Continuum AI: Integrating Foundational AI Agents with the Cognitive Computing Continuum

Lauri Lovén, Center for Ubiquitous Computing, University of Oulu, Finland.
lauri.loven (at) oulu.fi.

***Abstract*—In this vision paper, we introduce Continuum AI, a novel approach to integrating foundational AI agents across the AI-driven computing continuum. By leveraging the seamless interconnection of diverse computing resources, from edge devices to high-performance computing clusters, Continuum AI aims to address the challenges associated with the efficiency, scalability, and sustainability of AI systems. We discuss the current status of foundational models and the AI-driven computing continuum, highlighting existing challenges and opportunities. Furthermore, we explore the research challenges and anticipated benefits of incorporating foundational AI agents within the AI-driven computing continuum, ultimately paving the way for a new era of efficient, robust, and secure AI solutions.**

I. Motivation

The rapid advancements in artificial intelligence (AI) have brought forth a wide array of AI models and agents, revolutionizing various domains such as natural language processing, computer vision, and reinforcement learning. In particular, the development of foundational AI models, such as the Transformer architecture [1], has enabled impressive performance gains in multiple AI tasks. However, as AI models grow larger and more complex, their computational requirements and energy consumption have increased significantly [2], raising concerns regarding the efficiency, scalability, and sustainability of AI systems.

At the same time, the AI-driven computing continuum, which refers to the seamless integration of various computing resources ranging from edge devices to high-performance computing clusters [3], offers a promising avenue for addressing these concerns. The AI-driven computing continuum aims to deploy diverse AI models and agents across this spectrum of resources, optimizing their performance, scalability, and adaptability[4]. By doing so, AI models can become more context-aware and efficient, dynamically allocating resources based on task requirements, data availability, and other constraints [5].

This vision paper introduces the concept of Continuum AI, which seeks to populate the AI-driven computing continuum with foundational AI models that together optimize resource usage and pursue user-set goals. By integrating foundational AI agents across the computing continuum, Continuum AI aims to address the challenges associated with the efficiency, scalability, and sustainability of AI systems while unlocking new possibilities for the future of artificial intelligence.

The following sections will delve into the current status of foundational models and the AI-driven computing continuum, highlighting existing challenges and opportunities for Continuum AI. Additionally, we will discuss the research challenges and anticipated benefits in integrating foundational AI agents within the AI-driven computing continuum, paving the way for a new era of efficient, robust, and secure AI solutions.

II. Current Status

A. Foundational Models

Foundational models are large-scale AI models that serve as the building blocks for various AI applications and tasks. These models have made significant strides in recent years, particularly in the areas of natural language processing [6], computer vision [7], and reinforcement learning [8].

Transformers, a class of neural networks, have proven to be especially successful in handling a wide range of tasks, such as language translation, sentiment analysis, and text summarization [1]. Furthermore, pre-trained models like BERT [6] and GPT [9] have laid the groundwork for fine-tuning on specific tasks, thus reducing training time and resources.

Despite these successes, foundational models still face challenges, such as the need for vast amounts of data and computational resources for training [2], the potential for biases [10], and the lack of interpretability [11].

B. AI-Driven Cognitive Computing Continuum

The computing continuum refers to the seamless integration of various computing resources, ranging from edge devices and local servers to cloud infrastructure and high-performance computing clusters [3]. In the context of AI, the AI-driven computing continuum aims to deploy diverse AI models and agents across this spectrum of resources, optimizing their performance, scalability, and adaptability.

This approach enables AI models to be more context-aware and efficient, as they can dynamically allocate computing resources based on the task requirements, data availability, and other constraints [5]. For instance, AI models deployed on edge devices can offer low-latency and privacy-preserving solutions, while models leveraging cloud resources can handle more computationally intensive tasks.

Recent developments in the AI-driven computing continuum include the emergence of federated learning, which enables distributed AI models to collaboratively learn from decentralized data sources while preserving privacy [12]. Additionally, advancements in AI accelerators and specialized hardware, such as GPUs, TPUs, and FPGAs, have led to more efficient AI model training and inference [13].

Despite these advancements, several challenges remain in the AI-driven computing continuum, including the efficient allocation and management of resources, seamless interaction among AI models and agents, and ensuring the privacy and security of user data. By integrating foundational AI agents across the computing continuum, Continuum AI aims to address these challenges and unlock new possibilities for the future of artificial intelligence.

III. Research Challenges

The integration of foundational AI agents within the AI-driven computing continuum brings forth several research challenges. Addressing these challenges will yield significant benefits, enabling the development of more efficient, robust, and secure AI solutions.

Efficient Allocation and Management of Resources. One of the primary research challenges in Continuum AI is the development of efficient mechanisms for allocating and managing computing resources across the AI-driven computing continuum. This includes devising strategies to dynamically adapt resource usage based on task requirements, data availability, and other constraints [5].

Solving this challenge will enable more efficient AI model training and deployment, potentially reducing energy consumption, lowering costs, and improving model performance. This can pave the way for AI systems that can adapt to new challenges and environments with minimal human intervention.

Seamless Interaction Among AI Models and Agents. Another critical research challenge is facilitating seamless interaction and collaboration among diverse AI models, agents, and systems across the computing continuum. This requires the development of novel algorithms, communication protocols, and standards to ensure interoperability and knowledge sharing among AI models and agents [14].

Overcoming this challenge will foster the development of more robust and versatile AI solutions, as models can leverage the strengths of various agents to address complex tasks. This collaboration can expedite the transfer of knowledge and expertise among AI models, potentially reducing the time and effort needed for training and fine-tuning.

Scalability and Generalization Across the Continuum. Ensuring the scalability and generalization of AI models across the computing continuum is a crucial research challenge. This entails devising methods to facilitate the development of AI models that can easily scale across tasks, domains, and environments, while promoting broader applicability and improved generalization capabilities [15].

Addressing this challenge will pave the way for the development of AI systems that can adapt to new challenges and environments with minimal human intervention. It will enable the creation of AI solutions that are more versatile and capable of handling diverse tasks and domains.

Privacy and Security in the AI-Driven Computing Continuum. In an interconnected and distributed AI ecosystem, ensuring the privacy and security of user data is a critical research challenge. This involves leveraging techniques such as federated learning [12], differential privacy, and secure multiparty computation (MPC) techniques to develop AI models that adhere to stringent privacy and security requirements.

Overcoming this challenge will enable the development of AI systems that are not only powerful but also preserve user privacy and ensure data security. This can foster increased trust in AI solutions and promote their adoption in sensitive domains such as healthcare, finance, and government.

In conclusion, addressing these research challenges in Continuum AI will unlock numerous benefits, including enhanced efficiency, improved collaboration, scalability, and privacy and security considerations. By fostering a collaborative and adaptable AI ecosystem across the computing continuum, we can pave the way for more efficient, robust, and secure AI solutions that will shape the future of artificial intelligence.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[2] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in nlp," *arXiv preprint arXiv:1906.02243*, 2019.

[3] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud computing and emerging it platforms: Vision, hype, and reality for delivering computing as the 5th utility," *Future Generation computer systems*, vol. 25, no. 6, pp. 599–616, 2009.

[4] H. Kokkonen, L. Lovén, N. H. Motlagh, J. Partala, A. González-Gil, E. Sola, I. Angulo, M. Liyanage, T. Leppänen, T. Nguyen *et al.*, "Autonomy and intelligence in the computing continuum: Challenges, enablers, and future directions for orchestration," *arXiv preprint arXiv:2205.01423*, 2022.

[5] M. Satyanarayanan, P. Bahl, R. Caceres, and N. Davies, "The case for vm-based cloudlets in mobile computing," *IEEE pervasive Computing*, vol. 8, no. 4, pp. 14–23, 2009.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[8] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

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

[10] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," *Advances in neural information processing systems*, vol. 29, 2016.

[11] D. Gunning and D. Aha, "Darpa's explainable artificial intelligence (xai) program," *AI magazine*, vol. 40, no. 2, pp. 44–58, 2019.

[12] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *arXiv preprint arXiv:1610.05492*, 2016.

[13] N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers *et al.*, "In-datacenter performance analysis of a tensor processing unit," in *Proceedings of the 44th annual international symposium on computer architecture*, 2017, pp. 1–12.

[14] D. H. Wolpert and K. Tumer, "An introduction to collective intelligence," *arXiv preprint cs/9908014*, 1999.

[15] J. Hernández-Orallo, "Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement," *Artificial Intelligence Review*, vol. 48, pp. 397–447, 2017.