Why Tranformer?

1. All the tokens in parallel
2. Handling long set of tokens was challenging with other architecture
3. A word can be represented in different context in different place

I went to bank to withraw money
I went to river bank to get some fresh air

I have interest in biking
The interest in the loan is increasing

The embeddings may not be 100% correct representation

we update embedding based on the input passed

4.

Diffusion model ⟸

Image → Image

text → Image

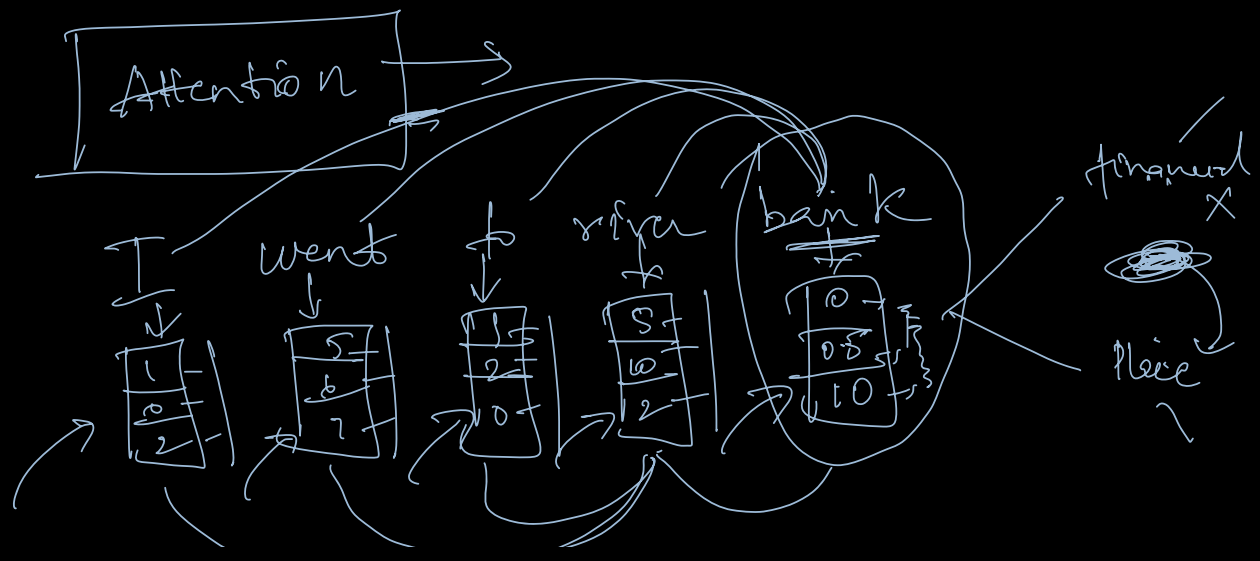
Image → Text

Voice → Text

Text → Voice

2017 → Attention is all you need ←

Transformer

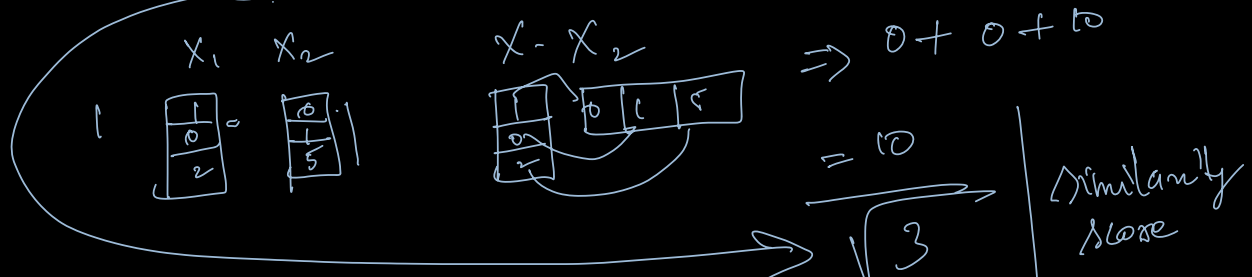


Updating the embeddings based on the Input passed by comparing the other tokens

How ?

→ similar to cosine similarity

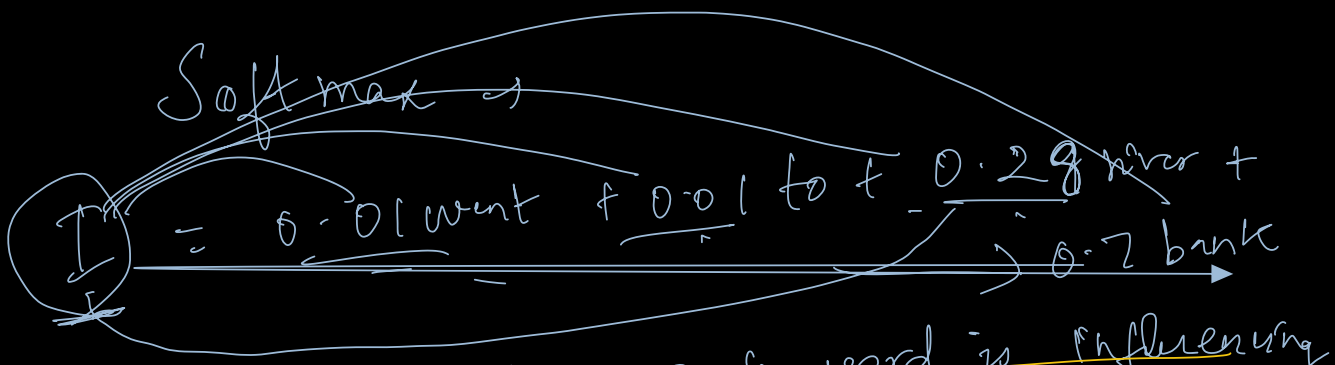
Scaled dot product



$len(x_1) = 3$

		went	to	river	bank
	1	0.5	0.6	0.1	20
went		1			
to			1		
river				1	
bank					1

$I = 0.5$ went $+ 0.6$ to $+ 10$ river $+ 20$ bank $= I$



to represent how each word is influencing the other word

Update each word embedding based on the influencing factor

How?

Key Queries Values

Multi Head Attention

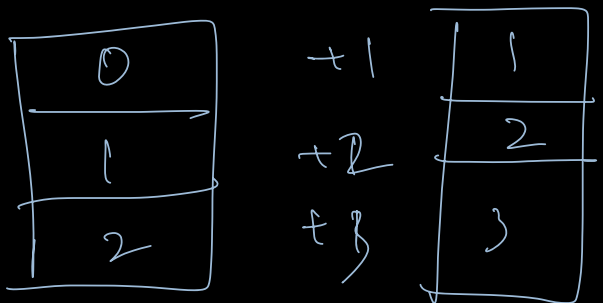
why?

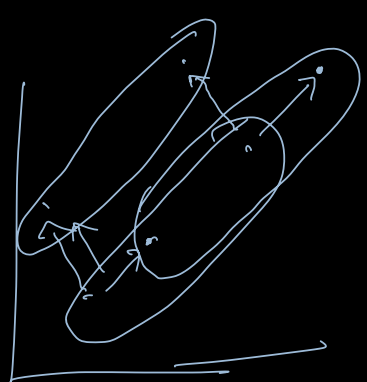
Decision tree

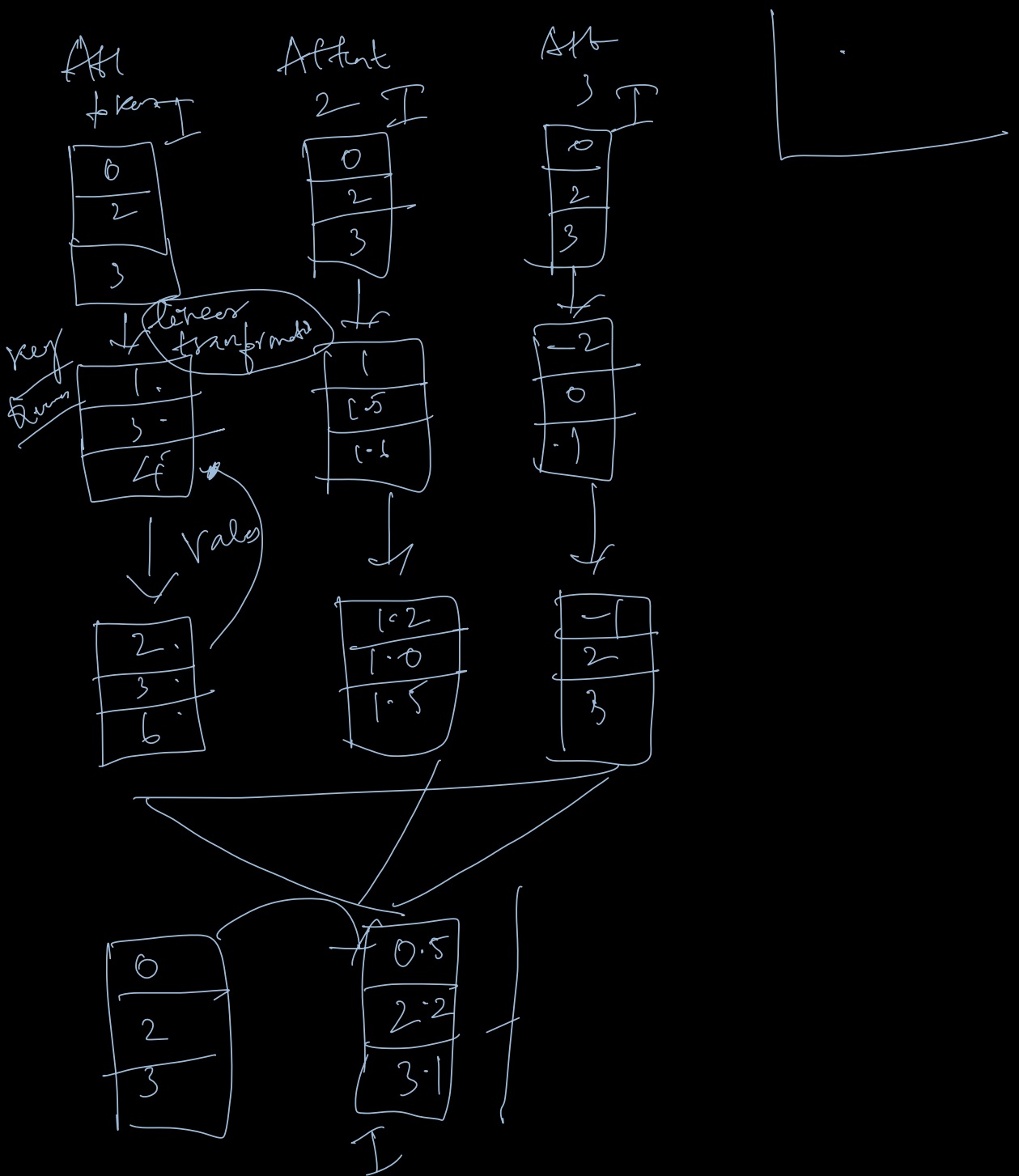
Random Forest

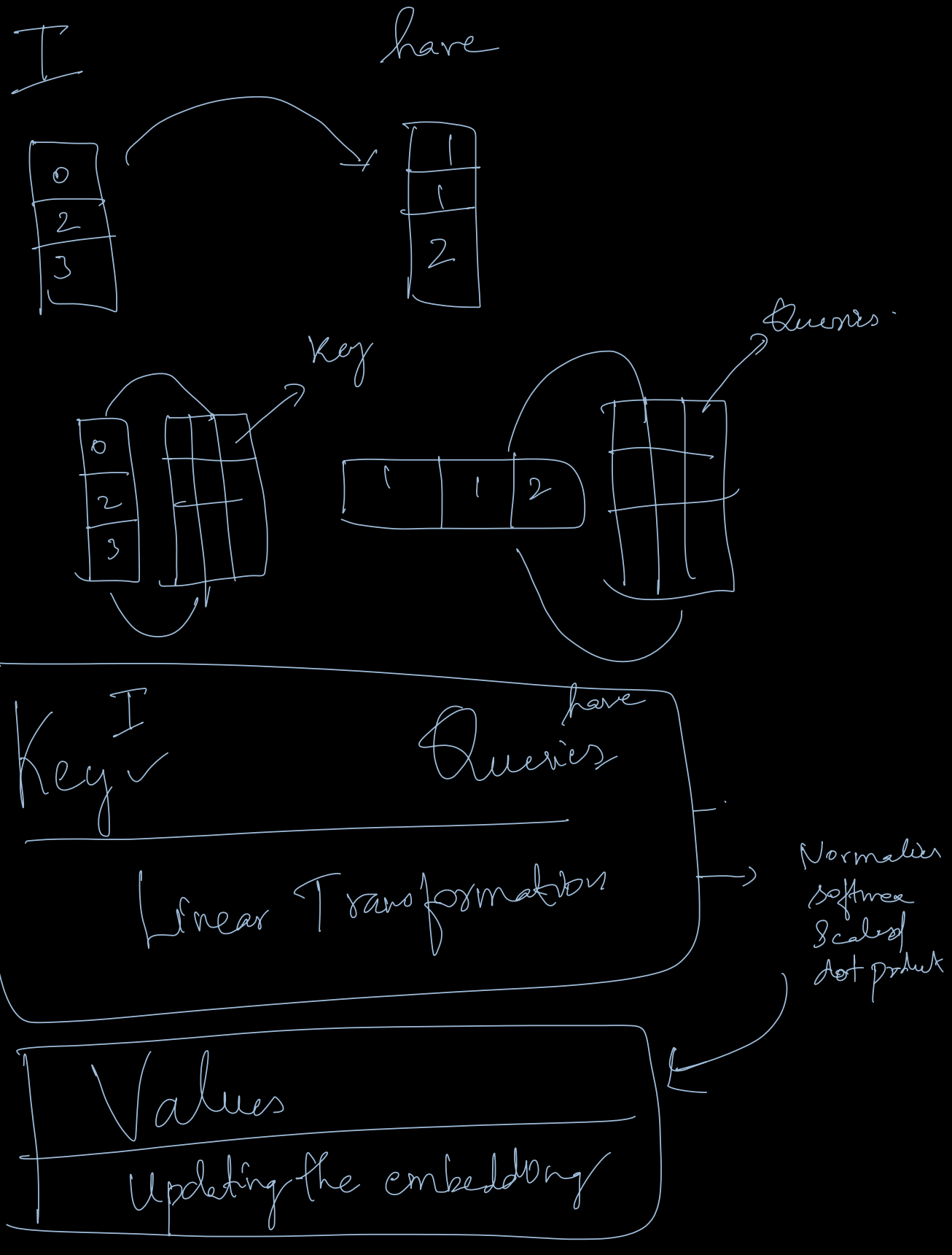
Stacking

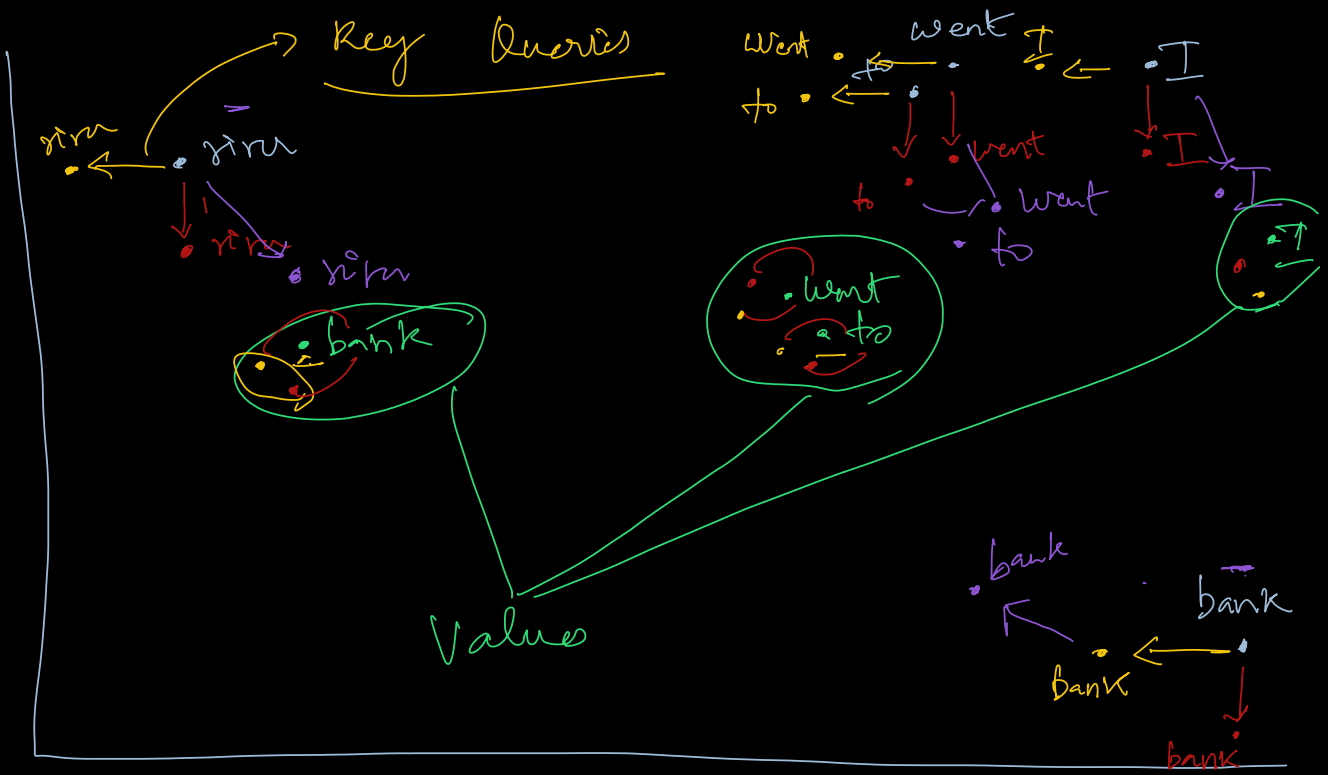
Linearly transform the embedding

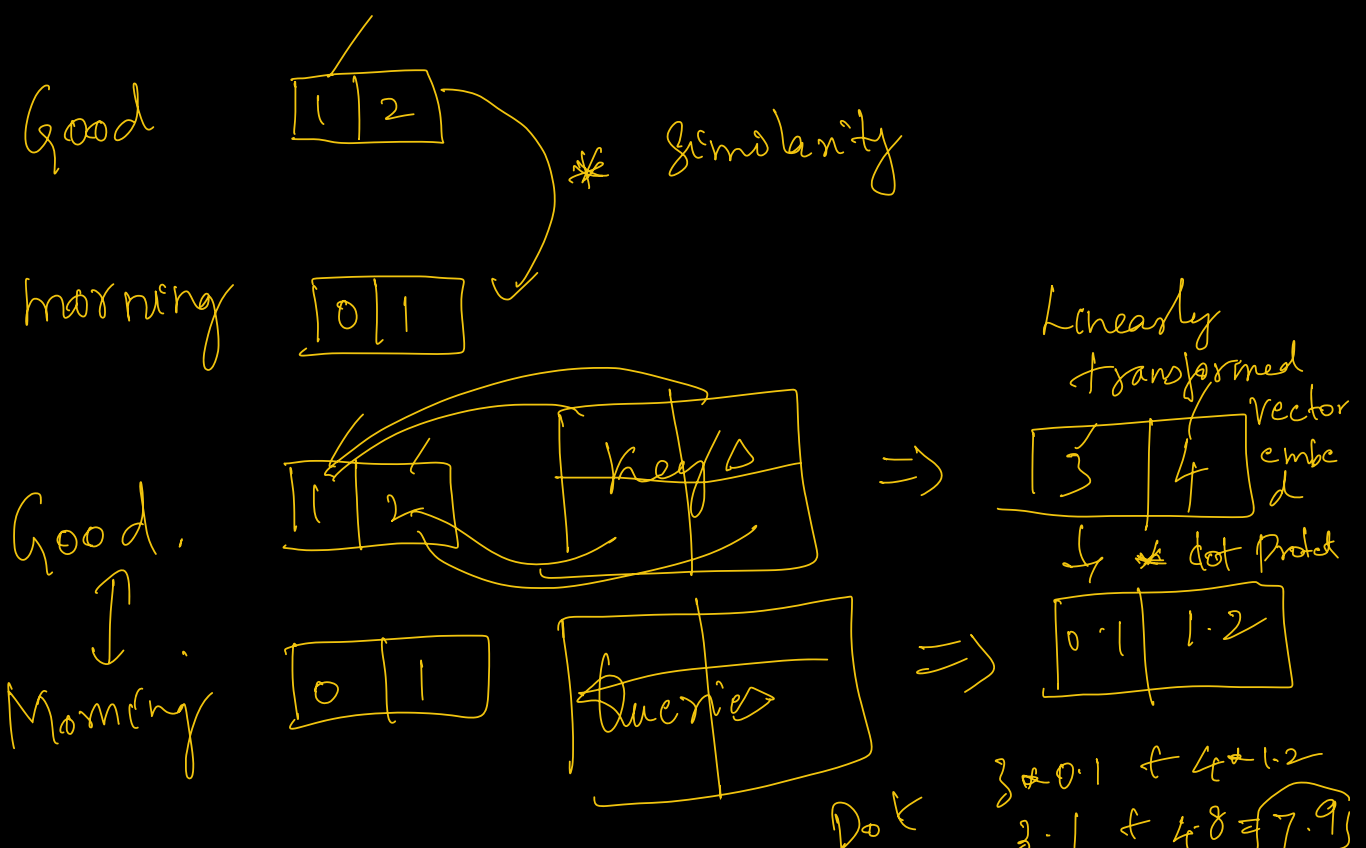




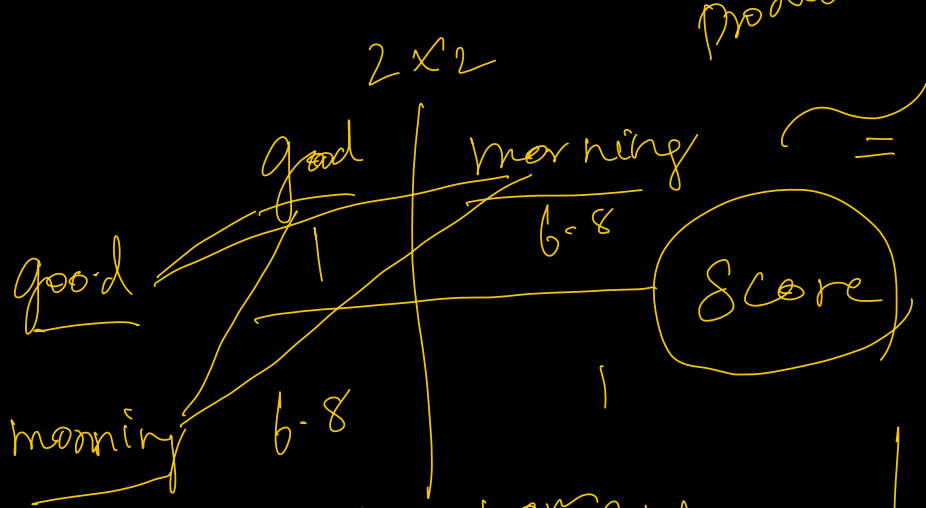

Att
token
I
0
2
3
Attent
2
I
0
2
3
Att
3
I
0
2
3
linear transformation
key
query
value
1.
3.
4
1
1.5
1.1
-2
0
-1
2.
3.
6.
1.2
1.0
1.5
-1
2
3
0.5
2.2
3.1
I

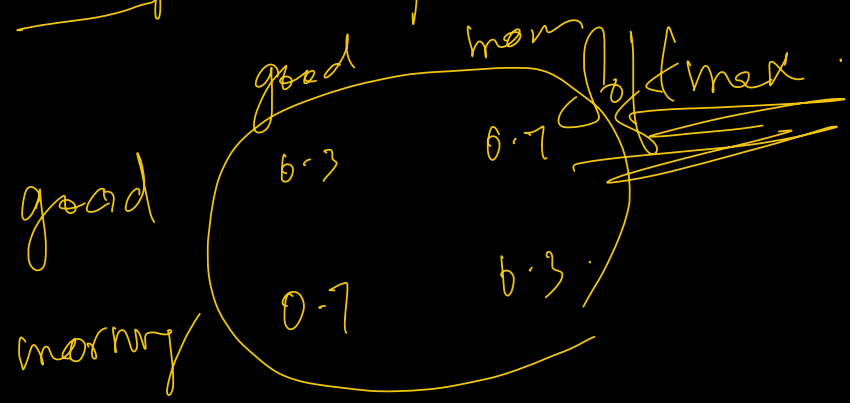

I
have
0
2
3
1
1
2
Key
Queries
0
2
3
1
1
2
Key
I
have
Queries
Linear Transformation
Normalized
Softmax
Scaled
dot product
Values
Updating the embedding


Key Queries
Went
went
I
to
went
I
went
to
bank
bank
bank
bank
bank
Values


Good
1 2
* Similarity
morning
0 1
Linearly transformed vector embed
Good.
I
Morning
1 2
Keys
3 4
* dot Product
0 1
Queries
0·1 1·2
Dot
3*0·1 + 4*1·2
3·1 + 4·8 = 7·9

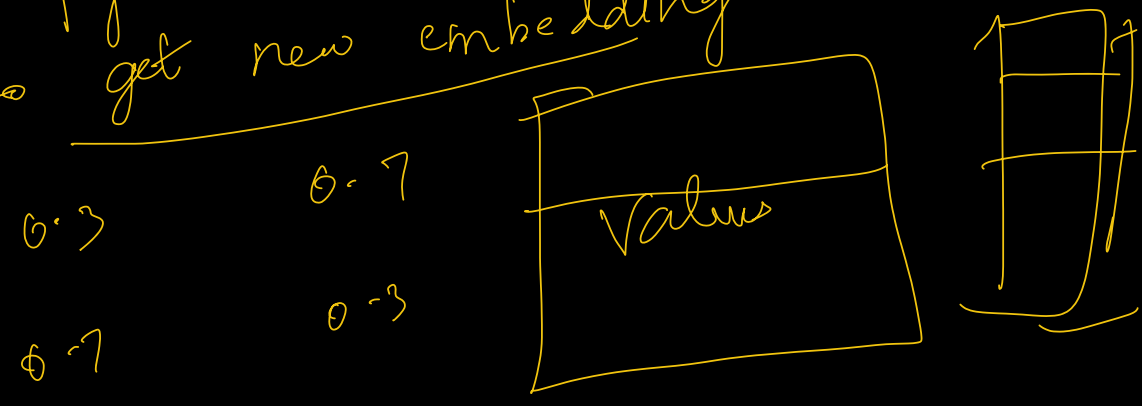
Scaled Dot product $\frac{7.9}{\sqrt{2}} = 6.8$

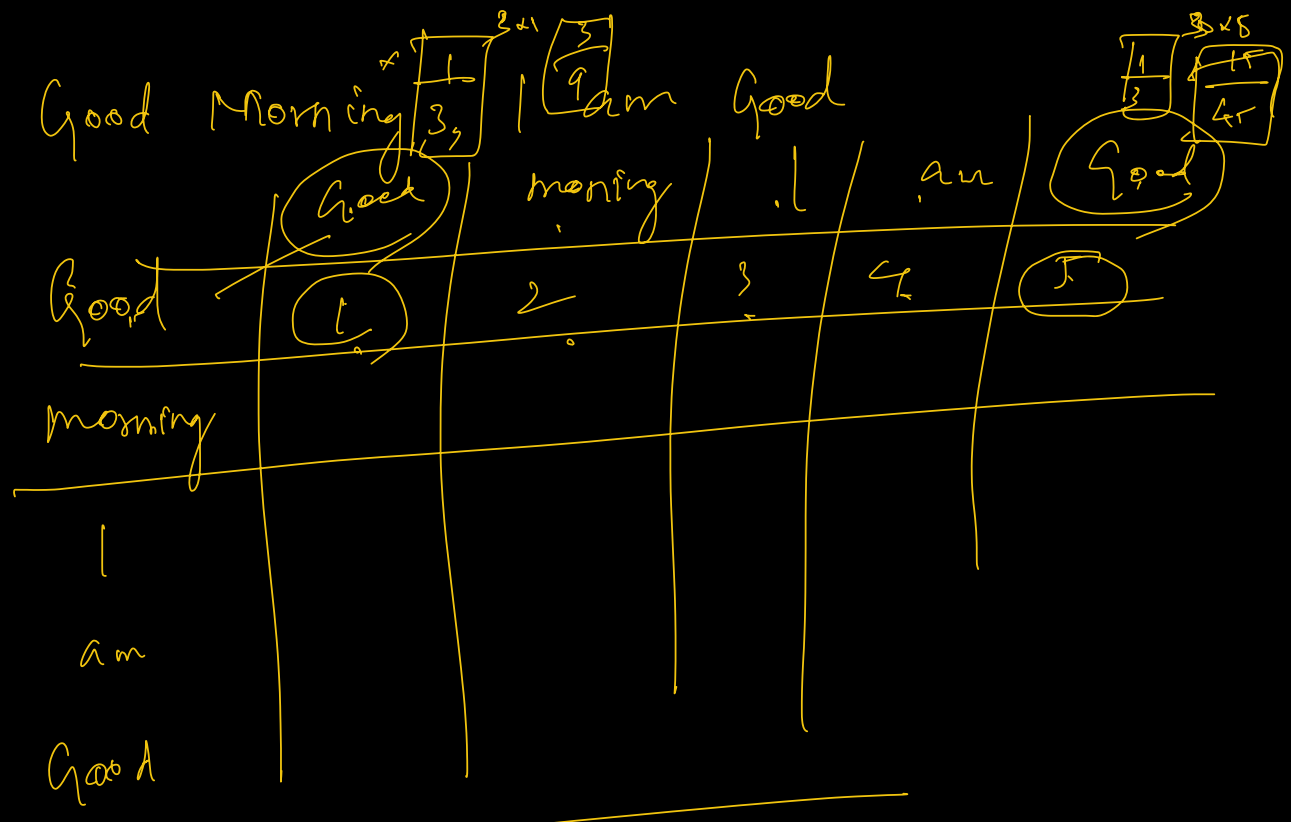




Normalize the similarity with addition of vector becomes (1)

multiply the values to get new embedding





Positional encoding

Odd 1, 3, 5, 7 . . . . cos

even 2, 4, 6, 8 . . . . sin

Odd → $\sin \frac{3 \cdot i}{}$ and Position

even → $\cos \frac{2 \cdot i}{}$

Good