On-Device Deep Learning Assignment #3

Name: Swapneel Dhananjay Wagholikar

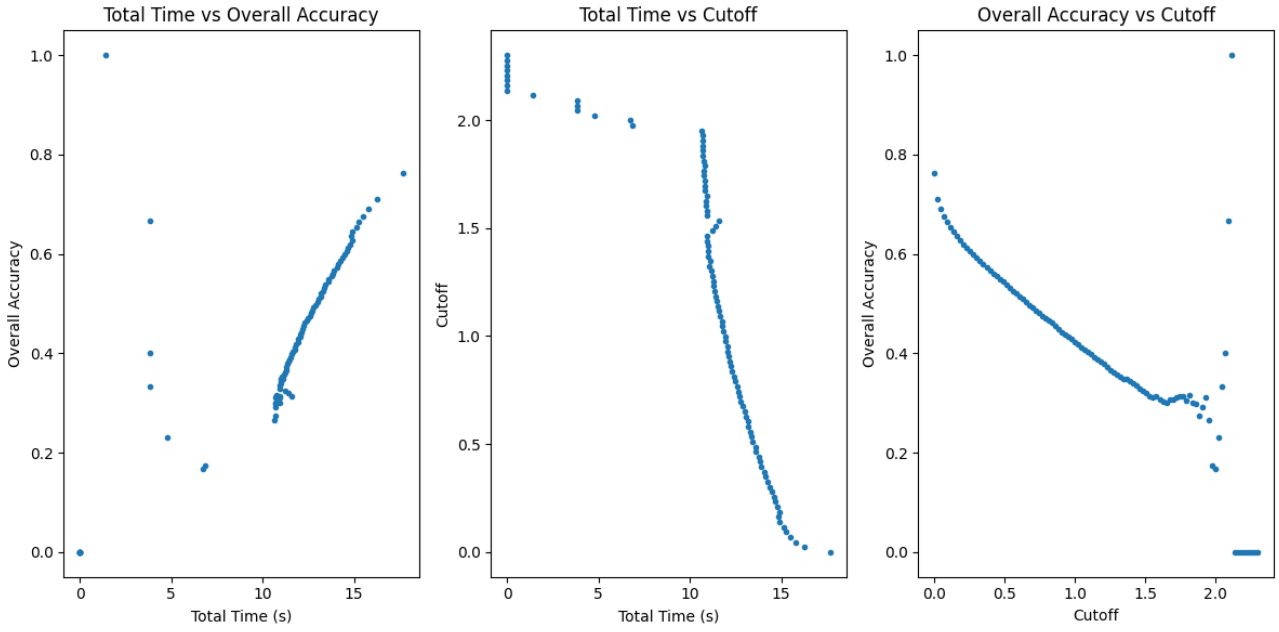
WPI ID: 257598983

Assignment3: Dynamic Network Inference

1a. As required, I calculated the accuracy and inference time for each layer at c=0.6. After that is done, overall accuracy and total inference time is taken as output from cutoff exit performance check function. Hereby attaching snapshot of the output terminal:

```
---------------------------------- Task 1a ----------------------------------
Exit Layer 0: Accuracy=0.4387911247130834, Inference Time=1.708237648010254 seconds
Exit Layer 1: Accuracy=0.545766590389016, Inference Time=1.7666170597076416 seconds
Exit Layer 2: Accuracy=0.5305280528052805, Inference Time=2.040156126022339 seconds
Exit Layer 3: Accuracy=0.5652901785714286, Inference Time=2.509432554244995 seconds
Exit Layer 4: Accuracy=0.5095057034220533, Inference Time=2.655341863632202 seconds
Exit Layer 5: Accuracy=0.5207803223070399, Inference Time=2.904303789138794 seconds
overall_accuracy 0.5149470899470899
total_time 13.584089040756226
```

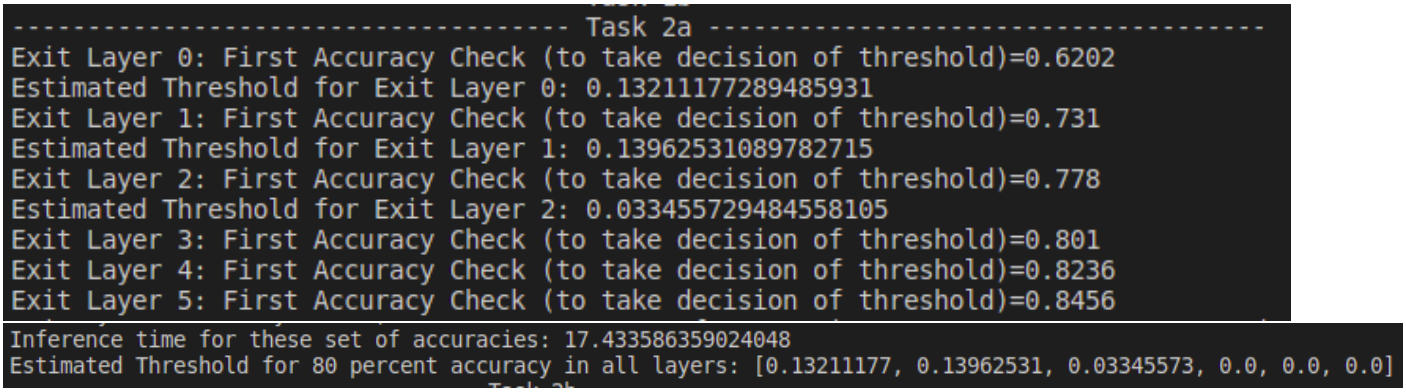
1b. In this part, all the 3 required graphs are plotted, and we can see the corresponding relations through that.



As we can observe, there is a tradeoff between overall accuracy and total inference time, and both have inverse correlation. So, we need to decide on our application that we need to give importance to accuracy or inference time and then need to decide the threshold value accordingly. Also, I came up with one metric that I used in part 2b of the assignment. We want the overall accuracy to be more and inference time to be less, so I designed a metric that is a ratio of overall accuracy

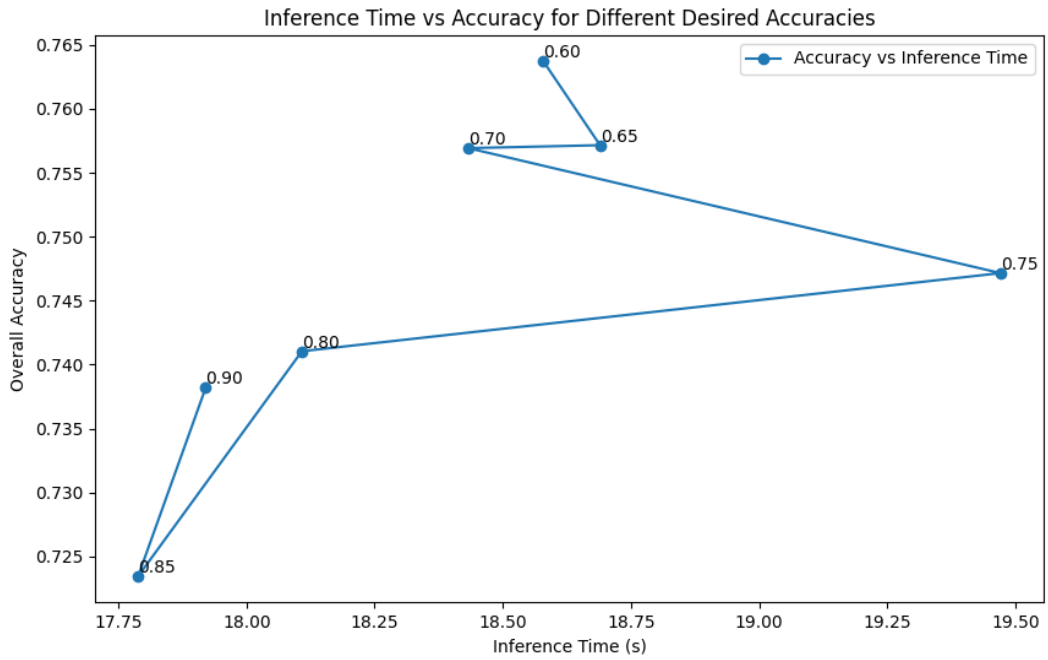
and total time. We can calculate this metric for all overall accuracy and total time points which we have and then check for which point we get this metric to be maximum and thus, we can select the best threshold.

2a. In this part, I have set the desired accuracy to 80%. After that, I estimated the threshold for each exit layer. For the first 3, I got some value of threshold, but for the last 3, accuracy was more than 80% for most of the samples (almost all the samples) and hence, we got the value of threshold as 0.0. Here is the snapshot of output terminal:



As you can see from the result, my code tried to update the threshold value for the first 3 layers and tried making it 80%. For this, I used the top 20% entropies and tried building a relation of entropy with desired accuracy. The results can be better if we get a more accurate relation to calculate this threshold.

2b. In this part, I have considered varying the 7 accuracy values [0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9] and overall accuracy is plotted against inference time. In my plot, we can infer about the relationship between overall accuracy and inference time and then I used the metric I explained in 1b here to find the best threshold value and then finally calculated the accuracy and inference time on test data.



Using the metric explained above, I have found the best threshold value and overall accuracy and inference time on Test data for this threshold value. Here is the snapshot of output terminal:

```
Best Thresholds: [0.03273254, 0.044205364, 0.007976418, 0.008089032, 0.001157964, 0.0020747662]
Test Accuracy using the Best Threshold: 0.7381940840685003
Inference Time on Test Data using the Best Threshold: 17.643163919448853 seconds
```

Final Test accuracy received is 73.81% and Test inference time is 17.64 seconds.