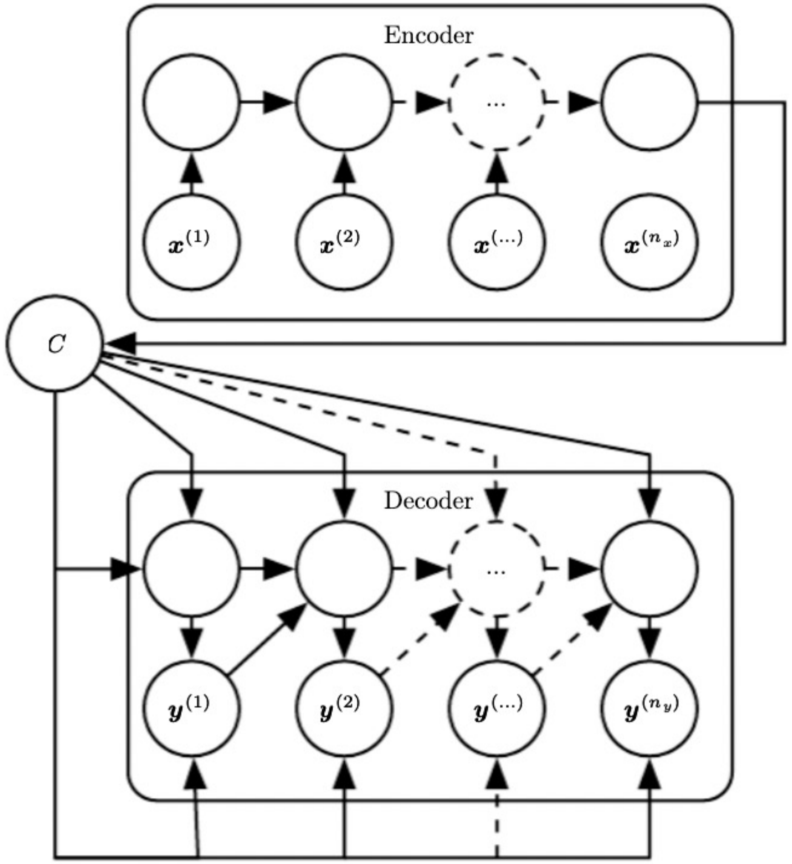
Transformer

Neural Networks Design And Application

Encoder-decoder for seq2seq



Attention mechanism

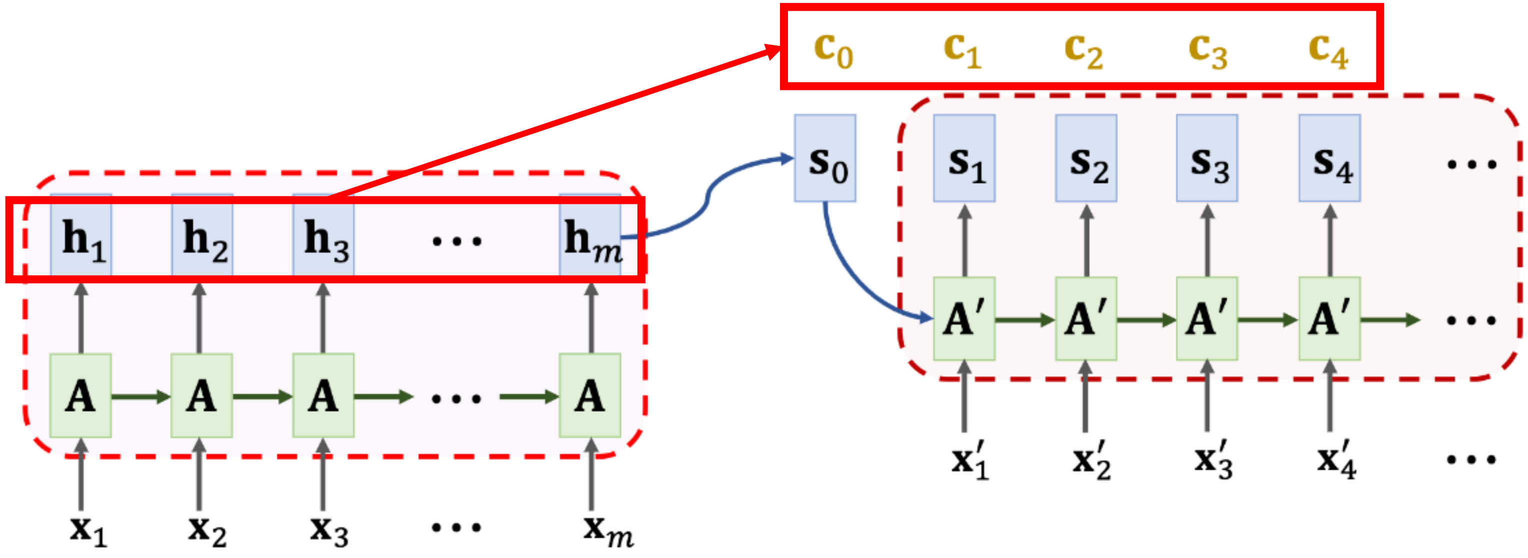


Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_8.pdf

Attention mechanism

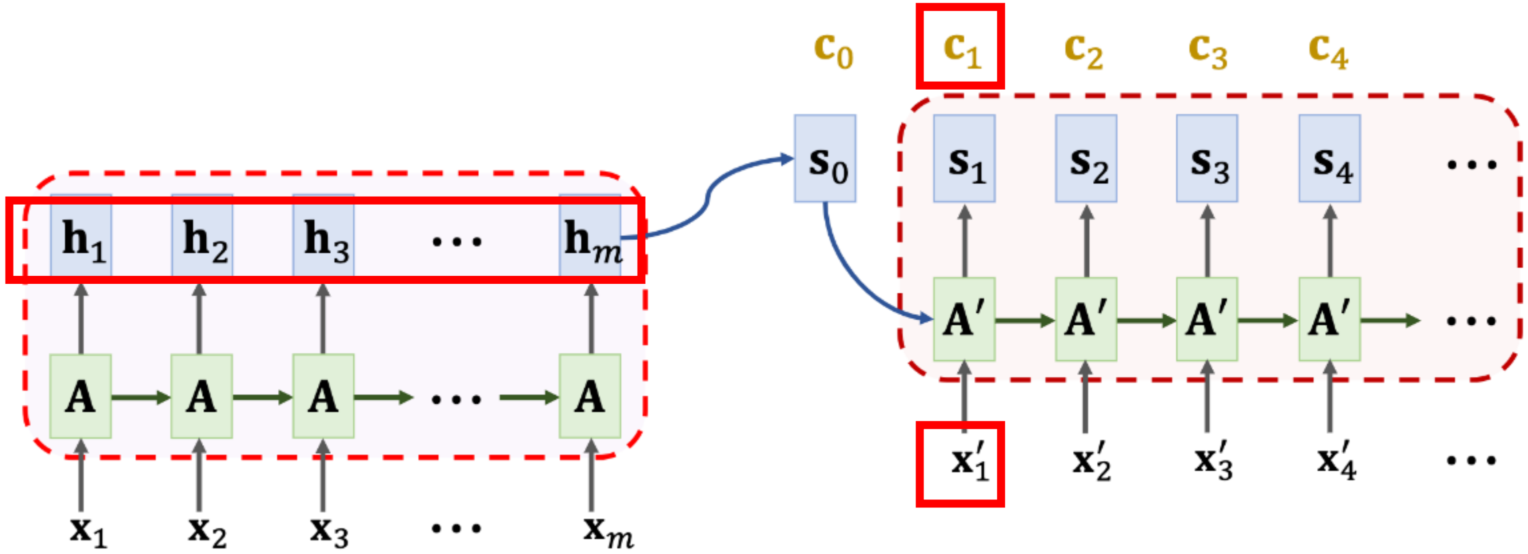


Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_8.pdf

Attention mechanism

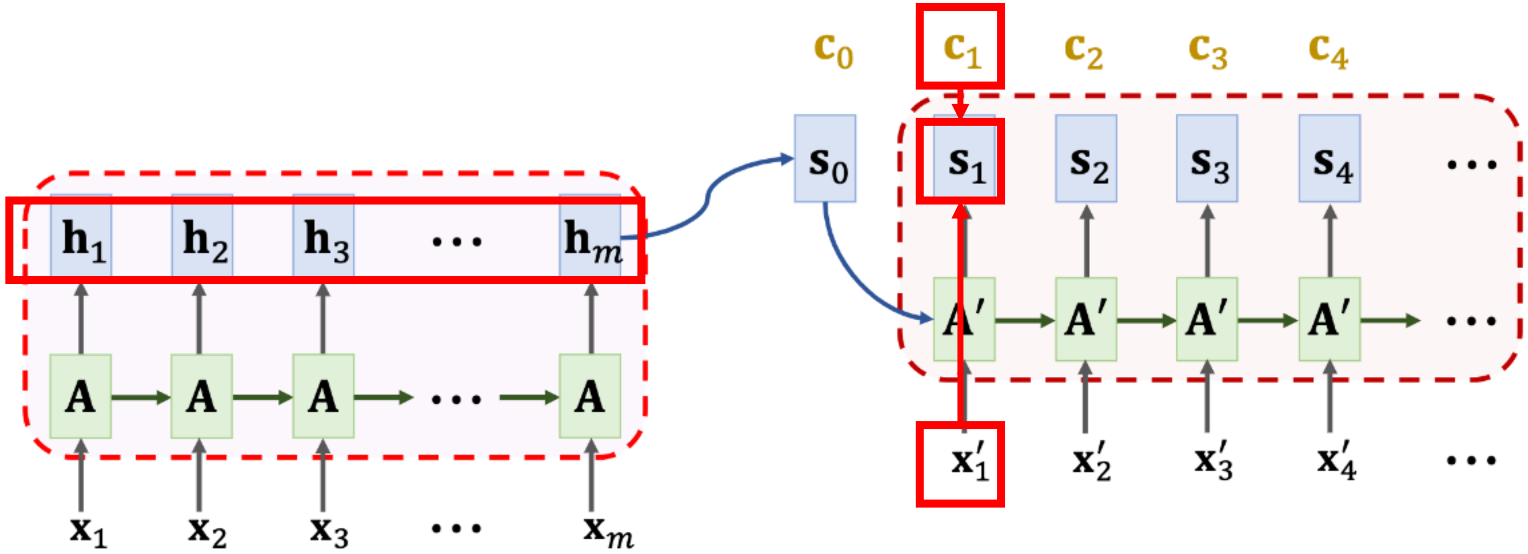


Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_8.pdf

Self-attention (intra-attention)

Weighted average

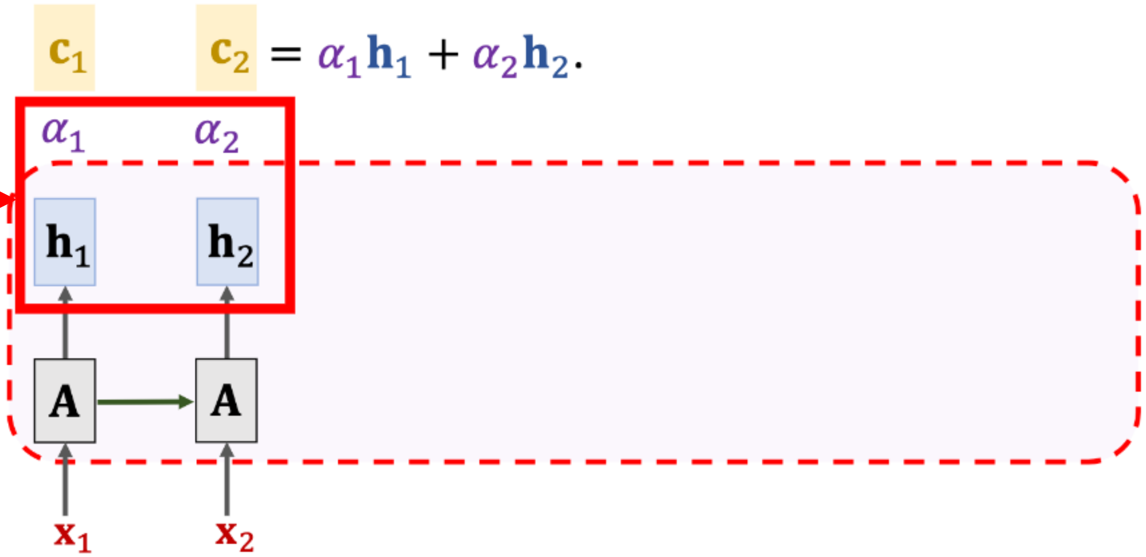


Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_9.pdf

Self-attention (intra-attention)

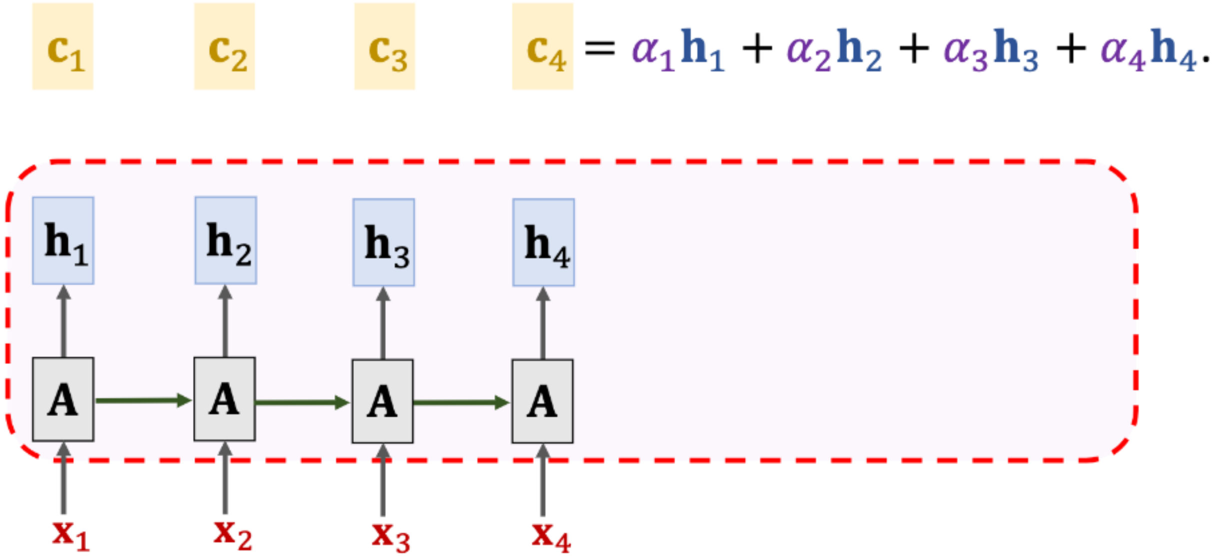


Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_9.pdf

Self-attention (intra-attention)

SimpleRNN:

$$\mathbf{h}_1 = \tanh\left(\mathbf{A} \cdot \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{h}_0 \end{bmatrix} + \mathbf{b}\right)$$

SimpleRNN + Self-Attention:

$$\mathbf{h}_1 = \tanh\left(\mathbf{A} \cdot \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{c}_0 \end{bmatrix} + \mathbf{b}\right)$$

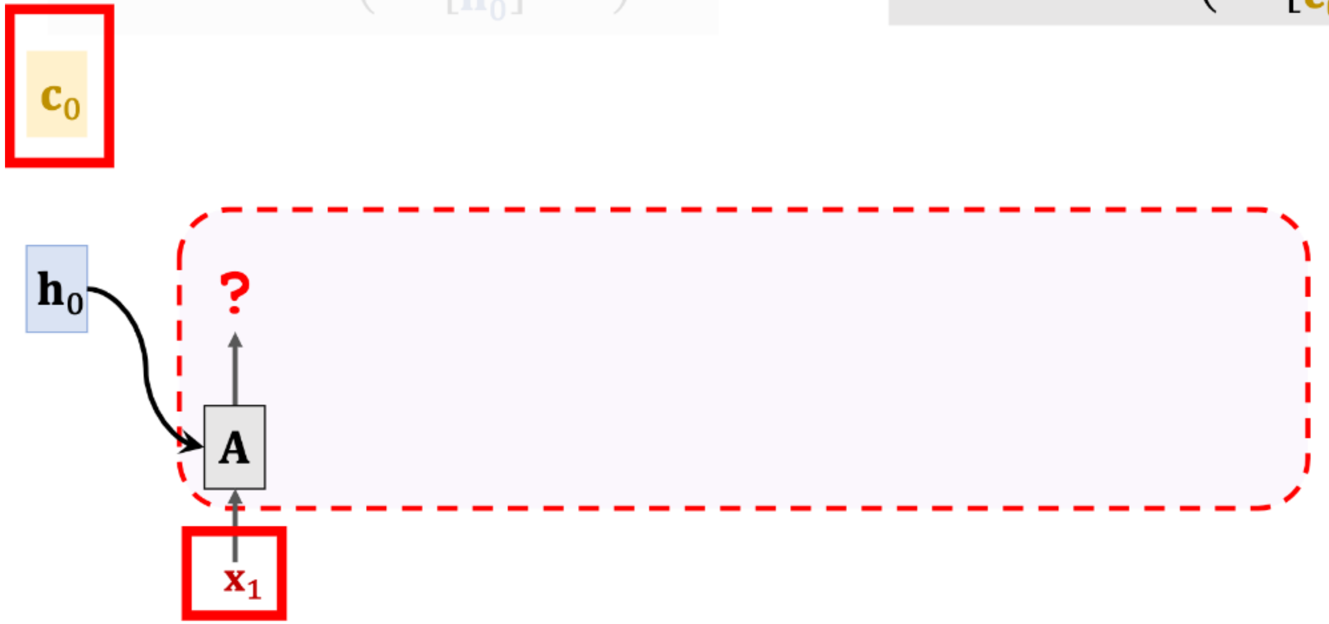


Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_9.pdf

Self-attention (intra-attention)

SimpleRNN:

$$\mathbf{h}_1 = \tanh\left(\mathbf{A} \cdot \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{h}_0 \end{bmatrix} + \mathbf{b}\right)$$

SimpleRNN + Self-Attention:

$$\boxed{\mathbf{h}_1} = \tanh\left(\mathbf{A} \cdot \boxed{\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{c}_0 \end{bmatrix}} + \mathbf{b}\right)$$

Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_9.pdf

Self-attention (intra-attention)

SimpleRNN:

$$\mathbf{h}_1 = \tanh\left(\mathbf{A}\cdot\begin{bmatrix}\mathbf{x}_1\\ \mathbf{h}_0\end{bmatrix}+\mathbf{b}\right)$$

SimpleRNN + Self-Attention:

$$\boxed{\mathbf{h}_1} = \tanh\left(\mathbf{A}\cdot\boxed{\begin{bmatrix}\mathbf{x}_1\\ \mathbf{c}_0\end{bmatrix}}+\mathbf{b}\right)$$

not h_{i-1}

Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_9.pdf

Self-attention (intra-attention)

SimpleRNN:

$$\mathbf{h}_1 = \tanh\left(\mathbf{A} \cdot \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{h}_0 \end{bmatrix} + \mathbf{b}\right)$$

SimpleRNN + Self-Attention:

$$\boxed{\mathbf{h}_1} = \tanh\left(\mathbf{A} \cdot \boxed{\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{c}_0 \end{bmatrix}} + \mathbf{b}\right)$$

not h_{i-1}

Self-attention builds connection over sequence elements

Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_9.pdf

Self-attention (intra-attention)

SimpleRNN:

$$\mathbf{h}_1 = \tanh\left(\mathbf{A} \cdot \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{h}_0 \end{bmatrix} + \mathbf{b}\right)$$

SimpleRNN + Self-Attention:

$$\boxed{\mathbf{h}_1} = \tanh\left(\mathbf{A} \cdot \boxed{\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{c}_0 \end{bmatrix}} + \mathbf{b}\right)$$

not h_{i-1}

Self-attention builds connection over sequence elements

Q: can we use self-attention without RNN?

Image from https://github.com/wangshusen/DeepLearning/blob/master/Slides/9_RNN_9.pdf

Transformer networks

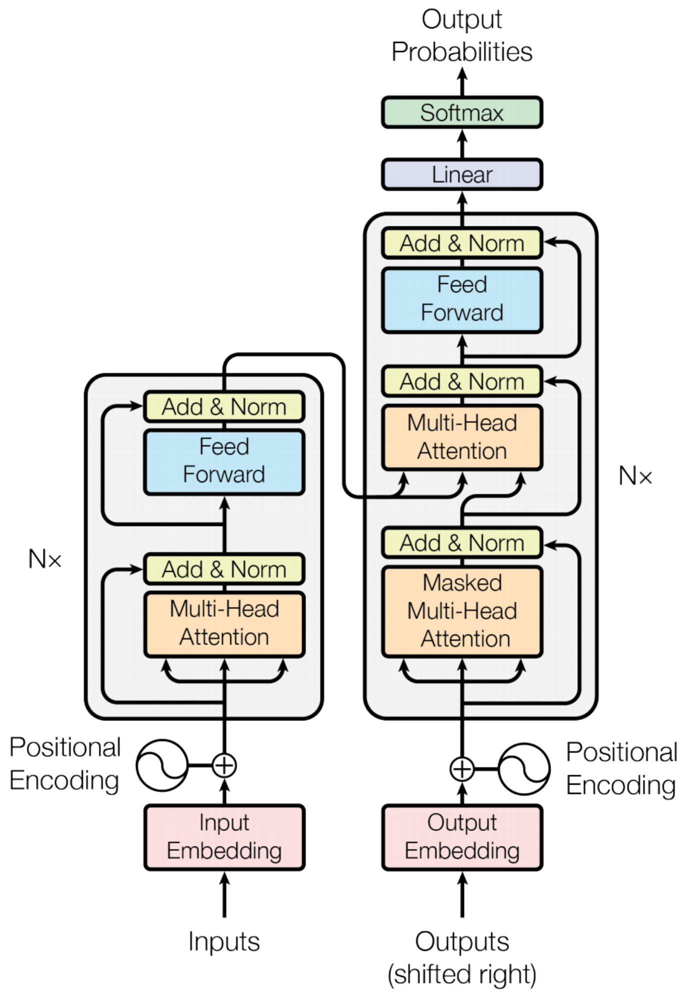


Figure 1: The Transformer - model architecture.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *arXiv preprint arXiv:1706.03762* (2017).

Transformer networks

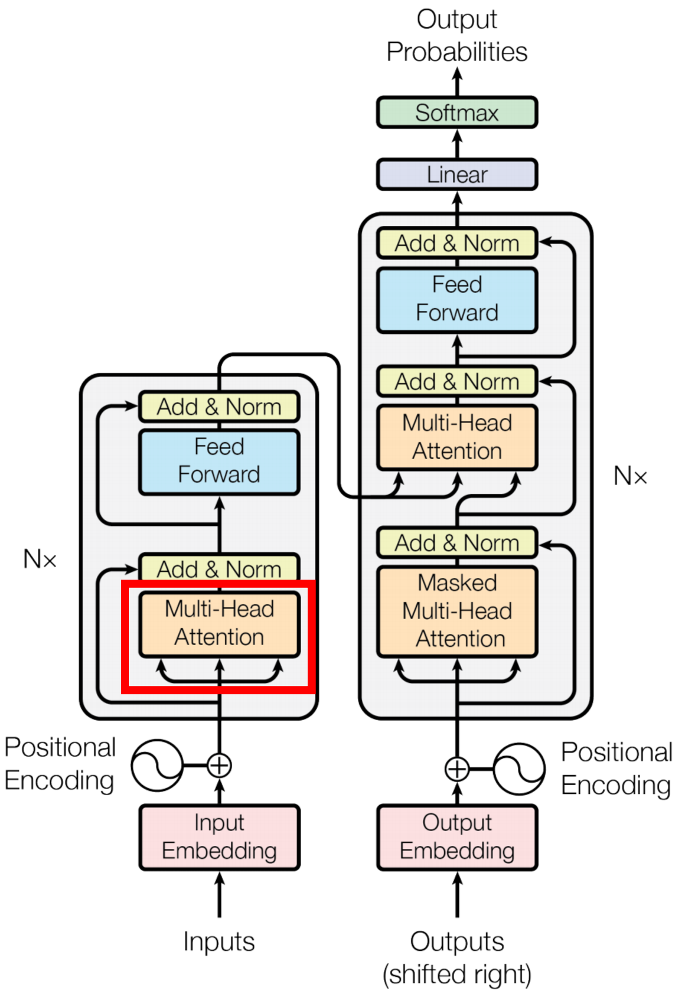


Figure 1: The Transformer - model architecture.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *arXiv preprint arXiv:1706.03762* (2017).

Transformer networks

3 inputs
1 output

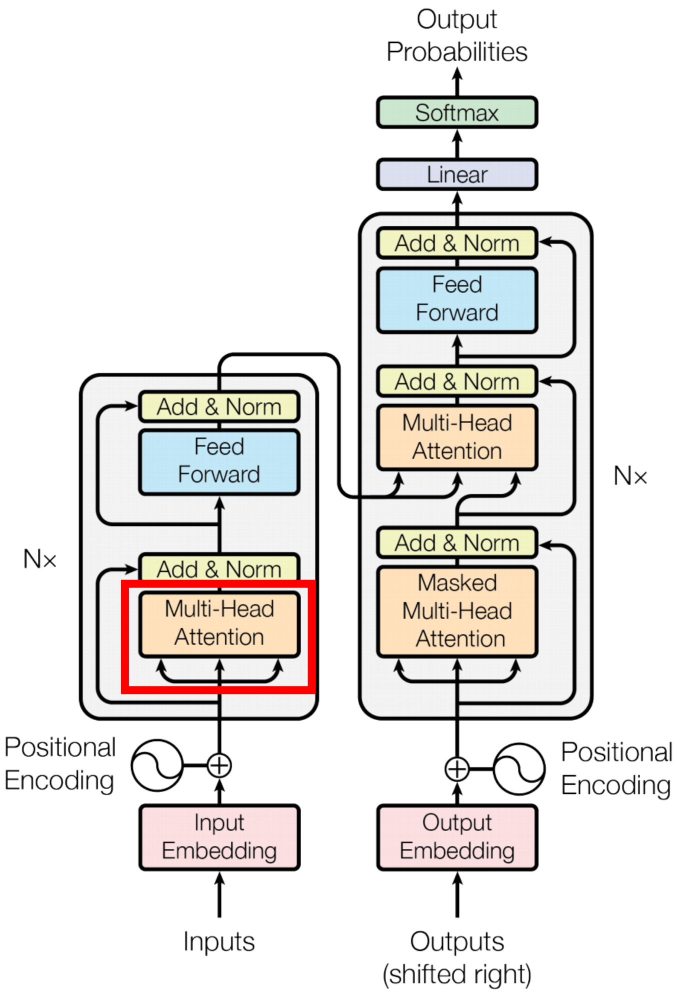
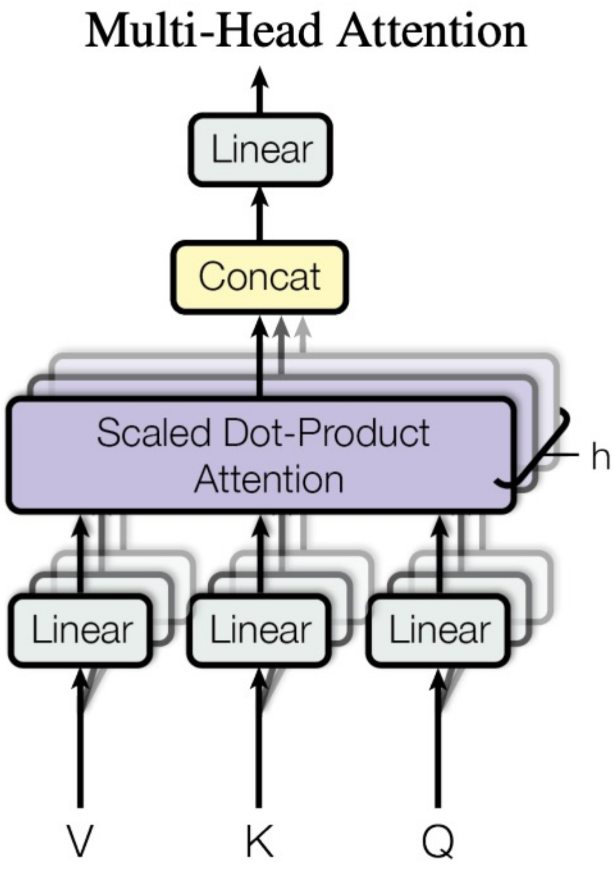


Figure 1: The Transformer - model architecture.

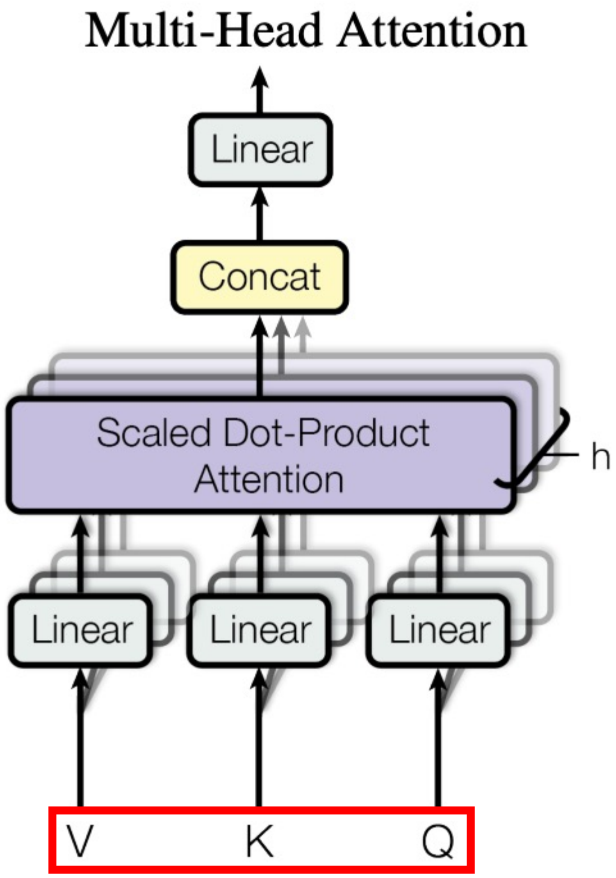
Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *arXiv preprint arXiv:1706.03762* (2017).

Transformer networks



Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *arXiv preprint arXiv:1706.03762* (2017).

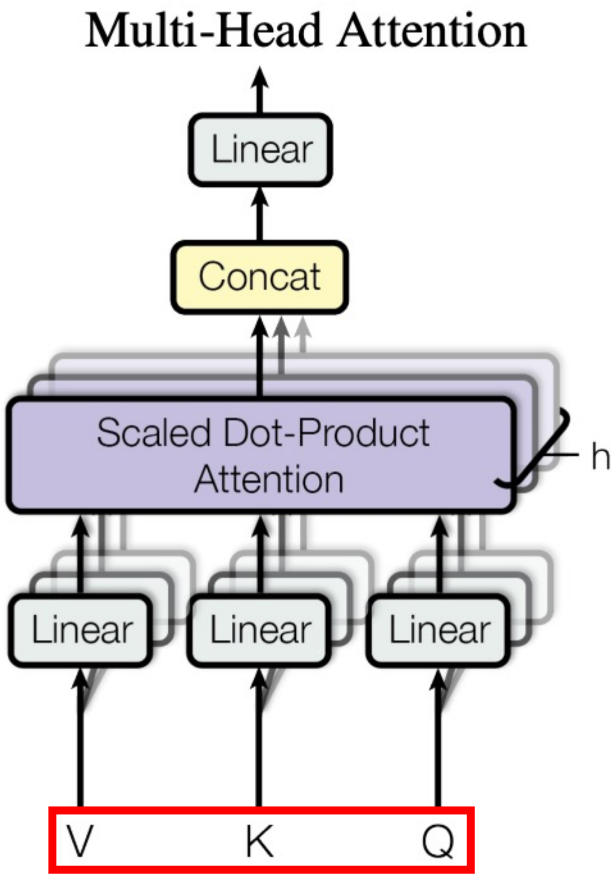
Transformer networks



V: value
K: key
Q: query

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *arXiv preprint arXiv:1706.03762* (2017).

Transformer networks



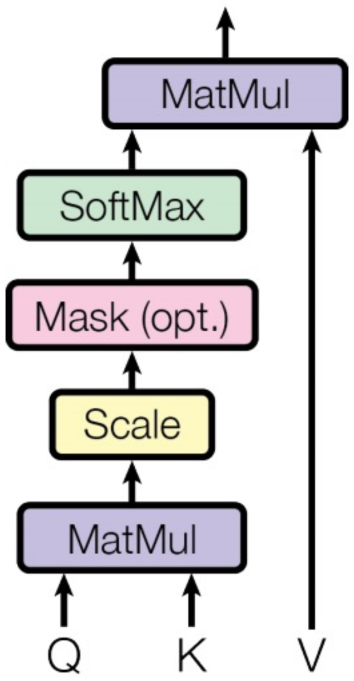
V: value
K: key
Q: query

Interpreted from information retrieval systems: https://stats.stackexchange.com/questions/421935/what-exactly-are-keys-queries-and-values-in-attention-mechanisms

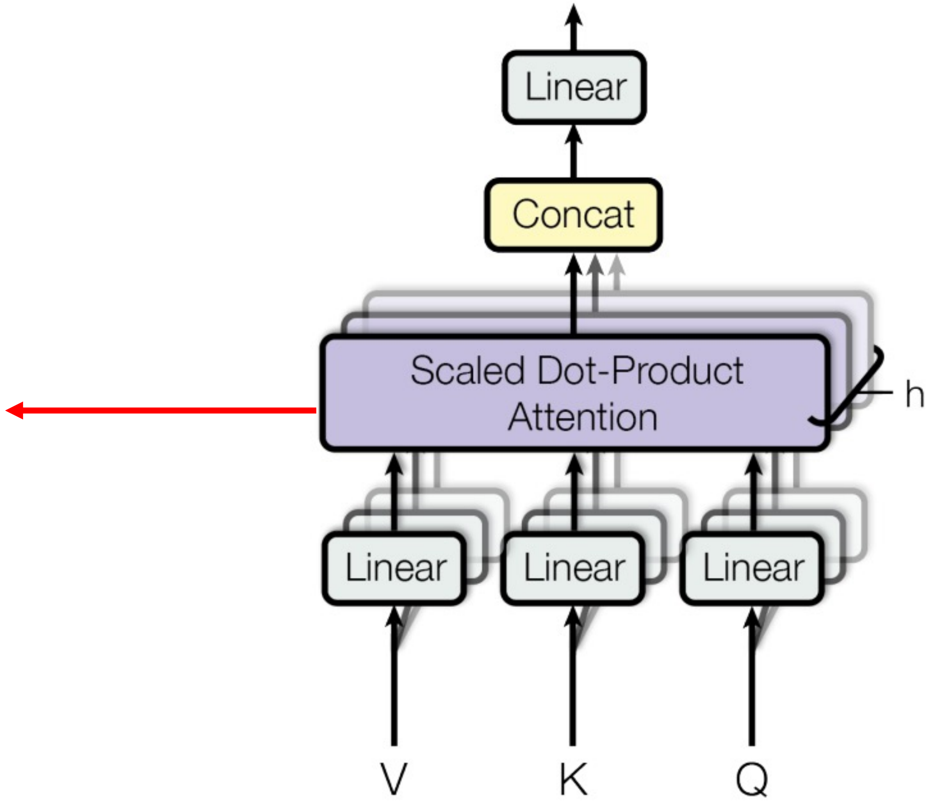
Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *arXiv preprint arXiv:1706.03762* (2017).

Transformer networks

Scaled Dot-Product Attention

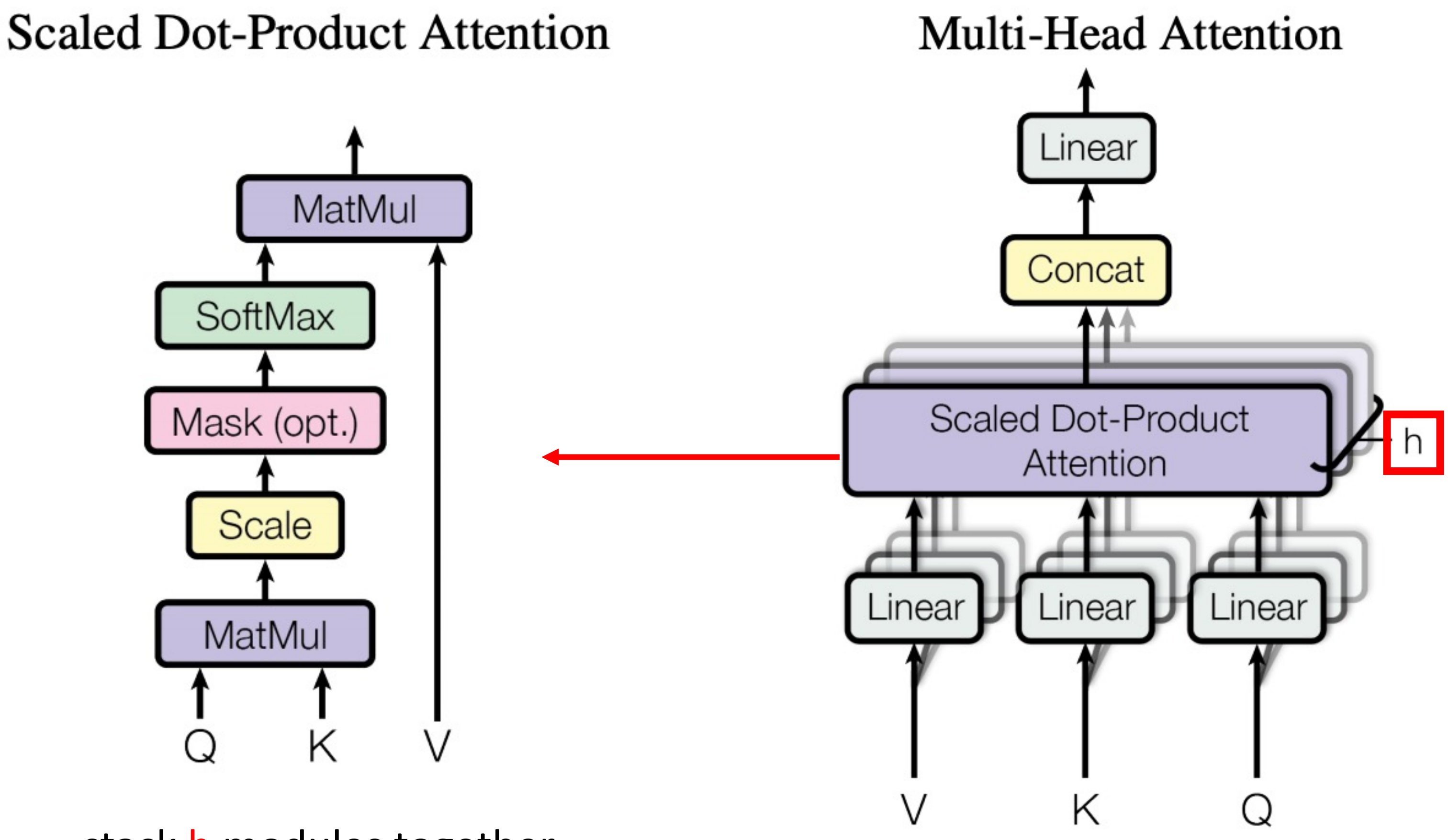


Multi-Head Attention



Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *arXiv preprint arXiv:1706.03762* (2017).

Transformer networks

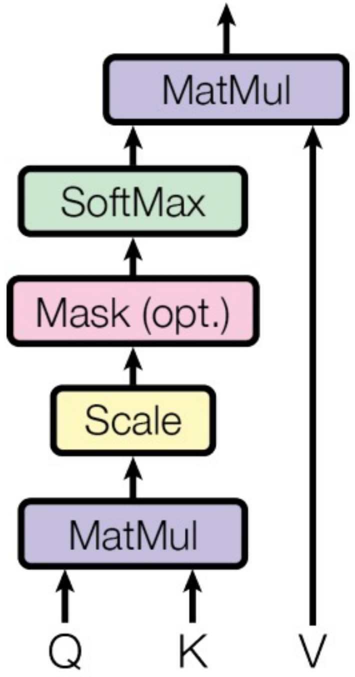


stack h modules together

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *arXiv preprint arXiv:1706.03762* (2017).

Transformer networks

Scaled Dot-Product Attention

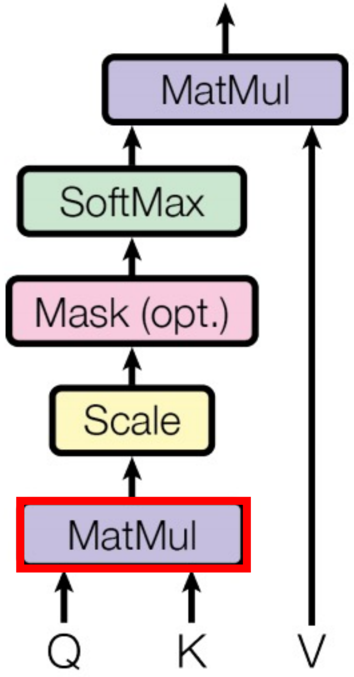


$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *arXiv preprint arXiv:1706.03762* (2017).

Transformer networks

Scaled Dot-Product Attention

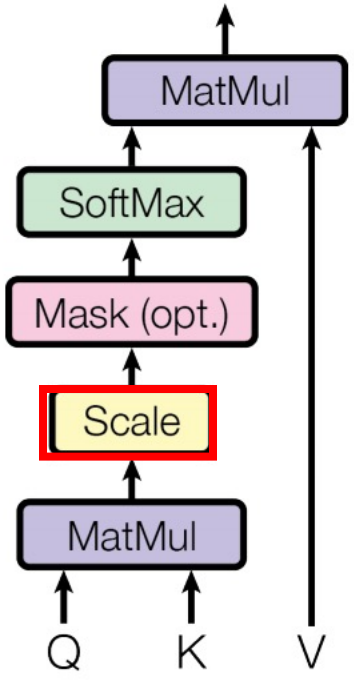


$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{\boxed{QK^T}}{\sqrt{d_k}}\right)V$$

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *arXiv preprint arXiv:1706.03762* (2017).

Transformer networks

Scaled Dot-Product Attention

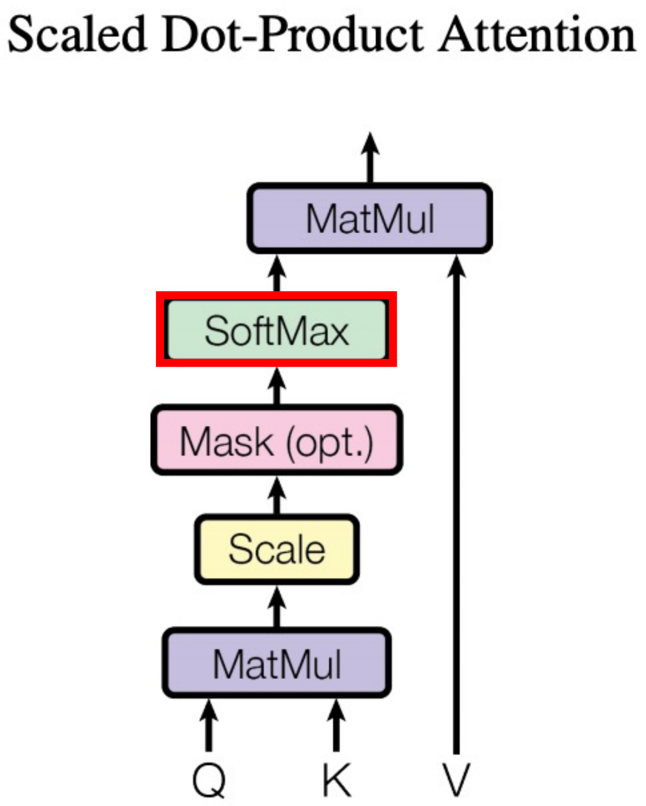


$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}(\frac{QK^T}{\boxed{\sqrt{d_k}}})V$$

Dimension of keys

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *arXiv preprint arXiv:1706.03762* (2017).

Transformer networks

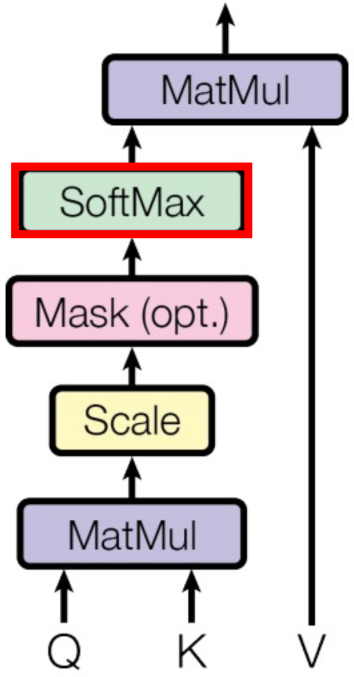


$$\mathrm{Attention}(Q,K,V) = \boxed{\mathrm{softmax}}(\frac{QK^T}{\sqrt{d_k}})V$$

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *arXiv preprint arXiv:1706.03762* (2017).

Transformer networks

Scaled Dot-Product Attention



weights for selection (keys in a retrieval system)

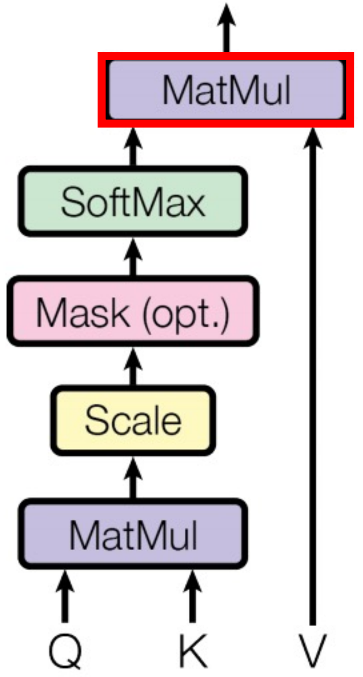
$$\mathrm{Attention}(Q,K,V) = \boxed{\mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})}V$$

An example: one-hot format

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *arXiv preprint arXiv:1706.03762* (2017).

Transformer networks

Scaled Dot-Product Attention

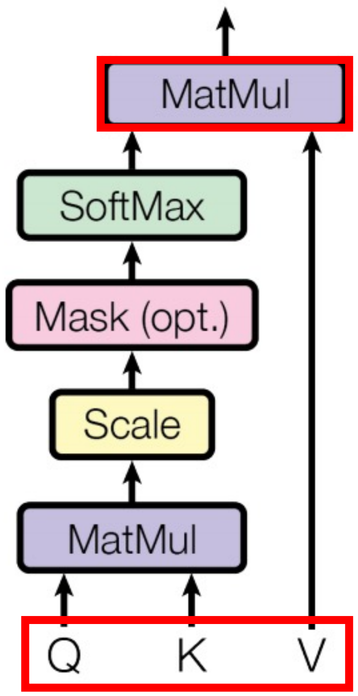


$$\mathrm{Attention}(Q,K,V) = \boxed{\mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V}$$

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *arXiv preprint arXiv:1706.03762* (2017).

Transformer networks

Scaled Dot-Product Attention



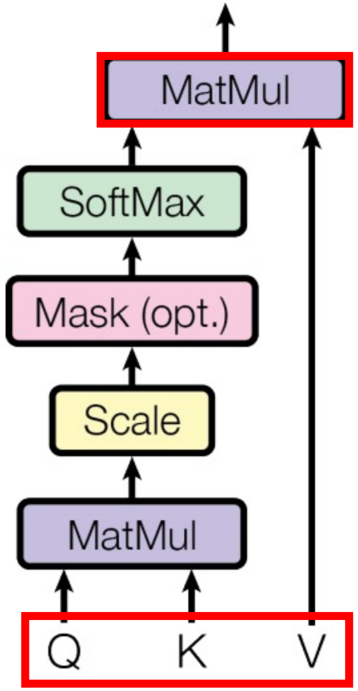
$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

How to generate

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *arXiv preprint arXiv:1706.03762* (2017).

Transformer networks

Scaled Dot-Product Attention



In addition to attention sub-layers, each of the layers in our encoder and decoder contains a fully connected feed-forward network, which is applied to each position separately and identically. This consists of two linear transformations with a ReLU activation in between.

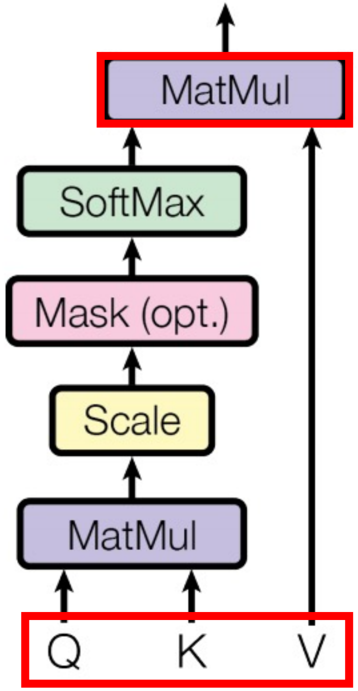
$$\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{2}$$

While the linear transformations are the same across different positions, they use different parameters from layer to layer. Another way of describing this is as two convolutions with kernel size 1. The dimensionality of input and output is $d_{\text{model}} = 512$, and the inner-layer has dimensionality $d_{ff} = 2048$.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *arXiv preprint arXiv:1706.03762* (2017).

Transformer networks

Scaled Dot-Product Attention



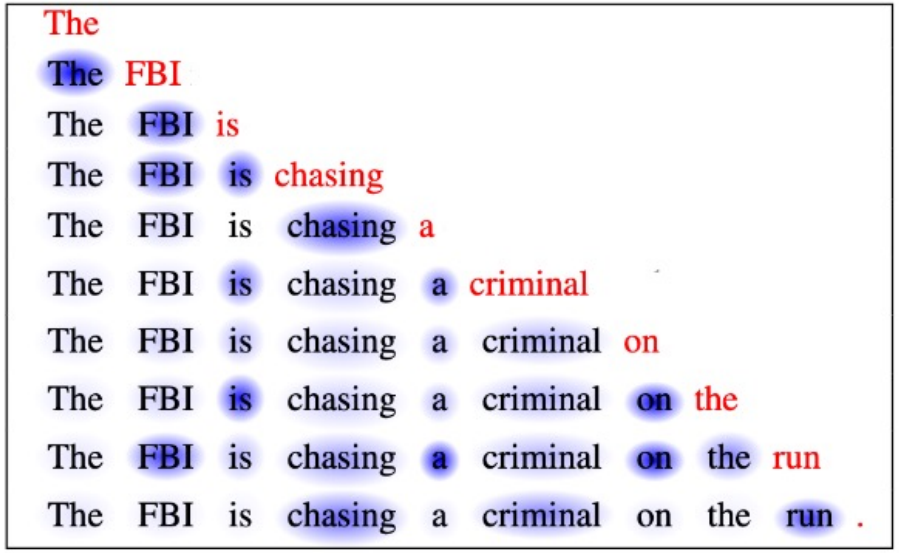
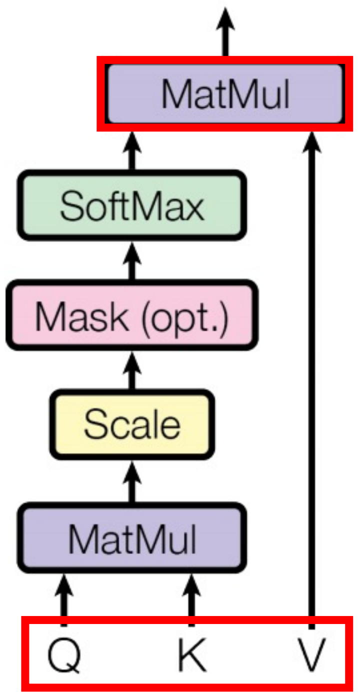


Figure is from the paper " Long Short-Term Memory-Networks for Machine Reading."

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *arXiv preprint arXiv:1706.03762* (2017).

Transformer networks

Scaled Dot-Product Attention



Greater weights (larger Softmax output)

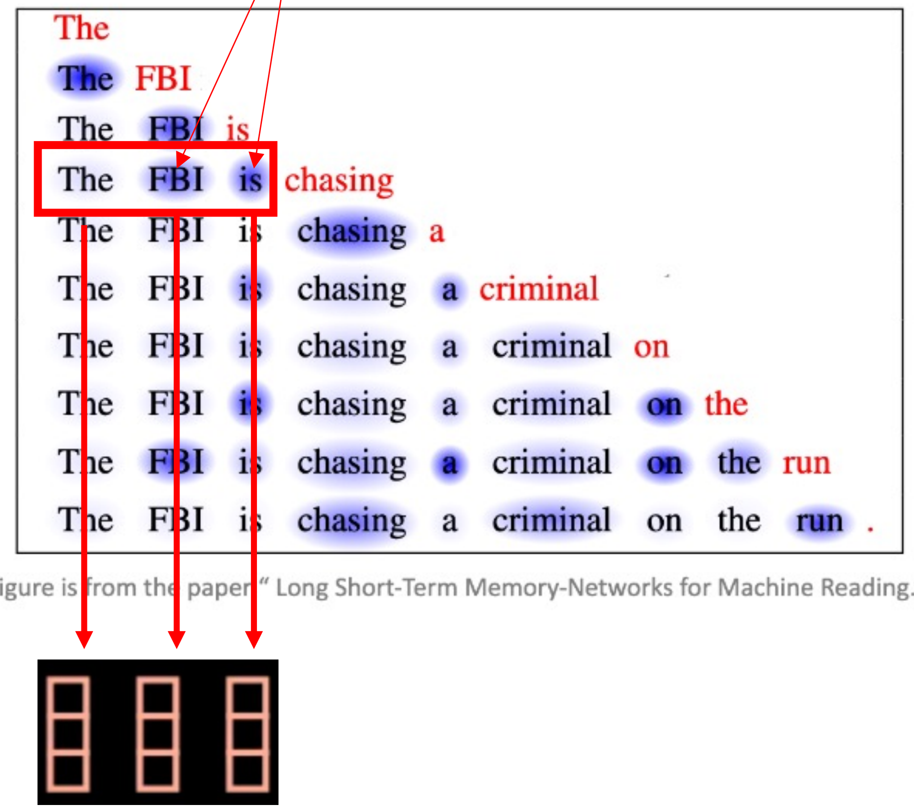


Figure is from the paper " Long Short-Term Memory-Networks for Machine Reading."

Embedding feature vectors (a matrix)

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *arXiv preprint arXiv:1706.03762* (2017).

Transformer networks

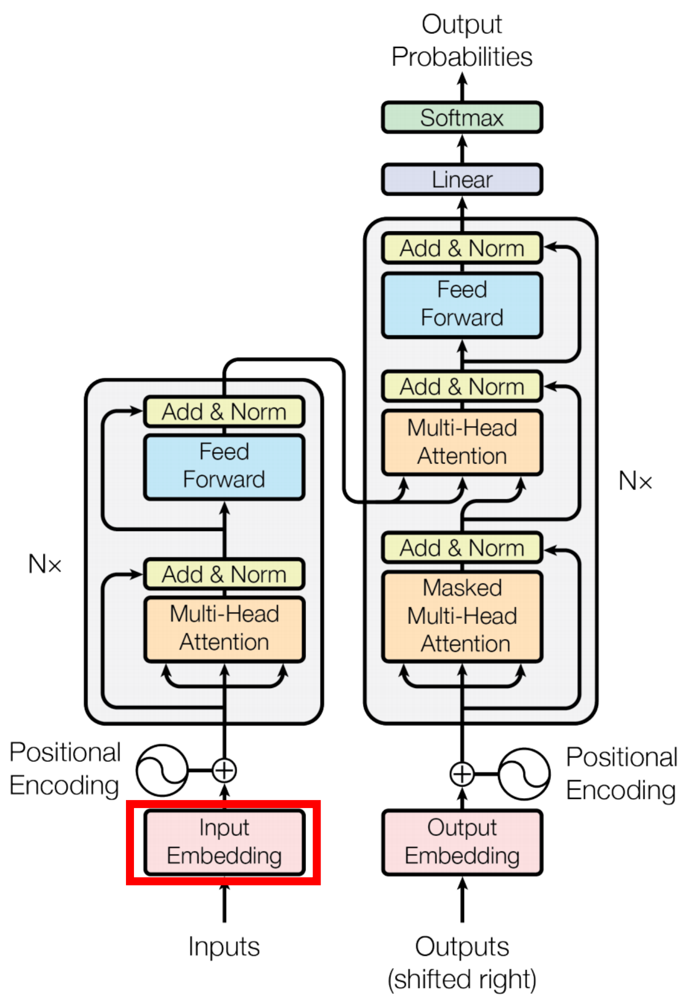


Figure 1: The Transformer - model architecture.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *arXiv preprint arXiv:1706.03762* (2017).

Transformer networks

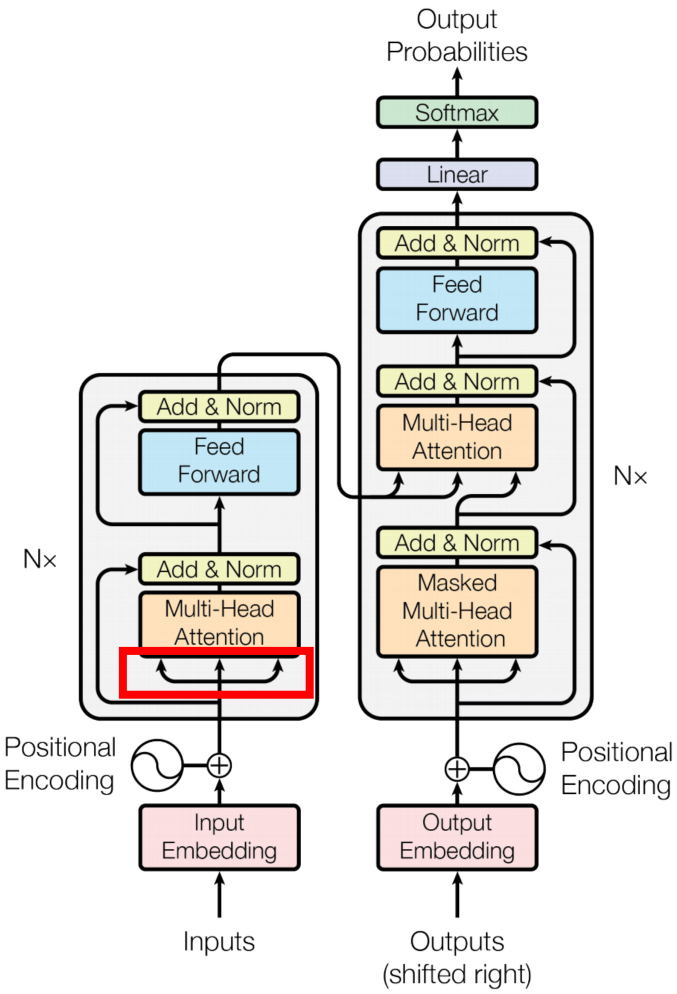


Figure 1: The Transformer - model architecture.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *arXiv preprint arXiv:1706.03762* (2017).

Transformer networks

Shortcut connection

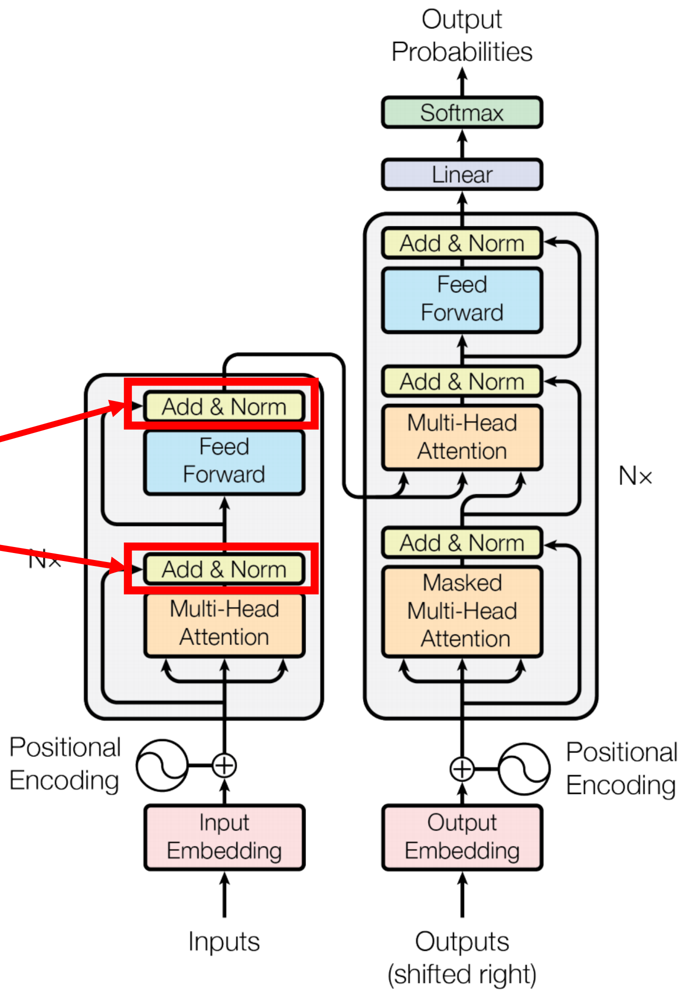


Figure 1: The Transformer - model architecture.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *arXiv preprint arXiv:1706.03762* (2017).

Transformer networks

Shortcut connection
(residual connection)

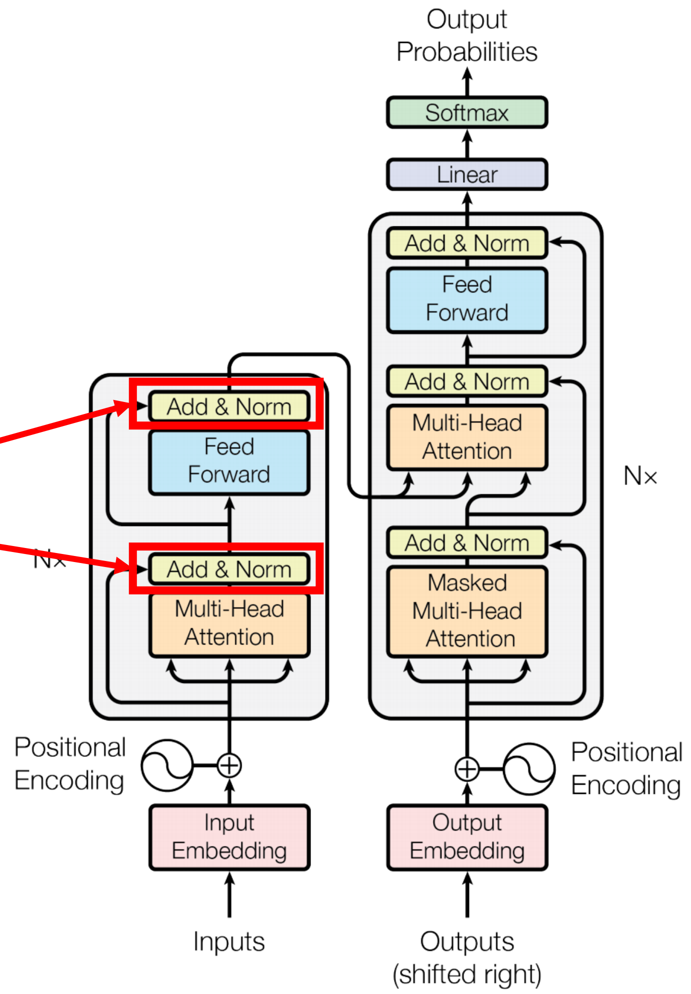


Figure 1: The Transformer - model architecture.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *arXiv preprint arXiv:1706.03762* (2017).

Transformer networks

No recurrent structure → not RNN

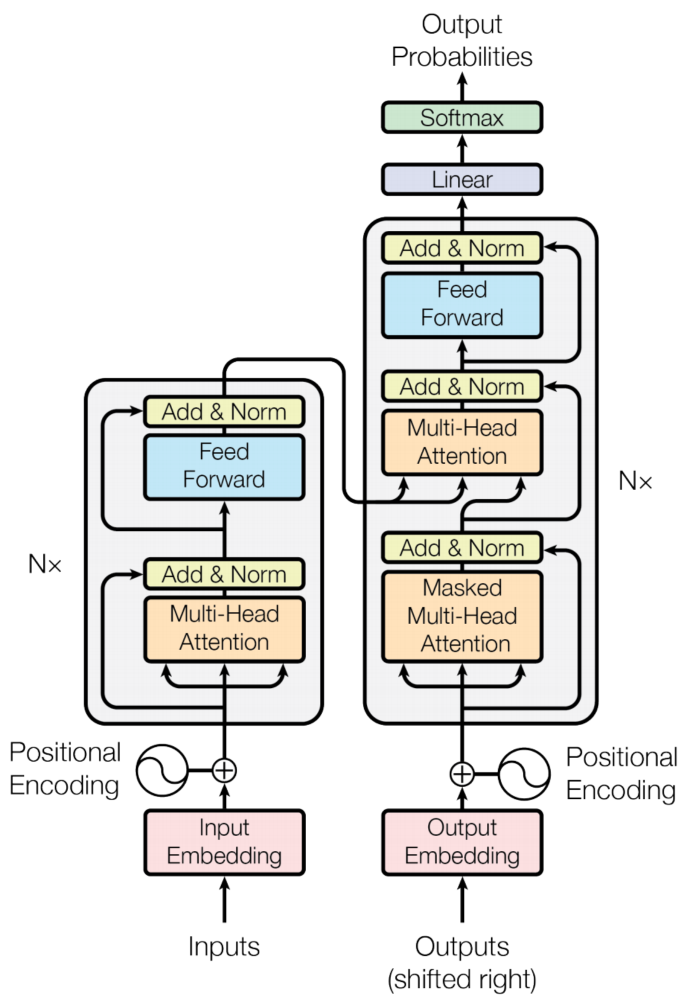


Figure 1: The Transformer - model architecture.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *arXiv preprint arXiv:1706.03762* (2017).

Transformer networks

No recurrent structure → not RNN

Q: any benefit?

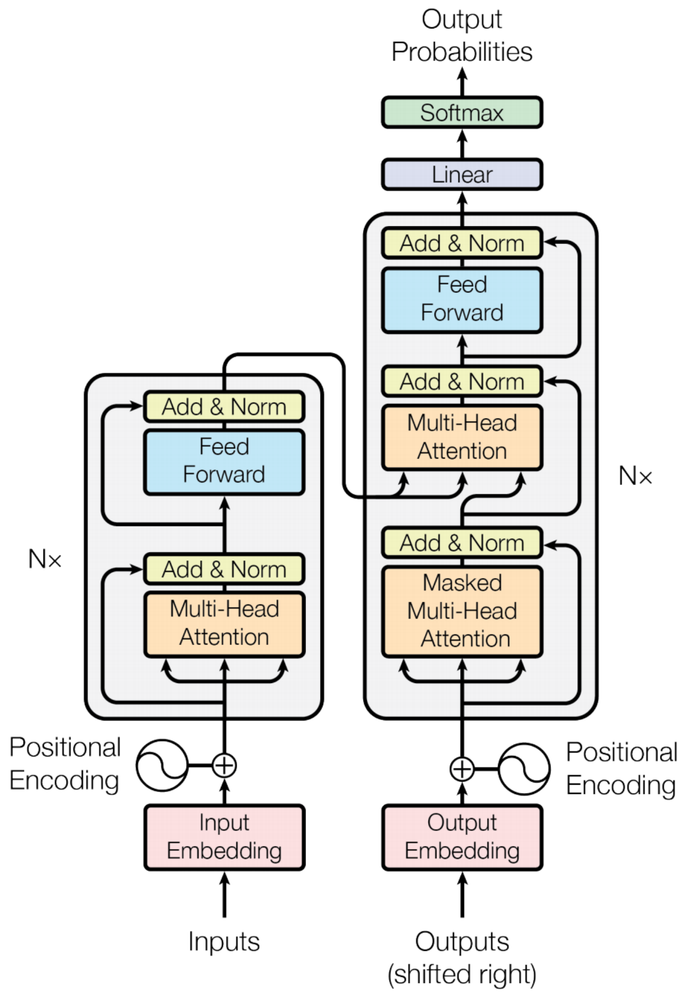


Figure 1: The Transformer - model architecture.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *arXiv preprint arXiv:1706.03762* (2017).

Transformer networks

No recurrent structure → not RNN

Q: any benefit?

1. No sequential dependence

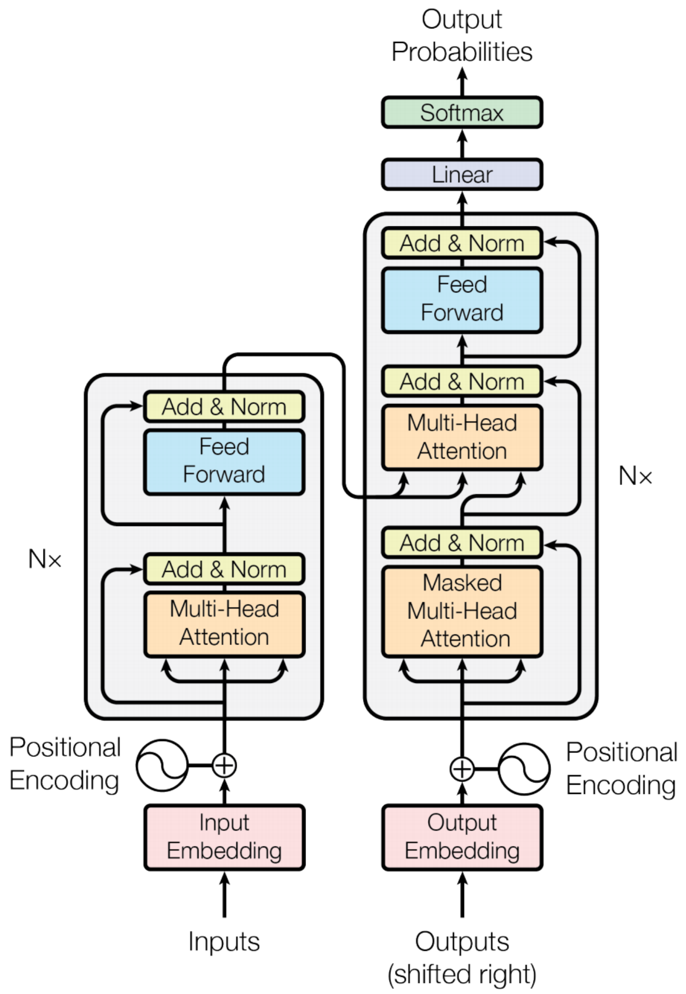


Figure 1: The Transformer - model architecture.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *arXiv preprint arXiv:1706.03762* (2017).

Transformer networks

No recurrent structure → not RNN

Q: any benefit?

1. No sequential dependence
2. Parallel processing

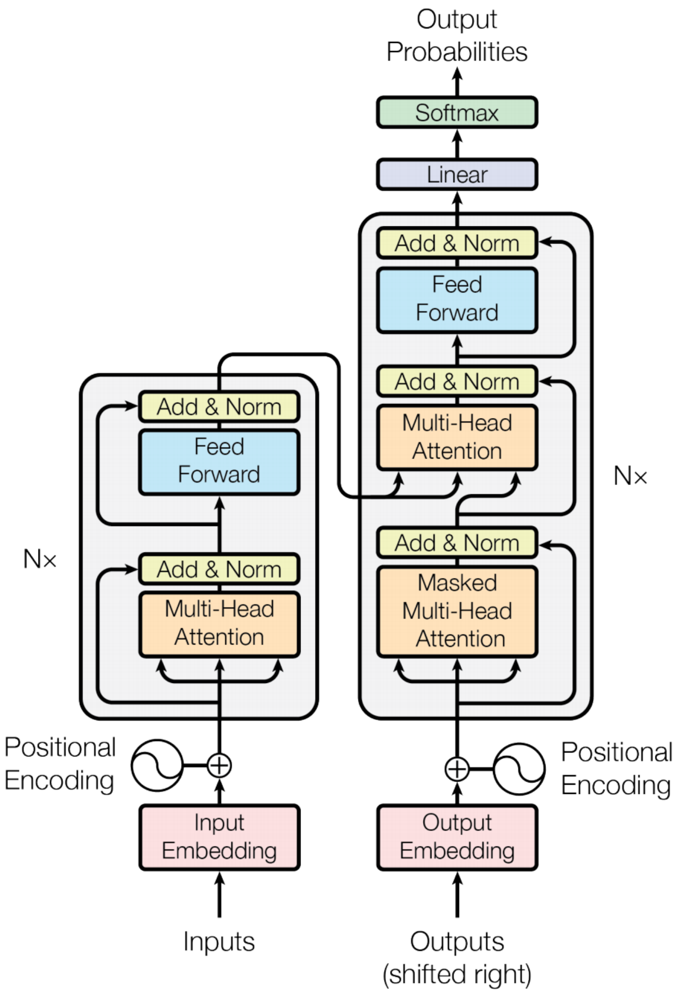


Figure 1: The Transformer - model architecture.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *arXiv preprint arXiv:1706.03762* (2017).

Transformer networks

No recurrent structure → not RNN

Q: any benefit?

1. No sequential dependence
2. Parallel processing
3. …

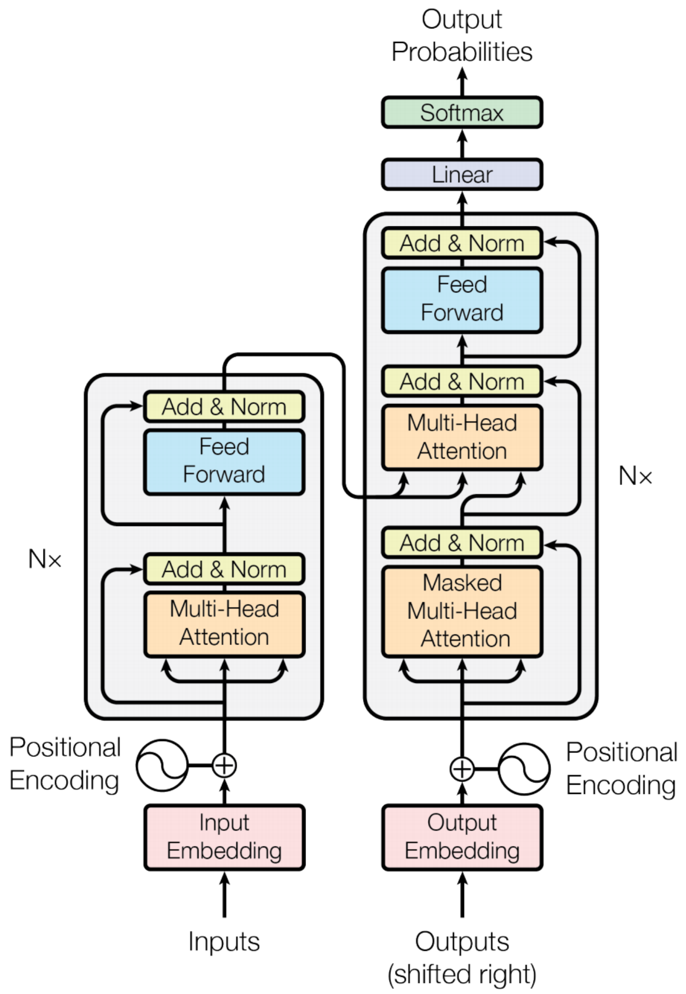


Figure 1: The Transformer - model architecture.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *arXiv preprint arXiv:1706.03762* (2017).