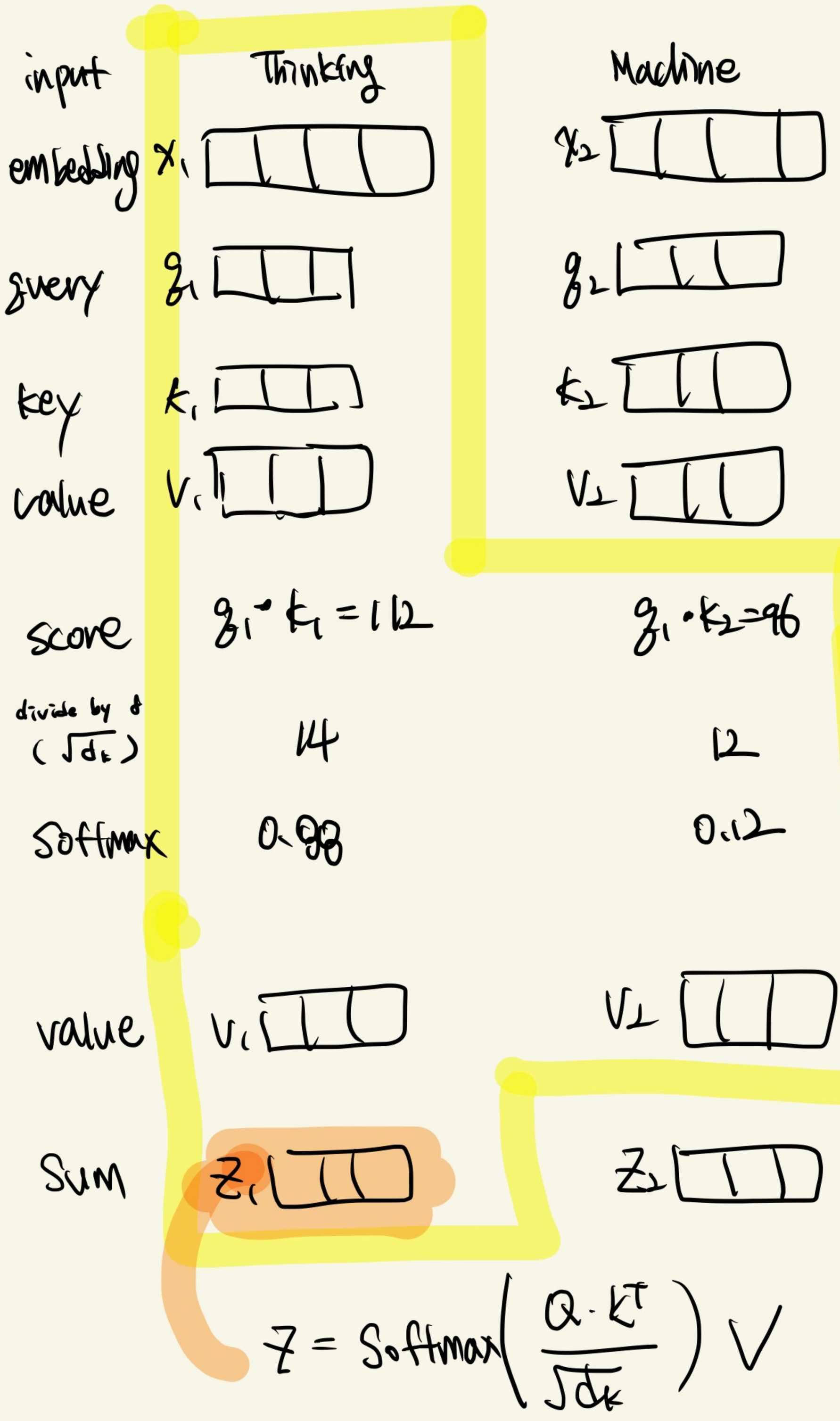
Blockdiagram_transformer-01.jpg

#1 $x_1 \cdots x_n \rightarrow z_1 \cdots z_n$ process



Blockdiagram_transformer-02.jpg

#2 query - key - value

Z

W_O

concatenate

Z_0 Z_1 Z_n

Q_0 K_0 V_0

Q_1 K_1 V_1

Q_n K_n V_n

W_0^Q W_0^K W_0^V

W_1^Q W_1^K W_1^V

W_n^Q W_n^K W_n^V

Head_num

Split into 8 heads

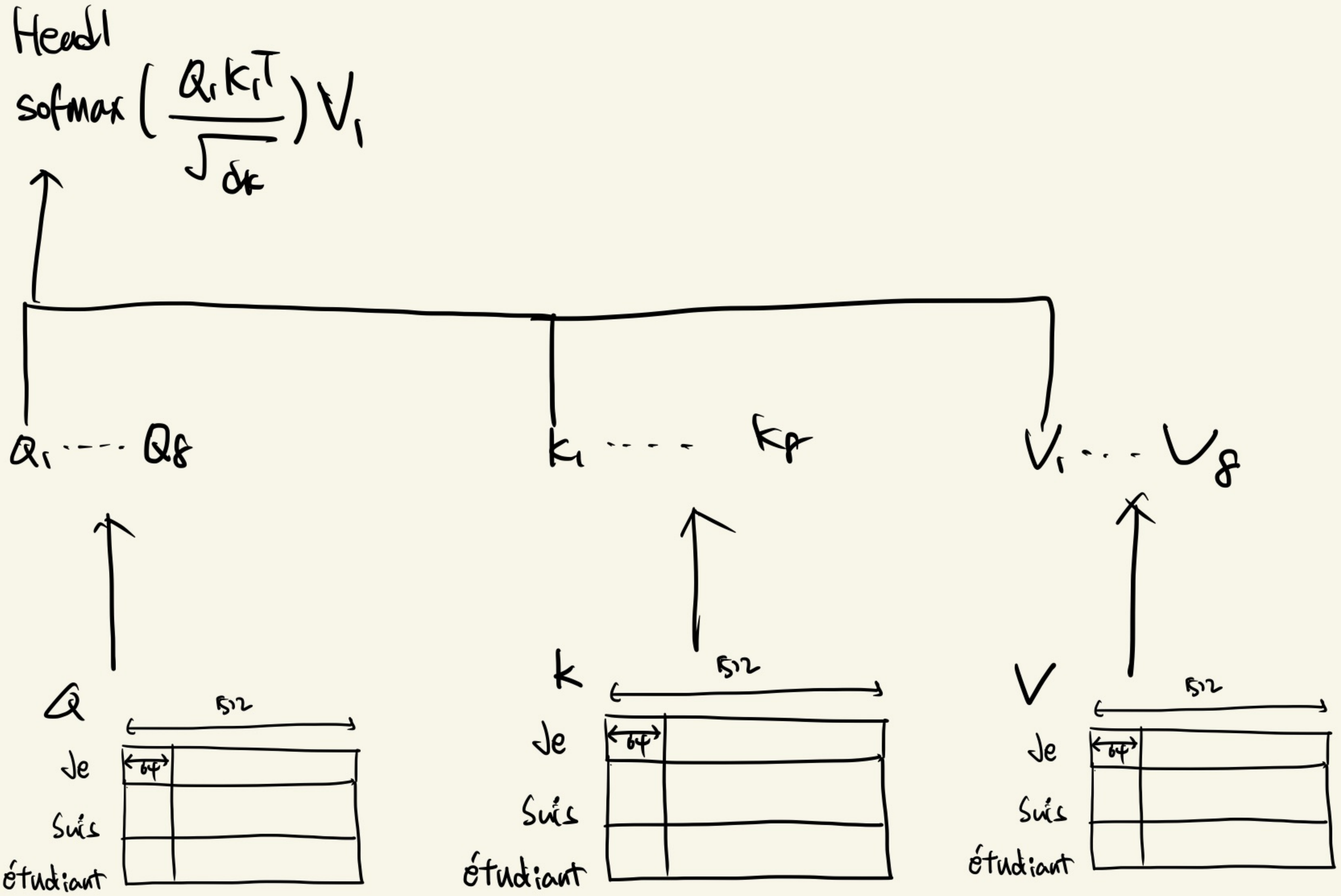
multiply X or R with weight matrix

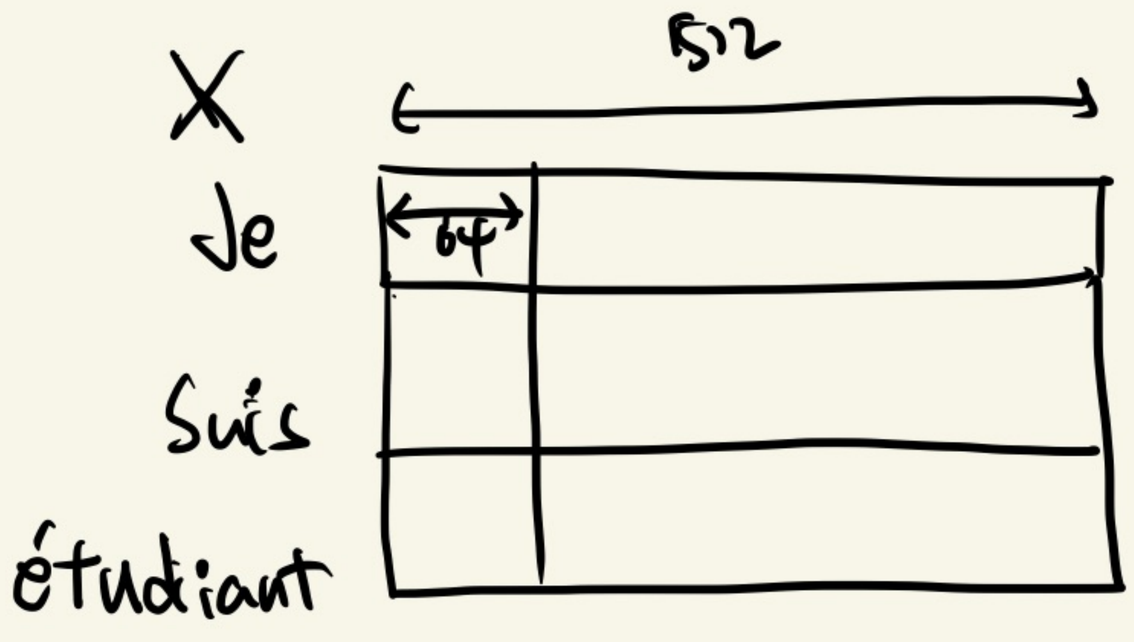
Embedded Vector

X or R

Sentenses

Blockdiagram_transformer-03.jpg





Blockdiagram_transformer-04.jpg

#3 Encoder

Add & Normalize

Feed Forward

Feed Forward

z_1

z_2

LayerNorm(X + Z)

z_1

z_2

Self-Attention

x_1

x_2

Encoder #1

Positional Encoding

x_1

Thinking

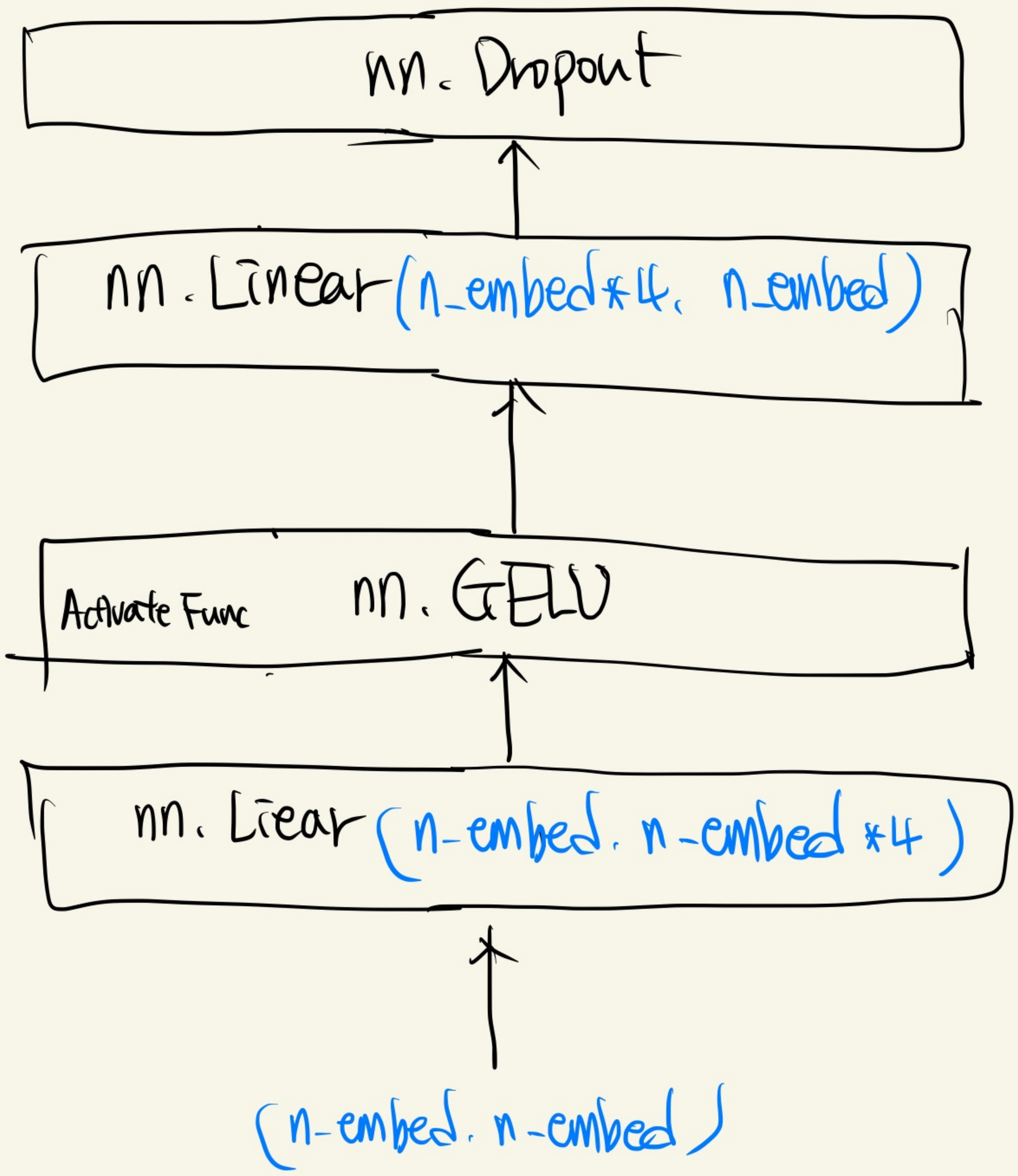
Positional Encoding

x_2

Machines.

Blockdiagram_transformer-05.jpg

#4. Feed Forward .



Blockdiagram_transformer-06.jpg

#4 Entire block

Kencdec Vencdec

Encoder #2

Add & Normalize

Feed Forward

Feed Forward

Add & Normalize

Self-Attention

Encoder #1

positional Encoding

x_1 Thinking

x_2 Machine

Softmax

Linear (Logit)

Decoder #2

Add & Normalize

Feed-Forward

Feed Forward

Add & Normalize

Encoder-Decoder Attention

Add & Normalize

self-Attention

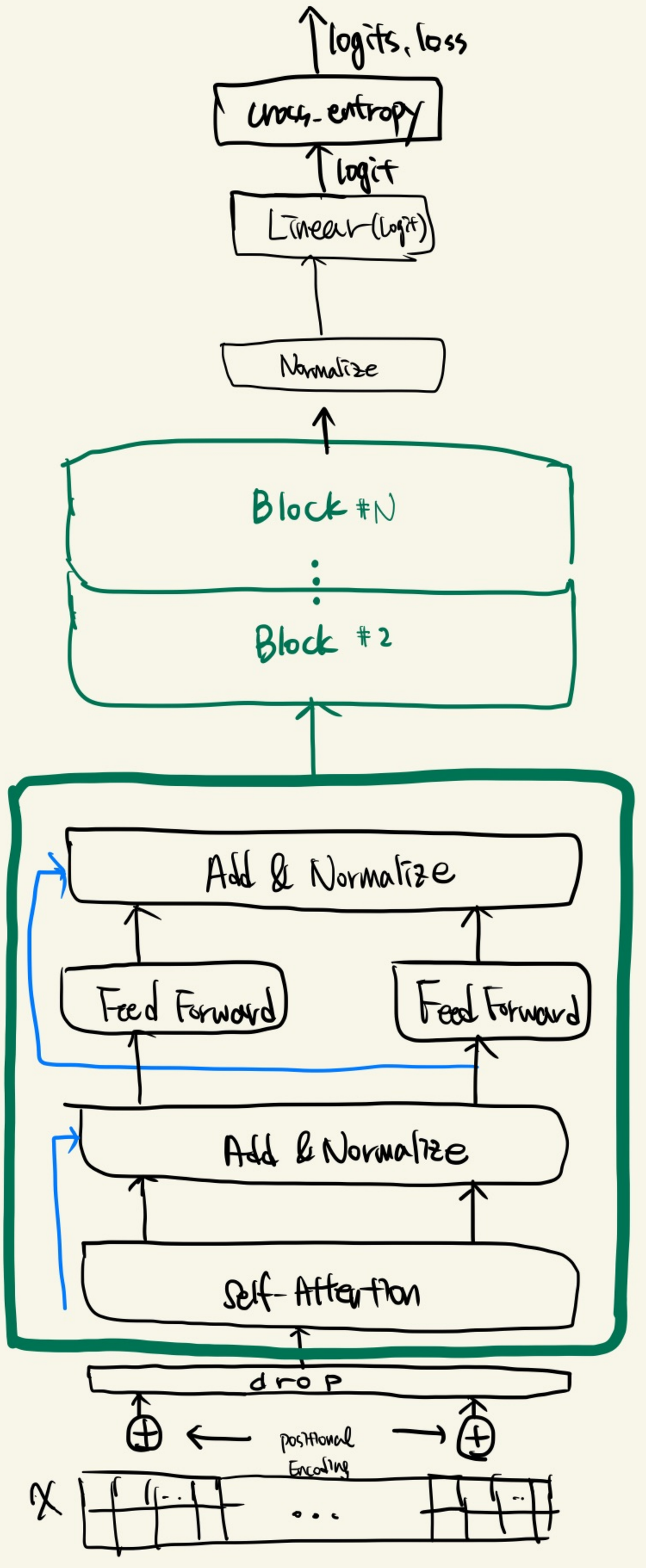
Decoder #1

* end of sentence 까지 반복

* earlier position was allowed (-inf)

Blockdiagram_transformer-07.jpg

#4 A5 GPT - decoder concept - unidirectional .

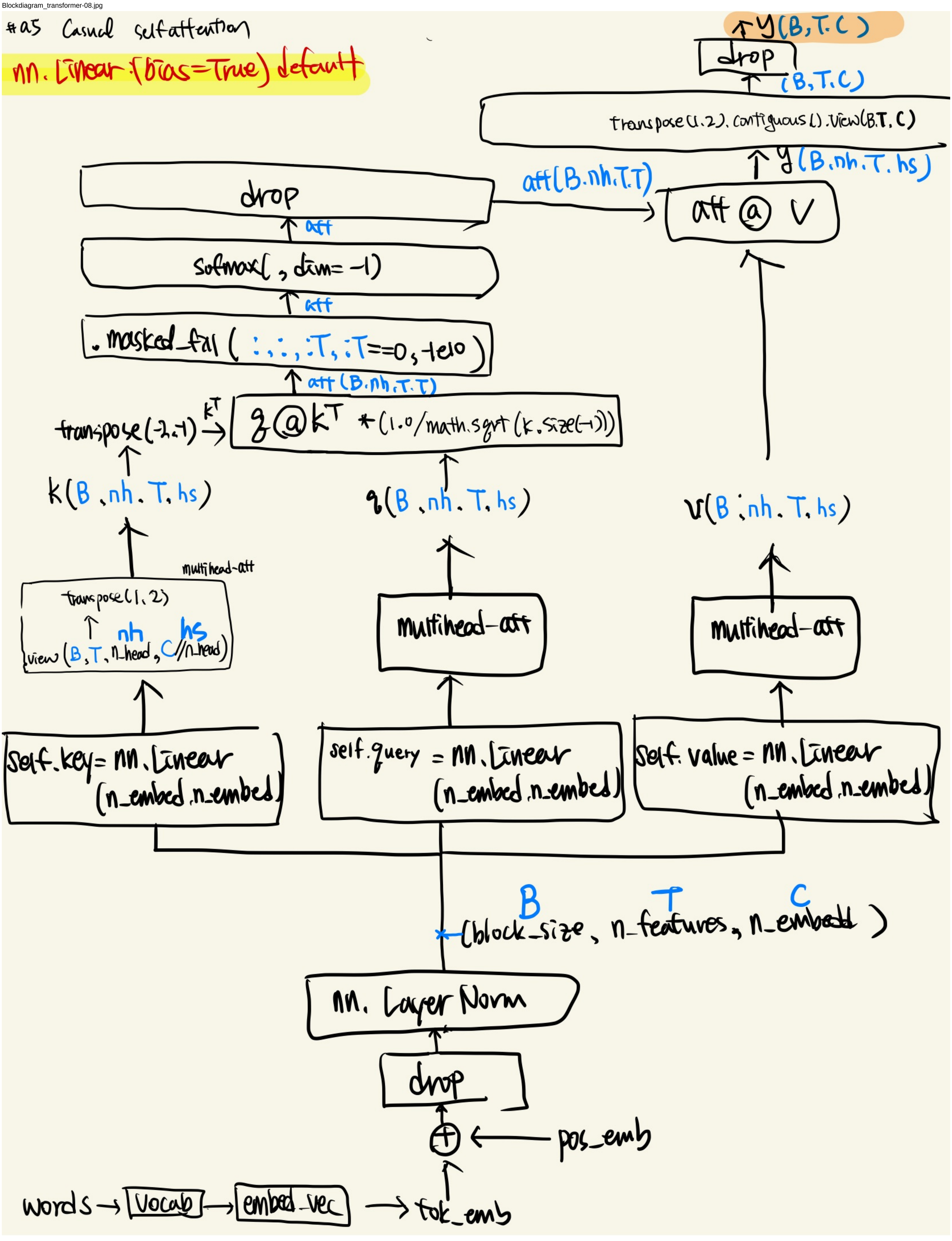


Block#1

Blockdiagram_transformer-08.jpg

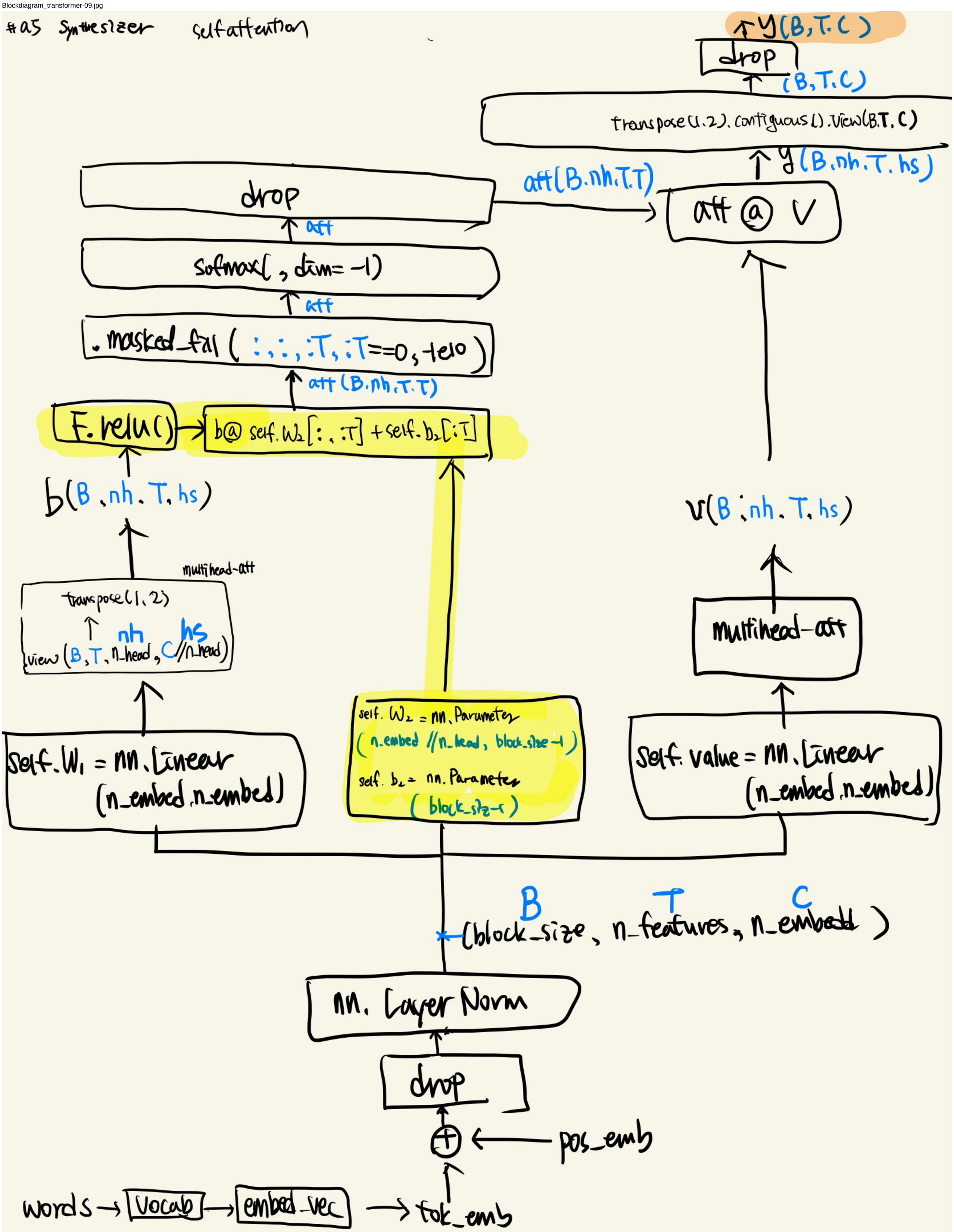
#a5 Casual selfattention

nn.Linear(bias=True) default



Blockdiagram_transformer-09.jpg

#a5 Synthesizer selfattention



Blockdiagram_transformer-10.jpg

#a5 Synthesizer — selfattention (equation version)