TensorFlow

PYTORCH

Self-Attention

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

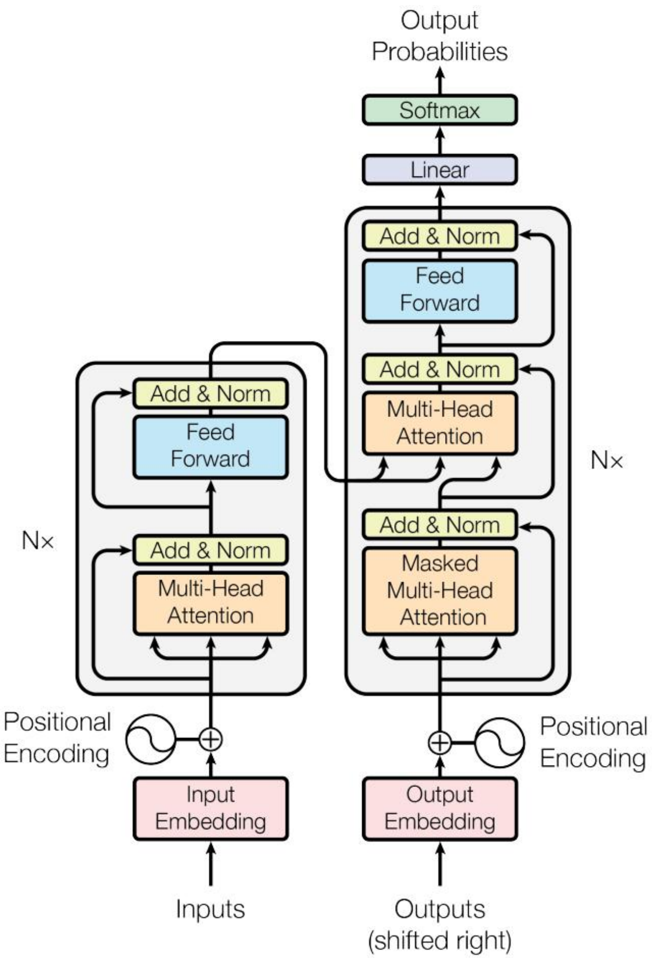
Llion Jones* Google Research llion@google.com

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

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Computation and Language

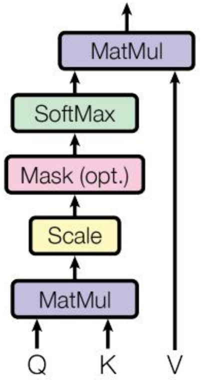


原文链接：https://arxiv.org/abs/1706.03762

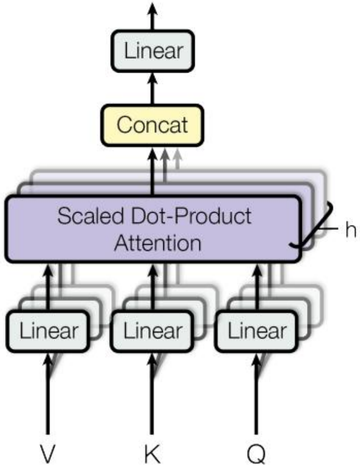
推荐博文：https://blog.csdn.net/qq_37541097/article/details/117691873

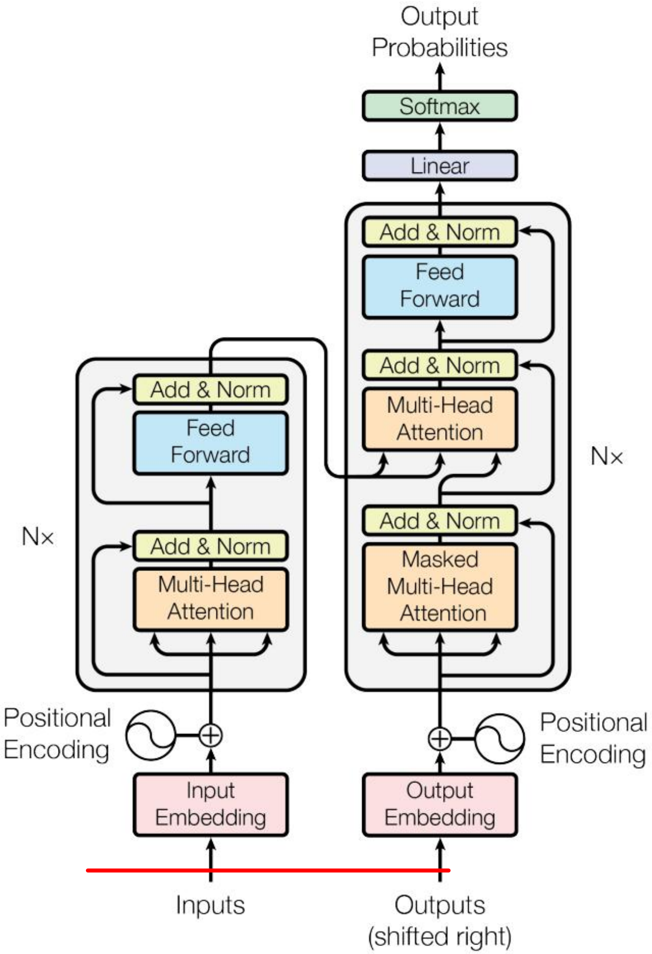
Self-Attention

Scaled Dot-Product Attention



Multi-Head Attention





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

$$\mathrm{MultiHead}(Q,K,V) = \mathrm{Concat}(\mathrm{head_1},\ldots,\mathrm{head_h})W^O$$
$$\text{where } \mathrm{head_i} = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

Self-Attention

To facilitate these residual connections, all sub-layers in the model, as well as the embeddinglayers, produce outputs of dimension $d_{model} = 512$

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

q: query (to match others)

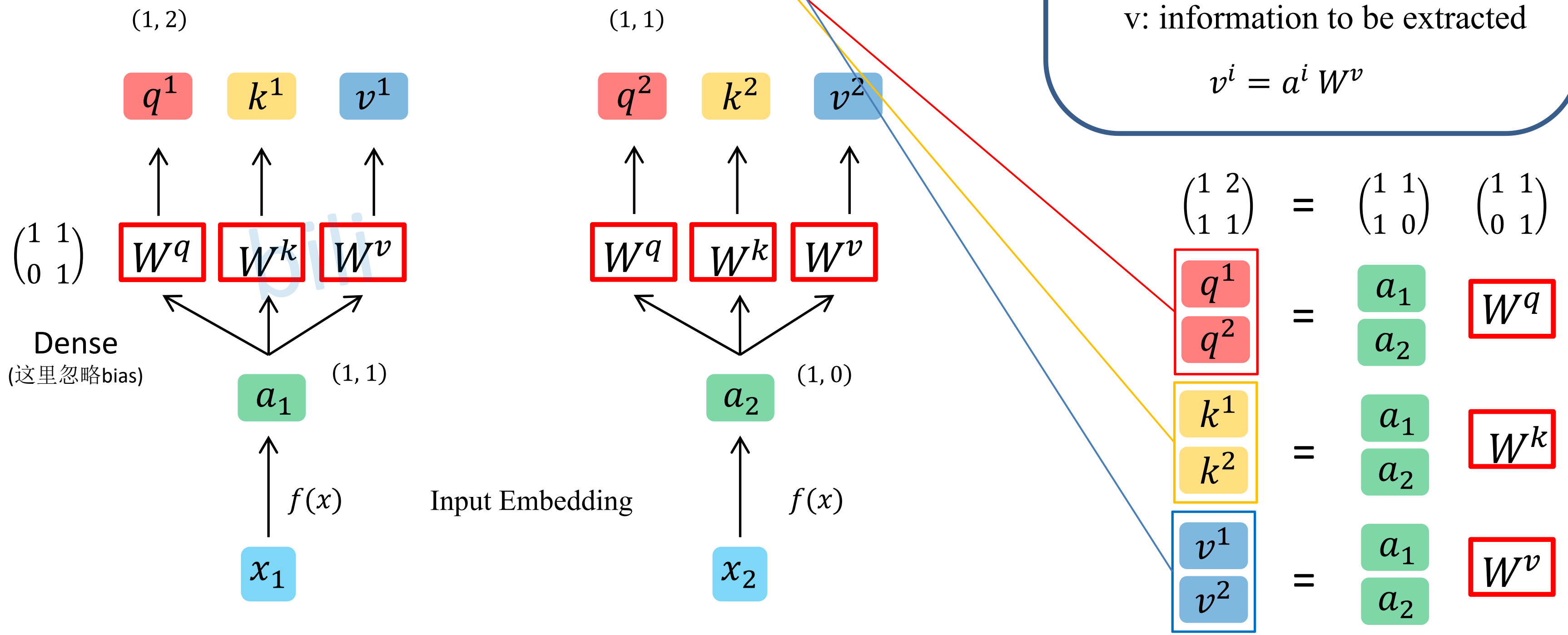
$$q^i = a^i W^q$$

k: key (to be matched)

$$k^i = a^i W^k$$

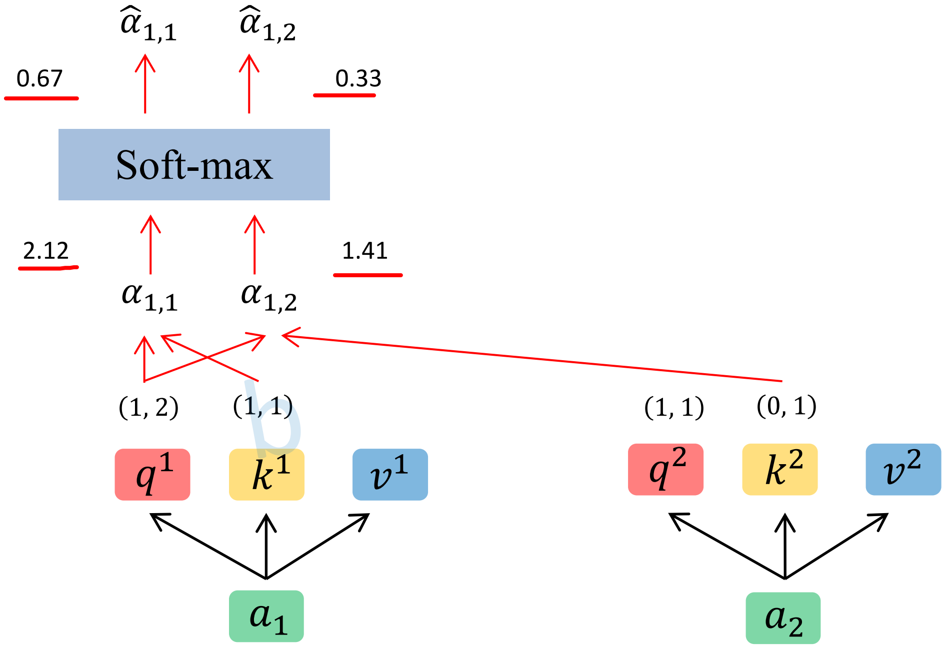
v: information to be extracted

$$v^i = a^i W^v$$



Self-Attention

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



Scaled Dot-Product Attension:

$$\alpha_{1,i} = q^1 \cdot k^i / \sqrt{d}$$

$$\alpha_{2,i} = q^2 \cdot k^i / \sqrt{d}$$

(d is the dim of k)

…….

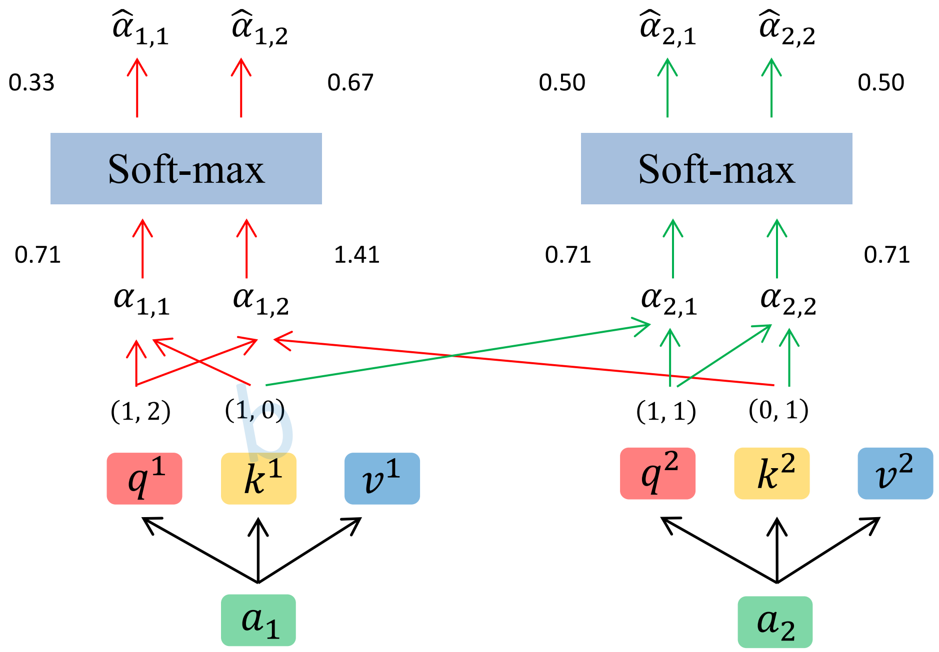
在论文中的解释是“进行点乘后的数值很大，导致通过softmax后梯度变的很小”

Soft-max

$$\hat{\alpha}_{1,i} = \frac{e^{\alpha_{1,i}}}{\sum_j e^{\alpha_{1,j}}}$$

Self-Attention

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



Scaled Dot-Product Attension:

$$\alpha_{1,i} = q^1 \cdot k^i / \sqrt{d}$$

$$\alpha_{2,i} = q^2 \cdot k^i / \sqrt{d}$$

(d is the dim of k)

....

$$\begin{pmatrix}0.71 & 1.41\\ 0.71 & 0.71\end{pmatrix} = \begin{pmatrix}1 & 2\\ 1 & 1\end{pmatrix}\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix} / 1.41$$

$$\begin{pmatrix}\alpha_{1,1} & \alpha_{1,2}\\ \alpha_{2,1} & \alpha_{2,2}\end{pmatrix} = \begin{matrix}q^1\\ q^2\end{matrix} \; k^1 \; k^2 \; /\sqrt{d}$$

Soft-max（行）

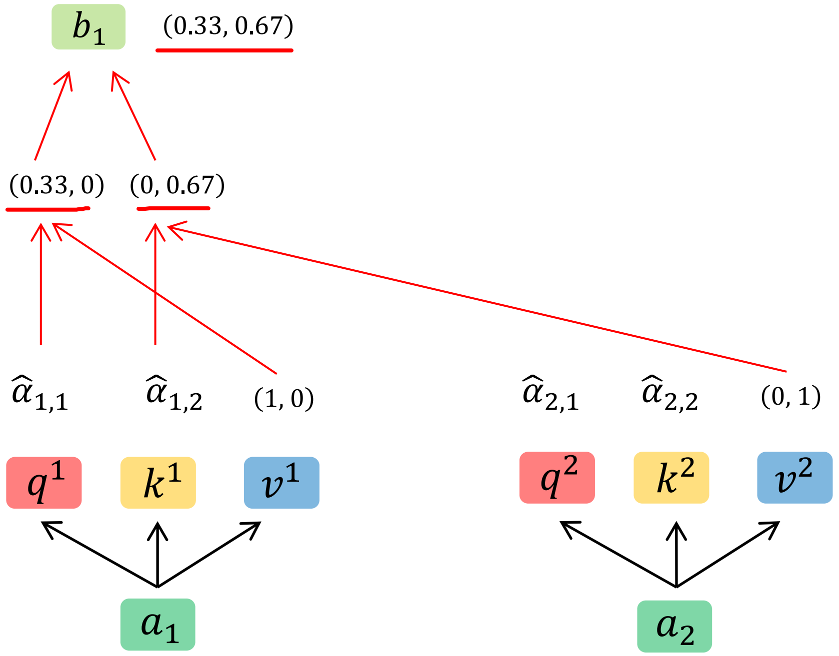
$$\begin{pmatrix}\hat{\alpha}_{1,1} & \hat{\alpha}_{1,2}\\ \hat{\alpha}_{2,1} & \hat{\alpha}_{2,2}\end{pmatrix} \quad \begin{pmatrix}0.33 & 0.67\\ 0.50 & 0.50\end{pmatrix}$$

Self-Attention

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

$$b^1 = \sum_i \hat{\alpha}_{1,i} \times v^i$$

$$b^2 = \sum_i \hat{\alpha}_{2,i} \times v^i$$



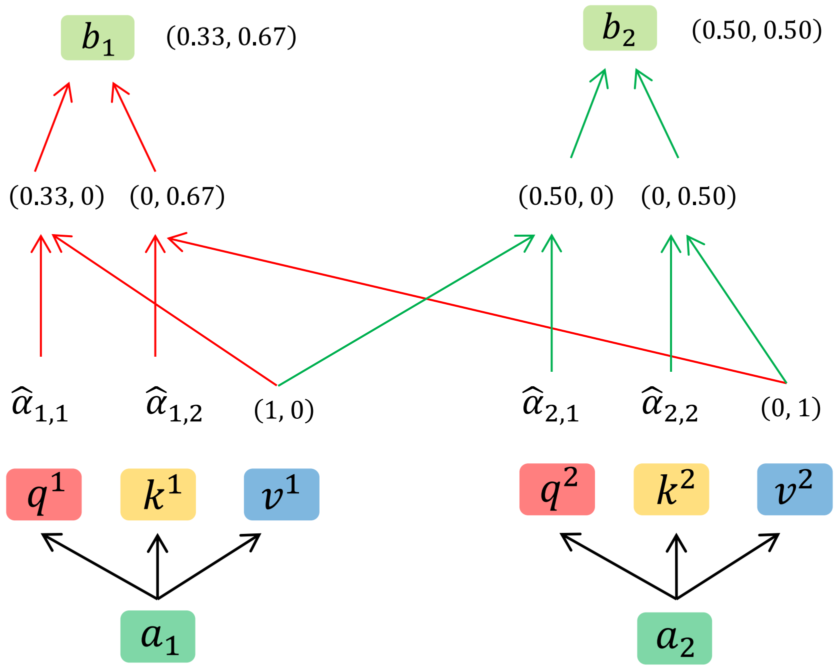
$$\begin{pmatrix} \hat{\alpha}_{1,1} & \hat{\alpha}_{1,2} \\ \hat{\alpha}_{2,1} & \hat{\alpha}_{2,2} \end{pmatrix} \quad \begin{pmatrix} 0.33 & 0.67 \\ 0.50 & 0.50 \end{pmatrix}$$

Self-Attention

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

$$b^1 = \sum_i \hat{\alpha}_{1,i} \times v^i$$

$$b^2 = \sum_i \hat{\alpha}_{2,i} \times v^i$$

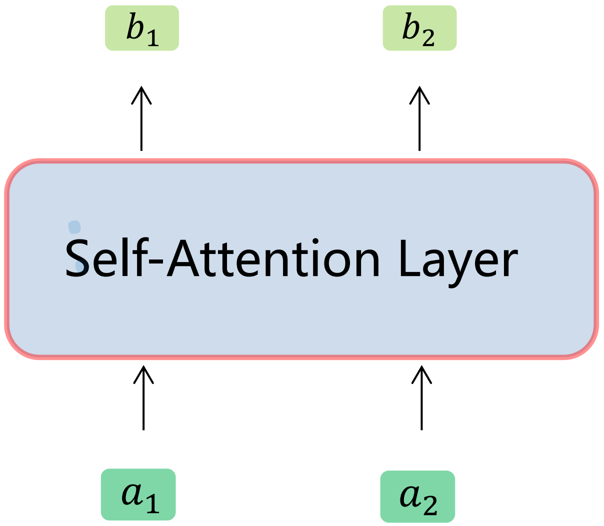


$$\begin{pmatrix} 0.33 & 0.67 \\ 0.50 & 0.50 \end{pmatrix} = \begin{pmatrix} 0.33 & 0.67 \\ 0.50 & 0.50 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\begin{matrix} b_1 \\ b_2 \end{matrix} = \begin{pmatrix} \hat{\alpha}_{1,1} & \hat{\alpha}_{1,2} \\ \hat{\alpha}_{2,1} & \hat{\alpha}_{2,2} \end{pmatrix} \begin{matrix} v^1 \\ v^2 \end{matrix}$$

Self-Attention

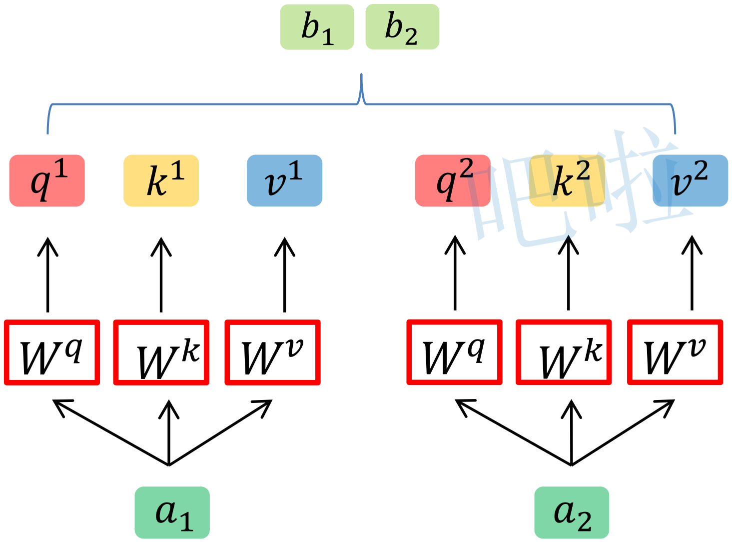
$$\mathrm{Attention}(Q,K,V)=\mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$



Multi-head Self-Attention

1个head的情况

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



Multi-head Self-Attention

2个head的情况

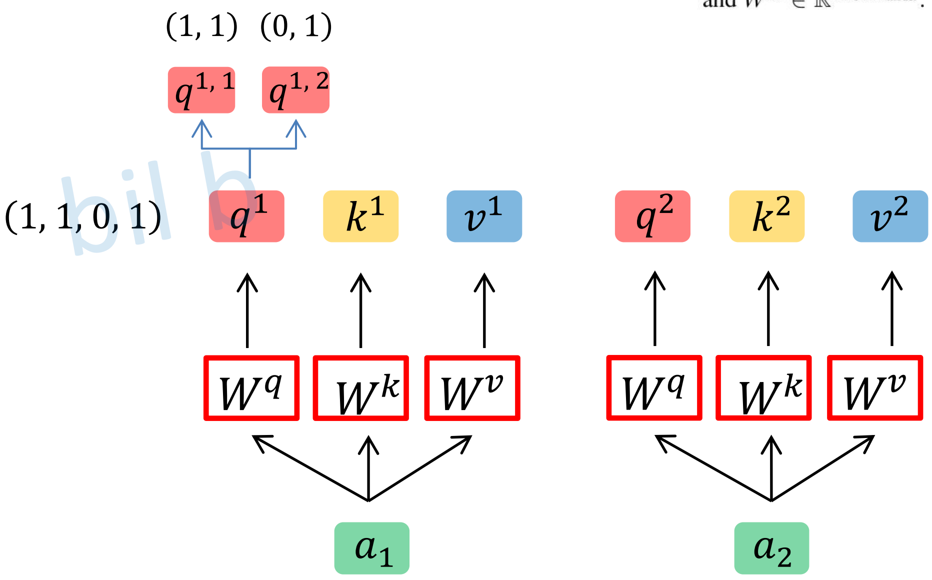
$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Concat}(\mathrm{head_1}, ..., \mathrm{head_h})W^O$$
$$\text{where } \mathrm{head_i} = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

Where the projections are parameter matrices $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$ and $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$.

$$d_k = d_v = d_{\text{model}}/h$$

2 2 4 2

线性映射



Multi-head Self-Attention

2个head的情况

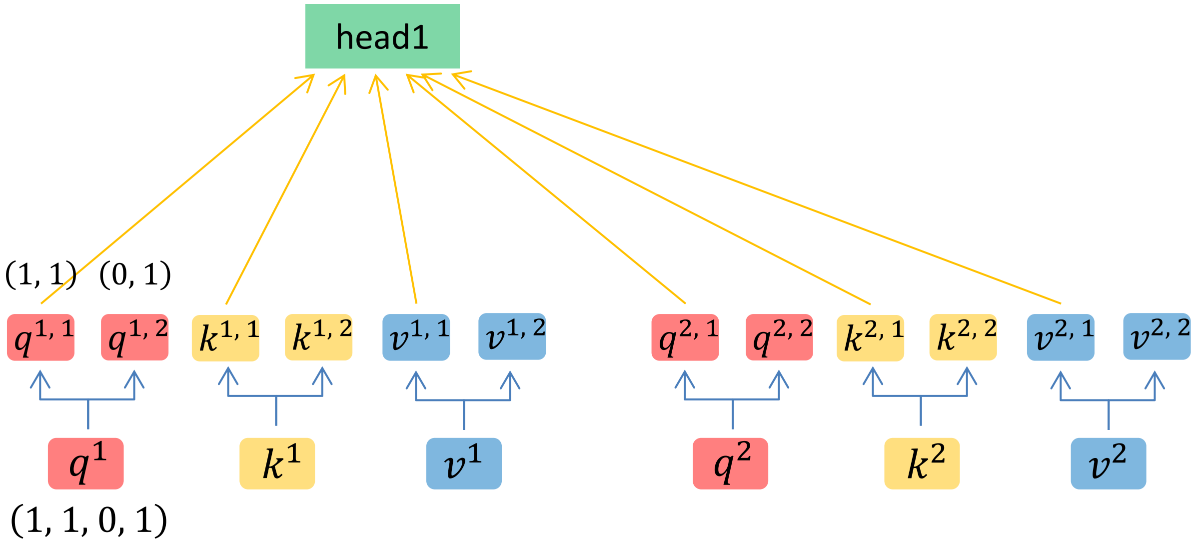
$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Concat}(\mathrm{head_1}, ..., \mathrm{head_h})W^O$$

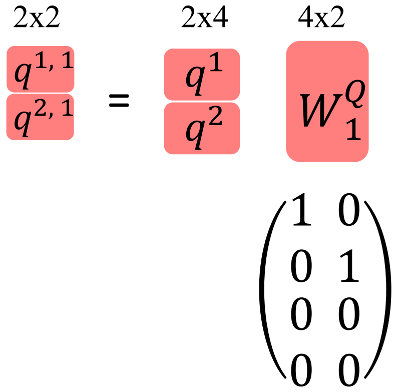
$$\text{where } \mathrm{head_i} = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

Where the projections are parameter matrices $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$ and $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$.

$$d_k = d_v = d_{\text{model}}/h$$

2 2 4 2





Multi-head Self-Attention

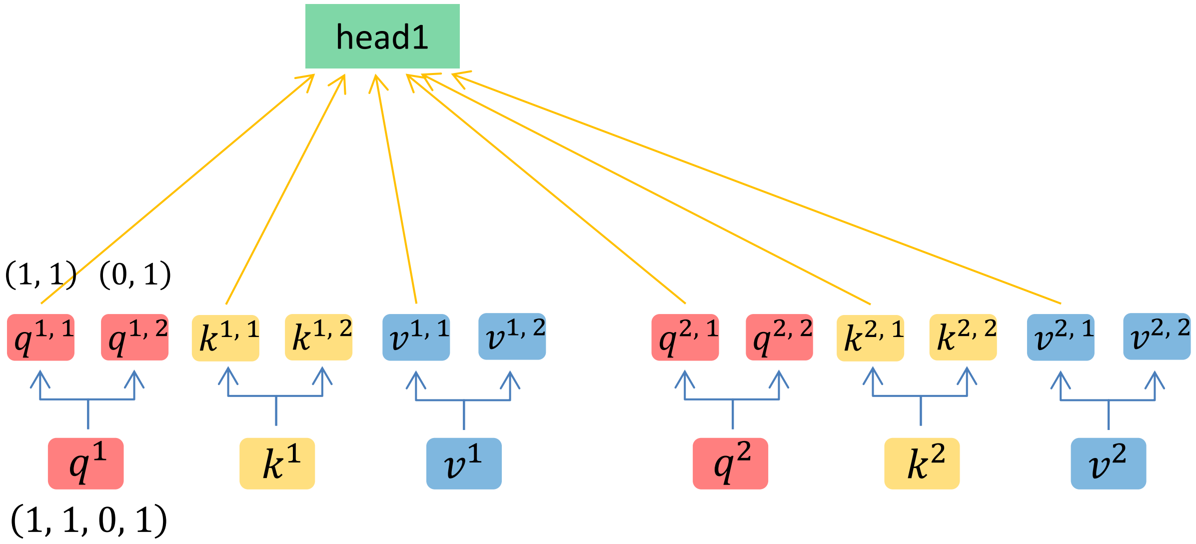
2个head的情况

$$\mathrm{MultiHead}(Q,K,V) = \mathrm{Concat}(\mathrm{head_1},...,\mathrm{head_h})W^O$$
$$\text{where } \mathrm{head_i} = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

Where the projections are parameter matrices $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$ and $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$.

$$d_k = d_v = d_{\text{model}}/h$$

2 2 4 2



2x2 2x4 4x2

$$\begin{bmatrix} q^{1,1} \\ q^{2,1} \end{bmatrix} = \begin{bmatrix} q^1 \\ q^2 \end{bmatrix} W_1^Q$$

$$\begin{bmatrix} k^{1,1} \\ k^{2,1} \end{bmatrix} = \begin{bmatrix} k^1 \\ k^2 \end{bmatrix} W_1^K$$

$$\begin{bmatrix} v^{1,1} \\ v^{2,1} \end{bmatrix} = \begin{bmatrix} v^1 \\ v^2 \end{bmatrix} W_1^V$$

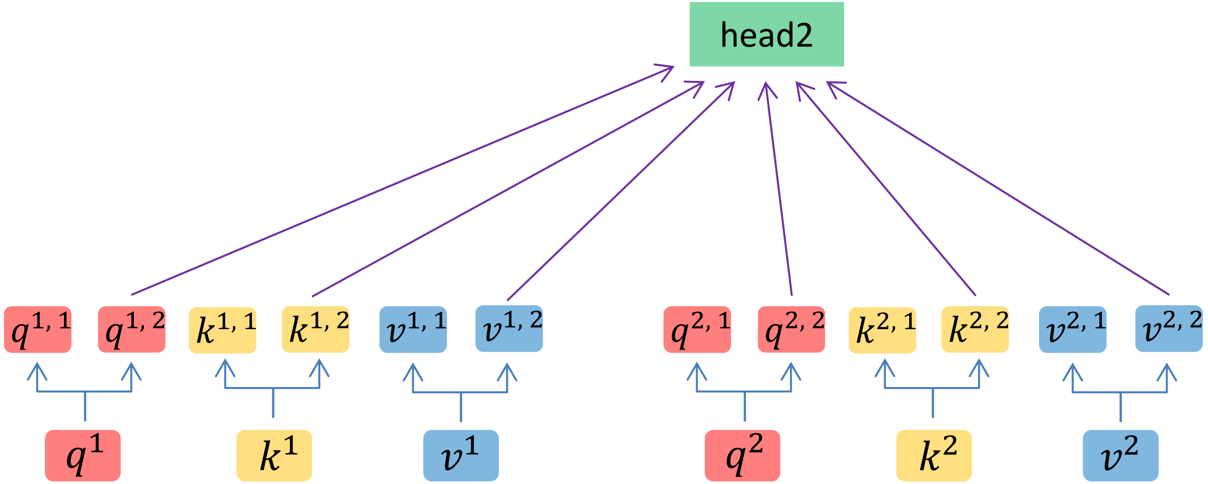
Multi-head Self-Attention

2个head的情况

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$$
$$\text{where head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

Where the projections are parameter matrices $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$ and $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$.

$$d_k = d_v = d_{\text{model}}/h$$
$$\;2 \quad\;\; 2 \quad\;\;\; 4 \quad 2$$



2x2 2x4 4x2

$$\begin{bmatrix} q^{1,2} \\ q^{2,2} \end{bmatrix} = \begin{bmatrix} q^1 \\ q^2 \end{bmatrix} W_2^Q$$

$$\begin{bmatrix} k^{1,2} \\ k^{2,2} \end{bmatrix} = \begin{bmatrix} k^1 \\ k^2 \end{bmatrix} W_2^K$$

$$\begin{bmatrix} v^{1,2} \\ v^{2,2} \end{bmatrix} = \begin{bmatrix} v^1 \\ v^2 \end{bmatrix} W_2^V$$

Multi-head Self-Attention

2个head的情况

$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Concat}(\mathrm{head_1}, ..., \mathrm{head_h})W^O$$
$$\text{where } \mathrm{head_i} = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

Where the projections are parameter matrices $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$ and $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$.

$$d_k = d_v = d_{\text{model}}/h$$

2 2 4 2

$b_{1,1}$ $b_{2,1}$

$q^{1,1}$ $k^{1,1}$ $v^{1,1}$ $q^{2,1}$ $k^{2,1}$ $v^{2,1}$

head1

$b_{1,2}$ $b_{2,2}$

$q^{1,2}$ $k^{1,2}$ $v^{1,2}$ $q^{2,2}$ $k^{2,2}$ $v^{2,2}$

head2

Multi-head Self-Attention

Concat

2个head的情况

$$\mathrm{MultiHead}(Q,K,V) = \mathrm{Concat}(\mathrm{head_1},\ldots,\mathrm{head_h})W^O$$
$$\text{where } \mathrm{head_i} = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

Where the projections are parameter matrices $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$ and $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$.

$$d_k = d_v = d_{\text{model}}/h$$

2 2 4 2

$b_{1,1}$ $b_{1,2}$ $b_{2,1}$ $b_{2,2}$

(1, 1, 0, 1) (0, 1, 1, 0)

$b_{1,1}$ $b_{2,1}$ $b_{1,2}$ $b_{2,2}$

(1, 1) (0, 1) (0, 1) (1, 0)

Multi-head Self-Attention

Fused

2个head的情况

$$\mathrm{MultiHead}(Q,K,V) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h)\underline{W^O}$$
$$\text{where } \mathrm{head}_i = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

Where the projections are parameter matrices $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$ and $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$.

$$\underset{2}{d_k} = \underset{2}{d_v} = \underset{4}{d_{\text{model}}}/\underset{2}{h}$$

b_1
b_2

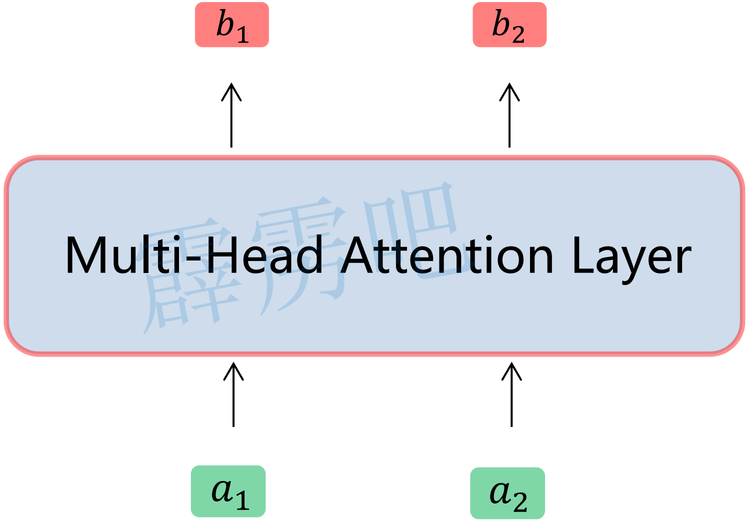
$b_{1,1}$ $b_{1,2}$ $(1, 1, 0, 1)$

$b_{2,1}$ $b_{2,2}$ $(0, 1, 1, 0)$

$$\underset{2\text{x}4}{\begin{pmatrix} b_1 \\ b_2 \end{pmatrix}} = \underset{2\text{x}4}{\begin{pmatrix} b_{1,1} & b_{1,2} \\ b_{2,1} & b_{2,2} \end{pmatrix}} \underset{4\text{x}4}{W^O}$$

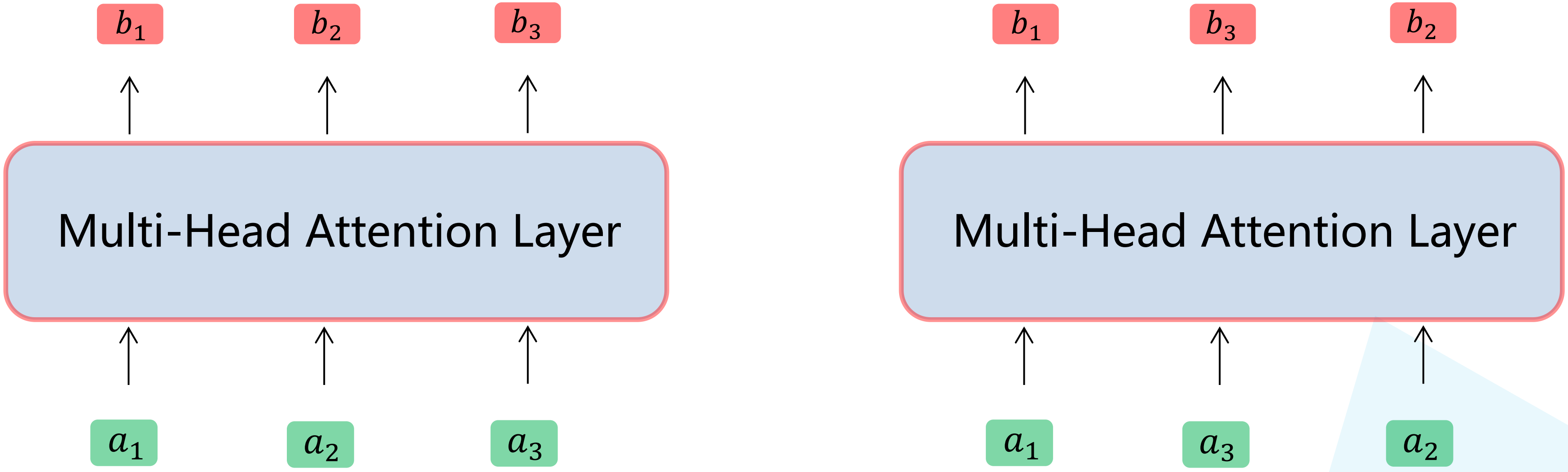
$$\begin{pmatrix} 1, 3, 0, 1 \\ 1, 1, 1, 2 \end{pmatrix} = \begin{pmatrix} 1, 1, 0, 1 \\ 0, 1, 1, 0 \end{pmatrix} \begin{pmatrix} 1, 1, 0, 0 \\ 0, 1, 0, 1 \\ 1, 0, 1, 1 \\ 0, 1, 0, 0 \end{pmatrix}$$

Multi-head Self-Attention



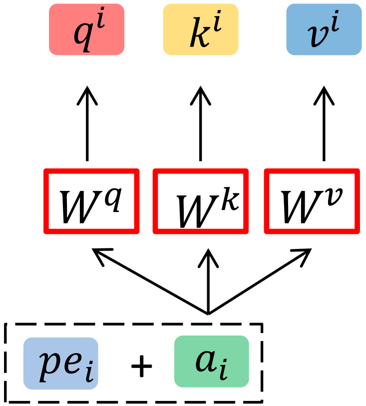
Multi-head Self-Attention

Positional Encoding



Multi-head Self-Attention

Positional Encoding



- 根据论文公式计算出位置编码
- 可训练的位置编码

沟通方式

1.github

https://github.com/WZMIAOMIAO/deep-learning-for-image-processing

2.bilibili

https://space.bilibili.com/18161609/channel/index

3.CSDN

https://blog.csdn.net/qq_37541097/article/details/103482003