On Spectral Properties of Gradient-based Explanation Methods

Amir Mehrpanah

Supervisors: Hossein Azizpour and Kevin Smith

ECCV 2024 — KTH Royal Institute of Technology

Today's Seminar Agenda

- What is explainability and why it matters?

Today’s Seminar Agenda

- What is explainability and why it matters?
- Research questions and literature around explainability.

Today’s Seminar Agenda

- What is explainability and why it matters?
- Research questions and literature around explainability.
- Our focus and approach in the field.

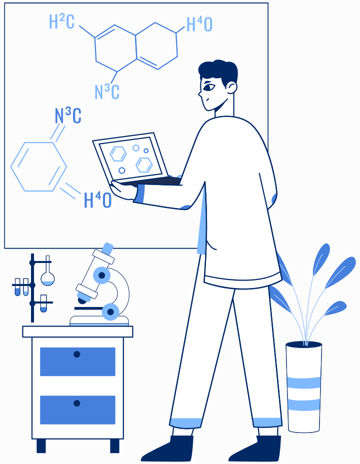
Today’s Seminar Agenda

- What is explainability and why it matters?
- Research questions and literature around explainability.
- Our focus and approach in the field.
- Touch upon our first step and future work.

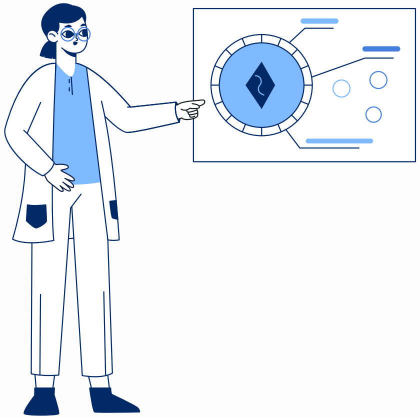
What is Explainability?

The Good Old Ways

Problem



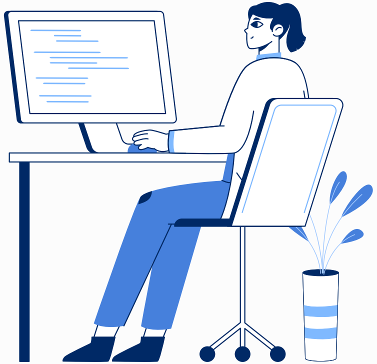
Years of Research



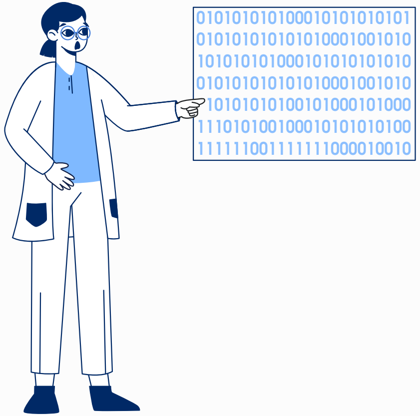
Understandable Outputs

The Machine Learning Era

Problem



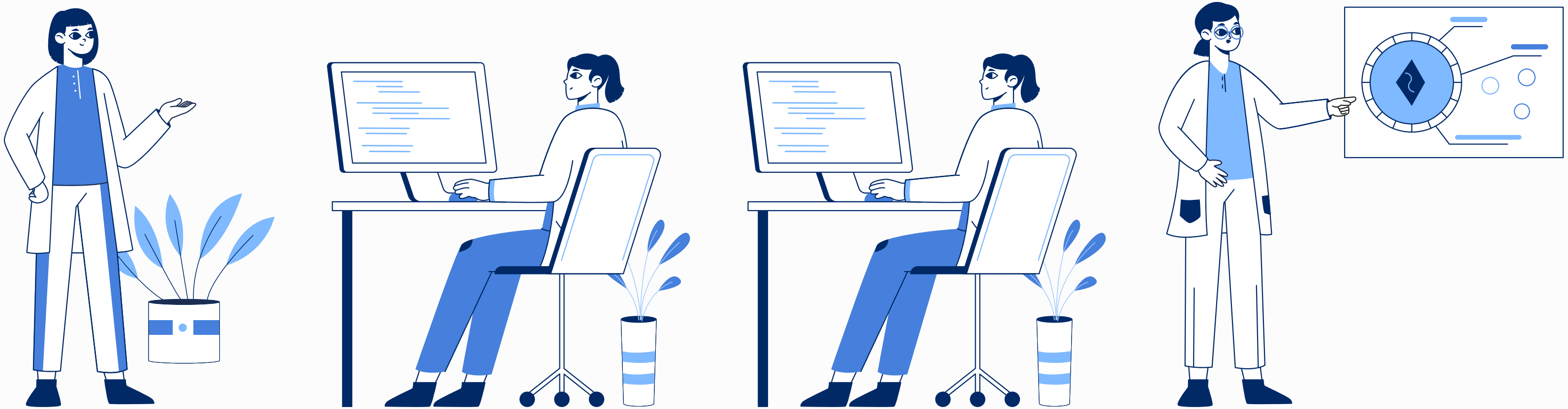
Machine Leaning



Black-box Outputs

Art By GreenSkyStudio

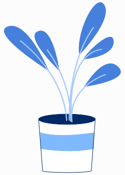
Explainable Machine Learning



Problem | Machine Leaning | Explainable ML | Explainable Outputs

What is an Explanation?

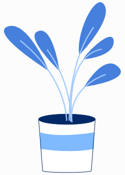
counterfactuals

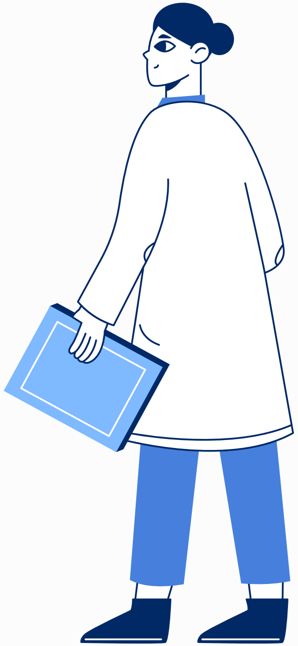


What is an Explanation?

counterfactuals

what if …?



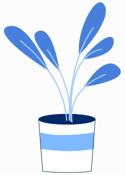


What is an Explanation?

counterfactuals

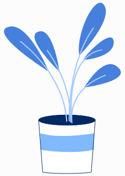
what if …?

hypothetical world



What is an Explanation?

what if time changes?

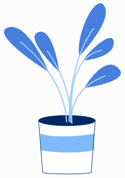


What is an Explanation?

what if time changes?

in a hypothetical world

the apple falls

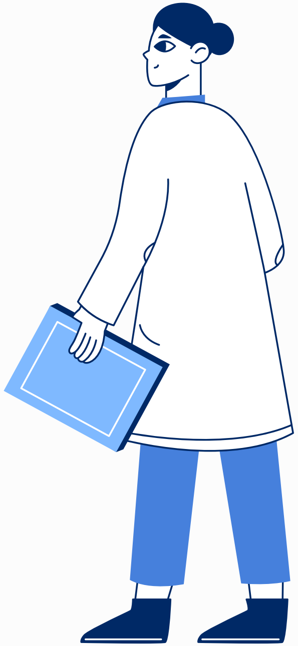


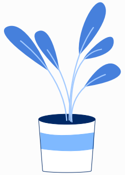
What is an Explanation?

what if time changes?

in a hypothetical linear world

$$y(t + \Delta t) = \Delta t \frac{\partial}{\partial t} y(t)$$



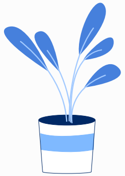


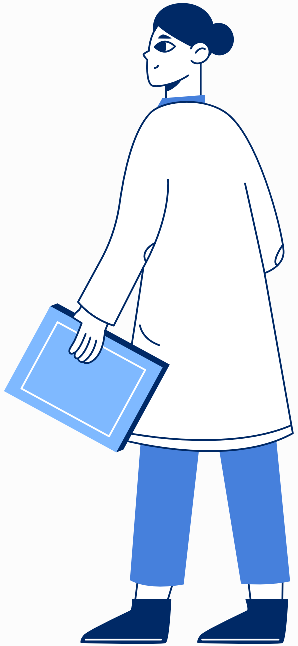
What is an Explanation?

what if input of a function changes?

in a hypothetical linear world

$$\nabla_x f(x)$$





Art By GreenSkyStudio

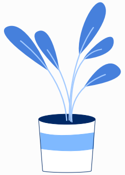
What is an Explanation?

explaining a function by its gradient

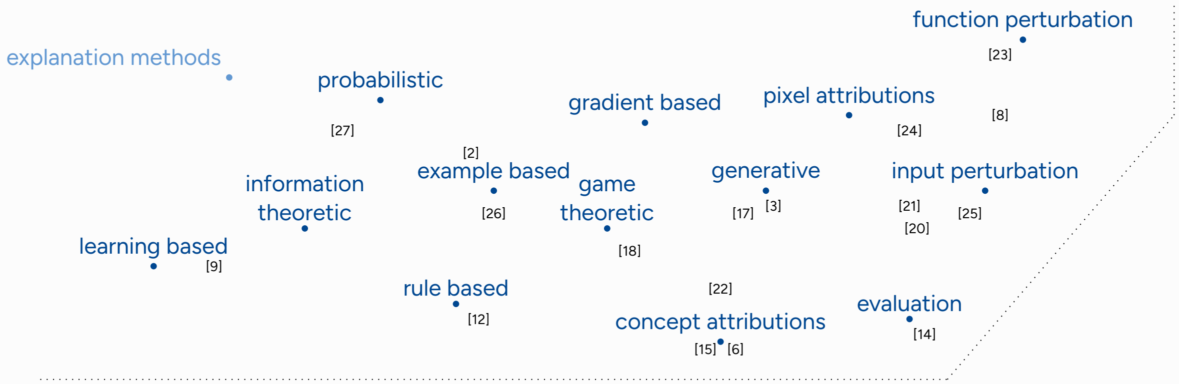
in a hypothetical linear world

by counterfactuals around x

$$\nabla_x f(x)$$

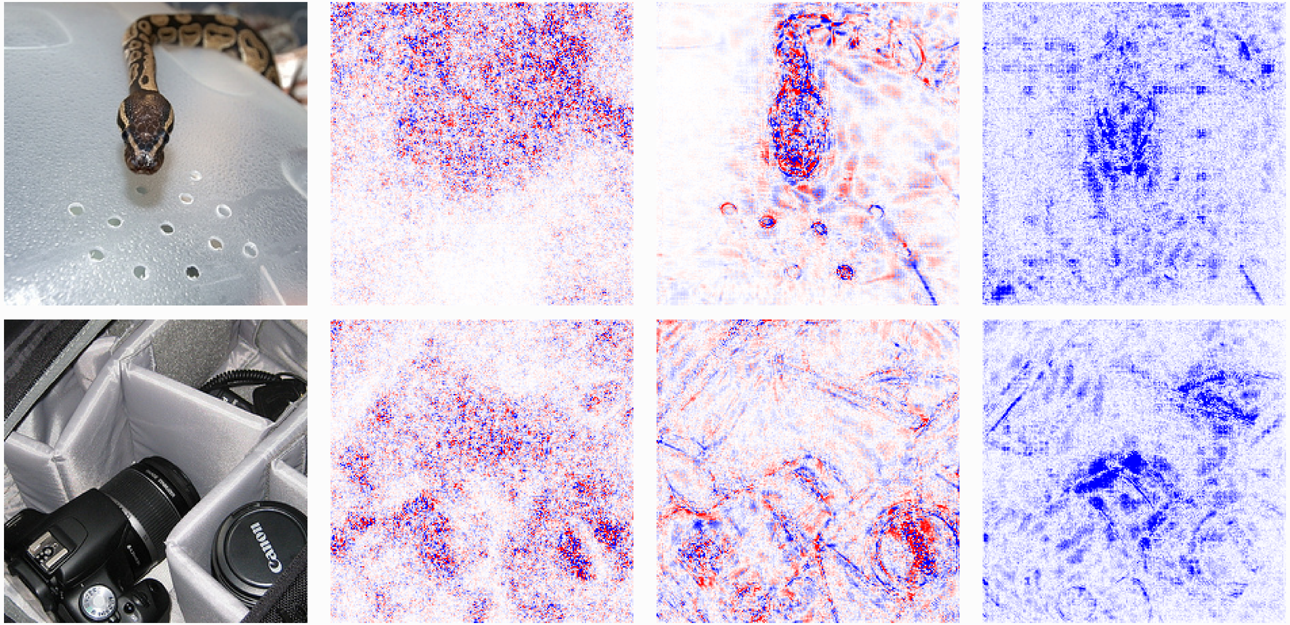


Developing Explainability Tools

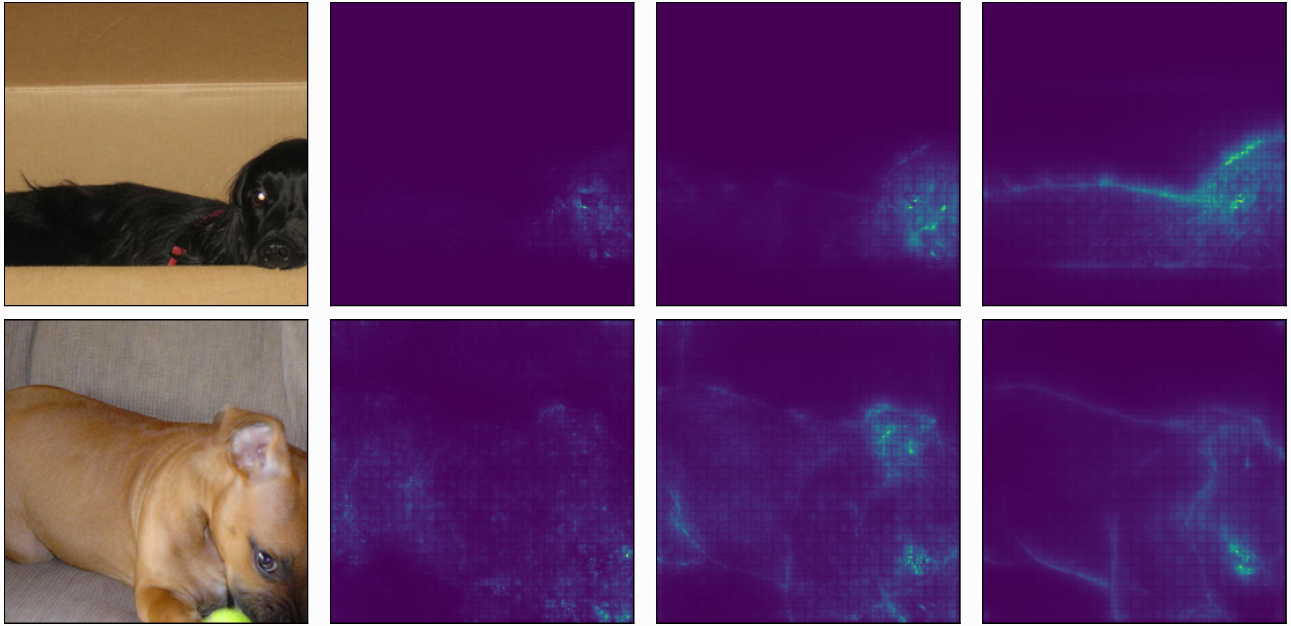


Unreliability of Explanation Methods

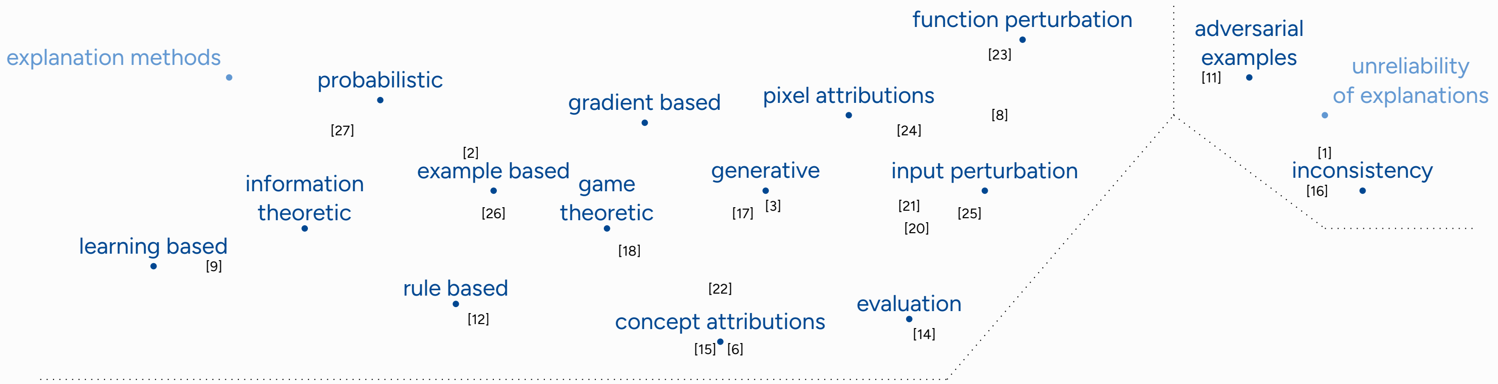
Unreliability of Explanation Methods



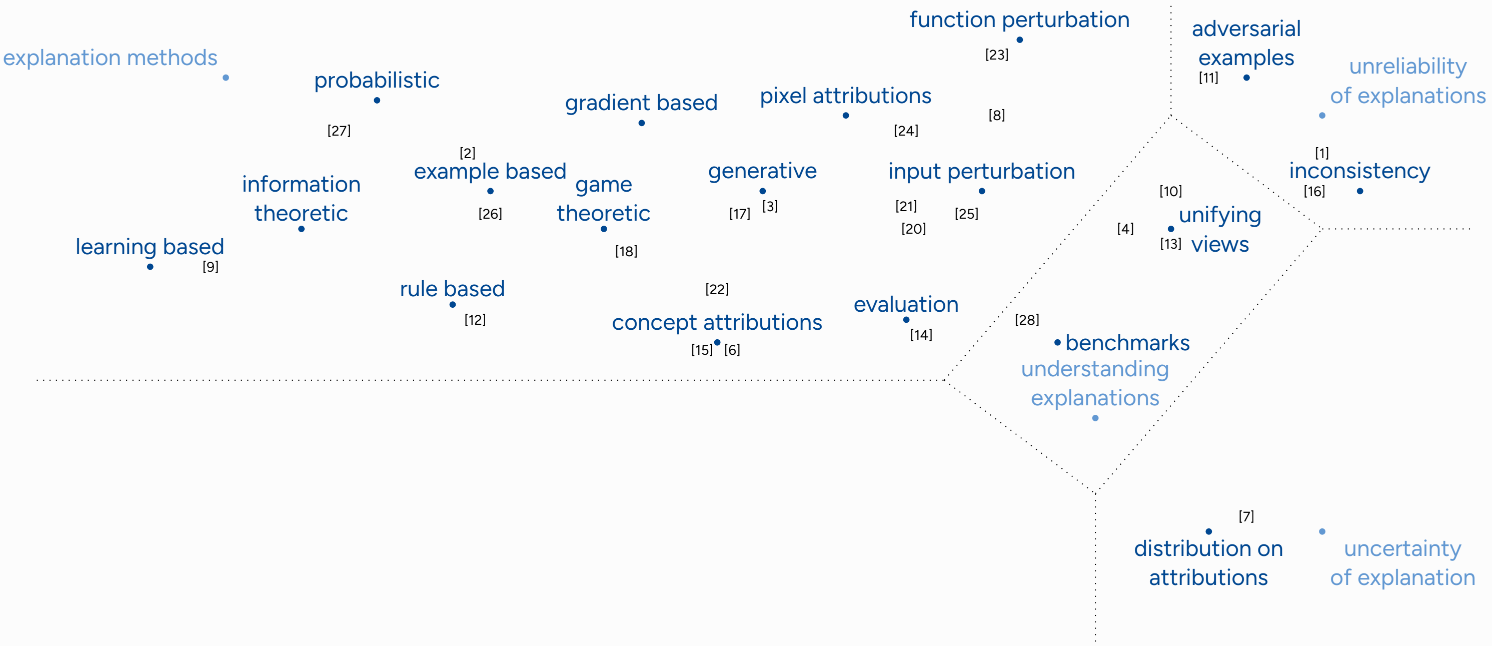
Unreliability of Explanation Methods



Flaws in Explainability Tools



Dealing with Flaws in Explainability



General Research Questions

- What is a definition for explanation?

General Research Questions

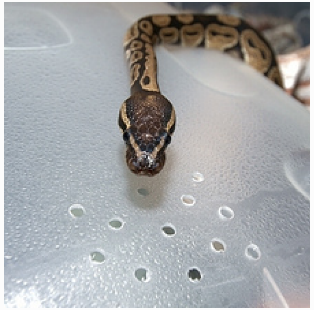
- What is a definition for explanation?
- How explanations can be evaluated?

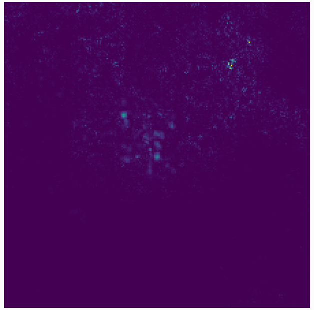
General Research Questions

- What is a definition for explanation?
- How explanations can be evaluated?
- How to make explanations less subjective?

General Research Questions

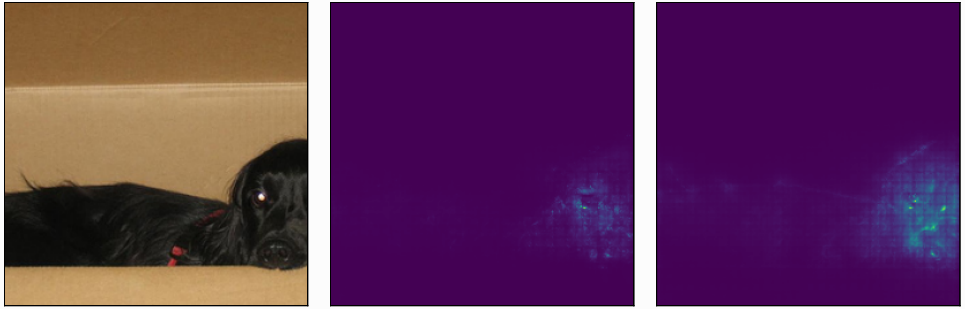
- What is a definition for explanation?
- How explanations can be evaluated?
- How to make explanations less subjective?
- Why the gradient of deep networks is sparse?





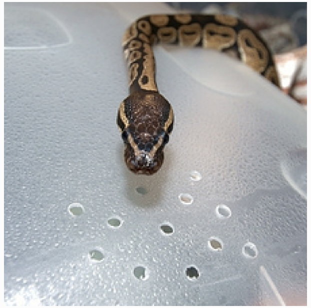
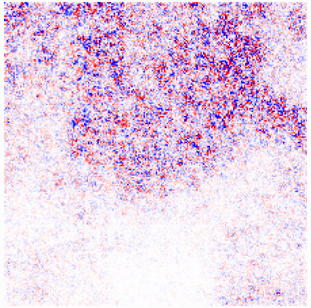
General Research Questions

- What is a definition for explanation?
- How explanations can be evaluated?
- How to make explanations less subjective?
- Why the gradient of deep networks is sparse?
- What causes the inconsistencies in explanations?

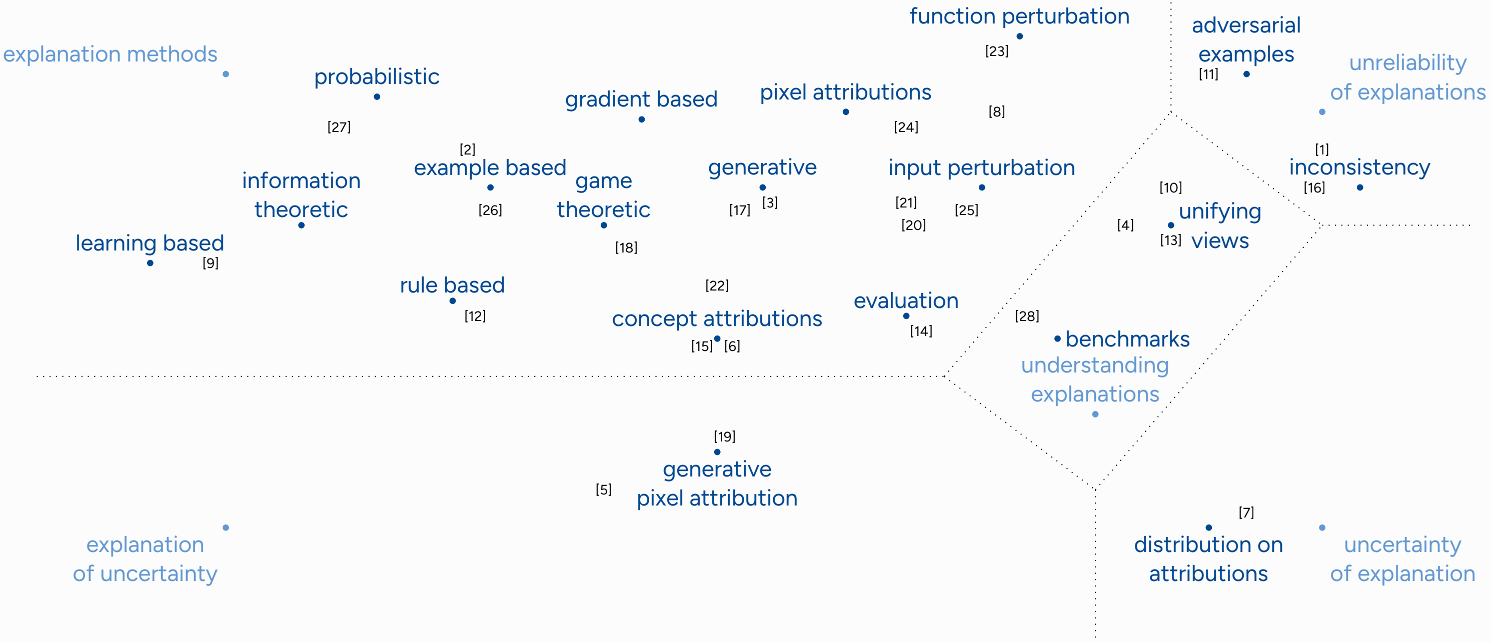


General Research Questions

- What is a definition for explanation?
- How explanations can be evaluated?
- How to make explanations less subjective?
- Why the gradient of deep networks is sparse?
- What causes the inconsistencies in explanations?
- Why the gradient sign changes with small noise?

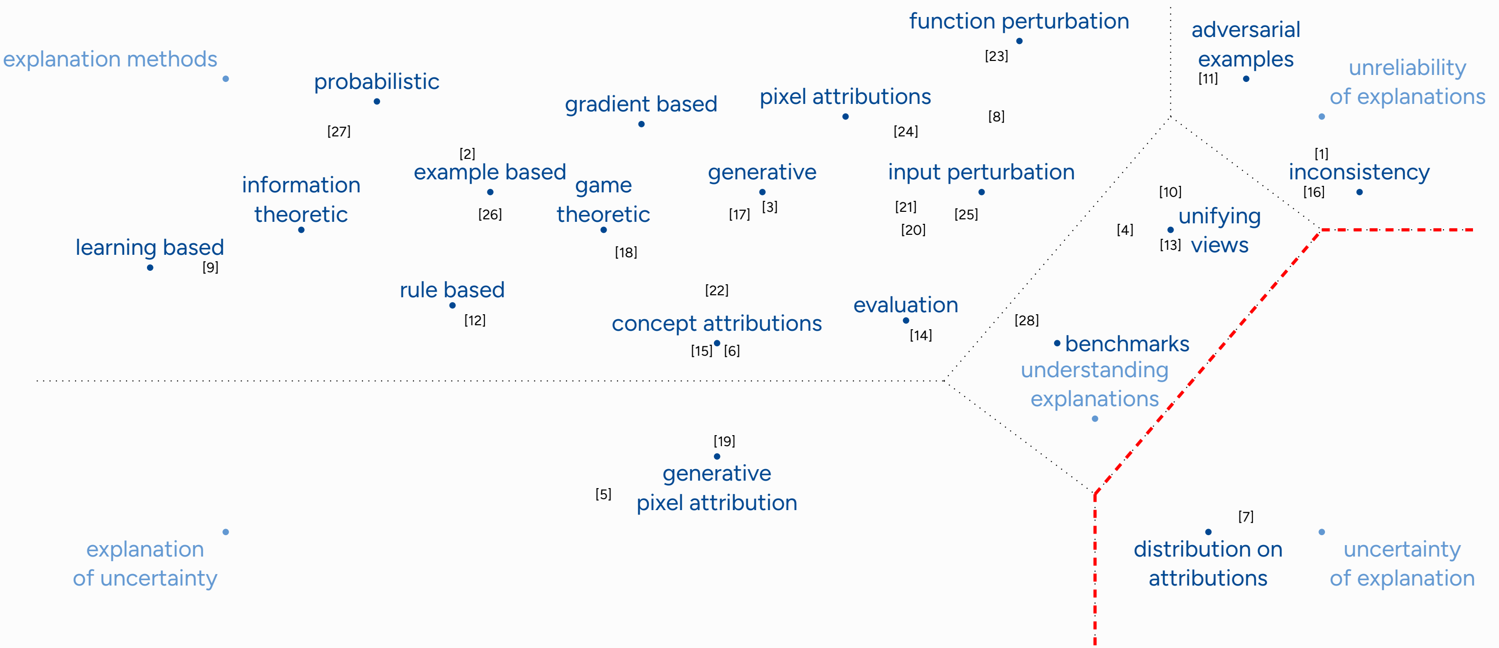



Repurposing Explainability Tools



Our Focus

Our Focus in the Literature



Our Research Questions

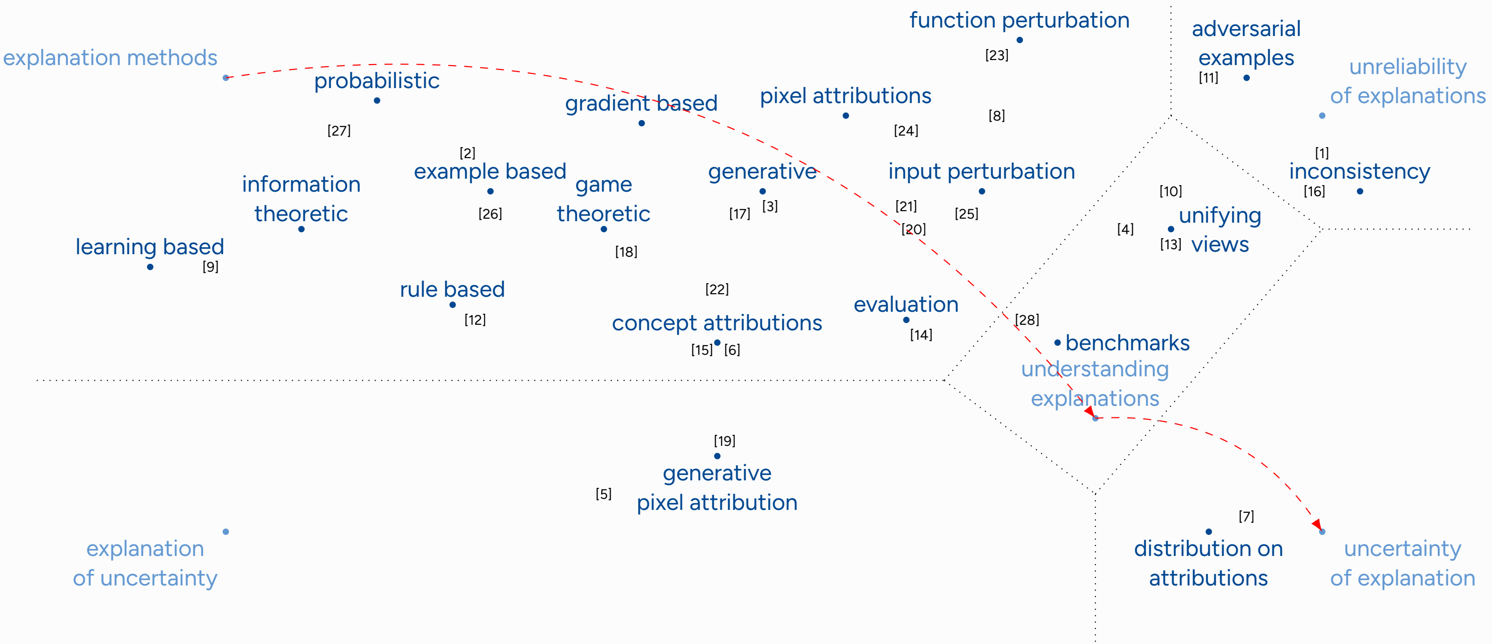
- What are the sources of uncertainty in explanations?

Our Research Questions

- What are the sources of uncertainty in explanations?
- How to quantify our uncertainty in explanation methods?

Our Focus

A High-level Trajectory is Worth 16x16 Papers



The First Step

Our Approach

- A probabilistic representation for explanations.

$$E = \nabla f(X) \quad s.t. \quad X \sim P$$

Our Approach

- A probabilistic representation for explanations.
- A spectral representation for explanations.

$$\mathcal{F}\{E\} = \mathcal{F}\{\nabla f(X)\} \quad s.t. \quad X \sim P$$

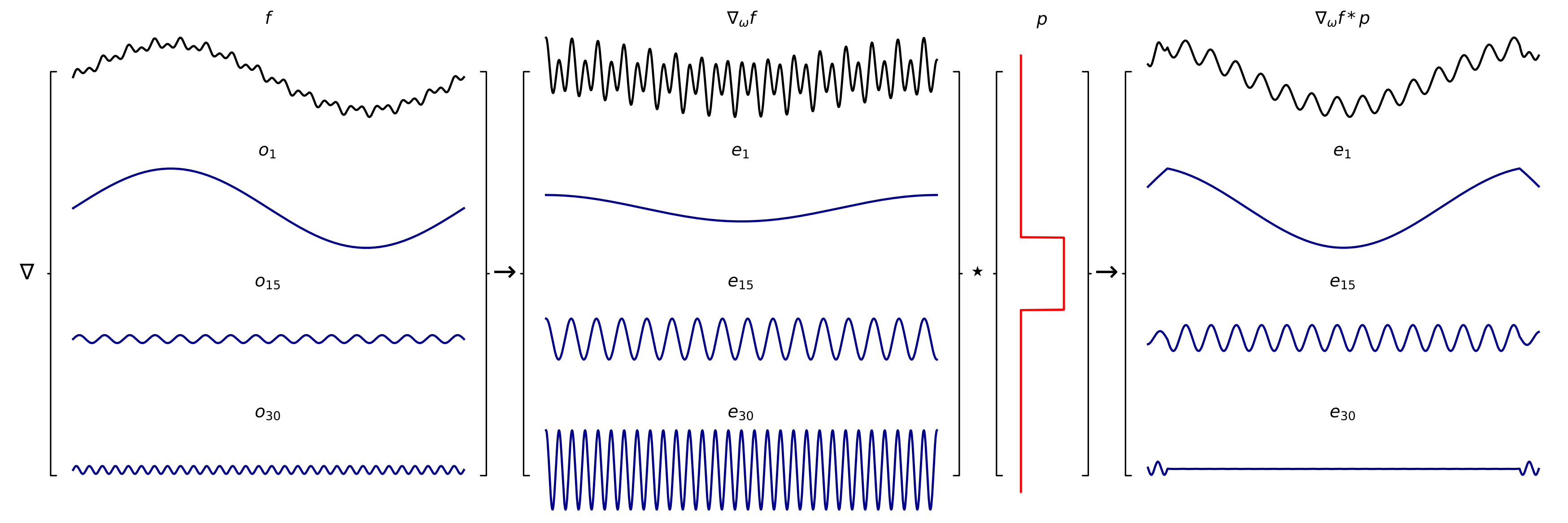
Our Findings

- Gradient operator amplifies the attribution of high-frequency features.

Our Findings

- Gradient operator amplifies the attribution of high-frequency features.
- Perturbation mitigates the attribution of high-frequency features.

Conceptual Visualization



Our Findings

- Gradient operator amplifies the attribution of high-frequency features.
- Perturbation mitigates the attribution of high-frequency features.
- Sign of the gradient depends on the chosen perturbation.

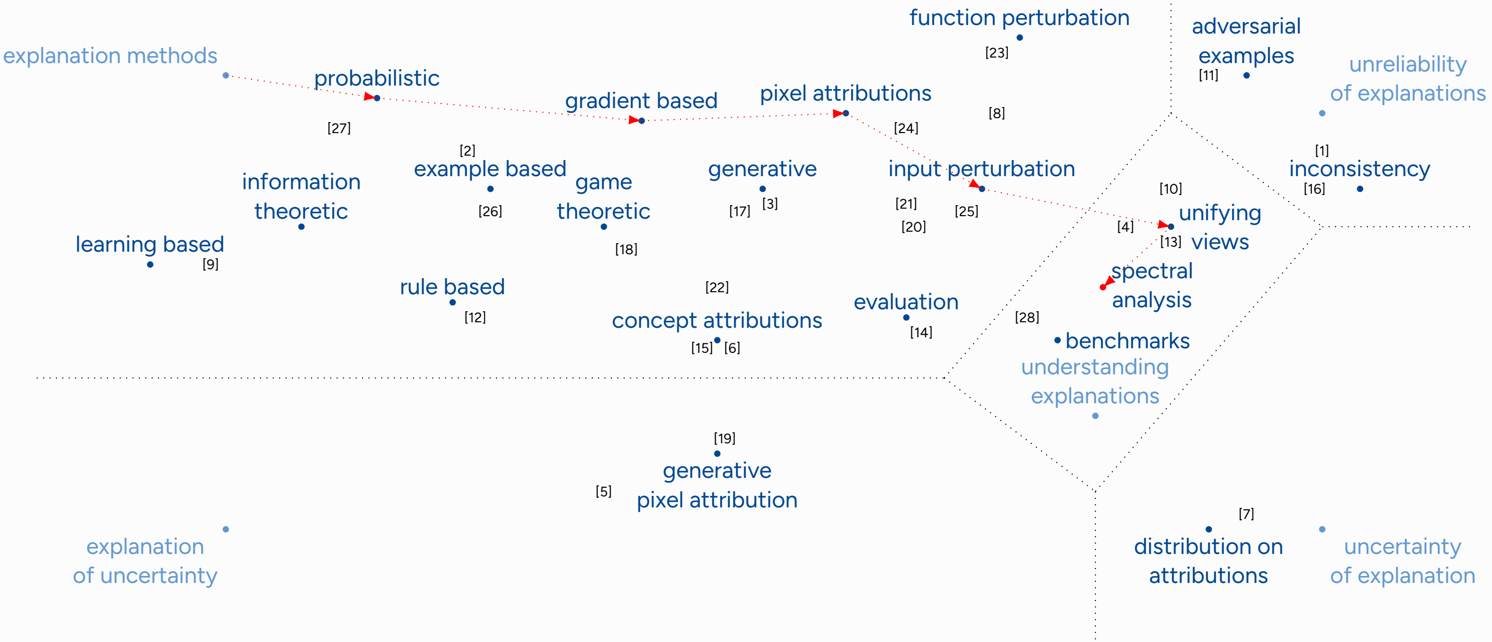
Our Findings

- Gradient operator amplifies the attribution of high-frequency features.
- Perturbation mitigates the attribution of high-frequency features.
- Sign of the gradient depends on the chosen perturbation.
- Gradient squared is a better design choice compared to gradient.

Our Findings

- Gradient operator amplifies the attribution of high-frequency features.
- Perturbation mitigates the attribution of high-frequency features.
- Sign of the gradient depends on the chosen perturbation.
- Gradient squared is a better design choice compared to gradient.
- A justification for the inconsistencies in the explanations.

Our Work in the Literature

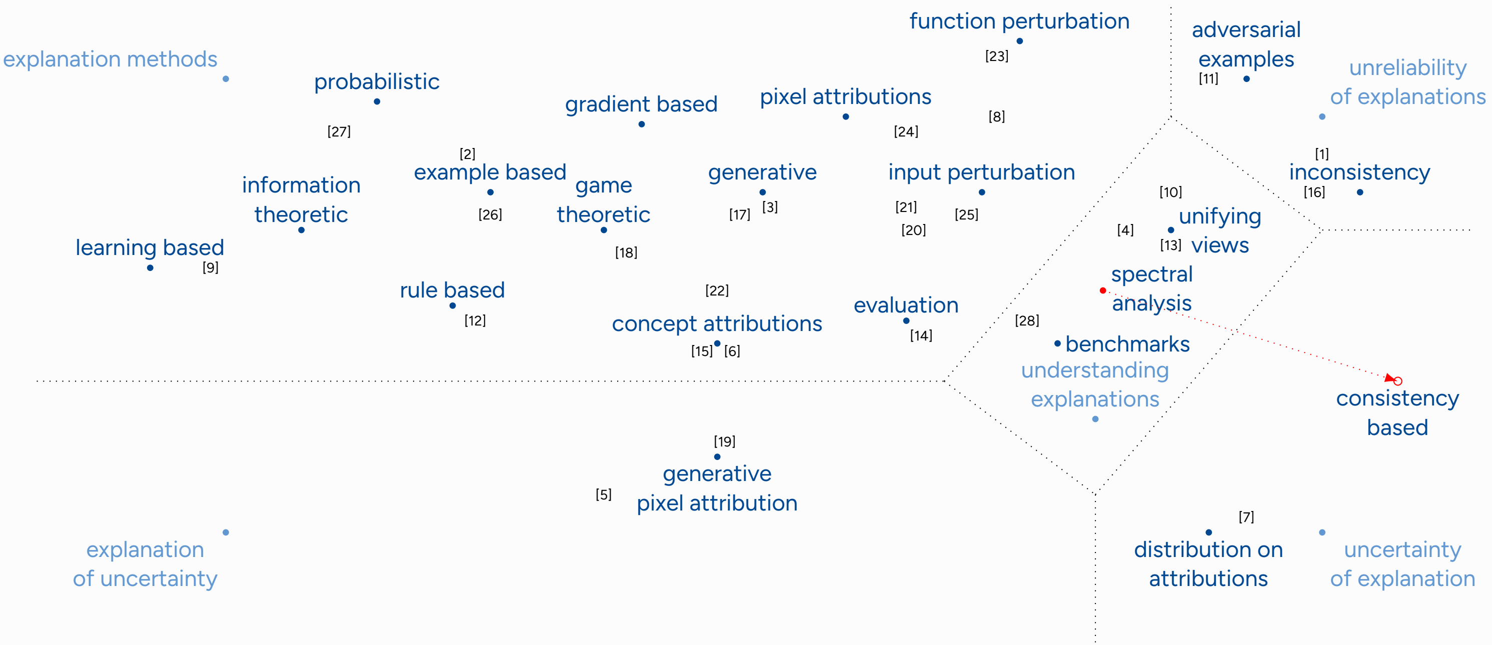


The Next Step

What is Next?

- Relate consistency of explanations to their uncertainty.

Our Next Work in the Literature



KTH

VETENSKAP
OCH KONST

Referenced Papers I

[1] Julius Adebayo et al. **Sanity Checks for Saliency Maps**. arXiv:1810.03292 [cs, stat]. Nov. 2020. DOI: 10.48550/arXiv.1810.03292. URL: http://arxiv.org/abs/1810.03292 (visited on 01/30/2023).

[2] Ajaya Adhikari et al. "LEAFAGE: Example-based and Feature importance-based Explanations for Black-box ML models". In: **2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)**. ISSN: 1558-4739. June 2019, pp. 1–7. DOI: 10.1109/FUZZ-IEEE.2019.8858846. URL: https://ieeexplore.ieee.org/abstract/document/8858846 (visited on 04/03/2024).

[3] Chirag Agarwal and Anh Nguyen. **Explaining image classifiers by removing input features using generative models**. arXiv:1910.04256 [cs, stat]. Oct. 2020. DOI: 10.48550/arXiv.1910.04256. URL: http://arxiv.org/abs/1910.04256 (visited on 04/16/2024).

[4] Sushant Agarwal et al. **Towards the Unification and Robustness of Perturbation and Gradient Based Explanations**. arXiv:2102.10618 [cs]. July 2021. DOI: 10.48550/arXiv.2102.10618. URL: http://arxiv.org/abs/2102.10618 (visited on 01/30/2023).

[5] Javier Antorán et al. **Getting a CLUE: A Method for Explaining Uncertainty Estimates**. arXiv:2006.06848 [cs, stat]. Mar. 2021. DOI: 10.48550/arXiv.2006.06848. URL: http://arxiv.org/abs/2006.06848 (visited on 01/14/2023).

[6] Andrew Bai et al. **Concept Gradient: Concept-based Interpretation Without Linear Assumption**. arXiv:2208.14966 [cs]. Aug. 2022. DOI: 10.48550/arXiv.2208.14966. URL: http://arxiv.org/abs/2208.14966 (visited on 02/02/2023).

[7] Kirill Bykov et al. **How Much Can I Trust You? – Quantifying Uncertainties in Explaining Neural Networks**. arXiv:2006.09000 [cs, stat]. June 2020. DOI: 10.48550/arXiv.2006.09000. URL: http://arxiv.org/abs/2006.09000 (visited on 02/13/2023).

[8] Kirill Bykov et al. **NoiseGrad: Enhancing Explanations by Introducing Stochasticity to Model Weights**. arXiv:2106.10185 [cs]. May 2022. DOI: 10.48550/arXiv.2106.10185. URL: http://arxiv.org/abs/2106.10185 (visited on 04/27/2023).

Referenced Papers II

[9] Jianbo Chen et al. **Learning to Explain: An Information-Theoretic Perspective on Model Interpretation**. arXiv:1802.07814 [cs, stat]. June 2018. DOI: 10.48550/arXiv.1802.07814. URL: http://arxiv.org/abs/1802.07814 (visited on 12/05/2023).

[10] Ian Covert, Scott Lundberg, and Su-In Lee. **Explaining by Removing: A Unified Framework for Model Explanation**. arXiv:2011.14878 [cs, stat]. May 2022. DOI: 10.48550/arXiv.2011.14878. URL: http://arxiv.org/abs/2011.14878 (visited on 05/23/2023).

[11] Amirata Ghorbani, Abubakar Abid, and James Zou. **Interpretation of Neural Networks is Fragile**. arXiv:1710.10547 [cs, stat]. Nov. 2018. DOI: 10.48550/arXiv.1710.10547. URL: http://arxiv.org/abs/1710.10547 (visited on 01/30/2023).

[12] Riccardo Guidotti et al. **Local Rule-Based Explanations of Black Box Decision Systems**. arXiv:1805.10820 [cs]. May 2018. DOI: 10.48550/arXiv.1805.10820. URL: http://arxiv.org/abs/1805.10820 (visited on 04/03/2024).

[13] Tessa Han, Suraj Srinivas, and Himabindu Lakkaraju. **Which Explanation Should I Choose? A Function Approximation Perspective to Characterizing Post Hoc Explanations**. arXiv:2206.01254 [cs]. Dec. 2022. DOI: 10.48550/arXiv.2206.01254. URL: http://arxiv.org/abs/2206.01254 (visited on 04/23/2024).

[14] Sara Hooker et al. **A Benchmark for Interpretability Methods in Deep Neural Networks**. arXiv:1806.10758 [cs, stat]. Nov. 2019. DOI: 10.48550/arXiv.1806.10758. URL: http://arxiv.org/abs/1806.10758 (visited on 01/12/2024).

[15] Been Kim et al. **Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)**. arXiv:1711.11279 [stat]. June 2018. DOI: 10.48550/arXiv.1711.11279. URL: http://arxiv.org/abs/1711.11279 (visited on 02/01/2023).

Referenced Papers III

[16] Pieter-Jan Kindermans et al. **The (Un)reliability of saliency methods**. arXiv:1711.00867 [cs, stat]. Nov. 2017. DOI: `10.48550/arXiv.1711.00867`. URL: `http://arxiv.org/abs/1711.00867` (visited on 02/14/2023).

[17] Shusen Liu et al. **Generative Counterfactual Introspection for Explainable Deep Learning**. arXiv:1907.03077 [cs, stat]. July 2019. DOI: `10.48550/arXiv.1907.03077`. URL: `http://arxiv.org/abs/1907.03077` (visited on 04/16/2024).

[18] Scott M. Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: **Advances in neural information processing systems** 30 (2017).

[19] Iker Perez et al. **Attribution of Predictive Uncertainties in Classification Models**. arXiv:2107.08756 [cs, stat]. June 2022. DOI: `10.48550/arXiv.2107.08756`. URL: `http://arxiv.org/abs/2107.08756` (visited on 01/18/2023).

[20] Vitali Petsiuk, Abir Das, and Kate Saenko. **RISE: Randomized Input Sampling for Explanation of Black-box Models**. arXiv:1806.07421 [cs]. Sept. 2018. DOI: `10.48550/arXiv.1806.07421`. URL: `http://arxiv.org/abs/1806.07421` (visited on 01/18/2023).

[21] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. **"Why Should I Trust You?": Explaining the Predictions of Any Classifier**. arXiv:1602.04938 [cs, stat]. Aug. 2016. DOI: `10.48550/arXiv.1602.04938`. URL: `http://arxiv.org/abs/1602.04938` (visited on 12/15/2022).

[22] Jessica Schrouff et al. **Best of both worlds: local and global explanations with human-understandable concepts**. arXiv:2106.08641 [cs]. Jan. 2022. DOI: `10.48550/arXiv.2106.08641`. URL: `http://arxiv.org/abs/2106.08641` (visited on 02/03/2023).

Referenced Papers IV

[23] Ramprasaath R. Selvaraju et al. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization”. In: **International Journal of Computer Vision** 128.2 (Feb. 2020). arXiv:1610.02391 [cs], pp. 336–359. ISSN: 0920-5691, 1573-1405. DOI: `10.1007/s11263-019-01228-7`. URL: `http://arxiv.org/abs/1610.02391` (visited on 11/09/2023).

[24] Daniel Smilkov et al. **SmoothGrad: removing noise by adding noise**. arXiv:1706.03825 [cs, stat]. June 2017. DOI: `10.48550/arXiv.1706.03825`. URL: `http://arxiv.org/abs/1706.03825` (visited on 04/27/2023).

[25] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. **Axiomatic Attribution for Deep Networks**. arXiv:1703.01365 [cs]. June 2017. DOI: `10.48550/arXiv.1703.01365`. URL: `http://arxiv.org/abs/1703.01365` (visited on 04/27/2023).

[26] **The effects of example-based explanations in a machine learning interface | Proceedings of the 24th International Conference on Intelligent User Interfaces**. URL: `https://dl.acm.org/doi/abs/10.1145/3301275.3302289` (visited on 04/03/2024).

[27] Eric Wang, Pasha Khosravi, and Guy Van den Broeck. **Probabilistic Sufficient Explanations**. arXiv:2105.10118 [cs]. May 2021. DOI: `10.48550/arXiv.2105.10118`. URL: `http://arxiv.org/abs/2105.10118` (visited on 01/29/2023).

[28] Mengjiao Yang and Been Kim. **Benchmarking Attribution Methods with Relative Feature Importance**. arXiv:1907.09701 [cs, stat]. Nov. 2019. DOI: `10.48550/arXiv.1907.09701`. URL: `http://arxiv.org/abs/1907.09701` (visited on 06/10/2023).