Deep Learning: Homework 2

Hugo Mantinhas 95592, João Silveira 95597

December 28, 2023

Work Division

Each member of the group worked together on all the questions in this assignment. For this reason, the workload was divided equally between the two members of the group.

Question 1

1.1

The computational complexity of the self-attention layer in a transformer with a single attention head is $O(L^2 * D)$, where L is the sequence length and D is the hidden size. Let's break down that computation:

- **Matrix multiplication of Q and K^T:** This operation has complexity $O(L^2 * D)$, since Q is of size $L \times D$ and K^T is of size $D \times L$.
- **Softmax:** This operation has complexity $O(L^2)$, since it is applied to a matrix of size $L \times L$.
- **Matrix multiplication of the result of the previous operation and V:** This operation has complexity $O(L^2 * D)$, since the result of the previous operation is of size $L \times L$ and V is of size $L \times D$.

Therefore, the overall complexity is therefore $O(L^2 * D)$.

This complexity can be problematic for long sequences due to its quadratic dependence on the sequence length L. As the sequence length increases, the computational cost grows quadratically. This makes it computationally expensive and impractical for very long sequences. To address this issue, various techniques such as attention prunin, approximations, or using specialized hardware have been explores to make transformer models more scalable to longer sequences.