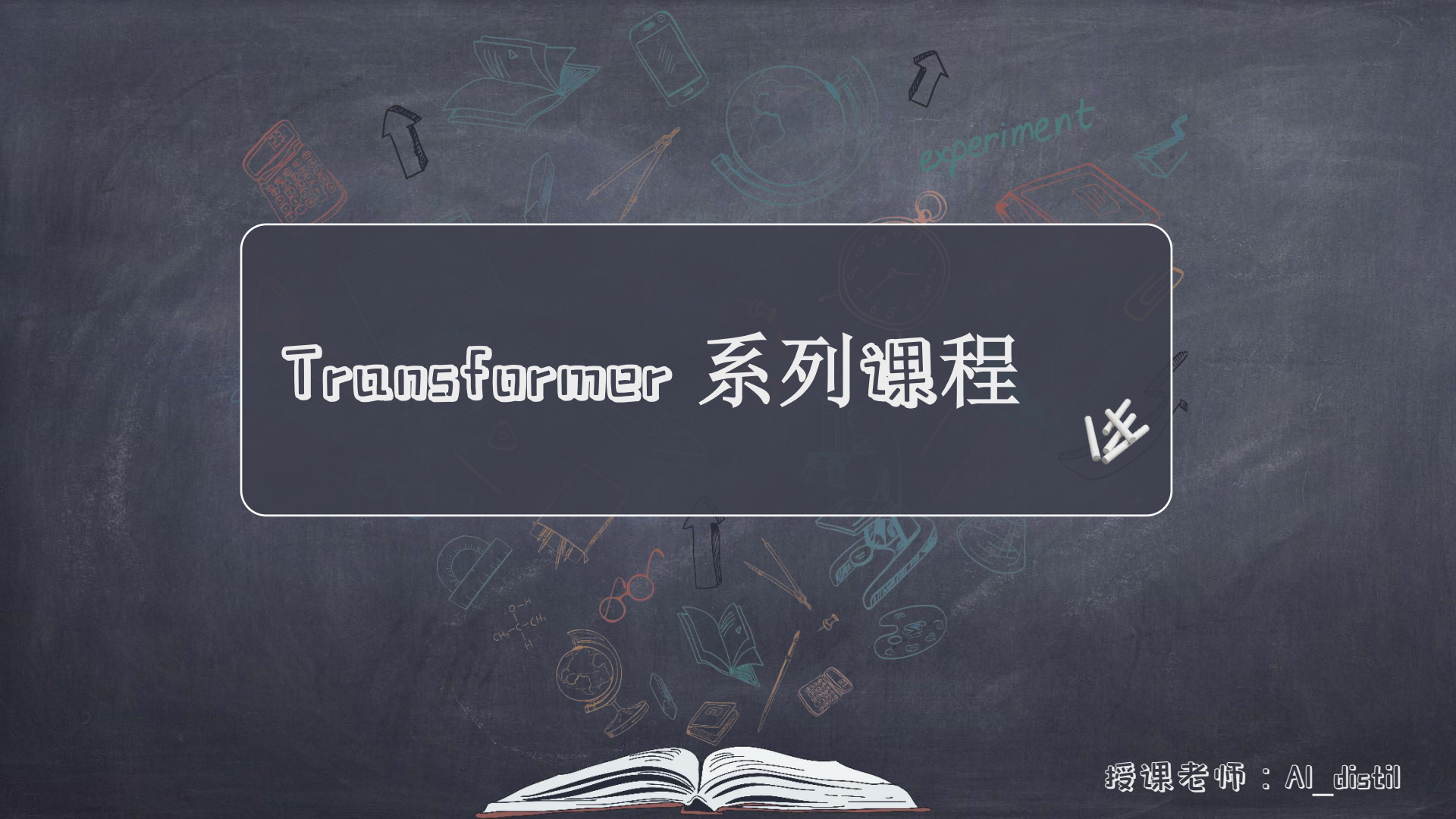
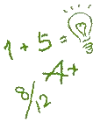


Transformer 系列课程

授课老师：AI_distil

目录
CONTENTS



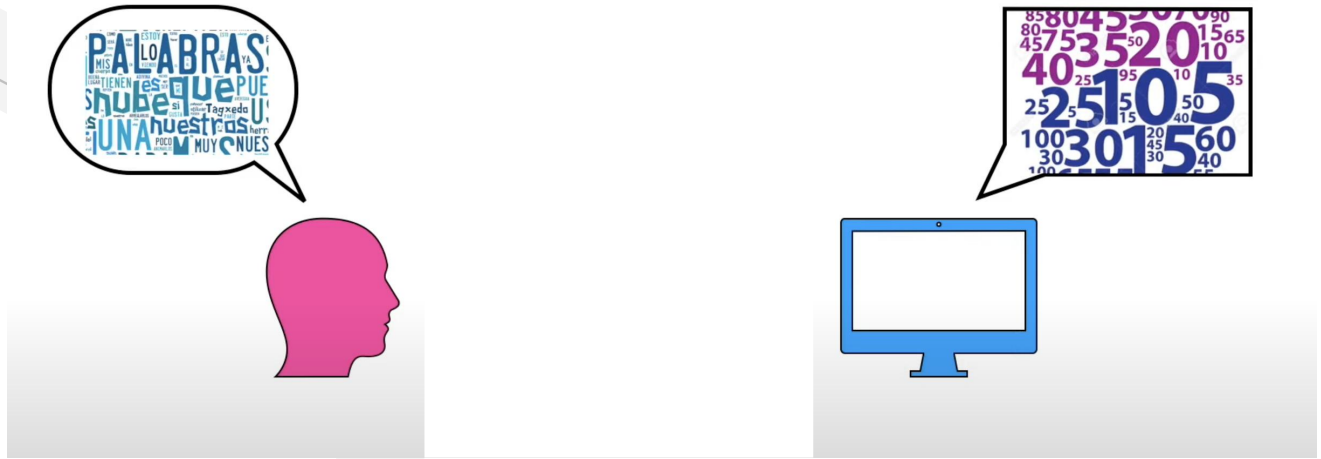
PART ONE

注意力机制

注意力机制

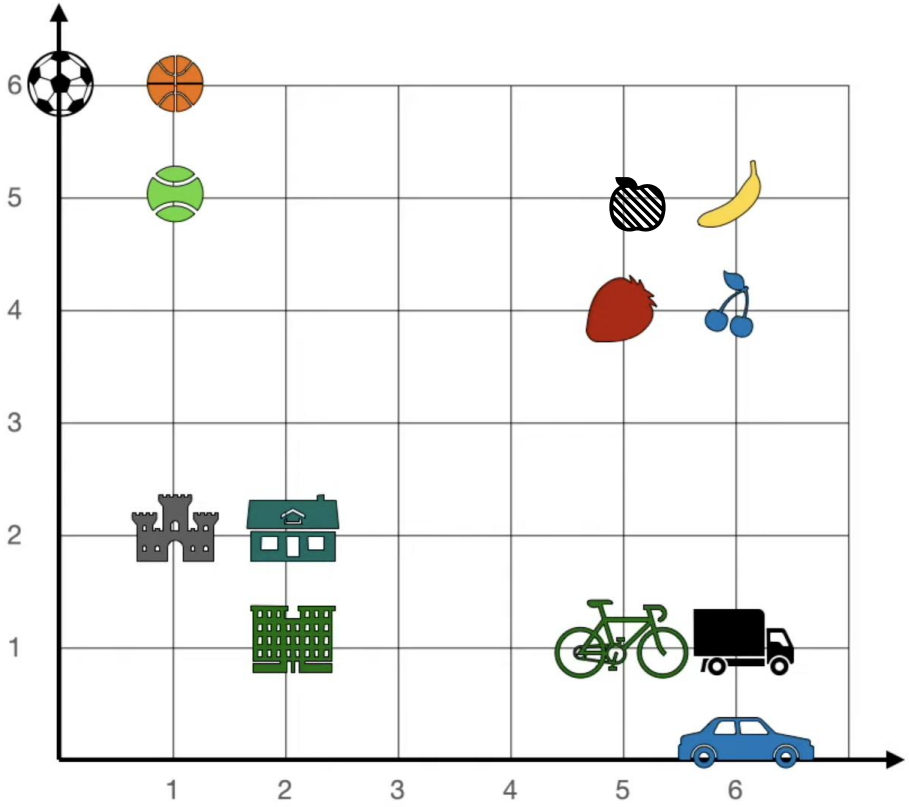
embeddings

embeddings 即词嵌入，
是NLP领域最重要部分



文本变成数字，embeddings 越好模型就会越好

embeddings



?

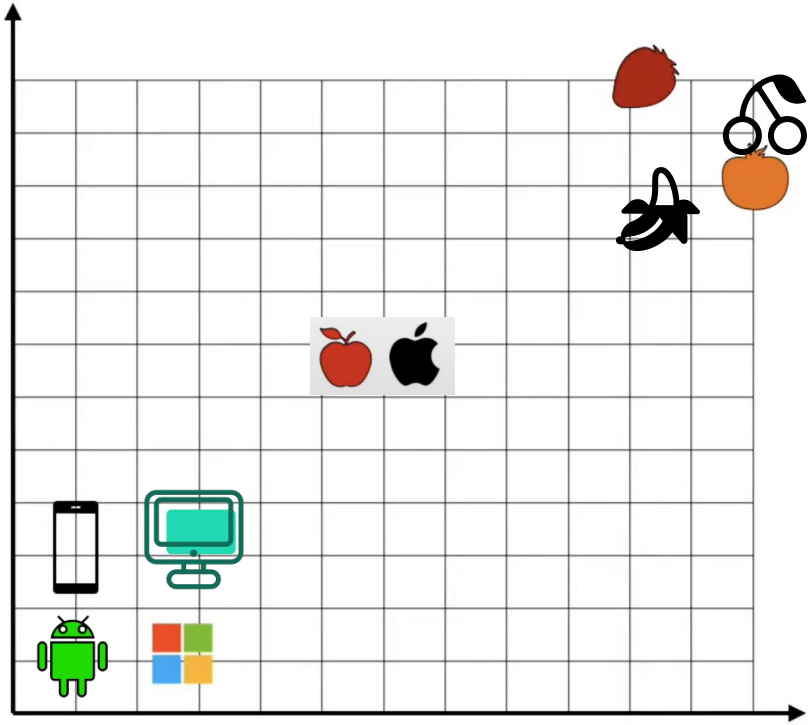
相似的单词会被赋予相似的数字

Word	Numbers	
Apple	5	5
Soccer	0	6
House	2	2
Car	6	0

Word	Numbers			
A	-0.82	-0.32	...	0.23
Aardvark	0.419	1.28	...	-0.06
...			...	
Zygote	-0.74	-1.02	...	1.35

4096

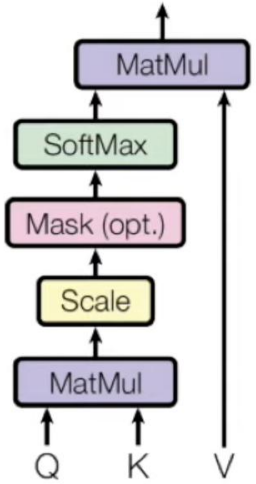
problems



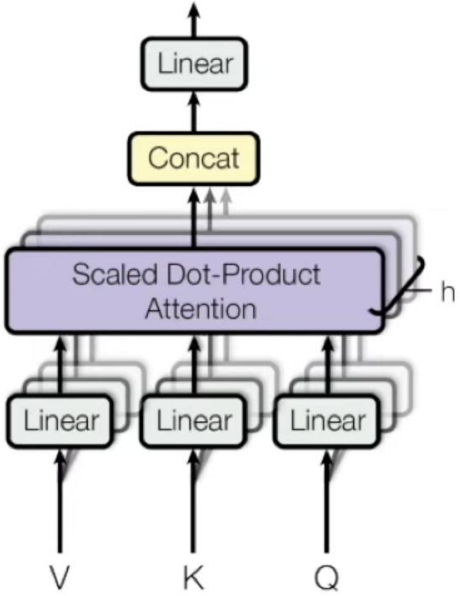
Apple ?

Attention

Scaled Dot-Product Attention



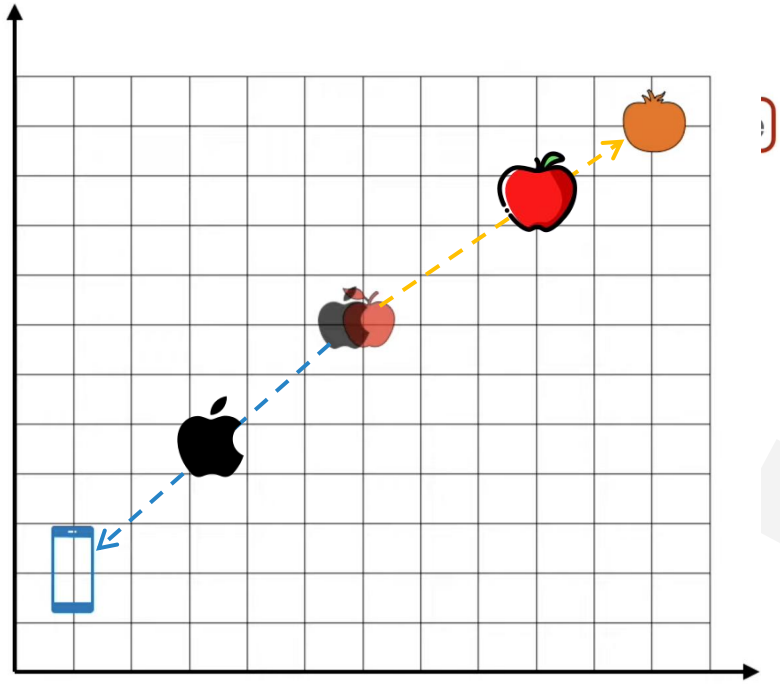
Multi-Head Attention

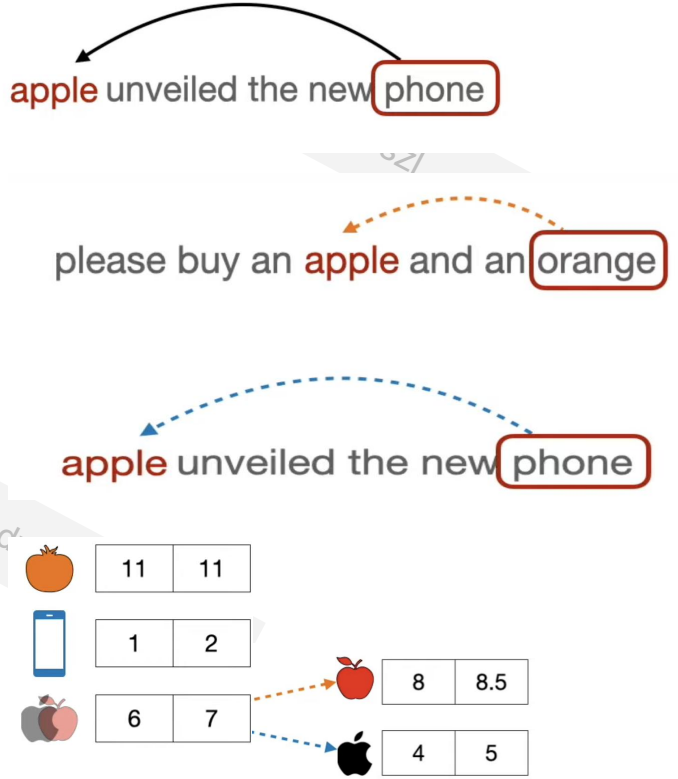


$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_\text{h})W^O$$
$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

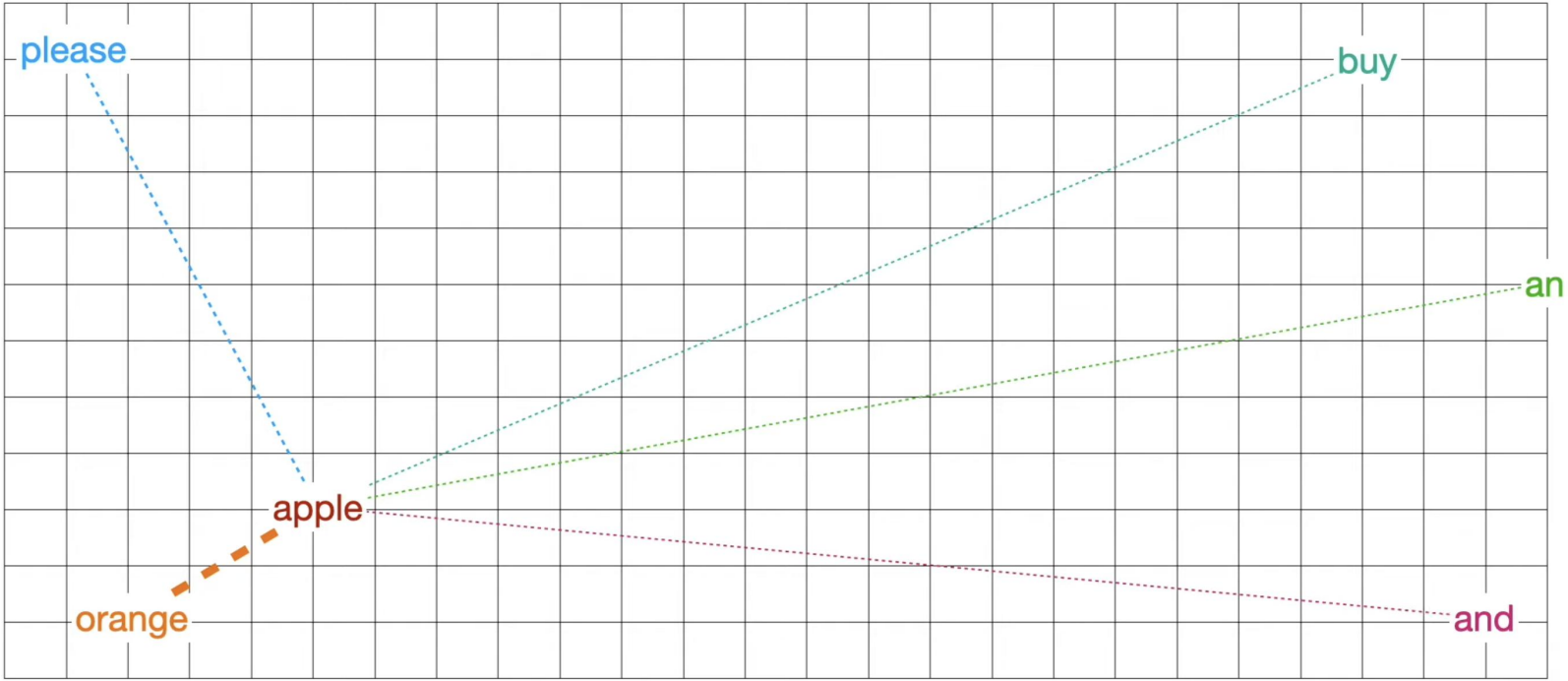
Attention



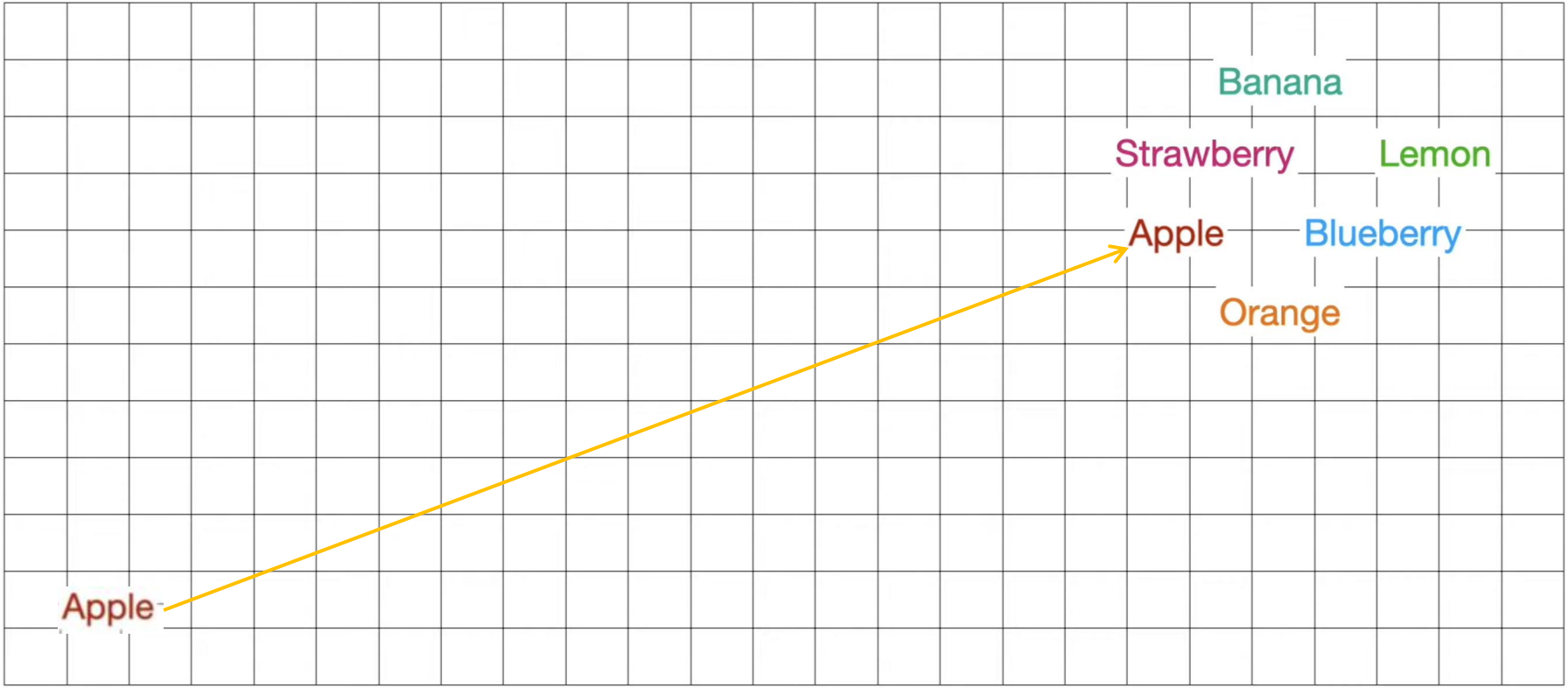

apple unveiled the new phone
please buy an apple and an orange
apple unveiled the new phone
11
11
1
2
6
7
8
8.5
4
5

Attention

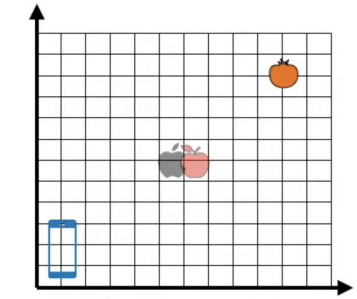

please buy an apple and an orange

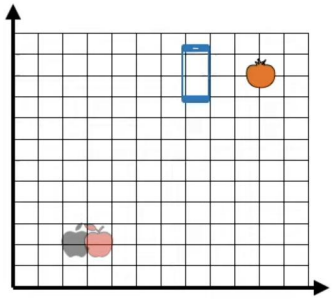

please
buy
an
apple
orange
and

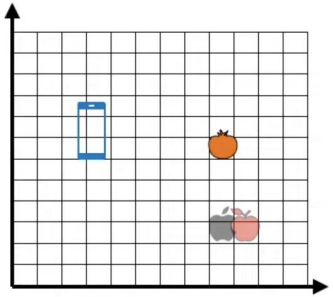
Attention

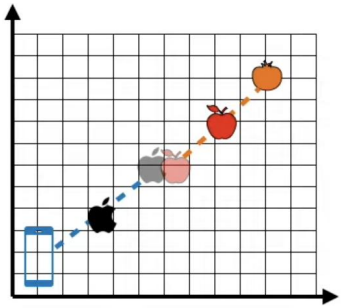


Multi-head attentiontention

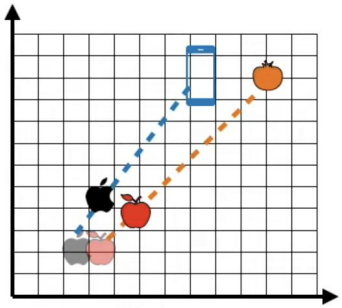




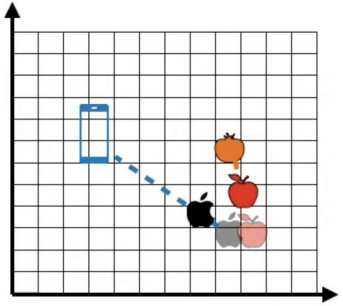




Good

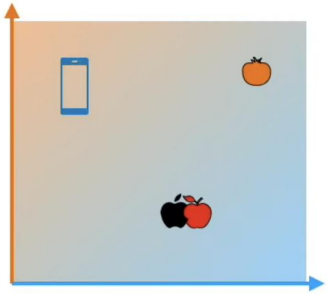


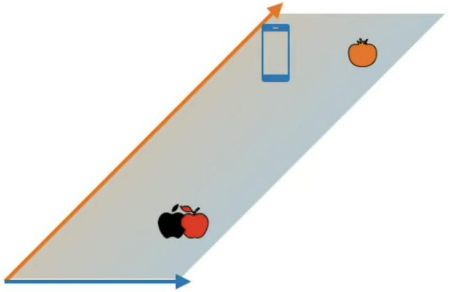
Bad

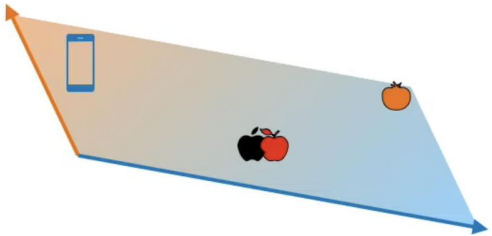


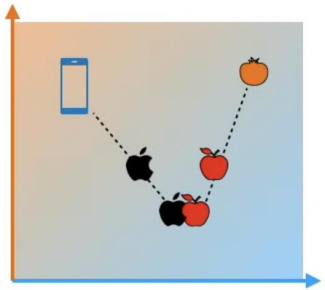
So-so

Linear transformatiuons

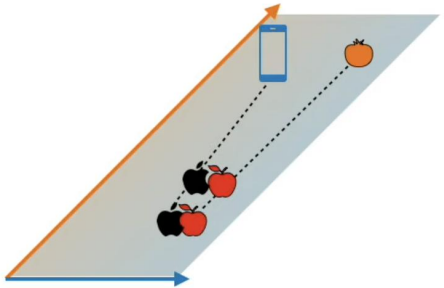




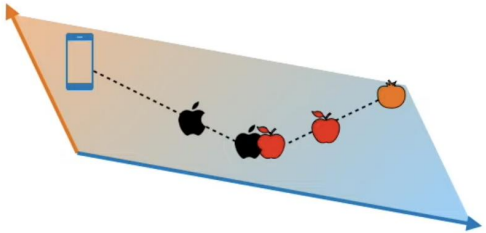




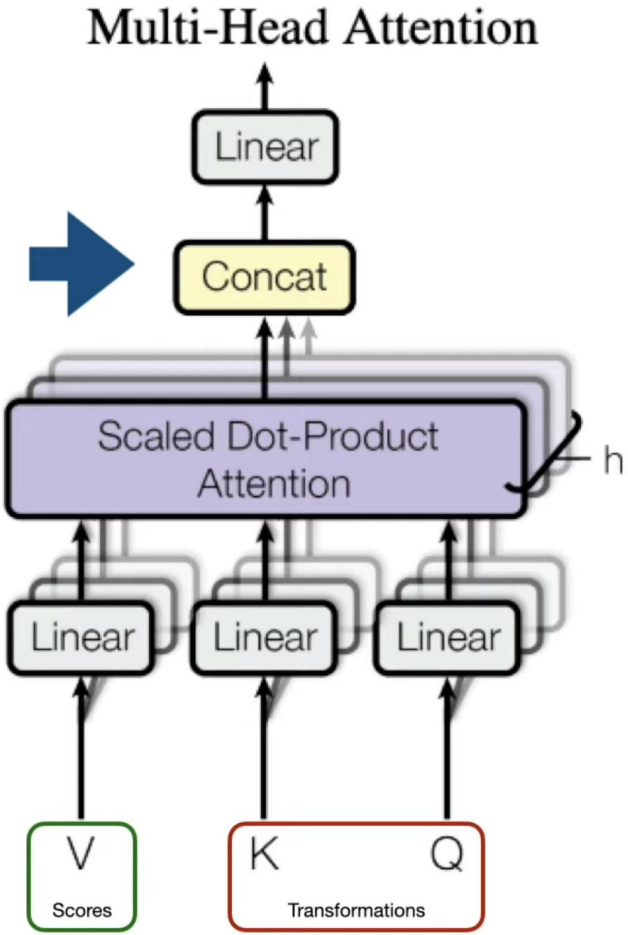
Okay

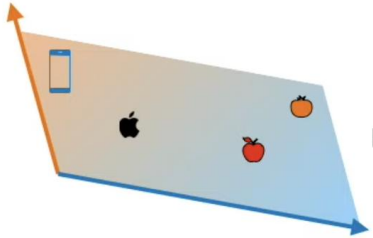


Bad

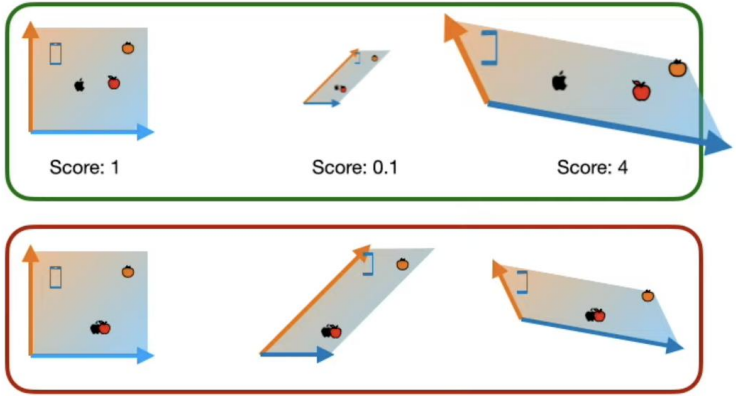


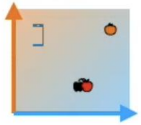
Good





Embedding that has context

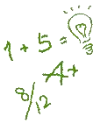




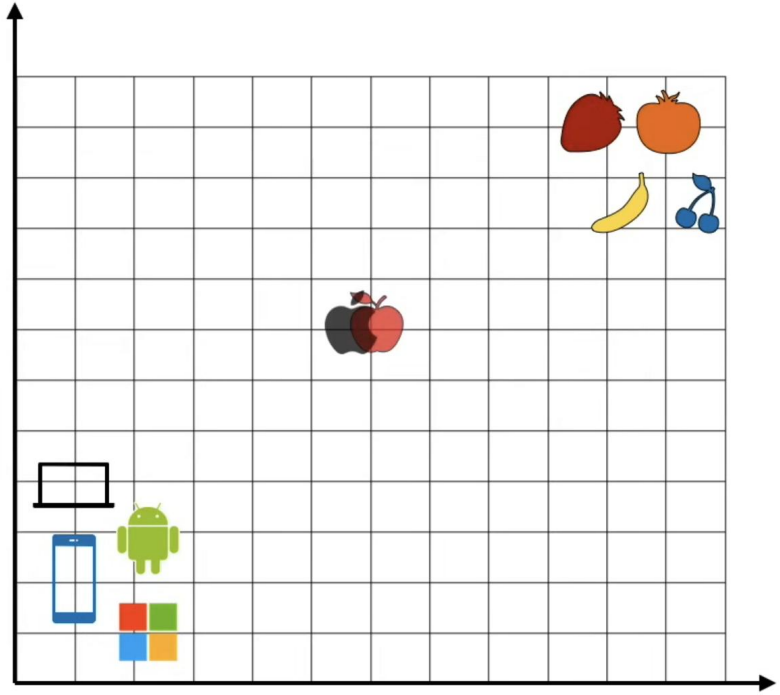
please buy an apple and an orange
apple unveiled the new phone

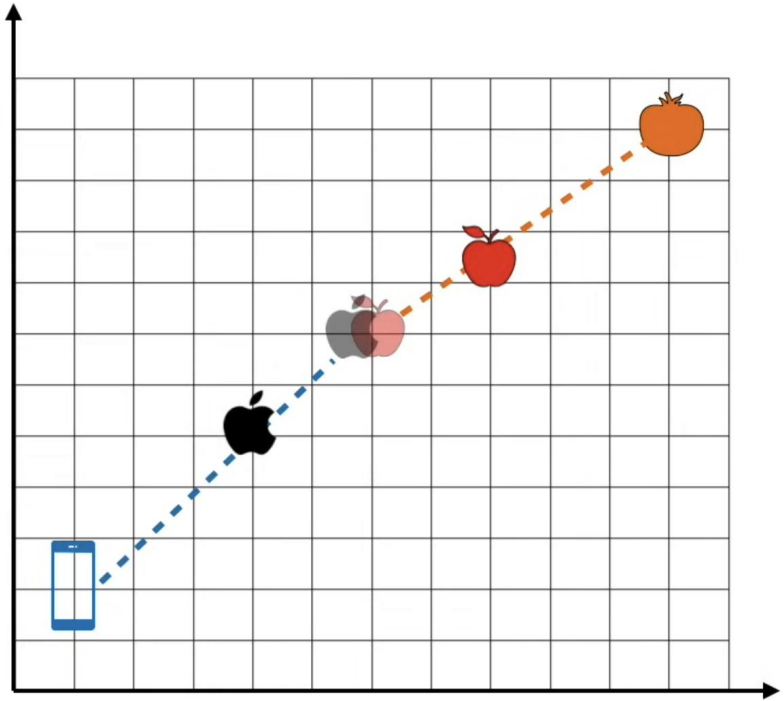
PART two

Q、K、V矩阵

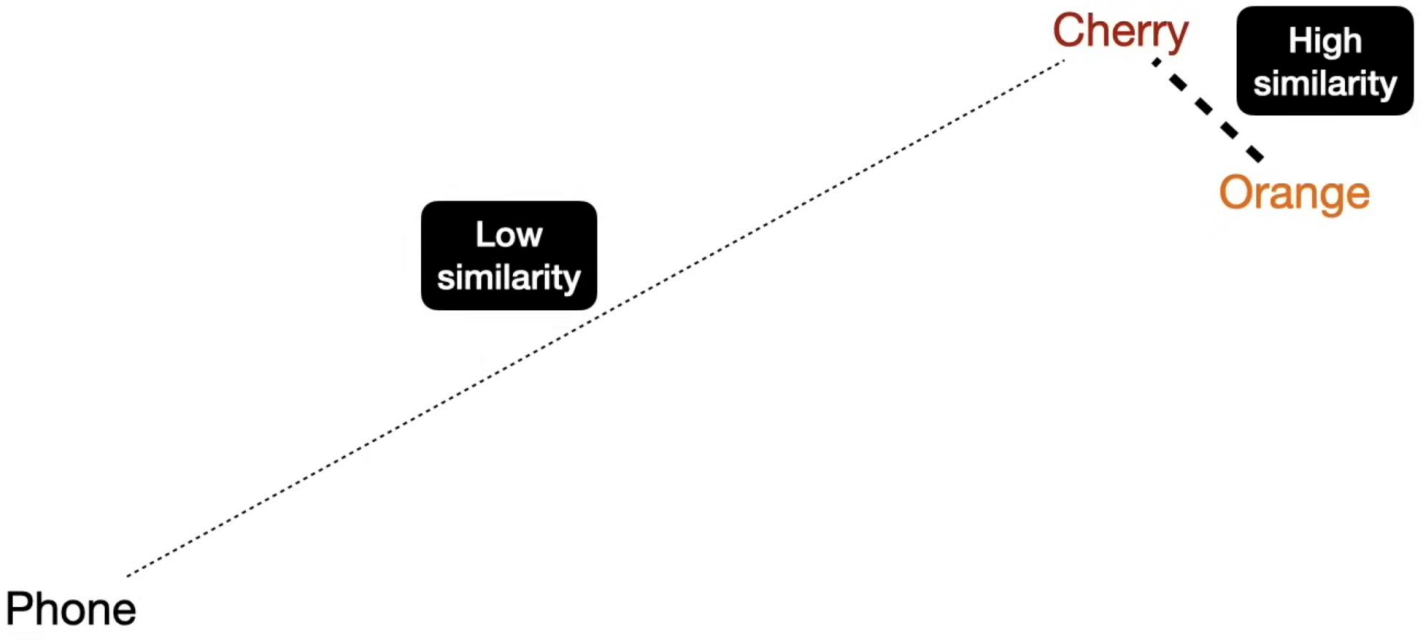


词义的相似性

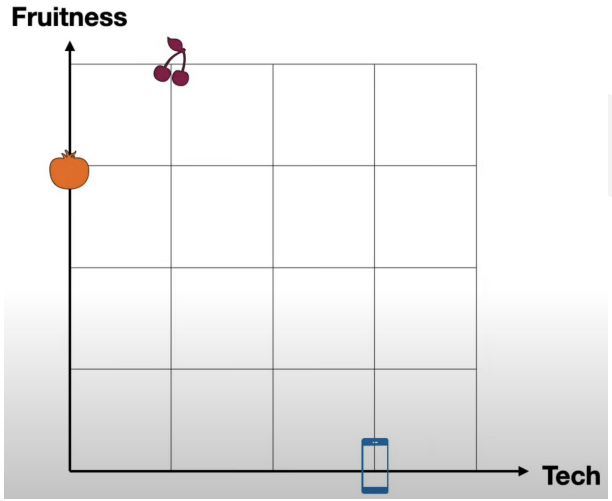


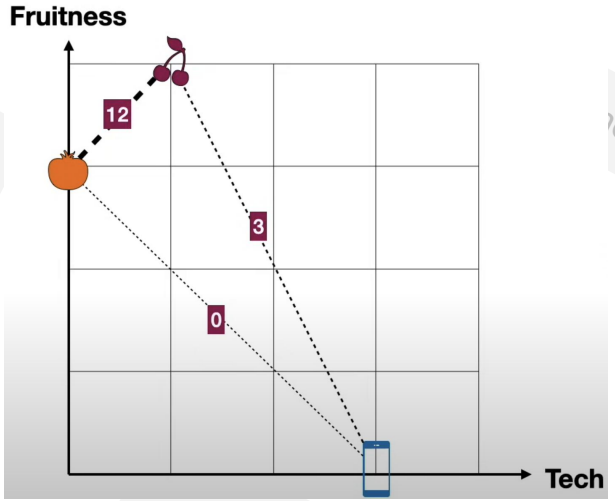


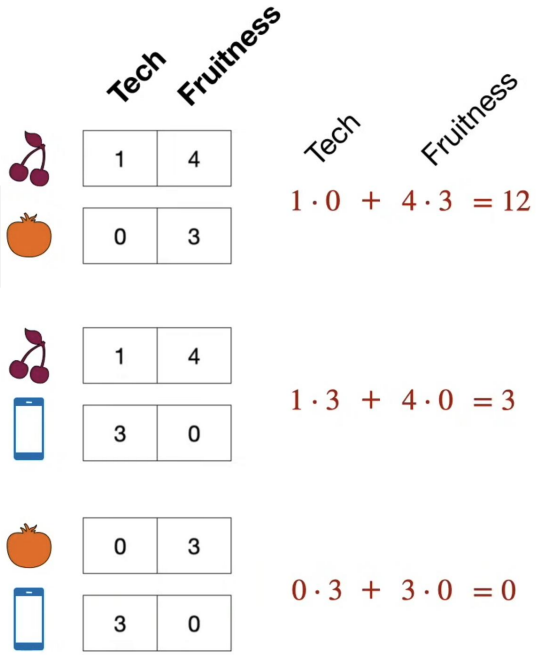
similarity



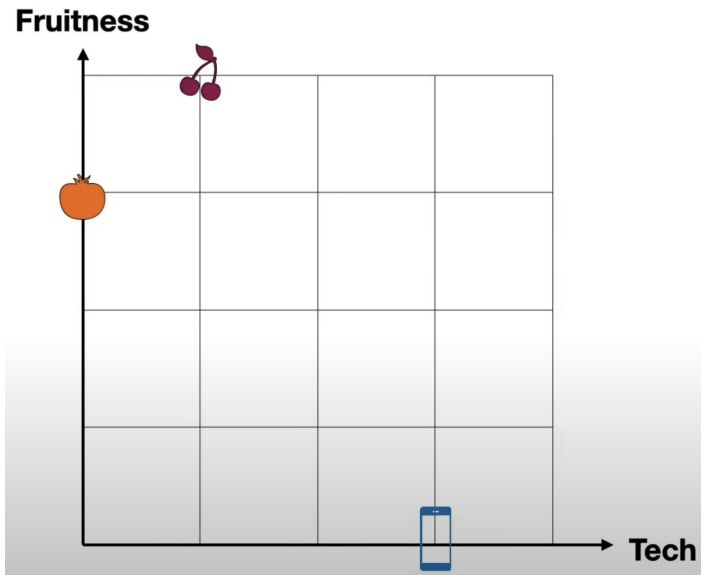
1、 点积

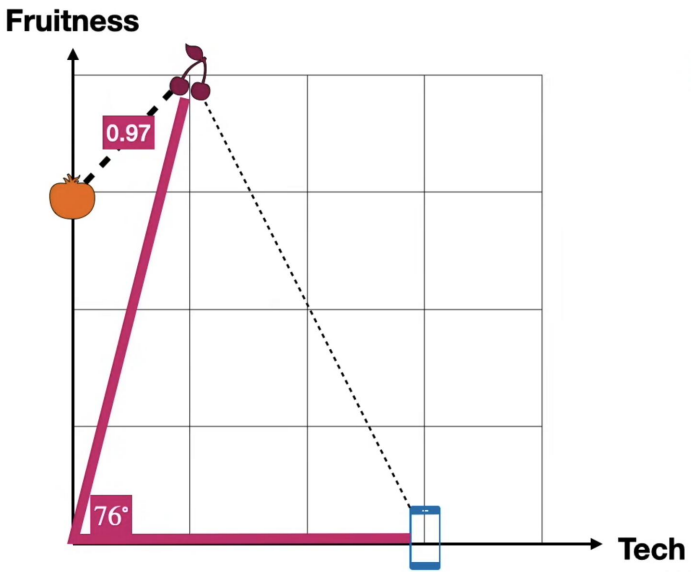

Fruitness
Tech


Fruitness
12
3
0
Tech


Tech
Fruitness
1
4
0
3
Tech
Fruitness
1 · 0 + 4 · 3 = 12
1
4
3
0
1 · 3 + 4 · 0 = 3
0
3
3
0
0 · 3 + 3 · 0 = 0

2、余弦相似度



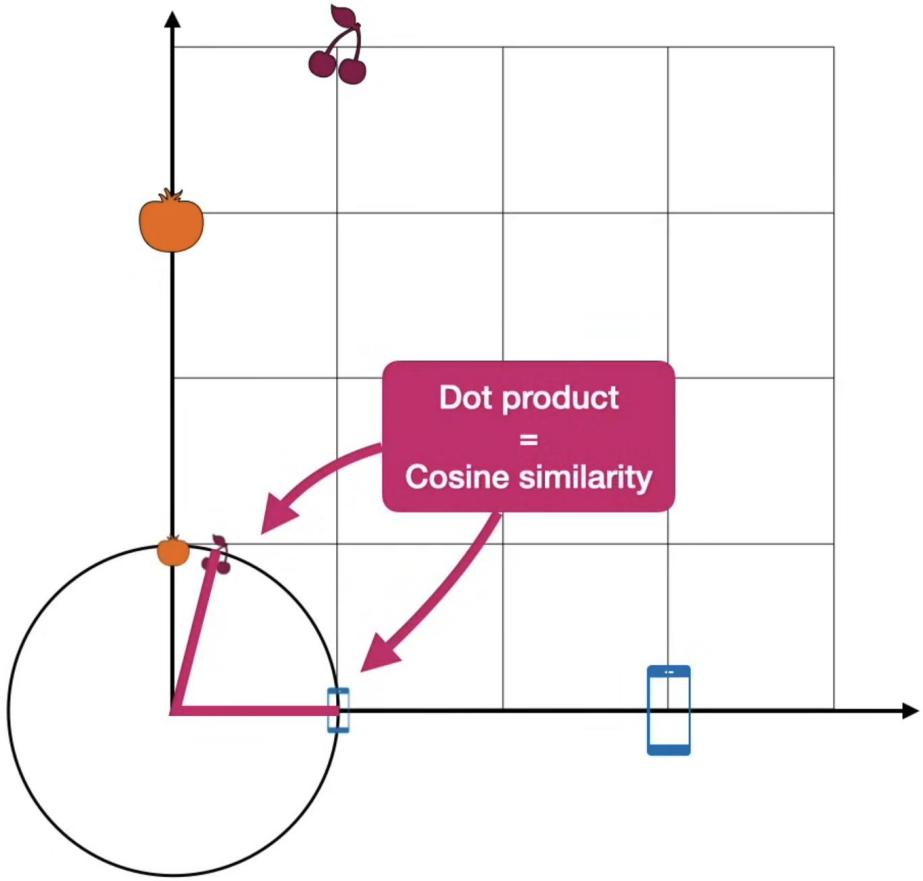


$cos(14°) = 0.97$

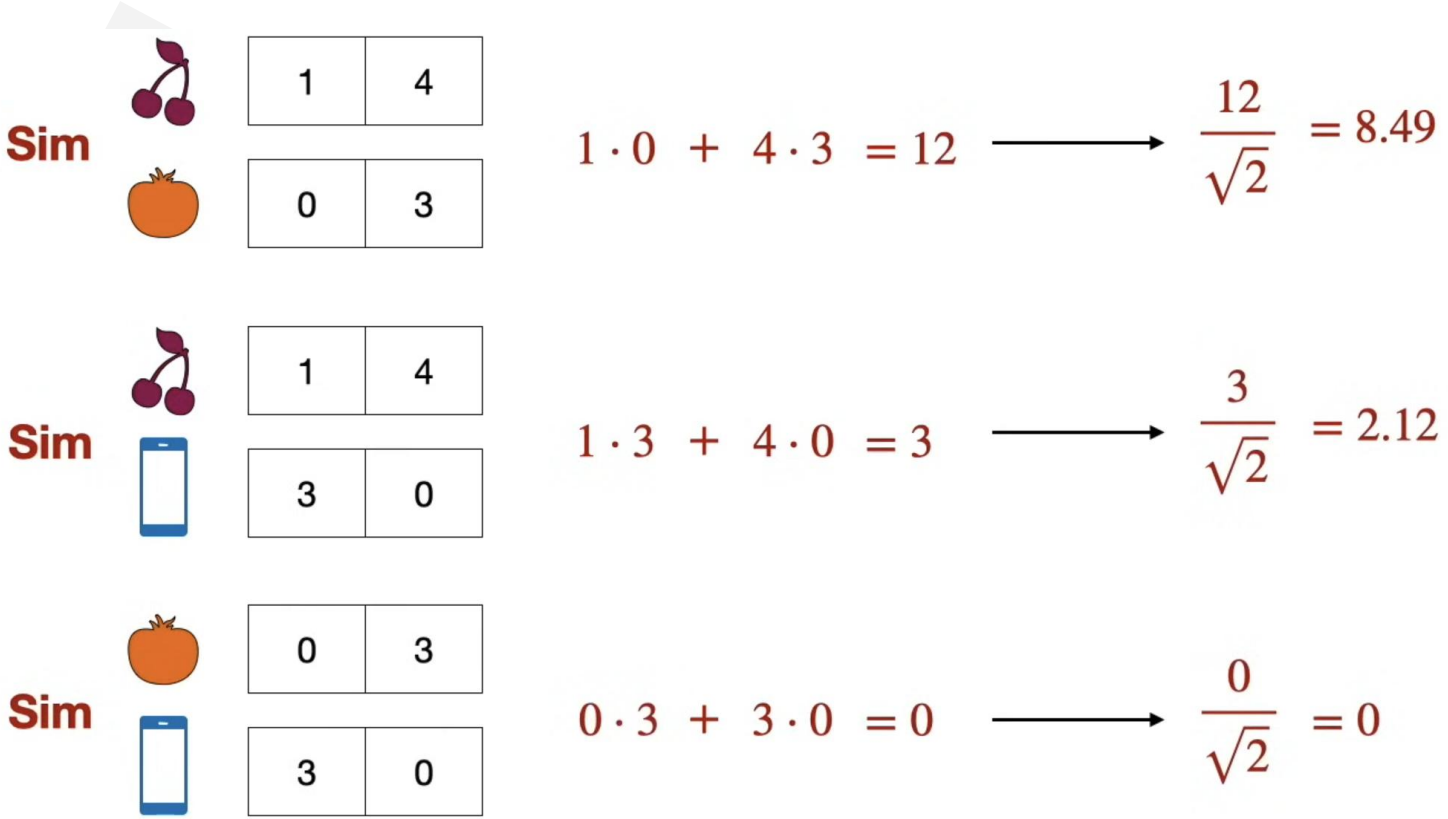
$cos(76°) = 0.24$

$cos(90°) = 0$

点积和余弦相似度

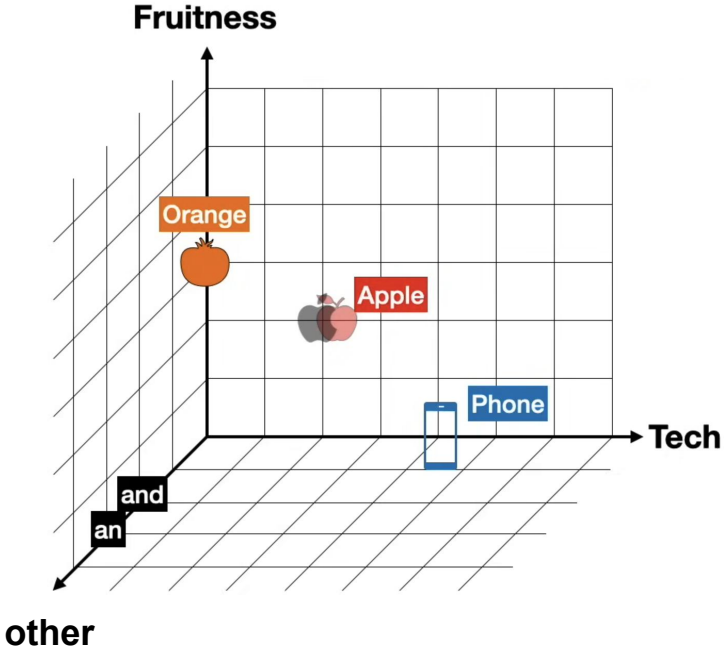


比例点积


Sim
1
4
0
3
$1 \cdot 0 + 4 \cdot 3 = 12$
$\frac{12}{\sqrt{2}} = 8.49$
Sim
1
4
3
0
$1 \cdot 3 + 4 \cdot 0 = 3$
$\frac{3}{\sqrt{2}} = 2.12$
Sim
0
3
3
0
$0 \cdot 3 + 3 \cdot 0 = 0$
$\frac{0}{\sqrt{2}} = 0$

例子

>an apple and an orange
>an apple phone



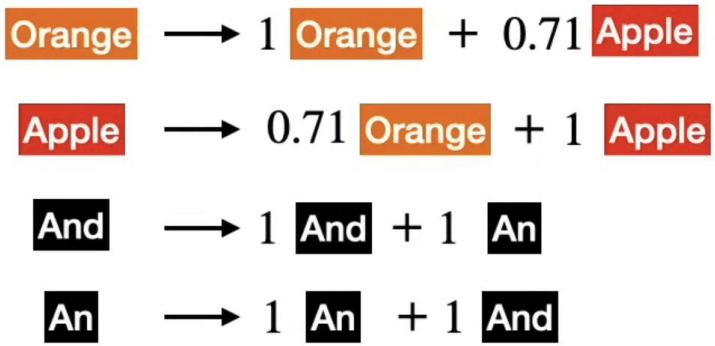
	Tech	Fruitness	Other
Orange	0	3	0
Phone	4	0	0
Apple	2	2	0
And	0	0	2
An	0	0	3

	Orange	Phone	Apple	And	An
Orange	1	0	0.71	0	0
Phone	0	1	0.71	0	0
Apple	0.71	0.71	1	0	0
And	0	0	0	1	1
An	0	0	0	1	1

计算过程

an apple and an orange

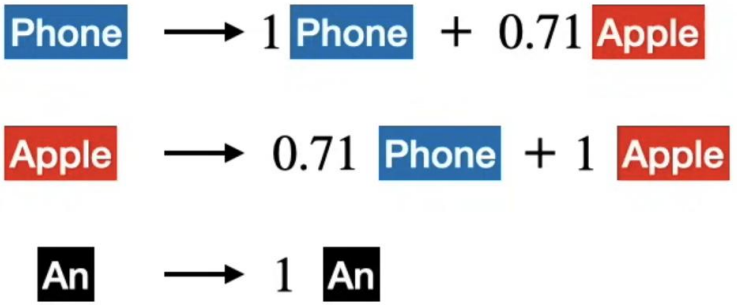
	Orange	Apple	And	An
Orange	1	0.71	0	0
Apple	0.71	1	0	0
And	0	0	1	1
An	0	0	1	1



计算过程

an apple phone

	Phone	Apple	An
Phone	1	0.71	0
Apple	0.71	1	0
An	0	0	1



归一化

$$\text{Orange} \longrightarrow \frac{1\ \text{Orange} + 0.71\ \text{Apple}}{1 + 0.71} = 0.58\ \text{Orange} + 0.42\ \text{Apple}$$

$$\text{Orange} \longrightarrow \frac{1\ \text{Orange} - 1\ \text{Motorcycle}}{1 - 1} =$$

soft Max

$$x \longrightarrow e^x$$

$$\text{Orange} \longrightarrow \frac{e^1\,\text{Orange} + e^{0.71}\,\text{Apple}}{e^1 + e^{0.71}} = 0.57\,\text{Orange} + 0.43\,\text{Apple}$$

$$\text{Orange} \longrightarrow \frac{e^1\,\text{Orange} + e^{-1}\,\text{Motorcycle}}{e^1 + e^{-1}} = 0.88\,\text{Orange} + 0.12\,\text{Motorcycle}$$

an apple and an orange

	Orange	Apple	And	An
Orange	1	0.71	0	0
Apple	0.71	1	0	0
And	0	0	1	1
An	0	0	1	1

Orange → 0.57 Orange + 0.43 Apple

Apple → 0.43 Orange + 0.57 Apple

And → 0.5 And + 0.5 An

An → 0.5 An + 0.5 And

an apple phone

	Phone	Apple	An
Phone	1	0.71	0
Apple	0.71	1	0
An	0	0	1

Phone → 0.57 Phone + 0.43 Apple

Apple → 0.43 Phone + 0.57 Apple

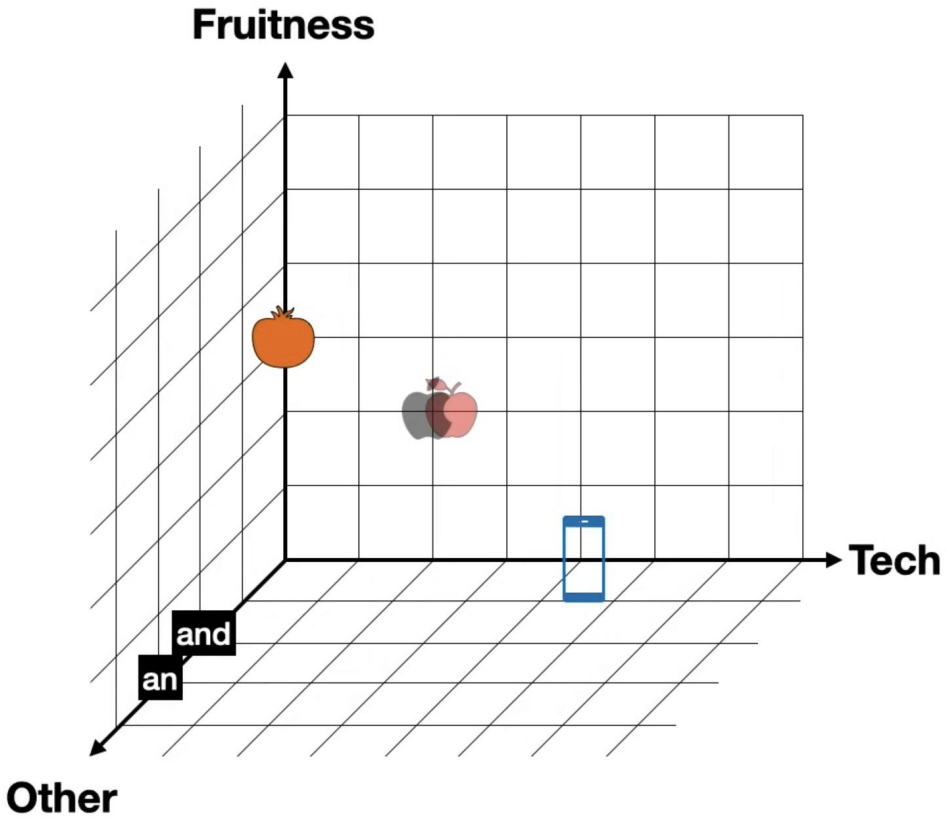
An → 1 An

真实结果

$$\frac{e^{1}\text{Orange}+e^{0.71}\text{Apple}+e^{0}\text{And}\ +\ e^{0}\ \text{An}}{e^{1}+e^{0.71}+e^{0}+e^{0}}$$

$$\text{Orange} \longrightarrow 0.4\ \text{Orange} + 0.3\ \text{Apple} + 0.15\ \text{And} + 0.15\ \text{An}$$

移动过程



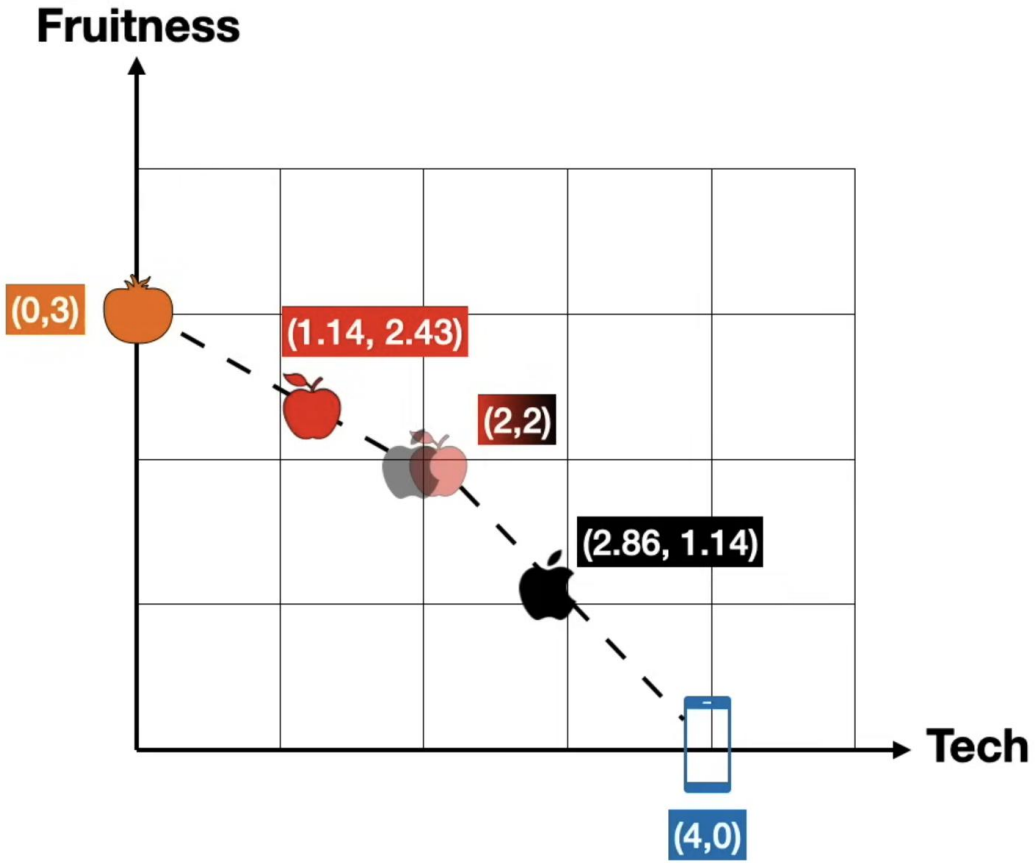
an apple and an orange

Apple → 0.43 Orange + 0.57 Apple

an apple phone

Apple → 0.43 Phone + 0.57 Apple

移动过程



an apple and an orange

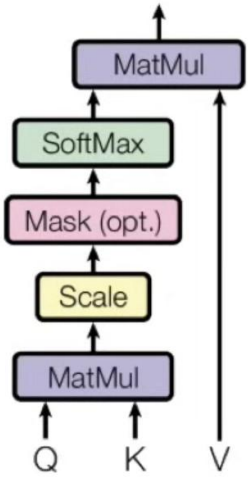
Apple → 0.43 Orange + 0.57 Apple

an apple phone

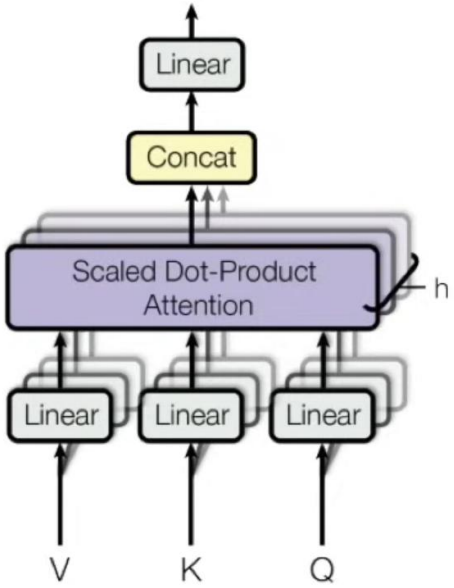
Apple → 0.43 Phone + 0.57 Apple

Q、K、V

Scaled Dot-Product Attention



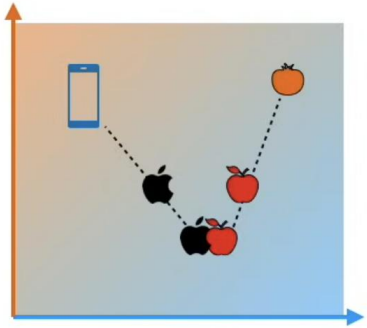
Multi-Head Attention



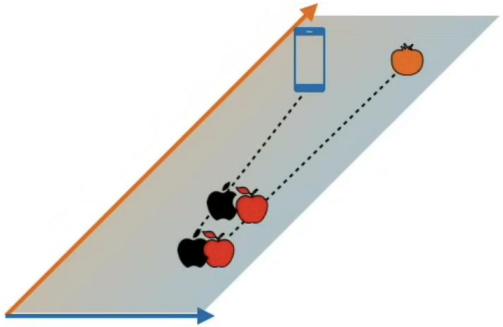
$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_\text{h})W^O$$
$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

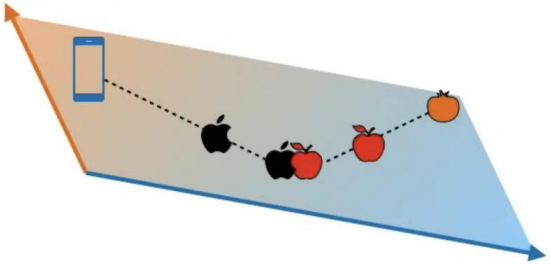
K、Q



Okay

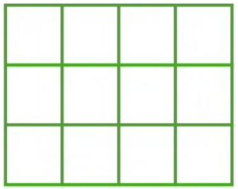


Bad

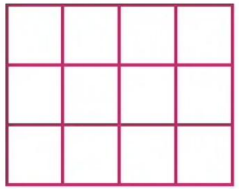


Good

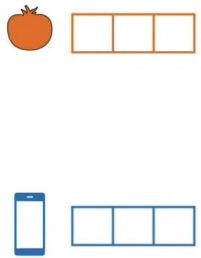
Keys

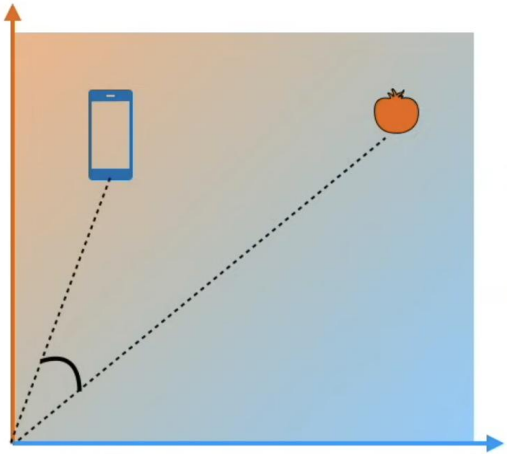


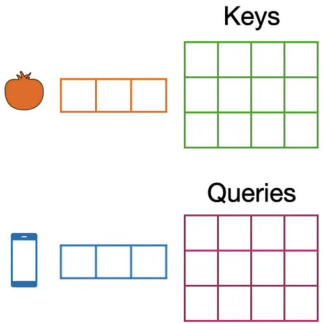
Queries

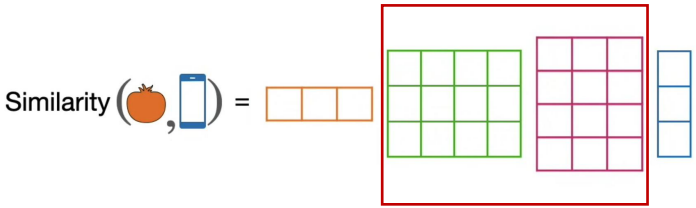


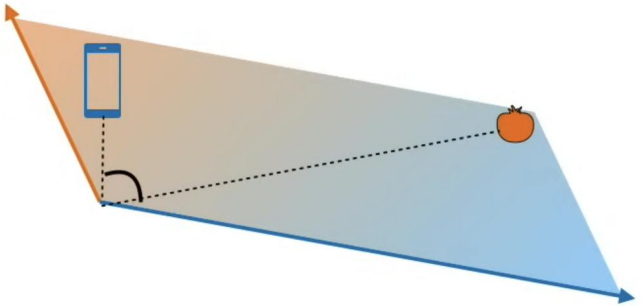
相似度



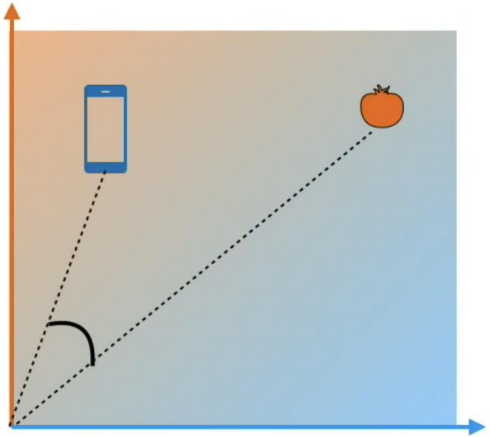


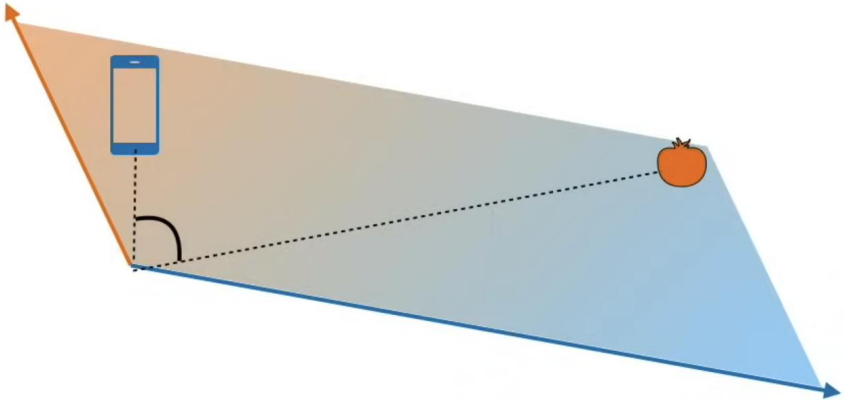


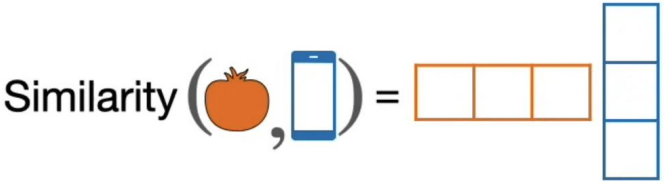


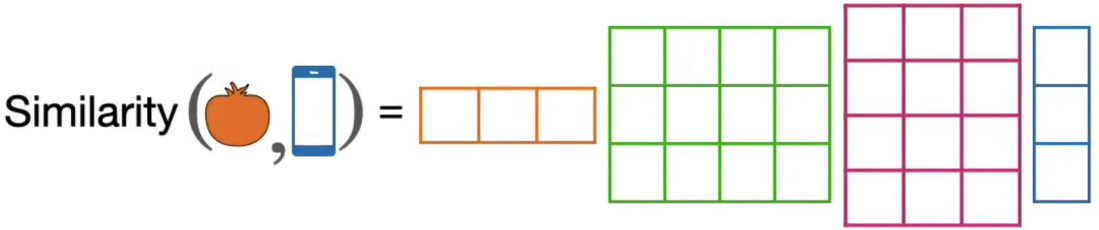


相似性的修改

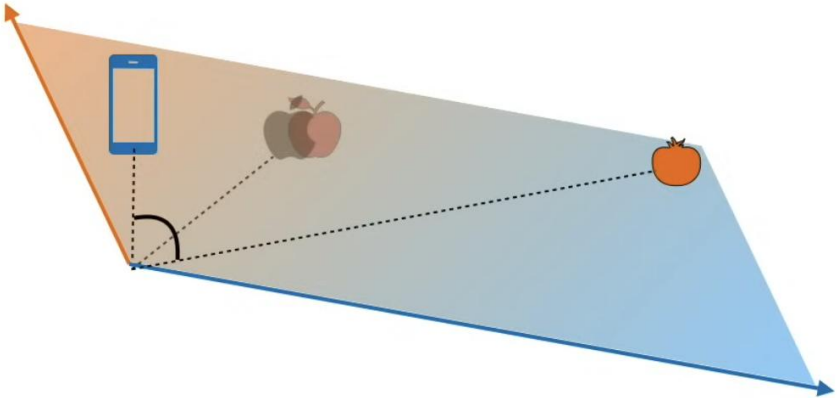




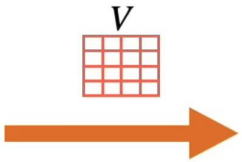


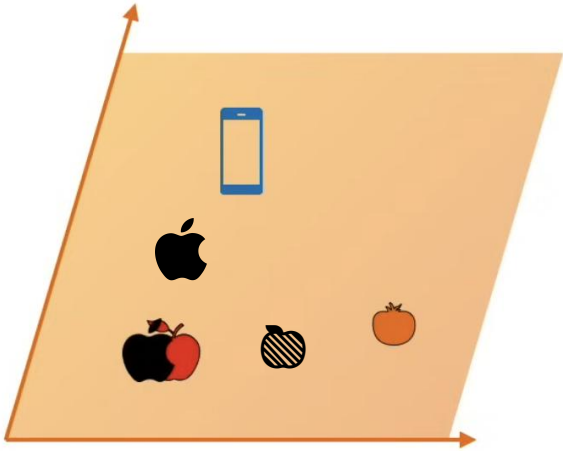


值矩阵



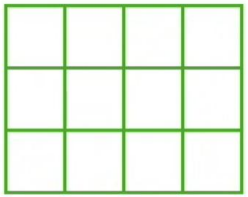
为了寻找相似性而做优化



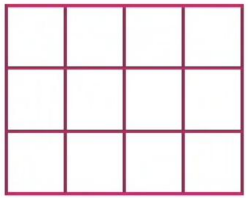


不断移动调整寻找最好的词嵌入矩阵

Keys



Queries



Q、K ==》embedding：

颜色、大小、气味……

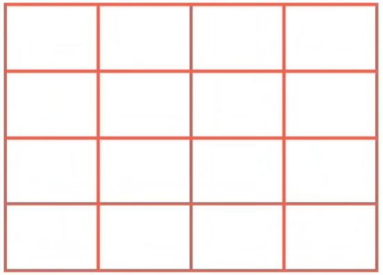
Values

实现过程

an apple and an orange

	Orange	Apple	And	An
Orange	0.4	0.3	0.15	0.15
Apple	0.3	0.4	0.15	0.15
And	0.15	0.15	0.5	0.5
An	0.15	0.15	0.5	0.5

Value matrix



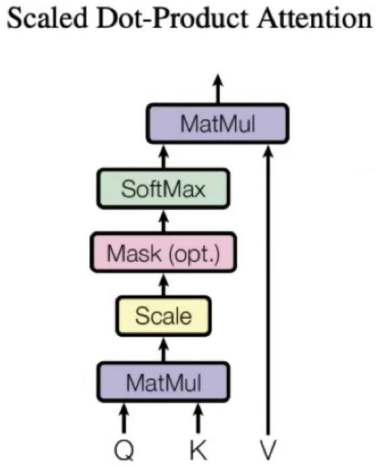
=

	Orange	Apple	And	An
Orange	v_{11}	v_{12}	v_{13}	v_{14}
Apple	v_{21}	v_{22}	v_{23}	v_{24}
And	v_{31}	v_{32}	v_{33}	v_{34}
An	v_{41}	v_{42}	v_{43}	v_{44}

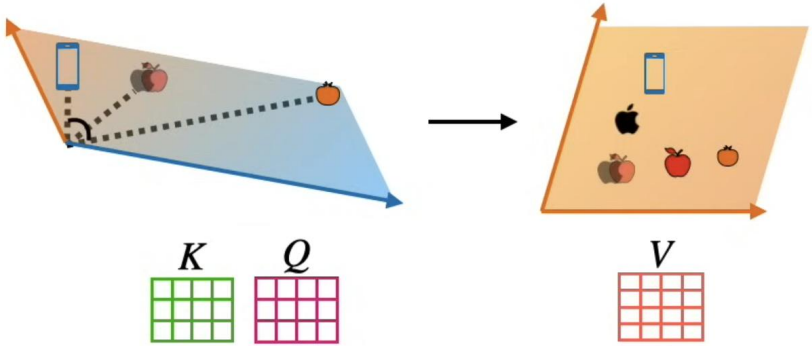
apple $\longrightarrow$ $0.3 \cdot$ orange
$+0.4 \cdot$ apple
$+0.15 \cdot$ and
$+0.15 \cdot$ an

apple $\longrightarrow$ $v_{21} \cdot$ orange
$+v_{22} \cdot$ apple
$+v_{23} \cdot$ and
$+v_{24} \cdot$ an

自注意力机制

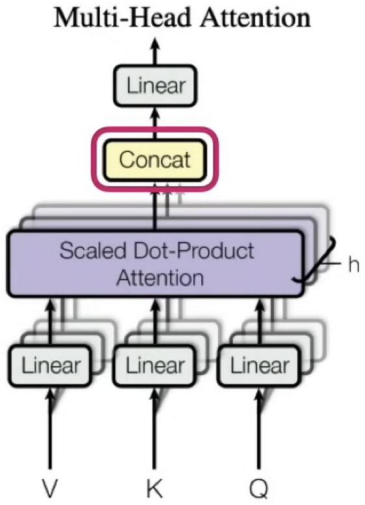


$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

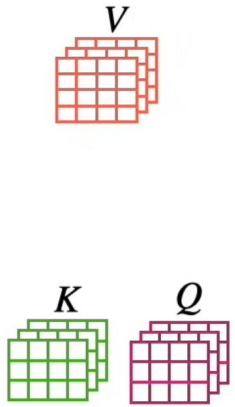


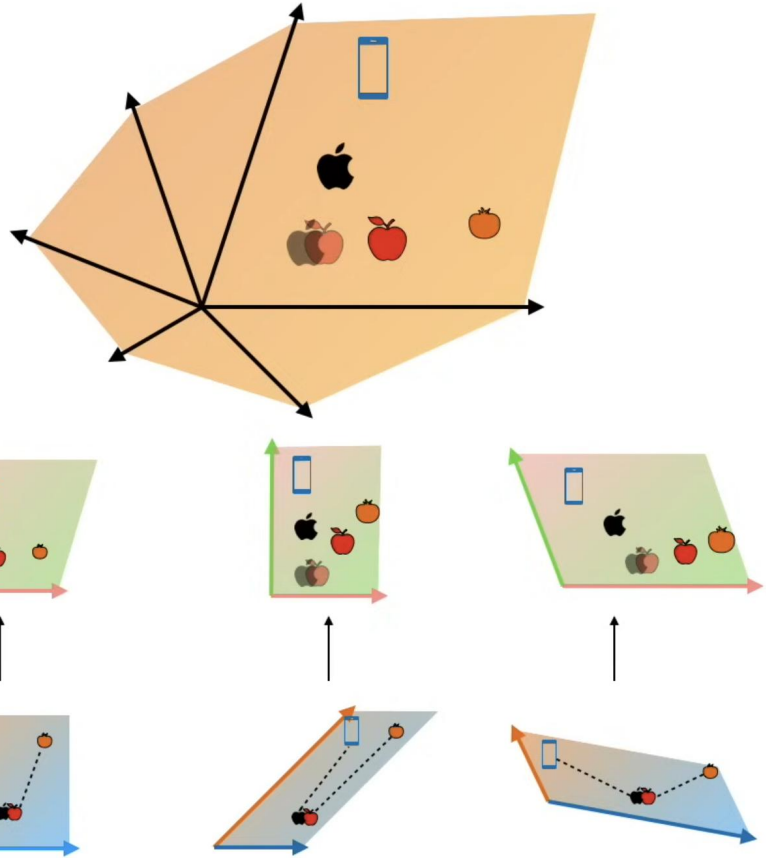
多头注意力

Multi-head attention



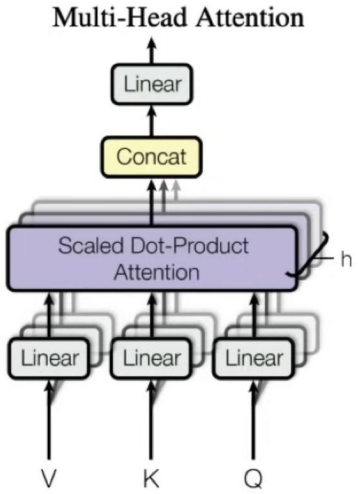
$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1,...,\text{head}_\text{h})W^O$$
$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$



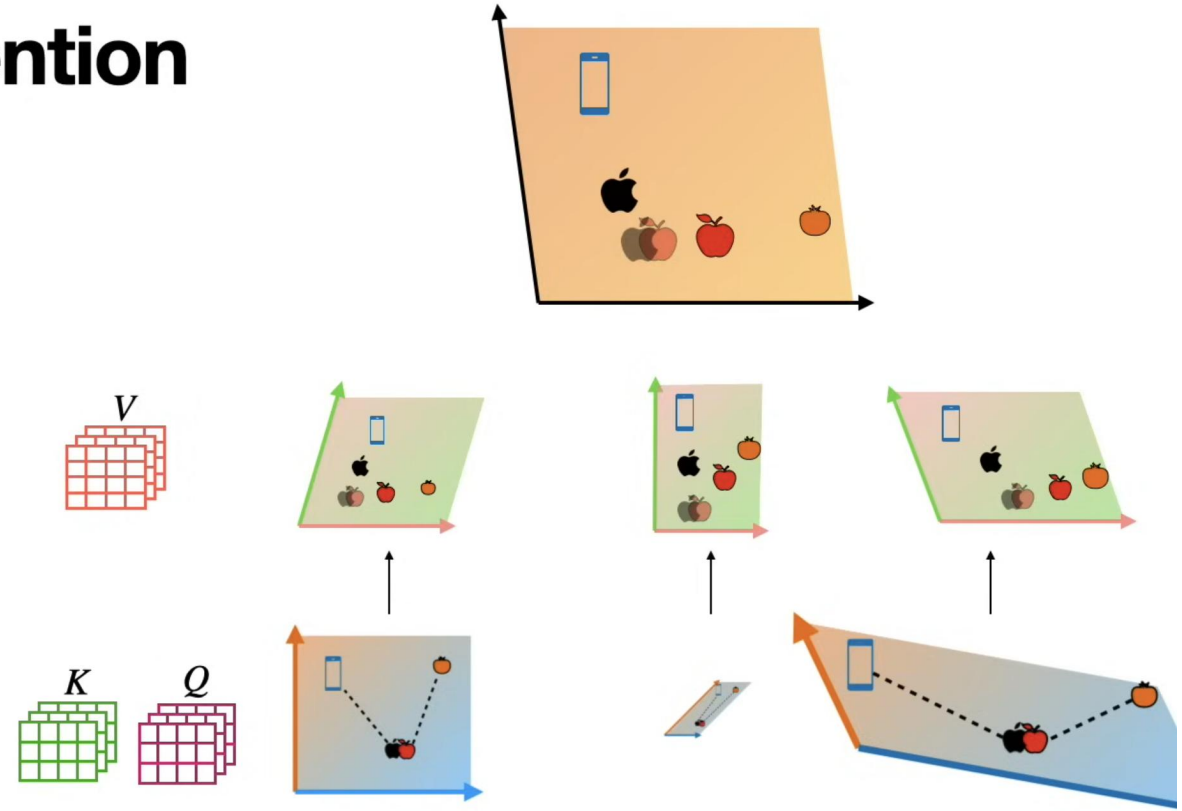


多头注意力

Multi-head attention

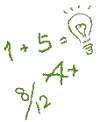


$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_\text{h})W^O$$
$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

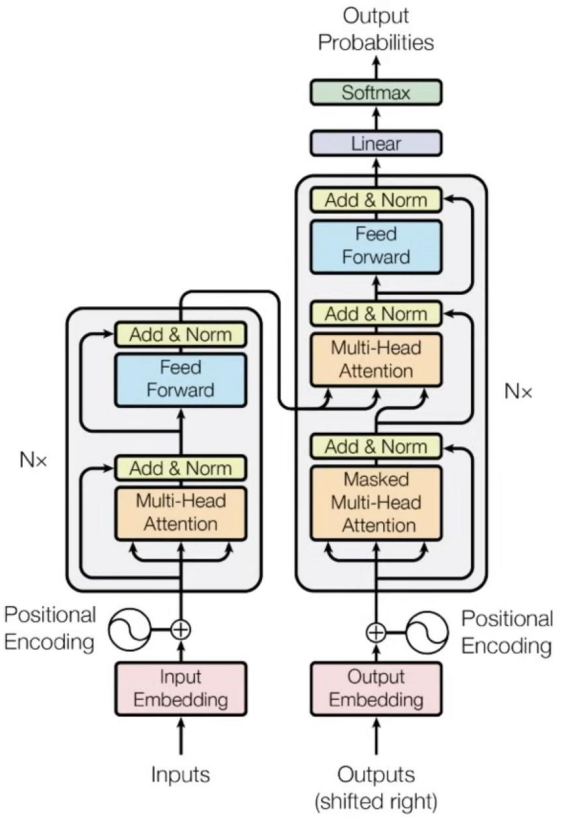


PART three

Transformer



transformer



one by one

最近怎___

怎样

怎会

怎么样

怎能

……

1-Gram

最近怎___

怎样

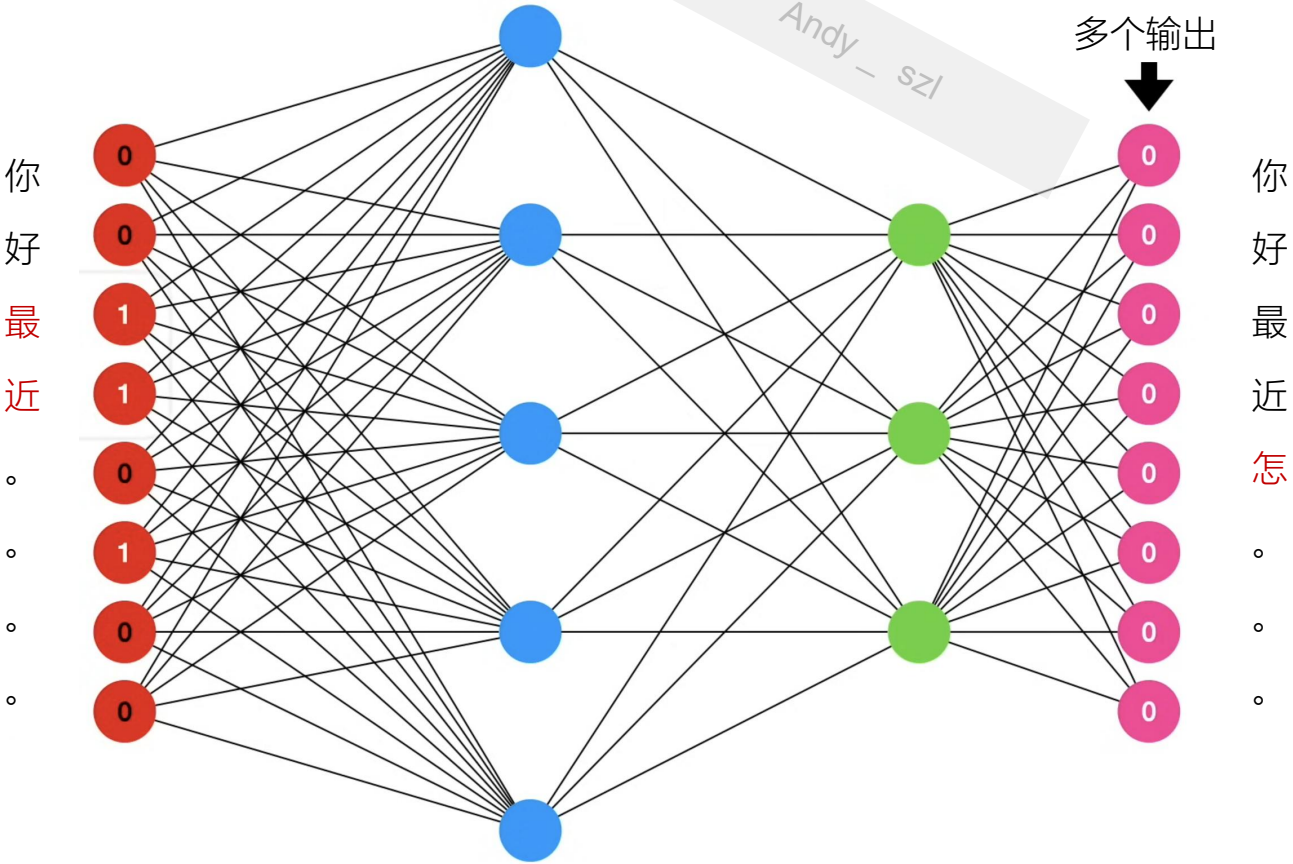
怎么样

……

3-Gram

神经网络

神经网络



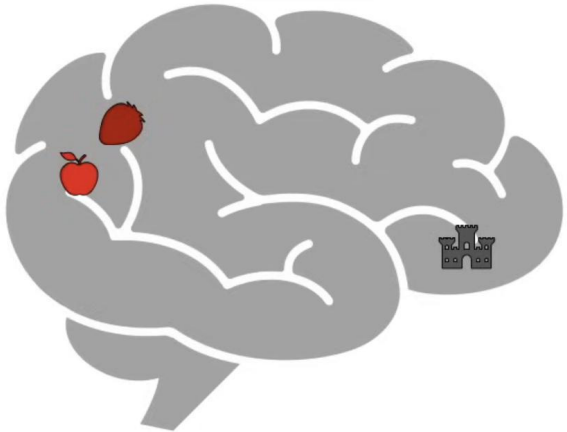
流程：tokenization

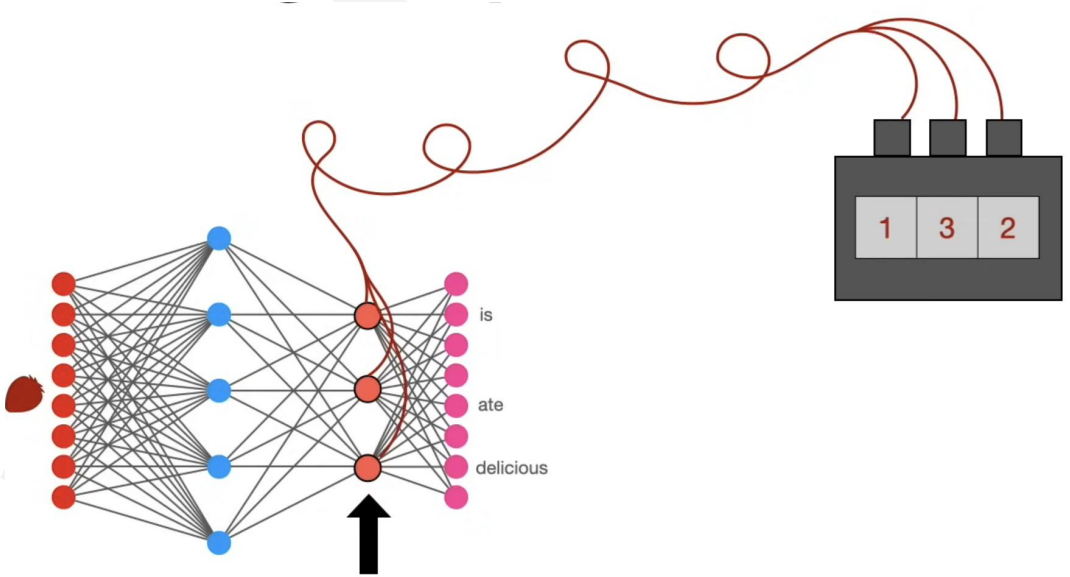
你最近怎么样 → 你 最近 怎么 样 ？

Embedding

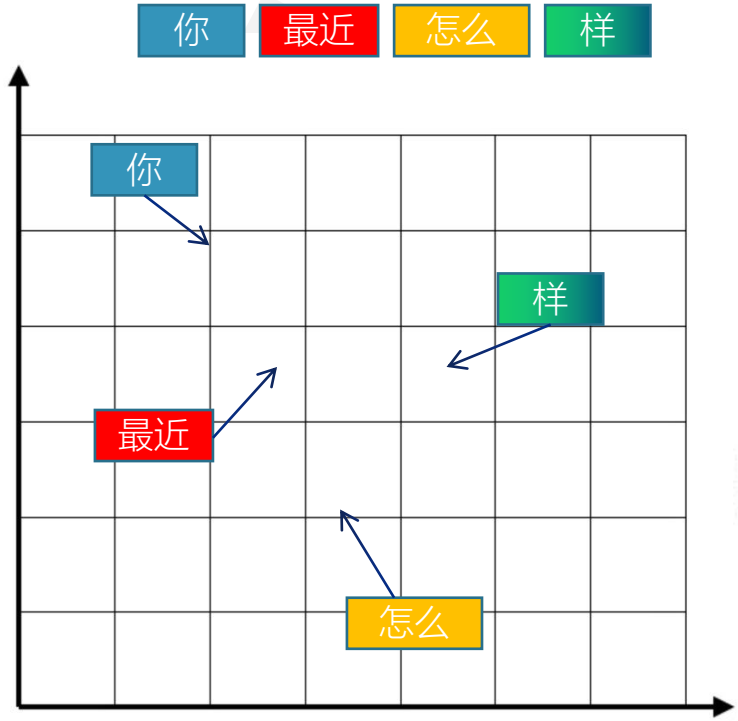
相似性高的词会给予更加相近的向量

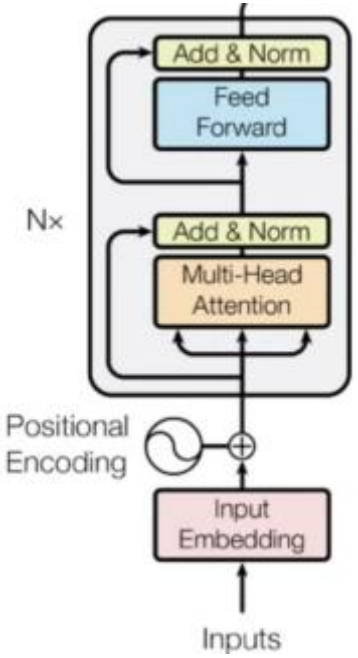
如何创建embedding

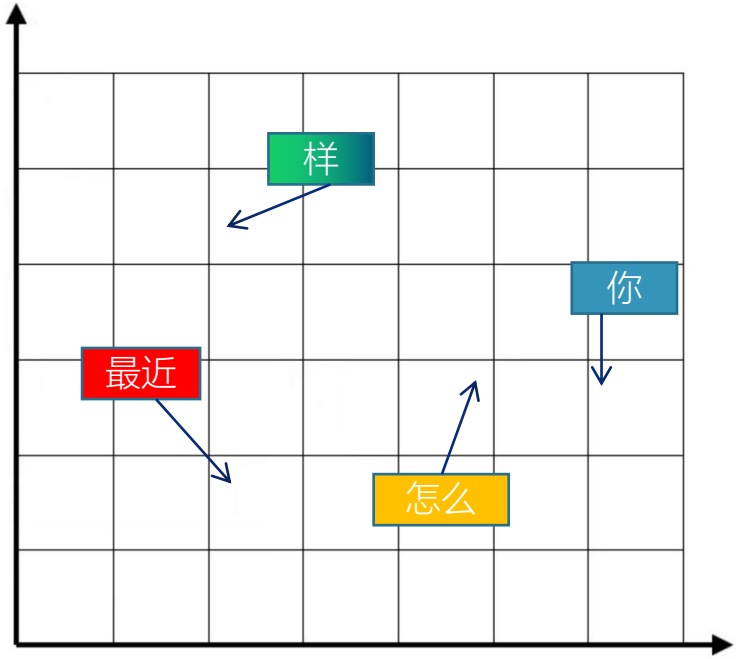




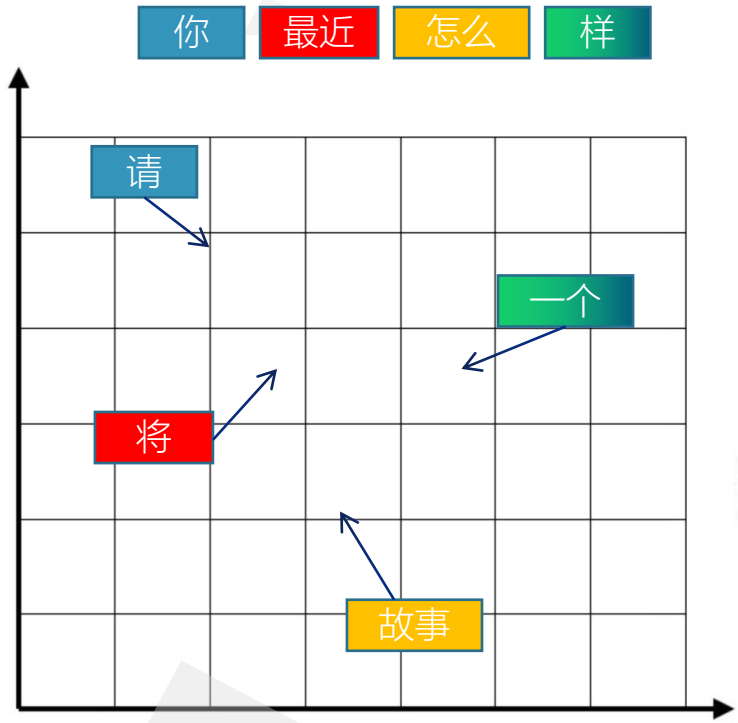
位置编码

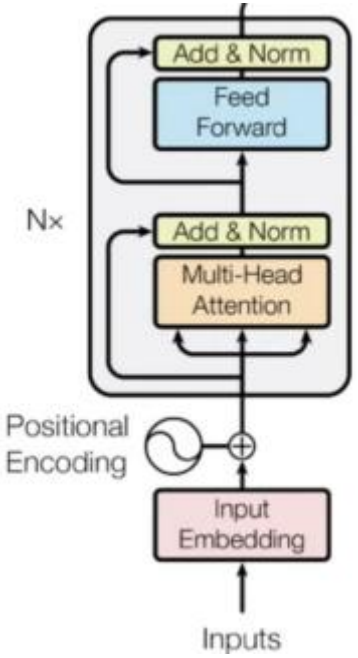

你
最近
怎么
样
你
样
最近
怎么

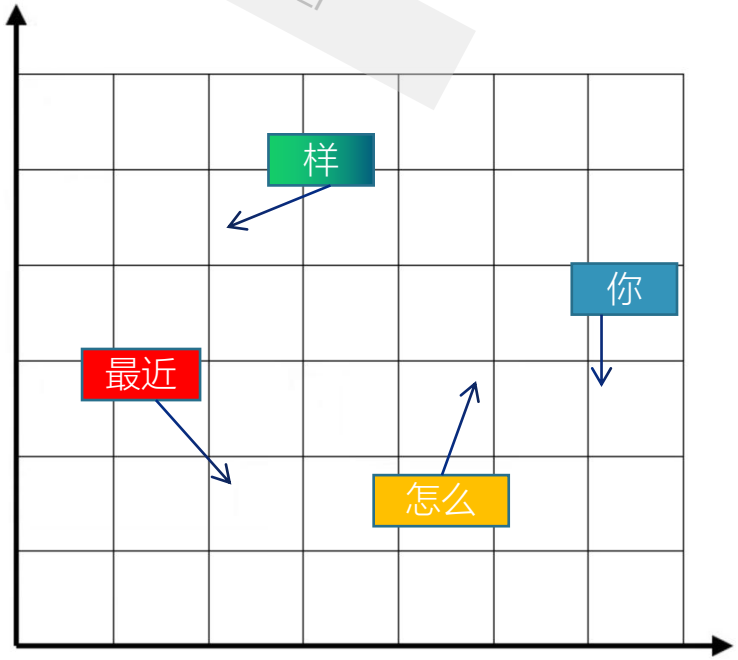

Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Input Embedding
Inputs


样
你
最近
怎么

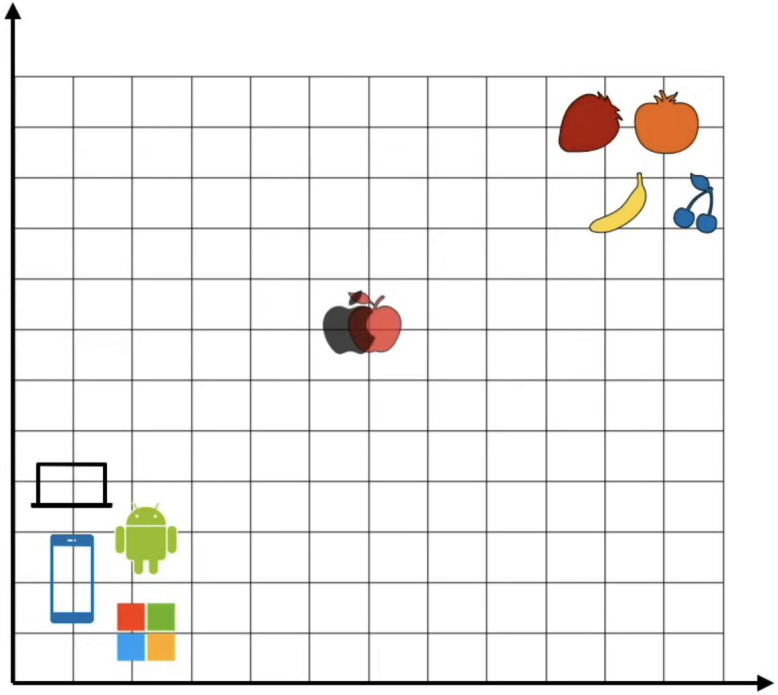
位置编码

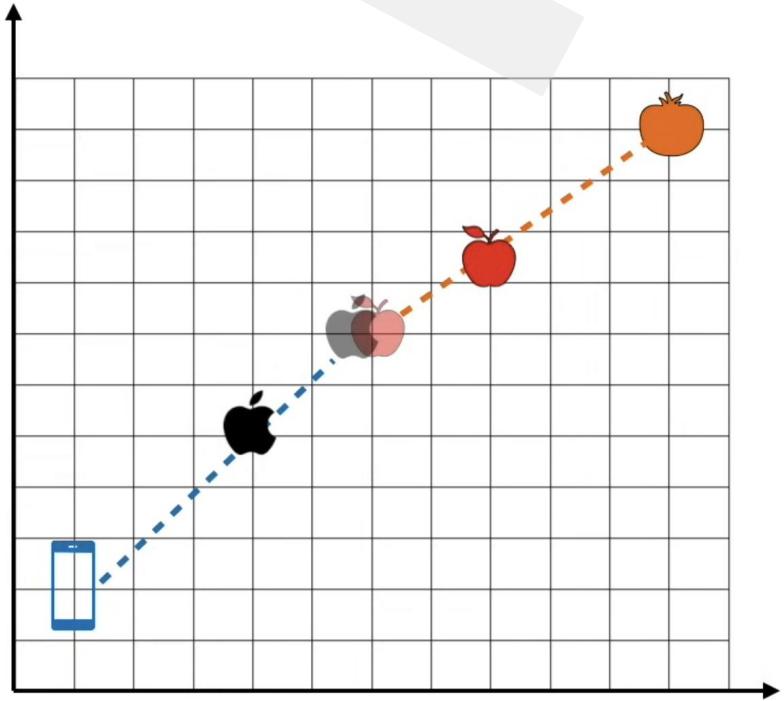




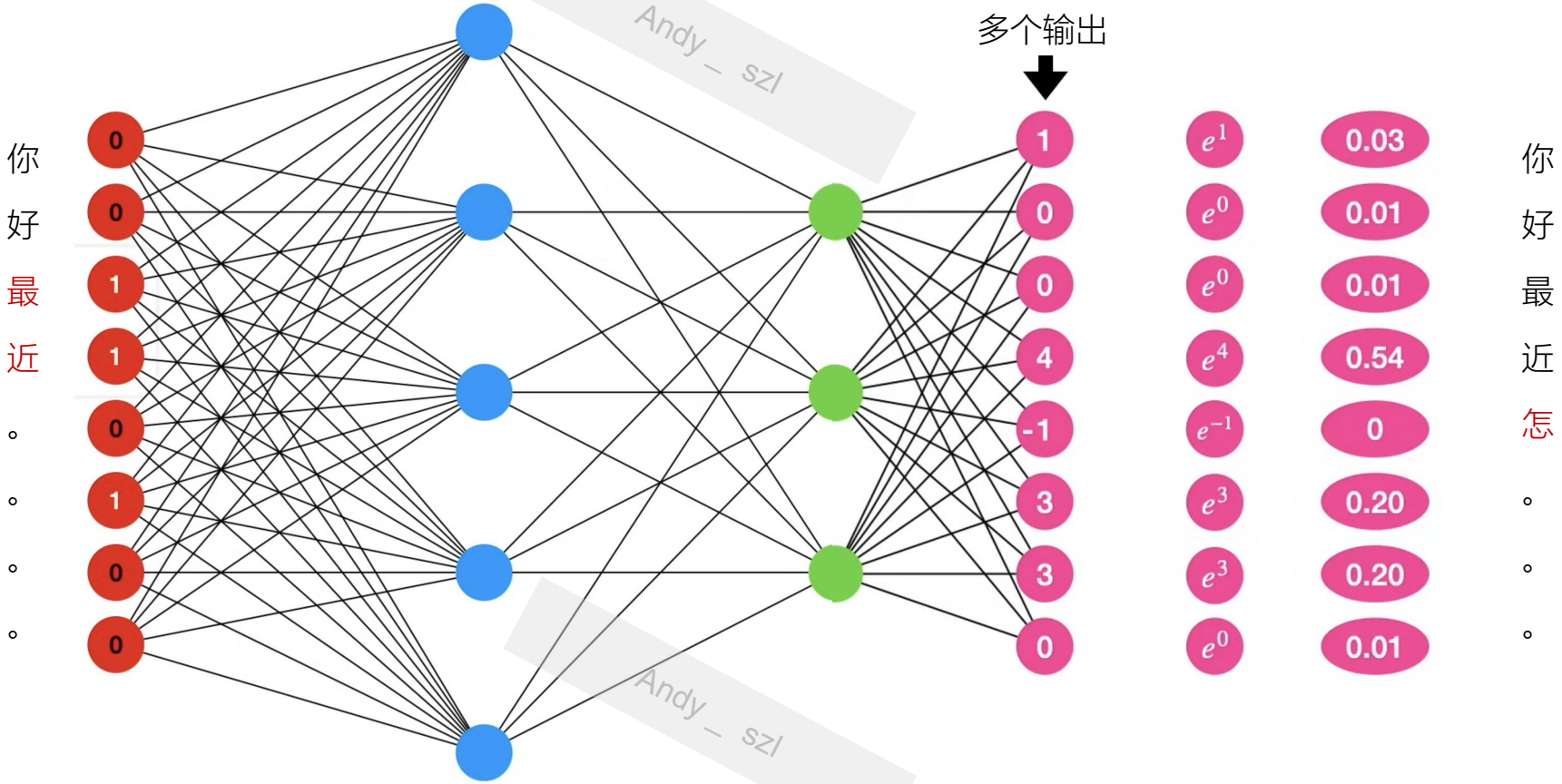


自注意力机制

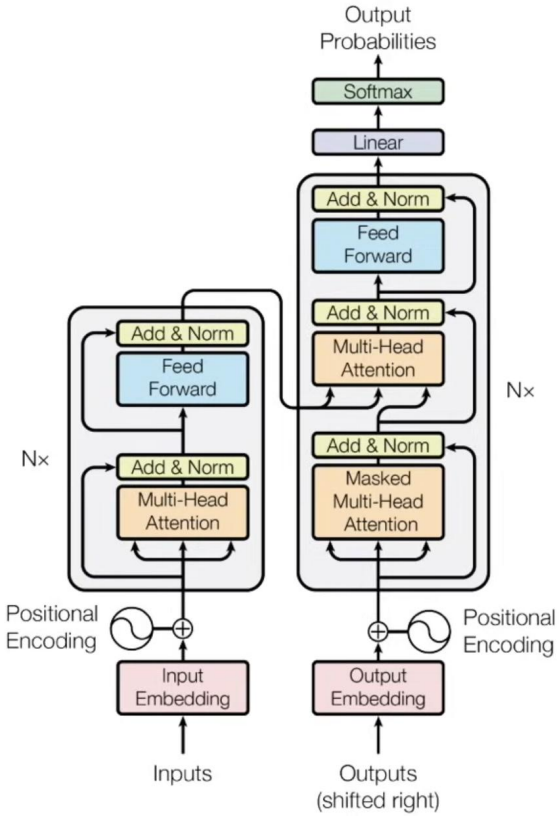




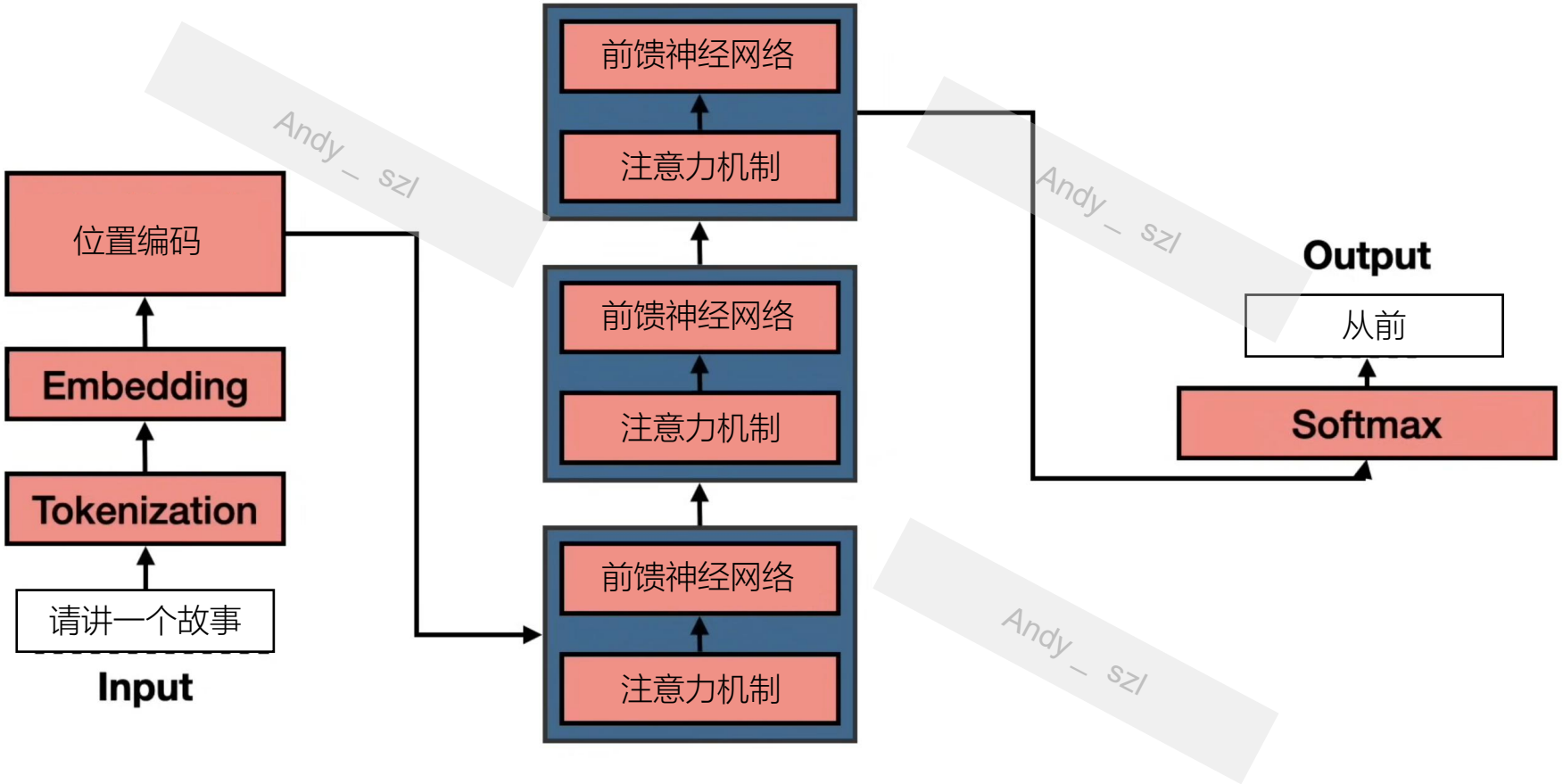
Soft Max


多个输出
你
好
最
近
。
。
。
。
0
0
1
1
0
1
0
0
1
0
4
-1
3
3
0
e^1
e^0
e^0
e^4
e^{-1}
e^3
e^3
e^0
0.03
0.01
0.01
0.54
0
0.20
0.20
0.01
你
好
最
近
怎
。
。
。
Andy – szl
Andy – szl

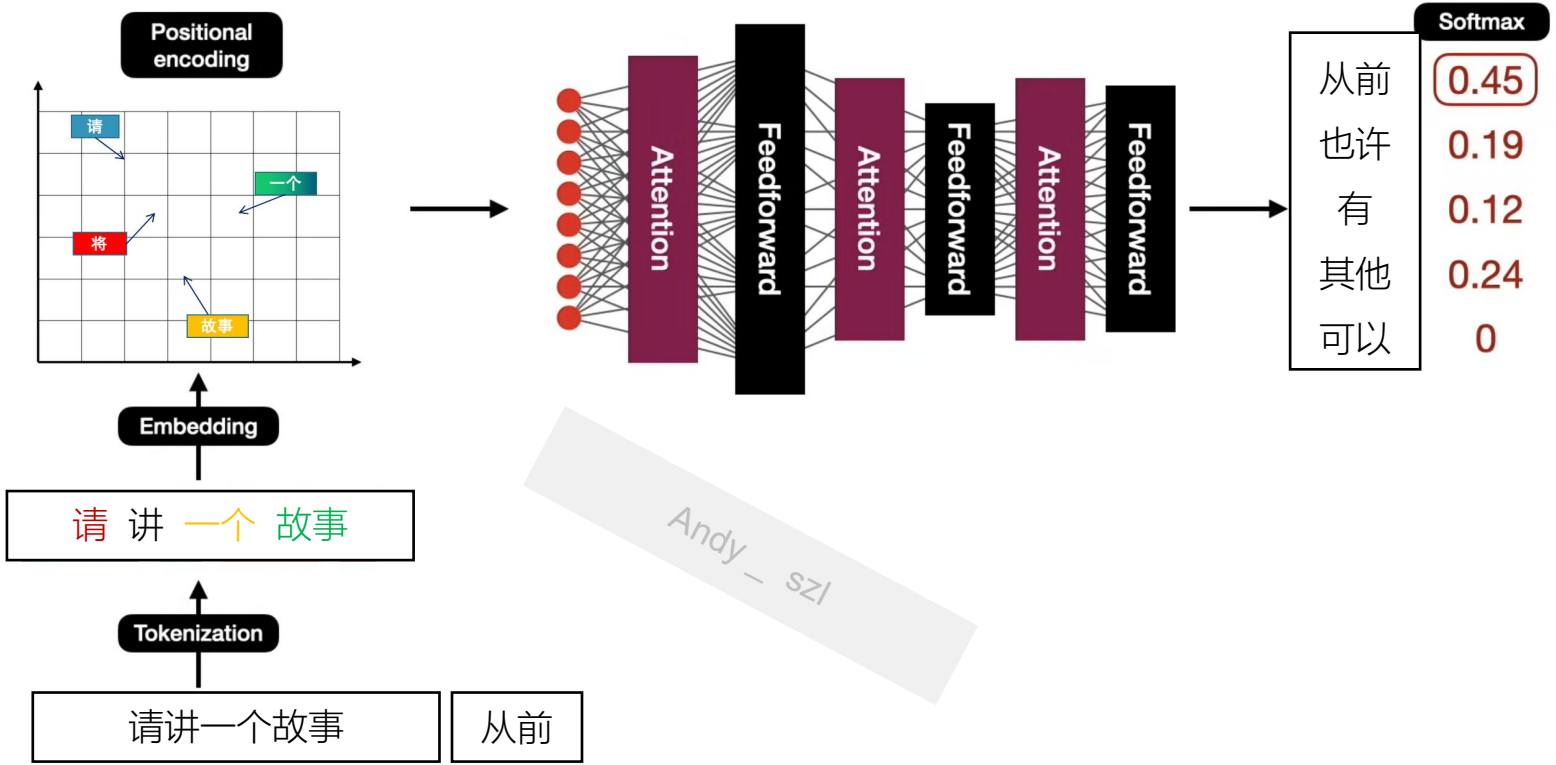
整个架构


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

整个架构


前馈神经网络
注意力机制
前馈神经网络
注意力机制
前馈神经网络
注意力机制
位置编码
Embedding
Tokenization
请讲一个故事
Input
Output
从前
Softmax
Andy _ szl
Andy _ szl
Andy _ szl

整个架构



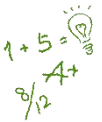
位置编码

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

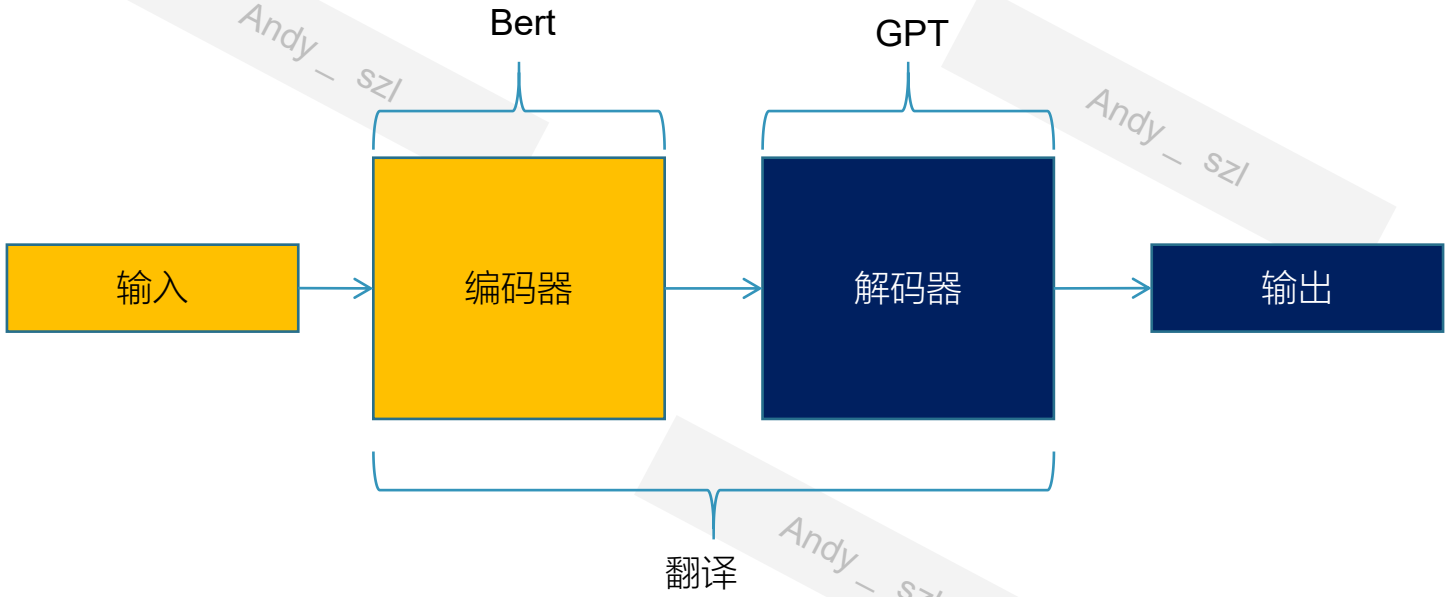
$$P(pos, 2i+1) = cos(\frac{pos}{10000^{2i/d_{model}}})$$

PART four

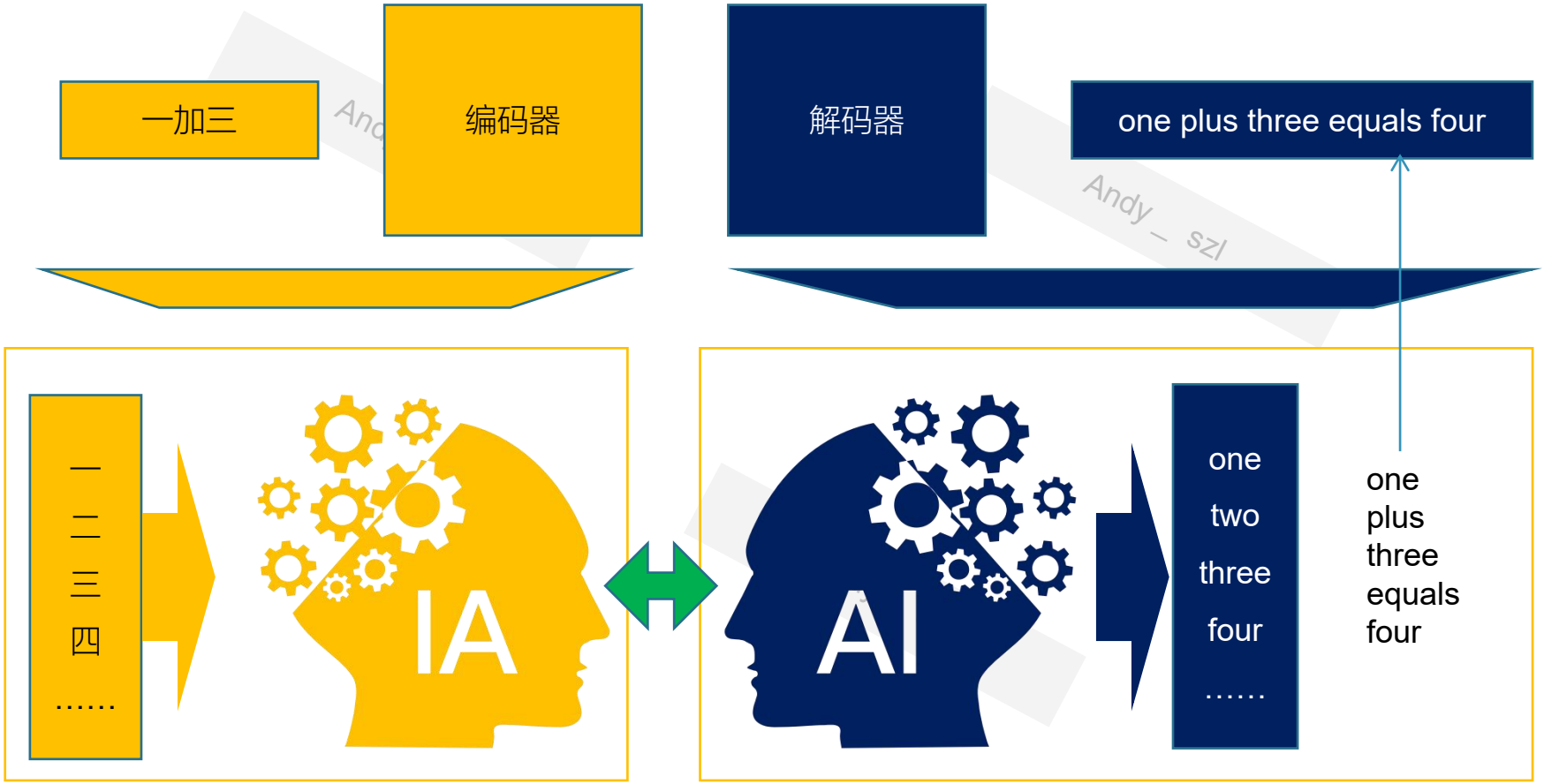
编码器和解码器



编码器和解码器



翻译示例


一加三
编码器
解码器
one plus three equals four
一
二
三
四
……
IA
AI
one
two
three
four
……
one
plus
three
equals
four

几个问题

1 解码器的输入是编码器的全部内容还是部分内容?

2 编码器和解码器使用的是一套embedding还是两套?

3 会有使用一套embedding的模型吗

4 使用两套embedding时，两套embedding如何建立联系呢

Thank you

致谢