Exercise - 3: Transformers

December 2, 2025

Questions

Q1. Consider a sequence of length 3. The query, key and value matrices (each row corresponds to a token) are

$$Q = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}, \qquad K = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{bmatrix}, \qquad V = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}.$$

Use the scaled dot-product attention formula with $d_k = 2$. Compute by hand (and give numerical answers):

(a) the score matrix $S = QK^\top$,

(b) the scaled scores $S/\sqrt{d_k}$,

(c) the row-wise softmax matrix $A = \text{softmax}(S/\sqrt{d_k})$,

(d) the final output $O = AV$.

Q2. Suppose you have two attention heads ($h = 2$) producing the following per-head outputs for one token:

$$\text{Head}_1 = [0.2, 0.5, 0.3], \quad \text{Head}_2 = [0.6, 0.1, 0.3].$$

(a) Write the concatenated vector before the output projection.

(b) If the output projection matrix W_O is

$$W_O = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix},$$

compute the projected output.

Q3. Why is masked attention used?

Hint

$$\text{Attention}(Q, K, V) = \text{softmax}\Big(\frac{QK^\top}{\sqrt{d_k}}\Big)V$$

Instructions

- Show all necessary steps clearly.