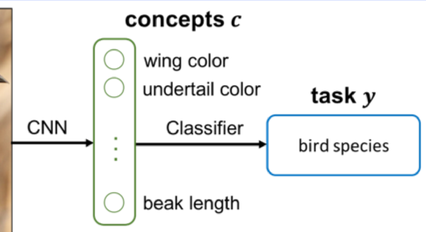
Faithful Concept Bottleneck Model

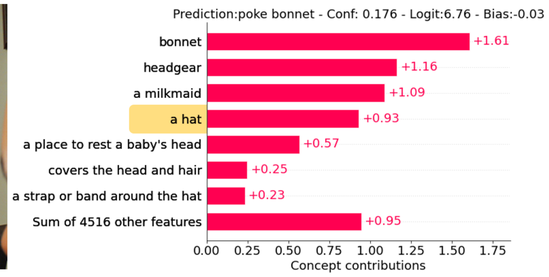
Henry Xu jix049@ucsd.edu

Mentor: Lily Weng lweng@ucsd.edu

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Overview





Intervene: activation of “a hat” 2.70 -> 0. New prediction: strainer ✓

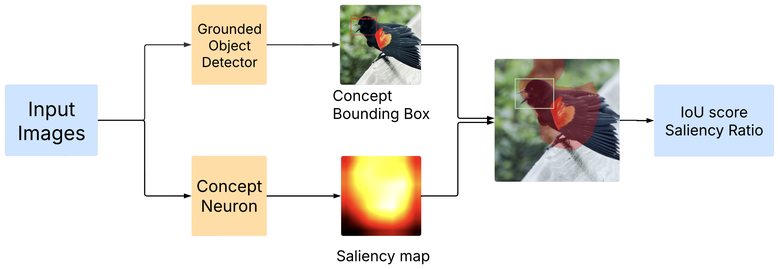
Concept Bottleneck Models (CBMs) provide interpretable prediction by introducing an intermediate Concept Bottleneck Layer (CBL), which encodes human-understandable concepts to explain models’ decision.

Existing approaches train concept neurons that fit the relevant class rather than the actual concept. We developed a faithfulness measurement method that combines Grounded Object Detection with saliency maps to ensure concept neurons pay ”attention” to the right areas, and we fine-tune CBMs to be more faithful with saliency loss.

Methods

Automated Concepts Generation: Prompt GPT4 with attribute framework and examples.
Resulted in more comprehensive concept sets and facilitates manual refinement.

What are useful visual concepts for distinguishing the bird species "Slaty backed Gull" from other birds in a photo? These features should be visually distinctable. For each item, you should be concise and precise, and use no more than five words. Each item should be a complete concept, not just a description. No punctuations. No ambiguous answers. The response should cover attributes listed in the example, you can put more than one answer to account for male and female difference or seasonal difference, or N/A for an attribute when applicable. If there are any visually important feature not covered by the attributes provided, add it as a feature of the closest body part.
For example:
Northern Flicker
bill color: gray bill
bill shape: long slightly curved bill



Faithfulness Measurement:

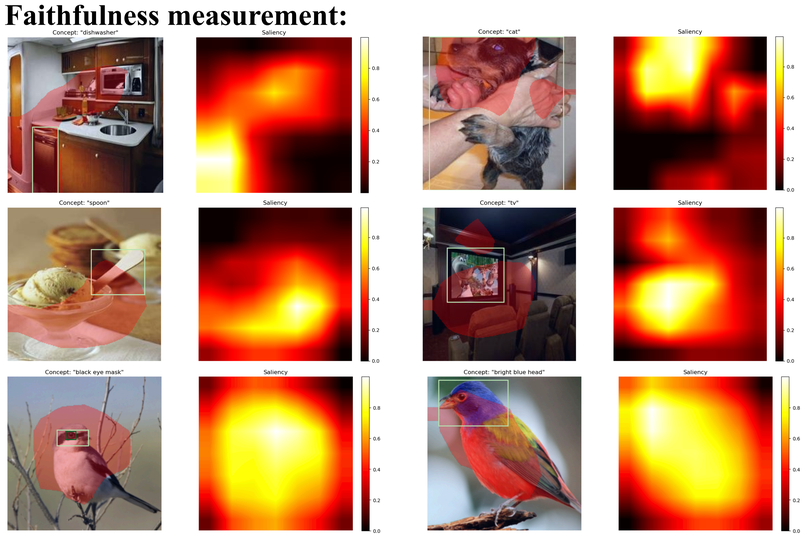
1. For each concept neuron, find top k activating images.
2. For each image, use GroundingDINO to annotate it with prompt ”class name . target concept”, obtain the GroundingDINO bounding boxes
3. Use Grad-CAM++ to obtain saliency maps.
4. Align concept bounding boxes with saliency maps and compute compute IoU score and correct/incorrect saliency ratio

Fine-tuning: In addition to the Binary Cross Entropy (BCE) loss for multi-label prediction, we defined a saliency loss to make sure the concept neurons pay attention to the correct parts of an image:

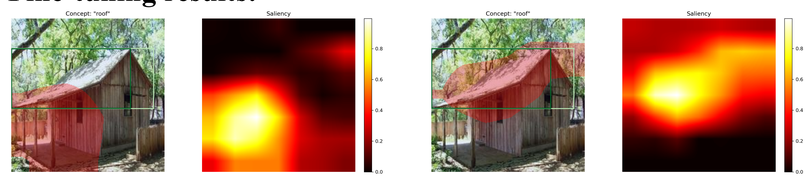
$$L_{sal} = RELU(\text{saliency outside bboxes} - \text{saliency inside bboxes} + 0.5)$$

Example Results

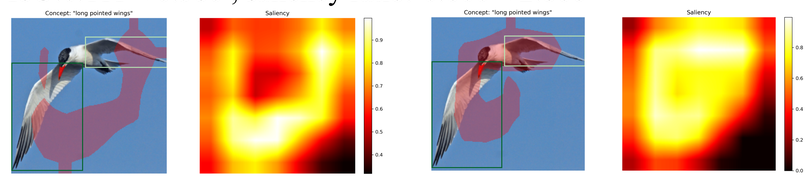
Faithfulness measurement:



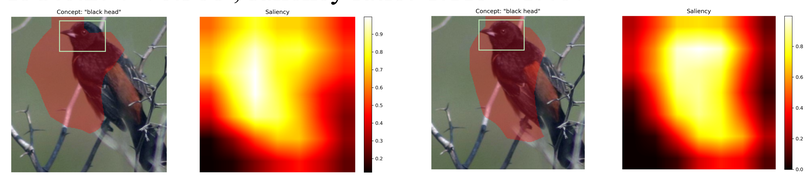
Fine-tuning results:



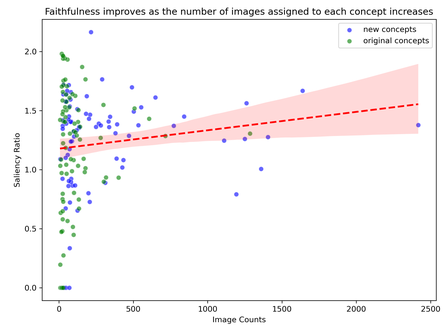
IoU:0.111→0.580 , saliency ratio: 0.600→2.866



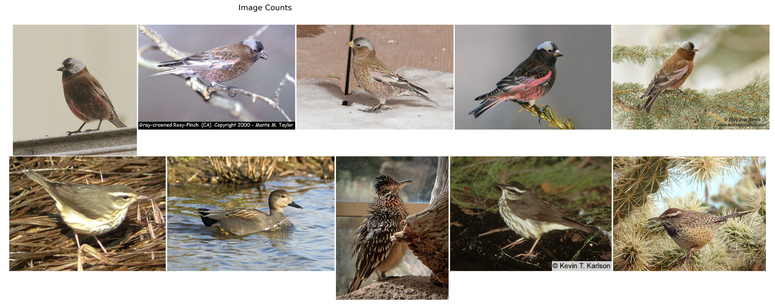
IoU:0.294→0.368 , saliency ratio: 1.115→1.467



IoU:0.150→0.176 , saliency ratio: 1.372→1.769



Concepts refinement: Faithfulness improves as concepts are correctly assigned to more classes.



Top activating images for neuron ”brown wings” in original(first row) and concepts refined model(second row). Despite brown wings is a common feature, all the images of the original model comes from a single bird species, indicating fitting to bird class instead of the true concept.