BERT

Bidirectional encoder representations from Transformers

https://jalammar.github.io/illustrated-bert/

1 Masked language modeling (MLM)

Mask out k% of the input words, and then predict the masked words

- Recommendation use k = 15%

between too little and too much masking.

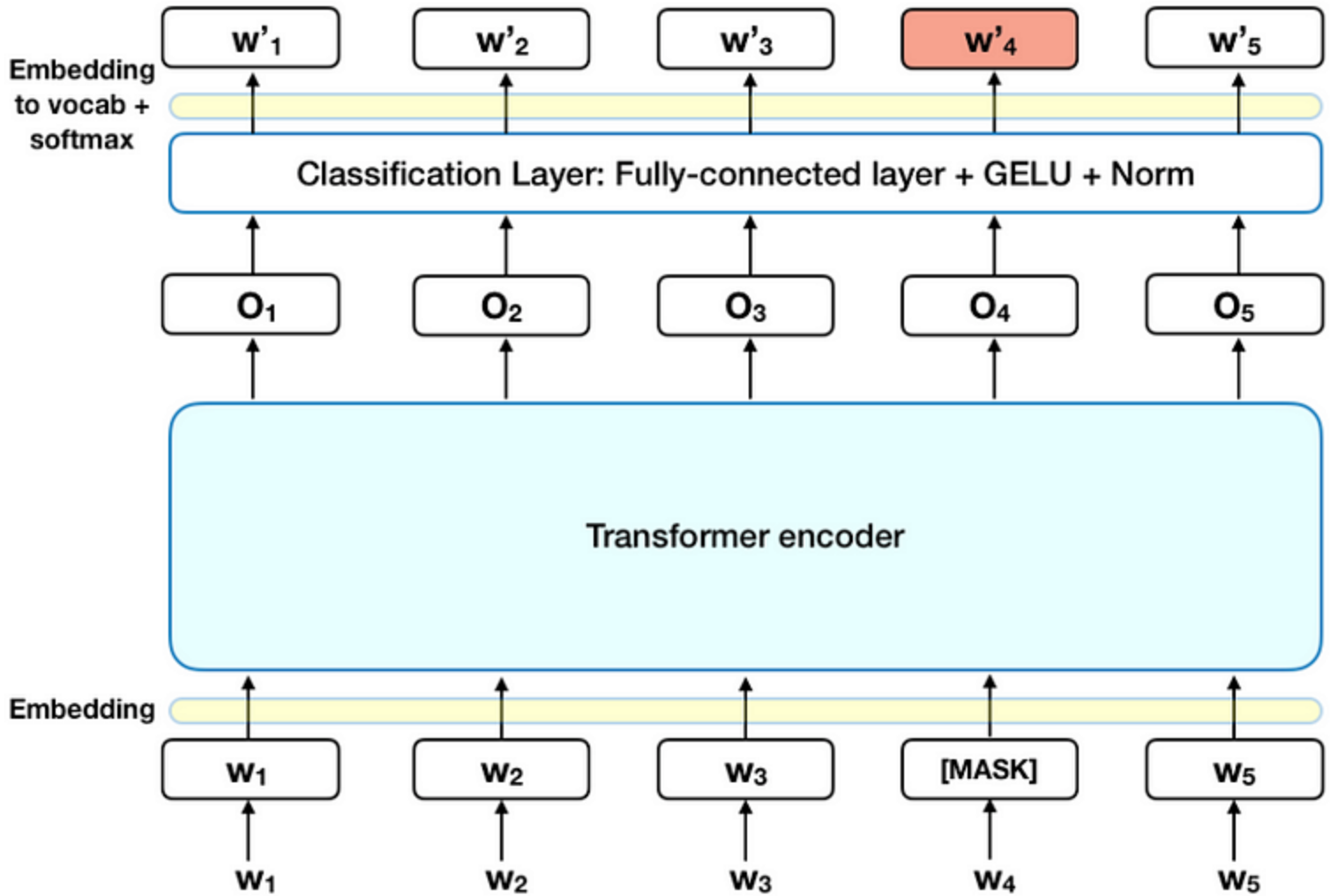
2 Next sentence prediction (NPS)

Binary classification task

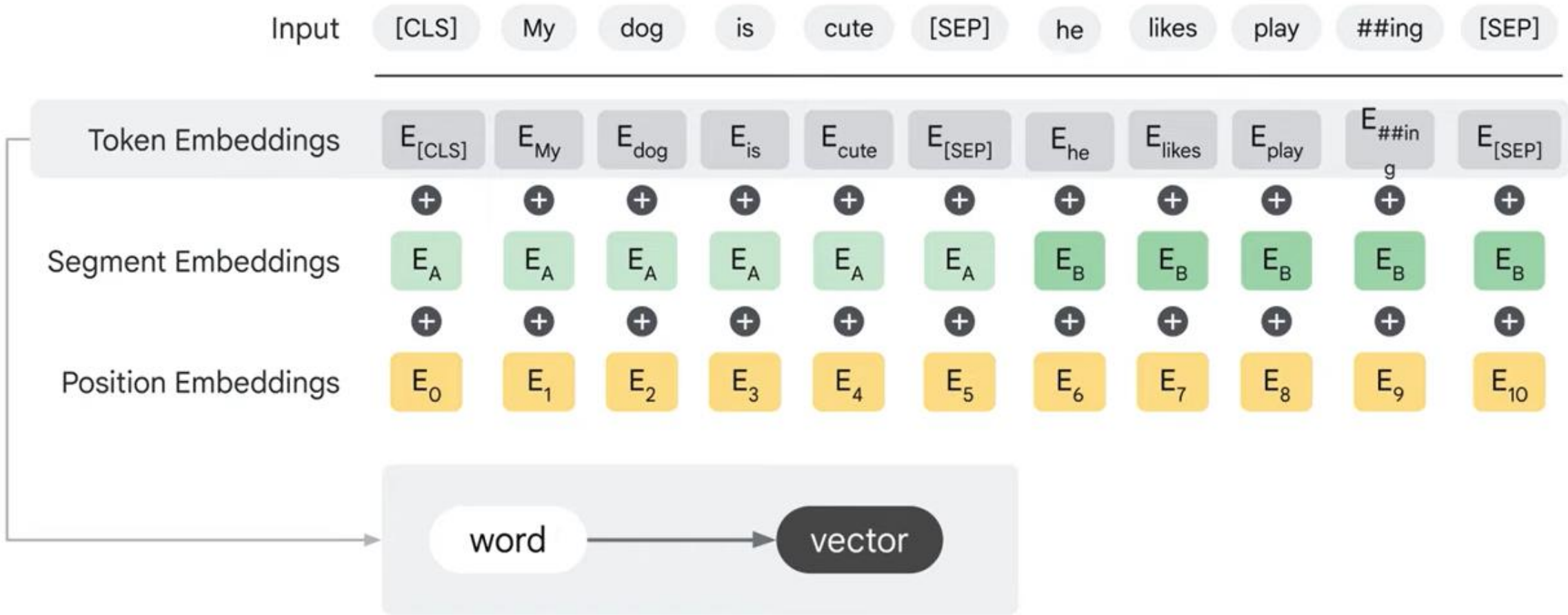
Learn the relationships between sentences and predict the next sentence given the first one.

Sentence A	The man went to the store.
Sentence B	He bought a gallon of milk.
Label	IsNextSentence

Sentence A	The man went to the store.
Sentence B	Penguins are flightless.
Label	NotNextSentence

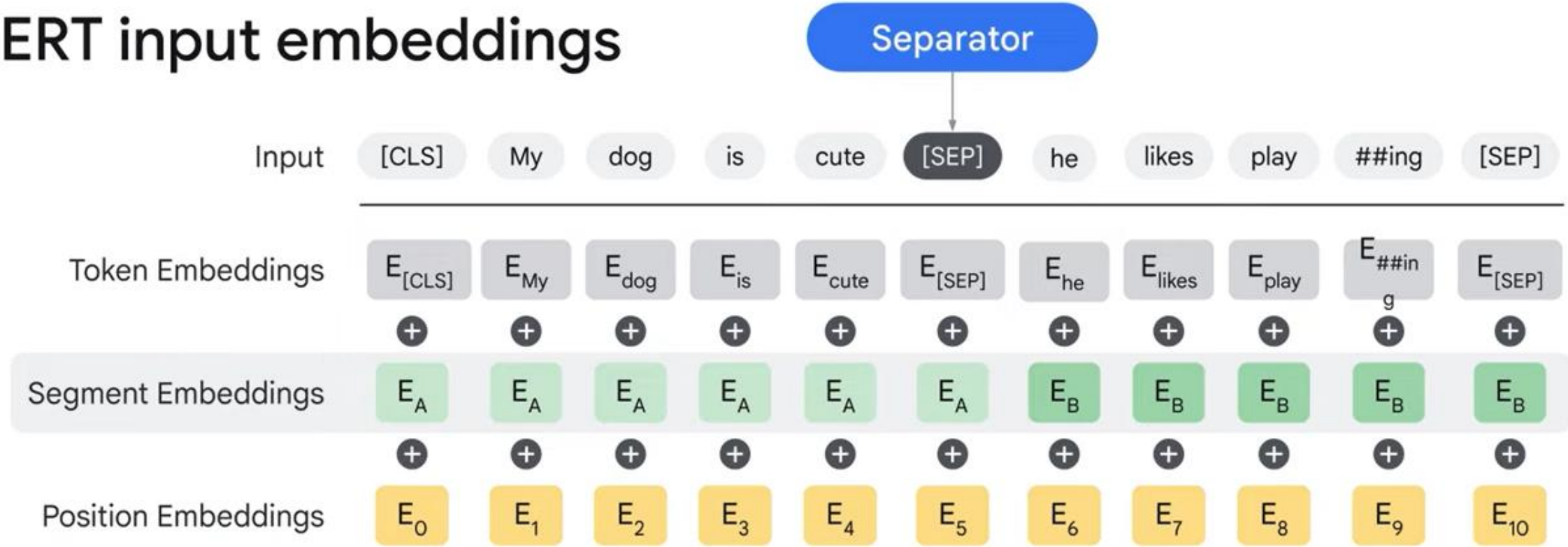

w'1
w'2
w'3
w'4
w'5
Embedding to vocab + softmax
Classification Layer: Fully-connected layer + GELU + Norm
O1
O2
O3
O4
O5
Transformer encoder
Embedding
w1
w2
w3
[MASK]
w5
w1
w2
w3
w4
w5

BERT input embeddings

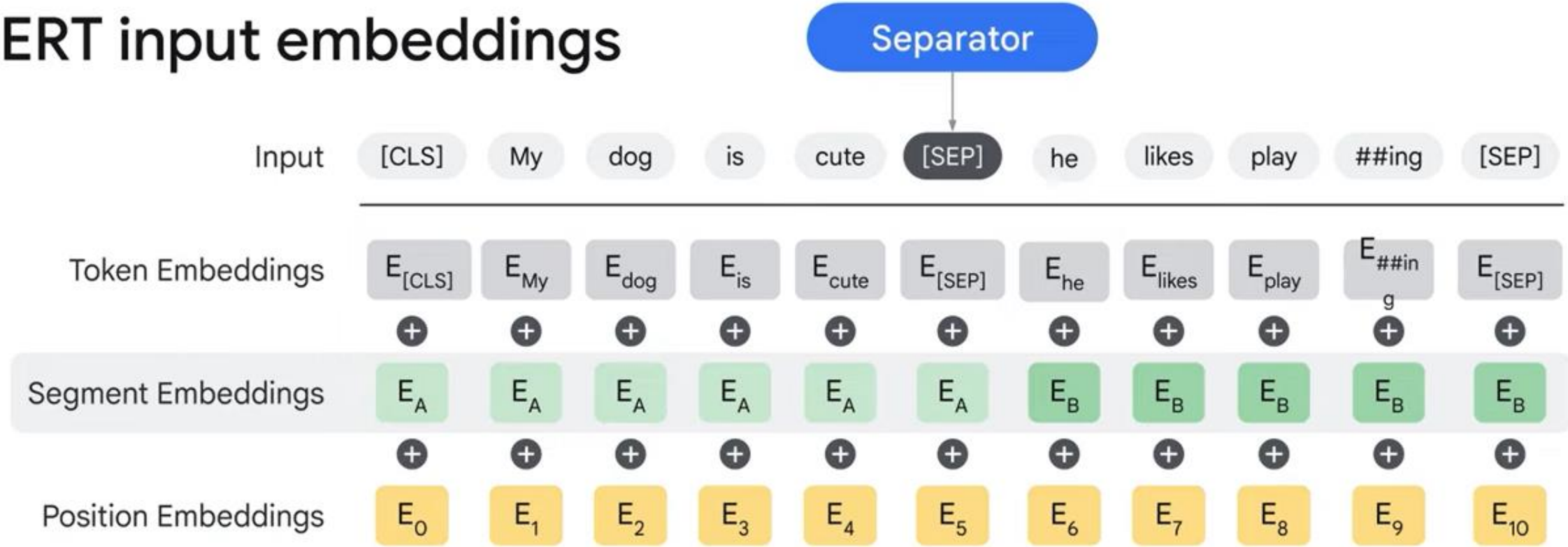


The token embeddings is a representation

BERT input embeddings

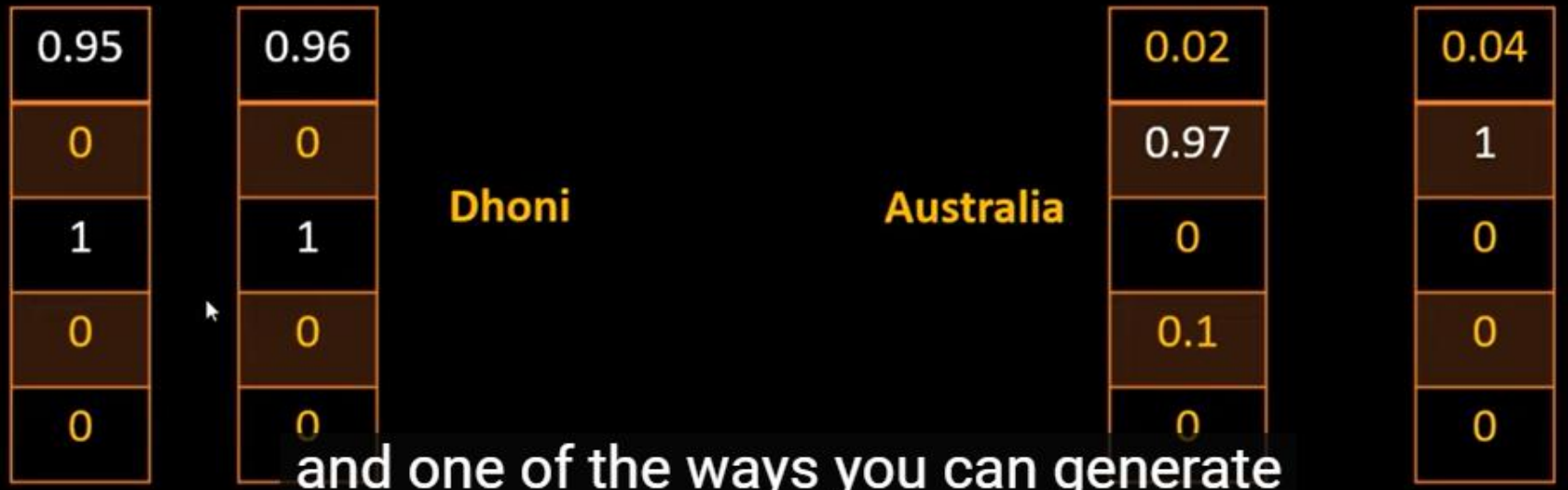


BERT input embeddings



The answer is to use segment embeddings.

	ashes	Australia	Bat	Cummins	cover	Dhoni	World cup	..	Zimbabwe
Person	0	0.02	0.1	**0.95**	0.03	**0.96**	0.67	...	0.04
Country	0	**0.97**	0	0	0	0	0	...	**1**
Healthy & Fit	0	0	0.3	**0.87**	0	**0.9**	0	...	0
Event	**1**	0.1	0	0	0.4	0	**1**	...	0
gear	0	0	**1**	0	0	0	0	...	0



and one of the ways you can generate

Issue with word2vec

Fixed embeddings

He didn't receive fair treatment

Fun fair in new York city this summer

sentence

BERT can generate contextualized embeddings

He didn't receive fair treatment → fair → [1, 0.9, 0.2, 1, 0.7]

Fun fair in new York city this summer → fair → [0, 0.1, 0, 0.5, 0.4]

it will generate it differently here

BERT can generate contextualized embeddings

it will generate a vector which is

BERT can generate embeddings for entire sentence

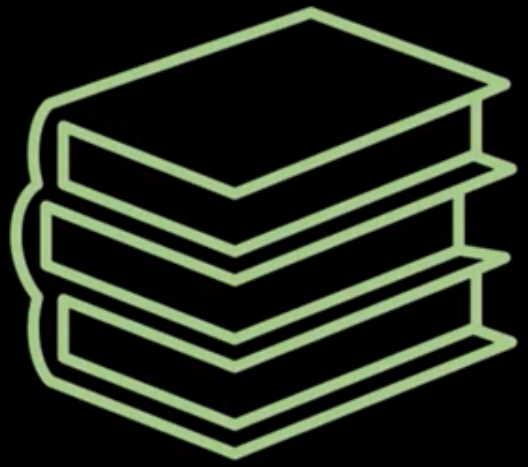
Amazing movie, I couldn't blink an eye for initial 45 minutes. It was that intense and interesting at the same time

768 length

be anything

Trained on

2500 M words



800 M words

2500 million words in the wikipedia

How was it trained?

Mased Language Model

Elon Reeve Musk is an ..... and business magnate. He is the founder of ..... Musk is one of the richest people in the world. Musk was raised in Pretoria, ... He briefly attended the University of Pretoria before moving to ..... aged 17 to attend Queen's University.

and they would generate this training

How was it trained?

Mased Language Model

Elon Reeve Musk is an ..... and business magnate. He is the founder of ..... Musk is one of the richest people in the world. Musk was raised in Pretoria, ... He briefly attended the University of Pretoria before moving to ..... aged 17 to attend Queen's University.

meaningful word and sentence embedding.

How was it trained?

Mased Language Model

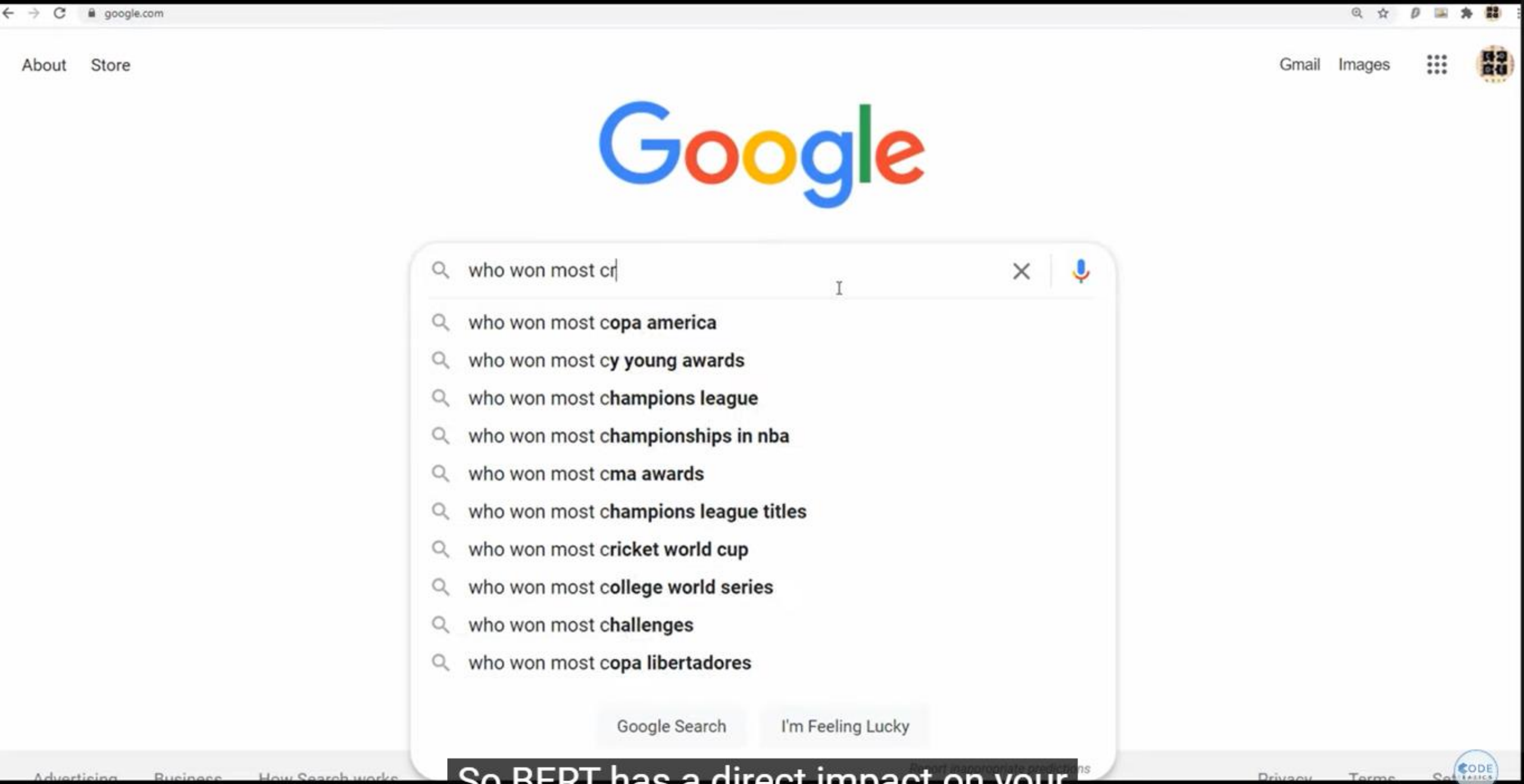
Elon Reeve Musk is an ..... and business magnate. He is the founder of ..... Musk is one of the richest people in the world. Musk was raised in Pretoria, ... He briefly attended the University of Pretoria before moving to ..... aged 17 to attend Queen's University.

Next sentence prediction

I am hungry → I would like to have pizza ✓

→ Table has four lags ✗

than you know table has four legs who



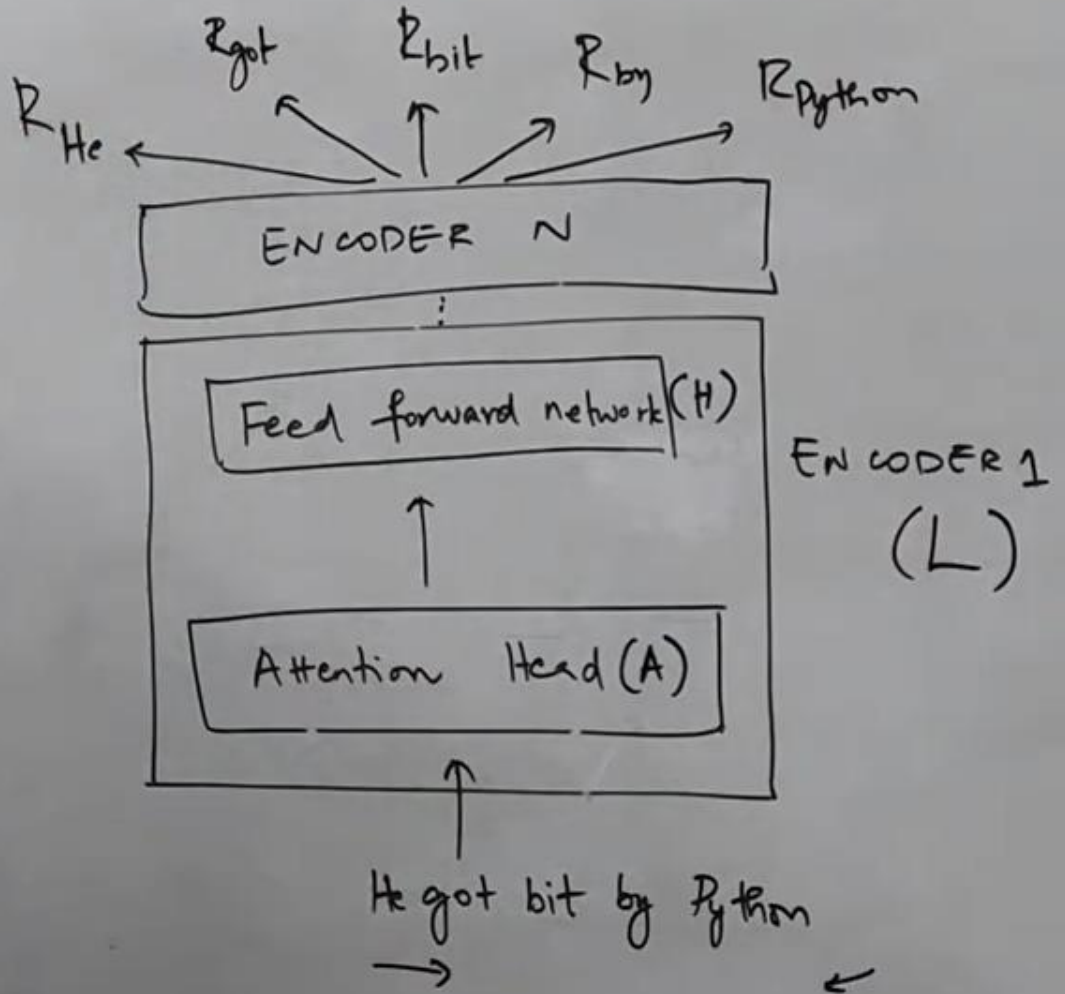
So BERT has a direct impact on your

Bidirectional Encoder Representations from Transformers

you're curious about

BERT

Bi-directional Encoder Representation from Transformers



CONFIGURATION OF BERT

Bi-directional Encoder Representation from Transformers

	BERT base	BERT large	BERT tiny	BERT mini	BERT small	BERT medium
Encoders (L)	12	24	2	4	4	8
Attention Heads (A)	12	16	2	4	4	8
Hidden Units (H)	768	1024	128	256	512	512