Large Language Models (LLMs) vs Transformers

Last Updated : 26 Aug, 2024

In recent years, advancements in artificial intelligence have led to the development of sophisticated models that are capable of understanding and generating human-like text. Two of the most significant innovations in this space are Large Language Models (LLMs) and Transformers. While they are often discussed together, they serve different purposes and operate under distinct principles.

This article will delve into the key differences between LLMs and Transformers, helping you understand how each contributes to the field of AI.

Table of Content

What Are Large Language Models (LLMs)?

Large Language Models (LLMs) are a type of AI model that has been trained on vast amounts of text data. These models are designed to understand, generate, and predict text. By analyzing patterns in the data, LLMs can generate coherent and contextually appropriate responses to text-based inputs.

Key Characteristics of LLMs:

Scale: LLMs are characterized by their large scale, with billions of parameters. The more parameters a model has, the more sophisticated it can be in understanding and generating language.

Training Data: LLMs are trained on diverse datasets that include books, articles, websites, and other text sources, enabling them to understand various topics and styles of writing.

Generalization: Due to their extensive training, LLMs can generalize across a wide range of topics, making them versatile tools for various applications, such as chatbots, content creation, and even coding assistance.

What Are Transformers?

Transformers are a type of neural network architecture that has revolutionized the field of natural language processing (NLP). Introduced in a 2017 paper titled "Attention is All You Need" by Vaswani et al., Transformers are designed to handle sequential data, such as text, by using a mechanism called self-attention.

Key Characteristics of Transformers:

Self-Attention Mechanism: The self-attention mechanism allows Transformers to weigh the importance of different words in a sentence when making predictions. This is crucial for understanding context, especially in long sentences.

Parallelization: Unlike traditional RNNs (Recurrent Neural Networks), which process data sequentially, Transformers can process multiple words at once, making them faster and more efficient.

Versatility: Transformers are not limited to language tasks; they can be applied to any problem involving sequential data, including tasks like image recognition and time-series forecasting.

LLMs vs. Transformers: A Comparative Analysis

1. Purpose of LLMs and Transformer

LLMs: Primarily focused on generating and understanding natural language, LLMs are built on various architectures, including Transformers.

Transformers: A neural network architecture used for various tasks, including but not limited to language modeling.

2. Architecture Design

LLMs: Can be based on different architectures, but many modern LLMs utilize the Transformer architecture to achieve state-of-the-art performance.

Transformers: A specific architecture that uses self-attention and is often employed as the backbone of LLMs.

3. Applications

LLMs: Used for a wide range of NLP tasks, from text generation and summarization to translation and sentiment analysis.

Transformers: Employed not just in NLP but also in other areas requiring the processing of sequential data, such as speech recognition and computer vision.

4. Training

LLMs: Require massive datasets and computational resources for training, often using Transformer architecture as a foundation.

Transformers: The architecture itself, which can be trained on various types of sequential data, is adaptable to different tasks.

5. Output

LLMs: Generate human-like text, making them suitable for applications that require natural language understanding and generation.

Transformers: Produce outputs depending on the task at hand, whether it be text, predictions, or other types of data sequences.

Summary Table

Here's a table summarizing the key differences between Large Language Models (LLMs) and Transformers:

Aspect	Large Language Models (LLMs)	Transformers

Purpose	Generate and understand natural language	Neural network architecture for sequential data
Architecture	Often based on the Transformer architecture	Uses self-attention and parallel processing
Applications	NLP tasks like text generation, summarization, etc.	NLP, image recognition, speech recognition, etc.
Training	Requires massive datasets and resources, often using Transformers	Adaptable to various tasks with different types of sequential data
Output	Human-like text	Task-dependent output (text, predictions, etc.)

When to Use Large Language Models (LLMs)?

1. Natural Language Generation:

Task: You need to generate human-like text, such as writing articles, generating dialogue for chatbots, or creating summaries.

Example: Generating content for blogs, answering customer queries in natural language.

2. Language Understanding:

Task: You require deep understanding of context and nuances in text, such as sentiment analysis, translation, or summarization.

Example: Analyzing social media posts to gauge public opinion.

3. Text Completion and Suggestion:

Task: You need to predict and complete text, such as auto-completion in text editors or code generation.

Example: Predicting the next word in a sentence or suggesting code snippets in an IDE.

4. Question Answering:

Task: You need to provide accurate answers to questions based on large bodies of text or data.

Example: Creating an AI assistant that answers queries about specific topics like legal advice or tech support.

When to Use Transformers?

1. Handling Sequential Data:

Advertisement: Your video starts in 24 seconds.

Task: You need to process data that has a sequential nature, such as time-series data, sequences of words, or sequences in music.

Example: Time-series forecasting for stock prices, processing DNA sequences in bioinformatics.

2. Building Custom Models:

Task: You need to create a custom neural network model for a specific task, not necessarily related to natural language, that requires handling sequences.

Example: Developing a model for machine translation, speech recognition, or even image captioning.

3. Speed and Efficiency:

Task: You require efficient processing of large datasets with parallelization capabilities.

Example: Training a model on a large dataset where speed is crucial, like in real-time applications.

4. General-Purpose Learning:

Task: You need a versatile model that can be adapted to a variety of tasks, including those beyond natural language processing.

Example: Adapting the Transformer architecture for tasks in computer vision or multi-modal learning.

Examples of Transformers

1. BERT (Bidirectional Encoder Representations from Transformers):

Use Case: Text classification, sentiment analysis, question answering, named entity recognition.

Characteristic: Focuses on understanding the context of a word in a sentence by looking at both its left and right context.

2. GPT (Generative Pre-trained Transformer):

Use Case: Text generation, summarization, translation.

Characteristic: Uses the Transformer architecture's decoder to generate coherent and contextually relevant text.

3. T5 (Text-To-Text Transfer Transformer):

Use Case: Translation, text summarization, text generation.

Characteristic: Converts every NLP problem into a text-to-text format, enabling a single model to handle multiple tasks.

4. ViT (Vision Transformer):

Use Case: Image classification, object detection.

Characteristic: Adapts the Transformer architecture for processing images, treating image patches as sequence data.

5. RoBERTa (Robustly Optimized BERT Pretraining Approach):

Use Case: Similar to BERT, with improvements in training methodology.

Characteristic: Focuses on better performance by optimizing the pre-training process.

Examples of Large Language Models (LLMs)

1. GPT-3 (Generative Pre-trained Transformer 3):

Use Case: Text generation, dialogue systems, content creation, coding assistance.

Characteristic: A massive model with 175 billion parameters, capable of generating highly coherent and contextually relevant text.

2. BERT (Bidirectional Encoder Representations from Transformers):

Use Case: Though primarily a Transformer, BERT is also used in language understanding tasks as part of larger LLM systems.

Characteristic: Focuses on understanding rather than generating text, often used in tasks requiring contextual understanding.

3. LLaMA (Large Language Model Meta AI):

Use Case: Text generation, conversation agents, and research in AI language capabilities.

Characteristic: A family of LLMs developed by Meta, designed for efficiency and versatility in natural language understanding and generation.

4. Turing-NLG (Natural Language Generation):

Use Case: Text generation, conversational AI, creative writing.

Characteristic: Developed by Microsoft, it's one of the largest models for natural language generation, with 17 billion parameters.

5. PaLM (Pathways Language Model):

Use Case: Text summarization, translation, question answering.

Characteristic: A large language model developed by Google, designed to handle complex tasks across multiple languages and domains.

Conclusion

While Large Language Models and Transformers are closely related, they serve distinct roles in the AI landscape. LLMs are powerful tools for language generation and understanding, often built on the Transformer architecture. On the other hand, Transformers are a versatile architecture that can be applied to a variety of tasks, not limited to language. Understanding the differences between these two concepts is crucial for anyone looking to explore the cutting edge of AI technology.

This comparison between LLMs and Transformers highlights the evolving nature of AI, where models and architectures are continually being refined and expanded to tackle increasingly complex problems.

What are LLMs?

Large language models (LLMs) are a category of foundation models trained on immense amounts of data making them capable of understanding and generating natural language and other types of content to perform a wide range of tasks.

LLMs have become a household name thanks to the role they have played in bringing generative AI to the forefront of the public interest, as well as the point on which organizations are focusing to adopt artificial intelligence across numerous business functions and use cases.

Outside of the enterprise context, it may seem like LLMs have arrived out of the blue along with new developments in generative AI. However, many companies, including IBM, have spent years implementing LLMs at different levels to enhance their natural language understanding (NLU) and natural language processing (NLP) capabilities. This has occurred alongside advances in machine learning, machine learning models, algorithms, neural networks and the transformer models that provide the architecture for these AI systems.

LLMs are a class of foundation models, which are trained on enormous amounts of data to provide the foundational capabilities needed to drive multiple use cases and applications, as well as resolve a multitude of tasks. This is in stark contrast to the idea of building and training domain specific models for each of these use cases individually, which is

prohibitive under many criteria (most importantly cost and infrastructure), stifles synergies and can even lead to inferior performance.

LLMs represent a significant breakthrough in NLP and artificial intelligence, and are easily accessible to the public through interfaces like Open AI's Chat GPT-3 and GPT-4, which have garnered the support of Microsoft. Other examples include Meta's Llama models and Google's bidirectional encoder representations from transformers (BERT/RoBERTa) and PaLM models. IBM has also recently launched its Granite model series on watsonx.ai, which has become the generative AI backbone for other IBM products like watsonx Assistant and watsonx Orchestrate.

In a nutshell, LLMs are designed to understand and generate text like a human, in addition to other forms of content, based on the vast amount of data used to train them. They have the ability to infer from context, generate coherent and contextually relevant responses, translate to languages other than English, summarize text, answer questions (general conversation and FAQs) and even assist in creative writing or code generation tasks.

They are able to do this thanks to billions of parameters that enable them to capture intricate patterns in language and perform a wide array of language-related tasks. LLMs are revolutionizing applications in various fields, from chatbots and virtual assistants to content generation, research assistance and language translation.

As they continue to evolve and improve, LLMs are poised to reshape the way we interact with technology and access information, making them a pivotal part of the modern digital landscape.

EbookHow to choose the right foundation model

Learn how to choose the right approach in preparing data sets and employing foundation models.

Read the ebook

Related content

Register for the ebook on AI data stores

Unlock the future of business with instructLab in watson.ia, discover how you can create a powerful, tailored AI solution that meets your unique needs.

How large language models work

LLMs operate by leveraging deep learning techniques and vast amounts of textual data. These models are typically based on a transformer architecture, like the generative pre-

trained transformer, which excels at handling sequential data like text input. LLMs consist of multiple layers of neural networks, each with parameters that can be fine-tuned during training, which are enhanced further by a numerous layer known as the attention mechanism, which dials in on specific parts of data sets.

During the training process, these models learn to predict the next word in a sentence based on the context provided by the preceding words. The model does this through attributing a probability score to the recurrence of words that have been tokenized—broken down into smaller sequences of characters. These tokens are then transformed into embeddings, which are numeric representations of this context.

To ensure accuracy, this process involves training the LLM on a massive corpora of text (in the billions of pages), allowing it to learn grammar, semantics and conceptual relationships through zero-shot and self-supervised learning. Once trained on this training data, LLMs can generate text by autonomously predicting the next word based on the input they receive, and drawing on the patterns and knowledge they've acquired. The result is coherent and contextually relevant language generation that can be harnessed for a wide range of NLU and content generation tasks.

Model performance can also be increased through prompt engineering, prompt-tuning, fine-tuning and other tactics like reinforcement learning with human feedback (RLHF) to remove the biases, hateful speech and factually incorrect answers known as "hallucinations" that are often unwanted byproducts of training on so much unstructured data. This is one of the most important aspects of ensuring enterprise-grade LLMs are ready for use and do not expose organizations to unwanted liability, or cause damage to their reputation.

LLM use cases

LLMs are redefining an increasing number of business processes and have proven their versatility across a myriad of use cases and tasks in various industries. They augment conversational AI in chatbots and virtual assistants (like IBM watsonx Assistant and Google's BARD) to enhance the interactions that underpin excellence in customer care, providing context-aware responses that mimic interactions with human agents.

LLMs also excel in content generation, automating content creation for blog articles, marketing or sales materials and other writing tasks. In research and academia, they aid in summarizing and extracting information from vast datasets, accelerating knowledge discovery. LLMs also play a vital role in language translation, breaking down language barriers by providing accurate and contextually relevant translations. They can even be used to write code, or "translate" between programming languages.

Moreover, they contribute to accessibility by assisting individuals with disabilities, including text-to-speech applications and generating content in accessible formats. From healthcare to finance, LLMs are transforming industries by streamlining processes, improving customer experiences and enabling more efficient and data-driven decision making.

Most excitingly, all of these capabilities are easy to access, in some cases literally an API integration away.

Here is a list of some of the most important areas where LLMs benefit organizations:

- **Text generation**: language generation abilities, such as writing emails, blog posts or other mid-to-long form content in response to prompts that can be refined and polished. An excellent example is retrieval-augmented generation (RAG).
- **Content summarization**: summarize long articles, news stories, research reports, corporate documentation and even customer history into thorough texts tailored in length to the output format.
- **AI assistants**: chatbots that answer customer queries, perform backend tasks and provide detailed information in natural language as a part of an integrated, self-serve customer care solution.
- **Code generation**: assists developers in building applications, finding errors in code and uncovering security issues in multiple programming languages, even "translating" between them.
- **Sentiment analysis**: analyze text to determine the customer's tone in order understand customer feedback at scale and aid in brand reputation management.
- **Language translation**: provides wider coverage to organizations across languages and geographies with fluent translations and multilingual capabilities.

LLMs stand to impact every industry, from finance to insurance, human resources to healthcare and beyond, by automating customer self-service, accelerating response times on an increasing number of tasks as well as providing greater accuracy, enhanced routing and intelligent context gathering.

LLMs and governance

Organizations need a solid foundation in governance practices to harness the potential of AI models to revolutionize the way they do business. This means providing access to AI

tools and technology that is trustworthy, transparent, responsible and secure. AI governance and traceability are also fundamental aspects of the solutions IBM brings to its customers, so that activities that involve AI are managed and monitored to allow for tracing origins, data and models in a way that is always auditable and accountable.