U-Net: Convolutional Networks for Biomedical Image Segmentation

(2015)

Olaf Ronneberger, Philipp Fischer, Thomas Brox
Resume

January 23, 2019

---

1 Introduction

The U-Net architecture is built upon the Fully Convolutional Network and modified in a way that it yields better segmentation in medical imaging. Compared to FCN-8, the two main differences are (1) U-net is symmetric and (2) the skip connections between the downsampling path and the upsampling path apply a concatenation operator instead of a sum and they reuse the activations from all the earlier layer in the encoder. These skip connections intend to provide local information to the global information while upsampling. Because of its symmetry, the network has a large number of feature maps in the upsampling path, which allows to transfer information. By comparison, the basic FCN architecture only had number of classes feature maps in its upsampling path.

2 The architecture

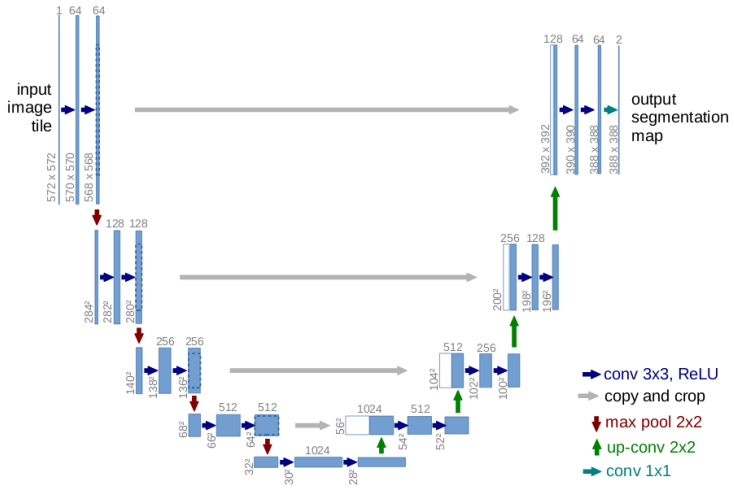

Fig. 1. U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows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

U-Net architecture is separated in 3 parts:

1. The contracting/downsampling path
2. Bottleneck
3. The expanding/upsampling path

Contracting/downsampling path The contracting path is composed of 4 blocks. Each block is composed of

- 3x3 Convolution Layer + activation function (with batch normalization)
- 3x3 Convolution Layer + activation function (with batch normalization)
- 2x2 Max Pooling

The number of feature maps doubles at each pooling, starting with 64 feature maps for the first block, 128 for the second, and so on. The purpose of this contracting path is to capture the context of the input image in order to be able to do segmentation. This coarse contextual information will then be transfered to the upsampling path by means of skip connections.

Bottleneck This part of the network is between the contracting and expanding paths. The bottleneck is built from simply 2 convolutional layers (with batch normalization), with dropout.

Expanding/upsampling path The expanding path is also composed of 4 blocks. Each of these blocks is composed of:

- Deconvolution layer with stride 2
- Concatenation with the corresponding cropped feature map from the contracting path
- 3x3 Convolution layer + activation function (with batch normalization)
- 3x3 Convolution layer + activation function (with batch normalization)

The purpose of this expanding path is to enable precise localization combined with contextual information from the contracting path.

Advantages

- The U-Net combines the location information from the downsampling path with the contextual information in the upsampling path to finally obtain a general information combining localisation and context, which is necessary to predict a good segmentation map.
- No dense layer, so images of different sizes can be used as input (since the only parameters to learn on convolution layers are the kernel, and the size of the kernel is independent from input image size).
- The use of massive data augmentation is important in domains like biomedical segmentation, since the number of annotated samples is usually limited.

3 Training

- They prefer to use large input tiles instead of large batches, so the input is one image.
- The loss function is pixel wise softmax with cross entropy over the final feature map. in the cross entropy loss they introduce a weight map to give some pixels more importance in the training.

$$p_k(\mathbf{x}) = \exp\left(a_k(\mathbf{x})\right) / \left(\sum_{k'=1}^{K} \exp\left(a_{k'}(\mathbf{x})\right)\right)$$

$$E = \sum_{\mathbf{x}\in\Omega} w(\mathbf{x}) \log\left(p_{\ell(\mathbf{x})}(\mathbf{x})\right)$$

- The wight maps contains two terms, w_c to balance the class frequencies, and the second term to force the network to learn small separation borders between cells, d_1 is the distance between the neighrest cell and d_2 is the distance to the second nearest cell.
- **data augmentation** they primarily use rotation and shift transformation, as well as deformation and gray value variations.