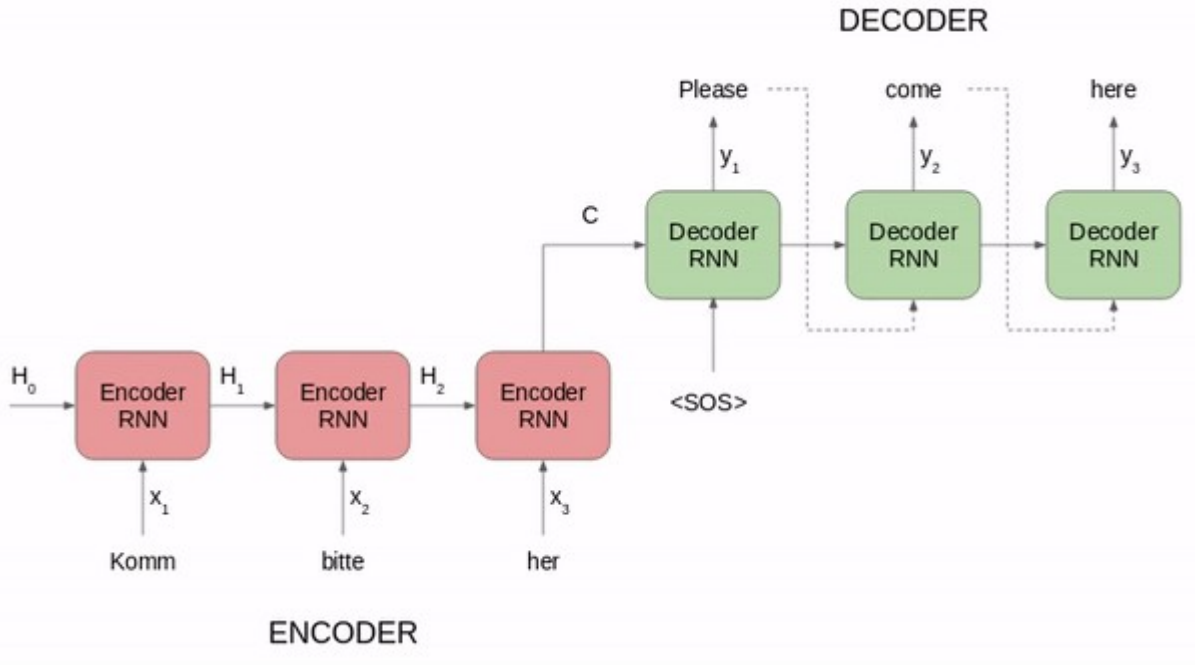
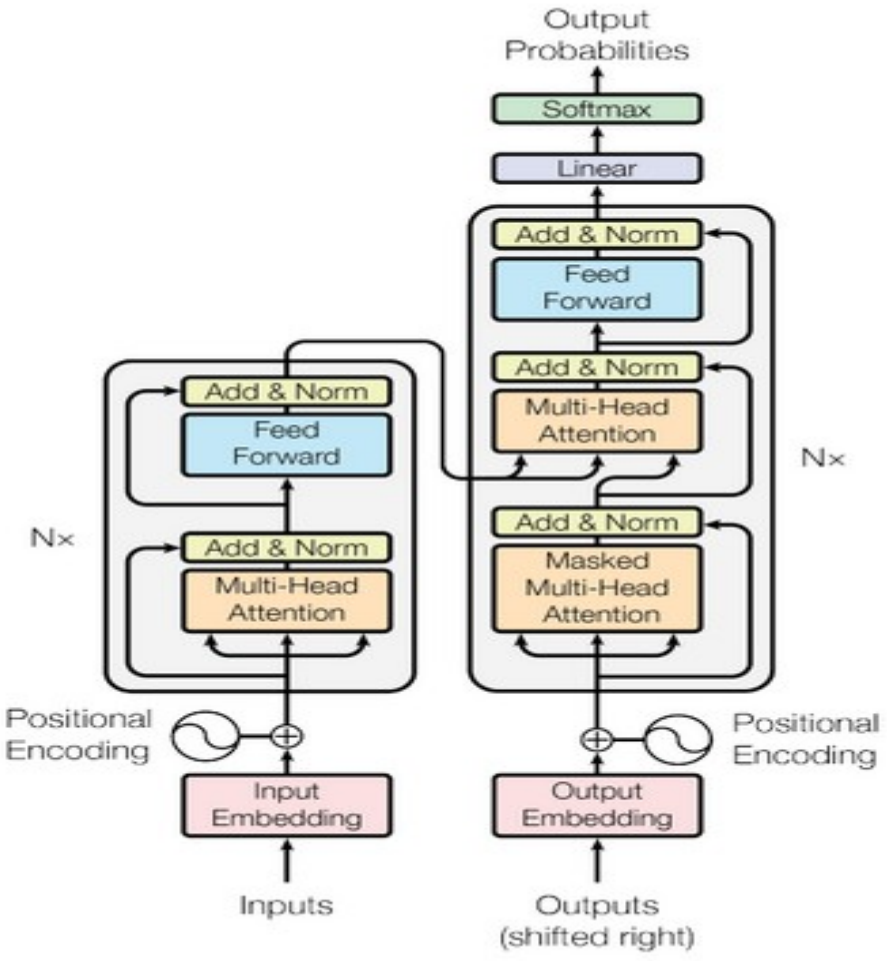
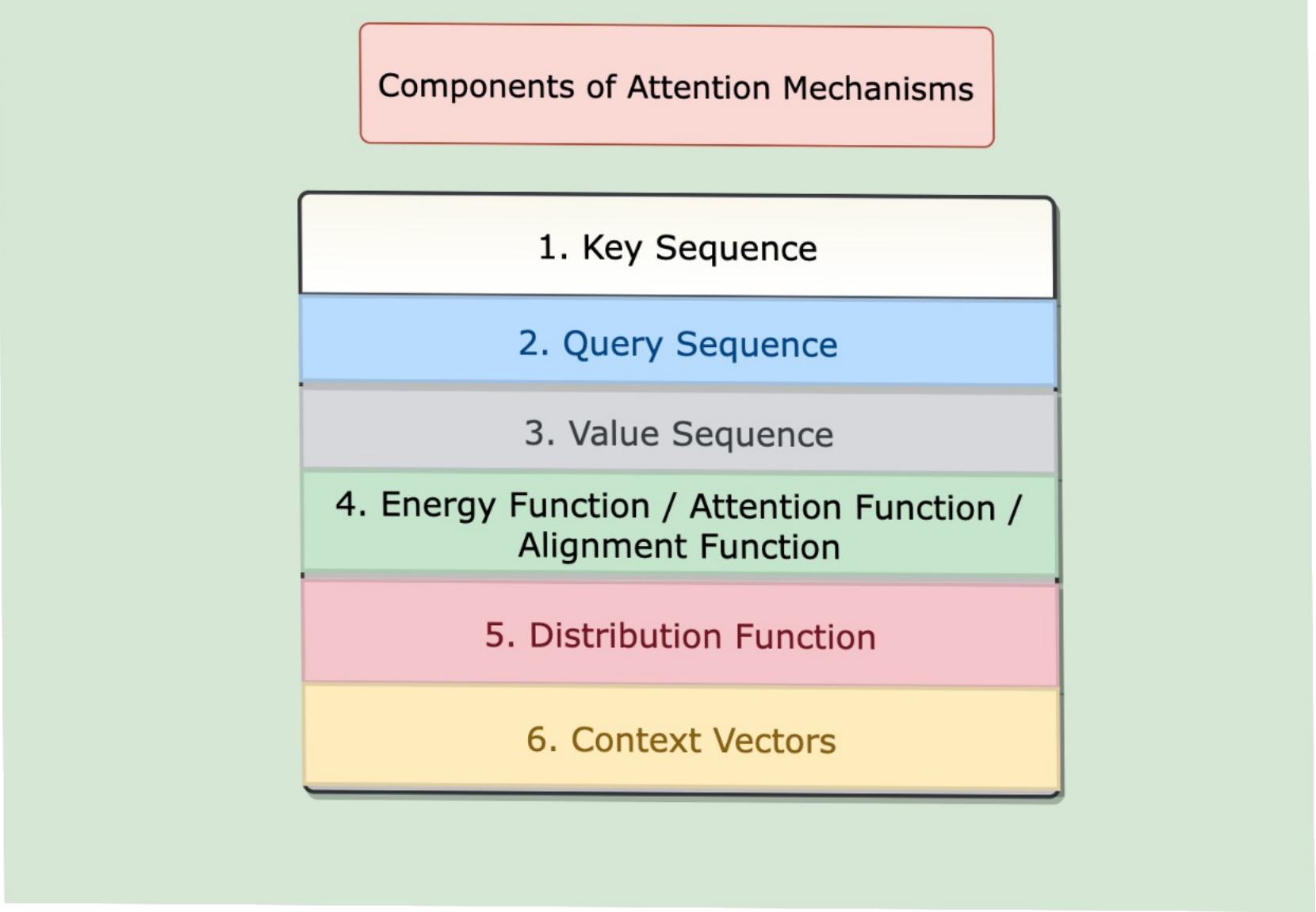
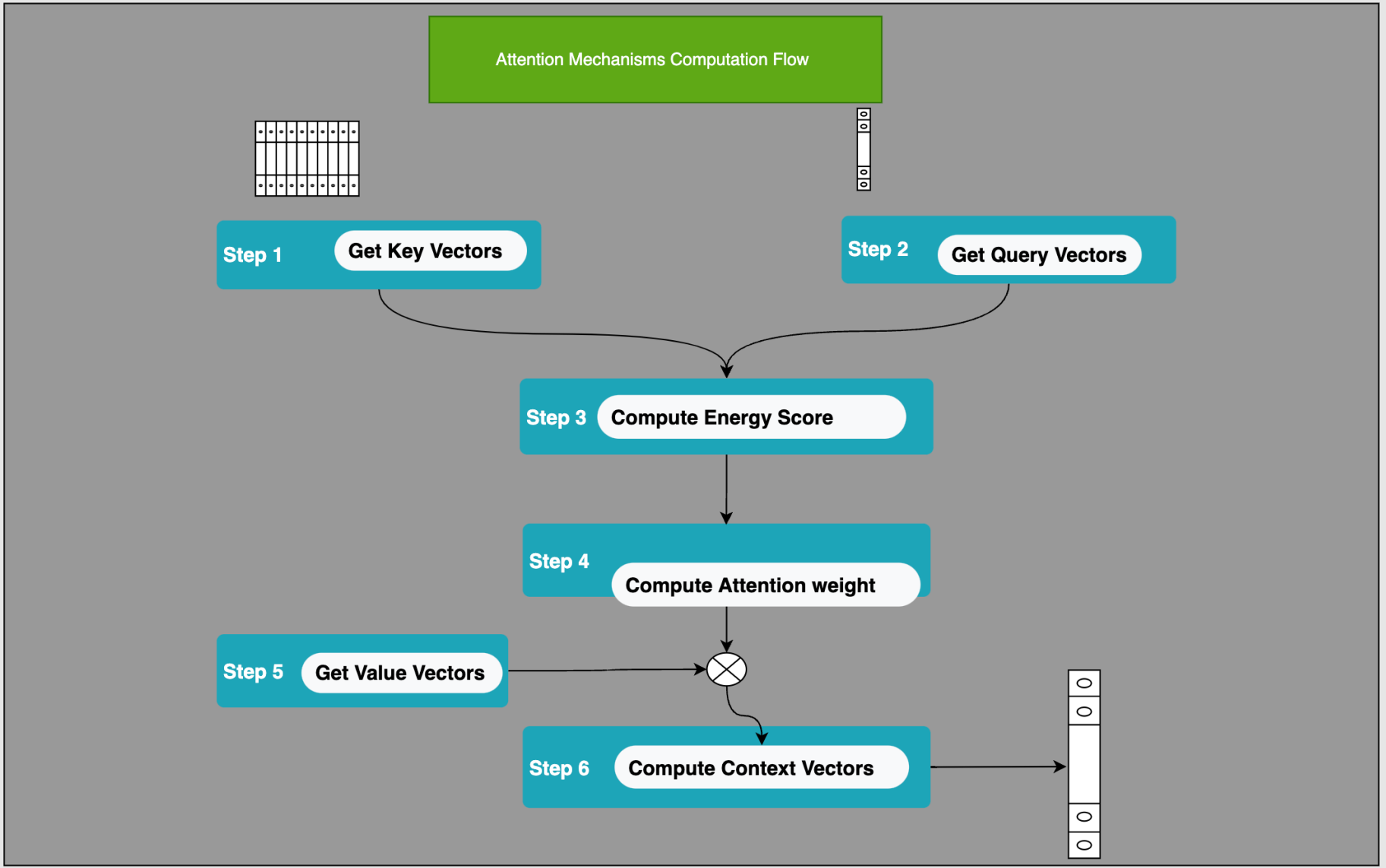
Transformers and its applications

RNN


DECODER
Please
come
here
y_1
y_2
y_3
C
Decoder RNN
Decoder RNN
Decoder RNN
H_0
Encoder RNN
H_1
Encoder RNN
H_2
Encoder RNN
<SOS>
x_1
x_2
x_3
Komm
bitte
her
ENCODER


Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)


Components of Attention Mechanisms
1. Key Sequence
2. Query Sequence
3. Value Sequence
4. Energy Function / Attention Function / Alignment Function
5. Distribution Function
6. Context Vectors


Attention Mechanisms Computation Flow
Step 1
Get Key Vectors
Step 2
Get Query Vectors
Step 3
Compute Energy Score
Step 4
Compute Attention weight
Step 5
Get Value Vectors
Step 6
Compute Context Vectors

Transformers and its applications (cont..)

Soft Attention

Attention score is used as weights in the weighted average context vector calculation. This is a differentiable function.

$$z_i = w_i * v_i$$

$$C = \sum_{i=1}^{d_k} (z_i)$$

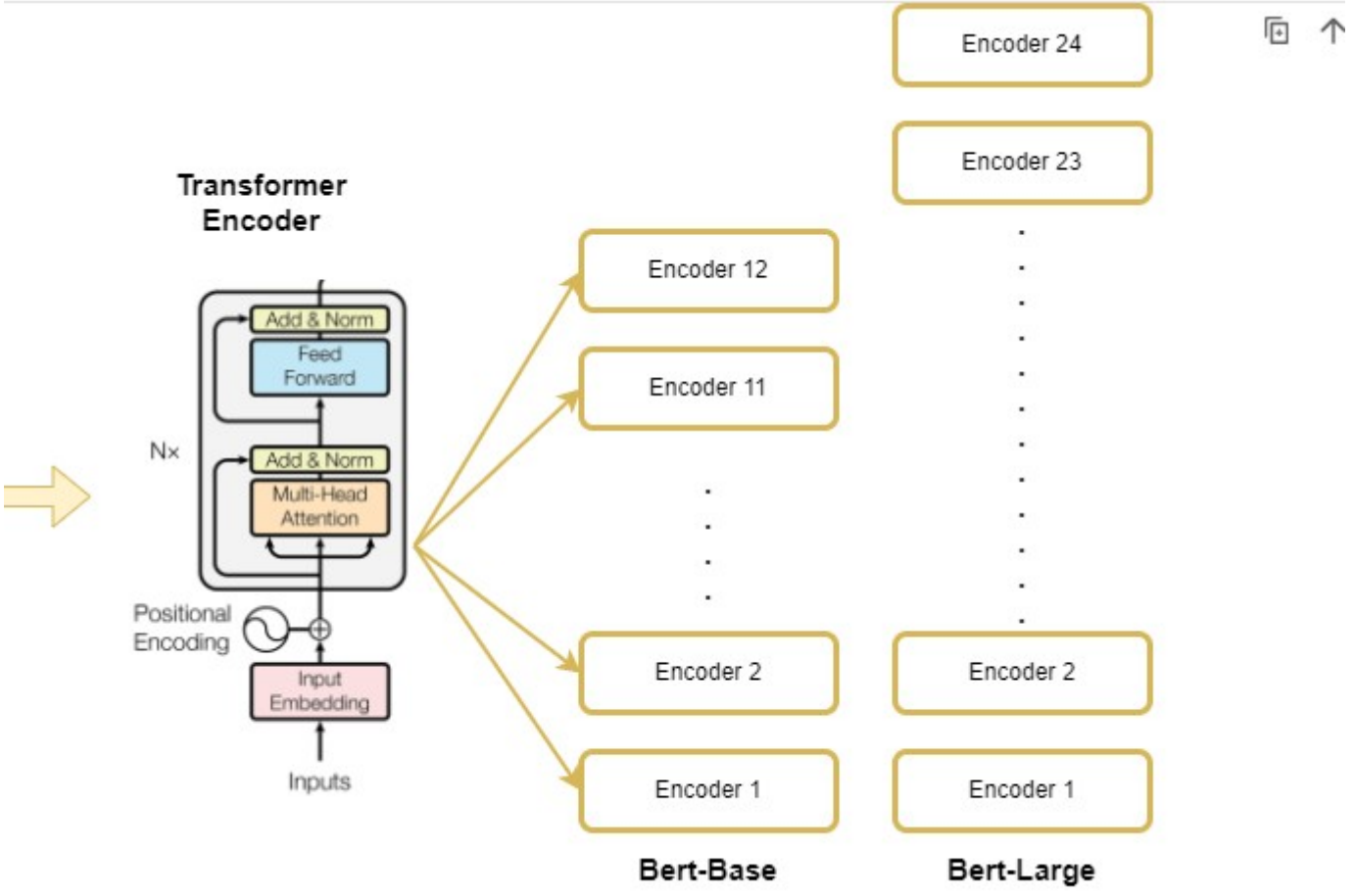
Hard Attention

Attention score is used as the probability of the i-th location getting selected. We could use a simple argmax to make the selection, but it is not differentiable and so complex techniques are employed.

$$p(s_{t,i} = 1 | s_t, v) = w_{t,i}$$

$$s_t{}^n \approx Multinoulli_L(w_i{}^n)$$

where s_t is a one hot encoder vector which is $= 1$ at position i


Transformer Encoder
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Input Embedding
Inputs
Encoder 12
Encoder 11
Encoder 2
Encoder 1
Bert-Base
Encoder 24
Encoder 23
Encoder 2
Encoder 1
Bert-Large

How BERT is different from RNN/LSTM?

RNN	BERT
It is built with recurrent blocks	It is built with transformer blocks
Each word in a sequence are passed one after the other	The entire sentence is passed at once
If N is the no.of words in a sequence, RNNs has to compute N sequential steps	It performs only one step for the entire sequence of length N
Similarly for the gradients to propagate from last word to the first word it takes N steps	In Bert, it is just one step process
Because of sequential processing there is no parallelization	It can be parallelization since it is not passed sequentially
They suffer from Vanishing gradients though LSTM/GRU can handle to an extent	Bert does not suffer from Vanishing gradient problem and can learn long term dependencies very well
No self attention	Has self attenion

BERT (cont..)

- Is BERT a word embedding model or pre-trained model?

- How was it pre-trained?
- Data- Wikipedia (2500 million words) and bookcorpus(800 million words)

BERT

- BERT was pre-trained on two tasks:
- 1. Predicting the masked works
- 2. Next Sentence prediction

Examples of predicting the masked works

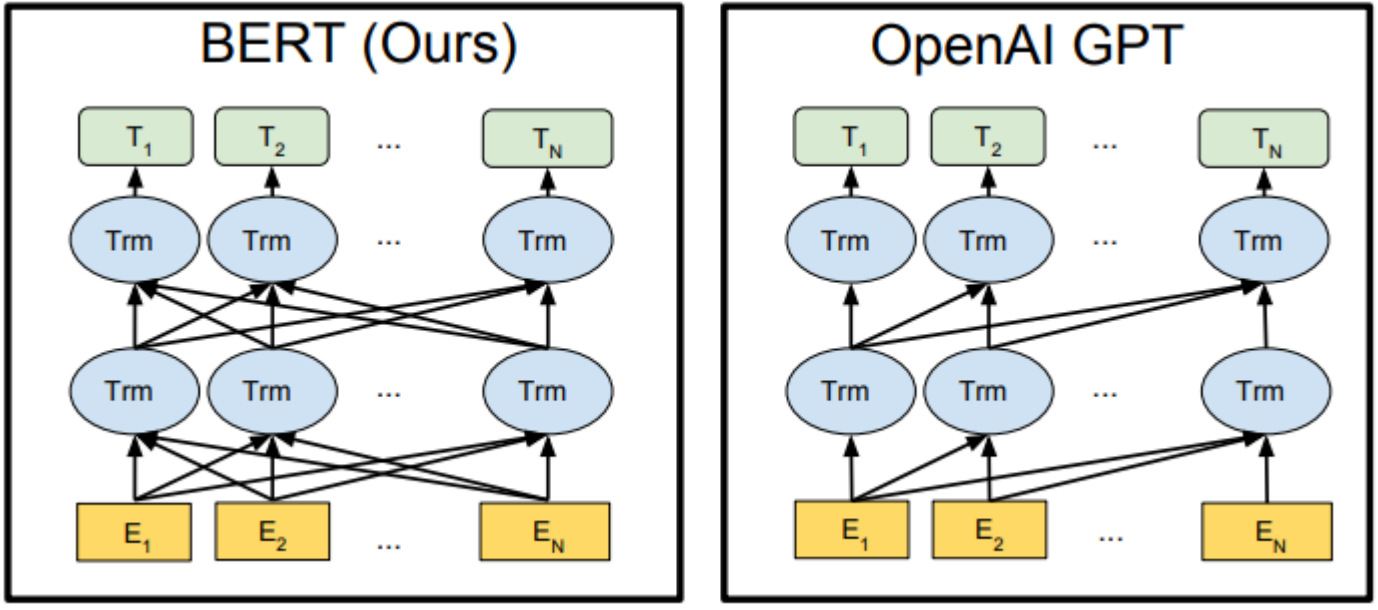
Actual sentence	Masked train sentence	Label
Today morning, I went for a tooth removal to my dentist	Today morning, I went for a **< MASK >** removal to my dentist	tooth
Did you go to school today	Did you **< MASK >** to school today	go

Examples of NSP

Sentence A	Sentence B	IsNext?
The Indian cricket team won the match against West Indies.	Obama served 2 terms as the president of USA.	No
Nadal won the 2022 Australian open.	With this victory he now has 21 GrandSlam titles to his name.	Yes

- **Which one is easier to predict?**
- Today morning, I went for a_____?
- Today morning, I went for a_____ removal to my dentist?
-

BERT is bi-directional



Types of BERT

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 on Transformer

Encoder only	Decoder only	Encoder + Decoder
BERT		Transformer
RoBerta	Transformer-XL	T5
Reformer	XLNet	XLM
FlauBert	GPT	XLM-RoBerta
CamemBert		BART
Longformer		

Comparison of different architectures

	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

Comparison of different architectures(cont..)

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