In NLP, there are multiple Scenario to capture multiple perspectives.

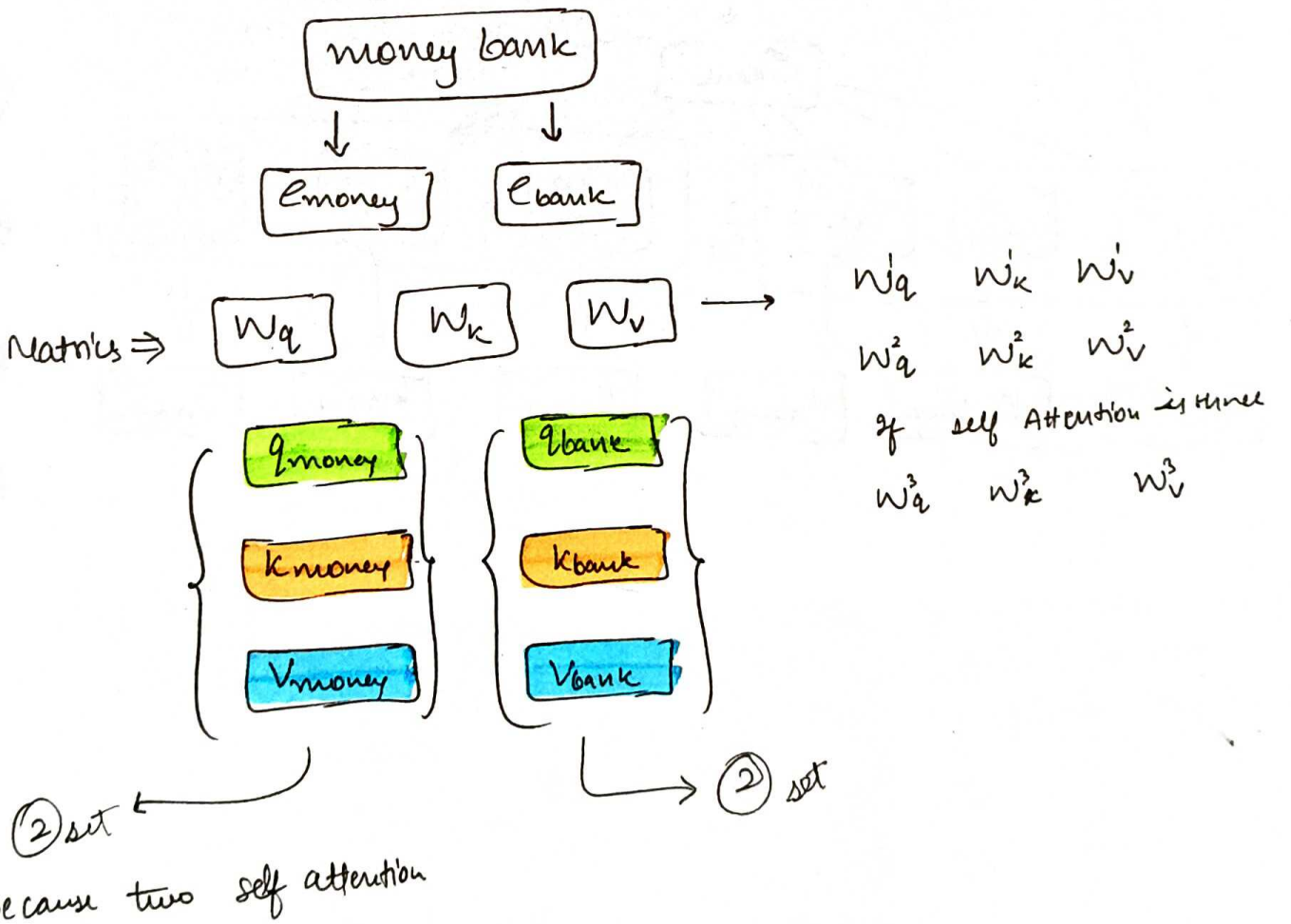
example → Document Summarization tool.

Multi-head Attention

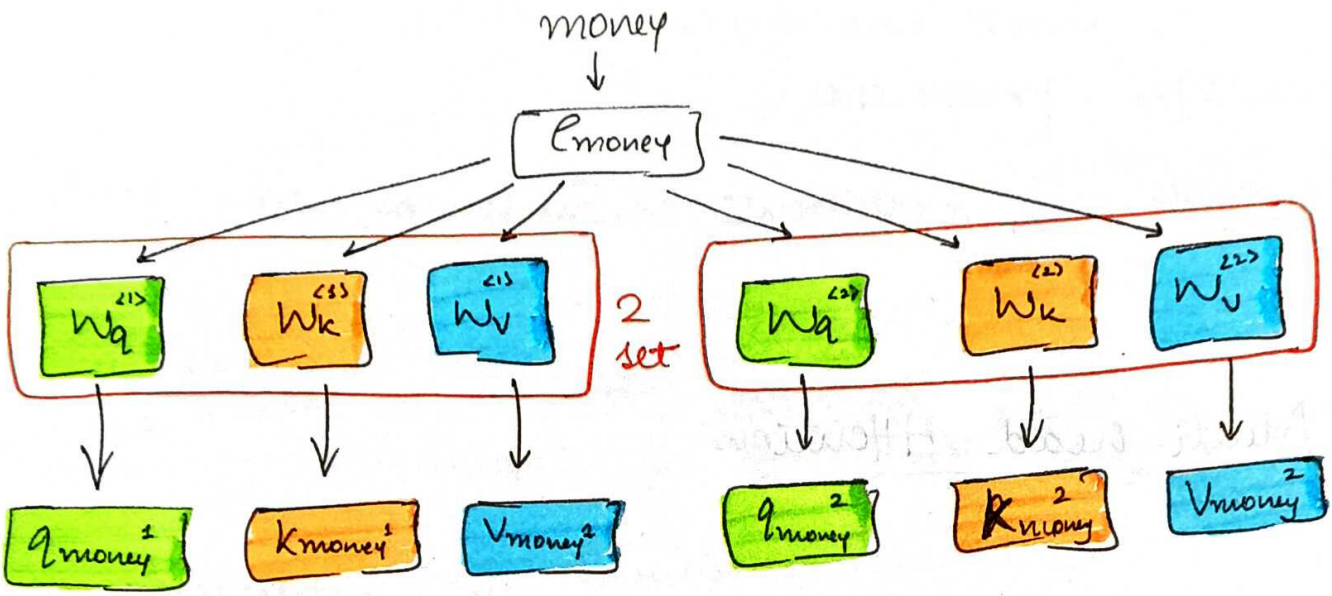
The man saw the astronomer with a telescope

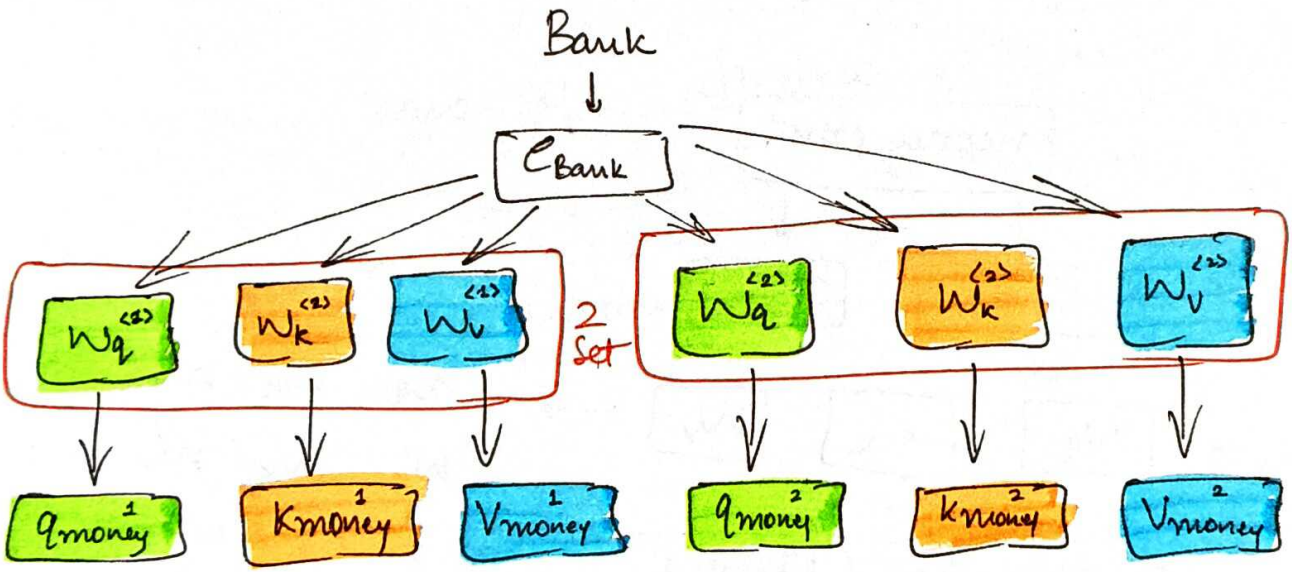
↳ 2 Meaning

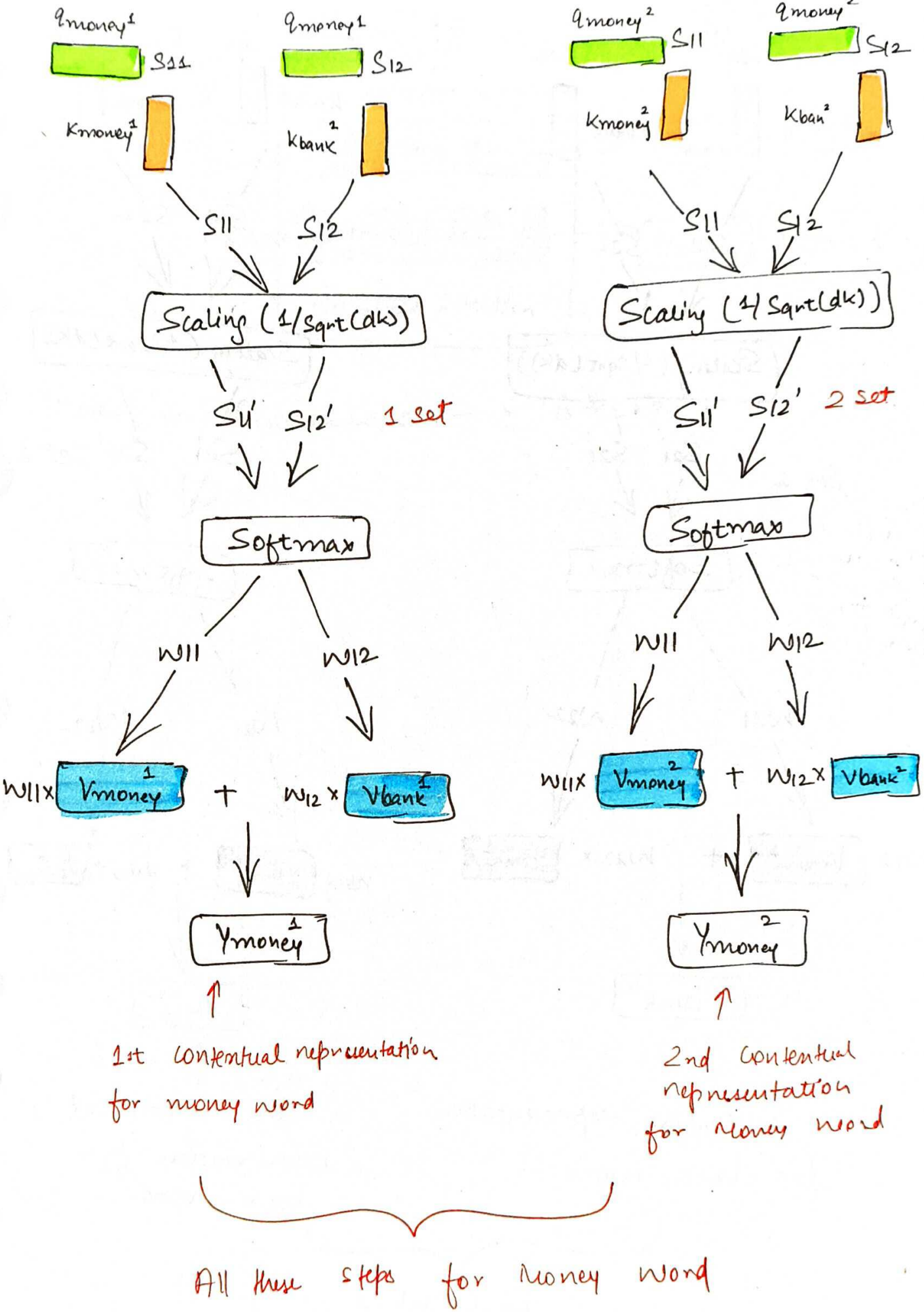
So, we use two self Attention.

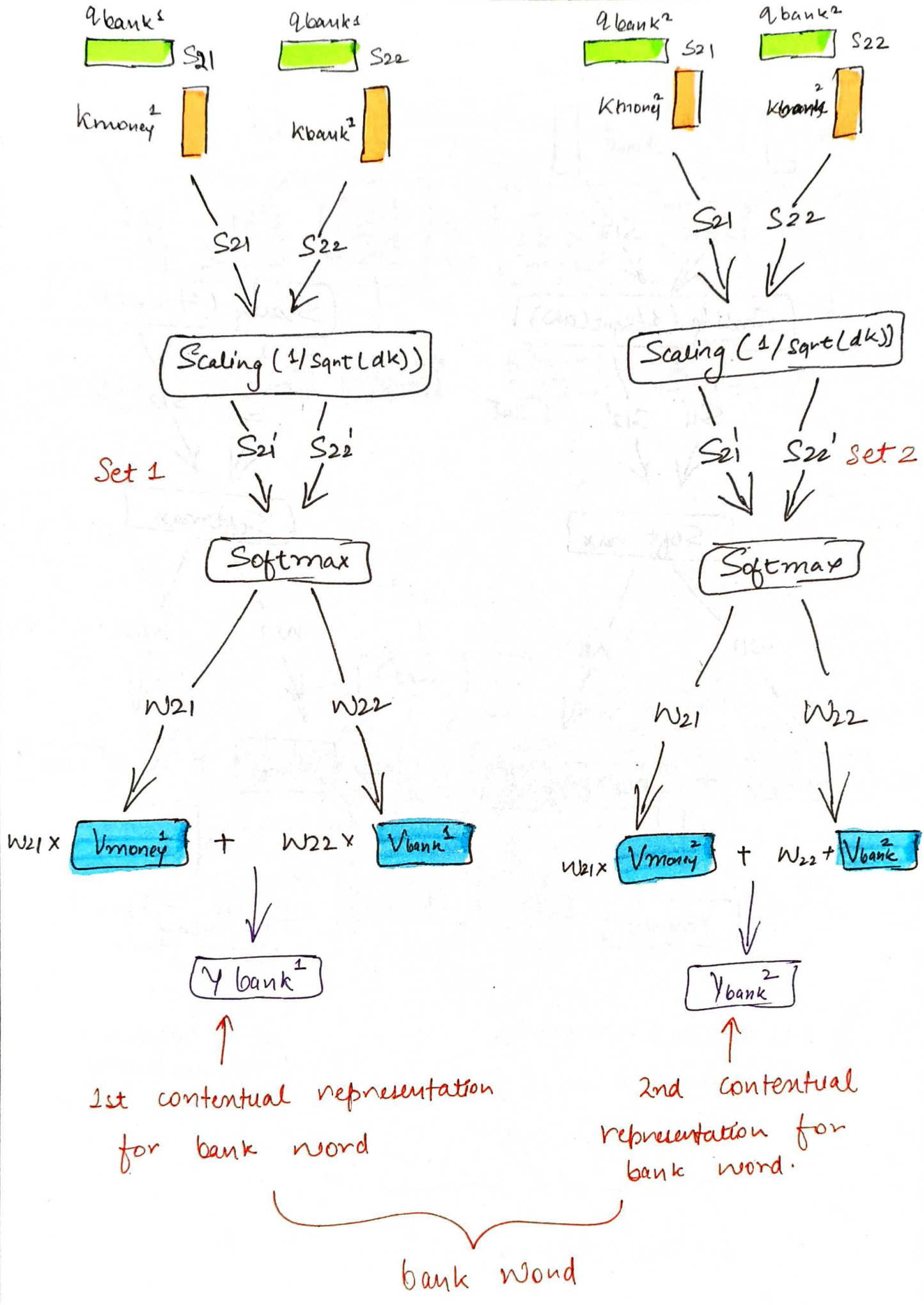


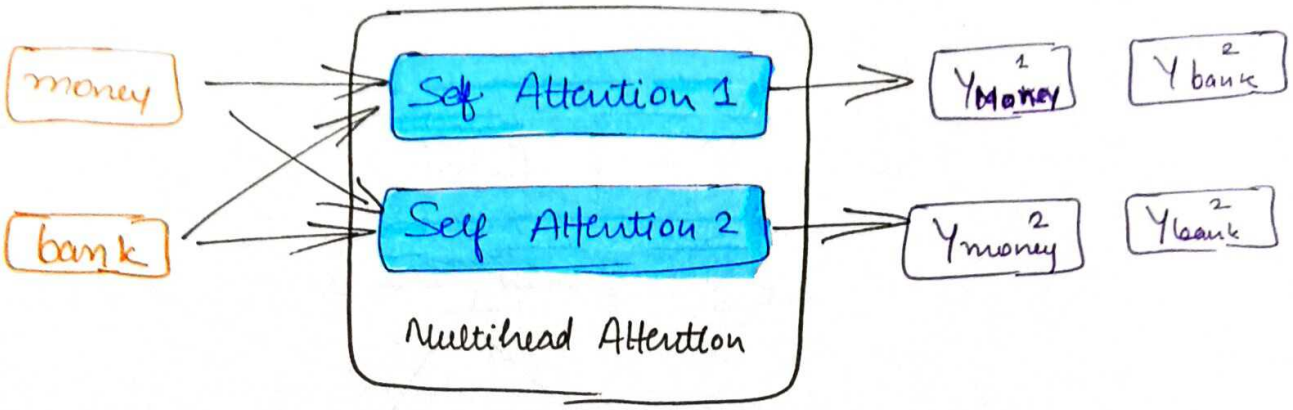
because two self attention


money
e_{money}
$W_q^{\langle 1 \rangle}$
$W_k^{\langle 1 \rangle}$
$W_v^{\langle 1 \rangle}$
2 set
$W_q^{\langle 2 \rangle}$
$W_k^{\langle 2 \rangle}$
$W_v^{\langle 2 \rangle}$
q_{money}^{1}
K_{money}^{1}
V_{money}^{2}
q_{money}^{2}
K_{money}^{2}
V_{money}^{2}

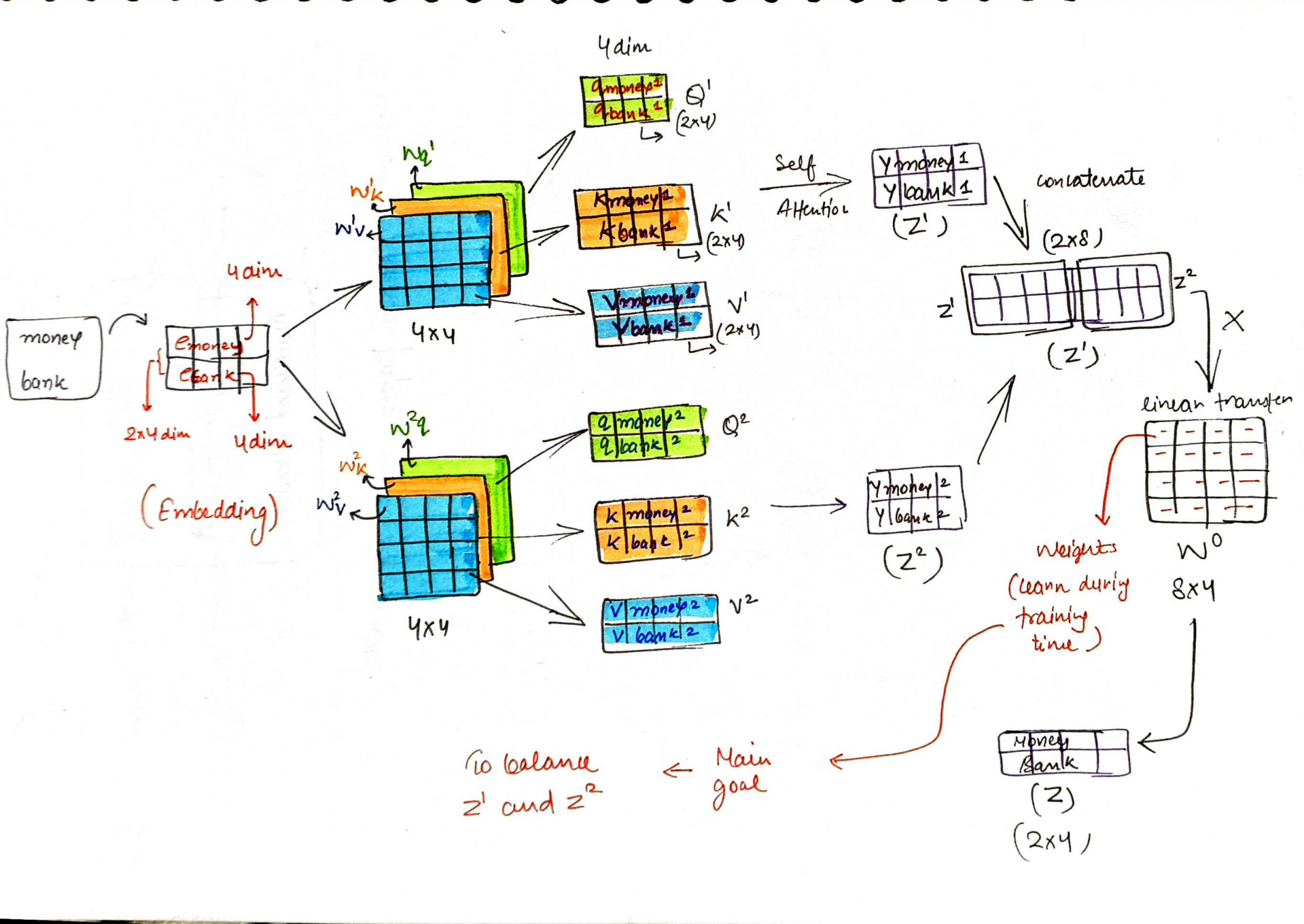

Bank
e_{Bank}
$W_q^{\langle 2 \rangle}$
$W_k^{\langle 2 \rangle}$
$W_v^{\langle 1 \rangle}$
2 set
$W_q^{\langle 2 \rangle}$
$W_k^{\langle 2 \rangle}$
$W_v^{\langle 2 \rangle}$
q_{money}^{1}
K_{money}^{1}
V_{money}^{1}
q_{money}^{2}
K_{money}^{2}
V_{money}^{2}


qmoney1
S11
qmoney1
S12
Kmoney1
Kbank2
S11
S12
Scaling (1/Sqrt(dk))
S11'
S12'
1 set
Softmax
W11
W12
W11X
Vmoney1
+
W12 X
Vbank1
Ymoney1
qmoney2
S11
qmoney2
S12
Kmoney2
Kbank2
S11
S12
Scaling (1/Sqrt(dk))
S11'
S12'
2 set
Softmax
W11
W12
W11X
Vmoney2
+
W12X
Vbank2
Ymoney2
1st contextual representation for money word
2nd contextual representation for money word
All these steps for Money word

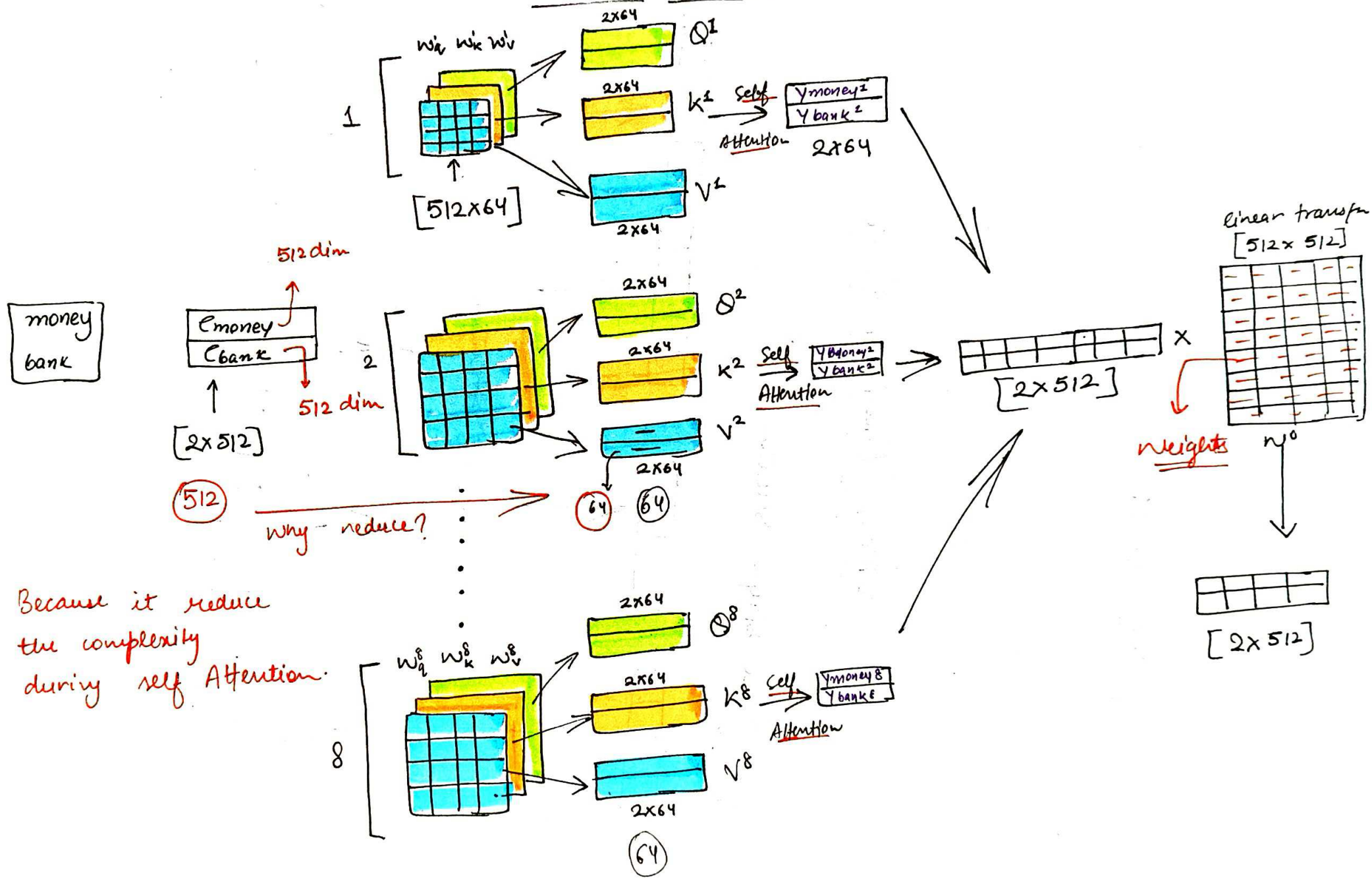

qbank¹
qbank¹
S21
S22
Kmoney¹
Kbank¹
S21
S22
Scaling (1/sqrt(dk))
S21'
S22'
Set 1
Softmax
W21
W22
W21 X
Vmoney¹
+
W22 X
Vbank¹
Y bank¹
1st contextual representation for bank word
qbank²
qbank²
S21
S22
Kmoney²
Kbank²
S21
S22
Scaling (1/sqrt(dk))
S21'
S22'
set 2
Softmax
W21
W22
W21 X
Vmoney²
+
W22 +
Vbank²
Ybank²
2nd contextual representation for bank word.
bank word



Transfomer Research paper → 8 heads


money
bank
4dim
Embedding
2x4 dim
4dim
4x4
4dim
Q'
(2x4)
K'
(2x4)
V'
(2x4)
Self Attention
(Z')
concatenate
(2x8)
Z'
Z²
(Z')
X
linear transform
W⁰
8x4
4x4
Q²
K²
V²
(Z²)
weights (learn during training time)
Money
Bank
(Z)
(2x4)
Main goal
To balance Z' and Z²

Attention is all you Need



Because it reduce the complexity during self Attention.

Benifit of Self Attention

Parallel Training

Drawback of Self Attention

word Order

eg:- "Nitish killed lion" is same "Lion killed Nitish"

Positional Encoding in Transformer

Proposing a Simple Solution