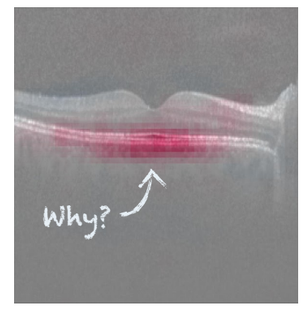
Improving Explainability of Disentangled Representations using Multipath-Attribution Mappings

Novel framework combining disentangled representations with multipath-attribution, yielding enhanced interpretability and generalisation on medical datasets.

Authors: Lukas Klein [1 2 3], João Carvalho [3], Mennatallah El-Assady [3], Paolo Penna [3], Joachim Buhmann [3], Paul Jäger [1 2]
Affiliations: [1] Helmholtz Imaging, [2] DKFZ, [3] ETH Zürich

Problem Statement

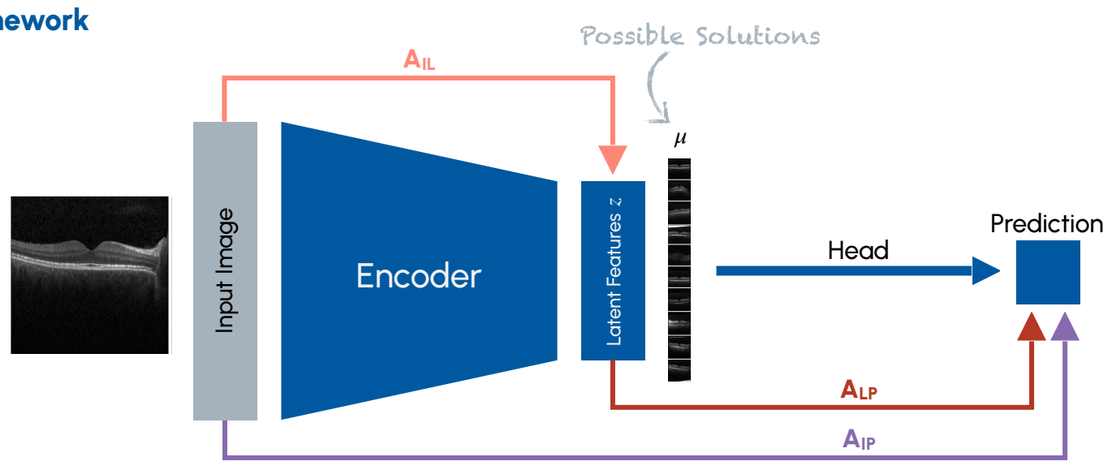
- Feature importance through **attribution maps fall short when explaining *why* a visual feature is used**, limiting causal statements and model interpretability.
- The attribution map on the right does not explain why the white retinal layer is important to the model's prediction of a healthy state. Could it be due to e.g. its shape, brightness, or thickness?



Our Approach

- Framework based on capturing semantically meaningful features in disentangled latent features.
- Verifying (PoC 1) and enhancing (PoC 2) interpretations through multi-path attribution maps.
- Allowing for qualitative shortcut detection w/o OOD test set and explaining why a model fails or generalises.

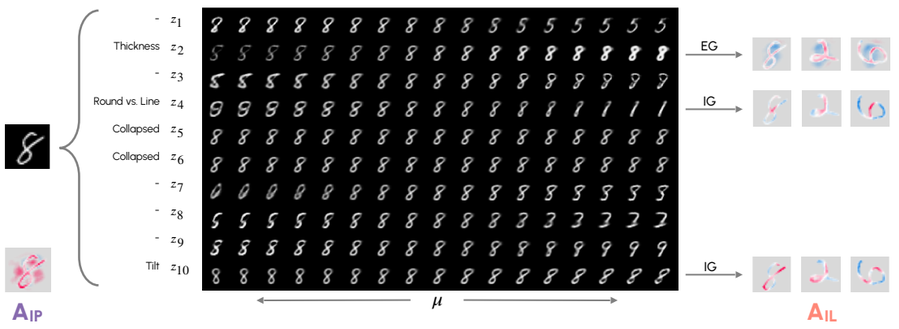
Framework



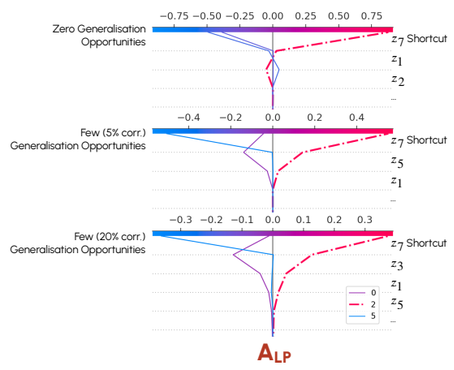
Attribution-paths:

A_{IP}: Input Image into Prediction. A_{LP}: Disentangled Latent Features into Prediction. A_{IL}: Input Image into dis. Latent Features.

PoC 1: Verifying the interpretation of disentangled Latent Features by A_{IL}.



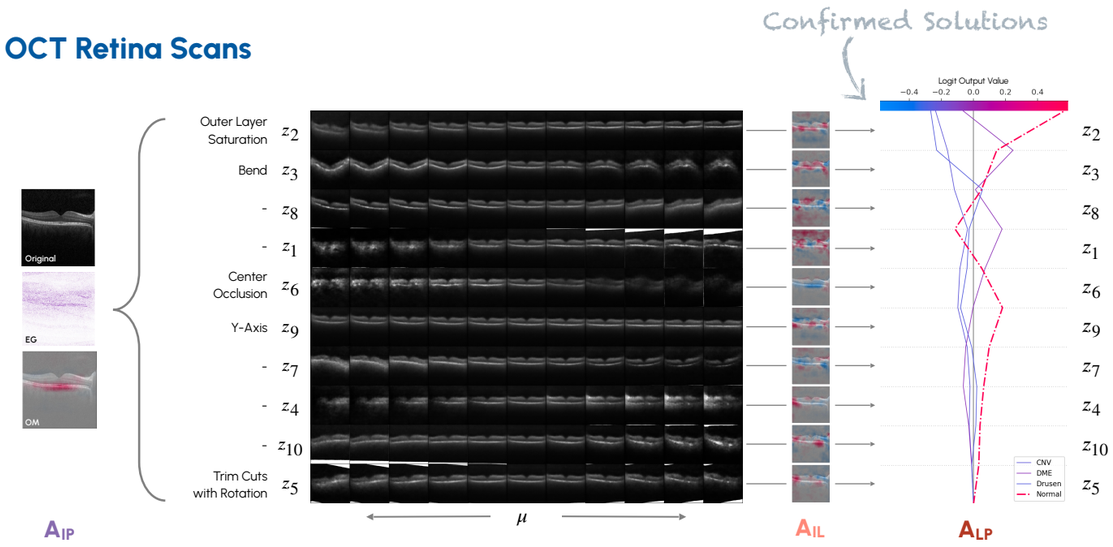
PoC 2: Using enhanced interpretability by A_{LP} for shortcut detection and generalisation evaluation.



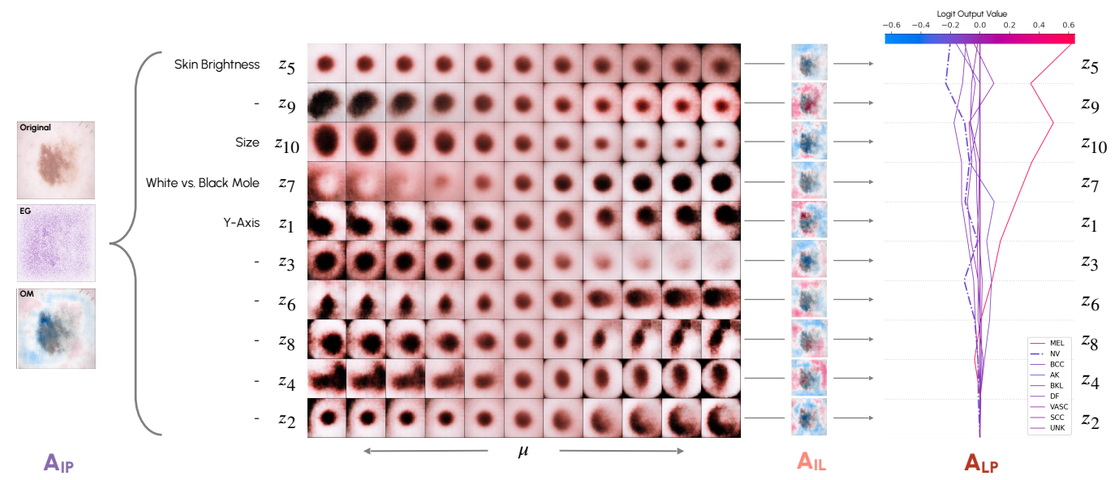
Experiments

- OCT Retina Scans: (1) disentangles a possible shortcut (2) enhances interpretability by explaining why the white retinal layer is used for prediction (3) corrects a wrong interpretation based on the classical attribution map, indicating that the trim cuts are used as shortcuts.
- ISIC Skin Lesions: (1) enables interpretability where classical attribution maps are uninformative (2) disentangles possible shortcuts (3) explains why the model fails.

OCT Retina Scans



ISIC Skin Lesions

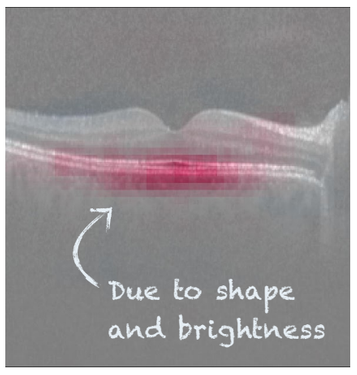


Results

The framework:

1. catalyses **more informative causality statements** than classical saliency-maps
2. facilitates qualitative **detection of shortcut learning**, and
3. enables **verification of model generalisation**,

all combined and in an interactive setting.



https://bit.ly/3w34pHF Paper

https://bit.ly/3Fcnz1M Code

HELMHOLTZ IMAGING · ETH zürich · dkfz. GERMAN CANCER RESEARCH CENTER IN THE HELMHOLTZ ASSOCIATION