Self-Attention

River Bank →

Money Bank →

Benefit

Parallel → drawback

words order

Nitish Killed Lion

Lion   killed   Nitish

RNN → sequential

Nitish killed lion

Unfold

River bank river   bank
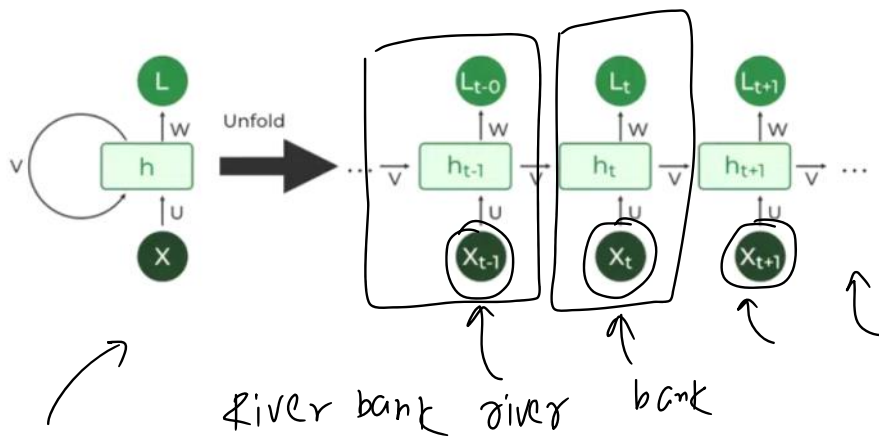
Proposing a simple solution

Smooth transition   continue
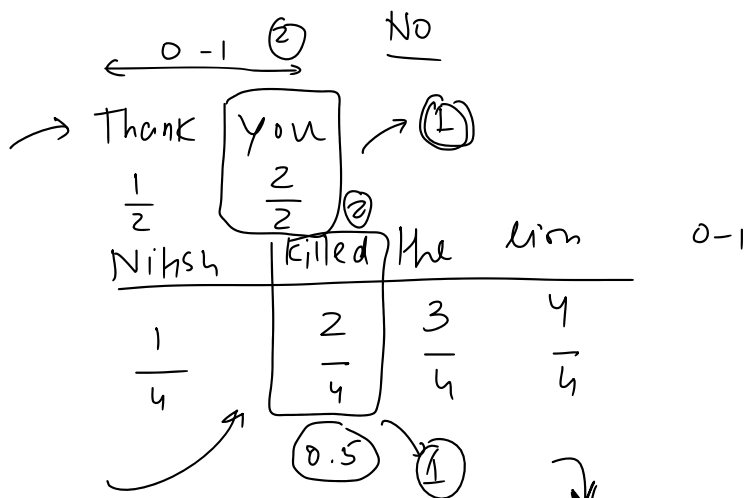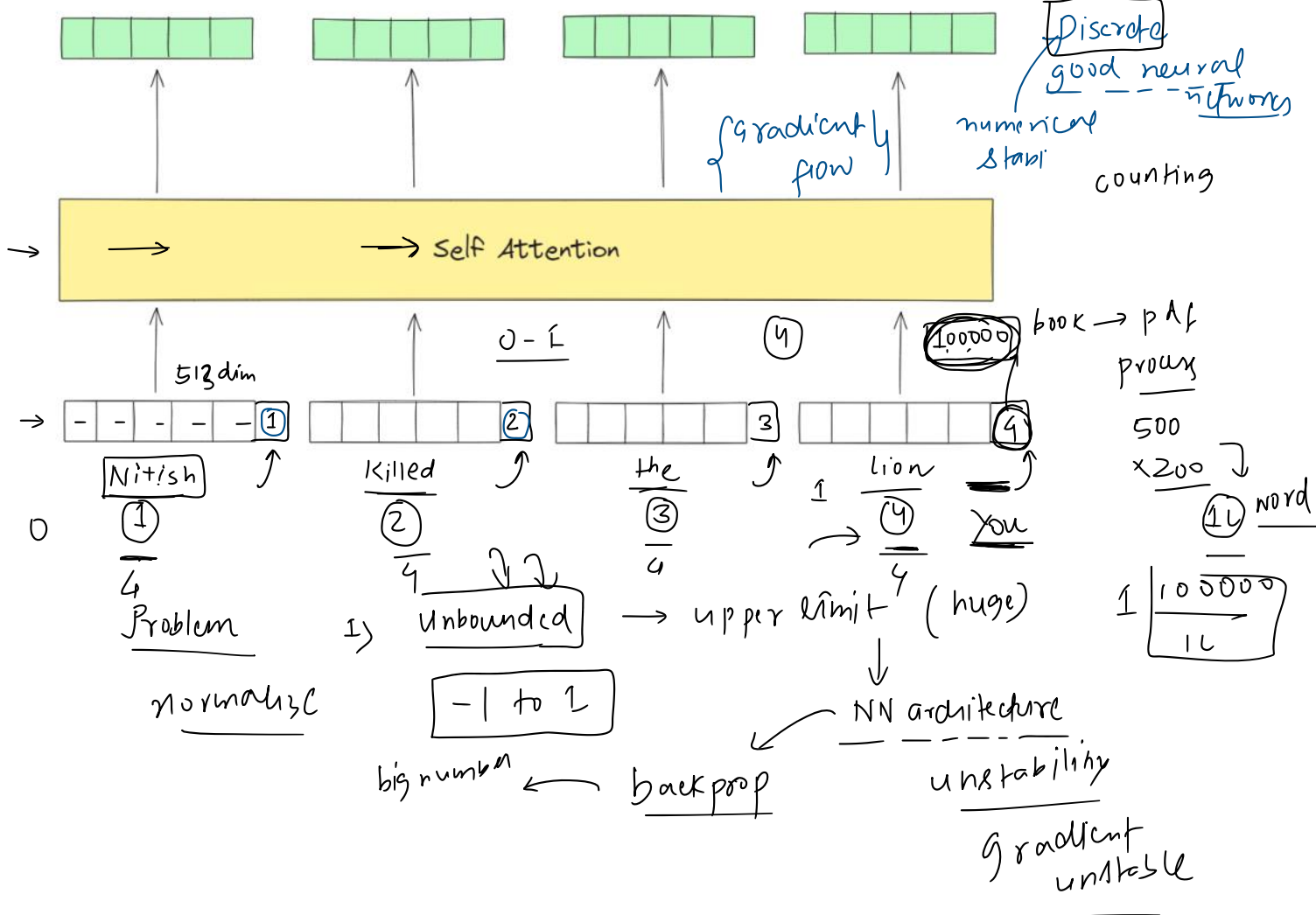
Problem 2
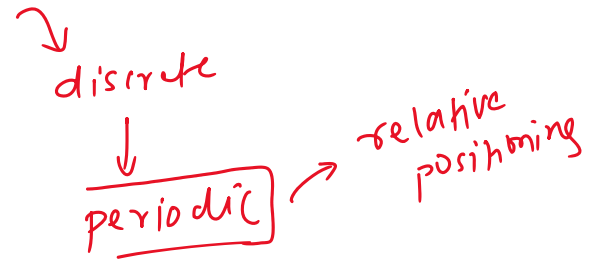
Discrete

good neural network

numerical stabl

counting

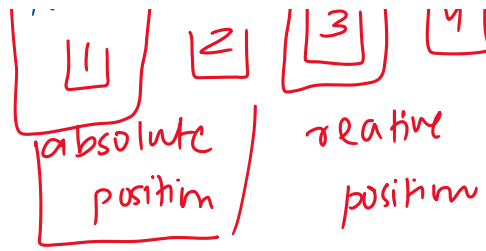Self Attention

gradient flow

book → pdf

proxy

500
×200

word

0 - L

4

100000

513 dim

| - | - | - | - | - | ① |
| Nitish | | | | | |

| | | | | ② |
| Killed | | | | |

| | | | 3 |
| the | | | |

| | | | ④ |
| lion | | | |

0

①
—
4
Problem

normalize

②
—
4

I) Unbounded → upper limit (huge)

−1 to 2

big number ← backprop

③
—
4

1 → ④
—
4
You

NN architecture

unstability

gradient unstable

1 | 100000
| 1 L

$1 \frac{100000}{1L}$

← 0 -1 ②   NO

→ Thank You → ①
   $\frac{1}{2}$ $\frac{2}{2}$ ②

Nitsh Killed the lion   0-1

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

0.5 ①

relative positioning

• — · — ● —

Problem 3 → can't   relative positioning

3 - 1 = 2

Nitish Killed the lion
  1   2   3   4

discrete

| 1 | 2 | 3 | 4 |

absolute position | relative position

discrete
↓
periodic → relative positioning
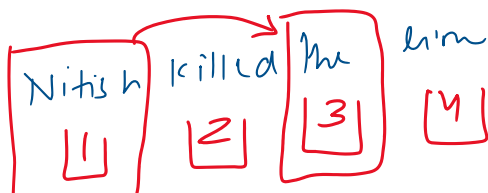
trigonometric
↓
sine
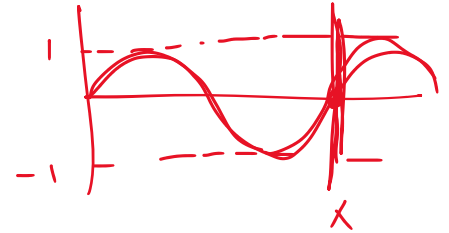
Problem
→ unbounded (bounded) ✓✓
→ discrete (continuous) ↘ function →
→ relative (periodic) ✓
   positn

positional → sine ⌐
encoder   ↓
            better
            solution

The sine function as a solution

periodic   [Bis problem]. $y = \sin(\text{pos})$

45

encoded value

position

35

$\begin{cases} \checkmark & 1) \quad \text{unbounded} \\ \checkmark & 2) \quad \text{discrete} \\ \checkmark & 3) \quad \text{relational} \\ & \qquad \text{positioning} \end{cases}$

encoded value

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

$y = \sin(1) = \boxed{0.84}$

$y = \sin(2) = \boxed{0.87}$

$y = \sin(3) = \boxed{0.14}$

$\qquad y = \sin(4) = \boxed{-0.90}$

Self Attention

0.84   0.87

Nitish   Killed   the   Lion
              (2)   3   4

0.11 $\longrightarrow$ 0.11
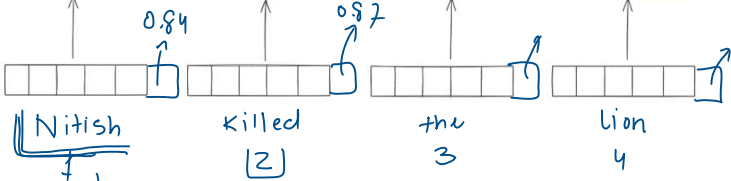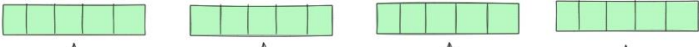
$y = \sin(1) = 0.84$ $y = \cos(1) = 0.5$

$y = \sin(\text{pos})$ $y = \cos(\text{pos})$

encoded value
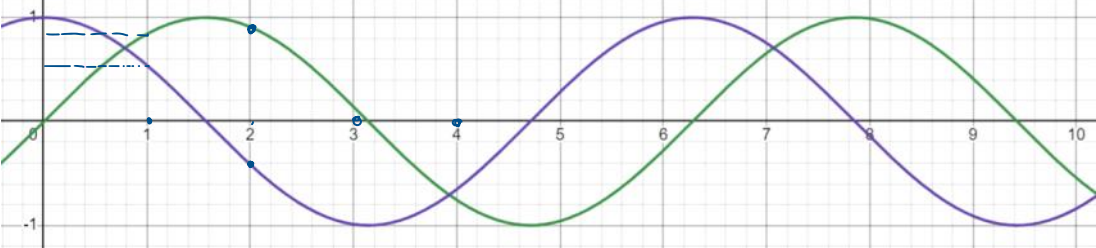
position

Self Attention
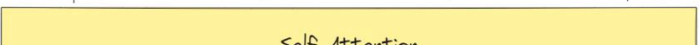
Nitish $\rightarrow$ [0.84, 0.5]

$y = \sin(2)$ $y = \cos(2)$

[0.87, -0.4]

scalar $\rightarrow$ vector

$[0.87, -0.4]$ ~

scalar → vector

Self Attention

0.84   0.5

1 Nitish   2 killed   3 the   4 lion

→ vector $[\_,\_,\_,\_]$ (4 numbe)
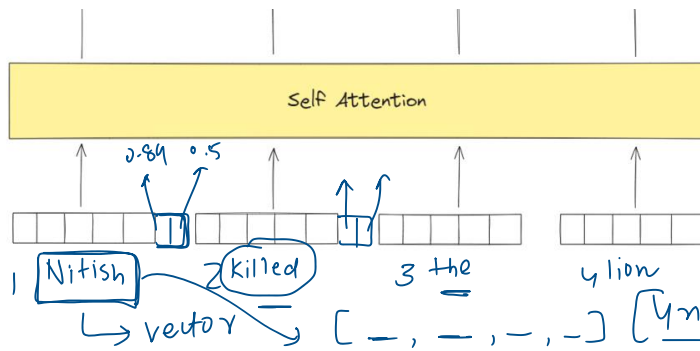
$y = \sin(pos)$ $y = \cos(pos)$

$y = \sin(pos/2)$ $y = \cos(pos/2)$
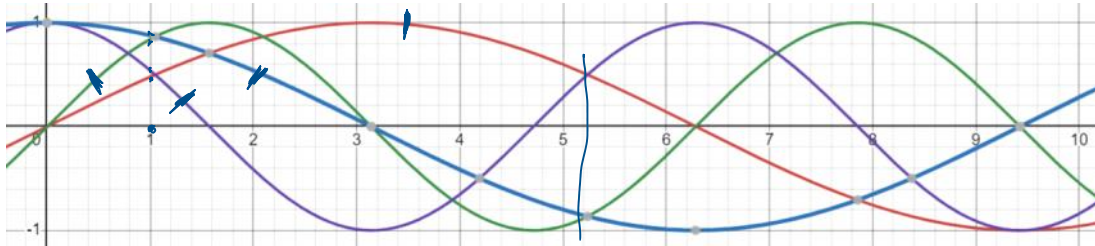
posit 6 number vector

$y = \sin(pos/3)$ $y = \cos(pos/3)$

frequency →

Self Attention

8 dim

512 dim

| 1 | 2 | 3 | 4 | 5 | 6 |
| 1 | 2 | 3 | 4 | 5 | 6 |

River Bank   vector → dim

| 1 | 2 | 3 | 4 | 5 | 6 |

| 1 | | | | | |

embd   +   pos encod
(6 dim)   (6 dim

= 6 dim vector

concatenation → addition

↳ 6 dim 6 dim

↳ 12 dim → parameters → training
increase   double
up

6 dim → 2

River → pos 1

positional encoding

Sine Cosine

$\frac{512}{2}$

positional vector

pos = 0   pos = 1

→ River   bank

| 1 | 2 | 3 | 4 | 5 | 6 |

| 0 | | | | | |

2i + 1

for i = 0 → pos = 0

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

$PE(0, \underline{1}) = \cos(\underline{0}/10000^{\underline{0}}) = \boxed{1}$
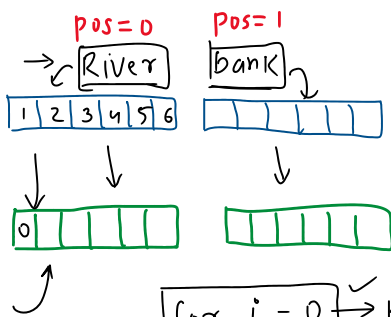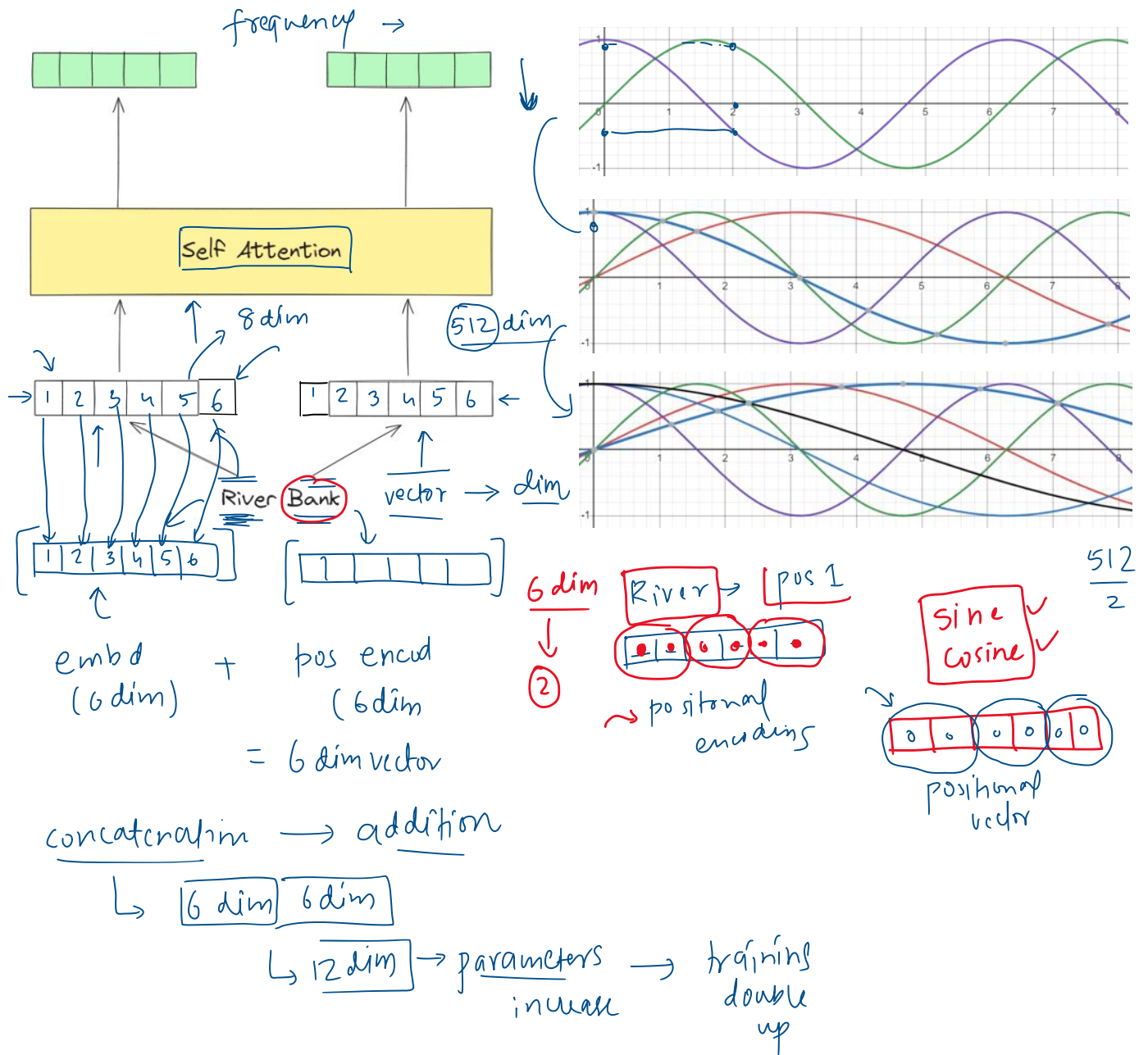
for i = 1 ✓

pos → position
pos = 0   pos = 1

dmodel → dim
of
embedding

$\boxed{dmodel = 6}$

$$PE_{(pos, 2i)} = \sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos, 2i+1)} = \cos(pos/10000^{2i/d_{\text{model}}})$$

i = 0 — $\boxed{dmodel/2}$ 1   0 - $\boxed{6/2}$ 3   0 - 3

$\boxed{i = 0, 1, 2}$

$\boxed{0 - 2}$

$\boxed{\text{for } i = 0}$

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

$PE(1, \underline{1}) = \cos(1/10000^0) = \underline{0.54}$

$\boxed{\text{for } i = 1}$

$\ldots \frac{1}{3}) = 0.04$

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

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

for i=1

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

$PE(1,3) = \dfrac{\cos\left(1 / 100000^{1/3} \right)}{} = \underline{0.99}$

for i=2 ✓

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

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

for i=2

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

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

River →

0	1	0	1	0	1

bank →

0.84	0.54	0.04	0.99	0.00	0.99
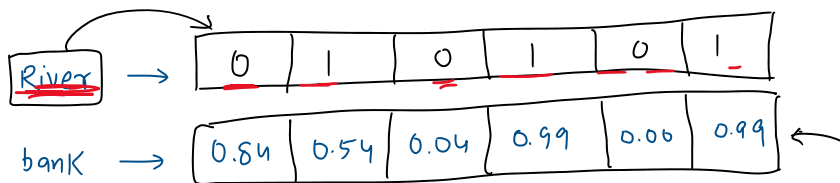
Interesting Observations

Sentence → 50 words

→ embedding → 128   d model = 128

50 → d model → 128
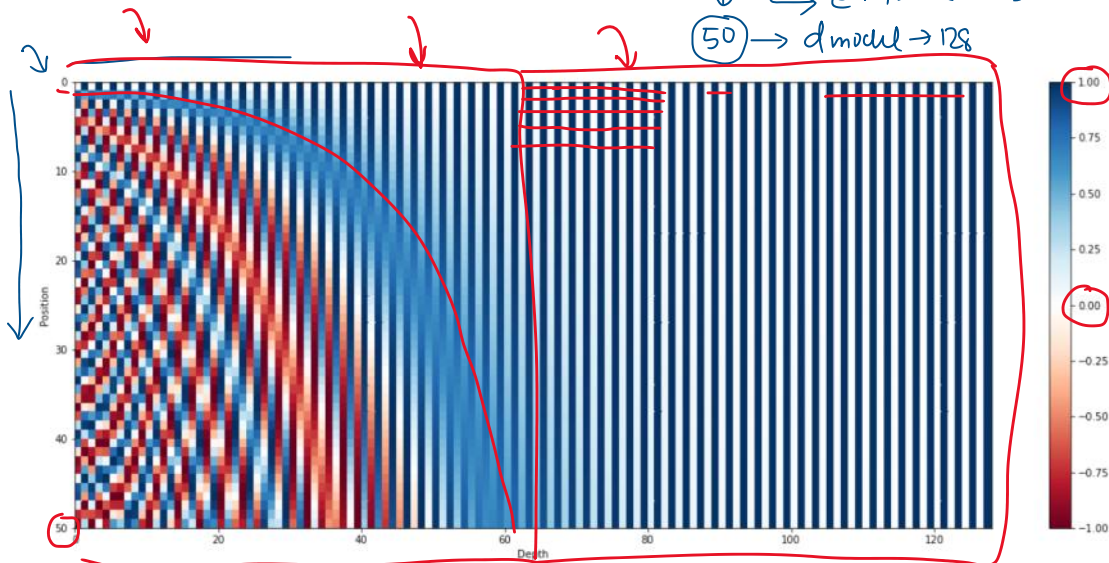
binary encoding   4 bit 8 bit

| 100 |

0 :	0 0 0 0		8 :	1 0 0 0
1 :	0 0 0 1		9 :	1 0 0 1
2 :	0 0 1 0		10 :	1 0 1 0
3 :	0 0 1 1		11 :	1 0 1 1
4 :	0 1 0 0		12 :	1 1 0 0
5 :	0 1 0 1		13 :	1 1 0 1
6 :	0 1 1 0		14 :	1 1 1 0
7 :	0 1 1 1		15 :	1 1 1 1
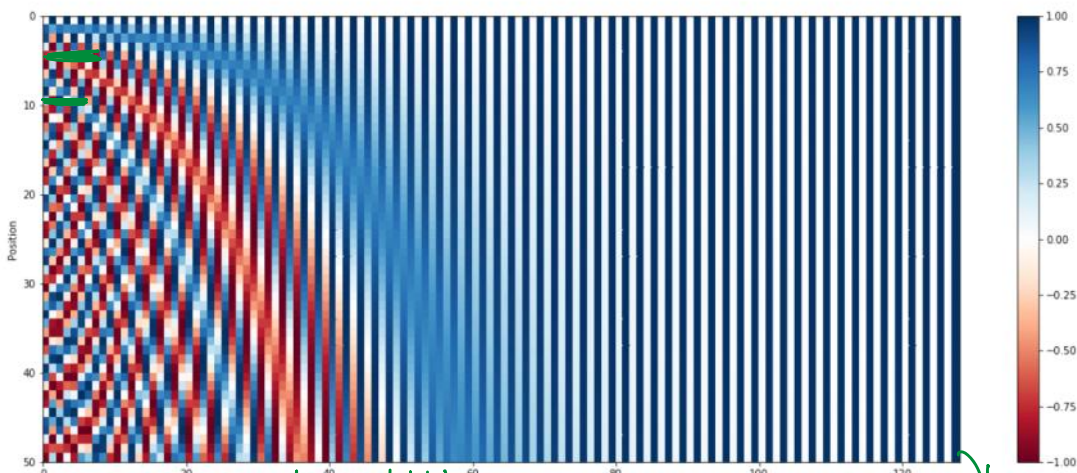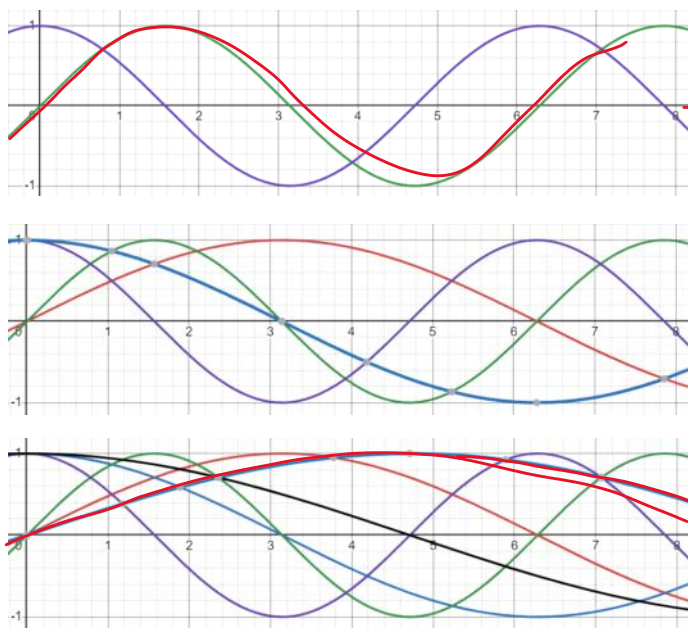
50 words
↓
50 vectors
↓
128 dim

V_{10}
V_{20}

→ 5 delta / distance   matrix →   1n delta

$V_5 \rightarrow$ matrix $\quad \rightarrow$ 5 delta / distance

property

$\quad \rightarrow V_{10}$

$V_{12} \rightarrow V_{17} \quad V_{21} \rightarrow V_{26}$

$V_{10} \rightarrow$ linear transformation $\quad \rightarrow$ matrix $\rightarrow$ 10 delta distance

$\quad \rightarrow V_{20}$ for each delta $\quad V_{30} \rightarrow V_{40}$

$V_{40} \rightarrow V_{50}$

linear / matrix