Question 1

0.1 Introduction

The proportion of cancer cells in tumor tissue is known as tumor purity. In high-throughput genomic analysis, we need to minimize the contamination of normal cells. Obtaining accurate estimates of tumor purity is therefore essential for accurate pathological assessment and sample selection.

In the given paper, author constructed a multi-instance deep learning model to predict the tumor purity of H&E stained digital histopathology slides. This model consists of three components: *feature extractor module*, *MIL pooling filter*, and *bag-level representation transformation module*.

In this question, we aim to implement a simple version of this model to classify the digit 0 and 7 in the MNIST data set. Every bag consists of 100 images with a fraction x of digit 0 and 1-x of digit 7; each image consists of 28 × 28 pixels. We donate the fraction x as the purity of each bag.

0.2 Model Construction and Results

At first, we constructed a simple Convolutional Neural Networks(CNN) to regression on digit 0 and digit 7. Fig. (1) gives the visualization of the network structure.

The basic structure of this network as:

1. convolution layer with $kernel\ size = 5$.
2. down sampling layer employed max-pooling with $kernel\ size = 2$.
3. convolution layer with with $kernel\ size = 3$.
4. down sampling layer employed max-pooling with $kernel\ size = 2$.
5. liner full connection layer with $dim = 800 \times 400$
6. liner full connection layer with $dim = 400 \times 50$
7. liner full connection layer with $dim = 50 \times 2$

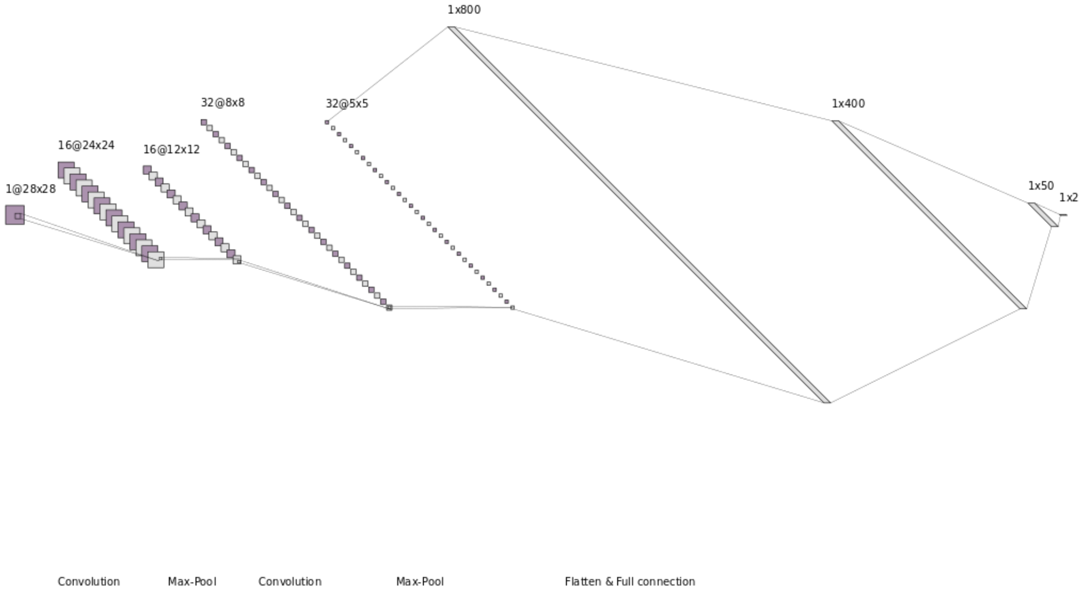


Figure 1: Network Structure

We trained the model using the SGD optimizer with a learning rate = 0.001, momentum = 0.9. The batch size was 100. We employed cross entropy as the loss function. Fig. (2) gives the network accuracy and loss during training. We evaluated our model on the test data set, then obtained 99% accuracy. Fig. (3) details the ROC curve of test data. We can find that the area under the ROC curve is approximately equal to 1, which means this model performances well.

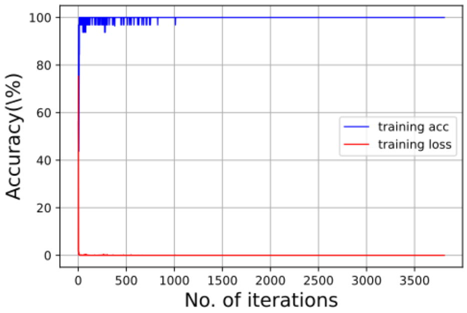


Figure 2: Training Accuracy and Loss

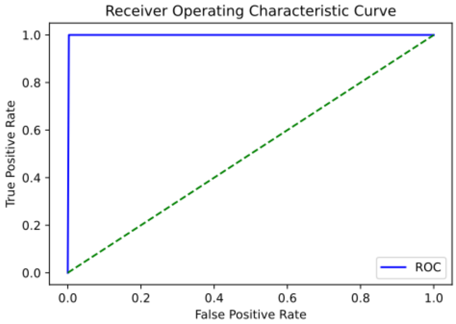
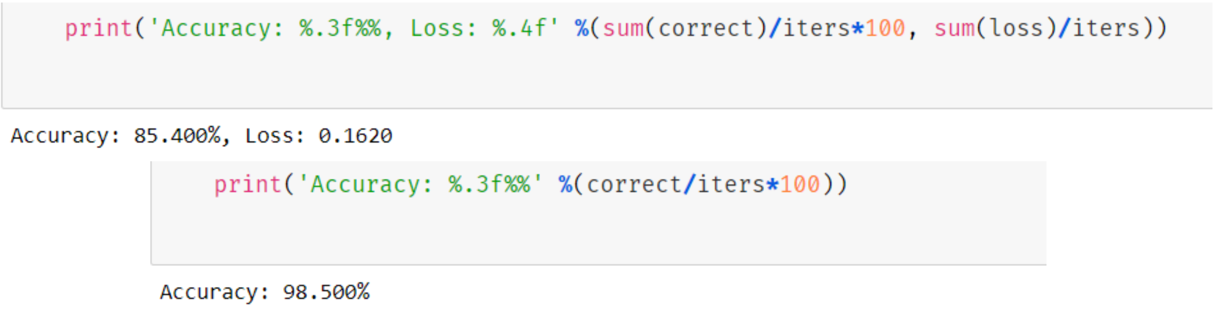


Figure 3: ROC Curve

After the model's training, we first construct the bags contain 0 and 7. Regard the number of 0 in each bag as the top-level label, or, the purity.

We called our CNN model for each sample in each bag to identify it. Then, sum up the times of predict 0 as the results of each bag. By comparing the number of zeros in the 100 images identified by the model with the actual number of zeros. That is, by comparing this predicted purity of each bag with the top-level label, we could get the final accuracy result. Here we use the absolute value between predict and label as the model loss.

In the end, this model implements the purity detection between digit 0 and 7 with accuracy = 85.4%. In machine learning, this may not be a very high accuracy rate. However, this is because the bag level's dimensionality is too large (i.e., purity in [0,100]). In fact, the average loss = 0.1620 shows that this model performance effectively. Therefore, if we tolerate the predicted purity in the interval of 1 above or below the true purity, the adjusted model obtains accuracy equal to 98.5%. Fig. (4) shows the accuracy of them.



```
print('Accuracy: %.3f%%, Loss: %.4f' %(sum(correct)/iters*100, sum(loss)/iters))
```

```
Accuracy: 85.400%, Loss: 0.1620
```

```
print('Accuracy: %.3f%%' %(correct/iters*100))
```

```
Accuracy: 98.500%
```

Figure 4: Model Accuracy

0.3 Conclusion

To conclude, in this question, we developed a CNN network to predict the purity of 0 in a 100 images bag. At the bottom level, the single network achieved a 99% correct rate for identifying the image. In the bag level, the adjusted model reached 98.5% accuracy for predict the bag purity.