

BERT: A Practical Overview

BERT (Bidirectional Encoder Representations from Transformers) is a language model architecture introduced by Google researchers in 2018. This document is an original summary of how it works and why it mattered, not a reproduction of the original paper.

What problem it solved

Before BERT, many popular language models read text in one direction only (left-to-right or right-to-left), which limited how much context they could use to understand a word. BERT was trained to consider context from both directions at once, giving it a richer understanding of word meaning based on everything around it.

Core architecture

- Built entirely from Transformer encoder layers (no decoder).
- Uses self-attention so every token can attend to every other token in the input.
- Comes in different sizes, most commonly a 12-layer 'base' version and a 24-layer 'large' version.

How it was pretrained

- Masked Language Modeling: random words in a sentence are hidden, and the model learns to predict them from context.
- Next Sentence Prediction: the model learns whether one sentence plausibly follows another.
- Both tasks were trained on large collections of unlabeled text, so no manual labeling was required at this stage.

Why it mattered

After pretraining, BERT could be fine-tuned on a specific task, such as question answering, sentiment analysis, or named entity recognition, usually with a small amount of labeled data and a lightweight extra layer on top. This 'pretrain then fine-tune' approach became the standard recipe for a large share of natural language processing work in the years that followed, and it directly influenced later encoder and encoder-decoder models.

Common use cases

- Search relevance and query understanding.
- Text classification, such as sentiment or topic labeling.
- Extractive question answering.
- Named entity recognition and token-level tagging.

Limitations

BERT is an encoder-only model, so it isn't naturally suited to open-ended text generation the way decoder-based models are. It also has a fixed, fairly short maximum input length compared to more recent architectures, and larger versions require meaningful compute to run at scale.