Assignment 3

1604827
CS904 - Computational Biology

March 19, 2020

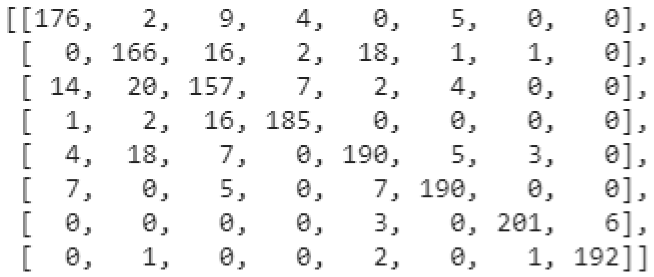
Question 1

Features

In this question I have started by extracting all the features that I use. These are the mean of each channel, the variance of each channel, the entropy of each channel, a histogram of each channel the PCA singular values of each channel and image moments of each image gabor and LBP.
The mean and variance probably wont be good features to use as they are very general parts of the image.
Entropy is an important feature as it relates the intensity of the image in places, this is useful as the places where target cells are are stained and thus have a higher intensity.
The histogram is taken with 20 points on the graph as this reduces some of the redundancies in using all 256 potential pixel intensities. When looking at the histograms I observed some patterns between the target number and the plots so I think that these will be a good factor to use.
I have taken the first 80 singular values this is because in testing the network and models I found that these produced results with the best RMSE, and this number of singular values explained a large amount of the variance. This means that the PCA will be a good and important feature.
I have also chosen to use two different image moments features these being Hu moments and Haralick features. Hu moments are useful as if two shapes are the same in an image, the image moment will necessarily be the same for those images. This is useful as it can be used to see how similar two images are, and hence is a good factor in prediction. Haralick features generate 14 statistical features based on the texture of an image, these have been shown to be useful in the problems similar to this one. These properties of the image moments make them important factors in this task.
The gabor feature gives a way to measure the texture of the image and proves to be a good predictor.

Results

For all of the models the SVM has had the hyper-parameters fine tuned to offer the best performance. And the data has been split into a 33% testing and 66%training sets.
Using only the mean of each of the images channels the accuracy is 77%.



```
[[176,   2,   9,   4,   0,   5,   0,   0],
 [  0, 166,  16,   2,  18,   1,   1,   0],
 [ 14,  20, 157,   7,   2,   4,   0,   0],
 [  1,   2,  16, 185,   0,   0,   0,   0],
 [  4,  18,   7,   0, 190,   5,   3,   0],
 [  7,   0,   5,   0,   7, 190,   0,   0],
 [  0,   0,   0,   0,   3,   0, 201,   6],
 [  0,   1,   0,   0,   2,   0,   1, 192]]
```

Figure 1: Confusion matrix produced by my model

Using only the variance of each of the images channels the accuracy is 70%.
Using only the PCA of each of the images channels the accuracy is 83%.
Using only the entropy of each of the images channels the accuracy is 75%.
Using only the histogram of each of the images channels the accuracy is 64%.
Using only the moments of each of the images channels the accuracy is 72%.
Using only the LBP of each of the images channels the accuracy is 66%.
Using the gabor features only an accuracy of 80% can be archived.
Combining all of the features an accuracy of 87.8% can be achieved.
Using the best two and three features gives an accuracy of 86.6% and 87.2% is achieved respectively.
Using the top three and the moments gives an accuracy of 87.6%.
My best model achieved an accuracy of 88% this is about 6% worse than the best performing CNN model. It is also slightly better than the models presented in the papers accuracy of 87.4%. However, as can be seen in the confusion matrix of both my results and the papers results, the types of images that the models performs worse on are the same.

Question 2

In this part several different models where trialed the performance of them can be seen in the table.

Model	Accuracy 5 epochs	Accuracy 8 epochs	Accuracy 12 epochs
VGG16	91.2	91.8	92.6
VGG19	89.6	91.3	92.0
resnet 18	89.8	91.1	91.8
resnet 34	91.9	91.2	92.4
resnet 50	91.8	91.4	93.1
resnet 101	93.7	92.8	92.7
resnet 152	92.2	93.5	94.2
squeezenet 1_0	89.7	89.9	91.0
squeezenet 1_1	87.7	87.0	88.2
densenet 121	92.3	94.5	94.5
densenet 169	92.2	93.1	94.2
densenet 201	92.6	93.1	93.0
densenet 161	91.2	92.7	92.1
alexnet	85.5	85.6	85.8

The confusion matrices for the top three performing models are shown below. As can be seen in them the stroma is the hardest to classify with tumor also being hard to classify. The complex class is often over predicted accounting for most of the misclassification of the stroma. The misclassification of the tumor is caused by the complex, lympho, debris and mucosa.

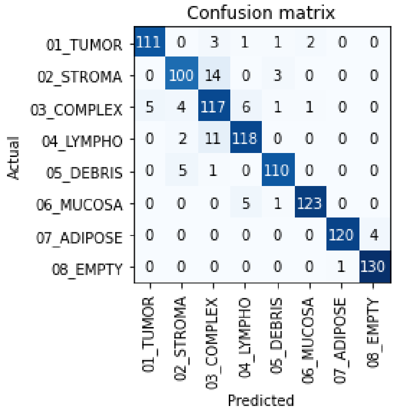


Figure 2: Densenet 121 confusion matrix

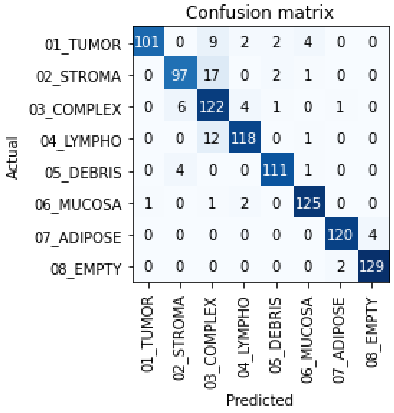


Figure 3: Densenet 169 confusion matrix

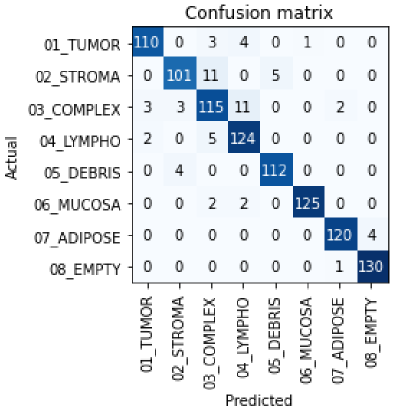


Figure 4: Resnet 152 confusion matrix

Question 3

In this part of the question the WSI was loaded and then split into patches. These patches can be used for classification.