Assignment 4 Finetuning Pretrained Multimodal Models on Vision Tasks (100 Pts)
Due Date: Tuesday 12/12 11:59PM EST

In this assignment, we will use pretrained CLIP (Contrastive Language-Image Pretraining) to perform zero-shot prediction and linear-probe evaluation on Intel Image Classification dataset. Note that the dataset has been used in our assignment 1 and can be easily downloaded as a zip file from the assignment 1 link. In particular, the primary steps of this assignment are provided as follows:

- Define the dataset in PyTorch. Intel Image Classification dataset is not built into torchvision.datasets. As we discussed in class, you must make sure that your own dataset class is compatible with the built-in one (just like CIFAR-100).
- **Perform zero-shot prediction using CLIP-pretrained vision backbone**. The detailed code snippet can be found on the official repository of CLIP. Make sure that you compute the similarity score as well as the ***TOP 10*** most similar labels.
- **Perform linear probe evaluation using CLIP-pretrained vision backbone**. Similar to the previous part, you can find the concrete example on the official repository of CLIP. Double check the GPU RAM and choose a backbone that can fit into the GPU memory. Use ***support vector machine (SVM)*** as the classifier.

In your submission, ***please include the printed results in the notebook or in a separate text file in case you use Python scripts***. Missing these outputs will result in significant deduction in grades.