Concept-Based Explanations in Computer Vision: Where Are We and Where Could We Go?

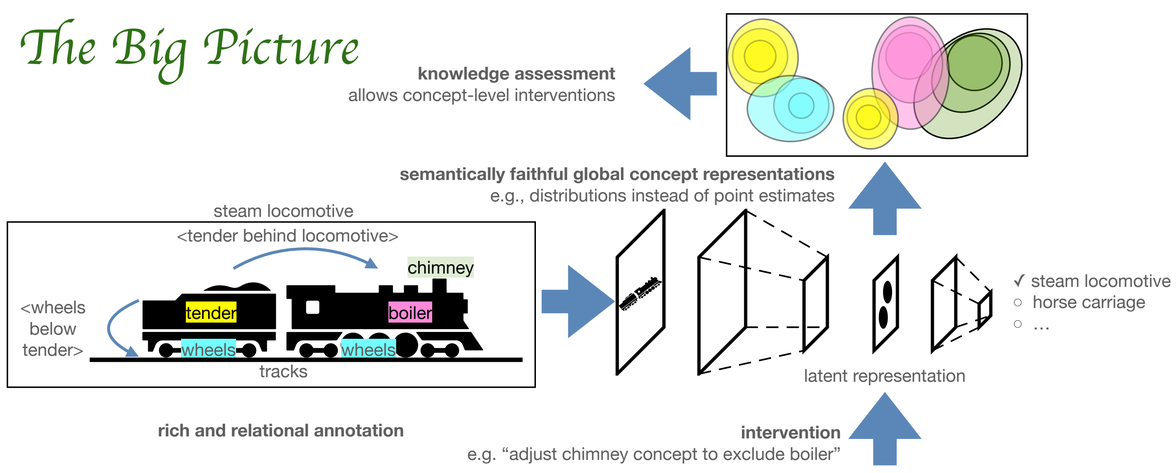
Jae Hee Lee[1], Georgii Mikriukov[2], Gesina Schwalbe[3], Stefan Wermter[1], Diedrich Wolter[3]
[1] University of Hamburg, [2] Anhalt University of Applied Sciences, [3] University of Lübeck

Motivation

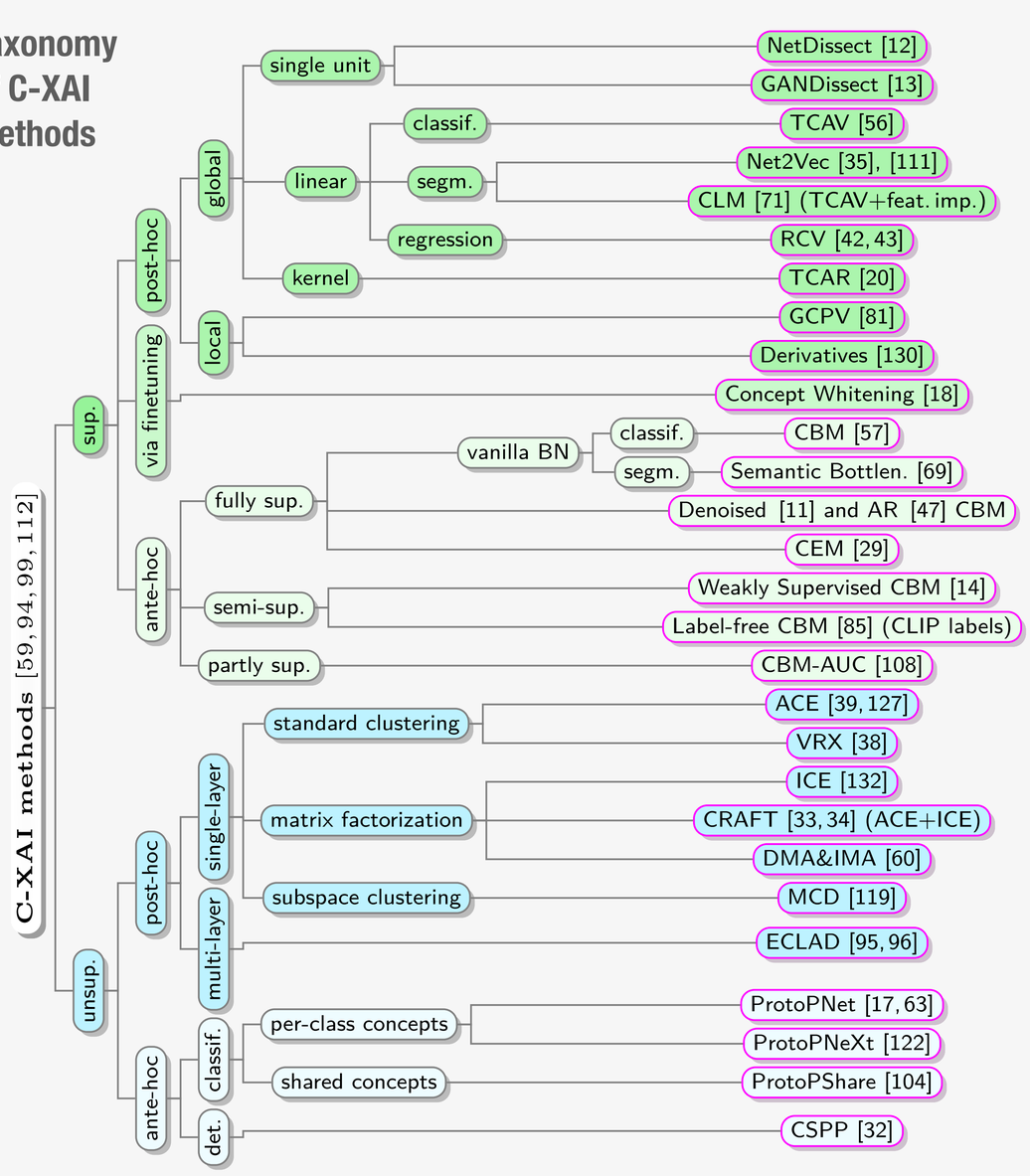
Concept-based XAI (C-XAI) explains how a vision model represents input in its intermediate layers using semantically meaningful concepts. Concepts act as a common alphabet between users and model.

Contributions

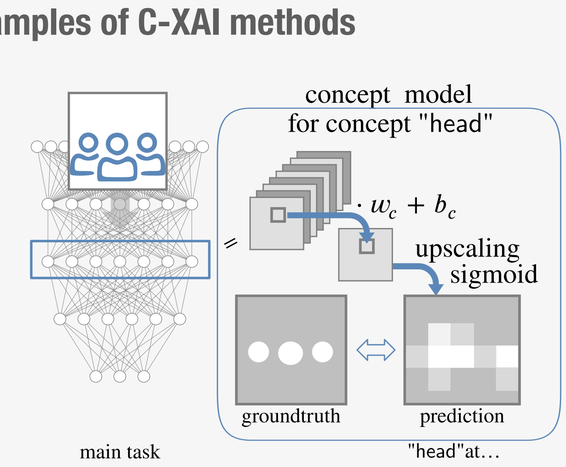
- Reviewing the state of the art in C-XAI.
- Discuss the state of the art and open challenges in
  1. extracting new concept types;
  2. beyond classical vector-based concept representations; and
  3. controlling concepts
- Discuss a potential role of ontological commitment in C-XAI.



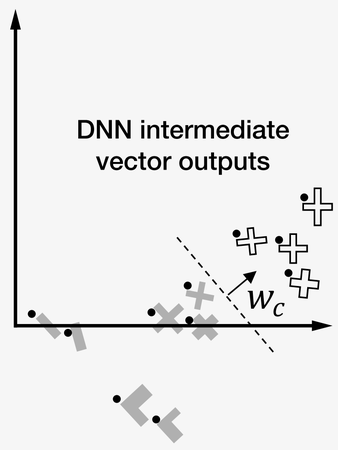
C-XAI Review



Examples of C-XAI methods



Net2Vec



TCAV

1. Extracting New Concept Types

Background

- Existing concept types are limited and the coverage can be extended.
  - image-level scene attributes *(e.g., sunny)*.
  - image qualities *(e.g., contrast)*.
  - attributes of image regions such as object *(e.g., person)* and object part classes *(e.g., beak)*.
  - object attributes such as material, texture, and color.

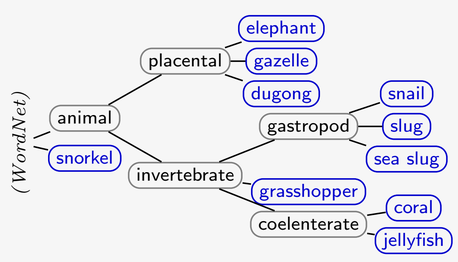
Challenges: How to extract …

A. temporal and multimodal concepts *(e.g., from videos)*?
B. concepts in a self-supervised way *(e.g., from videos)?*
C. concepts from new architectures *(e.g., ViTs)*?

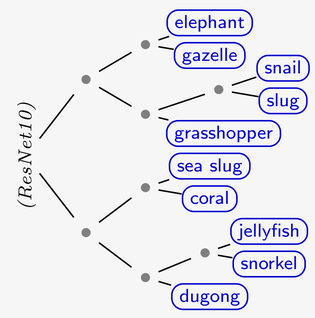
2. Improving Concept Representations

Background

- Existing concept representations: single neurons, vector-based representations, subspaces, latent space regions, or hierarchies of point estimates.
- Commonsense knowledge about concepts is essential for semantics.
  - *E.g.: Since* IsPartOf(head, person)*, the presence of a head implies the presence of a person.*
- Ontological commitment refers to the catalog of defined concepts and relations *(e.g.,* IsSimilarTo(cat, dog), IsSubclassOf(cat, animal), IsPartOf(head, person)*)*.
- Manually crafted, large ontologies aim to capture the ontological commitment of human common sense *(e.g., WordNet)*.
- To connect these sources of information to C-XAI one has to ground ontology concepts in network activation.



WordNet



ResNet10

Challenges: How to …

A. generalize concept representation *(e.g., to regions/distributions)*?
B. identify the ontological commitments in trained models?

3. Controlling Concepts

Background

- Given concepts $c_1, \ldots, c_n$, we can regard the corresponding concept vectors as a concept basis.
- The i^{th} coordinate of an activation x with respect to that concept basis captures the strength of the presence of concept c_i in x.
- Intervention on concept c_i: increasing/decreasing the i^{th} coordinate leads to increasing/decreasing the presence of the concept in x.

Challenges: How to …

A. apply logical constraints to the activations with varying expressivity of the logical constraints?
B. guide the model training and globally modify intermediate representations?
C. avoid catastrophic forgetting (i.e., retaining previously learned knowledge) in new tasks in a lifelong learning scenario?
D. identify side effects for a specific concept control mechanism and how to avoid them?