Rudy Shoushany, 2026, 14:1
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Multimodal Neural Networks: The Architectural Stepping Stone Toward Artificial General Intelligence

Rudy Shoushany

Abstract- The quest for Artificial General Intelligence (AGI) has shifted from specialized, narrow AI systems toward generalized foundation models capable of cross-domain reasoning. This paper explores the pivotal role of multimodal neural networks (MNNs) in this transition. By integrating diverse data streams—including text, vision, audio, and sensory inputs— MNNs mimic the human cognitive process of cross-modal alignment. We analyze current breakthroughs in native multimodal architectures, the shift from strong to weak semantic correlation learning, and the emergence of embodied AI as a critical path toward AGI. Our findings suggest that while MNNs provide the necessary perceptual framework for AGI, the integration of autonomous reasoning and self-correcting feedback loops remains the final frontier.

Keywords: Artificial General Intelligence (AGI); Multimodal Neural Networks; Foundation Models; Cross-Modal Representation Learning; Embodied AI; Multimodal Learning; Autonomous Reasoning; Self-Supervised Learning; Cognitive Architectures.

I. INTRODUCTION

Artificial General Intelligence (AGI) is defined as the ability of a machine to perform any intellectual task that a human can. Historically, AI development has been siloed into specific domains such as Computer Vision (CV) or Natural Language Processing (NLP). However, human intelligence is inherently multimodal; we do not perceive the world through a single lens but through a continuous fusion of sensory inputs. Multimodal neural networks (MNNs) represent a paradigm shift by attempting to create a unified representation space for these disparate data types. This paper argues that MNNs are not merely an incremental improvement but the essential architectural foundation for AGI.

II. THE MULTIMODAL NATURE OF INTELLIGENCE

Human cognition relies on the ability to form invariant representations of concepts across different sensory modalities. For instance, the concept of "fire" is simultaneously associated with a visual image, the sound of crackling, the warmth of heat, and the linguistic label.

Research in neuroscience suggests that specific neurons in the medial temporal lobe are activated by representations of an object across different modalities [1]. MNNs attempt to replicate this by using cross-modal alignment techniques, allowing a model to "understand" an image through its textual description and vice versa.

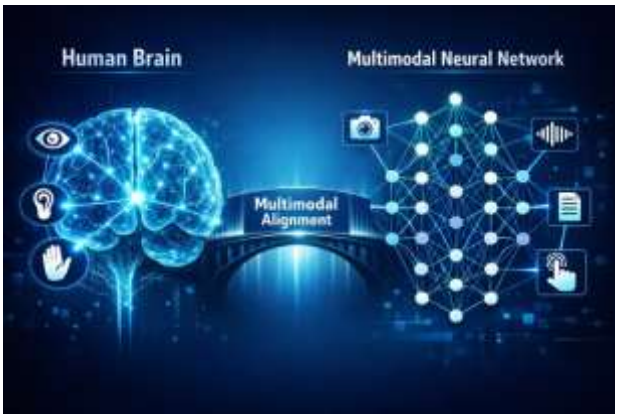

Figure 1: Conceptual Comparison of Human Brain and Multimodal Neural Network

Feature	Narrow AI	Multimodal AI (Foundation Models)	AGI (Target)
Data Input	Single modality (e.g., Text only)	Multiple modalities (Text, Image, Audio)	Universal sensory integration

Generalization	Task-specific	Cross-task within modalities	Autonomous cross-domain adaptation
Learning Style	Supervised	Self-supervised / Foundation-based	Continuous / Life-long learning
Reasoning	Pattern matching	Contextual association	Abstract logic and self-reflection

III. ARCHITECTURAL BREAKTHROUGHS: FROM FUSION TO NATIVE MULTIMODALITY

Early multimodal systems often used "late fusion" techniques, where separate models for vision and language were trained independently and their outputs combined. The current state-of-the-art has moved toward "native" multimodality, where a single transformer- based architecture is trained on interleaved multimodal data from the start.

The Role of Foundation Models

Foundation models like GPT-4V, Gemini, and BriVL (Bridging-Vision-and-Language) have demonstrated that scaling multimodal pre-training leads to emergent properties such as zero-shot reasoning and complex scene understanding [2]. These models leverage "weak semantic correlation" data—unstructured information from the internet—to learn broader associations than traditional human-annotated datasets allowed.

Embodied AI and the Physical World

A critical stepping stone to AGI is the transition from "passive" multimodality (viewing data) to "active" or "embodied" AI. By integrating sensorimotor data, MNNs allow agents to interact with the physical world, bridging the gap between digital intelligence and physical agency [3].

IV. 2026: THE TURNING POINT FOR AGI

As of early 2026, the AI community has witnessed a convergence of multimodal perception and agentic reasoning. Industry leaders and researchers suggest that 2026 marks a
"turning point" where AI systems begin to exhibit human-level performance in complex, multi-step cognitive tasks [4]. The integration of "thoughtful AI"—models that can simulate internal reasoning before acting—represents the latest evolution in the multimodal pipeline.

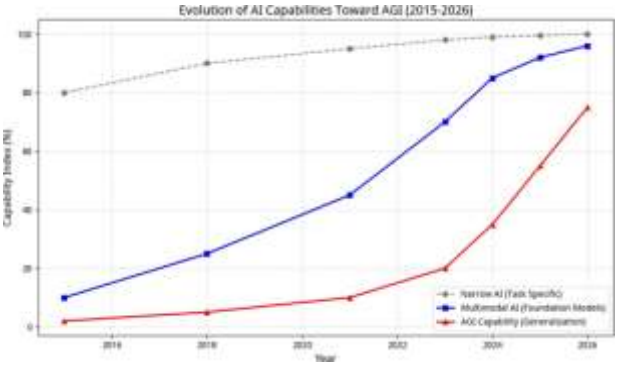

Figure 2: Evolution of AI Capabilities Toward AGI (2015-2026)

V. CHALLENGES AND FUTURE DIRECTIONS

Despite the progress, several hurdles remain: - Computational Efficiency: The energy and hardware requirements for training trillion-parameter multimodal models are immense. - High-Level Reasoning: While perception is largely solved, the ability to perform abstract mathematical or philosophical reasoning without "hallucination" is still a work in progress.

- **Ethical and Safety Frameworks:** As systems approach AGI, the need for robust alignment and safety protocols becomes paramount.

VI. CONCLUSION

Multimodal neural networks are the primary vehicle for the journey toward AGI. By providing a unified framework for perception and association, they lay the groundwork for machines that can understand

the world as humans do. While the "general" in AGI requires further advancements in reasoning and autonomy, the multimodal foundation is now firmly established.

Credit List

The following researchers and institutions have contributed significantly to the field of multimodal AI and AGI, as referenced in this paper:

1. Nanyi Fei, Zhiwu Lu, et al. - For their work on the BriVL foundation model and weak semantic correlation.
2. OpenAI & Google DeepMind Teams - For the development of GPT-4V, Gemini, and the advancement of native multimodal architectures.
3. Gyeong-Geon Lee, et al. - For research into the application of multimodal AGI in specialized domains like education.
4. Stanford Institute for Human-Centered AI (HAI) - For their ongoing analysis and predictions regarding the trajectory of AGI.
5. Y. Wang & A. Sun - For their comprehensive review of embodied AGI and its future directions.

REFERENCES

1. Fei, N., et al. (2022). "Towards artificial general intelligence via a multimodal foundation model." Nature Communications.
2. Lee, G. G., et al. (2023). "Multimodality of AI for Education: Towards Artificial General Intelligence." arXiv.
3. Wang, Y., & Sun, A. (2025). "Toward embodied AGI: A review of embodied AI and the road ahead." arXiv.
4. AI Multiple Research. (2026). "AGI/Singularity: Predictions Analyzed in 2026."
5. McKinsey & Company. (2025). "What is multimodal AI?"