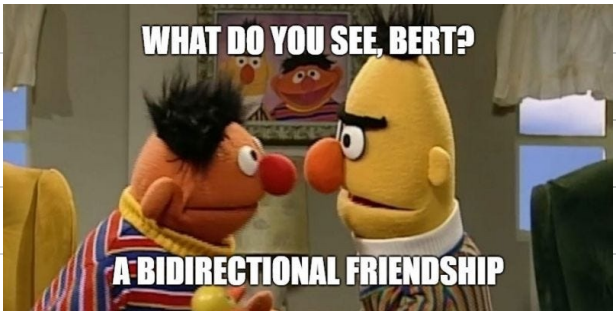
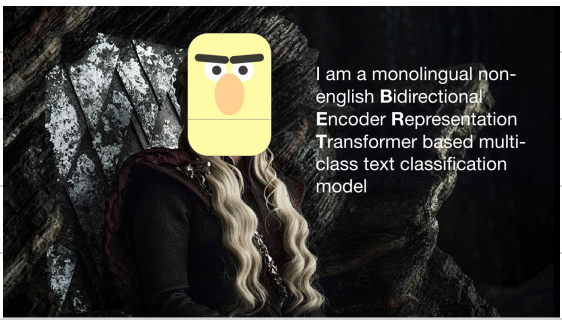
Session 11

BERT

May 08, 2024





Agenda

1. GRU
2. Transformer Recap End to End
3. Transformer in training
4. BERT
5. Masked Language Modelling

Transformer – End-To-End

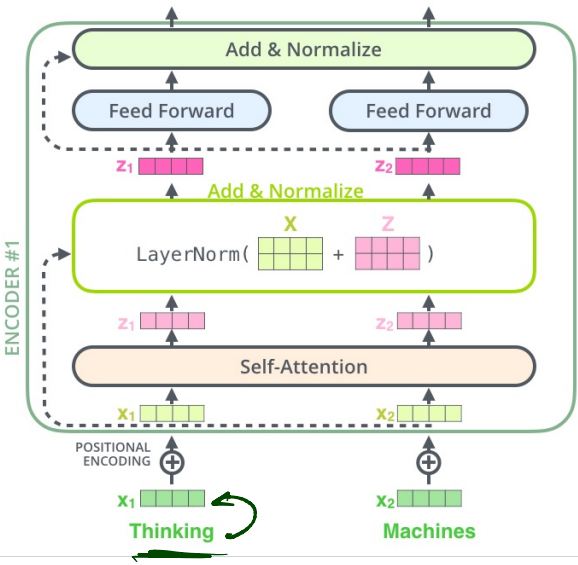
→ Inference

https://jalammar.github.io/illustrated-transformer/

"I love Pizza"

↓

"Me encanta la Pizza"

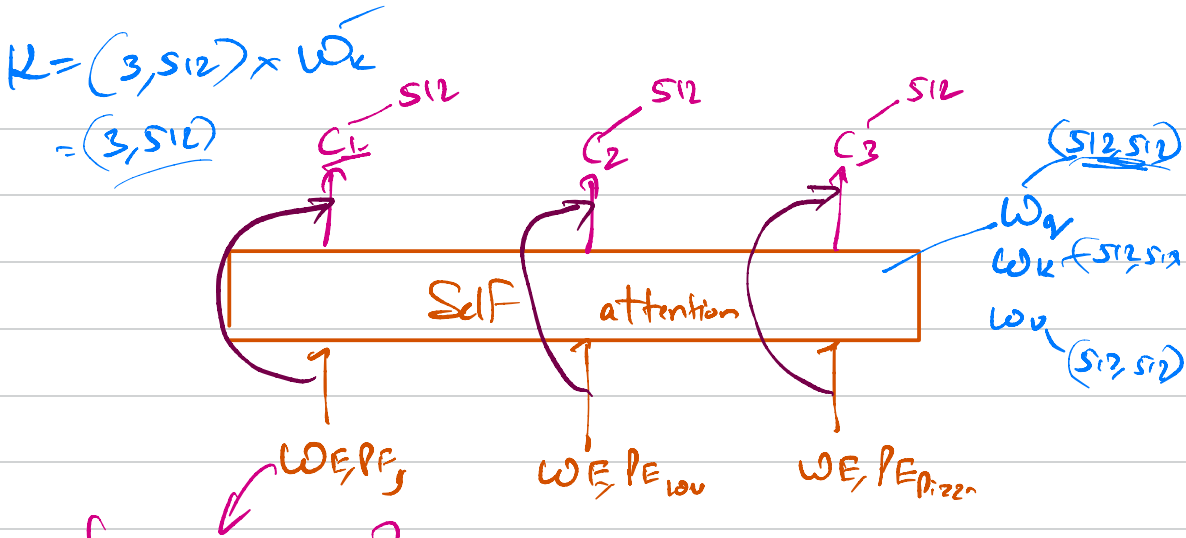


(1) (3, 512)

512

$$\begin{bmatrix} \text{Word}_1 \\ \text{Word}_2 \\ \text{Word}_3 \end{bmatrix} + \begin{bmatrix} PE_{w_1} \\ PE_{w_2} \\ PE_{w_3} \end{bmatrix} = \begin{bmatrix} W_{E,PE_I} \\ W_{E,PE_{love}} \\ W_{E,PE_{Pizza}} \end{bmatrix}$$

(3, 512)



[ ] 512

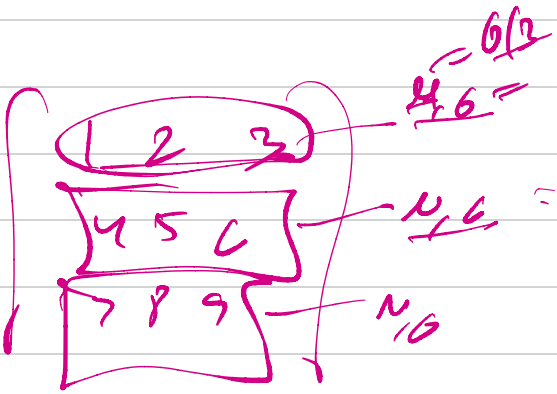
4 / 6

$WE, PE_Q + C_1$ — Layer normalization → (1, 512)

$WE, PE_{low} + C_2$ — " " → (1, 512)

$WE, PE_{Pizza} + C_3$ — " " → (1, 512)

output (3, 512) context vectors

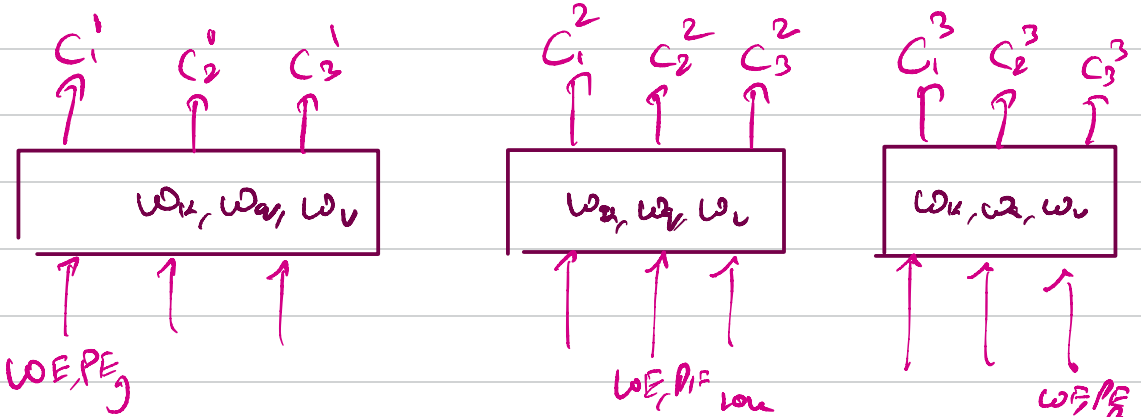


$\frac{1-2}{G_1}$, $\frac{2-2}{G_2}$, $\frac{3-2}{G_2}$

$\frac{4-5}{G_2}$, $\frac{5-5}{G_2}$, $\frac{6-5}{G_2}$

$\frac{7-8}{G_3}$, $\frac{8-8}{G_3}$, $\frac{9-8}{G_3}$

Assum head = 3



(1, 512)

head no

word no

I → C_1^1 C_1^2 C_1^3

love C_2^1 C_2^2 C_2^3

pizza C_3^1 C_3^2 C_3^3

(1, 1536)

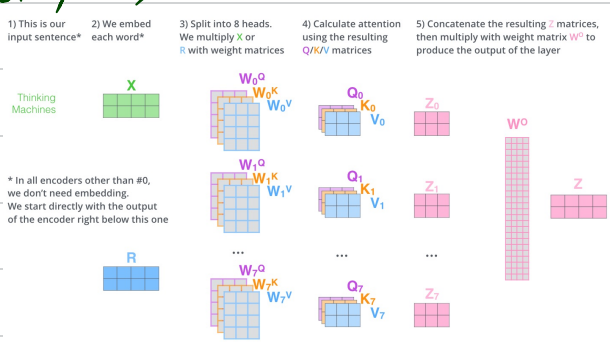
(1, 1536)

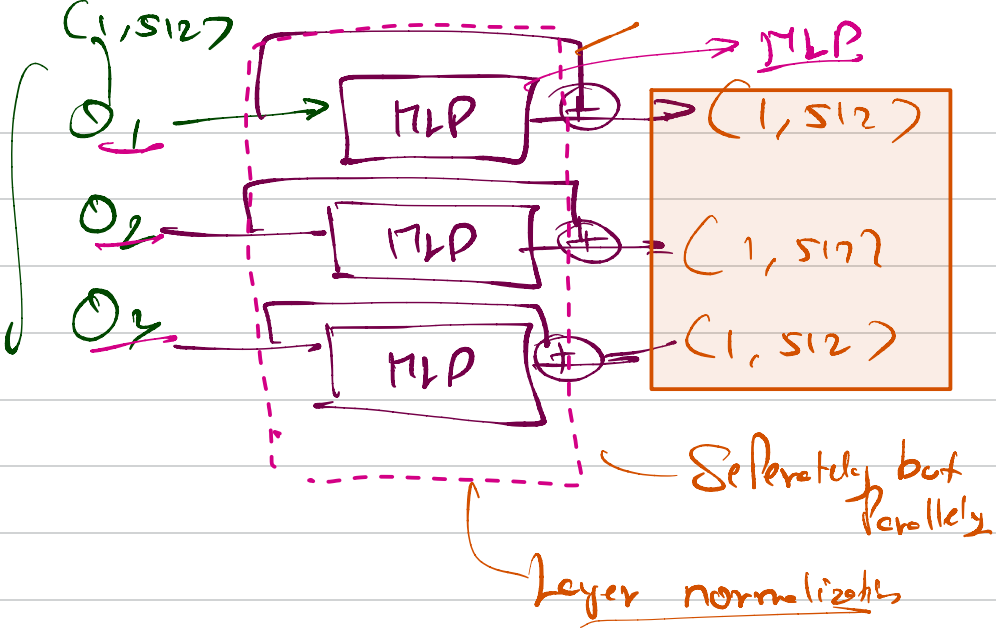
(1, 1536)

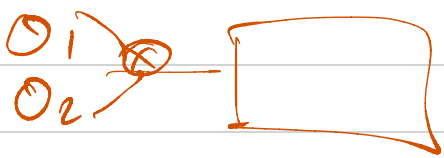
(1536, 512)

(3, 1536) · W_O

= (3, 512)

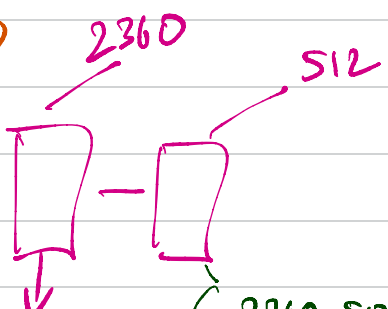






output = (3, 512)

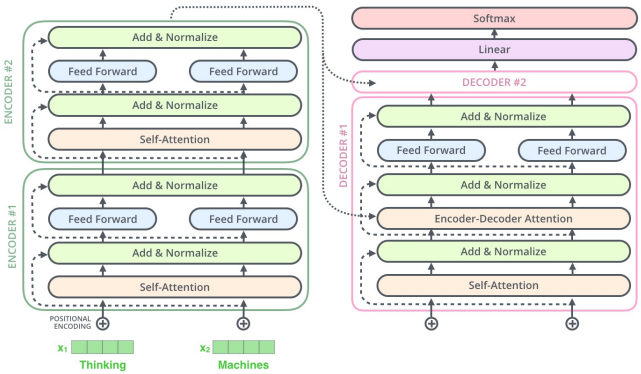
(2,512) × (512, 2360)



O1 → (1,512)

= 1, 2360

Decoder - Inferencing → Single Head



At t=1 <sos> → Word Embedding

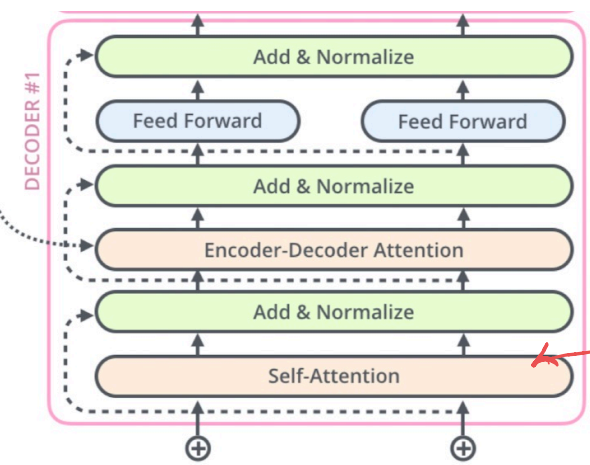
↓

(1, 512)

+

PE

$W_E, PE_{<sos>}$



C_1

Attention blo

W_K, W_Q, W_V

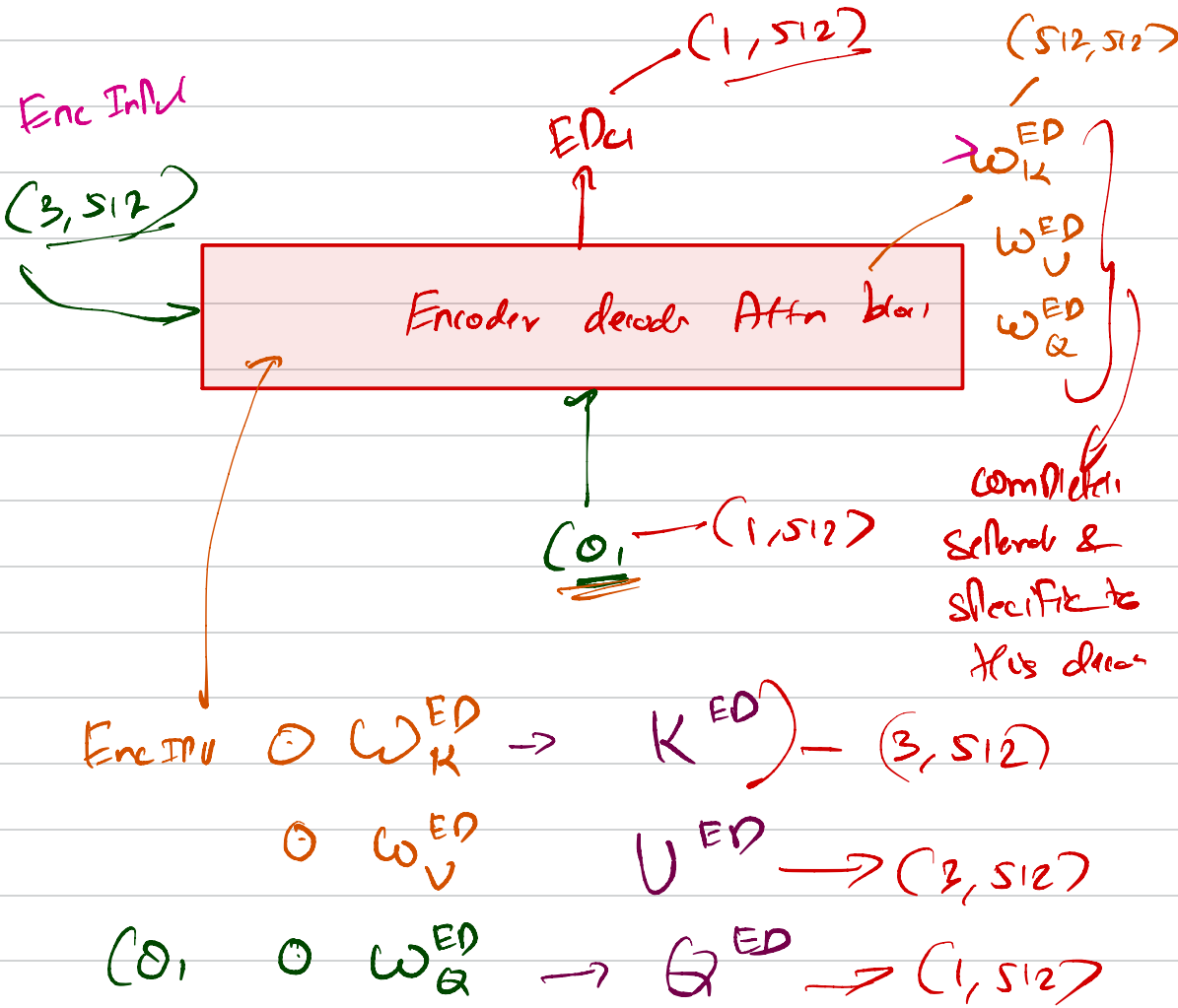
$W_E, Pos_{<sos>}$

Completely Separate

1,512 1,512

Layer normalize $(C_1 + W_{E,Pos_{<sos>}})$ $= C_{0,1}$ $(1, 512)$

Let's talk about encoder decoder attention block



What we will do, is compute similarity of each word of query (which for us is just one word) with respect to each word in key-matrix

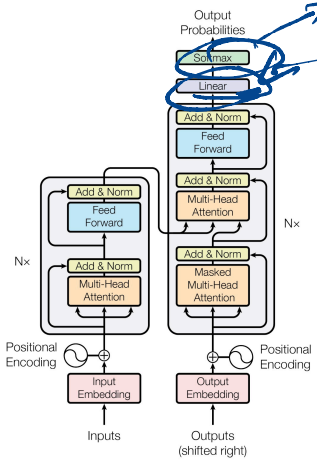
Compute similarity, which will be 3 numbers, corresponding to how similar each query word is wrt to each key word,

then s1 * v1 + s2 * v2 + s3 * v3; where s1, s2 and s3 are softmax out.

Layer Norm (EO_{t-1} + CO_t) → O_t

O_t → [MLP ⊕] → (1, 512) → FC1 → (1, 5177)

Vocabulary



Word at timestep = 2

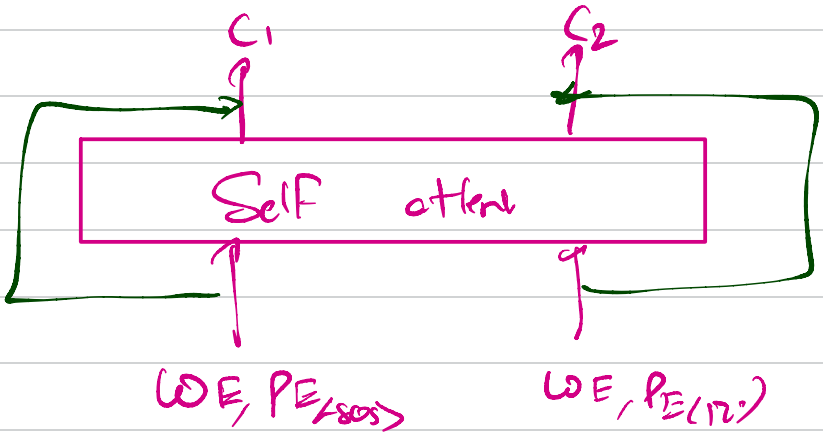
↓

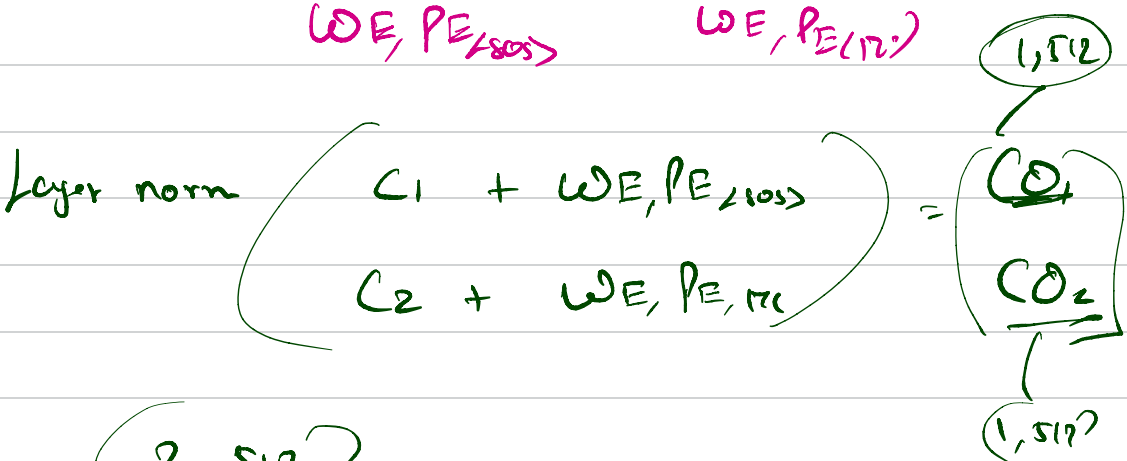
"He"

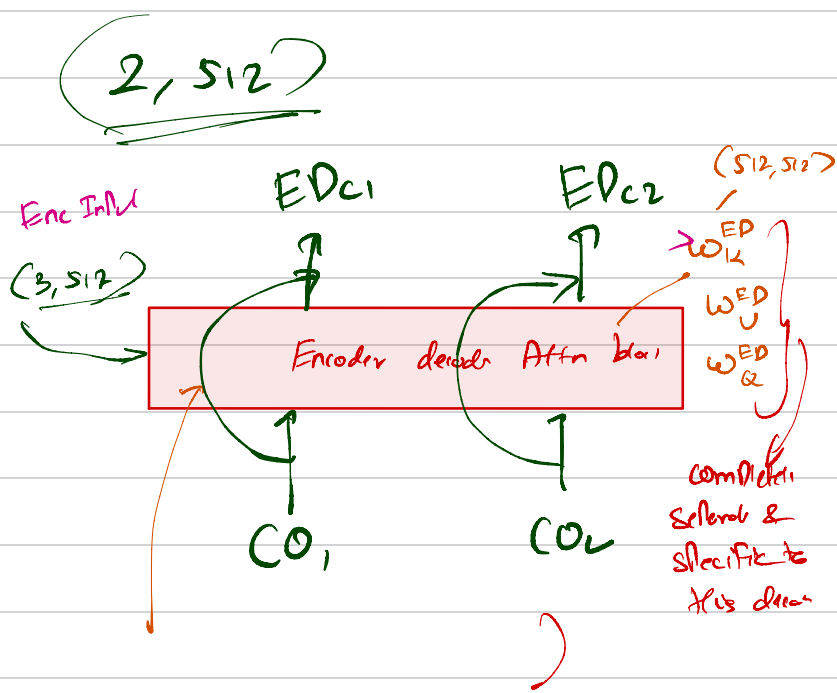
At timestep 2: <SOS> & "He"

<SOS> → word → + PE → $W_{E,PE_{<SOS>}}$

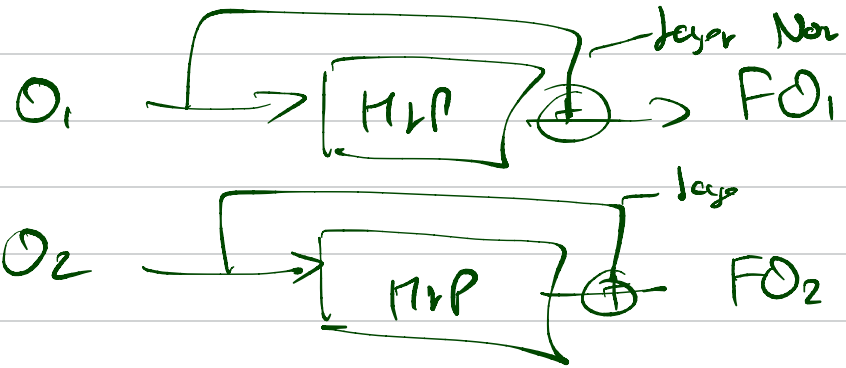
"He" → word → + PE → $W_{E,Pos_{He}}$

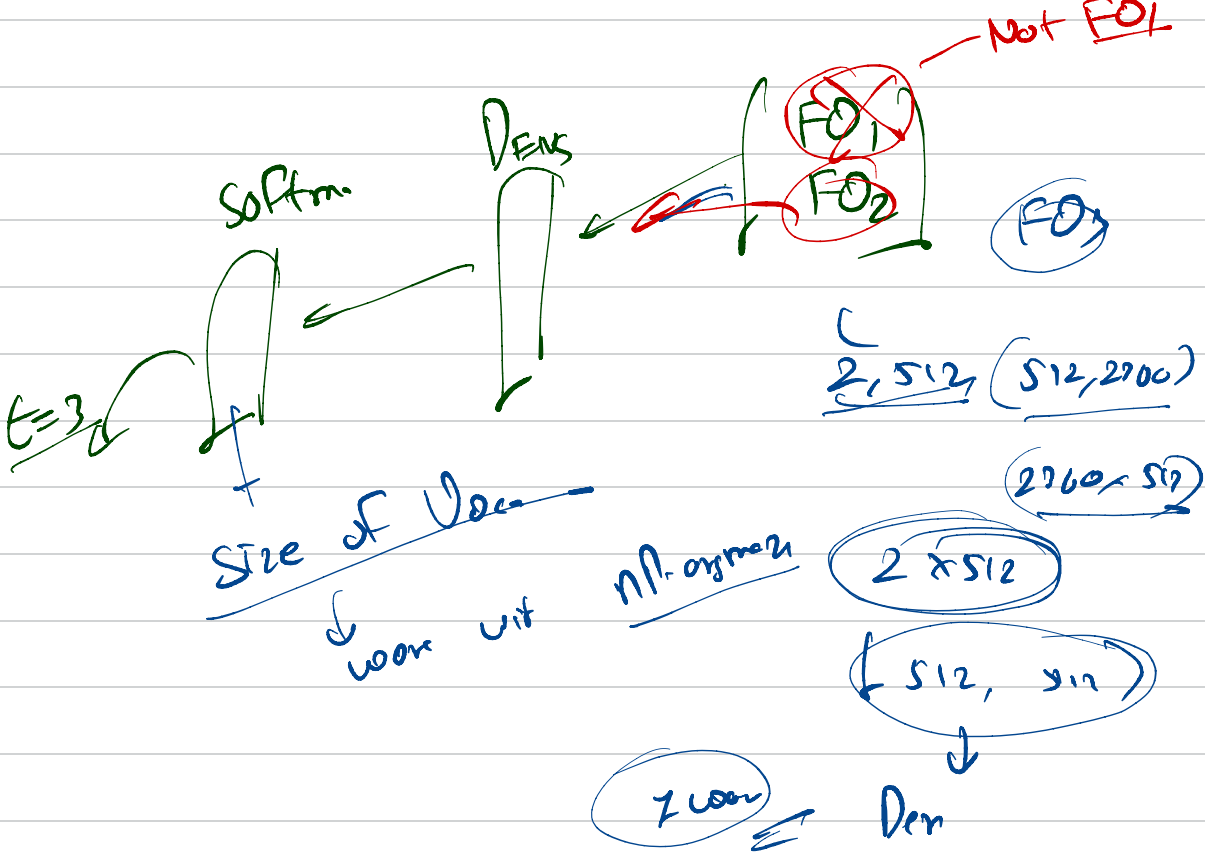

C1
C2
Self attent
WE, PE(105)
WE, PE(17)


Layer norm
C1 + WE, PE(105)
C2 + WE, PE, 17(
=
CO1
CO2
1, 512
1, 512


2, 512
Enc Inp
(3, 512
EDc1
EDc2
(512, 512)
Encoder decoder Attn block
CO1
CO2
compute selected & specific to this decoder

Layer Norm $\begin{pmatrix} C_1 + EDC_1 \\ C_2 + EDC_2 \end{pmatrix} = \begin{pmatrix} O_1 \\ O_2 \end{pmatrix}$

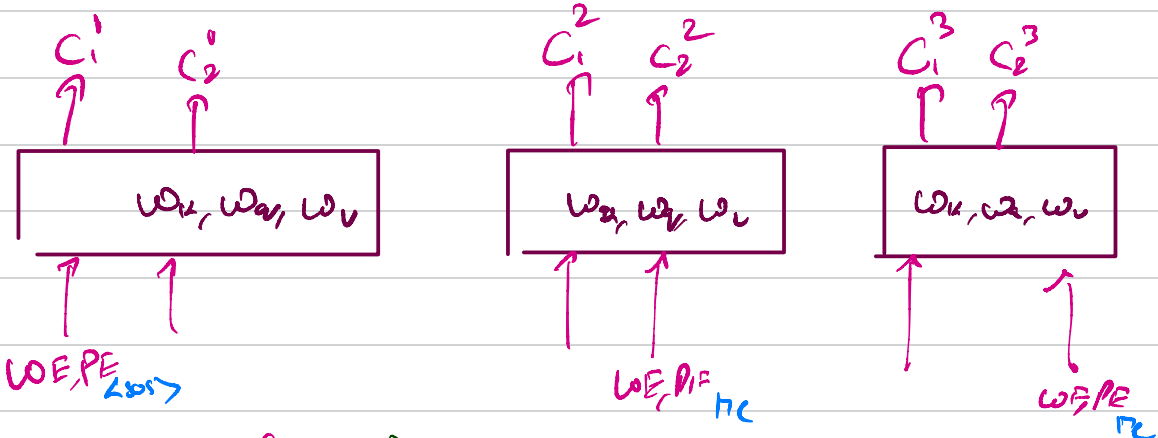




Multi head decoder

t=2 , <sos> , "Me"

Assum head = 3



(1, 512)

$$\begin{bmatrix} c_1^1 & c_1^2 & c_1^3 \\ c_2^1 & c_2^2 & c_2^3 \end{bmatrix} = \begin{matrix} (1, 1536) \\ (1, 1536) \end{matrix}$$

512 x 3

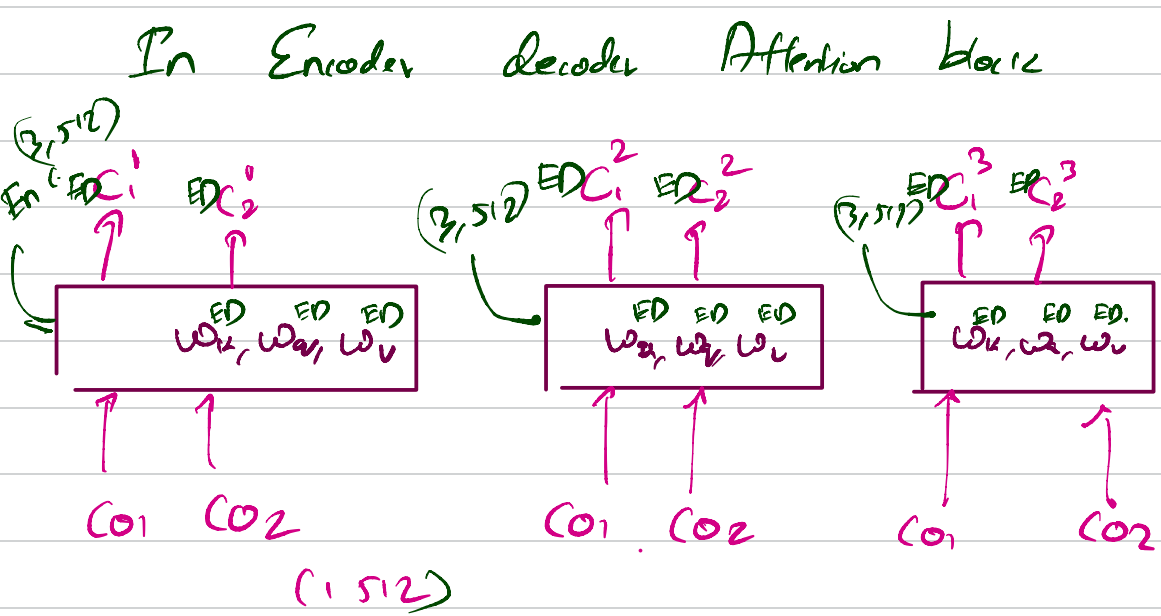
$$(2, 1536) \times W_0^D = \begin{matrix} 1, 512 \\ 1, 512 \end{matrix} \begin{matrix} CC_1 \\ CC_2 \end{matrix}$$

(1536, 512)

$$\text{Layernorm}\left(\begin{matrix} CC_1 + W_{E,PE_{(5013)}} \\ CC_2 + W_{E,PE_{(14)}} \end{matrix}\right) = \begin{matrix} (2,512) \\ CO_1 \\ CO_2 \end{matrix}$$

In multi head decoder, like multi head encoder, attention blocks are Separate, but feed forward layers are same.

In decoder, for each head, the self attention block and encoder decoder attention blocks are separate, feed forward layers are same



EDC_1^1 EDC_1^2 EDC_1^3 = (1, 1536) 512 x3

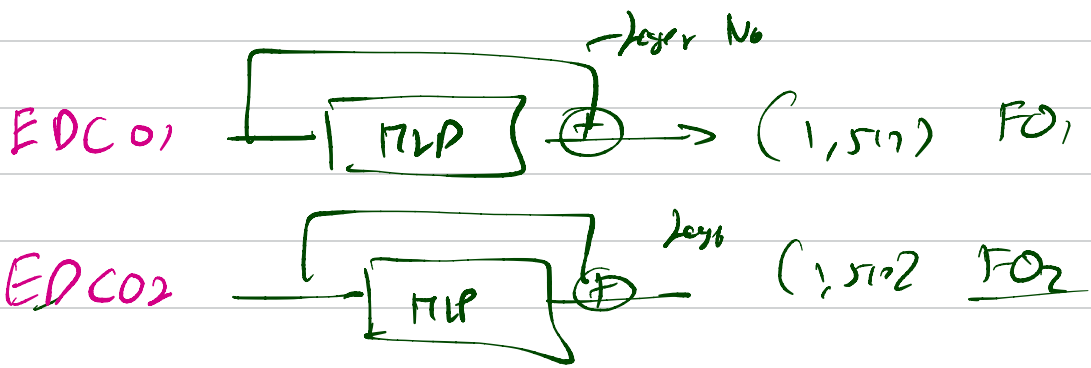
EDC_2^1 EDC_2^2 EDC_2^3 = (1, 1536)

(2, 1536) x W_0^{ED} (1536, 512) = 1, 512 → EDCC1

1, 512 → EDCC2

Layer Norm (EDCC1 + CO1 → EDCO1

EDCC2 + CO2 → EDCO2)



Untill you predict <EOS>

LSTM + Attention → Scale it up

PE

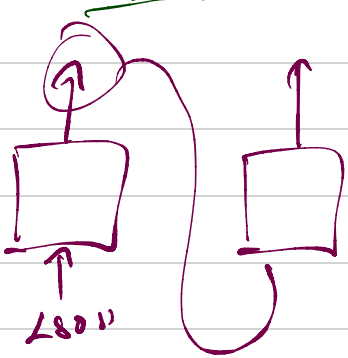
How training is done

Input

Ich habe in Miami ein Auto gekauft, war aber mit dem Preis nicht zufrieden

I bought a car in Miami but wasn't happy with the price

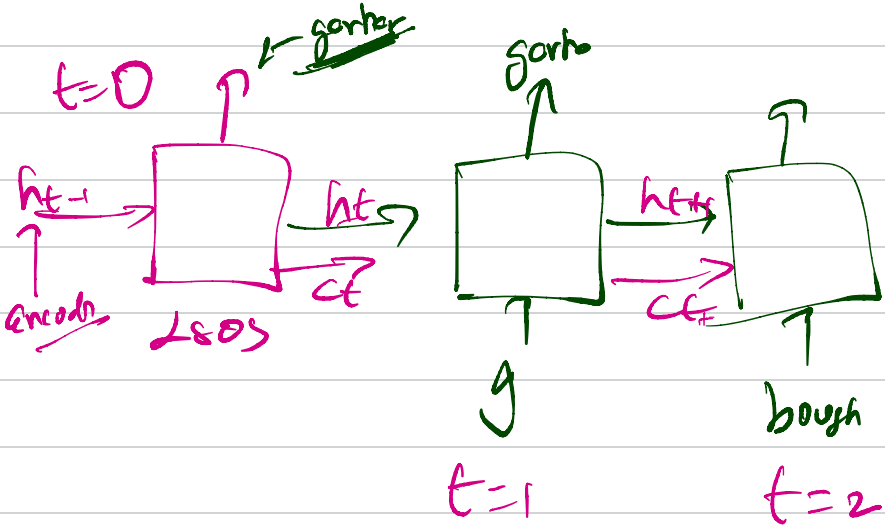
Target



Optional

ignore

Teacher Forcing → LSTM



TRANSFORMER

Input

Ich habe in Miami ein Auto gekauft, war aber mit dem Preis nicht zufrieden

I bought a car in Miami but wasn't happy with the price

Target

optional

Input → 14 word_ dim = 512

14, 512

Mask self attn

Enc-outputs

(4, 512)

Scaling of my decoder