Brain Cognition Driven Large Model and Small Model Integrated Robot for Dynamic Medical Analysis

Daobin Luo[1,2], Yuru Liu[1,2], Zhanyu Hu[1,2], Baoyu Zhang[1,2], Weishan Zhang[1,2] * Tao Chen[1,2], Shaohua Cao[1,2], Fei-Yue Wang[3], Xiaoli Li[4], Zengxiang Li[5]

[1]College of Computer Science and Technology, China University of Petroleum (East China)
[2]Shandong Key Laboratory of Intelligent Oil & Gas Industrial Software, China University of Petroleum (East China)
[3]Institute of Automation, Chinese Academy of Sciences, China
[4]Singapore University of Technology and Design, Singapore
[5]SingHealth Duke-NUS AI in Medicine Institute (AIMI), Singapore
luodaobin2001@gmail.com, liuyuru2021@163.com, huzhanyu0328@163.com, b22070002@s.upc.edu.cn, zhangws@upc.edu.cn, dongshou2015@s.upc.edu.cn, shaohuacao@upc.edu.cn, feiyue.wang@ia.ac.cn, xlli@i2r.a-star.edu.sg, shawn.li.zx@singhealth.com.sg

Abstract

Medical robots encounter substantial challenges in accommodating continuously changing data and dynamic task scenarios. To address this, in this demo, we design a **M**edical **Robot** driven by **B**rain **C**ognition based large model and small model Collaboration (BC-MRobot) for dynamic medical data analysis. BC-MRobot constructs a bidirectional optimization mechanism that achieves collaborative training between large model and small model. Furthermore, BC-MRobot incorporates an intelligent model factory that achieves data-aware model adaptation and customization, which can enhance its capabilities for handling ever changing data features in different analyzing tasks. BC-MRobot currently accommodates 4 data modalities encompassing 27 disease categories, which can be used for interactive medical analysis where different analyzing models can be generated dynamically at run time, with high analysis accuracy. **Video Link:** https://upcbdipt.github.io/#paper/bc-mrobot.

Introduction and Related Work

Medical robot for healthcare and eldercare services are attracting increasing attentions (Wang 2020; Asada et al. 2025; Min, Lai, and Ren 2025). As large models evolving quickly, which show remarkable capabilities in reasoning and decision-making, which provides us new opportunities to design a more intelligent robot(Lu et al. 2023; Fan et al. 2024; Teng et al. 2025). However, current medical robots face some critical limitations: (1) Centralized training may lead to data leakage risks. (2) Significant variations in medical data make robot difficult to adapt to heterogeneous data. (3) Low intelligence to adapt to dynamic scenarios and evolving data distributions, with low capabilities for continuous learning.

Federated learning provides new ideas for protecting data privacy. Personalized federated learning using hypernetwork can effectively deal with data heterogeneity problems (Scott, Zakerinia, and Lampert 2024). Furthermore, brain cognition has been igniting artificial intelligence since this discipline started. Research shows that increasing the endogenous potential of the hippocampus to generate new neurons throughout life rejuvenates learning and memory (Danciu et al. 2025), demonstrating that neurogenesis provides crucial neural plasticity for continuous learning. There exists research that uses such brain cognition ideas to integrate capabilities of large models and classical small models (Liu et al. 2025) to handle the evolving and changing data features. Ye et al. optimized human-machine interaction processes through a cognition-driven approach, which can be applied to medical robot design (Ye et al. 2022).

Therefore, based on our former work(Li et al. 2019) (Liu et al. 2025), this paper proposes a **M**edical **Robot** driven by **B**rain **C**ognition based large model and small model Collaboration (BC-MRobot) for dynamic medical data analysis in distributed medical environment. Our contributions are summarized as follows:

- A bidirectional collaborative optimization mechanism between large model and small model in federated learning scenarios is designed, which forms a closed loop for bidirectional knowledge transferring.
- We design an intelligent model factory to quickly customize personalized diagnostic models for specific medical tasks, which enables continuous improvement of intelligence of robots, and can enhance its capabilities of adaptation to evolving medical tasks.

BC-MRobot Overview

BC-MRobot adopts a cloud-edge-device architecture, where the cloud server accommodates large model and hypernetworks. The medical institutions participate in collaborative training from the edge side, and robots execute onsite analysis. Fig 1.(a) demonstrates a bidirectional collaborative optimization process between large model and small model. The model factory shown in Fig 1.(b) is designed for data-aware model customization and continuous model evolution.

*Corresponding author

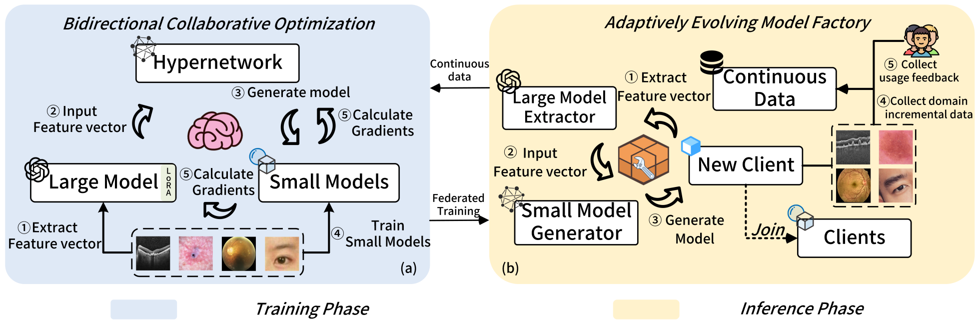

Figure 1: Overview of the design of BC-MRobot working mechanisms.

Bidirectional Collaborative Optimization BC-MRobot leverages the brain's self-reflection mechanism to construct a bidirectional optimization method between large and small models, where a high dimensional feature extractor is designed to extract high-level representations, a hypernetwork is designed to generate model weights and architectures, and small models are applied to specific tasks. The large model feature extractor comprises a large model with task-specific head removed and integrated LoRA adaptive module.

During federated training, the large model feature extractor performs deep feature embedding on data to extract representation vectors. Hypernetwork learns mapping relationships between representation vectors and model parameters to dynamically generate small model weights adapted to specific data characteristics, enabling knowledge transfer from large model to small model. Based on training gradients of small model, BC-MRobot updates the LoRA module parameters and hypernetwork weights by applying backpropagation, forming a bidirectional optimization loop between large model and small model.

Adaptively Evolving Model Factory BC-MRobot incorporates an adaptively and continuously evolving intelligent model factory, which utilizes the large model feature extractor and hypernetwork to provide dynamic model customization services. When data distribution changes, the large model feature extractor re-extracts features from client data, while the hypernetwork generates customized small model weights instantly based on these feature vectors. This allows newly joined medical institutions to directly customize small models without the need for re-training. Furthermore, the model factory periodically collects usage feedback and incremental data from medical institutions to train the large model feature extractor and hypernetwork parameters, achieving continuous model evolution.

Demonstration Logistics

We will show how BC-MRobot performs federated collaborative training and continuous model evolution capabilities when facing changing data and tasks, as well as its application in medical data analysis. We simulate the federated training process on an Ubuntu server equipped with dual RTX 5090 GPUs. A Kuavo 4PRO MaxB Robot is used for application demonstration. In this demo, Lingshu-7B is used as the multimodal medical diagnosis large model, and CNN models are employed as the diagnostic small models. Different number of layers and parameters will be generated dynamically by the model factory. More details can be found from the demonstration video via the video link.

BC-MRobot Platform The BC-MRobot platform has the following functionalities:

- Federated collaborative training management encompasses the configuration, initiation, and visualization of training tasks. Intuitive visualization of the bidirectional optimization process is designed to show the trajectory of training curves. The model factory achieves continuous model evolution through periodic federated collaborative training.
- Dynamic model customization services provide personalized model customization tailored to client-specific data distributions. We demonstrate adaptive evolution capabilities of the model factory by comparing personalized model with the generic model. The personalized models are more accurate for inference.

BC-MRobot for medical data analysis BC-MRobot identifies a user who waves to it, and automatically navigates to his location, and starts conversations. The user's medical images are matched through his identity. It can capture examination areas directly using its onboard cameras. BC-MRobot supports analyzing of images of fundus, retinal OCT , dermoscopy, ocular appearance, etc. After obtaining images, the robot performs preliminary diagnosis using its local small model, and uploads the images and preliminary diagnostic results to the cloud-based large model for higher accurate analysis. Finally, the system performs multimodal fusion analysis based on images, small model diagnostic results, and consultation processes to provide comprehensive diagnoses, and give further healthcare suggestions.

Conclusion

In this demo, we propose BC-MRobot, designing a novel bidirectional collaborative optimization mechanism between large and small models and a model factory enabling personalized model customization and continuous evolution. This demonstration showcases promising prospects of brain cognition based large and small models collaborative optimization for medical robots.

References

Asada, Y.; Horiguchi, T.; Yagi, K.; Komatsu, M.; Yamashita, A.; Ueta, R.; Yamaaki, N.; Shikida, M.; Nishio, S.; and Inagaki, M. 2025. Robot-Assisted Approach to Diabetes Care Consultations: Enhancing Patient Engagement and Identifying Therapeutic Issues. *Medicina*, 61(2): 352.

Danciu, D.-P.; Klawe, F. Z.; Kazarnikov, A.; Femmer, L.; Kostina, E.; Martin-Villalba, A.; and Marciniak-Czochra, A. 2025. Unraveling regulatory feedback mechanisms in adult neurogenesis through mathematical modelling. *npj Systems Biology and Applications*, 11(1): 82.

Fan, L.; Guo, C.; Tian, Y.; Zhang, H.; Zhang, J.; and Wang, F. 2024. Sora for foundation robots with parallel intelligence: three world models, three robotic systems. *Frontiers Inf. Technol. Electron. Eng.*, 25(7): 917–923.

Li, L.; Wang, X.; Wang, K.; Lin, Y.; Xin, J.; Chen, L.; Xu, L.; Tian, B.; Ai, Y.; Wang, J.; Cao, D.; Liu, Y.; Wang, C.; Zheng, N.; and Wang, F. 2019. Parallel testing of vehicle intelligence via virtual-real interaction. *Sci. Robotics*, 4(28).

Liu, Y.; Zhang, W.; Bao, Z.; Chai, X.; Gu, M.; Jiang, W.; Zhang, Z.; Tian, Y.; and Wang, F. 2025. Brain-Like Cognition-Driven Model Factory for IIoT Fault Diagnosis by Combining LLMs With Small Models. *IEEE Internet Things J.*, 12(16): 32296–32309.

Lu, Y.; Guo, C.; Dou, Y.; Dai, X.; and Wang, F. 2023. Could ChatGPT Imagine: Content Control for Artistic Painting Generation Via Large Language Models. *J. Intell. Robotic Syst.*, 109(2): 39.

Min, Z.; Lai, J.; and Ren, H. 2025. Innovating robot-assisted surgery through large vision models. *Nature Reviews Electrical Engineering*, 1–14.

Scott, J.; Zakerinia, H.; and Lampert, C. H. 2024. PeFLL: Personalized Federated Learning by Learning to Learn. In *ICLR*.

Teng, S.; Wang, Y.; Wang, X.; Li, J.; Li, Y.; Zhang, X.; Li, L.; Chen, L.; and Wang, F. 2025. The ParallelWorkforce: A Framework for Synergistic Collaboration in Digital, Robotic, and Biological Workers of Industry 5.0. *IEEE Trans. Comput. Soc. Syst.*, 12(4): 1627–1636.

Wang, F. 2020. Parallel Healthcare: Robotic Medical and Health Process Automation for Secured and Smart Social Healthcares. *IEEE Trans. Comput. Soc. Syst.*, 7(3): 581–586.

Ye, P.; Wang, X.; Zheng, W.; Wei, Q.; and Wang, F.-Y. 2022. Parallel cognition: Hybrid intelligence for human-machine interaction and management. *Frontiers of Information Technology & Electronic Engineering*, 23(12): 1765–1779.