Final Project- Hossein Moghimifam

BreastPathQ Challenge

1 Introduction

Approximately 14.5 million people have died due to cancer and it is estimated to surge up to 28 million by 2030. According to a study by the American Cancer Society (ACS), in the USA a total of 41,000 people died due to breast cancer, which makes it the second deadliest cancer among women after lung and bronchus cancer. There have been several efforts to use CNNs for breast cancer diagnosis using X-ray images [1] and WSIs [2] [3].

Neoadjuvant treatment (NAT) of breast cancer (BCa) can be considered as an option for patients with the locally advanced disease. Study of the tumor response to the NAT would provide useful information for patient management. Although the tumor size may not change, but the overall cellularity may reduce, making it an important factor when studying the response. Cellularity is defined as the percentage area of the overall tumor bed that is comprised of the tumor cells. In the current methods, tumor cellularity is manually estimated by pathologists on stained slides, the quality, and reliability of which might be impaired by the person assessing them.

In the diagnosis process, a biopsy followed by microscopic image analysis is a common procedure[1]. to visualize different components of the tissue, it is dyed with stains, in this case, hematoxylin and eosin (H&E). It allows pathologists to histologically inspect the structures and components of the breast tissue to determine the cancer cellularity.

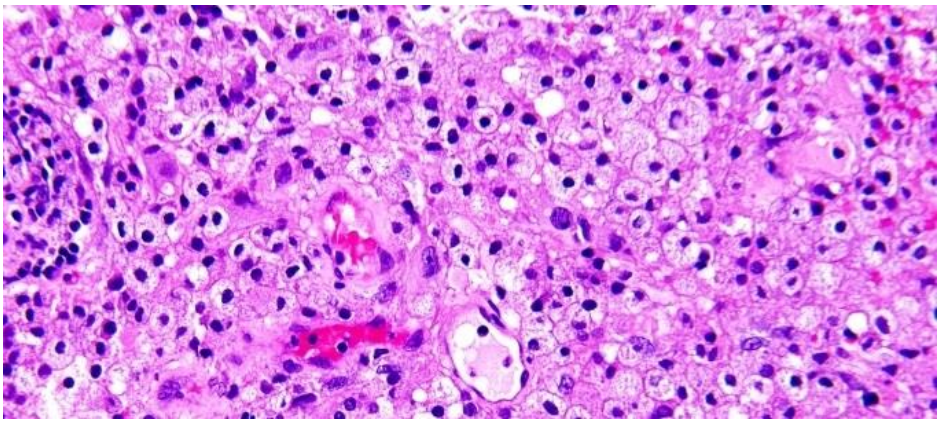

Figure 1- example of an H&E stained microscopic specimen [1]

The H&E stain would show the following response based on the cell in contact [1]:

- Nuclei: blue, black
- Cytoplasm: Pink
- Muscle fibers: deep red
- RBCs: orange red
- Fibrin: deep pink

In this report, I will describe a method for the determination of cancer cellularity from whole slide images (WSI) of breast cancer hematoxylin and eosin (H&E) stained pathological slides. This is a part of a "Grand Challenge" (BreastPathQ) [2] held by the SPIE (the international society for optics and photonics), along with Association of Physicists in Medicine (AAPM), and the National Cancer Institute (NCI). The publications for this challenge are not available yet, so there are no references for comparing the methods with the top achievers of the challenge. Furthermore, the challenge website has been unavailable since at least April 25th, making it impossible to asses the performance of the network on the test set.

2 Data and Evaluation Metrics

In this section, I will describe the dataset provided by the organizers of the competition and the evaluation metrics used to evaluate the performance of models.

2.1 Dataset

The dataset for this challenge was collected at the Sunnybrook Health Sciences Centre, Toronto. It comprises 96 whole slide images (WSI) which have been stained with H&E. WSIs were extracted from 64 patients with residual invasive breast cancer on resection specimens following neoadjuvant therapy. The specimens were handled according to routine clinical protocols and WSIs were scanned at 20X magnification (0.5 µm/pixel).

In the challenge, a training/validation set of 2,579 patches (2394 training, 185 evaluation) extracted from the above WSIs is provided. Each patch in the training and validation set has been assigned a tumor cellularity score by 1 expert pathologist, while the test set has reference standards from 2 pathologists. The test set has been prepared in an identical manner as the training set and contains 1,121 patches extracted from 25 WSIs.

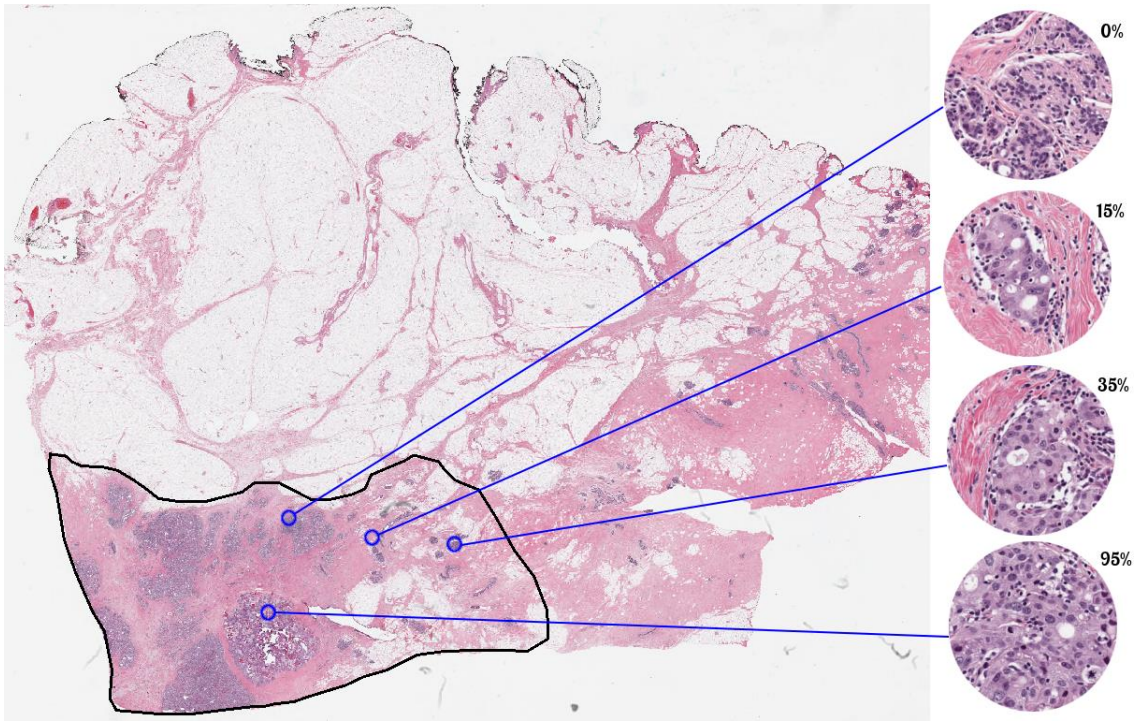

Figure 2- Extracted patch and their cancer cellularity

The labels of the test set for the original challenge is not known, and the challenge website calculates the test set score. Since the website is not available at the time of writing this report, I will randomly choose 15% of the training set as the test set. This would make the dataset distribution look like this:

Table 1- population of dataset

Set	*Population*
Training	2034
Validation	185
Test	360

2.2 Data preprocessing

The input data are RGB images of the resolution 512×512 pixels. The images are patches extracted from WSIs, so there is no need to remove excessive parts. They are the correct input size for famous networks like ResNet, VGG, and GoogleNet, so modification of these networks can be used readily. The only preprocessing that can help during the training phase is normalizing the images to the parameters that are used in pytorch pretrained models. As described in the documentation, the images have to be loaded in to a range of [0, 1] and then normalized using mean = [0.485, 0.456, 0.406] and std = [0.229, 0.224, 0.225]. The elements in the mean and std vector belong to the RGB channels respectively.

To classes have been defined to do the normalization and inverse it with respect to the above values. Inversing the normalization would be helpful for getting the original image for visualization purposes.

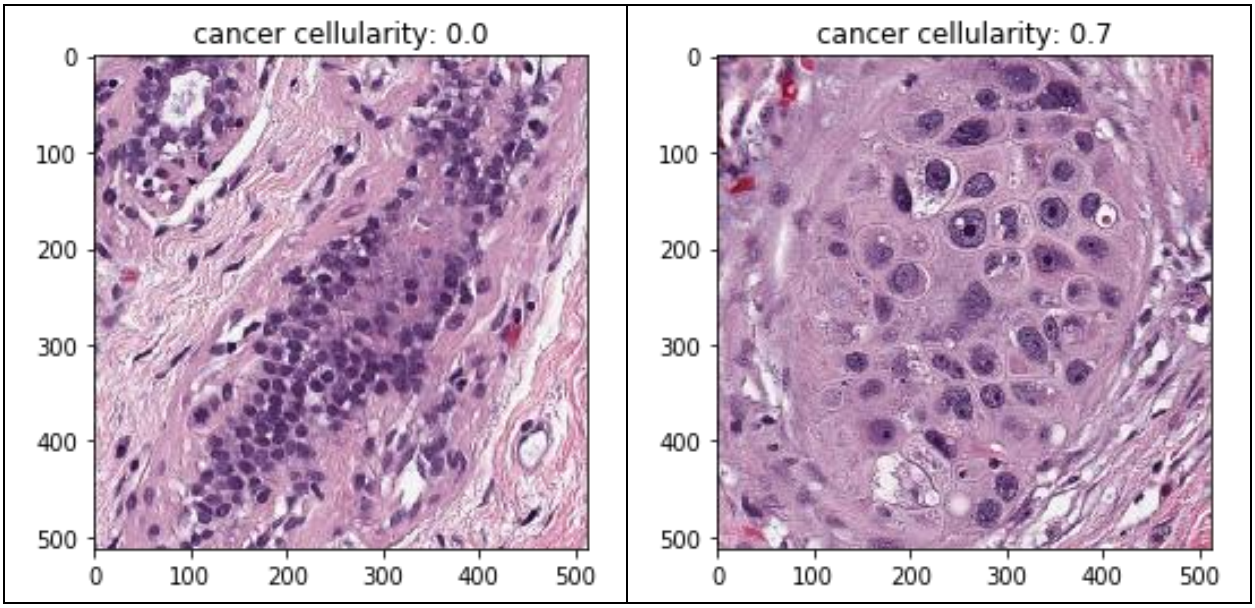

Figure 3- Examples of data with different cancer cellularity

2.3 Data augmentation

The images are microscopic images of the pathological slides, which can be put under the microscope with any angle. Rotating these images by any degree for augmentation should be fine, not in this case

though. They are scored by pathologists and all the pixels in the images should be present after the augmentation for the score to be valid. This limits the augmentation processes to horizontal and vertical flips. Augmentation would make the model more robust to the changes that might be seen in the test set.

Scaling and changing colors are the augmentation methods that make sense for this dataset. Scaling would change the magnification of the image and this would change the size of the features, which are important in pathological study of cells. Changing color would account for the small variation in the hue by dyes of different companies.

2.4 Evaluation Metrics

Method performance will be assessed using prediction probability (P_K) [1].

The definition of prediction probability is:

$$P_K = \frac{\left(\frac{P - Q}{P + Q + T}\right) + 1}{2}$$

, where P is the number of concordant pairs, Q the number of discordant pairs, and T the number of ties only in the submitted labels. This value would determine at what percentage the model would predict correctly. This metric rewards concordance and penalizes discordance and ties, meaning the higher the score, the better performing the model [1]. Python implication of

3 Method and Results

In this section, I will describe the approach to design the network for cellularity prediction. The general approach is to modify the pretrained models available in the Pytorch library. Based on the available computational power, I will investigate ResNet18, ResNet50, and AlexNet. Only the first one is doable on a CADE lab computer, so I will use Google Colab.

Output of all these networks is a vector of 1000 elements. In this project, we need to have a single value between 0 and 1 as the output. Adding a fully connected layer that has input of 1000 and output of 1 would cause loss of information. To avoid this problem, I would reduce the size of the vector step by step as shown in the figure below. The base model is the loaded pretrained model from the Pytorch library.

Learning rate for the fully connected (FC) layers should be higher than the base model because it is not pretrained. The learning rate for the FC layers was chosen to be 10 times the base model for all instances. Wherever learning rate is mentioned in this report, it refers to the base model learning rate.

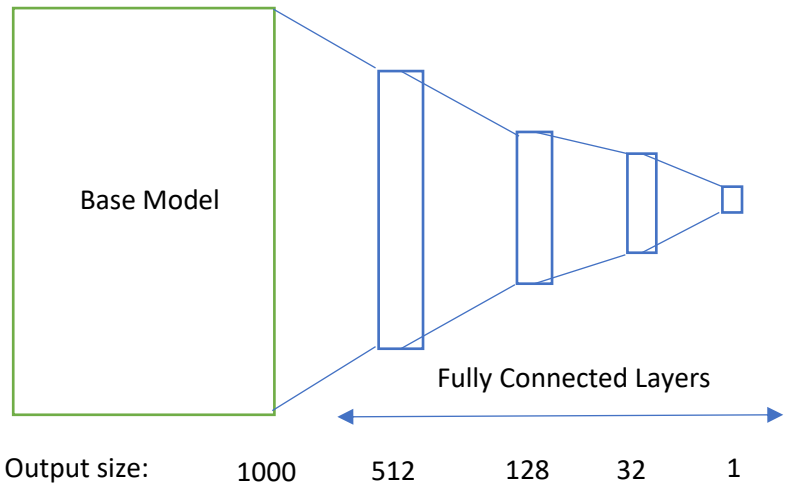

Figure 4- Neural Network Schematics

The performance of the modified networks on the validation set is investigated. The loss function for training is mean square error (MSE) since we care about finding a single value output and there is no classification problem.

3.1 ResNet18

ResNet variants have shown great performance in most of the deep learning problems. ResNet18 is the simplest of the pack. As described above, I will add fully connected layers to the pretrained ResNet18 and investigate its performance.

As starting point, the networks will be run for 20 epochs. Training loss and validation prediction probability (P_K) will be monitored. Batch size for the training set is 20, change of which had minimal effect on the performance.

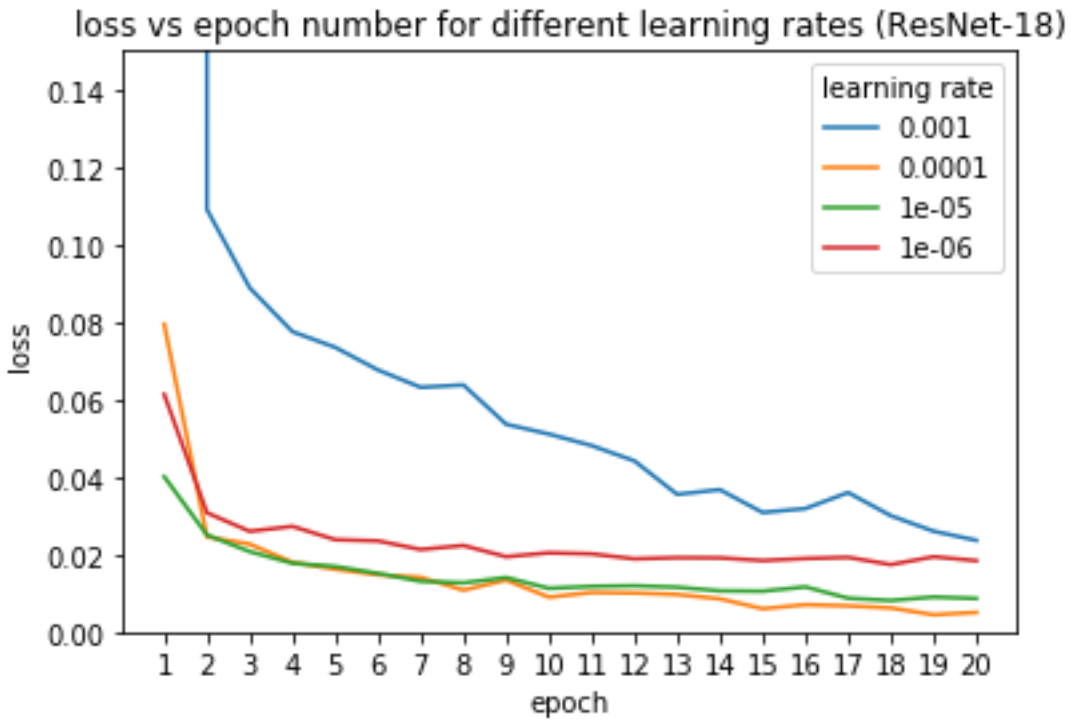

Figure 5- loss vs epoch for ResNet-18 (20 epochs)

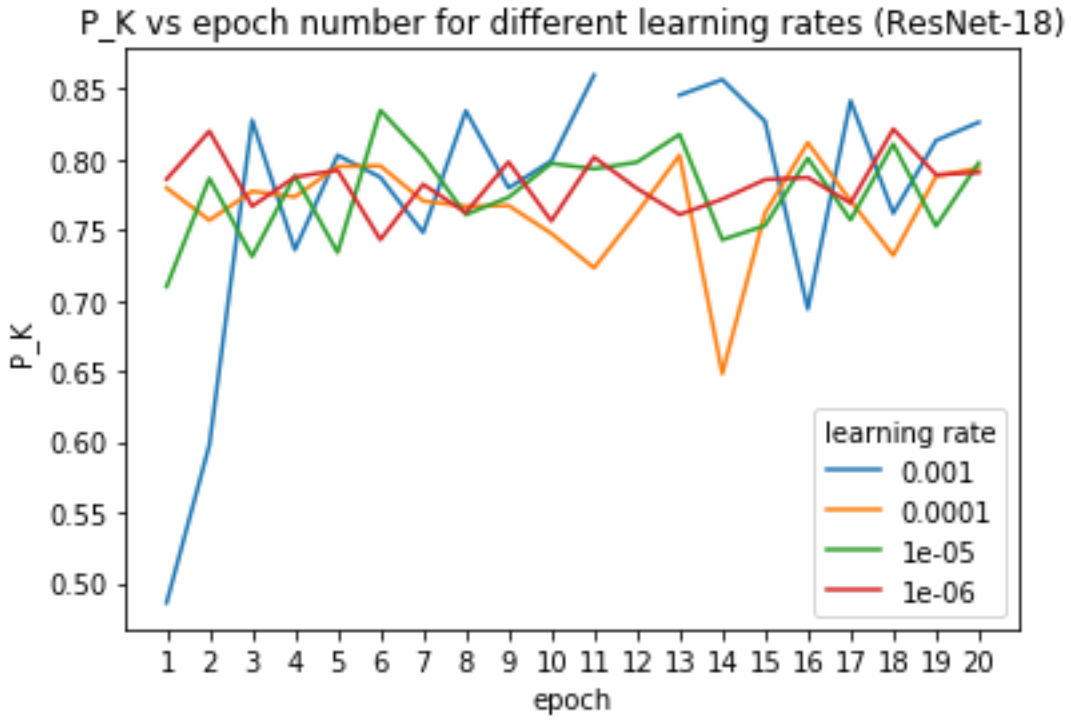

Figure 6- P_K vs epoch for ResNet-18 (20 epochs)

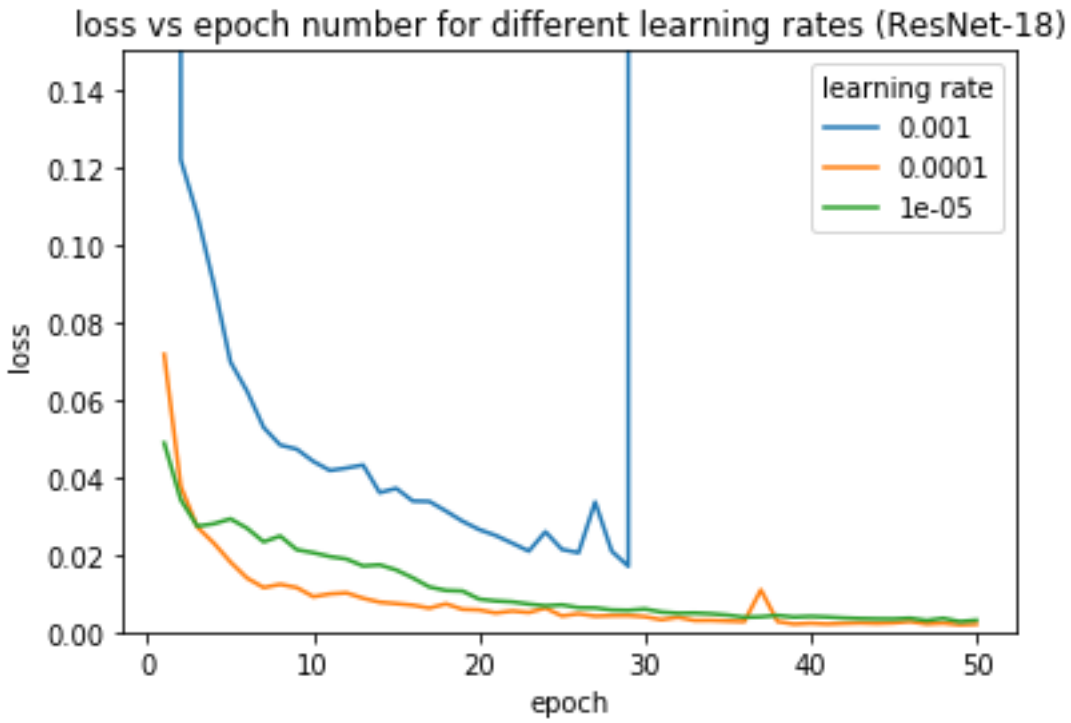

Figure 7- loss vs epoch for ResNet-18 (50 epochs)

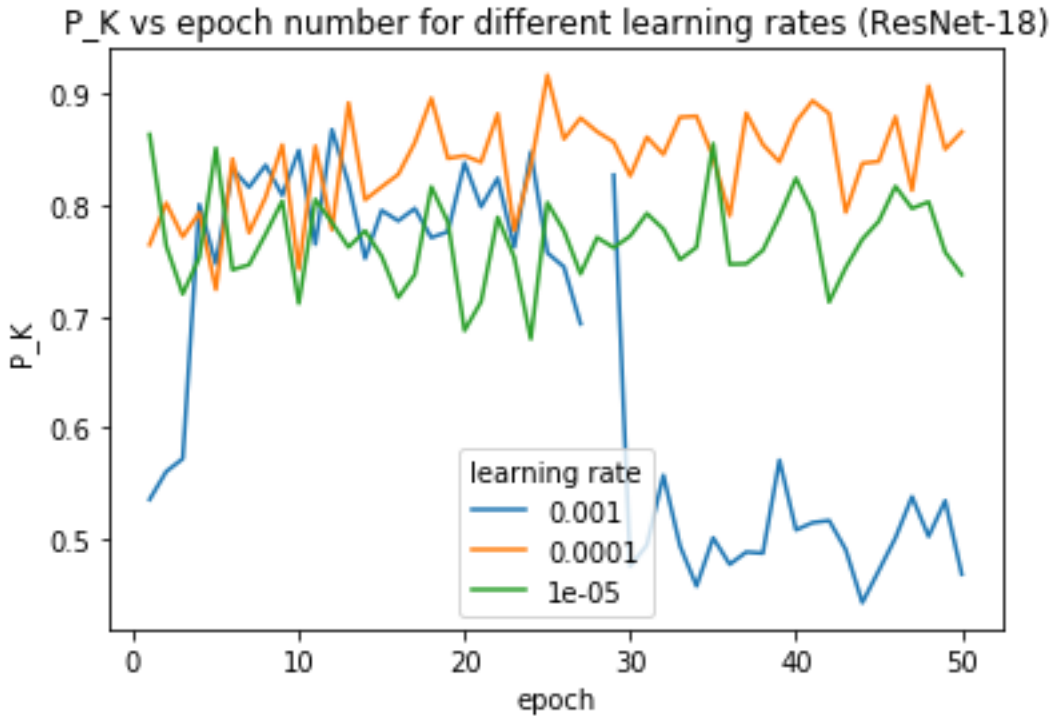

Figure 8- P_K vs epoch for ResNet-18 (50 epochs)

There are several points that can be inferred from above figures. The network can learn the problem to some extent and has a limit on how well it can perform on this dataset and this specific task. After some

point it would oscillate around a specific value and achieving higher or lower prediction probability is by chance even though the loss is decreasing. There are big jumps in the loss value, the biggest of which happens for the learning rate of 0.001 where is diverges. There is a single point where P_K is not defined. It is because of the definition of the scoring function. Ideally, the scoring function should receive the whole dataset as the input. But because of the computational limit, I'm feeding it as batches. If a single batch is all ties, it would return nan.

Although the network reaches its performance peak, based the loss and P_K figures, learning rate of 0.0001 shows the best performance. The model with the highest P_K was saved for evaluation on the test set.

3.2 ResNet50

The more complicated network of ResNet family is used to compare its performance with the least complicated one. Setup is the same as the previous network.

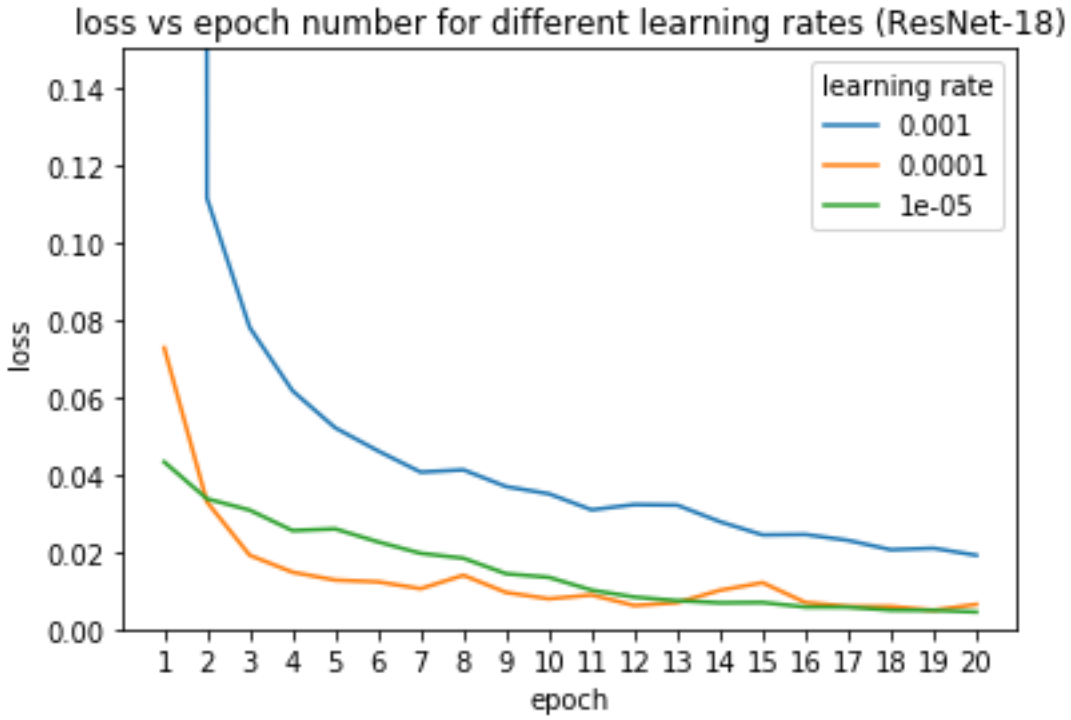

Figure 9- loss vs epoch for ResNet-50 (20 epochs)

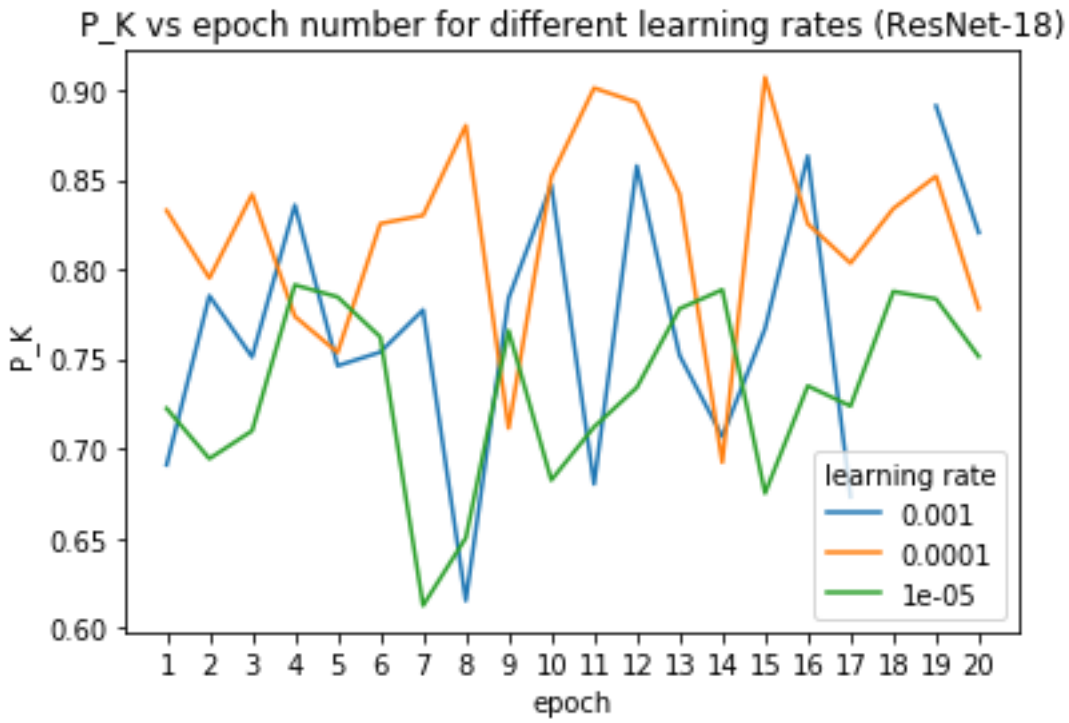

Figure 10- P_K vs epoch for ResNet-50 (20 epochs)

ResNet-50 shows approximately the same performance and the problems as ResNet-18. It achieves the same range of loss and score. Oscillations are higher for this network than the previous one though. Based on the figures, learning rate of 0.0001 shows the best performance and has the better performance than any of ResNet-18 networks in 20 epochs.

3.3 AlexNet

A different network was test with to see whether it would be able to overcome the problems of ResNets. With the same setup for 20 epochs, its performance is shown in the figures below.

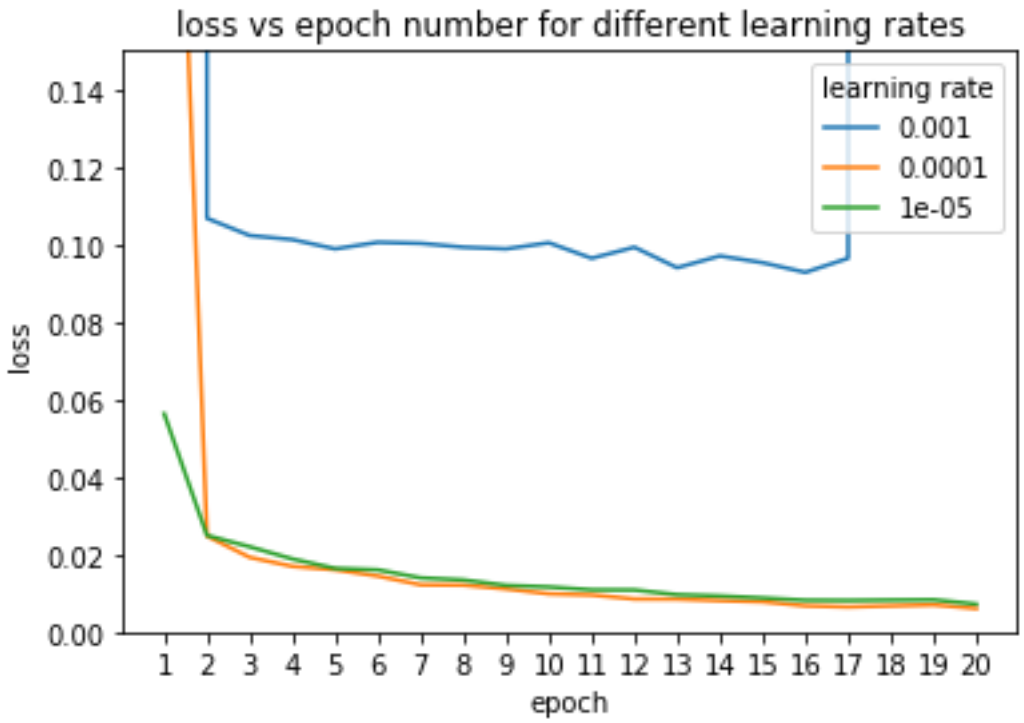

Figure 11- loss vs epoch for AlexNet (20 epochs)

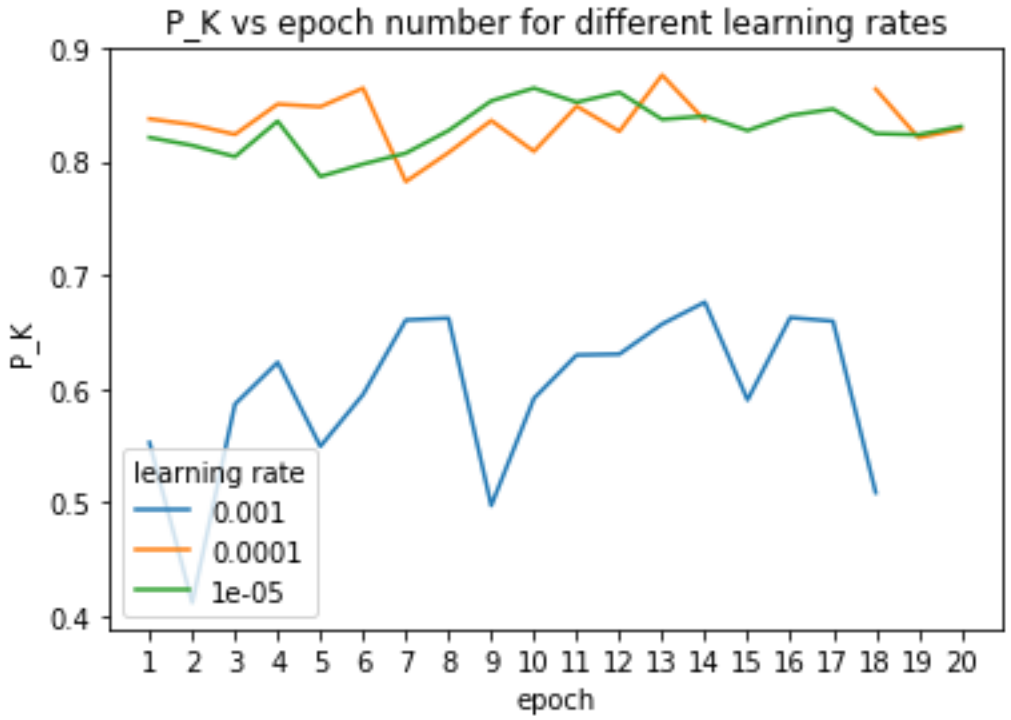

Figure 12- P_K vs epoch for AlexNet (20 epochs)

As it can be seen from the figures, AlexNet reaches a peak in score, but oscillates less than ResNet variants. Learning rate of 0.001 shows significantly lower performance, and learning rate of 0.0001 shows the best performance of the three.

3.4 Performance on the test set

Since the test set is a part of the training set put aside, I cannot compare performance of these networks on the actual test set of the challenge and with other state-of-the-art networks. To find the best performance of the networks on the test set, the trained network with the best performance in the previous part was run for 60 more epochs and any improved network was used to evaluate the performance on the test set.

Table 2- Performance of different models on the test set

Network	P_K *on the test set*
ResNet18	0.7992
ResNet50	0.8502
AlexNet	0.8369

4 Conclusion

Although all networks reached their peak during training, but the scores on the test set are on the acceptable side. ResNet50 showed the best performance among the three models followed by AlexNet and ResNet18. To have a meaningful learning, we need to use more complicated network such as combining ResNet features with inception [3]. Changing filter sizes to match the size of the cells in the images might help too.

Using segmentation seems to be a more practical approach in this problem. Deep neural networks can find features of the cells matching the ones with cancer. Using pixel labels of segmentation results for finding the cancer cellularity may also help improving the network performance. Also, using a pretrained network on pathological images can help improve the performance of the network.

5 References

[1] N. Wu, K. J. Geras, Y. Shen, J. Su, S. G. Kim, E. Kim, S. Wolfson, L. Moy and K. Cho, "BREAST DENSITY CLASSIFICATION WITH DEEP CONVOLUTIONAL NEURALNETWORKS," *Arxiv:1711.03674.*

[2] M. Peikari, M. J. Gangeh, J. Zubovits, G. Clarke and A. L. Martel, "Triaging Diagnostically Relevant Regions from Pathology Whole Slides of Breast Cancer: A Texture Based Approach," *IEEE Transactions on Medical Imaging,* vol. 35, no. 1, pp. 307 - 315, 2016.

[3] D. Wang, A. Khosla, R. Gargeya, H. Irshad and A. H. Beck, "Deep Learning for Identifying Metastatic Breast Cancer," *Arxiv:1606.05718.*

[4] M. Z. Alom, C. Yakopcic, T. M. Taha and V. K. Asari, "Breast Cancer Classification from Histopathological Images with Inception Recurrent Residual Convolutional Neural Network," *Arxiv:1811.04241.*

[5] [Online]. Available: https://laboratoryinfo.com/hematoxylin-and-eosin-staining/. [Accessed 1 5 2019].

[6] "BreastPathQ challenge," [Online]. Available: http://spiechallenges.cloudapp.net/competitions/14. [Accessed 20 4 2019].

[7] W. Smith, R. Dutton and N. Smith, "A measure of association for assessing prediction accuracy that is a generalization of non-parametric ROC area," *Statistical Medicine,* vol. 15, no. 11, pp. 1199-215, 1966.

[8] W. D. Smith, R. C. Dutton and T. N. Smith, "Measuring the Performance of Anesthetic Depth Indicators," *Clinical Science,* pp. 38-51, 1996.

[9] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha and V. K. Asari, "Improved Inception-Residual Convolutional Neural Network for Object Recognition," *Arxiv:1712.09888.*