Semantic Segmentation

Computer Vision

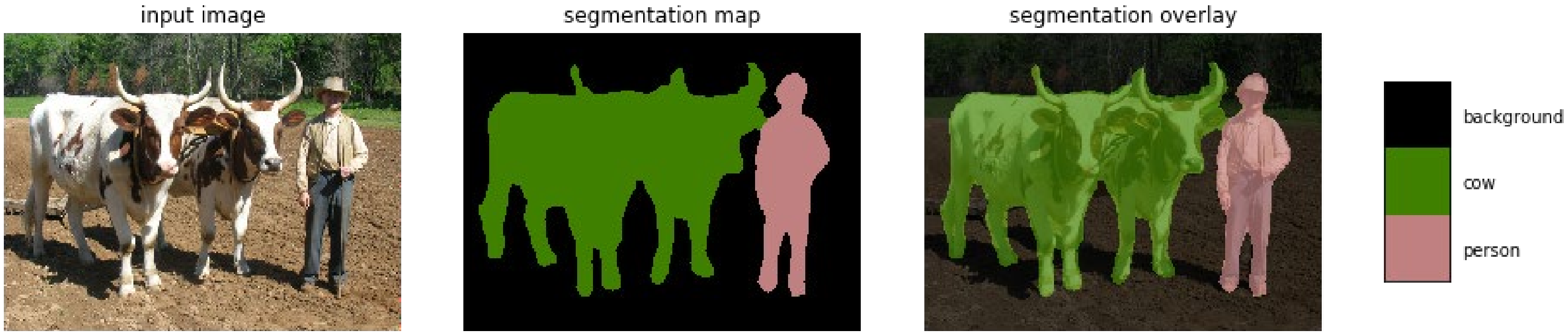
Outline

- Semantic Segmentation
- Approaches
- Sliding window, fully convolutional, upSampling
- U-Net
- Depthwise Separable Convolutions
- MobileNet
- Hands-on

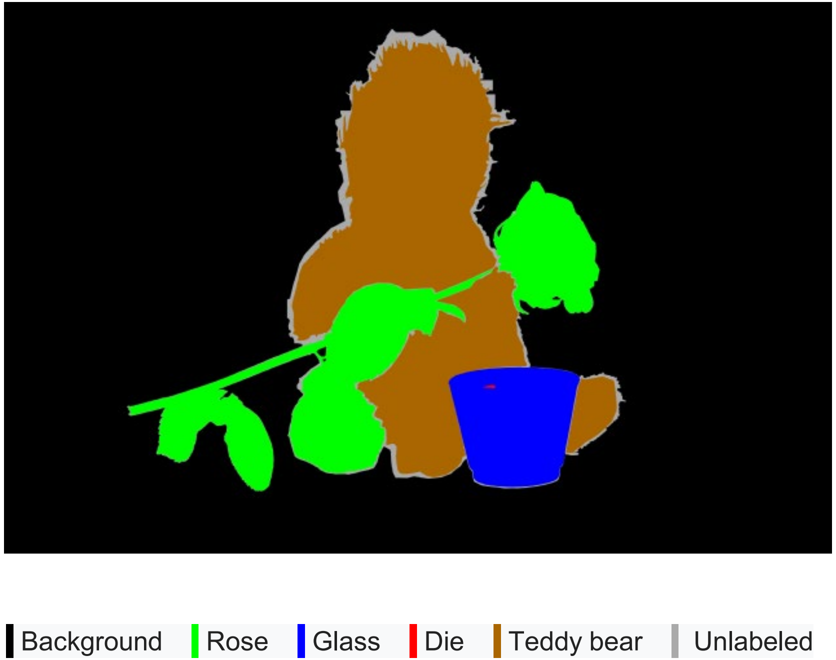
Semantic Segmentation

Semantic Segmentation

- Put label on each pixel in the image
- No need to specify the difference between the instances

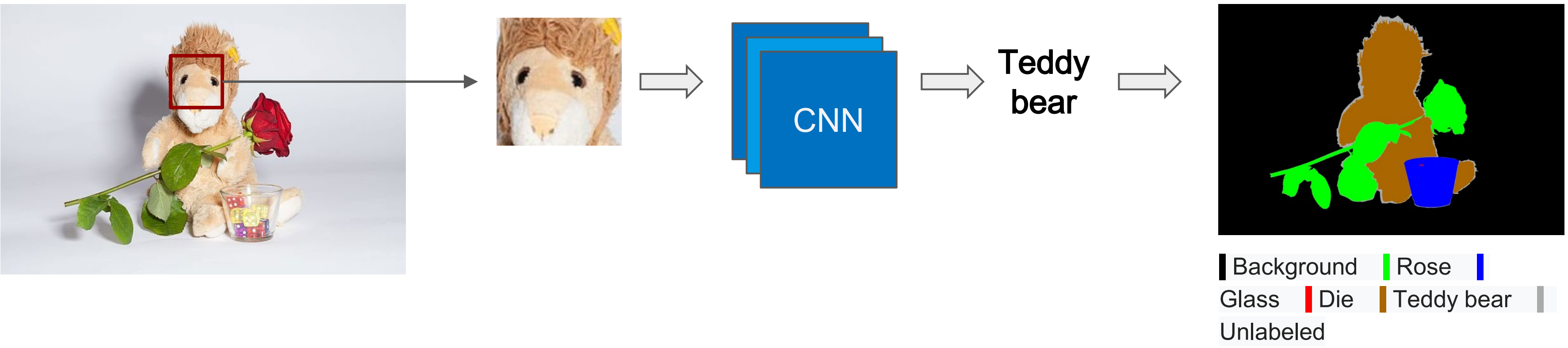


Example

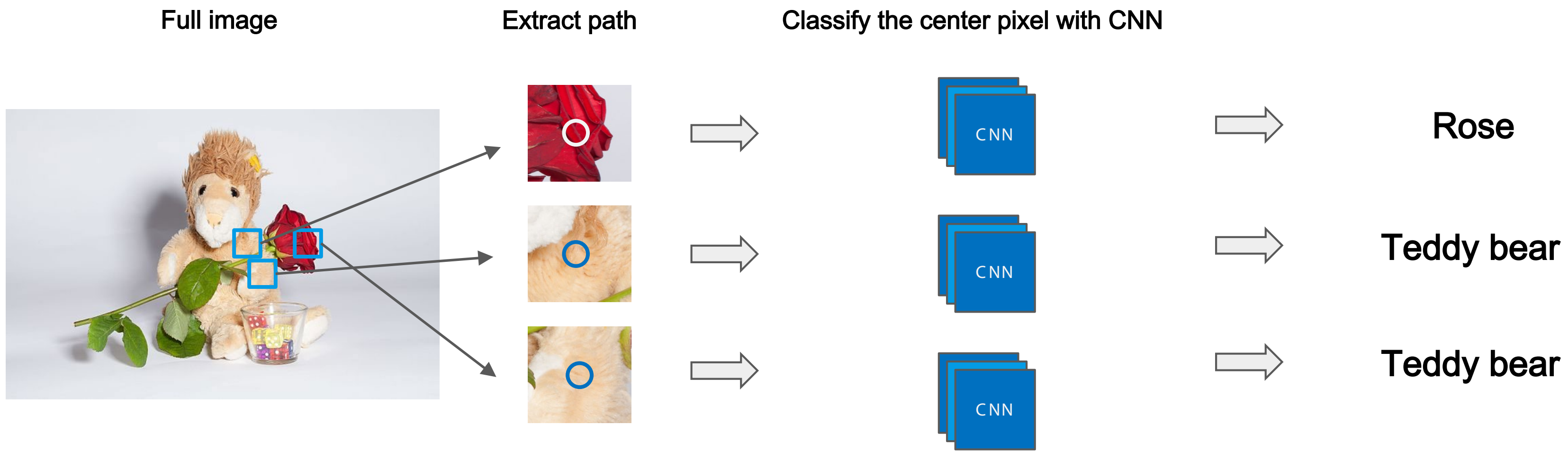


Semantic Segmentation

Extract path | Run through a CNN | Classify center pixel | Repeat for every pixel

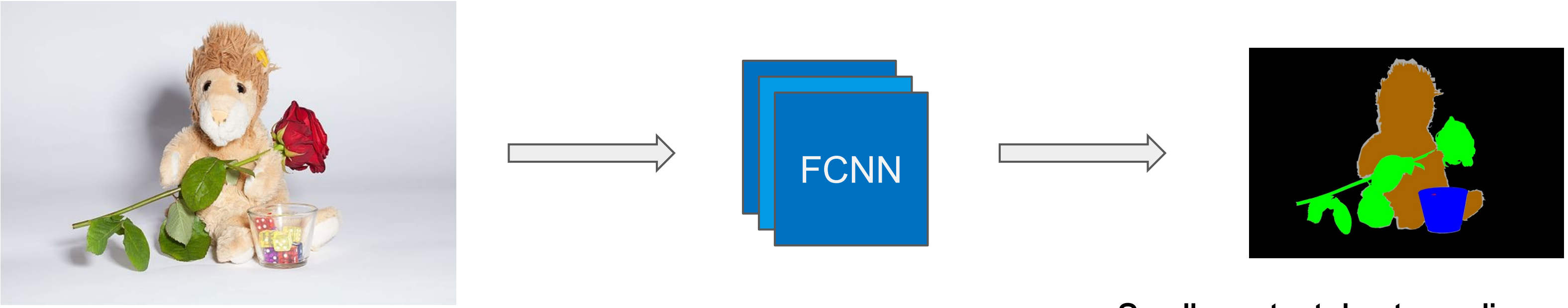


Semantic segmentation: sliding window



Semantic Segmentation

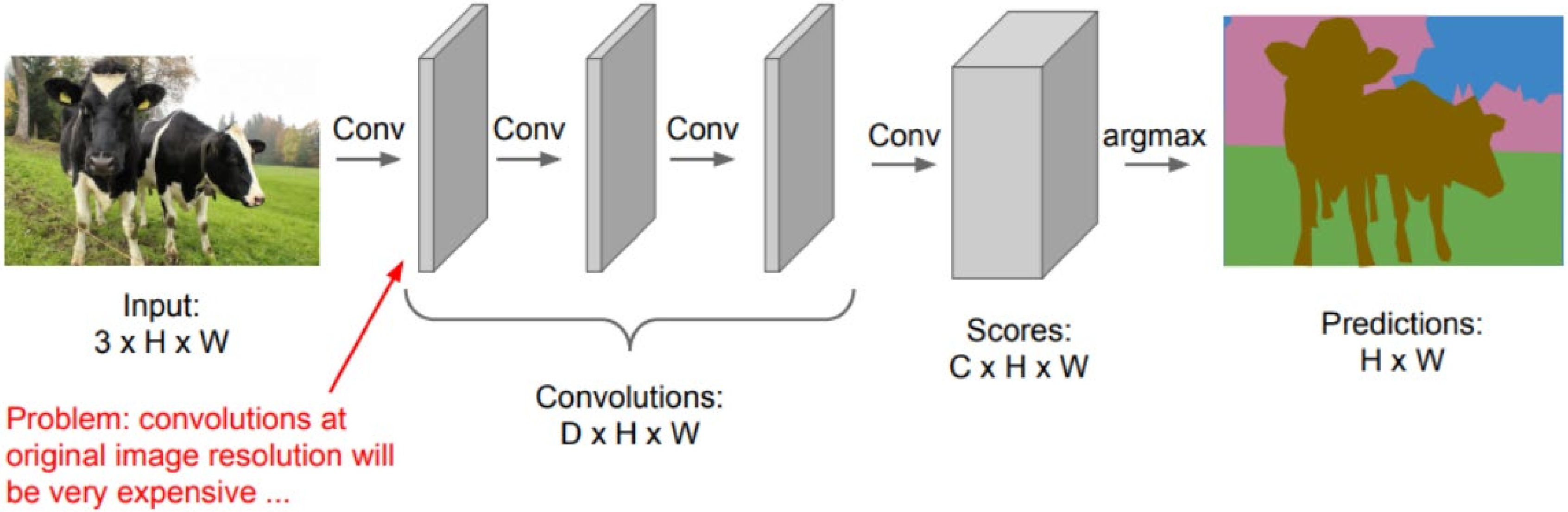
- Run through a fully convolutional network to get all pixels at once



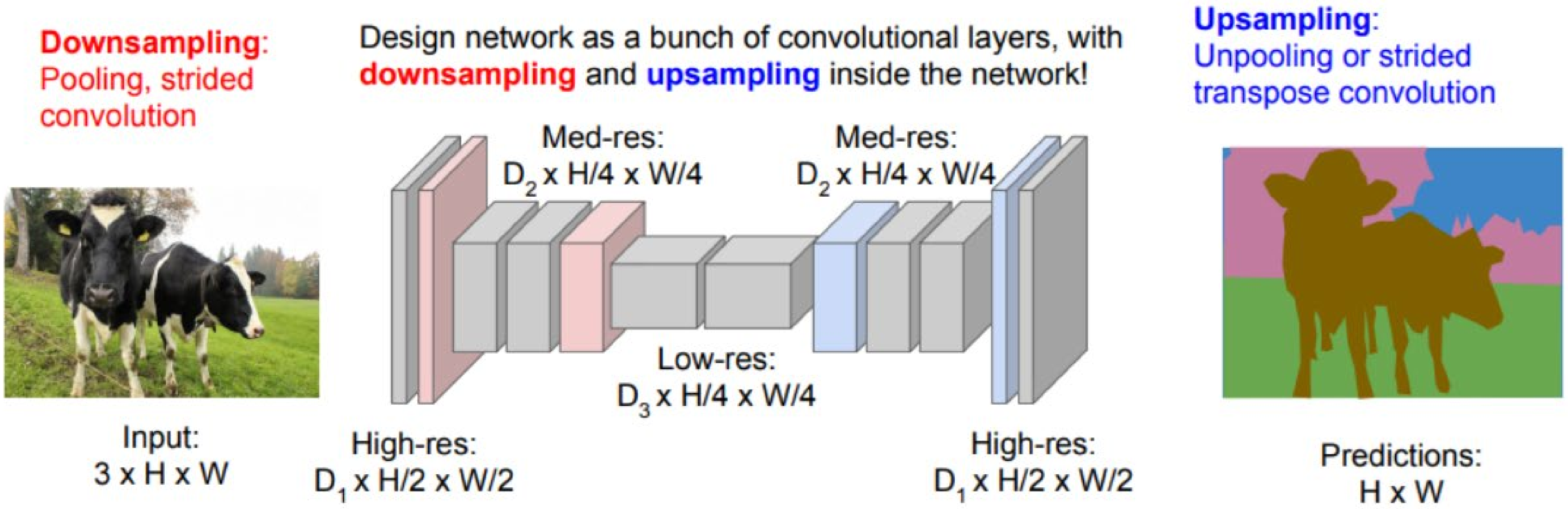
Smaller output due to pooling

Semantic Segmentation Idea : FCN

Design a network as a bunch of convolutional layers
to make predictions for pixels all at once!



Semantic Segmentation Idea : FCN

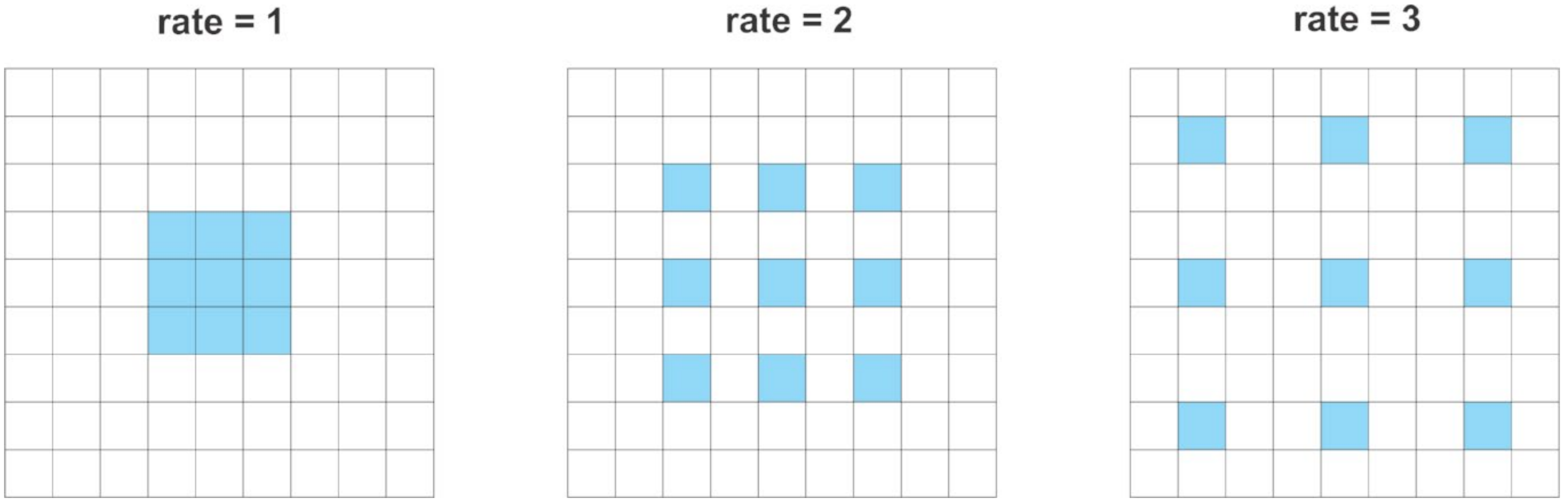


Long, Shelhamer, and Darrell, "Fully Convolutional Networks for Semantic Segmentation", CVPR 2015
Noh et al, "Learning Deconvolution Network for Semantic Segmentation", ICCV 2015

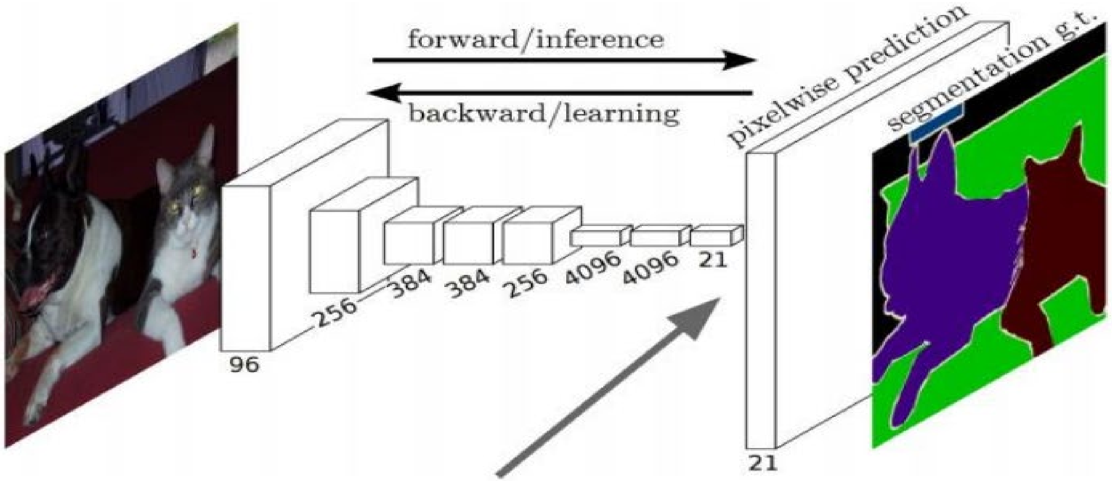
Atrous/Dilated Convolutions

Atrous (or dilated) convolutions are regular convolutions with a factor that allows us to expand the filter's field of view.

Consider a *3x3* convolution filter for instance. When the dilation rate is equal to 1, it behaves like a standard convolution. But, if we set the dilation factor to 2, it has the effect of enlarging the convolution kernel.



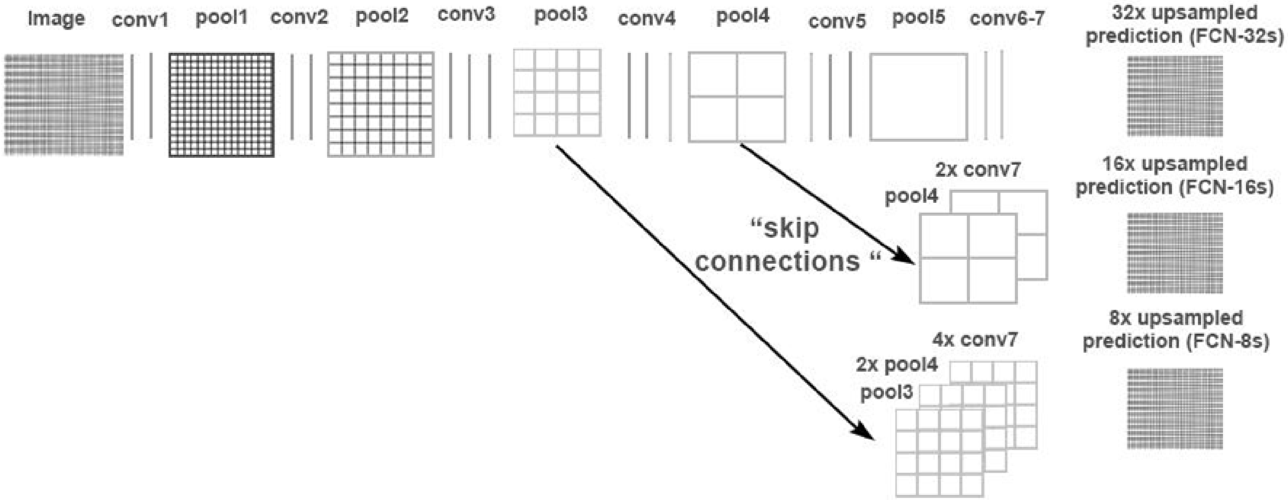
Semantic Segmentation : UpSampling



Learnable upsampling!

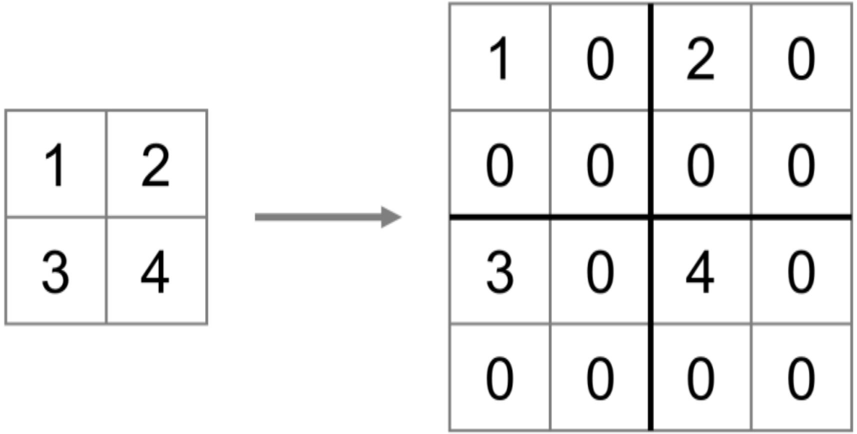
Long, Shelhamer, and Darrell, “Fully Convolutional Networks for Semantic Segmentation”, CVPR 2015

Semantic Segmentation : UpSampling



UpSampling

- Transposed convolution / Deconvolution
- Fractionally strided convolution
- Max-unpooling: Preserves spatial information



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

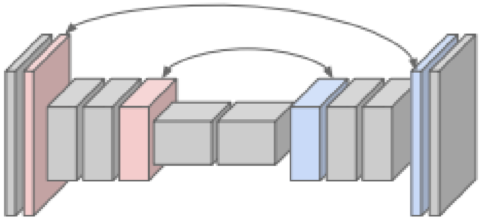
1	2
3	4

Input: 2 x 2

0	0	2	0
0	1	0	0
0	0	0	0
3	0	0	4

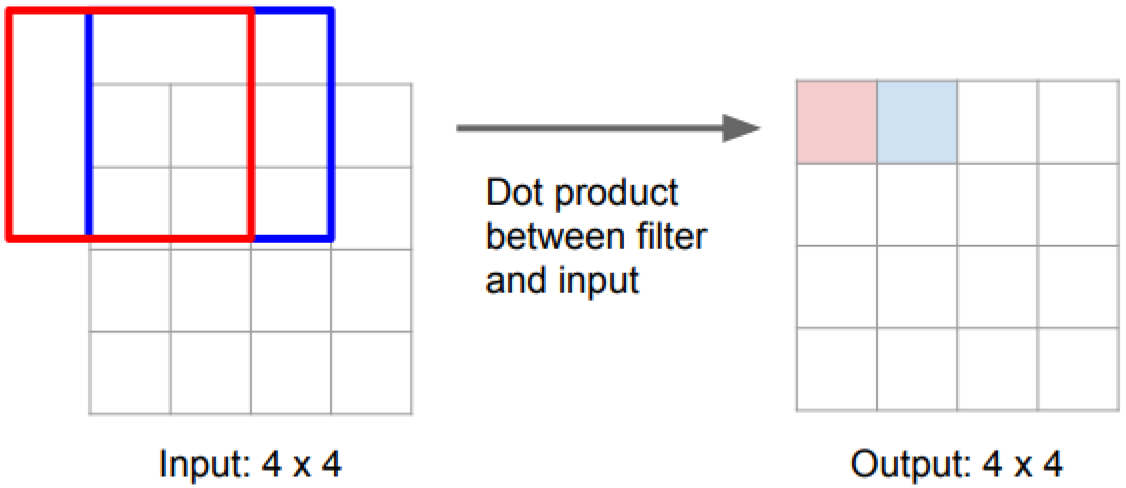
Output: 4 x 4

Corresponding pairs of downsampling and upsampling layers



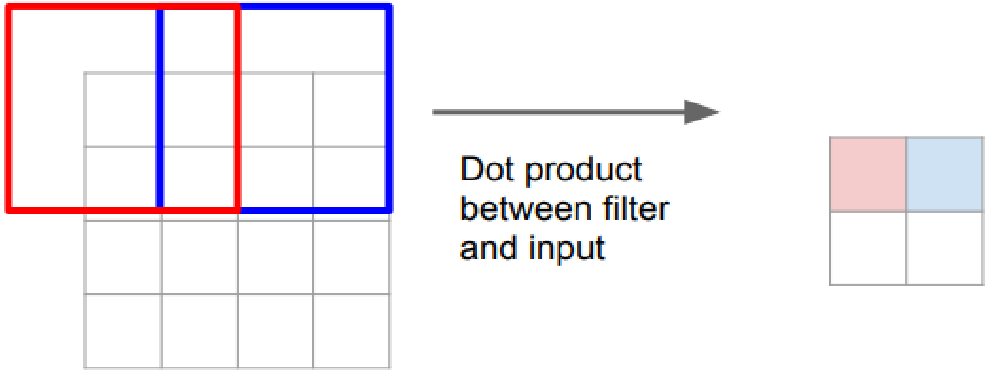
Learnable Upsampling: Transpose Convolution

Recall: Normal 3 x 3 convolution, stride 1 pad 1



Learnable Upsampling: Transpose Convolution

Recall: Normal 3 x 3 convolution, stride 2 pad 1



Input: 4 x 4

Output: 2 x 2

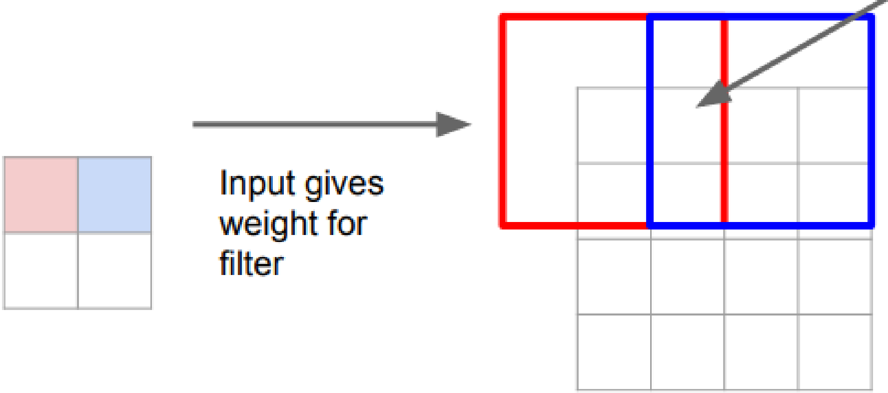
Filter moves 2 pixels in the input for every one pixel in the output

Stride gives ratio between movement in input and output

Learnable Upsampling: Transpose Convolution

3 x 3 **transpose** convolution, stride 2 pad 1

Other names:
-Deconvolution (bad)
-Upconvolution
-Fractionally strided convolution
-Backward strided convolution



Input: 2 x 2

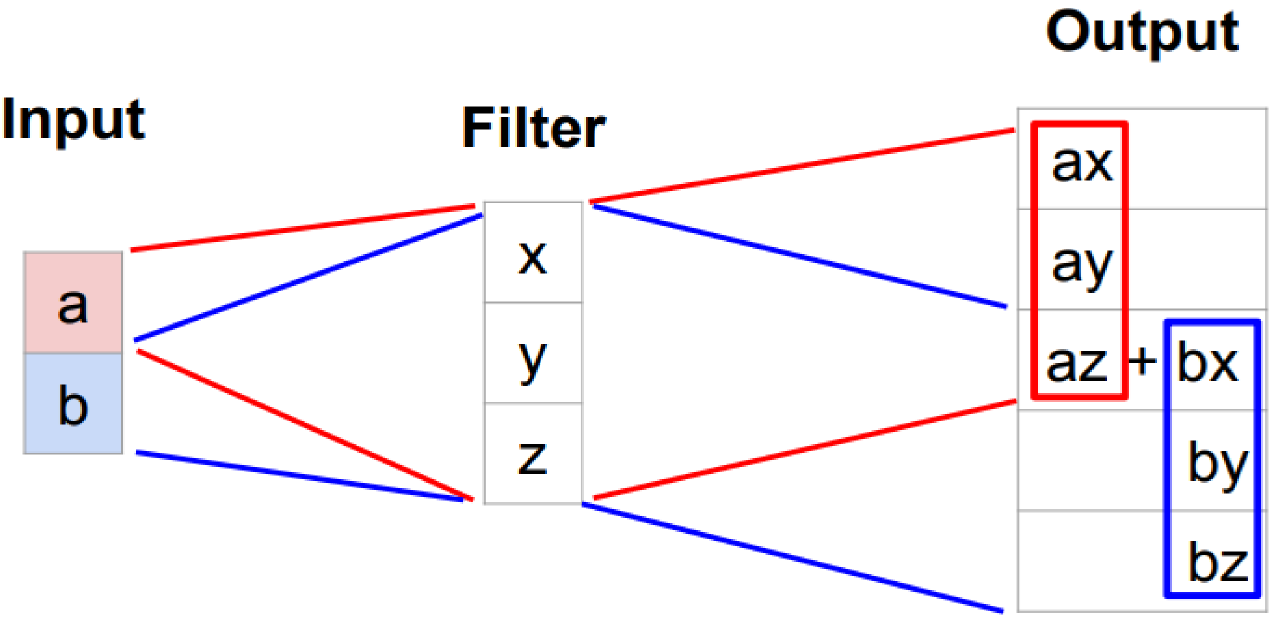
Output: 4 x 4

Sum where output overlaps

Filter moves 2 pixels in the <u>output</u> for every one pixel in the <u>input</u>

Stride gives ratio between movement in output and input

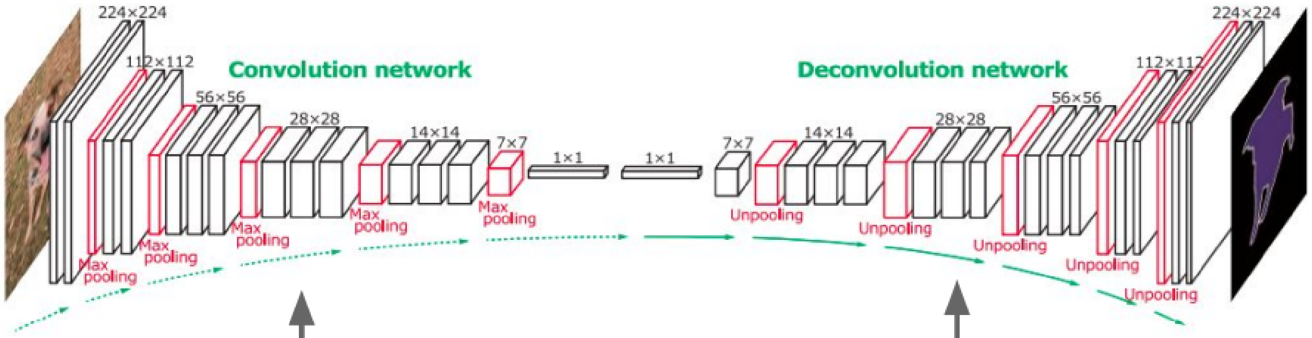
Learnable Upsampling: 1D Example



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

Need to crop one pixel from output to make output exactly 2x input

Semantic Segmentation: Upsampling



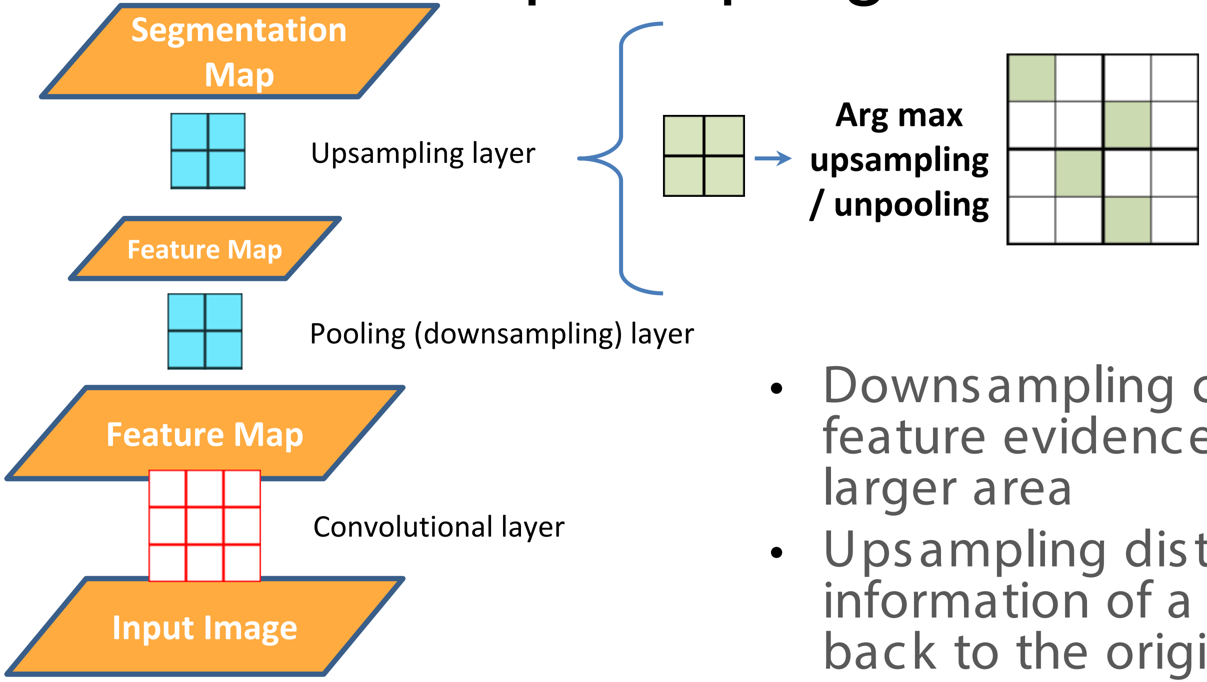
Normal VGG

"Upside down" VGG

Noh et al, "Learning Deconvolution Network for Semantic Segmentation", ICCV 2015

6 days of training on Titan X…

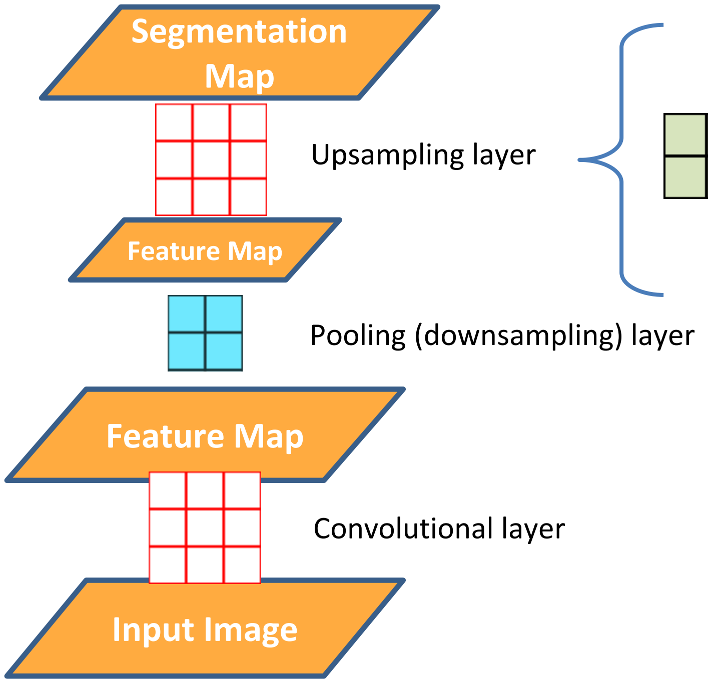
To produce a segmentation map downsampling is followed by upsampling

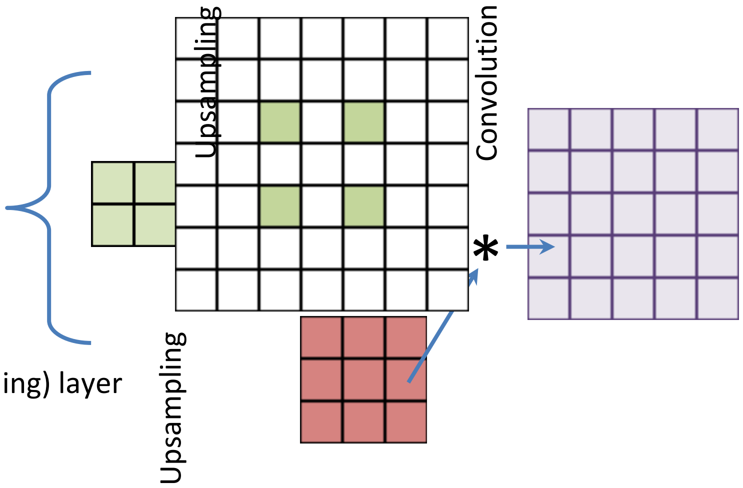


- Downsampling collects feature evidence from a larger area
- Upsampling distributes the information of a segment back to the original pixel domain

Upsampling can also be learned

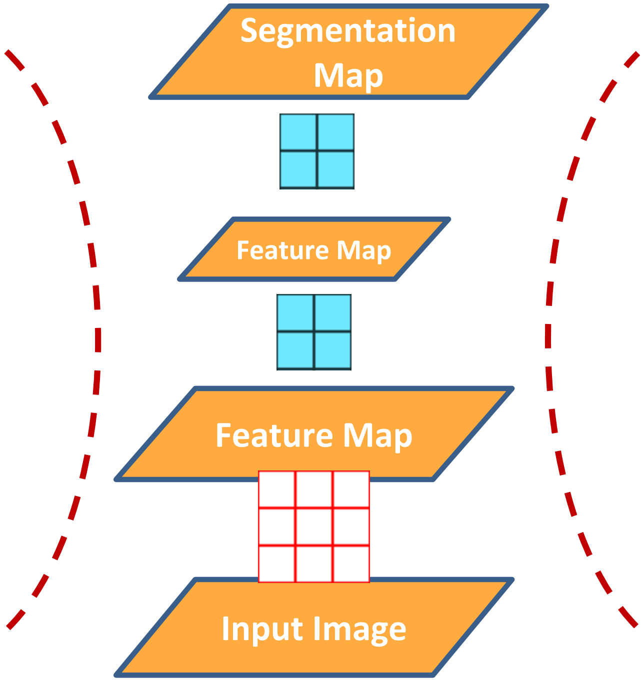
greatlearning
Learning for Life



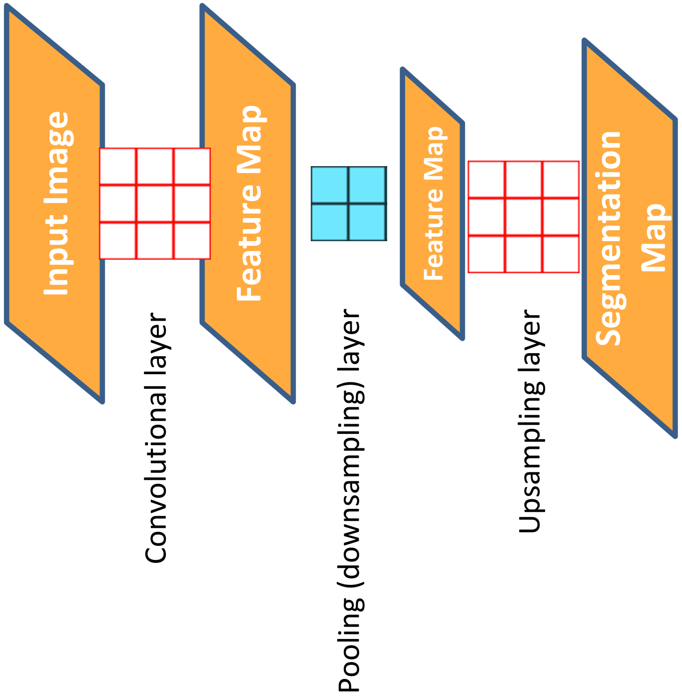


Transposed convolution

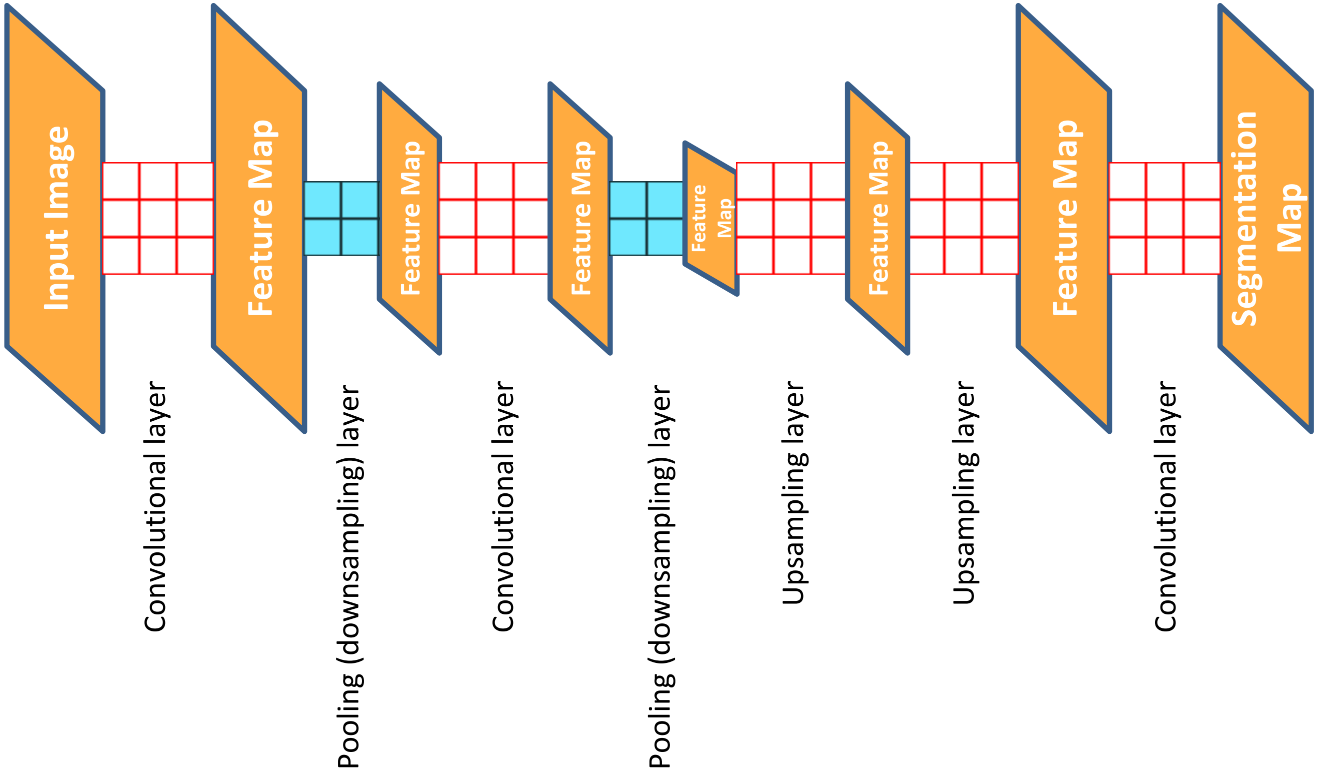
Downsampling and upsampling leads to an hour-glass structure



Let us re-arrange the layers horizontally

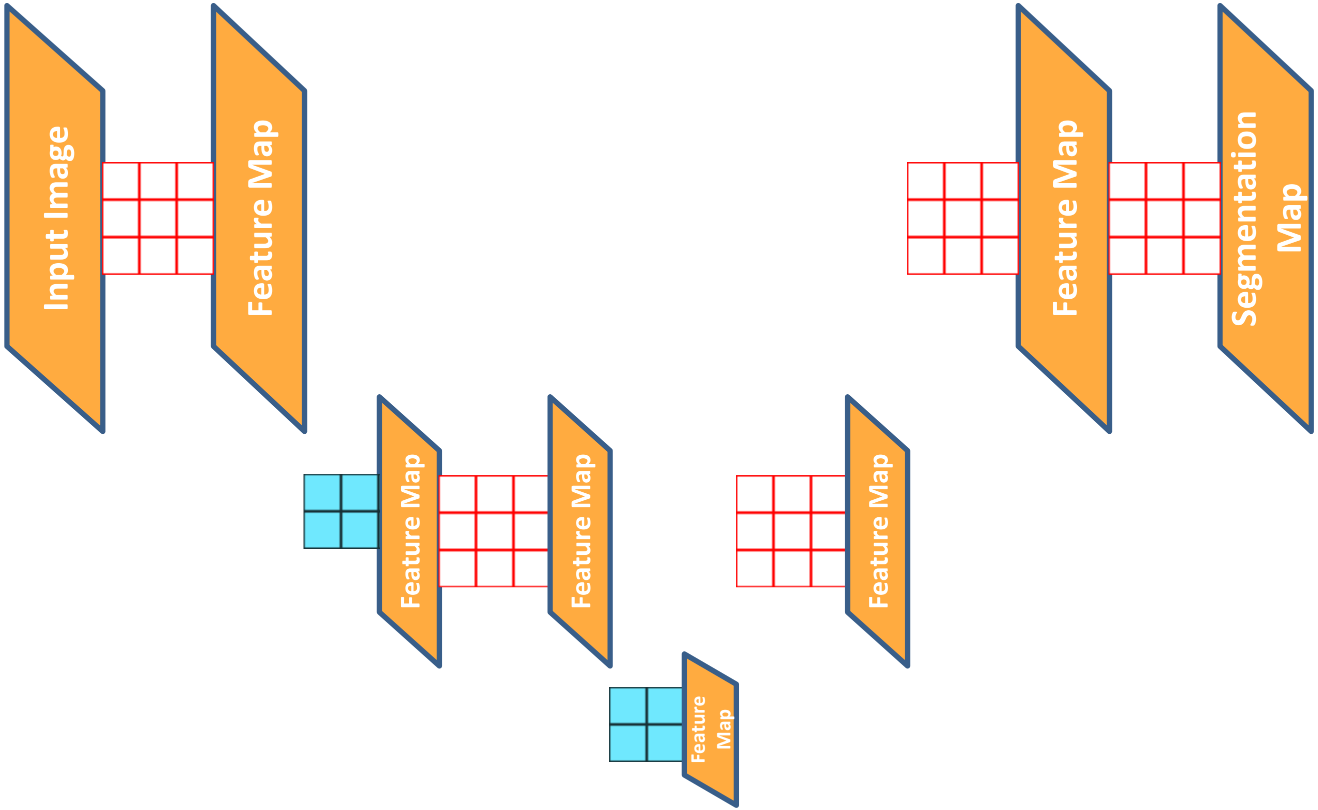


More layers can be added

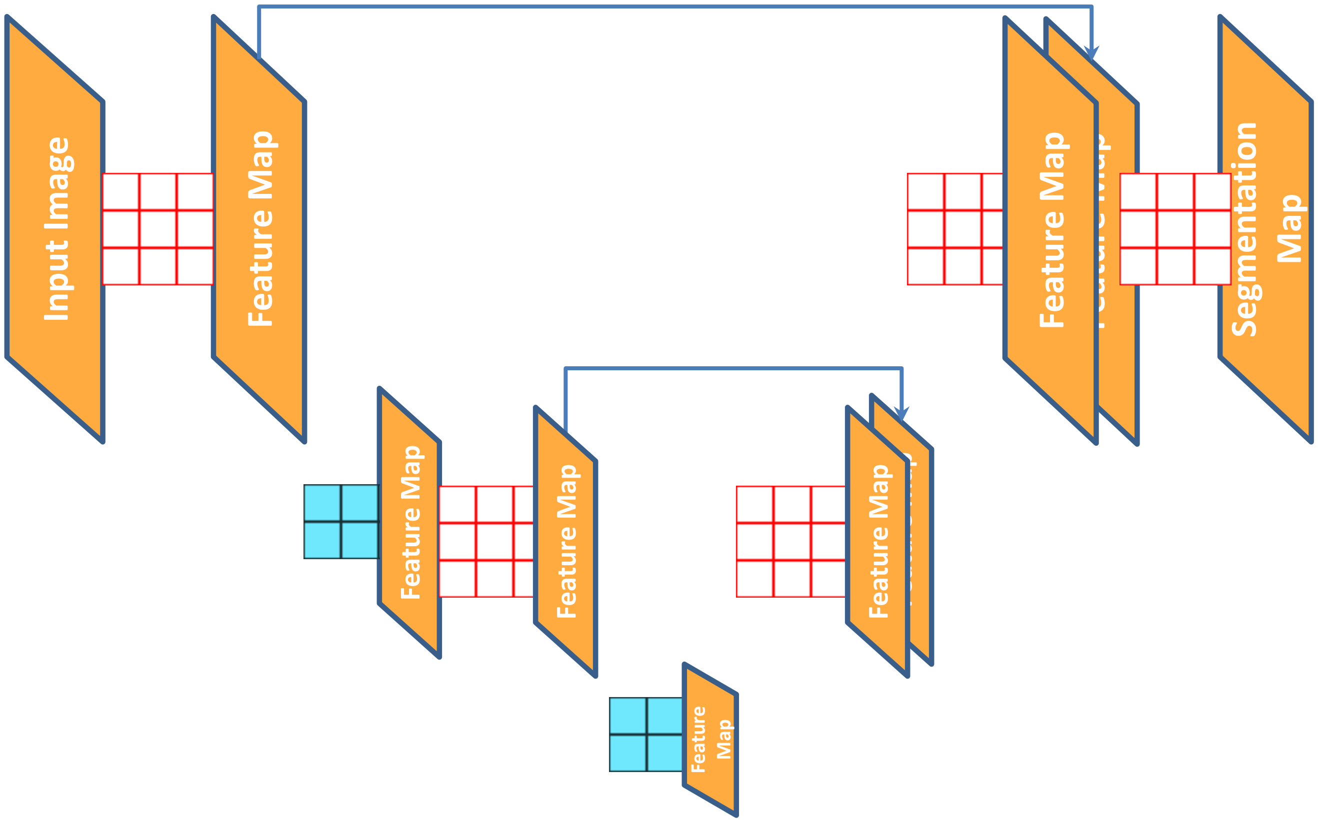


Visually rearrange layers in a big U

greatlearning
Learning for Life



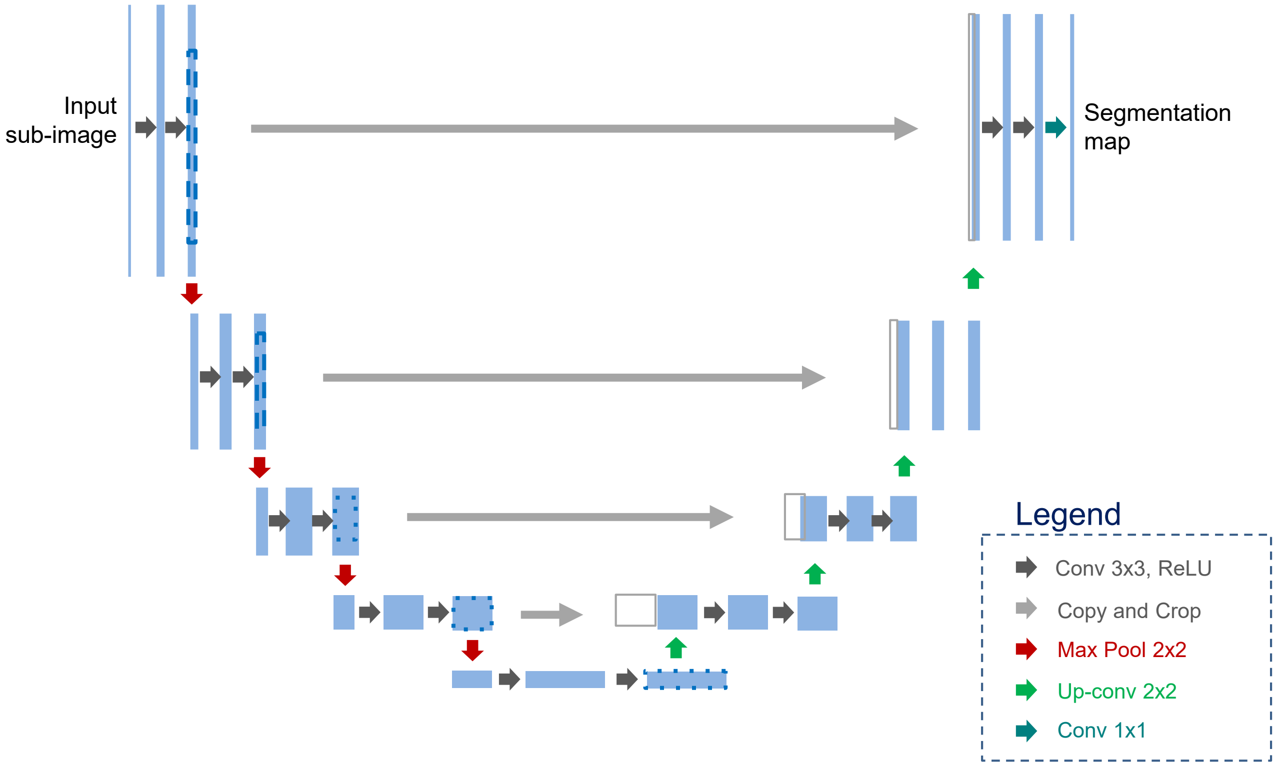
Concatenate previous feature maps for finer spatial context



U-Net

U-Net is based on the ideas described in the previous slides

greatlearning
Learning for Life



Source: "U-Net: Convolutional Networks for Biomedical Image Segmentation" Olaf Ronneberger, Philipp Fischer, Thomas Brox, 2015

U-Net

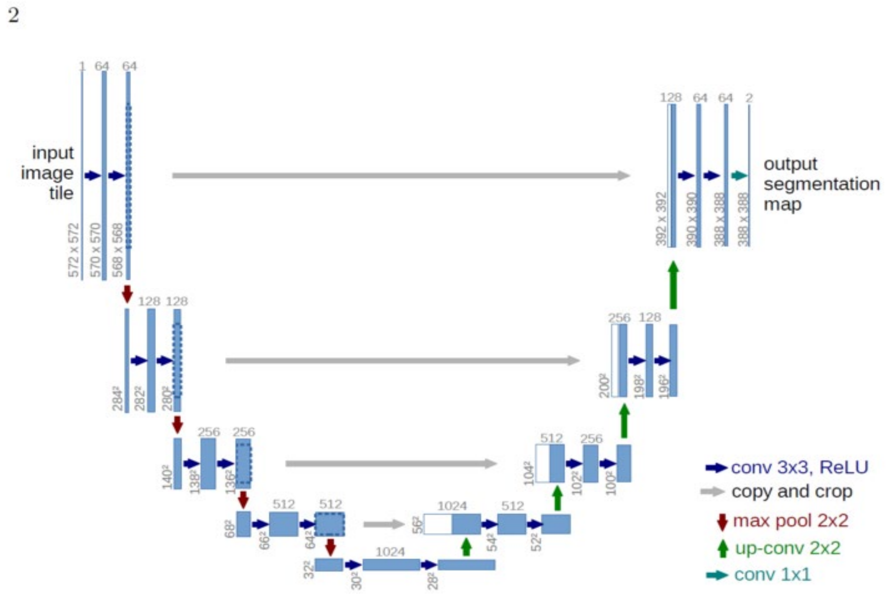


Fig. 1. U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations.

The architecture has -

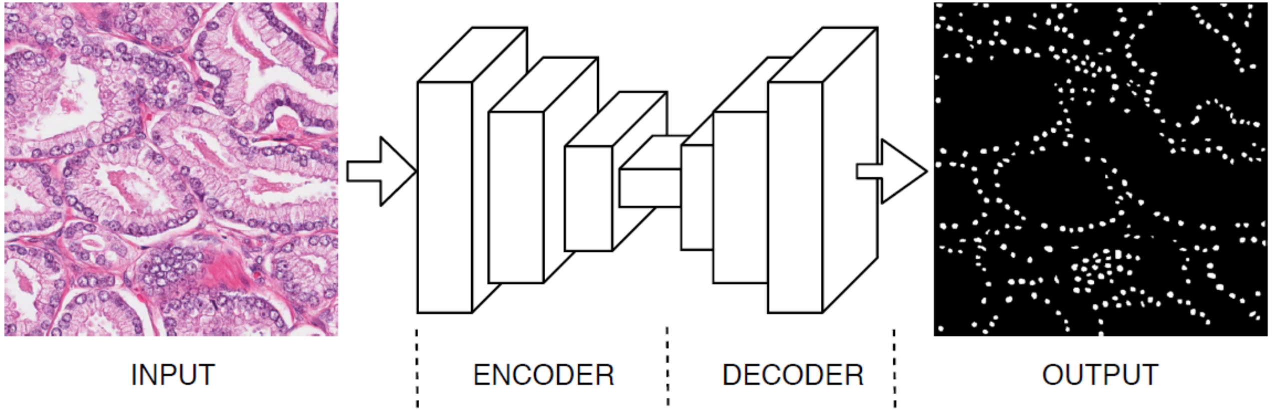
An Encoder - Downsampling part. It is used to get context in the image. It is just a stack of convolutional and max pooling layers.

A Decoder - Symmetric Upsampling part. It is used for precise localization. Transposed convolution is used for upsampling.

It is a fully convolutional network (FCN). it has Convolutional layers and it does not have any dense layer so it can work for image of any size.

A sample output for nucleus segmentation in pathology

greatlearning
Learning for Life



A general representation of fully convolutional networks. The encoder is composed of convolutional and pooling layers for downsampling and the decoder is composed of deconvolutional layers for upsampling.

Dice coefficient

Dice coefficient is defined as follows:

$$\frac{2 * |X \cap Y|}{|X| + |Y|}$$

X is the predicted set of pixels and Y is the ground truth.

A higher dice coefficient is better. A dice coefficient of 1 can be achieved when there is perfect overlap between X and Y. Since the denominator is constant, the only way to maximize this metric is to increase overlap between X and Y.

For more info on Dice coefficient: https://www.kaggle.com/c/carvana -image-masking-challenge#evaluation

References

https://arxiv.org/pdf/1803.02758.pdf

https://people.eecs.berkeley.edu/~jonlong/long_shelhamer_fcn.pdf

Paper - U-Net: Convolutional Networks for Biomedical Image Segmentation

Thank you!

Happy Learning :)