

转换器中的多层感知机

Multilayer Perceptron in Transformer

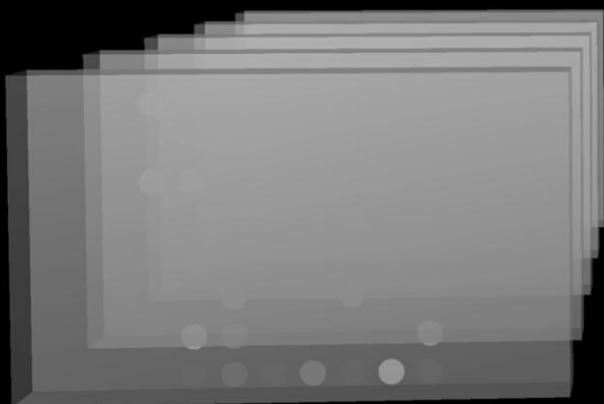
1 GPT-3中的MLP

- 主要存储世界知识 (facts)

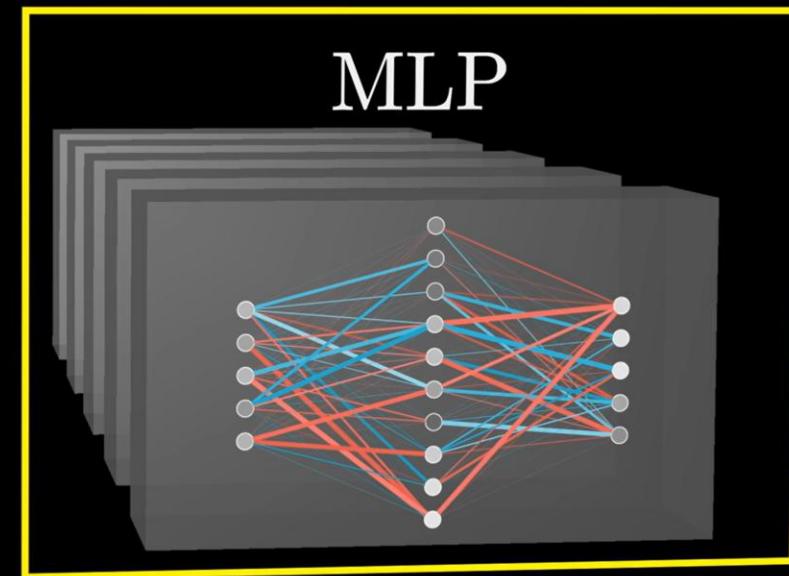
What are these?



Attention

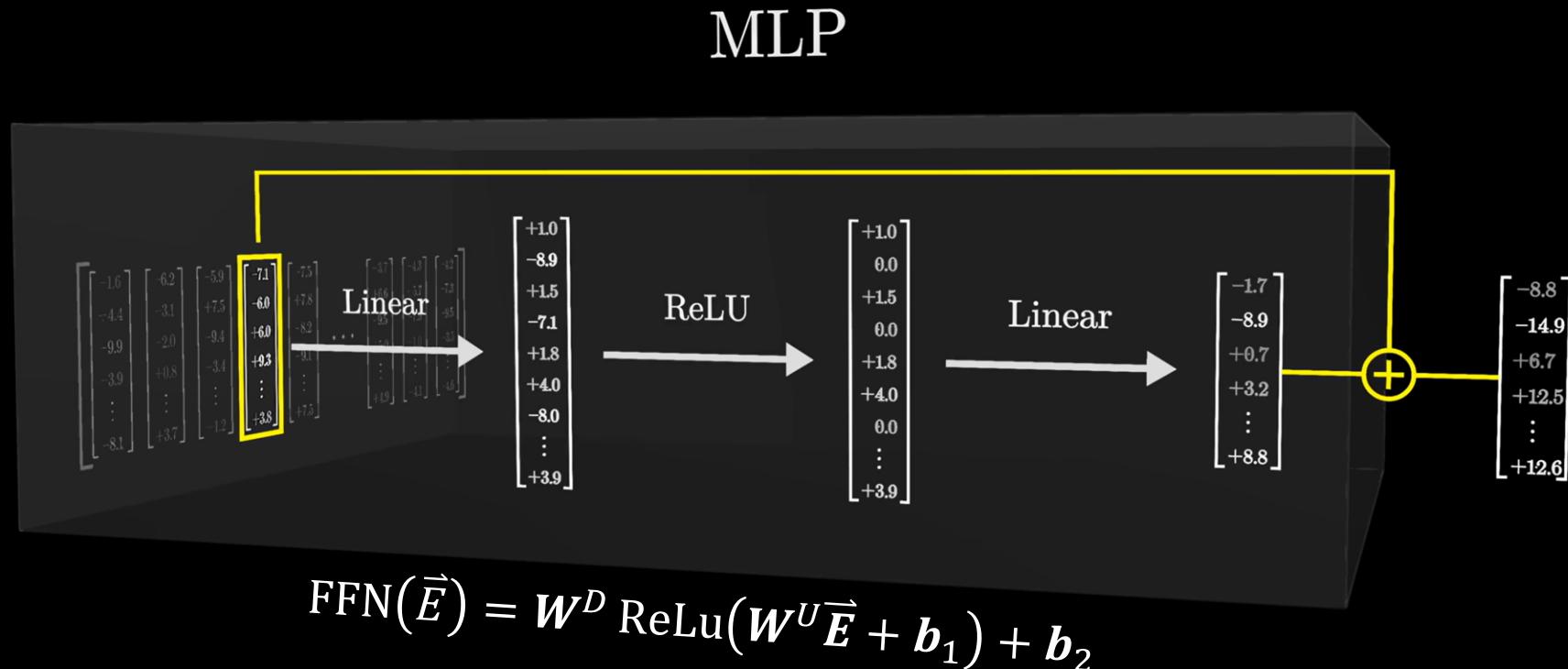


MLP



1 GPT-3中的MLP

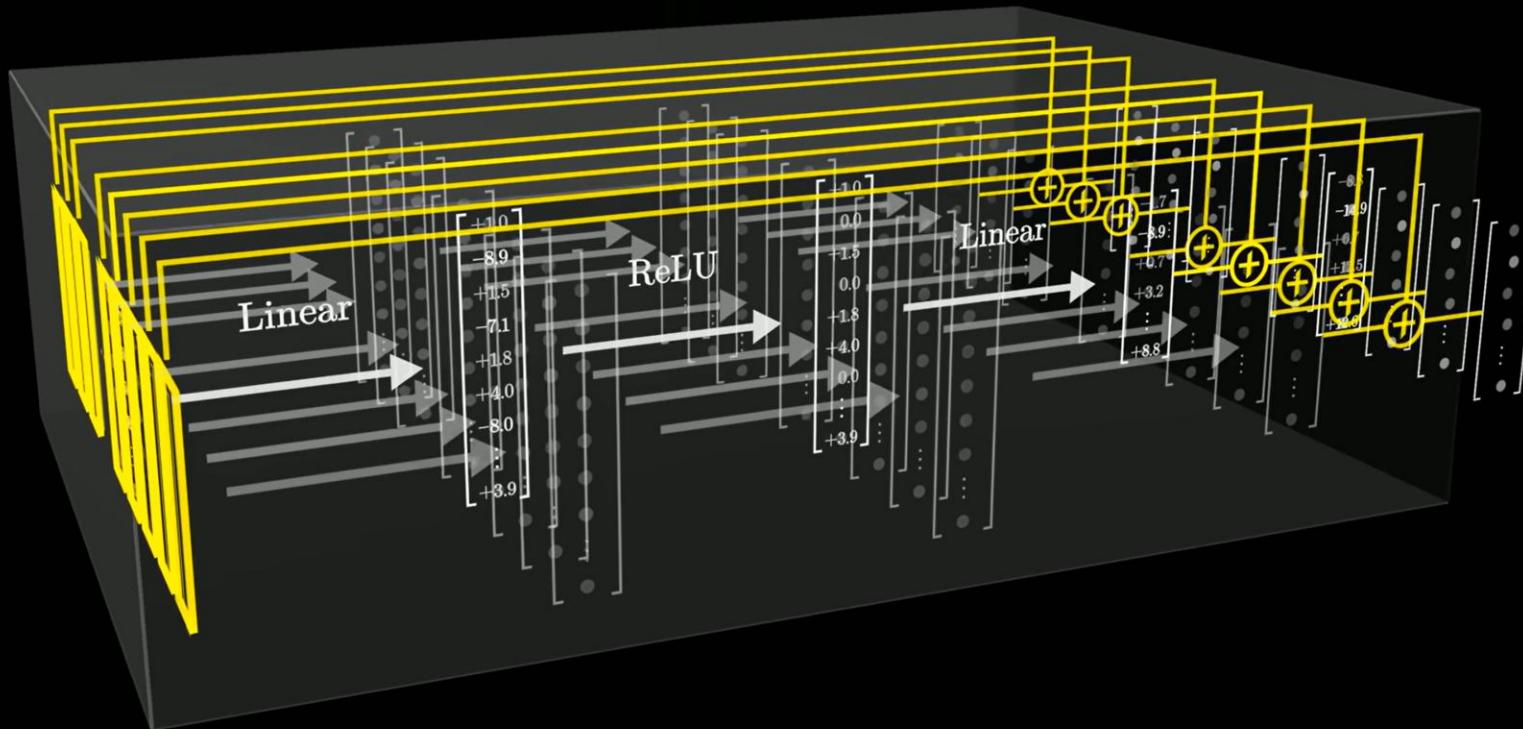
- 2层全连接神经网络



多层感知机 (MLP) 是最典型的全连接前向神经网络；“前向神经网络 (FFN)”强调信息单向传播，“全连接神经网络”强调层间连接方式，三者在多数教学语境下常指同一类模型，但严格来说 FFN 的概念更广。

1 GPT-3中的MLP

- 对每个经注意力机制处理后的嵌入向量同时处理

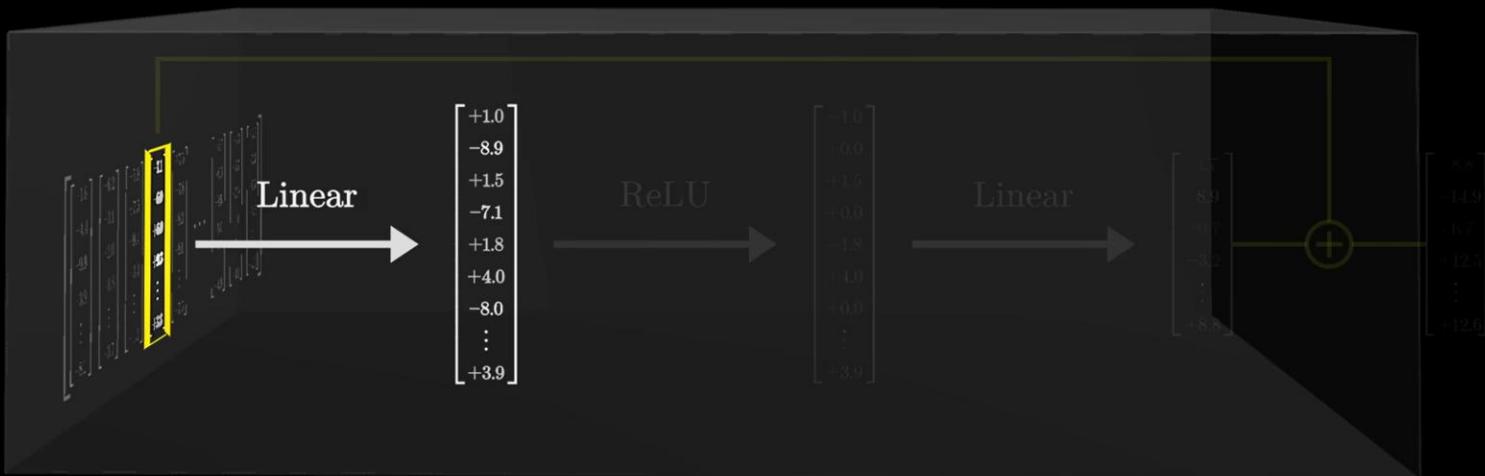


2 第一层

Up-projection

$$\overbrace{\begin{bmatrix} +1.7 & +8.0 & -7.2 & -7.1 & \cdots & -2.0 \\ -6.6 & +8.5 & -3.0 & +5.0 & \cdots & +7.6 \\ +2.4 & +5.0 & -3.0 & -4.6 & \cdots & -1.4 \\ +9.2 & +3.2 & +2.4 & -7.6 & \cdots & -1.0 \\ +1.6 & -1.8 & -5.2 & +8.0 & \cdots & -9.8 \\ +2.3 & -3.4 & +0.5 & +7.6 & \cdots & +8.1 \\ \boxed{+2.4} & \boxed{-9.6} & \boxed{+8.5} & \boxed{+3.8} & \cdots & \boxed{-6.5} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ +8.4 & +4.2 & -7.4 & -9.5 & \cdots & -9.3 \end{bmatrix}}^{\text{Up-projection}} = \begin{bmatrix} -199.5 \\ +36.6 \\ -142.0 \\ -177.5 \\ -0.1 \\ +118.9 \\ +65.9 \\ \vdots \\ -217.6 \end{bmatrix}$$

MLP

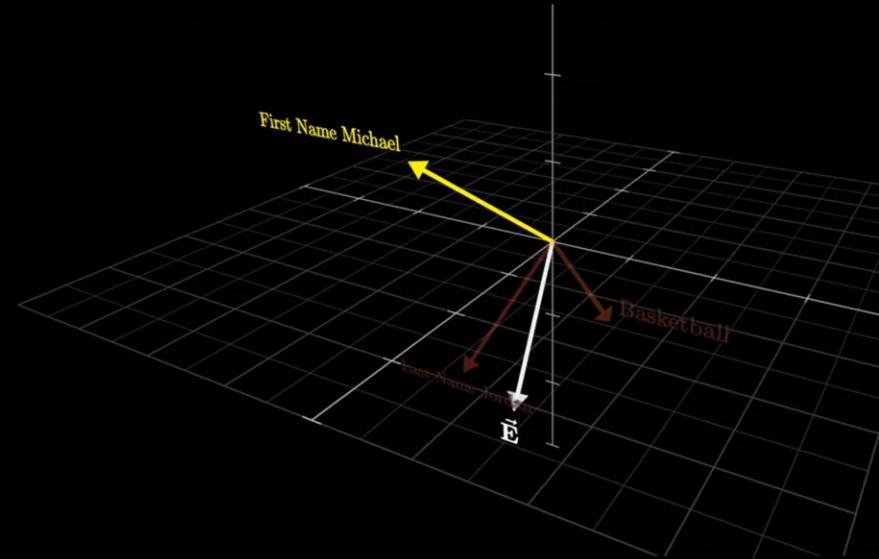


2 第一层

- 类比：“你问我猜”

Questions

$$\begin{bmatrix} \vec{\mathbf{R}}_0 \\ \vec{\mathbf{R}}_1 \\ \vec{\mathbf{R}}_2 \\ \vdots \\ \vec{\mathbf{R}}_n \end{bmatrix} \begin{bmatrix} \vec{\mathbf{E}} \end{bmatrix} = \begin{bmatrix} \vec{\mathbf{R}}_0 \cdot \vec{\mathbf{E}} \\ \vec{\mathbf{R}}_1 \cdot \vec{\mathbf{E}} \\ \vec{\mathbf{R}}_2 \cdot \vec{\mathbf{E}} \\ \vdots \\ \vec{\mathbf{R}}_n \cdot \vec{\mathbf{E}} \end{bmatrix}$$



MLP



2 第一层

- 类比：“你问我猜”

$$\begin{bmatrix} -5.0 & \text{Is it English?} & +6.8 \\ -7.4 & -4.4 & +1.7 & +9.3 & \cdots & +1.2 \\ -9.5 & +6.0 & \cdots & -2.2 \\ +7.2 & +4.9 & +1.1 & -7.2 & \cdots & -8.7 \\ -7.5 & -9.0 & -7.8 & -5.4 & \cdots & +4.2 \\ +1.2 & -9.7 & -8.5 & +9.3 & \cdots & +1.3 \\ -5.9 & -4.9 & +4.8 & -6.0 & \cdots & +1.6 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ +9.3 & +6.9 & -5.2 & -0.1 & \cdots & +2.4 \end{bmatrix} = \begin{bmatrix} -7.1 \\ -6.0 \\ +6.0 \\ +9.3 \\ \vdots \\ +3.8 \end{bmatrix} = \begin{bmatrix} -199.5 \\ +36.6 \\ -142.0 \\ -177.5 \\ -0.1 \\ +118.9 \\ +65.9 \\ \vdots \\ -218.2 \end{bmatrix}$$

MLP



2 第一层

- 类比：“你问我猜”

$$\begin{bmatrix} -5.0 & +7.1 & +0.8 & +1.0 & \cdots & +6.8 \\ \text{Part of source code?} \\ -9.5 & +6.0 & -5.3 & +6.1 & \cdots & -2.2 \\ +2.1 & -7.8 & -7.5 & -5.4 & \cdots & +4.2 \\ +1.2 & -9.7 & -8.5 & +9.3 & \cdots & +1.3 \\ -5.9 & -4.9 & +4.8 & -6.0 & \cdots & +1.6 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ +9.3 & +6.9 & -5.2 & -0.1 & \cdots & +2.4 \end{bmatrix} = \begin{bmatrix} -7.1 \\ -6.0 \\ +6.0 \\ +9.3 \\ \vdots \\ +3.8 \end{bmatrix}$$

$$= \begin{bmatrix} -199.5 \\ +36.6 \\ -142.0 \\ -177.5 \\ -0.1 \\ +118.9 \\ +65.9 \\ \vdots \\ -218.2 \end{bmatrix}$$

MLP



2 第一层

- 类比：“你问我猜”

$$\begin{bmatrix} -5.0 & +7.1 & +0.8 & +1.0 & \cdots & +6.8 \\ -7.4 & -4.4 & +1.7 & +9.3 & \cdots & +1.2 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ +7.2 & +4.9 & +1.1 & -7.2 & \cdots & -8.7 \\ -7.5 & -9.0 & -5.0 & +5.1 & \cdots & +4.2 \\ +1.2 & -9.7 & -8.5 & +9.3 & \cdots & +1.3 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ +9.3 & +6.9 & -5.2 & -0.1 & \cdots & +2.4 \end{bmatrix} \begin{bmatrix} -7.1 \\ -6.0 \\ +6.0 \\ +9.3 \\ \vdots \\ +3.8 \end{bmatrix} = \begin{bmatrix} -199.5 \\ +36.6 \\ -142.0 \\ -177.5 \\ -0.1 \\ +118.9 \\ +65.9 \\ \vdots \\ -218.2 \end{bmatrix}$$

是欧洲国家吗？

MLP



2 第一层

- 类比：“你问我猜”

$$\begin{bmatrix} -5.0 & +7.1 & +0.8 & +1.0 & \cdots & +6.8 \\ -7.4 & -4.4 & +1.7 & +9.3 & \cdots & +1.2 \\ -9.5 & +6.0 & -5.3 & +6.1 & \cdots & -2.2 \\ \text{In quotation marks?} & & & & & \\ \text{-7.5} & -9.0 & -7.8 & -5.4 & \cdots & +4.2 \\ +1.2 & -9.7 & -1.1 & -8.2 & \cdots & +1.3 \\ -5.9 & -4.9 & +4.8 & -6.0 & \cdots & +1.6 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ +9.3 & +6.9 & -5.2 & -0.1 & \cdots & +2.4 \end{bmatrix} = \begin{bmatrix} -7.1 \\ -6.0 \\ +6.0 \\ +9.3 \\ \vdots \\ +3.8 \end{bmatrix}$$

$$= \begin{bmatrix} -199.5 \\ +36.6 \\ -142.0 \\ -177.5 \\ -0.1 \\ +118.9 \\ +65.9 \\ \vdots \\ -218.2 \end{bmatrix}$$

MLP



2 第一层

- 类比：“你问我猜”

$$\begin{bmatrix} -5.0 & +7.1 & +0.8 & +1.0 & \cdots & +6.8 \\ -7.4 & -4.4 & +1.7 & +9.3 & \cdots & +1.2 \\ -9.5 & +6.0 & -5.3 & +6.1 & \cdots & -2.2 \\ +7.2 & +4.9 & +1.1 & -7.2 & \cdots & -8.7 \\ \text{Something metallic?} & & & & & \\ +1.2 & -9.7 & -8.5 & +9.3 & \cdots & +1.3 \\ -5.9 & \text{是某种金属吗?} & & & & +1.6 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ +9.3 & +6.9 & -5.2 & -0.1 & \cdots & +2.4 \end{bmatrix} = \begin{bmatrix} -7.1 \\ -6.0 \\ +6.0 \\ +9.3 \\ \vdots \\ +3.8 \end{bmatrix} = \begin{bmatrix} -199.5 \\ +36.6 \\ -142.0 \\ -177.5 \\ -0.1 \\ +118.9 \\ +65.9 \\ \vdots \\ -218.2 \end{bmatrix}$$

MLP



2 第一层

- 类比：“你问我猜”

$$\begin{bmatrix} -5.0 & +7.1 & +0.8 & +1.0 & \cdots & +6.8 \\ -7.4 & -4.4 & +1.7 & +9.3 & \cdots & +1.2 \\ -9.5 & +6.0 & -5.3 & +6.1 & \cdots & -2.2 \\ +7.2 & +4.9 & +1.1 & -7.2 & \cdots & -8.7 \\ -7.5 & -9.0 & -7.8 & -5.4 & \cdots & +4.2 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ +9.3 & +6.9 & -5.2 & -0.1 & \cdots & +2.4 \end{bmatrix} \begin{bmatrix} -7.1 \\ -6.0 \\ +6.0 \\ +9.3 \\ \vdots \\ +3.8 \end{bmatrix} = \begin{bmatrix} -199.5 \\ +36.6 \\ -142.0 \\ -177.5 \\ -0.1 \\ +118.9 \\ +65.9 \\ \vdots \\ -218.2 \end{bmatrix}$$

A four-legged animal?
是四足动物吗?

MLP



2 第一层

- 维度与参数数量

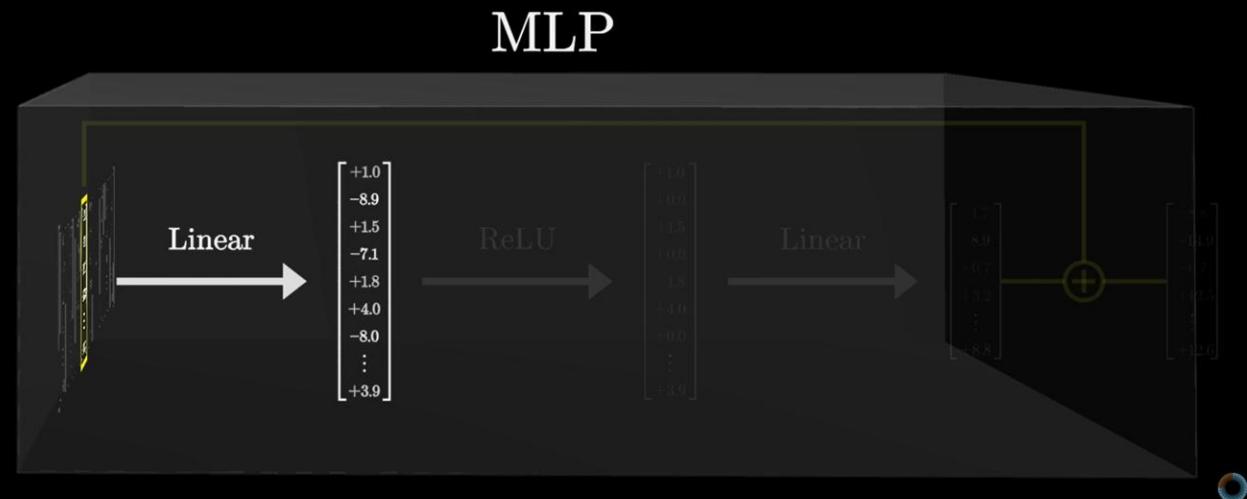
12,288

$$4 \times 12,288 = 49,152$$
$$\left\{ \begin{array}{|c|c|c|c|c|c|} \hline & -5.0 & +7.1 & +0.8 & +1.0 & \cdots & +6.8 \\ \hline & -7.4 & -4.4 & +1.7 & +9.3 & \cdots & +1.2 \\ \hline & -9.5 & +6.0 & -5.3 & +6.1 & \cdots & -2.2 \\ \hline & +7.2 & +4.9 & +1.1 & -7.2 & \cdots & -8.7 \\ \hline & -7.5 & -9.0 & -7.8 & -5.4 & \cdots & +4.2 \\ \hline & +1.2 & -9.7 & -8.5 & +9.3 & \cdots & +1.3 \\ \hline & -5.9 & -4.9 & +4.8 & -6.0 & \cdots & +1.6 \\ \hline & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \hline & +9.3 & +6.9 & -5.2 & -0.1 & \cdots & +2.4 \\ \hline \end{array} \right\}$$

Bias

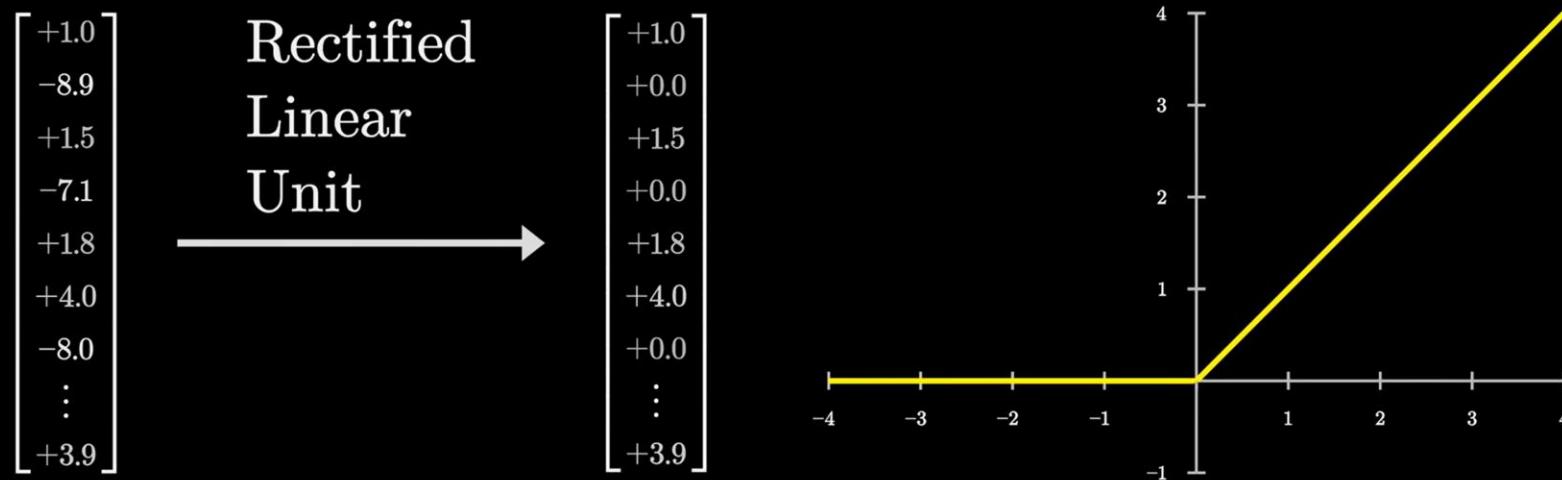
$$\begin{bmatrix} -7.1 \\ -6.0 \\ +6.0 \\ +9.3 \\ \vdots \\ +3.8 \end{bmatrix} + \begin{bmatrix} -1.0 \\ -5.7 \\ +5.0 \\ -8.6 \\ -4.7 \\ +6.0 \\ -6.1 \\ \vdots \\ +2.8 \end{bmatrix} = \begin{bmatrix} +1.0 \\ -8.9 \\ +1.5 \\ -7.1 \\ +1.8 \\ +4.0 \\ -8.0 \\ \vdots \\ +3.9 \end{bmatrix}$$

 GPT-3

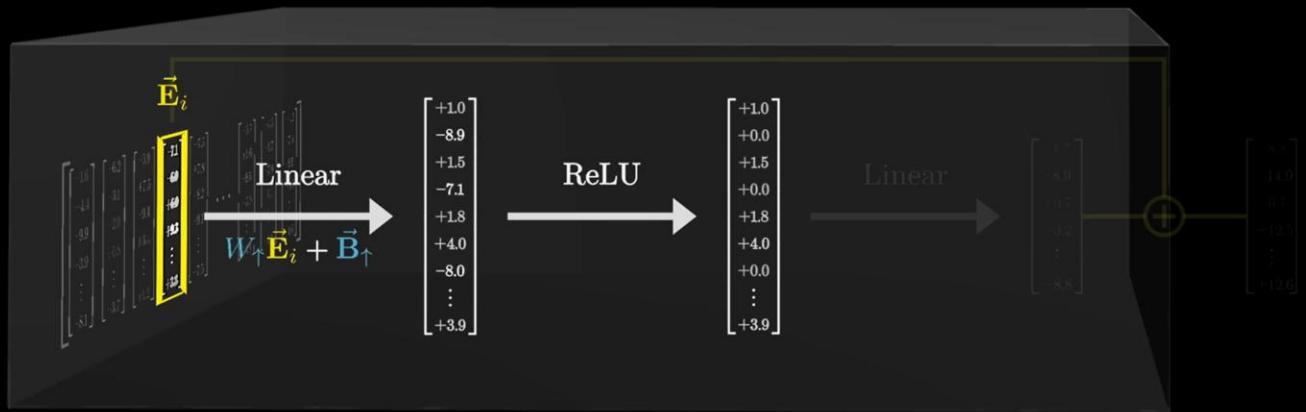


2 第一层

- 激活函数: ReLU



MLP



3 第二层

- “下投影”矩阵

“Down projection”

$$\underbrace{\begin{bmatrix} +0.5 & +8.4 & -4.7 & -8.6 & +4.7 & +5.4 & +8.1 & \cdots & -9.6 \\ -5.3 & +2.3 & +8.9 & +8.9 & +1.1 & +8.2 & +2.8 & \cdots & -0.3 \\ +2.1 & +1.0 & +8.4 & +8.3 & -2.1 & +9.2 & -6.5 & \cdots & -7.2 \\ +0.1 & -9.5 & +8.9 & +6.5 & -9.6 & -6.4 & -3.3 & \cdots & +6.1 \\ \vdots & \ddots & \vdots \\ -4.2 & -0.2 & +2.0 & -9.6 & +1.9 & -1.3 & +6.1 & \cdots & +7.8 \end{bmatrix}}_{\text{“Down projection”}} + \begin{bmatrix} +1.0 \\ +0.0 \\ +1.5 \\ +0.0 \\ +1.8 \\ +4.0 \\ +0.0 \\ \vdots \\ -1.6 \end{bmatrix} = \begin{bmatrix} -1.7 \\ -8.9 \\ +0.7 \\ +3.2 \\ \vdots \\ +8.8 \end{bmatrix} \Bigg\} 12,288$$

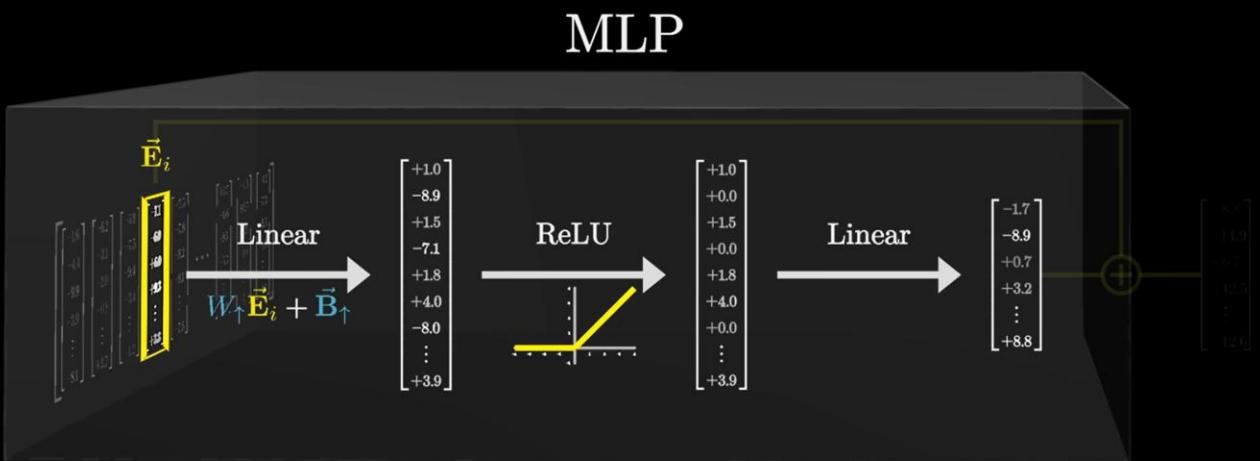
MLP



3 第二层

- 视作列向量组合

$$\begin{bmatrix} \vec{\mathbf{C}}_0 & \vec{\mathbf{C}}_1 & \vec{\mathbf{C}}_2 & \vec{\mathbf{C}}_3 & \vec{\mathbf{C}}_4 & \cdots & \vec{\mathbf{C}}_m \end{bmatrix} \begin{bmatrix} n_0 \\ n_1 \\ n_2 \\ n_3 \\ n_4 \\ \vdots \\ n_m \end{bmatrix} + \begin{bmatrix} \vec{\mathbf{B}} \end{bmatrix} = \begin{bmatrix} -1.7 \\ -8.9 \\ +0.7 \\ +3.2 \\ \vdots \\ +8.8 \end{bmatrix}$$



3 第二层

- 表示为对“你问我猜”的答案的加权组合

$$n_0 \vec{\mathbf{C}}_0 + n_1 \vec{\mathbf{C}}_1 + n_2 \vec{\mathbf{C}}_2 + n_3 \vec{\mathbf{C}}_3 + n_4 \vec{\mathbf{C}}_4 + \cdots + n_m \vec{\mathbf{C}}_m$$
$$\underbrace{\left[\begin{array}{c} \vec{\mathbf{C}}_0 \\ \vdots \\ \vec{\mathbf{C}}_1 \\ \vec{\mathbf{C}}_2 \\ \vec{\mathbf{C}}_3 \\ \vec{\mathbf{C}}_4 \\ \cdots \\ \vec{\mathbf{C}}_m \end{array} \right]}_{\text{A matrix of vectors}} \left[\begin{array}{c} n_0 \\ n_1 \\ n_2 \\ n_3 \\ n_4 \\ \vdots \\ n_m \end{array} \right] + \left[\begin{array}{c} \vec{\mathbf{B}} \end{array} \right] = \left[\begin{array}{c} -1.7 \\ -8.9 \\ +0.7 \\ +3.2 \\ \vdots \\ +8.8 \end{array} \right]$$



3 第二层

- 表示为对“你问我猜”的答案的加权组合

$$n_0 \vec{\mathbf{C}}_0 + n_1 \vec{\mathbf{C}}_1 + n_2 \vec{\mathbf{C}}_2 + n_3 \vec{\mathbf{C}}_3 + n_4 \vec{\mathbf{C}}_4 + \cdots + n_m \vec{\mathbf{C}}_m$$

$\overbrace{\quad\quad\quad\quad\quad\quad\quad\quad}$

$$\begin{array}{l} \text{篮球} \quad \xrightarrow{\text{Basketball}} \\ + \\ \text{芝加哥公牛} \xrightarrow{\text{Chicago Bulls}} \\ + \\ \text{23号球员} \xrightarrow{\text{Number 23}} \\ + \\ \text{1963年出生} \xrightarrow{\text{Born 1963}} \\ + \\ \vdots \end{array} \quad \begin{bmatrix} \vec{\mathbf{C}}_0 & \vec{\mathbf{C}}_1 & \vec{\mathbf{C}}_2 & \vec{\mathbf{C}}_3 & \vec{\mathbf{C}}_4 & \cdots & \vec{\mathbf{C}}_m \end{bmatrix} \begin{bmatrix} n_0 \\ n_1 \\ n_2 \\ n_3 \\ n_4 \\ \vdots \\ n_m \end{bmatrix} + \begin{bmatrix} \vec{\mathbf{B}} \end{bmatrix} = \begin{bmatrix} -1.7 \\ -8.9 \\ +0.7 \\ +3.2 \\ \vdots \\ +8.8 \end{bmatrix}$$

MLP

Diagram illustrating the MLP architecture:

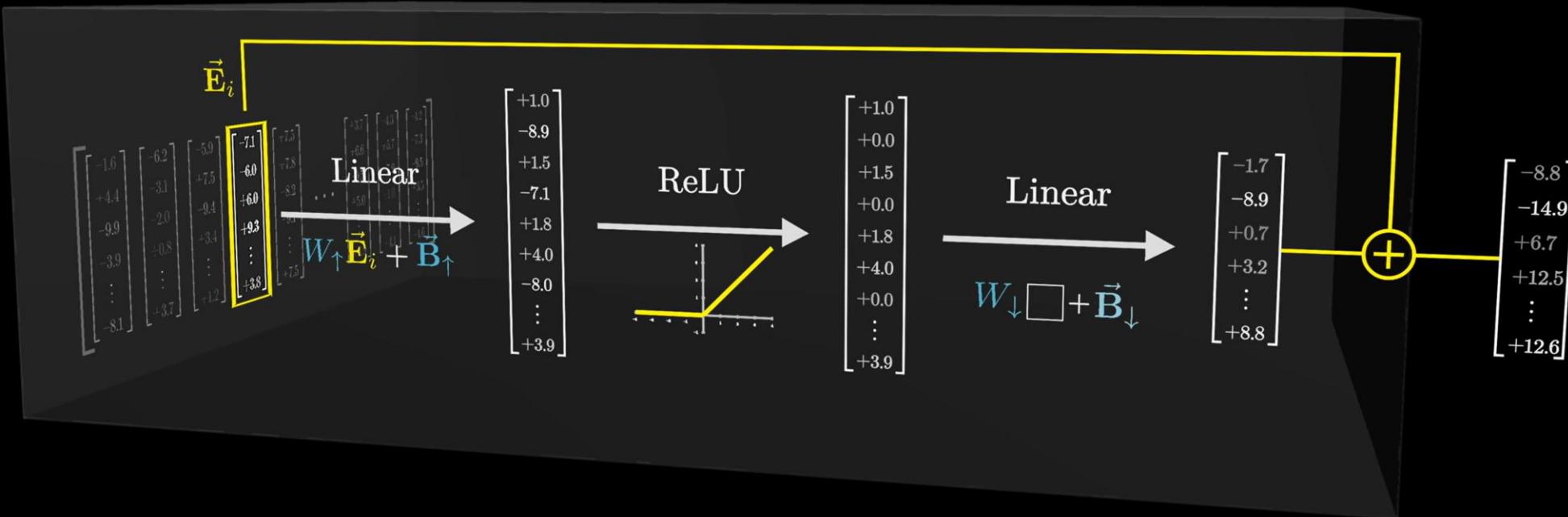
- The input vector $\vec{\mathbf{E}}_i$ is processed by a linear layer $W \uparrow \vec{\mathbf{E}}_i + \vec{\mathbf{B}} \uparrow$ to produce intermediate values: $[+1.0, -8.9, +1.5, -7.1, +1.8, +4.0, -8.0, \dots, +3.9]$.
- The output of the linear layer is passed through a ReLU activation function.
- The result of the ReLU activation is processed by another linear layer to produce the final output vector: $[+1.0, +0.0, +1.5, +0.0, +1.8, +4.0, +0.0, \dots, +3.9]$.
- The final output is calculated as the weighted sum of feature vectors $\vec{\mathbf{C}}$ plus the bias vector $\vec{\mathbf{B}}$: $[+1.0, -8.9, +0.7, +3.2, \dots, +8.8]$.

迈克尔·乔丹

3 第二层

- MLP完整结构

MLP



3 参数数量分析

- “上投影”矩阵参数数量：

$$4 \times 12,288 \times 12,288 = 603,979,776$$

$$4 \times 12,288 \left\{ \begin{matrix} & 12,288 \\ & \overbrace{\begin{bmatrix} -5.0 & +7.1 & +0.8 & +1.0 & \cdots & +6.8 \\ -7.4 & -4.4 & +1.7 & +9.3 & \cdots & +1.2 \\ -9.5 & +6.0 & -5.3 & +6.1 & \cdots & -2.2 \\ +7.2 & +4.9 & +1.1 & -7.2 & \cdots & -8.7 \\ -7.5 & -9.0 & -7.8 & -5.4 & \cdots & +4.2 \\ +1.2 & -9.7 & -8.5 & +9.3 & \cdots & +1.3 \\ -5.9 & -4.9 & +4.8 & -6.0 & \cdots & +1.6 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ +9.3 & +6.9 & -5.2 & -0.1 & \cdots & +2.4 \end{bmatrix}} \\ & \begin{bmatrix} -7.1 \\ -6.0 \\ +6.0 \\ +9.3 \\ \vdots \\ +3.8 \end{bmatrix} \end{matrix} \right. + \left. \text{Bias} \right\}$$

$\begin{bmatrix} -1.0 \\ -5.7 \\ +5.0 \\ -8.6 \\ -4.7 \\ +6.0 \\ -6.1 \\ \vdots \\ +2.8 \end{bmatrix} = \begin{bmatrix} +1.0 \\ -8.9 \\ +1.5 \\ -7.1 \\ +1.8 \\ +4.0 \\ -8.0 \\ \vdots \\ +3.9 \end{bmatrix}$

Up-projection	$49,152 \times 12,288$ $n_neurons * d_embed = 603,979,776$ per layer
Down-projection	
Unembedding	$50,257 \times 12,288$ $n_vocab * d_embed = 617,558,016$



3 参数数量分析

- “下投影”矩阵参数数量：

“Down projection”

$$\underbrace{\begin{bmatrix} +0.5 & +8.4 & -4.7 & -8.6 & +4.7 & +5.4 & +8.1 & \cdots & -9.6 \\ -5.3 & +2.3 & +8.9 & +8.9 & +1.1 & +8.2 & +2.8 & \cdots & -0.3 \\ +2.1 & +1.0 & +8.4 & +8.3 & -2.1 & +9.2 & -6.5 & \cdots & -7.2 \\ +0.1 & -9.5 & +8.9 & +6.5 & -9.6 & -6.4 & -3.3 & \cdots & +6.1 \\ \vdots & \ddots & \vdots \\ -4.2 & -0.2 & +2.0 & -9.6 & +1.9 & -1.3 & +6.1 & \cdots & +7.8 \end{bmatrix}}_{\text{Up-projection}} + \begin{bmatrix} +1.0 \\ +0.0 \\ +1.5 \\ +0.0 \\ +1.8 \\ +4.0 \\ +0.0 \\ \vdots \\ +3.9 \end{bmatrix} = \begin{bmatrix} -1.7 \\ -8.9 \\ +0.7 \\ +3.2 \\ \vdots \\ +8.8 \end{bmatrix} \quad \left. \right\} 12,288$$

Up-projection	$\frac{49,152}{n_neurons} * \frac{12,288}{d_embed} = 603,979,776$ per layer
Down-projection	$\frac{12,288}{d_embed} * \frac{49,152}{n_neurons} = 603,979,776$ per layer
Unembedding	$\frac{50,257}{n_vocab} * \frac{12,288}{d_embed} = 617,558,016$



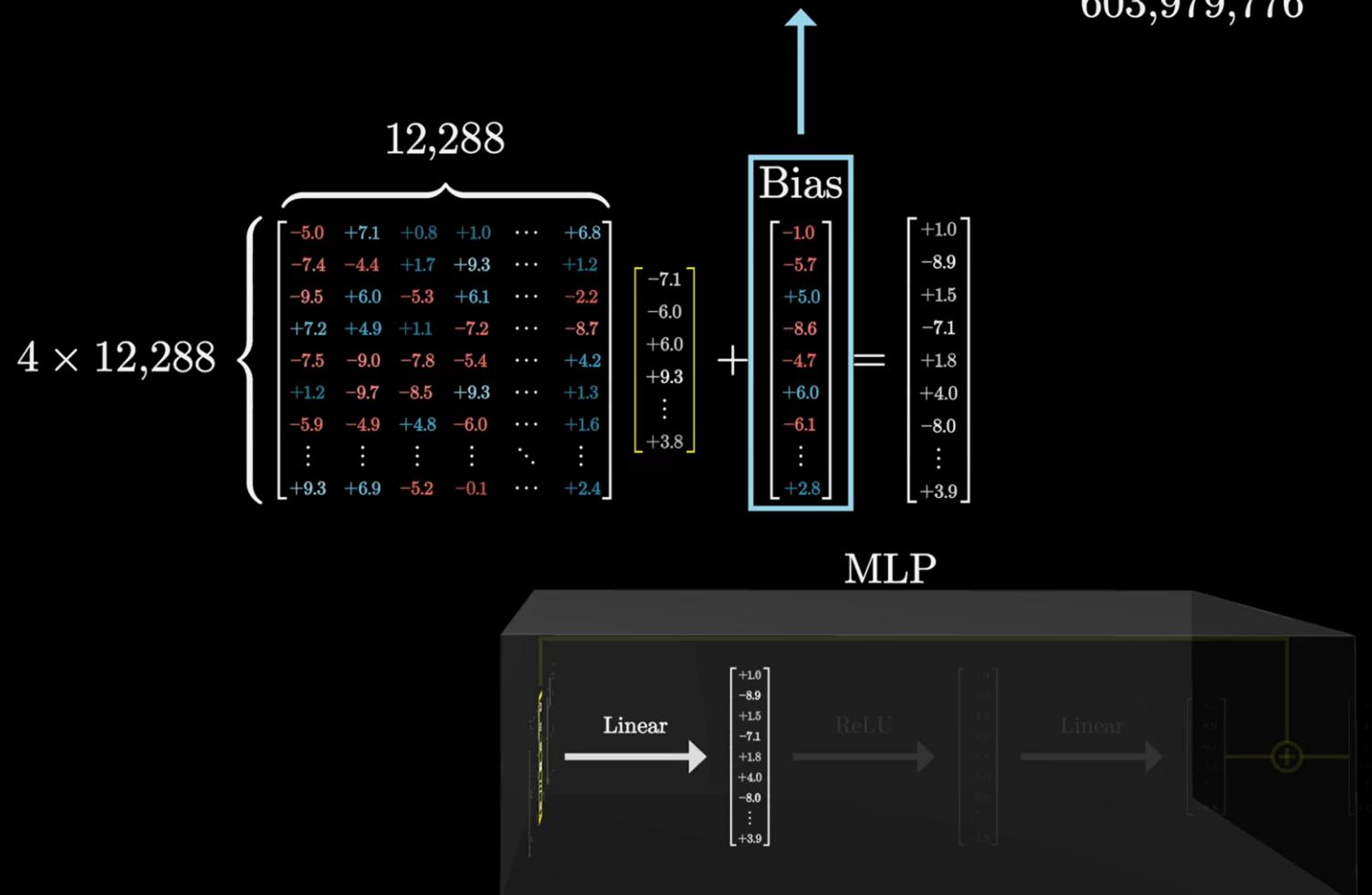
3 参数数量分析

- Bias部分参数可忽略

$$4 \times 12,288 \times 12,288 = 603,979,776$$

$$4 \times 12,288$$

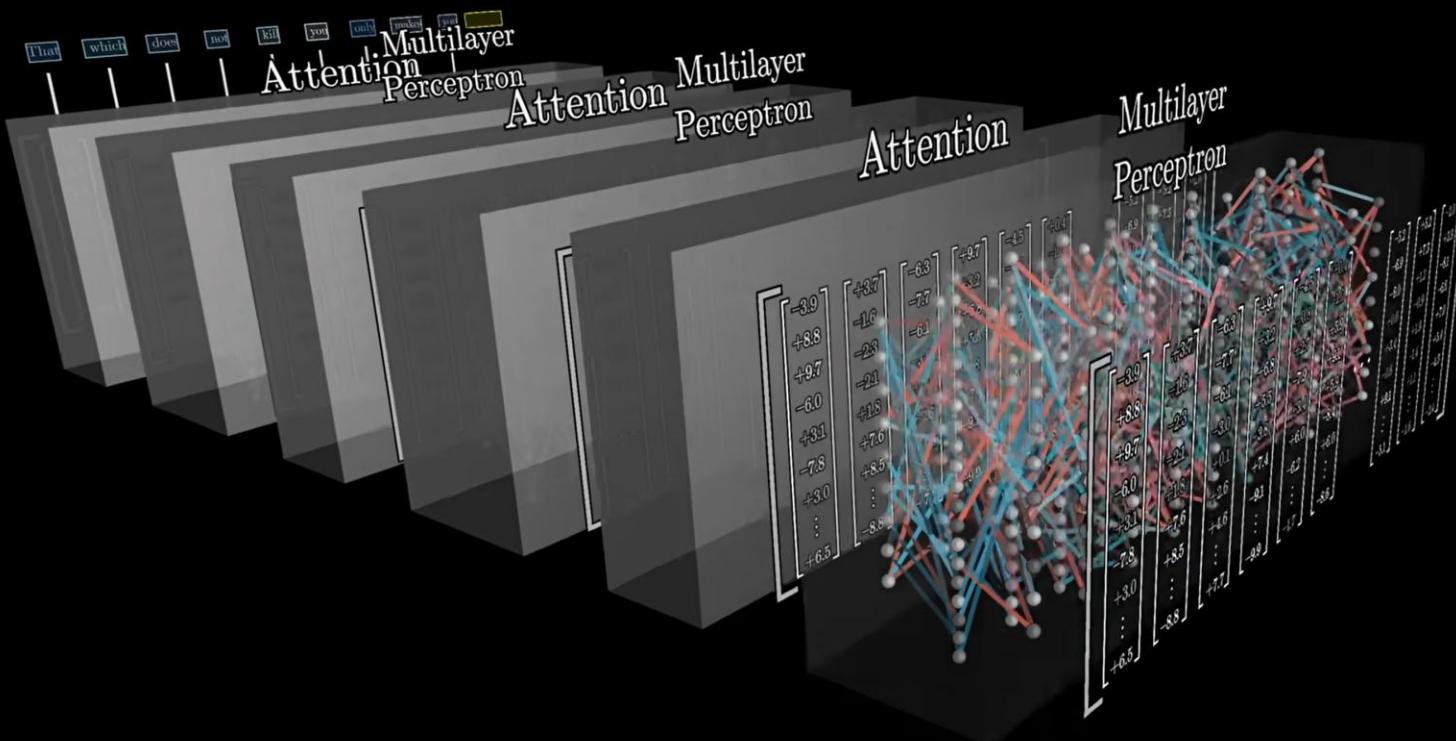
$$\frac{4 \times 12,288}{603,979,776} \approx 0.00008$$



3 参数数量分析

- × 96层网络

96 Layers



3 参数数量分析

- GPT-3参数总量

Total weights: 175,181,291,520

Organized into 27,938 matrices



GPT-3

Embedding	12,288	50,257			
	d_embed	*	n_vocab	=	617,558,016
Key	128	12,288	96	96	
	d_query	*	d_embed	*	n_heads * n_layers = 14,495,514,624
Query	128	12,288	96	96	
	d_query	*	d_embed	*	n_heads * n_layers = 14,495,514,624
Value	128	12,288	96	96	
	d_value	*	d_embed	*	n_heads * n_layers = 14,495,514,624
Output	12,288	128	96	96	
	d_embed	*	d_value	*	n_heads * n_layers = 14,495,514,624
Up-projection	49,152	12,288	96		
	n_neurons	*	d_embed	*	n_layers = 57,982,058,496
Down-projection	12,288	49,152	96		
	d_embed	*	n_neurons	*	n_layers = 57,982,058,496
Unembedding	50,257	12,288			
	n_vocab	*	d_embed	=	617,558,016



3 参数数量分析

- GPT-3参数总量

Total weights: 175,181,291,520

Organized into 27,938 matrices



GPT-3

Embedding	$12,288 \times 50,257$ $d_embed * n_vocab$	$= 617,558,016$
Key	$128 \times 12,288 \times 96 \times 96$ $d_query * d_embed * n_heads * n_layers$	$= 14,495,514,624$
Query	$128 \times 12,288 \times 96 \times 96$ $d_query * d_embed * n_heads * n_layers$	$= 14,495,514,624$
Value	$128 \times 12,288 \times 96 \times 96$ $d_value * d_embed * n_heads * n_layers$	$= 14,495,514,624$
Output	$12,288 \times 128 \times 96 \times 96$ $d_embed * d_value * n_heads * n_layers$	$= 14,495,514,624$
Up-projection	$49,152 \times 12,288 \times 96$ $n_neurons * d_embed * n_layers$	$= 57,982,058,496$
Down-projection	$12,288 \times 49,152 \times 96$ $d_embed * n_neurons * n_layers$	$= 57,982,058,496$
Unembedding	$50,257 \times 12,288$ $n_vocab * d_embed$	$= 617,558,016$



3 参数数量分析

- GPT-3参数总量

