Query Matrix

$$\begin{bmatrix} q_{money} \\ q_{bank} \\ q_{grows} \end{bmatrix} \xrightarrow{\text{dot}} \begin{bmatrix} S_{11} & S_{12} & S_{13} \\ S_{21} & S_{22} & S_{23} \\ S_{31} & S_{32} & S_{33} \end{bmatrix}$$

$$\begin{bmatrix} k_{money} & k_{bank} & k_{grows} \end{bmatrix}$$

key^T matrix

softmax ↓

$$\begin{bmatrix} Y_{money} \\ Y_{bank} \\ Y_{grows} \end{bmatrix} \xleftarrow{\text{dot}} \begin{bmatrix} W_{11} & W_{12} & W_{13} \\ W_{21} & W_{22} & W_{23} \\ W_{31} & W_{32} & W_{33} \end{bmatrix}$$

Contextual Embeddings

$$\begin{bmatrix} V_{money} \\ V_{bank} \\ V_{grows} \end{bmatrix}$$

V matrix

(A)

$$\boxed{\text{Attention}(Q,K,V) = \text{sofmax}\left(Q \cdot K^T\right) V}$$

↑

Is the mathematical representation of above process to calculate contextual embeddings.

Scaled Dot-Product :-

* Equation (A) is developed from scratch as you noticed guys! But when we headed to the reasearch paper "Attention All you Need". The formulation Mentioned is:

$$\boxed{\text{Attention}(Q,K,V) = \text{sofmax}\left(\frac{Q \cdot K^T}{\sqrt{d_K}}\right)V}$$

1. why they divided with $\sqrt{d_K}$?
2. What is d_K ?
3. How d_K is calculated ?

2A:- d_K is the dimension of key vector

1A:- To solve the "unstable Gradient" problem we divide with $\sqrt{d_K}$

I have two Questions running in my mind!

1. What is Unstable Gradient ?
2. Why we have divided with only $\sqrt{d_K}$, why not with d_K, d_Q and d_V ?

1. Let me answer what is Unstable gradient and vanishing gradient!

Ans:- What If I say:

* Low Dimensional Vector $\propto$ Low Variance

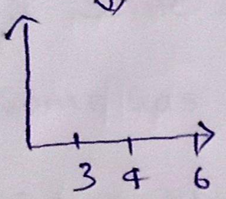
* High Dimensional Vector $\propto$ High Variance

I think your face might be : ☹

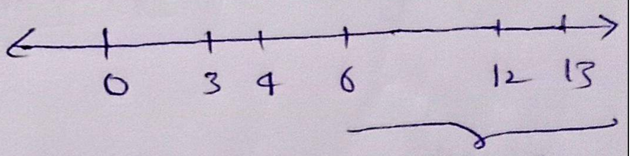
Let me explain.....

For example:- you have 3 sets of vectors

In 2D	In 3D
1. $[1,2]\cdot[1,1] = 3$	1. $[1,2,3]\cdot[1,1,1] = 6$
2. $[2,1]\cdot[1,2] = 2\cdot1+1\cdot2 = 4$	2. $[2,1,3]\cdot[1,2,3] = 13$
3. $[2,2]\cdot[1,2] = 2\cdot1+2\cdot2 = 6$	3. $[2,2,2]\cdot[1,2,3] = 12$



x = Spread



y = Spread

clearly, $y > x$

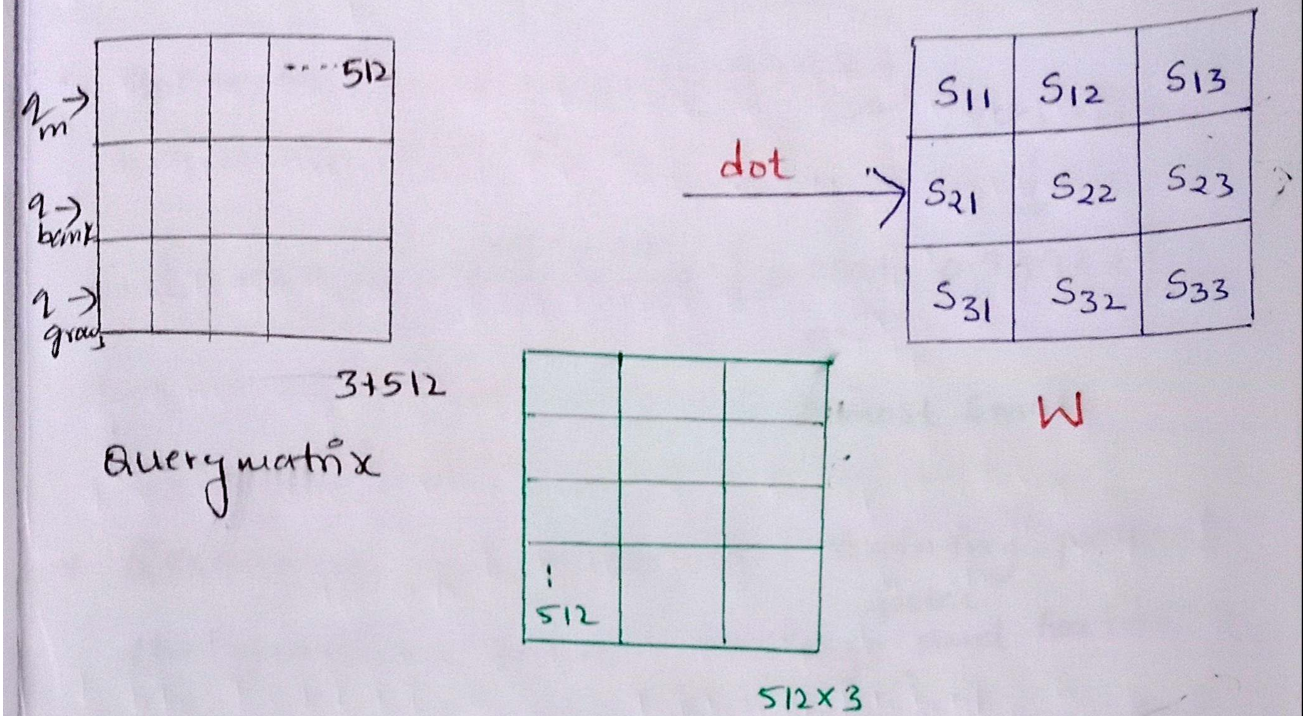
we know Variance is proportional to spread

$\Rightarrow Var(y) > Var(x)$

observe $\Rightarrow$ $dim(y\text{-case}) = 3 > dim(x\text{-case}) = 2$

* Think dimension as no. of elements in a vector

* Lets say the dimension of Query and key vector is 512



* Since over dimension is huge (512), then the values S11 ---- S33 in W has more variance. It means some among them in W have large value and some has low value.

* Now the problem is with softmax due to high variance data.

What is softmax?

input (x) ⟶ $e^{\boxed{x}}$ ⟶ probabilities

Let say we have [1,10] → we have to convert into probabilities [we have to use softmax]

$$1 \longrightarrow \cancel{\frac{e^1+e^{10}}{e^1}} = \frac{e^1}{e^1+e^{10}} = 0.0001$$

$$10 \longrightarrow \frac{e^{10}}{e^1+e^{10}} = 0.99987$$

Very high prob

$$\therefore [1, 10] \xrightarrow{\text{softmax}} [0.0001, 0.99987]$$

Almost small

* Because of this during the training process the gradient will not converge faster and training becomes very slow! for larger values. And for the smaller values the gradient allmost vanishes and smaller values will not update atall !

* Why we are having this problem?

It is because of high variance data! which is the consequence of High dimension(512).

But we can't avoid High dimension because embeddings having high dimension capture Semantic meaning with other words. Now what is our goal?

Goal: Irrespective of any dimension, variance should be almost similar!

Sol^n: what I am trying say is:

If it is 64-dimension and variance is 'x'

For 128 also $\approx x$

512 also $\approx x$

How can we achieve this?

The obvious answer is Normalization!

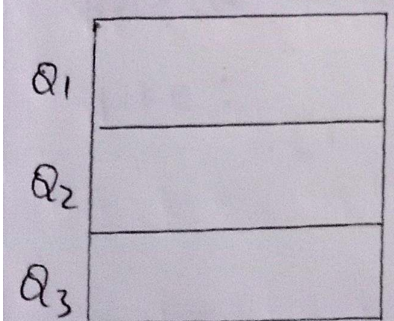
Let's take example:

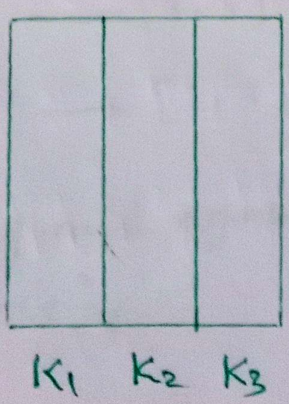
our problem with is with W matrix only right which contains high variance data.

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

[W]

* Let's focus on only 1st row, then we scale up our concept to complete W.





* Now we are focusing on variance of 1st row of W only.

Assume Q_1 is 1dim & K_i also 1 dim

$$Q_1 \longrightarrow K_1 \qquad [a] \text{——} [b] \Rightarrow S_{11} = ab$$
$$\quad \longrightarrow K_2 \qquad \text{——} [c] \Rightarrow S_{12} = ac$$
$$\quad \longrightarrow K_3 \qquad \text{——} [d] \Rightarrow S_{13} = ad.$$

S

* Think S_{11}, S_{12}, S_{13} are sample and W is population we want Var(W) (predict) based on sample!

Let assume ab, ac, ad are coming from a random variable 'x'. Lets consider the variance of S is var(x)

* What if the vectors dim = 2?

$$\left.\begin{aligned} [a\ b] \text{——} [c\ d] &\Rightarrow S_{11} = ac+bd \\ \text{——} [e\ f] &\Rightarrow S_{12} = ae+bf \\ \text{——} [g\ h] &\Rightarrow S_{13} = ag+bh \end{aligned}\right\}$$

Here we want for a general vector variance, not like:

$$\left.\begin{aligned} [1\ 1] \text{——} [1,2] &\Rightarrow 1+2=3 \\ \text{——} [1,3] &\Rightarrow 1+3=4 \\ \text{——} [2,1] &\Rightarrow 2+1=3 \end{aligned}\right\} \text{This is Sample variance}$$

Say ac+bd, ae+bf and ag+bh are coming from random variance 'y'.

* Now variance of 1st row is $Var(y)$

Is this relationship $Var(y) > Var(x)$ is True?

Ans: Is yes! bcz since high dimension data have high variance right!

Also roughly we can say $Var(y) \simeq 2\,Var(x)$

IF 3d?

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

$$\& \quad Var(z) > Var(y) > Var(x)$$

* If d-dimension?

$$\text{Variance} \simeq d\, \underbrace{Var(x)}_{dim=1}$$

$\therefore$ We figured out a linear relationship between variance and dimension

Do you agree!! ☺

$\therefore$ We have mathematically quantified the relation: "If dim $\uparrow$es $\longrightarrow$ Variance $\uparrow$es"

Summary:-

	problem	Ideally we want
1 dim $\longrightarrow$	$Var(x)$ $\longrightarrow$	$Var(x)$
2 dim $\longrightarrow$	2 $Var(x)$ $\longrightarrow$	$Var(x)$
3 dim $\longrightarrow$	3 $Var(x)$ $\longrightarrow$	$Var(x)$
⋮	We have to normalize these	
d dim $\longrightarrow$	d $Var(x)$ $\longrightarrow$	$Var(x)$

* I hope you understood the problem! ☺

Mathematical Rule:- If you have variable X with a variance of Var(x), and you create a new variable 'Y' by scaling 'x' with a constant 'c', so that $Y = cX$, the variance:

$$\boxed{Var(Y) = c^2 \, Var(x)}$$

It says...

If you have a random variable 'x' whose Variance is Var(x).

$$X \longrightarrow Var(x)$$

You defined a new variable $Y = cX$

Then $Var(y) \longrightarrow c^2 Var(x)$

1 dim $\longrightarrow$ $x = Var(x)$

2 dim $\longrightarrow$ $y \neq \cancel{2} x \longrightarrow Var(y) =$

$y = 2\,Var(x)$

$y' = \frac{y}{\sqrt{2}} \longrightarrow Var(y') = \left(\frac{1}{\sqrt{2}}\right)^2 y$

$= \frac{1}{\cancel{2}}(\cancel{2})(Var(x))$

$= Var(x)$

$\therefore$ If we normalize $\frac{y}{\sqrt{2}}$ then variance be_came $Var(x)$

3 dim $\longrightarrow$ $\frac{3\,Var(x)}{\sqrt{3}} \longrightarrow \frac{1}{\cancel{3}}(\cancel{3})\,Var(x)$

4 dim $\longrightarrow$ $\frac{4\,Var(x)}{\sqrt{4}} \longrightarrow \frac{1}{\cancel{4}}(\cancel{4})\,Var(x)$

$\vdots$

d dim $\longrightarrow$ $\frac{d\,Var(x)}{\sqrt{d}} \longrightarrow Var(x)$

$\therefore$ Irrespectiue of dimeusion variance become rouguly similar !!

$\therefore$ I hope you understood why we are normalizing with square-root and if you are clearly

observe the red circles we are dividing
with dimension itself.

** And that dimension should be dimension of K-vector because we are doing the dot-product with Key-Vector so we have to take d_k.

* ∴ We updated our First-principle approach by scaling with $\frac{1}{sqrt(d_k)}$

Let me summarize...

Next page

Summary - Self Attention

Q (3,n)

K (n,3)

$$Q \cdot K \xrightarrow{\text{dot}} \begin{bmatrix} s_{11} & s_{12} & s_{13} \\ s_{21} & s_{22} & s_{23} \\ s_{31} & s_{32} & s_{33} \end{bmatrix} \quad (3\times3)$$

$$\downarrow \text{scale } \frac{1}{\sqrt{d_k}}$$

$$\begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \end{bmatrix} (3\times3) \xleftarrow{\text{softmax}} \begin{bmatrix} s'_{11} & s'_{12} & s'_{13} \\ s'_{21} & s'_{22} & s'_{23} \\ s'_{31} & s'_{32} & s'_{33} \end{bmatrix} (3\times3)$$

$$\downarrow$$

$$\begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \end{bmatrix} \xrightarrow{\text{dot}} \begin{bmatrix} Y_{money} \\ Y_{bank} \\ Y_{grows} \end{bmatrix}$$

3 x 3

V-matrix 3xn

I am Very happy

Congrats !!