

Image recognition

deep learning 2

Raoul Grouls, 23 nov 2023

Recap

$$X = \{\vec{x}_1, \dots, \vec{x}_n \mid \vec{x} \in \mathbb{R}^d\}$$

Data

$$y = \{y_1, \dots, y_n \mid y \in \{0,1\}\}$$

Trainable Weights W, b

$$\text{(Non)linearity} \quad f(X) = WX + b \quad \sigma(X) = \max(0, X)$$

Predict

$$\hat{y} = f_n \circ \sigma \circ f_{n-1} \circ \dots \circ \sigma \circ f_1$$

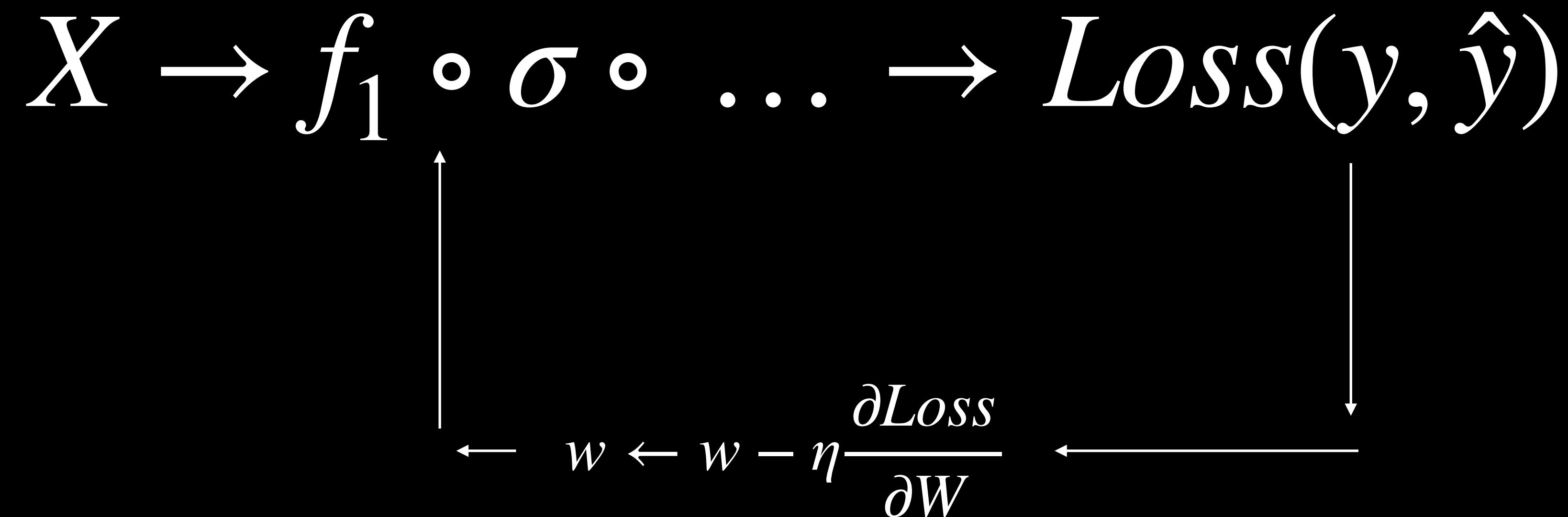
Loss

$$Loss(y, \hat{y})$$

Optimize

$$w \leftarrow w - \eta \frac{\partial Loss}{\partial W}$$

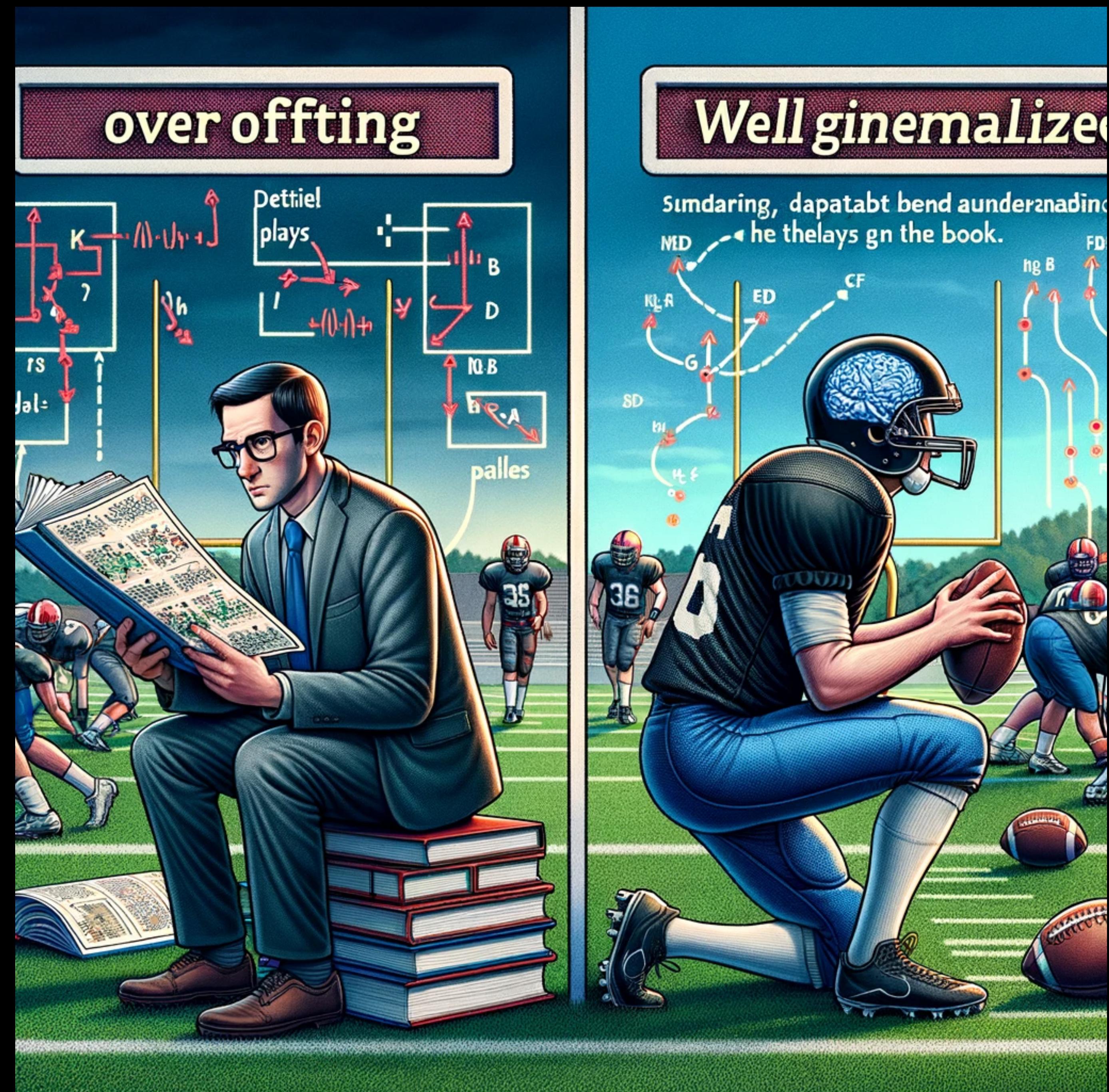
Recap



Overfitting

A model that corresponds too closely to a particular set of data, therefore failing to fit to unseen data.

- the error / performance metrics for the training data are very good, but bad on unseen data.
- Overfit models tend to have higher complexity (number of learnable parameters) than is warranted by the data complexity
- Small changes in the data can lead to large changes in the model.



Regularization

$$J(y, \hat{y} | W) = Loss(y, \hat{y}) + \eta \cdot Penalty(W)$$

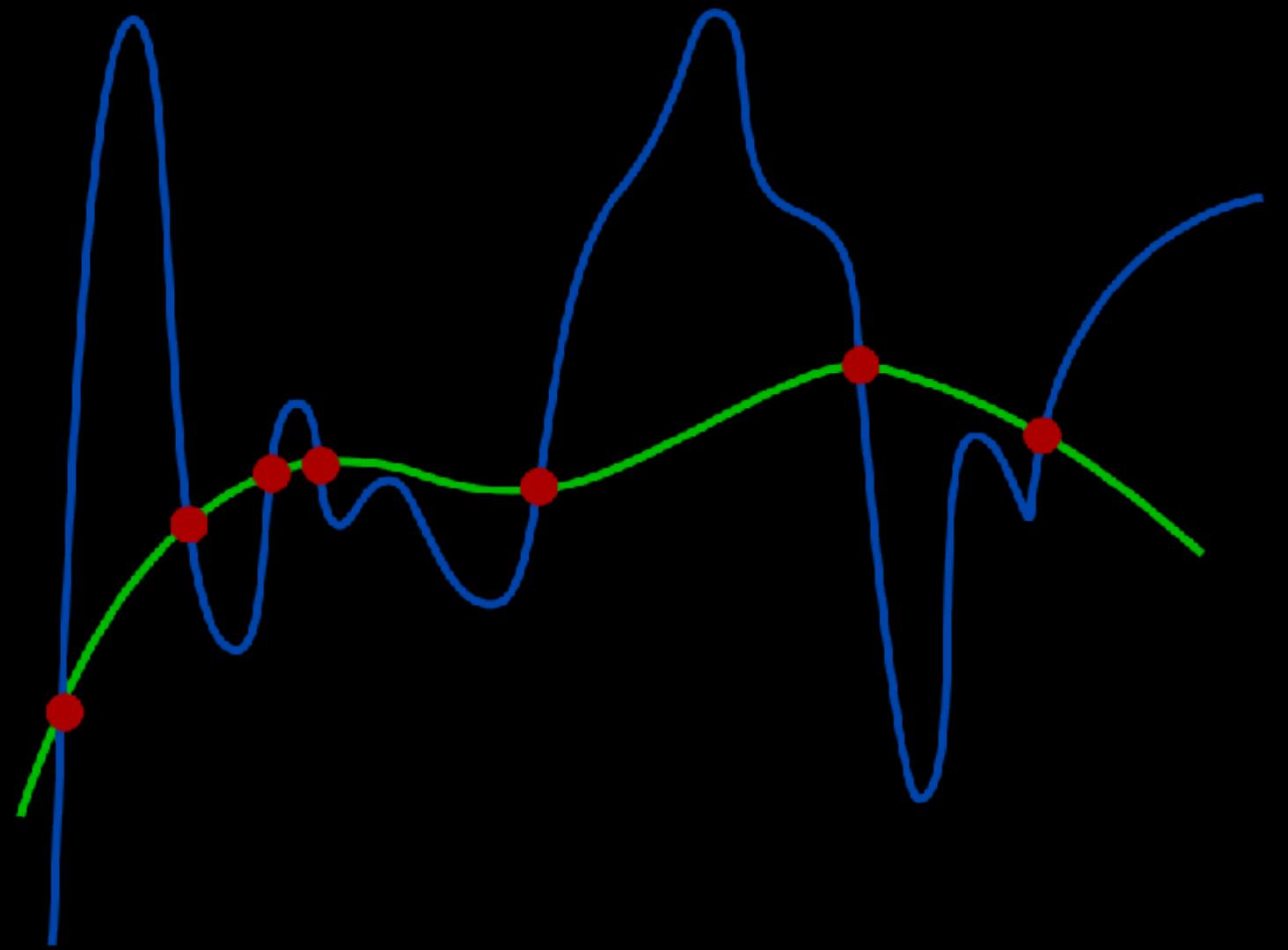
A process used to enforce simpler models to prevent overfitting.

Explicit Regularization:

- Involves directly adding a term to the optimization problem.
- These terms can include priors, penalties, or constraints.
- Imposes a cost on the optimization function.

Implicit Regularization:

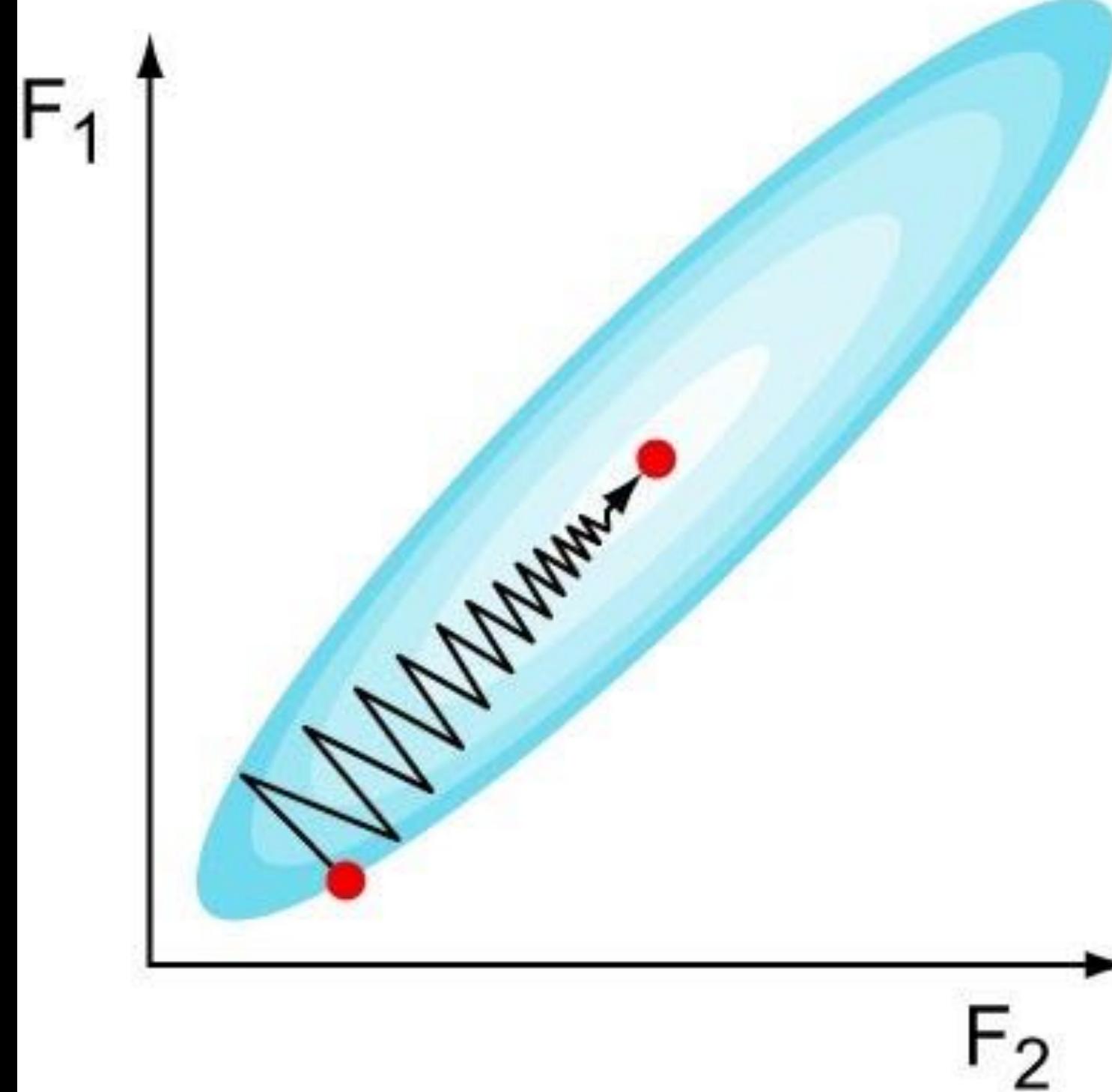
- All methods that are not directly added to the optimization problem.
- Examples include early stopping, introducing noise (eg with random batches, dropout), and discarding outliers.
- Ubiquitous in machine learning, including batched stochastic gradient descent and ensemble methods like random forests and gradient boosted trees.



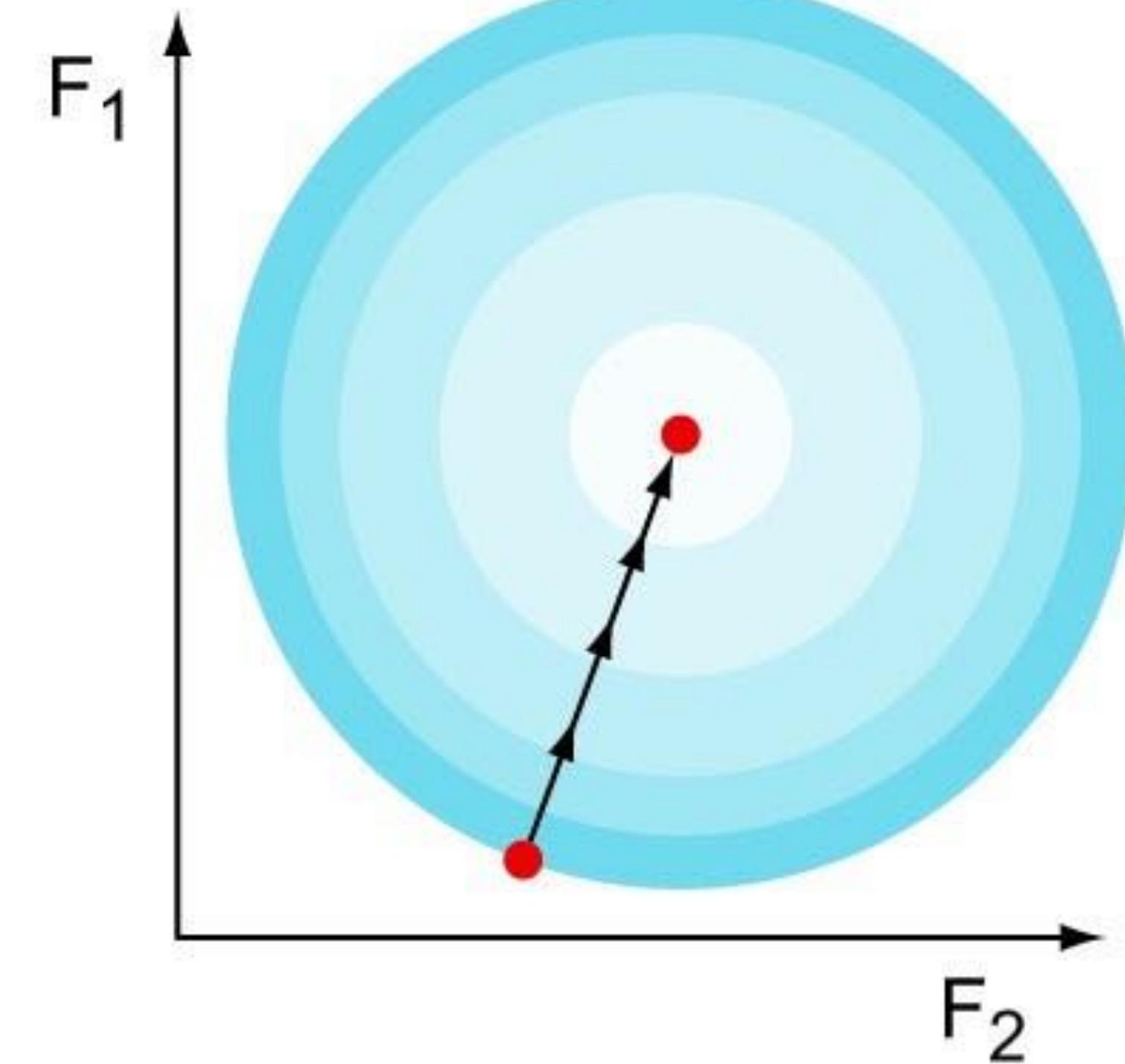
Normalization

Gradient descent with and without feature scaling

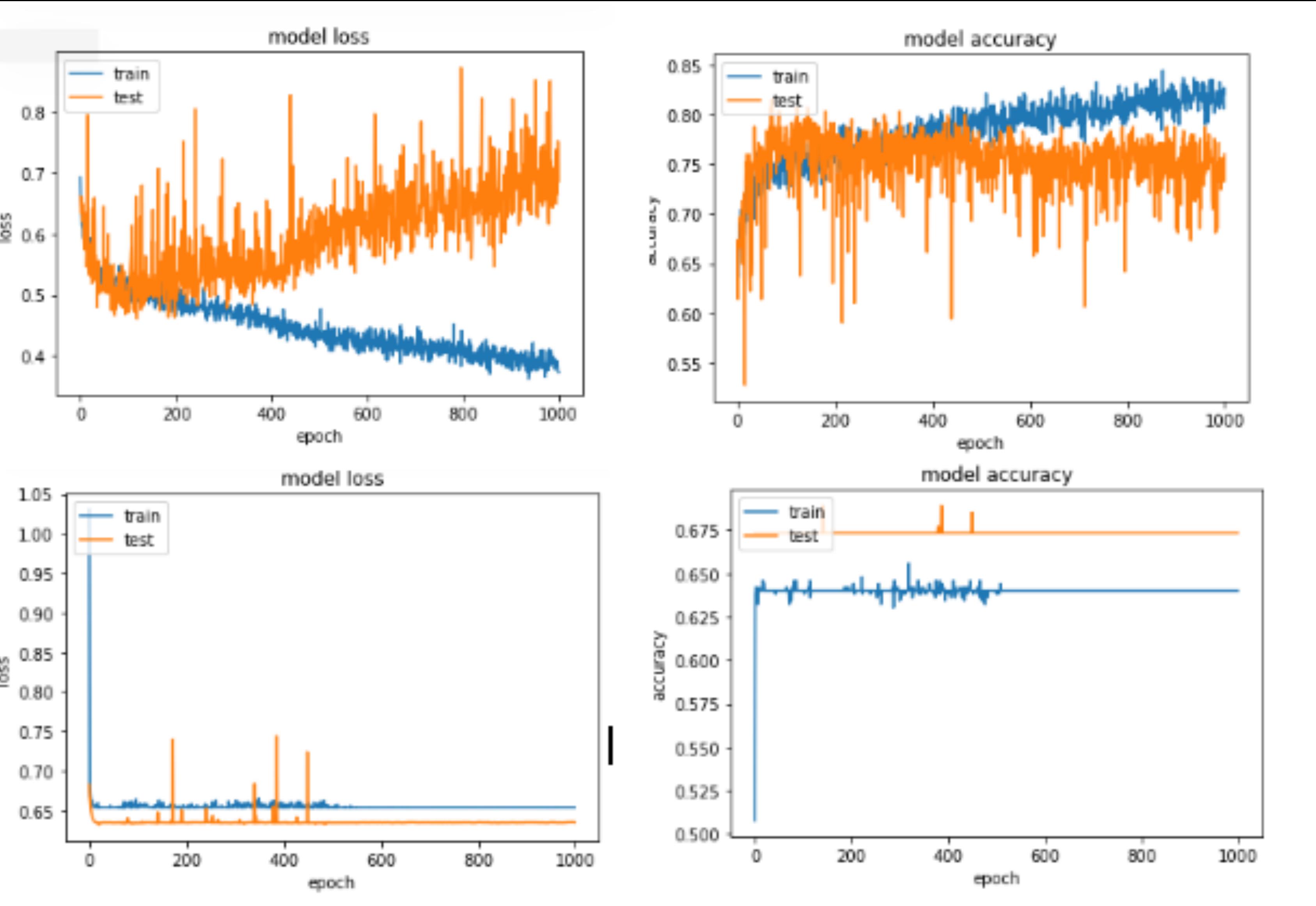
Non-normalized features



Normalized features



Spotting overfitting & underfitting



Dropout

- Randomly drop weights in order to avoid overspecialised parameters
- $\hat{w}_j = \begin{cases} w_j & \text{with } p(\theta) \\ 0 & \text{otherwise} \end{cases}$





Convolutions - the motivation

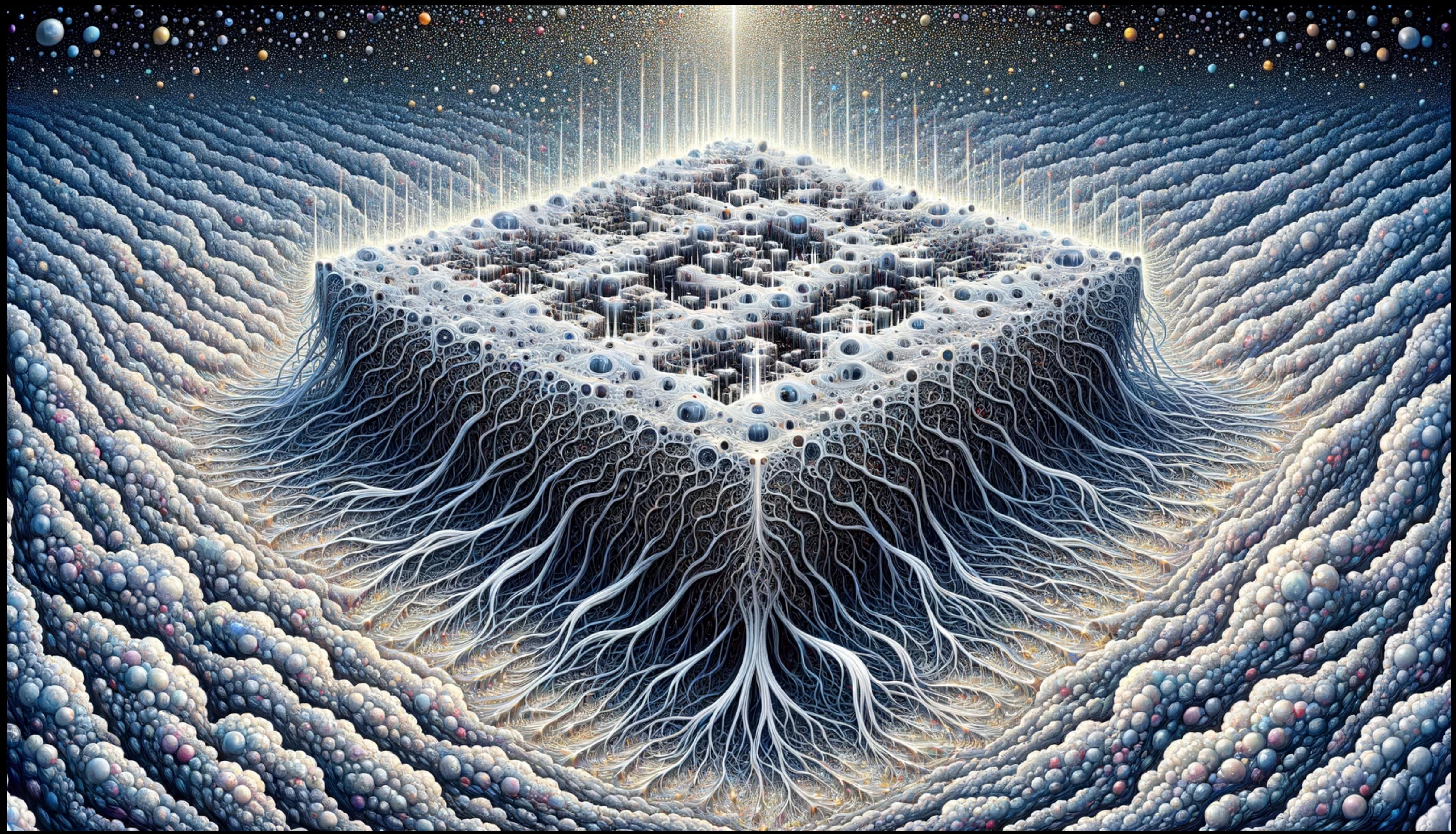
An image of size 28x28 has 784 pixels

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...



chatGPT on its visualisation

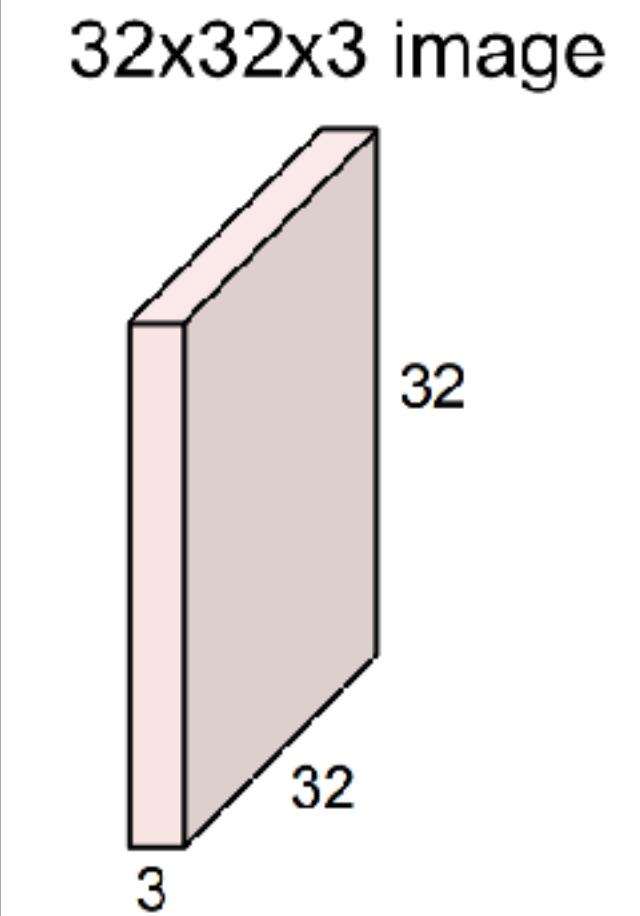
Creating a visualization of how a Convolutional Neural Network (CNN) "sees" images is a fascinating concept, but it's important to clarify that such visualizations are more like interpretations rather than literal views. A CNN processes images in a hierarchical manner, where early layers detect simple features like edges and colors, and deeper layers recognize more complex patterns or parts of objects.

To depict this, I'll create an artistic representation. The image will show a progression from the original input image on one side, transitioning through various stages of abstraction and feature detection, to represent the layers of a CNN. This transition will show how the image gets broken down into simpler features and then reconstructed into high-level representations. Let's create this visualization!

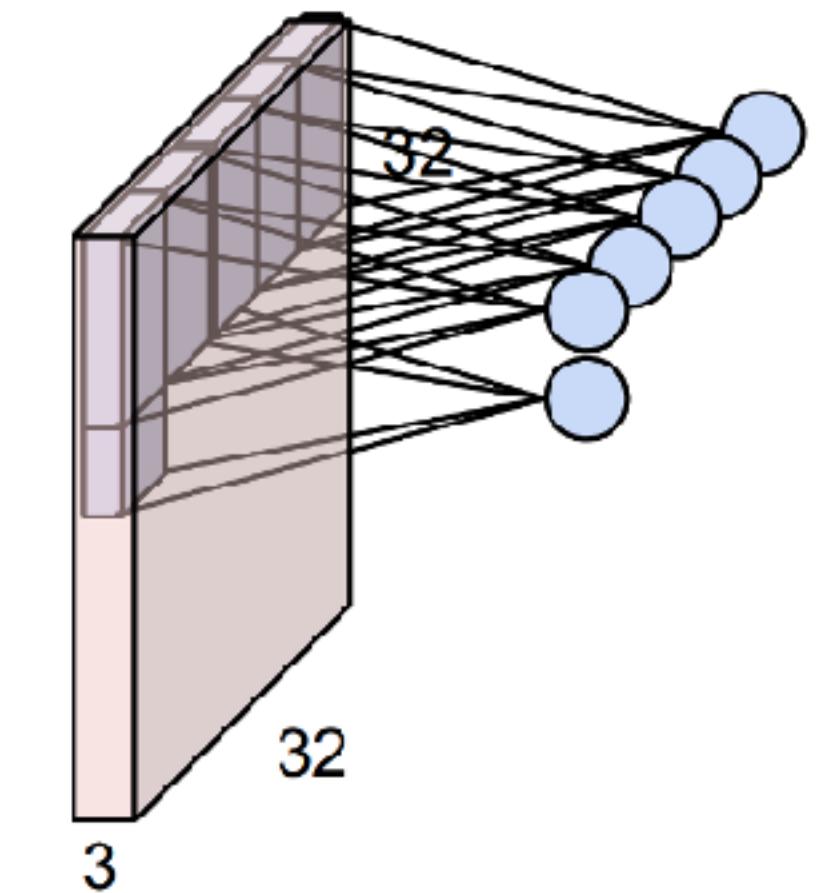
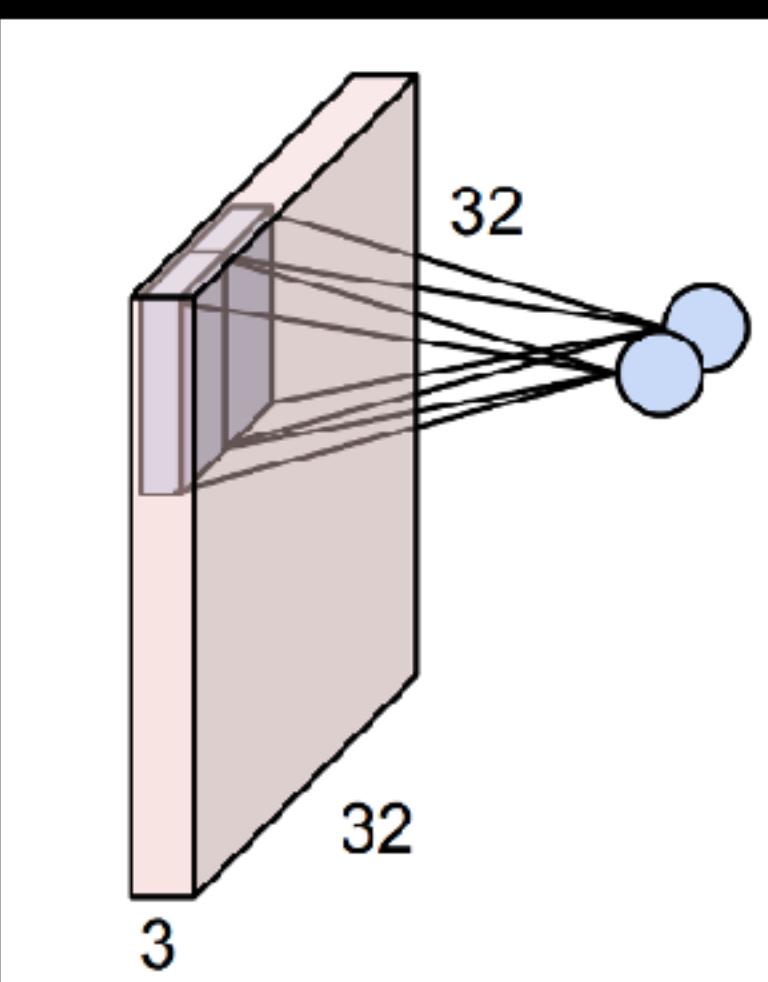
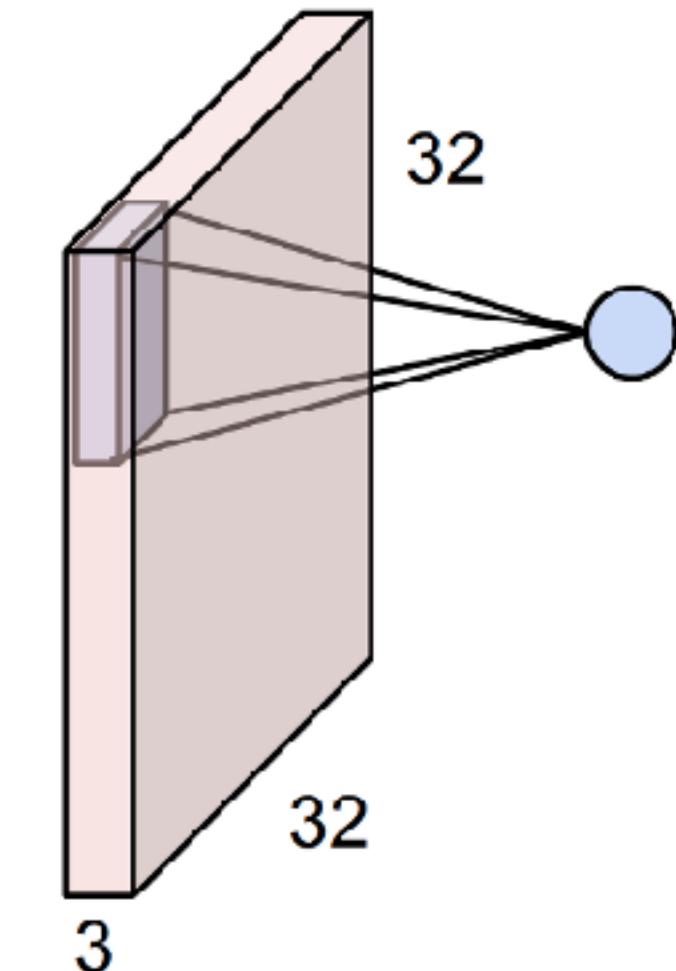


Convolutions

- Take a filter of size $(5 \times 5 \times 3)$
- This needs just 75 parameters
- Slide it over the image, like a scanner



5x5x3 filter

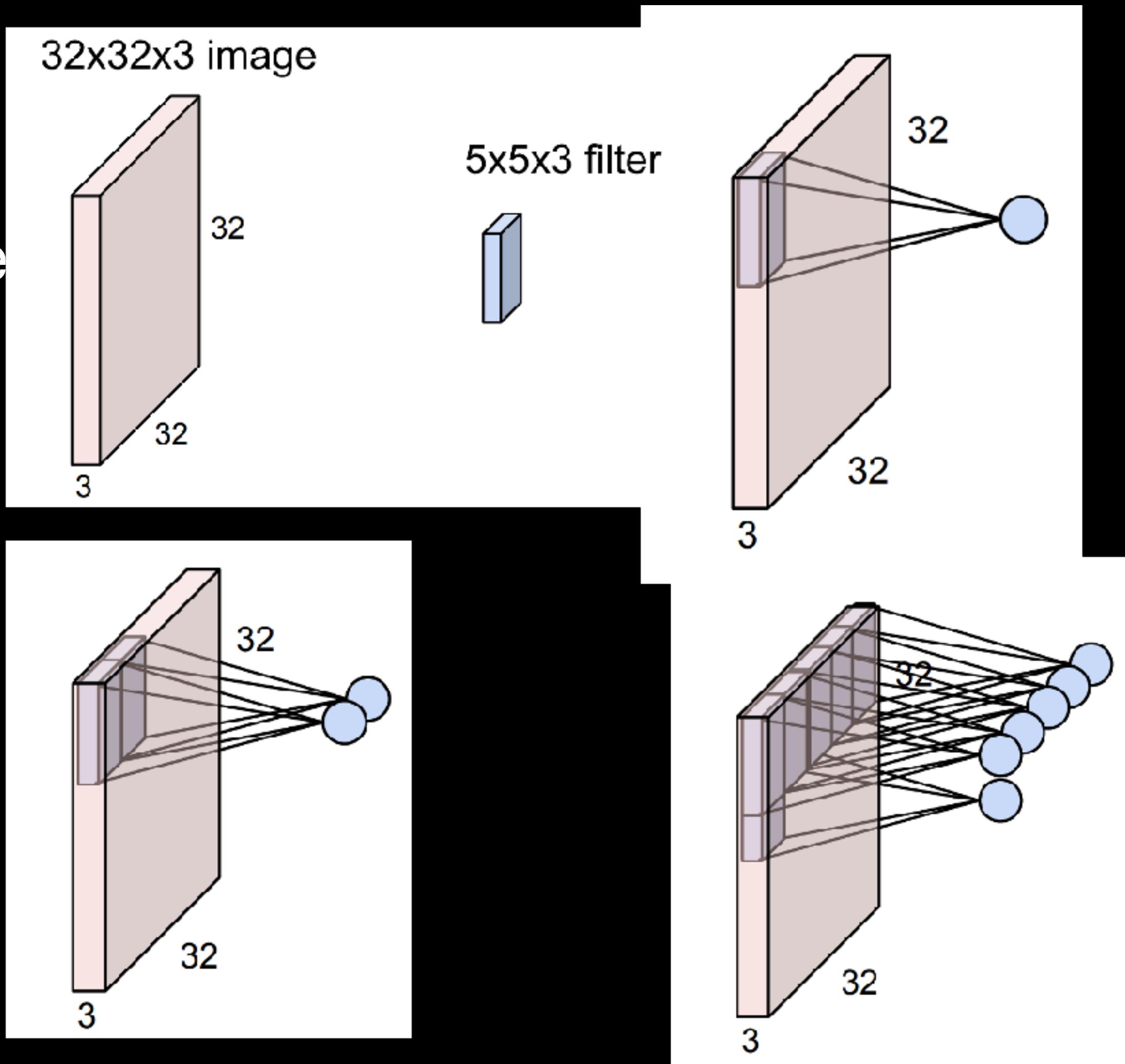


Convolutions

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

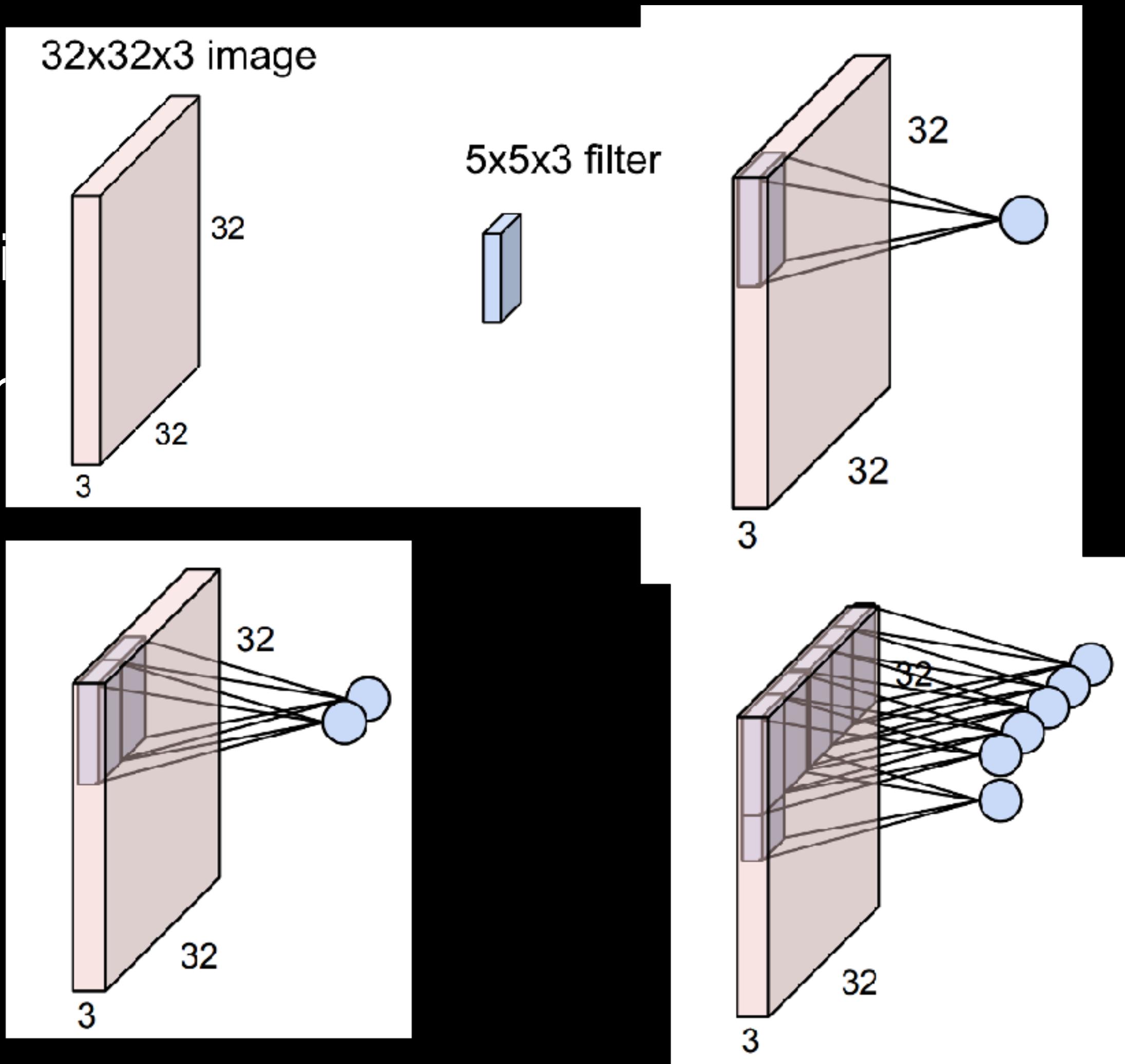
$$\begin{bmatrix} 1 & 10 \\ 2 & 20 \end{bmatrix} \cdot \begin{bmatrix} 1 & 2 \\ 2 & 0 \end{bmatrix} =$$

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



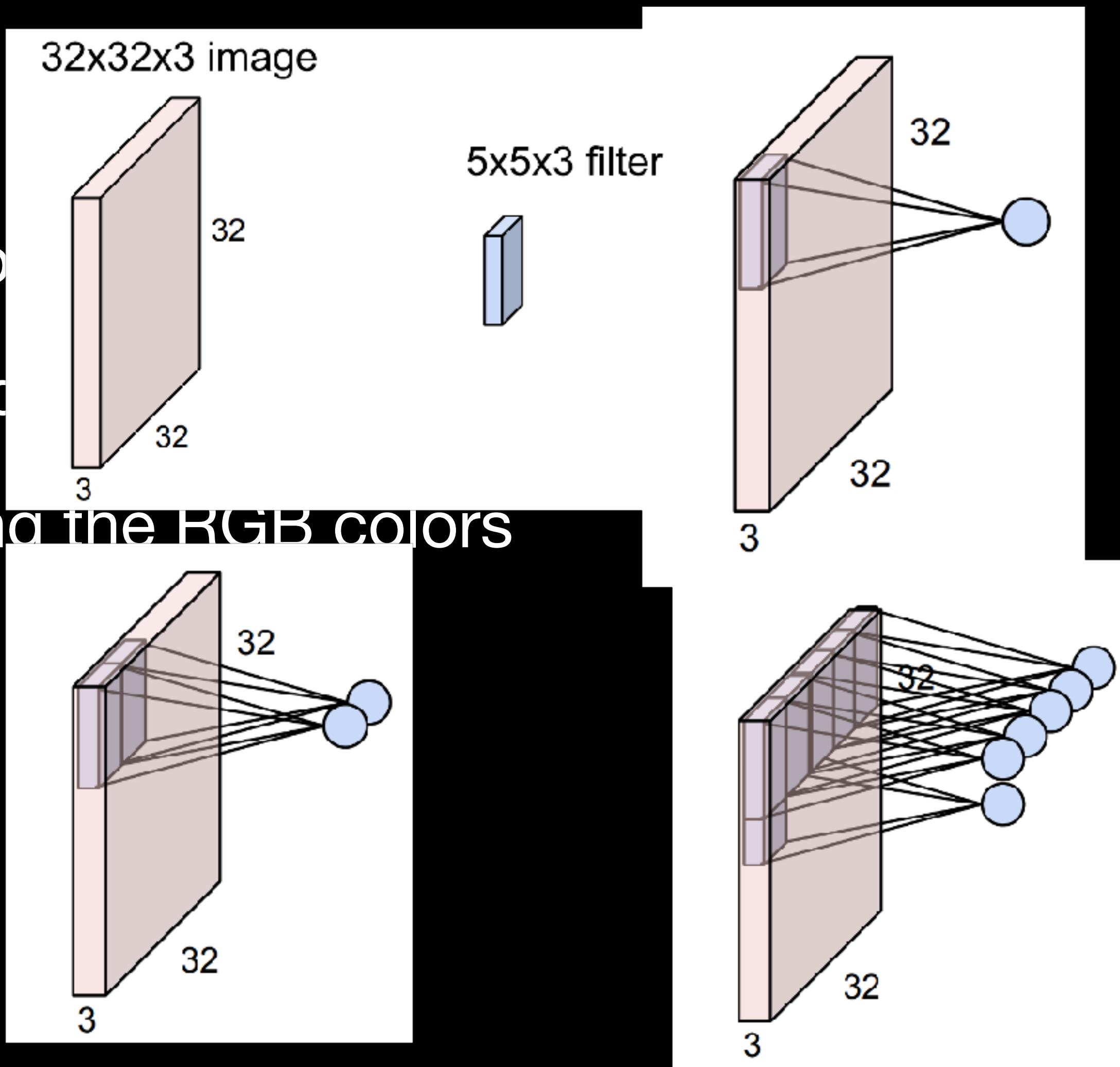
Convolutions

- So, with a 5×5 filter, every 5×5 image slice is multiplied by a number.
- This number contains weighted information.
- We call the result an activation map.



Convolutions

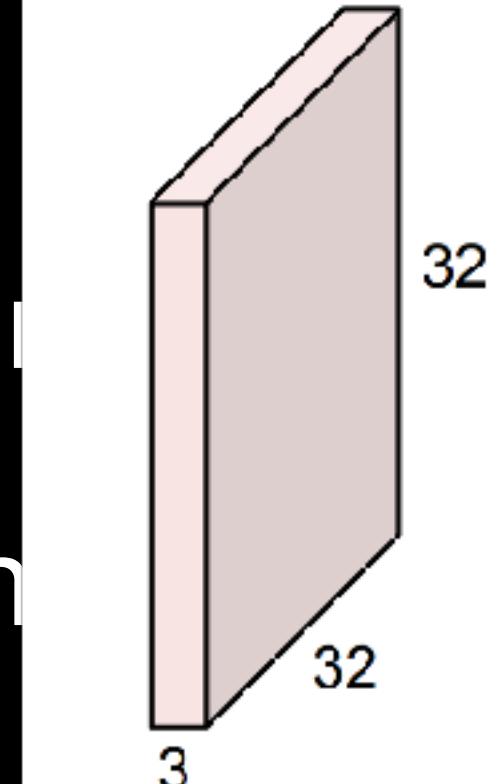
- The filter goes through the depth of the input
- Parameters of the filter are: width, height, depth
- In the first case, the depth is 3, representing the **RGB colors**



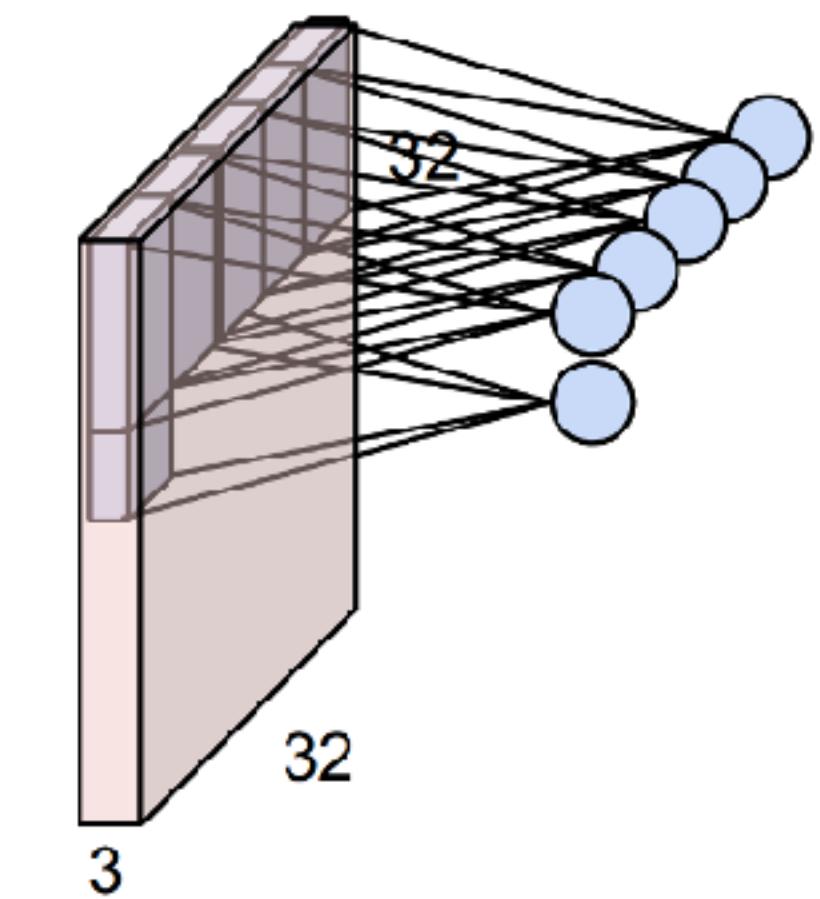
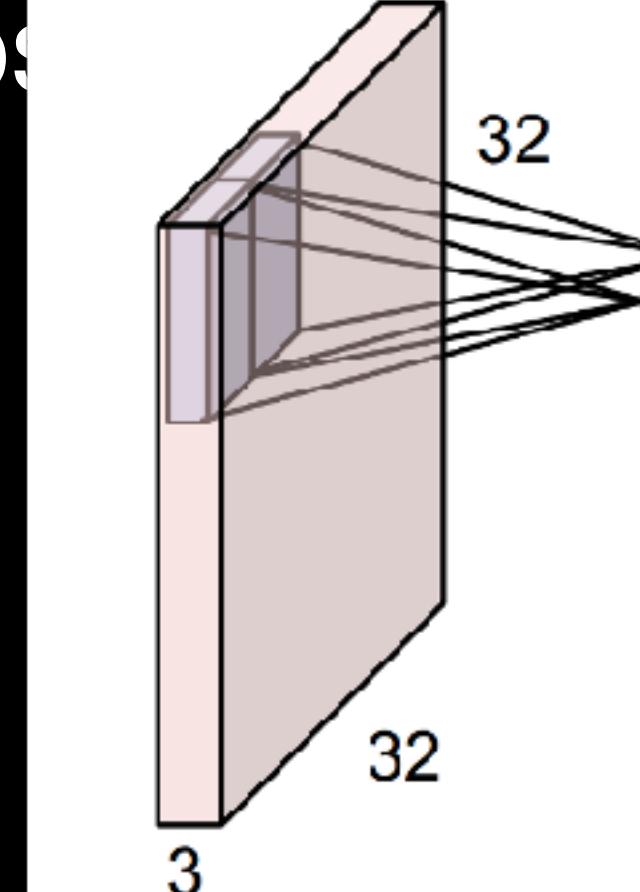
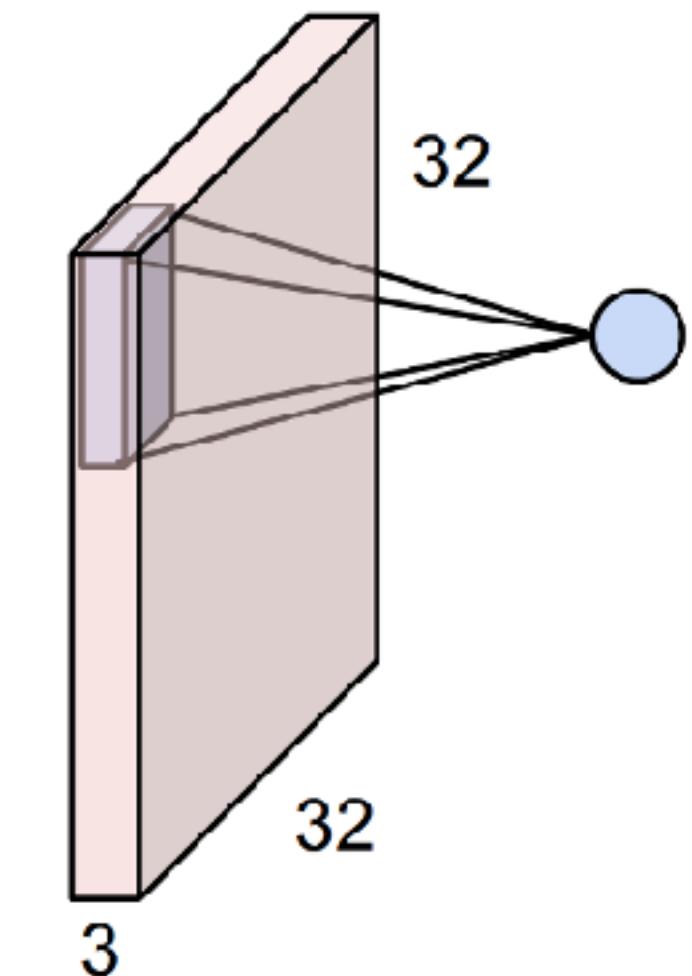
Convolutions

- Later on, the channel represents the amount of light at a certain wavelength.
- Typically, you might start with a matrix with dimensions $32 \times 32 \times 3$.
- After a few layers of convolutions, you end up with dimensions $32 \times 32 \times 512$, which means there are 512 activation maps.

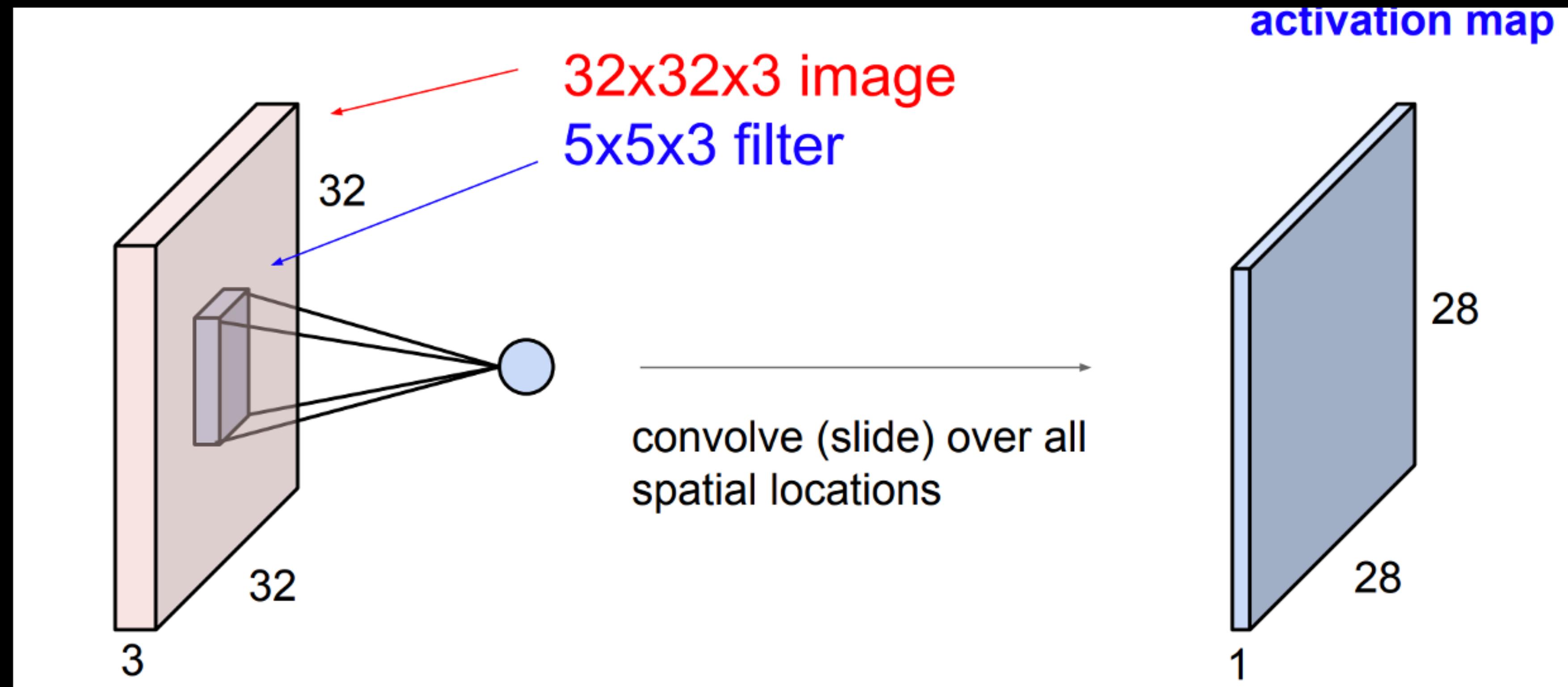
32x32x3 image



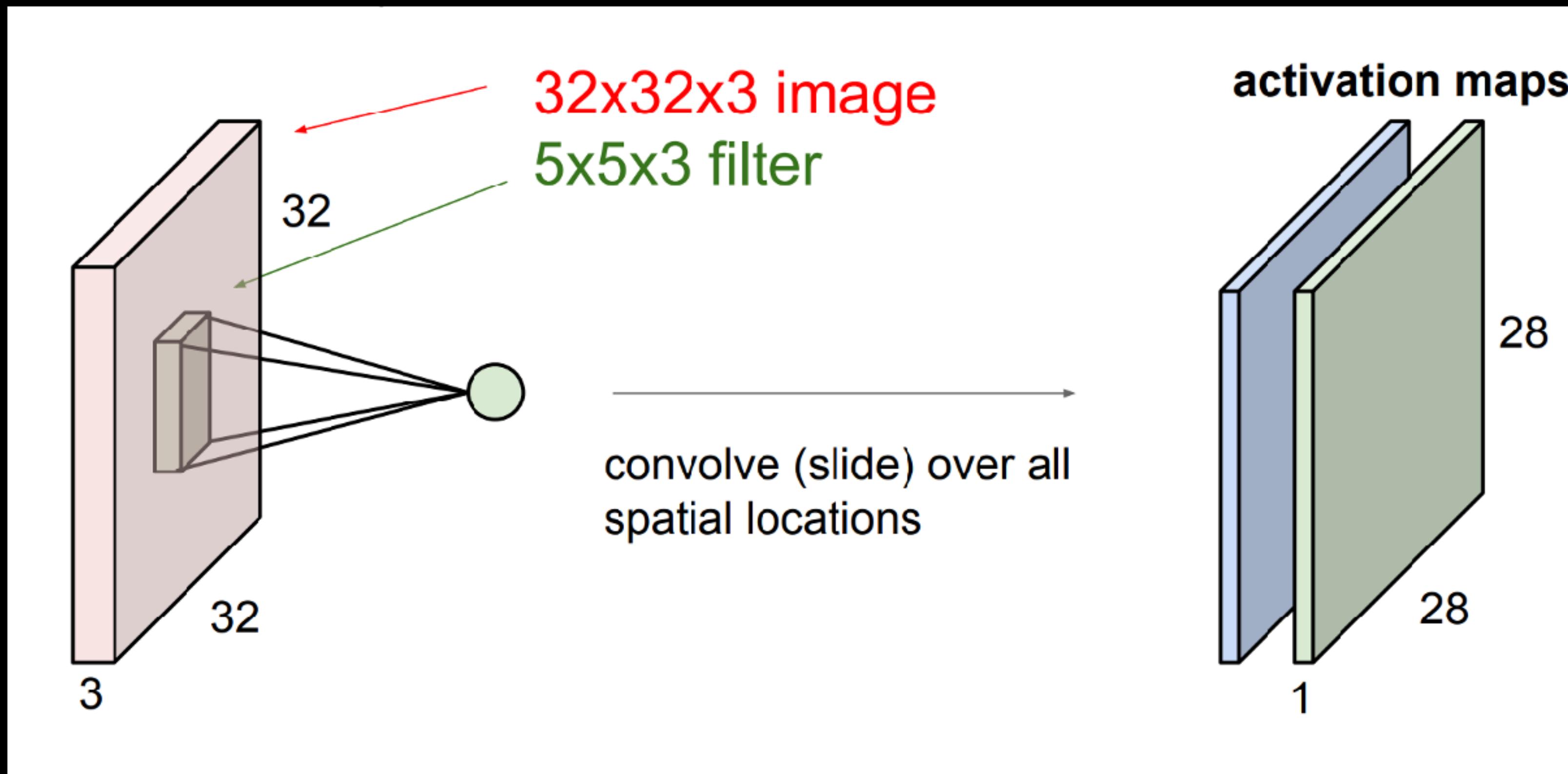
5x5x3 filter



Convolutions

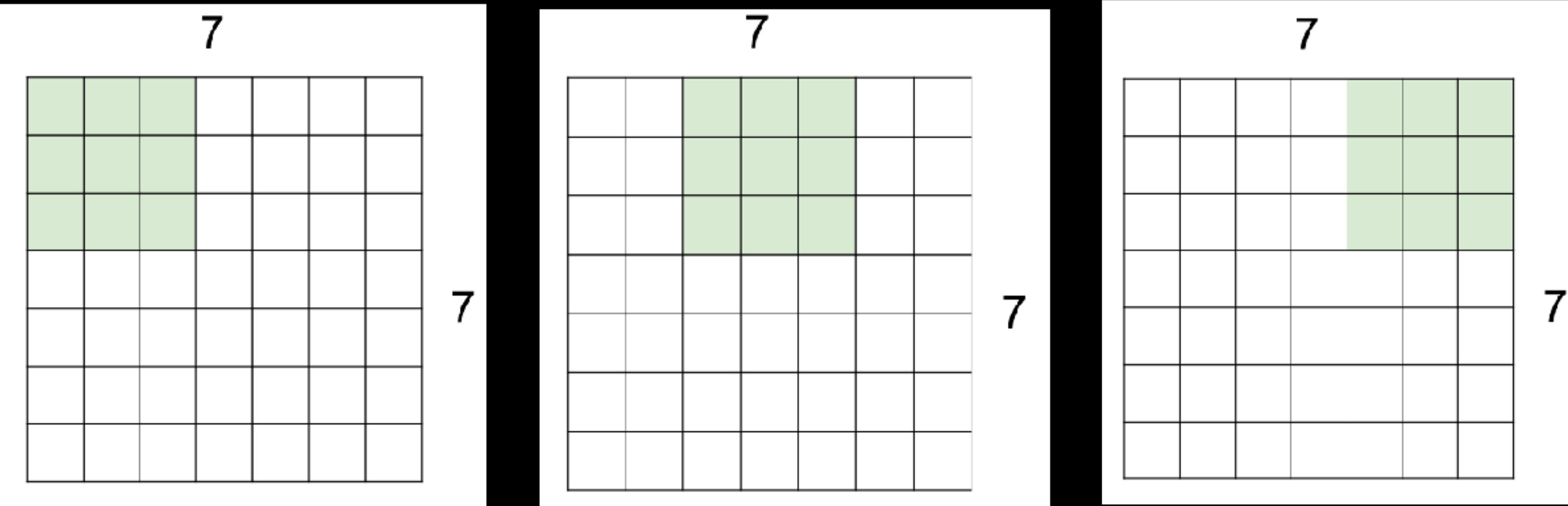


Convolutions



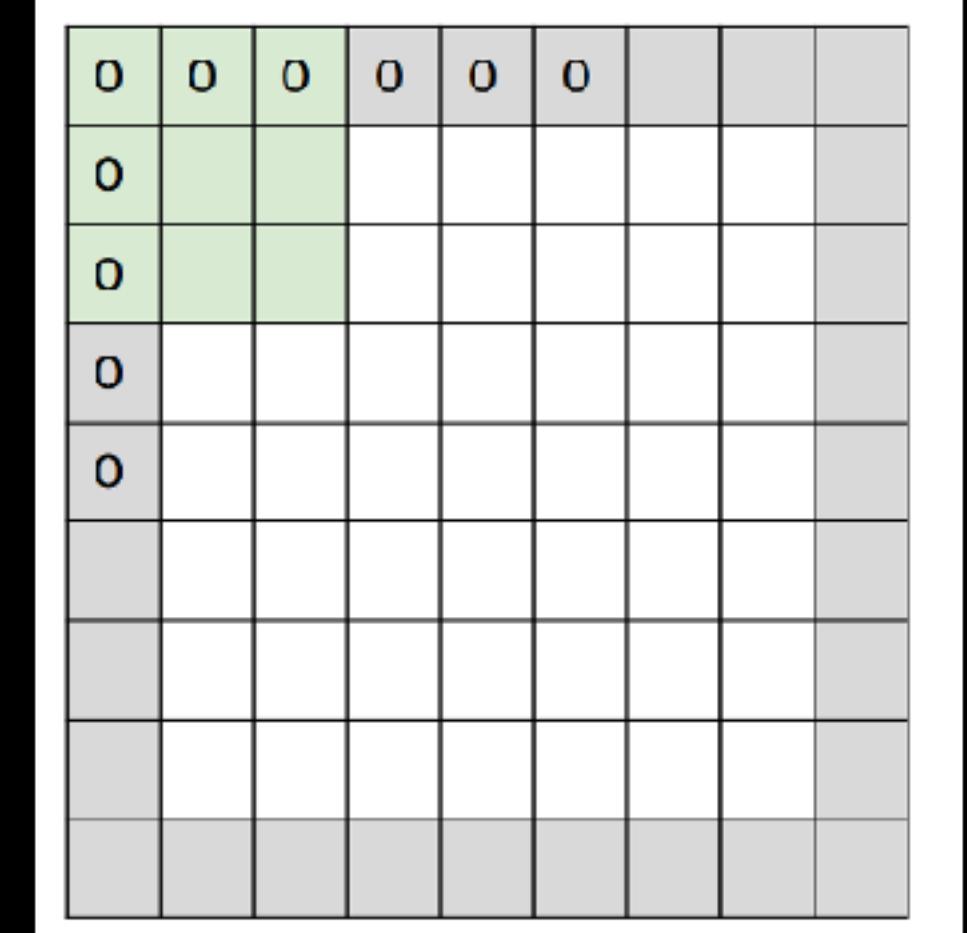
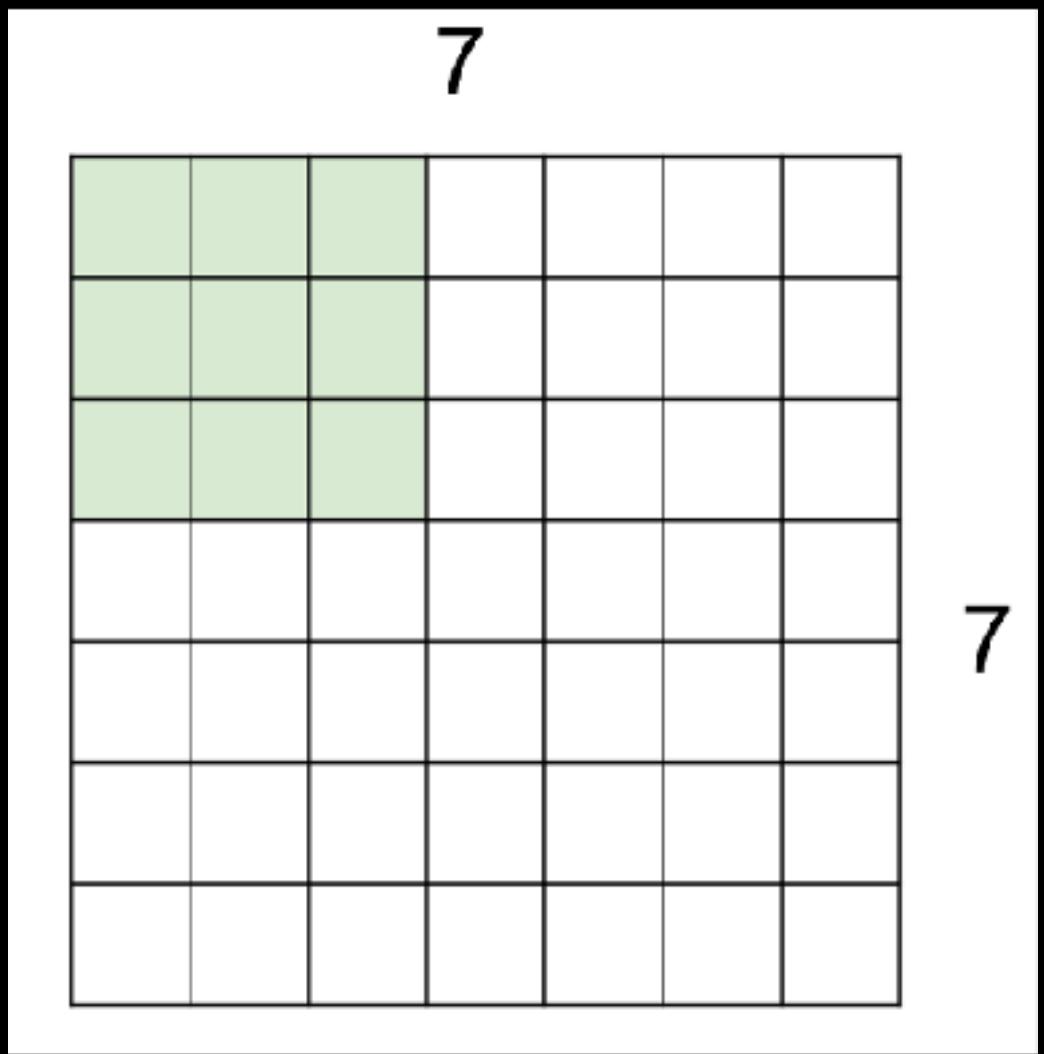
Stride

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



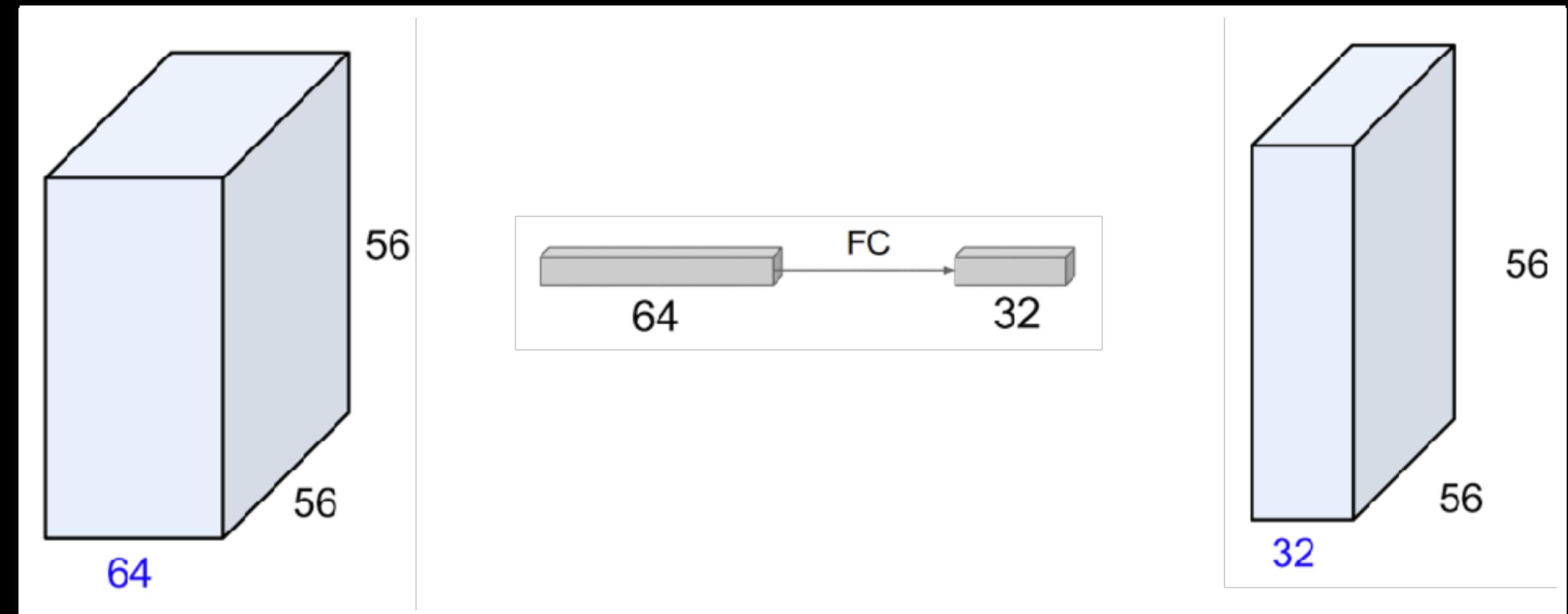
Padding

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



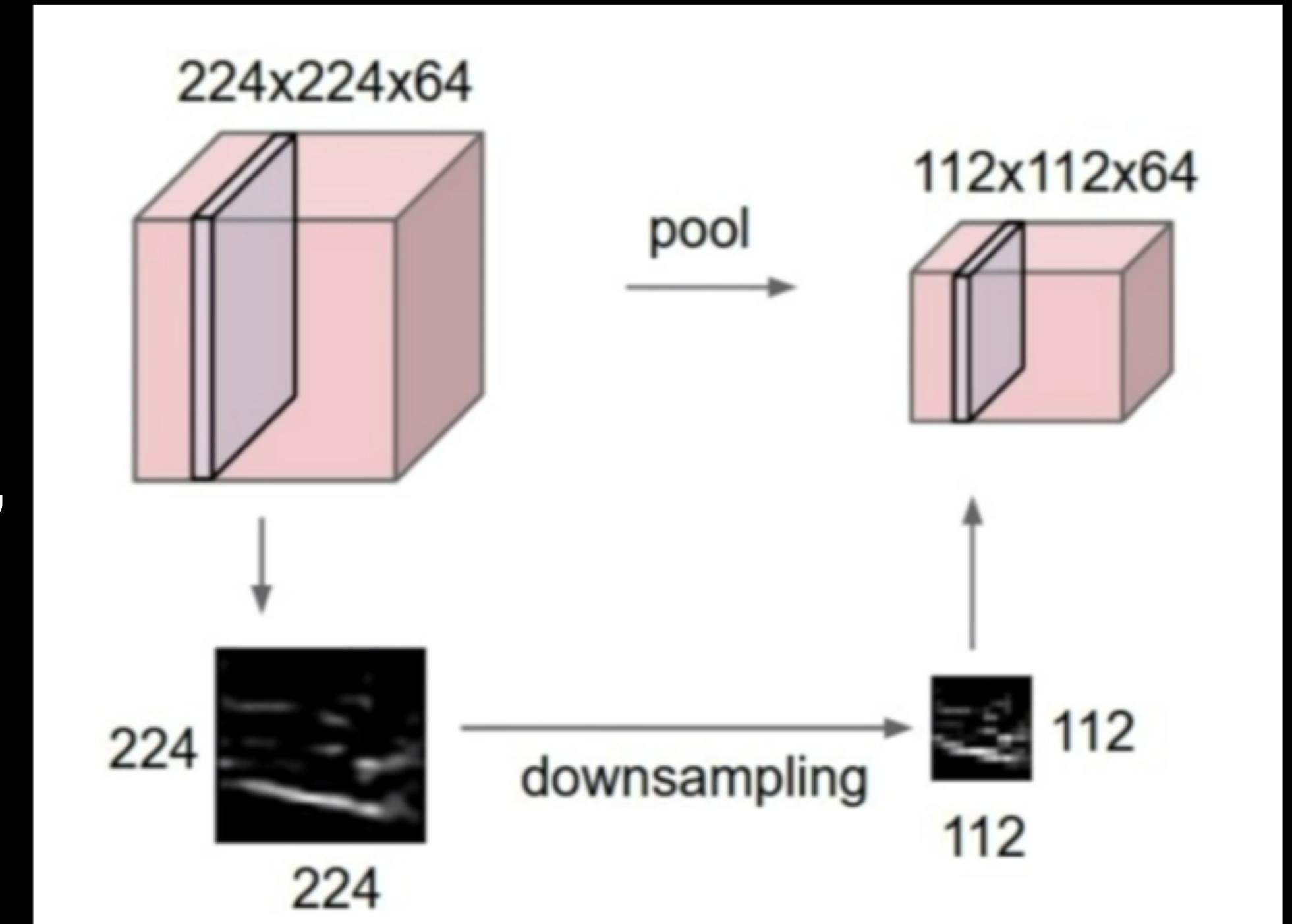
1x1 convolutions

- While they 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**.



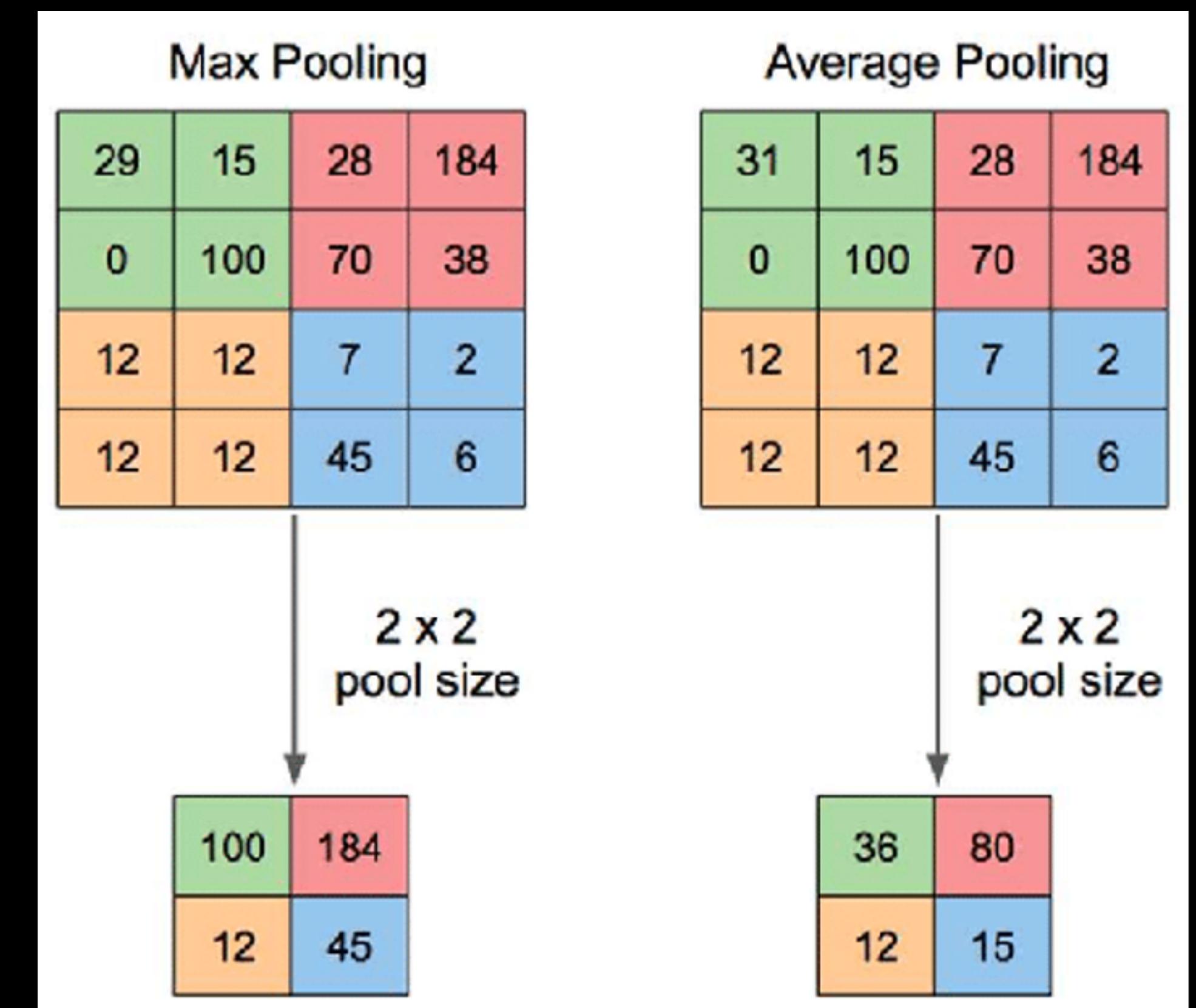
Pooling

- A way to downsample the input
- Convolutions with stride 2 also downsample, parameters



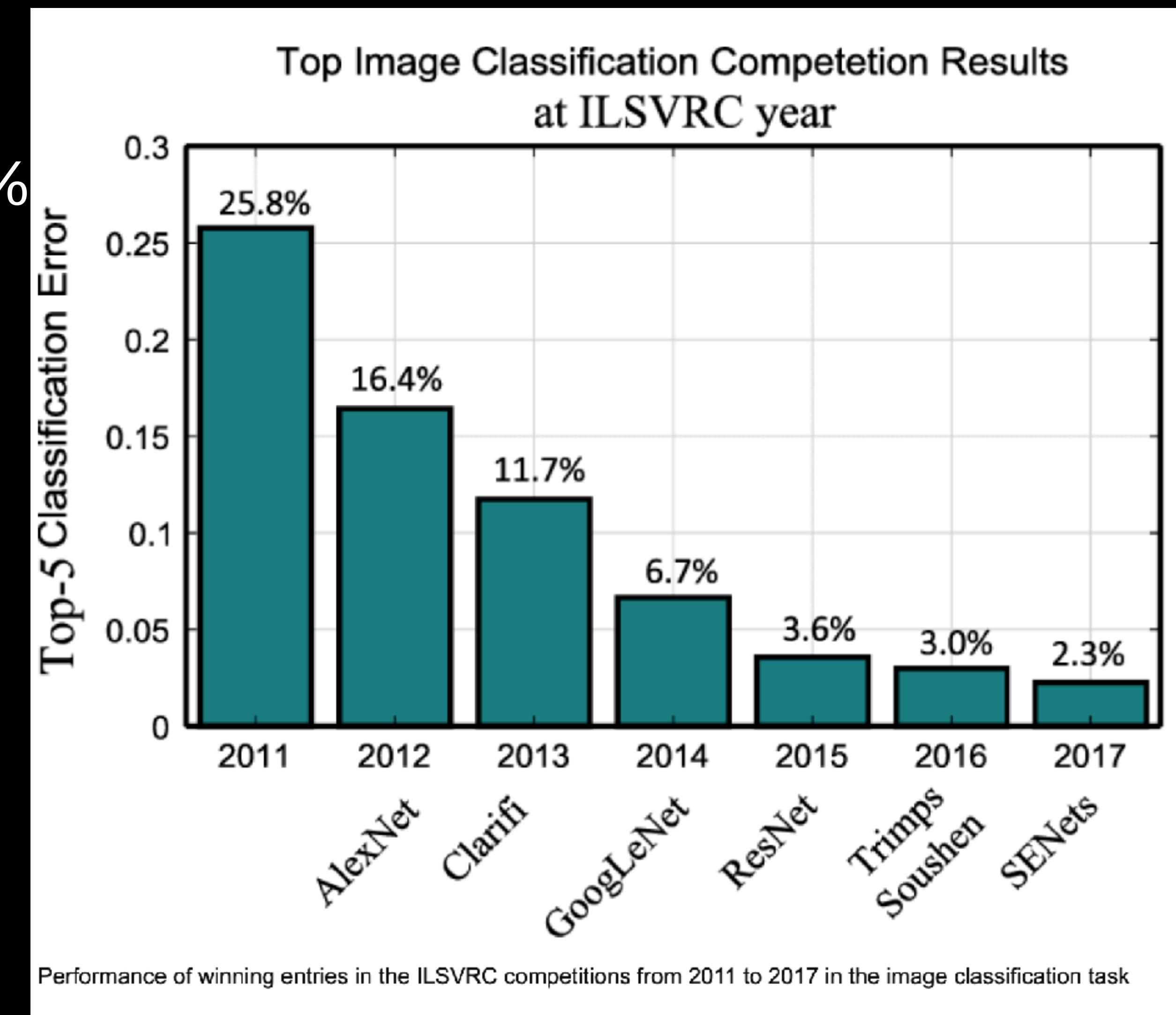
Pooling

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



Imagenet Challenge

- Human performance is about 5%



AlexNet

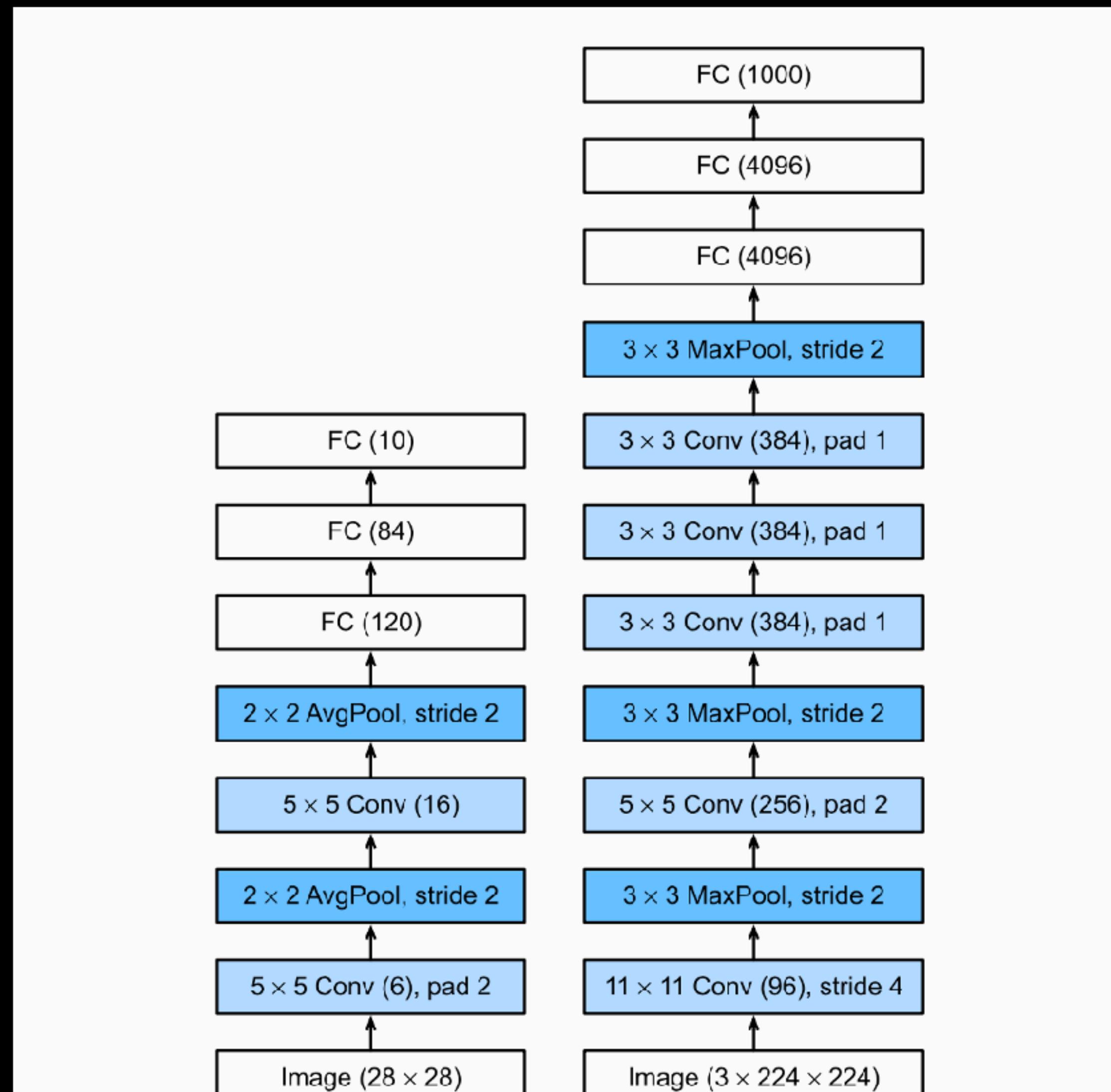
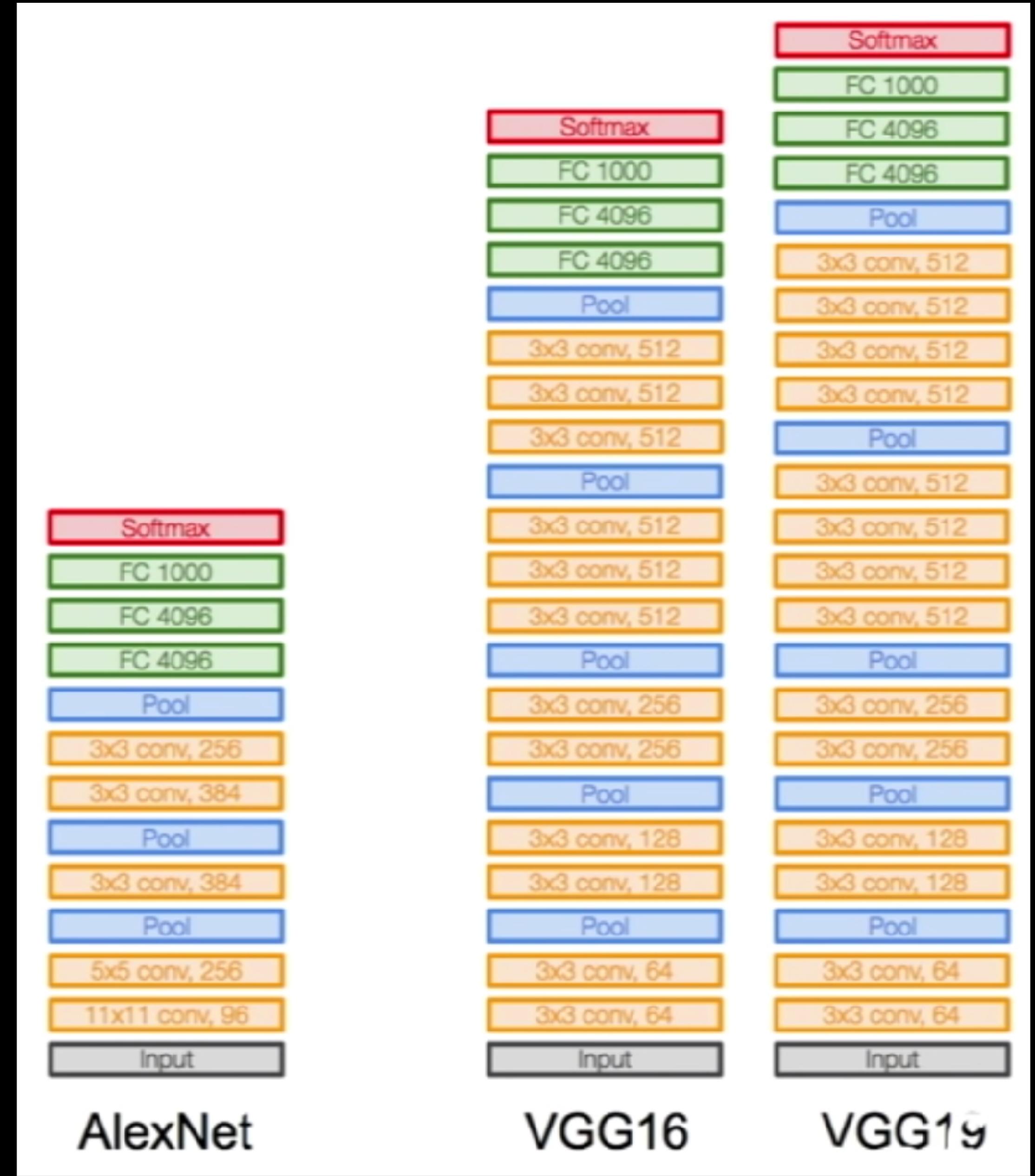


Fig. 7.1.2 From LeNet (left) to AlexNet (right).

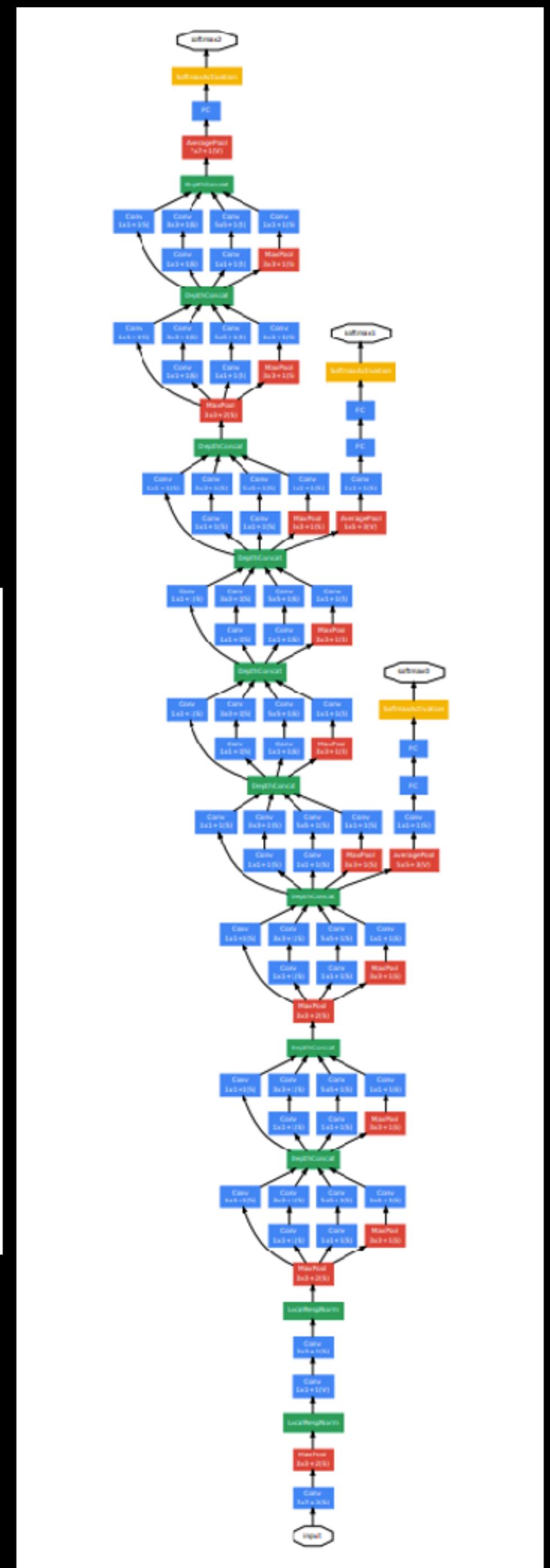
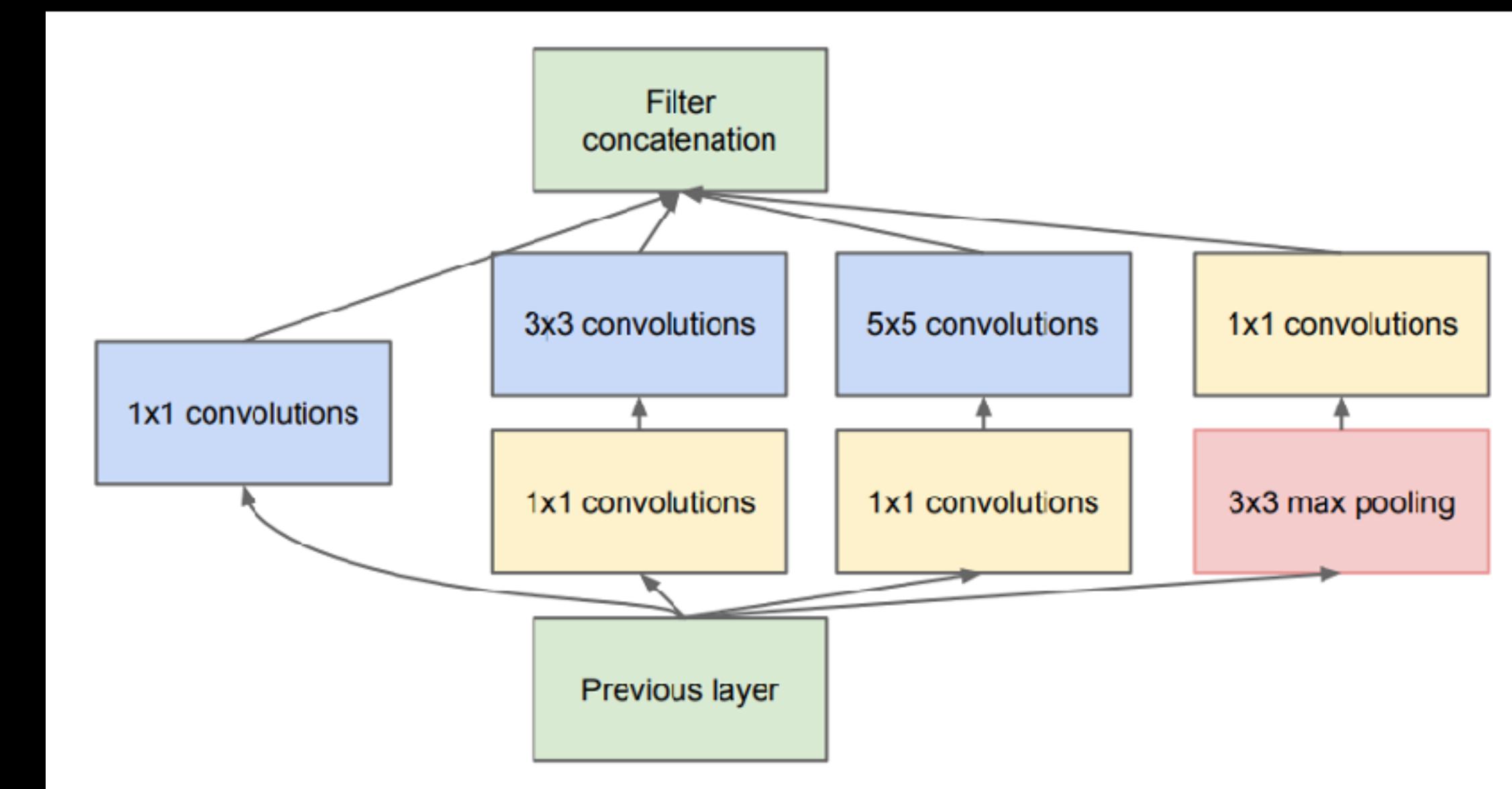
VGG

Changes in:
Filter size
channel numbers
depth
138 million parameters



GoogleNet

Parallel filters
Bottleneck layers
6.7 million parameters

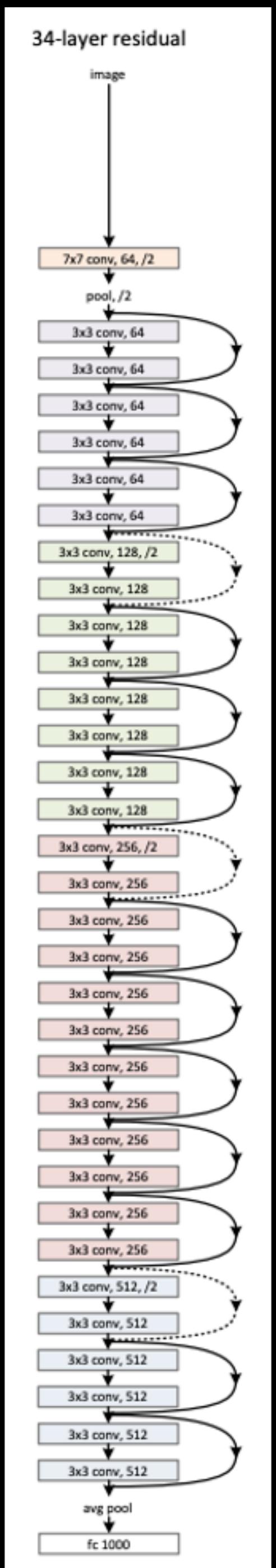


ResNet

batch normalisation

Residual blocks

11 million parameters for ResNet18



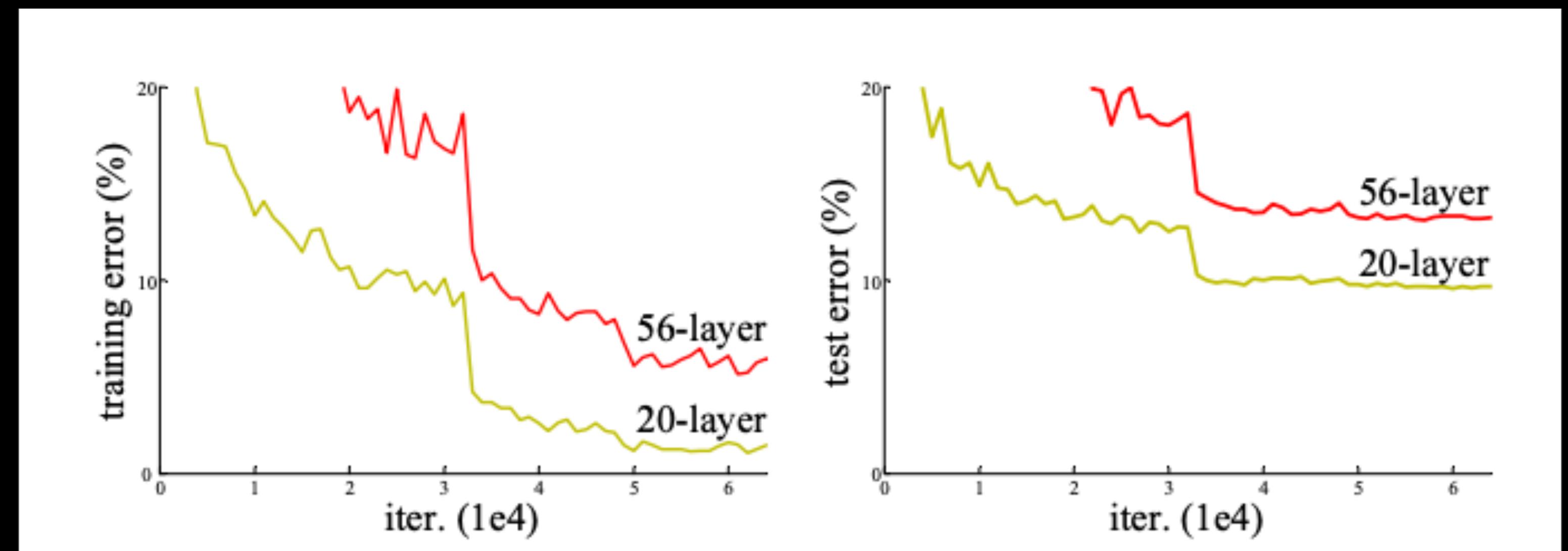
ResNet

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

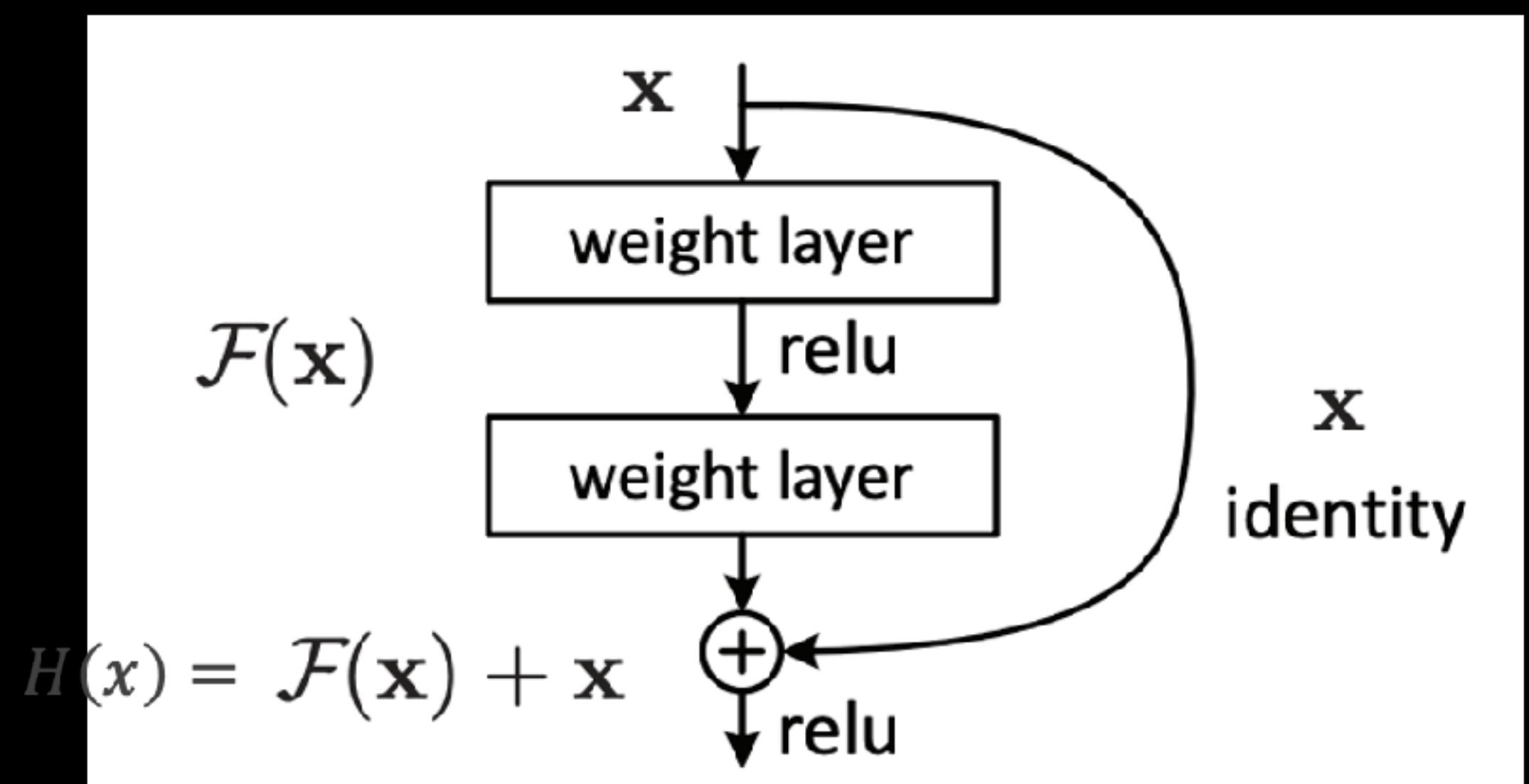


ResNet

Learn the residual, instead of the full mapping

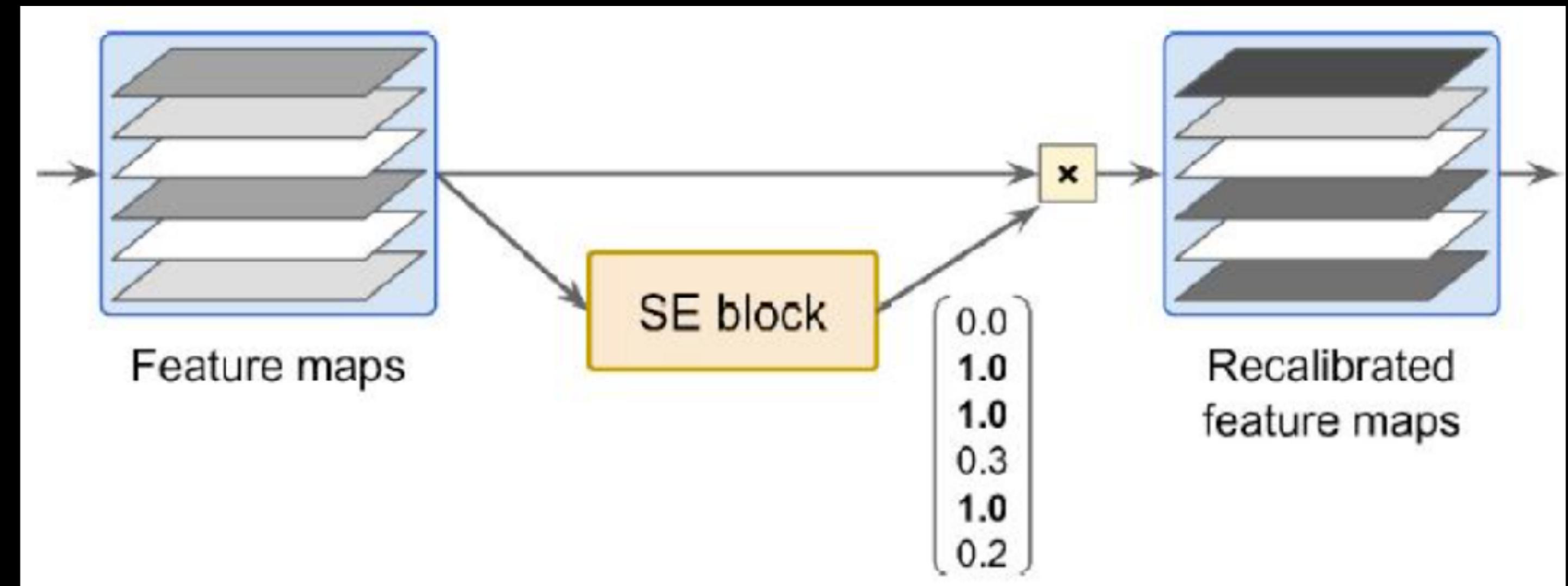
E.g., “use the same settings, but just change x ” is easier than starting from scratch.

It is easier for the gradient to flow back



SEnet

- 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

- **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.

