Zhou Wentao(1003770) Xiao Tianqi(1003684)

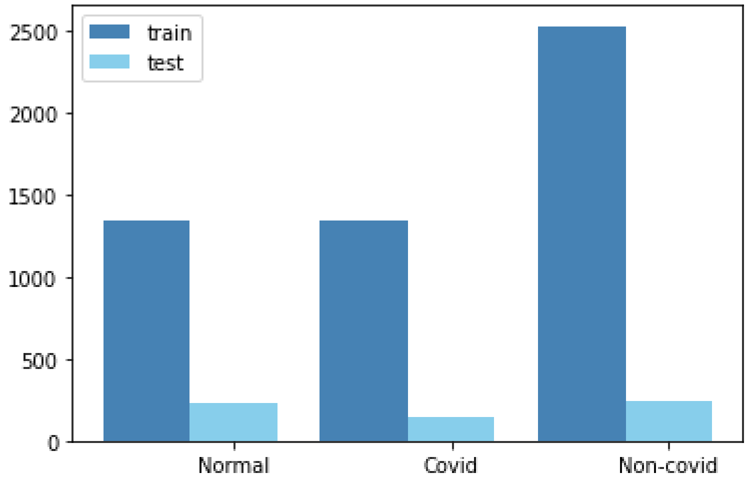
For convenience, we have modified the dataset a little bit, moving the covid and non-covid folder to the same level as the normal folder, skipping the infected folder.

The following are the questions in the instruction.

1. Your PDF report should explicitly mention what needs to be done to run your code.

   Using Google Colab to open Small Project.ipynb, run the cells one by one.

2. You are expected to explore the dataset slightly. Provide graphs showing the distribution of images among classes and discuss whether or not the dataset is balanced between classes, uniformly distributed, etc.



3. You are expected to discuss the typical data processing operations to be applied to your images. Typically, you have seen in the Demo Notebook, that our images needed some normalization was one, but why did we need it anyway? Are there other preprocessing operations that are needed for our images?

   All the pixels are normalized between 0 and 1, batch normalization is also applied at each convolutional layer to achieve faster convergence.

4. discuss the differences between the two architectures above and defend which architecture you decided to go for and why.

   We decided to use the model with one 3-classes classifier. From the bar chart above, we can see that the training set is extremely unbalanced. Around three-fourths of the training set are labeled infected. If we use a binary classifier for this, the model will tend to classify everything as infected, though the accuracy can easily reach over 75%, the f1 score will be low, which is undesirable. Therefore, we decided to use the three-class classifier and the cross-entropy loss. Now the label with the most training samples covers only about half of the whole training set. It is more balanced in this way. The performance of the model will be better regarding accuracy, precision, and recall.

5. Discuss the reasons behind this choice of architecture.

   The structure contains four convolutional layers with max pooling followed by three fully connected layers using softmax as the activation function for the final layer. At first, the model has only two convolutional layers and one fully-connected layer. However, the model is underfitting for the task, even the training accuracy is low. As a result, we gradually added more layers to the model. Meanwhile, we also increase the probability of the dropout as the difference between the accuracy of the training set and the test set is too big. The parameters of the model we use is the one with the highest accuracy testing with the test set during the 40 epochs.

6. You are expected to train your model using mini-batches of the training set and may freely decide on the value for a mini-batch size. Discuss it in your report.

   The mini-batch size is 16. It is not too big so that there is some randomness in the training set, also not too small so the result can be relatively stable.

7. Explain your choice of a loss function and its parameters, if any.

   The loss function is a cross-entropy loss which is the standard choice for a multi-class classification task.

8. Explain your choice of an optimizer and its parameters, if any (e.g. learning rate, momentum, etc.).

   The optimizer is Adam, it is the combination of momentum and RMSProp. The learning rate is set to 0.0005, which is not too big so that it is hard to find the optimum but also not too small so that it takes a lot of epochs to converge.

9. Explain your choice of initialization for your model parameters.

   The initialization is using Kaiming's normal distribution. It not only prevents gradient explosion but also prevents the symmetry of weights.

10. You might find it more difficult to differentiate between non-covid and covid x-rays, rather than between normal x-rays and infected (both covid and non-covid) people x-rays. Was that something to be expected? Discuss.

    Yes. It is very easy for us to distinguish the normal and infected x-rays. We just have to check whether there is a blur region in the lungs, but it is very hard to differentiate between covid and non-covid. The same logic can be applied to the model, there is more difference between normal and infected and less between covid and non-covid. Therefore, it will be harder to differentiate covid and non-covid than normal and infected.

11. Final question: would it be better to have a model with high overall accuracy or low true negatives/false positives rates on certain classes? Discuss.

It will be good to have high true negatives/false positives rates on covid and non-covid as it might be just a bit of a burden for normal to be classified as infected, but for covid being classified as normal might lead to a huge infection situation.