THE UNIVERSITY OF BRITISH COLUMBIA

Topics in AI (CPSC 532S):
Multimodal Learning with Vision, Language and Sound

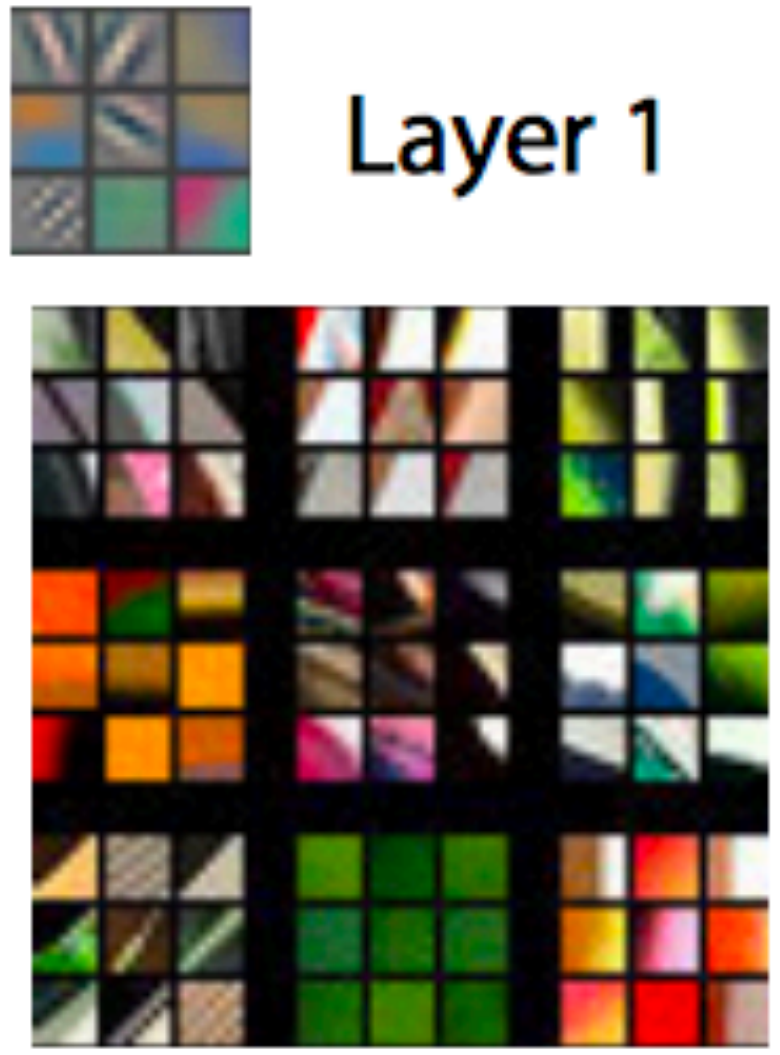
Lecture 7: Visualizing CNNs

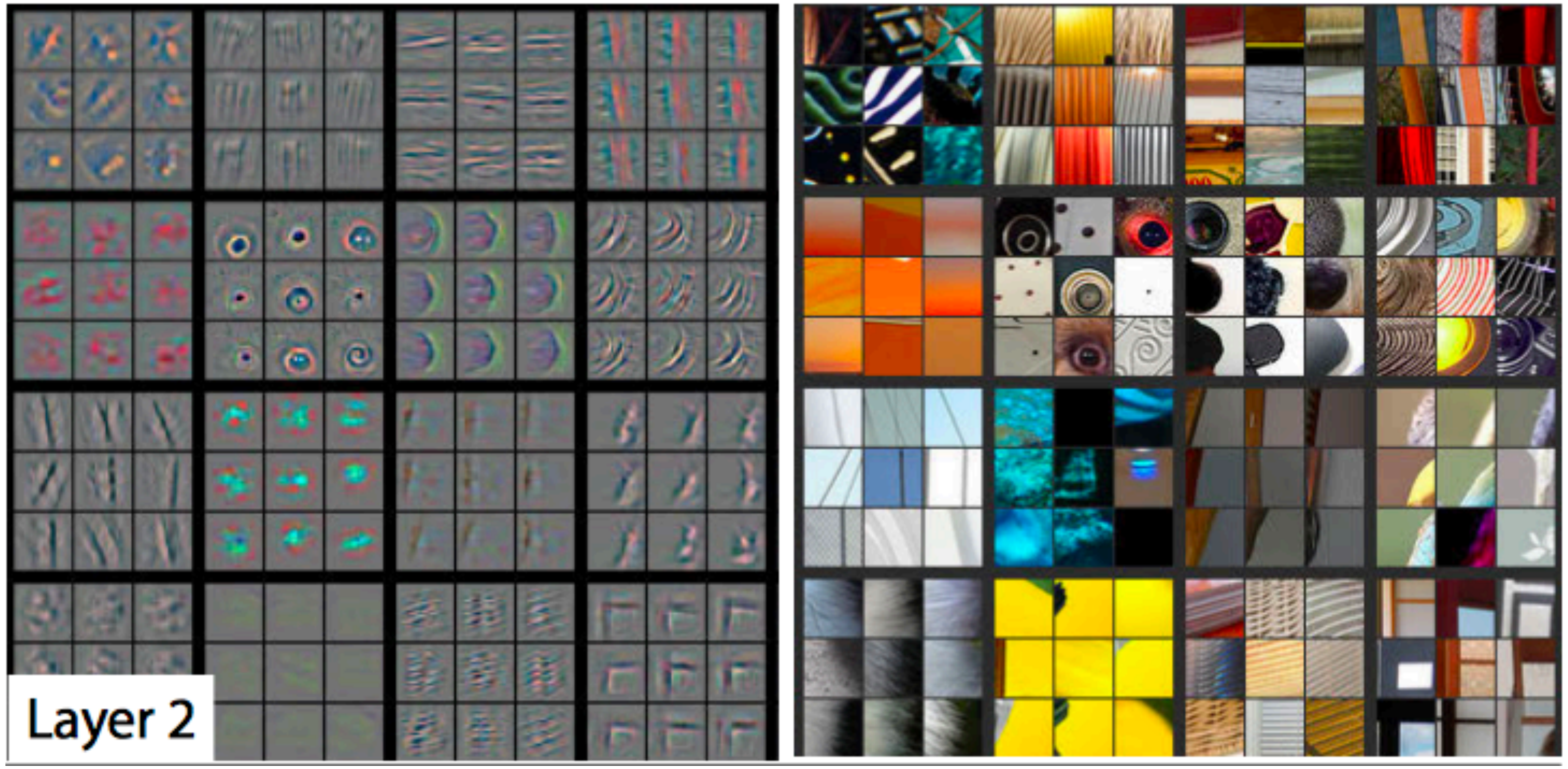
Logistics:

Assignment 2 was due **yesterday**

Assignment 3 will be posted soon … but …

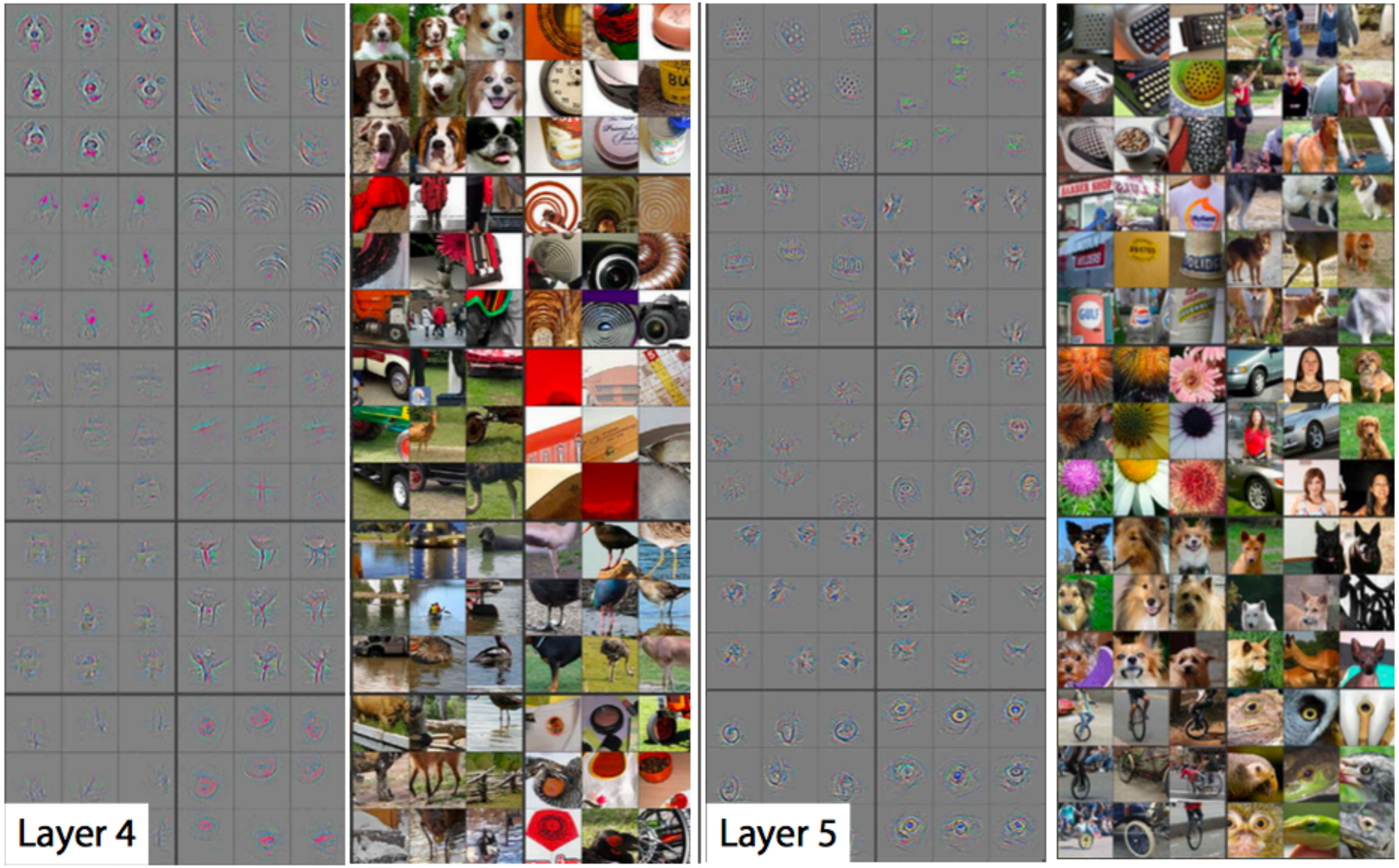
Recall …





[Zeiler and Fergus, 2013]

Recall …



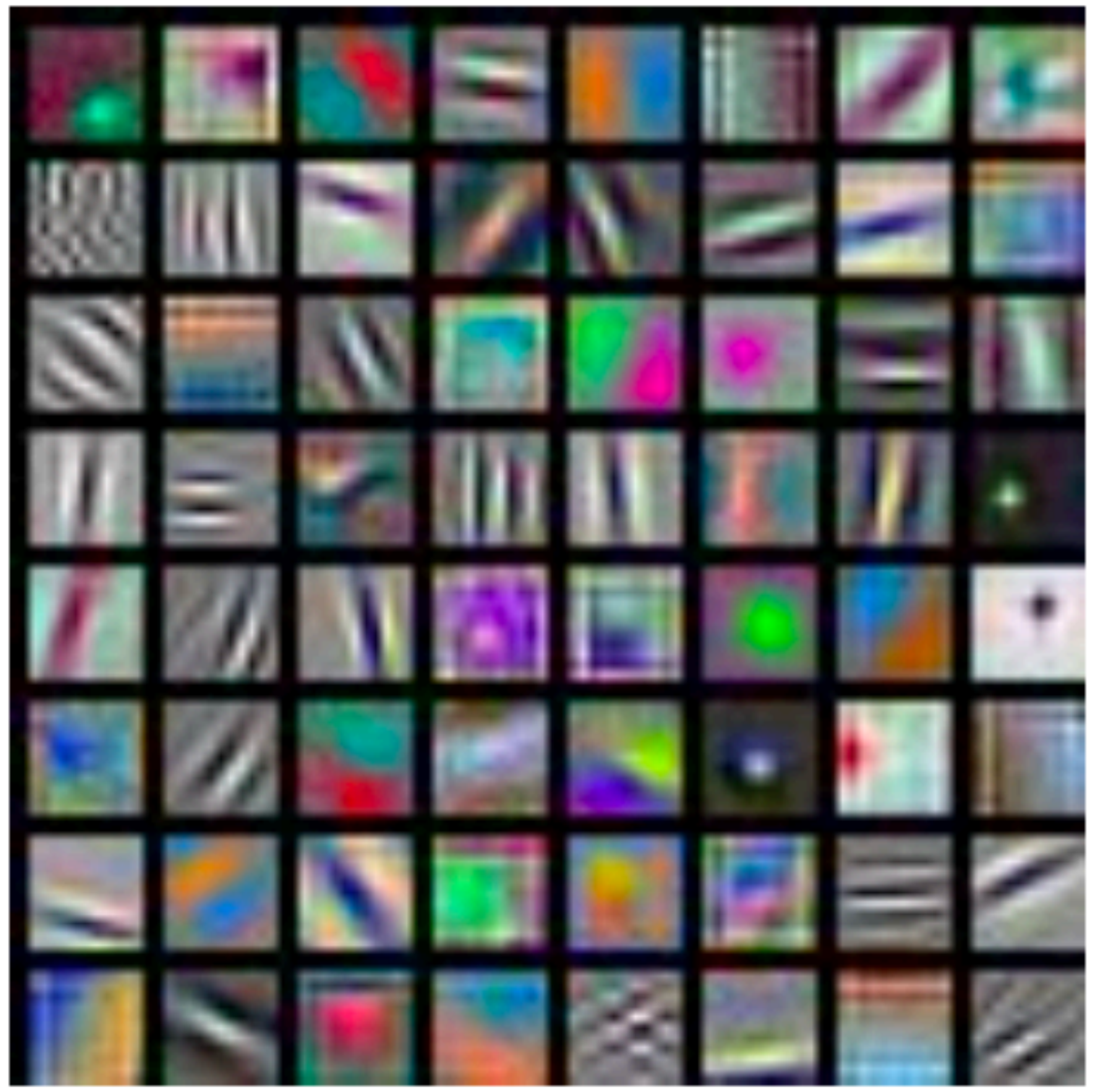
[Zeiler and Fergus, 2013]

Motivation …

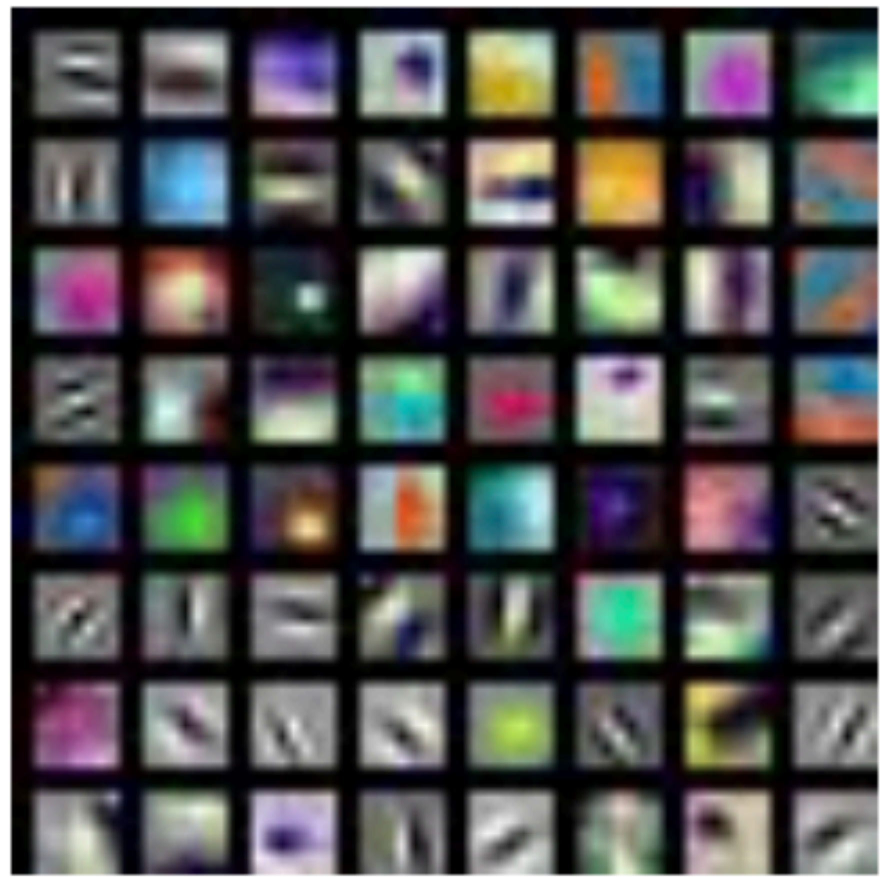
CNNs are big black boxes, lets get some intuition for how and why they work

First Layer Filters …

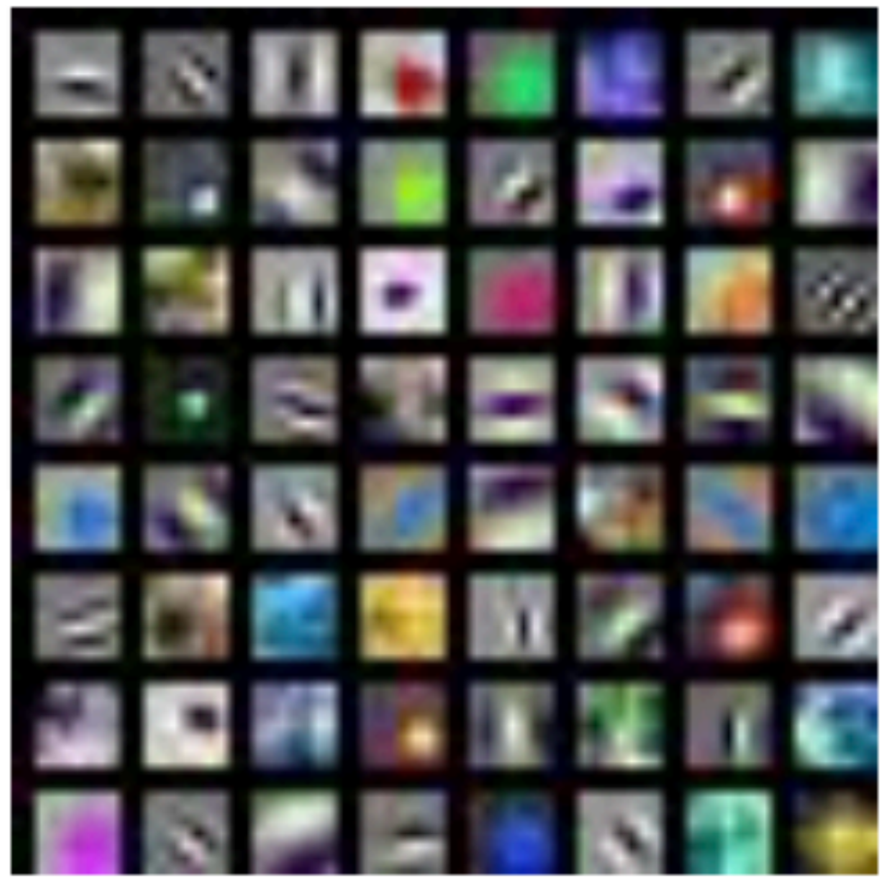
Directly **visualize filters** (only works for the first layer)



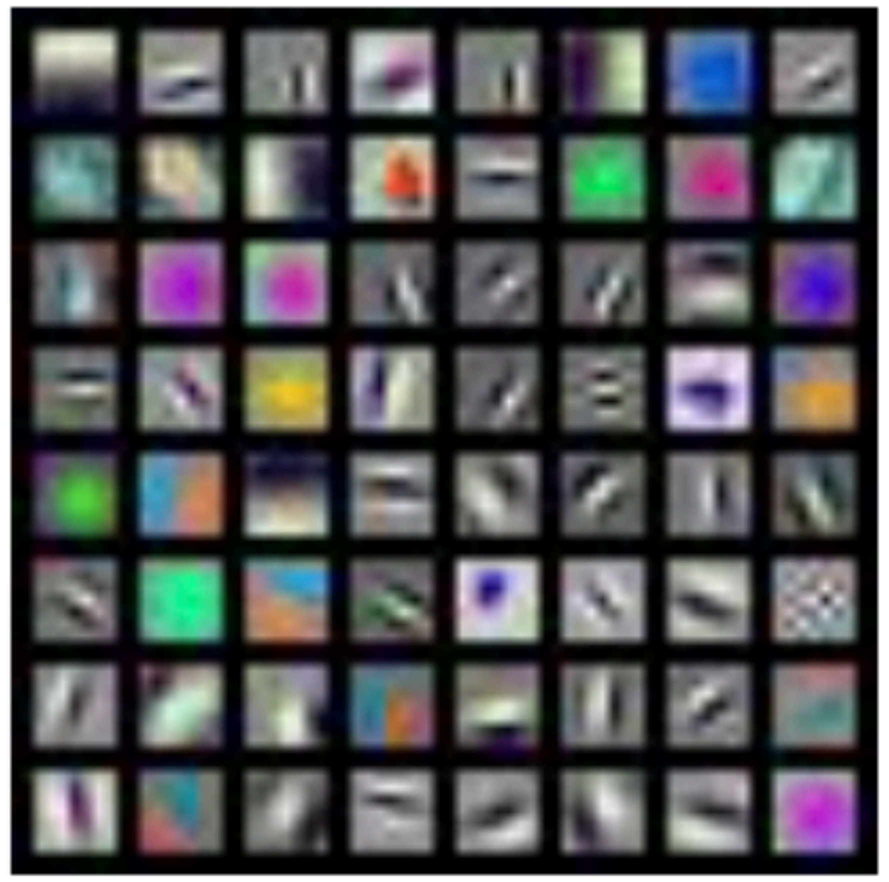
AlexNet:
64 x 3 x 11 x 11



ResNet-18:
64 x 3 x 7 x 7



ResNet-101:
64 x 3 x 7 x 7



DenseNet-121:
64 x 3 x 7 x 7

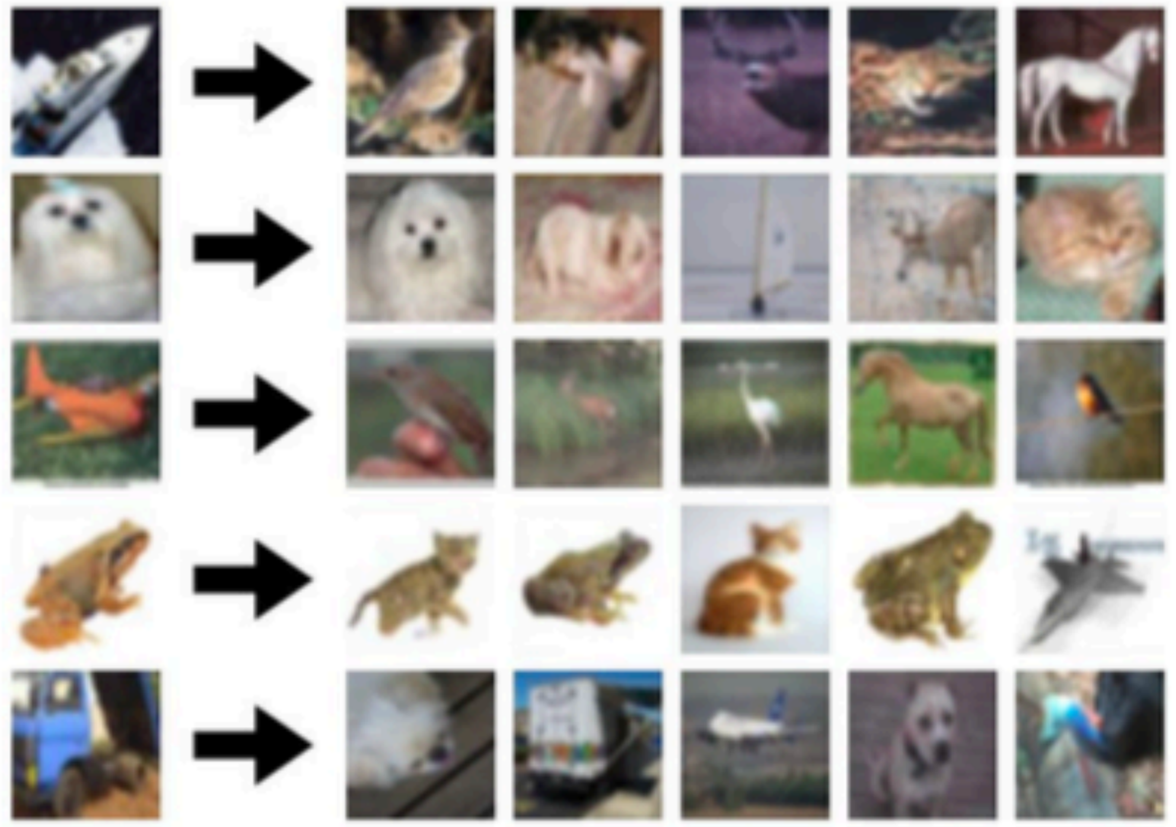
… surprisingly similar across variety of networks

… and nearly any dataset

* slide from Fei-Dei Li, Justin Johnson, Serena Yeung, **cs231n Stanford**

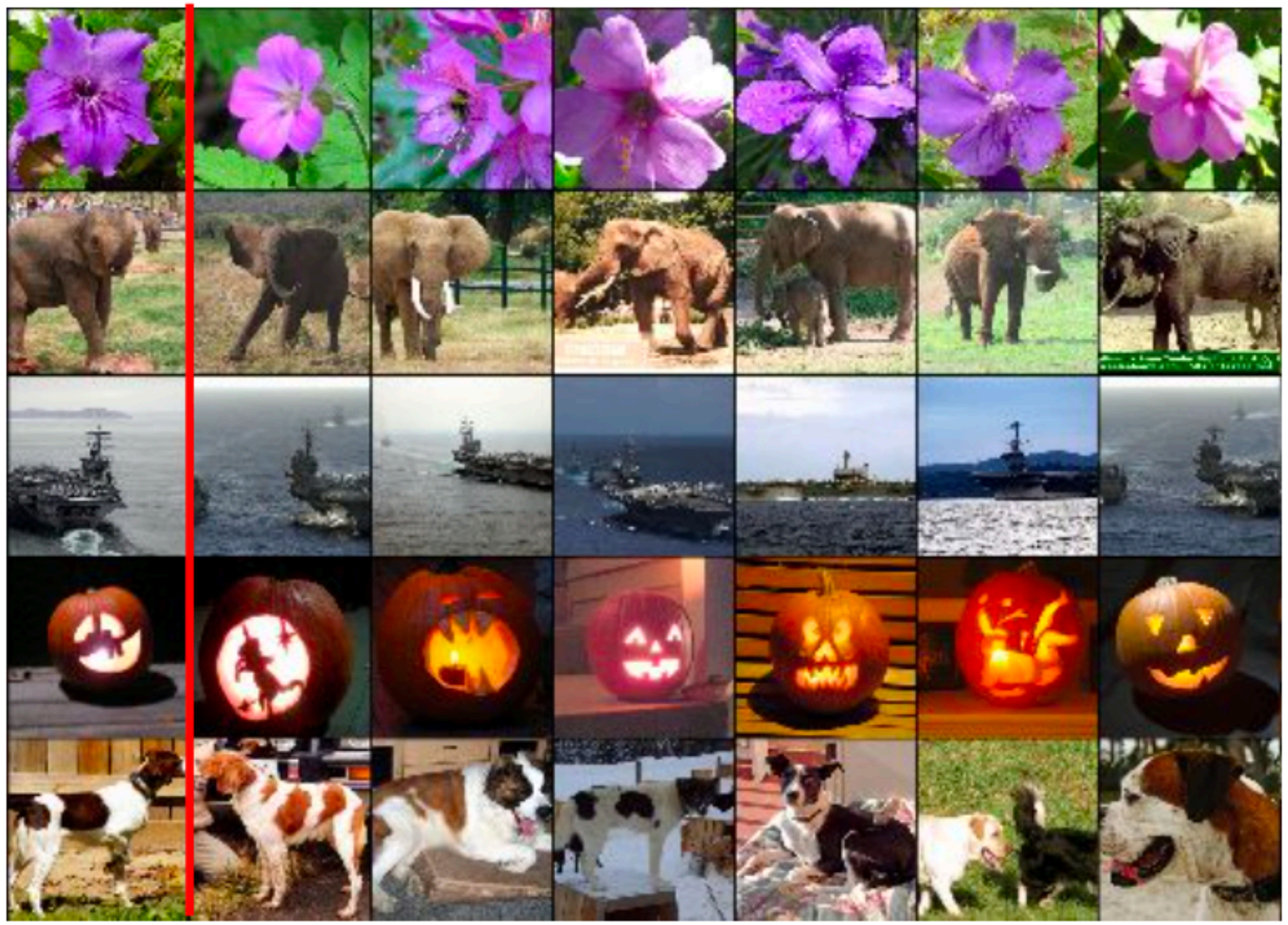
Last Layer

Recall: Nearest neighbors in pixel space



Test image

L2 Nearest neighbors in feature space

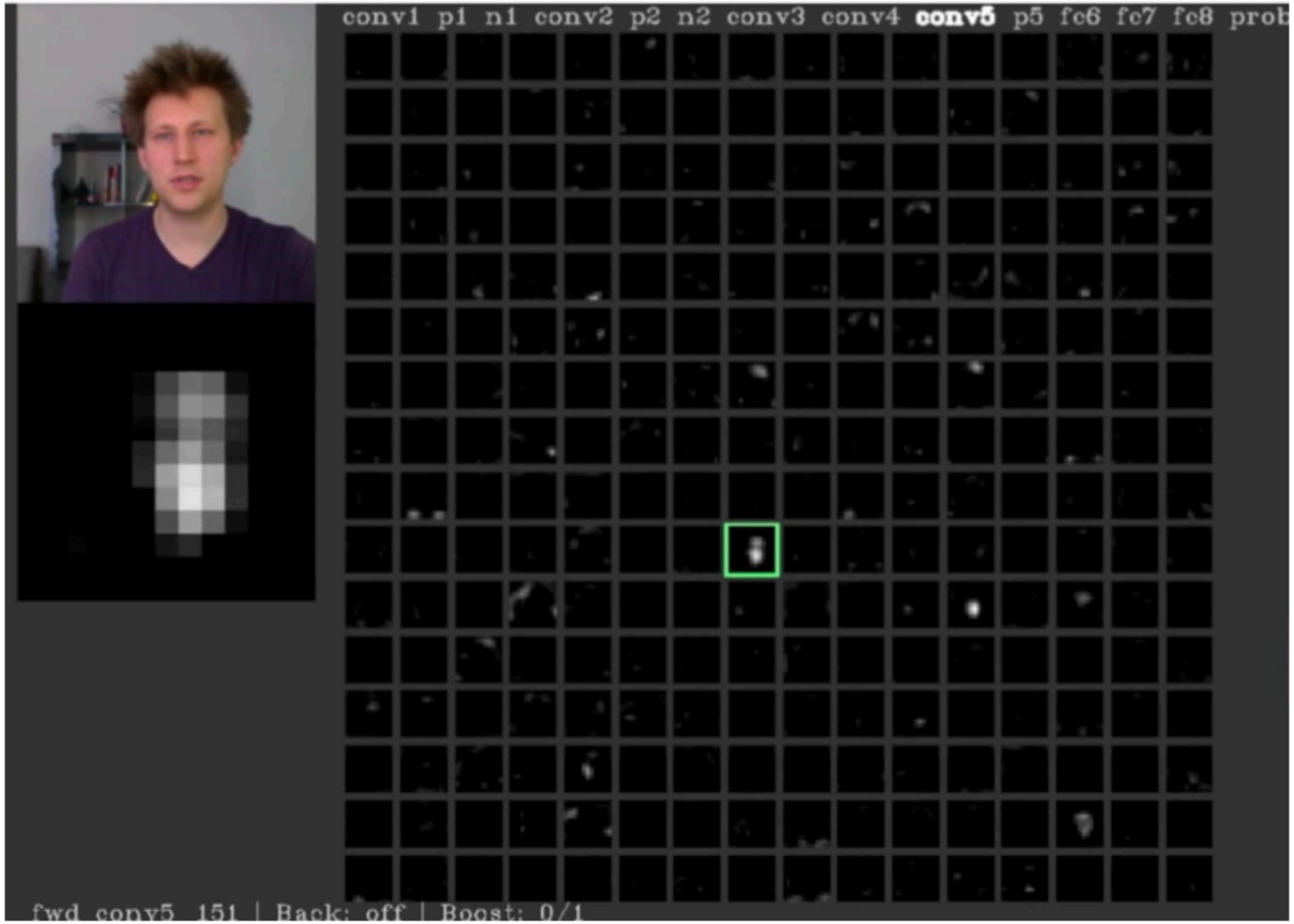


… you are doing this for **Assignment 2**

* slide from Fei-Dei Li, Justin Johnson, Serena Yeung, **cs231n Stanford**

Visualizing **Activations**

conv5 feature map of AlexNet is 128x13x13; visualize as 128 13x13 grayscale images


conv1 p1 n1 conv2 p2 n2 conv3 conv4 conv5 p5 fc6 fc7 fc8 prob
fwd conv5 151 | Back: off | Boost: 0/1

[Yosinski et al., 2014]

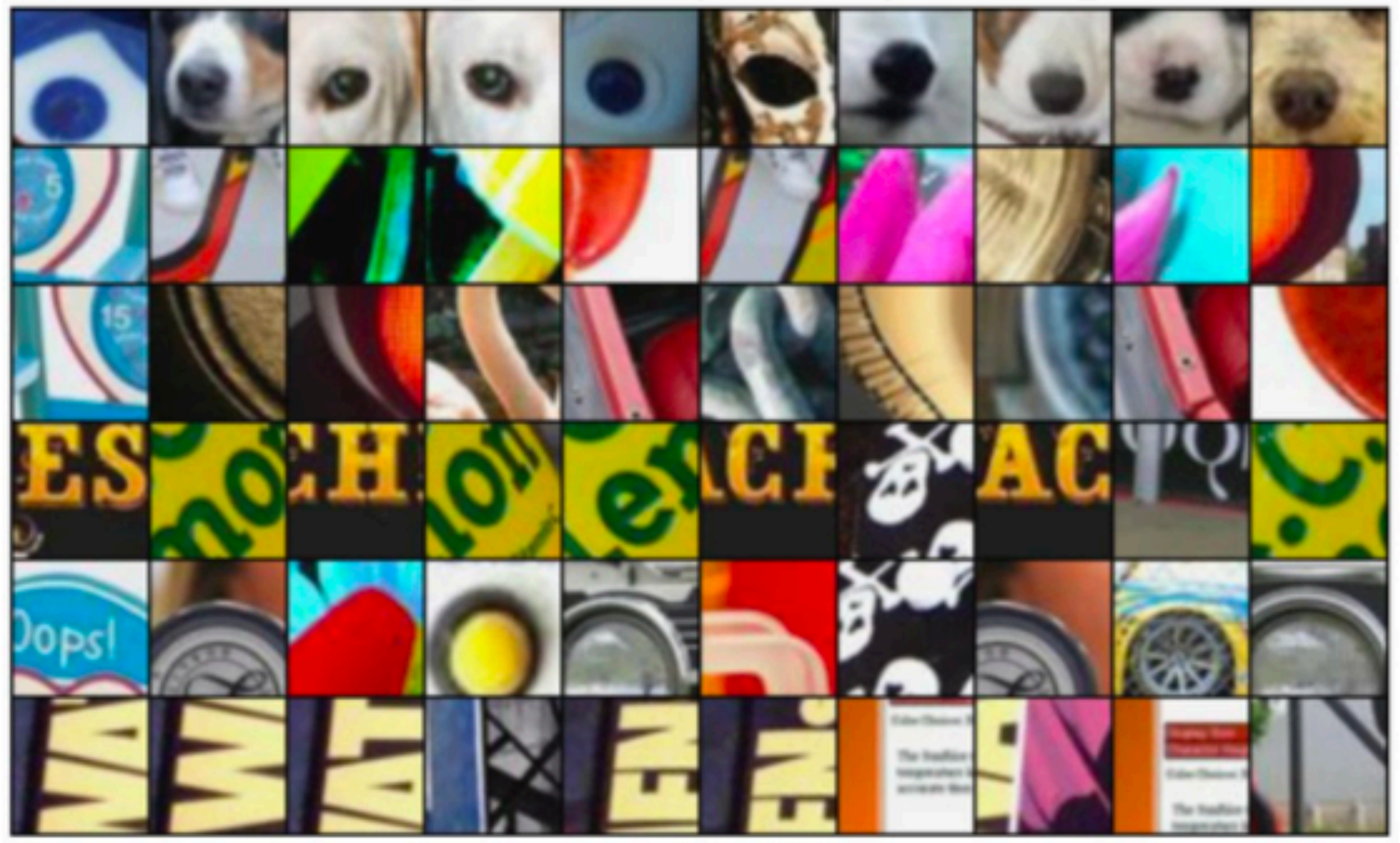
* slide from Fei-Dei Li, Justin Johnson, Serena Yeung, **cs231n Stanford**

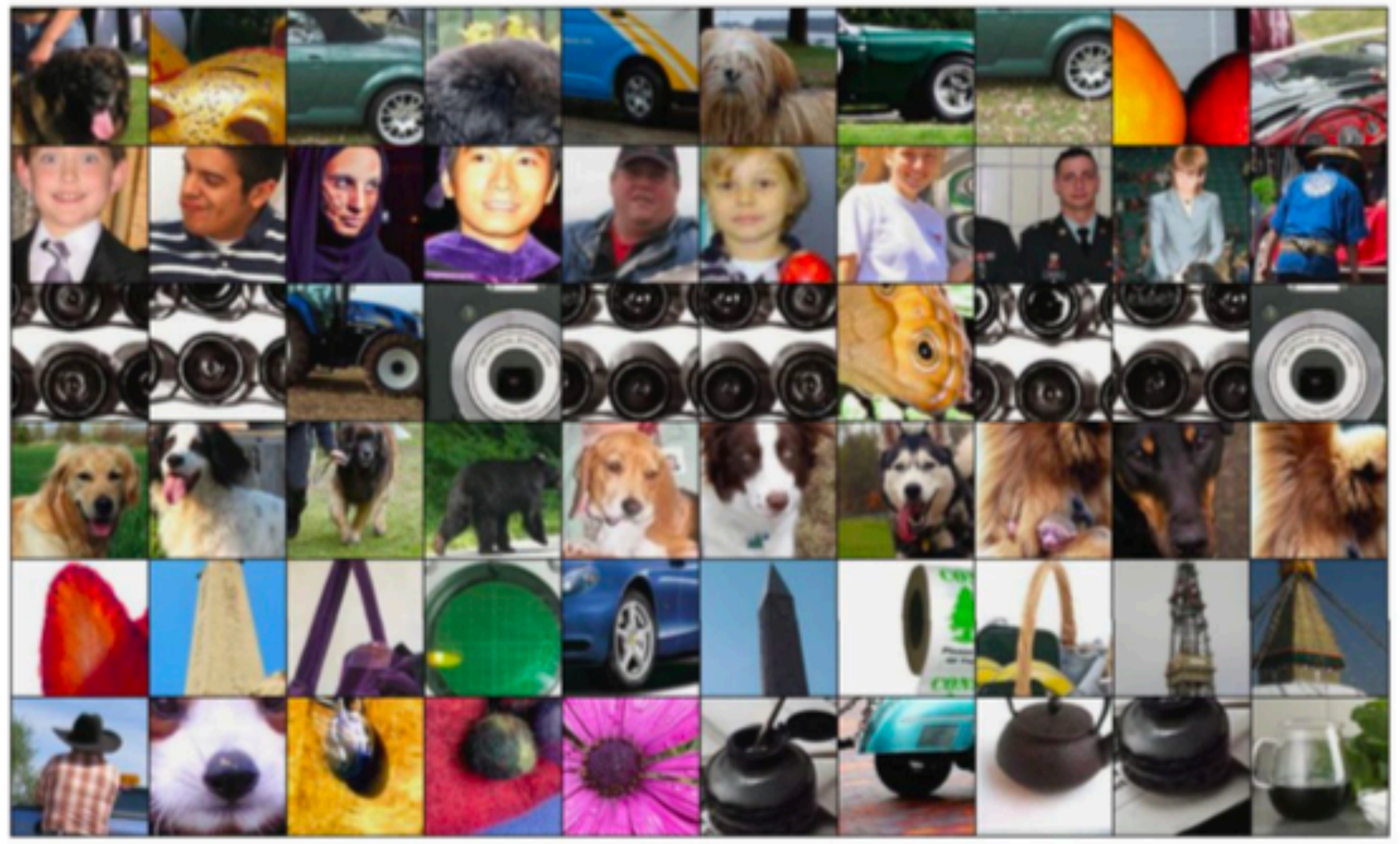
Maximally **Activating Patches**

— Pick a layer and a channel; e.g., cons5 of AlexNet is 128x13x13

— Run many images through the network

— Visualize image patches that correspond to maximal activation of the neuron





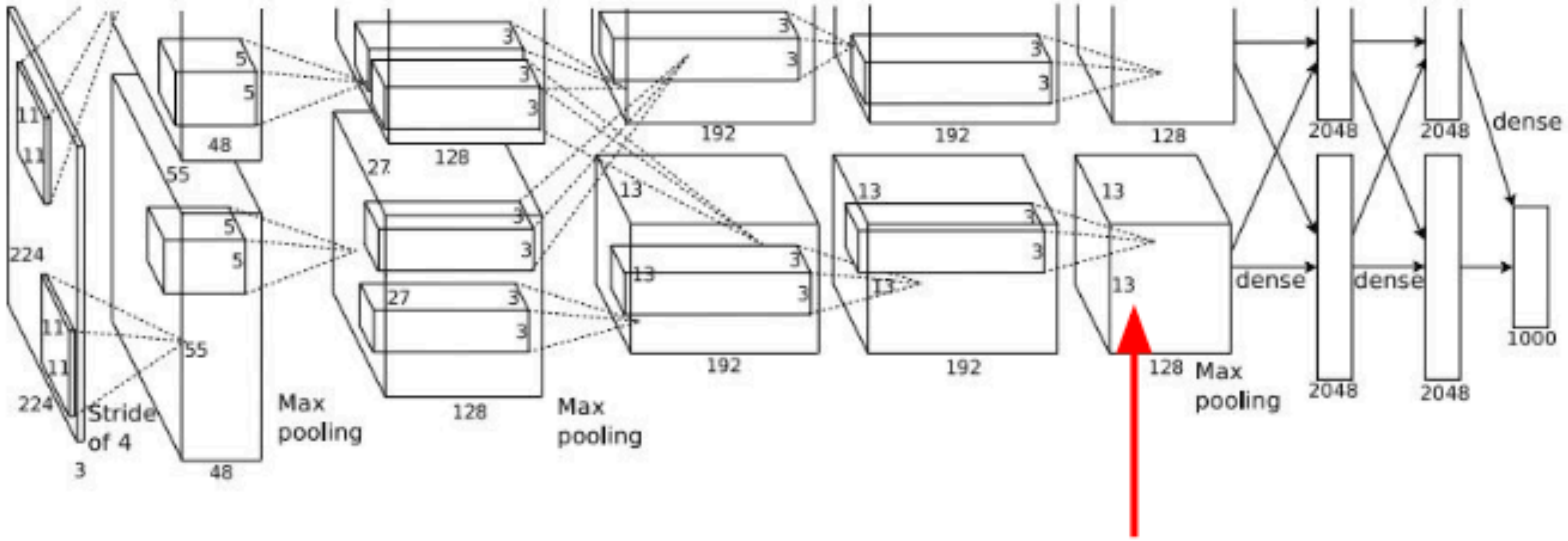
[Springenberg et al., 2015]

* slide from Fei-Dei Li, Justin Johnson, Serena Yeung, **cs231n Stanford**

Intermediate Features through (**Guided**) **BackProp**

— Pick a single intermediate neuron somewhere in the network, e.g., neuron in 128x13x13 conv5 feature map

— Compute **gradient of neuron value with respect to image pixels**

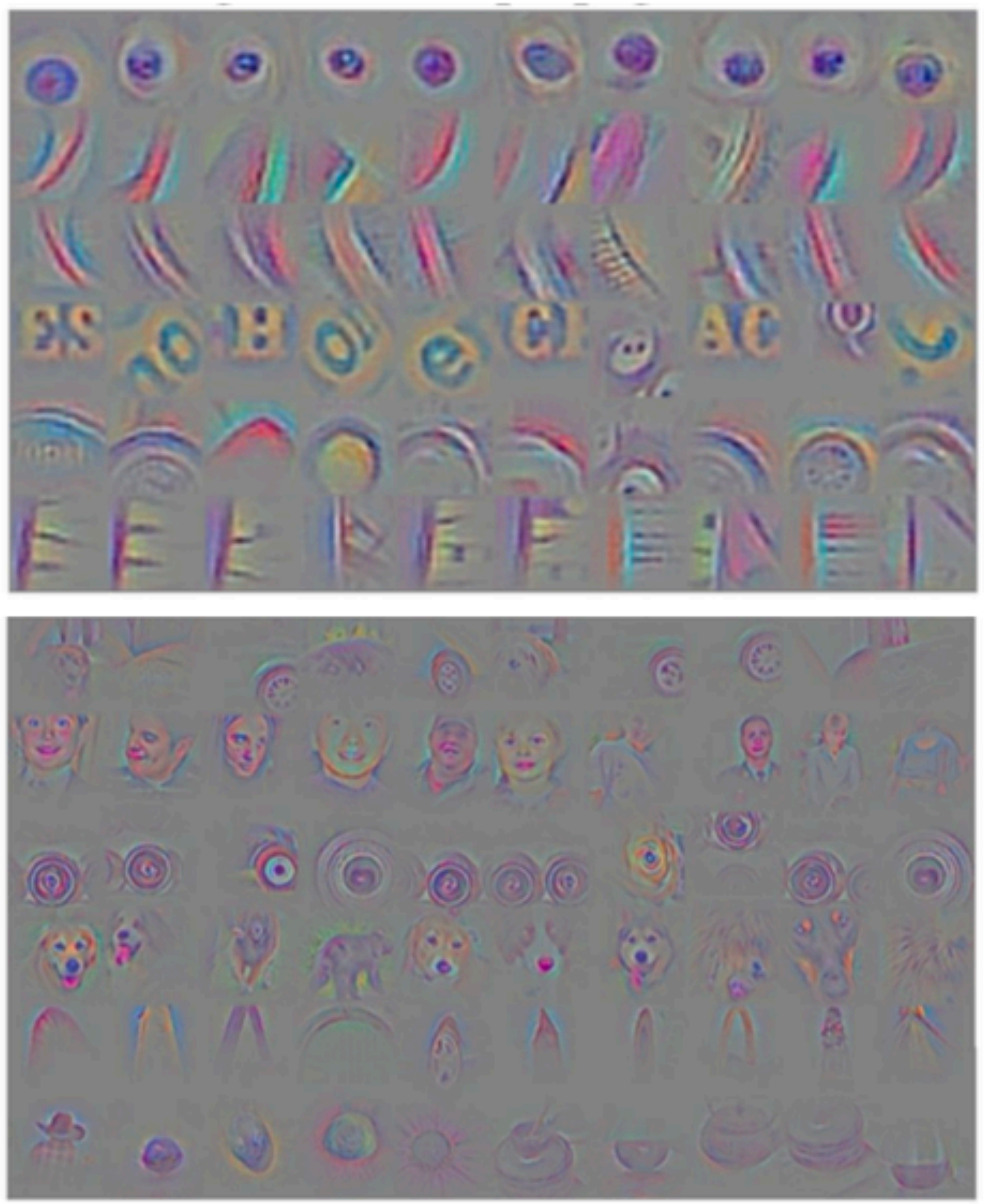


[Springenberg et al., 2015]

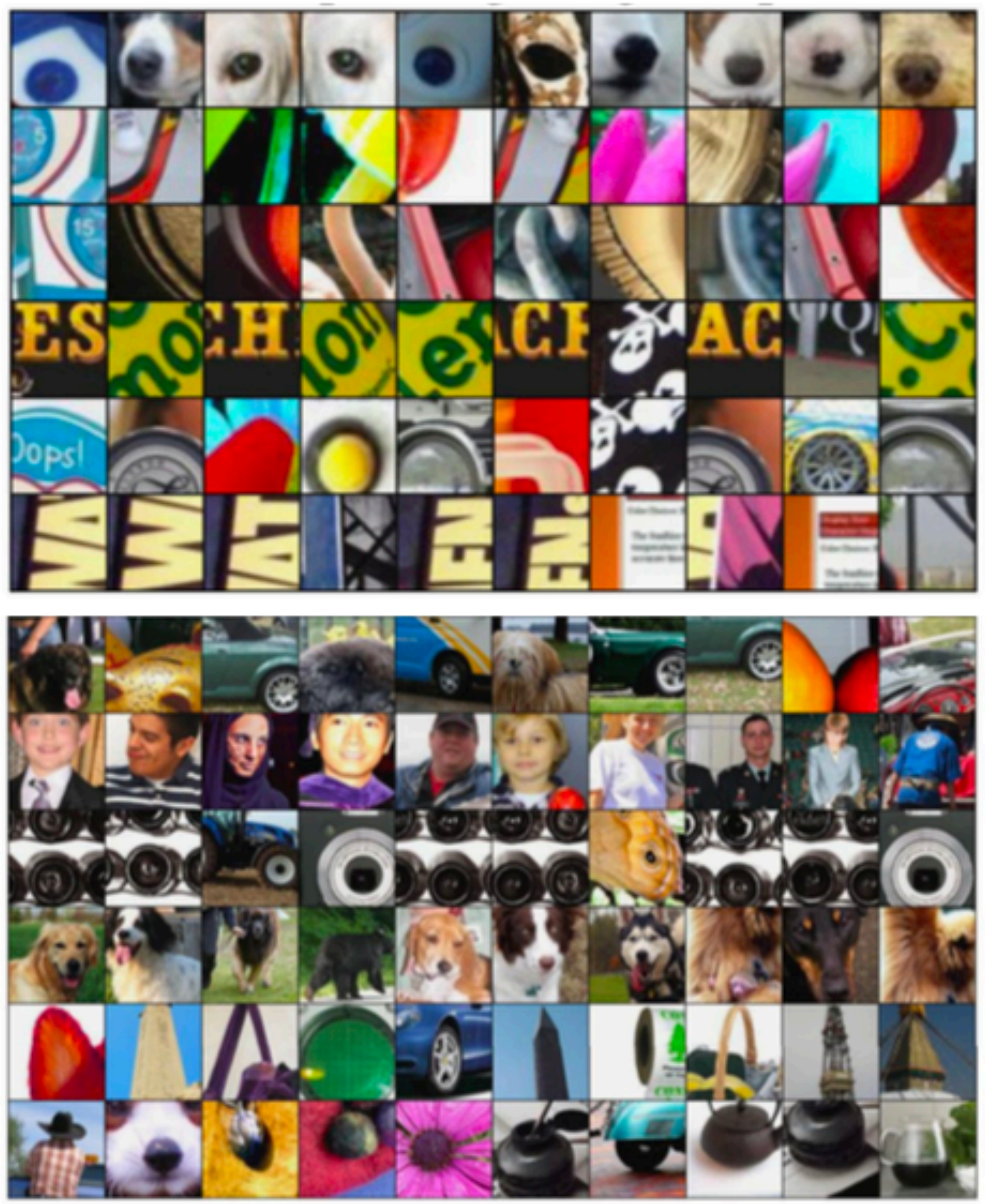
[Zeiler and Fergus, 2014]

* slide from Fei-Dei Li, Justin Johnson, Serena Yeung, **cs231n Stanford**

Intermediate Features through (**Guided**) **BackProp**



[Springenberg et al., 2015]



[Zeiler and Fergus, 2014]

* slide from Fei-Dei Li, Justin Johnson, Serena Yeung, **cs231n Stanford**

Gradient **Ascent**

(Guided) **BackProp**: find the part of an image that a neuron responds to

Gradient ascent: generate a synthetic image that maximally activates a neuron

$$\mathbf{I}^* = \arg\max_{\mathbf{I}} \underbrace{f(\mathbf{I})}_{\textbf{Neuron Value}} + \overbrace{R(\mathbf{I})}^{\text{Natural }\textbf{Image Regularizer}}$$

* slide from Fei-Dei Li, Justin Johnson, Serena Yeung, **cs231n Stanford**

Gradient **Ascent**

1. Initialize image with all zeros (can also start with an existing image)
2. Forward image to compute the current scores
3. BackProp to get gradient of the neuron with respect to image pixels
4. Make a small update to an image

$$\mathbf{I}^* = \arg\max_{\mathbf{I}} \underbrace{f(\mathbf{I})}_{\textbf{Neuron Value}} + \overbrace{R(\mathbf{I})}^{\text{Natural } \textbf{Image Regularizer}}$$

* slide from Fei-Dei Li, Justin Johnson, Serena Yeung, **cs231n Stanford**

Gradient **Ascent**

1. Initialize image with all zeros (can also start with an existing image)
2. Forward image to compute the current scores
3. BackProp to get gradient of the neuron with respect to image pixels
4. Make a small update to an image

Natural **Image Regularizer** $R(\mathbf{I}) = -\lambda||\mathbf{I}||_2^2$

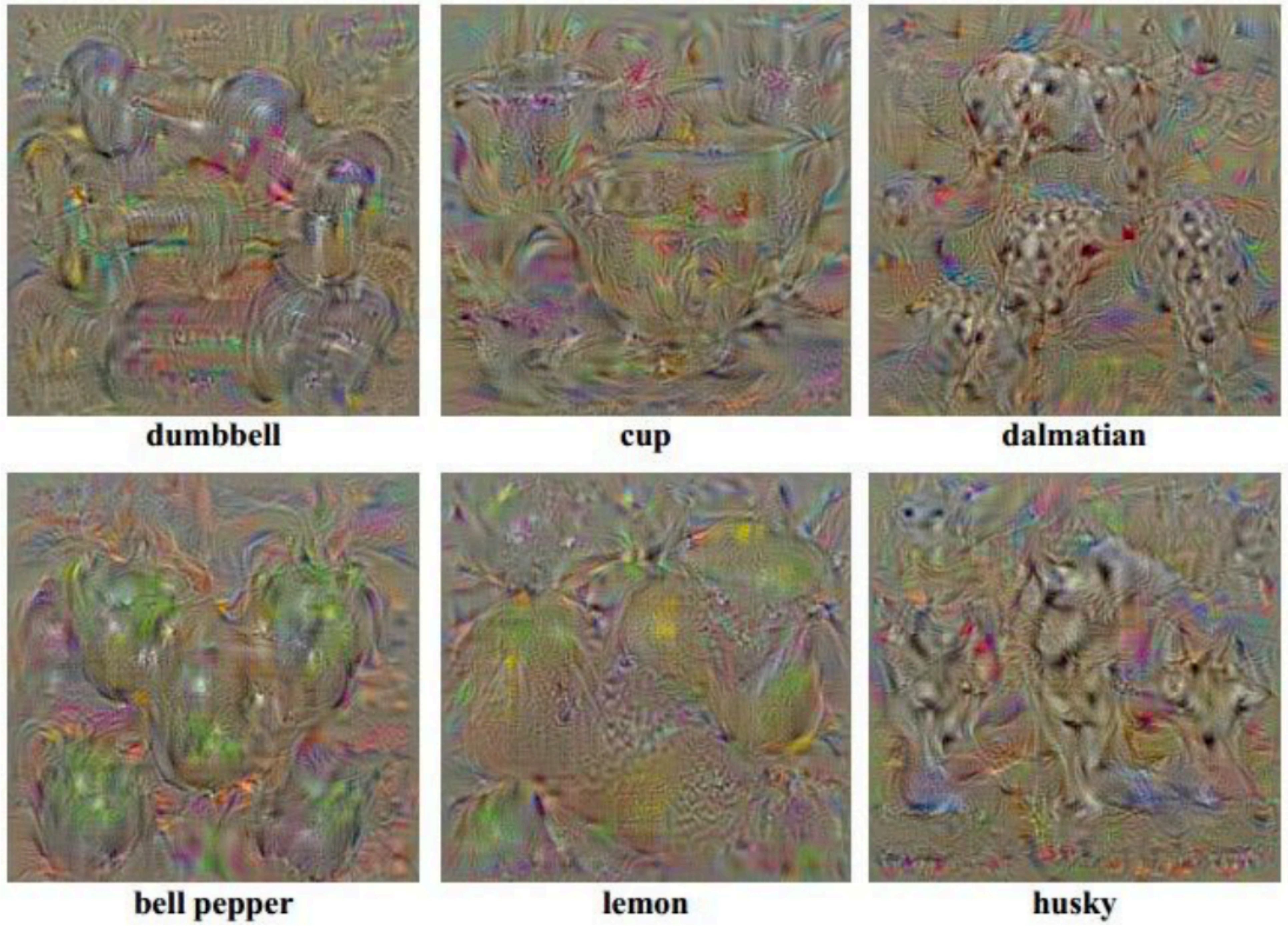
$$\mathbf{I}^* = \arg\max_{\mathbf{I}} f(\mathbf{I}) + R(\mathbf{I})$$

Score for class C before softmax

[Simonyan et al., 2014]

* slide from Fei-Dei Li, Justin Johnson, Serena Yeung, **cs231n Stanford**

Gradient **Ascent**



Natural **Image Regularizer** $R(\mathbf{I}) = -\lambda||\mathbf{I}||_2^2$

$$\mathbf{I}^* = \arg\max_{\mathbf{I}} \underbrace{f(\mathbf{I})}_{\text{Score for class C before softmax}} + \overbrace{R(\mathbf{I})}^{\text{Natural Image Regularizer}}$$

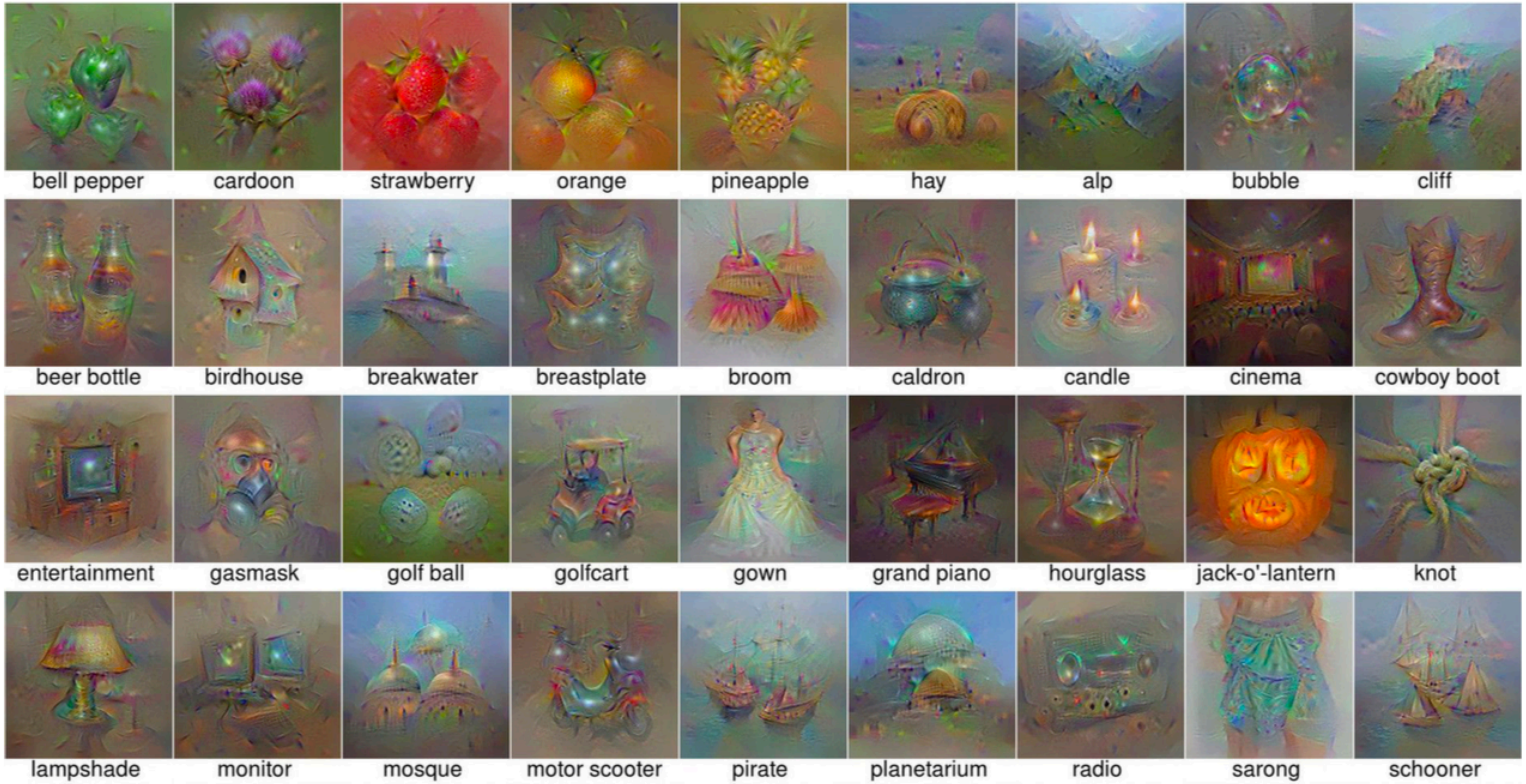
Score for class C before softmax

[Simonyan et al., 2014]

* slide from Fei-Dei Li, Justin Johnson, Serena Yeung, **cs231n Stanford**

Gradient **Ascent**

… with a few additional tweaks

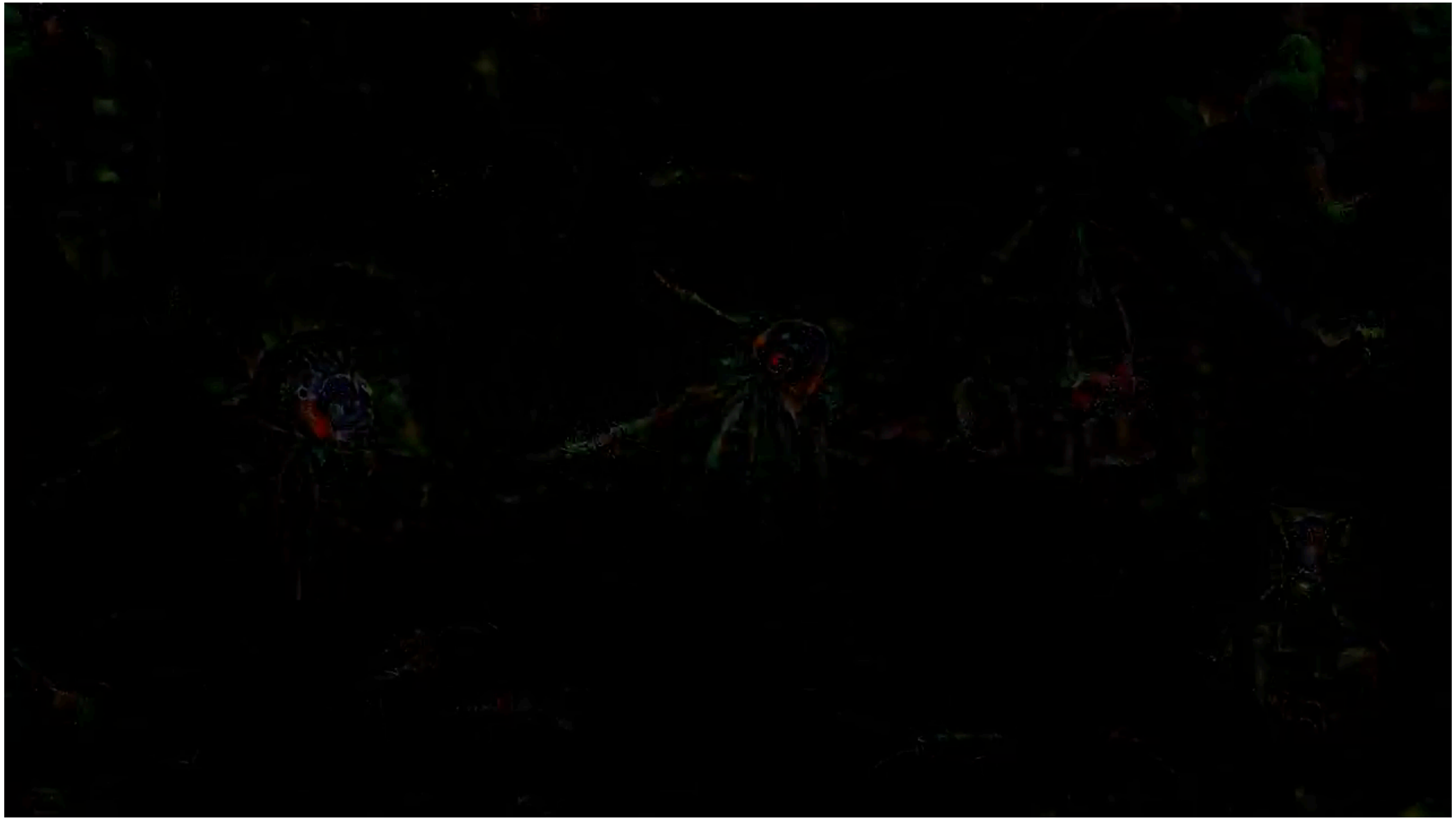


[Nguyen et al., 2015]

* slide from Fei-Dei Li, Justin Johnson, Serena Yeung, **cs231n Stanford**

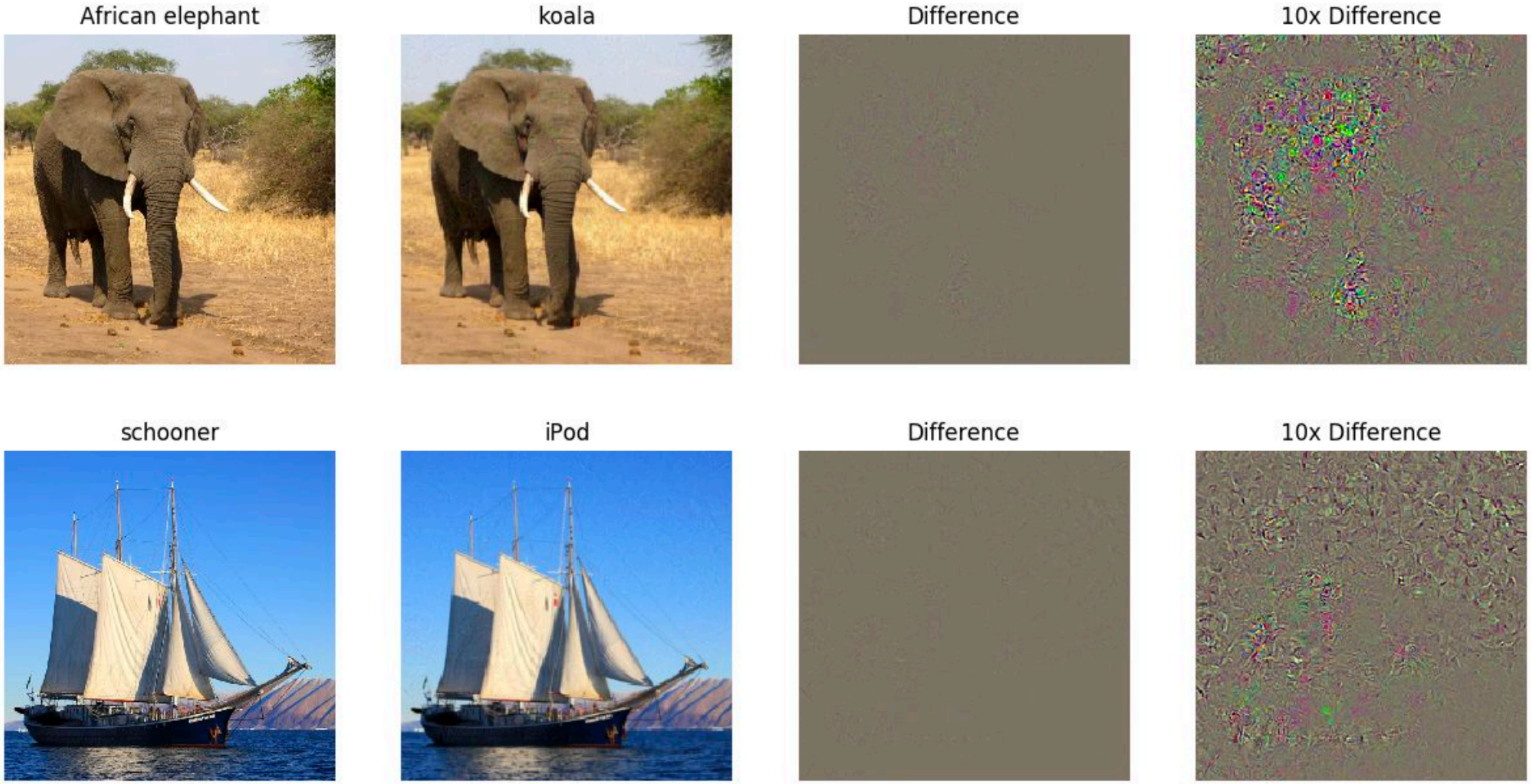
Deep **Dream**

[Mordvinsev, Olah, Tyka]



https://www.youtube.com/watch?v=DgPaCWJL7XI&t=11s

Fooling Images / **Adversarial** Examples



* slide from Fei-Dei Li, Justin Johnson, Serena Yeung, **cs231n Stanford**