

Semantic Segmentation

Computer Vision

This file is meant for personal use by rg.ravigupta91@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Proprietary Content. © Great Learning. All Rights Reserved. Unauthorized use or distribution is prohibited.

Outline

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

Semantic Segmentation

This file is meant for personal use by rg.ravigupta91@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Proprietary Content. © Great Learning. All Rights Reserved. Unauthorized use or distribution is prohibited.

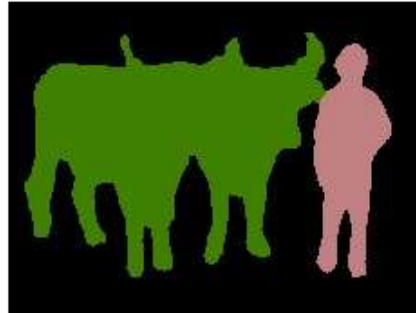
Semantic Segmentation

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

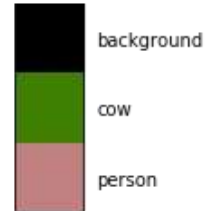
input image



segmentation map



segmentation overlay



Example



Background Rose Glass Die Teddy bear Unlabeled

This file is meant for personal use by rg.ravigupta91@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Proprietary Content. © Great Learning. All Rights Reserved. Unauthorized use or distribution is prohibited.

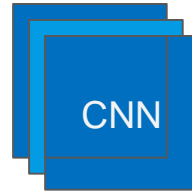
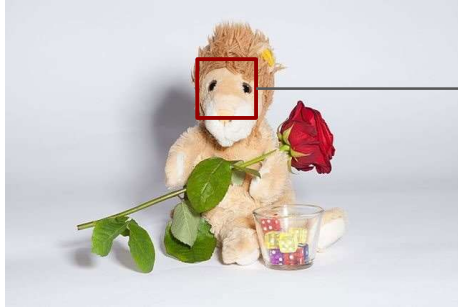
Semantic Segmentation

Extract
path

Run through a
CNN

Classify center
pixel

Repeat for every
pixel

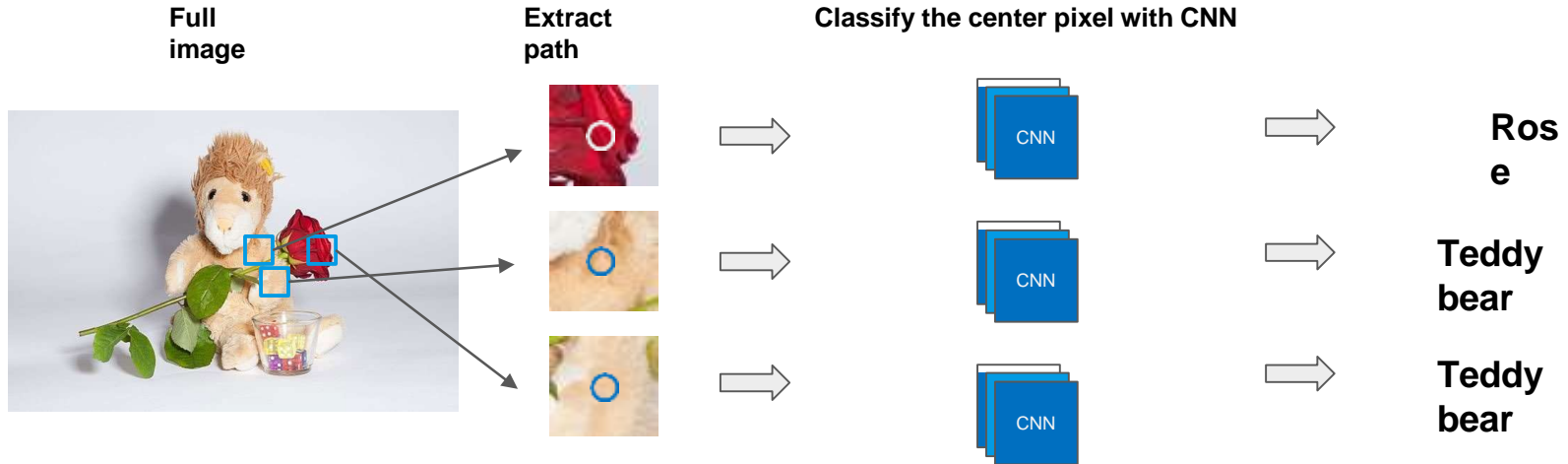


**Tedd
y
bear**



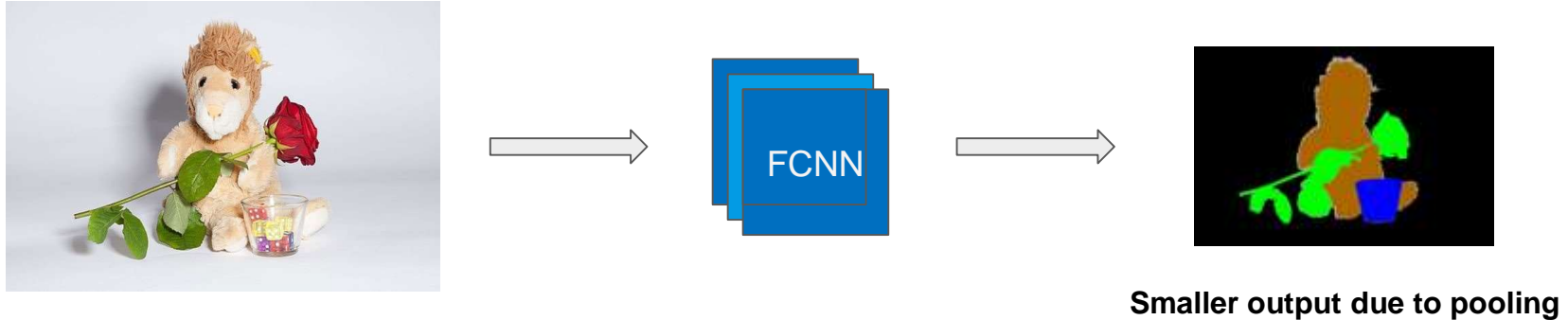
Background	Rose
Glass	Die
Unlabeled	Teddy bear

Semantic segmentation: sliding window



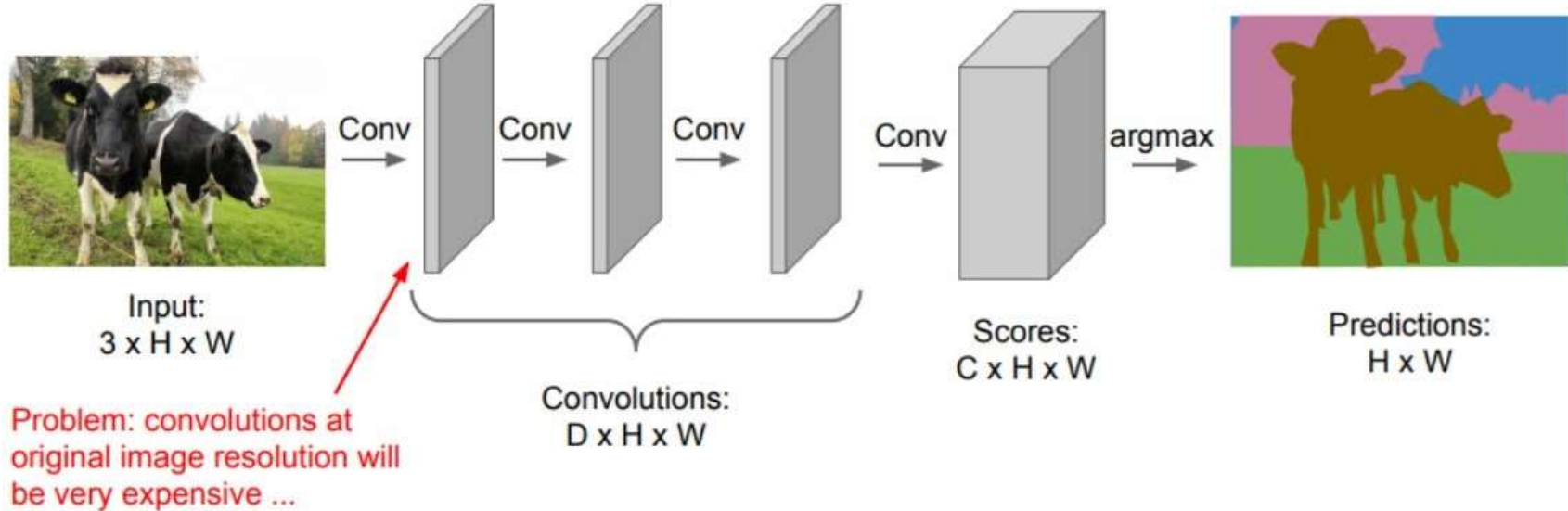
Semantic Segmentation

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



Semantic Segmentation Idea : FCN

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



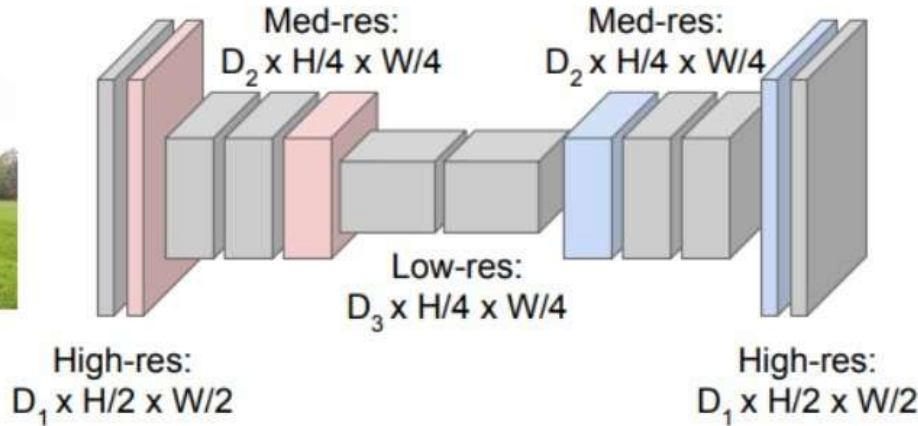
Semantic Segmentation Idea : FCN

Downsampling:
Pooling, strided
convolution



Input:
 $3 \times H \times W$

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



Upsampling:
Unpooling or strided
transpose convolution



Predictions:
 $H \times W$

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

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

This file is meant for personal use by rg.ravigupta91@gmail.com only.

Sharing or publishing the contents in part or full is liable for legal action.

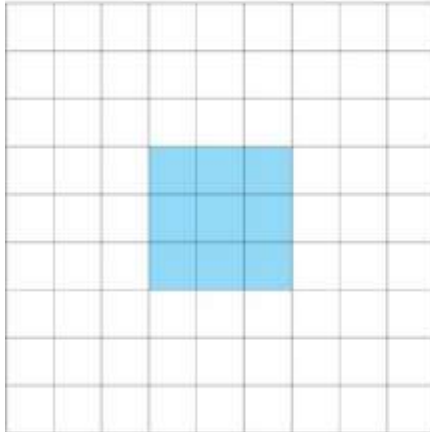
Proprietary Content. © Great Learning. All Rights Reserved. Unauthorized use or distribution is prohibited.

Atrous/Dilated Convolutions

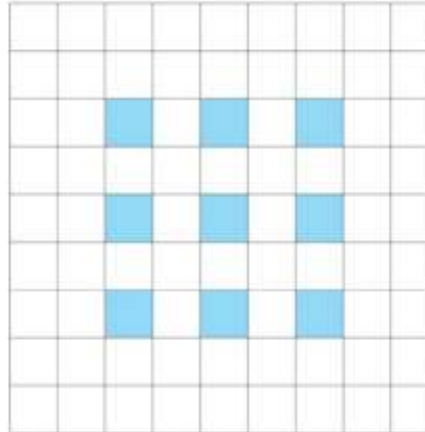
Atrous (or dilated) convolutions are regular convolutions with a factor that allows us to expand the filter's field of view.

Consider a 3×3 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.

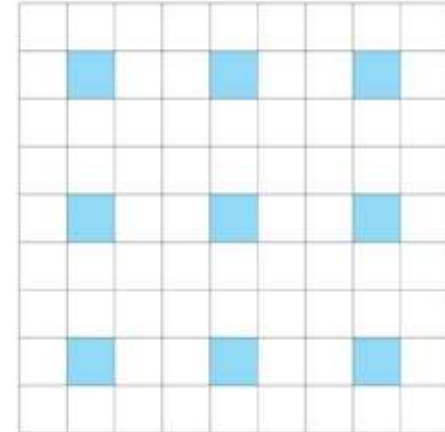
rate = 1



rate = 2

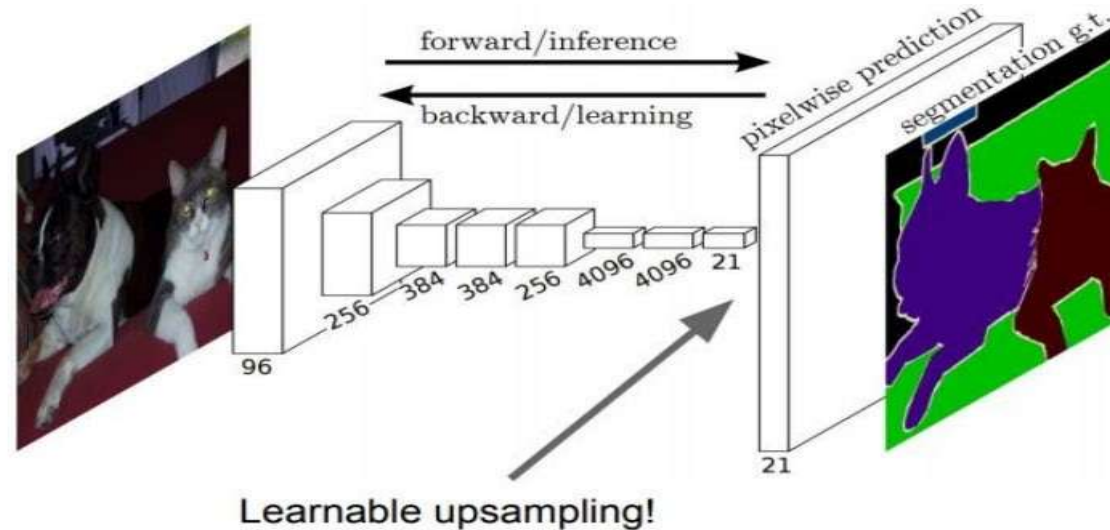


rate = 3



This file is meant for personal use by rg.ravigupta91@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

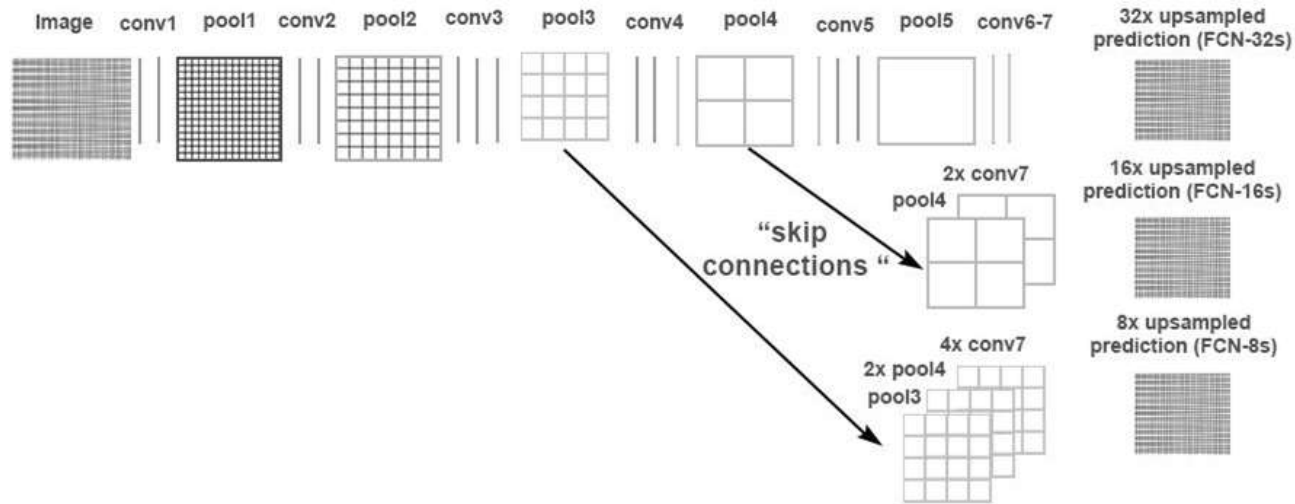
Semantic Segmentation : UpSampling



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

This file is meant for personal use by rg.ravigupta91@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

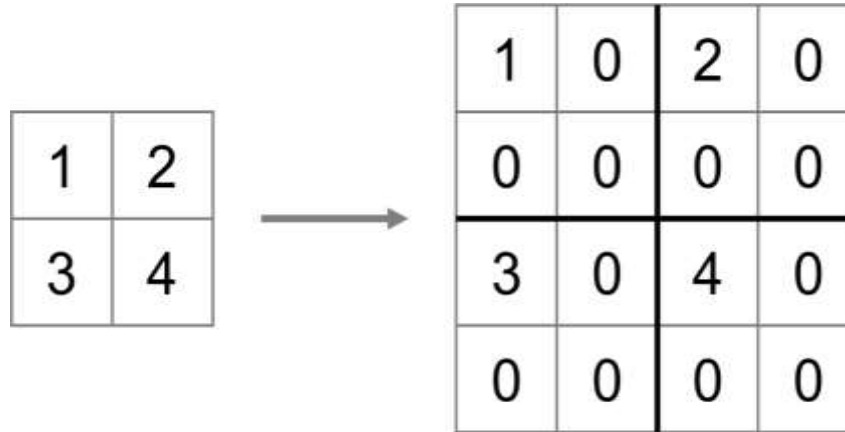
Semantic Segmentation : UpSampling



This file is meant for personal use by rg.ravigupta91@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

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

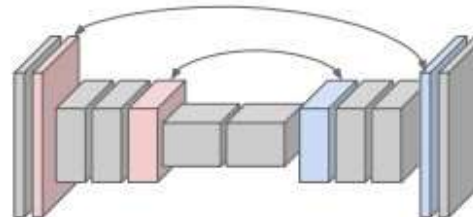
1	2
3	4

Input: 2 x 2

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

Output: 4 x 4

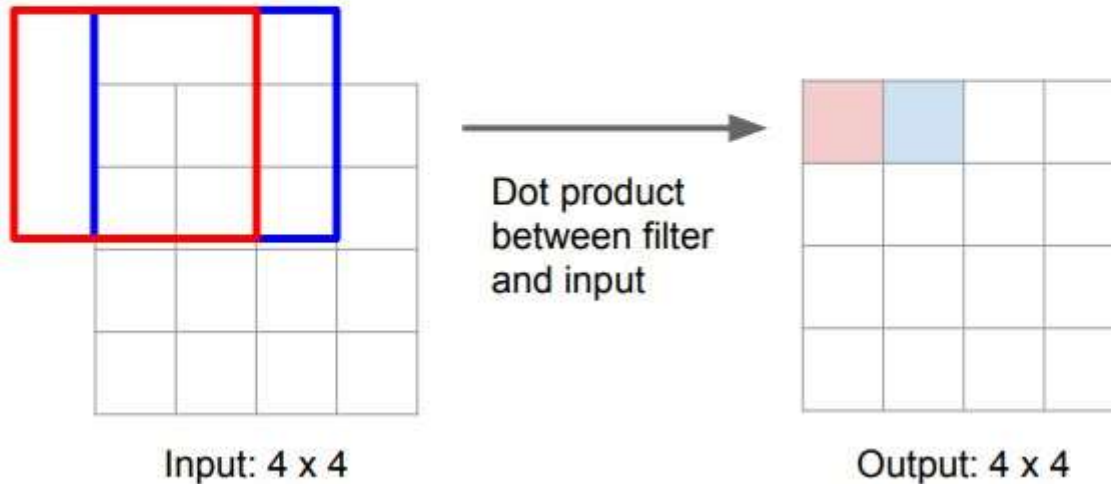
Corresponding pairs of
downsampling and
upsampling layers



This file is meant for personal use by rg.ravigupta91@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

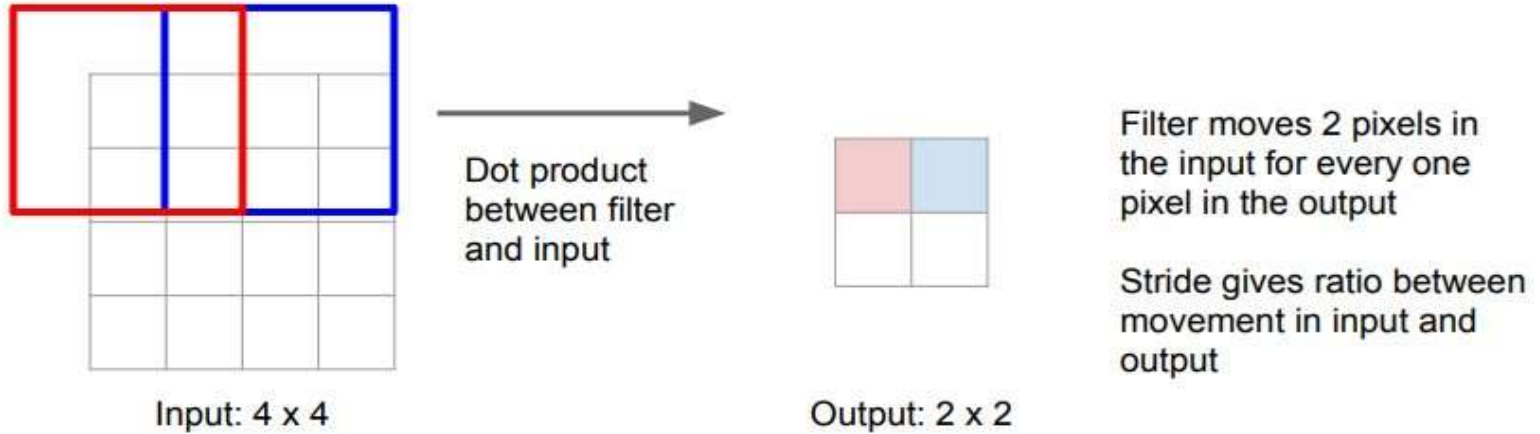
Learnable Upsampling: Transpose Convolution

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



Learnable Upsampling: Transpose Convolution

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

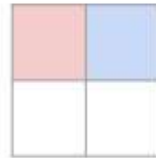


Learnable Upsampling: Transpose Convolution

Other names:

- Deconvolution (bad)
- Upconvolution
- Fractionally strided convolution
- Backward strided convolution

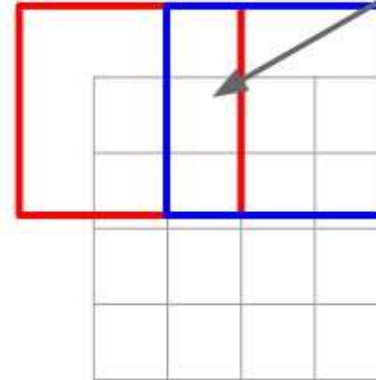
3 x 3 **transpose** convolution, stride 2 pad 1



Input: 2 x 2



Input gives weight for filter



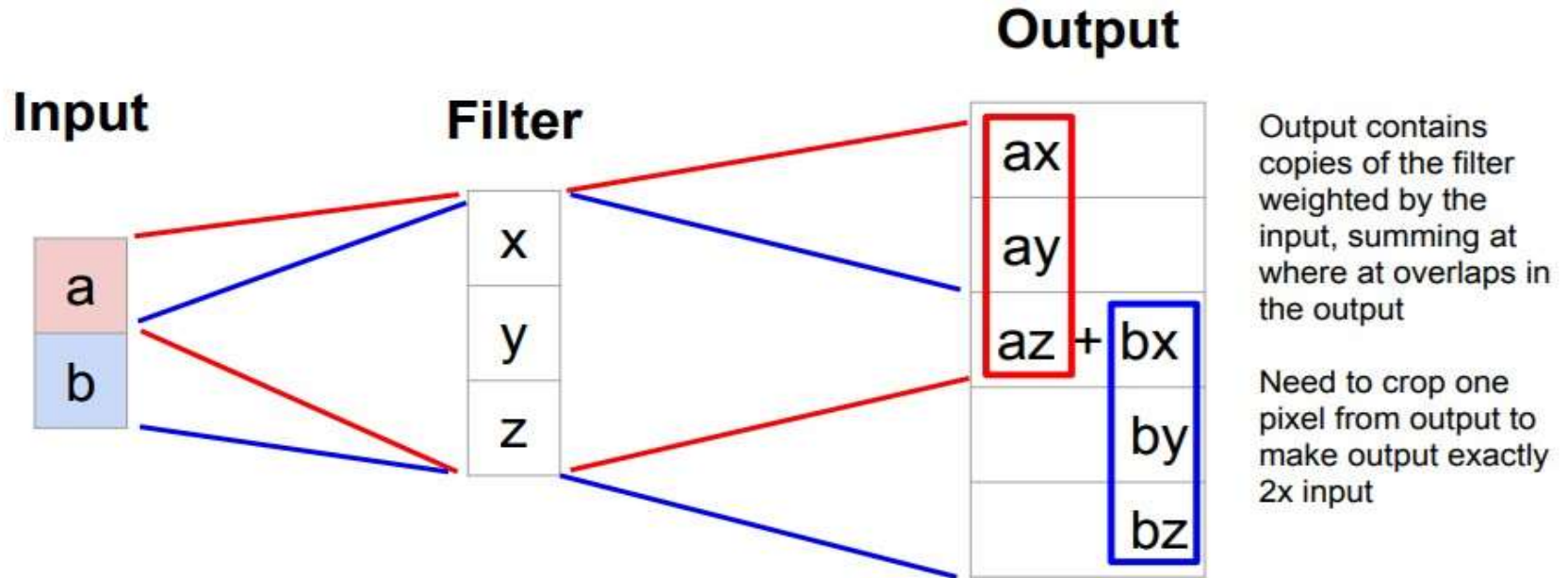
Output: 4 x 4

Sum where output overlaps

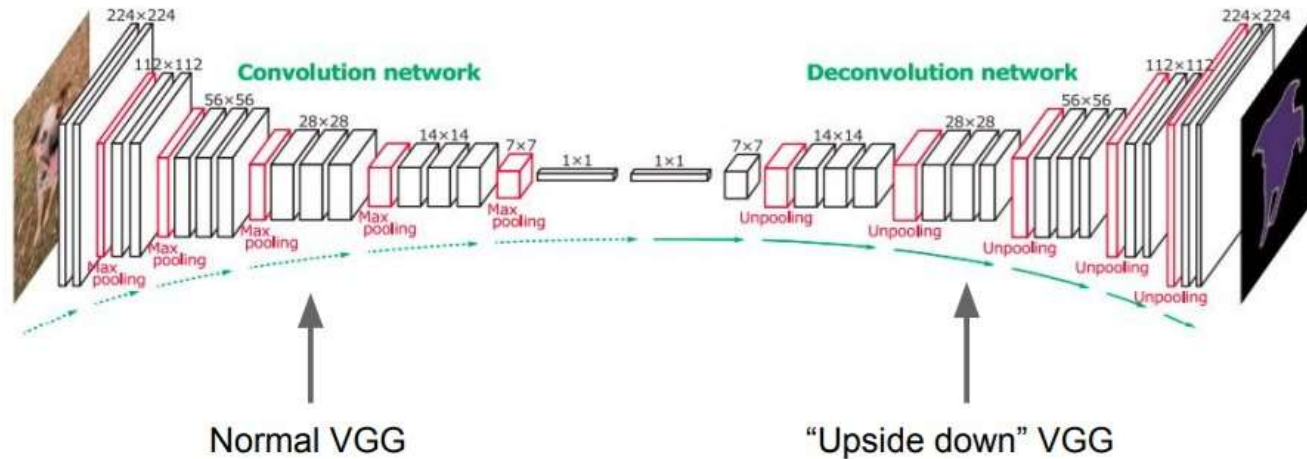
Filter moves 2 pixels in the output for every one pixel in the input

Stride gives ratio between movement in output and input

Learnable Upsampling: 1D Example



Semantic Segmentation: Upsampling



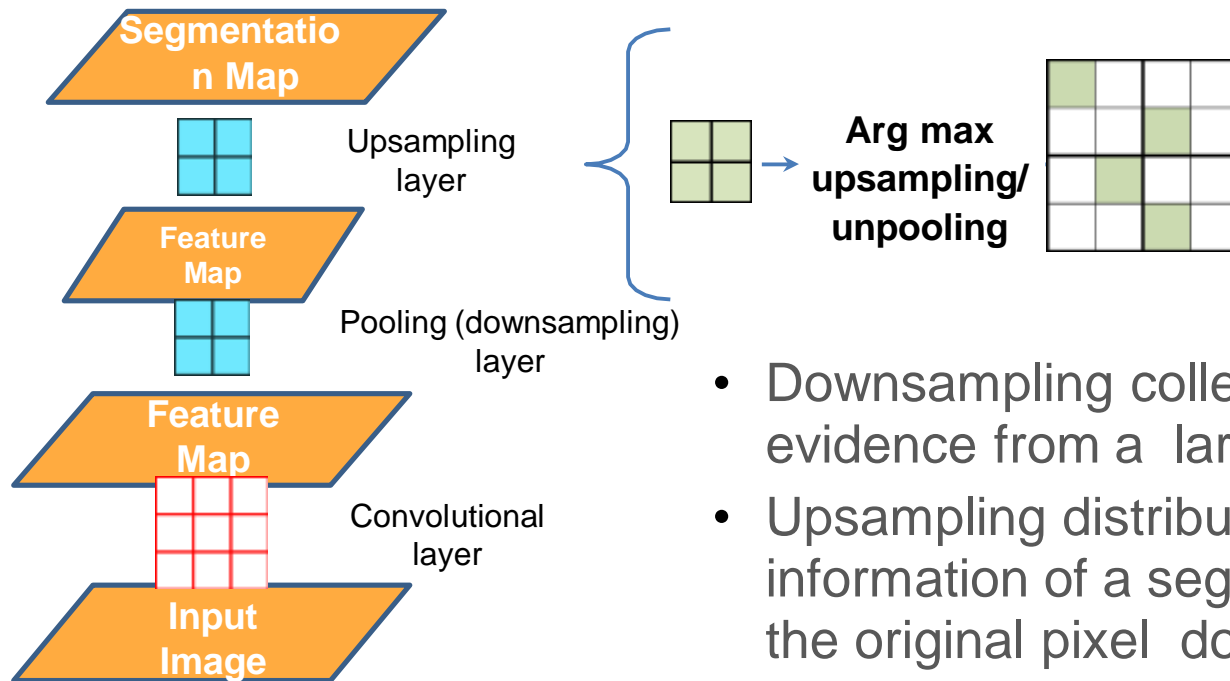
Noh et al, "Learning Deconvolution Network for Semantic Segmentation", ICCV 2015

6 days of training on Titan X...

This file is meant for personal use by rg.ravigupta91@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

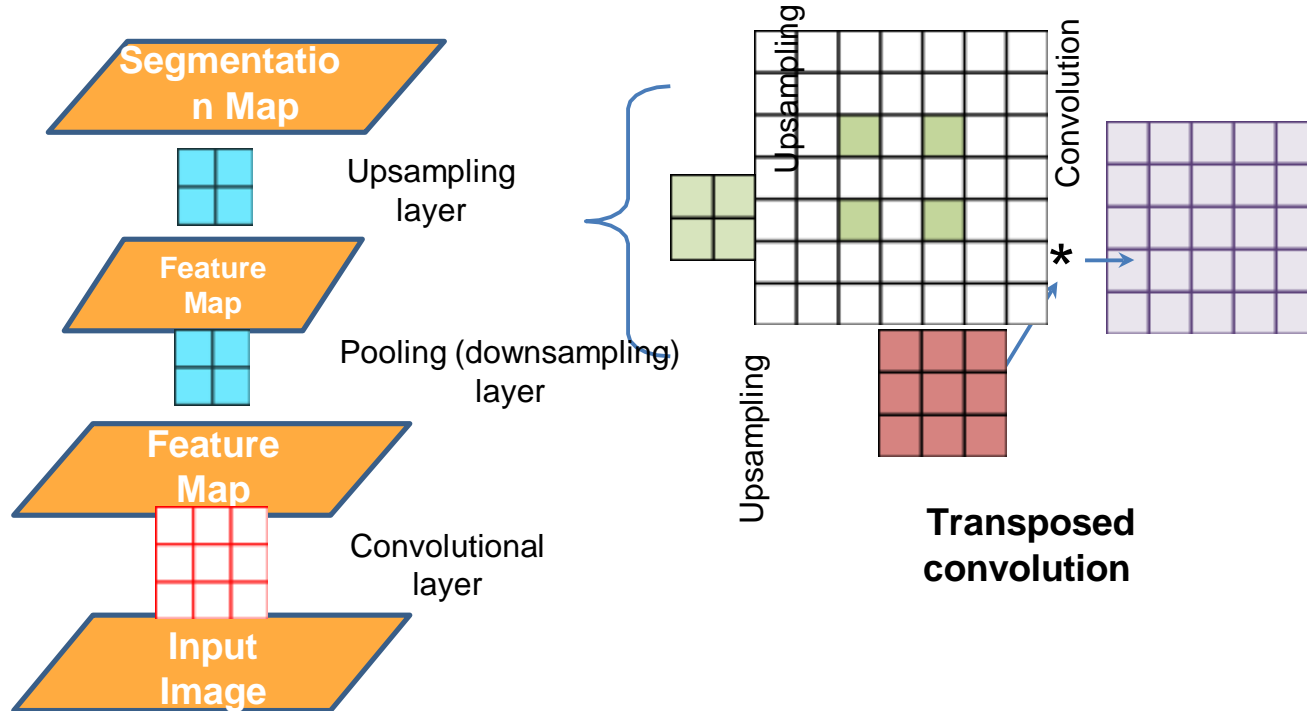
Copyright © Great Learning. All Rights Reserved. Unauthorized use or distribution is prohibited.

To produce a segmentation map downsampling is followed by upsampling

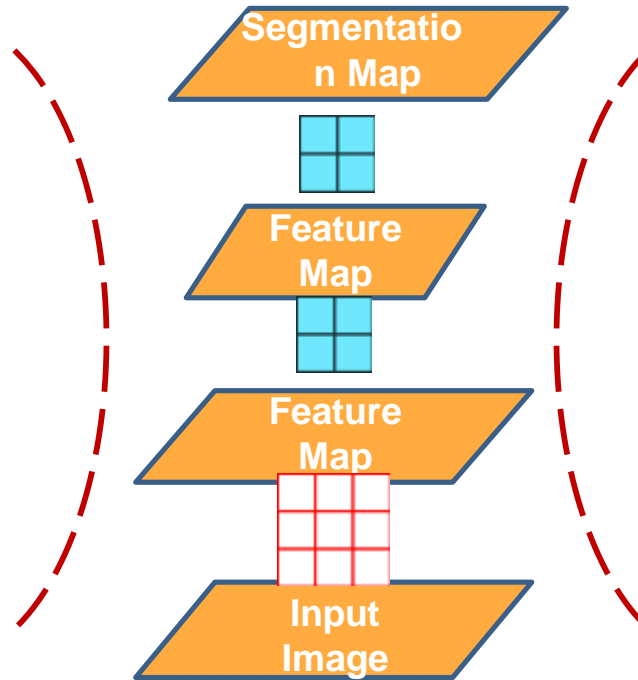


- Downsampling collects feature evidence from a larger area
- Upsampling distributes the information of a segment back to the original pixel domain

Upsampling can also be learned

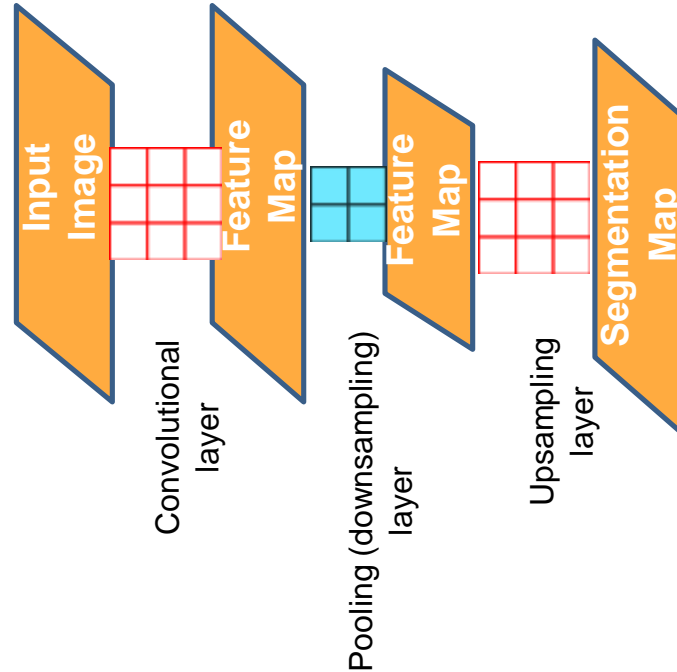


Downsampling and upsampling leads to an hour-glass structure



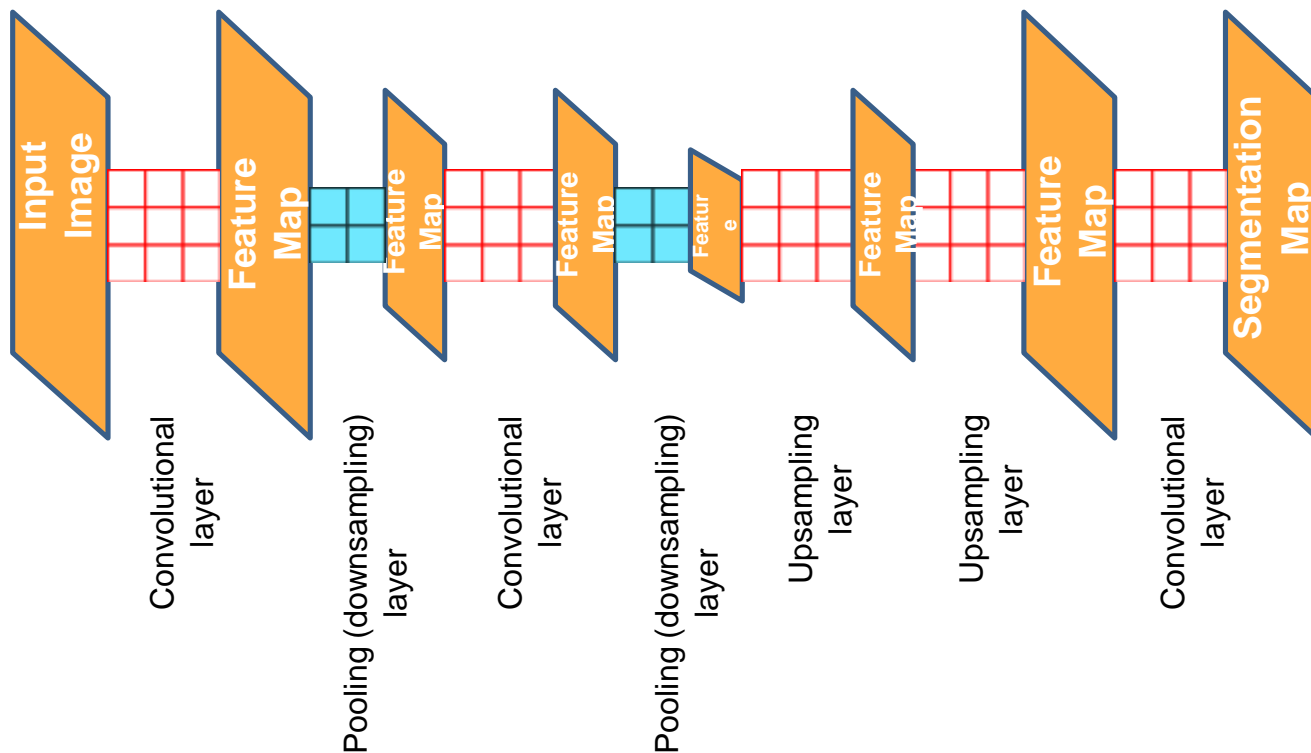
This file is meant for personal use by rg.ravigupta91@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Let us re-arrange the layers horizontally



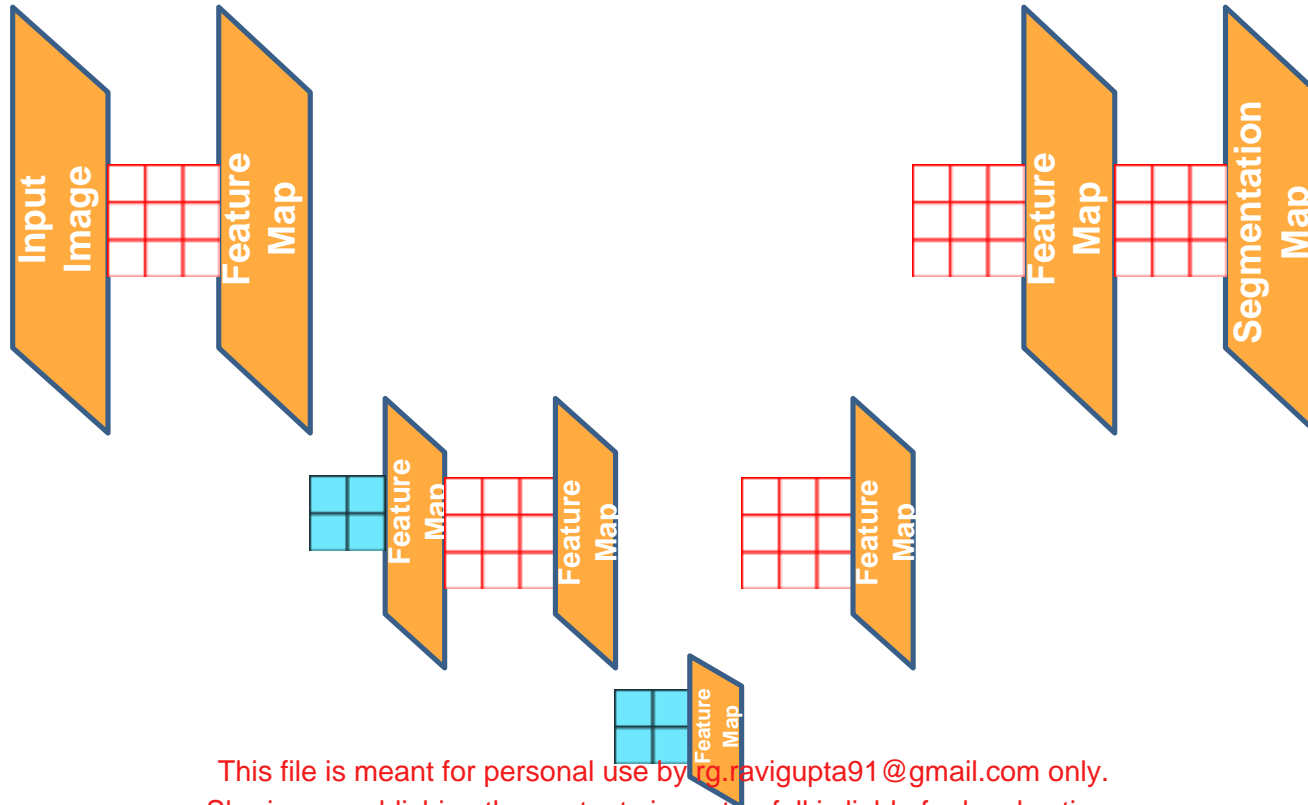
This file is meant for personal use by rg.ravigupta91@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

More layers can be added



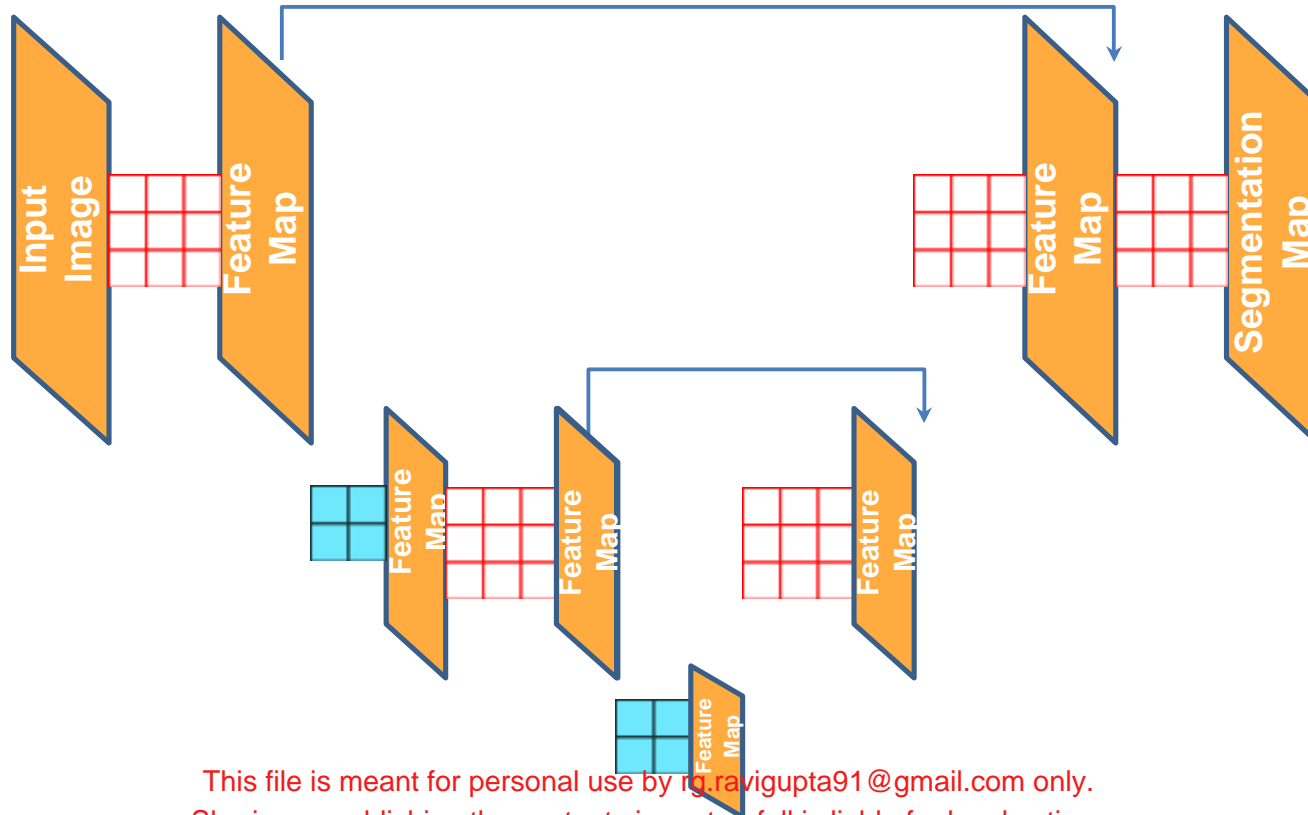
This file is meant for personal use by rg.ravigupta91@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Visually rearrange layers in a big U



This file is meant for personal use by rg.ravigupta91@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Concatenate previous feature maps for finer spatial context



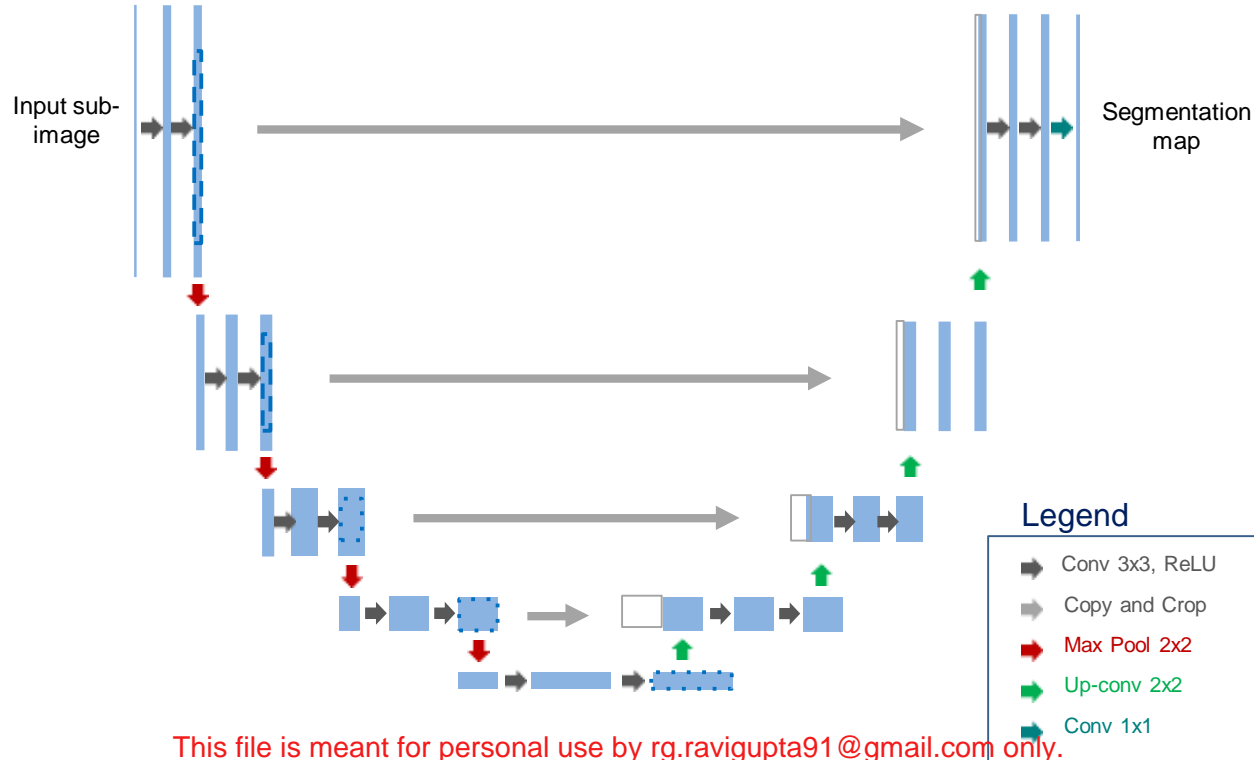
This file is meant for personal use by rg.ravigupta91@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

U-Net

This file is meant for personal use by rg.ravigupta91@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Proprietary Content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

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



U-Net

2

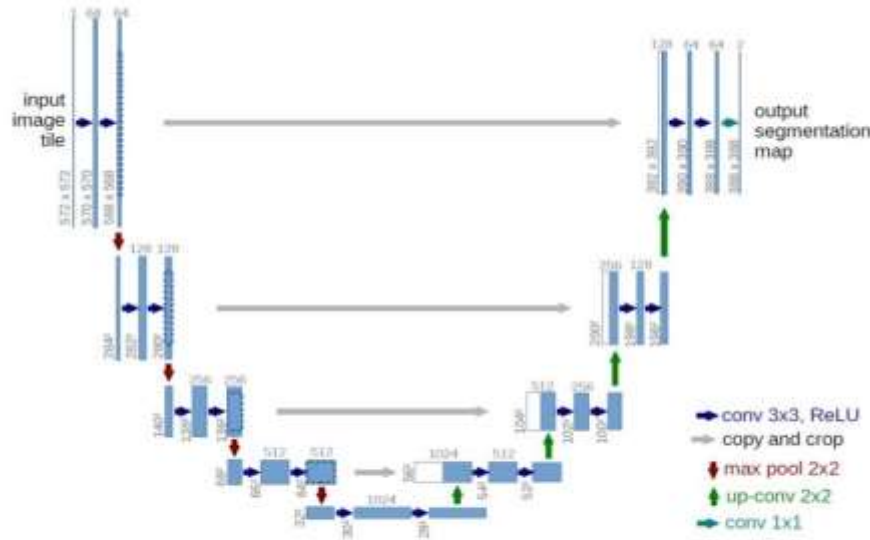


Fig. 1. U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows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

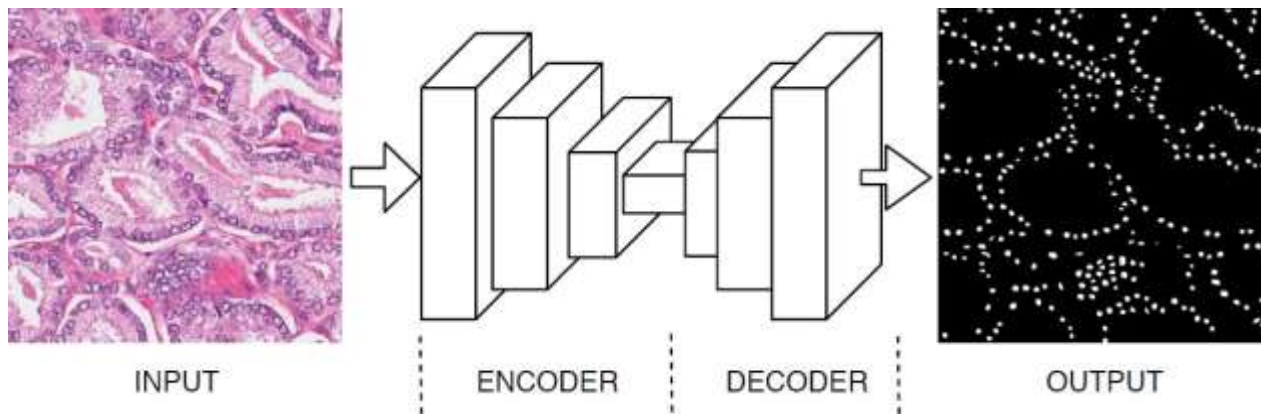
The architecture has -

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



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.

This file is meant for personal use by rg.ravigupta91@gmail.com only.

Sharing or publishing the contents in part or full is liable for legal action.

Dice coefficient

Dice coefficient is defined as follows:

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

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

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 :)