Contribution ID: **7** Type: **not specified**

Improving Explainability of Disentangled Representations using Multipath-Attribution Mappings

Explainable AI aims to render model behavior understandable by humans, which can be seen as an intermediate step in extracting causal relations from correlative patterns. Due to the high risk of possible fatal decisions in image-based clinical diagnostics, it is necessary to integrate explainable AI into these safety-critical systems. Current explanatory methods typically assign attribution scores to pixel regions in the input image, indicating their importance for a model's decision. However, they fall short when explaining why a visual feature is used. We propose a framework that utilizes interpretable disentangled representations for downstream-task prediction. Through visualizing the disentangled representations, we enable experts to investigate possible causation effects. Additionally, we deploy a multi-path attribution mapping for enriching and validating explanations. We demonstrate the effectiveness of our approach on a synthetic benchmark suite and two medical datasets.

I want to give an oral presentation.

yes

I want to present a poster.

yes

Primary author: Mr KLEIN, Lukas (DKFZ, Helmholtz Imaging)

Co-authors: Mr CARVALHO, João (ETH Zürich); Mrs EL-ASSADY, Mennatallah (ETH Zürich); Prof. BUHMANN, Joachim (ETH Zürich); Dr PENNA, Paolo (ETH Zürich); Dr JÄGER, Paul (DKFZ, Helmholtz Imaging)

Presenter: Mr KLEIN, Lukas (DKFZ, Helmholtz Imaging)

Session Classification: Poster Session