**Cross Attention**

======================================================================

Cross-Attention is a type of attention mechanism in Transformers that allows one sequence to attend to a different sequence — it's used to connect two different sequences.

In Cross-Attention, the decoder is generating an output sequence (e.g., a translation), and it needs to attend to the encoder’s output (the original input sentence).

So, instead of attending to itself, it attends to the **encoder’s output representations.**

----------------------------------------------------------------------------------------------------------

**Where is it used?**

Cross-Attention is used in the decoder of the Transformer architecture, especially in tasks like:

* Machine Translation (e.g., English → French)
* Image captioning
* Multimodal models (e.g., text attending to images)

----------------------------------------------------------------------------------------------------------

**How it works:**

* In Self-Attention, we use:



* In Cross-Attention, we use:



That means the decoder is asking:

*"For this decoder token, which encoder tokens are most relevant?"*

The decoder's query interacts with the encoder’s keys and values to gather useful context.

------------------------------------------------------------------------------------------------------------

**Visual Intuition:**

* Self-Attention: 🔄 Looks within its own sequence.
* Cross-Attention: 🔁 Looks at another sequence for information.

--------------------------------------------------------------------------------------------------------------

In summary, cross-attention acts as a bridge between the encoder and the decoder in Transformer models for sequence-to-sequence tasks. It enables the decoder to dynamically focus on the relevant parts of the encoded input sequence while generating the output sequence, leading to significant improvements in performance for tasks like machine translation and text generation.