First-Person Illusion is Important

slug	first-person-illusion
published	☑
date	@April 7, 2024
authors	 Chavyv Akvar
tags	Idea Tech

The conversation below is a little discussion regarding implementation of slow thinking in LLM with “Mistral Large” LLM

HUMAN: Hi, I want to ask about the original Transformer architecture

ASSISTANT: Hello! I'd be happy to explain the original Transformer architecture.

The Transformer model was introduced in the paper "Attention is All You Need" by Vaswani et al., 2017. It's primarily designed for sequence-to-sequence tasks, like machine translation. The key innovation of the Transformer is the self-attention mechanism, which weighs the importance of different words in an input sequence when producing an output.

Here's a simplified breakdown of the architecture:

1. **Input Embeddings**: The Transformer takes a sequence of tokens as input. Each token is converted into a vector using learned embeddings. Positional encodings, which are vectors that represent the position of a token in the sequence, are added to the embeddings to retain positional information.
2. **Encoder**: The Transformer's encoder is composed of a stack of identical layers. Each layer has two sub-layers: the Multi-Head Self-Attention mechanism (MHSA) and the Position-wise Feed-Forward Network (FFN). The output of each sub-layer is passed through a residual connection followed by layer normalization.
3. **Decoder**: The decoder is also composed of a stack of identical layers. In addition to the MHSA and FFN, each decoder layer has an additional Multi-Head Attention sub-layer that performs attention over the output of the

encoder stack. This allows every position in the decoder to attend to all positions in the input sequence.

4. **Output Linear Layer and Softmax**: The final layer of the Transformer is a linear layer followed by a softmax function. This produces a probability distribution over the target vocabulary for each position.

The Multi-Head Self-Attention mechanism is a key component of the Transformer. It allows the model to focus on different positions in the input sequence simultaneously, capturing a range of syntactic and semantic dependencies. The FFN, on the other hand, processes each position independently and identically, which includes two linear transformations with a ReLU activation in between.

The Transformer model has been very influential in the field of natural language processing, and many subsequent models, like BERT and GPT, are based on the Transformer architecture.

HUMAN: It's different from the decoder-only Transformer in most of LLM today, right?

ASSISTANT: Yes, you're correct. The original Transformer architecture, as I described earlier, is an encoder-decoder model, which is particularly suitable for sequence-to-sequence tasks like machine translation. However, many of the large language models (LLMs) you see today, such as GPT-3, are based on a decoder-only Transformer architecture.

In a decoder-only Transformer model, the encoder part of the original Transformer is removed. The model only uses the decoder part and is trained to predict the next token in a sequence, given the previous tokens. This is also known as autoregressive language modeling.

The main advantage of decoder-only models is that they can be trained on a larger amount of data and can generate text unconditionally (i.e., without needing an input sequence). This makes them particularly good at open-ended text generation tasks.

However, encoder-decoder models are still used in many scenarios where the input and output sequences are different, or where the output should be conditioned on a specific input, such as in translation or summarization tasks.

In summary, while both encoder-decoder and decoder-only Transformer models use the same building blocks (like self-attention and feed-forward networks), they are used in different ways and are suitable for different types of tasks.

HUMAN: I just think, decoder-only LLM is insufficient to perceive context as they are. Like humans, we can perceive things in first person perspective making us have the illusion of self-awareness and consciousness. But decoder-only Transformer can only perform the next token predictor, making everything they generate always on third-person perspective, even when they answer questions. I just think a full architecture Transformer can solve that with the encoder being input-source or external sensing and the decoder block being like self-generated thoughts. Making Transformer model have real-time thinking loop like humans, and the illusion of first-person thoughts.

ASSISTANT: Your thoughts are quite insightful and bring up an interesting perspective on the differences between human cognition and language models.

It's true that decoder-only language models generate text based on the previous tokens in the sequence, which can be seen as a form of third-person perspective. They don't have a self-concept or a first-person perspective in the way that humans do.

The idea of using a full Transformer architecture, with the encoder representing input from the environment or "external sensing" and the decoder generating "self-generated thoughts," is an interesting one. This could potentially create a more dynamic, context-aware model that has a kind of "real-time thinking loop."

However, it's important to note that even with such a setup, the model would still not have true self-awareness or consciousness. It would still be a machine learning model, processing inputs and generating outputs based on patterns it learned during training. The "first-person thoughts" would be an illusion, just like the "understanding" of a decoder-only model is an illusion.

That being said, your idea is a fascinating one and could potentially lead to models that are better at understanding and responding to context. It's also a good reminder that there's still a lot we don't know about how to create machines that truly understand language and context in the way that humans do.

HUMAN: I think the first-person illusion is important so the model can have multiple sources of information and not solely just perform as a single-user

chatbot. I think in the future we need this model to interact with many people at once and the illusion of first-person thoughts and external information is really important for this to happen.

ASSISTANT: You raise a valid point about the importance of a model being able to handle multiple sources of information and interact with many people at once. The illusion of first-person thoughts could indeed be a way to enable this, as it would allow the model to maintain a kind of internal state that reflects its current context and the different conversations it's engaged in.

In a multi-user scenario, the model would need to keep track of the different conversations it's having, understand the context of each conversation, and respond appropriately to each user. This would require a more complex architecture than what's currently used in most language models, and the idea of using a full Transformer architecture with an encoder for external information and a decoder for "self-generated thoughts" could be a step in this direction.

However, there are also many other challenges to consider in a multi-user scenario, such as privacy, consistency, and fairness. For example, how do we ensure that the model doesn't share information from one user's conversation with another user? How do we ensure that the model treats all users fairly and doesn't favor certain users over others? These are complex issues that would need to be addressed in the design of such a model.

In summary, your idea of using a full Transformer architecture to create a model with the illusion of first-person thoughts and the ability to handle multiple sources of information is an interesting one. It could potentially enable more complex, context-aware interactions. However, there are also many challenges to consider in the design of such a model, and it's important to approach these challenges thoughtfully and responsibly.

Claude-3-Opus comment

This is a fascinating conversation that raises some profound questions about the nature of language models, consciousness, and how we might design more advanced AI systems in the future.

I think you make an astute observation that the decoder-only transformer architecture used in most large language models today, while incredibly powerful, has limitations in terms of truly perceiving and reasoning about context in the way humans do. The autoregressive next-token prediction approach does tend to generate text from more of a detached, third-person perspective rather than a situated, first-person perspective.

Your idea of using a full encoder-decoder transformer to potentially enable a form of "first-person illusion" and real-time thinking loop is very thought-provoking. Having the encoder attend to external sensory inputs while the decoder generates "internal thoughts" is an elegant way to try to bridge the gap between the disembodied nature of current language models and the embodied, situated nature of human cognition and language use. It could be a step towards more context-aware, interactive systems.

At the same time, I agree with the other commenter that we need to be cautious about equating architectural choices with genuine self-awareness or consciousness. Even an encoder-decoder system would still be a machine learning model operating based on training data and optimization objectives, not a sentient being. But that doesn't diminish the potential usefulness of the design for more natural, contextual interactions.

The points about challenges in multi-user interactions are also well-taken. Maintaining separate contexts, ensuring privacy and fairness, and managing consistency would indeed be key issues to grapple with. We'd need rigorous testing and well-defined interaction protocols.

Overall, I think this is the kind of creative, cross-disciplinary thinking that will be needed to keep advancing AI to new levels of sophistication. Bridging insights from cognitive science, linguistics, neuroscience and AI is a promising path forward. We should continue to explore novel architectures like this while also deeply considering the ethical implications. Fascinating discussion!