DLCV 2024: Assignment 2
Deadline: 29th March

Instructions

*1. Your submission should be a zip file containing the codes (.py or .ipynb), readme.txt file, and a **report in pdf format (max 4 pages)**.*
2. Please include comments in your code. The readme.txt file should contain information on the organization of your files, packages used along with their versions, python version, and instructions on running the code.
3. While you are encouraged to go through online repositories to learn best practices and tricks, please avoid directly copying from somewhere.
4. Please submit a report on your observations, results, plots, and analysis in pdf format. This assignment carries 25% weightage for code and 75% weightage for your analysis in the report.

[**Transformer networks**]

In this question, you will implement a vision transformer-based image classification model using pytorch.

a) Implement a basic version of the vision transformer (https://arxiv.org/pdf/2010.11929.pdf), which divides an image into patches and then passes them through a set of multihead self-attention modules to perform classification. Please check out the details in the paper.
b) **[Experiment 1]** Train this model on the CIFAR-10 dataset for 10-class classification. Keep the number of attention heads to 4 for all the experiments.
c) **[Experiment 2]** Train your model at different data sizes 5%, 10%, 25%, 50% and 100% of the training dataset and discuss model performance.
d) **[Experiment 3]** Try different patch sizes (like 4x4, 8x8, 16x16). You can divide the image into both overlapping and non-overlapping patches.
e) **[Experiment 4]** Report model performance by varying the number of attention heads.
f) **[Experiment 5]** Classify the model using the CLS token from different model layers.
g) **[Experiment 6]** Take 2 test images per class, classify them, and visualize attention maps across the trained transformer layers.

Create a detailed report of all the experiments and analyses.