RDFIA : Pattern Recognition and Machine Learning for Image Understanding

Deep Learning Practical Work 1-e

Transformers

Srijita NANDI, Seydina KANDE
M2 Quantum Information

1 Introduction:

In this notebook, we focus on implementing a simplified version of the Vision Transformer (ViT) model, a powerful approach in computer vision that relies on transformer architecture. Traditionally used in Natural Language Processing (NLP), transformers excel at capturing relationships across sequences of data through self-attention mechanisms. Vision Transformers adapt this model to images by treating image patches as input tokens, allowing them to capture spatial dependencies between different regions of an image.

The ViT model divides images into fixed-size patches, projects these patches into embeddings, and then feeds them through transformer layers that perform self-attention across patches. Key components in this architecture include patch embedding (converting image patches into vector representations), positional encoding (retaining spatial information), and a classification token (CLS), which summarizes the entire image for classification tasks.

In this notebook, we implement a simplified ViT to explore these architectural elements without the complexities of large datasets, extensive data augmentation, or advanced regularization techniques, using the MNIST dataset for clarity and ease of experimentation. Additionally, we utilize the Timm library to compare our implementation against pre-built models, investigate challenges with transformer input formats, and analyze model performance. By building this basic version, we aim to understand the potential of ViT and its adaptability in visual tasks.

2 Notebook Questions and experimental results explanation

Done in the python notebook.

3 Conclusion:

In this notebook, we implemented a simplified version of the Vision Transformer (ViT) model and applied it to the MNIST dataset. Through this hands-on approach, we observed how a ViT processes image patches as tokens, leveraging self-attention to capture relationships across these patches. Our model demonstrated [specific results, such as accuracy and loss on test data – insert specific results here], showcasing how well a basic ViT can generalize to simple image data without extensive pretraining or augmentation.

Through experimentation, we learned about the importance of positional encoding for maintaining spatial coherence and observed the flexibility and limitations of transformer-based models in handling grayscale versus RGB images. The exercise also highlighted the computational demands of transformers, especially with an increasing number of tokens, and the potential ways to address these challenges, such as reducing patch numbers or exploring sparse attention.

Overall, this notebook provided a strong foundation in understanding how transformers can be adapted from NLP to vision tasks and the implications of their architecture on model performance. This exploration sets the stage for future work, such as applying ViT to more complex datasets or fine-tuning larger pretrained models, which can further leverage the strengths of self-attention in computer vision.