13 Mar 2025

POSITIONAL ENCODING IN TRANSFORMERS

- advantages of self attention
  - it can generate the contextual embedding of a given word
  - processing the each word in parallel.

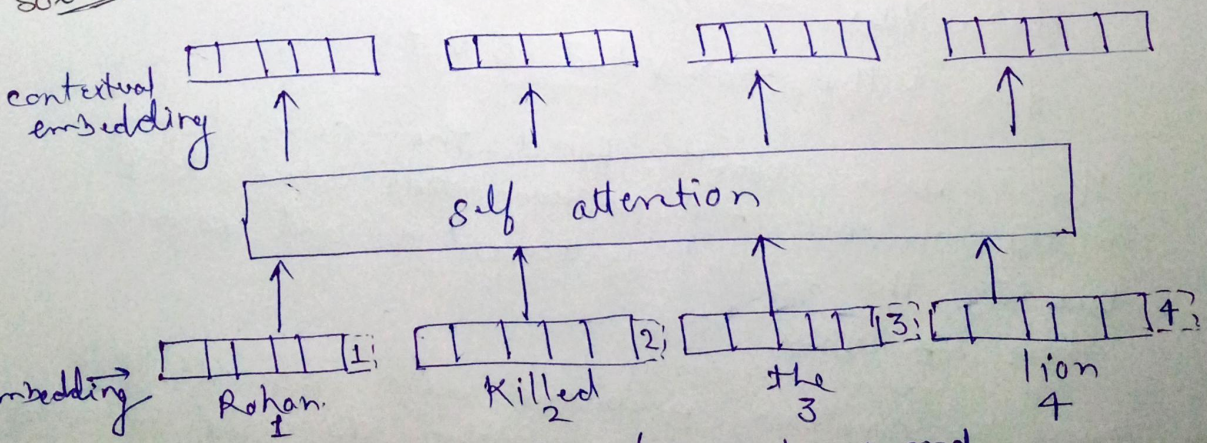
- disadvantages of self attention.
  - we can't capture the order of word in the sentence example.

"Rohan killed the lion" — sentence:1
"lion killed the Rohan" — sentence:2

But this problem we can't face in RNN

Hence we to add the positional encoding logic self attention to solve this problem.

Solution-1



- provide count number for each word.
- add number in the end of the embedding.
- But the problem for this approach is not work well for large number of word (book)

it work well if the counting range is from -1 to 1

Solution-2

- in the above approach we can modified it and make sure that the count must be in the range of -1 to 1
- this could be done with multiply with the number of word in the sentence. example

Sentence → Rohan killed the lion

Position → $\frac{1}{4}$ $\frac{2}{4}$ $\frac{3}{4}$ $\frac{4}{4}$

- But this approach has an also a problem example given sentences are from training data

Thank you ——— (i)

$\frac{1}{2}$ $\frac{2}{2}$

Rohan killed the lion ——— (ii)

$\frac{1}{4}$ $\frac{2}{4}$ $\frac{3}{4}$ $\frac{4}{4}$

for "you" → value is 1

for "killed" → value is 0.5

Here we can't find the consistency. in (i) sentence 2nd position has value is 1 but in the (ii) sentence the 2nd position has value 0.5

* there is another problems with above two approach is that it uses the discrete numbers (1, 2, 3 ...), this not good for Neural network (it prefers the smooth transitions using continuous numbers)

gradient flow becomes unstable.

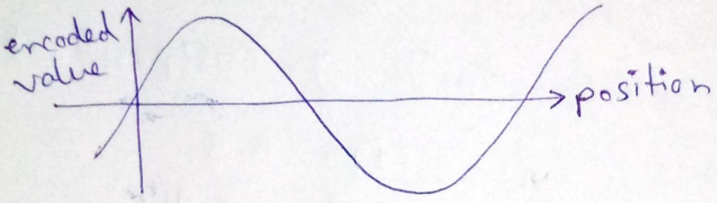
* one another problem is that we can't capture the relative positioning of each word

problems
- unbounded
- discrete
- relative positioning

Solution.
- bounded
- continuous.
- periodic function. to capture the relative ~~positing~~ positioning

Solution-3 use the trigonometric function (sin function)

- The sine function as a ~~solution~~ solution



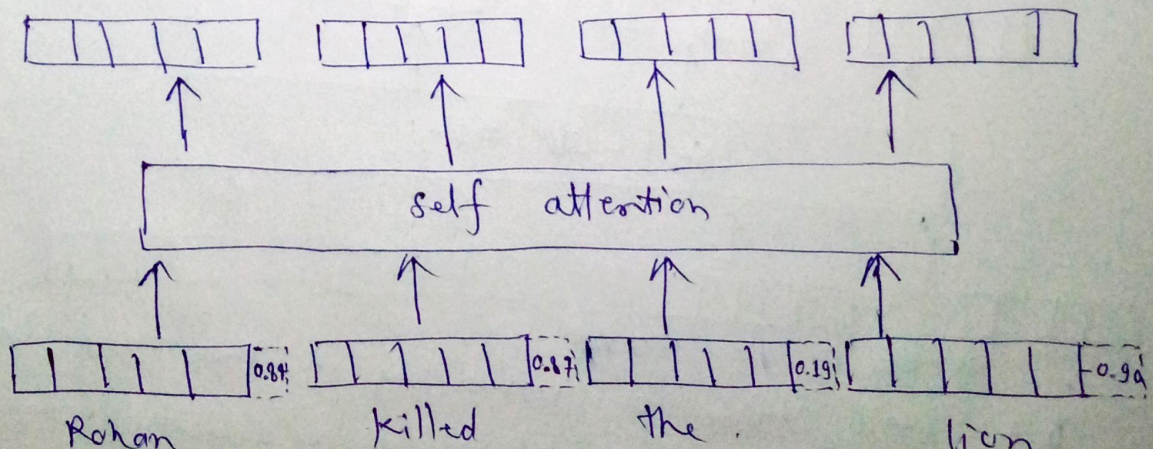
$y = \sin(\text{position})$

here y is encoded value.

example. "Rohan killed the lion"
1 2 3 4

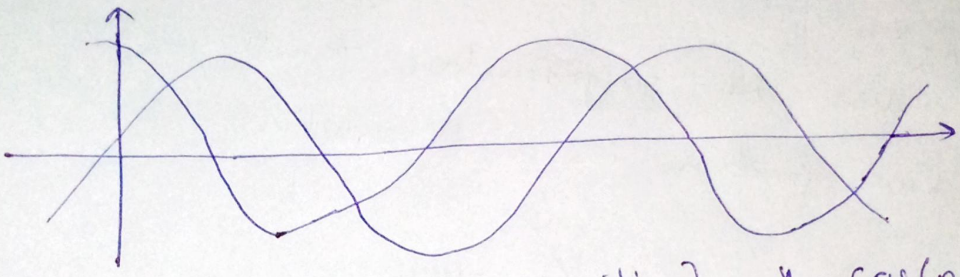
$y = \sin(1) = 0.84$ $y = \sin(2) = 0.87$

$y = \sin(3) = 0.19$ $y = \sin(4) = -0.90$



- but there is one problem with above approach is periodicity (two words has different position but have positional value) due sin function nature.

- to solve the above problem (periodicity) by using the trigonometric function (sin() and cos())



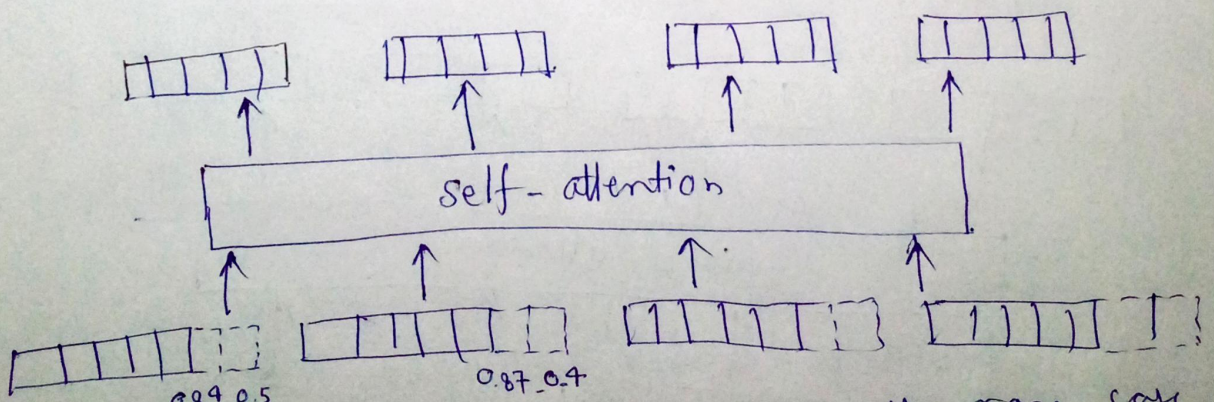
y = sin(position) y = cos(position)

y = sin(1) = 0.84
y = sin(2) = 0.87
y = sin(3) = 0.19
y = sin(4) = -0.90

y = cos(1) = 0.5
y = cos(2) = -0.4
y = cos(3) =
y = cos(4) =

Here each word represent as a vector.

Rohan → [0.84 | 0.5]



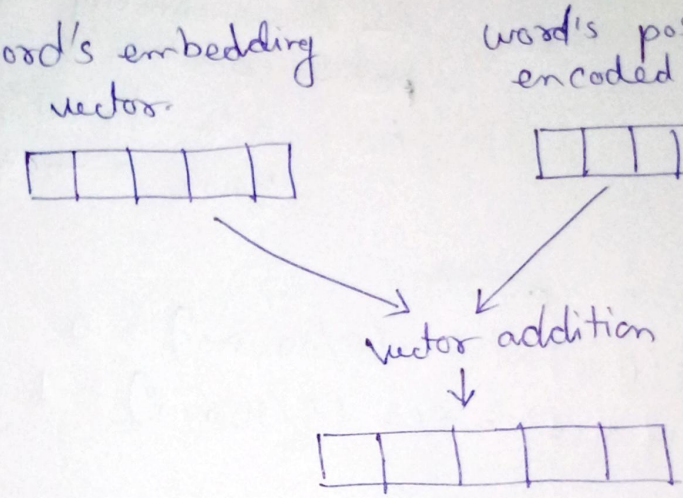
- But still have a posibility in the rare case that we got a problem (periodicity), to resolve this add more sin() and cos() functions.

$y = \sin(\text{position})$ $\quad$ $y = \cos(\text{position})$

$y = \sin(\text{position}/2)$ $\quad$ $y = \cos(\text{position}/2)$

- The size of the positional encoded vector is equal to the size of embedding of given word

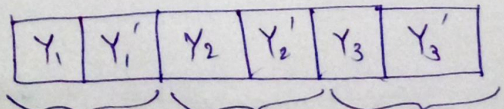
- in the research paper "Attention Is ALL You Need"



Here we can't ~~enter~~ concatenate the both vector we choose the vector addition between them.

How positional encoded vector calculate :

- let's take an example "River Bank"
  1 2
- Here we use the 3 different sin() and cos() function use.



$Y_1 = \sin(1)$ $Y_2 \sin(1/2)$ $Y_3 = \sin(1/3)$

$Y_1' = \cos(1)$ $Y_2' = \cos(1/2)$ $Y_3' = \cos(1/3)$

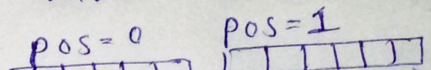
- Now same process for other word "Bank".
- in original research paper 256 pair of sin() and cos() function is used.

- But how at which frequency we change the sin() and cos() function.

$$PE(pos, 2i) = \sin\left(pos / 10000^{2i/d_{model}}\right)$$

$$PE(pos, 2i+1) = \cos\left(pos / 10000^{2i/d_{model}}\right)$$

take an example: "River bank"



Here d — dimensionality of embedding = 6
pos — position.
i — 0 to $d_{model}/2$
Here 0 to 3 (excluded)

pos = 0

for i = 0

$PE(0,0) = \sin(0/10000^{\circ}) = 0$

$PE(0,1) = \cos(0/10000^{\circ}) = 1$

for i = 1

$PE(0,2) = \sin(0/10000^{1/3}) = 0$

$PE(0,3) = \cos(0/10000^{1/3}) = 1$

for i = 2

$PE(0,4) = \sin(0/10000^{2/3}) = 0$

$PE(0,5) = \cos(0/10000^{2/3}) = 1$

pos = 1

for i = 0

$PE(1,0) = \sin(1/10000^{\circ}) = 0.84$

$PE(1,1) = \cos(1/10000^{\circ}) = 0.54$

for i = 1

$PE(1,2) = \sin\left(1/10000^{1/3}\right) = 0.04$

$PE(1,3) = \cos\left(1/10000^{1/3}\right) = 0.99$

for i = 2

$PE(1,4) = \sin\left(1/10000^{2/3}\right) = 0.00$

$PE(1,5) = \cos\left(1/10000^{2/3}\right) = 0.99$

River →	0	1	0	1	0	1

bank →	0.84	0.54	0.04	0.99	0.00	0.99