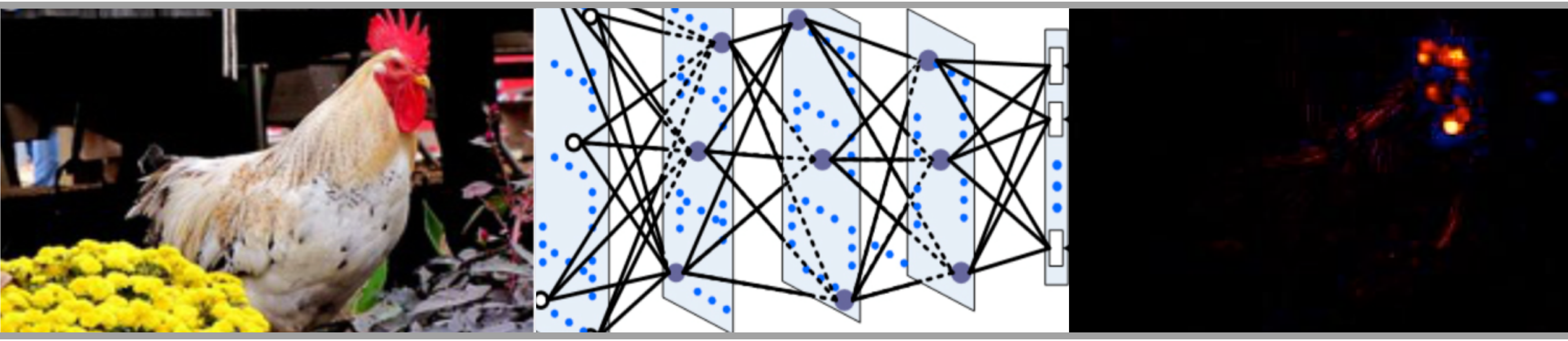
XXAI: eXtending XAI towards Actionable Interpretability

Wojciech Samek
AI Department, Fraunhofer HHI



CVPR'21 Tutorial on Interpretable Machine Learning for Computer Vision

20th June 2021, virtual event

ML Models = Black Boxes ?

Report No. 85-460-1

THE PERCEPTRON
A PERCEIVING AND RECOGNIZING AUTOMATON
(PROJECT PARA)

January, 1957

Prepared by: Frank Rosenblatt
Frank Rosenblatt,
Project Engineer

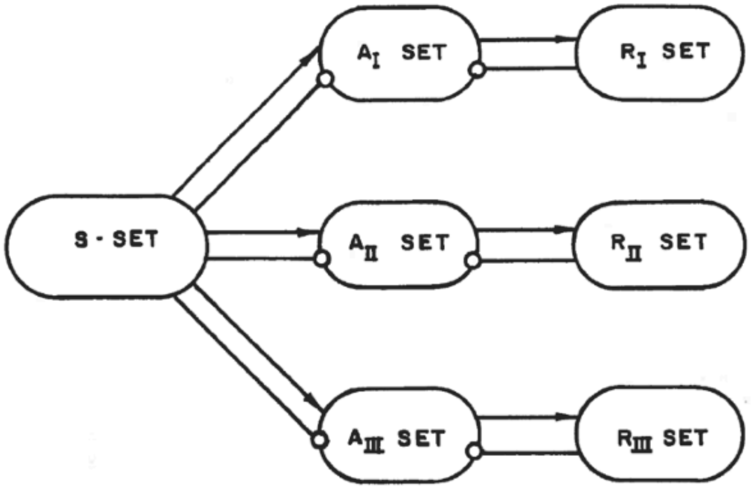
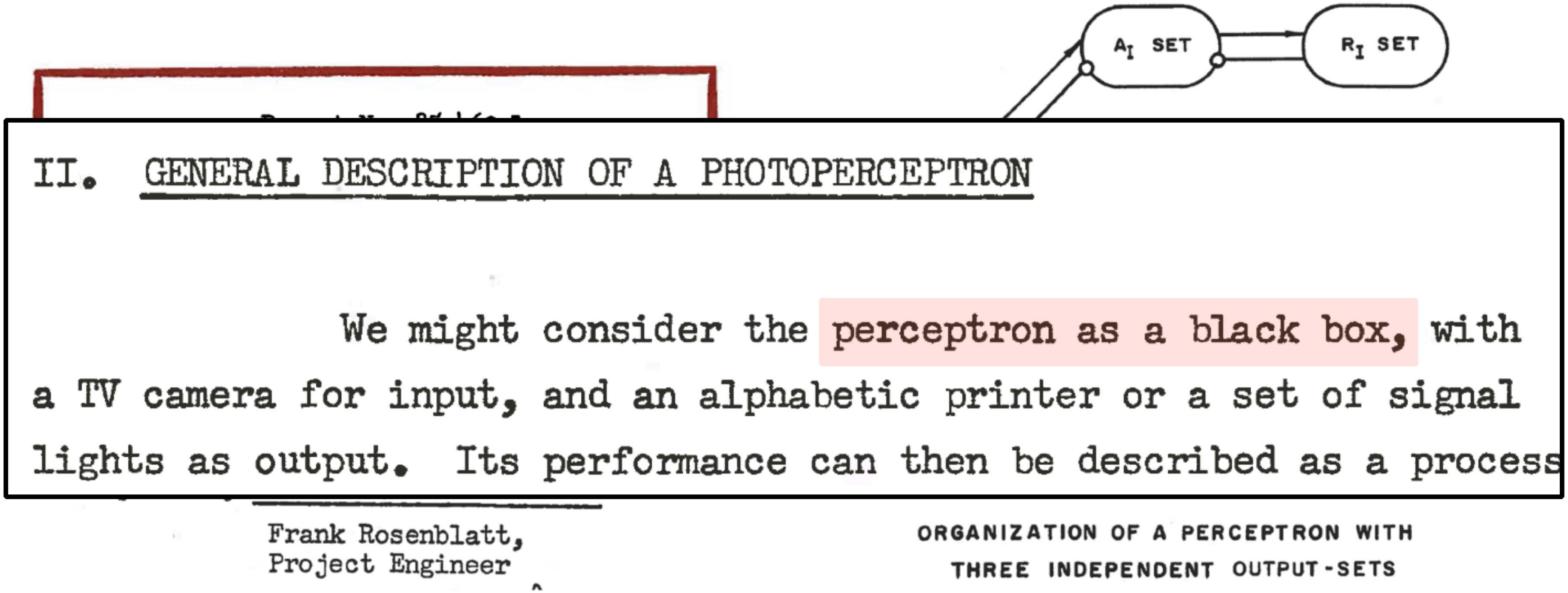


FIGURE 2

ORGANIZATION OF A PERCEPTRON WITH THREE INDEPENDENT OUTPUT-SETS

ML Models = Black Boxes ?



II. GENERAL DESCRIPTION OF A PHOTOPERCEPTRON

We might consider the perceptron as a black box, with a TV camera for input, and an alphabetic printer or a set of signal lights as output. Its performance can then be described as a process

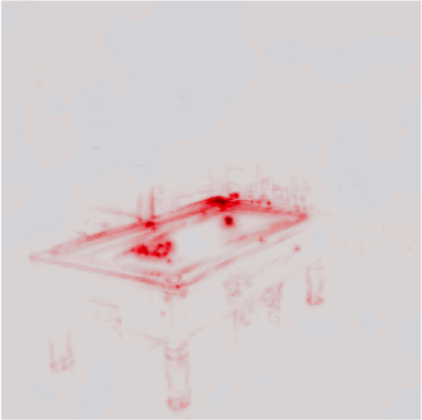
Frank Rosenblatt,
Project Engineer

ORGANIZATION OF A PERCEPTRON WITH THREE INDEPENDENT OUTPUT-SETS

Today: Post-hoc XAI

"why a given image is classified as a pool table"

some pool table



why it is classified as a pool table

Brief History

Visualization of neural networks using saliency maps
NJS **Morch**, U Kjems, LK Hansen… - Proceedings of ICNN …, **1995**

[PDF] How to explain individual classification decisions
D **Baehrens**, T Schroeter, S Harmeling… - The Journal of Machine …, 2010 - jmlr.org

Deep inside convolutional networks: Visualising image classification models and saliency maps
K Simonyan, A Vedaldi, A Zisserman - arXiv preprint arXiv:1312.6034, 2013 - arxiv.org

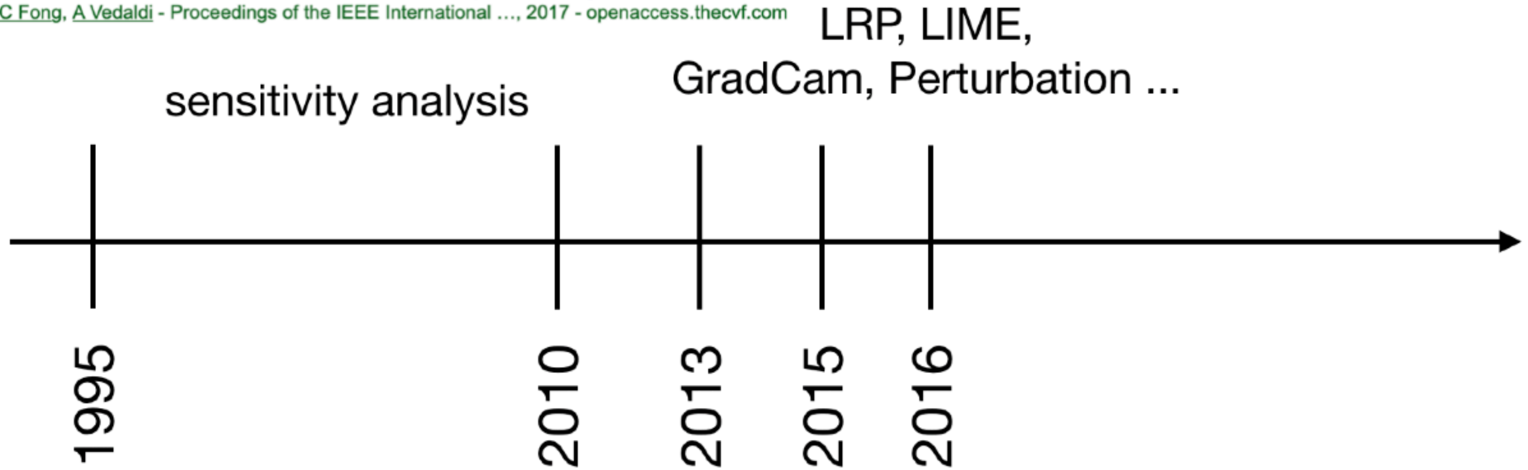
Brief History

[HTML] On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation
S Bach, A Binder, G Montavon, F Klauschen… - PloS one, 2015 - journals.plos.org

" Why should I trust you?" **Explaining** the predictions of any classifier
MT **Ribeiro**, S Singh, C Guestrin - Proceedings of the 22nd ACM …, 2016 - dl.acm.org

Grad-CAM: Why did you say that?
RR Selvaraju, A Das, R Vedantam, M Cogswell… - arXiv preprint arXiv …, 2016 - arxiv.org

Interpretable explanations of black boxes by **meaningful perturbation**
RC Fong, A Vedaldi - Proceedings of the IEEE International …, 2017 - openaccess.thecvf.com



Brief History

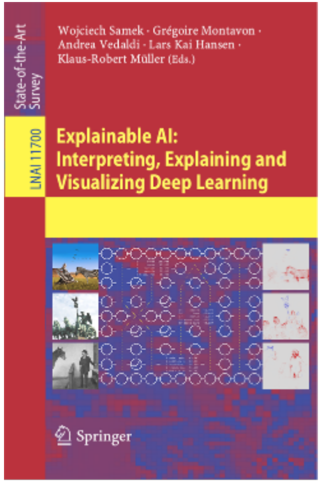
[HTML] Explaining nonlinear classification decisions with **deep taylor** decomposition
G **Montavon**, S Lapuschkin, A Binder, W Samek... - Pattern Recognition, 2017 - Elsevier

A unified approach to interpreting model predictions
SM Lundberg, SI Lee - Advances in neural information processing ..., 2017 - papers.nips.cc

Explaining recurrent neural network predictions in sentiment analysis
L **Arras**, G Montavon, KR Müller, W Samek - arXiv preprint arXiv ..., 2017 - arxiv.org

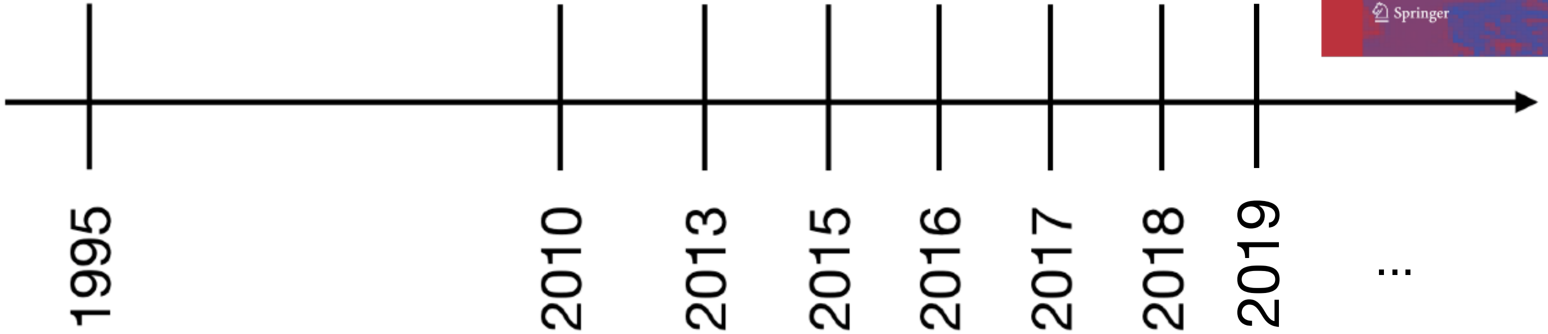
XAI for LSTMs

Theoretical frameworks for XAI

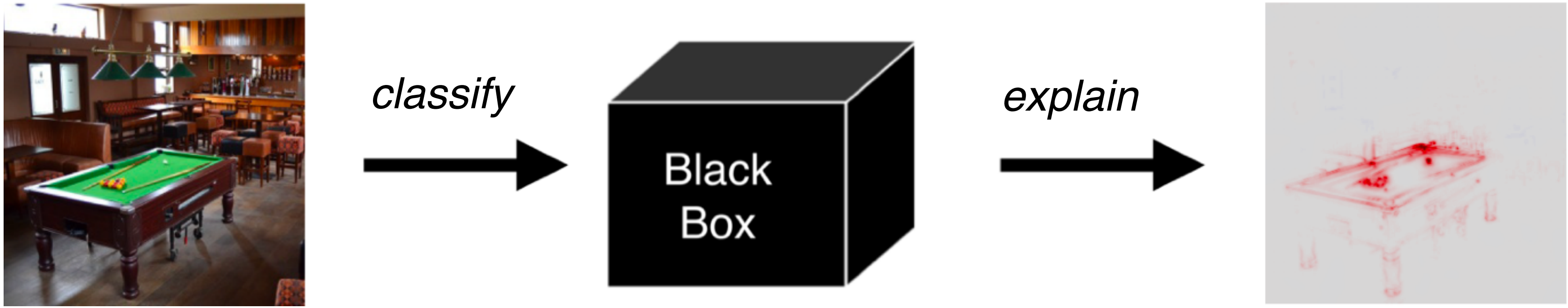


LRP, LIME, GradCam, Perturbation ...

sensitivity analysis

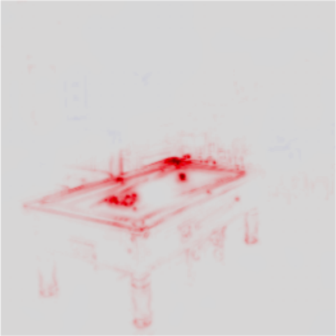


Explain? Yes We Can



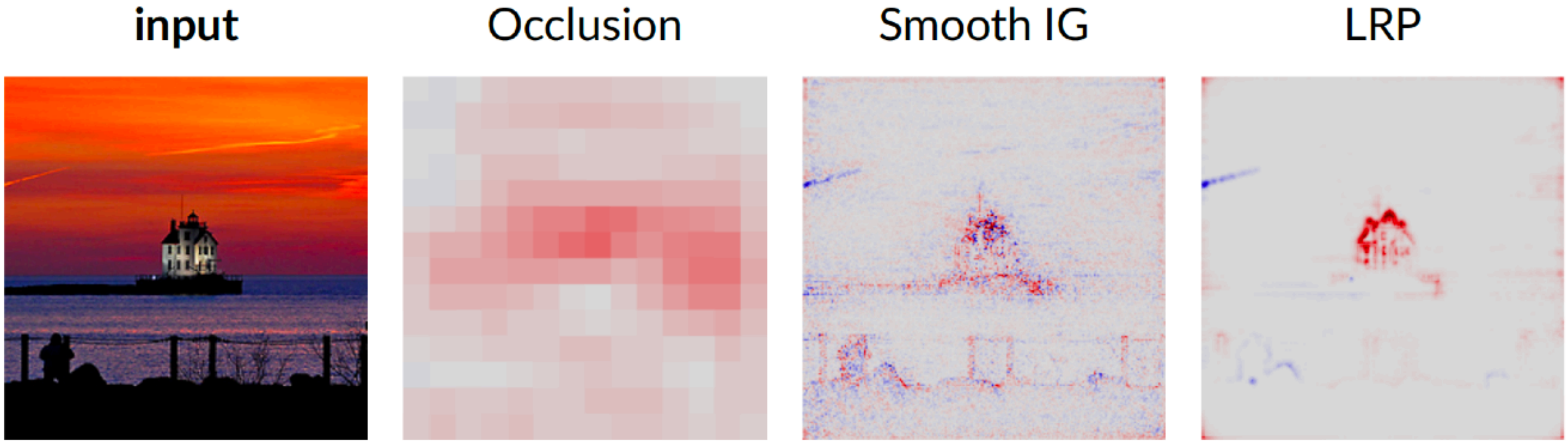
Explain? Yes We Can

And now ?



Are Our Explanations Good Enough?

What are good Explanations ?



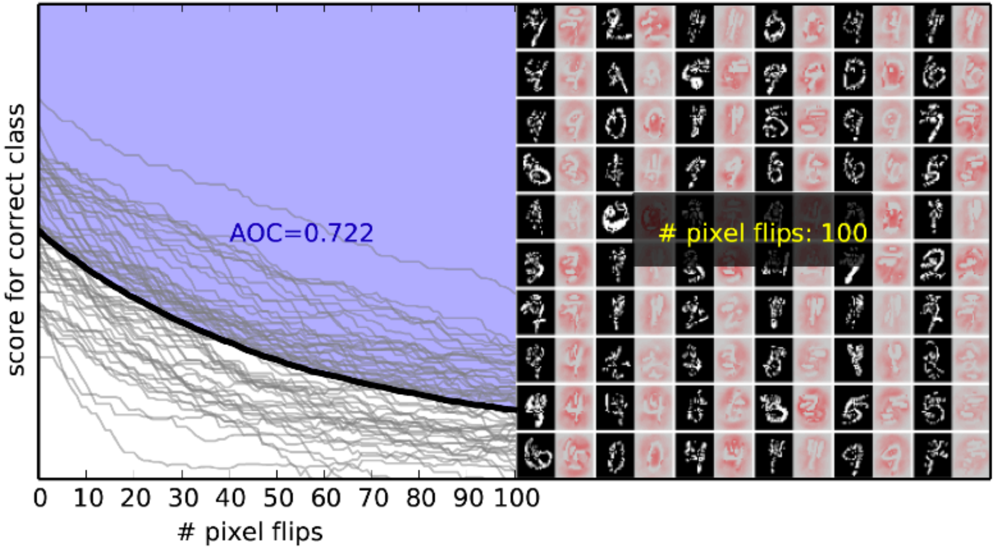
Which explanation technique should be preferred?

Some Desiderata for Explanations

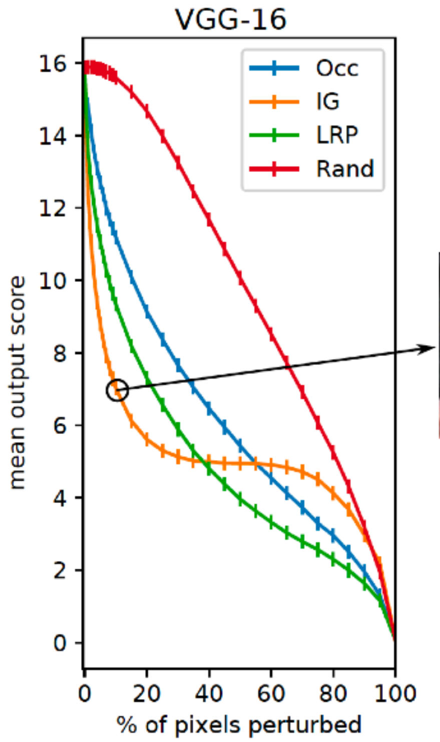
1. **Fidelity:** The explanation should reflect the quantity being explained and not something else.
2. **Understandability:** The explanation must be easily understandable by its receiver.
3. **Sufficiency:** The explanation should provide sufficient information on how the model came up with its prediction.
4. **Low Overhead:** The explanation should not cause the prediction model to become less accurate or less efficient.
5. **Runtime Efficiency:** Explanations should be computable in reasonable time.

Evaluating Fidelity: Pixel-Flipping

- The pixel-flipping procedure [9] destroys pixels from most to least relevant according to the explanation, and keeps track of the neural network output.
- The faster the output decreases, the better the explanation.



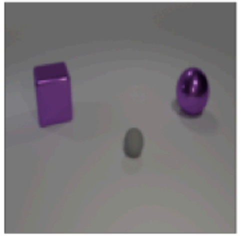
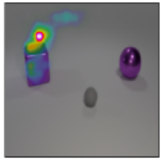
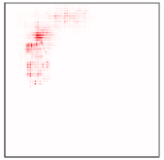
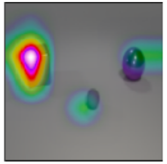
Evaluating Fidelity: Pixel-Flipping



- All explanation methods are more faithful than a random explanation.
- IG is the most faithful for the first few most relevant pixels, and then stagnates.
- Although not detected by VGG-16 anymore, the class-relevant patterns are still there after flipping (e.g. we can still see the dog). Did IG actually explain a *vulnerability* of VGG-16 instead of its typical behavior?

[Samek et al. 2021]

Evaluating Fidelity: Comparison with Ground Truth

What material is the large object that is left of the big purple metallic ball? *metal*			GT Unique First-non-empty
LRP [20]			0.97
SmoothGrad [43]			0.42
Grad-CAM [42]			0.38

[Arras et al. 2020]

Evaluating Sufficiency

- Example of a faithful, understandable, but *insufficient* explanation

  Q: *Why did the classifier predict this image to be a 'lighthouse'?*
  A: *Because the classifier found a lighthouse in the image.*

- Evaluating sufficiency:
  - Is the explanation actionable? (e.g. can we improve a model from the produced explanations).
  - Can we learn something general about the classifier? (e.g. what kind of features it uses).

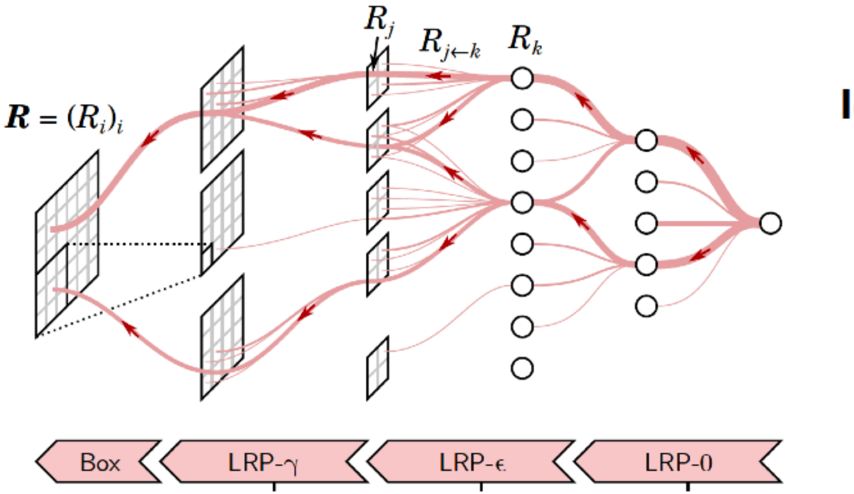
- Is it sufficient to explain a prediction in terms of individual pixels, or should we identify higher-order interactions?

Utilitarian Perspective

Explanations are good if they provide some additional (measurable) advantage.

Layer-wise Relevance Propagation

Layer-wise Relevance Propagation



$$R_j = \sum_k \frac{a_j w_{jk}}{\sum_{0,j} a_j w_{jk}} R_k$$

$$R_j = \sum_k \frac{a_j w_{jk}}{\epsilon_k + \sum_{0,j} a_j w_{jk}} R_k$$

$$R_j = \sum_k \frac{a_j (w_{jk} + \gamma w_{jk}^+)}{\sum_{0,j} a_j (w_{jk} + \gamma w_{jk}^+)} R_k$$

Ideas:

- Use the structure of the neural network to robustly compute relevance scores for the input features.
- Propagate the output of the network backwards by means of propagation rules.
- Propagation rules can be tuned for explanation quality. E.g. sensitive in top-layers, robust in lower layers.

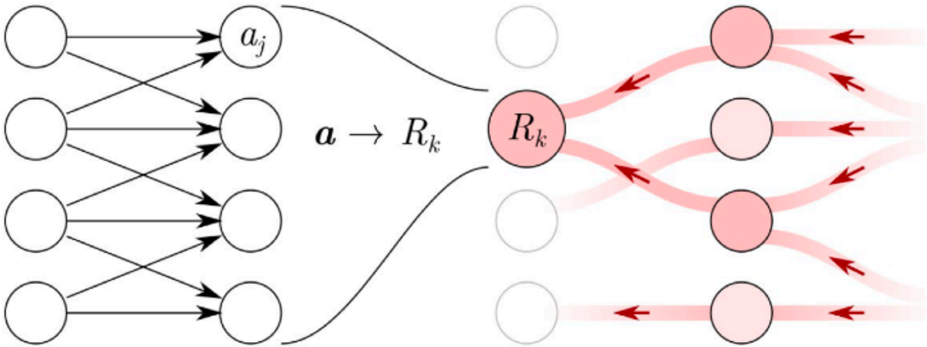
[Bach et al. 2015]

Can LRP be Justified Theoretically?

$$R_j = \sum_k \frac{a_j \cdot \rho(w_{jk})}{\epsilon + \sum_{0,j} a_j \cdot \rho(w_{jk})} R_k$$

Answer: Yes, using the deep Taylor decomposition framework.

Deep Taylor Decomposition

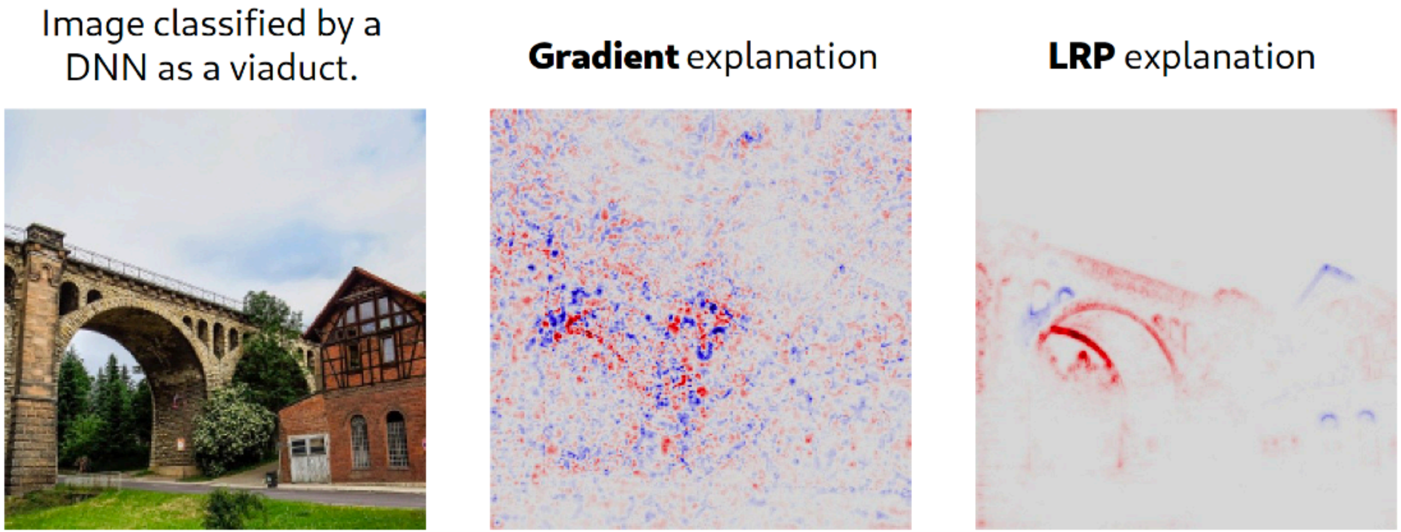


Key idea: Taylor expansions at each layer

$$R_k(\boldsymbol{a}) \approx \widehat{R}_k(\widetilde{\boldsymbol{a}}) + \underbrace{\textstyle\sum_j [\nabla \widehat{R}_k(\widetilde{\boldsymbol{a}})]_j \cdot (a_j - \widetilde{a}_j)}_{\textbf{LRP}} + \ldots$$

[Montavon et al. 2017]

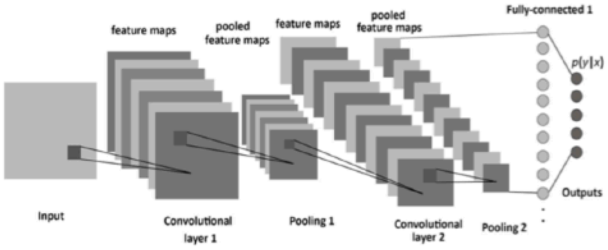
LRP is More Stable than Gradient



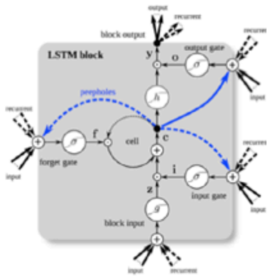
[Samek et al. 2021]

LRP for Different Types of Models

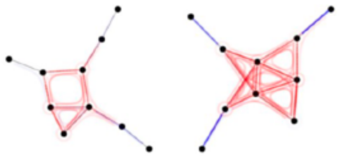
Convolutional NNs (Bach’15, Arras’17 …)



LSTM (Arras’17, Arras’19)



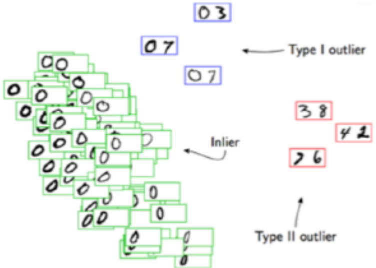
Graph neural networks (GNN-LRP)



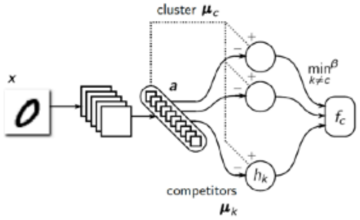
(Schnake'20)

BoW / Fisher Vector models
(Bach’15, Arras’16, Lapuschkin’16 …)

One-class SVM (Kauffmann’18)



Clustering (Kauffmann’19)

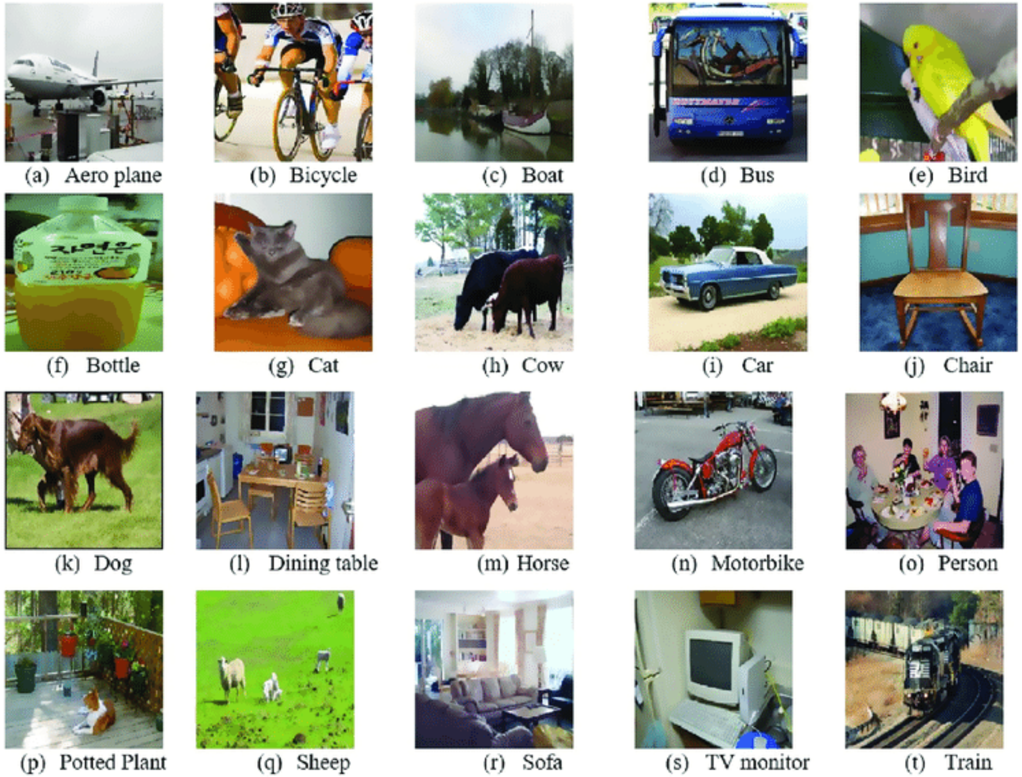


Similarity models (BiLRP)

(Eberle'20)

Towards Actionable Explanations with LRP

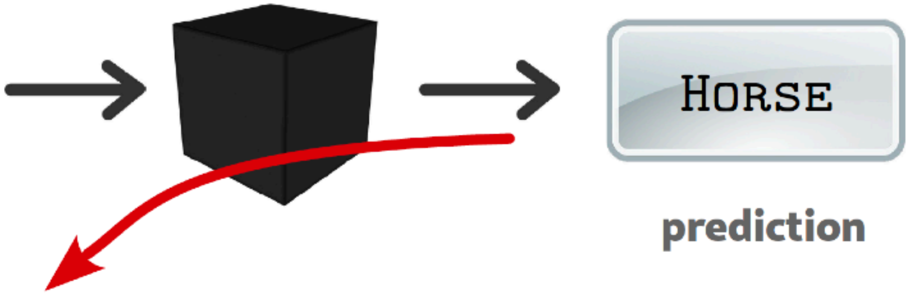
PASCAL VOC Challenge (2005 - 2012)



average precision of the Fisher Vector model on the PascalVOC dataset

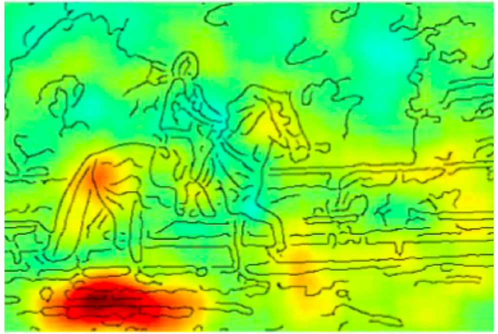
aer	bic	bir	boa	bot
79.08	66.44	45.90	70.88	27.64
bus	car	cat	cha	cow
69.67	80.96	59.92	51.92	47.60
din	dog	hor	mot	per
58.06	42.28	80.45	69.34	85.10
pot	she	sof	tra	tvm
28.62	49.58	49.31	82.71	54.33

Detecting Clever Hans



prediction

+ explanation



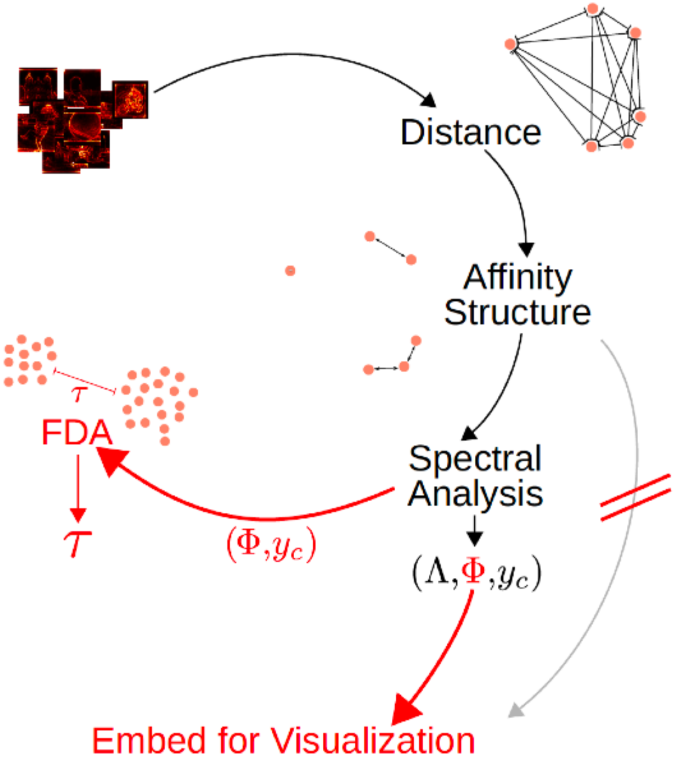
Unexpected: The classifier predicts correctly based on an **artifact** in the data (aka. '**Clever Hans**').

[Lapuschkin et al. 2019]

Detecting Clever Hans

Reason: This strategy works on the current data (many horses images have a copyright tag) → **spurious correlation**.

Automating Clever Hans Detection



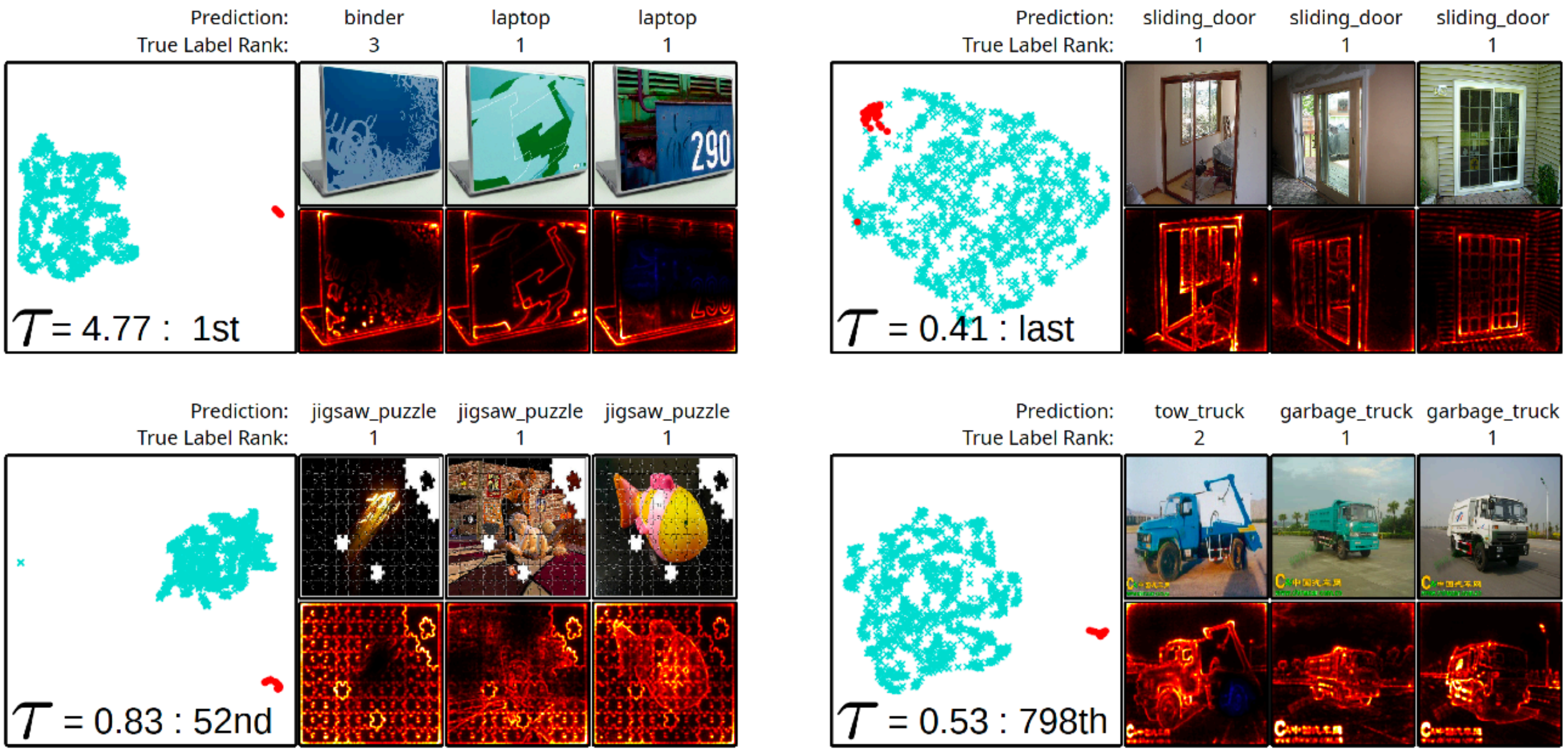
Extending SpRAy from [4]

- Further automating spurious cluster/class discovery by analyzing Φ with FDA[7]
- Visualizing the spectal embedding Φ, instead of affinity structure

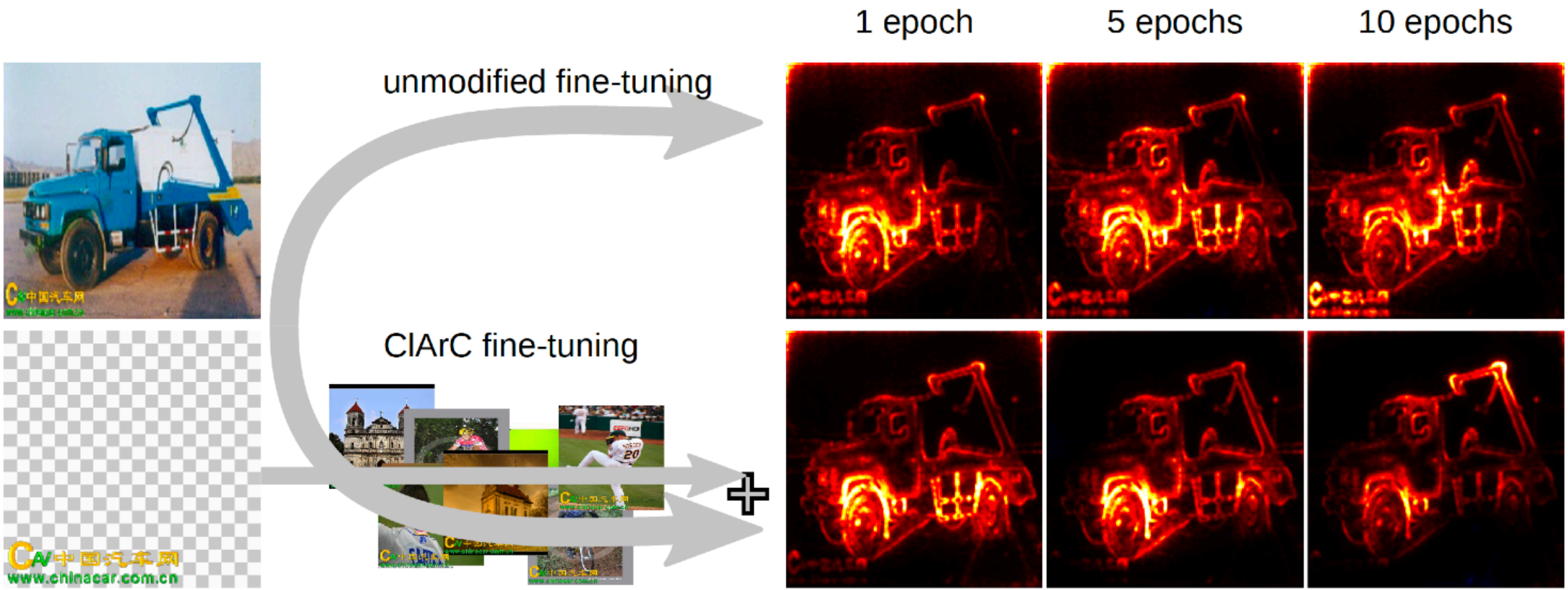
$$J(w) = \frac{w^\top S_b w}{w^\top S_w w}$$

(Anders et al. 2019)

Automating Clever Hans Detection



XAI-Based Model Improvement



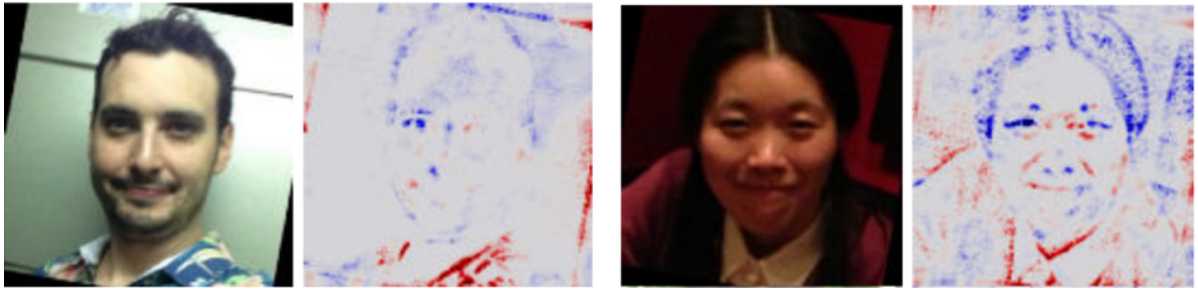
Isolate artefact, add to *other/all* classes, re-train model.

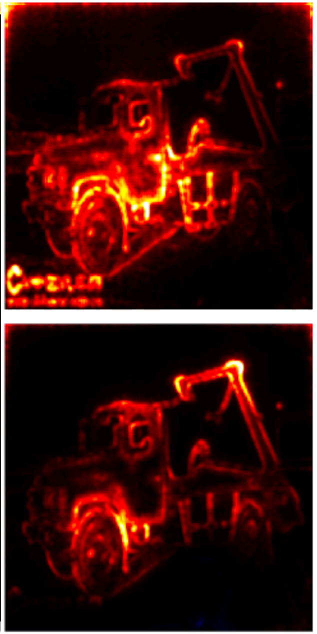
XAI-Based Model Improvement

1 epoch 5 epochs 10 epochs

P-ClArC Projective Class Artifact Compensation

Detect problem in CAV space -> project out (no retraining)

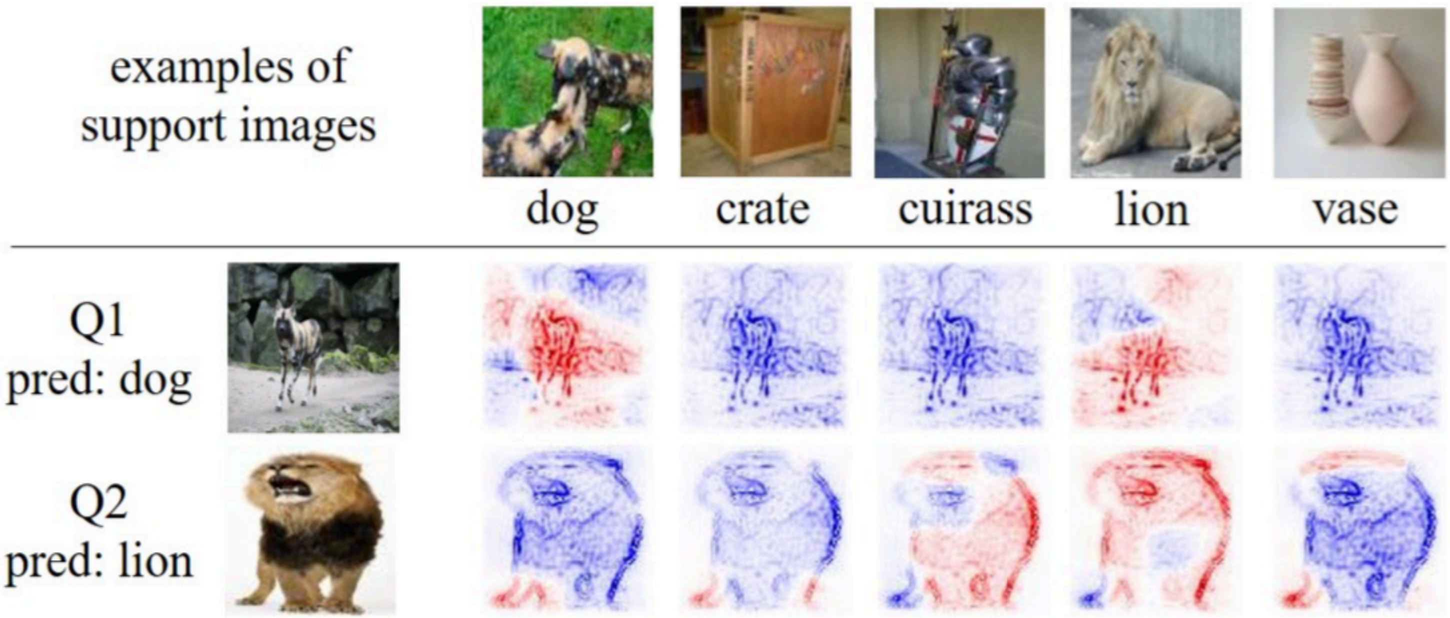




Isolate artefact, add to *other/all* classes, re-train model. [Anders et al. 2019]

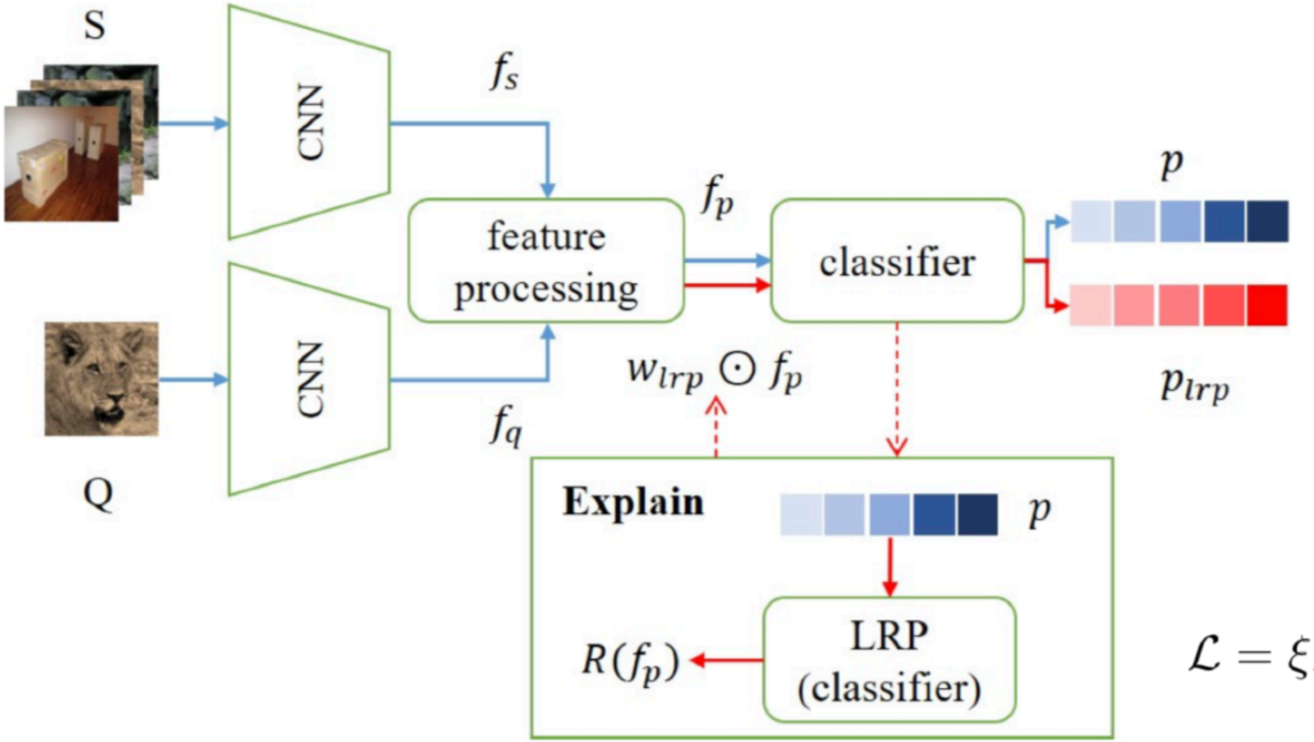
Explanation-Guided Training

Cross-domain few-shot classification task (CD-FSC)



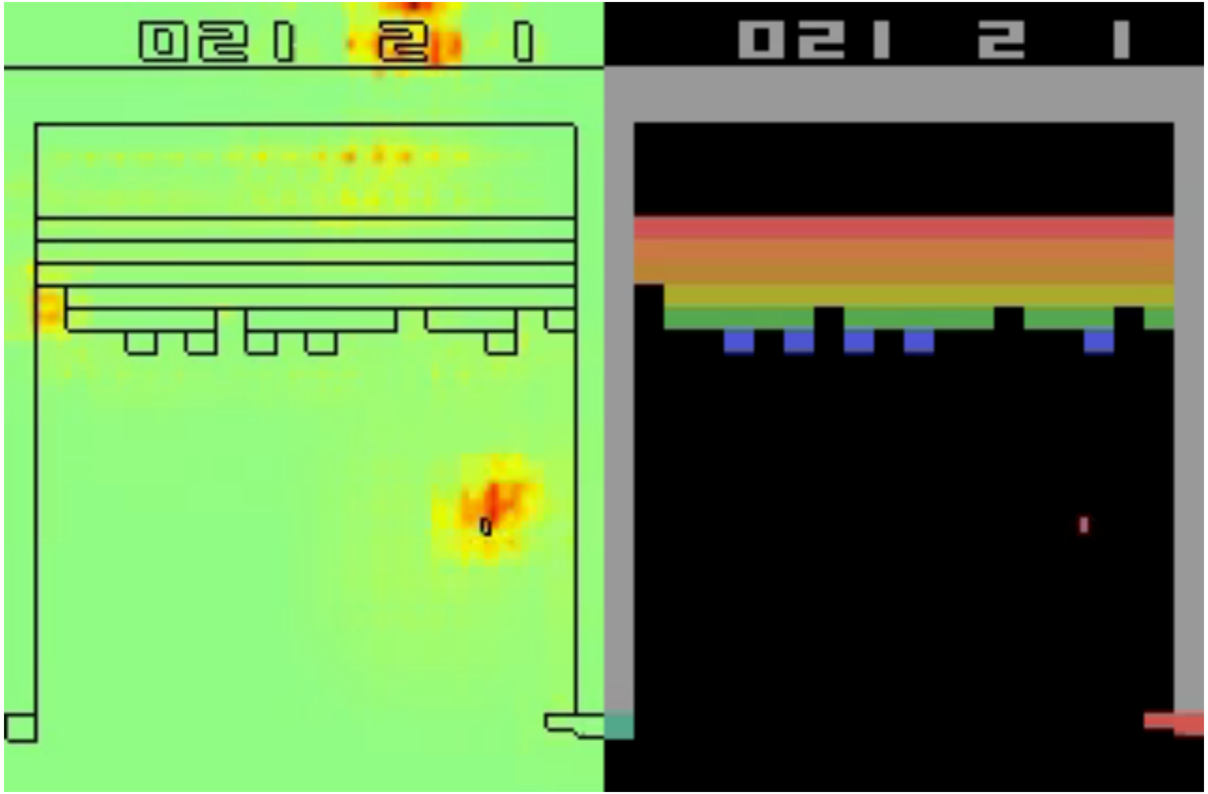
[Sun et al. 2021]

Explanation-Guided Training



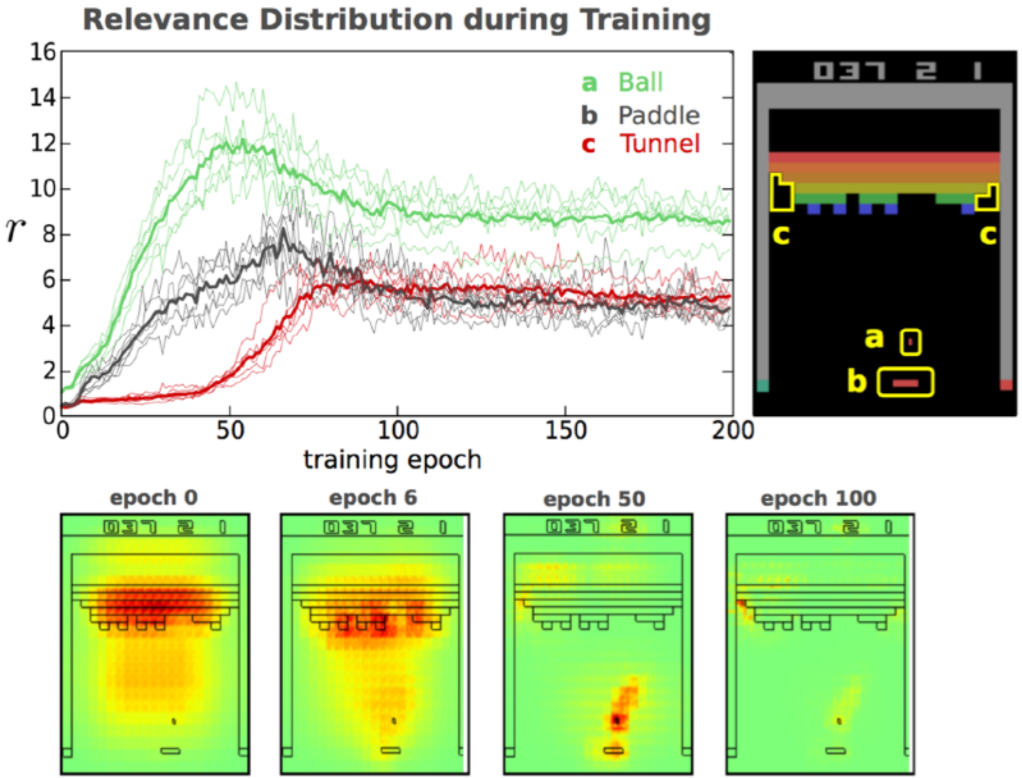
$$w_{lrp} = 1 + R(f_p)$$
$$f_{p-lrp} = w_{lrp} \odot f_p$$
$$\mathcal{L} = \xi \mathcal{L}_{ce}(y, p) + \lambda \mathcal{L}_{ce}(y, p_{lrp})$$

Understanding Learning Behaviour



(Lapuschkin et al., 2019)

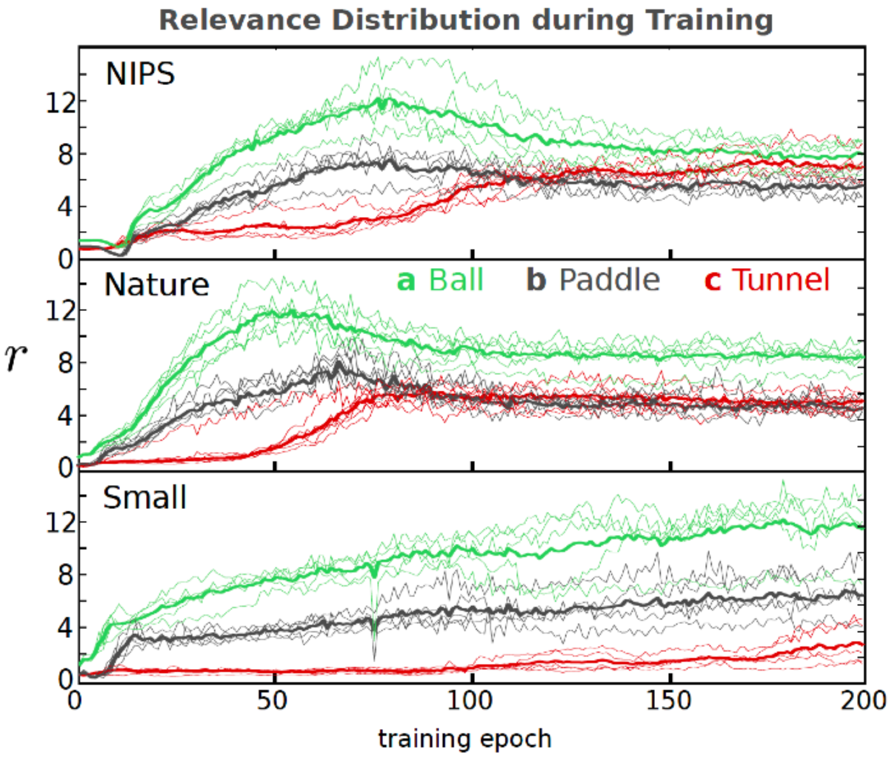
Understanding Learning Behaviour



model learns
1. track the ball
2. focus on paddle
3. focus on the tunnel

nature COMMUNICATIONS Unmasking Clever Hans predictors and assessing what machines really learn

Understanding Learning Behaviour

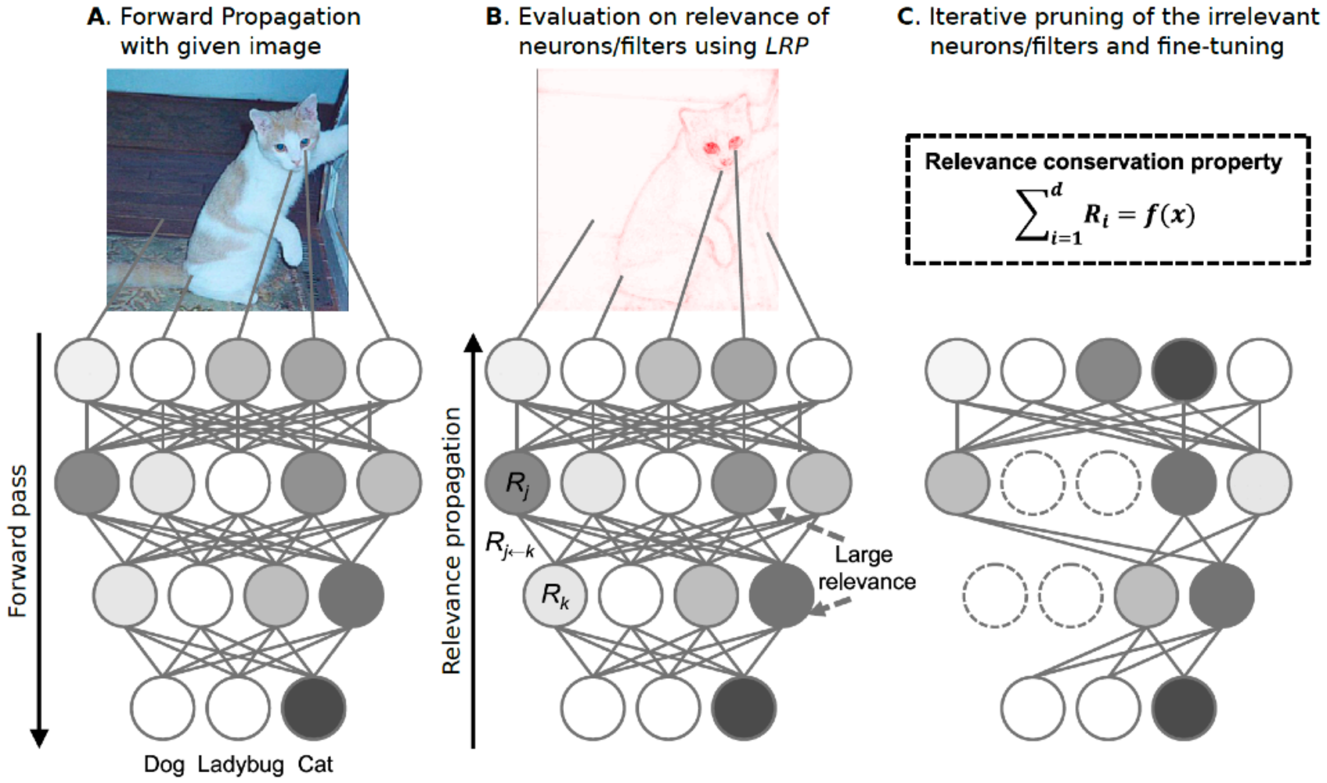


NIPS architecture	Nature architecture
C1 (4×8×8)→(16), [4×4]	C1 (4×8×8)→(32), [4×4]
C2 (16×4×4)→(32), [2×2]	C2 (32×4×4)→(64), [2×2]
	C3 (64×3×3)→(64), [1×1]
F1 (2592)→(256)	F1 (3136)→(512)
F2 (256)→(4)	F2 (512)→(4)

Small architecture
C1 (4×8×8)→(32), [4×4]
C2 (32×4×4)→(64), [2×2]
C3 (64×3×3)→(64), [1×1]
F1 (3136)→(4)

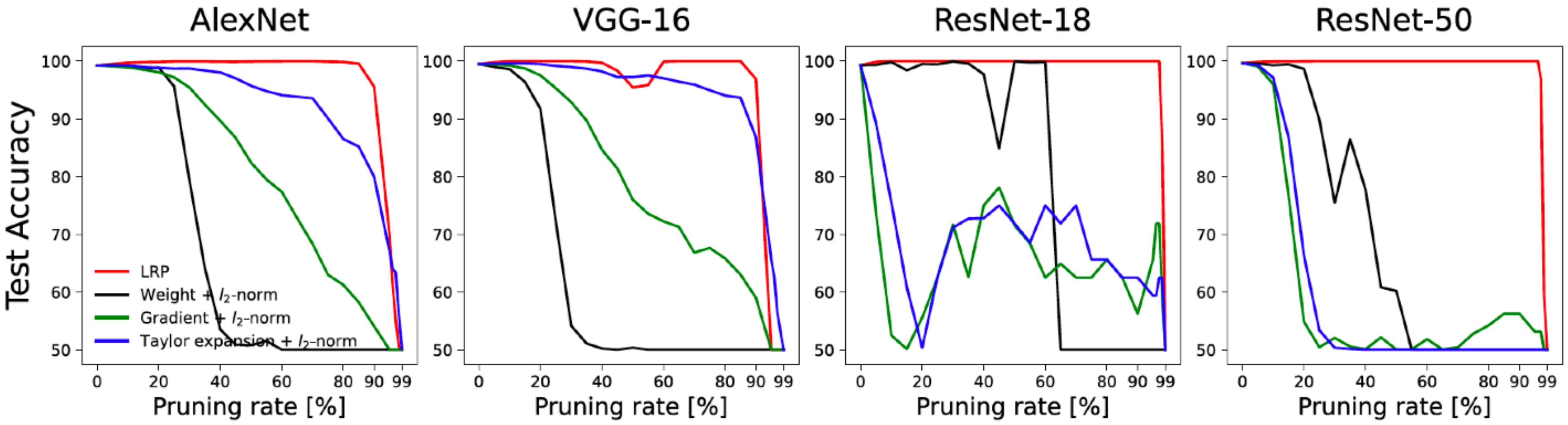
(Lapuschkin et al., 2019)

XAI-Based Pruning



(Yeom et al. 2021)

XAI-Based Pruning

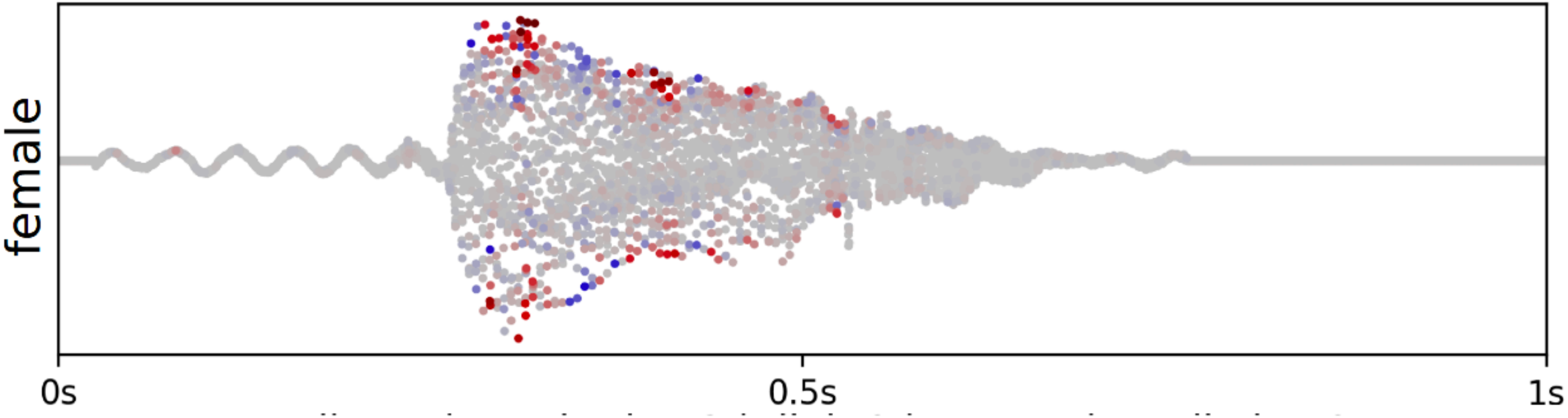


No fine-tuning

only 10 samples per class
(domain adaptation scenario)

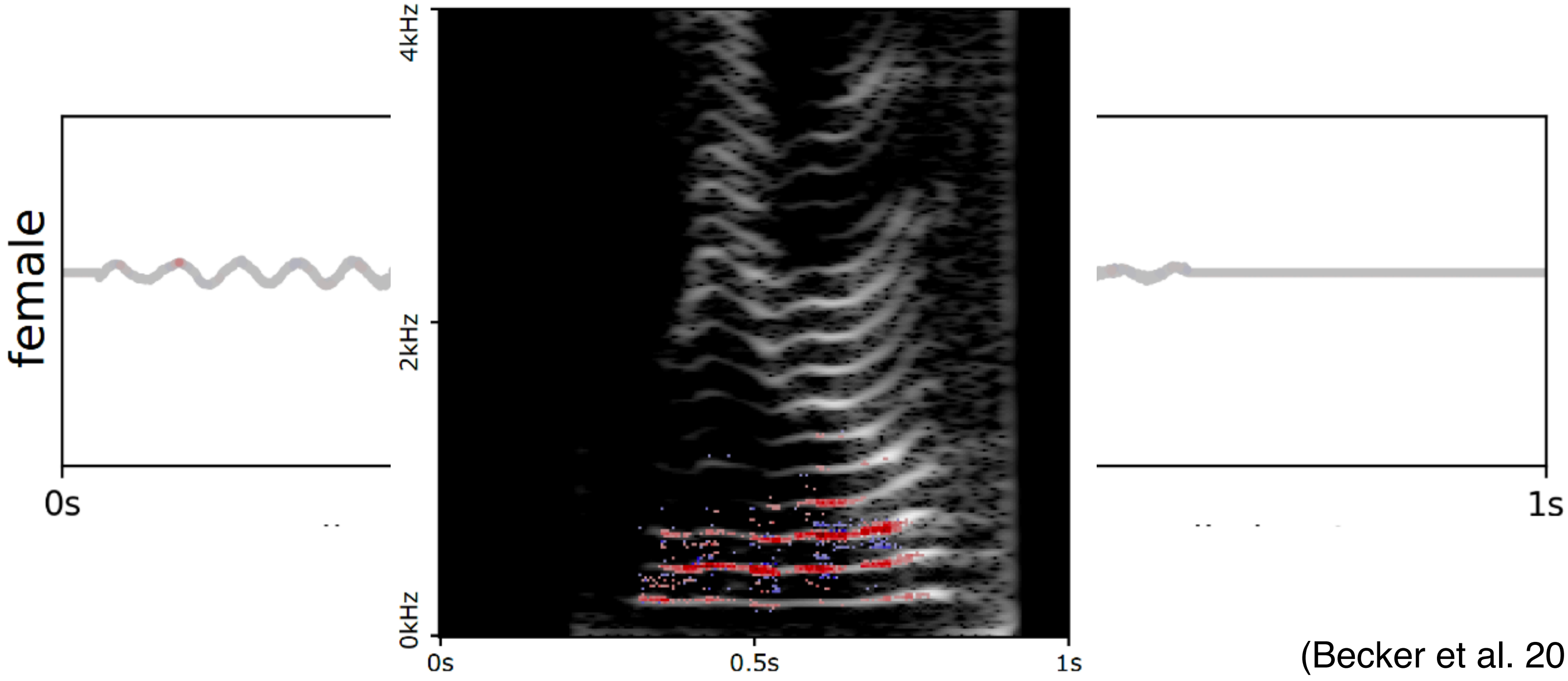
So are we done ?

Is This Explanation Actionable ?



(Becker et al. 2018)

Is This Explanation Actionable ?



(Becker et al. 2018)

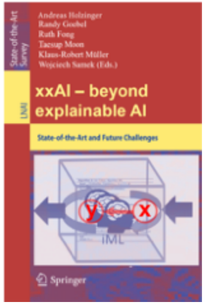
Conclusion

Explanations can be used beyond visualization purposed

Theoretical approaches to XAI exist (e.g. Deep Taylor, Shapley). That allows to compute meaningful explanations, also beyond deep nets.

Explanations need to be actionable (e.g. in scientific applications)

New book to come soon ...

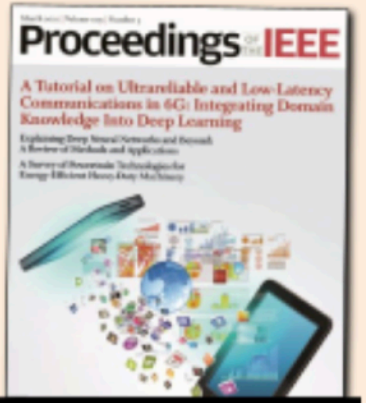


References

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, 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
- 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]
- 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

Methods Papers

- 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), 2019 [preprint, bibtex]
- 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, Lecture Notes in Computer Science, 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

- 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, G Montavon, W Samek, KR Müller. From Clustering to Cluster Explanations via Neural Networks
  arXiv:1906.07633, 2019
- O Eberle, J Büttner, F Kräutli, KR Müller, M Valleriani, G Montavon. Building and Interpreting Deep Similarity Models
  arXiv:2003.05431, 2020
- T Schnake, O Eberle, J Lederer, S Nakajima, K T. Schütt, KR Müller, G Montavon. XAI for Graphs: Explaining Graph Neural Network Predictions by Identifying Relevant Walks
  arXiv:2006.03589, 2020

References

Evaluation of Explanations

- L Arras, A Osman, W Samek. Ground Truth Evaluation of Neural Network Explanations with CLEVR-XAI
  arXiv:2003.07258, 2020 [preprint]
- 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

- 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
  arXiv:1805.11178, 2018
- 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]
- F Horst, D Slijepcevic, S Lapuschkin, AM Raberger, M Zeppelzauer, W Samek, C Breiteneder, WI Schöllhorn, B Horsak. On the Understanding and Interpretation of Machine Learning Predictions in Clinical Gait Analysis Using Explainable Artificial Intelligence
  arXiv:1912.07737, 2020 [preprint]
- 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]

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, 2020 [preprint, bibtex]
- J Sun, S Lapuschkin, W Samek, A Binder. Understanding Image Captioning Models beyond Visualizing Attention arXiv:2001.01037, 2020 [preprint]
- 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 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]

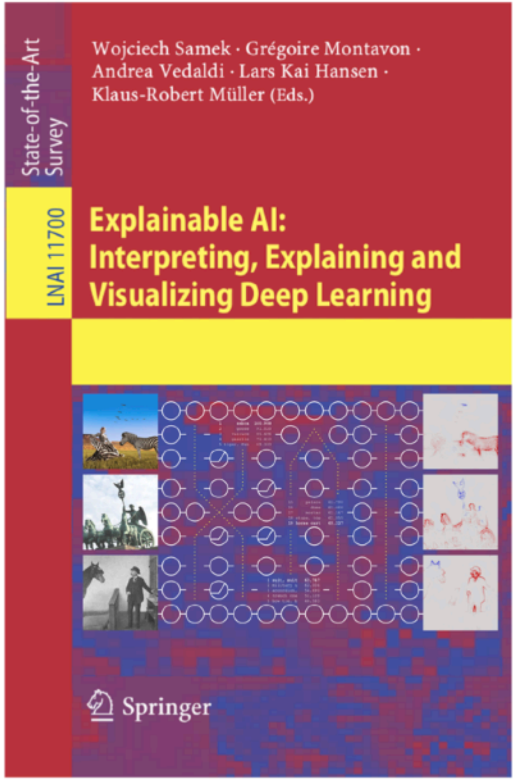
Interpretability and Causality

- 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
  medRxiv:2020.12.10.20225243, 2020

Model Improvement & Training Enhancement

- 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), 2020 [preprint, bibtex]

Our new book is out



Link to the book

https://www.springer.com/gp/book/9783030289539

Organization of the book

Part I Towards AI Transparency

Part II Methods for Interpreting AI Systems

Part III Explaining the Decisions of AI Systems

Part IV Evaluating Interpretability and Explanations

Part V Applications of Explainable AI

—> 22 Chapters

Thank you for your attention

`http://www.heatmapping.org`

- Tutorials
- Software
- Online Demos