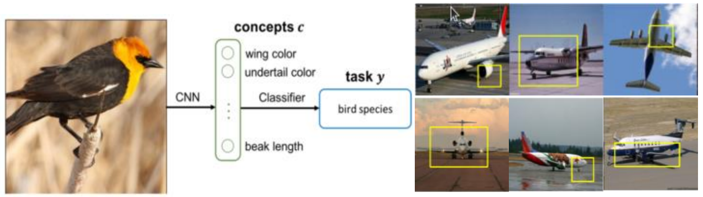
Fraunhofer IKS

Towards Human-Interpretable Prototypes for Visual Assessment of Image Classification Models

Poulami Sinhamahapatra, Lena Heidemann, Maureen Monnet, Karsten Roscher

AAAI 2023
Washington, DC, USA

Concept-based vs Prototype-based Learning



- Explicity specified concepts
- Need domain expert for concept annotations

- Implicitly learnt without supervision
- Finds dominant representative parts from whole dataset

→ **Can AI models be reliably explained using human-interpretable concepts?**

→ **Can we still learn human-interpretable concepts without requiring concept annotations?**

We need **prototype-based learning methods** which do not need concept annotations, are interpretable-by-design and global.

Motivation:

- How interpretable are these prototypes towards assessment of Image Classification models?
- What are the conditions these prototypes should fulfill towards a truly human-interpretable model?

Key Contributions:

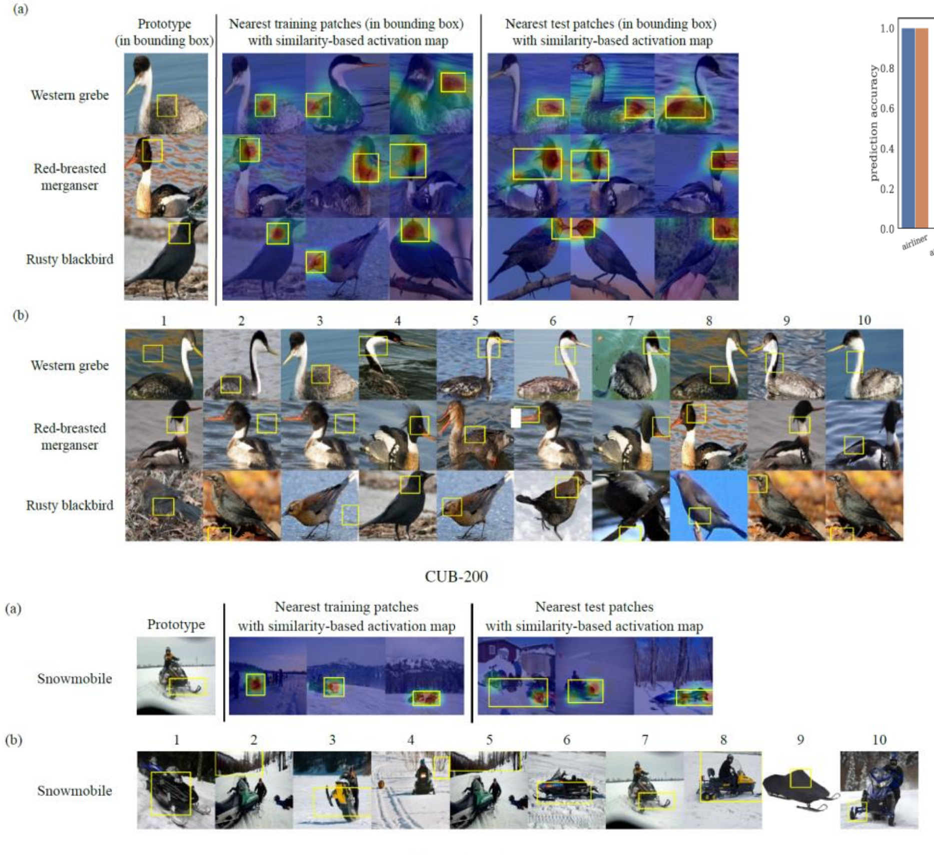
- ✓ Proposed a **Desiderata** for truly human-interpretable prototypes
- ✓ Designed a common setup to evaluate the existing methods (ProtoPNet, ProtoTree, PRP) in the light of these properties with real (CUB, IM30) and synthetic (3D Shapes) datasets
- ✓ Validated our findings quantitavely by conducting a user-study
- ✓ Demonstrated the application of prototype-based learning on real-world use-case of *OOD Detection*

***Desiderata* for the prototypes to be truly human-interpretable:**

- Human-understandable / interpretable
- Semantically disentangled
- Semantically transformation invariant
- Relevant to the learnt task

Full Paper

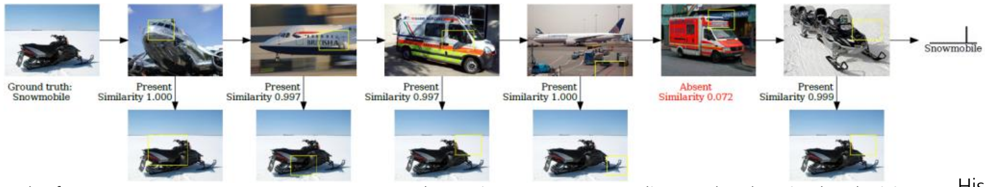
Qualitative Results (Real-world Dataset)


(a)
Prototype (in bounding box)
Nearest training patches (in bounding box) with similarity-based activation map
Nearest test patches (in bounding box) with similarity-based activation map
Western grebe
Red-breasted merganser
Rusty blackbird
(b)
1
2
3
4
5
6
7
8
9
10
Western grebe
Red-breasted merganser
Rusty blackbird
(a)
Prototype
Nearest training patches with similarity-based activation map
Nearest test patches with similarity-based activation map
Snowmobile
(b)
1
2
3
4
5
6
7
8
9
10
Snowmobile

CUB-200

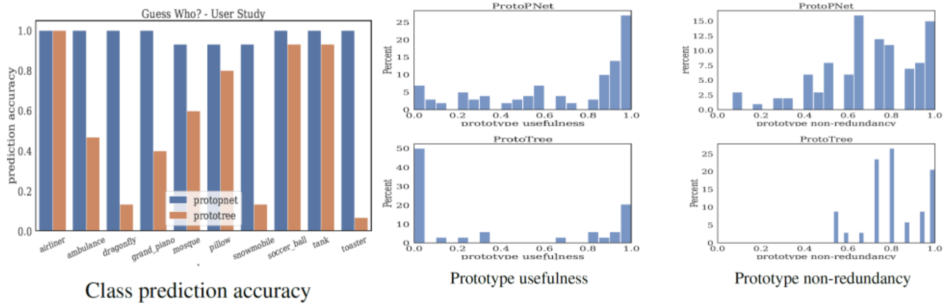
ImageNet-30

Results for ProtoPNet using CUB-200 (top) and ImageNet-30 (bottom): (a) shows for a prototype from a given class - the nearest training and test patches (yellow box) with similarity score based activation maps, (b) shows all the 10 prototypes (yellow box) learnt for the respective classes in (a).



Results from ProtoTree on ImageNet-30: For each test image, corresponding path taken in the decision tree towards final prediction is shown. The node and absent prototypes are shown in yellow and red.

Quantitative Results (User Study)



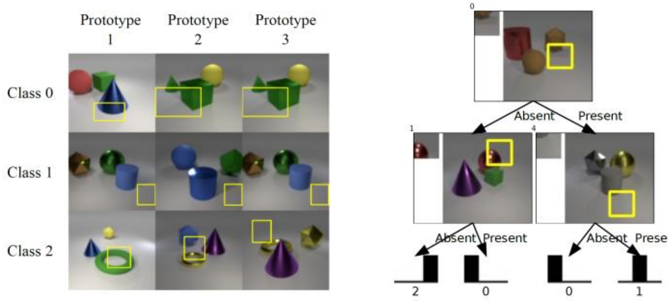
Class prediction accuracy

Prototype usefulness

Prototype non-redundancy

Quantitative results from user study conducted on prototypes from ImageNet-30 classes

Qualitative Results (Synthetic Dataset)

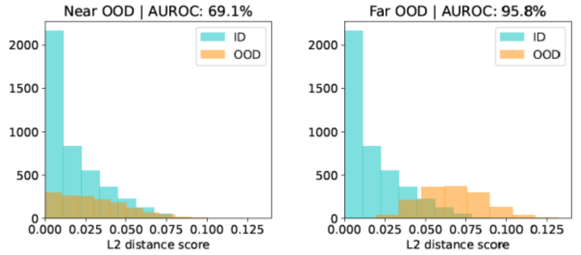


ProtoPNet

ProtoTree

Results using synthetic 3D-Shapes dataset (V1) showing in (a) prototypes from each class using ProtoPNet and (b) the node prototypes along with decision tree (depth=2) using ProtoTree. Yellow boxes show the prototypes.

Real-world application (OOD Detection)



Histogram showing distribution of L2 distances to closest prototypes for Near vs Far OOD samples for a model trained on first 150 classes of CUB-200 dataset.