Developing Multimodal Healthcare Foundation Model: From Data-driven to Knowledge-enhanced

Xiaochen Wang

The Pennsylvania State University
xcwang@psu.edu

Abstract

Foundation models in general domains have leveraged multimodal knowledge graphs to great effect, yet the healthcare sector lacks such comprehensive structures, presenting a significant gap in current research. Based on previous exploration with pure data-driven approaches, this proposal describes a two-stage project aiming to enhance multimodal healthcare foundation model with domain knowledge. The first stage is to construct a robust multimodal healthcare knowledge graph based on established healthcare taxonomies, such as UMLS, and enriched with data from multimodal clinical databases like MIMIC-CXR. This knowledge graph will incorporate medical images as cross-modal instances linked to healthcare terminologies, enhancing the depth and applicability of the graph. In the second stage, the knowledge graph will serve as a foundational tool in training healthcare foundation models with enhanced capabilities, particularly in reducing hallucination and managing concept ambiguity through the novel use of reinforcement learning techniques like Direct Preference Optimization (DPO). This research is expected to make significant contributions to the domain of healthcare AI by enabling more accurate, reliable, and explainable AI-driven diagnostics and interventions.

Introduction

Foundation models have revolutionized artificial intelligence, exhibiting profound success across numerous domains. Trained on vast datasets, these models excel in tasks like language understanding, image interpretation, and reasoning, showcasing their ability to leverage pre-trained knowledge effectively, especially in situations with minimal data. This capability has significantly enhanced efficiency and scalability across various industries.

Motivated by these successes, the healthcare research community has begun to explore the potential of healthcare foundation models. These models are designed to manage diverse healthcare modalities and address a broad range of downstream tasks, potentially transforming medical AI by providing comprehensive solutions for clinical applications. Nonetheless, the development of these models confronts substantial challenges due to the intrinsic complexities of healthcare data, particularly its knowledge-intensive nature.

Addressing medical tasks necessitates a deep reservoir of domain-specific knowledge (Wang et al. 2024a). Compared to their general-domain counterparts, healthcare foundation models require a broad spectrum of knowledge covering various subdomains. However, most existing approaches mechanically adopt the developmental pipelines used for general foundation models (Nazi and Peng 2024). Although some efforts incorporate knowledge-rich texts, such as medical textbooks and healthcare publications, the predominant training methodologies still follow general autoregressive patterns, without sufficient focus on integrating dense, structured knowledge.

Furthermore, policy regulations that limit the sharing of sensitive healthcare data underscore its decentralized nature. This constraint drastically diminishes the volume of data available for training effective healthcare foundation models, thus heightening the necessity to leverage substantial domain-specific knowledge as a compensatory measure.

Given the inherent complexities of healthcare data, including its temporality, high dimensionality, and multimodality, integrating well-structured knowledge into healthcare foundation models presents significant challenges. Achieving this integration requires a thorough exploration of directly utilizing multimodal healthcare data and delving into a wide array of domain-specific knowledge. These exploratory efforts form the core of the subsequent sections on previous and future work, providing a comprehensive framework for this research proposal.

Previous Work

My initial efforts in developing a multimodal healthcare foundation model centered on purely data-driven methods, devoid of any domain-specific knowledge integration. In 2023, I devised a targeted pretraining strategy to enhance healthcare foundation models using multimodal Electronic Health Records (EHRs) (Wang et al. 2023). This approach enabled the model to concurrently assimilate heterogeneous healthcare data, exploiting intricate correlations across various modalities and hierarchical levels within EHRs. Following this, a subsequent study focused on pretraining the healthcare foundation model with data from diverse sources, addressing the privacy concerns mentioned earlier (Wang et al. 2024c).

A pivotal shift from data-driven to knowledge-enhanced

methodologies characterizes my recent research, as evidenced by studies accepted at NeurIPS 2024 (Wang et al. 2024b) and EMNLP 2024 (Wang et al. 2024d). In the first of these, I introduced a clear problem formulation and a comprehensive benchmark for the task of federated knowledge injection. Here, the foundation model benefits from implicit knowledge sourced from distributed clients, which is parameterized through modality-specific encoders optimized during training with local datasets. In the latter study, I employed a Mix-of-Experts (MoE) mechanism alongside the Parameter-efficient Fine-tuning (PEFT) technique, enabling the foundation model to flexibly address various healthcare tasks across different modalities.

Despite these advancements, the knowledge remains implicitly parameterized, unverified by domain experts, and lacks explicit explainability—an essential aspect in healthcare. Building on the success of knowledge-enhanced, specialized healthcare AI initiatives (Luo et al. 2024), my forthcoming objective is to enrich the multimodal healthcare foundation model with explicit and explainable domain knowledge, thereby enhancing its clinical applicability and transparency.

Future Work

Construction of Multimodal Healthcare Knowledge Graph. A multimodal knowledge graph is a complex data structure that encapsulates extensive information about multimodal entities and their interrelationships. In general domains, multimodal knowledge graphs have been utilized to either explicitly develop or implicitly enhance foundation models. However, a multimodal healthcare knowledge graph, which would serve as an ideal basis for developing knowledge-enhanced healthcare foundation models, is currently lacking. This absence highlights a critical research gap that needs to be addressed to facilitate further exploration. Therefore, over the next few months, I will focus on establishing a multimodal healthcare knowledge graph. This graph will be constructed using well-established healthcare taxonomies, such as UMLS (Humphreys and Lindberg 1993), and will leverage images from existing multimodal clinical databases like MIMIC-CXR (Johnson et al. 2019). These images will be incorporated as cross-modal instances linked to healthcare terminologies, with further clarification of the relationships between these instances and the corresponding terminologies. The integration of these elements will ensure the quality and comprehensiveness of the knowledge and data included in the new graph. This comprehensive knowledge graph will be utilized in my subsequent research throughout my PhD candidacy.

Reinforcement-based Multimodal Knowledge Injection for Hallucination Mitigation. Following the establishment of the multimodal healthcare knowledge graph, the next step is to devise a methodology that enables healthcare foundation models to benefit from this resource. While existing research (Lee et al. 2024) has shown success in integrating multimodal knowledge graphs into foundation models in general domains, these approaches may fall short in addressing specific healthcare challenges such as the low tolerance for hallucination, concept ambiguity, and complex interrelations among concepts. In this study, I will focus on using the established graph to train a healthcare foundation model that exhibits reduced hallucination. Reinforcement learning techniques, such as Direct Preference Optimization (DPO), will be employed to discourage the model from generating responses that are inconsistent with the knowledge graph.

References

Humphreys, B. L.; and Lindberg, D. 1993. The UMLS project: making the conceptual connection between users and the information they need. *Bulletin of the Medical Library Association*, 81(2): 170.

Johnson, A. E.; Pollard, T. J.; Berkowitz, S. J.; Greenbaum, N. R.; Lungren, M. P.; Deng, C.-y.; Mark, R. G.; and Horng, S. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1): 317.

Lee, J.; Wang, Y.; Li, J.; and Zhang, M. 2024. Multimodal Reasoning with Multimodal Knowledge Graph. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 10767–10782. Bangkok, Thailand: Association for Computational Linguistics.

Luo, J.; Wang, X.; Wang, J.; Chang, A.; Wang, Y.; and Ma, F. 2024. CoRelation: Boosting Automatic ICD Coding Through Contextualized Code Relation Learning. *arXiv preprint arXiv:2402.15700*.

Nazi, Z. A.; and Peng, W. 2024. Large language models in healthcare and medical domain: A review. In *Informatics*, volume 11, 57. MDPI.

Wang, J.; Luo, J.; Ye, M.; Wang, X.; Zhong, Y.; Chang, A.; Huang, G.; Yin, Z.; Xiao, C.; Sun, J.; et al. 2024a. Recent Advances in Predictive Modeling with Electronic Health Records. *arXiv preprint arXiv:2402.01077*.

Wang, J.; Wang, X.; Lyu, L.; Chen, J.; and Ma, F. 2024b. FEDMEKI: A Benchmark for Scaling Medical Foundation Models via Federated Knowledge Injection. *arXiv preprint arXiv:2408.09227*.

Wang, X.; Luo, J.; Wang, J.; Yin, Z.; Cui, S.; Zhong, Y.; Wang, Y.; and Ma, F. 2023. Hierarchical pretraining on multimodal electronic health records. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2023, 2839. NIH Public Access.

Wang, X.; Luo, J.; Wang, J.; Zhong, Y.; Zhang, X.; Wang, Y.; Bhatia, P.; Xiao, C.; and Ma, F. 2024c. Unity in Diversity: Collaborative Pre-training Across Multimodal Medical Sources. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3644–3656.

Wang, X.; Wang, J.; Xiao, H.; Chen, J.; and Ma, F. 2024d. FEDKIM: Adaptive Federated Knowledge Injection into Medical Foundation Models. *arXiv preprint arXiv:2408.10276*.