large language model (LLM)

Authors

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland
*Corresponding author: alex.jung@aalto.fi

Abstract

The term large language model (LLM) refers to machine learning (ML) methods that analyze or generate text data, represented as sequences of tokens, using a high-dimensional model with billions of model parameters. Many current LLMs use a variant of a transformer that is trained via self-supervised learning. The training task is to predict words that are intentionally removed from a large text corpus, a construction that yields large training sets of labeled data points with little human supervision. A trained LLM maps an input sequence of tokens to a probability distribution over the next token. A prominent application is conversational artificial intelligence (AI), which generates human-like text for tasks ranging from answering questions to writing computer code.

Definition

An LLM is an umbrella term for machine learning (ML) methods that use high-dimensional ML models (with billions of model parameters) trained on large collections of text data. LLMs are used to analyze or generate sequences of tokens that constitute text data. Many current LLMs use some variant of a transformer that is trained via self-supervised learning, i.e., the training is based on the task of predicting a few words that are intentionally removed from a large text corpus. Thus, labeled data points can be constructed simply by selecting some words from a given text as labels and the remaining words as features of data points. This construction requires little human supervision and allows for sufficiently large training sets for LLMs (Brown et al., 2020; Devlin et al., 2019). A prominent application is conversational artificial intelligence (AI) which generates human-like text for tasks ranging from answering questions to writing and debugging computer code. Fig. 1 traces the stages an input sequence passes through.

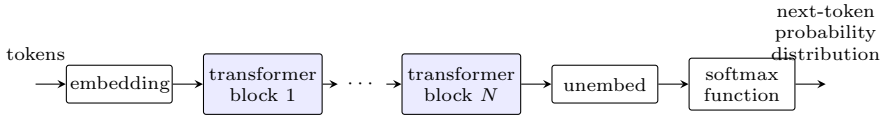

Figure 1: Signal-flow chart of an LLM. An input sequence of tokens is mapped to vectors by an embedding, refined by N stacked transformer blocks, mapped back to vocabulary scores by the unembedding, and turned into a probability distribution over the next token via softmax function

Cross References

token, transformer, NLP, foundation model, embedding, softmax function

References

1. Brown et al. (2020). Language Models are Few-Shot Learners. In: Adv. Neural Inf. Process. Syst., pp. 1877–1901.

2. Devlin et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proc. Conf. North American Chapter Assoc. Computational Linguistics (NAACL), pp. 4171–4186.