Bootstrapping Variational Information Pursuit with Large Language and Vision Models for Interpretable Image Classification

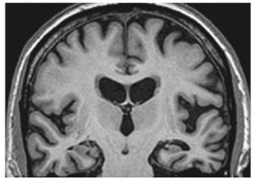
Aditya Chattopadhyay, Kwan Ho
Ryan Chan, René Vidal

VIDAL LAB
JOHNS HOPKINS WHITING SCHOOL of ENGINEERING
Penn Engineering GRASP LABORATORY General Robotics, Automation, Sensing & Perception Lab

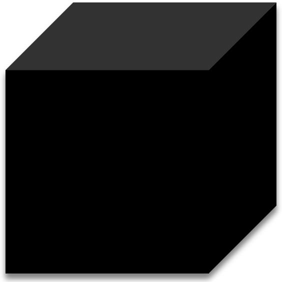
machine learning ■ trustworthy AI ■ computer vision ■ biomedical data science

Background: Need for interpretable models

Current models



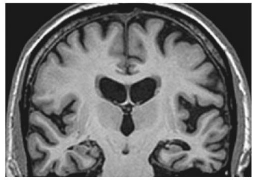
MRI Scan



Black-Box

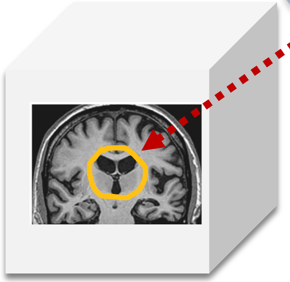
Patient has Alzheimer's disease with 98.6% probability

Desired models



MRI Scan

"Since this region is abnormally dilated…"



~~Black-Box~~

Patient has Alzheimer's disease with 98.6% probability

Background: Post-hoc explainability vs interpretable-by-design

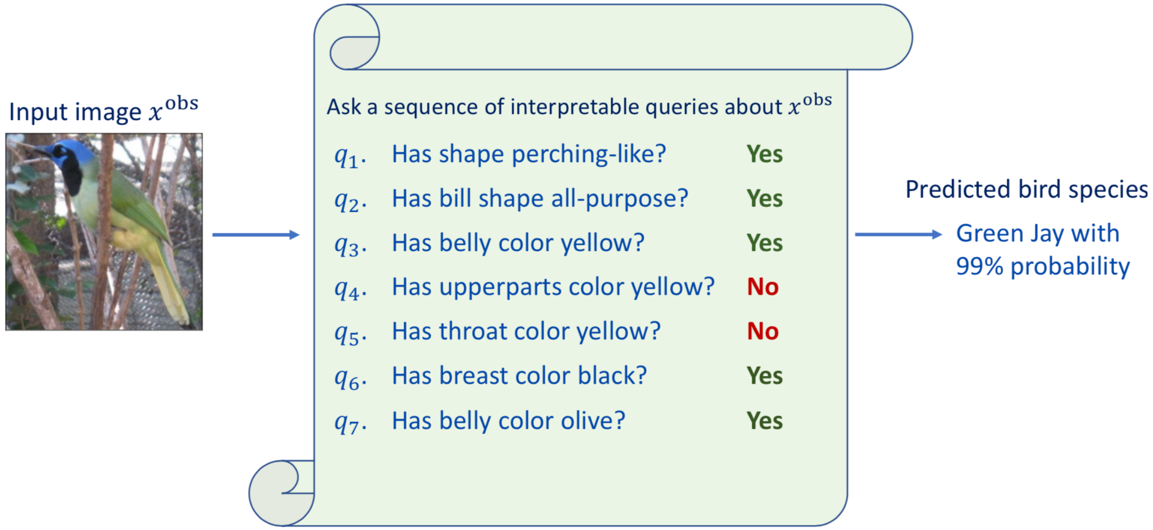
- Most current methods to explaining model decisions are post-hoc.
  - No need to retrain model, accuracy maintained.

- **Post-hoc** methods **don't provide faithful explanations.**[1]

- Need for models that are **interpretable-by-design** → provides **explanations** for decisions that are **interpretable to the user**.

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. Advances in neural information processing systems, 31.

Prior Work: Interpretable decisions via 20 Questions (20Q)



- An **interpretable-by-design** approach based on the 20-Question game.[1]
  1. Specify a large **set of interpretable queries**, Q, about the input.
  2. Given Q, **ask informative queries one at a time**. Each query choice depends on the query answers obtained so far.
  3. Once confident, make a **prediction based only** on the **obtained query-answers**.

1. Chattopadhyay, A., Slocum, S., Haeffele, B. D., Vidal, R., & Geman, D. (2022). Interpretable by design: Learning predictors by composing interpretable queries. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Motivation: How to apply 20Q for interpretable ML?

- **Applying the framework to any ML task requires:**

  1. ***Specification of query set:*** In prior work, queries were user-defined. Progress in **LLMs** make it possible to **automatically specify task-relevant query sets.**[1] ✅

  2. ***Mechanism for what to ask next:*** Use **Variational Information Pursuit (V-IP)**[2] a greedy algorithm, to select the **next most informative query** from Q.✅

  3. ***Mechanism to answer the queries:*** **A major bottleneck** → often **requires manually annotated data** for training classifiers to answer queries at test time. ❌

1. Oikarinen, T., Das, S., Nguyen, L. M., & Weng, T. W. (2023). Label-free Concept Bottleneck Models. In The Eleventh International Conference on Learning Representations.
2. Chattopadhyay, A., Chan, K. H. R., Haeffele, B. D., Geman, D., & Vidal, R. (2023). Variational Information Pursuit for Interpretable Predictions. In The Eleventh International Conference on Learning Representations.

Challenge: How to answer queries?

- Most datasets **don’t come** with **manually annotated query answers**.☹️

- **This work:** Interpretable predictions via V-IP + visual question-answering system trained to answer queries without any manual annotations.
  - Use pseudo-labels provided by pretrained Vision Language Models (VLMs) instead!

Our Contribution: Concept-QA

- Given, an image classification dataset, say Imagenet.

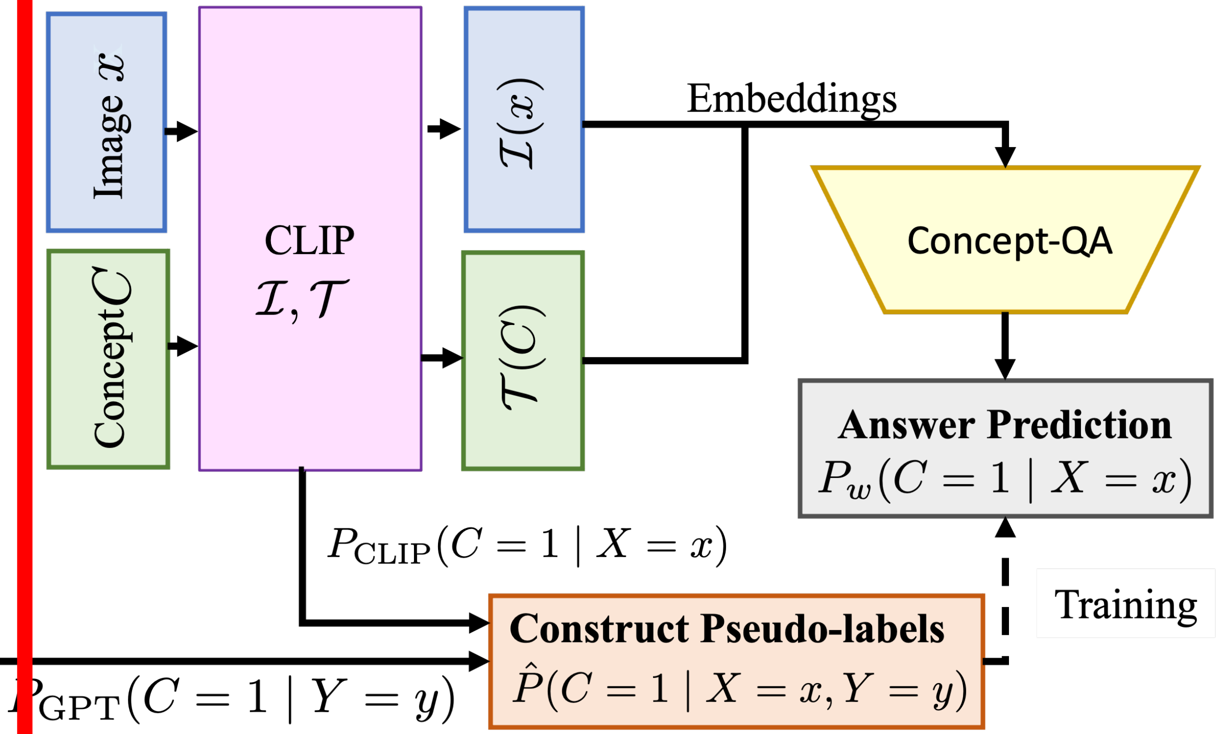
- A set of of task-relevant queries/concepts obtained from an LLM (say GPT4).

- Use pseudo-labels generated by a pretrained VLM to train Concept-QA
  – Can then use Concept-QA this to answer queries at test time

Our Proposal: Concept-QA

For every {class}:

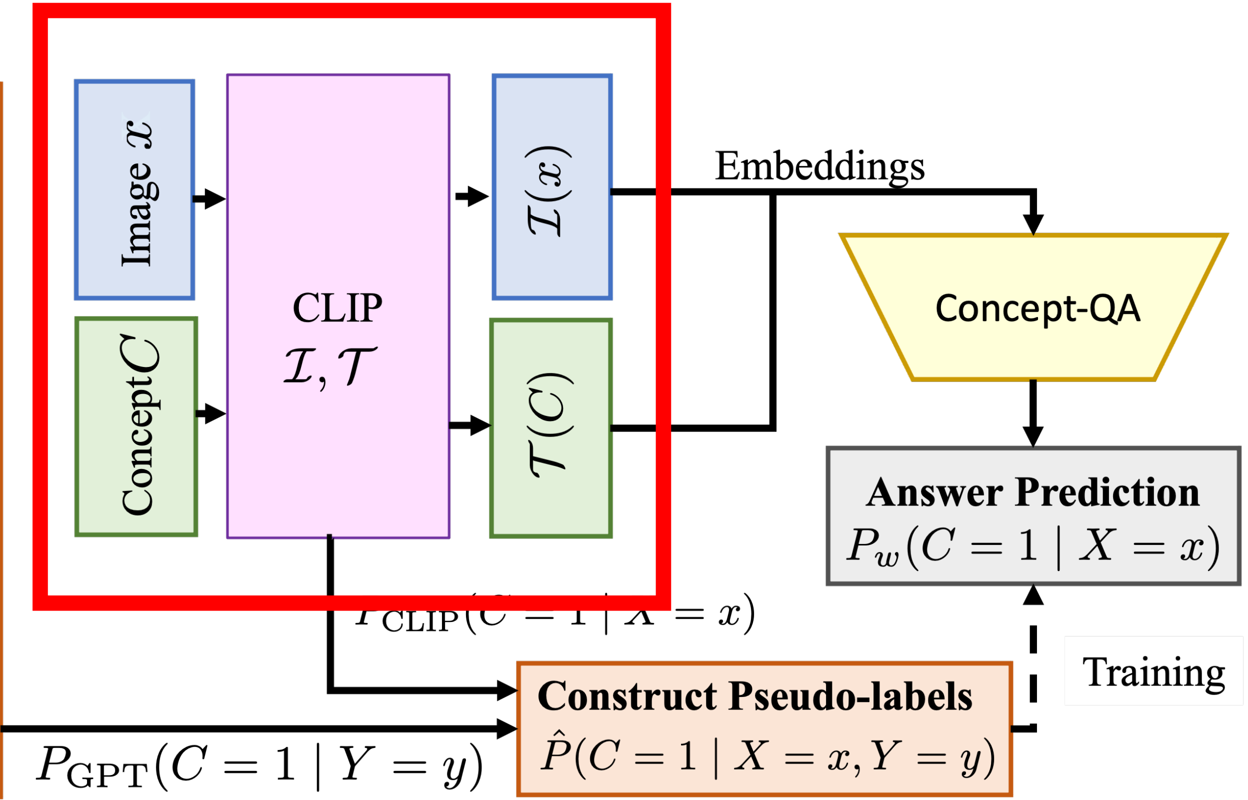
PROMPT to GPT:
Answer yes/no/depends for whether the following concepts are salient for recognizing a '{class=bear}': a predator, a prey, a pride, …
RESPONSE:
Description: A bear is a large mammal characterized by its robust…
a predator: Yes: Bears, with their sharp teeth..
a prey: Yes: Despite being a top predator…
a pride: No: Bears are typically solitary animals…
…



Our Proposal: Concept-QA

For every {class}:

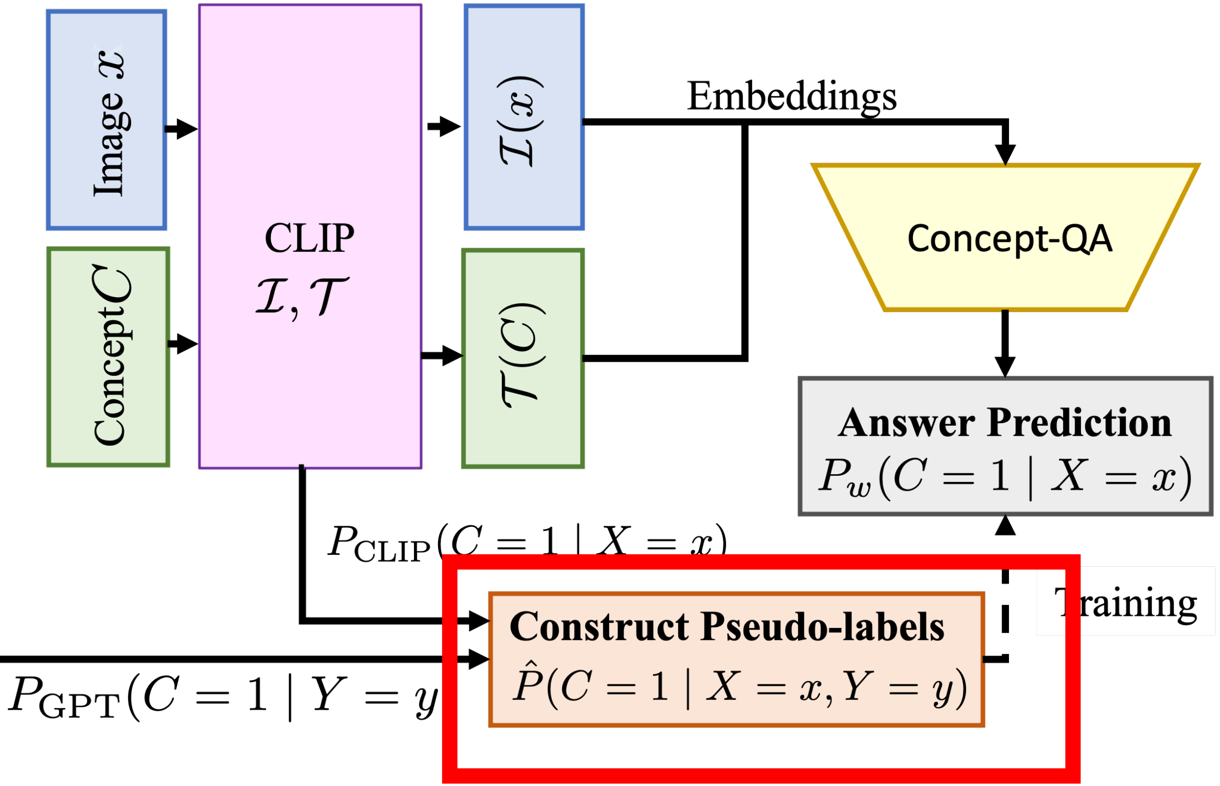
PROMPT to GPT:
Answer yes/no/depends for whether the following concepts are salient for recognizing a '{class=bear}': a predator, a prey, a pride, …
RESPONSE:
Description: A bear is a large mammal characterized by its robust…
a predator: Yes: Bears, with their sharp teeth..
a prey: Yes: Despite being a top predator…
a pride: No: Bears are typically solitary animals…
…



Our Proposal: Concept-QA

For every {class}:

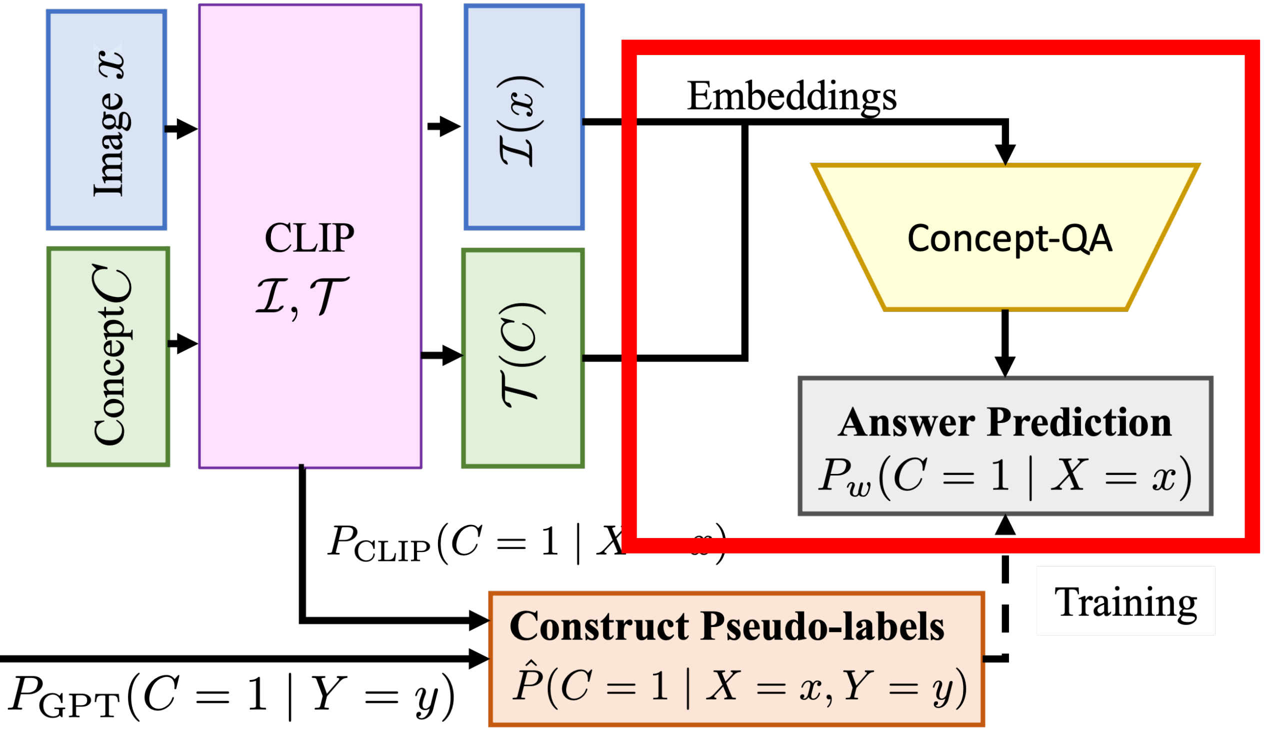
PROMPT to GPT:
Answer yes/no/depends for whether the following concepts are salient for recognizing a '{class=bear}': a predator, a prey, a pride, …
RESPONSE:
Description: A bear is a large mammal characterized by its robust…
a predator: Yes: Bears, with their sharp teeth..
a prey: Yes: Despite being a top predator…
a pride: No: Bears are typically solitary animals…
…



Our Proposal: Concept-QA

For every {class}:

PROMPT to GPT:
Answer yes/no/depends for whether the following concepts are salient for recognizing a '{class=bear}': a predator, a prey, a pride, …
RESPONSE:
Description: A bear is a large mammal characterized by its robust…
a predator: Yes: Bears, with their sharp teeth..
a prey: Yes: Despite being a top predator…
a pride: No: Bears are typically solitary animals…
…



What not directly use pretrained VLMs?

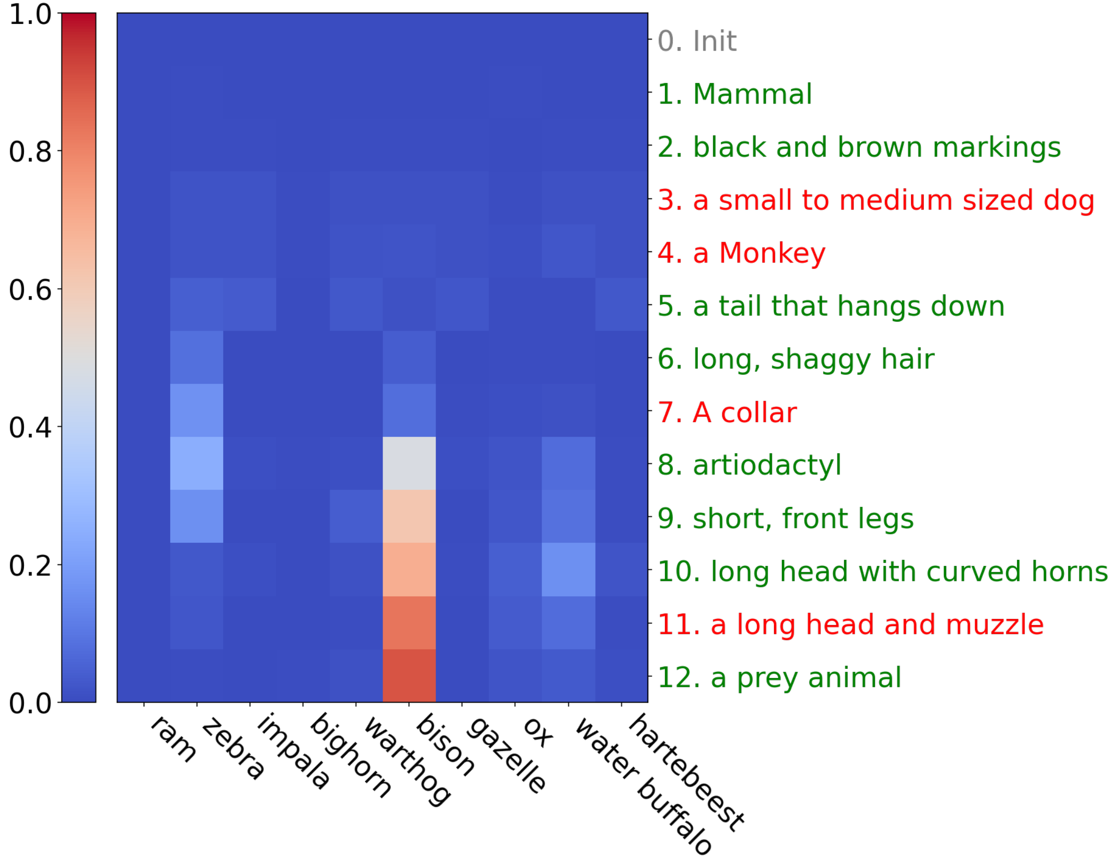
- Current **state-of-the-art VLMs** like Llava-1.5 and BLIP-2 are **too slow** to be used in an online sequential manner.
  - In contrast, **Concept-QA** is lightweight, **much faster** and **competitive**.

- **Concept-QA outperforms CLIP** in its ability to accurately answer queries.
  - Evaluated by manually annotating a subset of the dataset with query-answers.

1. Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry et al. "Learning transferable visual models from natural language supervision." In International conference on machine learning, pp. 8748-8763. PMLR, 2021.

Interpretable Predictions with VIP and Concept-QA

- **Task:** Image Classification (ImageNet)
- **Query set:** Queries about presence of different semantic concepts.



More Information,

Johns Hopkins University

Penn University of Pennsylvania

Research supported by the Army Research Office under the Multidisciplinary University Research Initiative contract W911NF-17-1-0304, the NSF grant 2031985 and by Simons Foundation Mathematical and Scientific Foundations of Deep Learning (MoDL) grant 135615.

Thank You!

https://github.com/adityac94/conceptqa_vip