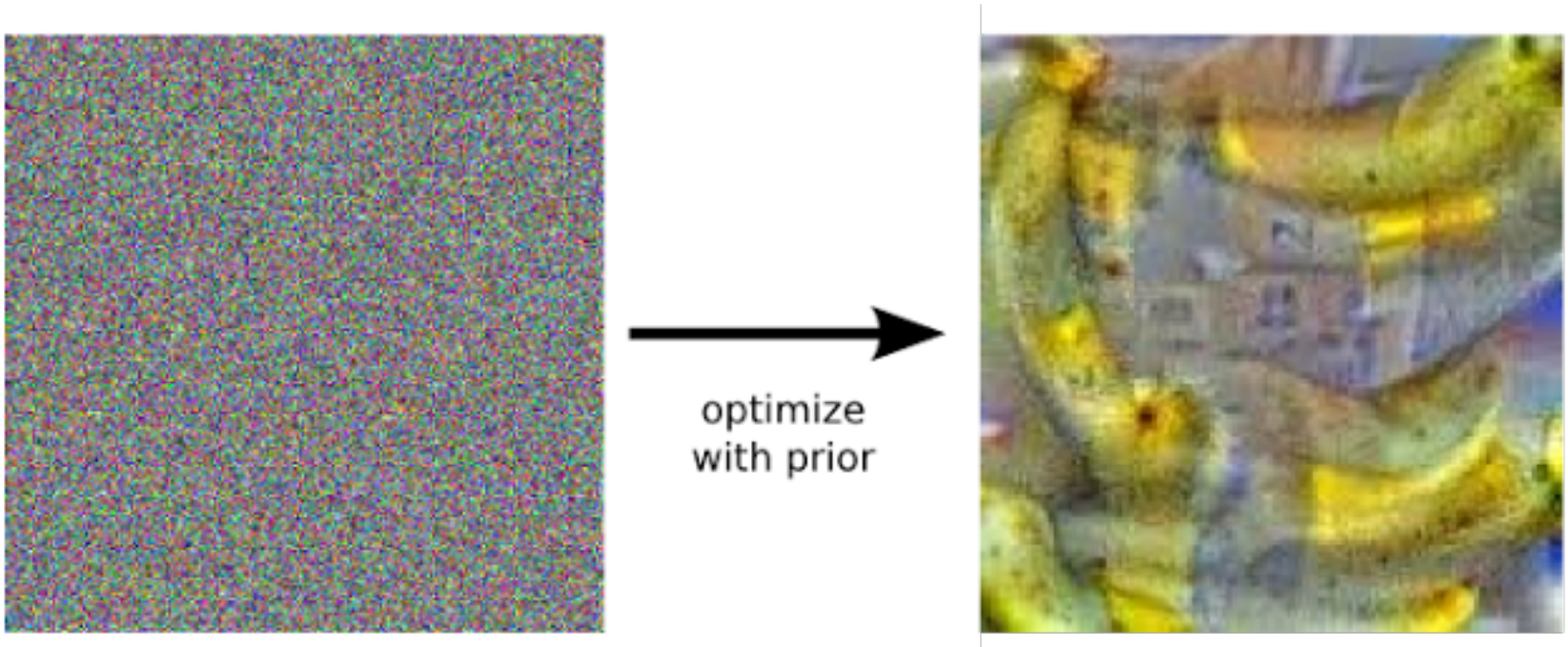
Inverting Neural Networks



[0]Taken from `http://neuralnetworksanddeeplearning.com`

Inverting Neural Networks

Instead of making predictions, change the input image to enhance a particular interpretation. [1]

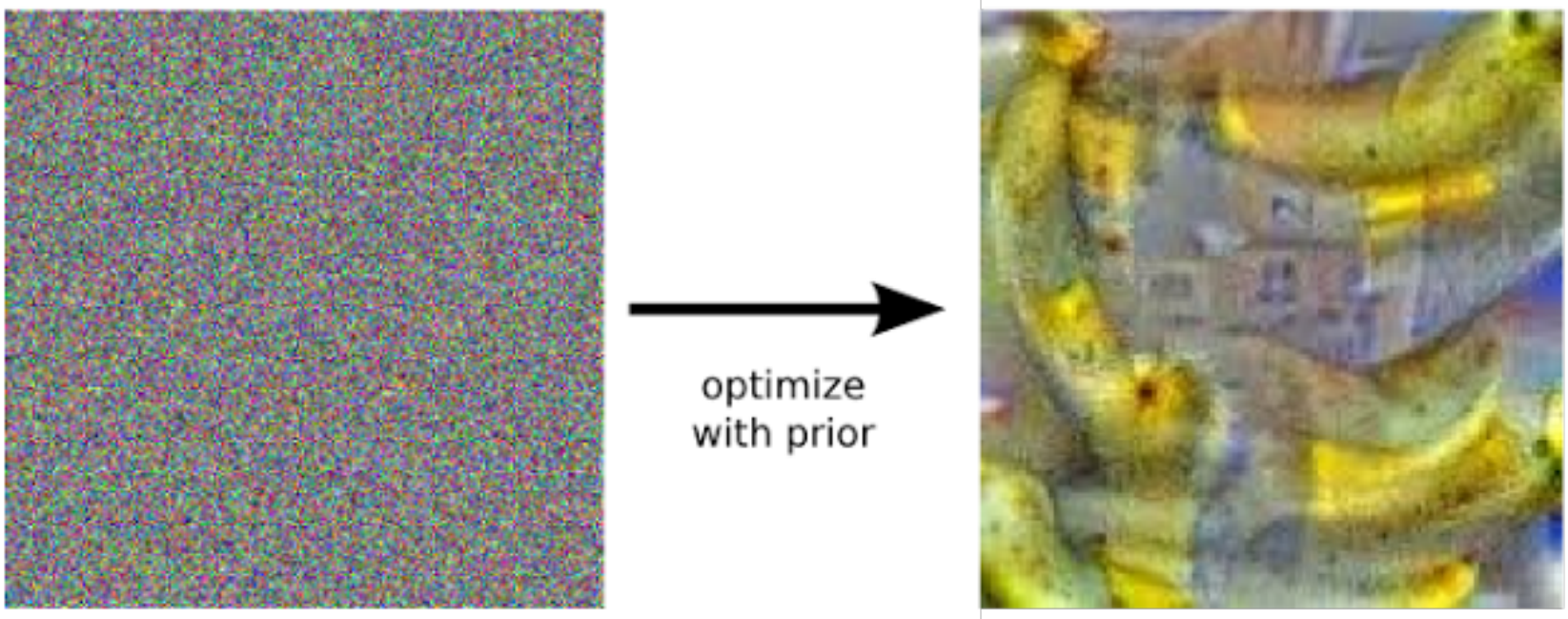


[1] Taken from `http://googleresearch.blogspot.ca/2015/06/-inceptionism-going-deeper-into-neural.html`

Generating Dreams

Idea: Pick a layer and use gradient ascent to maximize L^2 norm of activations. Some tricks:

- Random offset
- Normalize gradient
- Multiple scales



Generating Dreams

Generating Dreams

Apply algorithm iteratively to its own outputs + zoom after each iteration

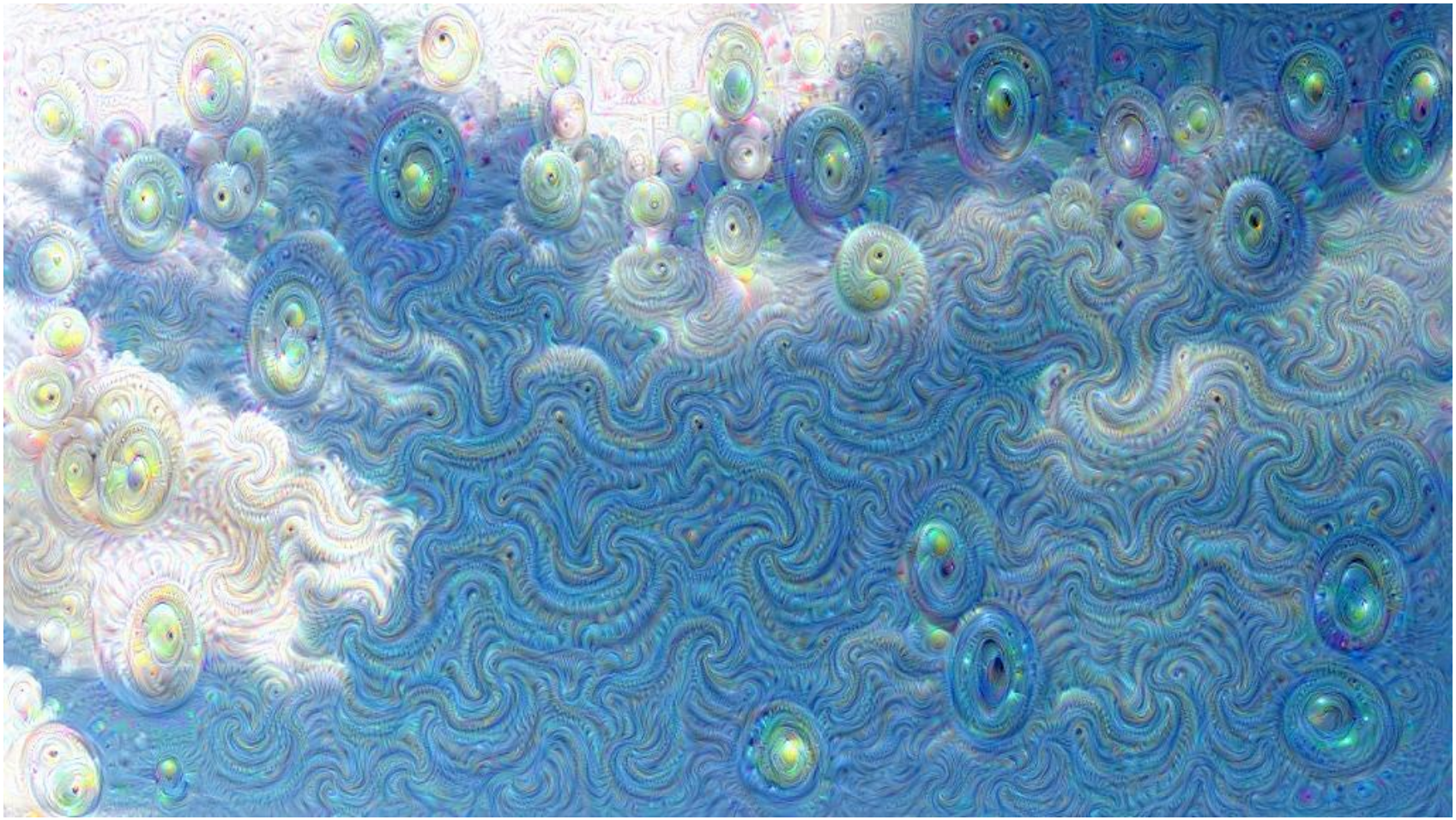
Generating Dreams

Apply algorithm iteratively to its own outputs + zoom after each iteration

Guided Dreams

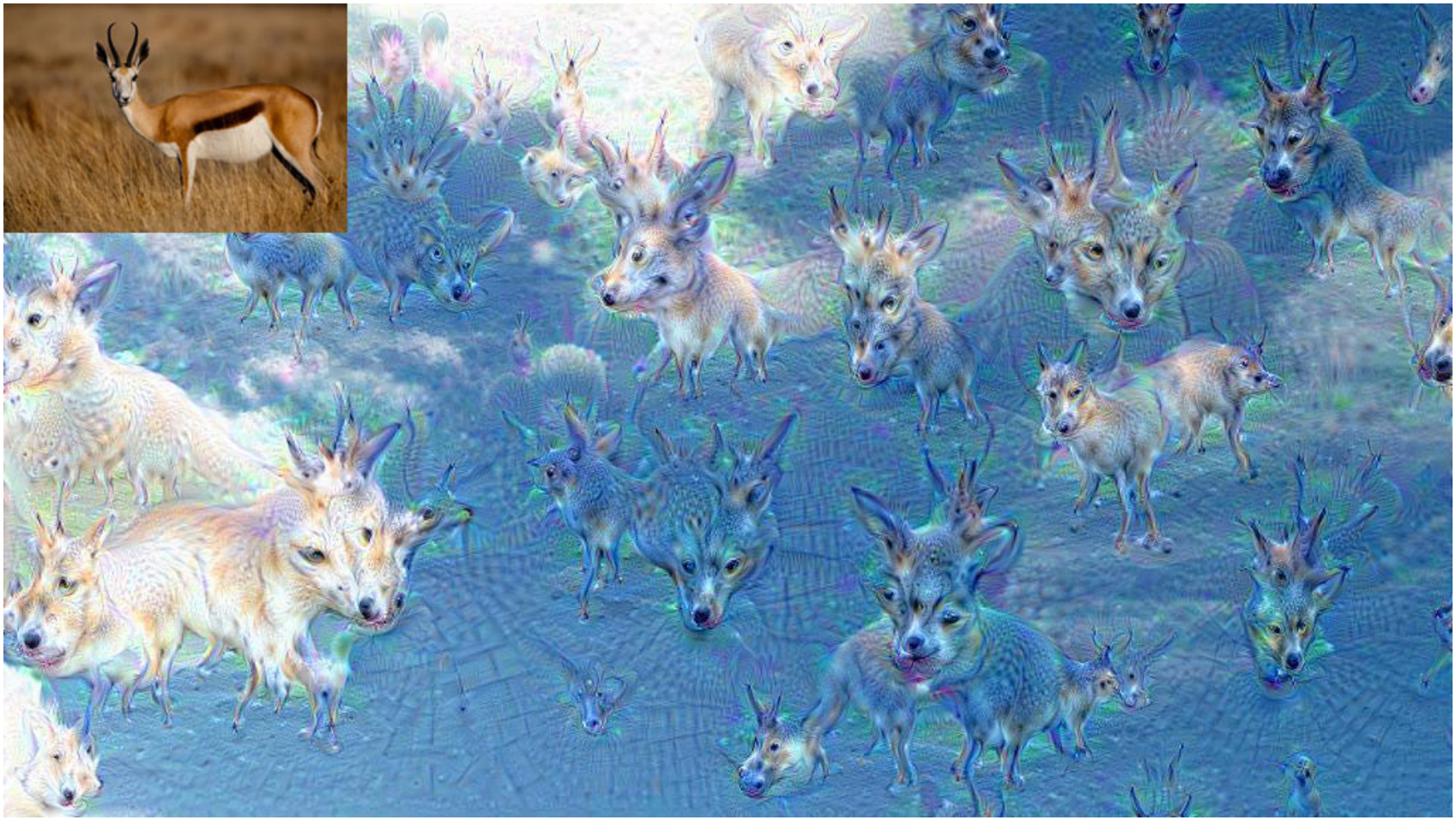
[1] http://googleresearch.blogspot.ca/2015/06/inceptionism-going-deeper-into-neural.html

Guided Dreams



[1] http://googleresearch.blogspot.ca/2015/06/inceptionism-going-deeper-into-neural.html

Guided Dreams



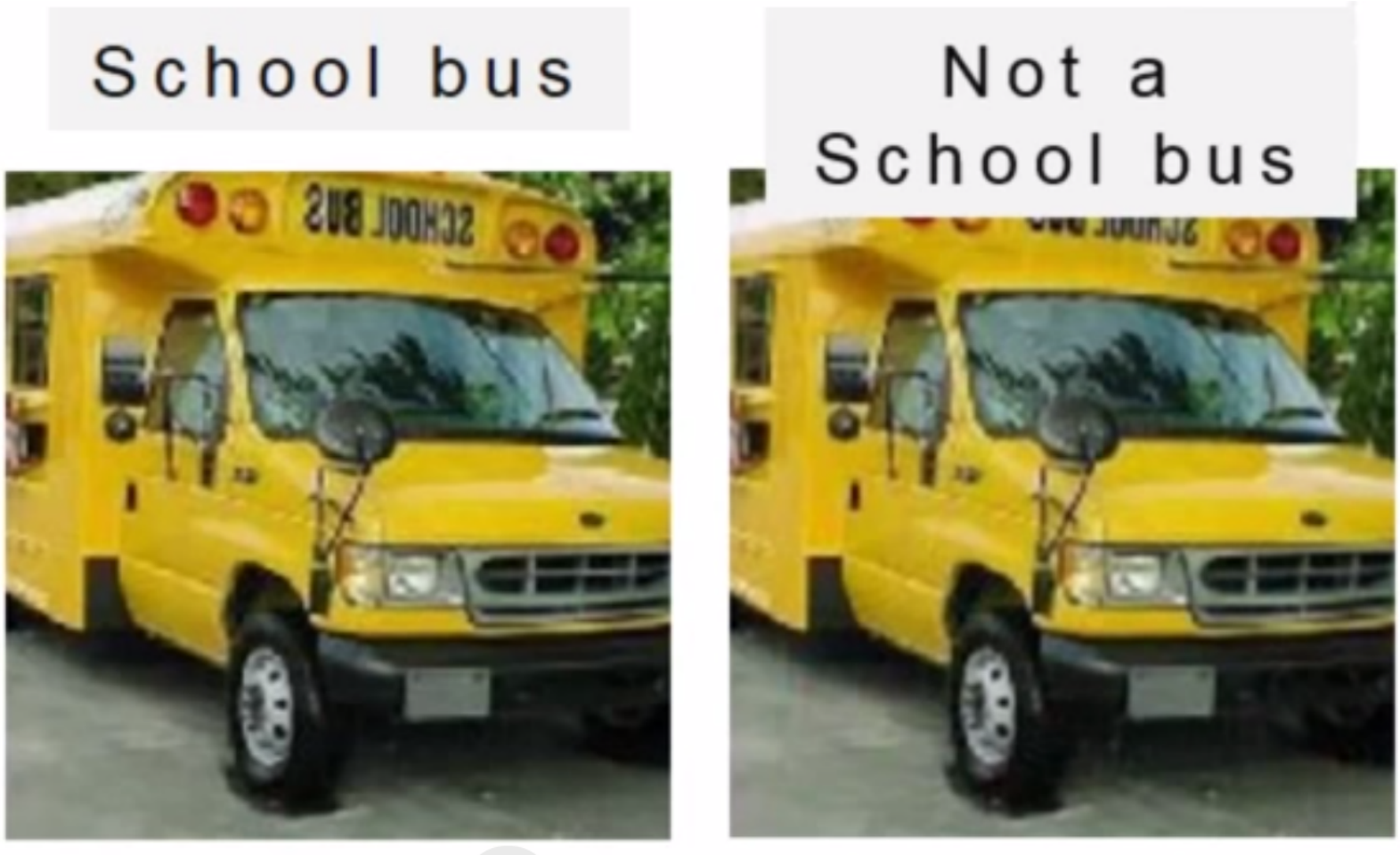
[1] http://googleresearch.blogspot.ca/2015/06/inceptionism-going-deeper-into-neural.html

Why is this useful?

- Dreams help visualize the model's representation of the world, exposing bias in how the model was trained.
- Predictive accuracy is not enough. Having a good representation is important.

Adversarial Images

Perturbing the input image in an imperceptible way can drastically change a DNN's classification



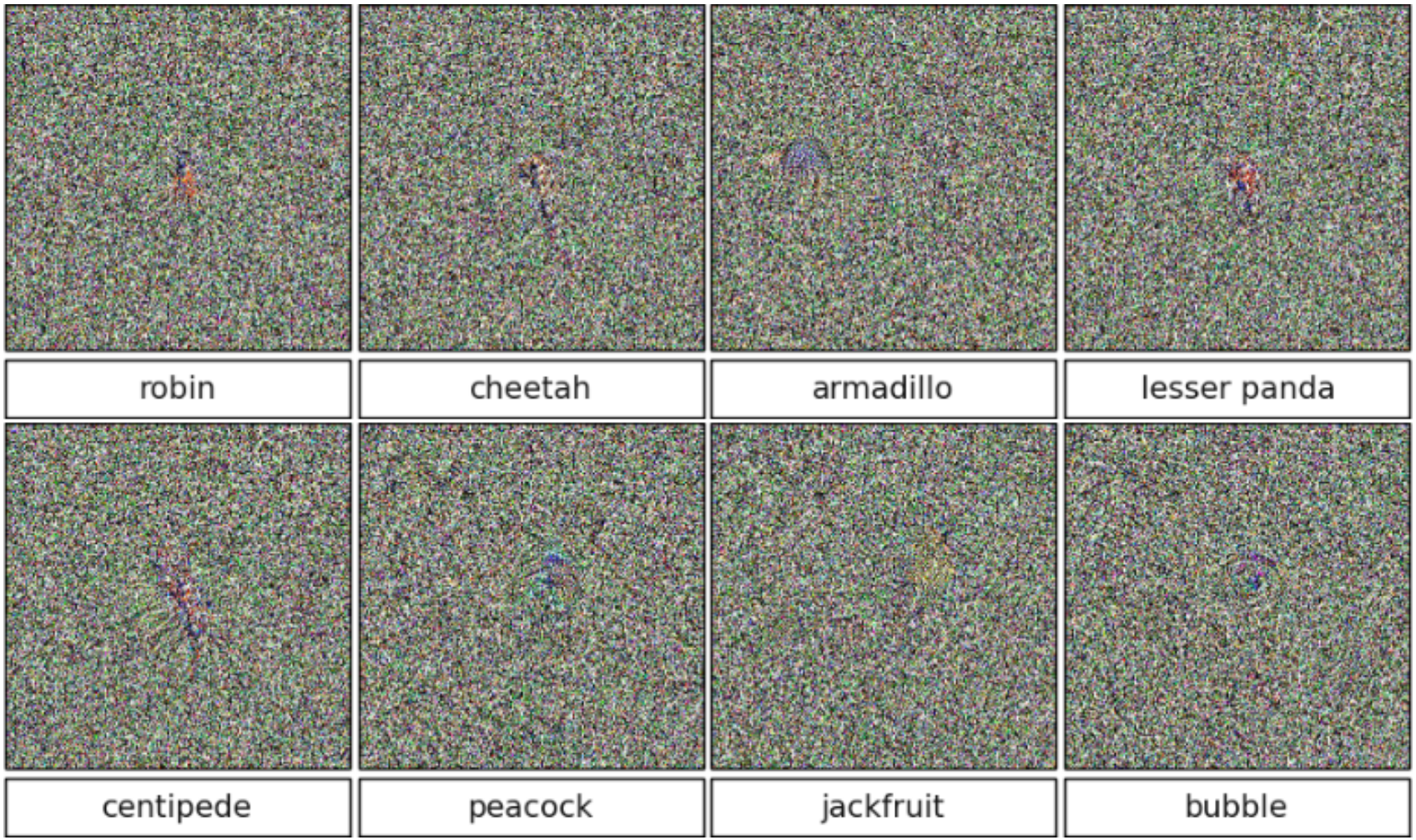
[1] Szegedy, Christian, et al. "Intriguing properties of neural networks."

Goal: invariant recognition

2

[Thomas Serre 2012]

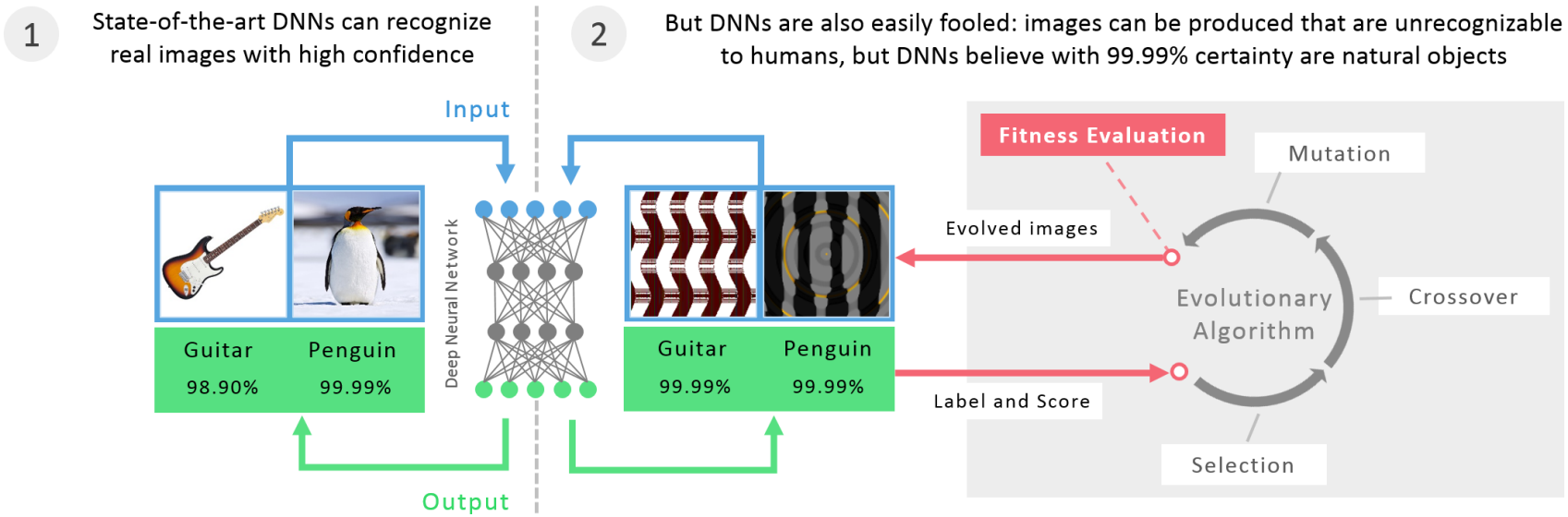
Fooling DNNs with unrecognizable images



[1] http://googleresearch.blogspot.ca/2015/06/inceptionism-going-deeper-into-neural.html

Fooling DNNs with unrecognizable images

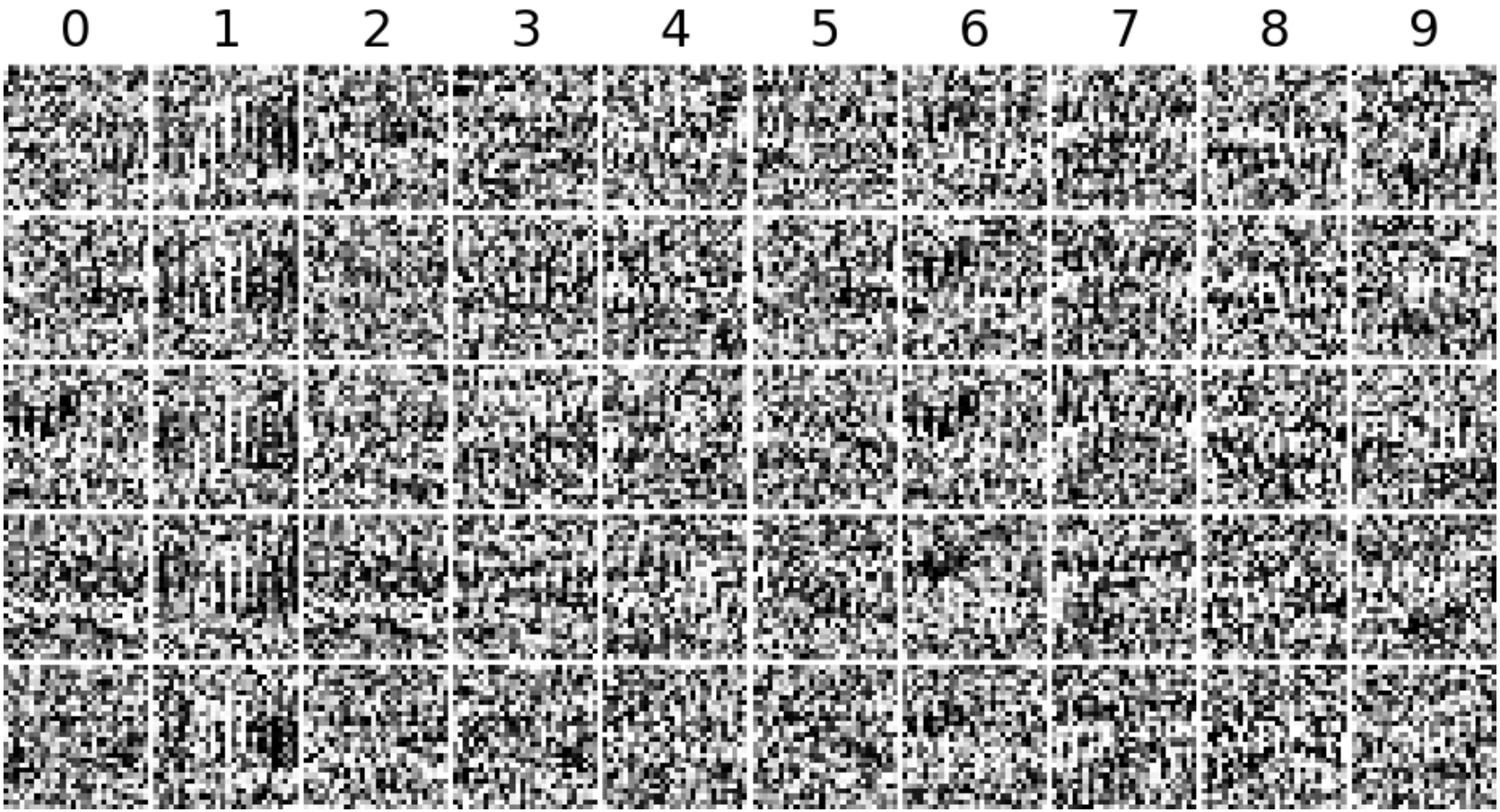
Produce 'adversarial images' using evolutionary algorithms.



[1] Nguyen et al. 2015

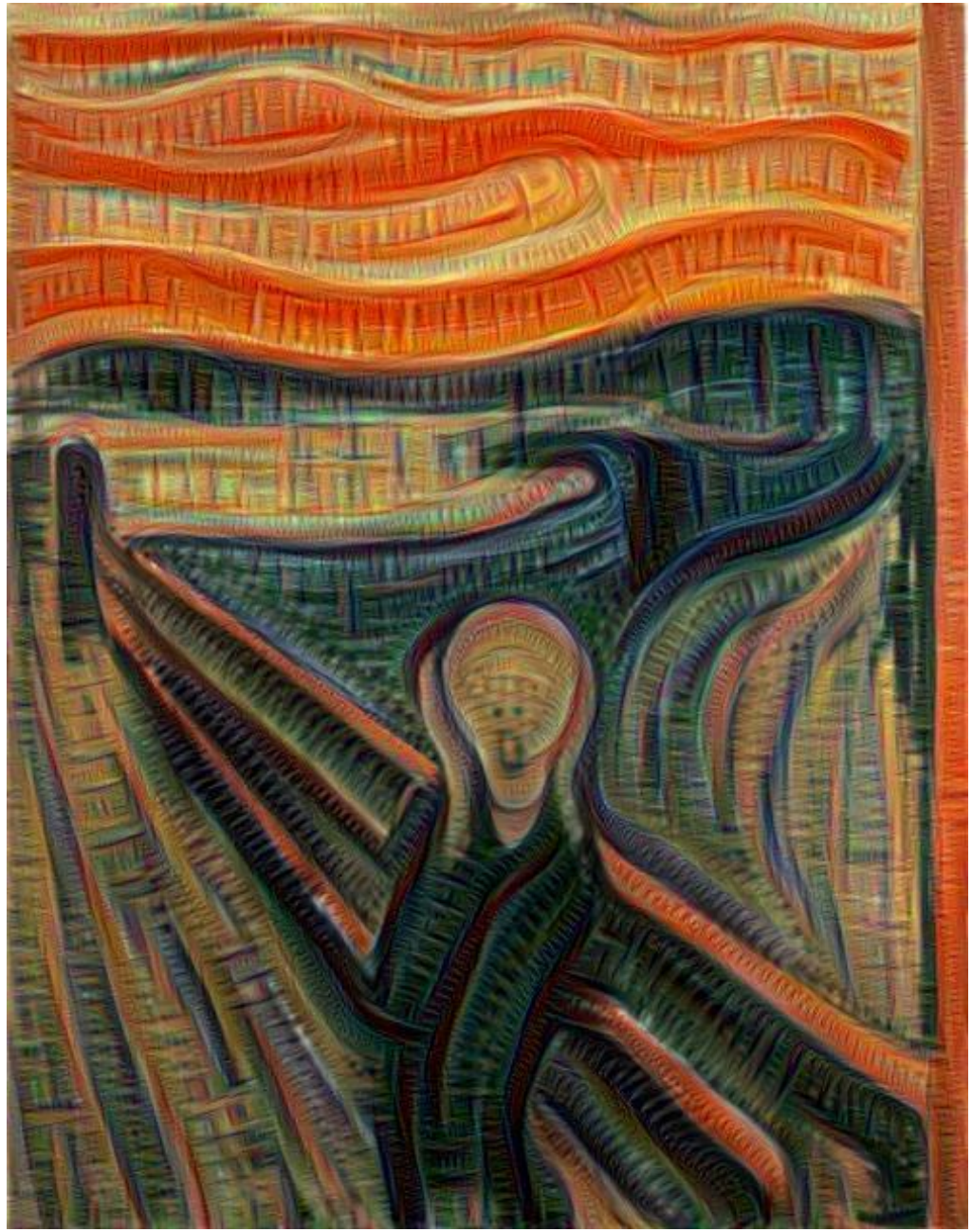
Fooling Neural Networks

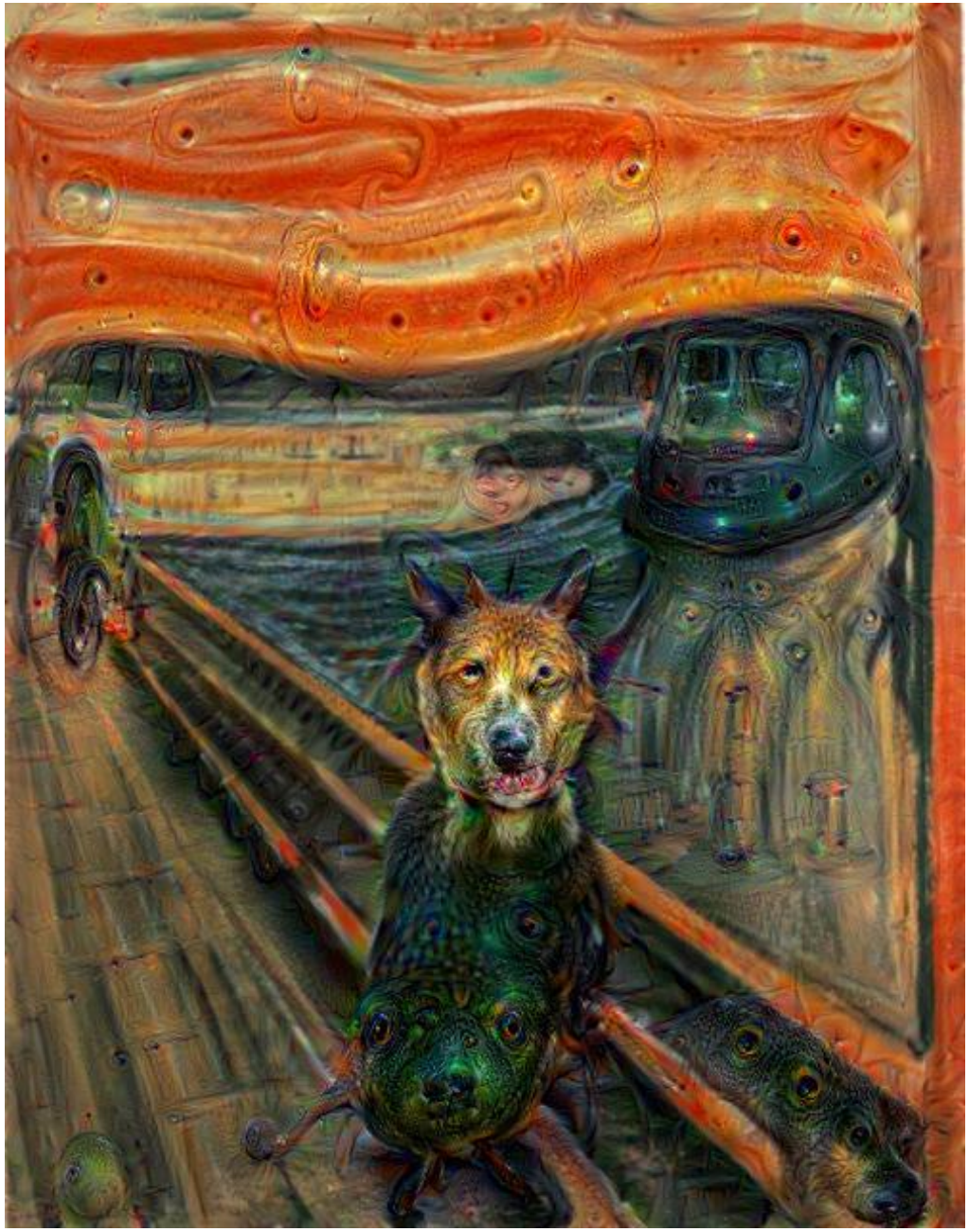
LeNet trained on MNIST is 99.99% certain that these images are digits:

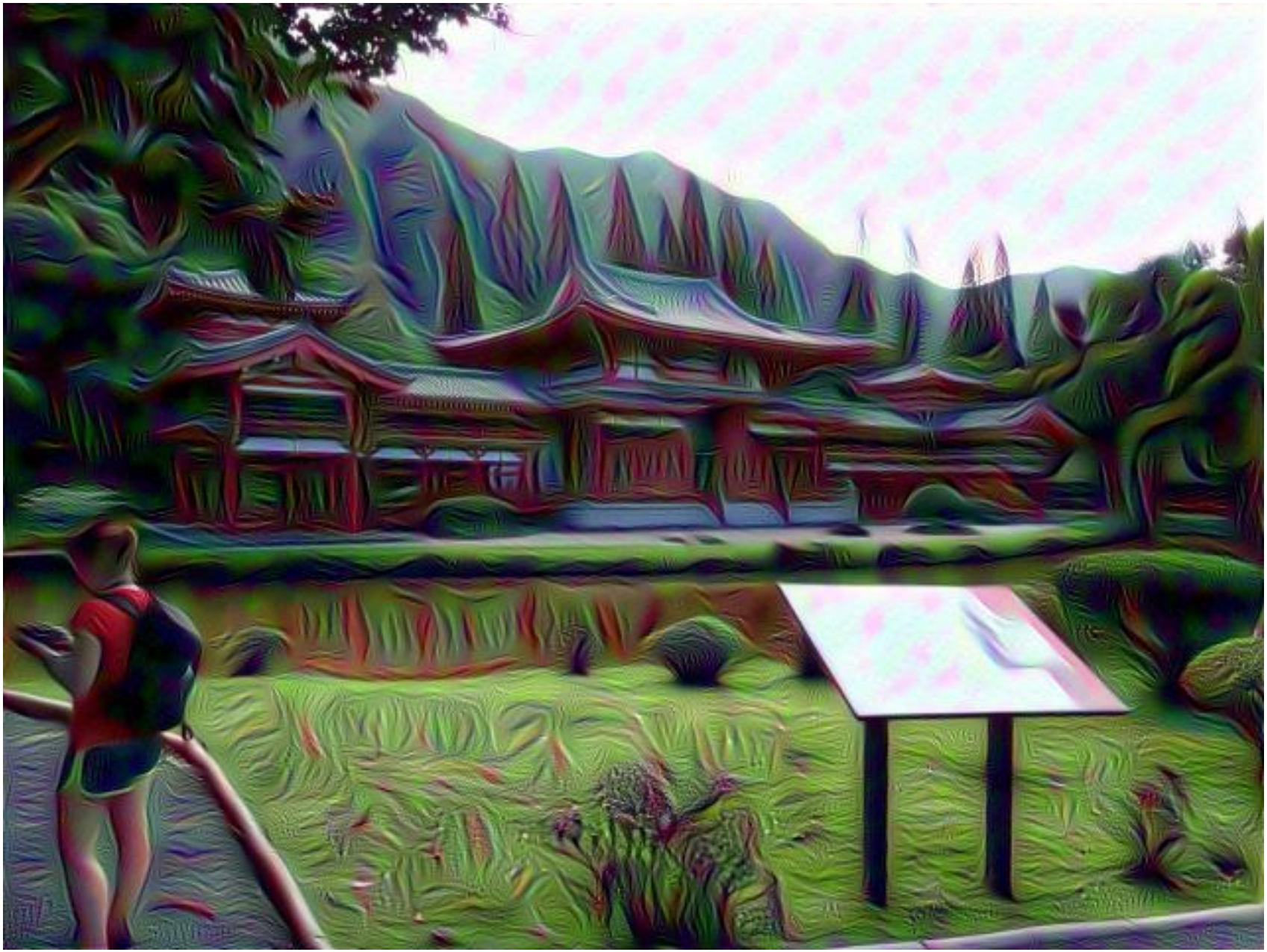


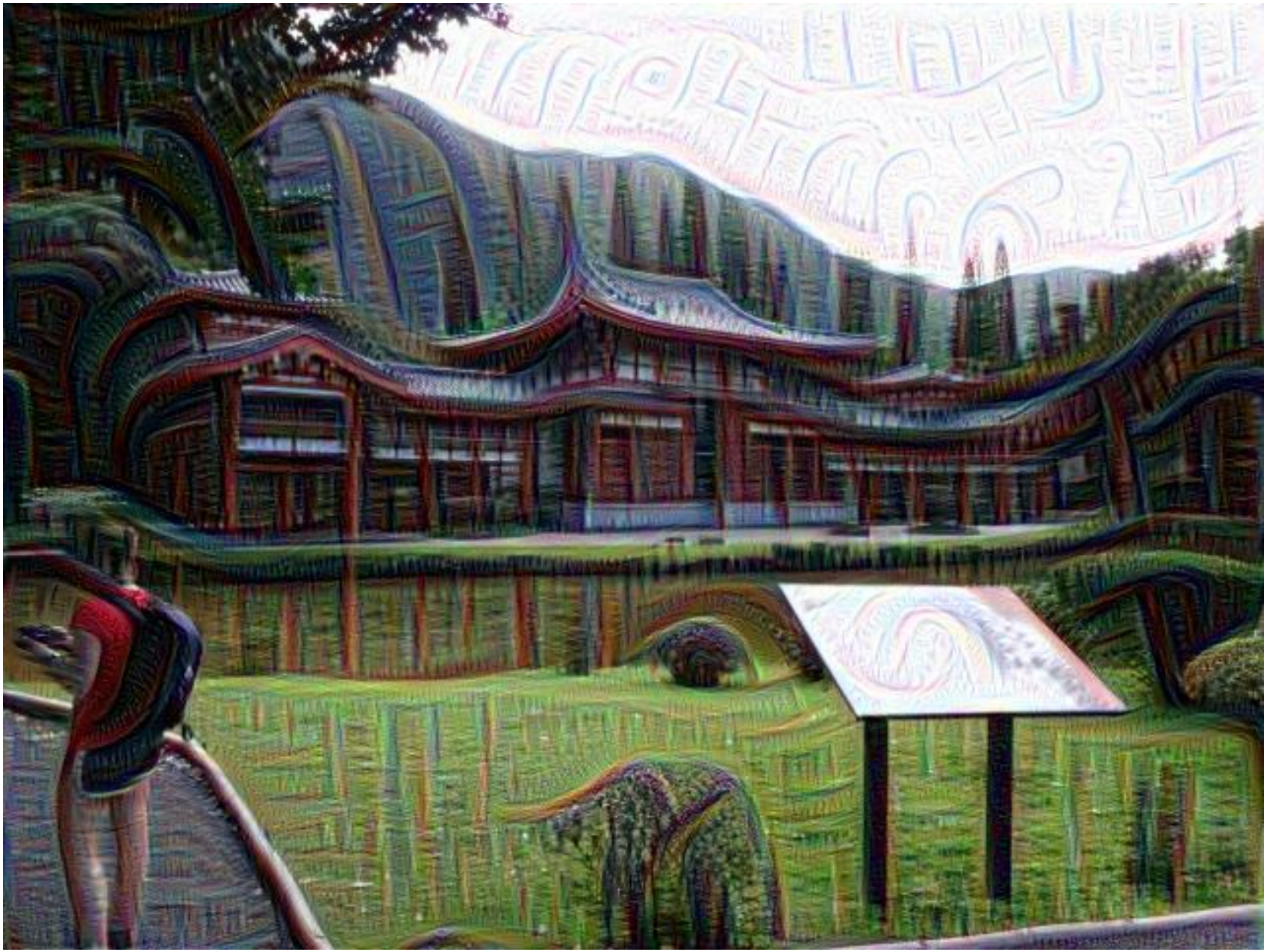
Why is this important?

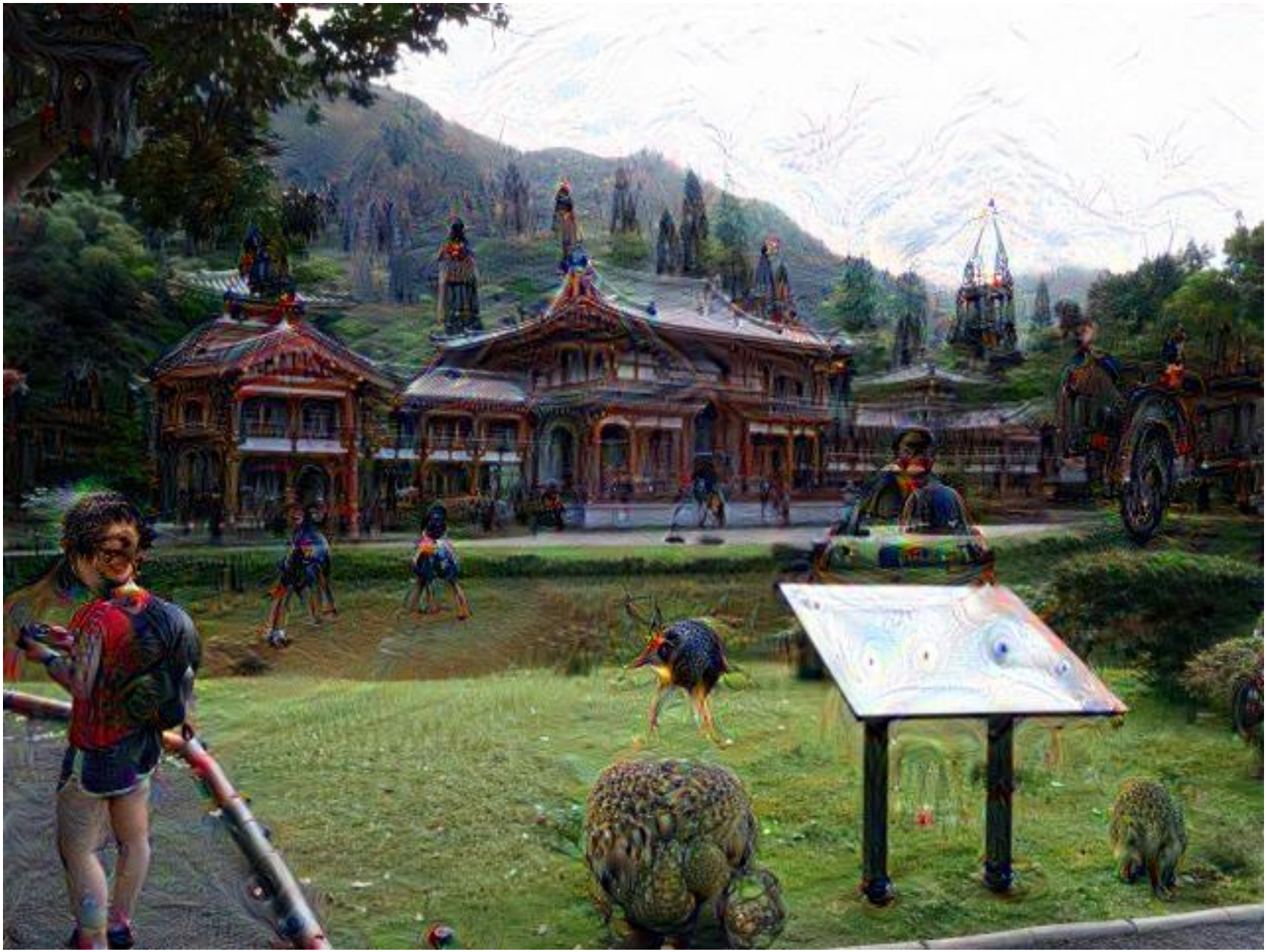
- Illuminates the model's representation of the world, exposing biases in the training set.
- Practical applications: False positives can be exploited.
  - recognition in security camera
  - image-based search engines

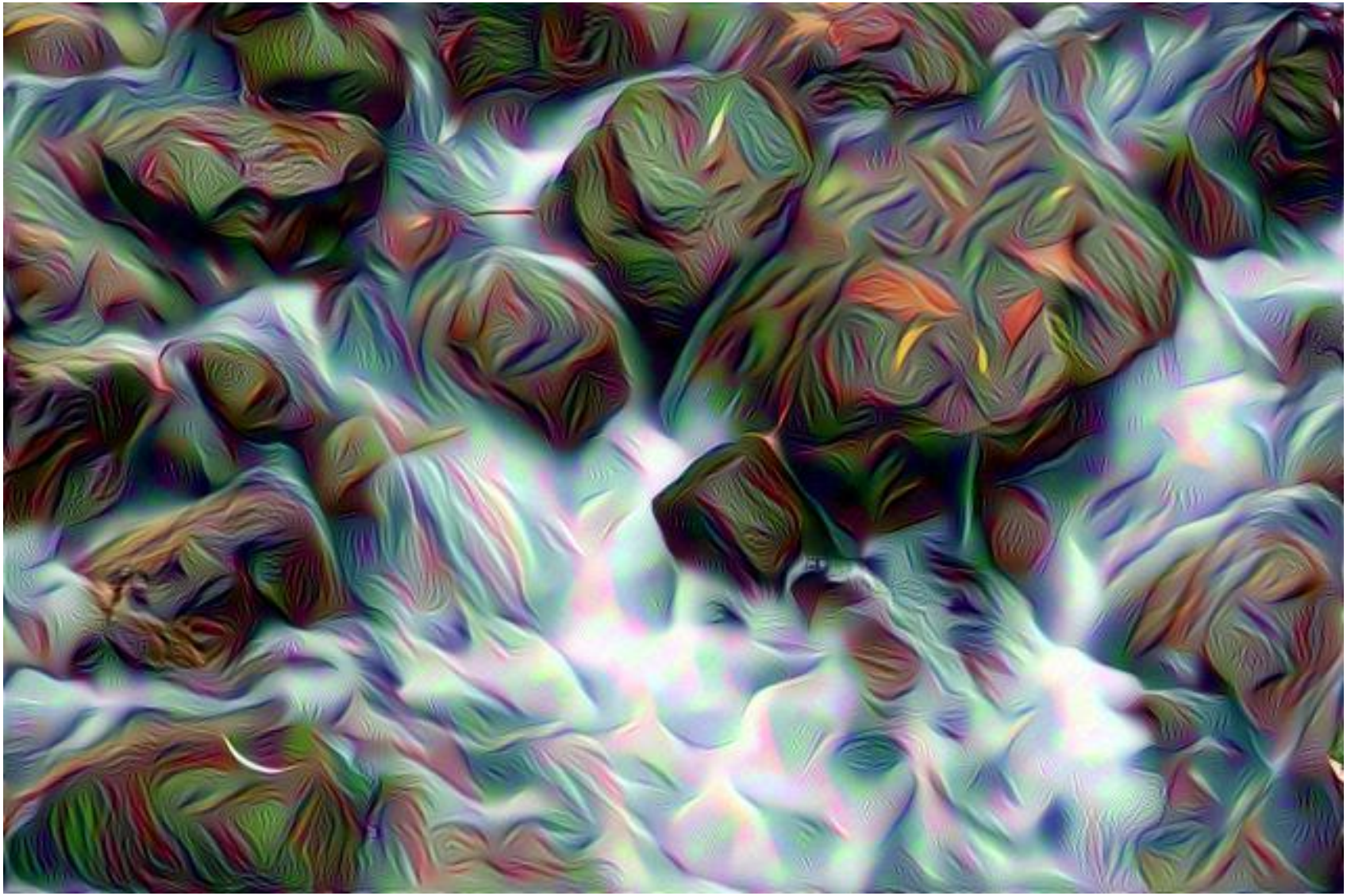


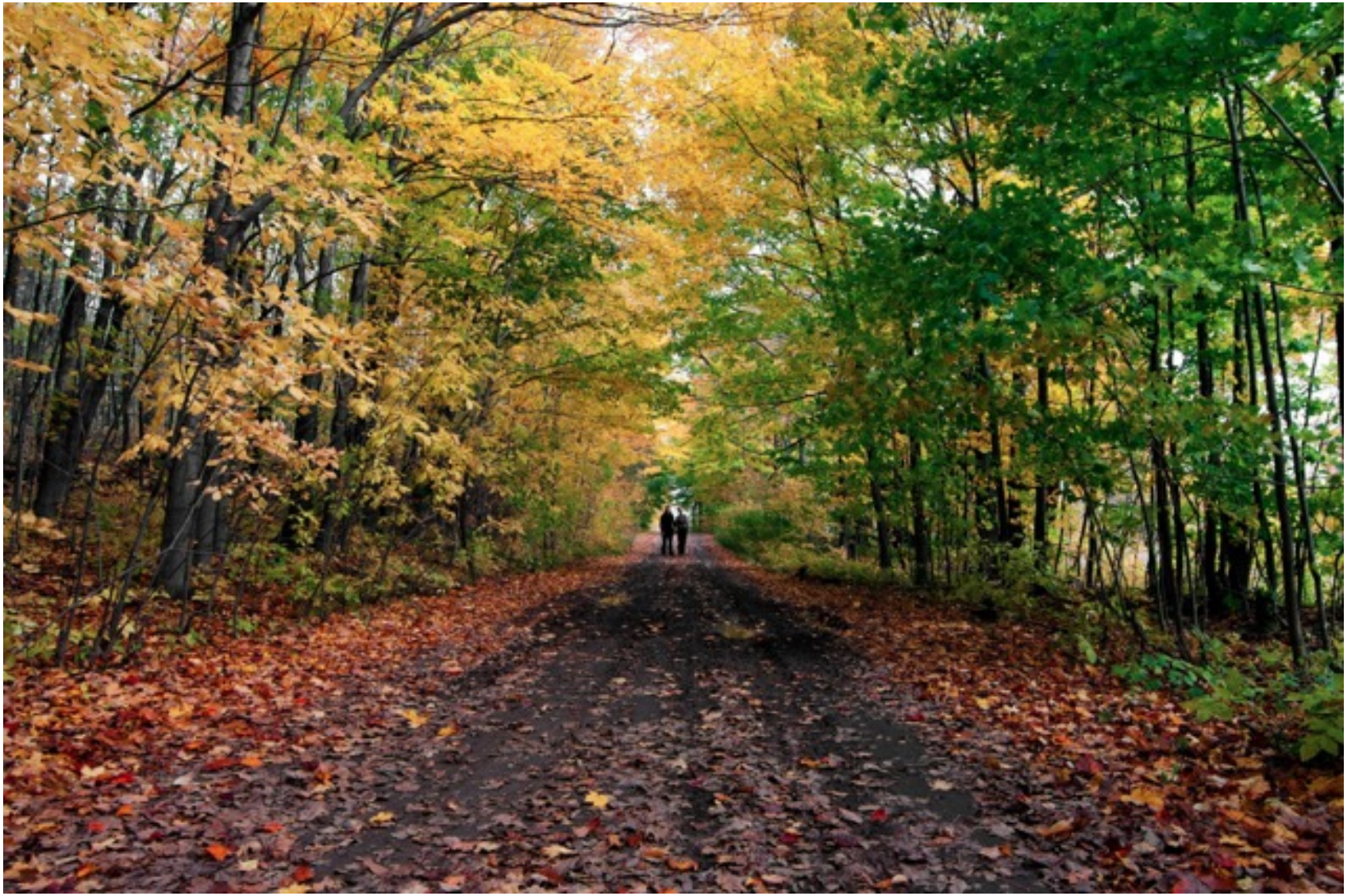


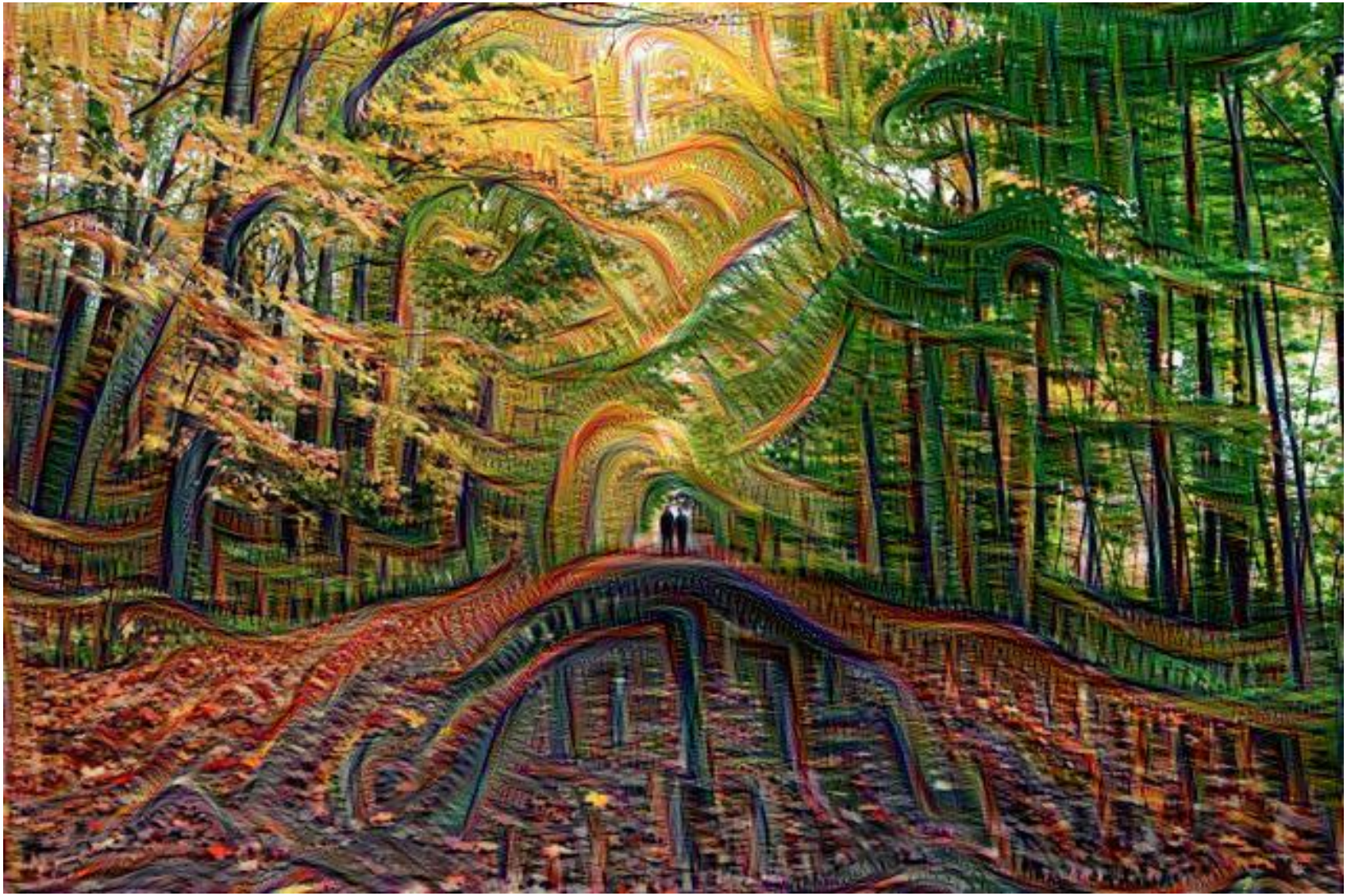


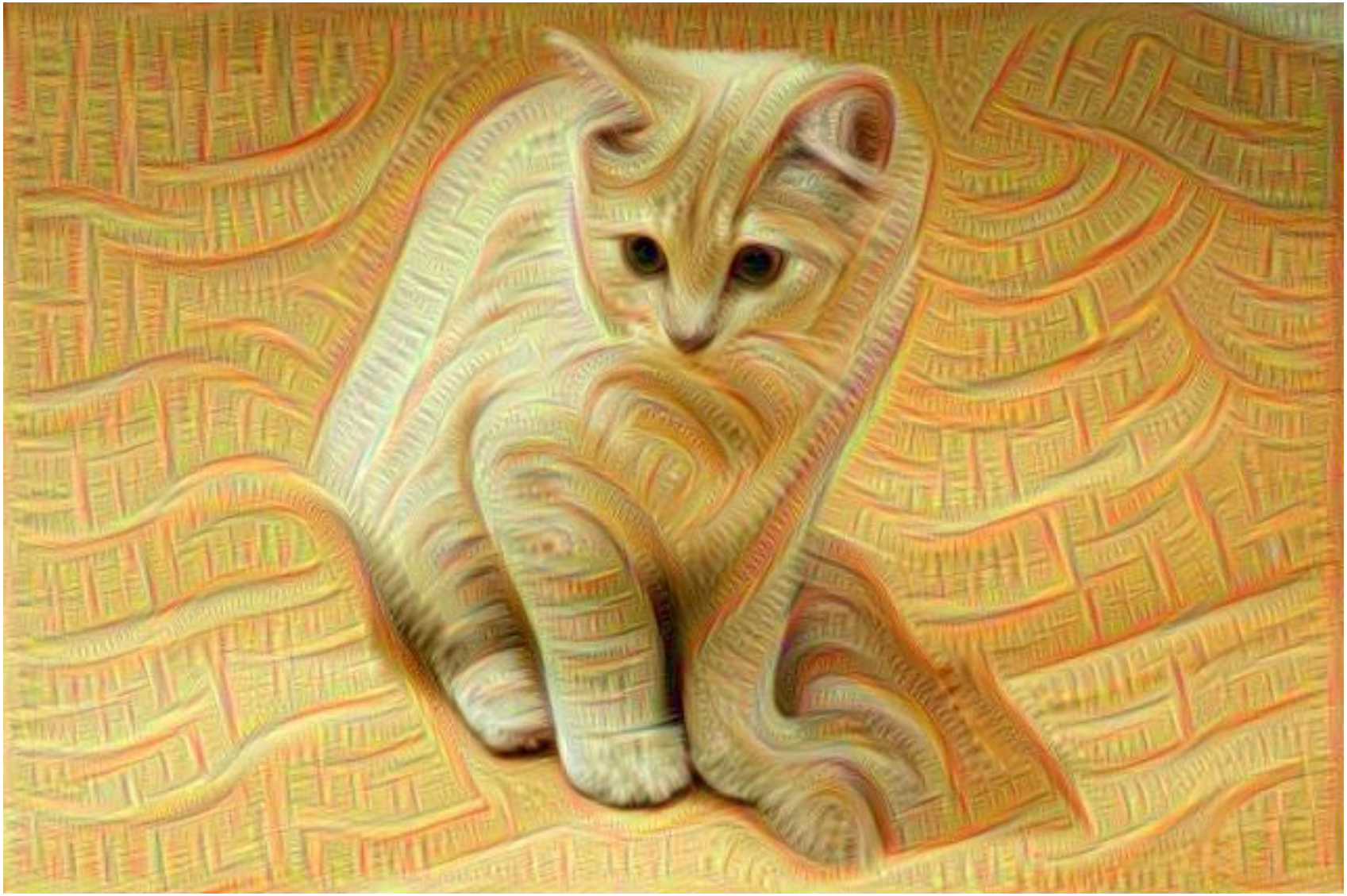


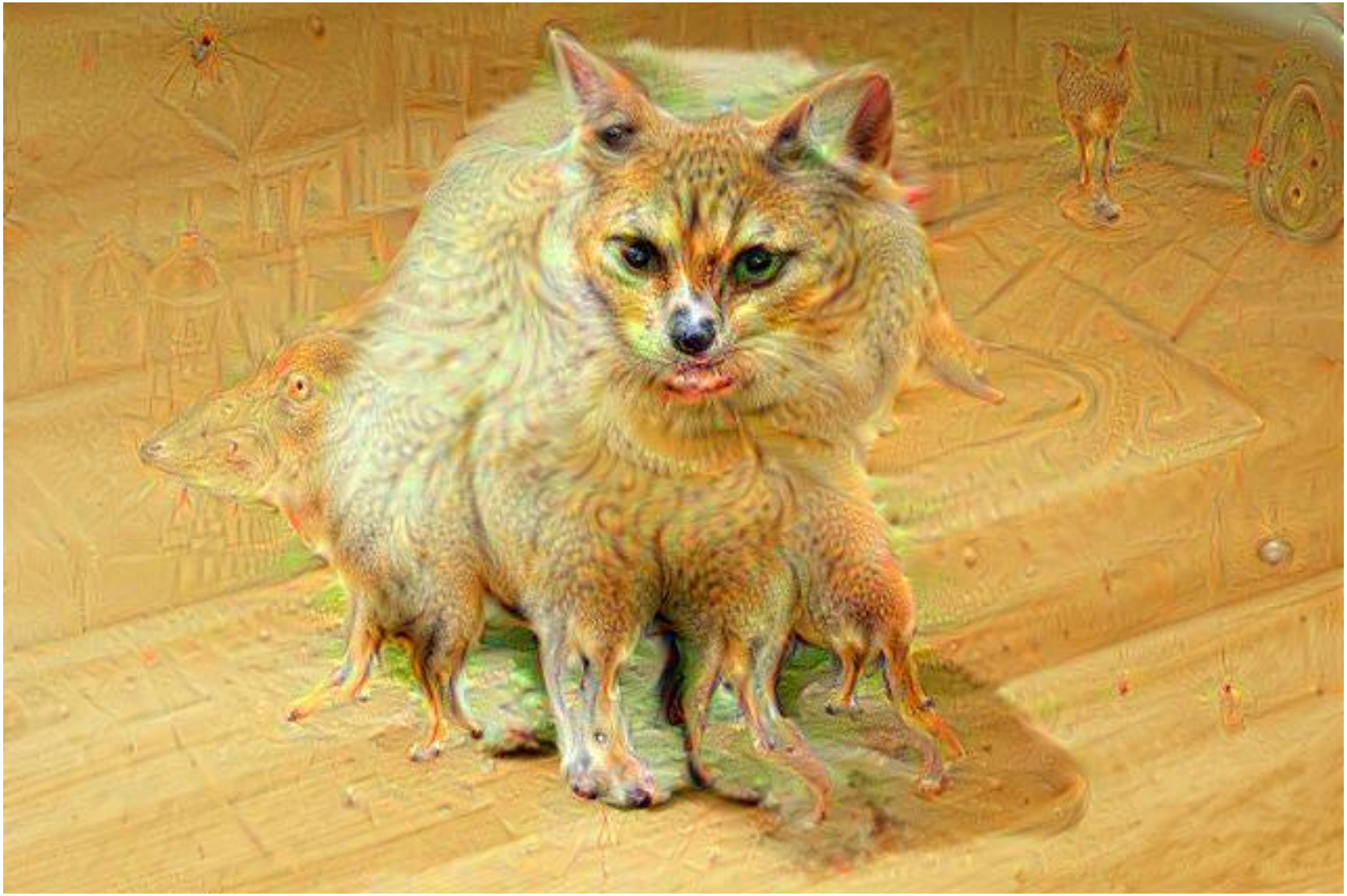












References

1. http://googleresearch.blogspot.ca/2015/06/inceptionism-going-deeper-into-neural.html
2. https://github.com/google/deepdream
3. https://github.com/VISIONAI/clouddream
4. http://www.pyimagesearch.com/2015/07/13/generating-art-with-guided-deep-dreaming
5. Nguyen A, Yosinski J, Clune J. Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. In Computer Vision and Pattern Recognition (CVPR '15), IEEE, 2015.
6. Szegedy, Christian, et al. "Intriguing properties of neural networks." arXiv preprint arXiv:1312.6199 (2013).