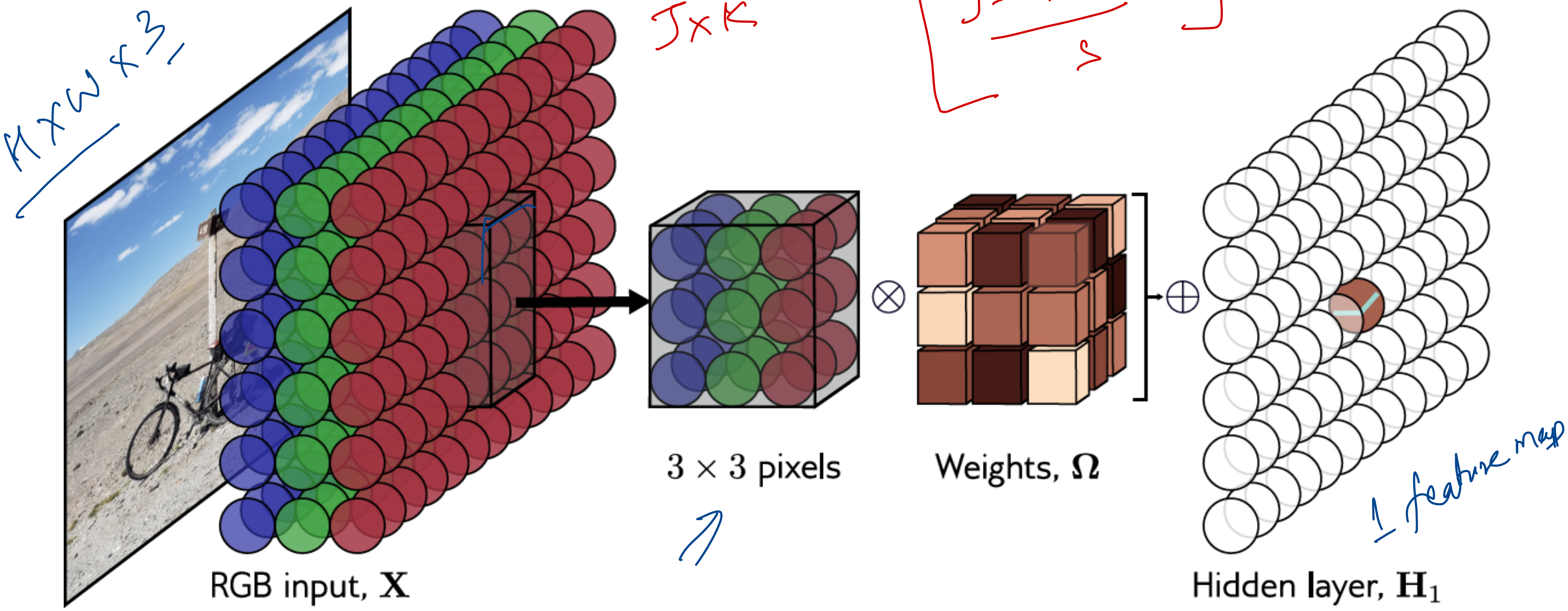
# Channels in 2D convolution

$$\left\lfloor \frac{J-M+2P}{S}+1 \right\rfloor \times \left\lfloor \frac{K-M+2P}{S}+1 \right\rfloor$$




Kernel size, stride, dilation all work as you would expect

# How many parameters?

- If there are $C_i$ input channels and kernel size K x K

$$\boldsymbol{\omega} \in \mathbb{R}^{C_i \times K \times K} \qquad \boxed{\boldsymbol{\beta} \in \mathbb{R}}$$

- If there are $C_i$ input channels and $C_o$ output channels

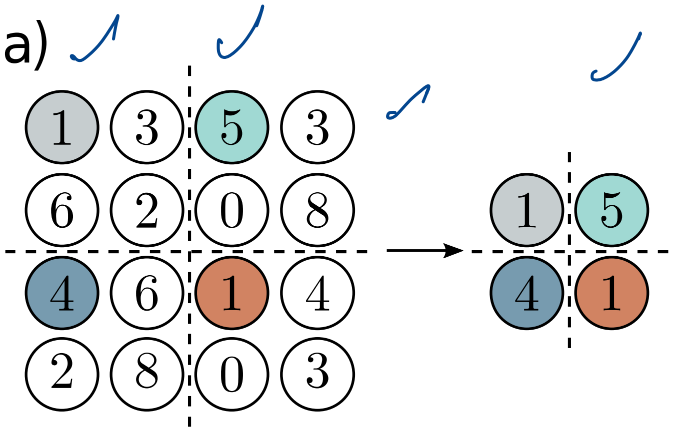
$$\boldsymbol{\omega} \in \mathbb{R}^{C_i \times C_o \times K \times K} \qquad \boxed{\boldsymbol{\beta} \in \mathbb{R}^{C_o}}$$

# Convolution #2

- 2D Convolution
- Downsampling and upsampling, 1x1 convolution
- Image classification
- Object detection
- Semantic segmentation
- Residual networks
- U-Nets and hourglass networks

# Downsampling

a)



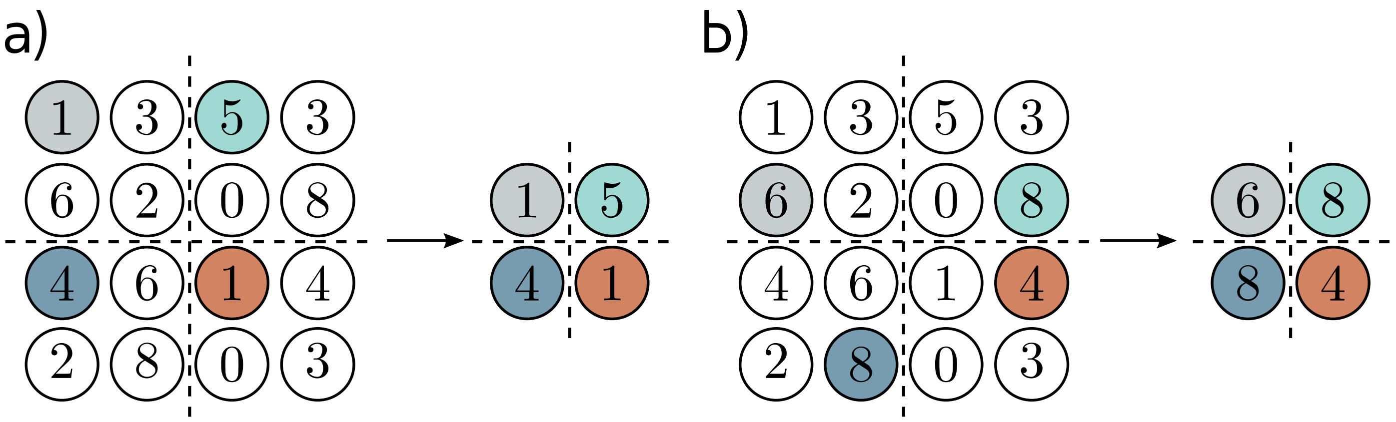
Sample every other position (equivalent to stride two)

# Downsampling

Zero-padding

stride = 1, 2

$$\left\lfloor \frac{I - M + 2P}{S} + 1 \right\rfloor$$




$$\left\lfloor \frac{4 - 2 + 0}{2} + 1 \right\rfloor = 2$$

Kernel size $2 \times 2$ Stride = 2

No zero-padding

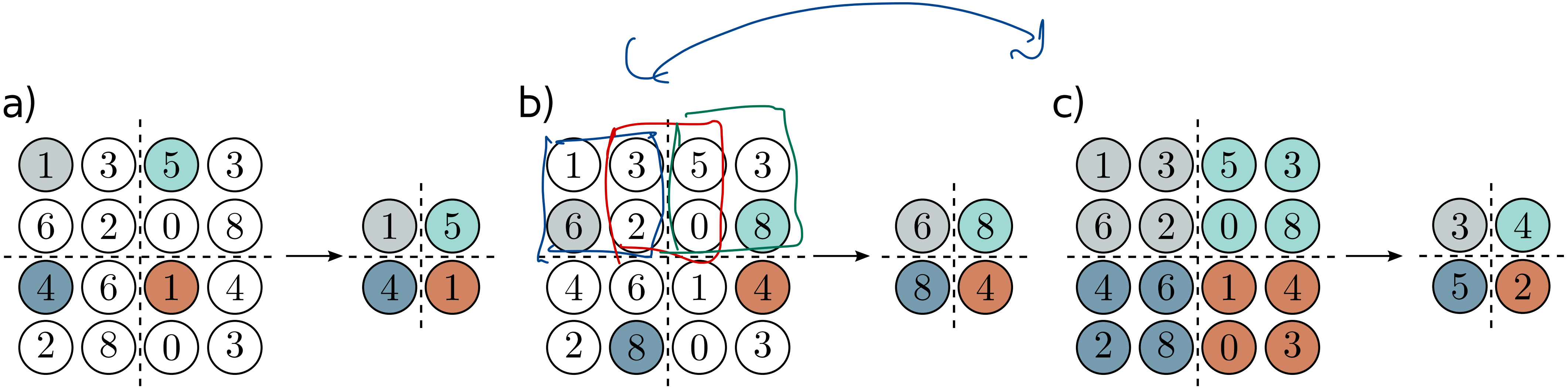
$I = 5$

$$\left\lfloor \frac{5 - 2 + 0}{2} + 1 \right\rfloor = 2$$

Sample every other position (equivalent to stride two)

Max pooling (partial invariance to translation)

# Downsampling



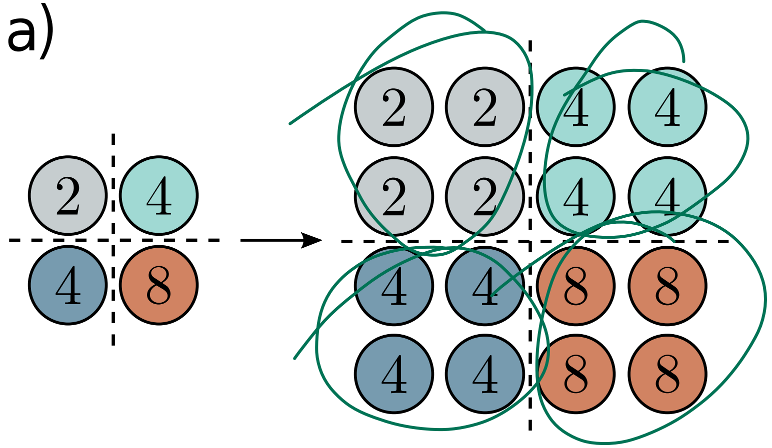

Sample every other position (equivalent to stride two)

Max pooling (partial invariance to translation)

Mean pooling

# Upsampling

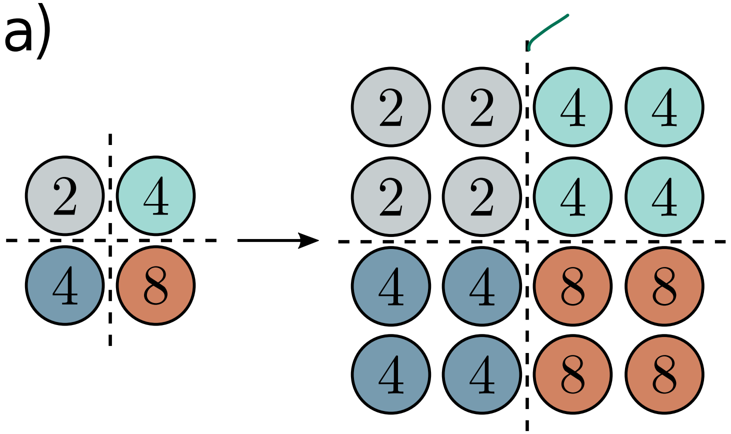
a)



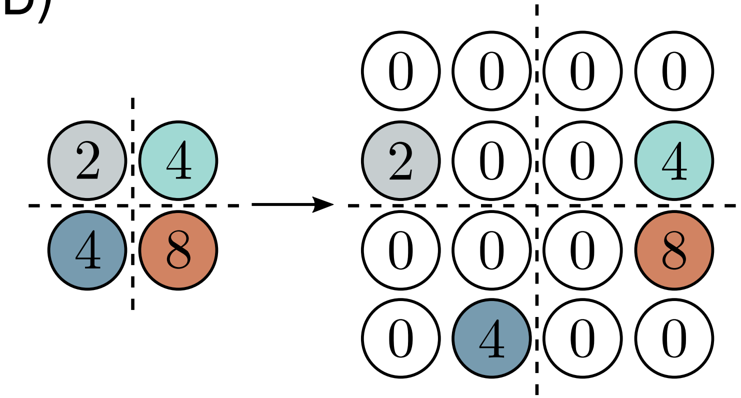

Duplicate

# Upsampling

a)




b)



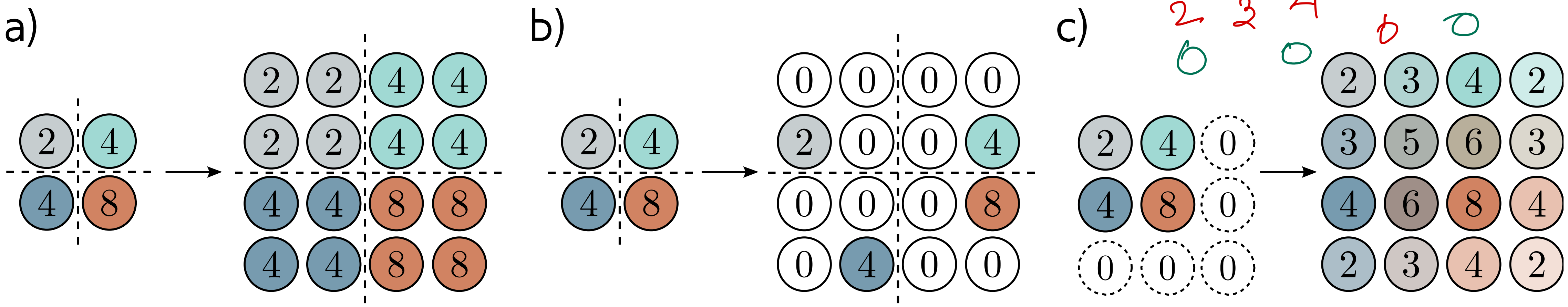

Duplicate

Max-upsampling

Max-unpooling

# Upsampling

2 3 4 2 0
3 4.5 6 3 0
4 6 8 4 0
2 3 4 2 0
0 0 0 0




Duplicate

Max-upsampling

Bilinear interpolation

# In-Network upsampling: “Max Unpooling”

**Max Pooling**
Remember which element was max!

| | | | |
|---|---|---|---|
| 1 | 2 | 6 | 3 |
| 3 | 5 | 2 | 1 |
| 1 | 2 | 2 | 1 |
| 7 | 3 | 4 | 8 |

Input: 4 x 4

| | |
|---|---|
| 5 | 6 |
| 7 | 8 |

Output: 2 x 2

→ … →

Rest of the network

**Max Unpooling**
Use positions from pooling layer

| | |
|---|---|
| 1 | 2 |
| 3 | 4 |

Input: 2 x 2

| | | | |
|---|---|---|---|
| 0 | 0 | 2 | 0 |
| 0 | 1 | 0 | 0 |
| 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 4 |

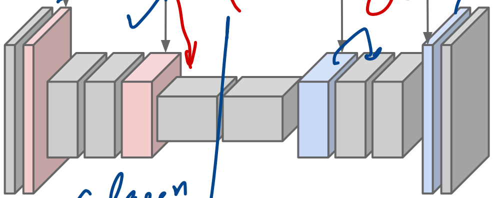
Output: 4 x 4

Corresponding pairs of downsampling and upsampling layers

224x224

encoding of the image

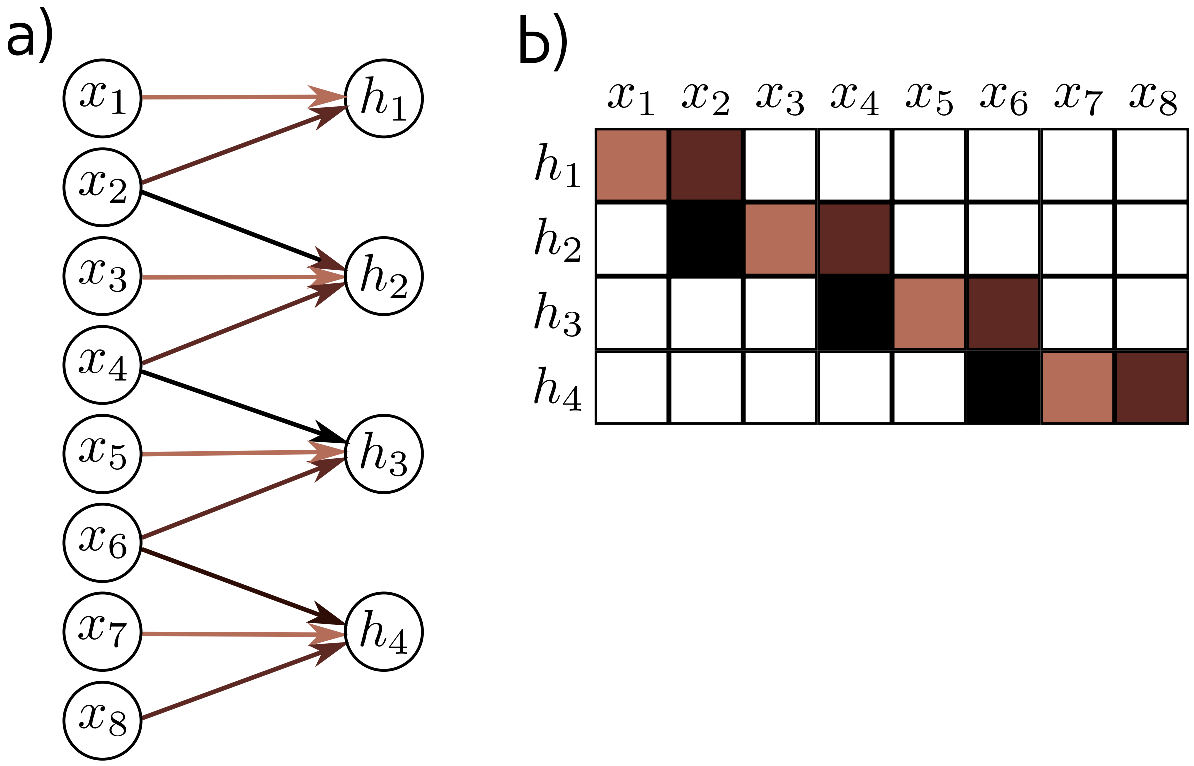
decoder



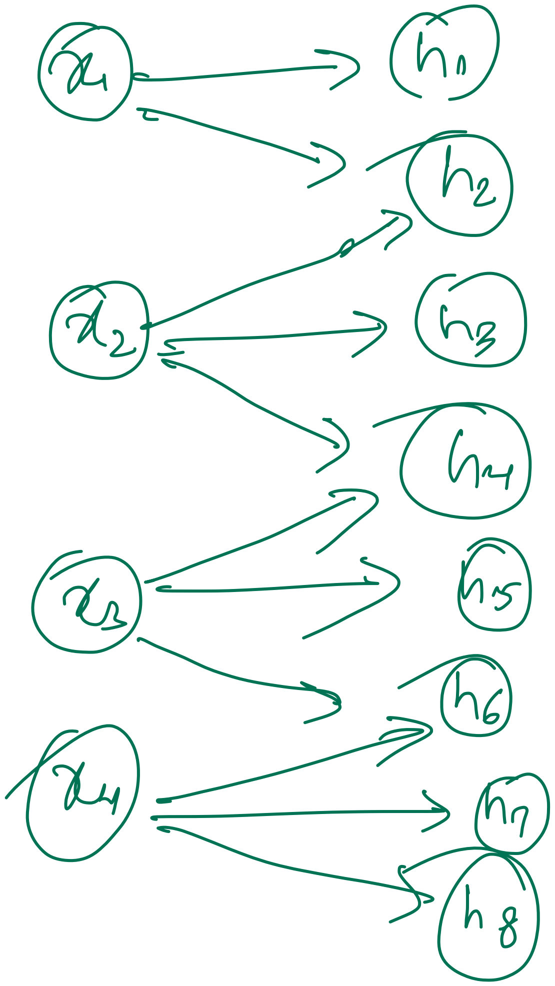
classif^n

semantic segmentation

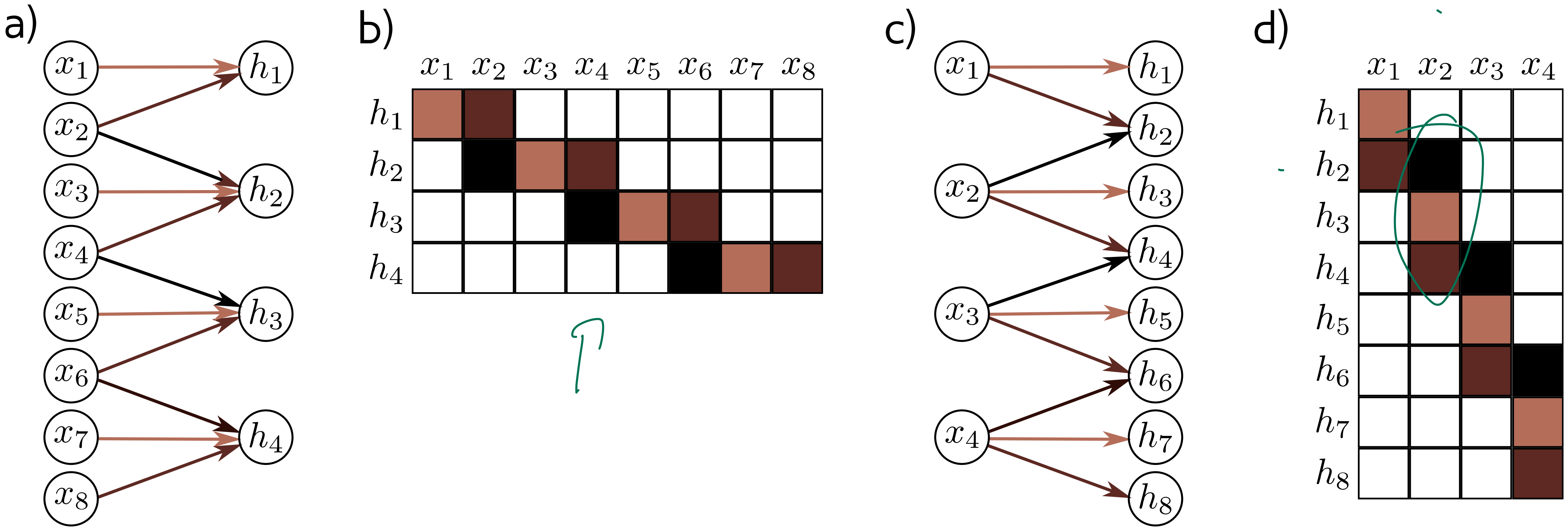
# Transposed convolutions




Kernel size 3, Stride 2 convolution



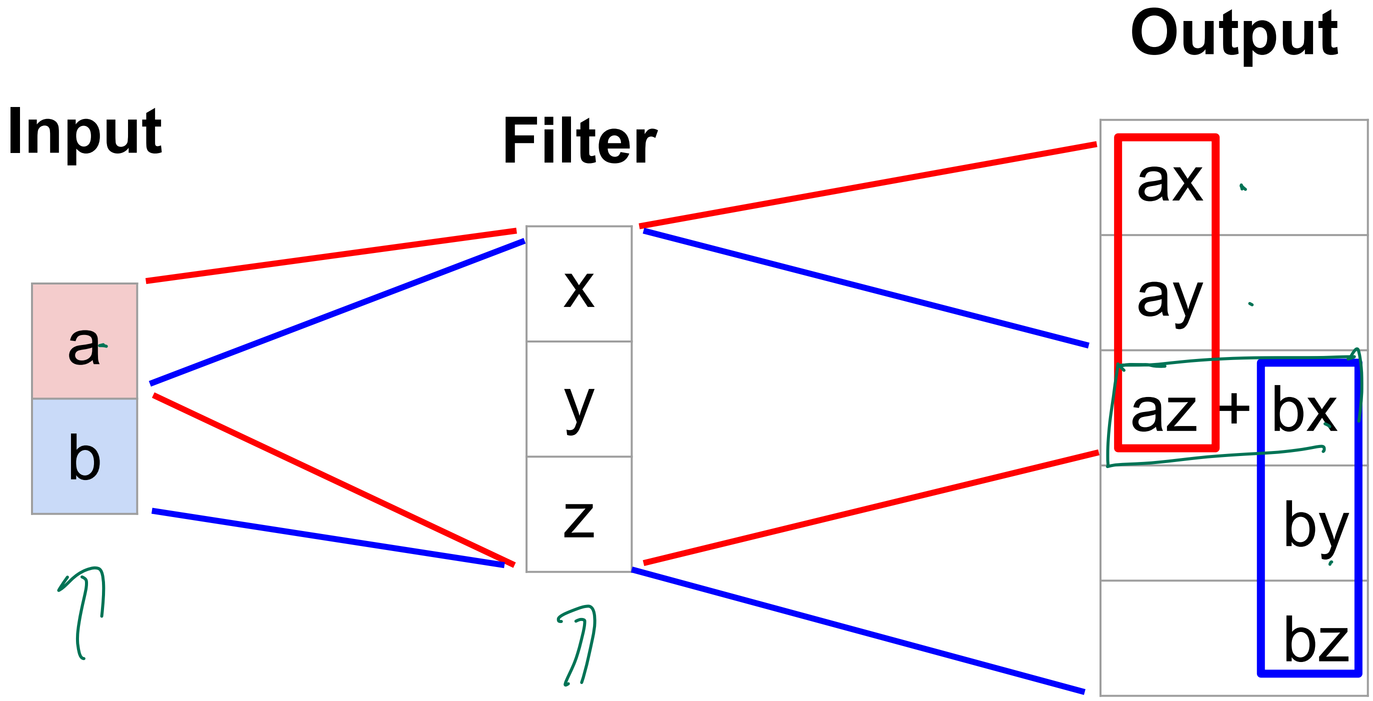
# Transposed convolutions




Kernel size 3, Stride 2 convolution

Transposed convolution

# Learnable Upsampling: 1D Example




Output contains copies of the filter weighted by the input, summing at where at overlaps in the output

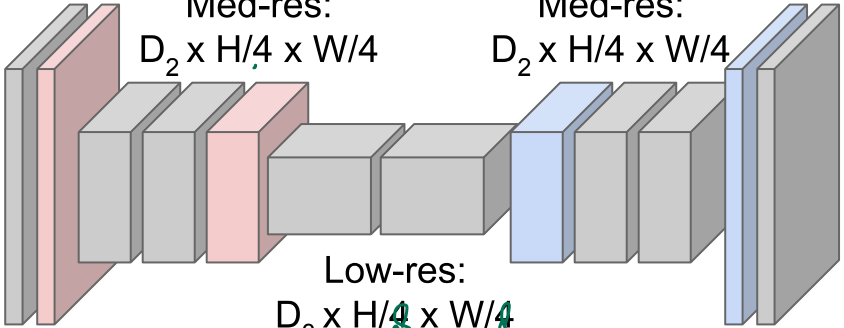
# Semantic Segmentation Idea: Fully Convolutional

**Downsampling**: Pooling, strided convolution

Design network as a bunch of convolutional layers, with **downsampling** and **upsampling** inside the network!

**Upsampling**: Unpooling or strided transposed convolution

Input:
3 x H x W



Med-res:
$D_2$ x H/4 x W/4

Med-res:
$D_2$ x H/4 x W/4

Low-res:
$D_3$ x H/4 x W/4

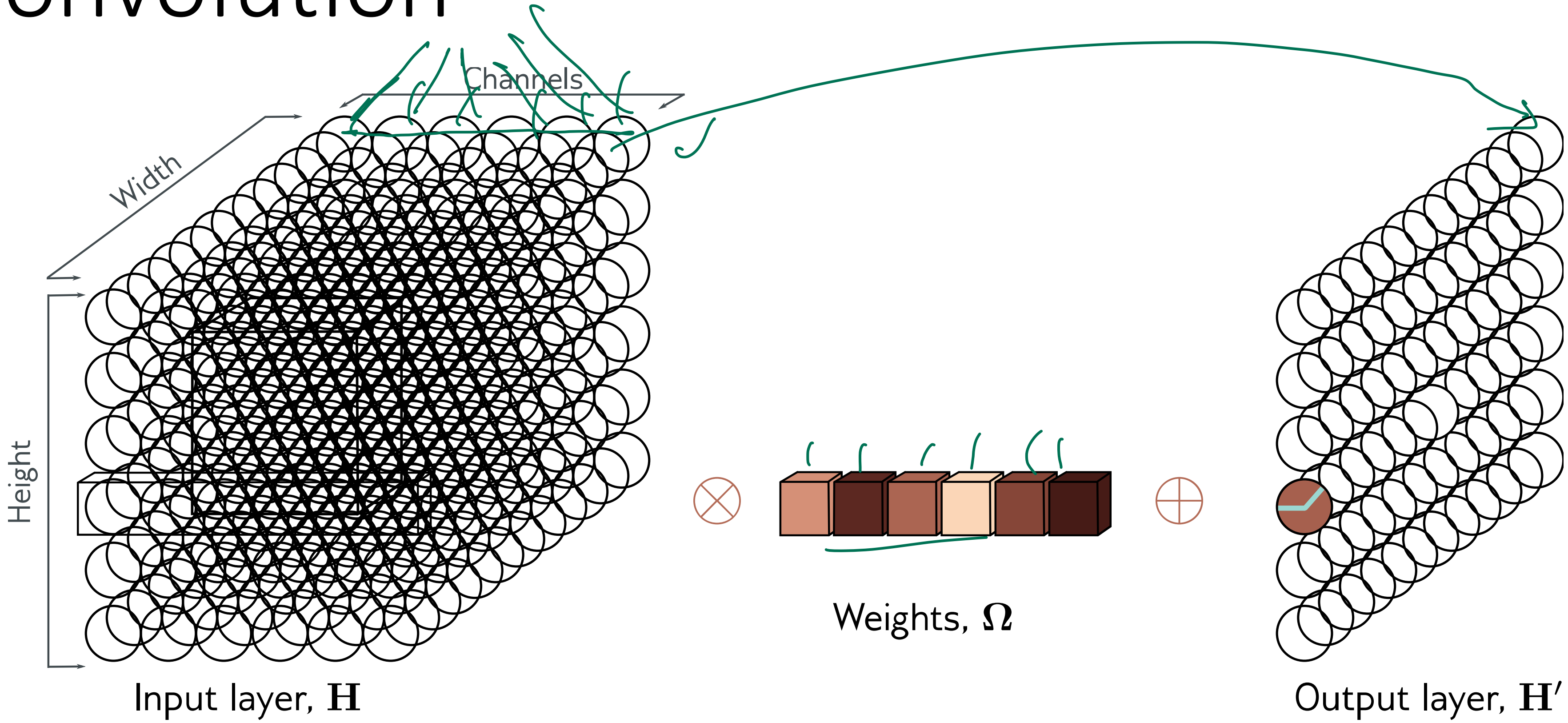
High-res:
$D_1$ x H/2 x W/2

High-res:
$D_1$ x H/2 x W/2

Predictions:
H x W

Long, Shelhamer, and Darrell, "Fully Convolutional Networks for Semantic Segmentation", CVPR 2015
Noh et al, "Learning Deconvolution Network for Semantic Segmentation", ICCV 2015

# 1x1 convolution

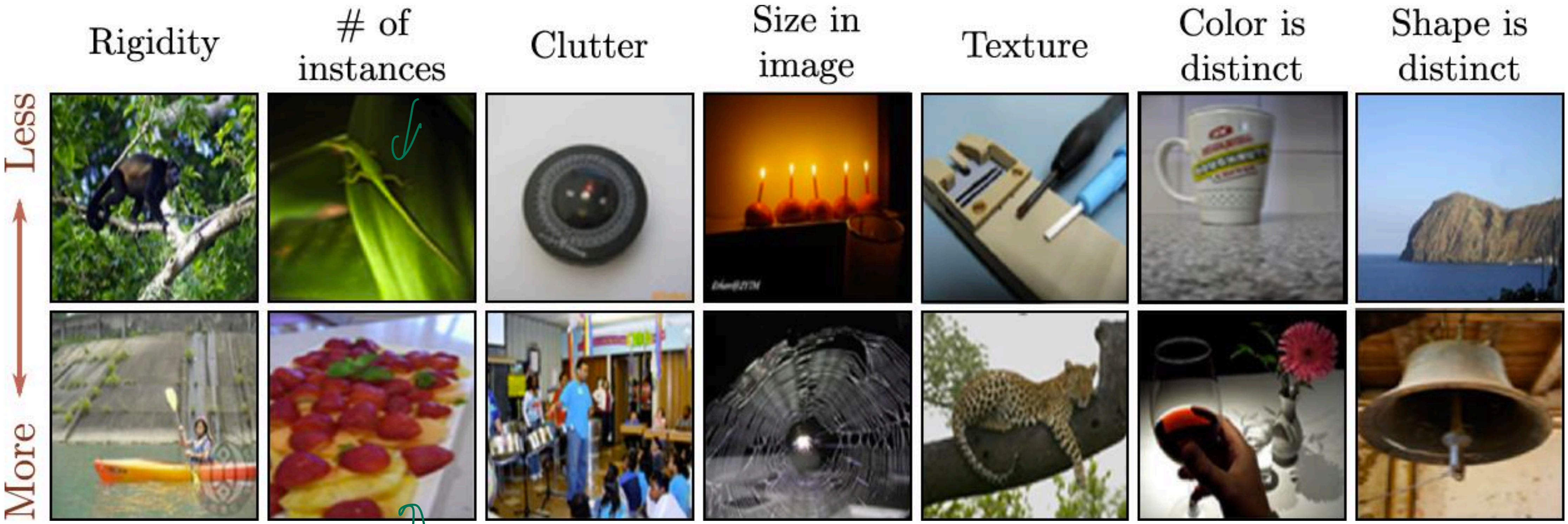



- Mixes channels
- Can change number of channels
- Equivalent to running same fully connected network at each position

# Convolution #2

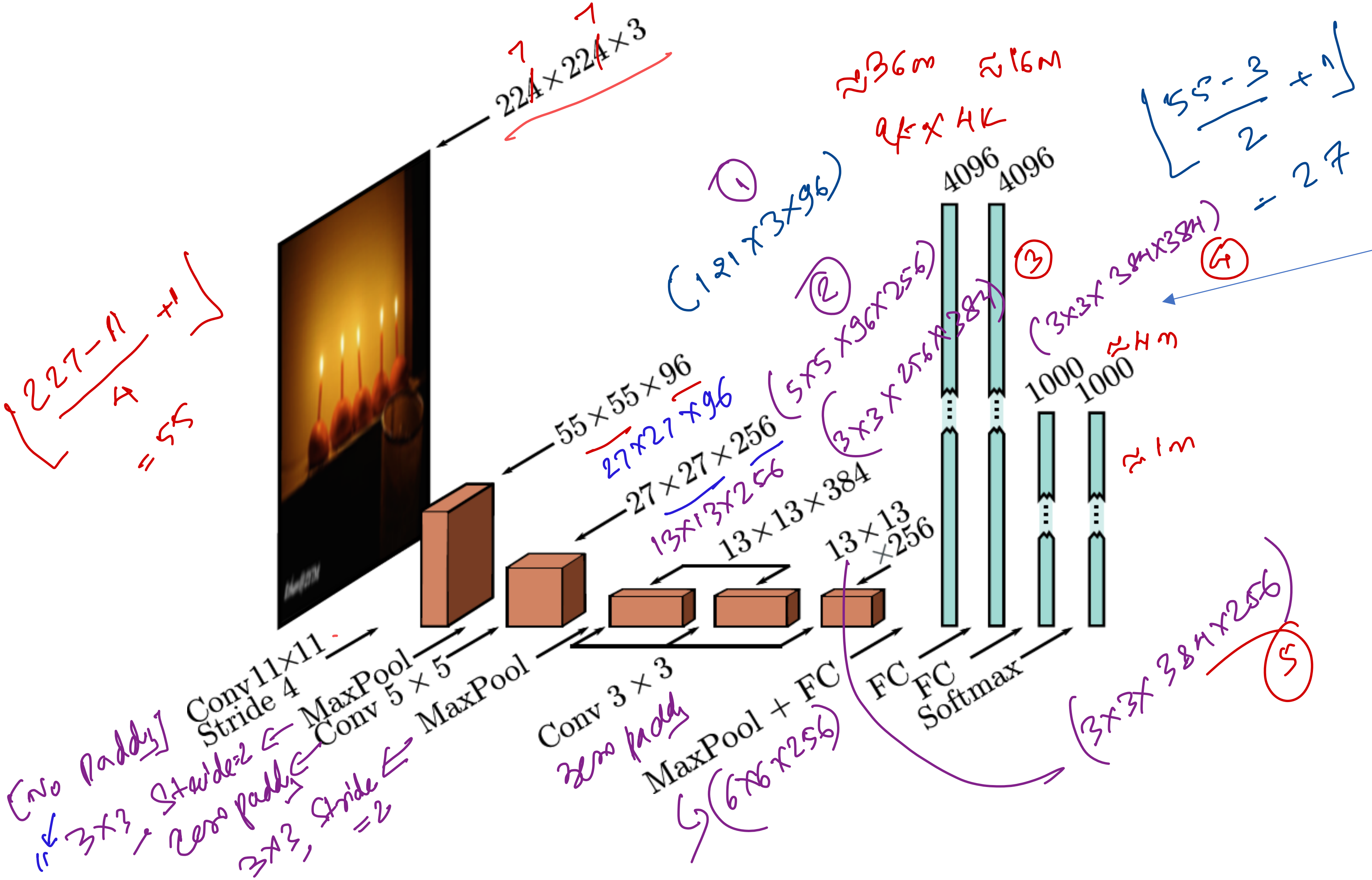
- 2D Convolution
- Downsampling and upsampling, 1x1 convolution
- Image classification
- Object detection
- Semantic segmentation
- Residual networks
- U-Nets and hourglass networks

# ImageNet database



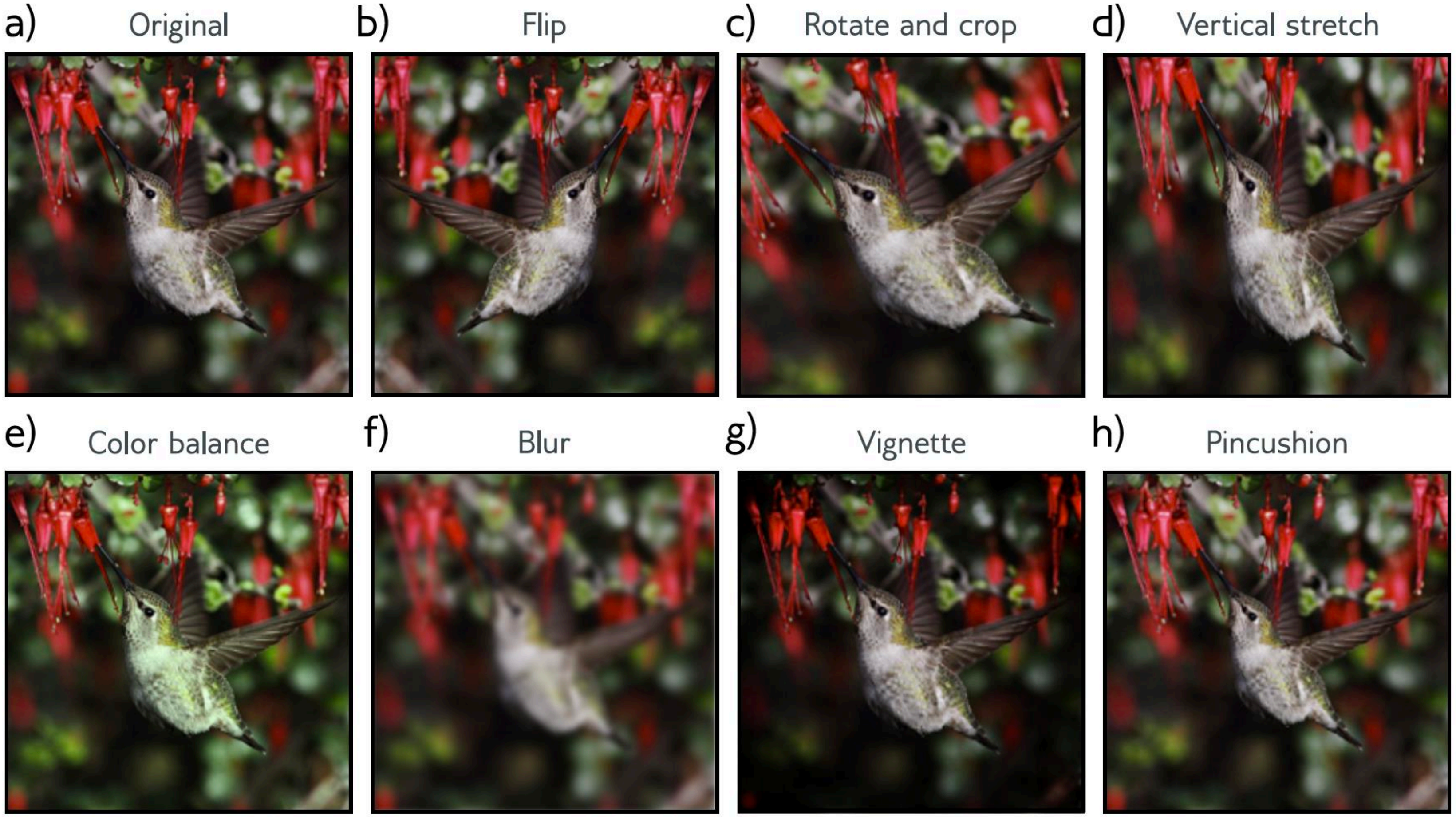

- 224 x 224 images
- 1,281,167 training images, 50,000 validation images, and 100,000 test images
- 1000 classes

# AlexNet (2012)



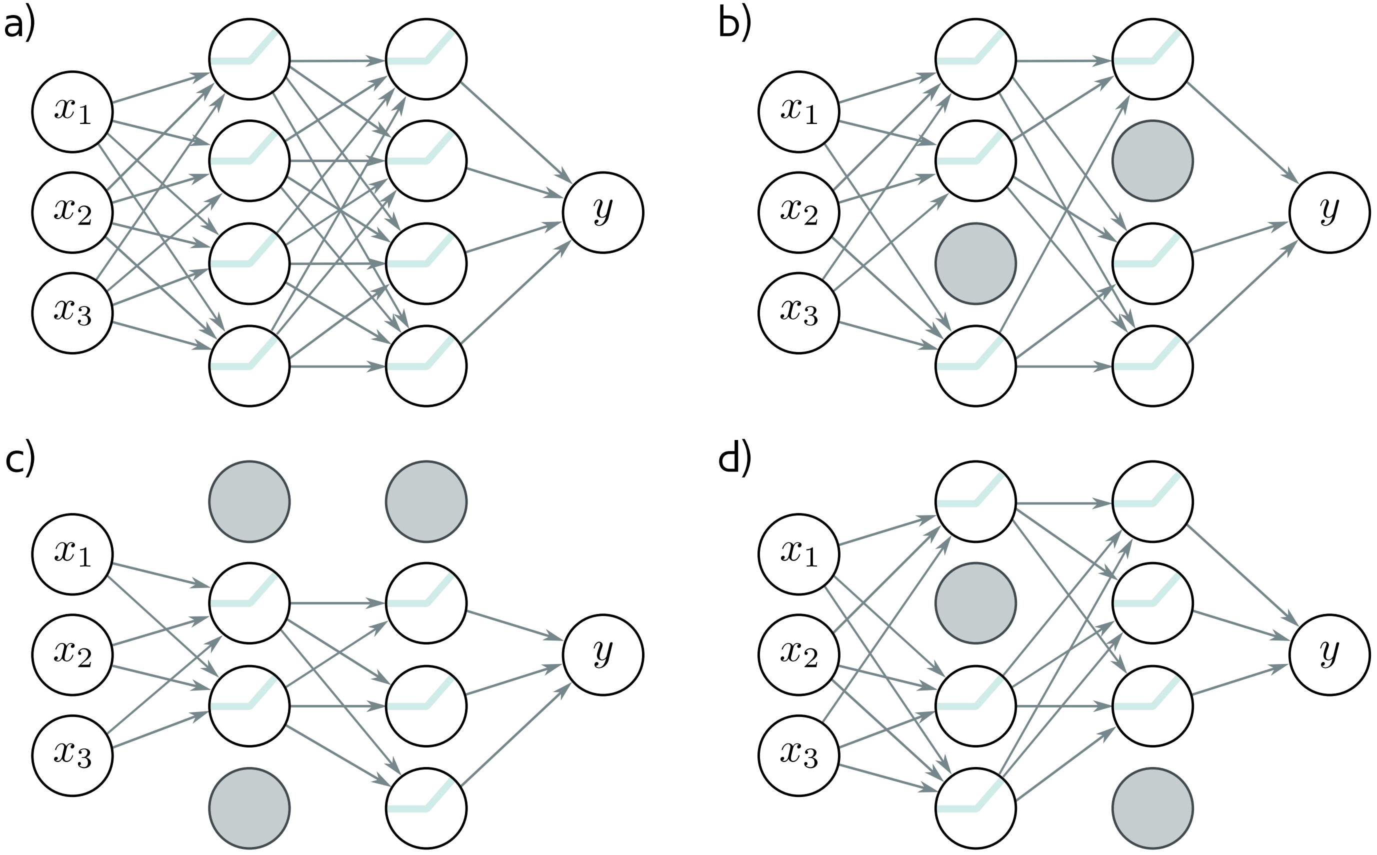

Almost all the 60 million parameters
parameters are in fully connected layers

# Data augmentation




- Data augmentation a factor of 2048 using (i) spatial transformations and (ii) modifications of the input intensities.

# Dropout




- Dropout was applied in the fully connected layers

# Details

- At test time average results from five different cropped and mirrored versions of the image
- SGD with a momentum coefficient of 0.9 and batch size of 128.
- L2 (weight decay) regularizer used.
- This system achieved a 16.4% top-5 error rate and a 38.1% top-1 error rate.

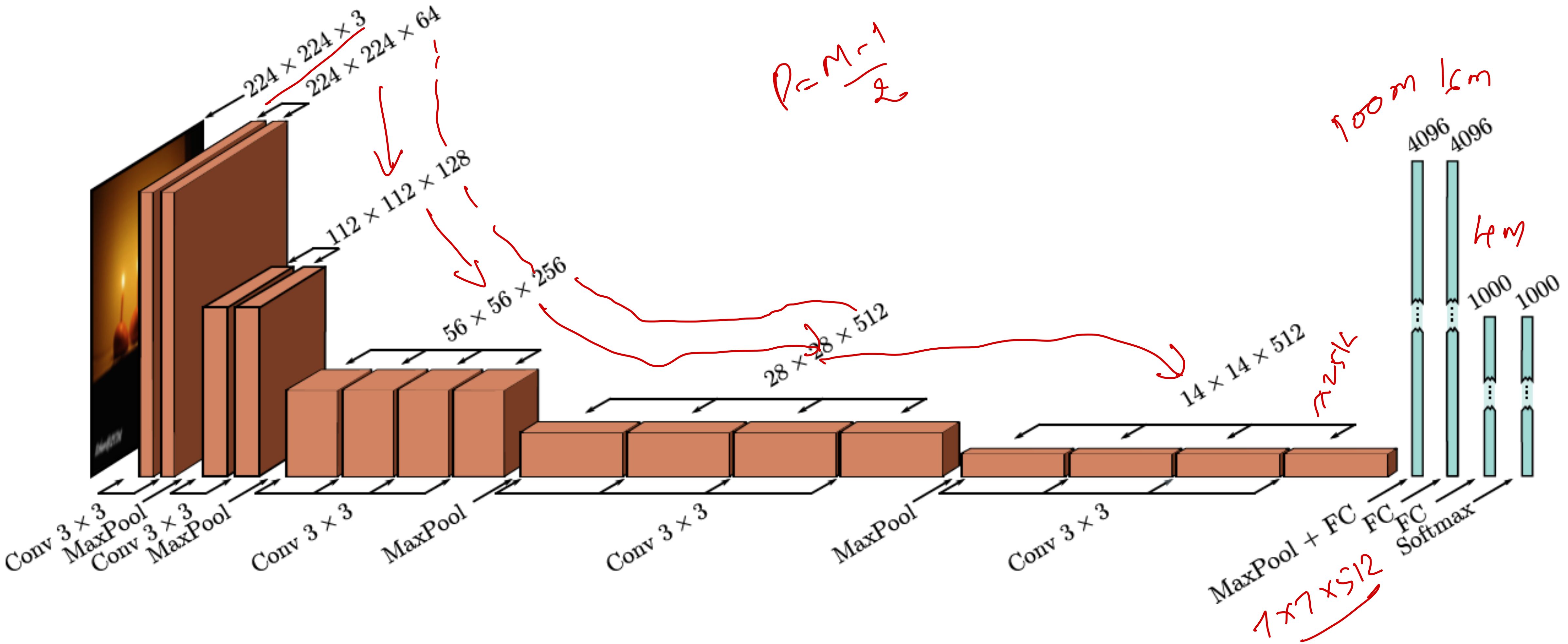
# VGG (2015)

-19

Max Pool: 2x2, stride=2 → no zero padding / valid conv

Conv: filter: 3x3, stride=1, "zero padding"

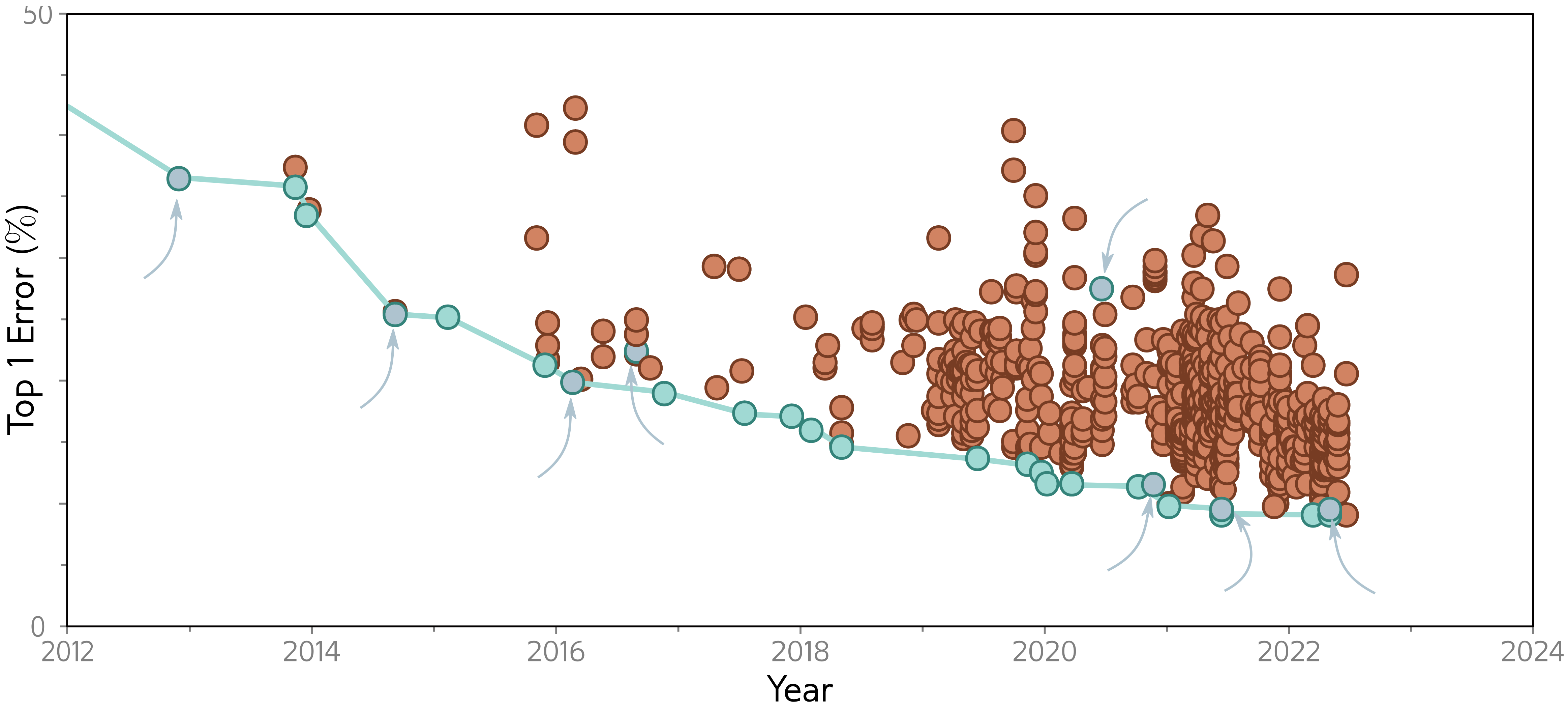
$P = \frac{M-1}{2}$



# Details

- 19 hidden layers
- 144 million parameters
- 6.8% top-5 error rate, 23.7% top-1 error rate

# ImageNet History



# Convolution #2

- 2D Convolution
- Downsampling and upsampling, 1x1 convolution
- Image classification
- Object detection
- Semantic segmentation
- Residual networks
- U-Nets and hourglass networks