Indian Institute of Technology Hyderabad
Deep Learning (AI5100): Assignment-3

Topic: Self Attention in computer vision
Assigned on: 11^{th} **April, 2024**
Deadline: 25^{th} **April, 2024**
Maximum Marks: 40

1 Instructions

- Answer all questions. We encourage best coding practices by not penalizing (i.e. you may not get full marks if you make it difficult for us to understand. Hence, use intuitive names for the variables, and comment your code liberally. You may use the text cells in the notebook for briefly explaining the objective of a code cell.)
- It is **expected** that you work on these problems individually. If you have any doubts please contact the TA or the instructor no later than 2 days prior to the deadline.
- You may use built-in implementations only for the basic functions such as `sqrt, log`, etc. from libraries such as `numpy` or `PyTorch`. Other high-level functionalities are expected to be implemented by the students. (Individual problem statements will make this clear. We can use the optimizers provided by the libraries such as PyTorch.)
- For plots, you may use `matplotlib` and generate clear plots that are complete and easy to understand.
- You are expected to submit the Python Notebooks saved as <your-roll-number>.ipynb
- If you are asked to report your observations, use the mark down text cells in the notebook.
- Late submission policy: < 1 day delay $\rightarrow$ 10% penalty, $\{> 1$, but $< 2\}$ days delay $\rightarrow$ 25% penalty, $\{> 2$, but $< 3\}$ days delay $\rightarrow$ 50% penalty, $\{> 3$, but $< 4\}$ days delay $\rightarrow$ 75% penalty, and for a submission beyond 4 days, there won't be any reward.

2 Questions

1. **Self-Attention for Object Recognition with CNNs**: Implement a sample CNN with one or more self-attention layer(s) for performing object recognition over CIFAR-10 dataset. You have to implement the self-attention layer yourself and use it in the forward function defined by you. All other layers (fully connected, nonlinearity, conv layer, etc.) can be bulit-in implementations. The network can be a simpler one (e.g., it may have 1x Conv, 4x [Conv followed by SA], 1x Conv, and 1x GAP). Please refer to the reading material provided here or any other similar one. [10 Marks]
2. **Object Recognition with Vision Transformer**: Implement and train an Encoder only Transformer (ViT-like) for the above object recognition task. In other words, implement multi-headed self-attention for the image classification (i.e., appending a $< class >$ token to the image patches that are accepted as input tokens). Compare the performance of the two implementations (try to keep the number of parameters to be comparable and use the same amount of training and testing data). [10 Marks]