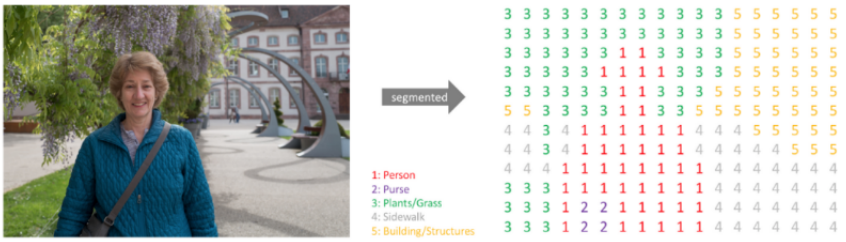
U-Net Review

Eunsoo Lim

Feb.24 2022

1 Introduction

U-net was published by Olaf Ronneberger in 2015 and targets classification on a biological domain. In the biological domain, It is not enough to classify with just a bounding box. In order to classify more accurately, you must employ a pixel-wise methodology called segmentation. Let me explain semantic segmentation briefly.



Semantic segmentation is one of the essential parts of computer vision and is dividing the pixels into meaningful units. Specifically, it is a predicting where each pixel is included in the class in. As a Representative model of semantic segmentation, There are FCN, DeepLab, U-Net so on. Below is an analysis of the U-net.

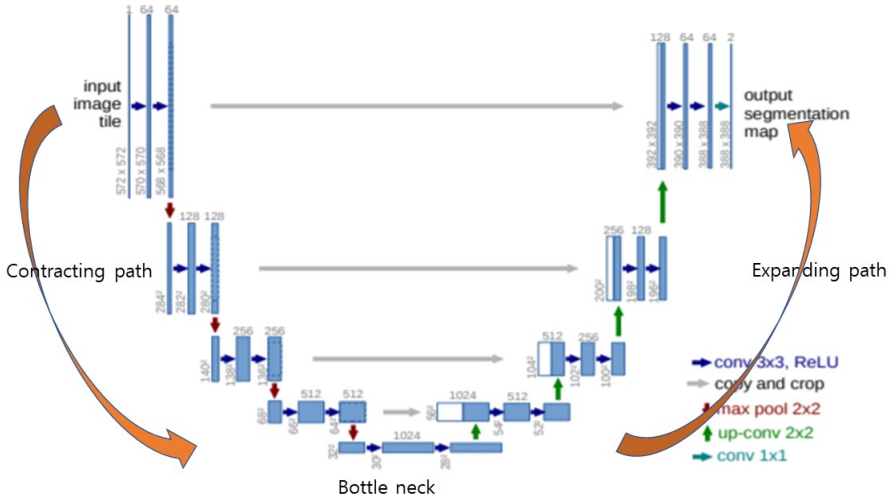
2 Analysis

From now on, The architecture, training, pros and cons of the U-net paper will be presented sequentially. The architecture session looks into the U-net model character in detail. Next, the training session will handle the loss function, optimization, and augmentation. Finally, the pros and cons of U-net will be shown.

2.1 Architecture

U-net model can be divided into 3 types largely. First of all, the contracting path is extracting context information. And, The expanding path is to distinguish each pixel-wise by fusing context and localization

information.



- **Contracting Path**

  Generate feature map repeating the down-sampling process. This role is extracting contextual information. The number of channels is supposed to double according to 3x3 Convolution.

  1. 3X3 Convolution Layer + ReLu + BatchNorm (No padding,Stride 1)
  2. 3X3 Convolution Layer + ReLu + BatchNorm (No padding,Stride 1)
  3. 2×2 Max-polling Layer (Stride 2)

- **Bottle Neck**

  It is the part of switch. The last Dropout Layer is the technical part for robust and generalization of the model

  1. 3X3 Convolution Layer + ReLu + BatchNorm (No padding,Stride 1)
  2. 3X3 Convolution Layer + ReLu + BatchNorm (No padding,Stride 1)
  3. Dropout Layer

- **Expanding Path**

  This part's role is concatenating contextual information from the contracting path and localization information. The classification vector of each pixel is created by the last layer which is a 1x1 Convolution Layer.

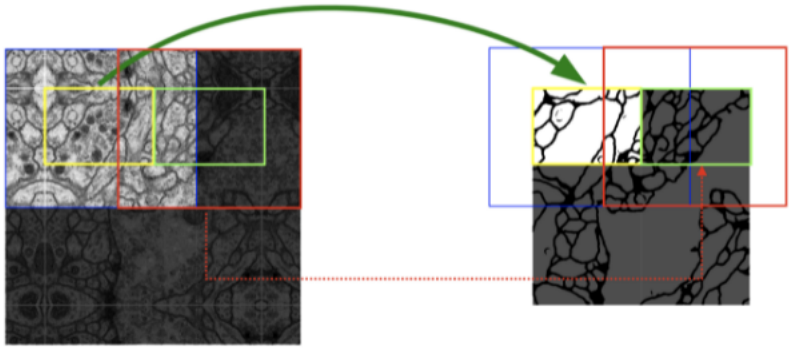
  1. 2×2 Deconvolution layer (Stride 2)
  2. Cropping and Concatenating with feature map

3. 3×3 Convolution Layer + ReLu + BatchNorm (No Padding, Stride 1)

4. 3×3 Convolution Layer + ReLu + BatchNorm (No Padding, Stride 1)

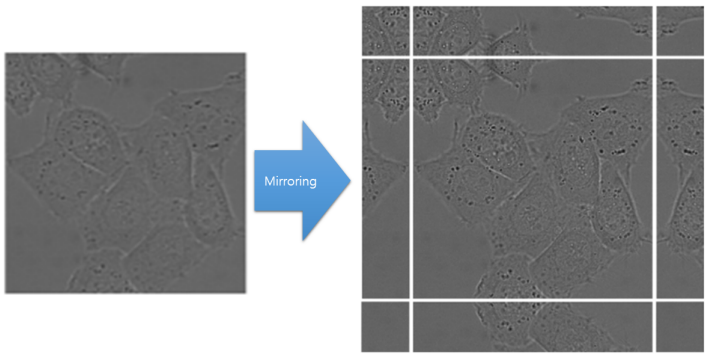
2.2 Training

- **Overlap-tile strategy**



As inferred from an upper image, the input image should be divided as patches, called tile as well. If you input the blue region, then the yellow region will be extracted. In the case of the red region, the result would be the green region. In other words, this strategy is cropping and segmentation, so that there are parts that overlap.

- **Mirroring Extrapolate**



When predicting the boundary line, it is normal to employ padding in convolution deep learning. U-net applies the mirroring images by making a copy of boundary pixels and reversing left and right. This is because, In my opinion, U-net was generated to target the biomedical domain in which cells usually composed of symmetry up, down, left and right are main.

- **Weight Loss**

  When the model is learning, It is most important to separate bounding lines. Therefore, U-net is applied Weight-Map. Loss value increases according to Weight-Map how each pixel is close to the bounding line.

  In order to take advance of memory maximally, U-net prefers big-size images and uses the momentum method.

$$Loss = \sum_x w(x) log(p_{l(x)}(x))$$
$$p_k(x) = exp(a_k(x))/(\sum_i^K exp(a_i(x)))$$
$$w(x) = w_c(x) + w_0 \cdot exp(-\frac{(d_1(x)+d_2(x))^2}{2\sigma^2})$$

  In the upper loss formula,*w(x)* takes a huge value when it is close to the boundary line. Then, the weight of its pixel is getting greater at closing to the boundary and the model is able to learn boundary pixels.

- **Data Augmentation** Since there are small data usually, the model could be sturdy by using data augmentation such as rotation, shift, elastic distortion.

2.3 summary

Positive points	Negative points	Suggestion
Skip architecture which fuse context and local information, improving accuracy	The optimal depth of the model for the data-set is unknown	I think the number of convolutions is small. It could reduce the capability of extracting information.I suggest adaptive increasing the number of depth and channel according to number of input data-set account.
Even small data is not a problem due to augmentation	A limited structure in which only encoder and decoder with the same depth are skip connected. Then, the connection power between encoder and decoder can be weak	In order to prevent memory's lack problem, I think using residual learning(like ResNet) can be substitution
Using patches makes the model faster	The number of filters at the encoder is smaller than the number of filters at decoder, so that makes Asymmetric structure	For improving accuracy of segmentation, concatenate middle feature maps and use batch normalization.

Table 1: U-net summary

3 Conclusion

According to the result of the experiment, U-net is the best ranking in the EM segmentation challenge(march 6th,2015). I guess skip architecture is an important part of making the loss of information reduced. Anyway, it is clear that U-net is an effective and practical model. I researched other segmentation models and summarize their character.

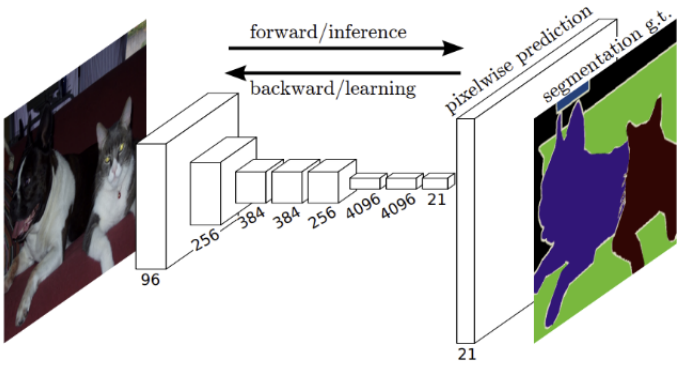
cf. How to measure pixel accuracy

If you consider segmentation as image classification, you can think the criteria for evaluation of classification for each pixel. At this time, when evaluating prediction map by class, consider it as binary classification and evaluate it by pixel and channel. When evaluating as an image classification problem for each pixel, distinguish whether the correct answer class is correct for each pixel, that is, True/-False. For example, assuming that the 2x2 area in the center is foreground and one of the prediction results is missed in a map with a 4x4 size, the Accuracy can be obtained by checking the Error Metrics (TP+TN)/(FP+FN+TP+TN).True positive (TP)+ True negative (TN) is 15 excluding one incorrectly predicted column as the number of samples classified correctly. And one compartment of False case 1 is False Negative (FN) because the foreground is predicted as the background. Therefore, the denominator term is 16. Therefore, Pixel Accuracy can be calculated as 15/16.

3.1 Further Research

I researched additional segmentation model and introduce them briefly.

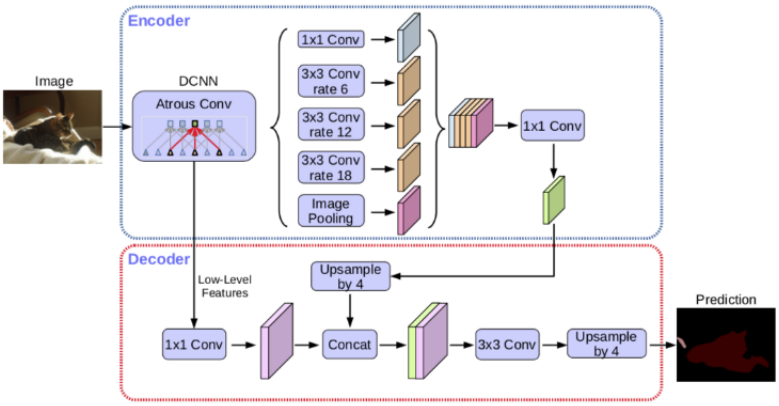
- **FCN**



Models for initial segmentation can be obtained by adding Bilinear interpolation tasks. Although the dense map could be derived from the coarse map through interpolation, the information on the

predicted dense map is still rough because the size of the feature map is fundamentally too small. Skip architecture (defective local information of the shallow layer and semantic information of the deep layer in the deep neural network) allowed us to achieve more sophisticated Segmentation results. This study laid the foundation for Dense Prediction or Semantic Segmentation through the End-to-End method of the Full-Convolution Model, which has influenced many related studies since then.

- **DeepLab**



The picture above is DeepLabV3+. U-Net seemed intuitive in structure, but DeepLabV3+ seems a little more complicated than that. It is the encoder and decoder in the figure above that act as Contracting path and Expansive path in U-Net. feature map came through an encoder is 16 times smaller than the resolution of the original resolution. Deeplab used atrous convolution. The ASPP(Atrous Spatial Pyramid Pooling) is composed on application of the Atrous convolution widely on the size. Deeplab extracts featrue from ASSP block and acquires segmentation mask throught upsampling.