

ATTENTION IS ALL YOU NEED

完全基于注意力的网络 (变形金刚)

北京大学 中文信息处理 唐乾桐



为什么选择报告这篇论文

Why this paper?

为什么选择报告这篇论文

Why this paper?

思考

为什么选择报告这篇论文

Why this paper?

思考：语言学在基于神经网络的nlp研究中能做什么？

为什么选择报告这篇论文

Why this paper?

思考：语言学在基于神经网络的nlp研究中能做什么？

评测

学到了哪些语言学知识？
语言学导向的评测方法和评测数据集

建模

直接参与模型的研发
(但似乎还没有先例)

解释

模型的可解释性研究

为什么选择报告这篇论文

Why this paper?

评测

学到了哪些语言学知识？

语言学导向的评测方法和评测数据集

eg.

Analogical Reasoning

Challenge Set

-
-
-

建模

为什么选择报告这篇论文

Why this paper?



建模

直接参与模型的研发



解释

模型的可解释性研究

为什么选择报告这篇论文

Why this paper?

建模

解释

都需要

对模型本身有一定深度的了解

为什么选择报告这篇论文

Why this paper?

Attention is all you need

2017.06

Google

为什么选择报告这篇论文

Why this paper?

Attention is all you need

2017.06

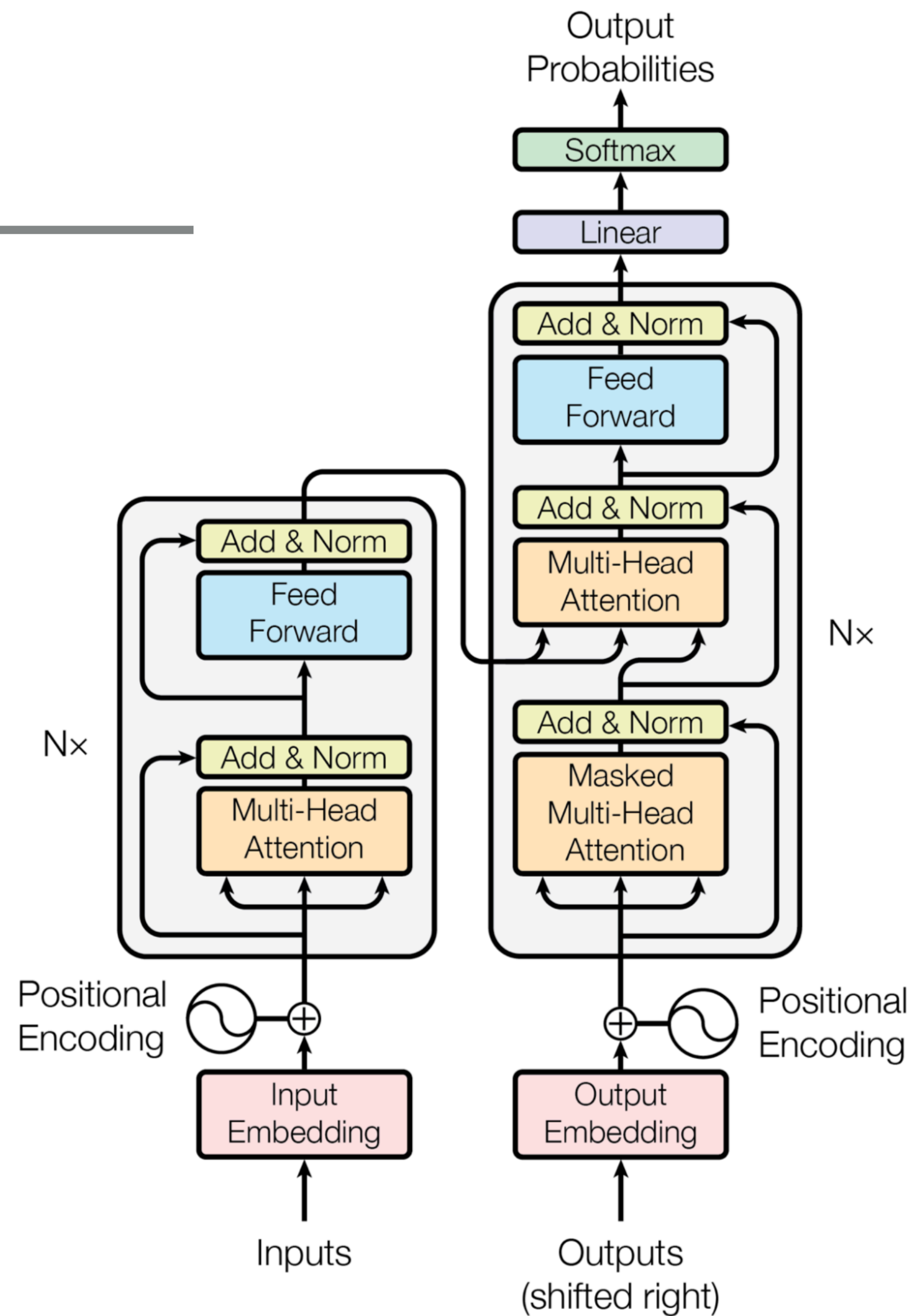
Google

论文模型

Overview

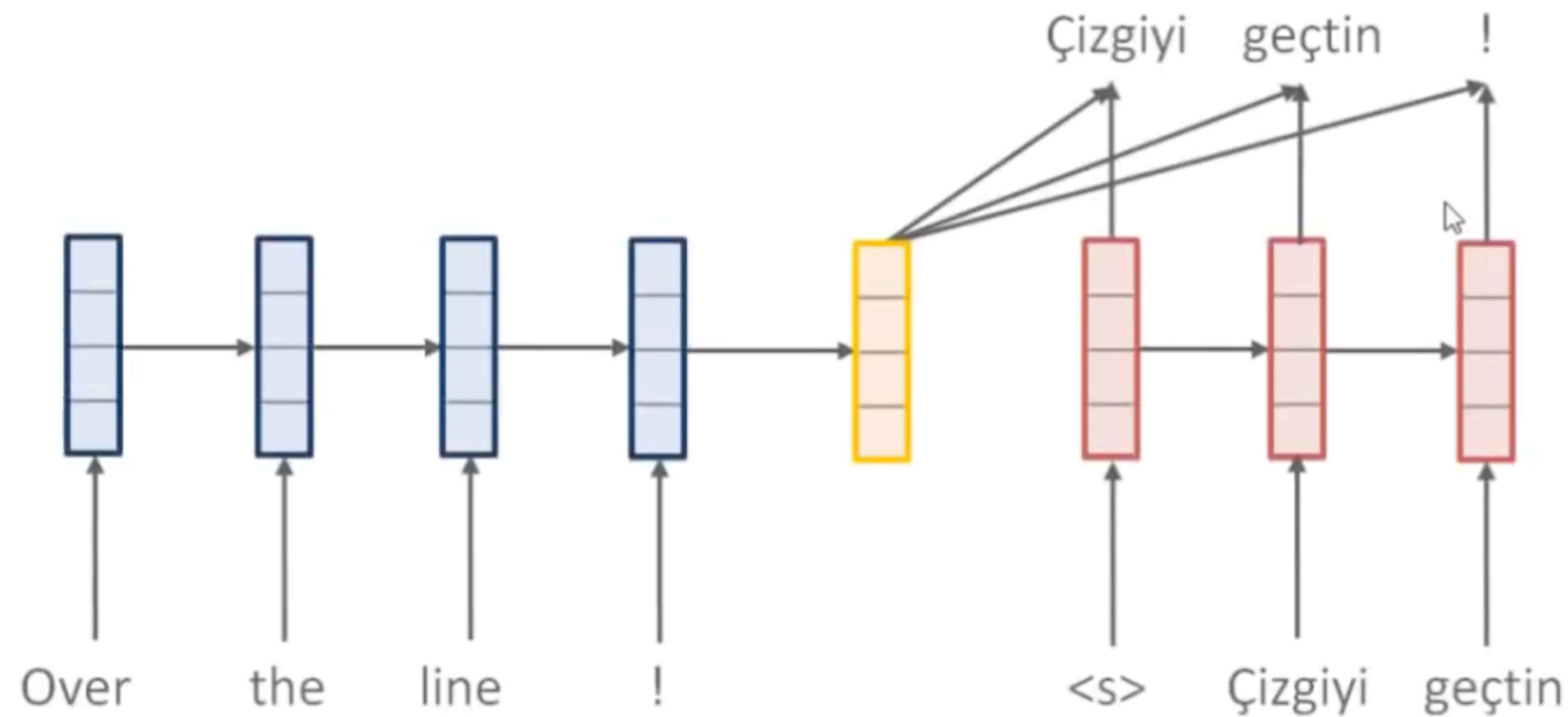
论文模型

Overview

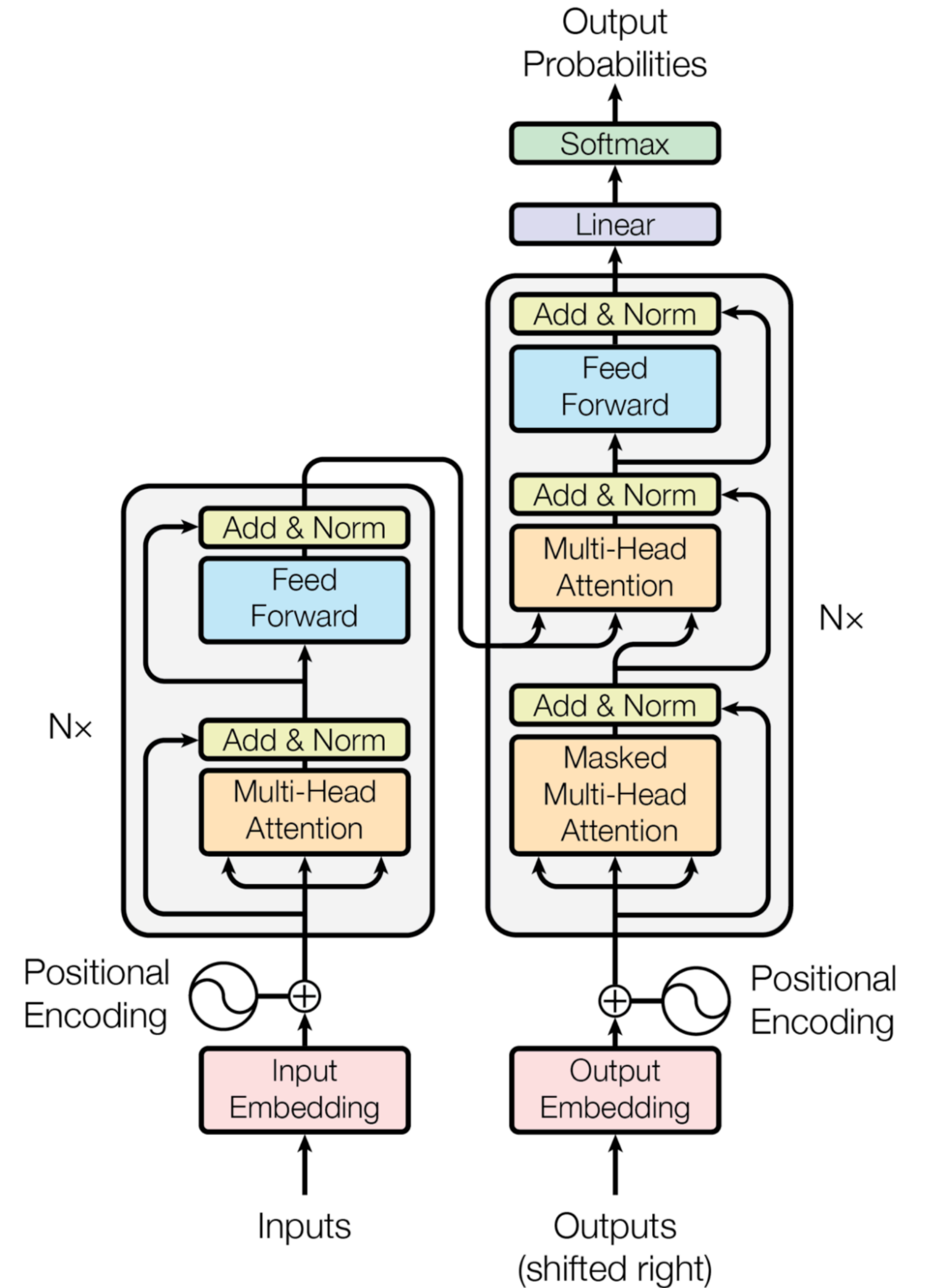


论文模型

Overview

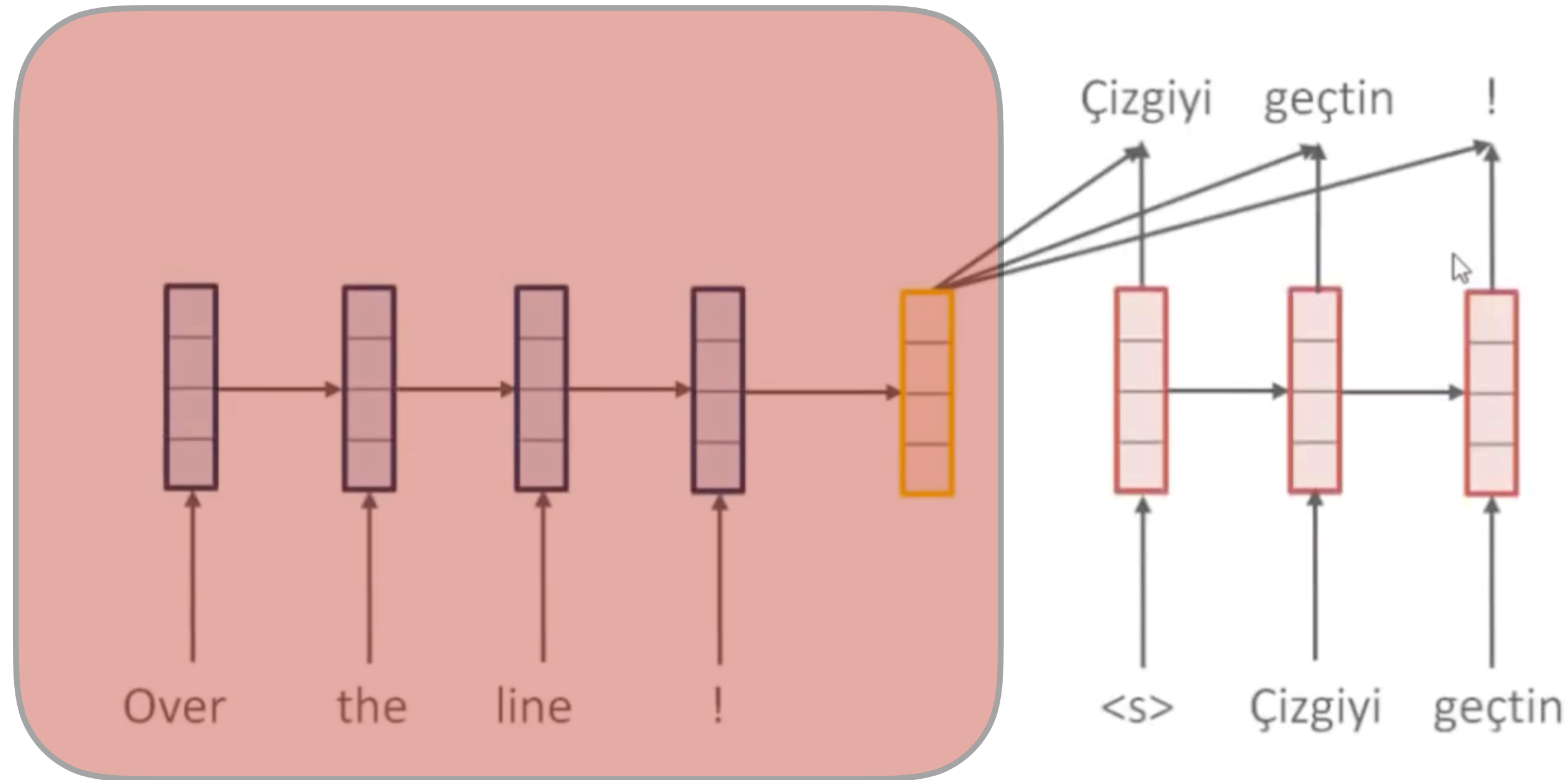


RNN encoder-decoder

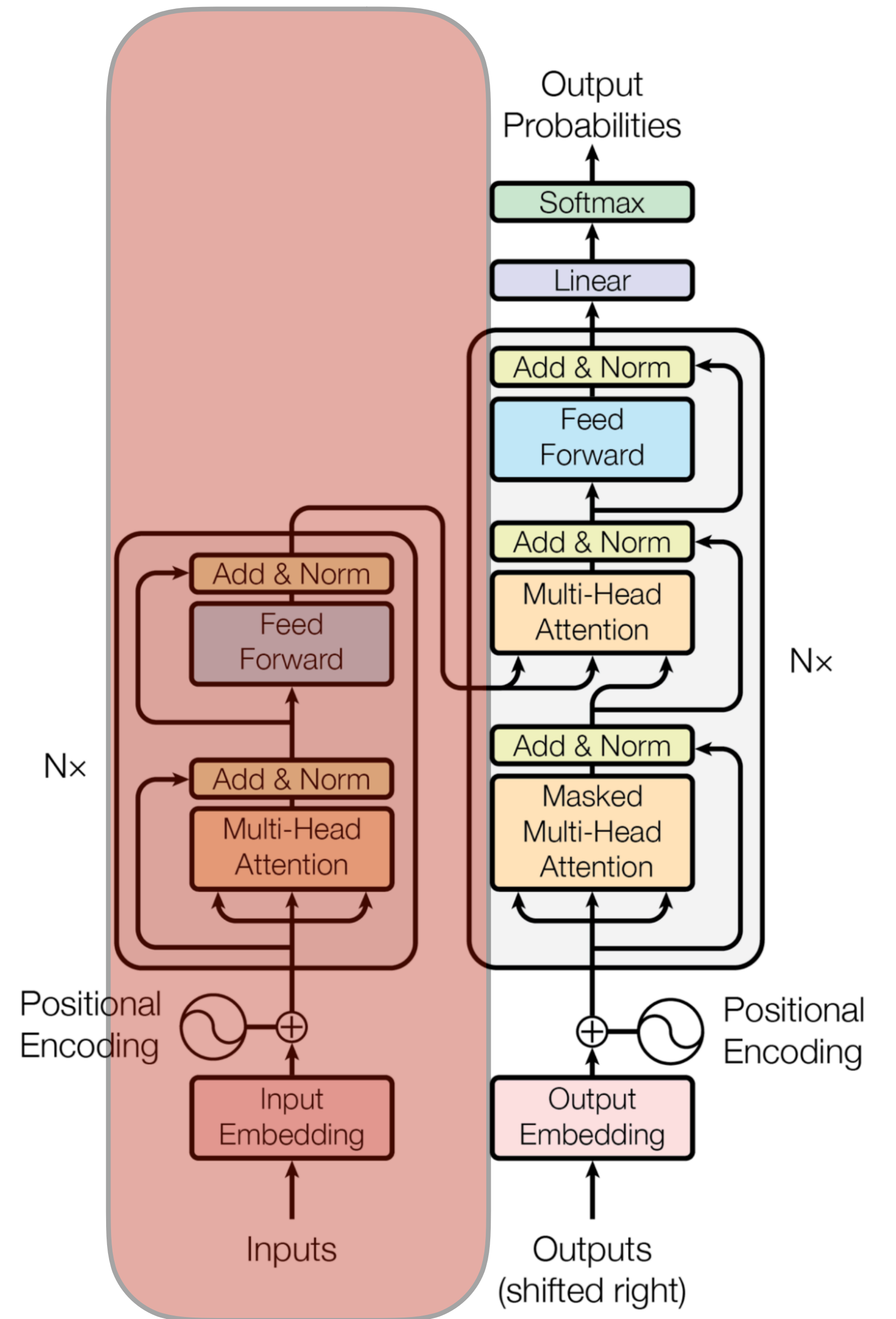


论文模型

Overview

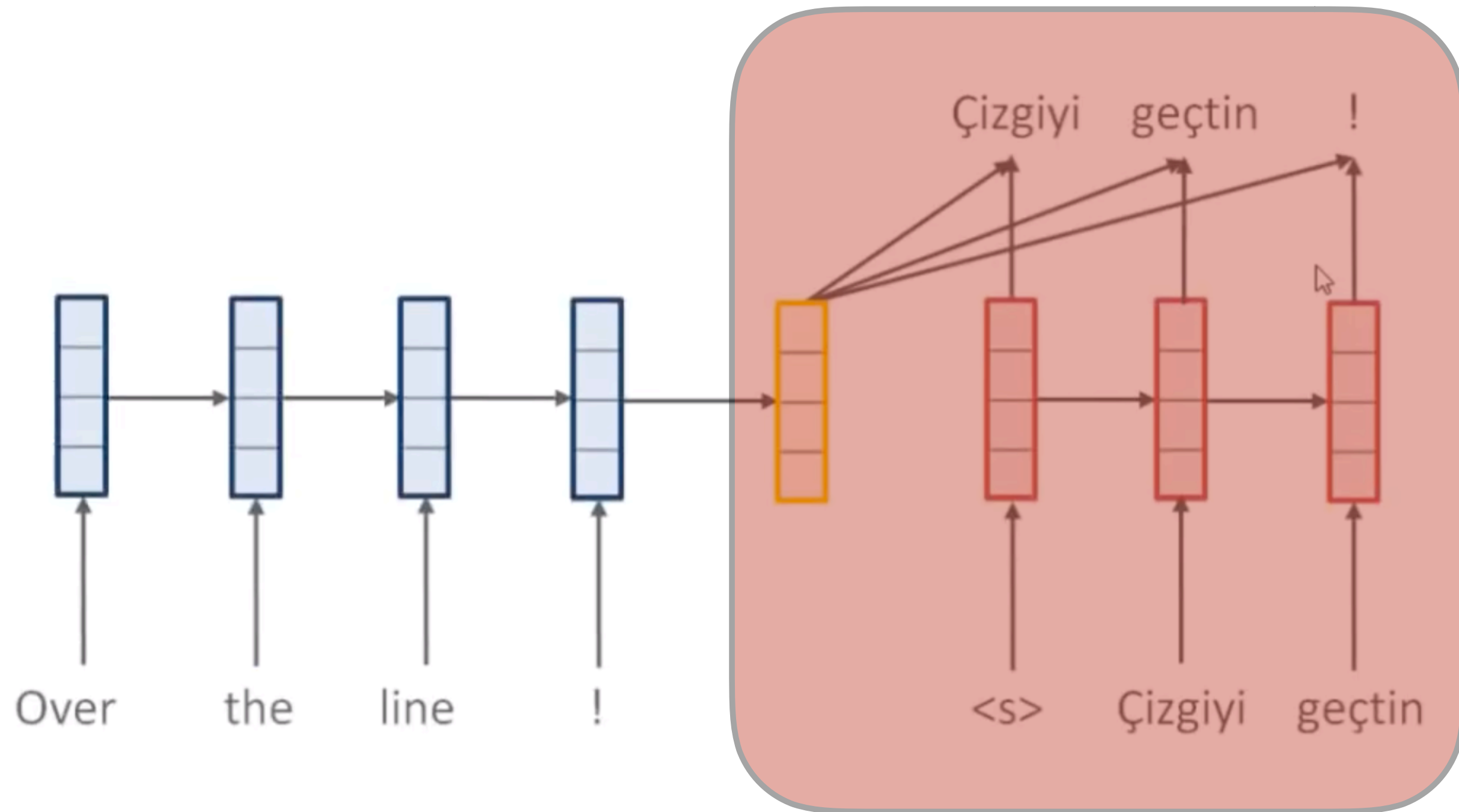


RNN encoder-decoder

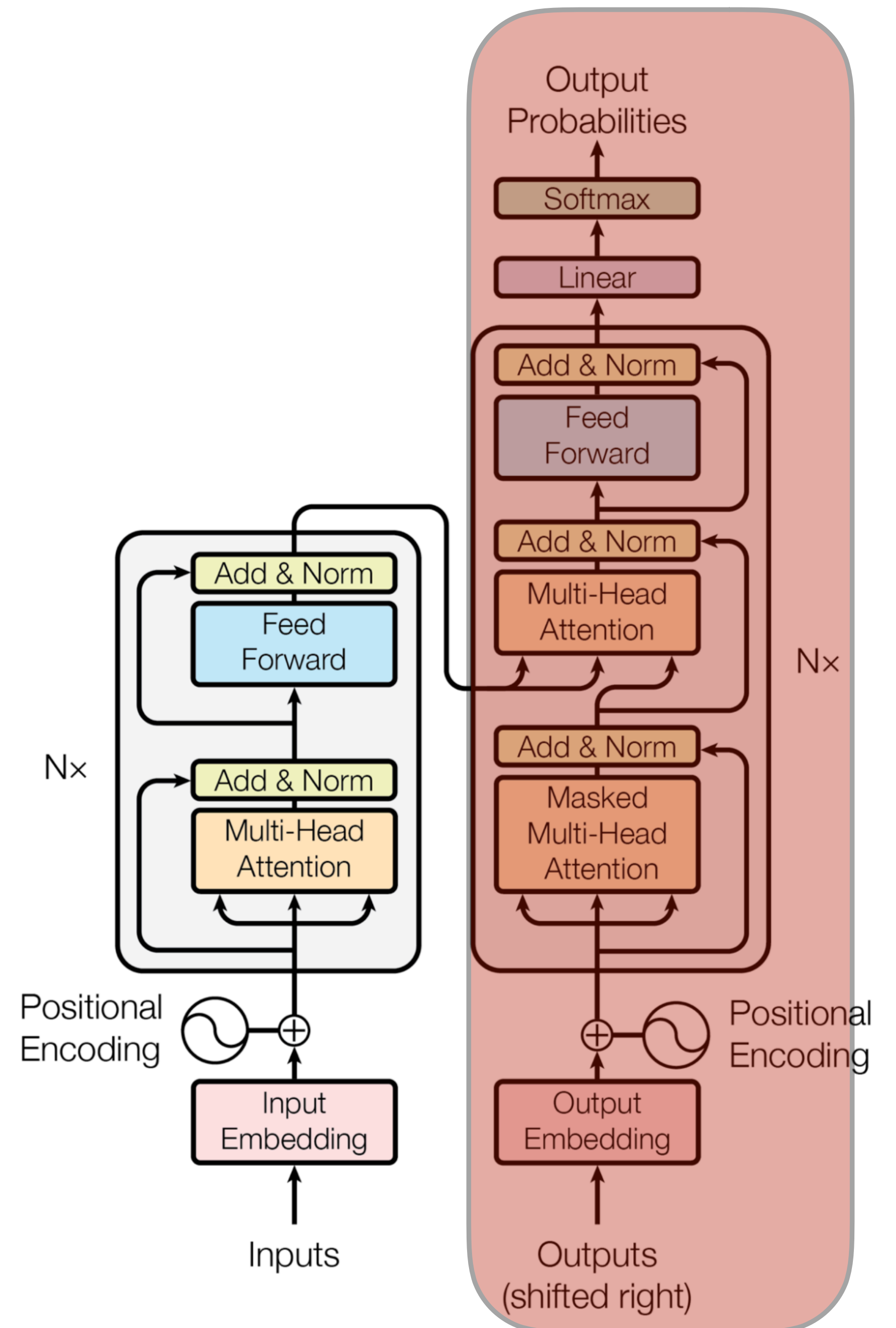


论文模型

Overview

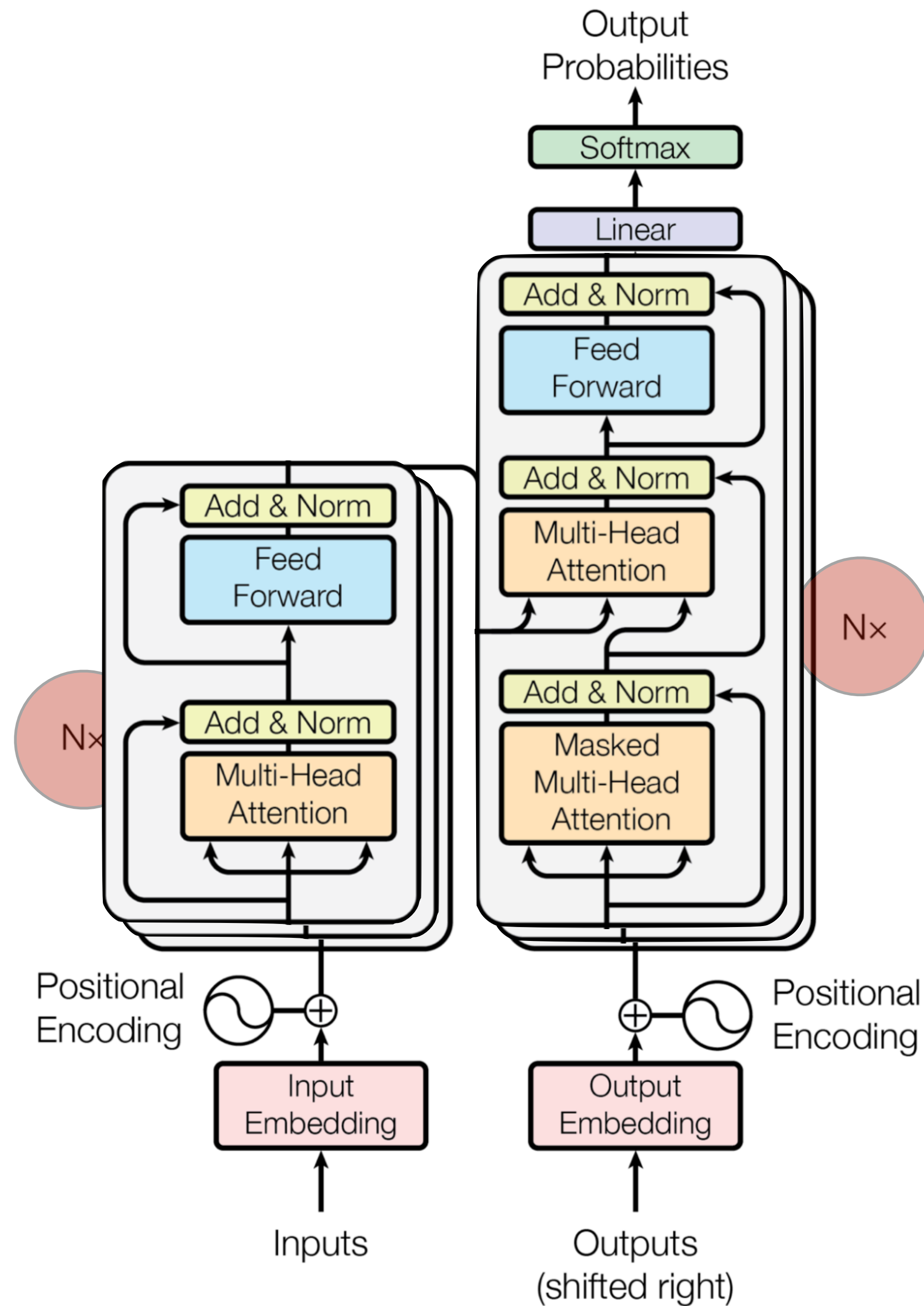


RNN encoder-decoder



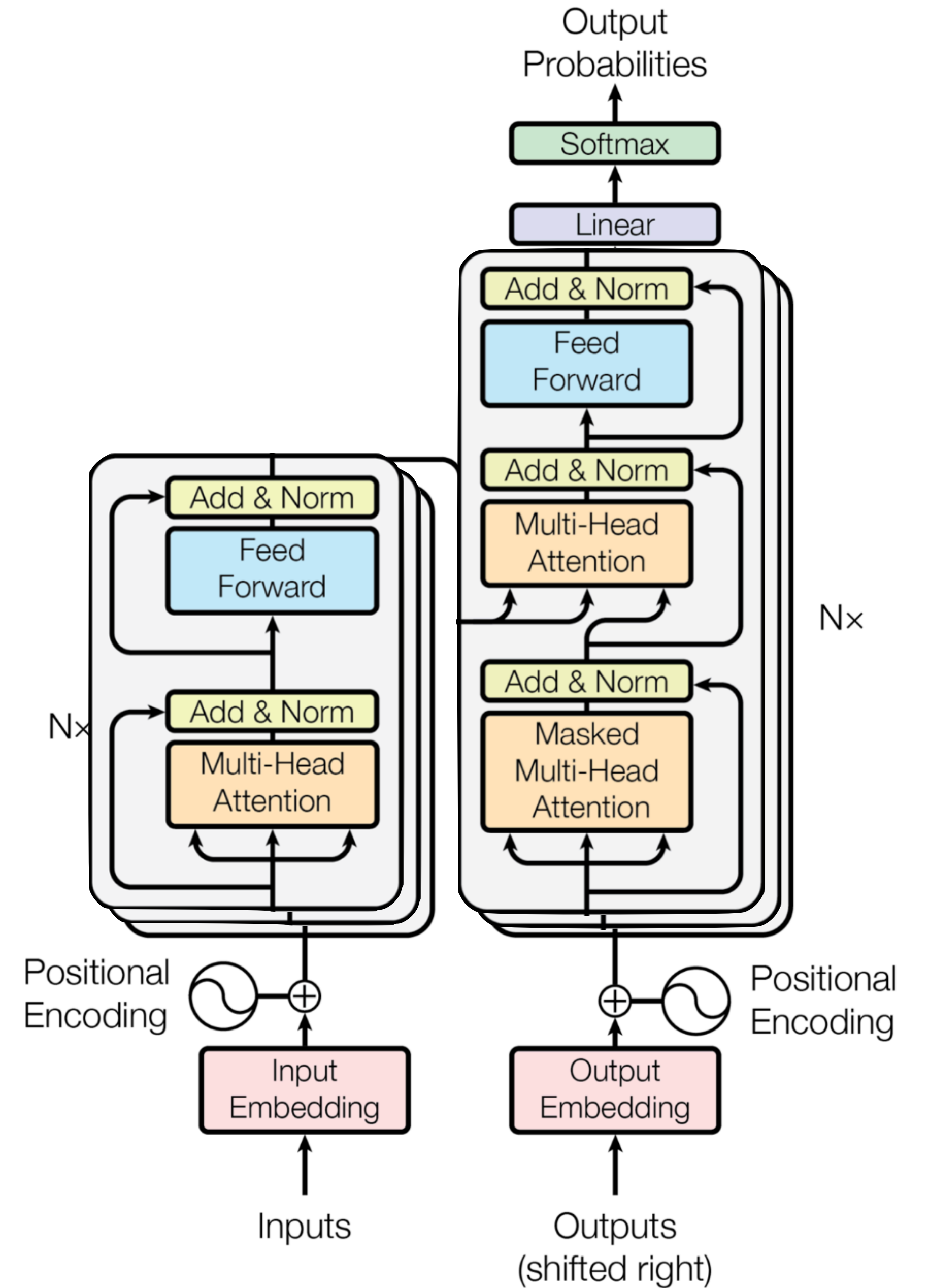
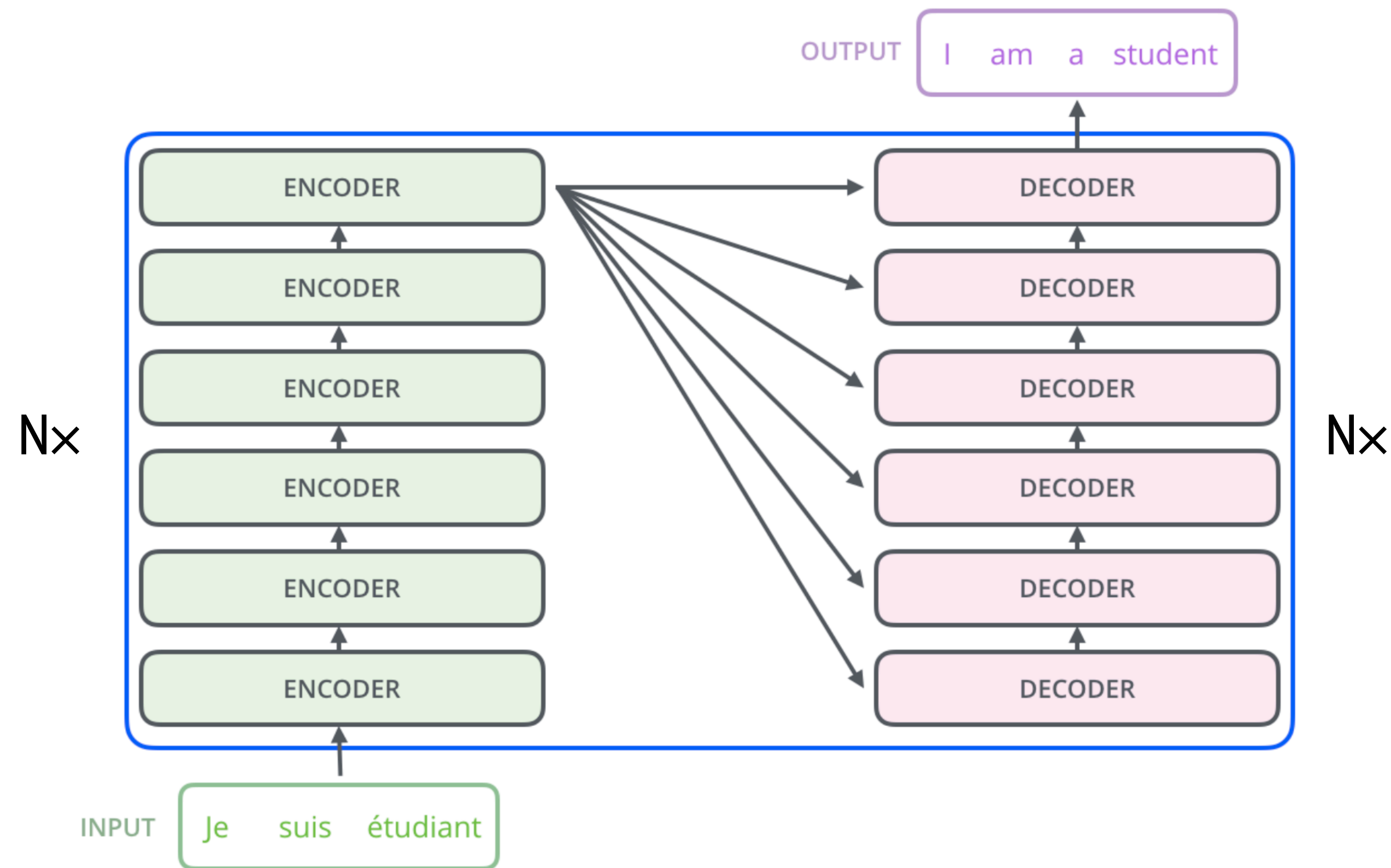
论文模型

Overview



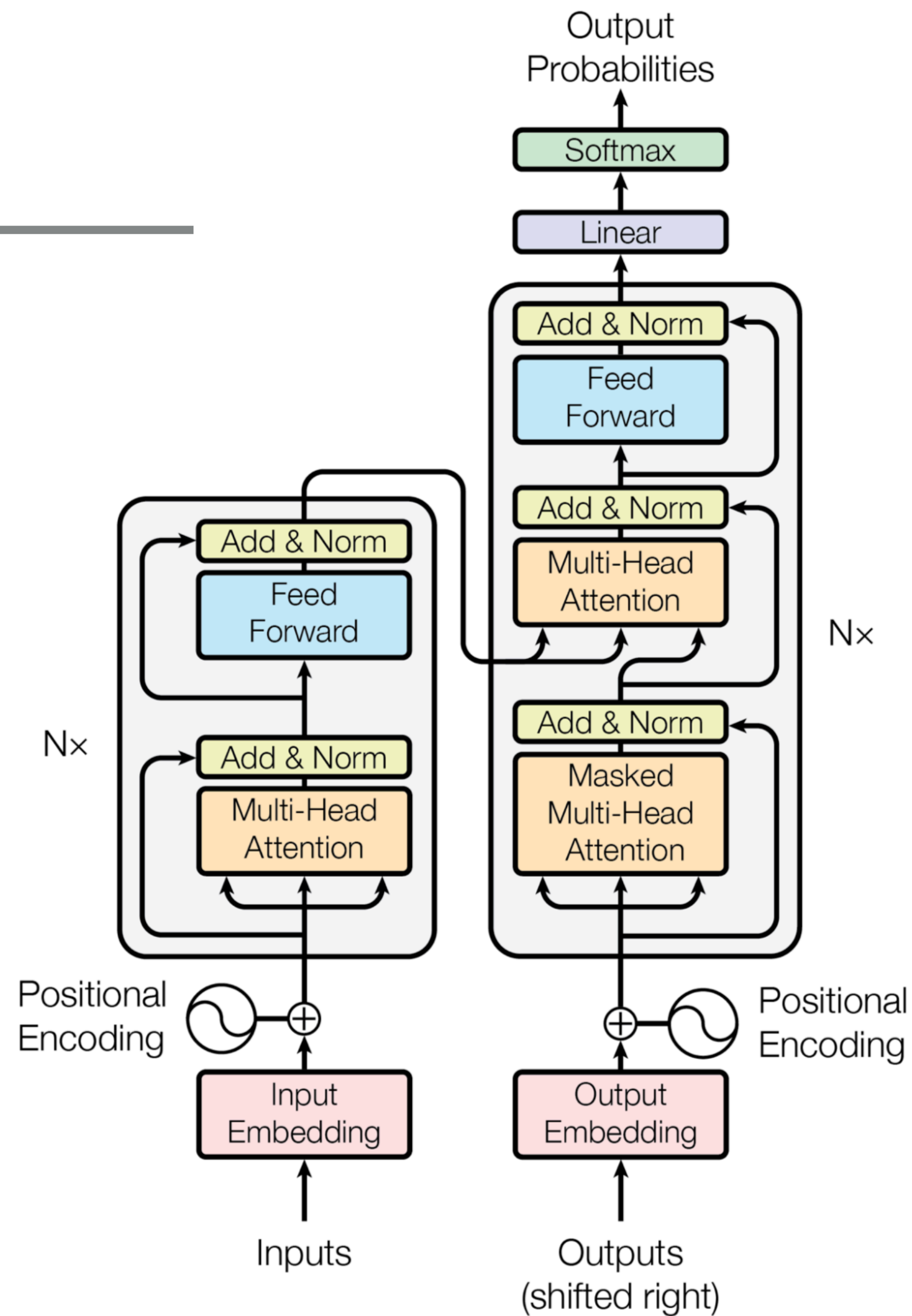
论文模型

Overview



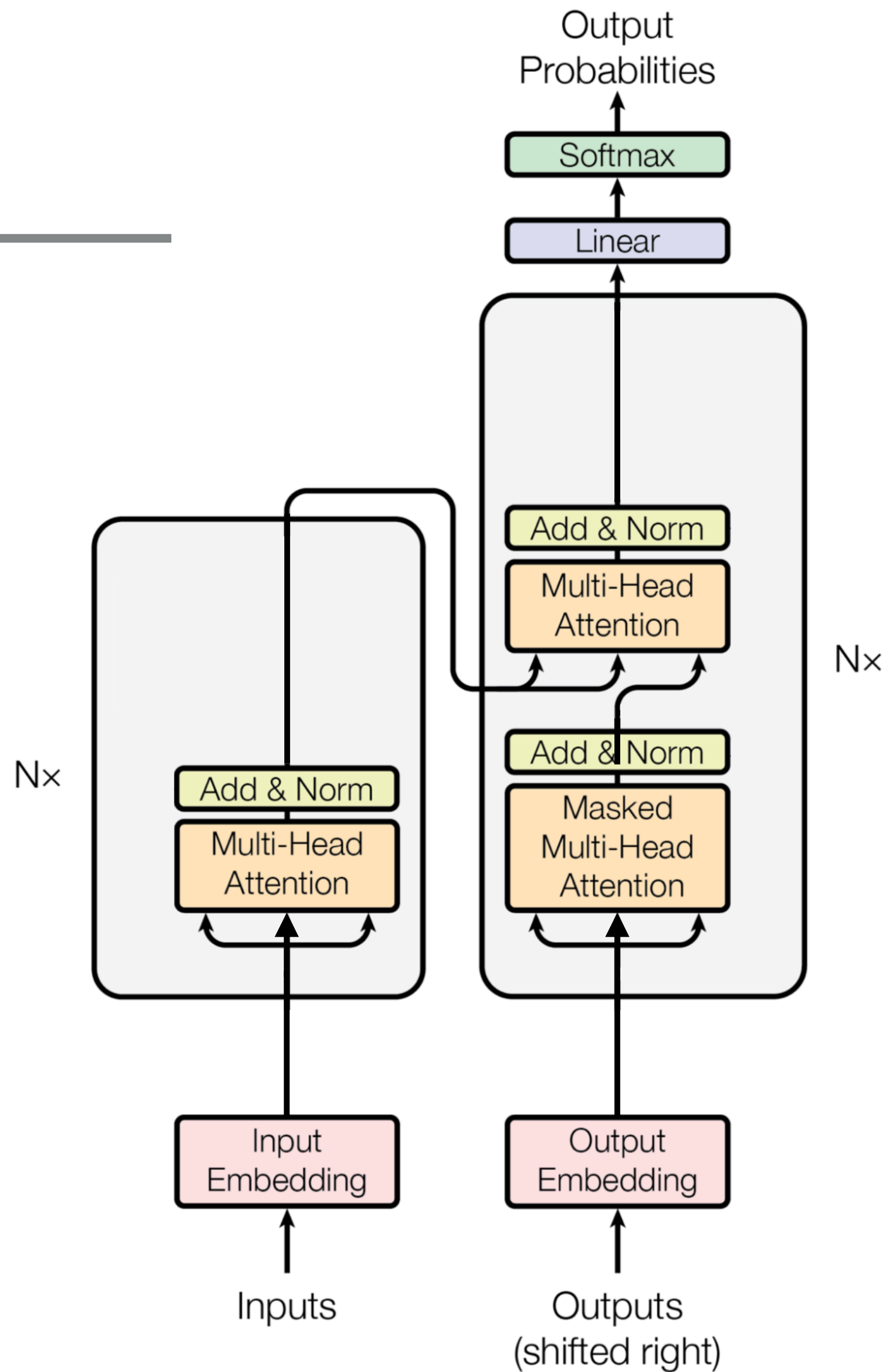
论文模型

Overview



论文模型

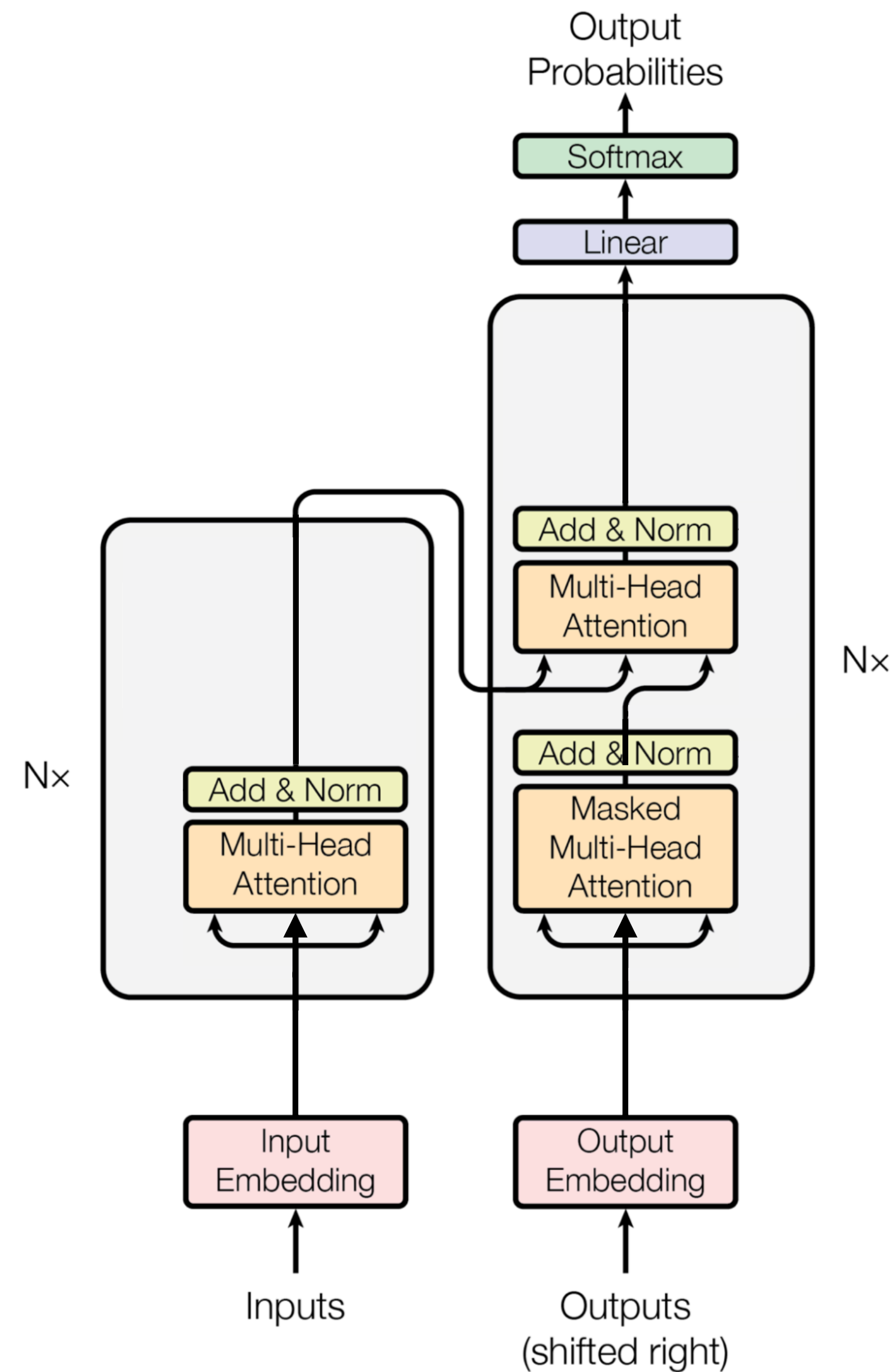
Overview



论文模型

Overview

Attention



注意力

Attention

注意力

Attention

多年以后，奥雷连诺上校站在行刑队面前，准会想起父亲带他去参观冰块的那个遥远的下午。

注意力

Attention

多年以后，奥雷连诺上校站在行刑队面前，准会想起父亲带他去参观冰块的那个遥远的下午。

注意力

Attention

多年以后，奥雷连诺上校站在行刑队面前，准会想起父亲带他去参观冰块的那个遥远的下午。

0.6

0.1 0.1 0.2

注意力

Attention

多年 以后 ， 奥雷连诺上校 站在 行刑队 面前， 准会 想起 父亲 带 他 去 参观 冰块 的 那个 遥远的下午。

0 0 0 0.6 0 0 0 0 0 0 0.1 0.1 0.2 0 0 0 0

注意力

Attention

多年以后，奥雷连诺上校站在行刑队面前，准会想起父亲带他去参观冰块的那个遥远的下午。

Many years later as he faced the firing squad, Colonel Aureliano Buendía was to remember that distant afternoon when his father took ___ to discover ice.

注意力

Attention

多年以后，奥雷连诺上校站在行刑队面前，准会想起父亲带他去参观冰块的那个遥远的下午。

0.9

0.2

0.3

Many years later as he faced the firing squad, Colonel Aureliano Buendía was to remember that distant afternoon when his father took ___ to discover ice.

0.5

注意力

Attention

Attention

多年以后，奥雷连诺上校站在行刑队面前，准会想起父亲带他去参观冰块的那个遥远的下午。

Many years later as **he** ~~faced~~ the firing squad, Colonel Aureliano Buendía was to remember that distant afternoon when **his** father took **him** to discover ice.

0.2

0.9

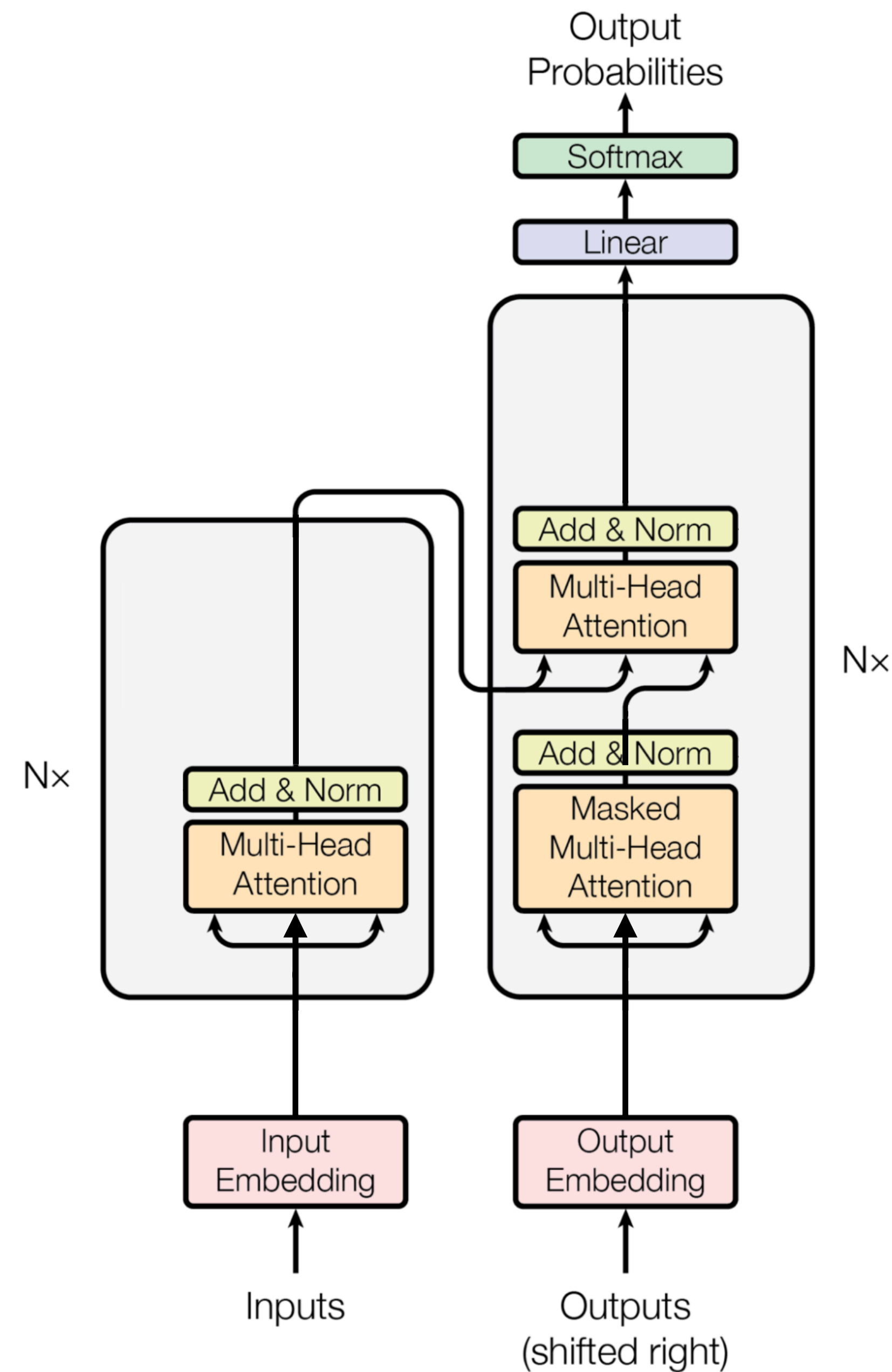
0.3

0.5

注意力

Attention

Attention



注意力

Attention

Attention

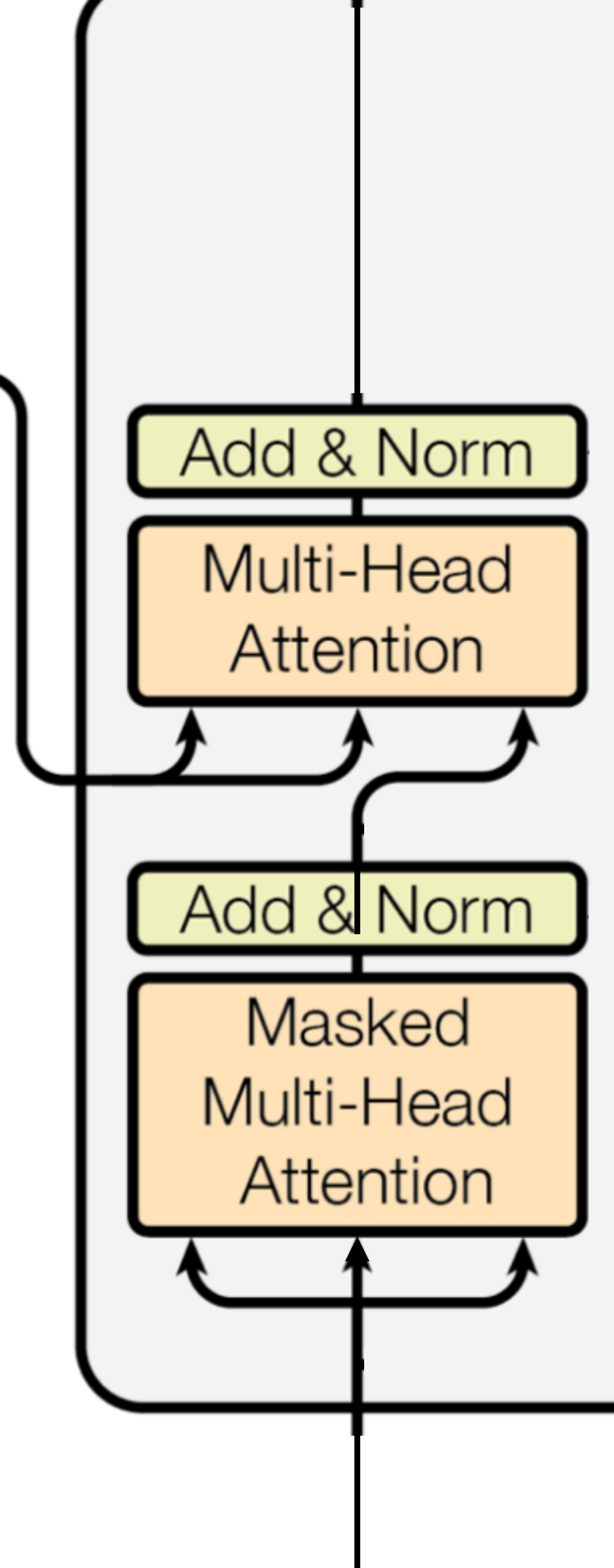
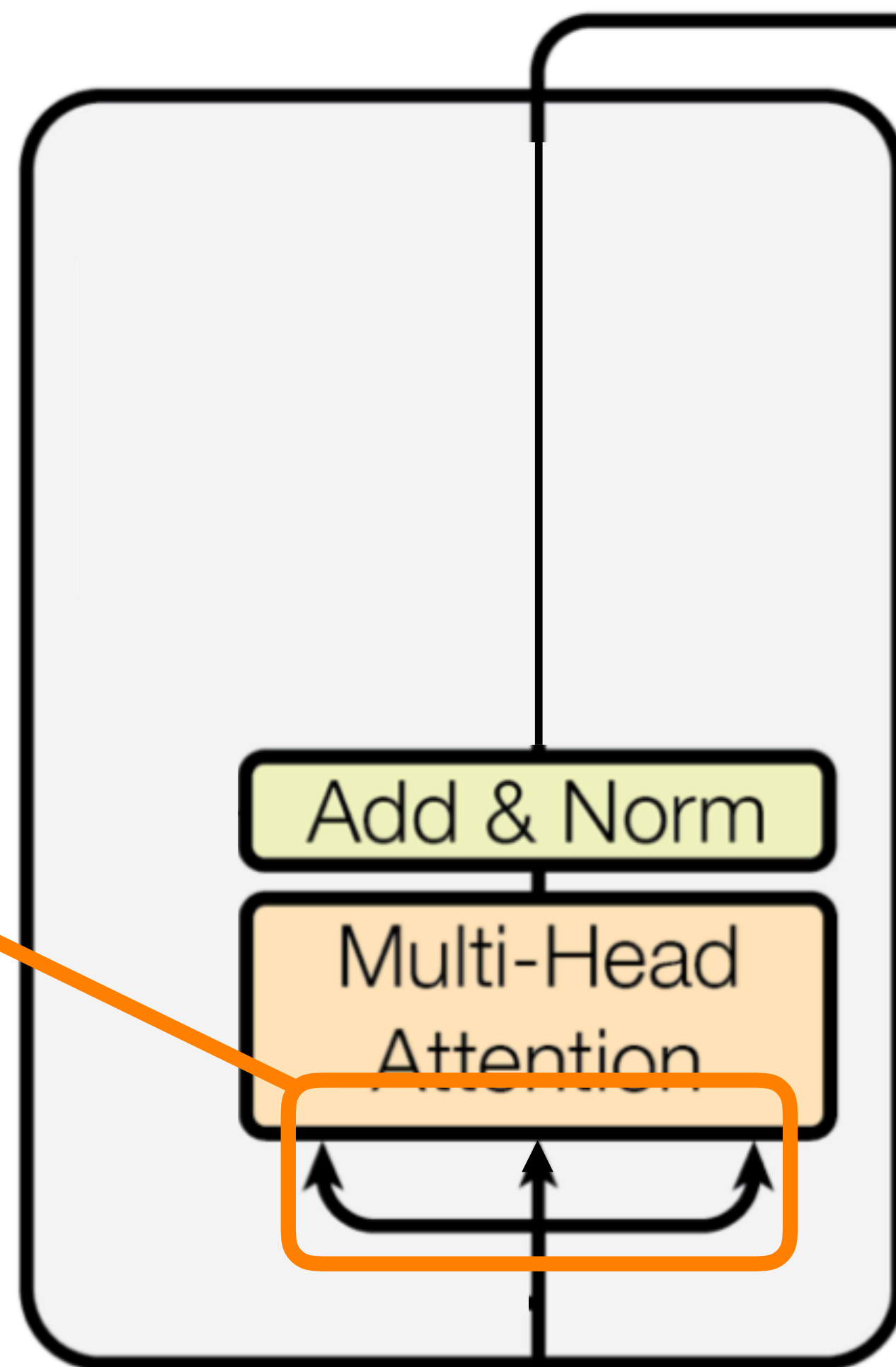
Query

Key

Value



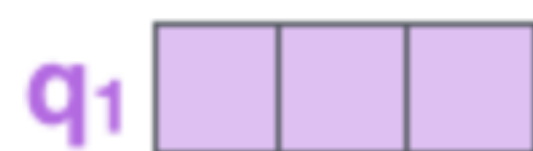
$N \times$



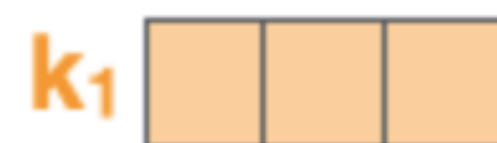
注意力

Attention

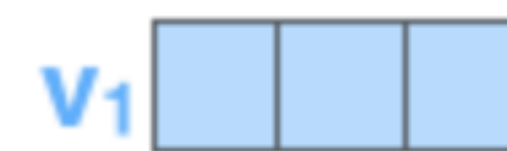
Attention



Query



Key



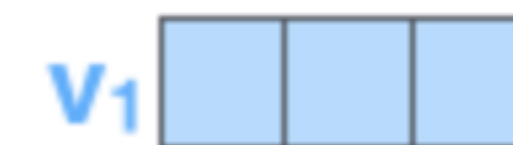
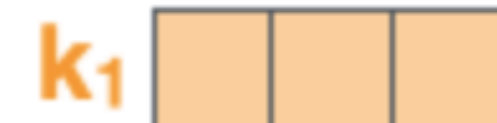
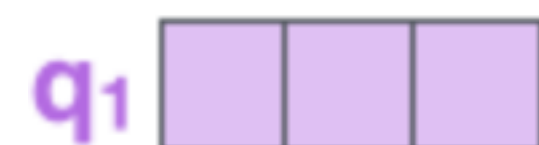
Value

注意力

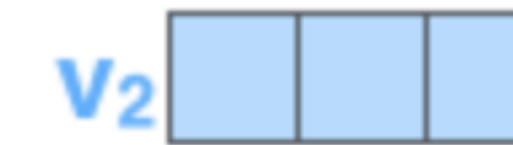
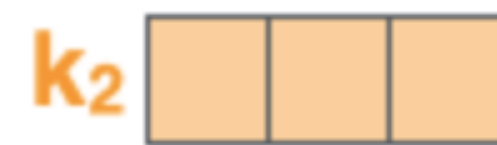
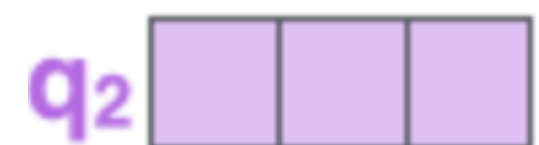
Attention

Attention

word1



word2



Query

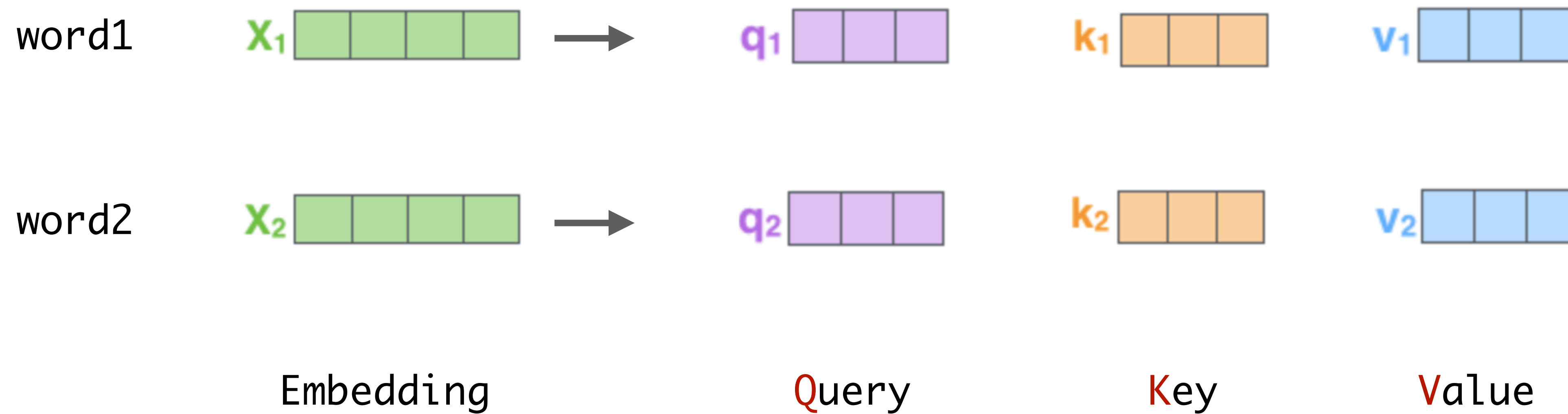
Key

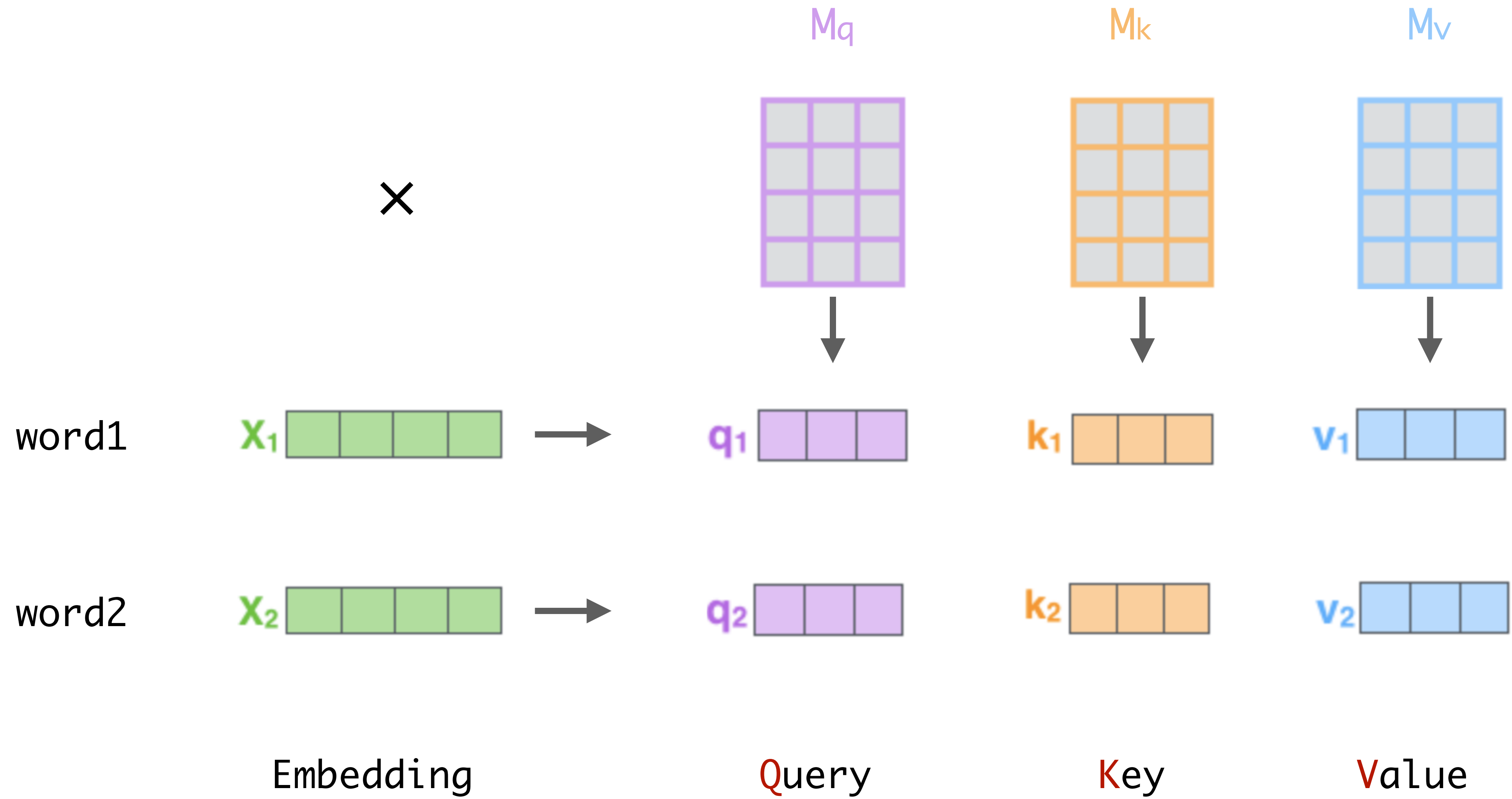
Value

注意力

Attention

Attention





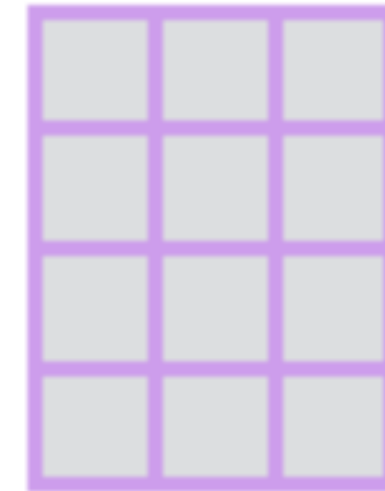
Attention

需要训练的参数:

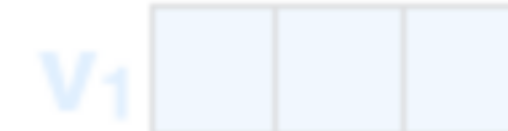
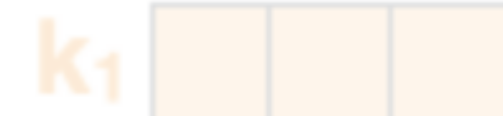
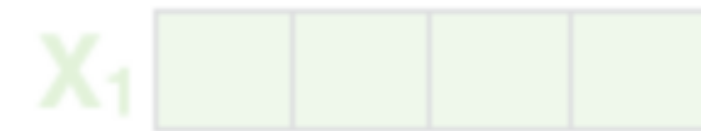
M_q

M_k

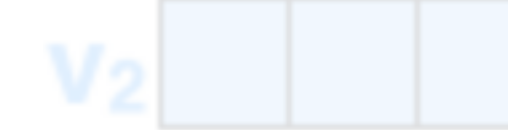
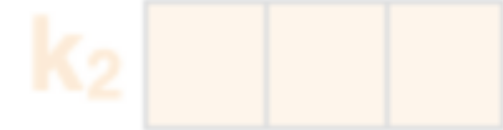
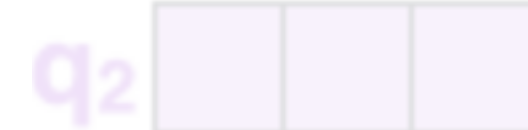
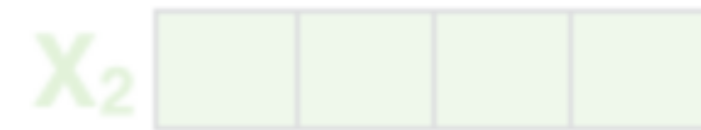
M_v



word1



word2

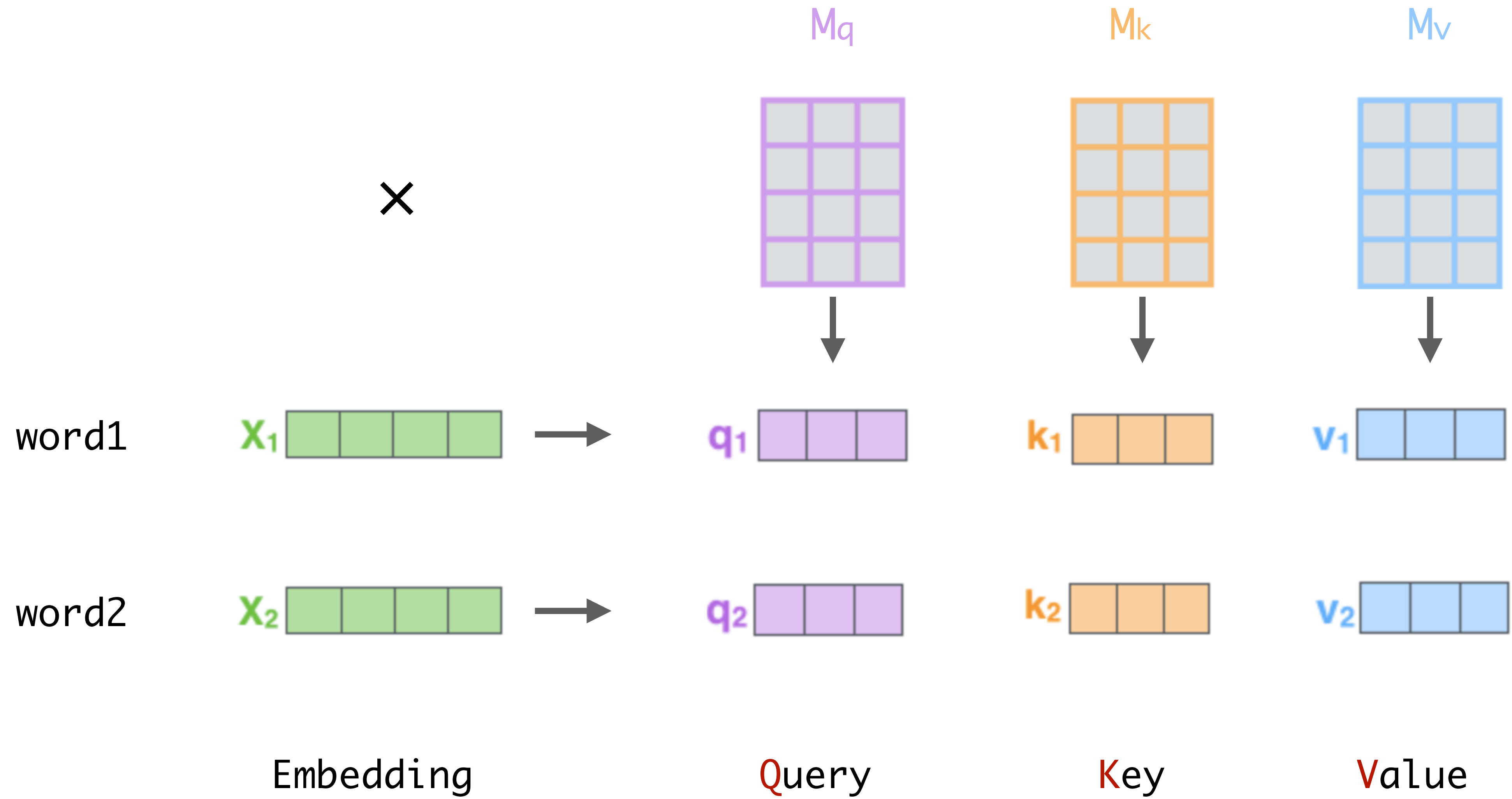


Embedding

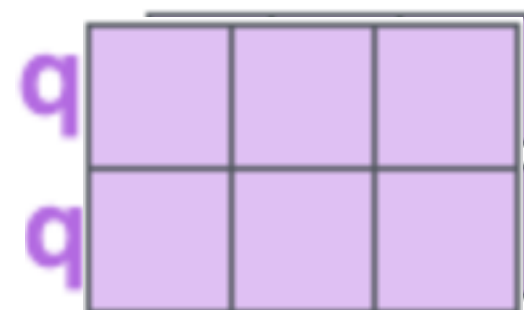
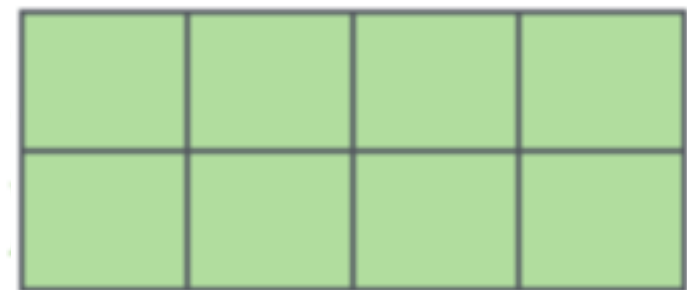
Query

Key

Value



word1
Sentence
word2



Embedding

Query

Key

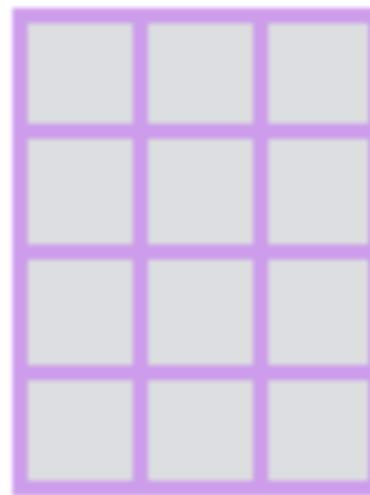
Value

×

M_q

M_k

M_v



Attention =

$$\text{Softmax} \left(\frac{\begin{array}{c} \text{Query} \\ \begin{array}{|c|c|c|} \hline \square & \square & \square \\ \hline \square & \square & \square \\ \hline \end{array} \times \begin{array}{c} \text{Key}' \\ \begin{array}{|c|c|} \hline \square & \square \\ \hline \square & \square \\ \hline \square & \square \\ \hline \end{array} \end{array} \right) \begin{array}{c} \text{Value} \\ \begin{array}{|c|c|c|} \hline \square & \square & \square \\ \hline \square & \square & \square \\ \hline \end{array} \end{array}$$

The diagram illustrates the attention mechanism. It shows a purple 2x3 grid labeled "Query" multiplied by an orange 3x2 grid labeled "Key'". The result of this multiplication is a blue 2x3 grid labeled "Value". The entire operation is enclosed in large parentheses, with a horizontal line underneath the multiplication and the square root of d_k below the line. The word "Softmax" is written to the left of the parentheses.

Attention =

Query Key' Value

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

Attention =

Scaled Dot-Product Attention

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

Diagram illustrating the Scaled Dot-Product Attention mechanism:

- Q** (Query): A 2x3 grid of light purple squares.
- K** (Key): A 2x3 grid of light orange squares.
- V** (Value): A 2x3 grid of light blue squares.
- The operation QK^T is shown with a multiplication symbol (\times) between the Q and K grids.
- The result of QK^T is divided by $\sqrt{d_k}$ (indicated by a horizontal line and a square root symbol).
- The result is then passed through a **softmax** function (indicated by a large parenthesis).

Attention =

$$\text{Softmax} \left(\frac{\begin{array}{c} \text{Query} \\ \begin{array}{|c|c|c|} \hline \square & \square & \square \\ \hline \square & \square & \square \\ \hline \end{array} \times \begin{array}{c} \text{Key}' \\ \begin{array}{|c|c|} \hline \square & \square \\ \hline \square & \square \\ \hline \square & \square \\ \hline \end{array} \end{array} \right) \begin{array}{c} \text{Value} \\ \begin{array}{|c|c|c|} \hline \square & \square & \square \\ \hline \square & \square & \square \\ \hline \end{array} \end{array}$$

The diagram illustrates the attention mechanism. It shows a purple 2x3 grid labeled "Query" multiplied by an orange 3x2 grid labeled "Key'". The result is a blue 2x3 grid labeled "Value". The entire operation is enclosed in large parentheses, with a horizontal line under the multiplication and the square root of d_k below it. The word "Softmax" is written to the left of the parentheses.

Attention =

$$\text{Softmax} \left(\frac{\begin{matrix} Q_1 & \begin{matrix} \square & \square & \square \end{matrix} \\ Q_2 & \begin{matrix} \square & \square & \square \end{matrix} \end{matrix} \times \begin{matrix} K_1 & K_2 \\ \begin{matrix} \square & \square \end{matrix} \\ \begin{matrix} \square & \square \end{matrix} \\ \begin{matrix} \square & \square \end{matrix} \end{matrix}}{\sqrt{d_k}} \right) \begin{matrix} \begin{matrix} \square & \square & \square \\ \square & \square & \square \end{matrix} \\ V_1 \\ V_2 \end{matrix}$$

The diagram illustrates the attention mechanism. It shows a 2x3 matrix of purple squares (Q) with labels Q1 and Q2 on the left, multiplied by a 3x2 matrix of orange squares (K) with labels K1 and K2 on the left. The result is a 2x2 matrix of blue squares (V) with labels V1 and V2 on the right. A horizontal line is drawn under the product of the Q and K matrices, and the expression is enclosed in large parentheses. The label 'Softmax' is positioned to the left of the opening parenthesis, and the label 'Attention =' is at the top left of the image.

Attention =

$$\text{Softmax} \left(\frac{\begin{array}{c} \text{Q1} \\ \text{Q2} \end{array} \begin{array}{|c|c|} \hline \text{K1} & \text{K2} \\ \hline 112 & 96 \\ \hline 88 & 120 \\ \hline \end{array}}{\sqrt{d_k}} \right) \begin{array}{|c|c|c|} \hline & & \text{V1} \\ \hline & & \text{V2} \\ \hline \end{array}$$

The diagram illustrates the attention mechanism. It shows a matrix of query vectors (Q1, Q2) and key vectors (K1, K2) being multiplied together. The result is then passed through a softmax function, and the output is multiplied by the value vectors (V1, V2). The matrix of query and key vectors is a 2x2 grid with values 112, 96, 88, and 120. The matrix of value vectors is a 2x3 grid.

Attention =

$$\text{Softmax} \left(\frac{\begin{array}{cc} & \begin{array}{cc} W_1 & W_2 \end{array} \\ \begin{array}{c} W_1 \\ W_2 \end{array} & \begin{array}{|c|c|} \hline 112 & 96 \\ \hline 88 & 120 \\ \hline \end{array} \end{array} \right) \begin{array}{|c|c|c|} \hline & & \\ \hline & & \\ \hline \end{array} \begin{array}{c} V_1 \\ V_2 \end{array}$$

$\sqrt{d_k}$

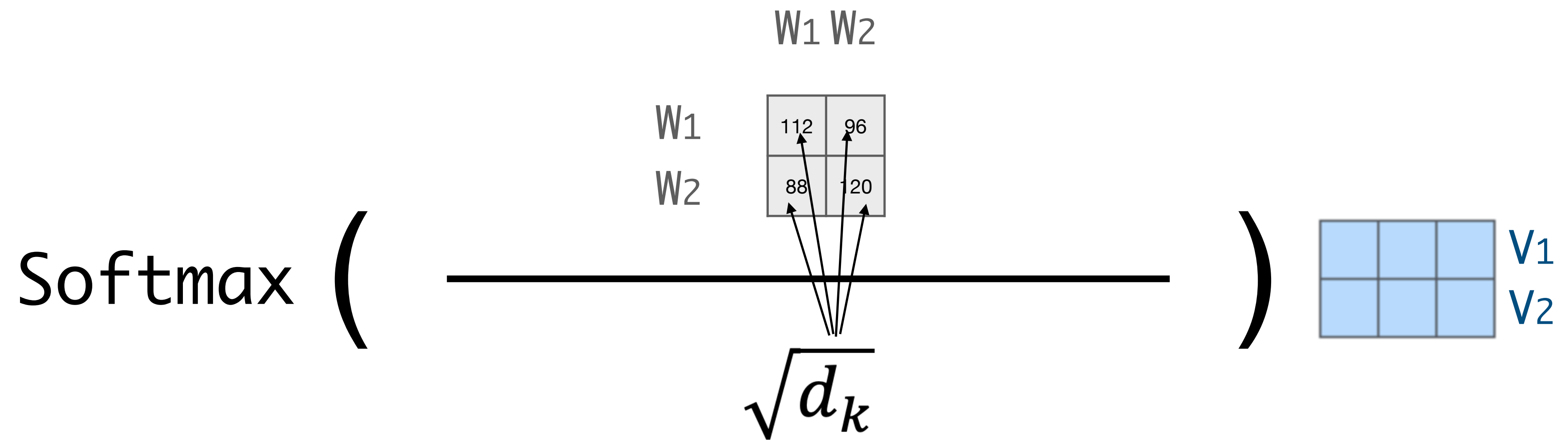
Attention =

$$\text{Softmax} \left(\frac{\begin{matrix} & W_1 & W_2 \\ W_1 & \begin{matrix} 112 & 96 \\ 88 & 120 \end{matrix} \\ W_2 & & \end{matrix}}{\sqrt{d_k}} \right) \begin{matrix} V_1 \\ V_2 \end{matrix}$$

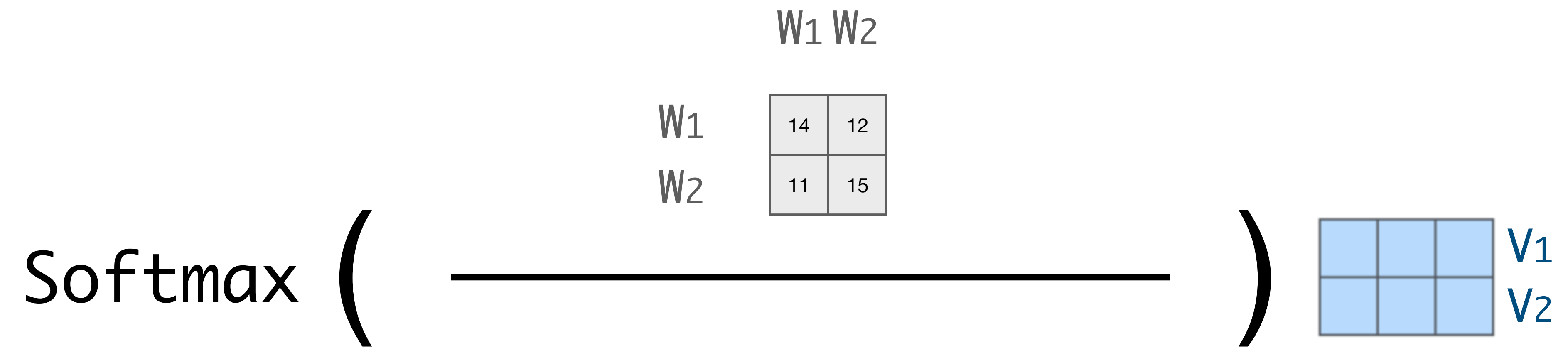
Scaling factor

The diagram illustrates the attention mechanism. It shows two weight matrices, W1 and W2, which are multiplied together to produce a 2x2 matrix of values. This matrix is then divided by the square root of the dimensionality of the key, $\sqrt{d_k}$, which is labeled as the scaling factor. The result is passed through a Softmax function, which produces a 2x3 matrix of values, labeled V1 and V2.

Attention =



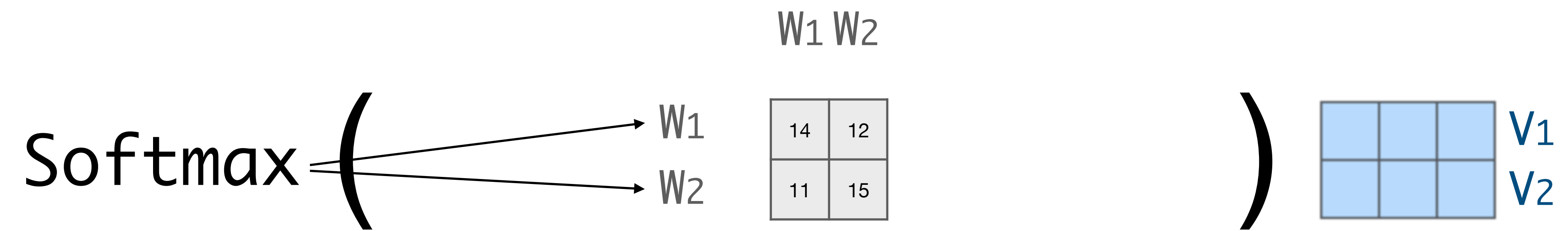
Attention =



Attention =

$$\text{Softmax} \left(\begin{array}{c} W_1 \\ W_2 \end{array} \begin{array}{|c|c|} \hline W_1 & W_2 \\ \hline 14 & 12 \\ \hline 11 & 15 \\ \hline \end{array} \right) \begin{array}{|c|c|c|} \hline & & V_1 \\ \hline & & V_2 \\ \hline \end{array}$$

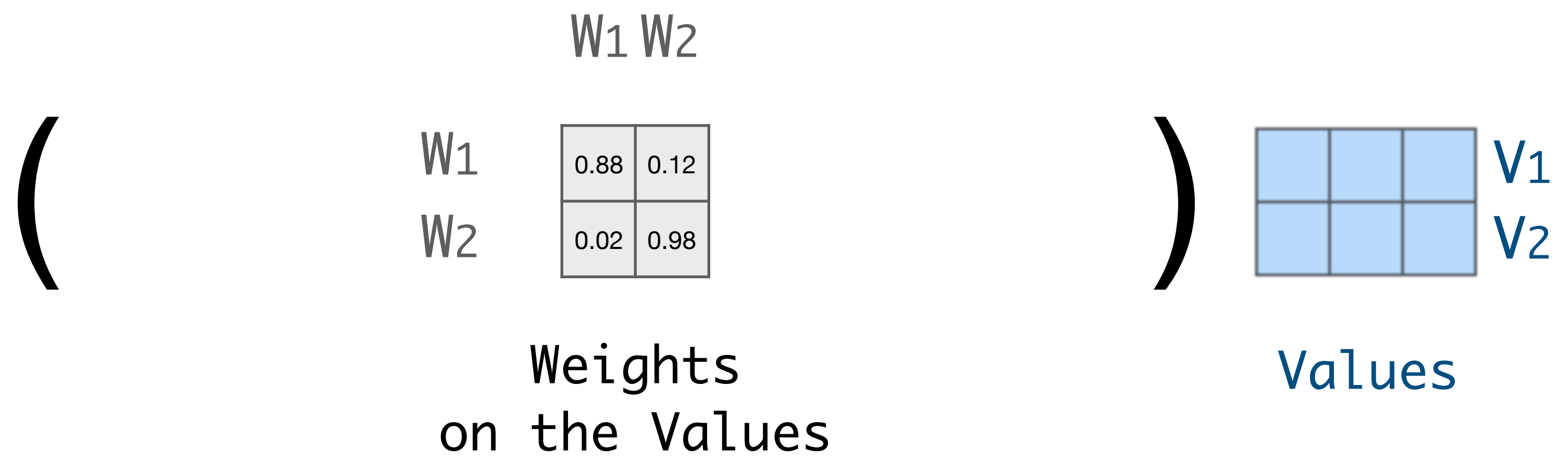
Attention =



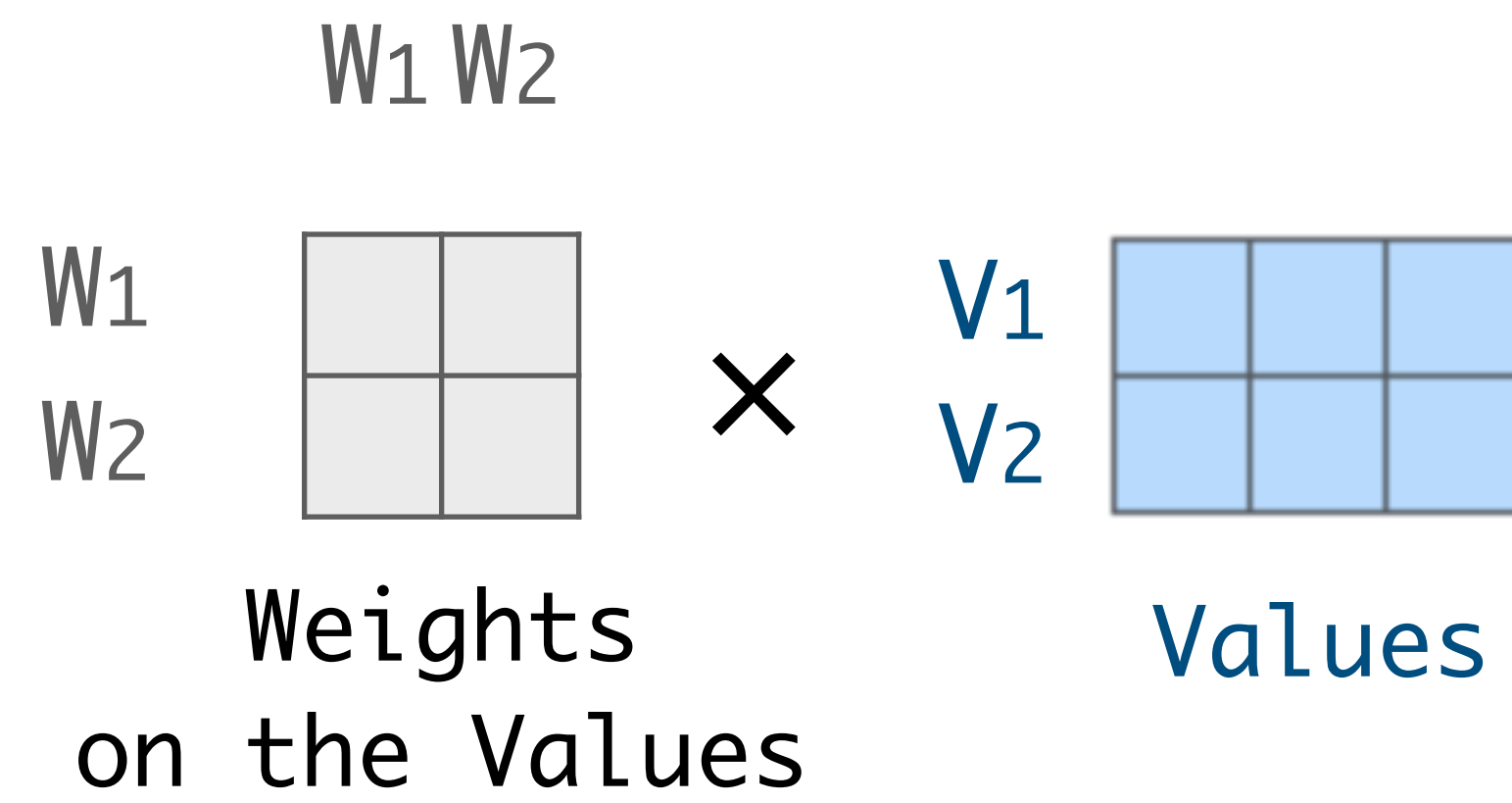
Attention =

$$\left(\begin{array}{c} W_1 \\ W_2 \end{array} \begin{array}{cc} W_1 & W_2 \\ \hline 0.88 & 0.12 \\ \hline 0.02 & 0.98 \end{array} \right) \begin{array}{ccc} \square & \square & \square \\ \hline \square & \square & \square \end{array} \begin{array}{c} V_1 \\ V_2 \end{array}$$

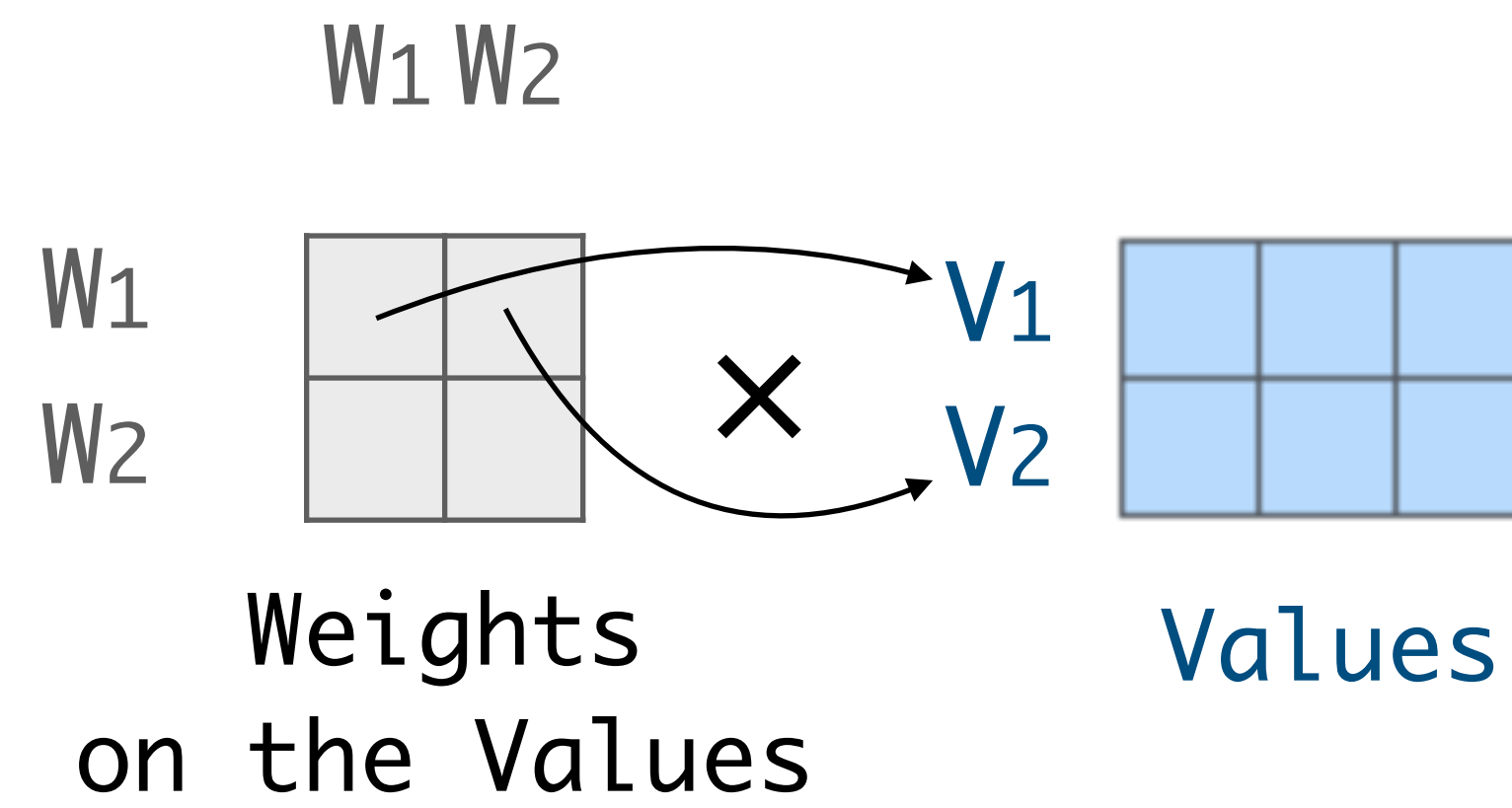
Attention =



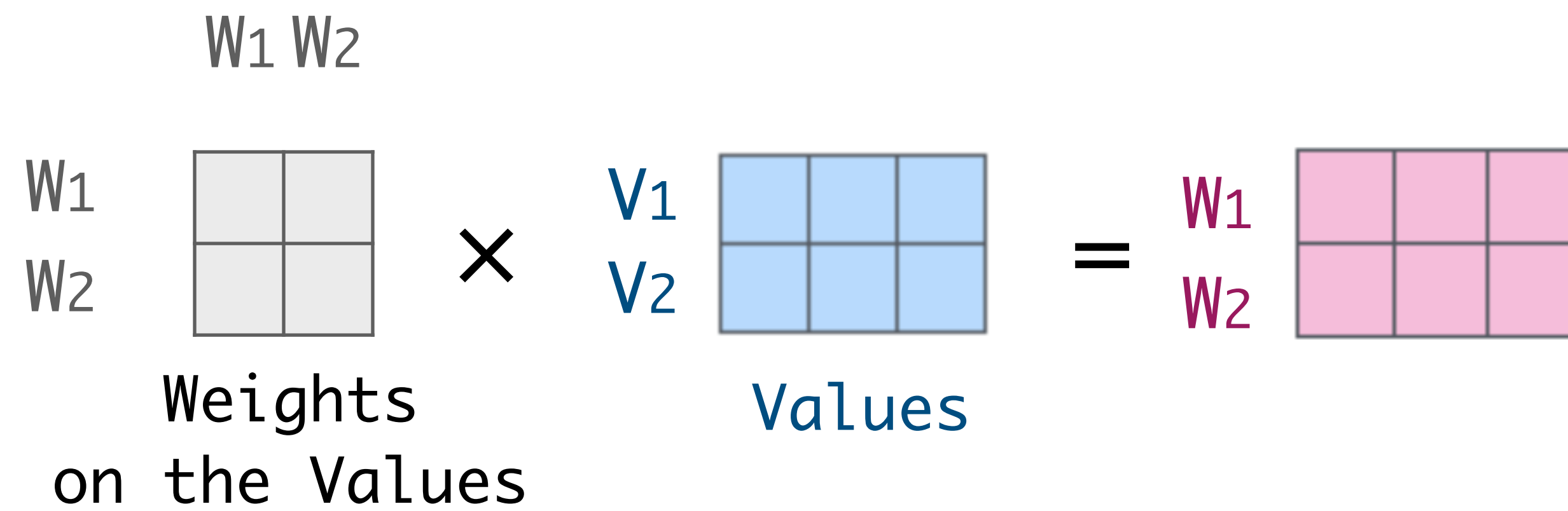
Attention =



Attention =



Attention =

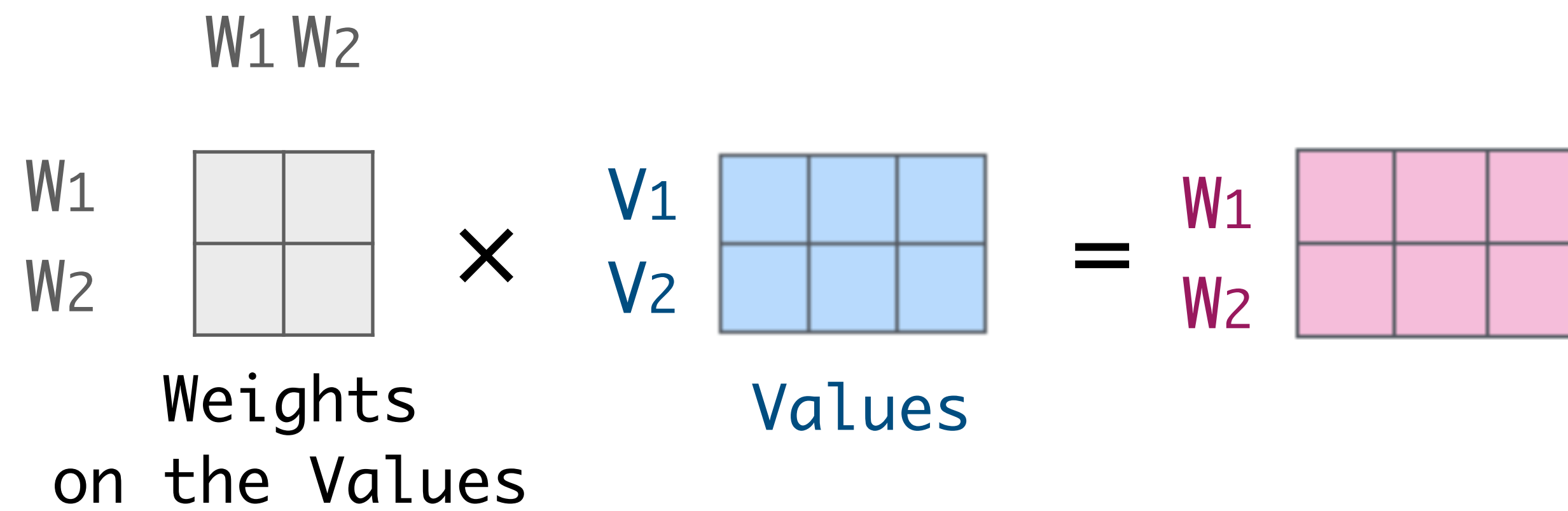


Attention =

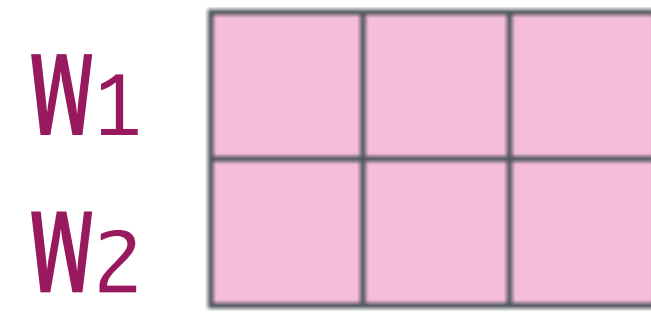
Attention做了什么？

1. 在每个token的embedding里，加入了所有token的信息
 2. 这些token的信息，是想以“注意力”为权重，加进每个token的
 3. 但具体这些权重是不是真的符合人的“注意力”，还要再往后看。
- 如果符合，那这个模型就是完全可解释的了。

Attention =

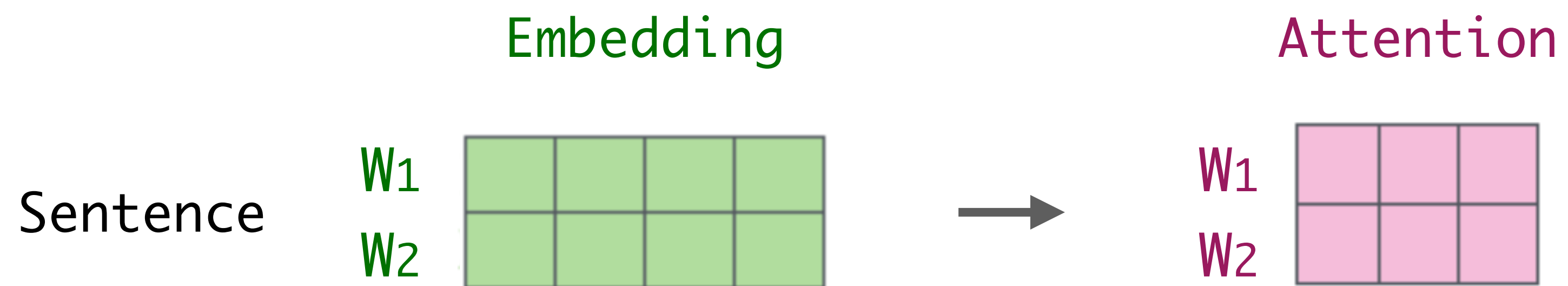


Attention



注意力

Attention

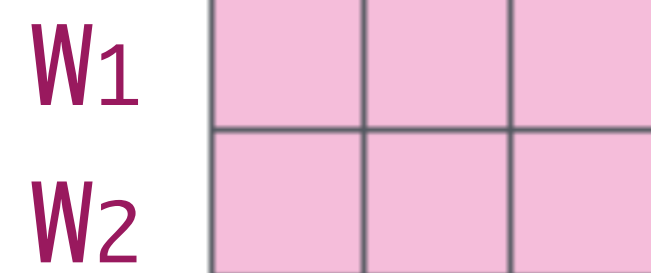


$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

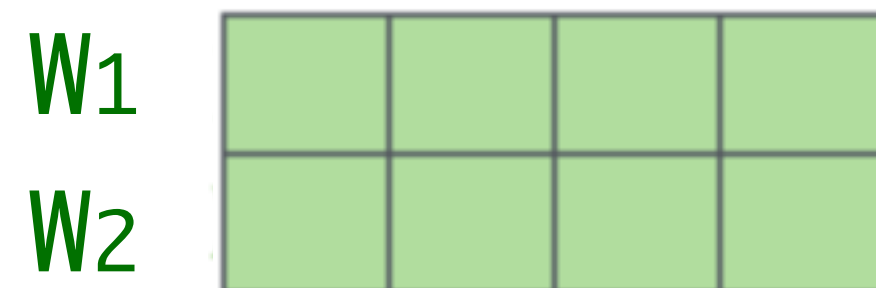
注意力

Attention

Attention

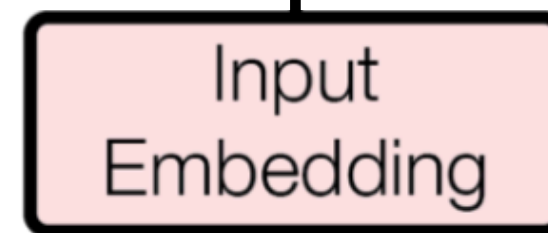
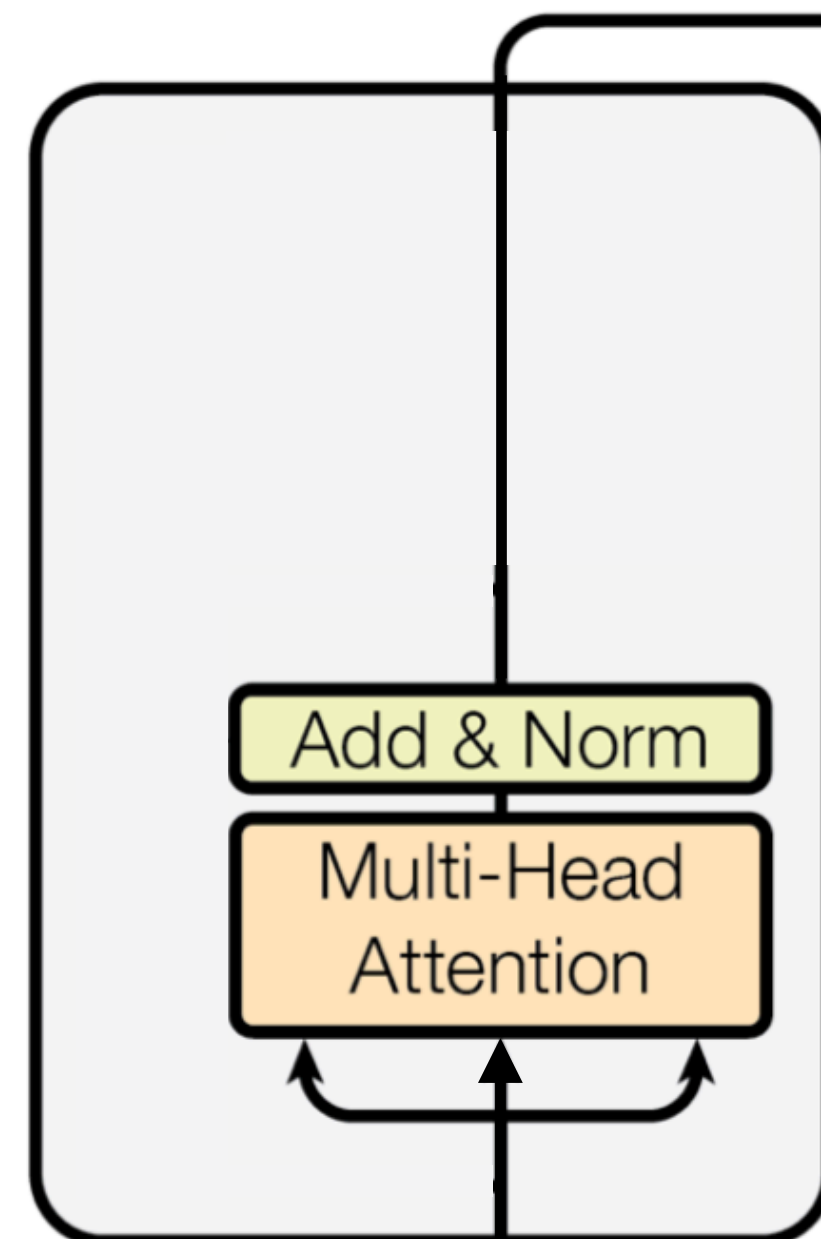


Embedding

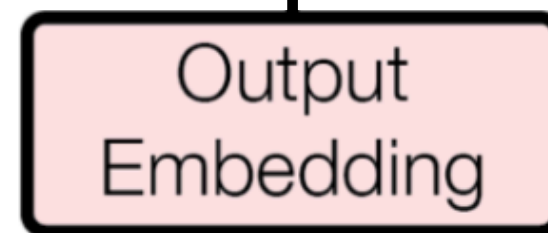
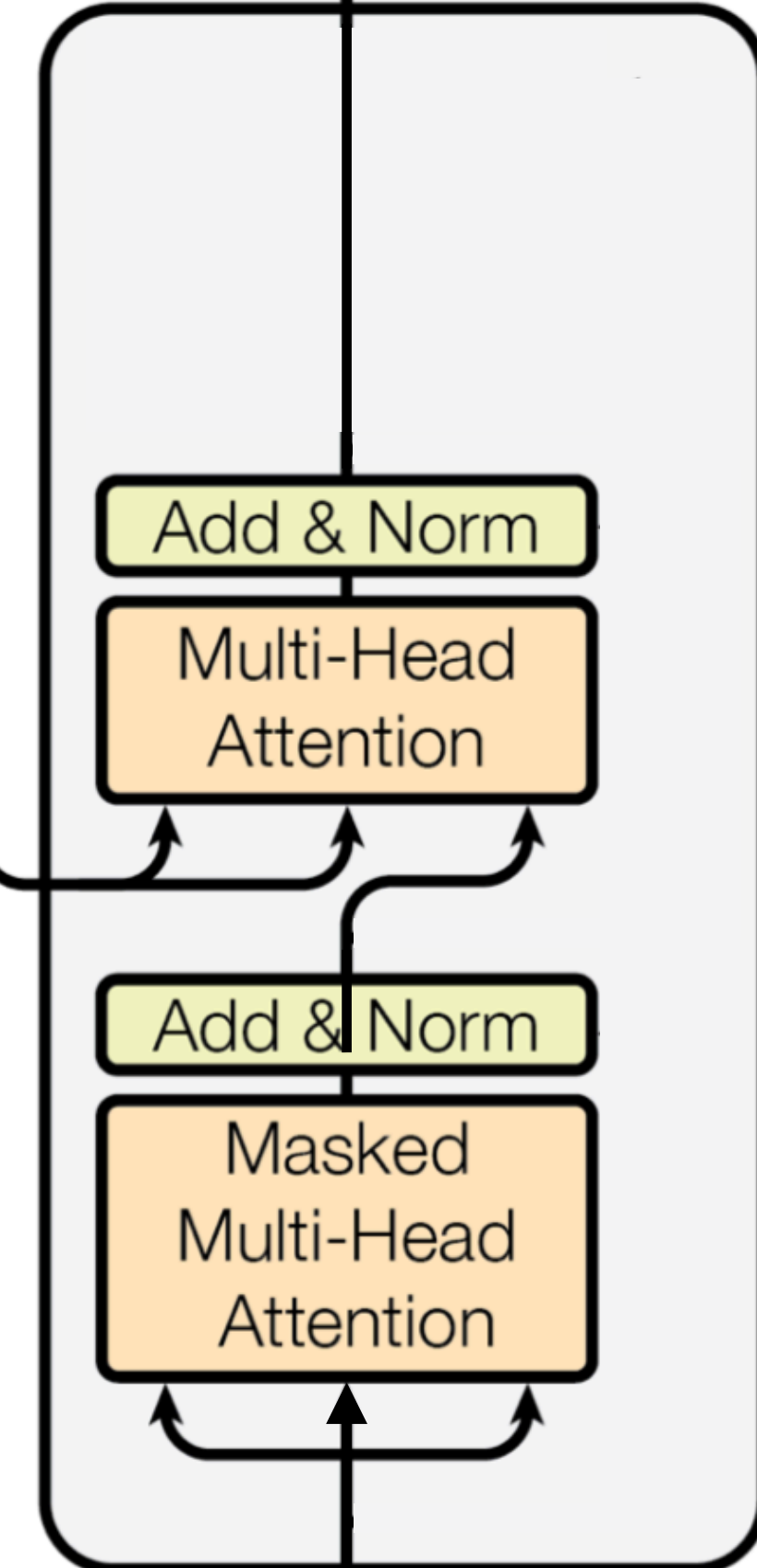


Sentence

$N \times$



Inputs

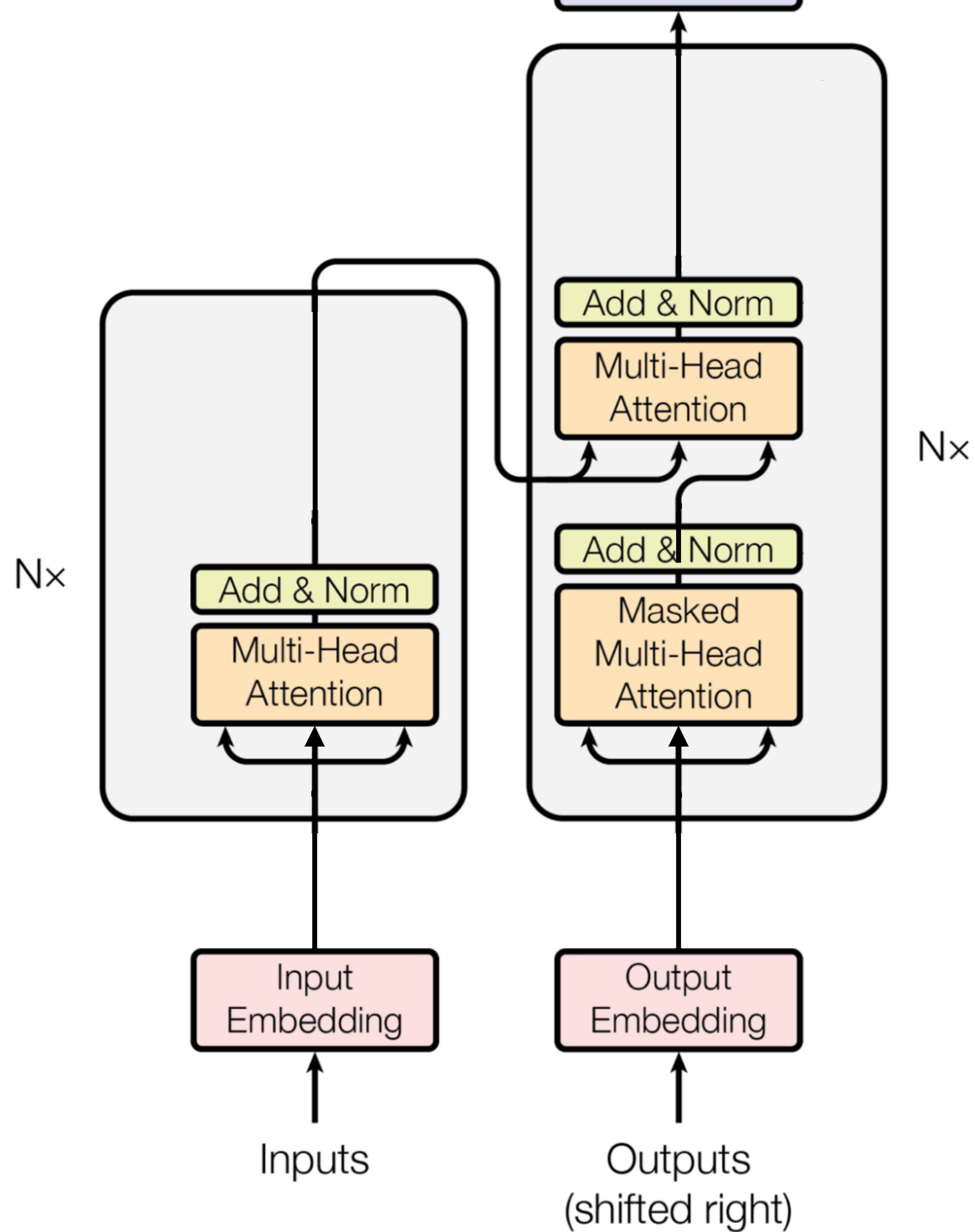
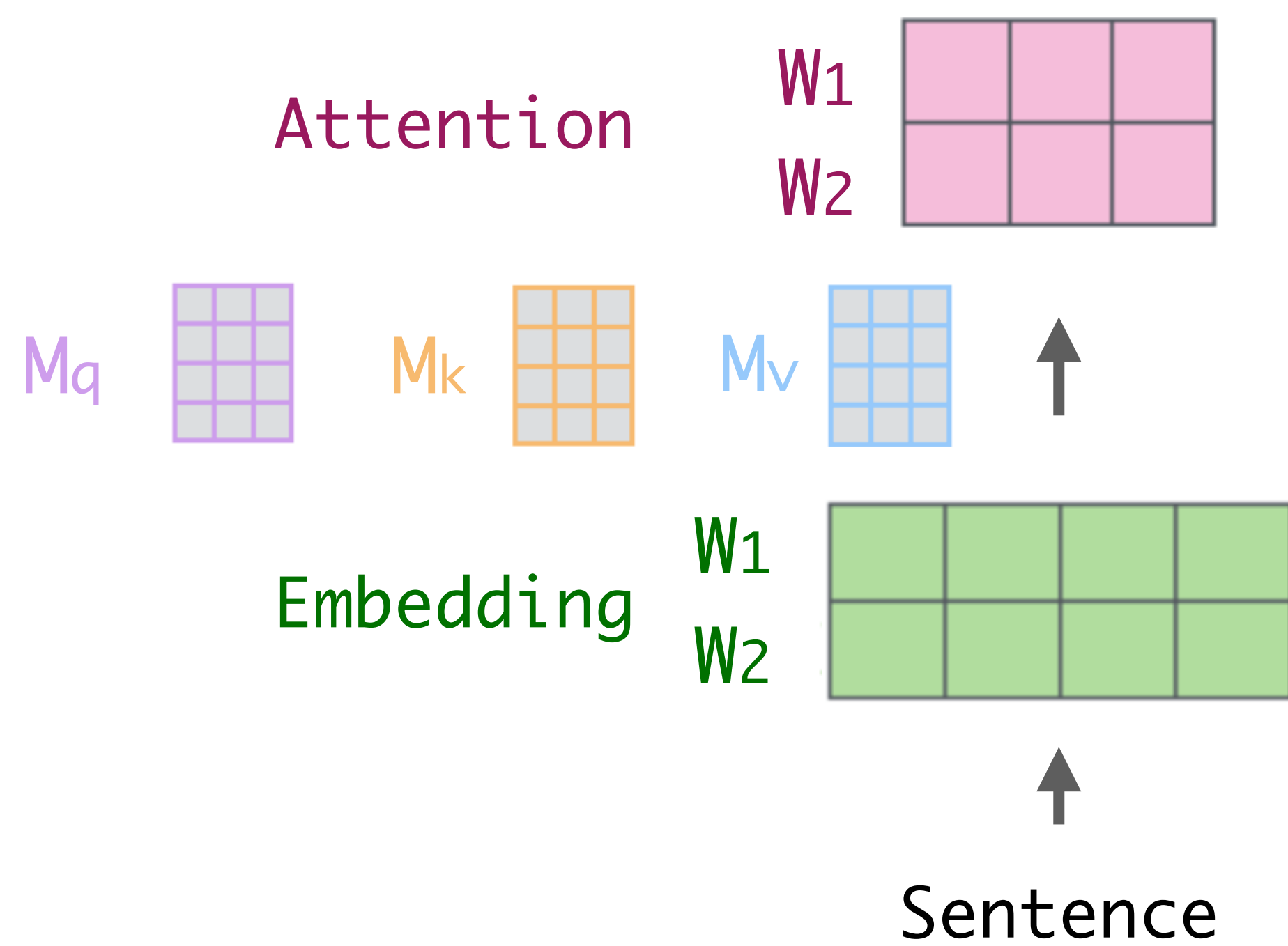


Outputs
(shifted right)

$N \times$

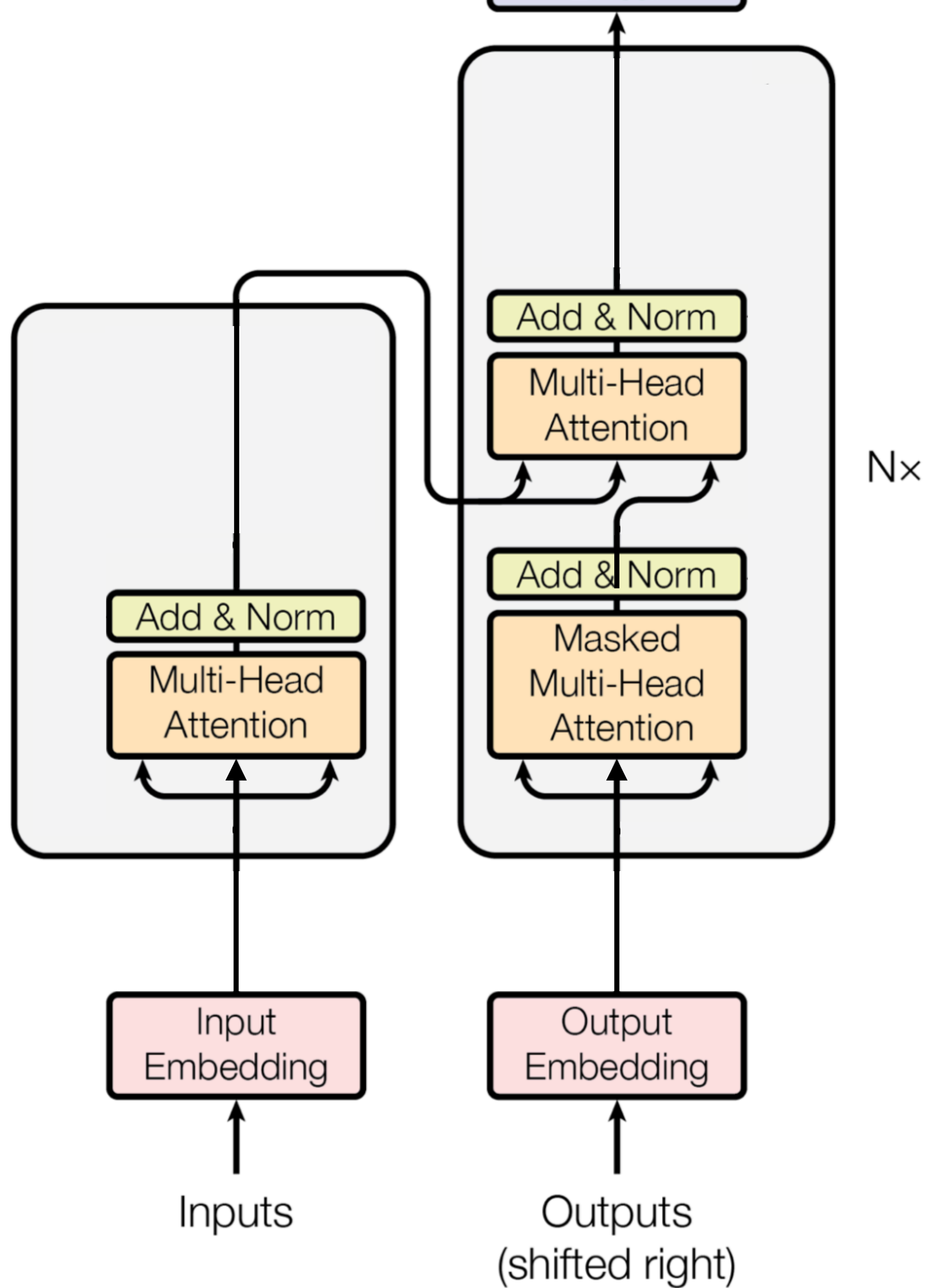
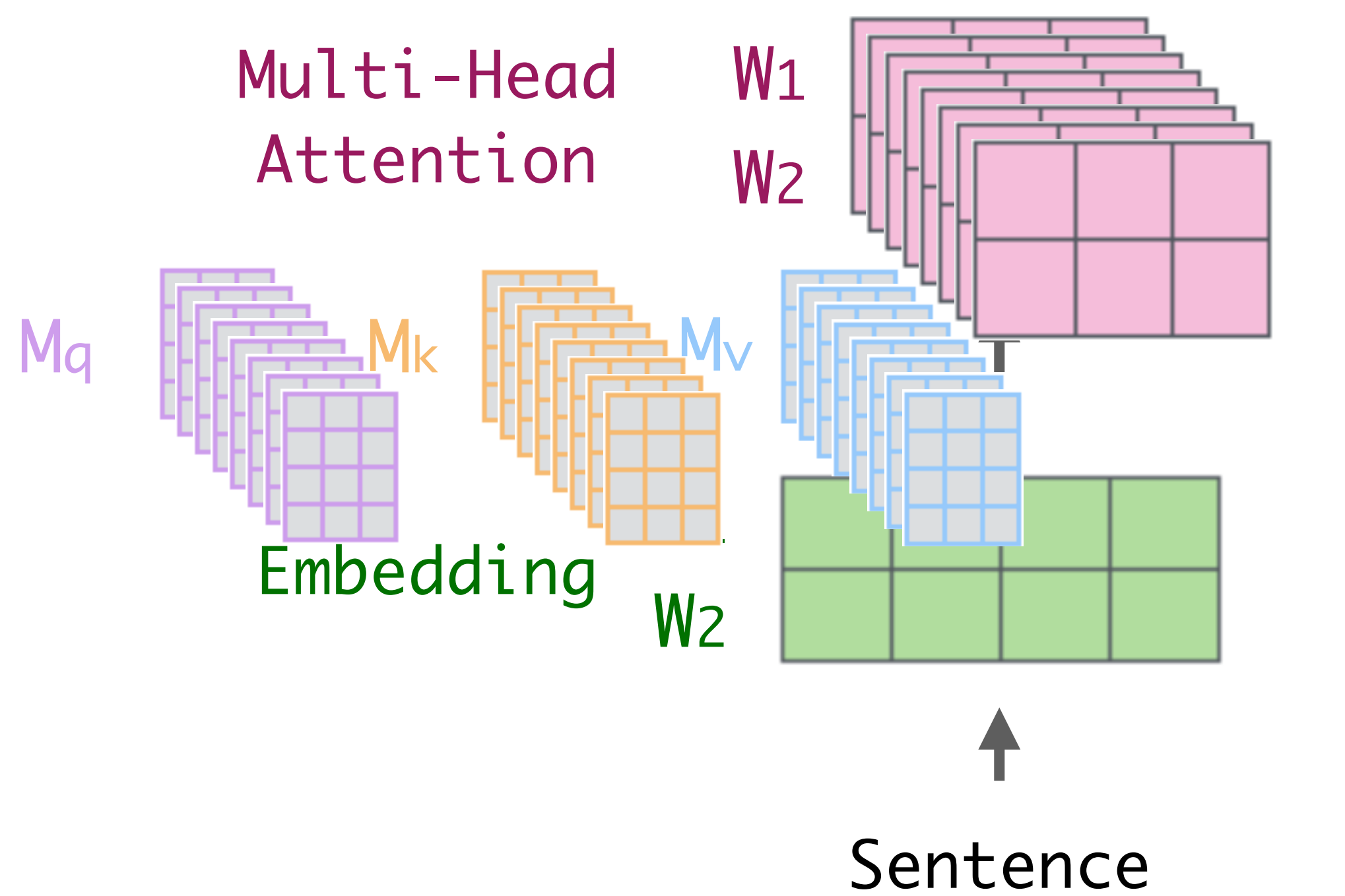
注意力

Attention



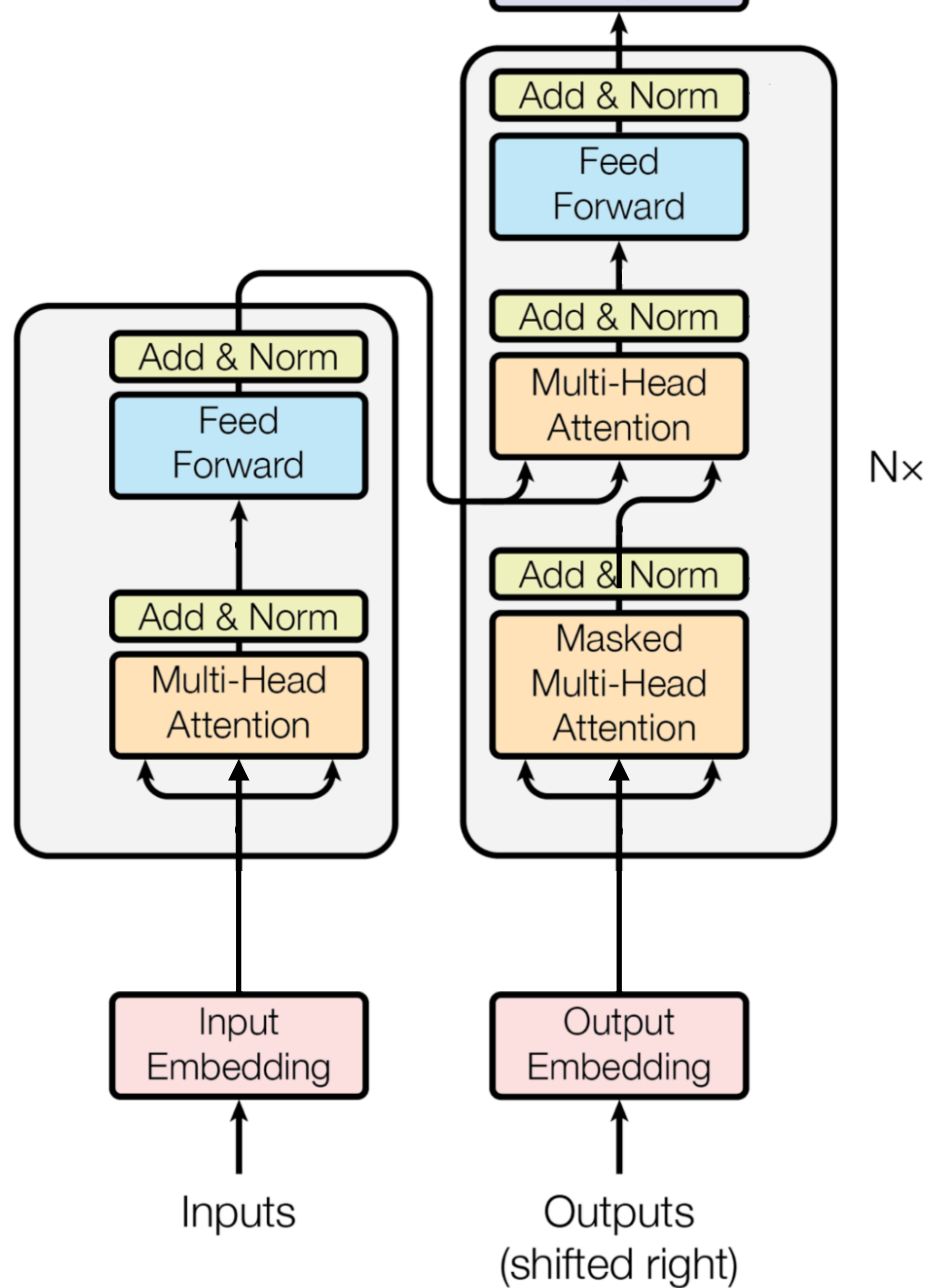
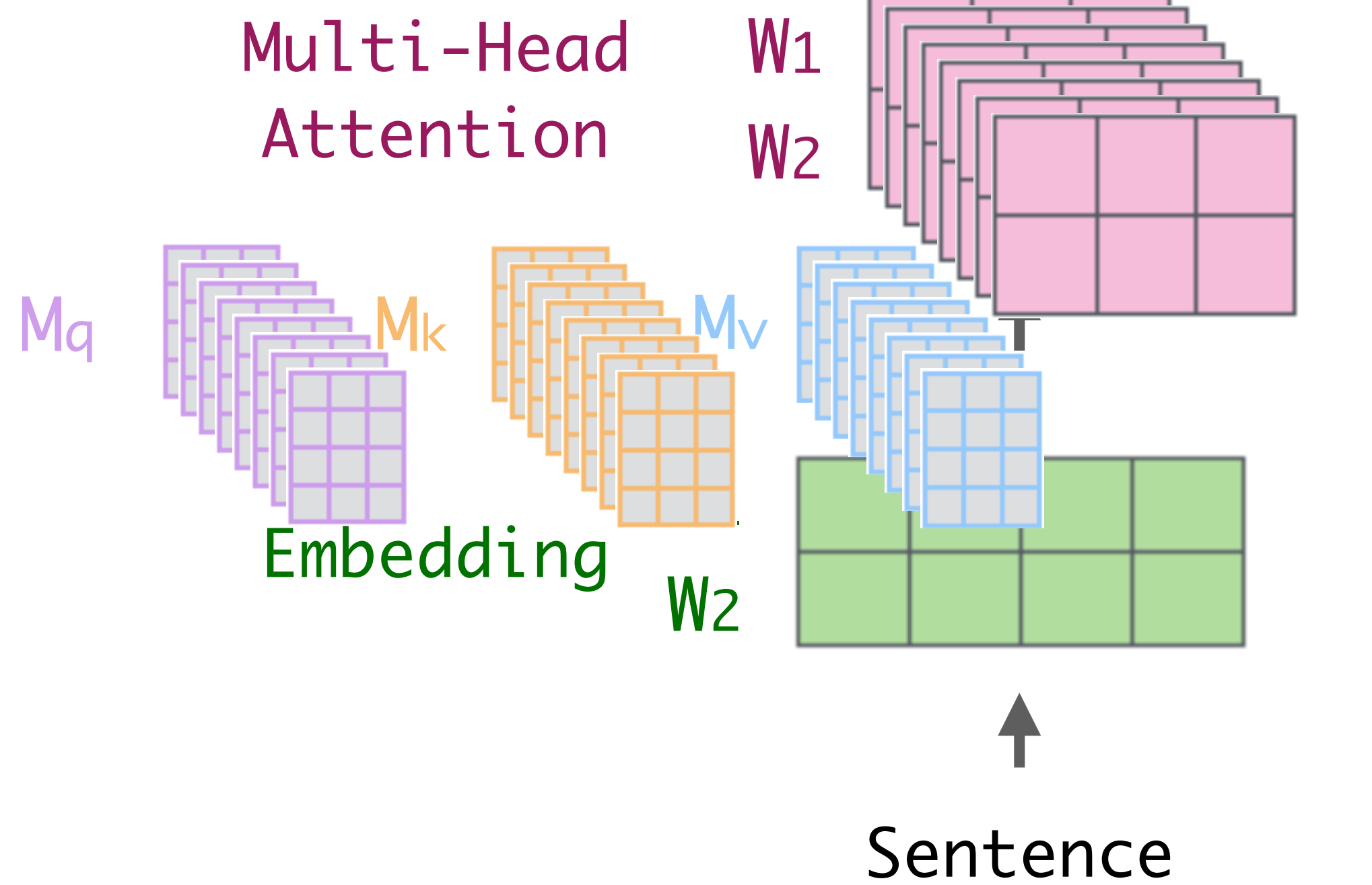
注意力

Attention



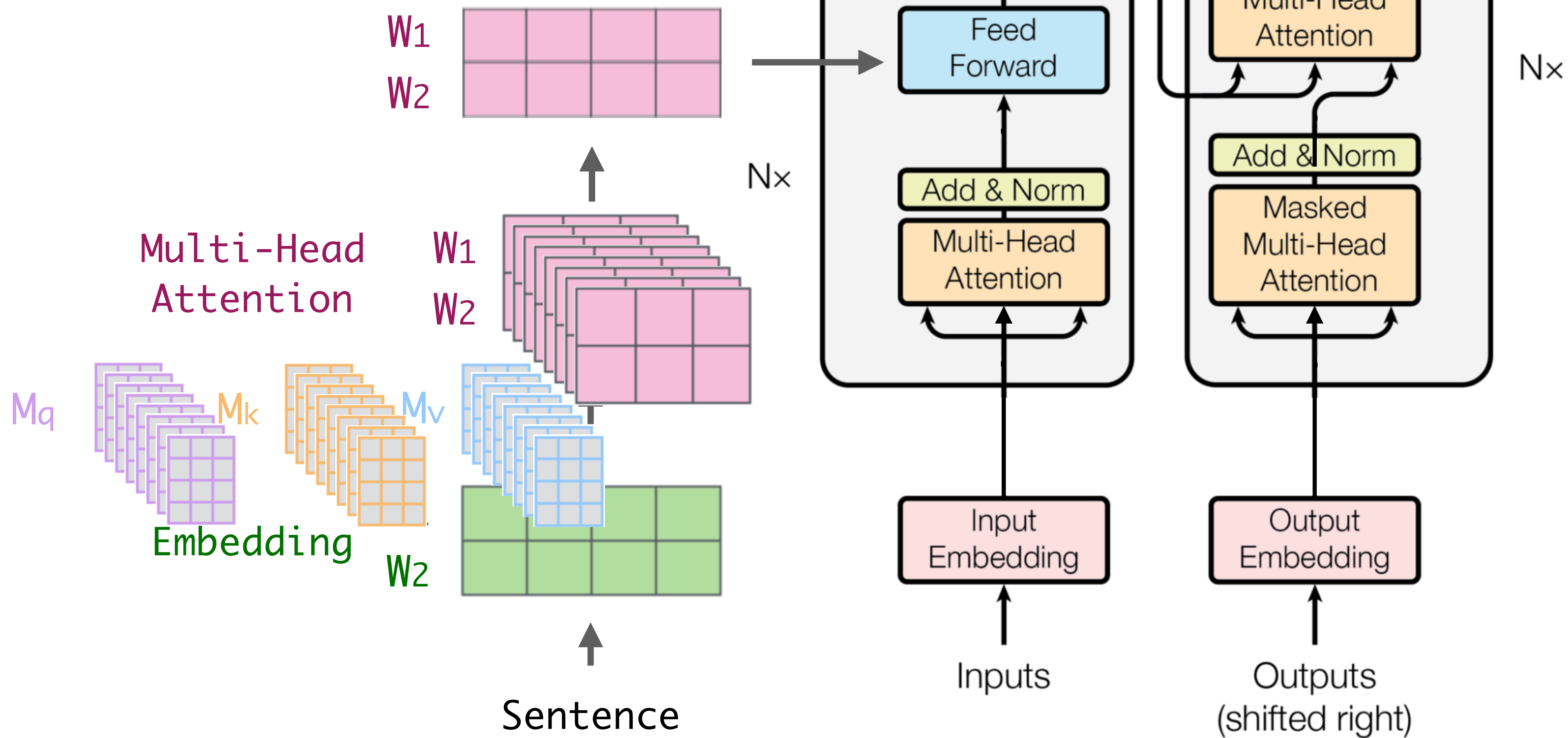
注意力

Attention



注意力

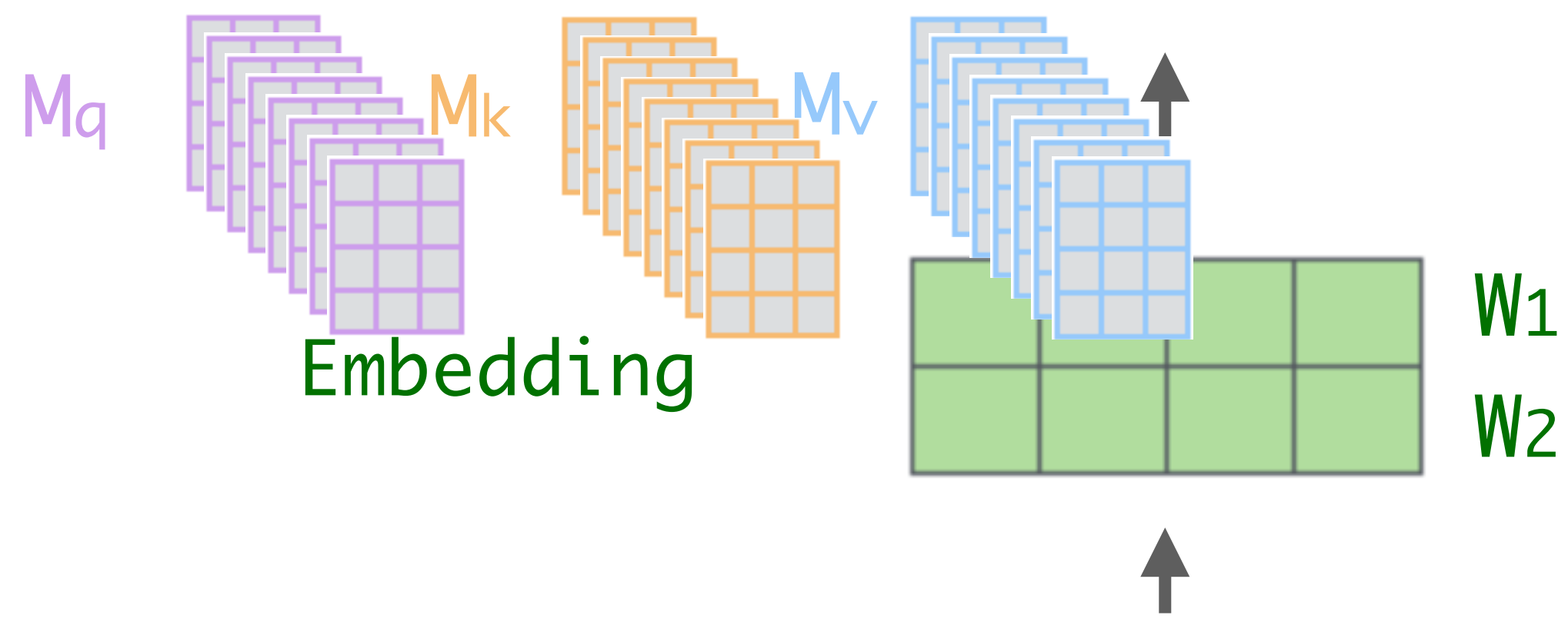
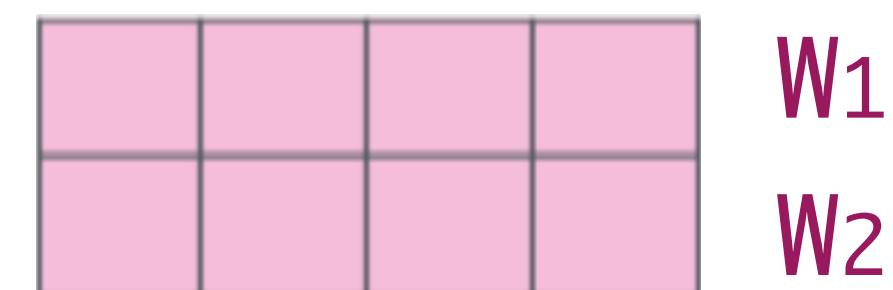
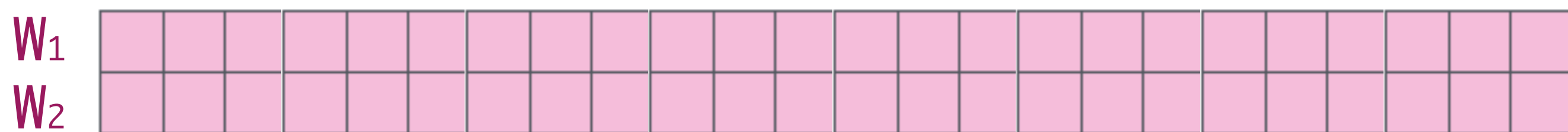
Attention



注意力

Attention

Multi-Head
Attention

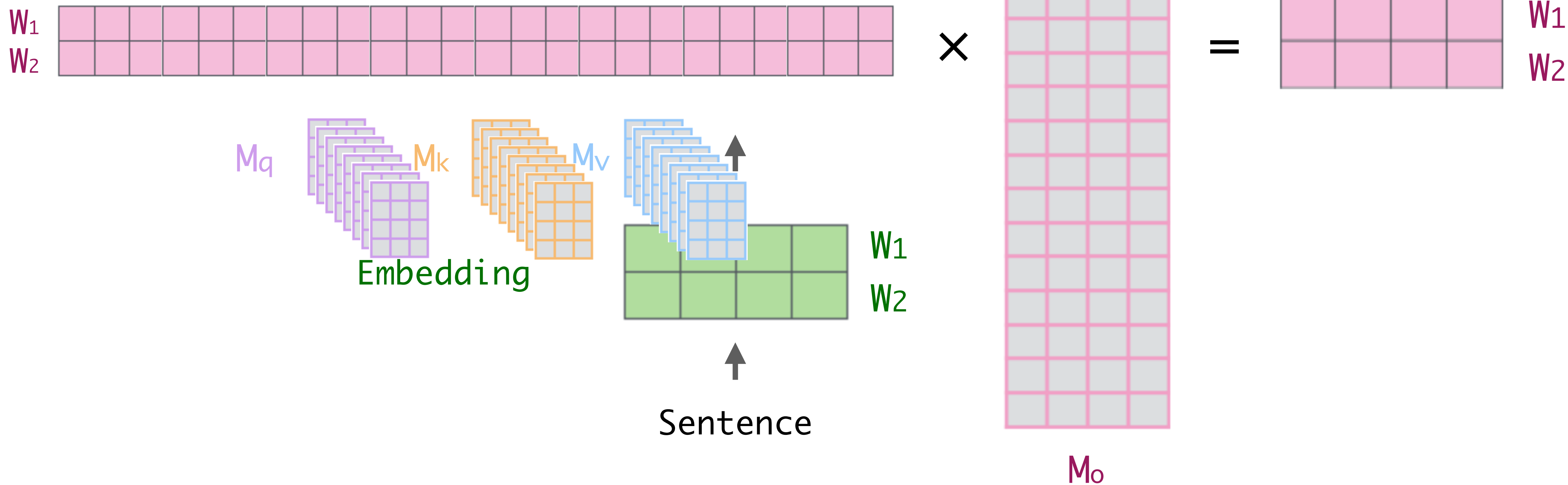


Sentence

注意力

Attention

Multi-Head
Attention



注意力

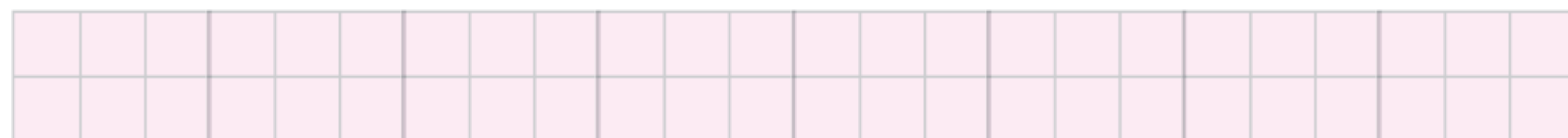
Attention

Multi-Head
Attention

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O$$

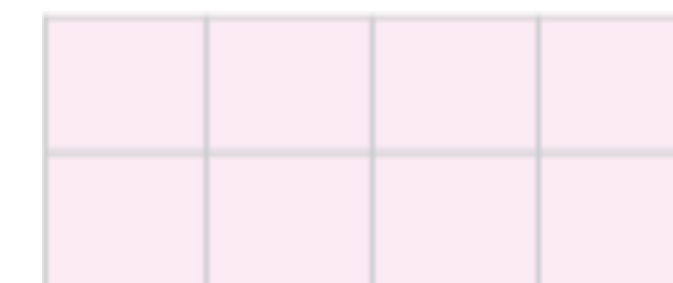
$$\text{where head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

W_1
 W_2

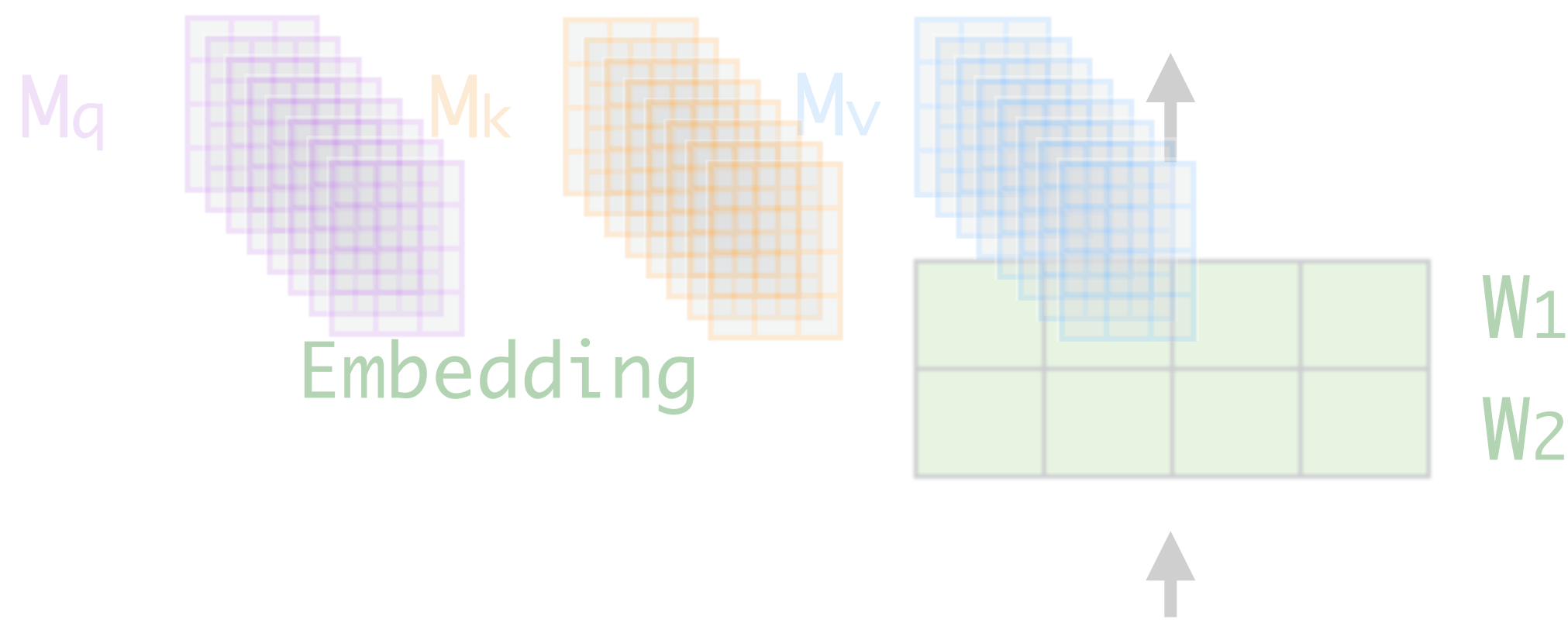


\times

$=$



W_1
 W_2

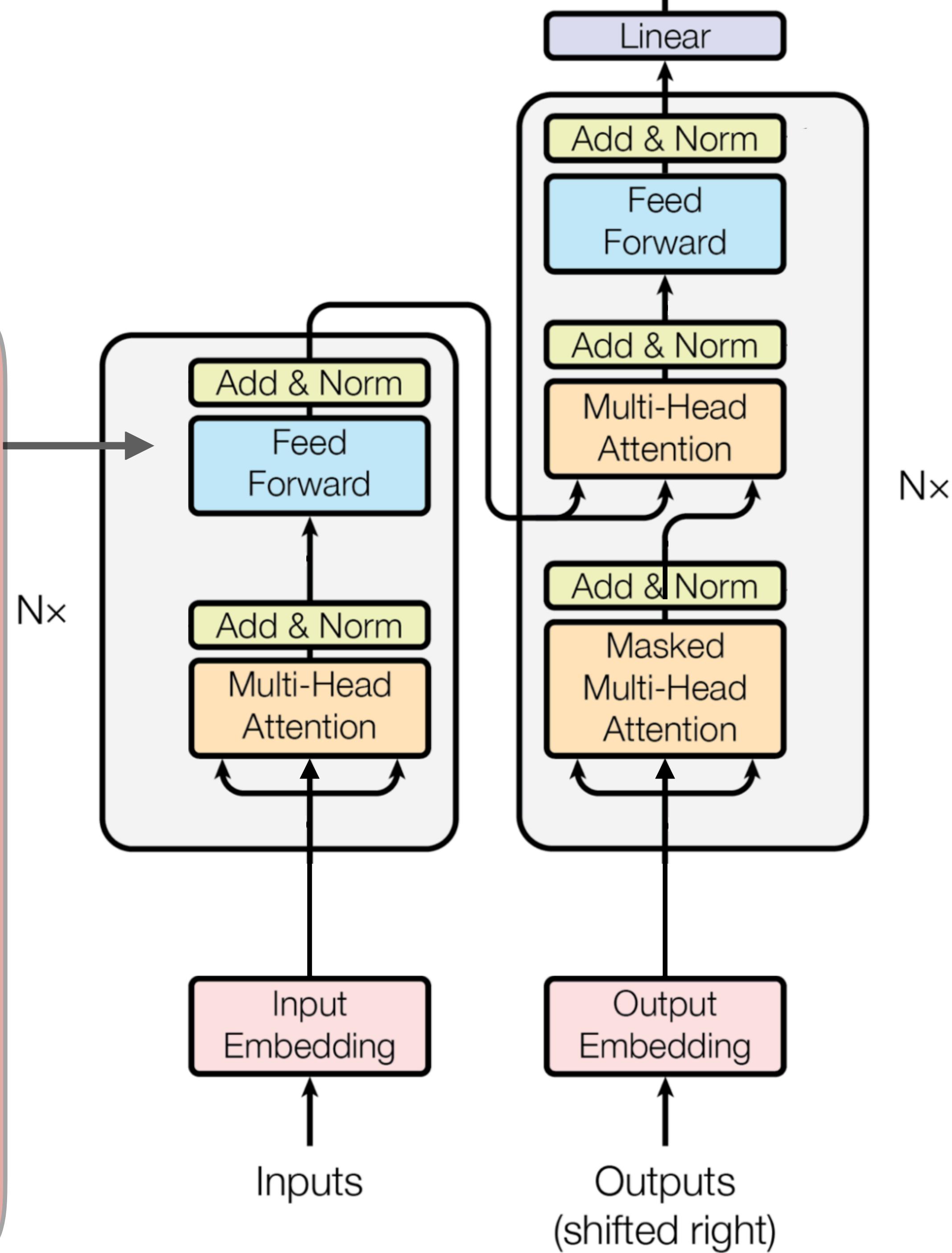
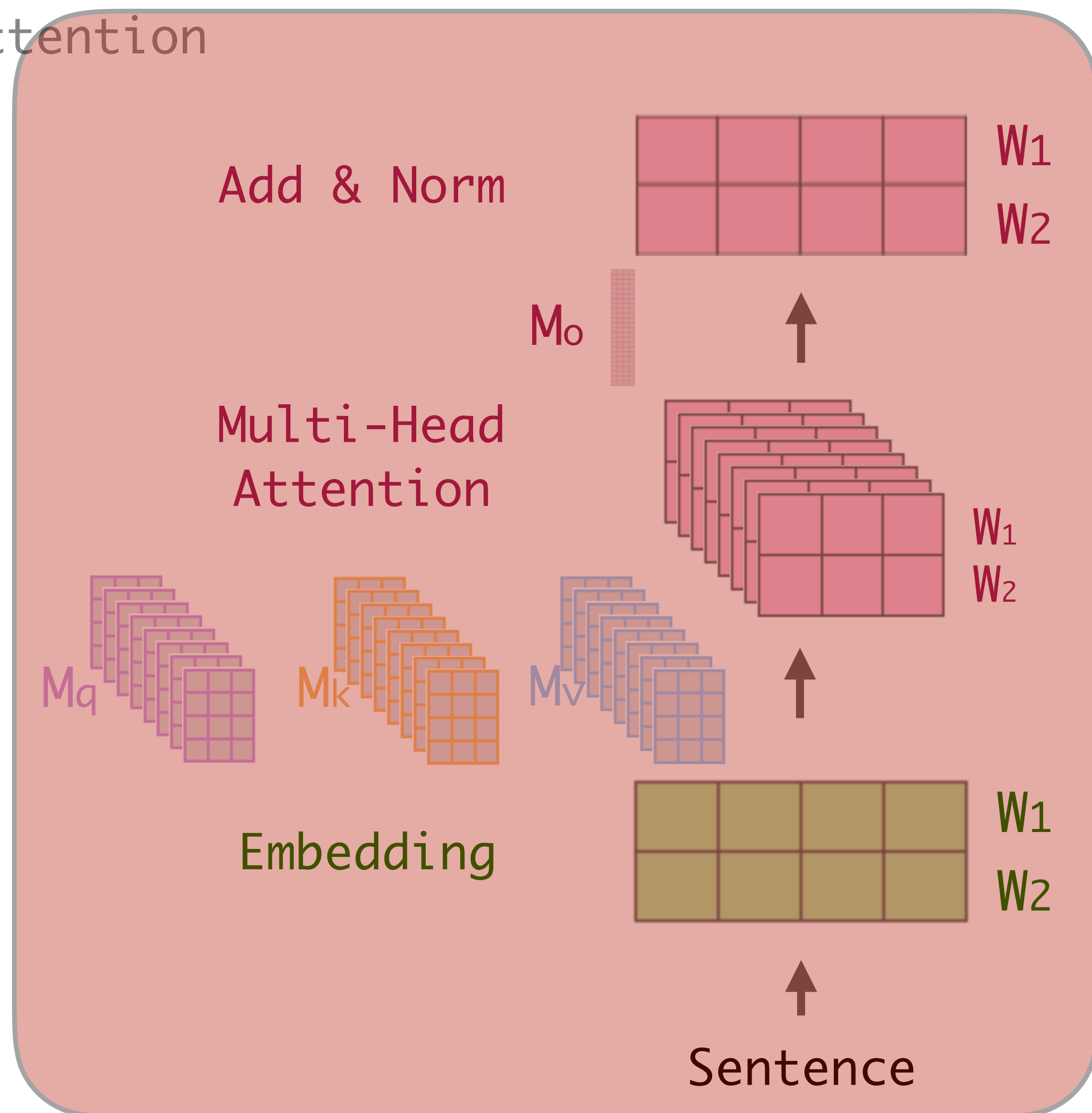


Sentence

M_o

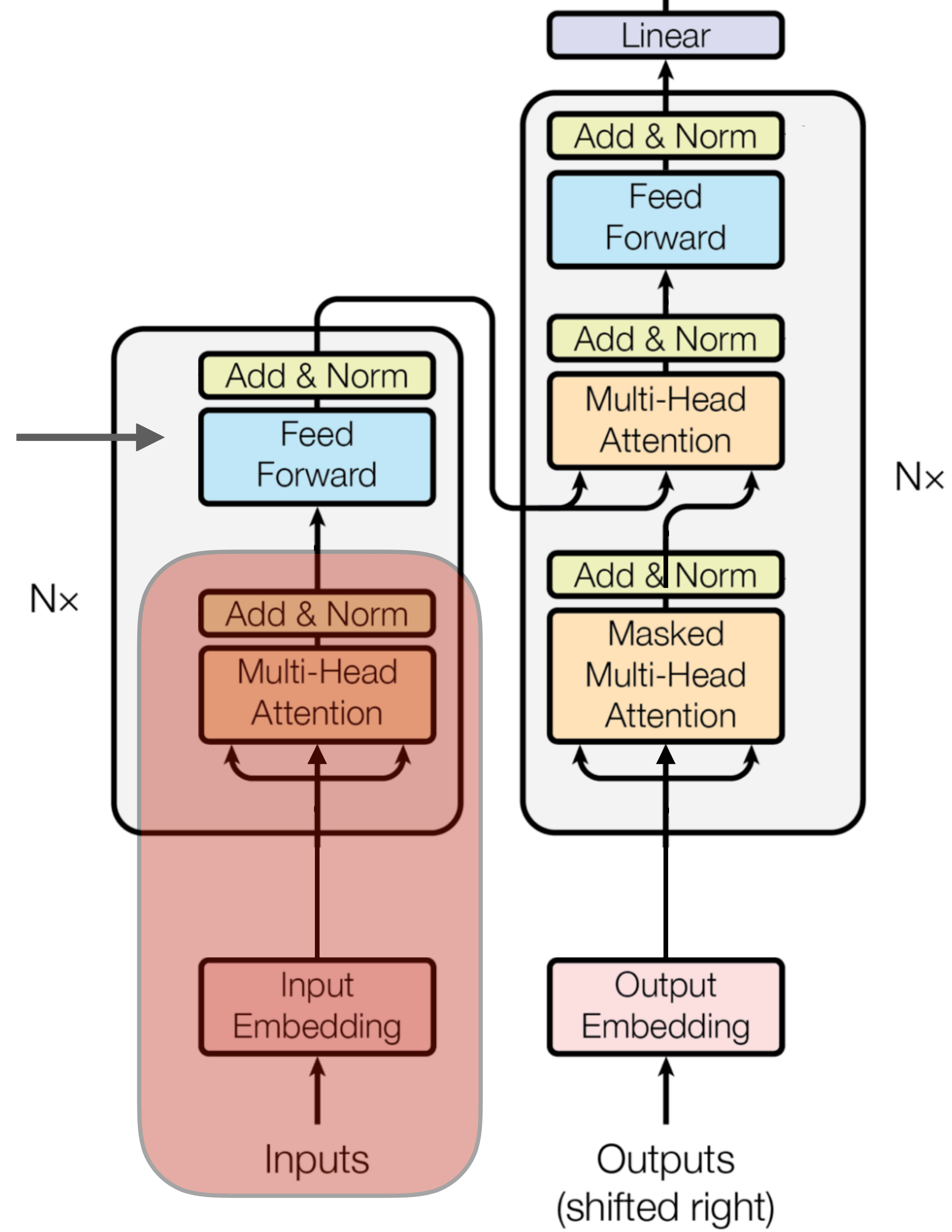
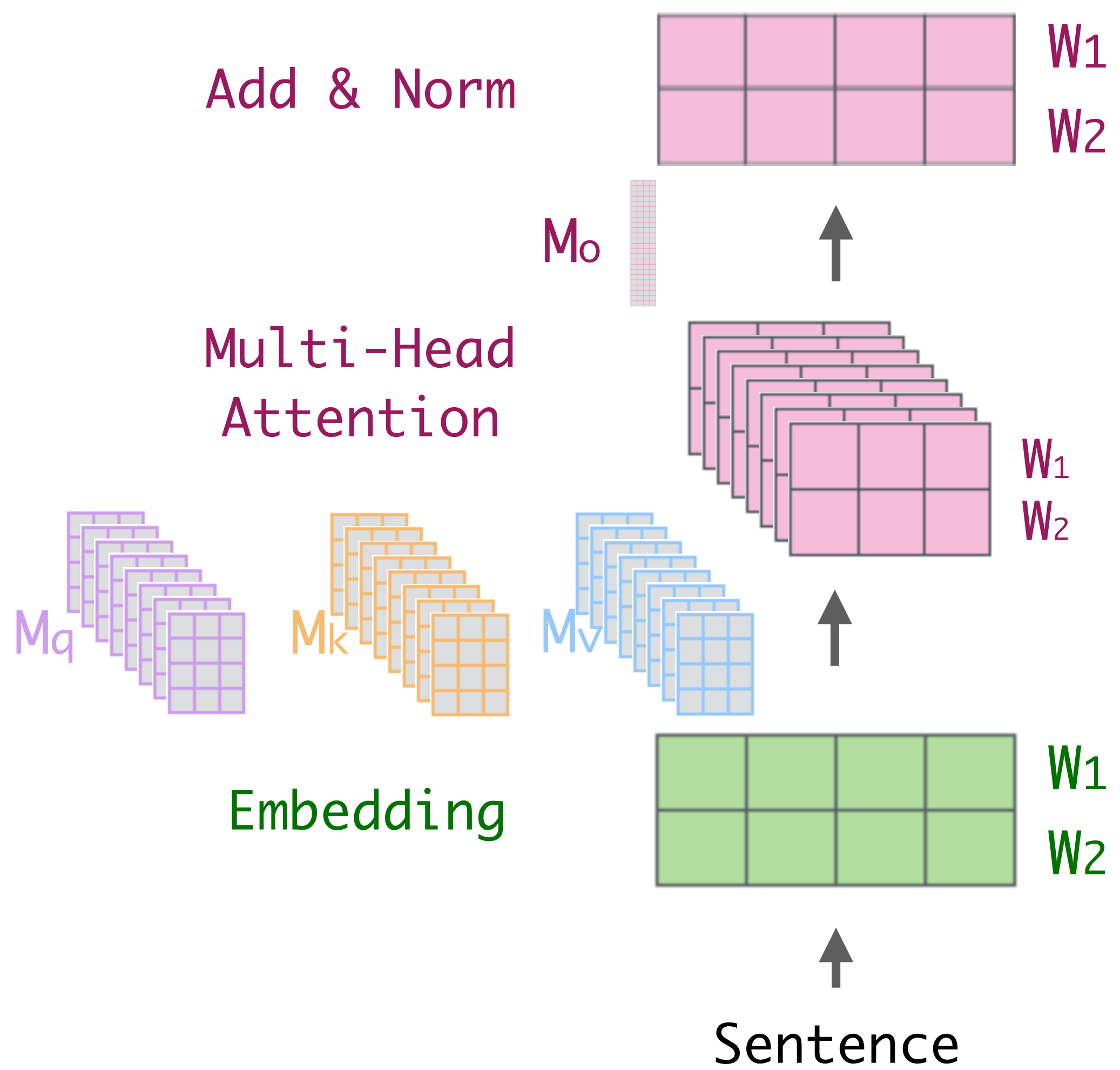
注意力

Attention



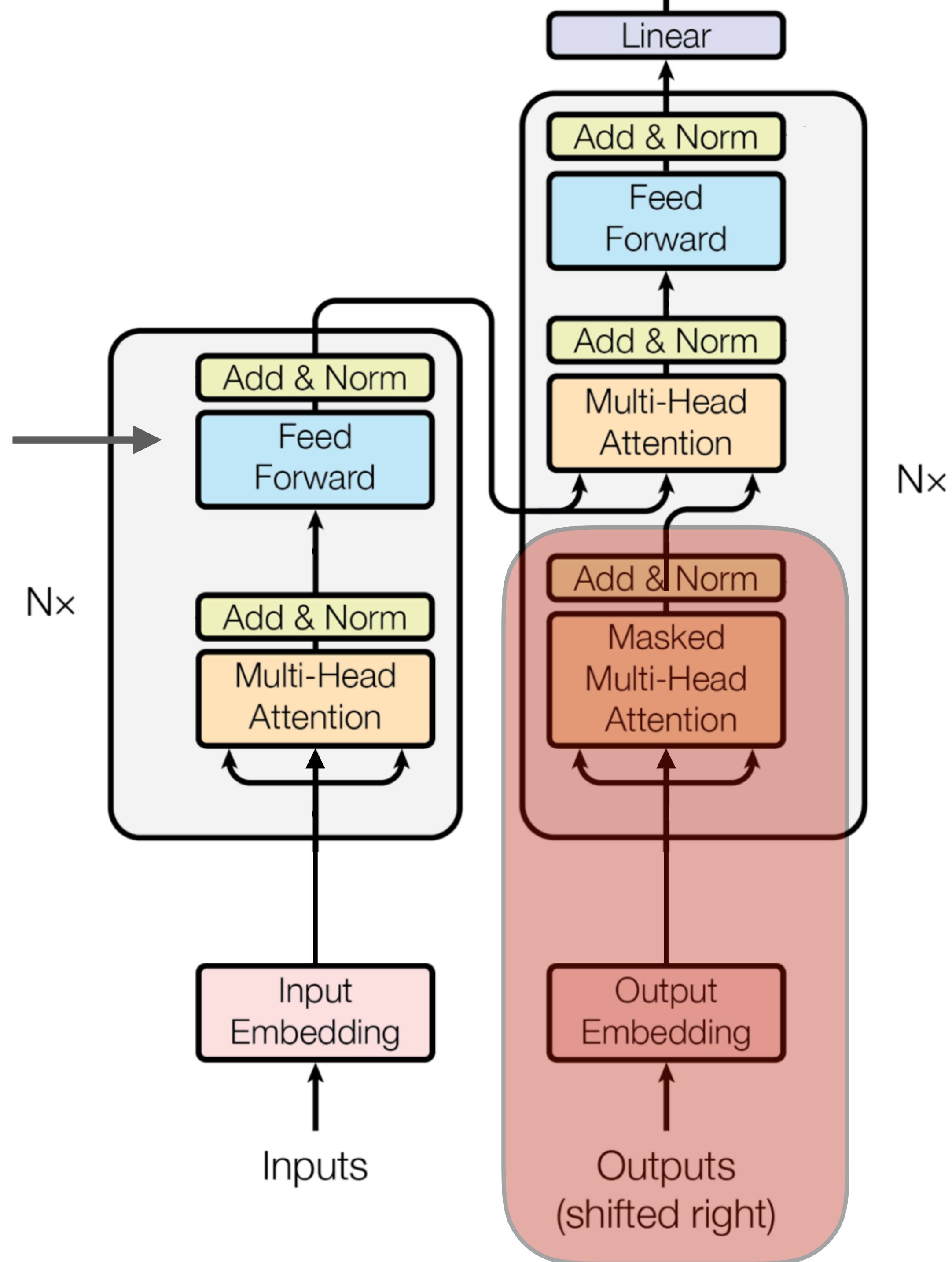
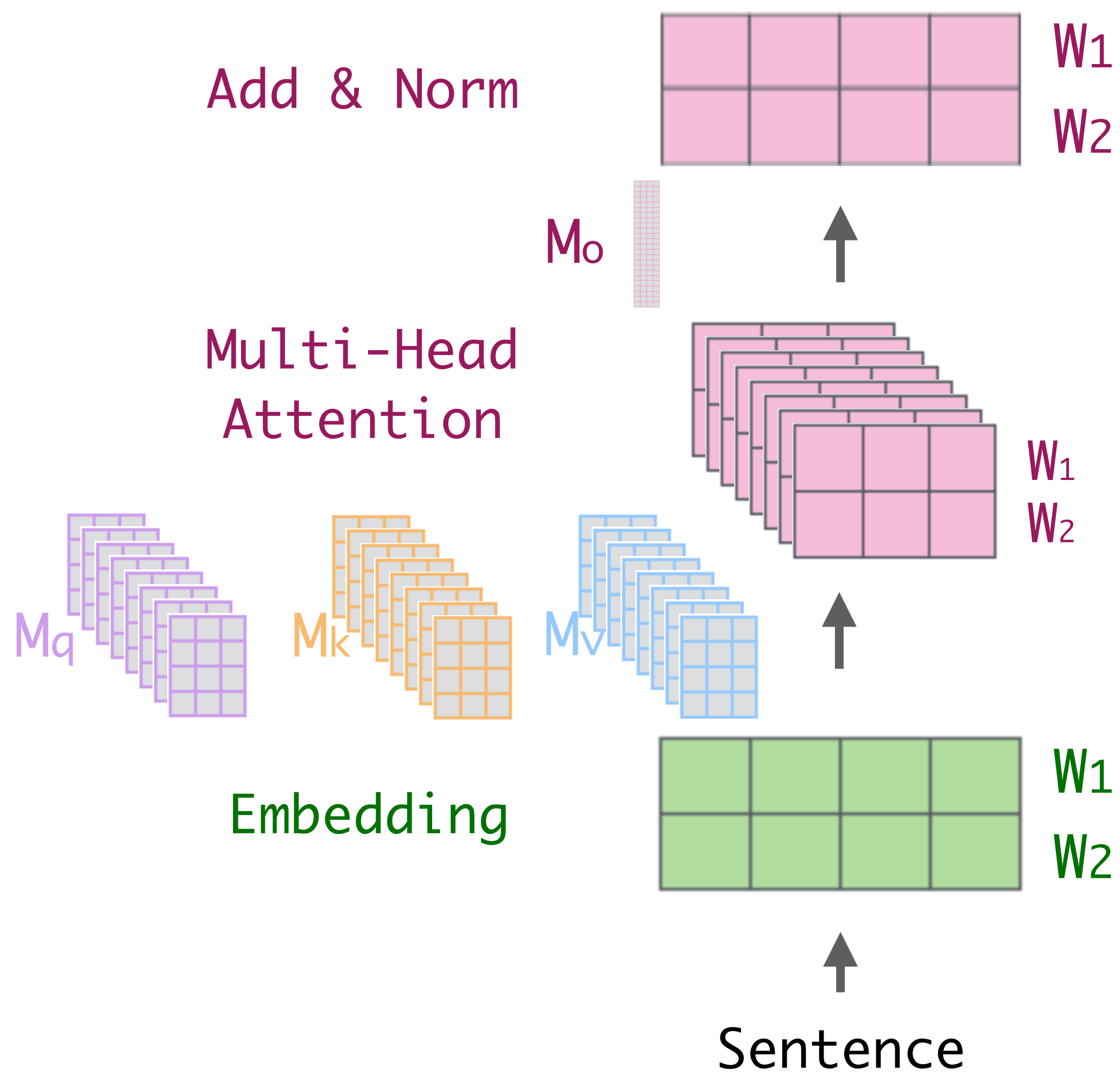
注意力

Attention



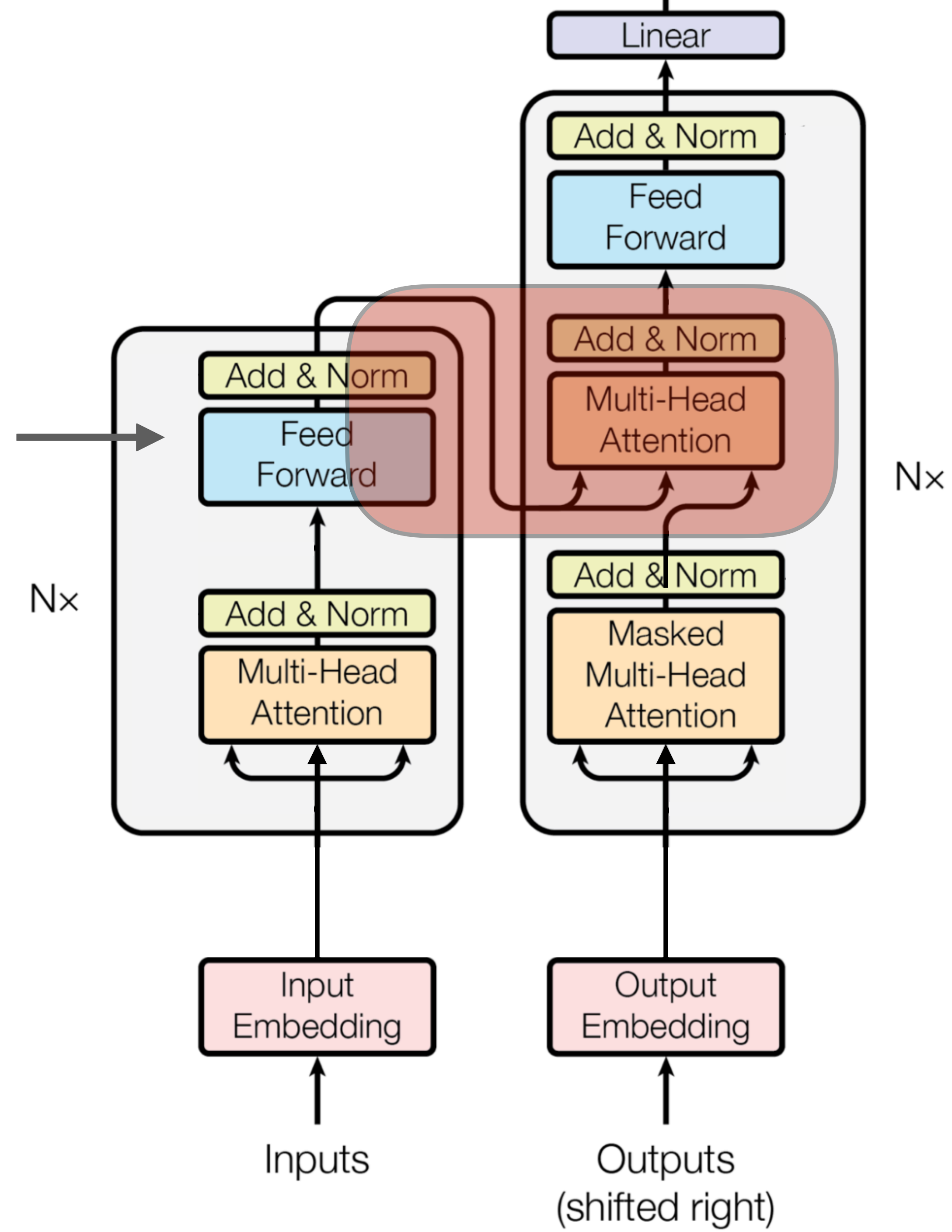
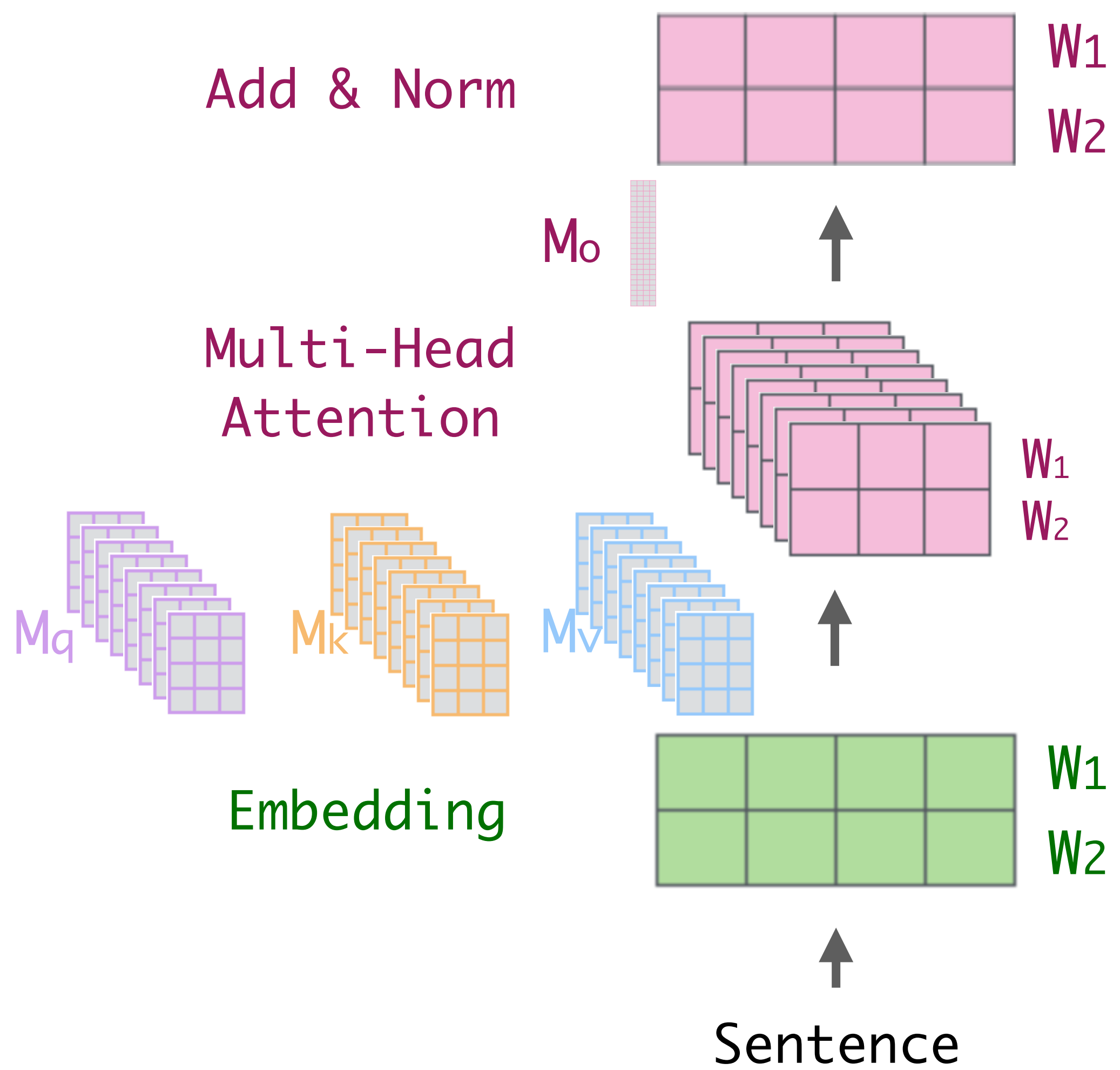
注意力

Attention

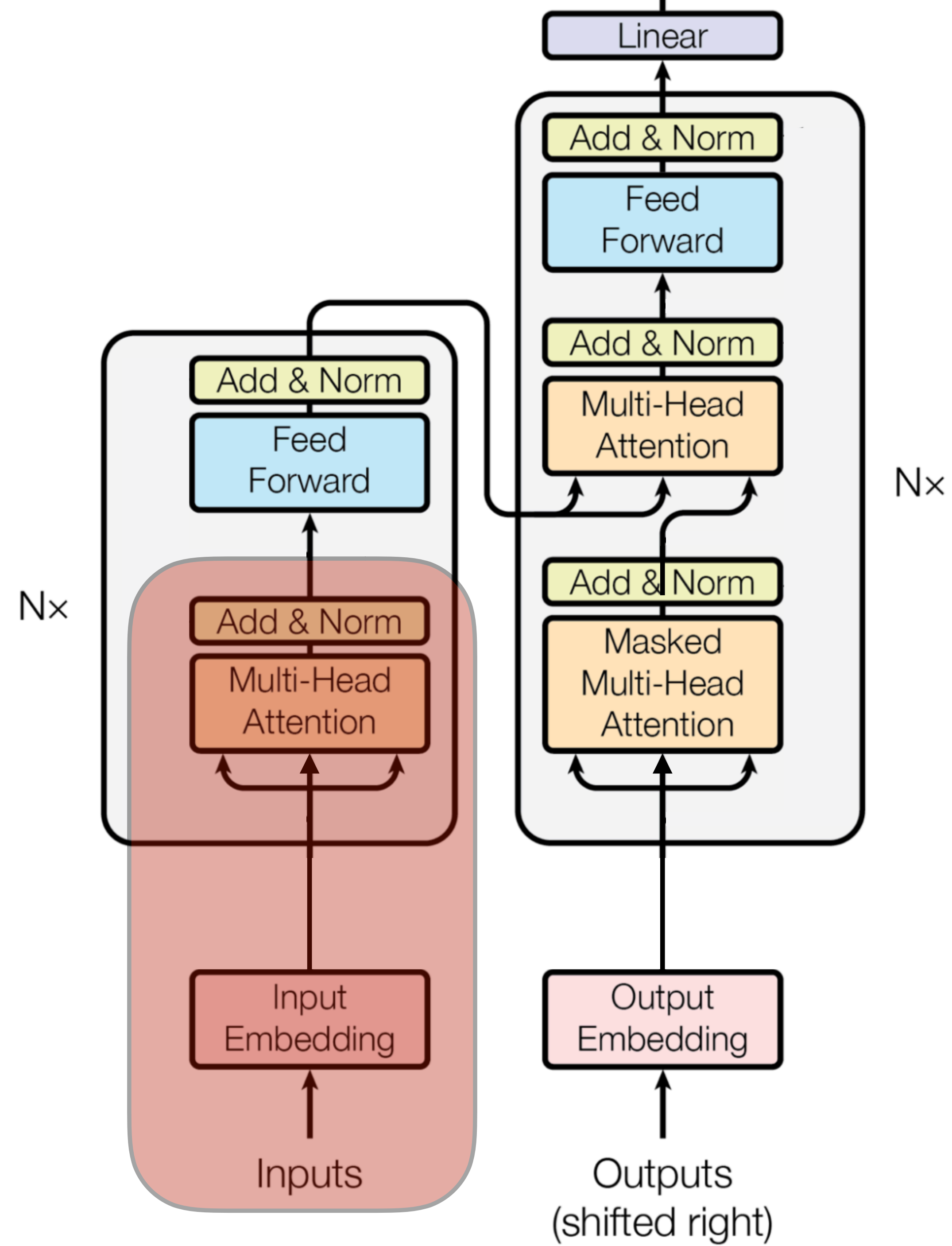
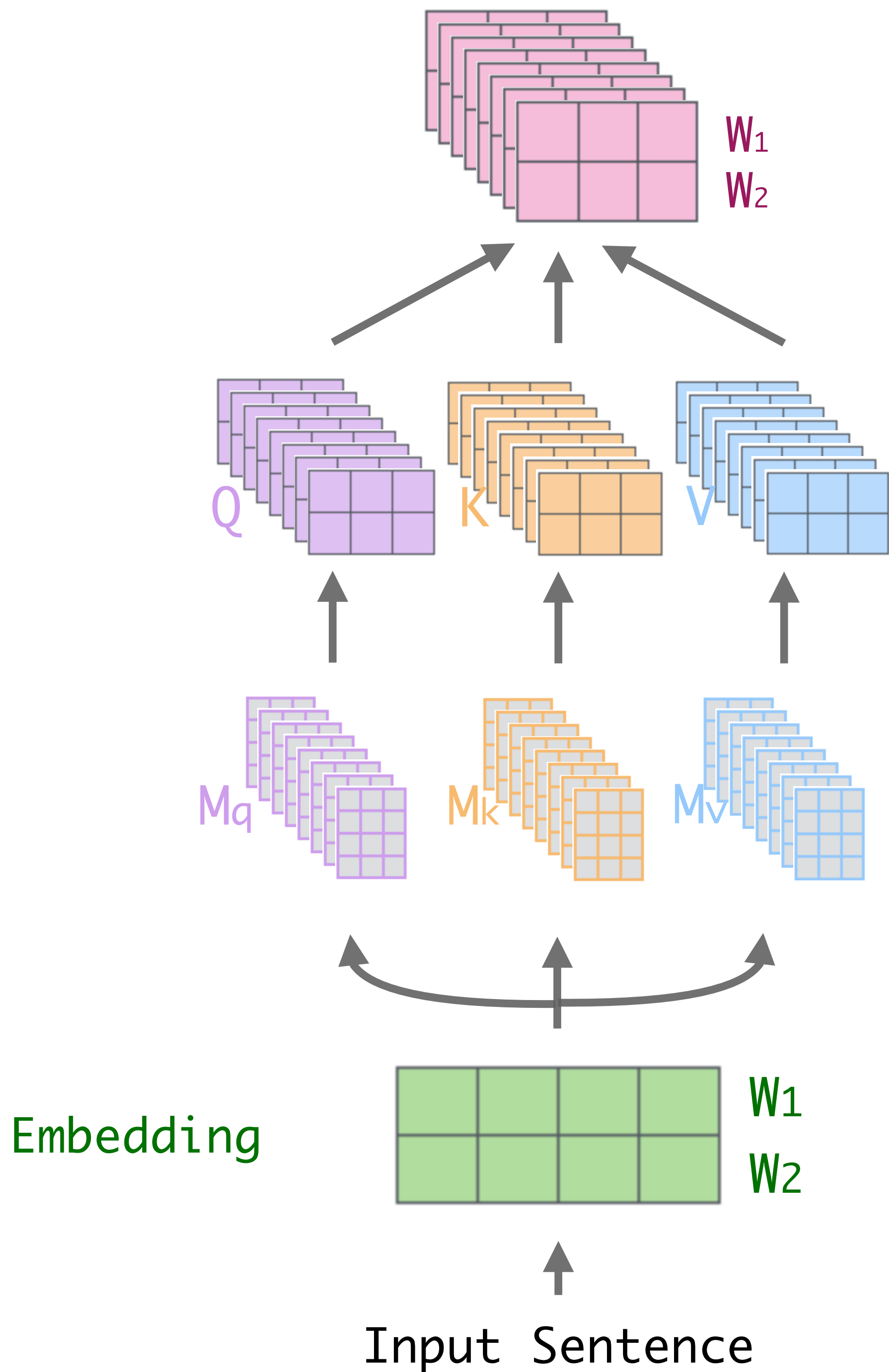


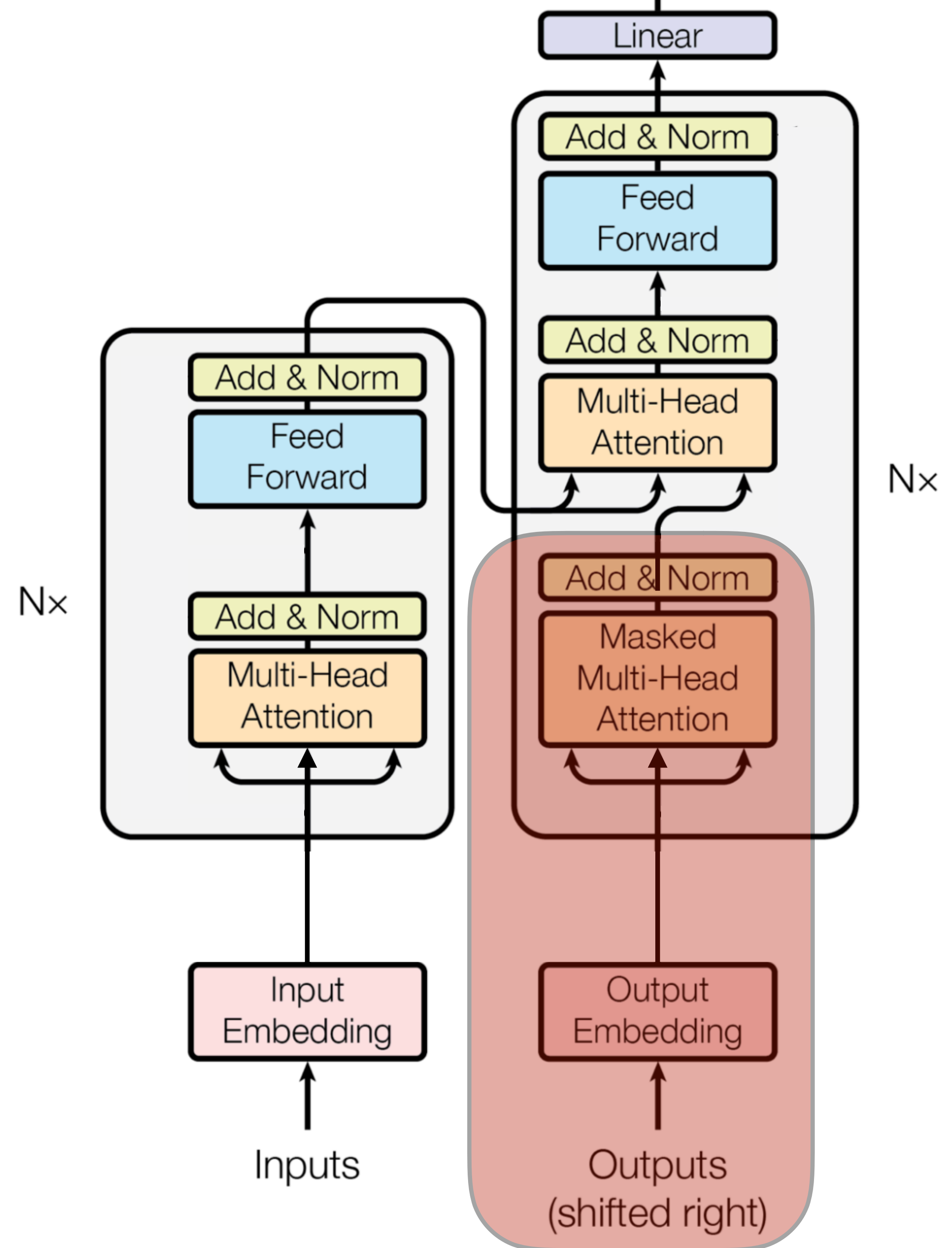
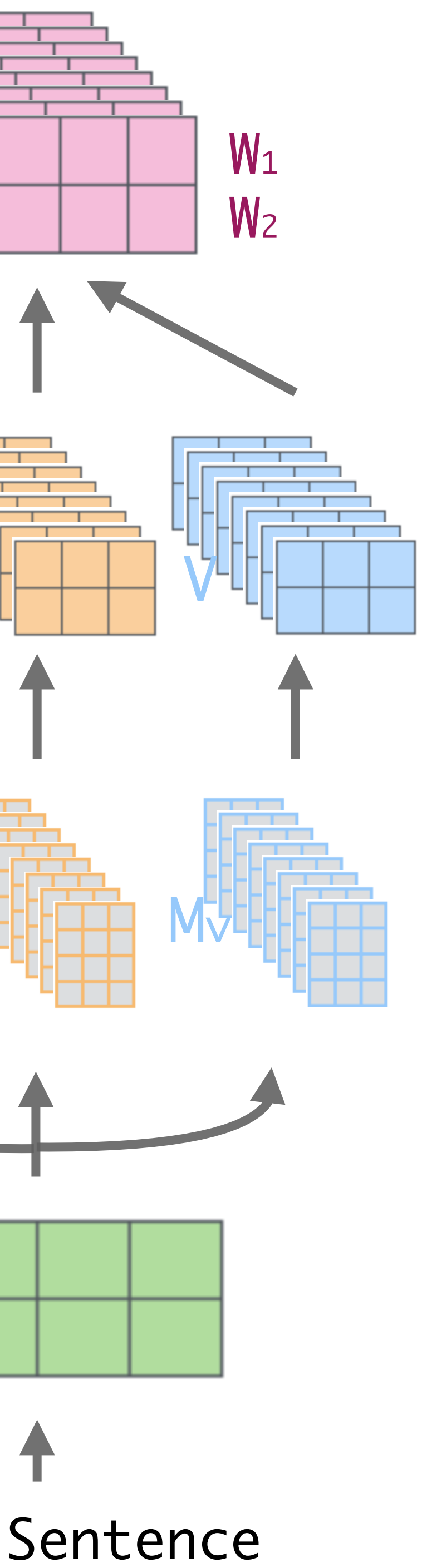
注意力

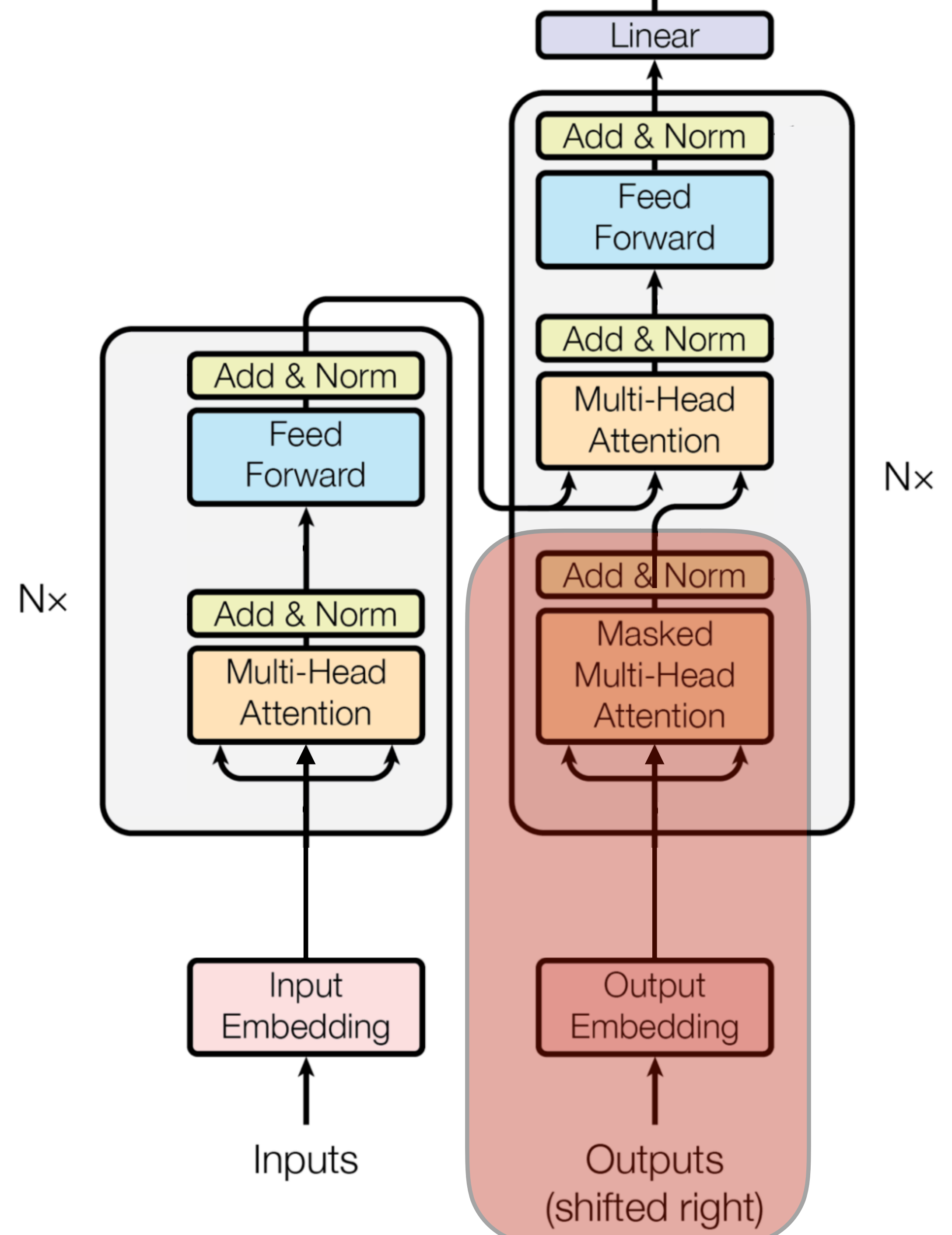
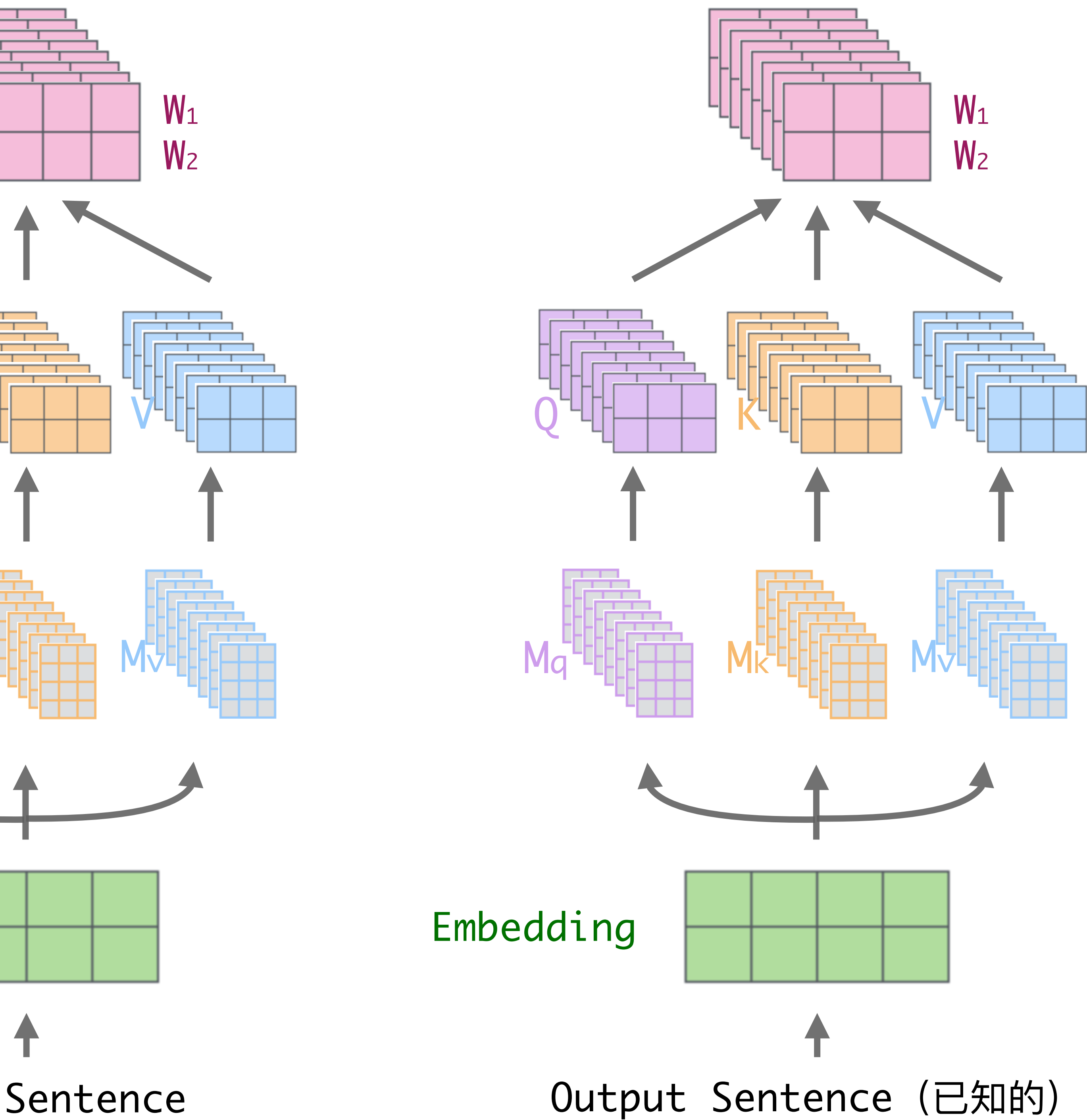
Attention

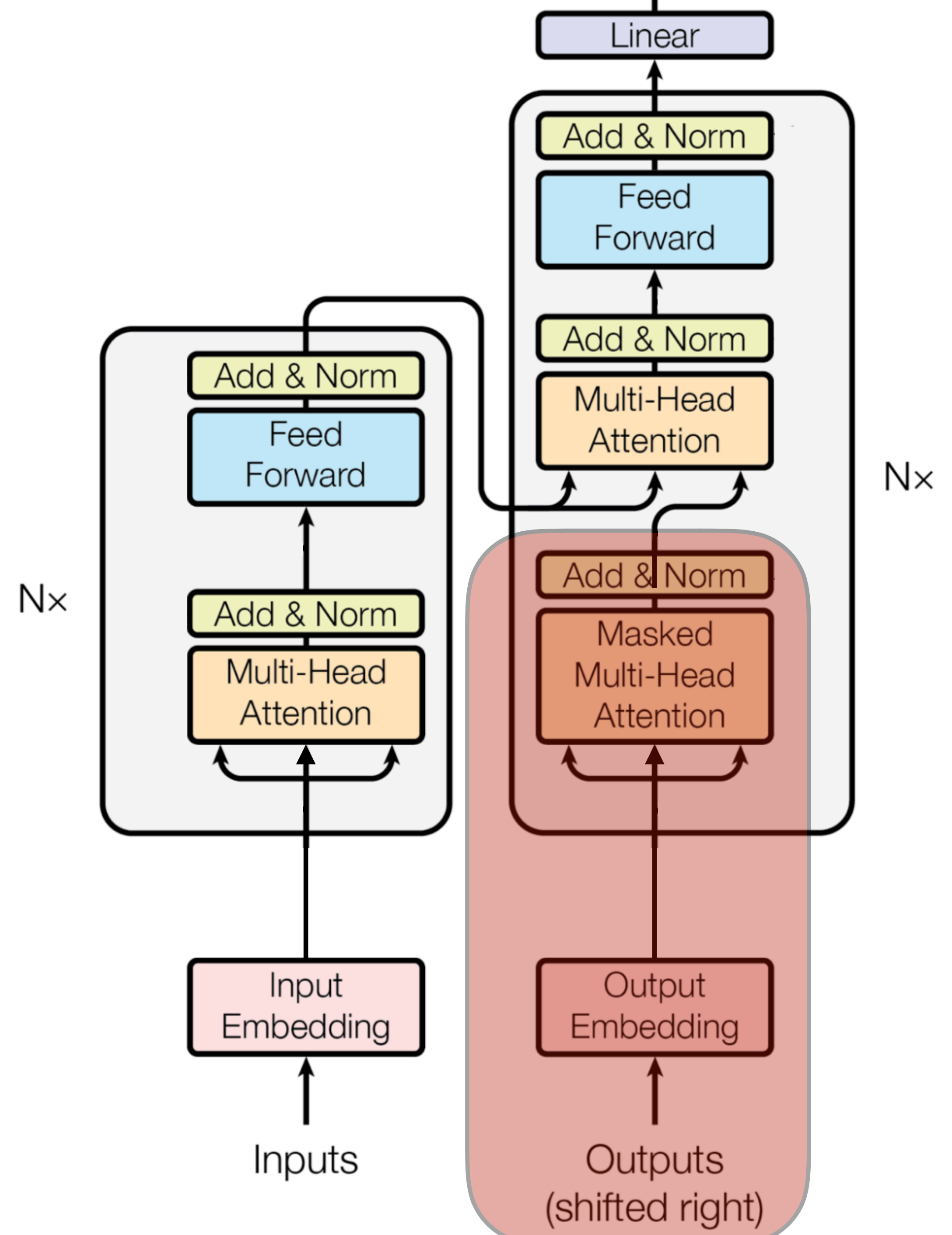
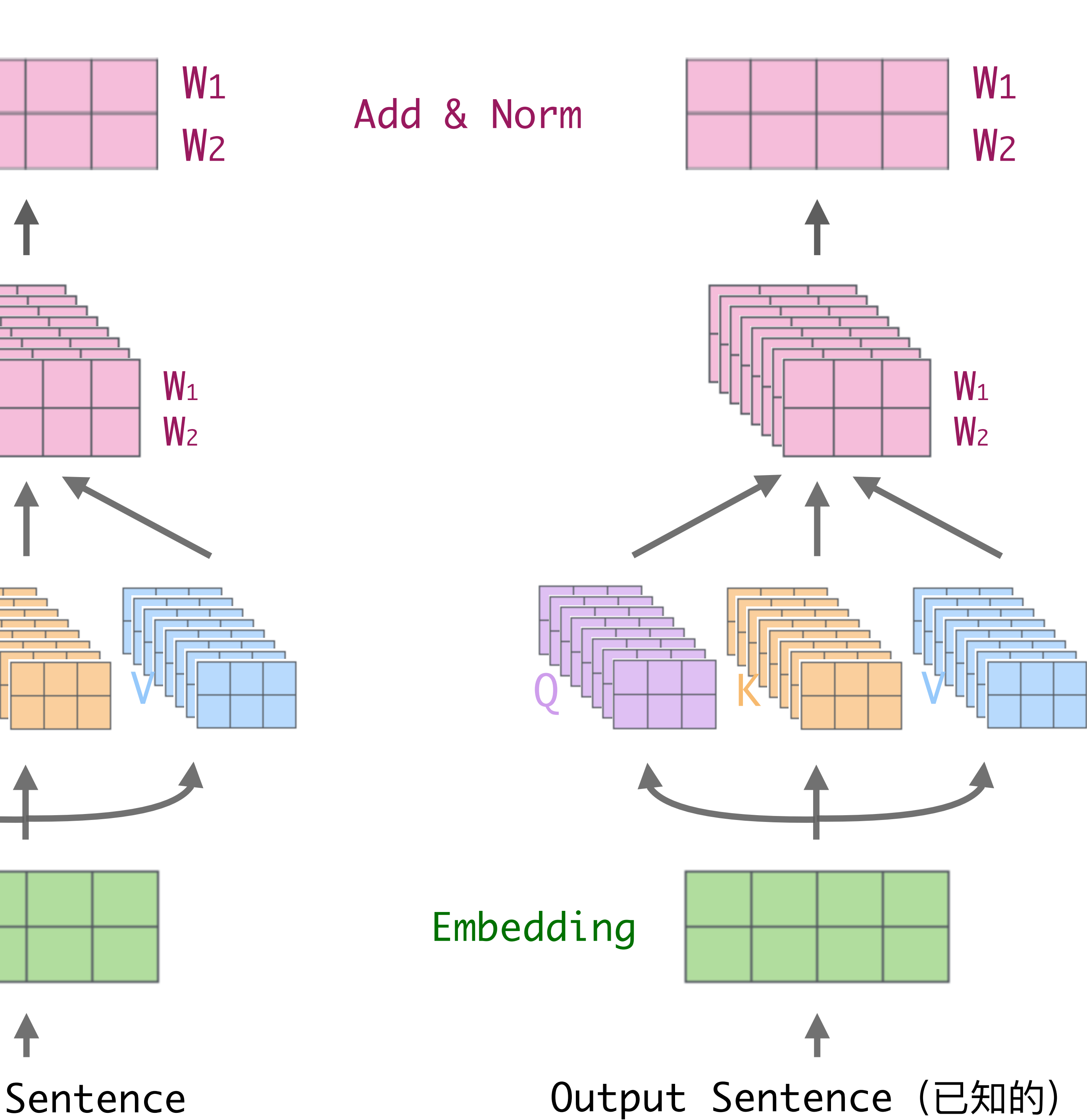


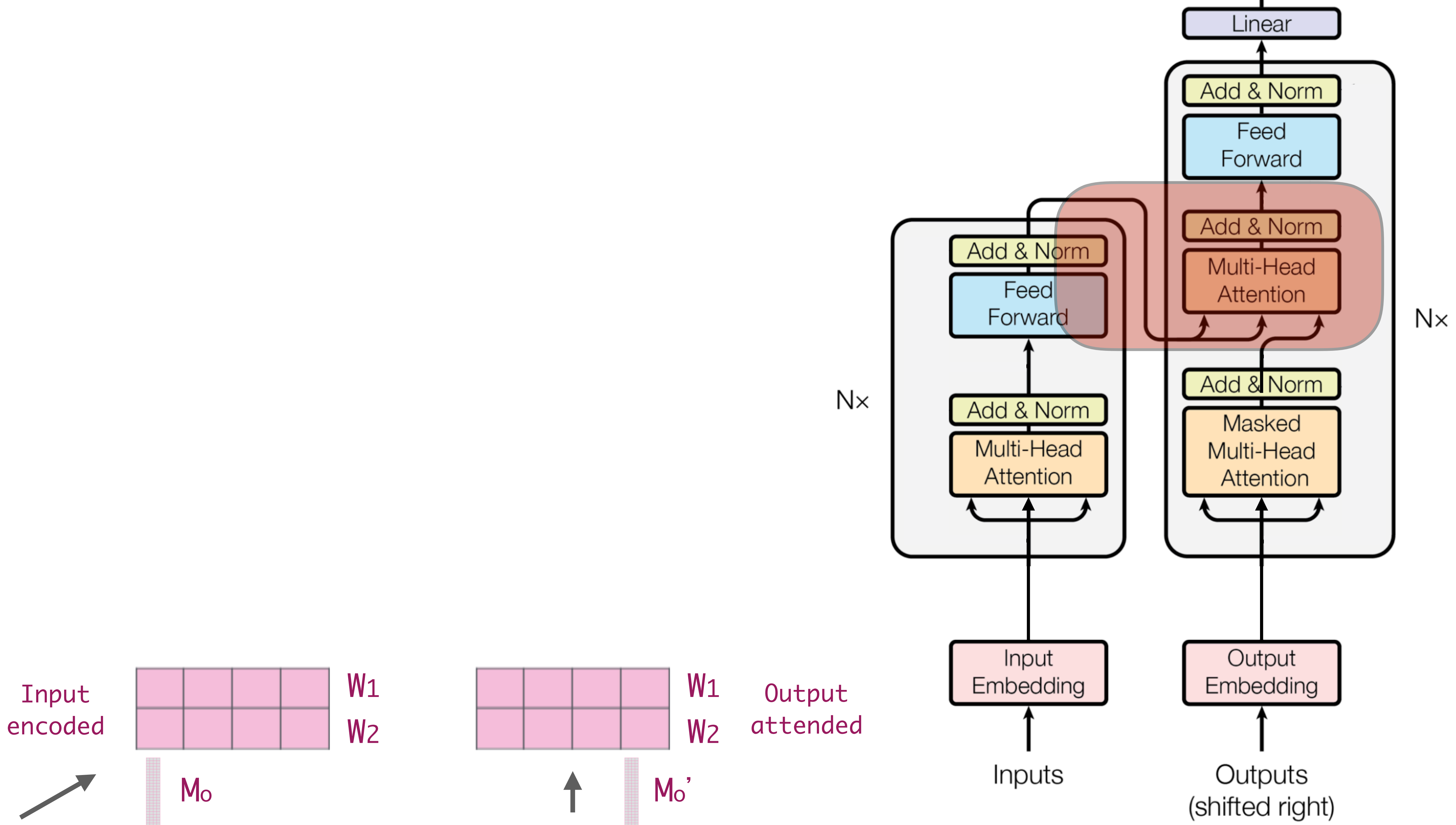
Multi-Head Attention

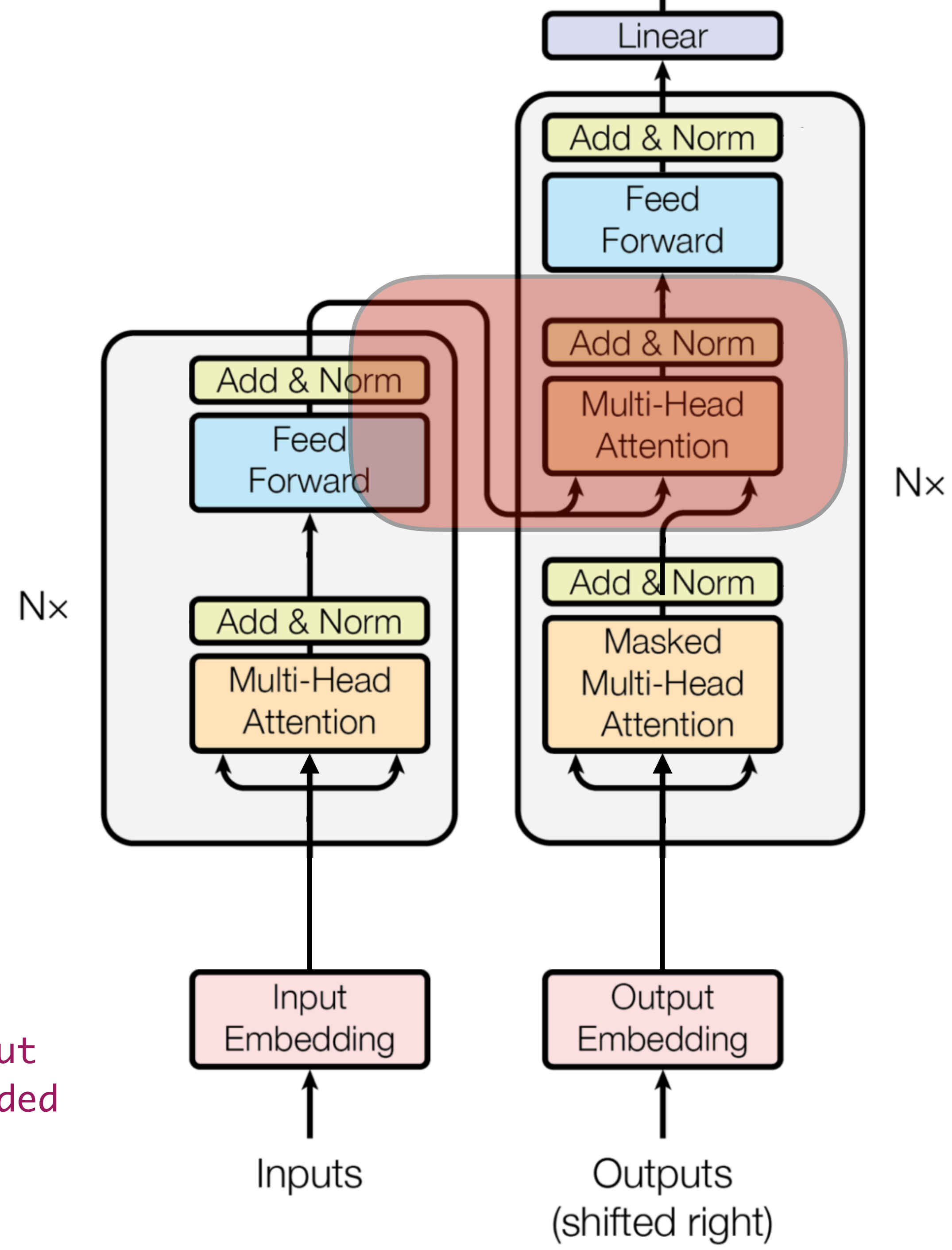
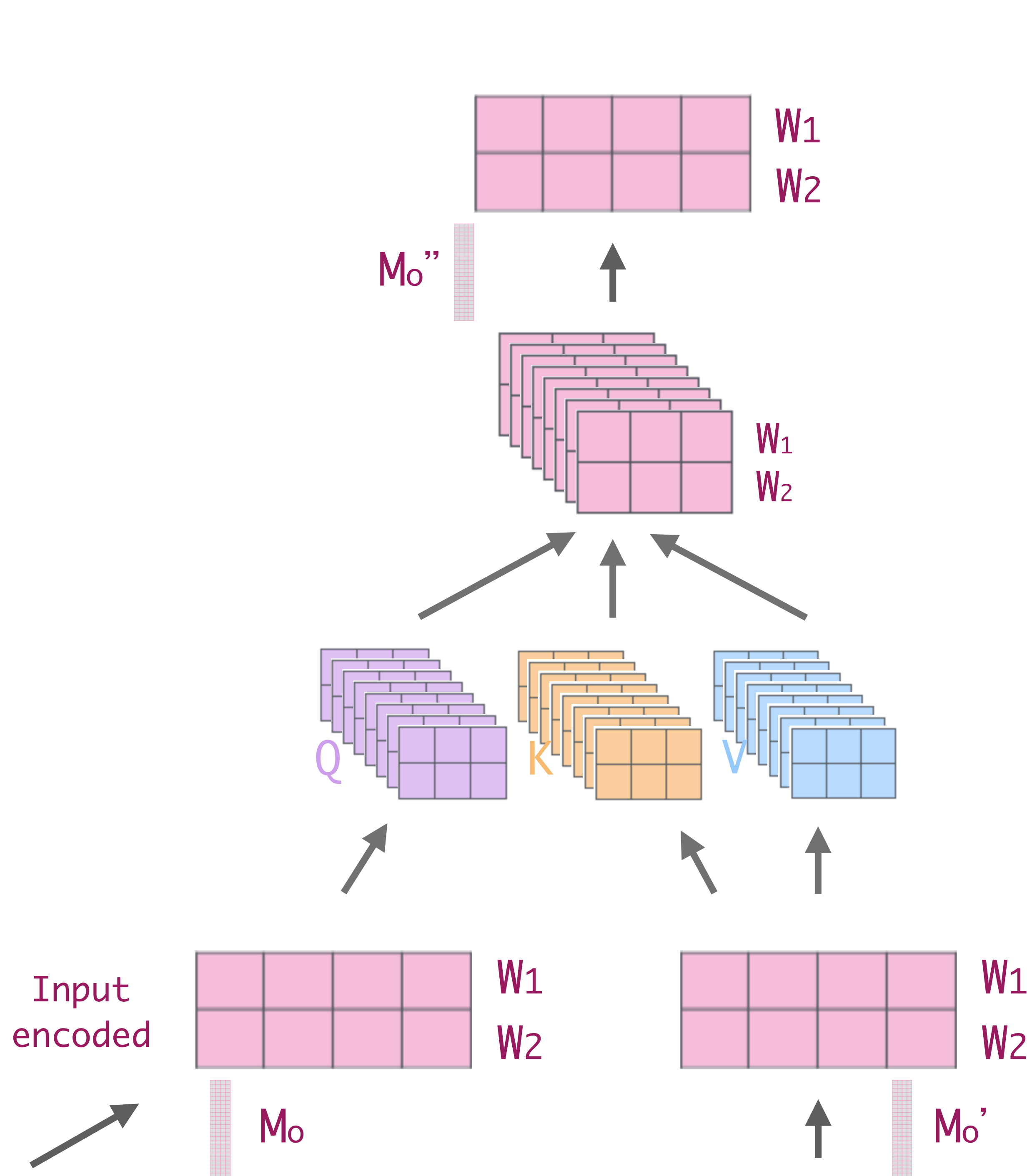


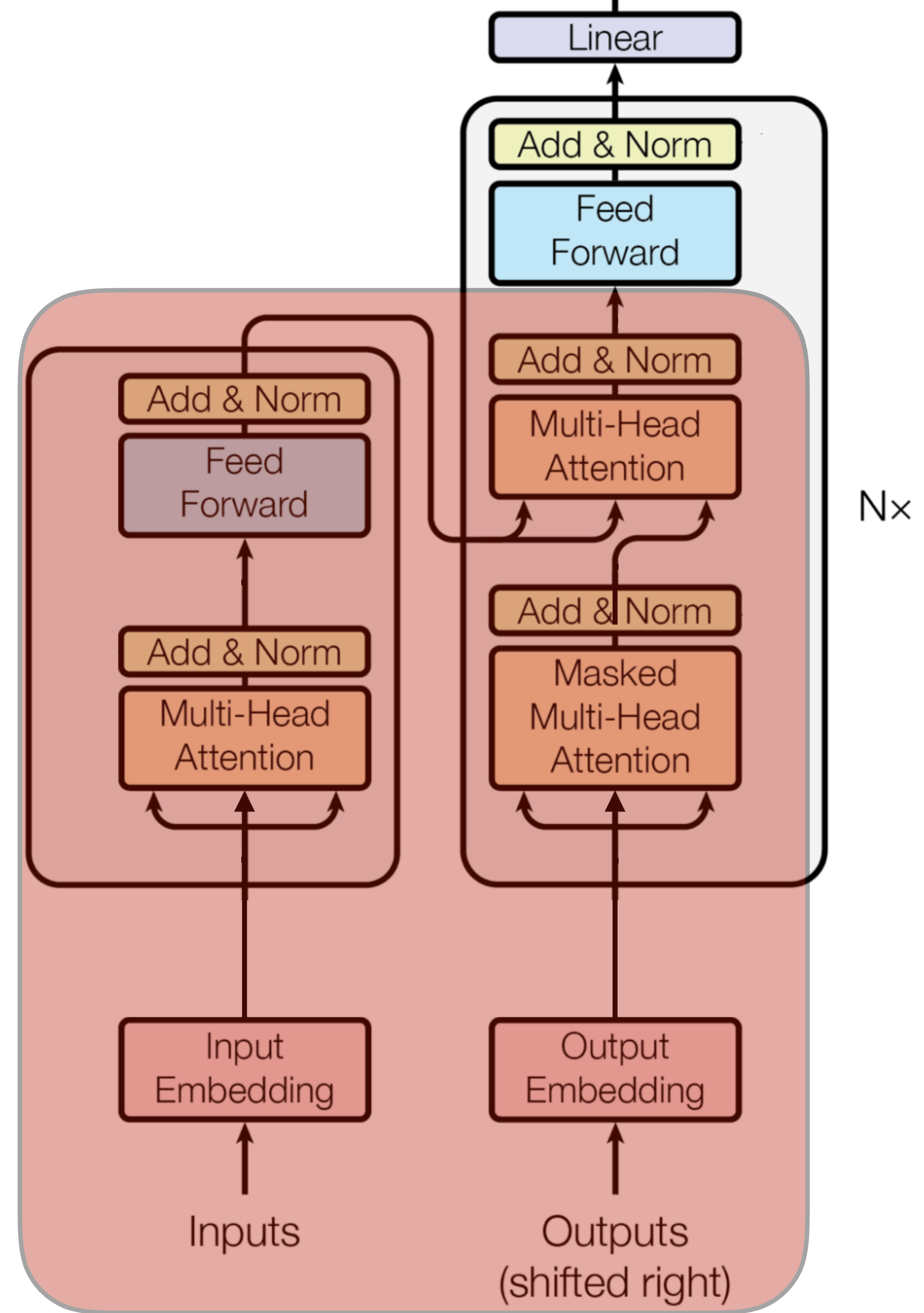
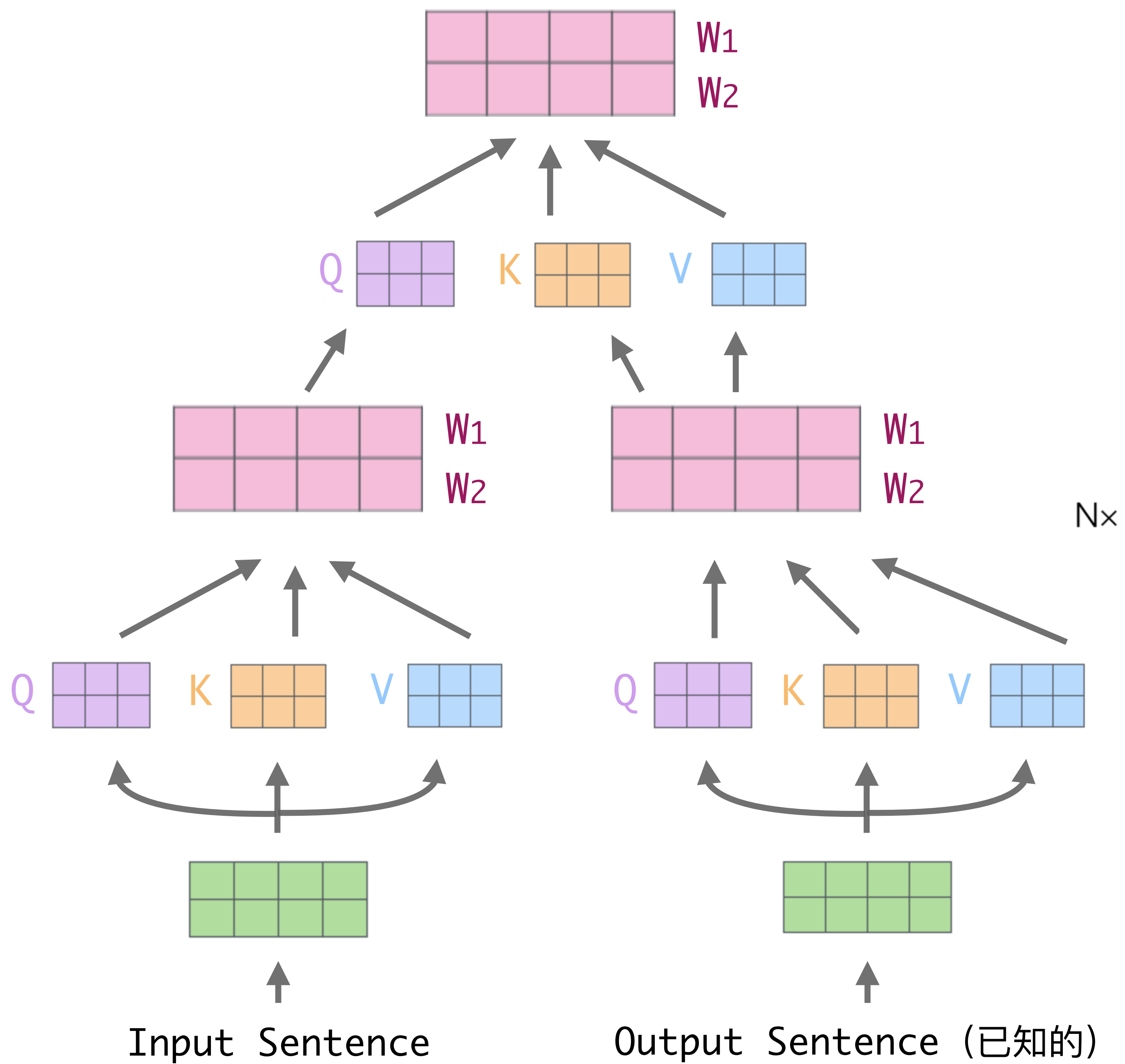






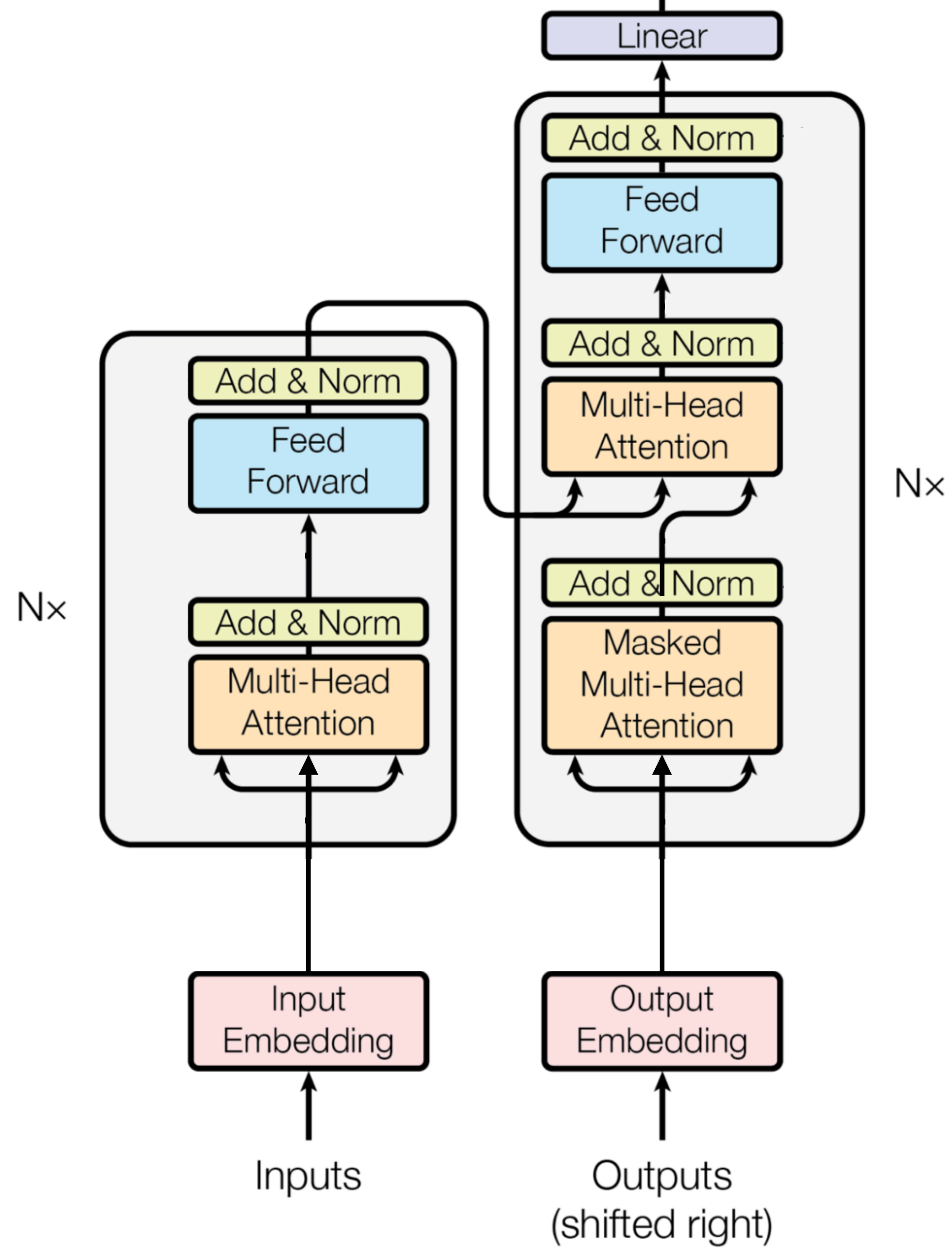






注意力

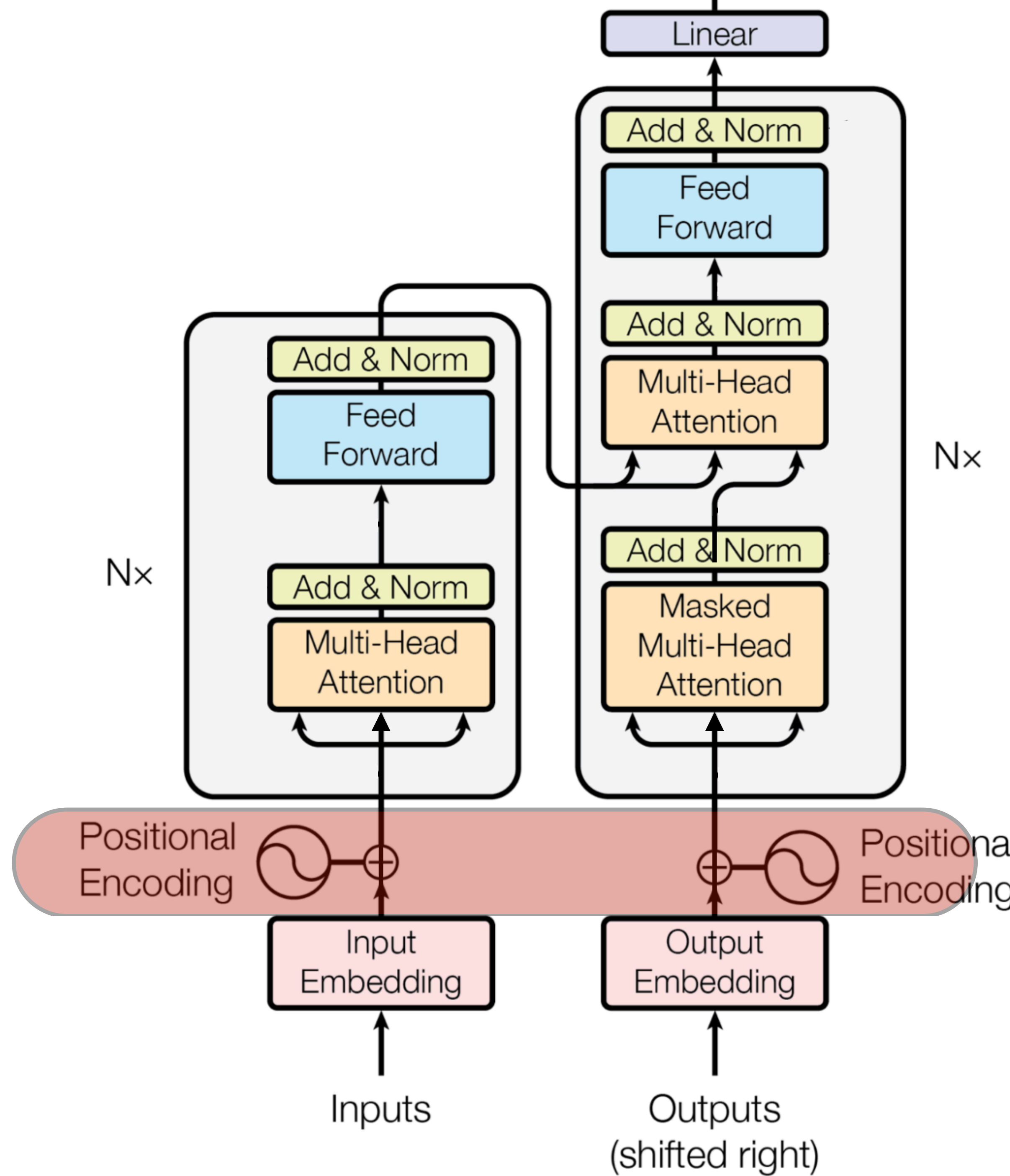
Attention



注意力

Attention

位置编码



位置编码

Positional encoding

位置编码

Positional encoding

为什么需要单独对token的位置进行编码？

位置编码长什么样？

为什么要长这样？

位置编码

Positional encoding

为什么需要单独对token的位置进行编码？

由于模型不包含递归和卷积结构，因而不能有效利用序列的顺序特征。我们需要加入序列中各个Token间相对位置或Token在序列中绝对位置的信息。

位置编码

Positional encoding

为什么需要单独对token的位置进行编码？

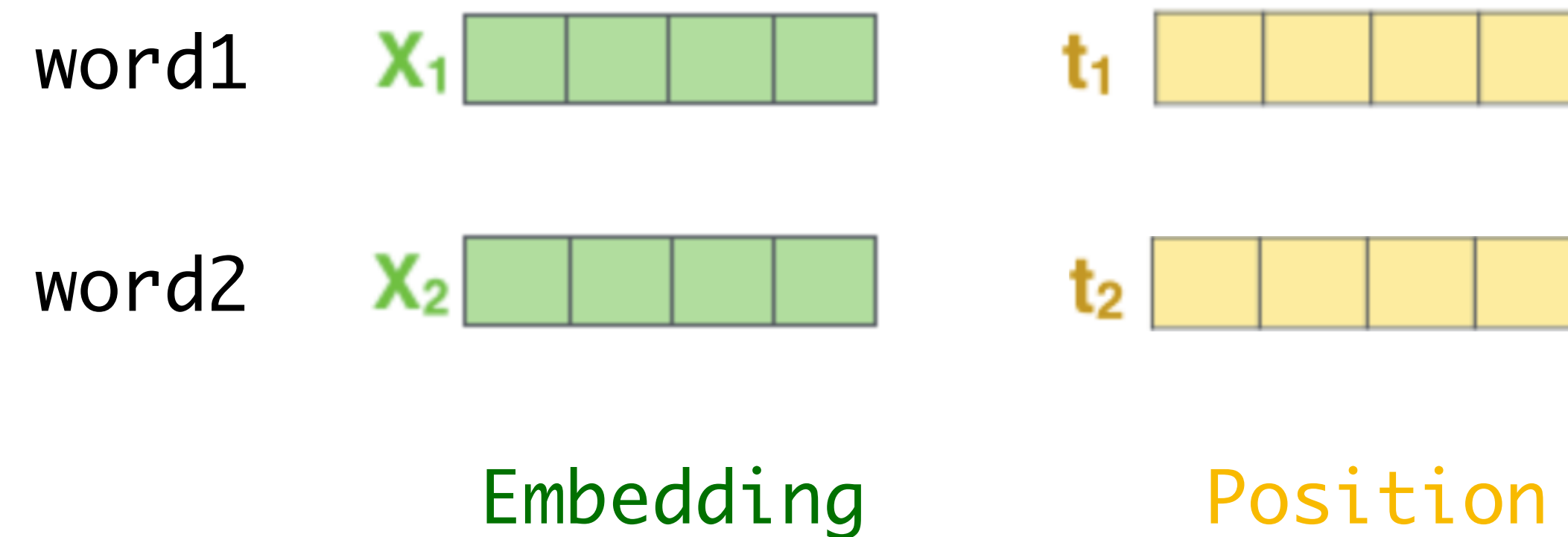
位置编码长什么样？

为什么要长这样？

位置编码

Positional encoding

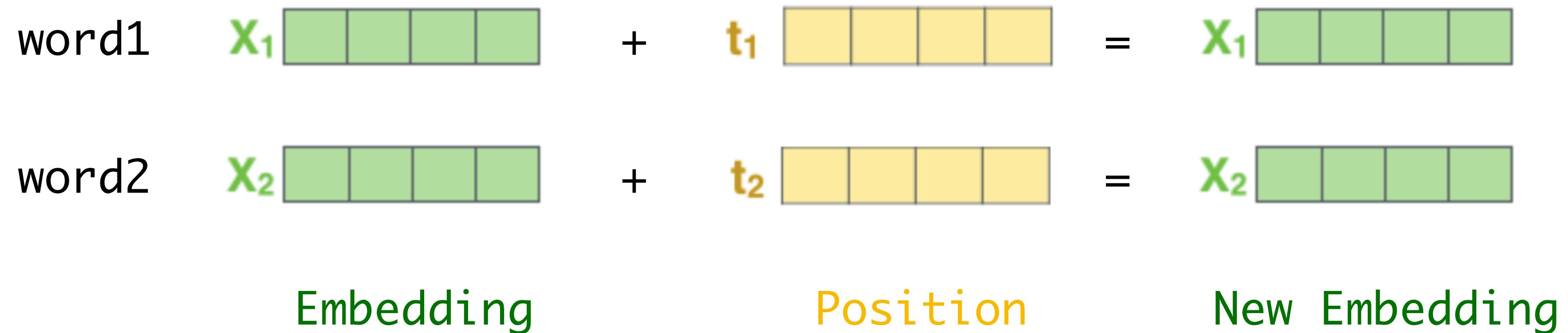
位置编码长什么样？



位置编码

Positional encoding

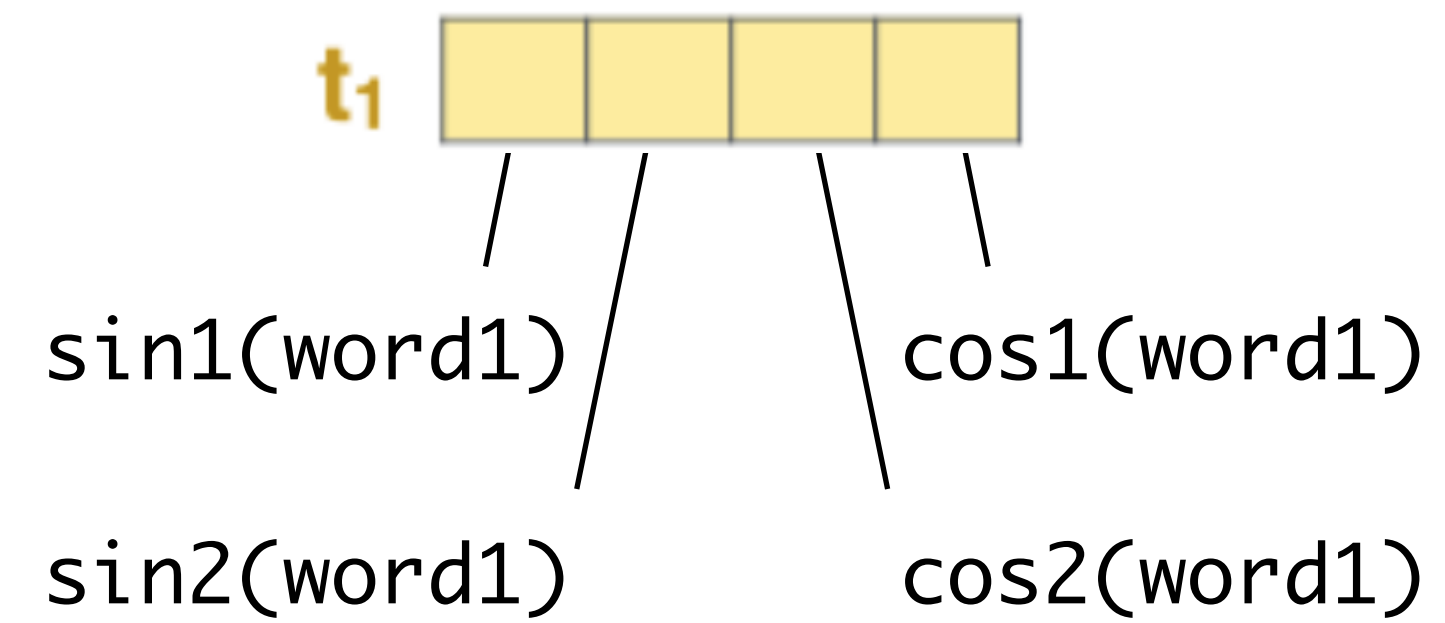
位置编码长什么样？



位置编码

Positional encoding

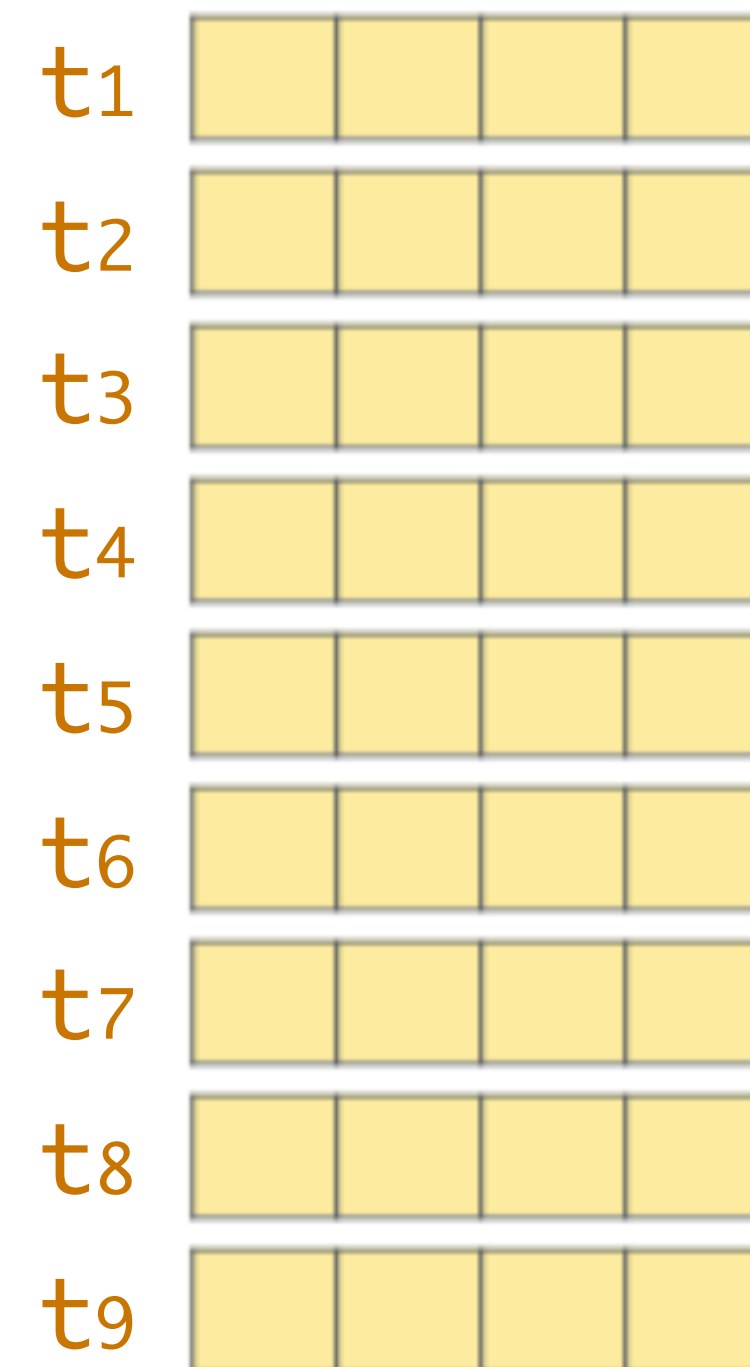
位置编码长什么样？



位置编码

Positional encoding

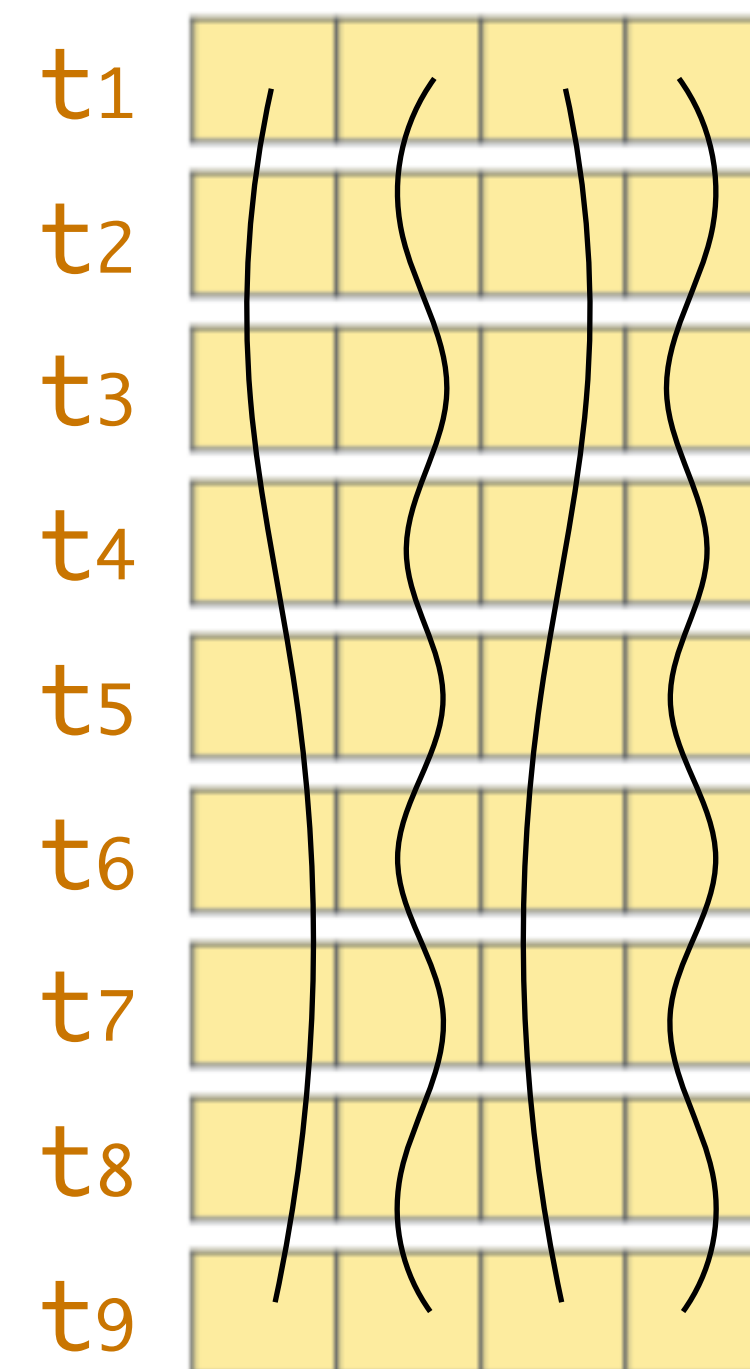
位置编码长什么样？



位置编码

Positional encoding

位置编码长什么样？



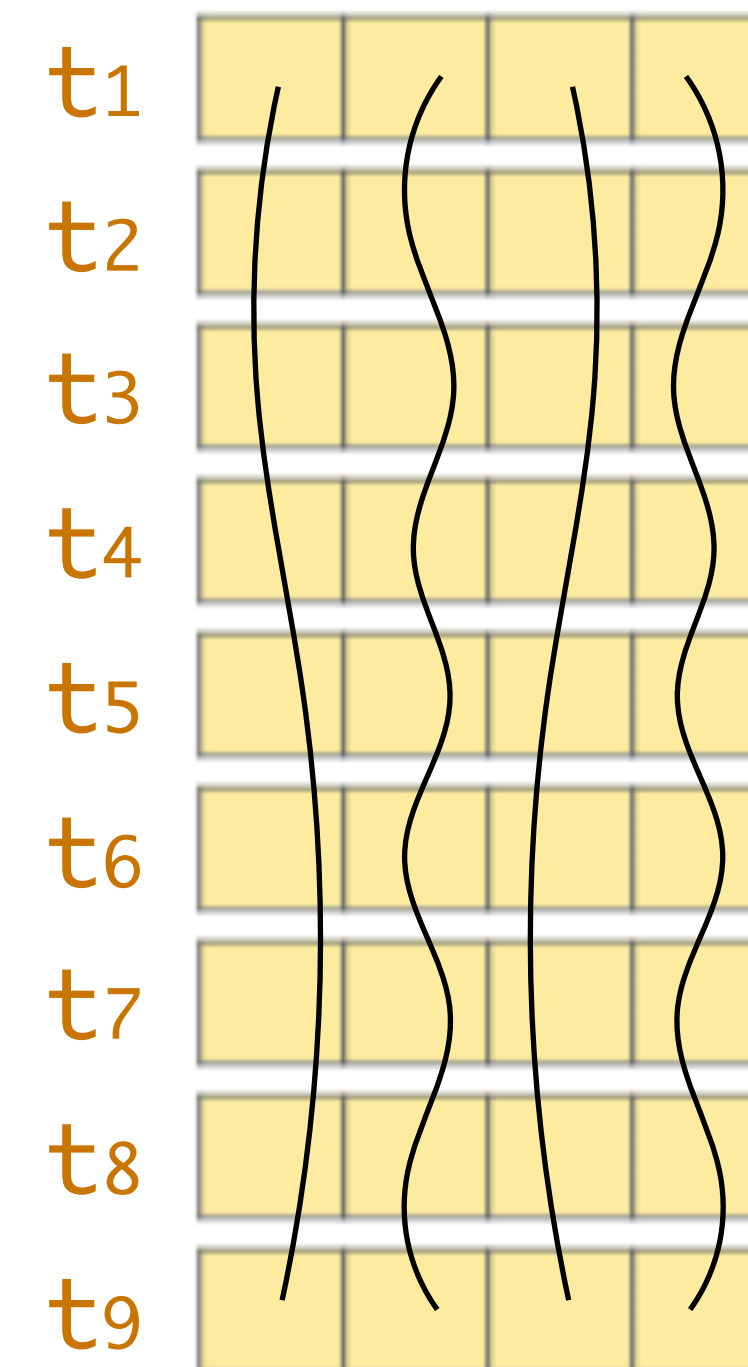
位置编码

Positional encoding

位置编码长什么样？

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

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



位置编码

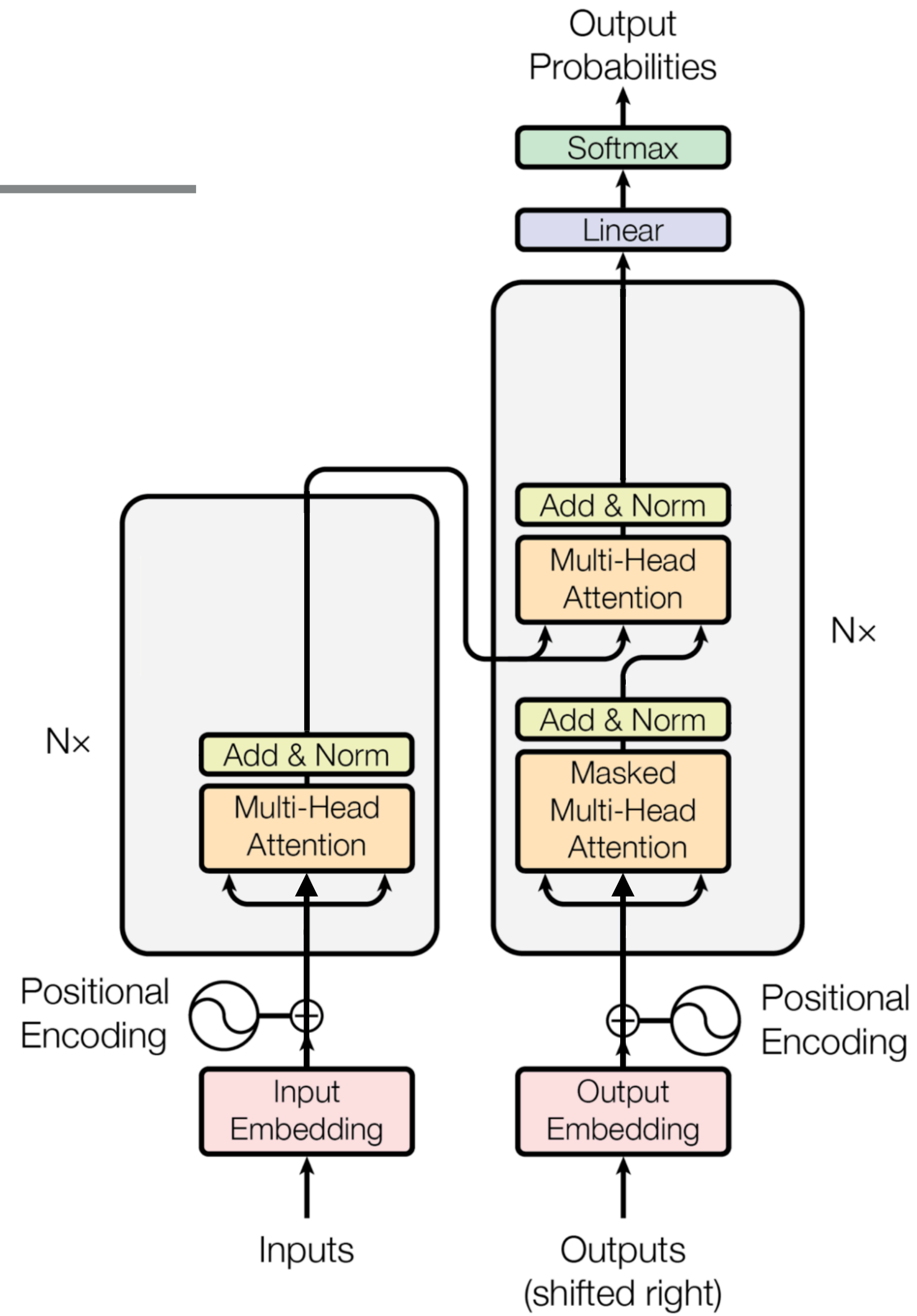
Positional encoding

为什么需要单独对token的位置进行编码？

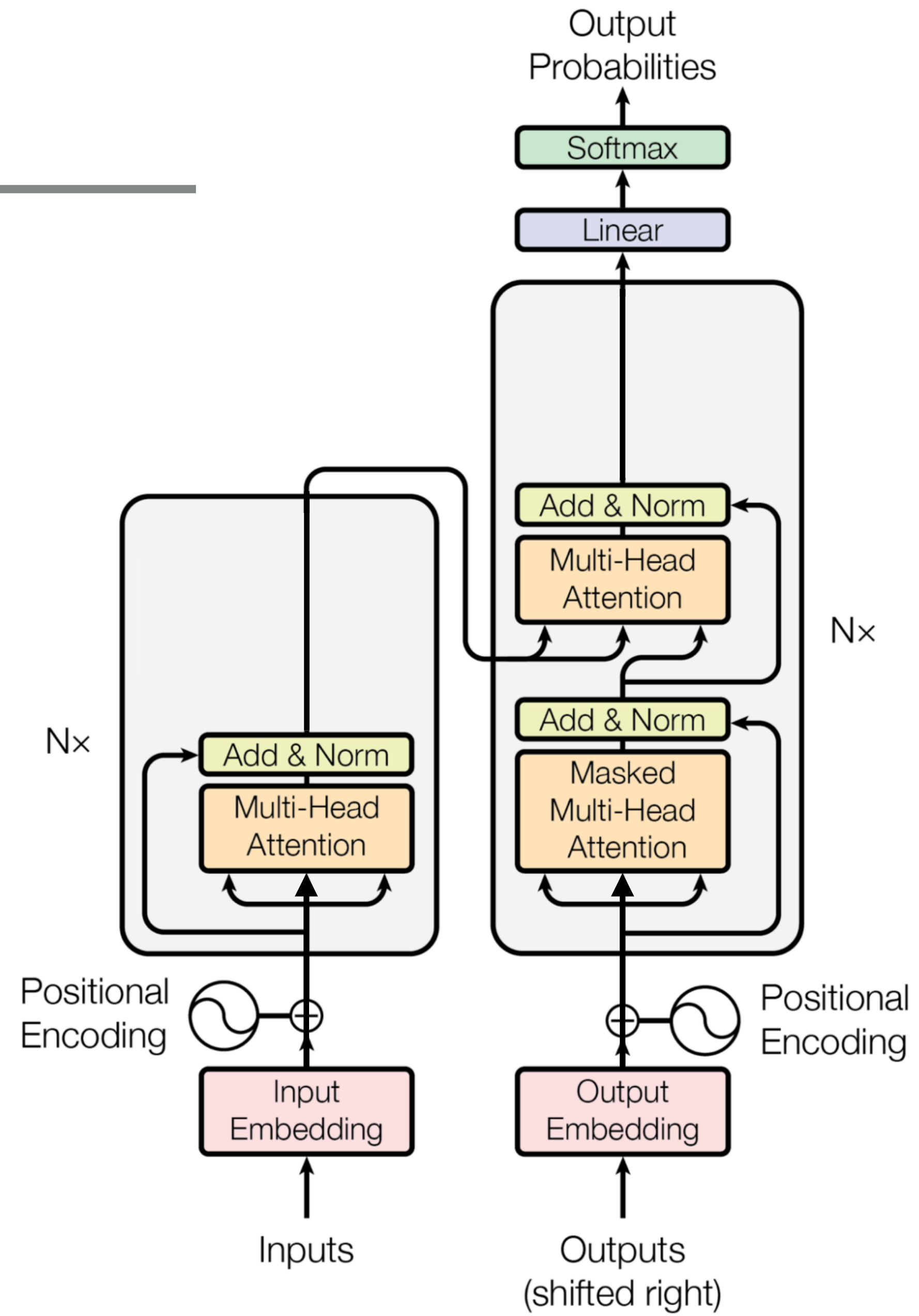
位置编码长什么样？

为什么要长这样？

位置编码



位置编码



残差连接

Residual connection

残差连接

Residual connection

为什么要做残差连接？

怎么做残差连接？

残差连接

Residual connection

为什么要做残差连接？

主要是为了防止梯度爆炸和梯度消失
因为transformer有点深

残差连接

Residual connection

为什么要做残差连接？

怎么做残差连接？

残差连接

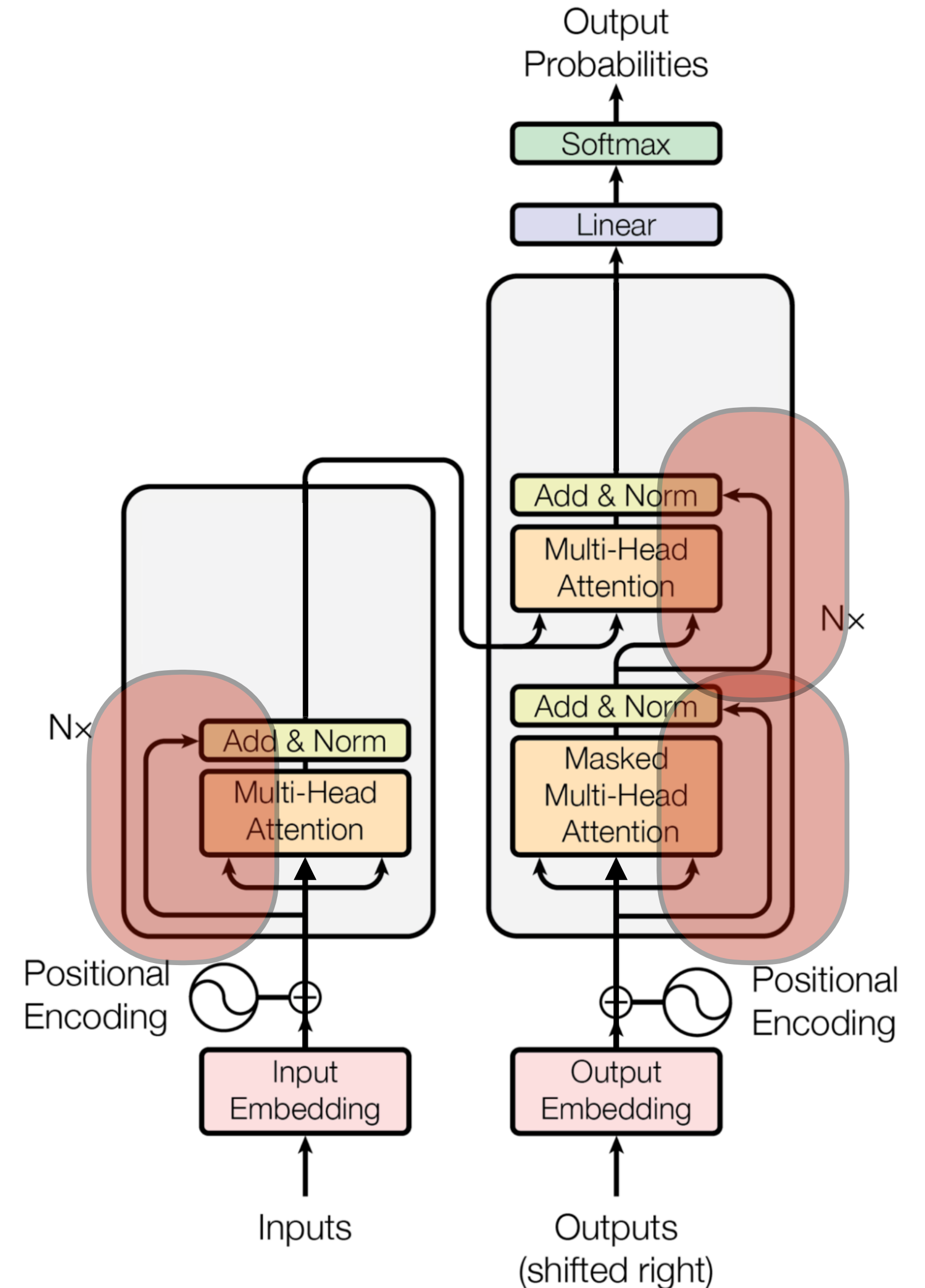
Residual connection

怎么做残差连接?

shortcut

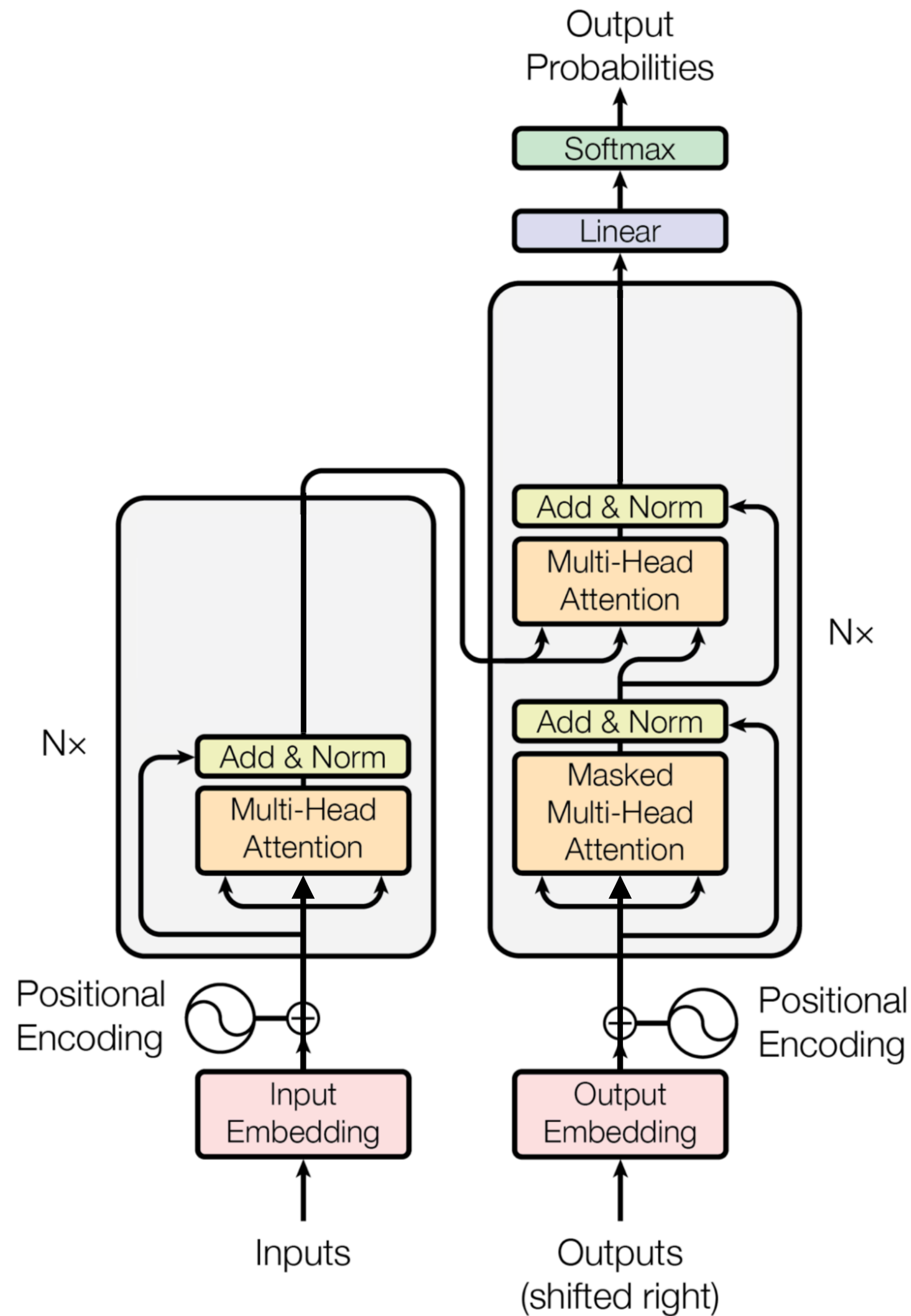
$$H(x) = F(x) + x$$

训练使 $F(x)$ 趋近于0



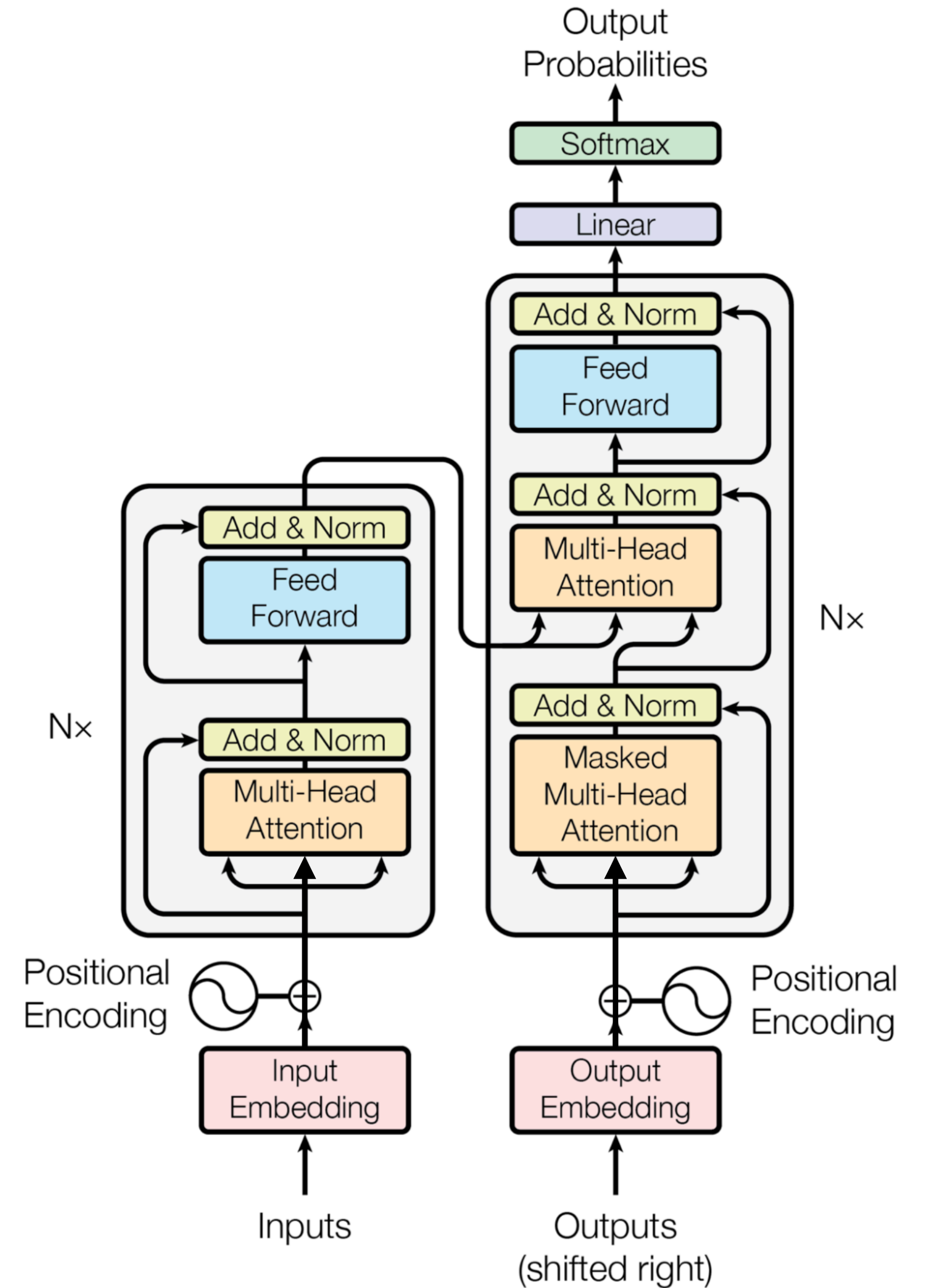
残差连接

Residual connection



残差连接

Residual connection



简单分析一下

Simple analysis

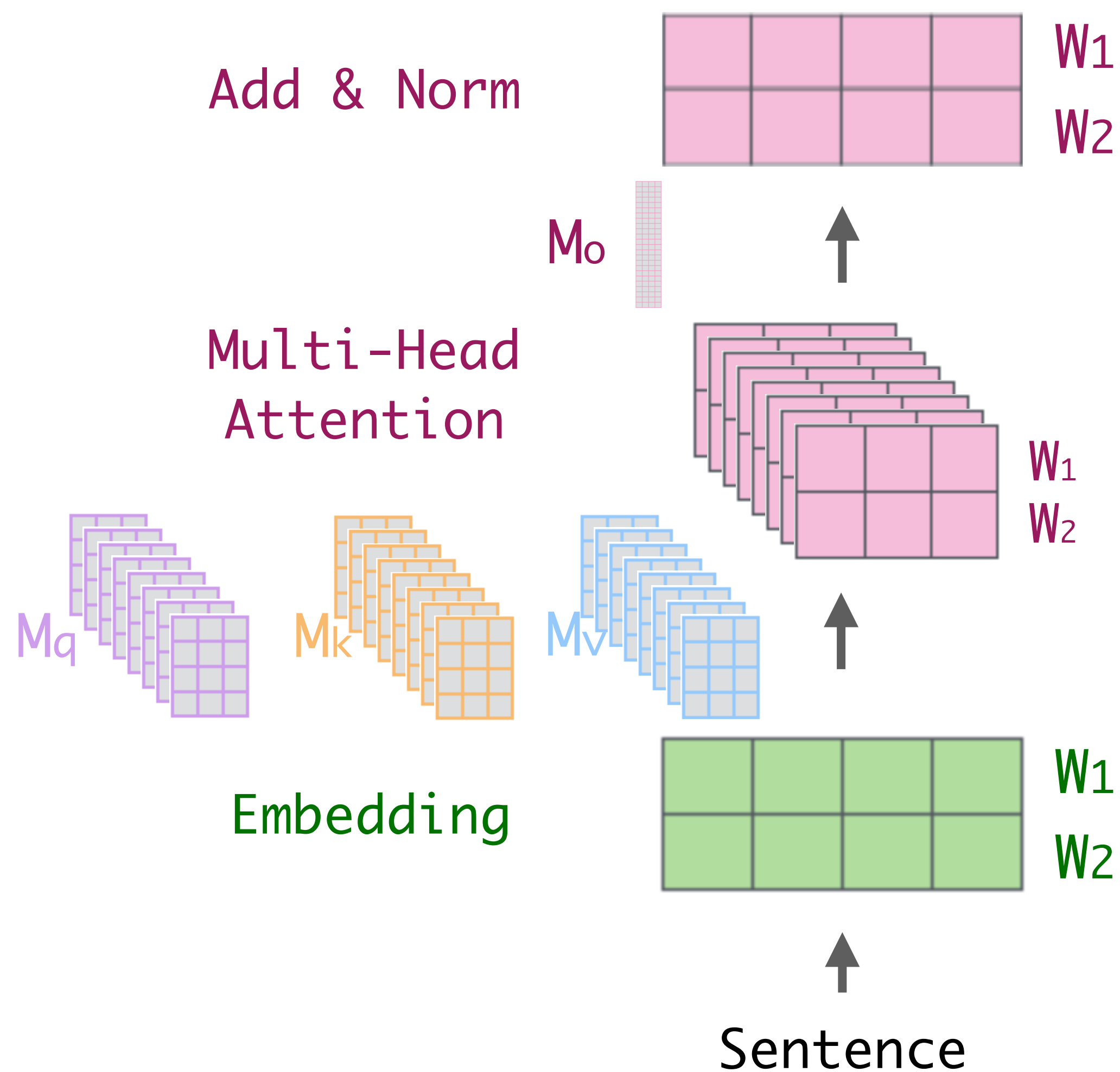
简单分析一下

Simple analysis

Attention做了什么？

简单分析一下

Simple analysis



Attention做了什么？

1. 在每个token的embedding里，加入了所有token的信息。
2. 这些token的信息不是随便加的，而是想以“注意力”为权重，加进每个token。
3. Multi-Head的每一层都只attend了一种 linguistic regularity? 还是分布式地attend的？
4. 如果不是分布式的，那么就具备可解释性了。
5. 如果是分布式的，那么或许可以通过加入启发性语言知识等方法，让它的每一层都着重attend一种 linguistic regularity，并可以以此做一个评测。
6. 那么，我们来看一下是不是分布式的：

注意力可视化

ATTENTION IS ALL YOU NEED

完全基于注意力的网络 (变形金刚)

北京大学 中文信息处理 唐乾桐

