

Exercise 1 - Dot-Product Attention

You are given a set of vectors

$$\mathbf{h}_1 = (1, 2, 3)^\top, \quad \mathbf{h}_2 = (1, 2, 1)^\top, \quad \mathbf{h}_3 = (0, 1, -1)^\top$$

and an alignment source vector $\mathbf{s} = (1, 2, 1)^\top$. Compute the resulting dot-product attention weights α_i for $i = 1, 2, 3$ and the resulting context vector \mathbf{c} .

Exercise 2 - Attention in Transformers

Transformers use a scaled dot product attention mechanism given by

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

where $Q \in \mathbb{R}^{n_q \times d_k}$, $K \in \mathbb{R}^{n_k \times d_k}$, $V \in \mathbb{R}^{n_k \times d_v}$.

- Is the softmax function here applied row-wise or column-wise? What is the shape of the result?
- What is the value of d ? Why is it needed?
- What is the computational complexity of this attention mechanism? How many additions and multiplications are required? Assume the canonical matrix multiplication and not counting $\exp(x)$ towards computational cost.
- In the masked variant of the module, a masking matrix is added before the softmax function is applied. What are its values and its shape? For simplicity, assume $n_q = n_k$.

Exercise 3 - Scaled Dot-Product Attention by Hand

Consider the matrices Q , K , V given by

$$Q = \begin{bmatrix} 1 & 2 \\ 3 & 1 \end{bmatrix}, \quad K = \begin{bmatrix} 2 & 1 \\ 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad V = \begin{bmatrix} 1 & 2 & -2 \\ 1 & 1 & 2 \\ 0 & 1 & -1 \end{bmatrix}.$$

Compute the context matrix C using the scaled dot product attention.