HistNet: Histopathology Segmentation using Context Aggregation

Abstract

Digital histopathology has seen a significant improvement towards aiding pathologists diagnosis by the use of deep learning segmentation models. The segmentation task allows for segregating benign and malignant cells from digitized tissue whole slide images (WSI). However, for reliable deployment, it is essential to produce accurate object boundaries. Pathologist performs this challenging task using context to extract complex features from a WSI. We leverage this idea and propose a model that aggregates context through use of parallel dilated and standard convolutions, embedded within each encoder-decoder block of the segmentation network. We show that, the proposed architecture not only beats the baseline by a large margin but also achieves an accuracy improvement over state-of-the-art segmentation architecture by upto 1.95% DICE Score.

1. Introduction

Colon cancer is a leading cause of cancer related deaths worldwide, in US alone over 52,000 deaths were reported in 2016 [1]. The most reliable method for diagnosis is analysis of whole colorectal slides of patient tissue samples using a microscope by a pathologist. However, a pathologist needs years of experience to gain expertise in whole slides image (WSI) analysis. The manual process is time consuming, exhausting and requires a quantitative assessment of pathology slides. Furthermore, the diagnosis is based on expert opinion which can be subjective [6].

With the advent of digital WSI scanners [12], the diagnostic fields has seen a major paradigm shift. This has led to the development of Digital Pathology that uses machine learning techniques to assist pathologists in their diagnosis. Specifically, WSI pixels are segmented into malignant and benign regions, which can be used to diagnose the tissue sample. Segmentation of WSI images allows not only quantitative assessment but also helps in understanding the underlying morphological features. Deep learning models are especially suited for these tasks as they are able to learn expert knowledge after being trained on manually annotated WSIs. They have been used for different digital histopathology tasks including breast cancer [16] [10] [14], colorectal cancer [3] [7] and prostate cancer [8].

Segmentation of slide images is challenging due to several reasons. There are large variations in glandular structures, which make identification of relevant region difficult. Also, slide images have large variations due to the process of staining and digitization. Presence of artifacts on whole slide images further increase the complexity of the task. Pathologists alleviate these issue by observing contextual information in histopathological samples combined with localized morphological information.

Figure 1 shows sample tiles from a WSI along with binary masks for malignant (foreground) and benign (background) tissues. Tumors are generally heterogeneous consisting of a mixture of tumor nests, lymphoid aggregates, stroma, and other normal cell structures. To obtain a diagnosis, pathologist use tissue morphology, cell size, cell distribution and colour. This is done in reference to features of the neighborhood tissue cells, to extract contextual information together with cell morphological features.

In this paper, we aim to improve segmentation accuracy on colorectal WSI by mimicking pathologist. We achieve this by combining local and global information using context aggregation. Our proposed model uses an encoder-decoder architecture similar to U-Net [13]. The encoder network incorporates inception like module that combines information from different receptive fields. This module uses atrous (dilated) [20] convolution to improve models ability to capture context. The basic block of the module consists of standard convolution in parallel with convolutional operations. Furthermore, the scale diversity is increased by concatenating output of parallel branches with different number of stacked basic blocks. Such a structure allows aggregating context and local information at multiple levels. We show that by using these blocks, the network is able to significantly improves segmentation performance.

Secondly, to improve the model’s ability to finely delineate morphological features on WSI, we add modified ResNet block after each upsampling. Similar to encoder blocks, the modified ResNet blocks consists of two parallel convolutions, of which one is dilated convolution to incorporated context. This allows the network to access information from different fields of view (FoV) even during upsampling stage. We show qualitatively as well as quantitatively
To summarize, the contributions of the paper are as follows:

1. We propose to use dilated convolution in parallel to standard convolution in encoder as well as decoder stages of the segmentation network to allow combine information from different FoV.

2. Our model uses a new inception like module in encoder to aggregate context at multiple levels.

3. We adapt U-Net decoder to better exploit local as well a global information for recovering accurate object boundaries, using a modified ResNet module.

4. We show through experimental evaluation that histopathology segmentation benefits from aggregating context, embedded within the basic blocks of encoder-decoder structure.

2. Related Work

2.1. Semantic Segmentation

FCN [11] introduced an end-to-end trainable convolutional network for dense prediction tasks. It used fractionally strided convolution to upsample feature maps from an altered classification network. SegNet [2] as well as U-Net [13] have a similar encoder-decoder network, however with much larger decoder than FCN. This allows gradual upsampling of the feature maps by considering local as well global features. To fuse lower level spatial information with abstract features, U-net decoder concatenates corresponding feature maps from encoder though a skip connection. These architectures provide evidence that decoder architecture plays a significant role in performance on the segmentation task, which we further exploit in our work. Based on these observations, we focus on improving the decoder architecture to allow efficient recovery of spatial information during upsampling.

PSPNet [22] and DeepLabV3+ [4] both exploit contextual relationships and global information. PSPNet uses a pyramid pooling module that extracts information from different scales, while DeepLabV3 uses multiple parallel di-
lated (atrous) convolutions with different rates to capture multi-scale information in a Atrous Spatial Pyramid Pooling block. Similar to DeepLabV3 our model uses dilated convolution to incorporate information from large FOV. However, in our model, the proposed parallel convolutions are embedded within each encoder and decoder blocks. This allows to aggregate context in each stage of downsampling as well as upsampling.

Recent works [19, 23, 21] have proposed segmentation models for real-time segmentation, however at the cost of accuracy. This is not desirable for digital histopathology, as accuracy is essential for reliable deployment.

2.2. Histopathology Segmentation

WSIs typically have an exceptionally high resolutions, patch sampling is widely used to deal with this. [17] uses this strategy to train a patch level classification model on randomly sampled patches. During inference patch-level predictions are used to create a tumor probability heatmap. [10] samples sub-images from regions of interest on the WSI. To speed up computation, these tiles are then processed through a FCN model without a decoder. The model generates probability maps in much smaller sizes as its output, which are then reconstructed and stitched to produce a dense whole-slide probability map. [10] proposed an end-to-end trainable multi-stage model, to embed relationships between neighborhood patches in the dense prediction. [16] uses a segmentation model on feature map generated by assembling feature vectors from a patch classifier. Although these approaches make processing significantly faster, they lead to substantial loss of resolution in the output, which maybe detrimental to understand the underlying reasons for a specific diagnosis.

Some works [3, 18, 5] propose a multi-task network for segmentation instance of glandular structures, at patch-level. Multi-task architecture allows the model to learn better discriminative features. [18] uses three parallel network architectures for foreground segmentation, object detection and edge detection which makes this approach computation extremely slow. Similar to our work, model proposed by [5] extensively uses dilated convolutions. However, the decoder consists of stacked convolutions unlike our network architecture which uses dilated ResNet modules.

3. Method

The proposed HistNet consists of encoder-decoder structure similar to U-Net. The network spatially downsamples the feature map encoding high level features and retrieves spatial boundary information by stage-wise upsampling the feature map. Also, the model uses skip connection between encoder and decoder to provide decoder with local feature information for sharper boundary generation.

Novel to our proposed network, we use modified dilated inception modules in the encoder. These modules enable the encoder to combine information from different FOV as the feature maps are downsampled, aggregating context to extract complex features to identify tissue cell morphology. Furthermore, the decoder up-sampling stage uses modified ResNet blocks, these blocks are also designed to access information at different FOV. This multi-scale information adds to the ability of the decoder to produce sharper object boundaries. The overview of the network architecture can be seen in Figure 2.

3.1. Dilated Inception Module

Inspired by Inception block which combines information from different scale, we use a modified version which uses an additional dilated convolution in parallel to existing convolutions. The dilation rate of this convolution adds additional ability to extract multi-scale information. Figure 3 compares inception module with our proposed inception module. Standard inception modules has three convolution layers and one concatenation layer, whereas, the modified modules consist of three standard and two dilated convolution layers. To aggregate information, the output of the two parallel convolutions are summed element-wise.

HistNet uses three inception modules as show in Figure 2. Each of the encoder modules having different dilation rate. Empirically, we found that having lower dilation rate downstream in the encoder is beneficial with feature map resolution is beneficial. This maybe because smaller feature map size (deep in the encoder) have already consolidated information by extracted complex features. Thus, the three dilated inception modules uses dilation rates of 4, 3, and 2 from top to bottom.

3.2. Dilated ResNet Module

The proposed network uses modified dilated ResNet modules at each decoder stage. Similar to dilated inception block, these modules also use dilated convolution in parallel to standard convolution. Figure 4 shows the structure of ResNet blocks with and without dilated convolutions. Standard ResNet unit consists of two convolution layer whereas in the modified block consists of two standard and two dilated convolutions. The output of the parallel convolutions is combined using element-wise addition.

To allow the network to conserve relevant information extracted from different FOV, instead of summing the skip connection, we concatenate the branches. The concatenation is followed by a compression layer. The compression layer reduces the number of feature channels using 1x1 convolutions. We use a constant dilation rate of 2 for all the dilated convolutions in the decoder, we found that increasing
Figure 2: HistNet detailed structure and annotations

Figure 3: Inception modules; (a) represents the inception module using standard convolution, and (b) represents the modified inception module using standard and dilated convolution

4. Experiments

4.1. Setting

4.1.1 Dataset

We perform experiments using two colorectal histopathology datasets which consists of WSI, namely Colonoscopy Tissue Segment dataset from DigestPath 2019 [9] and GlaS@MICCAI2015 [15]. Both the datasets are designed for binary segmentation task to identify benign and malignant tissue.

Colonoscopy Tissue Segment - DigestPath2019 [9] DigestPath 2019 dataset consists of 250 images samples of tissue from 93 WSI containing lesion morphology and another 410 images of tissue from 231 WSI which are malignant free. The challenge only releases the training set and keeps testing set secret. Thus we divided the training dataset into three subset for training, validation and testing which are mutually exclusive. Each of these image has a dimension of 5000x5000 pixels. To deal with the large image size, we create 512x512 tiles from the tissue images. In total, we
Figure 4: ResNet modules with 1x1 compression; (a) represents the ResNet modules using standard convolution, (b) represents the ResNet modules using standard and dilated convolution.

extract 12,922 tiles, of which we use 70% for training, 15% for validation and testing.

GlaS@MICCAI2015 [15] GlaS@MICCAI2015 dataset contains 85 training images and 80 test images. The test images are split between Test A and Test B, with 60 and 20 images in each split. The images are mostly of size 775x522 pixels. We subdivided the training into two subsets namely training and validation which are mutually exclusive. In training we have used 60 images and rest of the images (25) used for validation. We resized the RGB and labeled image into $512 \times 512$. To prevent over-fitting, we augment training data using random flipping, affine transformation, Gaussian blur, color distortion and gamma transform.

4.1.2 Implementation Details

All the models were implemented in Keras framework. In the preprocessing step, the RGB images were subtracted by mean computed from the training data set only. We reduce the feature map dimension by a factor of two at each stage in the encoder and increase it by the same factor in the decoder. Number of feature map channels obtain at the end of each encoder-decoder stage are 16,16,32,64,128,256, where 256 corresponds to lowest resolution feature map.

We use Adam optimizer and step learning rate with an initial learning rate of $10^{-3}$; decreasing the learning rate by a factor 0.1 for every 150 epochs. All models were trained until convergence was achieved with respect to training and validation loss. We used NVIDIA Titan X for training the model.

4.2. Evaluation

To evaluate the effectiveness of our approach we compare HistNet with DeepLabV3+ which aims at aggregates multi-scale context information as well as baseline U-Net model. Table 1 and 2 provide the performance results on the two colorectal histopathology datasets.

HistNet model is a able to achieve superior performance compared to both U-Net as well as DeepLabV3+ on both the colorectal histopathology dataset. For DigestPath2019 dataset, the model achieves increase in DICE score of 6.65% as compared to U-Net and 1.95% in comparison to DeepLabV3+, where as evaluating on GlaS@MICCAI2015 dataset, HistNet improves DICE score by 5.58% and 0.8% with respect to U-Net and DeepLabV3+ respectively. Based on this evaluation we can infer that combining multi-scale information within encoder decoder blocks works favorably for histopathology image segmentation.

Table 3 reports number of parameters in HistNet model in comparison to U-Net and DeepLabV3+. Inspite of achieving higher DICE score, HistNet uses significantly less number of parameters as compared to other two benchmark models.

We also evaluate the models qualitatively. Figure 1 shows the segmentation mask produced by HistNet and the two benchmark models. U-Net performs poorly, possibly since it does not use context aggregation unlike the other two models. Importantly, segmentation boundaries produced by HistNet are finer and much closer to the ground truth as compared to DeepLabV3+ which also combines multi-scale context information. This further reinforces the ability of our model to extract relevant morphological information using context aggregation.

4.3. Ablation Study

In this section, we show results of our ablation study on DigestPath2019 colonoscopy tissue segmentation dataset. We investigate the contributions of HistNet model design to the performance improvement. Table 4 gives the results of our analysis.

We first analyze performance improvement obtain by adding HistNet encoder and decoder to U-Net architecture. Replacing U-Net encoder with our proposed HistNet encoder we obtain an accuracy (Dice score) gain of 2.27%, furthermore using HistNet encoder as well as decoder together adds another 4.38% gain.

We also experiment by removing the dilated convolution from both the encoder and decoder network (HistNet wo/dilation), that is using standard inception blocks in encoder and ResNet blocks in decoder, as shown in Figure 3 and 4. This allows us to evaluate the improvement in the

HistNet uses 31 standard convolutions and 17 dilated convolutions, including encoder and decoder networks. Thus to maintain similarity we train DeepLabV3+ with ResNet-50 backbone.
<table>
<thead>
<tr>
<th>Model</th>
<th>Mean DICE %</th>
<th>Mean IOU %</th>
</tr>
</thead>
<tbody>
<tr>
<td>U-Net</td>
<td>85.43 ± 4.78</td>
<td>76.6 ± 5.23</td>
</tr>
<tr>
<td>HistNet</td>
<td>92.08 ± 3.65</td>
<td>86.27 ± 4.71</td>
</tr>
<tr>
<td>DeepLabV3+</td>
<td>90.13 ± 7.59</td>
<td>84.48 ± 7.33</td>
</tr>
</tbody>
</table>

Table 1: Results on Colonoscopy tissue segment dataset - DigestPath2019

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean DICE %</th>
<th>Mean IOU %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test-A</td>
<td>85.47±1.01</td>
<td>75.31±1.13</td>
</tr>
<tr>
<td>Test-B</td>
<td>81.14±2.32</td>
<td>69.64±1.65</td>
</tr>
<tr>
<td>HistNet</td>
<td>90.58±0.73</td>
<td>83.25±1.17</td>
</tr>
<tr>
<td>Test-A</td>
<td>85.81±1.96</td>
<td>76.64±2.39</td>
</tr>
<tr>
<td>Test-B</td>
<td>82.06±0.05</td>
<td>75.92±2.38</td>
</tr>
<tr>
<td>DeepLabV3+</td>
<td>89.78±0.03</td>
<td>82.06±0.05</td>
</tr>
</tbody>
</table>

Table 2: Results on GIas@MICCAI2015 dataset

<table>
<thead>
<tr>
<th>Model</th>
<th>No. of parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>U-Net</td>
<td>7.7M</td>
</tr>
<tr>
<td>HistNet</td>
<td>3.4M</td>
</tr>
<tr>
<td>DeepLabV3+</td>
<td>59.3M</td>
</tr>
</tbody>
</table>

Table 3: Comparing number of parameters

<table>
<thead>
<tr>
<th>Modified Encoder</th>
<th>Modified Decoder</th>
<th>Dilated 3x3 Conv</th>
<th>Mean DICE %</th>
<th>Mean IOU %</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓ *</td>
<td>85.43</td>
<td>76.60</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>91.38</td>
<td>85.26</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>92.08</td>
<td>86.27</td>
</tr>
</tbody>
</table>

Table 4: Ablation Study. * Dilated 3x3 conv is used only in the modified encoder

model due to parallel standard and dilated convolutions. As shown in Table 4, HistNet obtain an improvement of 0.7% in DICE score and 1.01% in IOU score.

5. Conclusion

We proposed a model architecture to continuously aggregate context in order to accurately segment malignant tissue cell of colorectal histopathology WSI. The model combines multi-scale FOV using parallel convolutions within encoder and decoder blocks. Evaluating on two histopathology datasets, we show that HistNet is able to generate fine boundaries around tissue cells, as compared to existing segmentation models. Quantitatively, our model achieves an improvement of about 2% DICE score and 1.8% IOU score. The results reinforce the importance of context aggregation in histopathology image analysis.

References


