IP-CRR: Information Pursuit for Interpretable Classification of Chest Radiology Reports

Yuyan Ge[1], Kwan Ho Ryan Chan[1], Pablo Messina[2], René Vidal[1]

[1] University of Pennsylvania, USA
[2] Pontifical Catholic University of Chile, CENIA, iHEALTH, Chile

MICCAI 2025
28th International Conference on Medical Image Computing and Computer Assisted Intervention
23-27 September 2025
Daejeon Convention Center
Daejeon REPUBLIC OF KOREA

Introduction

- **Motivation:** AI in radiology has the potential to improve **diagnostic accuracy**, enhance **efficiency** and **reduce workload**
- **Challenge:** **Lack of interpretability** limits clinical adoption
- **Solution:** Information Pursuit for classifying Chest Radiology reports (IP-CRR), an **interpretable-by-design framework** that mimics the doctors' decision–making process
- **Key idea**: Ask a sequence of interpretable queries about the report, make a prediction based on the query-answer pairs, which will naturally serve as an explanation for the prediction

Radiology report x: There is a dialysis catheter overlying the right chest with the tip in the cavoatrial junction. Heart size is normal. The mediastinal and hilar contours are stable. The pulmonary vasculature is normal. Lungs are clear. No pleural effusion or pneumothorax is seen. There are no acute osseous abnormalities.

Black-box model

Classify the following radiology report with whether a specific disease is present or not.
Report: {report x}
Disease: Cardiomegaly

Black-box model

Prediction: Present ✗

IP-CRR

Ask a sequence of interpretable queries about the report x:

q_1: Moderately enlarged heart shadow?	Unknown
q_2: Slightly enlarged heart contour?	No
q_3: Normal shape heart?	Yes
q_4: Stable cardio mediastinal contour?	Yes

Prediction: Cardiomegaly is **not** present with 99% confidence ✓

Methods

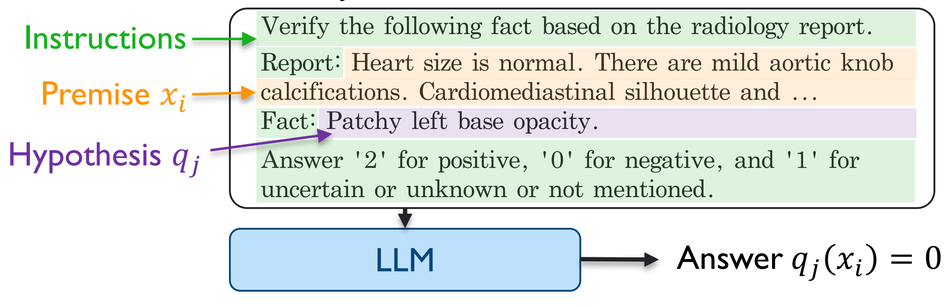
Query set generation

- **Extract representative facts** from a large set of reports:
  1) Extract a list of facts from each report using language models ChatGPT 3.5/4 and T5-small[1]
  2) Encode each fact into embedding space using CXRFE[1]
  3) Cluster embeddings to get representative facts

Query answering

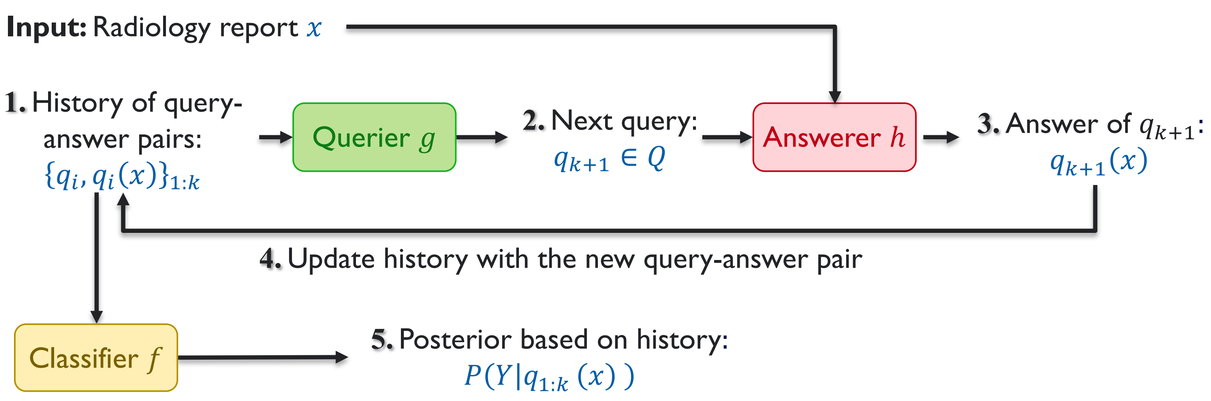
- Answer a query by using Natural Language Inference (NLI) to **check if the fact is entailed by the report**

For every $x_i \in X$ and $q_j \in Q$:



Classification based on Information Pursuit

- **Querier**: selects the next **most informative** query
- **Classifier**: predicts disease from query-answer pairs



- Given a sample CRR $x^{obs} \in \mathcal{X}$, IP selects the queries:

$$q_1 = \underset{q \in Q}{\operatorname{argmax}}\, I(q(X); Y);$$

$$q_{k+1} = \underset{q \in Q}{\operatorname{argmax}}\, I(q(X); Y \mid q_{1:k}(x^{\text{obs}})).$$

- X, Y: Random variables pertaining to data and labels respectively
- Q: Query set
- $q(X)$: Answer to query evaluated at X

- To implement IP, we follow a variational approach V-IP[2]:

$$\min_{f,g} \quad \mathbb{E}_{X,S}\left[D_{KL}(P(Y \mid X) \parallel \hat{P}(Y \mid q(X), S))\right]$$

$$\text{where} \quad q := g(S) \in Q$$

$$\hat{P}(Y \mid q(X), S) := f(\{q, q(X)\} \cup S).$$

- $S = q_{1:k}(x)$: History of query-answer pairs

Results

- Experiments on six medical conditions in MIMIC-CXR dataset: lung opacity (LO), cardiomegaly (CA), etc.
- IP-CRR converges to the max performance with only about **30/520 of queries**

Table 1: Predictive performance of five methods for six medical conditions.

Methods	Average Precision ↑						F1 Score ↑					
	LO	SD	PE	CM	CA	PN	LO	SD	PE	CM	CA	PN
Flan-T5-large	0.527	0.445	0.616	0.380	0.073	0.190	0.663	0.321	0.754	0.543	0.139	0.299
CXR-BERT (FT-Last)	0.900	0.969	0.945	0.864	0.361	0.449	0.829	0.912	0.887	0.789	0.223	0.449
CXR-BERT (FT-All)	0.984	0.970	0.962	0.964	0.992	0.641	0.987	0.978	0.953	0.982	0.991	0.541
CBM	0.947	0.934	0.874	0.791	0.345	0.432	0.884	0.853	0.801	0.738	0.241	0.431
IP-CRR (Our)	**0.972**	**0.959**	**0.925**	**0.892**	**0.578**	**0.468**	**0.918**	**0.889**	**0.860**	**0.811**	**0.350**	**0.451**

Black-box methods Interpretable methods

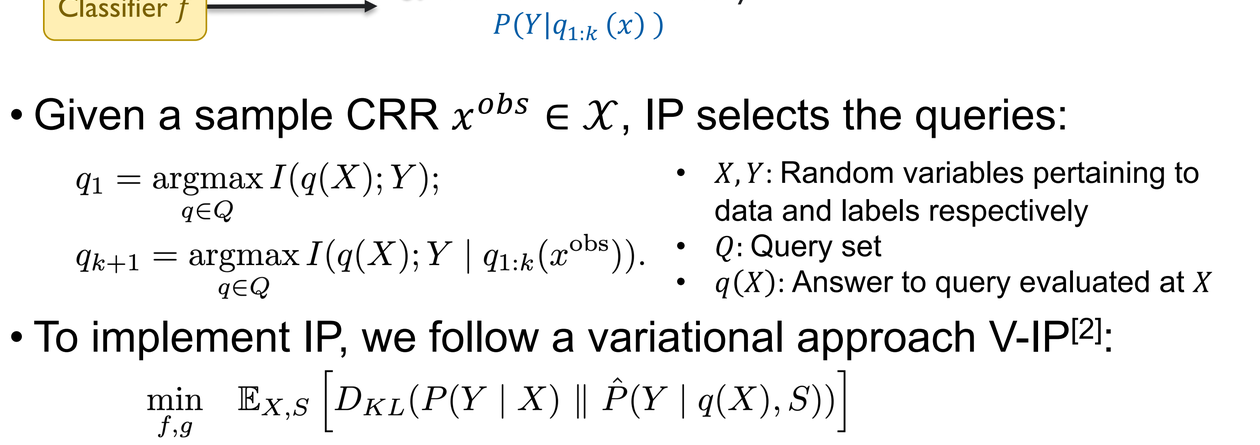


Figure 1: Examples of query-answer chains from IP-CRR. Each row in the colored matrix represents the posterior $P(Y|q_{1:k}(x))$ after selecting the query q_k. A query is shown in red if the answer is "no", in green if "yes", and in black if "unknown". "N" or "P" means negative or positive prediction of the disease, respectively.

References:

[1] Messina, P., Vidal, R., et al.: Extracting and encoding: Leveraging large language models and medical knowledge to enhance radiological text representation. In: ACL Findings (2024)
[2] Chattopadhyay, A., Chan, K.H.R., et al.: Variational information pursuit for interpretable predictions. In: ICLR (2023)

MICCAI Paper Code