CS753 Project Report
Prostate Cancer Grade Assessment

Kartik Khandelwal 160070025
Sriram YV 16D070017

July 5, 2022

Contents

1 Introduction

Prostate Cancer is the second most common cancer worldwide. The cancer is highly treatable in its early stages, hence its early success full diagnosis is the key to decrease its mortality. Our Project aim to determine the grade (ISUP) of the cancer given the Whole Slide Image (WSI) of the cancer biopsy. ISUP grade ranges takes integer values from 0 to 5 with a higher value implying a more aggressive cancer. The task has been launched as a Kaggle competition as a part of the PANDA workshop at MICCAI 2020.

The task at it's core is an image classification task but we can't simply use of-the-shelf classifier trained on some classification dataset like ImageNet and finetune for our task because the WSI are very high resolution (10^9 pixels), hence which makes using the whole image for training highly inefficient. Sub sampling the images is not an option since it can lead to loss of information critical for assessment. Thus, the challenge is to be able to identify key regions in the image and use them to extract meaningful information.

2 Proposed Approach

We plan to use Deep Convolutional Neural Network for feature extraction and classification but first we need to design algorithms to identify key regions in the image which can be fed into the Deep CNN's for feature extraction. We explore three different approaches for the same which are explained in the next 3 subsections.

2.1 Approach 1 : Tiled Image

Based on this kaggle notebook, we extract top-N^2 $d \times d$ square tiles from the image on the basis of the number of tissue pixels in the image. These square tiles are then re arranged in row major order to create a $N * d \times N * d$ dimension bigger image which is then fed into the CNN for classification. Figure 1 shows the WSI of a prostate biopsy and figure 2 shows the tiled images for tile size $\in \{128, 256, 512\}$ while keeping the total size of the tiled image fixed (1536×1536) by changing N. The trade-off here is that for larger d, while we retain more locality but it occurs at the cost of including lesser information (tissue pixels) in our tiled image.

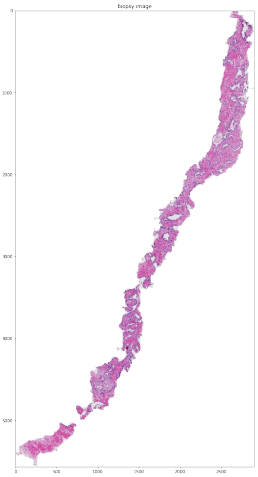


Figure 1: Example of WSI of prostate biopsy

2.2 Approach 2 : Attention Based Multiple Instance Learning

Multiple Instance Learning(MIL) is a variation of supervised learning where a single class is assigned to a bag of instances. For our task, each tile corresponds to an instance and the collection of tiles forming the bag, the label of the bag being the ISUP grade of the cancer for the WSI.

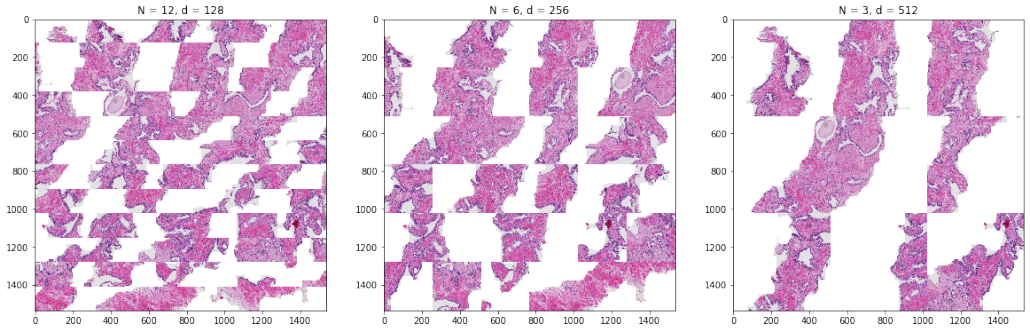


Figure 2: Example of tiled images for different sized tiles

Based on [1], we train a Multi Layer perceptron as the attention network along with the feature extractor Deep CNN network. The task of the attention network is to output weights for each of the instance feature vectors which are outputted by the Deep CNN network. The Instance feature vectors are then combined into a single feature vector using weighted linear combination where the weights corresponds to those given by the attention network. The idea is to assign higher weights to key instances (those containing cancerous cells, if any) which allow to predict the ISUP grade. Figure 3 shows an illustration of the method (image taken from [1]).

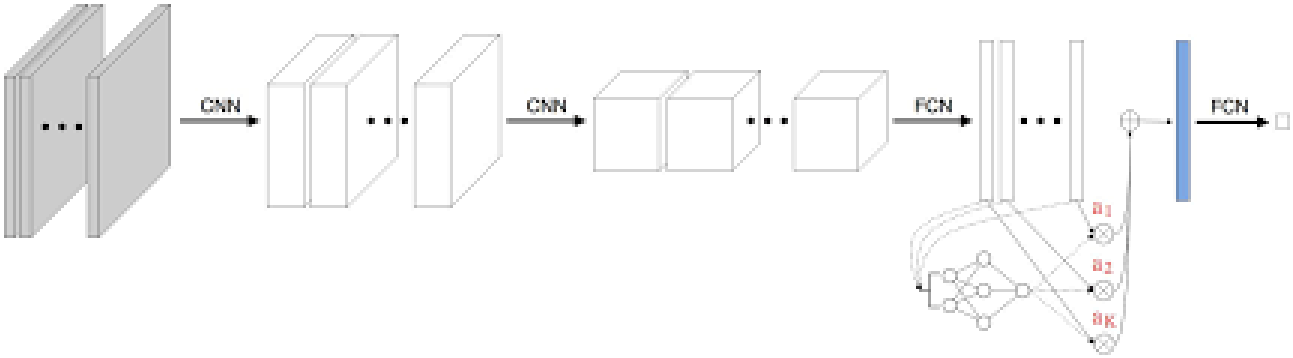


Figure 3: Attetion based Multiple Instance Learning

2.3 Approach 3: Segmentation based Multi-task learning

The segmentation masks provided with most train images can be put into use to improve the learned model. We can take up a multi step approach, wherein we segment out the cancerous tissue and pass the masked image to classification DNN. To be able to achieve this we would need to learn or use a very accurate segmentation to obtain accurate segmentation before passing to classification DNN. This further becomes a challenge due to the presence of inaccurate masks in the given dataset.

Instead we tried a multi-task learning approach, wherein learning to predict the segmentation aids the learning of ISUP grade classification. Features from intermediate layers of classification DNN fused with upsampled outputs at intermediate nodes of decoder to retain spatial information. We used the decoder from FCN with upsampling and unpool operations. The loss due to classification task and DNN task were differently weighted with high preference given to classification task.

The images in the datasets are from two different centres. The masks from Radboud are automatically labelled and some of the masks are inaccurate (some images with high ISUP grade have no cancerous tissue in the given masks). So, this required some data cleaning at the start of the experiments.

3 Implementation Details

3.1 Deep CNN Model architecture

We used EfficientNet-b0 as our feature extractor. EfficientNets are a family of models proposed in [2] which using intelligent scalling of network depth, width and resolution are able to superpass state-of-the-art accuracy with upto **10x** improvement in efficiency. Figure 4 shows a comparison of different scaling methods for images (taken from [2]). GeM[5] is used as the Global pooling layer.

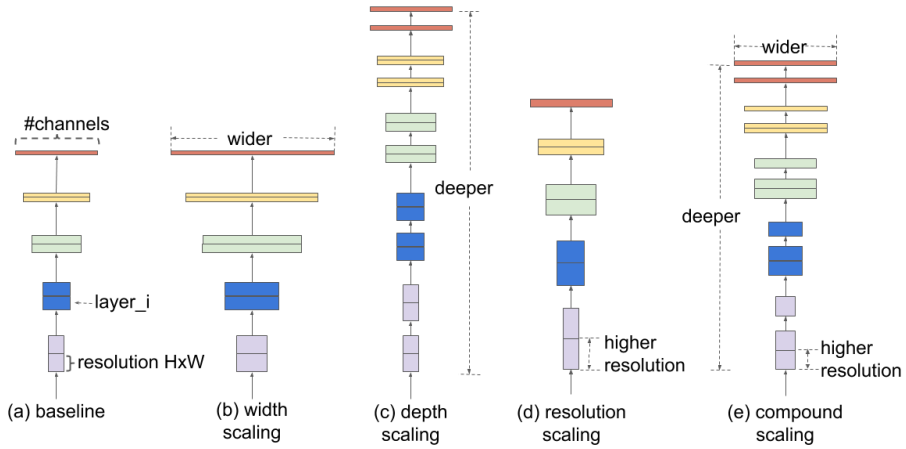


Figure 4: A comparison of different scaling methods

3.2 Training Specifics

We used pytorch as our ML framework. Nvidia Apex[3] was used for mixed precision training to be able to load large batches, Adam optimizer used for gradient descent. We use gradual Warm-up[4] followed by cosine annealing of learning rate which was initialized to 3×10^{-4}. All the models were trained for 30 epochs and the model with the best validation cohen kappa score was used for testing.

3.3 Code Link

We used pytorch as our ML framework. All our code is available at https://github.com/Kartik14/PANDA. The different approaches are available under different branches on the repository. We used the code from this kaggle notebook as starter code.

4 Experiments & Results

4.1 Dataset

We used the dataset released as part of the kaggle competition. It can be found here. The dataset contains 10617 train images along with 1000 test images from two different medical centres 'Karolinska' and 'Radboud'. The images are available as multi-level tiff images. For each train image, segmentation mask is also provided labeling each pixel as background, benign or cancerous for 'Karolinski' and as background, stroma, healthy, Gleason-3, Gleason-4 and Gleason-5 for 'Radnoud'.

4.2 Evaluation Metric

Quadratic Weighted Kappa (qwk) is used for as evaluation metric which measures the agreement between prediction and target. This metric typically varies from 0 (random agreement) to 1 (complete agreement). In the event that there is less agreement between the raters than expected by chance, the metric may go below 0. It attempts to adjust for the probability of the algorithm and labels assigning items to the same category "by chance". Weights are used to penalise more in cases where the disagreement is large, since there is an implicit order in the ISUP grade.

The quadratic weighted kappa is calculated as follows. First, an N x N histogram matrix O is constructed, such that $O_{i,j}$ corresponds to the number of isup_grades i (actual) that received a predicted value j. An N-by-N matrix of weights, w, is calculated based on the difference between actual and predicted values: $w_{i,j} = \frac{(i-j)^2}{(N-1)^2}$. An N-by-N histogram matrix of expected outcomes, E, is calculated assuming that there is no correlation between values. This is calculated as the outer product between the actual histogram vector of outcomes and the predicted histogram vector, normalized such that E and O have the same sum. From these three matrices, the quadratic weighted kappa is calculated as:

$$\kappa = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}}$$

4.3 Results

4.3.1 Approach 1 : Tiled Image

Figure 5 shows the confusion matrix for three different variation of d(tile size) and N(Number of tiles) on validation set (test set is hidden by kaggle until the end of competition). The qwk for the three different approaches are similar with the best for $d = 128$ at 0.8791.

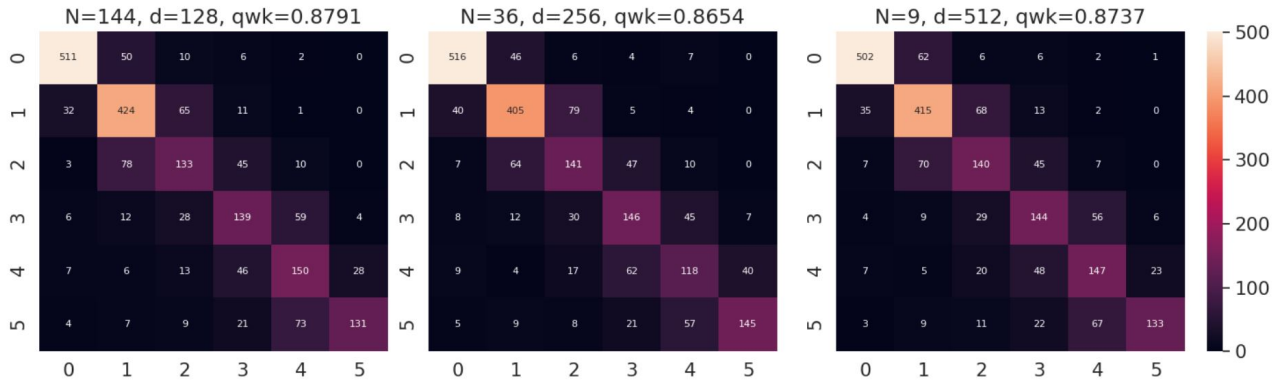


Figure 5: Histogram for different values of d and N

4.3.2 Approach 2 : Attention based MIL

Figure 6 shows the confusion matrix for approach 2. The qwk value for the validation set is 0.8640 similar to those achieved by approach 1. Figure 7 displays the attention maps along with the WSI and the segmentaion mask (provided in the dataset). It can be seen that the model is able to give more weight to the key regions fairly well. More attention maps are shown in Appendix

4.3.3 Approach 3: Segmentation based MTL

We found no improvement in classification on training jointly with the segmentation task. In fact, the test qwk dropped to 0.84 in this case. It appears that the features passed from classification DNN are insufficient to learn the segmentation. A closer look at segmentations produced would help analyze and improve the segmentation networks.

4.3.4 Overall Results

Table 1 compared the validation and test qwk obtained for different methods. The test qwk is the best for tiled image with $N = 36, d = 256$ despite it being a comparatively lower validation qwk. Also, surprisingly the test qwk for Attention based MIL is significantly smaller (0.77) even though its validation qwk is comparable to tiled image based method.

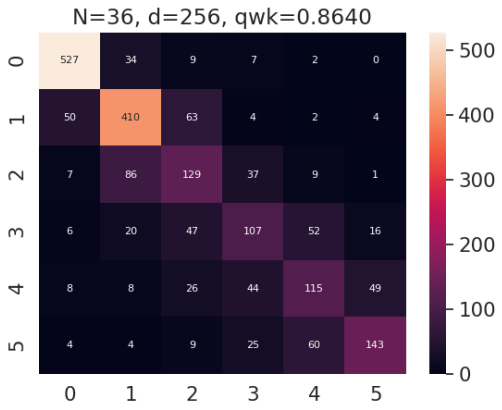
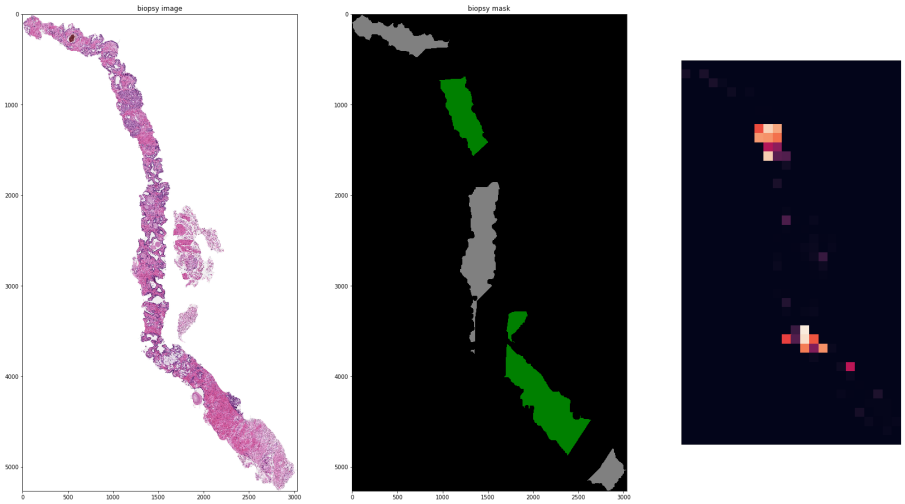


Figure 6: Histogram for attention based MIL method



Method		validation qwk	test qwk
Tiled Image	$N = 144, d = 128$	0.8791	0.85
	$N = 36, d = 256$	0.8654	0.86
	$N = 9, d = 512$	0.8737	0.83
Attention based MIL		0.8640	0.77
Segmentation based MTL		-	0.84

Table 1: Overall results comparing qwk for different methods

5 Challenges & Future Work

One of the major challenge we faced was to train the model fast. The images were very large and even after using EfficientNet b0 with mixed precision training, we were only able to afford a batch size of 4 on a single 11 GB gpu. This restricted us from being able to quickly iterate over different methods. For future work, we plan to look more closely at the attention based methods and understand the difference between the val and test qwk.

6 Conclusion

We have explored three different methods for the task of Prostate Cancer grade assessment and have been able to achieve a qwk score of 0.86 on the kaggle public leader board test set.

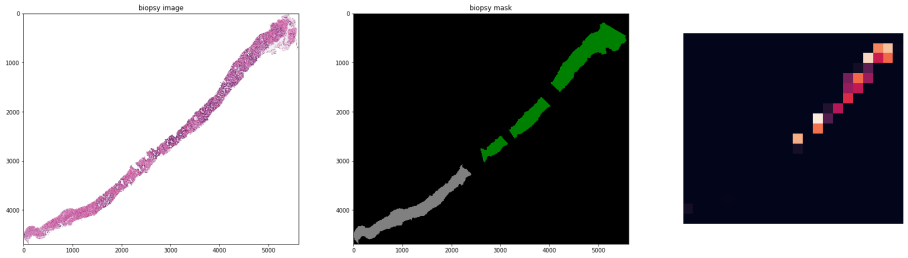


Figure 7: Attention maps obtained from approach 2, From left to right, WSI, Segmentation mask, Attention mask

References

[1] Maximilian Ilse and Jakub M. Tomczak and Max Welling *Attention-based Deep Multiple Instance Learning, 2018*

[2] Mingxing Tan and Quoc V. Le *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, 2019*

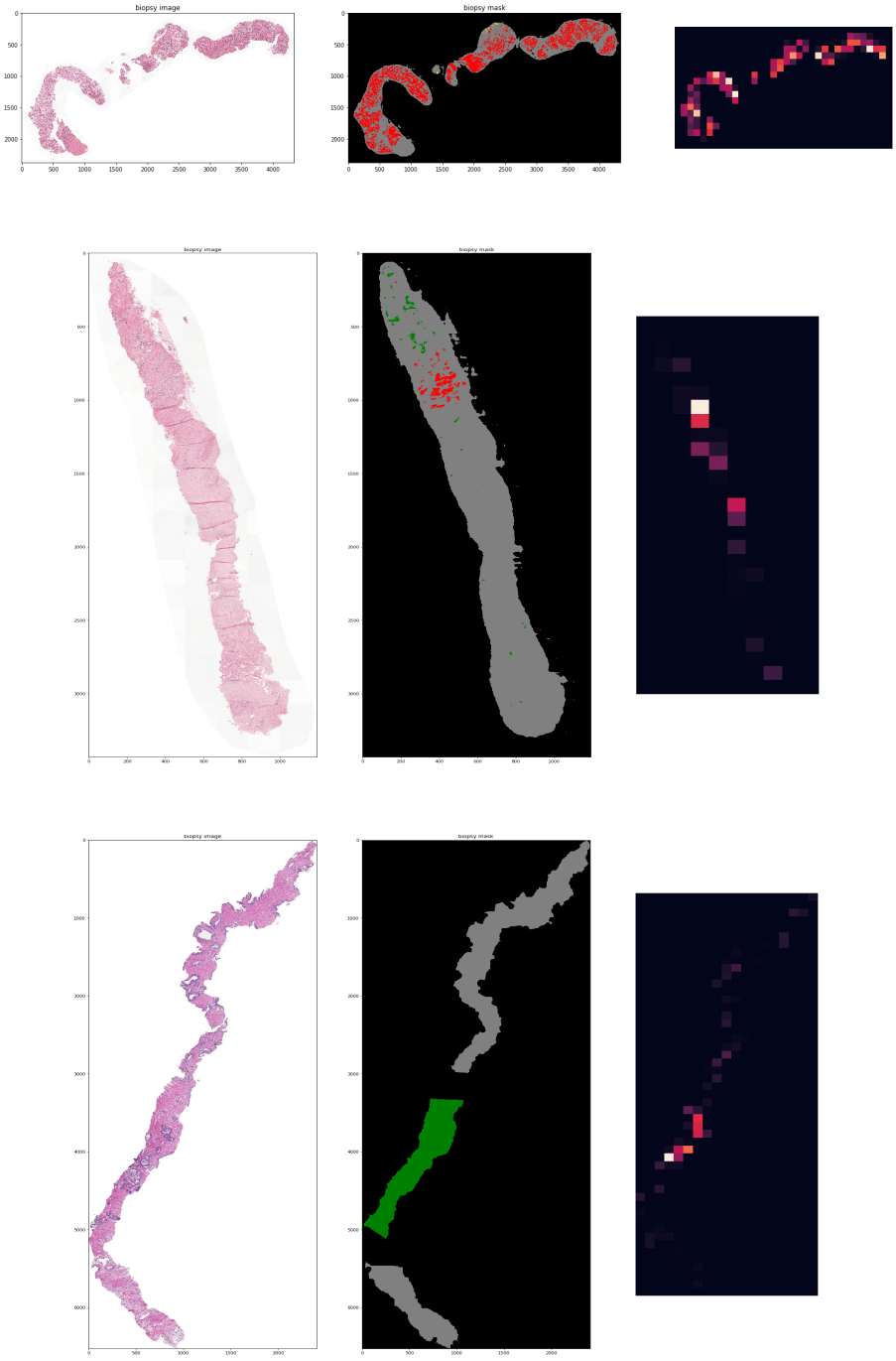
[3] https://github.com/NVIDIA/apex

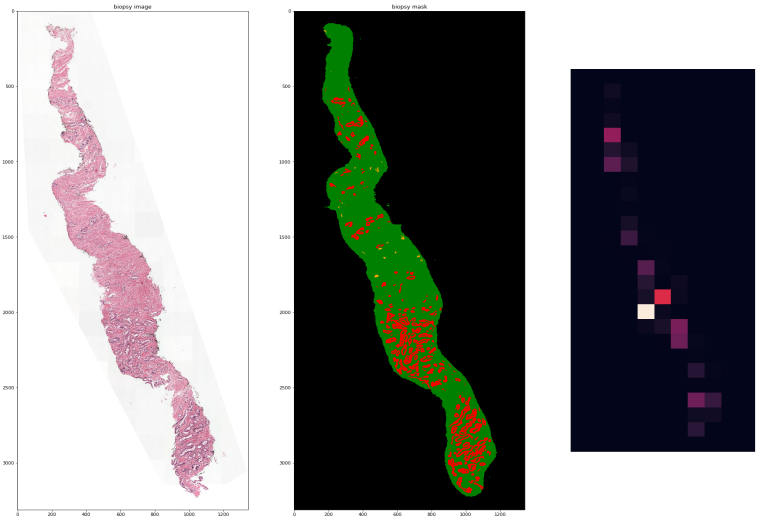
[4] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, Kaiming He *Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour*

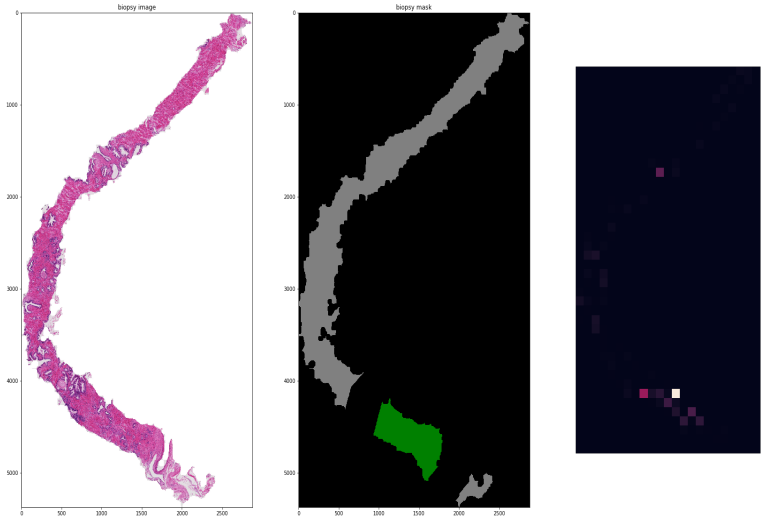
[5] Filip Radenovic, Giorgos Tolias and Ondrej Chum *Fine-tuning CNN Image Retrieval with No Human Annotation, 2017*

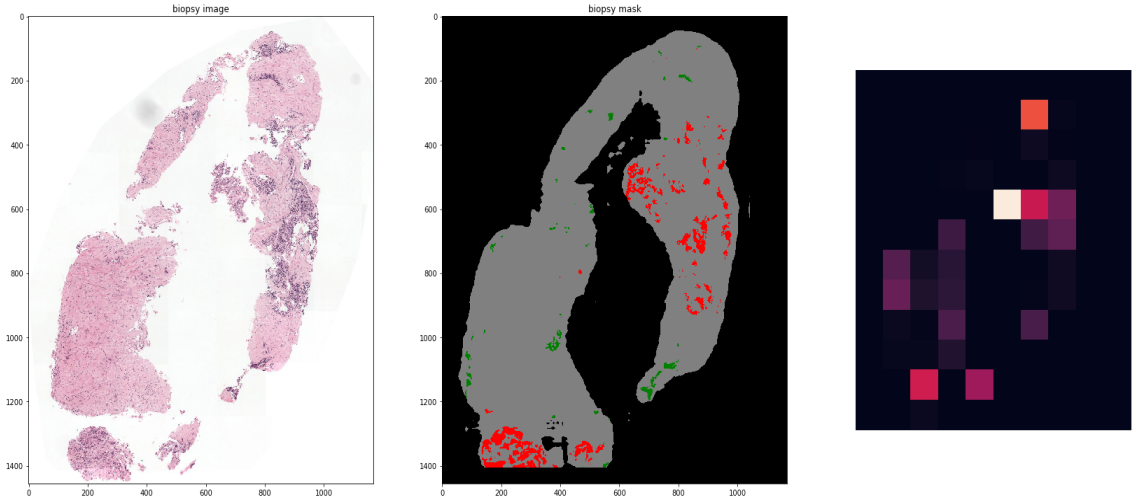
A Appendix

A.1 Attention Maps obtained using Attention based MIL




biopsy image
biopsy mask


biopsy image
biopsy mask


biopsy image
biopsy mask