Xplainer: From X-Ray Observations to Explainable Zero-Shot Diagnosis

Yu-Chuan Cheng
yu-chuan.cheng@stud.uni-heidelberg.de

Feb 28, 2025

Introduction

In recent years, artificial intelligence (AI) has transformed medical imaging, pushing the boundaries of diagnostic accuracy and efficiency. Among the most promising developments is Xplainer (Pellegrini et al., 2023), a model designed for an explainable zero-shot diagnosis based on X-ray observations. This innovative approach uses contrastive learning, prompt engineering, and interpretability techniques to enhance model transparency while improving generalization across datasets.

Xplainer builds upon models like CLIP (Radford et al., 2021) and other vision-language frameworks. By applying descriptive features and prompt engineering, it improves diagnostic precision, particularly for lung diseases such as pneumonia. However, despite its advances, questions remain about its clinical viability and real-world implementation.

From CLIP to Xplainer: A Historical Perspective

1. **Deep Visual-Semantic Alignments (Karpathy & Fei-Fei, 2014):** Pioneered CNN-RNN embeddings for image descriptions.
2. **Fine-Grained Visual Representations (Reed, Akata, Lee, & Schiele, 2016):** Improved image-text alignment using symmetric embeddings.
3. **Latent Language Learning (Andreas, Klein, & Levine, 2017):** Demonstrated how language fosters compositional generalization.
4. **CLIP (2021):** Introduced contrastive learning for vision-language tasks, maximizing cosine similarity between matching image-text pairs.
5. **Biomedical Vision-Language Processing (Boecking et al., 2022):** Applied contrastive loss in medical imaging, improving text-based recognition.
6. **Xplainer**

The foundation of Xplainer can be traced back to key breakthroughs in the concepts. The first three papers are the basis of CLIP and the fourth paper is the baseline of Xplainer. Building on these advancements, Xplainer refines vision-language alignment by integrating medical descriptors and domain-specific knowledge. Unlike previous approaches, it moves beyond predefined pathologies, allowing for greater adaptability in zero-shot learning scenarios.

Methodology

Xplainer adapts the classification-by-description approach of contrastive image-text models to the multi-label medical diagnosis task. Instead of directly predicting a diagnosis, it

prompts the model to classify the presence of descriptive observations a radiologist would look for in an X-ray scan and uses the descriptor probabilities to estimate the likelihood of a diagnosis.

Xplainer leverages BioVil, a pre-trained CLIP model for X-rays, and applies contrastive observation-based prompting. This approach (Figure 1) learns representations of X-ray images and their corresponding descriptive observations by maximizing the similarity between the image and text, thus linking image features with textual semantics. When making a diagnosis, the system first calculates the probabilities of descriptive observations based on the input image and then estimates the likelihood of a diagnosis using these probabilities. For instance, to diagnose pneumonia, Xplainer evaluates the probabilities of descriptive observations such as "lung opacity" or "signs of inflammation" and uses these to estimate the likelihood of pneumonia.

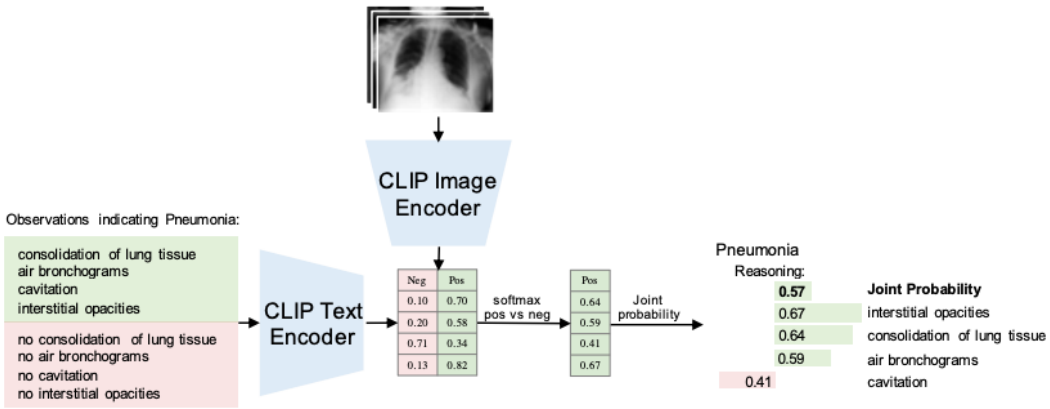

Figure 1: System pipeline (Pellegrini et al., 2023)

Key Findings and Takeaways

This comparison table 1 highlights three methods for the diagnosis of medical images. Xplainer excels in **generalization, localization, and interpretability** but assumes simultaneous and equally important descriptors. Report-guided methods (Seibold, Reiß, Sarfraz, Stiefelhagen, & Kleesiek, 2022) adapt to new pathologies and use unstructured reports with a global-local dual encoder but are costly to train. BioViL is the basis of Xplainer. Therefore, we put it here and show its disadvantages for comparison.

References

Andreas, J., Klein, D., & Levine, S. (2017). Learning with latent language. In *North american chapter of the association for computational linguistics.* Retrieved from `https://api.semanticscholar.org/CorpusID:37390552`

Boecking, B., Usuyama, N., Bannur, S., Castro, D. C., Schwaighofer, A., Hyland, S., ... Oktay, O. (2022). Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing. In (Vol. 13696, pp. 1–21). doi: 10.1007/978-3-031-20059-5_1

Method	Advantages	Disadvantages
Xplainer	- Stronger generalization ability - Fine localization capabilities - Built-in interpretability	- No considering interdependencies - All descriptors the same importance
Report-Guided	- Can adopt new pathology - Only need unstructured report - Has more detail	- High training cost
BioViL (Baseline)	- Strong versatility - Support unlabeled data training	- Fail outside of their predefined set - Lack of fine-grained alignment - Weak interpretability

Table 1: Sum up three Method.

Karpathy, A., & Fei-Fei, L. (2014). Deep visual-semantic alignments for generating image descriptions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3128-3137. Retrieved from `https://api.semanticscholar.org/CorpusID:8517067`

Pellegrini, C., Keicher, M., Özsoy, E., Jiraskova, P., Braren, R., & Navab, N. (2023, June). *Xplainer: From X-Ray Observations to Explainable Zero-Shot Diagnosis* (No. arXiv:2303.13391). arXiv. doi: 10.48550/arXiv.2303.13391

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning.* Retrieved from `https://api.semanticscholar.org/CorpusID:231591445`

Reed, S. E., Akata, Z., Lee, H., & Schiele, B. (2016). Learning deep representations of fine-grained visual descriptions. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 49-58. Retrieved from `https://api.semanticscholar.org/CorpusID:7102424`

Seibold, C., Reiß, S., Sarfraz, M. S., Stiefelhagen, R., & Kleesiek, J. (2022). Breaking with fixed set pathology recognition through report-guided contrastive training. *ArXiv*, *abs/2205.07139*. Retrieved from `https://api.semanticscholar.org/CorpusID:248811422`