# Topic: Comparing Performance and Transparency of Multimodal Foundation Models in Healthcare

Artificial intelligence (AI) has demonstrated its efficiency and universality in various domains. However, the black-box nature of Deep Learning (DL) algorithms and foundation models (FM) restricts their clinical applications despite having underlying statistical principles [1]. Enhancing its transparency and explainability is a way to improve the confidence of both physicians and patients. It contributes to ferreting out the dominant features in making the diagnosis. This property helps physicians retrace clinical cases and judgments, as well as benefits patients in understanding what they are experiencing. Following the need to explore how AI makes decisions and what it learns from the networks, explainable Artificial Intelligence (XAI) is proposed to reveal the decision process, making explanations and interpretations visible [2]. XAI calls for confidence, privacy, ethics, and fairness [3], demonstrating its compatibility with applicable medicine.

The majority of the papers for XAI in medical image diagnosis apply attribution-based explainability methods that aim to determine the contribution of an input feature to the target neuron [1]. It can be further separated into three categories based on various mechanisms [4]: backpropagation-based like IxG [5], activation-based exemplified by techniques such as Grad-CAM [6] and Grad-CAM++ [7], and perturbation-based methods with occlusion [8] and RISE [9]. The backpropagation-based methods are based on the gradients [10, 11, 12, 13, 14]. Activation-based methods determine the importance by weighting activation maps [15, 16, 17, 18]. Perturbation-based methods assign importance by observing the change in input with perturbing [19, 20]. This mechanical difference leads to different visual quality, with IxG, Grad-CAM, ablation-CAM, and occlusion showing better readability [4].

Foundation models (FM) are a type of AI that are trained on large-scale datasets. With pre-trained weights, they can be data and computation efficient when adapted to downstream tasks with a limited number of task-specific datasets. In the context of medical data, those datasets refer to medical images and corresponding reports, and vary between modalities [21] like MRI and CT. Multimodal FMs like CLIP [22] therefore show superiority in processing medical data compared to vision FMs and language FMs. However, CLIP is pre-trained using general-domain data that are available on the Internet [22], showing a large domain gap between universal information and medical data that can be hardly compensated by fine-tuning for downstream tasks. On the contrary, domain-specific FMs serve as a solid foundation for biomedical tasks [23], and medical-CLIPs [24, 25, 26] are proposed to bridge the gap with medical-domain pre-training, which leads to better performances compared to general CLIP. Additionally, CLIPs are mostly used for classification, image text retrieval, and visual question answering (VQA) tasks. Since applying CLIPs to visual and language tasks is difficult because of the requirement of complex multimodal reasoning [27], fine-tuning that merges a cross-modal interaction module should be taken into account.

This sets the scope for our work. We build upon the generally pre-trained CLIP [22] and domain-specific pre-trained CLIPs (PMC-CLIP [24], CT-CLIP [25], BiomedCLIP [26]) and fine-tune them for different downstream tasks with task-specific datasets. The application of these CLIPs enables the comparison between multimodal foundation models using different pre-training schemes, leading to further discussion of foundation models in the healthcare domain. To ulteriorly compare the difference between these two types of CLIPs, we intend to adopt XAI for transparency, explainability, and interpretability. As attribution-based explainability methods are most commonly used in medical imaging, we consider introducing IxG [5], IntGrad [13], and occlusion [8] for difference visualization since they showcase better interpretability [4].

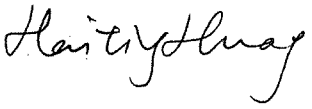
The proposed work comprises the following key components:

- Prepare all the task-specific medical datasets for downstream tasks, e.g. VQA-RAD [28] and SLAKE [29] for visual question answering (VQA) task; MedMNIST [30] for image classification task; ROCO [31] and PMC-15M [26] for image-text retrieval task.
- Start with classification tasks: fine-tune the CLIPs with the MedMNIST dataset, and evaluate for classification tasks. Evaluation metrics: AUROC and accuracy.
- Regarding VQA tasks, we employ two MedVQA methods: QCR [32] and MEVF [33], fine-tune the CLIPs with SLAKE and VQA-RAD datasets, and incorporate the fine-tuned CLIPs into MedVQA methods. Evaluation metrics: accuracy.

- For Image-text retrieval tasks, fine-tune the CLIPs with ROCO and PMC-15M datasets, and image-text retrieval tasks can reference from discriminability-captioning [27, 34]. Evaluation metrics: recall@k (k=1, 5, 10).
- Bias studies: for each task, train monomodal models (chosen based on the visual encoder and text encoder in CLIPs) with corresponding task datasets, to identify the potential bias in the tasks and datasets.
- After adapting the CLIPs to each task, applying explainability methods e.g. IxG, IntGrad, occlusion.
- Qualitative evaluation can be achieved by showing saliency maps.
- (Optional) Quantitative evaluation of the explainability methods with Quantus [35].

This comprehensive analysis will demonstrate the effectiveness of domain-specific foundation models by applying explanations, facilitating a deeper understanding of model behavior in medical imaging.

**Supervisors:** Dr. Dario Zanca, Dr. Emmanuelle Salin, and Prof. Dr. Björn Eskofier
**Student:** Haiting Huang
**Start – End:** November 15th 2024 - May 14th 2025



# References


[1] Amitojdeep Singh, Sourya Sengupta, and Vasudevan Lakshminarayanan. Explainable deep learning models in medical image analysis. *Journal of imaging*, 6(6):52, 2020.

[2] Jaegul Choo and Shixia Liu. Visual analytics for explainable deep learning. *IEEE computer graphics and applications*, 38(4):84–92, 2018.

[3] Peter Kieseberg, Edgar Weippl, and Andreas Holzinger. Trust for the doctor-in-the-loop. *ERCIM news*, 104(1):32–33, 2016.

[4] Sukrut Rao, Moritz Böhle, and Bernt Schiele. Towards better understanding attribution methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10223–10232, 2022.

[5] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMlR, 2017.

[6] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[7] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018.

[8] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014.

[9] V Petsiuk. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

[10] Karen Simonyan. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[11] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[12] Suraj Srinivas and François Fleuret. Full-gradient representation for neural network visualization. *Advances in neural information processing systems*, 32, 2019.

[13] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[14] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018.

[15] Harish Guruprasad Ramaswamy et al. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 983–991, 2020.

[16] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888, 2021.

[17] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25, 2020.

[18] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[19] Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. *Advances in neural information processing systems*, 30, 2017.

[20] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision*, pages 3429–3437, 2017.

[21] Bobby Azad, Reza Azad, Sania Eskandari, Afshin Bozorgpour, Amirhossein Kazerouni, Islem Rekik, and Dorit Merhof. Foundational models in medical imaging: A comprehensive survey and future vision. *arXiv preprint arXiv:2310.18689*, 2023.

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[23] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

[24] Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 525–536. Springer, 2023.

[25] Ibrahim Ethem Hamamci, Sezgin Er, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Muhammed Furkan Dasdelen, Bastian Wittmann, Enis Simsar, Mehmet Simsar, et al. A foundation model utilizing chest ct volumes and radiology reports for supervised-level zero-shot detection of abnormalities. *arXiv preprint arXiv:2403.17834*, 2024.

[26] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, et al. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*, 2(3):6, 2023.

[27] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*, 2021.

[28] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.

[29] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE, 2021.

[30] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

[31] Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christoph M Friedrich. Radiology objects in context (roco): a multimodal image dataset. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*, pages 180–189. Springer, 2018.

[32] Li-Ming Zhan, Bo Liu, Lu Fan, Jiaxin Chen, and Xiao-Ming Wu. Medical visual question answering via conditional reasoning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2345–2354, 2020.

[33] Binh D Nguyen, Thanh-Toan Do, Binh X Nguyen, Tuong Do, Erman Tjiputra, and Quang D Tran. Overcoming data limitation in medical visual question answering. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*, pages 522–530. Springer, 2019.

[34] Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. Discriminability objective for training descriptive captions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6964–6974, 2018.

[35] Anna Hedström, Leander Weber, Daniel Krakowczyk, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina Marina M.-C. Höhne. Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, 24(34):1–11, 2023.