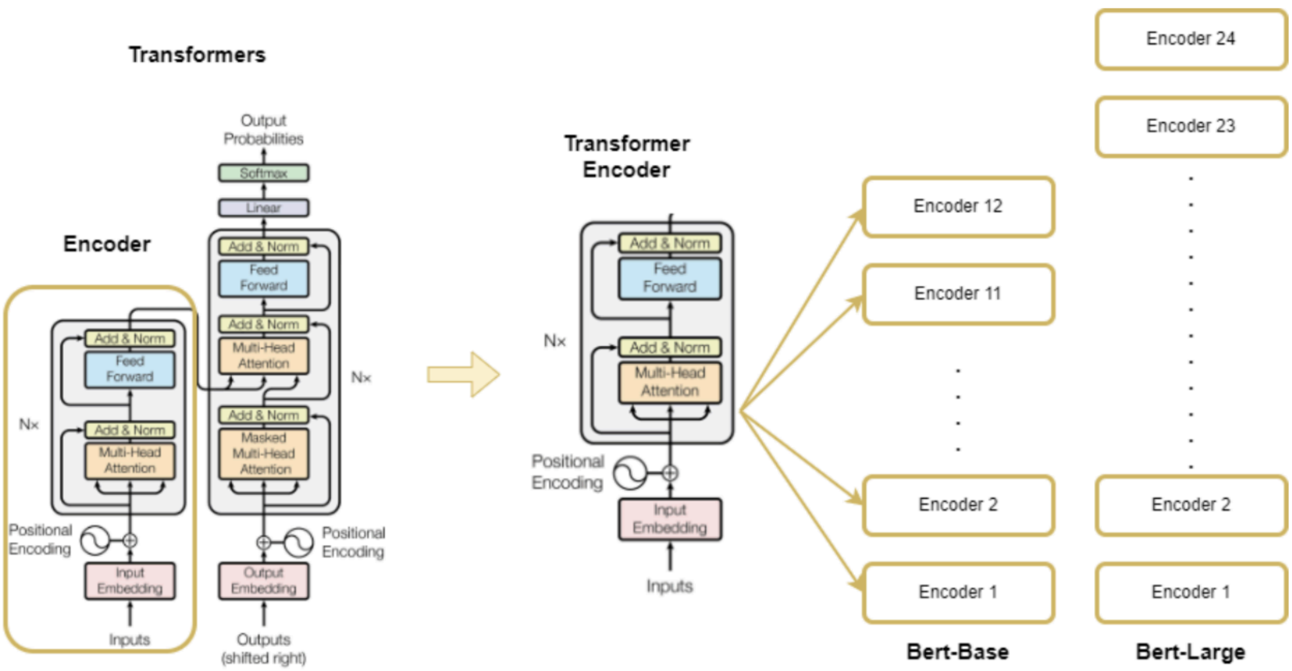
Lecture 11: Bert

Bert

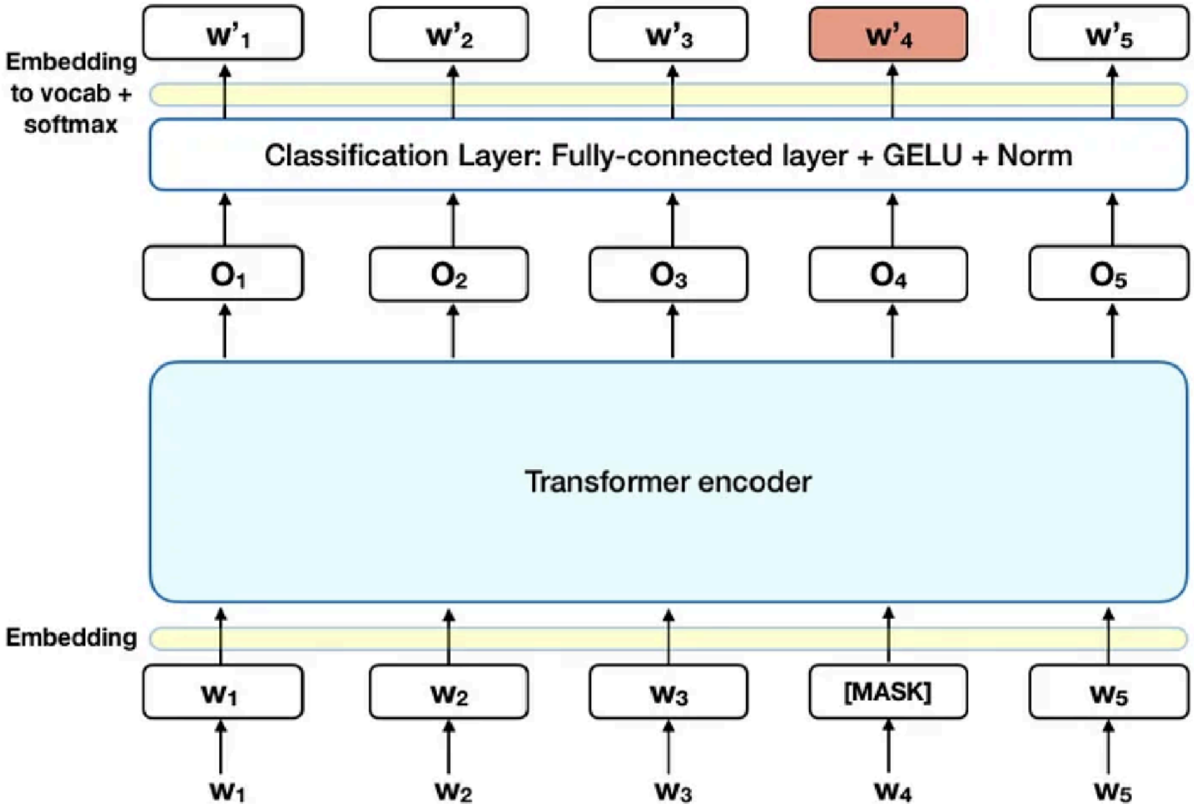
- BERT stands for BiDirectional Encoder Representation from Transformers.
- BERT uses only the Encoder portion of the transformer's architecture which is the reason it's called **Encoder Representation from transformers.**



Bert uses two training mechanisms namely **Masked Language Modeling (MLM)** and **Next Sentence Prediction (NSP)** to overcome the dependency challenge.

Masked Language Modeling (MLM)

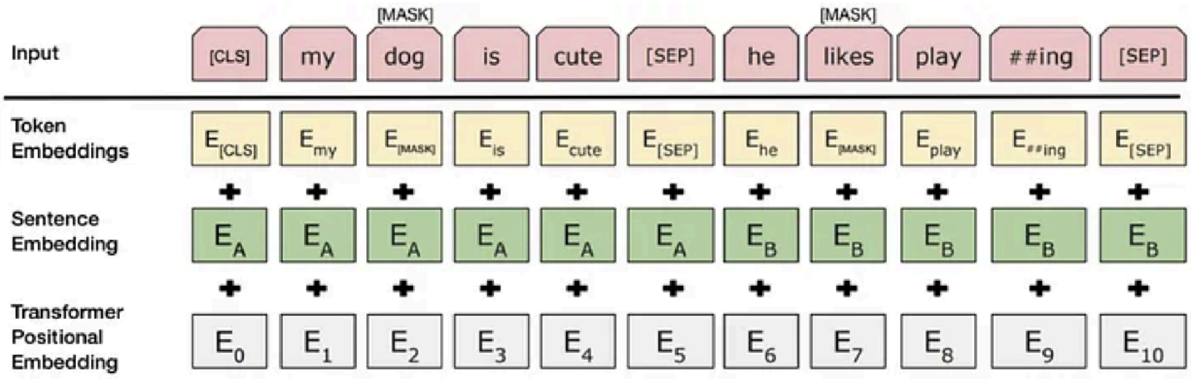
Before feeding the input vector into the encoder, 15% of the words in the sequence will be masked with a [MASK] token. The goal of MLM is to predict the masked word with respect to all other words in the sentence.


w'1
w'2
w'3
w'4
w'5
Embedding to vocab + softmax
Classification Layer: Fully-connected layer + GELU + Norm
O1
O2
O3
O4
O5
Transformer encoder
Embedding
w1
w2
w3
[MASK]
w5
w1
w2
w3
w4
w5

Encoded output of the Transformer encoder is sent to a fully connected classification layer. The results of the classifier would then be multiplied by an embedding matrix to convert to the vocabulary

Next Sentence Prediction (NSP)

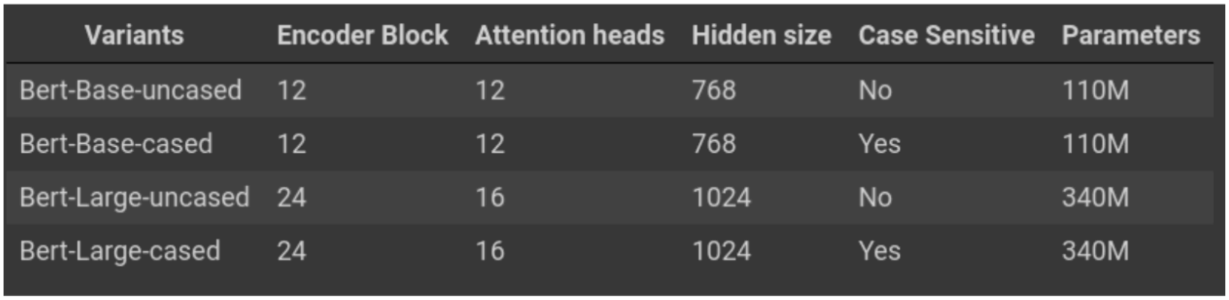
While training, the model would receive pairs of sentences as inputs. The model eventually learns it and predicts whether the second sentence in the pair is the consecutive sentence or not.



We can understand, a[CLS] token at the beginning and a [SEP] token at the end of a sentence are added.

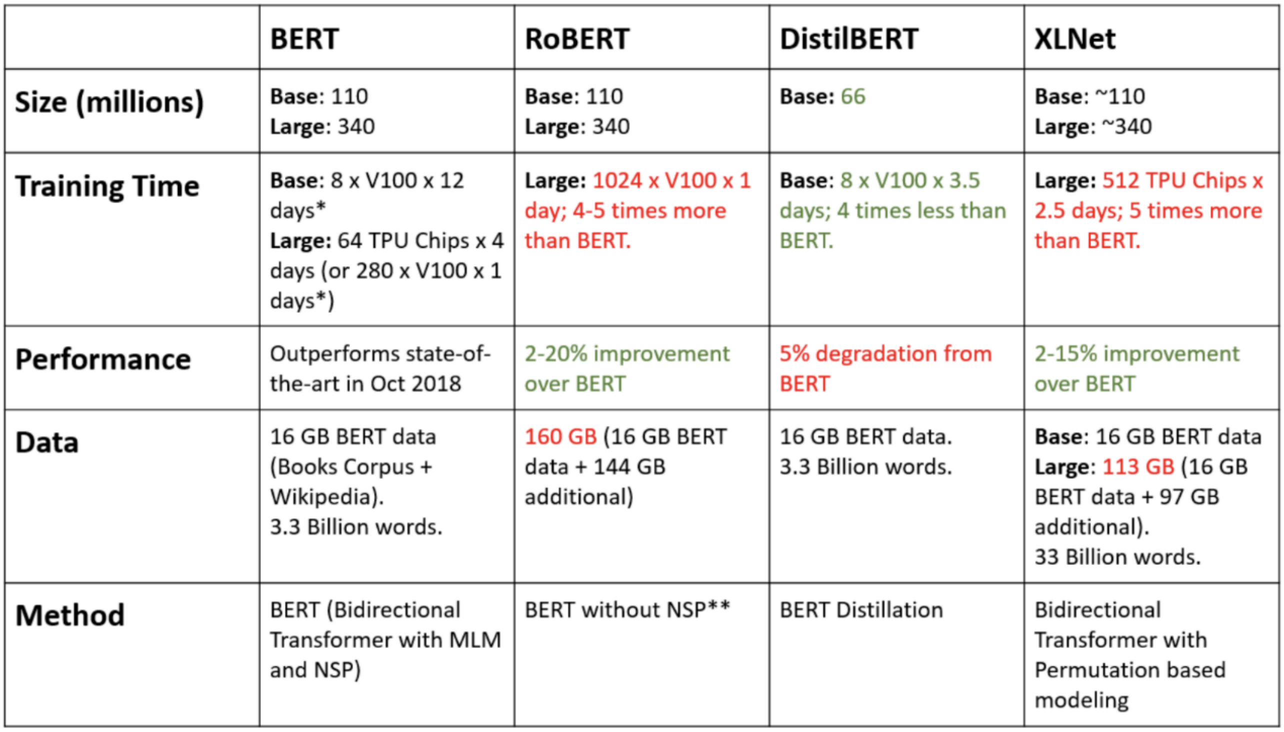
To predict whether the second sentence is connected to the first sentence, the entire input sequence would be sent to a Transformer encoder and then the output of the [CLS] token is transformed into a 2×1 shaped vector, using a simple classification layer.

There are 4 variants of BERT based on the number of encoder blocks and Attention heads



Variants	Encoder Block	Attention heads	Hidden size	Case Sensitive	Parameters
Bert-Base-uncased	12	12	768	No	110M
Bert-Base-cased	12	12	768	Yes	110M
Bert-Large-uncased	24	16	1024	No	340M
Bert-Large-cased	24	16	1024	Yes	340M

More architectures based on Transformers



	BERT	RoBERT	DistilBERT	XLNet
Size (millions)	**Base**: 110 **Large**: 340	**Base**: 110 **Large**: 340	**Base:** 66	**Base**: ~110 **Large**: ~340
Training Time	**Base**: 8 x V100 x 12 days* **Large:** 64 TPU Chips x 4 days (or 280 x V100 x 1 days*)	**Large:** 1024 x V100 x 1 day; 4-5 times more than BERT.	**Base**: 8 x V100 x 3.5 days; 4 times less than BERT.	**Large:** 512 TPU Chips x 2.5 days; 5 times more than BERT.
Performance	Outperforms state-of-the-art in Oct 2018	2-20% improvement over BERT	5% degradation from BERT	2-15% improvement over BERT
Data	16 GB BERT data (Books Corpus + Wikipedia). 3.3 Billion words.	160 GB (16 GB BERT data + 144 GB additional)	16 GB BERT data. 3.3 Billion words.	**Base**: 16 GB BERT data **Large**: 113 GB (16 GB BERT data + 97 GB additional). 33 Billion words.
Method	BERT (Bidirectional Transformer with MLM and NSP)	BERT without NSP**	BERT Distillation	Bidirectional Transformer with Permutation based modeling