Attention Is All You Need

An Intuitive Understanding of Self-Attention

arXiv > cs > arXiv:1706.03762

Sea
Help

Computer Science > Computation and Language

[Submitted on 12 Jun 2017 (v1), last revised 2 Aug 2023 (this version, v7)]

Attention Is All You Need

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks in an encoder-decoder configuration. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Motivation - Why Attention?

Human attention is selective: **we focus on important parts,** ignoring irrelevant details.

Example: In the phrase **"A Lannister Always Pays His Debts,"** the word "Lannister" carries most of the context.

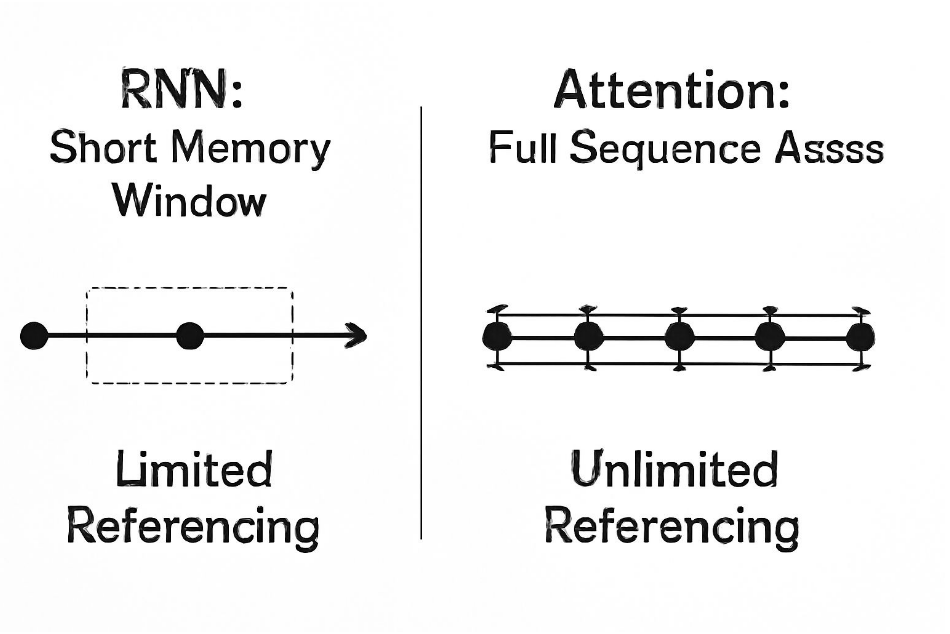
Attention mechanism mimics this by focusing on relevant parts of input data.

Why Were Attention Mechanisms Needed?

RNNs, LSTMs, and GRUs have limited memory windows, struggling with long sequences.

Attention provides "**infinite**" context access, allowing models to attend to all previous tokens.

Enables transformers to capture **long-range dependencies** efficiently.



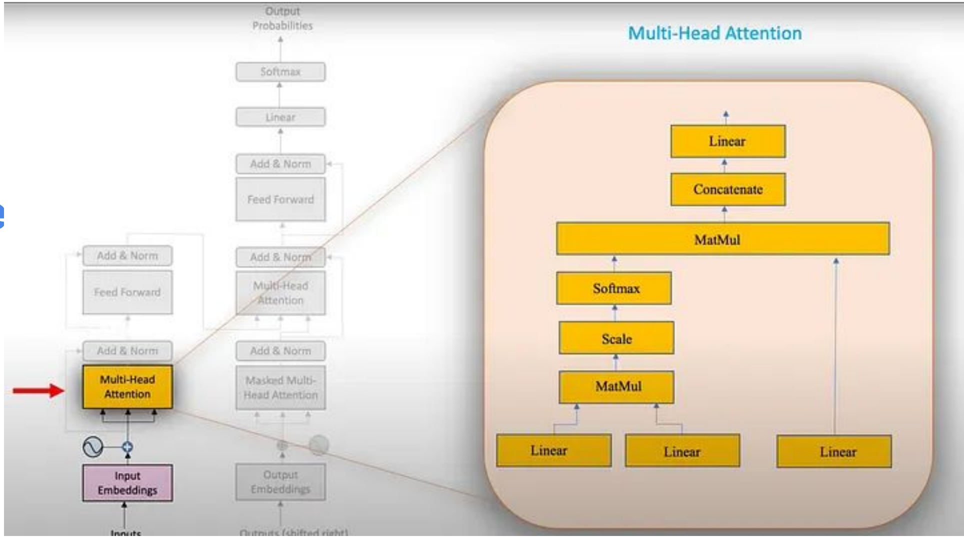
What is Self-Attention?

Self-attention relates words within the same sentence, deciding what needs focus.

Key components: **Query, Key, and Value vectors.**

Three key questions:

- What are Query, Key & Value?
- What is Positional Encoding?
- What inputs do Query, Key & Value receive?



Understanding Query, Key & Value

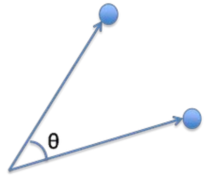
Query: What we are searching for (like search text).

Key: Metadata or title of content (like video titles).

Value: The actual content or information.

Attention measures similarity between Query and each Key to weigh Values accordingly.

$$sim(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$



$$Similarity(A,B) = \frac{A.B^T}{scaling}$$

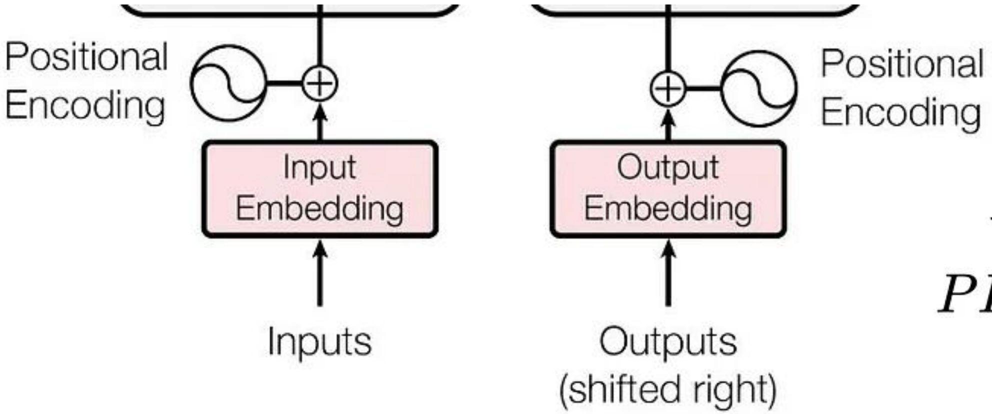
$$Similarity(Q,K) = \frac{Q.K^T}{scaling}$$

Role of Positional Encoding

Transformers process entire sequences at once, losing word order info.

Positional encoding adds unique position information to each word embedding.

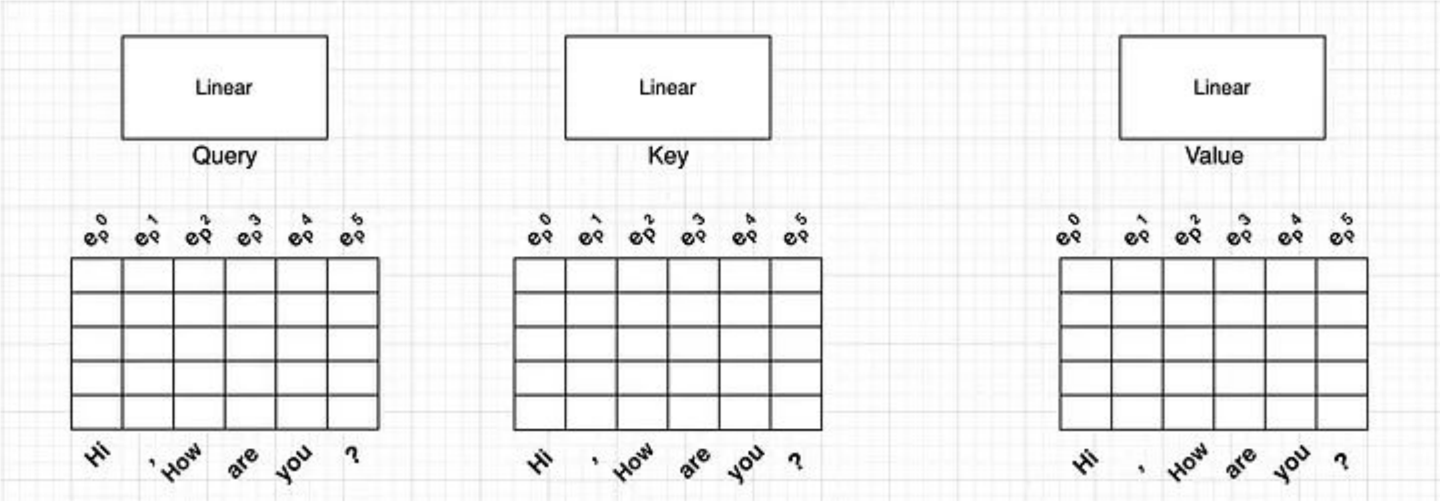
Uses **sinusoidal waves** to encode relative and absolute positions.



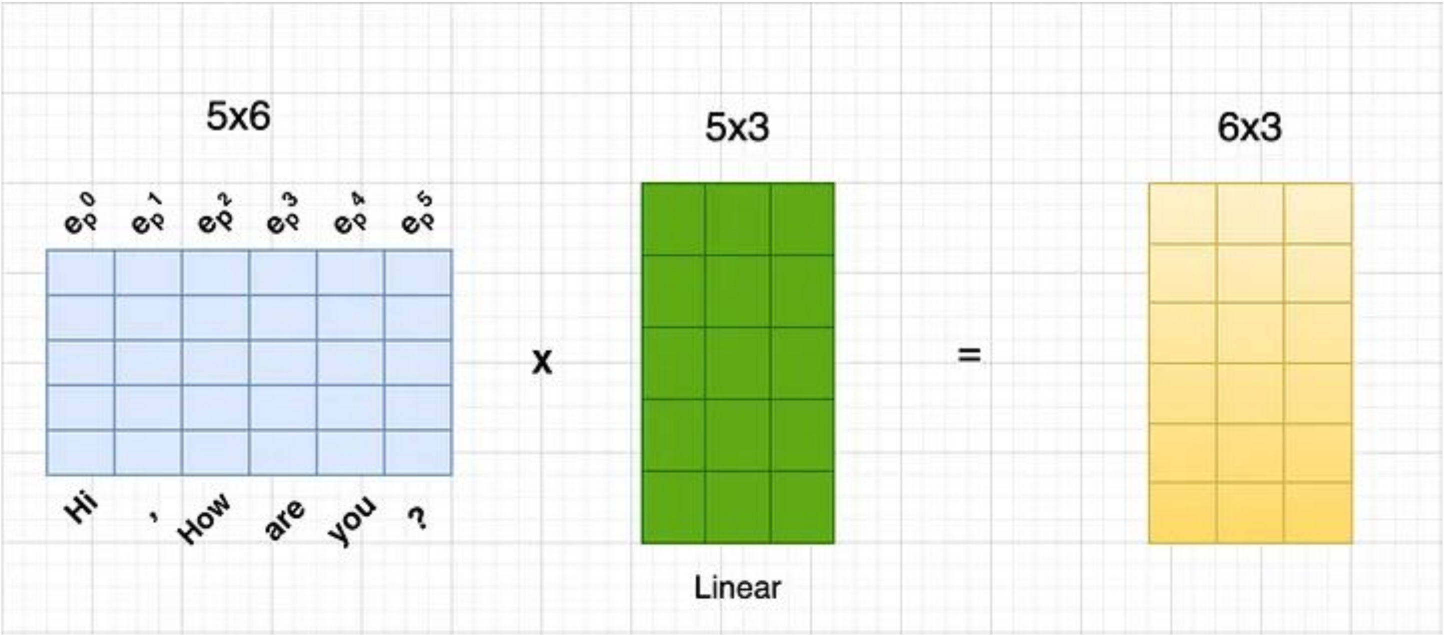
$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

“Hi, How are you?” and you want your transformer to output “I am fine”

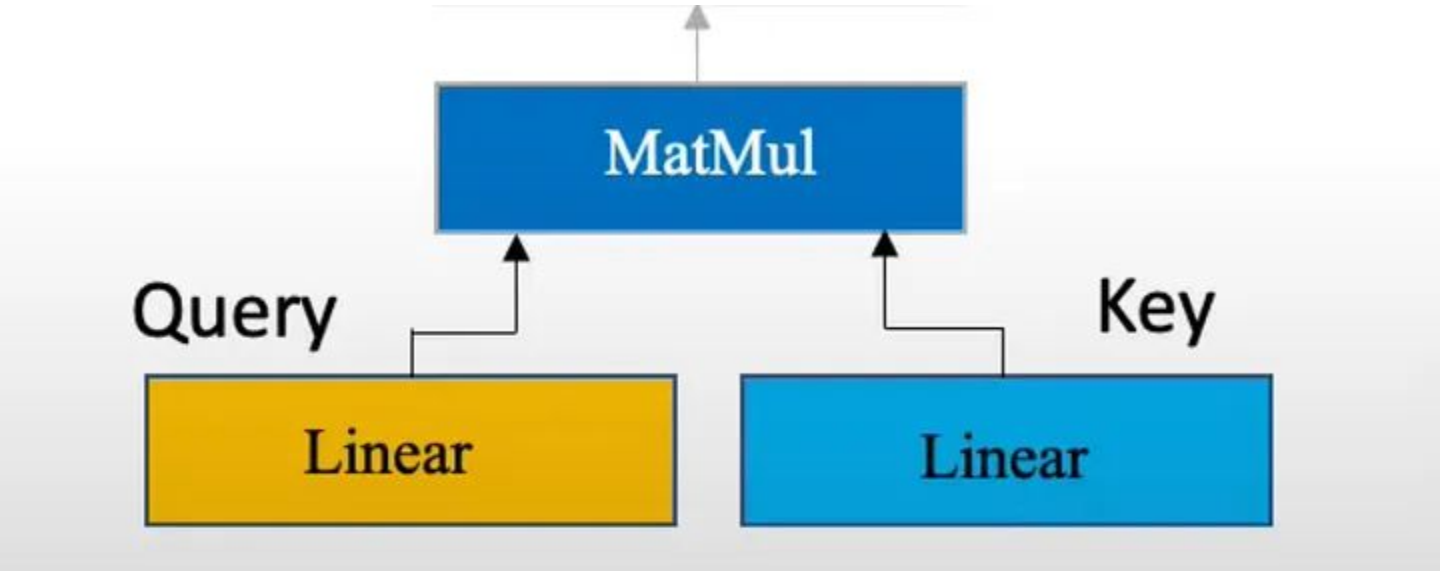
Step 1 - Input sentence embedding + positional encoding → three copies going to Query, Key, and Value layers.



Step 1 - Input sentence embedding + positional encoding → three copies going to Query, Key, and Value layers.

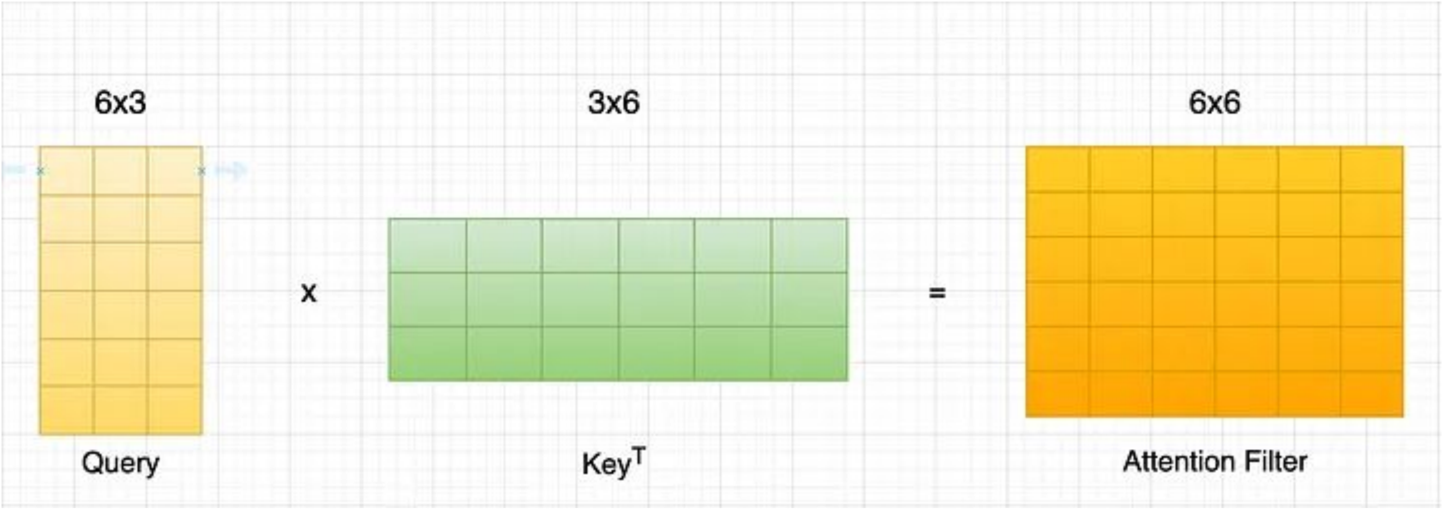


Step 2 - Query and Key used to calculate attention scores via dot product and scaling.



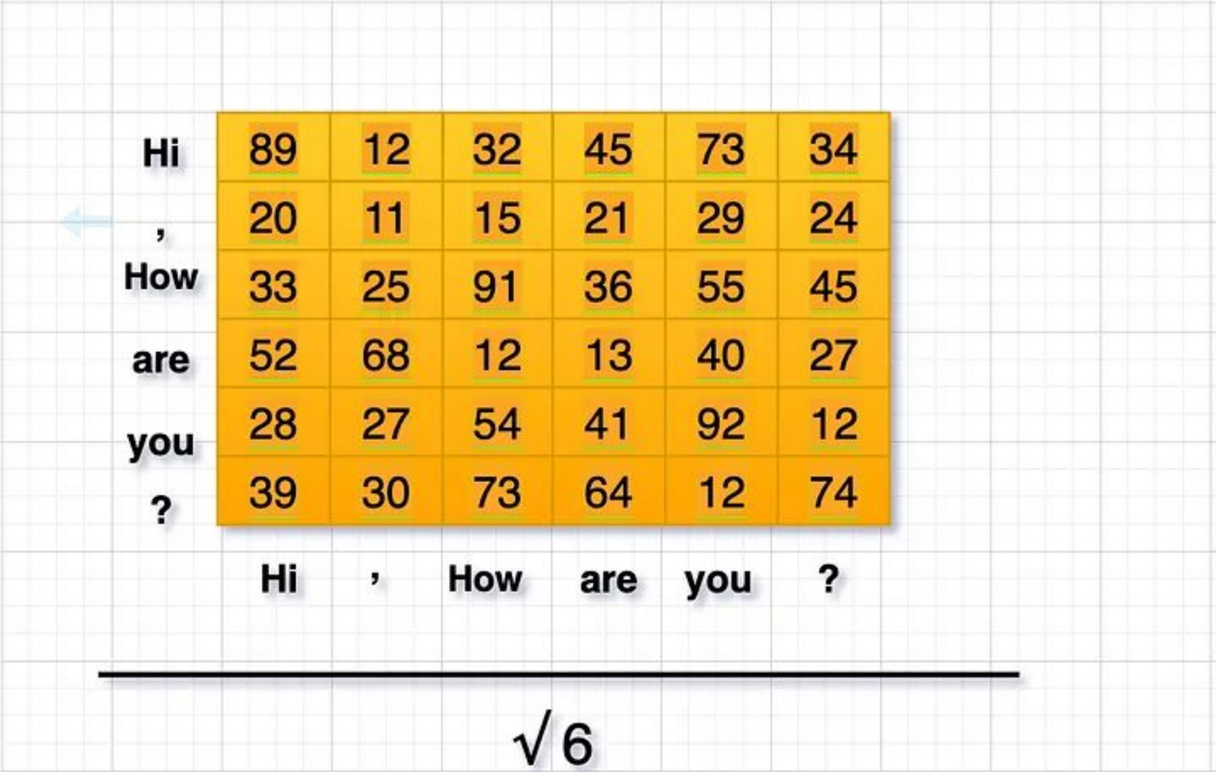
Step 2 - Query and Key used to calculate attention scores via dot product and scaling.

The output of this dot product can be called an ***Attention filter.***

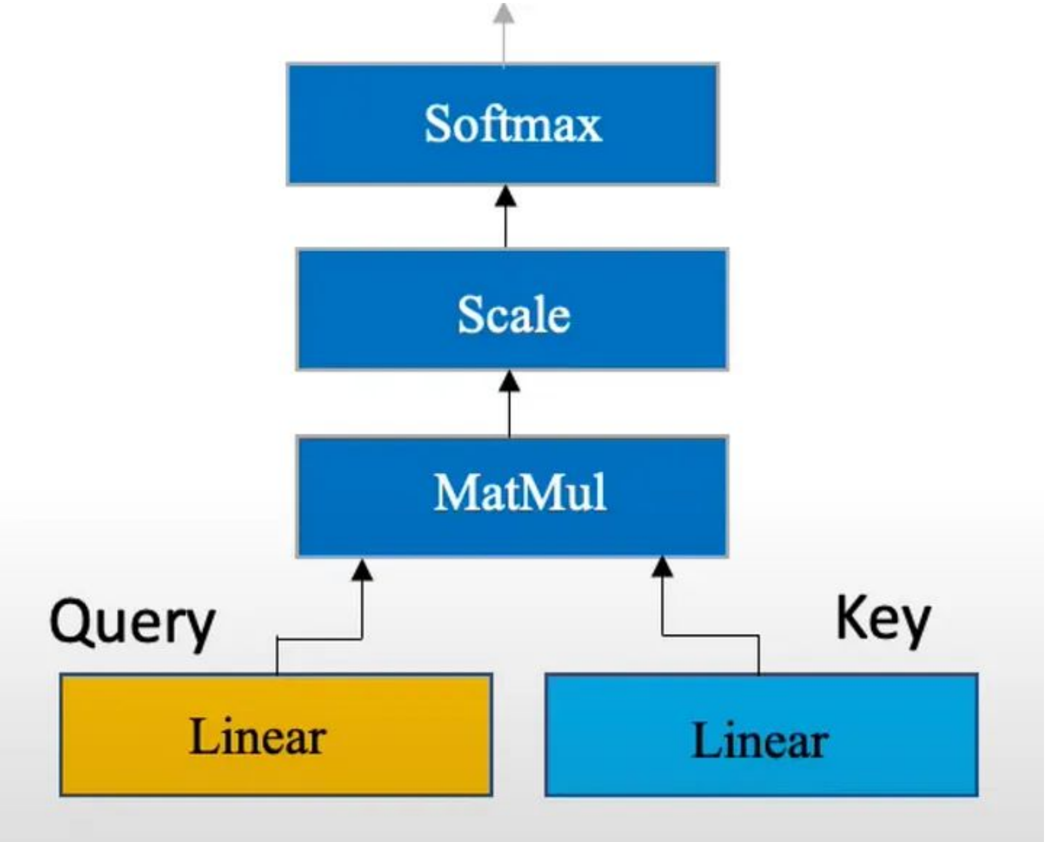


Hi	89	12	32	45	73	34
,	20	11	15	21	29	24
How	33	25	91	36	55	45
are	52	68	12	13	40	27
you	28	27	54	41	92	12
?	39	30	73	64	12	74
	Hi	,	How	are	you	?

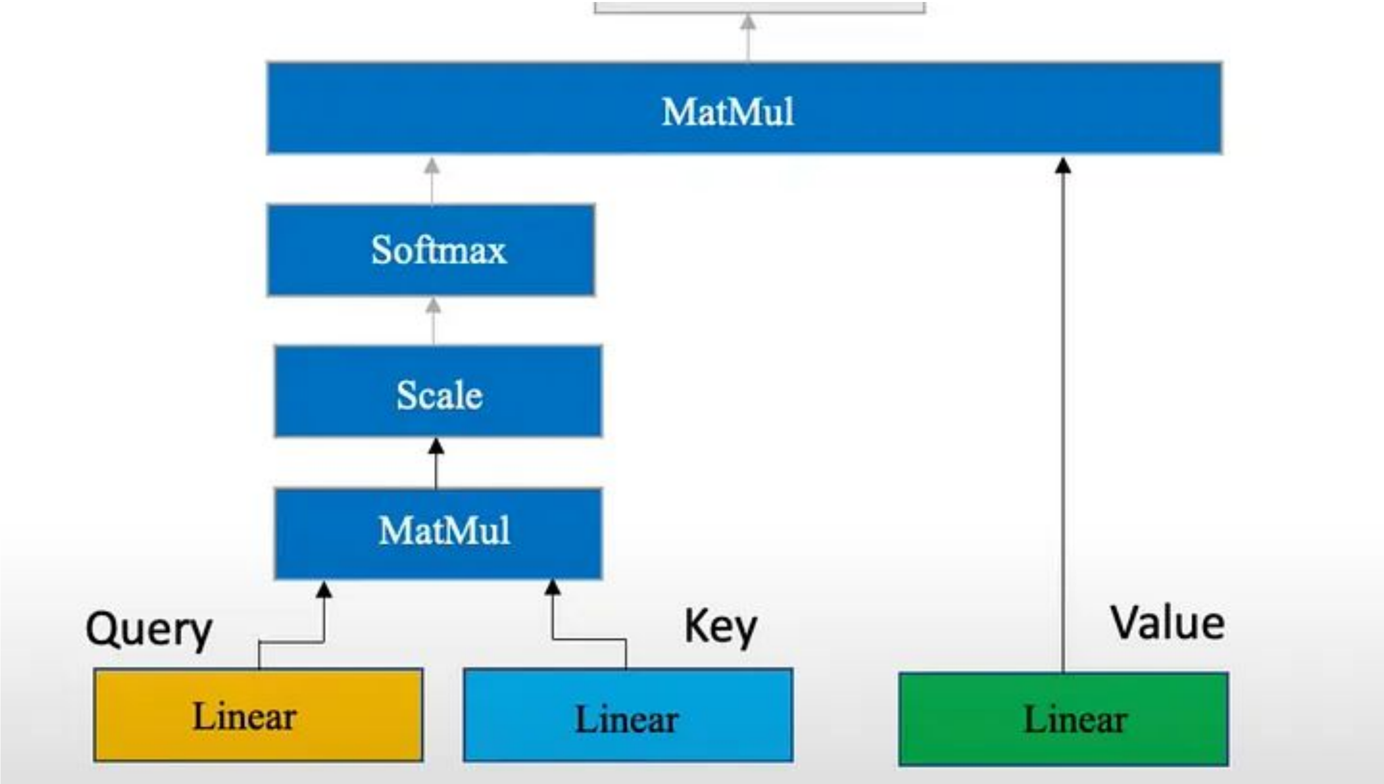
The authors of the “Attention is all you need” paper divided the attention score by the square root of the dimension of the key vector, in our case i.e. 6.

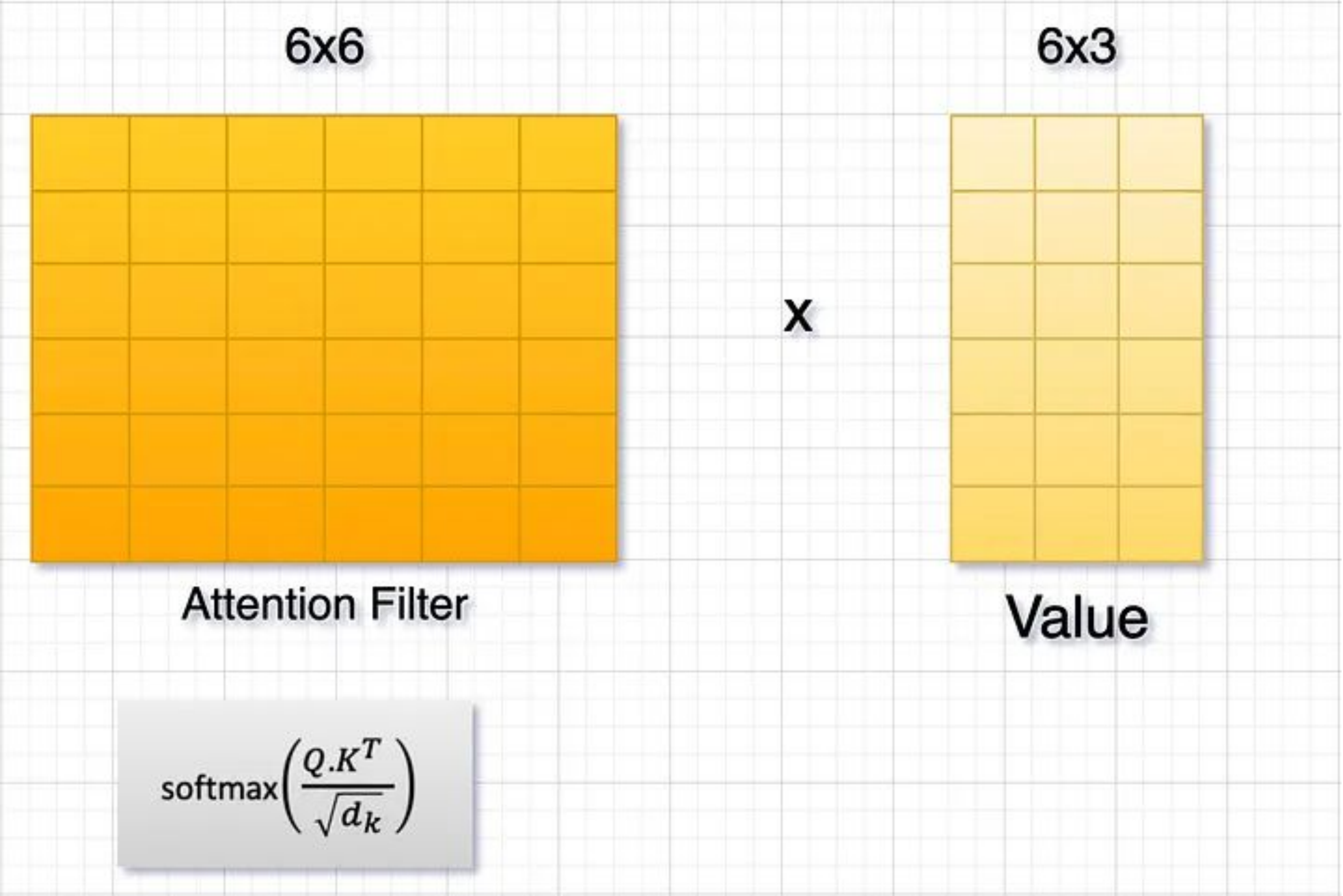


Step 3 - Softmax normalizes scores to form attention weights.



Step 4 - Attention weights multiply Value vectors to produce output.




6x6
6x3
x
Attention Filter
Value
softmax$\left(\frac{Q.K^T}{\sqrt{d_k}}\right)$

Intuition

Attention Filter Original Image Filtered Image

Final Formula of Attention Mechanism

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

Self-Attention Step-by-Step

1. Input sentence embedding + positional encoding → three copies going to Query, Key, and Value layers.
2. Query and Key used to calculate attention scores via dot product and scaling.
3. Softmax normalizes scores to form attention weights.
4. Attention weights multiply Value vectors to produce output.
5. Output passes through a linear layer for final representation.

Why Attention Filters Matter

Attention filters help emphasize relevant features, discard noise.

Analogous to image processing, where filters highlight important patterns.

Result: Better context understanding improving predictions.

Summary & Impact

Attention enables long-term dependency modeling beyond RNN limits.

Self-attention creates dynamic relationships within input sequences.

Powers state-of-the-art models like Transformers in NLP and beyond.

Quiz Time!

What is the main purpose of the attention mechanism in transformer models?

A) To increase model size

B) To focus on relevant parts of the input sequence

C) To reduce input data

D) To eliminate all noise from the input

Quiz Time!

In the self-attention mechanism, what does the "Query" vector represent?

A) The content being searched

B) The metadata or titles

C) The input sentence embedding with positional information

D) The output prediction

Quiz Time!

Why is positional encoding necessary in transformer models?

A) To increase the dimensionality of embeddings

B) To provide information about the order of words since transformers process all input simultaneously

C) To reduce model computation time

D) To remove irrelevant words from the input

Quiz Time!

What operation is used to calculate the similarity score between Query and Key vectors in self-attention?

A) Addition

B) Dot product (scaled)

C) Subtraction

D) Concatenation

Quiz Time!

What does the output of the self-attention layer represent?

A) The original input word embeddings without modification

B) A weighted combination of Value vectors based on attention scores

C) The sum of Query and Key vectors

D) A randomly shuffled version of input embeddings