The Science and Engineering of LLMs

Assignment 1: The Transformer Architecture and Attention Mechanism

April 1, 2025

Part 1: Analyzing the Design of Transformer

1. Scaled Dot-Product Attention.

The output of Transformer's self-attention is computed as follows:

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

Where d_k is each key vector's dimension.

1. Assume that q and k are both d_k-dimensional vectors, and each component of q and k are independent variables with mean 0 and variance 1. Prove that $Var\left(\frac{q \cdot k}{\sqrt{d_k}}\right) = 1$.
2. In self-attention, we use different matrices to transform the representation of tokens into Q, K and V. Now, suppose $Q = K$. From the characteristics of matrix $QK^\top$, specify at least two consequences of this.

2. Concatenation vs. Summation of Embeddings.

Different from the token ordering in RNNs, transformers naturally lack the ability to represent token positions, so one needs to use positional encoding or positional embedding as a compensation. Assume that we have a token embedding $e_i \in \mathbb{R}^{d_e}$, its corresponding positional encoding vector $p_i \in \mathbb{R}^{d_e}$, and two transform matrices $W_e \in \mathbb{R}^{d_h \times d_e}$ and $W_p \in \mathbb{R}^{d_h \times d_e}$. If we denote concatenation on the first dimension as $[;]$, please express $[W_e^\top; W_p^\top]^\top [e_i; p_i]$ by a summation of two terms relating to e_i and p_i, respectively.