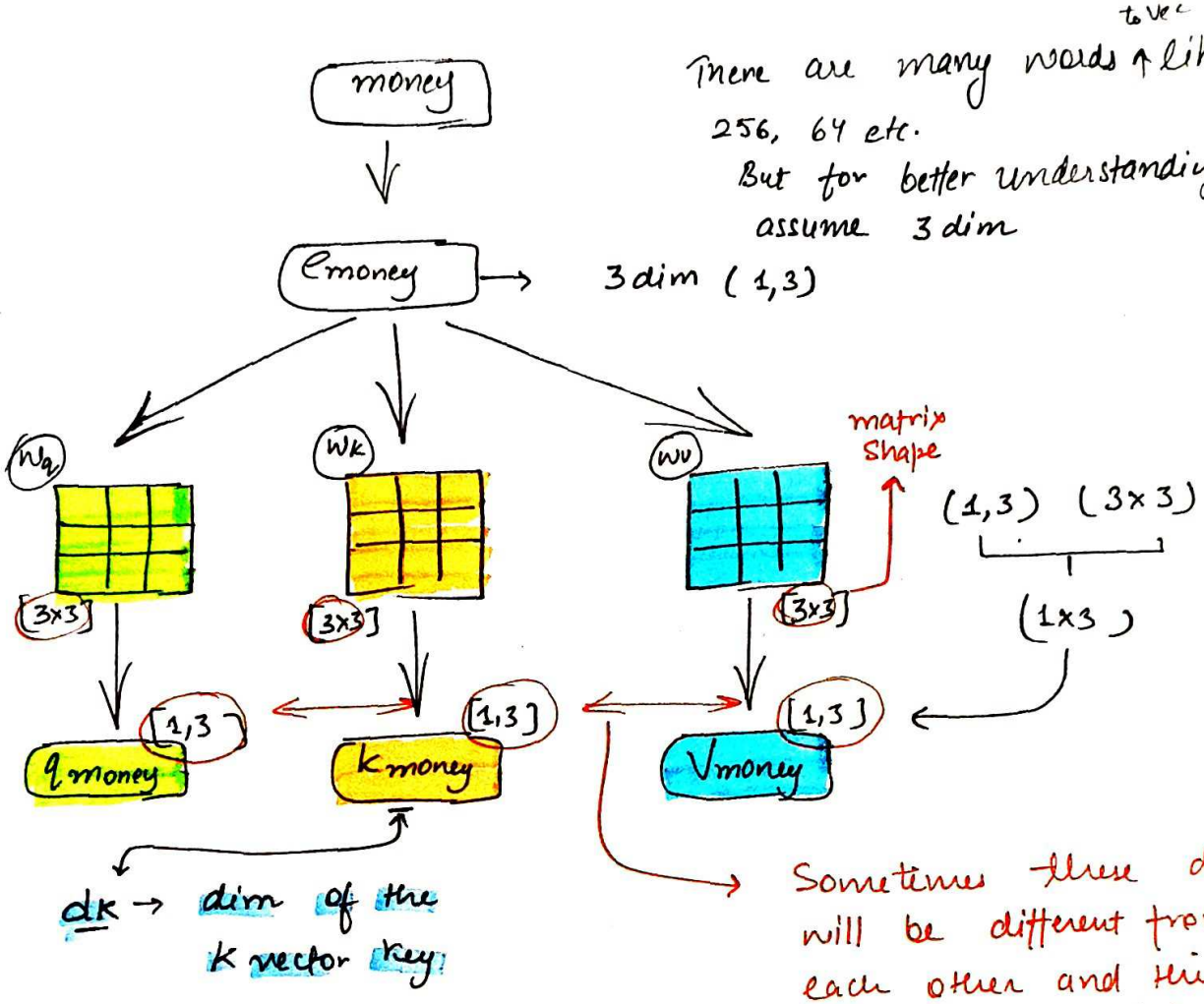
Scaler Dot Product Attention



There are many words ↑ like 512, 256, 64 etc. to vec
But for better understanding we assume 3 dim

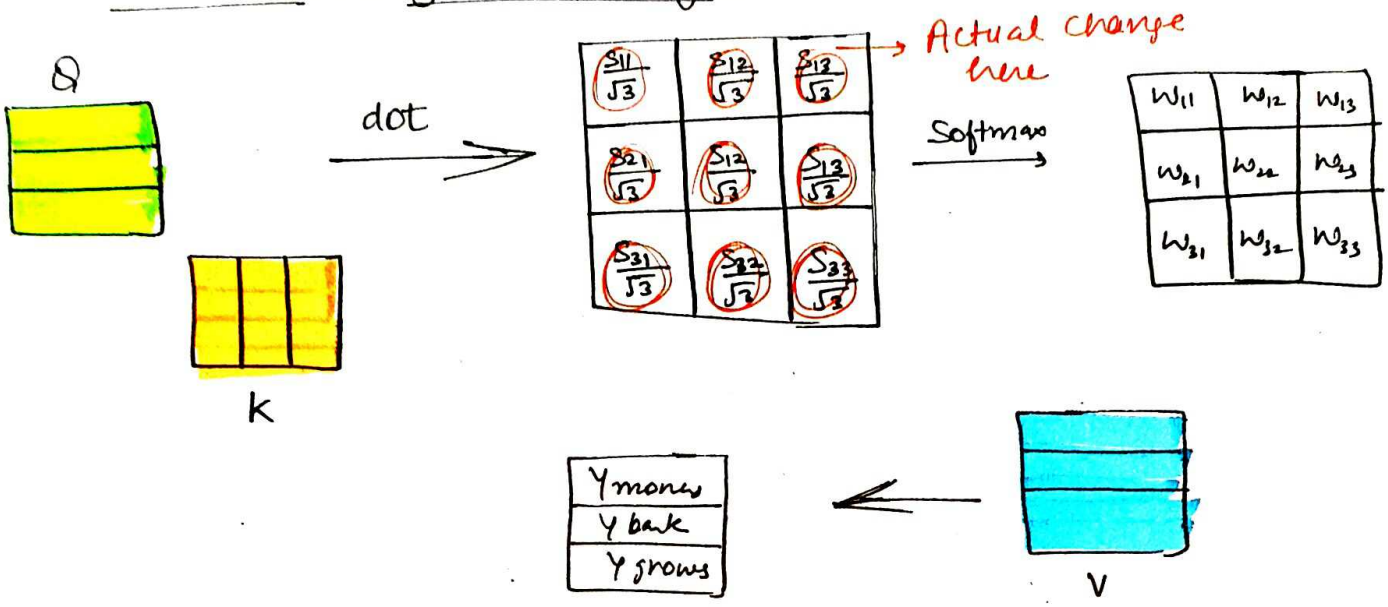
In our case,

$d_k = d_q = d_v = 3$

$$\text{Attention}(Q, k, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$ → Scaled formula

$$\text{Attention}(Q, k, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{3}}\right)V$$ in our case

Actual meaning in diagram



Why?

↳ dot-product ka Nature

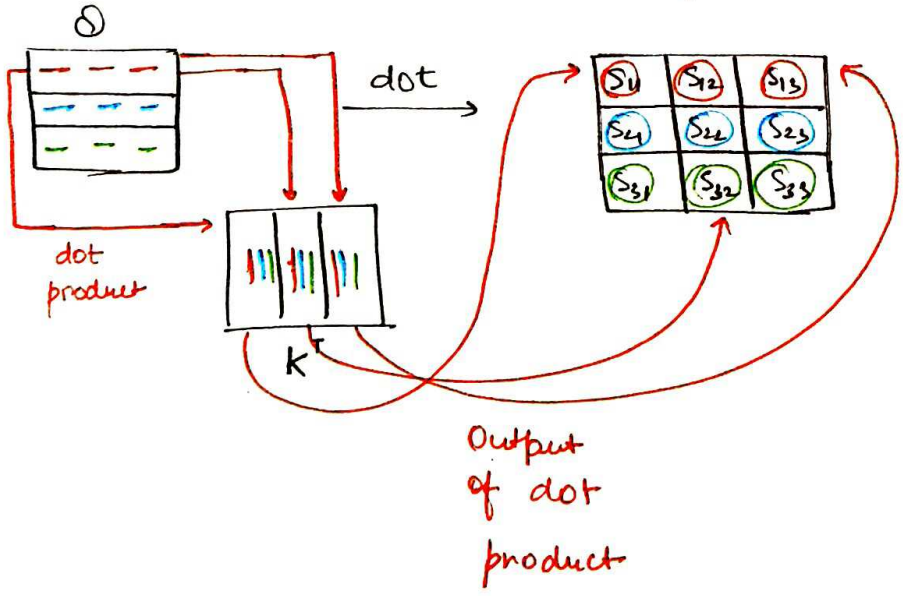
In this formula

$[Q K^T]$

matrix matrix

dot product

How many vectors in this dot product?



Total 9 vector-vector dot products which means we get 9 Number.

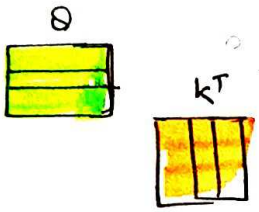
If we have Number then we can easily calculate mean and variance.

Now, what is the meaning of "Dot-product ka Nature"?

(i) low dimension vector → dot product of low dimension vector → is → low variance

(ii) Higher dimension vector → dot product of higher dimension vector → is → Higher variance

Example



Let say,
we have (3) dim vector
$Q(1,3)$
$(1,3)$
k^T
→ (9) dot product values → 3 dim vector is very low dim
then variance of 9 dot product value is also low.

Let say,
we have (512) dim vector → Still we get 9 dot product value
↳ But 512 dim vector is very high so, dot product variance is very high.

Another example

2-D

1) [1,2] [3,2]
2) [ ] [ ]
3) [ ] [ ]
4) [ ] [ ]
5) [ ] [ ]

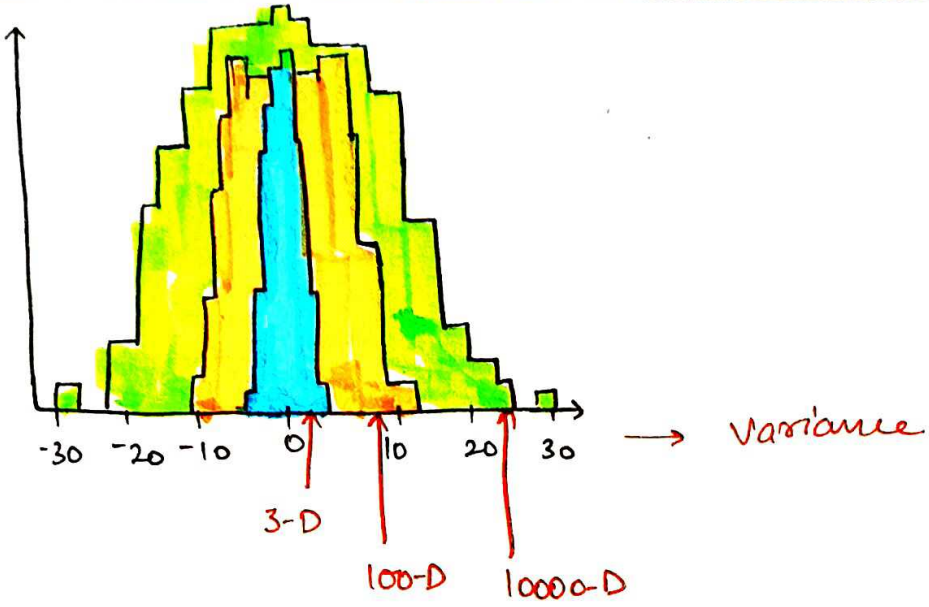
a
b
c
d
e
→ σ_1^2

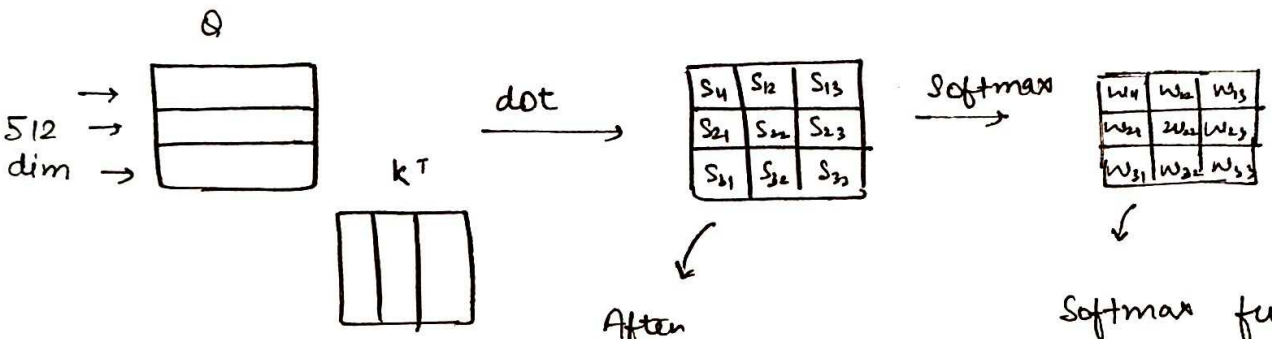
3-D

1) [1,2,3] [3,2,1]
2) [ ] [ ]
3) [ ] [ ]
4) [ ] [ ]
5) [ ] [ ]

f
g
h
i
j
→ σ_2^2

σ_2^2 variance is greater than σ_1^2





After dot product variance will be higher.
(like some values are very high and some value are very low)

Softmax function give small value to small probability and large value to large prob.

during backpropagation gradient descent only focus on large value after Softmax probab and neglect small value.
This will cause vanishing gradient problem.
Traing process → kharab ho jya.

Softmax: [4, 5] low variance

Small probab [0.45, 0.59]

[1, 10] large variance

[0.001, 0.999]
↙ ↘
Small prob large probab.

Real life example :- large variance height student in class. Teacher ask to hand raise for doubt. Because of Large height student small height student's hand not see. Only large Height student doubt solve and small height student's doubt is ignore. So, Training on the basis of large student.

Now, let's imagine. Student's is almost same in the class. And Hand of every student is easily seen by Teacher. Doubt is solved of every student. So, Training on the basis of all student. Variety of question is increase.

Problem :-

S_{11}	S_{12}	S_{13}
S_{21}	S_{22}	S_{23}
S_{31}	S_{32}	S_{33}

} variance is very high

→ softmax

W_{11}	W_{12}	W_{13}
W_{21}	W_{22}	W_{23}
W_{31}	W_{32}	W_{33}

} Probability is (99%) (1%)

max probab ← 99% ; 1% ← min probab.

Vanishing gradient problem. { during backpropagation Gradient descent focus on (99%) and ignore (1%).

One solution is to use low dim vector from starting but in low vector we cannot get good information.

another solution is Scaled the value of doted producted which variance is very high. After scaling variance will be low

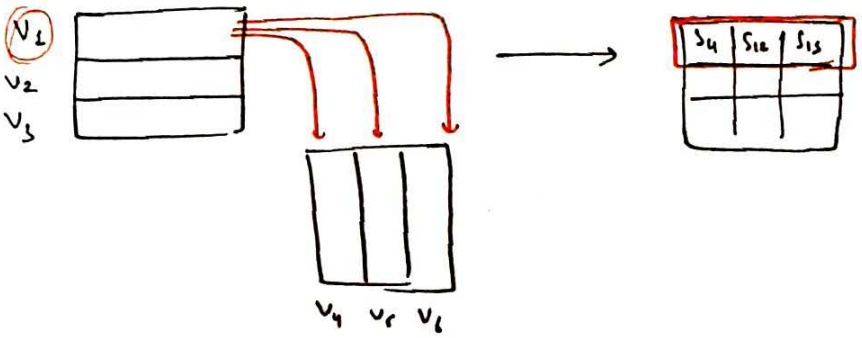
→ (i) [10, 20, 30, 40, 50, 60, 70]
variance is 400.0

(ii) [1, 2, 3, 4, 5, 6, 7]
variance is 4.0

Now, we have to find that scaling factor or number which is divide with value of dotted produ.

If dimension ↑↑ then how much variance ↑↑

In math quantify :



$V_1 \rightarrow V_4$
$\rightarrow V_5$
$\rightarrow V_6$

Let say all these vector are 1-D

1-D

[a] → [b] → ab
[c] → ac
[d] → ad

here we don't want Sample variance

we want expected variance (population Variance) (Var X)

Now Increase dim from 1-D to 2-D

[a b]

[c d] → ac+bd
[e f] → ae+bf
[g h] → ag+bh

Y → Var(Y)

Var(Y) > Var(X)

Now Increase dim from 2-D to 3-D

[a b c]

[d e f] → ad+be+cf
[g h i] → ag+bh+ci
[k l m] → ak+bl+cm

Z → Var(Z)

$Var(z) > Var(y) > Var(x)$

$Var(z) \simeq 3\,Var(x)$

If d dim then $dm \to d\,Var(x)$

* Somehow $\begin{bmatrix} [ac+bd] \\ [ae+bf] \\ \vdots \end{bmatrix}$ ← ise divide karna hai ki every time $Var(x)$ hi aaye

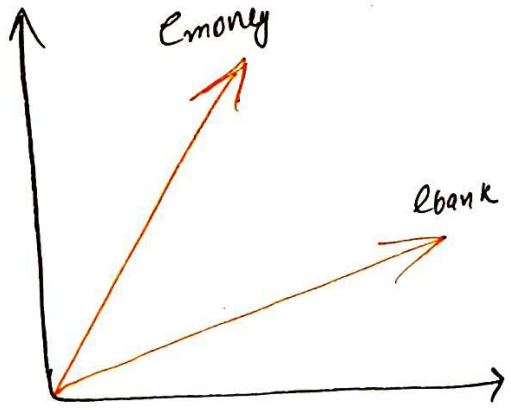
$[a\ b]$ $[c\ d]$ / $\sqrt{2}$; $[e\ f]$ / $\sqrt{2}$; $[g\ h]$ / $\sqrt{2}$ → $\sqrt{dk}$ } $2\,Var(x)$

dim

$\frac{1}{2} Var(y) \to \frac{1}{2} 2\,Var(x) = Var(x)$

$d_k \to$ dimension

Self Attention Geometric Intuition



money bank

Emoney → [2, 7]

Ebank → [9, 3]