Transformers II

Dr. Parlett-Pelleriti

Outline

- Overview of Transformer Architecture
- Attention
- Attention in Transformers
- Multi-Headed Attention
- Masking and Cross-Attention
- Regularization in Transformers
- Transformers You Might Know

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

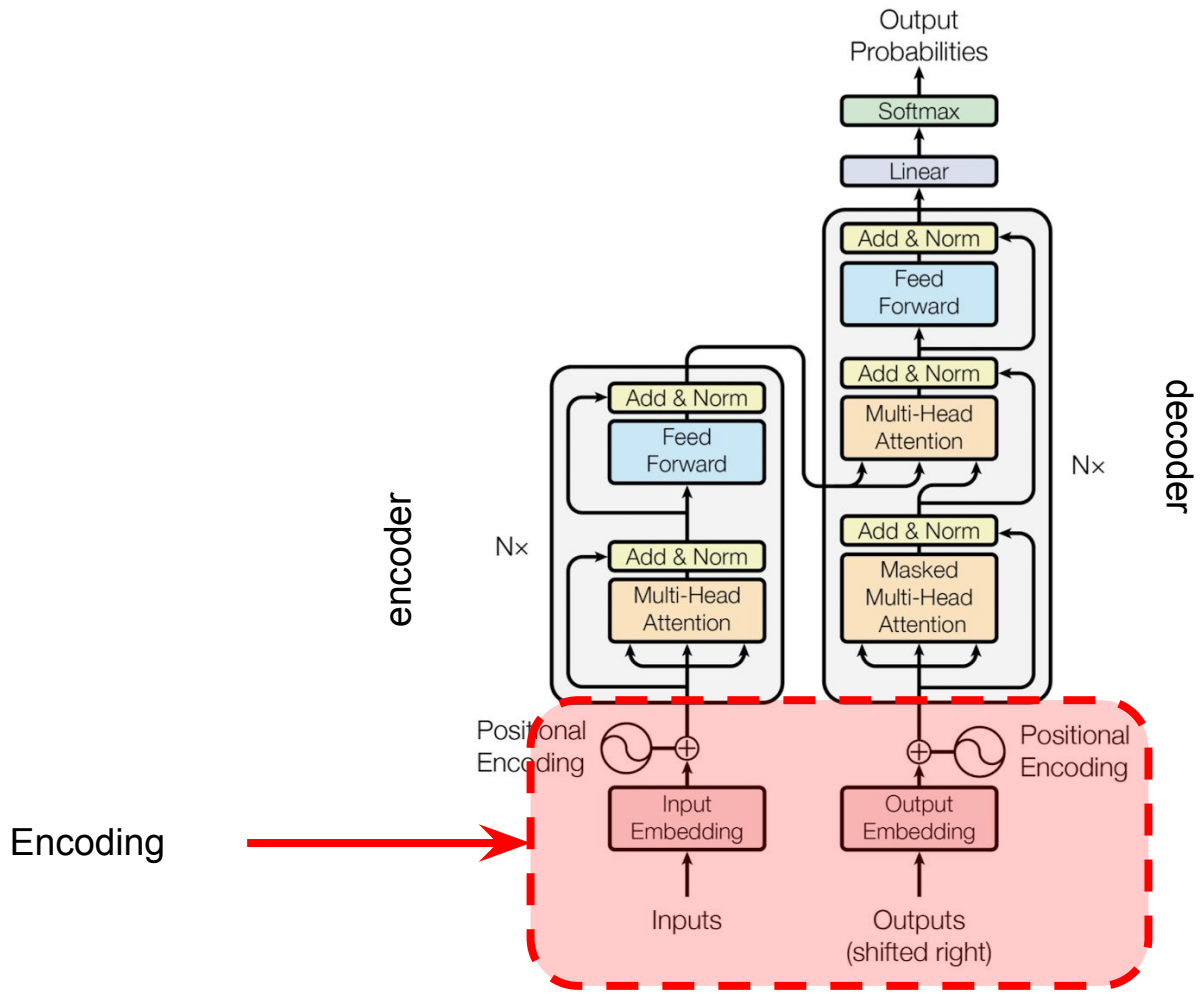


Figure 1: The Transformer - model architecture.

Image from: Attention is all you Need (Vaswani et. al 2017)

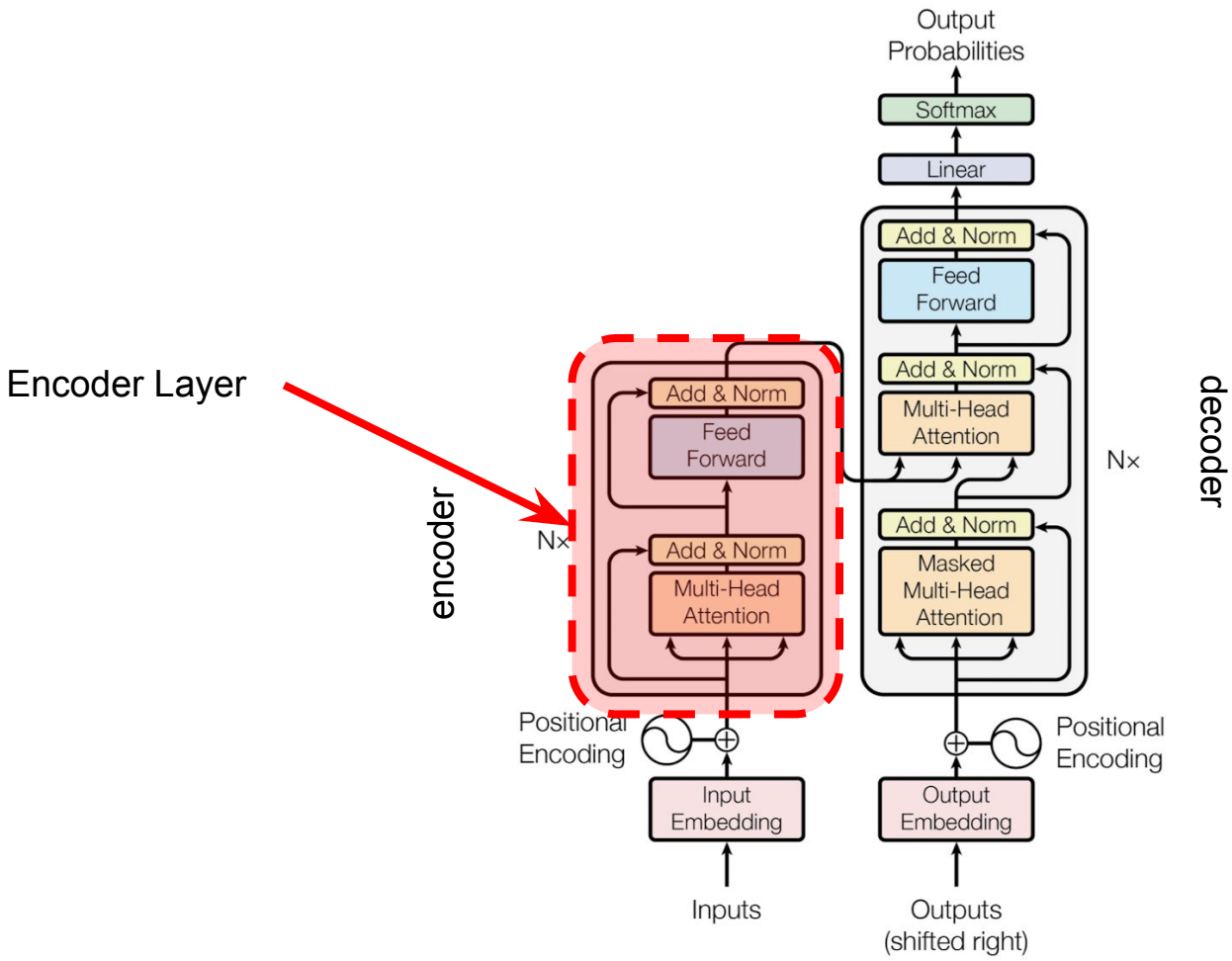


Figure 1: The Transformer - model architecture.

Image from: Attention is all you Need (Vaswani et. al 2017)

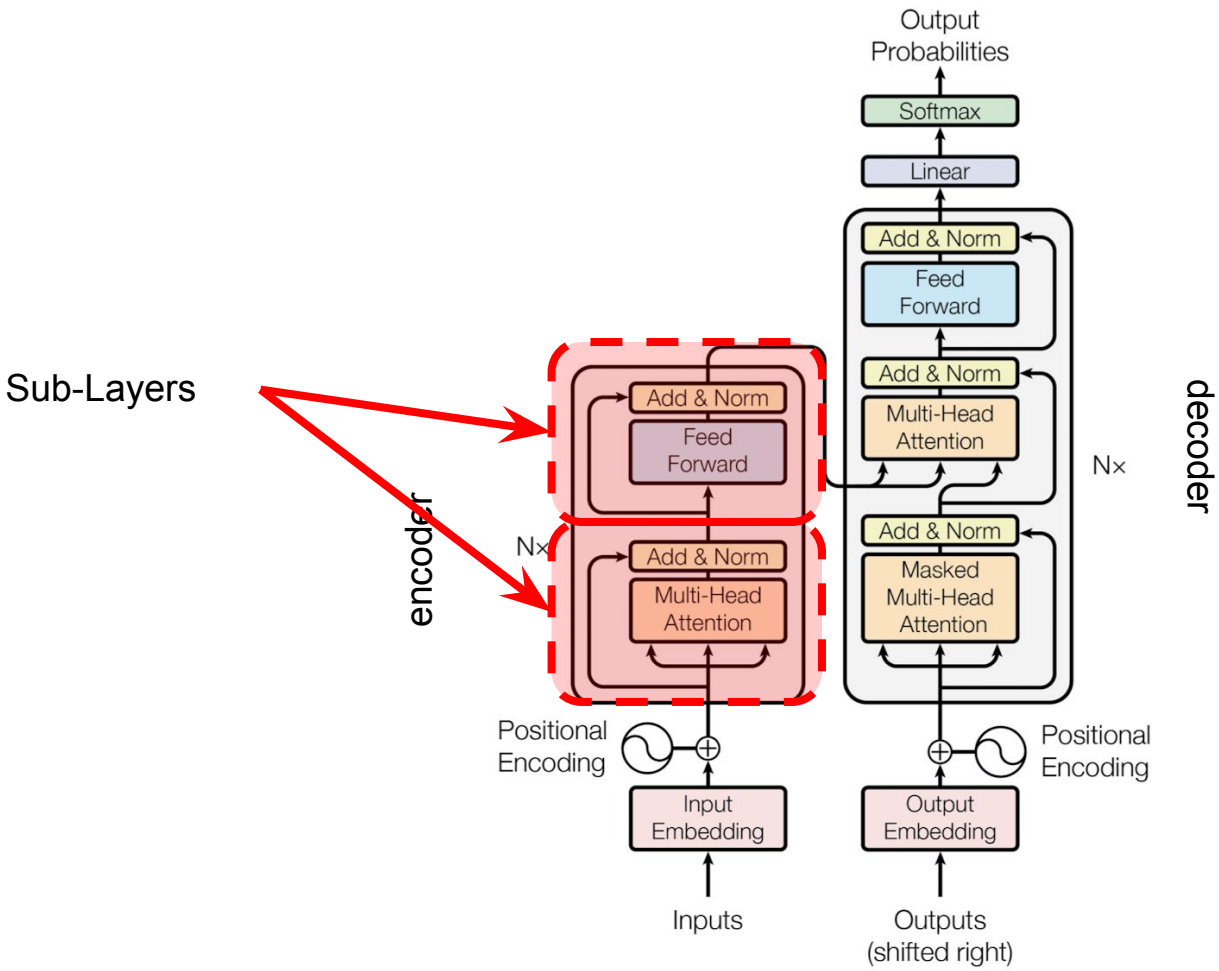


Figure 1: The Transformer - model architecture.

Image from: Attention is all you Need (Vaswani et. al 2017)

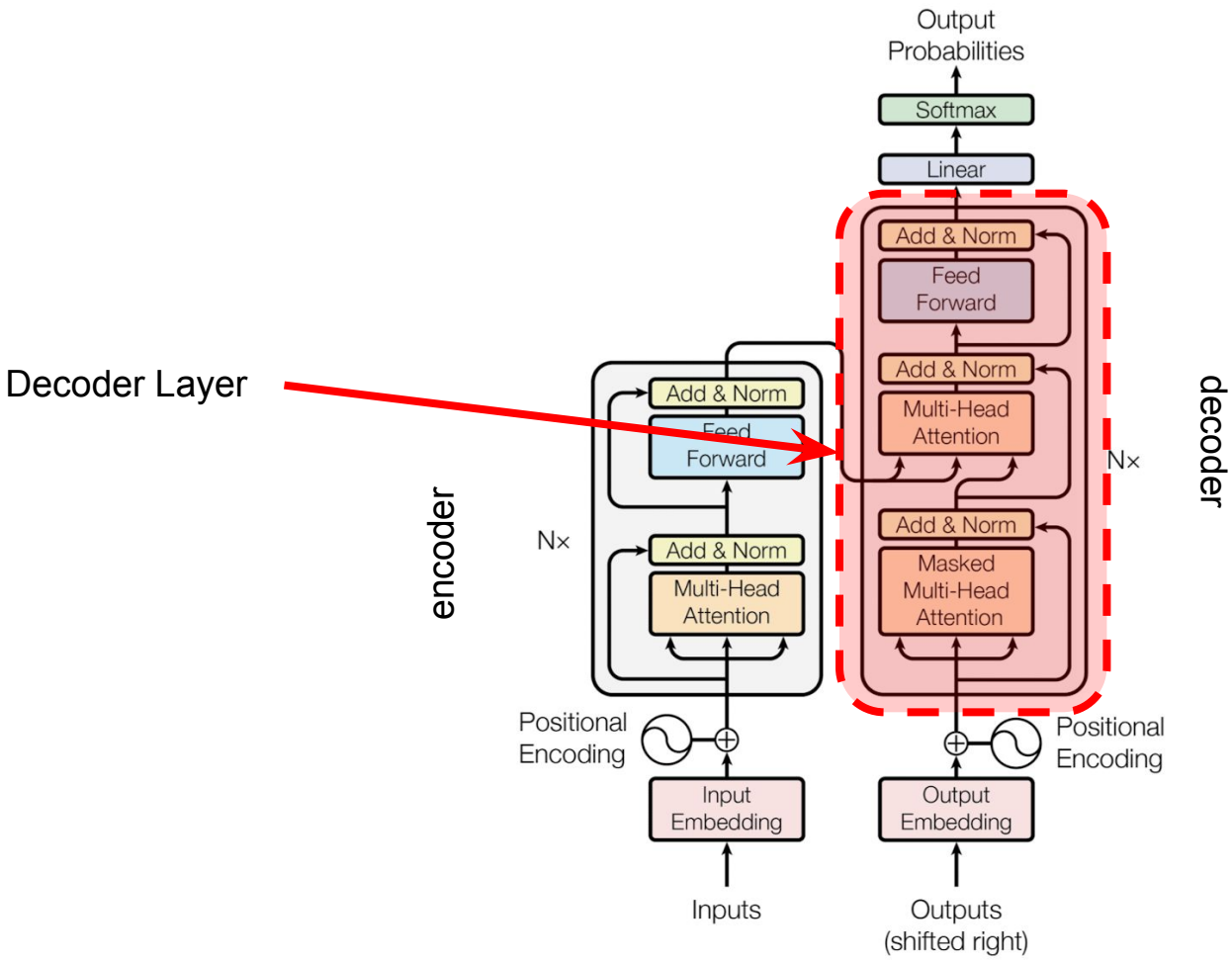


Figure 1: The Transformer - model architecture.

Image from: Attention is all you Need (Vaswani et. al 2017)

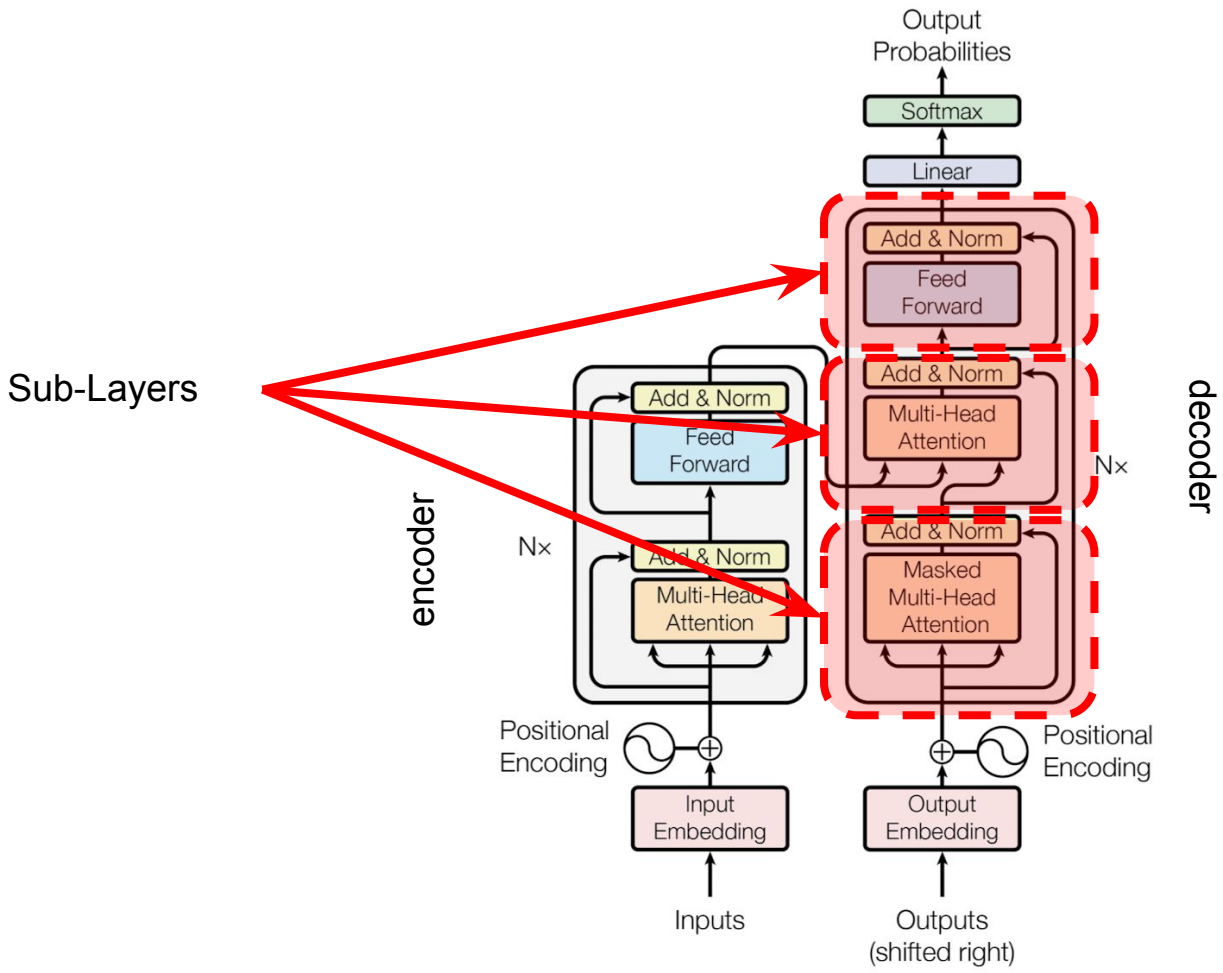


Figure 1: The Transformer - model architecture.

Image from: Attention is all you Need (Vaswani et. al 2017)

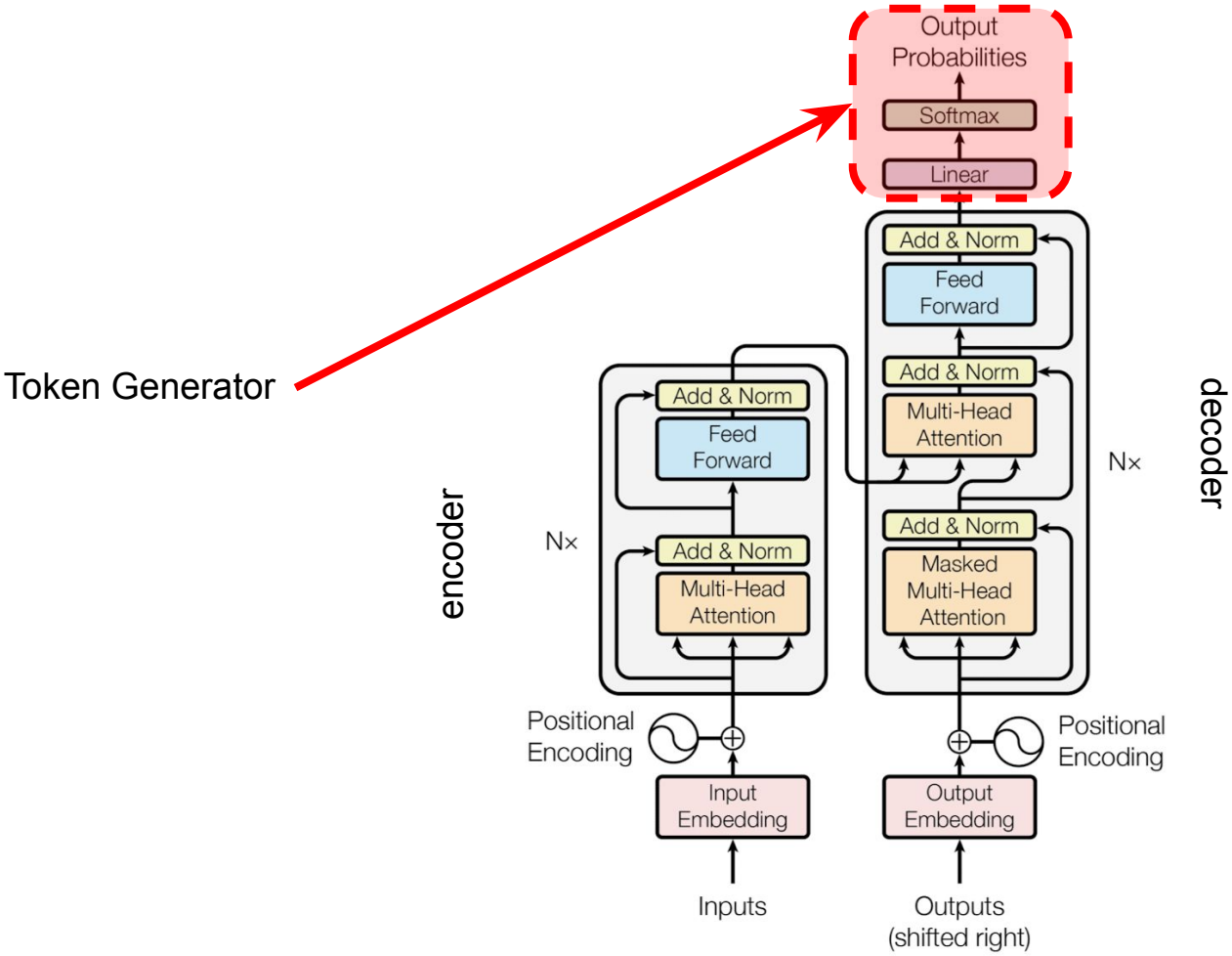


Figure 1: The Transformer - model architecture.

Image from: Attention is all you Need (Vaswani et. al 2017)

Vicki
@vboykis

They don’t tell you this in the paper (well they do but you have to read it like 15 times)

Attention

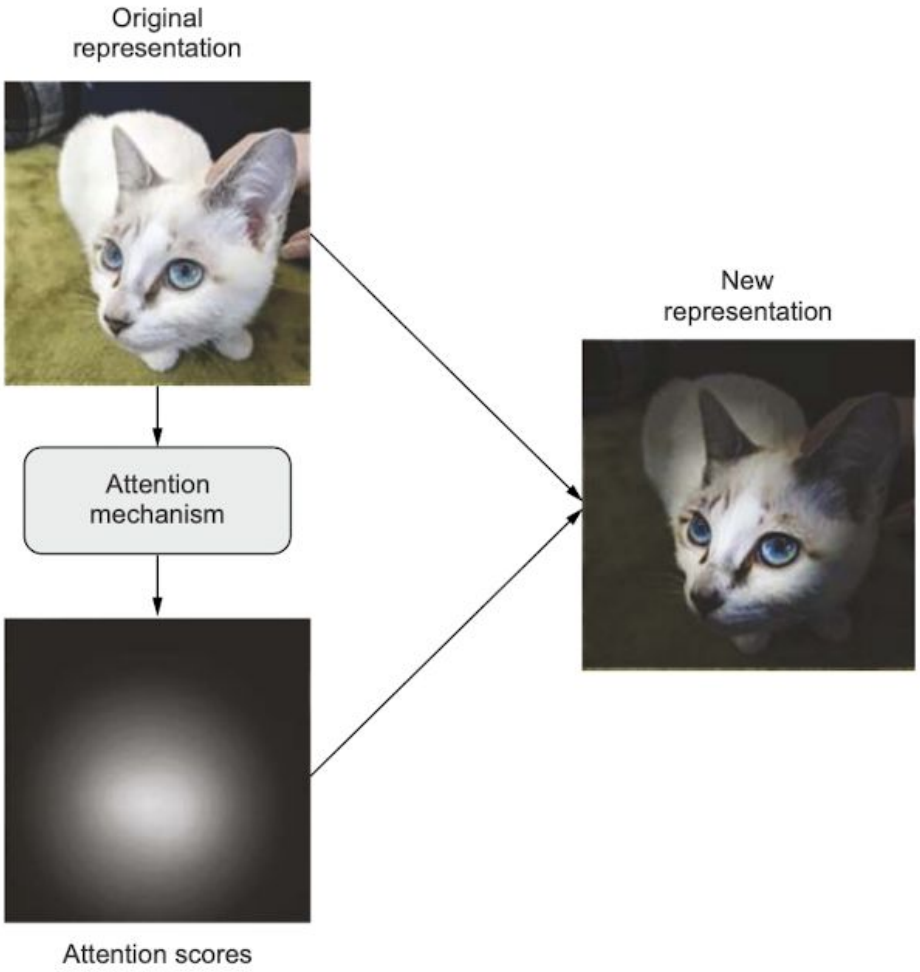


Image from: Deep Learning with Python (Chollet)

Attention

Attention

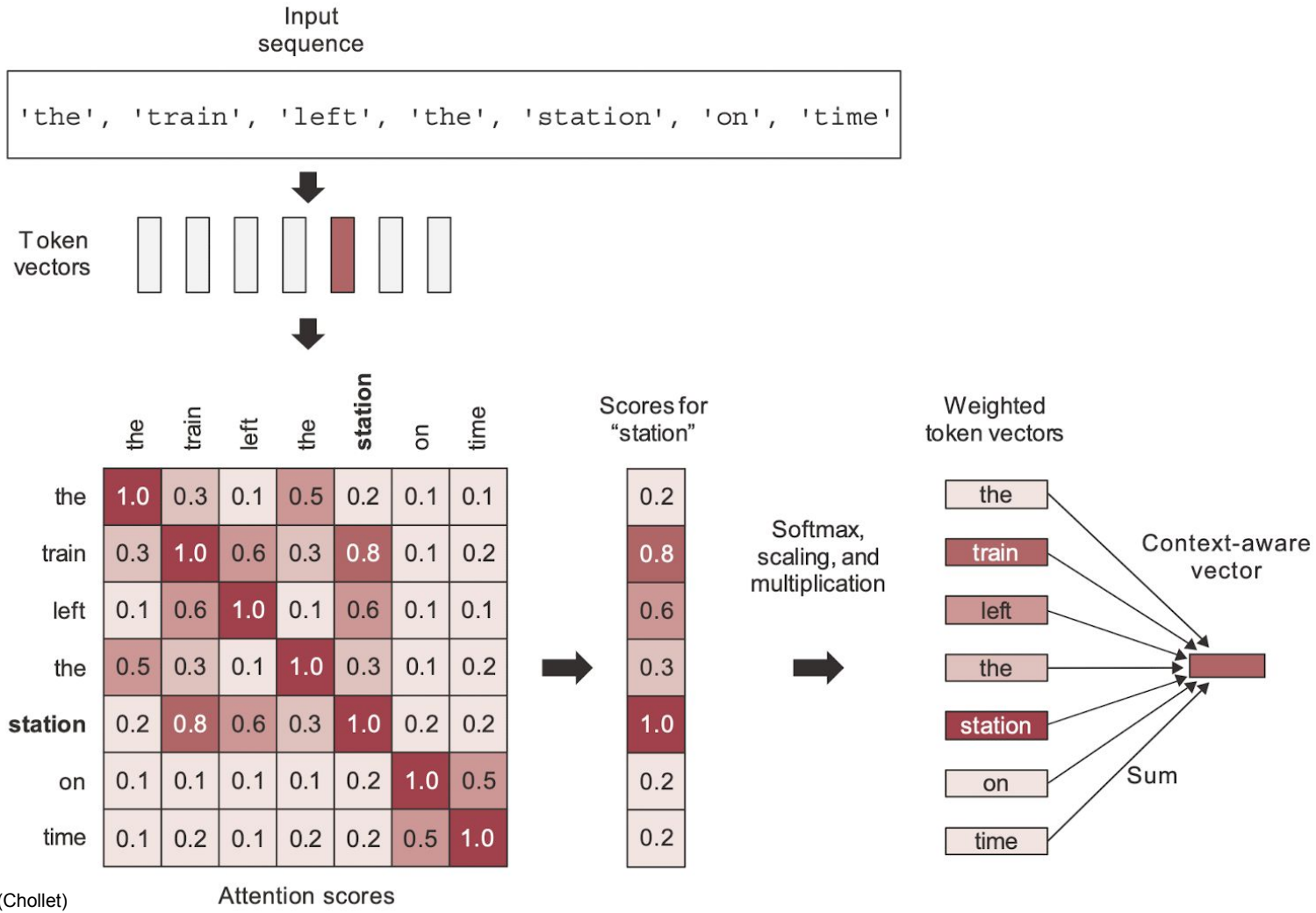


Image from: Deep Learning with Python (Chollet)

Attention

$$x'_i = \sum_{j=1}^{n} w_{ji} \cdot x_j$$

Attention

Attention weights

$$x'_i = \sum_{j=1}^{n} w_{ji} \cdot x_j$$

Attention

Attention weights

Each embedding

$$x'_i = \sum_{j=1}^{n} w_{ji} \cdot x_j$$

Attention

Attention weights

Each embedding

$$x'_i = \sum_{j=1}^{n} w_{ji} \cdot x_j$$

Context aware vector

Attention

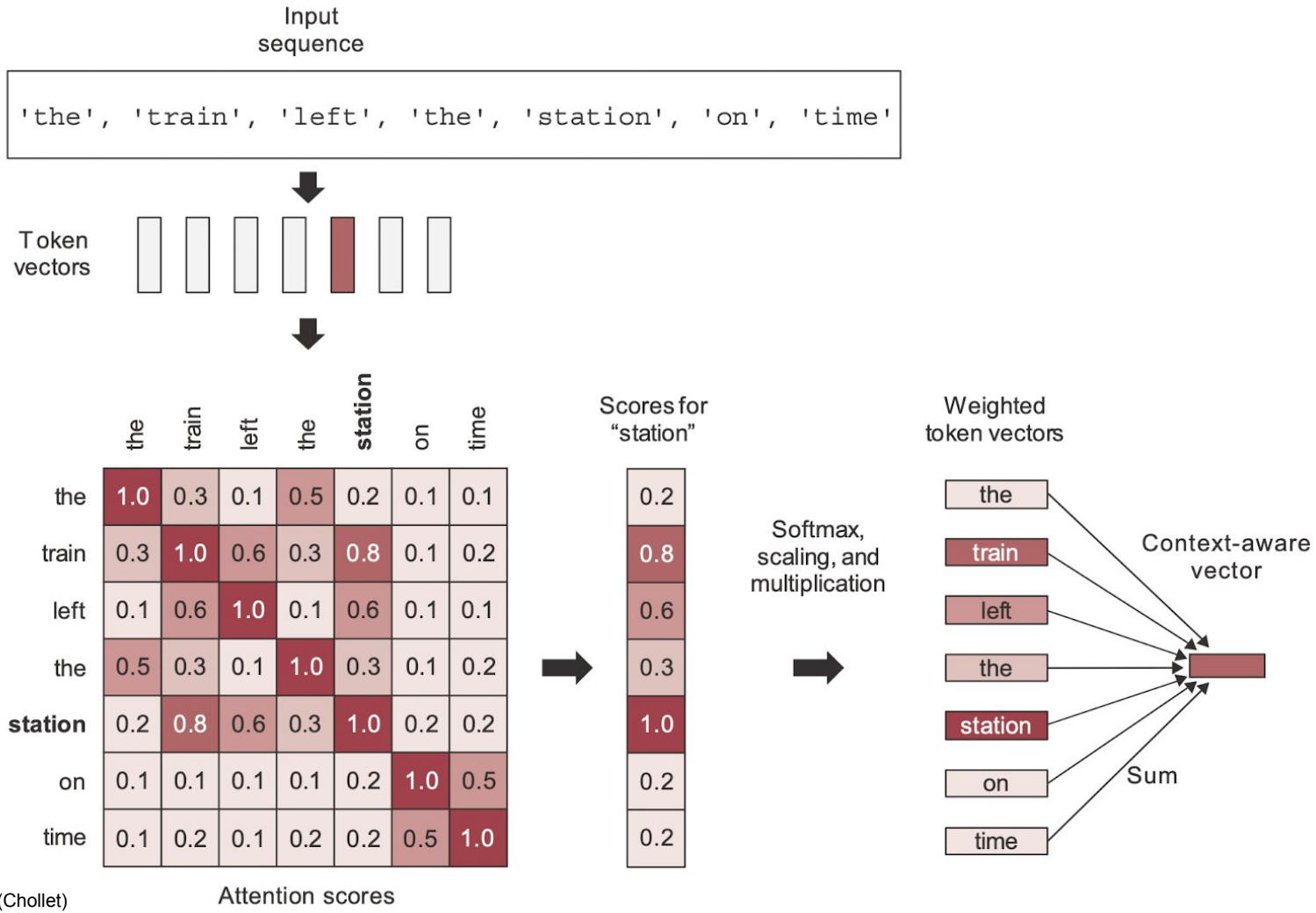


Image from: Deep Learning with Python (Chollet)

Attention in Transformers

Queries, Keys, and Values

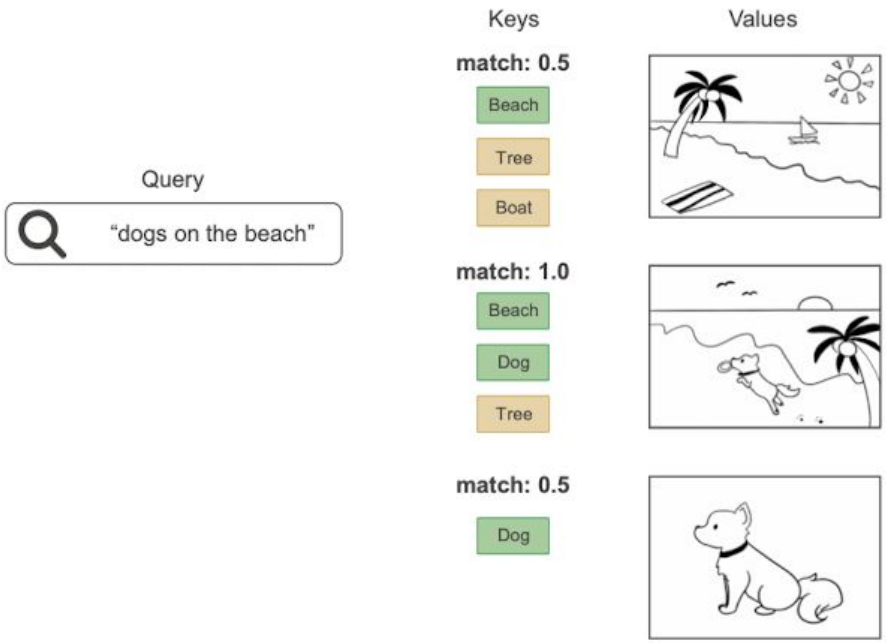


Figure 11.7 Retrieving images from a database: the "query" is compared to a set of "keys," and the match scores are used to rank "values" (images).

Image from: Deep Learning with Python (Chollet)

Queries, Keys, and Values

$$Q = E * W_q$$

$$K = E * W_k$$

$$V = E * W_v$$

Queries, Keys, and Values

Learnable Parameters

$$Q = E * W_q$$

$$K = E * W_k$$

$$V = E * W_v$$

Queries, Keys, and Values

Learnable Parameters

$$Q = E * W_q$$

$$K = E * W_k$$

$$V = E * W_v$$

Must be the same size

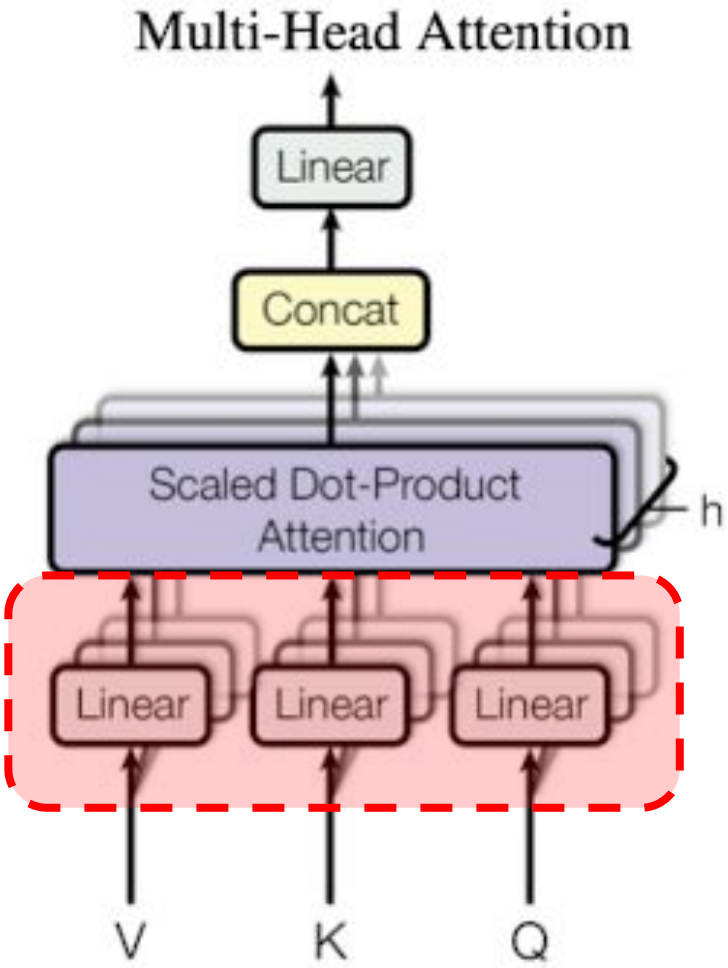


Image from: Attention is all you Need (Vaswani et. al 2017)

Queries, Keys, and Values

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

Vicki
@vboykis

They don’t tell you this in the paper (well they do but you have to read it like 15 times)

Queries, Keys, and Values

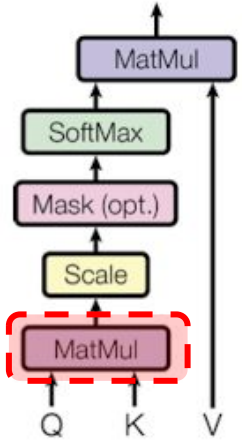
Attention Weights

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

Queries, Keys, and Values

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

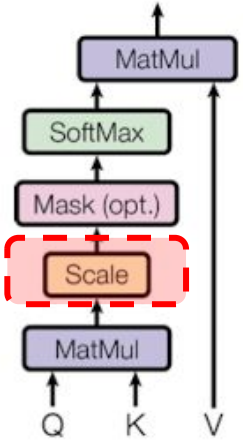
Scaled Dot-Product Attention



Queries, Keys, and Values

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

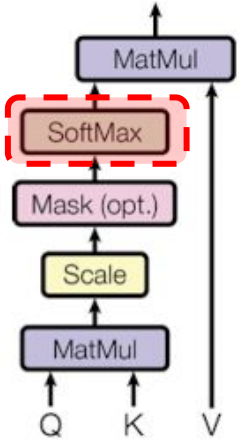
Scaled Dot-Product Attention



Queries, Keys, and Values

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

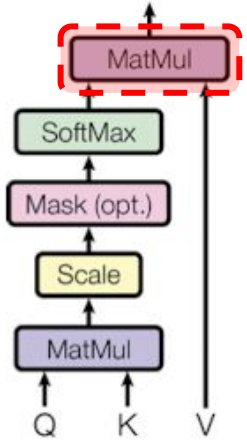
Scaled Dot-Product Attention



Queries, Keys, and Values

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

Scaled Dot-Product Attention



Scaled Dot-Product Attention

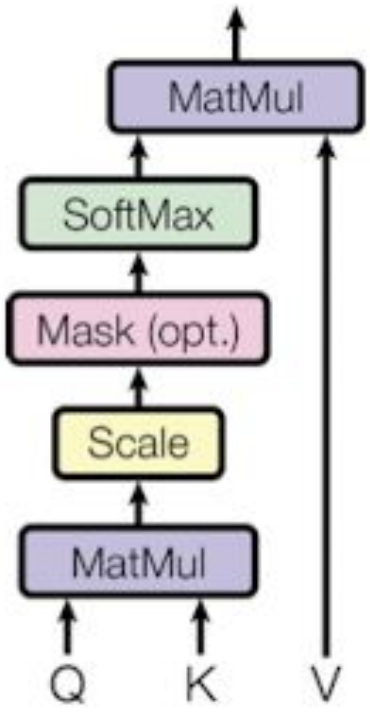


Image from: Attention is all you Need (Vaswani et. al 2017)

Masking

Scaled Dot-Product Attention

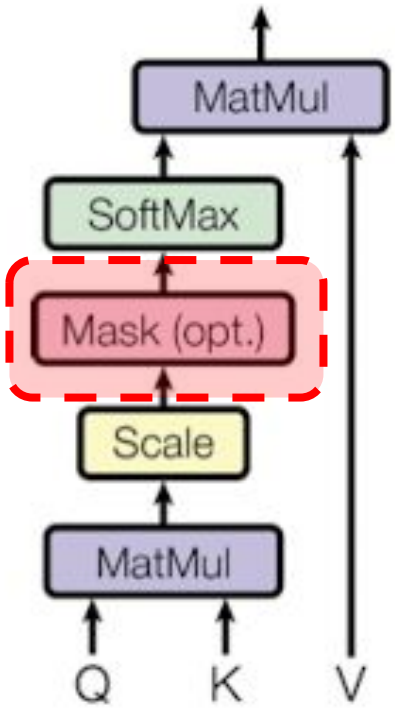


Image from: Attention is all you Need (Vaswani et. al 2017)

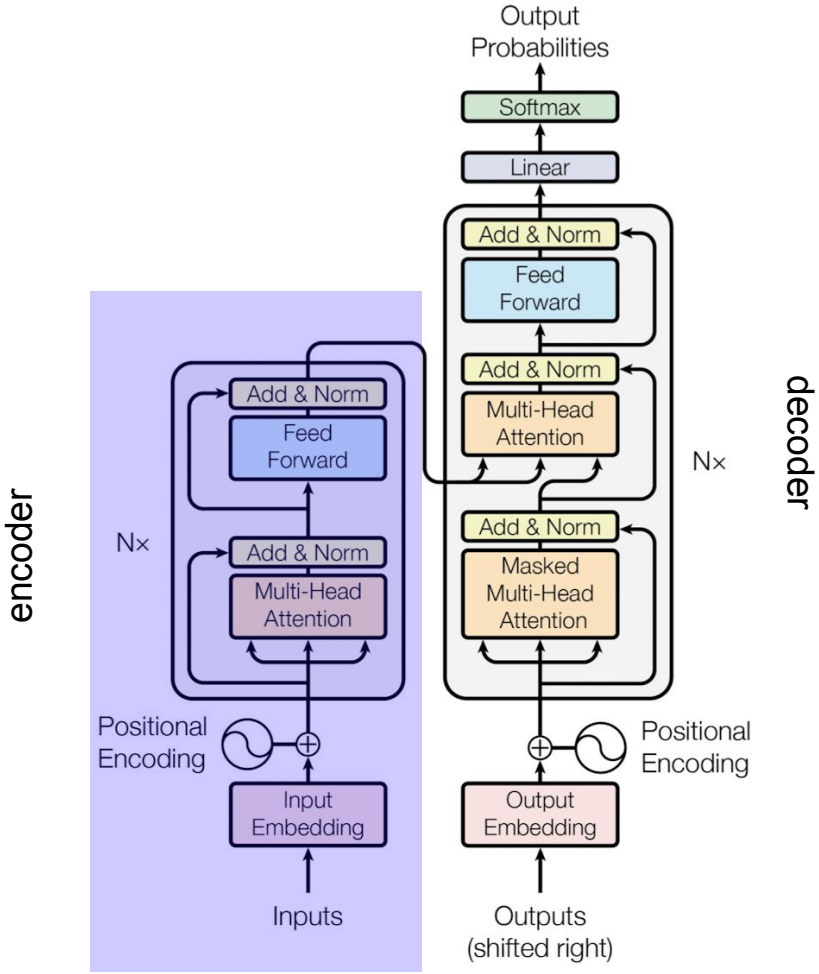


Figure 1: The Transformer - model architecture.

Image from: Attention is all you Need (Vaswani et. al 2017)

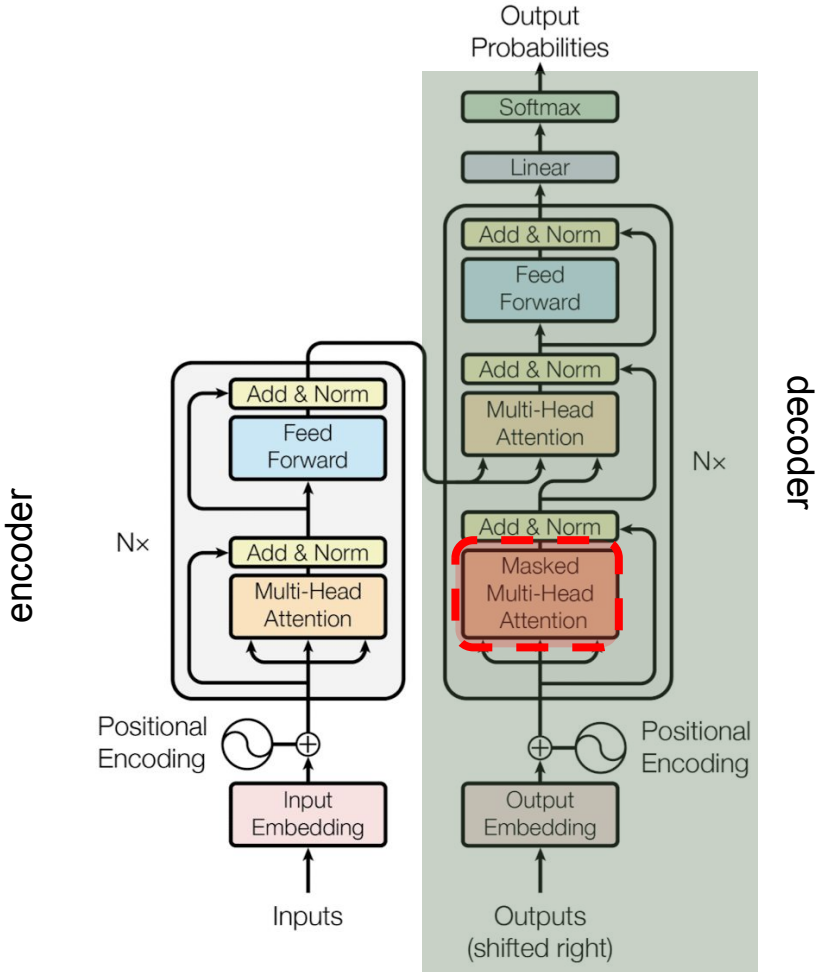


Figure 1: The Transformer - model architecture.

Image from: Attention is all you Need (Vaswani et. al 2017)

$$\begin{bmatrix} 1 & 0.2 & 0.6 \\ 0.1 & 1 & 0.8 \\ 0.2 & 0.1 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & -inf & -inf \\ 0.1 & 1 & -inf \\ 0.2 & 0.1 & 1 \end{bmatrix}$$

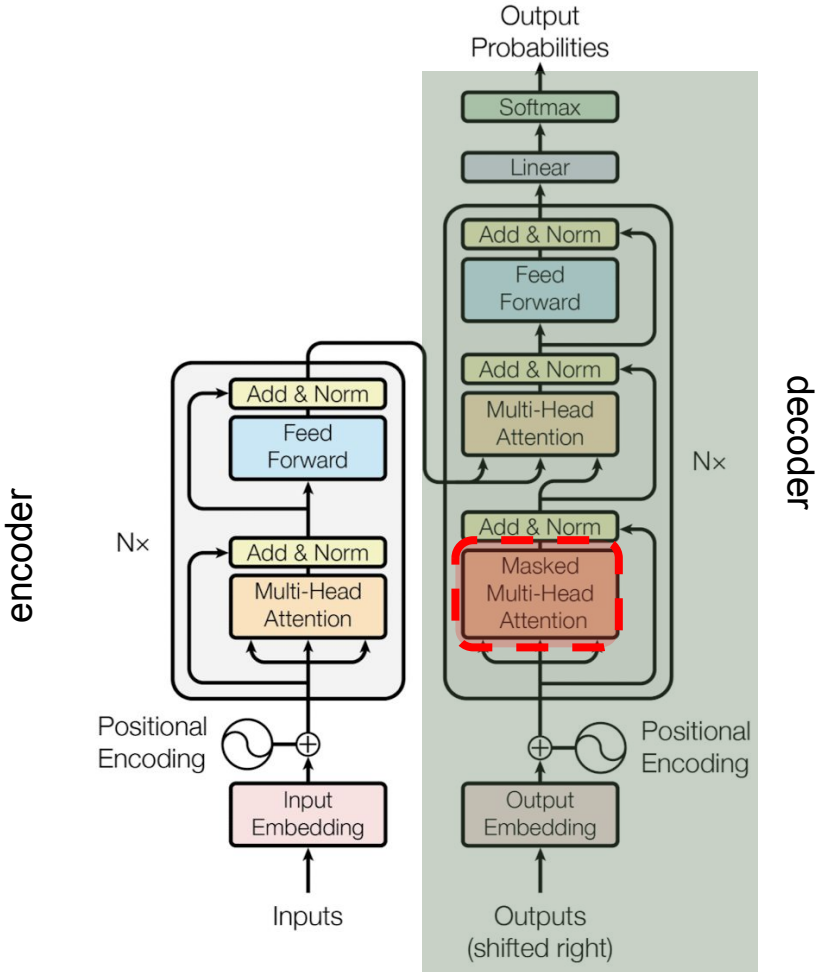


Figure 1: The Transformer - model architecture.

Image from: Attention is all you Need (Vaswani et. al 2017)

Cross-Attention

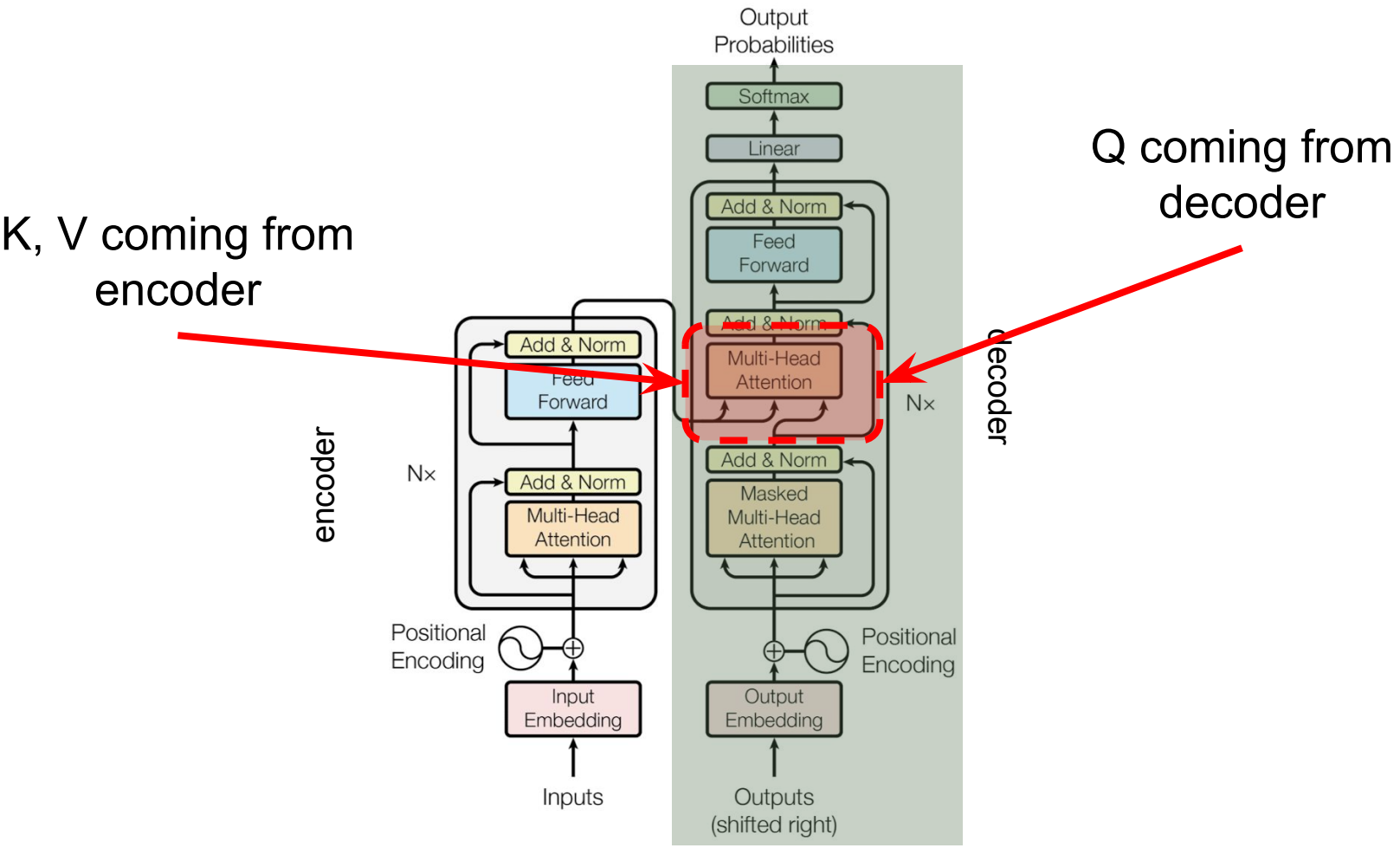


Figure 1: The Transformer - model architecture.

Image from: Attention is all you Need (Vaswani et. al 2017)

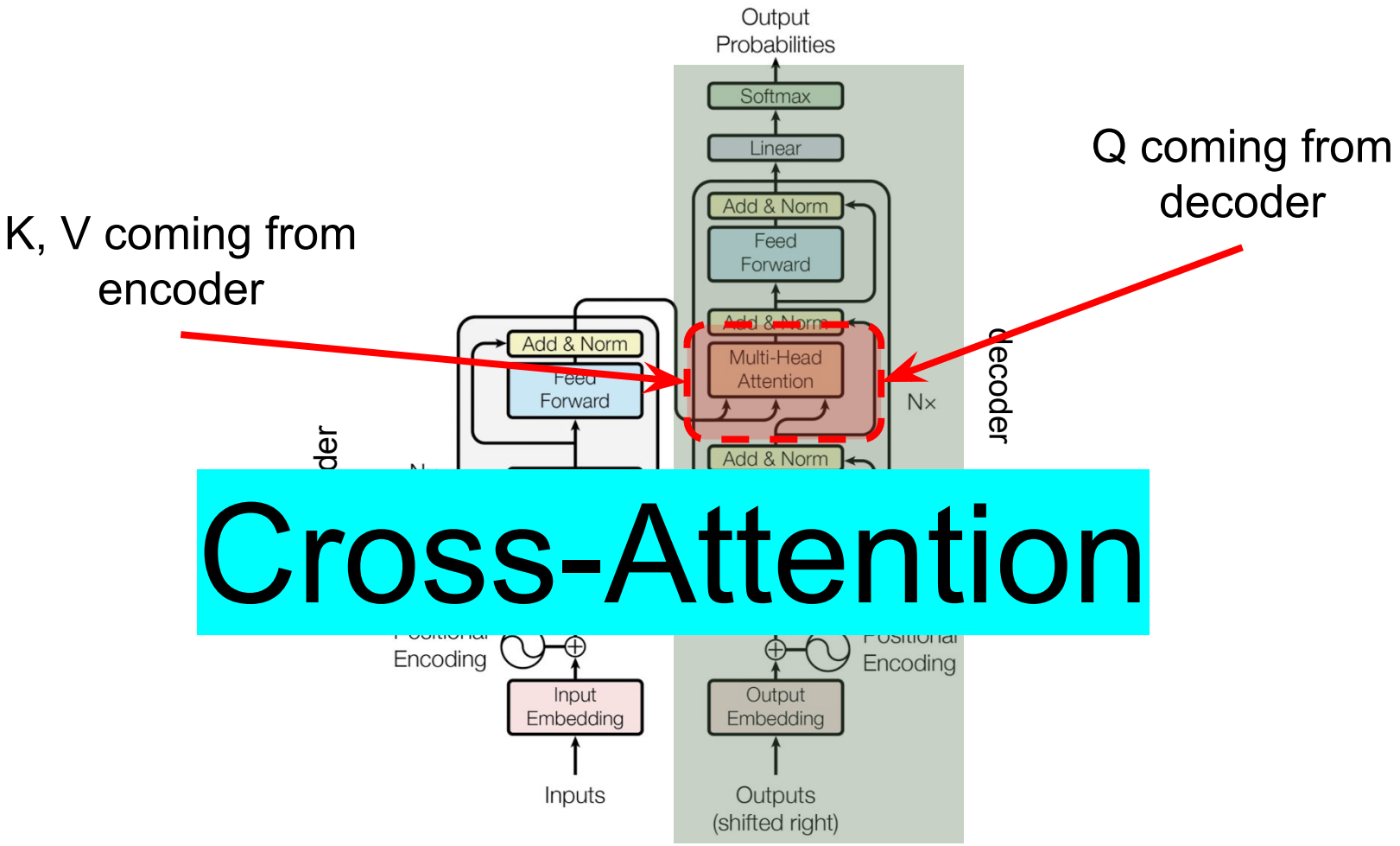


Figure 1: The Transformer - model architecture.

Image from: Attention is all you Need (Vaswani et. al 2017)

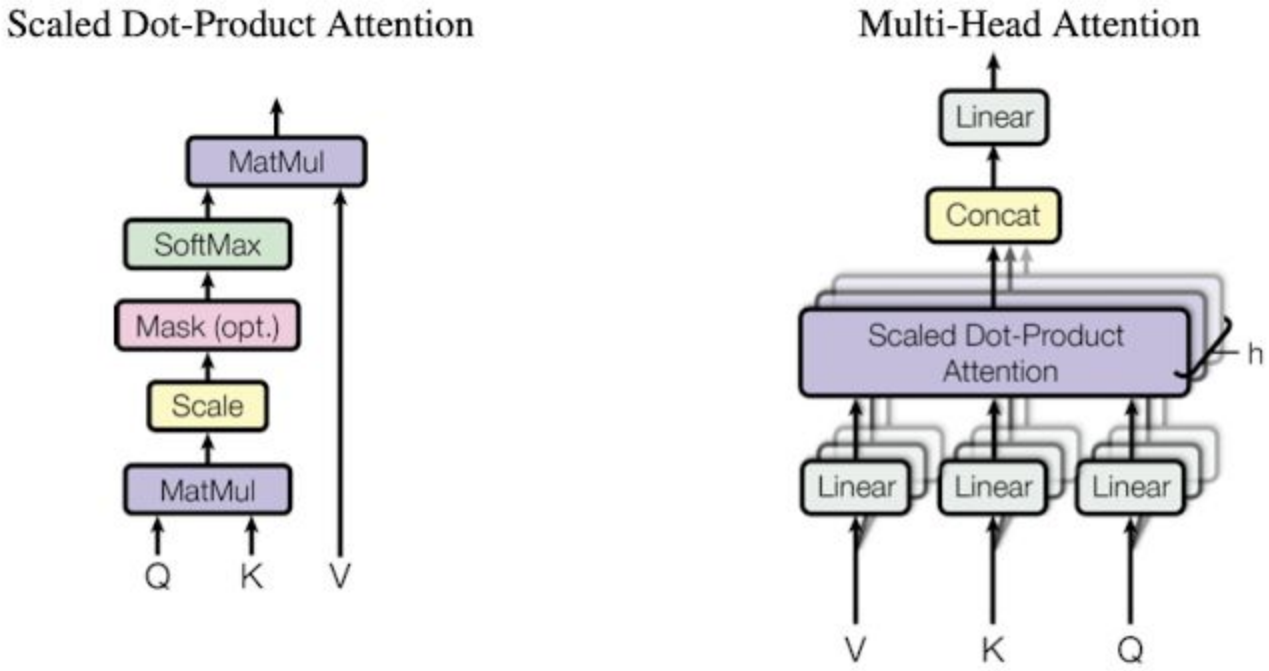


Figure 2: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel.

Image from: Attention is all you Need (Vaswani et. al 2017)

Regularization Methods

- Layer Norms
- Residual Connections

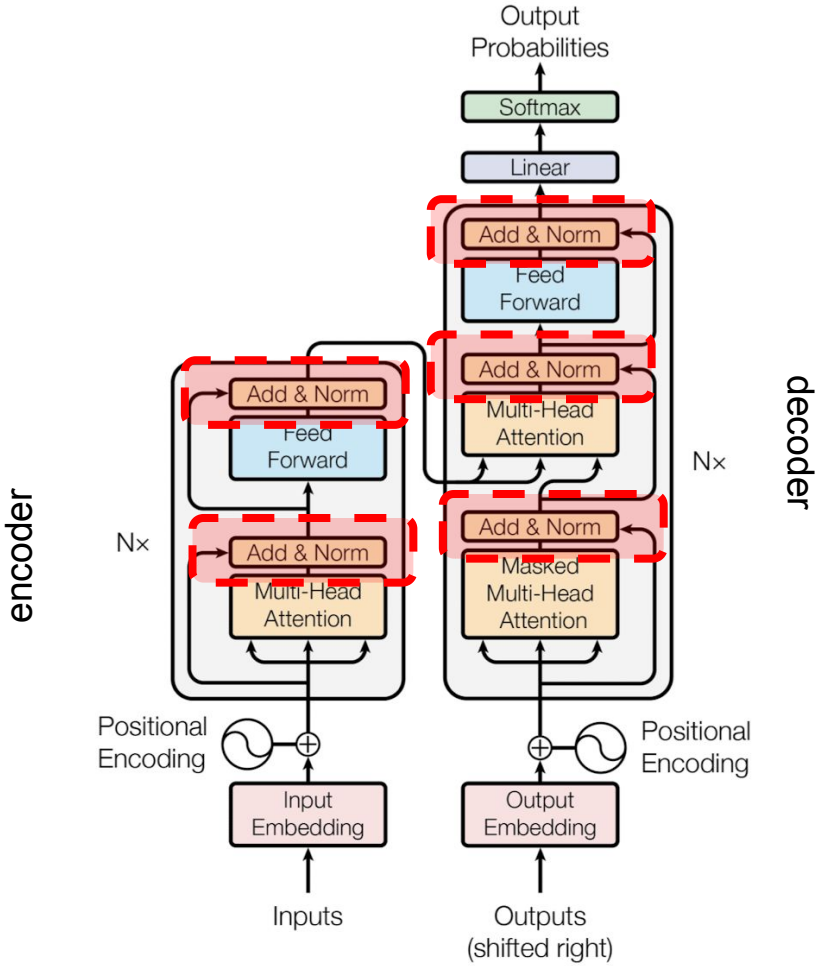


Figure 1: The Transformer - model architecture.

Image from: Attention is all you Need (Vaswani et. al 2017)

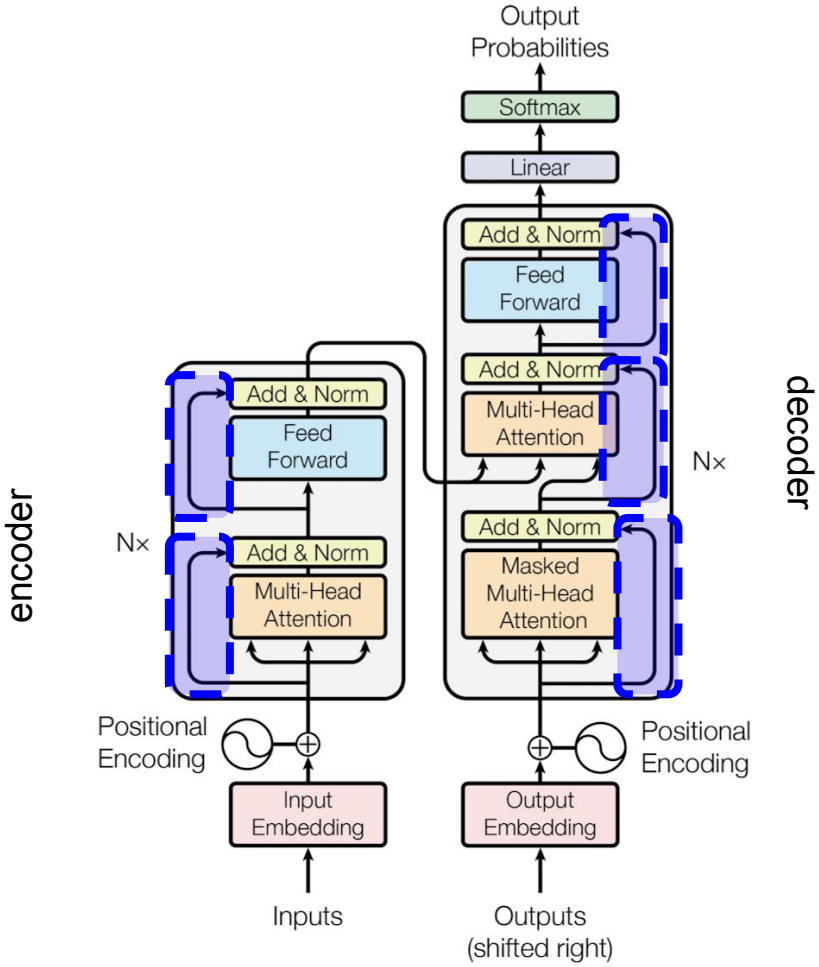


Figure 1: The Transformer - model architecture.

Image from: Attention is all you Need (Vaswani et. al 2017)

Overview

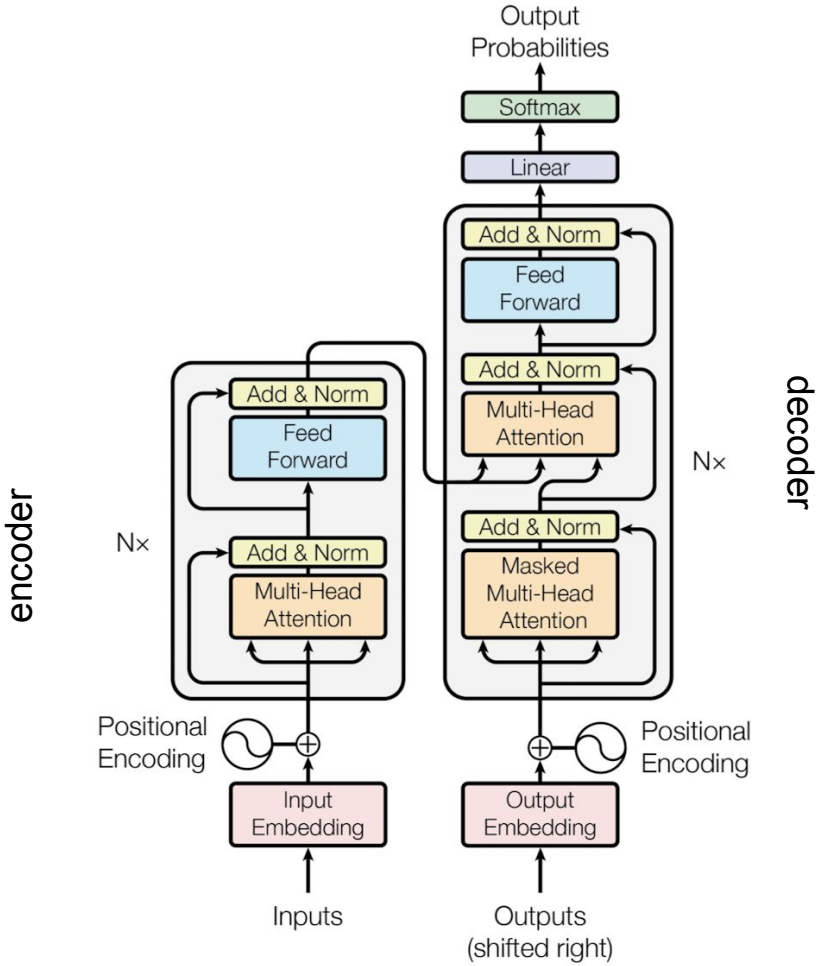


Figure 1: The Transformer - model architecture.

Image from: Attention is all you Need (Vaswani et. al 2017)

Transformer Models You Might Know

- **Encoder:** BERT
- **Decoder:** GPT
- **Encoder/Decoder:** Machine Translation