**1. Read the file**



**2. Operate the file**

a. Read the files related to 4 labels respectively

b. Split the file into training and test set

c. Label the training set

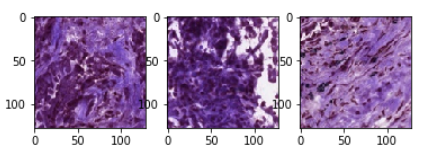
d. Label the test set



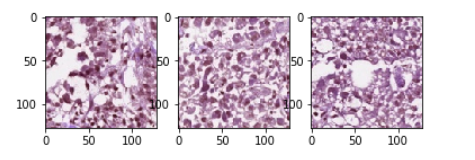
Here, the reason for chosen the training set from [100:484] is that after checking the data labels, we find that they are unbalanced, the minimum number of label is for other cases, which size is 484, hence, at the beginning, we just select the data until 484 to ensure the balancing problem.

**3. Visualize the data**

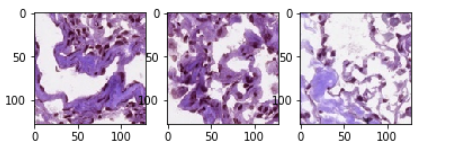
a. artifacts



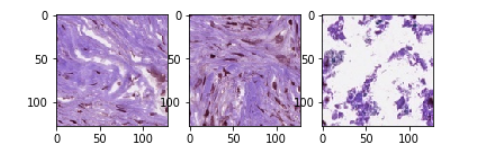
b. cancer



c. normal

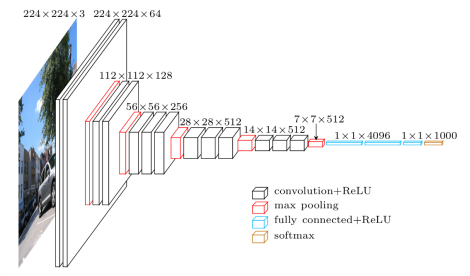


d. other

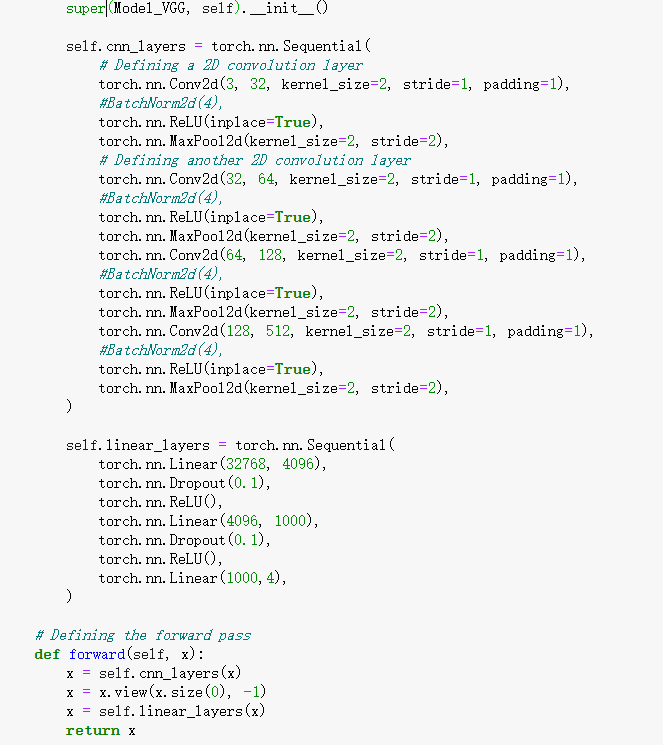


**4. Build the model**

**VGG model**

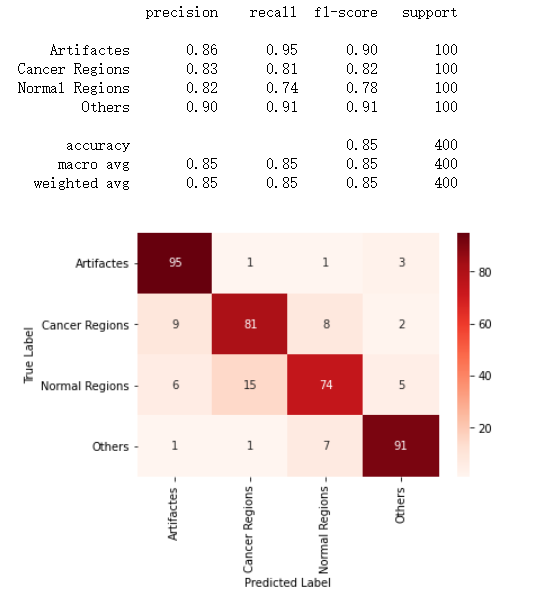
****

Following the structure of the VGG model, I just set up the model as below



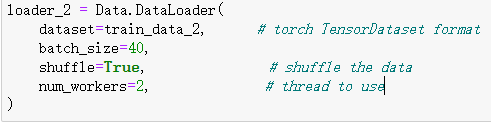
**5. Classification report**

By using the VGG model, we can achieve the accuracy score for about 85 percent.

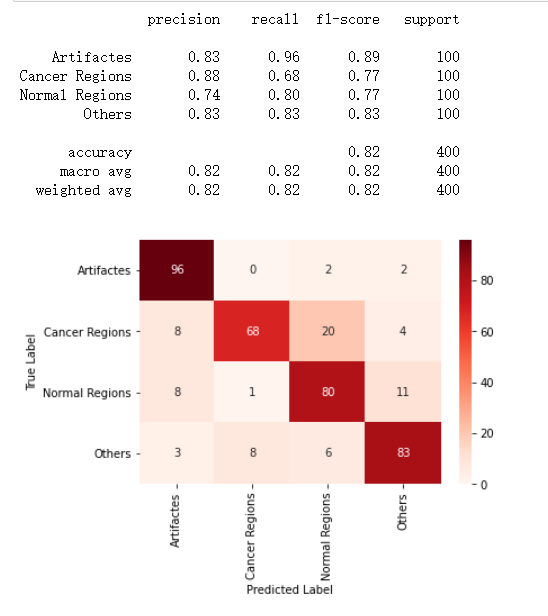


**6. Changing the batch size**

Since for the first time, we don’t set the batch size, we regard whole dataset as one batch, the advantage of controlling the batch size is to prevent overfitting, at this time, we set the batch size as 40 to see the effect



The classification report is as follows



**7. Add the batch normalization**

To improve the model, we come across the idea of adding the batch normalization, the use of batch normalization is to remove the effect of outliers. Running the model on my presentation teammate’s computer, we find that the accuracy is improved a lot, from previous 85 percent to 92 percent.

