View on Web

The Finesse in Fusion - The Power of Multimodal AI

18th Oct,2023



If we are still wondering about the tantalizing future of Artificial Intelligence (AI), it's time to turn our attention to Multimodal AI. As humans, we are naturally adept at soaking up ideas and weaving context from the symphony of images, sounds, videos, and text. Multimodal AI integrates multiple sensory or data modalities, such as text, images, speech, and gestures, to enhance AI systems' capabilities. It allows machines to simultaneously process and understand information from different sources, enabling more comprehensive and human-like interactions. **While super chatbots like ChatGPT can churn out poetry and even pass the US bar exam, it is still a soloist in the orchestra of disruptive innovation. AI can be a versatile player in the orchestra or a true doppelganger of the human mind only when it is multimodal.**

ChatGPT maker OpenAI announced last month that its GPT-3.5 and GPT-4 models can analyze images and translate them into words, while its mobile apps will have speech synthesis so that users can have full-scale conversations with chatbots. In multimodal AI, numerous data types work together to help AI establish content and better interpret context,

something that was lacking in earlier AI.

How does multimodal AI differ from other AI?

Data is the fundamental difference between multimodal AI and traditional single-modal AI. Generally, single-modal AIs work with a single data source or type. A financial AI, for example, analyzes business financial data and economic and industry data to spot financial problems or make financial projections. In other words, single-modal AIs are focused on specific tasks. **In contrast, multimodal AI ingests and processes data from various sources, including video, images, speech, sound, and text, allowing the user to perceive the environment or situation in more detail. Multimodal AI thus simulates human perception more closely.**

Use Cases & Applications

Multimodal AI has a wide range of use cases compared to unimodal AI. Here are a few examples of how multimodal AI can be used:

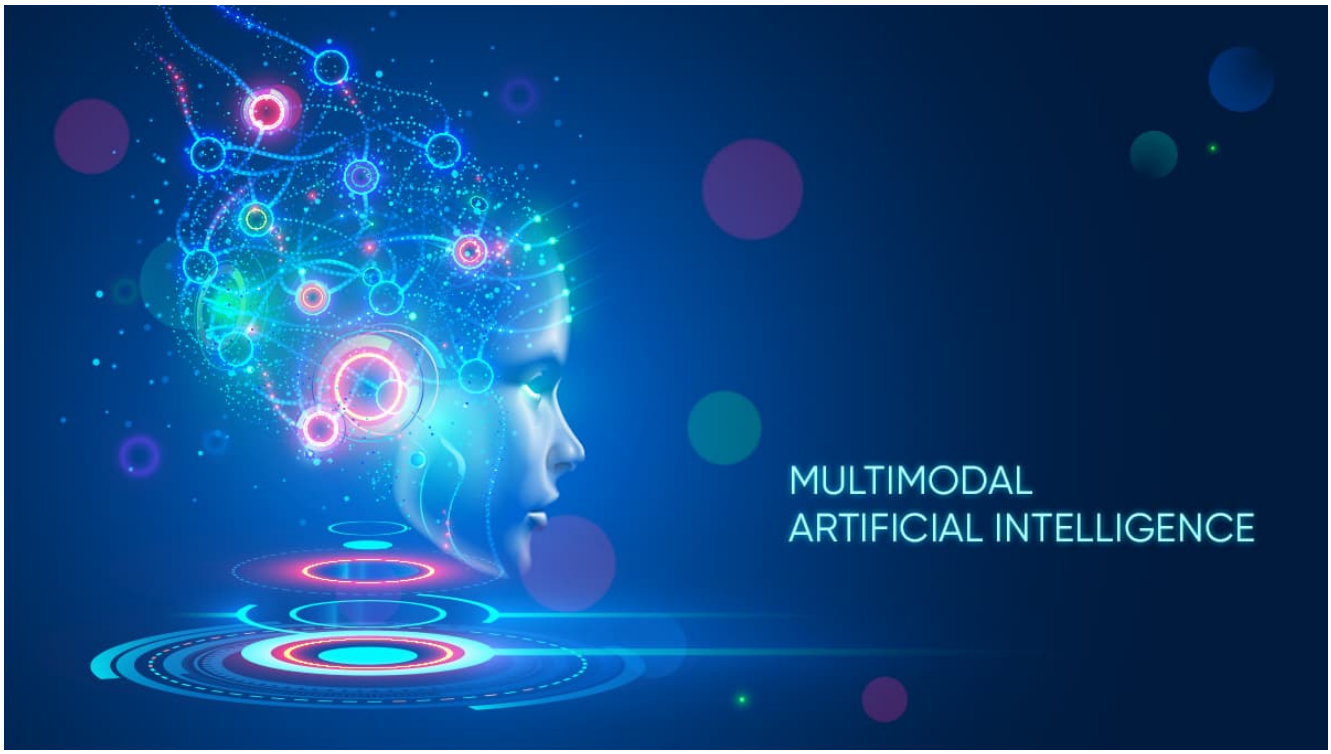
Computer vision: There is much more to computer vision than simply identifying objects in the future. By combining multiple data types, AI can better identify the context of an image. A dog image and dog sounds, for example, are more likely to result in an accurate identification of an object as a dog. Another possibility is to combine **facial recognition** with Natural Language Processing (NLP) to better identify an individual.

Industry: Multimodal AI has a wide range of applications in the workplace. Manufacturing processes can be overseen and optimized by multimodal AI, product quality can be improved, or maintenance costs can be reduced by using multimodal AI. A healthcare vertical uses multimodal AI to analyze vital signs, diagnostic data, and records of patients to improve treatment. The automotive vertical uses multimodal AI to monitor a driver for fatigue indicators, such as closed eyes and lane departures, to recommend rest or a change of drivers.

Language processing: **NLP** tasks such as sentiment analysis are performed by multimodal AI. By combining signs of stress in a user's voice with signs of anger in their facial expression, a system can tailor or temper responses according to the user's needs. It is also possible for AI to improve pronunciation and speech in other languages when text is combined with the sound of speech.

Robotics: Robots must interact with real-world environments, with humans, and with a wide variety of objects, such as pets, cars, buildings, their access points, etc. Multimodal AI is

crucial to robotics development. Multimodal AI uses data from cameras, microphones, GPS, and other sensors to create a detailed understanding of the environment.



Challenges and Limitations of Multimodal AI

Data Collection and Annotation Challenges: Collecting and annotating diverse and high-quality multimodal datasets can be a daunting task. It requires meticulous coordination and expertise to gather data from multiple sources and ensure consistent labeling across different modalities.

Domain Adaptation and Generalization Issues: Multimodal AI systems often struggle with adapting to different domains and generalizing their learnings across diverse data sources. The representations and features extracted from one modality may not easily translate or transfer to another.

Learning nuance: It can be challenging to teach an AI to distinguish between different meanings from identical input. Consider a person who says, "Wonderful." The AI understands the word, but "wonderful" can be interpreted as sarcastic disapproval. Using other contexts, such as speech inflections or facial cues, can help create an accurate response.

Decision-making complexity: Developing neural networks through training can be complex, making it difficult for humans to understand how AI makes decisions and evaluates data. Even extensively trained models use a finite data set, and it is impossible to predict how

unknown, unseen, or other new data might affect the AI and its decisions. As a result, multimodal AI can be unreliable or unpredictable.

Harnessing Multimodal AI for the Future

Multimodal AI holds immense promise in revolutionizing how machines perceive and understand the world. Despite the challenges and limitations, ongoing research and advancements in algorithms, the exploration of new modalities, and ethical considerations will pave the way for even more powerful multimodal AI systems. Multimodal AI will undoubtedly shape the future of AI technologies, leading to more intelligent, adaptable, and responsible systems that can better assist, understand, and engage with humans. **But don't worry, humans will have the last laugh - after all, AI can't laugh!**



AUTHOR:

Jayajit Dash

Senior Manager- Corporate Communications (Marketing)

© Copyright CSM Technologies, Level-6, E/56, Infocity, Chandrasekharpur, Bhubaneswar, Odisha - 751024