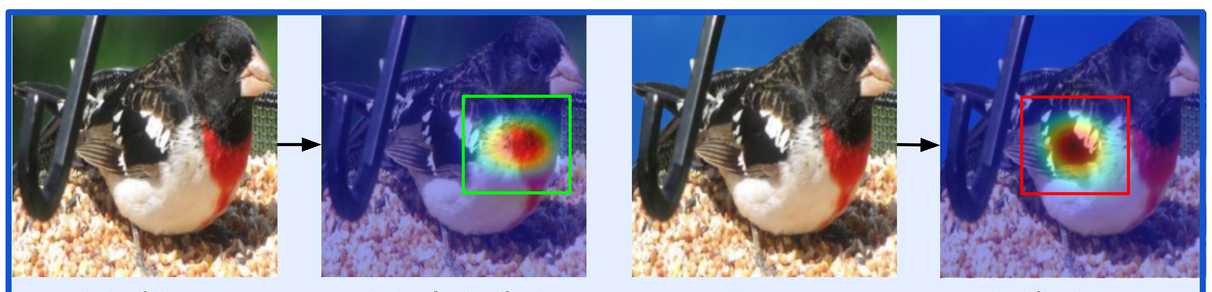
Interpretability Benchmark for Evaluating Spatial Misalignment of Prototypical Parts Explanations

Mikołaj Sacha[1,2], Bartosz Jura[3], Dawid Rymarczyk[1,4], Łukasz Struski[1], Jacek Tabor[1], Bartosz Zieliński[1,4]

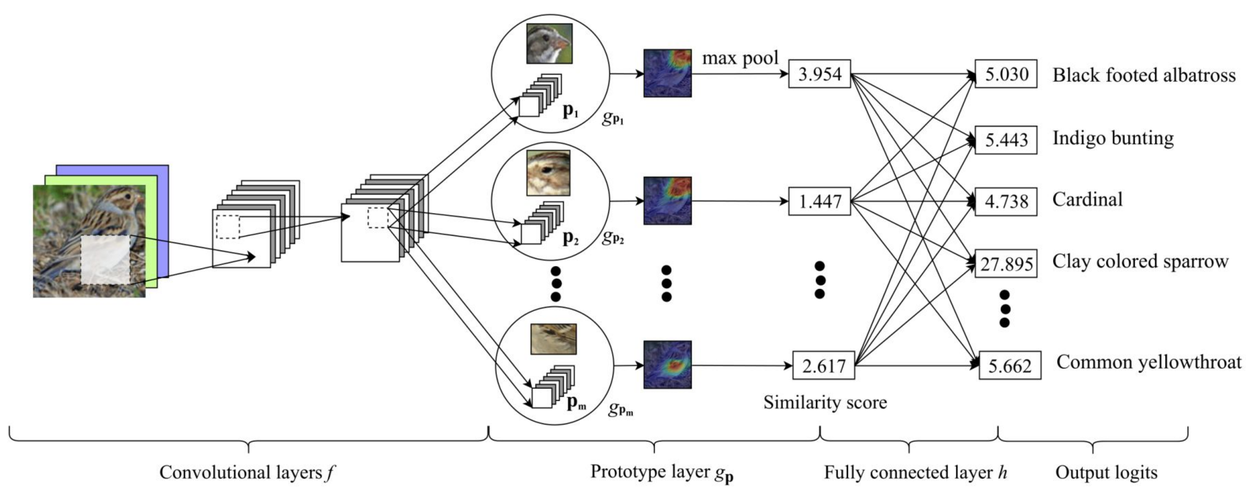
[1] Faculty of Mathematics and Computer Science, Jagiellonian University
[2] Doctoral School of Exact and Natural Sciences, Jagiellonian University
[3] Łukasiewicz Research Network – Poznań
[4] Ardigen SA
[5] IDEAS NCBR

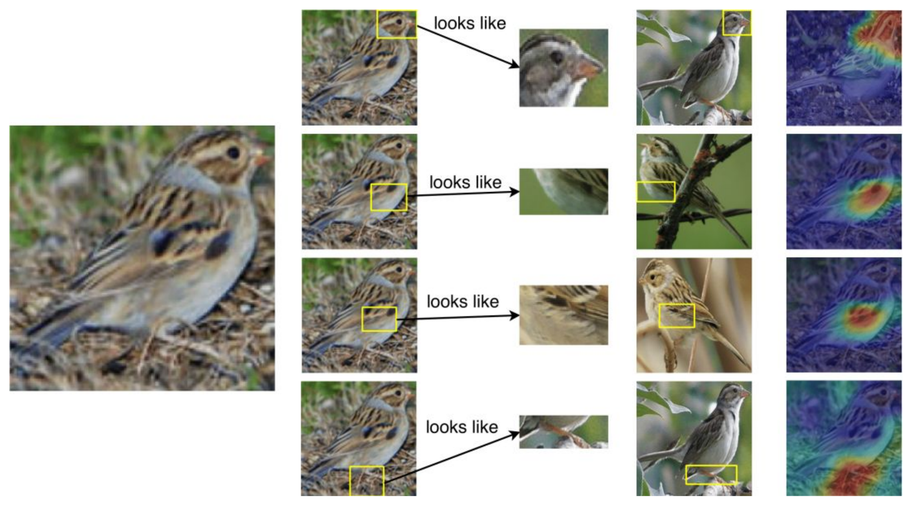
- **Prototypical part-based networks** perform interpretable image classification by comparing regions of images to *prototypical parts*.
- We examine ***spatial misalignment*** between the explanations provided by prototypical part-based networks and their actual inference mechanism.
- **We measure spatial misalignment** for popular types of prototypical part-based network and **introduce a method for enforcing learning truthful explanations.**



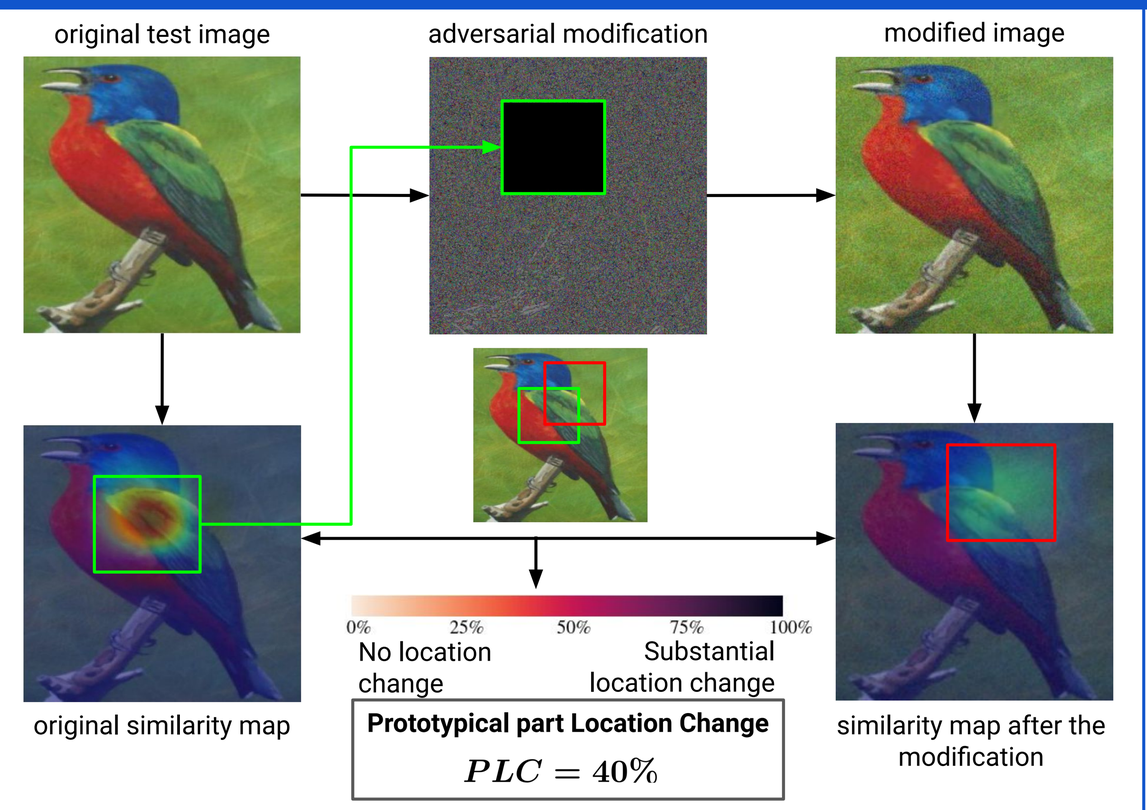
The receptive field of the prototypical part activation region can depend on parts of the image outside this region. This leads to misleading explanation given by the prototypical part similarity map.

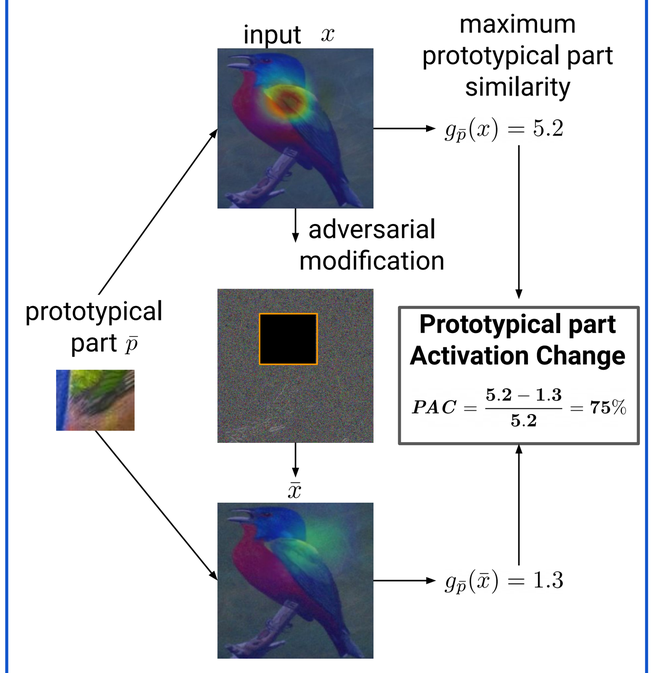
Preliminary: Prototypical Part-Based Networks [1]

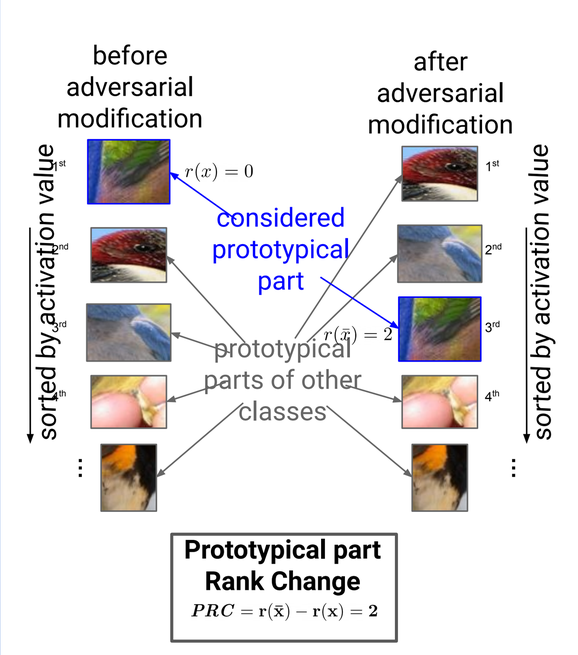




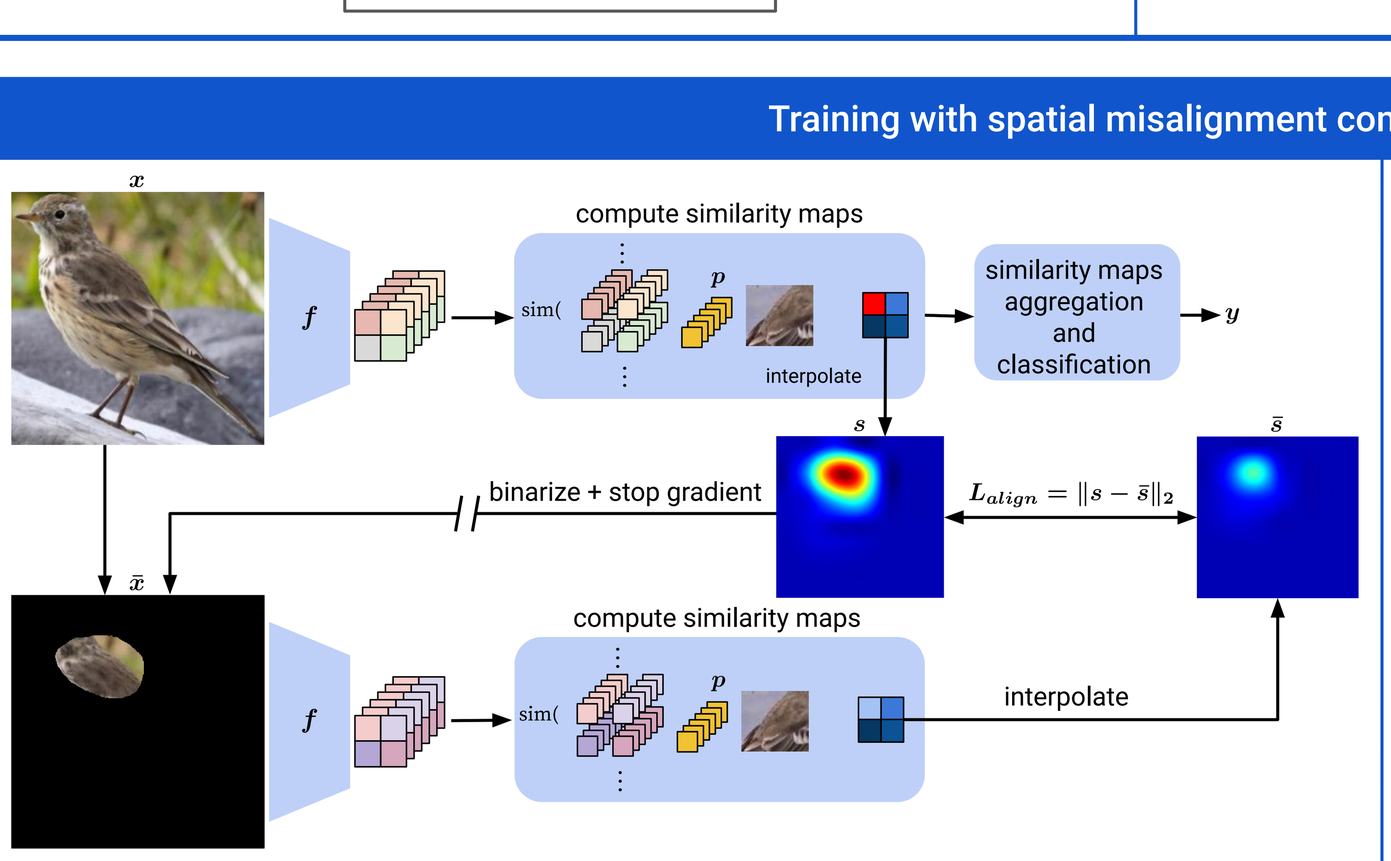
Interpretability Benchmark

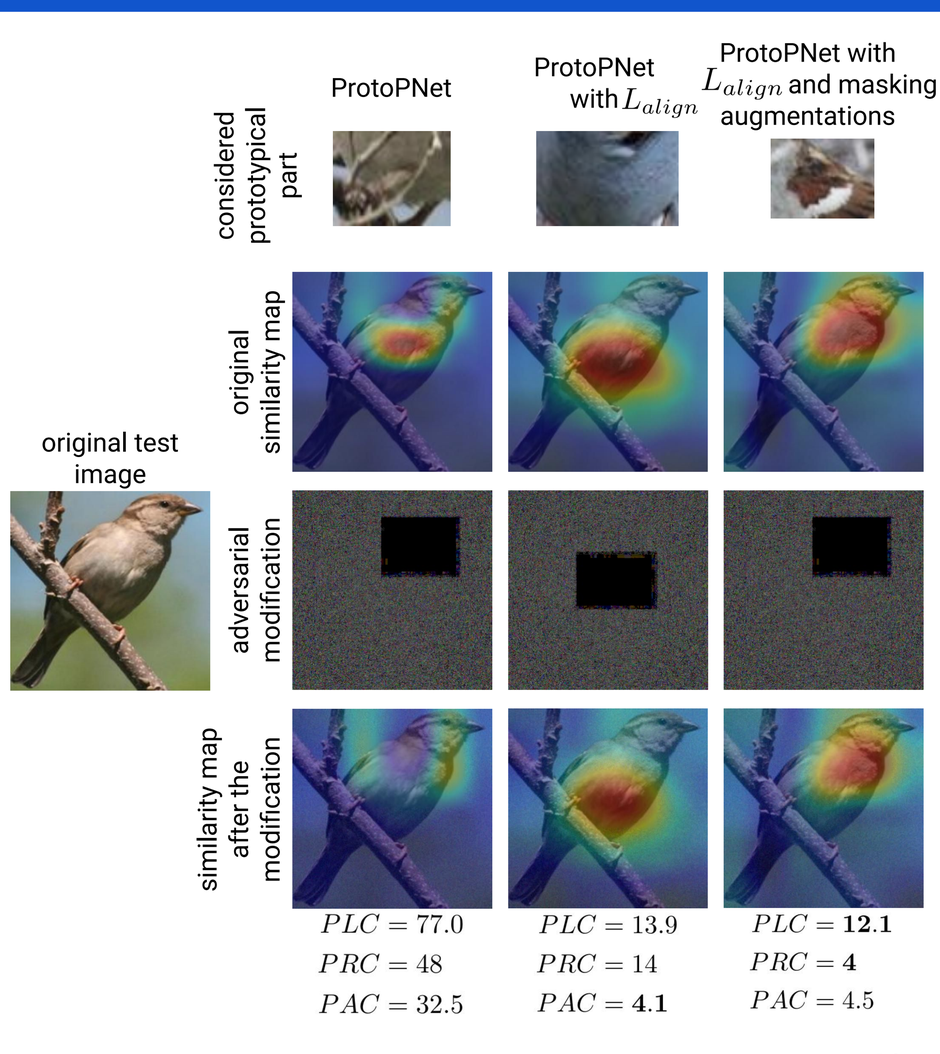


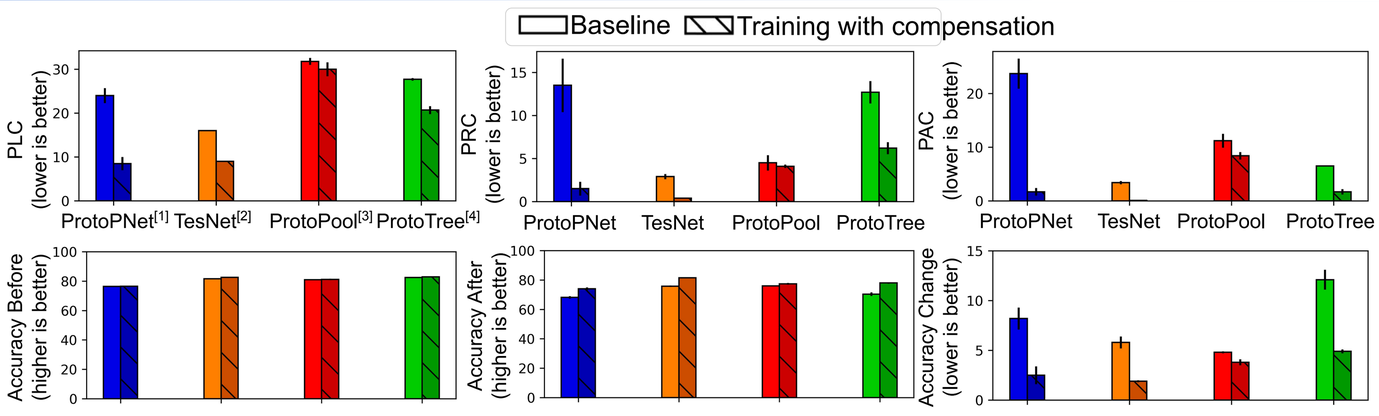




Training with spatial misalignment compensation







References

[1] *This Looks Like That: Deep Learning for Interpretable Image Recognition.* Chen, C. et al. NeurIPS 2019
[2] *Interpretable Image Recognition by Constructing Transparent Embedding Space.* Wang, J. et al. ICCV 2021
[3] *Interpretable image classification with differentiable prototypes assignment.* Rymarczyk, D. et al. Computer Vision-ECCV 2022
[4] *Neural Prototype Trees for Interpretable Fine-grained Image Recognition.* Nauta, M. et al. CVPR 2021

arxiv gmum.net repository

IDEAS NCBR · group of machine gmum learning research · JAGIELLONIAN UNIVERSITY IN KRAKÓW