Approach towards problem statement

Before starting the project need to keep one thing in mind it may be heavy on computation side so that's why in my code I have changed the runtime to GPU on Google Colab.

Problem statement: Dataset contains 1252 CT scans that are positive for SARS-CoV-2 infection (COVID-19) and 1230 CT scans for patients non-infected by SARS-CoV-2, 2482 CT scans in total. These data have been collected from real patients in hospitals from Sao Paulo, Brazil. The aim is to identify if a person is infected by SARS-CoV-2 through the analysis of his/her CT scans.

According to problem statement, it contains 1252 CT scan images for covid positive cases and 1230 images for negative positive cases,

Hence data was uploaded directly from google drive to colab.

First point to keep in mind is dataset should not be imbalanced which is checked by counting the no. of images.

Later on, data frame is created which contains following attributes file (images name and path), disease id 0 represents covid positive and 1 represents negative along with disease type.

Just to verify covid positive and covid negative, images are plotted using matplotlib.pyplot,

As mentioned in problem statement there is a need to resize the images because there are lots of images having different configuration, standard size for image after resizing (64, 64, 3) where 3 are number of channels for RGB.

Now model creation started, first of all through sklearn.train_test_split() is used to split the dataframe into training and validation, whereas 20% of data is as validation.

For image classification we are using CNN- Convulational Neural Network approach, where as we are going to use Resnet50 for transfer learning, as we know resnet50 are one of the most accurate neural network to work on image classification.

For hidden layers "relu" as acitivation function is used, for output layer it is "softmax", Adam is used as optimizer and categorical-crossentropy is used as loss function.

According to problem statement Data Augmentation is done by using following parameters:

- Width-shift-range for random horizontal shifts.
- Height-shift-range for random vertical shifts.
- Zoom range, it is the range to zoom on images.
- Horizontal and vertical flips are also used.

Checkpoint at “model.h5” is mentioned along with early stopping, and model has been compiled and model gets fit where 500 epochs are mentioned and early stopping takes place,

Later on after training the model need to predict some result hence random image is taken and checked the probability of result.

Hence at last multiple metrics are used to get accuracy between true value and predicted value.

As it is a classification problem hence confusion matrix is used, where one can see correctly predicted values on diagonals.

Accuracy score between true and predicted value is also mentioned,

Later on some graphs has been plotted under model accuracy, first one is between accuracy and epochs second one is between loss and epochs.

At last classification report which contains various parameters to judge about model such as f1 score or performance metrics, precision and recall.