Understanding Tokens in the Context of Large Language Models like BERT and T5

FHIRFLY · Follow
3 min read · Sep 6, 2023

Introduction

Tokens are fundamental building blocks in the world of Natural Language Processing (NLP) and Large Language Models (LLMs) like BERT, GPT, and T5. These tokens serve as the input units that these models read, process, and generate outputs from. This article aims to provide a detailed understanding of what tokens are, how they function within LLMs, and why they are crucial for the performance and capabilities of these models.

What is a Token?

In the context of LLMs, a token can be as small as a single character or as long as a word. In some sophisticated models, tokens can even represent entire phrases. Essentially, tokens are the pieces of text that language models read and analyze. They serve as the primary input data that these models use to perform tasks ranging from text classification to language generation.

Tokenization Process

Before an LLM can process text, it must be broken down into tokens through a process called tokenization. Different models have their ways of tokenizing text:

- **Word-based Tokenization**: Each word is a separate token.

- . Example: “ChatGPT is great!” becomes [‘ChatGPT’, ‘is’, ‘great’, ‘!’]
- **Subword Tokenization**: Words can be broken down into smaller units or subwords.
- . Example: In BERT, ‘ChatGPT’ might be broken down into [‘Chat’, ‘##G’, ‘##PT’]
- **Character-based Tokenization**: Each character is a token.
- . Example: ‘Chat’ becomes [‘C’, ‘h’, ‘a’, ‘t’]

Tokens in Transformer Models

BERT (Bidirectional Encoder Representations from Transformers)

BERT typically uses WordPiece tokenization. In this approach, common words or subwords are preserved, but less common words are broken down into subwords or individual characters. For example, the sentence “ChatGPT is great!” could be tokenized into tokens like `[‘[CLS]’, ‘Chat’, ‘##G’, ‘##PT’, ‘is’, ‘great’, ‘!’, ‘[SEP]’]`.

T5 (Text-To-Text Transfer Transformer)

T5 uses a variant of Byte Pair Encoding (BPE) for its tokenization. It aims to convert all NLP problems into a text-to-text format. The tokenization could look like `[‘Chat’, ‘G’, ‘PT’, ‘ is’, ‘ great’, ‘!’]`.

Why are Tokens Important?

Flexibility

Tokenization allows LLMs to understand and manipulate language. Subword tokenization allows models to deal with a broad range of words, even those not seen during training.

Efficiency

By breaking down text into smaller units, LLMs can process and analyze data much more efficiently.

Contextual Understanding

Tokens in models like BERT and T5 can capture contextual relationships between words, enabling nuanced language understanding and generation.

Conclusion

Understanding the concept of tokens is crucial for anyone looking to dive deep into the realm of Large Language Models. Tokens serve as the foundational units of data that enable these sophisticated models to read, understand, and generate human-like text. They play an integral role in the flexibility, efficiency, and capabilities of models like BERT and T5.

Artificial Intelligence | ChatGPT | Data Science | Large Language Models | Machine Learning

Follow

Written by FHIRFLY

74 Followers

SECURE. PRIVATE. AVAILABLE. CONFIDENTIAL. INTEGRAL. INTEROPERABLE. OUT OF THE DARKNESS COMES LIGHT.

More from FHIRFLY