Mini-project n° 3 – ViT vs CNN

The objective of this project is to compare the performances of two different architectures for the task of image classification: Vision Transformers (ViT) and Convolutional Neural Networks (CNN).

- Choose two or more labelled datasets for the classification task. Motivate your choices.
- Make sure to explain the ViT architecture in detail.
- Choose one or multiple ViT and CNN architectures. Motivate your choices.
- You are expected to code the models yourselves, using the usual pytorch modules (Conv2d, MaxPool2d, Linear, attention modules...), i.e. there is no need to code the attention mechanism or the convolution mechanism.
- One of the final results of the projects should be a summary table comparing the performances of the different models on different datasets.
- Draw conclusions on weather the choice of a CNN or a ViT would be influenced by factors other than performance : the dataset, the computational budget, training and inference time, the memory budget, the interpretability of the models etc.
- How do your conclusions compare to those in the literature ?

BONUS: Are there any explainability techniques that give interpretable explanations on these models ? Are the explanations different in the ViT and the CNN models ?