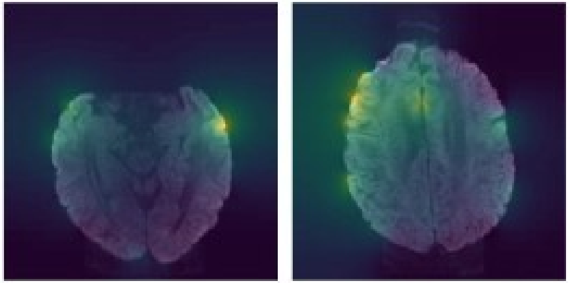
Exploring the Explanation Landscape

Overview

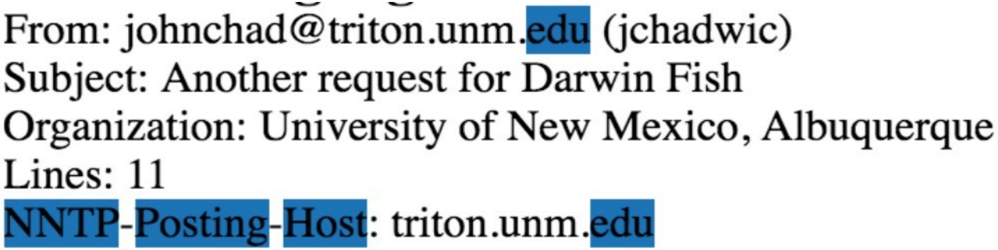
- Reliability pillars
- OpenXAI
- Is OpenXAI all you need?
- New directions

Natural images Fong et al. 2017



MRI brain scans Agarwal et al. 2021

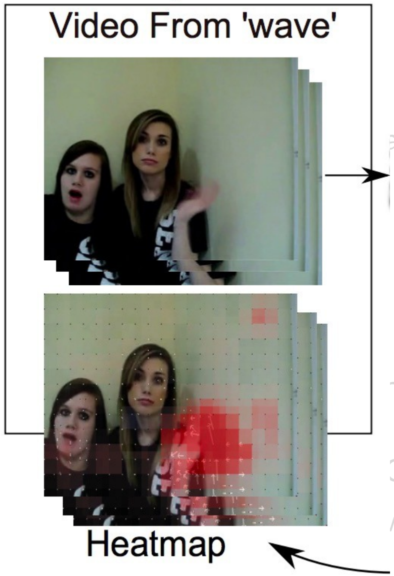
er et al. 2014
ngenberg et al. 2015
vorks. *Sundararajan*
ion. *Zhou et al. 2016*
of any classifier. *Ribe*
Smootl ... milkov et al. 2017
MP: Inte ... l Perturbation. Fong et al. 2017
SHAP: A Unified Approach to Interpreting Model Predictions. *Lundberg et al. 2017*
PDA: Visualizing deep neural network decisions: Prediction difference analysis. Zintgraf et al. 2017


From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

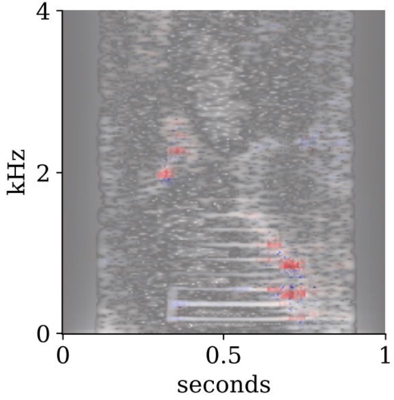
Text Ribeiro et al. 2016

FIDO: Explaini ... ual generation. Chang et al. 201
Expected Grad... s Using Attribution Priors. Erion
FG-Vis: Interpretable and Fine-Grained Visual Explanations for Co *2019*
Understanding Deep Networks via Extren
MP-G: Removing input features via a generative mo

...



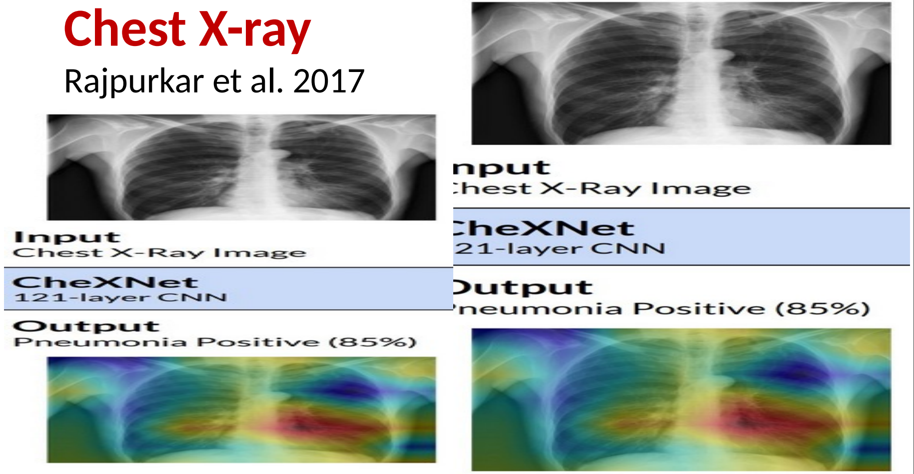
Videos Srinivasan et al. 2017



Audio Becker et al. 2019

Chest X-ray

Rajpurkar et al. 2017



t al. 2014

nberg et al. 2015

ks. Sundararajan et al. 2018

. Zhou et al. 2016

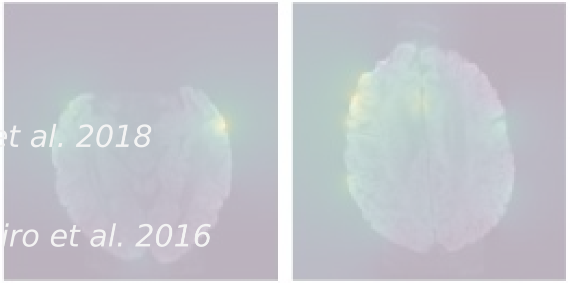
any classifier. Ribeiro et al. 2016

. 2017

. Fong et al. 2017

Lundberg et al. 2017

ysis. Zintgraf et al. 2017



MRI brain scans

Agarwal et al. 2021

4

Video From 'wave'

FIDO: Explaini

Text

Ribeiro et al. 201

Expected Gradients: Learning Explainable Mo

FG-Vis: Interpretable and Fine-Graine
2019

Understanding Deep Networks via Ex

MP-G: Removing input features via a generati

Detecting biases

Lapuschkin et al. 2016

How do we evaluate the **reliability** of state-of-the-art explanation methods?

Reliability Pillars

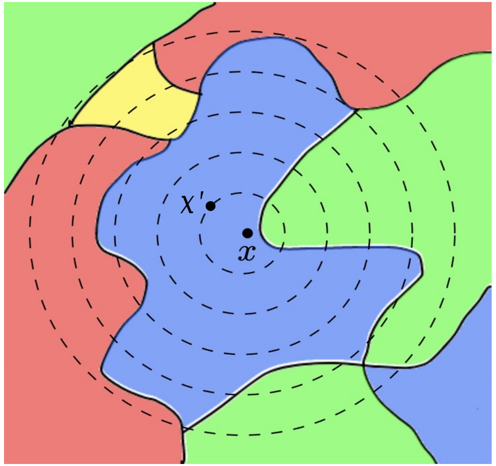
C. Agarwal, M. Zitnik, H. Lakkaraju, Probing GNN Explainers: A Rigorous Theoretical and Empirical Analysis of GNN Explanation Methods, AISTAT

Pillar 1: Faithfulness

Prediction model
to be explained

Masking function

Pillar 2: Stability



C. Agarwal et al., Rethinking Stability for Attribution-based Explanations, Oral presentation @ ICLR 2022 PAIR^2Struct workshop

Pillar 3: Counterfactual Fairness

-

J. Dai et al., Fairness via Explanation Quality: Evaluating Disparities in the Quality of Post hoc Explanations, AIES'2022

How do we **pick** an explanation method from the XAI landscape?

OpenXAI

- OpenXAI provides an automated end-to-end pipeline that simplifies and standardizes the evaluation of post hoc explanation methods

- OpenXAI promotes transparency and reproducibility in benchmarking explanation methods

https://github.com/AI4LIFE-GROUP/OpenXAI

C. Agarwal et al., OpenXAI: Towards a Transparent Evaluation of Model Explanations, NeurIPS Datasets and Benchmark Track'2022

OpenXAI's Key Components

- A flexible synthetic data generator and a collection of diverse 7 real-world datasets, 16 pre-trained models, and 6 state-of-the-art explanation methods

- Open-source implementations of 22 quantitative metrics for evaluating faithfulness, stability (robustness), and fairness of explanation methods

- First-ever public XAI leaderboards to benchmark explanation methods

XAI ready Dataloaders and Models

```
from openxai import Dataloader
loader_train, loader_test = Dataloader.return_loaders(data_name='german',
download=True)
inputs, labels = iter(loader_test).next()
```

XAI ready Dataloaders and Models

```
from openxai import Dataloader
loader_train, loader_test = Dataloader.return_loaders(data_name='german',
download=True)
inputs, labels = iter(loader_test).next()
```

OpenXAI provides pre-trained models for readily benchmarking explanation methods.

```
from openxai import LoadModel
model = LoadModel(data_name='german', ml_model='ann')
```

OpenXAI Explainers

- OpenXAI provides ready-to-use implementations of six state-of-the-art feature attribution methods

```
from openxai import Explainer
exp_method = Explainer(method='LIME')
explanations = exp_method.get_explanations(model, X=inputs, y=labels)
```

OpenXAI Explainers

```
@abstractmethod
def get_explanations(self, model, X: torch.Tensor, y: torch.Tensor):
    """
    Generate explanations for given input/s.
    Parameters
    ----------
    model: pre-trained ML model
    X: torch.tensor
        Input in two-dimensional shape (m, n).
    y: torch.tensor
        Labels
    Returns
    ----------
    torch.Tensor
        Explanation vector/matrix.
    """
    pass
```

OpenXAI's Evaluation

- OpenXAI provides implementations and ready-to-use APIs for a set of 22 quantitative metrics proposed by prior research to evaluate the faithfulness, stability, and fairness of explanation methods

```
from openxai import Evaluator
metric_evaluator = Evaluator(inputs, labels, model, explanations)
score = metric_evaluator.eval(metric='RIS')
```

OpenXAI's Leaderboard

Explore Leaderboards

German Credit

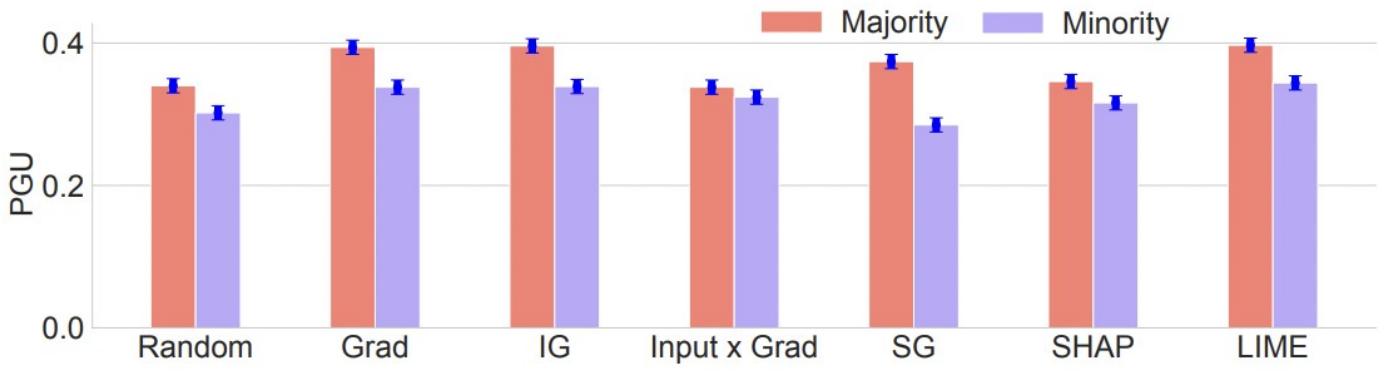
Faithfulness

Method	FA ↑	RA	SA	SRA	RC	PRA	PGI	PGU
Vanilla Gradient	0.950	0.950	0.846	0.846	1.000	1.000	0.149	0.174
SmoothGrad	0.950	0.950	0.606	0.606	1.000	1.000	0.202	0.124
Integrated Gradient	0.950	0.950	0.846	0.846	1.000	1.000	0.148	0.173
LIME	0.938	0.810	0.938	0.814	0.998	0.989	0.156	0.169
Gradient x Input	0.785	0.161	0.382	0.071	0.890	0.875	0.171	0.161
SHAP	0.130	0.007	0.112	0.006	-0.053	0.488	0.135	0.180

Exploring the landscape using OpenXAI

- LIME produces more faithful (+24.9%) explanations

- Across all real-world datasets, SmoothGrad achieves 63.2% higher RRS values



Is OpenXAI all you need?

- How to benchmark different non-perturbation-based explanation methods?

- Benchmarking explanations on other modalities
  - Vision (Quantus)
  - NLP (e-ViL)
  - Graphs (GraphXAI)

C. Agarwal et al., Evaluating Explainability for Graph Neural Networks, Nature Scientific Data’2023
M. Kayser et al., e-ViL: A Dataset and Benchmark for Natural Language Explanations in Vision-Language Tasks, ICCV’2021

New directions

- Training models using Explanation Feedbacks
- Differentiable Explainable Curricula for RL Agents
- Learning Hierarchical and Multi-modal Explanations

Thank you!

- Email: chiragagarwall12@gmail.com
- Website: http://chirag126.github.io/

@_cagarwal

Questions?