Interpretability Metrics for Image Segmentation Models

Annamária Szenkovits

Babeș–Bolyai University, Faculty of Mathematics and Computer Science, Cluj-Napoca

16th Conference on Mathematics and Informatics with Applications
November 29, 2025

Overview

Measuring interpretability?

- ideally: feedback from real human evaluators
    - expensive
- **Quantification** and automated evaluation of explanation quality
    - definition of metrics
    - no clear consensus
    - dynamically evolving field: new models and new metrics appear frequently

Literature review – some metric types [4]

- 361 reviewed papers from 2014–2020: explainable models
- **12 metrics** proposed for the **qualitative evaluation** of model-generated explanations

- 3 main categories:
    - **content**: examines the faithfulness and completeness of the explanation compared to the explained black-box model
    - **presentation**: concerns the format and layout of the explanation
    - **user**: evaluates the explanation's effect on the user and how it meets user needs

Literature review – some metric types [4]

Metric type	Metric description
fidelity	how accurately the explanation reflects the behavior of the explained black-box model
completeness	to what extent the explanation covers the behavior of the explained black-box model
consistency	how consistent and deterministic the explanation is

Table: Metrics related to the **content** of explanations

Literature review – some metric types [4]

Metric type	Metric description
fidelity	how accurately the explanation reflects the behavior of the explained black-box model
completeness	to what extent the explanation covers the behavior of the explained black-box model
consistency	how consistent and deterministic the explanation is

Table: Metrics related to the **content** of explanations

Literature review – some metric types [4]

Metric type	Metric description
compactness	the size of the explanation
composition	the visual format of the explanation

Table: Metrics related to the **presentation** of explanations

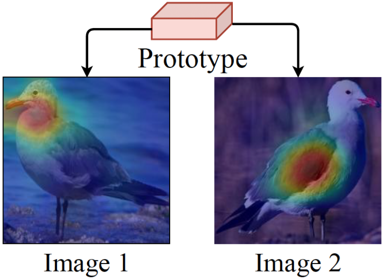
Literature review – some metric types [4]

Metric type	Metric description
context	how relevant the explanation is to the user and their needs
coherence	how well the explanation aligns with the user's prior knowledge

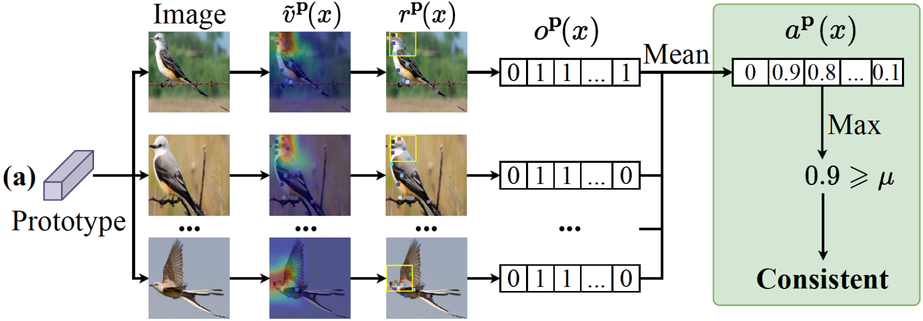
Table: Metrics measuring the **user-related** aspects of explanations

Prototype consistency in image classification models

- starting point: [2] — introduced **interpretability metrics** for prototype-based image classification models (e.g. ProtoPNet [1])
- measuring prototype **consistency**: how consistently a given prototype is activated in the same object regions (e.g. bird beak) across different images



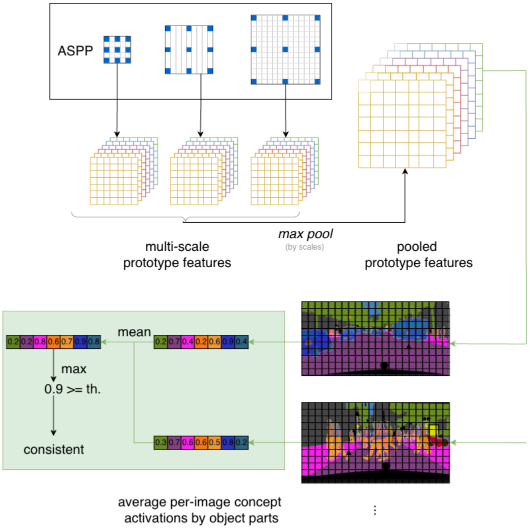
Prototype consistency in image classification models – ProtoPNet



Concept consistency in image segmentation models

- adapting the presented metric to segmentation
- dataset used: **Cityscapes-Panoptic-Parts** [3]
- an **extended version** of the Cityscapes dataset: within each semantic class, the parts of the objects are also annotated, e.g.:
    - **persons**: torso, head, arm, leg
    - **cars**: wheel, windshield, license plate

Concept consistency in image segmentation models



Preliminary results

threshold	number of consistent concepts
0.6	107/256
0.7	82/256
0.8	7/256

Table: Number of consistent concepts of the segmentation model at different thresholds, on the Cityscapes dataset, with a training mIoU score of 0.64

Next steps

- definition of additional metrics
- applying the metrics to the ProtoSeg model for comparison
- experiments on other datasets (e.g. Pascal VOC)

References I

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su.
This looks like that: deep learning for interpretable image recognition.
Advances in Neural Information Processing Systems, 32, 2019.

Qihan Huang, Mengqi Xue, Wenqi Huang, Haofei Zhang, Jie Song, Yongcheng Jing, and Mingli Song.
Evaluation and improvement of interpretability for self-explainable part-prototype networks.
In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2011–2020, 2023.

References II

Panagiotis Meletis, Xiaoxiao Wen, Chenyang Lu, Daan de Geus, and Gijs Dubbelman.
Cityscapes-Panoptic-Parts and PASCAL-Panoptic-Parts datasets for Scene Understanding.
arXiv preprint arXiv:2004.07944, 2020.

Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert.
From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai.
ACM Computing Surveys, 55(13s):1–42, July 2023.

Thank you for your attention!

This research is supported by the project “**Romanian Hub for Artificial Intelligence – HRIA**”, Smart Growth, Digitization and Financial Instruments Program, 2021–2027, MySMIS no. 351416, and the Hungarian Academy of Sciences through the **Domus Scholarship** no. 62/22/2025/HTMT.