Advanced Deep Learning *ECS 289L, Spring 2024*

HW2: Finetuning Vision Transformer (ViT)

This homework aims to familiarize you with fine-tuning deep learning models for downstream tasks. We will use the CIFAR-10 dataset to train and fine-tune a Vision Transformer (ViT) model. To expedite the training process, we will reduce the training samples of CIFAR-10 to 1000 per class. You will need to install the timm library (version 0.5.4). You can use Google Colab with T4 GPUs for this homework. Clone this repository and use the the *train.ipynb* notebook to train your models. Please read the following paper for details of ViT architecture: Vision Transformer. For each of the following parts, please train the model for 5 epochs.

1. Train a ViT-base model from scratch for 5 epochs and report both the loss and Top-1 Accuracy on CIFAR-10 train and validation sets at every epoch or several iterations. Use the default parameters for training from scratch. This should take less than 15 minutes on Google Colab.

2. Fine-tuning Pretrained ViT: Repeat the previous part, but initialize the model from a pretrained model on the ImageNet dataset. The code is set up to download the pretrained model by simply passing an argument. Note that the pretrained linear layer of the ImageNet model has 1000 classes, so it is a linear layer (the last layer or the classifier) of size 192×1000, so you need to replace it with a new linear layer for only 10 classes. Finetune the whole model without freezing any layer. Compare the training accuracy and loss with those obtained from training from scratch. Justify your observations in a few sentences.

3. Suppose we want to reduce the inference time to deploy the ViT model on a small device, e.g., an iPhone. One solution is to train a smaller model; however, there is a trade-off between accuracy and model size. First, finetune *vit_tiny* on CIFAR-10 and compare the accuracy and throughput to *vit_base*. Throughput is measured as the number of images per second processed at the test time.

4. Knowledge Distillation (KD) is a well-known method to improve the smaller model. In KD, we train a large model (teacher, e.g., *vit_base*) first in the standard way, and then use it to generate pseudo-groundtruth in training a smaller model (student, e.g., *vit_tiny*). Please use the model from part (2) as the teacher and distill it to a *vit_tiny* pretrained on ImageNet. You will finetune the whole student model in this case without freezing any layer. Compare the accuracy of the student model trained with the teacher model or from scratch.