Measuring interpretability in chest radiograph classification

ProtoCXR: every prediction grounded in a real training case

Tran Quoc Hung · Le Tuan Hung · Nguyen Ngoc Dung · Luong Thi Tra My · Tran Hoang Linh
Academic supervisors: Hoang Tuan Anh and Joshua Dwight
School of Science, Engineering & Technology | RMIT University Vietnam

The problem

Deep learning models now match radiologists on many chest radiograph tasks, but a clinician cannot safely act on a diagnosis they cannot interrogate.

The common remedy is a saliency map drawn after training, colouring in where the model supposedly looked. Such maps barely change even when the model's weights are destroyed, so they may not be describing the model at all.

Our approach

ProtoCXR makes the explanation part of the computation rather than a separate estimate.

It classifies a radiograph by comparing its regions against learned visual prototypes, each projected onto a real training patch. Every prediction arrives with a “this looks like that” comparison to a confirmed case.

How it works

Backbone
Feature map

→

Prototype layer
Occurrence weighted

→

Linear classifier
Sum of prototype scores

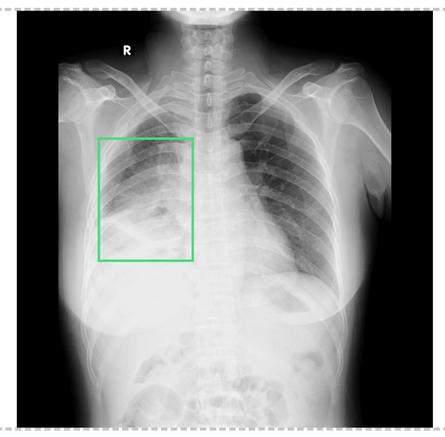
Prototype push
Every prototype is projected onto its nearest real training patch, so each explanation points at a confirmed case rather than an abstract vector.

Deterministic anchor
The matched patch is an arg min over distances, computed in fp32 so the same input always yields the same explanation.

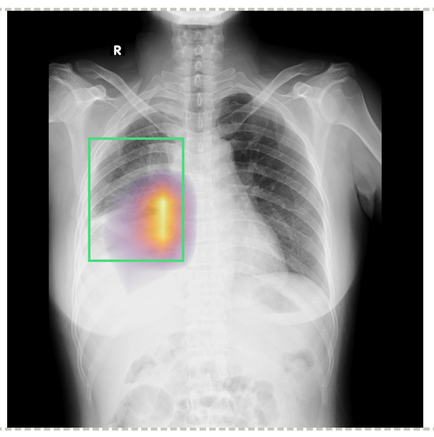
Exact attribution
The class score is an additive sum of per prototype contributions, so each prototype's share is read off the model, never estimated afterwards.

Trained in four stages
Prototype initialisation, a warm phase with the backbone frozen, a joint phase with periodic push, then classifier fine tuning.

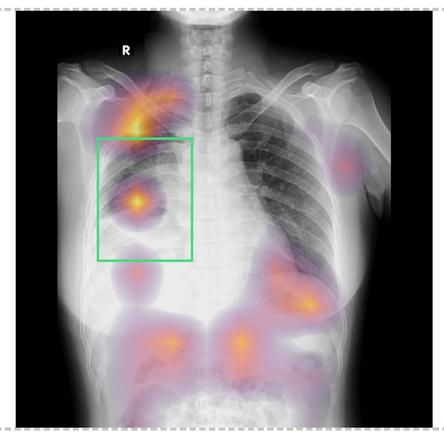
Where the model actually looked



Radiograph + radiologist box



ProtoCXR explanation



Saliency baseline

Same image, same trained model, same class. Only the explanation method differs.

What we measured

Prototype grounding
Distance from each prototype to its nearest real training patch

Lesion localisation
Agreement with boxes drawn by radiologists

Region coverage
Share of activation falling inside the lesion

Model dependence
Does the map change when the weights are destroyed

Where it falls short

Boundary agreement
Localisation overlap is low in absolute terms for every method tested, including the saliency baselines. Models find roughly the right region without matching boundaries.

Accuracy
Classification sits below black box models trained on a single dataset. The contribution here is measured interpretability, not peak predictive performance.

Evaluation
Metrics are functionally grounded. A clinical reader study relating explanation quality to radiologist judgement is left for future work.

The explanation is the computation

Interpretability should be measured, not illustrated. ProtoCXR is a reproducible protocol and an inherently interpretable baseline for chest radiograph research.

Full project on the RMIT Showcase