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

Chantal Pellegrini*[1], Matthias Keicher*[1], Ege Özsoy*[1],
MD Petra Jiraskova[2], MD Rickmer Braren[2], and Nassir Navab[1]

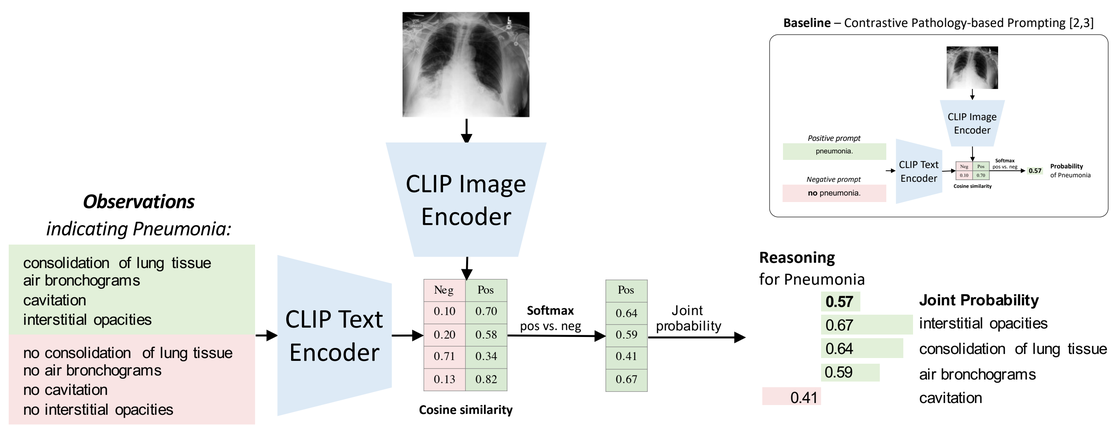
[1]Chair for Computer-Aided Medical Procedures and Augmented Reality, Technical University of Munich, Germany
[2]Department of Diagnostic and Interventional Radiology, School of Medicine, Technical University of Munich, Germany
*Contributed equally.

Motivation

Automated diagnosis prediction from medical images provides valuable support for clinical decision-making. However, existing methods not only rely on large amounts of annotated data, they are often a black-box. In this work, we introduce Xplainer, a **zero-shot classification-by-description approach**, drawing inspiration from how a radiologist interprets an X-ray. Rather than making a direct diagnosis, it identifies and classifies descriptive observations in the image, building a transparent path to the final prediction. This design not only makes our model **inherently explainable**, but also allows for **adaptation to new diseases** with known symptoms without the need for additional training or annotated data.

Method

Xplainer - Zero-shot Classification-by-description



Xplainer builds upon BioVil [1], which is a contrastive language-image pretraining (CLIP) model, trained on radiological images and reports. Instead of predicting clinical findings directly, we first **predict visual observations associated with each finding** and then form a **joint probability**. For each finding, the list of observations is initially created by ChatGPT and then refined by experienced radiologists.

Zero-shot Inference:

1. Compute the image embedding for the X-ray image.
2. Compute the text embeddings for observations (and their absence) for each pathology:
   There is/are (no) <**observation**> *indicating* <**pathology**>
3. Compute the cosine similarity between each image and text embedding.
4. Estimate the softmax probability for the presence of each observation in the X-ray.
5. Finally, determine the likelihood of each pathology by computing the joint probability:

$$\log(P(p)) = \frac{1}{N}\sum_{i=1}^{N}\log(P(o_i))$$

Data

MIMIC III: Large dataset of over 200,000 Chest X-ray images paired with free-text reports. Used for self-supervised, contrastive language-image pretraining.
CheXpert: Multi-label classification with 14 classes (12 pathologies, "No Finding" and "Support Devices") of Chest X-rays. Encompasses 200 validation and 500 test samples.
ChestX-ray14: Multi-label classification of 14 pathologies; test set of 25,596 Chest X-rays.

References

[1] Benedikt Boecking et al. "Making the most of text semantics to improve biomedical vision–language processing". In: *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*. Springer. 2022, pp. 1–21.

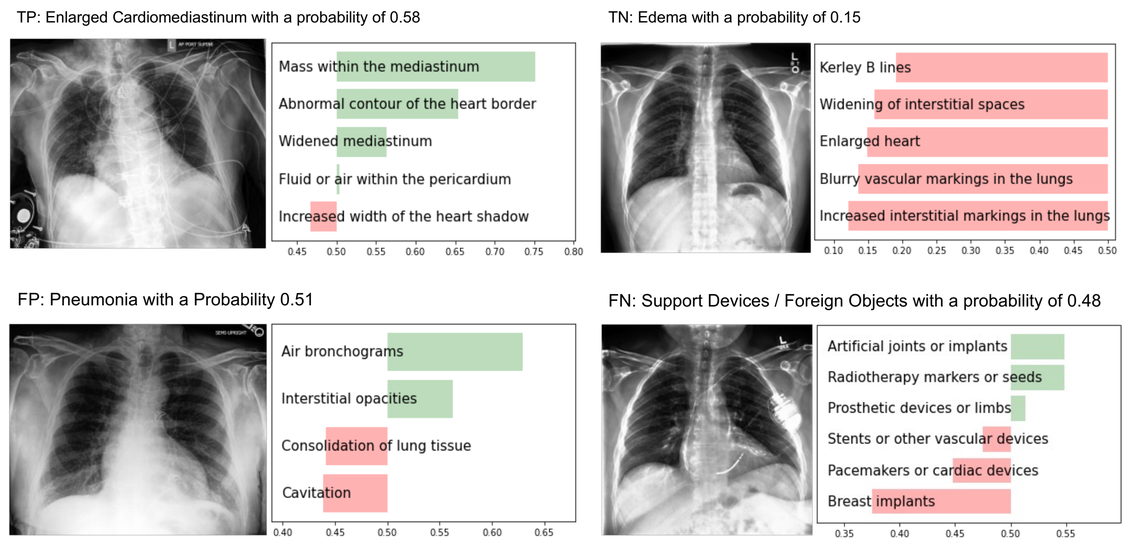
[2] Constantin Seibold et al. "Breaking With Fixed Set Pathology Recognition Through Report-Guided Contrastive Training". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V*. Singapore, Singapore: Springer-Verlag, 2022, 690–700. isbn: 978-3-031-16442-2.

Acknowledgement

The authors gratefully acknowledge the financial support by the Federal Ministry of Education and Research of Germany (BMBF) under project DIVA (FKZ 13GW0469C) and the Bavarian Research Foundation (BFS) under project PandeMIC (grant AZ-1429-20C).

Results

Explanations



Zero-Shot Diagnosis

- We evaluated on the CheXpert and ChestX-ray14 datasets for multi-label classification.
- Xplainer achieved state-of-the-art (SOTA) out-of-domain results on both datasets.
- Our performance improvement goes hand in hand with explainability.

	CLIP pretraining data	CheXpert		ChestX-ray14
		val	test	test
CheXzero [2]	MIMIC	–	74.73	–
Seibold et al. [3]	MIMIC	78.86	–	71.23
Seibold et al. [3]	MIMIC, PadChest, ChestX-ray14	83.24	–	78.33 (in domain)
Xplainer	MIMIC	**84.92**	**80.58**	**71.73**

Ablation on Prompting Styles

- Observation-based prompting outperforms pathology-based prompting by 9%.
- Contrastive prompting outperforms basic prompting with thresholding.
- Specifying the pathology reduces ambiguity and further improves performance by 7%.
- Using a report-style formulation for prompts results in a slight improvement.

	AUC
Contrastive pathology-based Prompting *((no) <pathology>)*	76.14
Observation-based Prompting:	
Basic Prompt *(<observation>)*	58.65
Contrastive Prompt *((no) <observation>)*	77.00
+ pathology Indication *(indicating <pathology>)*	84.35
+ Report Style *(There is/are)*	84.92

Radiologist Refinement

- Experienced radiologists improved or removed incorrect and irrelevant descriptors.
- Manually refining with domain knowledge results in a slight performance improvement.
- The already promising results achieved by only relying on ChatGPT demonstrate the potential of integrating large generic language models into medical image analysis.

	CheXpert Val	CheXpert Test	ChestX-ray14
ChatGPT Prompts	83.61	79.94	71.40
Refined Prompts	**84.92**	**80.58**	**71.73**

Conclusion

We introduce **Xplainer**, a novel and effective zero-shot approach for chest X-ray diagnosis that **achieves SOTA results** in detecting common lung findings. The compositionality of our classification-by-description method offers **intuitive explanations** and **fine-grained class customization**. Our work highlights the potential of contrastive pretraining combined with observation-based prompting for medical zero-shot classification, where labeled data is limited and explainability is crucial.

Code Demo