Preparing for LLMs

Video 2: Pre trained transformers : BERT

Analytics Vidhya

5. BERT

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

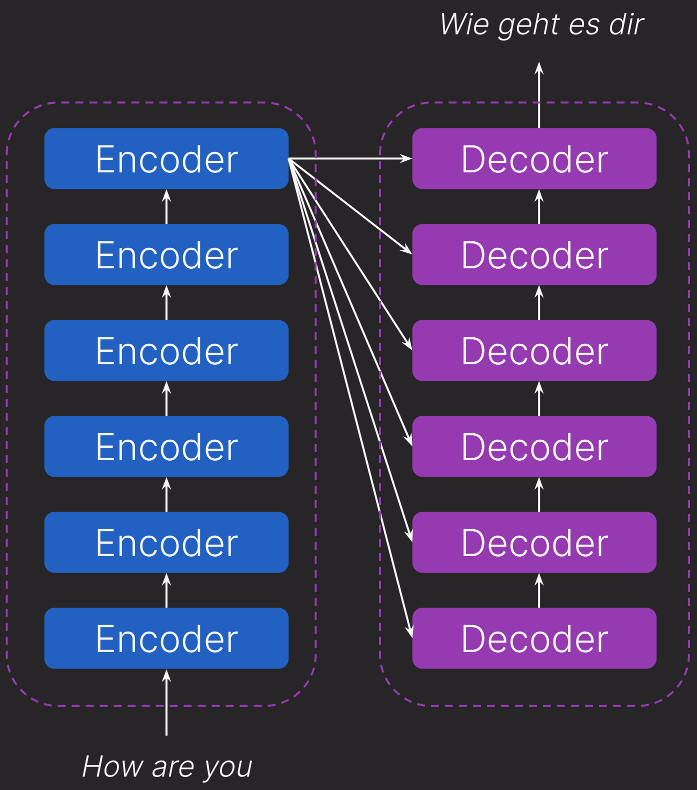
This method considers both left and right context of words in a sentence.

Ice melts __________ **it gets warm.**

Left

Right

Transformer

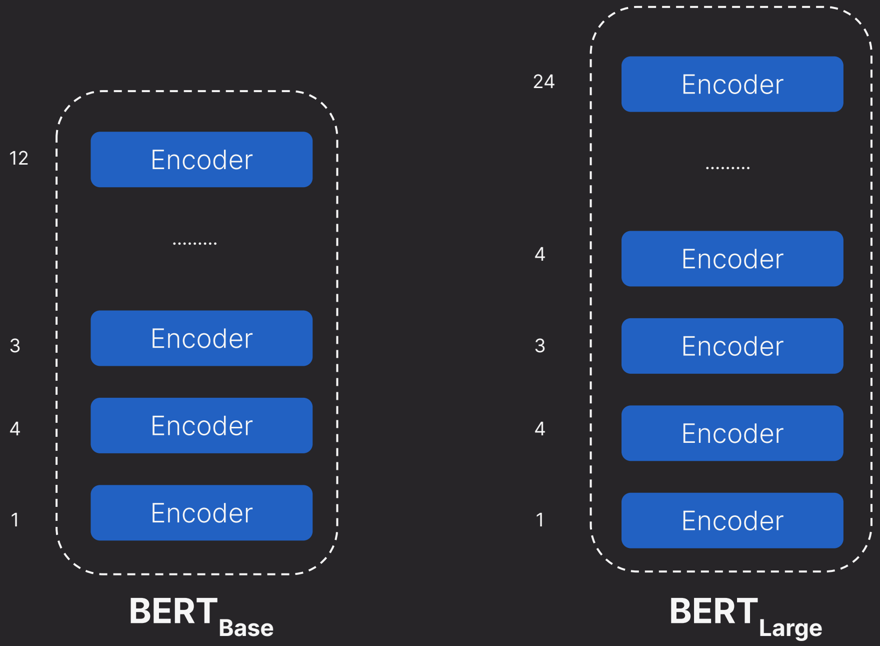


Analytics Vidhya

Working of BERT

12 Encoders units

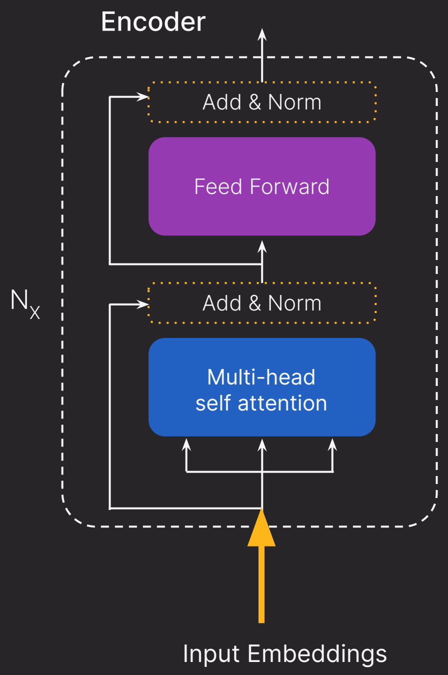
110 millions parameters



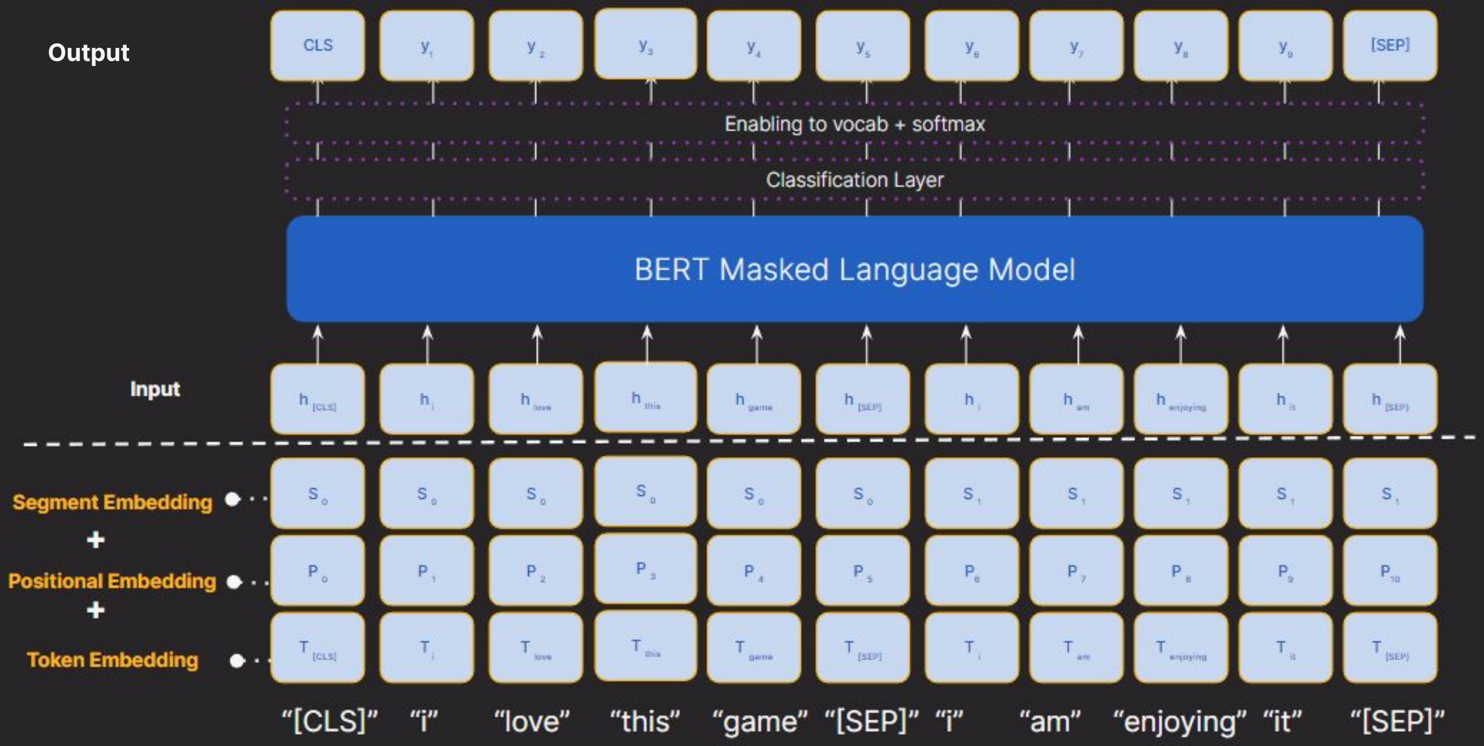
24 Encoders units

340 millions parameters

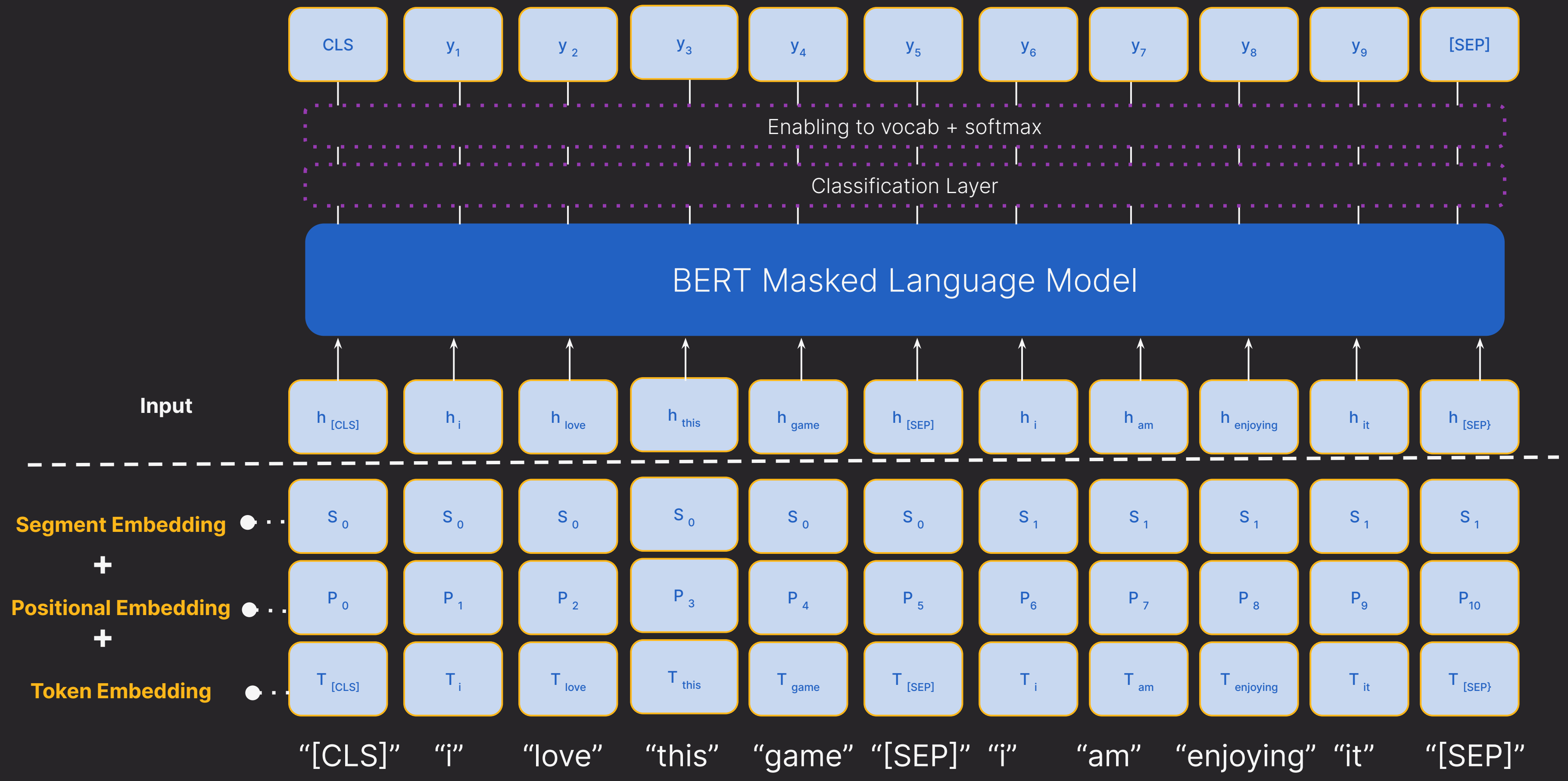
Working of BERT



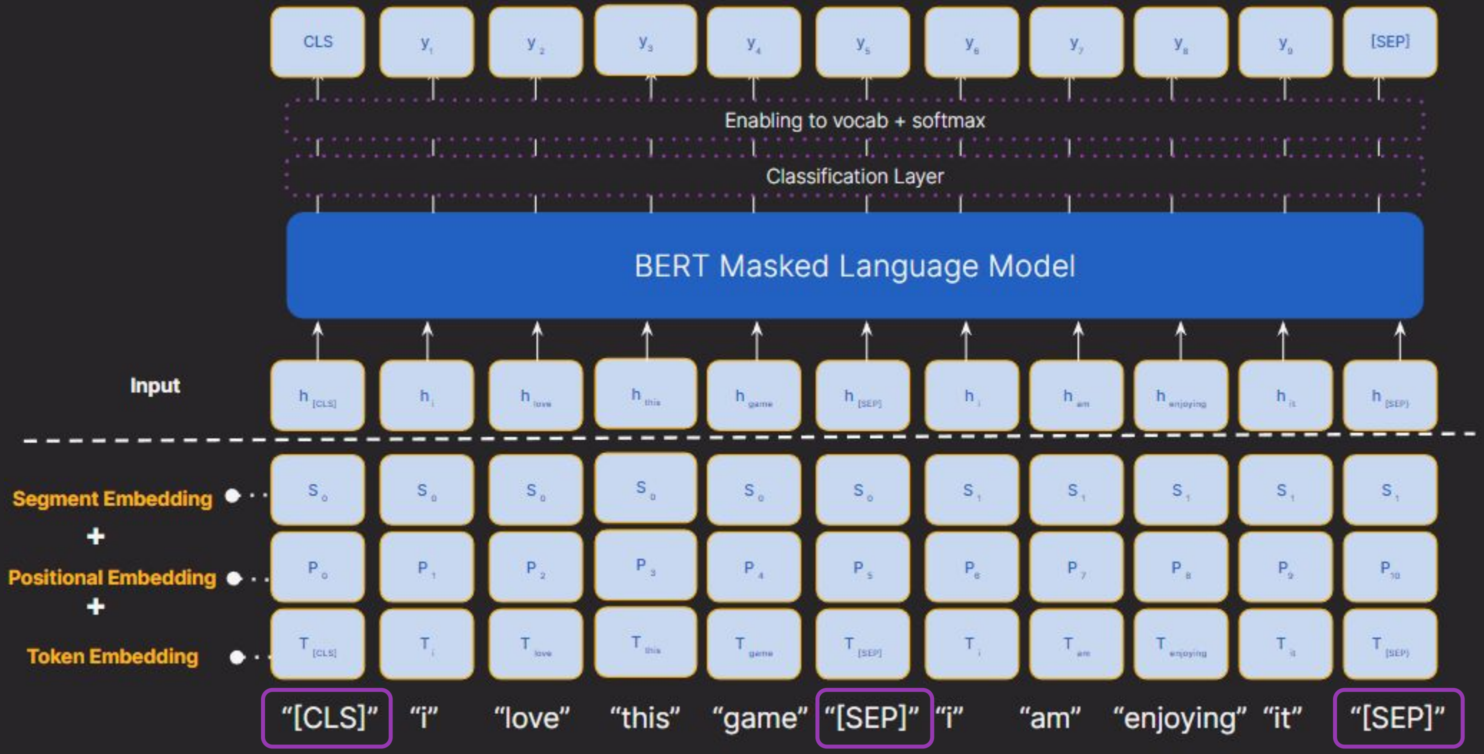
Working of BERT



Analytics Vidhya

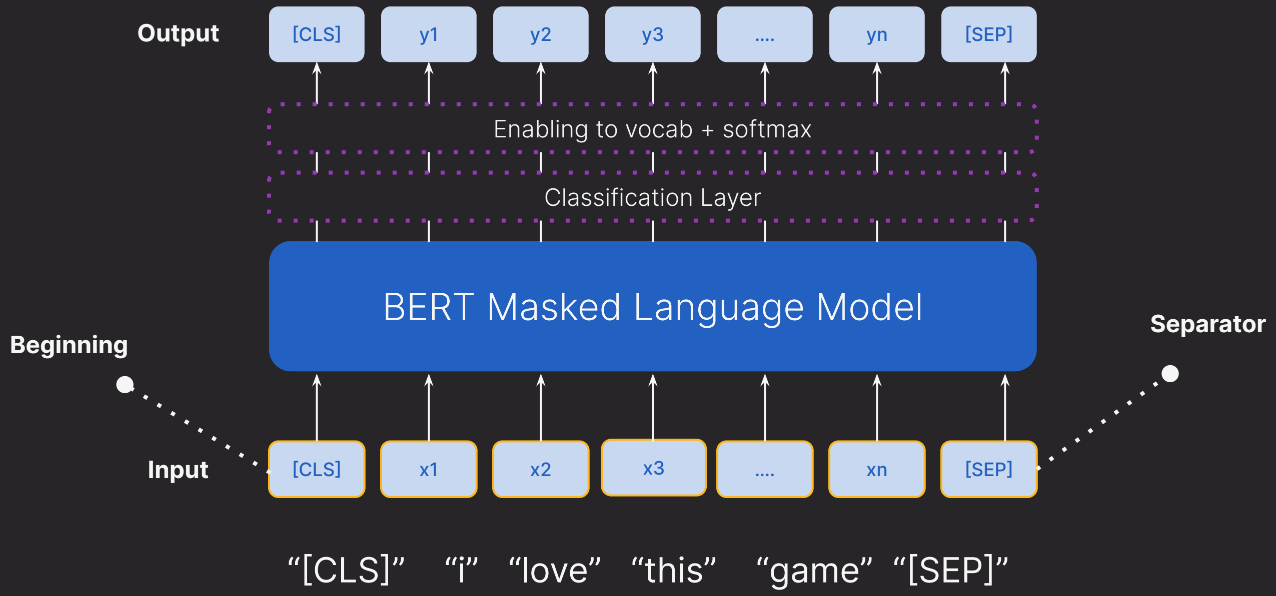


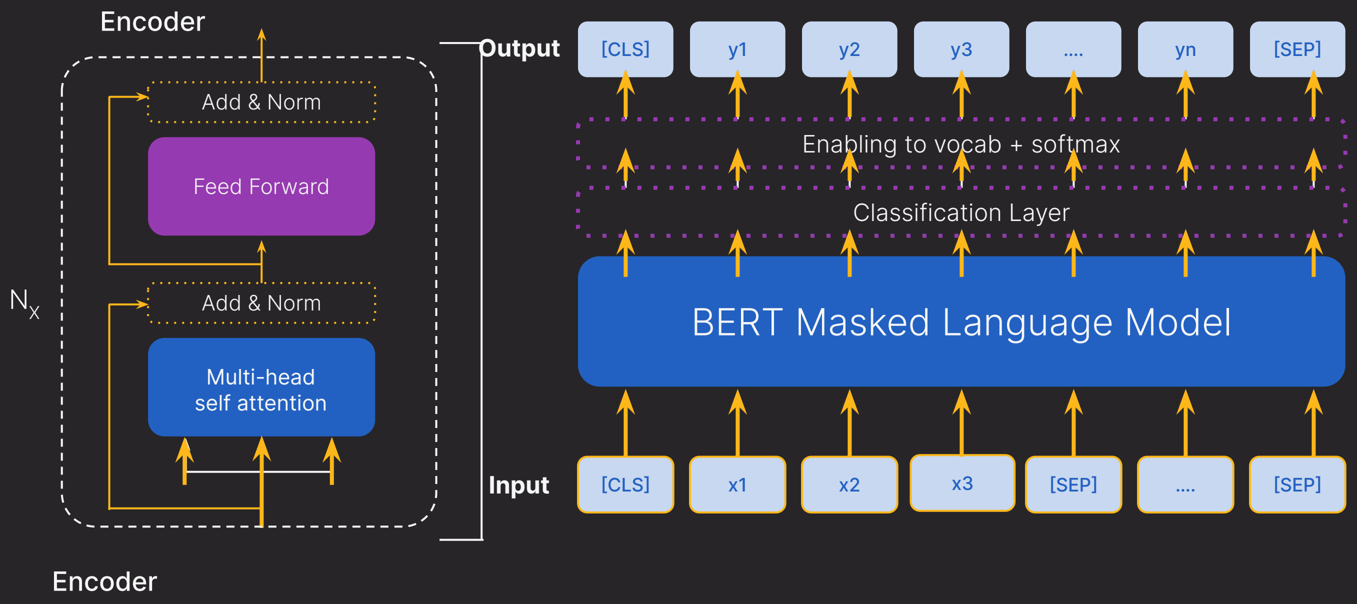
Working of BERT

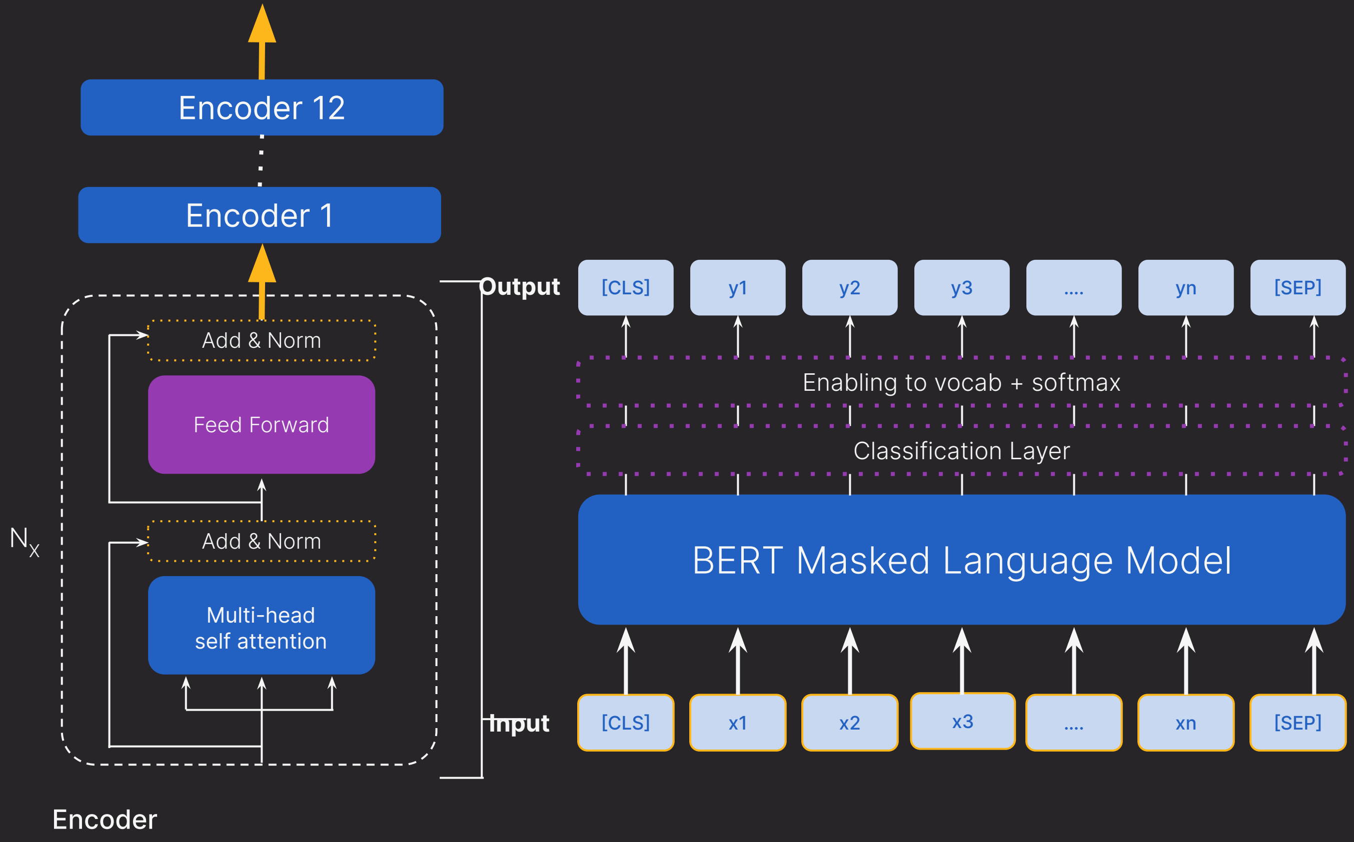


Working of BERT

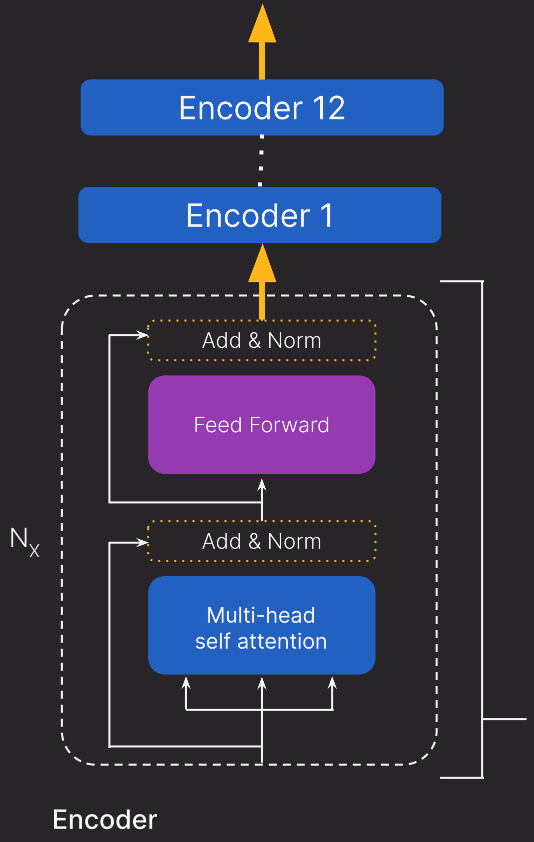
- Pre-training is done on a large amount of text to learn the contextual meaning.




Encoder
Add & Norm
Feed Forward
Add & Norm
Multi-head
self attention
N_X
Encoder
Output
Input
[CLS]
y1
y2
y3
....
yn
[SEP]
Enabling to vocab + softmax
Classification Layer
BERT Masked Language Model
[CLS]
x1
x2
x3
[SEP]
....
[SEP]

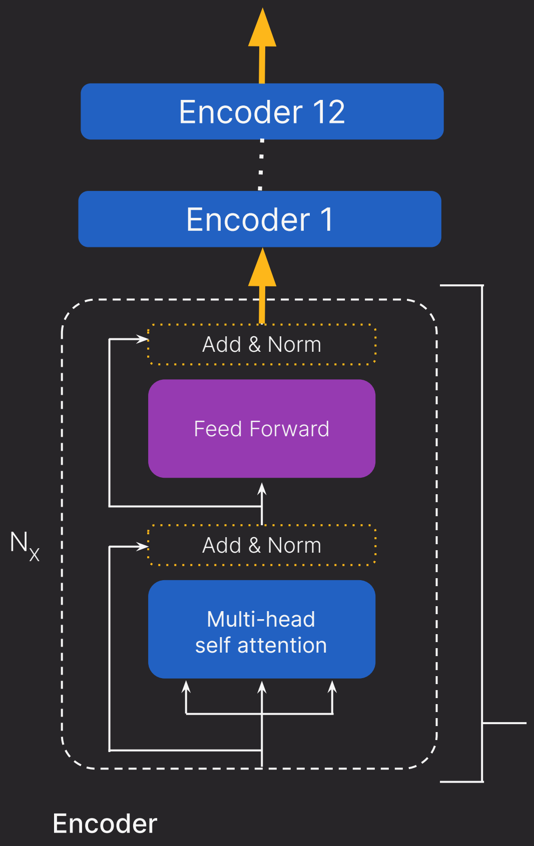

Encoder 12
Encoder 1
Add & Norm
Feed Forward
Add & Norm
Multi-head self attention
N_x
Encoder
Output
Input
[CLS]
y1
y2
y3
....
yn
[SEP]
Enabling to vocab + softmax
Classification Layer
BERT Masked Language Model
[CLS]
x1
x2
x3
....
xn
[SEP]

BERT



1 Pre - Training

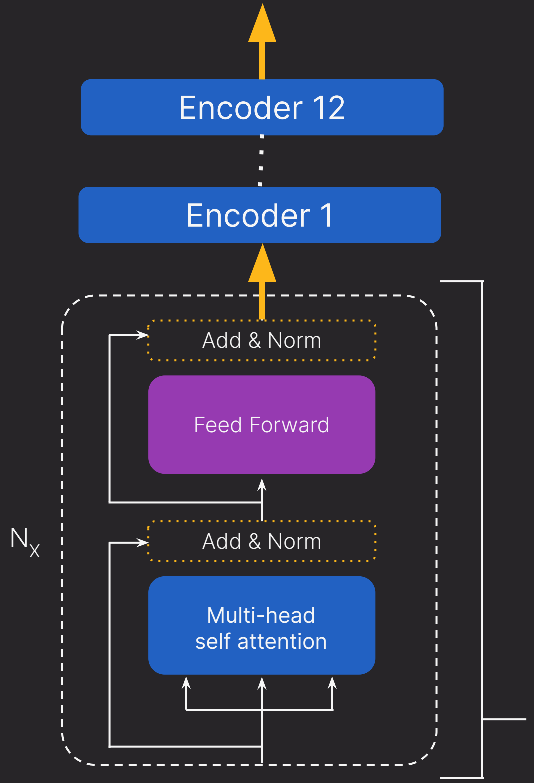
2 Fine-Tuning



BERT

1 Pre - Training

- Wikipedia of **2500 million** words
- Book corpus data of **800 million** words



Encoder

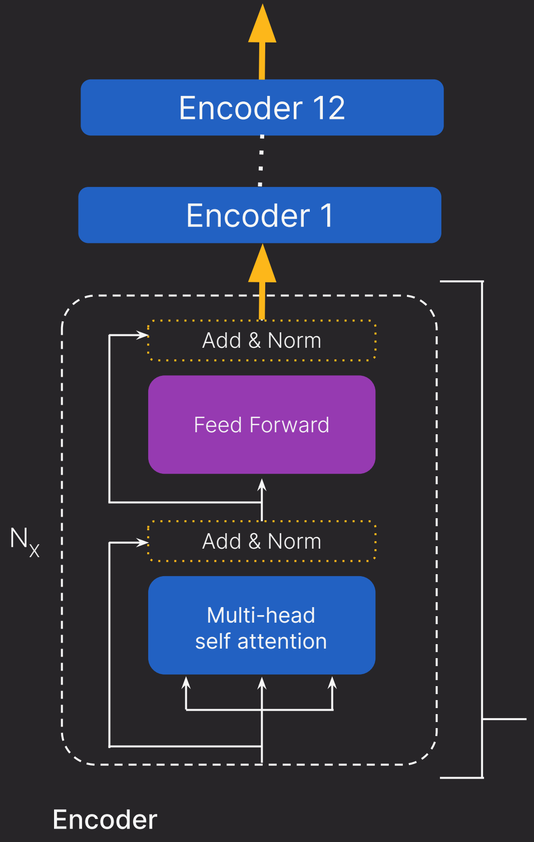
BERT

1 Pre - Training

Challenge of defining prediction goal

- **Masked Language Model (MLM)**
- **Next Sentence Prediction (NSP)**

BERT



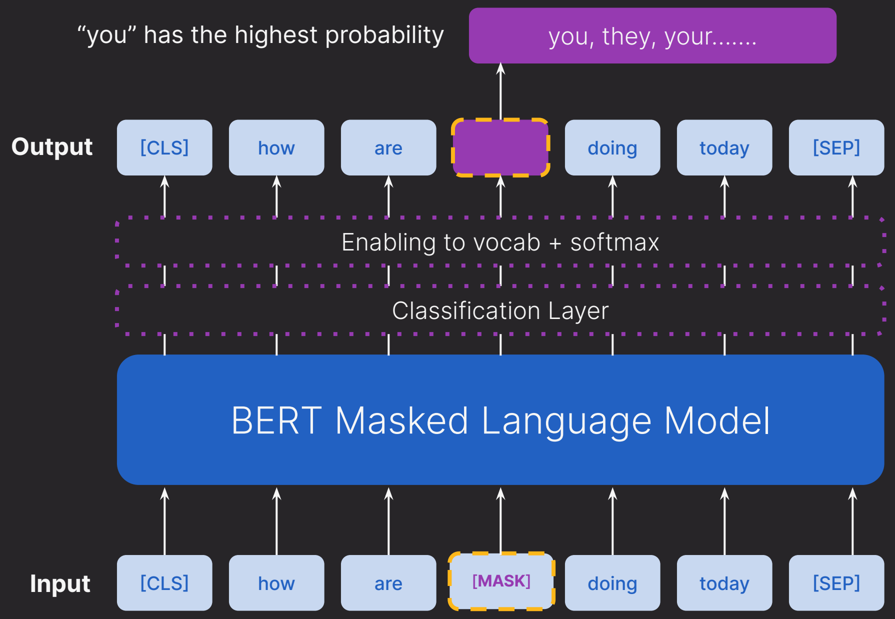
1 Pre - Training

Challenge of defining prediction goal

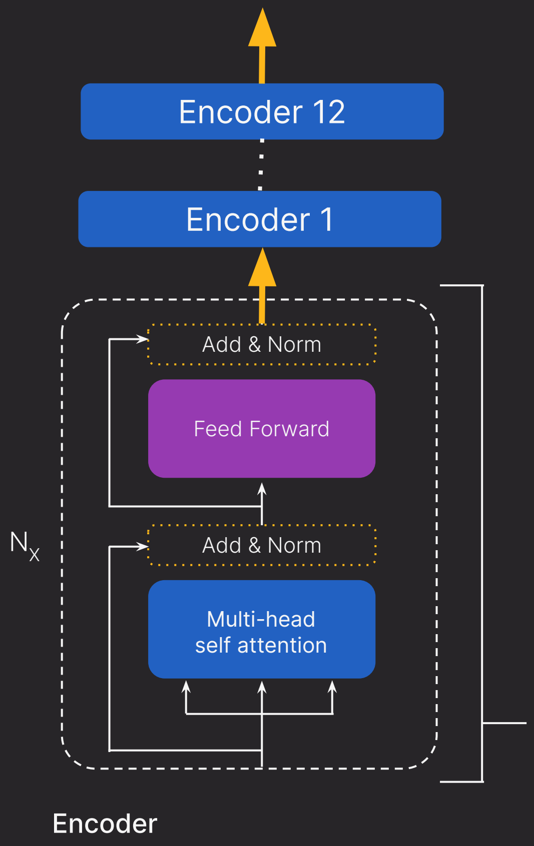
- Masked Language Model (MLM)
- Next Sentence Prediction (NSP)

Masked Language Model (MLM)

- BERT **masks and replaces some words** in input sequences with symbols like [MASK], to be predicted by the model.



BERT



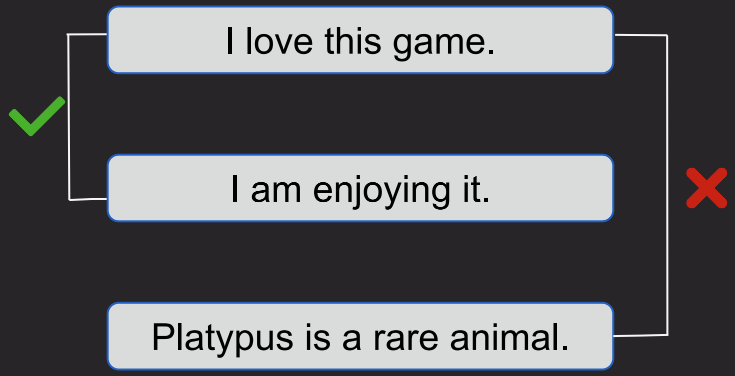
1 Pre - Training

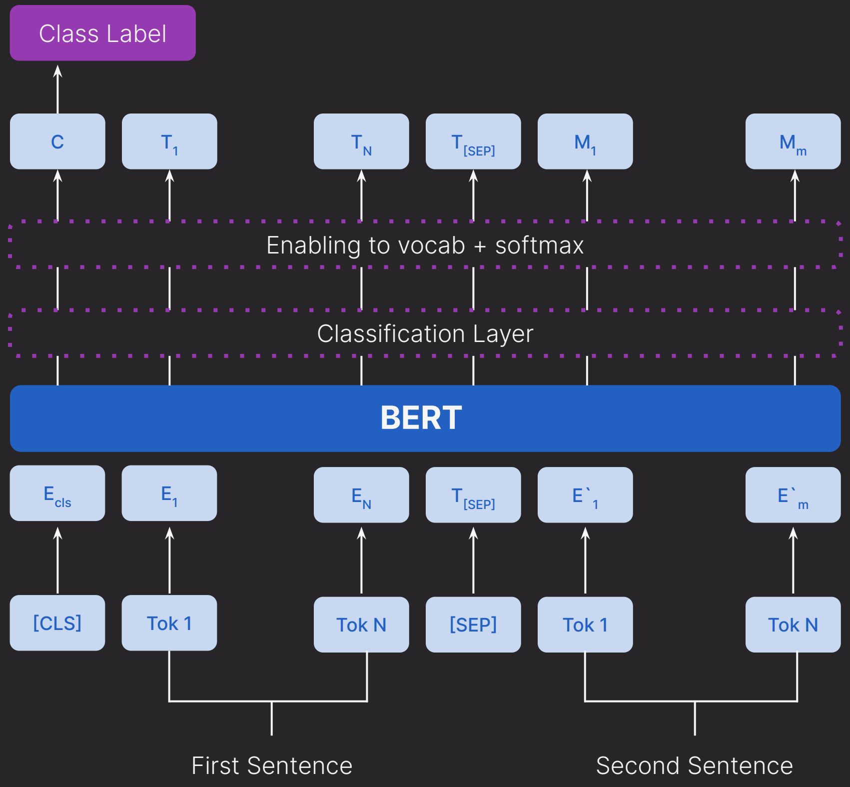
Challenge of defining prediction goal

- **Masked Language Model (MLM)**
- **Next Sentence Prediction (NSP)**

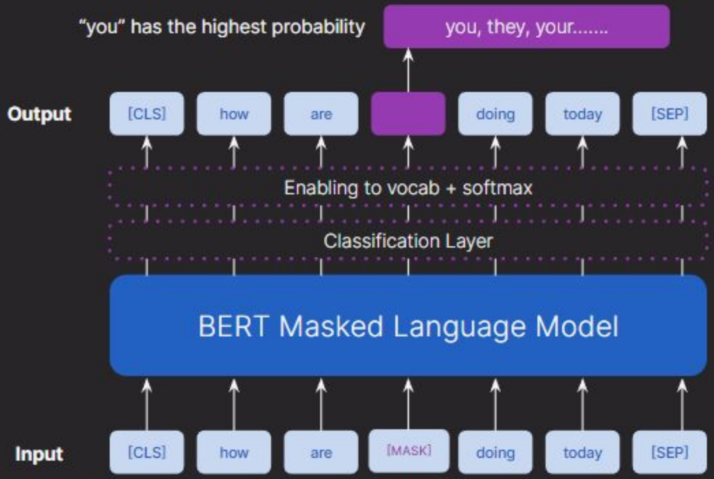
Next Sentence Prediction (NSP)

- For a pair of sentences, BERT has to identify whether one sentence logically follows the next or not.



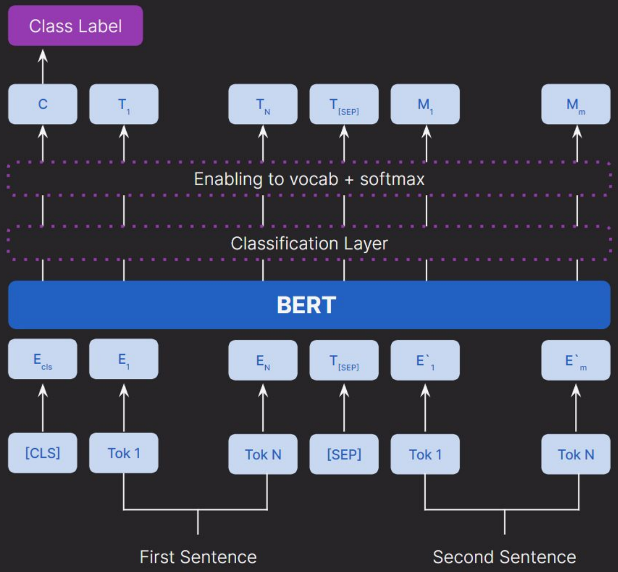


Analysis



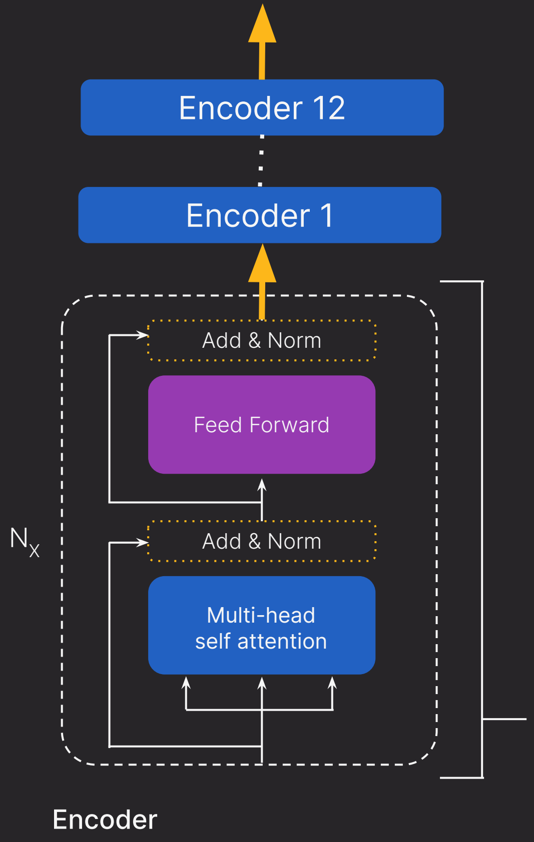
Masked Language Model (MLM)

Understanding context within a sentence



Next Sentence Prediction (NSP)

Understand relationships between sentences



BERT

1 Pre - Training

2 Fine-Tuning

Task specific dataset

- ✓ Test Classification
- ✓ Sentiment Analysis
- ✓ NER
- ✓ POS tagging

2 Fine-Tuning

- Output Layer Tuning
- Full Model Fine Tuning
- Feature Extraction

Fine tuning

Output Layer Tuning

- Task specific outer layer is trained while freezing the lower layers

Full Model Fine Tuning

- Fine tuning all the layers of the model.
- Effective for large dataset.

Feature Extraction

1. Domain Adaptive helps the model in Feature Extraction
2. Fine tuning on specific tasks

GLUE

General **L**anguage **U**nderstanding **E**valuation

Measure performance of models on various language understanding challenges.

Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP
1	Microsoft Alexander v-team	Turing ULR v6		91.3	73.3	97.5	94.2/92.3	93.5/93.1	76.4/90.9
2	JDExplore d-team	Vega v1		91.3	73.8	97.9	94.5/92.6	93.5/93.1	76.7/91.1
3	Microsoft Alexander v-team	Turing NLR v5		91.2	72.6	97.6	93.8/91.7	93.7/93.3	76.4/91.1
4	DIRL Team	DeBERTa + CLEVER		91.1	74.7	97.6	93.3/91.1	93.4/93.1	76.5/91.0
5	ERNIE Team - Baidu	ERNIE		91.1	75.5	97.8	93.9/91.8	93.0/92.6	75.2/90.9
6	AliceMind & DIRL	StructBERT + CLEVER		91.0	75.3	97.7	93.9/91.9	93.5/93.1	75.6/90.8
7	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.8	71.5	97.5	94.0/92.0	92.9/92.6	76.2/90.8
8	HFL iFLYTEK	MacALBERT + DKM		90.7	74.8	97.0	94.5/92.6	92.8/92.6	74.7/90.6
9	PING-AN Omni-Sinitic	ALBERT + DAAF + NAS		90.6	73.5	97.2	94.0/92.0	93.0/92.4	76.1/91.0
10	T5 Team - Google	T5		90.3	71.6	97.5	92.8/90.4	93.1/92.8	75.1/90.6
11	Microsoft D365 AI & MSR AI & GATECH	MT-DNN-SMART		89.9	69.5	97.5	93.7/91.6	92.9/92.5	73.9/90.2
12	Huawei Noah's Ark Lab	NEZHA-Large		89.8	71.7	97.3	93.3/91.0	92.4/91.9	75.2/90.7
13	LG AI Research	ANNA		89.8	68.7	97.0	92.7/90.1	93.0/92.8	75.3/90.5

Click on a submission to see more information

Preparing for LLMs

Video 3: Using Pre trained transformers : BERT

Analytics Vidhya