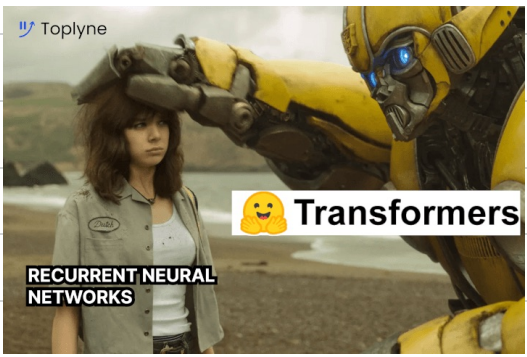
Session 10

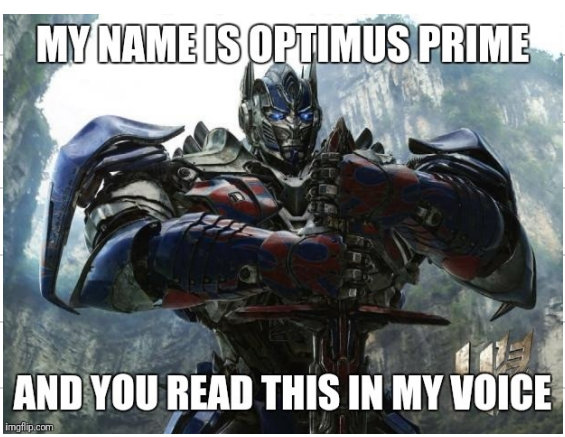
Transformers

May 06, 2024



Agenda

1. Seq2Seq Learning
2. Neural Machine Translation
3. Transformer
   a. Encoder
   b. Decoder



Self attention.

→ Seperate For K, Q, V

Input → (10, 512) → W_Q → (10, 512)

W_K → (10, 512)

W_V → (10, 512)

w_1 = Thinking → (1, 512)

w_2 = Machine → (1, 512)

(2, 512)

Create $W(K, W_Q, W_V)$ → (512, 512)

(512, 512)

↓

(512, 512)

Thinking → (1, 512)

Machine → (1, 512)

K_1

K_2

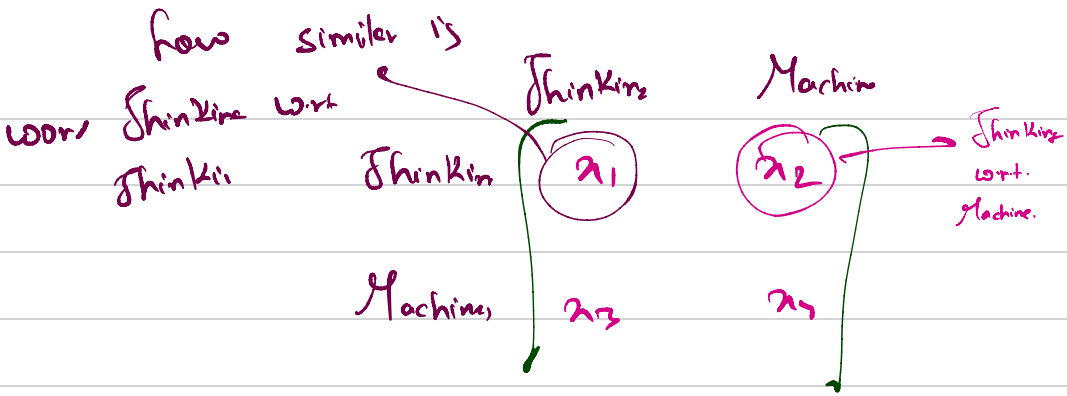
K = (2, 512)

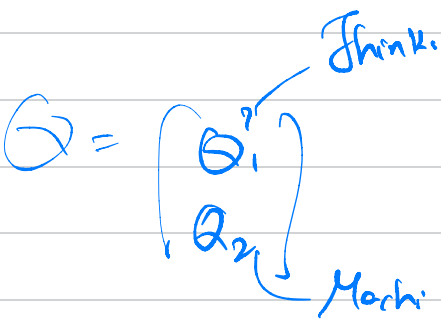
Q = (2, 512)

V = (2, 512)

Attention Scores = $\text{Softmax}\left(\frac{K \cdot Q^T}{\sqrt{512}}\right)$ = [ ] 2×2

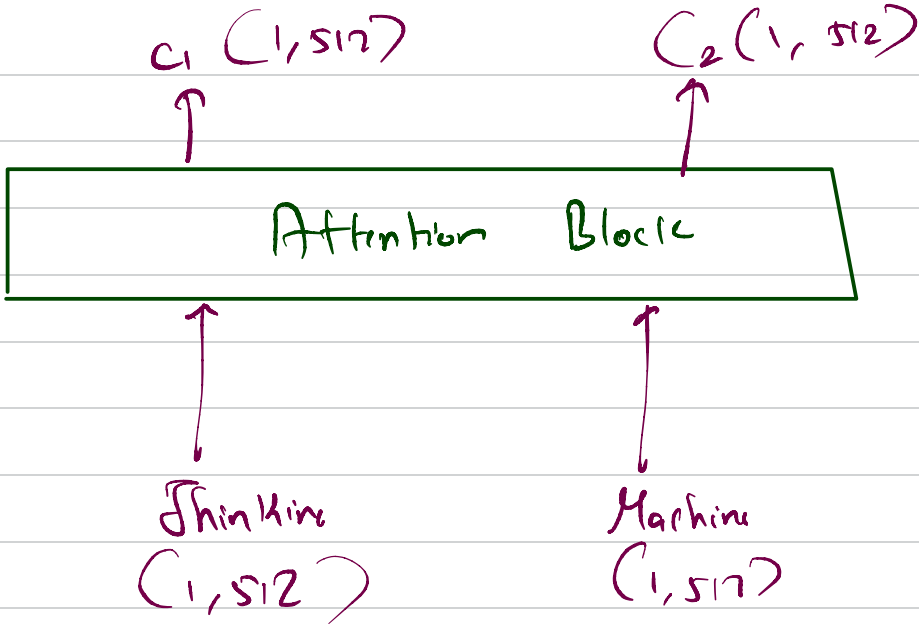
(2, 512) (512, 2)





Content Vector 1 (Thinking) = $V_1 \times x_1 \oplus V_2 \times x_2 = (1, 512)$

Content Vector 2 (Machines) = $V_1 \times x_3 \oplus V_4 \times x_4 = (1, 512)$



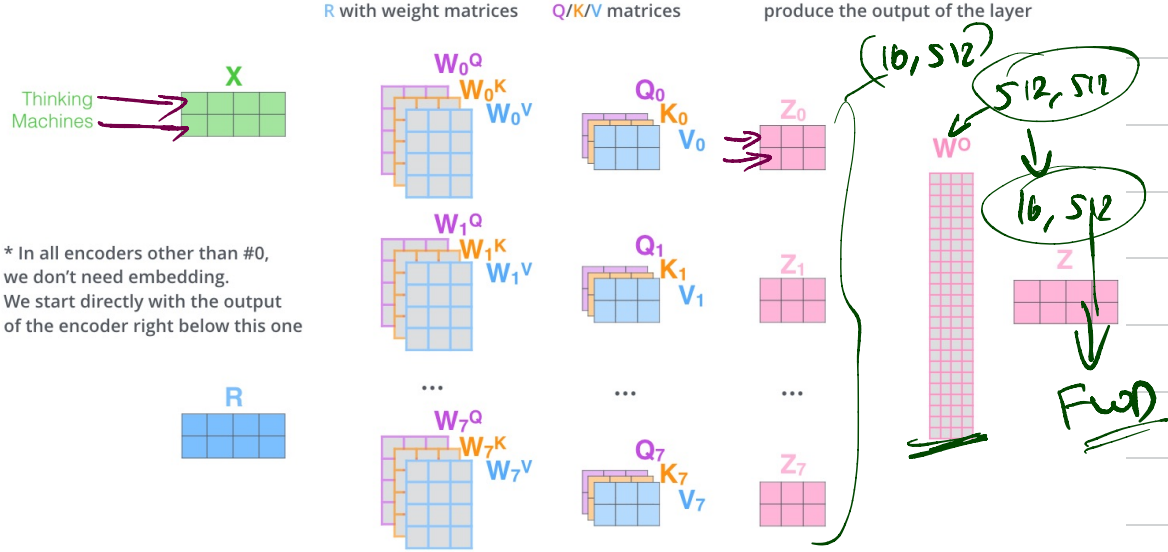
1) This is our input sentence*

2) We embed each word*

3) Split into 8 heads. We multiply X or R with weight matrices

4) Calculate attention using the resulting Q/K/V matrices

5) Concatenate the resulting Z matrices, then multiply with weight matrix W^O to produce the output of the layer



* In all encoders other than #0, we don't need embedding. We start directly with the output of the encoder right below this one

Output

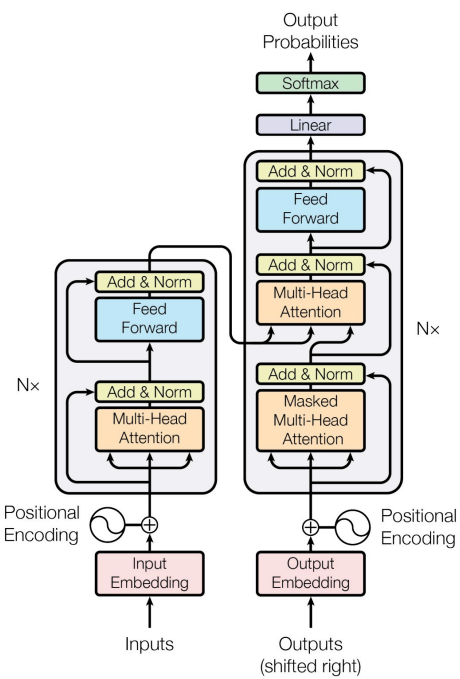
1st attention block = (2, 512)

↓

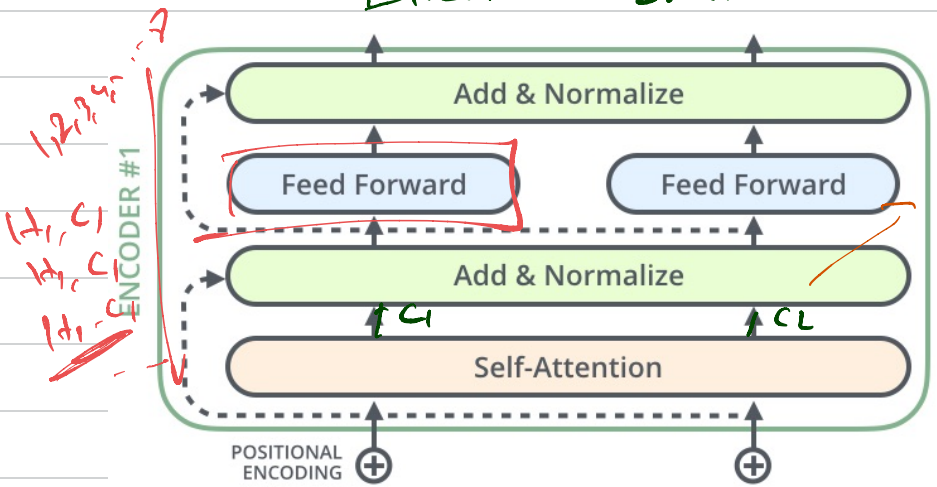
8th attention block = (2, 512)

↓

(16, 512)

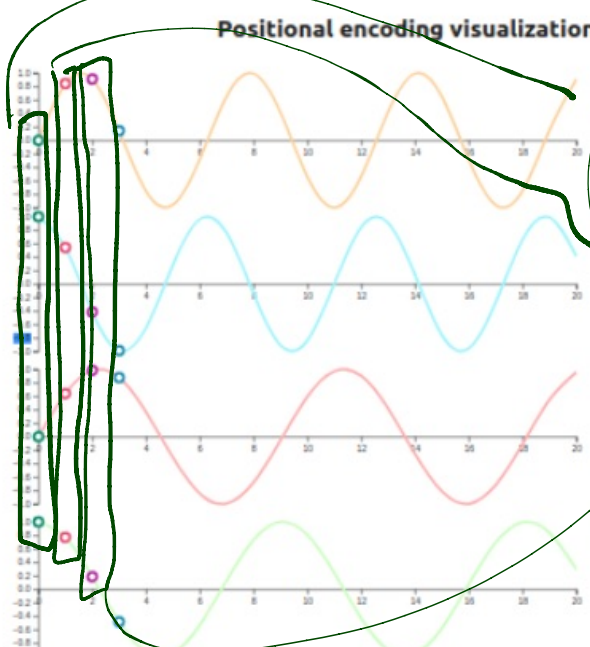


Encoder Block



Layer norm

Sine Cosine values



	p0	p1	p2	p3	
	0.000	0.841	0.909	0.141	i=0
	1.000	0.540	-0.416	-0.990	i=1
	0.000	0.638	0.983	0.875	i=2
	1.000	0.770	0.186	-0.484	i=3

Positional Encoding

$$PE_{(pos,2i)} = sin(\frac{pos}{10000^{2i/d_{model}}})$$

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

Settings: d = 50

The value of each positional encoding depends on the *position* (*pos*) and *dimension* (*d*). We calculate result for every *index* (*i*) to get the whole vector.

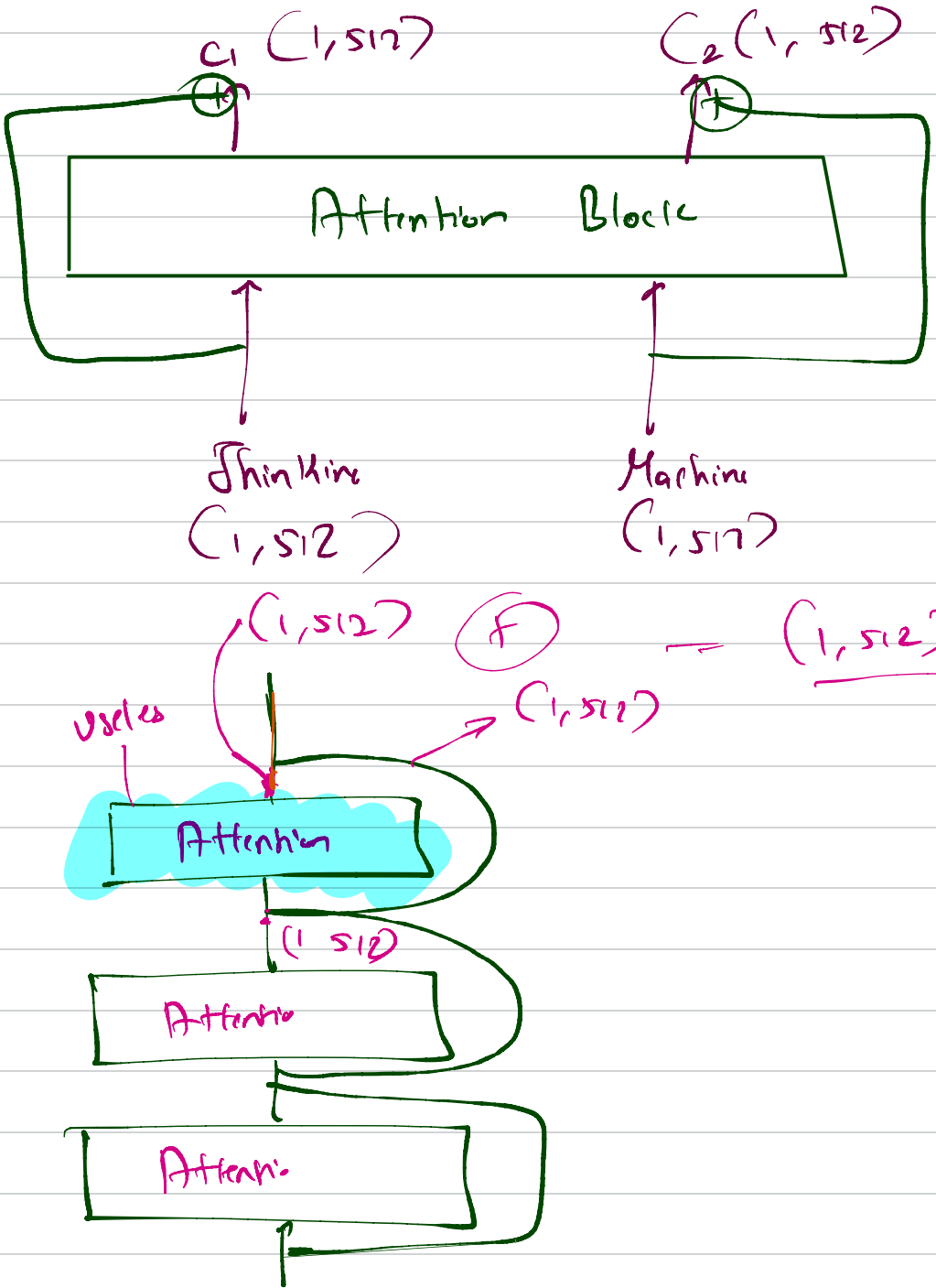
$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$

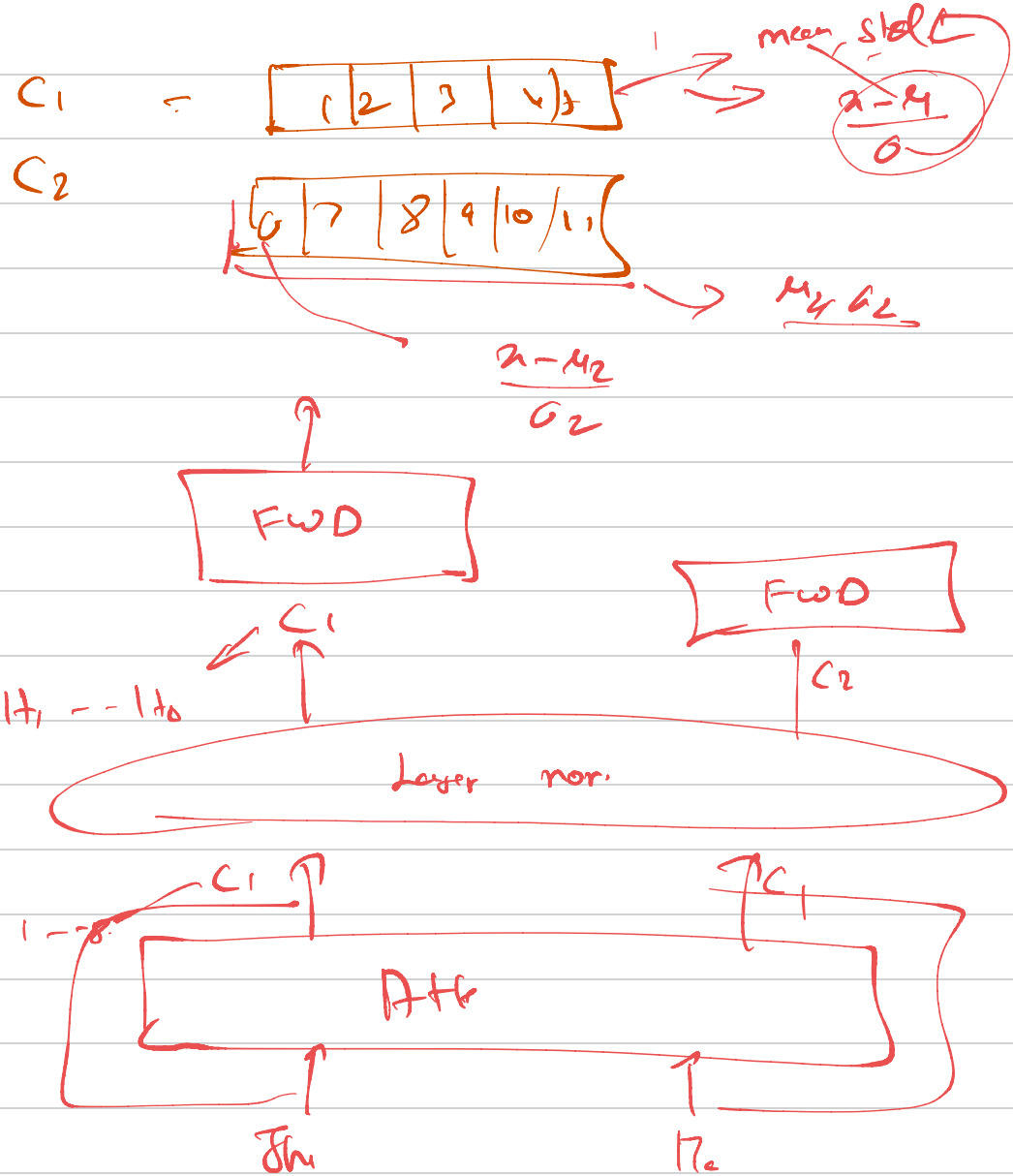
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

pos0 — pos1 — pos2 — pos3

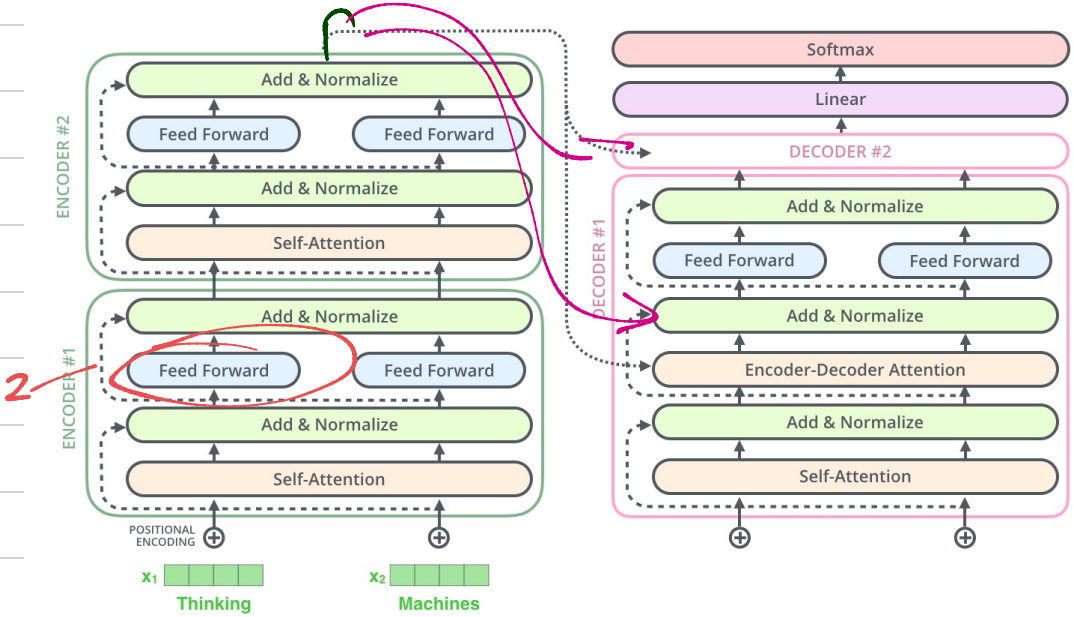
w_1 w_2 w_3 (w_4) w_5

w_1 w_2 - - - w_{80} w_{10}


C1 (1, 512)
C2 (1, 512)
Attention Block
Thinking
(1, 512)
Machine
(1, 512)
(1, 512)
f
= (1, 512)
(1, 512)
Useless
Attention
(1 512)
Attentio
Attenti.


C1
1 2 3 4
mean std
x-μ1 / σ
C2
6 7 8 9 10 11
μ2 σ2
x-μ2 / σ2
FWD
FWD
C1
C2
Layer nor.
C1
C1
DH
Ih
Ic

(2, 512)



2

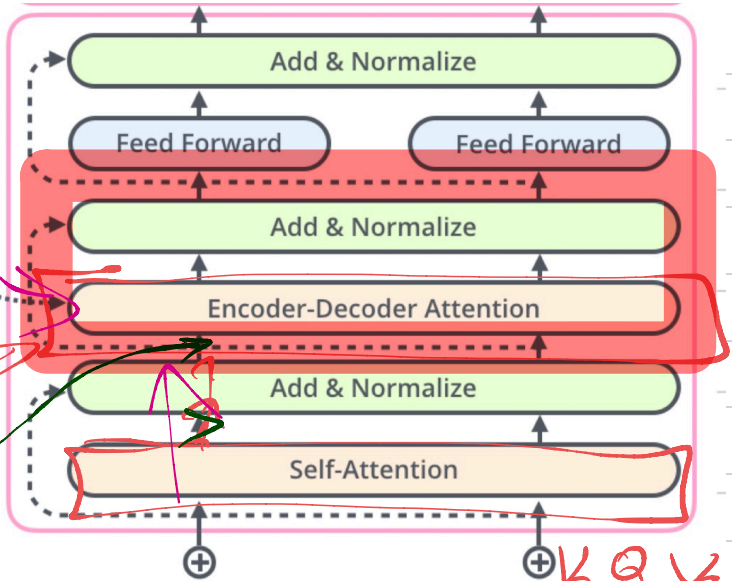
$C_1 \rightarrow (1, 512)$

(2, 512)

$W_{KQV} \rightarrow (2, 512)$

$[512, 512]$

$(512, 512) * W_{VD} \rightarrow (2, 512)$



K, V, Q

Encoder

K, Q, V

<SOS> → (1, 512)

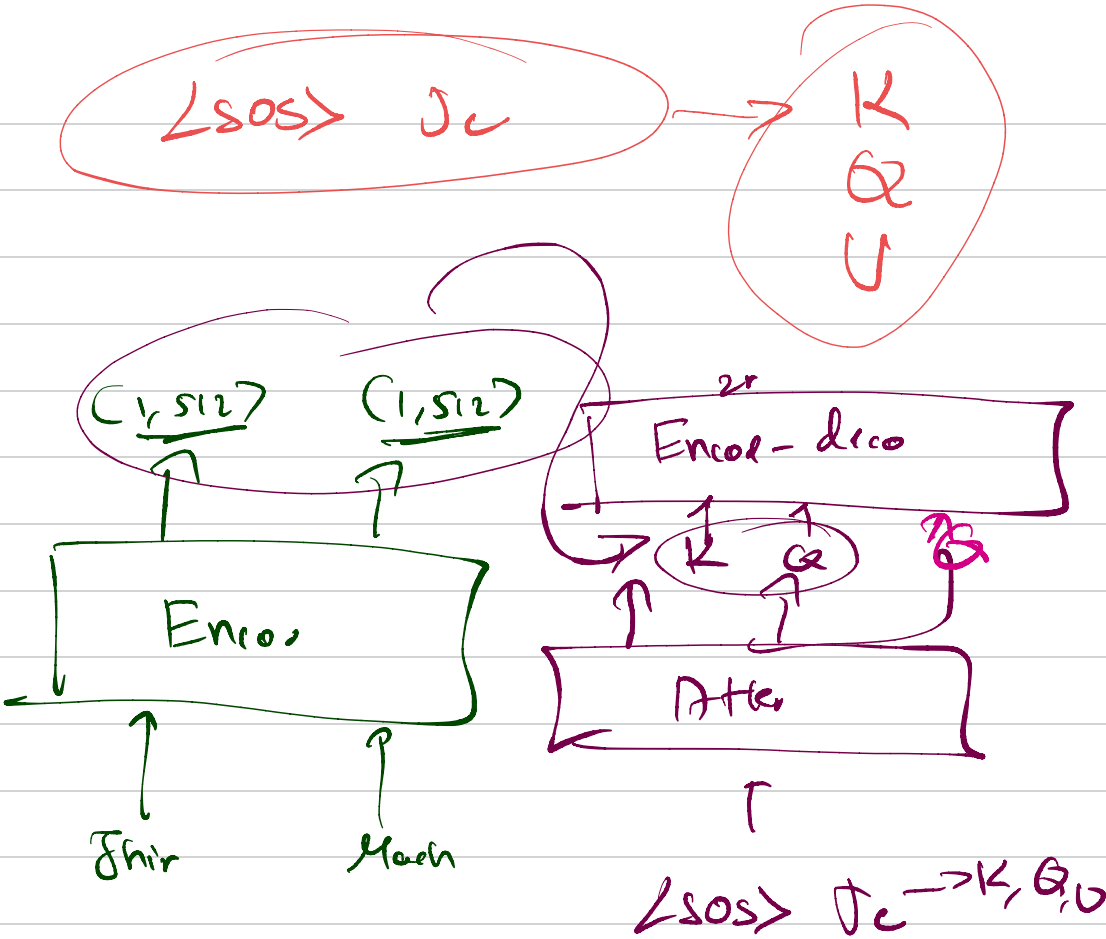
I love apple → K, V

Je aime bateau

<SOS>

t = 2 → Je

I in Fren

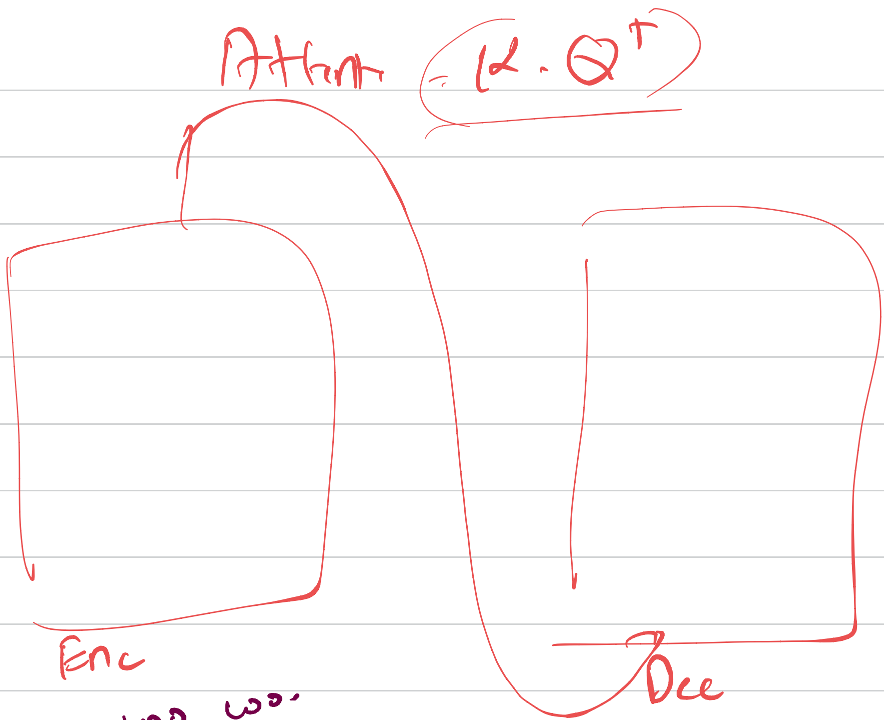


Input → (128K, 512)

W_Q = (512, 512) | K = (128K, 512)

W_K = (512, 512) | Q = (128K, 512)

W_V = (512, 512) | V = (128K, 512)



100 wo...

gone

I ha — . . . economic

Economy

$Q^T \cdot K$

$\frac{\partial \text{ Economy}}{\partial \text{ Economique}}$ → max Rel in loss

$\frac{\partial \text{ Economy}}{\partial}$ / ha — . . . gone