10 Mar 2025

SCALED DOT PRODUCT ATTENTION

- in the research paper "Attention Is All You Need" the attention is

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

- in denominator term we use it for handle the training unstability

Here, d_k – dimensionality of the K(key) vector

low dimensional vector ⟶ dot product ⟶ low variance

high dimensional vector ⟶ dot product ⟶ high variance

example

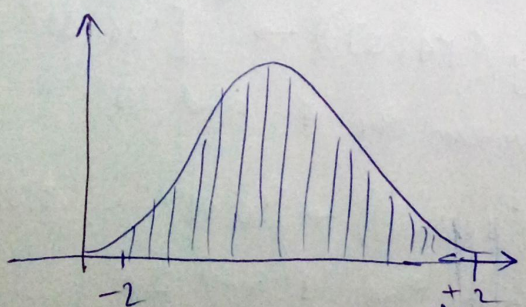
$[1,2] \cdot [2,3] = a$
$[1,4] \cdot [5,6] = b$
$[8,9] \cdot [10,2] = c$

variance = σ_1^2

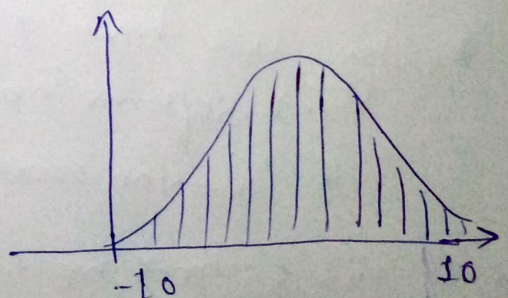
$<$

$[1,2,3] \cdot [3,2,1] = f$
$[3,4,5] \cdot [6,7,8] = g$
$[10,9,11] \cdot [8,7,6] = h$

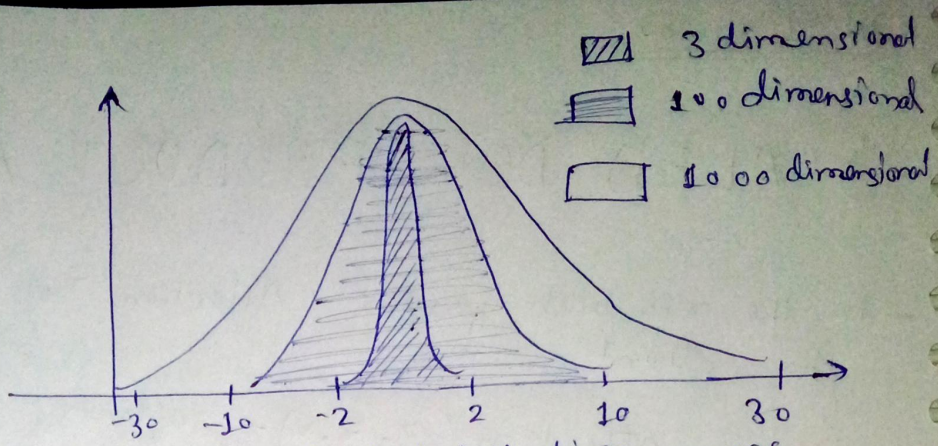
variance = σ_2^2



each vector have ~~three~~ dim



each vector have 100 dim



comparison of dot product distribution across dimensions.

- Here we notice that vector's dimension increases then variance increases.

* if the variance is very high then it create a problem. ~~फिर क्या~~ variance ?

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

3 x 3

- the above matrix का variance.

* why we got high variance of above matrix?

because of number of dimension in vectors (K(key), V(value), Q(query)) is high.

* why we have a problem with high variance?

(i) softmax(np.array([4, 5])) → ['26.89%', '73.11%']

(low variance between the '4' and '5'

(ii) softmax(np.array([1, 10])) → ['0.01%', '99.99%']

high variance between the '1' and '10'

- Hence the variance is very high then the probability difference between the 'W' matrix is high so during training and backpropagation all focus is align toward the high probabilities and low probabilities get ignored.
- and due to these low probabilities we can face the vanishing gradient problem. due to that parameter in 'W' matrix ~~not~~ negligibly update.

* So what we do to decrease the variance of 'S' matrix?
simple scale that matrix (divide each value with. $\sqrt{d_k}$)

{ if we have a random variable X with a variance of Var(X), and you can create a new variable Y by scaling X with a constant c, so that Y = cX, the variance of Y (Var(Y)) is related to the variance of X by the square of the scaling factor c. mathematically, this relationship is expressed as:

$$Var(Y) = c^2 Var(X)$$ }

architecture of Scaled dot product Attention: