Transformers

Transformers are a type of neural network architecture that have revolutionized Natural Language Processing (NLP) tasks. Unlike traditional algorithms, transformers use attention mechanisms to focus on the most important parts of an input sequence, weighing the relevance of each element. This allows them to understand context and meaning, even in languages with complex grammatical structures.

Key Components and Concepts :

- **Attention Mechanism :** This mechanism allows the model to prioritize the most relevant parts of the input data. It assigns weights to each element in the sequence based on its importance to other elements. Multi-head attention allows the model to attend to different parts of the input simultaneously, enhancing its ability to understand nuanced meanings.
- **Encoder-Decoder Architecture :** The encoder processes the input sequence and produces a continuous representation, or embedding, of the input, which is then passed to the decoder to generate the output sequence2. The encoder comprises multiple self-attention and feed-forward layers, allowing the model to effectively process and understand the input sequence. The decoder uses the embeddings produced by the encoder and its internal states to generate the output sequence.
- **Self-Attention :** Allows the model to focus on relevant parts of the input sequence, capturing long-range dependencies effectively.

Use Cases :

- **Machine Translation :** Transformers provide a robust framework for translating text between languages with remarkable accuracy.
- **Text Summarization :** They generate concise summaries of long texts while retaining essential information.
- **Question Answering :** Transformers provide accurate answers to user queries based on context.
- **Sentiment Analysis :** They are used to understand the sentiment behind a piece of text, which is crucial for businesses and social media monitoring.
- *Image generation, speech recognition, recommender systems, chatbots*

BERT / GPT

BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) are two influential models in natural language processing (NLP), both built upon the Transformer architecture. Despite their shared foundation, they differ in design, training objectives, and applications.

BERT Overview

Introduced by Google in 2018, BERT is an encoder-only model that processes text bidirectionally, meaning it considers context from both preceding and following words simultaneously. This approach enables BERT to grasp the full context of a word based on its surroundings. BERT is pre-trained using two primary tasks:

- *Masked Language Modeling (MLM) :* Randomly masks words in a sentence and trains the model to predict the missing words based on context.
- *Next Sentence Prediction (NSP) :* Trains the model to understand the relationship between two sentences by predicting if one sentence logically follows another.

These training methods allow BERT to generate deep contextual representations, making it particularly effective for tasks that require understanding the intricacies of language, such as sentiment analysis, named entity recognition, and question answering.

GPT Overview

Developed by OpenAI, GPT is a series of decoder-only models designed for text generation. GPT processes text unidirectionally, typically from left to right, predicting the next word in a sequence based on the preceding context. The training involves:

- *Autoregressive Language Modeling :* Trains the model to predict the next word in a sentence, enabling it to generate coherent and contextually relevant text.

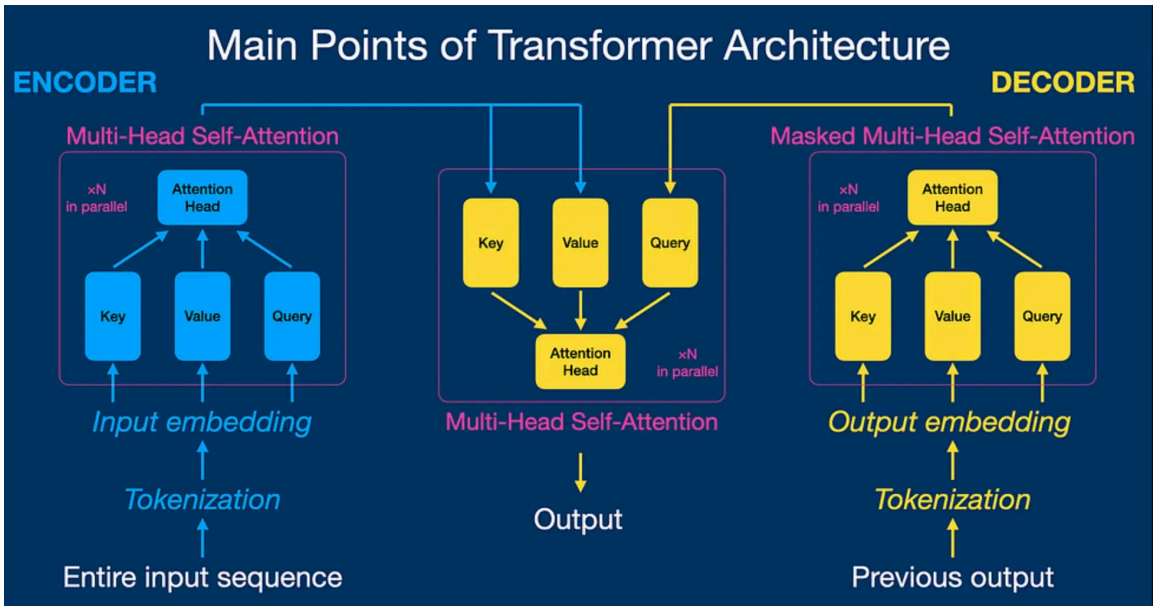
This training approach makes GPT adept at tasks like text completion, translation, and creative writing. The GPT series has evolved over time, with GPT-3 and GPT-4 being notable for their large scale and versatility in various NLP applications.

Key Differences

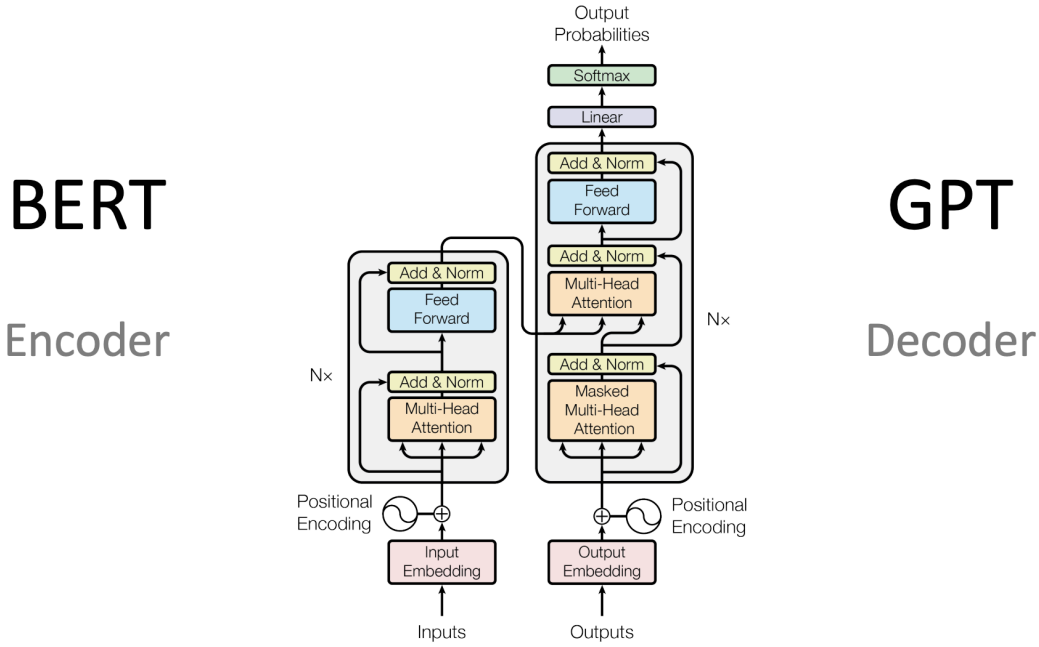
- *Architecture :* BERT utilizes an encoder-only structure, processing text bidirectionally, while GPT employs a decoder-only structure, processing text unidirectionally.
- *Training Objectives :* BERT focuses on understanding language through MLM and NSP tasks, whereas GPT is geared towards generating language via autoregressive modeling.
- *Applications :* BERT excels in tasks that require deep language comprehension, such as classification and understanding tasks. In contrast, GPT is more suited for tasks involving text generation, like drafting content or translating text.

Aspect	BERT	GPT
Architecture	Utilizes a bidirectional Transformer architecture, processing input text in both directions simultaneously. Employs an encoder-only architecture.	Employs a unidirectional Transformer architecture, processing text from left to right. Based on a decoder-only transformer model.
Training Objective	Trained using a masked language model (MLM) task, predicting masked words based on the surrounding context.	Trained using a causal language model (CLM) task, predicting the next word in a sequence.
Understanding	Captures context from both the left and right of a word, providing a comprehensive understanding of sentence structure and semantics.	Understands context only from the left of a word, which may limit its ability to fully grasp the relationships between words in some cases.
Use Cases	Excels at sentence and token-level classification tasks, sentiment analysis,	Suitable for tasks like text generation, translation, and conversational AI.

	and question answering. Extensions like sBERT are used for semantic search.	
Real-World Example	Used in Google Search to understand the context of search queries.	Employed to generate human-like text responses in applications like chatbots and content creation.
Training data size	Trained on 3 TB of data.	Trained on 45 TB of data.
Parameters	BERT (Large) totals 340 million parameters.	GPT-4 has 1.5 billion parameters. GPT-1 consists of 117 million parameters.
Directionality	Bidirectional, processes text in both directions simultaneously.	Unidirectional, processes text from left to right.



Transformer Architecture



BERT and GPT Architecture