Foundation Models in Medical Imaging: A Survey

Anonymous CVPR submission

Paper ID *****

Abstract

Foundation models, a class of pre-trained deep learning architectures, have gain huge popularity in the field of medical imaging. These models, pre-trained on large scale of data, have been successfully adapted to address the unique challenges posed by medical images. In this paper, we look into the evolution, taxonomy, characteristics, and applications of foundation models in the context of medical imaging. Our chronological exploration provides insights into their future development,limitations and impact on clinical workflows.

1. Introduction

Medical imaging plays a critical role in disease diagnosis, treatment planning, and patient management. However, the scarcity of annotated medical image data and the complexity of various imaging modalities hinder the direct application of conventional deep learning models. Foundation models, pre-trained on large-scale, diverse datasets, offer a promising solution. These models capture generic features from non-medical domains, which can be fine-tuned for specific medical tasks. Let us explore the key aspects of foundation models and their significance in medical imaging research.

1.1. The Foundation Model Paradigm

Foundation models represent a departure from task-specific architectures. Unlike traditional deep learning models designed for specific applications, foundation models are pre-trained on extensive non-medical data. This pre-training phase equips them with a rich understanding of general features, such as edges, textures, and object representations. Consequently, foundation models serve as a knowledge base that can be fine-tuned for specialized tasks, including medical image analysis.

1.2. Transfer Learning and Adaptation

The power of foundation models lies in their transferability. By leveraging pre-existing knowledge, these models adapt swiftly to new domains. In medical imaging, where labeled data is often scarce, foundation models provide a head start. Researchers can fine-tune them using smaller, domain-specific datasets, tailoring their representations to suit specific clinical challenges. This transfer learning paradigm bridges the gap between general features and medical context.

2. Current Challenges

Heterogeneous Data Integration: Each imaging modality captures different aspects of the human body, leading to heterogeneous data representations. Integrating these diverse data sources in a coherent and informative manner remains a significant challenge. Different modalities often provide complementary information about the same underlying pathology. Effectively leveraging this complementary information for improved diagnostic and prognostic insights is an ongoing area of exploration.

3D Images: CT, MRI

2D Images: X-Ray, ultrasound, camera images (endoscope, fundus, dermoscope, etc.) ,Pathology

Task-Specific Requirements: Medical tasks such as image classification, anomaly detection, and disease prognosis demand specialized model architectures to address the intricacies of each task, making it challenging to create a universal model that performs optimally across all tasks.

Clinical Trust and explainability: Clinicians require transparent and interpretable AI systems to confidently incorporate AI-generated insights into their decision-making processes. Lack of interpretability can hinder the widespread adoption of AI in clinical settings.

3. Taxonomy

There many ways to categorize foundation models in medical imaging. Like in Paper [1], they decided FM into two main categories: Textually Prompted Models (TPMs) and Visually Prompted Models (VPMs) based on the architecture of the models. Also, considering the function of the models, we can divide them into conversational models, segmentation models, registeration models, etc. [11]

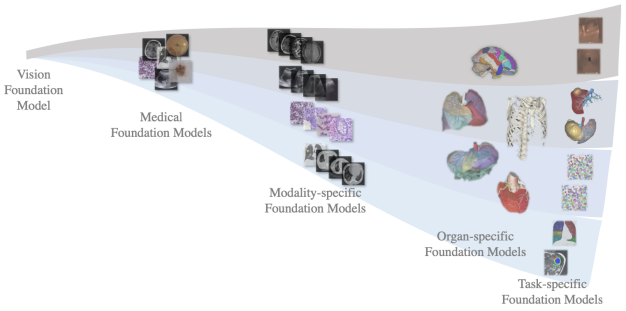


Figure 1. The taxonomy of foundation models in medical image analysis.

- **General Vision Models:** These models are trained on large-scale datasets and serve to bridge differences between various imaging modalities. They exhibit broad context reasoning and generalization abilities, making them suitable for multiple medical imaging tasks.
- **Modality-Specific Models:** Tailored to a specific imaging modality (e.g., X-rays, MRI, CT), these models excel within their designated modality but may lack generalization capabilities.
- **Organ/Task-Specific Models:** These models are trained for specific anatomical structures or medical tasks, such as lung segmentation, lesion diagnosis, or clinical parameter estimation. They perform exceptionally well on their designated tasks.
- **Hybrid Models:** Combining the strengths of both general vision models and modality-specific models, hybrid models offer a balance of generalization and task-specific performance.

4. LLM in medical imaging

In recent years, there has been a significant development in the field of large language models. These models, based on deep learning techniques, aim to comprehend and generate human language. Typically comprising billions or even hundreds of billions of parameters, large language models demonstrate exceptional performance in natural language processing tasks, including text generation, machine translation, question answering systems, and summarization.

Representative works in the realm of large language models include OpenAI's GPT series (such as GPT-4), Meta's LLama model, and Google's Gemini model, among others. These models have achieved significant milestones within their respective domains, propelling substantial advancements in the field of natural language processing.

Lately many scientist are working on to apply LLM to image tasksThere many ways to do thisfor example, designing a conversational model to get text prompts and output clinical results, or finding some ways to connect vision and text embeddings.

Method	Base-model
BioBert	Bert
BioMedGPT	GPT
PMC-LLaMA	LLama
Radiology-Llama2	LLama2
ClinicalGPT	GPT
XrayGPT [6]	GPT
ChatDoctor [3]	GPT
DeID-GPT [4]	GPT-4

Table 1. Overview of text-prompt foundation models in medical imaging

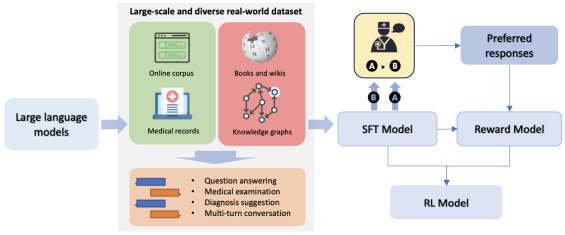


Figure 1: The overview of ClinicalGPT.

Figure 2. A overview of ClinicalGPT

4.1. Conversational models

Considering the huge computational resources required to pretrain large language models, some works focusing on collecting high quality sft datasets to finetune pretrained LLMs(Mostly LLama family), some works using this method are Radiology-LLAMA2 [5], PMC-LLama [9], ClinicalGPT [7].

Their ability to swiftly generate coherent and clinically relevant reports could be transformative, particularly in busy radiology departments where timely and accurate reporting is of the essence. And they could be developed into a conversational assistant that helps doctors in real-time. This would enable a more dynamic interaction, where the model could assist in tasks ranging from quick data retrieval to offering second opinions on diagnoses.

4.2. Multimodal models

Other method(ChatCAD [8],BiomedGPT [10]) focusing on using other domain-specific models to tackle with different clinical data, and then use LLM to intergrate all the results and get final reports.

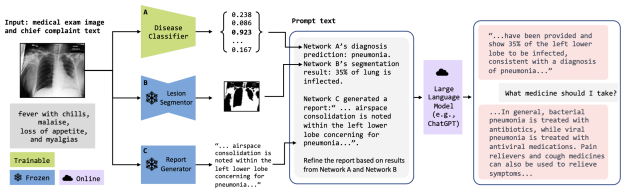


Figure 3. A overview of ChatCAD

5. Foundation models in medical image segmentation

Segmentation is one of the most common task in medical imaging. The process of segmentation involves partitioning an image into multiple segments to extract meaningful information. In the context of medical imaging, this could include segmenting organs, tumors, lesions, or other relevant structures. The foundation model can learn to recognize these features and accurately delineate them within the images, providing valuable assistance to healthcare professionals in their diagnostic and treatment-related tasks.

Traditional Algorithms:

- **Graph-based Methods**: For instance, Graph Cut and Watershed Transform. These methods utilize graph theory techniques to segment images by defining nodes and edges within a graph, transforming the image segmentation problem into a graph theory problem.
- **Active Contour Models**: For example, the Level Set method, which is based on curve evolution for image segmentation. This method effectively segments objects within images by evolving curves.
- **Energy Minimization-based Methods**: For example, Markov Random Field, which views image segmentation as an energy minimization problem. By defining an energy function, this method seeks the optimal solution for image segmentation and is commonly used for pixel-wise segmentation tasks.

Deep Learning Algorithms:

- **Convolutional Neural Networks (CNN)**: CNNs, through their convolutional and pooling layers, automatically learn features and identify different structures within images. They excel in medical image segmentation, with popular architectures like U-Net, SegNet, and DeepLab being based on CNNs. These networks effectively capture image features and are used to segment different structures within medical images.
- **Fully Convolutional Networks (FCN)**: FCNs are specifically designed for semantic segmentation, extending convolutional neural networks to pixel-level label predictions. Unlike traditional CNNs, FCNs can accept arbitrary-sized input images and produce segmentation results of the same size, making them well-suited for image segmentation tasks.
- **Encoder-Decoder Architectures**: These models typically consist of an encoder (for feature extraction) and a decoder (for generating segmentation masks). For example, U-Net follows this architecture, where the encoder gradually extracts features, and the decoder maps these features back to the input image size, generating segmentation masks.
- **Vision Transformers** ViTs have emerged as a recent trend in the field of computer vision. Unlike Convolutional Neural Networks (CNNs), ViTs possess the capability to comprehend global image patterns without being constrained by fixed receptive fields. This allows them to capture both local details and broader contexts within images. Furthermore, hierarchical transformers, which combine the strengths of CNNs and transformers, serve to further enhance the segmentation capabilities of ViTs.

5.1. SAM

The Segment Anything Model (SAM) [1]represents a significant advancement in the realm of computer vision and image segmentation. This model is predicated on the capability to delineate and segment any object or region within a given image with a high degree of precision and flexibility. In the context of academic research, SAM embodies an innovative approach that addresses the challenges associated with traditional segmentation methods, which often require extensive manual input or are limited to predefined categories of objects.

At its core, SAM leverages the power of deep learning architectures, particularly convolutional neural networks (CNNs), to achieve its segmentation tasks. These networks are trained on vast datasets to recognize and differentiate between a multitude of features and patterns present in digital images. The training process involves the use of annotated images where the boundaries of various segments are clearly defined. Through this process, SAM learns to identify these boundaries and apply them to new, unseen images.

Referring to the exist survey paper [2],we can see SAM outperforms most current zero shot models in both 2D and 3D (CT,MR,X-ray) datasets. We know SAM is good at zero-shot, but SAM is not good compared to current SOTA when both finetuned. Maybe its because its has pretrained on so many photos from different domains, and the little amount of finetuning data is not strong enough to change its knowledge a lot.

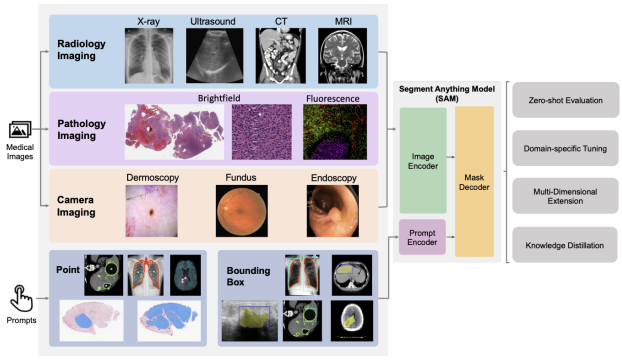
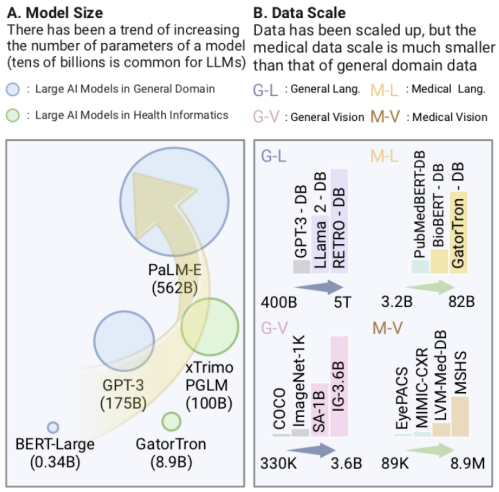


Figure 4. Overview of sam's application



6. Discussion

LLM-based foundation models As we introduced in section 3, the huge boosting of Large Language models (i.e. GPT series) leads to a new era of natural language processing. However, the method can't be directly applied to medical image processing. Computer vision scientists introduced many methods to do so.

Conversational method Conversational models is convenient because doctors can directly interact it with natural language Conversational models face the challenge of understanding context and handling ambiguous queries effectively. Ensuring accurate responses in complex medical scenarios remains an ongoing research challenge. Additionally, they require careful fine-tuning on conversational datasets to perform optimally.

6.1. Future Work

6.1.1 Multimodal Fusion

Medical imaging often involves the integration of data from various modalities such as MRI, CT scans, and ultrasound. One promising avenue for future work involves leveraging foundation models to effectively fuse information from these different modalities, thereby enhancing diagnostic accuracy and clinical decision-making. We think the development of robust multimodal fusion techniques within foundation models stands to revolutionize the field of medical imaging, offering a pathway towards more accurate and comprehensive clinical insights derived from diverse imaging sources.

Multimodal Feature Fusion: Developing techniques to fuse features extracted from different imaging modalities within foundation models, allowing for a more comprehensive and informative representation of the underlying physiological or pathological processes.

Cross-Modal Knowledge Transfer: Exploring methods to transfer knowledge learned from one modality to enhance the analysis of another, enabling foundation models to benefit from the strengths of each modality and potentially compensate for their individual limitations.

6.1.2 Interpretability and Explainability

In the context of medical image analysis, the interpretability and explainability of foundation models are crucial for fostering trust among clinicians and ensuring the responsible integration of AI into clinical practice. Many foundation models, especially deep learning architectures, are often perceived as "black box" systems, making it challenging for clinicians to understand the rationale behind their predictions.

Clinical Correlation and Contextual Information: Integrating clinical knowledge and contextual information into the interpretability framework, allowing foundation models to provide explanations that align with known medical principles and patient-specific factors.

6.1.3 Call for high quality training data

The number of parameter for large language model is surging fast, which requires large scale data to pretrain. However, the medical data scale is much smaller than the general nlp domain, which limits the increasement for parameter scale for foundation models. Thus, call for high quality training data is important.

7. Conclusion

In this paper, we look into various foundation models in medical imaging. We first propose some current challgenge for medical image area and foundation models. Then We survey different categories in foundation model in section 3, and introduce LLM and semgentation models in section 34. Furthermore, we discuss about these models and propose some furture improvment in this field.

References

[1] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 3

[2] Ho Hin Lee, Yu Gu, Theodore Zhao, Yanbo Xu, Jianwei Yang, Naoto Usuyama, Cliff Wong, Mu Wei, Bennett A. Landman, Yuankai Huo, Alberto Santamaria-Pang, and Hoifung Poon. Foundation models for biomedical image segmentation: A survey, 2024. 3

[3] Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge, 2023. 2

[4] Zhengliang Liu, Yue Huang, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Yiwei Li, Peng Shu, Fang Zeng, Lichao Sun, Wei Liu, Dinggang Shen, Quanzheng Li, Tianming Liu, Dajiang Zhu, and Xiang Li. Deid-gpt: Zero-shot medical text de-identification by gpt-4, 2023. 2

[5] Zhengliang Liu, Yiwei Li, Peng Shu, Aoxiao Zhong, Longtao Yang, Chao Ju, Zihao Wu, Chong Ma, Jie Luo, Cheng Chen, Sekeun Kim, Jiang Hu, Haixing Dai, Lin Zhao, Dajiang Zhu, Jun Liu, Wei Liu, Dinggang Shen, Tianming Liu, Quanzheng Li, and Xiang Li. Radiology-llama2: Best-in-class large language model for radiology, 2023. 2

[6] Omkar Thawkar, Abdelrahman Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Shahbaz Khan. Xraygpt: Chest radiographs summarization using medical vision-language models, 2023. 2

[7] Guangyu Wang, Guoxing Yang, Zongxin Du, Longjun Fan, and Xiaohu Li. Clinicalgpt: Large language models finetuned with diverse medical data and comprehensive evaluation, 2023. 2

[8] Sheng Wang, Zihao Zhao, Xi Ouyang, Qian Wang, and Dinggang Shen. Chatcad: Interactive computer-aided diagnosis on medical image using large language models, 2023. 2

[9] Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-llama: Towards building open-source language models for medicine, 2023. 2

[10] Kai Zhang, Jun Yu, Eashan Adhikarla, Rong Zhou, Zhiling Yan, Yixin Liu, Zhengliang Liu, Lifang He, Brian Davison, Xiang Li, Hui Ren, Sunyang Fu, James Zou, Wei Liu, Jing Huang, Chen Chen, Yuyin Zhou, Tianming Liu, Xun Chen, Yong Chen, Quanzheng Li, Hongfang Liu, and Lichao Sun. Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks, 2024. 2

[11] Shaoting Zhang and Dimitris Metaxas. On the challenges and perspectives of foundation models for medical image analysis, 2023. 1