Deconstructing Every AI Theorical Topic in the Context of Multimodality and Reconstructing System-Level AI from These New Findings

Yiming Huang*
Faculty of Information Technology, Beijing University of Technology
(Dated: 2022.5.1)

I. INTRODUCTION & BACKGROUND

Artificial Intelligence(AI) originated in the middle of the last century. With a series of landmark events like the theory of Alan Turing and meeting at Dartmouth College, making a kind of algorithm, model, or system to imitate humans even reach the intelligence of humans comes to serious discussion. From that time to the beginning of our century, Deep learning finally beats the Symbolism of Machine Learning and its success reveals that **auto-adapting structure with learnable parameters is more "intelligent"** to some degree. Meanwhile, system-level planning and controlling meet it thus born a new paradigm of learning: Reinforce Learning. As Deep Learning goes further in its application of computer vision(CV) and natural language processing(NLP), without doubt, the natural idea of combining these two modalities comes to researchers' and developers' minds. It is true that the intelligent body process multi-modalities at the same time, thus **multi-modal deep learning is a closer way to intelligence. To be brief, we researchers need to discuss AI in multimodal circumstances**. It is the precondition of my main future work.

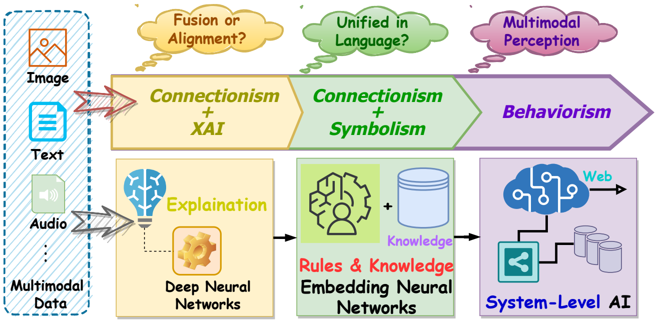

FIG. 1. Illustration of my research proposal. Coincidentally, my research interest in three periods just happens to be the three typical paradigms (Symbolism, Connectionism, and Behaviorism) in the AI field.

II. MY SHORT-TERM RESEARCH INTEREST

On the other hand, deep learning is often criticized as an uninterpretable black-box system that provides neither a sufficient guarantee to users nor basic principles of why it reaches such intelligence. However, in my opinion, a simple fact indicates that we can not easily analyze deep neural networks in math equations. That fact is Modern Biology does not give extinct explanations of how the human brain works and how intelligence is formed. My further idea is **that the current pure-mathematical analysis gives an explanation of parameterized function, but not intelligence**. On the contrary, using networks to explain networks might be a potential way for exploring, recent work like using GPT-4 to explain GPT-2 [1] and probing networks in NLP are enlightening to this idea. Considering different modalities, fusion, and alignment between modalities is also complicated. A readily available answer provided by the Creator is our brain, neural science might be the reference to our topic. In addition, the science of language and psychology might also contain some cues. **My short-term research interest will focus on this direction—Explainable AI in multimodal case**. I believe exploring this topic can obtain a deeper insight into intelligence.

III. MY MIDDLE-TERM RESEARCH INTEREST

In recent months, large language models(LLM) like ChatGPT [2] shows their extraordinary capabilities in many areas. Some researchers believe its intelligence is the so-called "emergence" of the huge system [3], but I also associate a simple fact to argue. That is we human beings learn far less corpus than ChatGPT, so true intelligence might not be such a huge costing model. An intuitive idea is to embed know-based memory in the model to replace the pretraining. This idea is inspired by inheritance, which is the "pretraining" of us human beings. **This proposed knowledge-embedding model/system will be dominated by the modality of language**, due to the concrete symbols in language modality containing more information and pre-embedding knowledge mainly forming by language. **The abovementioned is my Middle-Term research interest. In another saying, I plan to use knowledge embedding to enhance/replace pre-training/finetuning/prompts of the so-called multimodal big model today**.

* huangyiming2002@126.com

IV. MY LONG-TERM RESEARCH INTEREST

Until now, famous works in the deep-learning area can be mainly attributed to the end-to-end model which can be regarded as a kind of parameterized function. However, recent works like VISPROG [4] and Generative Agents [5] remind us researchers that future AI shall not be a simple function, but an integrated system! Like the "environment" concept in Reinforce Learning, the huge World Wide Web is an excellent environment for us to train an AI system. **If we can link our AI system to the Web through APIs, it can learn from the multimodal corpus with in-time real feedback(quite a few will be produced by people online). How intelligent will this system be!** With the aim to learn this biggest "dataset", I believe that simple RL paradigms and stacked Transformer blocks are not competent enough. Only as sufficient exploring demonstrated in Sec. II and Sec. III done, the new finding might inspire us to develop such a system. **Therefore, my research interest in the long term is to accumulate enough profound insights to build this AI system, which could be the prototype of Artificial Life(AL)**.

[1] OpenAI, Language models can explain neurons in language models, https://openai.com/research/language-models-can-explain-neurons-in-language-models (2023).

[2] OpenAI, Chatgpt: A large language model, https://openai.com/research/chatgpt (2022).

[3] R. Schaeffer, B. Miranda, and S. Koyejo, Are emergent abilities of large language models a mirage?, CoRR **abs/2304.15004**, 10.48550/arXiv.2304.15004 (2023), 2304.15004.

[4] T. Gupta and A. Kembhavi, Visual programming: Compositional visual reasoning without training, CoRR **abs/2211.11559**, 10.48550/arXiv.2211.11559 (2022), 2211.11559.

[5] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, Generative agents: Interactive simulacra of human behavior, CoRR **abs/2304.03442**, 10.48550/arXiv.2304.03442 (2023), 2304.03442.