Tom Nuno Wolf 1,2,3, Fabian Bongratz 1,2,3, Anne-Marie Rickmann 1,2, Sebastian Pölsterl<sup>2</sup>, Christian Wachinger<sup>1,2,3</sup>

Department of Radiology, Technical University of Munich, Munich, Germany
 Lab for AI in Medical Imaging, Ludwig-Maximilians-University, Munich, Germany
 Munich Center for Machine Learning (MCML)



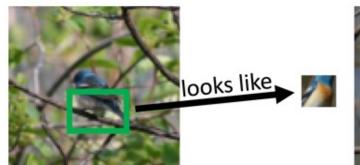




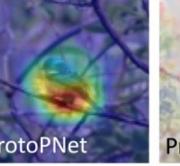
## **Keep The Faith:**

### Faithful Explanations in Convolutional Neural Networks for Case-Based Reasoning

#### **Case-Based Reasoning:**

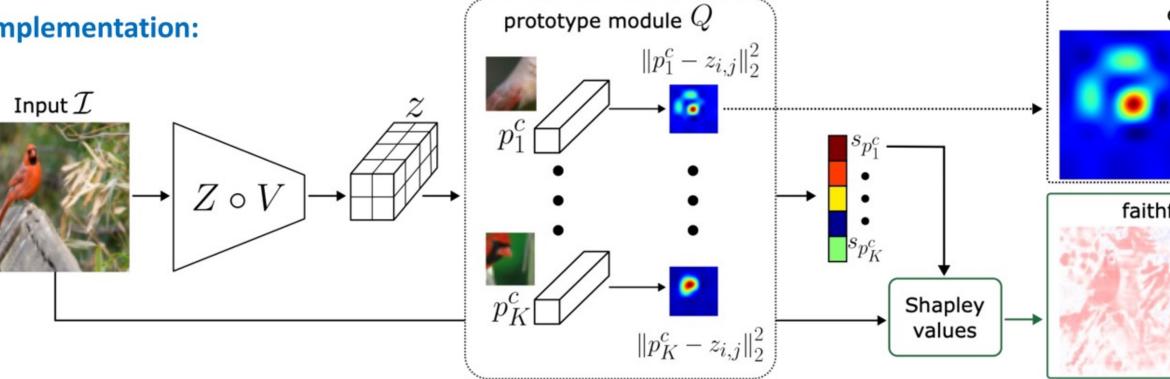


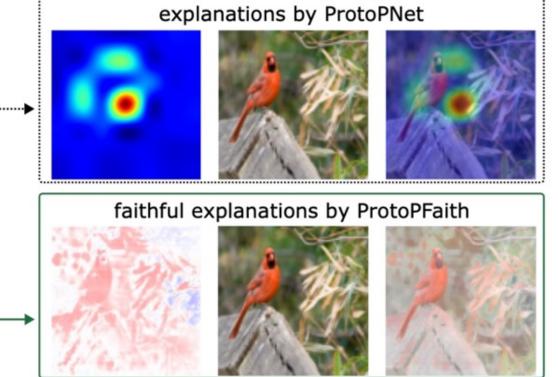






#### Implementation:

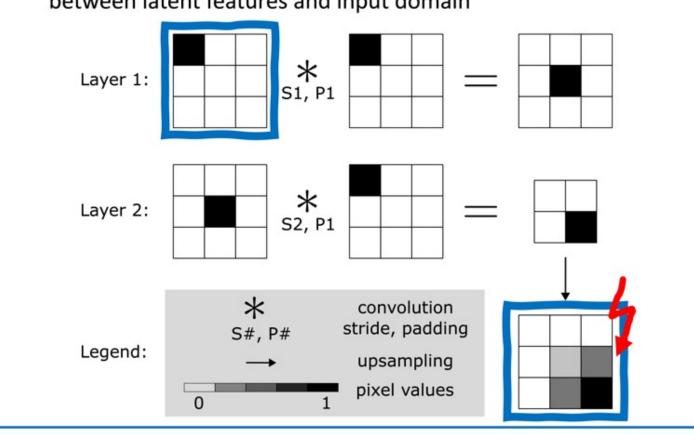




# XAI unveiled

#### **Motivation:**

 Assumption in ProtoPNet-like [1] architectures: spatial dependency between latent features and input domain



#### Method:

- Convert trained ProtoPNet into Lightweight Probabilistic Neural Network [2]
- Extract explanations following DASP [3] over similarity scores s
- Explanations are based on Shapley values, which satisfy all axioms that we define to be required for faithfulness
- Extraction of vanilla explanations still possible for the same model

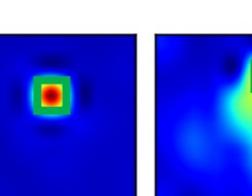
#### **Requirements:**

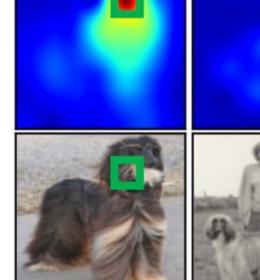
 Closed-form solution for propagation of normal distributions through all layers

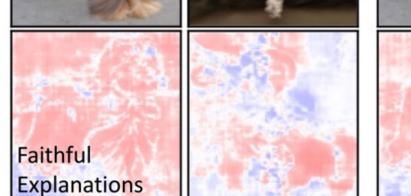
[1] C. Chen, O. Li, D. Tao, et al.: "This Looks Like That: Deep Learning for Interpretable Image Recognition", NeurIPS 2019 [2] J. Gast, S. Roth: "Lightweight probabilistic deep networks", CVPR 2018 [3] M. Ancona, C. Oztireli, M. Gross: "Explaining deep neural networks with a polynomial time algorithm for shapley value approximation", ICML 2019 [4] M. Nauta, R. Van Bree, C. Seifert: "Neural prototype trees for interpretable fine-grained image recognition", CVPR 2021
[5] E. Kim, S. Kim, M. Seo, S. Yoon: "XProtoNet: Diagnosis in Chest Radiography With Global and Local Explanations", CVPR 2021

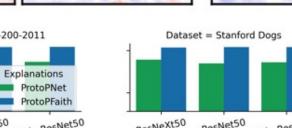
## **Results:**

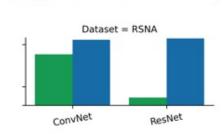
**Explanations** 











Bounding box of vanilla

prototype explanations

#### **Key Takeaways:**

- Theoretical violations manifest in experimental results
- Findings generalize to other implementations of case-based
- reasoning, e.g. ProtoTrees [4] and XProtoNet [5]
- Faithful explanations are difficult to interpret







