What is BERT? An Intro to BERT Models

Discover the inner workings of BERT, one of the first and most successful LLMs.

Contents

Nov 2, 2023 · 11 min read

Javier Canales Luna
Data Science | Climate Activist | Writer | Environmental Law and Policy | Education

TOPICS

Artificial Intelligence

Scientific breakthroughs rarely take place in a vacuum. Instead, they are often the penultimate step of a staircase built on accumulated human knowledge. To understand the success of large language models (LLMs), such as ChatGPT and Google Bart, we need to go back in time and talk about BERT.

Developed in 2018 by Google researchers, BERT is one of the first LLMs. With its astonishing results, it rapidly became a ubiquitous baseline in NLP tasks, including general language understanding, question & answer, and named entity recognition.

Interested in learning more about LLMs? Start Chapter 1 of our Large Language Models (LLMs) Concepts course today.

It's fair to say that BERT paved the way for the generative AI revolution we are witnessing these days. Despite being one of the first LLMs, BERT is still widely used, with thousands of open-source, free, and pre-trained BERT models available for specific use cases, such as sentiment analysis, clinical note analysis, and toxic comment detection.

Curious about BERT? Keep reading the article, where we will explore the architecture of Ber, the inner workings of the technology, some of its real-world applications, and its limitations.

What is BERT?

BERT (standing for Bidirectional Encoder Representations from Transformers) is an open-source model developed by Google in 2018. It was an ambitious experiment to test the performance of the so-called **transformer** –an innovative neural architecture presented by Google researchers in the famous paper **Attention is All You Need** in 2017– on natural language (NLP) tasks.

The key to the success of BERT is its transformer architecture. Before transformers kicked in, modeling natural language was a very challenging task. Despite the rise of sophisticated neural networks –namely, **recurrent** or **convolutional** neural networks– the results were only partly successful.

The main challenge lies in the mechanism neural networks used to predict the missing word in a sentence. At that time, state-of-the-art neural networks relied on the **encoder-decoder architecture**, a powerful yet time-and-resource-consuming mechanism that is unsuitable for parallel computing.

With these challenges in mind, Google researchers developed the transformer, an innovative neural architecture based on the attention mechanism, as explained in the following section.

How Does BERT Work?

Let's take a look at how BERT works, covering the technology behind the model, how it's trained, and how it processes data.

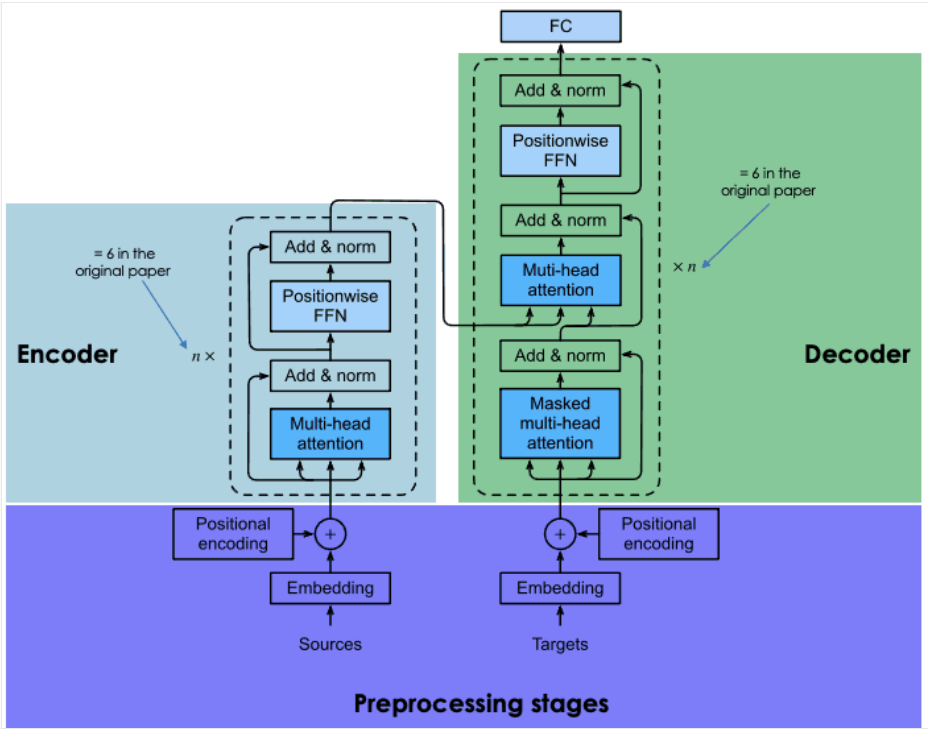
Core architecture and functionality

Recurrent and convolutional neural networks use sequential computation to generate predictions. That is, they can predict which word will follow a sequence of given words once trained on huge datasets. In that sense, they were considered unidirectional or context-free algorithms.

By contrast, transformer-powered models like BERT, which are also based on the encoder-decoder architecture, are bidirectional because they predict words based on the previous words and the following words. This is achieved through the self-attention mechanism, a layer that is incorporated in both the encoder and the decoder. The goal of the attention layer is to capture the contextual relationships existing between different words in the input sentence.

Nowadays, there are many versions of pre-trained BERT, but in the original paper, Google trained two versions of BERT: BERTbase and BERTlarge with different neural architectures. In essence, BERTbase was developed with 12 transformer layers, 12 attention layers, and 110 million parameters, while BERTlarge used 24 transformer layers, 16 attention layers, and 340 million parameters. As expected, BERTlarge outperformed its smaller brother in accuracy tests.

To know in detail how the encoder-decoder architecture works in transformers, we highly recommend you to read our **Introduction to Using Transformers and Hugging Face.**



An explanation of the architecture of transformers

Pre-training and fine-tuning

Transformers are trained from scratch on a huge corpus of data, following a time-consuming and expensive process (that only a limited group of companies, including Google, can afford).

In the case of BERT, it was pre-trained over four days on Wikipedia (~2.5B words) and Google's BooksCorpus (~800M words). This allows the model to acquire knowledge not only in English but also in many other languages from around the world.

To optimize the training process, Google developed new hardware, the so-called TPU (Tensor Processing Unit), specifically designed for machine learning tasks.

To avoid unnecessary and costly interactions in the training process, Google researchers used **transfer learning** techniques to separate the (pre)training phase from the fine-tuning

phase. This allows developers to choose pre-trained models, refine the input-output pair data of the target task, and retrain the head of the pre-trained model by using domain-specific data. This feature is what makes LLMs like BERT the **foundation model** of endless applications built on top of them,

The role of Masked Language Modelling in BERT's processing

The key element to achieving bidirectional learning in BERT (and every LLM based on transformers) is the attention mechanism. This mechanism is based on masked language modeling (MLM). By masking a word in a sentence, this technique forces the model to analyze the remaining words in both directions in the sentence to increase the chances of predicting the masked word. MLM is based on techniques already tried in the field of computer vision, and it's great for tasks that require a good contextual understanding of an entire sequence.

BERT was the first LLM to apply this technique. In particular, a random 15% of the tokenized words were masked during training. The result shows that BERT could predict the hidden words with high accuracy.

Curious about masked language modeling? Check our **Large Language Models (LLMs) Concepts Course** to learn all the details about this innovative technique.

What is BERT Used for? BERT's Impact on NLP

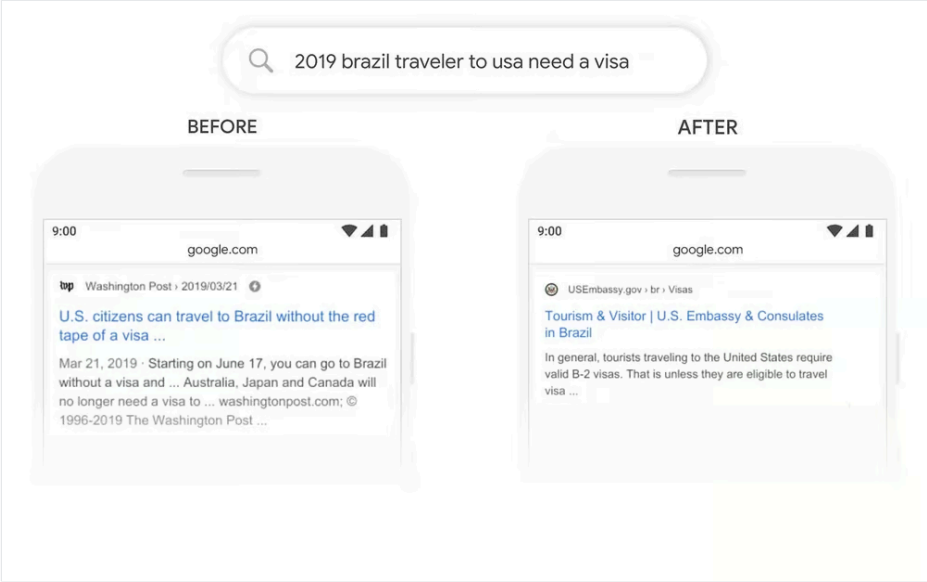
Powered by transformers, BERT was able to achieve state-of-the-art results in multiple NLP tasks. Here are some of the tests where BERT excels:

- **Question answering.** BERT has been one of the first transformer-powered chatbots, delivering impressive results.
- **Sentiment analysis.** For example, BERT has been successful in predicting positive or negative punctuation for movie reviews.
- **Text generation.** A precursor of next-generation chatbots, BERT was already able to create long texts with simple prompts.
- **Summarizing text.** Equally, BERT was able to read and summarize texts from complex domains, including law and healthcare.
- **Language translation.** BERT has been trained on data written in multiple languages. That makes it a multilingual model, which translates into great suitability for language translation.
- **Autocomplete tasks.** BERT can be used for autocomplete tasks, for example, in emails or messaging services.

Real-World Applications of BERT

Many LLMs have been tried in experimental sets, but not many have been incorporated into well-established applications. This is not the case with BERT, which is used every day by millions of people (even though we may not be aware of that).

A great example is Google Search. In 2020, Google announced that it had adopted BERT through Google Search in over 70 languages. That means that Google uses BERT to rank content and display featured snippets. With the attention mechanism, Google can now use the context of your question to provide useful information, as shown in the following example.



Source: Google

BERT's variants and adaptations

But this is only a part of the story. The success of BERT is greatly due to its open-source nature, which has allowed developers to access the source code of the original BERT and create new features and improvements.

This has resulted in a good number of BERT's variants. Below, you can find some of the most well-known variants:

- **RoBERTa.** Short for "Robustly Optimized BERT Approach", RoBERTa is a BERT variant created by Meta in collaboration with Washington University. Considered a more powerful version than the original BERT, RoBERTa was trained with a dataset 10 times bigger than the one used to train BERT. As for its architecture, the most significant difference is the use of dynamic masking learning instead of static masking learning. This technique, which involved duplicating training data and masking it 10 times, each time with a different mask strategy, allowed RoBERTa to learn more robust and generalizable representations of words.
- **DistilBERT.** Since the launch of the first LLMs in the late 2010s, there has been a consolidated trend of building bigger and heavier LLMs. This makes sense, as there seems to be a direct relationship between model size and model accuracy. Yet, it's also true that the bigger the model, the more resources it requires to run, and hence, fewer people can afford to use it. DistilBERT aims at making BERT more accessible by offering a smaller, faster, cheaper, and lighter variant. Based on the architecture of the original BERT, DistilBERT uses knowledge distillation techniques during pre-training to reduce the size by 40% while retaining 97% of its language understanding capabilities and being 60% faster.
- **ALBERT.** Stands for A Lite BERT, ALBERT was specifically designed to increase the efficiency of BERTlarge during pre-training. As training bigger models often results in memory limitations, longer training times, and unexpected model degradation, ALBERT creators developed two parameter-reduction techniques to reduce memory consulting and increase speed during training.

If you want to know more about the open-source LLM movement, we highly recommend you to read our post with the Top Open-Source LLM in 2023

Fine-tuning BERT for specific tasks

One of the greatest things about BERT, and LLMs in general, is that the pre-training process is separated from the fine-tuning process. That means that developers can take pre-trained versions of BERT, and customize them for their specific use cases.

In the case of BERT, there are hundreds of fine-tuned versions of BERT developed for a wide diversity of NLP tasks. Below, you can find a very, very limited list of fine-tuned versions of BERT:

- **BERT-base-chinese.** A version of BERTbase trained for NLP tasks in Chinese

- **BERT-base-NER.** A version of BERTbase customized for named entity recognition
- **Symps_disease_bert_v3_c41.** A symptom-to-disease classification model for a natural language chatbot.
- **BERT for Patents.** is a model trained by Google on 100M+ patents worldwide. It is based on BERTlarge.

Understanding BERT's Limitations

BERT comes with the traditional limitations and problems associated with LLMs. The predictions of BERT are always based on the quantity and quality of data used for training it. If training data is limited, poor, and biased, BERT may throw inaccurate, harmful results or even so-called LLM hallucinations.

In the case of the original BERT, this is even more likely, as the model was trained without Reinforcement Learning from Human Feedback (RLHF), a standard technique used by more advanced models, like ChatGPT, LLaMA 2, and Google Bard, to enhance AI safety. RLHF involves using human feedback to monitor and steer the learning process of the LLM during training, thereby ensuring effective, safer, and trustful systems.

Furthermore, although it can be considered a small model compared to other state-of-the-art LLMs, like ChatGPT, it still requires a considerable amount of computing power to run it, let alone train it from scratch. Therefore, developers with limited resources may not be able to use it.

The Future of BERT and NLP

BERT was one of the first modern LLMs. But far from old-fashioned, BERT is still one of the most successful and widely used LLMs. Thanks to its open-source nature, today, there are multiple variants and hundreds of pre-trained versions of BERT designed for specific NLP tasks.

If you're interested in keeping up with BERT and recent NLP developments, DataCamp is here to help. Check out our curated materials and stay tuned on the current generative AI revolution!

- Natural Language Processing Tutorial
- Understanding Text Classification in Python
- An Introduction to Using Transformers and Hugging Face
- What is Deep Learning? A Tutorial for Beginners
- Deep Learning with PyTorch Course
- Mastering Natural Language Processing (NLP) with PyTorch: Comprehensive Guide

AUTHOR
Javier Canales Luna

I am a freelance data analyst, collaborating with companies and organisations worldwide in data science projects. I am also a data science instructor with 2+ experience. I regularly write data-science-related articles in English and Spanish, some of which have been published on established websites such as DataCamp, Towards Data Science and Analytics Vidhya As a data scientist with a background in political science and law, my goal is to work at the interplay of public policy, law and technology, leveraging the power of ideas to advance innovative solutions and narratives that can help us address urgent challenges, namely the climate crisis. I consider myself a self-taught person, a constant learner, and a firm supporter of multidisciplinary. It is never too late to learn new things.

TOPICS

Artificial Intelligence

Start Your NLP Journey Today!

TRACK

Natural Language Processing in Python

0 min

Learn how to transcribe, and extract exciting insights from books, review sites, and online articles with Natural Language Processing (NLP) in Python.

See Details →

Start Course

See More →

Related

BLOG

What is an LLM? A Guide on Large Language Models and...

BLOG

Understanding and Mitigating Bias in Large Language Model...

BLOG

Exploring BLOOM: A Comprehensive Guide to the...

See More →

LEARN

Learn Python