Understanding the Complexity of Transformers in AI: The Enigma of AI Thinking Since AlexNet

The field of artificial intelligence (AI) has witnessed remarkable advancements, but the opacity of how AI models operate has also increased. Since AlexNet's groundbreaking success in 2012, AI systems have become increasingly difficult to interpret, leading to ongoing debates about the possibility of ever truly understanding how AI "thinks." Among the most complex systems today are transformers, a type of model that has revolutionized natural language processing (NLP), yet operates in ways that even its creators struggle to fully explain.

This article delves into the intricate architecture of transformers, the nature of AI’s thought process, and why unraveling the internal logic of these models might be impossible. We will first explore AlexNet’s role in ushering in a new era of AI opacity before examining how transformers have compounded these challenges.

The AlexNet Revolution: A New Age of Uninterpretable AI

Before 2012, AI was already making waves, but it was AlexNet, developed by Alex Krizhevsky and his team, that significantly altered the landscape. This deep convolutional neural network (CNN) was a key breakthrough, winning the ImageNet Large Scale Visual Recognition Challenge by a huge margin. The significance of AlexNet’s victory cannot be overstated—it was one of the first demonstrations that deep learning could outperform traditional machine learning techniques in image recognition.

However, AlexNet’s success also marked the point at which neural networks became more opaque. The architecture was too complex for simple inspection, and trying to interpret its reasoning process became an exercise in frustration for many AI researchers. Although researchers could measure its performance, the "why" behind its decisions often remained elusive.

This opacity was a direct consequence of deep learning. The deeper the network, the more complex the transformations performed by its layers. Even though we could trace individual neuron activations, understanding the rationale behind their combinations became virtually impossible.

The Rise of Transformers: From Attention Mechanisms to Unfathomable Depths

While CNNs like AlexNet dominated early deep learning research, another paradigm shift occurred with the introduction of the transformer model in the paper Attention is All You Need by Vaswani et al. in 2017. Transformers marked a radical departure from previous architectures, relying on an attention mechanism that could weigh the importance of different elements in an input sequence, rather than processing inputs in a fixed order like recurrent neural networks (RNNs).

This shift to attention-based architectures allowed transformers to handle vast amounts of data and excel in tasks like machine translation, text generation, and question-answering systems. Models like BERT, GPT-3, and ChatGPT are direct descendants of the original transformer architecture.

However, the success of transformers also introduced new levels of complexity that made understanding AI’s thought processes even more elusive than in AlexNet's era. While the attention mechanism theoretically provides more transparency by showing which parts of the input the model is focusing on, the overall reasoning of transformers is distributed across millions—sometimes billions—of parameters. This vastness makes it extremely difficult to track how individual decisions are made or to interpret the model’s internal logic in any meaningful way.

Understanding Attention: A Double-Edged Sword

The attention mechanism is often cited as an attempt to make AI models more interpretable. By quantifying which parts of the input sequence (e.g., words in a sentence) are given more "attention," it provides a glimpse into what the

model is considering important at any given moment. In some sense, it’s like being able to see which neurons are firing most strongly in a biological brain.

However, attention isn’t necessarily synonymous with interpretability. The fact that a model pays more attention to certain words doesn’t explain why it has made a specific decision or what patterns it is abstracting across the input data. Attention scores are only a surface-level phenomenon; the deeper logic of why the model behaves the way it does is still hidden within its layers of complex interactions.

The issue is further compounded when models like GPT-3, which have hundreds of billions of parameters, are trained on enormous datasets. The training process involves so many minute adjustments to the model's weights that even minor differences in input can lead to vastly different outputs. This raises the question: Is it possible for humans to ever truly understand the intricate workings of such a model?

Why We May Never Fully Understand AI’s Thought Process

At the core of the issue lies a fundamental challenge: AI models like transformers operate on a scale of complexity that far exceeds human cognition. Each layer in a transformer consists of hundreds of neurons interacting in non-linear ways, and with each successive layer, these interactions become increasingly difficult to parse.

Here are several key reasons why understanding AI, particularly transformers, may be a fundamentally impossible task:

High Dimensionality: Transformers process data in extremely high-dimensional spaces. While humans are adept at thinking in three or four dimensions, transformers operate in hundreds or even thousands of dimensions. Visualizing or even conceptualizing how decisions are made in such high-dimensional spaces is beyond our natural capabilities.

Distributed Representations: In transformers, information is not stored in discrete units but is distributed across many neurons. Each neuron participates in a multitude of functions, and the specific role of any given neuron might be difficult or impossible to isolate. It’s not like a decision tree, where we can follow a clear, hierarchical path to a decision.

Non-Linear Interactions: Neural networks, including transformers, rely heavily on non-linear activation functions. This means that even small changes in input can lead to disproportionately large changes in output. Tracking these non-linear interactions through multiple layers makes it incredibly difficult to understand how the model arrived at its conclusion.

Training on Massive Datasets: Modern transformers are trained on massive datasets—like the entirety of the internet for GPT models. This sheer scale makes it difficult to even know what information the model has seen, much less how it has processed and abstracted that information into its internal representations.

The Black Box Problem: Even with interpretability techniques such as attention maps, layer-wise relevance propagation, or saliency maps, these tools only scratch the surface of understanding. The "black box" nature of AI—where the inner workings remain inscrutable—remains largely intact despite these efforts.

The Limits of Explainable AI (XAI)

Explainable AI (XAI) has emerged as a field aimed at making AI systems more interpretable. While some success has been achieved in developing tools that can provide insights into model decisions, XAI has its own limitations. Many of the techniques focus on post-hoc explanations, meaning they try to make sense of the model’s decisions after they’ve been made. These explanations may not accurately reflect the true reasoning process of the model, which further complicates the issue of trust and reliability.

Additionally, efforts to make models more interpretable often come at the expense of accuracy. Simpler models like decision trees or linear models are easier to interpret but perform worse on complex tasks compared to transformers. There’s an inherent trade-off between interpretability and performance, which raises the question of whether it’s even worth trying to make state-of-the-art models more interpretable.