Image recognition

deep learning 2

R.Grouls, 5 mei 2022

Neural networks

recap

$$X = \{x_1, \ldots, x_n\}$$

$$f(X) = WX + b$$

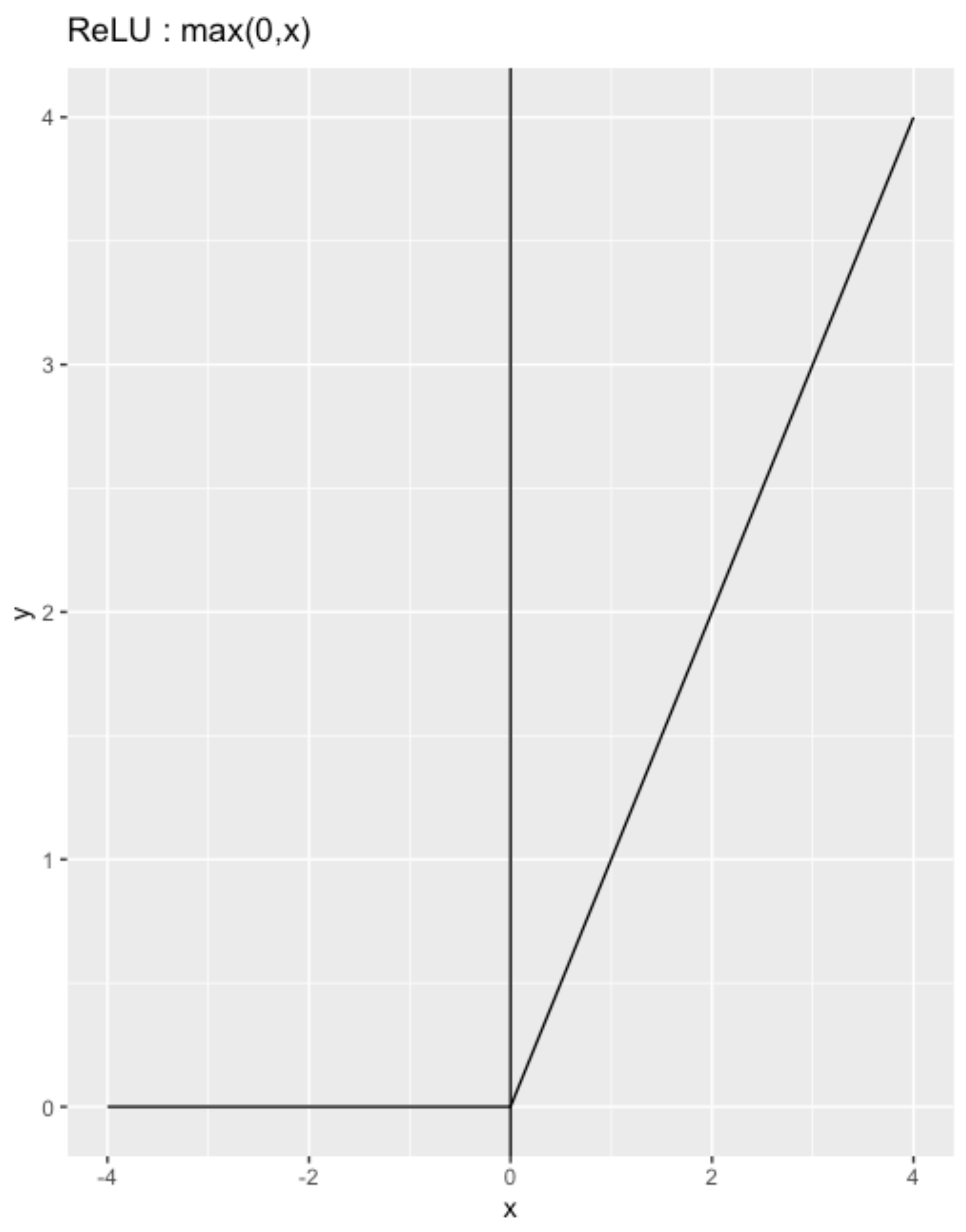
$$\hat{y} = \sigma(f(X))$$

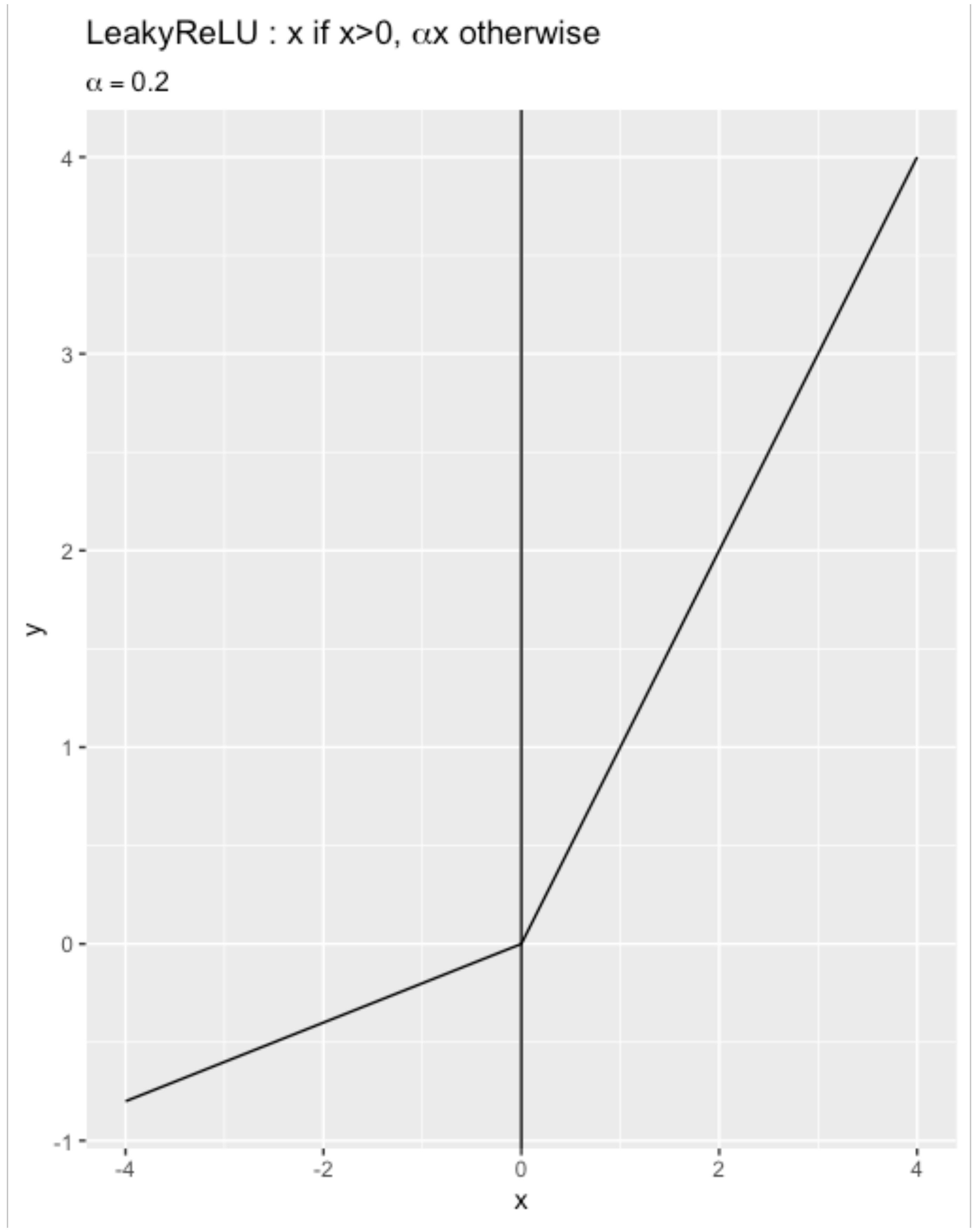
$$X \rightarrow \sigma(f(X)) \rightarrow Loss(y, \hat{y})$$

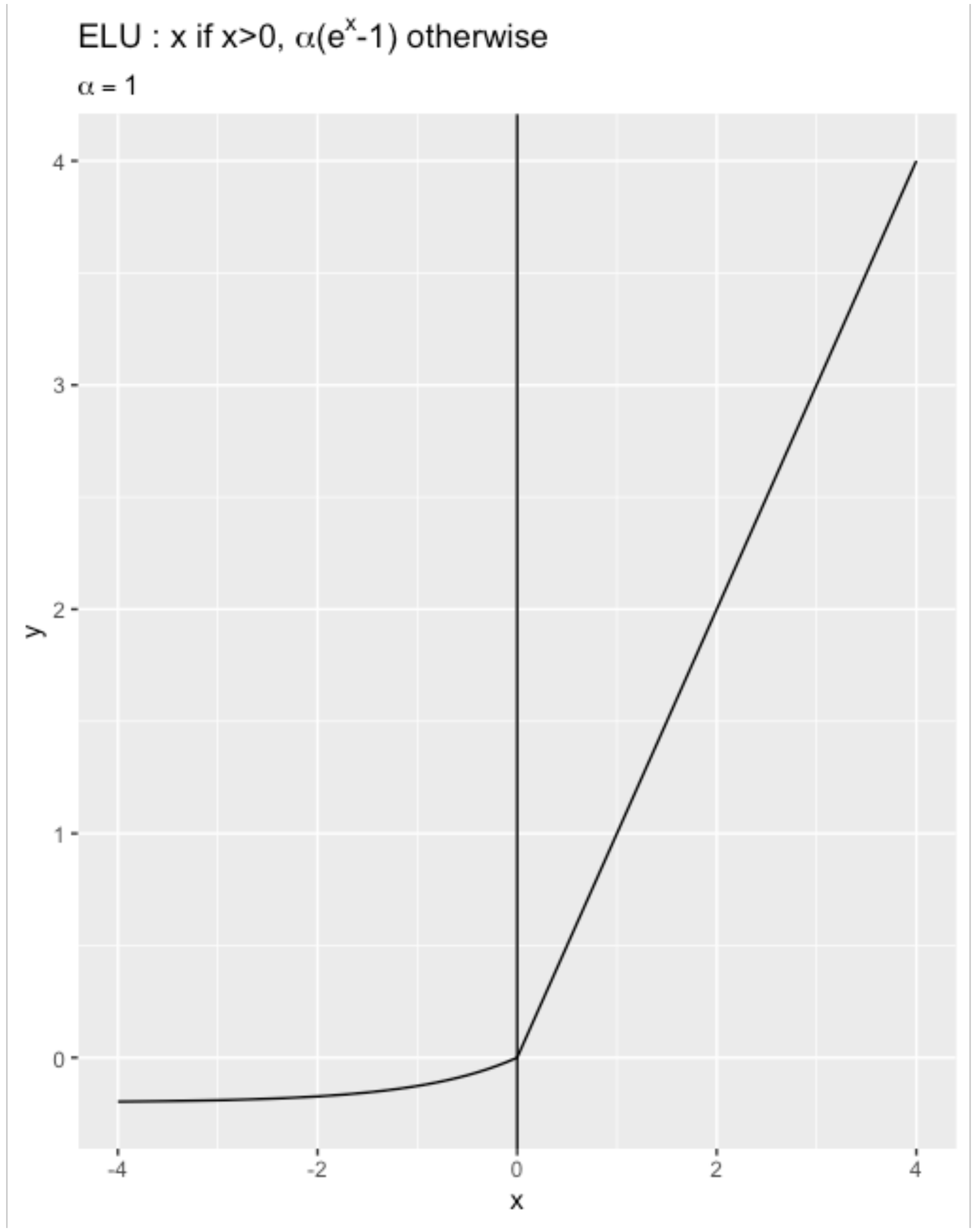
Backpropagation

Activations

non-linearities







Convolutions - the motivation

The curse of dimensionality

An image of size 28x28 has 784 pixels

With 256x256 you are already at 65536 pixels

Treating every pixels as a feature will blow up the amount of parameters for your model:

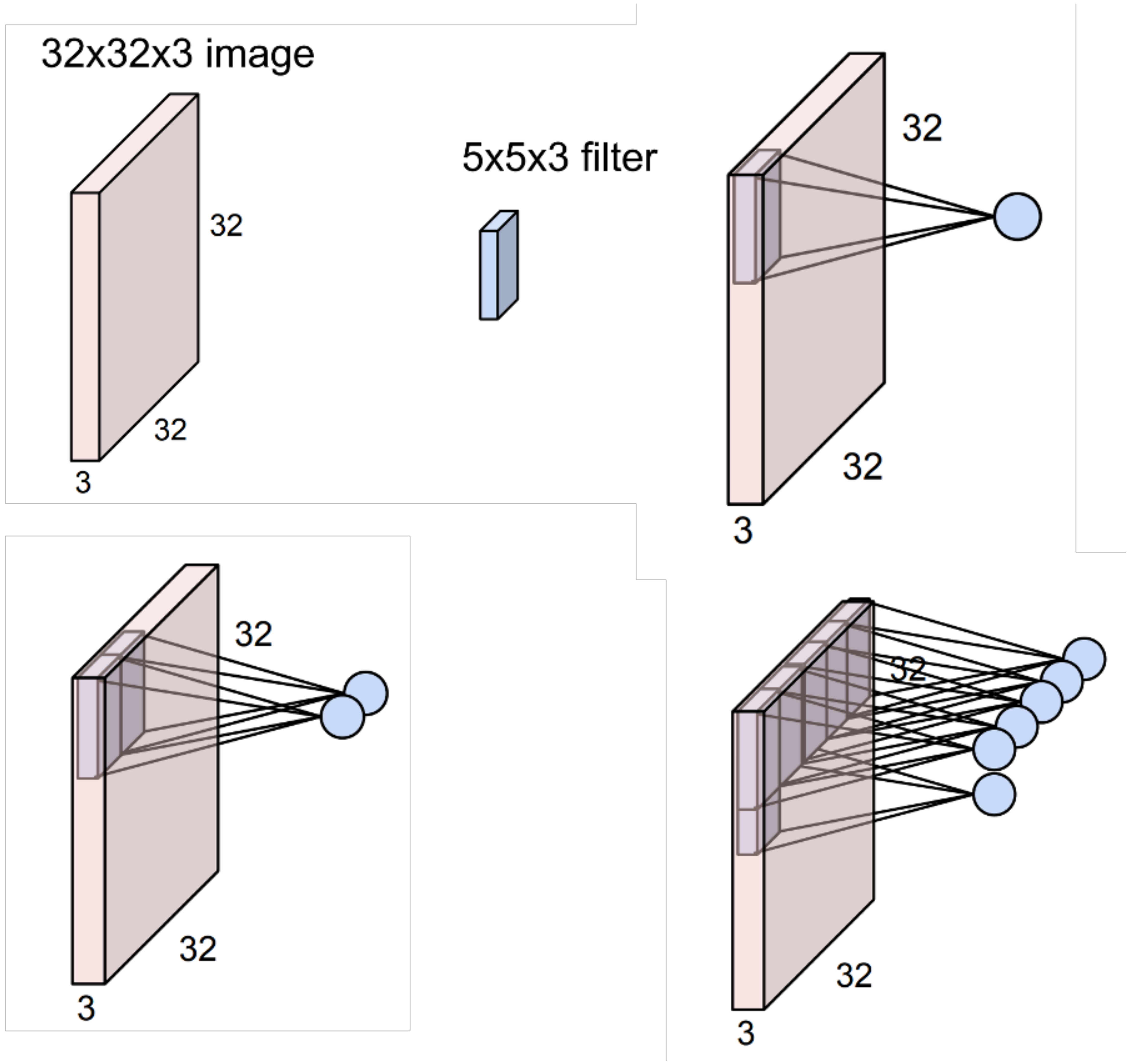
Assuming you halve the dimension, a batch with dimensions (32, 65536) will need a (65536, 32000) weight matrix.

That are over 2×10^9 parameters, just for the first layer…

Convolutions

Divide and conquer

- Take a filter of size (5x5x3)
- This needs just 75 parameters
- Slide it over the image, like a scanner



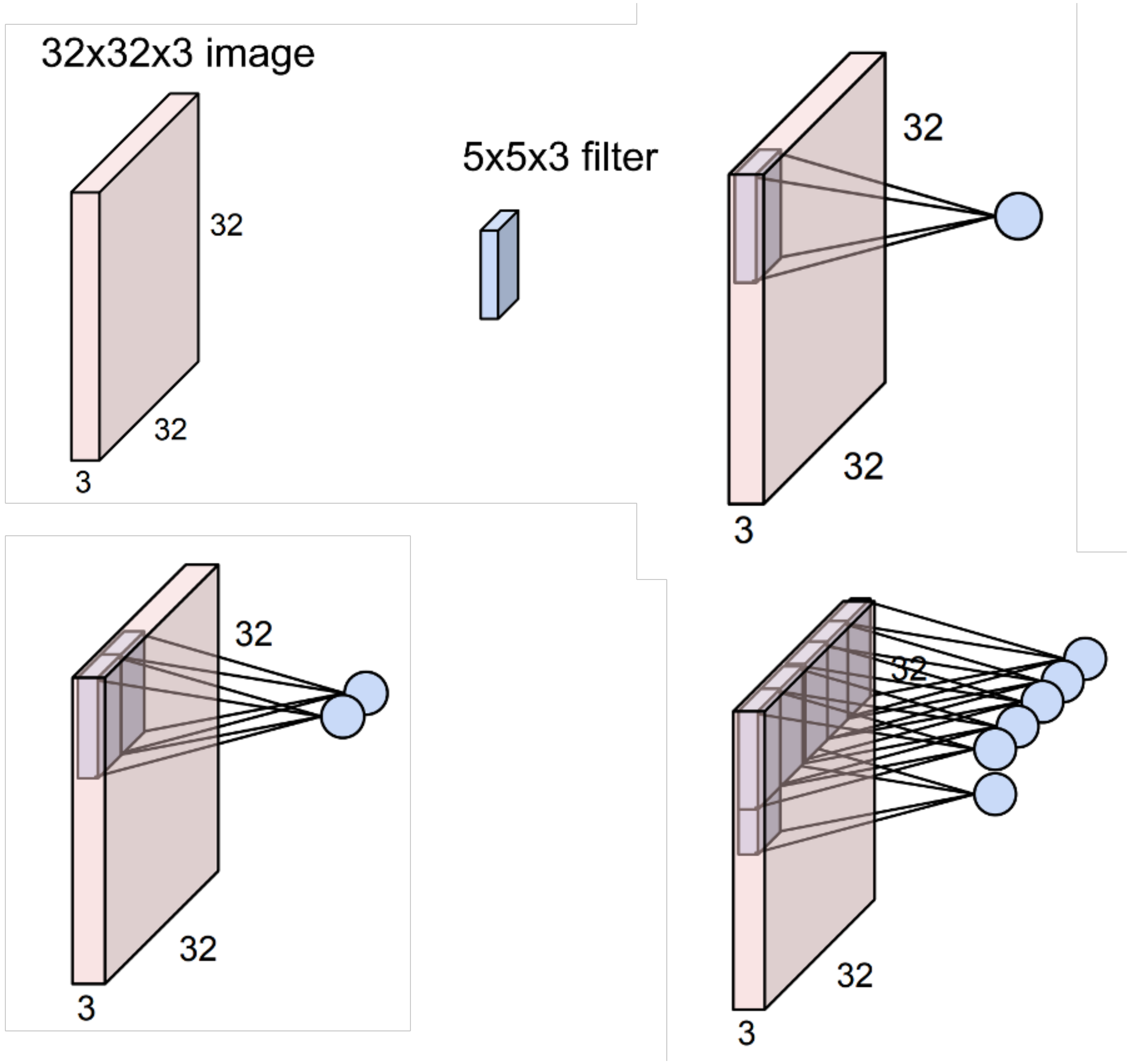
Convolutions

Divide and conquer

- Calculate the dot product between the filter and the image
- E.g.,

$$\begin{bmatrix} 1 & 10 \\ 2 & 20 \end{bmatrix} \cdot \begin{bmatrix} 1 & 2 \\ 2 & 0 \end{bmatrix} =$$

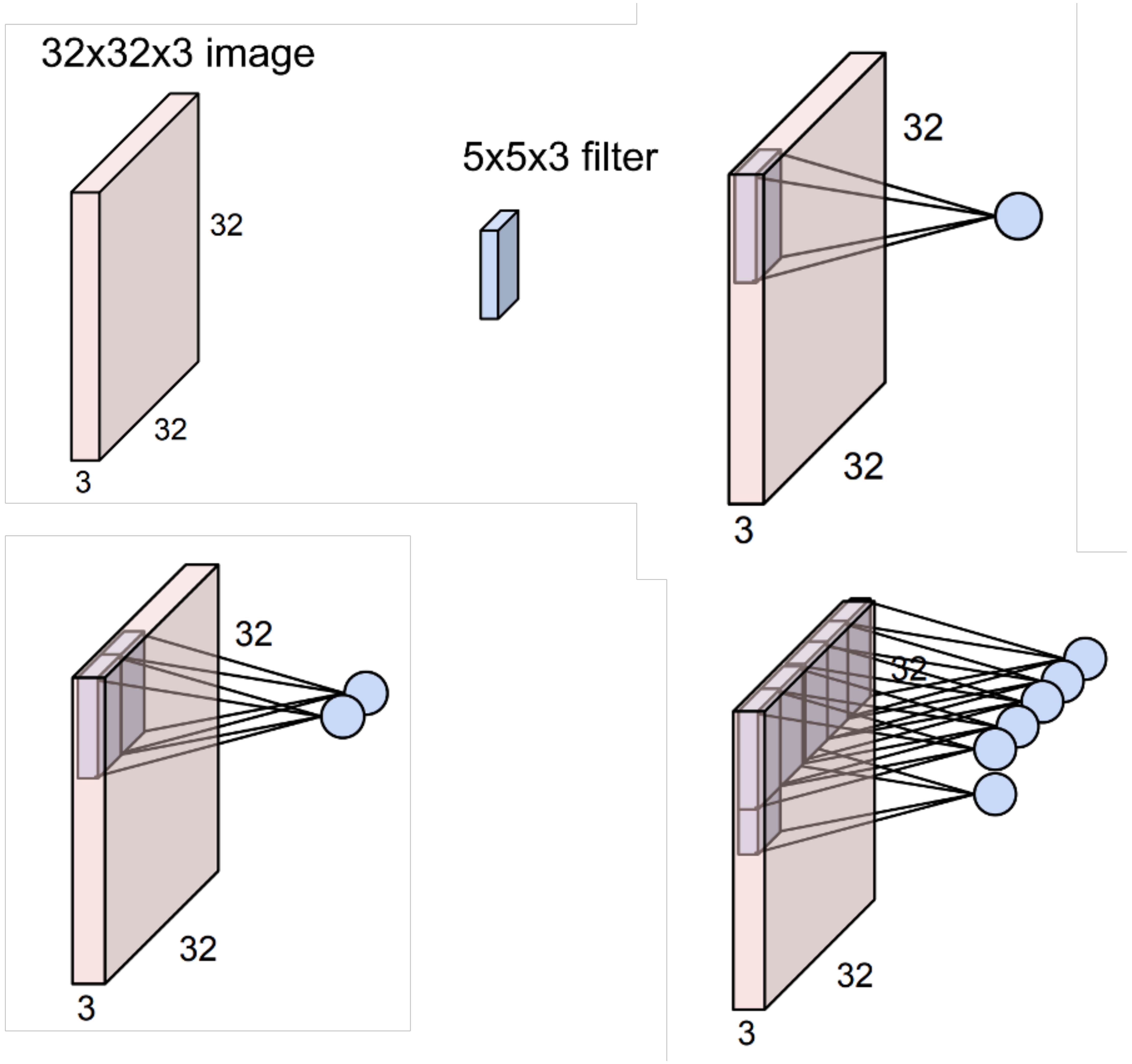
$$(1 * 1) + (10 * 2) + (2 * 2) + 0 = 25$$



Convolutions

Divide and conquer

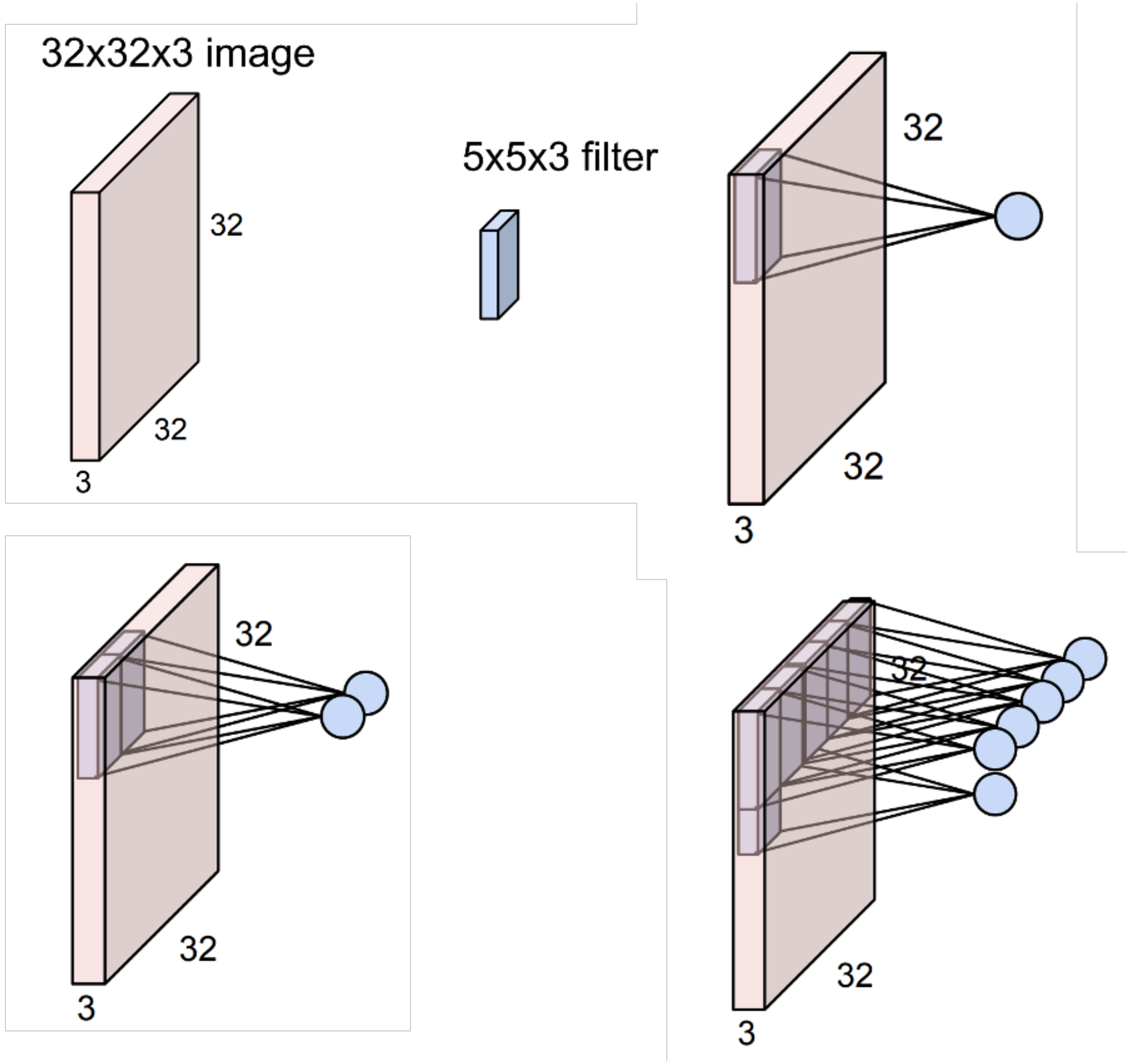
- So, with a 5x5 filter, every 5x5 image slice is reduced to a single number
- This number contains weighted information about it's neighborhood
- We call the result an activation map.



Convolutions

Divide and conquer

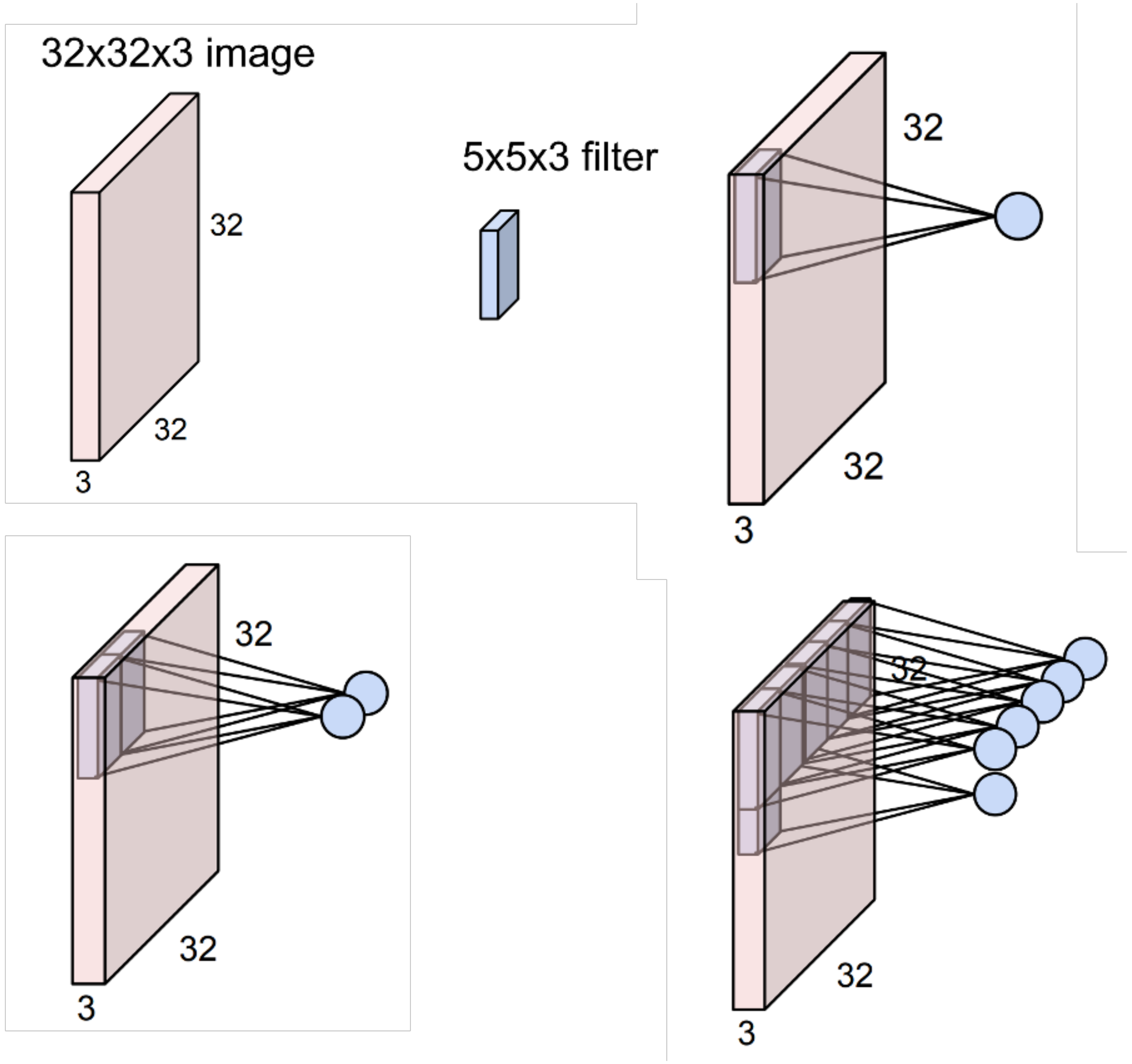
- The filter goes through the depth of the input
- Parameters of the filter are: width, height, depth
- In the first case, the depth is 3, representing the RGB colors



Convolutions

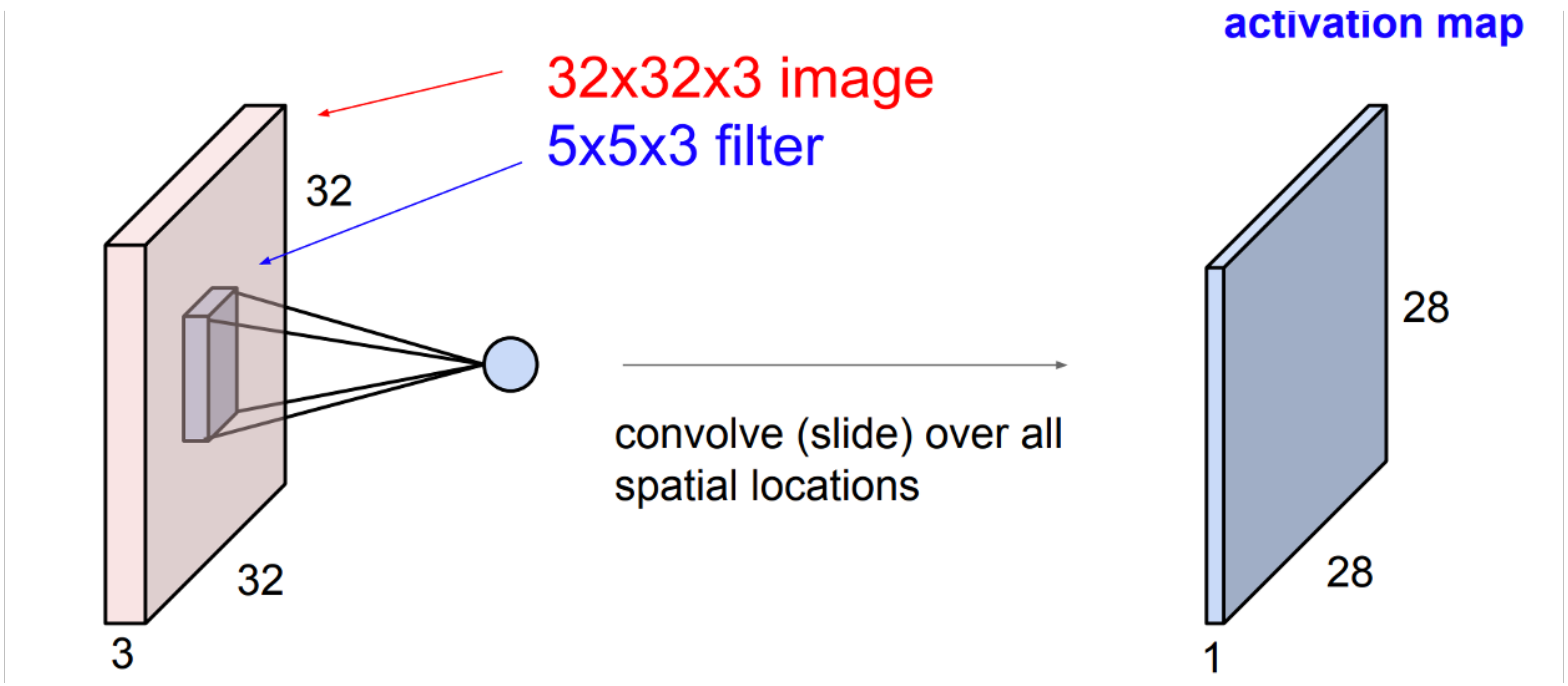
Divide and conquer

- Later on, the channel represents the amount of filters the model has available.
- Typically, you might start with a matrix with dimensions like (256, 256, 3)
- After a few layers of convolutions, you end up with dimensions like (5, 5, 512) which means there are 512 activation maps



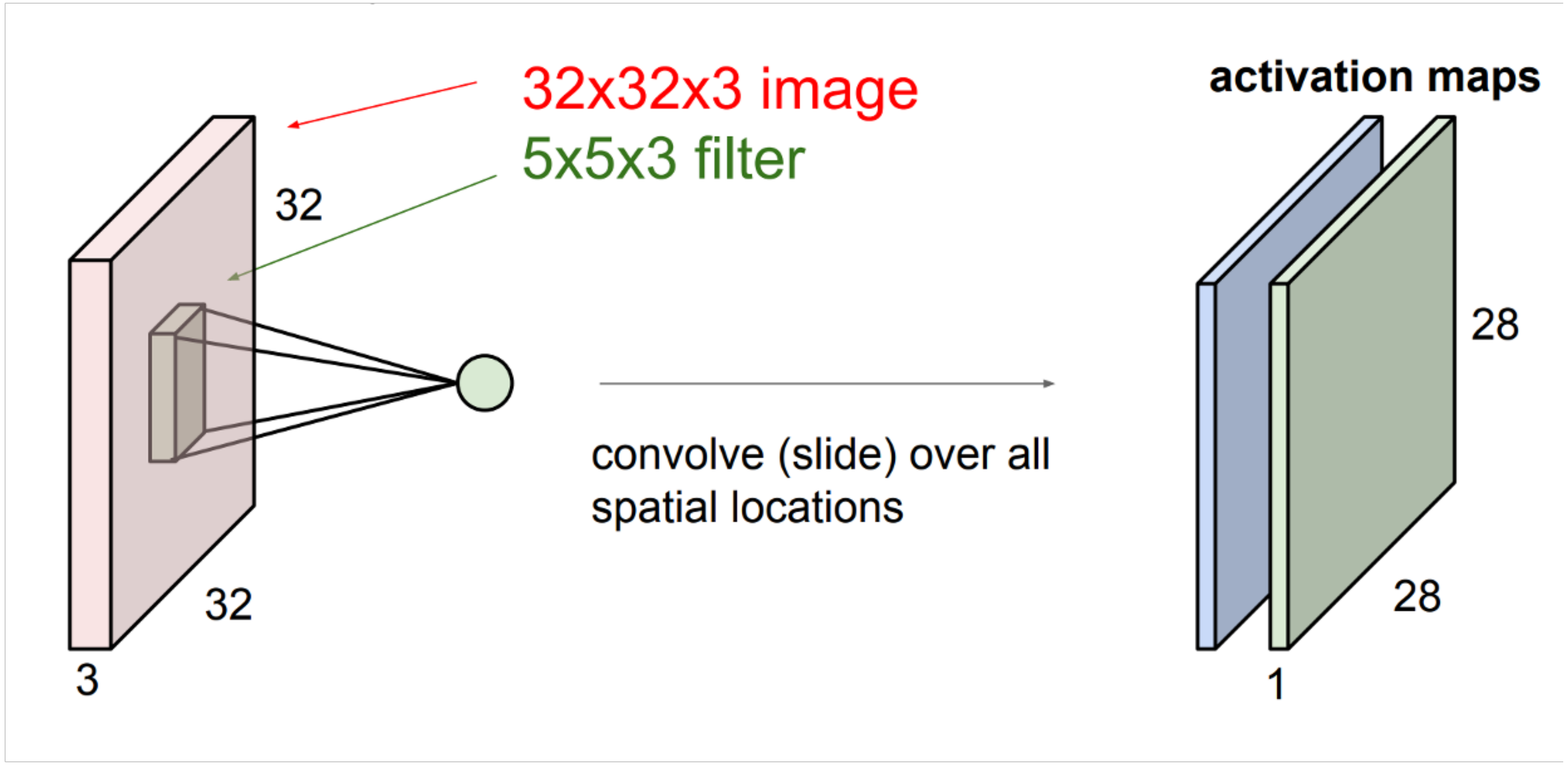
Convolutions

Divide and conquer



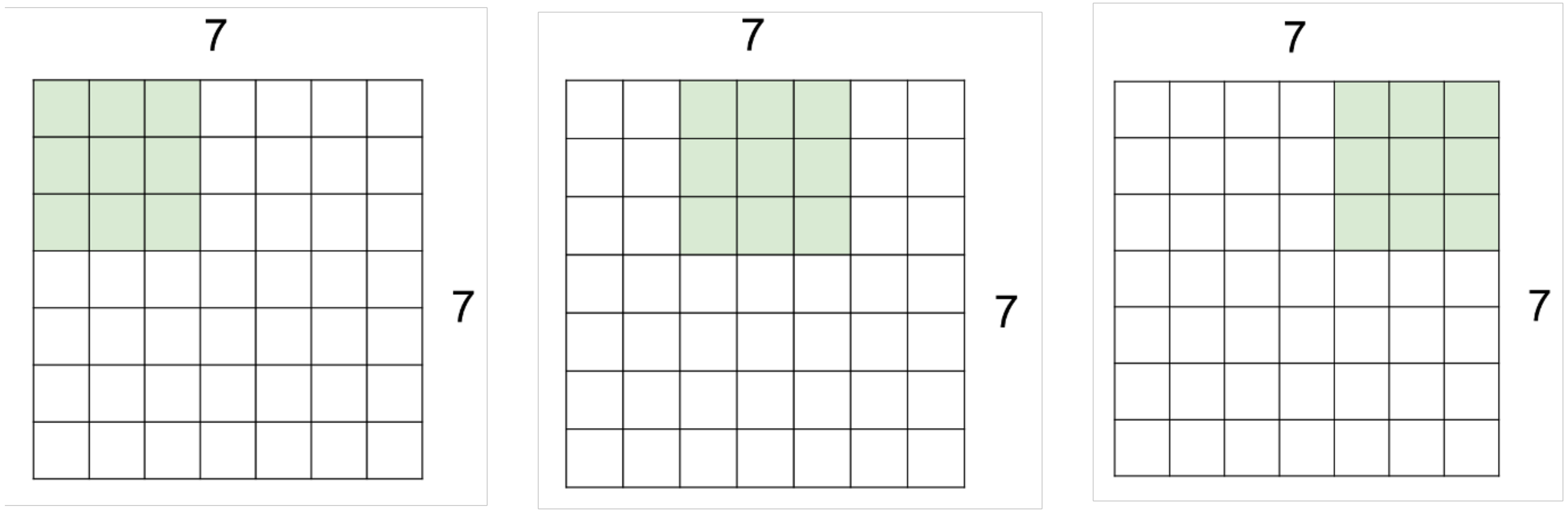
Convolutions

Divide and conquer



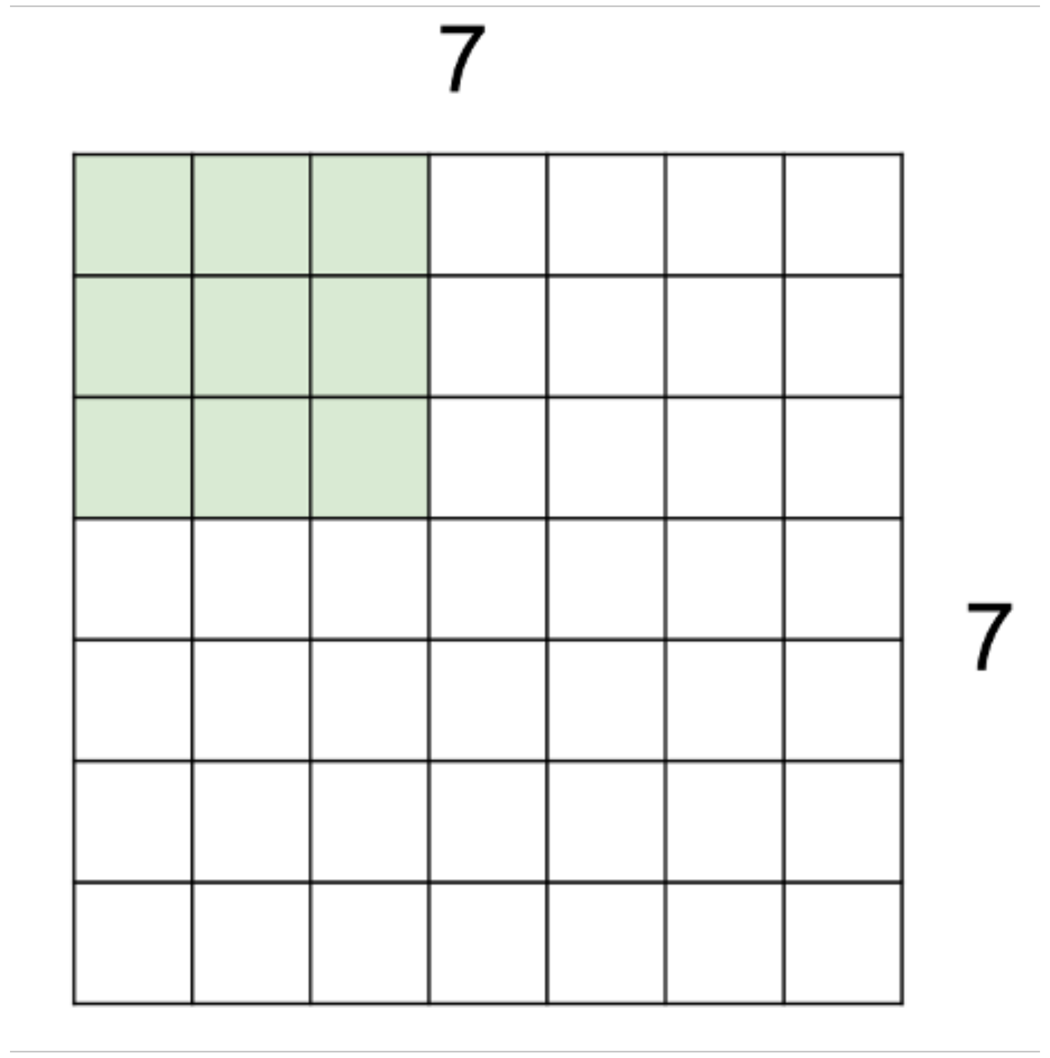
Stride

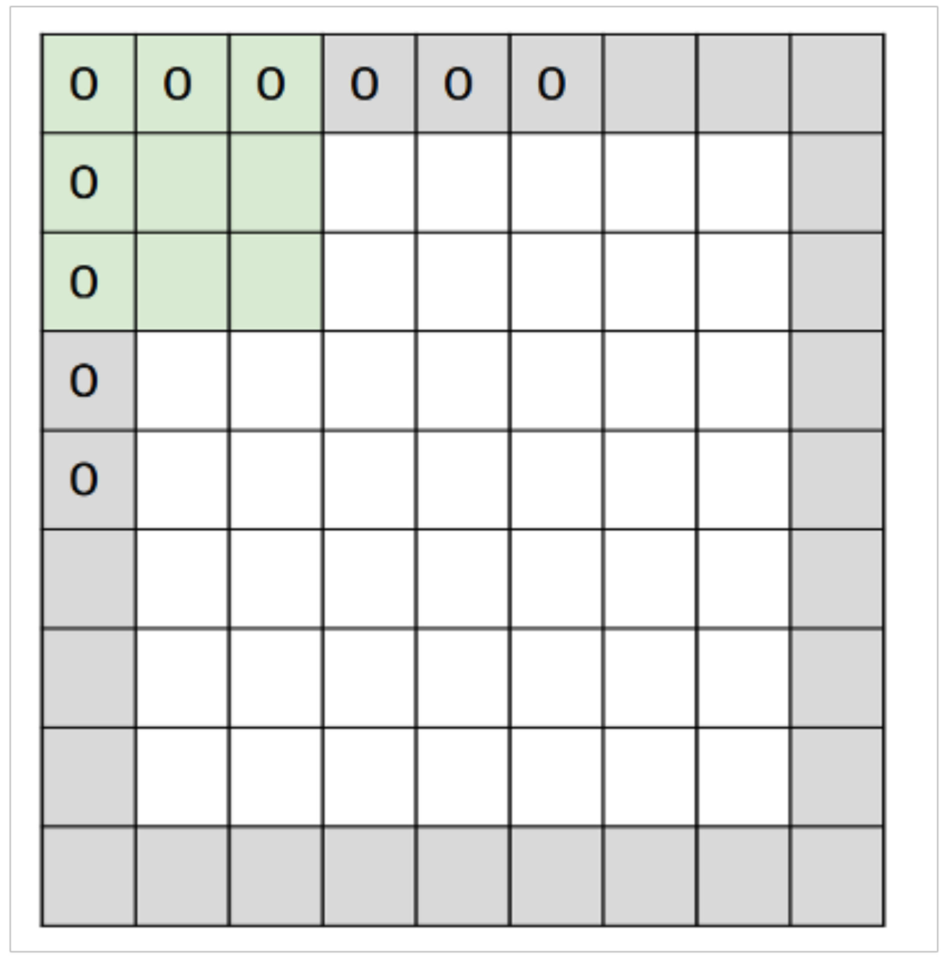
- How many pixels to step while sliding the filter
- Note that this shrinks the size if stride > 1



Padding

- no padding: ignore the last pixel if the filter does not fit
- add a row of zeros to make the slide fit.



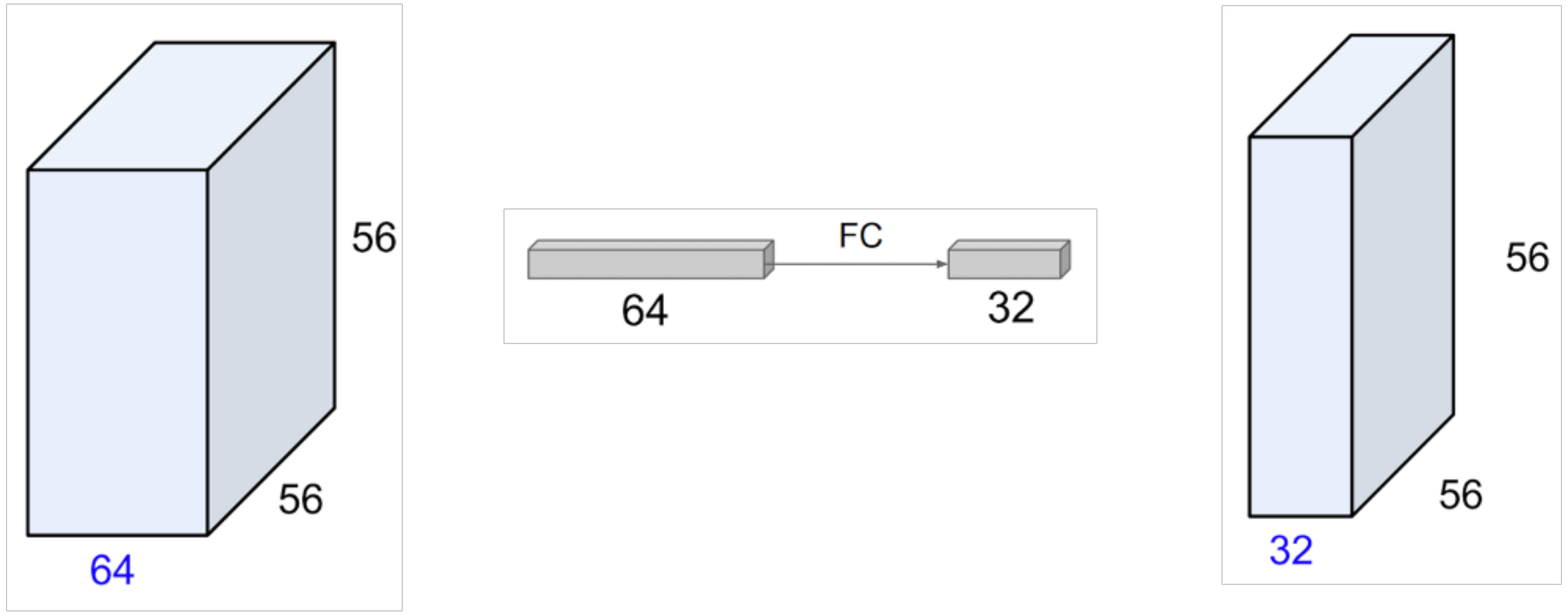


1x1 convolutions

- While the can not capture spatial patterns, they will capture patterns along the **depth dimension**

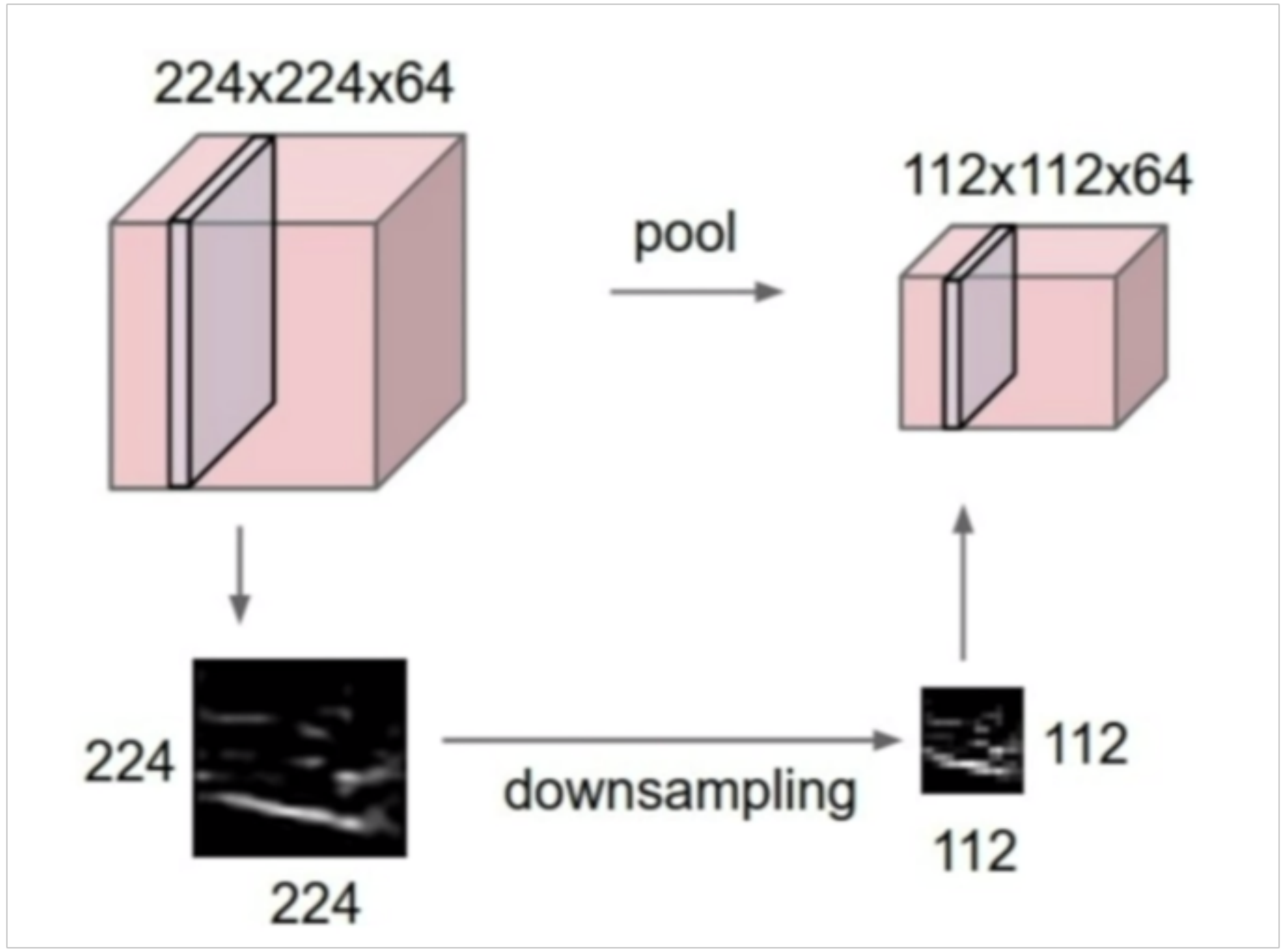
- When configured to reduce the depth, they **reduce dimensionality**

- We can add extra activations, adding more **non-linearities**.



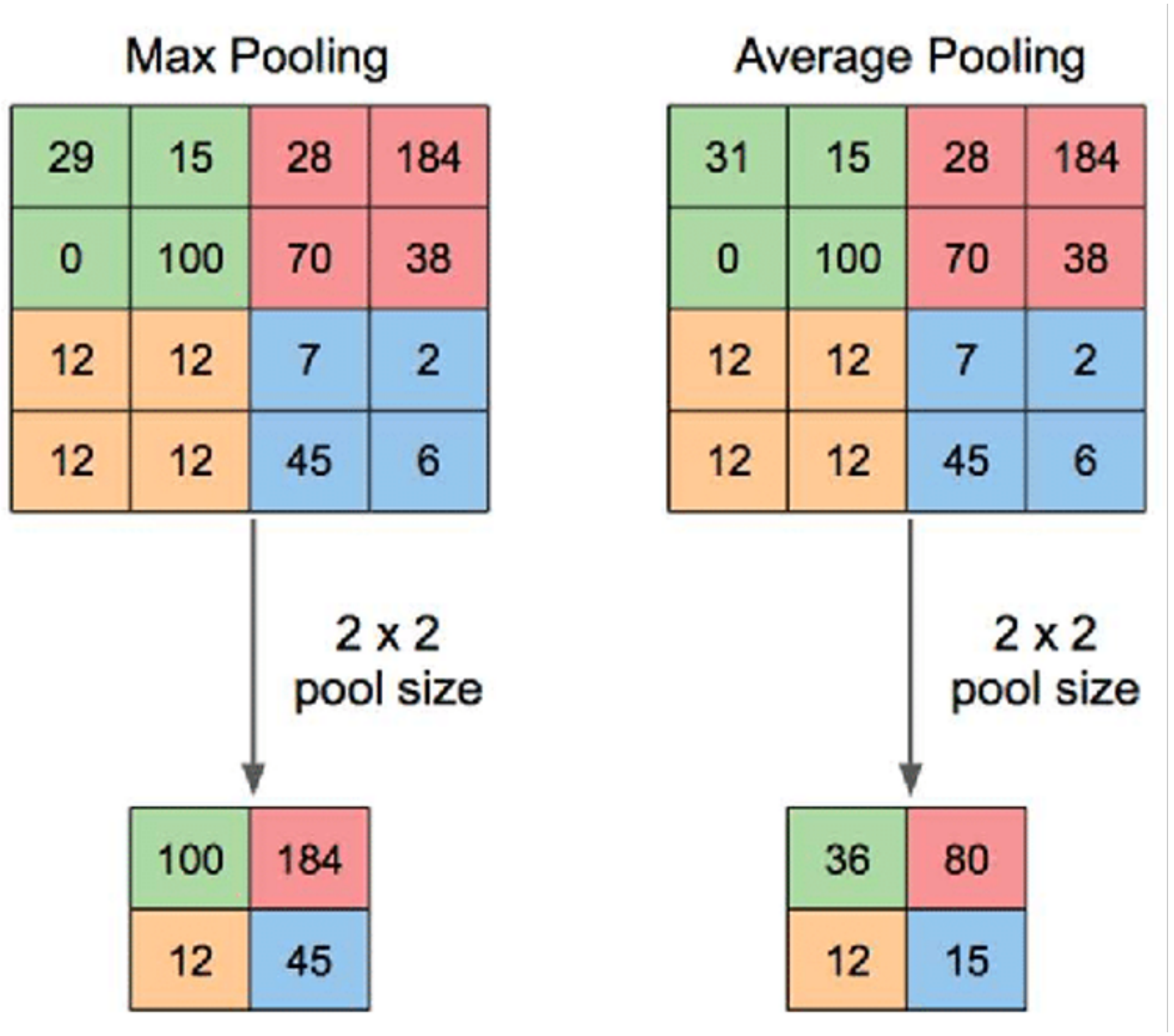
Pooling

- A way to downsample the input
- Convolutions with stride 2 also downsample, but pooling has no learnable parameters



Pooling

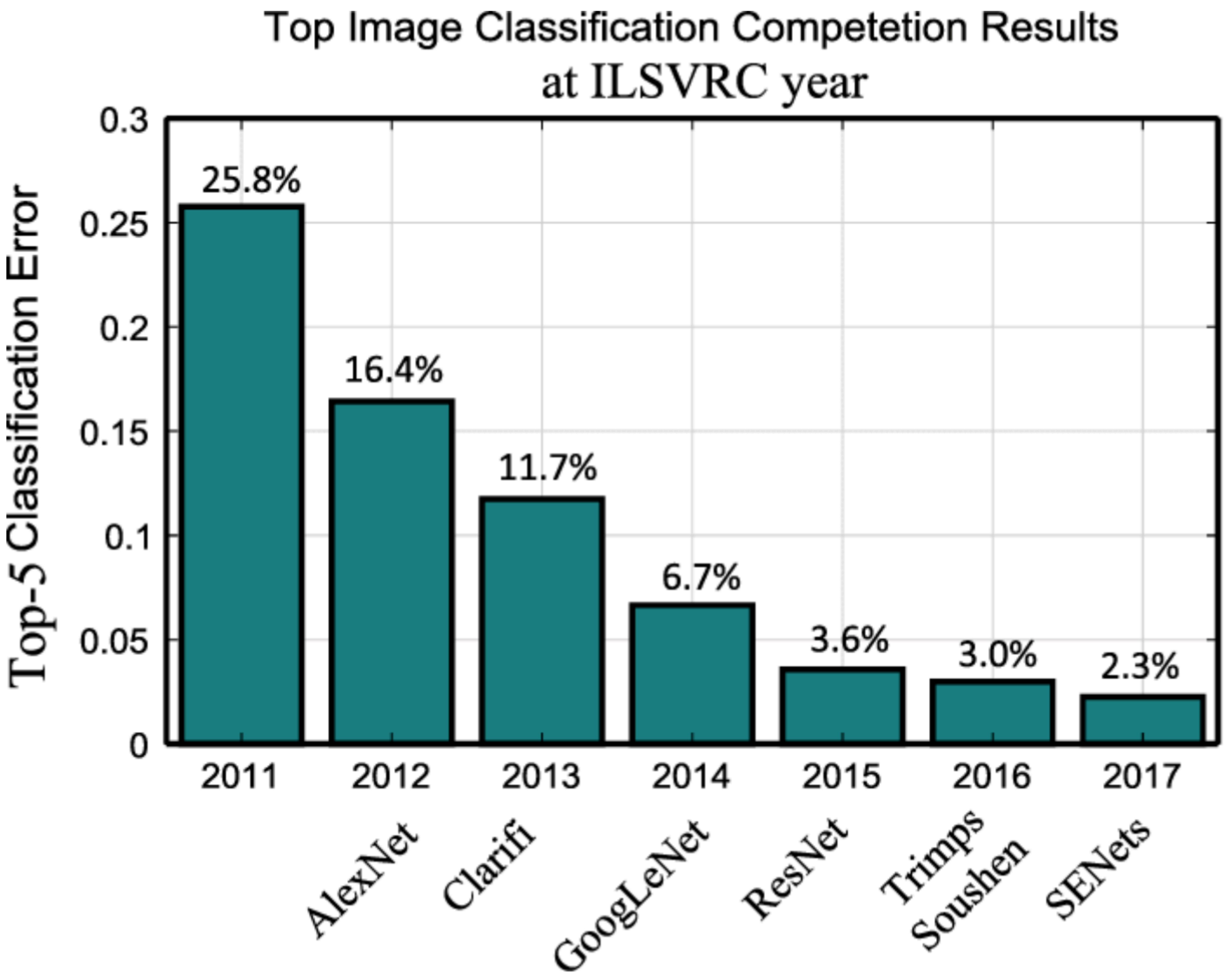
- max pooling: focus on the highest value
- average pooling: smooth the pixels



Imagenet Challenge

Superhuman performance

- Human performance is about 5%



Performance of winning entries in the ILSVRC competitions from 2011 to 2017 in the image classification task

AlexNet

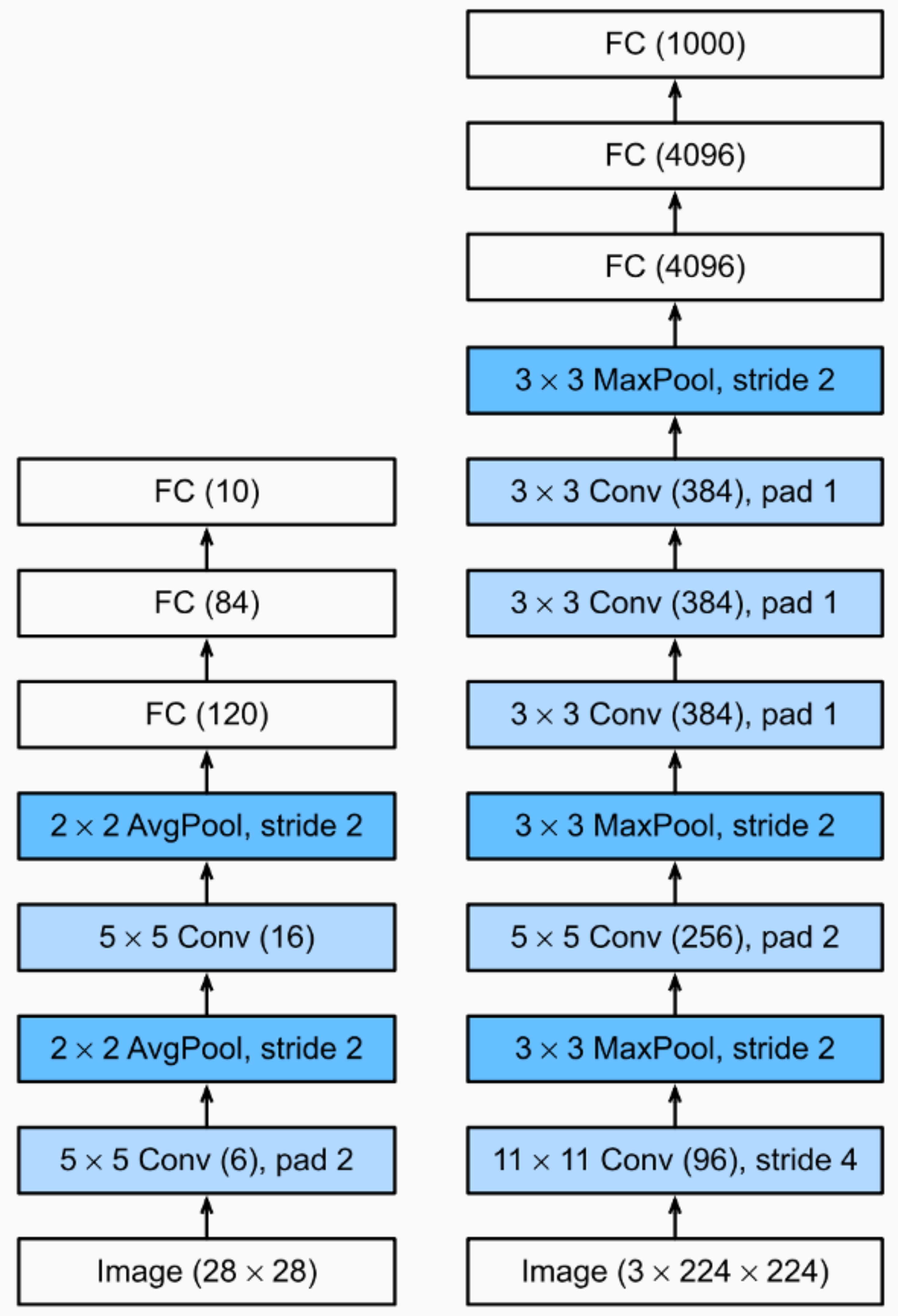


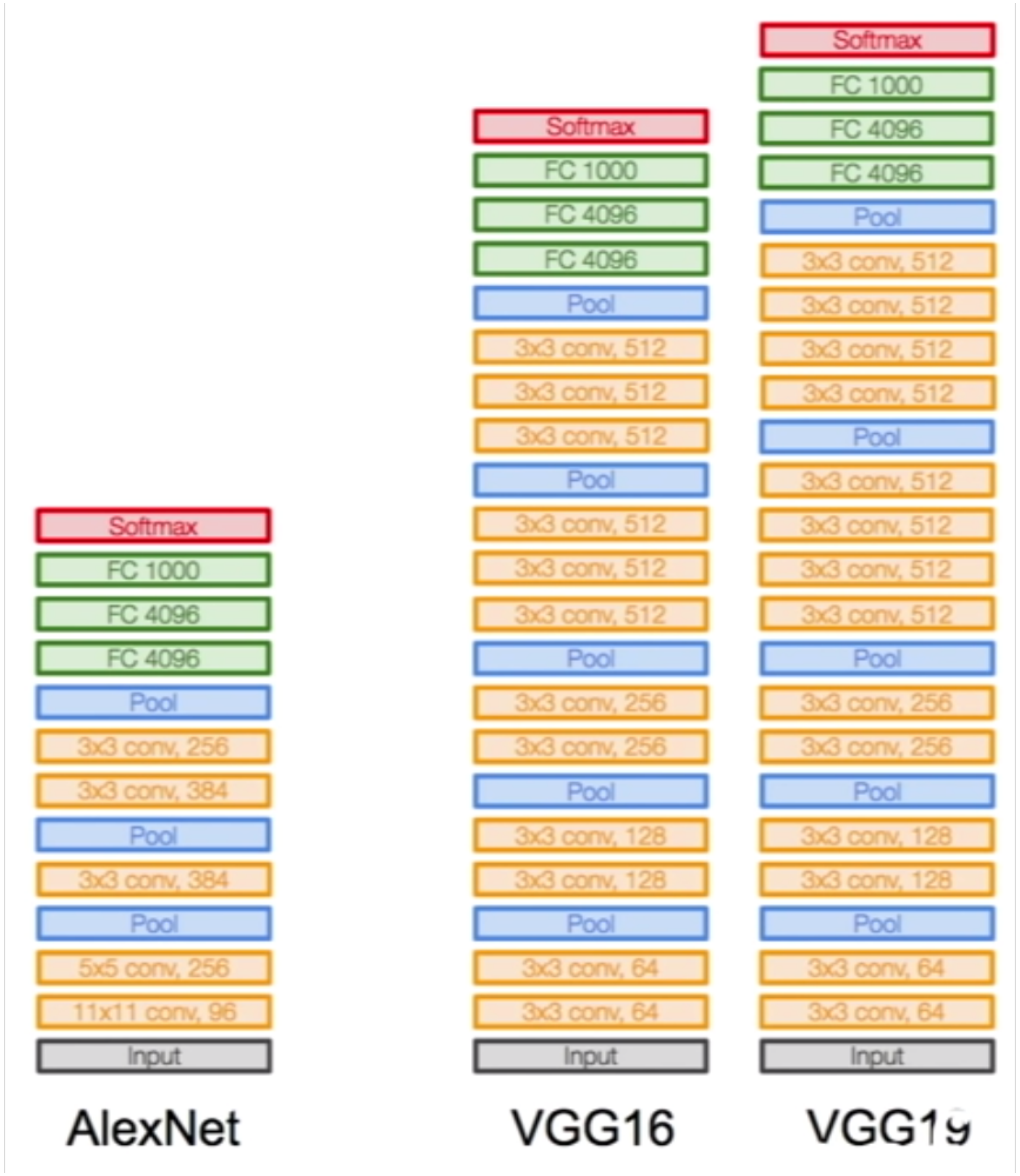
Fig. 7.1.2 From LeNet (left) to AlexNet (right).

VGG

Changes in:

- Filter size
- channel numbers
- depth

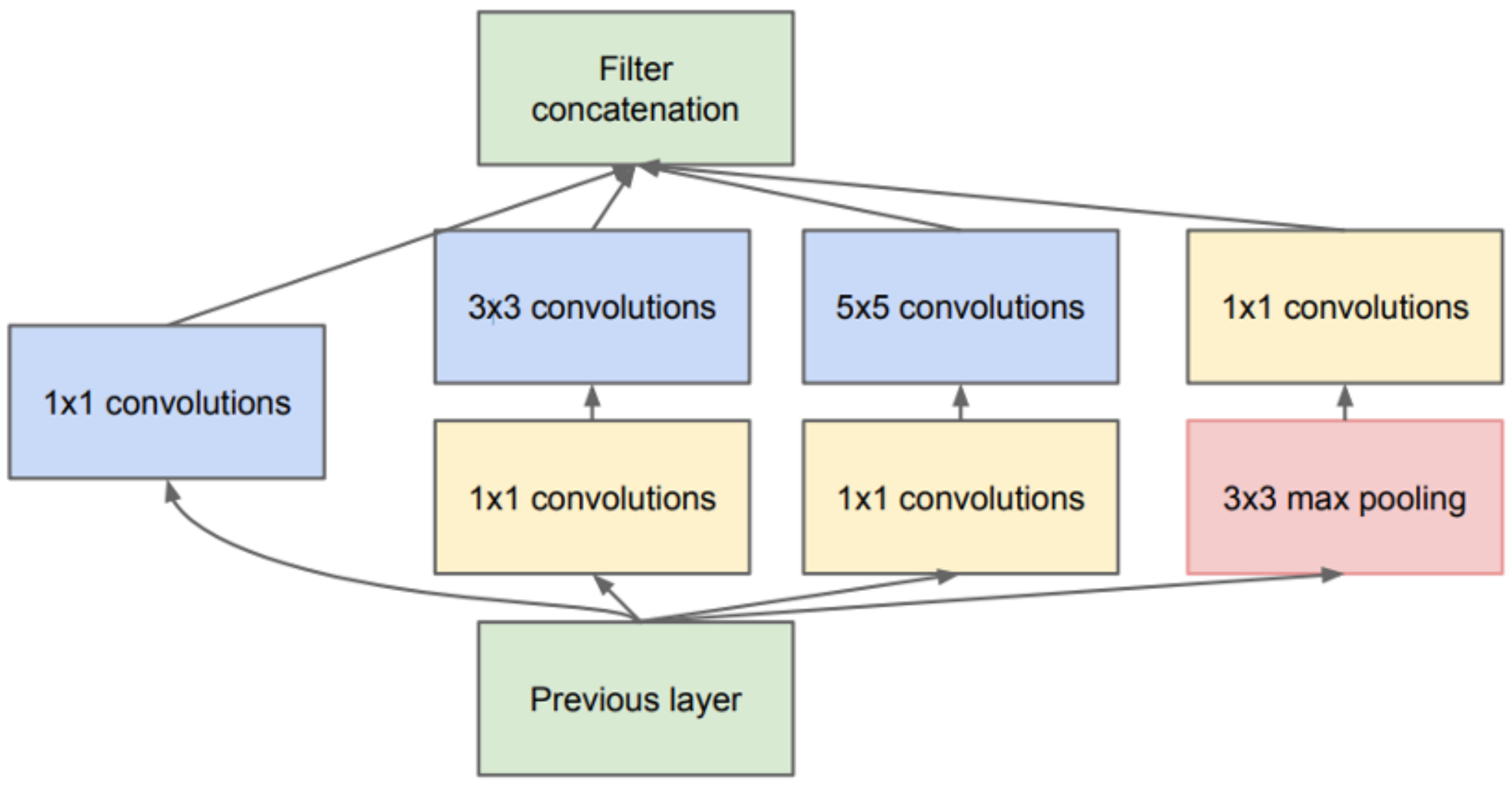
138 million parameters



GoogleNet

- Parallel filters
- Bottleneck layers

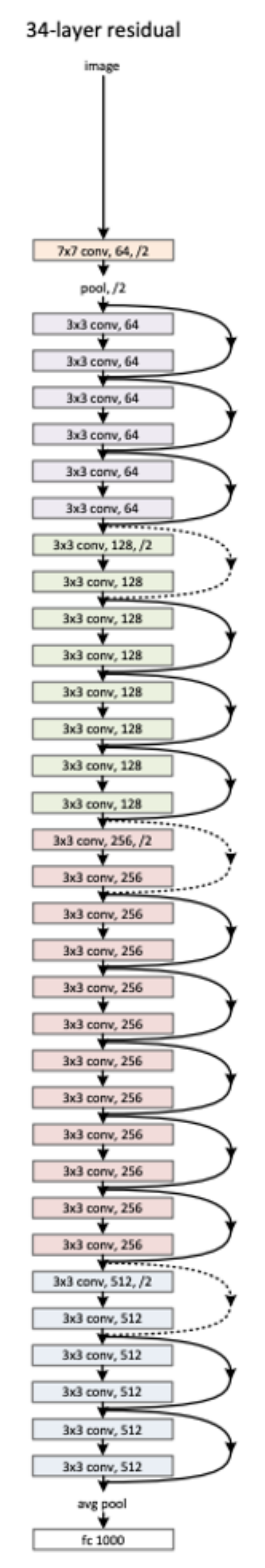
6.7 million parameters



ResNet

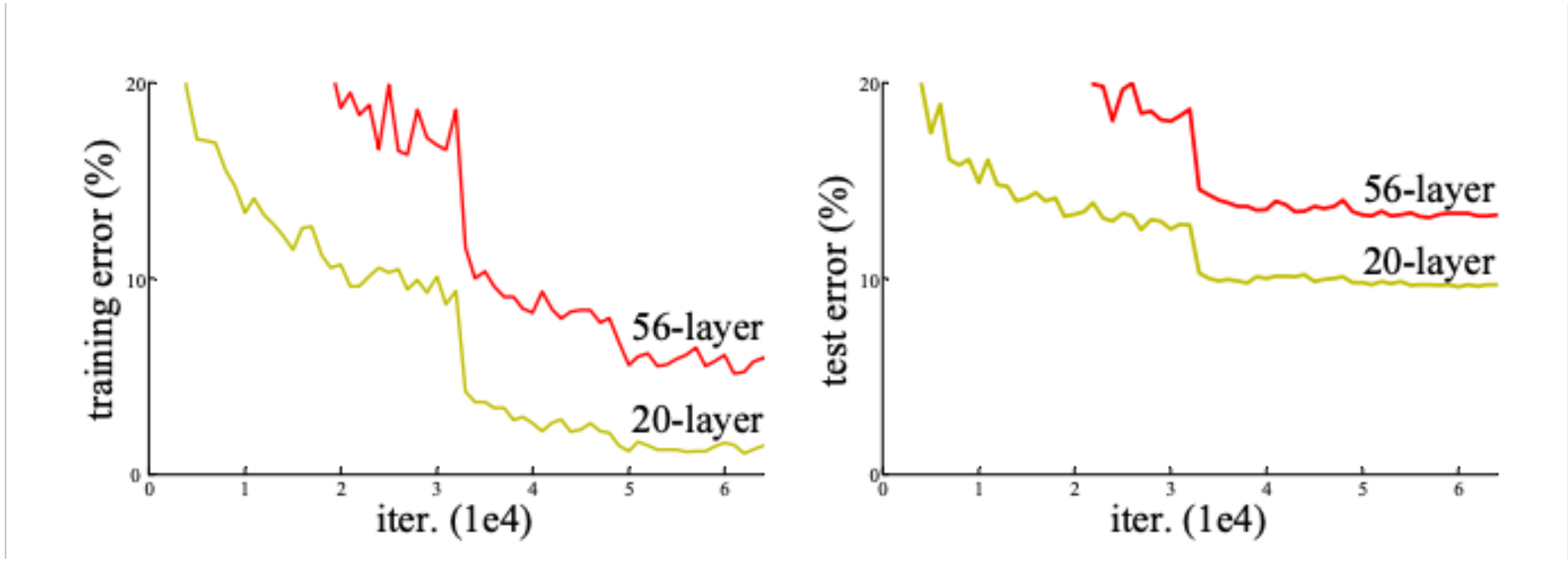
- batch normalisation
- Residual blocks

11 million parameters
for ResNet18



ResNet

- **What they noticed**: a deeper net (56 layers) was performing worse than a shallower net (20 layers), but not because of overfitting
- **Hypothesis:** a deeper net is harder
- **Idea**: don't learn the mapping, but learn the residual
- **Residual**: what you need to change from the previous representation to achieve the mapping

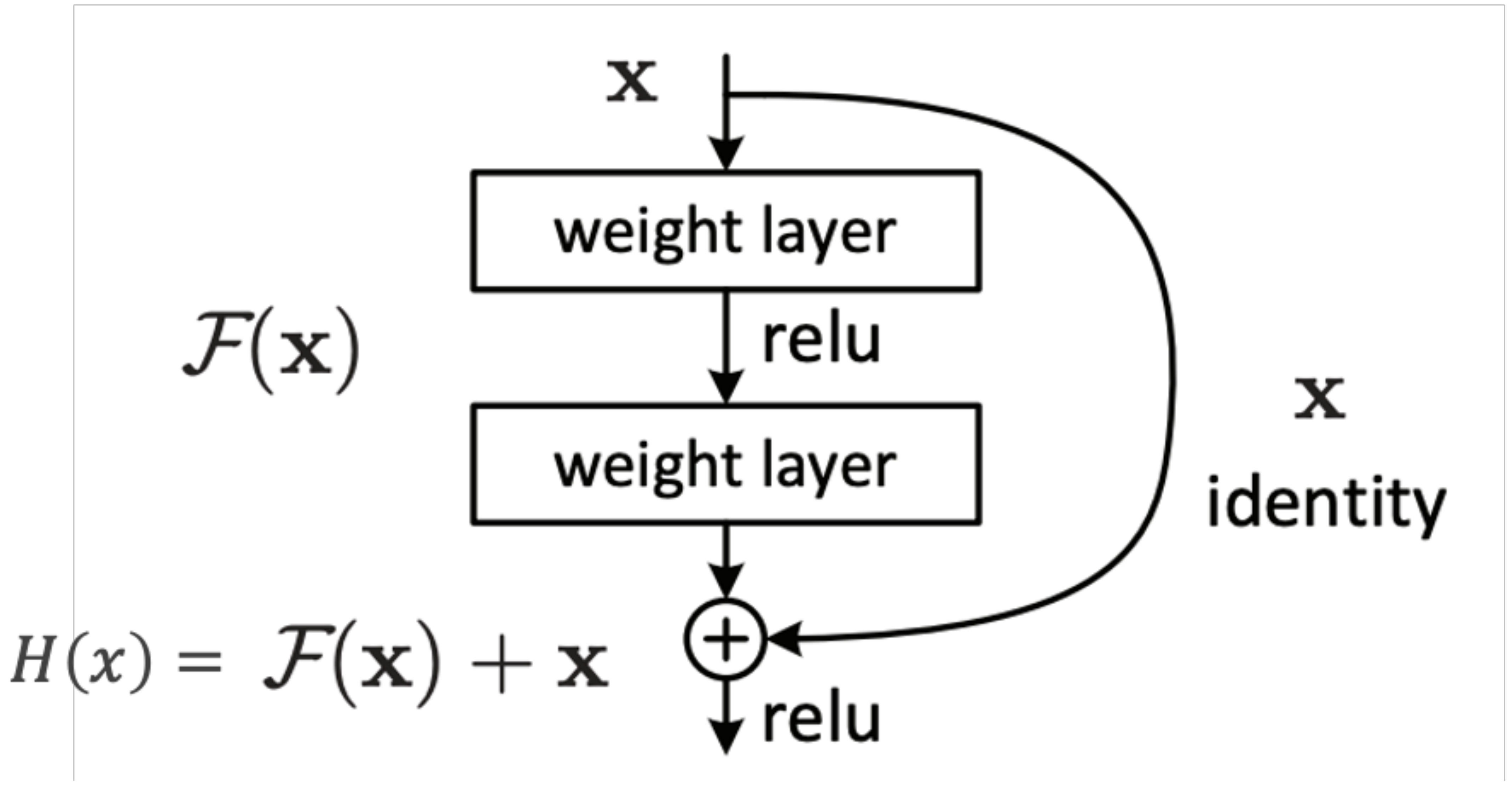


ResNet

Learn the residual, instead of the full mapping

E.g.,“use the same settings, but just change x” is easier than starting from scratch.

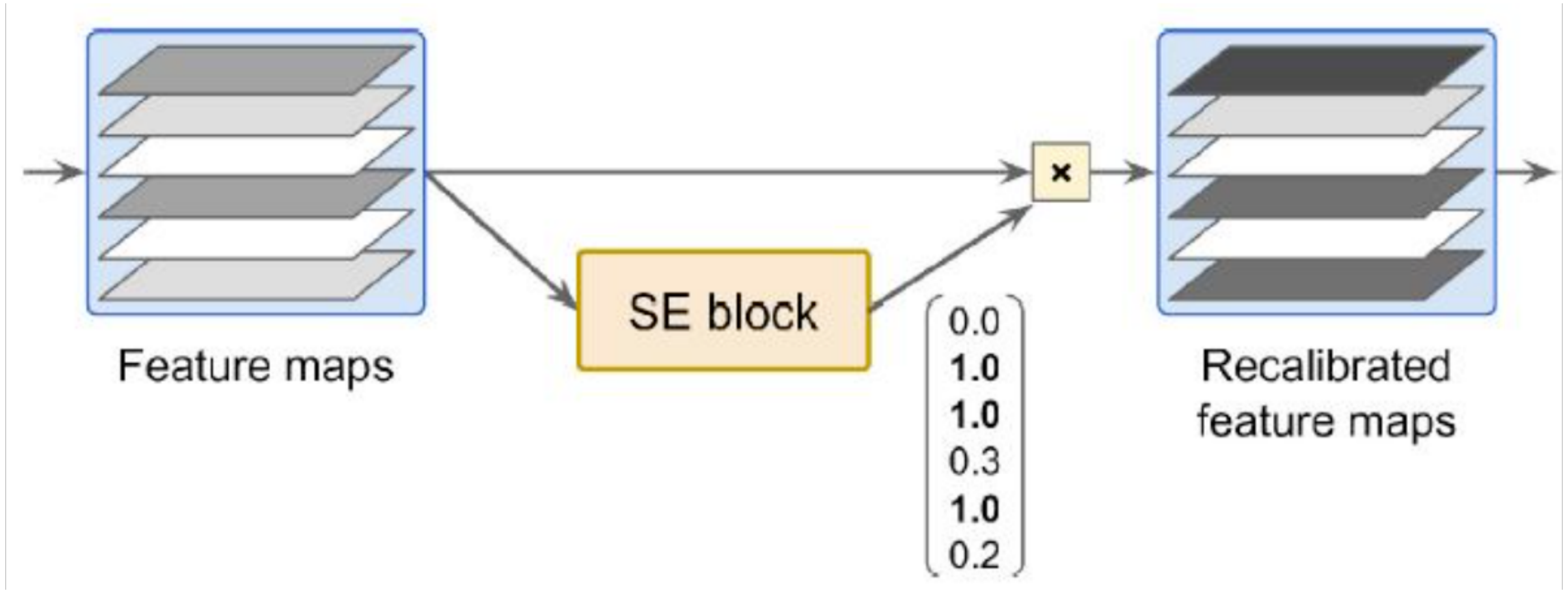
It is easier for the gradient to flow back



SEnet

Squeeze and excite

- Which features are most likely to be activated together?
- E.g., if you see a left eye, you expect a right eye. Even if it is hidden in the image.



SEnet

Squeeze and excite

- **global average pool**: calculates the average of each activation map
- **Squeeze**: reduce dimensionality, typically by a fraction of 16
- **Excite**: restore the amount of dimensions.

The process can be compared to writing an abstract of a text.