Transformers

Embeddings

Suppose we have a simple time series with 4 values

$$\mathbf{y} = [-2, 2, 0, 1]$$

at times

$$t = [1, 2, 3, 4].$$

We choose an embedding dimension d and create weights W and biases b. Suppose $d = 3$.

Define

$$X = \begin{bmatrix} y & t \end{bmatrix},$$

so we have $F = 2$ features (including time).

Thus, W is $F \times d = 2 \times 3$, and b is $d \times 1$.

Suppose

$$W = \begin{bmatrix} 2 & 0 & -1 \\ 1 & -2 & 0 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}.$$

The embedding is

$$E = XW + \mathbf{1}b^T.$$

Positional Encoding

We create positional encoding using

$$\sin\left(\frac{\text{position}}{10000^{2i/d}}\right), \quad \cos\left(\frac{\text{position}}{10000^{2i/d}}\right).$$

For $d = 3$, we use

$$P = \begin{bmatrix} \sin 1 & \cos 1 & 0 \\ \sin 2 & \cos 2 & 0 \\ \sin 3 & \cos 3 & 0 \\ \sin 4 & \cos 4 & 0 \end{bmatrix}.$$

The position-aware embedding is

$$Z = P + E.$$

Self-Attention

Steps:

1. Create queries $Q = ZW_q + \mathbf{1}b_q^T$.
2. Create keys $K = ZW_k + \mathbf{1}b_k^T$.

3. Compute dot products: $S = \frac{QK^T}{\sqrt{d}}$.
4. Apply softmax row-wise to get attention weights A.
5. Create values $V = ZW_v + \mathbf{1}b_v^T$.
6. Compute outputs $O = AV$.

Attention Block Diagram