#### 

**Machine Learning's Revolution in Computer Science (2018-2023)**

Phanindra kumar Tirumalasetty

Department of Data Science, University of Maryland, Baltimore County(UMBC)

DATA 606: Capstone in Data Science

Dr. Tony Diana

May 4, 2024

# Abstract

This study discusses the effect of Machine Learning (ML) on Computer Science from 2018 to 2023, focusing on the transformer model. In “Attention is All You Need,” published in 2017, the introduction of Transformer models significantly transformed this field by overcoming the weaknesses of Recurrent Neural Networks(RNNs) and Long Short-Term Memory Networks (LSTMs). Transformers enable parallel processing and include a mechanism that allows for attention. The progress development and usage of Transformer-based models like BERT and GPT-3 are highlighted in this paper and have changed how we think about natural language processing (NLP). This paper also examines how Large Language Models (LLMs) are used in different sectors, their challenges such as hallucinations, prospects like quantum computing integration, and new architectures like Hyena.

# Introduction

The world of Artificial intelligence(AI) research has been climbing slowly and steadily for decades. Since the 1980s, we have known about artificial neural networks and continued developing them. But then, in recent years, suddenly, we got things like GPT, DALL-E, and Tesla’s Full Self-Driving (FSD) system, which are all powerful AI. Did the machines suddenly get smarter?

The answer is yes. Image captioning with Machine Learning algorithms was a significant feat five years ago. Today, the field works on the inverse – generating images or videos based on a sentence. This quick pace of change is due to several critical breakthroughs made between 2018 and 2023. However, the key to success was a newly innovative software design for neural networks, the Transformer model. Published initially in the paper “Attention is All You Need” in 2017, the Transformer became a distinguishing novelty secret weapon for enhancing AI capabilities – something deep learning has never yet achieved.

This paper discusses the shift from neural networks (particularly RNNs or Recurrent Neural Networks) to Transformers and how they paved the way for large language models (LLMs). It also provides an overview of where we stand now regarding Artificial Intelligence, with predictions about future directions this field may take and Machine Learning techniques that could be used alongside these methods, leading towards more complex AIs.

# 

# Literature review

Vaswani, A., and seven other Google researchers (2017) introduced a novel neural network model, Transformer, which changed how we look at natural language processing. This differs from traditional recurrent neural networks (RNNs), where its self-attention mechanisms allow it to simultaneously examine all parts of a sentence. When this approach was adopted, there was great improvement in various natural language processing applications such as machine translation, text summarization, and question answering.

In a study by Devlin J in 2018, Bidirectional Encoder Representations from Transformers (BERT) were introduced as an advanced language model that works well in different NLP tasks. Unsupervised learning based on the BERT pre-training approach is demonstrated in masked language modeling and next-sentence prediction to show the efficacy of extensive unsupervised learning in NLP.

GPT-3 (Generative Pre-trained Transformer 3), developed by Brown Tom B and a few other researchers in 2020 from Open AI, is a powerful language model that can perform various language tasks with little domain-specific training data requirement. This work has extended the frontiers of large language models, thus showing how transformer-based architectures can be scaled up for use in general-purpose AI applications.

# From Neural Networks to Transformers: The Rise of Attention-based Models

## What is a Neural network?

Neural networks take inspiration from the human brain and consist of nodes or neurons linked in a web-like arrangement. Typically, a neural network has three layers: input, hidden, and output. The input layer receives data, the hidden layers process it, and the output layer produces the final result. Then comes Deep learning. It is an area within machine learning that employs complicated structures called deep neural networks, consisting of many layers with connected nodes or neurons. These layers allow the model to learn from large amounts of data, making it useful for image and speech recognition tasks. “Deep Learning waves have lapped at the shores of computational linguistics for several years now, but 2015 seems like the year when the full force of the tsunami hit the major Natural Language Processing (NLP) conferences.” - (Dr. Christopher D. Manning et al., 2015)

## What can Neural networks do for you?

Neural networks have become a powerful tool across a surprising range of tasks. They can use image data to recognize faces in photos or guide self-driving cars. Natural language processing is another area where they shine, allowing them to understand and respond to human language, whether for chatbots, translating languages, or writing different creative text formats.

## Can Neural Networks Remember? Exploring RNNs and LSTMs for Sequential Data?

Recurrent Neural Networks (RNNs) are a significant method in traditional machine learning; they were designed mainly for working with sequential data like sentences in a text or time series. While traditional neural networks excel at static data like images, they struggle with sequential information like speech or text. This limitation arises because they process information in a single pass, making it challenging to capture long-range dependencies. That is where RNNs come in. RNNs are a particular neural network explicitly designed to handle sequential data. They incorporate a feedback loop that allows them to consider past information when processing the present input.

However, they need help with long-range dependencies due to the vanishing gradient problem, where the influence of initial inputs diminishes in long sequences. To address this limitation, LSTMs, or long-short-term memory networks, are a more advanced type of RNN with a different structure with gates to control the flow of information through the network. These gates allow LSTMs to retain important information over long periods, which makes them useful for language modeling and text generation, among other tasks.

## The Fall of RNNs and Rise of Transformers:

So, what led to this Transformer takeover? As impressive as RNNs are, they have their limitations. They process information step-by-step, like reading a book word by word. This approach worked well for them until they had to deal with long-range dependencies – relationships between words or data points far apart in a sequence. In other words, it is like recalling what happened at the beginning of a very long novel when you’re close to its end — things get blurry. The second issue was that RNNs suffered from slowness and high computational cost, especially with large datasets.

# Introduction to Transformers

If there were ever an origin story for AI in this century, it would be Google’s 2017 research paper, "Attention Is All You Need.” This deep learning architecture, known as transformers, has since dominated the field of AI (2018-2023) after being introduced by this paper. Unlike RNNs or LSTMs, which take in data one word at a time, transformers take in entire sequences simultaneously. This parallel processing makes them efficient and adept at capturing complex relationships in data, a crucial factor in tasks like language translation and summarization.

## Why Transformers Surpassed Previous Models

1. **Parallel Processing:** Unlike Recurrent Neural Networks, which process the data sequentially, Transformers analyze everything simultaneously. This enables them to grasp far-ranging relationships in data, which RNNs have traditionally struggled with.
2. **Attention Mechanism:** Transformers can concentrate on the most significant parts of the data irrespective of their position in the sequence. This is like highlighting critical points across a lengthy document, leading to a better understanding.
3. **Efficiency:** Parallel processing and focused attention make Transformers faster and more efficient than RNNs, especially when handling large datasets.
4. **Pre-training and Transfer Learning:** Transformers can be pre-trained on massive datasets, making them powerful tools for transfer learning across various tasks. Due to their sequential nature, this is more challenging for RNNs.

## Why, exactly, was the transformer such a massive breakthrough?

Before transformers, RNNs dominated language AI, processing text one word at a time. Transformers, with their core concept of "attention," revolutionized the field. They ditched RNNs, analyzing all words simultaneously (parallel processing) and relying only on the attention mechanism to capture long-distance word relationships. This led to a more global understanding of the text and increased efficiency and scalability due to the perfect timing with the rise of powerful GPUs.

# Applications of Transformers: Conquering the World One Sequence at a Time

With their outstanding capabilities, Transformers have rapidly become the preferred models for several tasks, particularly in natural language processing (NLP).

## Enter Large Language Models (LLMs)

Transformers are the backbone of many Large Language Models (LLMs). They are trained on massive amounts of text data. As a result, they can generate coherent and fluent text.LLMs are good at natural language processing tasks like translation, text summarization, and conversational agents. LLMs do well because they are pre-trained on a large corpus of text data and can be fine-tuned for specific tasks. GPT is an example of a Large Language Model. These models are called “large” because they have billions of parameters that shape their responses. For instance, GPT-3, the giant version of GPT, has 175 billion parameters and was trained on a massive corpus of text data.

Figure1

*Brief overview of research milestones leading to modern LLMs.*



## 

## Types of Large Language Models:

### LLMs are few shot learners: GPT-3

GPT-3, or Generative Pre-trained Transformer 3rd generation, is a neural network machine learning model trained using internet data to generate text. It requires little input text but gives vast amounts of relevant and sophisticated machine-generated text as output. OpenAI developed GPT -3. The deep learning neural network that powers GPT-3 has over 175 billion machine learning parameters. The GPT-3 paper “Language Models are Few-shot Learners” showed that LLMs improve at few-shot learning by scaling up LLMs in terms of parameter and dataset size.

### What is BERT?

BERT is a machine learning framework for natural language processing. Google developed it in 2018 to enhance the understanding of unlabelled text in different tasks by learning to predict what might come before and after any given word or sentence.BERT is used for a wide variety of language tasks. It can help chatbots answer questions and predict text when writing an email. BERT can quickly summarize long legal contracts and differentiate words with multiple meanings based on the surrounding text.

### GPT-4

GPT-4 by OpenAI is a text generation model that can generate creative text formats such as poems, code, scripts, musical composition, and more. This language model is excellent in conversational language and can also be very flexible. However, its factual accuracy cannot always be guaranteed. According to OpenAi, “It is impossible to tune every corner case. Sometimes, GPT may exhibit bad accuracy or be unwilling to cooperate.”

### ***Claude***

From Google AI: LLM Claude scored very high on reasoning and factual accuracy. This model is incredible as Claude can pick up any fact from its internal knowledge base and process it to suit many tasks. Therefore, the model used appropriate logic and context.

### Gemini

Several versions of this language model have been developed for different purposes. The most powerful of these is the Gemini Ultra, which handles complex tasks, while the versatile one you’re now interacting with within Bard is called Gemini Pro. There’s also a Gemini Nano specially designed to be more efficient on mobile devices. Multimodal abilities make Gemini stand out among others – it can handle text and image data.

## 

## What can an LLM do for you?

Large Language Models (LLMs) are potent language tools that can do many things. Here are a few examples:

1. **Search:** LLMs can improve the quality of search results by providing the user with more relevant and accurate information. Search engines like Google and Bing already use LLMs to provide better user results.
2. **Generate Content (Write or Edit):** Generating content based on prompts provided by a user is one of the most common use cases for LLMs
3. **Answering Questions:** It combines “Search” and “Summarize.” LLMs can power question-answering systems in many areas, such as customer service, education, and healthcare.
4. **Language translation:** LLMs can help improve the accuracy and speed of language translation by understanding the softness of different languages and doing natural translations. (Cellstrat, 2023)
5. **Content creation:** An LLM can create various types of content, such as product descriptions, marketing copy, or even summarizations for long documents.
6. **Code generation:** People have started experimenting with LLMs to generate codes that could help programmers with their work or automate some repetitive tasks.

**Hallucinations of LLMs**

LLMs hallucinate when they generate text that is:

1. Untrue: LLMs can provide information that can be proven wrong, even if it sounds like it makes sense.
2. Senseless: The resultant text may be grammatical but meaningless or deviate from the original question.
3. Disjointed: The answer might not correspond with the initial question or conversation.

**Reasons behind Hallucination:**

1. Incomplete or biased training data: Flawed or monotonous data used to train LLMs can result in outputting biased responses or ones that are simply untrue.
2. Ambiguous user prompts: Unclear prompts may need to be clarified so that the LLM can provide responses based on mistaken interpretations of intent.
3. The statistical nature of LLMs: An LLM is a complex pattern matcher. It can assemble what seems like logical data, but based on its learning materials, it turns out to be inaccurate.

# State of AI environment in future trends

Imagine AI as a toolbox. Transformers were the shiny new hammers that could seemingly fix everything. Now, researchers realize a hammer isn't always the best tool. That's why they're building a whole new toolbox filled with specialized wrenches, screwdrivers, and maybe even a laser cutter.

All good things must end. The robust transformer architecture faces challenges. Its training demands vast computing resources, causing chip shortages. Additionally, processing long sequences becomes computationally expensive due to the attention mechanism's quadratic scaling with sequence length. This limits the context window, the maximum sequence a model can handle. Researchers are actively seeking alternative architectures to overcome these limitations and improve efficiency for long sequences.

Challengers to the throne - While transformers dominate AI, researchers are exploring new architectures to address their limitations. Hyena, a subquadratic architecture, uses convolutions and element-wise multiplication instead of attention. (Poli, Michael, et al.,2023) This makes Hyena faster and better suited for long sequences, achieving state-of-the-art performance with less computation. HyenaDNA, built on Hyena, tackles genomics with a 1-million-token context window.

Quantum computing is yet another impending paradigm shift. Although they are still in their infancy, quantum computers have the power to completely transform artificial intelligence by processing data in radically different ways. Qubits, which can concurrently exist in both states, are the basis of quantum computers as opposed to classical computers, which use bits (0s and 1s). Because of this, they can now complete some computations tenfold quicker than they could with traditional computers. (Rao Rahul, 2024)

The future of AI may involve a combination of specialized classical architectures like Hyena and the immense processing power of quantum computers, creating a groundbreaking toolbox for tackling tomorrow's challenges.

# References

Poli, M. (2023, April 19). Hyena Hierarchy: Towards Larger Convolutional Language Models. https://arxiv.org/pdf/2302.10866

Rao, R. (2024, April 22). *Quantum computers can run powerful AI that works like the brain*. Scientific American.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023, August 2). *Attention is all you need*. arXiv.org. https://arxiv.org/abs/1706.03762

Brown, T. B. (n.d.). Language models are few-shot learners. http://arxiv.org/pdf/2005.14165

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, May 24). *Bert: Pre-training of deep bidirectional Transformers for language understanding*. arXiv.org. https://arxiv.org/abs/1810.04805

Manning, C. D. (2015). Computational Linguistics and Deep Learning - ACL Anthology. https://aclanthology.org/J15-4006.pdf

CellStrat. (2023, April 25). *Real-world use cases for large language models (llms)*. Medium. https://cellstrat.medium.com/real-world-use-cases-for-large-language-models-llms-d71c3a577bf2