HW2 Model Compression

Torch Summary:

```
       BatchNorm2d-68             [-1, 39, 7, 7]              78
              ReLU-69             [-1, 39, 7, 7]               0
        BasicBlock-70             [-1, 39, 7, 7]               0
        AvgPool2d-71             [-1, 39, 1, 1]               0
           Linear-72                   [-1, 10]             400
================================================================
Total params: 99,603
Trainable params: 99,603
Non-trainable params: 0
----------------------------------------------------------------
Input size (MB): 0.00
Forward/backward pass size (MB): 2.28
Params size (MB): 0.38
Estimated Total Size (MB): 2.66
----------------------------------------------------------------
100%|
Accuracy of the network on the 10000 test images: 94.75 %
```

Compression Method:

1. **Create a much smaller ResNet-like model as a student network.**
2. **Use the given resnet-50 model as the teacher network.**
3. **Distill the knowledge from the teacher to the student, with KL divergence loss.**
4. **Afterwards, prune the student model based on the group norm importance. The model is pruned iteratively for 30 steps, and fine tuning is done after each step.**
5. **Random cropping, horizontal flip, and erasing are used during the training stage for both knowledge distillation and pruning.**

References:

Student model & Random Erasing method:

Paper: [1708.04896] Random Erasing Data Augmentation (arxiv.org)

Code: GitHub - zhunzhong07/Random-Erasing: Random Erasing Data Augmentation. Experiments on CIFAR10, CIFAR100 and Fashion-MNIST

KL Divergence Loss:

Code: GitHub - haitongli/knowledge-distillation-pytorch: A PyTorch implementation for exploring deep and shallow knowledge distillation (KD) experiments with flexibility

Pruning Method:

Paper: [2301.12900] DepGraph: Towards Any Structural Pruning (arxiv.org)

Code: GitHub - VainF/Torch-Pruning: [CVPR-2023] Towards Any Structural Pruning