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

Ford.

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?

評测

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

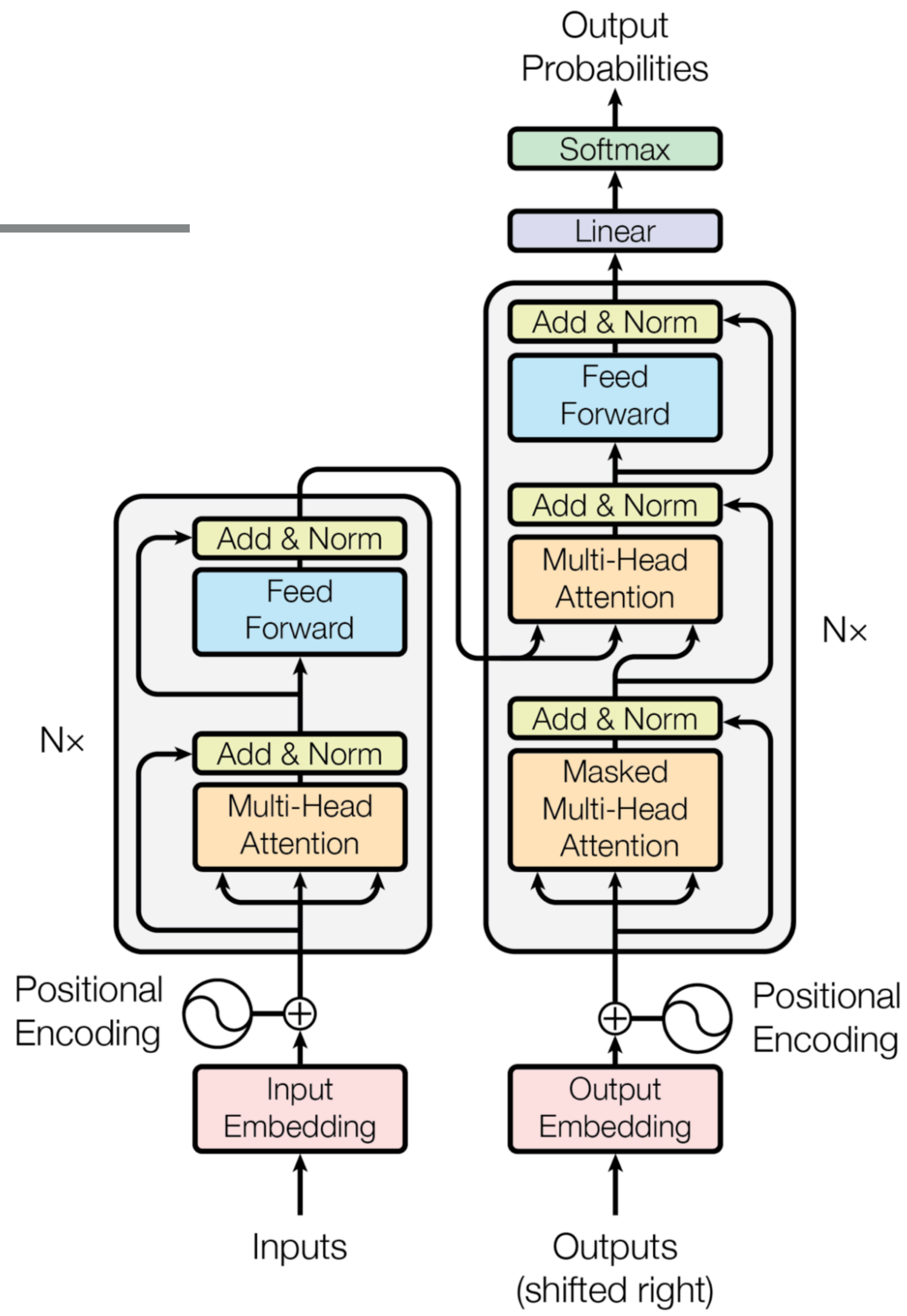
2017.06 Google

论文的模型

Overview

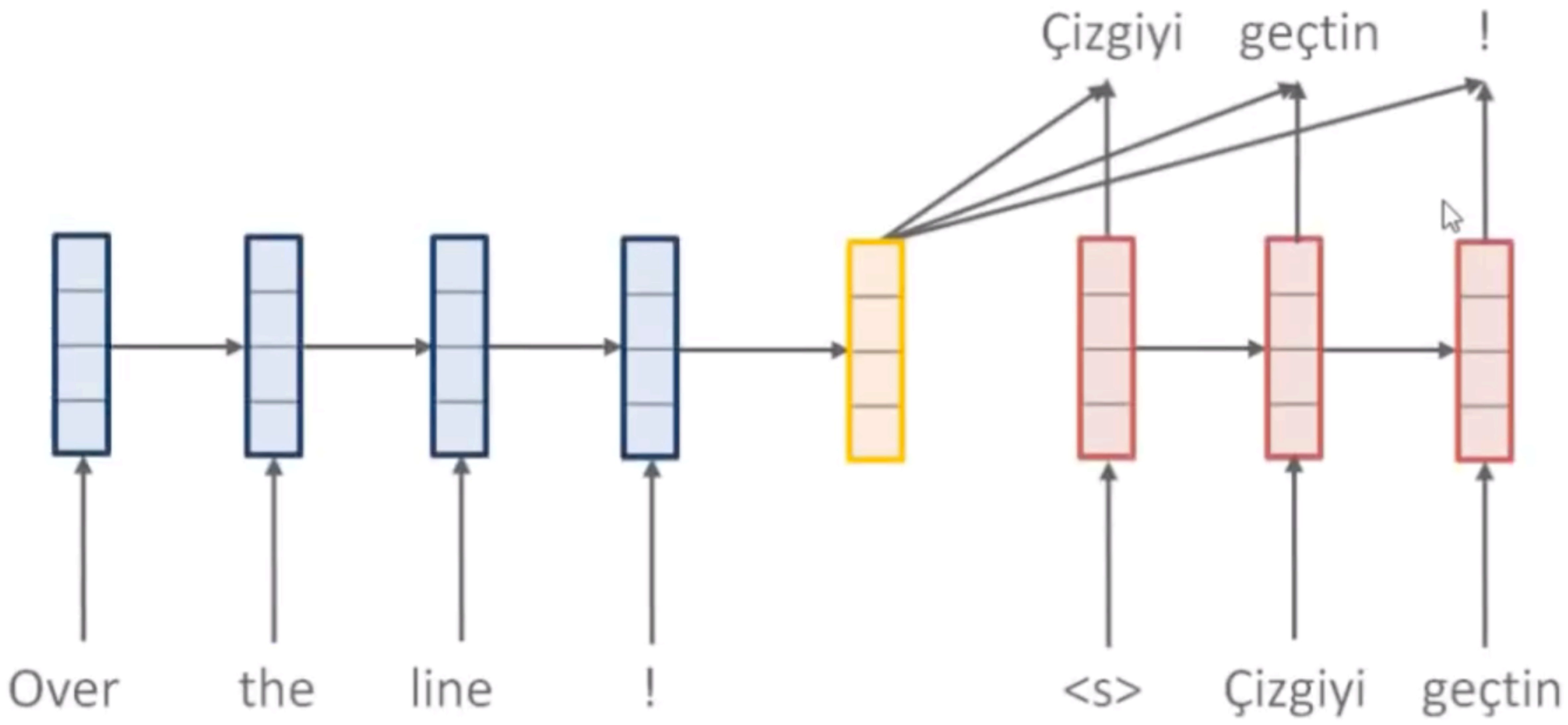
论文的模型

Overview

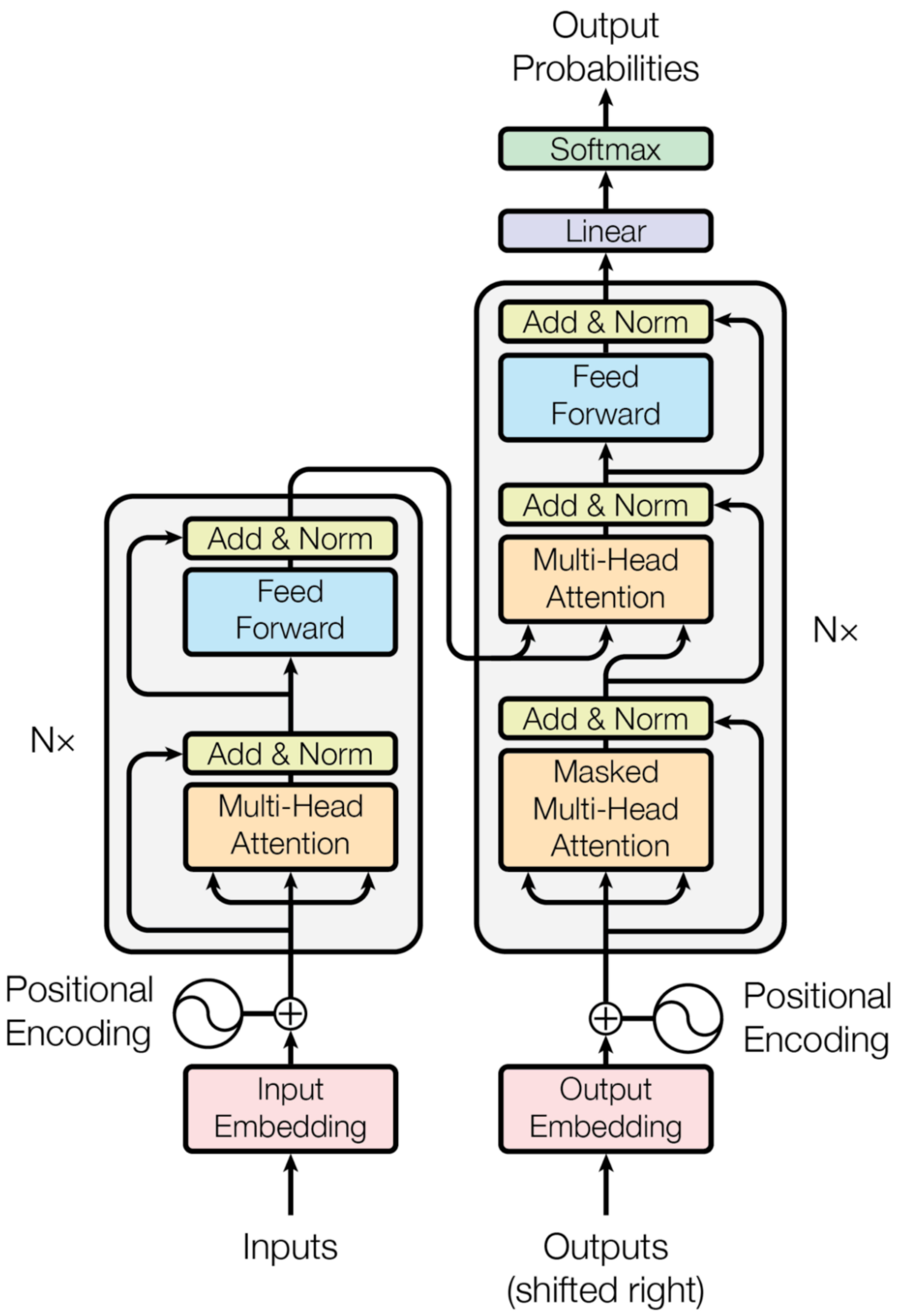


论文的模型

Overview

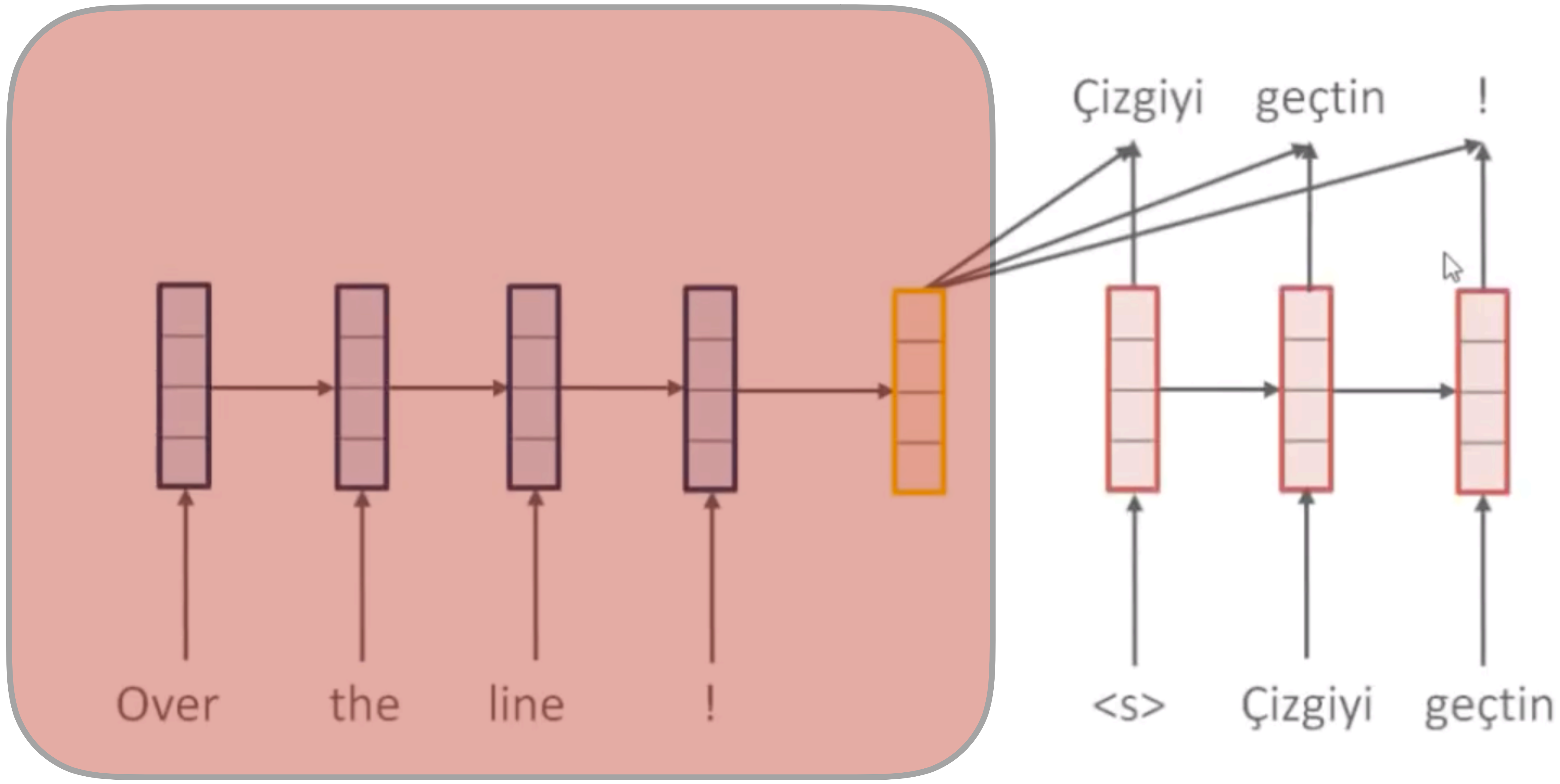


RNN encoder-decoder

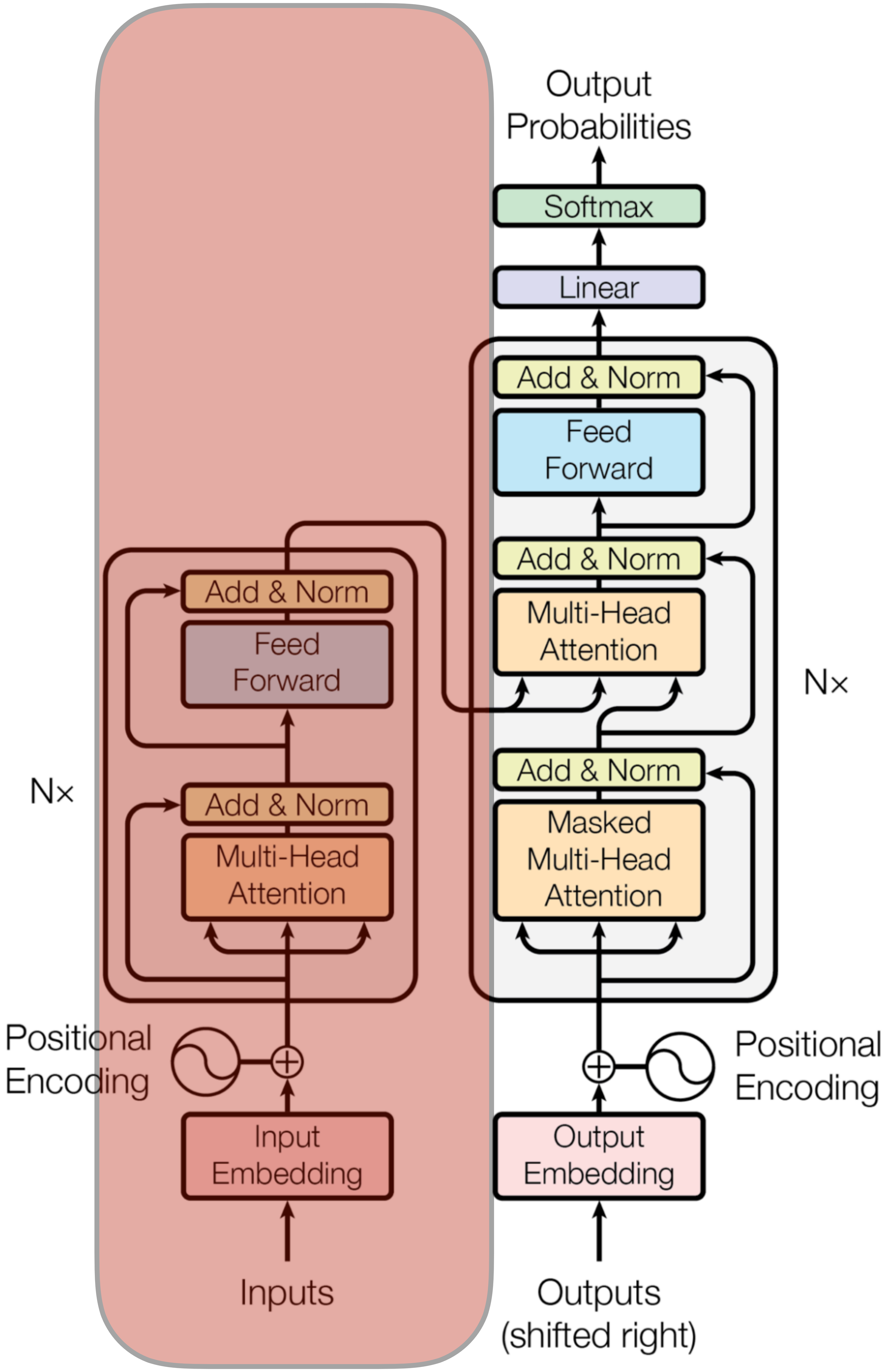


论文的模型

Overview

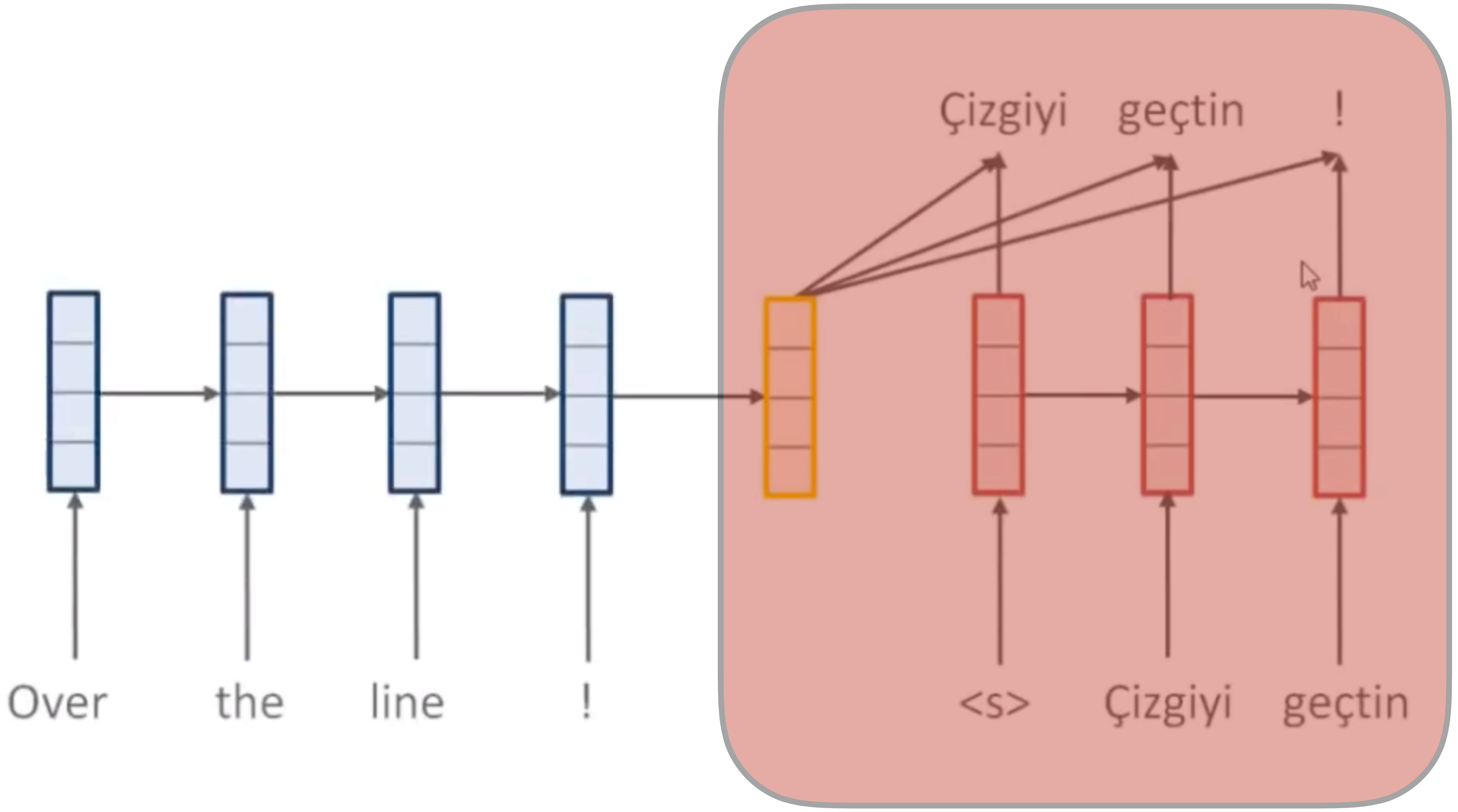


RNN encoder-decoder

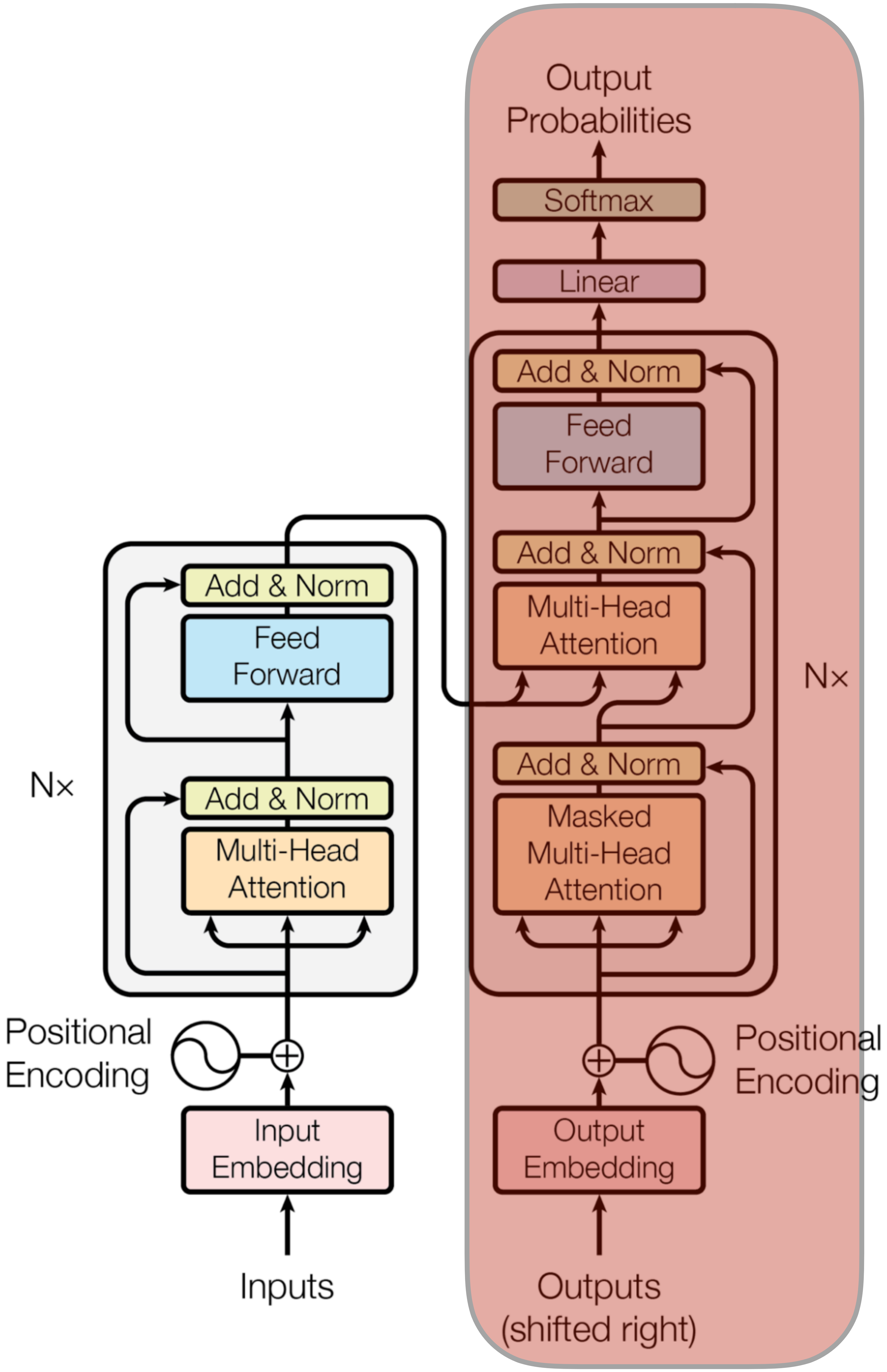


论文的模型

Overview

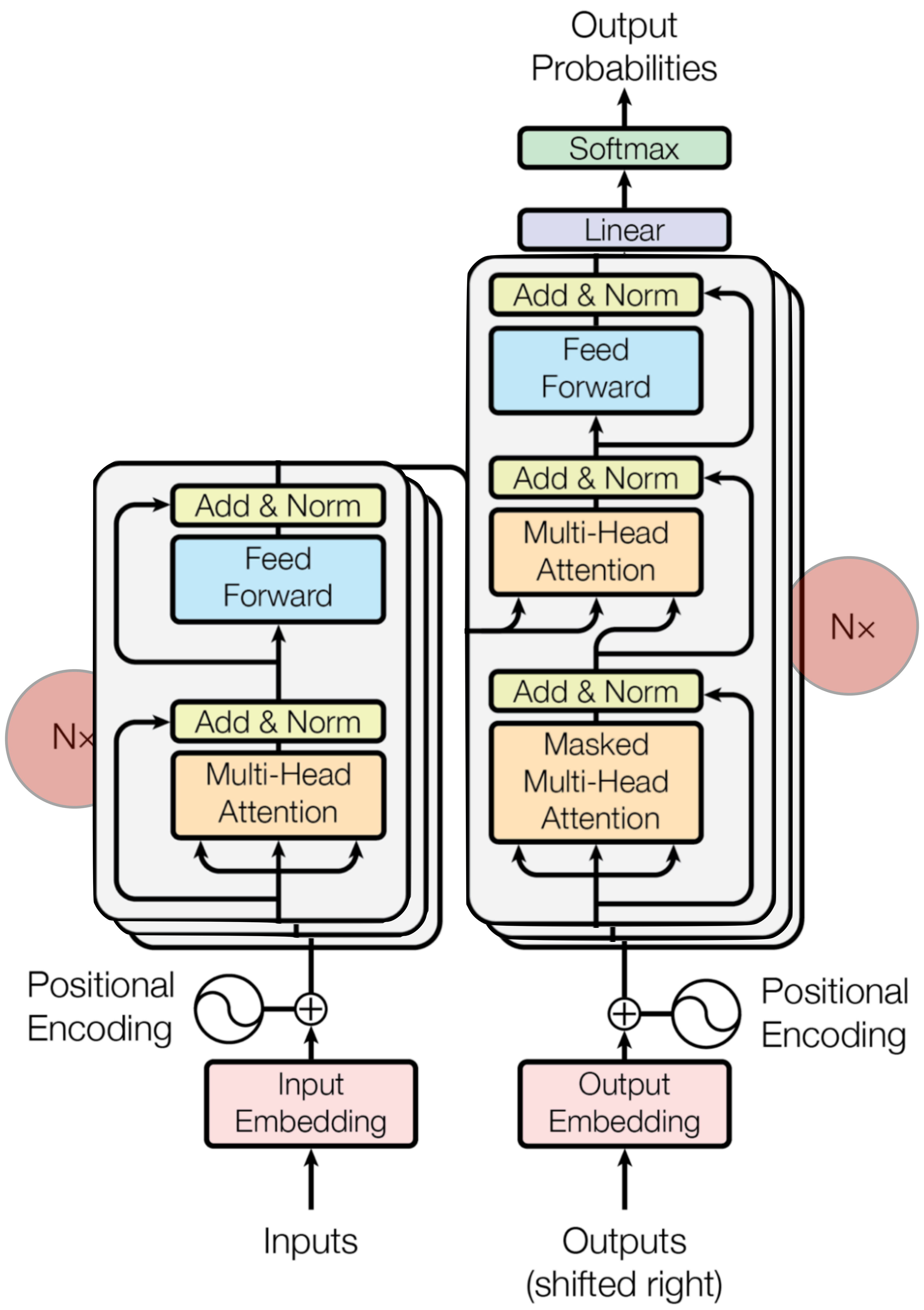


RNN encoder-decoder



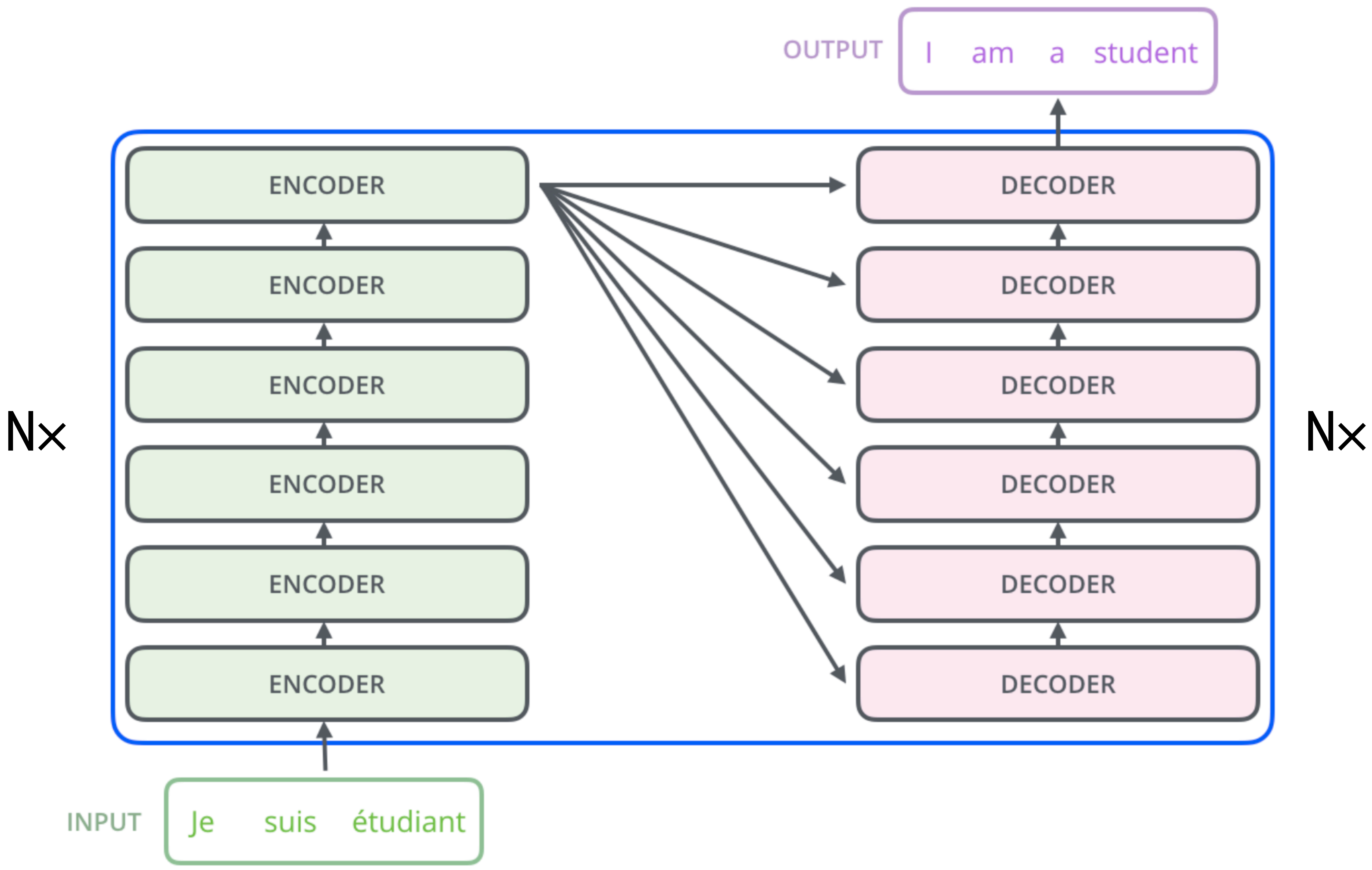
论文的模型

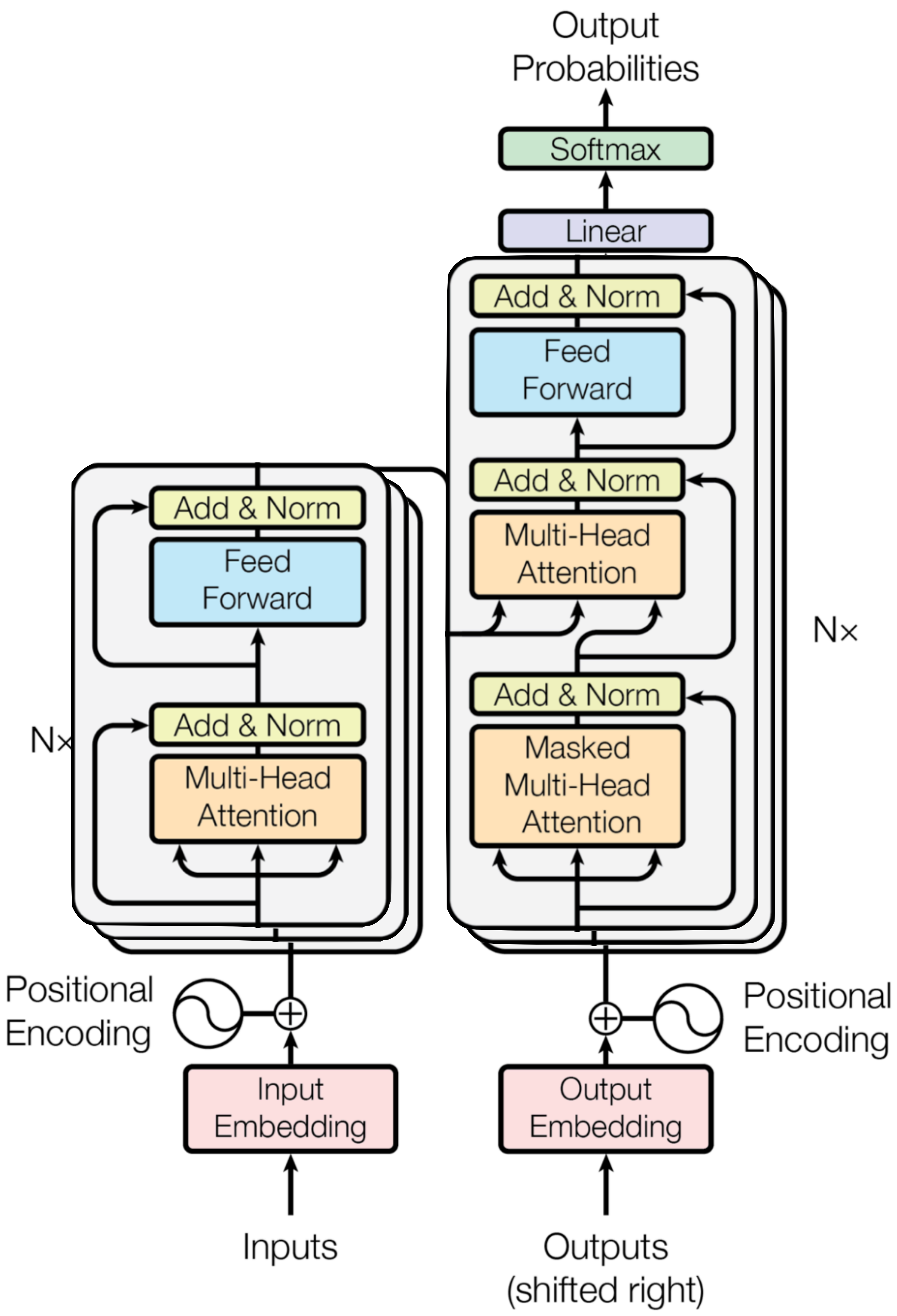
Overview



论文的模型

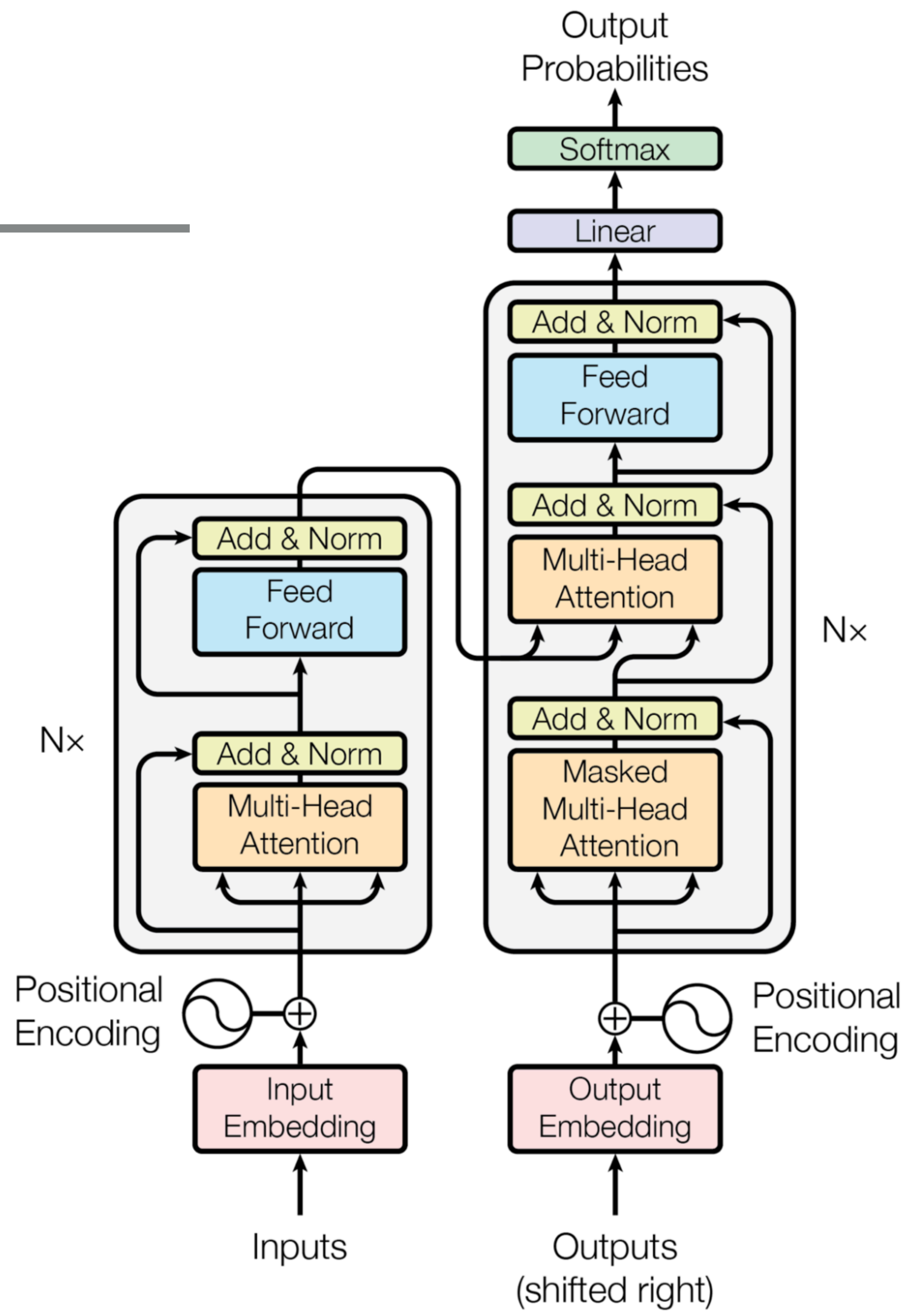
Overview





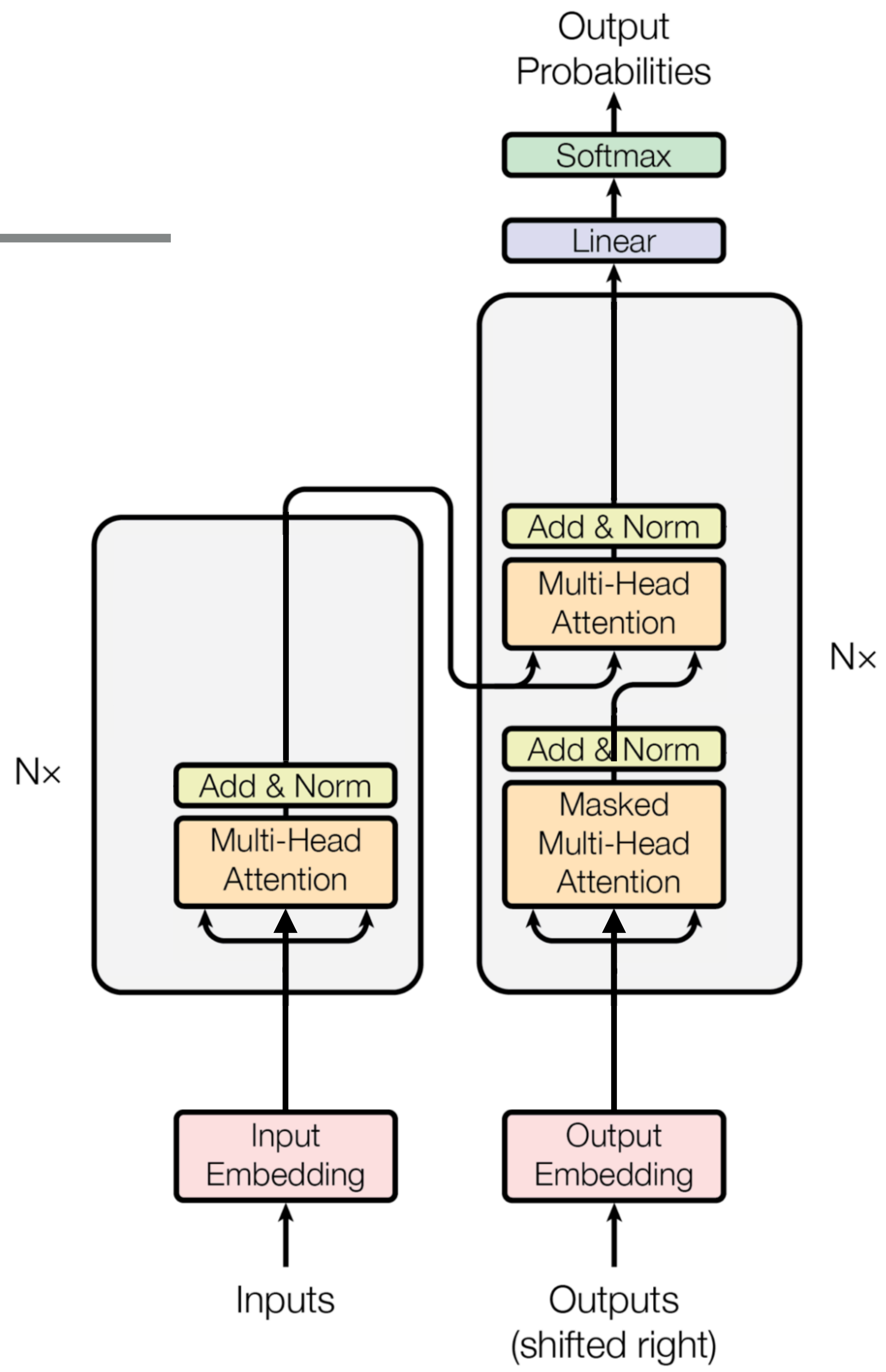
论文的模型

Overview



论文的模型

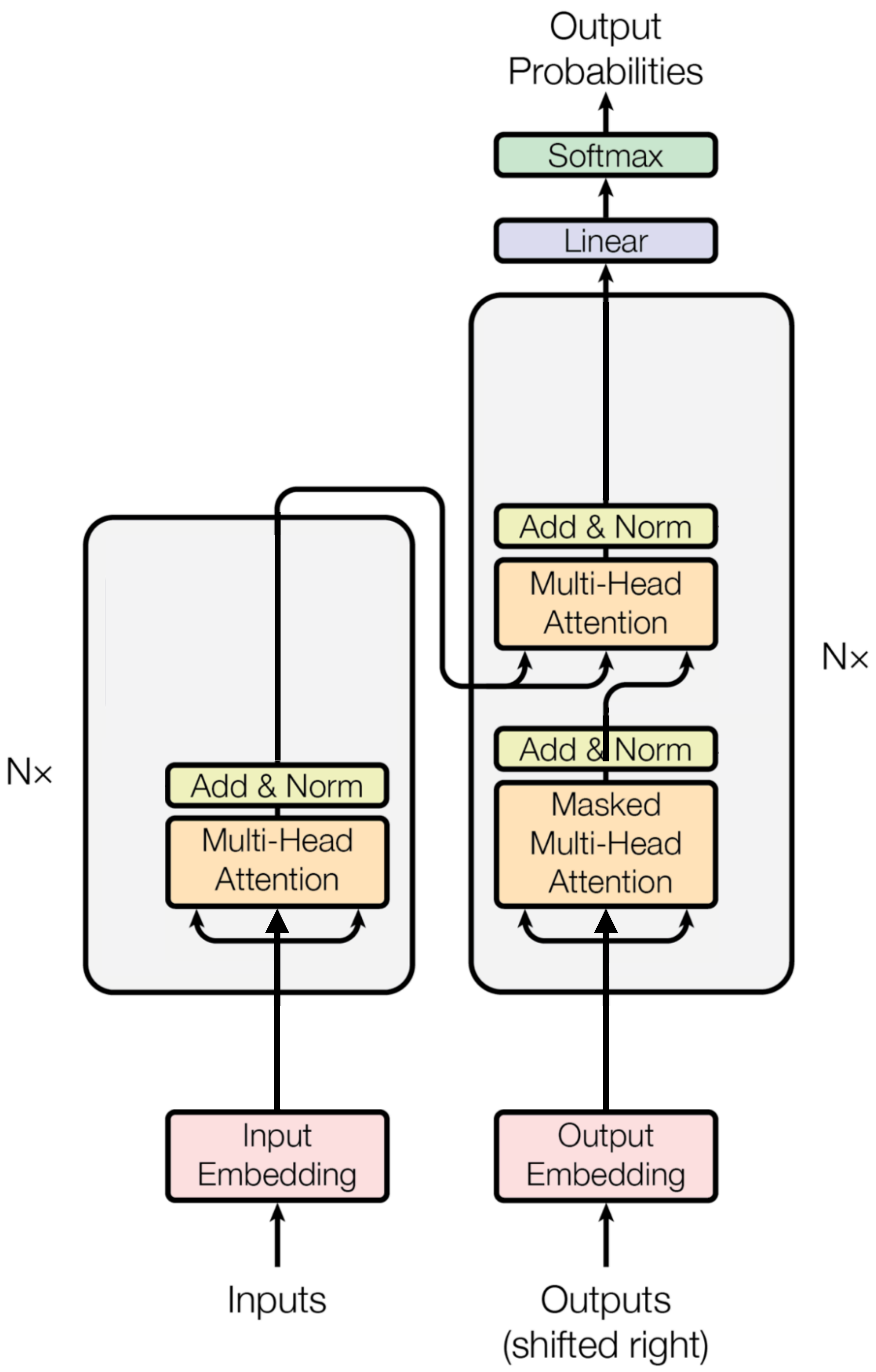
Overview



论文的模型

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。

0.9

0.2 0.3

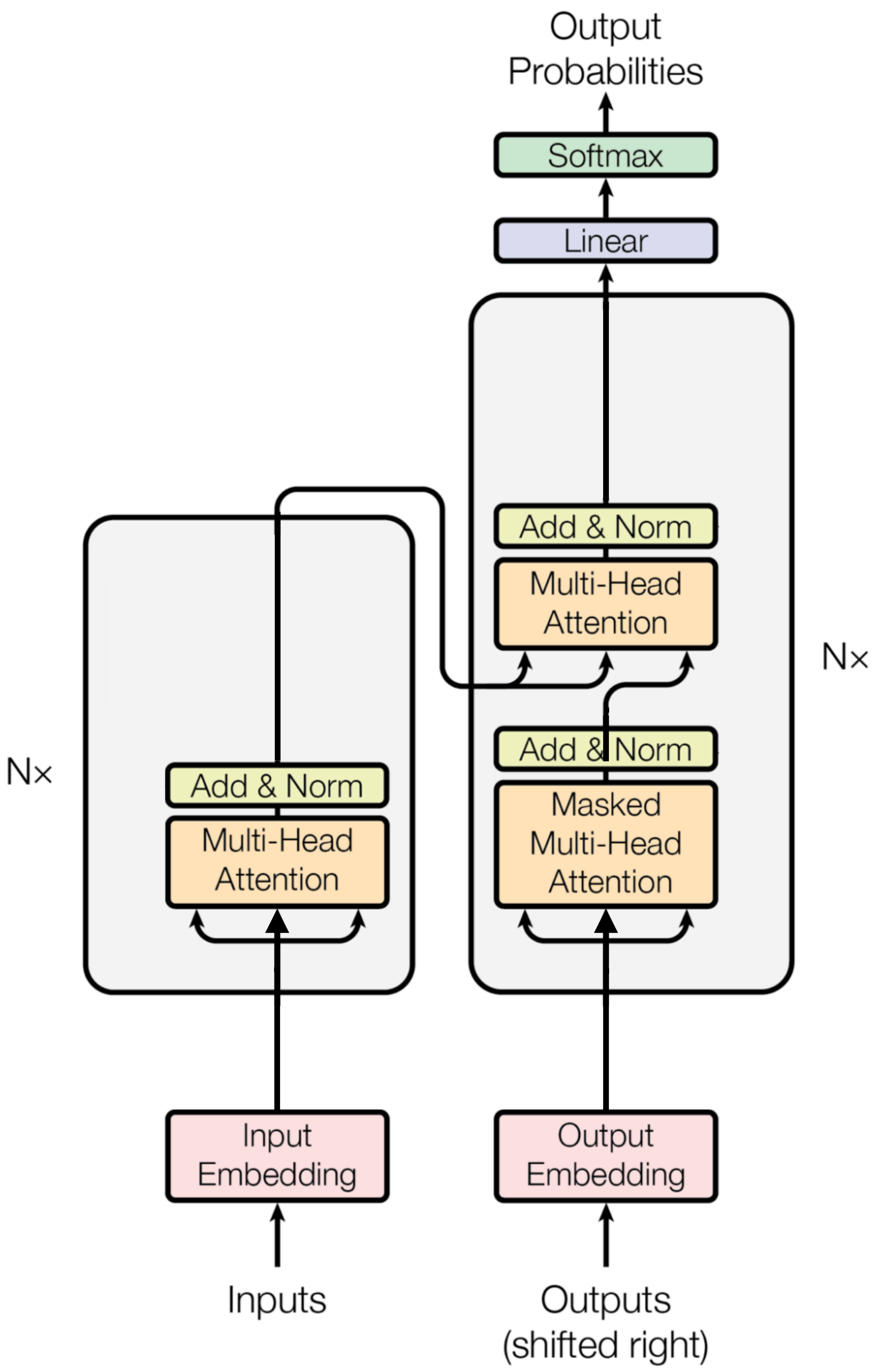
Many years later as he faced the firing squad, Colonel Aureliano Buendía was to remember that distant afternoon when his father took him to discover ice.

0.5

注意力

Attention

Attention



注意力

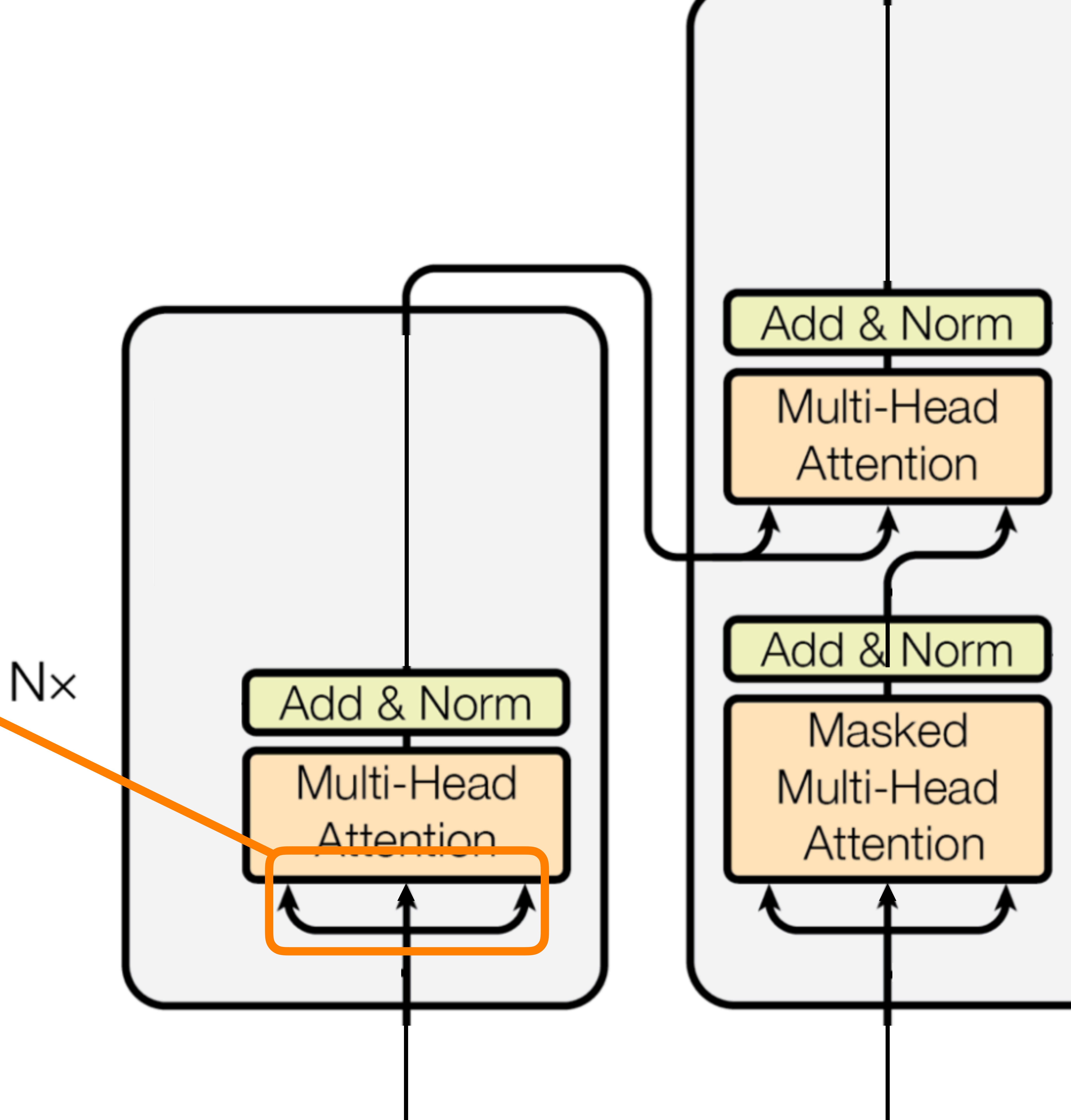
Attention

Attention

Query

Key

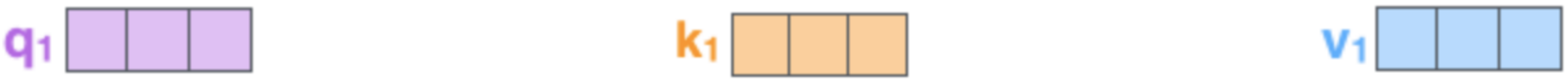
Value



注意力

Attention

Attention

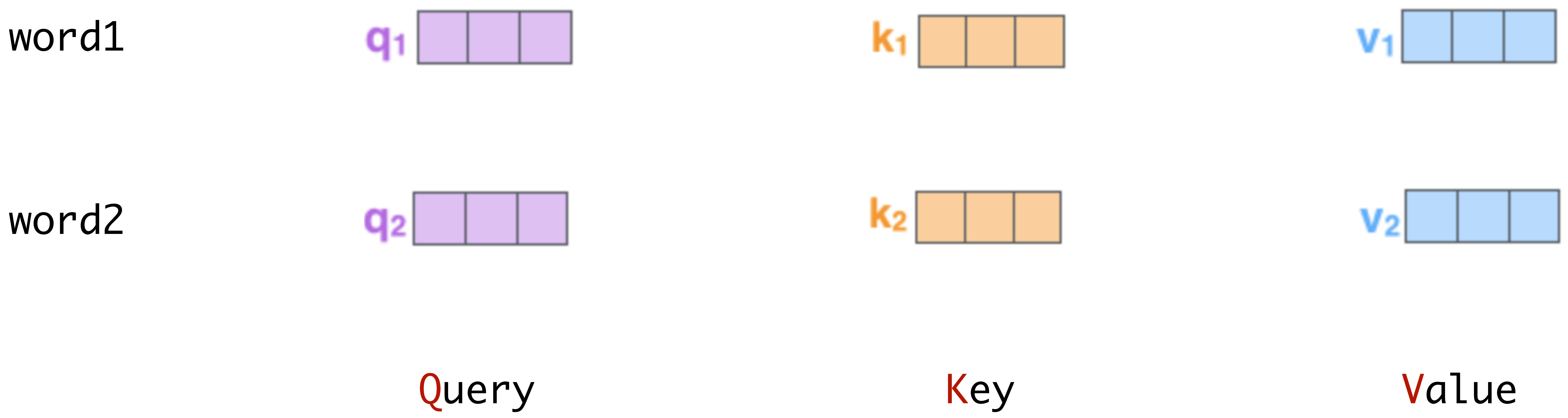


Query Key Value

注意力

Attention

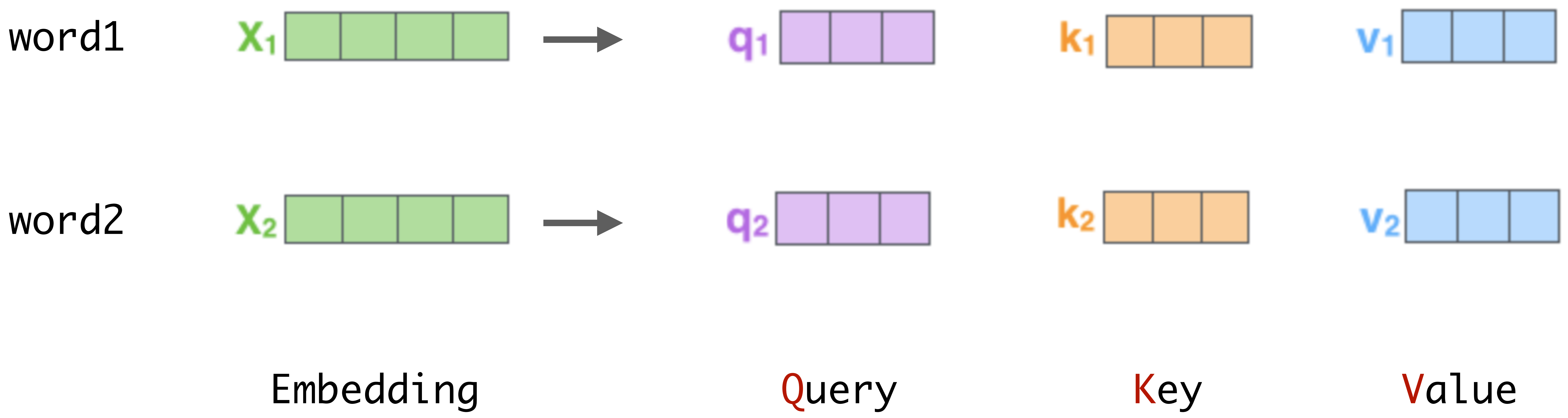
Attention

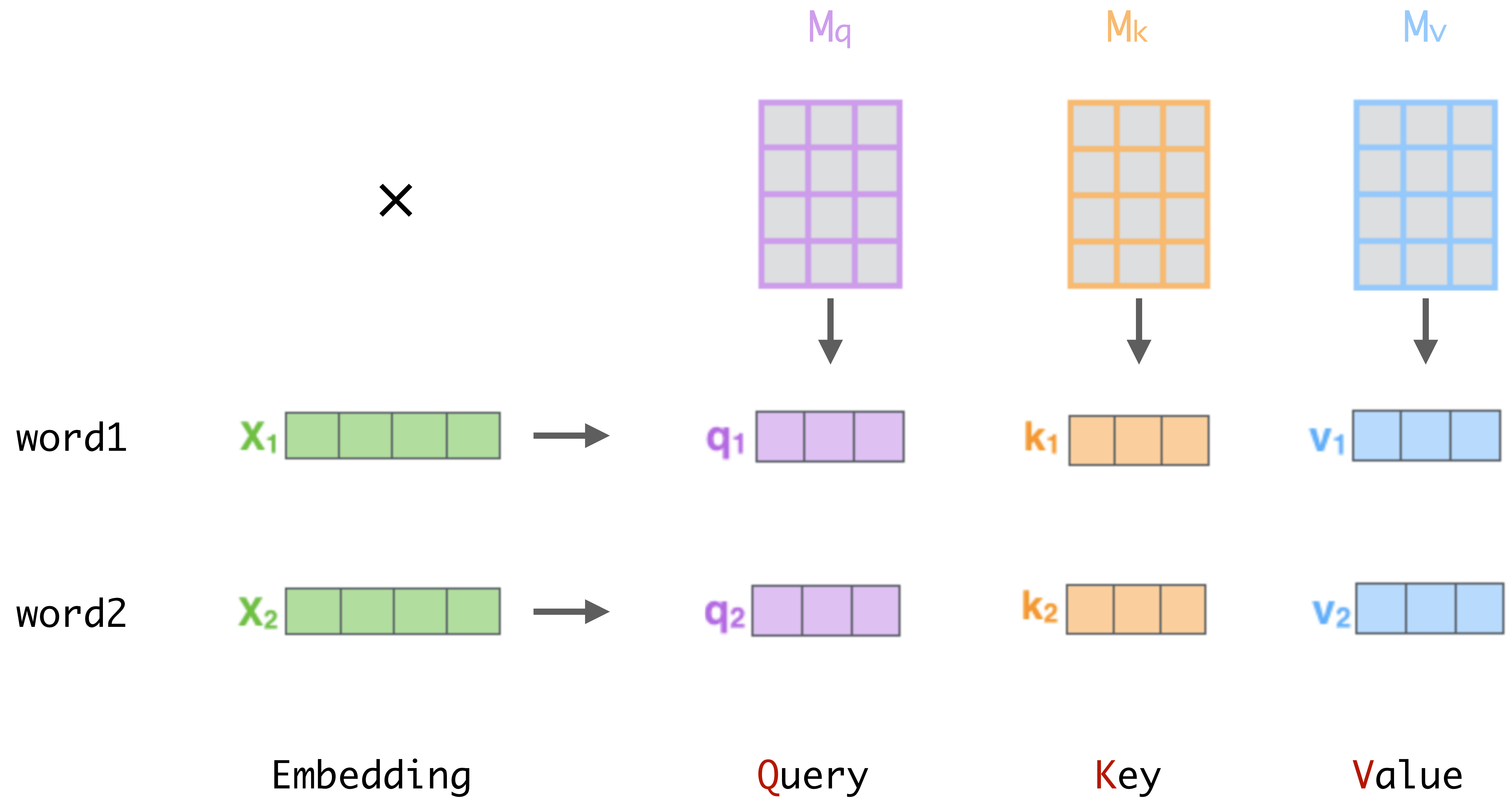


注意力

Attention

Attention

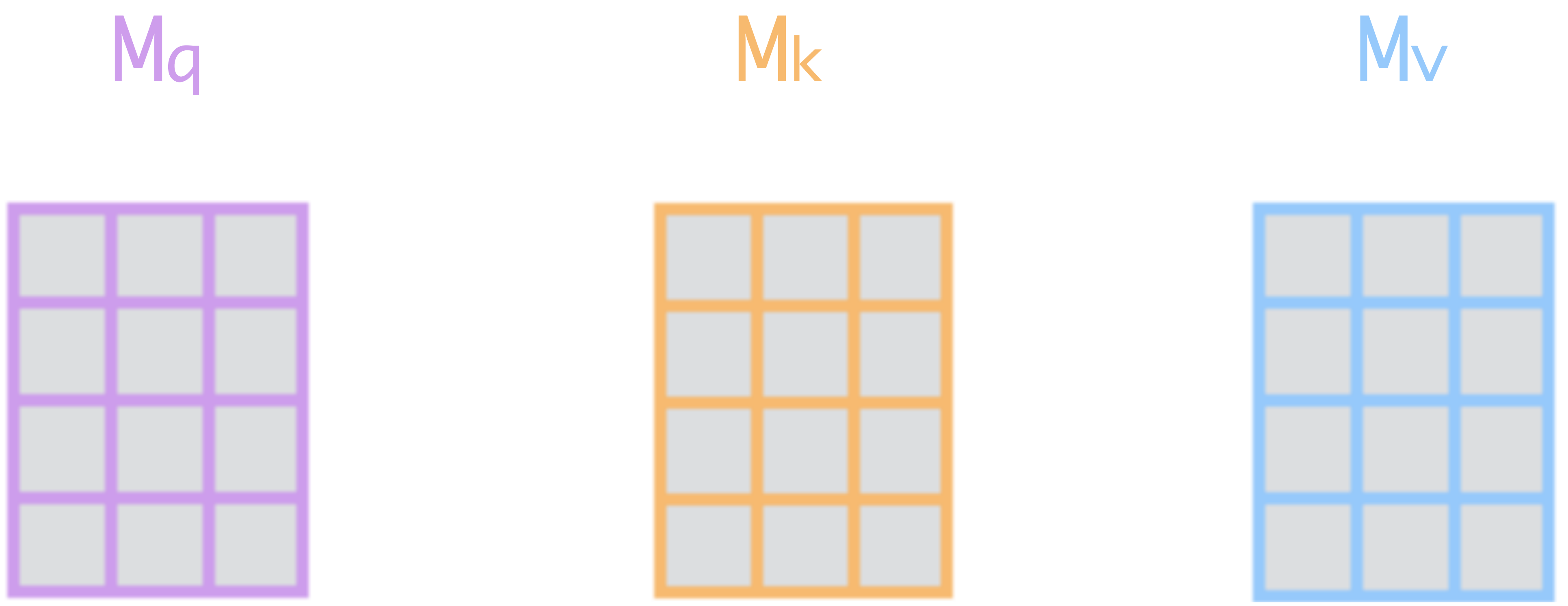


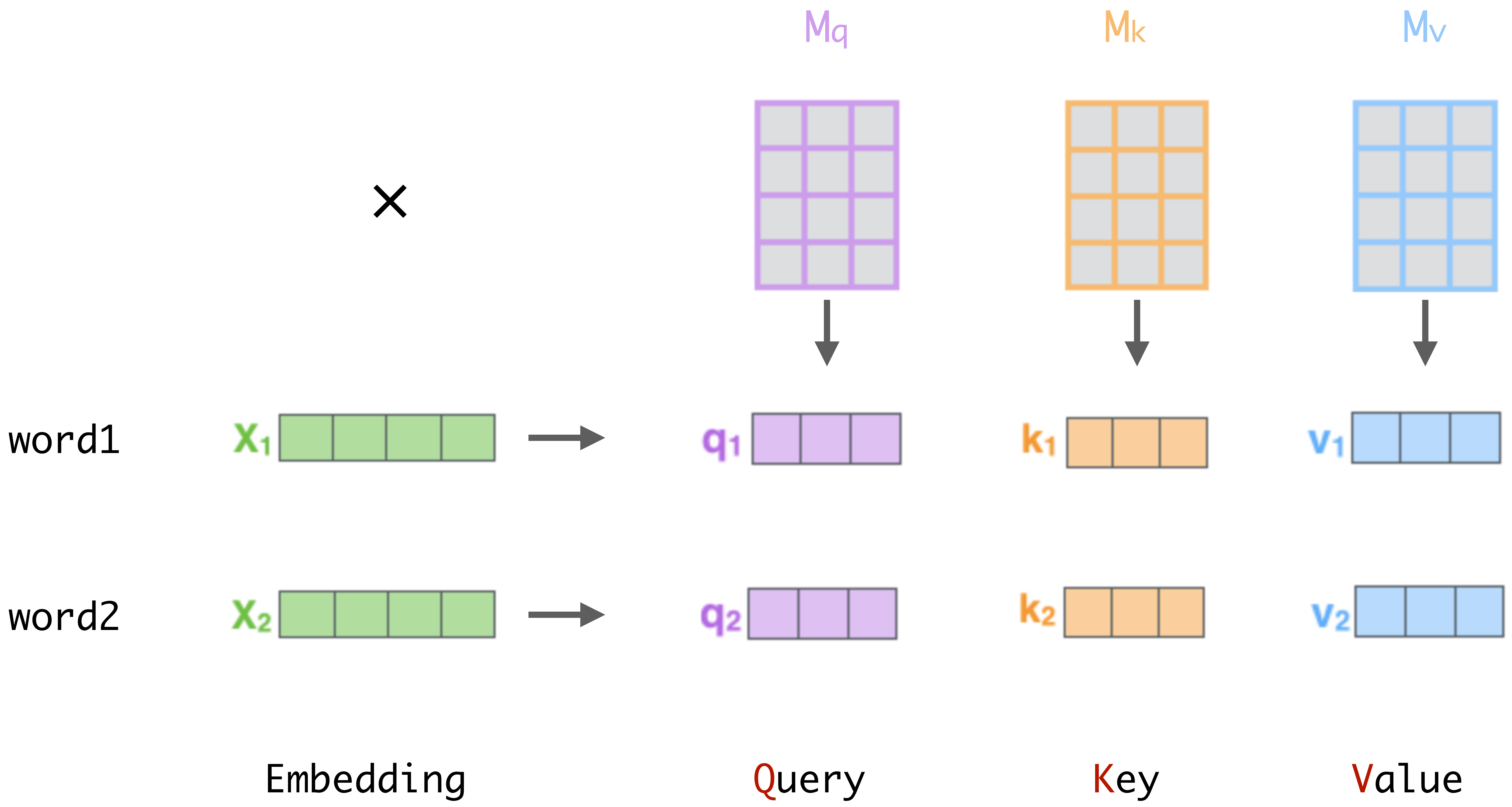


Mq
Mk
Mv
×
word1
X1
q1
k1
v1
word2
X2
q2
k2
v2
Embedding
Query
Key
Value

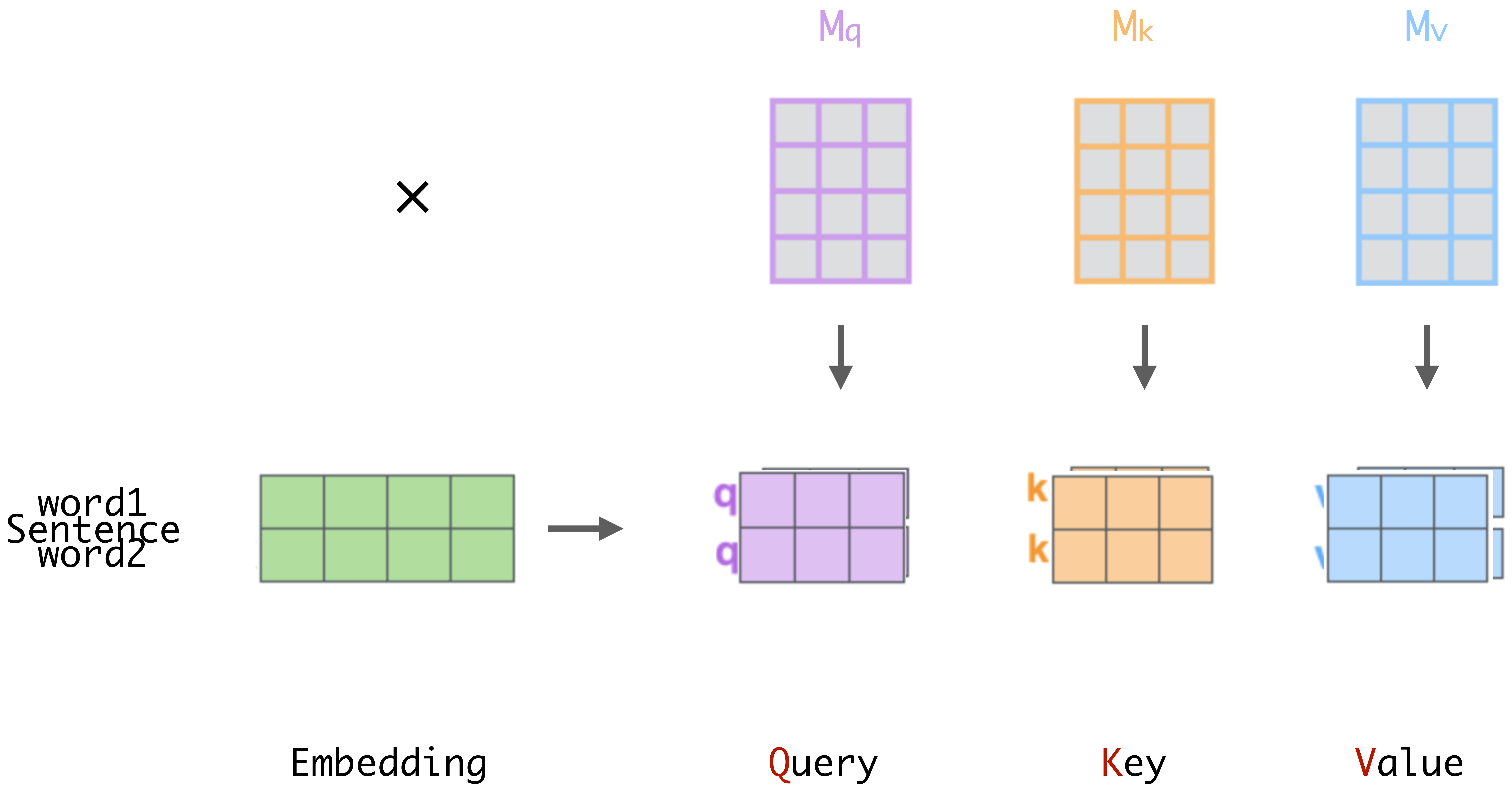
Attention

需要训练的参数：

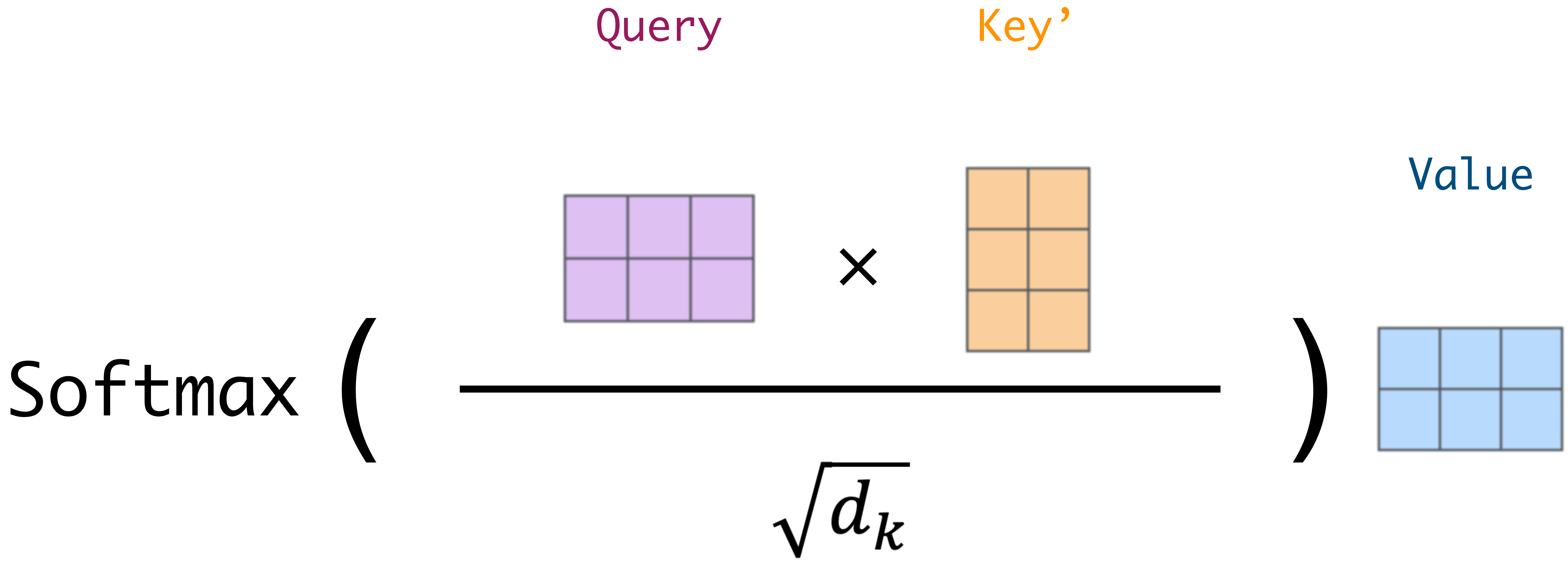




Mq
Mk
Mv
×
word1
X1
q1
k1
v1
word2
X2
q2
k2
v2
Embedding
Query
Key
Value


Mq
Mk
Mv
×
word1
Sentence
word2
q
q
k
k
Embedding
Query
Key
Value

Attention =

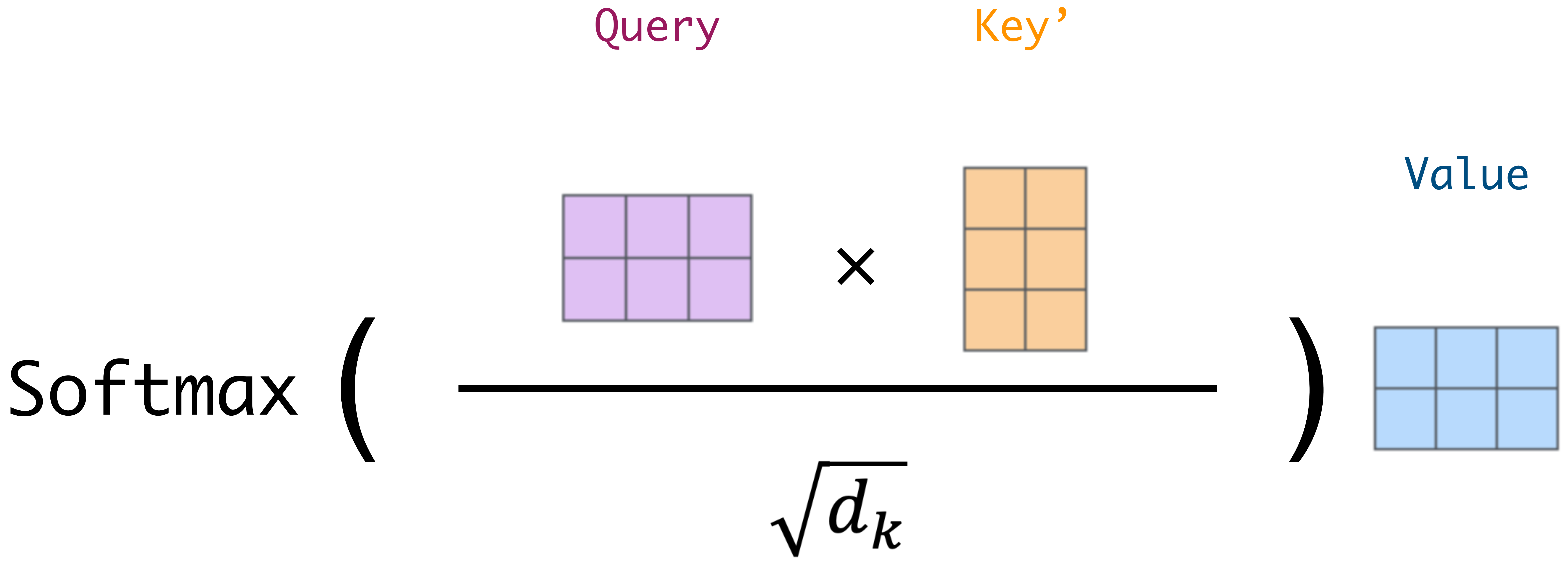


$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

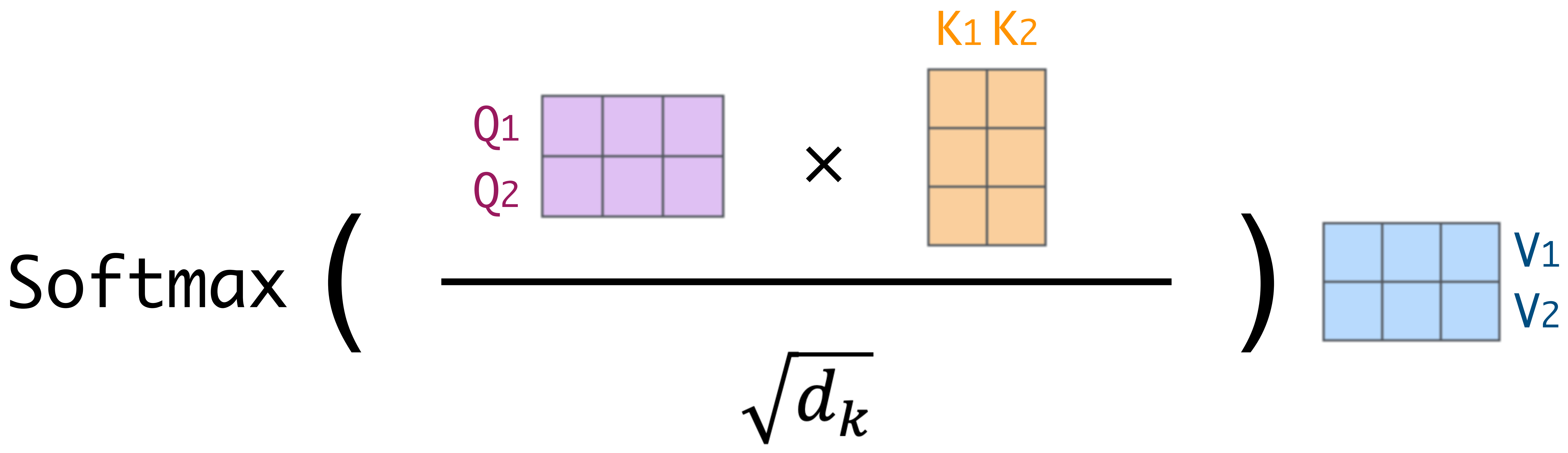
Scaled Dot-Product Attention

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

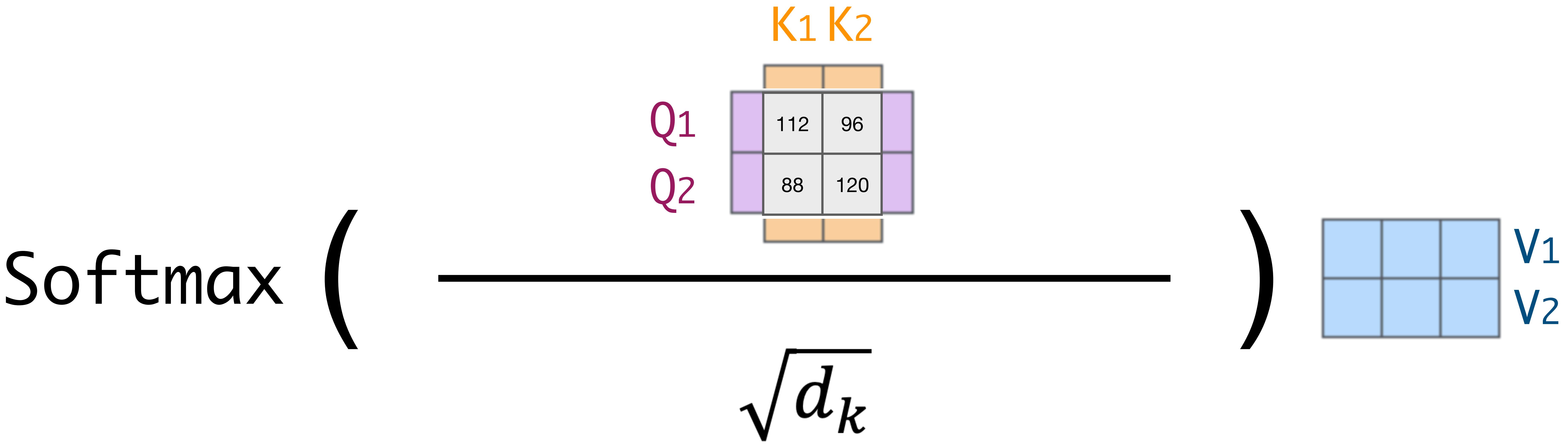
Attention =



Attention =



Attention =



Attention =

$$\mathrm{Softmax}\left(\frac{\begin{array}{cc} & \begin{array}{cc} W_1 & W_2 \end{array} \\ \begin{array}{c} W_1 \\ W_2 \end{array} & \begin{bmatrix} 112 & 96 \\ 88 & 120 \end{bmatrix} \end{array}}{\sqrt{d_k}} \right) \begin{array}{cc} \begin{bmatrix} \; & \; & \; \\ \; & \; & \; \end{bmatrix} & \begin{array}{c} V_1 \\ V_2 \end{array} \end{array}$$

Attention =

$$\mathrm{Softmax}\left(\frac{\begin{array}{c|cc} & W_1 & W_2 \\ \hline W_1 & 112 & 96 \\ W_2 & 88 & 120 \end{array}}{\sqrt{d_k}}\right)\begin{bmatrix} \; & \; & \; \\ \; & \; & \; \end{bmatrix}\begin{matrix} V_1 \\ V_2 \end{matrix}$$

Scaling factor

Attention =

$$\text{Softmax}\left(\frac{\begin{array}{ccc} & W_1 & W_2 \\ W_1 & 112 & 96 \\ W_2 & 88 & 120 \end{array}}{\sqrt{d_k}}\right) \begin{array}{cc} \begin{bmatrix} \; & \; & \; \\ \; & \; & \; \end{bmatrix} & \begin{array}{c} V_1 \\ V_2 \end{array} \end{array}$$

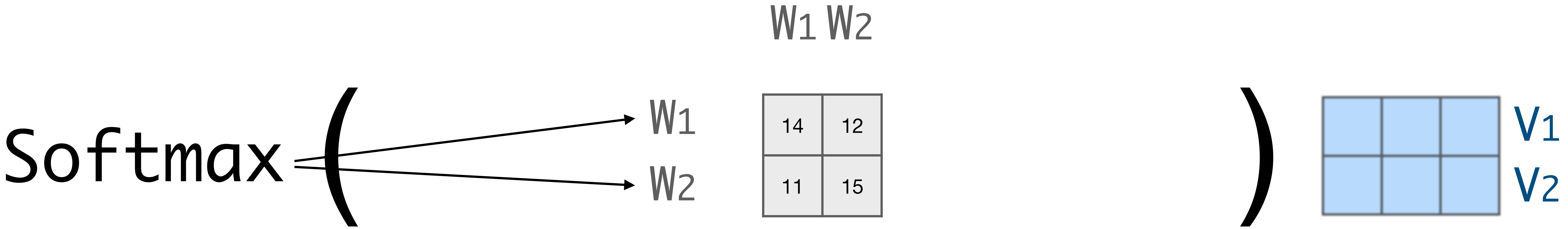
Attention =

$$\text{Softmax}\left(\frac{\begin{array}{cc} & \begin{array}{cc} W_1 & W_2 \end{array} \\ \begin{array}{c} W_1 \\ W_2 \end{array} & \begin{bmatrix} 14 & 12 \\ 11 & 15 \end{bmatrix} \end{array}}{\qquad\qquad\qquad}\right) \begin{array}{cc} \begin{bmatrix} \; & \; & \; \\ \; & \; & \; \end{bmatrix} & \begin{array}{c} V_1 \\ V_2 \end{array} \end{array}$$

Attention =

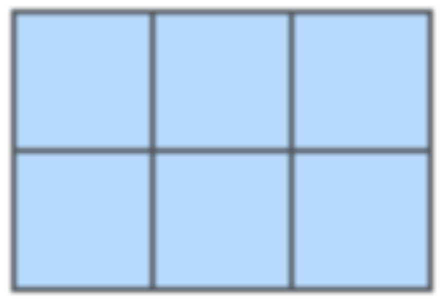
$$\mathrm{Softmax}\left(\begin{array}{cc} & \begin{array}{cc} W_1 & W_2 \end{array} \\ \begin{array}{c} W_1 \\ W_2 \end{array} & \begin{bmatrix} 14 & 12 \\ 11 & 15 \end{bmatrix} \end{array} \right) \begin{array}{cc} \begin{bmatrix} \; & \; & \; \\ \; & \; & \; \end{bmatrix} & \begin{array}{l} V_1 \\ V_2 \end{array} \end{array}$$

Attention =



Attention =

$$\left(\begin{array}{cc} & \begin{array}{cc} W_1 & W_2 \end{array} \\ \begin{array}{c} W_1 \\ W_2 \end{array} & \begin{bmatrix} 0.88 & 0.12 \\ 0.02 & 0.98 \end{bmatrix} \end{array} \right)$$

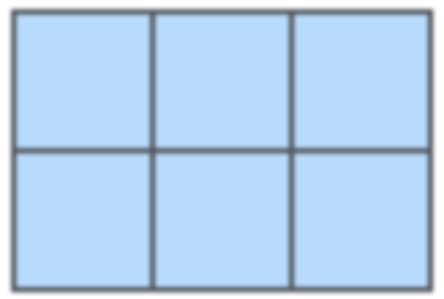


V_1
V_2

Attention =

$$\left(\begin{array}{ccc} & W_1 & W_2 \\ W_1 & 0.88 & 0.12 \\ W_2 & 0.02 & 0.98 \end{array} \right)$$

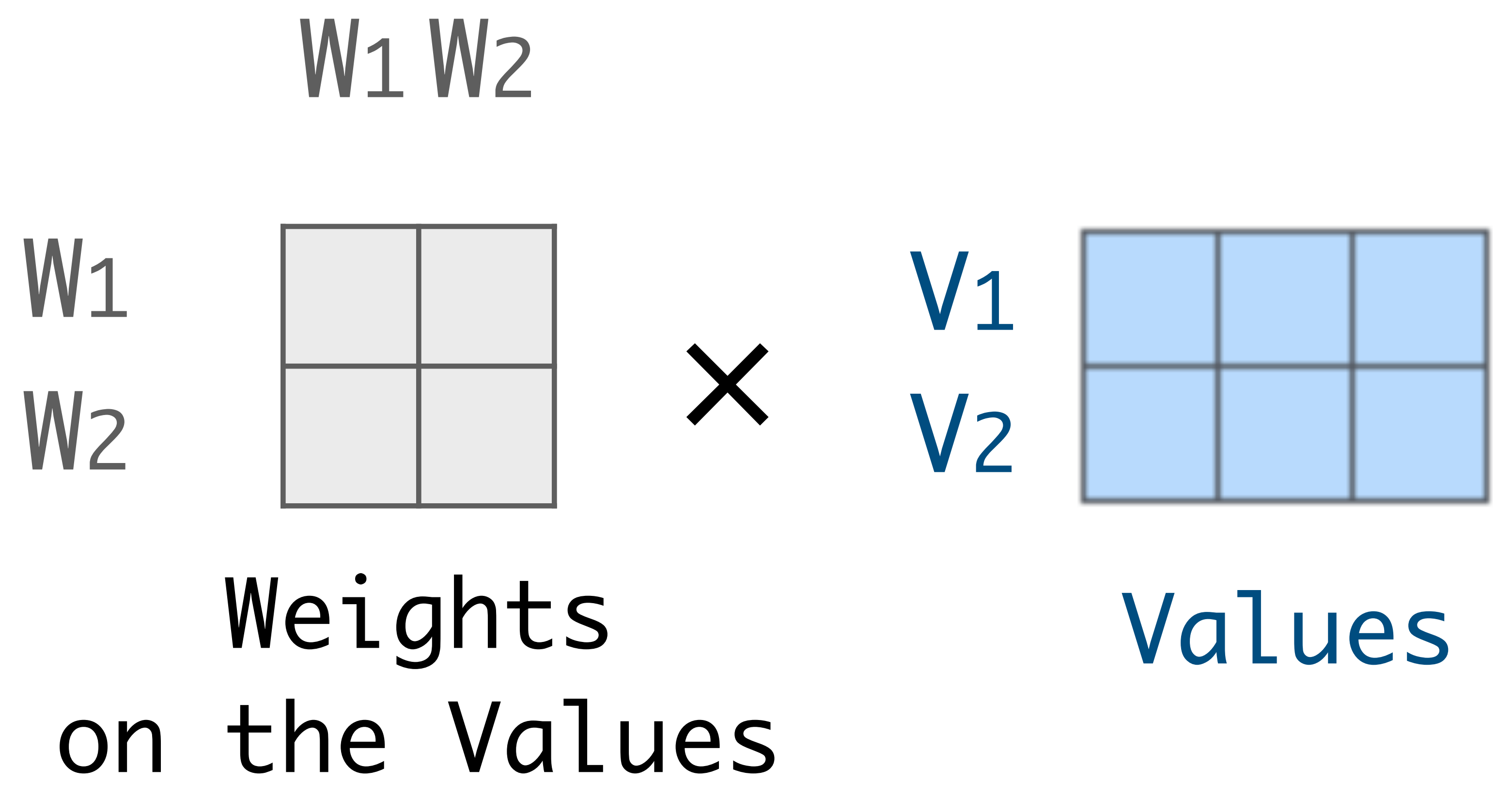
Weights
on the Values



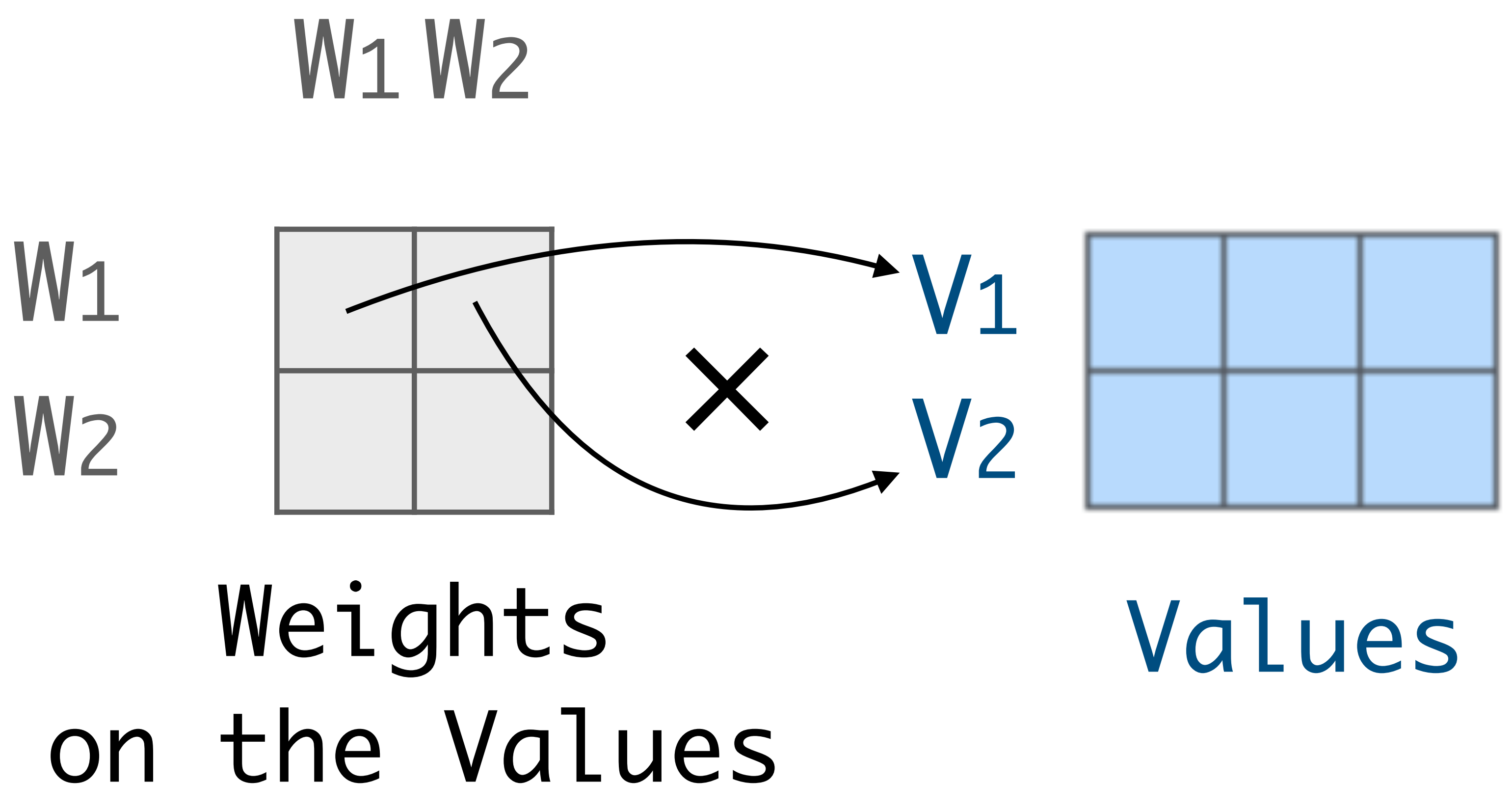
V_1
V_2

Values

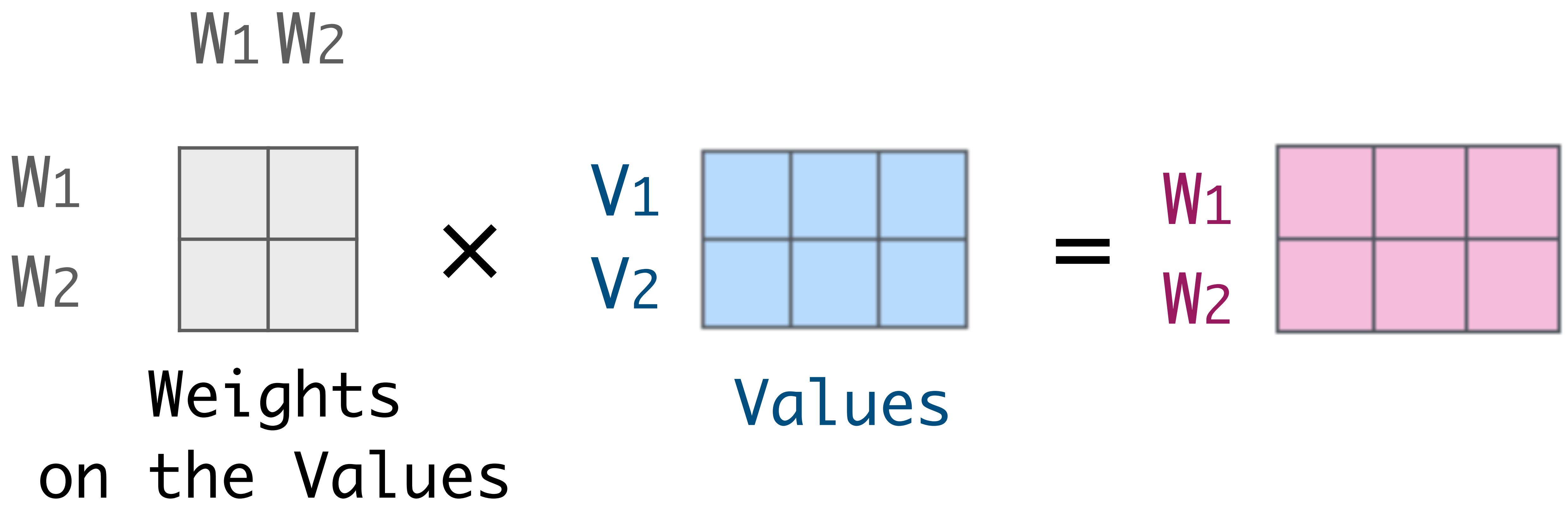
Attention =



Attention =



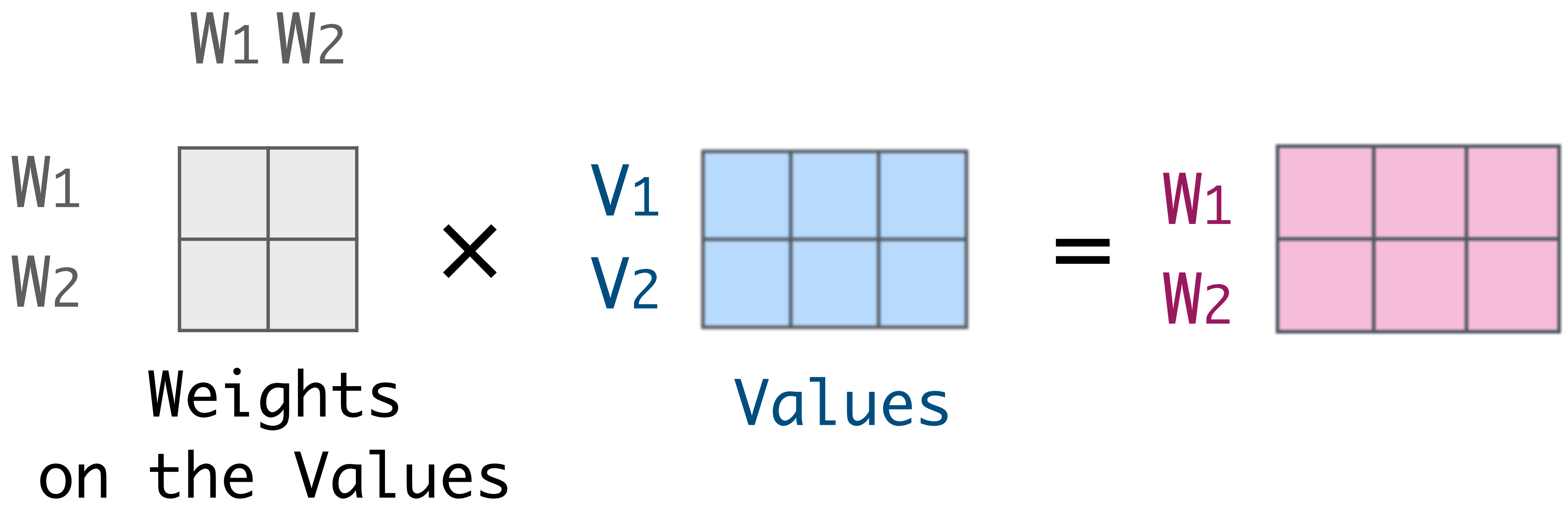
Attention =



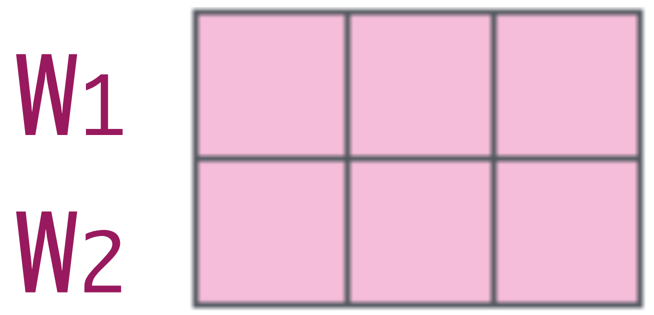
Attention做了什么？

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 =

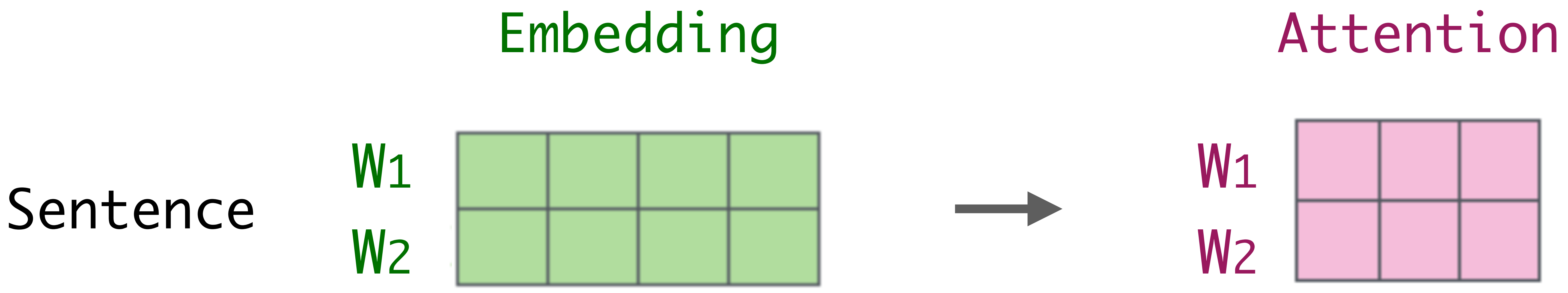


Attention



注意力

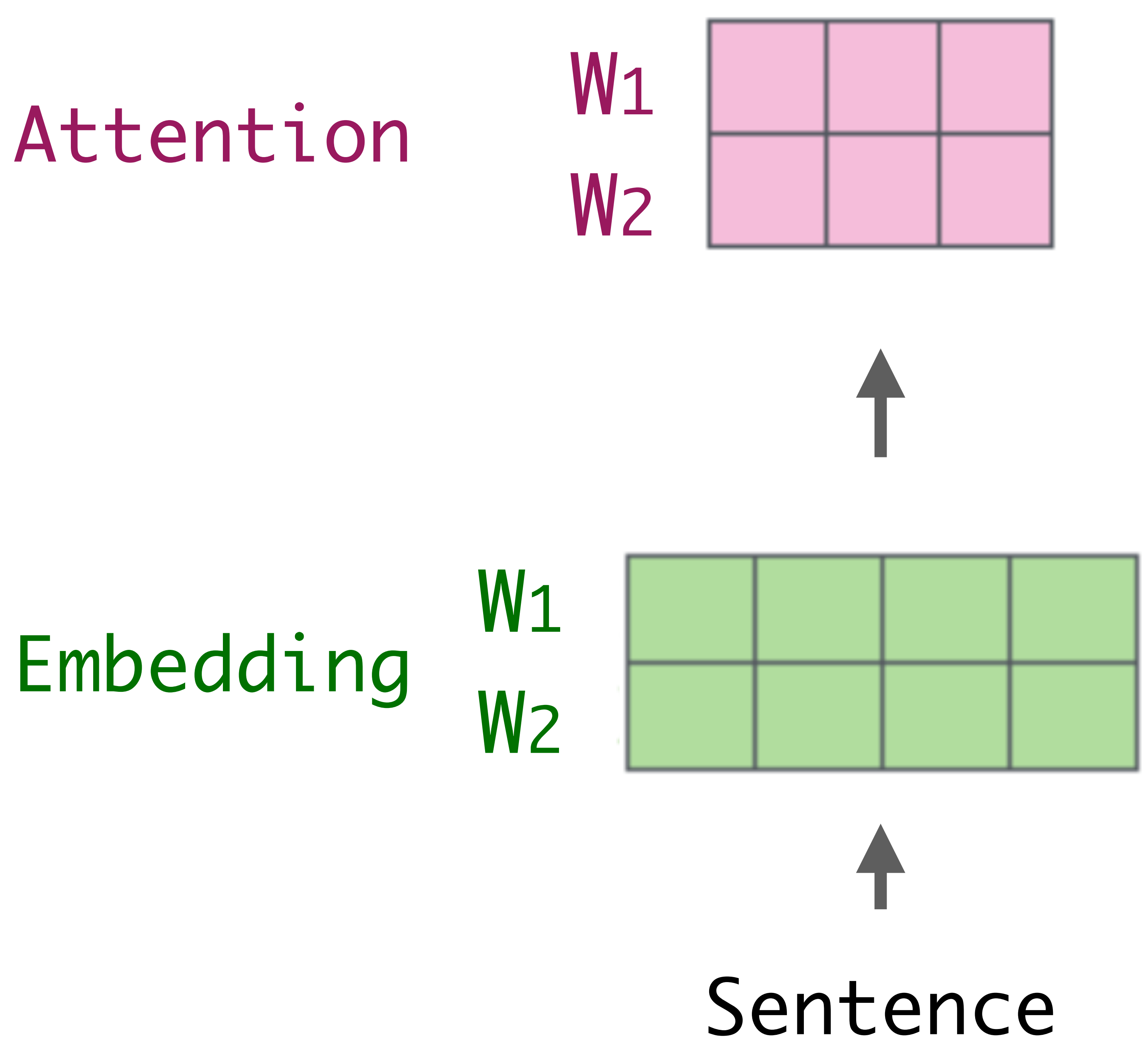
Attention

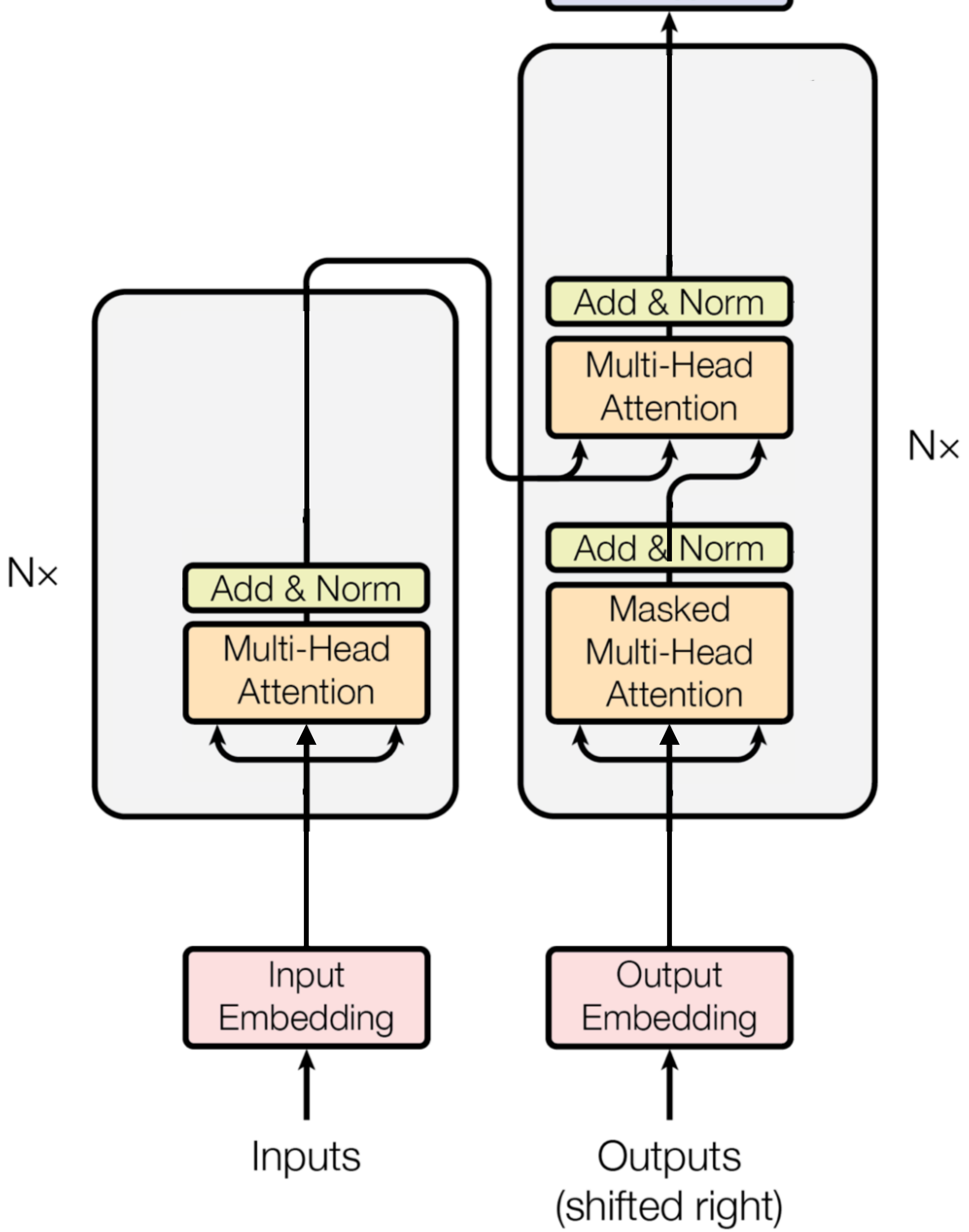


$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

注意力

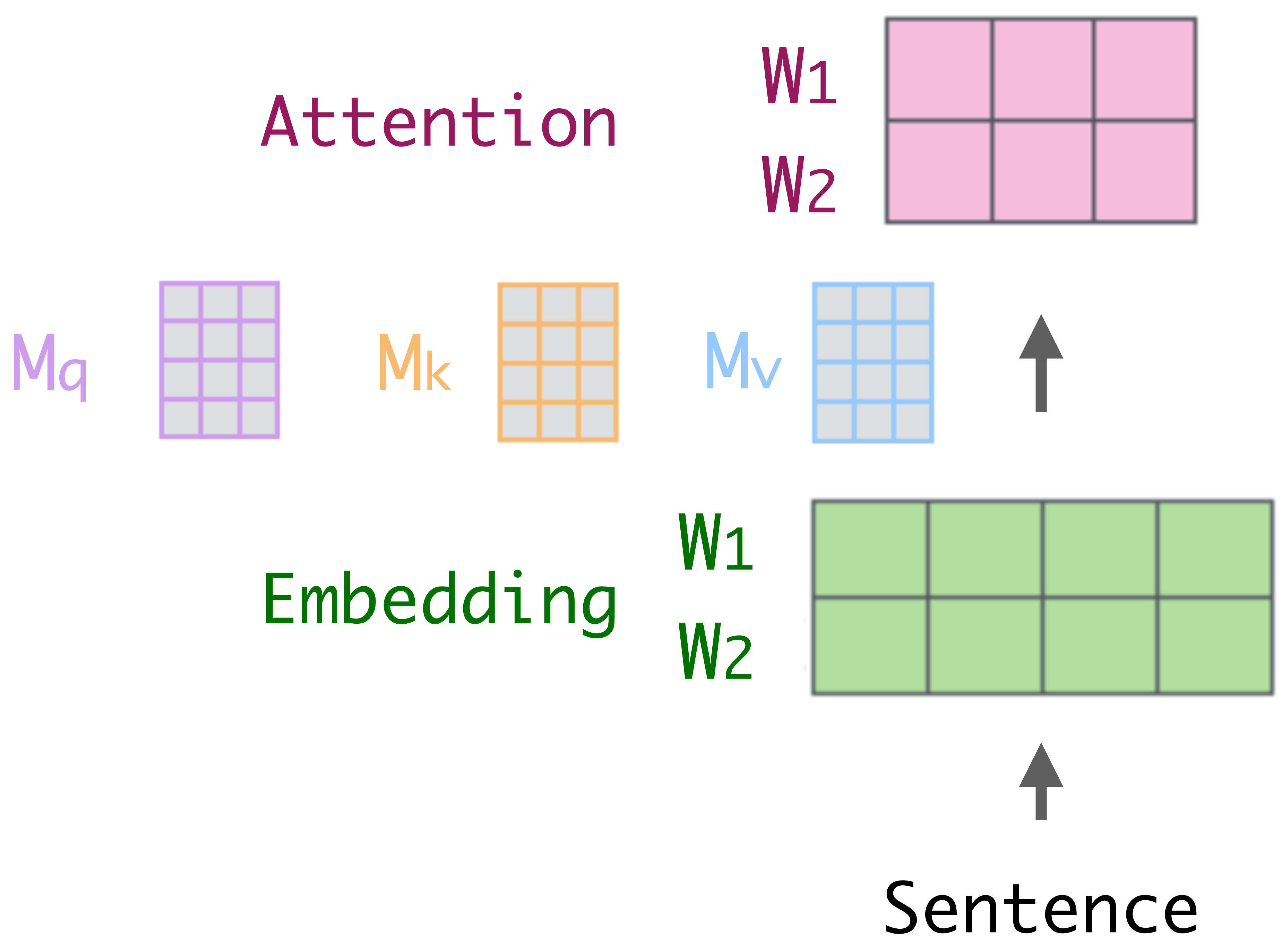
Attention

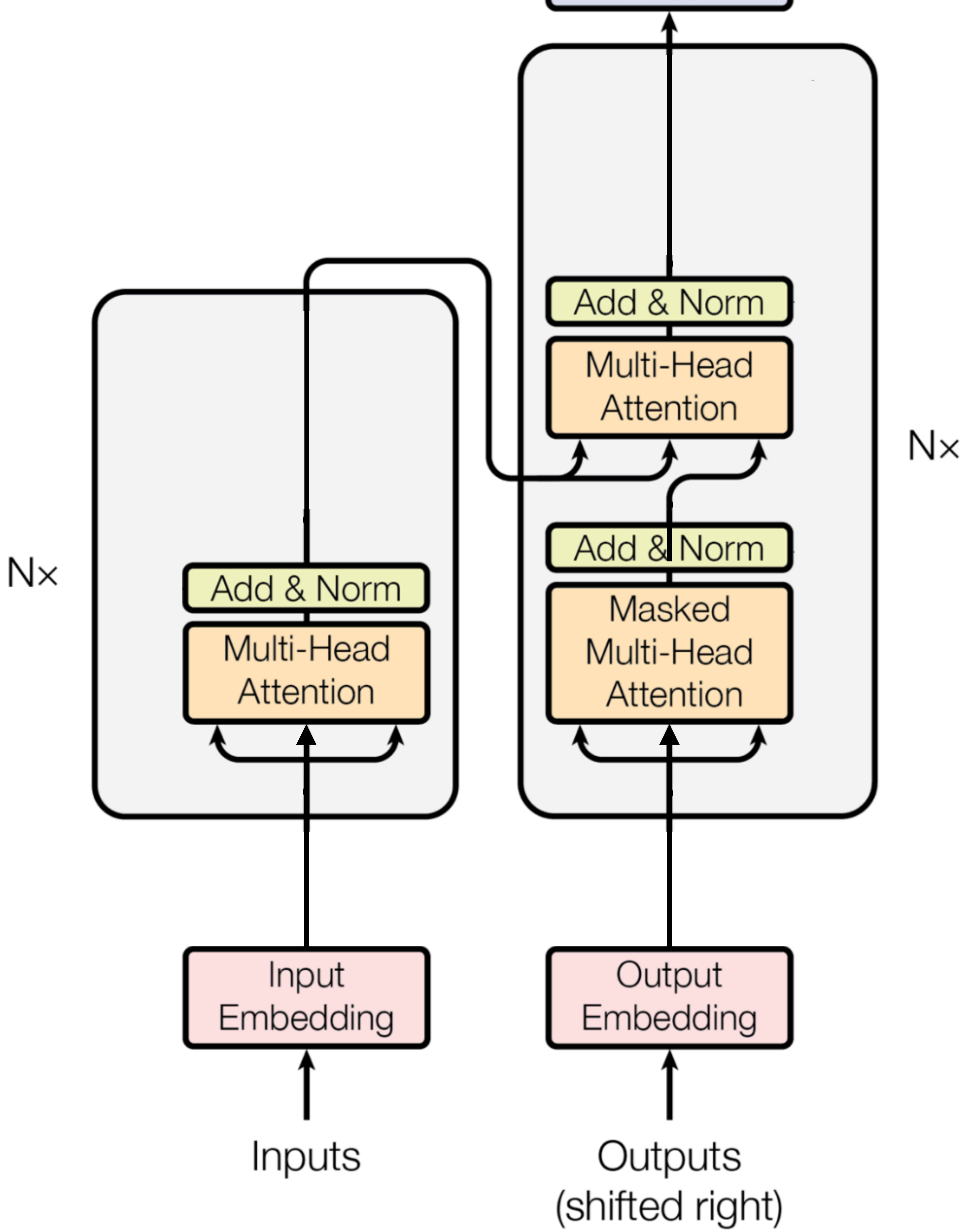




注意力

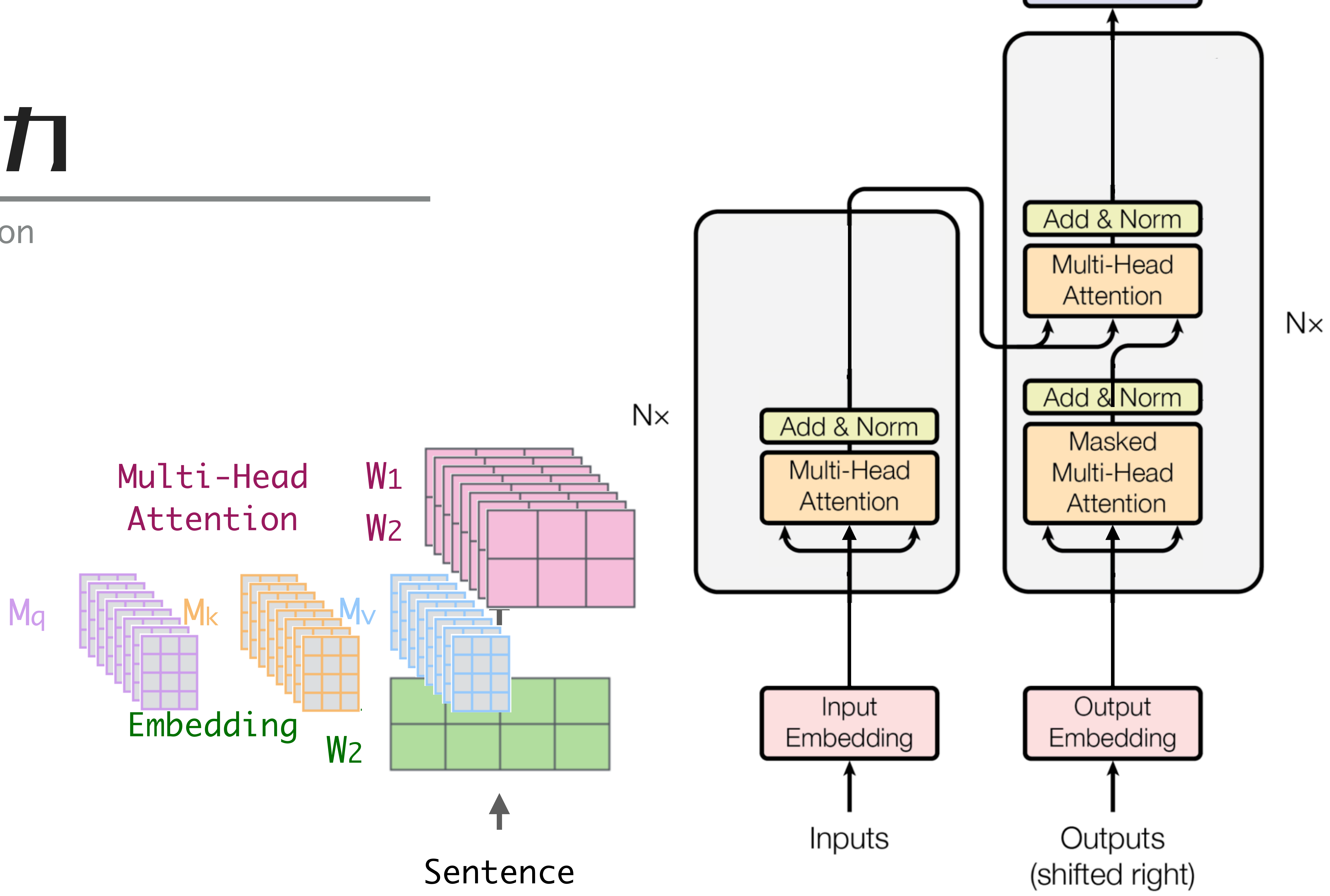
Attention





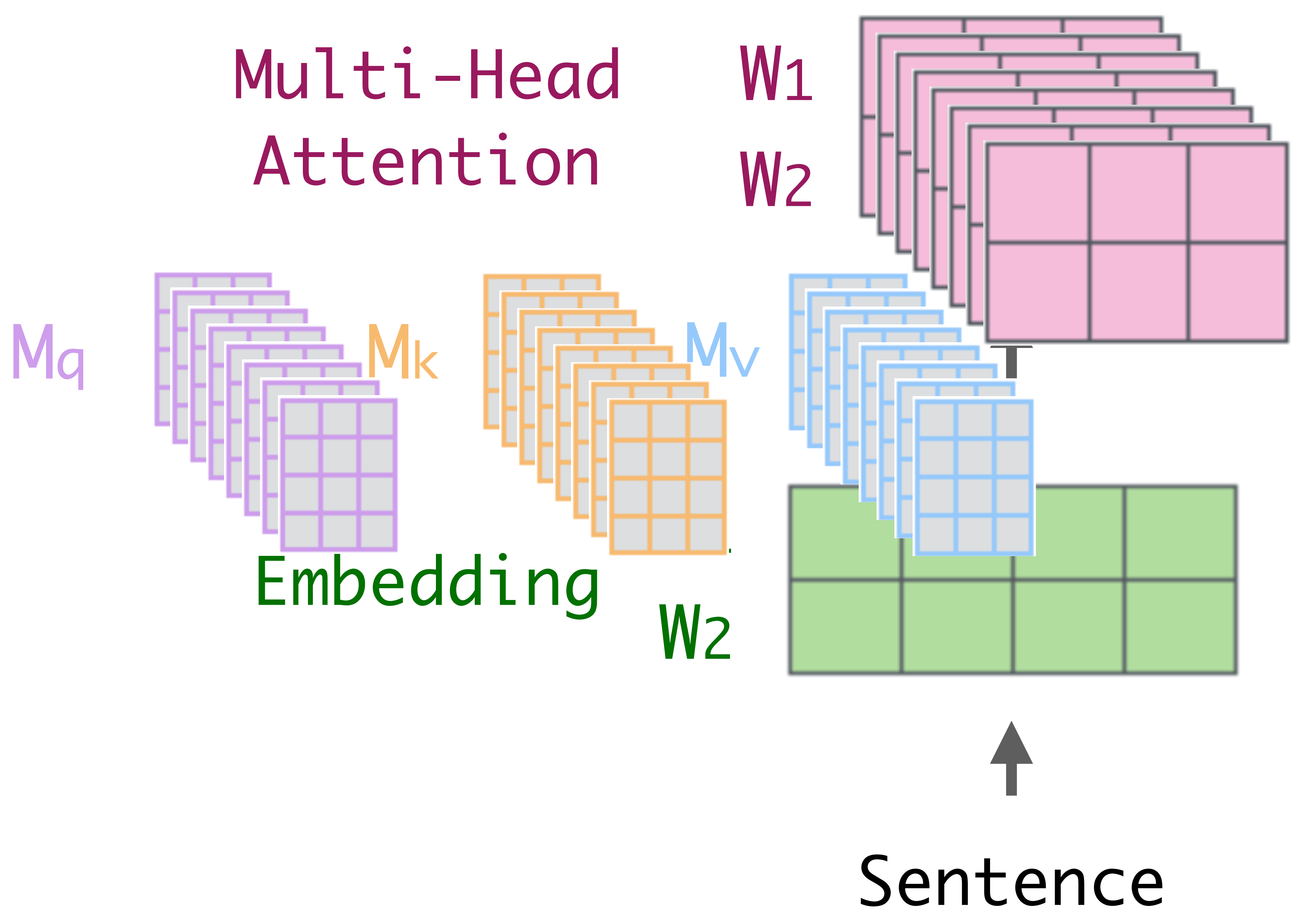
注意力

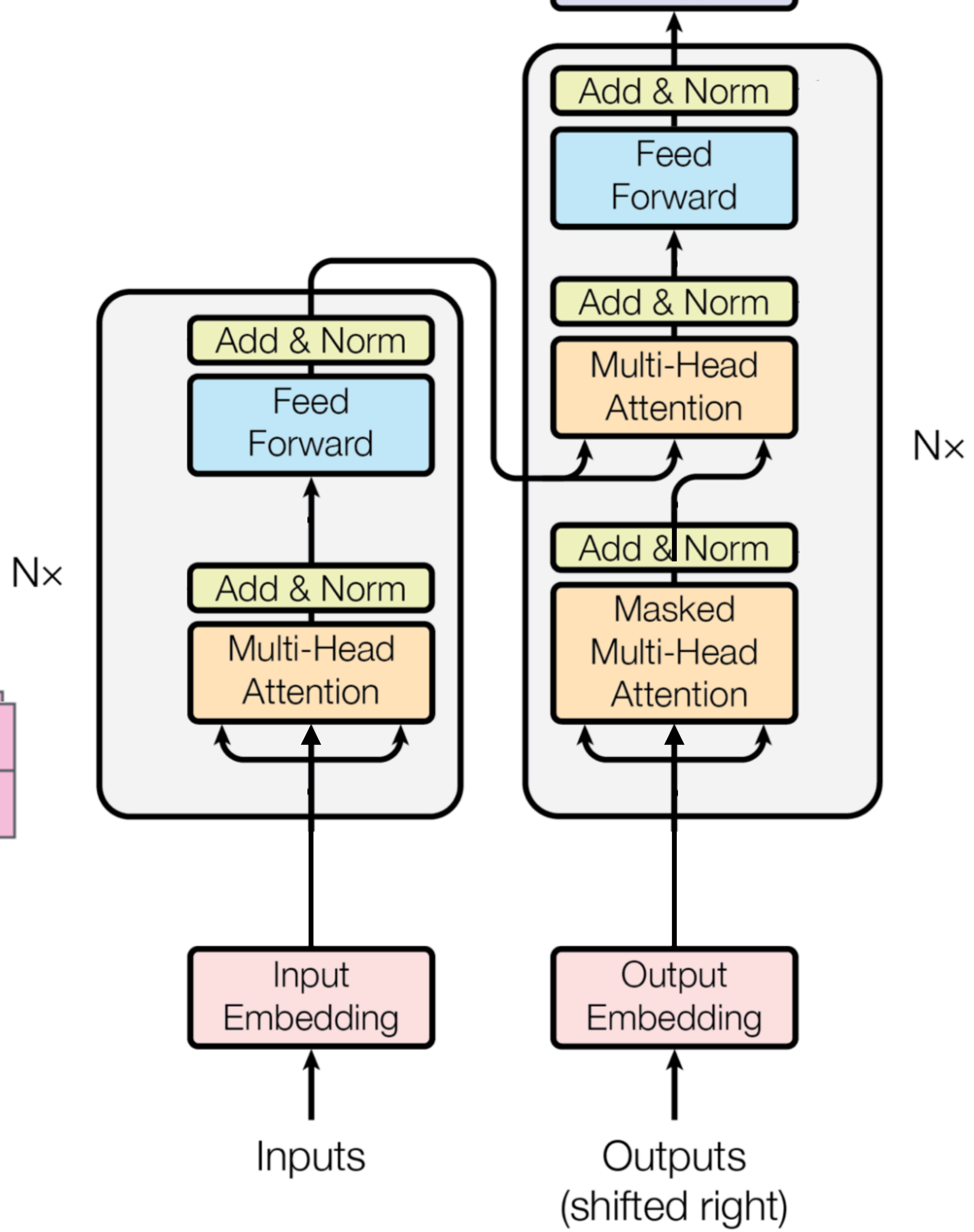
Attention



注意力

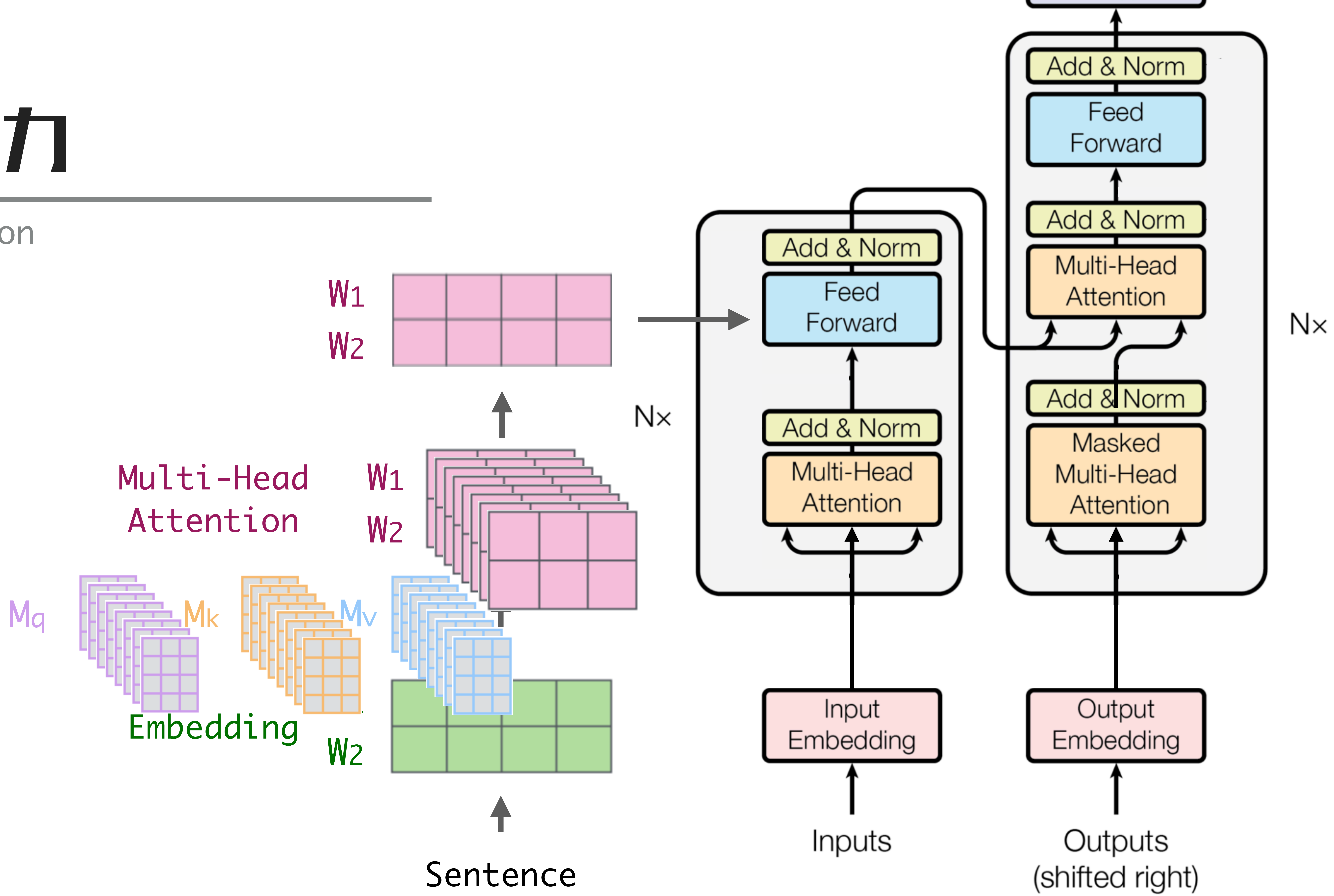
Attention





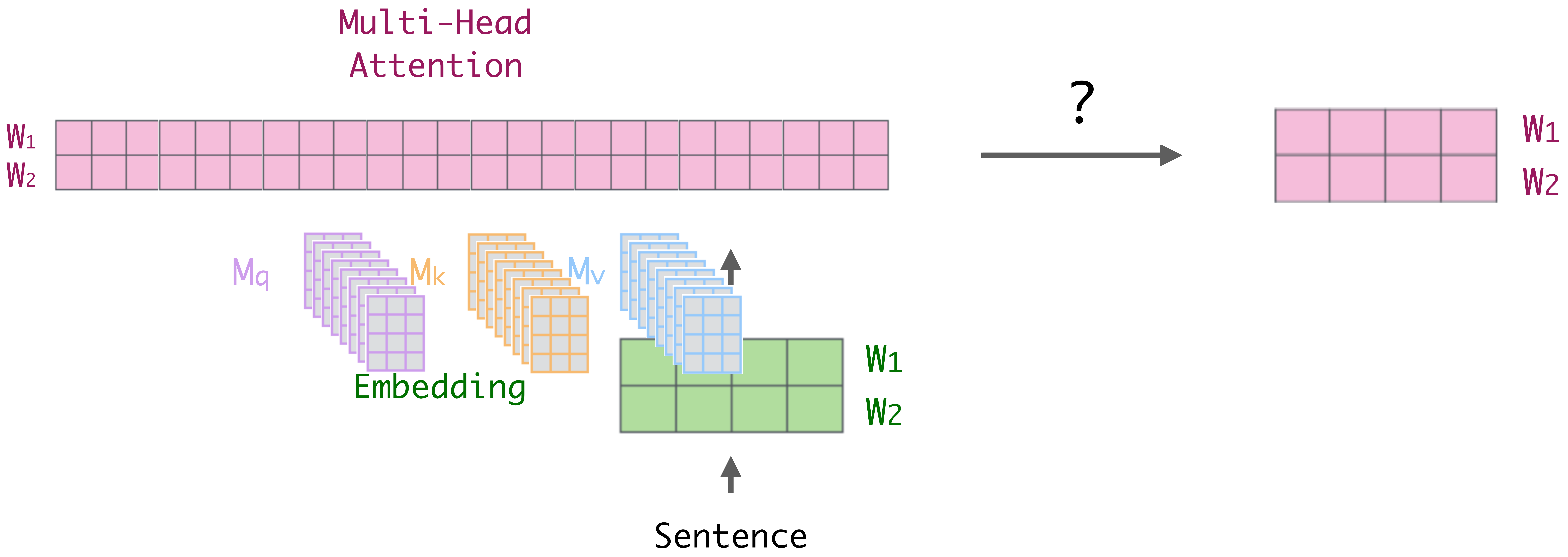
注意力

Attention



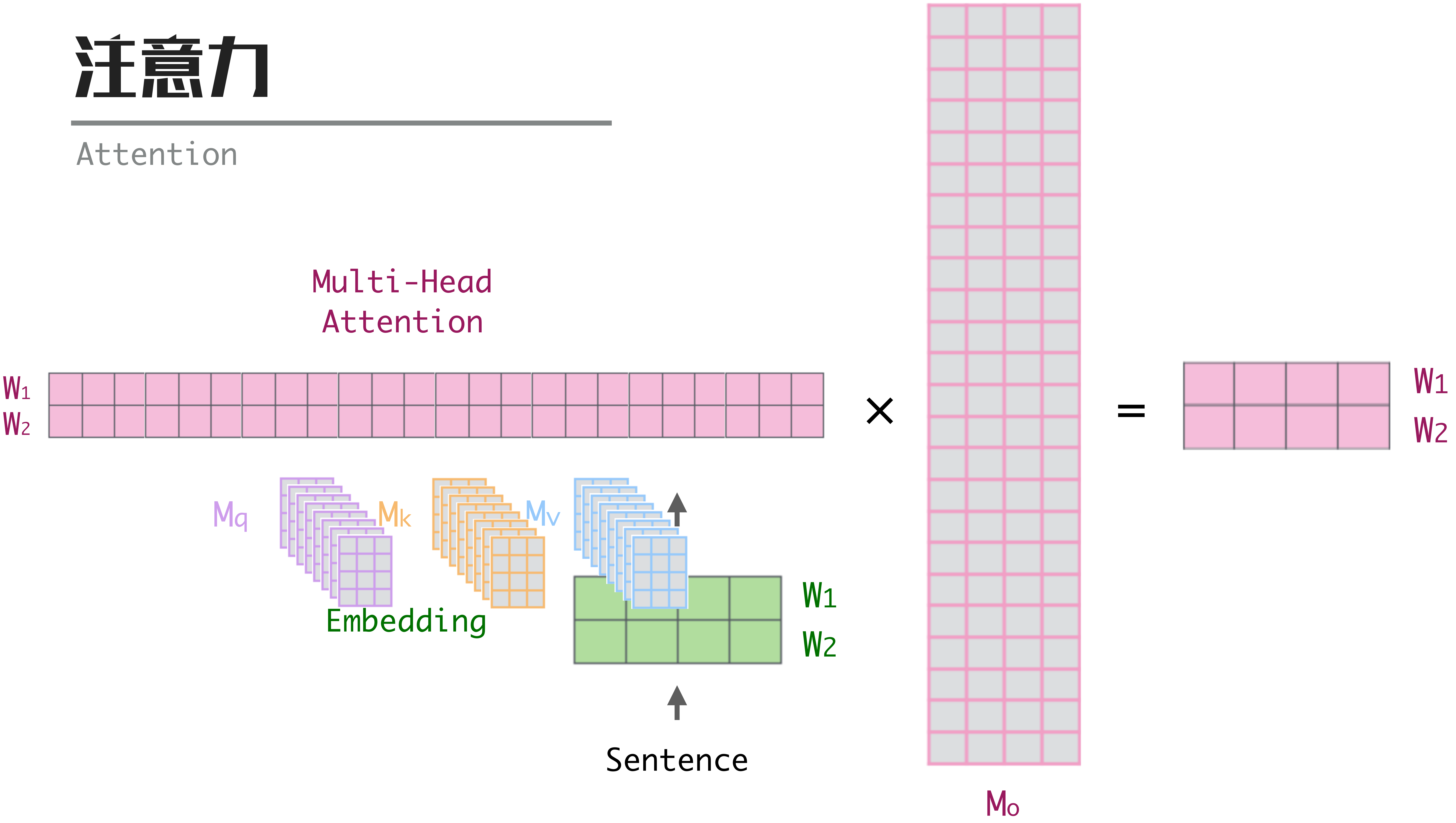
注意力

Attention



注意力

Attention



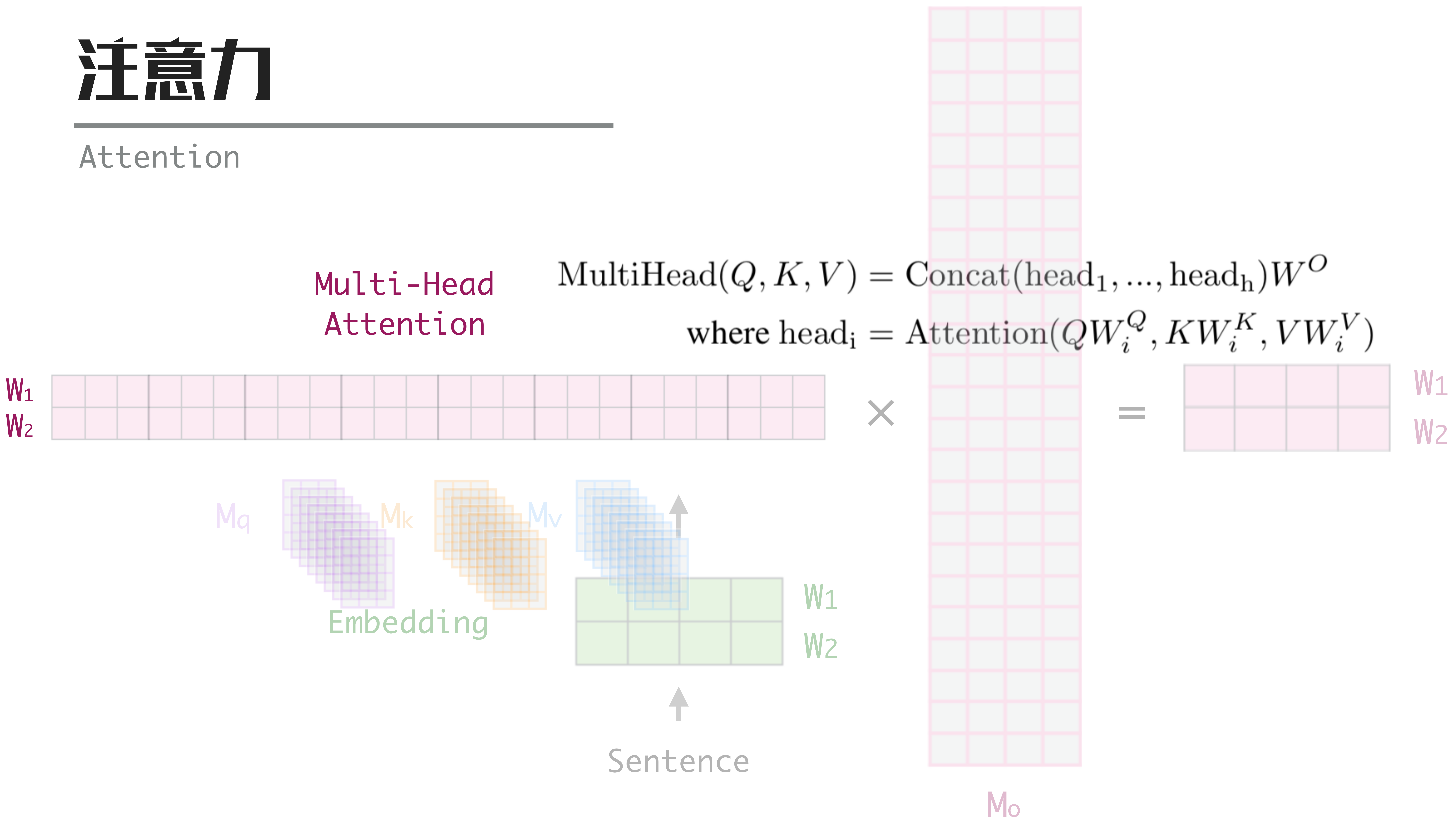
注意力

Attention

Multi-Head Attention

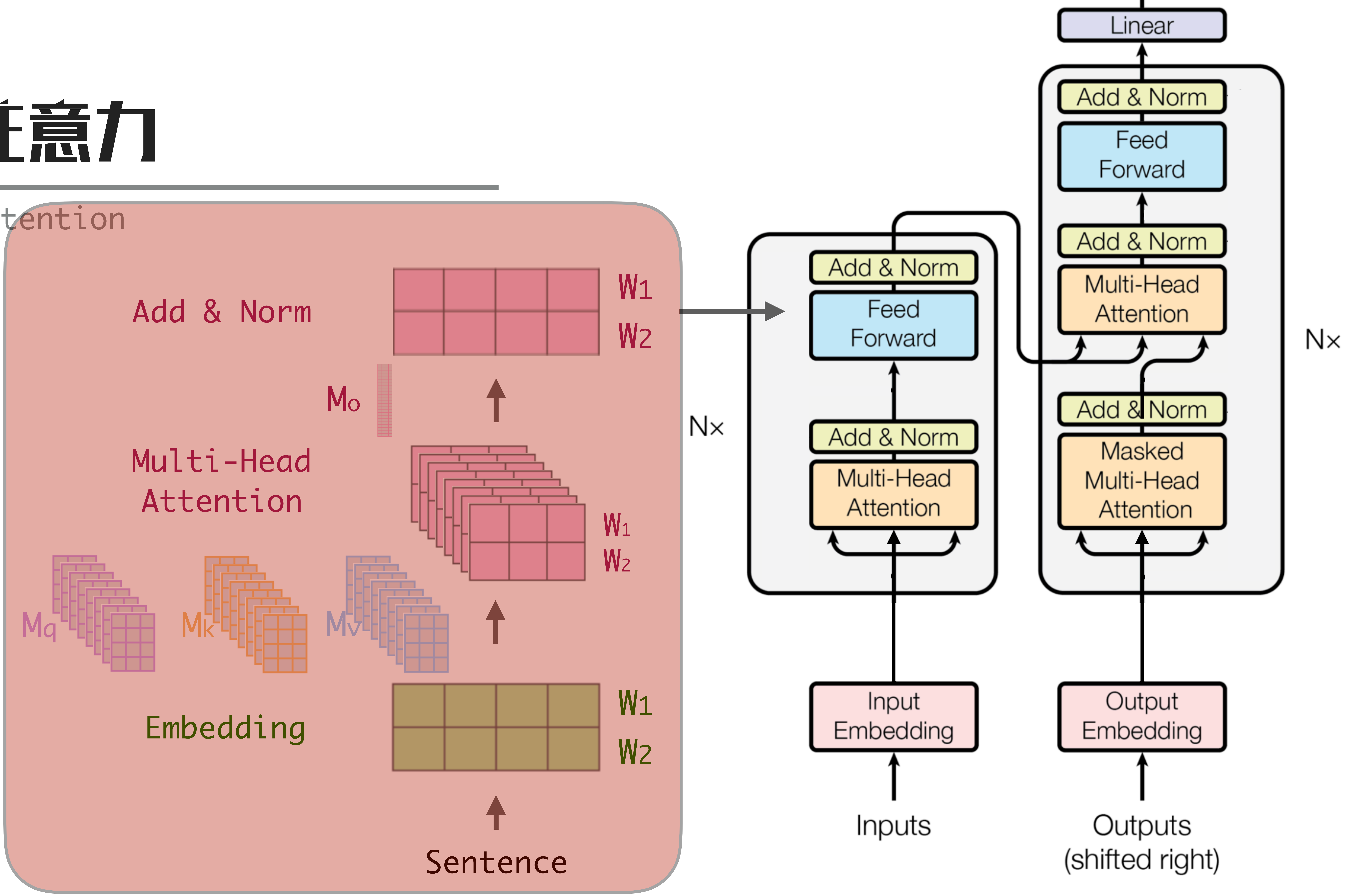
$$\mathrm{MultiHead}(Q,K,V) = \mathrm{Concat}(\mathrm{head_1},\dots,\mathrm{head_h})W^O$$

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



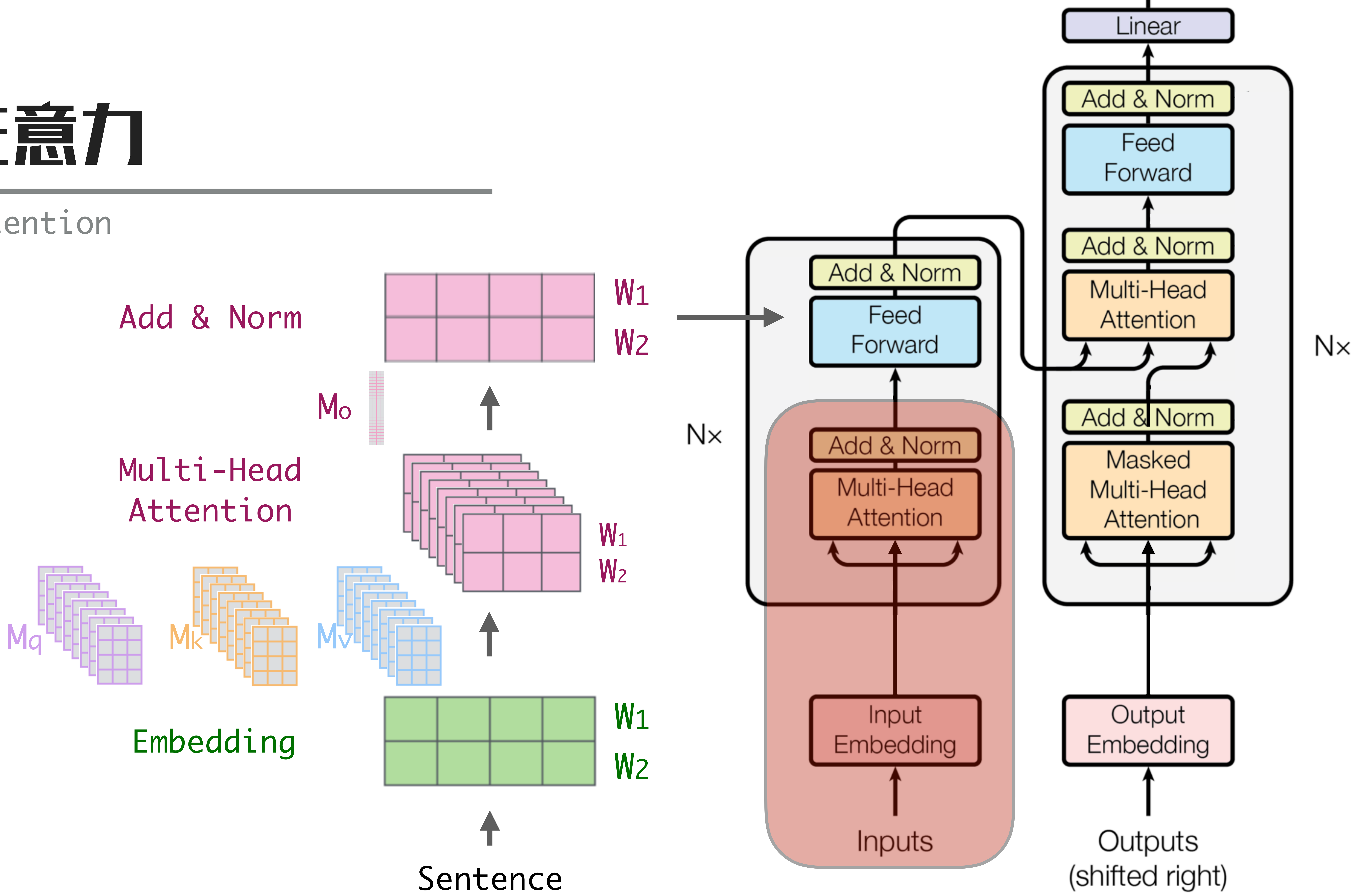
注意力

Attention



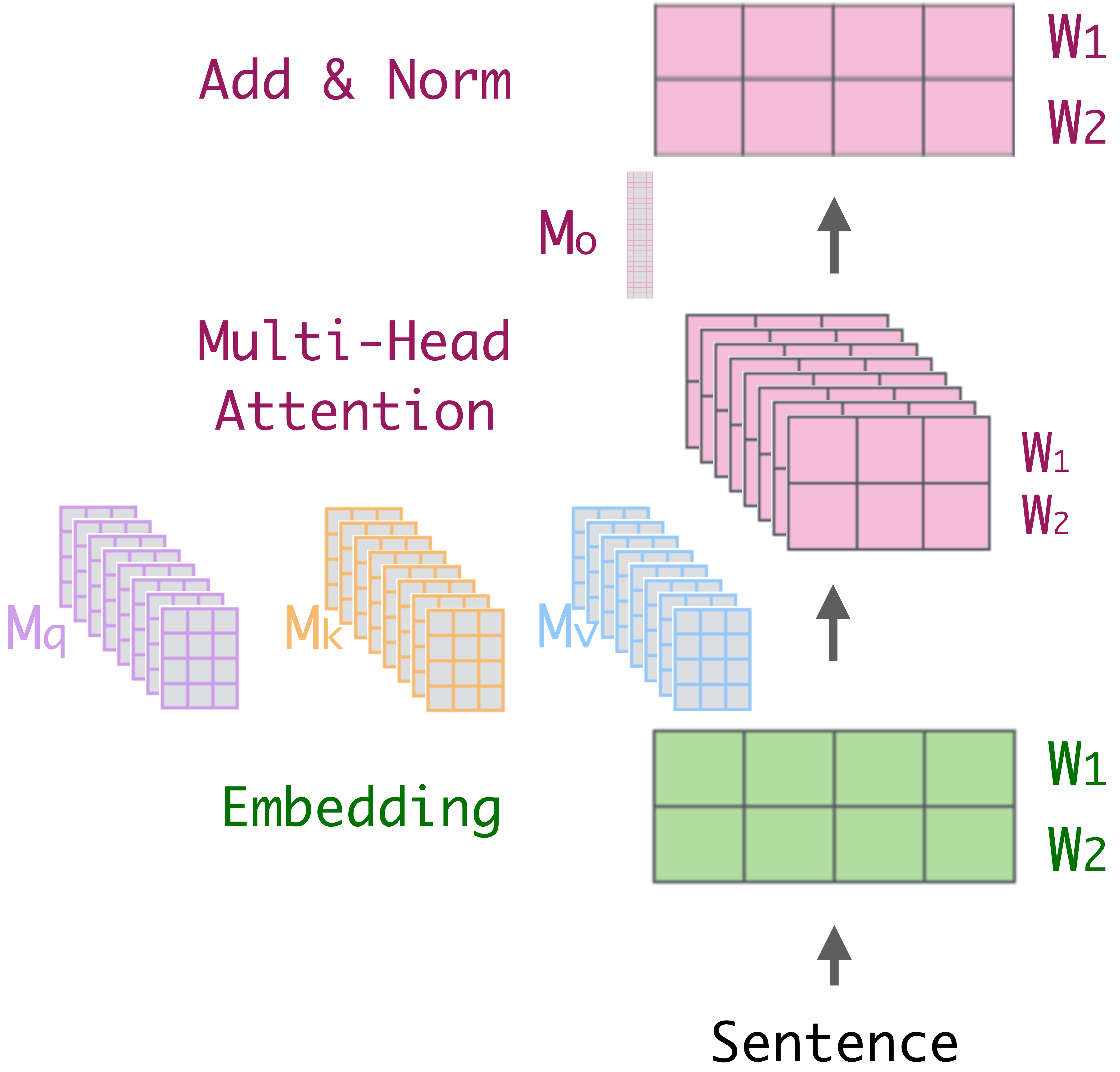
注意力

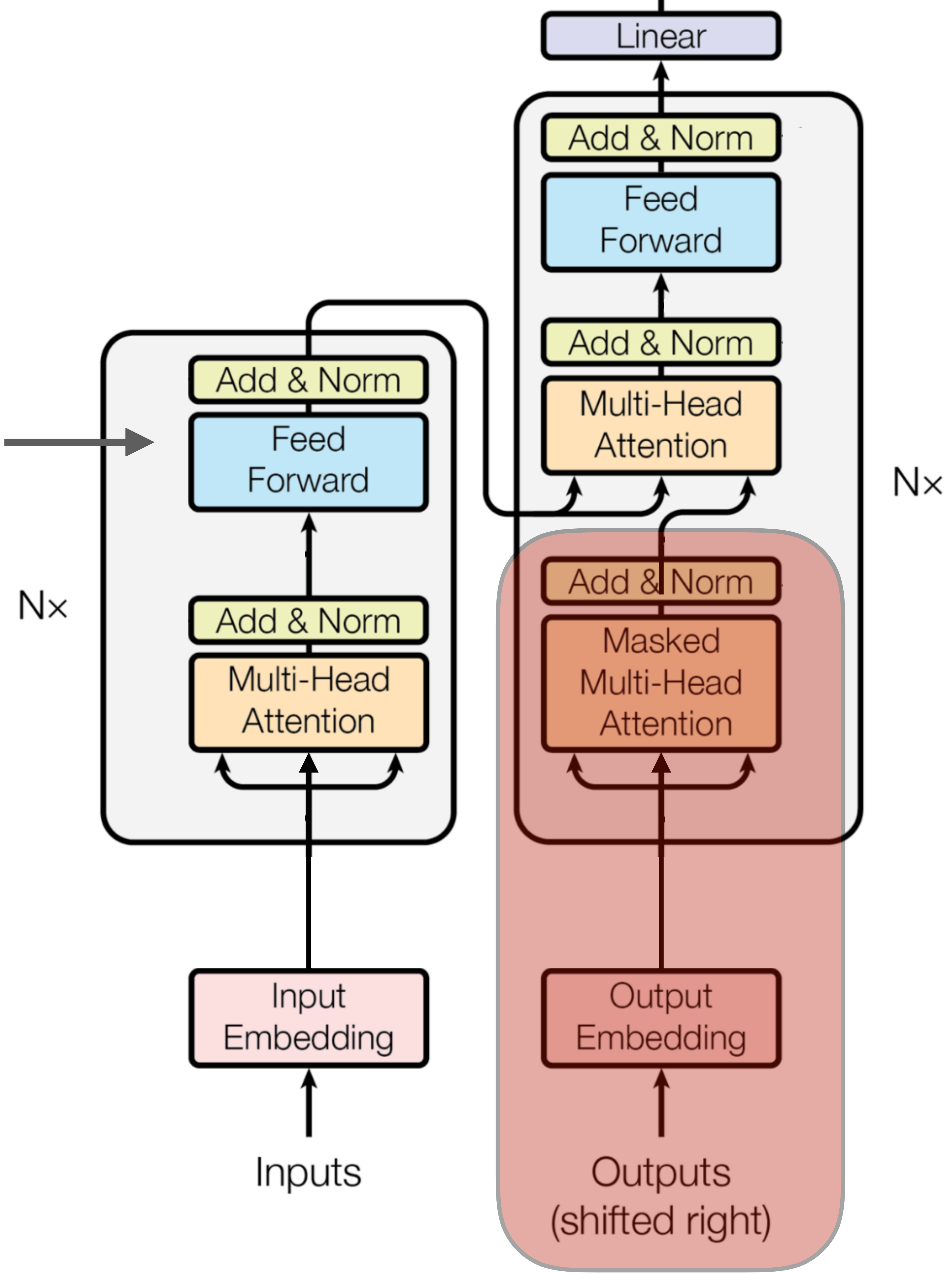
Attention



注意力

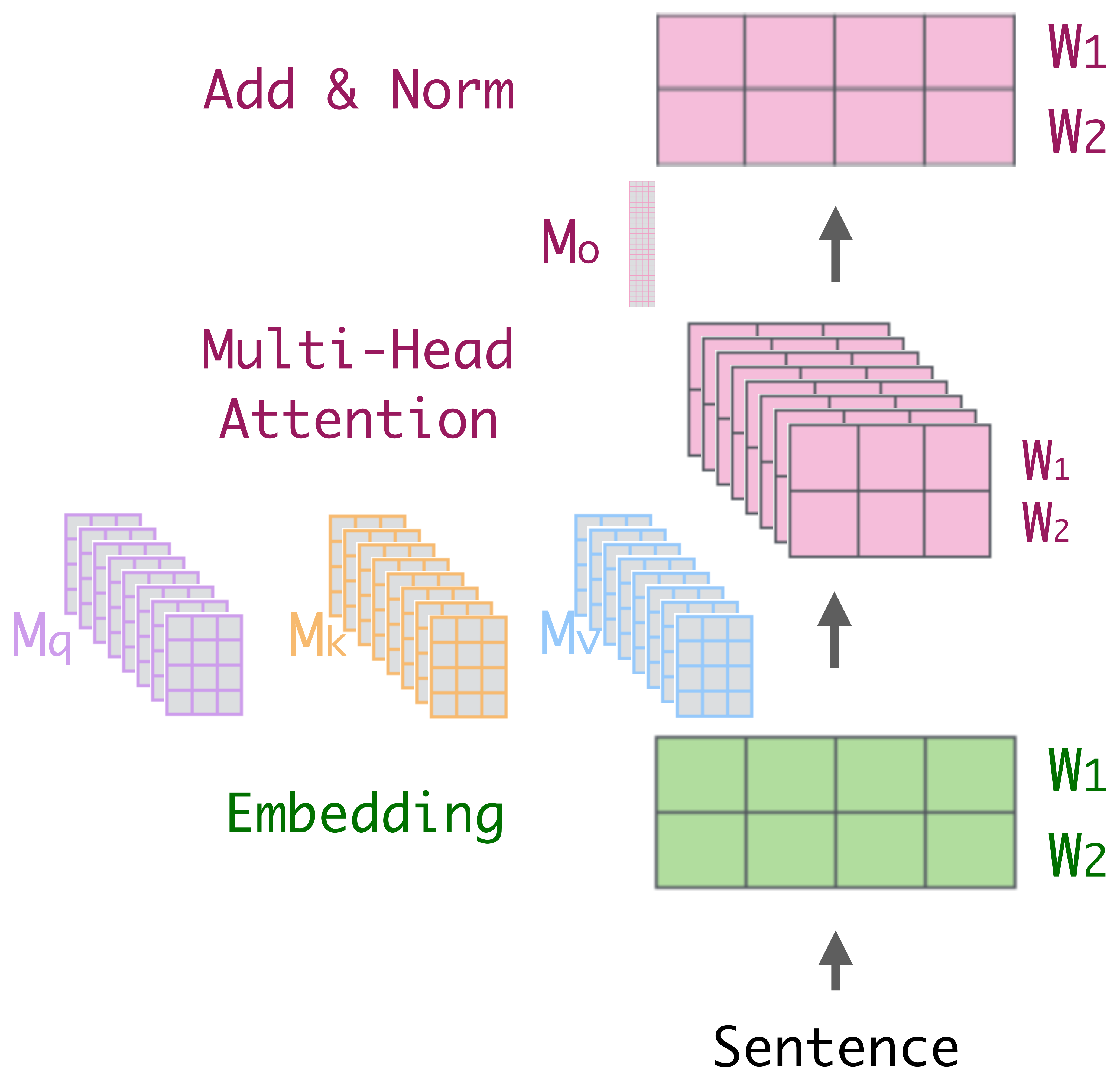
Attention

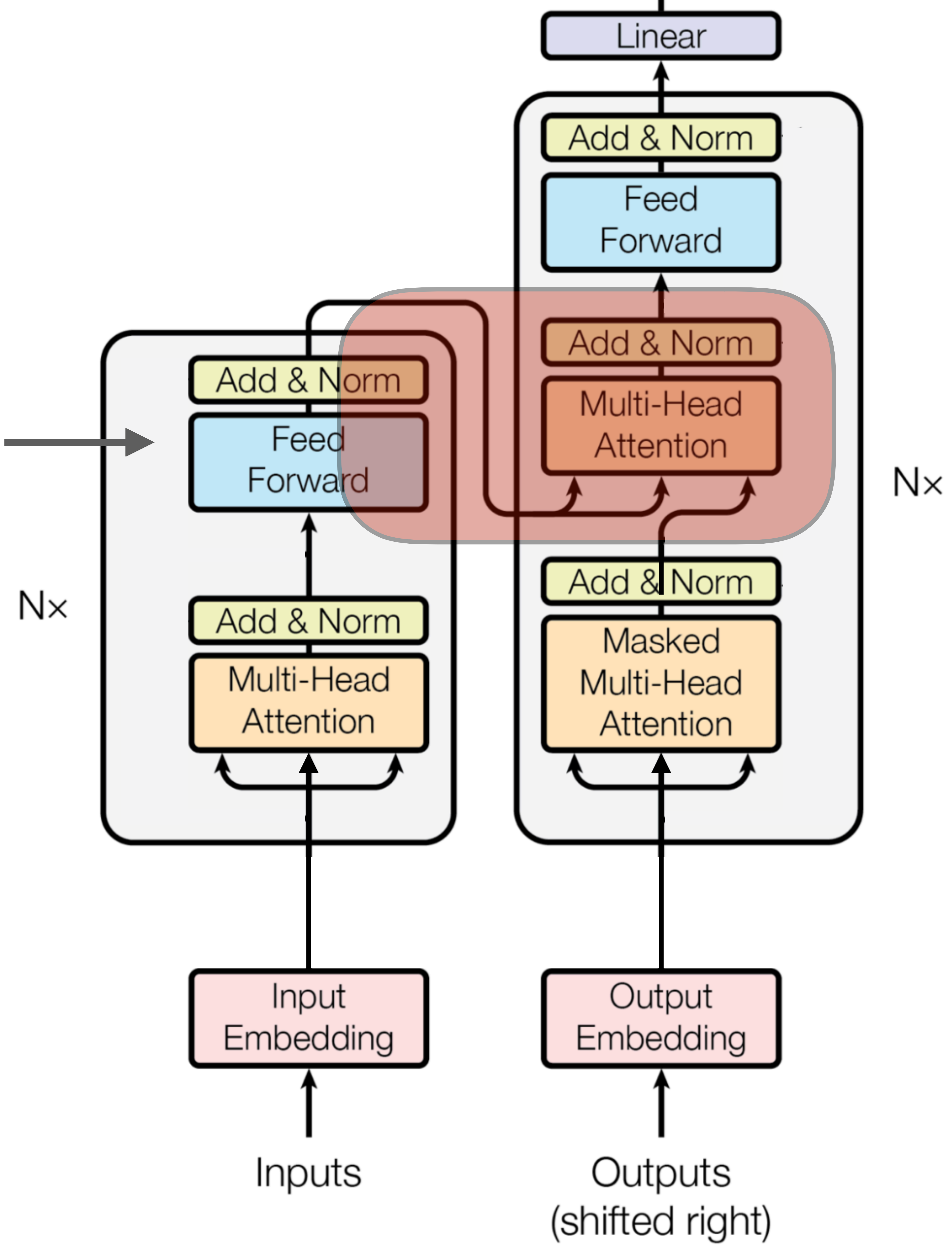


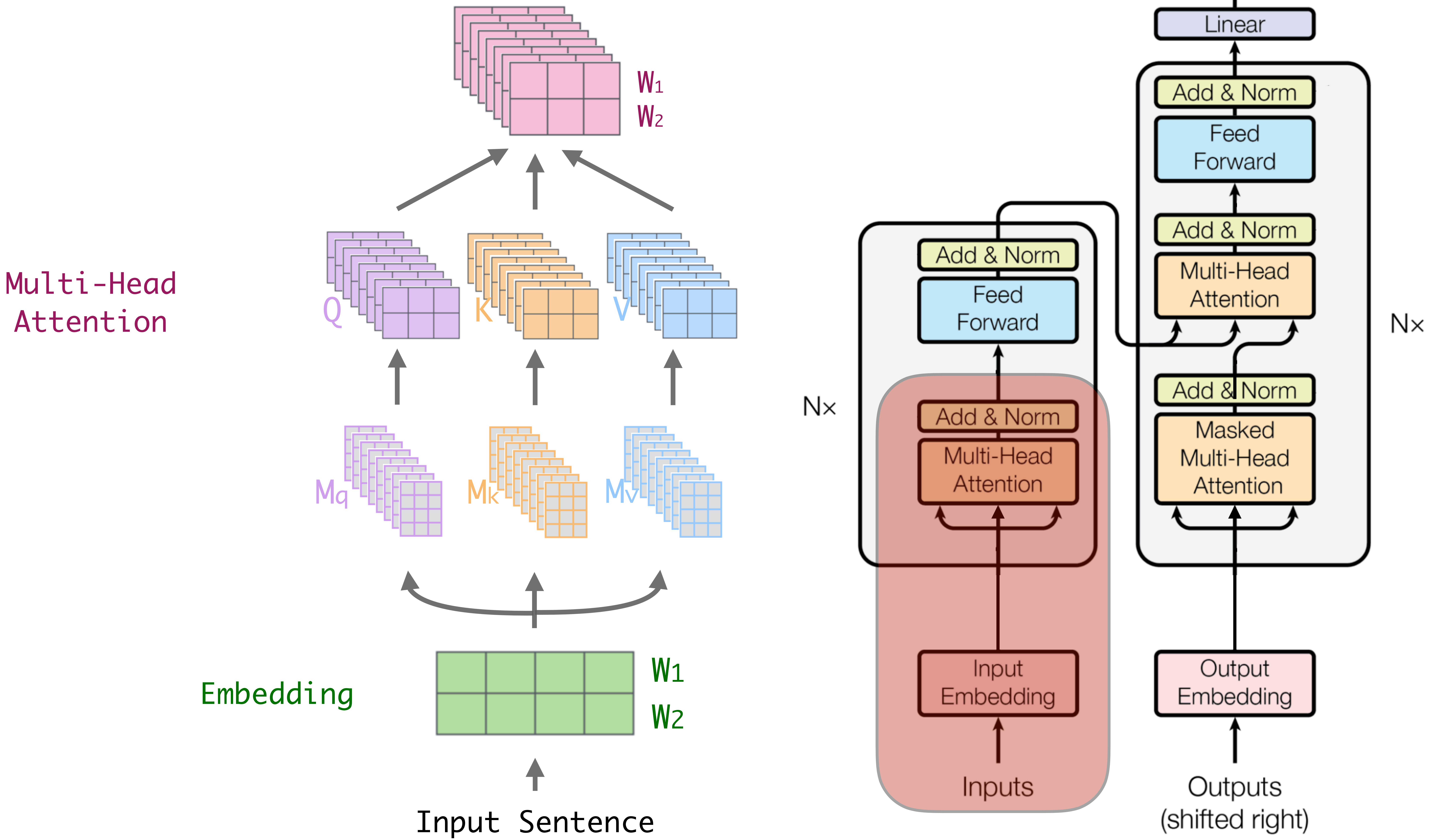


注意力

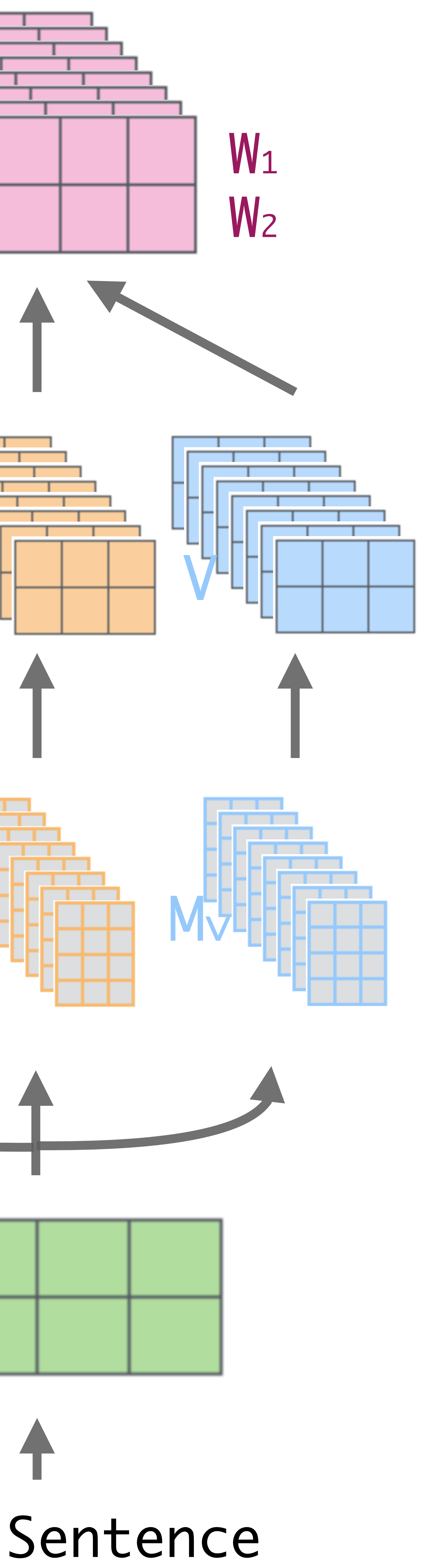
Attention

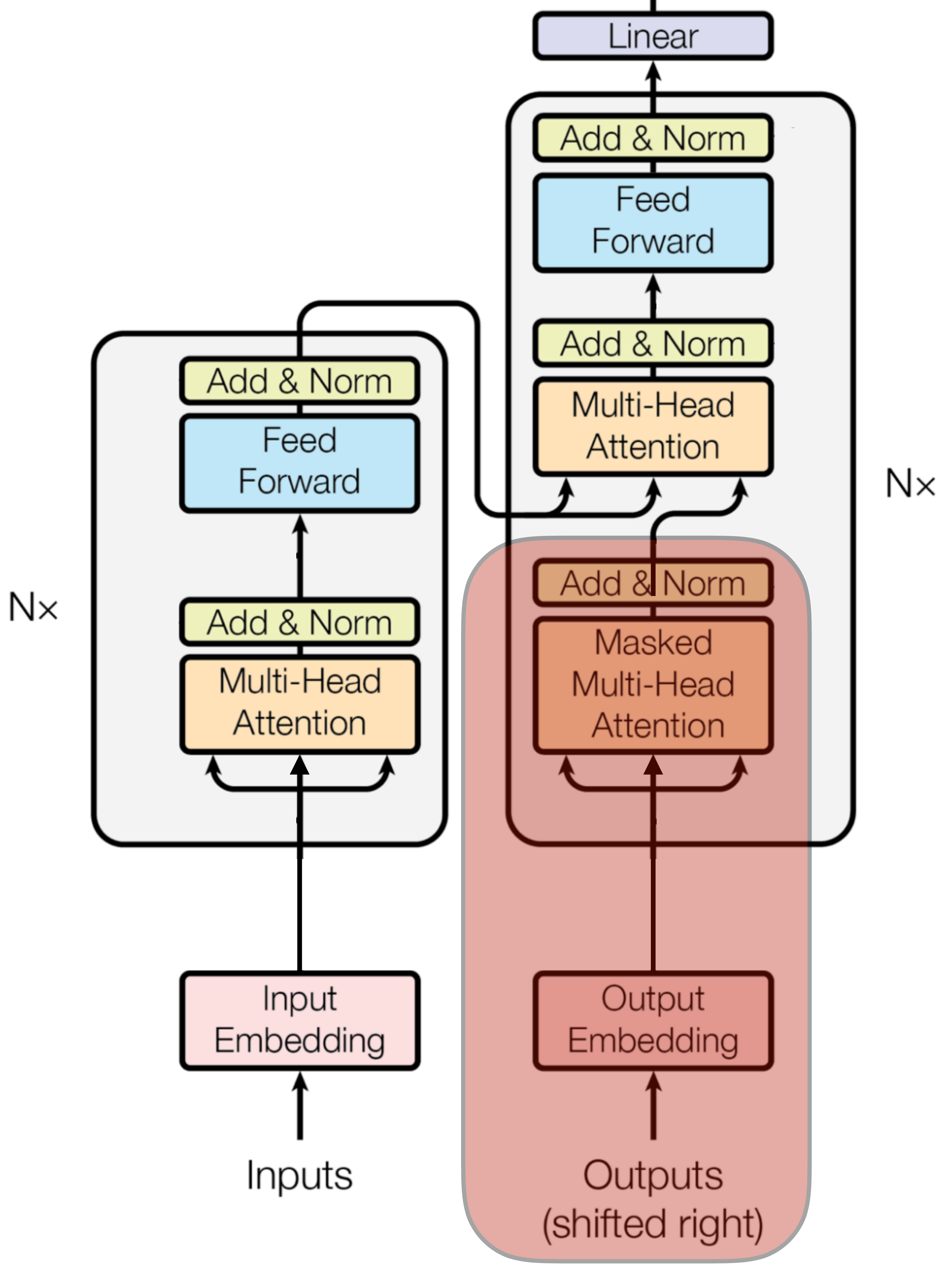




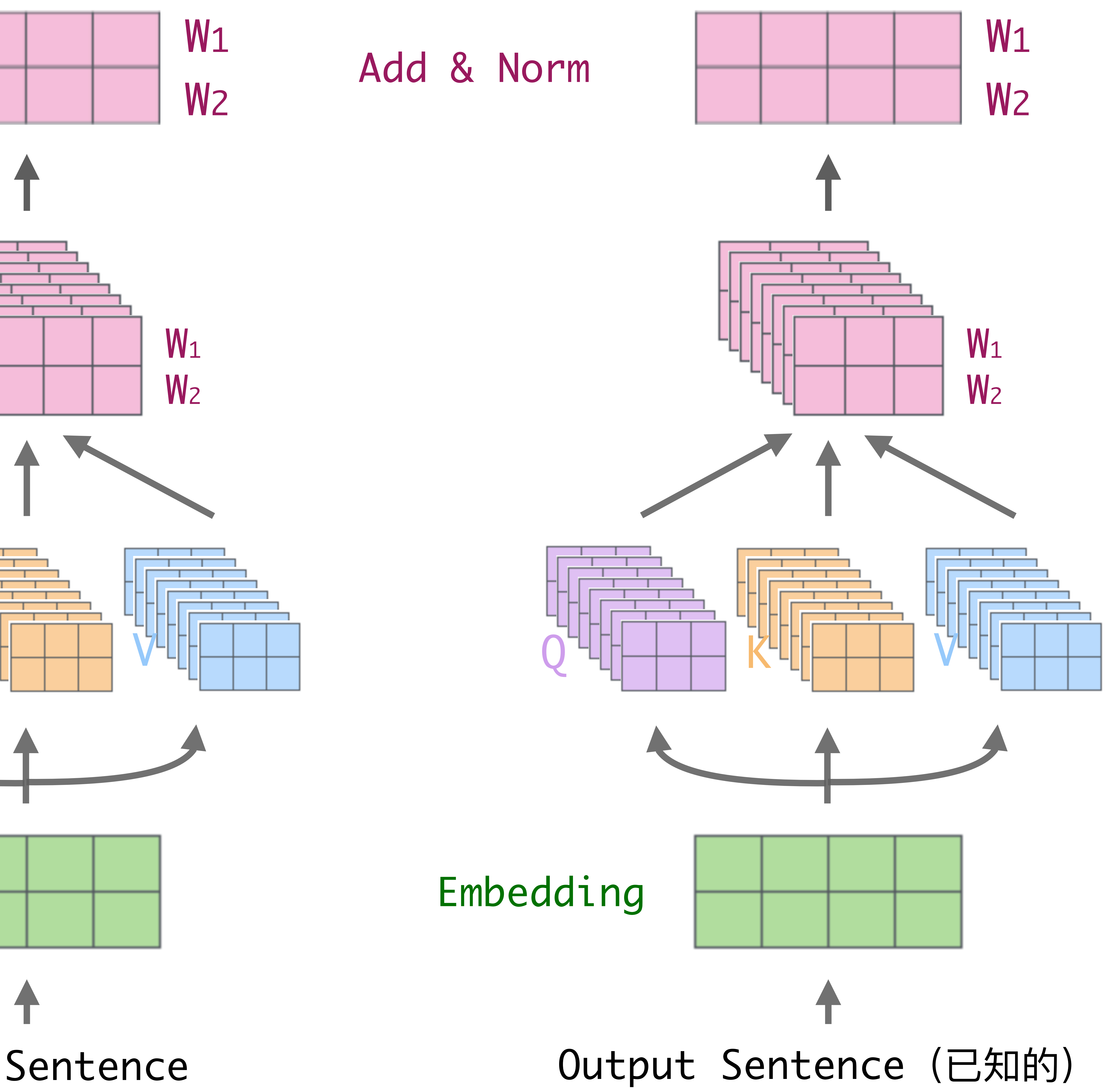


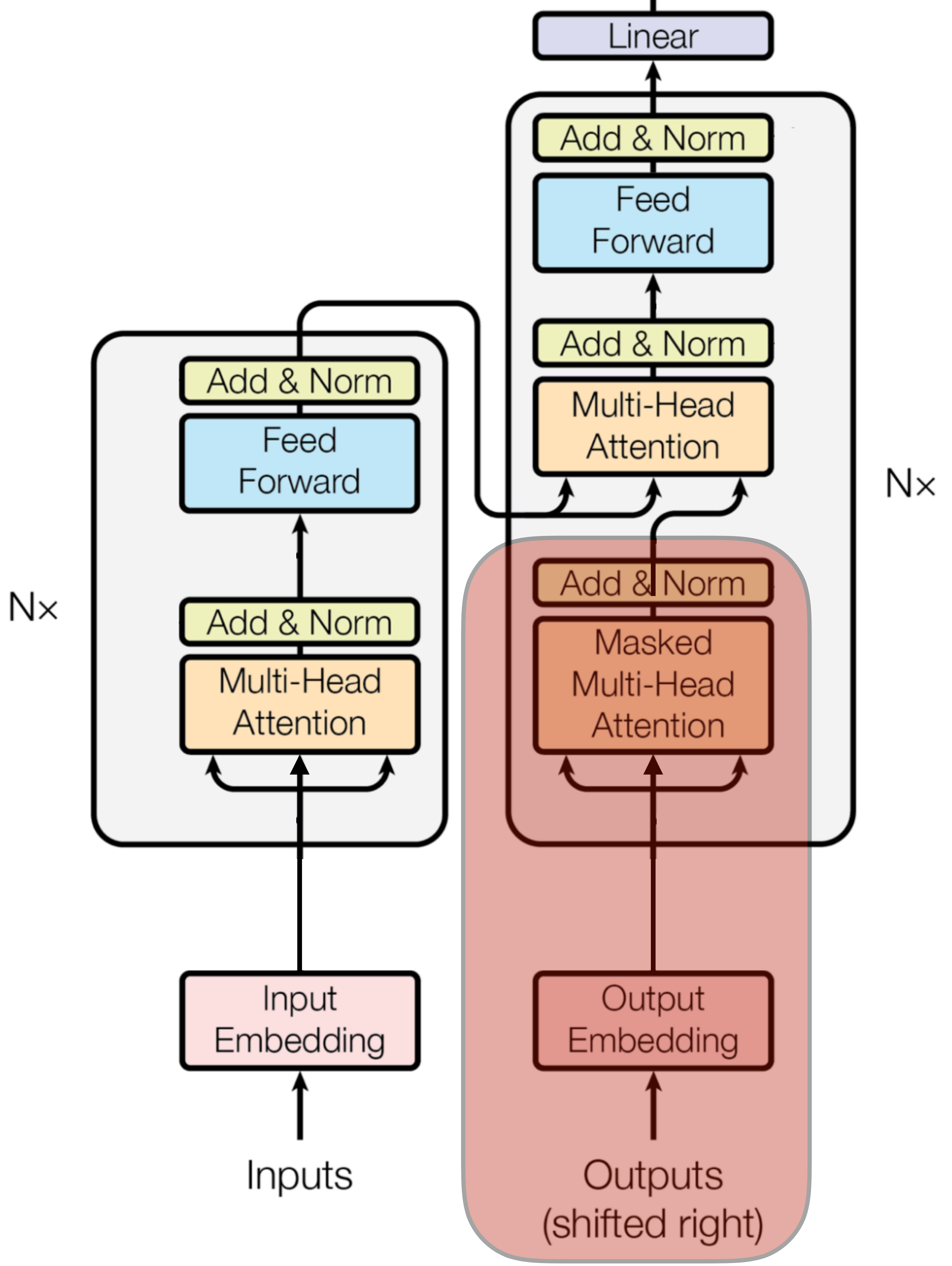
Multi-Head Attention
Embedding
Input Sentence
W1
W2
Q
K
V
Mq
Mk
Mv
W1
W2
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Output Embedding
Outputs (shifted right)
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Input Embedding
Inputs
Nx

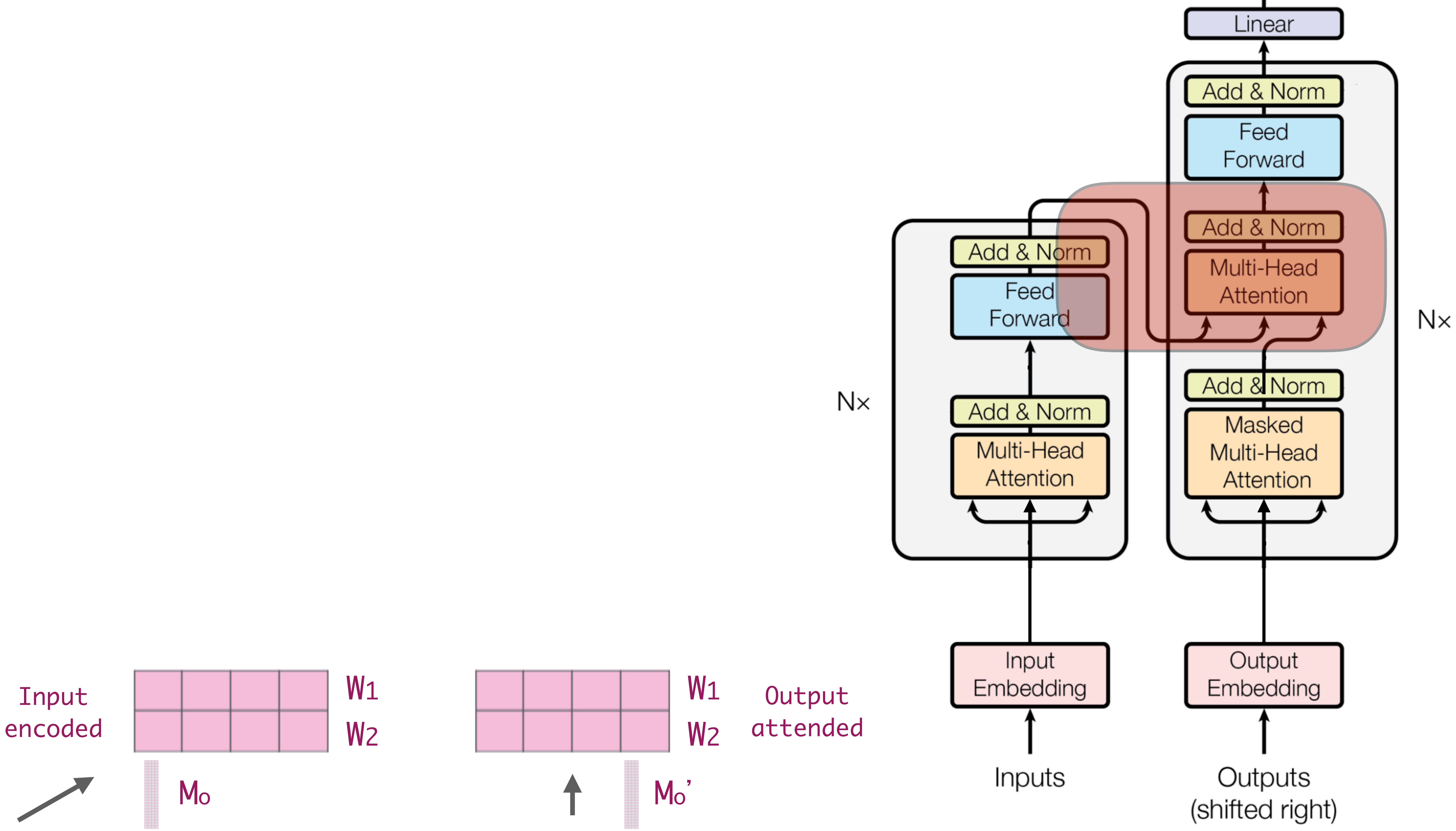

W1
W2
V
MV
Sentence


Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Add & Norm
Feed Forward
Nx
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)

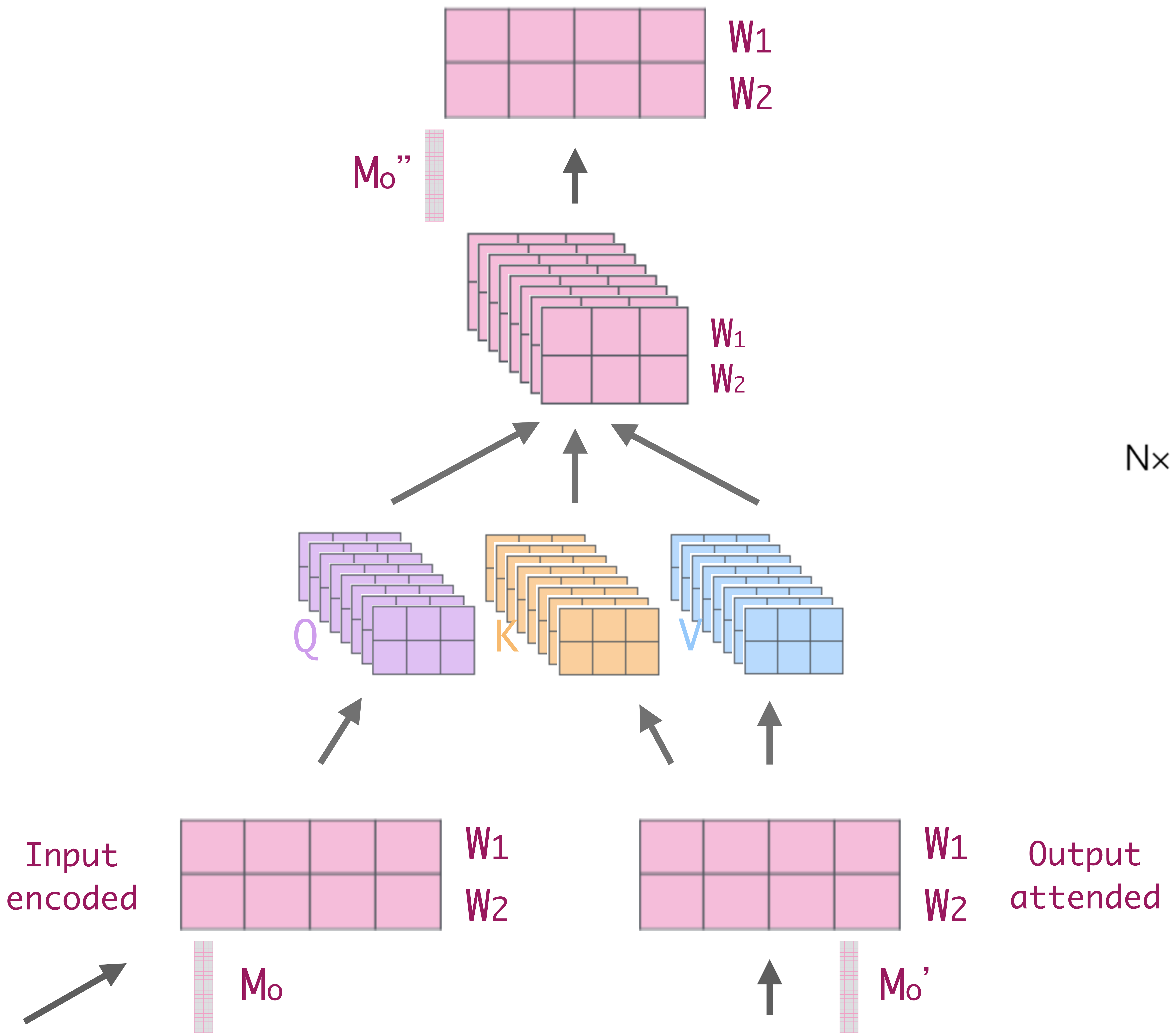
W1
W2
V
Mv
Sentence
W1
W2
Q
K
V
Mq
Mk
Mv
Embedding
Output Sentence (已知的)
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Add & Norm
Masked Multi-Head Attention
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)

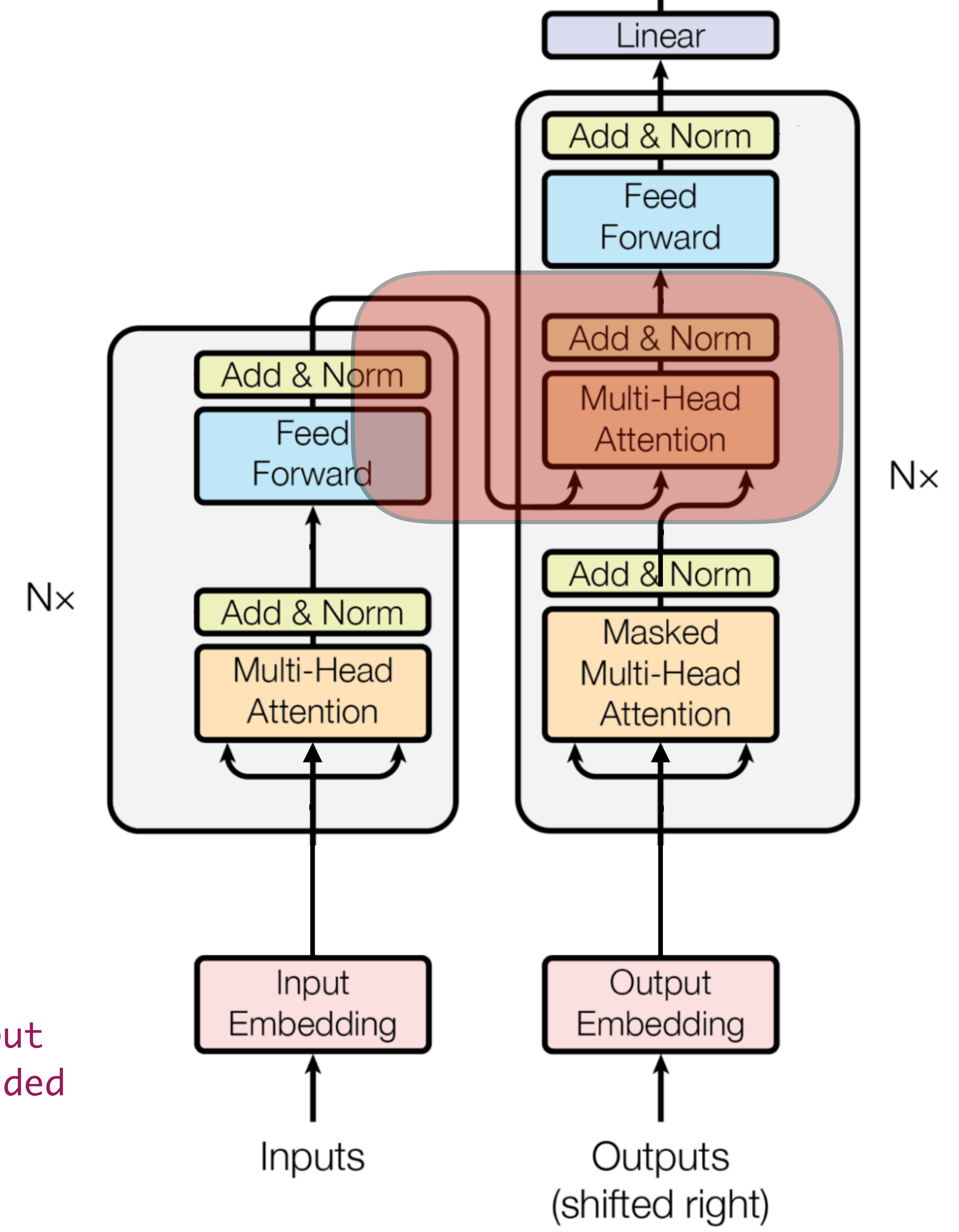

W1
W2
Add & Norm
W1
W2
W1
W2
W1
W2
V
Q
K
V
Embedding
Sentence
Output Sentence (已知的)


Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Add & Norm
Masked Multi-Head Attention
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)



Input
encoded
W1
W2
M0
W1
W2
Output
attended
M0'
Linear
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Add & Norm
Masked
Multi-Head
Attention
Nx
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Nx
Input
Embedding
Output
Embedding
Inputs
Outputs
(shifted right)

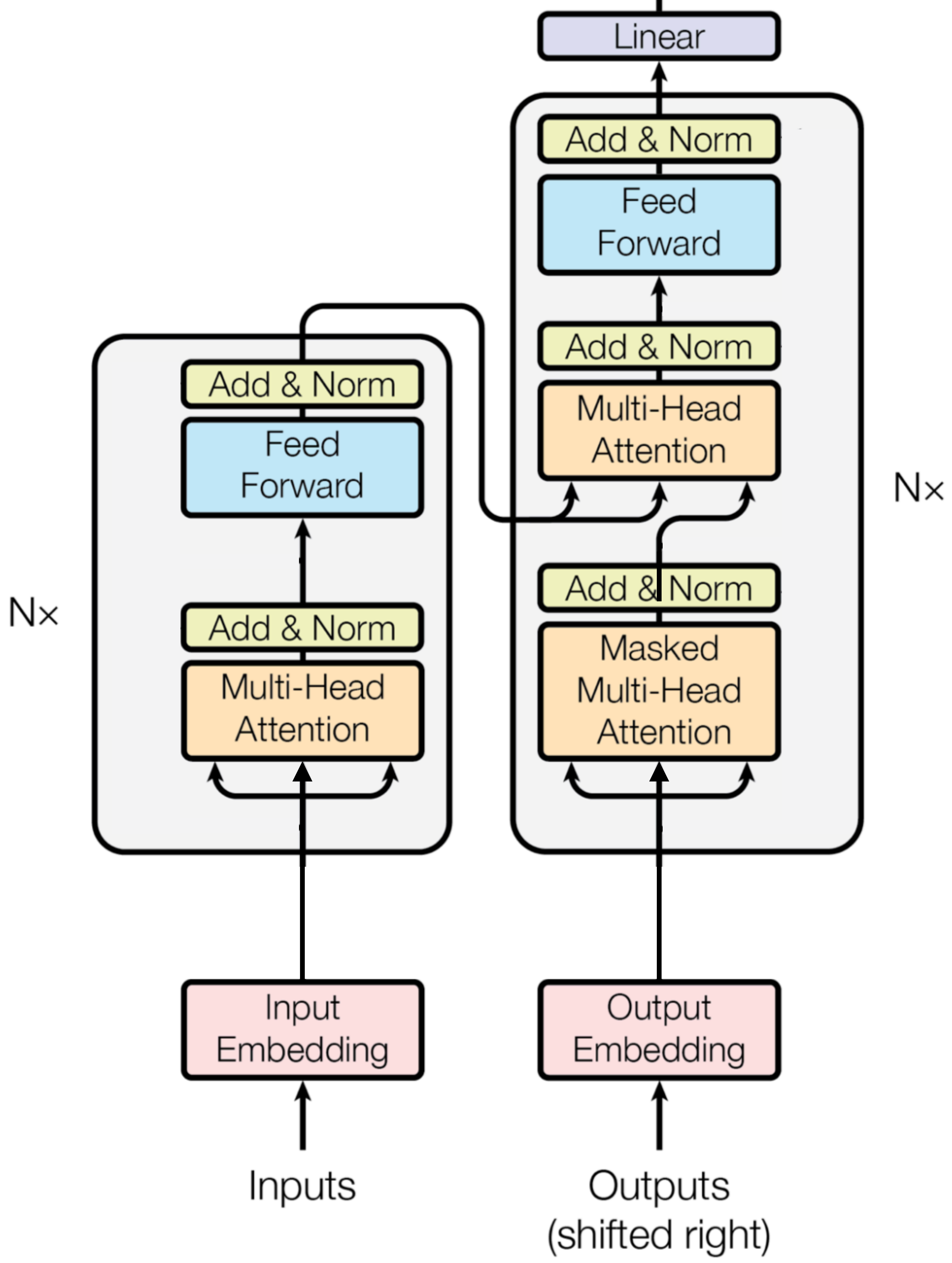

W1
W2
Mo"
W1
W2
Q
K
V
Input encoded
W1
W2
Mo
W1
W2
Output attended
Mo'


Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)


W1
W2
Q
K
V
W1
W2
W1
W2
Q
K
V
Q
K
V
Input Sentence
Output Sentence (已知的)
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Feed Forward
Nx
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)

注意力

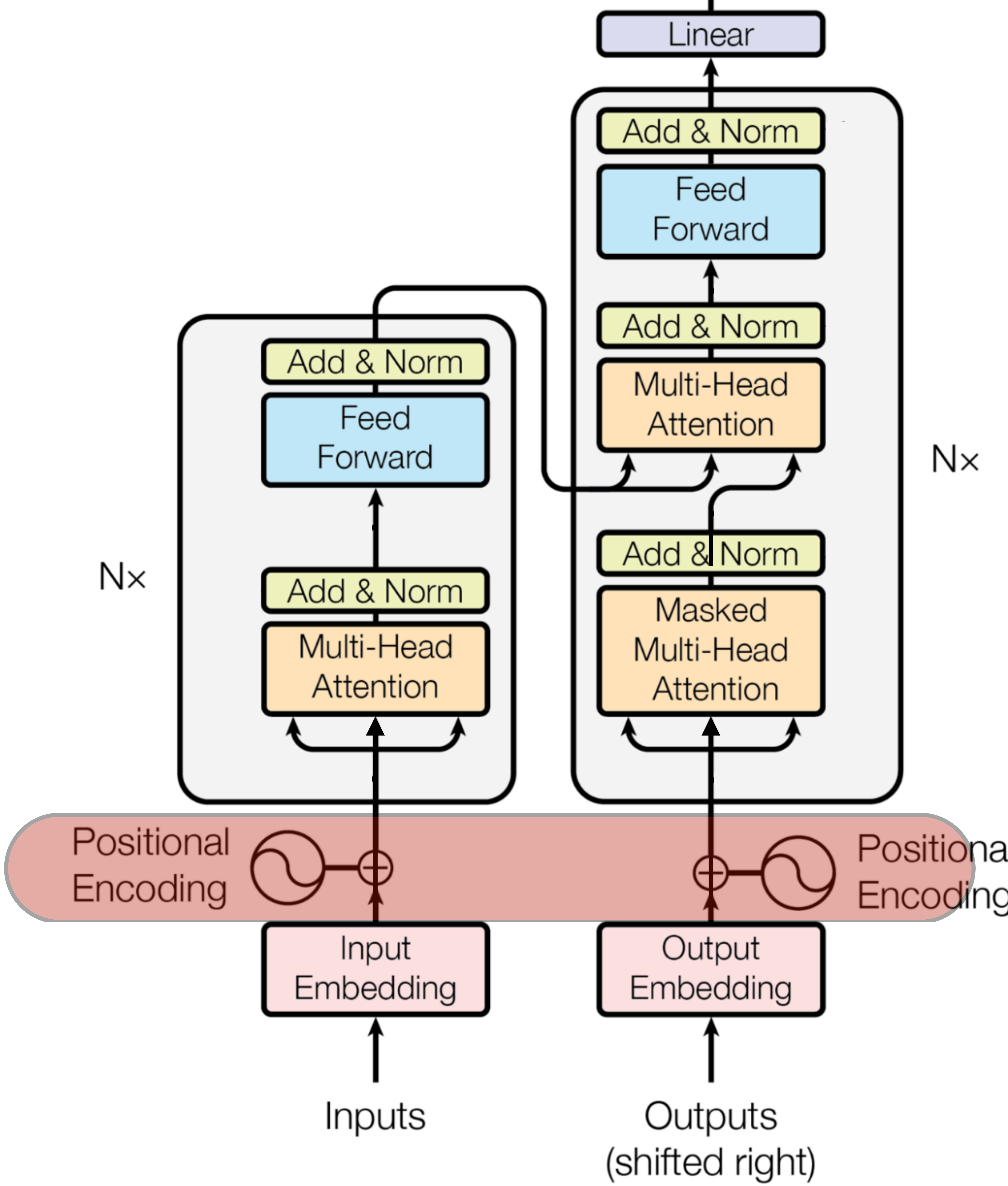
Attention



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?

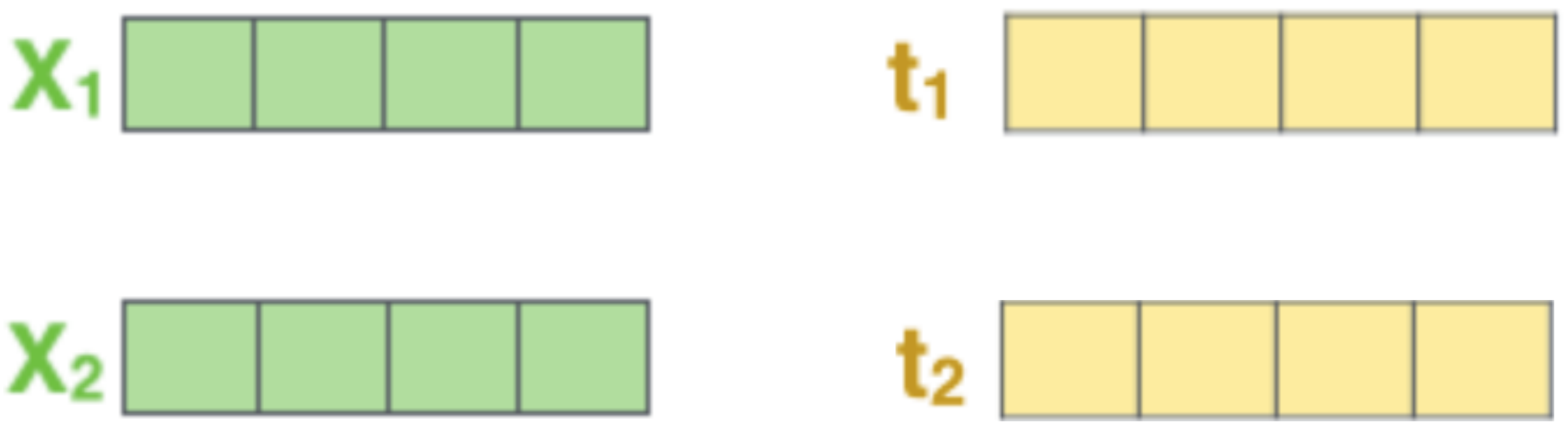
位置编码

Positional encoding

位置编码长什么样?

word1

word2

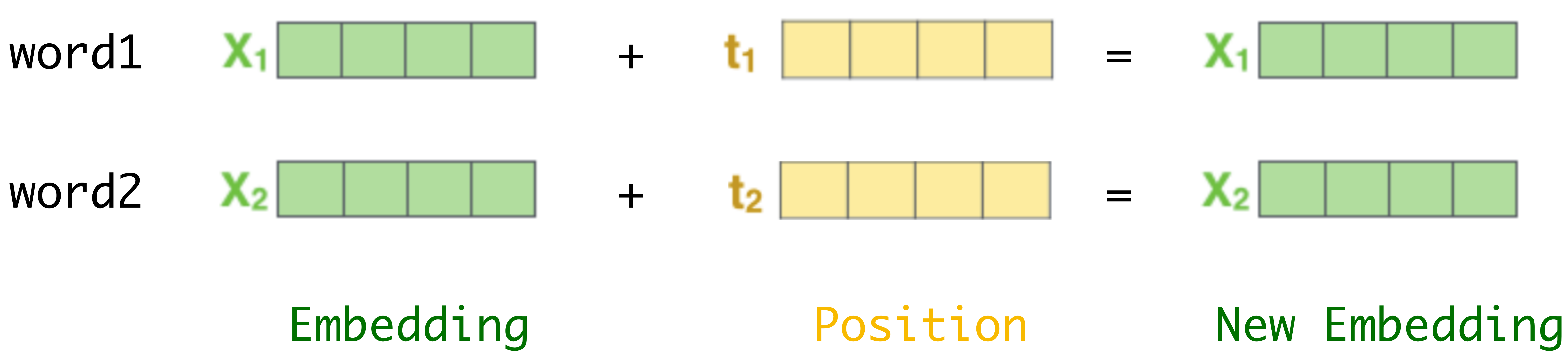


Embedding　　Position

位置编码

Positional encoding

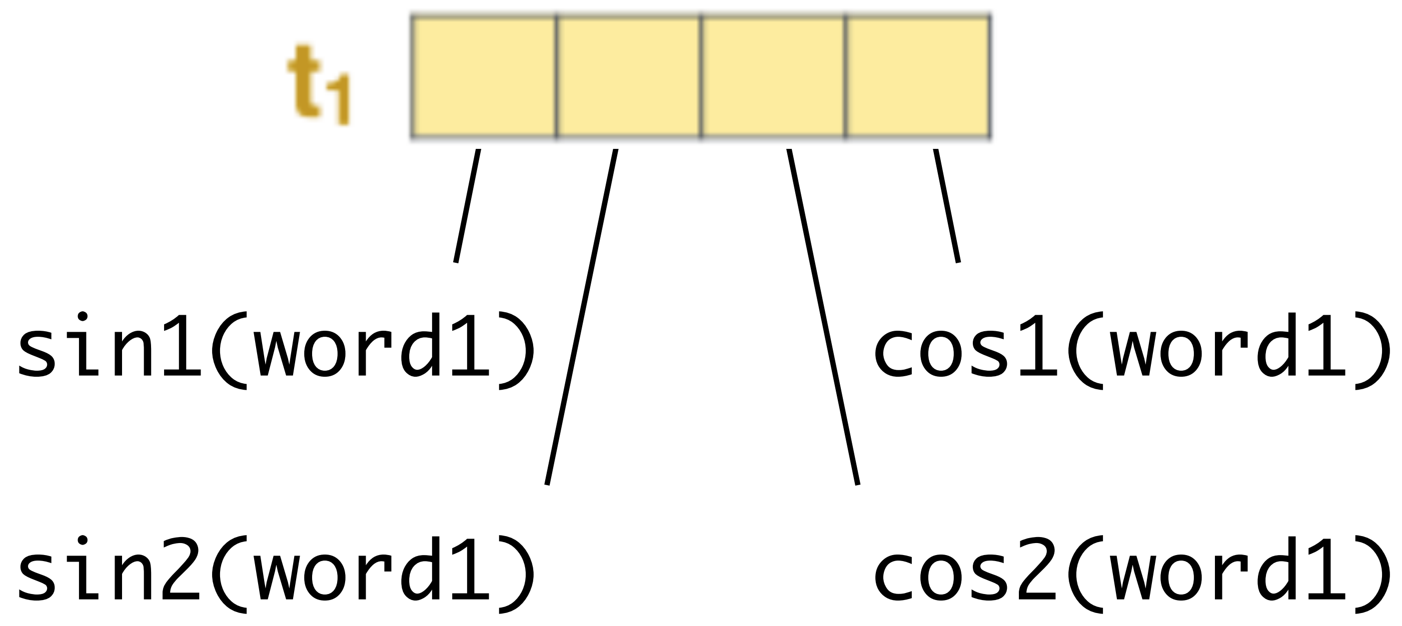
位置编码长什么样?



位置编码

Positional encoding

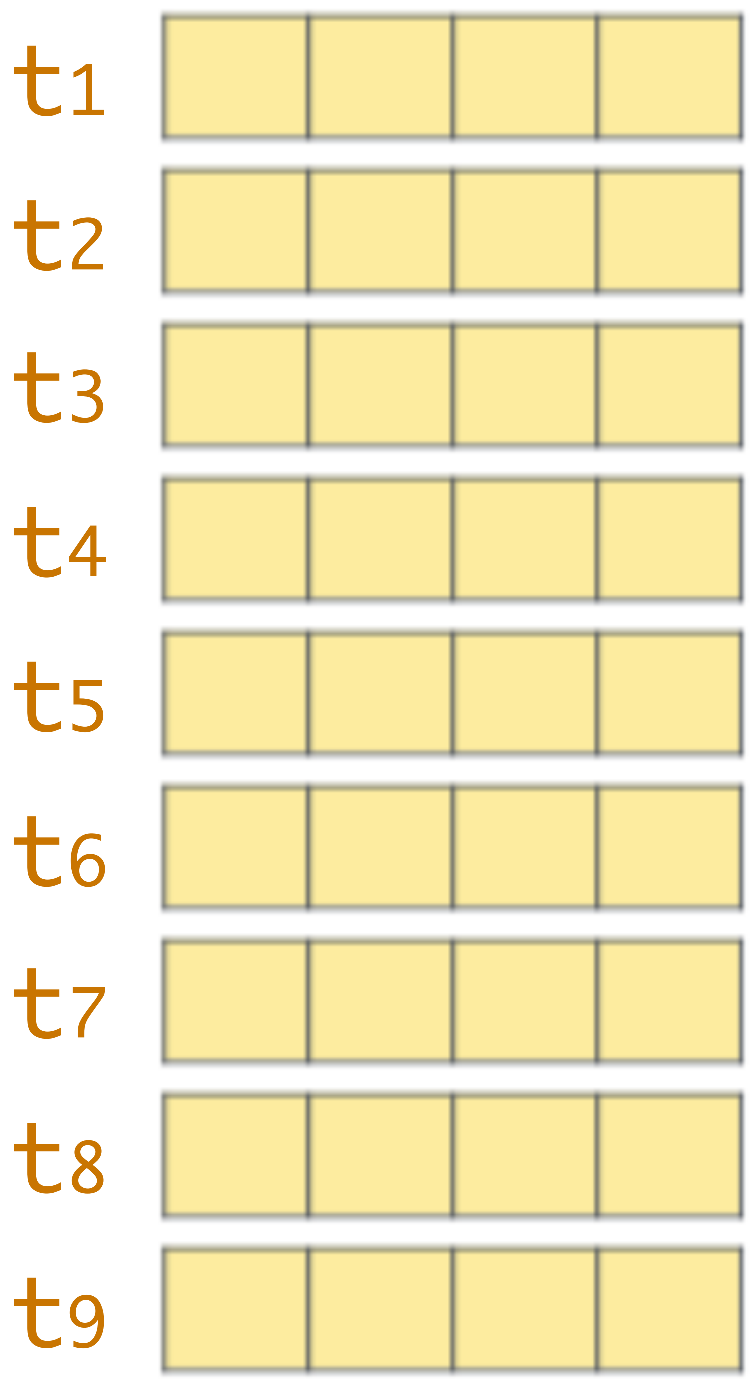
位置编码长什么样?



位置编码

Positional encoding

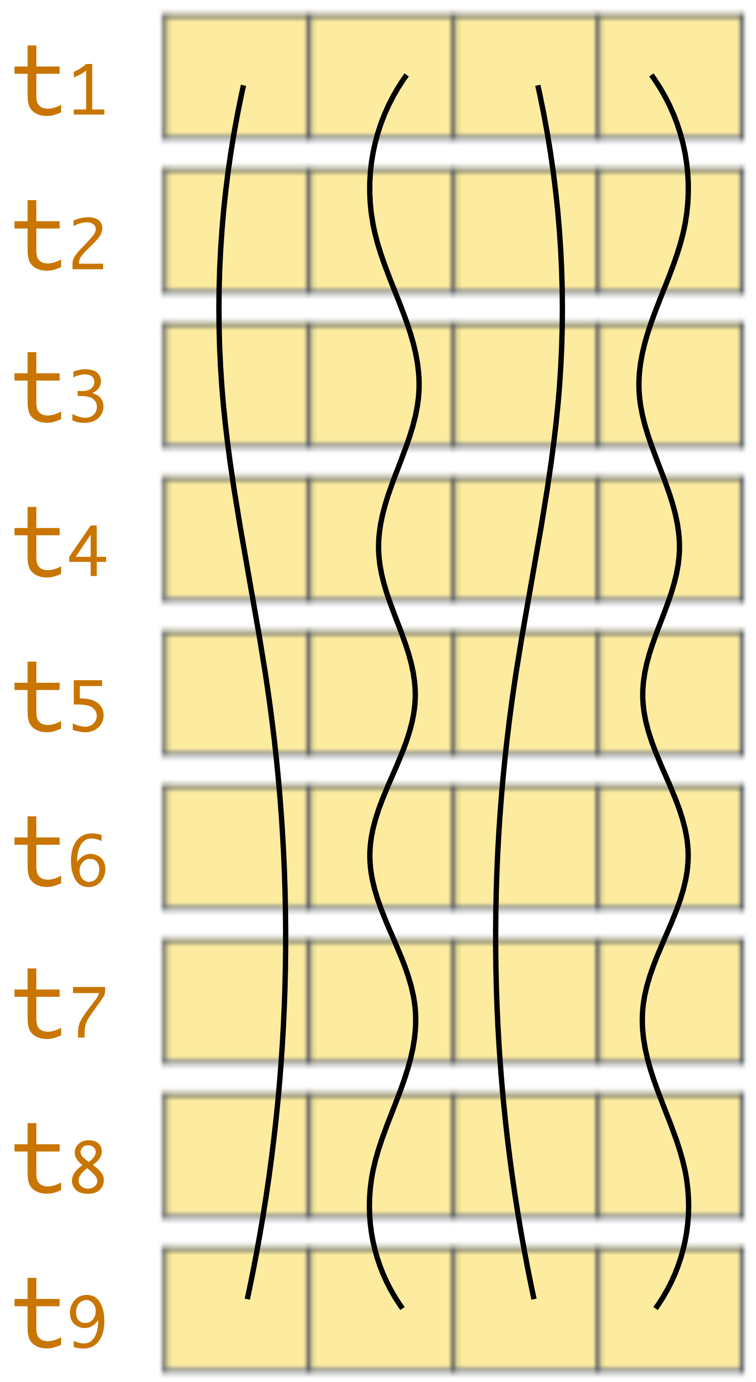
位置编码长什么样?



位置编码

Positional encoding

位置编码长什么样?



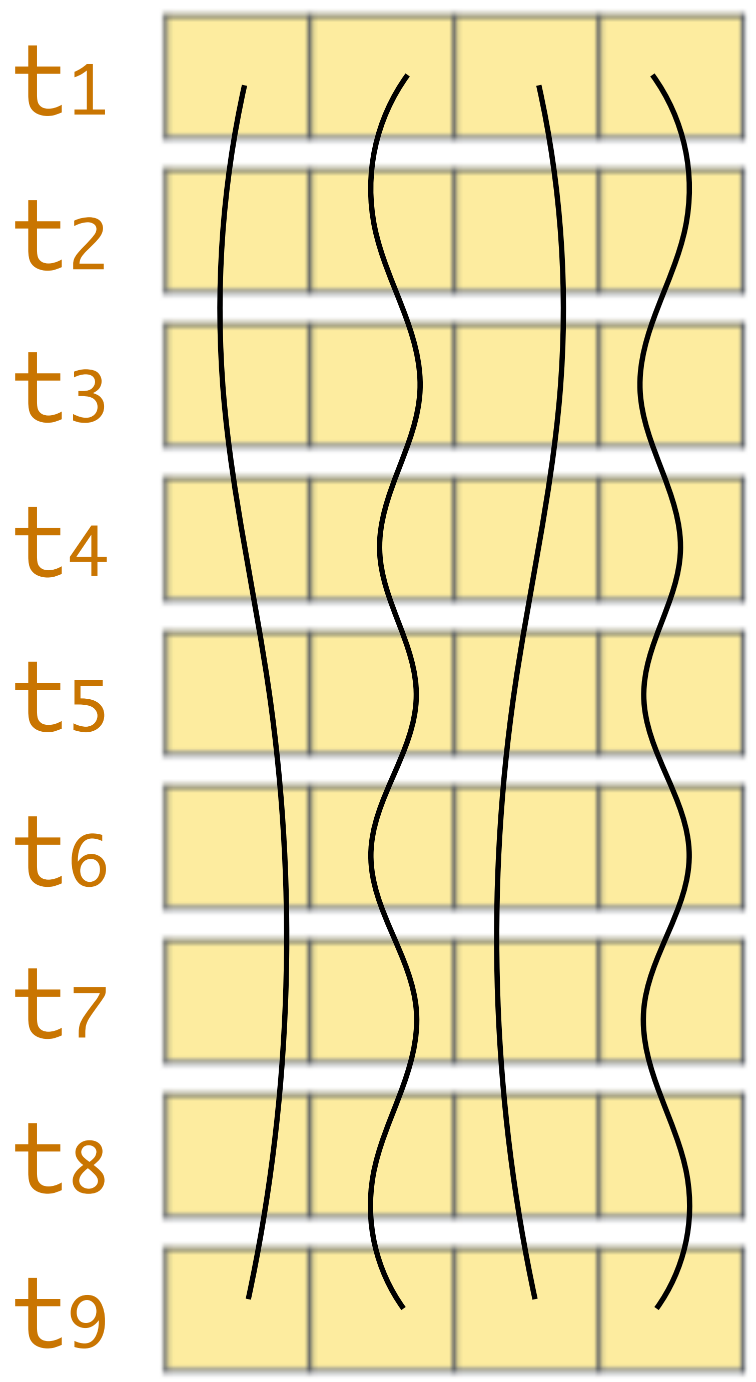
位置编码

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

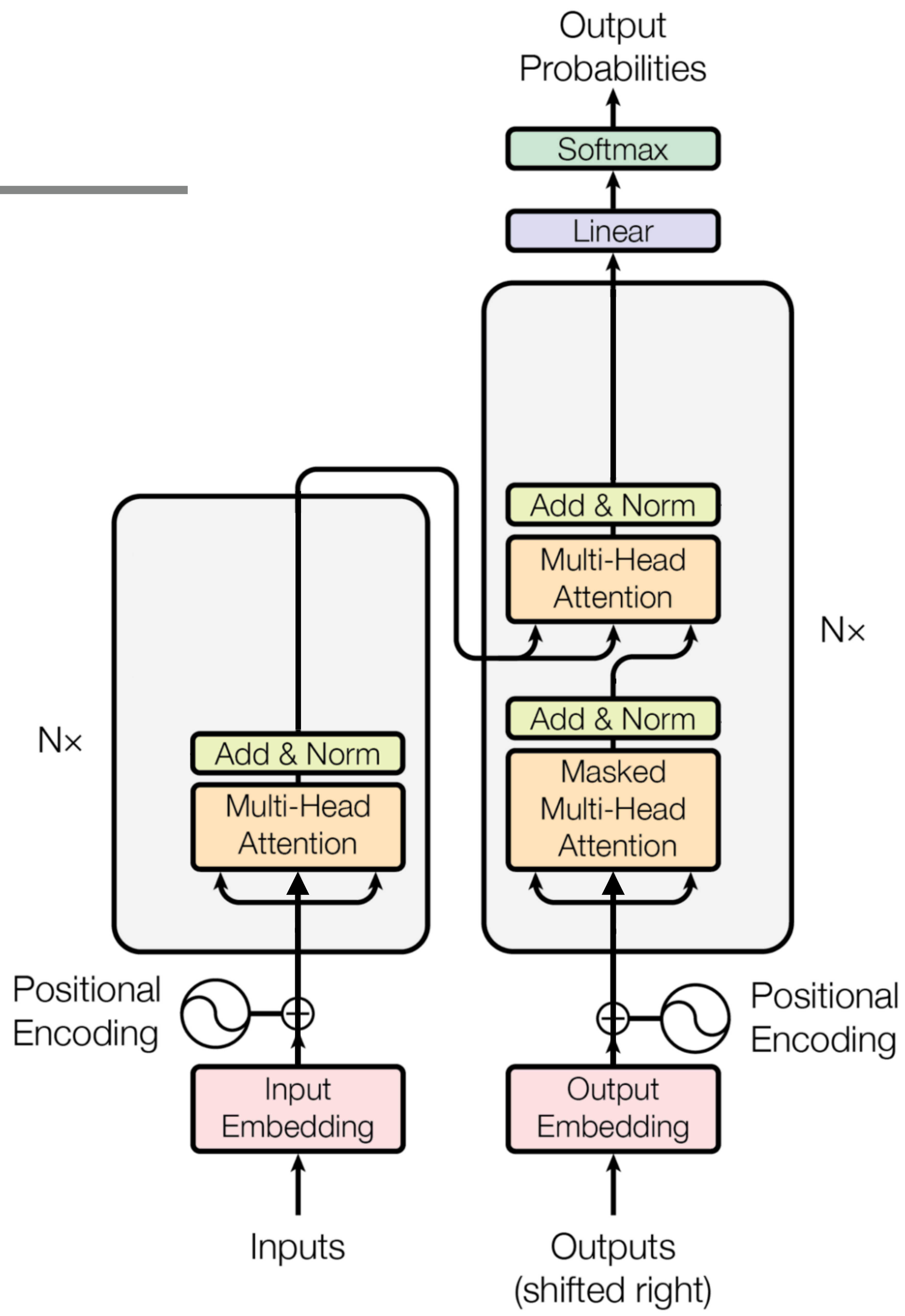
Positional encoding

为什么需要单独对token的位置进行编码?

位置编码长什么样?

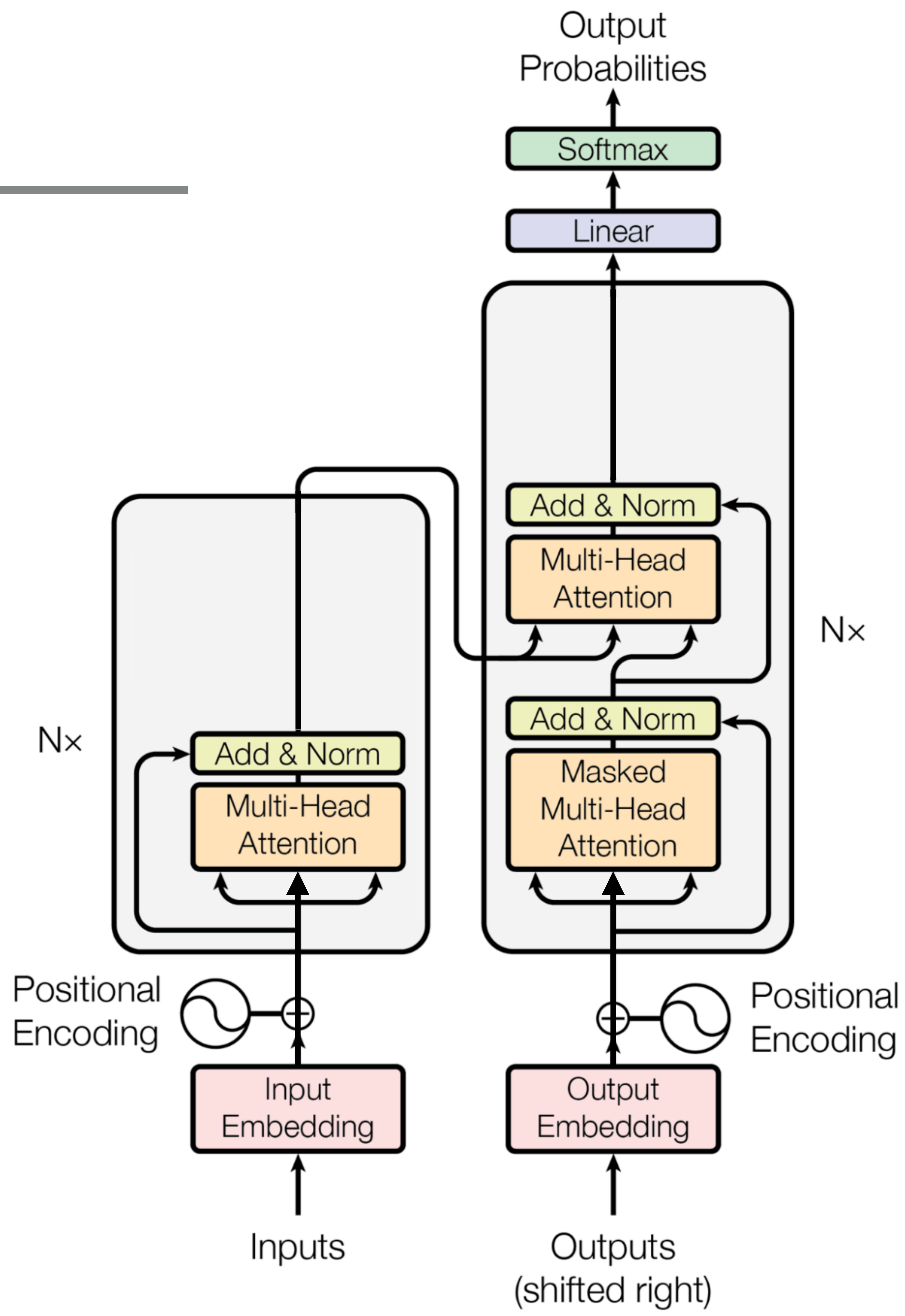
为什么要长这样?

位置编码



Output
Probabilities
Softmax
Linear
Add & Norm
Multi-Head
Attention
Nx
Add & Norm
Masked
Multi-Head
Attention
Nx
Add & Norm
Multi-Head
Attention
Positional
Encoding
Positional
Encoding
Input
Embedding
Output
Embedding
Inputs
Outputs
(shifted right)

位置编码



Output
Probabilities
Softmax
Linear
Add & Norm
Multi-Head
Attention
Nx
Add & Norm
Masked
Multi-Head
Attention
Nx
Add & Norm
Multi-Head
Attention
Positional
Encoding
Positional
Encoding
Input
Embedding
Output
Embedding
Inputs
Outputs
(shifted right)

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

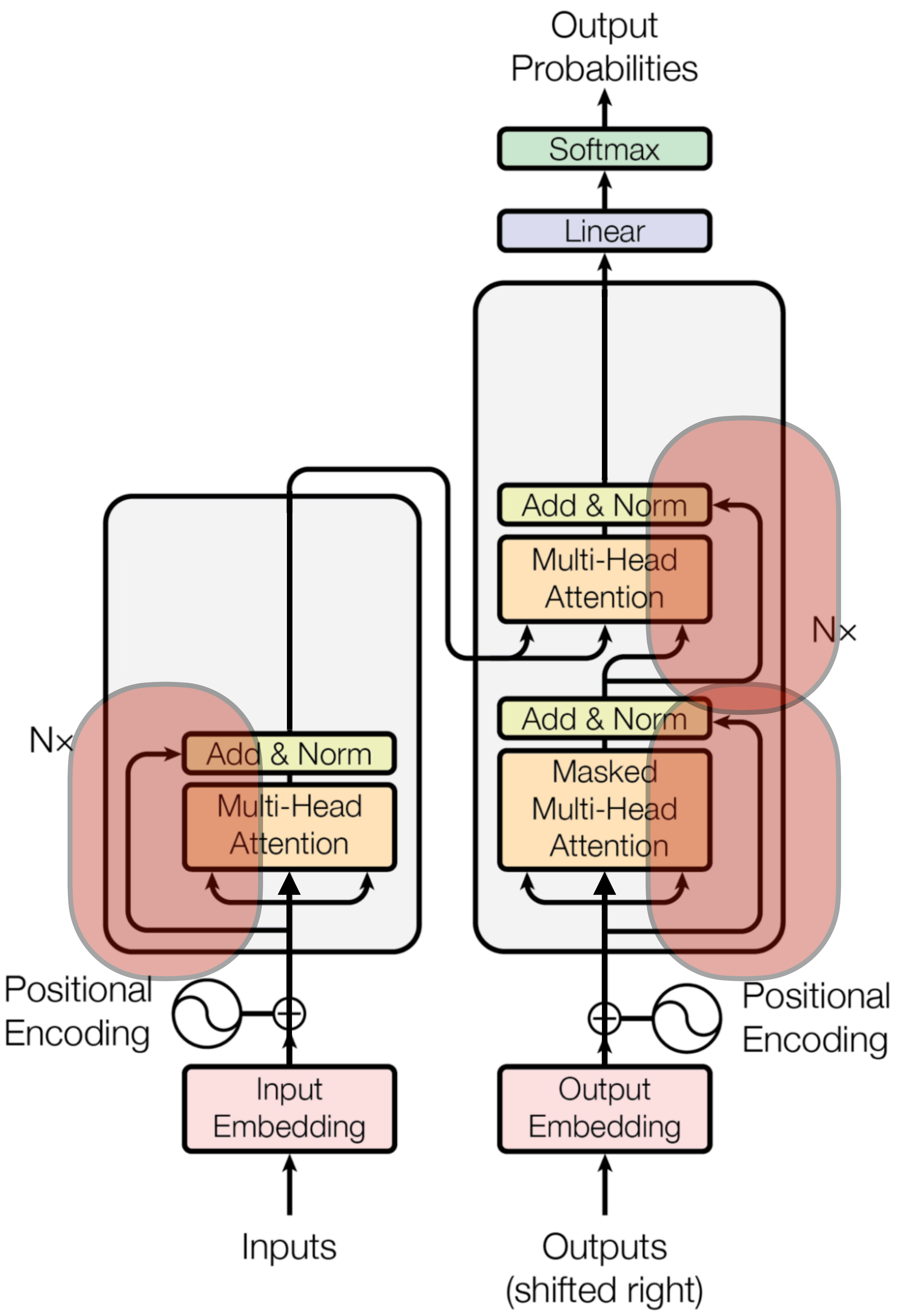
Residual connection

怎么做残差连接?

shortcut

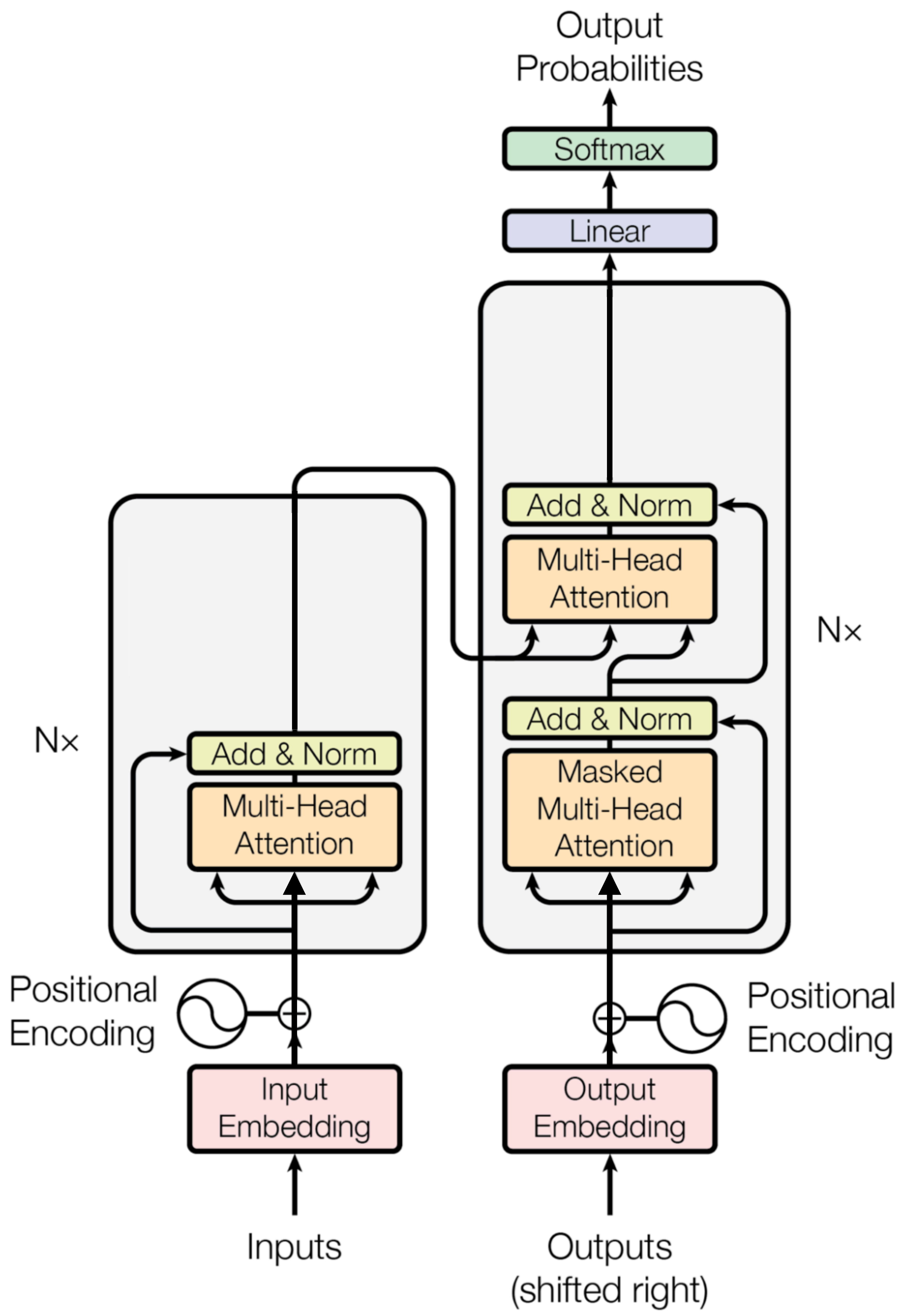
$$H(x) = F(x) + x$$

训练使F(x)趋近于0



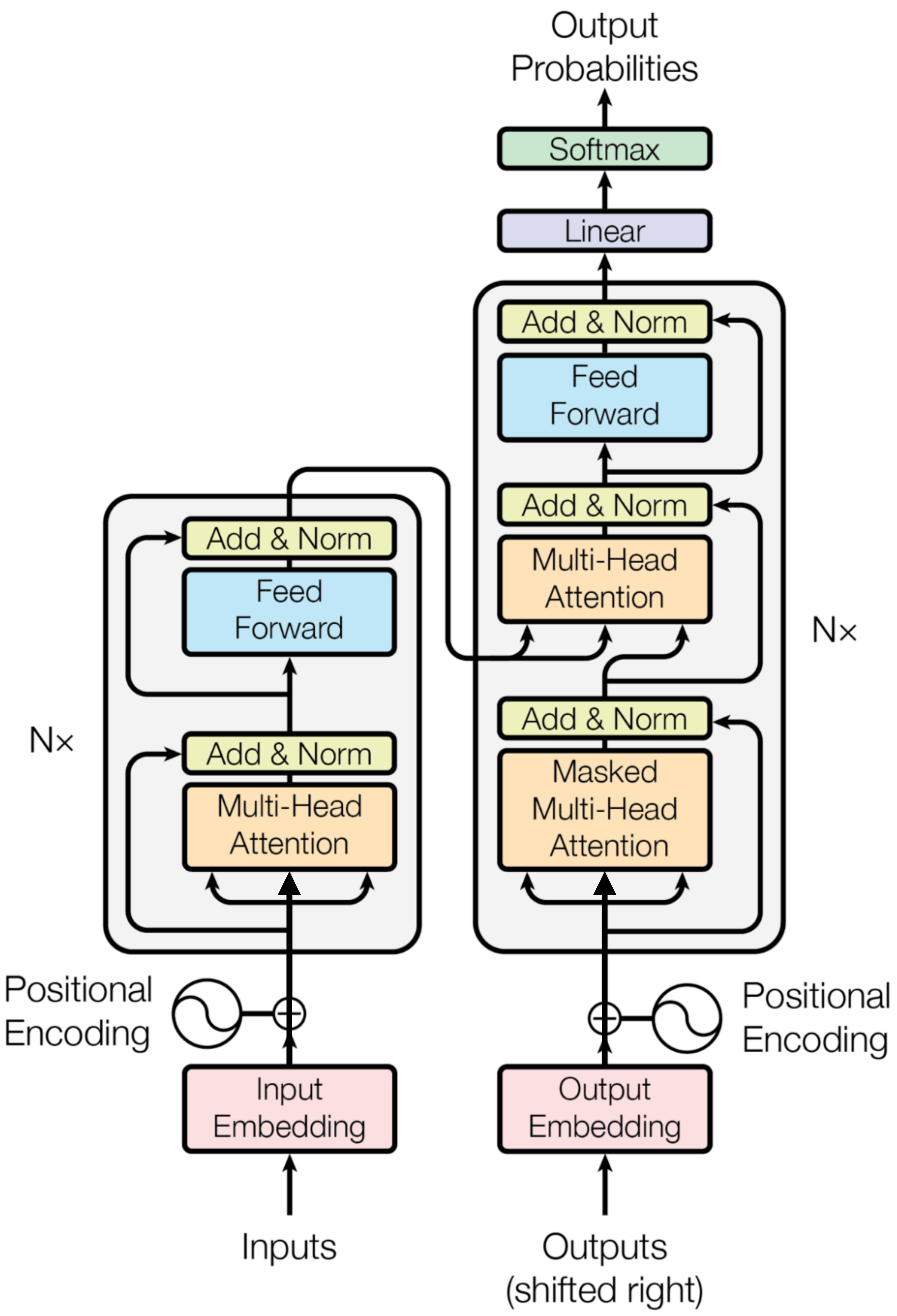
残差连接

Residual connection



残差连接

Residual connection



简单分析一下

Simple analysis

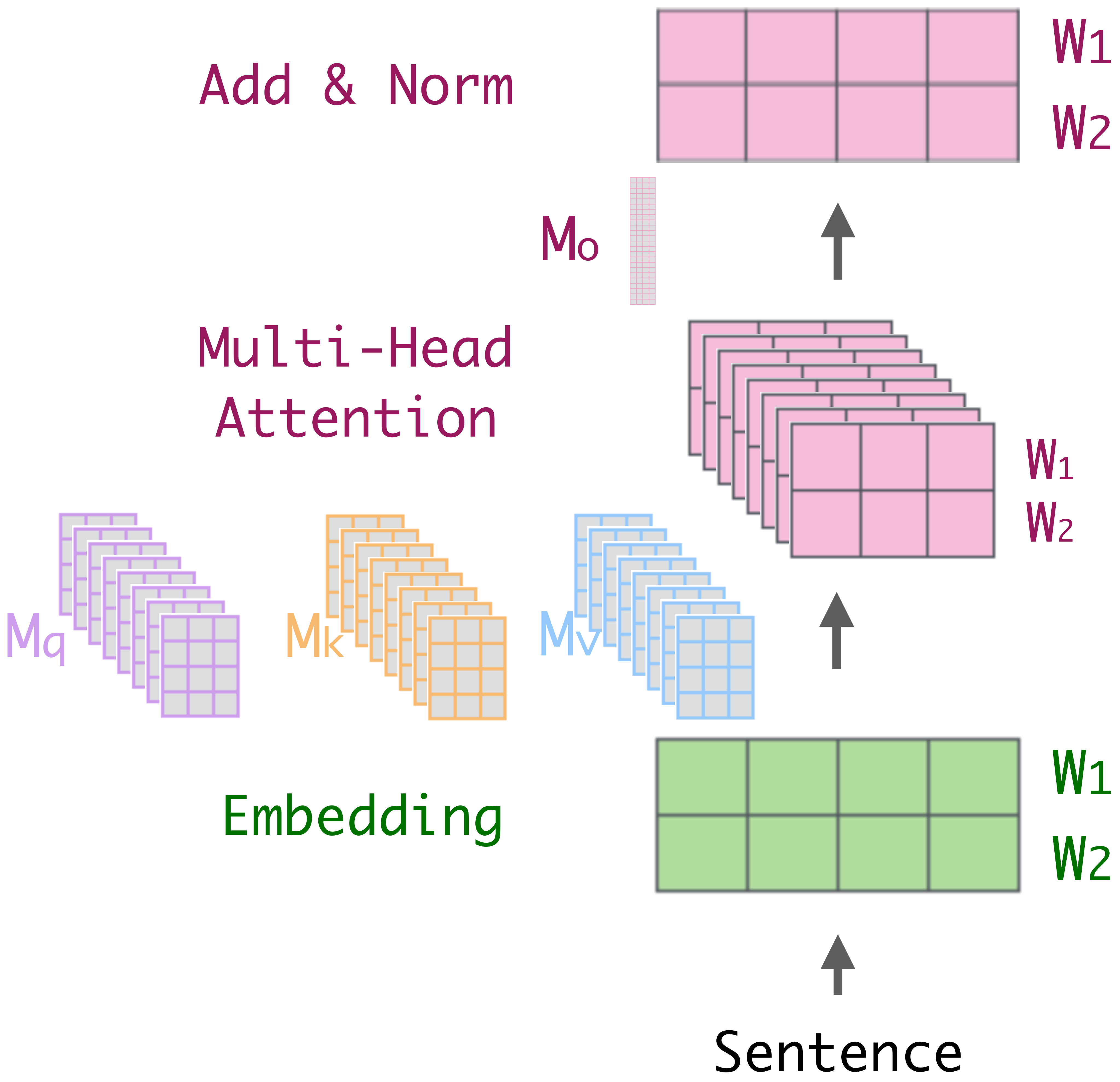
简单分析一下

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

Ford.