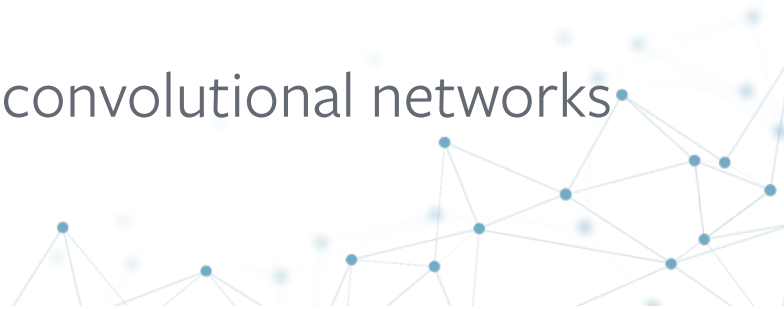
Introduction to Computer Vision: Convolutional Networks (Part 2)

Laurens van der Maaten

Summary from last time

- Loss measures discrepancy between a prediction and the true label
- ERM minimizes the loss averaged over the training data
- Logistic regression is an linear ERM model that uses logistic loss
- Multi-layer networks iteratively apply learned linear functions and fixed non-linear functions: gradients computed by chain rule (backprop)
- Convolutions filter signals and are the basis of convolutional networks



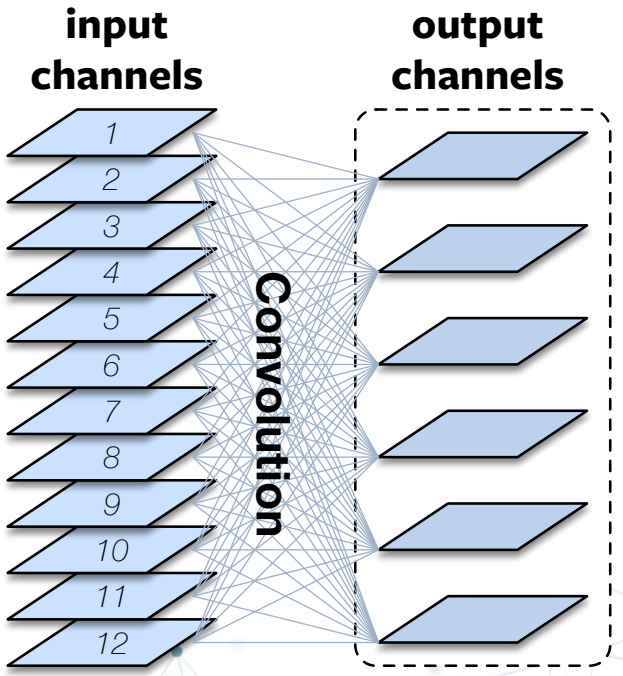
Convolutional networks

- Convolutional network layers generally have **multiple input channels** (for example, RGB) and **multiple output channels**

Convolutional networks

- Convolutional network layers generally have **multiple input channels** (for example, RGB) and **multiple output channels**

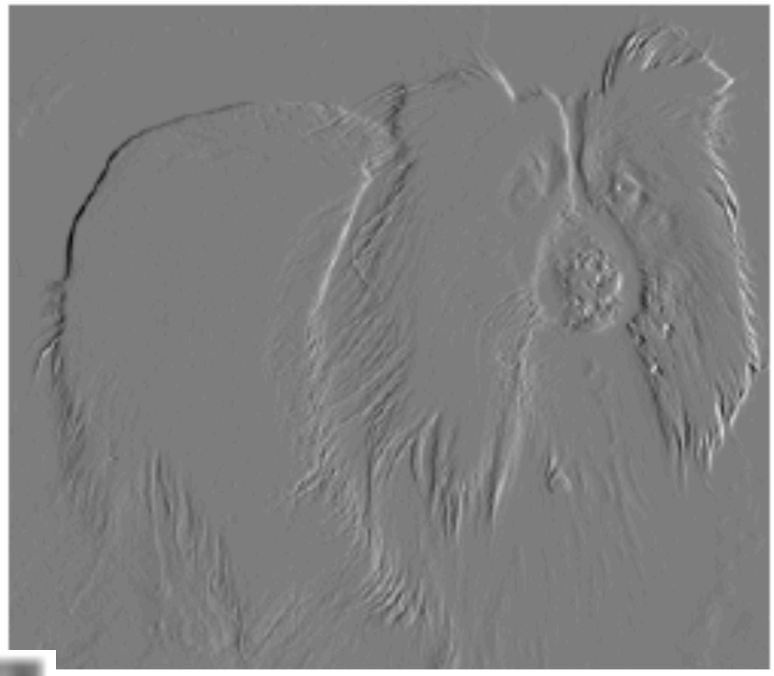
- To produce a single output channel:
  - **Convolve** each input channel with some **kernel**
  - Note that the kernel is **different** for each input channel
  - **Sum** the convolved input channels to produce the output channel



A simple convolutional network layer

* Credits: Ian Goodfellow

A simple convolutional network layer



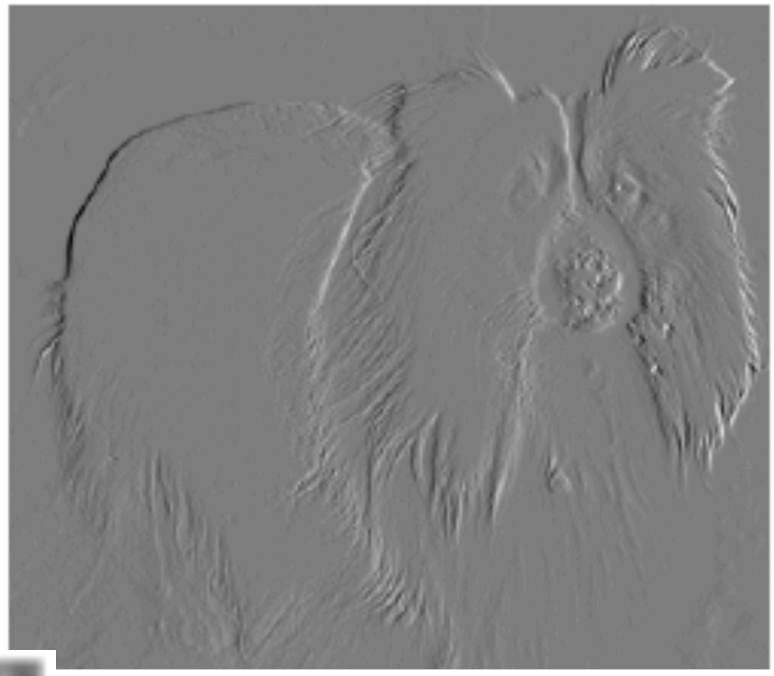
* Credits: Ian Goodfellow

A simple convolutional network layer

Only has 1 input and 1 output!

Input

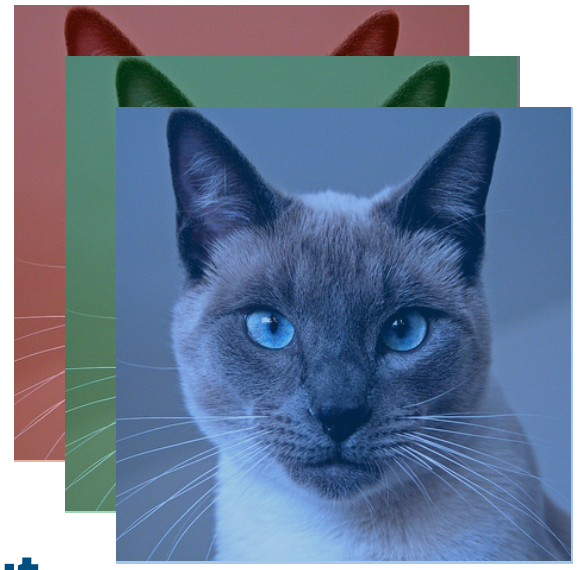
Kernel



Output

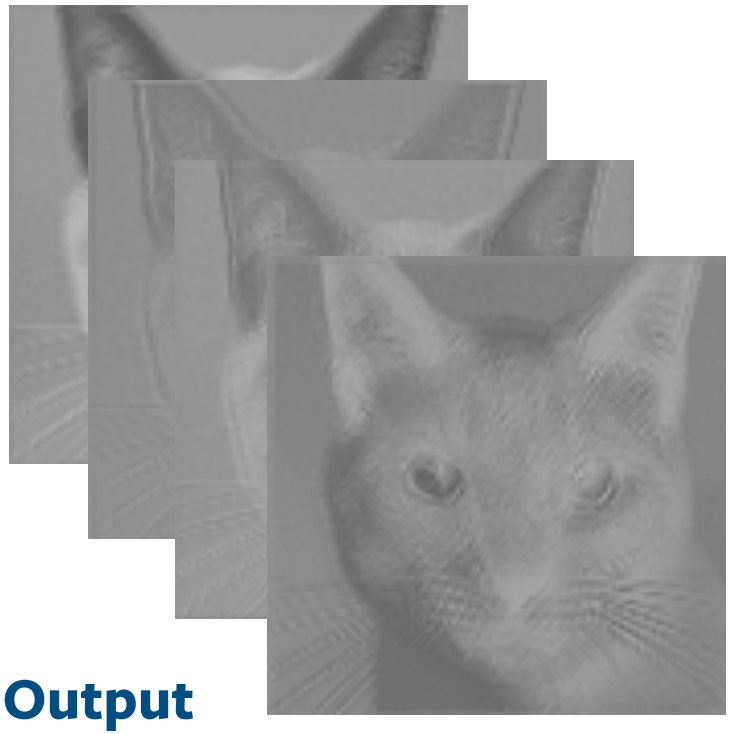
* Credits: Ian Goodfellow

A more complicated convolutional layer



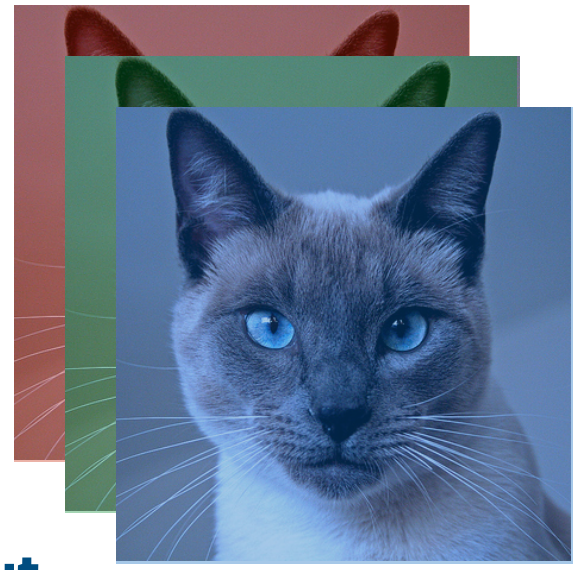
Input

filter



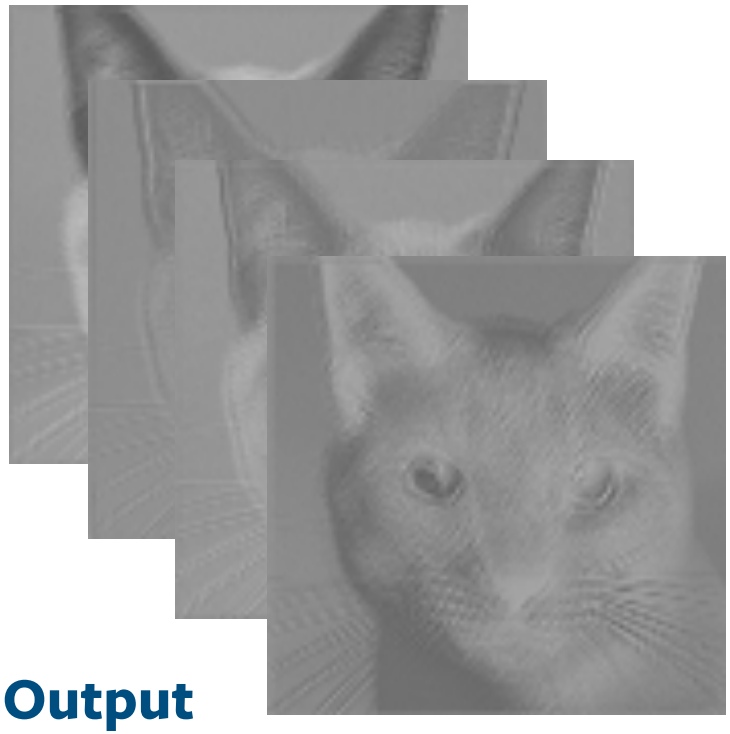
Output

A more complicated convolutional layer



Input

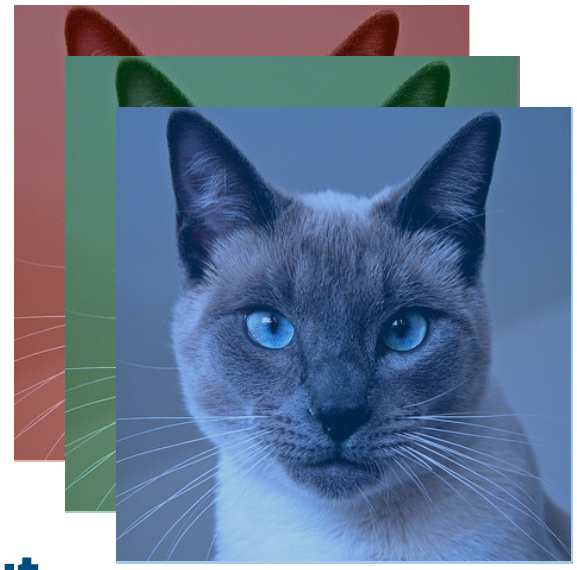
filter



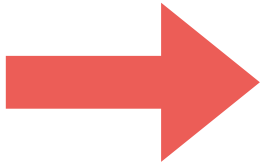
Output

How many parameters does this convolutional layer have?

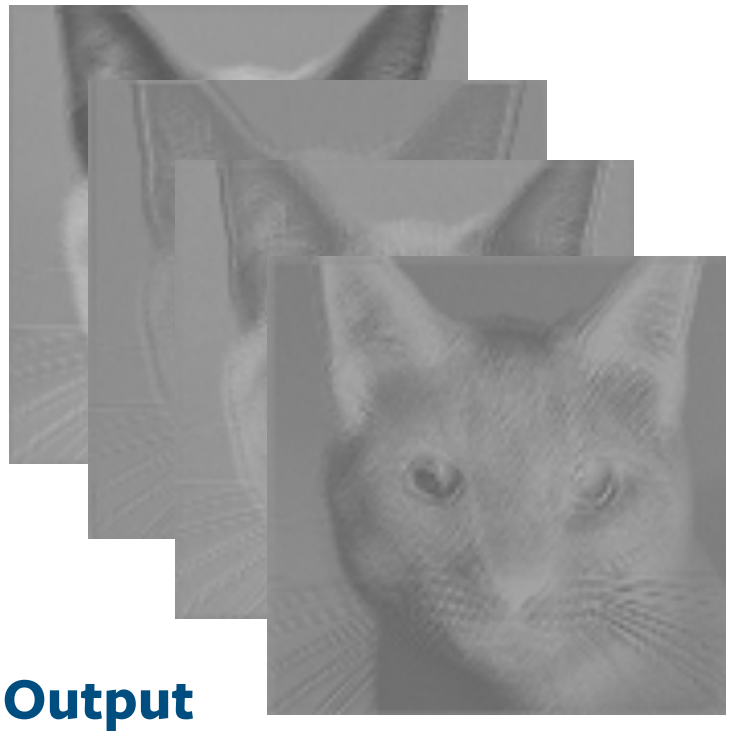
A more complicated convolutional layer



Input



filter



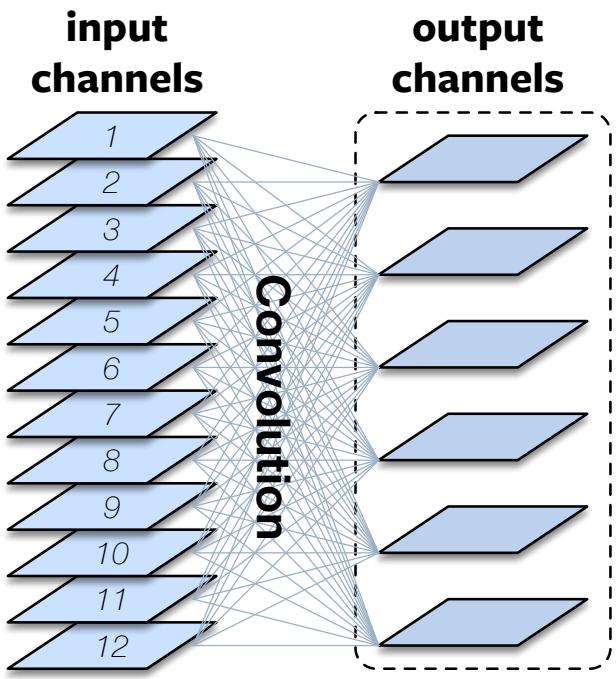
Output

How many parameters does this convolutional layer have?

3 x 4 x kernel height x kernel width

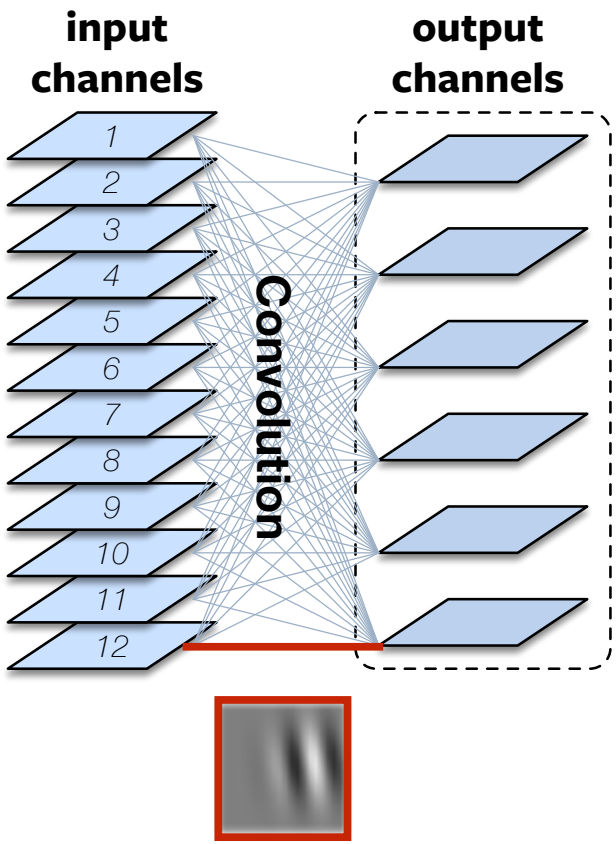
More details on convolutional layers

- Convolutional layers have five main hyperparameters:
  - Number of input channels
  - Number of output channels
  - Kernel size
  - Kernel stride
  - Input padding



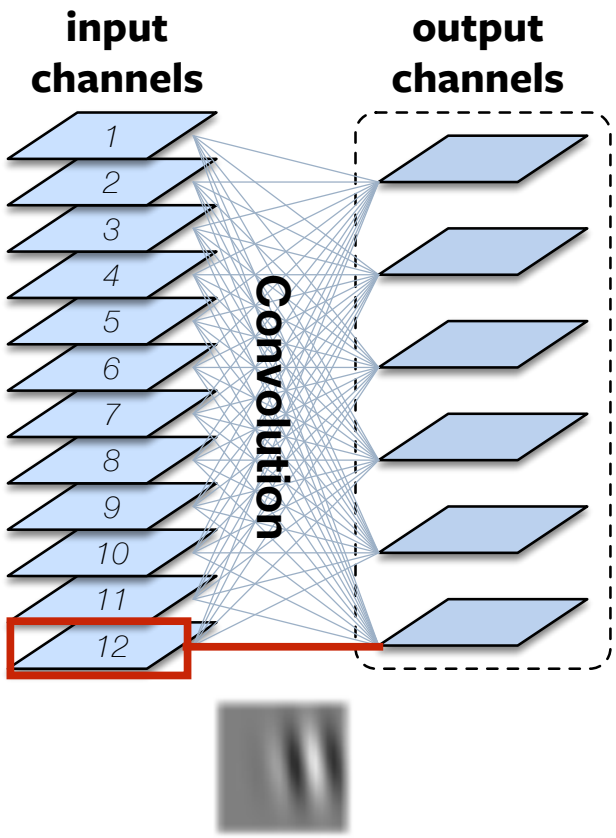
More details on convolutional layers

- Convolutional layers have five main hyperparameters:
  - Number of input channels
  - Number of output channels
  - Kernel size
  - Kernel stride
  - Input padding



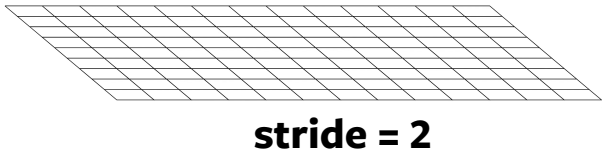
More details on convolutional layers

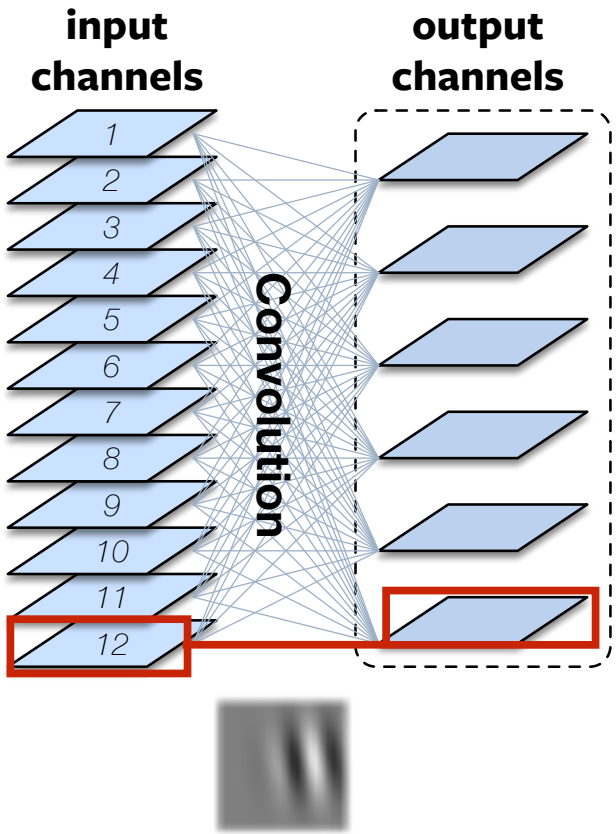
- Convolutional layers have five main hyperparameters:
  - Number of input channels
  - Number of output channels
  - Kernel size
  - Kernel stride
  - Input padding

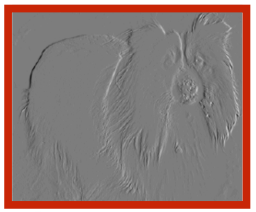


More details on convolutional layers

- Convolutional layers have five main hyperparameters:
  - Number of input channels
  - Number of output channels
  - Kernel size
  - Kernel stride
  - Input padding

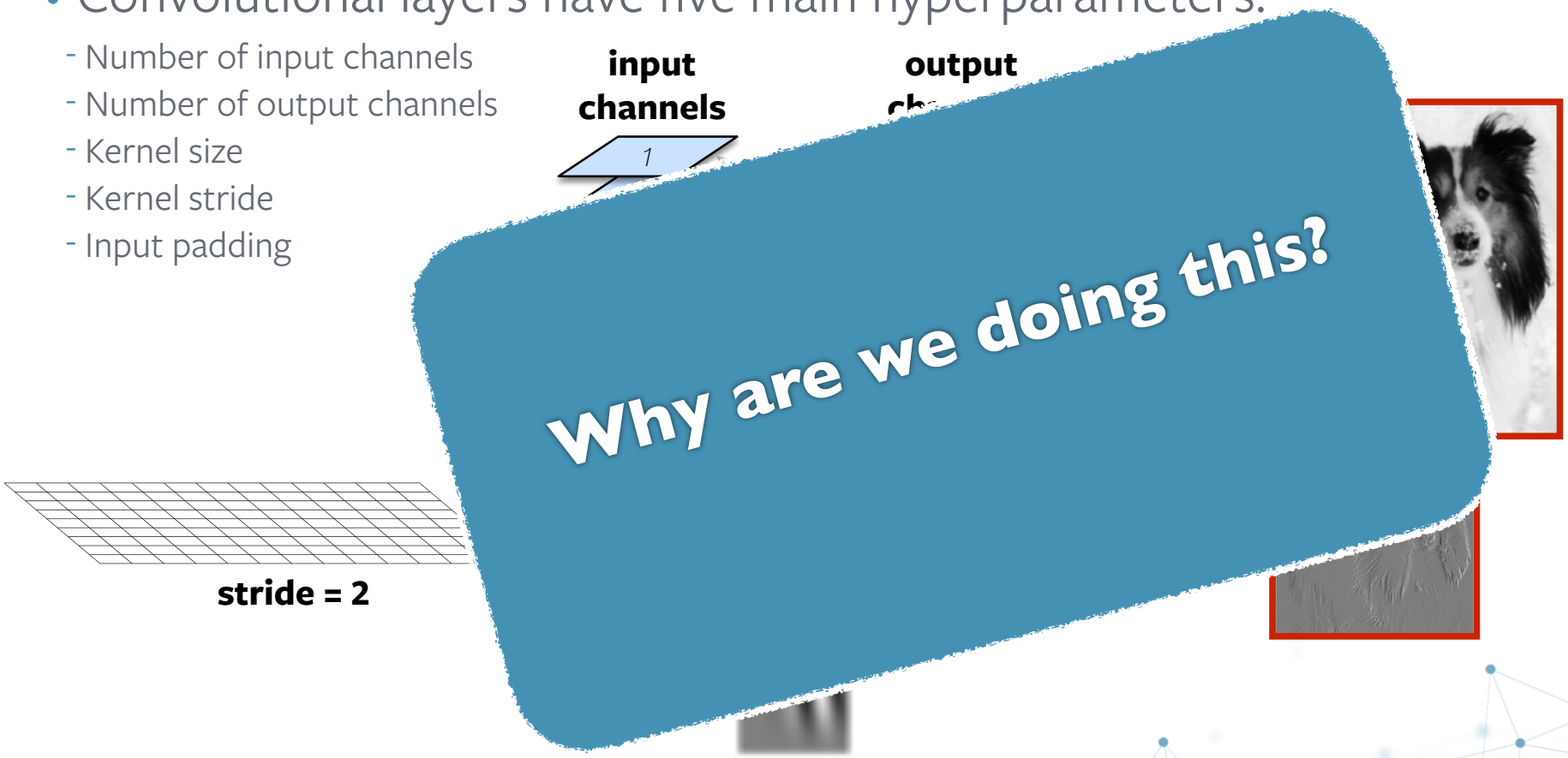






More details on convolutional layers

- Convolutional layers have five main hyperparameters:
  - Number of input channels
  - Number of output channels
  - Kernel size
  - Kernel stride
  - Input padding



More details on convolutional layers

- Convolutional layers have five main hyperparameters:
  - Number of input channels
  - Number of output channels
  - Kernel size
  - Kernel stride
  - Input padding

Is that all we need to build a modern conv net?

Is that all we need to build a modern conv net?

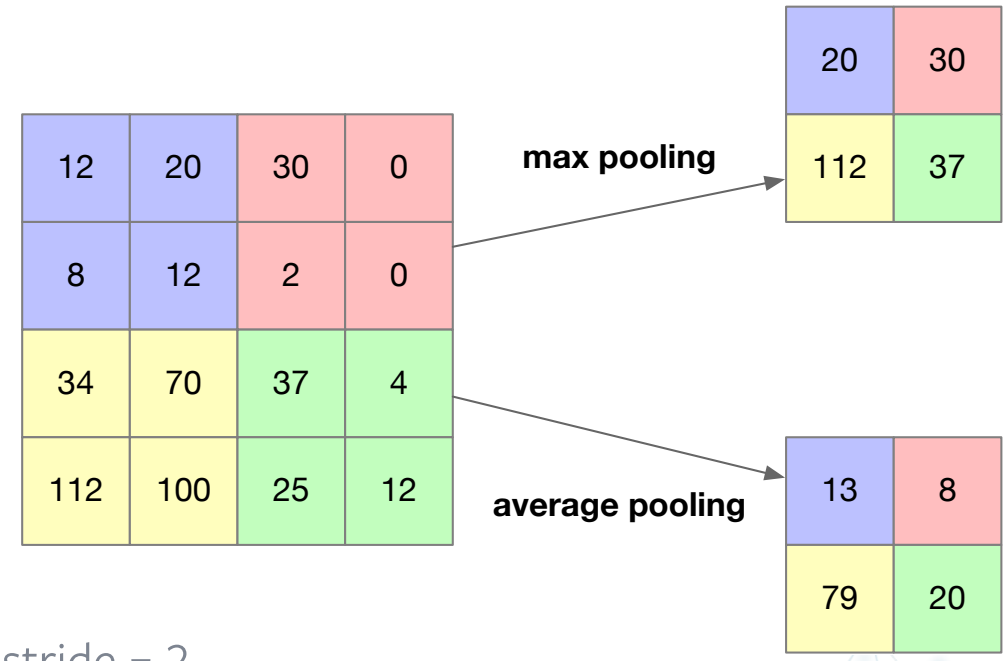
No! We also need:

- **Pooling layers:** These layers allow us to remove spatial structure to produce pooled (summarized) features
- **Rectified Linear Units:** Common non-linearity for convolutional networks
- **Batch Normalization:** Makes training more stable and we can train networks more quickly / reliably

Pooling

- Input and output of a convolution layer both have **spatial structure**

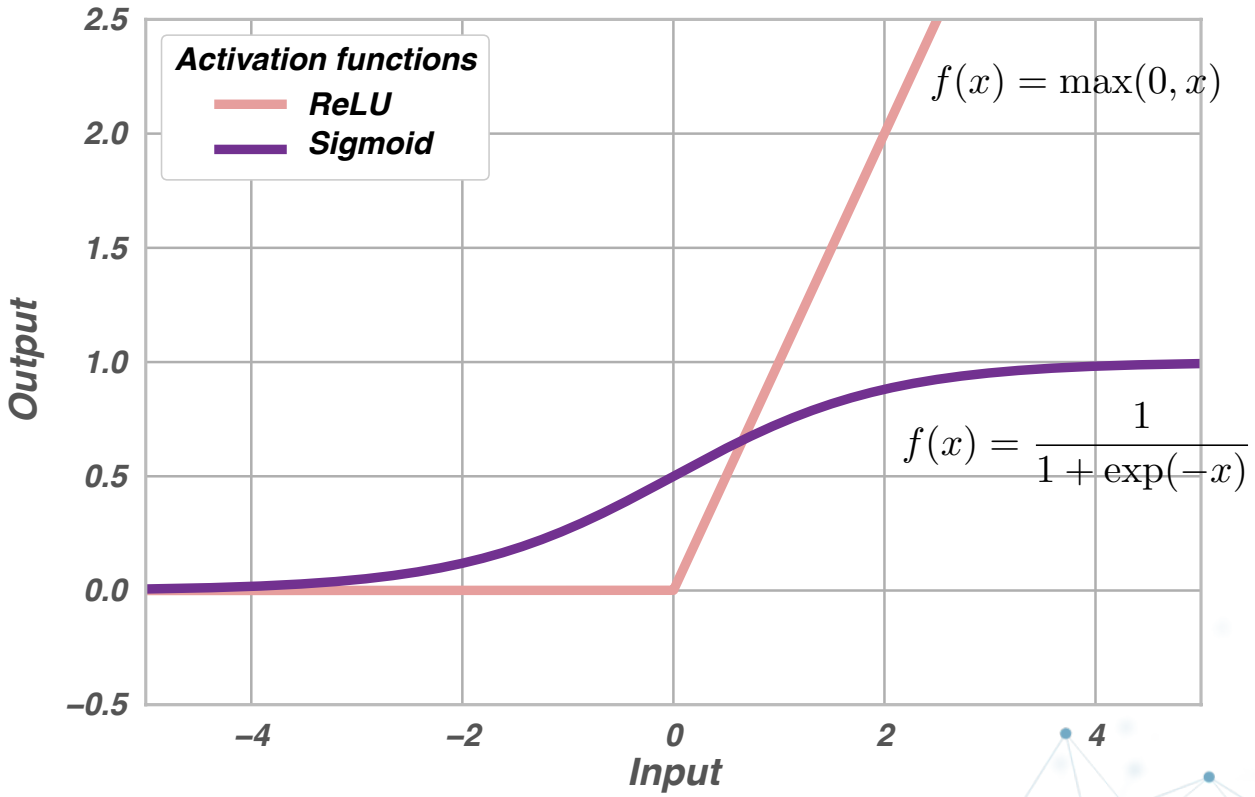
- We gradually remove the spatial structure via **pooling**:



* This pool uses size = 2, stride = 2.

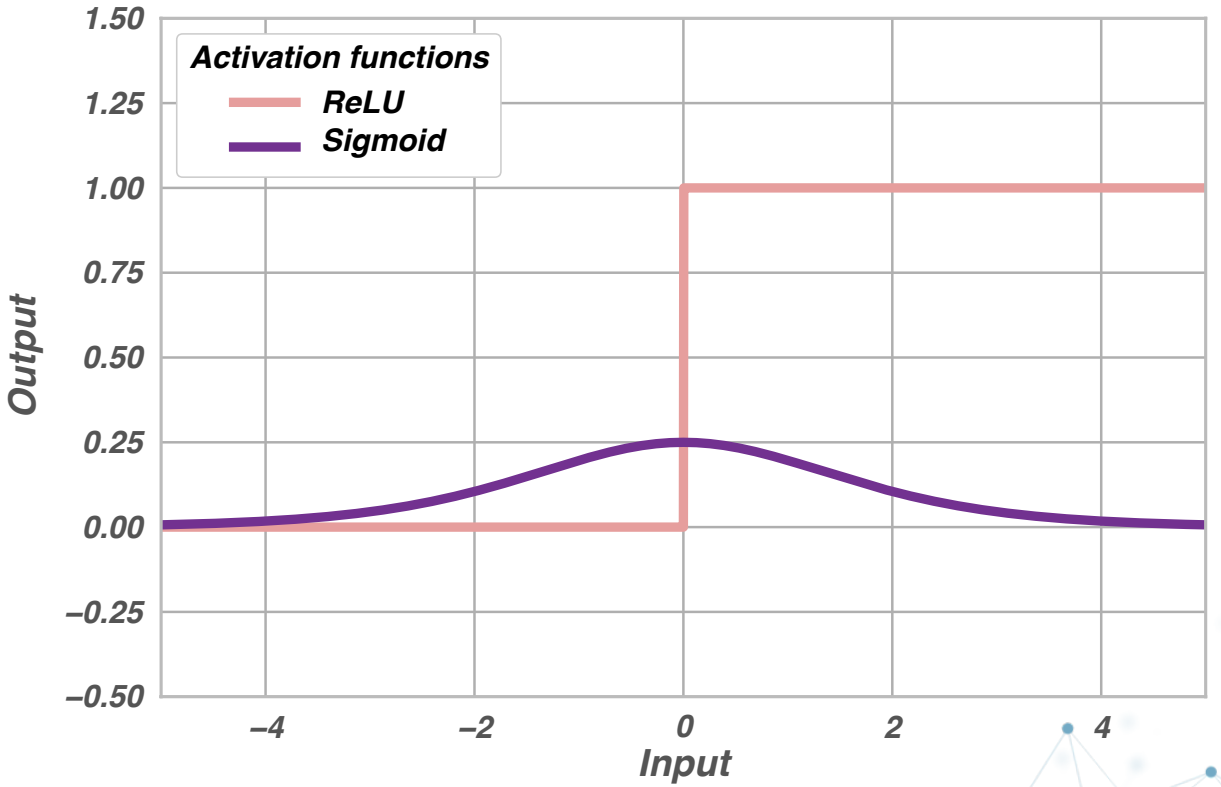
Rectified linear units

- **Rectified linear units** (ReLUs) have "better" gradients than sigmoids:



Rectified linear units

- **Rectified linear units** (ReLUs) have "better" gradients than sigmoids:

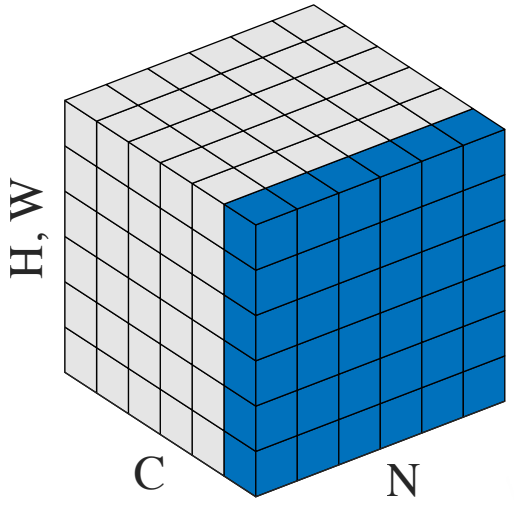


Batch normalization

- When using ReLU, gradients may easily blow up or go to zero
- **Batch normalization** make gradients better behaved:

Z-score input: $\hat{\mathbf{x}} = \frac{\mathbf{x} - \mu}{\sigma}$

Affine transform: $\mathbf{y} = \gamma\hat{\mathbf{x}} + \beta$



* Note: Batch normalization works differently in training and test time.

Batch normalization

- Makes network learning more stable and produces better final models:

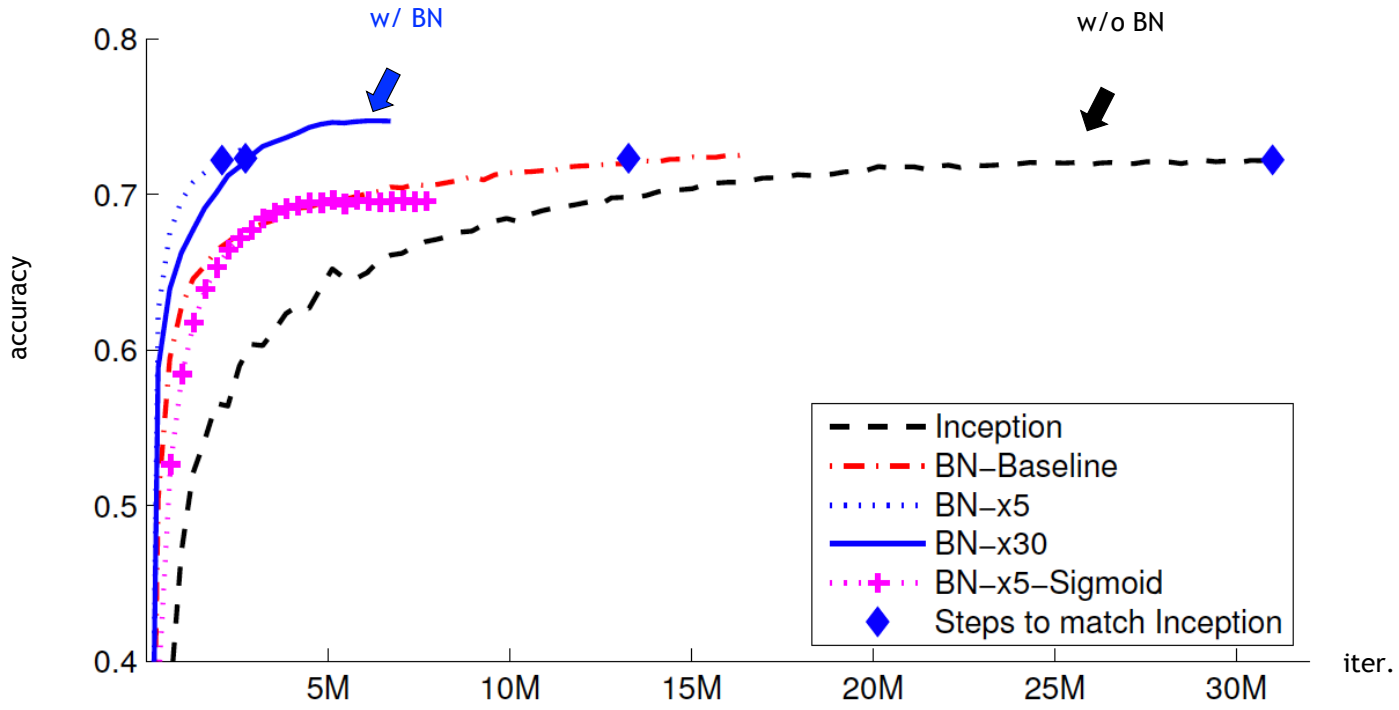
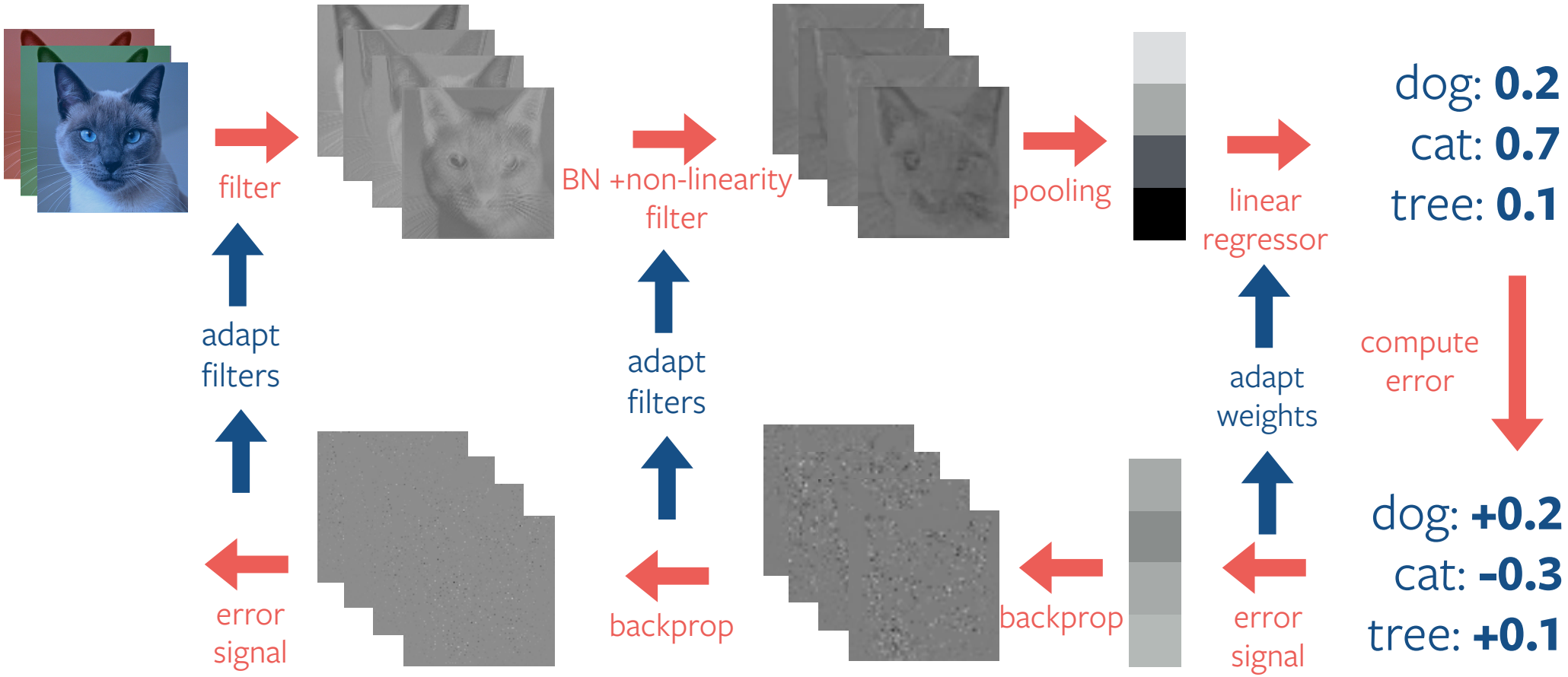
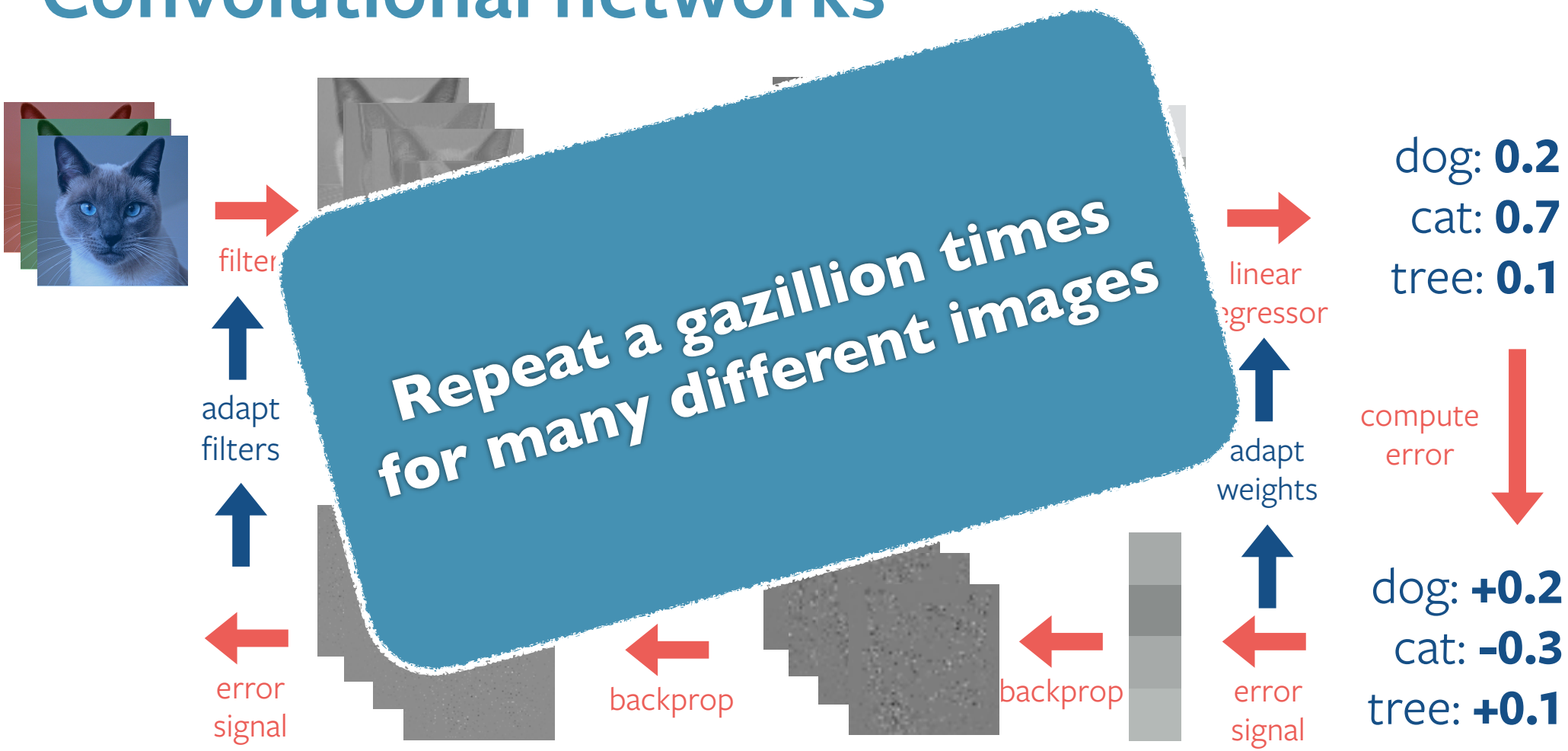


Figure credit: Ioffe & Szegedy

Convolutional networks

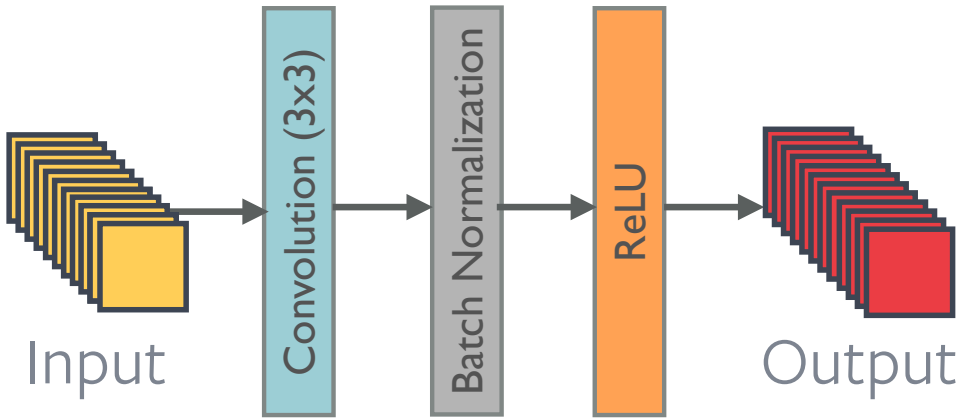


Convolutional networks



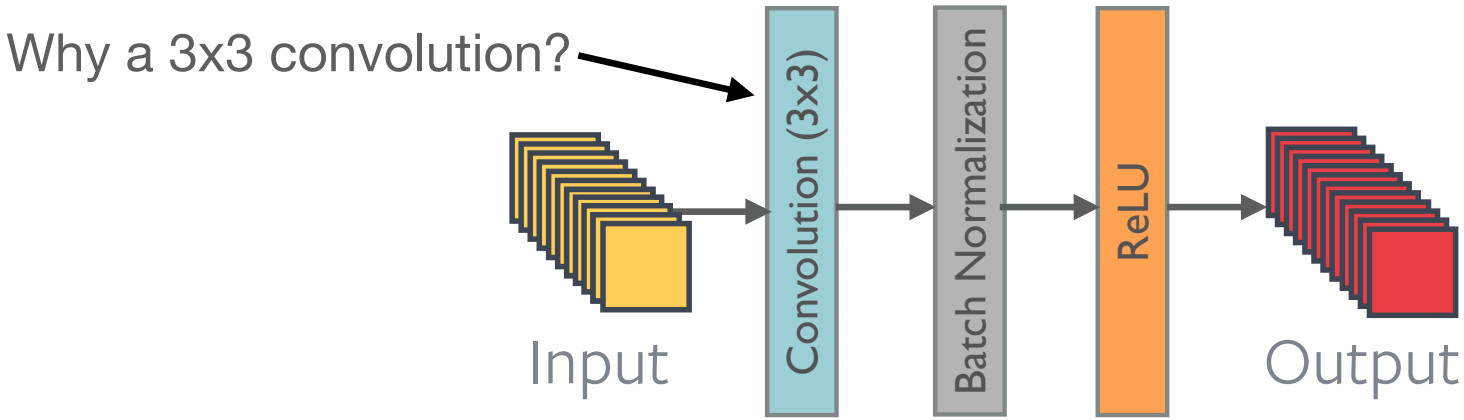
Building block

- Common building block comprises Conv2D + BatchNorm + ReLU:



Building block

- Common building block comprises Conv2D + BatchNorm + ReLU:



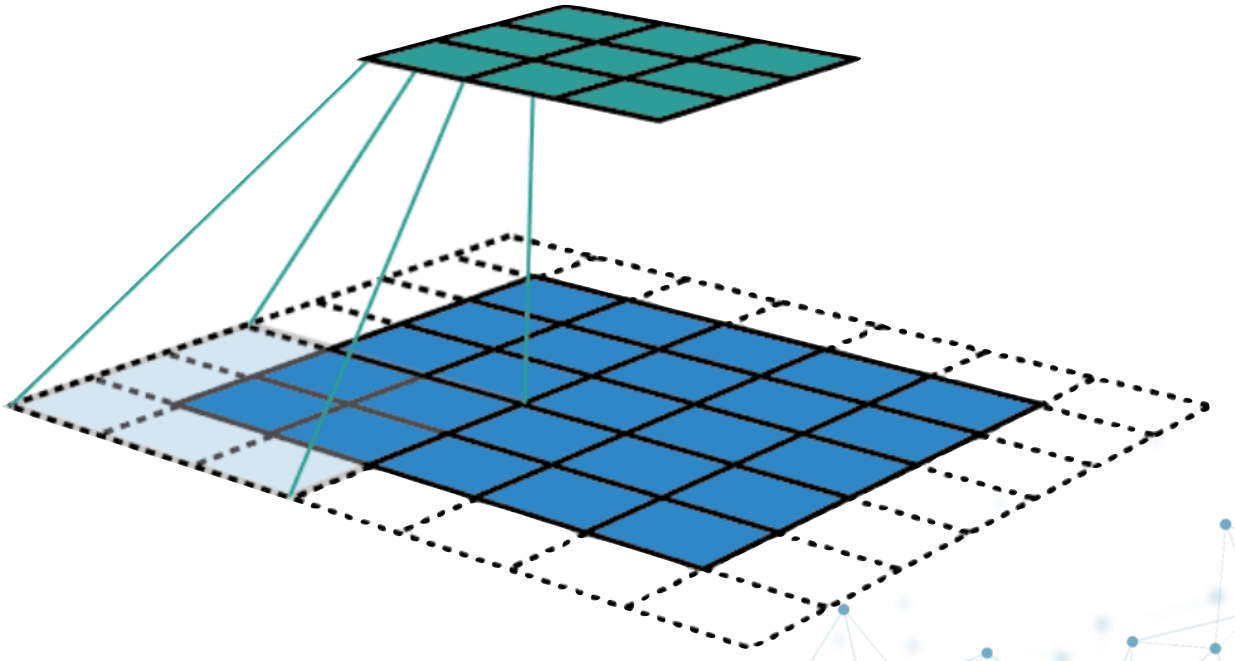
Receptive fields

- The **receptive field** of a feature is the part of the input that "influenced" the feature

Filter sizes

- The **receptive field** of a feature is the part of the input that "influenced" the feature:

Feature map

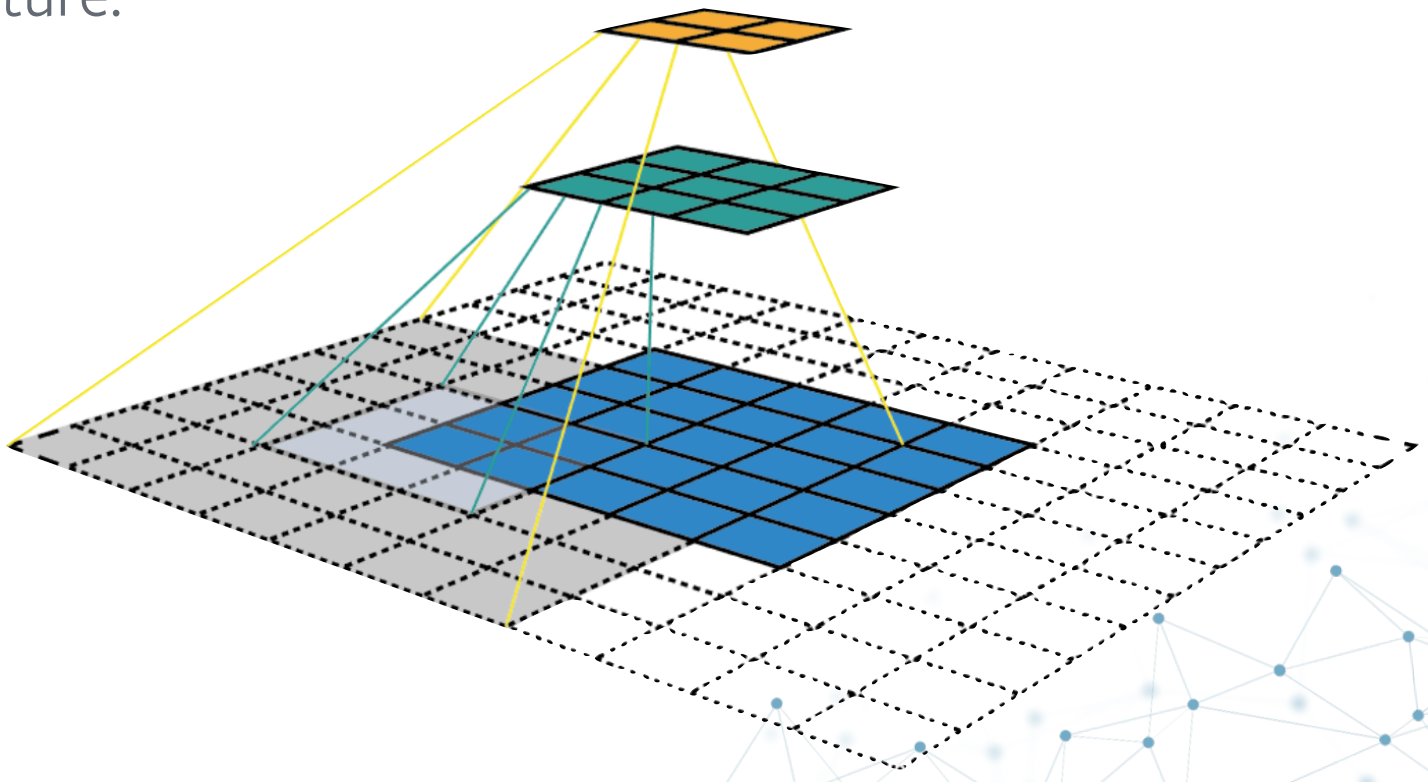


Input image

* Credits: Dang Ha The Hien

Receptive fields

- The **receptive field** of a feature is the part of the input that "influenced" the feature:



* Credits: Dang Ha The Hien

Receptive fields

- What is the receptive field of one **5x5 filter**?
  - And how much compute does it take?

Receptive fields

- What is the receptive field of one **5x5 filter**?
  - And how much compute does it take?

- What is the receptive field of **two 3x3 filters**?
  - How much compute does that take?

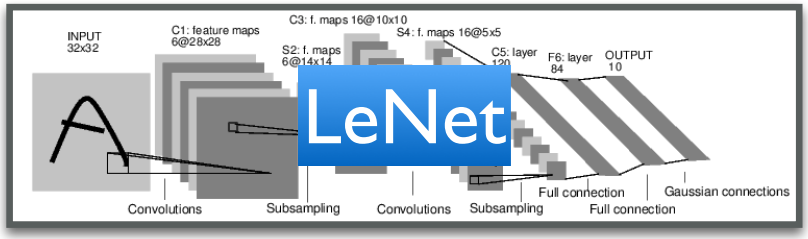
Receptive fields

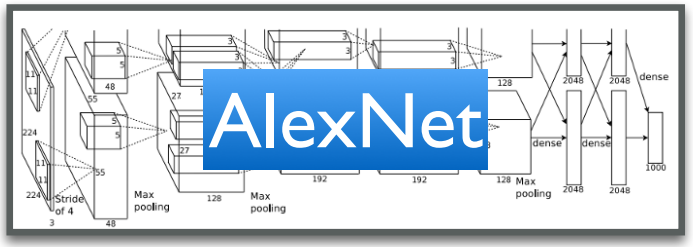
- What is the receptive field of one **5x5 filter**?
  - And how much compute does it take?

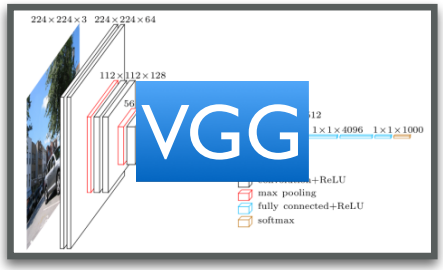
- What is the receptive field of **two 3x3 filters**?
  - How much compute does that take?

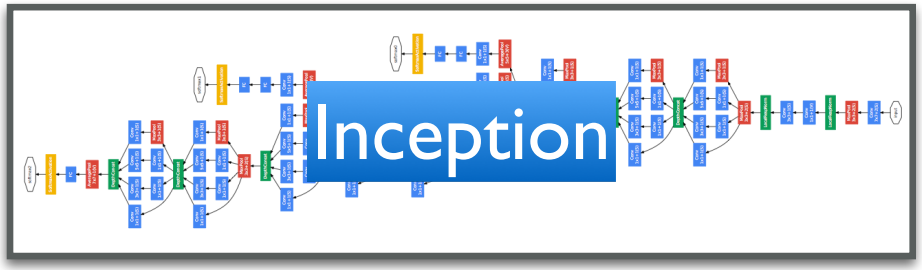
- Many current architecture use primarily **3x3 filters** for this reason

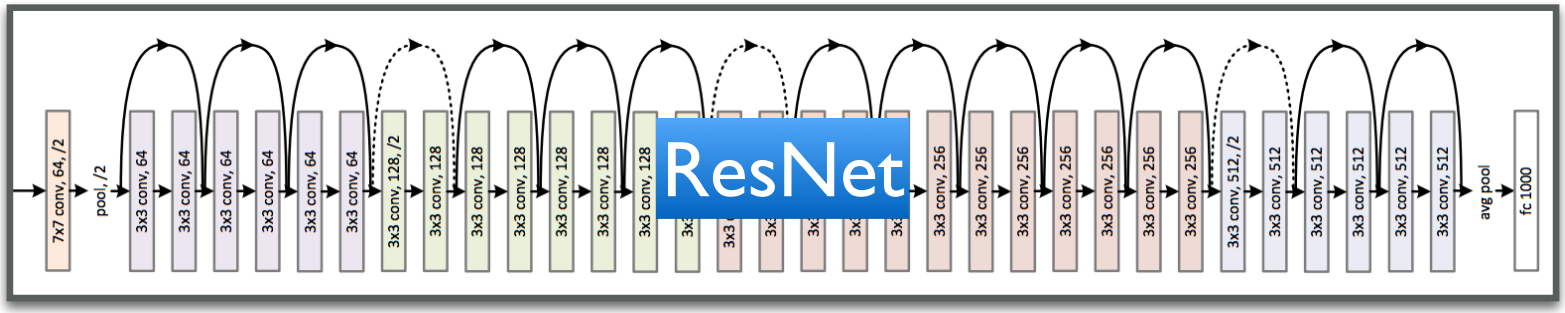
Convolutional networks







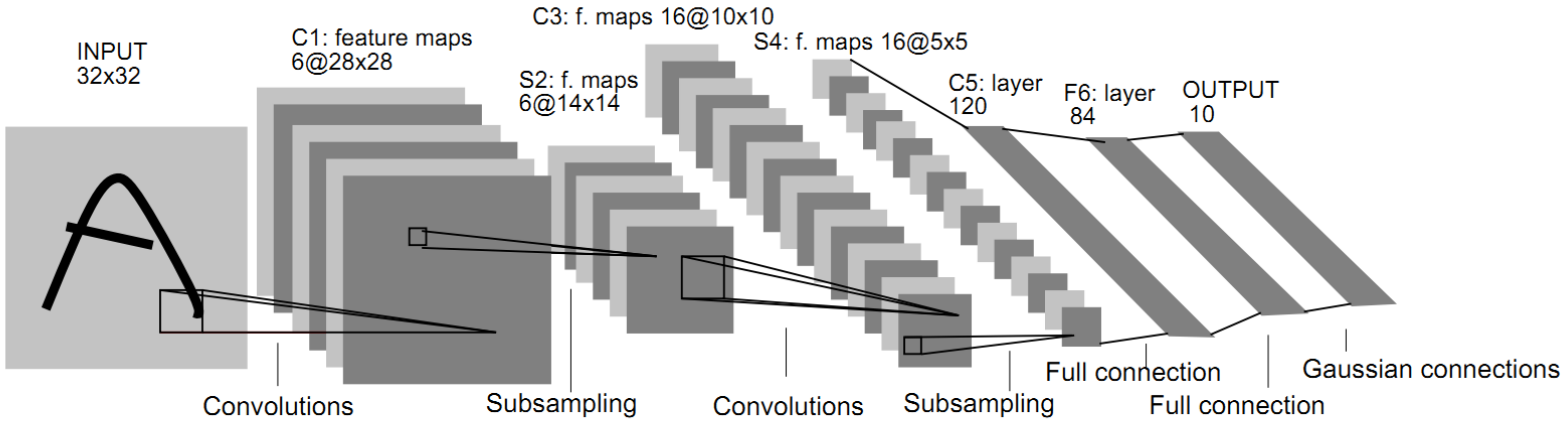




* Slide credits: Gao Huang

LeNet

- First model that really worked by using convolutions:

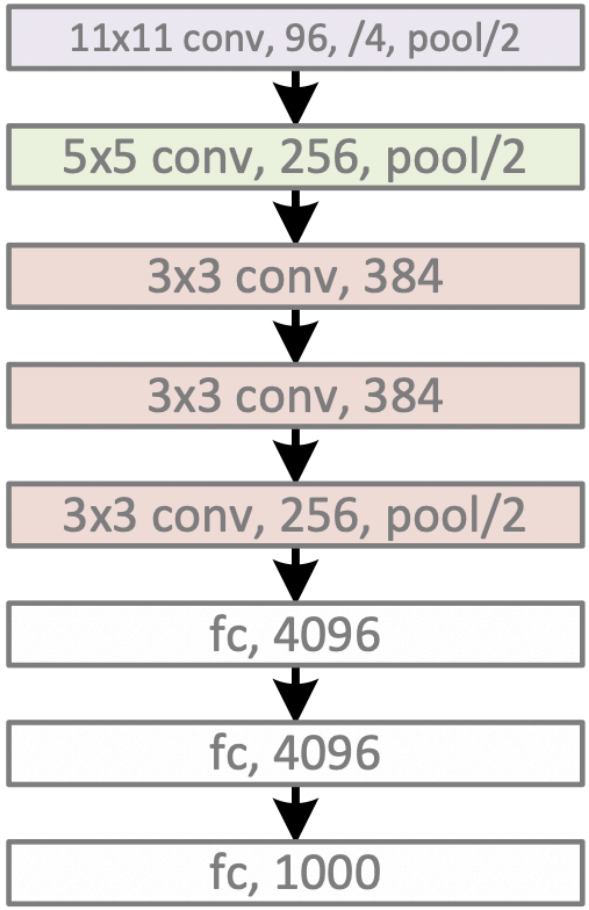


- Compared to today's networks, few convolutional layers and more fully-connected layers

AlexNet

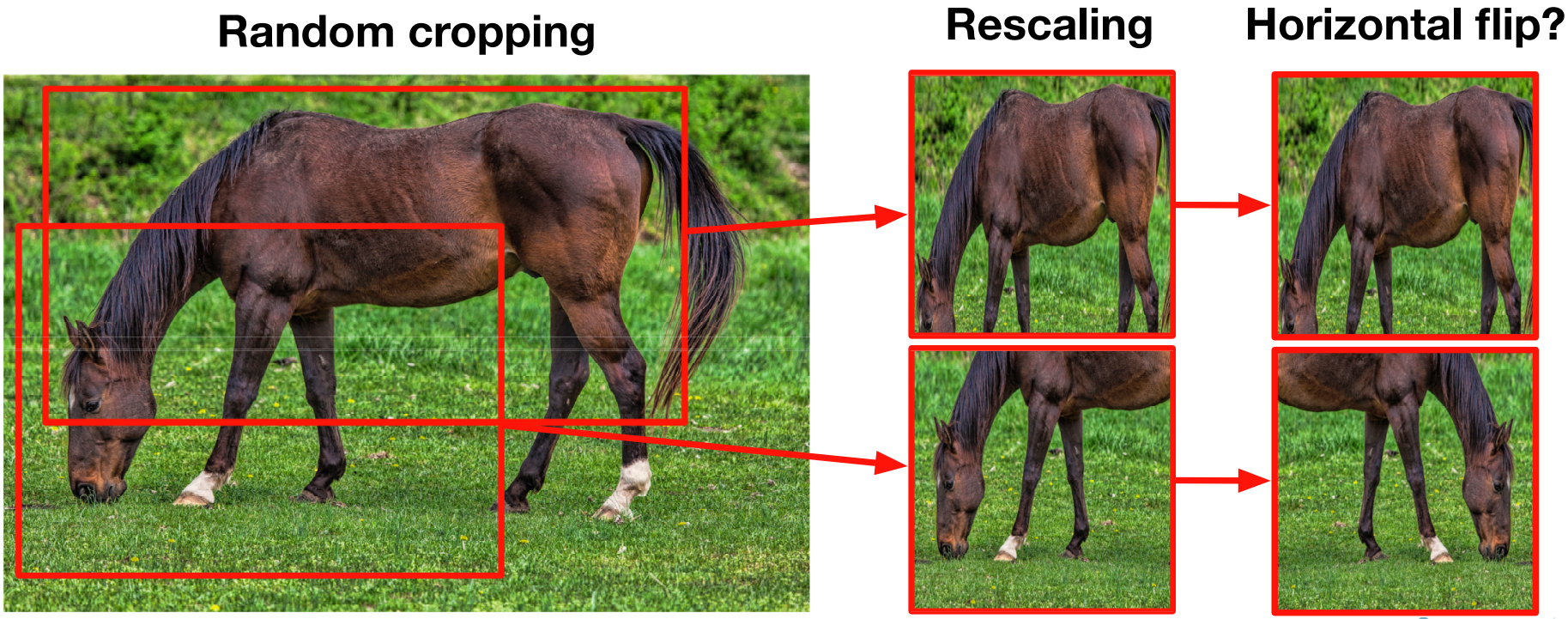
- LeNet-style model with three new components:
  - **Rectified linear units**
  - **Dropout** to reduce overfitting (currently less popular)
  - **Data augmentation** (still very important)

- AlexNet played a key role in popularizing convolutional networks



Data augmentation

- Create many (hard) training examples from a single annotated image:



* Note: This is only used at training time! (Why?)

Summary

- Convolutional layers perform many convolutions and sum the results to create output channels
- Pooling gradually eliminates spatial structure from the inputs
- Rectified linear units facilitate learning by having good gradients
- Batch normalization controls the scale of the gradients
- You now understand all the components to build AlexNets!

Reading material

- S. Ioffe and C. Szegedy. **Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift**. In *International Conference on Machine Learning* (*ICML*), 2015.

- K. He, X. Zhang, S. Ren, and J. Sun. **Deep Residual Learning for Image Recognition**. In *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), 2016.

Questions?