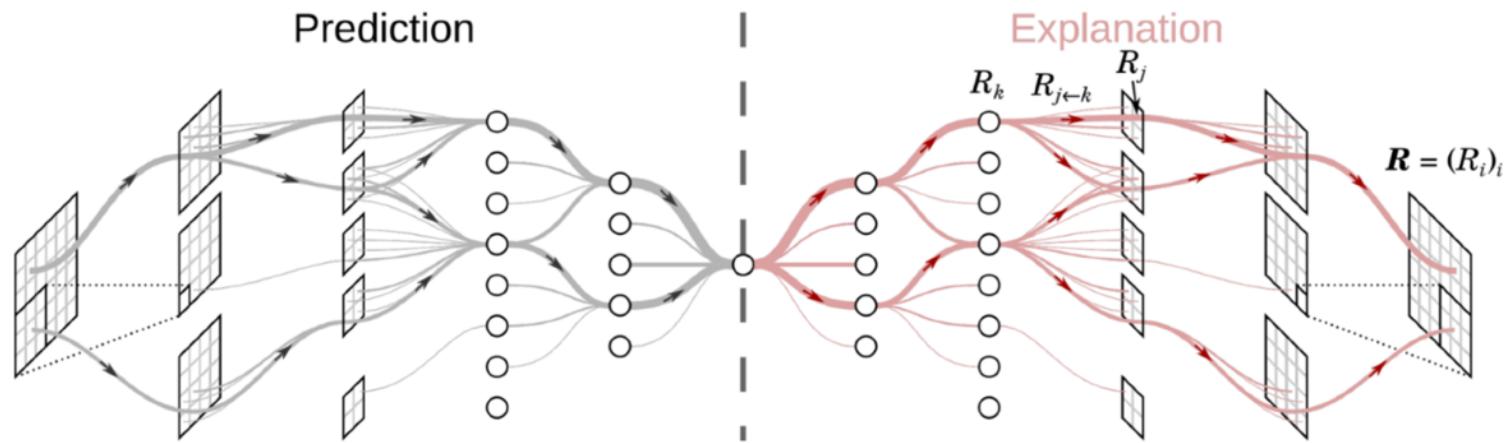


Concept-Level Explainable AI

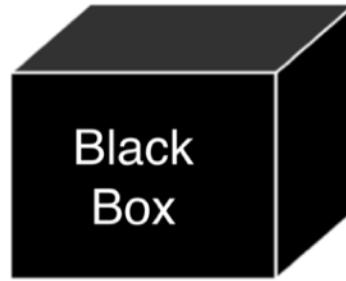
Wojciech Samek
TU Berlin & Fraunhofer HHI



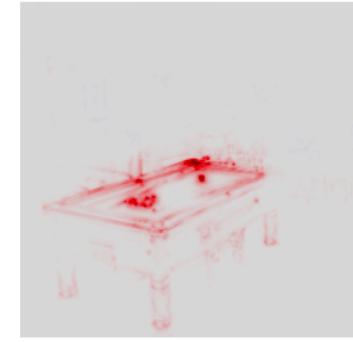
Explain? Yes We Can



classify



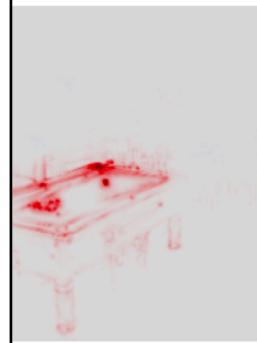
explain



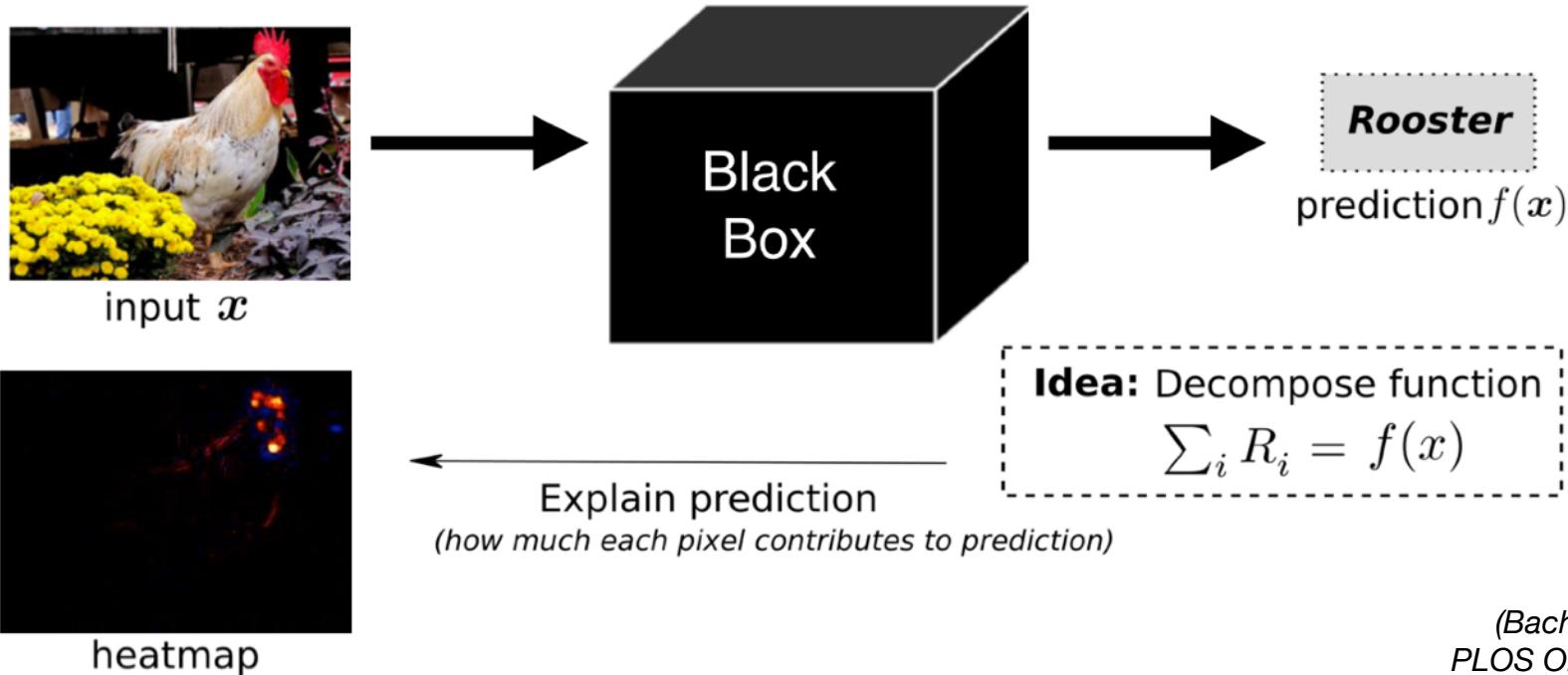
Explain? Yes We Can



| | | | | |
|-------------------------------|------------------------------|--------------------------|----------------------------|-----------------------------|
| Baehrens'10 Gradient | Sundarajan'17 Int Grad | Zintgraf'17 Pred Diff | Ribeiro'16 LIME | Haufe'15 Pattern |
| Zurada'94 Gradient | Symonian'13 Gradient | Zeiler'14 Occlusions | Fong'17 M Perturb | Kindermans'17 PatternNet |
| Poulin'06 Additive | Lundberg'17 Shapley | Bazen'13 Taylor | Montavon'17 Deep Taylor | Shrikumar'17 DeepLIFT |
| Zeiler'14 Deconv | Landecker'13 Contrib Prop | Bach'15 LRP | Zhang'16 Excitation BP | |
| Caruana'15 Fitted Additive | Springenberg'14 Guided BP | Zhou'16 GAP | Selvaraju'17 Grad-CAM | |



XAI 1.0: Layer-wise Relevance Propagation

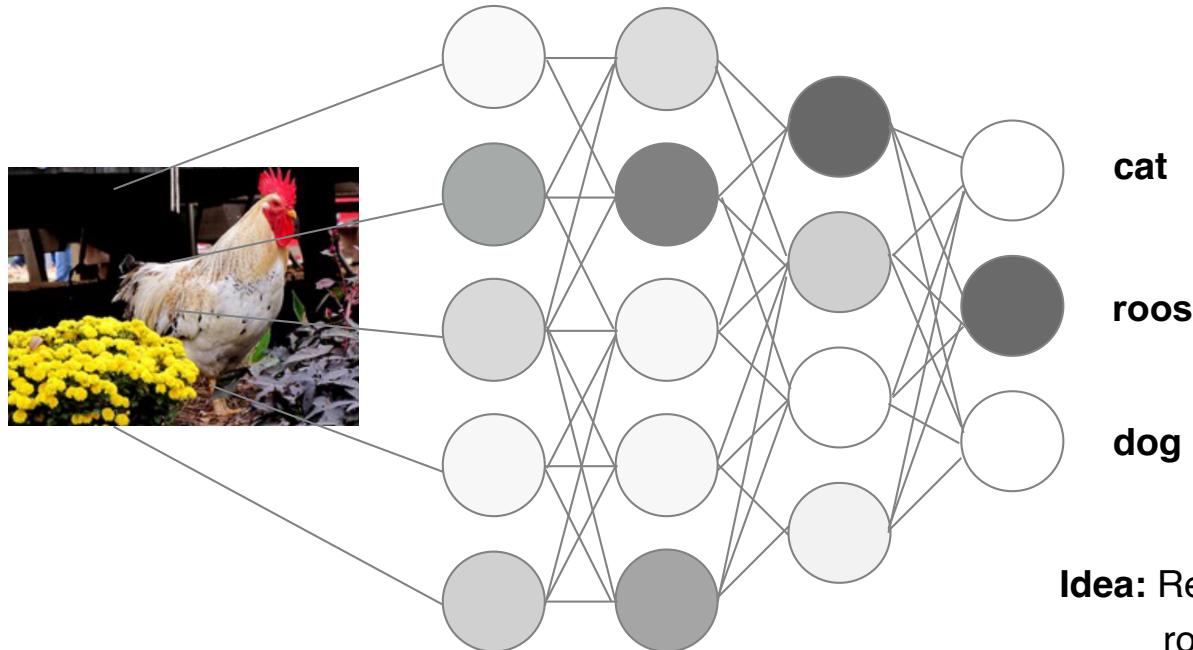


(Bach et al.,
PLOS ONE, 2015)

Layer-wise Relevance Propagation is a general approach to explain predictions of AI.

XAI 1.0: Layer-wise Relevance Propagation

Classification

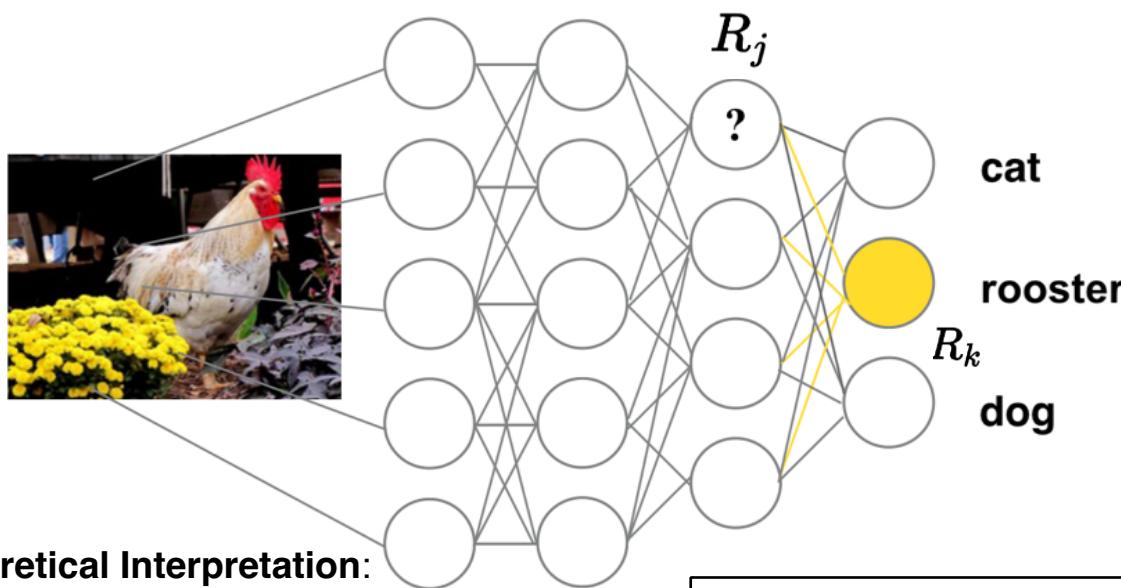


Idea: Redistribute the evidence for class
rooster back to image space.

5

XAI 1.0: Layer-wise Relevance Propagation

Explanation



how much has j contributed to activation of k

LRP:

(1) decompose

$$R_{j \leftarrow k} = \frac{z_{jk}}{z_k} R_k$$

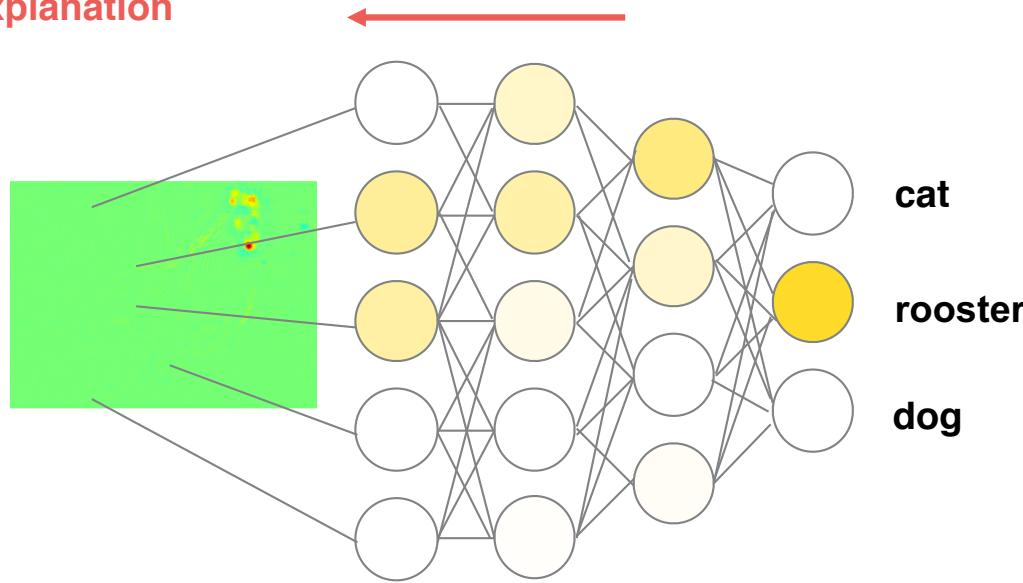


(2) aggregate

$$R_j = \sum R_{j \leftarrow k}$$

XAI 1.0: Layer-wise Relevance Propagation

Explanation



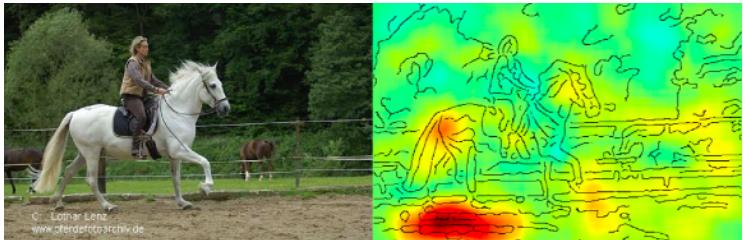
Layer-wise relevance conservation

$$\sum_i R_i = \dots = \sum_i R_i^{(l)} = \sum_j R_j^{(l+1)} = \dots = f(x)$$

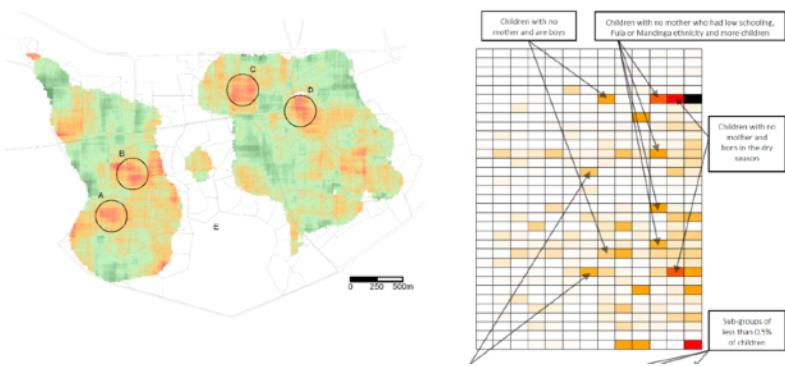
7

More Than Just Explaining ...

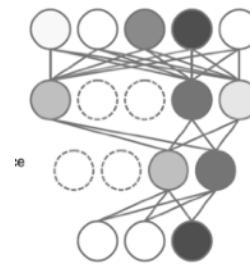
Validation & Debugging



Scientific Insights



Pruning & Quantization

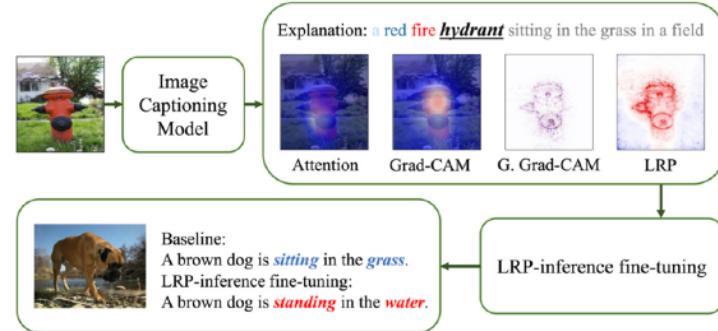


| | | | |
|------|-------|------|-------|
| 1.36 | -0.56 | 0 | -0.56 |
| 0 | 0 | 0 | -0.56 |
| 0 | -0.56 | 0 | 0 |
| 1.36 | 0 | 1.36 | -0.56 |

| | | | |
|-------|-------|-------|-------|
| -0.03 | 0.01 | 0.01 | -0.06 |
| 0.02 | 0.07 | 0.08 | -0.04 |
| 0.03 | -0.08 | 0.10 | -0.03 |
| 0.01 | 0.05 | -0.04 | 0.04 |

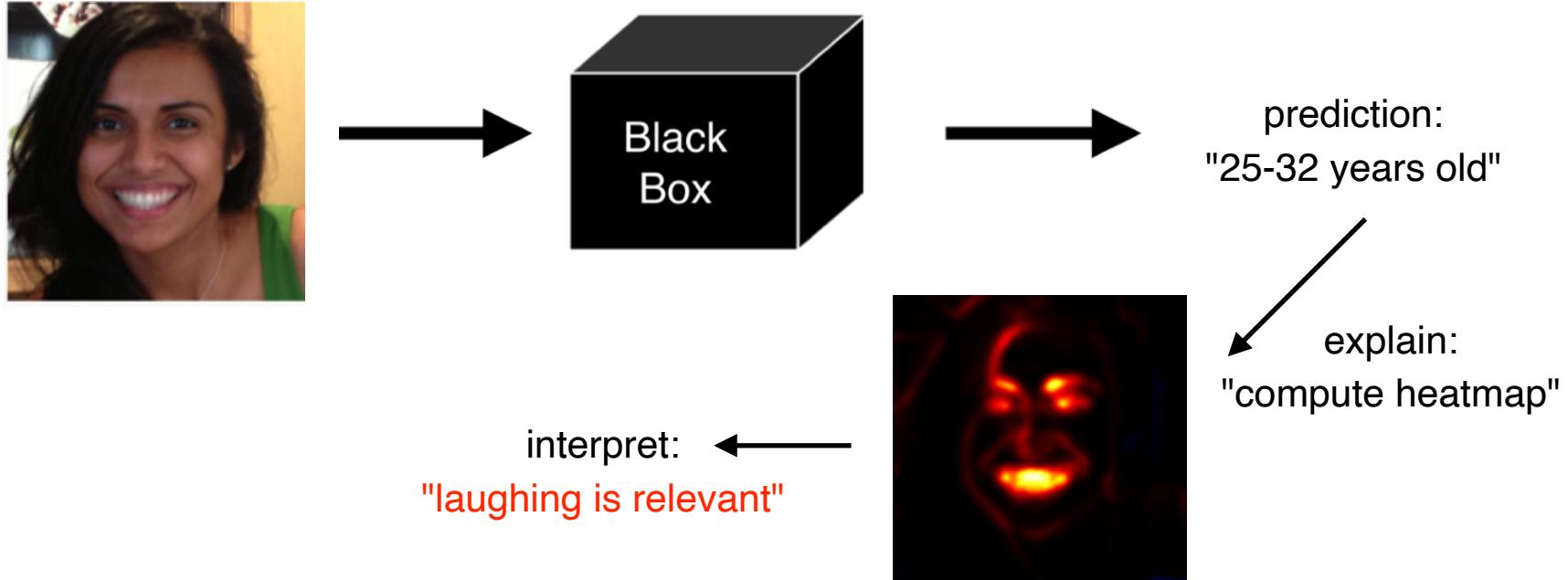
| | | | |
|-------|-------|-------|-------|
| -0.04 | 0.01 | 0.01 | -0.03 |
| 0.02 | -0.07 | 0.08 | -0.02 |
| 0.03 | -0.04 | 0.10 | -0.03 |
| 0.01 | 0.05 | -0.05 | 0.02 |

XAI-Based Model Improvement

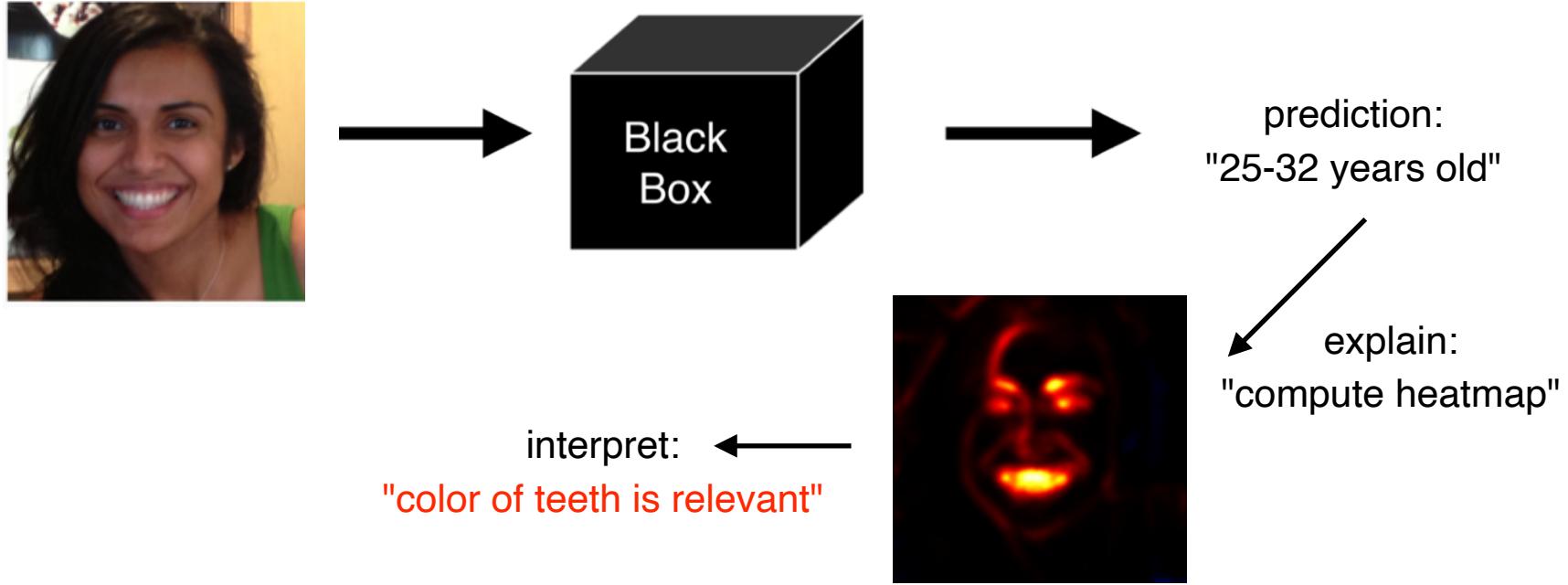


How understandable is the explanation ?

Limitations of Attribution Maps

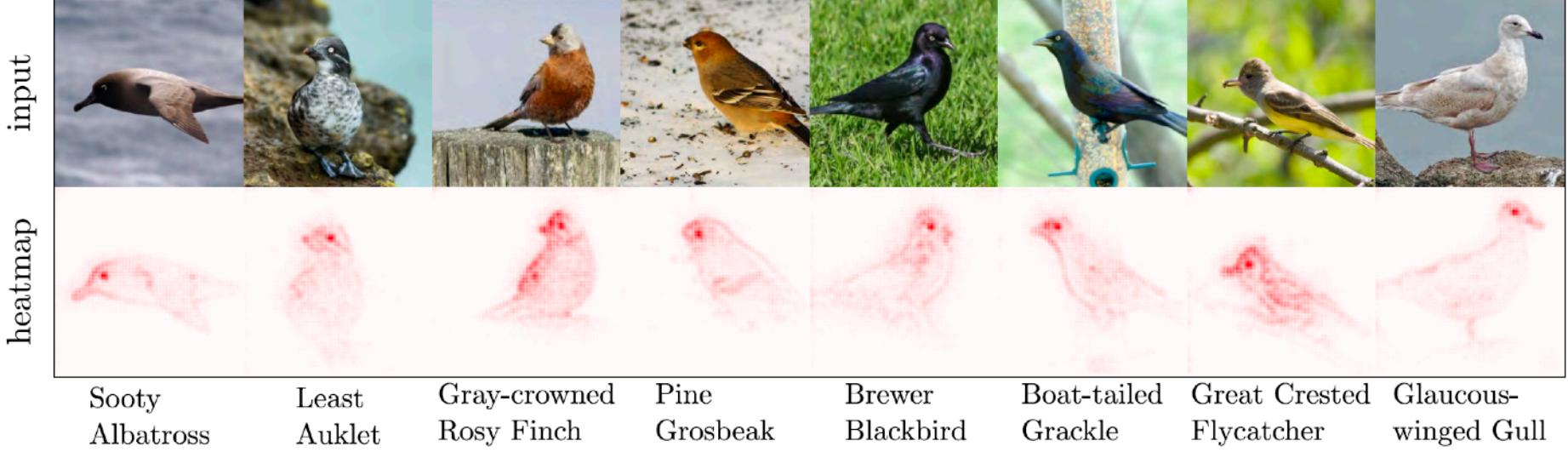


Limitations of Attribution Maps



Problem: We know *where* something relevant is not *what* it is.

Limitations of Attribution Maps



Problem: Loss of explanatory value due to superposition of concepts (all relevant).

Entering XAI 2.0

From “Where” to “What”: Towards Human-Understandable Explanations through Concept Relevance Propagation

Reduan Achtibat^{1,*}

Maximilian Dreyer^{1,*}

Ilona Eisenbraun¹

Sebastian Bosse¹

Thomas Wiegand^{1,2,3}

Wojciech Samek^{1,2,3,†}

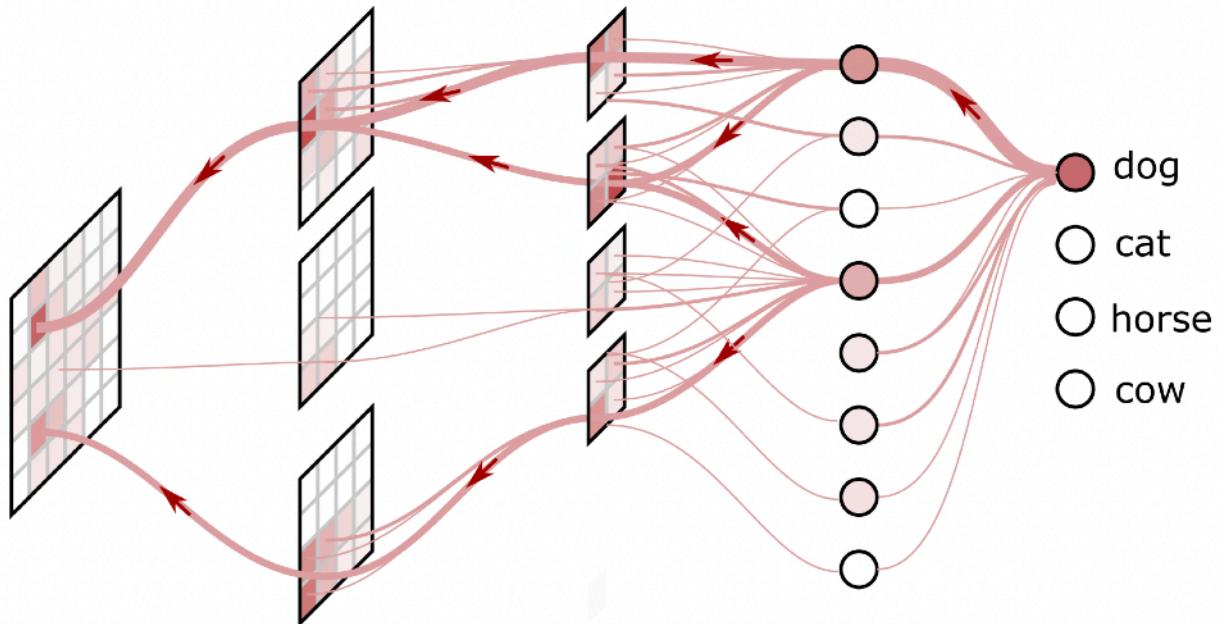
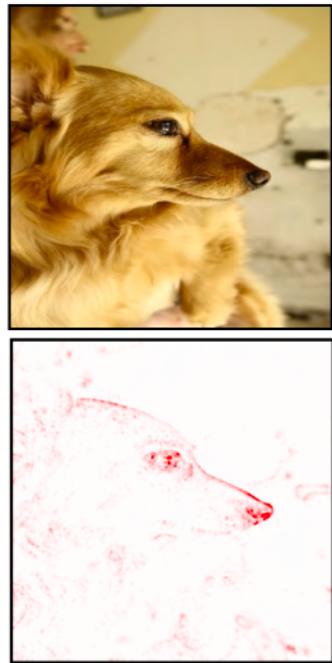
Sebastian Lapuschkin^{1,†}

<https://arxiv.org/abs/2206.03208>

13

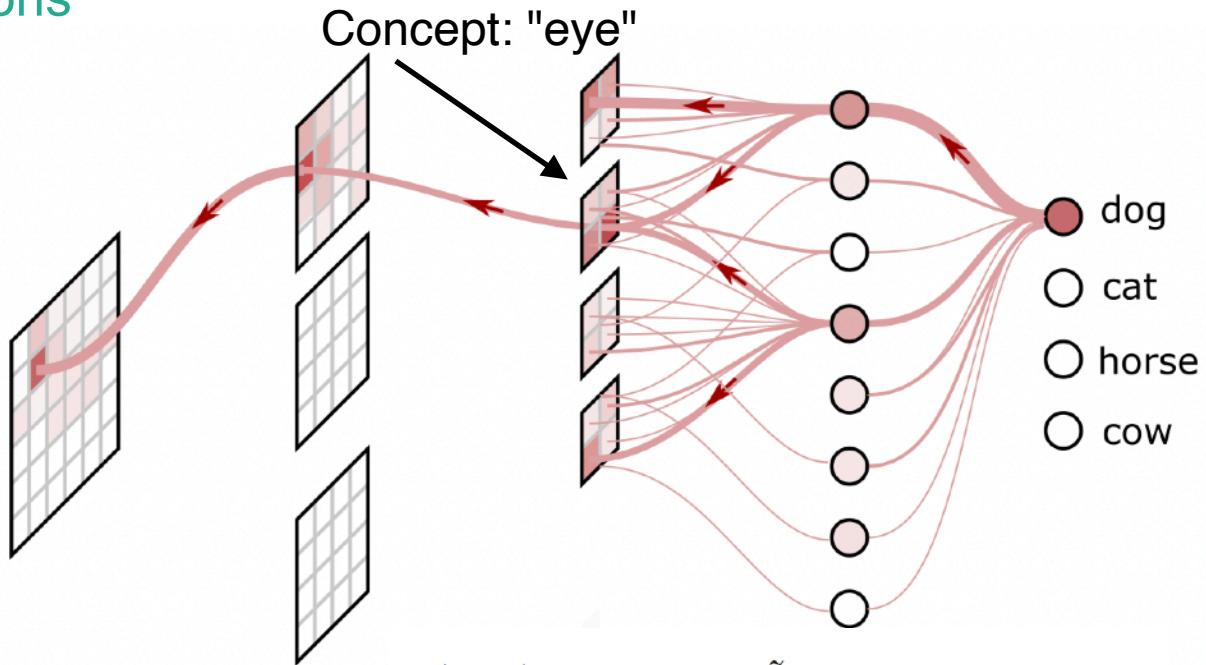
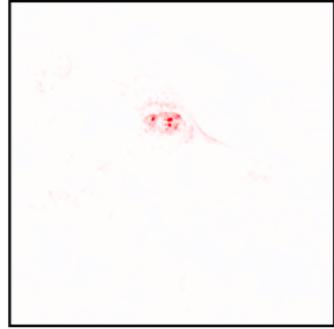
Concept Relevance Propagation

(i) Conditional Explanations



Concept Relevance Propagation

(i) Conditional Explanations

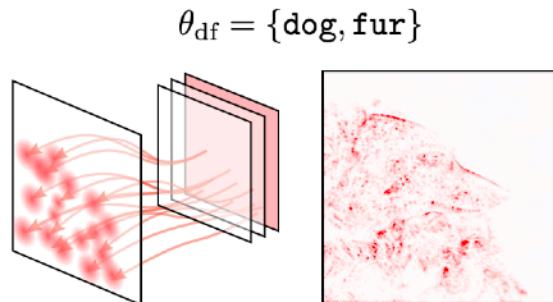
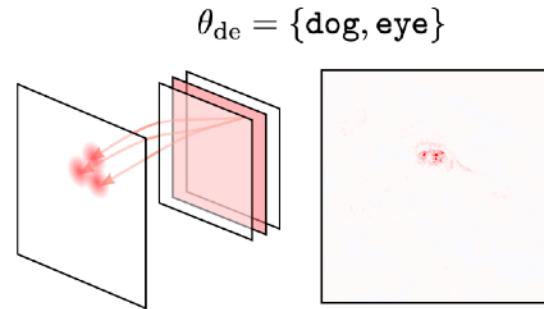
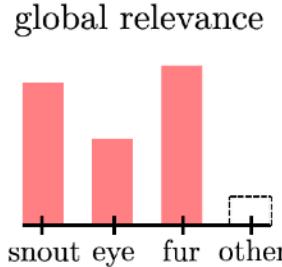
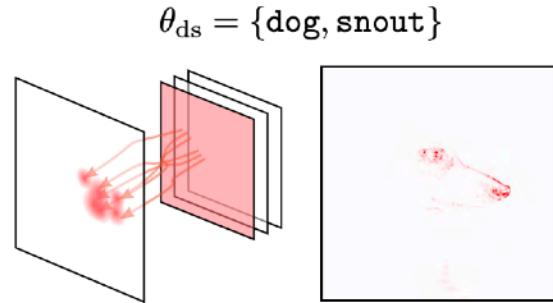
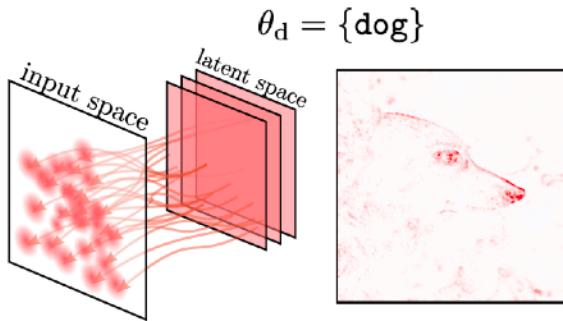


$$R_{i \leftarrow j}^{(l-1,l)}(\mathbf{x}|\theta \cup \theta_l) = \frac{z_{ij}}{z_j} \cdot \sum_{c_l \in \theta_l} \delta_{j c_l} \cdot R_j^l(\mathbf{x}|\theta)$$

15

Concept Relevance Propagation

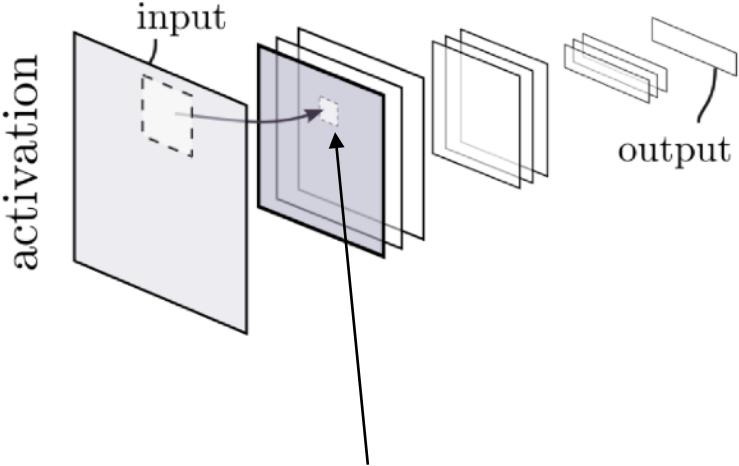
(i) Conditional Explanations



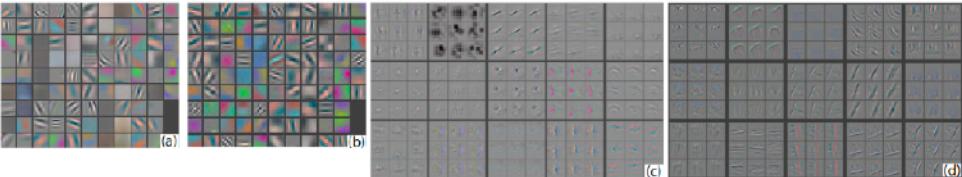
But usually we do not know what concept the channel is encoding?



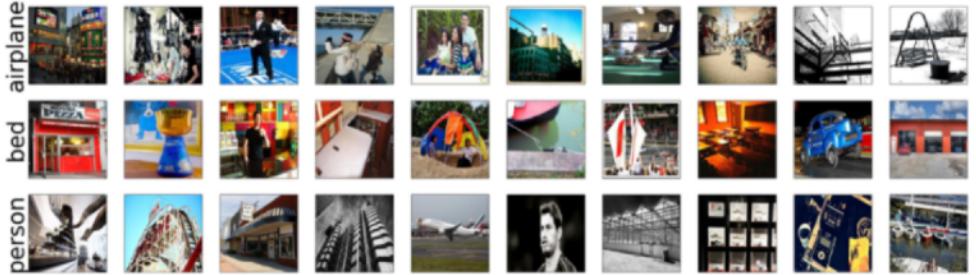
Addressing the "What"-Question



(Zeiler et al., 2014) feature visualization



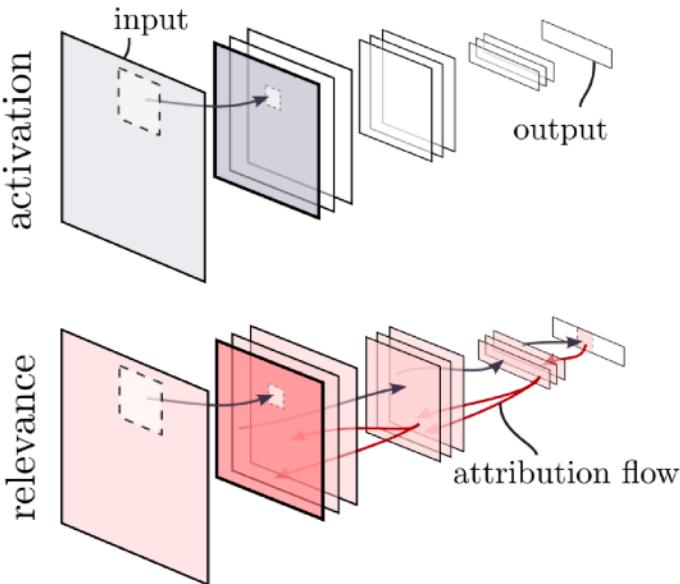
(Chen et al., 2020) data-based activation maximization



Relevance Maximization

or why Activation Maximization is not the Answer

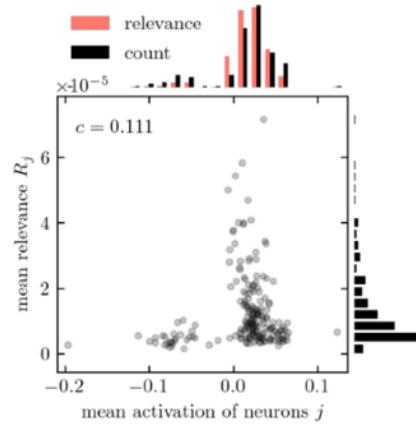
activation vs relevance flow



activation = stimulation
without task-context

relevance = usefulness
within a task-context

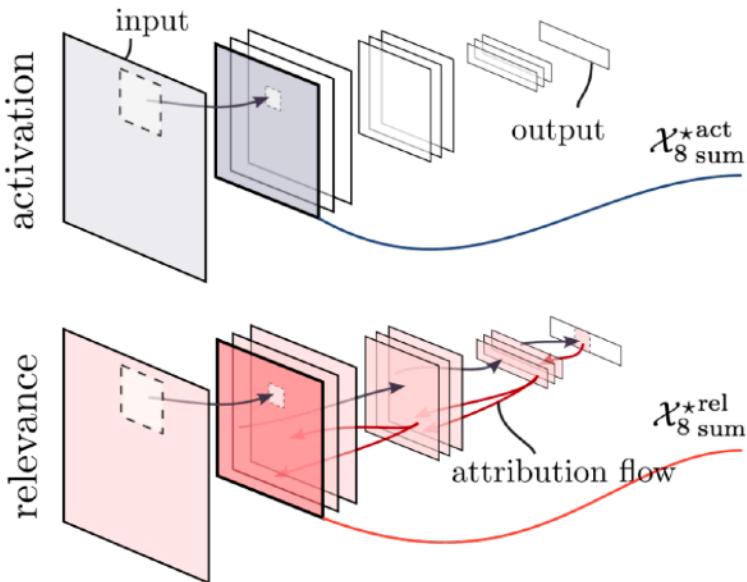
practically no correlation
between activation and
relevance



Concept Relevance Propagation

(ii) Understand the "What"-Question through RelMax

activation vs relevance flow → result in different example sets



within task-context

20

More Insights Into Reference Samples



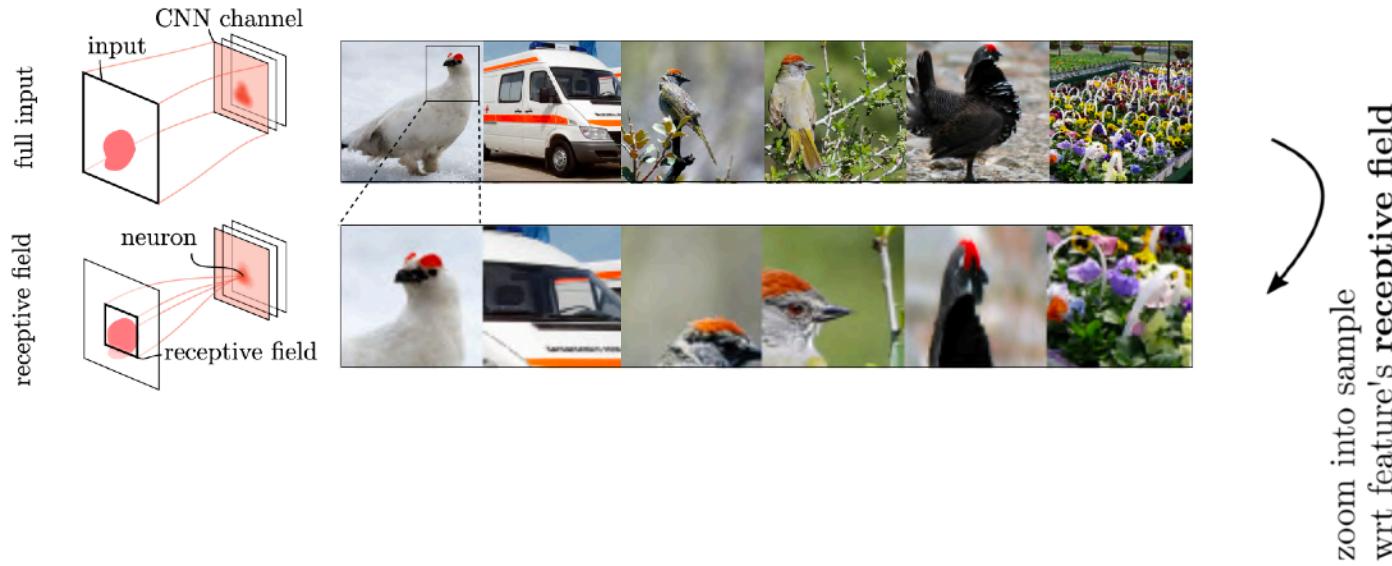
typical scenario in literature:

provide **full input-sized** explanatory examples.



which one is / are the relevant feature(s) ?

More Insights Into Reference Samples

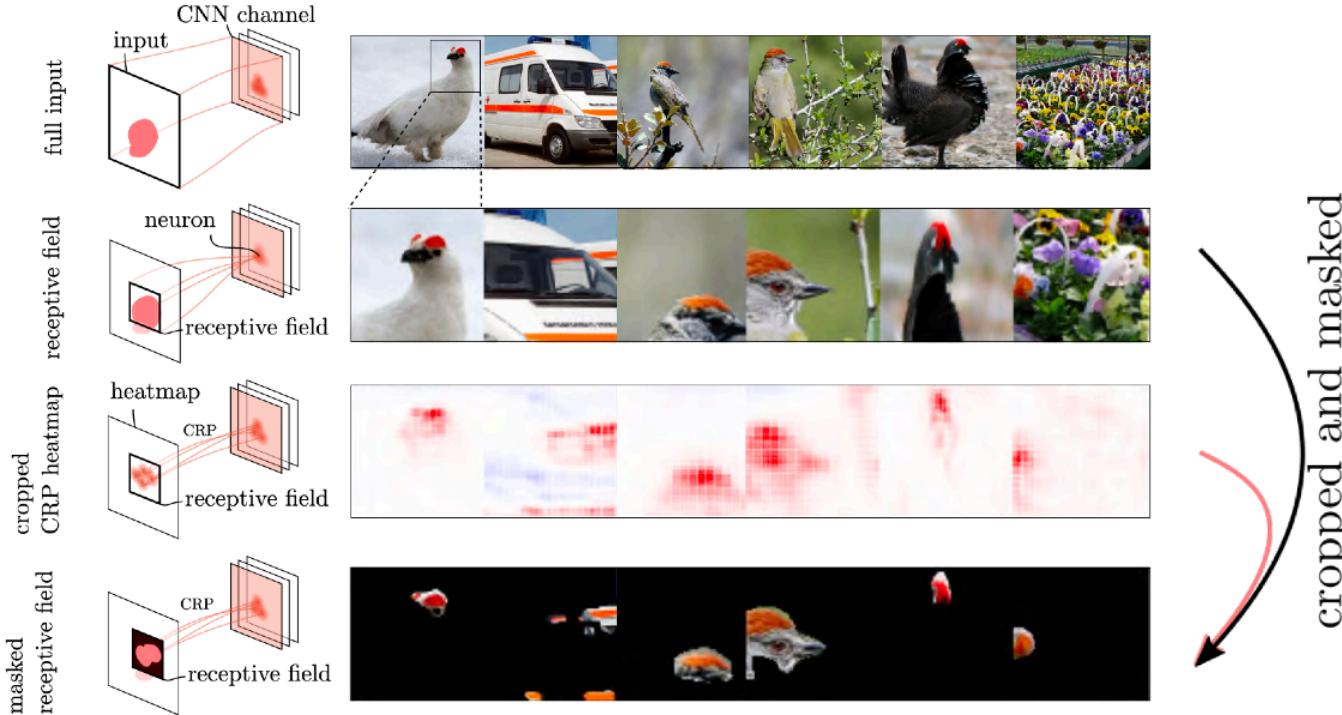


More Insights Into Reference Samples



Concept Relevance Propagation

(iii) Highlight the Key Feature of a Concept

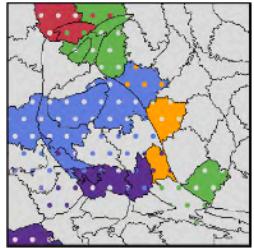


Concept Atlas

input



concept atlas

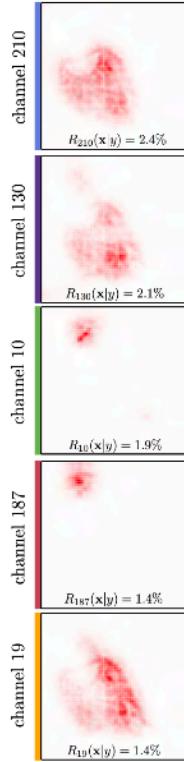


- most relevant
- 2nd most relevant

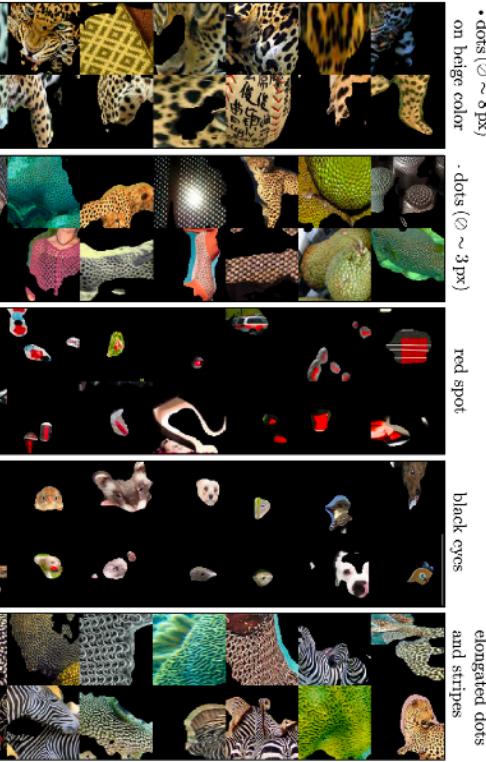
channels

| | |
|-----|-------|
| 210 | 187 |
| 130 | 19 |
| 10 | other |

conditional heatmap



masked reference samples

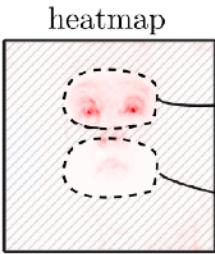
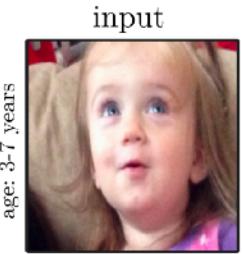


Concept Relevance Propagation

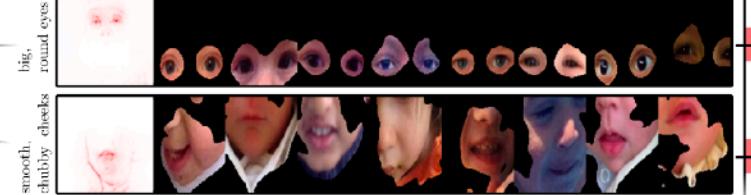
explain individual prediction

local XAI

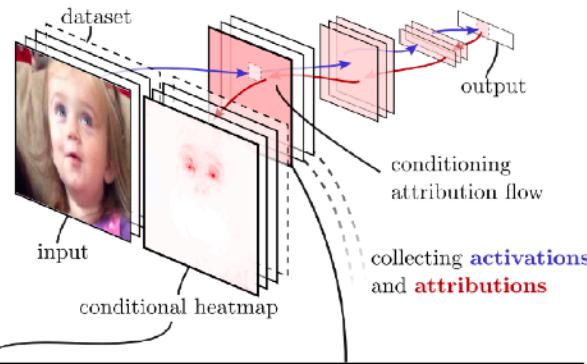
Where is the model looking at?



big, round eyes



glocal XAI
What features is the model using here?

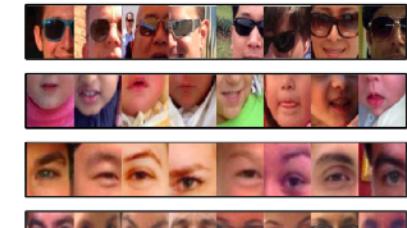


explain model representation

global XAI

What features exist?

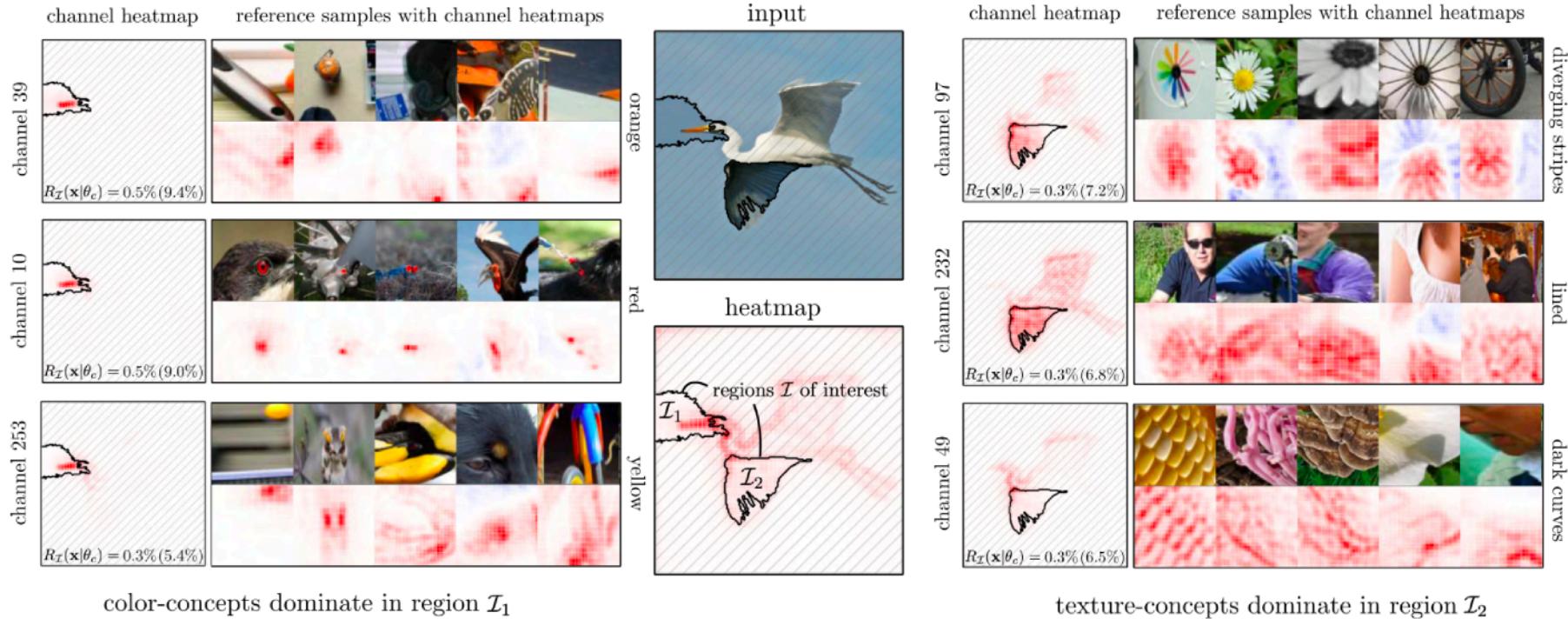
feature visualization



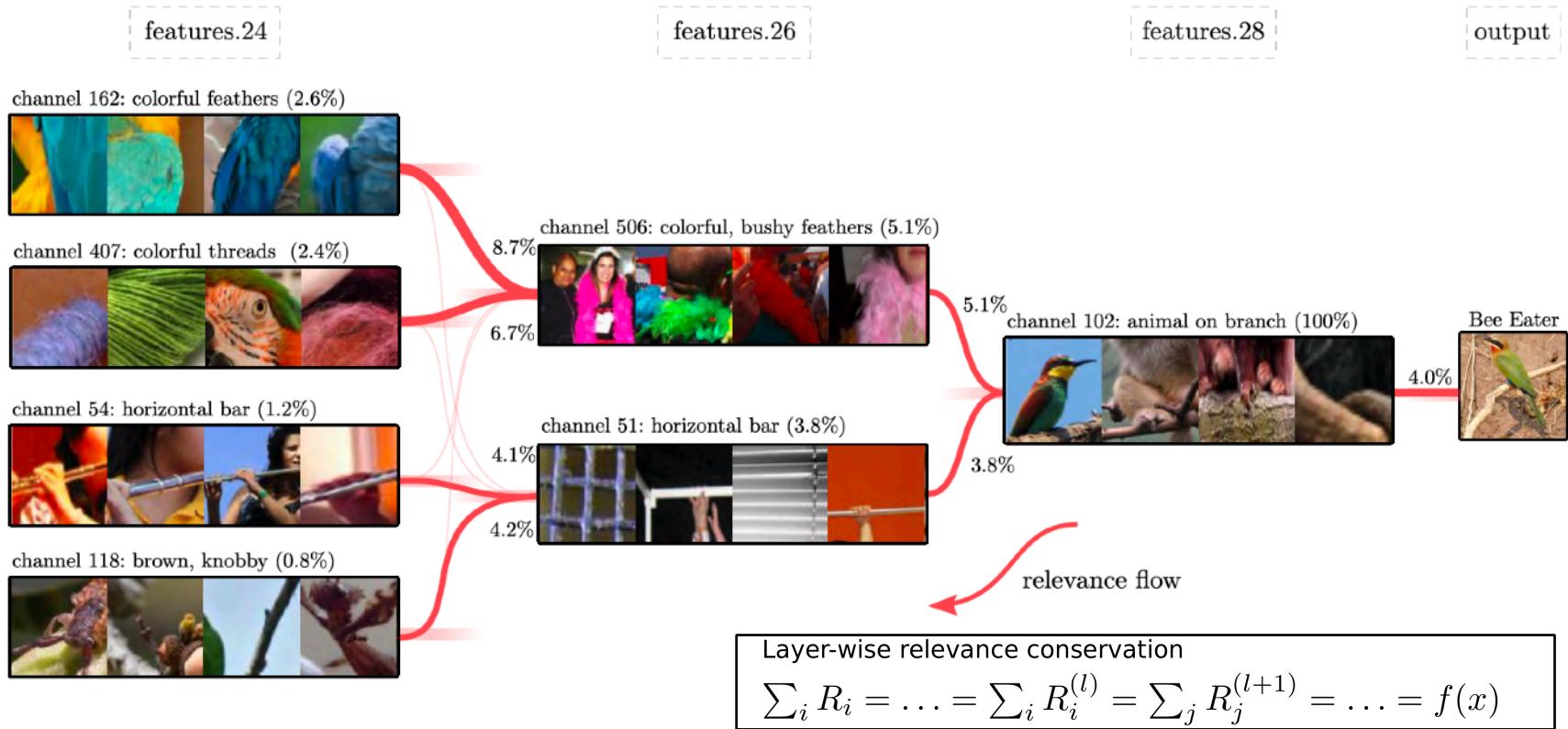
(Achtibat et al. 2022)

26

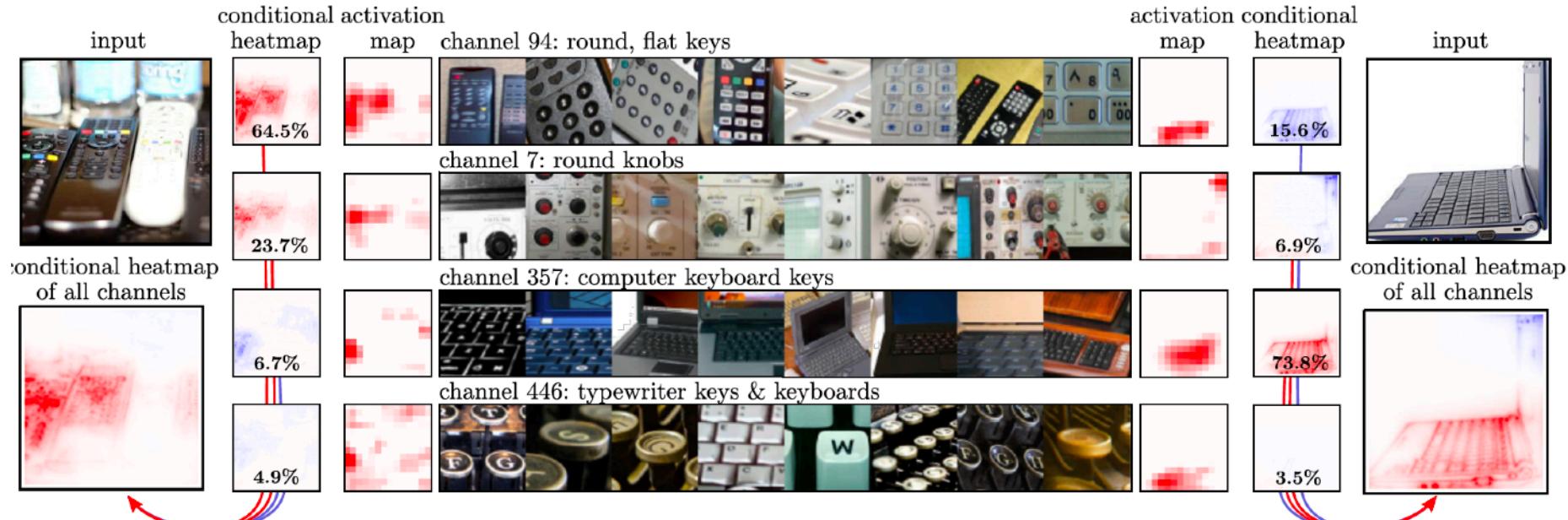
Local Analyses as a Bottom-Up Complement



Concept Composition



Understanding Concept Spaces



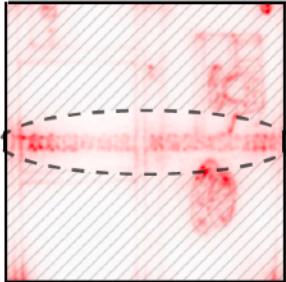
Identifying Clever Hans

Concept-based Reverse Search

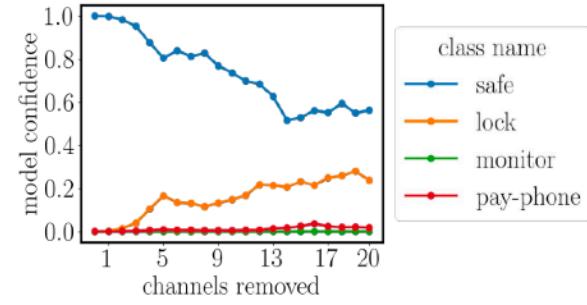
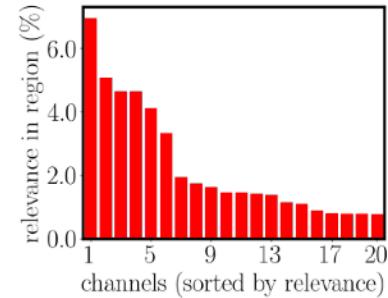
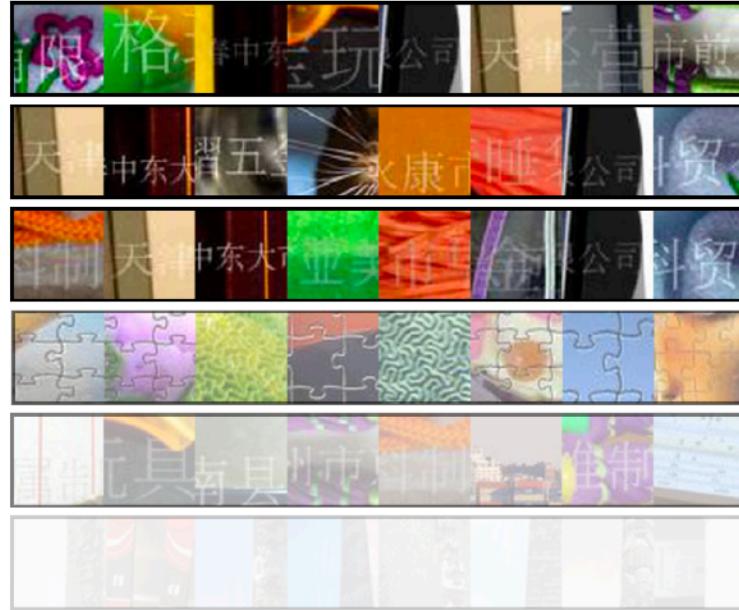
input



input heatmap



most relevant channels in region



Identifying Clever Hans

Concept-based Reverse Search

whistle



mob



screw



mosquito net



can opener



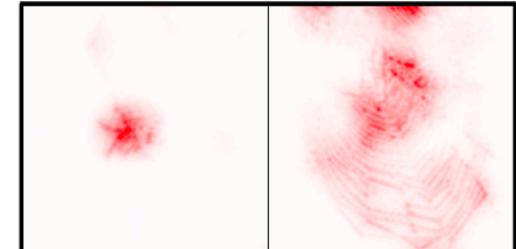
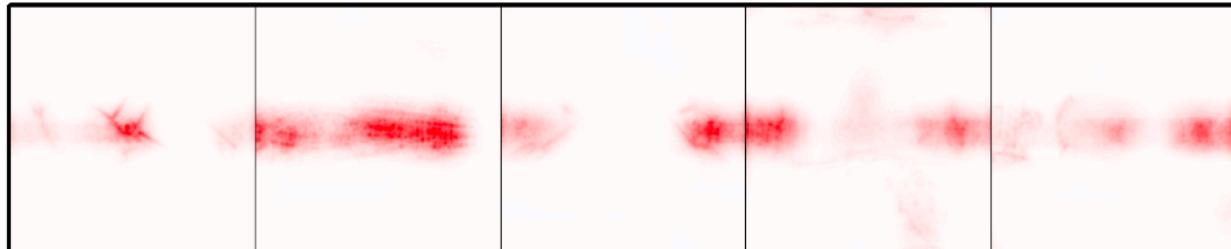
puma



spiderweb



conditional heatmap $R(\mathbf{x}|\theta = \{c_{361}, y\})$



Fixing the Model: Adapt encoding space globally [Anders, Weber, et al. 2022] or rather outcome-dependently?

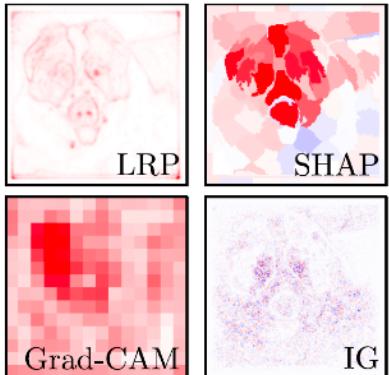
31

Human Evaluation

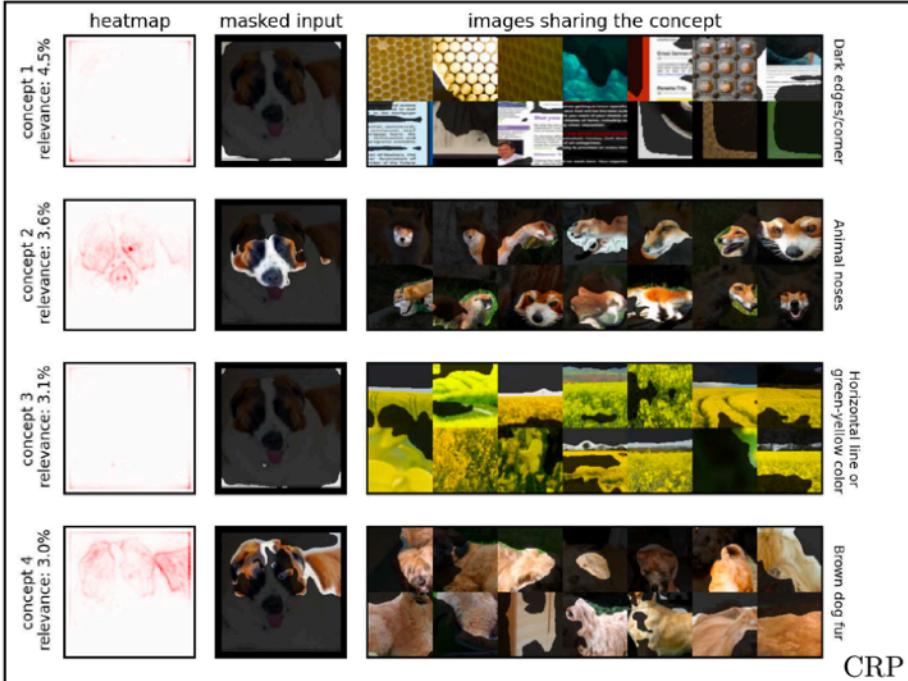
a input with artifact



b local explanations



c global CRP explanation



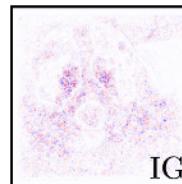
Human Evaluation

a input with artifact

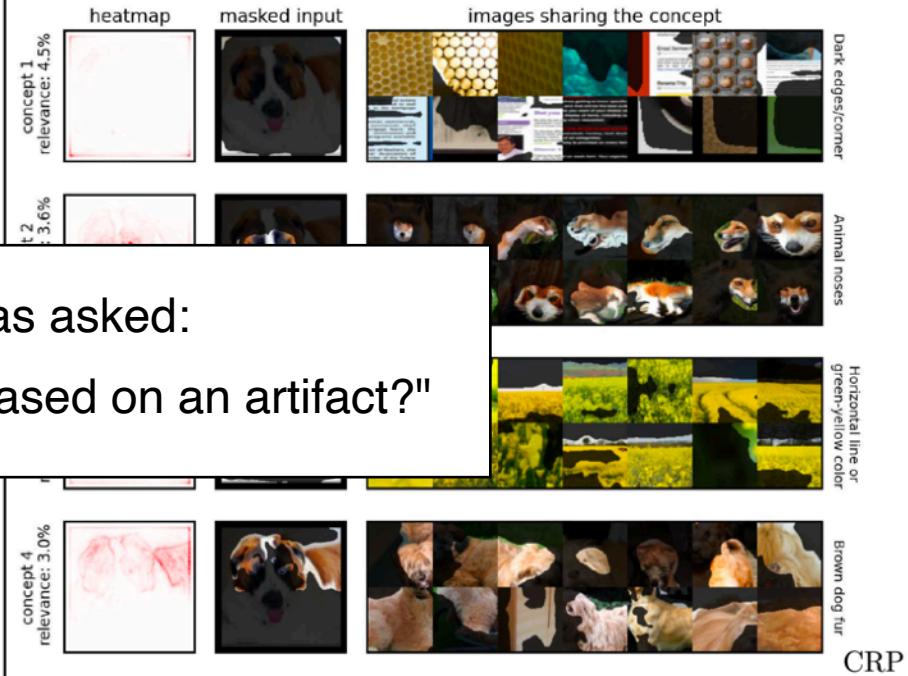


Human was asked:

"Was the prediction based on an artifact?"



c global CRP explanation

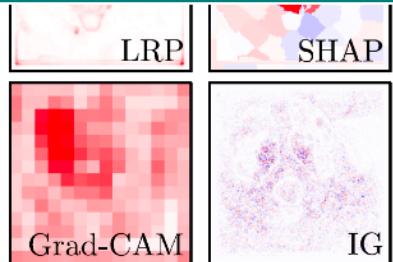


Human Evaluation

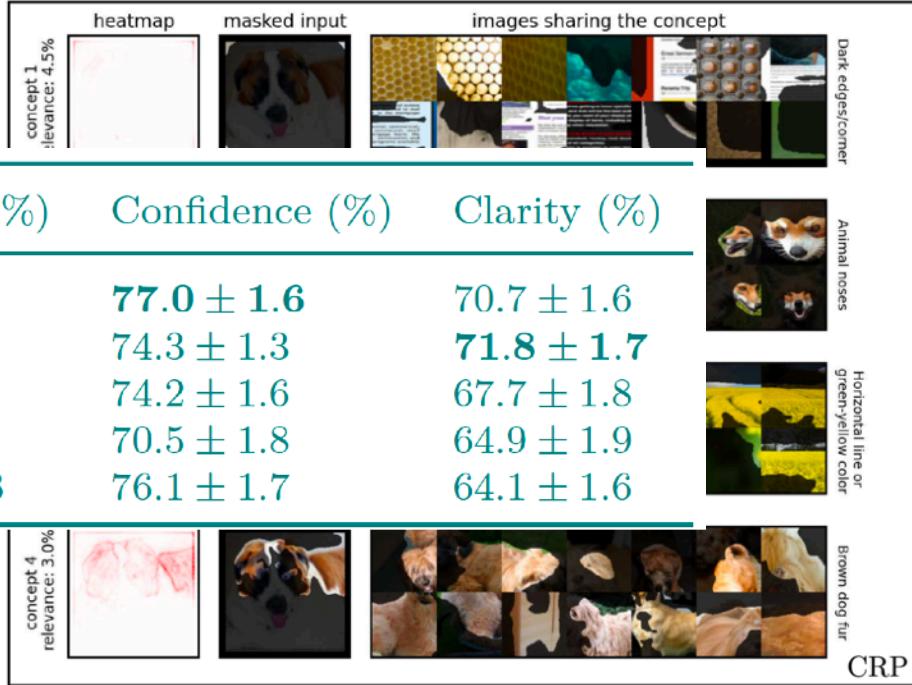
a input



| Method | Accuracy (%) | F1-Score (%) | Confidence (%) | Clarity (%) |
|------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|
| IG | 51.7 ± 1.9 | 52.7 ± 2.9 | 77.0 ± 1.6 | 70.7 ± 1.6 |
| LRP | 56.6 ± 2.9 | 61.6 ± 2.4 | 74.3 ± 1.3 | 71.8 ± 1.7 |
| SHAP | 58.3 ± 2.7 | 62.2 ± 2.4 | 74.2 ± 1.6 | 67.7 ± 1.8 |
| Grad-CAM | 63.7 ± 3.4 | 67.4 ± 2.3 | 70.5 ± 1.8 | 64.9 ± 1.9 |
| CRP (ours) | 80.9 ± 3.4 | 82.3 ± 1.8 | 76.1 ± 1.7 | 64.1 ± 1.6 |



c global CRP explanation

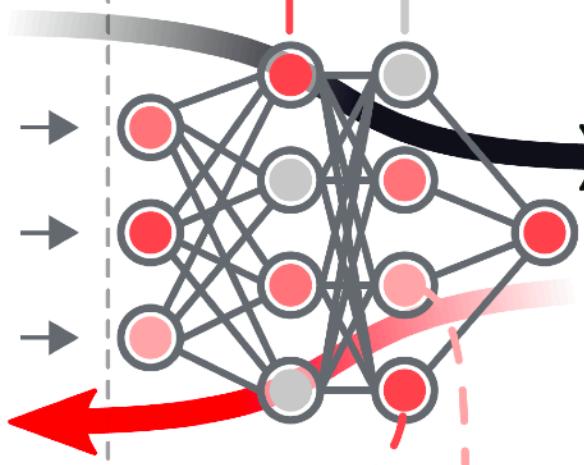
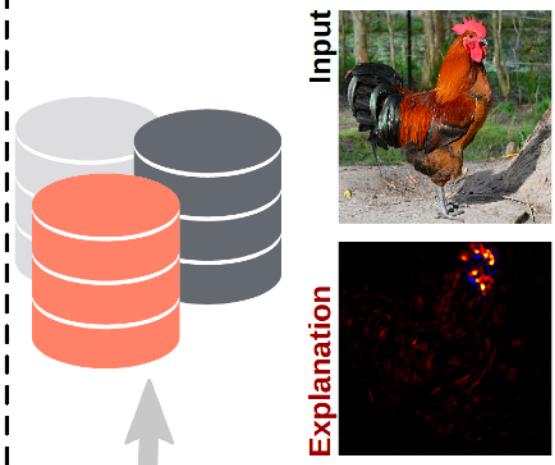


Conclusion

MODEL LEVEL XAI

relevant and irrelevant
model components

DATA(SET) LEVEL XAI

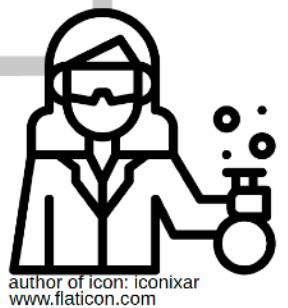


HUMAN LEVEL XAI

... "mainly
because of its
red comb and
throat wattles."

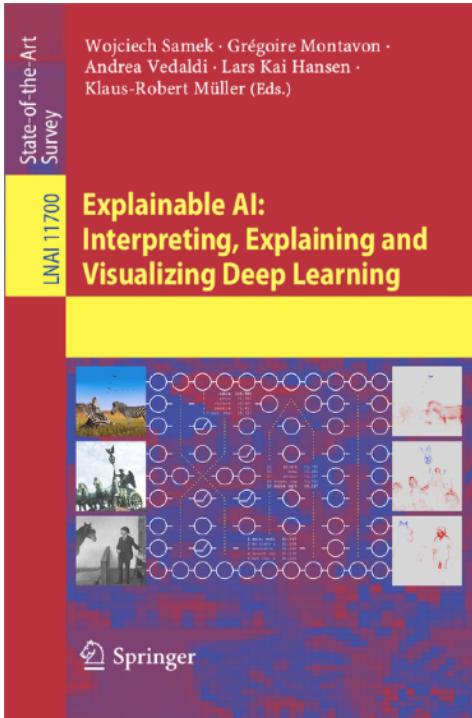


Secondary
indicators are
feather-like
structures."

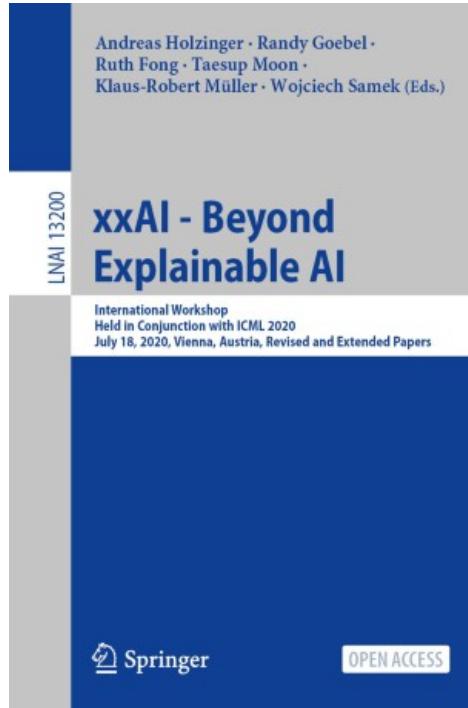


author of icon: iconixar
www.flaticon.com

From XAI to XXAI



(2019)



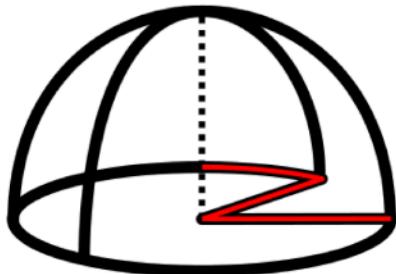
(2022)

New directions in XAI:

- Explain & Improve
- Concept-Level XAI
- Regression, RL, Unsup. L.
- Non-interpretable domains
- Beyond Explaining

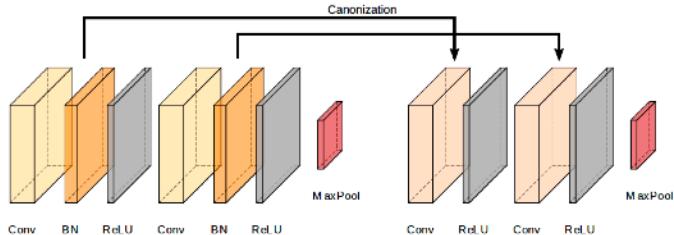
...

Toolboxes



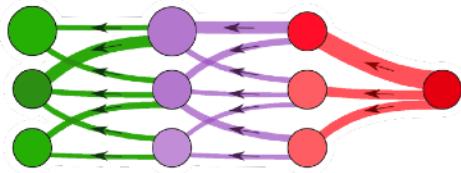
zennit

Canonization



QUANTUS

iNNvestigate



ExplainableAI.jl

Refs: [Alber et al. 2019; Anders, Neumann, et al. 2021; Motzkus et al. 2022; Hedström et al. 2022; Hill 2022]

38

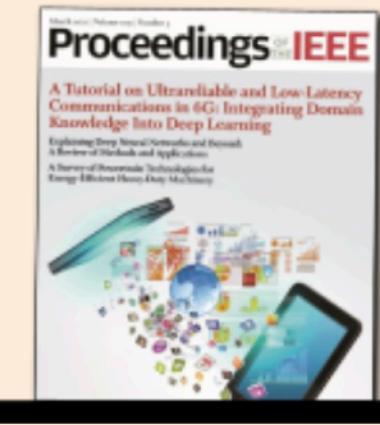
References

W Samek, G Montavon, S Lapuschkin, C Anders, KR Müller

Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications

Proceedings of the IEEE, 109(3):247-278, 2021

With the broader and highly successful usage of machine learning (ML) in industry and the sciences, there has been a growing demand for explainable artificial intelligence (XAI). Interpretability and explanation methods for gaining a better understanding of the problem-solving abilities and strategies of nonlinear ML, in particular, deep neural networks, are, therefore, receiving increased attention. In this work, we aim to: 1) provide a timely overview of this active emerging field, with a focus on “post hoc” explanations, and explain its theoretical foundations; 2) put interpretability algorithms to a test both from a theory and comparative evaluation perspective using extensive simulations; 3) outline best practice aspects, i.e., how to best include interpretation methods into the standard usage of ML; and 4) demonstrate successful usage of XAI in a representative selection of application scenarios. Finally, we discuss challenges and possible future directions of this exciting foundational field of ML.



References

Tutorial / Overview Papers

- W Samek, G Montavon, S Lapuschkin, C Anders, KR Müller. [Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications](#)
Proceedings of the IEEE, 109(3):247-278, 2021 [[preprint](#), [bibtex](#)]
- A Holzinger, A Saranti, C Molnar, P Biece, W Samek. [Explainable AI Methods - A Brief Overview](#)
xxAI - Beyond Explainable AI, Springer LNAI, 13200:13-38, 2022 [[bibtex](#)]
- G Montavon, W Samek, KR Müller. [Methods for Interpreting and Understanding Deep Neural Networks](#)
Digital Signal Processing, 73:1-15, 2018 [[bibtex](#)]
- W Samek, T Wiegand, KR Müller. [Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models](#)
ITU Journal: ICT Discoveries, 1(1):39-48, 2018 [[preprint](#), [bibtex](#)]
- W Samek, KR Müller. [Towards Explainable Artificial Intelligence](#)
in Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer LNCS, 11700:5-22, 2019 [[preprint](#), [bibtex](#)]
- G Montavon, A Binder, S Lapuschkin, W Samek, KR Müller. [Layer-Wise Relevance Propagation: An Overview](#)
in Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer LNCS, 11700:193-209, 2019 [[preprint](#), [bibtex](#), [demo code](#)]

References

- S Bach, A Binder, G Montavon, F Klauschen, KR Müller, W Samek. [On Pixel-wise Explanations for Non-Linear Classifier Decisions by Layer-wise Relevance Propagation](#)
PLOS ONE, 10(7):e0130140, 2015 [[preprint](#), [bibTex](#)]
- G Montavon, S Lapuschkin, A Binder, W Samek, KR Müller. [Explaining NonLinear Classification Decisions with Deep Taylor Decomposition](#)
Pattern Recognition, 65:211–222, 2017 [[preprint](#), [bibTex](#)]
- M Kohlbrenner, A Bauer, S Nakajima, A Binder, W Samek, S Lapuschkin. [Towards best practice in explaining neural network decisions with LRP](#)
Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN), 1-7, 2019 [[preprint](#), [bibTex](#)]
- W Samek, L Arras, A Osman, G Montavon, KR Müller. [Explaining the Decisions of Convolutional and Recurrent Neural Networks](#)
Mathematical Aspects of Deep Learning, Cambridge University Press, 2021
- A Binder, G Montavon, S Lapuschkin, KR Müller, W Samek. [Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers](#)
Artificial Neural Networks and Machine Learning – ICANN 2016, Part II, LNCS, Springer-Verlag, 9887:63-71, 2016 [[preprint](#), [bibTex](#)]
- PJ Kindermans, KT Schütt, M Alber, KR Müller, D Erhan, B Kim, S Dähne. [Learning how to explain neural networks: PatternNet and PatternAttribution](#)
Proceedings of the International Conference on Learning Representations (ICLR), 2018
- L Rieger, P Chormai, G Montavon, LK Hansen, KR Müller. [Structuring Neural Networks for More Explainable Predictions](#)
in Explainable and Interpretable Models in Computer Vision and Machine Learning, 115-131, Springer SSCML, 2018

References

Explaining Beyond DNN Classifiers

- S Letzgus, P Wagner, J Lederer, W Samek, KR Müller, G Montavon. [Toward Explainable AI for Regression Models](#)
[Signal Processing Magazine, 2022](#) [[preprint](#), [bibtex](#)]
- G Montavon, J Kauffmann, W Samek, KR Müller. [Explaining the Predictions of Unsupervised Learning Models](#)
[xxAI - Beyond Explainable AI, Springer LNAI, 13200:117-138, 2022](#) [[preprint](#), [bibtex](#)]
- A Ali, T Schnake, O Eberle, G Montavon, KR Müller, L Wolf. [XAI for Transformers: Better Explanations through Conservative Propagation](#)
[arXiv:2202.07304, 2022](#)
- J Kauffmann, KR Müller, G Montavon. [Towards Explaining Anomalies: A Deep Taylor Decomposition of One-Class Models](#)
[Pattern Recognition, 107198, 2020](#) [[preprint](#)]
- L Arras, J Arjona, M Widrich, G Montavon, M Gillhofer, KR Müller, S Hochreiter, W Samek. [Explaining and Interpreting LSTMs](#)
[in Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer LNCS, 11700:211-238, 2019](#) [[preprint](#), [bibtex](#)]
- J Kauffmann, M Esders, L Ruff, G Montavon, W Samek, KR Müller. [From Clustering to Cluster Explanations via Neural Networks](#)
[arxiv:1906.07633v2, 2021](#) [[demo code](#)]
- O Eberle, J Büttner, F Kräutli, KR Müller, M Valleriani, G Montavon. [Building and Interpreting Deep Similarity Models](#)
[IEEE Transactions on Pattern Analysis and Machine Intelligence, 44\(3\):1149-1161, 2022](#)
- T Schnake, O Eberle, J Lederer, S Nakajima, K T. Schütt, KR Müller, G Montavon. [Higher-Order Explanations of Graph Neural Networks via Relevant Walks](#)
[IEEE Transactions on Pattern Analysis and Machine Intelligence, Early Access, 2021](#) [[demo code](#), [arxiv](#)]

References

Evaluation of Explanations

- L Arras, A Osman, W Samek. [CLEVR-XAI: A Benchmark Dataset for the Ground Truth Evaluation of Neural Network Explanations](#)
[Information Fusion, 81:14-40, 2022 \[preprint\]](#), [bibTex](#)
- W Samek, A Binder, G Montavon, S Bach, KR Müller. [Evaluating the Visualization of What a Deep Neural Network has Learned](#)
[IEEE Transactions on Neural Networks and Learning Systems, 28\(11\):2660-2673, 2017 \[preprint\]](#), [bibTex](#)
- L Arras, A Osman, KR Müller, W Samek. [Evaluating Recurrent Neural Network Explanations](#)
[Proceedings of the ACL Workshop on BlackboxNLP, 113-126, 2019 \[preprint\]](#), [bibTex](#)
- G Montavon. [Gradient-Based Vs. Propagation-Based Explanations: An Axiomatic Comparison](#)
in Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer LNCS, 11700:253-265, 2019 [\[bibTex\]](#)

References

Detecting Model and Dataset Artefacts

- S Lapuschkin, S Wäldchen, A Binder, G Montavon, W Samek, KR Müller. [Unmasking Clever Hans Predictors and Assessing What Machines Really Learn](#)
Nature Communications, 10:1096, 2019 [[preprint](#), [bibtex](#)]
- S Lapuschkin, A Binder, G Montavon, KR Müller, W Samek. [Analyzing Classifiers: Fisher Vectors and Deep Neural Networks](#)
Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2912-2920, 2016 [[preprint](#), [bibtex](#)]
- CJ Anders, T Marinc, D Neumann, W Samek, KR Müller, S Lapuschkin. [Analyzing ImageNet with Spectral Relevance Analysis: Towards ImageNet un-Hans'ed](#)
arXiv:1912.11425, 2019
- J Kauffmann, L Ruff, G Montavon, KR Müller. [The Clever Hans Effect in Anomaly Detection](#)
arXiv:2006.10609, 2020

References

Software Papers

- A Hedström, L Weber, D Bareeva, F Motzkus, W Samek, S Lapuschkin, MMC Höhne [Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanation](#)
arXiv:2202.06861, 2022 [[preprint](#), [bibtex](#)]
- CJ Anders, D Neumann, W Samek, KR Müller, S Lapuschkin [Software for Dataset-wide XAI: From Local Explanations to Global Insights with Zennit, CoRelAy, and ViRelAy](#).
arXiv:2106.13200, 2021 [[preprint](#), [bibtex](#)]
- M Alber, S Lapuschkin, P Seegerer, M Hägele, KT Schütt, G Montavon, W Samek, KR Müller, S Dähne, PJ Kindermans [iNNvestigate neural networks!](#)
Journal of Machine Learning Research, 20(93):1–8, 2019 [[preprint](#), [bibtex](#)]
- M Alber. [Software and Application Patterns for Explanation Methods](#)
in Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer LNCS, 11700:399–433, 2019 [[bibtex](#)]
- S Lapuschkin, A Binder, G Montavon, KR Müller, W Samek [The Layer-wise Relevance Propagation Toolbox for Artificial Neural Networks](#)
Journal of Machine Learning Research, 17(114):1–5, 2016 [[preprint](#), [bibtex](#)]

References

Application to Sciences

- I Sturm, S Bach, W Samek, KR Müller. [Interpretable Deep Neural Networks for Single-Trial EEG Classification](#)
[Journal of Neuroscience Methods, 274:141–145, 2016](#) [[preprint](#), [bibtex](#)]
- M Hägele, P Seegerer, S Lapuschkin, M Bockmayr, W Samek, F Klauschen, KR Müller, A Binder. [Resolving Challenges in Deep Learning-Based Analyses of Histopathological Images using Explanation Methods](#)
[Scientific Reports, 10:6423, 2020](#) [[preprint](#), [bibtex](#)]
- A Binder, M Bockmayr, M Hägele, S Wienert, D Heim, K Hellweg, A Stenzinger, L Parlow, J Budczies, B Goeppert, D Treue, M Kotani, M Ishii, M Dietel, A Hocke, C Denkert, KR Müller, F Klauschen. [Towards computational fluorescence microscopy: Machine learning-based integrated prediction of morphological and molecular tumor profiles](#)
[Nature Machine Intelligence, 3:355-366, 2021](#) [[preprint](#), [bibtex](#)]
- F Horst, S Lapuschkin, W Samek, KR Müller, WI Schöllhorn. [Explaining the Unique Nature of Individual Gait Patterns with Deep Learning](#)
[Scientific Reports, 9:2391, 2019](#) [[preprint](#), [bibtex](#)]
- D Slijepcevic, F Horst, B Horsak, S Lapuschkin, AM Raberger, A Kranzl, W Samek, C Breiteneder, WI Schöllhorn, M Zeppelzauer. [Explaining Machine Learning Models for Clinical Gait Analysis](#)
[ACM Transactions on Computing for Healthcare, 3\(2\):1-21, 2022](#) [[preprint](#), [bibtex](#)]
- AW Thomas, HR Heekeren, KR Müller, W Samek. [Analyzing Neuroimaging Data Through Recurrent Deep Learning Models](#)
[Frontiers in Neuroscience, 13:1321, 2019](#) [[preprint](#), [bibtex](#)]
- P Seegerer, A Binder, R Saitenmacher, M Bockmayr, M Alber, P Jurmeister, F Klauschen, KR Müller. [Interpretable Deep Neural Network to Predict Estrogen Receptor Status from Haematoxylin-Eosin Images](#)
[Artificial Intelligence and Machine Learning for Digital Pathology, Springer LNCS, 12090, 16-37, 2020](#) [[bibtex](#)]
- SM Hofmann, F Beyer, S Lapuschkin, M Loeffler, KR Müller, A Villringer, W Samek, AV Witte. [Towards the Interpretability of Deep Learning Models for Human Neuroimaging](#)
[bioRxiv 2021.06.25.449906, 2021](#) [[bibtex](#)]

References

Application to Text

- L Arras, F Horn, G Montavon, KR Müller, W Samek. "[What is Relevant in a Text Document?](#)": An Interpretable Machine Learning Approach
PLOS ONE, 12(8):e0181142, 2017 [[preprint](#), [bibtex](#)]
- L Arras, G Montavon, KR Müller, W Samek. [Explaining Recurrent Neural Network Predictions in Sentiment Analysis](#)
Proceedings of the EMNLP Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, 159-168, 2017 [[preprint](#), [bibtex](#)]
- L Arras, F Horn, G Montavon, KR Müller, W Samek. [Explaining Predictions of Non-Linear Classifiers in NLP](#)
Proceedings of the ACL Workshop on Representation Learning for NLP, 1-7, 2016 [[preprint](#), [bibtex](#)]
- F Horn, L Arras, G Montavon, KR Müller, W Samek. [Exploring text datasets by visualizing relevant words](#)
arXiv:1707.05261, 2017

References

Application to Images & Faces

- S Lapuschkin, A Binder, KR Müller, W Samek. [Understanding and Comparing Deep Neural Networks for Age and Gender Classification](#) Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW), 1629-1638, 2017 [[preprint](#), [bibtex](#)]
- C Seibold, W Samek, A Hilsmann, P Eisert. [Accurate and Robust Neural Networks for Face Morphing Attack Detection](#) Journal of Information Security and Applications, 53:102526, 2020 [[preprint](#), [bibtex](#)]
- S Bach, A Binder, KR Müller, W Samek. [Controlling Explanatory Heatmap Resolution and Semantics via Decomposition Depth](#) Proceedings of the IEEE International Conference on Image Processing (ICIP), 2271-2275, 2016 [[preprint](#), [bibtex](#)]
- A Binder, S Bach, G Montavon, KR Müller, W Samek. [Layer-wise Relevance Propagation for Deep Neural Network Architectures](#) Proceedings of the 7th International Conference on Information Science and Applications (ICISA), 6679:913-922, Springer Singapore, 2016 [[preprint](#), [bibtex](#)]
- F Arbabzadah, G Montavon, KR Müller, W Samek. [Identifying Individual Facial Expressions by Deconstructing a Neural Network](#) Pattern Recognition - 38th German Conference, GCPR 2016, Lecture Notes in Computer Science, 9796:344-354, 2016 [[preprint](#), [bibtex](#)]

References

Application to Video

- C Anders, G Montavon, W Samek, KR Müller. [Understanding Patch-Based Learning of Video Data by Explaining Predictions in Explainable AI: Interpreting, Explaining and Visualizing Deep Learning](#), Springer LNCS 11700:297-309, 2019 [[preprint](#), [bibtex](#)]
- V Srinivasan, S Lapuschkin, C Hellge, KR Müller, W Samek. [Interpretable human action recognition in compressed domain](#) Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1692-1696, 2017 [[preprint](#), [bibtex](#)]

Application to Speech

- S Becker, M Ackermann, S Lapuschkin, KR Müller, W Samek. [Interpreting and Explaining Deep Neural Networks for Classification of Audio Signals](#)
arXiv:1807.03418, 2018

References

Application to Neural Network Pruning

- S Ede, S Baghdadian, L Weber, A Nguyen, D Zanca, W Samek, S Lapuschkin. [Explain to Not Forget: Defending Against Catastrophic Forgetting with XAI](#)
arXiv:2205.01929, 2022 [[preprint](#), [bibtex](#)]
- D Becking, M Dreyer, W Samek, K Müller, S Lapuschkin. [ECQx: Explainability-Driven Quantization for Low-Bit and Sparse DNNs](#)
xxAI - Beyond Explainable AI, Springer LNAI, 13200:271-296, 2022 [[preprint](#), [bibtex](#)]
- S Yeom, P Seegerer, S Lapuschkin, A Binder, S Wiedemann, KR Müller, W Samek. [Pruning by Explaining: A Novel Criterion for Deep Neural Network Pruning](#)
Pattern Recognition, 115:107899, 2021 [[preprint](#), [bibtex](#)]

Interpretability and Causality

- A Rieckmann, P Dworzynski, L Arras, S Lapuschkin, W Samek, OA Arah, NH Rod, CT Ekstrom. [Causes of Outcome Learning: A causal inference-inspired machine learning approach to disentangling common combinations of potential causes of a health outcome](#)
International Journal of Epidemiology, dyac078, 2022 [[preprint](#), [bibtex](#)]

References

Model Improvement & Training Enhancement

- L Weber, S Lapuschkin, A Binder, W Samek. [Beyond Explaining: Opportunities and Challenges of XAI-Based Model Improvement](#)
arXiv:2203.08008, 2022 [[preprint](#)], [bibtex](#)]
- J Sun, S Lapuschkin, W Samek, A Binder. [Explain and Improve: LRP-Inference Fine Tuning for Image Captioning Models](#)
Information Fusion, 77:233-246, 2022 [[preprint](#)], [bibtex](#)]
- F Pahde, L Weber, CJ Anders, W Samek, S Lapuschkin. [PatClArC: Using Pattern Concept Activation Vectors for Noise-Robust Model Debugging](#)
arXiv:2202.03482, 2022 [[preprint](#)], [bibtex](#)]
- J Sun, S Lapuschkin, W Samek, Y Zhao, NM Cheung, A Binder. [Explanation-Guided Training for Cross-Domain Few-Shot Classification](#)
Proceedings of the 25th International Conference on Pattern Recognition (ICPR), 7609-7616, 2021 [[preprint](#)], [bibtex](#)]

Concept-Level Explanations

- R Achtibat, M Dreyer, I Eisenbraun, S Bosse, T Wiegand, W Samek, S Lapuschkin. [From "Where" to "What": Towards Human-Understandable Explanations through Concept Relevance Propagation](#)
arXiv:2206.03208, 2022 [[preprint](#)], [bibtex](#)]

Thank you for your attention

<http://www.heatmapping.org>

- ▶ Tutorials
- ▶ Software
- ▶ Online Demos

