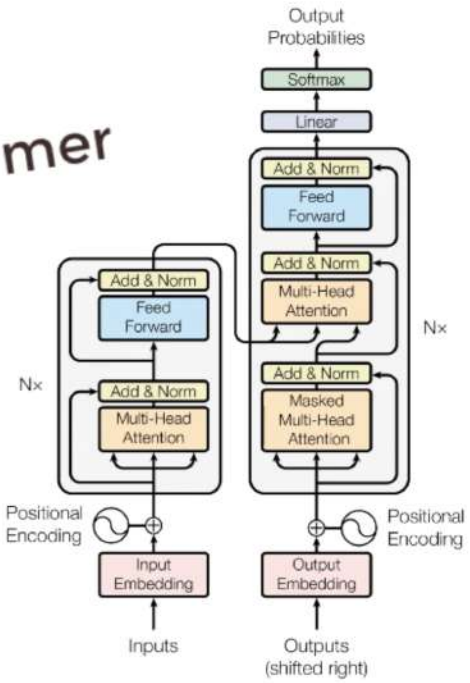
BERT

LSTM Vs Transformer

Transformer

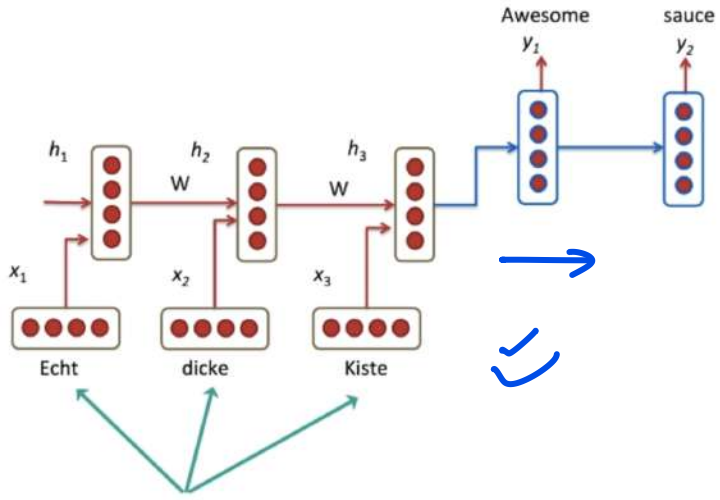


Source: Attention is All you Need (Vaswani et al., 2017)

LSTM Vs Transformer

LSTM Networks

1. Slow
2. Not truly Bidirectional



True meaning of source words not entirely captured

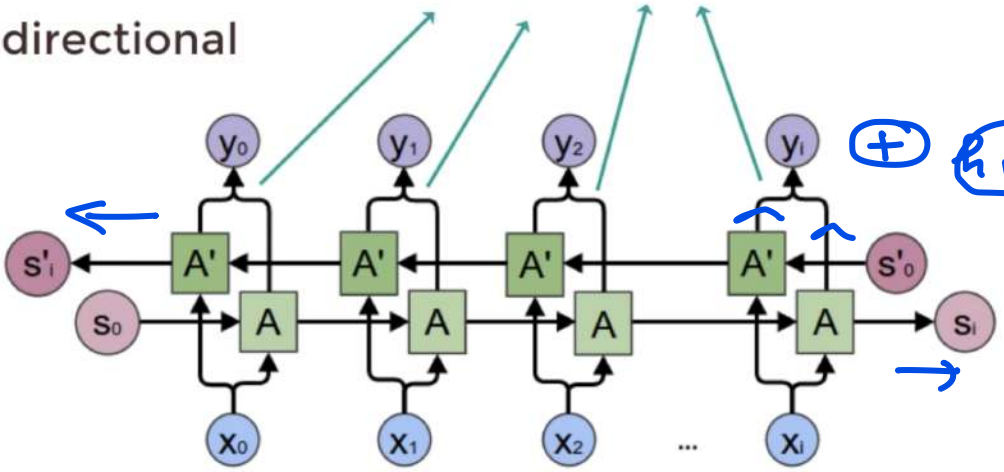
Source [Blog]: Attention & Memory in Deep Learning (Britz, 2017)

LSTM Vs Transformer

LSTM Networks

1. Slow
2. Not truly Bidirectional

Simple concatenation

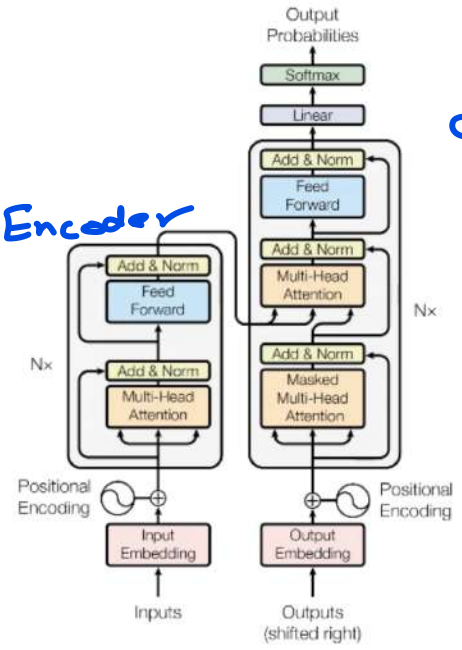


Source [Blog]: Understanding LSTMs (Colah, 2015)

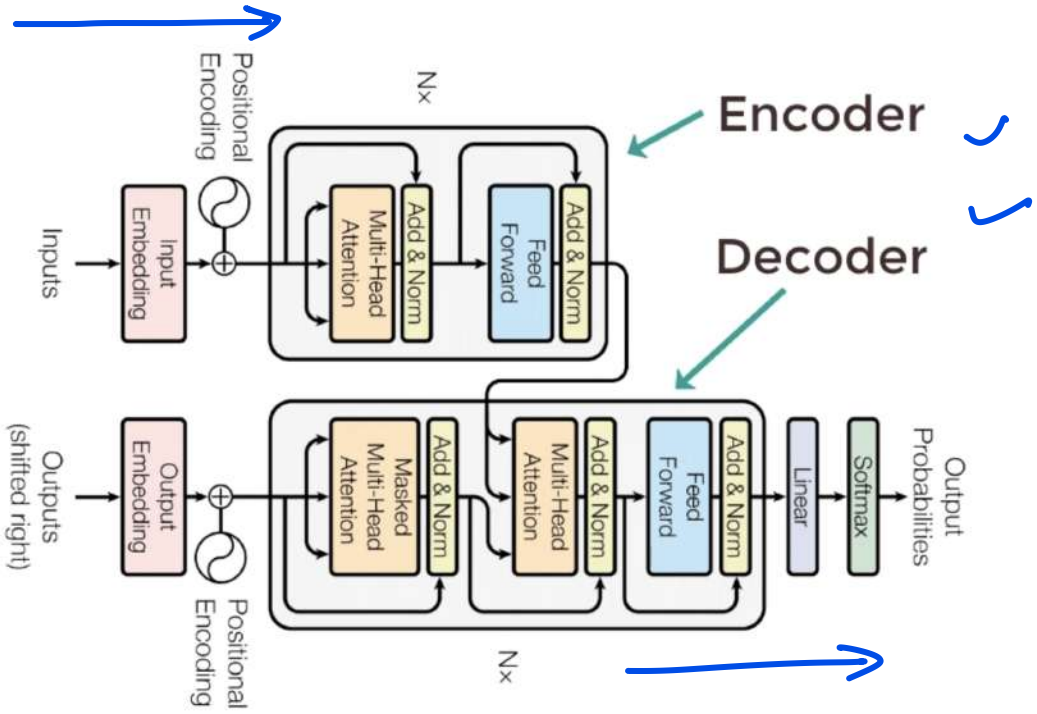
LSTM Vs Transformer

Transformer

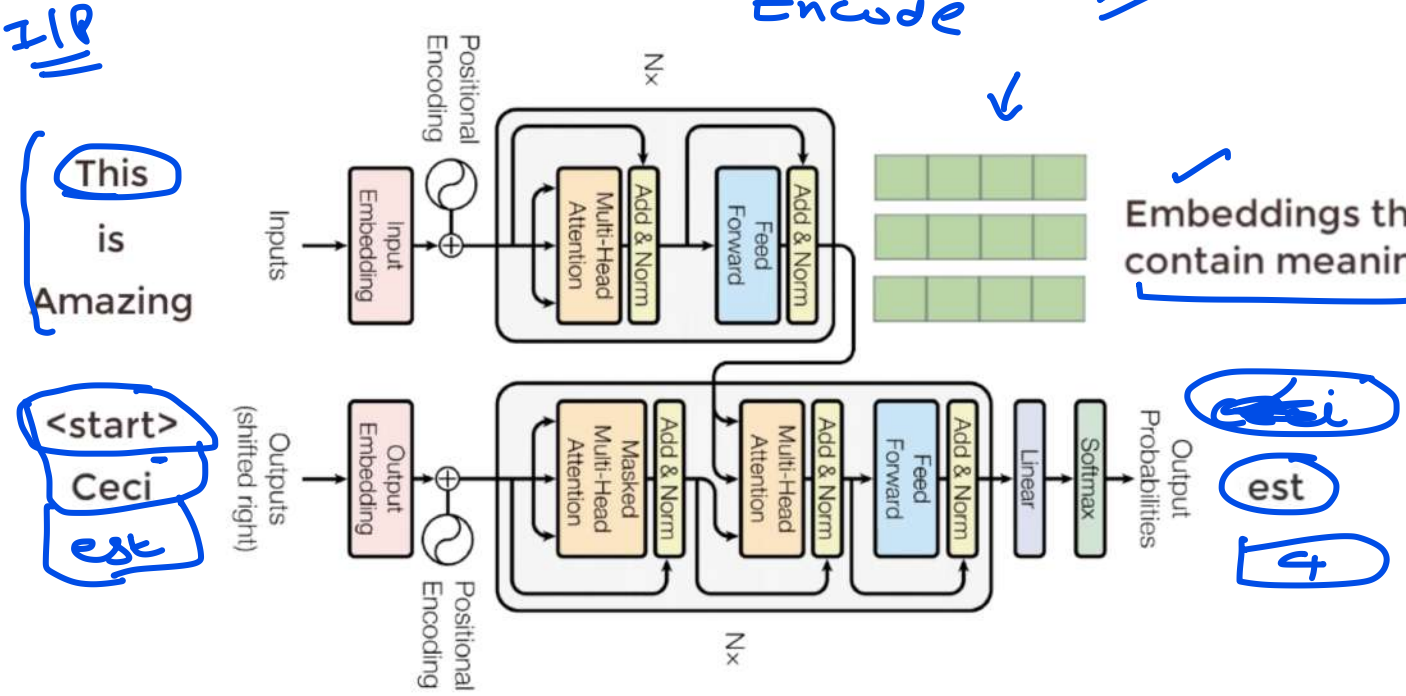
1. ~~Slow~~ Faster
2. ~~Not truly Bidirectional~~
   Deeply Bidirectional



Transformer Flow



Transformer Flow



Encode

This is Amazing

<start> Ceci est

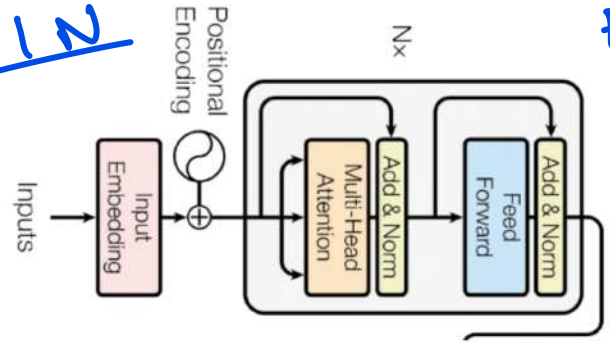
Embeddings that contain meaning

Ceci est 4

Transformer Flow

Eng → French

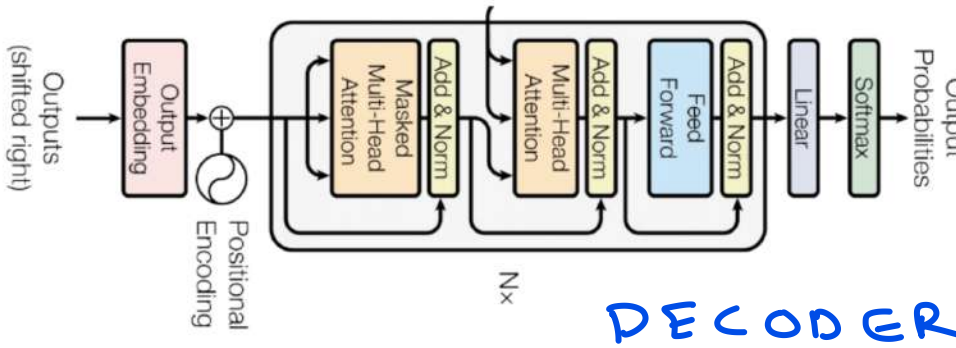
ENCODER training



NLU

What is English? What is context?

What is language!

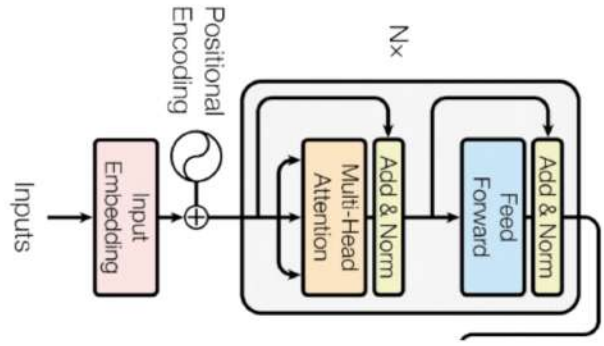


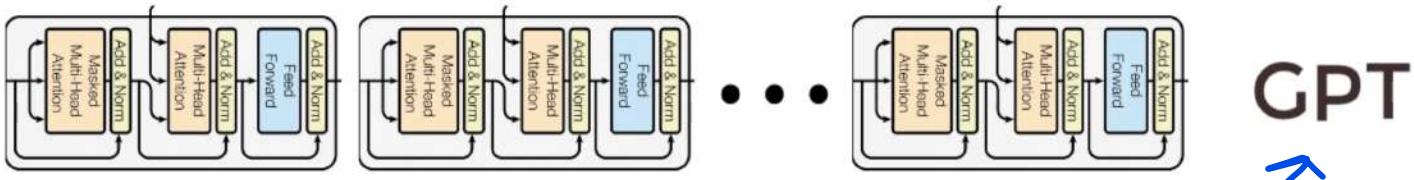
How to map English words to French words?

What is language!

DECODER

Transformer Flow



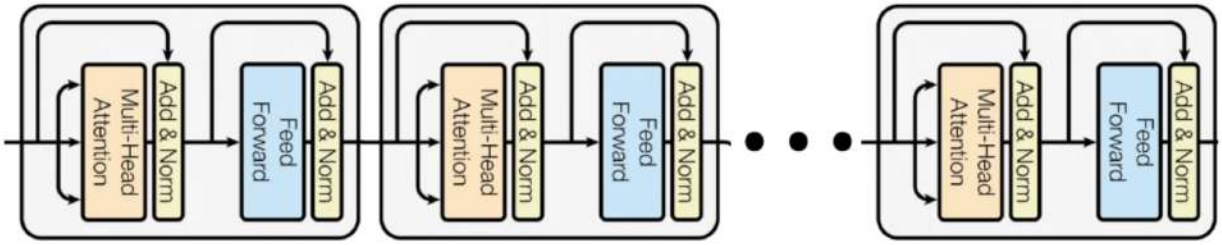


GPT

Decoder + Decoder + ... + Decode

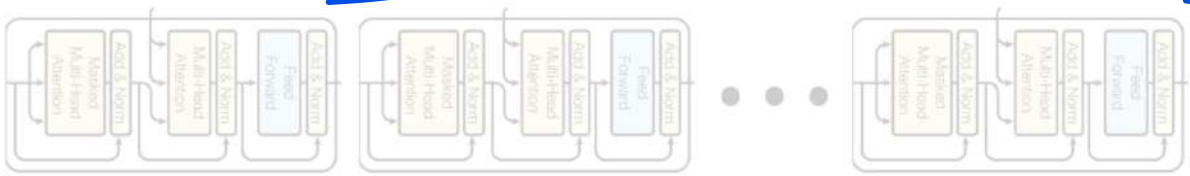
Transformer Flow

Encoder + Encoder + .. + Encoder



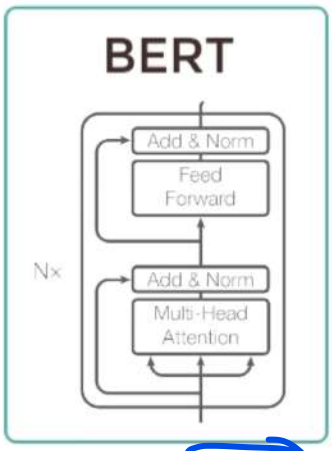
BERT

Bidirectional **E**ncoder **R**epresentation from **T**ransformers



GPT

Bidirectional Encoder Representation from Transformers



base + fine

Problems to Solve

eng-french

- Neural Machine Translation
- Question Answering
- Sentiment Analysis
- Text summarization

Encoder

Needs Language understanding

How to solve Problems (BERT Training)

- Pretrain BERT to understand langauge
- Fine tune BERT to learn specific task

Bidirectional Encoder Representation from Transformers

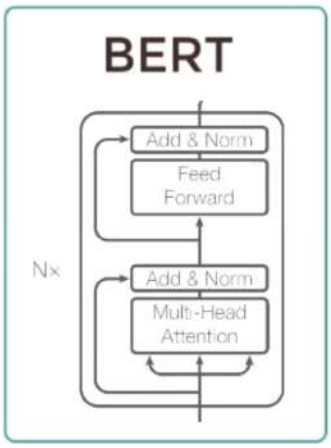
Pretraining (Pass 1) : "What is language? What is context?"

Masked Language Model (MLM)

The [MASK1] brown fox [MASK2] over the lazy dog.

Next Sentence Prediction (NSP)

A: Ajay is a cool dude.
B: He lives in Ohio

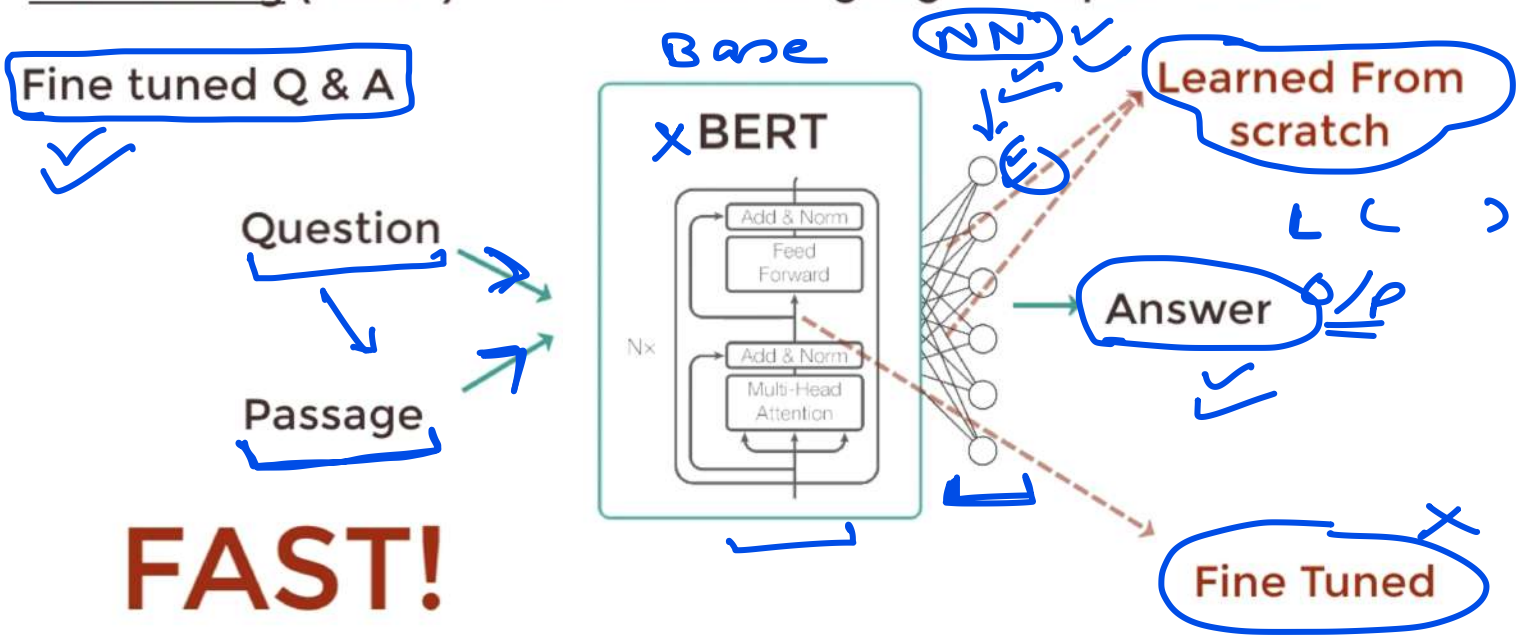


[MASK1] = quick
[MASK2]= jumped

Yes. Sentence B follows sentence A

Bidirectional Encoder Representation from Transformers

Fine Tuning (Pass 1): "How to use language for specific task?"

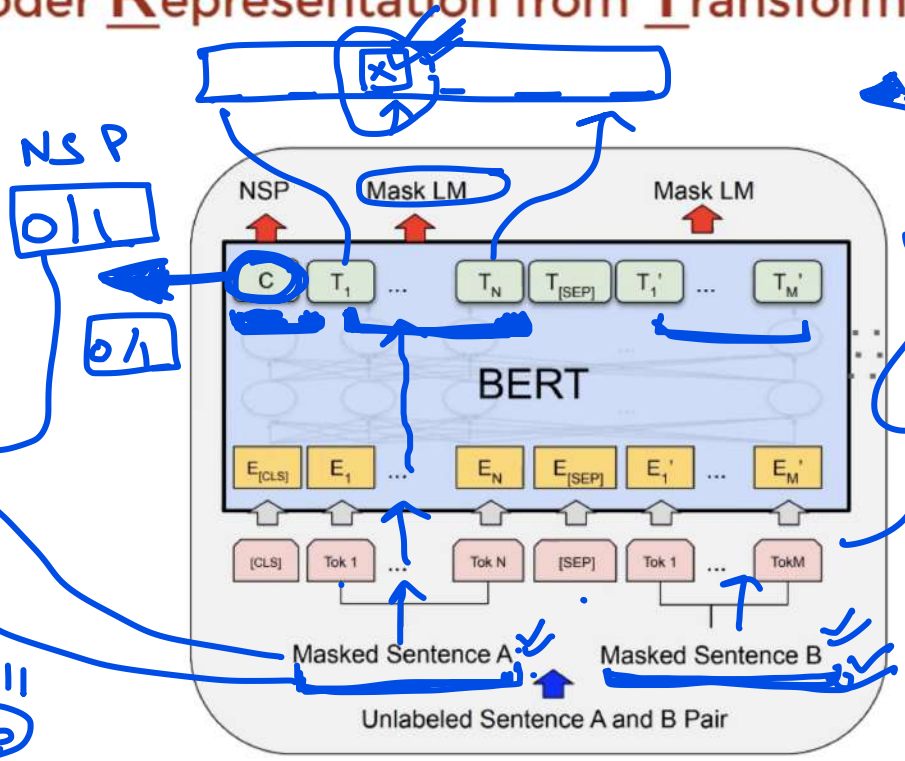


Bidirectional Encoder Representation from Transformers

Pretraining (Pass 2)

Problems to train on simultaneously:

1. Masked Language Modeling (Mask LM)
2. Next Sentence Prediction (NSP)



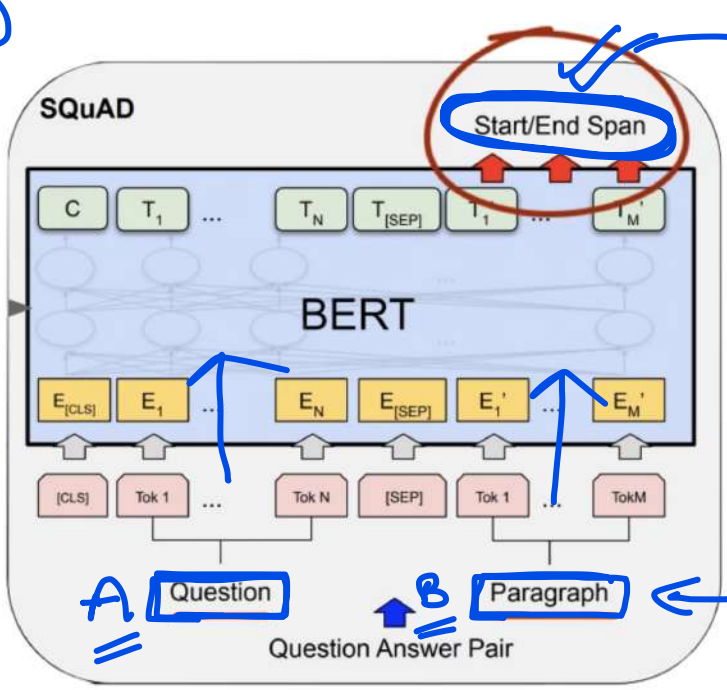
Source: BERT: Pre-training of deep bidirectional Transformers for language understanding (Devlin et al., 2019)

Bidirectional Encoder Representation from Transformers

Fine Tuning (Pass 2)

Change output to display text in which answer exists

Change inputs to take in Question, Passage



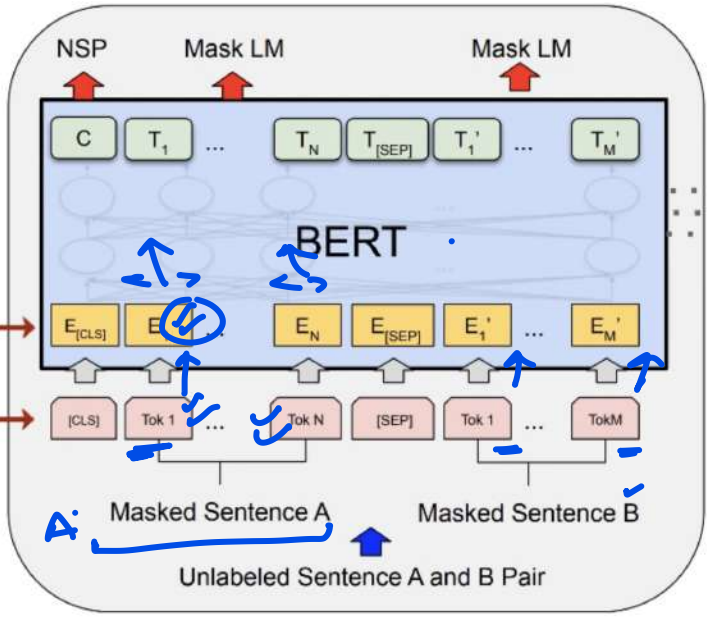
Bidirectional Encoder Representation from Transformers

Pretraining (Pass 3)

$E_1 = T_1 + S_1 + P_1$

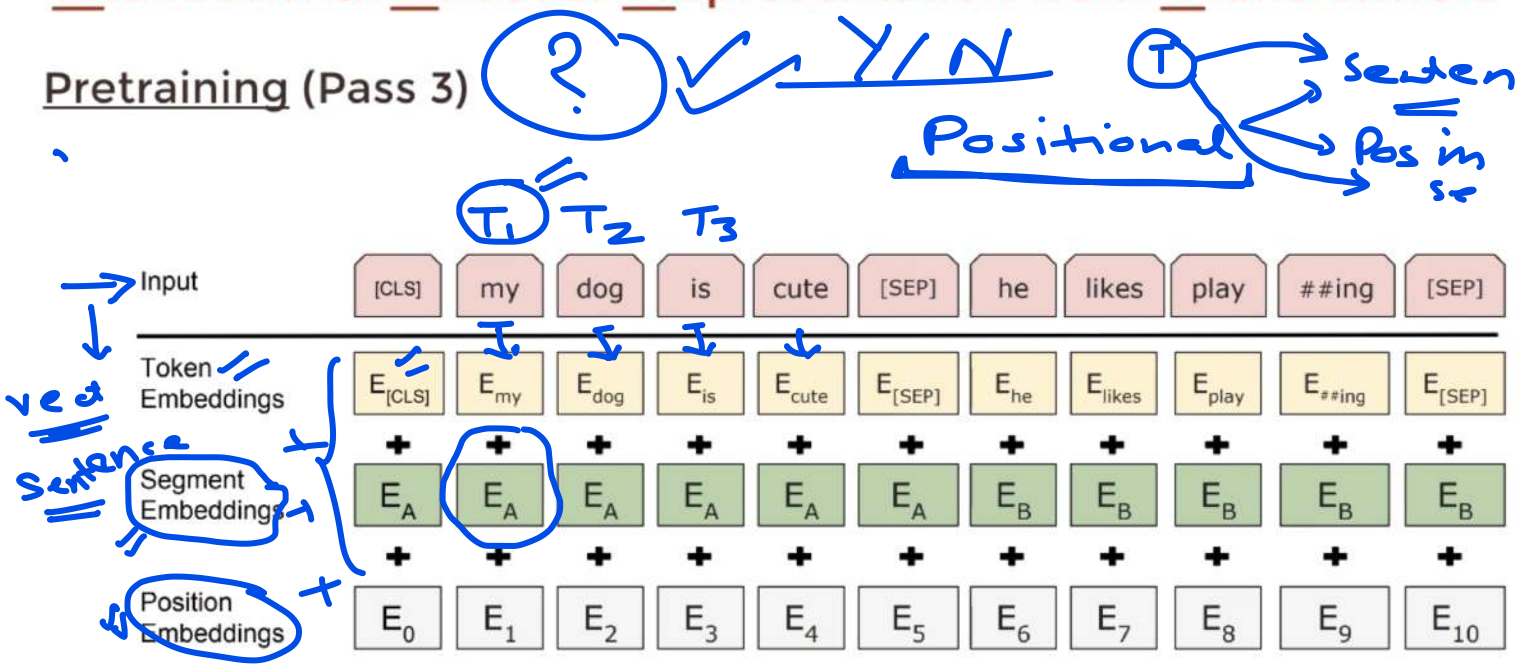
initial embeddings

word token inputs



Bidirectional Encoder Representation from Transformers

Pretraining (Pass 3)



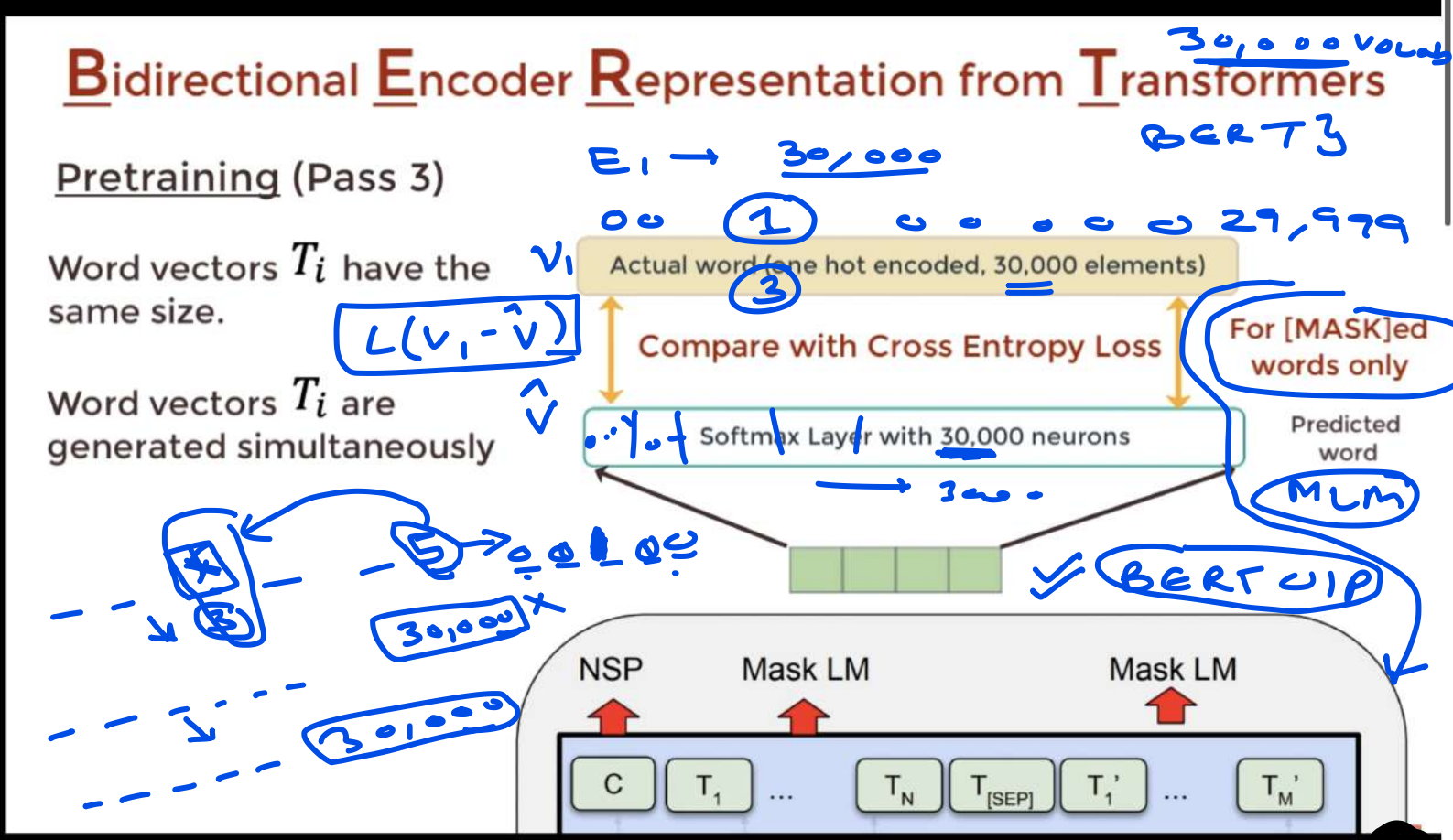
Segment + Position embeddings add "ordering" for inputs

Bidirectional Encoder Representation from Transformers

Pretraining (Pass 3)

Word vectors T_i have the same size.

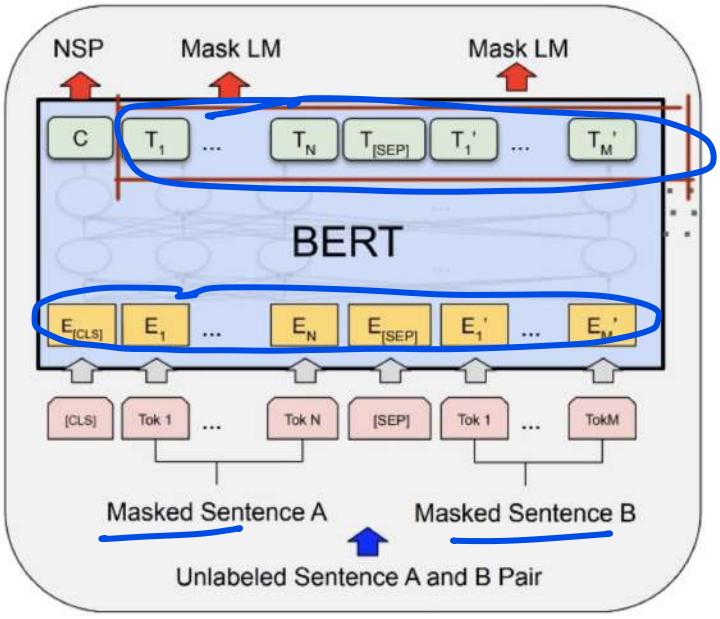
Word vectors T_i are generated simultaneously

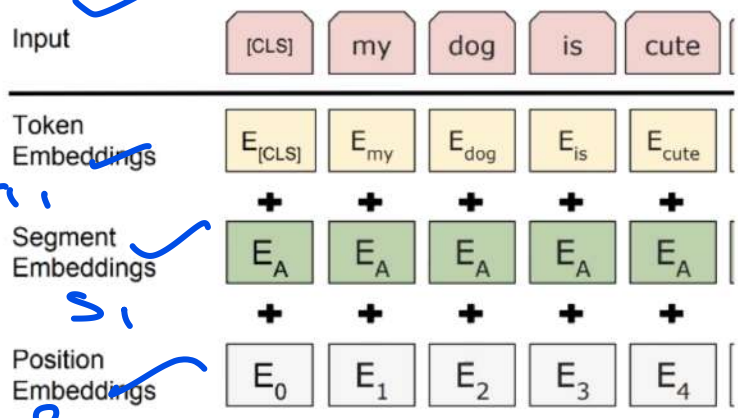


Bidirectional Encoder Representation from Transformers

BERT → Pretrain → NLU

Pretraining (Summary)

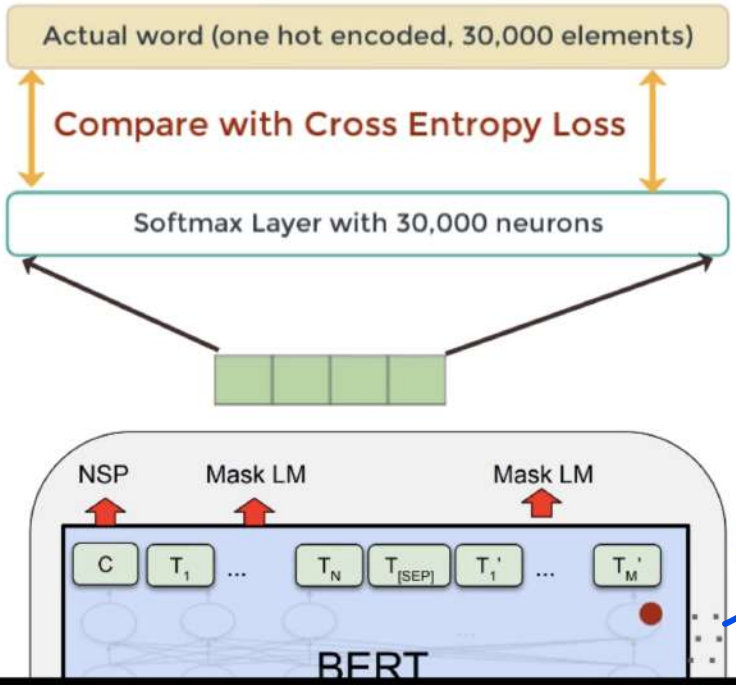


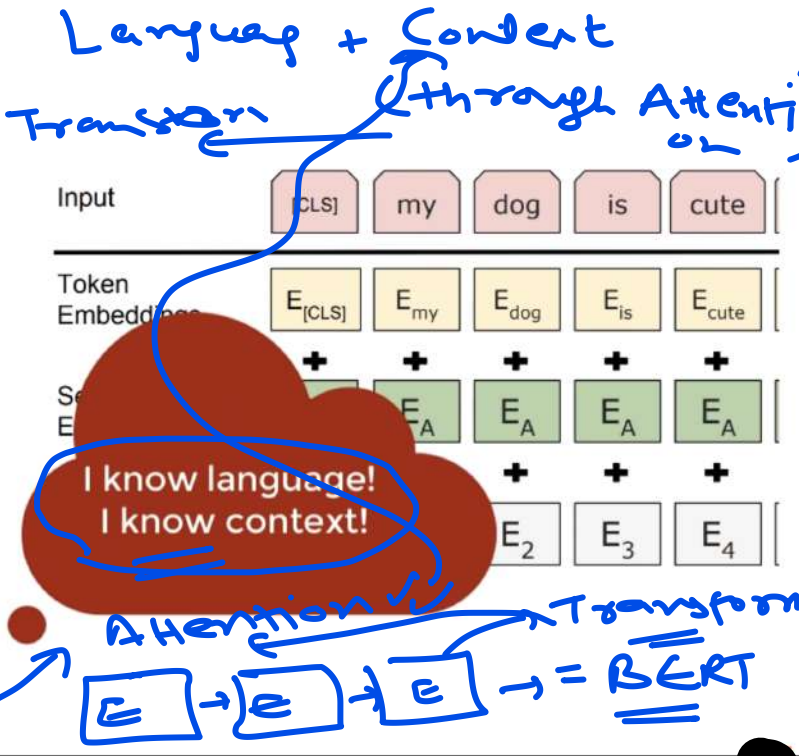


$E_1 = T_1 + S_1 + P_1$

Bidirectional Encoder Representation from Transformers

Pretraining (Summary)

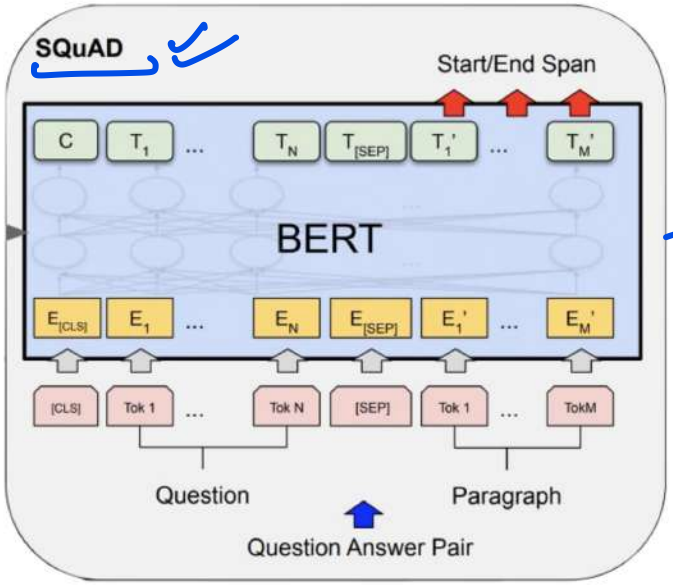




Bidirectional Encoder Representation from Transformers

Fine Tuning (Summary)

Q/A = BERT 30min



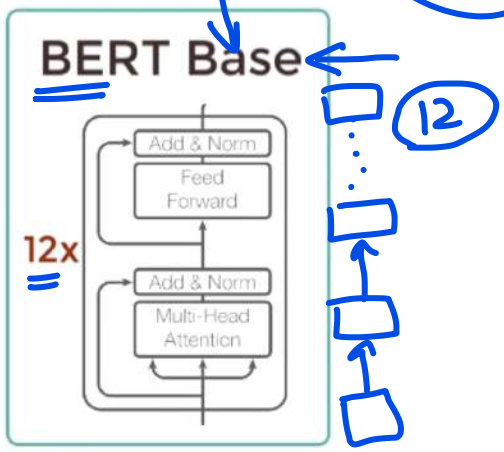
"Stanford Question & Answer Dataset"

- 30 minute training on single cloud TPU with 91% F1 score.

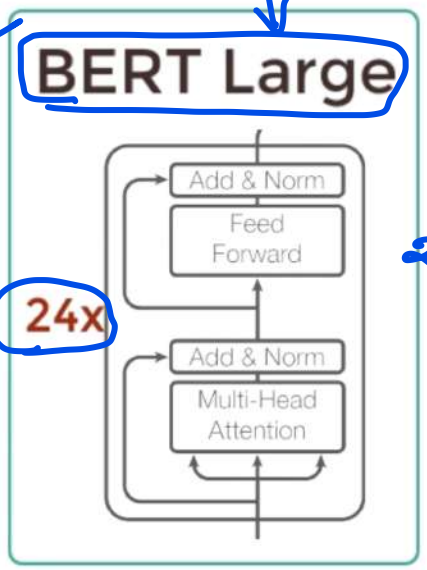
Tensor Processing Unit

Bidirectional Encoder Representation from Transformers

Performance } Score



110M Parameters



24 Encoder layers

Much more capable than Base

340M Parameters