# Convolutional Neural Networks (cont.)

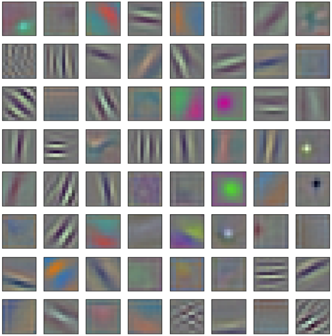
Nan Ye

School of Mathematics and Physics
The University of Queensland

# Visualizing CNNs

**Visualize filters**

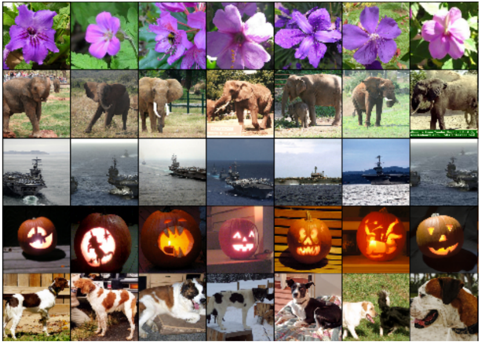
- We can visualize filters in a convolutional layer by viewing the weight array of each filter as an image.
- The first layer of AlexNet (dimension 64x3x11x11)



- Visualizing filters at deeper layers can be done, but not very interesting (the kernel sizes are often very small).

## Visualize learned features: nearest neighbors

- We can visualize the learned features (the last hidden layer) by retrieving training images which are nearest to the test image in the feature space.
- Nearest neighbors using AlexNet features

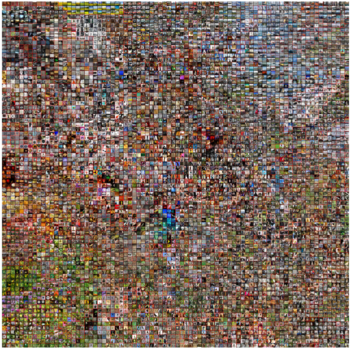


Test images in 1st column, others are nearest training images

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

## Visualize learned features: 2D embeddings

- We can also visualize the learned features by embedding the features of images in a plane.
- 2D embeddings of AlexNet features

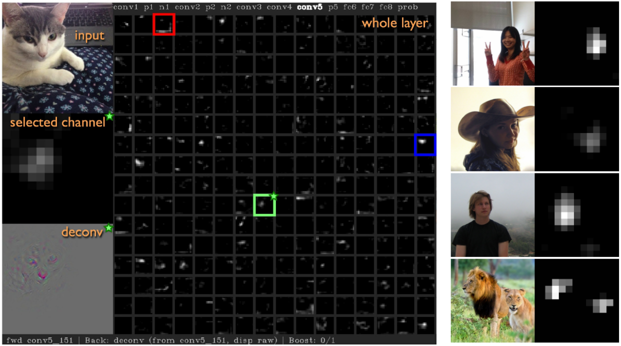


6400 images placed at their embedded locations for their features

https://cs.stanford.edu/people/karpathy/cnnembed/

## Visualize activations

- We can visualize how a convolutional layer responds to an input image by displaying its feature maps as images.
- Activations of a 256x13x13 convolutional layer


conv1 p1 n1 conv2 p2 n2 conv3 conv4 conv5 p5 fc6 fc7 fc8 prob
input
whole layer
selected channel
deconv
fwd conv5_151 | Back: deconv (from conv5_151, disp raw) | Boost: 0/1

The green channel is able to detect animal or human faces

Yosinski, Clune, Nguyen, Fuchs, and Lipson, Understanding neural networks through deep visualization, 2015

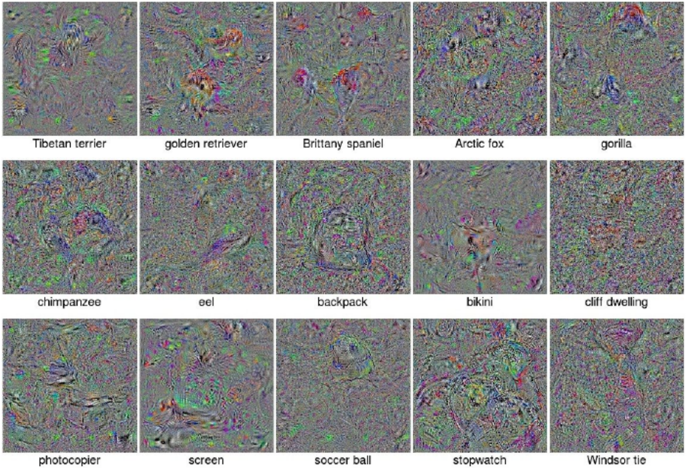
## Visualize maximally activating natural image

- For a neuron $a$, search for an image $x$ that maximizes the regularized activation

  $$a(x) - R(x),$$

  where $R(x)$ is a regularizer that encourages the image to be natural.
- This can be solved by gradient ascent.
- A more general version is to apply a transformation after one or a few gradient ascent steps.

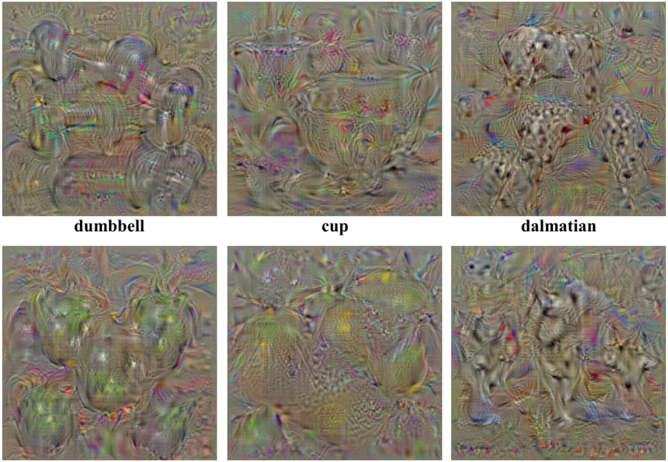
- Images constructed without regularization



- Deep neural networks are easily fooled!

Nguyen, Yosinski, and Clune, Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images, 2015

- Images constructed using $R(x) = \|x\|_2$ for output neurons



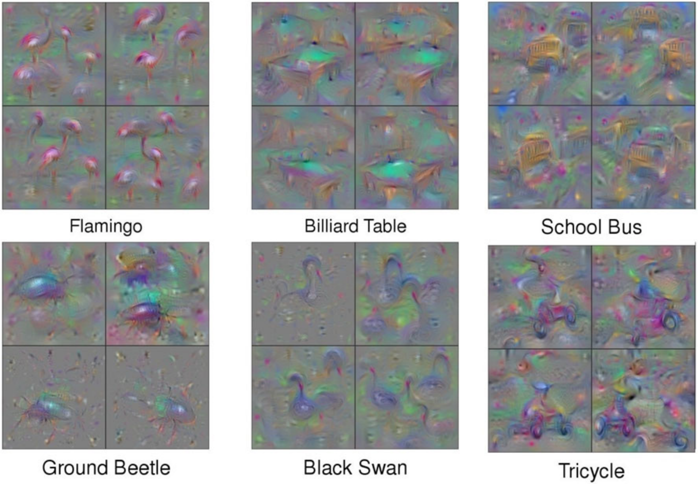
**dumbbell** **cup** **dalmatian**

**bell pepper** **lemon** **husky**

Simonyan, Vedaldi, and Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, 2013

- Better result can be achieved by using a combination of various regularizers (good image priors)
    - L2 regularizer
    - Gaussian blur
    - Clip pixels with small values to 0
    - Clip pixels with small gradients to 0

- Four images are generated for each output neuron using four different combinations of priors



Flamingo

Billiard Table

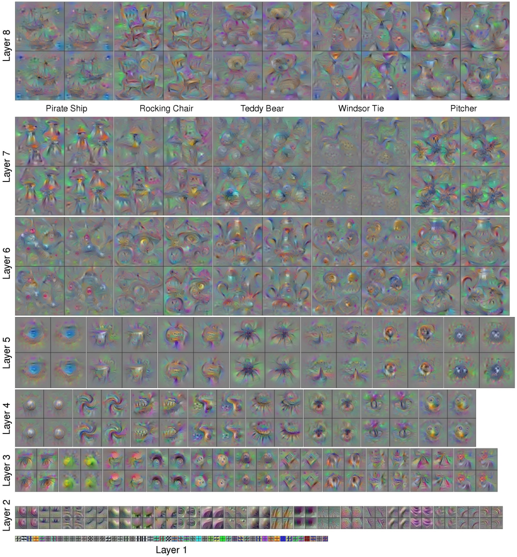
School Bus

Ground Beetle

Black Swan

Tricycle

Yosinski, Clune, Nguyen, Fuchs, and Lipson, Understanding neural networks through deep visualization, 2015



# Visualization of all layers

Yosinski, Clune, Nguyen, Fuchs, and Lipson, Understanding neural networks through deep visualization, 2015

# What You Need to Know...

- Visualizing CNNs
  - Filters as images
  - Learned features: nearest neighbors, 2D embeddings
  - Activations: activating natural images, maximally activating synthetic (natural) images