Numerical on Attention is all you need

Input + Positional Encoding

- Word embeddings (for simplicity, chosen small numbers):
  - **Hi → [1, 0]**
  - **How → [0, 1]**
- Positional encoding (toy example, just add [0.1, 0.1] for 1st word, [0.2, 0.2] for 2nd):
  - **Hi → [1.1, 0.1]**
  - **How → [0.2, 1.2]**
- These embeddings will be copied into **Query (Q), Key (K), and Value (V).**

Create Q, K, V

- We multiply each embedding with small weight matrices (assume fixed ones for simplicity):
  - $Q = Embedding \times W_Q$
  - $K = Embedding \times W_K$
  - $V = Embedding \times W_V$

For easy numbers, let's assume weight matrices are identity (so Q=K=V = embedding).

So:

- Hi → Q=[1.1, 0.1], K=[1.1, 0.1], V=[1.1, 0.1]
- How → Q=[0.2, 1.2], K=[0.2, 1.2], V=[0.2, 1.2]

Compute Attention Scores ($Q \cdot K^T$)

Dot product of Query (Q) of a word with Key (K) of all words:

For **Hi** (Q=[1.1, 0.1]):

- Score(Hi→Hi) = 1.1×1.1 + 0.1×0.1 = **1.22**
- Score(Hi→How) = 1.1×0.2 + 0.1×1.2 = **0.34**

For **How** (Q=[0.2, 1.2]):

- Score(How→Hi) = 0.2×1.1 + 1.2×0.1 = **0.34**
- Score(How→How) = 0.2×0.2 + 1.2×1.2 = **1.48**

Scale Scores

- Scale by $\sqrt{(\text{dimension})} = \sqrt{2} \approx 1.41$

So:

- Hi→Hi = 1.22 / 1.41 ≈ **0.86**
- Hi→How = 0.34 / 1.41 ≈ **0.24**
- How→Hi = 0.34 / 1.41 ≈ **0.24**
- How→How = 1.48 / 1.41 ≈ **1.05**

Apply Softmax (Get Weights)

For **Hi row [0.86, 0.24]**:

- exp(0.86)=2.36, exp(0.24)=1.27 → Softmax = [0.65, 0.35]

For **How row [0.24, 1.05]**:

- exp(0.24)=1.27, exp(1.05)=2.86 → Softmax = [0.31, 0.69]

So weights =

- Hi pays **65% attention to itself, 35% to How**
- How pays **31% attention to Hi, 69% to itself**

Weighted Sum with Values (Final Output)

Multiply weights with Value vectors:

For **Hi output**:

$= 0.65 \times [1.1, 0.1] + 0.35 \times [0.2, 1.2]$

$= [0.65 \times 1.1 + 0.35 \times 0.2, 0.65 \times 0.1 + 0.35 \times 1.2]$

$= [0.91, 0.47]$

For **How output**:

$= 0.31 \times [1.1, 0.1] + 0.69 \times [0.2, 1.2]$

$= [0.48, 0.87]$

Intuition

The word **Hi** still keeps most of its meaning (0.91, 0.47) but borrows some context from **How**.

The word **How** is enriched (0.48, 0.87) by mixing with **Hi**.

This mixing **= Self-Attention** → words become context-aware.