Cognitive Convergence: Exploring Artificial General Intelligence (AGI): Perspectives and Multimodal Model Applications

Ms. Priya Godbole

Assistant Professor, Department of Computer Science, Dr. S. C. Gulhane Prerna College of Commerce, Science, and Arts, Nagpur, MS. (India)

Abstract **Industry revolution has emerged as being significantly powered by artificial intelligence (AI) in past decades. Artificial general Intelligence (AGI) adapts the concepts from deep learning to simulate human cognitive abilities into systems for smart manufacturing. Artificial General Intelligence (AGI) represents a significant advancement in the field of artificial intelligence, embodying the capability to understand, learn, and apply knowledge across a wide range of tasks similar to human intelligence. The convergence of AGI with multimodal models, combining information from various sensory modalities, holds transformative potential across industries such as neuroscience, healthcare, biomedicine, entertainment, robotics, and more. This paper aims to provide the perspective of AGI and its fundamental components, and the integration of AGI into multimodal models.**

Keywords- Artificial General Intelligence (AGI), Multimodal Models, Deep Learning, Artificial Neural Networks, Self-Supervised Learning, Large language models (LLMs), Human-like Intelligence, Cognitive Abilities, Computer vision

I. INTRODUCTION

The evolution of AI over the past few years has been remarkable but real intelligence that can be used in a variety of contexts is still exclusive. The phrase "narrow AI" refers to the development of systems that perform particular "intelligent" actions in particular scenarios, according to Ray Kurzweil [1]. For a restricted AI system, some degree of human upgrading or reconfiguration is typically required to enable the system to preserve its level of intelligence if the context or the behavior specification is even a little changed. Recent years have seen the emergence of a somewhat diverse community of researchers dedicated to the explicit interest in AGI, as demonstrated by conference series like AGI [2], BICA [3], and Advances in Cognitive Systems [4], as well as a large number of special tracks and symposia on Human-Level Intelligence [5], Integrated Intelligence [6] and related topics.

High-level cognitive abilities like creativity, abstract reasoning, and problem-solving are also important features of the human brain [7]. Humanity has been working toward building artificial general intelligence (AGI) systems since the middle of the 20th century that are capable of reasoning, problem-solving, and creativity and have intelligence comparable to or even greater than that of a human. Alan Turing and other pioneers established early concepts about computers and their ability to emulate human cognition in the 1940s [8].

AGI, often known as "strong AI," is capable of generalized cognitive functions such as the generalized transfer of learned knowledge from one context to others and goes beyond narrow task-specific AI. In numerous AI research domains, including computer vision (CV) and natural language processing (NLP), deep learning has gained outstanding success. Deep residual networks (ResNets) [9], for instance, have previously outperformed humans at image classification. RoBERT [10] a language model has also performed better than people in several natural language tasks. Relationship networks [11] developed by DeepMind have superhuman performance on a dataset for relational reasoning. The study of attention has long been a focus of research in the domains of psychology and neuroscience, and its application to artificial intelligence greatly accelerates the development of AGI. The "Transformer" model, which is based on a self-attention mechanism of Artificial neural networks, has served as the inspiration for several cutting-edge artificial neural networks, including BERT [12] and GPT [13]. The Vision Transformer (ViT) [14] model, which represents an image as a series of patches, exhibits state-of-the-art performance in a variety of computer vision (CV) tasks by incorporating self-attention processes into image processing.

The capacity to apply knowledge and abilities in many contexts or disciplines, as well as the ability to synthesize information from other domains or modalities, are important indicators of cognitive intelligence. The incorporation of AGI with multimodal models, which can receive and comprehend data from various sensory inputs including text, images, and audio, is a crucial development that has the potential to significantly change the field of AI applications. The correlation between two different modalities is often modeled in pre-training data by multimodal (visual and textual) foundation models [15][16], which typically accept image-text pairings as input.

Despite encouraging outcomes on quick learning/transfer and cross-modal comprehension tasks from existing multimodal foundation models, most of them [15][16][17] assume that there is a high semantic association between the input image-text pairs (such as image-caption pairings) and anticipate that there will be exact matches between the objects or regions in a picture and the words in the text. The most recent multimodal foundation models [15][16] often make use of object detectors to extract significant picture regions and use a single-tower network design to more accurately mimic the fine-grained region–word matching. We create the Bridging-Vision and Language (BriVL) model by self-supervised learning [17-21] using a great deal of multimodal data to address the aforementioned problems. Modeling weak semantic correlation data by image-text matching as compared to modeling strong semantic correlation data by direct image-to-text "translation" in previous works [15] would help us obtain a more cognitive model. Despite merely being pre-trained with an image-text matching learning target, our BriVL has already met some of the essential requirements for an AGI system due to its great generalization capabilities [22].

II. CHARACTERISTICS OF AGI

A. *Scalable*: Large language models (LLMs) represent some of the earliest models that show human-level performance across a range of tasks.[23] For LLMs like the GPT-2 and GPT-3, there is a correlation between the number of neurons and cognitive ability. It has been revealed that GPT-3 outperforms humans on several benchmarks for natural language processing, including tasks involving question-answering, language translation, and text completion.[24] Its size and ability to process natural language have made it a potent tool for a variety of uses, such as chatbots, content creation, and language translation. It will be interesting to watch how the link between the number of parameters and cognitive ability changes as scholars work to advance LLMs and improve AGI.

B. *Multimodality and Interdisciplinary Composition*: The ability of the human brain to simultaneously process and combine data from various sensory modalities is truly remarkable. The ability to perceive and understand their surroundings through a variety of sensory inputs, including sight, sound, touch, taste, and smell, is provided by this remarkable quality. Additionally, the skillful management of multimodal inputs enables people to make more accurate and comprehensive assessments of their environment, promoting smart communication and connection with others. The smart acquisition of knowledge from various modalities therefore has the potential to improve human cognitive ability. GPT-4 not only exhibits a high degree of proficiency in a variety of fields, including literature, medicine, law, mathematics, the physical sciences, and programming, but it also demonstrates an exceptional understanding of complex ideas by combining abilities and concepts from several fields with utility.

C. *Text-to-image and image-to-text generation*: Among the most well-known models for handling picture descriptions (image-to-text generation) and text-to-image generation tasks are CLIP [25], DALL-E [26] and their successor GLIDE [27], VisualGPT [28] and Diffusion [29]. A pre-training technique called CLIP trains distinct picture and text encoders and learns to anticipate which photos in a dataset are connected to specific descriptions. Notably, CLIP possesses multimodal neurons that activate when exposed to both the classifier label text and the associated image, comparable to the "Halle Berry" neuron in humans [30], indicating a fused multimodal representation. The generated images in GLIDE, an extension of DALL-E, are created using a diffusion model, although CLIP is still used to rank them [25]. The VisualGPT is the development of GPT-2 from a single language model to a multimodal model with a self-resurrecting activation unit that gives sparse activations to avoid mistakenly overwriting linguistic knowledge.

D. *Visual question answering*: A key use of multimodal learning is visual question answering, which calls for a model to accurately answer a text-based query based on an image. The vision encoder, text encoder, multimodal fusion, and decoder modules can be built using a variety of sub-architectures using METER, a basic framework for training effective end-to-end vision-language transformers. The Unified Vision-Language pretrained Model (VLMo) [31], uses a modular transformer network to concurrently learn a dual encoder and a fusion encoder. A shared self-attention layer and a pool of modality-specific experts are present in each network block, providing a great deal of flexibility for tuning.

E. *Video-language modeling* In 2021, Microsoft's Project Florence-VL released ClipBERT [32], a transformer model that combines a Convolutionary Neural Network (CNN) with minimally sampled frames. It is end-to-end optimized to handle common video-language problems. Masked Visual token Modeling and Sparse Attention have been added in later iterations of ClipBERT, such as VIOLET [33] and SwinBERT [34], to advance current techniques in video question answering video retrieval, and video captioning.

F. *Multimodal learning with auditory data*: The latest development from Meta AI, Data2vec [35], offers a novel self-supervised learning framework that does not require conventionally labeled data. The Kosmos-1 [36] large language model from Microsoft handles text, visual, and aural input. It understands general modalities and exhibits contextual learning and instruction following using multimodal web-based corpora. Its skills include language comprehension, image captioning, answering visual questions, and image identification, demonstrating the ability for cross-modal transfer, which makes it easier for knowledge to be shared between language and multimodal inputs. For instance, GPT-4 performs better in textual tasks than ChatGPT because it incorporates multimodality [37].

G. *Alignment*: Although a fact that several LLMs, like as BERT [12], GPT [23] GPT-2 [37] GPT-3, and Text-to-Text Transfer Transformer (T5) [38], have excelled at particular tasks, they nevertheless fall short of real AGI due to their capacity to perform unexpected behaviors. Reinforcement learning from human feedback (RLHF) has been used recently in large language models (LLMs) like Sparrow, InstructGPT, ChatGPT, and GPT-4 to address the problem of alignment with human instructions

III. TECHNOLOGICAL ADVANCES IN AGI

The main strategies that language models, like LLMs, rely on are zero-shot prompting, few-shot prompting, in-context learning, and instruction.

A. *In-context learning*: In the sense of AGI, in-context learning refers to the model's ability to perceive and carry out new tasks by giving it a finite set of input-output pair examples [39] within prompts or simply by giving it a task description. While in-context learning shows parallels to explicit fine-tuning at the prediction, representation, and attention behavior levels, prompts help the model understand the task's structure and patterns. This lessens the possibility of overfitting downstream labeled training data and enables to generalize and execute new tasks more efficiently without additional training or fine-tuning [40]. Recent developments in large-scale AGI models, particularly GPT-4, have shown such an intriguing capacity.

B. *Prompt and instruction tuning*: In many downstream applications, the pre-trained models can achieve zero-shot learning owing to the prompt and instruction tuning-based techniques [41]. Producing accurate and safe outputs based on instructions is a necessary condition for AGI models to perform at a level comparable to that of humans. Untrue and harmful outputs must be effectively managed as these models are employed increasingly frequently. Leading the way in this regard is InstructGPT. Supervised training is carried out with the aid of human-provided demonstrations and prompts to enhance the caliber of model outputs. Following that, humans compile and grade the various models' results according to their merit. The models are further improved using RLHF [42], a method that uses human preferences as rewards to direct the learning process.

C. *Evolution of AGI*: AGI is capable of adapting to new situations, transferring domain knowledge, and displaying human-like cognitive abilities beyond streamlined and formatted task-solving workflows in the current literature, in contrast to "narrow AI", which is designed to perform specific tasks.[43][44] Overall, AGI might exhibit exceptional adaptation and versatility. Though actual AGI has not yet been achieved by science, advances in artificial intelligence and its subfields, such as deep learning, have set the stage for future research and the pursuit of AGI.

D. *Deep learning and modern AGI*: AI has made significant progress through the development of deep learning, which was made possible by ground-breaking improvements in computing power and the accessibility of enormous datasets. AGI is getting closer to being a reality due to advancements in computer vision, natural language processing, and reinforcement learning. The birth of pre-trained language models, including BERT [12] and its various domain-specific variants, larger models like GPT-3 [45], and vision transformer (ViT) based models in computer vision, which revolutionized language modeling by utilizing self-attention mechanisms. Although these models are not yet AGI, they are a huge step in the right direction. A variety of use cases, including essay writing, question answering, search, translation, data augmentation, computer-aided diagnosis, and data de-identification, have been deployed using ChatGPT's chatbot interface, which allowed millions of users to interact with AI in a more natural way [45]. Advanced math and logical thinking are both possible using ChatGPT. Additionally, the model does exceptionally well on common tests like the GRE, LSAT, and USMLE [46] GPT-4 is expected to address a previously unresolved variety of issues and has a wide range of applications. Its development is evidence of the significant advancements made in the search for AGI.

E. The Architecture of AGI: Werbos's invention of the back-propagation algorithm in 1975, which is used in artificial neural networks [47], revolutionized the field by making it possible to effectively train neural networks with multiple layers, such as the perceptron. Deep learning has gained popularity because of advancements in technology, including the creation of graphics processing units (GPUs) and tensor processing units (TPUs), which have made it possible to train deep neural networks effectively. This advancement has stepped up the study and development of AGI by enabling the creation of more potent neural networks that can handle more difficult tasks. GPT-3 was trained to utilize large-scale distributed training over several GPUs and consumed a significant number of computational resources and energy, training a GPT model demands strong hardware and parallel processing techniques. Distributed computing methods are required for creating AGI models like GPT-4. TensorFlow, PyTorch, and Horovod are examples of distributed computing frameworks that make it easier to deploy these strategies, even though the precise distributed computing systems used to train GPT models may not be publicly known [48]. These frameworks allow researchers and developers to divide the training process among several devices, control device synchronization, and communication, and effectively utilize the available computational resources

IV. LIMITATIONS OF AGI

The creation of novel machine learning techniques, such as more effective teaching techniques, in-context learning algorithms, and reasoning paradigms, is necessary for the development of AGI. Brain-inspired AI methods try to provide computers the ability to learn from unstructured data without having to label it and to quickly generalize from a small number of instances, which is essential for giving computers the ability to learn and adapt to new tasks and contexts. The development of AGI will have moral and societal effects to take into account, including bias, privacy, and security concerns. It is crucial to make sure that AGI is created and applied in a way that benefits society as a whole and is consistent with human values as it grows more potent and prevalent.

CONCLUSION

AGI's ability to replicate human cognitive versatility across a multitude of tasks holds the promise of revolutionizing industries such as healthcare, entertainment, education, and beyond. By transcending the limitations of task-specific AI, AGI infuses machines with the capacity to learn, reason, and adapt in a manner reminiscent of human thought processes. The convergence of AGI and multimodal models symbolizes not just technological advancement, but the fusion of human and machine cognition to shape a future where AI serves as a truly versatile collaborator. In the Industry revolution, AGI's role as a transformative tool is examined with its convergence with multimodal models for manufacturing smart tools. AGI, with its potential to propel innovation and redefine societal paradigms, necessitates a harmonious balance between technological prowess and ethical considerations

REFERENCES

[1] R. Kurzweil, "The singularity is near When humans transcend biology", Penguin, 2005.

[2] S. Adams, I. Arel, J. Bach, R. Coop, R. Furlan, B. Goertzel, J. S. Hall, A. Samsonovich, M.Scheutz, M. Schlesinger, M.; et al., "Mapping the landscape of human-level artificial general intelligence. AI Magazine vol. 33(1), 2012, pp. 25-42.

[3] J. S. Albus, "Engineering of mind: An introduction to the science of intelligent systems" Wiley, 2001.

[4] N. Alvarado, S. Adams, S. Burbeck, and C. Latta, "Beyond the Turing test: Performance metrics for evaluating a computer simulation of the human mind", In The 2nd International Conference on Development and Learning, 2002, pp. 147–152, IEEE.

[5] N. Alvarado, S. Adams, S. Burbeck, and C. Latta, "Beyond the Turing test: Performance metrics for evaluating a computer simulation of the human mind", In The 2nd International Conference on Development and Learning, 2002, pp. 147–152, IEEE.

[6] J. R. Anderson, and C. Lebiere, "The Newell test for a theory of cognition. Behavioral and Brain Sciences", vol. 26(05), 2003, pp. 587–601.

[7] K. Teffer, K. Semendeferi, "Human prefrontal cortex: evolution, development, and pathology", Prog Brain Res. Vol. 195, 2012, pp. 191–218.

[8] "Turing AM", Computing Machinery and Intelligence. Springer; 2009.

[9] K. He, X. Zhang, S. Ren, & J. Sun, "Deep residual learning for image recognition", In IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 770–778.

[10] Y. Liu et al., "Roberta: A robustly optimized Bert pretraining approach", arXiv preprint arXiv:1907.11692, 2019.

[11] A. Santoro et al., "Simple neural network module for relational reasoning", In Advances in Neural Information Processing Systems, 2017, pp. 4967–4976.

[12] J. Devlin, MW. Chang MW, K. Lee, K. Toutanova, "Bert: pre-training of deep bidirectional transformers for language understanding", arXiv preprint arXiv:181004805, 2018.

[13] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, "Improving language understanding by generative pre-training", 2018.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov et al., "An image is worth 16x16 words: transformers for image recognition at scale', arXiv preprint arXiv:201011929, 2020.

[15] X. Li, et al., "Oscar: Object-semantics aligned pre-training for vision-language tasks", In European Conference on Computer Vision, 2020, pp. 121–137.

[16] A. Radford, et al., "Learning transferable visual models from natural language supervision', In International Conference on Machine Learning, 2021, pp. 8748–8763.

[17] J. Lin, J. et al., "A Chinese multimodal trainer", arXiv preprint arXiv:2103.00823, 2021.

[18] Z. Wu, Y. Xiong, S. Yu, and D. Lin, D, "Unsupervised feature learning via nonparametric instance

discrimination", In IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3733–3742.

[19] A. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding", arXiv preprint arXiv, 2018, pp. 1807.03748.

[20] R. D. Hjelm, et al. "Learning deep representations by mutual information estimation and maximization", International Conference on Learning Representations 2019.

[21] C. Zhuang, A. L. Zhai, & D. Yamins. "Local aggregation for unsupervised learning of visual embeddings", In International Conference on Computer Vision, 2019, pp. 6002–6012.

[22] Fei, Nanyi, et al. "Towards artificial general intelligence via a multimodal foundation model." Nature Communications 13.1 2022.

[23] Y. Liu, T. Han, S Ma, et al. "Summary of chatgpt /gpt-4 research and perspective towards the future of large language models", arXiv preprint arXiv, vol. 230401852, 2023.

[24] T. Brown, B. Mann, N. Ryder, et al. "Language models are few-shot learners", Adv Neural Inf Process Syst, vol. 33, 2020, pp. 1877–1901.

[25] A. Radford, JW Kim, C. Hallacy, et al. "Learning transferable visual models from natural language supervision", International Conference on Machine Learning, PMLR, vol. 8748–8763, 2021.

[26] A. Ramesh, M. Pavlov, G. Goh, et al. "Zero-shot text-to-image generation. International Conference on Machine Learning", PMLR, 2021, pp. 8821–8831.

[27] A. Nichol, P. Dhariwal, A. Ramesh, et al. "Glide: towards photorealistic image generation and editing with text-guided diffusion mode", arXiv preprint arXiv, vol. 211210741, 2021.

[28] J. Chen, H. Guo, K. Yi, B. Li, M. Elhoseiny. "VisualGPT: data-efficient adaptation of pre-trained language models for image captioning", In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18030–18040.

[29] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer. "High-resolution image synthesis with latent diffusion models", In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10684–10695.

[30] RQ Quiroga, L Reddy, G Kreiman, C Koch, I Fried. "Invariant visual representation by single neurons in the human brain", Nature, vol. 435, no. 7045, 2005, pp. 1102–1107.

[31] H. Bao, W. Wang, L. Dong, et al. "Vlmo: unified vision-language pre-training with mixture-of- modality-experts", Adv Neural Inf Process Syst, vol. 35, 2022, pp. 32897–32912.

[32] J. Lei, L. Li, L. Zhou, et al. "Less is more: clipbert for video-and-language learning via sparse sampling", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 7331–7341.

[33] TJ Fu, L. Li, Z. Gan, et al. "Violet: end-to-end video-language transformers with masked visual-token modelling", arXiv preprint arXiv, 2021, vol. 211112681.

[34] K. Lin, L. Li, CC Lin, et al. "Swinbert: end-to-end transformers with sparse attention for video captioning", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 17949–17958.

[35] A. Baevski, WN Hsu, Q. Xu, A. Babu, J. Gu, M. Auli. "Data2vec: a general framework for self-supervised learning in speech, vision and language", International Conference on Machine Learning, PMLR, 2022, pp. 1298–1312.

[36] S. Huang, L. Dong, W. Wang, et al. "Language is not all you need: aligning perception with language models", arXiv preprint arXiv, vol. 230214045, 2023.

[37] S. Bubeck, V. Chandrasekaran, R. Eldan, et al. "Sparks of artificial general intelligence: early experiments with GPT-4", arXiv preprint arXiv, vol. 230312712. 2023.

[38] C. Raffel, N. Shazeer, A. Roberts, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer", J Mach Learn Res, vol. 21, no. (1), 2020, pp. 5485–5551.

[39] H. Dai, Z. Liu, W. Liao, et al. "ChatAug: leveraging ChatGPT for text data augmentation", arXiv preprint arXiv, vol. 230213007, 2023.

[40] D. Dai, Y. Sun, L. Dong, Y. Hao, Z. Sui, F. Wei. "Why can GPT learn in context? Language models secretly perform gradient descent as Meta optimizers", arXiv preprint arXiv, vol. 221210559, 2022.

[41] V. Sanh, A. Webson, C. Raffel, et al. "Multitask prompted training enables zero-shot task generalization", arXiv preprint arXiv, vol. 211008207, 2021.

[42] PF Christiano, J. Leike, T. Brown, M. Martic, S. Legg, D. Amodei. "Deep reinforcement learning from human preferences", Adv Neural Inf Process Syst, vol. 30, 2017.

[43] B. Goertzel. "Artificial general intelligence: concept, state of the art, and prospects", Journal of Artificial General Intelligence, vol. 5, no. (1):1, 2014.

[44] H. Hodson. "DeepMind and Google: the battle to control artificial intelligence", The Economist, 2019, pp. 13–613, 2019.

[45] Z. Liu, X. Yu, L. Zhang, et al. "DeID-GPT: zero-shot medical text de-identification by GPT-4", arXiv preprint arXiv, vol. 230311032, 2023.

[46] OpenAI, GPT-4 Technical Report. 2023.

[47] P. Werbos. "Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences", PhD Thesis, Committee on Applied Mathematics. Cambridge, MA: Harvard University; 1974.

[48] Lin Zhao, Lu Zhang, Zihao Wu, Yuzhong Chen, Haixing Dai, Xiaowei Yu, Zhengliang Liu, Tuo Zhang, Xintao Hu, Xi Jiang, Xiang Li, Dajiang Zhu, Dinggang Shen, Tianming Liu. "When brain-inspired AI meets AGI", Meta-Radiology vol. 1, no. 100005, 2023.