Evaluating the Fidelity of Explanations for Convolutional Neural Networks in Alzheimer's Disease Detection

Bjarne C. Hiller

2025-03-09

University of Rostock

Who are we?

University of Rostock

Bjarne Hiller

Martin Becker

Sebastian Bader

Thomas Kirste

DZNE

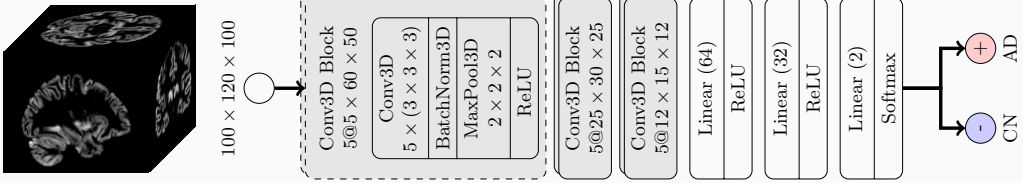
Martin Dyrba

Devesh Singh

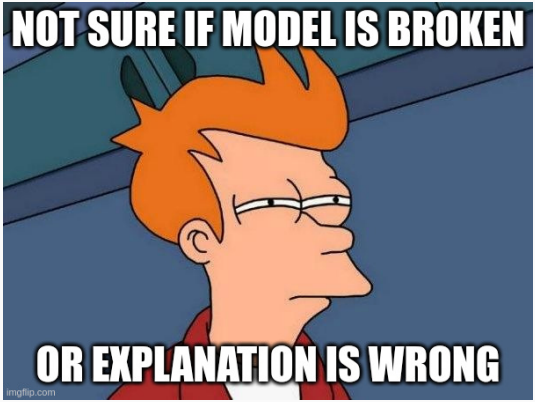
Deep Learning for Medical Image processing

Data

Model


100 × 120 × 100
Conv3D Block
5@5 × 60 × 50
Conv3D
5 × (3 × 3 × 3)
BatchNorm3D
MaxPool3D
2 × 2 × 2
ReLU
Conv3D Block
5@25 × 30 × 25
Conv3D Block
5@12 × 15 × 12
Linear (64)
ReLU
Linear (32)
ReLU
Linear (2)
Softmax
+
-
AD
CN

eXplainable AI and Feature Attribution Methods


NOT SURE IF MODEL IS BROKEN
OR EXPLANATION IS WRONG
imgflip.com

Perturbation Tests: Insertion and Deletion

Attribution Maps

Relevance per ROI

results

Conclusion

Take-Aways

1. Perturbation tests offer a model-agnostic fidelity metric.
2. The baseline should be chosen carefully w.r.t. to context.
3. Attribution Maps need interpretation to actually explain anything.