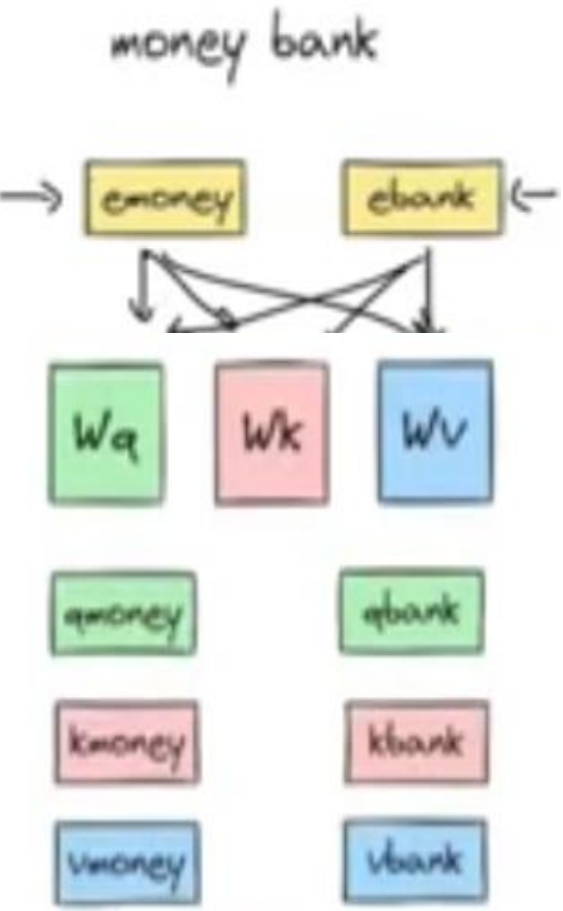
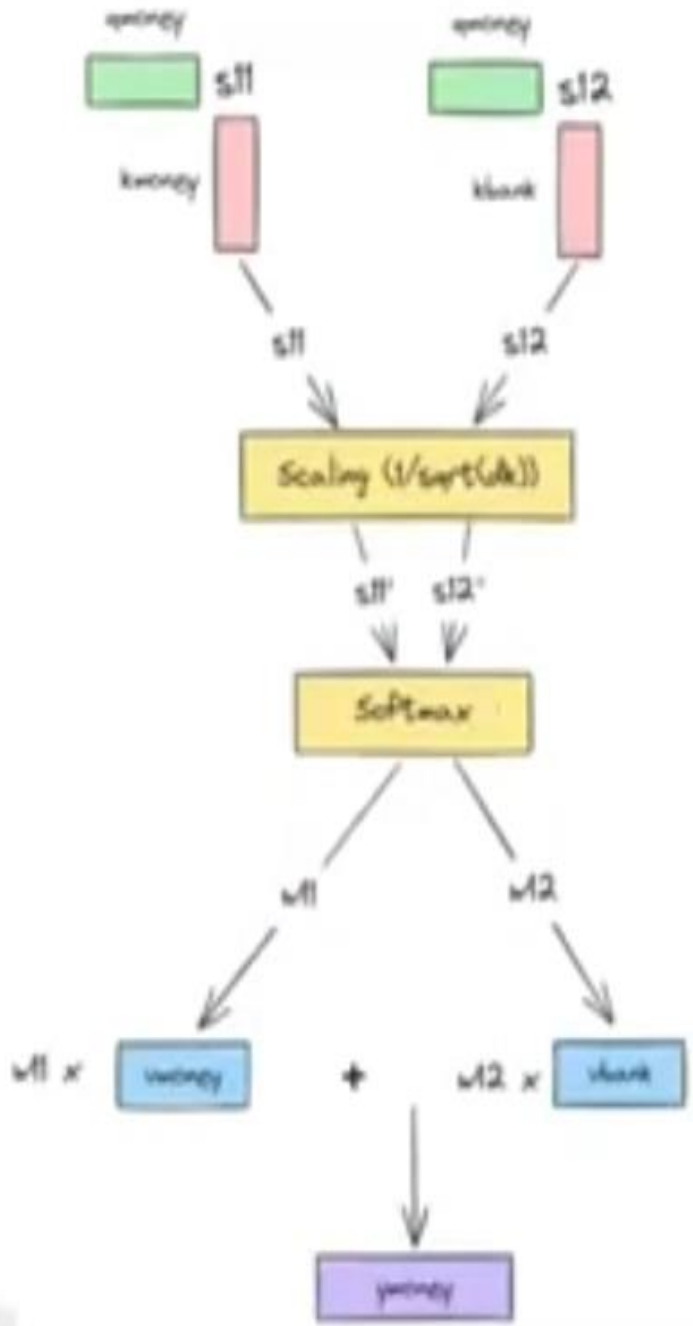
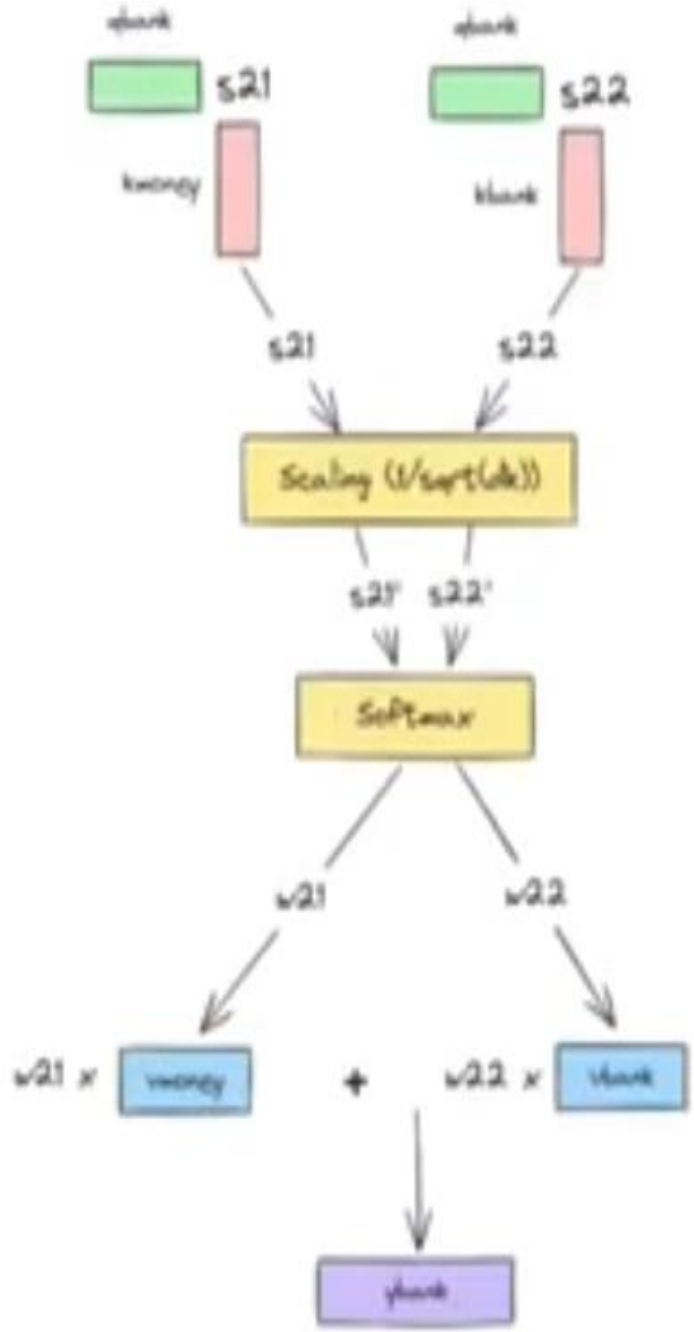
Multi-head Attention/Positional Encoding

Dr. Muhammad Safyan


money bank
emoney
ebank
Wq
Wk
Wv
qmoney
qbank
kmoney
kbank
vmoney
vbank


qmoney
s11
qmoney
s12
kmoney
kbank
s11
s12
Scaling (1/sqrt(dk))
s11'
s12'
Softmax
v11
v12
v11 x
vmoney
+
v12 x
vbank
ymoney


qbank
s21
qbank
s22
kmoney
kbank
s21
s22
Scaling (1/sqrt(dk))
s21'
s22'
Softmax
v21
v22
v21 x
vmoney
+
v22 x
vbank
ybank

Problem in self-Attention

The man saw the astronomer with a telescope

	The	man	saw	the	astronomer	with	a	telescope
The								
man								
saw								
the								
astronomer								
with								
a								
telescope								

- Self attention can hold single translation/perspective at a time.
- Depends upon
- Similarity between
  - man and telescope
  - Saw and telescope
- In second scenario
  - Similarity may be between
  - Astronomer and Telescope
  - Saw and telescope
- Documentation summary

Paragraph

The future of AI in Pakistan presents a dynamic and promising landscape, marked by rapid advancements and a burgeoning ecosystem of innovation. With a robust talent pool of engineers and IT professionals, Pakistan is poised to become a significant player in the global AI arena. The government's proactive stance on AI, exemplified by initiatives like the National Strategy for Artificial Intelligence, aims to harness AI's potential across various sectors, including healthcare, education, agriculture, and urban infrastructure. Pakistani an startups and tech giants are increasingly incorporating AI to solve complex societal challenges, improve efficiency, and enhance service delivery. Moreover, Pakistan's focus on ethical AI and data security aims to create a sustainable and responsible growth trajectory. As AI becomes more integrated into daily life and industry, Pakistan's unique blend of technological prowess, entrepreneurial spirit, and societal needs will likely shape a distinctive path in the AI domain, fostering innovation that is not only technologically advanced but also socially inclusive and impactful.

Summary-1

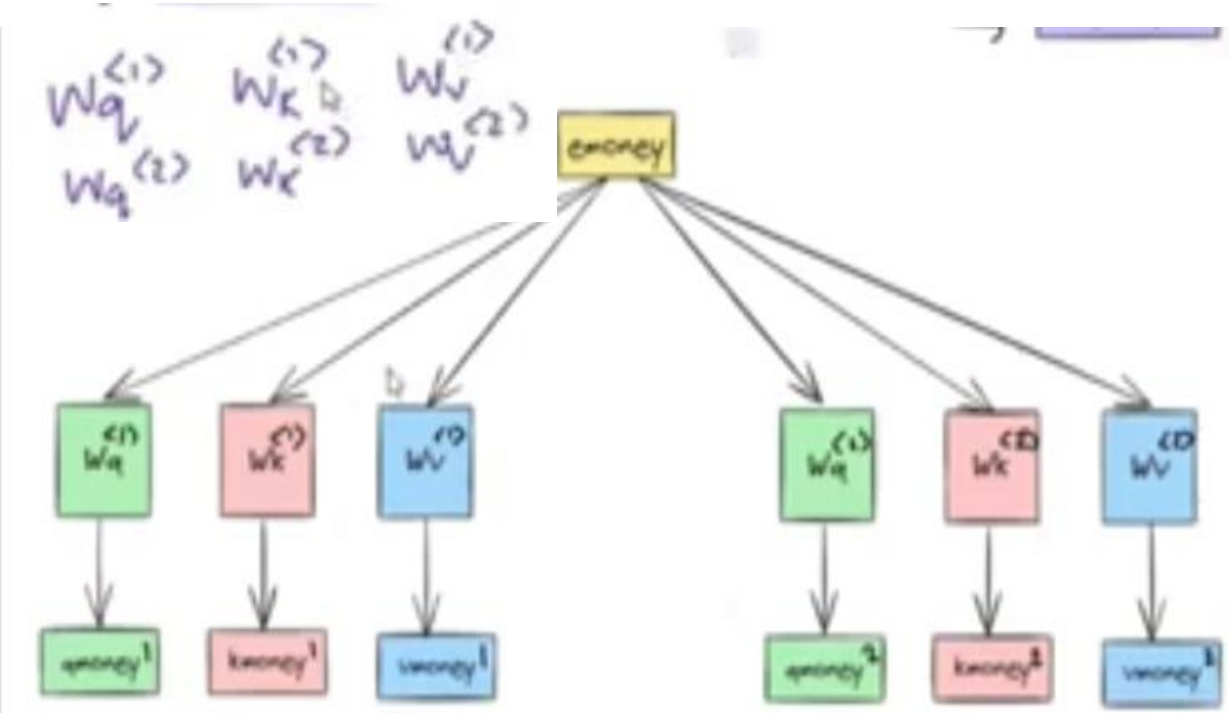
government initiatives to apply AI across various sectors like healthcare and education. With a focus on innovation, ethical AI, and data security, Pakistan aims to integrate AI to address societal challenges, enhance efficiency, and promote inclusive growth. This approach positions Pakistan to uniquely contribute to global AI advancements while ensuring sustainable and responsible development within its own borders.

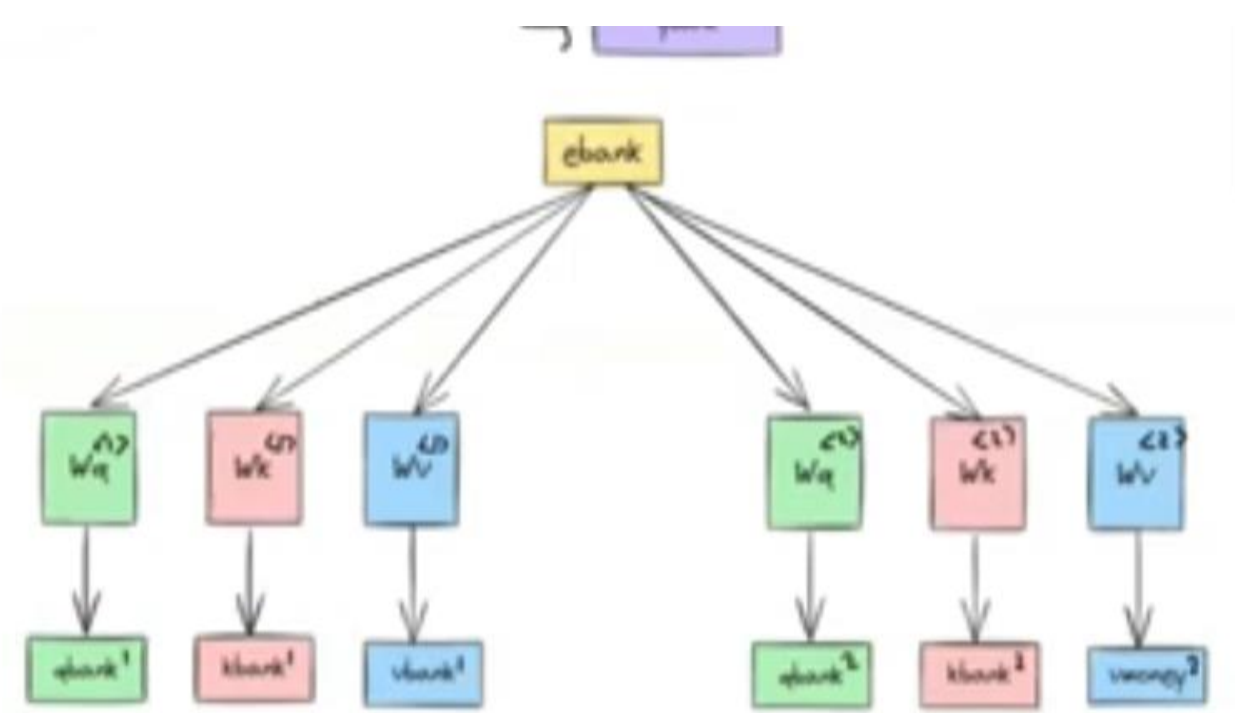
Summary-2

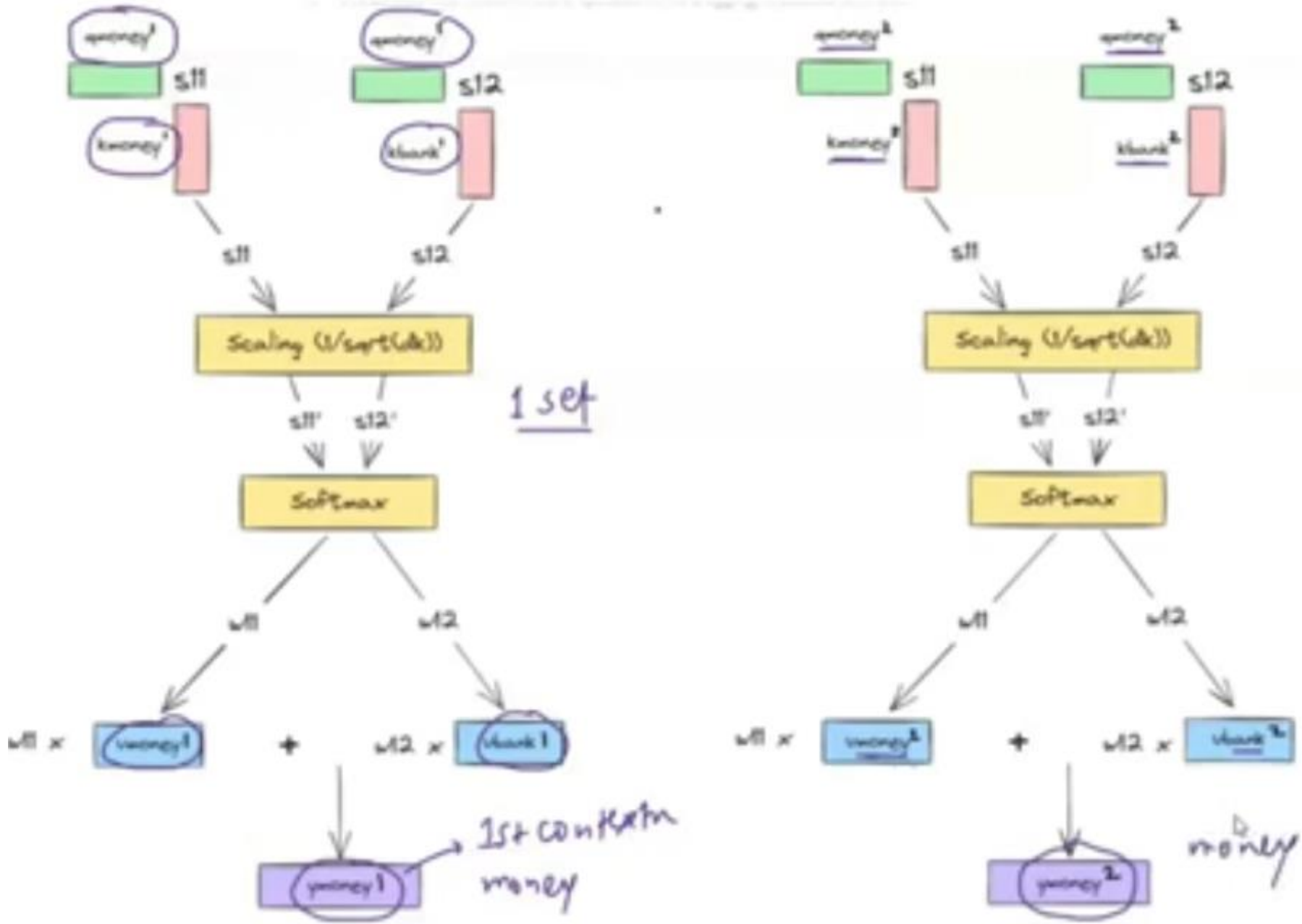
Pakistan's AI future holds promise as it harnesses a burgeoning talent pool and government initiatives to pioneer AI-driving innovation. With a focus on sectors like healthcare and education, Pakistan aims to leverage AI for a digital development. The emphasis on ethical AI and data security underscores pakistanian commitment to responsible technological advancement. This approach positions pakistan not only as a global AI hub but also as a trailblazer in addressing societal challenges through cutting-edge technology

Multi-head Attention

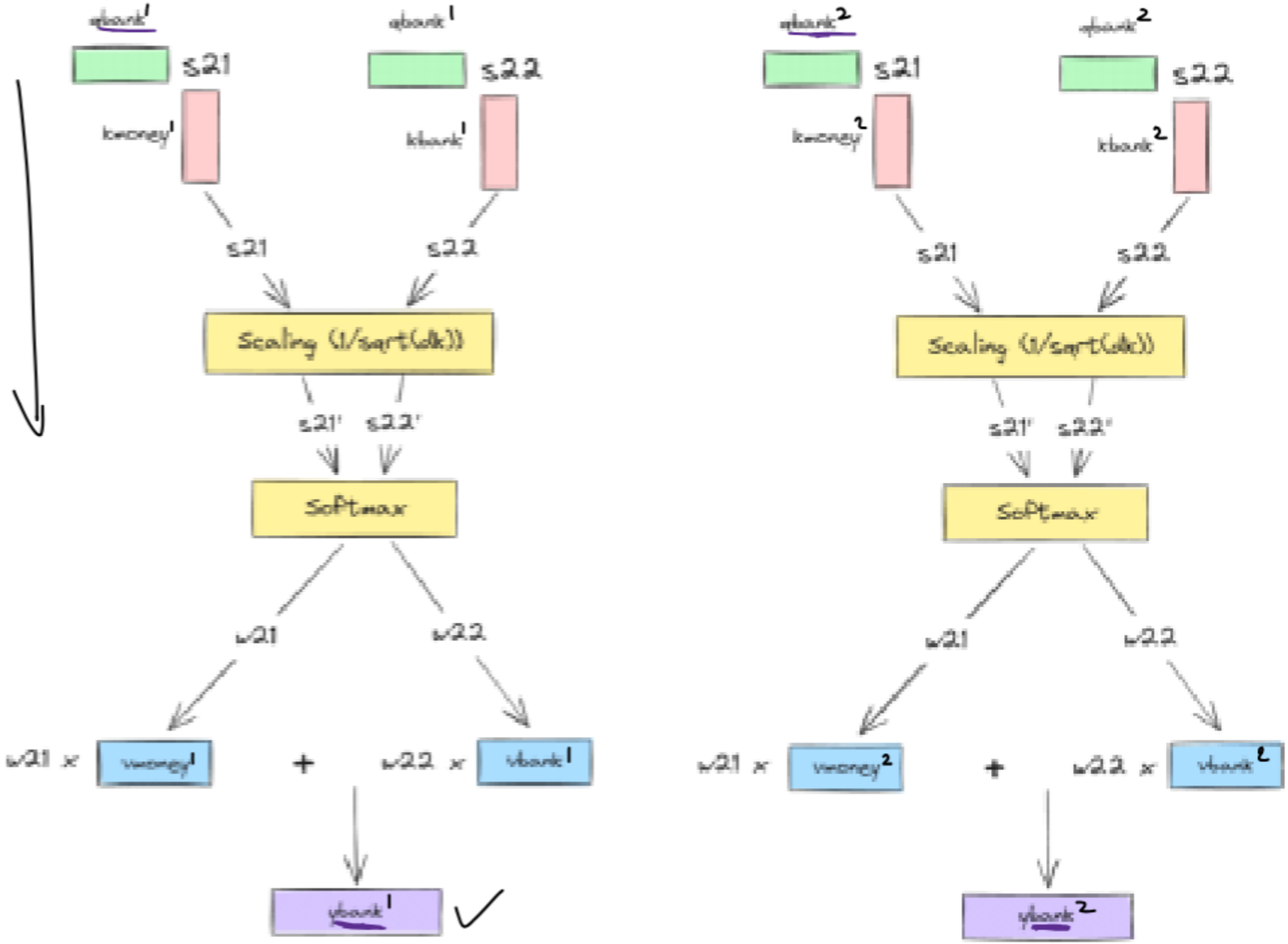
- What if we use two self-attention model instead of one attention model.
- Use two sets of
  - W1q, w1k, w1v
  - W2q, w2k, w2v
- Whole process will be done twice and you will be two contextual vectors

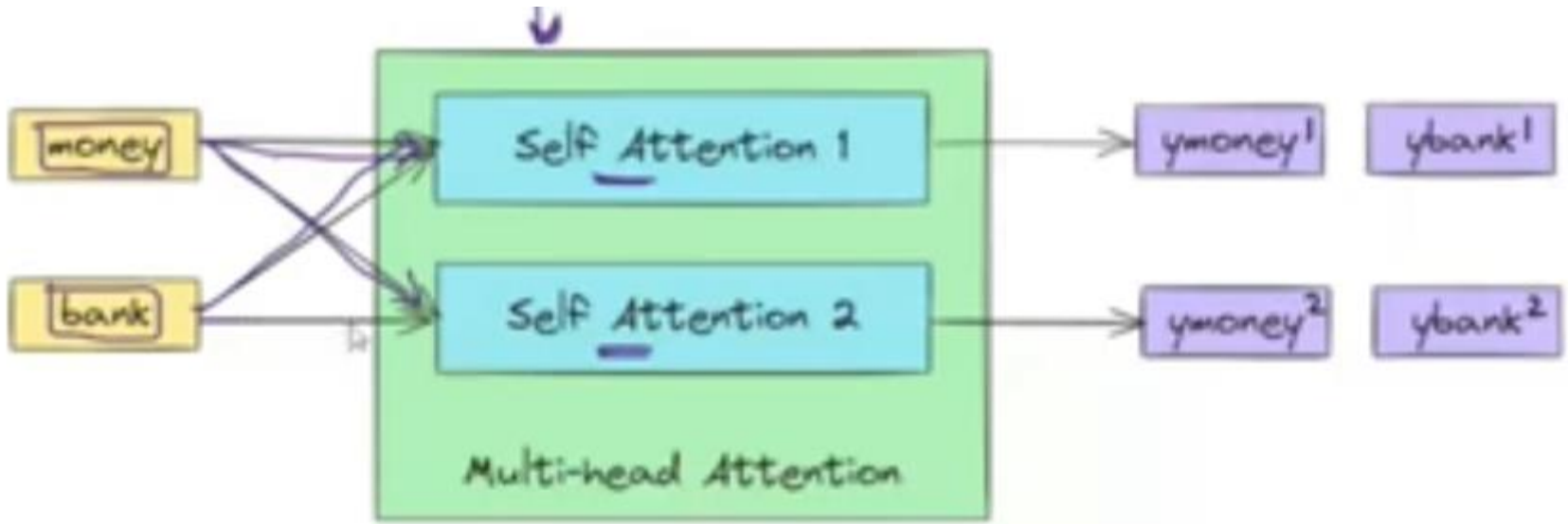

$W_q^{(1)}$ $W_k^{(1)}$ $W_v^{(1)}$
$W_q^{(2)}$ $W_k^{(2)}$ $W_v^{(2)}$
emoney
$W_q^{(1)}$ $W_k^{(1)}$ $W_v^{(1)}$ $W_q^{(2)}$ $W_k^{(2)}$ $W_v^{(2)}$
q_{emoney}^1 k_{emoney}^1 v_{emoney}^1 q_{emoney}^2 k_{emoney}^2 v_{emoney}^2


ebank
$W_q^{(1)}$ $W_k^{(1)}$ $W_v^{(1)}$ $W_q^{(2)}$ $W_k^{(2)}$ $W_v^{(2)}$
q_{ebank}^1 k_{ebank}^1 v_{ebank}^1 q_{ebank}^2 k_{ebank}^2 v_{ebank}^2

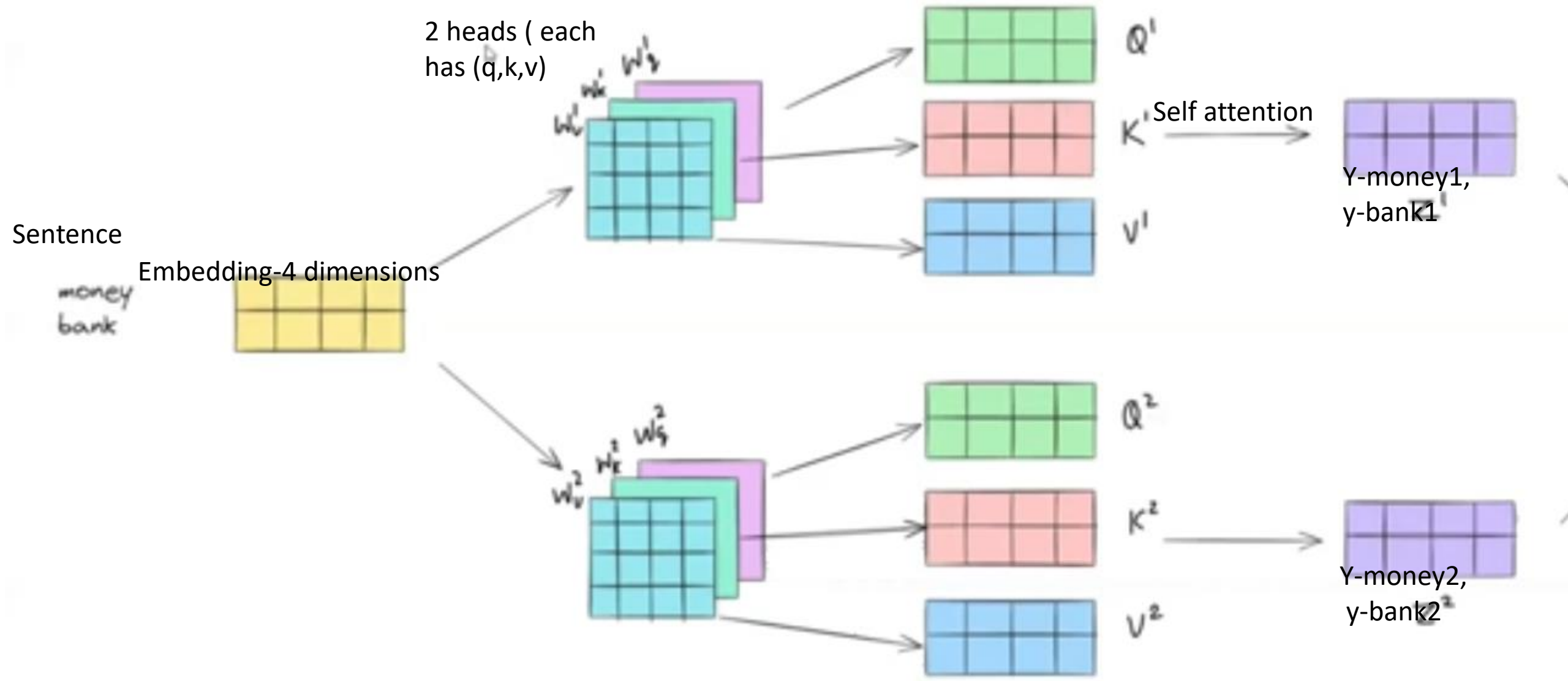


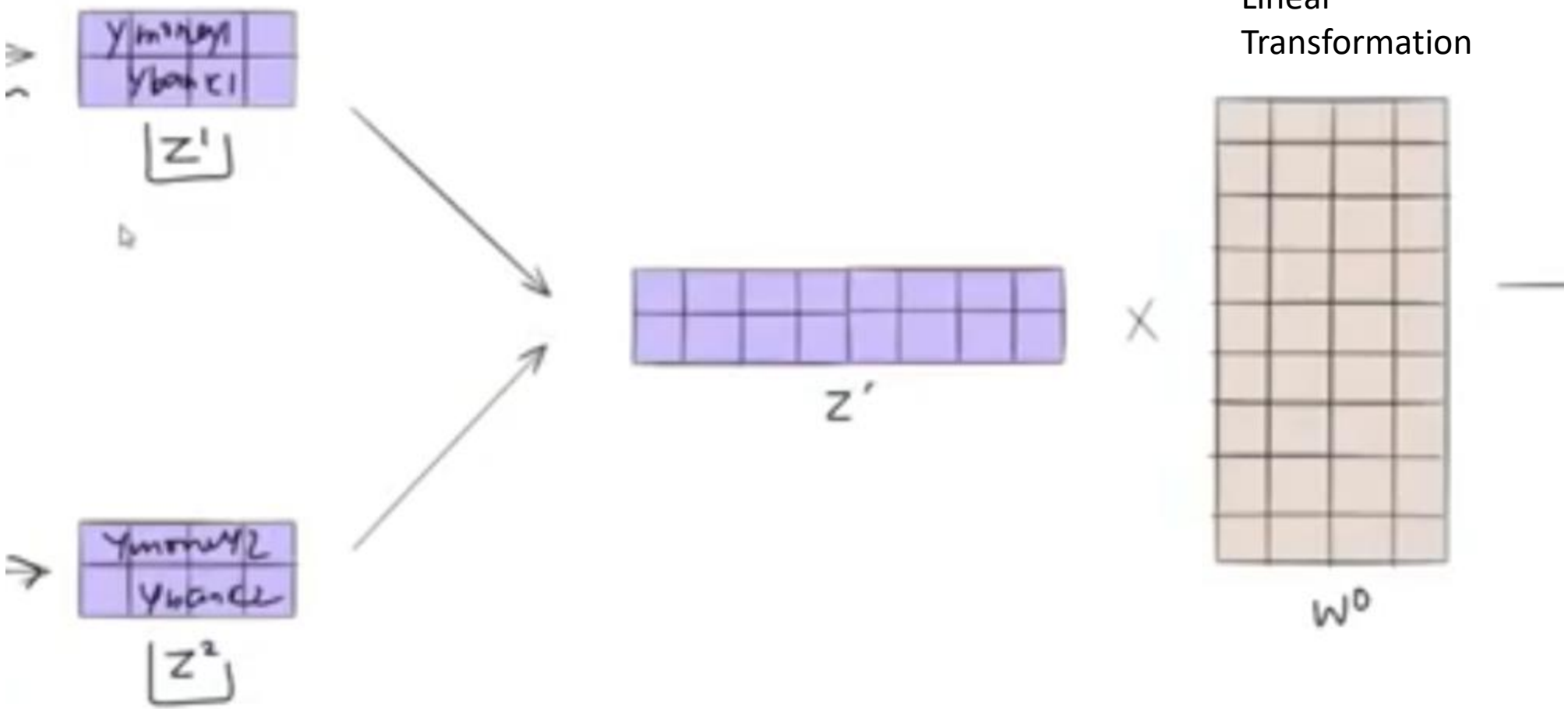
qmoney¹
s11
kmoney¹
qmoney¹
s12
kbank¹
s11
s12
Scaling (1/sqrt(dk))
s11'
s12'
1 set
Softmax
w11
w12
w11 x
vmoney1
+
w12 x
vbank1
ymoney1
1st contextn
money
qmoney²
s11
kmoney²
qmoney²
s12
kbank²
s11
s12
Scaling (1/sqrt(dk))
s11'
s12'
Softmax
w11
w12
w11 x
vmoney²
+
w12 x
vbank²
ymoney²
money

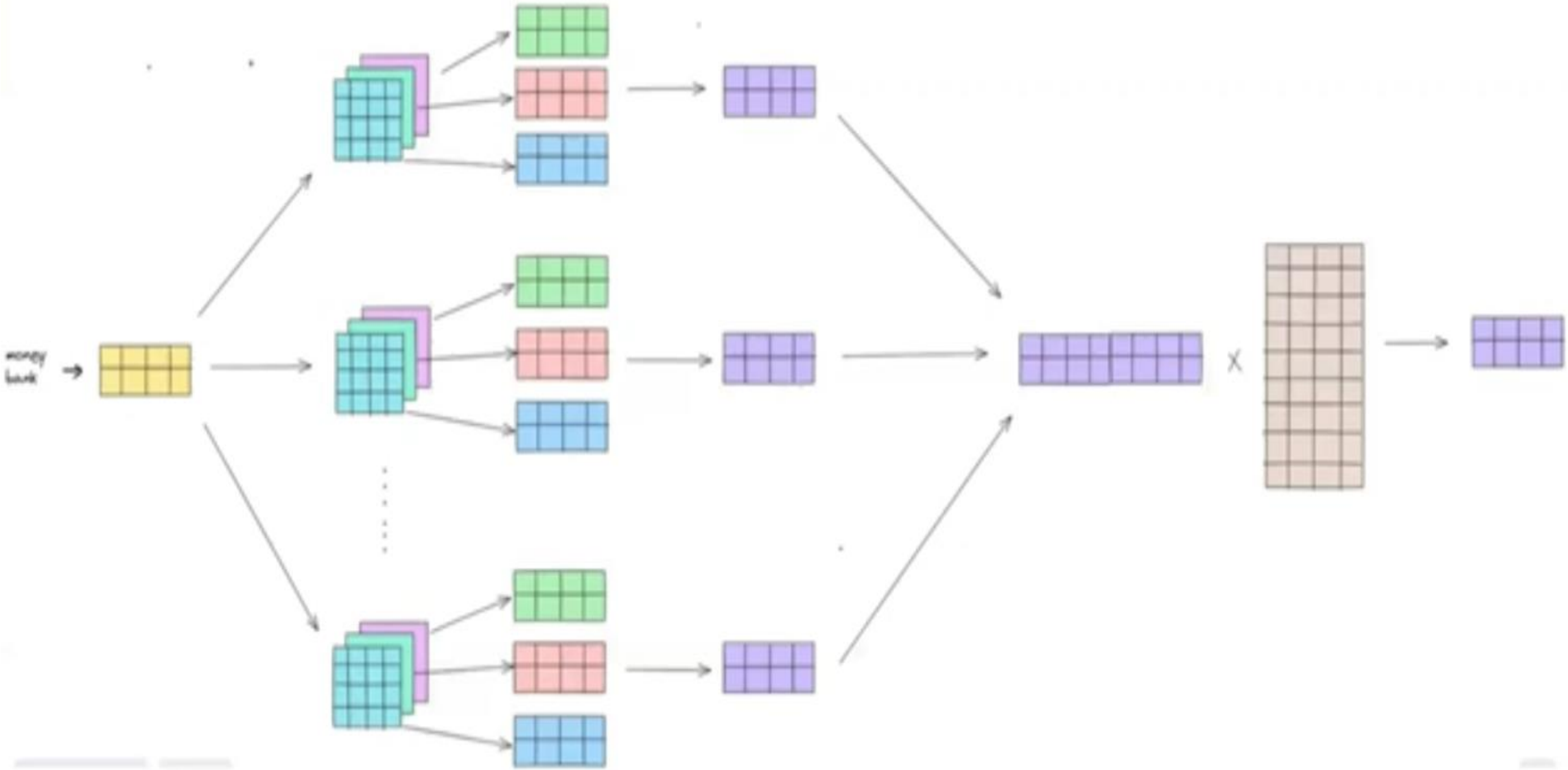

qbank1
qbank1
s21
s22
kmoney1
kbank1
s21
s22
Scaling (1/sqrt(dk))
s21'
s22'
SoPtmax
w21
w22
w21 x
vmoney1
+
w22 x
vbank1
ybank1
qbank2
qbank2
s21
s22
kmoney2
kbank2
s21
s22
Scaling (1/sqrt(dk))
s21'
s22'
SoPtmax
w21
w22
w21 x
vmoney2
+
w22 x
vbank2
ybank2

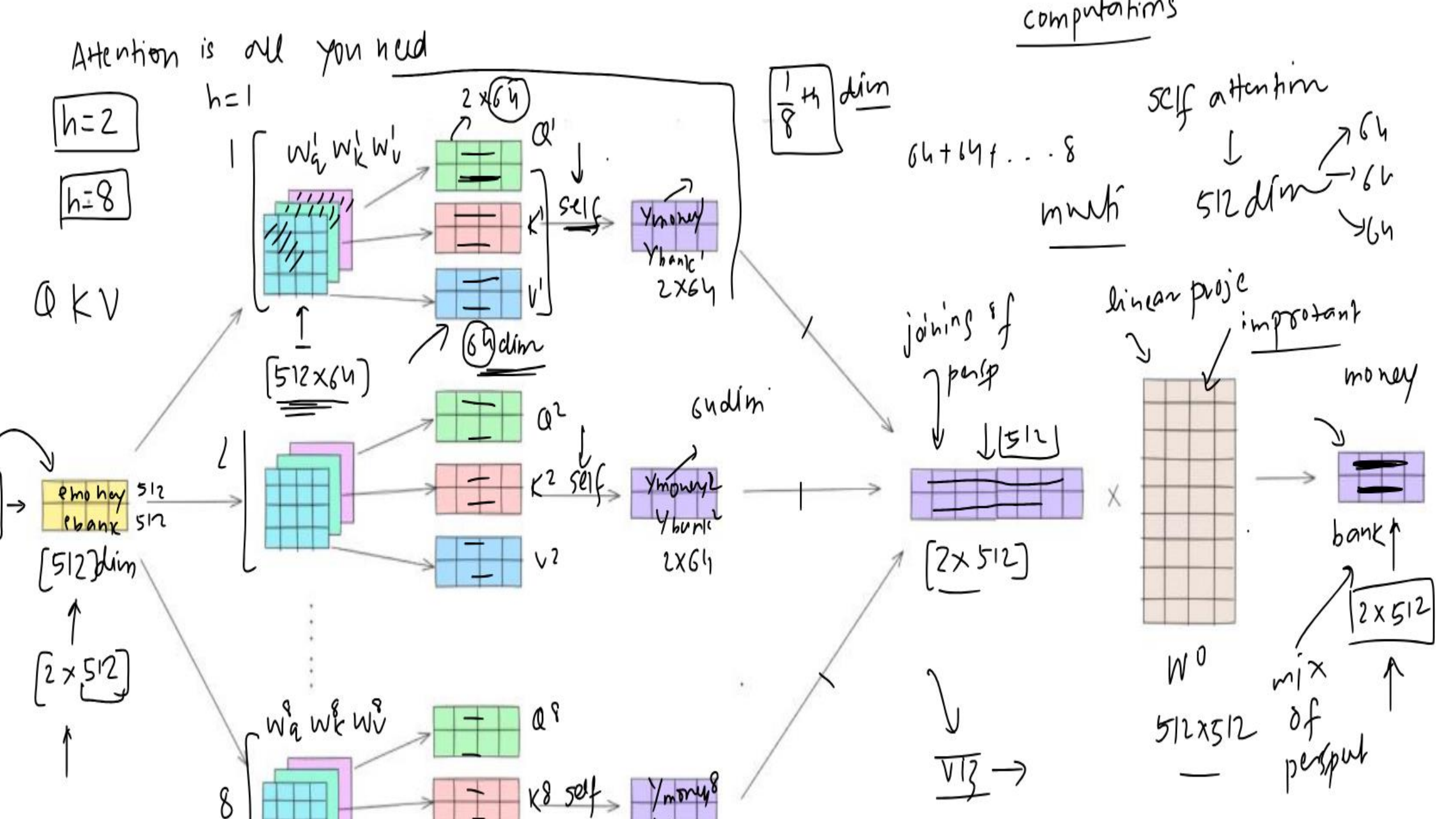


Transformer use 8 heads

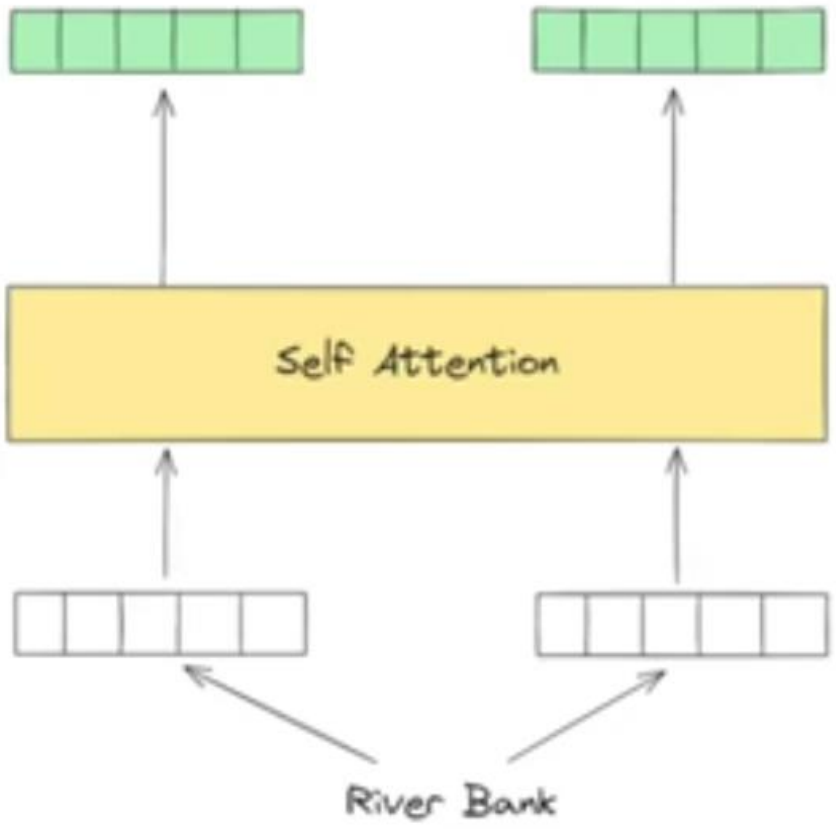

2 heads (each has (q,k,v)
Sentence
Embedding-4 dimensions
money
bank
Q1
K1
V1
Self attention
Y-money1,
y-bank1
Q2
K2
V2
Y-money2,
y-bank2

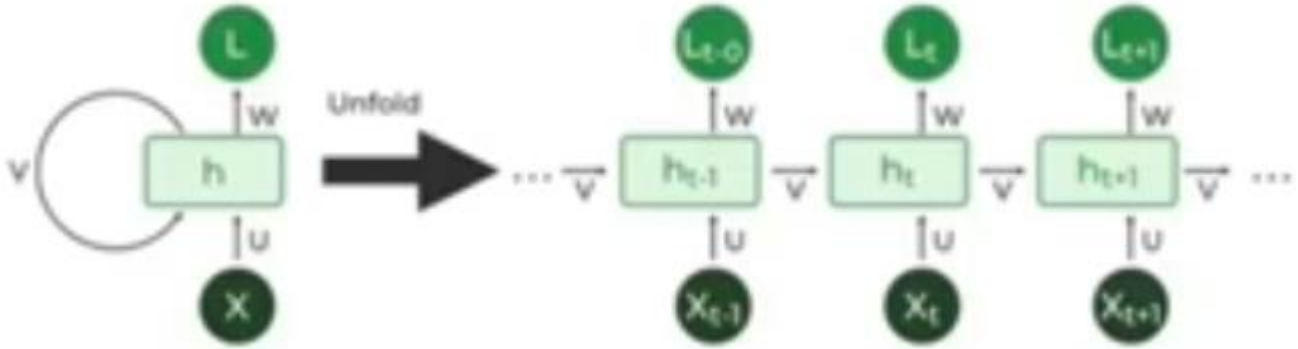

Linear
Transformation
Z¹
Z²
Z'
W⁰


empty
book


Attention is all you need
computations
h=1
h=2
h=8
1/8 th dim
64+64+...8
multi
self attention
512 dim
64
64
64
Q K V
2x64
Q1
K1
V1
self
Ymoney
Ybank'
2x64
W1q W1k W1v
[512x64]
64 dim
Q2
K2 self
Ymoney2
Ybank2
2x64
V2
64dim
Q8
K8 self
Ymoney8
W8q W8k W8v
emoney
ebank
512
512
[512]dim
[2x512]
joining of persp
[2x512]
linear proje
important
money
bank
2x512
W0
512x512
mix of perspt
VIZ →

Positional Encoding why?





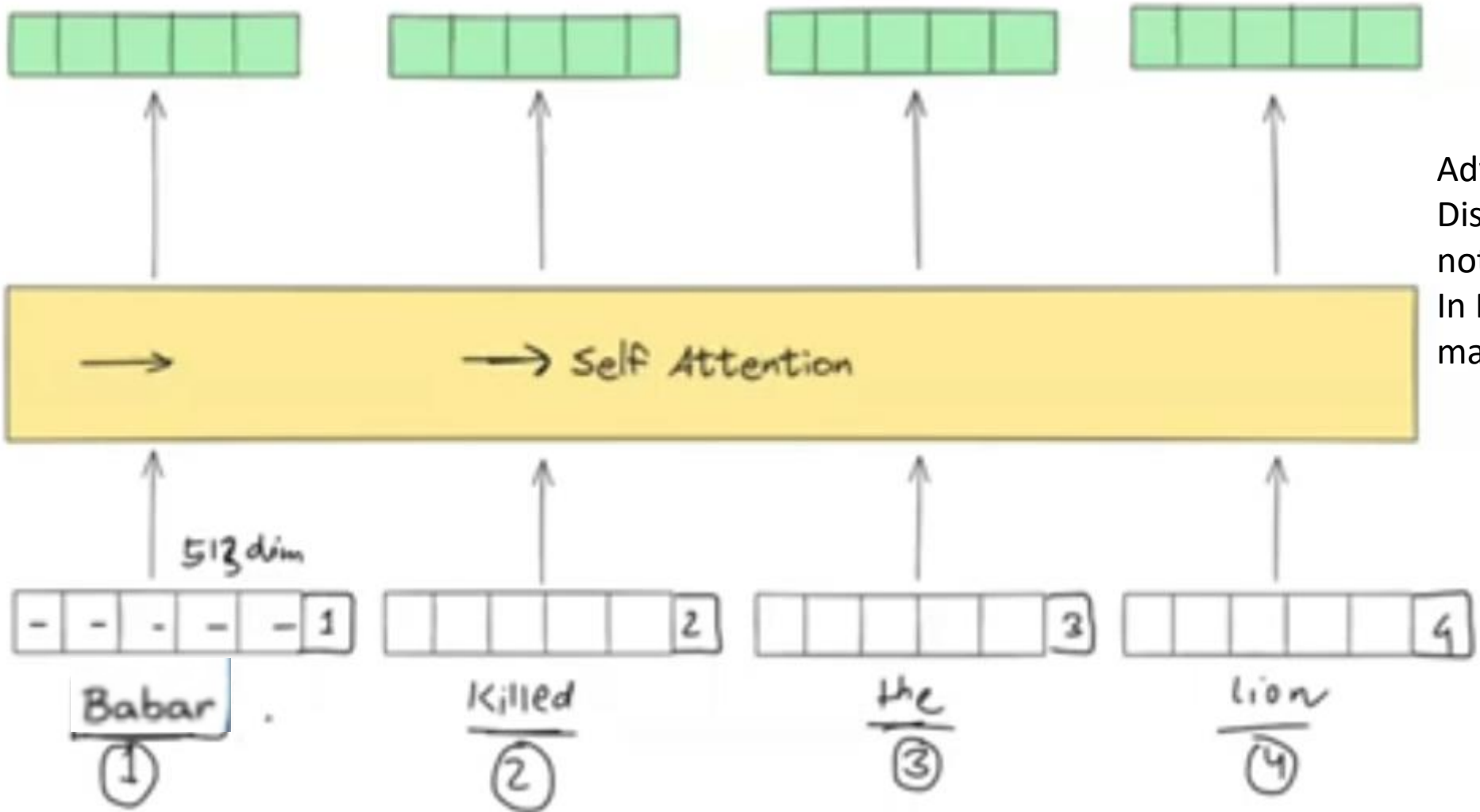
Benefit of Self-attention:
Parallelism
Drawback:
Word order is no captured.

POSITIONAL ENCODING

- Babar killed Loin
- Loin Killed Babar
- There is no order of word understood by Self-attention model. But order does matter.
- This problem is not in RNN.

Positional encoding

- Solution is in positional encoding. That preserve the position of the encoding.



Advantage: Parallel
Disadvantage: does not hold order
In NLP order does matter.

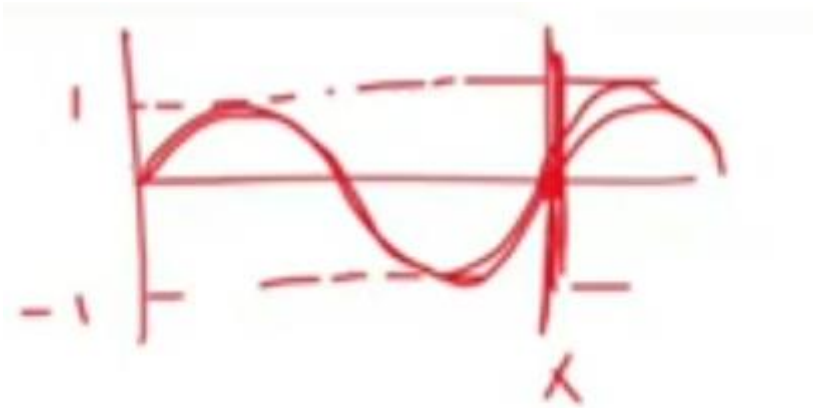
problems

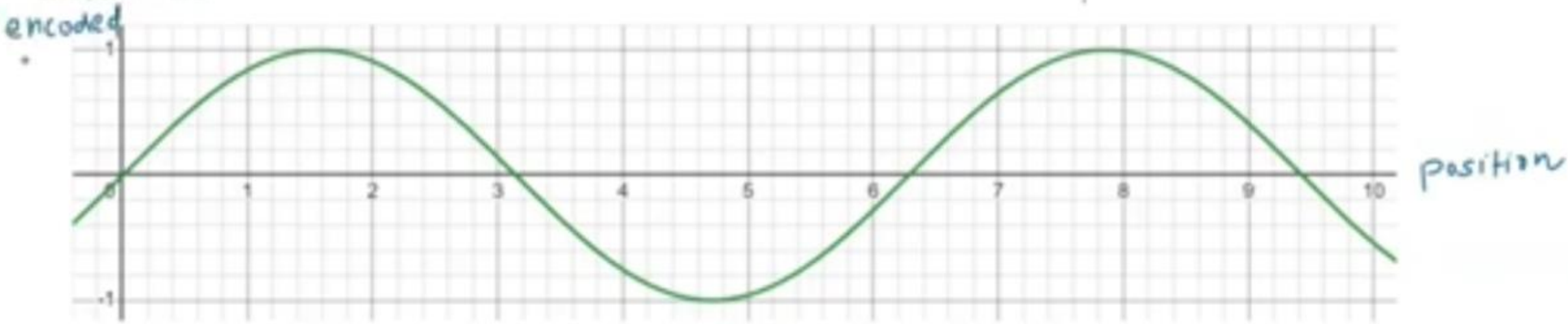
- Unbounded
  - Pdf book
  - Gradient become unstable
- Normalize → NO

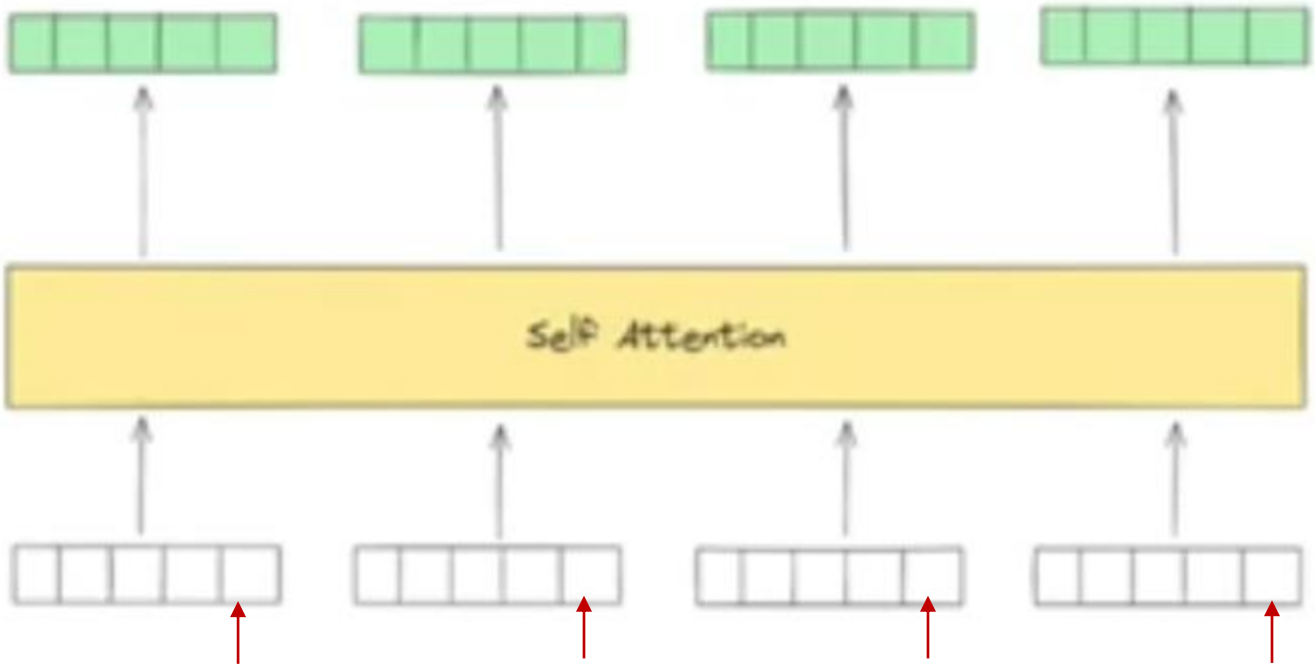
  e.g .
  - Thank you
  - Babar killed the loin

- Number are discrete. That is problem with NN
  - Numerical instability
- Can not captured relative position
  - Discrete sequence is not good
- Need a number that is bounded, continuous, periodic
  - Trigonometric function e.g sine, cos

- Periodic function
- Positional encoding







$y = \sin(1) = 0.84$

$y = \sin(2) = 087$

$y = \sin(3) = 0.14$

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

encoded value

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

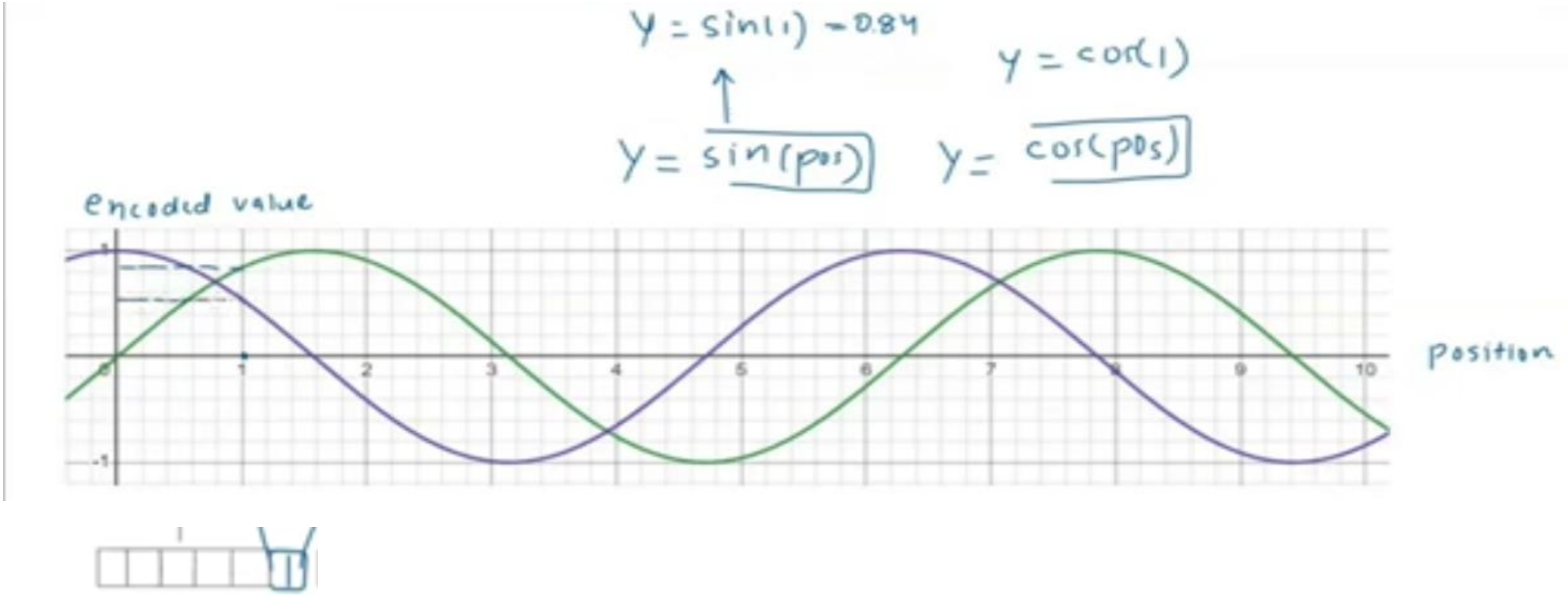
$y = \sin(1) = 0.84$

$y = \sin(2) = 0.87$

$y = \sin(3) = 0.14$

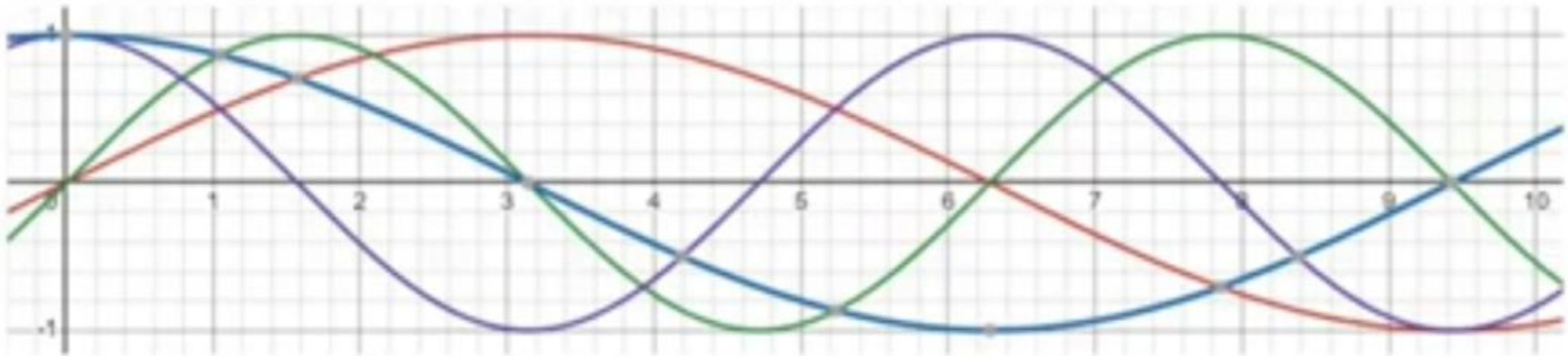
$y = \sin(4) = -0.90$

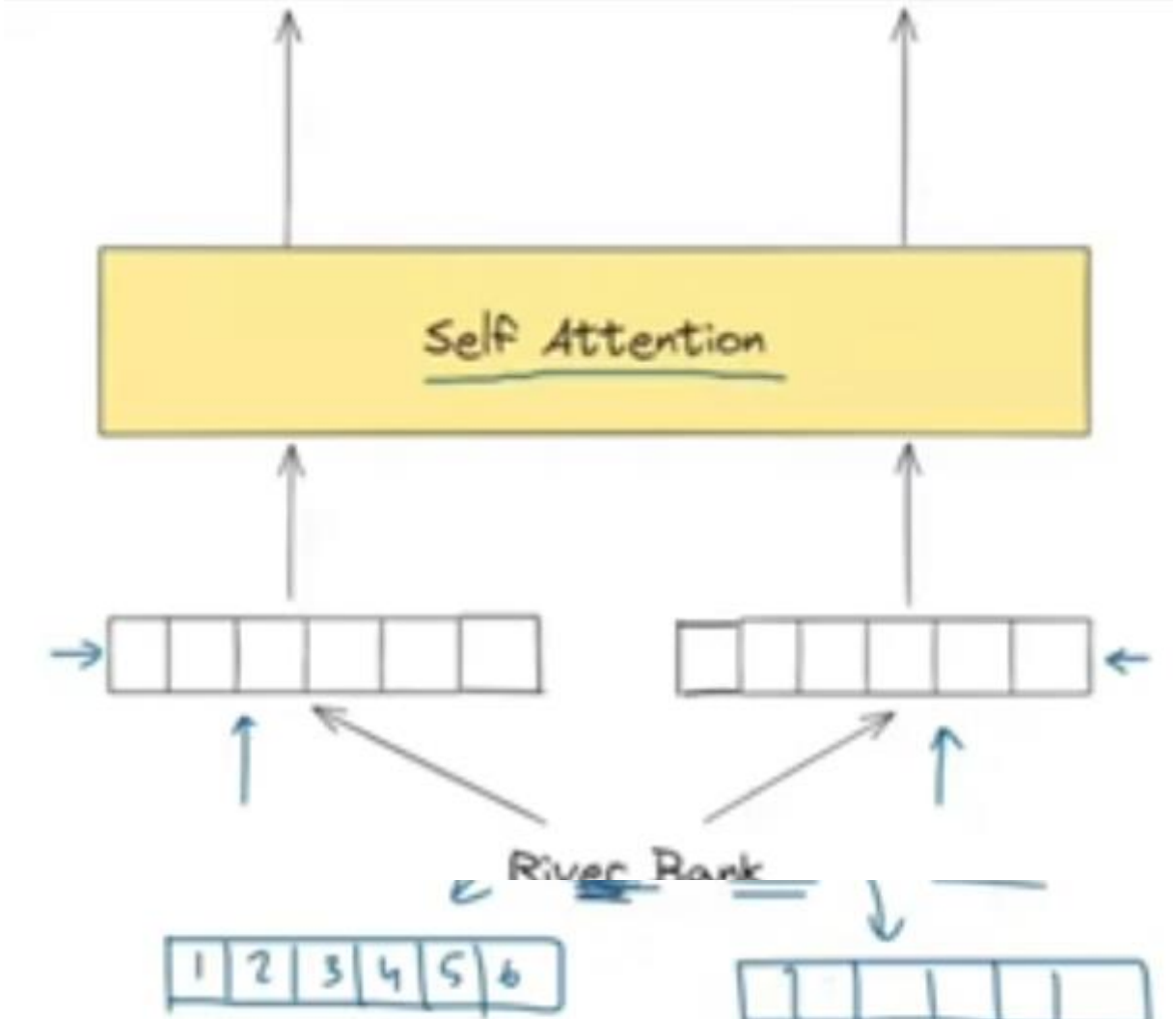
- Take two periodic functions instead of one.
- So position become vector



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

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





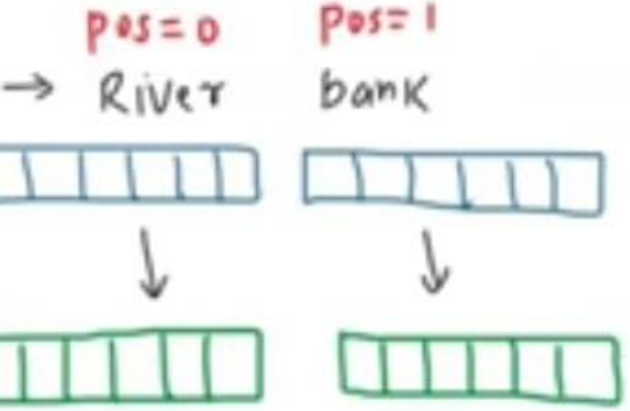
Positional encoding dimension should be same as of word.

Add them
Not concate

How to decide frequency?

embd (6 dim) + pos encod (6 dim

ORGINAL PAPER FORMUAL



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

i = 0 – d_model/2 − 1

for i = 0

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

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

for i = 0

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

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

for i = 1

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

for i = 1

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

for i=1 ✓

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

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

for i=1

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

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

for i=2 ✓

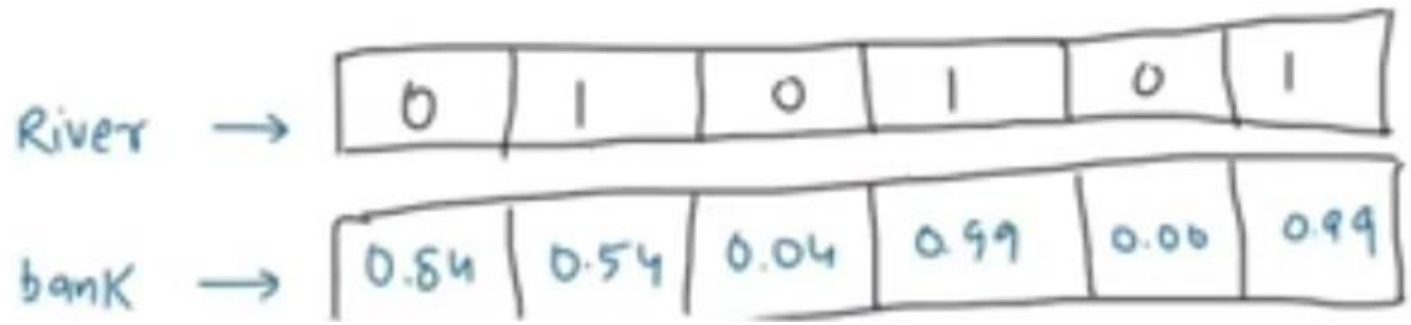
$PE(0,4) = \sin\left(0 / 10000^{2/3}\right) = 0$

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

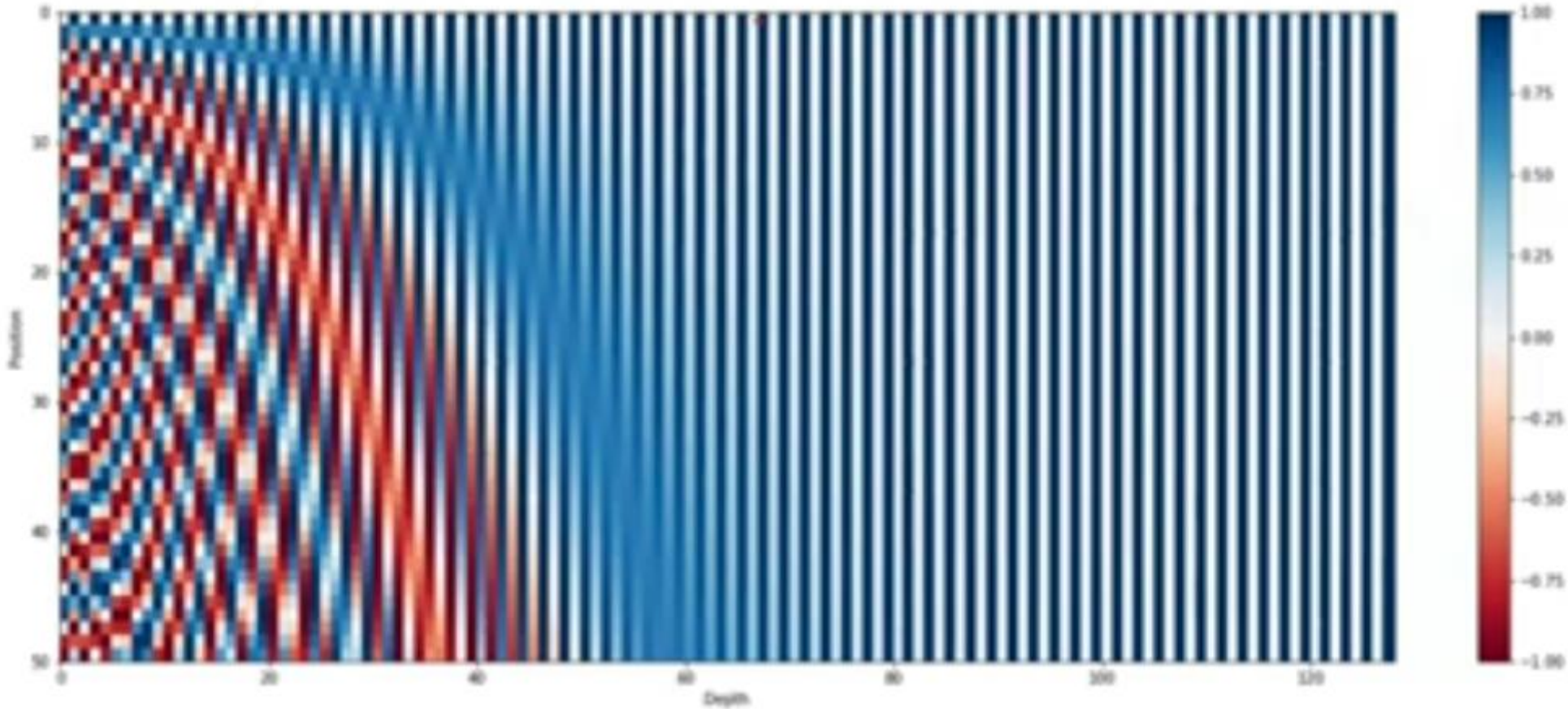
for i=2

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

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



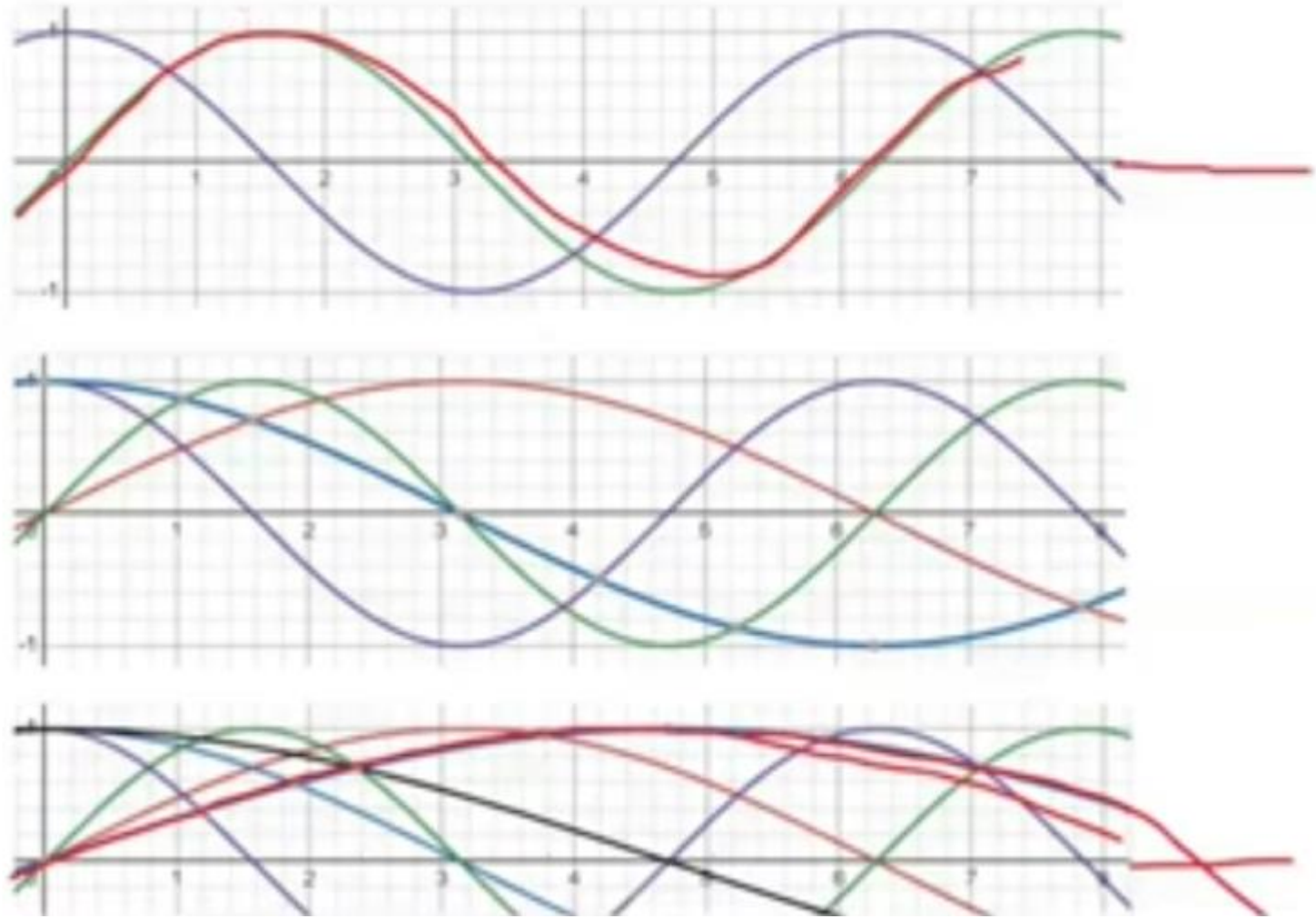
Sentence → 50 words
↓ ↳ embedding → 128 $d_{model} = 128$

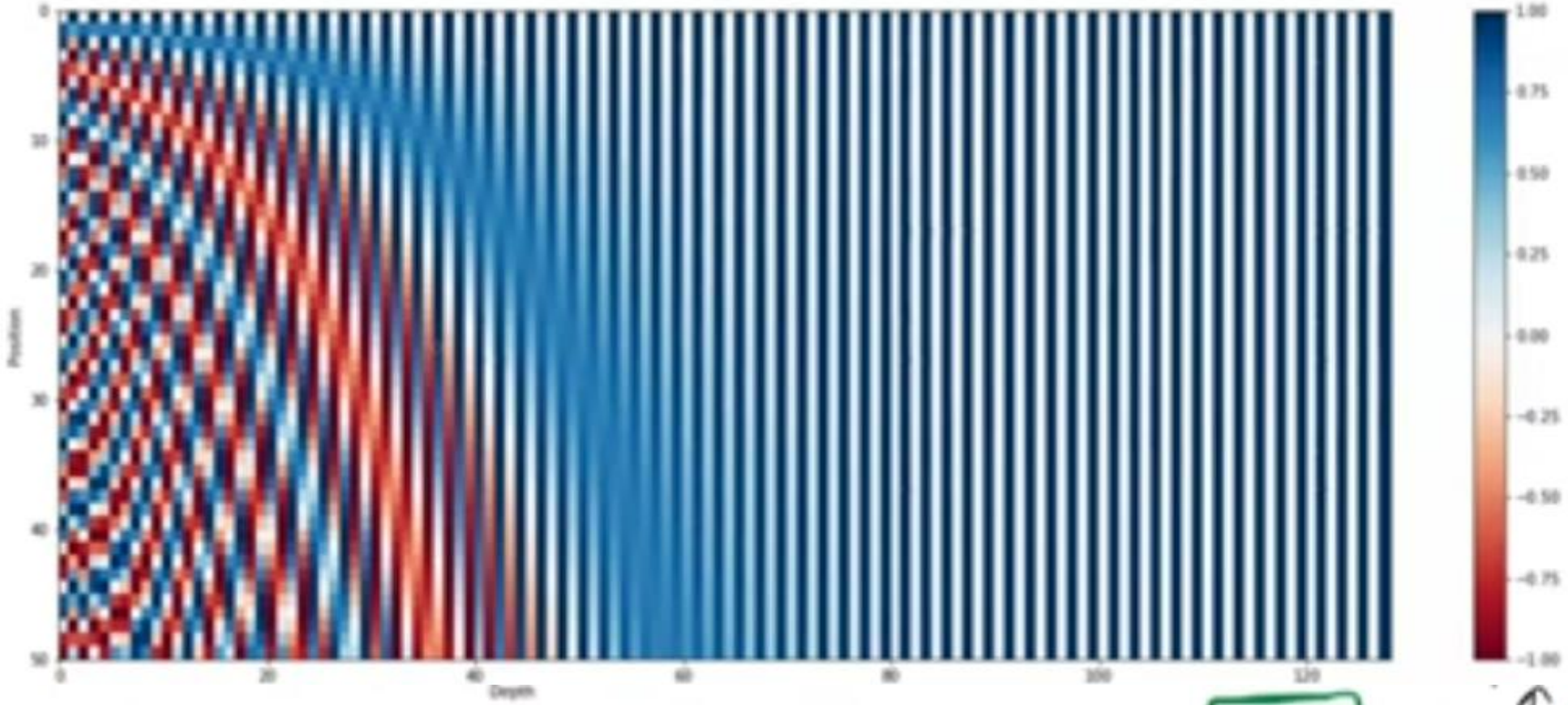


100

binary

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

properly

V_{10} → linear transformation → matrix → 10 delta distance

V_{10} ↔ V_{20}

$V_{30} \rightarrow V_{40}$

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

https://blog.timodenk.com/linear-relationships-in-the-transformers-positional-encoding/