Problem Statement: Entropy-based early exit on BrancyNet

In this assignment, we will test the performance of an Early Exit method. We will provide a pre-trained BrancyNet model that is trained on CIFAR-10 data. In this code, we can exit from five different layers and one final output layer. We will exit from different layers according to a certain threshold. We will make this decision based on entropy. Note that certain portions of the data will exit each layer that will be decided according to the threshold we set.

Task 1:

a) First, look at the model and understand how the inputs and outputs are taken from each layer. Let's assume we have a fixed threshold (cutoff), ‘c”. Calculate the entropy for each sample in a batch after each layer and decide whether we want to exit that sample from a certain layer. Remove the samples that exited from a layer from the pool of samples in the batch when moving to the next layer. Calculate accuracy and inference time at each layer for c = 0.6.

b) Plot total time vs overall accuracy, total time vs cutoff, and overall accuracy vs cutoff for 100 different ‘c’ ranging from the minimum to the maximum possible entropy (hint: the maximum is not 1). Describe which threshold you think the model works best in terms of both inference time and accuracy. Note that, for accuracy calculation, you need to take a weighted mean of the accuracy scores for each layer output (weighted according to the number of samples that exited after each layer). Also, for inference time calculation, you need to calculate the total time needed by all layers for the inference to be completed.

Task 2:

a) Let’s think we wish to have a minimum accuracy in each layer if we wish to exit from that layer. For this task, we might think that we want 80% accuracy in each layer before exit. Additionally, assume that in all exit layers, the threshold is different. Your first task is to estimate the threshold using the validation data. Also, calculate the inference time.
*Hint: You should calculate the entropy on the validation data (not testing data) to get the threshold.

b) Now, let’s vary the desired minimum accuracy on the training data. Get the threshold for each accuracy and calculate the inference time. Plot inference time vs accuracy and find out the best threshold. Using the best threshold, find the accuracy and inference time on test data.