



SHAP Interpretability for LLaVA-Rad

Thesis BA/FA/MA
Supervisor Jakob Hofmann
Examiner Prof. Dr.-Ing. Bin Yang

Date
10. April 2026

Motivation

LLaVA-Rad is a state-of-the-art vision-language model specifically fine-tuned for radiological image understanding and report generation. While these multimodal large language models show impressive performance on medical imaging tasks, their decision-making processes remain largely opaque. This lack of interpretability poses significant challenges in clinical settings where understanding *why* a model makes a particular diagnosis is crucial for physician trust and patient safety.

SHAP (SHapley Additive exPlanations) provides a theoretically grounded framework for explaining model predictions by attributing importance scores to input features. However, applying SHAP to multimodal vision-language models presents unique challenges due to the complex interplay between visual and textual representations. This thesis will develop and evaluate SHAP-based interpretability methods specifically tailored for LLaVA-Rad, enabling clinicians to understand which image regions and linguistic patterns drive diagnostic predictions.

Objectives

- Implement SHAP-based explanation methods for LLaVA-Rad's vision-language architecture, handling both image patches and text tokens.
- Evaluate interpretability quality through quantitative metrics (faithfulness, localization) and qualitative assessment.

Prerequisites

- Good programming skills in Python and PyTorch
- *Optional*: Prior experience working with Large Language Models
- *Optional*: Participated in the ISS Deep Learning Lab or took the Deep Learning exam with good results

If this topic has sparked your interest, write me an email and we can discuss the proposal in more detail. Please include your current transcript and CV.

References

- [1] B. Boecking, N. Usuyama, S. Bannur, D. Coelho de Castro, A. Schwaighofer, S. Hyland, H. Sharma, M. T. Wetscherek, T. Naumann, A. Nori, J. Alvarez Valle, H. Poon and O. Oktay, "MS-CXR: Making the Most of Text Semantics to Improve Biomedical Vision-Language Processing," *PhysioNet*, Nov. 2024, version 1.1.0. [Online]. Available: <https://doi.org/10.13026/9g2z-jg61>
- [2] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark and S. Horng, "MIMIC-CXR: A large publicly available database of labeled chest radiographs," *Scientific Data*, vol. 6, no. 1, p. 317, 2019. [Online]. Available: <https://doi.org/10.1038/s41597-019-0322-0>
- [3] J. M. Zambrano Chaves, S.-C. Huang, Y. Xu, H. Xu, N. Usuyama, S. Zhang, F. Wang, Y. Xie, M. Khademi, Z. Yang, H. Awadalla, J. Gong, H. Hu, J. Yang, C. Li, J. Gao, Y. Gu, C. Wong, M.-H. Wei, T. Naumann, M. Chen, M. Lungren, A. Chaudhari, S. Yeung, C. Langlotz, S. Wang and H. Poon, "Towards a Clinically Accessible Radiology Foundation Model: Open-Access and Lightweight, with Automated Evaluation," arXiv preprint arXiv:2403.08002, 2024. [Online]. Available: <https://arxiv.org/abs/2403.08002>
- [4] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 4765–4774. [Online]. Available: <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>

Jakob Hofmann
jakob.hofmann@iss.uni-stuttgart.de