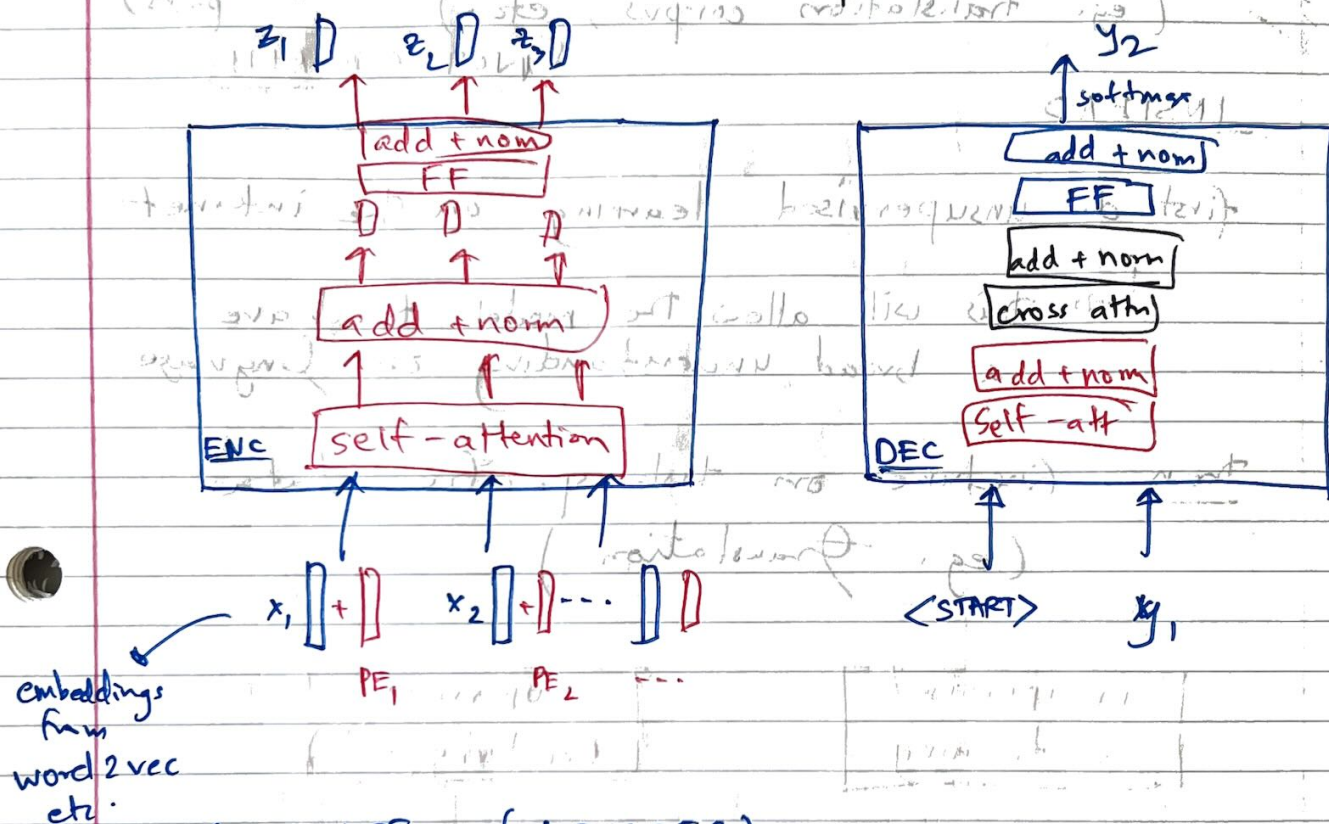


* BERT (Bidirectional Encoder Representations from Transformers Devlin et al. NAACL 2019)

recall

standard transformer model from previous classes



SIDENOTE (DECODER)

* in practice : instead of top-1 prediction of softmax

↳ consider top-k tokens and sample from that probability
eg. $\begin{bmatrix} \dots \end{bmatrix}$ leads to variance of outputs.

↳ consider a prob. threshold and get output that add to it
sample from that

* BERT: key-idea "TRANSFER LEARNING"

it's difficult to collect annotated data (input-output pairs)
(eg. translation corpus, etc.)

NOT SCALABLE

INSTEAD

first do unsupervised learning on the internet

↳ this will allow the model to have a broad understanding of language

then fine-tune on task-specific data
(eg. translation)

unsupervised
pretraining

supervised
fine-tuning (tasks)

* BERT: ENCODER

uses the encoder architecture from Transformers

BERT vs GPT :

↳ uses Transformer
Encoder

↳ uses Transformer
Decoder

* BERT pre-training

- ① Masked Language Modeling (MLM)
- ② Next Sentence Prediction (NSP)

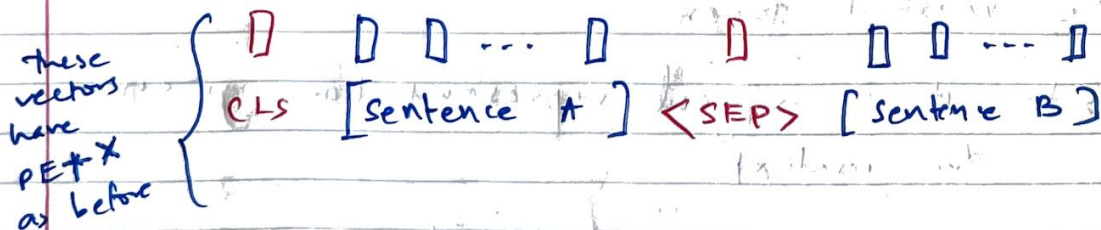
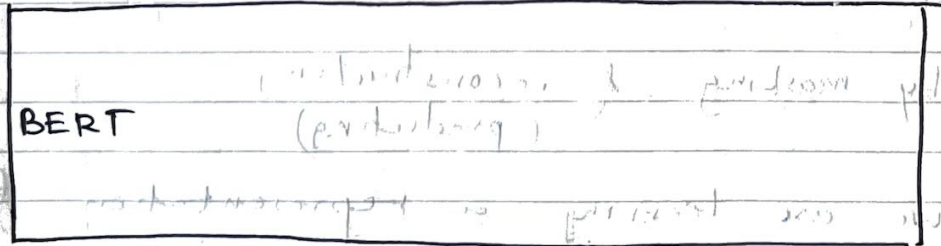
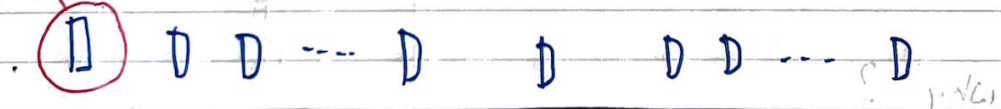
both objectives are used simultaneously to train BERT

(Note: later on, people realized that NSP is not that important)

BERT input: two sentences with a separator token
(is A before B?) (and [CLS] token)

#2

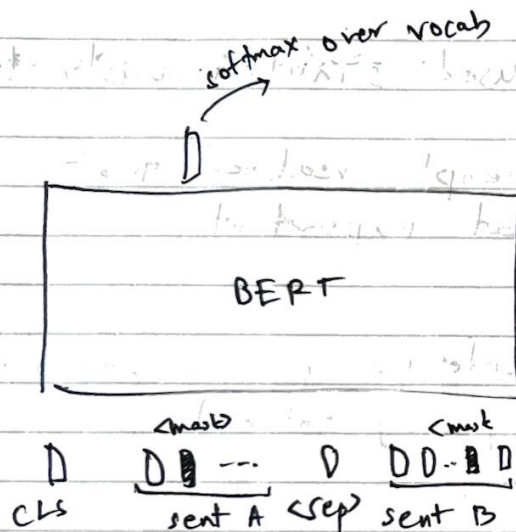
Binary Classifier



The final [CLS] embedding then contains "info that represents the "entire sentence"

#1 MLM

- mask some % of input tokens (at random)
- try to predict those tokens in the output



similar to
SSL idea
in vision
(patch reconstruction)

why?

- by masking & reconstructing (predicting)

we are learning a representation for the masked token

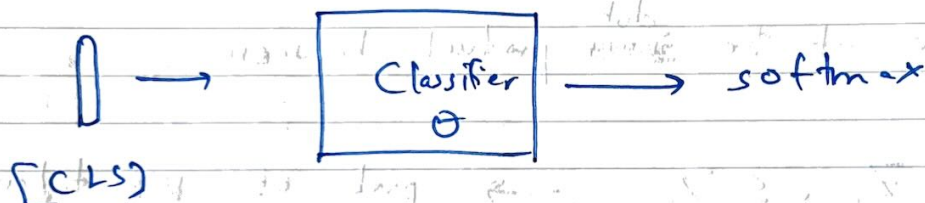
s.t. it is a good representation given the context

* Finetuning Step

- task specific

* Classification Problem (eg. sentiment analysis)

- take CLS token,



* Question Answering

(very important NLP task)

one format

Input : question Q : Will I get an A+ in CMSC 675 ?

Output : answer A : No

another format

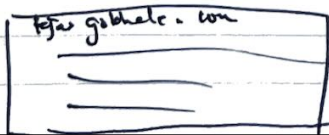
"Reading Comprehension"

Question + Passage (Document)
eg. Wiki

↓
Answer

Q : where did Tejas receive his PhD ?

P:



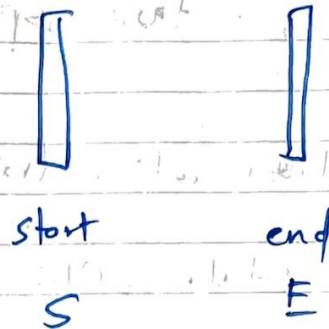
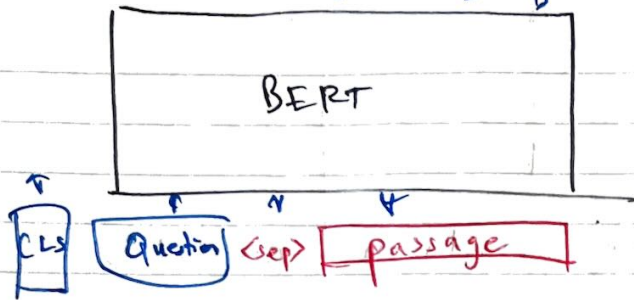
"span" in passage

A : $\langle \text{start idx} \rangle \quad \langle \text{end idx} \rangle$

answer is here

D DDD DDDDD

introduce two vectors



look at the ~~dot~~ product between

$\langle Z_i, S \rangle \rightarrow$ prob. of ~~the~~ token i being the start of the answer

$\langle Z_i, E \rangle \rightarrow$ " end of answer

Learn S, E
Fine tune BERT