Very Short and Accurate Explanations by Design

Prisha Shroff, Hancheng Min, René Vidal

University of Pennsylvania

ICCV OCT 19-23, 2025 HONOLULU HAWAII

Introduction

- The lack of transparency of "black-box" models raises the question of whether their predictions can be trusted
- Existing explainable AI methods:

Post-Hoc

- Aim to explain the prediction after it has been made.
- Do not reliably or faithfully represent the model's decision-making process

Interpretable-by-design methods

- Map the input image to a set of interpretable concepts from which a prediction is made
- Concept Bottleneck Models use a linear predictor, reducing accuracy and interpretability

- Information Pursuit (IP) [1] is an explainable-by-design method based on selecting a small set of most informative concepts that are sufficient for prediction.
- Implementing IP requires three ingredients:

Querier

- Selects which concepts to query and in what order

Answerer

- Predicts whether a concept selected by the querier is present in the image

Classifier

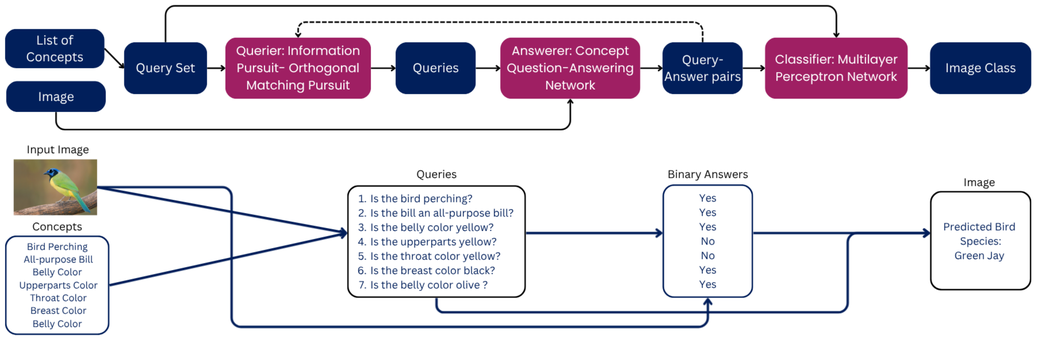
- Predicts the class from the sequence of query-answer pairs

- Existing IP-based models use either an inefficient querier, or a non-zero shot and inaccurate answerer

Current Approaches	Efficient Query Selection?	Accurate Zero-Shot Answers?
Concept Bottleneck Models (CBMs)	No selection	Not zero-shot
Information Pursuit (IP) [1]	No	Not zero-shot
Variational Information Pursuit (VIP) [2]	Somewhat	Not zero-shot
Variational Information Pursuit - Concept Question Answering Network (VIP-Concept QA) [3]	Somewhat	Yes
Information Pursuit - Orthogonal Matching Pursuit (IP-OMP) [4]	Yes	No
Information Pursuit- Orthogonal Matching Pursuit-Concept Question Answering Network (IP-OMP-Concept QA)	Yes	Yes

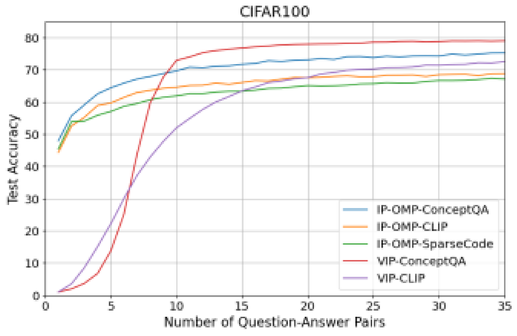
- IP-OMP uses OMP in CLIP embedding space to represent an image as a sparse code of a few queries, and bypasses the need to train a separate querier

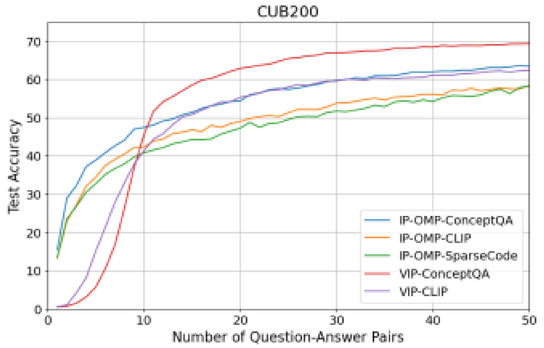
Our Method: IP-OMP-Concept QA



Experimental Results

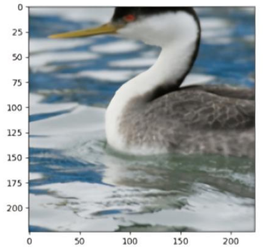
IP-OMP-ConceptQA is better for short explanation lengths





- When #queries < 9, IP-OMP-ConceptQA outperforms all other methods
- When #queries > 9, IP-OMP-Concept QA is competitive with VIP-Concept QA (drop of accuracy is less than 4%)
- When #queries < 11, IP-OMP-ConceptQA performs the **best** against all algorithms
- IP-OMP-Concept QA is **competitive** with VIP-Concept QA (drop of accuracy is less than **6%**)

Qualitative Evaluation of Explanations



	IP-OMP Queries:	Sparse Code Answers:	Concept QA Answers:
1.	a black or dark grey plumage	0.15	1.0 (yes)
2.	a large blue-grey bird	-0.23	-1.0 (no)
3.	a large, black spider	-0.25	-1.0 (no)
4.	a seabird	0.18	1.0 (yes)
5.	a small songbird	-0.28	-1.0 (no)
6.	all black coloration	-0.23	-1.0 (no)
7.	large webbed feet	0.28	1.0 (yes)
8.	three toes on each foot	0.30	-1.0 (no)
9.	white and black coloration	0.19	1.0 (yes)
10.	white body	-0.36	-1.0 (no)
11.	white undertail coverts	0.30	1.0 (yes)
12.	white underwings	-0.16	-1.0 (no)
13.	yellowish underparts	-0.18	-1.0 (no)
14.	a black bill with a yellow tip	0.19	-1.0 (no)
15.	a blue-gray body	0.31	-1.0 (no)
16.	a dark, glossy bill	0.16	-1.0 (no)
17.	a medium-sized duck	0.16	-1.0 (no)
18.	a small, blue bird	0.16	-1.0 (no)
19.	white stripes on the wings	0.12	-1.0 (no)
20.	yellowish legs	0.12	-1.0 (no)

- A positive coefficient (yes, the query is present) is shown in green and a negative coefficient (no, the query is not present) is shown in red.
- Sparse coefficients (answers from IP-OMP) are between (-0.3,0.3), making it uninterpretable
- Concept QA provides binary answers in {-1,1), which are interpretable

Paper Contributions

- Developed IP-OMP-ConceptQA
  - Interpretable-by-design method
  - Efficient query selection method (querier)
  - Accurate zero-shot answering system (answerer)
  - Produces an interpretable representation (sequence of query-answer pairs) for classification
- IP-OMP-ConceptQA outperforms existing methods in scenarios where very short explanations are desired

References

[1] Chattopadhyay et al. Interpretable by design: Learning predictors by composing interpretable queries. TPAMI 2022.
[2] Chattopadhyay et al. Variational Information Pursuit for Interpretable Predictions, ICLR 2023.
[3] Chattopadhyay et al. Bootstrapping Variational Information Pursuit with Foundation Models for Interpretable Image Classification. ICLR 2024.
[4] Chattopadhyay et al. Information Maximization Perspective of Orthogonal Matching Pursuit with Applications to Explainable AI. NeurIPS 2023.