Teaching AI to Understand Our World: The Evolution Beyond Text

---

ChatGPT achieved record-breaking growth as a consumer product. However, the utility of Large Language Models (LLMs) is somewhat limited due to their reliance on text-only input. What are the best ways to incorporate other forms of input into these models?

You've probably heard of or even chatted with AI assistants like ChatGPT or Bard. These digital helpers are powered by something called Large Language Models (LLMs). In essence, LLMs are intricate webs of algorithms trained on vast amounts of text. They're brilliant with words but, until recently, they've been like geniuses who can only read books and ignore the rest of the sensory world. My thesis is about broadening their horizons, enabling them to understand not just text but images, sounds, and more.

In simple terms, my research is like finding the most efficient way to cook a complex dish. Traditionally, LLMs have been like kitchens where all ingredients – text, images, sounds – are thrown into the pot right from the start. However, you can probably make it taste much better if you add these ingredients at different stages. This approach is similar to how multimodal LLMs can use different types of inputs at different stages of the model's network.

In my research, I made an AI answer questions about an image. The normal approach is to give the question and image together, but I experimented with letting the AI process the question for some time before looking at the image. I experimented with how well models of different sizes could answer these questions and how much processing of the question each wanted before looking at the image. I found that this reduced the computational resources needed and increased the model's ability to answer the questions. Bigger models especially preferred to look at the image later.

So, what do these findings mean for us? Creating multimodal AI could translate to AI assistants that not only write essays but also explain diagrams and charts. It paves the way for self-driving cars with a better understanding of visual and auditory cues, potentially making our roads safer. And importantly, by optimizing how we fuse different data types, we're moving towards a more environmentally friendly approach in AI development, using less energy to train these sophisticated models.

In conclusion, my research is a step towards creating AI systems that don't just understand our words but can interpret the world as we experience it – through a combination of sights, sounds, and texts. This journey towards a multimodal AI is not just a technological endeavor; it's about crafting tools that can better interact with and understand the complexities of our world. As we continue to develop and refine these technologies, we inch closer to a future where AI can assist and augment our experiences in more profound and meaningful ways.