Topics:

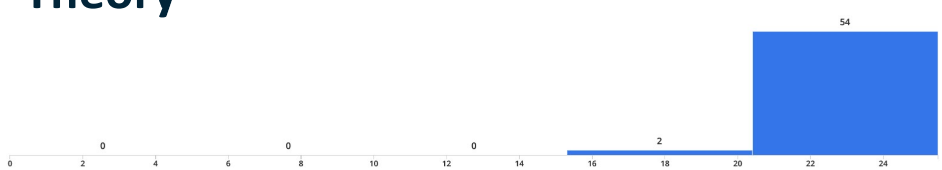
- Visualization
- Advanced Computer Vision Architectures

CS 4644-DL / 7643-A
ZSOLT KIRA

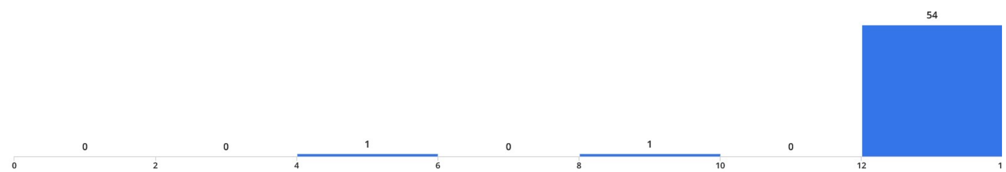
- **Assignment 1 grades out**
  - 1 week from release for re-grade requests

CS4644

Theory

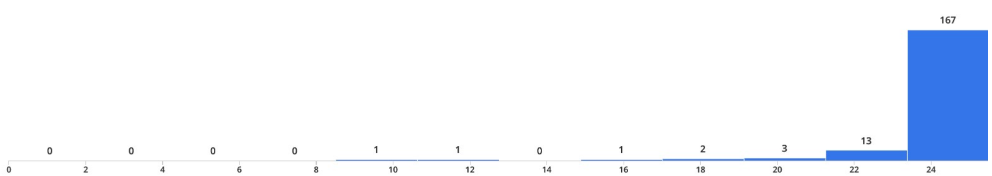


Minimum	Median	Maximum	Mean	Std Dev
18.5	25.5	28.5	25.54	2.14

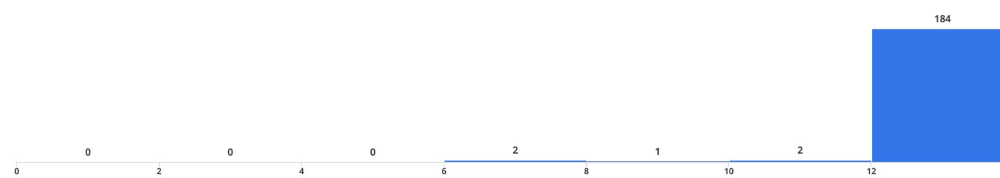


Minimum	Median	Maximum	Mean	Std Dev
4.0	14.0	14.0	13.72	1.46

CS7643



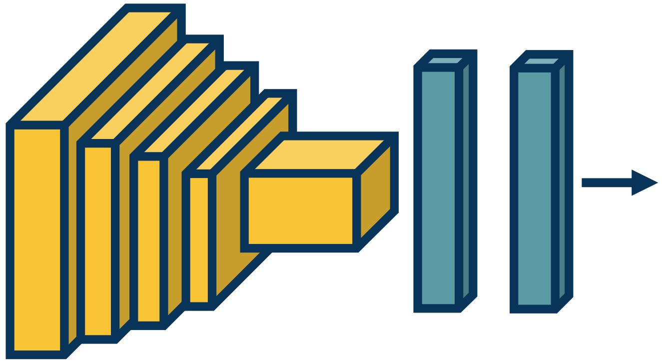
Minimum	Median	Maximum	Mean	Std Dev
9.6	25.5	28.5	25.58	2.69



Minimum	Median	Maximum	Mean	Std Dev
6.0	14.0	14.0	13.82	0.96

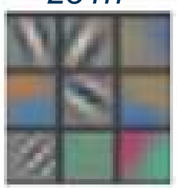
- **Assignment 2**
  - Due Feb 24th!

- **Projects!**
  - Meta projects up
    - Note: Typically advanced
  - Start forming teams!



Weights

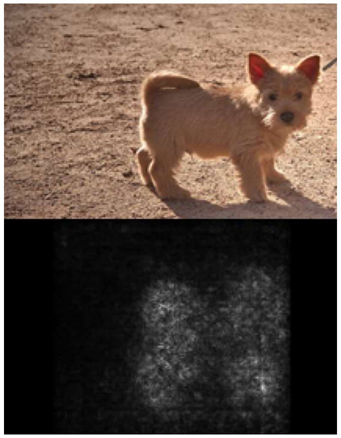
Fei-Fei Li, Justin Johnson, Serena Yeung, from CS 231n



Zeiler & Fergus, 2014

Activations

Gradients



Simonyan et al, 2013

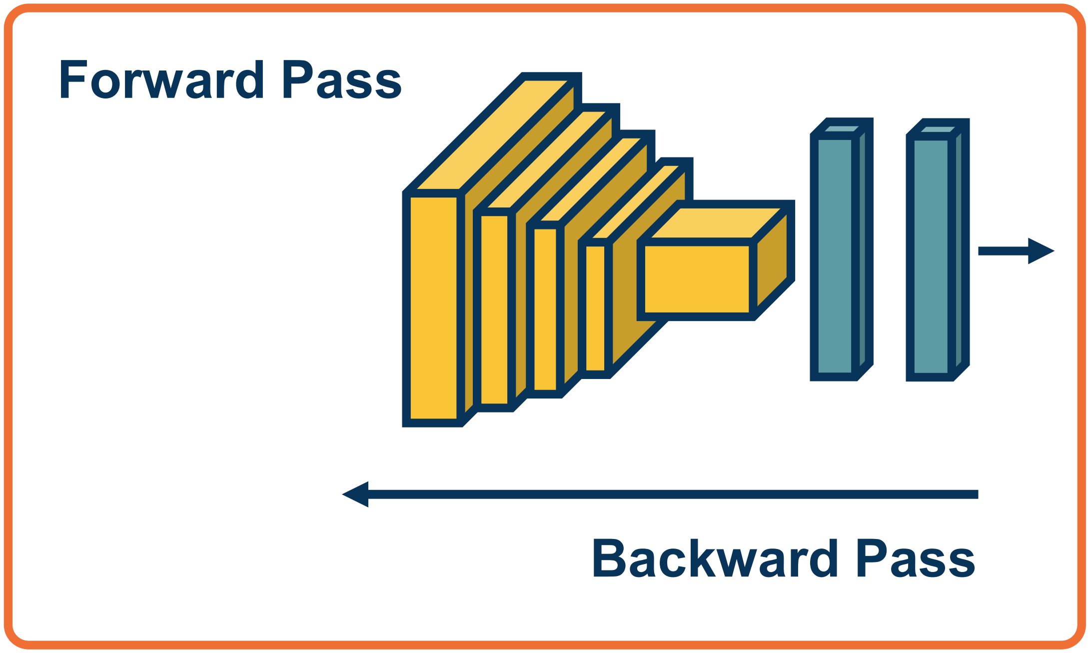
Robustness

Hendrycks & Dietterich, 2019

Visualizing Neural Networks

Gradient-Based Visualizations

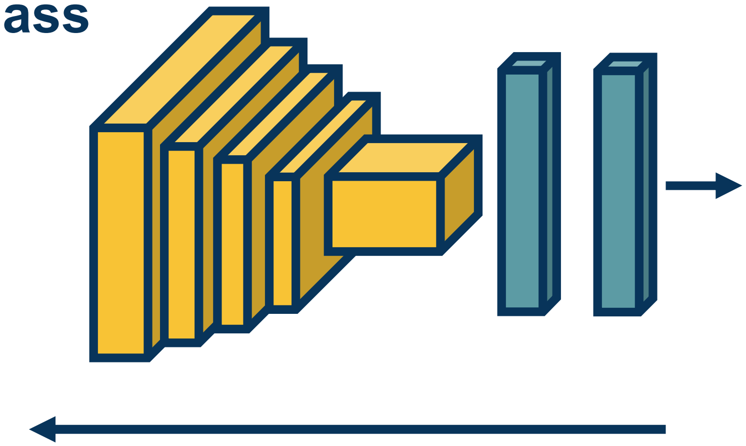
Given a **trained** model, we can perform forward pass given an input to get scores, softmax probabilities, loss and then backwards pass to get gradients



- Note: We are keeping parameters/weights **frozen**
  - Do not use gradients w.r.t. weights to perform updates

Backwards pass gives us **gradients** for all layers: How the loss changes as we change different parts of the input

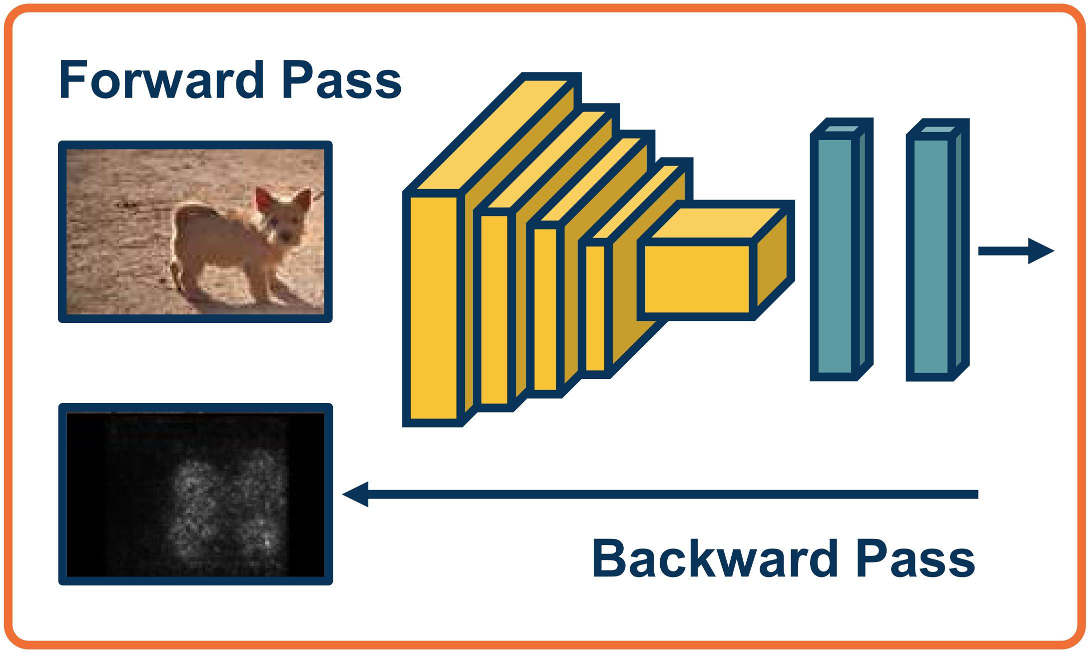
This can be **useful not just for optimization**, but also to understand what was learned



- Gradient of **loss** with respect to **all layers** (including input!)
- Gradient of **any layer** with respect to **input** (by cutting off computation graph)

Idea: We can backprop to the image

- Sensitivity of loss to individual pixel changes
- Large sensitivity implies important pixels
- Called **Saliency Maps**



In practice:

- Instead of loss, find gradient of classifier **scores** (pre-softmax)
- Take absolute value of gradient
- Sum across all channels

From: Simonyan et al., "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps", 2013

Gradient of Loss w.r.t. Image

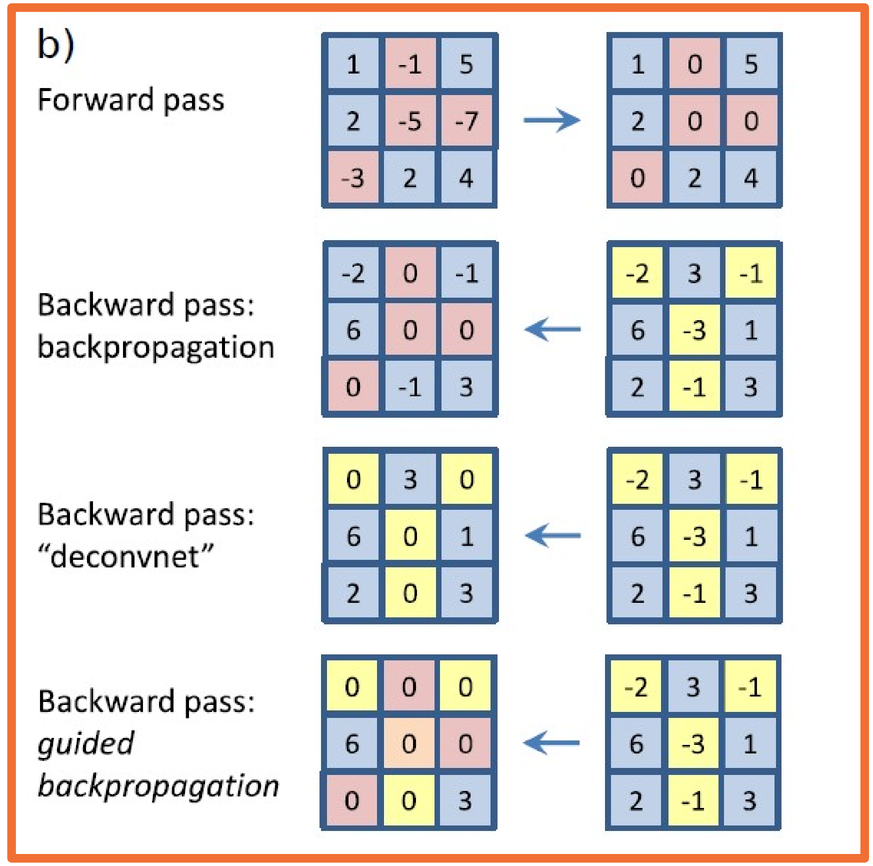
Guided Backprop

Normal backprop not always best choice

Example: You may get parts of image that **decrease** the feature activation

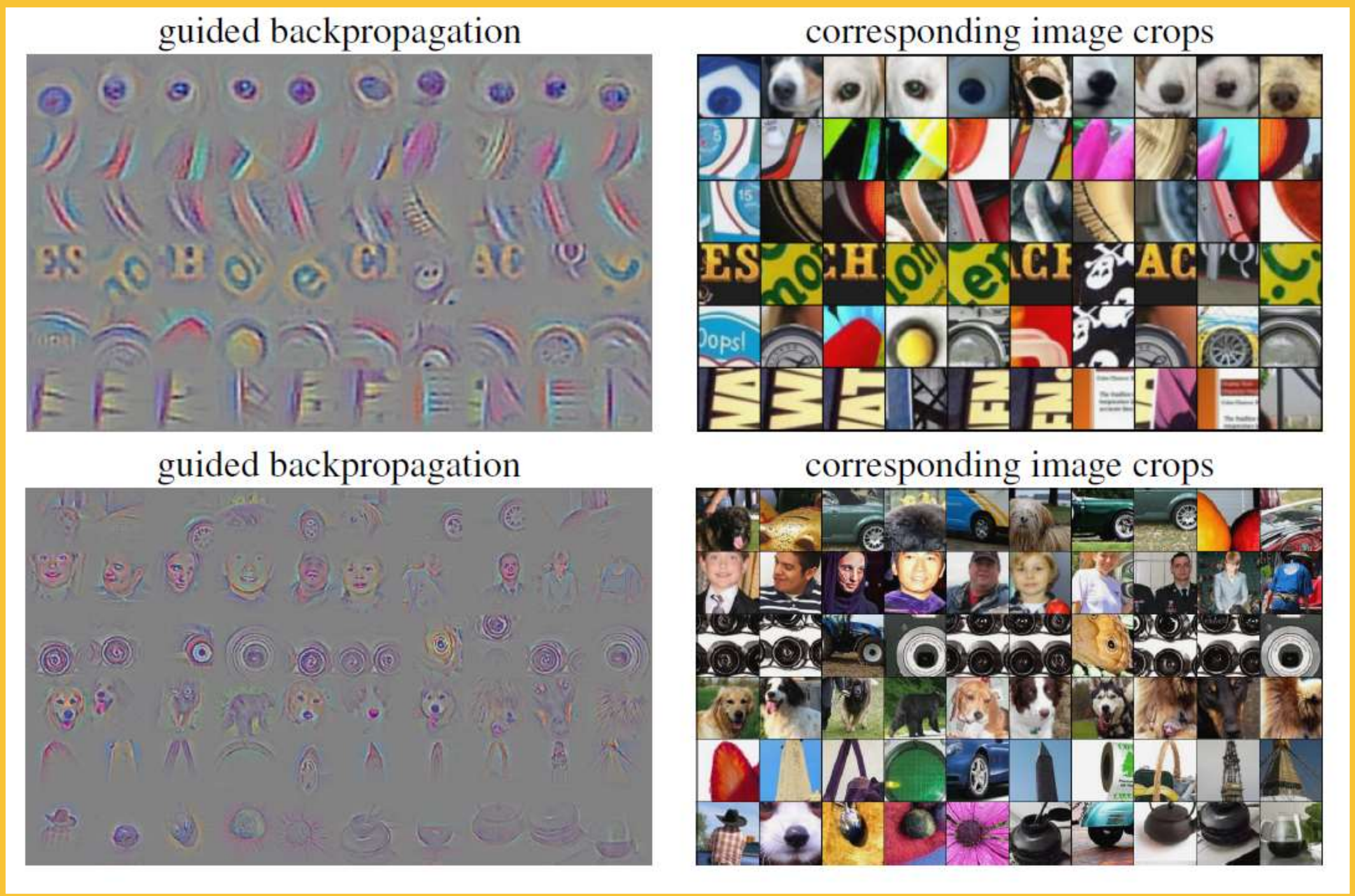
- There are probably lots of such input pixels

Guided backprop can be used to improve visualizations



From: Springenberg et al., “Striving For Simplicity: The All Convolutional Net”

Guided Backprop Results



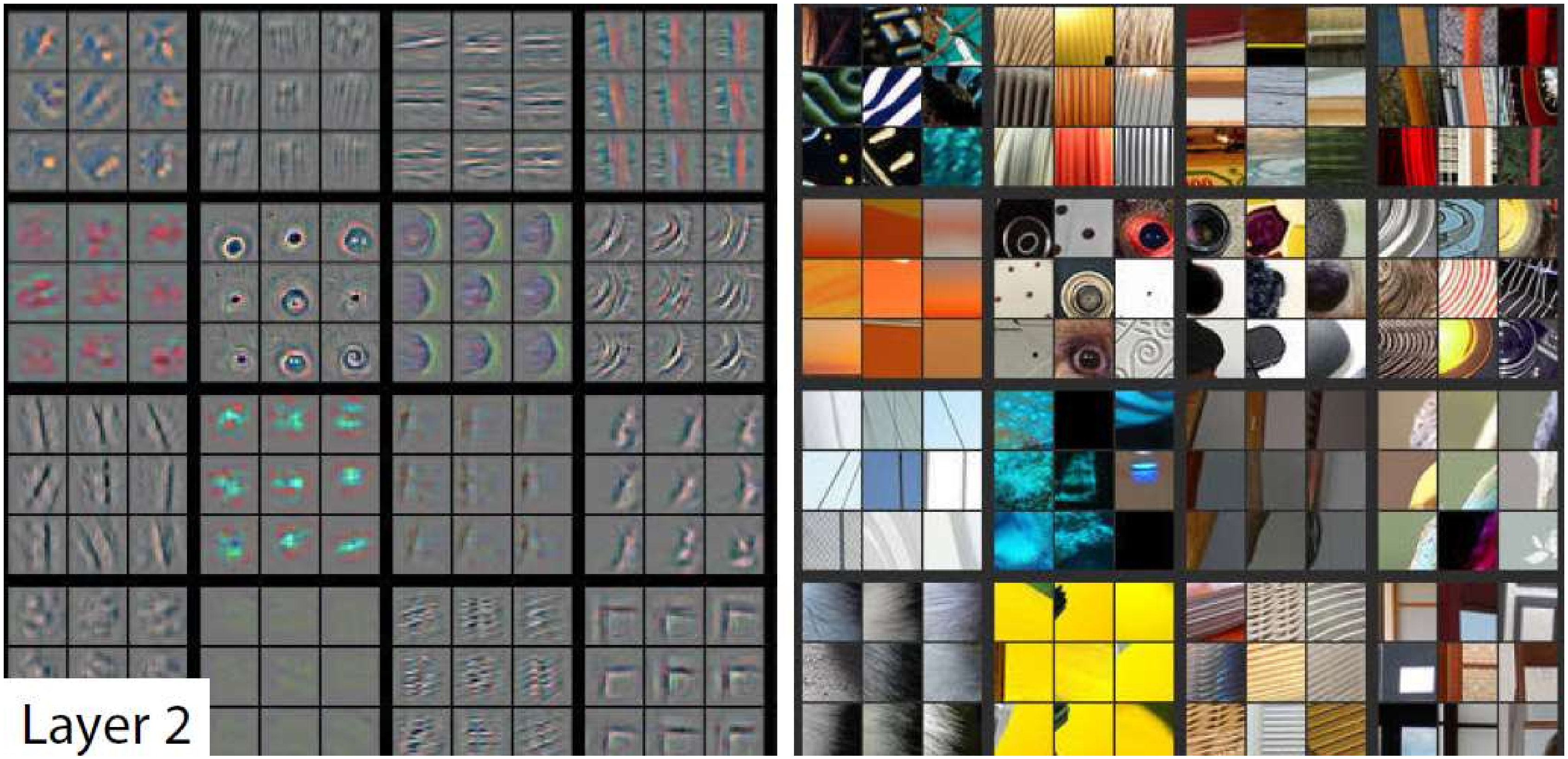
From: Springenberg et al., “Striving For Simplicity: The All Convolutional Net"

VGG Layer-by-Layer Visualization

Note: These images were created by a slightly different method called **deconvolution,** which ends up being similar to guided backprop

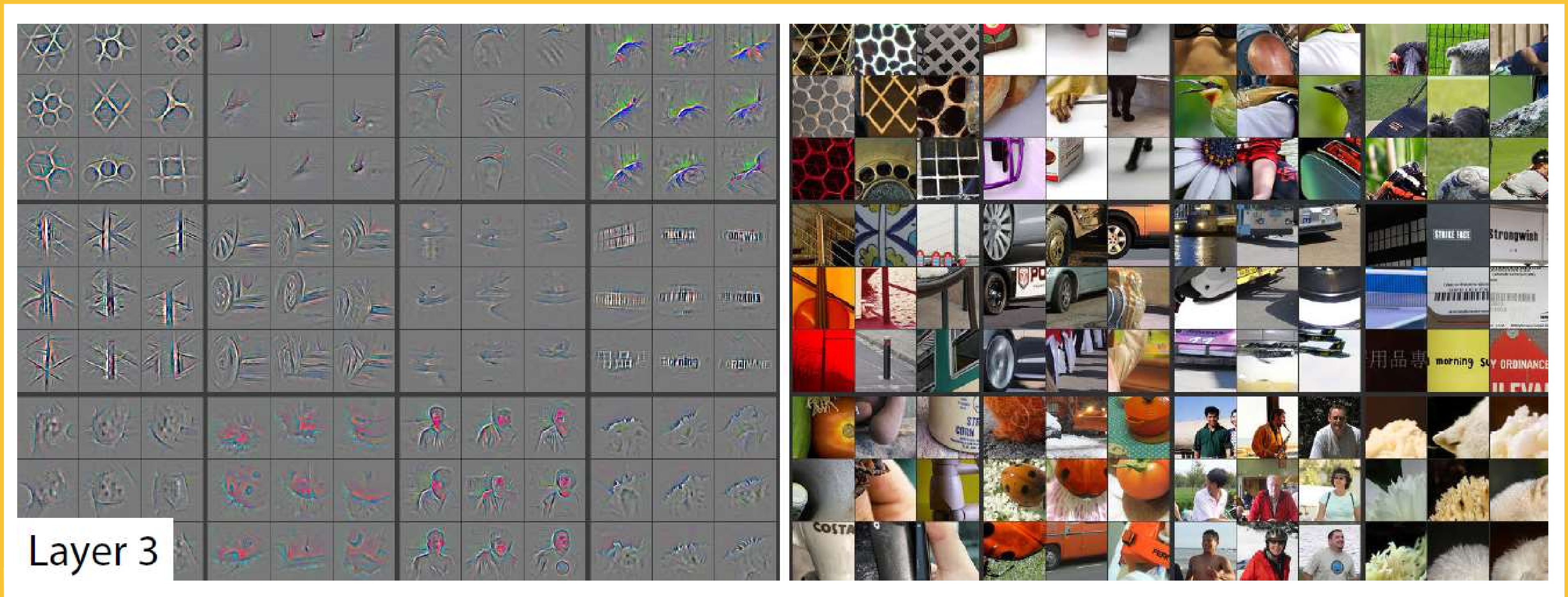
From: "Visualizing and Understanding Convolutional Networks, Zeiler & Fergus, 2014.

VGG Layer-by-Layer Visualization



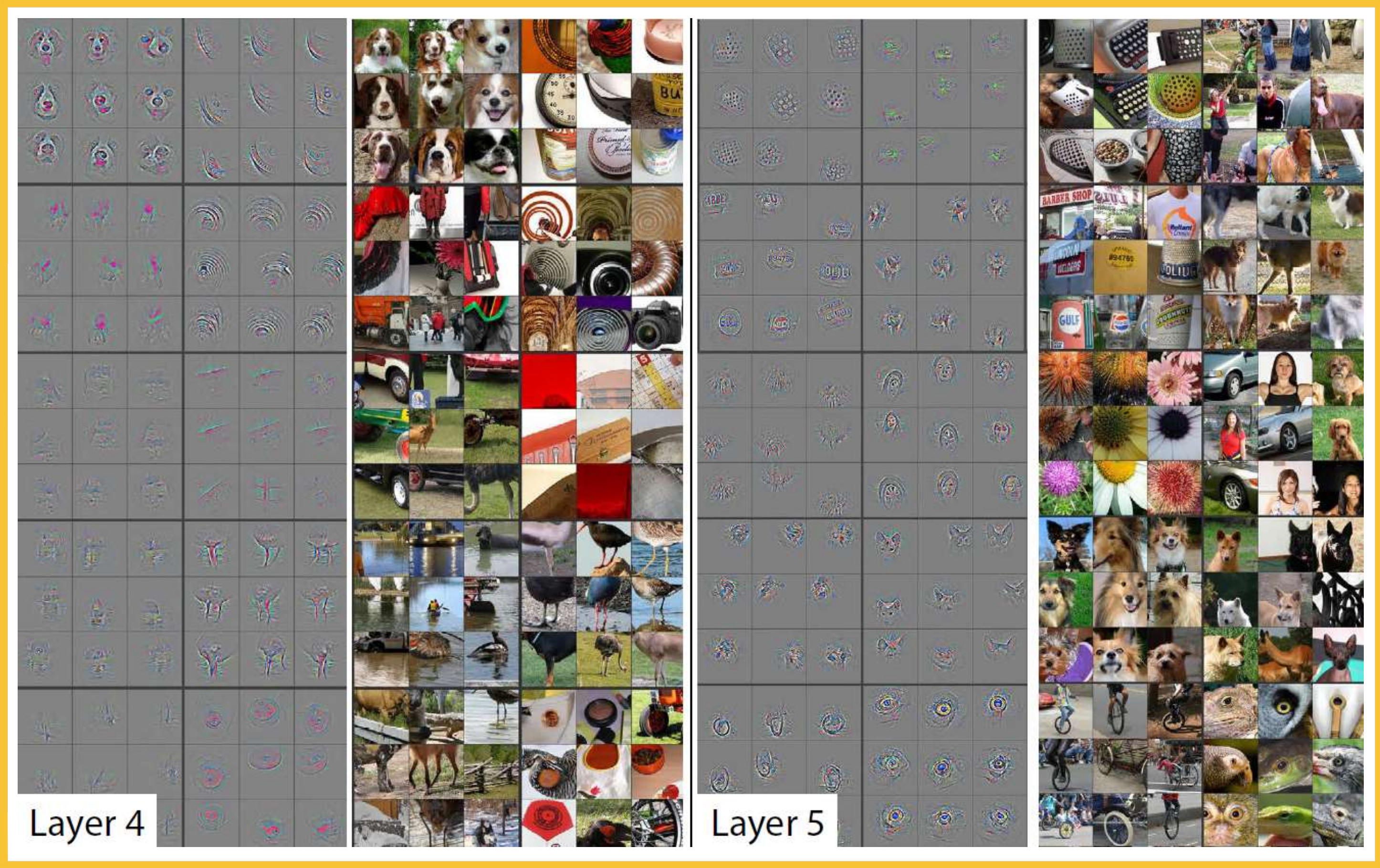
From: “Visualizing and Understanding Convolutional Networks, Zeiler & Fergus, 2014.

VGG Layer-by-Layer Visualization



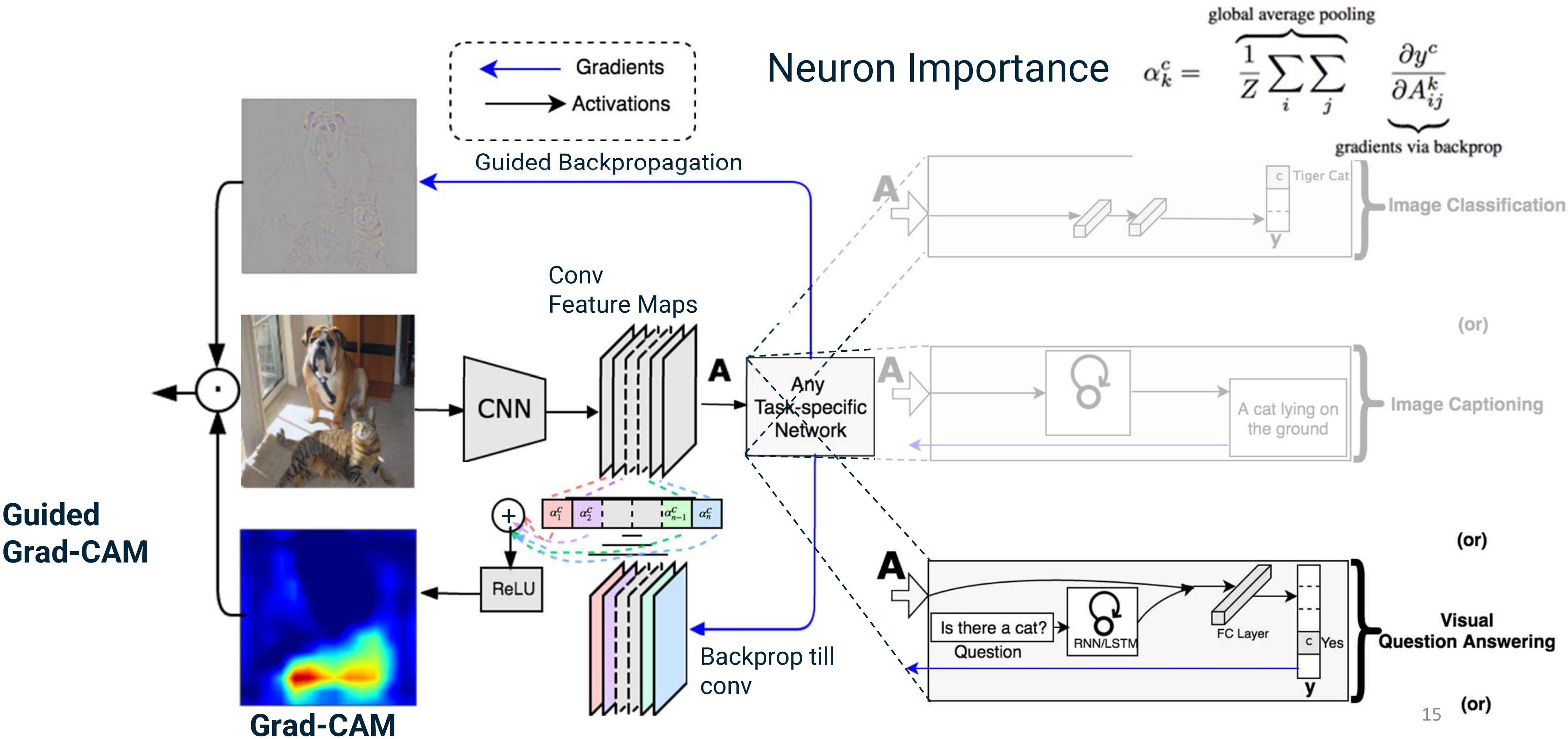
From: "Visualizing and Understanding Convolutional Networks, Zeiler & Fergus, 2014.

VGG Layer-by-Layer Visualization



From: "Visualizing and Understanding Convolutional Networks, Zeiler & Fergus, 2014.

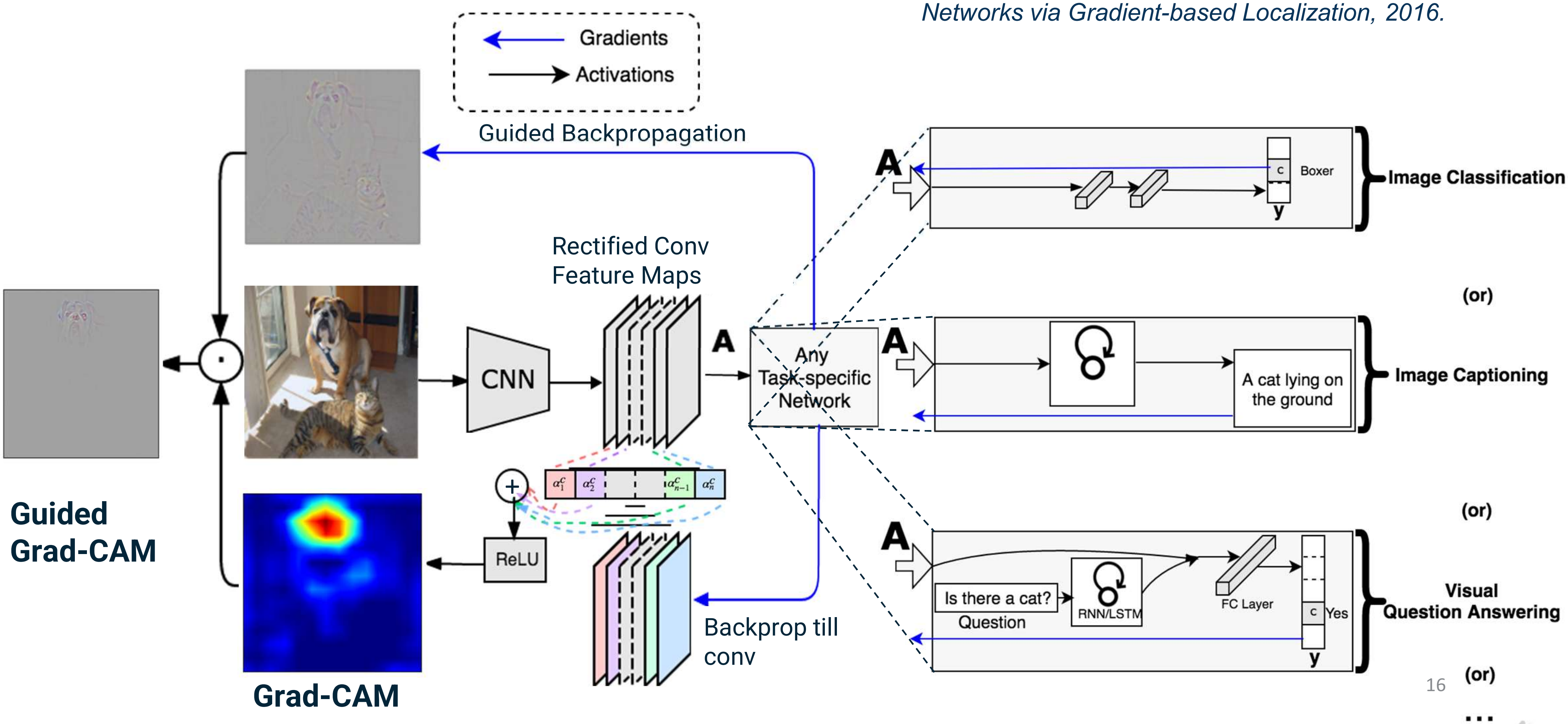
Georgia Tech



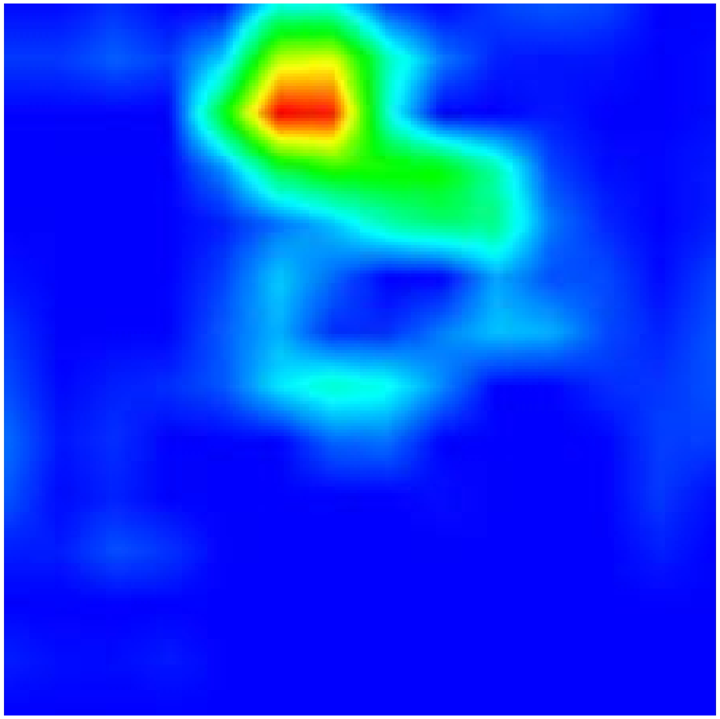
Selfvaraju et al., Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, 2016.

GradCAM

Selfvaraju et al., Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, 2016.



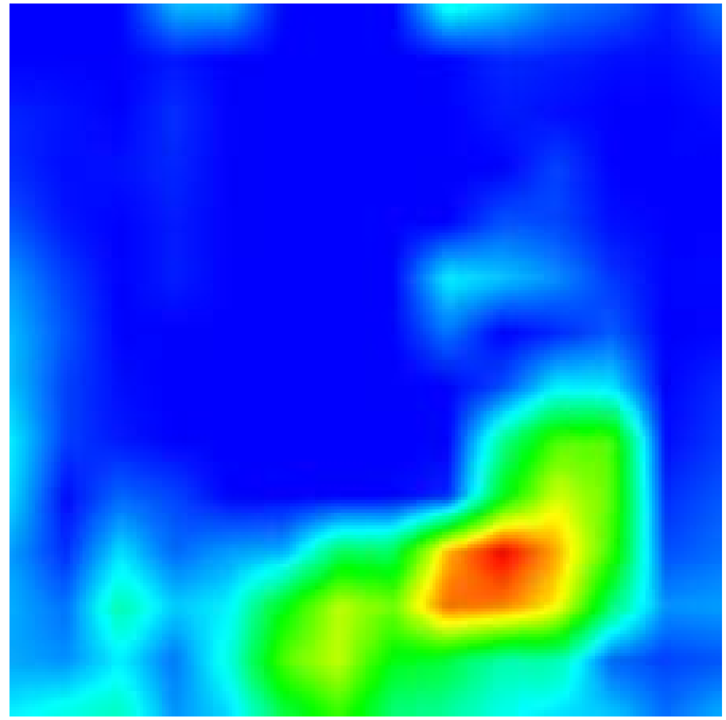
Grad-CAM

What animal is in this picture? Dog

Selfvaraju et al., Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, 2016.

Grad-CAM

What animal is in this picture? Cat

Selfvaraju et al., Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, 2016.

Grad-CAM

Summary

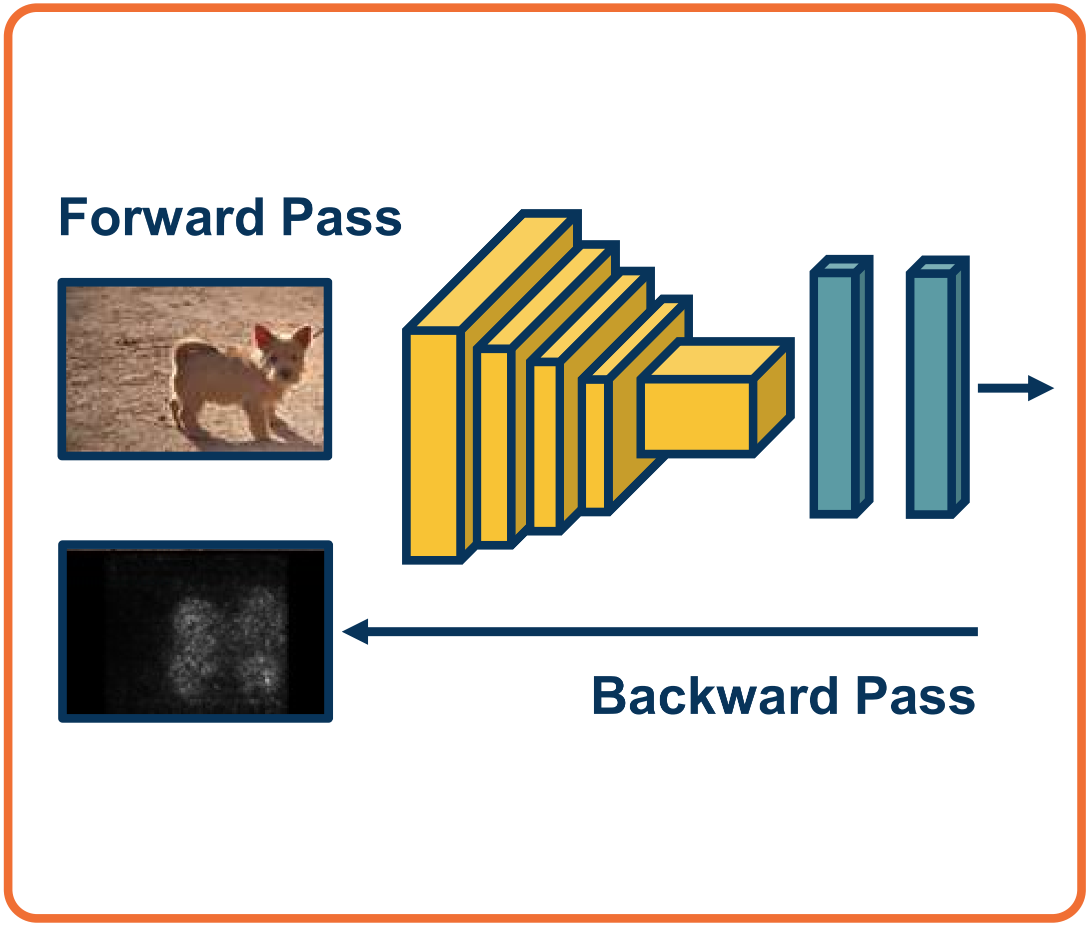
- Gradients are important **not just for optimization**, but also for **analyzing** what neural networks **have learned**
- Standard backprop **not always the most informative** for visualization purposes
- Several ways to **modify the gradient flow** to improve visualization results

Optimizing the Input Images

Idea: Since we have the gradient of scores w.r.t. inputs, can we *optimize* the image itself to maximize the score?

Why?

- Generate images from scratch!
- Adversarial examples



From: Simonyan et al., "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps", 2013

Gradient Ascent on the Scores

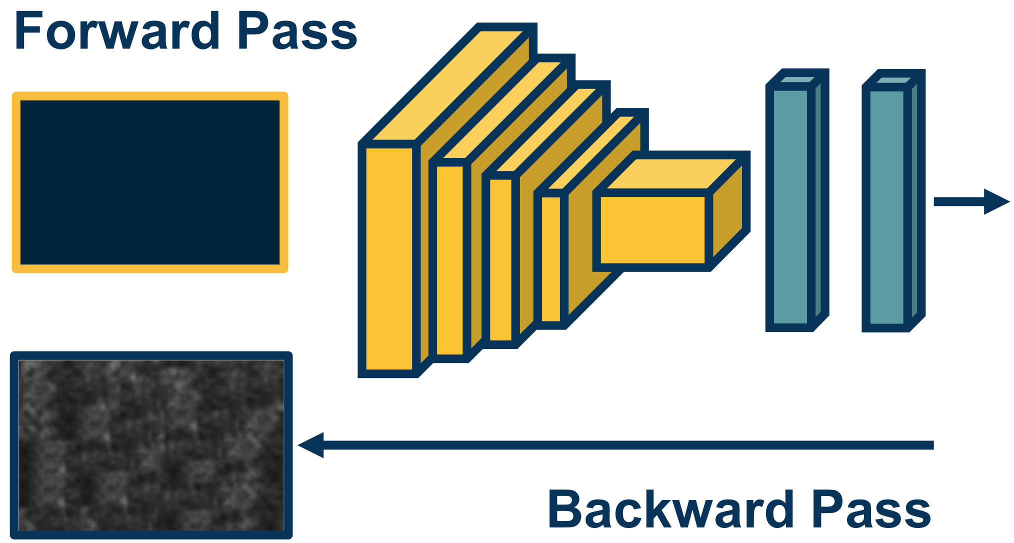
We can perform **gradient ascent** on image

- Start from random/zero image
- Use scores to avoid minimizing other class scores instead

Often need **regularization term** to induce statistics of natural imagery

- E.g. small pixel values, spatial smoothness

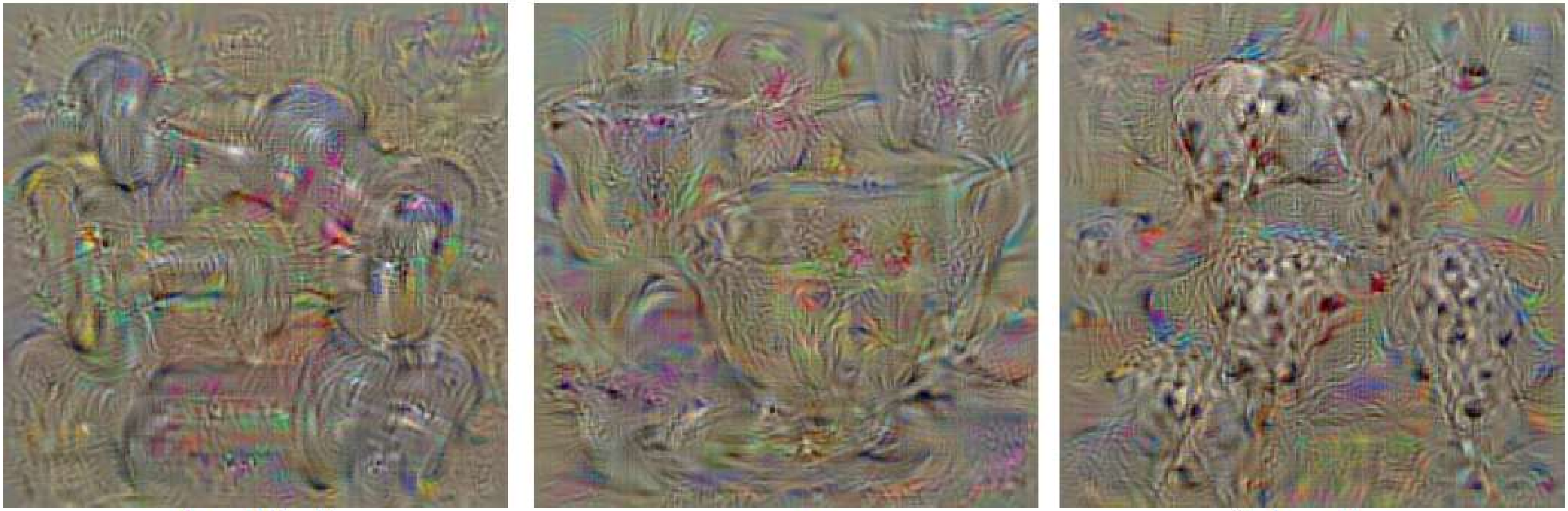
$$\arg\max_{I} S_c(I) - \lambda \left|\left|I\right|\right|_2^2$$



$$I = I + \alpha \frac{\partial S_c}{\partial I}$$

From: Simonyan et al., "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps", 2013

Example Images



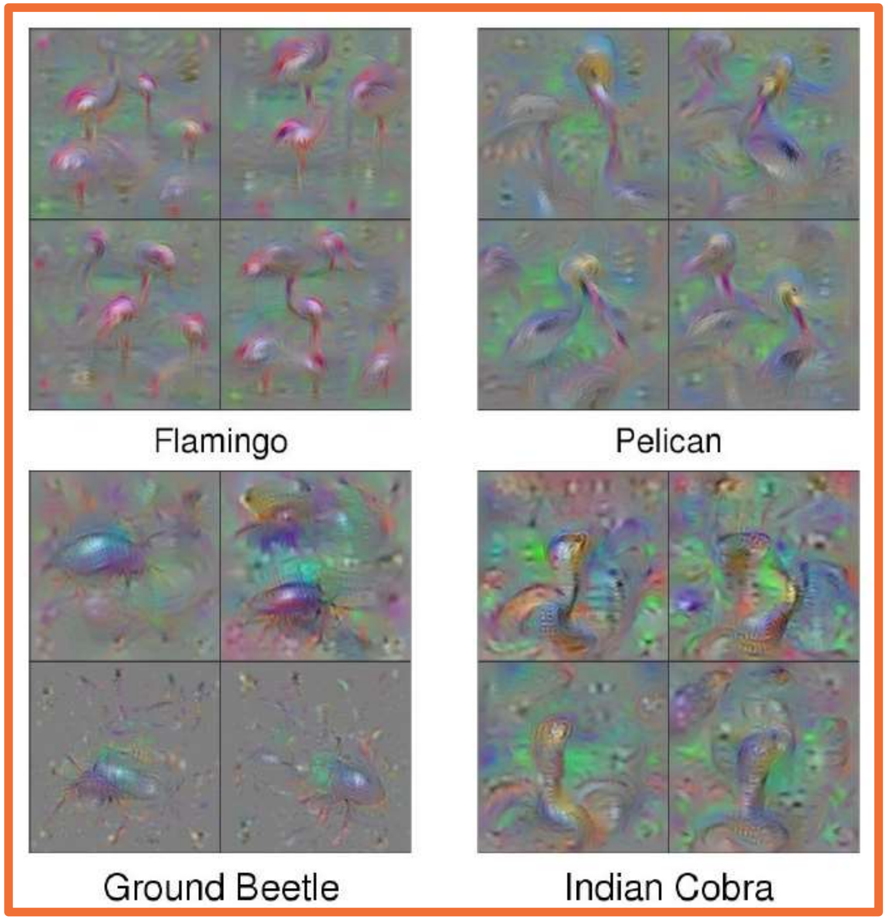
dumbbell   cup   dalmatian

Note: You might have to squint!

From: Simonyan et al., "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps", 2013

Can improve results with **various tricks**:

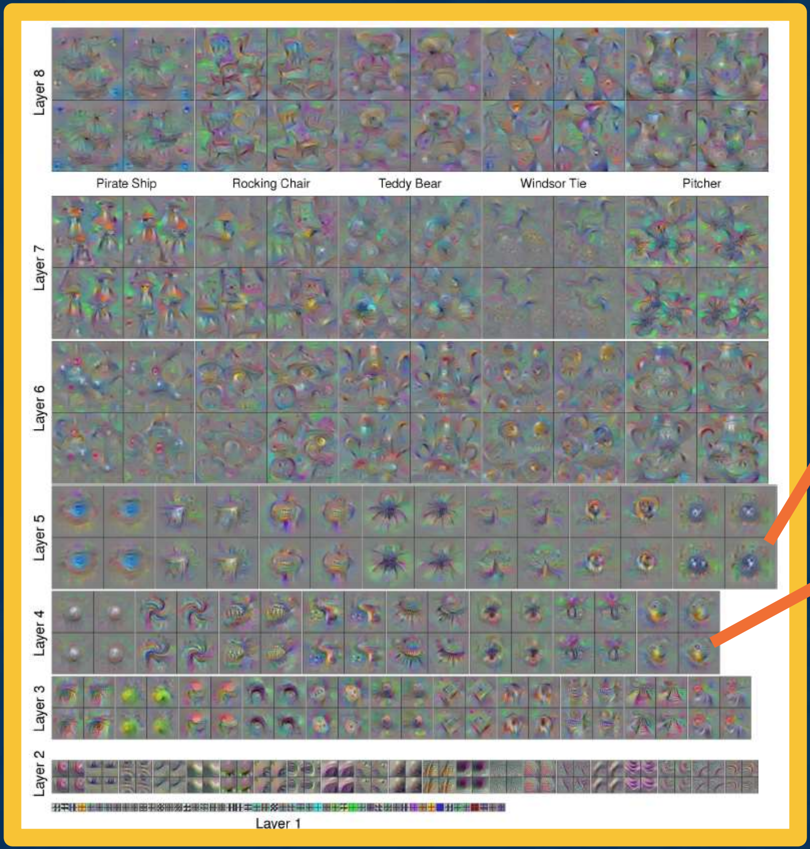
- Clipping or normalization of small values & gradients
- Gaussian blurring



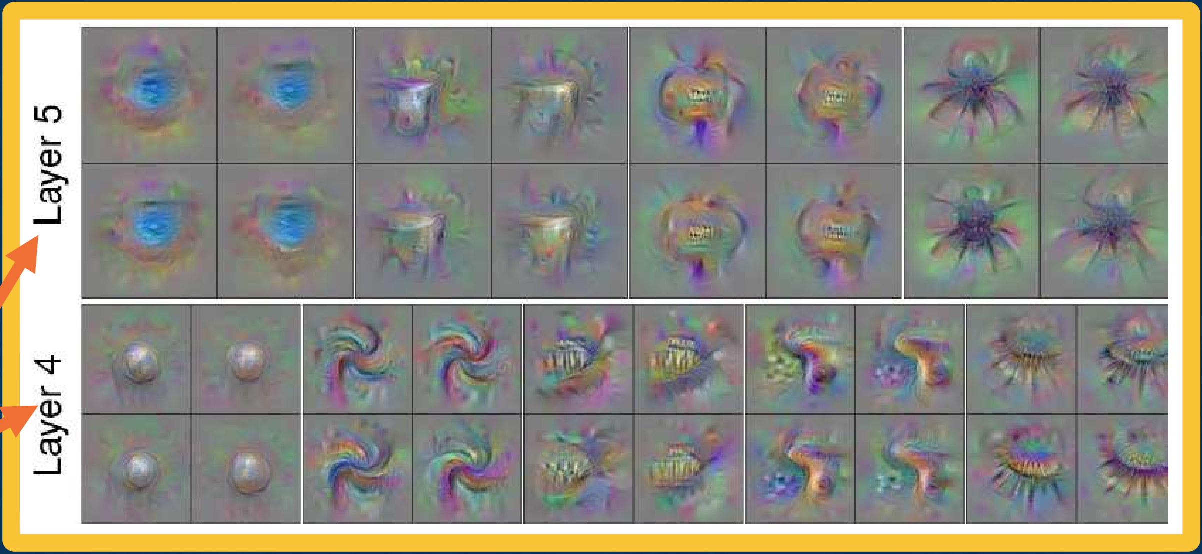
From: Yosinski et al., "Understanding Neural Networks Through Deep Visualization", 2015

Example Images

Improved Results



Note: Can generate input images to maximize any arbitrary activation!



From: Yosinski et al., "Understanding Neural Networks Through Deep Visualization", 2015

Georgia Tech

Summary

We can optimize the input image to **generate** examples to increase class scores or activations

This can show us a great deal about what examples (not in the training set) **activate the network**

Testing Robustness

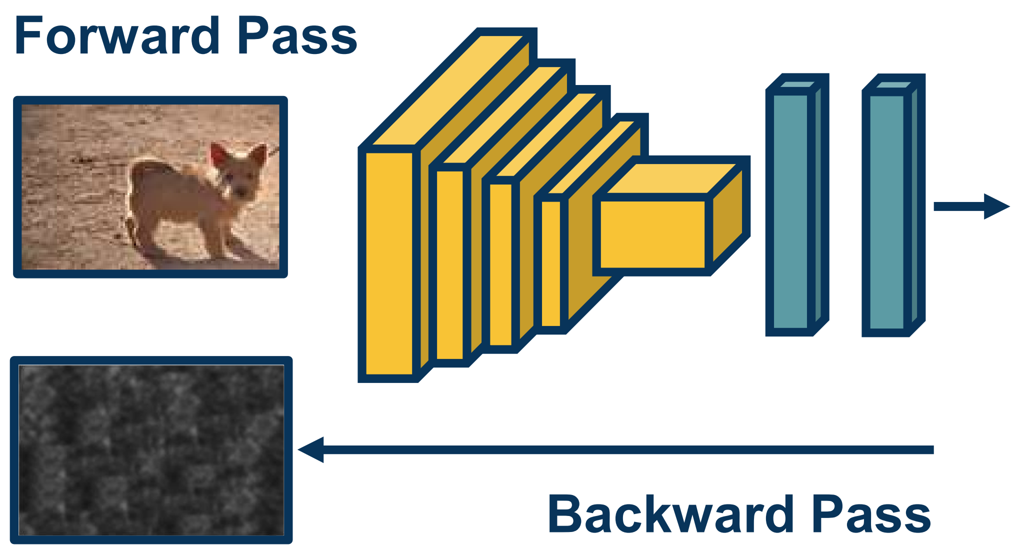
Gradient Ascent on the Scores

- We can perform **gradient ascent** on image
- Rather than start from zero image, why not real image?
- And why not optimize the score of an **arbitrary** (incorrect!) class

Surprising result: You need very small amount of pixel changes to make the network confidently wrong!

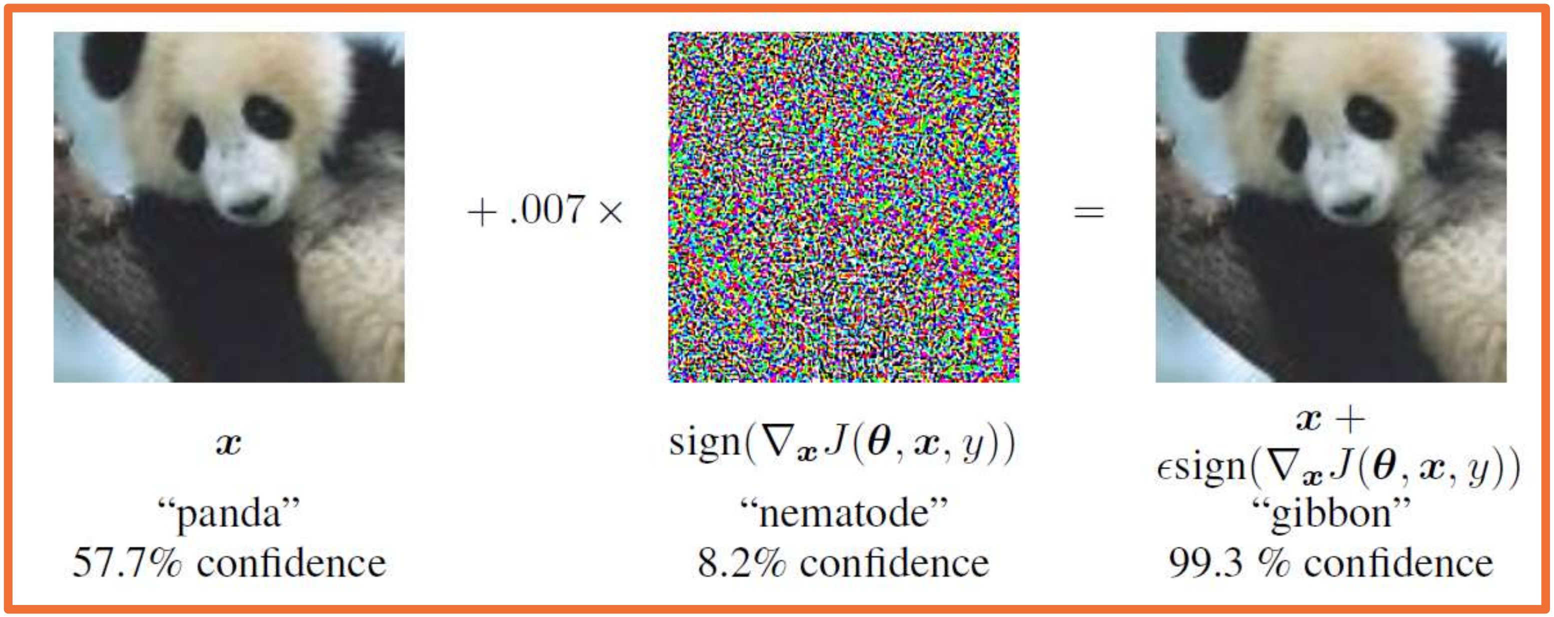
$$\boldsymbol{argmax}\ \boldsymbol{S_c(I)} - \boldsymbol{\lambda} \left|\left|\boldsymbol{I}\right|\right|_2^2$$

where $\boldsymbol{c} = \boldsymbol{cat}$



From: Simonyan et al., "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps", 2013

Example of Adversarial Noise

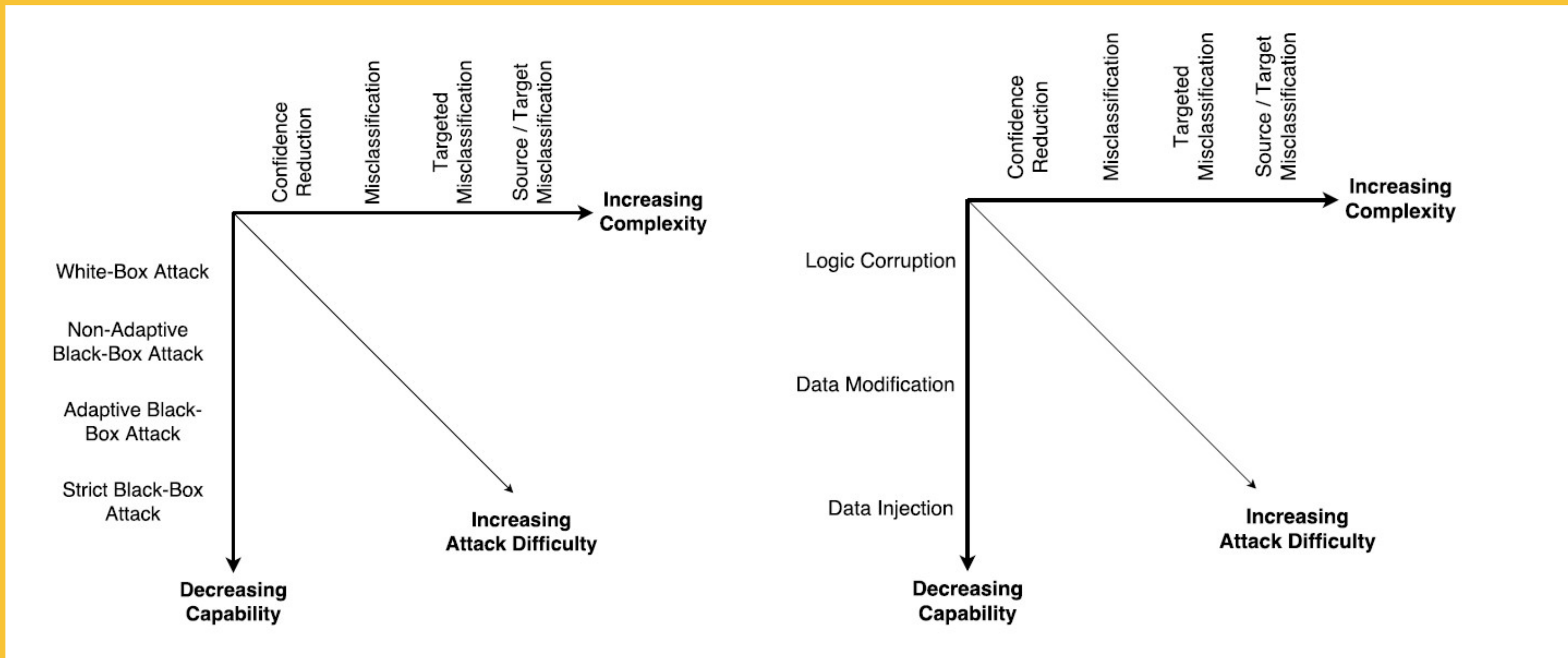


Note this problem is not specific to deep learning!

- Other methods also suffer from it
- Can show how **linearity** (even at the end) can bring this about
  - Can add many small values that add up in right direction

From: Goodfellow et al., "Explaining and Harnessing Adversarial Examples", 2015

Variations of Attacks



Single-Pixel Attacks!

Su et al., "One Pixel Attack for Fooling Deep Neural Networks", 2019.

White vs. Black-Box Attacks of Increasing Complexity

Chakraborty et al., Adversarial Attacks and Defences: A Survey, 2018

Georgia Tech

Summary of dversarial Attacks/Defenses

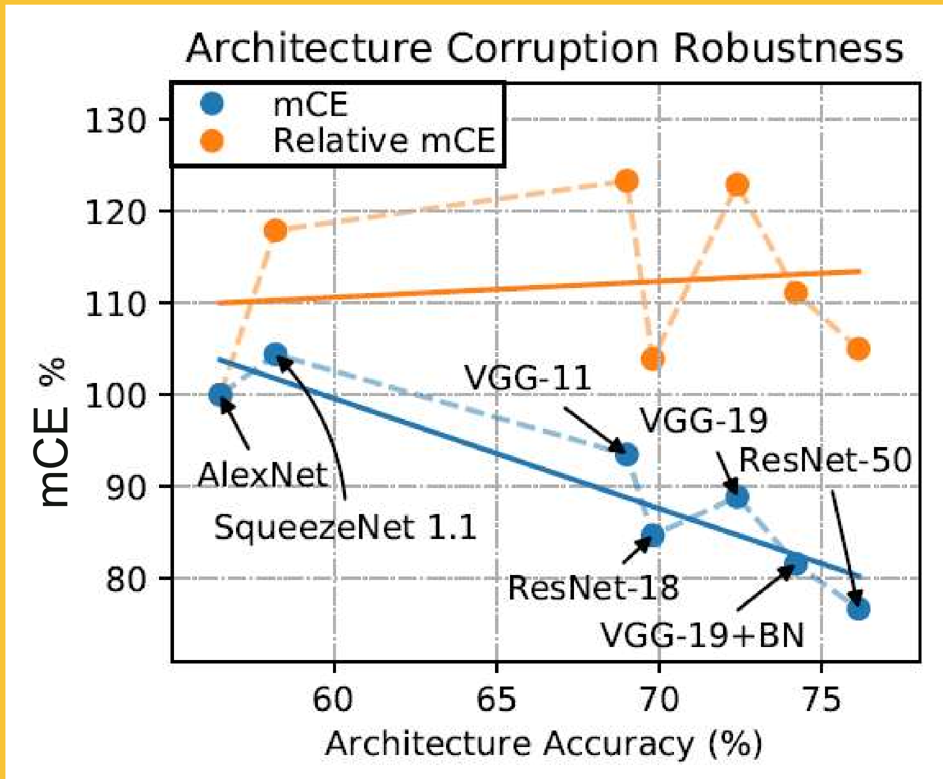
Similar to other security-related areas, it's an active **cat-and-mouse game**

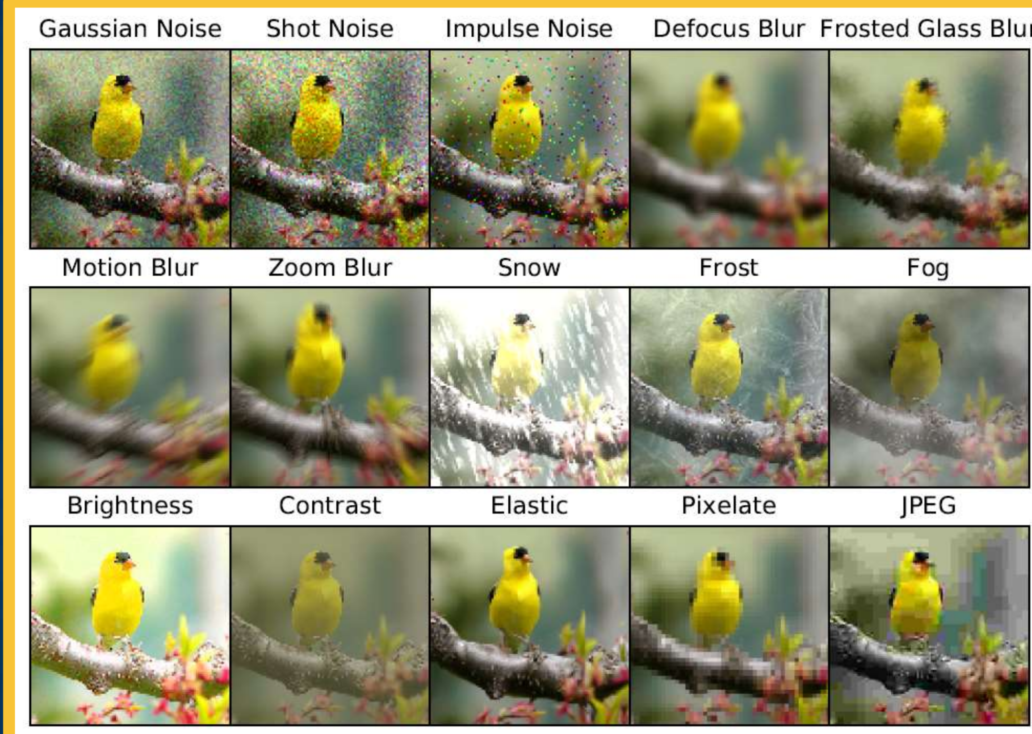
Several defenses such as:

- Training with adversarial examples
- Perturbations, noise, or re-encoding of inputs

There are **not universal methods** that are robust to all types of attacks

Other Forms of Robustness Testing





$$\mathrm{CE}_c^f = \left(\sum_{s=1}^{5} E_{s,c}^f\right) \Big/ \left(\sum_{s=1}^{5} E_{s,c}^{\mathrm{AlexNet}}\right).$$

Hendrycks & Dietterich, "Benchmarking Neural Network Robustness to Common Corruptions and Perturbations", 2019.

We can try to understand the **biases of CNNs**

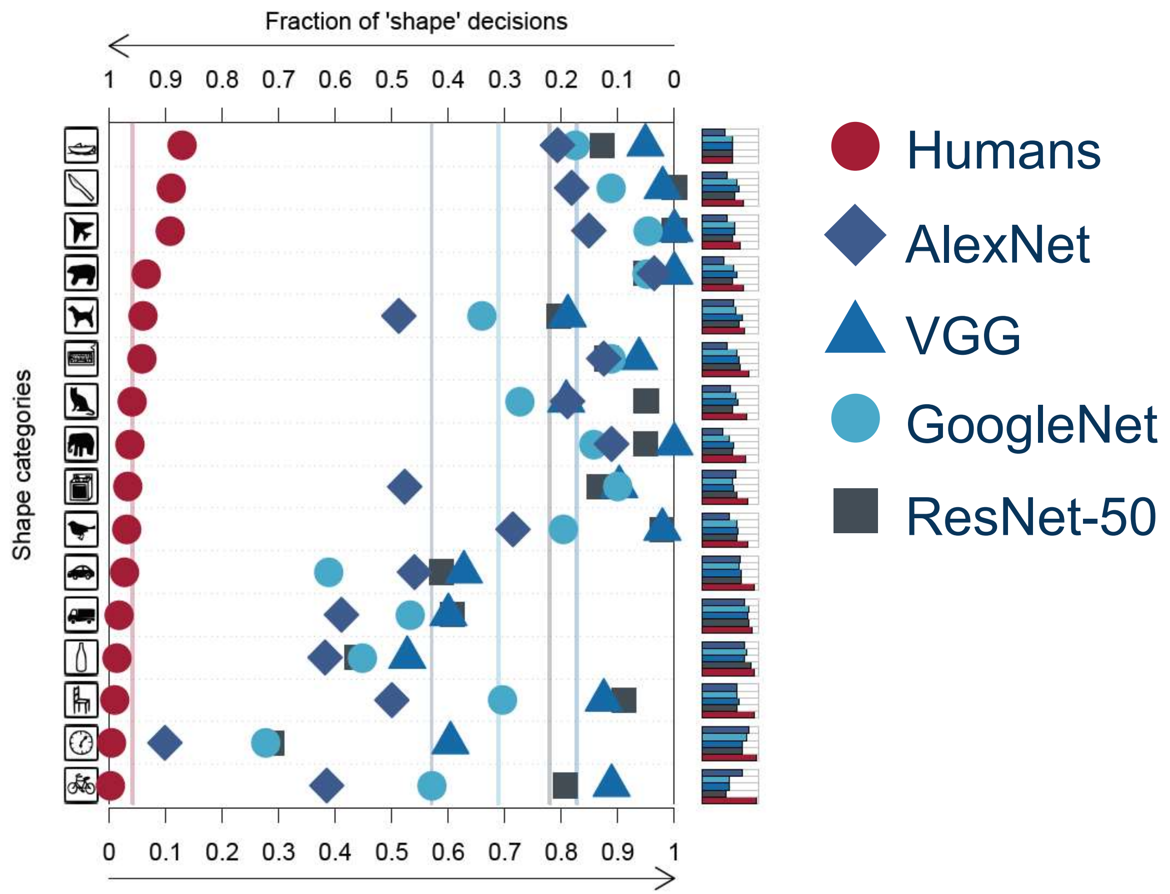
- Can compare to those of humans

Example: Shape vs. Texture Bias

Geirhos, "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness", 2018.

(a) Texture image		(b) Content image		(c) Texture-shape cue conflict	
81.4%	**Indian elephant**	71.1%	**tabby cat**	63.9%	**Indian elephant**
10.3%	indri	17.3%	grey fox	26.4%	indri
8.2%	black swan	3.3%	Siamese cat	9.6%	black swan

Shape vs. Texture Bias



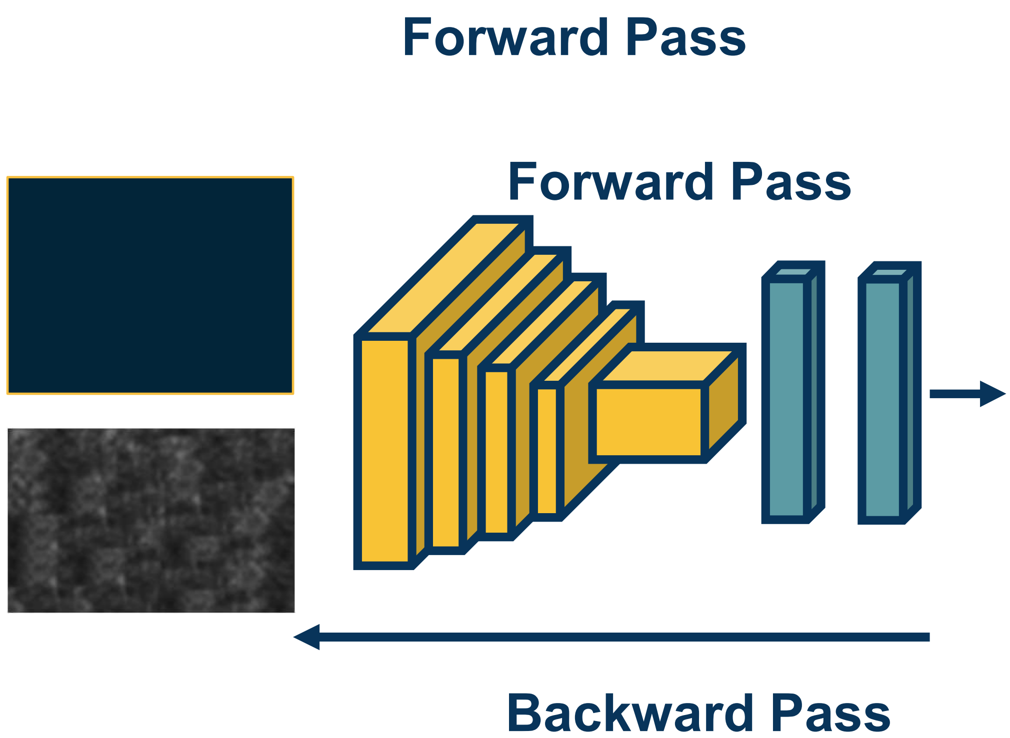
Geirhos, "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness", 2018.

Summary

- Various ways to test the **robustness** and **biases** of neural networks
- Adversarial examples have **implications** for understanding and trusting them
- Exploring the **gain of different architectures** in terms of robustness and biases can also be used to understand what has been learned

Style Transfer

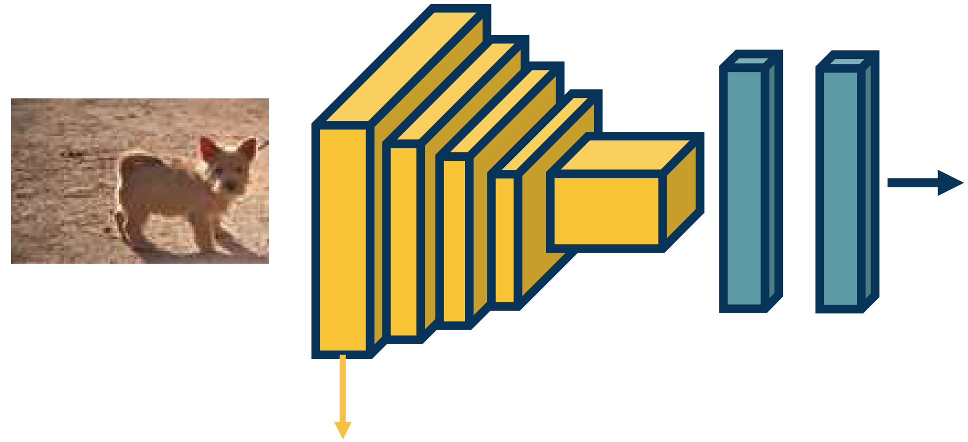
- We can generate images through backprop
  - Regularization can be used to ensure we match image statistics
- **Idea:** What if we want to preserve the content of the image?
  - Match features at different layers!
  - We can have a loss for this



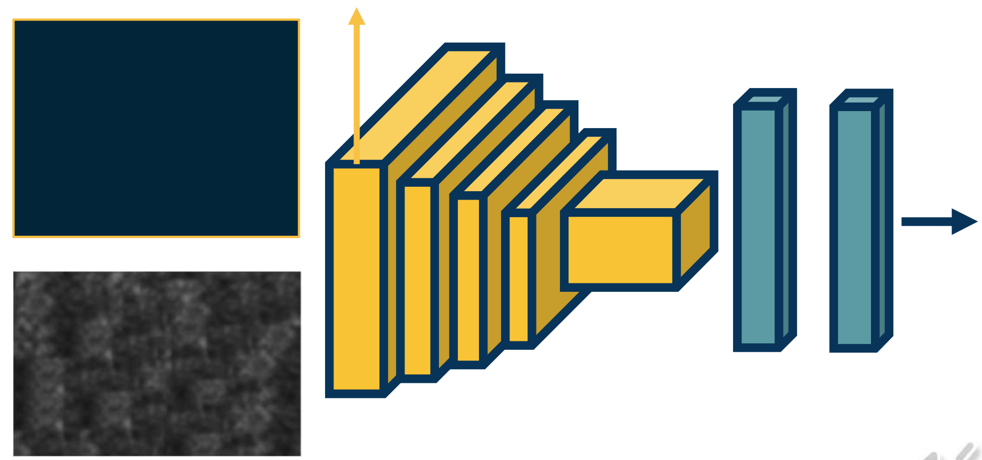
Generating Images with Content

Matching Features to Replicate Content

- We can generate images through backprop
  - Regularization can be used to ensure we match image statistics
- **Idea:** What if we want to preserve the content of a particular image C?
  - Match features at different layers!
  - We can have a loss for this



$$L_{content} = (F_C^1 - F_P^1)^2$$

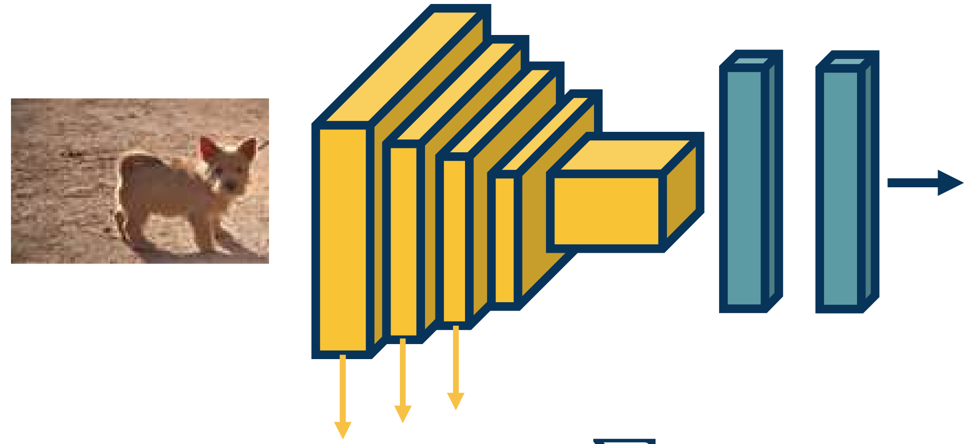


Georgia Tech

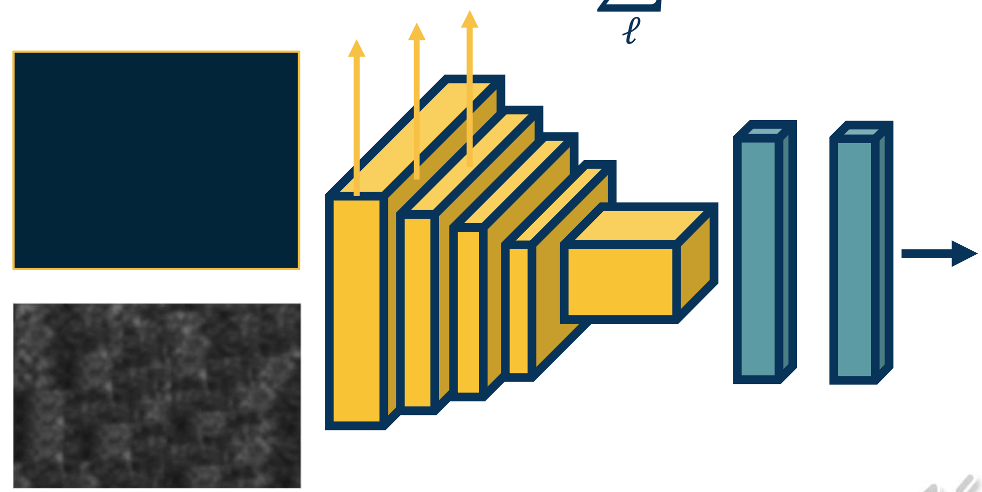
Multiple Content Losses

- How do we deal with multiple losses?
  - Remember, backwards edges going to same node *summed*

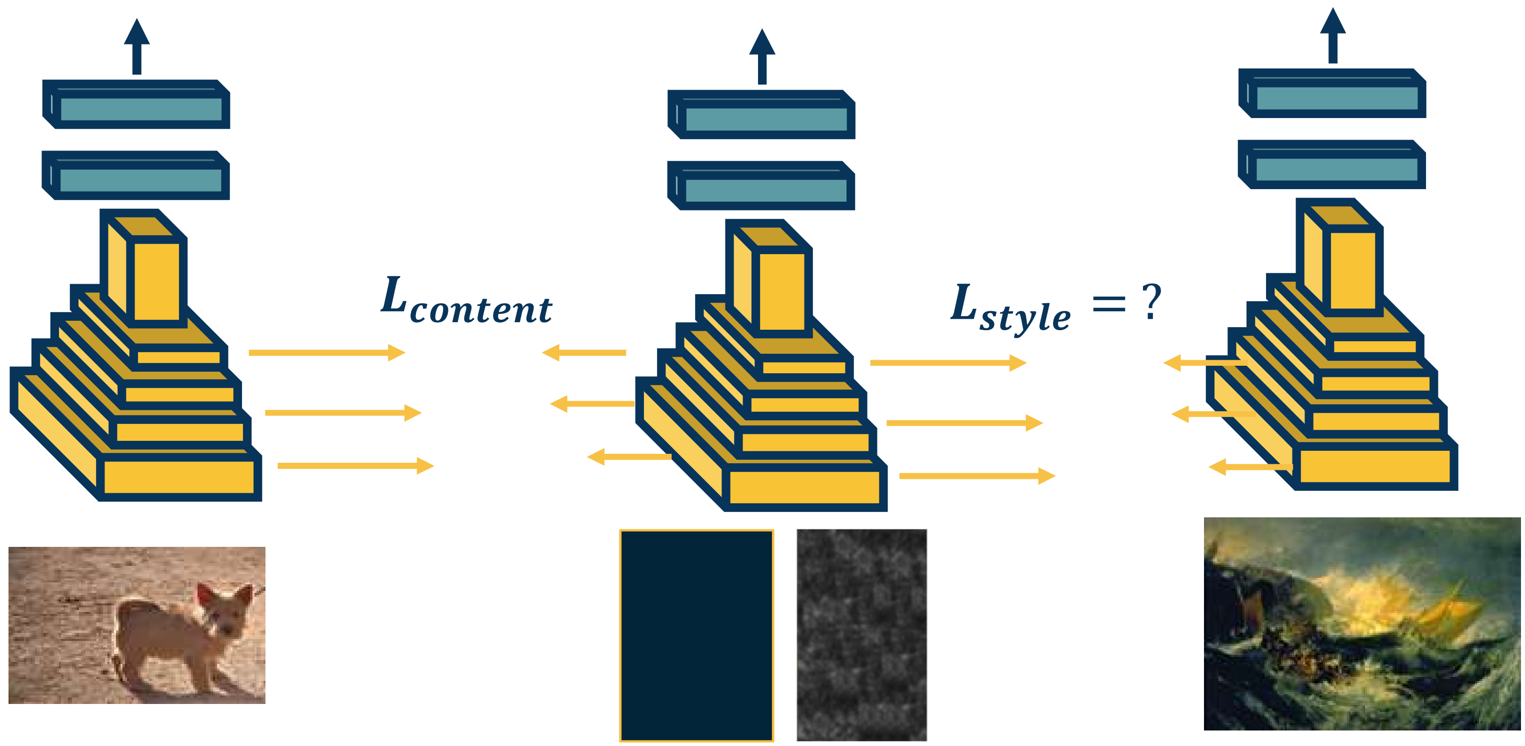
- We can have this content loss at many different layers and sum them too!



$$L_{content} = \sum_{\ell} (F_C^{\ell} - F_P^{\ell})^2$$



- **Idea:** Can we have the *content* of one image and *texture* (style) of another image?
  - **Yes!**

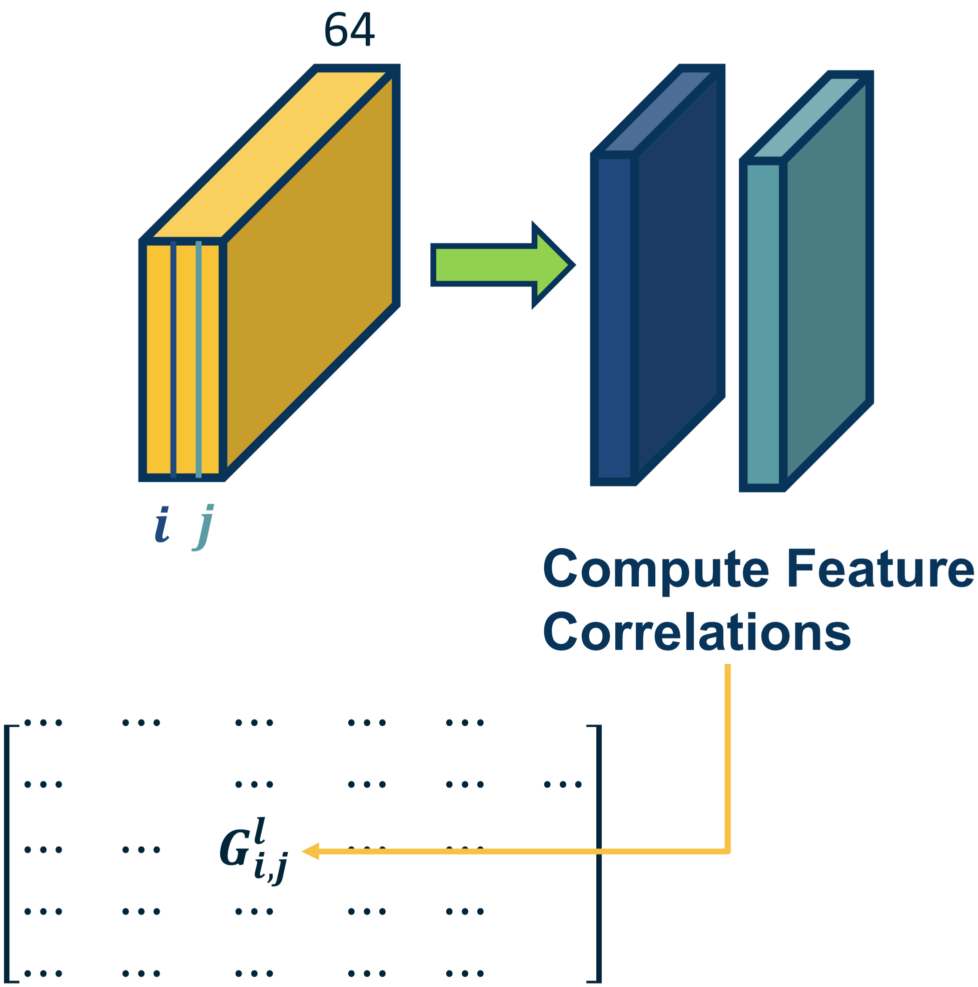


Replicating Content and Style

- How do we represent similarity in terms of textures?

- Long history in image processing!
  - Key ideas revolve around summary *statistics*
  - Should ideally remove most spatial information

- Deep learning variant: Feature correlations!
  - Called a Gram Matrix



$$G_S^{\ell}(i,j) = \sum_k F_S^{\ell}(i,k) F_S^{\ell}(j,k)$$

where i,j are particular **channels** in the output map of layer ℓ and k is the position (convert the map to a vector)

$$L_{style} = \sum_{\ell} \left(G_S^{\ell} - G_P^{\ell}\right)^2$$

$$L_{total} = \alpha L_{content} + \beta L_{style}$$

Gradient Ascent on the Scores

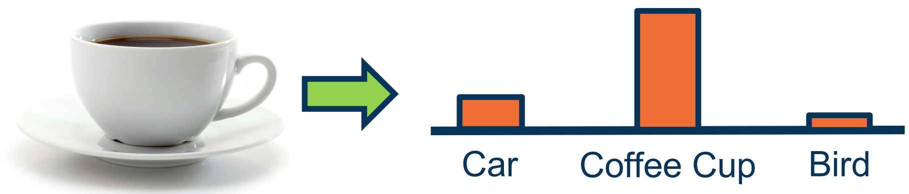
Gradient Ascent on the Scores

Gradient Ascent on the Scores

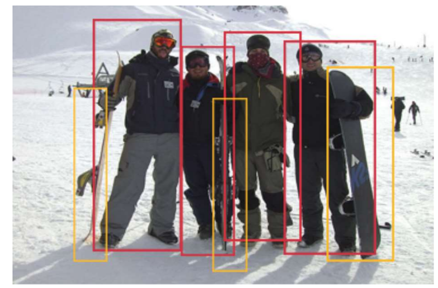
Summary

- Generating images through optimization is a powerful concept!
- Besides fun and art, methods such as stylization also useful for understanding what the network has learned
- Also useful for other things such as data augmentation

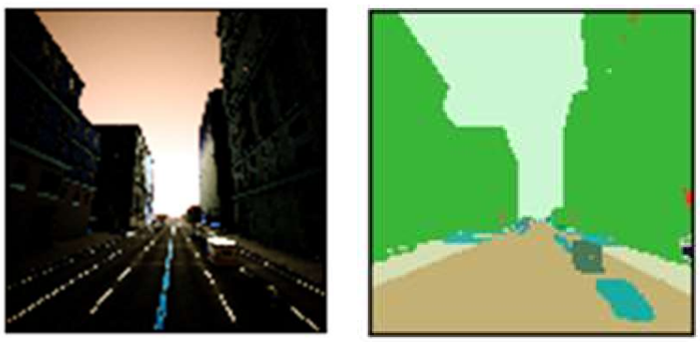
Image Segmentation Networks



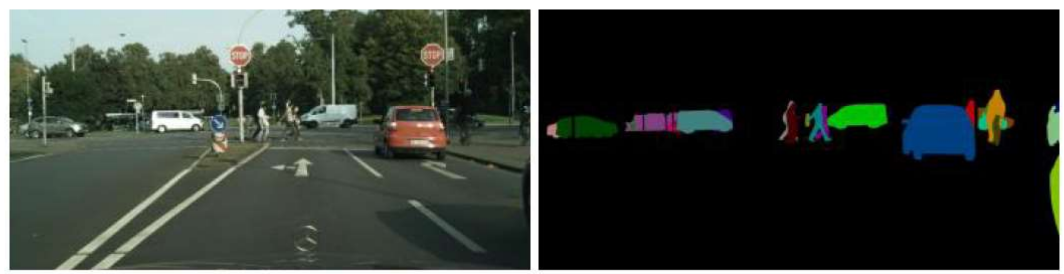
Classification
(Class distribution per image)



Object Detection
(List of bounding boxes with class distribution per box)



Semantic Segmentation
(Class distribution per pixel)



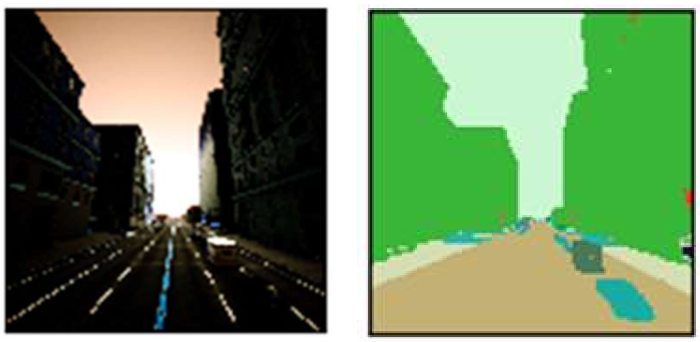
Instance Segmentation
(Class distribution per pixel with unique ID)

Computer Vision Tasks

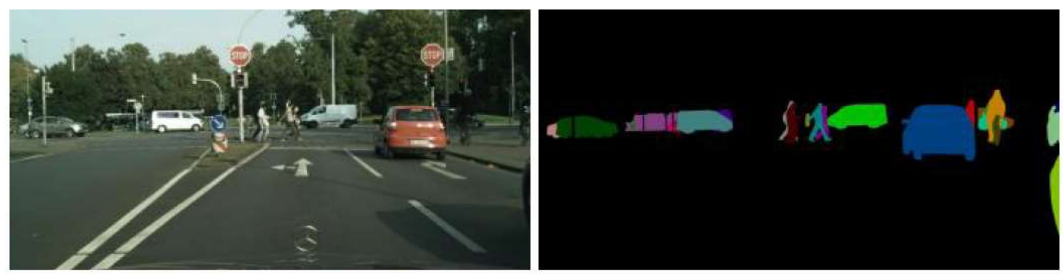
Segmentation Tasks

Given an image, output another image

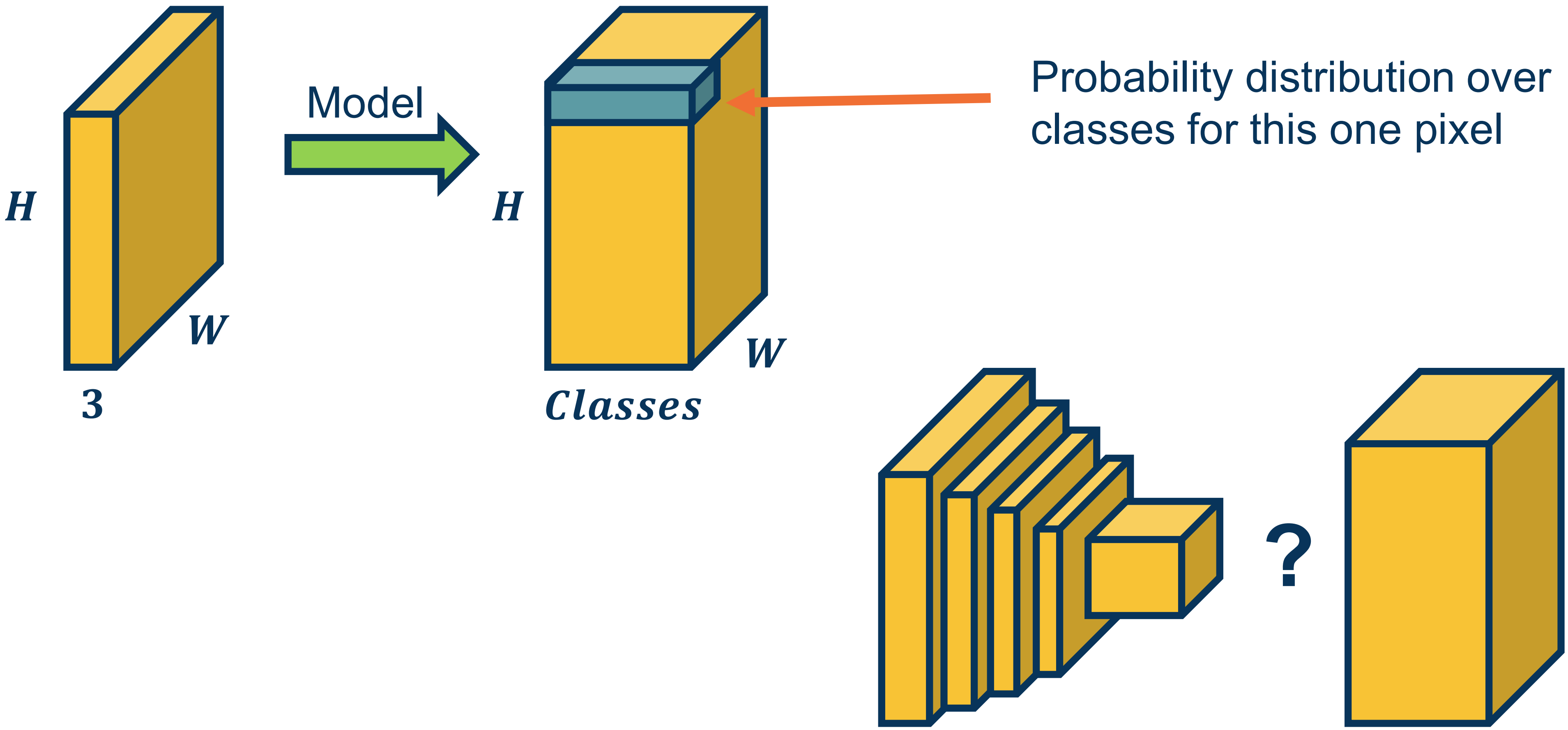
- Each output contains class distribution per pixel
- More generally an image-to-image problem



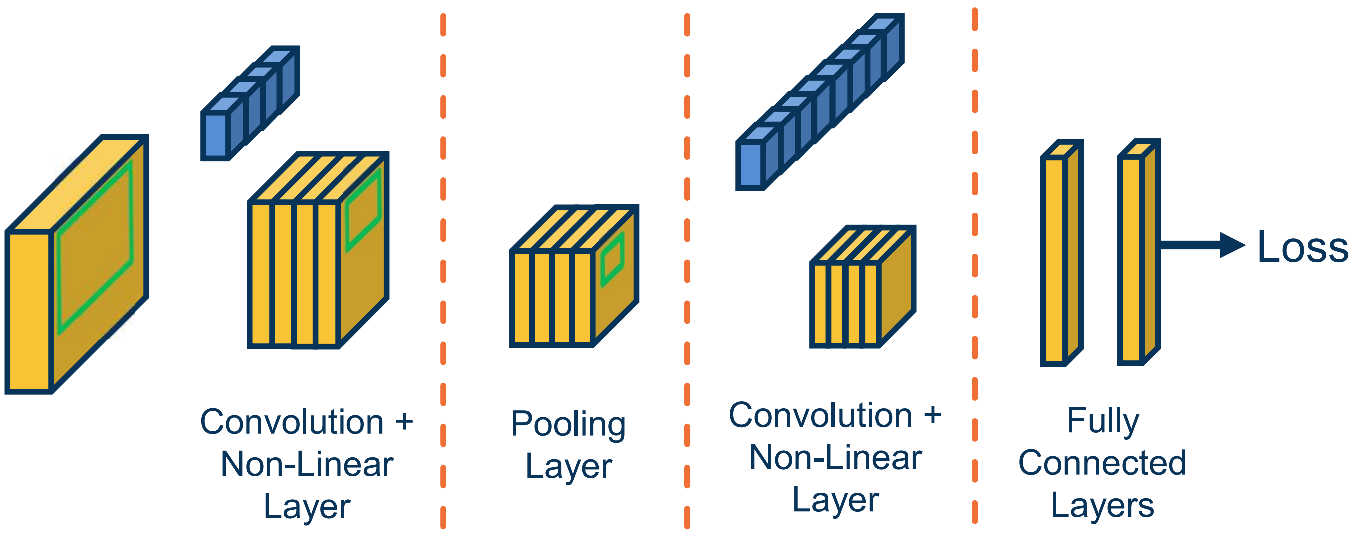
Semantic Segmentation
(Class distribution per pixel)



Instance Segmentation
(Class distribution per pixel with unique ID)



Input & Output

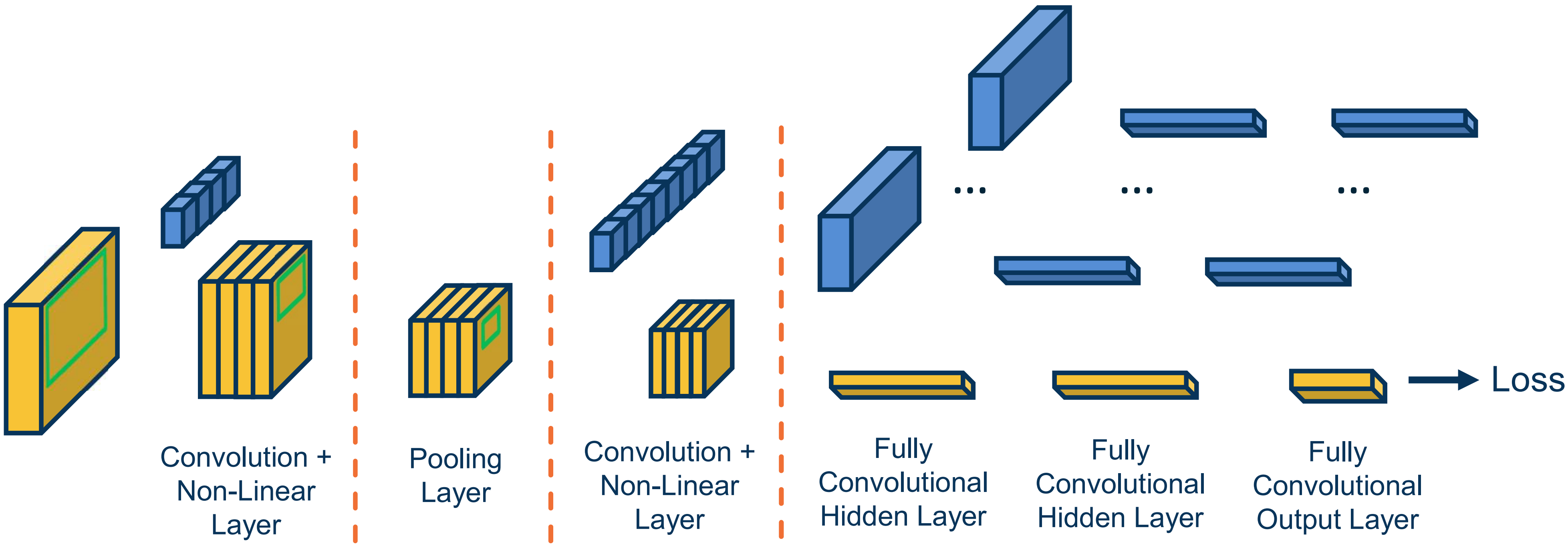


Fully connected layers no longer explicitly retain spatial information (though the network can still learn to do so)

Idea: Convert fully connected layer to convolution!

Idea 1: Fully-Convolutional Network

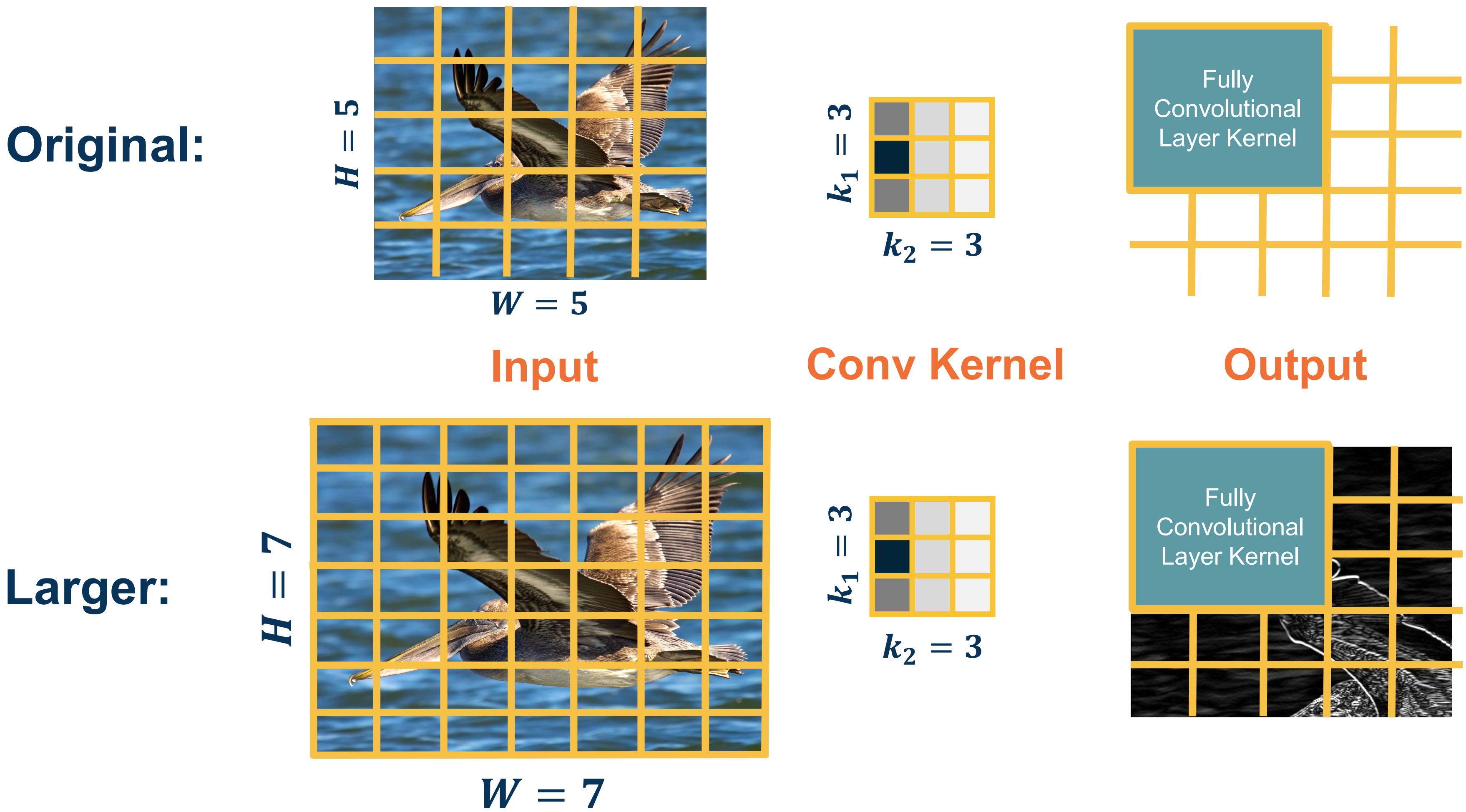
Georgia Tech



Each kernel has the size of entire input! (output is 1 scalar)

- This is equivalent to Wx+b!
- We have one kernel per output node

Converting FC Layers to Conv Layers



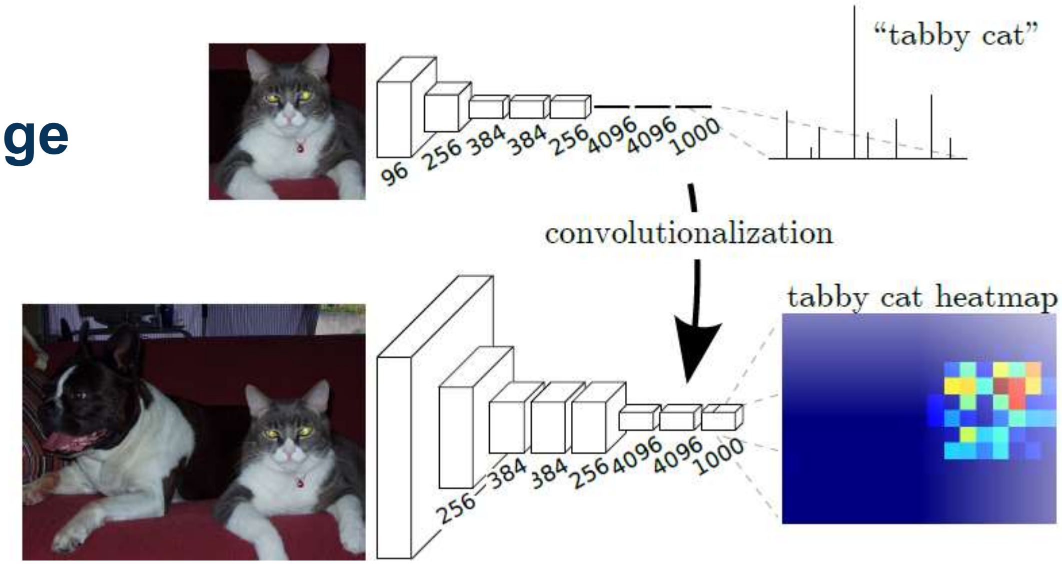
Same Kernel, Larger Input

Why does this matter?

- We can stride the “fully connected” classifier across larger inputs!
- Convolutions work on arbitrary input sizes (because of striding)

Original sized image

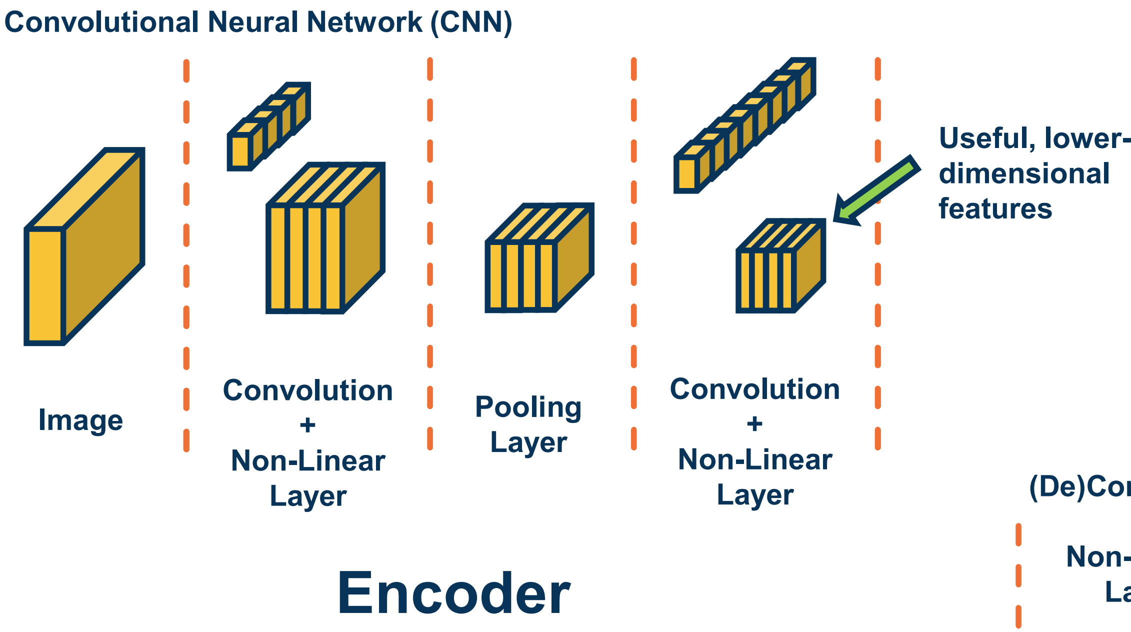
Larger Image



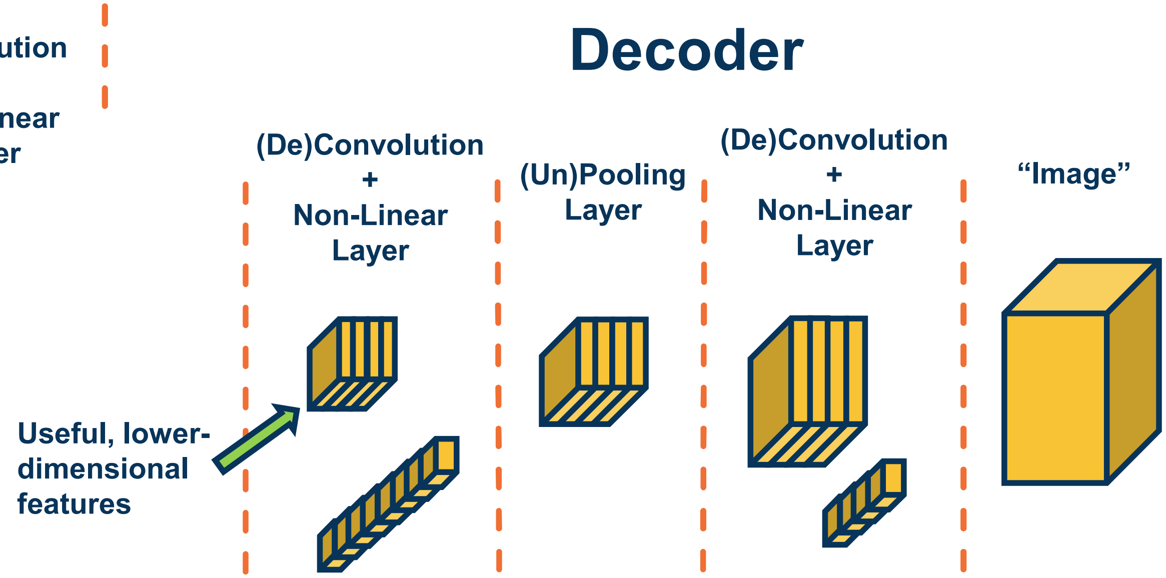
Larger Output Size!

Larger Output Maps

Long, et al., “Fully Convolutional Networks for Semantic Segmentation”, 2015



We can develop learnable or non-learnable upsampling layers!



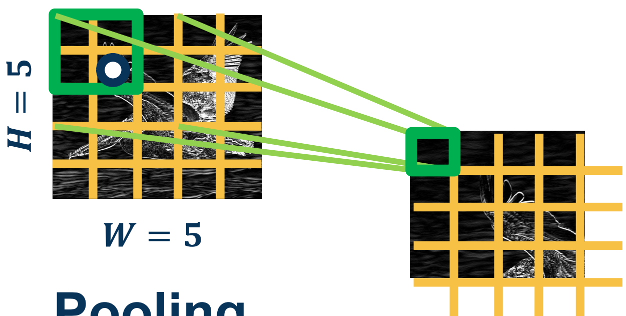
Idea 2: “De”Convolution and UnPooling

Max Unpooling

Example : Max pooling

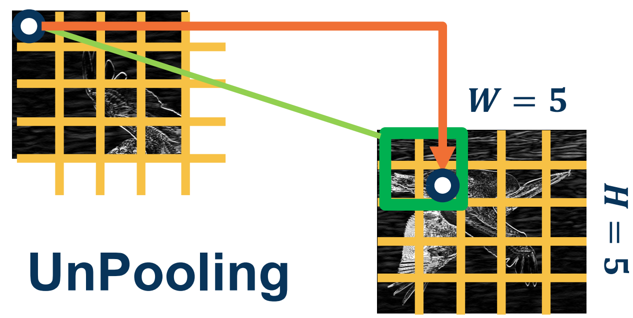
- Stride window across image but perform per-patch **max operation**

$$X(0:1,0:1) = \begin{bmatrix} 100 & 150 \\ 100 & 200 \end{bmatrix} \Rightarrow \max(0:1,0:1) = 200$$



Pooling

Copy value to position chosen as max in encoder, fill reset of this window with zeros



UnPooling

Idea: Remember max elements in encoder! Copy value from equivalent position, rest are zeros

Georgia Tech

$$X = \begin{bmatrix} 120 & 150 & 120 \\ 100 & 50 & 110 \\ 25 & 25 & 10 \end{bmatrix} \xrightarrow[\text{2x2 max pool}]{} Y = \begin{bmatrix} 150 & 150 \\ 100 & 110 \end{bmatrix}$$

Encoder

Decoder

$$X = \begin{bmatrix} 300 & 450 \\ 100 & 250 \end{bmatrix} \xrightarrow{\text{2x2 max unpool}} Y = \begin{bmatrix} 0 & 300 & - \\ 0 & 0 & - \\ - & - & - \end{bmatrix}$$

Max Unpooling Example (one window)

Max Unpooling Example

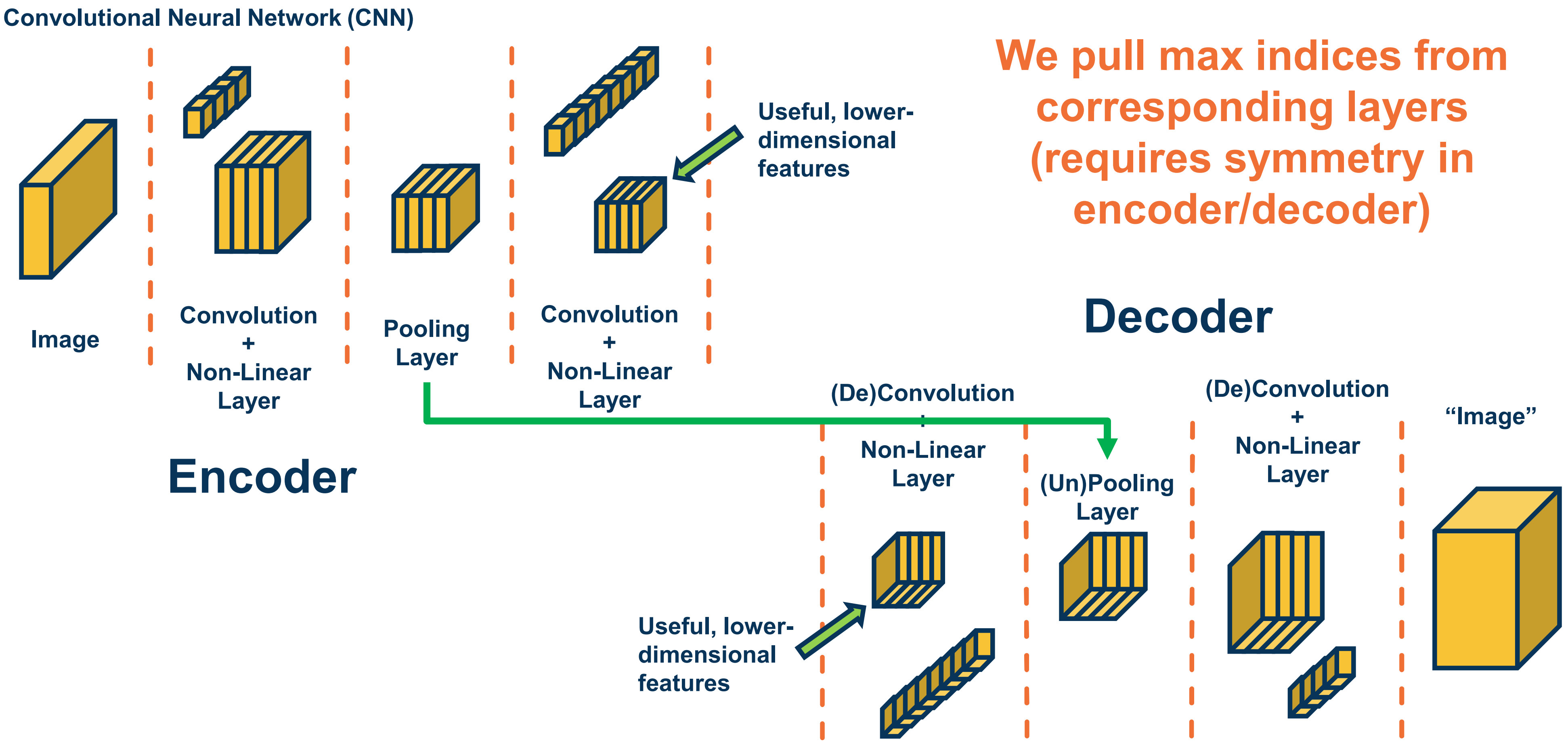
$$\mathbf{X}_{\mathbf{enc}} = \begin{bmatrix} 120 & 150 & 120 \\ 100 & 50 & 110 \\ 25 & 25 & 10 \end{bmatrix} \xrightarrow{\text{2x2 max pool}} \mathbf{Y}_{\mathbf{enc}} = \begin{bmatrix} 150 & 150 \\ 100 & 110 \end{bmatrix}$$

Encoder

Contributions from multiple windows are summed

Decoder

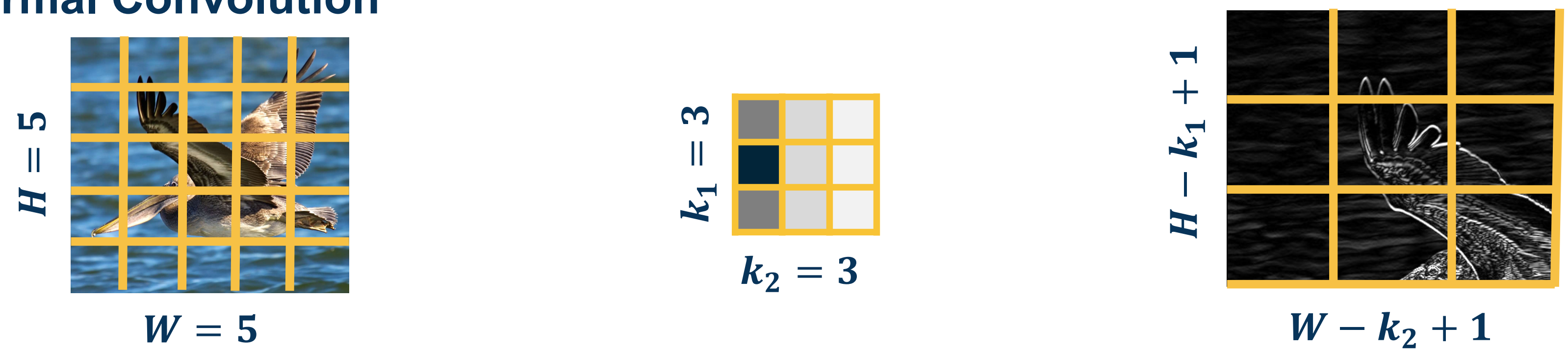
$$\mathbf{X}_{\mathbf{dec}} = \begin{bmatrix} 300 & 450 \\ 100 & 250 \end{bmatrix} \xrightarrow{\text{2x2 max unpool}} \mathbf{Y}_{\mathbf{dec}} = \begin{bmatrix} 0 & 300 + 450 & 0 \\ 100 & 0 & 250 \\ 0 & 0 & 0 \end{bmatrix}$$



Symmetry in Encoder/Decoder

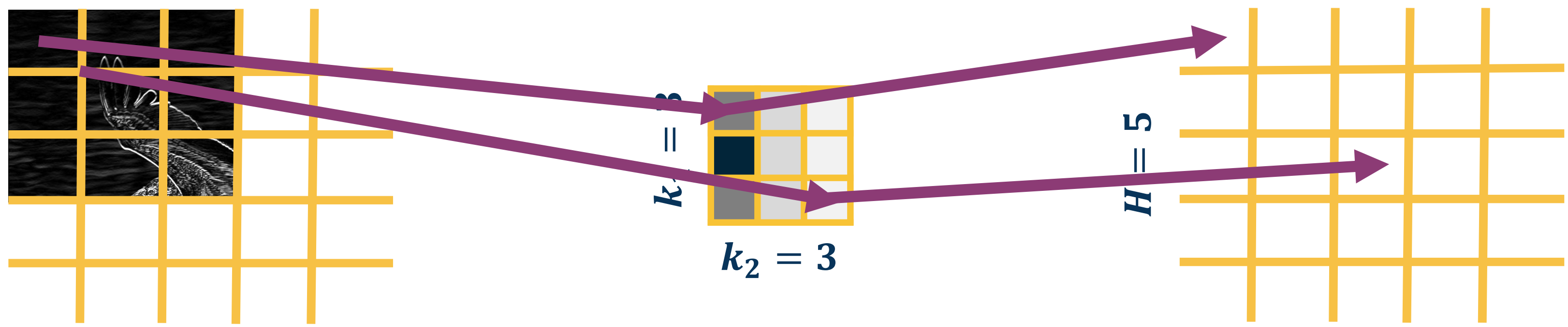
How can we *upsample* using convolutions and learnable kernel?

Normal Convolution



Transposed Convolution (also known as "deconvolution", fractionally strided conv)

Idea: Take each input pixel, multiply by learnable kernel, "stamp" it on output



"De"Convolution (Transposed Convolution)

Georgia Tech

$$X = \begin{bmatrix} 120 & 150 & 120 \\ 100 & 50 & 110 \\ 25 & 25 & 10 \end{bmatrix} \qquad K = \begin{bmatrix} 1 & -1 \\ 2 & -2 \end{bmatrix}$$

Contributions from multiple windows are summed

$$\begin{bmatrix} 120 & -120 & 0 & 0 \\ 240 & -240 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

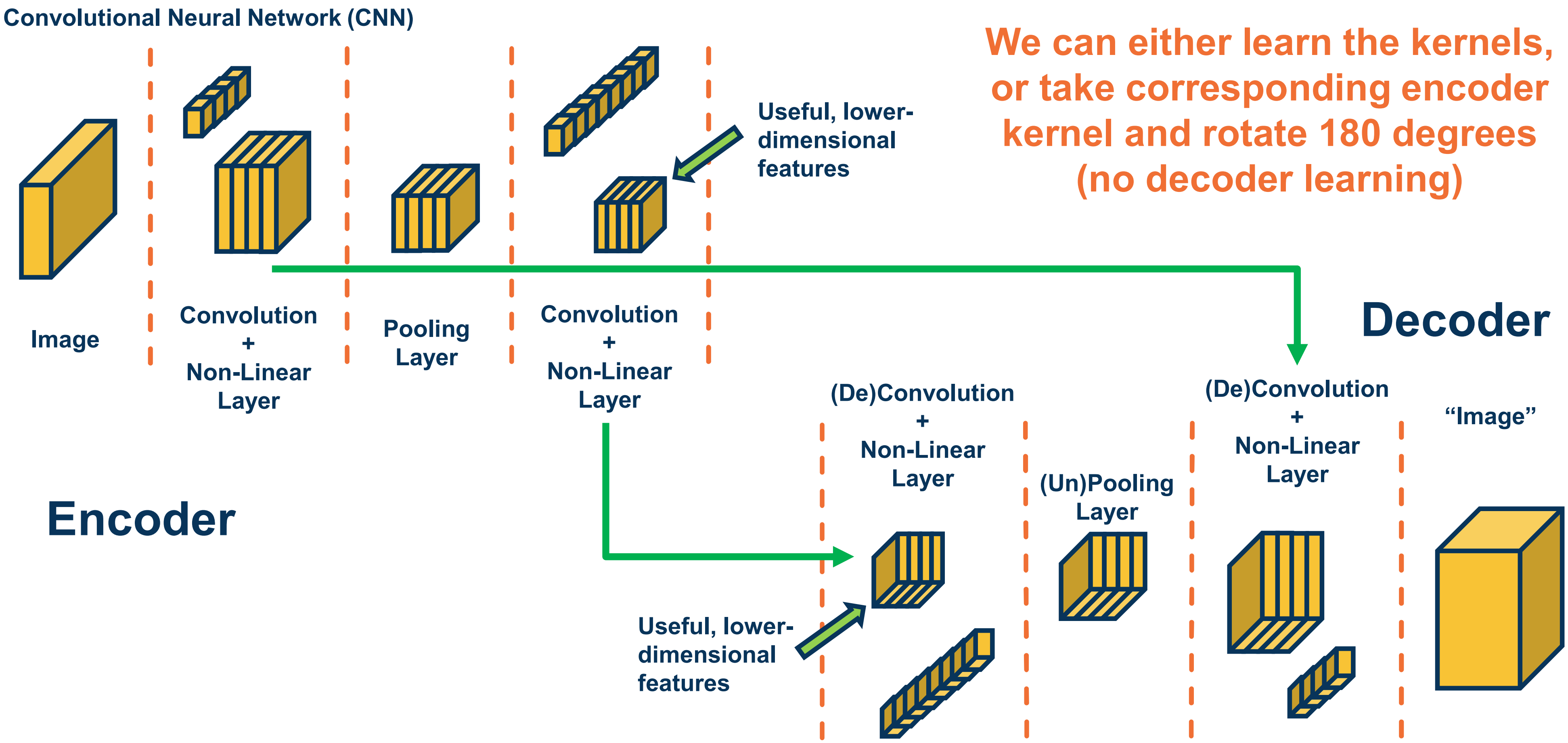
Incorporate $X(0,0)$

$$\begin{bmatrix} 120 & -120+150 & -150 & 0 \\ 240 & -240+300 & -300 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

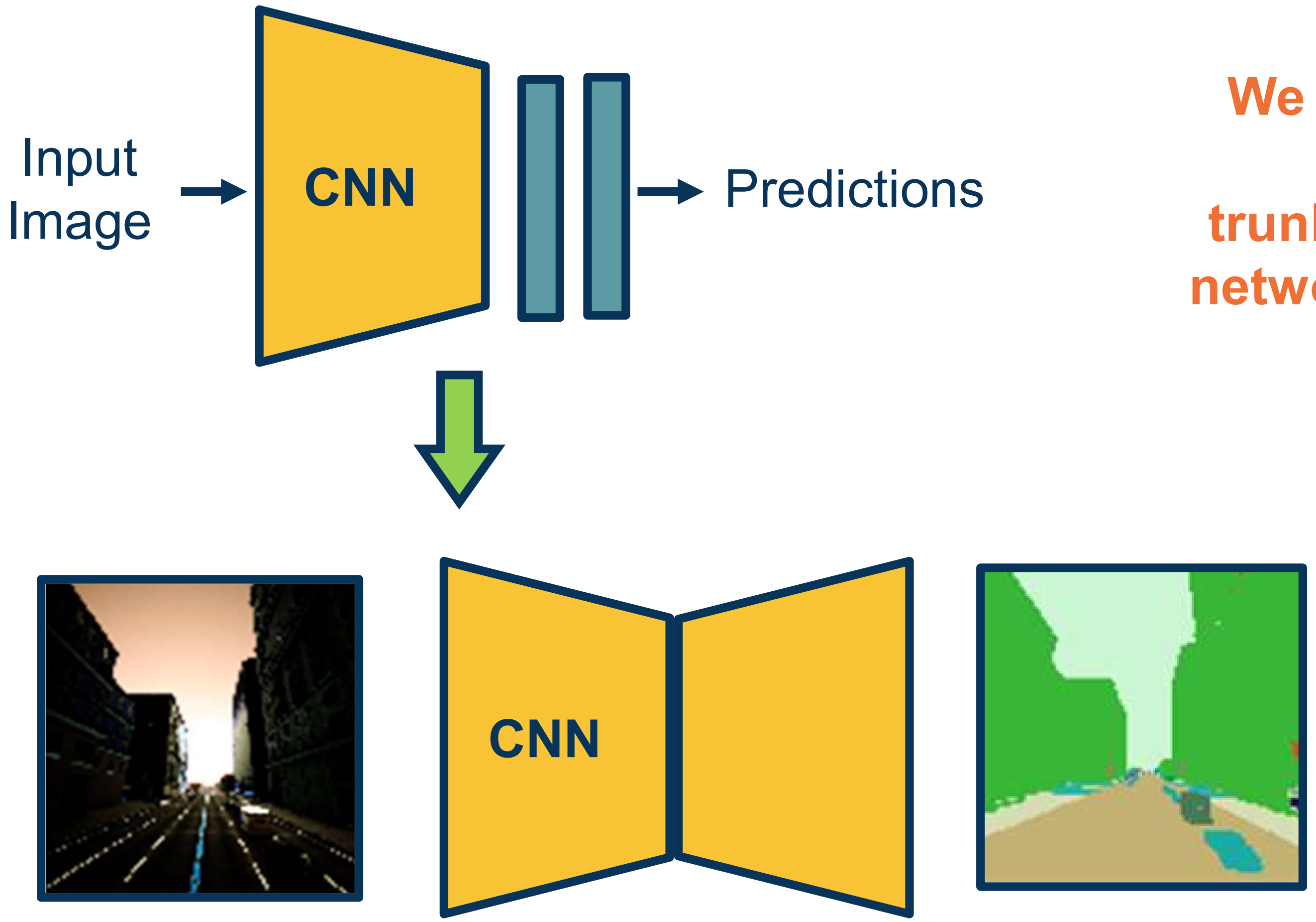
Incorporate $X(1,0)$

Transposed Convolution Example

We can either learn the kernels, or take corresponding encoder kernel and rotate 180 degrees (no decoder learning)



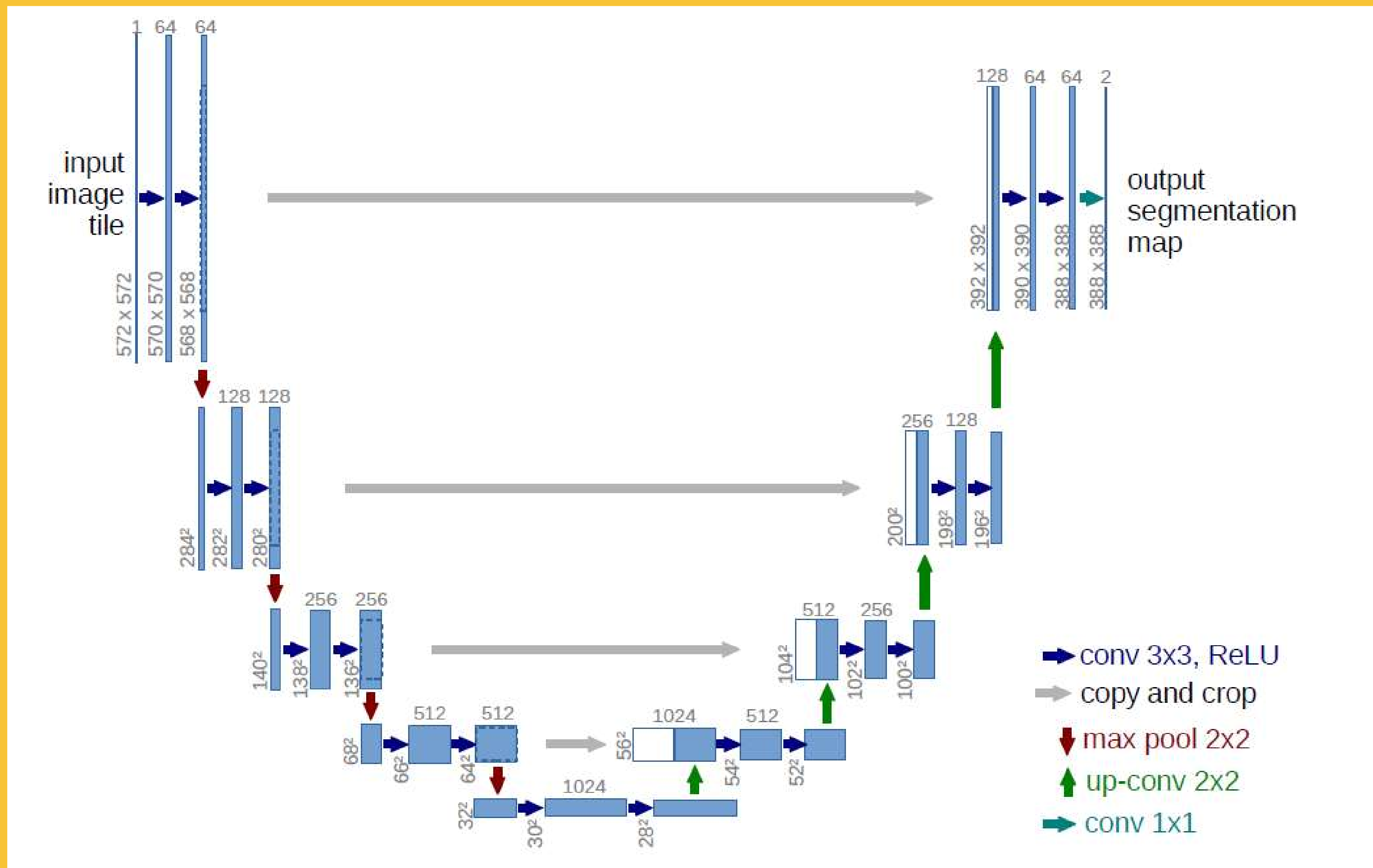
Symmetry in Encoder/Decoder



We can start with a pre-trained trunk/backbone (e.g. network pretrained on ImageNet)!

U-Net

You can have skip connections to bypass bottleneck!



Ronneberger, et al., "U-Net: Convolutional Networks for Biomedical Image Segmentation", 2015

Georgia Tech

Summary

- Various ways to get **image-like outputs,** for example to predict segmentations of input images
- Fully convolutional layers essentially apply the striding idea to the output classifiers, supporting arbitrary input sizes
  - (without output size depending on what the input size is)
- We can have various upsampling layers that actually increase the size
- Encoder/decoder architectures are popular ways to leverage these to perform general image-to-image tasks