Exercise Session 5

Attention Transformers and Tokenization

Prepared by Abdulkadir Gokce and Xiuying Wei

Overview

Task 1. Computational Cost of Multi-Head Self-Attention.

Consider a single multi-head attention block in a transformer encoder with the following parameters:

$$\text{sequence length } n, \quad \text{model dimension } d_{\text{model}}, \quad \text{number of heads } h.$$

Each head uses key-, query-, and value-projections of dimension $d_k = d_v = d_{\text{model}}/h$. Assume that the input to the attention block is a tensor of shape $(1, n, d_{\text{model}})$ (assume batch size of 1), and that the multi-head attention consists of:

- Three linear projections $\mathbf{W}^Q$, $\mathbf{W}^K$, $\mathbf{W}^V \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$ (no bias),
- Reshape/split into h heads of dimension d_k each,
- Scaled dot-product + softmax attention per head,
- Concatenation of the h head outputs,
- Final output projection $\mathbf{W}^O \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$ (no bias).

Calculate the following quatities:

1. **Number of learnable parameters:** Estimate the total number of floating-point parameters that must be stored for this single multi-head attention block (ignore any biases). Express your answer in terms of d_{model} and h.
2. **FLOPs:** Count the total number of floating-point operations (one multiply or one add counts as one FLOP) required to compute the forward pass of this attention block. For simplicity, you may:
   - Count each $m \times k$ by $k \times p$ matrix-multiply as $2\,m\,k\,p$ FLOPs.
   - Count each element-wise addition or multiplication outside of the matrix-multiplies (e.g., scaling by $1/\sqrt{d_k}$, softmax exponentials and divisions) approximately as $O(n\,n\,h)$ FLOPs, and you may omit lower-order terms.

   Express your answer in terms of n, d_{model}, and h.

3. **Efficiency bottlenecks (FLOPs aspect):** Building on the FLOPs above, discuss the efficiency bottlenecks that arise when processing the whole sequence. Elaborate on factors that can slow the attention module. **Hint**: Consider the heavy components—linear projections and the scaled dot-product attention—and explain when each becomes expensive as a function of the sequence length n, model dimension d_{model}, and number of heads h.
