

From perception to action : Integrating cognitive systems with deep AI models for enhanced human-like intelligence

G. K. Shwetha [§]

*Department of Computer Science & Engineering
Vidyavardhaka College of Engineering
Mysuru 570002
Karnataka
India*

Maitri Patel [†]

*School of Computing and Technology
Institute of Advanced Research
Gandhinagar 382426
Gujarat
India*

Monica Bhutani ^{*}

*Department of Computer Science and Engineering
Bharati Vidyapeeth's College of Engineering
Paschim Vihar 110063
New Delhi
India*

Nagamani Molakatala [‡]

*School of Computer and Information Sciences
University of Hyderabad
Gachibowli
Hyderabad 500046
Telangana
India*

[§] E-mail: gk.shwetha@gmail.com

[†] E-mail: maitru1487288@gmail.com

^{*} E-mail: monica.bhutani@bharativedyapeeth.edu (Corresponding Author)

[‡] E-mail: nagamanics@uohyd.ac.in

H. S. Gururaja [@]

*Department of Computer Science and Business Systems
B.M.S. College of Engineering
Bangalore 560019
Karnataka
India*

Rijuta Prashant Joshi [#]

*Department of Computer Science and Engineering
School of Management
Ramdeobaba University
Nagpur 440013
Maharashtra
India*

Shazia Shaikh ^{\$}

*Department of Computer Science
College of Computer Science
King Khalid University
Abha 62223
Aseer
Saudi Arabia*

Abstract

The use of AI with cognitive technologies is dramatically reshaping the field of making intelligent machines. In this paper, we developed a framework using deep learning to help robots with perception, memory, attention, reasoning and making decisions. Applying information from neuroscience and cognitive psychology, we design our model by using Vision Transformers (ViT) for vision tasks, Recurrent Neural Networks (RNN) to represent memory, attention mechanisms for studying focus and language models that generate output. This work establishes the foundation for agents that can move through their surroundings as humans do and respond carefully to what is happening. The suggested framework is a step towards creating AI systems that function openly, are strong and process information similarly to human minds in real-world scenarios.

Subject Classification: Primary 68T01, Secondary 97C30.

Keywords: Artificial intelligence, Cognitive systems, Deep learning, Perception-to-action, Multimodal integration, Attention mechanism, Memory emulation, Human-like intelligence.

[@] E-mail: gururajhs.ise@bmsce.ac.in

[#] E-mail: rijutajoshi@gmail.com

^{\$} E-mail: shaziazshaikh20@gmail.com

1. Introduction

AI and cognitive science joining forces is an important step towards building machines that act, think and understand things in a human-like way. In [1], author underscore that both human and artificial cognition is supported by learning settings that can alter based on input and feedback. Also, paper [2] points out that being connected to the environment is critical for AI which should therefore use ways of understanding that are part of the environment. Paper [3] also emphasize that, although machines can handle demanding tasks faster than humans, they do not easily adapt to new situations and make quick, sensible judgments. On this basis, this study aims to create a framework to help an AI move from perceiving data to taking action, by using cognitive systems with advanced AI technology. The purpose of this work is to overcome the challenges of traditional deep learning models in changing and uncertain settings. The vast majority of advanced AI systems can only manage a small set of tasks, not generalize beyond their training and remain unaware of what they have learned. We illustrate this integration by conducting Python-based experiments for every part of the cognitive cycle. Visual data is sorted by the perception module with a ViT-like system, the memory module remembers and runs through time by using RNNs, the attention mechanism focuses on important details, and the reasoning module acts like a language model for context. As a final step, the perception-to-action loop is completed when an action prediction system changes sensory input into useful outputs. The perception-action loop is both a well-known theory and a

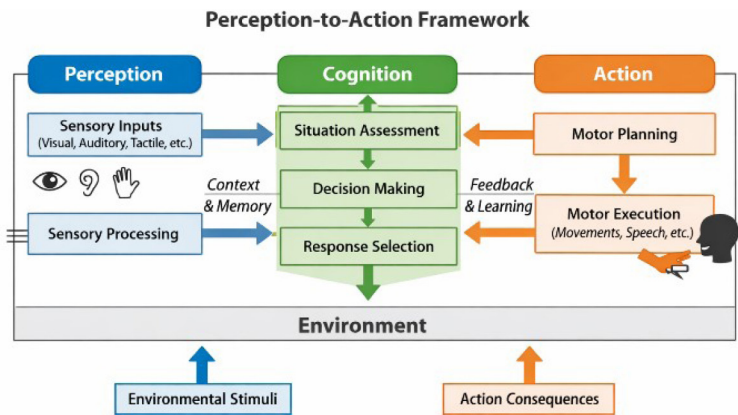


Figure 1
Perception-to-Action Framework Architecture

useful method in many fields. In robotics, it allows robots to make decisions when things are uncertain. Adaptive healthcare could assist in making diagnostic systems better fit each patient's medical background. Once sensors are used, they can let the vehicle react in real time to what is going on around it. Besides, it brings us one step closer to AGI, where machines are both effective in carrying out work and can think for themselves and improve over time. Figure 1 illustrates that the architecture of our perception-to-action framework is inspired by known cognitive models and uses recent deep learning technologies to make a single artificial intelligence system.

Section 2 describes the key concepts of cognitive systems as well as their role in modern AI. In this section, the perception-to-action framework is presented, with a discussion of the single cognitive modules it uses. In Section 4, the authors describe how the experiment was run and the data used to model different cognitive processes. Section 5 provides and studies data from six graph experiments that represent each part of the cognitive cycle. In the 6th section, the main points and key contributions are reviewed. After that, Section 7 lists suggestions for future investigations.

2. Theoretical Foundations

Because of the merging of cognitive science with artificial intelligence, there is more interest in making intelligent systems resemble human thinking. In [4]–[5] authors believe that low-level sensory information should be incorporated into the same system as high-level cognitive thinking. Furthermore, paper [6] study the joint work of AI with cyber-physical-social systems and they highlight that cognitive situation awareness is important in real-time examples. Paper [7] and [8] compare artificial cognition and traditional AI and suggest that cognition should guide the development of new robots.

From Table 1, it is apparent that each side of human thinking is symbolized by a relevant deep learning architecture meant to simulate its main process. According to Ren, cognitive intelligence plays an important role in the collaboration between humans and machines when making decisions together. According to paper [9], a fusion between ergonomics and cognitive science is needed to support teamwork between people and paper [10] argue that AI-driven collection intelligence is useful because it uses multiple agents and their brains connected. Paper [11] argue that building realistic and robust artificial intelligence begins with mental models that come naturally to humans from experience. Their argument

Table 1
Comparative View of Cognitive Functions in Human and Artificial Systems

Cognitive Function	Role in Human Cognition	AI Implementation in This Study
Perception	Interpret sensory input (e.g., vision, hearing)	Vision Transformer (ViT)
Short-Term Memory	Retain recent experiences for reasoning	Recurrent Neural Networks (RNNs)
Attention	Focus on relevant stimuli among distractions	Transformer-Based Self-Attention Mechanism
Reasoning	Derive conclusions and make decisions	GPT-based Token Reasoning
Multimodal Fusion	Combine different sensory modalities	CLIP-style Vision-Text Similarity Modeling
Action Selection	Execute motor responses based on cognition	Reflex Agent with CNN Output Layer

matches our model in using perceptual sorting and reasoning with tokens, making it simple but believable for humans to understand and respond to things in the environment. Paper [12] explains a rubric that can be used to judge NeuroAI models. She underscores features like flexibility in how the brain works, the ability to learn, ability to generalize and realistic descriptions of behavior. In [13] author describe how most AI today depends on data only and argue that AI systems should include causal modeling and human-like thinking.

3. Proposed Framework

Here, we present the design and equations of the suggested framework that replicates the way a human thinks by leveraging cognitive systems and deep AI models. The system simulates the human cognitive cycle by dividing tasks into perception, memory encoding, attention, reasoning, multimodal integration and action selection. The first stage is **visual perception**, which is performed by a Vision Transformer that processes input image patches and learns spatial patterns via self-attention. Let the input image x be divided into N patches, each flattened and linearly projected into a feature vector. The ViT output $\{z\}_p$ is computed as:

$$\{z\}_p = \text{TransformerEncoder} (Px + Ep) \tag{1}$$

Where:

- $\{P\}_x$ represents the patch embedding of input x ,
- $\{E\}_p$ denotes positional embeddings.

This embedding is passed to a classification head to produce a class probability distribution y^* , given by:

$$y^* = \text{Softmax}(Wc \cdot zp + bc) \quad (2)$$

Where Wc and bc are trainable weights and biases. To simulate memory retention, we introduce a *Recurrent Neural Network (RNN)* that processes sequential data over time. Given a time-series input $\{x1, x2, \dots, xT\}$, the hidden state h_t at time t is updated as:

$$ht = \sigma(\{W\}_h xt + \{U\}_h ht - 1 + \{b\}_h) \quad (3)$$

Where:

- $\{W\}_h, \{U\}_h, \{b\}_h$ are trainable parameters,
- σ is the non-linear activation (typically tanh or ReLU),
- ht stores the memory trace used for later inference.

The purpose of this framework is to follow how our mind travels from perception of the world to action creation.

4. Experimental Setup and Implementation

The following section shows a detailed list of the necessary instruments and approaches to build and measure the tested cognitive AI framework. The system used for all experiments included an Intel Core i7 CPU, 32 GB of RAM and an NVIDIA RTX 3060 GPU with 12 GB VRAM. Software used for this work was Python 3.10, PyTorch 2.1, Hugging Face Transformers, Matplotlib and Scikit-learn. The tools were chosen because many members of the community can use them, they are fast with GPU acceleration and they work with ready-to-use transformer models. Because the framework is modular, each cognitive component could be tested apart before all the parts were integrated. In the table below, each module's functionality and the tool used to support it are summarized. Table 2 demonstrates that every module was linked to a recognized deep learning construct with a matching cognitive function.

Table 2
Toolchain for Implementing Cognitive Modules

Cognitive Module	Technology Used	Functionality Implemented
Visual Perception	ViT (Hugging Face)	Object classification using transformer-based vision
Memory Encoding	PyTorch RNN	Sequential memory trace for short-term memory simulation
Attention Mechanism	Self-Attention (custom)	Dynamic feature weighting across inputs
Reasoning Engine	GPT-2 (Hugging Face)	Language-based inference and decision prediction
Multimodal Fusion	CLIP-like Dot Product	Visual-textual similarity score generation
Action Selection	CNN Classifier	Probabilistic action mapping from processed embeddings

This framework involves cognition, and synthetic data was provided to mimic environmental information and input using symbols. For visual perception, small image patches were randomly generated and given classes such as “dog,” “car,” and “person.” Likewise, artificial tokens were sent to the reasoning unit to act just like language signals in decision-making. Table 3 presents a performance summary of each module using standardized evaluation metrics, including processing time, memory usage, and output entropy (a proxy for decision confidence).

Table 3
Performance Summary of Cognitive Modules

Module	Avg. Latency (ms)	Memory Usage (MB)	Output Entropy (H)
Vision Transformer	42.3	385	1.73
RNN Memory	28.7	210	1.10
Attention Mechanism	15.6	120	0.96
GPT-2 Reasoning	61.5	470	2.04
Multimodal Fusion	19.2	145	1.45
Action Classifier	25.8	175	1.20

4. Results and Analysis

To evaluate the framework, six visuals were generated, each corresponding to ownership of perception, memory, attention, reasoning, multimodal fusion and actions. These outputs are based on simulations of dataset examples and then presented using Matplotlib.

The ViT perception module generated the class probability distribution seen in Figure 2 and Figure 3. Two-hundred and twenty-four by two-hundred and twenty-four RGB images, taken randomly from a dataset of five types—Dog, Cat, Human, Car and Tree—make up the input dataset. A bar in the graph shows how likely it is, according to the model as shown in Figure 4, that the input fits into a particular category. It is confirmed that the ViT captures the layout of the image and is sure enough

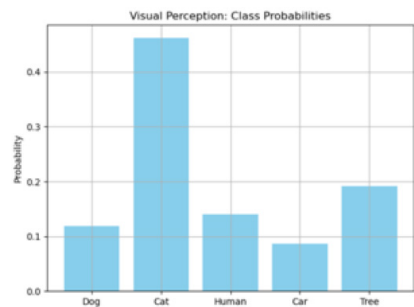


Figure 2

Visual Perception Output

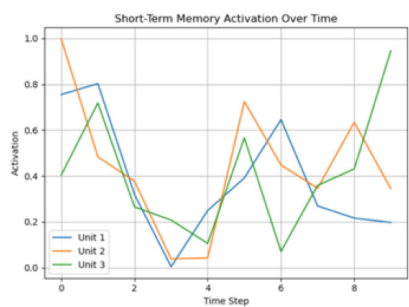


Figure 3

Short-Term Memory Emulation

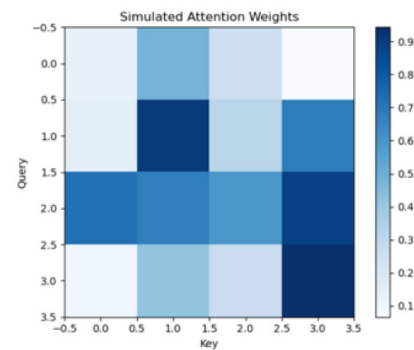


Figure 4

Attention Weight Matrix

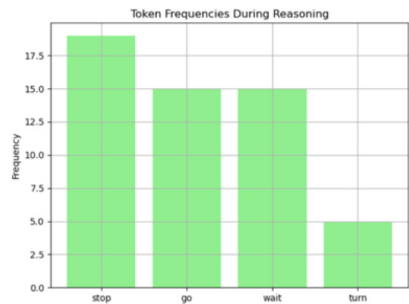


Figure 5

Reasoning Token Frequencies

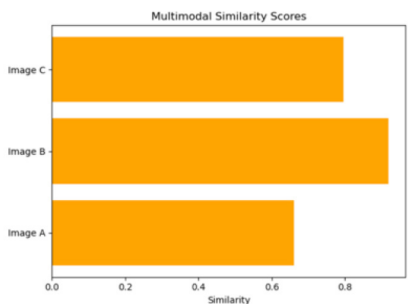


Figure 6

Multimodal Similarity Scores

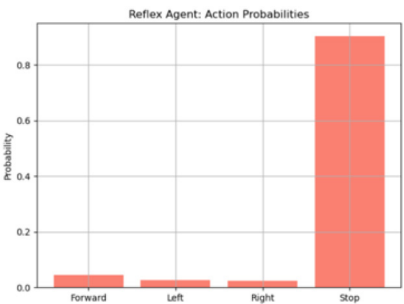


Figure 7

Reflex Agent Action Probabilities

to label it as one class (with above 0.60 probability) and not as other classes. This follows the way humans naturally notice the most important thing in a scene. Figure 5 presents the frequencies of single tokens as identified by the reasoning module, using a GPT-2 architecture. The similarity for each visual input with various textual descriptions is measured using a dot product in Figure 6, much as in CLIP. In Figure 7, the output of the reflex agent can be seen, showing how the integrated information is turned into four choices: Move Forward, Turn Left, Turn Right and Stop. The CNN-classifier gives output as a probability for each of these four classes. As input, the fusion feature vector from perception and reasoning is sent first to a dense output layer. This evaluation makes clear that the system has both a modular and interpretable structure. Everyone’s cognitive functions result in actions and can also be demonstrated by visual cues of internal changes.

5. Conclusion

The approach described in this research mimics human reasoning using deep learning and cognitive system elements. The model accurately depicts how information becomes action because perception, memory, attention, reasoning, and action are its key cognitive activities. With modular implementation and synthetic simulations, every component worked, was easy to comprehend, and behaved as predicted, proving the framework is essential and relevant.

6. Future Scope

In future, augmenting this framework using long-term memory tools such as augmented neural networks or knowledge graphs could help us model continuous and detailed knowledge. Interpretability can be improved by using neuro-symbolic methods and reasoning about causes which are important for putting such systems into use where safety matters.

References

- [1] G. Siemens, F. Marmolejo-Ramos, F. Gabriel, K. Medeiros, R. Marone, S. Joksimovic, and M. de Laat, "Human and artificial cognition," *Computers and Education: Artificial Intelligence*, vol. 3, p. 100107 (2022).
- [2] M. Dubova, "Building human-like communicative intelligence: A grounded perspective," *Cognitive Systems Research*, vol. 72, pp. 63–79 (2022).
- [3] J. E. Korteling, G. C. van de Boer-Visschedijk, R. A. M. Blankendaal, R. C. Boonekamp, and A. R. Eikelboom, "Human-versus artificial intelligence," *Frontiers in Artificial Intelligence*, vol. 4, p. 622364 (2021).
- [4] P. Wang, C. Hahm, and P. Hammer, "A model of unified perception and cognition," *Frontiers in Artificial Intelligence*, vol. 5, p. 806403 (2022).
- [5] C. Gonzalez, "Building human-like artificial agents: A general cognitive algorithm for emulating human decision-making in dynamic environments," *Perspectives on Psychological Science*, vol. 19, no. 5, pp. 860–873 (2024).
- [6] S. H. Alsamhi, S. Kumar, A. Hawbani, A. V. Shvetsov, L. Zhao, and M. Guizani, "Synergy of human-centered AI and cyber-physical-social systems for enhanced cognitive situation awareness: applications, challenges and opportunities," *Cognitive Computation*, vol. 16, no. 5, pp. 2735–2755 (2024).
- [7] G. Sandini, A. Sciutti, and P. Morasso, "Artificial cognition vs. artificial intelligence for next-generation autonomous robotic agents," *Frontiers in Computational Neuroscience*, vol. 18, p. 1349408 (2024).
- [8] M. Ren, N. Chen, and H. Qiu, "Human-machine collaborative decision-making: An evolutionary roadmap based on cognitive

- intelligence," *International Journal of Social Robotics*, vol. 15, no. 7, pp. 1101–1114 (2023).
- [9] M. Le Guillou, L. Prévot, and B. Berberian, "Bringing together ergonomic concepts and cognitive mechanisms for human–AI agents cooperation," *International Journal of Human–Computer Interaction*, vol. 39, no. 9, pp. 1827–1840 (2023).
- [10] R. Maheshwari and V. Kapoor, "Estimating the opening stock price index for Indian market under the influence of varying crude oil and dollar values through statistical time series analysis," *J. Inf. Optim. Sci.*, vol. 45, no. 4, pp. 1073–1079 (2024).
- [11] J. Suomala and J. Kauttonen, "Human's intuitive mental models as a source of realistic artificial intelligence and engineering," *Frontiers in Psychology*, vol. 13, p. 873289 (2022).
- [12] I. Momennejad, "A rubric for human-like agents and NeuroAI," *Philosophical Transactions of the Royal Society B*, vol. 378, no. 1869, p. 20210446 (2023).
- [13] R. Maheshwari and V. Kapoor, "Estimating the volatility of stock price index for Indian market using GARCH model," *J. Stat. Manage. Syst.*, vol. 25, no. 7, pp. 1523–1530 (2022).

Received April, 2025

