Attention

Traditional sequence models like RNNs and LSTMs suffer from limitations:

- Difficulty in handling long sequences due to vanishing gradients.
- Fixed-length context vectors make it hard to capture all relevant information.
- They process input sequentially, which is computationally inefficient.

Attention mechanisms address these problems by allowing the model to look at **all positions in the input sequence** and **selectively focus** on the most relevant parts.

3. What is Attention Mechanism?

In simple terms, **attention** is a process of assigning **weights** to different parts of the input based on their relevance to a specific task (e.g., predicting the next word in a sentence).

3.1 Core Components

The standard attention mechanism involves three main components:

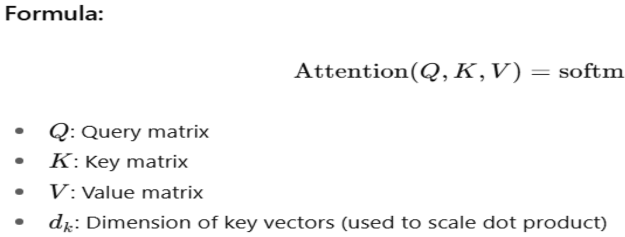
- **Query (Q)**: Represents the current position we're focusing on.
- **Key (K)**: Represents all possible input positions.
- **Value (V)**: The actual content at each input position.

The attention score between a query and key is calculated using a similarity function (often dot product), and the result is used to weight the corresponding value.

4. Scaled Dot-Product Attention

The most common type of attention is **Scaled Dot-Product Attention**, introduced in the Transformer model.

Formula:



Formula:

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

- Q: Query matrix
- K: Key matrix
- V: Value matrix
- d_k: Dimension of key vectors (used to scale dot product)

Step-by-Step Process:

1. Compute dot products between the query and all keys: QKTQK^TQKT
2. Scale the results by dk\sqrt{d_k}dk to avoid large values.
3. Apply softmax to get normalized attention weights.
4. Multiply the weights with values VVV to get the output.

5. Example of Attention Mechanism

Let’s consider a machine translation example:

Input Sentence (English):
`"The cat sat on the mat"`

Target Output (French):
`"Le chat était assis sur le tapis"`

When translating the word `"chat"` (French for "cat"), the model should focus more on the word `"cat"` in the English sentence.

Using attention, the model calculates alignment scores between `"chat"` and all English words:

- Attention score with `"cat"` → **high**
- Attention score with `"The"`, `"sat"`, `"mat"` → **low**

The output for `"chat"` is mostly influenced by the representation of `"cat"` due to its high attention weight.

6. Types of Attention Mechanisms

6.1 Soft vs. Hard Attention

- **Soft Attention**: Differentiable, assigns probabilities to all input positions (used in most deep learning models).
- **Hard Attention**: Non-differentiable, chooses a single input position (requires reinforcement learning).

6.2 Self-Attention

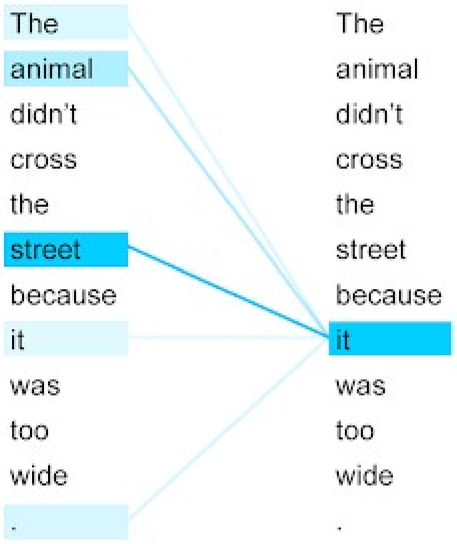
- Each word attends to **all other words** in the same sequence.
- Core part of Transformer architecture.

6.3 Multi-Head Attention

- Runs multiple attention mechanisms in parallel.
- Allows the model to learn different types of relationships simultaneously.

7. Advantages of Attention Mechanisms

- **Handles long-range dependencies** better than RNNs.
- **Parallelizable** (especially in self-attention).
- **Improves performance** in tasks like translation, summarization, and image captioning.


The
animal
didn't
cross
the
street
because
it
was
too
wide
.
The
animal
didn't
cross
the
street
because
it
was
too
wide
.