TUM EI 70360: Machine Learning and Optimization
Fall 2023
Lecturer: Reinhard Heckel
Teaching Assistant: Tobit Klug

Problem Set 12
Issued: Tuesday, Jan. 16, 2024
Due: Thursday, Jan. 25, 2024

Problem 1 (Number of parameters and compute of the GPT model)**.** Consider the decoder-only transformer considered in class with the following parameters. For all tasks omit sub-leading terms such as non-linearities, biases, and normalization.

`vocab_size`: size of the vocabulary
`dim_embd`: dimension or size of the embeddings
`n_heads`: number of heads
`n_layers`: number of layers
`context_length`: the (maximal) context length, that is the maximal number of input tokens

1. Compute the total number of non-embedding trainable parameters N of the transformer, i.e., the total number of trainable parameters excluding the number of parameters associated with embeddings.
2. Now compute the total number of non-embedding compute in floating point operations (FLOPs) per tokens for a forward pass. Make the following assumption: The multiplication of two matrices $\mathbf{A}$ and $\mathbf{B}$, where $\mathbf{A}$ is $i \times j$ and $\mathbf{B}$ is $j \times k$ gives a matrix $\mathbf{C}$ of dimension $i \times k$. For each element of $\mathbf{C}$, you perform j multiply-accumulate operations (since you're multiplying and then summing up j pairs of elements). Each of these involves one multiplication and one addition. Therefore, for each element of $\mathbf{C}$, you have $2j$ FLOPs. Extending this to the entire matrix $\mathbf{C}$ (which has ik elements), the total number of FLOPs is $2ijk$.
3. Assume that the cost of a backward pass is roughly two times a forward pass, and assume that we have a transformer in the regime `dim_embd` $>>$ `context_length`$/12$. Justify the common statement 'the non-embedding compute per token for training is roughly $6N$'.