Summary of U-Net: Convolutional Networks for Biomedical Image Segmentation

Charul Daudkhane, Matriculation: 396089

1. Abstract

The original U-Net paper [1] introduces a new neural network architecture along with a training strategy for fast and precise image segmentation on Biomedical Images. This architecture is a modification and extension to the existing ”fully convolutional network” [2] and can be trained end-to-end on few images by making strong use of data augmentation. The contracting and expanding parts of the U-Net are responsible for context capturing and precise localization respectively.
Results indicate that U-Net model outperformed the prior best(sliding-window convolutional network) on the *ISBI challenge for segmentation of neuronal structures in electron microscopic stacks*[6] and won the *Cell Tracking Challenge at ISBI 2015*[7] for two transmitted light microscopy categories.

2. Introduction

The performance of any deep convolution neural network depends on the size of the network and the availability of the training data. The Imagenet classification breakthrough by Krizhevsky et al [3] was due to the availability of the ImageNet dataset (1 million training images) which was trained on a 8 layer large network with millions of parameters. In many biomedical image processing applications, the image classification task also includes pixel-wise localization and doing so on a small dataset is very challenging. Ciresan et al. [4] introduced a possible solution for localizing pixels with less available training data. The network used a sliding-window setup to predict the class label of each pixel and used the local region (patch) around that pixel as the training data. However, the resulting network was very slow as the entire network was trained separately for every patch and the network suffered a trade-off between the localization accuracy and the use of context. The U-Net paper [1] introduces an elegant architecture build upon ”fully convolutional network [2] which provides exceptional result for microscopic image segmentation with less available training data. The usual contracting network in [2] is supplemented by additional layers and the pooling operators are replaced by upsampling operators, thus increasing the resolution of the output. An important modification includes the use of a larger number of feature channels allowing the network to propagate context information to higher resolution layers.

For seamless segmentation of arbitrarily large images, the network uses an overlap-tile strategy as show in Figure 2. This strategy ensures that the image resolution is not limited to the GPU memory. The missing context at the border region is extrapolated by mirroring the input image.

In many cell segmentation tasks, it is important to learn the separation of touching objects belonging to the same class. The paper proposes the use of weighted loss, where a larger weight is applied in the loss function to the background labels between touching cells.

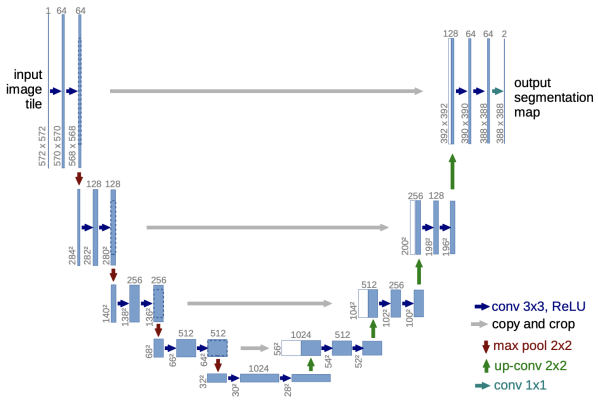

Figure 1. U-net architecture for 32x32 pixels in the lowest resolution where the blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box.White boxes represent copied feature maps and arrows denote the different operations.

3. Introduction

The U-Net architecture as described in Figure 1, consists of the following:

1. **Contracting Path (Left Side)**
   It comprises of 4 blocks each with the following layers:
   - Two 3x3 unpadded convolutions, followed by a

rectified linear unit (ReLU)
- 2x2 max pooling operation with stride 2 for down sampling

2. **Expansive Path (Right Side)**
   It comprises of 4 blocks each with the following layers:
   - Up-convolutional layer for upsampling the feature map via a 2x2 convolution
   - Copy and crop connection from the corresponding feature map from the contracting path
   - Two 3x3 convolutions, each followed by a ReLU activation function.

Every down-sampling step doubles the number of feature channels (from 64 to 512 for the 4th block). The contracting path is responsible for capturing the context of the input image and outputs a feature map for the expansive path.

The up-convolutions in the expansive path halve the number of feature channels(from 512 to 64) in order to perform concatenation with the block on the left. It enables precise localization along with the contextual information from the contracting path. The final layer outputs a class label for each 64 component feature vector via 1x1 convolution thus yielding a total of 23 convolutional layers. Figure 2 describes the segmented output from the U-Net architecture.

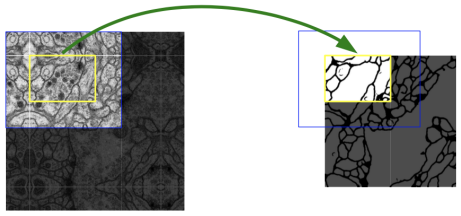

Figure 2. The overlap-tile strategy for seamless segmentation of arbitrary large images. Prediction of the segmentation in the yellow area, requires image data within the blue area as input and the missing pixel data is extrapolated via mirroring.

4. Network Architecture

The paper [1] describes an efficient training process for image segmentation. In order to maximize the GPU usage, large input tile size is preferred over large batch size and hence a batch size is kept to one.

The activation function used during the training is a pixel wise soft-max over the feature map and is denoted by the following equation:

where $a_k(x)$ denotes the activation in feature channel k at the pixel position x. The value for $p_k(x) \approx 1$or the class k

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

that has the maximum activation $a_k(x)$and $p_k(x) \approx 0$or all other k.

The loss function used for weight adjustment is cross entropy and is given by the following equation:

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

where l is the true label of each pixel and w is a weight map assigned to some pixels for giving more importance during training.
In order to make the network learn the separation borders between the touching cells (Figure 5c and d) and to compensate for the difference in frequency of pixels for certain class in the training data set, a weight map is computed for each ground truth segmentation.This weight map is computed by the following equation:

$$w(\mathbf{x}) = w_c(\mathbf{x}) + w_0 \cdot \exp\left(-\frac{(d_1(\mathbf{x}) + d_2(\mathbf{x}))^2}{2\sigma^2}\right)$$

where d1 and d2 are distances to the borders for first and second nearest cell and w_c is the weight map for class frequency balancing. For the weight initialization strategy, weights are drawn from a Gaussian distribution with a standard deviation of $\sqrt{2/N}8]$, where N denotes the incoming nodes for one neuron.

5. Data Augmentation

In Biomedical image processing applications, the amount of available training data is less so in order make the neural network robust and learn invariance properties, data augmentation is needed. The most importance invariance in microscopic images includes the shift and rotation, robustness to deformations and gray value variation. The paper [1] uses random elastic deformations of the training samples as the key to train the segmentation network. These deformations are generated using random displacement sampled from a Gaussian distribution with 10 pixels standard devia-

tion. The Drop-out layers at the end of the contracting path also contribute towards data augmentation. The importance of data augmentation for learning invariance is shown in Dosovitskiy et al. [5] in the scope of unsupervised feature learning.

6. Experiments

The paper summarizes the performance of U-Net architecture on three dataset for image segmentation challenge :

1. **Segmentation of neuronal structures in electron microscopic recordings**
   The EM segmentation challenge [6] at ISBI 2012 consisted of 30 images (512x512 pixels) from serial section transmission electron microscopy of the Drosophila first instar larva ventral nerve cord (VNC). The available training data consisted of fully annotated ground truth segmentation map for cells (white) and membranes (black). The performance of the network was evaluated on "warping error", the "Rand error" and the "pixel error" [6]. Figure 3 summarizes the results of different architecture on this challenge. The U-Net (averaged over 7 rotated versions of the input data) achieved a warping error of 0.0003529 and a rand-error of 0.0382.

Rank	Group name	Warping Error	Rand Error	Pixel Error
	** human values **	0.000005	0.0021	0.0010
1.	u-net	**0.000353**	0.0382	0.0611
2.	DIVE-SCI	0.000355	0.0305	0.0584
3.	IDSIA [1]	0.000420	0.0504	0.0613
4.	DIVE	0.000430	0.0545	**0.0582**
⋮				
10.	IDSIA-SCI	0.000653	**0.0189**	0.1027

Figure 3. Ranking on the EM segmentation challenge sorted by warping error.

2. **Cell segmentation task in light microscopic images:**
   The ISBI cell tracking segmentation challenge 2014 and 2015 [7] was conducted on two datasets :
   - "PhC-U3732 containing Glioblastoma-astrocytoma U373 cells on a polyacrylimide substrate recorded by phase contrast microscopy (see Figure 5a,b)
   - "DIC-HeLa3 containing HeLa cells on a flat glass recorded by differential interference contrast (DIC) microscopy (see Figure 6, Figure 5c,d)

PhC-U3732 consisted of 35 partially annotated training images for segmentation. U-Net performed significantly better with an average IOU (intersection over union) of 92 % Figure 45 summarizes the results of the ISBI cell tracking challenge

"DIC-HeLa3 consisted of 20 partially annotated training images and U-NET achieved an average IOU of 77.5 %

Name	PhC-U373	DIC-HeLa
IMCB-SG (2014)	0.2669	0.2935
KTH-SE (2014)	0.7953	0.4607
HOUS-US (2014)	0.5323	-
second-best 2015	0.83	0.46
u-net (2015)	**0.9203**	**0.7756**

Figure 4. Segmentation results (IOU) on the ISBI cell tracking challenge 2015.

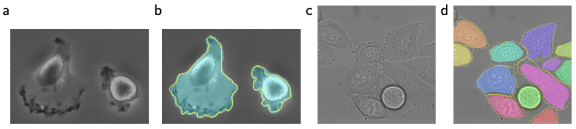

Figure 5. Result on the ISBI cell tracking challenge. (a) part of an input image of the PhC-U373 data set. (b) Segmentation result (cyan mask) with manual ground truth (yellow border) (c) input image of the DIC-HeLa data set. (d) Segmentation result (random colored masks) with manual ground truth (yellow border).

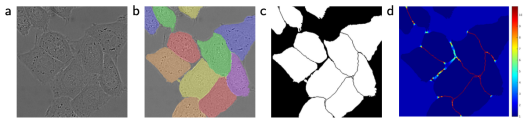

Figure 6. HeLa cells on glass recorded with DIC (differential interference contrast) microscopy. (a) raw image (b) overlay with ground truth segmentation. Different colors indicate different instances of the HeLa cells. (c) generated segmentation mask (d) map with a pixel-wise loss weight to force the network to learn the border pixels.

7. Conclusion

The U-Net architecture performed exceptionally well in different Biomedical image segmentation applications. These results could be attributed to the *U-shaped design of the network allowing to captures the localization and contextual*

information together and the extensive use of data augmentation with elastic deformation requiring the training time of only 10 hours on a NVidia Titan GPU (6 GB).

7.1. Citations and References

1. O Ronneberger, P Fischer, T Brox : U-Net: Convolutional Networks for Biomedical Image Segmentation(2015)

2. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation (2014), arXiv:1411.4038 [cs.CV]

3. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep con- volutional neural networks. In: NIPS. pp. 11061114 (2012)

4. Ciresan, D.C., Gambardella, L.M., Giusti, A., Schmidhuber, J.: Deep neural net- works segment neuronal membranes in electron microscopy images. In: NIPS. pp. 28522860 (2012)

5. Dosovitskiy, A., Springenberg, J.T., Riedmiller, M., Brox, T.: Discriminative un- supervised feature learning with convolutional neural networks. In: NIPS (2014)

6. WWW: Web page of the em segmentation challenge, http://brainiac2.mit.edu/ isbi_challenge

7. WWW: Web page of the cell tracking challenge, http://www.codesolorzano.com/celltrackingchallenge/Cell_Tracking_Challenge/Welcome.html

8. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human- level performance on imagenet classification (2015), arXiv:1502.01852 [cs.CV]