Towards Human-Interpretable Prototypes for Visual Assessment of Image Classification Models
Poulami Sinhamahapatra, Lena Heidemann, Maureen Monnet, Karsten Roscher

Concept-based vs Prototype-based Learning

- Explicitly 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 quantitatively by conducting a user-study
- Demonstrated the application of prototype-based learning on real-world use-case of OOD Detection

Full Paper