



Search Medium

Write



So, Let's BERT!



Gal Hever · Follow

4 min read · Aug 23, 2020



3



Introduction

GPT is a fine-tunable pre-trained model that is based on the Transformer, but this Transformer was trained only on a forward language model. ELMo's language model was bi-directional but trained with LSTM. BERT that was proposed by Devlin et al., at 2018, compared to the previous models contains both; it has a Transformer-based model whose language model looks both forward and backwards and it uses Transformers instead of RNNs to process text and combines context from both directions.

What is BERT?

Bidirectional Encoder Representations from Transformers (BERT) is a self-supervised approach for pre-training a deep transformer encoder. BERT compared to the previous models proposes a transformer-based model whose language model looks both forward and backwards and also enjoys from the transformer's benefits. Each layer in BERT applies self-attention, passes its results through a feed-forward network, and then hands it off to the next encoder that learns a representation for each token.

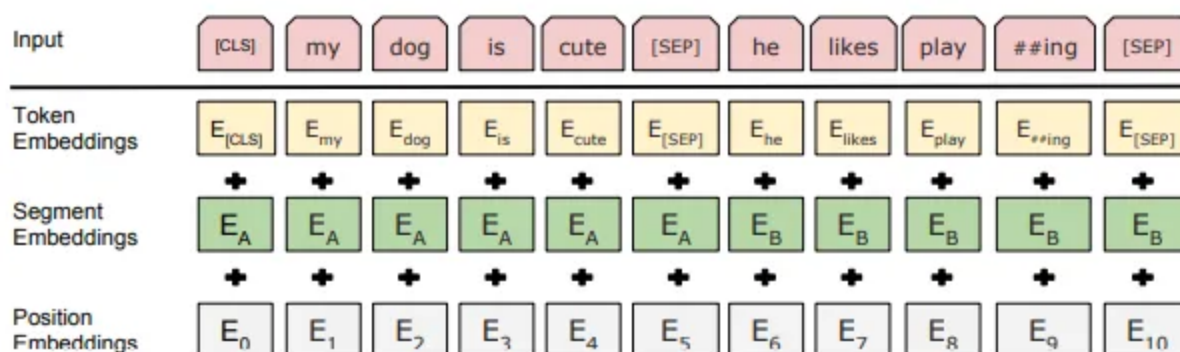
So, How does it work?

Model's input

Given a sequence of tokens $X = (x_1, x_2, \dots, x_n)$, BERT pads the input sentence with [CLS] and [SEP] tokens. Then it trains an encoder that produces a contextualized vector representation for each token: $\text{encoder}(x_1, x_2, \dots, x_n) = x_1, x_2, \dots, x_n$. Finally, the representation of each token in the

sequence is constructed by summing the corresponding token, segment, and position embeddings.

The [SEP] token indicates when the next sentence starts for the NSP task. The [CLS] token is added to sequence A and sequence B to form the input, where the target of [CLS] is whether sequence B indeed follows sequence A in the corpus. [CLS] stands for classification tasks and it supposes to summarize the sentence and the output of the last layer of this token will be used for the classification step.



Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

In the end of the pre-processing level the input will be composed of:

- Token Embedding — The original sequence that was padded with [CLS] and [SEP].
- Segment Embedding — Indicates the sequence that the token belongs to.
- Position Embedding — Indicates the order of the tokens in the sequence.

Training the Model

BERT pre-trains the model parameters by two tasks, the masked language model (MLM) and the next sentence prediction (NSP).

MLM — In this task the the 15% of the input's tokens are substituted randomly. Of those, 80% are replaced with [MASK], 10% are replaced with a random token, and 10% are kept unchanged. The task is to predict the original tokens from the modified input according to its context.

NSP — In this task the mission is to predict whether two sequences are following or not.

The loss is trained on two tasks on the same time: Masked LM (MLM) + Next sentence prediction (NSP). The learning rate is warmed up over the first 10,000 steps to a peak value of $1e-4$, and then linearly decayed.

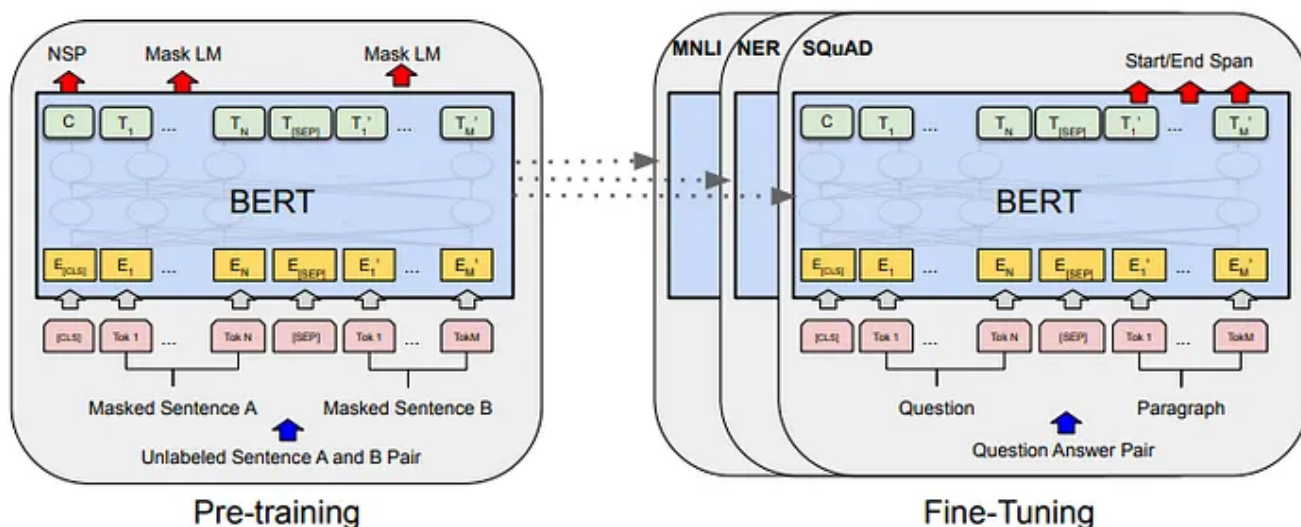
Different ways to use BERT

There are two ways to use the model:

1. Pre-training approach — to create contextualized word embeddings.
2. Fine-tuning approach- fine-tune the pre-trained model and then feed these embeddings to existing model.

There are two sizes for the model:

- BERT BASE — 12 encoder layers
- BERT LARGE — 24 encoder layers



Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

Disadvantages of the Model

- **No relationship between masked words** — A masked token that the model used for prediction will not be available for another prediction.
- **MLM is not a real task** — The training is not useful in the real life.
- **Maximum sequence length is limited to 512 tokens** — Can't deal with really long sequence (e.g; a full book).
- **Not an auto-regressive model** — Just 15% of the data is used for the training set.
- **Can't deal with span** — BERT model have achieved high performance on supervised dataset that masks individual tokens. However, when it engaged with tasks that involved reasoning about relationships between spans of text such as question answering it was more challenging target.
- **Can't deal with sentence generation**

End Notes

Language model pre-training method has been found to be effective for solving many NLP tasks. In particular the pretrained BERT gained lately huge leverage and obtains new state-of-the-art results on eleven tasks. The tasks that BERT has been applied to are typically modeled as classification problems and sequence labeling tasks except of the SQuAD question answering (Rajpurkar et al., 2016) task, in which the objective is to find the starting point and ending point of an answer span.

References

[<https://arxiv.org/abs/1810.04805>]

The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning)

Discussions: Hacker News (98 points, 19 comments), Reddit r/MachineLearning (164 points, 20 comments) Translations...

[jalammar.github.io](https://github.com/jalammar)

More from the list: "NLP"

Curated by Himanshu Birla



Jon Gi... in Towards Data ...

Characteristics of Word Embeddings

★ · 11 min read · Sep 4, 2021



Jon Gi... in Towards Data ...

The Word2vec Hyperparameters

★ · 6 min read · Sep 3, 2021



Jon Gi... in

The Word2vec

★ · 15 min read

[View list](#)**Written by Gal Hever**

108 Followers

Data Scientist

[Follow](#)**More from Gal Hever**

Gal Hever

Getting Started with NVIDIA NeMo ASR

NVIDIA NeMo—Quick Start Guide



Gal Hever

Sentiment Analysis with Pytorch—Part 4—LSTM\BiLSTM Model

Introduction

3 min read · Apr 20, 2021

8 min read · Apr 11, 2020



89



74

2



Gal Hever

Coreference Resolution Models

A Review of the Latest Models

12 min read · Aug 10, 2020



17



59

2

[See all from Gal Hever](#)

Gal Hever


Sentiment Analysis with Pytorch— Part 1—Data Preprocessing

Introduction

8 min read · Apr 8, 2020

Recommended from Medium

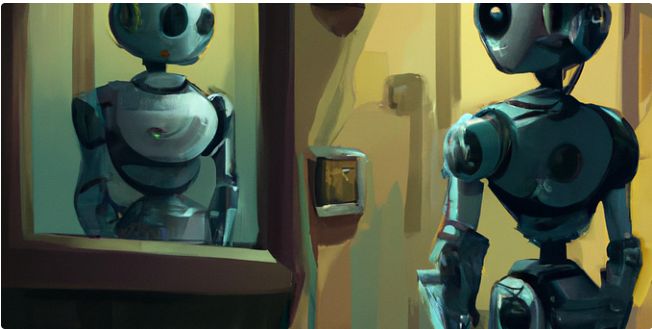



 Tomas Vykruta

Understanding Causal LLM's, Masked LLM's, and Seq2Seq: A...

In the world of natural language processing (NLP), choosing the right training approach i...

7 min read · Apr 30



 Thomas van Dongen in Towards Data Science

Demystifying efficient self-attention

A practical overview

20 min read · Nov 7, 2022

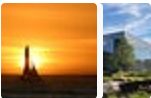


Lists



Staff Picks

465 stories · 317 saves



Stories to Help You Level-Up at Work

19 stories · 235 saves



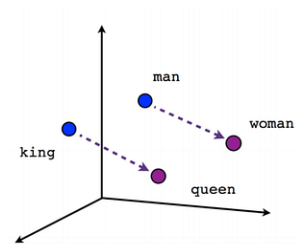
Self-Improvement 101

20 stories · 643 saves

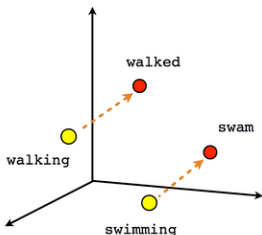


Productivity 101

20 stories · 597 saves



Male-Female



Verb tense

[CLS]	my	dog	is	cute	[SEP]	he	likes	play	#
$E_{[CLS]}$	E_{my}	E_{dog}	E_{is}	E_{cute}	$E_{[SEP]}$	E_{he}	E_{likes}	E_{play}	$E_{\#}$
+	+	+	+	+	+	+	+	+	+
E_A	E_A	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B
+	+	+	+	+	+	+	+	+	+
E_0	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9



Maninder Singh

Accelerate Your Text Data Analysis with Custom BERT Word...

One thing is for sure the way humans interact with each other naturally is one of the most...

4 min read · Apr 24



156



Zain ul Abideen

A Comparative Analysis of LLMs like BERT, BART, and T5

Exploring Language Models

6 min read · Jun 26



20



1



Michael Humor in CoinsBench

What are the LLaMA model weights?

The LLaMA models released by Meta AI were trained with different transformer...

4 min read · May 3



3



1



Avinash Patil

Embeddings: BERT better than ChatGPT4?

In this study, we compared the effectiveness of semantic textual similarity methods for...

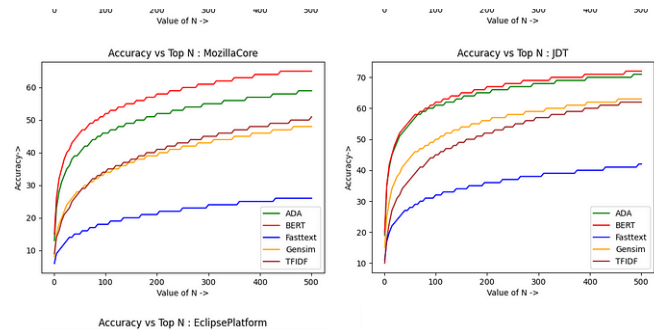
4 min read · Sep 19



3



1



See more recommendations