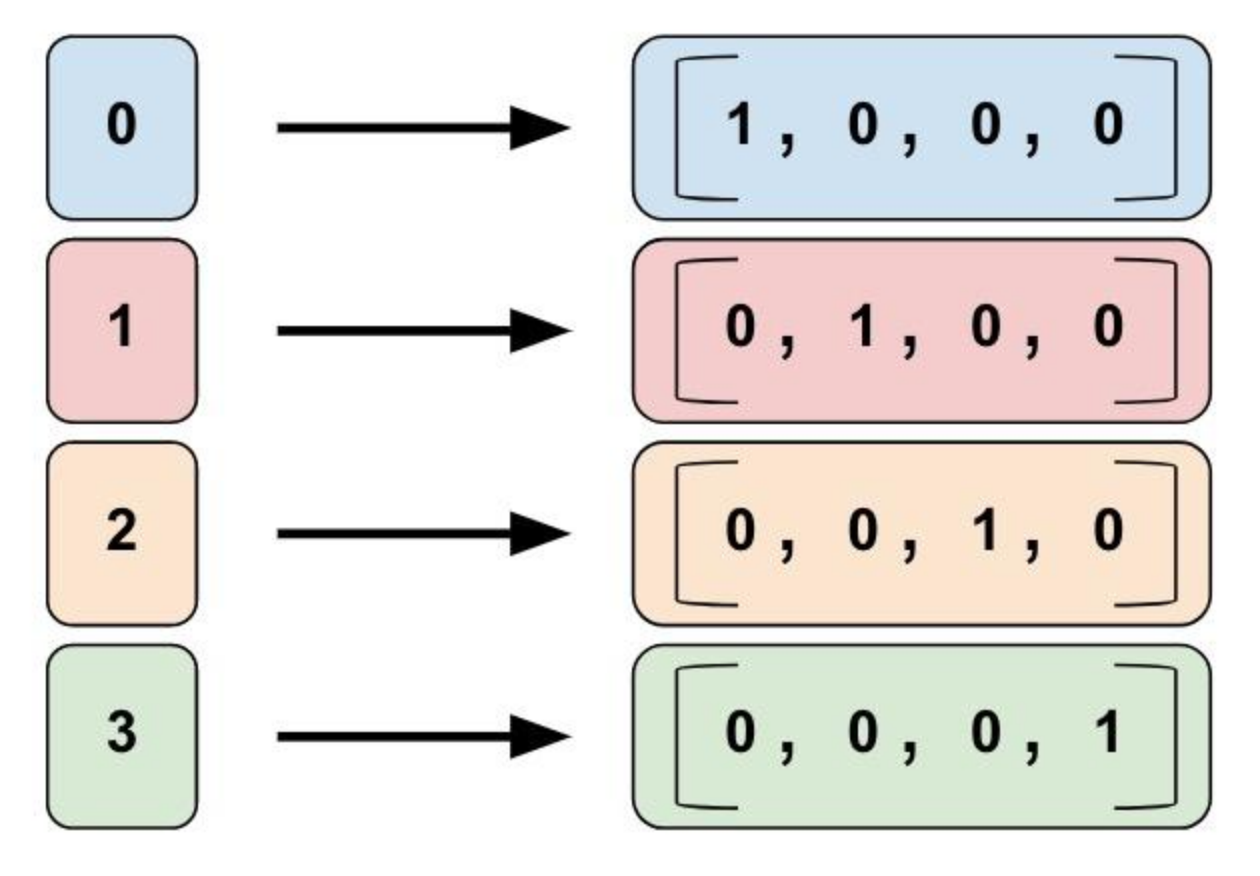
BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS

Pedro Lelis • Cientista de Dados • pedrolelis@ciandt.com

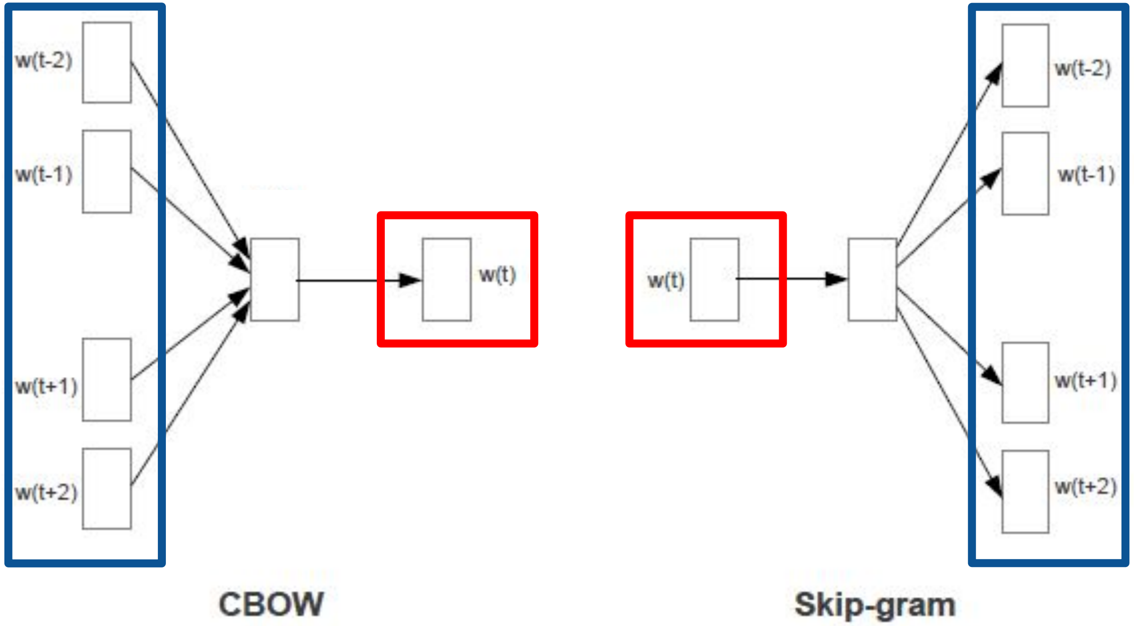
BERT

Pedro Lelis • Cientista de Dados • pedrolelis@ciandt.com

Word Embedding

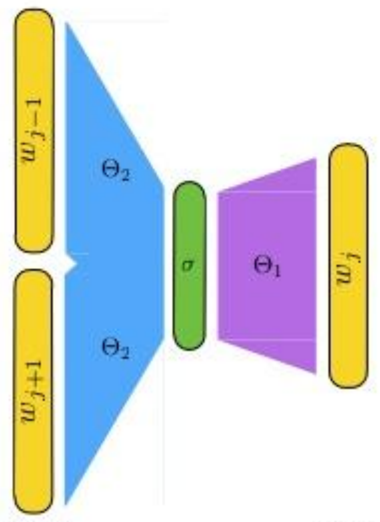

0
1
2
3
[1 , 0 , 0 , 0]
[0 , 1 , 0 , 0]
[0 , 0 , 1 , 0]
[0 , 0 , 0 , 1]

"The fat cat sat on the mat."



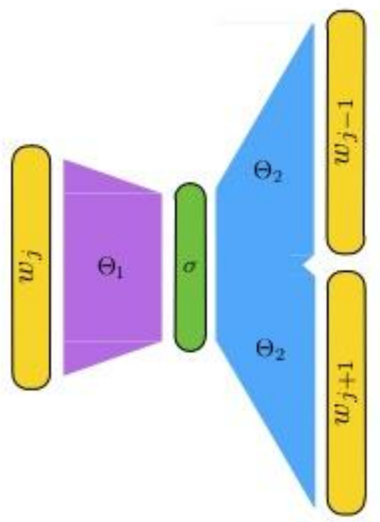
CBOW

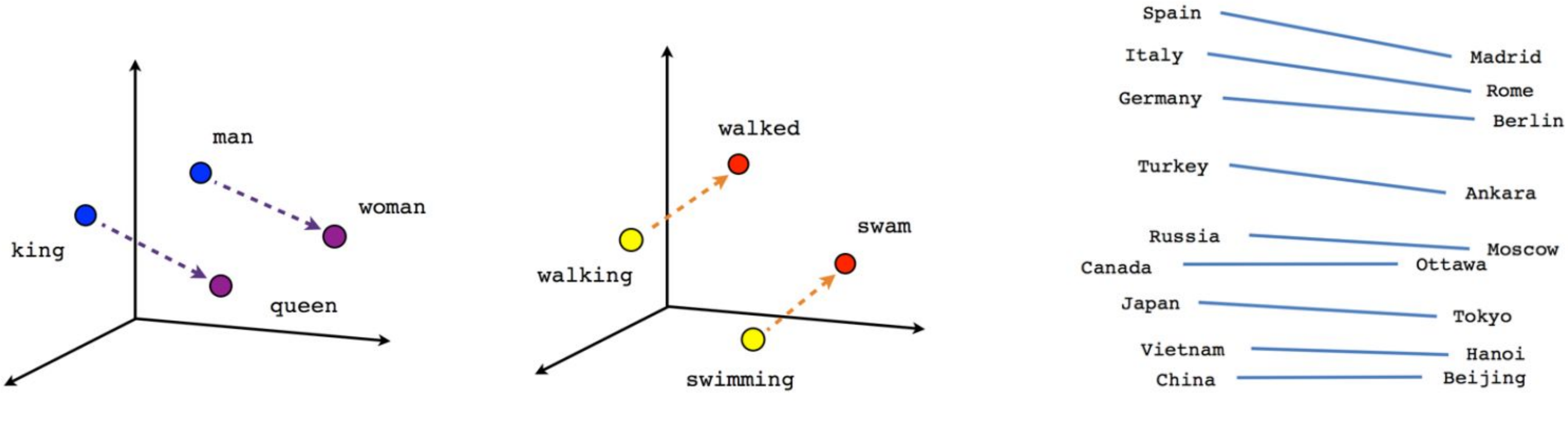
$$\max p\,(w|C)$$



Skip-gram

$$\max p\,(C|w)$$



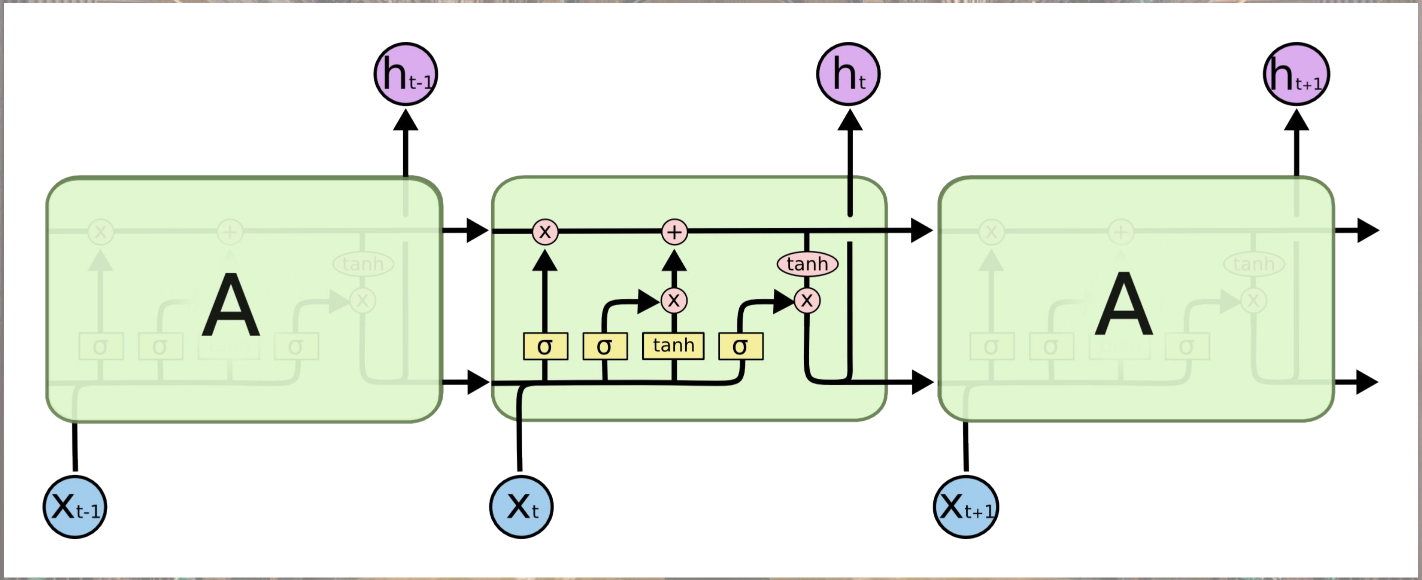


Male-Female

Verb tense

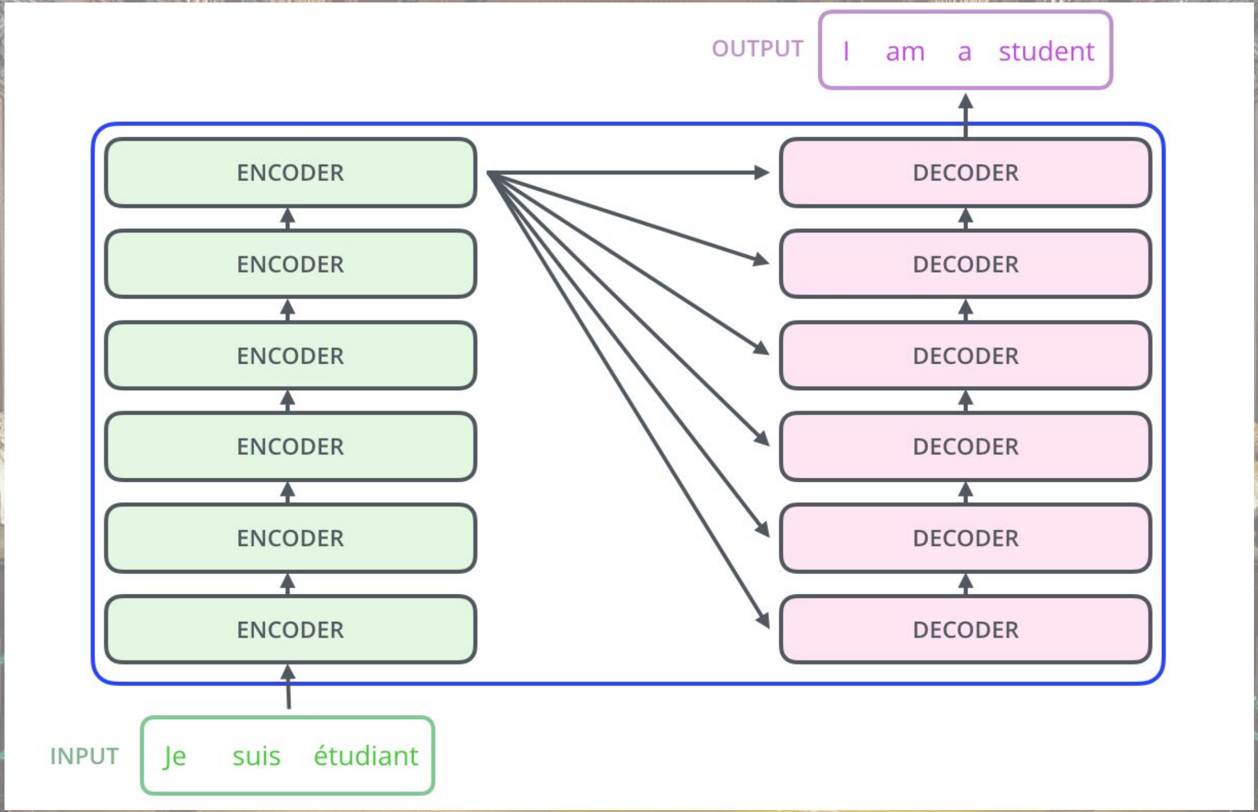
Country-Capital

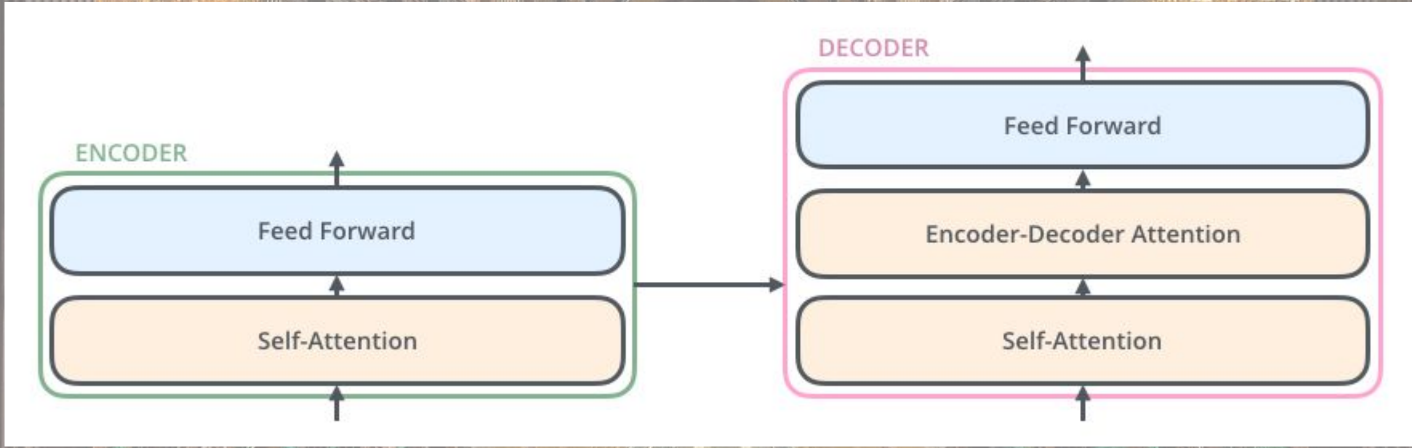
Long Short-Term Memory

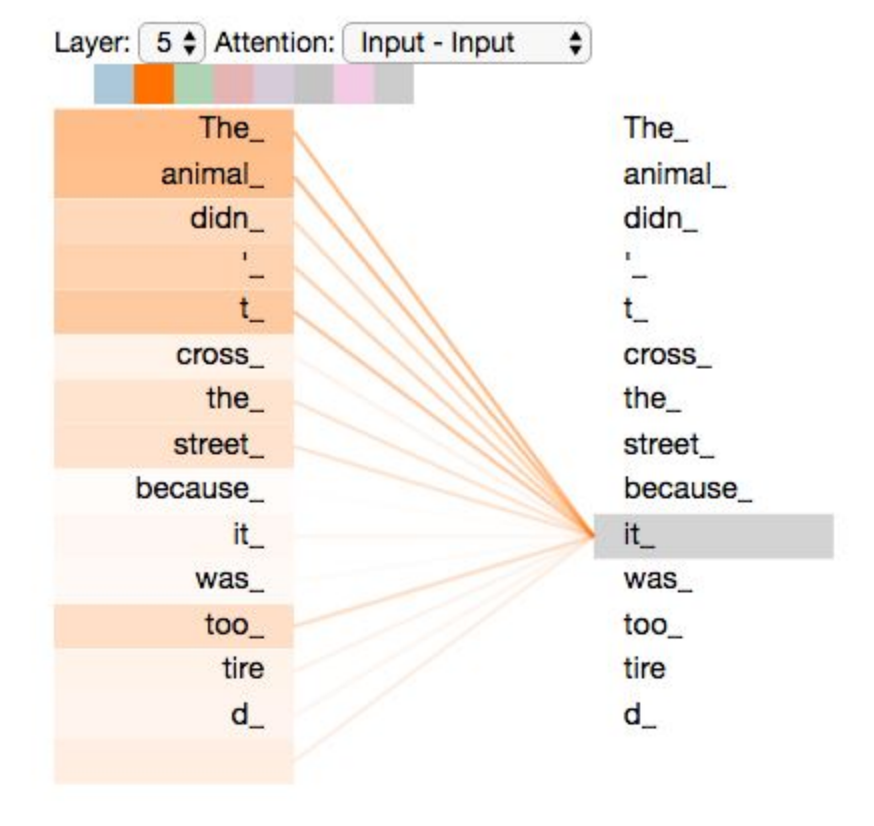


h t-1
h t
h t+1
A
A
σ
σ
tanh
σ
tanh
x
+
x
x
X t-1
X t
X t+1

Transformer


OUTPUT
I am a student
DECODER
DECODER
DECODER
DECODER
DECODER
DECODER
ENCODER
ENCODER
ENCODER
ENCODER
ENCODER
ENCODER
INPUT
Je suis étudiant

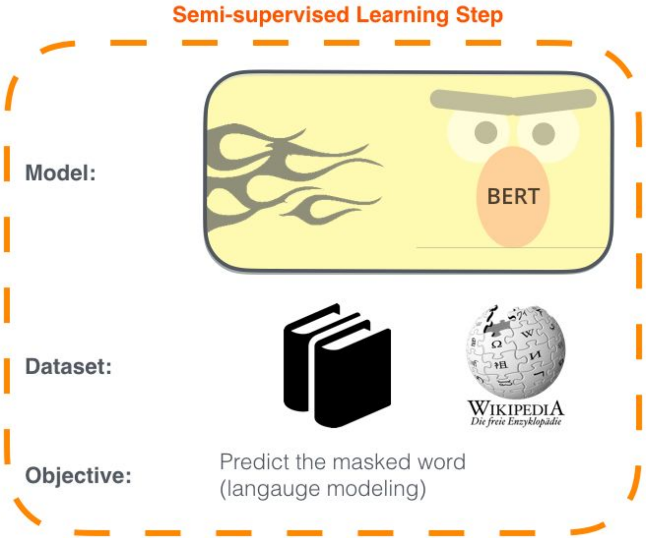

ENCODER
Feed Forward
Self-Attention
DECODER
Feed Forward
Encoder-Decoder Attention
Self-Attention


Layer: 5
Attention: Input - Input
The_
animal_
didn_
'_
t_
cross_
the_
street_
because_
it_
was_
too_
tire
d_
The_
animal_
didn_
'_
t_
cross_
the_
street_
because_
it_
was_
too_
tire
d_

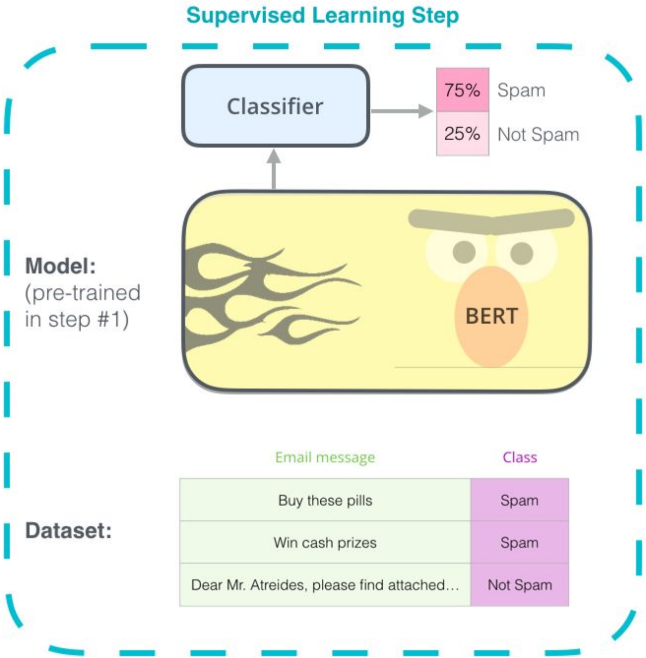
Semi-supervised Sequence Learning

1 - Semi-supervised training on large amounts of text (books, wikipedia..etc).

The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.

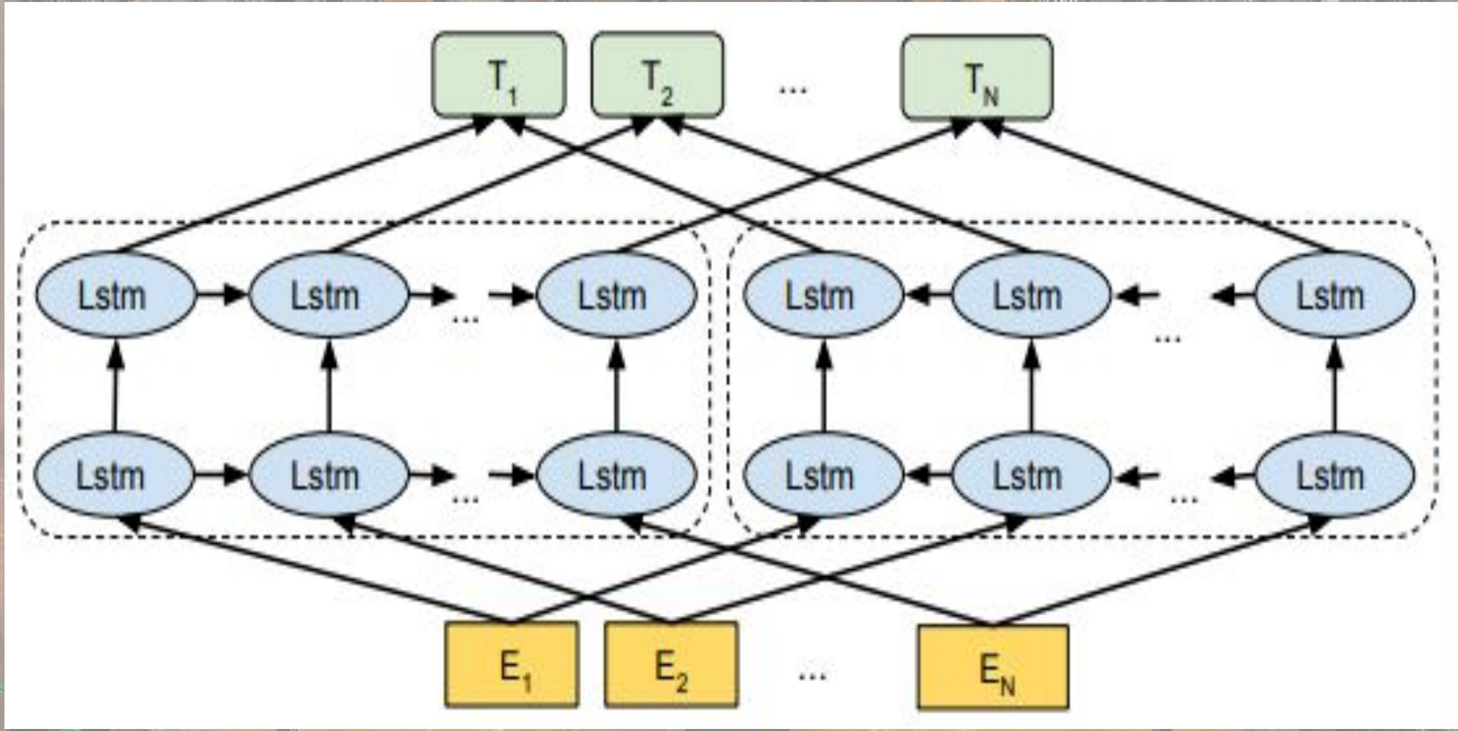


2 - Supervised training on a specific task with a labeled dataset.



Email message	Class
Buy these pills	Spam
Win cash prizes	Spam
Dear Mr. Atreides, please find attached…	Not Spam

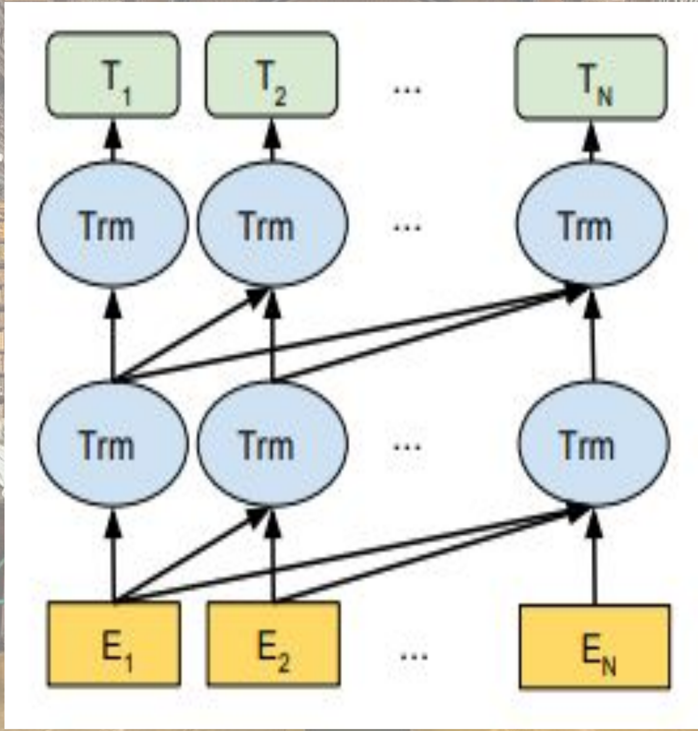
ELMo



T1
T2
...
TN
Lstm
Lstm
Lstm
Lstm
Lstm
Lstm
...
...
Lstm
Lstm
Lstm
Lstm
Lstm
Lstm
...
...
E1
E2
...
EN

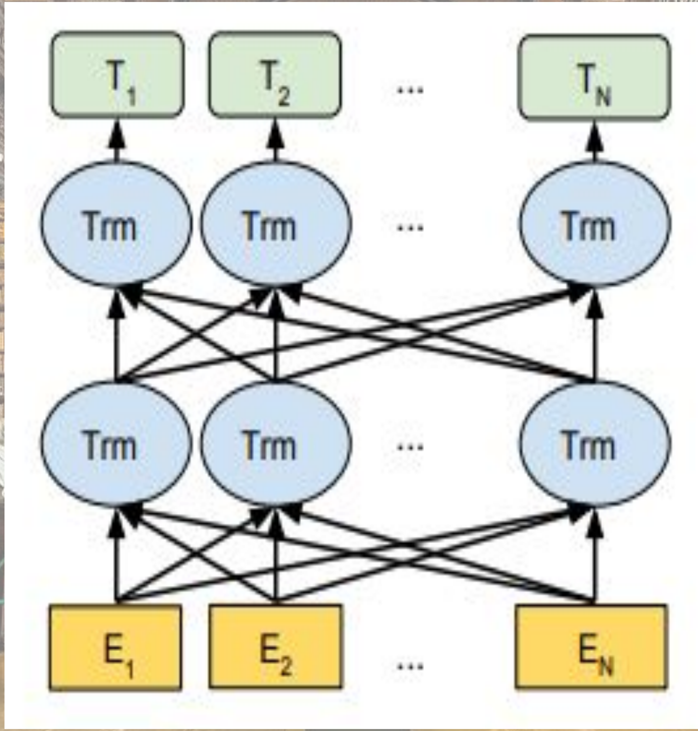
ULMFiT

OpenAI Transformer



T1
T2
...
TN
Trm
Trm
...
Trm
Trm
Trm
...
Trm
E1
E2
...
EN

BERT


T1
T2
...
TN
Trm
Trm
...
Trm
Trm
Trm
...
Trm
E1
E2
...
EN

Input = [CLS] the man went to [MASK] store [SEP]

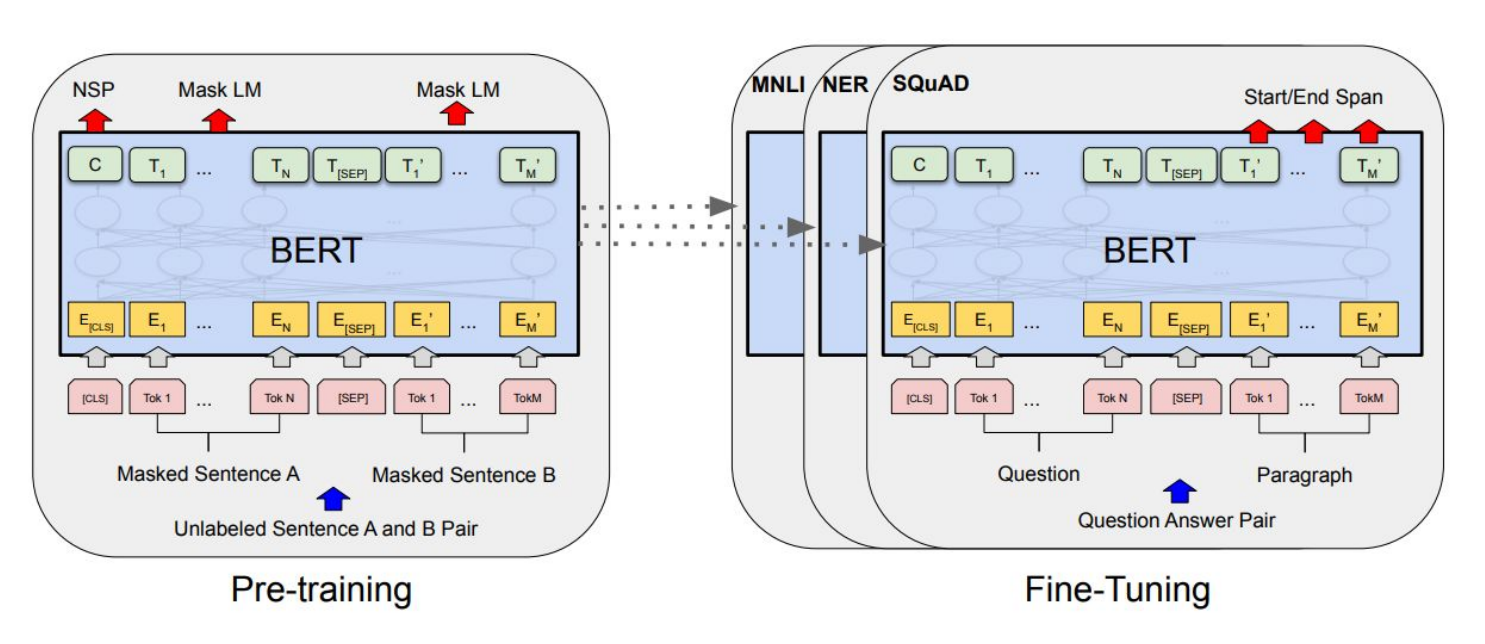
he bought a gallon [MASK] milk [SEP]

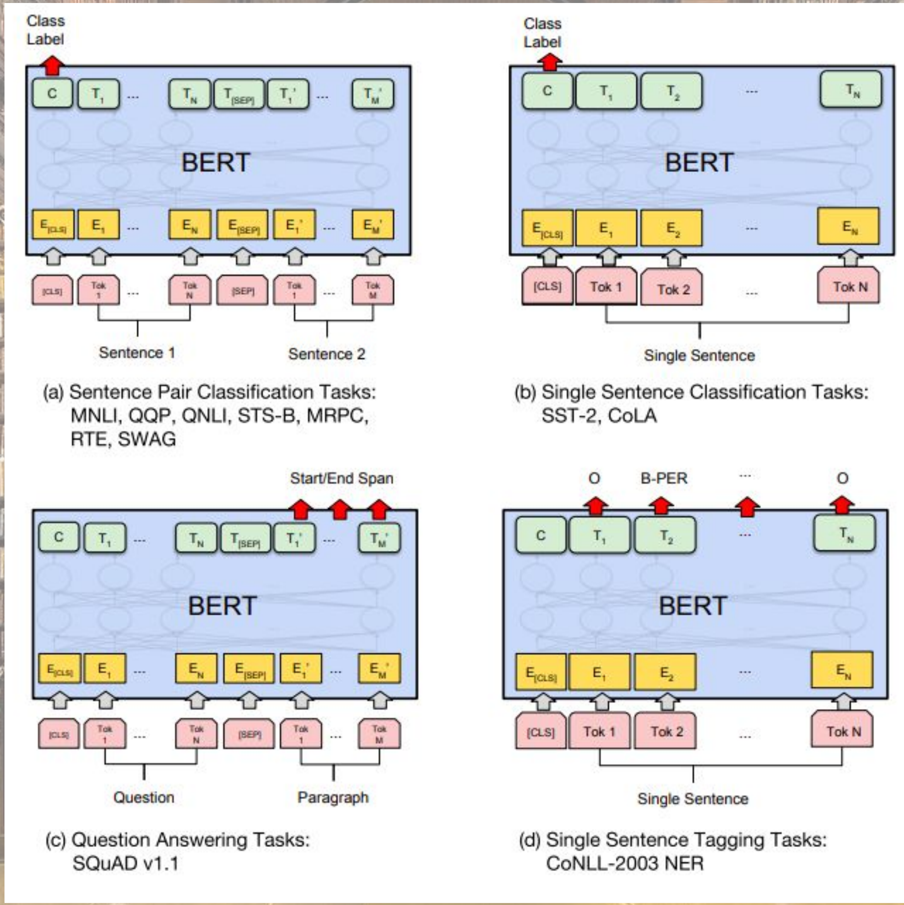
Label = IsNext

Input = [CLS] the man [MASK] to the store [SEP]

penguin [MASK] are flight ##less birds [SEP]

Label = NotNext


NSP
Mask LM
Mask LM
C
T_1
T_N
$T_{[SEP]}$
T_1'
T_M'
BERT
$E_{[CLS]}$
E_1
E_N
$E_{[SEP]}$
E_1'
E_M'
[CLS]
Tok 1
Tok N
[SEP]
Tok 1
TokM
Masked Sentence A
Masked Sentence B
Unlabeled Sentence A and B Pair
Pre-training
MNLI
NER
SQuAD
Start/End Span
C
T_1
T_N
$T_{[SEP]}$
T_1'
T_M'
BERT
$E_{[CLS]}$
E_1
E_N
$E_{[SEP]}$
E_1'
E_M'
[CLS]
Tok 1
Tok N
[SEP]
Tok 1
TokM
Question
Paragraph
Question Answer Pair
Fine-Tuning


Class Label
BERT
Sentence 1
Sentence 2
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG
Class Label
BERT
[CLS]
Tok 1
Tok 2
Tok N
Single Sentence
(b) Single Sentence Classification Tasks: SST-2, CoLA
Start/End Span
BERT
Question
Paragraph
(c) Question Answering Tasks: SQuAD v1.1
O
B-PER
O
BERT
[CLS]
Tok 1
Tok 2
Tok N
Single Sentence
(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	**Average**
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

Table 1: GLUE Test results, scored by the evaluation server (https://gluebenchmark.com/leaderboard). The number below each task denotes the number of training examples. The "Average" column is slightly different than the official GLUE score, since we exclude the problematic WNLI set.[8] BERT and OpenAI GPT are single-model, single task. F1 scores are reported for QQP and MRPC, Spearman correlations are reported for STS-B, and accuracy scores are reported for the other tasks. We exclude entries that use BERT as one of their components.

BERT End to End (Fine-tuning + Predicting) with Cloud TPU

Predicting Movie Reviews with BERT on TF Hub

The Illustrated Word2vec - Jay Alammar

Understanding LSTM Networks - Christopher Olah

The Illustrated Transformer - Jay Alammar

The Illustrated BERT, ELMo, and co. - Jay Alammar

CI&T NO TDCSP

Pedro Lelis

Moved by curiosity, I'm an entrepreneurial scientist who seeks to improve people's lives through Artificial Intelligence solutions.

Main researches and interesses: Fairness in Machine Learning, Few-shot Learning, Transformer (Attention), Generative Adversarial Network, Reinforcement Learning and Memory-Augmented Neural Network.

https://forms.gle/eWuLqj3yoQNg15SC8

<PEOPLE>
<COMMUNITIES>
<KNOWLEDGE>

All together_