Blood Vessel Segmentation from Fundus Photographs

Meeran Ismail

Stanford University, Stanford CA 94305, USA

1 Introduction

1.1 Problem Motivation

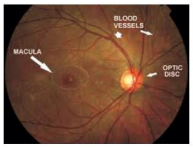

Fig. 1. Labeled Fundus Photo

Most cases of blindness in American adults occur due to late-stage diabetic retinopathy [5]. Analyzing blood vessels in retinal images is key for early diagnosis of diabetic retinopathy, but is often difficult to discern in blurry retinal images. Automated segmentation would vessels stand out, and could thus aid less experienced physicians in diagnosing the disease.

1.2 Problem Statement and Dataset

The goal of this project will be to build a deep neural network to compute an image mask of blood vessels in fundus photos. The dataset used for this is the DRIVE dataset [1], a commonly referenced dataset for segmentation of blood vessels from fundus photography. The dataset consists of 20 training examples and 20 testing examples, each of which consists of a fundus photograph and an accompanying image mask labeling where the blood vessels are (see Figure 2):

Data Augmentation Because we only had 20 training examples, and Ronneberger et al. recommends "extensive use of data augmentation," [2], we decided to augment the pictures in two ways: first, given that the optic disk appears on either the left or right of the image with equal frequency, we horizontally flipped each picture. Second given the natural rotation of the eye, we also slightly rotated each original picture by a random amount that wouldn't exceed 0.2 degrees. The result was 40 additional training images from the augmentation, resulting in 60 total training images, a scale that has had empirical success on other datasets, such as segmenting neuronal structures in electron microscopic stacks [2].

Fig. 2. Retinal photo (left), image mask with segmented blood vessels (right)

2 Approach

Previous models on this dataset include a variety of supervised approaches, such as a hierachal Markovian approach, and unsupervised approaches, such as support vector machines and random forest classifiers using a hybrid of features from a variety of different filters [3]. Savu et. al. uses a sliding window architecture that predicts each pixel based on the image patch that pixel centers [4]. Ronneberger et. al., however, proposes the U-Net architecture [2], which can train on very few examples and outperforms the sliding window approach on other biological dataset; we thus propose that model to be used on fundus photography.

2.1 U-Net Architecture

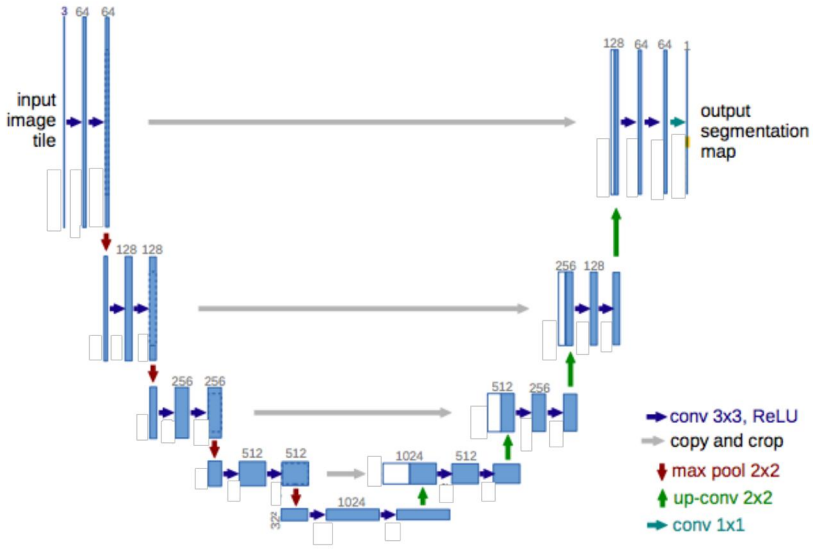

Fig. 3. U-Net Architecture [2]; the number on top of each layer represents the number of feature channels in that layer.

First, two 3x3 convolutions are applied on the 512x512 input image, increasing the number of feature channels to 64. Then, a contracting path of 4 downsampling steps is applied, each downsampling step consisting of two 3x3 convolutions and a 2x2 max pooling operation. Two additional 3x3 convolutions are applied after this contracting path; then, the resultant layer is put through an expansive path (symmetrical to the contracting path) of 4 upsampling steps, each consisting of a 2x2 "up-convolution" [2](4) concatenated with the result of the downsampling step of the same depth (see Fig. 2 above), followed by 2 3x3 convolutions. 2 3x3 convolutions are then applied on the previous layer before a 1x1 convolutional filter with a sigmoid output is applied, resulting in an image mask, where each pixel is a sigmoid output over the final image mask, labelling each pixel with the probability that it belongs to a blood vessel. With the exception of the final 1x1 filter with a sigmoid activation, all convolutions and up-convolutions use SAME padding and are followed by RELU activations. Additionally, dropout is applied at the end of the 3rd and 4th downsampling steps, with $keep\_prob = 0.5$, and both the max-pooling operations and the up-convolutions use strides of length 2 in both the x and y directions. [2]

The contracting path is designed to capture context in the image [2]. Each downsampling step doubles the number of feature channels via the two convolutions, while halving both the x and y dimensions of that layer through max-pooling.

The expansive path, which is designed for "precise localization" [2], does the opposite; in each upsampling step, the "up-convolution" (covered in the next section) doubles the x and y dimensions of the layer it is applied to while halving the number of feature channels, and applying 2 3x3 convolutions in the upsampling step also halves the number of feature channels.

A problem that can arise is the loss of context information; because each downsampling step doubles the number of feature channels, yet each upsampling step reduces the number of feature channels by a factor of 4, a significant amount of the context information learned through the contracting path that would be useful in helping with localization is lost at each step. To solve for this, during each upsampling step, after the up-convolution is applied, the layer from the downsampling step with the same number of feature channels is concatenated before the 2 3x3 convolutions are applied, thus providing context learned during the contracting path, and ensuring that each upsampling step only halves the number of feature channels. For example, during the first upsampling step, the up-convolution reduces the number of channels from 1024 to 512; then, the downsampling step with 512 channels is concatenated to the result of the up-convolution, to create a 1024-channel layer that is reduced to 512 channels again after the 2 3x3 convolutions are applied.

"Up-Convolution" Up-convolutions are a critical part of the expansive path of the U-Net, as they are required to expand the layer and upsample. Given a $n \times n$ input layer, and a $k \times k$ filter, an up-convolution is performed by taking each element of the input layer, scaling the filter by that element, and setting the $k \times k$

submatrix of the output that corresponds to the position of that input element equal to the scaled filter. To determine the position of the output layer that corresponds to a given element, start by choosing the top-left $k \times k$ submatrix to correspond to the top-left element of the input layer. Then, when moving over to the next input element in either direction, shift the $k \times k$ submatrix of the output by *stride_length* positions in the same direction, and apply the same operations. If the up-convolution result on one element overlaps on that of another element in the output, sum the results in the output cells that overlap. Figure 4 is a 1-D example that demonstrates this more clearly. As is the case with normal convolution, the number of up-convolution filters directly determines the number of channels in the output layer of the up-convolution.

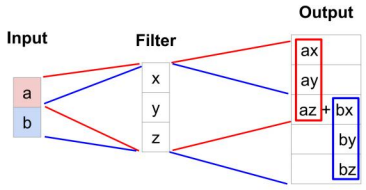

Fig. 4. 1-D up-convolution

2.2 Loss Function and Evaluation

The loss function used is the summed binary cross-entropy loss over each pixel, i.e., if we let Y and $\hat{Y}$ denote the ground-truth and generated image masks respectively, then

$$J = -\sum_{i=1}^{512}\sum_{j=1}^{512} Y_{ijk} \log \hat{Y}_{ijk} + (1 - Y_{ijk})\, log\left(1 - \hat{Y}_{ijk}\right)$$

The model will be evaluated on mean pixel accuracy, the most commonly referenced metric in the Aguirre-Ramos et. al. review of previous approaches on segmenting blood vessels [3].

3 Results, Discussion, and Further Approaches

3.1 Results

After training this U-Net for 5 epochs at 2000 steps per epoch on a K80 GPU, this model outperforms the approaches summarized by Aguirre-Ramos et. al. [3] on accuracy on the test set, as seen in Figure 5:

Visualizations of some of the segmentations predicted by the U-Net can be seen in Figure 6.

Model	*Test Accuracy*	*Training Loss*	*Test Loss*
U-Net	**0.9640**	**0.0485**	**0.1438**
Sliding Window ConvNet [4]	0.9013	-	-
SVMs [3]	0.9510	-	-
Hierarchal Markovian (Unsupervised) [3]	0.9439	-	-
Random Forest on Gabor Filter Features [3]	0.9464	-	-

Fig. 5. Results of U-Net, as well as comparisons to other approaches.

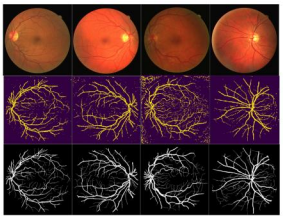

Fig. 6. Top Row: fundus photos from test set, Middle Row: model-generated image masks, Bottom Row: ground truth image masks (each column is one test example)

3.2 Discussion

While this model outperforms previous approaches, there is still room for improvement. First, as can be seen in the results table, training loss is signicantly lower than test loss; this is a sign that overfitting might be happening, and so applying increased dropout or more data augmentation might help. Second, a quick glance at Figure 6 shows that the bulk of the error comes from false positives, especially when blood vessels are very faint in a field-of-view that isn't very bright to begin with. This can be mitigated by adjusting the loss function, and placing increased penalty on false positives. Additionally, many of the false positives occur outside of the field of view of the fundus photograph. Thus, given that field-of-view masks is defined automatically from the photo, if include the field-of-view mask as an additional input layer, that is only applied at the very end of the U-Net to "drop-out" all the pixels that aren't in the field of view, many false positives can be fixed instantaneously. Finally, evaluate different metrics that have become more prevalent in image segmentation, such as the DICE coefficient, and possibly retrain the U-Net to minimize the DICE coefficient instead of cross-entropy.

3.3 Code Repo

`https://github.com/meeranmismail/unet_fundus_segmentation`

This repo was originally forked from `https://github.com/zhixuhao/unet`.

References

1. J.J. Stall, M.D. Abramoff, M. Niemeijer, M.A. Viergever, B. van Ginneken, Ridge based vessel segmentation in color images of the retina, IEEE Transactions on Medical Imaging, 2004, vol. 23, pp. 501-209
2. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation `https://arxiv.org/pdf/1505.04597`, 18 May 2015.
3. Hugo Aguirre-Ramos, Juan Gabriel Avina-Cervantes, Ivan Cruz-Aceves, Jos Ruiz-Pinales, Sergio Ledesma, Blood vessel segmentation in retinal fundus images using Gabor filters, fractional derivatives, and Expectation Maximization, Applied Mathematics and Computation, 2018, vol. 339, pp. 568-587.
4. M. Savu, D. Popescu and L. Ichim, "Blood vessel segmentation in eye fundus images", 2017 International Conference on Smart Systems and Technologies (SST), Osijek, 2017, pp. 245-249.
5. Klein R, Klein B. Vision disorders in diabetes. In: National Diabetes Data Group, ed. Diabetes in America. 2nd ed. Bethesda, MD: National Institutes of Health, National Institute