

转换器中的注意力

Attention in Transformer

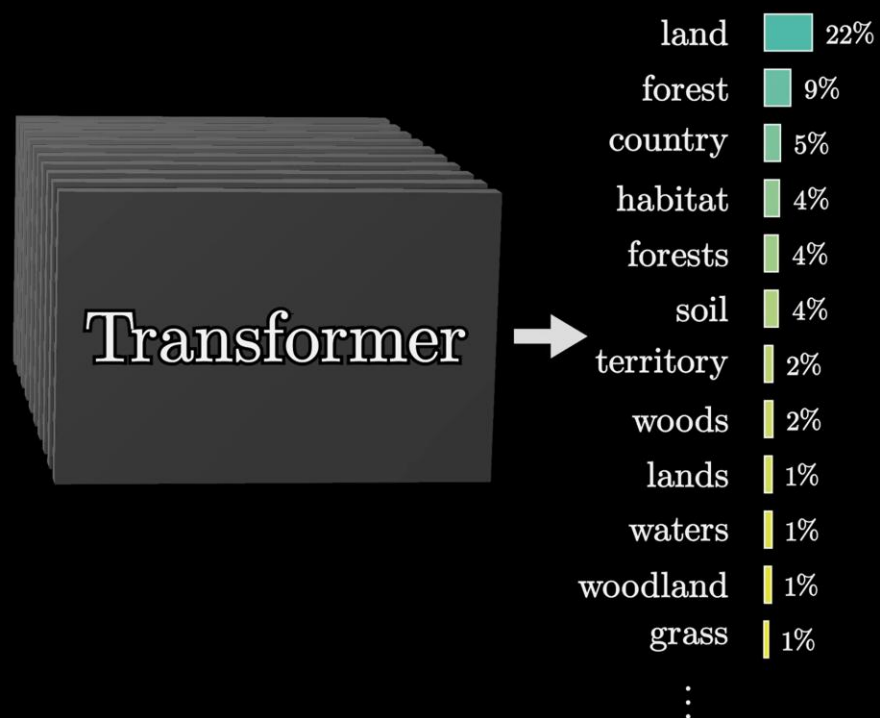
# 1. 回顾

2. 启发示例：注意力的作用
3. 注意力机制：查询（Query）、键（Key）
4. 掩码（Masking）
5. 注意力机制：值（Value）
6. 参数数量
7. 多头注意力
8. Transformer的深度

## 1.1 GPT工作方式

- 预测下一个词汇的概率分布

Behold, a wild pi creature,  
foraging in its native \_\_\_\_\_



## 1.2 词元 (token)

- 文本中，词元通常是单词、单词片段或常见字母组合

### A Convenient Lie

Let's pretend that tokens are always simply words

---

### The Truth

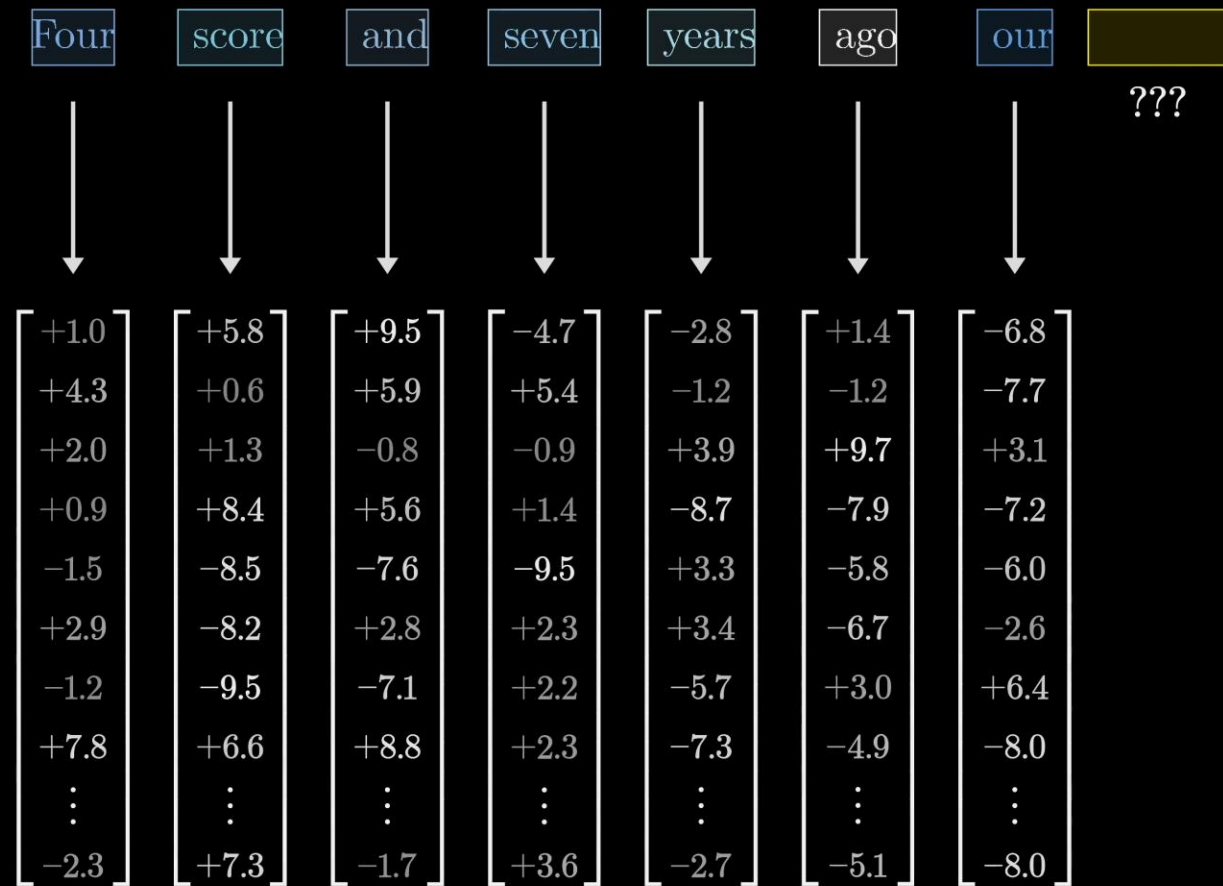
This process (known fancifully as tokenization) frequently subdivides words





## 1.3 词嵌入 (embedding)

- 将词编码量为向量

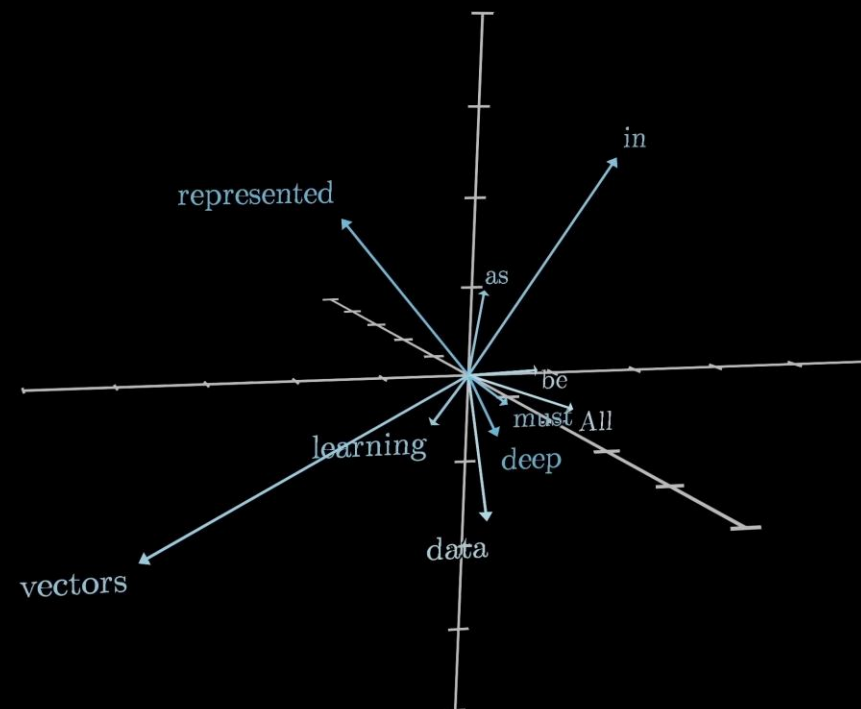


## 1.4 词嵌入的几何意义

- 词向量空间的方向具备语义上的含义

Words  $\xrightarrow{\text{"Embedding"}}$  Vectors

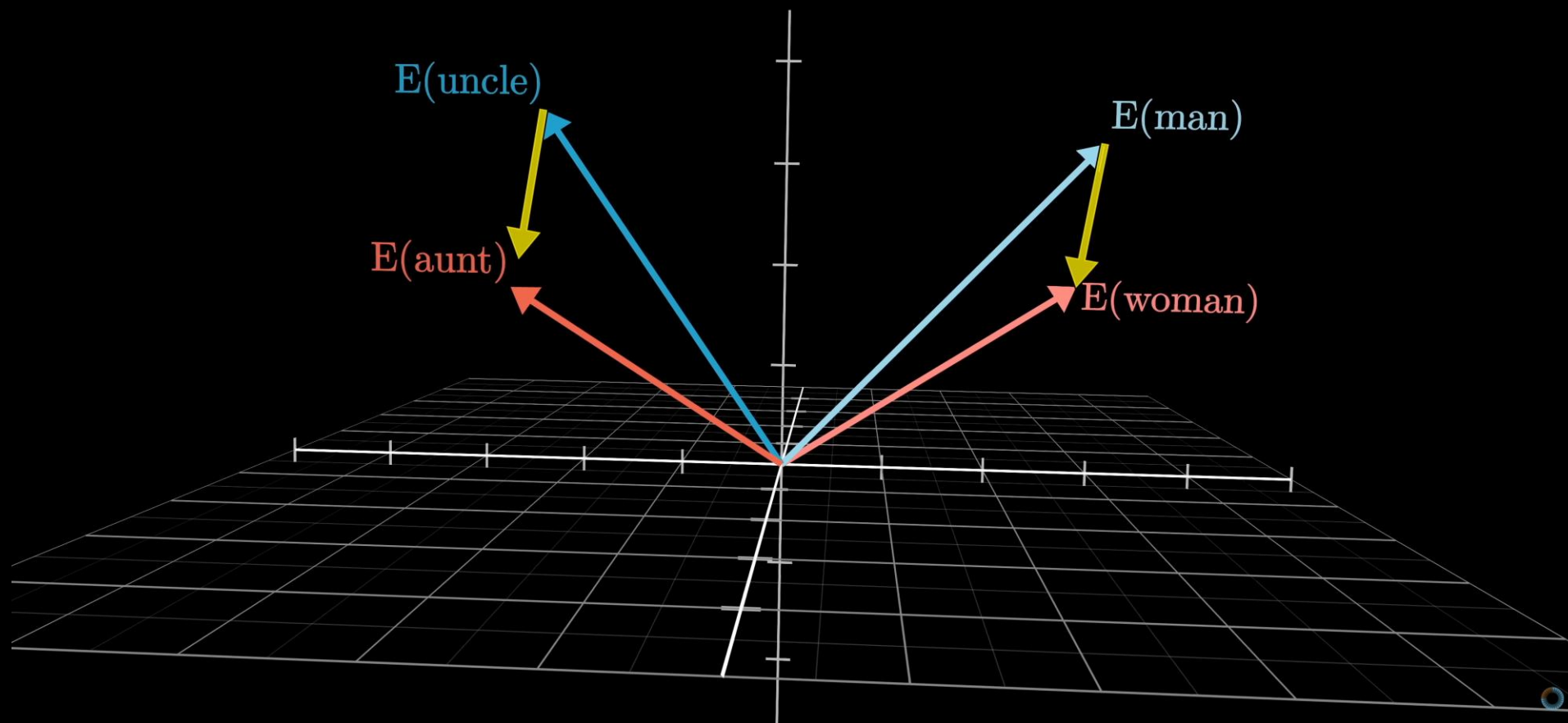
All  
data  
in  
deep  
learning  
must  
be  
represented  
as  
vectors



## 1.4 词嵌入的几何意义

- 词向量空间的方向具备语义上的含义

$$E(\text{aunt}) - E(\text{uncle}) \approx E(\text{woman}) - E(\text{man})$$



1. 回顾
- 2. 启发示例：注意力的作用**
3. 注意力机制：查询（Query）、键（Key）
4. 掩码（Masking）
5. 注意力机制：值（Value）
6. 参数数量
7. 多头注意力
8. Transformer的深度

## 2.1 例1: 多义词

- 同一个词的嵌入向量是一样的, 没有考虑上下文

北美鼯(qú)鼯鼠



American shrew mole

一摩尔二氧化碳

$6.02 \times 10^{23}$   
One mole of carbon dioxide

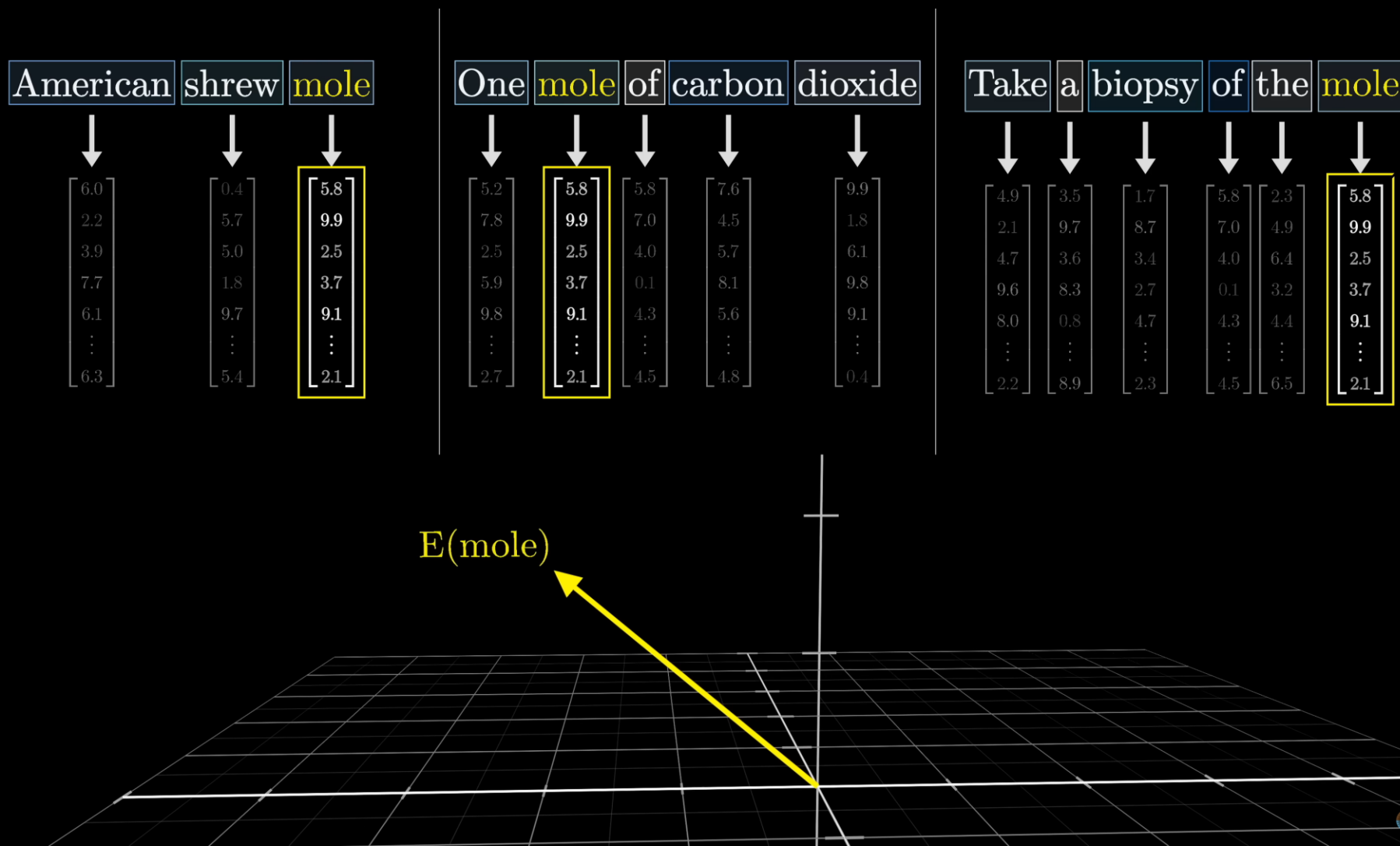
对痣进行活组织检查



Take a biopsy of the mole

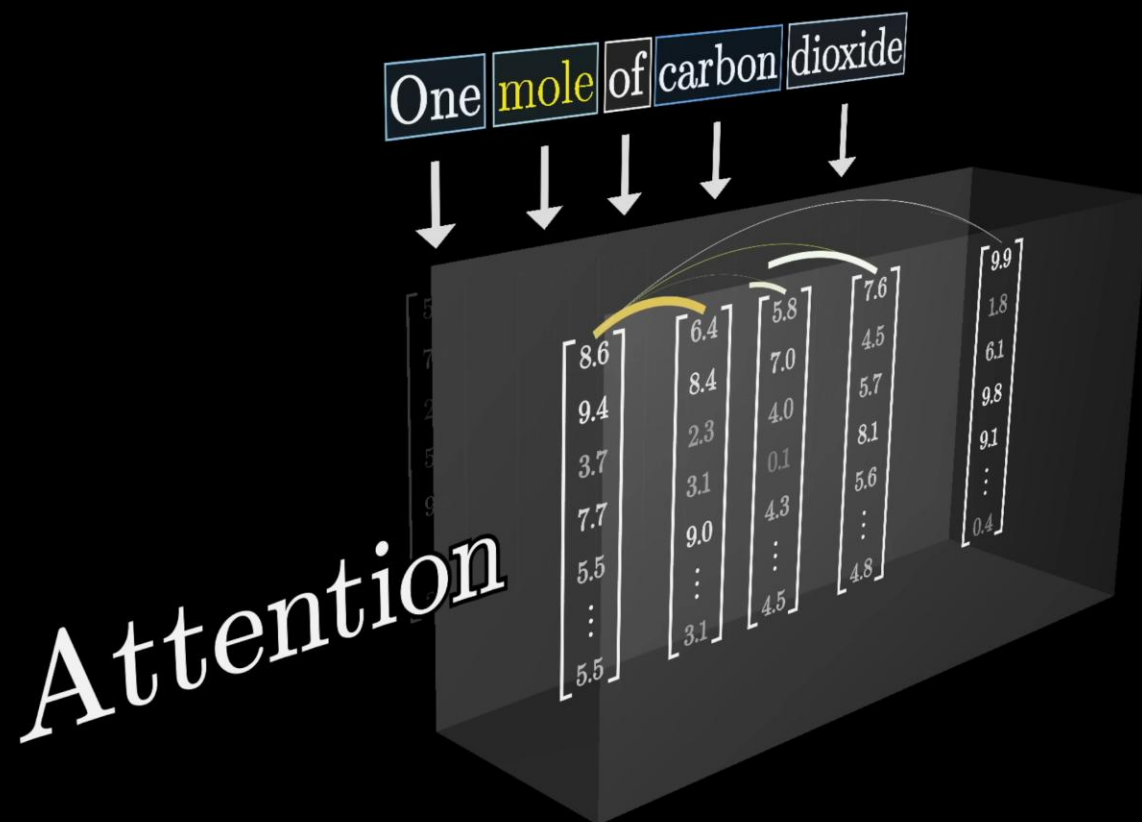
## 2.1 例1：多义词

- 同一个词的嵌入向量是一样的，没有考虑上下文



## 2.1 例1：多义词

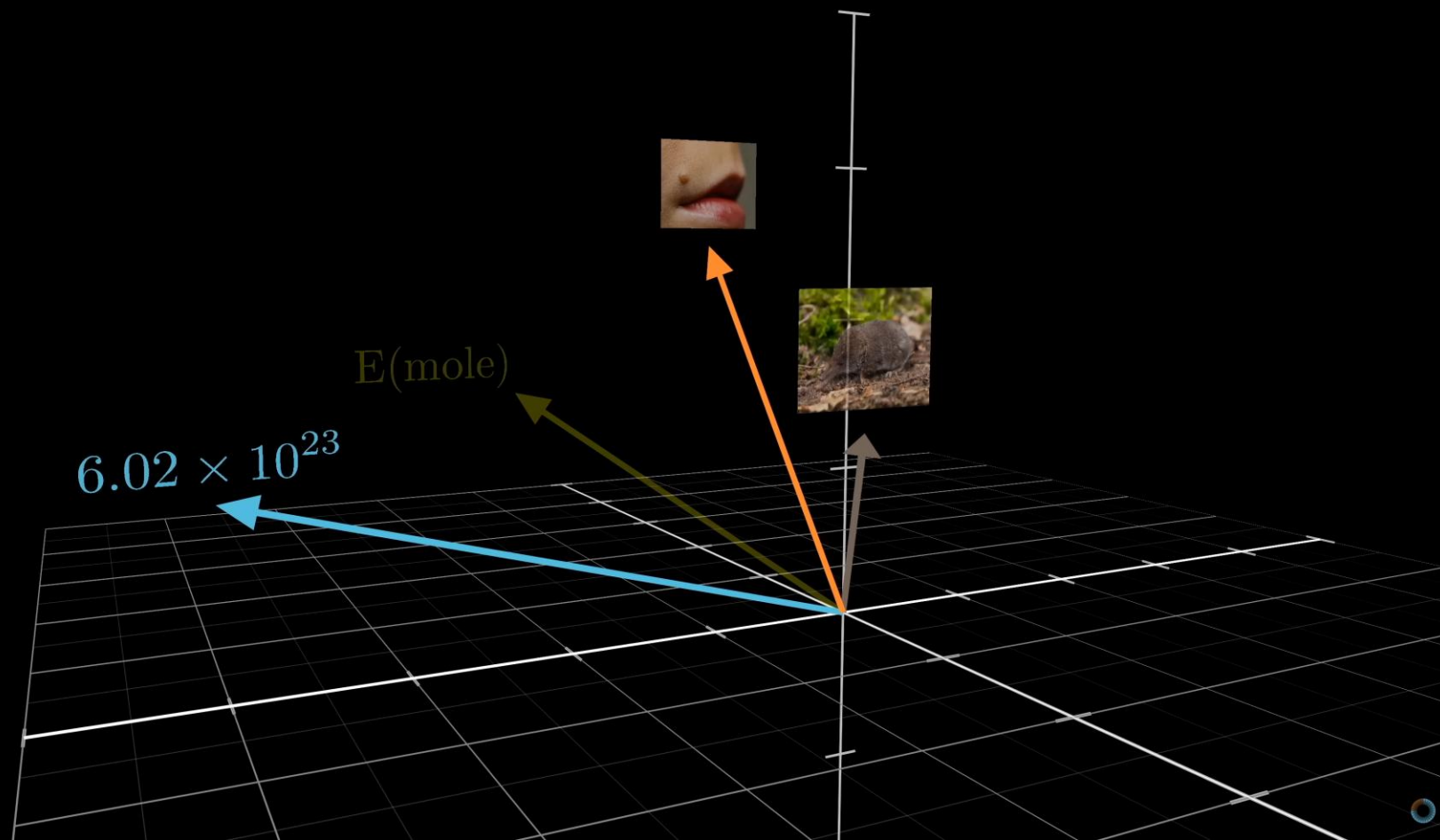
- 注意力机制可以利用上下文信息，来调整向量



## 2.1 例1: 多义词

- 几何意义上, 初始词向量 + 增量 = 指向所需方向

mole

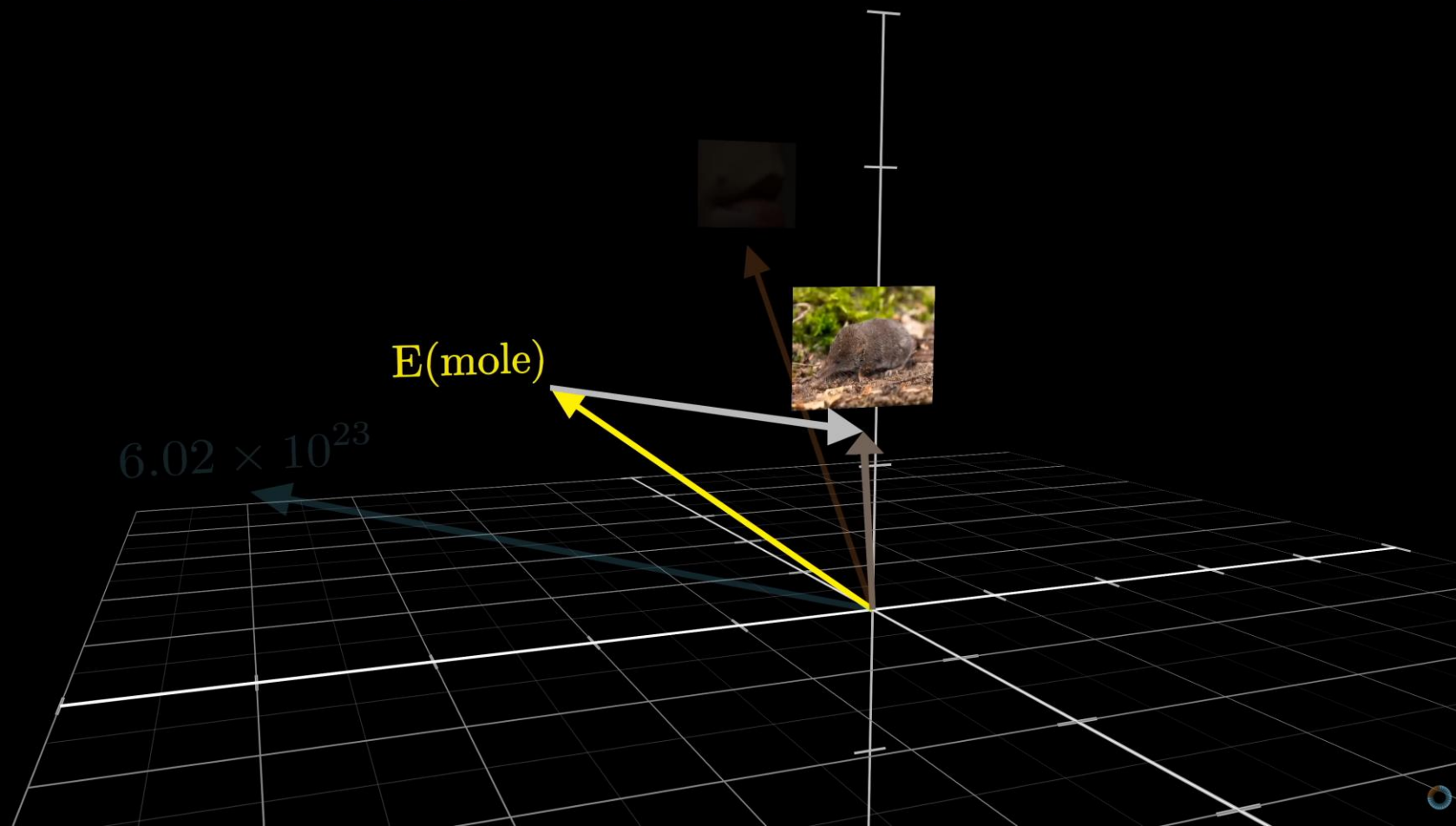




## 2.1 例1：多义词

- 几何意义上，注意力机制将计算出所需要的增量

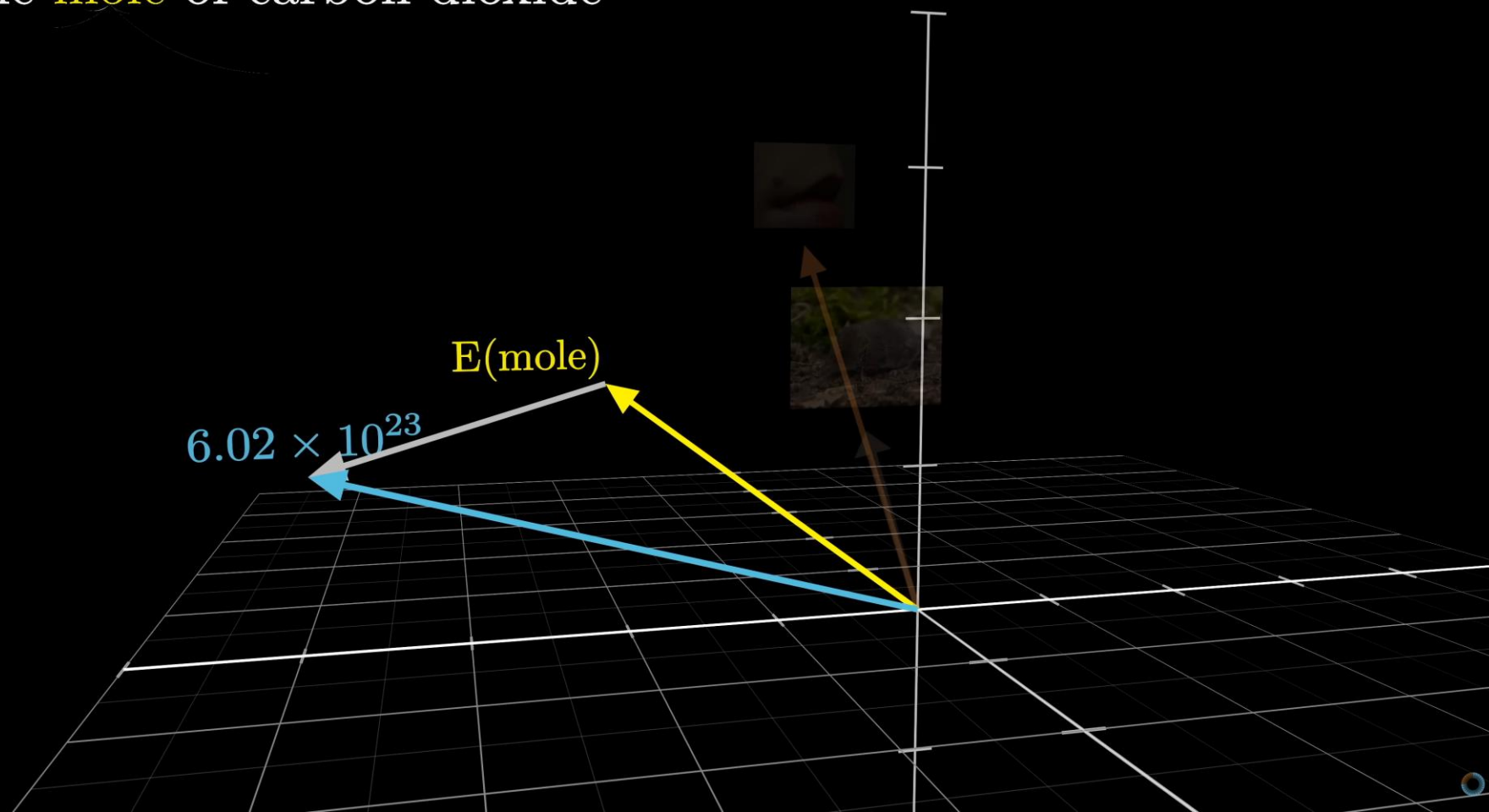
American shrew **mole**



## 2.1 例1: 多义词

- 几何意义上, 注意力机制将计算出所需要的增量

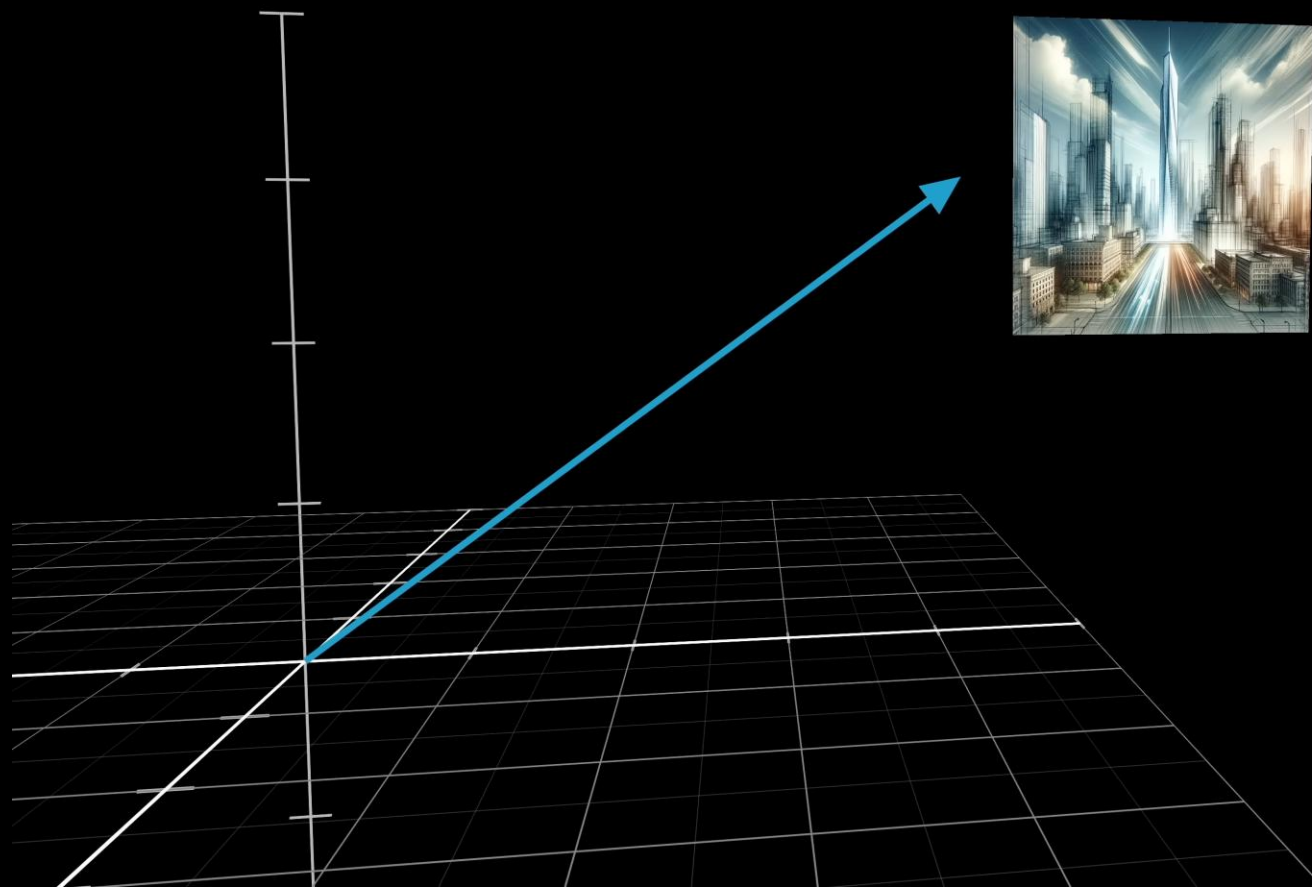
One **mole** of carbon dioxide



## 2.2 例2：修饰

- 注意力机制根据上下文更新向量

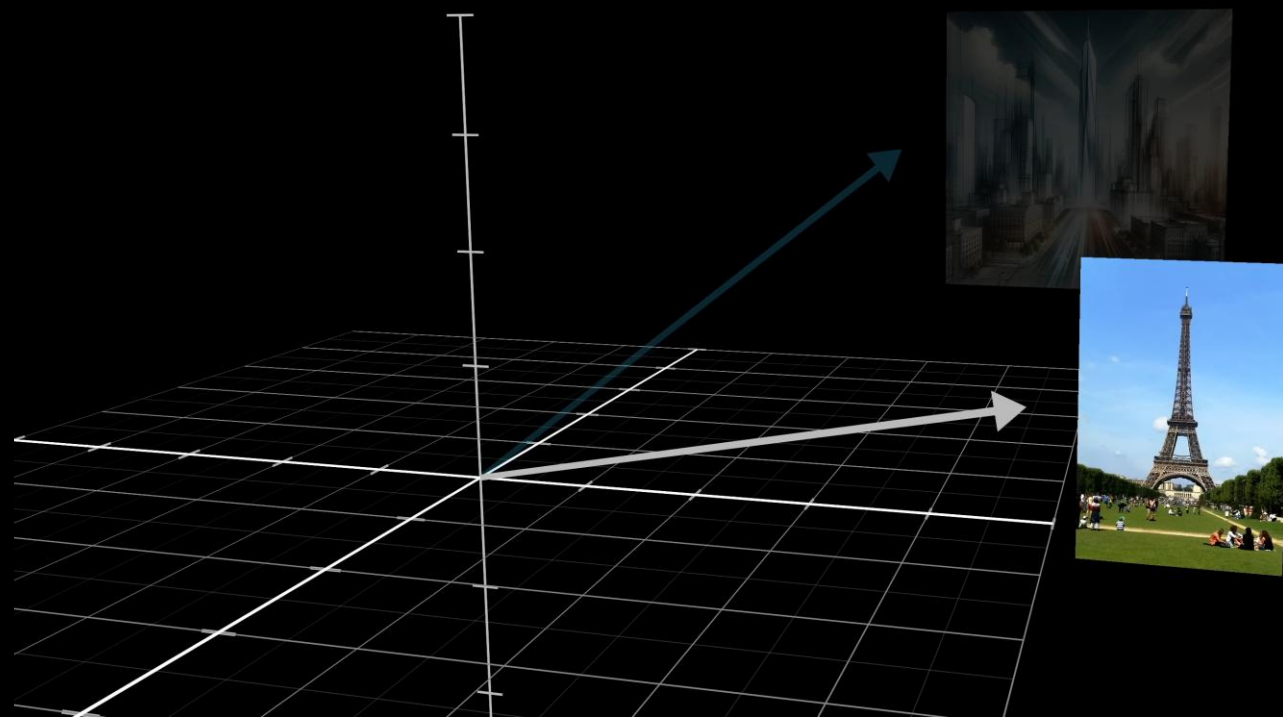
Tower



## 2.2 例2：修饰

- 如果“tower”前面出现了“Eiffel”，那么注意力机制应该更新向量，使其与“Eiffel tower”、“Paris”和“France”等相关向量更相关

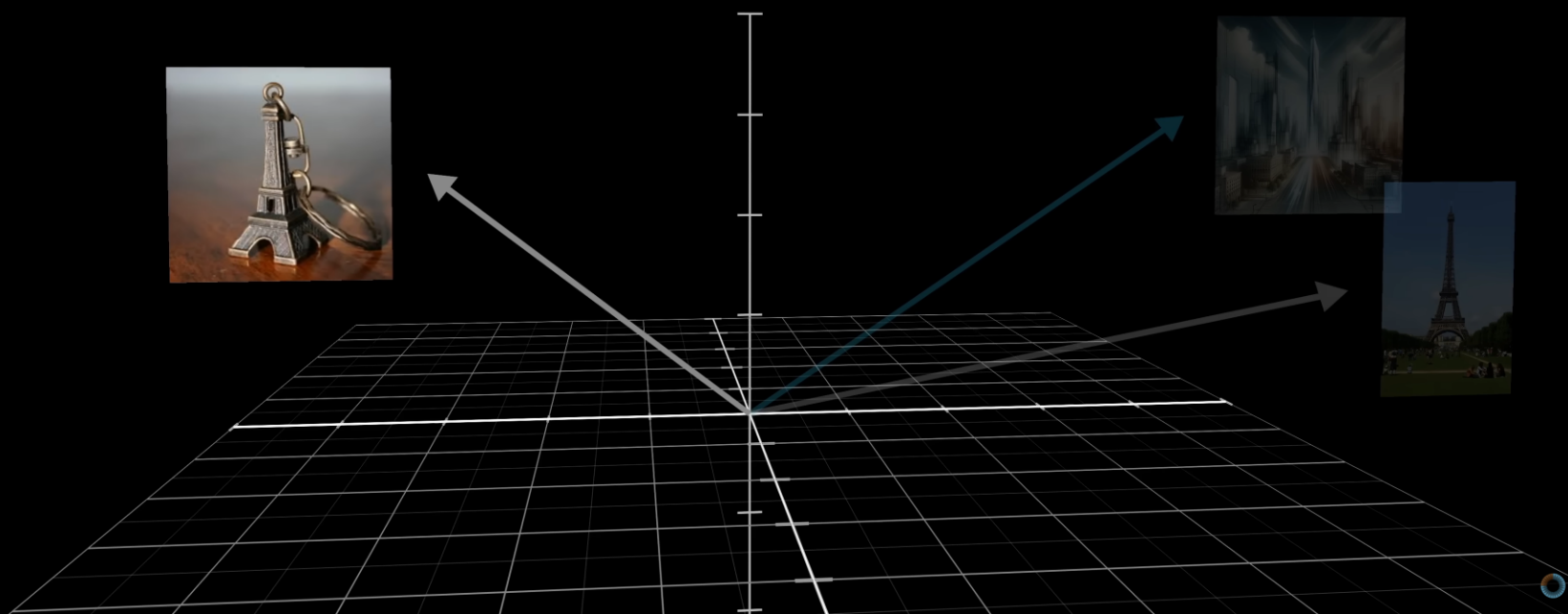
### Eiffel Tower



## 2.2 例2：修饰

- 如果“tower”前面出现了“miniature”，那么向量应该进一步更新，不再与“large”和“tall”等概念相关

### Miniature Eiffel Tower



## 2.3 例3：长上下文

- 例如在长篇推理小说中，最后要猜测谁是凶手

### Short distance

American shrew mole

### Long distance

Harry Potter was a highly unusual boy in many ways. For one thing, he hated the summer holidays more than any other time of year. For another, he really wanted to do his homework but was forced to do it in secret, in the dead of night. And he also happened to be a wizard.

It was nearly midnight, and he was lying on his stomach in bed, the blankets drawn right over his head like a tent, a flashlight in one hand and a large leather-bound book (A History of Magic by Bathilda Bagshot) propped open against the pillow. Harry moved the tip of his eagle-feather quill down the page, frowning as he looked for something that would help him write his essay, "Witch Burning in the Fourteenth Century Was Completely Pointless discuss."

The quill paused at the top of a likely-looking paragraph. Harry Pushed his round glasses up the bridge of his nose, moved his flashlight closer to the book, and read:

Non-magic people (more commonly known as Muggles) were particularly afraid of magic in medieval times, but not very good at recognizing it. On the rare occasion that they did catch a real witch/or wizard, burning had no effect whatsoever. The witch or wizard would perform a basic Flame Freezing Charm and then pretend to shriek with pain while enjoying a gentle, tickling sensation. Indeed, Wendelin the Weird enjoyed being burned so much that she allowed herself to be caught no less than fortyseven times in various disguises.

Harry put his quill between his teeth and reached underneath his pillow for his ink bottle and a roll of parchment. Slowly and very carefully he unscrewed the ink bottle, dipped his quill into it, and began to write, pausing every now and then to listen, because if any of the Dursleys heard the scratching of his quill on their way to the bathroom, he'd probably find himself locked in the cupboard under the stairs for the rest of the summer.

The Dursley family of number four, Privet Drive, was the reason that Harry never enjoyed his summer holidays. Uncle Vernon, Aunt Petunia, and their son, Dudley, were Harry's only living relatives. They were Muggles, and they had a very medieval attitude toward magic. Harry's dead parents, who had been a witch and wizard themselves, were never mentioned under the Dursleys' roof. For years, Aunt Petunia and Uncle Vernon had hoped that if they kept Harry as downtrodden as possible, they would be able to squash the magic out of him. To their fury, they had been unsuccessful. These days they lived in terror of anyone finding out that Harry had spent most of the last two years at Hogwarts School of Witchcraft and Wizardry. The most they could do, however, was to lock away Harry's spellbooks, wand, cauldron, and broomstick at the start of the summer break, and forbid him to talk to the neighbors.

This separation from his spellbooks had been a real problem for Harry, because his teachers at Hogwarts had given him a lot of holiday work. One of the essays, a particularly nasty one about shrinking potions, was for Harry's least favorite teacher, Professor

## 2.3 例3：长上下文

- 不仅需要前几个词，而是结合整篇文章的全部信息

"Each of you has what on the surface appears an unassailable alibi." she said, as the storm cleared and the crimson rays of dusk poured through the window where everyone was gathered. "But only one of you could have known about the loose screw on the window, while also having known where the second key was hidden. I am left, inescapably then, to the conclusion that therefore, the murderer was ???



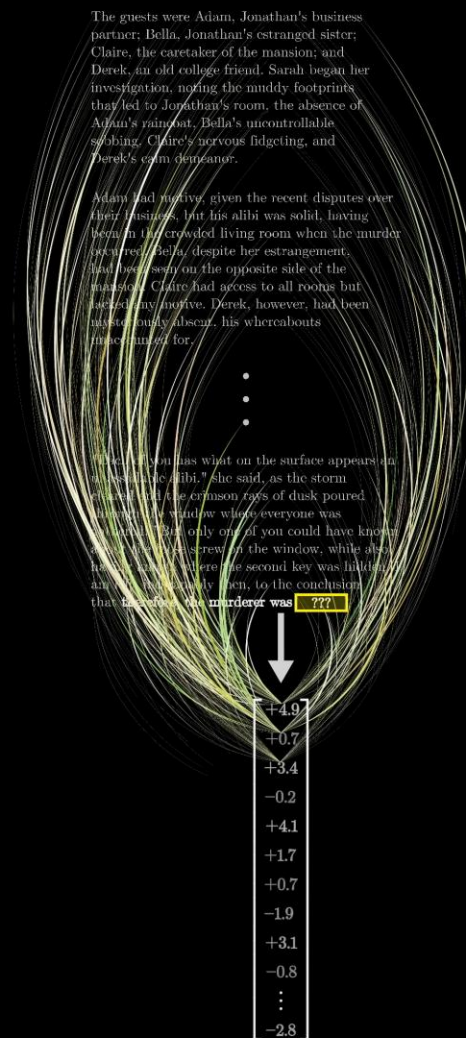
$\begin{bmatrix} +2.2 \\ -6.2 \\ +5.1 \\ -8.3 \\ +8.8 \\ -0.6 \\ -3.2 \\ -7.6 \\ -1.3 \\ +3.2 \\ \vdots \\ -8.0 \end{bmatrix}$





## 2.3 例3：长上下文

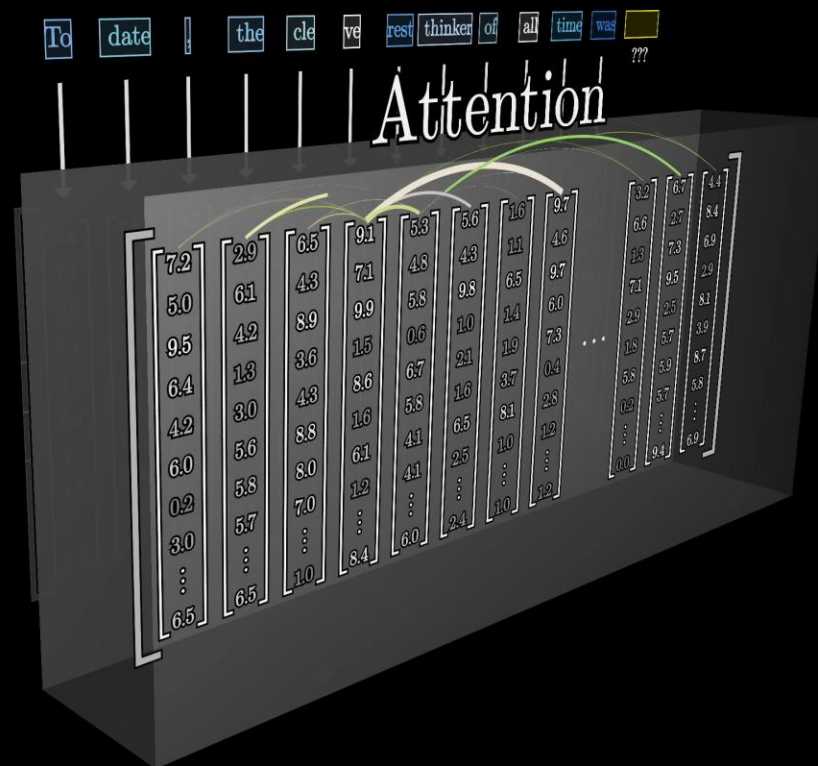
- 不仅需要前几个词，而是结合整篇文章的全部信息





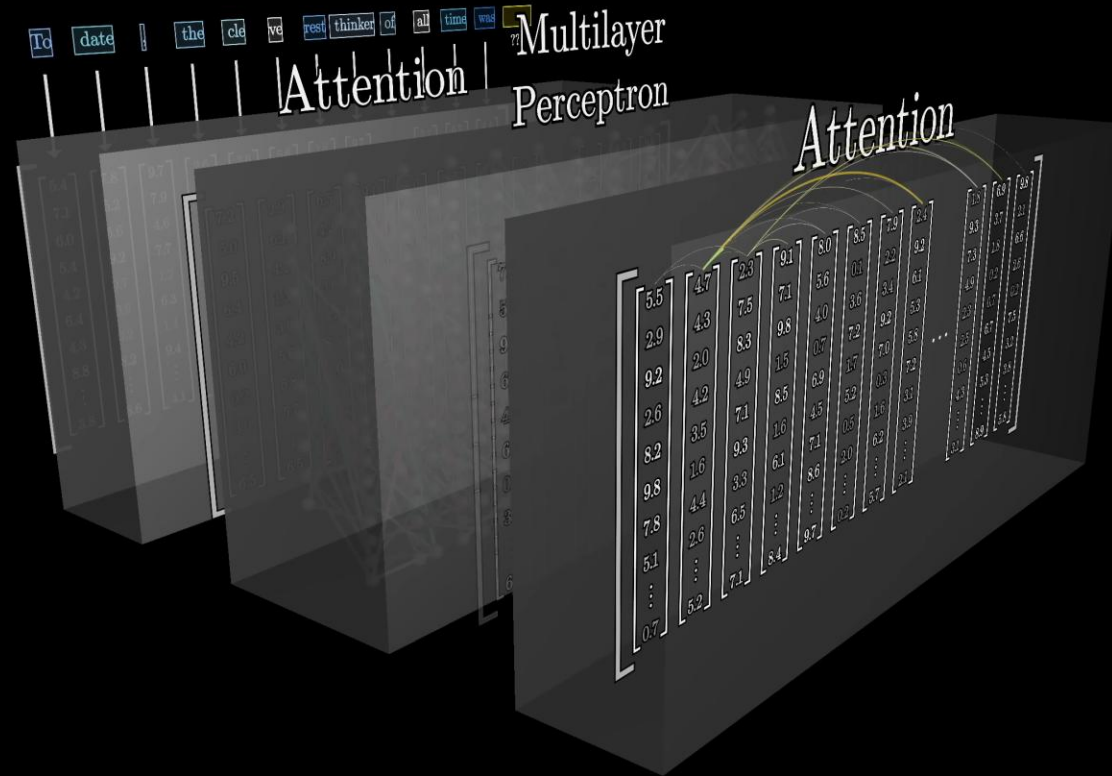
## 2.3 例3：长上下文

- 注意力机制和深层网络让信息在词向量之间传递



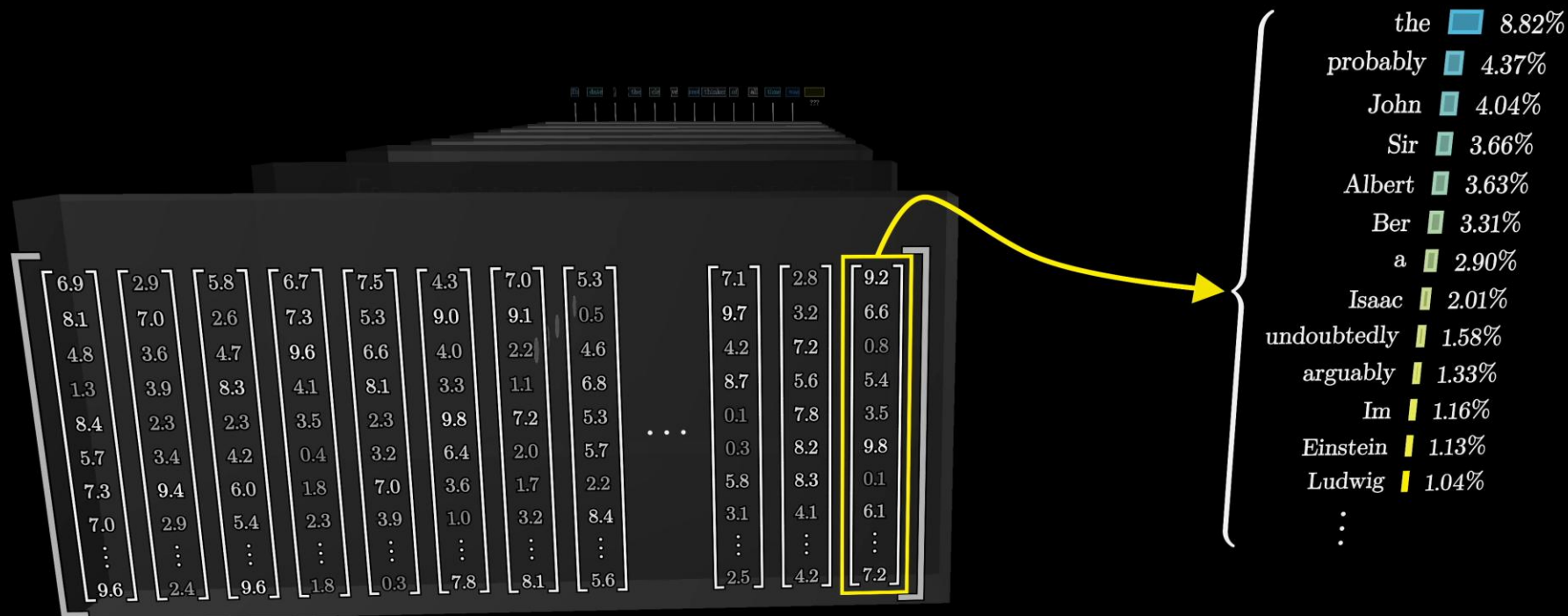
## 2.3 例3：长上下文

- 从而获取更长、更远、更丰富的上下文信息



## 2.3 例3：长上下文

- 这些信息最终将汇聚到最后一个词上，完成对“凶手”的预测



1. 回顾
2. 启发示例：注意力的作用
- 3. 注意力机制：查询（Query）、键（Key）**
4. 掩码（Masking）
5. 注意力机制：值（Value）
6. 参数数量
7. 多头注意力
8. Transformer的深度

### 3.1 注意力机制的简化示例

a fluffy blue creature roamed the verdant forest

一只 毛茸茸的 蓝色 生物 漫步于 那 郁郁葱葱的 森林



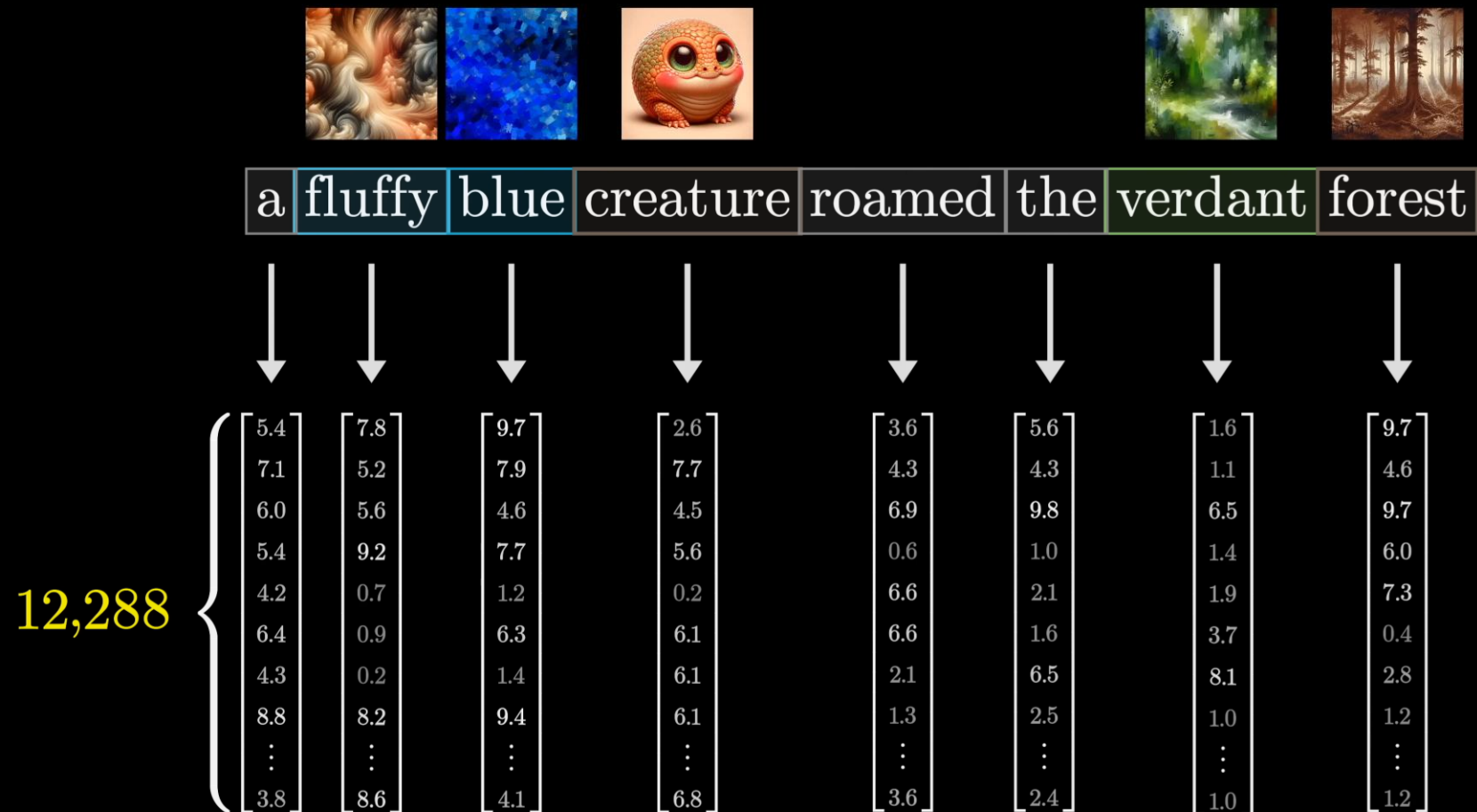
### 3.1 注意力机制的启发性示例

- 假设目前我们只关心形容词对名词的更新



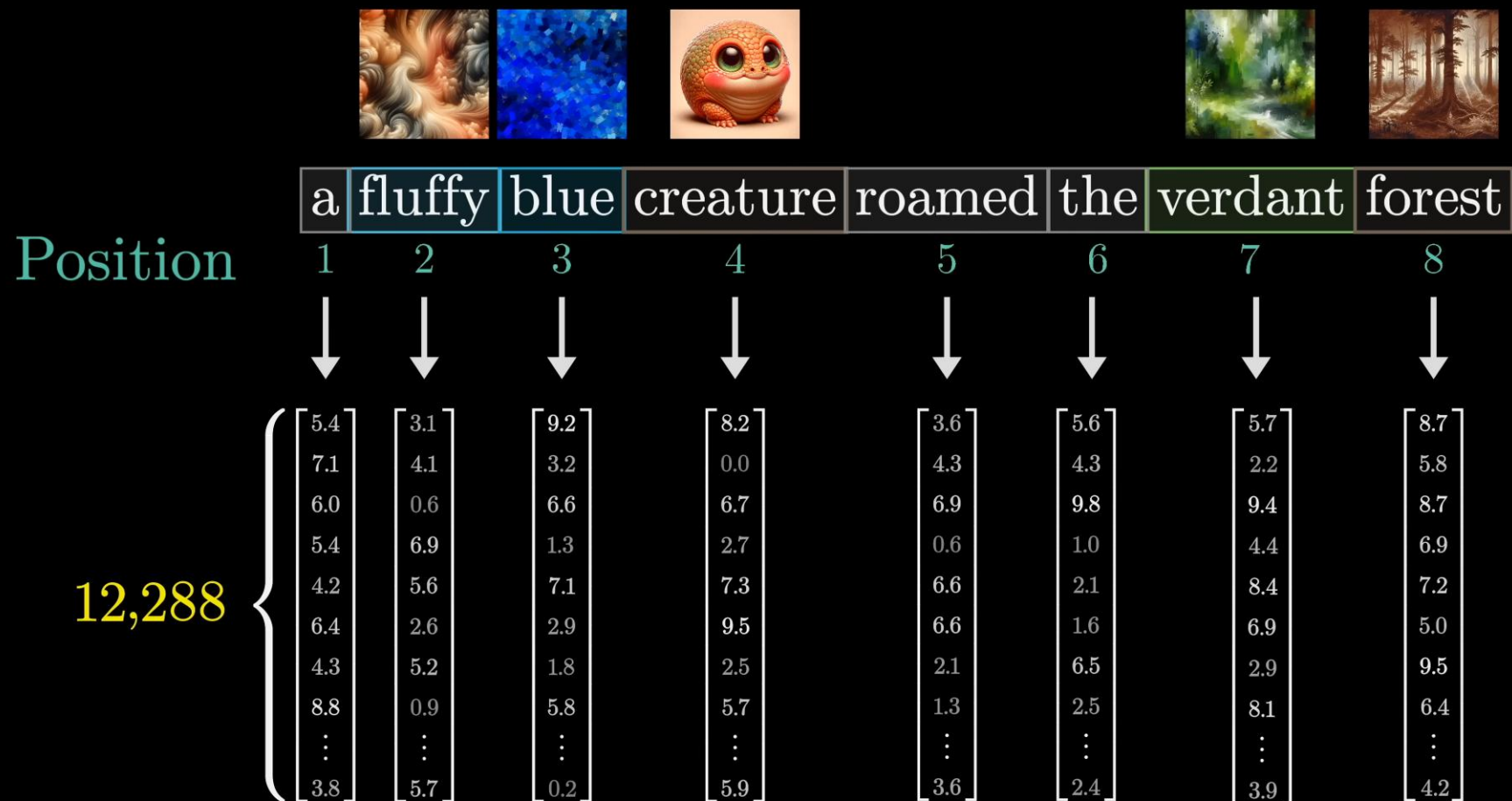
### 3.1 注意力机制的启发性示例

- 词嵌入只考虑单个词元



### 3.1 注意力机制的启发性示例

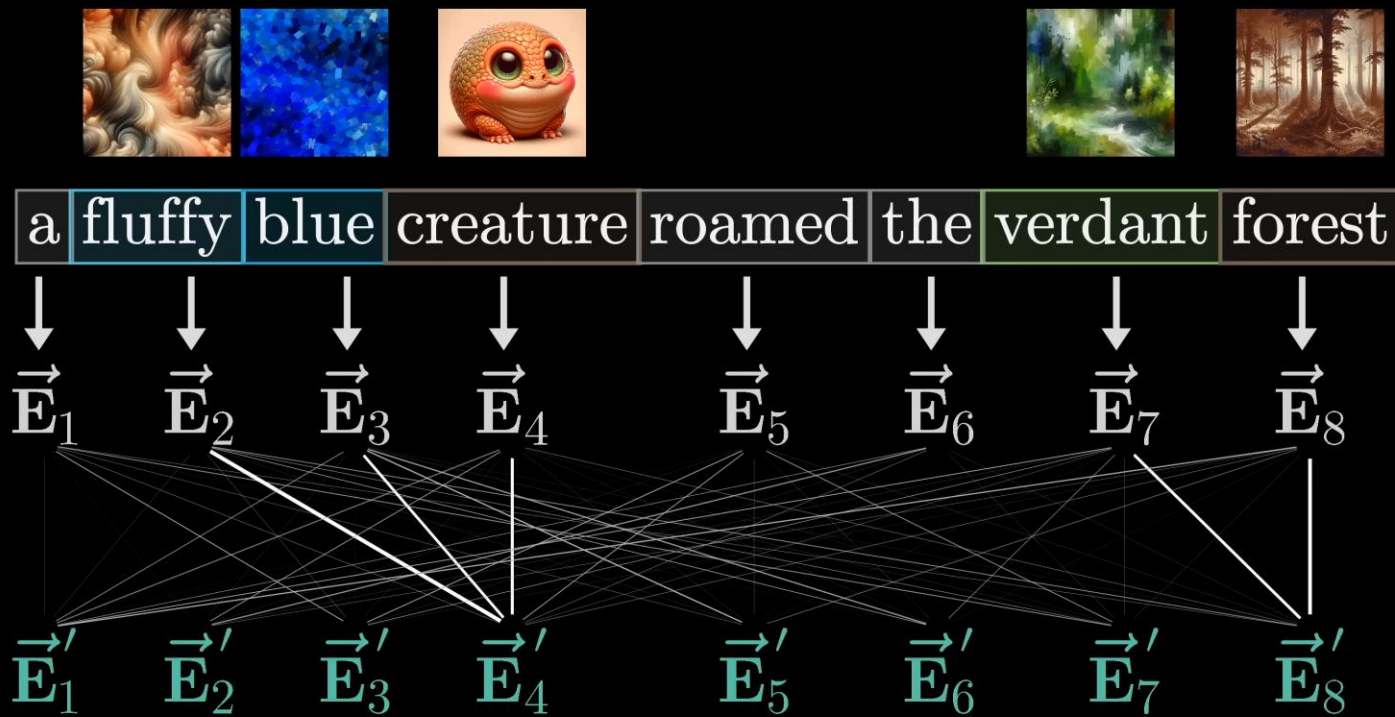
- 同时，位置信息也被编入了嵌入向量





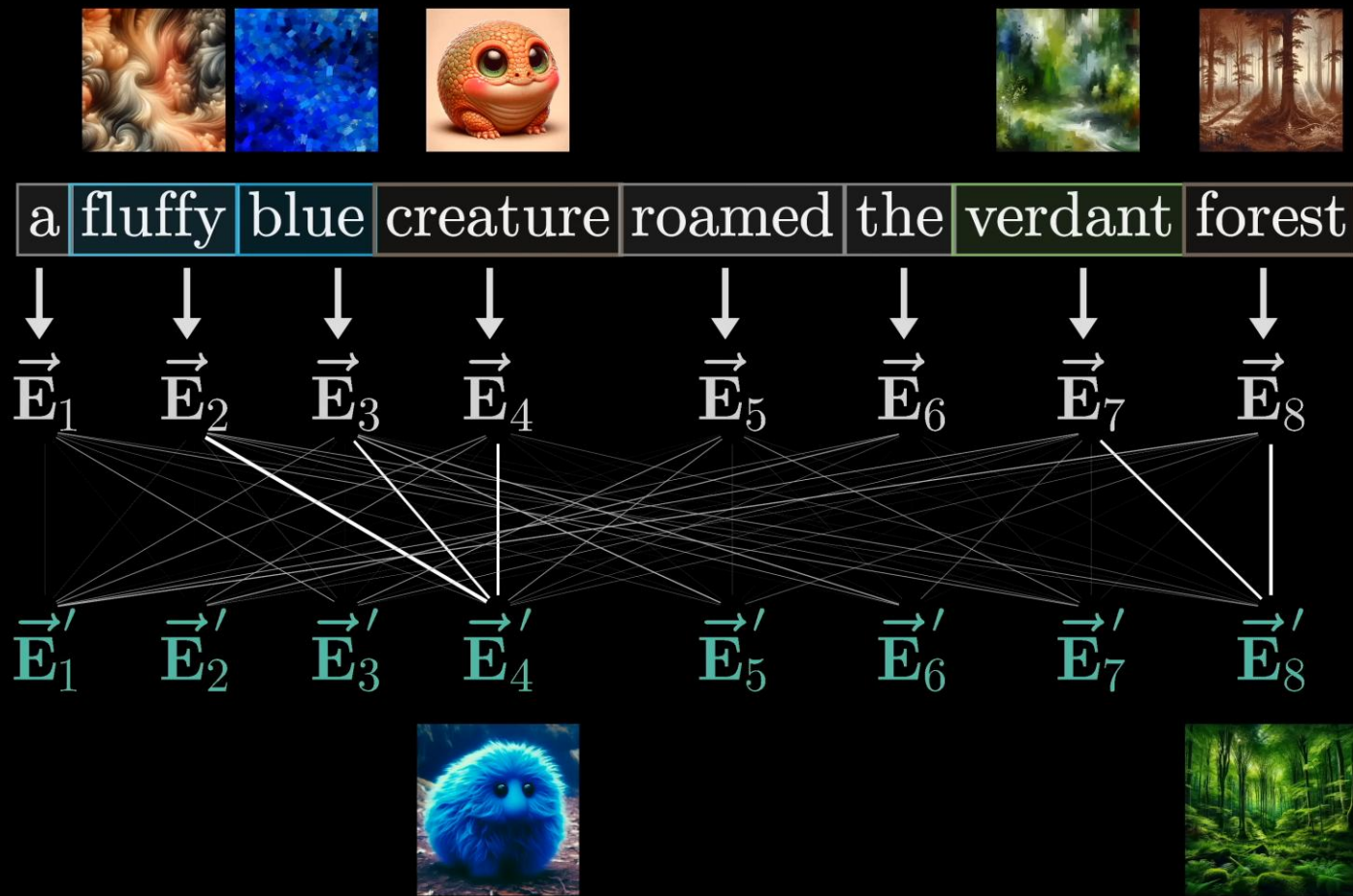
### 3.1 注意力机制的启发性示例

- 注意力的目标是通过计算，获得嵌入向量的更新



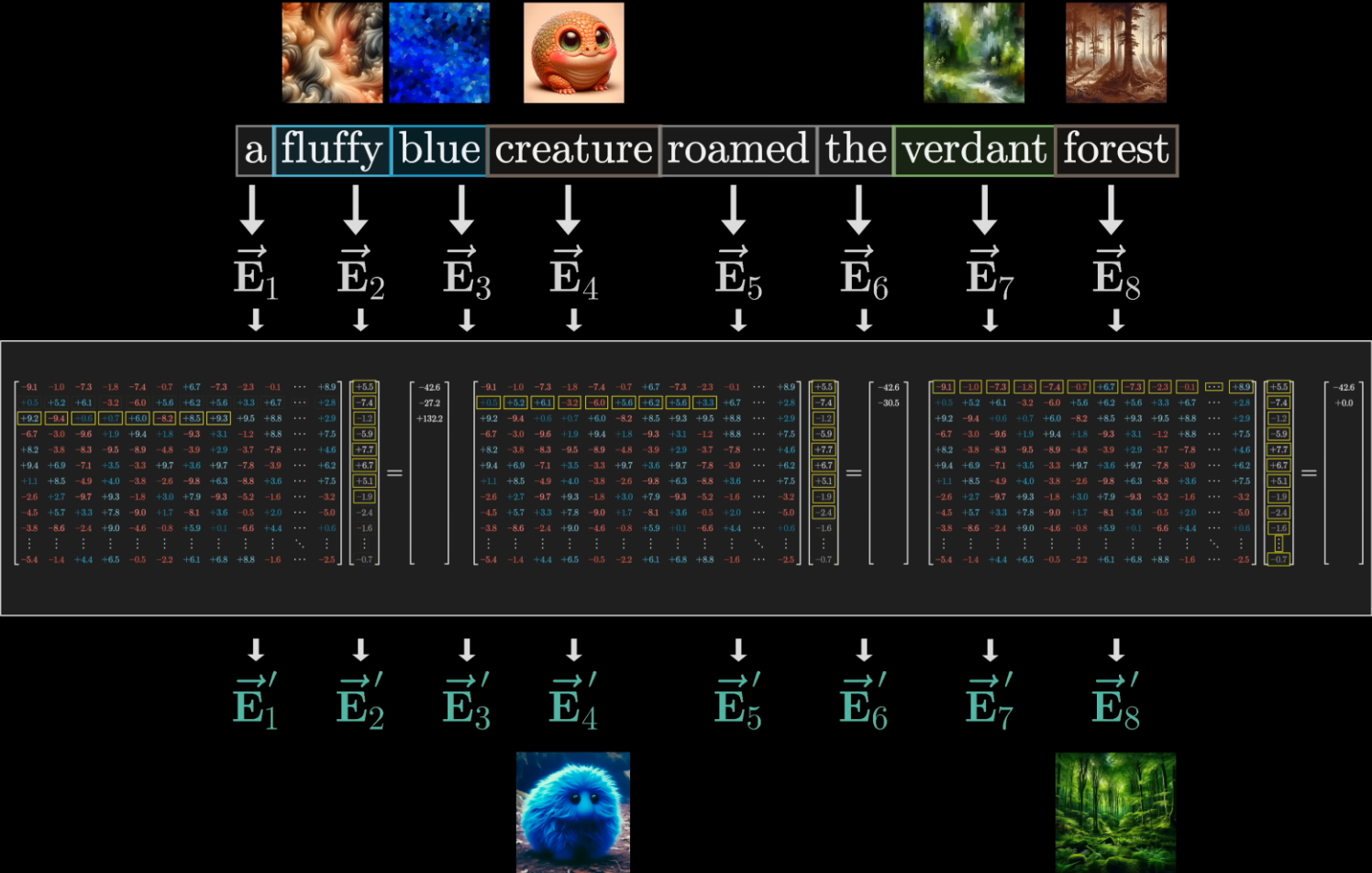
### 3.1 注意力机制的启发性示例

- 此例中，即为被修饰的名词



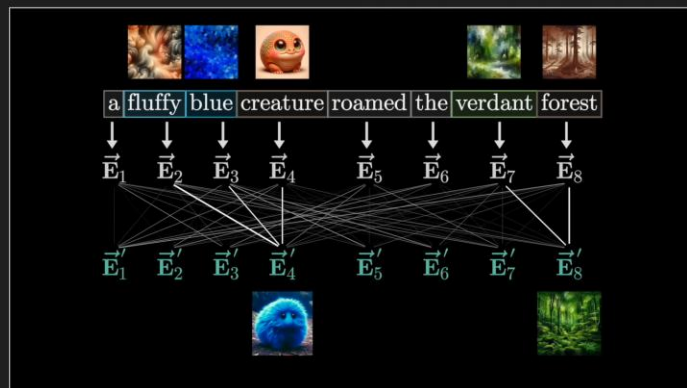
### 3.1 注意力机制的启发性示例

- 深度学习中，我们希望采用矩阵计算来实现这些功能
- 这些矩阵包含的是可调节的参数，即为“权重”矩阵

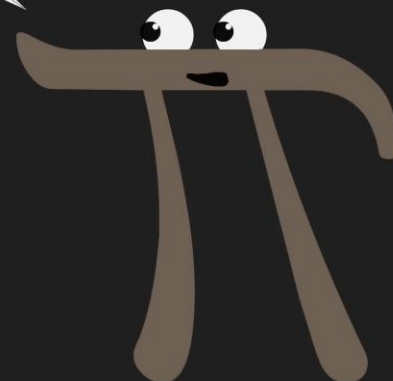
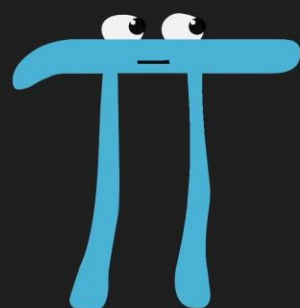
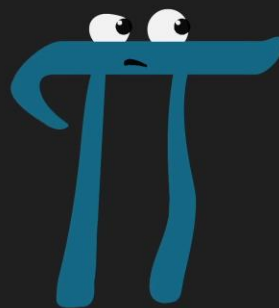
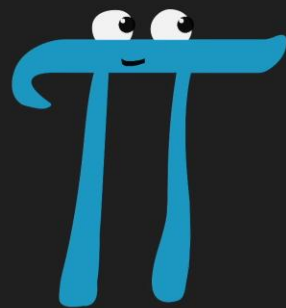


### 3.1 注意力机制的启发性示例

- 再次强调，这个例子是一个虚构的启发性示例
- 大模型真正的可解释性被掩盖在海量的参数和计算中

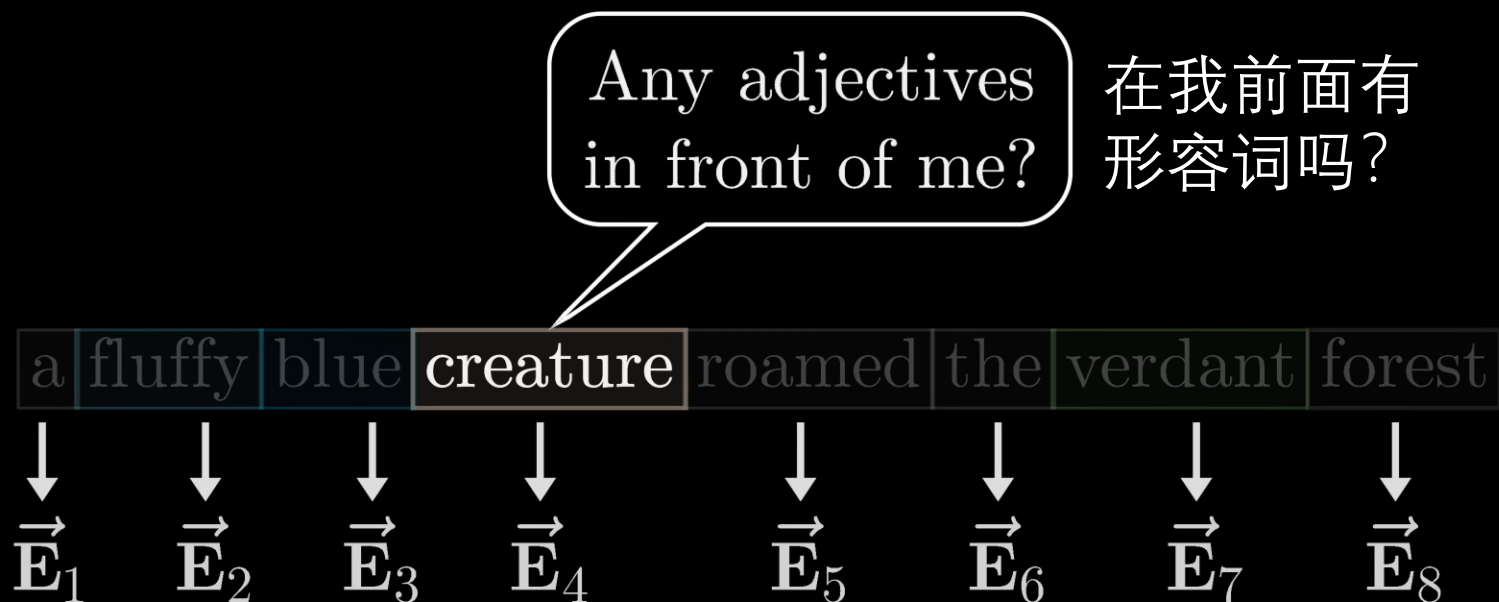


This is a made-up  
motivating example



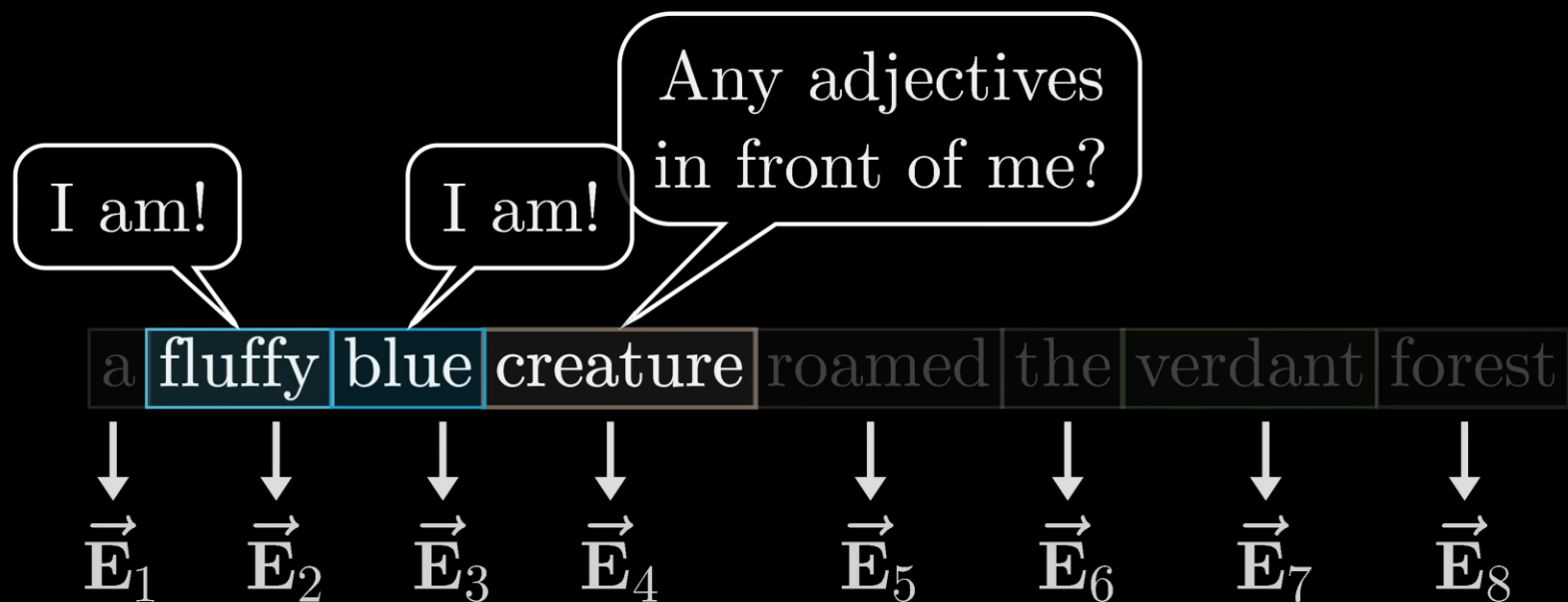
## 3.2 查询 (Query)

- 首先，名词问出了一个“问题”：查询 (Query, 简称为Q)



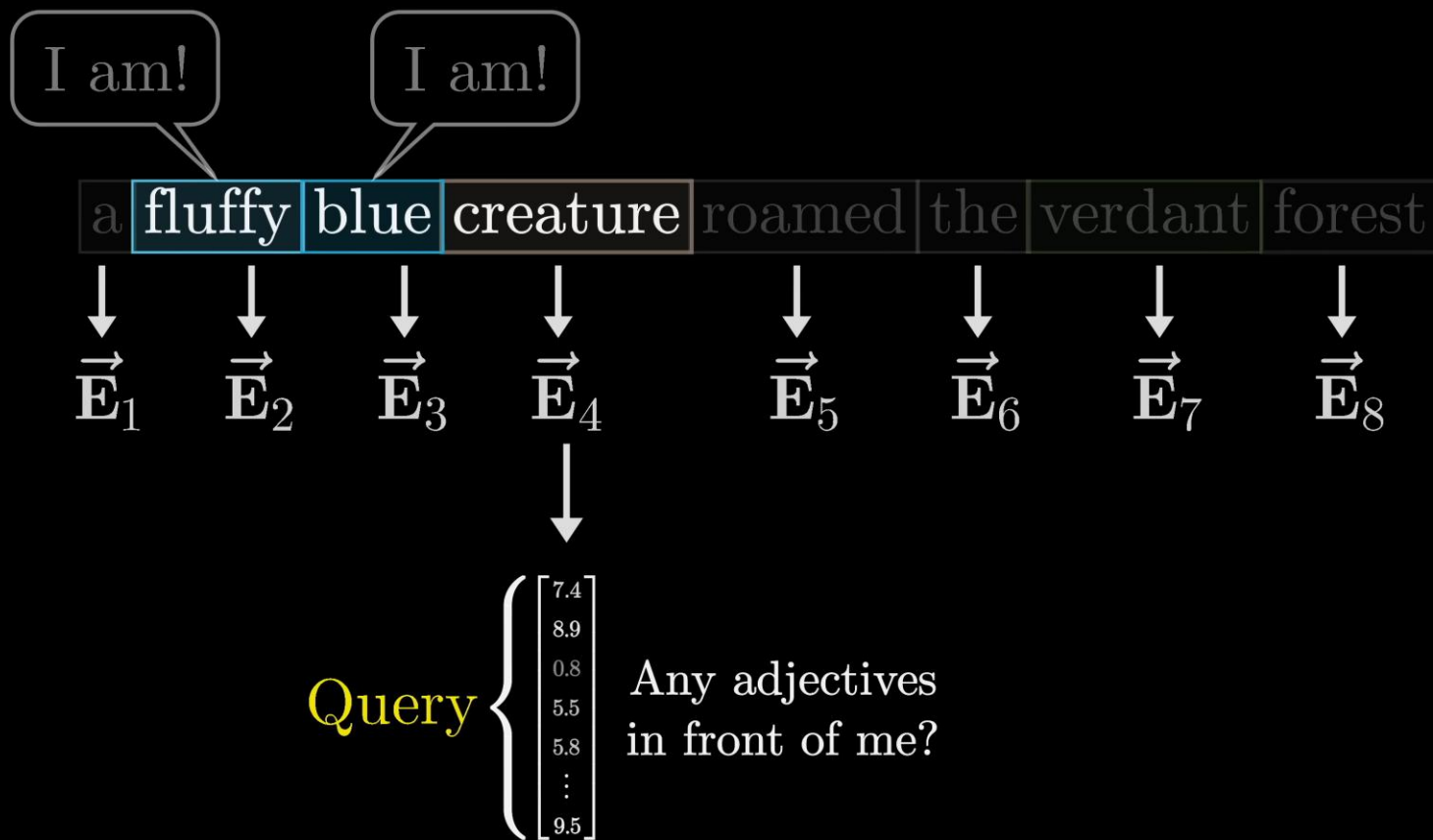
## 3.2 查询

- 接着，形容词给出了“回答”：键（Key，简写为K）



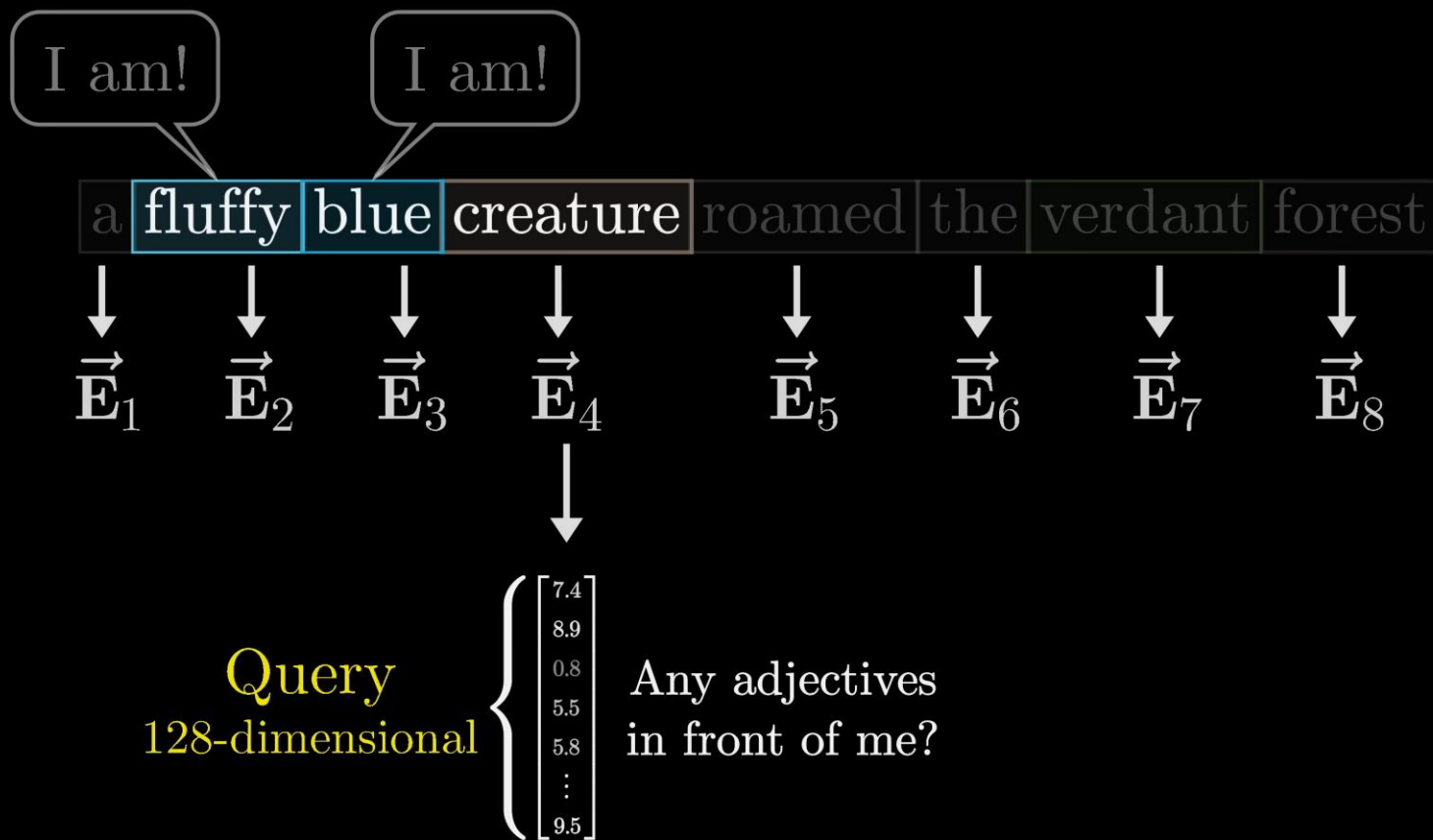
## 3.2 查询

- “问题”被编码为了一个向量 $Q_i$



## 3.2 查询

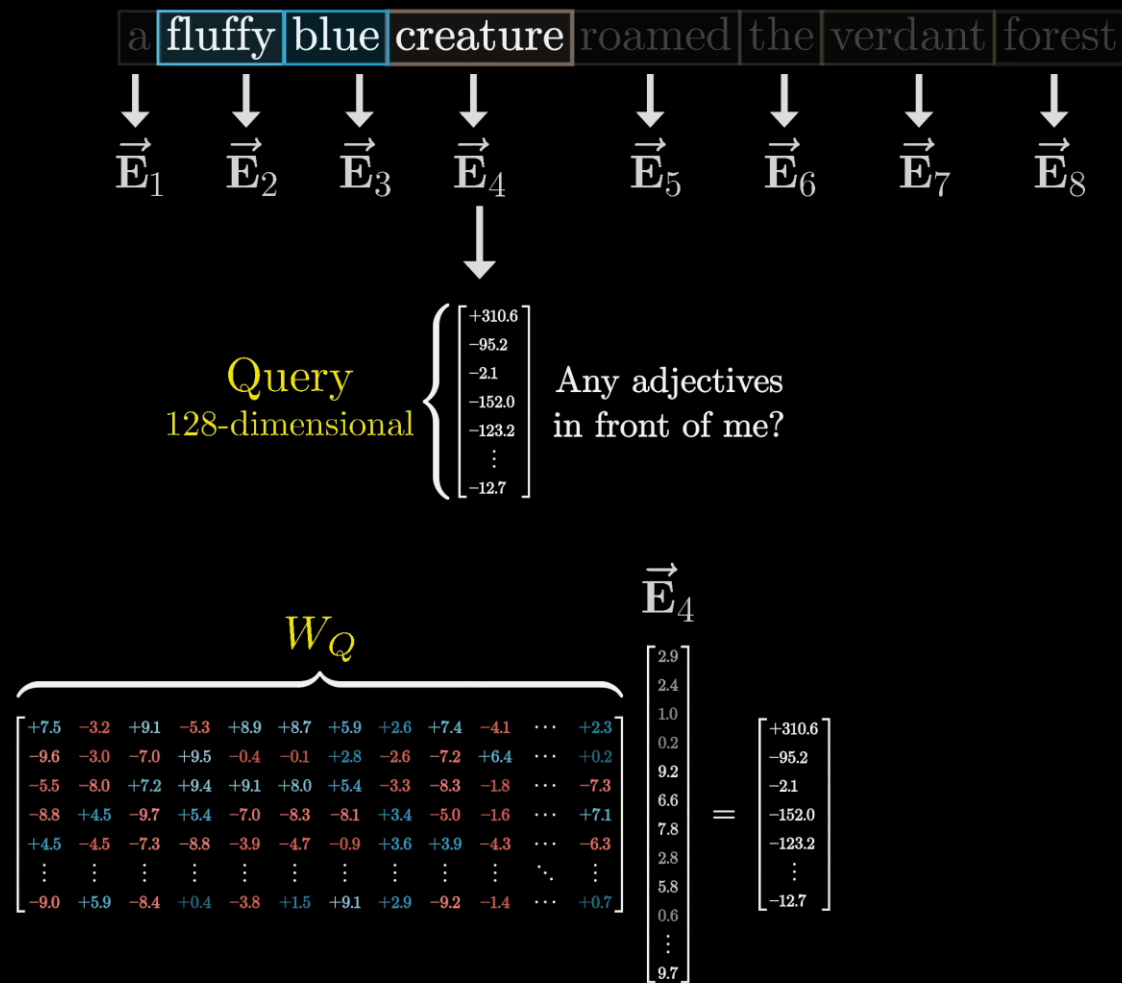
- “问题”被编码为了一个向量 $Q_i$





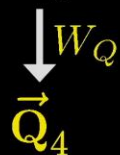
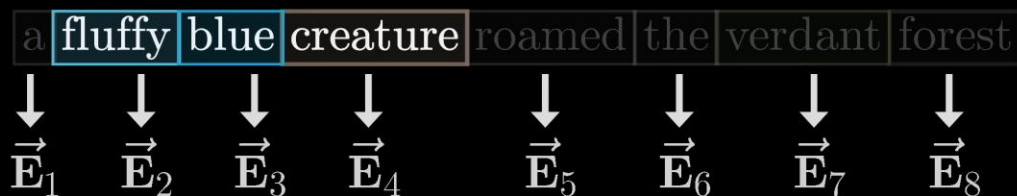
## 3.2 查询

- $Q_i$  向量由一个权重矩阵  $W_Q$  与嵌入向量  $E_i$  相乘得到



## 3.2 查询

- $Q_i$  向量由一个权重矩阵  $W_Q$  与嵌入向量  $E_i$  相乘得到

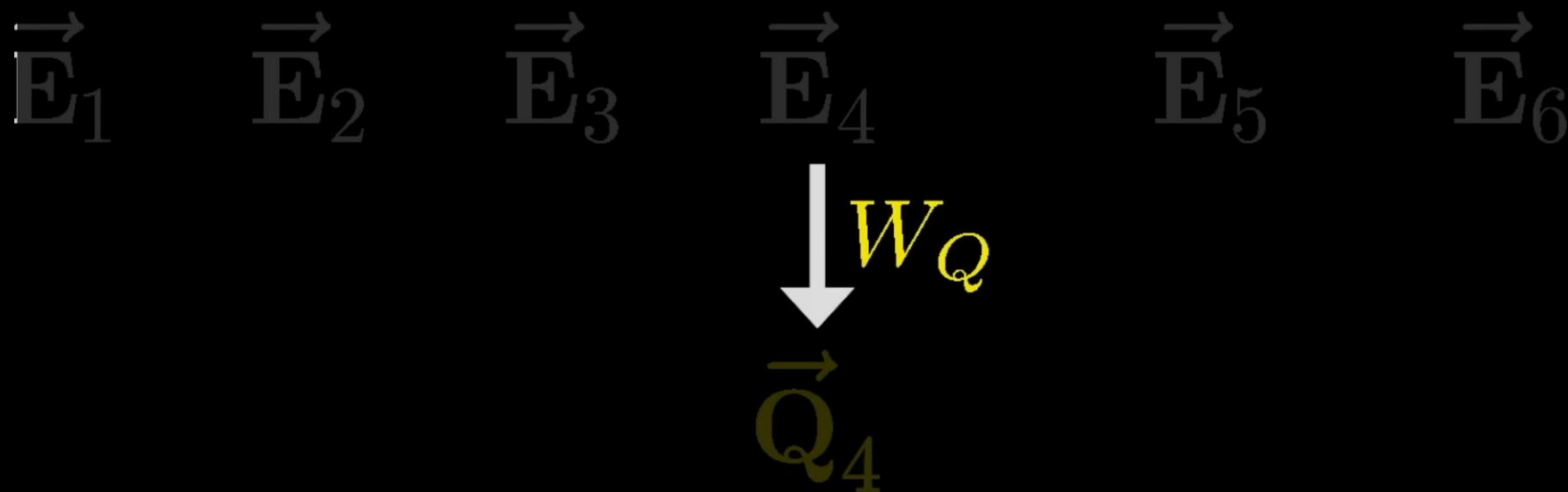


Any adjectives  
in front of me?

$$\overbrace{\begin{bmatrix} +7.5 & -3.2 & +9.1 & -5.3 & +8.9 & +8.7 & +5.9 & +2.6 & +7.4 & -4.1 & \cdots & +2.3 \\ -9.6 & -3.0 & -7.0 & +9.5 & -0.4 & -0.1 & +2.8 & -2.6 & -7.2 & +6.4 & \cdots & +0.2 \\ -5.5 & -8.0 & +7.2 & +9.4 & +9.1 & +8.0 & +5.4 & -3.3 & -8.3 & -1.8 & \cdots & -7.3 \\ -8.8 & +4.5 & -9.7 & +5.4 & -7.0 & -8.3 & -8.1 & +3.4 & -5.0 & -1.6 & \cdots & +7.1 \\ +4.5 & -4.5 & -7.3 & -8.8 & -3.9 & -4.7 & -0.9 & +3.6 & +3.9 & -4.3 & \cdots & -6.3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ -9.0 & +5.9 & -8.4 & +0.4 & -3.8 & +1.5 & +9.1 & +2.9 & -9.2 & -1.4 & \cdots & +0.7 \end{bmatrix}}^{W_Q} \vec{E}_4 = \vec{Q}_4$$
$$\begin{bmatrix} 2.9 \\ 2.4 \\ 1.0 \\ 0.2 \\ 9.2 \\ 6.6 \\ 7.8 \\ 2.8 \\ \vdots \\ 5.8 \\ 0.6 \\ \vdots \\ 9.7 \end{bmatrix} = \begin{bmatrix} +310.6 \\ -95.2 \\ -2.1 \\ -152.0 \\ -123.2 \\ \vdots \\ -12.7 \end{bmatrix}$$

## 3.2 查询

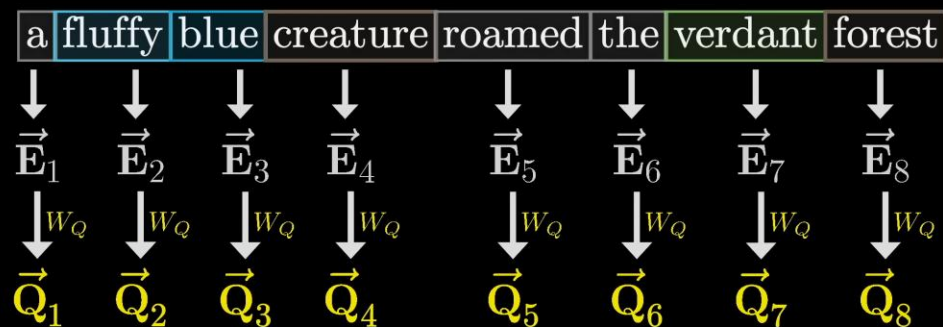
- 这个计算被记为：向量+箭头+权重+结果向量



Any adjectives  
in front of me?

## 3.2 查询

- 将 $W_Q$ 与每一个嵌入向量 $E_i$ 相乘，得到各自的 $Q_i$ 向量
- $W_Q$ 作为权重，也是通过数据训练得到的

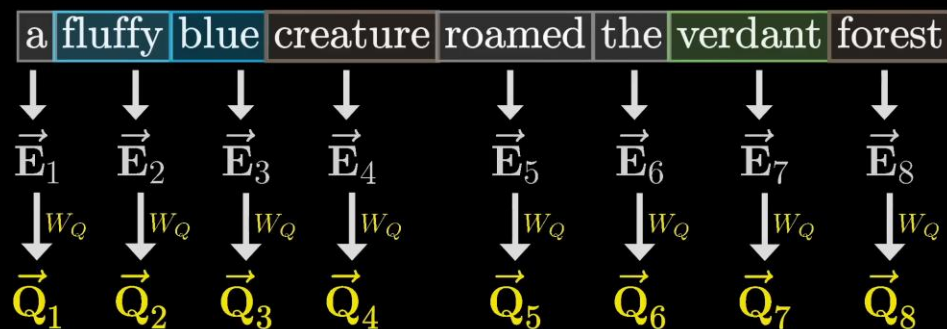


Any adjectives  
in front of me?

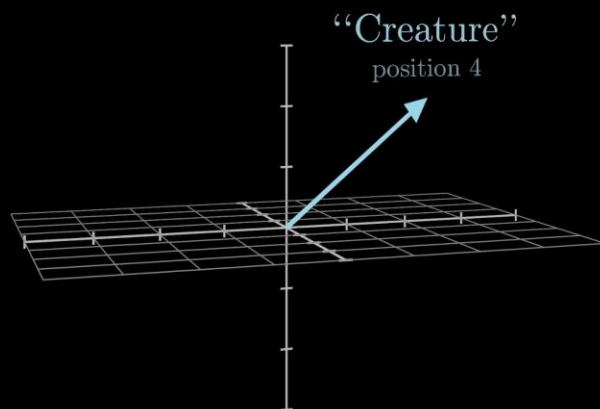
$$\overbrace{\begin{bmatrix} +7.5 & -3.2 & +9.1 & -5.3 & +8.9 & +8.7 & +5.9 & +2.6 & +7.4 & -4.1 & \cdots & +2.3 \\ -9.6 & -3.0 & -7.0 & +9.5 & -0.4 & -0.1 & +2.8 & -2.6 & -7.2 & +6.4 & \cdots & +0.2 \\ -5.5 & -8.0 & +7.2 & +9.4 & +9.1 & +8.0 & +5.4 & -3.3 & -8.3 & -1.8 & \cdots & -7.3 \\ -8.8 & +4.5 & -9.7 & +5.4 & -7.0 & -8.3 & -8.1 & +3.4 & -5.0 & -1.6 & \cdots & +7.1 \\ +4.5 & -4.5 & -7.3 & -8.8 & -3.9 & -4.7 & -0.9 & +3.6 & +3.9 & -4.3 & \cdots & -6.3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ -9.0 & +5.9 & -8.4 & +0.4 & -3.8 & +1.5 & +9.1 & +2.9 & -9.2 & -1.4 & \cdots & +0.7 \end{bmatrix}}^{W_Q} \begin{bmatrix} 2.9 \\ 2.4 \\ 1.0 \\ 0.2 \\ 9.2 \\ 6.6 \\ 7.8 \\ 2.8 \\ 5.8 \\ 0.6 \\ \vdots \\ 9.7 \end{bmatrix}^{\vec{E}_i} = \begin{bmatrix} +310.6 \\ -95.2 \\ -2.1 \\ -152.0 \\ -123.2 \\ \vdots \\ -12.7 \end{bmatrix}^{\vec{Q}_i}$$

## 3.2 查询

- $W_Q$ 的作用：将名词从嵌入向量空间，映射到查询/键空间，体现为特定的查询含义  
即此例中“在我的位置之前有啥形容词吗？”



Embedding space  
12,288-dimensional



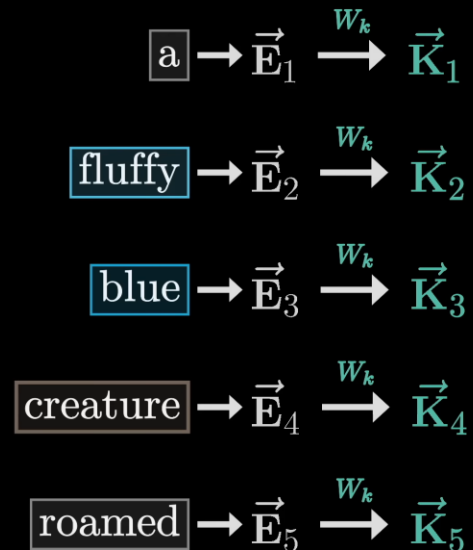
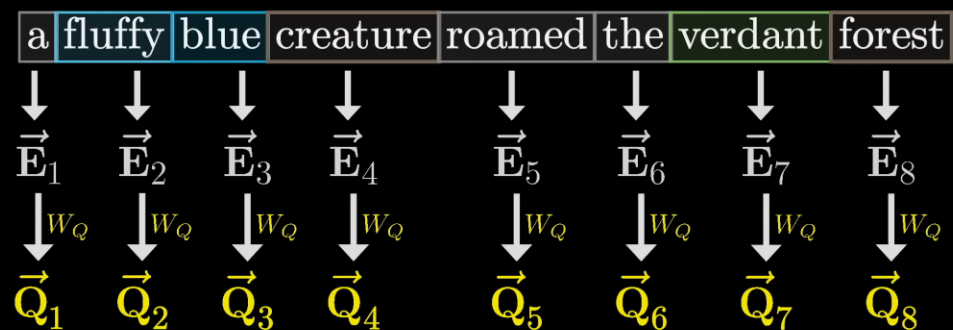
$W_Q$

Query/Key space  
128-dimensional



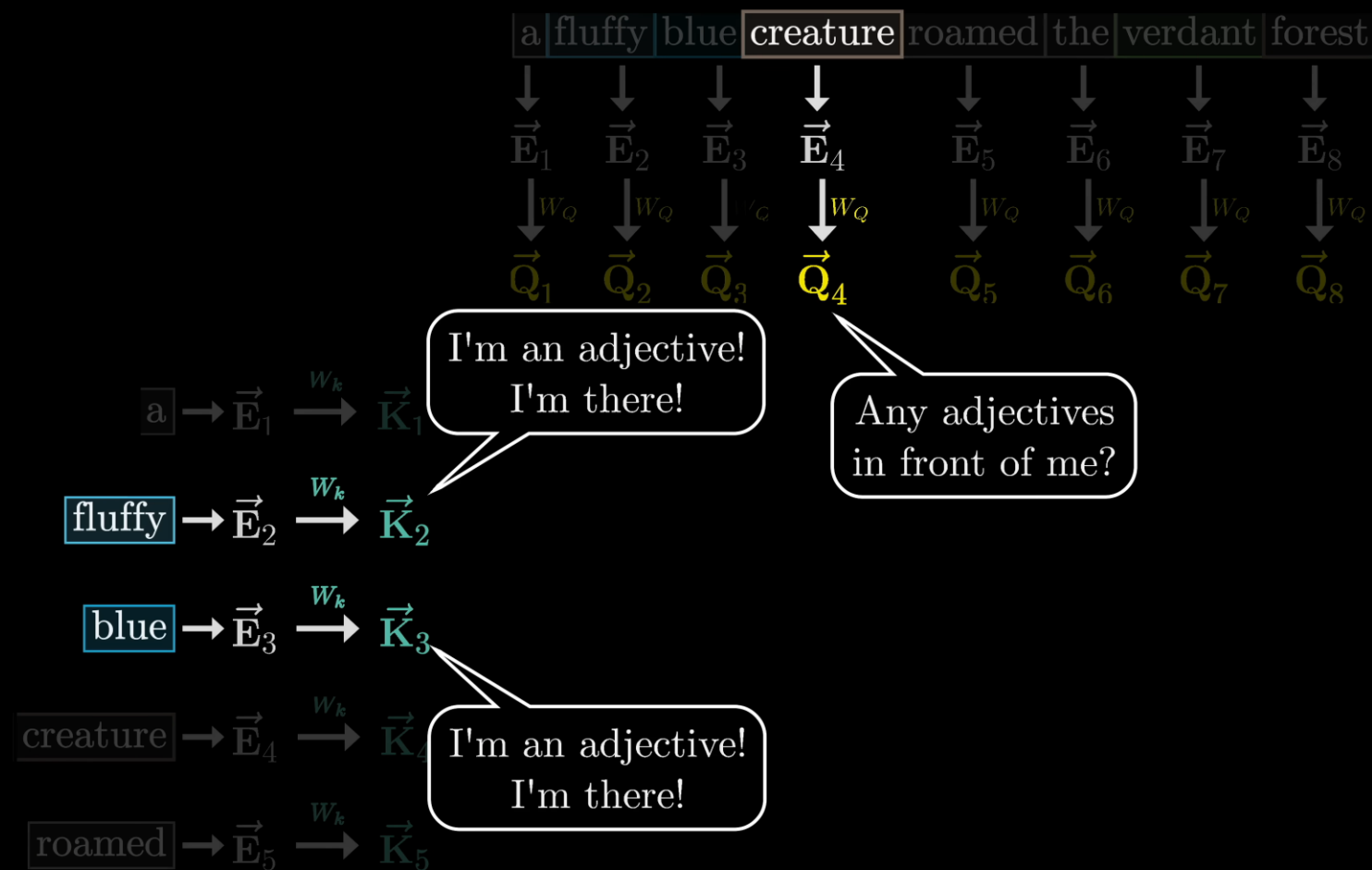
### 3.3 键 (Key)

- 与“查询”相对应的是“键 (Key, 简称为K)”
- $K_i$  向量, 由一个权重矩阵  $W_K$  与每一个嵌入向量相乘得到



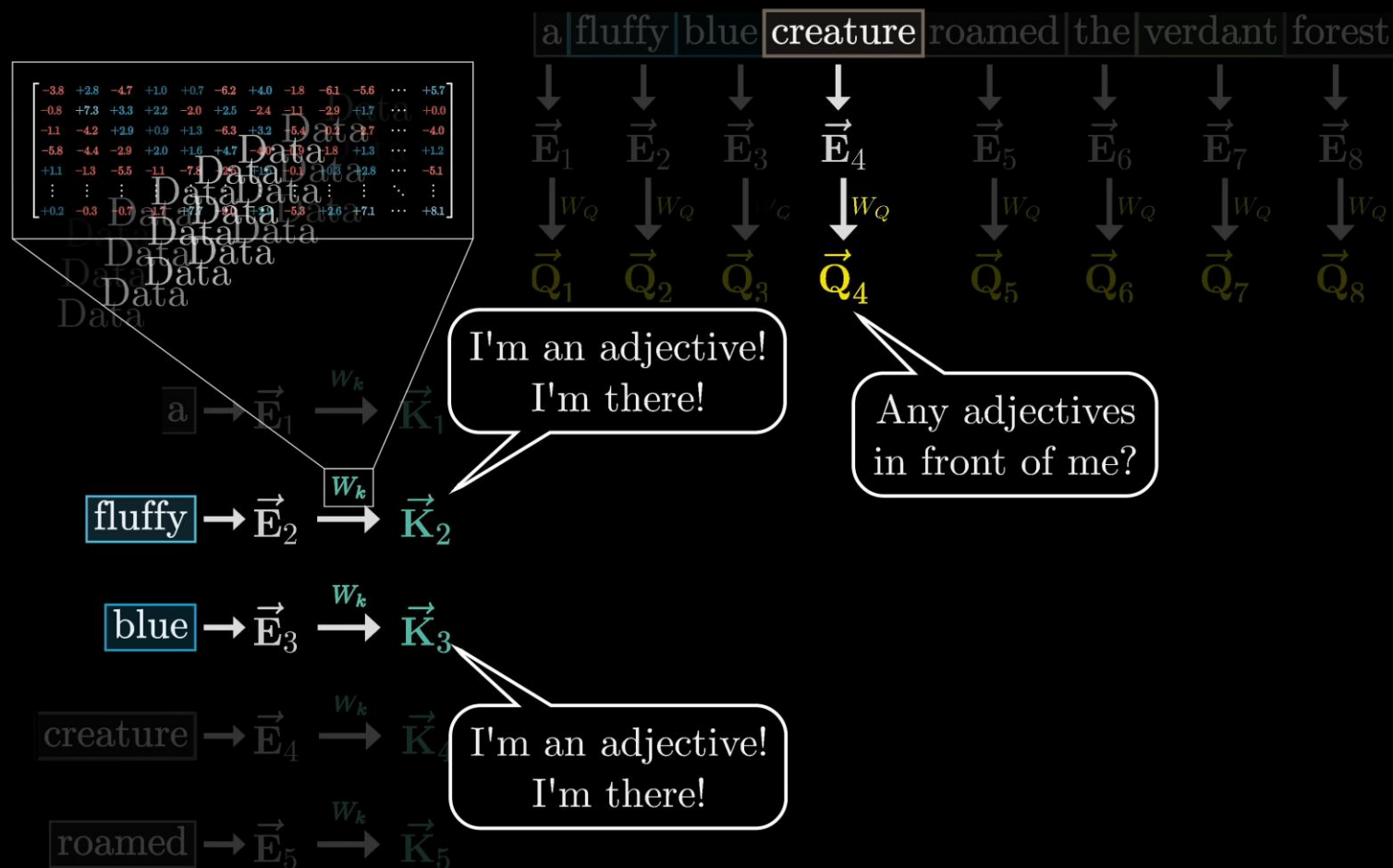
### 3.3 键

- 可以理解为对 $Q_i$ 提出问题的回答



### 3.3 键

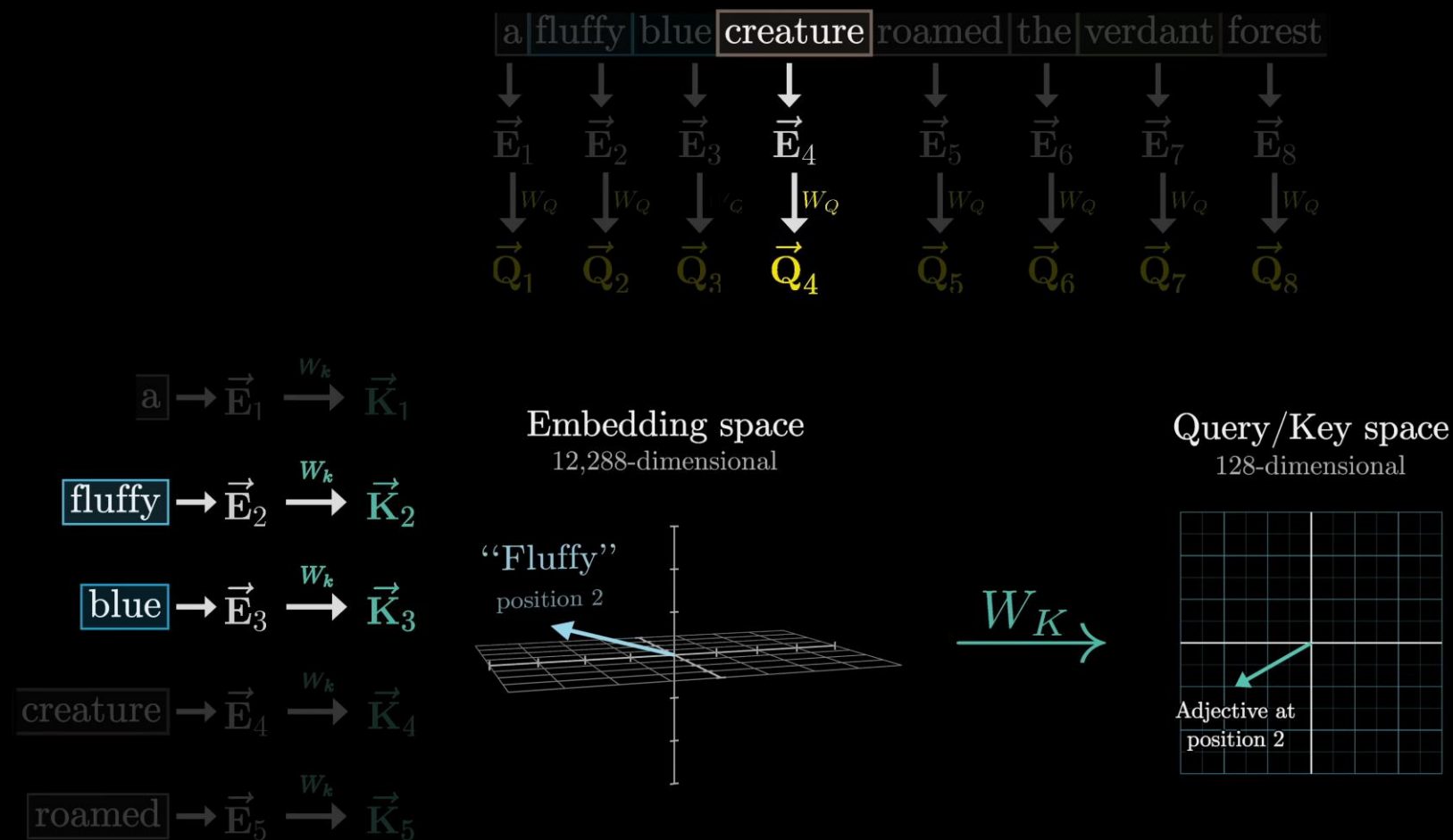
- $W_K$ 作为权重，也是通过数据训练得到的





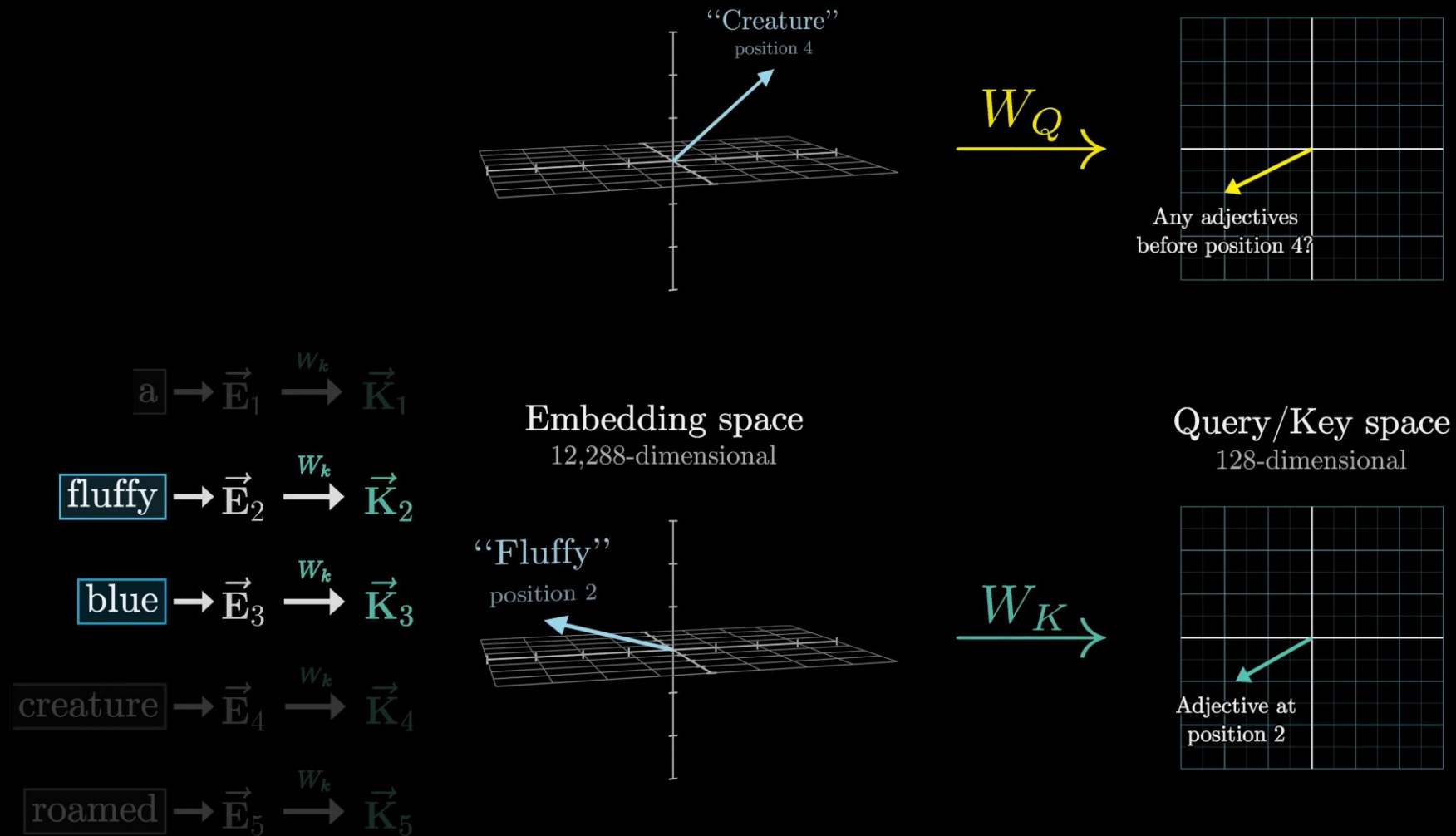
### 3.3 键

- $W_K$ 的作用：将名词从嵌入向量空间，映射到查询/键空间，体现为对特定查询的回应
- 即此例中“‘我’fluffy，是形容词，在2号位，”



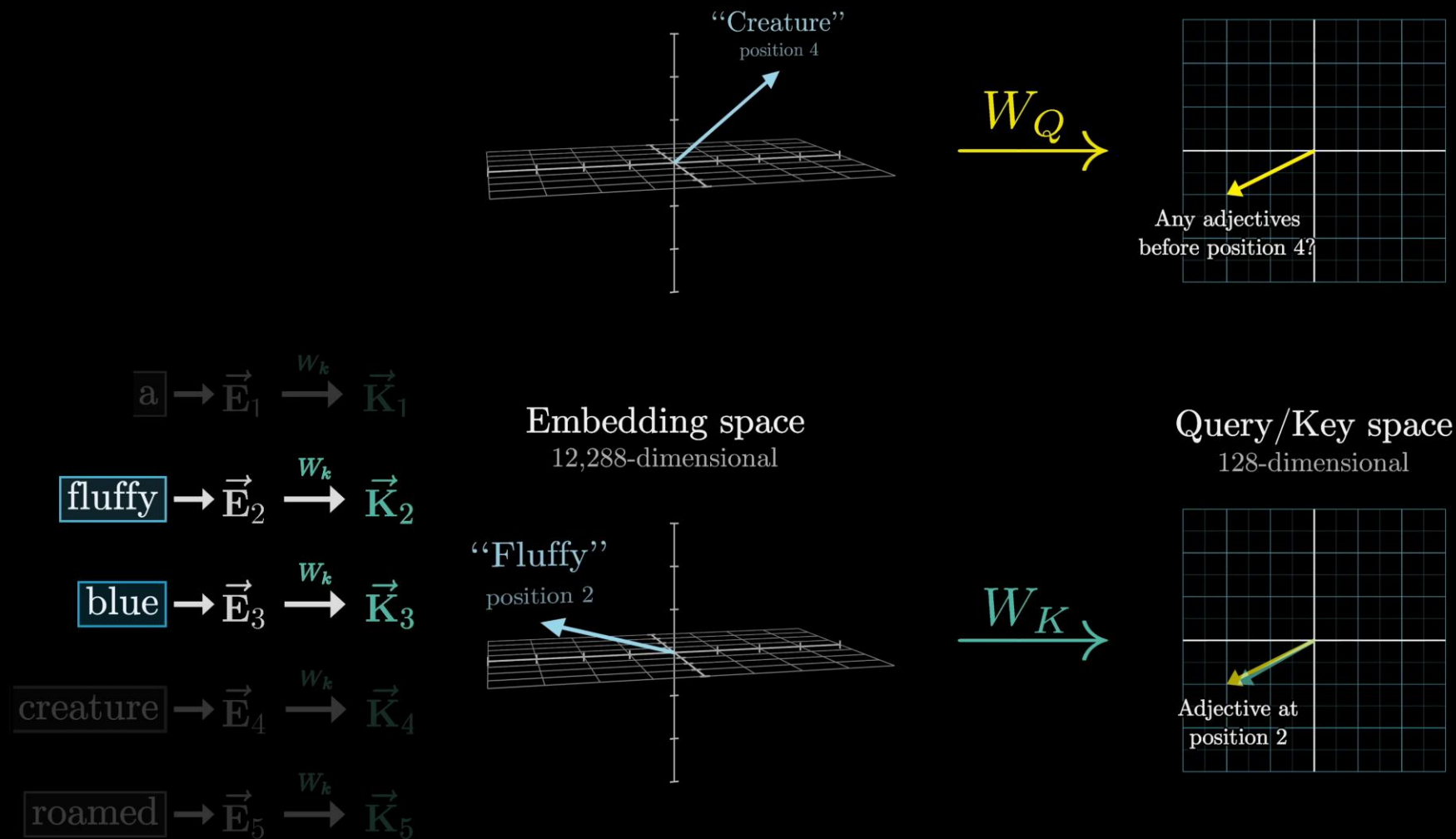
### 3.3 键

- 几何上理解：如果查询得到了键很好的响应，则键与查询向量方向一致。



### 3.3 键

- 几何上理解：如果查询得到了键很好的响应，则键与查询向量方向一致。



### 3.4 注意力

- 计算每一对 $Q_i$ 、 $K_j$ 向量之间的点乘

|                                                              |                                   |                                   |                                   |                                   |                                   |                                   |                                   |                                   |  |
|--------------------------------------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|--|
|                                                              | a                                 | fluffy                            | blue                              | creature                          | roamed                            | the                               | verdant                           | forest                            |  |
|                                                              | $\downarrow$<br>$\vec{E}_1$       | $\downarrow$<br>$\vec{E}_2$       | $\downarrow$<br>$\vec{E}_3$       | $\downarrow$<br>$\vec{E}_4$       | $\downarrow$<br>$\vec{E}_5$       | $\downarrow$<br>$\vec{E}_6$       | $\downarrow$<br>$\vec{E}_7$       | $\downarrow$<br>$\vec{E}_8$       |  |
|                                                              | $\downarrow^{W_Q}$<br>$\vec{Q}_1$ | $\downarrow^{W_Q}$<br>$\vec{Q}_2$ | $\downarrow^{W_Q}$<br>$\vec{Q}_3$ | $\downarrow^{W_Q}$<br>$\vec{Q}_4$ | $\downarrow^{W_Q}$<br>$\vec{Q}_5$ | $\downarrow^{W_Q}$<br>$\vec{Q}_6$ | $\downarrow^{W_Q}$<br>$\vec{Q}_7$ | $\downarrow^{W_Q}$<br>$\vec{Q}_8$ |  |
| a $\rightarrow \vec{E}_1 \xrightarrow{W_k} \vec{K}_1$        |                                   |                                   |                                   |                                   |                                   |                                   |                                   |                                   |  |
| fluffy $\rightarrow \vec{E}_2 \xrightarrow{W_k} \vec{K}_2$   |                                   |                                   |                                   |                                   |                                   |                                   |                                   |                                   |  |
| blue $\rightarrow \vec{E}_3 \xrightarrow{W_k} \vec{K}_3$     |                                   |                                   |                                   |                                   |                                   |                                   |                                   |                                   |  |
| creature $\rightarrow \vec{E}_4 \xrightarrow{W_k} \vec{K}_4$ |                                   |                                   |                                   |                                   |                                   |                                   |                                   |                                   |  |
| roamed $\rightarrow \vec{E}_5 \xrightarrow{W_k} \vec{K}_5$   |                                   |                                   |                                   |                                   |                                   |                                   |                                   |                                   |  |
| the $\rightarrow \vec{E}_6 \xrightarrow{W_k} \vec{K}_6$      |                                   |                                   |                                   |                                   |                                   |                                   |                                   |                                   |  |
| verdant $\rightarrow \vec{E}_7 \xrightarrow{W_k} \vec{K}_7$  |                                   |                                   |                                   |                                   |                                   |                                   |                                   |                                   |  |
| forest $\rightarrow \vec{E}_8 \xrightarrow{W_k} \vec{K}_8$   |                                   |                                   |                                   |                                   |                                   |                                   |                                   |                                   |  |

### 3.4 注意力

- 计算每一对 $Q_i$ 、 $K_j$ 向量之间的点乘

|                                                                             | a                                                                | fluffy                                                           | blue                                                             | creature                                                         | roamed                                                           | the                                                              | verdant                                                          | forest                                                           |  |
|-----------------------------------------------------------------------------|------------------------------------------------------------------|------------------------------------------------------------------|------------------------------------------------------------------|------------------------------------------------------------------|------------------------------------------------------------------|------------------------------------------------------------------|------------------------------------------------------------------|------------------------------------------------------------------|--|
|                                                                             | $\downarrow$<br>$\vec{E}_1$<br>$\downarrow_{W_Q}$<br>$\vec{Q}_1$ | $\downarrow$<br>$\vec{E}_2$<br>$\downarrow_{W_Q}$<br>$\vec{Q}_2$ | $\downarrow$<br>$\vec{E}_3$<br>$\downarrow_{W_Q}$<br>$\vec{Q}_3$ | $\downarrow$<br>$\vec{E}_4$<br>$\downarrow_{W_Q}$<br>$\vec{Q}_4$ | $\downarrow$<br>$\vec{E}_5$<br>$\downarrow_{W_Q}$<br>$\vec{Q}_5$ | $\downarrow$<br>$\vec{E}_6$<br>$\downarrow_{W_Q}$<br>$\vec{Q}_6$ | $\downarrow$<br>$\vec{E}_7$<br>$\downarrow_{W_Q}$<br>$\vec{Q}_7$ | $\downarrow$<br>$\vec{E}_8$<br>$\downarrow_{W_Q}$<br>$\vec{Q}_8$ |  |
| $\boxed{\text{a}} \rightarrow \vec{E}_1 \xrightarrow{W_k} \vec{K}_1$        | $\vec{K}_1 \cdot \vec{Q}_1$                                      | $\vec{K}_1 \cdot \vec{Q}_2$                                      | $\vec{K}_1 \cdot \vec{Q}_3$                                      | $\vec{K}_1 \cdot \vec{Q}_4$                                      | $\vec{K}_1 \cdot \vec{Q}_5$                                      | $\vec{K}_1 \cdot \vec{Q}_6$                                      | $\vec{K}_1 \cdot \vec{Q}_7$                                      | $\vec{K}_1 \cdot \vec{Q}_8$                                      |  |
| $\boxed{\text{fluffy}} \rightarrow \vec{E}_2 \xrightarrow{W_k} \vec{K}_2$   | $\vec{K}_2 \cdot \vec{Q}_1$                                      | $\vec{K}_2 \cdot \vec{Q}_2$                                      | $\vec{K}_2 \cdot \vec{Q}_3$                                      | $\vec{K}_2 \cdot \vec{Q}_4$                                      | $\vec{K}_2 \cdot \vec{Q}_5$                                      | $\vec{K}_2 \cdot \vec{Q}_6$                                      | $\vec{K}_2 \cdot \vec{Q}_7$                                      | $\vec{K}_2 \cdot \vec{Q}_8$                                      |  |
| $\boxed{\text{blue}} \rightarrow \vec{E}_3 \xrightarrow{W_k} \vec{K}_3$     | $\vec{K}_3 \cdot \vec{Q}_1$                                      | $\vec{K}_3 \cdot \vec{Q}_2$                                      | $\vec{K}_3 \cdot \vec{Q}_3$                                      | $\vec{K}_3 \cdot \vec{Q}_4$                                      | $\vec{K}_3 \cdot \vec{Q}_5$                                      | $\vec{K}_3 \cdot \vec{Q}_6$                                      | $\vec{K}_3 \cdot \vec{Q}_7$                                      | $\vec{K}_3 \cdot \vec{Q}_8$                                      |  |
| $\boxed{\text{creature}} \rightarrow \vec{E}_4 \xrightarrow{W_k} \vec{K}_4$ | $\vec{K}_4 \cdot \vec{Q}_1$                                      | $\vec{K}_4 \cdot \vec{Q}_2$                                      | $\vec{K}_4 \cdot \vec{Q}_3$                                      | $\vec{K}_4 \cdot \vec{Q}_4$                                      | $\vec{K}_4 \cdot \vec{Q}_5$                                      | $\vec{K}_4 \cdot \vec{Q}_6$                                      | $\vec{K}_4 \cdot \vec{Q}_7$                                      | $\vec{K}_4 \cdot \vec{Q}_8$                                      |  |
| $\boxed{\text{roamed}} \rightarrow \vec{E}_5 \xrightarrow{W_k} \vec{K}_5$   | $\vec{K}_5 \cdot \vec{Q}_1$                                      | $\vec{K}_5 \cdot \vec{Q}_2$                                      | $\vec{K}_5 \cdot \vec{Q}_3$                                      | $\vec{K}_5 \cdot \vec{Q}_4$                                      | $\vec{K}_5 \cdot \vec{Q}_5$                                      | $\vec{K}_5 \cdot \vec{Q}_6$                                      | $\vec{K}_5 \cdot \vec{Q}_7$                                      | $\vec{K}_5 \cdot \vec{Q}_8$                                      |  |
| $\boxed{\text{the}} \rightarrow \vec{E}_6 \xrightarrow{W_k} \vec{K}_6$      | $\vec{K}_6 \cdot \vec{Q}_1$                                      | $\vec{K}_6 \cdot \vec{Q}_2$                                      | $\vec{K}_6 \cdot \vec{Q}_3$                                      | $\vec{K}_6 \cdot \vec{Q}_4$                                      | $\vec{K}_6 \cdot \vec{Q}_5$                                      | $\vec{K}_6 \cdot \vec{Q}_6$                                      | $\vec{K}_6 \cdot \vec{Q}_7$                                      | $\vec{K}_6 \cdot \vec{Q}_8$                                      |  |
| $\boxed{\text{verdant}} \rightarrow \vec{E}_7 \xrightarrow{W_k} \vec{K}_7$  | $\vec{K}_7 \cdot \vec{Q}_1$                                      | $\vec{K}_7 \cdot \vec{Q}_2$                                      | $\vec{K}_7 \cdot \vec{Q}_3$                                      | $\vec{K}_7 \cdot \vec{Q}_4$                                      | $\vec{K}_7 \cdot \vec{Q}_5$                                      | $\vec{K}_7 \cdot \vec{Q}_6$                                      | $\vec{K}_7 \cdot \vec{Q}_7$                                      | $\vec{K}_7 \cdot \vec{Q}_8$                                      |  |
| $\boxed{\text{forest}} \rightarrow \vec{E}_8 \xrightarrow{W_k} \vec{K}_8$   | $\vec{K}_8 \cdot \vec{Q}_1$                                      | $\vec{K}_8 \cdot \vec{Q}_2$                                      | $\vec{K}_8 \cdot \vec{Q}_3$                                      | $\vec{K}_8 \cdot \vec{Q}_4$                                      | $\vec{K}_8 \cdot \vec{Q}_5$                                      | $\vec{K}_8 \cdot \vec{Q}_6$                                      | $\vec{K}_8 \cdot \vec{Q}_7$                                      | $\vec{K}_8 \cdot \vec{Q}_8$                                      |  |

### 3.4 注意力

- 点乘的结果越大， $Q_i$ 与 $K_j$ 的更为对齐

|                                                                             | a                                                                | fluffy                                                           | blue                                                             | creature                                                         | roamed                                                           | the                                                              | verdant                                                          | forest                                                           |  |
|-----------------------------------------------------------------------------|------------------------------------------------------------------|------------------------------------------------------------------|------------------------------------------------------------------|------------------------------------------------------------------|------------------------------------------------------------------|------------------------------------------------------------------|------------------------------------------------------------------|------------------------------------------------------------------|--|
|                                                                             | $\downarrow$<br>$\vec{E}_1$<br>$\downarrow_{W_Q}$<br>$\vec{Q}_1$ | $\downarrow$<br>$\vec{E}_2$<br>$\downarrow_{W_Q}$<br>$\vec{Q}_2$ | $\downarrow$<br>$\vec{E}_3$<br>$\downarrow_{W_Q}$<br>$\vec{Q}_3$ | $\downarrow$<br>$\vec{E}_4$<br>$\downarrow_{W_Q}$<br>$\vec{Q}_4$ | $\downarrow$<br>$\vec{E}_5$<br>$\downarrow_{W_Q}$<br>$\vec{Q}_5$ | $\downarrow$<br>$\vec{E}_6$<br>$\downarrow_{W_Q}$<br>$\vec{Q}_6$ | $\downarrow$<br>$\vec{E}_7$<br>$\downarrow_{W_Q}$<br>$\vec{Q}_7$ | $\downarrow$<br>$\vec{E}_8$<br>$\downarrow_{W_Q}$<br>$\vec{Q}_8$ |  |
| $\boxed{\text{a}} \rightarrow \vec{E}_1 \xrightarrow{W_k} \vec{K}_1$        | $\vec{K}_1 \cdot \vec{Q}_1$                                      | $\vec{K}_1 \cdot \vec{Q}_2$                                      | $\vec{K}_1 \cdot \vec{Q}_3$                                      | $\vec{K}_1 \cdot \vec{Q}_4$                                      | $\vec{K}_1 \cdot \vec{Q}_5$                                      | $\vec{K}_1 \cdot \vec{Q}_6$                                      | $\vec{K}_1 \cdot \vec{Q}_7$                                      | $\vec{K}_1 \cdot \vec{Q}_8$                                      |  |
| $\boxed{\text{fluffy}} \rightarrow \vec{E}_2 \xrightarrow{W_k} \vec{K}_2$   | $\vec{K}_2 \cdot \vec{Q}_1$                                      | $\vec{K}_2 \cdot \vec{Q}_2$                                      | $\vec{K}_2 \cdot \vec{Q}_3$                                      | $\vec{K}_2 \cdot \vec{Q}_4$                                      | $\vec{K}_2 \cdot \vec{Q}_5$                                      | $\vec{K}_2 \cdot \vec{Q}_6$                                      | $\vec{K}_2 \cdot \vec{Q}_7$                                      | $\vec{K}_2 \cdot \vec{Q}_8$                                      |  |
| $\boxed{\text{blue}} \rightarrow \vec{E}_3 \xrightarrow{W_k} \vec{K}_3$     | $\vec{K}_3 \cdot \vec{Q}_1$                                      | $\vec{K}_3 \cdot \vec{Q}_2$                                      | $\vec{K}_3 \cdot \vec{Q}_3$                                      | $\vec{K}_3 \cdot \vec{Q}_4$                                      | $\vec{K}_3 \cdot \vec{Q}_5$                                      | $\vec{K}_3 \cdot \vec{Q}_6$                                      | $\vec{K}_3 \cdot \vec{Q}_7$                                      | $\vec{K}_3 \cdot \vec{Q}_8$                                      |  |
| $\boxed{\text{creature}} \rightarrow \vec{E}_4 \xrightarrow{W_k} \vec{K}_4$ | $\vec{K}_4 \cdot \vec{Q}_1$                                      | $\vec{K}_4 \cdot \vec{Q}_2$                                      | $\vec{K}_4 \cdot \vec{Q}_3$                                      | $\vec{K}_4 \cdot \vec{Q}_4$                                      | $\vec{K}_4 \cdot \vec{Q}_5$                                      | $\vec{K}_4 \cdot \vec{Q}_6$                                      | $\vec{K}_4 \cdot \vec{Q}_7$                                      | $\vec{K}_4 \cdot \vec{Q}_8$                                      |  |
| $\boxed{\text{roamed}} \rightarrow \vec{E}_5 \xrightarrow{W_k} \vec{K}_5$   | $\vec{K}_5 \cdot \vec{Q}_1$                                      | $\vec{K}_5 \cdot \vec{Q}_2$                                      | $\vec{K}_5 \cdot \vec{Q}_3$                                      | $\vec{K}_5 \cdot \vec{Q}_4$                                      | $\vec{K}_5 \cdot \vec{Q}_5$                                      | $\vec{K}_5 \cdot \vec{Q}_6$                                      | $\vec{K}_5 \cdot \vec{Q}_7$                                      | $\vec{K}_5 \cdot \vec{Q}_8$                                      |  |
| $\boxed{\text{the}} \rightarrow \vec{E}_6 \xrightarrow{W_k} \vec{K}_6$      | $\vec{K}_6 \cdot \vec{Q}_1$                                      | $\vec{K}_6 \cdot \vec{Q}_2$                                      | $\vec{K}_6 \cdot \vec{Q}_3$                                      | $\vec{K}_6 \cdot \vec{Q}_4$                                      | $\vec{K}_6 \cdot \vec{Q}_5$                                      | $\vec{K}_6 \cdot \vec{Q}_6$                                      | $\vec{K}_6 \cdot \vec{Q}_7$                                      | $\vec{K}_6 \cdot \vec{Q}_8$                                      |  |
| $\boxed{\text{verdant}} \rightarrow \vec{E}_7 \xrightarrow{W_k} \vec{K}_7$  | $\vec{K}_7 \cdot \vec{Q}_1$                                      | $\vec{K}_7 \cdot \vec{Q}_2$                                      | $\vec{K}_7 \cdot \vec{Q}_3$                                      | $\vec{K}_7 \cdot \vec{Q}_4$                                      | $\vec{K}_7 \cdot \vec{Q}_5$                                      | $\vec{K}_7 \cdot \vec{Q}_6$                                      | $\vec{K}_7 \cdot \vec{Q}_7$                                      | $\vec{K}_7 \cdot \vec{Q}_8$                                      |  |
| $\boxed{\text{forest}} \rightarrow \vec{E}_8 \xrightarrow{W_k} \vec{K}_8$   | $\vec{K}_8 \cdot \vec{Q}_1$                                      | $\vec{K}_8 \cdot \vec{Q}_2$                                      | $\vec{K}_8 \cdot \vec{Q}_3$                                      | $\vec{K}_8 \cdot \vec{Q}_4$                                      | $\vec{K}_8 \cdot \vec{Q}_5$                                      | $\vec{K}_8 \cdot \vec{Q}_6$                                      | $\vec{K}_8 \cdot \vec{Q}_7$                                      | $\vec{K}_8 \cdot \vec{Q}_8$                                      |  |

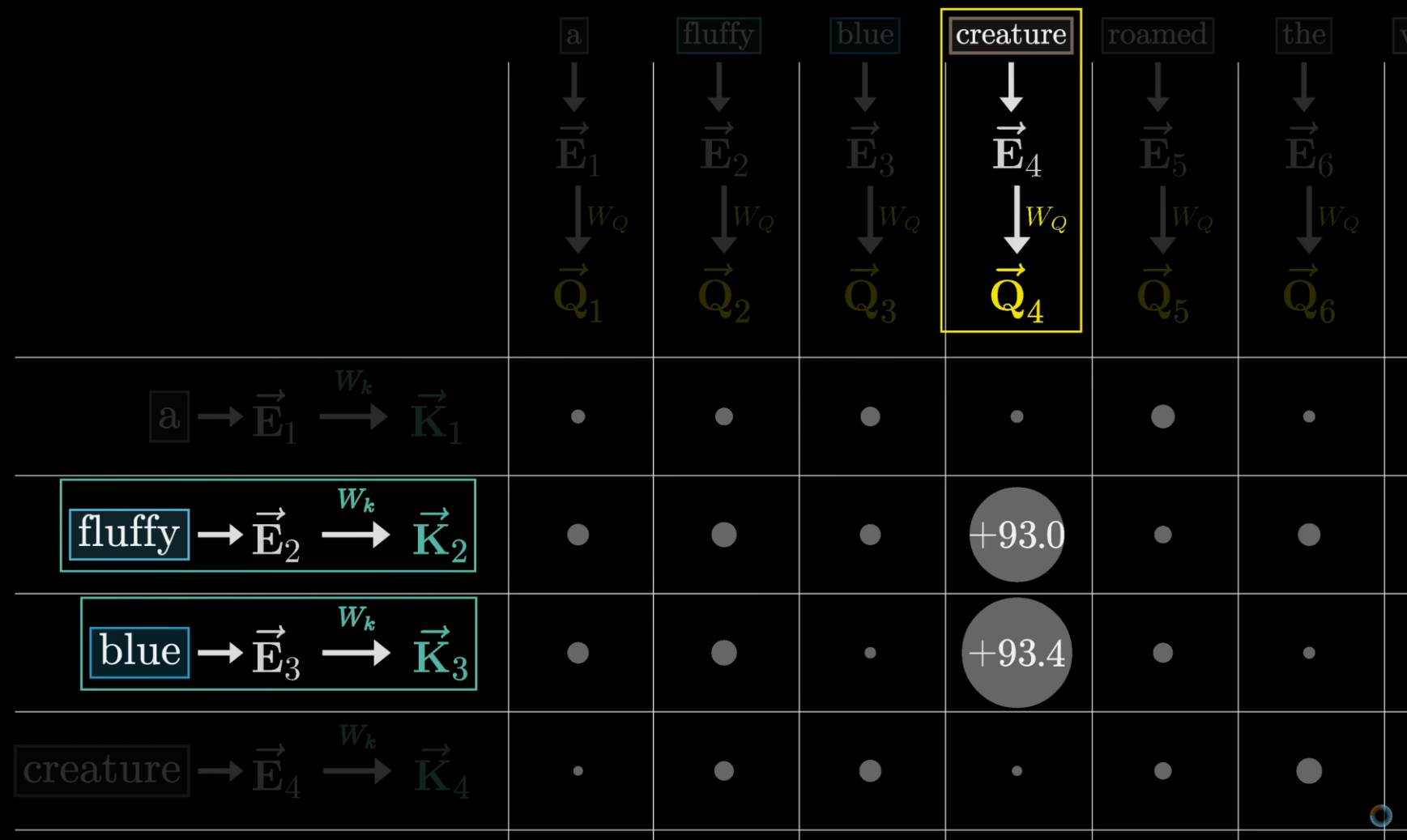
### 3.4 注意力

- 本例中，名词creature和forest在各自的修饰词fluffy、blue以及verdant上得到较大的响应

|                                                                             | a                                                                | fluffy                                                           | blue                                                             | creature                                                         | roamed                                                           | the                                                              | verdant                                                          | forest                                                           |  |
|-----------------------------------------------------------------------------|------------------------------------------------------------------|------------------------------------------------------------------|------------------------------------------------------------------|------------------------------------------------------------------|------------------------------------------------------------------|------------------------------------------------------------------|------------------------------------------------------------------|------------------------------------------------------------------|--|
|                                                                             | $\downarrow$<br>$\vec{E}_1$<br>$\downarrow_{W_Q}$<br>$\vec{Q}_1$ | $\downarrow$<br>$\vec{E}_2$<br>$\downarrow_{W_Q}$<br>$\vec{Q}_2$ | $\downarrow$<br>$\vec{E}_3$<br>$\downarrow_{W_Q}$<br>$\vec{Q}_3$ | $\downarrow$<br>$\vec{E}_4$<br>$\downarrow_{W_Q}$<br>$\vec{Q}_4$ | $\downarrow$<br>$\vec{E}_5$<br>$\downarrow_{W_Q}$<br>$\vec{Q}_5$ | $\downarrow$<br>$\vec{E}_6$<br>$\downarrow_{W_Q}$<br>$\vec{Q}_6$ | $\downarrow$<br>$\vec{E}_7$<br>$\downarrow_{W_Q}$<br>$\vec{Q}_7$ | $\downarrow$<br>$\vec{E}_8$<br>$\downarrow_{W_Q}$<br>$\vec{Q}_8$ |  |
| $\boxed{\text{a}} \rightarrow \vec{E}_1 \xrightarrow{W_k} \vec{K}_1$        | $\vec{K}_1 \bullet \vec{Q}_1$                                    | $\vec{K}_1 \bullet \vec{Q}_2$                                    | $\vec{K}_1 \bullet \vec{Q}_3$                                    | $\vec{K}_1 \bullet \vec{Q}_4$                                    | $\vec{K}_1 \bullet \vec{Q}_5$                                    | $\vec{K}_1 \bullet \vec{Q}_6$                                    | $\vec{K}_1 \bullet \vec{Q}_7$                                    | $\vec{K}_1 \bullet \vec{Q}_8$                                    |  |
| $\boxed{\text{fluffy}} \rightarrow \vec{E}_2 \xrightarrow{W_k} \vec{K}_2$   | $\vec{K}_2 \bullet \vec{Q}_1$                                    | $\vec{K}_2 \bullet \vec{Q}_2$                                    | $\vec{K}_2 \bullet \vec{Q}_3$                                    | $\vec{K}_2 \bullet \vec{Q}_4$                                    | $\vec{K}_2 \bullet \vec{Q}_5$                                    | $\vec{K}_2 \bullet \vec{Q}_6$                                    | $\vec{K}_2 \bullet \vec{Q}_7$                                    | $\vec{K}_2 \bullet \vec{Q}_8$                                    |  |
| $\boxed{\text{blue}} \rightarrow \vec{E}_3 \xrightarrow{W_k} \vec{K}_3$     | $\vec{K}_3 \bullet \vec{Q}_1$                                    | $\vec{K}_3 \bullet \vec{Q}_2$                                    | $\vec{K}_3 \bullet \vec{Q}_3$                                    | $\vec{K}_3 \bullet \vec{Q}_4$                                    | $\vec{K}_3 \bullet \vec{Q}_5$                                    | $\vec{K}_3 \bullet \vec{Q}_6$                                    | $\vec{K}_3 \bullet \vec{Q}_7$                                    | $\vec{K}_3 \bullet \vec{Q}_8$                                    |  |
| $\boxed{\text{creature}} \rightarrow \vec{E}_4 \xrightarrow{W_k} \vec{K}_4$ | $\vec{K}_4 \bullet \vec{Q}_1$                                    | $\vec{K}_4 \bullet \vec{Q}_2$                                    | $\vec{K}_4 \bullet \vec{Q}_3$                                    | $\vec{K}_4 \bullet \vec{Q}_4$                                    | $\vec{K}_4 \bullet \vec{Q}_5$                                    | $\vec{K}_4 \bullet \vec{Q}_6$                                    | $\vec{K}_4 \bullet \vec{Q}_7$                                    | $\vec{K}_4 \bullet \vec{Q}_8$                                    |  |
| $\boxed{\text{roamed}} \rightarrow \vec{E}_5 \xrightarrow{W_k} \vec{K}_5$   | $\vec{K}_5 \bullet \vec{Q}_1$                                    | $\vec{K}_5 \bullet \vec{Q}_2$                                    | $\vec{K}_5 \bullet \vec{Q}_3$                                    | $\vec{K}_5 \bullet \vec{Q}_4$                                    | $\vec{K}_5 \bullet \vec{Q}_5$                                    | $\vec{K}_5 \bullet \vec{Q}_6$                                    | $\vec{K}_5 \bullet \vec{Q}_7$                                    | $\vec{K}_5 \bullet \vec{Q}_8$                                    |  |
| $\boxed{\text{the}} \rightarrow \vec{E}_6 \xrightarrow{W_k} \vec{K}_6$      | $\vec{K}_6 \bullet \vec{Q}_1$                                    | $\vec{K}_6 \bullet \vec{Q}_2$                                    | $\vec{K}_6 \bullet \vec{Q}_3$                                    | $\vec{K}_6 \bullet \vec{Q}_4$                                    | $\vec{K}_6 \bullet \vec{Q}_5$                                    | $\vec{K}_6 \bullet \vec{Q}_6$                                    | $\vec{K}_6 \bullet \vec{Q}_7$                                    | $\vec{K}_6 \bullet \vec{Q}_8$                                    |  |
| $\boxed{\text{verdant}} \rightarrow \vec{E}_7 \xrightarrow{W_k} \vec{K}_7$  | $\vec{K}_7 \bullet \vec{Q}_1$                                    | $\vec{K}_7 \bullet \vec{Q}_2$                                    | $\vec{K}_7 \bullet \vec{Q}_3$                                    | $\vec{K}_7 \bullet \vec{Q}_4$                                    | $\vec{K}_7 \bullet \vec{Q}_5$                                    | $\vec{K}_7 \bullet \vec{Q}_6$                                    | $\vec{K}_7 \bullet \vec{Q}_7$                                    | $\vec{K}_7 \bullet \vec{Q}_8$                                    |  |
| $\boxed{\text{forest}} \rightarrow \vec{E}_8 \xrightarrow{W_k} \vec{K}_8$   | $\vec{K}_8 \bullet \vec{Q}_1$                                    | $\vec{K}_8 \bullet \vec{Q}_2$                                    | $\vec{K}_8 \bullet \vec{Q}_3$                                    | $\vec{K}_8 \bullet \vec{Q}_4$                                    | $\vec{K}_8 \bullet \vec{Q}_5$                                    | $\vec{K}_8 \bullet \vec{Q}_6$                                    | $\vec{K}_8 \bullet \vec{Q}_7$                                    | $\vec{K}_8 \bullet \vec{Q}_8$                                    |  |

### 3.4 注意力

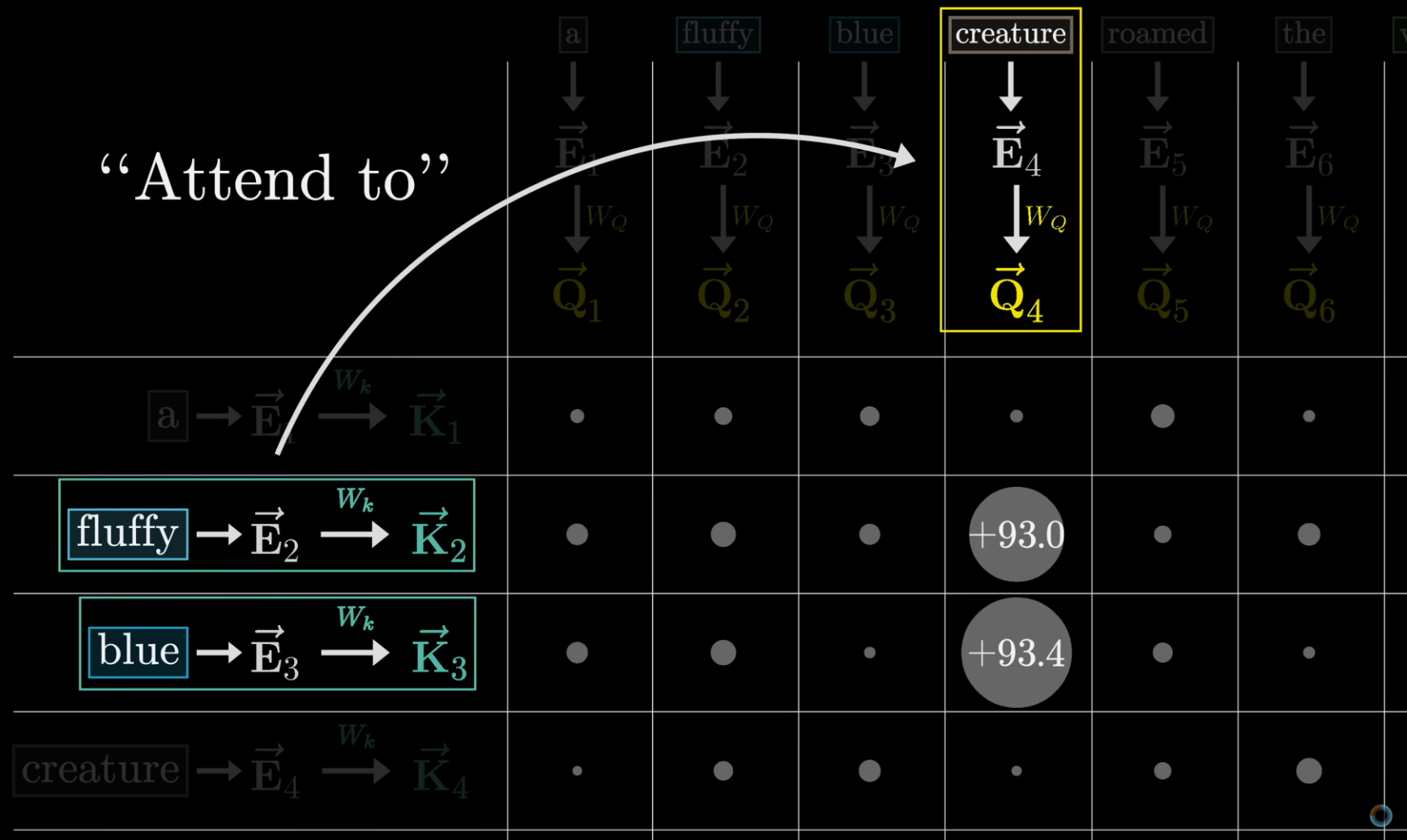
- fluffy和blue的词嵌入“注意到 (attend to) ”creature的词嵌入





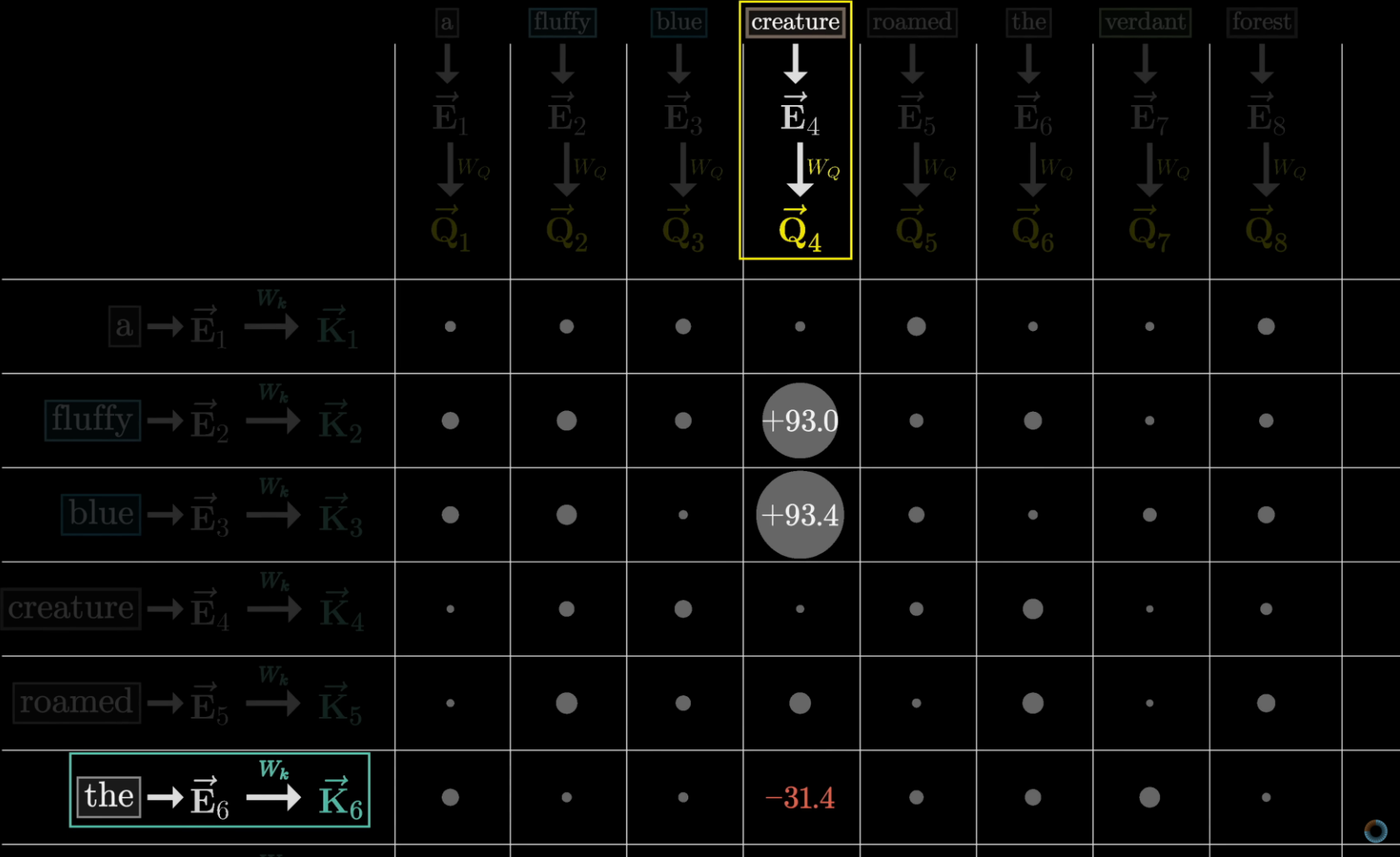
### 3.4 注意力

- fluffy和blue的词嵌入“注意到 (attend to)”creature的词嵌入



### 3.4 注意力

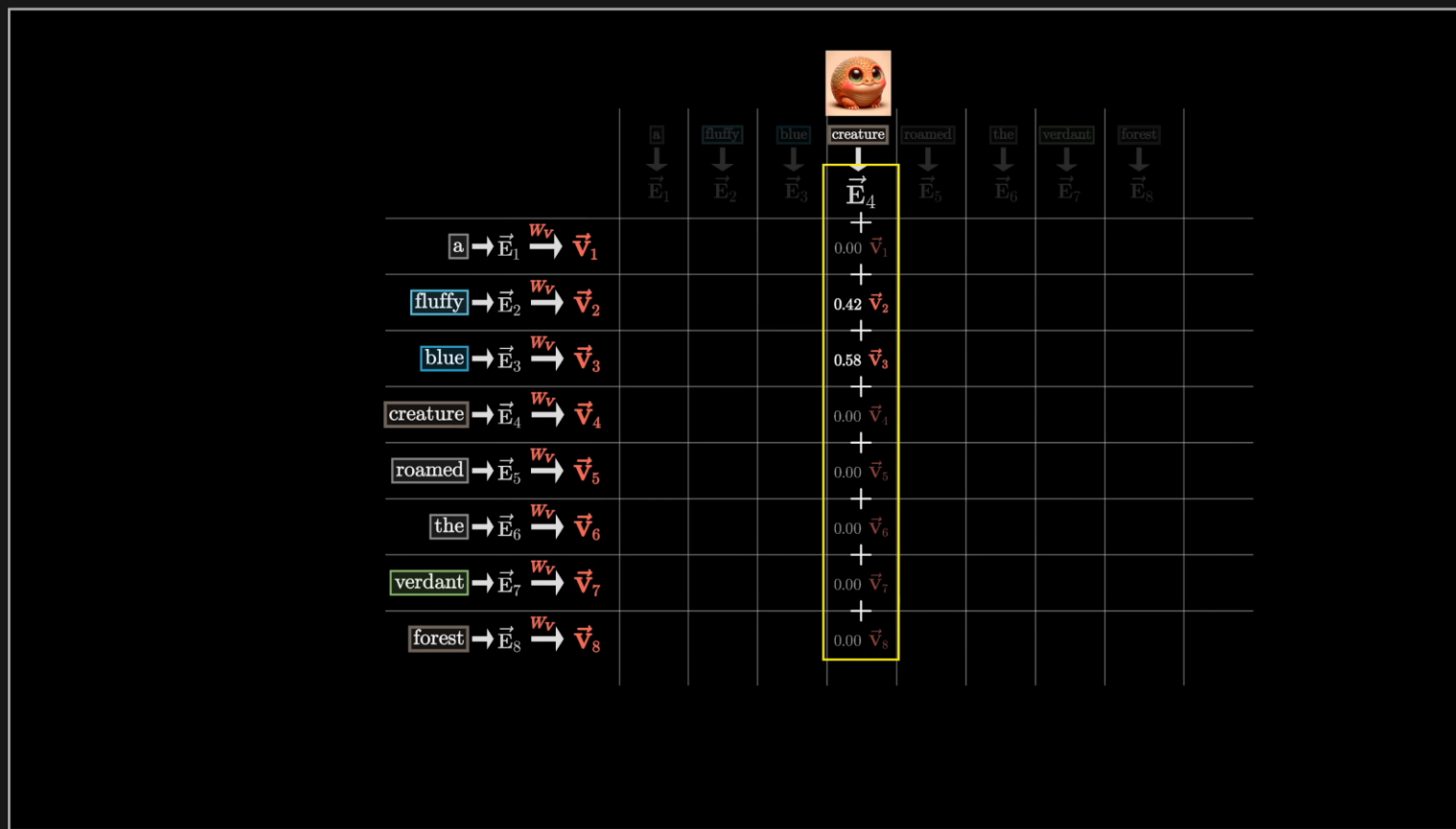
- 相反，点乘的结果越小，则两个词无关



### 3.5 注意力模式

- 稍后我们将利用点乘作为嵌入向量更新的权重

In a minute...

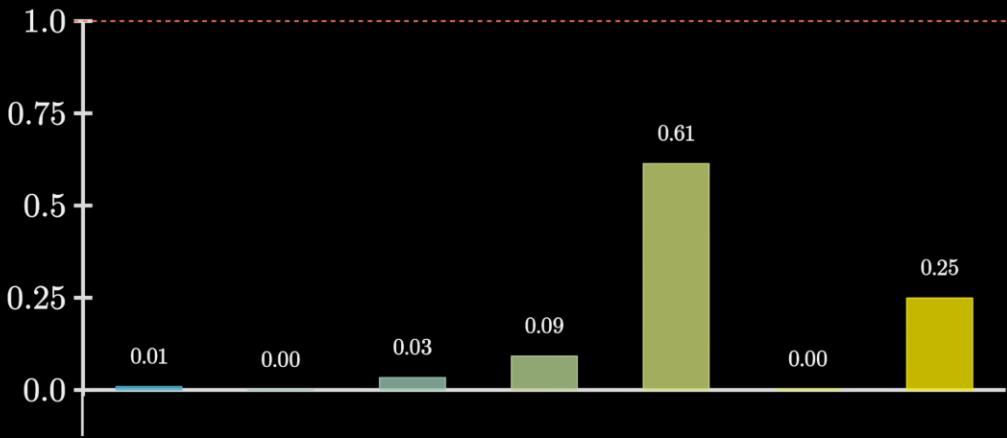


### 3.5 注意力模式

- 因此需要将其进行归一化：即均为正值且和为一

| creature                     | roamed                       | the                          | verdant                      | forest                       |
|------------------------------|------------------------------|------------------------------|------------------------------|------------------------------|
| $\downarrow \vec{E}_4$       | $\downarrow \vec{E}_5$       | $\downarrow \vec{E}_6$       | $\downarrow \vec{E}_7$       | $\downarrow \vec{E}_8$       |
| $\downarrow_{W_Q} \vec{Q}_4$ | $\downarrow_{W_Q} \vec{Q}_5$ | $\downarrow_{W_Q} \vec{Q}_6$ | $\downarrow_{W_Q} \vec{Q}_7$ | $\downarrow_{W_Q} \vec{Q}_8$ |
| -27.8                        | -5.2                         | -89.3                        | -45.2                        | -36.1                        |
| +93.0                        | -15.7                        | -15.7                        | -15.7                        | -15.7                        |
| +93.4                        | -51.5                        | -51.5                        | -51.5                        | -51.5                        |
| +4.9                         | -32.4                        | -92.3                        | -9.5                         | -28.1                        |
| -55.4                        | +0.6                         | -64.7                        | -96.7                        | -18.9                        |
| -31.4                        | +5.5                         | +0.6                         | -4.6                         | -96.8                        |
| -59.8                        | -79.0                        | -97.9                        | +3.7                         | +93.8                        |
| -90.6                        | -75.6                        | -89.0                        | -70.8                        | +4.7                         |

We want these to  
act like weights



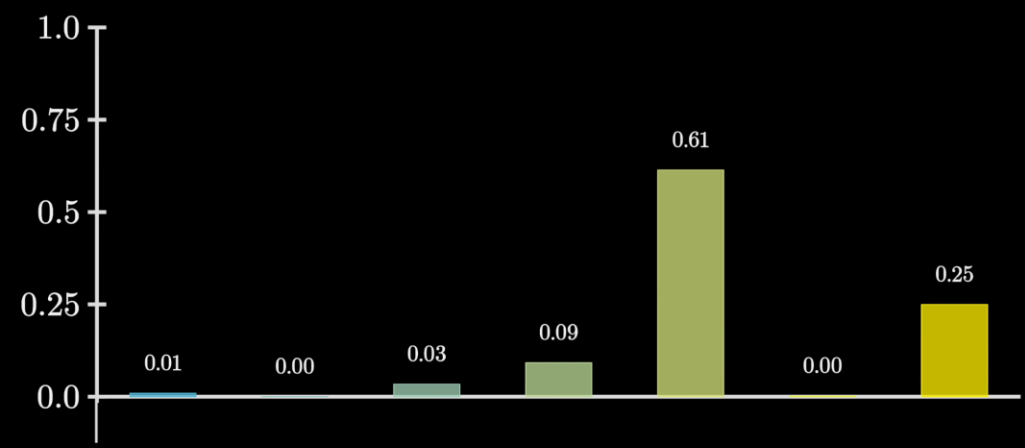
### 3.5 注意力模式

- 因此需要将其进行归一化：即均为正值且和为一

| creature                     | roamed                       | the                          | verdant                      | forest                       |
|------------------------------|------------------------------|------------------------------|------------------------------|------------------------------|
| $\downarrow \vec{E}_4$       | $\downarrow \vec{E}_5$       | $\downarrow \vec{E}_6$       | $\downarrow \vec{E}_7$       | $\downarrow \vec{E}_8$       |
| $\downarrow_{W_Q} \vec{Q}_4$ | $\downarrow_{W_Q} \vec{Q}_5$ | $\downarrow_{W_Q} \vec{Q}_6$ | $\downarrow_{W_Q} \vec{Q}_7$ | $\downarrow_{W_Q} \vec{Q}_8$ |
| -27.8                        | -5.2                         | -89.3                        | -45.2                        | -36.1                        |
| +93.0                        | -55.7                        | -56.7                        | -45.2                        | -36.1                        |
| +93.4                        | -55.7                        | -56.7                        | -45.2                        | -36.1                        |
| +4.9                         | -32.4                        | -92.3                        | -9.5                         | -28.1                        |
| -55.4                        | +0.6                         | -64.7                        | -96.7                        | -18.9                        |
| -31.4                        | +5.5                         | +0.6                         | -4.6                         | -96.8                        |
| -59.8                        | -79.0                        | -97.9                        | +3.7                         | +93.8                        |
| -90.6                        | -75.6                        | -89.0                        | -70.8                        | +4.7                         |

We want these to act like weights

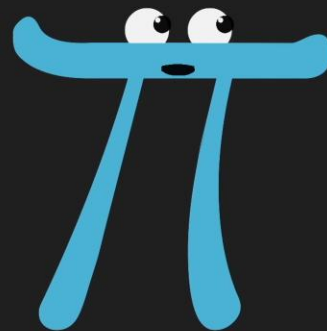
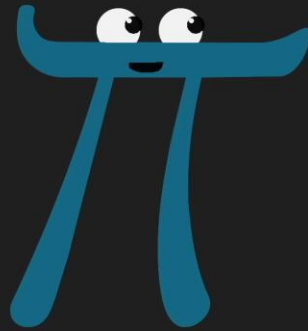
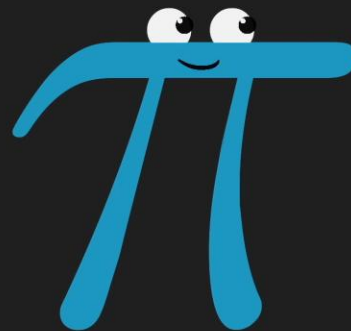
0.01 + 0.00 + 0.03 + 0.09 + 0.61 + 0.00 + 0.25 = 1.00



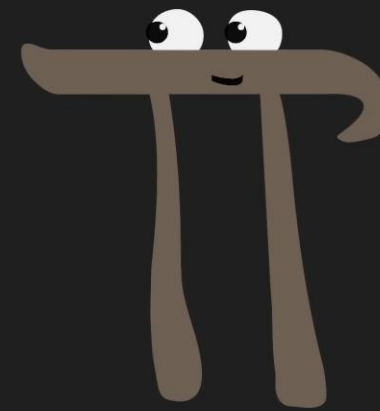
## 3.5 注意力模式

- 归一化方法：Softmax

|          |                                                     | a                                                         | fluffy                                                    | blue                                                      | creature                                                  | roamed                                                    | the                                                       | verdant                                                   | forest                                                    |                                      |
|----------|-----------------------------------------------------|-----------------------------------------------------------|-----------------------------------------------------------|-----------------------------------------------------------|-----------------------------------------------------------|-----------------------------------------------------------|-----------------------------------------------------------|-----------------------------------------------------------|-----------------------------------------------------------|--------------------------------------|
|          |                                                     | $\downarrow \vec{E}_1$<br>$\downarrow w_0$<br>$\vec{Q}_1$ | $\downarrow \vec{E}_2$<br>$\downarrow w_0$<br>$\vec{Q}_2$ | $\downarrow \vec{E}_3$<br>$\downarrow w_0$<br>$\vec{Q}_3$ | $\downarrow \vec{E}_4$<br>$\downarrow w_0$<br>$\vec{Q}_4$ | $\downarrow \vec{E}_5$<br>$\downarrow w_0$<br>$\vec{Q}_5$ | $\downarrow \vec{E}_6$<br>$\downarrow w_0$<br>$\vec{Q}_6$ | $\downarrow \vec{E}_7$<br>$\downarrow w_0$<br>$\vec{Q}_7$ | $\downarrow \vec{E}_8$<br>$\downarrow w_0$<br>$\vec{Q}_8$ |                                      |
| a        | $\rightarrow \vec{E}_1 \xrightarrow{w_1} \vec{K}_1$ | -0.7                                                      | -83.7                                                     | -24.7                                                     | -27.8                                                     | -5.2                                                      | -89.3                                                     | -45.2                                                     | -36.1                                                     |                                      |
| fluffy   | $\rightarrow \vec{E}_2 \xrightarrow{w_1} \vec{K}_2$ | 73.4                                                      | -2.9                                                      | -5.1                                                      | +93.0                                                     |                                                           |                                                           |                                                           |                                                           | We want these to<br>act like weights |
| blue     | $\rightarrow \vec{E}_3 \xrightarrow{w_1} \vec{K}_3$ | 53.4                                                      | -5.7                                                      | +1.8                                                      | +93.4                                                     |                                                           |                                                           |                                                           |                                                           |                                      |
| creature | $\rightarrow \vec{E}_4 \xrightarrow{w_1} \vec{K}_4$ | -21.5                                                     | 29.7                                                      | 56.1                                                      | +4.9                                                      | -32.4                                                     | -92.3                                                     | -9.5                                                      | -28.1                                                     |                                      |
| roamed   | $\rightarrow \vec{E}_5 \xrightarrow{w_1} \vec{K}_5$ | -20.1                                                     | -40.9                                                     | -87.8                                                     | -55.4                                                     | +0.6                                                      | -64.7                                                     | -96.7                                                     | -18.9                                                     |                                      |
| the      | $\rightarrow \vec{E}_6 \xrightarrow{w_1} \vec{K}_6$ | -97.9                                                     | -33.3                                                     | -22.6                                                     | -31.4                                                     | +5.5                                                      | +0.6                                                      | -4.6                                                      | -96.8                                                     |                                      |
| verdant  | $\rightarrow \vec{E}_7 \xrightarrow{w_1} \vec{K}_7$ | -41.2                                                     | 55.5                                                      | -42.3                                                     | -59.8                                                     | -79.0                                                     | -97.9                                                     | +3.7                                                      | +93.8                                                     |                                      |
| forest   | $\rightarrow \vec{E}_8 \xrightarrow{w_1} \vec{K}_8$ | -58.9                                                     | 75.5                                                      | 91.1                                                      | -90.6                                                     | -75.6                                                     | -89.0                                                     | 70.8                                                      | -4.7                                                      |                                      |

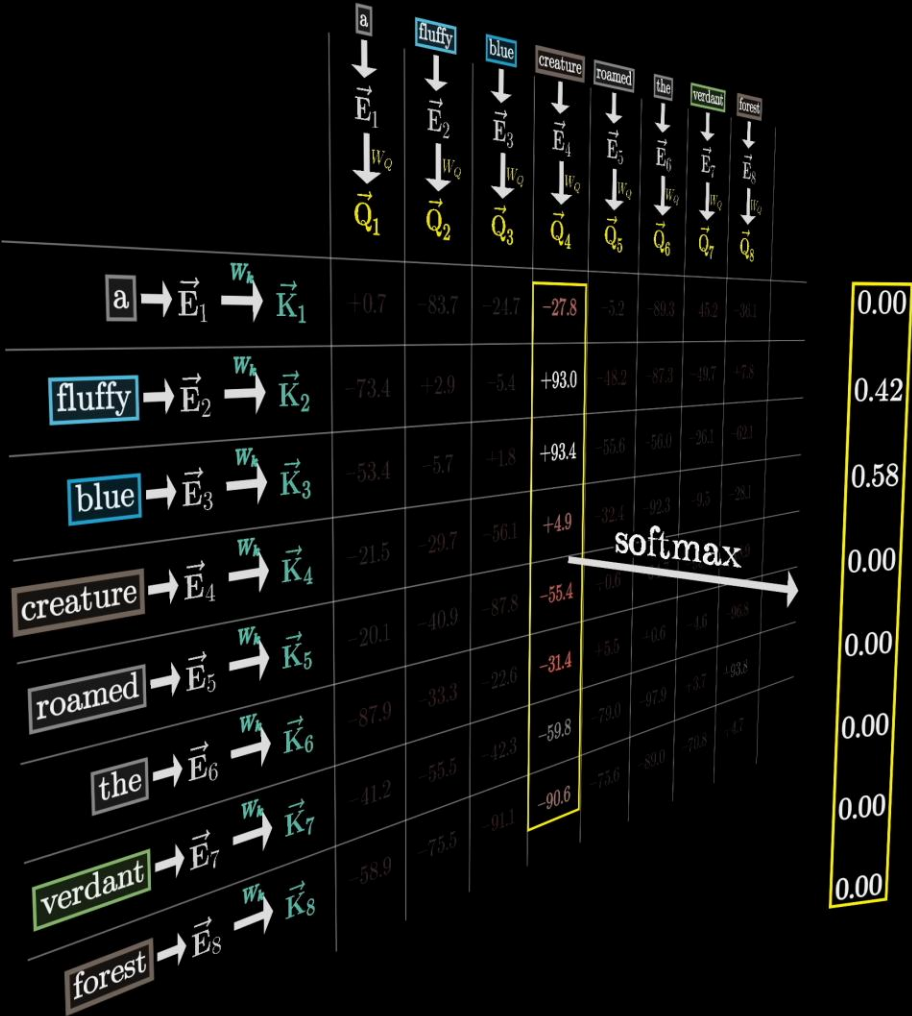


Softmax!



### 3.5 注意力模式

- 逐列进行Softmax归一化



### 3.5 注意力模式

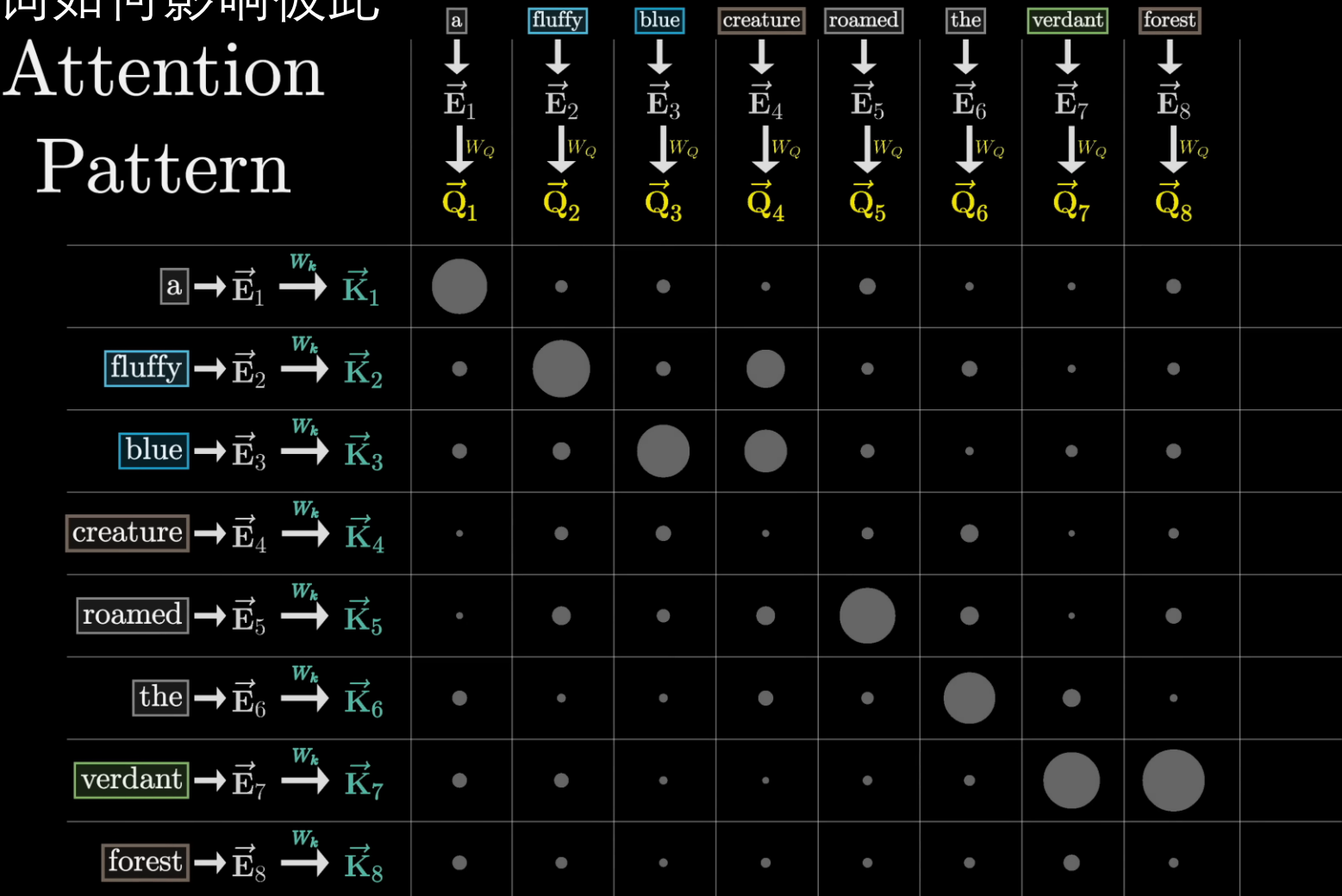
- 逐列进行Softmax归一化





### 3.5 注意力模式

- 将归一化的值填入表格，得到了注意力模式
- 注意力模式描述了词如何影响彼此



### 3.6 注意力公式

- 矩阵  $Q$ 、 $K$  分别由每个词嵌入的查询和键向量构成

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

Diagram illustrating the components of the Attention formula:

- The matrix  $Q$  (Query) is represented by a row vector of query vectors:  $[Q_1 \ Q_2 \ Q_3 \ Q_4 \ Q_5 \ \cdots \ Q_n]$ .
- The matrix  $K$  (Key) is represented by a row vector of key vectors:  $[K_1 \ K_2 \ K_3 \ K_4 \ K_5 \ \cdots \ K_n]$ .

## 3.6 注意力公式

- $QK^T$ 表示（归一化之前）的注意力模式

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

|          | $Q_1$           | $Q_2$           | $Q_3$           | $Q_4$           | $Q_5$           | $\dots$ | $Q_n$           |  |
|----------|-----------------|-----------------|-----------------|-----------------|-----------------|---------|-----------------|--|
| $K_1$    | $Q_1 \cdot K_1$ | $Q_2 \cdot K_1$ | $Q_3 \cdot K_1$ | $Q_4 \cdot K_1$ | $Q_5 \cdot K_1$ | $\dots$ | $Q_n \cdot K_1$ |  |
| $K_2$    | $Q_1 \cdot K_2$ | $Q_2 \cdot K_2$ | $Q_3 \cdot K_2$ | $Q_4 \cdot K_2$ | $Q_5 \cdot K_2$ | $\dots$ | $Q_n \cdot K_2$ |  |
| $K_3$    | $Q_1 \cdot K_3$ | $Q_2 \cdot K_3$ | $Q_3 \cdot K_3$ | $Q_4 \cdot K_3$ | $Q_5 \cdot K_3$ | $\dots$ | $Q_n \cdot K_3$ |  |
| $K_4$    | $Q_1 \cdot K_4$ | $Q_2 \cdot K_4$ | $Q_3 \cdot K_4$ | $Q_4 \cdot K_4$ | $Q_5 \cdot K_4$ | $\dots$ | $Q_n \cdot K_4$ |  |
| $K_5$    | $Q_1 \cdot K_5$ | $Q_2 \cdot K_5$ | $Q_3 \cdot K_5$ | $Q_4 \cdot K_5$ | $Q_5 \cdot K_5$ | $\dots$ | $Q_n \cdot K_5$ |  |
| $\vdots$ | $\vdots$        | $\vdots$        | $\vdots$        | $\vdots$        | $\vdots$        | $\dots$ | $\vdots$        |  |

### 3.6 注意力公式

- 分母中的 $d_k = 128$ 是查询/键的维度

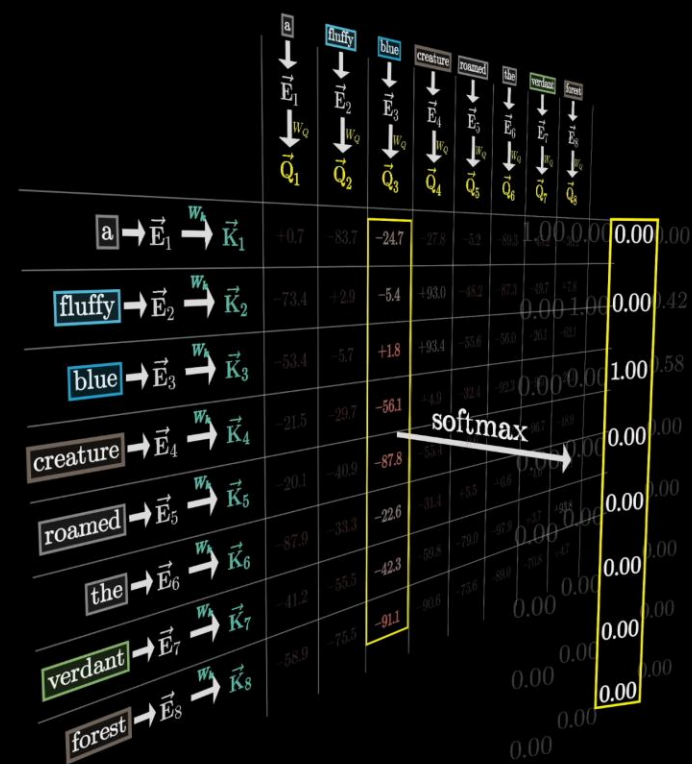
$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

|          | $Q_1$                              | $Q_2$                              | $Q_3$                              | $Q_4$                              | $Q_5$                              | $\dots$ | $Q_n$                              |  |
|----------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|---------|------------------------------------|--|
| $K_1$    | $\frac{Q_1 \cdot K_1}{\sqrt{d_k}}$ | $\frac{Q_2 \cdot K_1}{\sqrt{d_k}}$ | $\frac{Q_3 \cdot K_1}{\sqrt{d_k}}$ | $\frac{Q_4 \cdot K_1}{\sqrt{d_k}}$ | $\frac{Q_5 \cdot K_1}{\sqrt{d_k}}$ | $\dots$ | $\frac{Q_n \cdot K_1}{\sqrt{d_k}}$ |  |
| $K_2$    | $\frac{Q_1 \cdot K_2}{\sqrt{d_k}}$ | $\frac{Q_2 \cdot K_2}{\sqrt{d_k}}$ | $\frac{Q_3 \cdot K_2}{\sqrt{d_k}}$ | $\frac{Q_4 \cdot K_2}{\sqrt{d_k}}$ | $\frac{Q_5 \cdot K_2}{\sqrt{d_k}}$ | $\dots$ | $\frac{Q_n \cdot K_2}{\sqrt{d_k}}$ |  |
| $K_3$    | $\frac{Q_1 \cdot K_3}{\sqrt{d_k}}$ | $\frac{Q_2 \cdot K_3}{\sqrt{d_k}}$ | $\frac{Q_3 \cdot K_3}{\sqrt{d_k}}$ | $\frac{Q_4 \cdot K_3}{\sqrt{d_k}}$ | $\frac{Q_5 \cdot K_3}{\sqrt{d_k}}$ | $\dots$ | $\frac{Q_n \cdot K_3}{\sqrt{d_k}}$ |  |
| $K_4$    | $\frac{Q_1 \cdot K_4}{\sqrt{d_k}}$ | $\frac{Q_2 \cdot K_4}{\sqrt{d_k}}$ | $\frac{Q_3 \cdot K_4}{\sqrt{d_k}}$ | $\frac{Q_4 \cdot K_4}{\sqrt{d_k}}$ | $\frac{Q_5 \cdot K_4}{\sqrt{d_k}}$ | $\dots$ | $\frac{Q_n \cdot K_4}{\sqrt{d_k}}$ |  |
| $K_5$    | $\frac{Q_1 \cdot K_5}{\sqrt{d_k}}$ | $\frac{Q_2 \cdot K_5}{\sqrt{d_k}}$ | $\frac{Q_3 \cdot K_5}{\sqrt{d_k}}$ | $\frac{Q_4 \cdot K_5}{\sqrt{d_k}}$ | $\frac{Q_5 \cdot K_5}{\sqrt{d_k}}$ | $\dots$ | $\frac{Q_n \cdot K_5}{\sqrt{d_k}}$ |  |
| $\vdots$ | $\vdots$                           | $\vdots$                           | $\vdots$                           | $\vdots$                           | $\vdots$                           | $\dots$ | $\vdots$                           |  |

### 3.6 注意力公式

- Softmax函数对输入的矩阵逐列做归一化，得到注意力模式

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



1. 回顾
2. 启发示例：注意力的作用
3. 注意力机制：查询（Query）、键（Key）
4. **掩码（Masking）**
5. 注意力机制：值（Value）
6. 参数数量
7. 多头注意力
8. Transformer的深度

## 4.1 模型训练

- 同时预测输入文本中每一个词的下一个词
- 这样，一个样本就能提供多次训练的机会

the → ???

the fluffy → ???

the fluffy blue → ???

the fluffy blue creature → ???

the fluffy blue creature roamed → ???

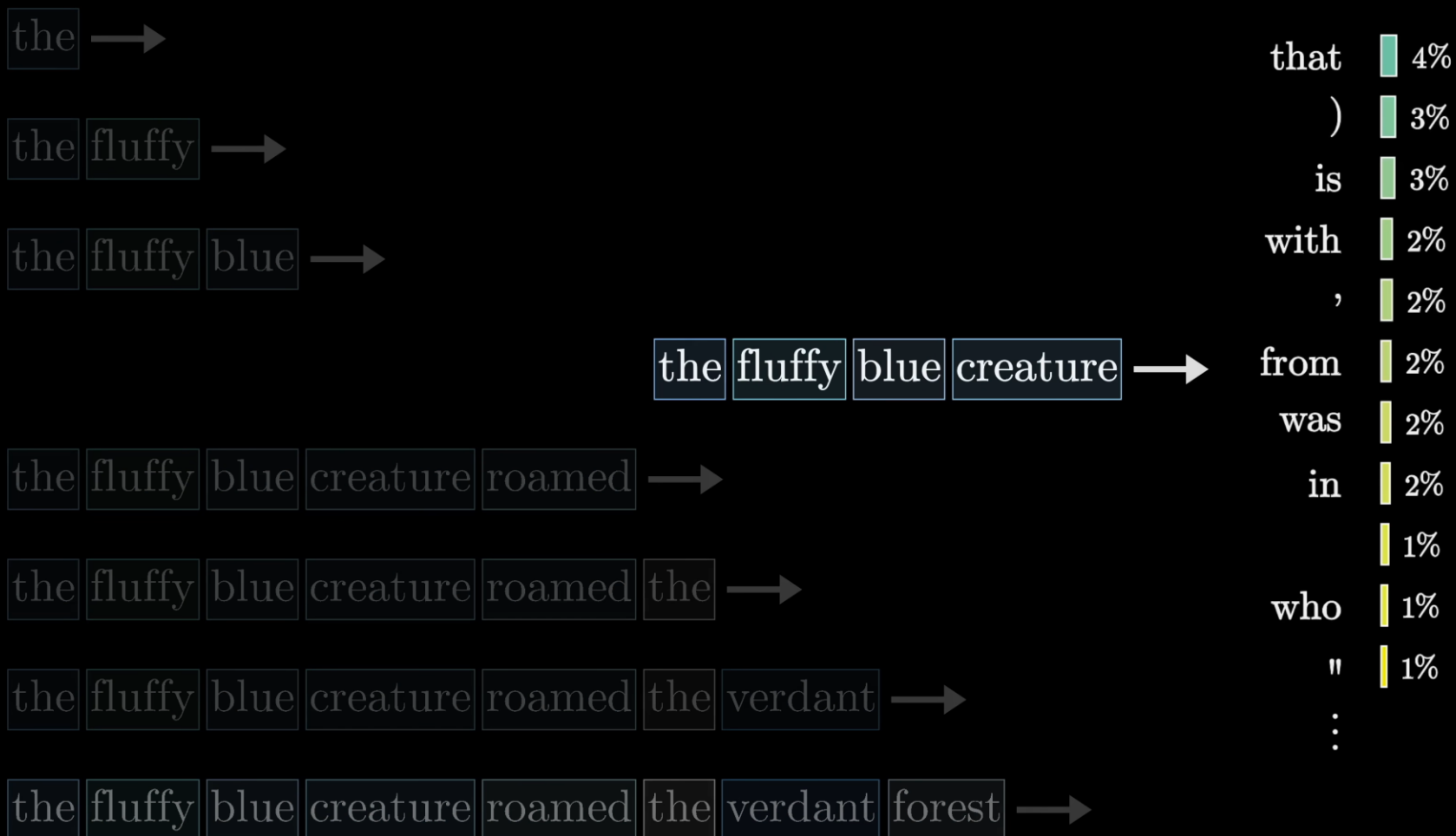
the fluffy blue creature roamed the → ???

the fluffy blue creature roamed the verdant → ???

the fluffy blue creature roamed the verdant forest → ???

## 4.1 模型训练

- 根据预测结果与实际出现的词的差异，来更新模型参数（权重）





## 4.1 模型训练

- 根据预测结果与实际出现的词的差异，来更新模型参数（权重）

the →

the fluffy →

the fluffy blue →

the fluffy blue creature →

the fluffy blue creature roamed →

the fluffy blue creature roamed the →

the fluffy blue creature roamed the verdant →

the fluffy blue creature roamed the verdant forest →

streets 8%

forest 6%

land 5%

halls 4%

fields 1%

world 1%

city 1%

earth 1%

room 1%

woods 1%

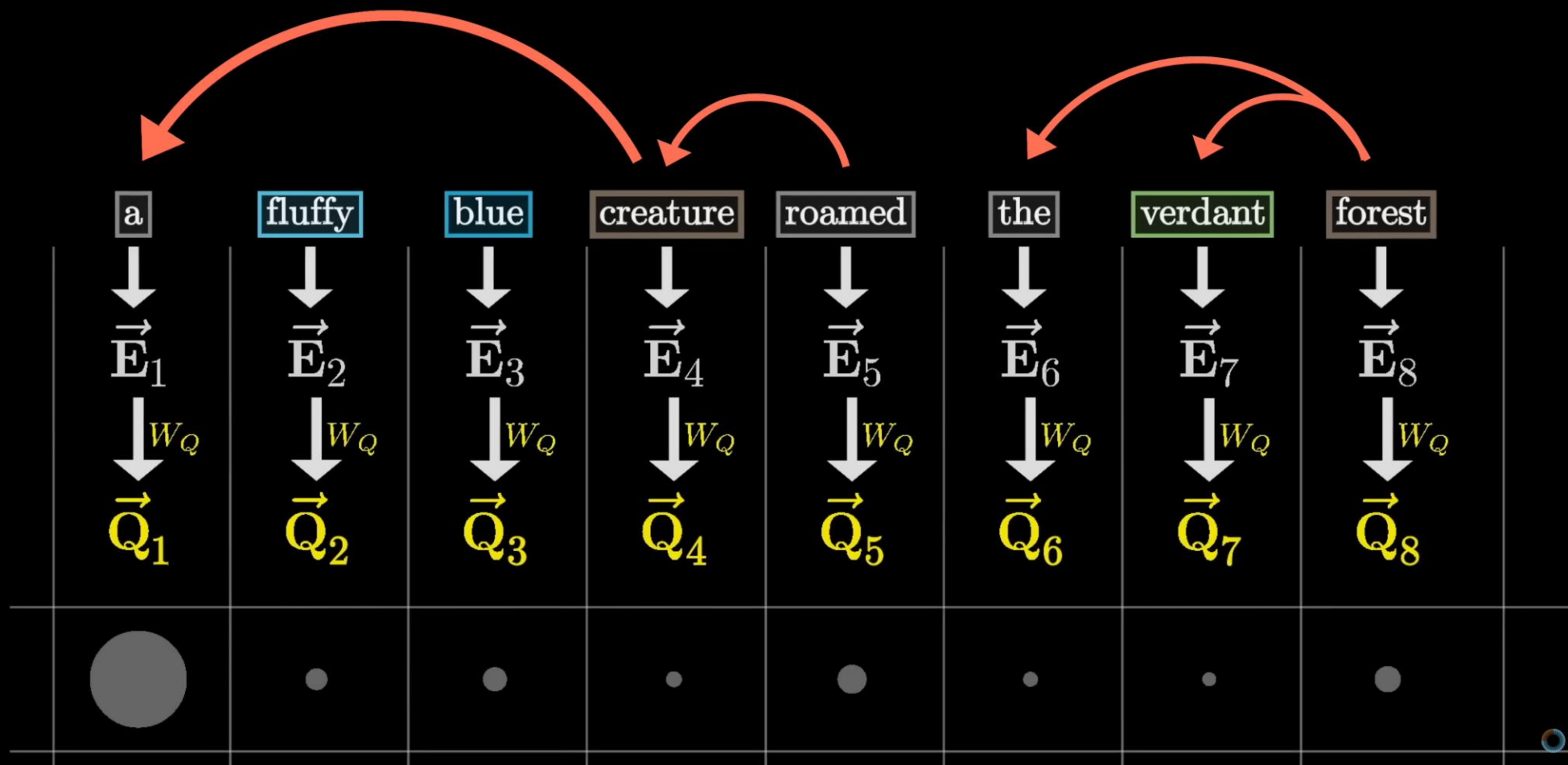
countryside 1%

vast 1%

⋮

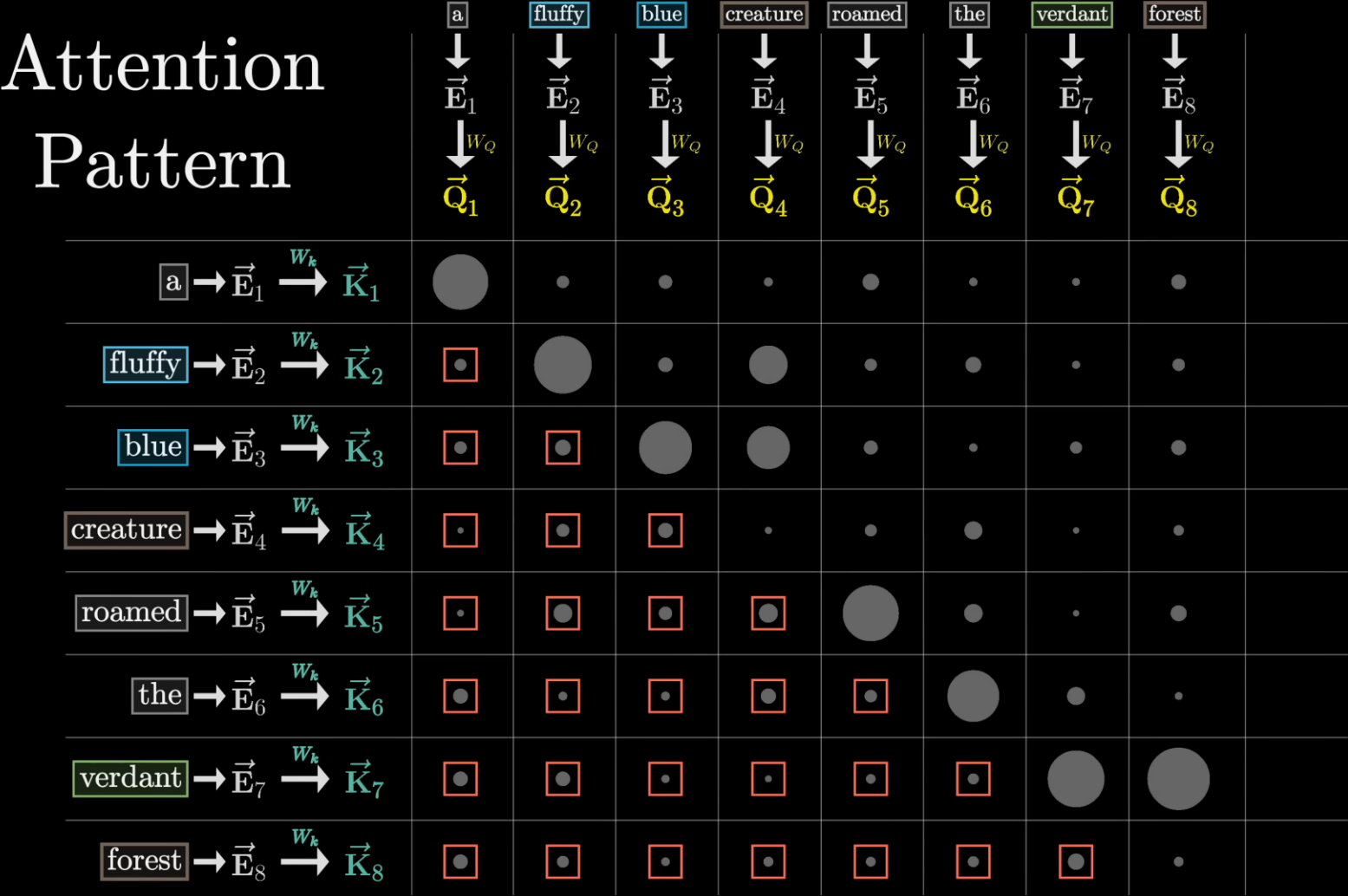
## 4.1 模型训练

- 因此，就不能够让后面的词影响之前的词语，否则就会“泄露答案”



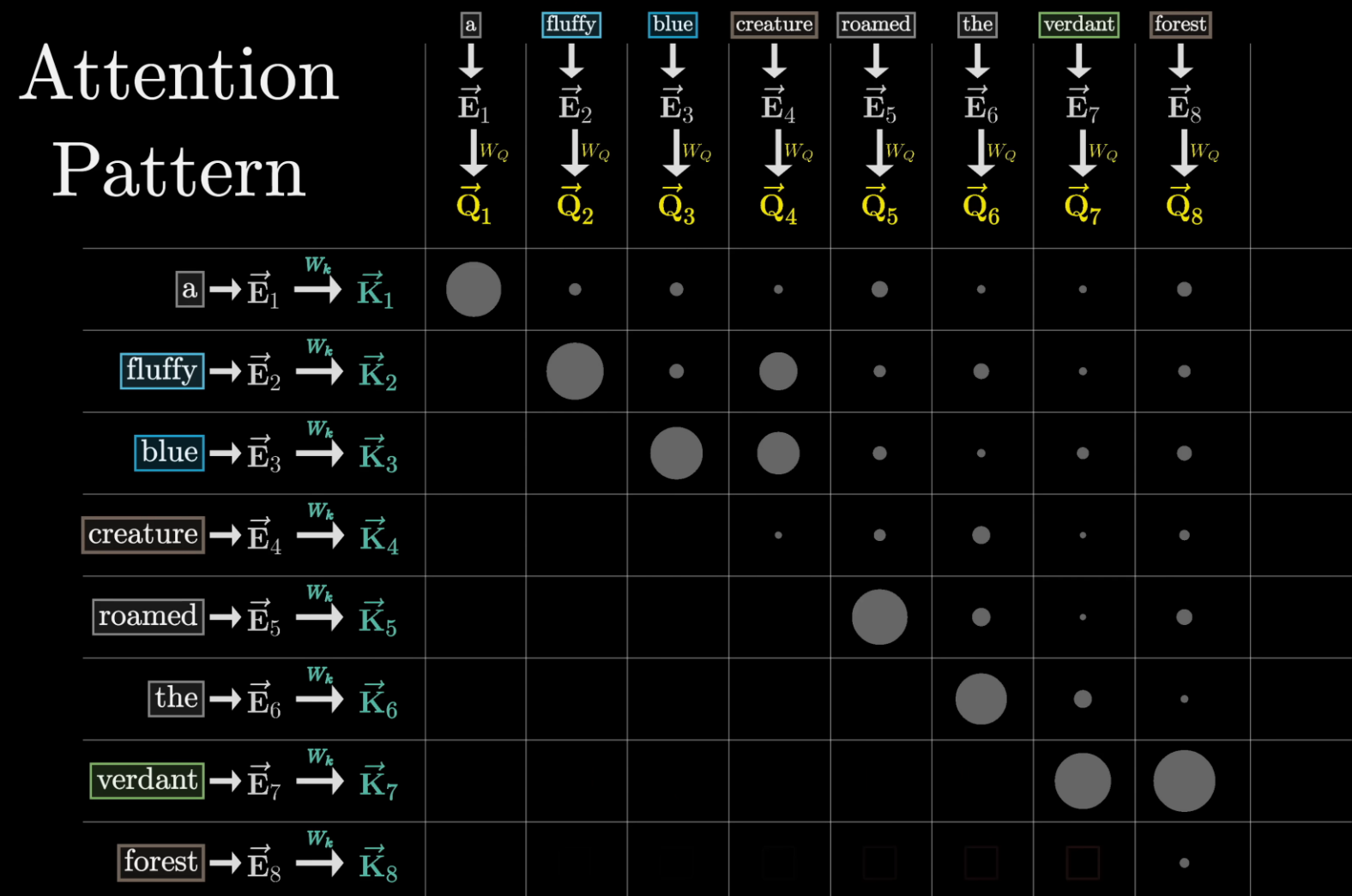
## 4.2 掩码

- 在注意力模式矩阵中，将的下三角阵“归零”



## 4.2 掩码

- 在注意力模式矩阵中，将的下三角阵“归零”



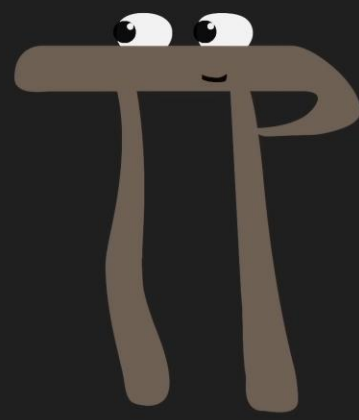
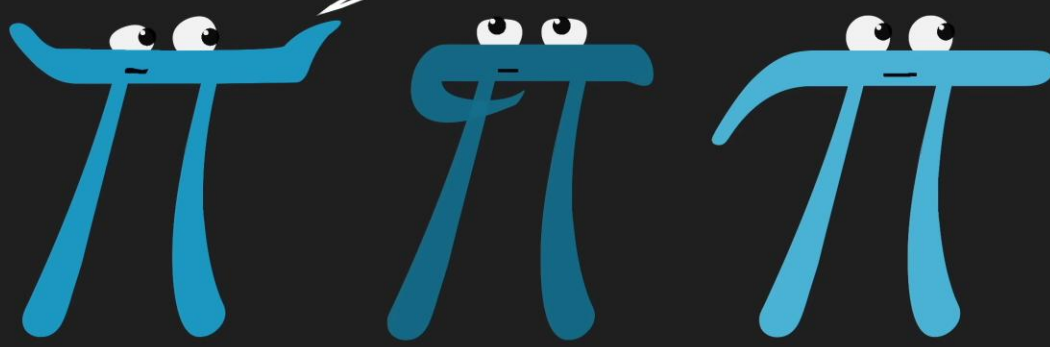
## 4.2 掩码

- 但不能直接设为0

| Attention Pattern                                                     | $\begin{matrix} \text{[a]} \\ \downarrow \vec{E}_1 \\ \downarrow w_q \\ \vec{Q}_1 \end{matrix}$ | $\begin{matrix} \text{[fluffy]} \\ \downarrow \vec{E}_2 \\ \downarrow w_q \\ \vec{Q}_2 \end{matrix}$ | $\begin{matrix} \text{[blue]} \\ \downarrow \vec{E}_3 \\ \downarrow w_q \\ \vec{Q}_3 \end{matrix}$ | $\begin{matrix} \text{[creature]} \\ \downarrow \vec{E}_4 \\ \downarrow w_q \\ \vec{Q}_4 \end{matrix}$ | $\begin{matrix} \text{[roamed]} \\ \downarrow \vec{E}_5 \\ \downarrow w_q \\ \vec{Q}_5 \end{matrix}$ | $\begin{matrix} \text{[the]} \\ \downarrow \vec{E}_6 \\ \downarrow w_q \\ \vec{Q}_6 \end{matrix}$ | $\begin{matrix} \text{[verdant]} \\ \downarrow \vec{E}_7 \\ \downarrow w_q \\ \vec{Q}_7 \end{matrix}$ | $\begin{matrix} \text{[forest]} \\ \downarrow \vec{E}_8 \\ \downarrow w_q \\ \vec{Q}_8 \end{matrix}$ |
|-----------------------------------------------------------------------|-------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------|
| $\text{[a]} \rightarrow \vec{E}_1 \xrightarrow{w_k} \vec{K}_1$        | ●                                                                                               | •                                                                                                    | •                                                                                                  | •                                                                                                      | •                                                                                                    | •                                                                                                 | •                                                                                                     |                                                                                                      |
| $\text{[fluffy]} \rightarrow \vec{E}_2 \xrightarrow{w_k} \vec{K}_2$   | □                                                                                               | ●                                                                                                    | •                                                                                                  | •                                                                                                      | •                                                                                                    | •                                                                                                 | •                                                                                                     |                                                                                                      |
| $\text{[blue]} \rightarrow \vec{E}_3 \xrightarrow{w_k} \vec{K}_3$     | □                                                                                               | □                                                                                                    | ●                                                                                                  | •                                                                                                      | •                                                                                                    | •                                                                                                 | •                                                                                                     |                                                                                                      |
| $\text{[creature]} \rightarrow \vec{E}_4 \xrightarrow{w_k} \vec{K}_4$ | □                                                                                               | □                                                                                                    | □                                                                                                  | •                                                                                                      | •                                                                                                    | •                                                                                                 | •                                                                                                     |                                                                                                      |
| $\text{[roamed]} \rightarrow \vec{E}_5 \xrightarrow{w_k} \vec{K}_5$   | □                                                                                               | □                                                                                                    | □                                                                                                  | □                                                                                                      | ●                                                                                                    | •                                                                                                 | •                                                                                                     |                                                                                                      |
| $\text{[the]} \rightarrow \vec{E}_6 \xrightarrow{w_k} \vec{K}_6$      | □                                                                                               | □                                                                                                    | □                                                                                                  | □                                                                                                      | □                                                                                                    | ●                                                                                                 | •                                                                                                     |                                                                                                      |
| $\text{[verdant]} \rightarrow \vec{E}_7 \xrightarrow{w_k} \vec{K}_7$  | □                                                                                               | □                                                                                                    | □                                                                                                  | □                                                                                                      | □                                                                                                    | □                                                                                                 | ●                                                                                                     | ●                                                                                                    |
| $\text{[forest]} \rightarrow \vec{E}_8 \xrightarrow{w_k} \vec{K}_8$   | □                                                                                               | □                                                                                                    | □                                                                                                  | □                                                                                                      | □                                                                                                    | □                                                                                                 | □                                                                                                     | •                                                                                                    |

Then they wouldn't be normalized 它们将不再是归一化的

Set them to 0?



## 4.2 掩码

- 需将未归一化的注意力模式矩阵的下三角阵设为 $-\infty$

Unnormalized  
Attention Pattern

|           |           |           |           |           |       |
|-----------|-----------|-----------|-----------|-----------|-------|
| +3.53     | +0.80     | +1.96     | +4.48     | +3.74     | -1.95 |
| $-\infty$ | -0.30     | -0.21     | +0.82     | +0.29     | +2.91 |
| $-\infty$ | $-\infty$ | +0.89     | +0.67     | +2.99     | -0.41 |
| $-\infty$ | $-\infty$ | $-\infty$ | +1.31     | +1.73     | -1.48 |
| $-\infty$ | $-\infty$ | $-\infty$ | $-\infty$ | +3.07     | +2.94 |
| $-\infty$ | $-\infty$ | $-\infty$ | $-\infty$ | $-\infty$ | +0.31 |

softmax  
→

Normalized  
Attention Pattern

|  |  |  |  |  |  |
|--|--|--|--|--|--|
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

## 4.2 掩码

- 才能够使softmax函数的输出为0，且保证注意力模式满足归一化要求

Unnormalized  
Attention Pattern

|           |           |           |           |           |       |
|-----------|-----------|-----------|-----------|-----------|-------|
| +3.53     | +0.80     | +1.96     | +4.48     | +3.74     | -1.95 |
| $-\infty$ | -0.30     | -0.21     | +0.82     | +0.29     | +2.91 |
| $-\infty$ | $-\infty$ | +0.89     | +0.67     | +2.99     | -0.41 |
| $-\infty$ | $-\infty$ | $-\infty$ | +1.31     | +1.73     | -1.48 |
| $-\infty$ | $-\infty$ | $-\infty$ | $-\infty$ | +3.07     | +2.94 |
| $-\infty$ | $-\infty$ | $-\infty$ | $-\infty$ | $-\infty$ | +0.31 |

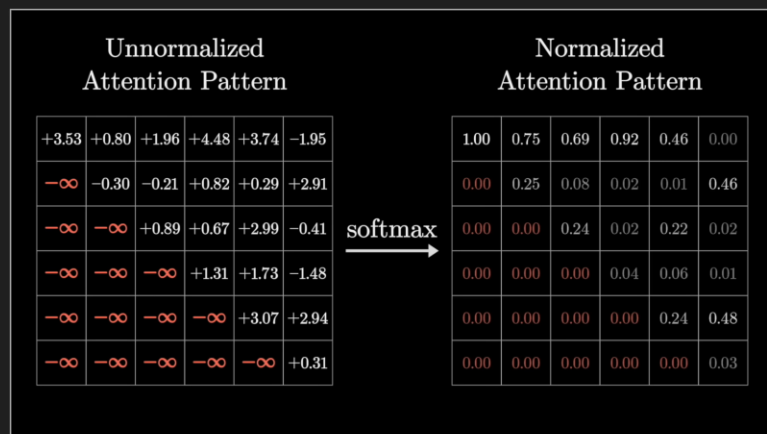
softmax  
→

Normalized  
Attention Pattern

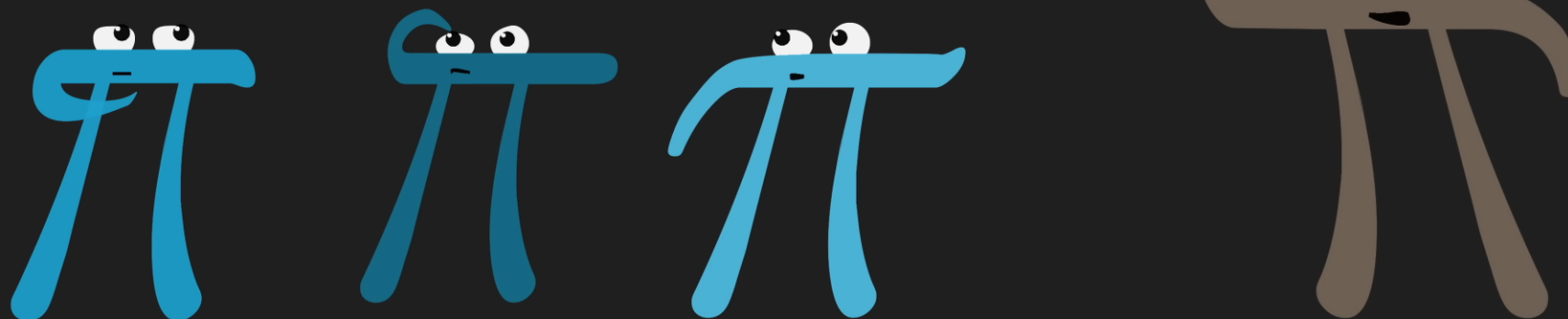
|      |      |      |      |      |      |
|------|------|------|------|------|------|
| 1.00 | 0.75 | 0.69 | 0.92 | 0.46 | 0.00 |
| 0.00 | 0.25 | 0.08 | 0.02 | 0.01 | 0.46 |
| 0.00 | 0.00 | 0.24 | 0.02 | 0.22 | 0.02 |
| 0.00 | 0.00 | 0.00 | 0.04 | 0.06 | 0.01 |
| 0.00 | 0.00 | 0.00 | 0.00 | 0.24 | 0.48 |
| 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.03 |

## 4.2 掩码

- 这种操作被称为“掩码 (masking)”



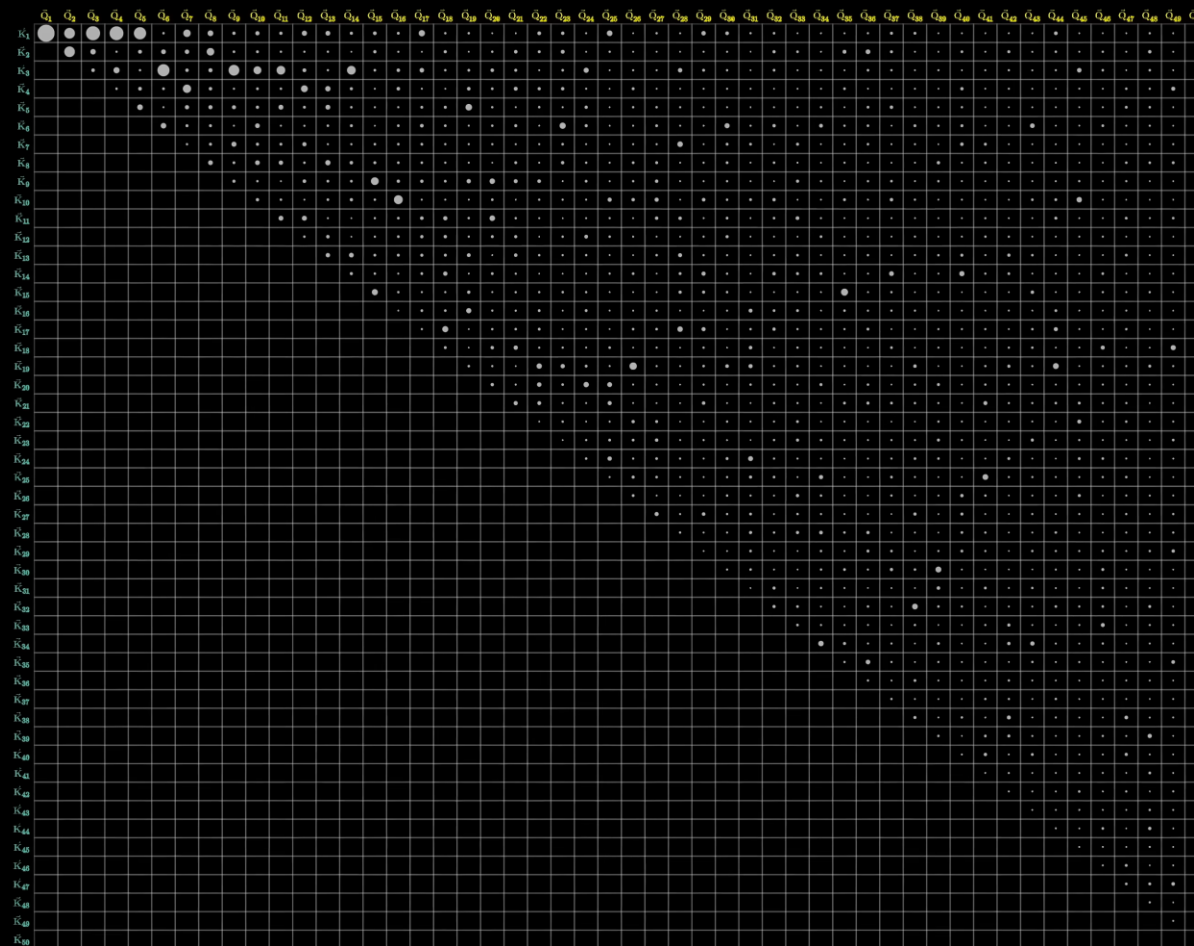
This is called  
“masking”





## 4.3 上下文长度瓶颈

- 注意力模式矩阵的大小等于上下文长度的平方
- 这成为了大模型上下文长度的瓶颈



## 4.3 上下文长度瓶颈

- 对此，现在有了很多新的Transformer架构，在此略过
  - Sparse Attention Mechanisms
  - Blockwise Attention
  - Linformer
  - Reformer
  - Ring attention
  - Longformer
  - Adaptive Attention Span
  - ⋮

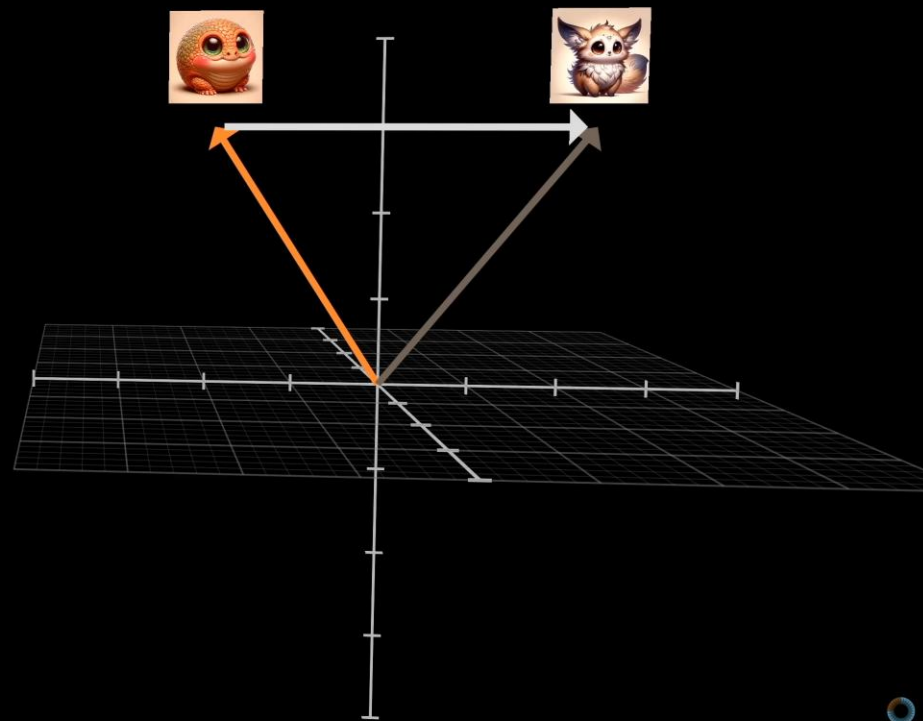
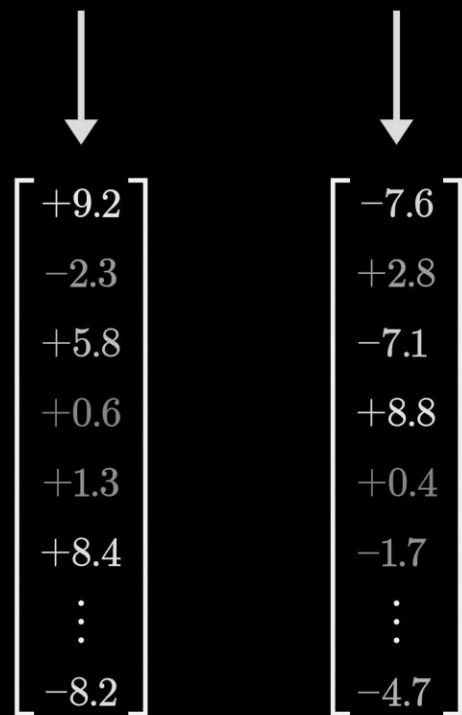


1. 回顾
2. 启发示例：注意力的作用
3. 注意力机制：查询（Query）、键（Key）
4. 掩码（Masking）
- 5. 注意力机制：值（Value）**
6. 参数数量
7. 交叉注意力
8. 多头注意力
9. Transformer的深度

## 5.1 更新的几何意义

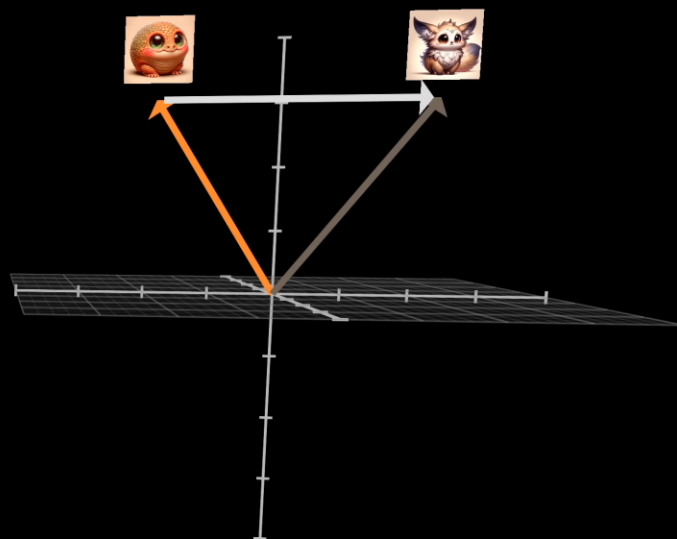
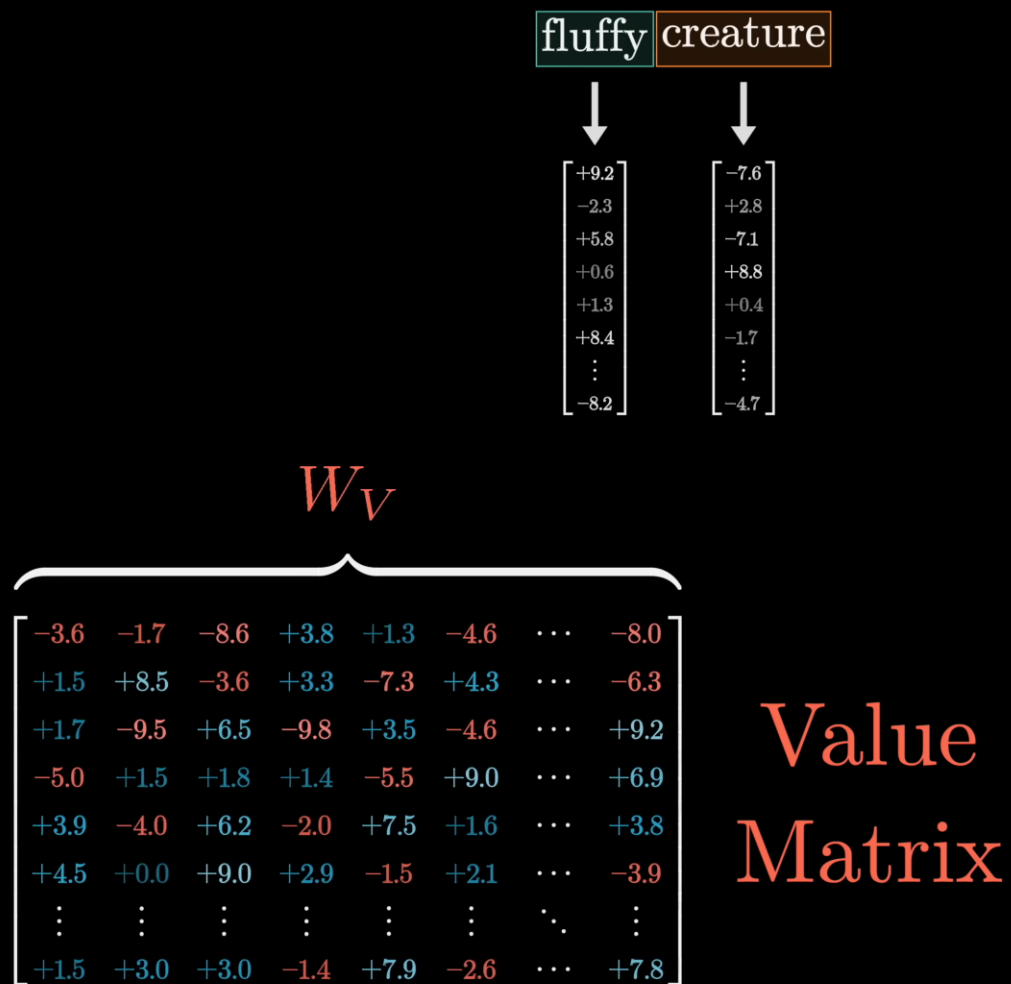
- 回到形容词修饰名词的例子中

fluffy creature



## 5.1 更新的几何意义

- 用值矩阵 (value matrix)  $W_V$  乘以嵌入向量, 得到值 (value) 向量

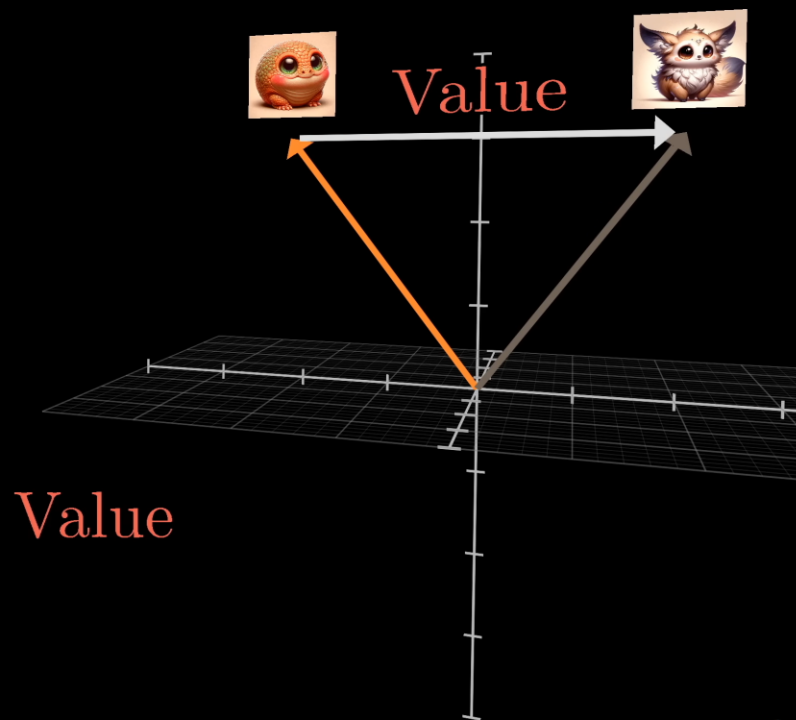


## 5.1 更新的几何意义

- 用值矩阵 (value matrix)  $W_V$  乘以嵌入向量, 得到值 (value) 向量

The diagram illustrates the calculation of a value vector. At the top, the words "fluffy" and "creature" are shown in boxes. Arrows point from "fluffy" to a vertical vector of values:  $\begin{bmatrix} +9.2 \\ -2.3 \\ +5.8 \\ +0.6 \\ +1.3 \\ +8.4 \\ \vdots \\ -8.2 \end{bmatrix}$ . An arrow from "creature" points to another vertical vector:  $\begin{bmatrix} -7.6 \\ +2.8 \\ -7.1 \\ +8.8 \\ +0.4 \\ -1.7 \\ \vdots \\ -4.7 \end{bmatrix}$ . Below these, a large matrix  $W_V$  is shown, with a bracket above it. A curved arrow connects the "fluffy" vector to the  $W_V$  matrix. The matrix  $W_V$  is a 10x10 grid of values. To the right of the matrix is an equals sign followed by a vertical vector of values, which is the result of the multiplication.

$$W_V \begin{bmatrix} +9.2 \\ -2.3 \\ +5.8 \\ +0.6 \\ +1.3 \\ +8.4 \\ \vdots \\ -8.2 \end{bmatrix} = \begin{bmatrix} -52.4 \\ +89.3 \\ -80.2 \\ -17.8 \\ +7.3 \\ +223.8 \\ \vdots \\ -41.0 \end{bmatrix}$$



## 5.1 更新的几何意义

- 嵌入向量+值向量=更新嵌入向量

Diagram illustrating the update of an embedding vector using a value vector.

Input words and their corresponding value vectors (columns):

- blue:  $\begin{bmatrix} +1.0 \\ +4.3 \\ +2.0 \\ +0.9 \\ -1.5 \\ +2.9 \\ \vdots \\ +7.8 \end{bmatrix}$
- fluffy:  $\begin{bmatrix} +9.2 \\ -2.3 \\ +5.8 \\ +0.6 \\ +1.3 \\ +8.4 \\ \vdots \\ -8.2 \end{bmatrix}$
- creature:  $\begin{bmatrix} -7.6 \\ +2.8 \\ -7.1 \\ +8.8 \\ +0.4 \\ -1.7 \\ \vdots \\ -4.7 \end{bmatrix}$

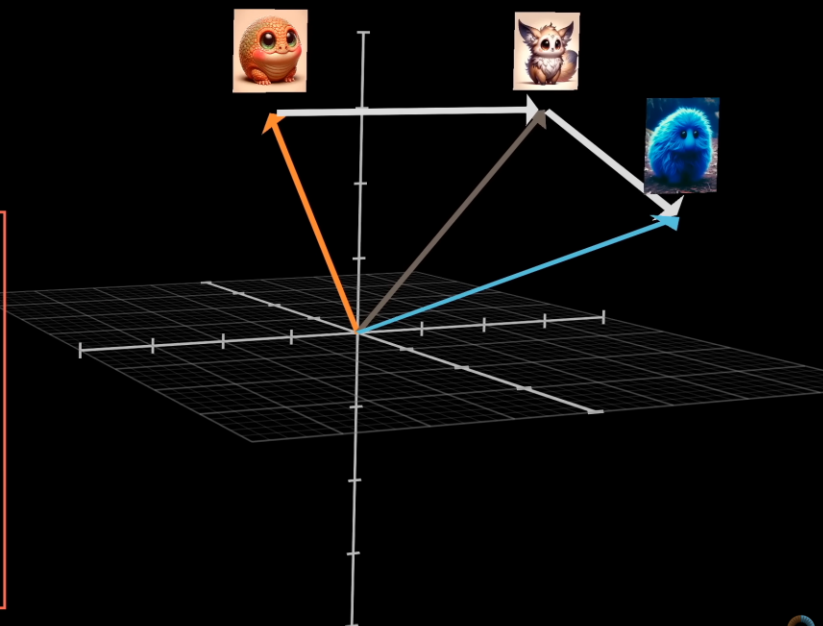
The value vector for "blue" is multiplied by the weight matrix  $W_V$  (indicated by a blue arrow) to produce the update vector.

Weight matrix  $W_V$  (rows correspond to the dimensions of the embedding vector):

$$\begin{bmatrix} -3.6 & -1.7 & -8.6 & +3.8 & +1.3 & -4.6 & \cdots & -8.0 \\ +1.5 & +8.5 & -3.6 & +3.3 & -7.3 & +4.3 & \cdots & -6.3 \\ +1.7 & -9.5 & +6.5 & -9.8 & +3.5 & -4.6 & \cdots & +9.2 \\ -5.0 & +1.5 & +1.8 & +1.4 & -5.5 & +9.0 & \cdots & +6.9 \\ +3.9 & -4.0 & +6.2 & -2.0 & +7.5 & +1.6 & \cdots & +3.8 \\ +4.5 & +0.0 & +9.0 & +2.9 & -1.5 & +2.1 & \cdots & -3.9 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ +1.5 & +3.0 & +3.0 & -1.4 & +7.9 & -2.6 & \cdots & +7.8 \end{bmatrix}$$

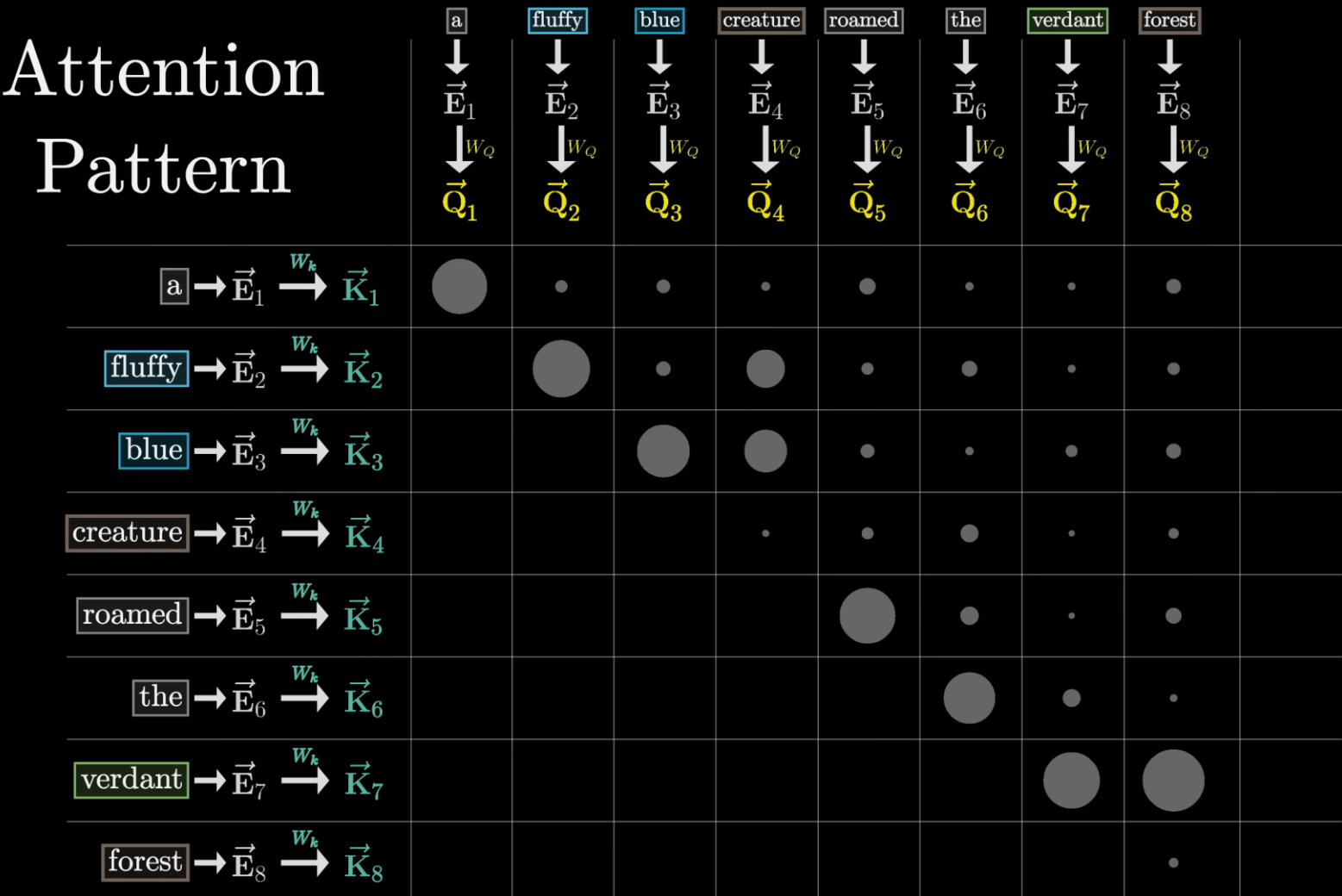
The resulting update vector (shown in a red box) is:

$$\begin{bmatrix} -103.0 \\ +13.1 \\ +12.6 \\ +95.7 \\ +11.2 \\ +14.4 \\ \vdots \\ +62.1 \end{bmatrix}$$



## 5.2 嵌入向量更新

- 注意力模式矩阵中





## 5.2 嵌入向量更新

- 此时不再需要Q或K了，我们关注嵌入向量 (E) 和注值 (V)

| Value matrix                                 |  | a                                                                                   | fluffy                                                                              | blue                                                                                | creature                                                                            | roamed                                                                              | the                                                                                  | verdant                                                                               | forest                                                                                |  |
|----------------------------------------------|--|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|--|
| $W_V$                                        |  | $\downarrow$<br>$\vec{E}_1$                                                         | $\downarrow$<br>$\vec{E}_2$                                                         | $\downarrow$<br>$\vec{E}_3$                                                         | $\downarrow$<br>$\vec{E}_4$                                                         | $\downarrow$<br>$\vec{E}_5$                                                         | $\downarrow$<br>$\vec{E}_6$                                                          | $\downarrow$<br>$\vec{E}_7$                                                           | $\downarrow$<br>$\vec{E}_8$                                                           |  |
| a $\rightarrow \vec{E}_1 \rightarrow$        |  |  |  |  |  |  |   |    |    |  |
| fluffy $\rightarrow \vec{E}_2 \rightarrow$   |  |                                                                                     |  |  |  |  |   |    |    |  |
| blue $\rightarrow \vec{E}_3 \rightarrow$     |  |                                                                                     |                                                                                     |  |  |  |   |    |    |  |
| creature $\rightarrow \vec{E}_4 \rightarrow$ |  |                                                                                     |                                                                                     |                                                                                     |  |  |   |    |    |  |
| roamed $\rightarrow \vec{E}_5 \rightarrow$   |  |                                                                                     |                                                                                     |                                                                                     |                                                                                     |  |   |    |    |  |
| the $\rightarrow \vec{E}_6 \rightarrow$      |  |                                                                                     |                                                                                     |                                                                                     |                                                                                     |                                                                                     |  |  |  |  |
| verdant $\rightarrow \vec{E}_7 \rightarrow$  |  |                                                                                     |                                                                                     |                                                                                     |                                                                                     |                                                                                     |                                                                                      |  |  |  |
| forest $\rightarrow \vec{E}_8 \rightarrow$   |  |                                                                                     |                                                                                     |                                                                                     |                                                                                     |                                                                                     |                                                                                      |                                                                                       |  |  |

## 5.2 嵌入向量更新

- 此时不再需要Q或K了，我们关注嵌入向量（E）和注值（V）

| Value matrix                                                 |  | a                           | fluffy                      | blue                        | creature                    | roamed                      | the                         | verdant                     | forest                      |  |
|--------------------------------------------------------------|--|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|--|
| $W_V$                                                        |  | $\downarrow$<br>$\vec{E}_1$ | $\downarrow$<br>$\vec{E}_2$ | $\downarrow$<br>$\vec{E}_3$ | $\downarrow$<br>$\vec{E}_4$ | $\downarrow$<br>$\vec{E}_5$ | $\downarrow$<br>$\vec{E}_6$ | $\downarrow$<br>$\vec{E}_7$ | $\downarrow$<br>$\vec{E}_8$ |  |
| a $\rightarrow \vec{E}_1 \xrightarrow{W_V} \vec{V}_1$        |  |                             |                             |                             |                             |                             |                             |                             |                             |  |
| fluffy $\rightarrow \vec{E}_2 \xrightarrow{W_V} \vec{V}_2$   |  |                             |                             |                             |                             |                             |                             |                             |                             |  |
| blue $\rightarrow \vec{E}_3 \xrightarrow{W_V} \vec{V}_3$     |  |                             |                             |                             |                             |                             |                             |                             |                             |  |
| creature $\rightarrow \vec{E}_4 \xrightarrow{W_V} \vec{V}_4$ |  |                             |                             |                             |                             |                             |                             |                             |                             |  |
| roamed $\rightarrow \vec{E}_5 \xrightarrow{W_V} \vec{V}_5$   |  |                             |                             |                             |                             |                             |                             |                             |                             |  |
| the $\rightarrow \vec{E}_6 \xrightarrow{W_V} \vec{V}_6$      |  |                             |                             |                             |                             |                             |                             |                             |                             |  |
| verdant $\rightarrow \vec{E}_7 \xrightarrow{W_V} \vec{V}_7$  |  |                             |                             |                             |                             |                             |                             |                             |                             |  |
| forest $\rightarrow \vec{E}_8 \xrightarrow{W_V} \vec{V}_8$   |  |                             |                             |                             |                             |                             |                             |                             |                             |  |

## 5.2 嵌入向量更新

- 根据注意力的大小，对值进行加权求和，得到嵌入向量的更新增量

| Value matrix                                                 | a                           | fluffy                      | blue                        | creature                    | roamed                      | the                         | verdant                     | forest                      |  |
|--------------------------------------------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|--|
| $W_V$                                                        | $\downarrow$<br>$\vec{E}_1$ | $\downarrow$<br>$\vec{E}_2$ | $\downarrow$<br>$\vec{E}_3$ | $\downarrow$<br>$\vec{E}_4$ | $\downarrow$<br>$\vec{E}_5$ | $\downarrow$<br>$\vec{E}_6$ | $\downarrow$<br>$\vec{E}_7$ | $\downarrow$<br>$\vec{E}_8$ |  |
| a $\rightarrow \vec{E}_1 \xrightarrow{W_V} \vec{V}_1$        |                             |                             |                             | 0.00                        |                             |                             |                             |                             |  |
| fluffy $\rightarrow \vec{E}_2 \xrightarrow{W_V} \vec{V}_2$   |                             |                             |                             | 0.42                        |                             |                             |                             |                             |  |
| blue $\rightarrow \vec{E}_3 \xrightarrow{W_V} \vec{V}_3$     |                             |                             |                             | 0.58                        |                             |                             |                             |                             |  |
| creature $\rightarrow \vec{E}_4 \xrightarrow{W_V} \vec{V}_4$ |                             |                             |                             | 0.00                        |                             |                             |                             |                             |  |
| roamed $\rightarrow \vec{E}_5 \xrightarrow{W_V} \vec{V}_5$   |                             |                             |                             | 0.00                        |                             |                             |                             |                             |  |
| the $\rightarrow \vec{E}_6 \xrightarrow{W_V} \vec{V}_6$      |                             |                             |                             | 0.00                        |                             |                             |                             |                             |  |
| verdant $\rightarrow \vec{E}_7 \xrightarrow{W_V} \vec{V}_7$  |                             |                             |                             | 0.00                        |                             |                             |                             |                             |  |
| forest $\rightarrow \vec{E}_8 \xrightarrow{W_V} \vec{V}_8$   |                             |                             |                             | 0.00                        |                             |                             |                             |                             |  |

## 5.2 嵌入向量更新

- 根据注意力的大小，对值进行加权求和，得到嵌入向量的更新增量

| Value matrix                                                      | <div>a</div> <div>↓</div> <div><math>\vec{E}_1</math></div> | <div>fluffy</div> <div>↓</div> <div><math>\vec{E}_2</math></div> | <div>blue</div> <div>↓</div> <div><math>\vec{E}_3</math></div> | <div>creature</div> <div>↓</div> <div><math>\vec{E}_4</math></div> | <div>roamed</div> <div>↓</div> <div><math>\vec{E}_5</math></div> | <div>the</div> <div>↓</div> <div><math>\vec{E}_6</math></div> | <div>verdant</div> <div>↓</div> <div><math>\vec{E}_7</math></div> | <div>forest</div> <div>↓</div> <div><math>\vec{E}_8</math></div> |  |
|-------------------------------------------------------------------|-------------------------------------------------------------|------------------------------------------------------------------|----------------------------------------------------------------|--------------------------------------------------------------------|------------------------------------------------------------------|---------------------------------------------------------------|-------------------------------------------------------------------|------------------------------------------------------------------|--|
| <div>a</div> → $\vec{E}_1$ $\xrightarrow{w_v}$ $\vec{v}_1$        |                                                             |                                                                  |                                                                | 0.00 $\vec{v}_1$                                                   |                                                                  |                                                               |                                                                   |                                                                  |  |
| <div>fluffy</div> → $\vec{E}_2$ $\xrightarrow{w_v}$ $\vec{v}_2$   |                                                             |                                                                  |                                                                | 0.42 $\vec{v}_2$                                                   |                                                                  |                                                               |                                                                   |                                                                  |  |
| <div>blue</div> → $\vec{E}_3$ $\xrightarrow{w_v}$ $\vec{v}_3$     |                                                             |                                                                  |                                                                | 0.58 $\vec{v}_3$                                                   |                                                                  |                                                               |                                                                   |                                                                  |  |
| <div>creature</div> → $\vec{E}_4$ $\xrightarrow{w_v}$ $\vec{v}_4$ |                                                             |                                                                  |                                                                | 0.00 $\vec{v}_4$                                                   |                                                                  |                                                               |                                                                   |                                                                  |  |
| <div>roamed</div> → $\vec{E}_5$ $\xrightarrow{w_v}$ $\vec{v}_5$   |                                                             |                                                                  |                                                                | 0.00 $\vec{v}_5$                                                   |                                                                  |                                                               |                                                                   |                                                                  |  |
| <div>the</div> → $\vec{E}_6$ $\xrightarrow{w_v}$ $\vec{v}_6$      |                                                             |                                                                  |                                                                | 0.00 $\vec{v}_6$                                                   |                                                                  |                                                               |                                                                   |                                                                  |  |
| <div>verdant</div> → $\vec{E}_7$ $\xrightarrow{w_v}$ $\vec{v}_7$  |                                                             |                                                                  |                                                                | 0.00 $\vec{v}_7$                                                   |                                                                  |                                                               |                                                                   |                                                                  |  |
| <div>forest</div> → $\vec{E}_8$ $\xrightarrow{w_v}$ $\vec{v}_8$   |                                                             |                                                                  |                                                                | 0.00 $\vec{v}_8$                                                   |                                                                  |                                                               |                                                                   |                                                                  |  |

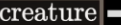
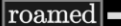
## 5.2 嵌入向量更新

- 根据注意力的大小，对值进行加权求和，得到嵌入向量的更新增量

| Value matrix<br>$W_V$                                             | <div>a</div> <div>↓</div> <div><math>\vec{E}_1</math></div> | <div>fluffy</div> <div>↓</div> <div><math>\vec{E}_2</math></div> | <div>blue</div> <div>↓</div> <div><math>\vec{E}_3</math></div> | <div>creature</div> <div>↓</div> <div><math>\vec{E}_4</math></div> | <div>roamed</div> <div>↓</div> <div><math>\vec{E}_5</math></div> | <div>the</div> <div>↓</div> <div><math>\vec{E}_6</math></div> | <div>verdant</div> <div>↓</div> <div><math>\vec{E}_7</math></div> | <div>forest</div> <div>↓</div> <div><math>\vec{E}_8</math></div> |  |
|-------------------------------------------------------------------|-------------------------------------------------------------|------------------------------------------------------------------|----------------------------------------------------------------|--------------------------------------------------------------------|------------------------------------------------------------------|---------------------------------------------------------------|-------------------------------------------------------------------|------------------------------------------------------------------|--|
| <div>a</div> → $\vec{E}_1$ $\xrightarrow{W_V}$ $\vec{V}_1$        |                                                             |                                                                  |                                                                | <div>0.00</div> $\vec{V}_1$                                        |                                                                  |                                                               |                                                                   |                                                                  |  |
| <div>fluffy</div> → $\vec{E}_2$ $\xrightarrow{W_V}$ $\vec{V}_2$   |                                                             |                                                                  |                                                                | 0.42 $\vec{V}_2$                                                   |                                                                  |                                                               |                                                                   |                                                                  |  |
| <div>blue</div> → $\vec{E}_3$ $\xrightarrow{W_V}$ $\vec{V}_3$     |                                                             |                                                                  |                                                                | 0.58 $\vec{V}_3$                                                   |                                                                  |                                                               |                                                                   |                                                                  |  |
| <div>creature</div> → $\vec{E}_4$ $\xrightarrow{W_V}$ $\vec{V}_4$ |                                                             |                                                                  |                                                                | <div>0.00</div> $\vec{V}_4$                                        |                                                                  |                                                               |                                                                   |                                                                  |  |
| <div>roamed</div> → $\vec{E}_5$ $\xrightarrow{W_V}$ $\vec{V}_5$   |                                                             |                                                                  |                                                                | <div>0.00</div> $\vec{V}_5$                                        |                                                                  |                                                               |                                                                   |                                                                  |  |
| <div>the</div> → $\vec{E}_6$ $\xrightarrow{W_V}$ $\vec{V}_6$      |                                                             |                                                                  |                                                                | <div>0.00</div> $\vec{V}_6$                                        |                                                                  |                                                               |                                                                   |                                                                  |  |
| <div>verdant</div> → $\vec{E}_7$ $\xrightarrow{W_V}$ $\vec{V}_7$  |                                                             |                                                                  |                                                                | <div>0.00</div> $\vec{V}_7$                                        |                                                                  |                                                               |                                                                   |                                                                  |  |
| <div>forest</div> → $\vec{E}_8$ $\xrightarrow{W_V}$ $\vec{V}_8$   |                                                             |                                                                  |                                                                | <div>0.00</div> $\vec{V}_8$                                        |                                                                  |                                                               |                                                                   |                                                                  |  |

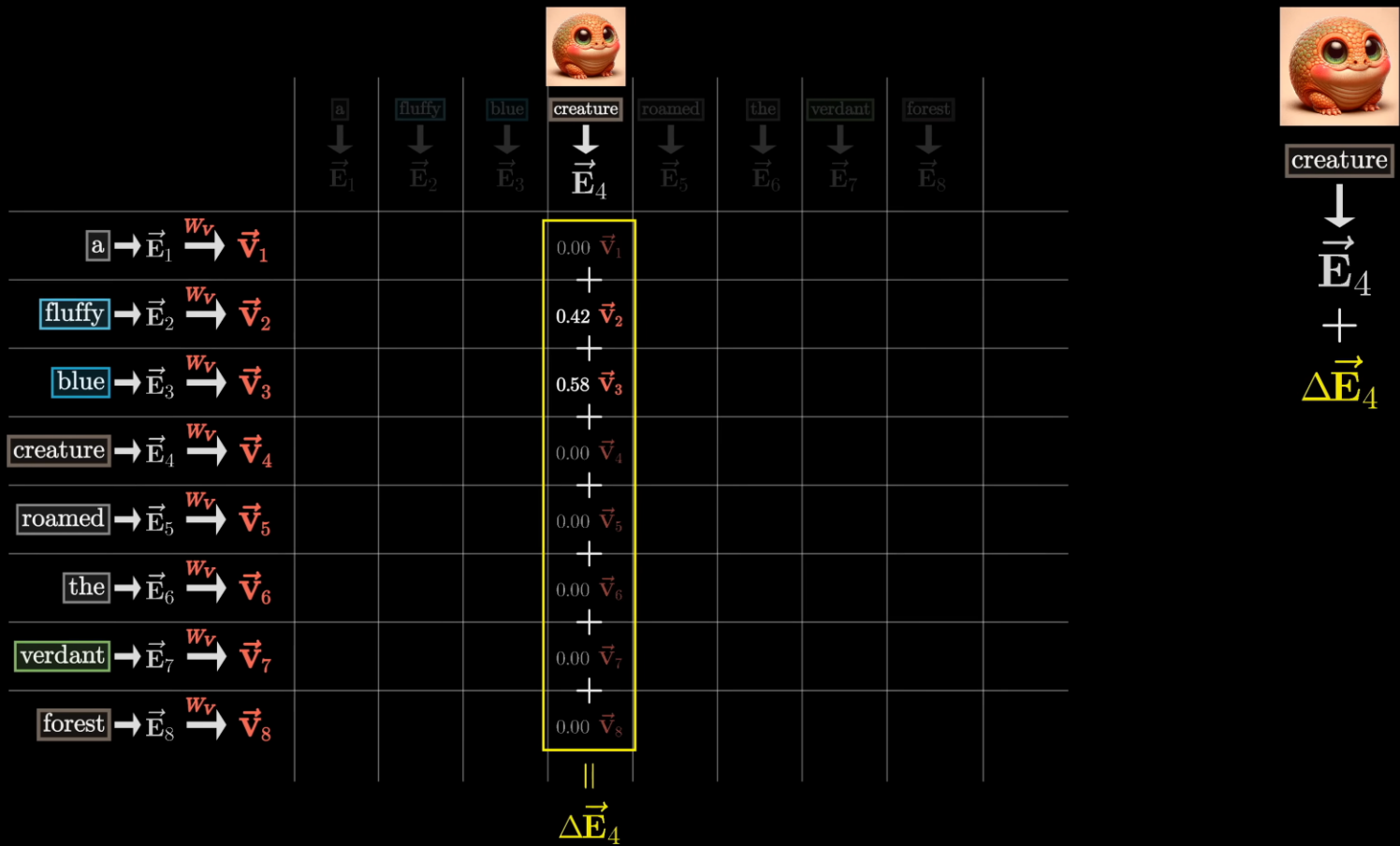
## 5.2 嵌入向量更新

- 根据注意力的大小，对值进行加权求和，得到嵌入向量的更新增量 $\Delta E$

|                                                                                      |                                                                                     |                                                                                     |                                                                                     |                                                                                                                                                                                                                                                                                                                  |                                                                                     |                                                                                     |                                                                                     |                                                                                     |
|--------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
|                                                                                      |  |  |  |                                                                                                                                                                                                                               |  |  |  |  |
|                                                                                      | $\downarrow$<br>$\vec{E}_1$                                                         | $\downarrow$<br>$\vec{E}_2$                                                         | $\downarrow$<br>$\vec{E}_3$                                                         | $\downarrow$<br>$\vec{E}_4$                                                                                                                                                                                                                                                                                      | $\downarrow$<br>$\vec{E}_5$                                                         | $\downarrow$<br>$\vec{E}_6$                                                         | $\downarrow$<br>$\vec{E}_7$                                                         | $\downarrow$<br>$\vec{E}_8$                                                         |
|    | $\rightarrow \vec{E}_1$                                                             |                                                                                     |                                                                                     | $\xrightarrow{w_v} \vec{v}_1$                                                                                                                                                                                                                                                                                    |                                                                                     |                                                                                     |                                                                                     |                                                                                     |
|    | $\rightarrow \vec{E}_2$                                                             |                                                                                     |                                                                                     | $\xrightarrow{w_v} \vec{v}_2$                                                                                                                                                                                                                                                                                    |                                                                                     |                                                                                     |                                                                                     |                                                                                     |
|    | $\rightarrow \vec{E}_3$                                                             |                                                                                     |                                                                                     | $\xrightarrow{w_v} \vec{v}_3$                                                                                                                                                                                                                                                                                    |                                                                                     |                                                                                     |                                                                                     |                                                                                     |
|    | $\rightarrow \vec{E}_4$                                                             |                                                                                     |                                                                                     | $\xrightarrow{w_v} \vec{v}_4$                                                                                                                                                                                                                                                                                    |                                                                                     |                                                                                     |                                                                                     |                                                                                     |
|    | $\rightarrow \vec{E}_5$                                                             |                                                                                     |                                                                                     | $\xrightarrow{w_v} \vec{v}_5$                                                                                                                                                                                                                                                                                    |                                                                                     |                                                                                     |                                                                                     |                                                                                     |
|    | $\rightarrow \vec{E}_6$                                                             |                                                                                     |                                                                                     | $\xrightarrow{w_v} \vec{v}_6$                                                                                                                                                                                                                                                                                    |                                                                                     |                                                                                     |                                                                                     |                                                                                     |
|    | $\rightarrow \vec{E}_7$                                                             |                                                                                     |                                                                                     | $\xrightarrow{w_v} \vec{v}_7$                                                                                                                                                                                                                                                                                    |                                                                                     |                                                                                     |                                                                                     |                                                                                     |
|  | $\rightarrow \vec{E}_8$                                                             |                                                                                     |                                                                                     | $\xrightarrow{w_v} \vec{v}_8$                                                                                                                                                                                                                                                                                    |                                                                                     |                                                                                     |                                                                                     |                                                                                     |
|                                                                                      |                                                                                     |                                                                                     |                                                                                     | <div>0.00 <math>\vec{v}_1</math><br/>+<br/>0.42 <math>\vec{v}_2</math><br/>+<br/>0.58 <math>\vec{v}_3</math><br/>+<br/>0.00 <math>\vec{v}_4</math><br/>+<br/>0.00 <math>\vec{v}_5</math><br/>+<br/>0.00 <math>\vec{v}_6</math><br/>+<br/>0.00 <math>\vec{v}_7</math><br/>+<br/>0.00 <math>\vec{v}_8</math></div> |                                                                                     |                                                                                     |                                                                                     |                                                                                     |
|                                                                                      |                                                                                     |                                                                                     |                                                                                     | $\parallel$<br>$\Delta \vec{E}_4$                                                                                                                                                                                                                                                                                |                                                                                     |                                                                                     |                                                                                     |                                                                                     |

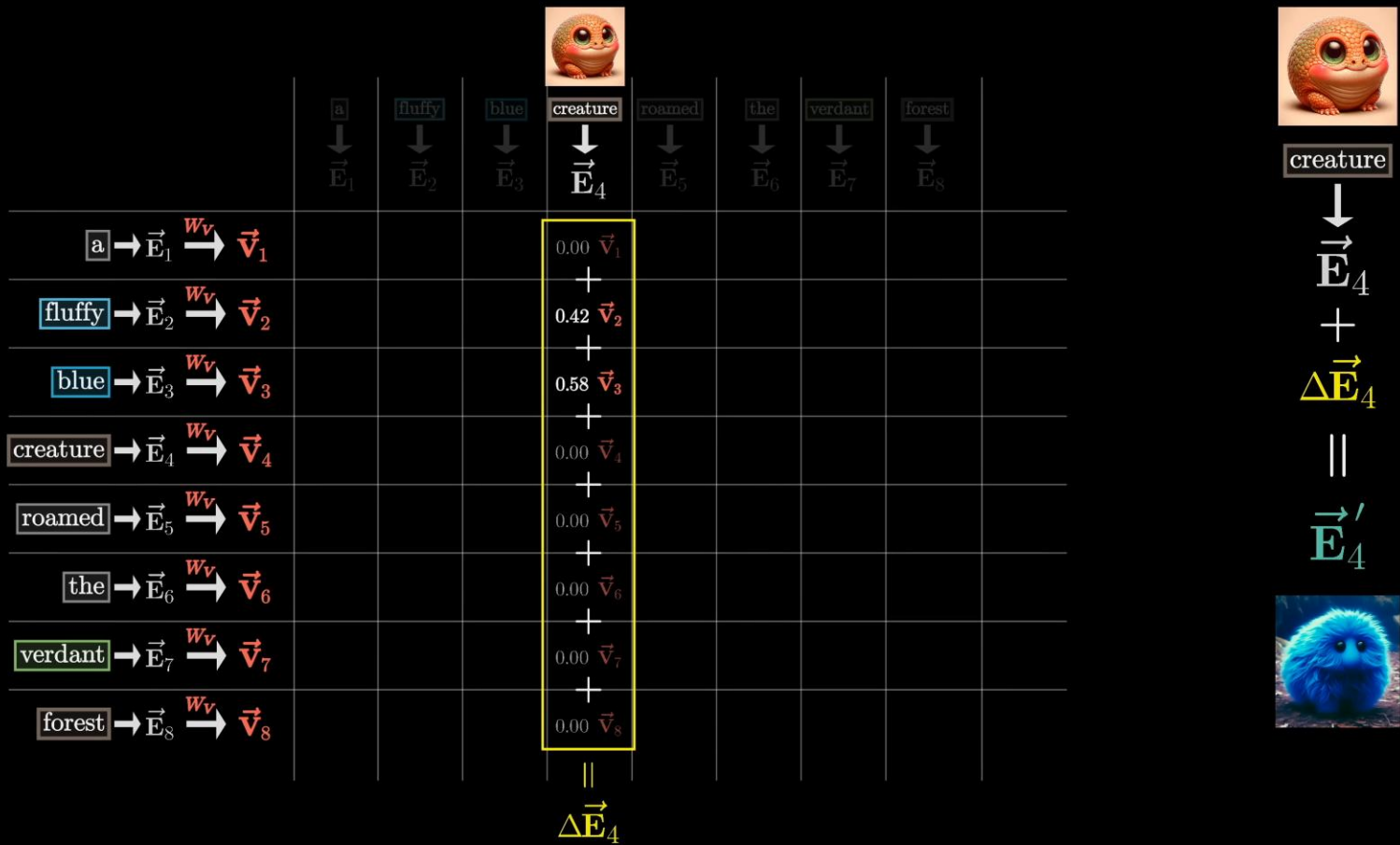
# 5.2 嵌入向量更新

- 更新嵌入向量  $(E'_i) = \text{嵌入向量 } (E_i) + \text{更新增量 } (\Delta E_i)$



# 5.2 嵌入向量更新

- 更新嵌入向量 ( $E'_i$ ) = 嵌入向量 ( $E_i$ ) + 更新增量 ( $\Delta E_i$ )





# 5.2 嵌入向量更新

- 对每个词元都进行这样的更新

|                                                                   | <div>a</div> <div>↓</div> <div><math>\vec{E}_1</math></div> | <div>fluffy</div> <div>↓</div> <div><math>\vec{E}_2</math></div> | <div>blue</div> <div>↓</div> <div><math>\vec{E}_3</math></div> | <div>creature</div> <div>↓</div> <div><math>\vec{E}_4</math></div> | <div>roamed</div> <div>↓</div> <div><math>\vec{E}_5</math></div> | <div>the</div> <div>↓</div> <div><math>\vec{E}_6</math></div> | <div>verdant</div> <div>↓</div> <div><math>\vec{E}_7</math></div> | <div>forest</div> <div>↓</div> <div><math>\vec{E}_8</math></div> |  |
|-------------------------------------------------------------------|-------------------------------------------------------------|------------------------------------------------------------------|----------------------------------------------------------------|--------------------------------------------------------------------|------------------------------------------------------------------|---------------------------------------------------------------|-------------------------------------------------------------------|------------------------------------------------------------------|--|
| <div>a</div> → $\vec{E}_1$ $\xrightarrow{w_v}$ $\vec{v}_1$        | 1.00 $\vec{v}_1$                                            | 0.00 $\vec{v}_1$                                                 | 0.00 $\vec{v}_1$                                               | 0.00 $\vec{v}_1$                                                   | 0.00 $\vec{v}_1$                                                 | 0.00 $\vec{v}_1$                                              | 0.00 $\vec{v}_1$                                                  | 0.00 $\vec{v}_1$                                                 |  |
| <div>fluffy</div> → $\vec{E}_2$ $\xrightarrow{w_v}$ $\vec{v}_2$   | 0.00 $\vec{v}_2$                                            | 1.00 $\vec{v}_2$                                                 | 0.00 $\vec{v}_2$                                               | 0.42 $\vec{v}_2$                                                   | 0.00 $\vec{v}_2$                                                 | 0.00 $\vec{v}_2$                                              | 0.00 $\vec{v}_2$                                                  | 0.00 $\vec{v}_2$                                                 |  |
| <div>blue</div> → $\vec{E}_3$ $\xrightarrow{w_v}$ $\vec{v}_3$     | 0.00 $\vec{v}_3$                                            | 0.00 $\vec{v}_3$                                                 | 1.00 $\vec{v}_3$                                               | 0.58 $\vec{v}_3$                                                   | 0.00 $\vec{v}_3$                                                 | 0.00 $\vec{v}_3$                                              | 0.00 $\vec{v}_3$                                                  | 0.00 $\vec{v}_3$                                                 |  |
| <div>creature</div> → $\vec{E}_4$ $\xrightarrow{w_v}$ $\vec{v}_4$ | 0.00 $\vec{v}_4$                                            | 0.00 $\vec{v}_4$                                                 | 0.00 $\vec{v}_4$                                               | 0.00 $\vec{v}_4$                                                   | 0.00 $\vec{v}_4$                                                 | 0.00 $\vec{v}_4$                                              | 0.00 $\vec{v}_4$                                                  | 0.00 $\vec{v}_4$                                                 |  |
| <div>roamed</div> → $\vec{E}_5$ $\xrightarrow{w_v}$ $\vec{v}_5$   | 0.00 $\vec{v}_5$                                            | 0.00 $\vec{v}_5$                                                 | 0.00 $\vec{v}_5$                                               | 0.00 $\vec{v}_5$                                                   | 0.01 $\vec{v}_5$                                                 | 0.00 $\vec{v}_5$                                              | 0.00 $\vec{v}_5$                                                  | 0.00 $\vec{v}_5$                                                 |  |
| <div>the</div> → $\vec{E}_6$ $\xrightarrow{w_v}$ $\vec{v}_6$      | 0.00 $\vec{v}_6$                                            | 0.00 $\vec{v}_6$                                                 | 0.00 $\vec{v}_6$                                               | 0.00 $\vec{v}_6$                                                   | 0.99 $\vec{v}_6$                                                 | 1.00 $\vec{v}_6$                                              | 0.00 $\vec{v}_6$                                                  | 0.00 $\vec{v}_6$                                                 |  |
| <div>verdant</div> → $\vec{E}_7$ $\xrightarrow{w_v}$ $\vec{v}_7$  | 0.00 $\vec{v}_7$                                            | 0.00 $\vec{v}_7$                                                 | 0.00 $\vec{v}_7$                                               | 0.00 $\vec{v}_7$                                                   | 0.00 $\vec{v}_7$                                                 | 0.00 $\vec{v}_7$                                              | 1.00 $\vec{v}_7$                                                  | 1.00 $\vec{v}_7$                                                 |  |
| <div>forest</div> → $\vec{E}_8$ $\xrightarrow{w_v}$ $\vec{v}_8$   | 0.00 $\vec{v}_8$                                            | 0.00 $\vec{v}_8$                                                 | 0.00 $\vec{v}_8$                                               | 0.00 $\vec{v}_8$                                                   | 0.00 $\vec{v}_8$                                                 | 0.00 $\vec{v}_8$                                              | 0.00 $\vec{v}_8$                                                  | 0.00 $\vec{v}_8$                                                 |  |
|                                                                   |                                                             |                                                                  |                                                                |                                                                    |                                                                  |                                                               |                                                                   |                                                                  |  |
|                                                                   | $\Delta \vec{E}_1$                                          | $\Delta \vec{E}_2$                                               | $\Delta \vec{E}_3$                                             | $\Delta \vec{E}_4$                                                 | $\Delta \vec{E}_5$                                               | $\Delta \vec{E}_6$                                            | $\Delta \vec{E}_7$                                                | $\Delta \vec{E}_8$                                               |  |



# 5.2 嵌入向量更新

- 对每个词元都进行这样的更新

|                                         | <div>a</div> <div>↓</div> <div><math>\vec{E}_1</math></div> | <div>fluffy</div> <div>↓</div> <div><math>\vec{E}_2</math></div> | <div>blue</div> <div>↓</div> <div><math>\vec{E}_3</math></div> | <div>creature</div> <div>↓</div> <div><math>\vec{E}_4</math></div> | <div>roamed</div> <div>↓</div> <div><math>\vec{E}_5</math></div> | <div>the</div> <div>↓</div> <div><math>\vec{E}_6</math></div> | <div>verdant</div> <div>↓</div> <div><math>\vec{E}_7</math></div> | <div>forest</div> <div>↓</div> <div><math>\vec{E}_8</math></div> |
|-----------------------------------------|-------------------------------------------------------------|------------------------------------------------------------------|----------------------------------------------------------------|--------------------------------------------------------------------|------------------------------------------------------------------|---------------------------------------------------------------|-------------------------------------------------------------------|------------------------------------------------------------------|
| $\vec{E}_1 \xrightarrow{w_v} \vec{v}_1$ | 1.00 $\vec{v}_1$                                            | 0.00 $\vec{v}_1$                                                 | 0.00 $\vec{v}_1$                                               | 0.00 $\vec{v}_1$                                                   | 0.00 $\vec{v}_1$                                                 | 0.00 $\vec{v}_1$                                              | 0.00 $\vec{v}_1$                                                  | 0.00 $\vec{v}_1$                                                 |
| $\vec{E}_2 \xrightarrow{w_v} \vec{v}_2$ | 0.00 $\vec{v}_2$                                            | 1.00 $\vec{v}_2$                                                 | 0.00 $\vec{v}_2$                                               | 0.42 $\vec{v}_2$                                                   | 0.00 $\vec{v}_2$                                                 | 0.00 $\vec{v}_2$                                              | 0.00 $\vec{v}_2$                                                  | 0.00 $\vec{v}_2$                                                 |
| $\vec{E}_3 \xrightarrow{w_v} \vec{v}_3$ | 0.00 $\vec{v}_3$                                            | 0.00 $\vec{v}_3$                                                 | 1.00 $\vec{v}_3$                                               | 0.58 $\vec{v}_3$                                                   | 0.00 $\vec{v}_3$                                                 | 0.00 $\vec{v}_3$                                              | 0.00 $\vec{v}_3$                                                  | 0.00 $\vec{v}_3$                                                 |
| $\vec{E}_4 \xrightarrow{w_v} \vec{v}_4$ | 0.00 $\vec{v}_4$                                            | 0.00 $\vec{v}_4$                                                 | 0.00 $\vec{v}_4$                                               | 0.00 $\vec{v}_4$                                                   | 0.00 $\vec{v}_4$                                                 | 0.00 $\vec{v}_4$                                              | 0.00 $\vec{v}_4$                                                  | 0.00 $\vec{v}_4$                                                 |
| $\vec{E}_5 \xrightarrow{w_v} \vec{v}_5$ | 0.00 $\vec{v}_5$                                            | 0.00 $\vec{v}_5$                                                 | 0.00 $\vec{v}_5$                                               | 0.00 $\vec{v}_5$                                                   | 0.01 $\vec{v}_5$                                                 | 0.00 $\vec{v}_5$                                              | 0.00 $\vec{v}_5$                                                  | 0.00 $\vec{v}_5$                                                 |
| $\vec{E}_6 \xrightarrow{w_v} \vec{v}_6$ | 0.00 $\vec{v}_6$                                            | 0.00 $\vec{v}_6$                                                 | 0.00 $\vec{v}_6$                                               | 0.00 $\vec{v}_6$                                                   | 0.99 $\vec{v}_6$                                                 | 1.00 $\vec{v}_6$                                              | 0.00 $\vec{v}_6$                                                  | 0.00 $\vec{v}_6$                                                 |
| $\vec{E}_7 \xrightarrow{w_v} \vec{v}_7$ | 0.00 $\vec{v}_7$                                            | 0.00 $\vec{v}_7$                                                 | 0.00 $\vec{v}_7$                                               | 0.00 $\vec{v}_7$                                                   | 0.00 $\vec{v}_7$                                                 | 0.00 $\vec{v}_7$                                              | 1.00 $\vec{v}_7$                                                  | 1.00 $\vec{v}_7$                                                 |
| $\vec{E}_8 \xrightarrow{w_v} \vec{v}_8$ | 0.00 $\vec{v}_8$                                            | 0.00 $\vec{v}_8$                                                 | 0.00 $\vec{v}_8$                                               | 0.00 $\vec{v}_8$                                                   | 0.00 $\vec{v}_8$                                                 | 0.00 $\vec{v}_8$                                              | 0.00 $\vec{v}_8$                                                  | 0.00 $\vec{v}_8$                                                 |
|                                         | <br>$\Delta \vec{E}_1$                                      | <br>$\Delta \vec{E}_2$                                           | <br>$\Delta \vec{E}_3$                                         | <br>$\Delta \vec{E}_4$                                             | <br>$\Delta \vec{E}_5$                                           | <br>$\Delta \vec{E}_6$                                        | <br>$\Delta \vec{E}_7$                                            | <br>$\Delta \vec{E}_8$                                           |

$\vec{E}_1$   
+  
 $\Delta \vec{E}_1$   
||  
 $\vec{E}'_1$

$\vec{E}_2$   
+  
 $\Delta \vec{E}_2$   
||  
 $\vec{E}'_2$

$\vec{E}_3$   
+  
 $\Delta \vec{E}_3$   
||  
 $\vec{E}'_3$

$\vec{E}_4$   
+  
 $\Delta \vec{E}_4$   
||  
 $\vec{E}'_4$

$\vec{E}_5$   
+  
 $\Delta \vec{E}_5$   
||  
 $\vec{E}'_5$

$\vec{E}_6$   
+  
 $\Delta \vec{E}_6$   
||  
 $\vec{E}'_6$

$\vec{E}_7$   
+  
 $\Delta \vec{E}_7$   
||  
 $\vec{E}'_7$

$\vec{E}_8$   
+  
 $\Delta \vec{E}_8$   
||  
 $\vec{E}'_8$

# 5.2 嵌入向量更新

- 由此，通过注意力模块，将得到一系列更精确的嵌入向量

|                                         | <div>a</div> <div>↓</div> <div><math>\vec{E}_1</math></div> | <div>fluffy</div> <div>↓</div> <div><math>\vec{E}_2</math></div> | <div>blue</div> <div>↓</div> <div><math>\vec{E}_3</math></div> | <div>creature</div> <div>↓</div> <div><math>\vec{E}_4</math></div> | <div>roamed</div> <div>↓</div> <div><math>\vec{E}_5</math></div> | <div>the</div> <div>↓</div> <div><math>\vec{E}_6</math></div> | <div>verdant</div> <div>↓</div> <div><math>\vec{E}_7</math></div> | <div>forest</div> <div>↓</div> <div><math>\vec{E}_8</math></div> |
|-----------------------------------------|-------------------------------------------------------------|------------------------------------------------------------------|----------------------------------------------------------------|--------------------------------------------------------------------|------------------------------------------------------------------|---------------------------------------------------------------|-------------------------------------------------------------------|------------------------------------------------------------------|
| $\vec{E}_1 \xrightarrow{w_v} \vec{v}_1$ | 1.00 $\vec{v}_1$                                            | 0.00 $\vec{v}_1$                                                 | 0.00 $\vec{v}_1$                                               | 0.00 $\vec{v}_1$                                                   | 0.00 $\vec{v}_1$                                                 | 0.00 $\vec{v}_1$                                              | 0.00 $\vec{v}_1$                                                  | 0.00 $\vec{v}_1$                                                 |
| $\vec{E}_2 \xrightarrow{w_v} \vec{v}_2$ | 0.00 $\vec{v}_2$                                            | 1.00 $\vec{v}_2$                                                 | 0.00 $\vec{v}_2$                                               | 0.42 $\vec{v}_2$                                                   | 0.00 $\vec{v}_2$                                                 | 0.00 $\vec{v}_2$                                              | 0.00 $\vec{v}_2$                                                  | 0.00 $\vec{v}_2$                                                 |
| $\vec{E}_3 \xrightarrow{w_v} \vec{v}_3$ | 0.00 $\vec{v}_3$                                            | 0.00 $\vec{v}_3$                                                 | 1.00 $\vec{v}_3$                                               | 0.58 $\vec{v}_3$                                                   | 0.00 $\vec{v}_3$                                                 | 0.00 $\vec{v}_3$                                              | 0.00 $\vec{v}_3$                                                  | 0.00 $\vec{v}_3$                                                 |
| $\vec{E}_4 \xrightarrow{w_v} \vec{v}_4$ | 0.00 $\vec{v}_4$                                            | 0.00 $\vec{v}_4$                                                 | 0.00 $\vec{v}_4$                                               | 0.00 $\vec{v}_4$                                                   | 0.00 $\vec{v}_4$                                                 | 0.00 $\vec{v}_4$                                              | 0.00 $\vec{v}_4$                                                  | 0.00 $\vec{v}_4$                                                 |
| $\vec{E}_5 \xrightarrow{w_v} \vec{v}_5$ | 0.00 $\vec{v}_5$                                            | 0.00 $\vec{v}_5$                                                 | 0.00 $\vec{v}_5$                                               | 0.00 $\vec{v}_5$                                                   | 0.01 $\vec{v}_5$                                                 | 0.00 $\vec{v}_5$                                              | 0.00 $\vec{v}_5$                                                  | 0.00 $\vec{v}_5$                                                 |
| $\vec{E}_6 \xrightarrow{w_v} \vec{v}_6$ | 0.00 $\vec{v}_6$                                            | 0.00 $\vec{v}_6$                                                 | 0.00 $\vec{v}_6$                                               | 0.00 $\vec{v}_6$                                                   | 0.99 $\vec{v}_6$                                                 | 1.00 $\vec{v}_6$                                              | 0.00 $\vec{v}_6$                                                  | 0.00 $\vec{v}_6$                                                 |
| $\vec{E}_7 \xrightarrow{w_v} \vec{v}_7$ | 0.00 $\vec{v}_7$                                            | 0.00 $\vec{v}_7$                                                 | 0.00 $\vec{v}_7$                                               | 0.00 $\vec{v}_7$                                                   | 0.00 $\vec{v}_7$                                                 | 0.00 $\vec{v}_7$                                              | 1.00 $\vec{v}_7$                                                  | 1.00 $\vec{v}_7$                                                 |
| $\vec{E}_8 \xrightarrow{w_v} \vec{v}_8$ | 0.00 $\vec{v}_8$                                            | 0.00 $\vec{v}_8$                                                 | 0.00 $\vec{v}_8$                                               | 0.00 $\vec{v}_8$                                                   | 0.00 $\vec{v}_8$                                                 | 0.00 $\vec{v}_8$                                              | 0.00 $\vec{v}_8$                                                  | 0.00 $\vec{v}_8$                                                 |
|                                         | $\Delta \vec{E}_1$                                          | $\Delta \vec{E}_2$                                               | $\Delta \vec{E}_3$                                             | $\Delta \vec{E}_4$                                                 | $\Delta \vec{E}_5$                                               | $\Delta \vec{E}_6$                                            | $\Delta \vec{E}_7$                                                | $\Delta \vec{E}_8$                                               |

$\vec{E}_1$  $\vec{E}_2$  $\vec{E}_3$  $\vec{E}_4$  $\vec{E}_5$  $\vec{E}_6$  $\vec{E}_7$  $\vec{E}_8$

$+$  $+$  $+$  $+$  $+$  $+$  $+$  $+$

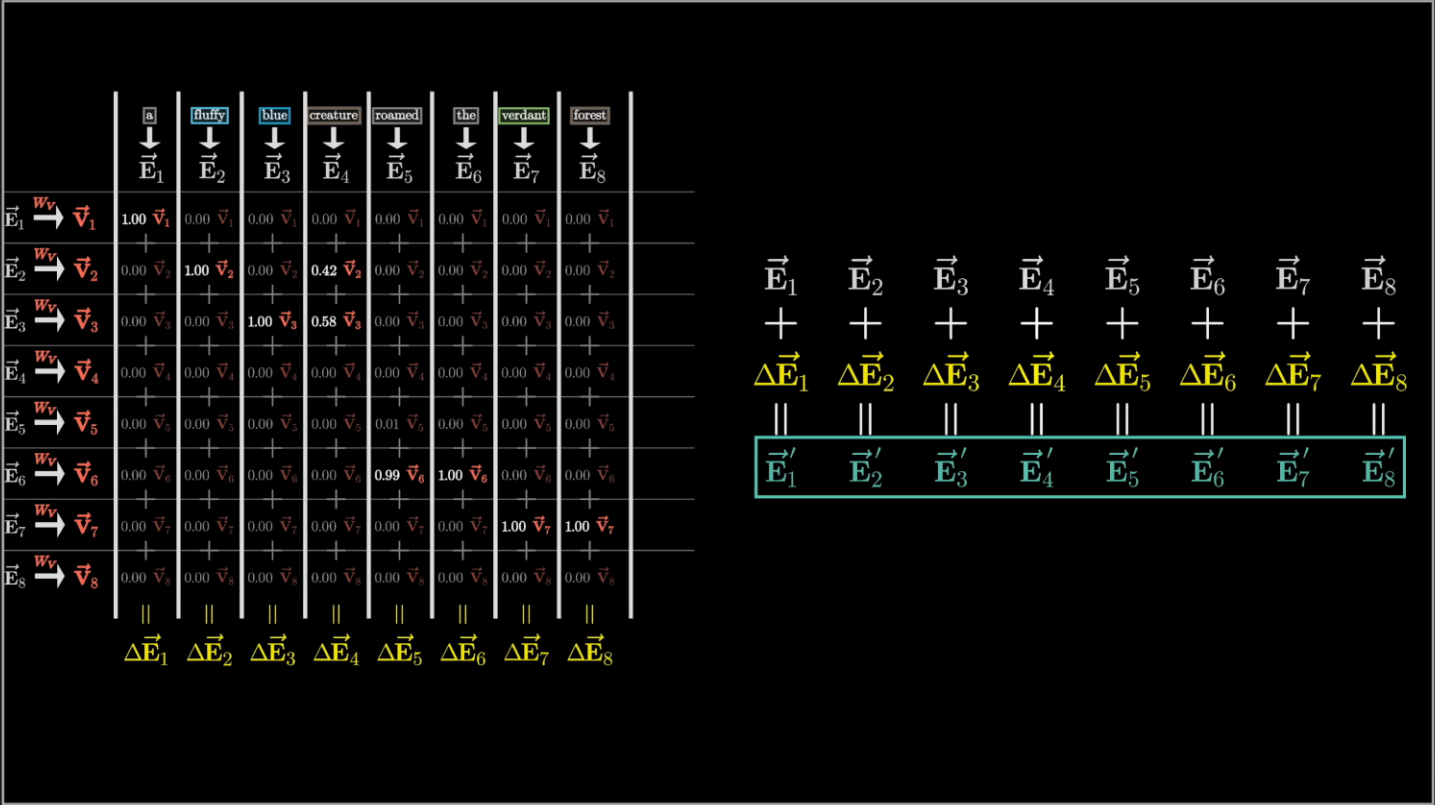
$\Delta \vec{E}_1$  $\Delta \vec{E}_2$  $\Delta \vec{E}_3$  $\Delta \vec{E}_4$  $\Delta \vec{E}_5$  $\Delta \vec{E}_6$  $\Delta \vec{E}_7$  $\Delta \vec{E}_8$

$\vec{E}'_1$  $\vec{E}'_2$  $\vec{E}'_3$  $\vec{E}'_4$  $\vec{E}'_5$  $\vec{E}'_6$  $\vec{E}'_7$  $\vec{E}'_8$

# 5.2 嵌入向量更新

- 以上为注意力模块的“一个头”

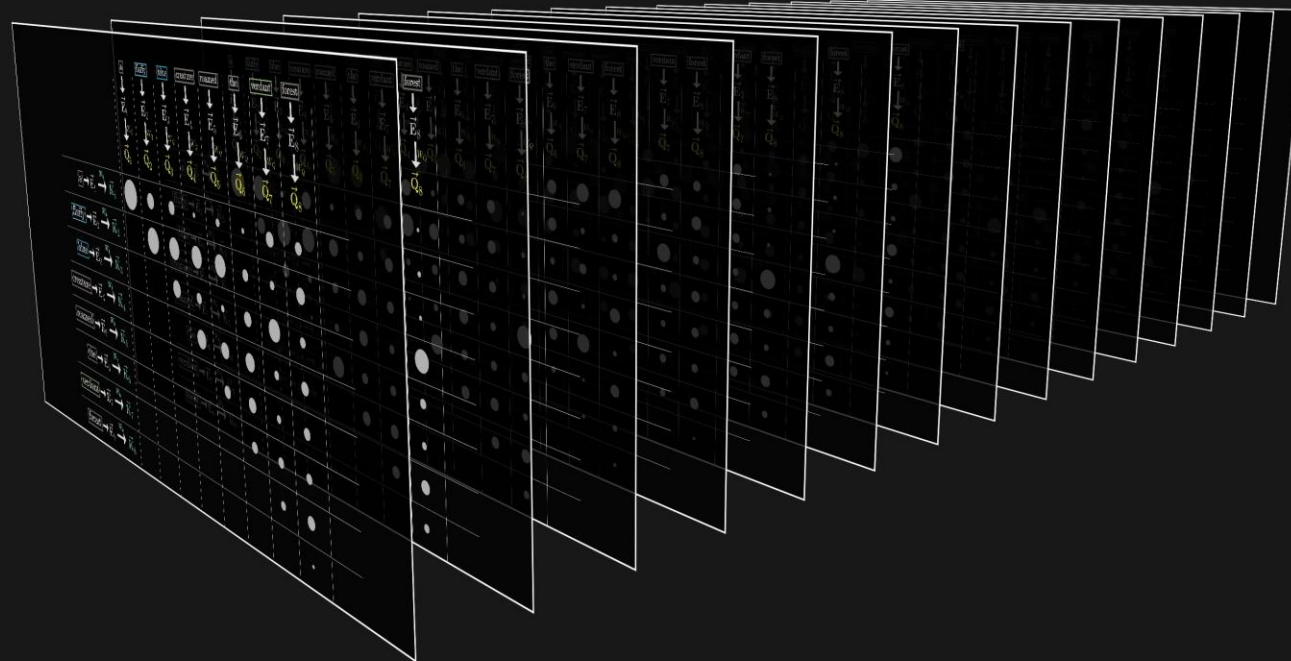
## One head of attention



## 5.2 嵌入向量更新

- 将上述程序并列重复多次，构成“多头注意力”（详见第7节）

### Multi-headed attention



1. 回顾
2. 启发示例：注意力的作用
3. 注意力机制：查询（Query）、键（Key）
4. 掩码（Masking）
5. 注意力机制：值（Value）
- 6. 参数数量**
7. 多头注意力
8. Transformer的深度

## 6.1 Q、K和V

- 参数：Q、K和V矩阵

Query

|      |      |      |      |      |      |      |      |     |      |
|------|------|------|------|------|------|------|------|-----|------|
| -3.7 | +3.9 | -2.4 | -6.3 | -9.4 | -8.6 | +3.6 | -0.9 | ... | +0.7 |
| +7.9 | +9.7 | -5.6 | +3.2 | -4.7 | -9.5 | +5.1 | -3.6 | ... | -2.3 |
| +1.7 | +6.6 | +2.6 | +7.4 | -4.5 | +5.9 | -6.2 | +9.0 | ... | +3.7 |
| ⋮    | ⋮    | ⋮    | ⋮    | ⋮    | ⋮    | ⋮    | ⋮    | ⋱   | ⋮    |
| -5.6 | +8.9 | +4.6 | -4.9 | -5.7 | +0.4 | -9.4 | -5.8 | ... | -1.5 |

Key

|      |      |      |      |      |      |      |      |     |      |
|------|------|------|------|------|------|------|------|-----|------|
| -2.5 | -0.7 | -4.4 | +1.7 | +7.2 | -7.6 | +0.3 | -7.3 | ... | +4.3 |
| -2.1 | +1.3 | -6.3 | -7.0 | -0.2 | -2.9 | +8.7 | +5.3 | ... | +4.9 |
| +8.0 | -8.2 | +1.0 | +1.7 | +9.1 | -4.1 | -5.1 | -7.9 | ... | -9.6 |
| ⋮    | ⋮    | ⋮    | ⋮    | ⋮    | ⋮    | ⋮    | ⋮    | ⋱   | ⋮    |
| +8.5 | +3.4 | +5.6 | -4.3 | +1.7 | -8.6 | -0.3 | +9.5 | ... | +7.5 |

How many  
parameters?

Value

|      |      |      |      |      |      |      |      |     |      |
|------|------|------|------|------|------|------|------|-----|------|
| -3.2 | +9.1 | -5.3 | +8.9 | +8.7 | +5.9 | +2.6 | +7.4 | ... | -4.1 |
| +6.9 | +2.3 | -9.6 | -3.0 | -7.0 | +9.5 | -0.4 | -0.1 | ... | +2.8 |
| -2.6 | -7.2 | +6.4 | -6.1 | +0.2 | -5.5 | -8.0 | +7.2 | ... | +9.4 |
| +9.1 | +8.0 | +5.4 | -3.3 | -8.3 | -1.8 | -5.3 | -7.3 | ... | -8.8 |
| +4.5 | -9.7 | +5.4 | -7.0 | -8.3 | -8.1 | +3.4 | -5.0 | ... | -1.6 |
| +1.1 | +7.1 | +4.5 | -4.5 | -7.3 | -8.8 | -3.9 | -4.7 | ... | -0.9 |
| +3.6 | +3.9 | -4.3 | -2.4 | -6.3 | +5.7 | -8.8 | +3.9 | ... | +5.5 |
| +5.5 | -4.8 | -2.5 | +1.7 | -4.5 | -2.6 | -6.0 | -0.8 | ... | -9.0 |
| ⋮    | ⋮    | ⋮    | ⋮    | ⋮    | ⋮    | ⋮    | ⋮    | ⋱   | ⋮    |
| +5.9 | -8.4 | +0.4 | -3.8 | +1.5 | +9.1 | +2.9 | -9.2 | ... | -1.4 |



# 6.1 Q、K和V

- Q、K的参数数量

Query

$128 \times 12,288 = 1,572,864$

12,288

128 {

|      |      |      |      |      |      |      |      |     |      |
|------|------|------|------|------|------|------|------|-----|------|
| -3.7 | +3.9 | -2.4 | -6.3 | -9.4 | -8.6 | +3.6 | -0.9 | ... | +0.7 |
| +7.9 | +9.7 | -5.6 | +3.2 | -4.7 | -9.5 | +5.1 | -3.6 | ... | -2.3 |
| +1.7 | +6.6 | +2.6 | +7.4 | -4.5 | +5.9 | -6.2 | +9.0 | ... | +3.7 |
| ⋮    | ⋮    | ⋮    | ⋮    | ⋮    | ⋮    | ⋮    | ⋮    | ⋱   | ⋮    |
| -5.6 | +8.9 | +4.6 | -4.9 | -5.7 | +0.4 | -9.4 | -5.8 | ... | -1.5 |

Key

$128 \times 12,288 = 1,572,864$

12,288

128 {

|      |      |      |      |      |      |      |      |     |      |
|------|------|------|------|------|------|------|------|-----|------|
| -2.5 | -0.7 | -4.4 | +1.7 | +7.2 | -7.6 | +0.3 | -7.3 | ... | +4.3 |
| -2.1 | +1.3 | -6.3 | -7.0 | -0.2 | -2.9 | +8.7 | +5.3 | ... | +4.9 |
| +8.0 | -8.2 | +1.0 | +1.7 | +9.1 | -4.1 | -5.1 | -7.9 | ... | -9.6 |
| ⋮    | ⋮    | ⋮    | ⋮    | ⋮    | ⋮    | ⋮    | ⋮    | ⋱   | ⋮    |
| +8.5 | +3.4 | +5.6 | -4.3 | +1.7 | -8.6 | -0.3 | +9.5 | ... | +7.5 |

Value

|     |     |     |     |     |     |     |     |     |      |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|------|
| 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0  |
| 1.1 | 1.2 | 1.3 | 1.4 | 1.5 | 1.6 | 1.7 | 1.8 | 1.9 | 2.0  |
| 2.1 | 2.2 | 2.3 | 2.4 | 2.5 | 2.6 | 2.7 | 2.8 | 2.9 | 3.0  |
| 3.1 | 3.2 | 3.3 | 3.4 | 3.5 | 3.6 | 3.7 | 3.8 | 3.9 | 4.0  |
| 4.1 | 4.2 | 4.3 | 4.4 | 4.5 | 4.6 | 4.7 | 4.8 | 4.9 | 5.0  |
| 5.1 | 5.2 | 5.3 | 5.4 | 5.5 | 5.6 | 5.7 | 5.8 | 5.9 | 6.0  |
| 6.1 | 6.2 | 6.3 | 6.4 | 6.5 | 6.6 | 6.7 | 6.8 | 6.9 | 7.0  |
| 7.1 | 7.2 | 7.3 | 7.4 | 7.5 | 7.6 | 7.7 | 7.8 | 7.9 | 8.0  |
| 8.1 | 8.2 | 8.3 | 8.4 | 8.5 | 8.6 | 8.7 | 8.8 | 8.9 | 9.0  |
| 9.1 | 9.2 | 9.3 | 9.4 | 9.5 | 9.6 | 9.7 | 9.8 | 9.9 | 10.0 |



# 6.1 Q、K和V

- 理论上，V矩阵参数数量

Query  
1,572,864

|      |      |      |      |      |      |      |      |     |      |
|------|------|------|------|------|------|------|------|-----|------|
| -3.7 | +3.9 | -2.4 | -5.3 | -9.4 | -6.6 | +3.8 | -8.9 | ... | +0.1 |
| +7.9 | +0.7 | -6.8 | +5.3 | -4.7 | -8.5 | +5.1 | -3.6 | ... | -2.9 |
| +1.7 | +6.6 | +9.6 | +7.4 | -4.5 | +5.9 | -6.3 | +9.0 | ... | +3.7 |
| ⋮    | ⋮    | ⋮    | ⋮    | ⋮    | ⋮    | ⋮    | ⋮    | ⋱   | ⋮    |
| -5.0 | +6.9 | +4.6 | -4.9 | -5.7 | +0.4 | -9.4 | -5.8 | ... | -1.5 |

Key

1,572,864

|      |      |      |      |      |      |      |      |     |      |
|------|------|------|------|------|------|------|------|-----|------|
| -2.6 | -0.7 | -4.4 | +1.7 | +7.3 | -7.6 | +6.3 | -7.3 | ... | +4.3 |
| -2.1 | +1.3 | -6.3 | -7.9 | -6.3 | -2.9 | +8.7 | +5.3 | ... | +4.9 |
| +8.0 | -8.3 | +1.9 | +1.7 | +9.1 | -4.1 | -5.1 | -7.9 | ... | -9.0 |
| ⋮    | ⋮    | ⋮    | ⋮    | ⋮    | ⋮    | ⋮    | ⋮    | ⋱   | ⋮    |
| +8.5 | +3.4 | +5.6 | -4.3 | +1.7 | -6.6 | -6.3 | +9.6 | ... | +7.5 |

Value

d\_input  
12,288

d\_output  
12,288

12,288

12,288

$$\begin{bmatrix} -3.2 & +9.1 & -5.3 & +8.9 & +8.7 & +5.9 & +2.6 & +7.4 & \cdots & -4.1 \\ +6.9 & +2.3 & -9.6 & -3.0 & -7.0 & +9.5 & -0.4 & -0.1 & \cdots & +2.8 \\ -2.6 & -7.2 & +6.4 & -6.1 & +0.2 & -5.5 & -8.0 & +7.2 & \cdots & +9.4 \\ +9.1 & +8.0 & +5.4 & -3.3 & -8.3 & -1.8 & -5.3 & -7.3 & \cdots & -8.8 \\ +4.5 & -9.7 & +5.4 & -7.0 & -8.3 & -8.1 & +3.4 & -5.0 & \cdots & -1.6 \\ +1.1 & +7.1 & +4.5 & -4.5 & -7.3 & -8.8 & -3.9 & -4.7 & \cdots & -0.9 \\ +3.6 & +3.9 & -4.3 & -2.4 & -6.3 & +5.7 & -8.8 & +3.9 & \cdots & +5.5 \\ +5.5 & -4.8 & -2.5 & +1.7 & -4.5 & -2.6 & -6.0 & -0.8 & \cdots & -9.0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ +5.9 & -8.4 & +0.4 & -3.8 & +1.5 & +9.1 & +2.9 & -9.2 & \cdots & -1.4 \end{bmatrix} \begin{bmatrix} +0.2 \\ +0.7 \\ +3.6 \\ -4.4 \\ -7.3 \\ -2.1 \\ +9.0 \\ -6.2 \\ \vdots \\ +0.9 \end{bmatrix} = \begin{bmatrix} -198.6 \\ +73.1 \\ -28.2 \\ +119.4 \\ +215.7 \\ +91.8 \\ -29.1 \\ -5.6 \\ \vdots \\ -5.1 \end{bmatrix}$$

# 6.1 Q、K和V

- 理论上，V矩阵参数数量

Query  
1,572,864

|      |      |      |      |      |      |      |      |     |      |
|------|------|------|------|------|------|------|------|-----|------|
| -3.7 | +8.9 | -2.4 | -6.3 | -9.4 | -8.6 | +8.8 | -8.9 | ... | +6.7 |
| +7.0 | +9.7 | -1.6 | +8.3 | -4.7 | -9.5 | +8.1 | -8.8 | ... | -2.3 |
| +1.7 | +6.6 | +8.6 | +7.4 | -4.5 | +6.9 | -6.2 | +9.0 | ... | +8.7 |
| ⋮    | ⋮    | ⋮    | ⋮    | ⋮    | ⋮    | ⋮    | ⋮    | ⋱   | ⋮    |
| -6.6 | +8.9 | +4.6 | -4.9 | -6.7 | +8.4 | -9.4 | -6.8 | ... | -1.5 |

Key  
1,572,864

|      |      |      |      |      |      |      |      |     |      |
|------|------|------|------|------|------|------|------|-----|------|
| -2.5 | -0.7 | -4.4 | +1.7 | +7.3 | -7.6 | +6.3 | -7.3 | ... | +4.5 |
| -2.1 | +1.3 | -6.3 | -7.8 | -8.3 | -2.9 | +8.7 | +6.9 | ... | +4.9 |
| +8.0 | -8.3 | +1.8 | +1.7 | +9.1 | -4.1 | -6.1 | -7.9 | ... | -9.6 |
| ⋮    | ⋮    | ⋮    | ⋮    | ⋮    | ⋮    | ⋮    | ⋮    | ⋱   | ⋮    |
| +8.6 | +3.4 | +6.6 | -4.8 | +1.7 | -8.6 | -8.8 | +9.6 | ... | +7.0 |

Value

12,288 × 12,288 = 150,994,944

?

12,288

12,288

|      |      |      |      |      |      |      |      |     |      |
|------|------|------|------|------|------|------|------|-----|------|
| -3.2 | +9.1 | -5.3 | +8.9 | +8.7 | +5.9 | +2.6 | +7.4 | ... | -4.1 |
| +6.9 | +2.3 | -9.6 | -3.0 | -7.0 | +9.5 | -0.4 | -0.1 | ... | +2.8 |
| -2.6 | -7.2 | +6.4 | -6.1 | +0.2 | -5.5 | -8.0 | +7.2 | ... | +9.4 |
| +9.1 | +8.0 | +5.4 | -3.3 | -8.3 | -1.8 | -5.3 | -7.3 | ... | -8.8 |
| +4.5 | -9.7 | +5.4 | -7.0 | -8.3 | -8.1 | +3.4 | -5.0 | ... | -1.6 |
| +1.1 | +7.1 | +4.5 | -4.5 | -7.3 | -8.8 | -3.9 | -4.7 | ... | -0.9 |
| +3.6 | +3.9 | -4.3 | -2.4 | -6.3 | +5.7 | -8.8 | +3.9 | ... | +5.5 |
| +5.5 | -4.8 | -2.5 | +1.7 | -4.5 | -2.6 | -6.0 | -0.8 | ... | -9.0 |
| ⋮    | ⋮    | ⋮    | ⋮    | ⋮    | ⋮    | ⋮    | ⋮    | ⋱   | ⋮    |
| +5.9 | -8.4 | +0.4 | -3.8 | +1.5 | +9.1 | +2.9 | -9.2 | ... | -1.4 |

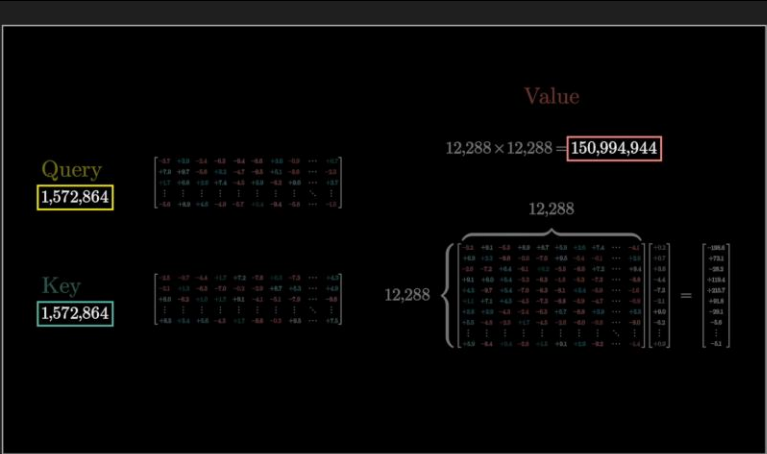
|      |
|------|
| +0.2 |
| +0.7 |
| +3.6 |
| -4.4 |
| -7.3 |
| -2.1 |
| +9.0 |
| -6.2 |
| ⋮    |
| +0.9 |

=

|        |
|--------|
| -198.6 |
| +73.1  |
| -28.2  |
| +119.4 |
| +215.7 |
| +91.8  |
| -29.1  |
| -5.6   |
| ⋮      |
| -5.1   |

# 6.1 Q、K和V

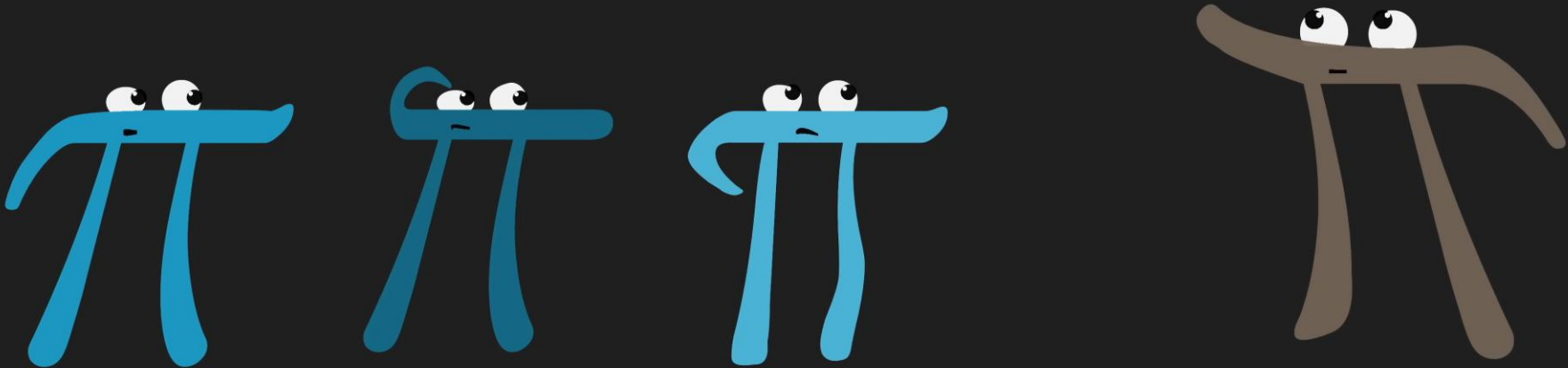
- 实际中，设定：V矩阵的参数数量 = Q、K的参数数量之和



# Value params

||

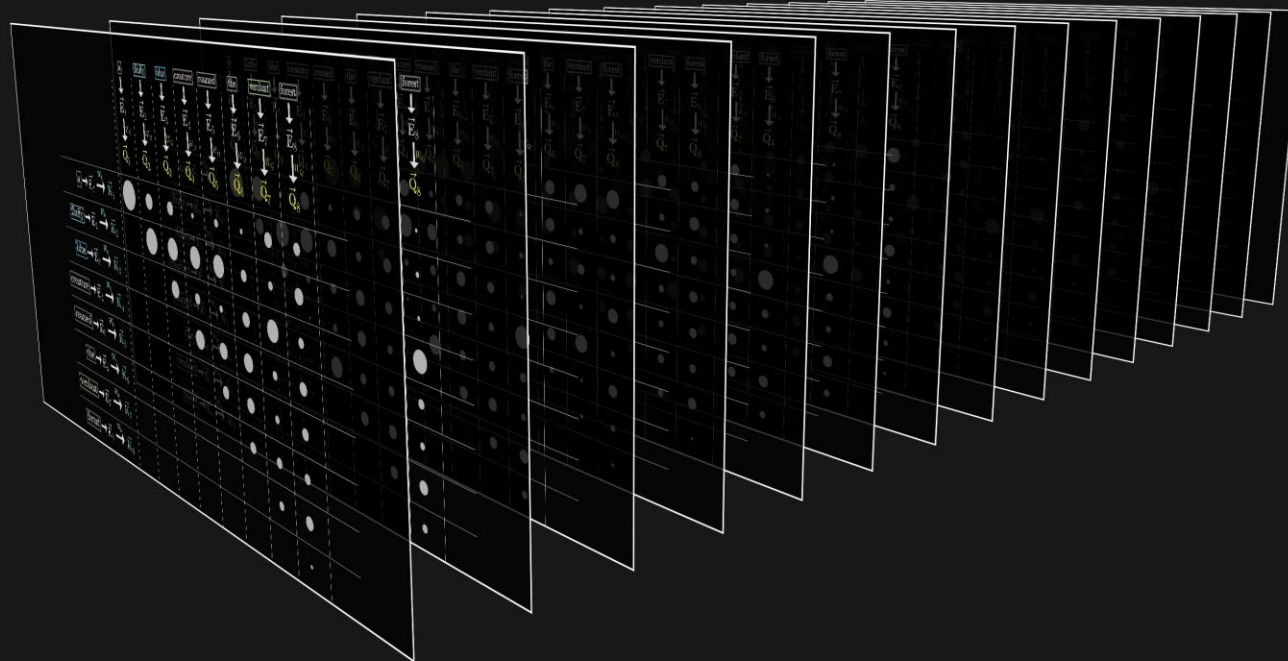
(# Query params) + (# Key params)



## 6.1 Q、K和V

- 有利于更高效地实现“多头注意力”

### Multi-headed attention



# 6.2 V矩阵的设计

- 将V矩阵拆为两个小矩阵

Value

~~12,288 × 12,288 = 150,994,944~~

12,288

12,288

-3.2

+9.1

-5.3

+8.9

+8.7

+5.9

+2.6

+7.4

...

-4.1

+6.9

+2.3

-9.6

-3.0

-7.0

+9.5

-0.4

-0.1

...

+2.8

-2.6

-7.2

+6.4

-6.1

+0.2

-5.5

-8.0

+7.2

...

+9.4

+9.1

+8.0

+5.4

-3.3

-8.3

-1.8

-5.3

-7.3

...

-8.8

+4.5

-9.7

+5.4

-7.0

-8.3

-8.1

+3.4

-5.0

...

-1.6

+1.1

+7.1

+4.5

-4.5

-7.3

-8.8

-3.9

-4.7

...

-0.9

+3.6

+3.9

-4.3

-2.4

-6.3

+5.7

-8.8

+3.9

...

+5.5

+5.5

-4.8

-2.5

+1.7

-4.5

-2.6

-6.0

-0.8

...

-9.0

⋮

⋮

⋮

⋮

⋮

⋮

⋮

⋮

⋮

⋮

+5.9

-8.4

+0.4

-3.8

+1.5

+9.1

+2.9

-9.2

...

-1.4

+0.2

+0.7

+3.6

-4.4

-7.3

-2.1

+9.0

-6.2

⋮

+0.9

=

-198.6

+73.1

-28.2

+119.4

+215.7

+91.8

-29.1

-5.6

⋮

-5.1

105

# 6.2 V矩阵的设计

- 将V矩阵拆为两个小矩阵

Value

12,288

+5.0

-3.3

...

+7.2

-8.9

-4.9

...

-7.8

-3.0

+4.8

...

+2.4

+4.2

-5.8

...

+3.5

+7.5

+0.9

...

-9.3

+4.2

-9.7

...

+0.6

+8.4

-8.1

...

-9.4

-3.1

+2.4

...

-5.7

⋮

⋮

⋮

⋮

+8.9

-9.8

...

+2.0

12,288

-7.8

+8.9

-5.3

+3.8

-8.7

+4.6

+7.6

-4.5

...

-2.5

+4.9

-5.2

-6.5

-1.0

-3.9

+6.7

-5.2

+0.0

...

+2.7

+7.3

+8.7

+5.0

+4.0

+9.3

+9.8

-1.0

-8.5

...

-6.9

⋮

⋮

⋮

⋮

⋮

⋮

⋮

⋮

⋮

⋮

-1.8

+1.0

-4.5

-0.9

-1.9

-5.0

+0.1

-3.8

...

+0.5

+0.2

+0.7

+3.6

-4.4

-7.3

-2.1

+9.0

-6.2

⋮

+0.9

=

-198.6

+73.1

-28.2

+119.4

+215.7

+91.8

-29.1

-5.6

⋮

-5.1



## 6.2 V矩阵的设计

- 两个小矩阵作为一个整体，构成线性变换

# Linear map

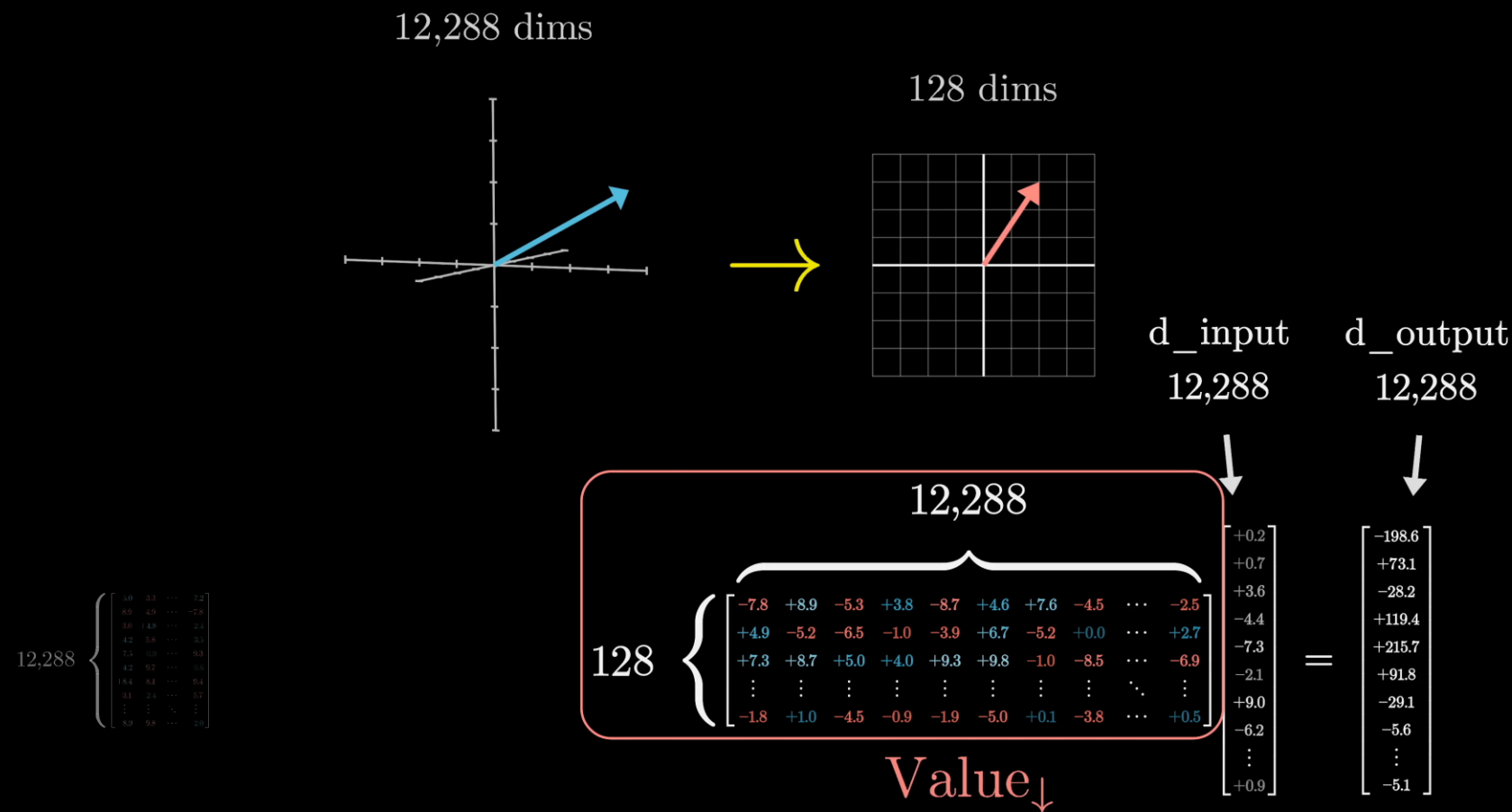
$$12,288 \left\{ \begin{bmatrix} +5.0 & -3.3 & \cdots & +7.2 \\ -8.9 & -4.9 & \cdots & -7.8 \\ -3.0 & +4.8 & \cdots & +2.4 \\ +4.2 & -5.8 & \cdots & +3.5 \\ +7.5 & +0.9 & \cdots & -9.3 \\ +4.2 & -9.7 & \cdots & +0.6 \\ +8.4 & -8.1 & \cdots & -9.4 \\ -3.1 & +2.4 & \cdots & -5.7 \\ \vdots & \vdots & \ddots & \vdots \\ +8.9 & -9.8 & \cdots & +2.0 \end{bmatrix} \overbrace{\begin{bmatrix} -7.8 & +8.9 & -5.3 & +3.8 & -8.7 & +4.6 & +7.6 & -4.5 & \cdots & -2.5 \\ +4.9 & -5.2 & -6.5 & -1.0 & -3.9 & +6.7 & -5.2 & +0.0 & \cdots & +2.7 \\ +7.3 & +8.7 & +5.0 & +4.0 & +9.3 & +9.8 & -1.0 & -8.5 & \cdots & -6.9 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ -1.8 & +1.0 & -4.5 & -0.9 & -1.9 & -5.0 & +0.1 & -3.8 & \cdots & +0.5 \end{bmatrix}}^{12,288} \begin{bmatrix} +0.2 \\ +0.7 \\ +3.6 \\ -4.4 \\ -7.3 \\ -2.1 \\ +9.0 \\ -6.2 \\ \vdots \\ +0.9 \end{bmatrix} = \begin{bmatrix} -198.6 \\ +73.1 \\ -28.2 \\ +119.4 \\ +215.7 \\ +91.8 \\ -29.1 \\ -5.6 \\ \vdots \\ -5.1 \end{bmatrix}$$





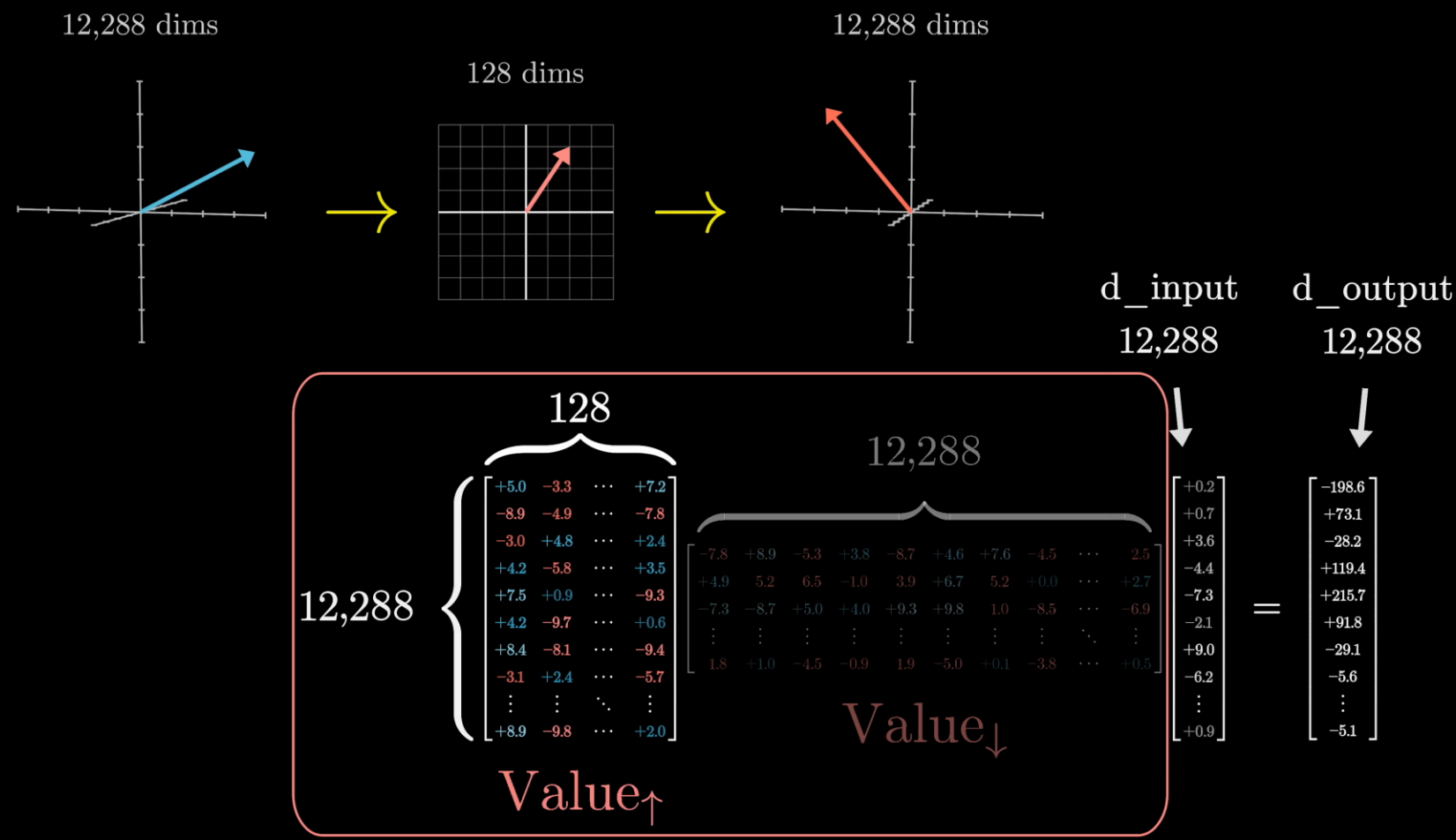
# 6.2 V矩阵的设计

- 第一步：“值↓”矩阵将12288维的嵌入空间，映射到128维



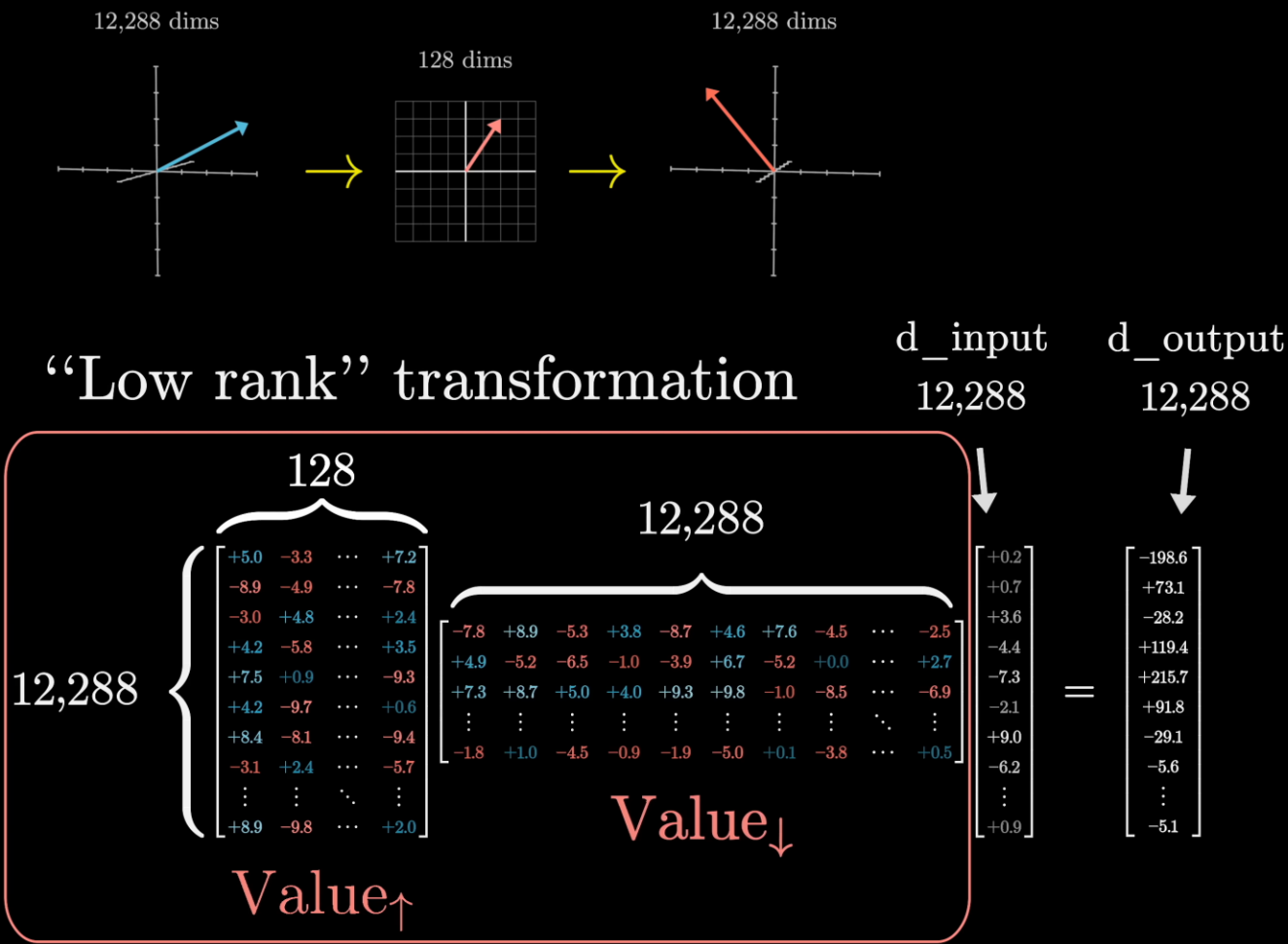
## 6.2 V矩阵的设计

- 第二步：“值↑”矩阵将128维的，映射回12288维



# 6.2 V矩阵的设计

- 拆分的过程，称为“低秩变换”



## 6.3 单头注意力参数数量

- 参数数量

Total weights: 175,181,291,520  
Organized into 27,938 matrices



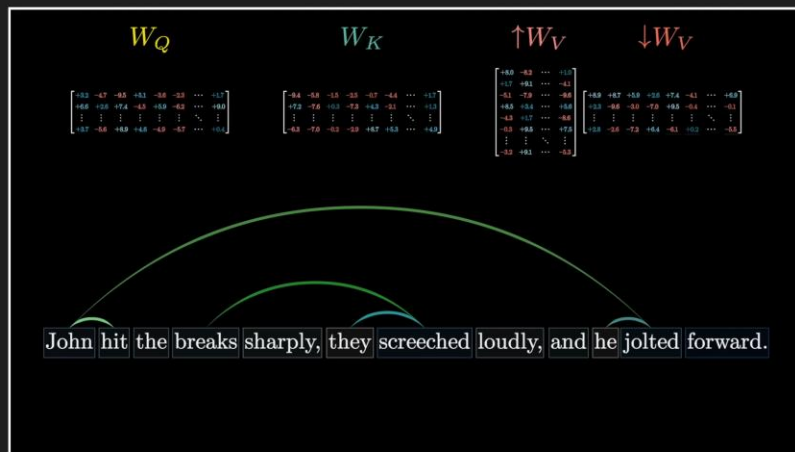
|                    |                                                                             |
|--------------------|-----------------------------------------------------------------------------|
| Embedding          | $\overset{12,288}{d\_embed} * \overset{50,257}{n\_vocab} = 617,558,016$     |
| Key                | $\overset{128}{d\_query} * \overset{12,288}{d\_embed} = 1,572,864$ per head |
| Query              | $\overset{128}{d\_query} * \overset{12,288}{d\_embed} = 1,572,864$ per head |
| Value <sub>↓</sub> | $\overset{128}{d\_value} * \overset{12,288}{d\_embed} = 1,572,864$ per head |
| Value <sub>↑</sub> | $\overset{12,288}{d\_embed} * \overset{128}{d\_value} = 1,572,864$ per head |
| Up-projection      | 6,291,456                                                                   |
| Down-projection    |                                                                             |
| Unembedding        | $\overset{50,257}{n\_vocab} * \overset{12,288}{d\_embed} = 617,558,016$     |
|                    |                                                                             |



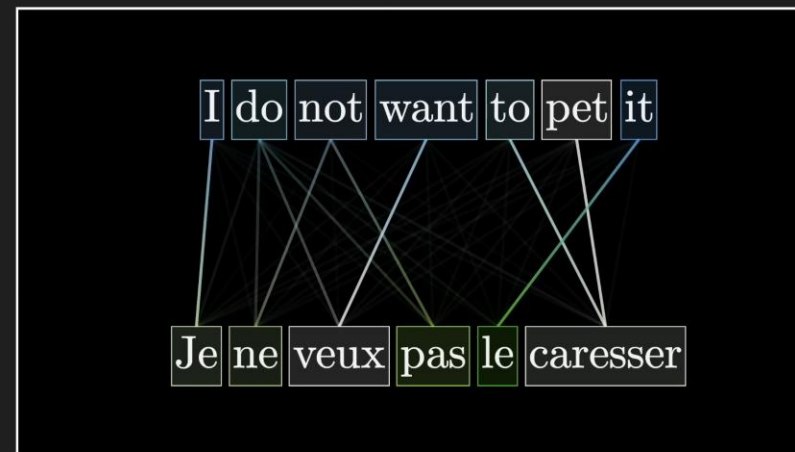
## 6.4 交叉注意力

- 前文所述均为“自注意力”
- 这里补充“交叉注意力”

### Self-attention



### Cross-attention



## 6.4 交叉注意力

- 交叉注意力描述两种序列的关系
- 例如，对于两种语言文本的翻译

I do not want to pet it

Je ne veux pas le caresser



## 6.4 交叉注意力

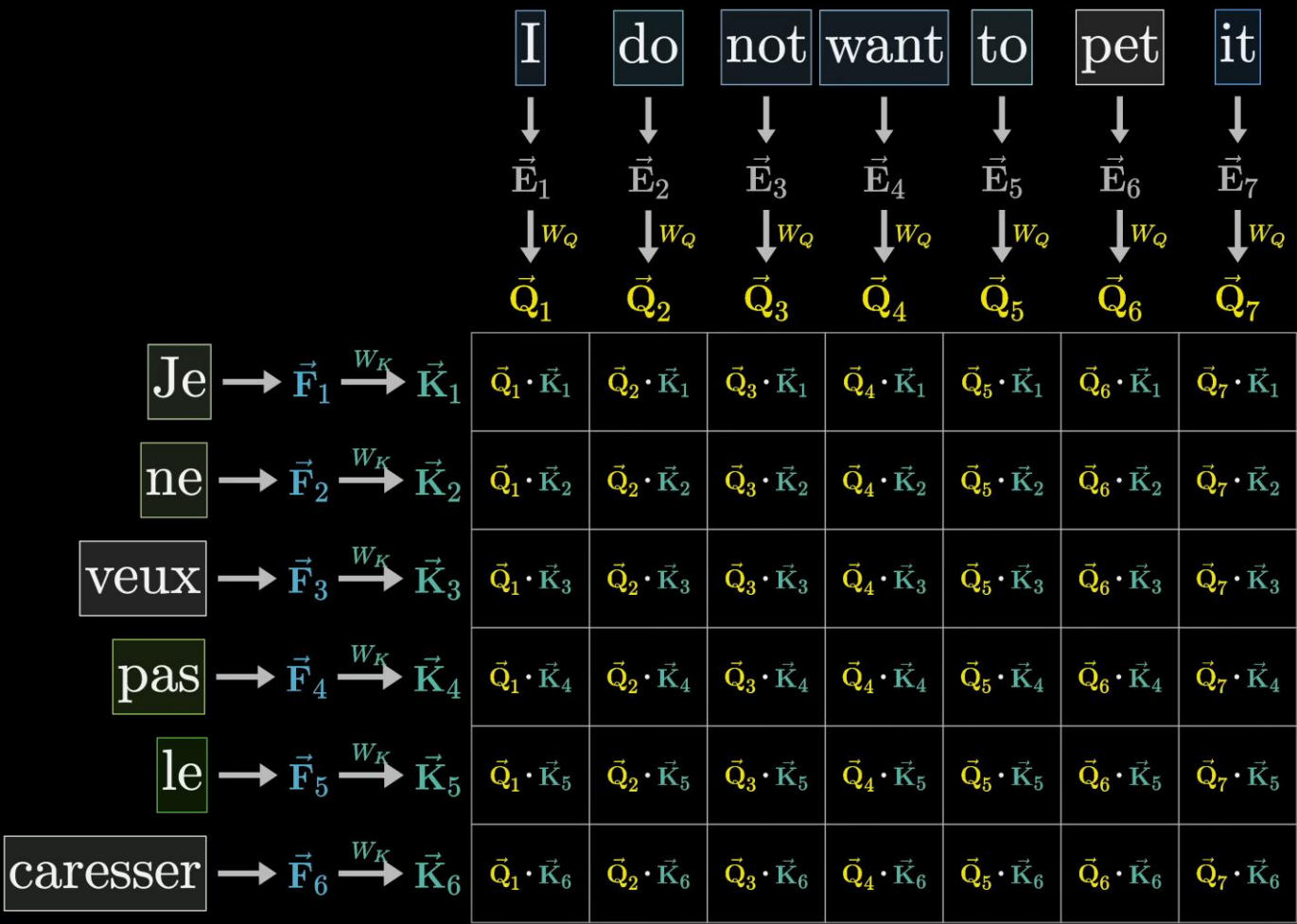
- 又如，录音转文字



or maybe audio input of speech, and an ongoing transcription

# 6.4 交叉注意力

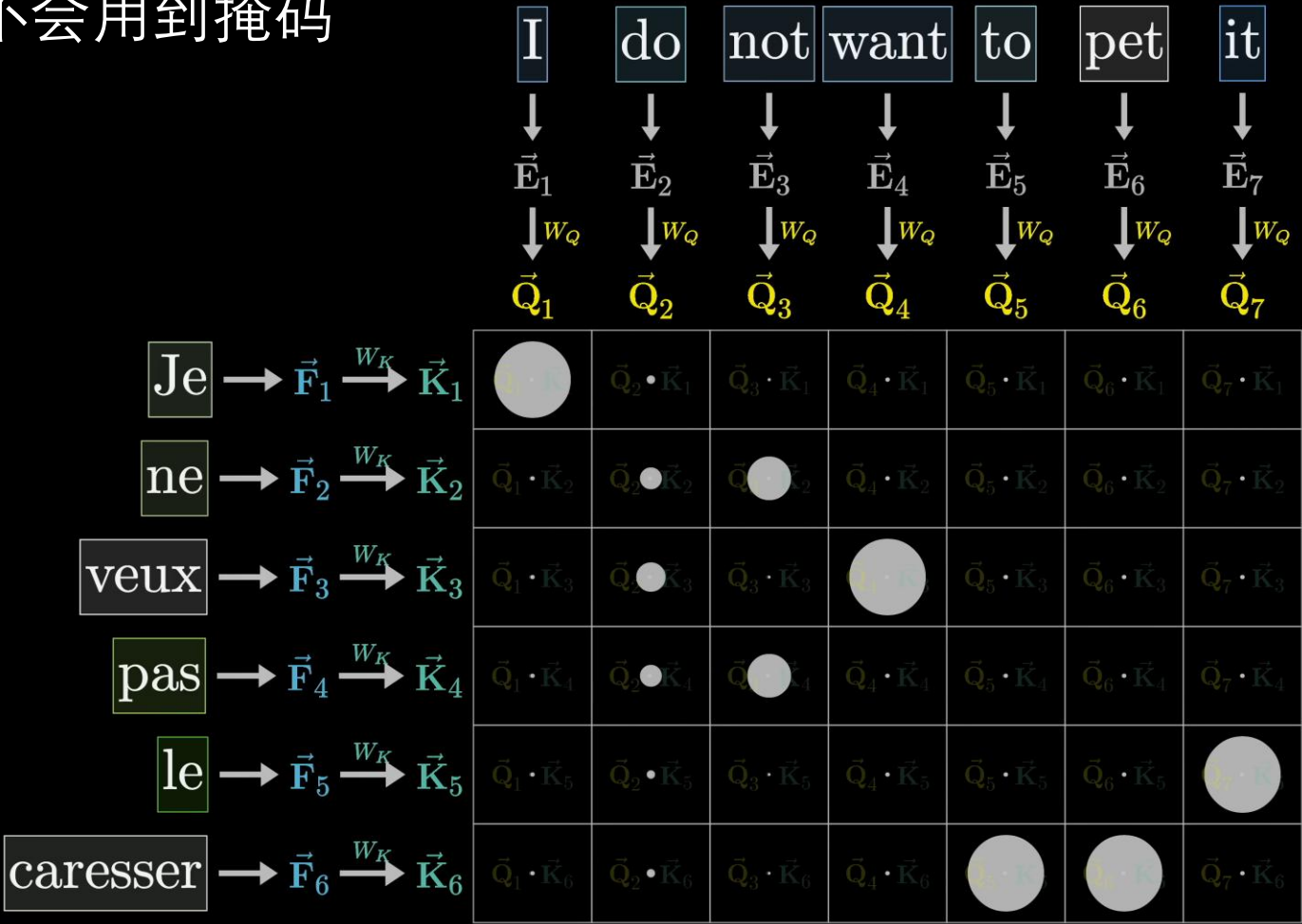
- 交叉注意力与自注意力非常类似





# 6.4 交叉注意力

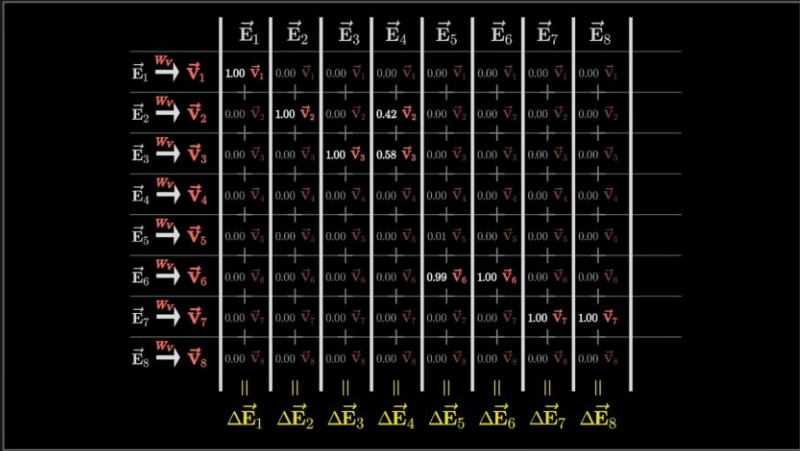
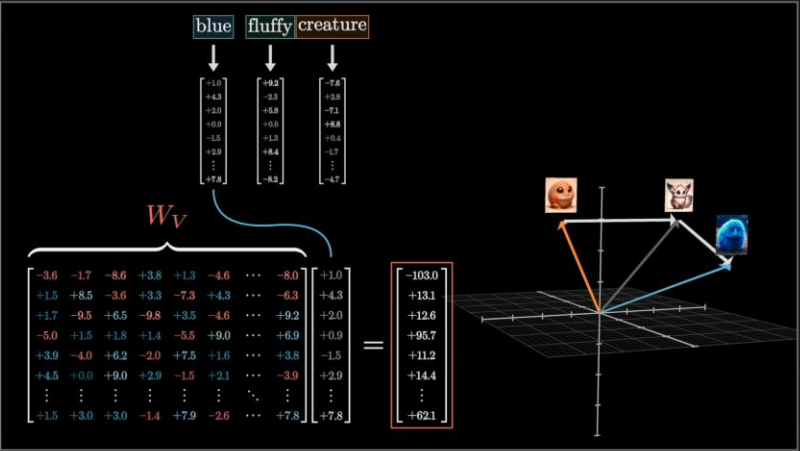
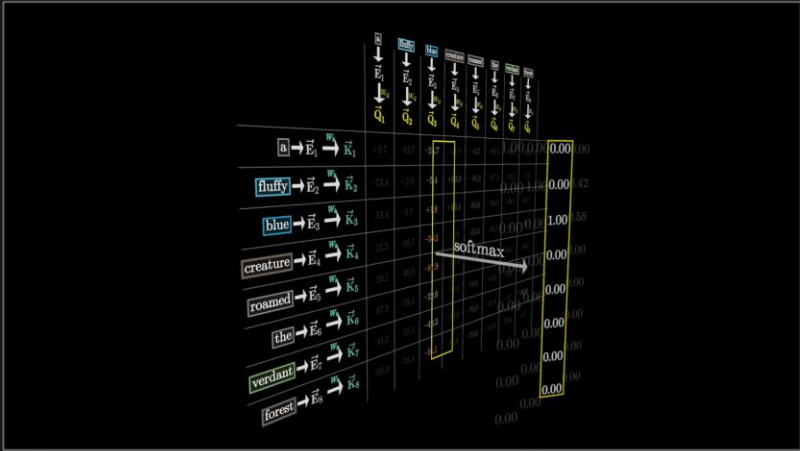
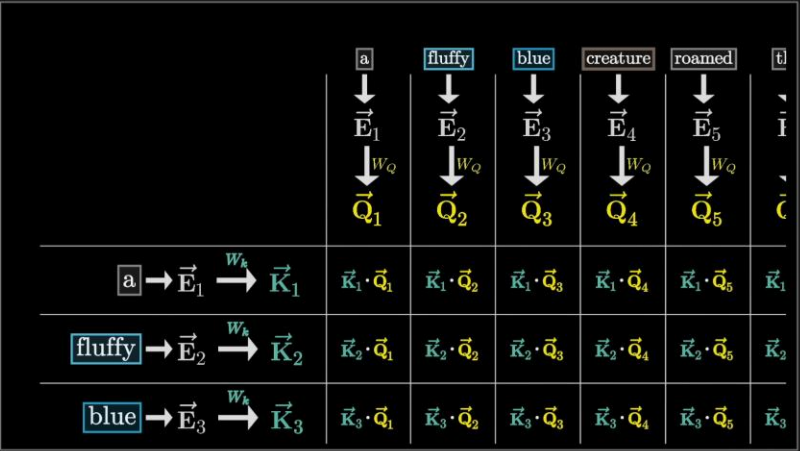
- 后者关注同一文本的词元之间的联系
- 前者关注不同语言的词之间的联系
- 交叉注意力通常不会用到掩码



1. 回顾
2. 启发示例：注意力的作用
3. 注意力机制：查询（Query）、键（Key）
4. 掩码（Masking）
5. 注意力机制：值（Value）
6. 参数数量
7. 多头注意力
8. Transformer的深度

# 7.1 不同语义的注意力

- 前面在单头注意力中，我们分别了解了查询、键、值和注意力模式

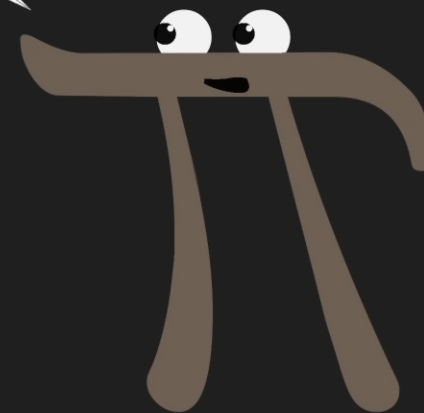
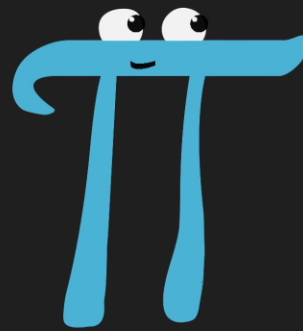
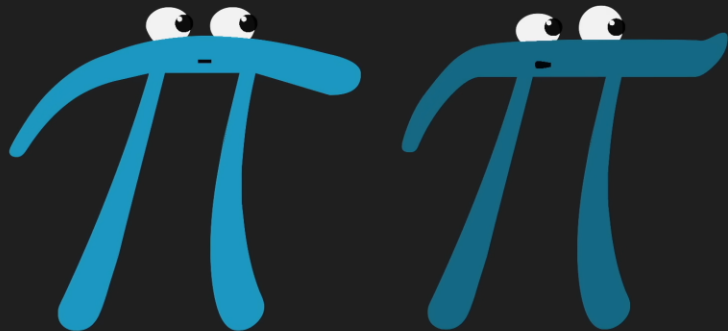


## 7.1 不同语义的注意力

- 而多头注意力就是重复并列“堆叠”

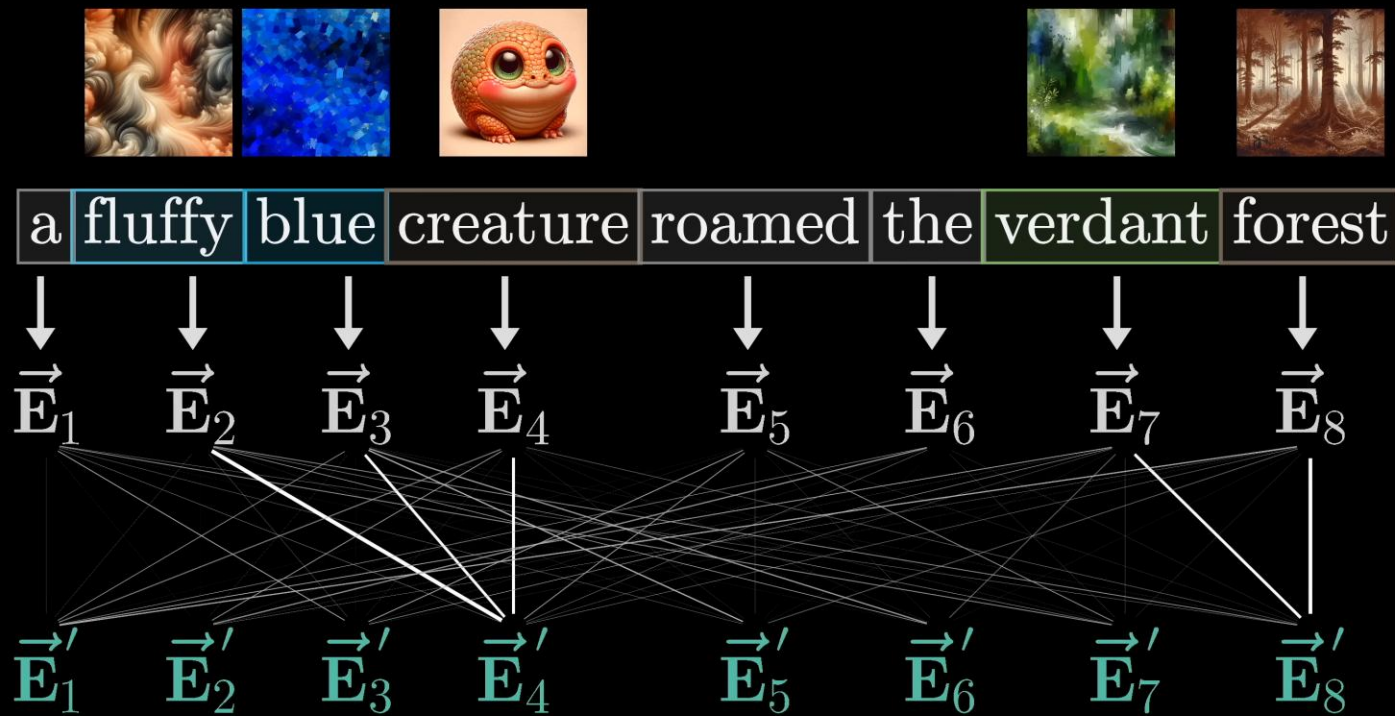


Now do that about  
10,000 times



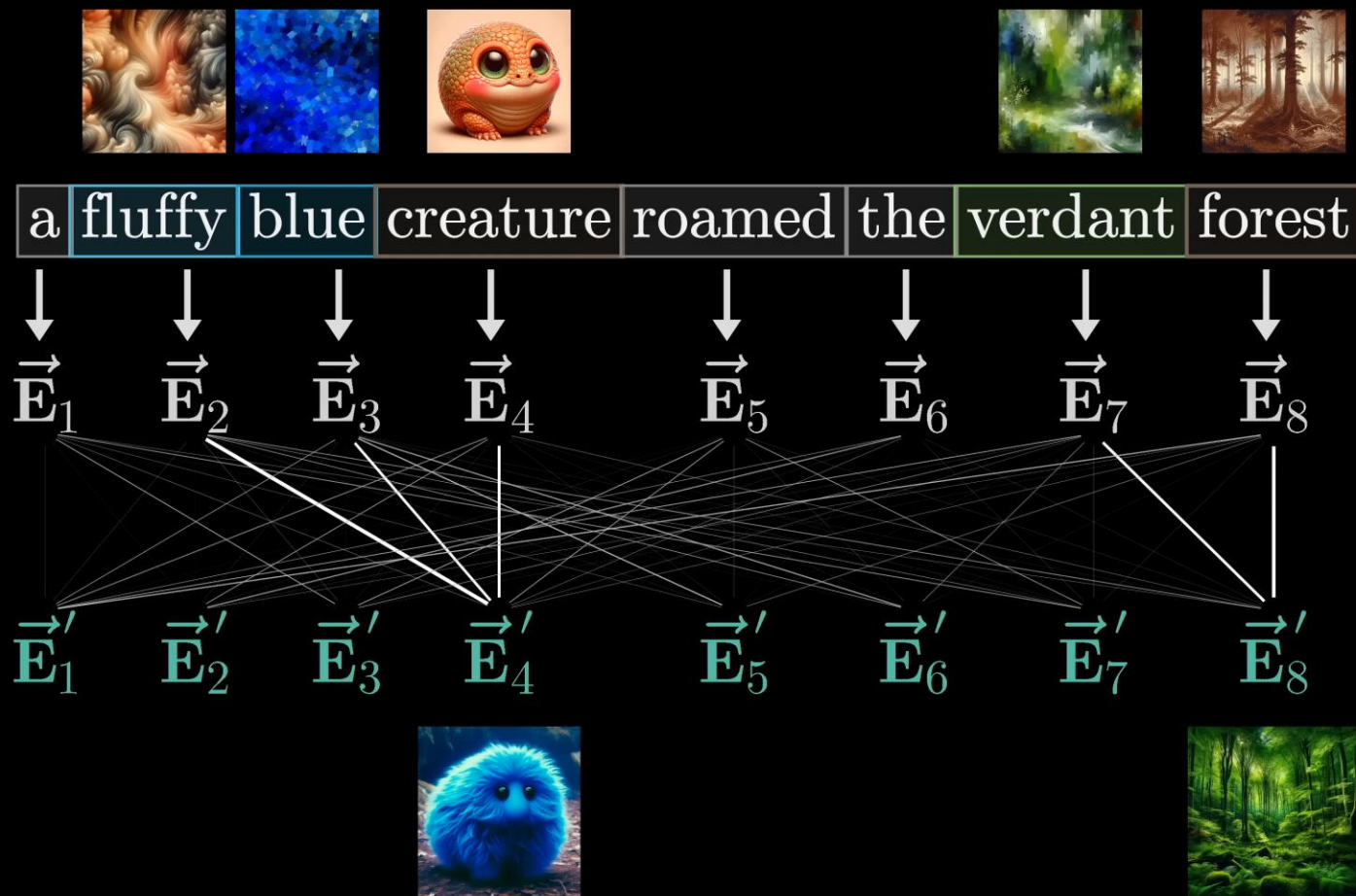
## 7.1 不同语义的注意力

- 此前的例子中，注意力描述了形容词对名词的修饰关系



## 7.1 不同语义的注意力

- 但语义的关系远不止于此



## 7.1 不同语义的注意力

- 例如，动词对名词的作用关系



... when suddenly they crashed the car into a tree ...

## 7.1 不同语义的注意力

- 例如，动词对名词的作用关系

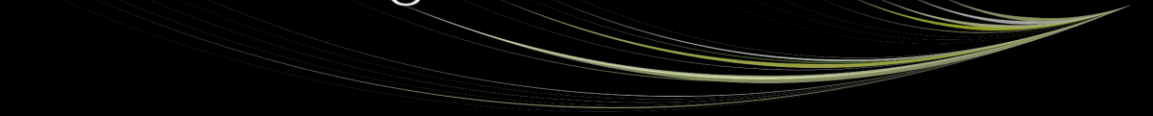




## 7.1 不同语义的注意力

- 又如，非语法的，语义上的关系

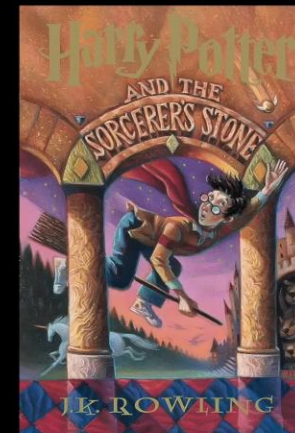
... wizard ... Hogwarts ... Hermione ... Harry



## 7.1 不同语义的注意力

- 又如，非语法的，语义上的关系

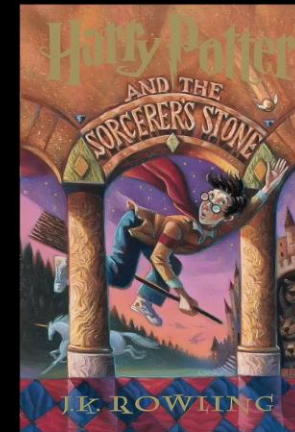
... wizard ... Hogwarts ... Hermione ... Harry



## 7.1 不同语义的注意力

- 又如，非语法的，语义上的关系

... wizard ... Hogwarts ... Hermione ... Harry

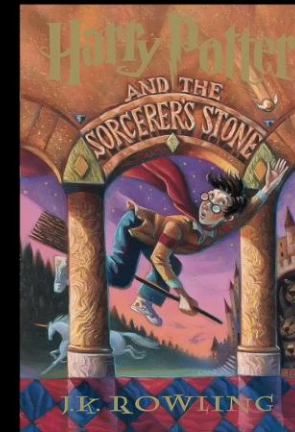


... Queen ... Sussex ... William ... Harry

## 7.1 不同语义的注意力

- 又如，非语法的，语义上的关系

... wizard ... Hogwarts ... Hermione ... Harry



... Queen ... Sussex ... William ... Harry



## 7.2 多头注意力

- 用不同的“头”来描述不同种类潜在的语义关系

 $W_Q$ 

$$\begin{bmatrix} +3.2 & -4.2 & +2.3 & -1.4 & -7.2 & -4.0 & \cdots & +1.4 \\ +1.8 & +1.5 & +3.0 & +3.0 & -1.4 & +7.9 & \cdots & -2.6 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ -1.3 & +7.8 & +6.1 & +4.0 & -7.9 & +8.3 & \cdots & +4.2 \end{bmatrix}$$

 $W_K$ 

$$\begin{bmatrix} +9.9 & -6.9 & +7.3 & -6.7 & +2.3 & -7.4 & \cdots & +6.9 \\ +6.1 & +1.4 & -1.8 & -8.5 & +3.9 & -0.9 & \cdots & +4.4 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ +7.3 & +9.4 & +7.0 & -9.7 & -2.8 & +4.6 & \cdots & -6.5 \end{bmatrix}$$

 $\uparrow W_V$ 

$$\begin{bmatrix} +0.4 & -8.8 & \cdots & -5.9 \\ -9.5 & +5.8 & \cdots & -5.5 \\ -3.1 & +8.5 & \cdots & +4.0 \\ -9.3 & -6.6 & \cdots & +2.4 \\ +1.5 & -5.2 & \cdots & +8.6 \\ +2.3 & +0.7 & \cdots & +1.8 \\ \vdots & \vdots & \ddots & \vdots \\ +4.6 & -3.7 & \cdots & -2.0 \end{bmatrix}$$

 $\downarrow W_V$ 

$$\begin{bmatrix} -5.7 & -6.2 & +8.8 & +4.7 & -0.2 & -5.4 & \cdots & -4.9 \\ -8.8 & -1.3 & -3.7 & +3.9 & -2.4 & -6.3 & \cdots & -9.4 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ -8.6 & +3.6 & -0.9 & +0.7 & +7.9 & +9.7 & \cdots & -5.6 \end{bmatrix}$$

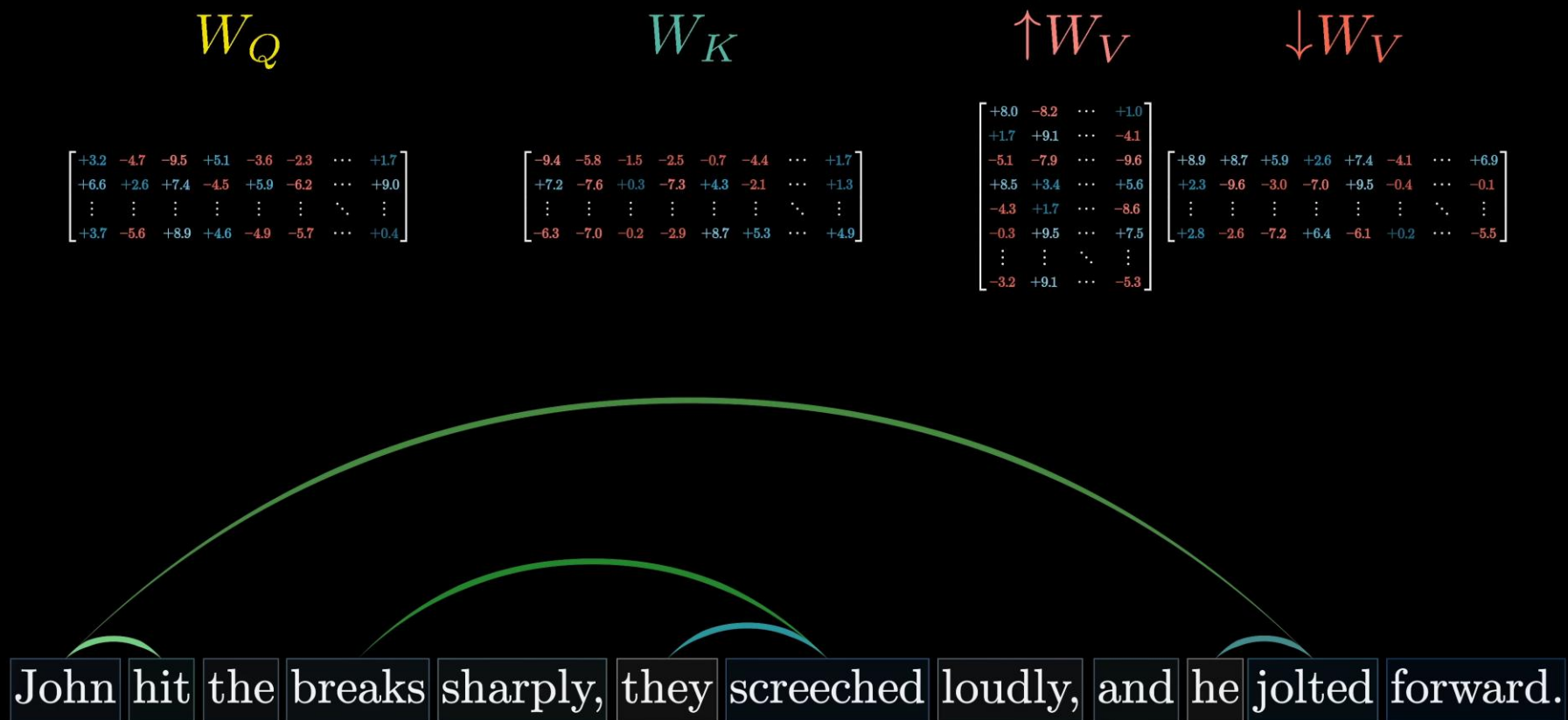
John hit the breaks sharply, they screeched loudly, and he jolted forward.

约翰猛踩刹车，刹车发出刺耳的声音，他颠簸着前进



## 7.2 多头注意力

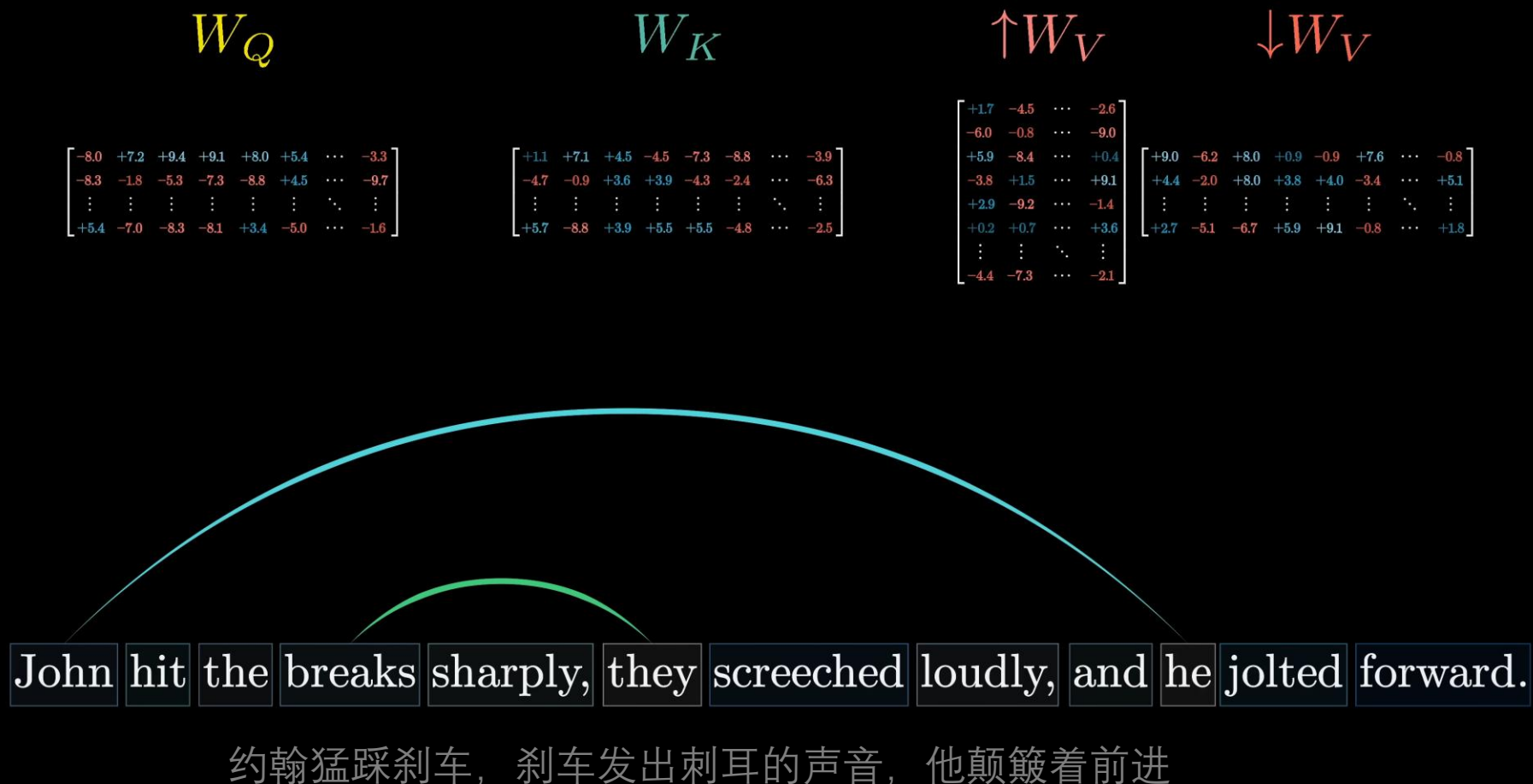
- 每个“头”的查询、键、值矩阵和注意力模式都不相同（从数据中训练得到）



约翰猛踩刹车，刹车发出刺耳的声音，他颠簸着前进

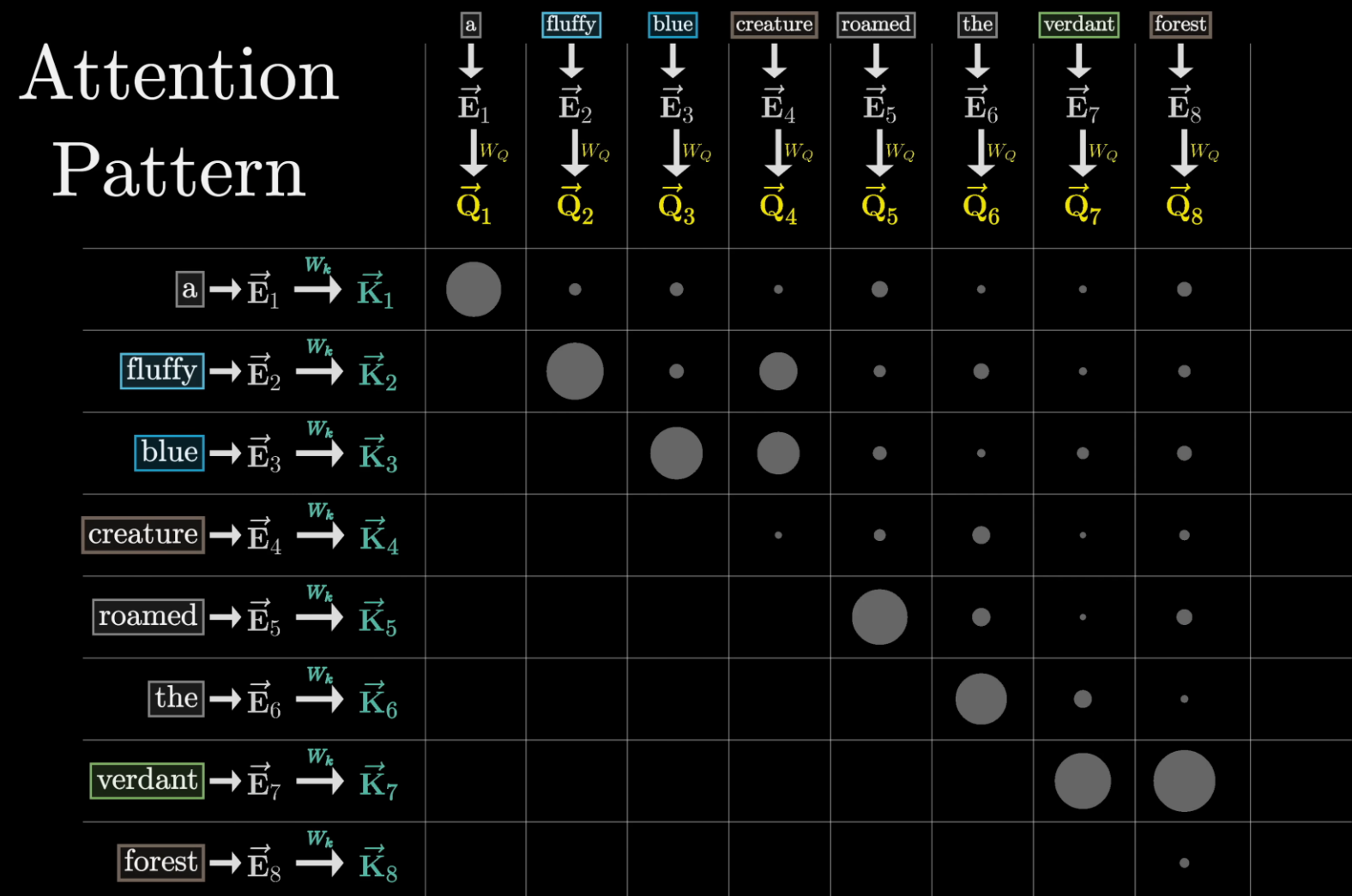
## 7.2 多头注意力

- 每个“头”的查询、键、值矩阵和注意力模式都不相同（从数据中训练得到）



## 7.2 多头注意力

- 每个“头”的查询、键、值矩阵和注意力模式都不相同（从数据中训练得到）

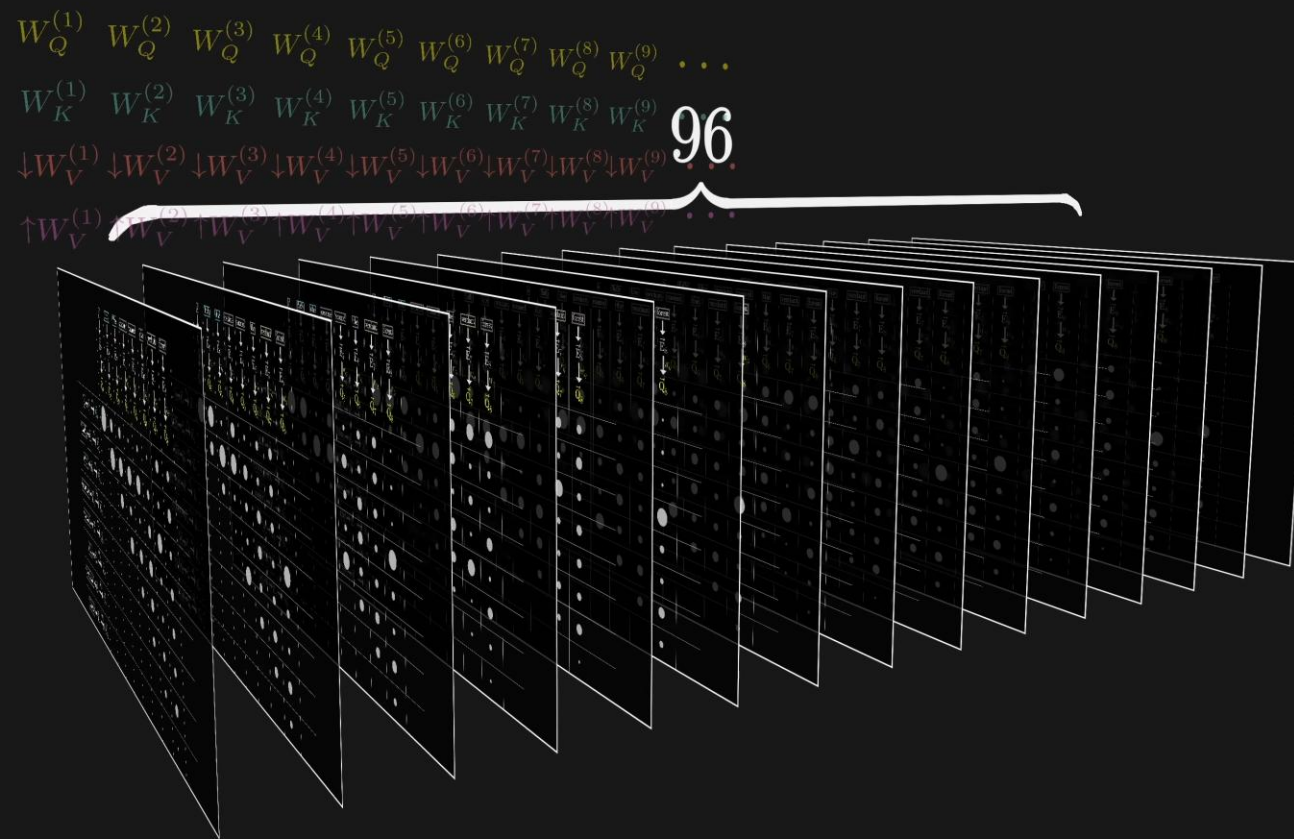




## 7.2 多头注意力

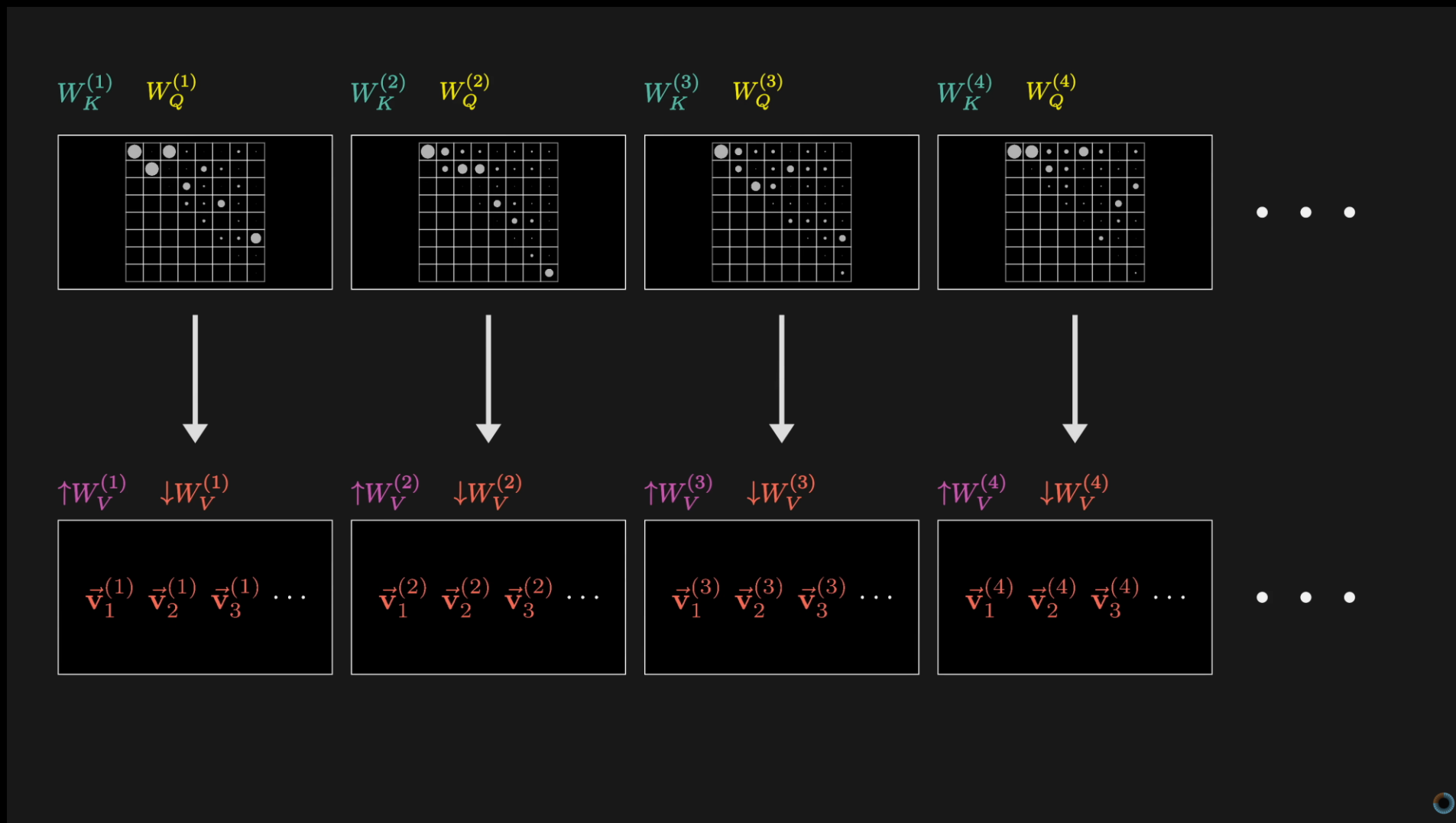
- GPT-3中，每个注意力模块，由96个“头”组成

### Multi-headed attention



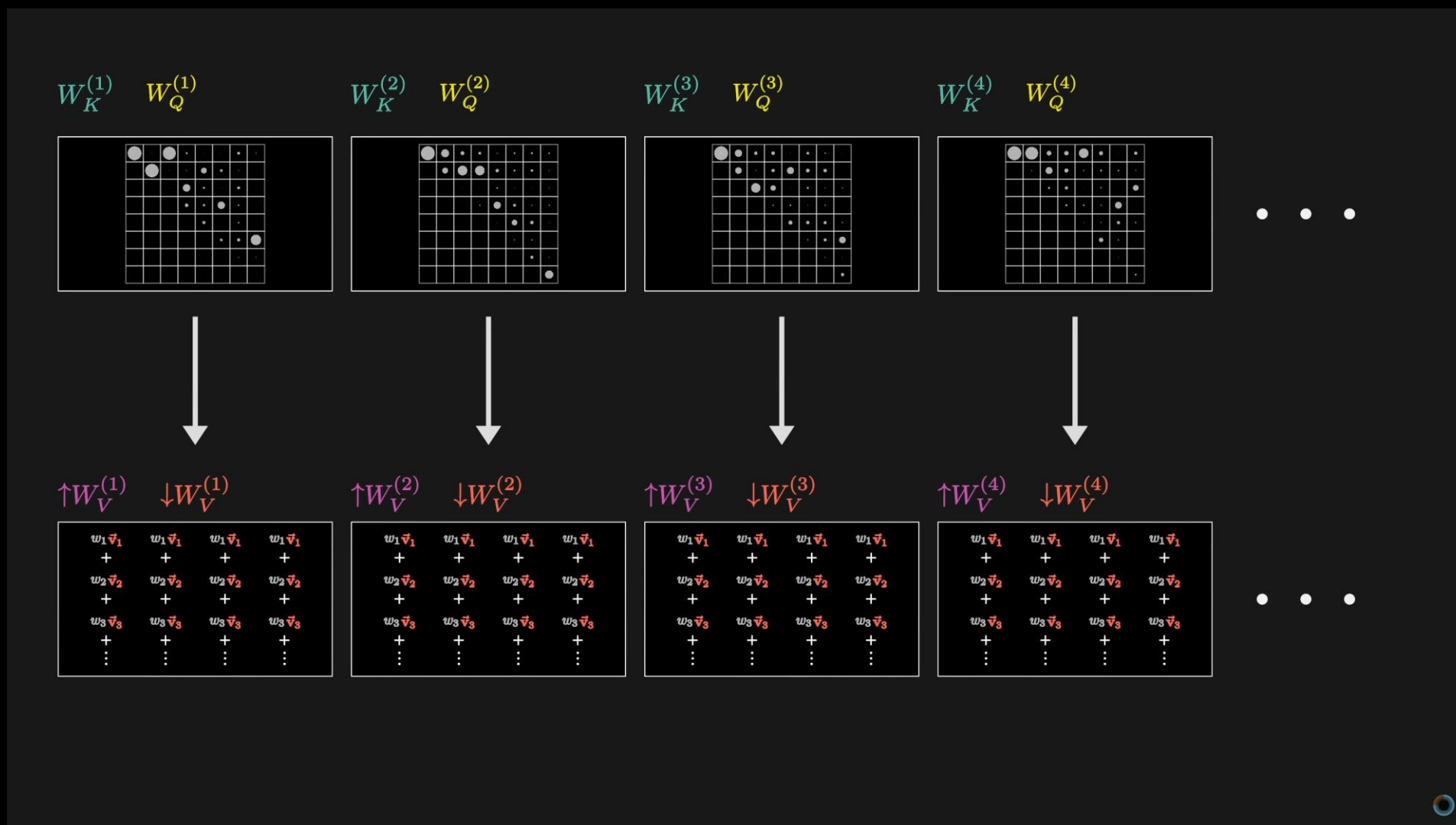
## 7.2 多头注意力

- 每个“头”，由一组查询、键、值矩阵构成



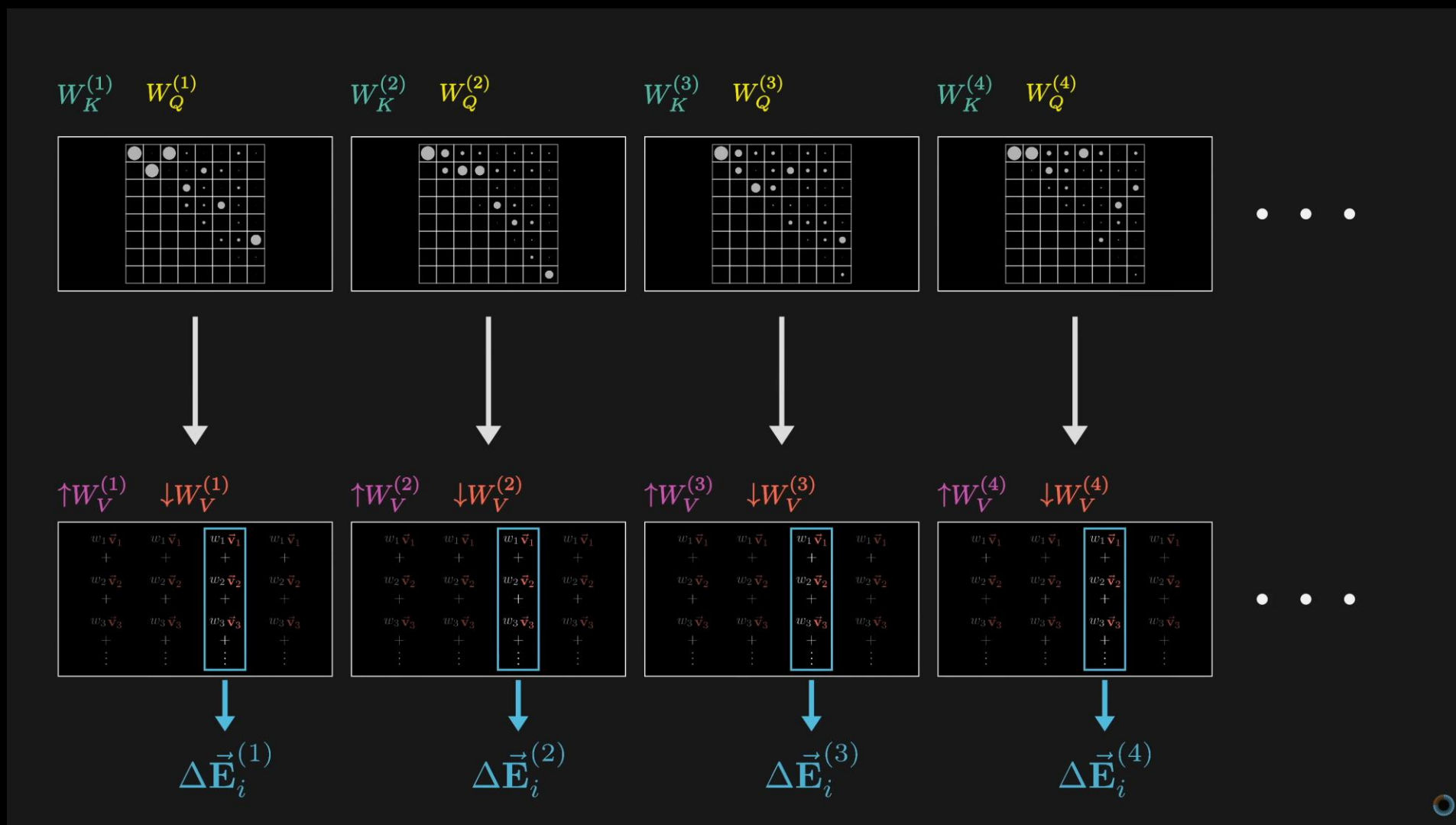
## 7.2 多头注意力

- 进而每个“头”，都会根据注意力模式，分别作加权求和



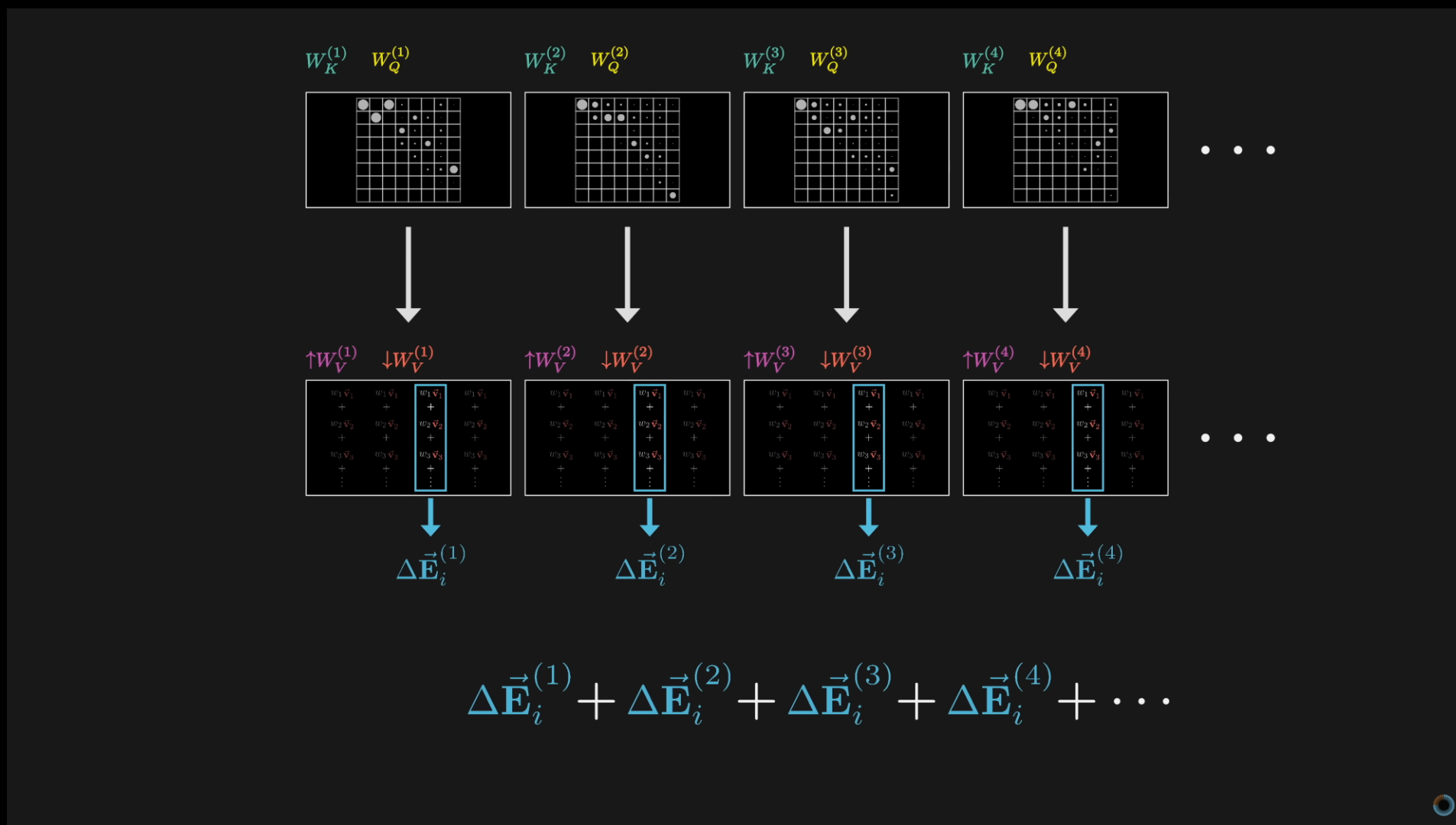
## 7.2 多头注意力

- 对每个词元，每个头都会给出一个嵌入向量的更新



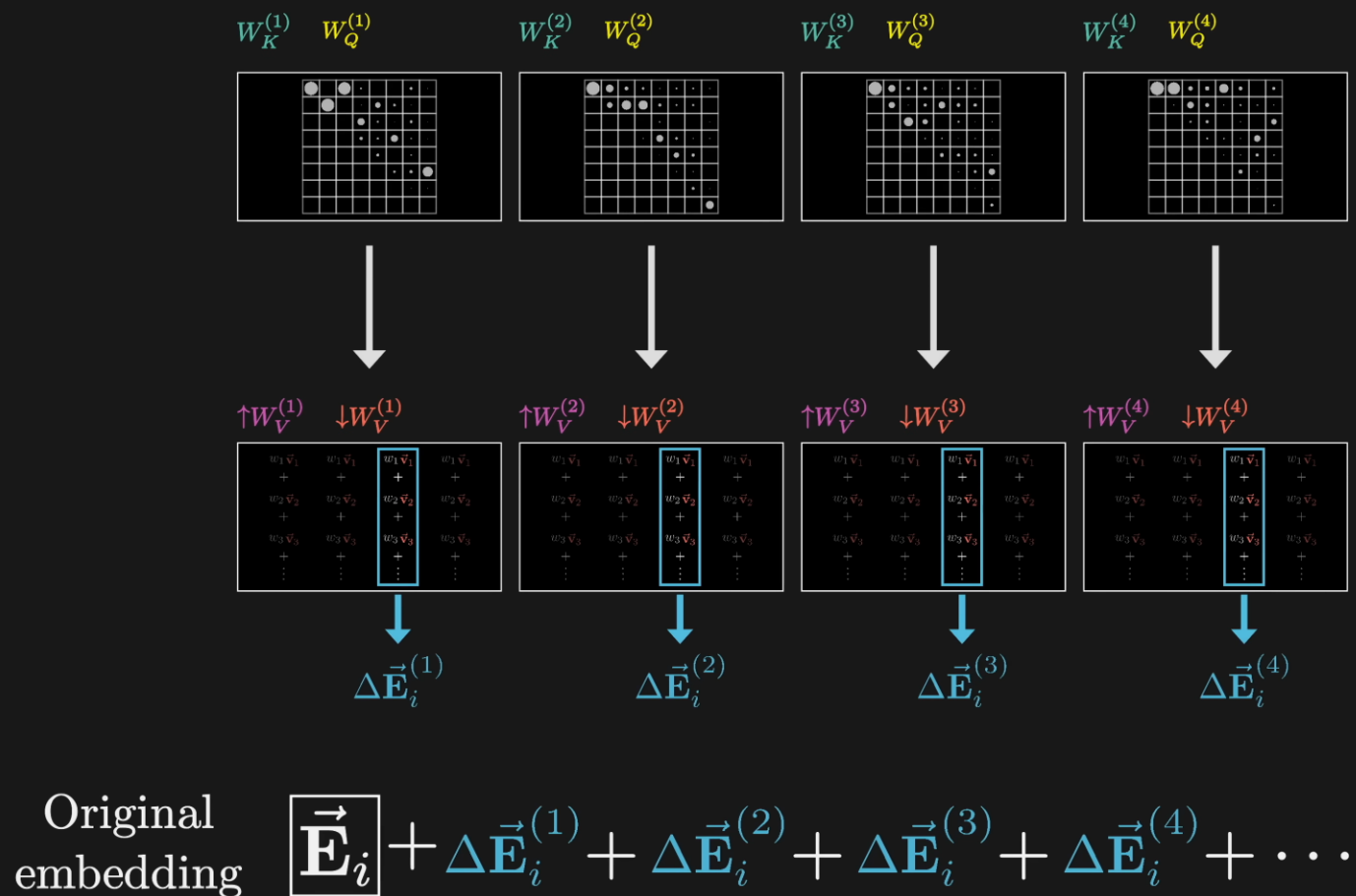
## 7.2 多头注意力

- 将每个头的更新加起来作为总的更新



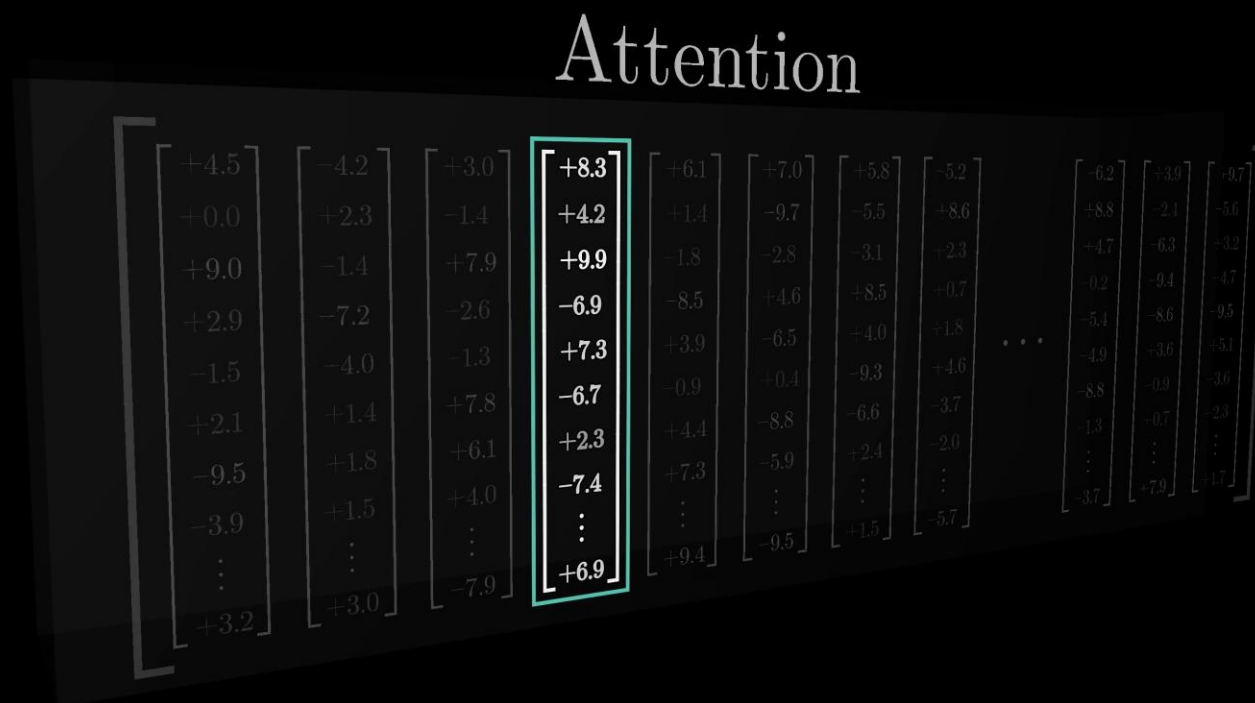
## 7.2 多头注意力

- 总的更新与原始的嵌入向量相加



## 7.2 多头注意力

- 由此构成了注意力模块的一系列输出



New  
embedding

$$\vec{\mathbf{E}}_i + \Delta \vec{\mathbf{E}}_i^{(1)} + \Delta \vec{\mathbf{E}}_i^{(2)} + \Delta \vec{\mathbf{E}}_i^{(3)} + \Delta \vec{\mathbf{E}}_i^{(4)} + \dots$$



## 7.2 多头注意力

- 头数的增加，参数的数量增加为96倍

Total weights: 175,181,291,520  
Organized into 27,938 matrices



|                    |                                                                             |
|--------------------|-----------------------------------------------------------------------------|
| Embedding          | $\overset{12,288}{d\_embed} * \overset{50,257}{n\_vocab} = 617,558,016$     |
| Key                | $\overset{128}{d\_query} * \overset{12,288}{d\_embed} = 1,572,864$ per head |
| Query              | $\overset{128}{d\_query} * \overset{12,288}{d\_embed} = 1,572,864$ per head |
| Value <sub>↓</sub> | $\overset{128}{d\_value} * \overset{12,288}{d\_embed} = 1,572,864$ per head |
| Value <sub>↑</sub> | $\overset{12,288}{d\_embed} * \overset{128}{d\_value} = 1,572,864$ per head |
| Up-projection      |                                                                             |
| Down-projection    |                                                                             |
| Unembedding        | $\overset{50,257}{n\_vocab} * \overset{12,288}{d\_embed} = 617,558,016$     |
|                    |                                                                             |





## 7.2 多头注意力

- 头数的增加，参数的数量增加为96倍

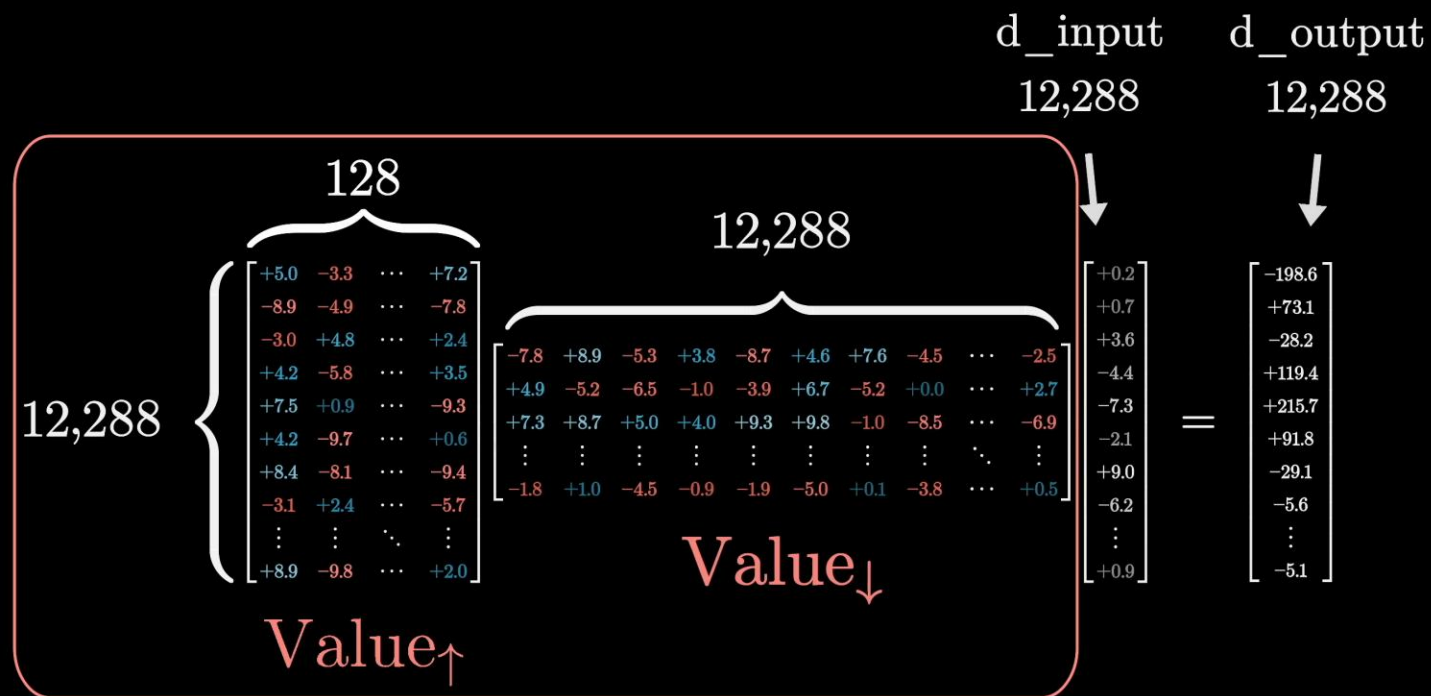
Total weights: 175,181,291,520  
Organized into 27,938 matrices



|                    |                                                                                                         |
|--------------------|---------------------------------------------------------------------------------------------------------|
| Embedding          | $\overset{12,288}{d\_embed} * \overset{50,257}{n\_vocab} = 617,558,016$                                 |
| Key                | $\overset{128}{d\_query} * \overset{12,288}{d\_embed} * \overset{96}{n\_heads} = 150,994,944$ per layer |
| Query              | $\overset{128}{d\_query} * \overset{12,288}{d\_embed} * \overset{96}{n\_heads} = 150,994,944$ per layer |
| Value <sub>↓</sub> | $\overset{128}{d\_value} * \overset{12,288}{d\_embed} * \overset{96}{n\_heads} = 150,994,944$ per layer |
| Value <sub>↑</sub> | $\overset{12,288}{d\_embed} * \overset{128}{d\_value} * \overset{96}{n\_heads} = 150,994,944$ per layer |
| Up-projection      | 603,979,776                                                                                             |
| Down-projection    |                                                                                                         |
| Unembedding        | $\overset{50,257}{n\_vocab} * \overset{12,288}{d\_embed} = 617,558,016$                                 |

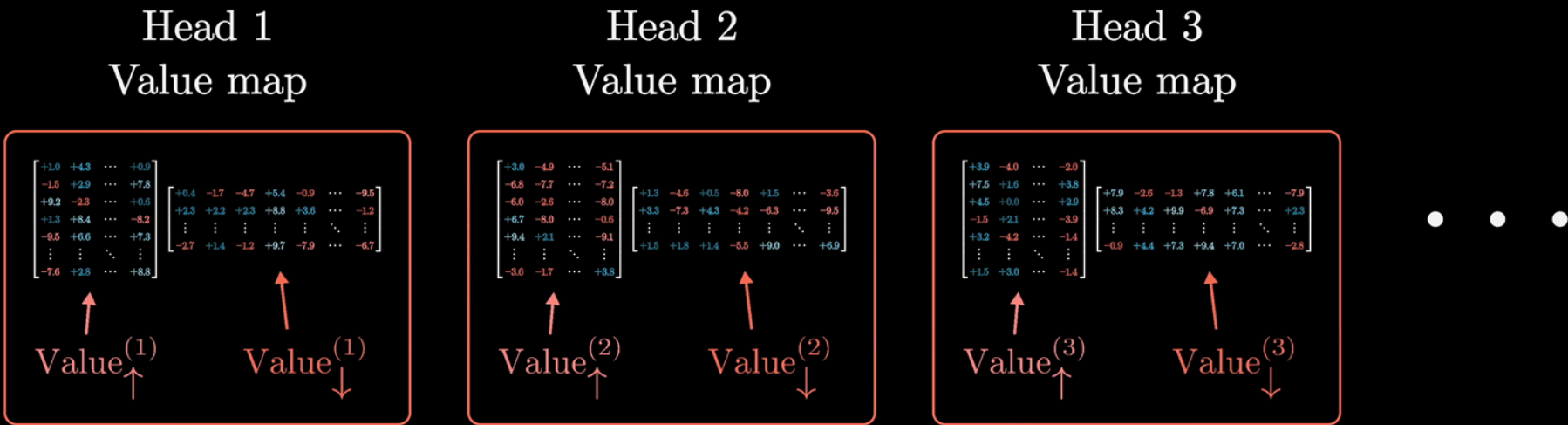
## 7.3 输出矩阵

- 前面提到，值矩阵被拆分为“值↓”和“值↑”



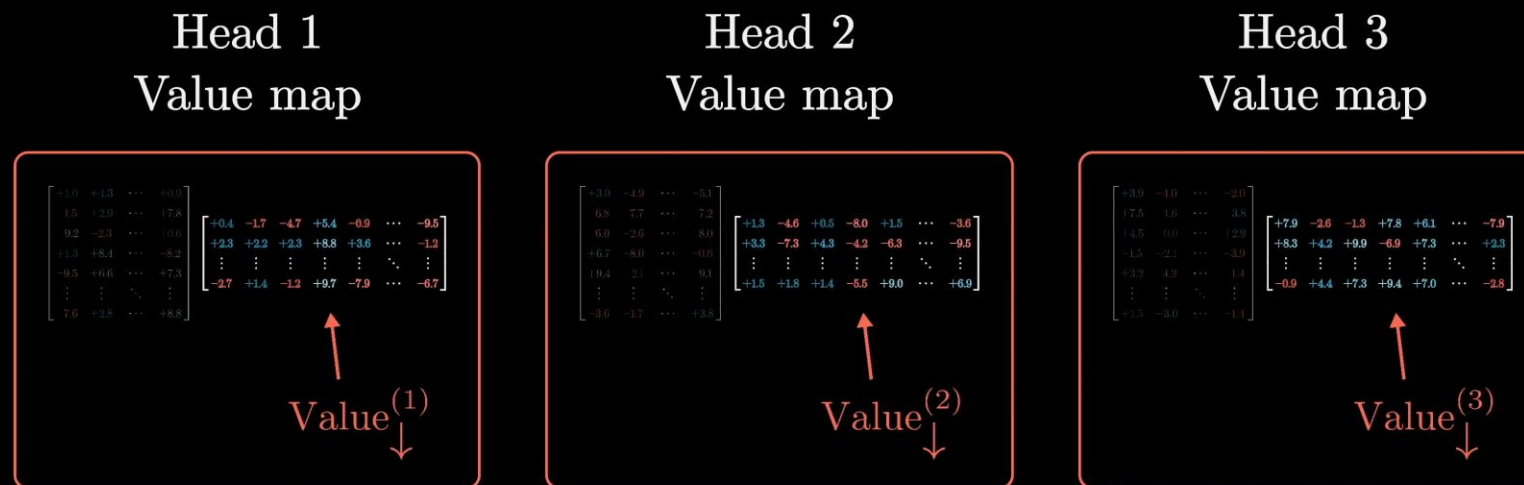
# 7.3 输出矩阵

- 每个头均有一组“值↓”和“值↑”



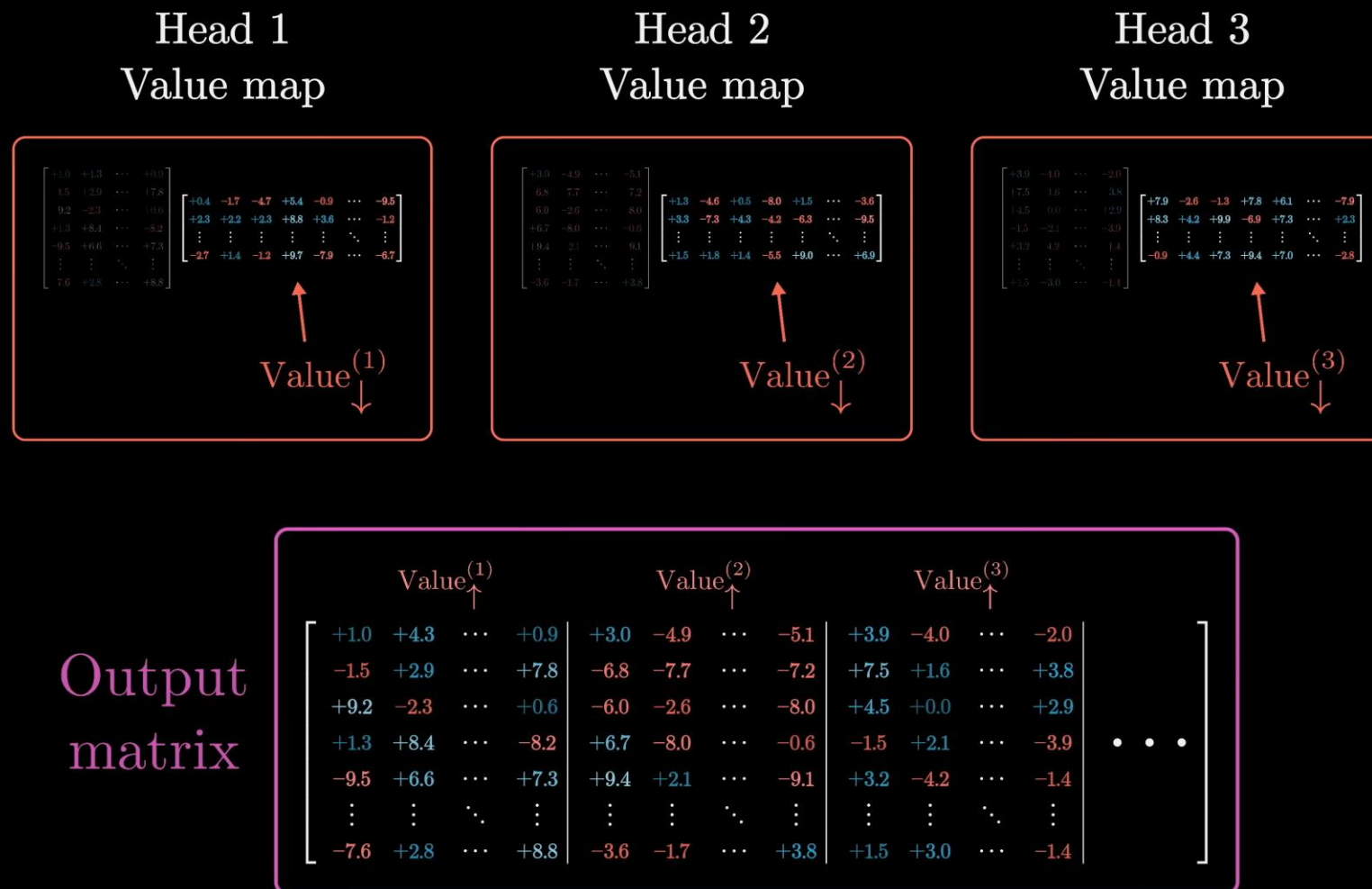
## 7.3 输出矩阵

- 在论文\*中, 每个头的“值↓”抽取出来, 重新组合成一个输出矩阵



## 7.3 输出矩阵

- 在论文\*中, 每个头的“值↓”抽取出来, 重新组合成一个输出矩阵



## 7.3 输出矩阵

- 在论文\*中, 每个头的“值↓”抽取出来, 重新组合成一个输出矩阵

Multi-head attention allows the model to jointly attend to information from different representation subspaces at different positions. With a single attention head, averaging inhibits this.

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

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

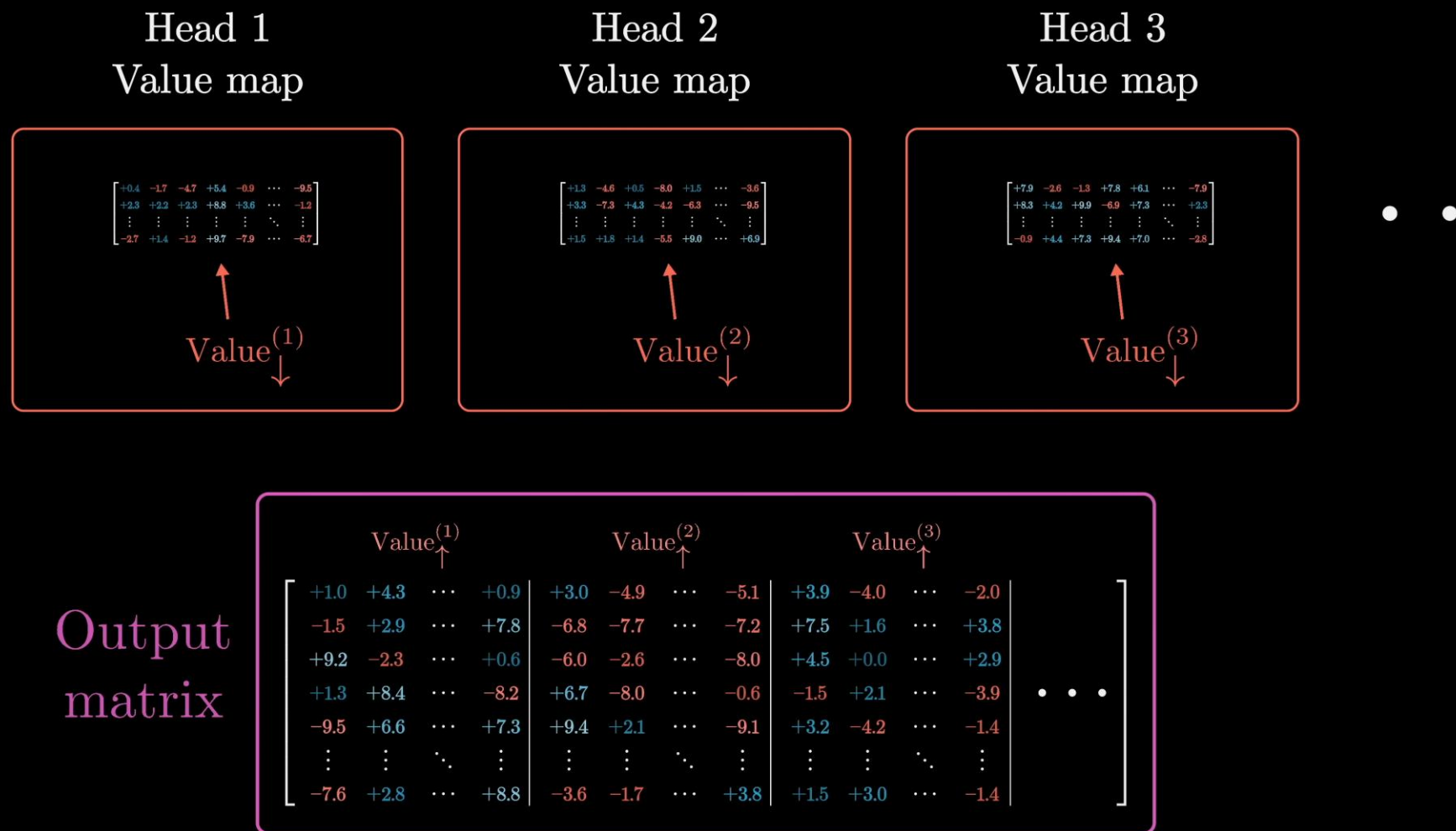
Where the projections are parameter matrices  $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$ ,  $W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$ ,  $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$  and  $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$ .

Output  
matrix

| Value <sup>(1)</sup><br>↑ |      |   |      | Value <sup>(2)</sup><br>↑ |      |   |      | Value <sup>(3)</sup><br>↑ |      |   |      | ⋮ |
|---------------------------|------|---|------|---------------------------|------|---|------|---------------------------|------|---|------|---|
| +1.0                      | +4.3 | ⋯ | +0.9 | +3.0                      | -4.9 | ⋯ | -5.1 | +3.9                      | -4.0 | ⋯ | -2.0 |   |
| -1.5                      | +2.9 | ⋯ | +7.8 | -6.8                      | -7.7 | ⋯ | -7.2 | +7.5                      | +1.6 | ⋯ | +3.8 |   |
| +9.2                      | -2.3 | ⋯ | +0.6 | -6.0                      | -2.6 | ⋯ | -8.0 | +4.5                      | +0.0 | ⋯ | +2.9 |   |
| +1.3                      | +8.4 | ⋯ | -8.2 | +6.7                      | -8.0 | ⋯ | -0.6 | -1.5                      | +2.1 | ⋯ | -3.9 |   |
| -9.5                      | +6.6 | ⋯ | +7.3 | +9.4                      | +2.1 | ⋯ | -9.1 | +3.2                      | -4.2 | ⋯ | -1.4 |   |
| ⋮                         | ⋮    | ⋮ | ⋮    | ⋮                         | ⋮    | ⋮ | ⋮    | ⋮                         | ⋮    | ⋮ | ⋮    |   |
| -7.6                      | +2.8 | ⋯ | +8.8 | -3.6                      | -1.7 | ⋯ | +3.8 | +1.5                      | +3.0 | ⋯ | -1.4 |   |

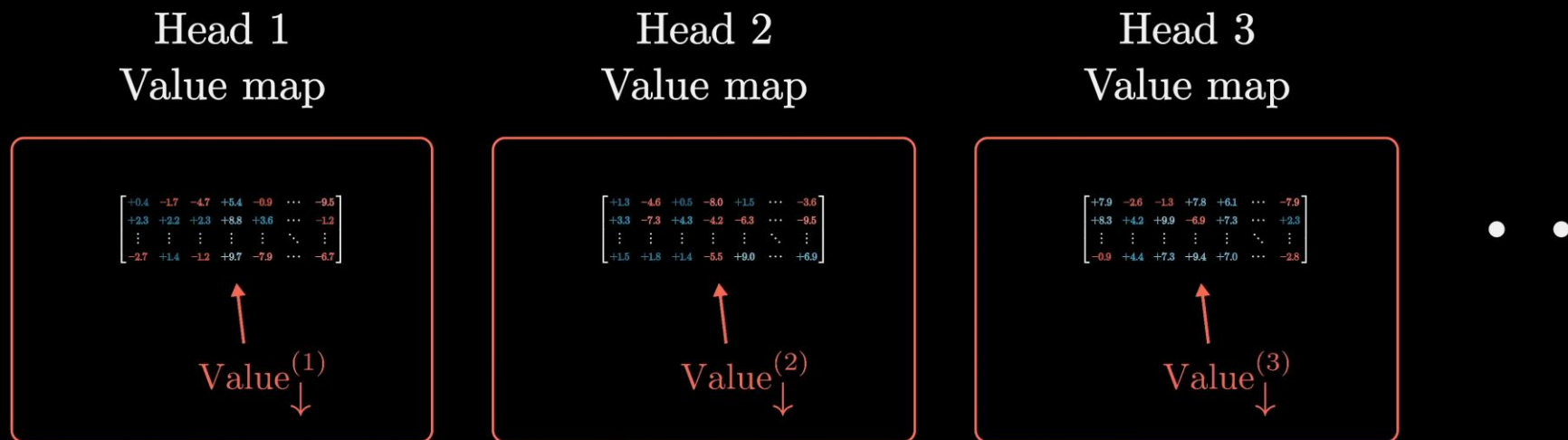
## 7.3 输出矩阵

- “值↑”才是论文\*中的“值”矩阵



## 7.3 输出矩阵

- “值↑”才是论文\*中的“值”矩阵



Multi-head attention allows the model to jointly attend to information from different representation subspaces at different positions. With a single attention head, averaging inhibits this.

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

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

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  $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$ .

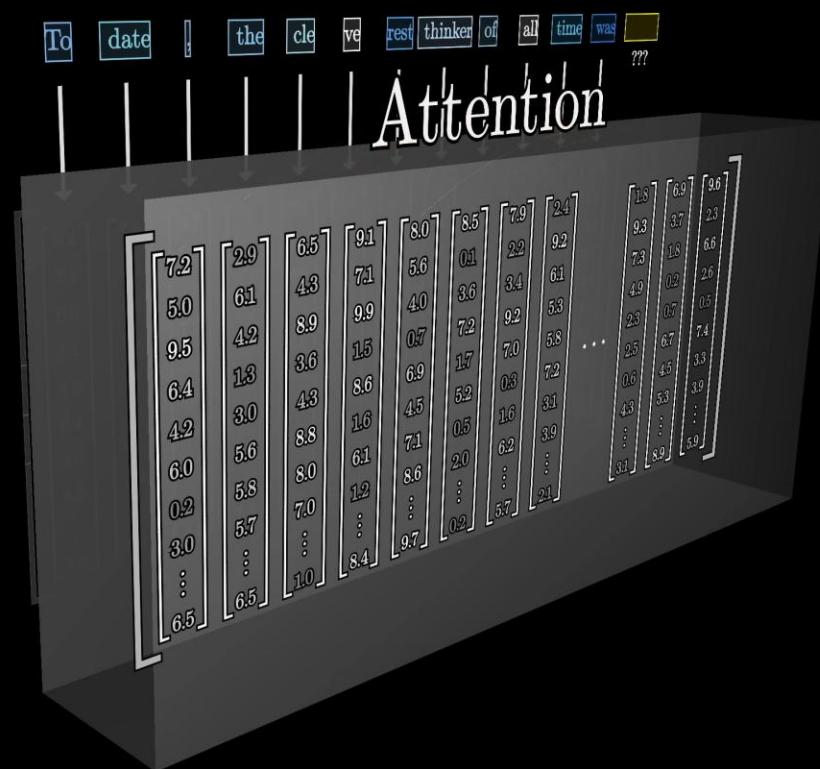
Vaswani, A. "Attention is all you need." *Advances in Neural Information Processing Systems* (2017).



1. 回顾
2. 启发示例：注意力的作用
3. 注意力机制：查询（Query）、键（Key）
4. 掩码（Masking）
5. 注意力机制：值（Value）
6. 参数数量
7. 多头注意力
8. Transformer的深度

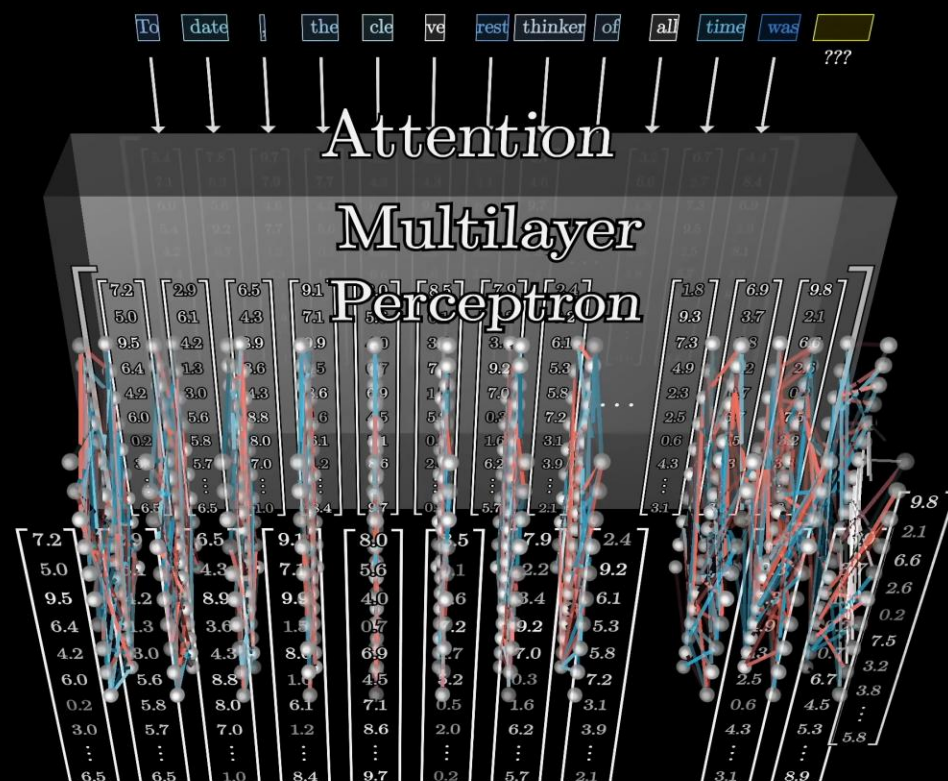
## 8.1 深度网络

- Transformer不止有一个注意力块



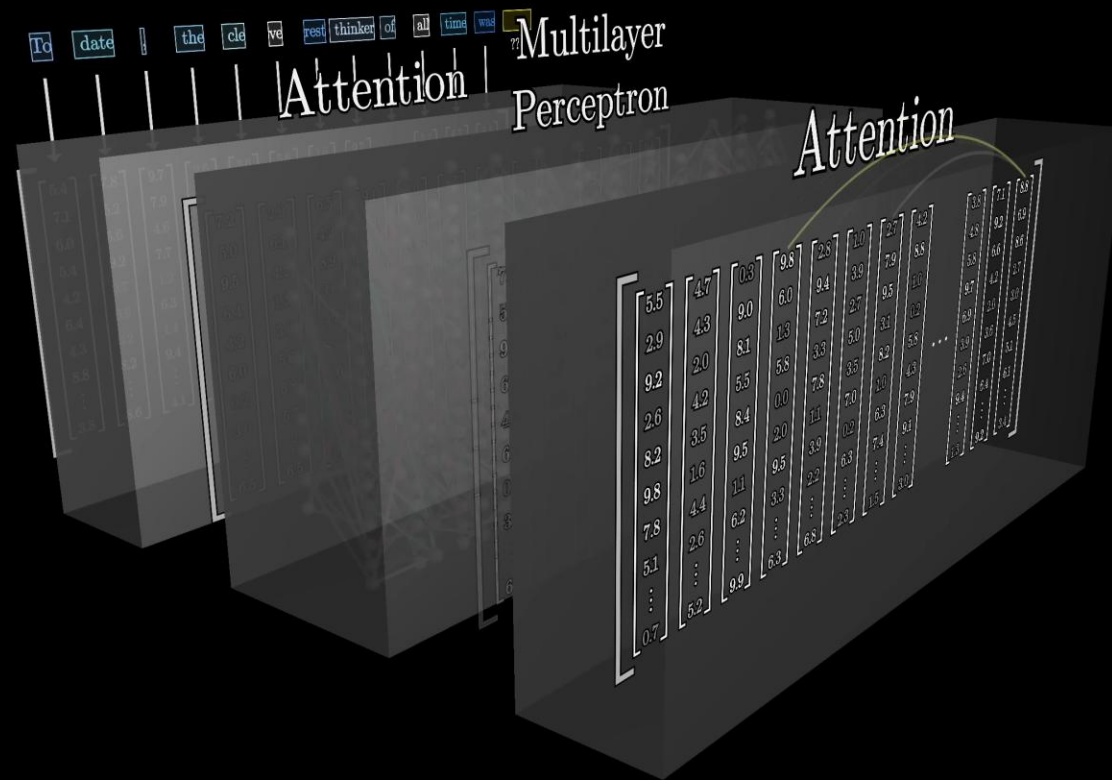
## 8.1 深度网络

- 注意力后面接着多层感知机块



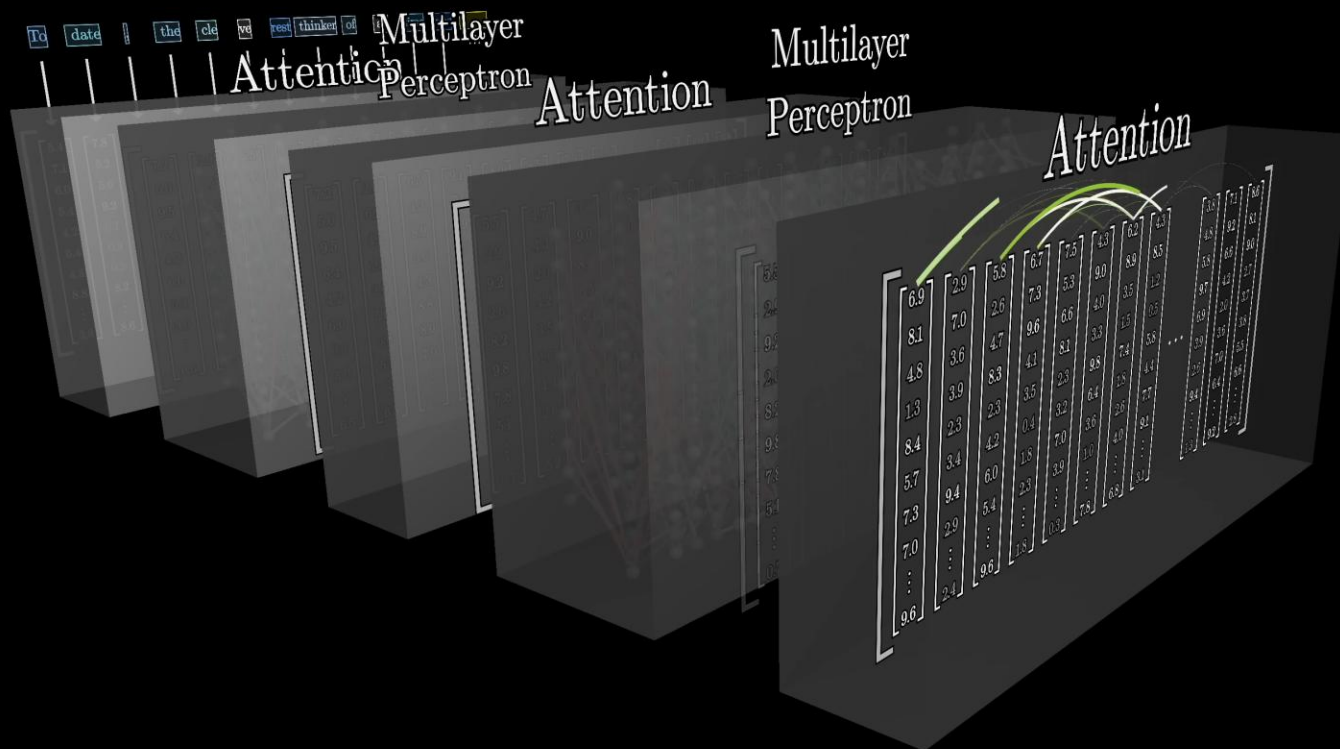
## 8.1 深度网络

- (注意力 + 多层感知机块)  $\times L$



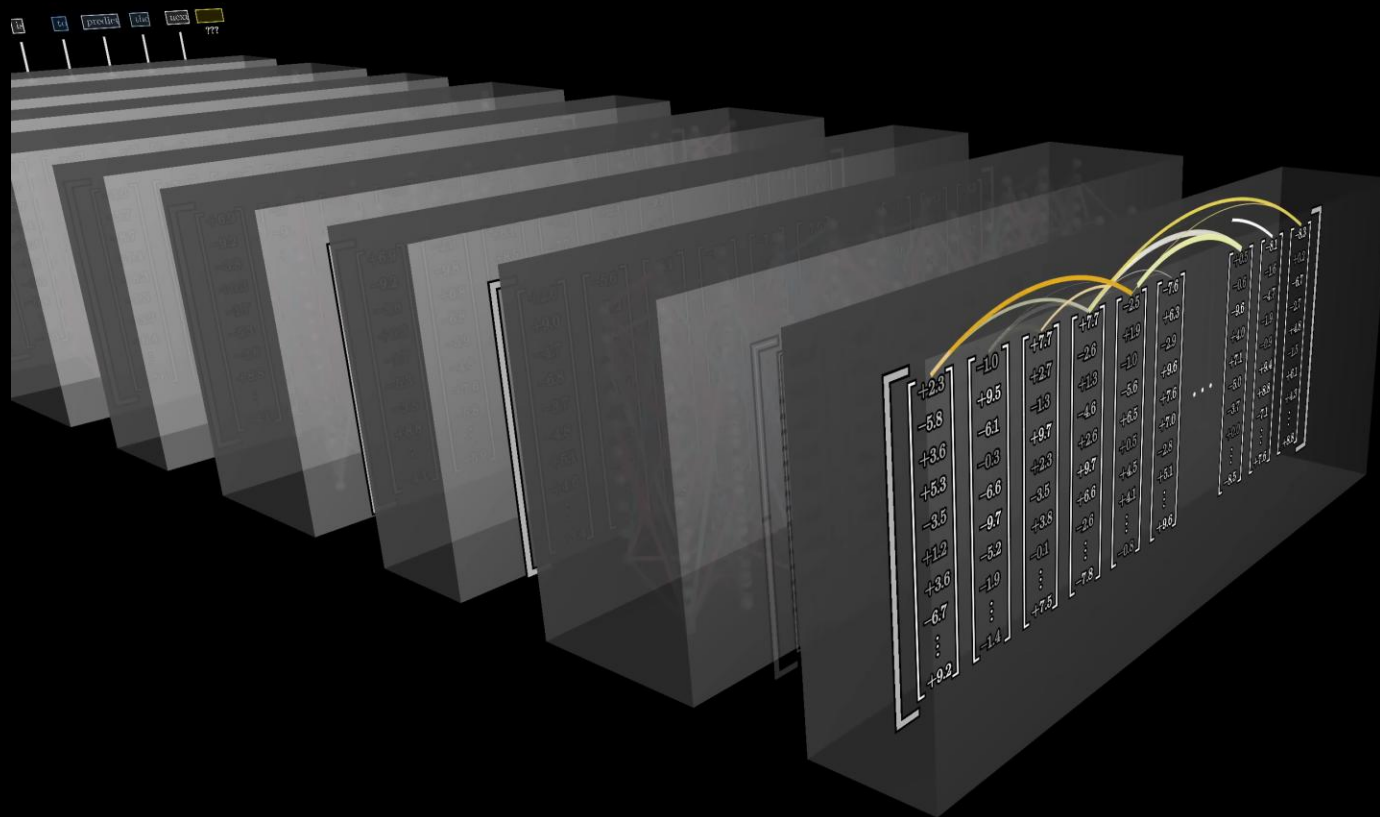
## 8.1 深度网络

- 嵌入向量经过注意力块的更新后，还能够进一步在后面的块中更新



## 8.1 深度网络

- 从而吸收文本中更深刻的含义

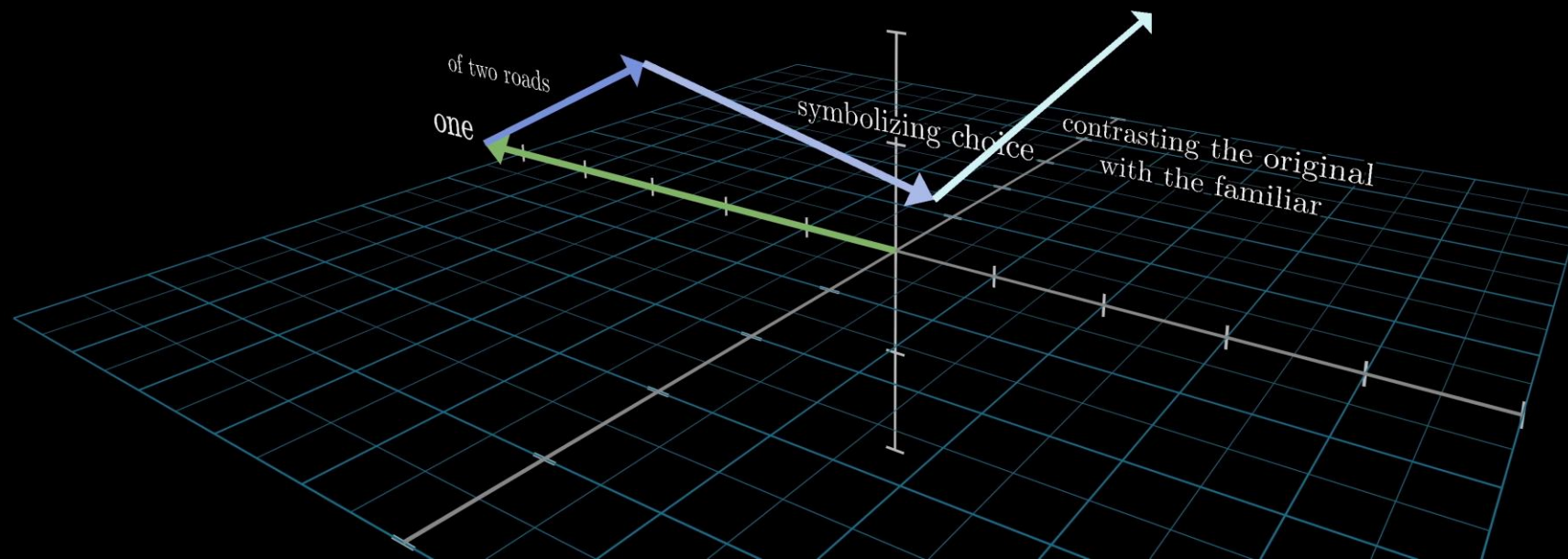


## 8.1 深度网络

- 在最深层的块中，理解更隐晦、更具诗意的文学修养
- 或更高级、更抽象的科学原理等等

Two roads diverged in a wood, and I—  
I took the **one** less traveled by,

一片树林里分出两条路，  
而我选择了人迹更少的一条...  
——普希金《未选择的路》





## 8.2 参数数量

- 层数的增加，参数的数量增加96倍

Total weights: 175,181,291,520  
Organized into 27,938 matrices



|                 |                                                                                                         |
|-----------------|---------------------------------------------------------------------------------------------------------|
| Embedding       | $\overset{12,288}{d\_embed} * \overset{50,257}{n\_vocab} = 617,558,016$                                 |
| Key             | $\overset{128}{d\_query} * \overset{12,288}{d\_embed} * \overset{96}{n\_heads} = 150,994,944$ per layer |
| Query           | $\overset{128}{d\_query} * \overset{12,288}{d\_embed} * \overset{96}{n\_heads} = 150,994,944$ per layer |
| Value           | $\overset{128}{d\_value} * \overset{12,288}{d\_embed} * \overset{96}{n\_heads} = 150,994,944$ per layer |
| Output          | $\overset{12,288}{d\_embed} * \overset{128}{d\_value} * \overset{96}{n\_heads} = 150,994,944$ per layer |
| Up-projection   |                                                                                                         |
| Down-projection |                                                                                                         |
| Unembedding     | $\overset{50,257}{n\_vocab} * \overset{12,288}{d\_embed} = 617,558,016$                                 |





## 8.2 参数数量

- 层数的增加，参数的数量增加96倍

Total weights: 175,181,291,520  
Organized into 27,938 matrices



|                 |                                                                                                                            |
|-----------------|----------------------------------------------------------------------------------------------------------------------------|
| Embedding       | $\overset{12,288}{d\_embed} * \overset{50,257}{n\_vocab} = 617,558,016$                                                    |
| Key             | $\overset{128}{d\_query} * \overset{12,288}{d\_embed} * \overset{96}{n\_heads} * \overset{96}{n\_layers} = 14,495,514,624$ |
| Query           | $\overset{128}{d\_query} * \overset{12,288}{d\_embed} * \overset{96}{n\_heads} * \overset{96}{n\_layers} = 14,495,514,624$ |
| Value           | $\overset{128}{d\_value} * \overset{12,288}{d\_embed} * \overset{96}{n\_heads} * \overset{96}{n\_layers} = 14,495,514,624$ |
| Output          | $\overset{12,288}{d\_embed} * \overset{128}{d\_value} * \overset{96}{n\_heads} * \overset{96}{n\_layers} = 14,495,514,624$ |
| Up-projection   |                                                                                                                            |
| Down-projection |                                                                                                                            |
| Unembedding     | $\overset{50,257}{n\_vocab} * \overset{12,288}{d\_embed} = 617,558,016$                                                    |

## 8.2 参数数量

- 注意力机制参数总量约为60B

Total weights: 175,181,291,520  
Organized into 27,938 matrices



|                 |                                                                                                                            |
|-----------------|----------------------------------------------------------------------------------------------------------------------------|
| Embedding       | $\overset{12,288}{d\_embed} * \overset{50,257}{n\_vocab} = 617,558,016$                                                    |
| Key             | $\overset{128}{d\_query} * \overset{12,288}{d\_embed} * \overset{96}{n\_heads} * \overset{96}{n\_layers} = 14,495,514,624$ |
| Query           | $\overset{128}{d\_query} * \overset{12,288}{d\_embed} * \overset{96}{n\_heads} * \overset{96}{n\_layers} = 14,495,514,624$ |
| Value           | $\overset{128}{d\_value} * \overset{12,288}{d\_embed} * \overset{96}{n\_heads} * \overset{96}{n\_layers} = 14,495,514,624$ |
| Output          | $\overset{12,288}{d\_embed} * \overset{128}{d\_value} * \overset{96}{n\_heads} * \overset{96}{n\_layers} = 14,495,514,624$ |
| Up-projection   | 57,982,058,496                                                                                                             |
| Down-projection |                                                                                                                            |
| Unembedding     | $\overset{50,257}{n\_vocab} * \overset{12,288}{d\_embed} = 617,558,016$                                                    |

## 8.2 参数数量

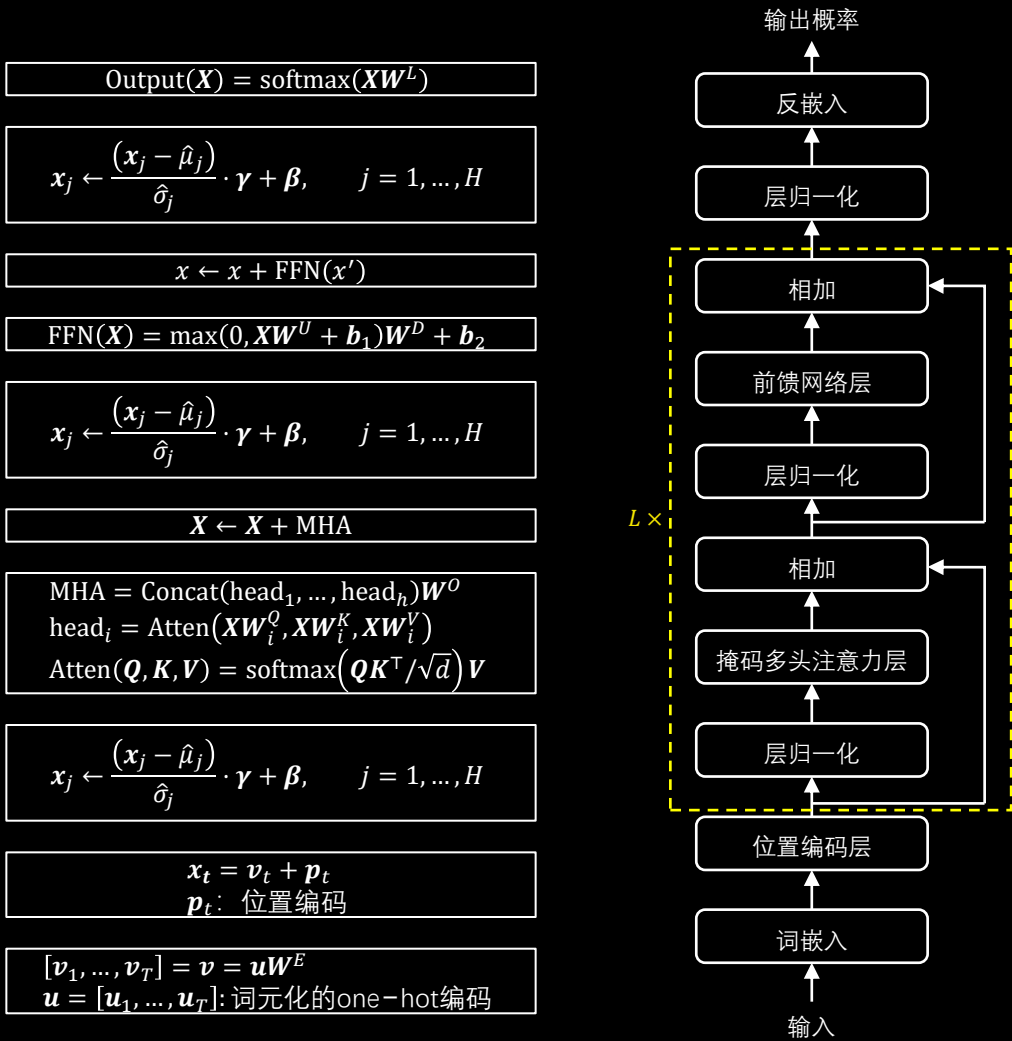
- 注意力机制参数总量的1/3

About 1/3 of What  
Attention is ~~All~~ You Need



## 8.2 参数数量

### GPT-3模型构架与参数总量



$$W^E \in \mathbb{R}^{V \times H}, W^L \in \mathbb{R}^{V \times H}$$
$$\gamma, \beta \in \mathbb{R}^H$$
$$W_i^Q, W_i^K, W_i^V, \text{head}_i \in \mathbb{R}^{H \times H_i}$$
$$\text{MHA} \in \mathbb{R}^{H \times H}$$
$$W^U \in \mathbb{R}^{H \times H'}, W^D \in \mathbb{R}^{H' \times H}, b_1 \in \mathbb{R}^H, b_2 \in \mathbb{R}^H$$

ganized into 27,938 matrices

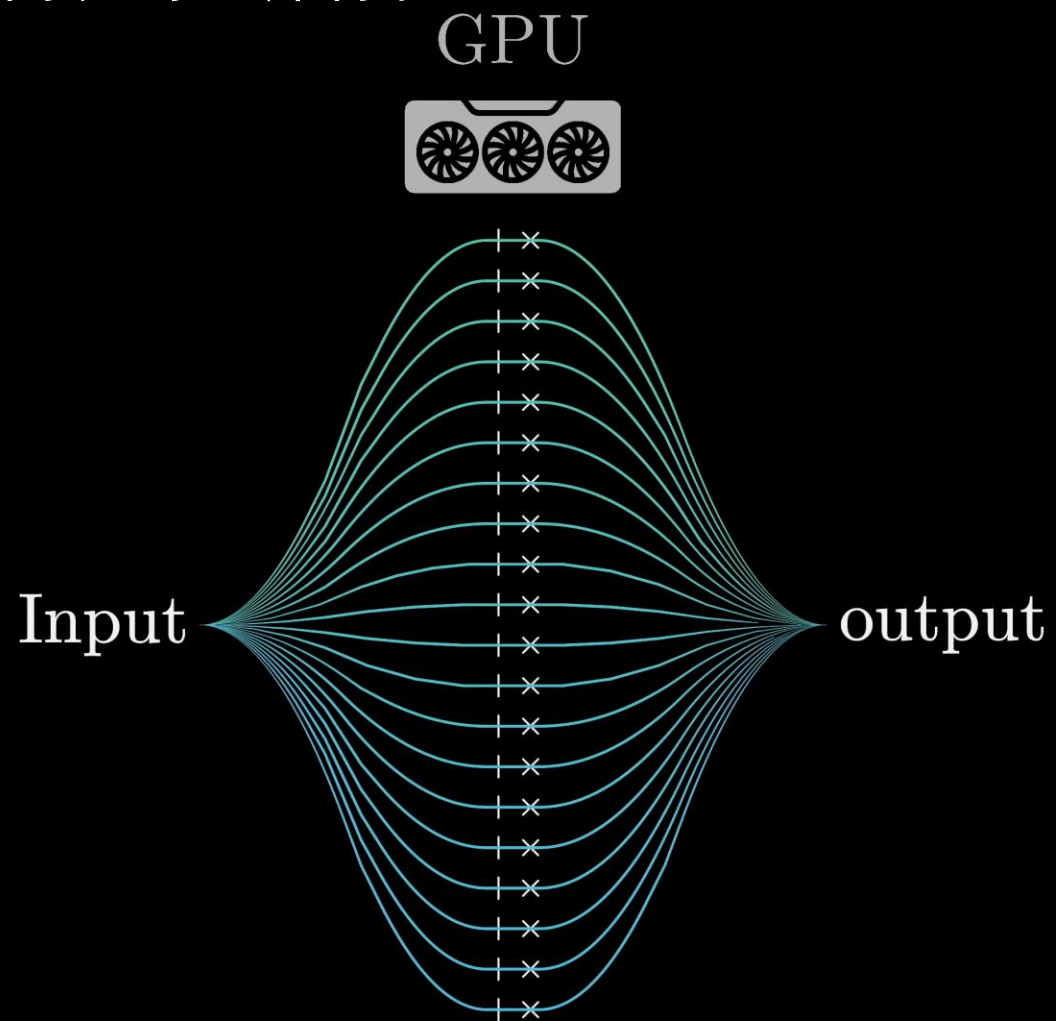
 GPT-3

Total weights:  
175,181,291,520

|                       |                                                  |                    |                  |
|-----------------------|--------------------------------------------------|--------------------|------------------|
| $W^E$ Embedding       | 12,288<br>d_embed * n_vocab                      | 50,257             | = 617,558,016    |
| $W_i^K$ Key           | 128<br>d_query * d_embed * n_heads * n_layers    | 12,288<br>96<br>96 | = 14,495,514,624 |
| $W_i^Q$ Query         | 128<br>d_query * d_embed * n_heads * n_layers    | 12,288<br>96<br>96 | = 14,495,514,624 |
| $W_i^V$ Value         | 128<br>d_value * d_embed * n_heads * n_layers    | 12,288<br>96<br>96 | = 14,495,514,624 |
| $W^O$ Output          | 12,288<br>d_embed * d_value * n_heads * n_layers | 128<br>96<br>96    | = 14,495,514,624 |
| $W^U$ Up-projection   | 49,152<br>n_neurons * d_embed * n_layers         | 12,288<br>96       | = 57,982,058,496 |
| $W^D$ Down-projection | 12,288<br>d_embed * n_neurons * n_layers         | 49,152<br>96       | = 57,982,058,496 |
| $W^L$ Unembedding     | 50,257<br>n_vocab * d_embed                      | 12,288             | = 617,558,016    |

## 8.3 总结

- 注意力机制的成功源于高度的可并行化

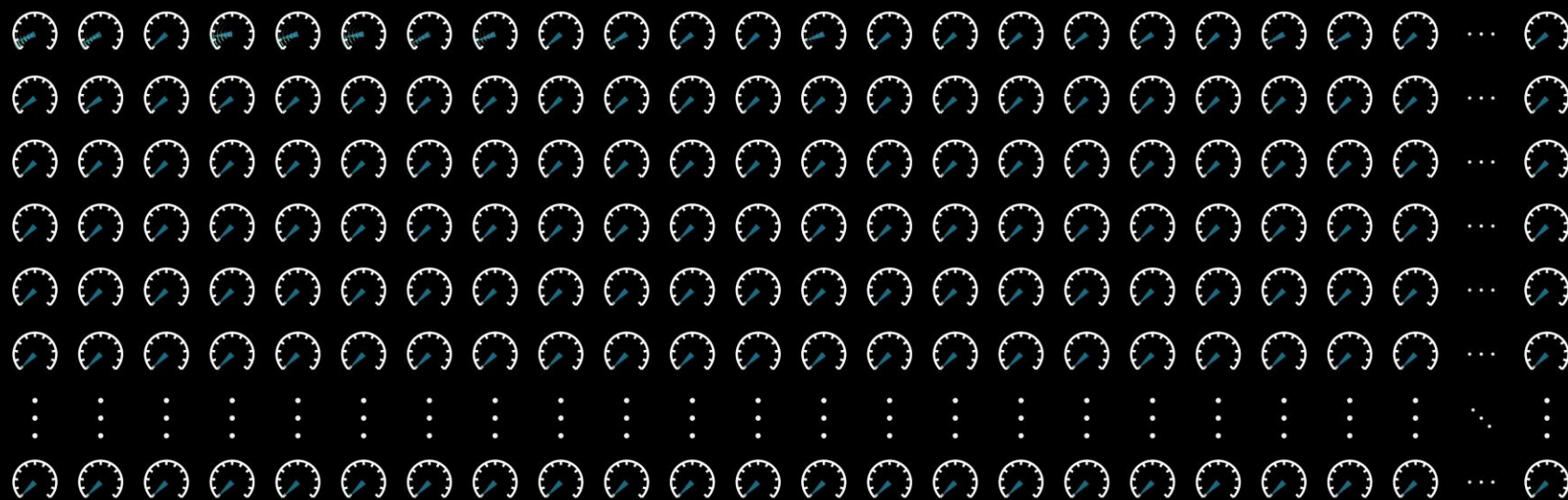


## 8.3 总结

- 使得大模型可以做到足够大——参数数量大
- 大模型的成功源于参数数量爆发带来的“智能涌现”

# Scale matters!

Total parameters: 175,181,291,520



# 8.3 总结

- GPT-3模型与经典Transformer的构架对比

