Resident Health Agent Personalization via Collaborative Edge Fine-Tuning

Leo Zeng
Centre for Perceptual and Interactive Intelligence
Hong Kong, China

Abstract

The aging population represents a growing demographic trend globally, with many countries witnessing a rise in the number of older adults. This shift has spurred the development of accessible infrastructures and increased the demand for healthcare services, particularly for those who wish to receive care at home [6]. Among the innovative solutions emerging to address this need, resident health agents stand out as prominent machine-learning-driven applications designed for domestic settings. These agents enable local collection and analysis of sensor data, facilitating personalized health monitoring. Despite their widespread adoption in various smart health applications, including elderly care [1, 5] and Alzheimer's disease monitoring [2], tailoring these agents to meet individual user profiles and specific application requirements remains a significant challenge.

Traditional cloud offloading approaches that utilize abundant cloud computing resources for model refinement require uploading raw data and models to external servers, which unavoidably raises privacy concerns [4]. An alternative approach focuses on on-device fine-tuning, keeping both the model and data locally, but this method is constrained by the limited computational power of edge devices. Cloud-based federated learning offers a privacy-preserving training mechanism [3]; however, it still relies on the processing capabilities of the device itself, which can be a limiting factor.

To address these limitations, we leverage the observation that families typically possess multiple edge devices at home with underutilized computing resources. By pooling these devices, we can augment the available computing resources for personalized agent model fine-tuning. For instance, as illustrated in Fig. 1, a smart home might include a smartphone, desktop computer, tablet, and smartwatch. Upon deployment, we select an appropriate subset of devices as training participants, considering the trade-off between the computing resources they offer and the communication overhead involved in data transmission. With the selected devices, we devise a workload partitioning strategy to map model layers to devices. To meet memory requirements, we employ pipeline parallelism to coordinate the parallel training process across edge devices. During runtime, we initiate the collaborative fine-tuning process to train a personalized health agent until convergence.

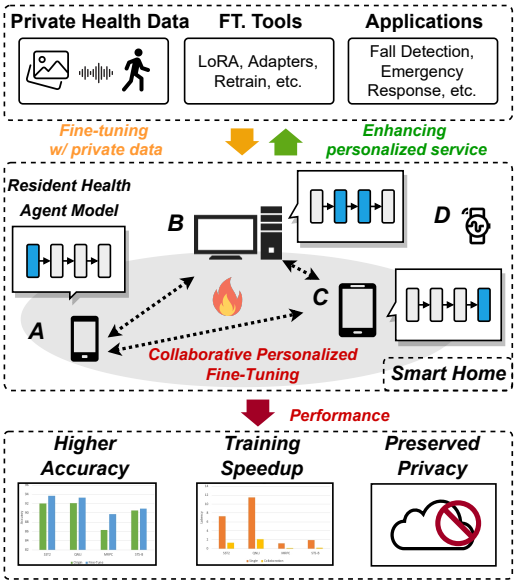

Figure 1: Overview of the proposed solution.

Preliminary evaluations in sentiment analysis and natural language inference show that collaborative edge fine-tuning enjoys the following performance benefits, as in Fig. 1. First, by fine-tuning resident health agents toward dedicated users, the model accuracy is upgraded at most 1.89, resulting in personalized services of improved satisfaction. Second, by leveraging more edge devices for resource augmentation, the fine-tuning process obtains up to 3.2x speedups, making agent personalization a much more agile procedure. Third, collaborative edge fine-tuning prunes the reliance on cloud servers, which avoids cloud-related privacy leakage from the basis. In the future, we intend to refine the collaboration scheduler design to maximize the edge resource utilization for further speedup. Besides, we are also deploying the system to our real-world healthcare testbed and will collect more real-world measurements for system verification.

Acknowledgments

This work is supported by the Centre for Perceptual and Interactive Intelligence (CPII) Ltd under the Innovation and Technology Commission (ITC)'s InnoHK.

References

[1] Manasa Kalanadhabhatta, Adrelys Mateo Santana, Lynnea Mayorga, Tauhidur Rahman, Deepak Ganesan, and Adam Grabell. 2024. Multi-stakeholder Perspectives on Mental Health Screening Tools for Children. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–15.

[2] Xiaomin Ouyang, Xian Shuai, Yang Li, Li Pan, Xifan Zhang, Heming Fu, Sitong Cheng, Xinyan Wang, Shihua Cao, Jiang Xin, et al. 2024. ADMarker: A Multi-Modal Federated Learning System for Monitoring Digital Biomarkers of Alzheimer's Disease. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*. 404–419.

[3] Xiaomin Ouyang, Zhiyuan Xie, Heming Fu, Sitong Cheng, Li Pan, Neiwen Ling, Guoliang Xing, Jiayu Zhou, and Jianwei Huang. 2023. Harmony: Heterogeneous multi-modal federated learning through disentangled model training. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*. 530–543.

[4] Brian Testa, Yi Xiao, Harshit Sharma, Avery Gump, and Asif Salekin. 2023. Privacy against real-time speech emotion detection via acoustic adversarial evasion of machine learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 3 (2023), 1–30.

[5] Karine Tung, Steven De La Torre, Mohamed El Mistiri, Rebecca Braga De Braganca, Eric Hekler, Misha Pavel, Daniel Rivera, Pedja Klasnja, Donna Spruijt-Metz, and Benjamin M Marlin. 2022. BayesLDM: A Domain-specific Modeling Language for Probabilistic Modeling of Longitudinal Data. In *2022 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*. IEEE, 78–90.

[6] Yi Xiao, Harshit Sharma, Zhongyang Zhang, Dessa Bergen-Cico, Tauhidur Rahman, and Asif Salekin. 2024. Reading between the heat: Co-teaching body thermal signatures for non-intrusive stress detection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 4 (2024), 1–30.