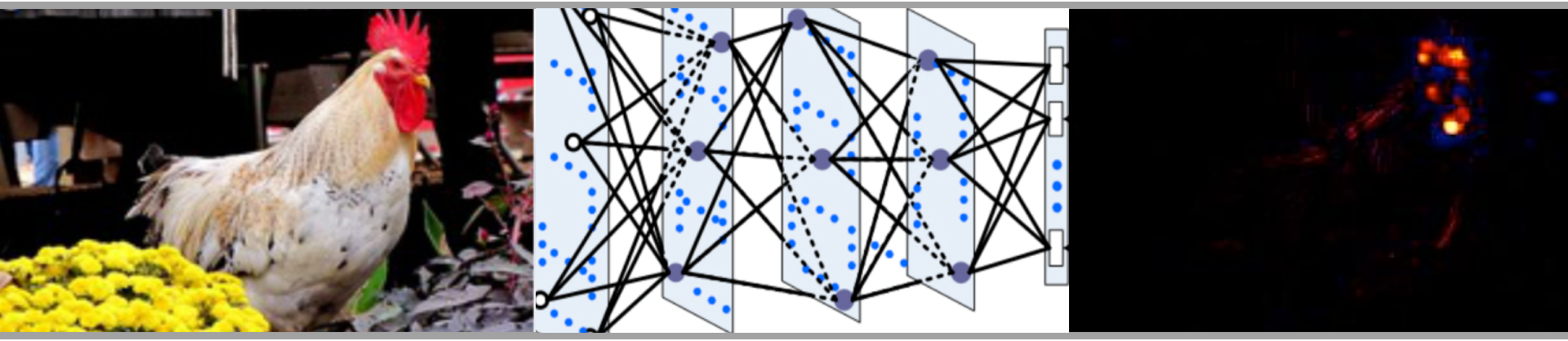
Fraunhofer
Heinrich Hertz Institute

BIFOLD
Berliner Zentrum für MASCHINELLES LERNEN
ellis

Recent Advances in XAI

Wojciech Samek
ML Group, Fraunhofer HHI

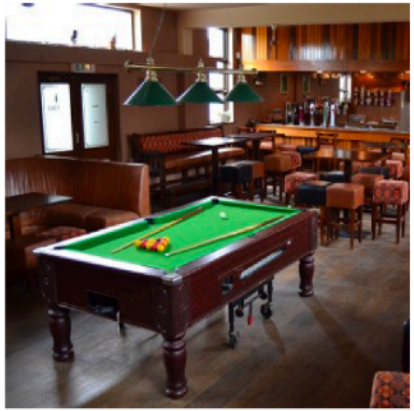


3rd Workshop Advances in Interpretable ML and AI (AIMLAI)

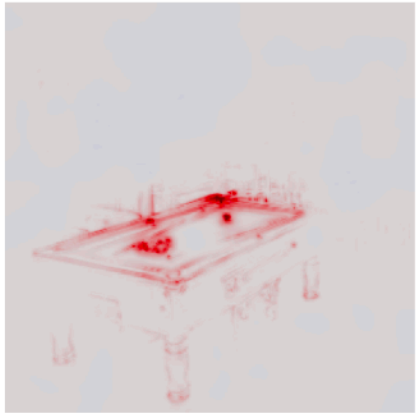
19th October 2020, virtual event

Towards Explainable AI

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



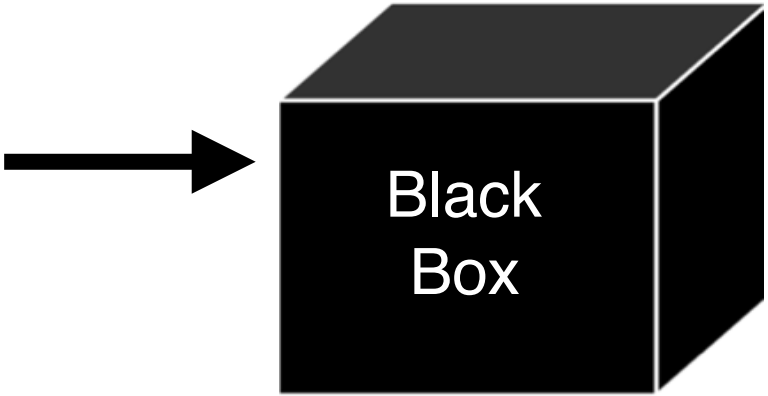
some pool table



why it is classified as a pool table

Why Explanations ?

input x



Rooster

prediction $f(x)$

trust & verification

improve system

learn from the system

legal aspects

Explanation Methods

Perturbation-Based

Occlusion-Based (Zeiler & Fergus 14)

Meaningful Perturbations (Fong & Vedaldi 17)

…

Function-Based

Sensitivity Analysis (Simonyan et al. 14)

(Simple) Taylor Expansions

Gradient x Input (Shrikumar et al. 16)

…

Surrogate- / Sampling-Based

LIME (Ribeiro et al. 16)

SmoothGrad (Smilkov et al. 16)

…

Structure-Based

LRP (Bach et al. 15)

Deep Taylor Decomposition (Montavon et al. 17)

Excitation Backprop (Zhang et al. 16)

…

Layer-wise Relevance Propagation

input x

Black Box

Rooster

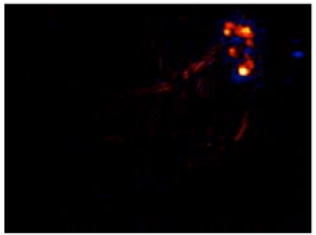
prediction $f(x)$

Idea: Decompose function

$\sum_i R_i = f(x)$

Explain prediction

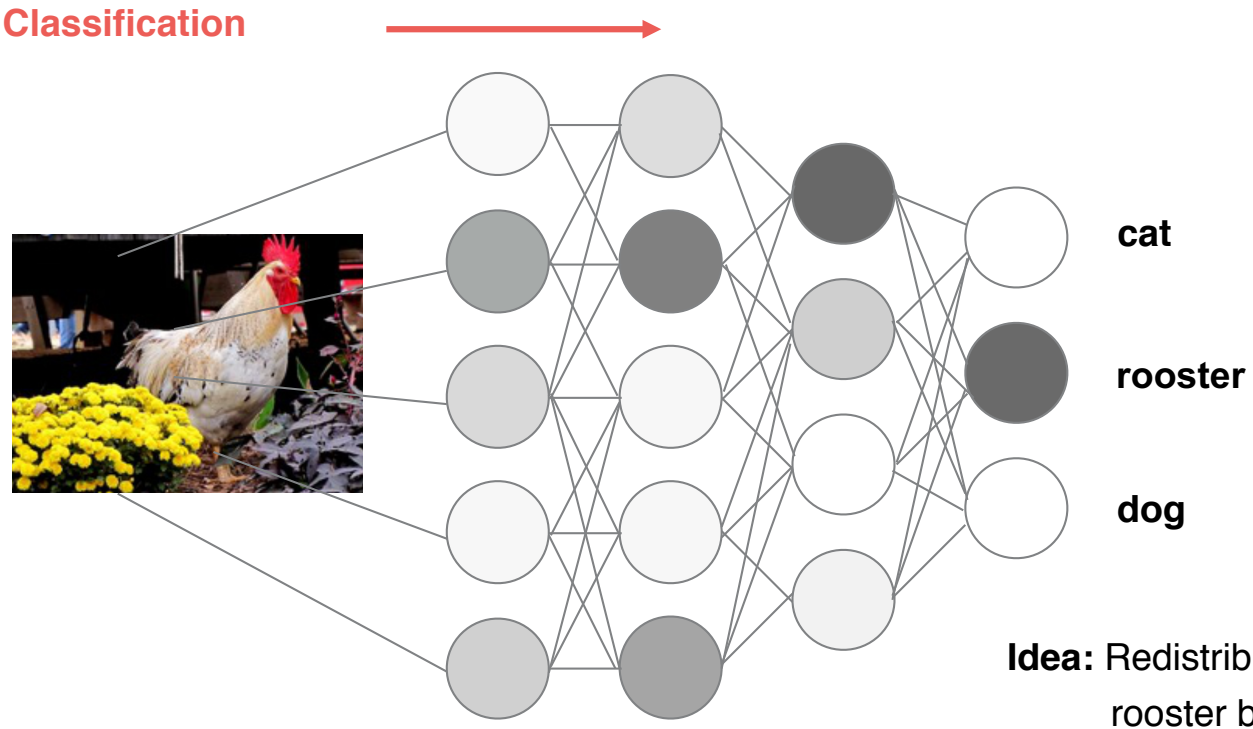
(how much each pixel contributes to prediction)



heatmap

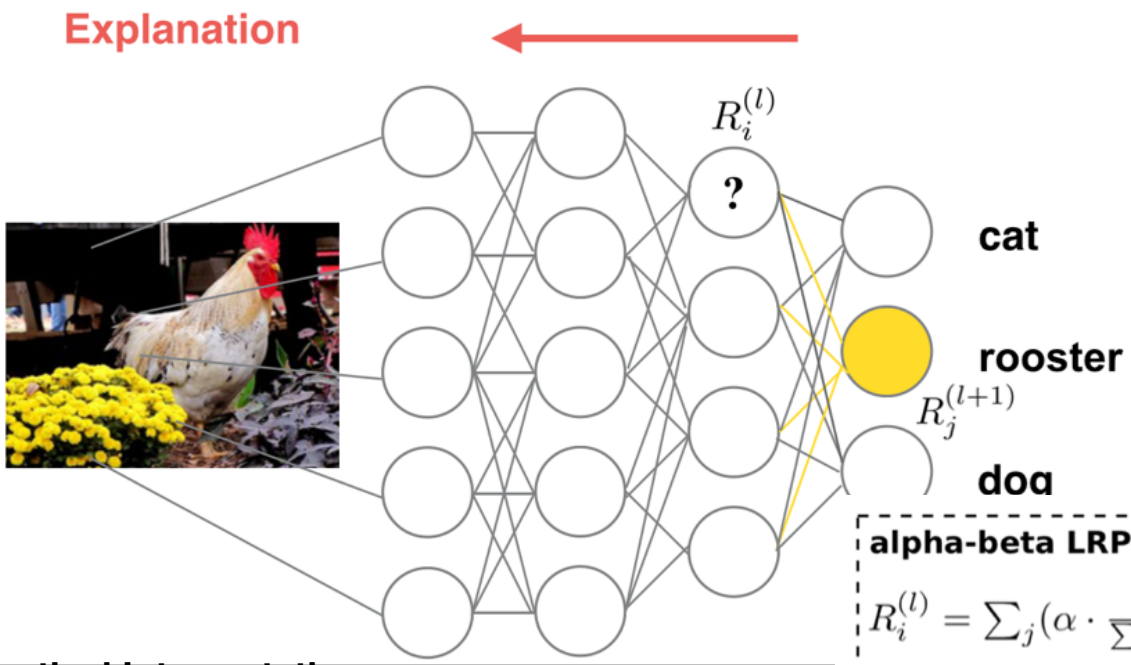
(Bach et al., PLOS ONE, 2015)

Layer-wise Relevance Propagation



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

Layer-wise Relevance Propagation



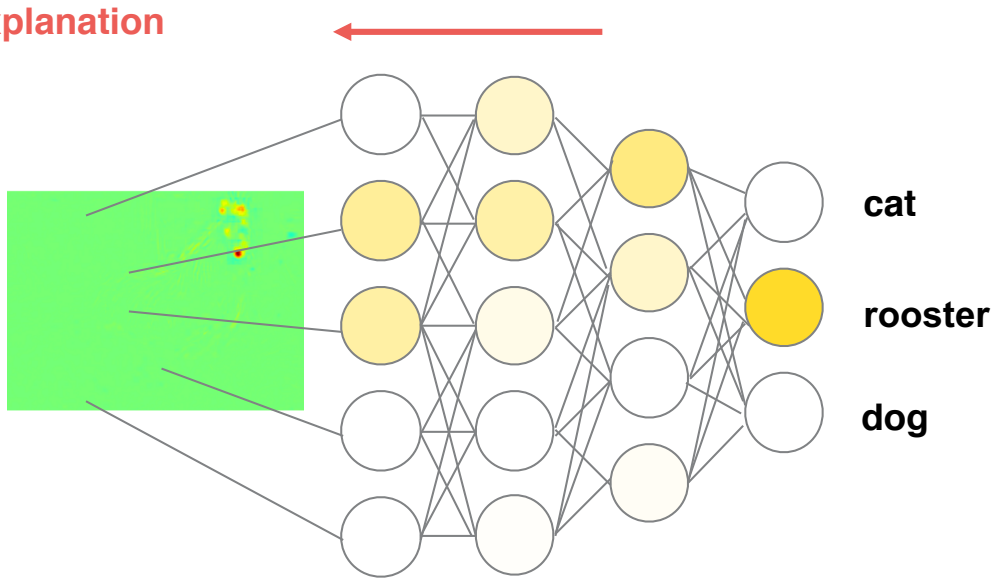
alpha-beta LRP rule (Bach et al. 2015)

$$R_i^{(l)} = \sum_j (\alpha \cdot \frac{(x_i \cdot w_{ij})^+}{\sum_{i'} (x_{i'} \cdot w_{i'j})^+} + \beta \cdot \frac{(x_i \cdot w_{ij})^-}{\sum_{i'} (x_{i'} \cdot w_{i'j})^-}) R_j^{(l+1)}$$

where $\alpha + \beta = 1$

Theoretical Interpretation:
Deep Taylor Decomposition

Layer-wise Relevance Propagation

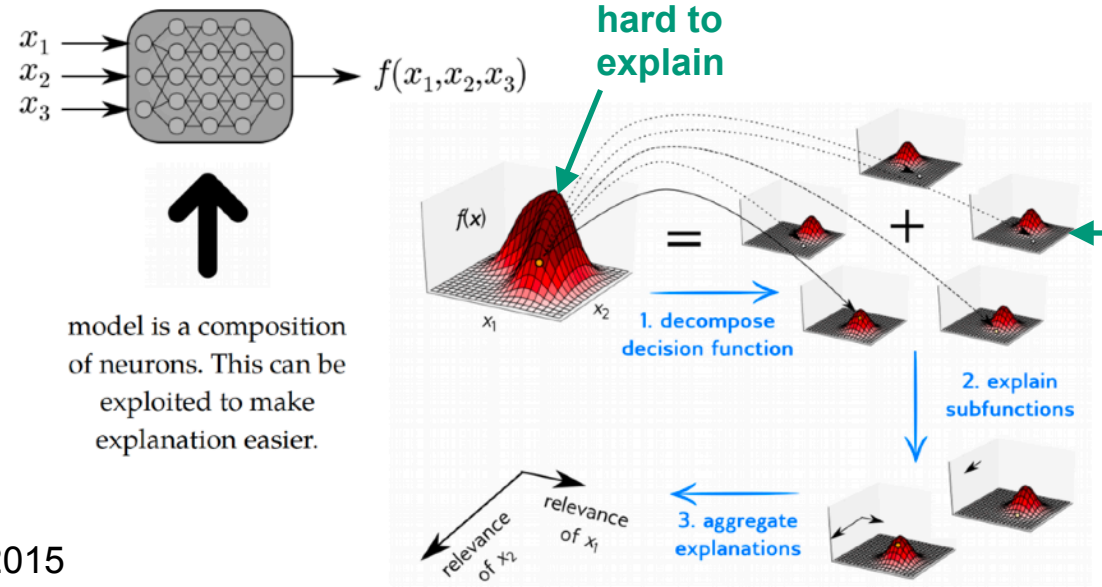


Layer-wise relevance conservation

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

Deep Taylor's View on LRP

LRP's idea: To robustly explain a model, leverage the neural network structure of the decision function.



(Bach et al., 2015
Montavon et al. 2017)

Taylor Decomposition

$$\boldsymbol{x} \mapsto f(\boldsymbol{x})$$

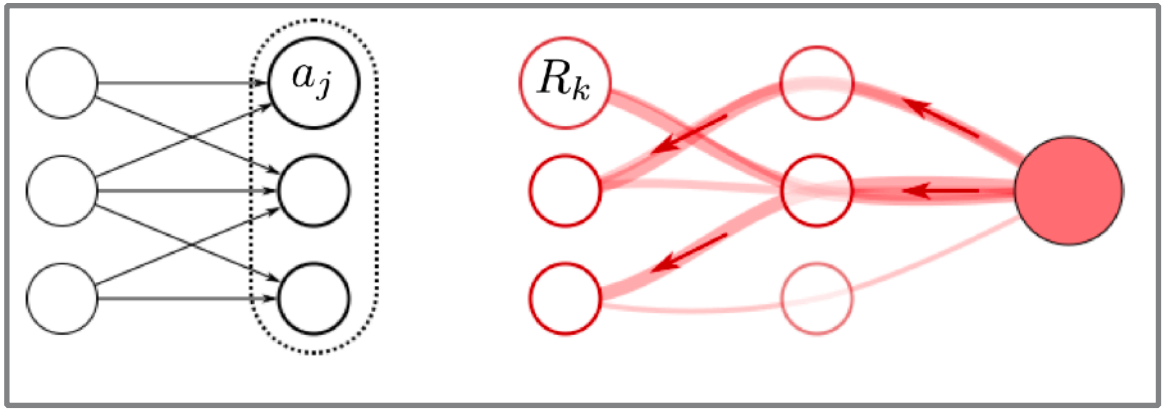
$$f(\boldsymbol{x}) = f(\widetilde{\boldsymbol{x}}) + \sum_{i=1}^{d} [\nabla f(\widetilde{\boldsymbol{x}})]_i \cdot (x_i - \widetilde{x}_i) + \mathcal{O}(\boldsymbol{x}\boldsymbol{x}^\top)$$

Idea: Use Taylor expansion to redistributed relevance from output to input

Limitations:
- difficult to find good root point
- gradient shattering

Deep Taylor Decomposition

$$\boldsymbol{a} \mapsto R_k(\boldsymbol{a})$$



$$R_k(\boldsymbol{a}) = R_k(\widetilde{\boldsymbol{a}}) + \sum_j [\nabla R_k(\widetilde{\boldsymbol{a}})]_j \cdot (a_j - \widetilde{a}_j) + \mathcal{O}(\boldsymbol{a}\boldsymbol{a}^\top)$$

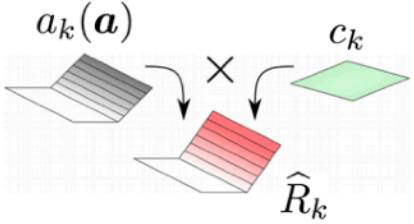
Idea: Use Taylor expansion to redistributed relevance from one layer to another

Advantage:
- easy to find good root point
- no gradient shattering

Deep Taylor Decomposition

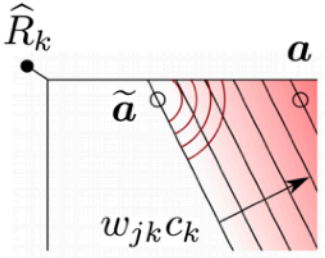
1. Relevance model

$$\widehat{R}_k(\boldsymbol{a}) = \max(0, \textstyle\sum_j a_j\, w_{jk})\, c_k$$



2. Taylor expansion

$$\widehat{R}_k(\boldsymbol{a}) = \widehat{R}_k(\widetilde{\boldsymbol{a}}) + \textstyle\sum_j \underbrace{(a_j - \widetilde{a}_j) \cdot w_{jk}\, c_k}_{R_{j \leftarrow k}} + 0$$



3. Choosing the reference point

$$\widetilde{\boldsymbol{a}}^{(k)} = \mathbf{0} \iff \rho = (\cdot), \epsilon = 0 \qquad \text{(LRP-0)}$$

$$\widetilde{\boldsymbol{a}}^{(k)} = \boldsymbol{a} - t \cdot \boldsymbol{a} \iff \rho = (\cdot), \epsilon = (t^{-1} - 1) \cdot a_k \qquad \text{(LRP-}\epsilon\text{)}$$

$$\widetilde{\boldsymbol{a}}^{(k)} = \boldsymbol{a} - t \cdot \boldsymbol{a} \odot \mathbf{1}_{\boldsymbol{w}_k \succ \mathbf{0}} \iff \rho = \max(0, \cdot) \qquad \text{(LRP-}\gamma\text{)}$$

(Montavon et al., 2017)

Deep Taylor Decomposition

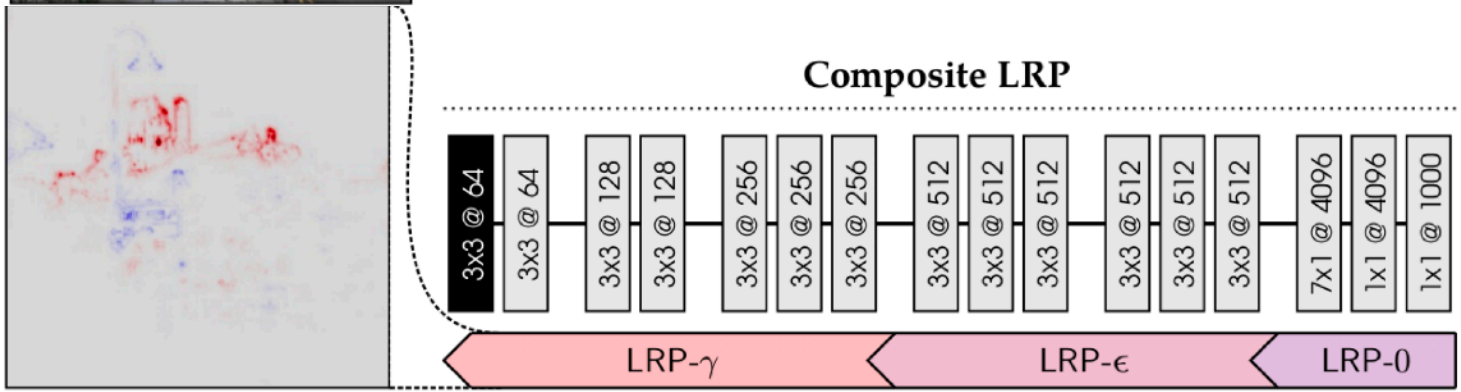
Name	Formula	Usage	DTD
LRP-0 [7]	$R_j = \sum_k \frac{a_j w_{jk}}{\sum_{0,j} a_j w_{jk}} R_k$	upper layers	✓
LRP-ϵ [7]	$R_j = \sum_k \frac{a_j w_{jk}}{\epsilon + \sum_{0,j} a_j w_{jk}} R_k$	middle layers	✓
LRP-γ	$R_j = \sum_k \frac{a_j (w_{jk} + \gamma w_{jk}^+)}{\sum_{0,j} a_j (w_{jk} + \gamma w_{jk}^+)} R_k$	lower layers	✓
LRP-$\alpha\beta$ [7]	$R_j = \sum_k \left(\alpha \frac{(a_j w_{jk})^+}{\sum_{0,j} (a_j w_{jk})^+} - \beta \frac{(a_j w_{jk})^-}{\sum_{0,j} (a_j w_{jk})^-} \right) R_k$	lower layers	$\times^\star$
flat [30]	$R_j = \sum_k \frac{1}{\sum_j 1} R_k$	lower layers	$\times$
w^2-rule [36]	$R_i = \sum_j \frac{w_{ij}^2}{\sum_i w_{ij}^2} R_j$	first layer ($\mathbb{R}^d$)	✓
$z^{\mathcal{B}}$-rule [36]	$R_i = \sum_j \frac{x_i w_{ij} - l_i w_{ij}^+ - h_i w_{ij}^-}{\sum_i x_i w_{ij} - l_i w_{ij}^+ - h_i w_{ij}^-} R_j$	first layer (pixels)	✓

(* DTD interpretation only for the case $\alpha = 1, \beta = 0$.)

Best Practice for LRP

Principle: Explain each layer type (input, conv., fully connected layer) with the optimal rule according to DTD.

(Montavon et al., 2019)
(Kohlbrenner et al., 2019)



Evaluating Explanations

Perturbation Analysis
[Bach'15, Samek'17, Arras'17, …]

Pointing Game
[Zhang'16]

Using Axioms
[Montavon'17, Sundararajan'17, Lundberg'17, …]

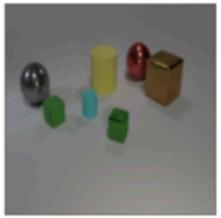
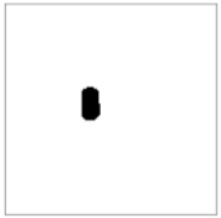
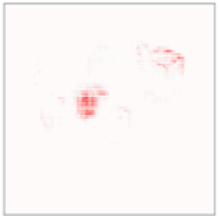
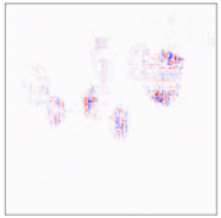
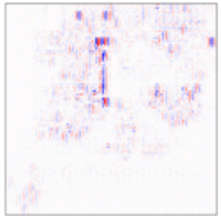
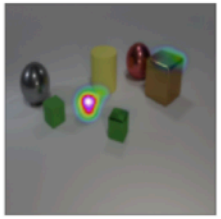
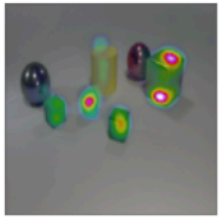
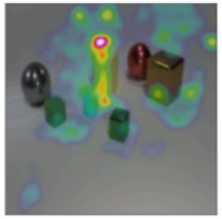
Task Specific Evaluation
[Poerner'18]

Solve other Tasks
[Arras'17, Arjona-Medina'18, …]

Using Ground Truth
[Arras'19, Osman'20]

Human Judgement
[Ribeiro'16, Nguyen'18 …]

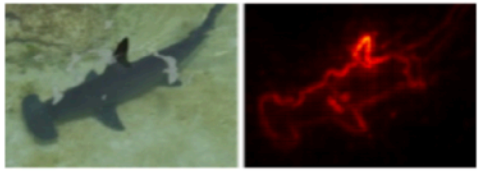
Evaluating Explanations

Question, Answer	Image	One Object Mask	All Objects Mask
The cyan rubber thing is what size? *small*			
Method	LRP	IG	GI
raw heatmap			
overlayed heatmap			

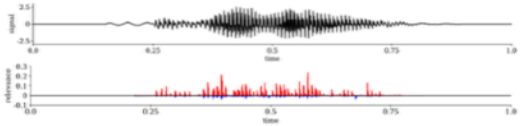
Using Ground Truth
[Osman'20]

LRP Applied to Different Problems

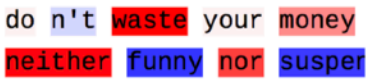
General Images (Bach' 15, Lapuschkin'16)



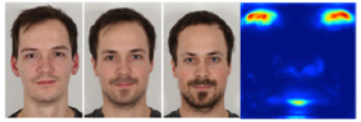
Speech (Becker'18)



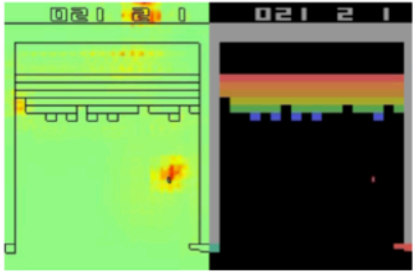
Text Analysis (Arras'16 &17)



Morphing Attacks (Seibold'18)



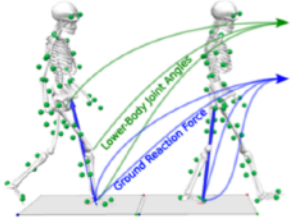
Games (Lapuschkin'19)



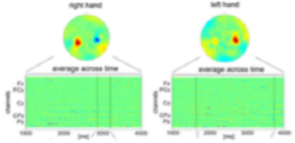
VQA (Samek'19)

Video (Anders'19)

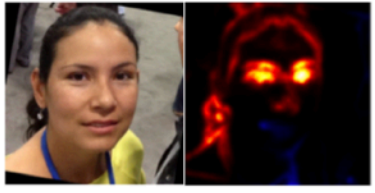
Gait Patterns (Horst'19)



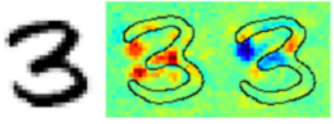
EEG (Sturm'16)



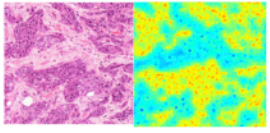
Faces (Lapuschkin'17)



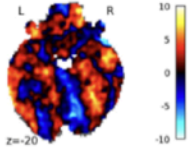
Digits (Bach' 15)



Histopathology (Hägele'19)

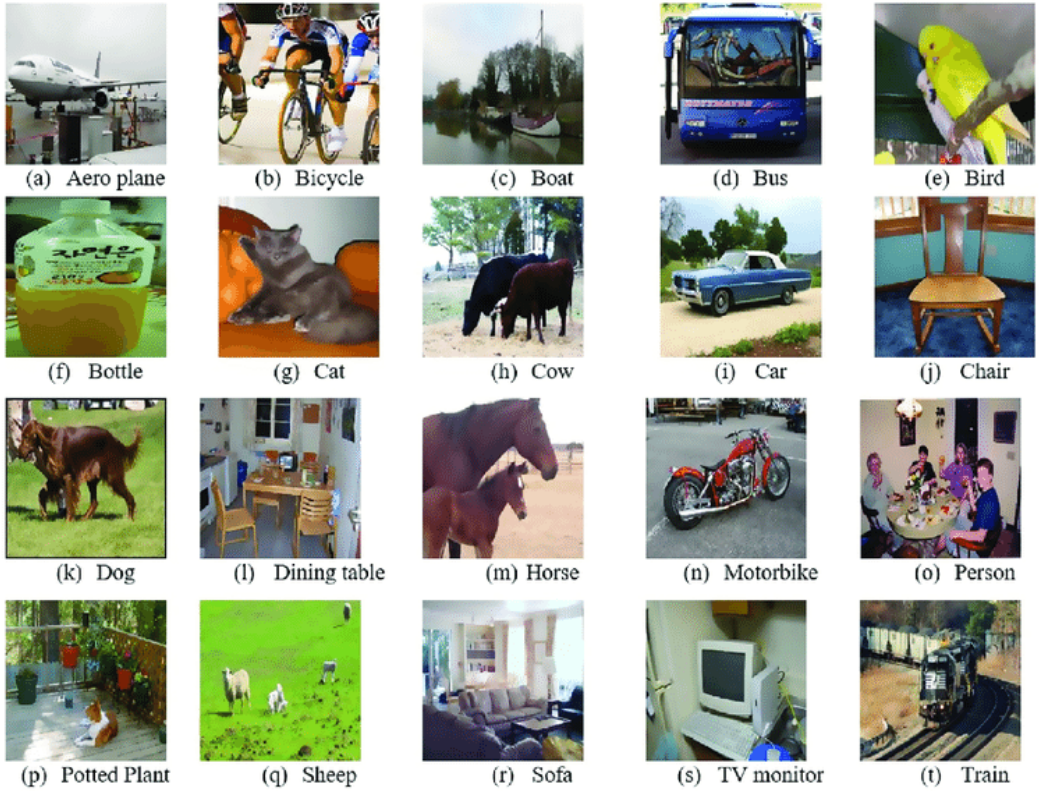


fMRI (Thomas'18)



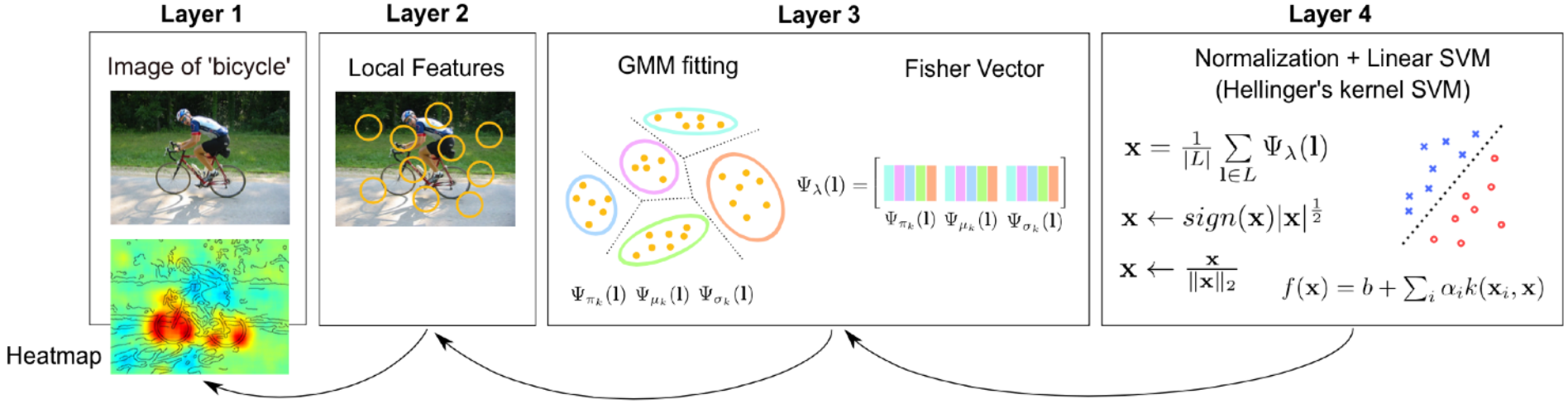
Beyond Deep Classifiers: Explaining Other Models

PASCAL VOC Challenge (2005 - 2012)



	mean
SRN+ [?]	88.8
SFA_NET [?]	87.5
SE [?]	86.5
LIG_DCNN_FEAT_ALL [?]	85.4
S&P_OverFeast_Fast_Bayes [?]	82.8
NUSPSL_CTX_GPM_SCM [?]	82.2
BCE_loss [?]	82.1
Resnet [?]	80.7
CNN_SIGMOID [?]	79.7
NUSPSL_CTX_GPM [?]	78.6
NUS_Context_SVM [?]	78.3
NLPR_PLS_SSVW [?]	78.3
Semi-Semantic Visual Words & Partial Least Squares [?]	78.3
Bayes_Ridge_CNN [?]	77.0
NUSPSL_CTX_GPM_SVM [?]	76.7
Bayes_Ridge_Deep [?]	74.7
CVC_UVA_UNITN [?]	74.3
UvA_UNITN_MostTellingMonkey [?]	73.4
CNNsSVM [?]	72.2
CVC_CLS [?]	71.0
MSRA_USTC_HIGH_ORDER_SVM [?]	70.5
MSRA_USTC_PATCH [?]	70.2
ITI_FK_FUSED_GRAY-RGB-HSV-OP-SIFT [?]	67.1
LIRIS_CLSDET [?]	66.8
ITI_FK_BS_GRAYSIFT [?]	63.2
BPACAD_COMB_LF_AK_WK [?]	61.4
NLPR_IVA_SVM_BOWDect_Convolution [?]	61.1

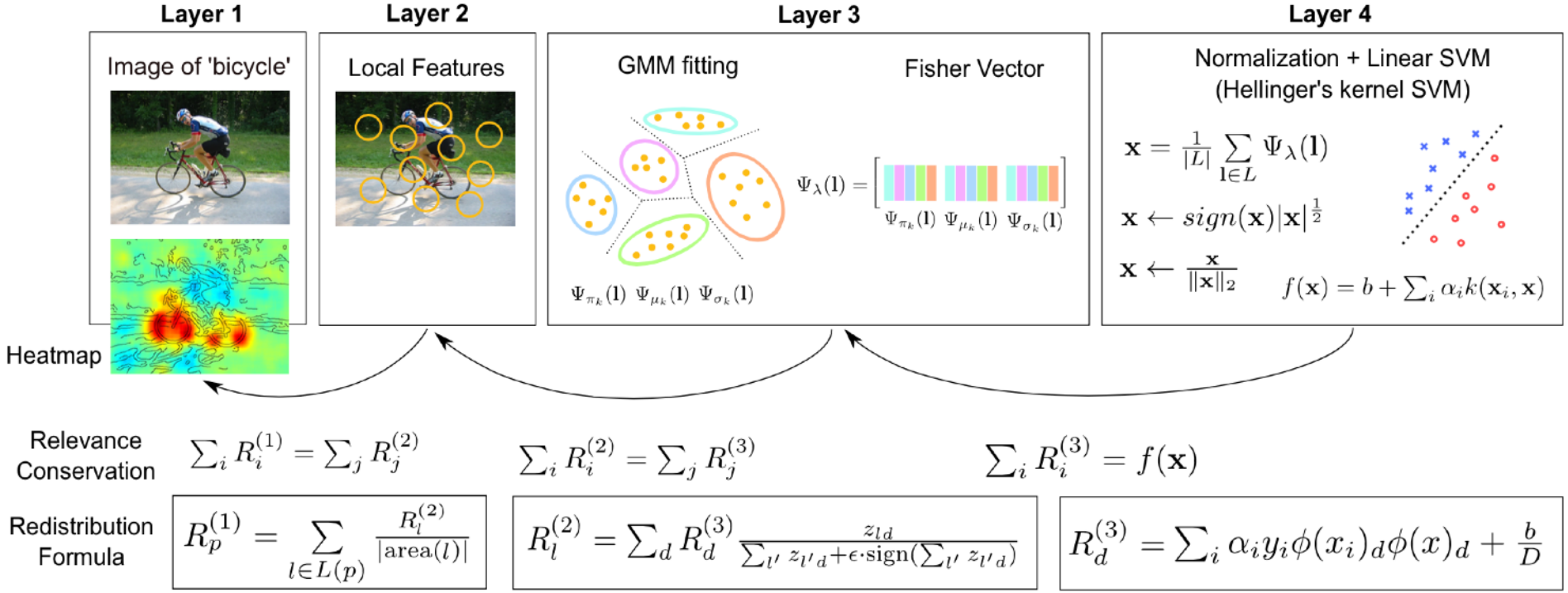
Explaining BoW-Based Classifiers



LRP Principle: (Proportional) Redistribution + Conservation

(Bach et al., 2015)

Explaining BoW-Based Classifiers

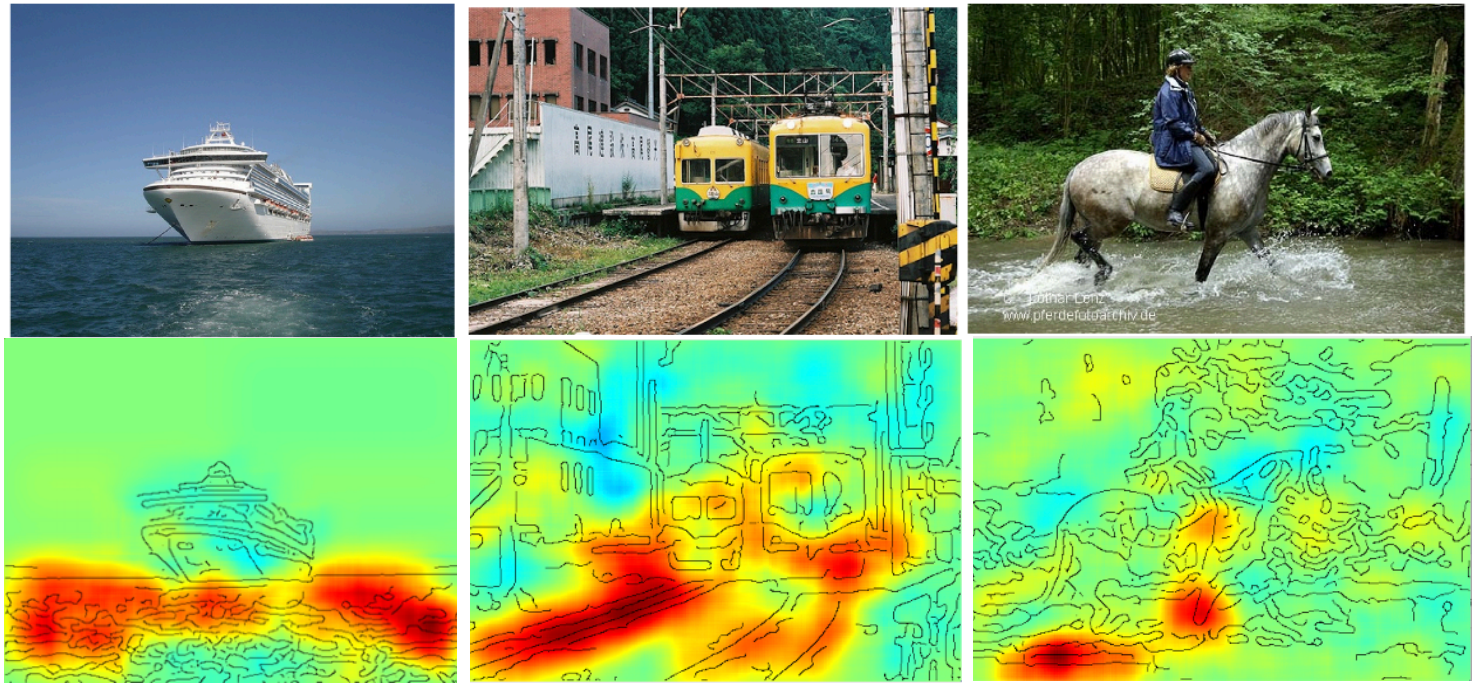


Unmasking Clever Hans Predictors

Leading method (Fisher-Vector / SVM Model) of PASCAL VOC challenge

Unmasking Clever Hans Predictors

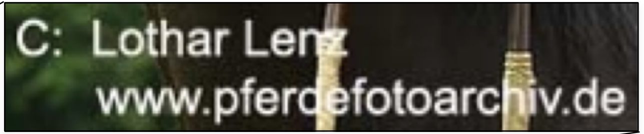
Leading method (Fisher-Vector / SVM Model) of PASCAL VOC challenge



(Lapuschkin et al., 2019)

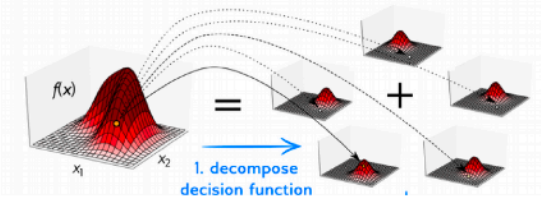
Unmasking Clever Hans Predictors

‘horse’ images in PASCAL VOC 2007

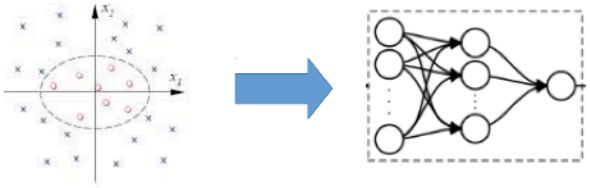


The “Neuralization” Trick

LRP’s idea: To robustly explain a model, leverage the neural network structure of the decision function.



NEON’s idea: When the ML model is not a neural network (e.g. a kernel machine), convert it into a neural network first (‘neuralize’ it).



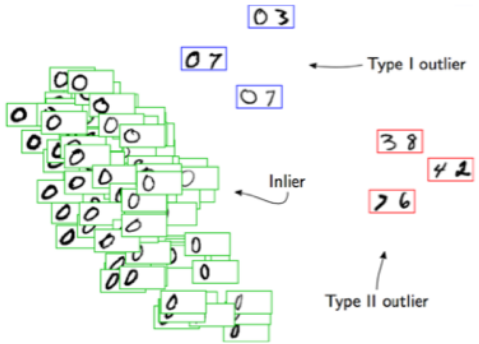
The “Neuralization” Trick

NEON (Neuralization-Propagation)

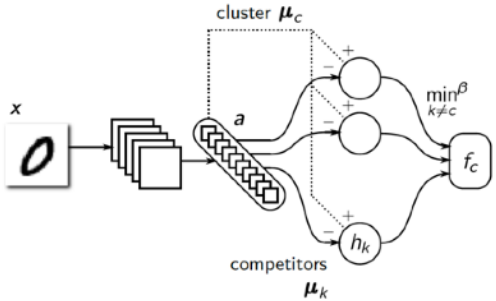
Explain ML algorithm (e.g., SVM, k-Means) in two steps:

1. Convert it into a neural network (‘neuralize it’)
2. Explain the neural network with propagation methods (LRP)

One-class SVM (Kauffmann’18)



Clustering (Kauffmann’19)



Neuralizing Clustering

Class or cluster membership probabilities are often modeled via the 'softmax' function:

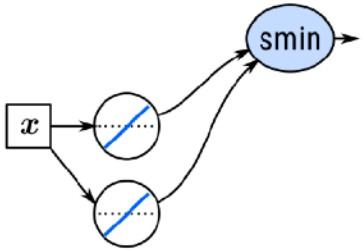
$$p_k = \frac{\exp(\mathbf{w}_k^\top \mathbf{a})}{\sum_j \exp(\mathbf{w}_j^\top \mathbf{a})}$$

Because softmax saturates at 0 and 1, it doesn't capture the full evidence for/against the class. The log-likelihood ratio $\ell_k = \log(p_k/(1-p_k))$ does not saturate.

This quantity can be rewritten as a strictly equivalent two-layer neural network:

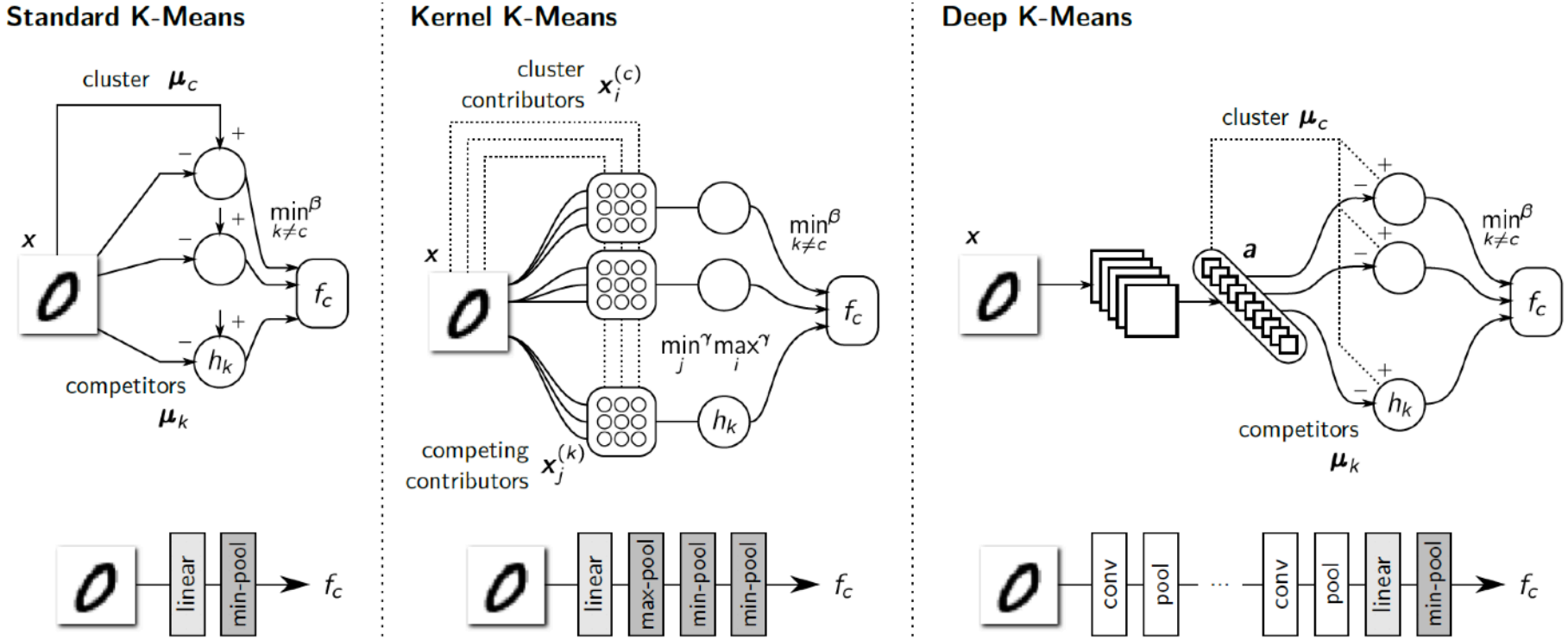
$$h_j = (\mathbf{w}_k - \mathbf{w}_j)^\top \mathbf{a} \qquad \text{(layer 1)}$$

$$\ell_k(\mathbf{a}) = \underbrace{-\log \sum_{j \neq k} \exp(-h_j)}_{\text{smin}} \qquad \text{(layer 2)}$$



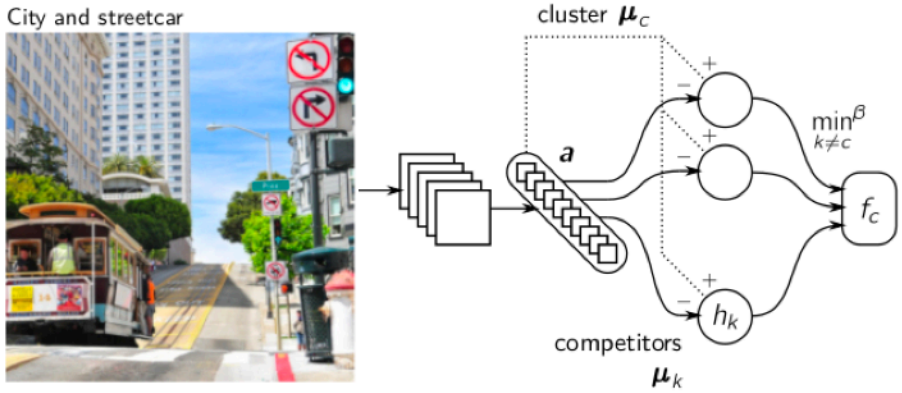
(Kauffmann et al. 2019)

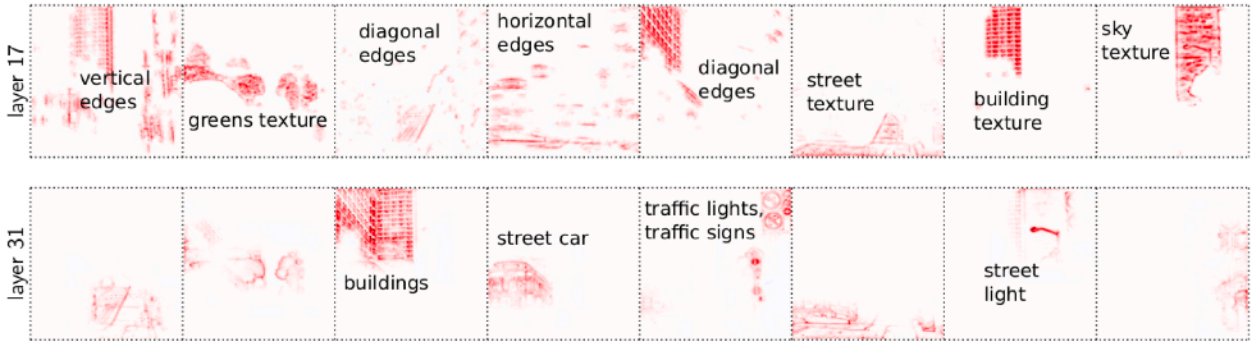
Neuralizing K-Means



(Kauffmann et al. 2019)

K-Means on VGG-16 Features

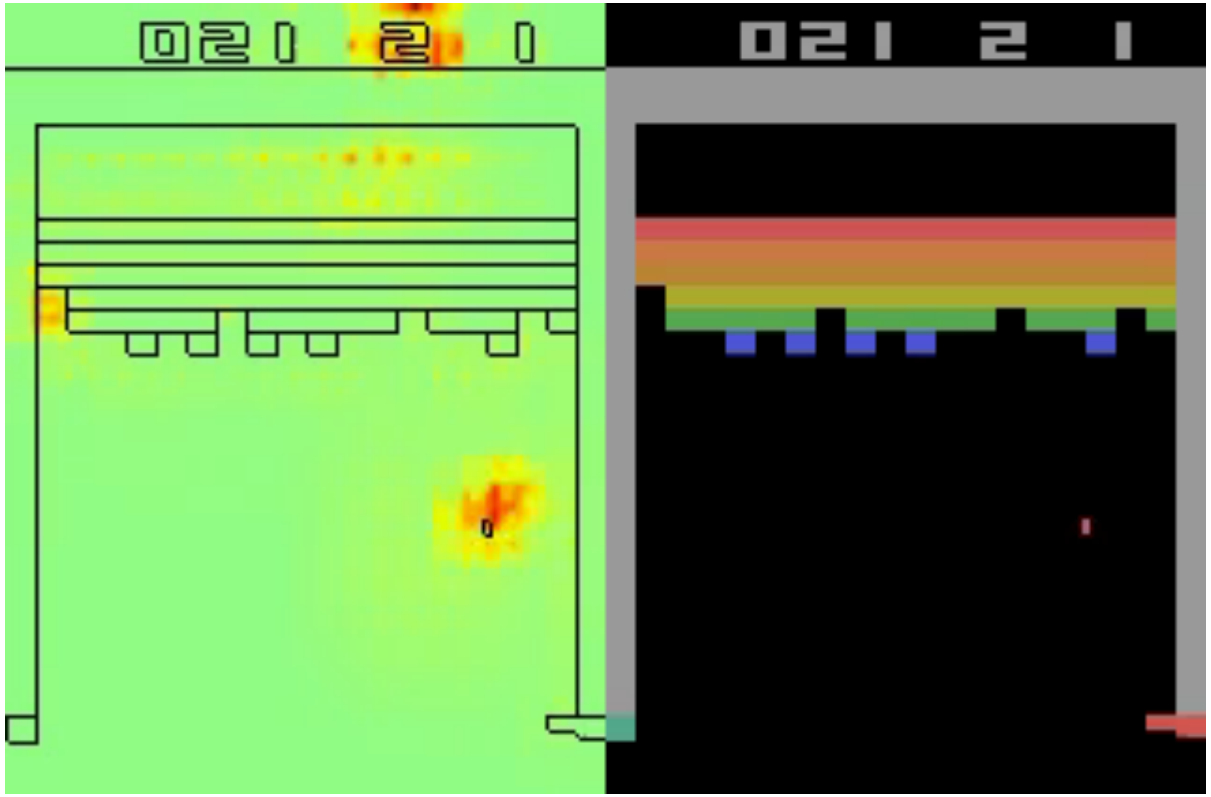




(Kauffmann et al. 2019)

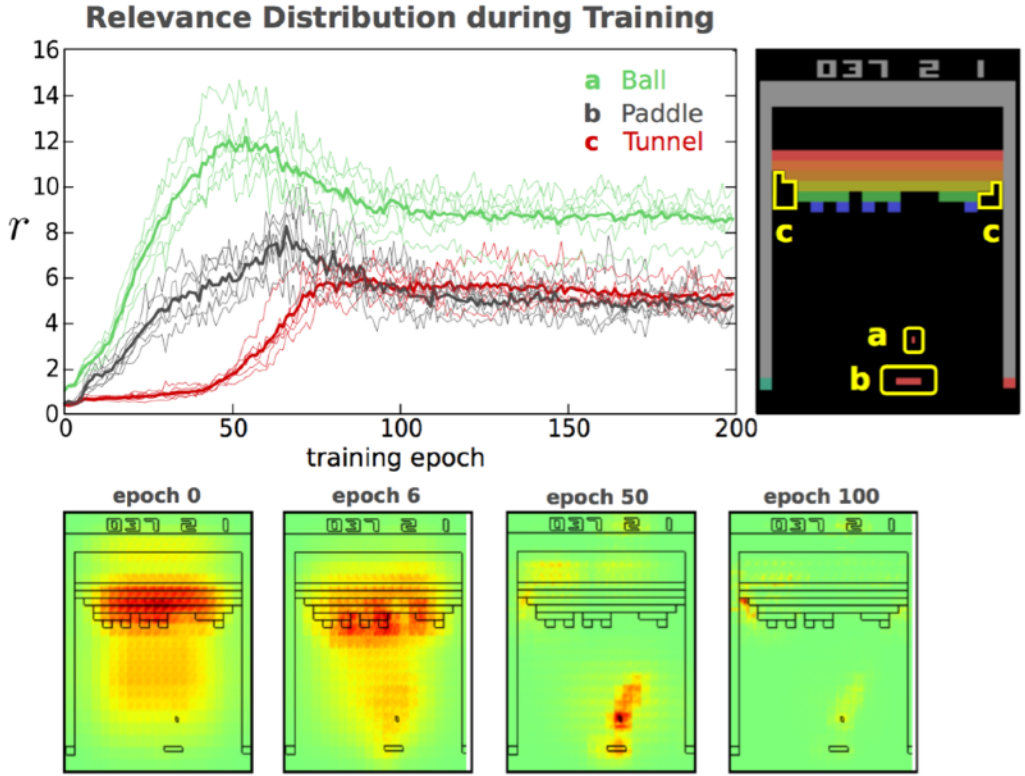
Beyond Deep Classifiers:
XAI Beyond Visualization

Understanding Learning Behaviour



(Lapuschkin et al., 2019)

Understanding Learning Behaviour

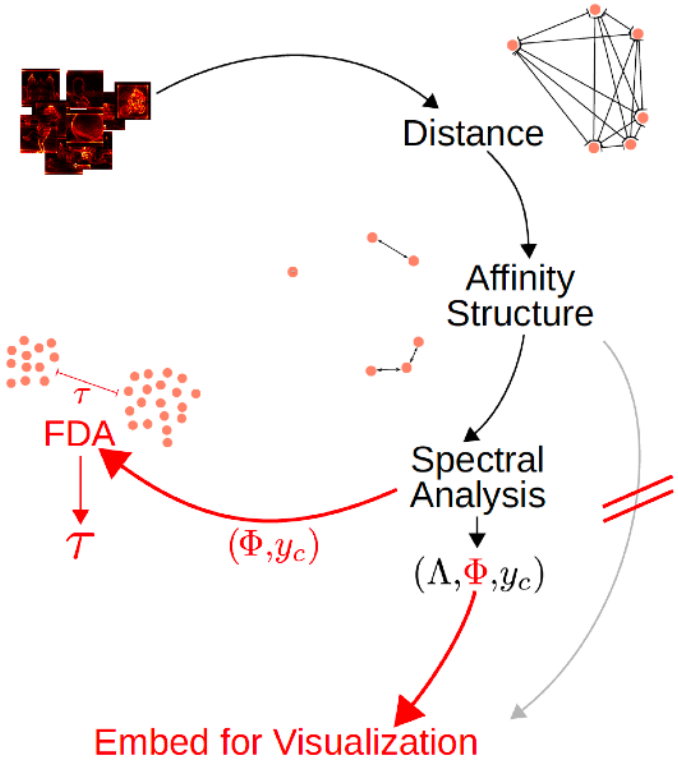


model learns
1. track the ball
2. focus on paddle
3. focus on the tunnel

nature COMMUNICATIONS

Unmasking Clever Hans predictors and assessing what machines really learn

Automating Clever Hans Detection



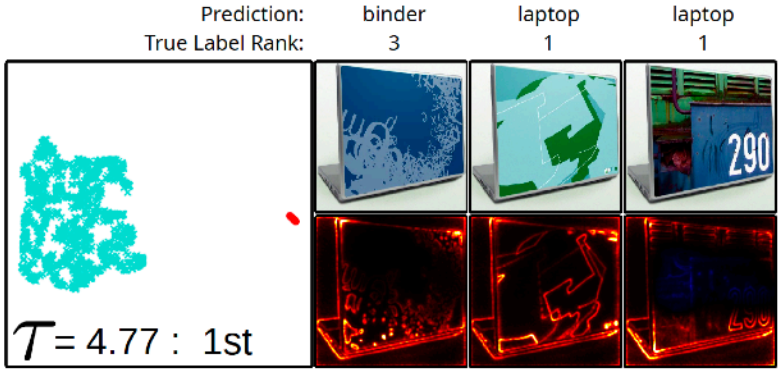
Extending SpRAy from [4]

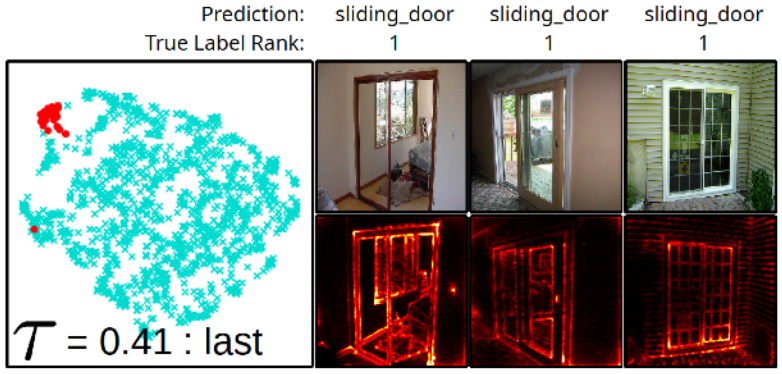
- Further automating spurious cluster/class discovery by analyzing Φ with FDA[7]
- Visualizing the spectal embedding Φ, instead of affinity structure

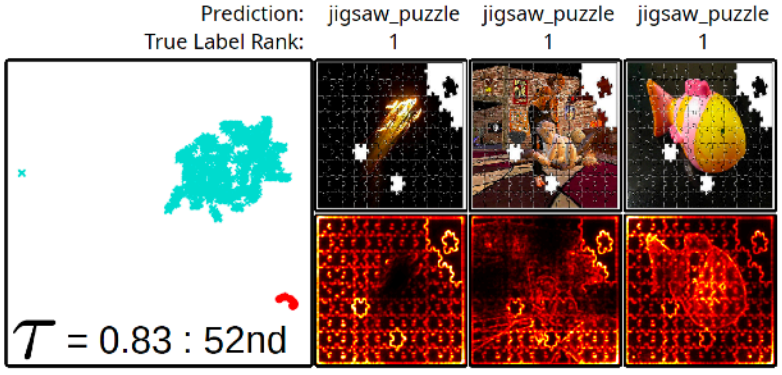
$$J(w) = \frac{w^\mathsf{T} S_b w}{w^\mathsf{T} S_w w}$$

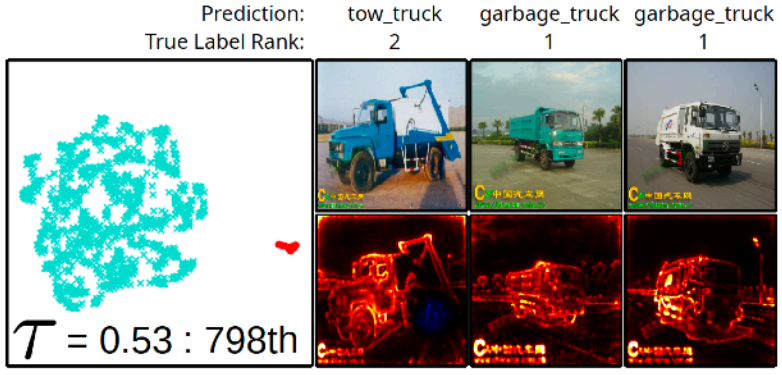
(Anders et al. 2019)

Automating Clever Hans Detection

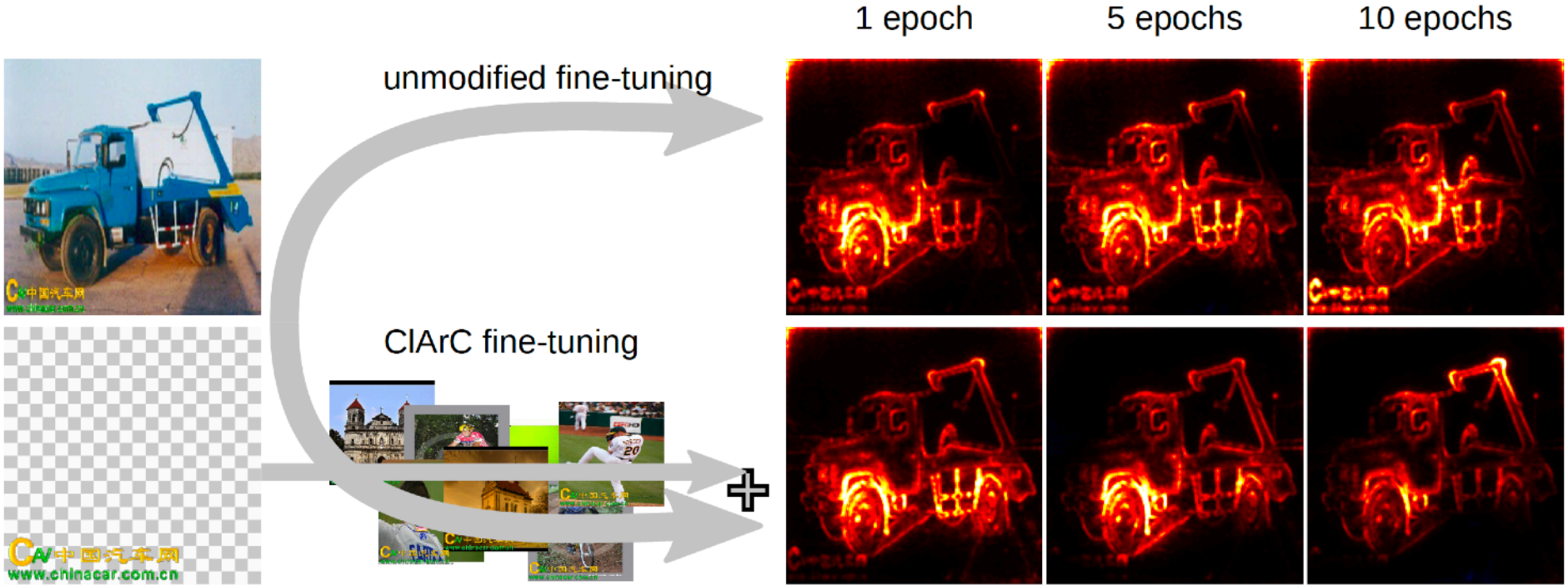








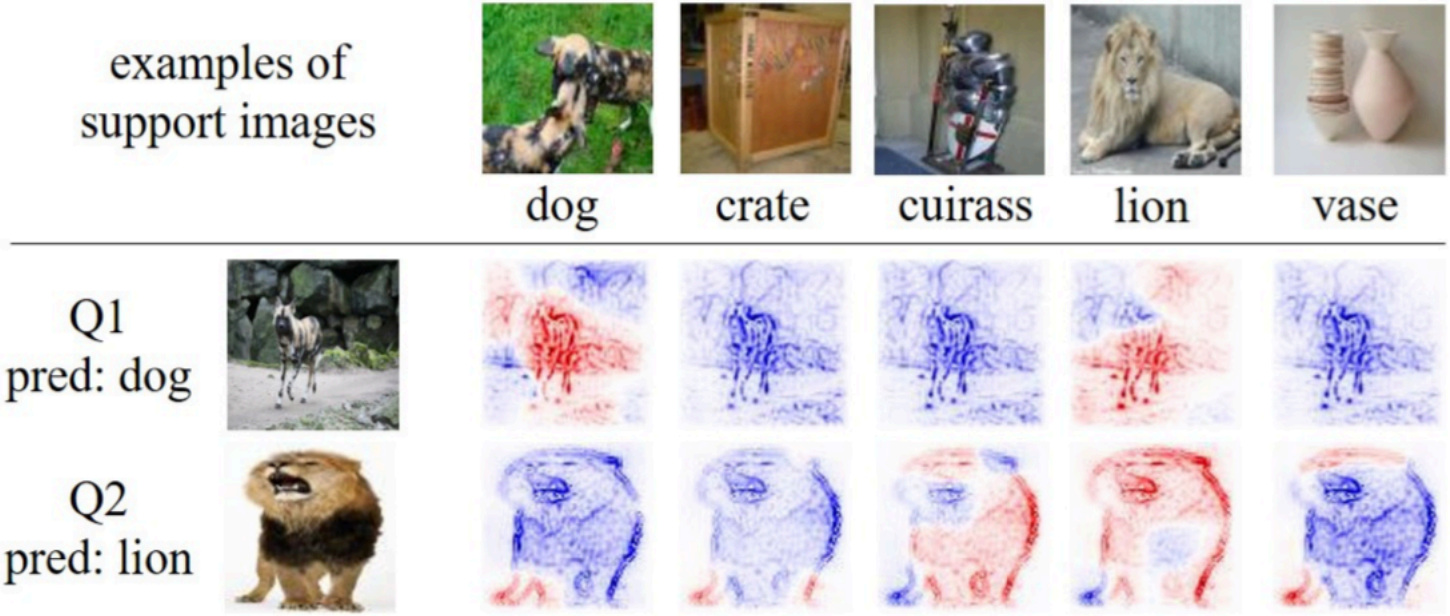
Unhansing



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

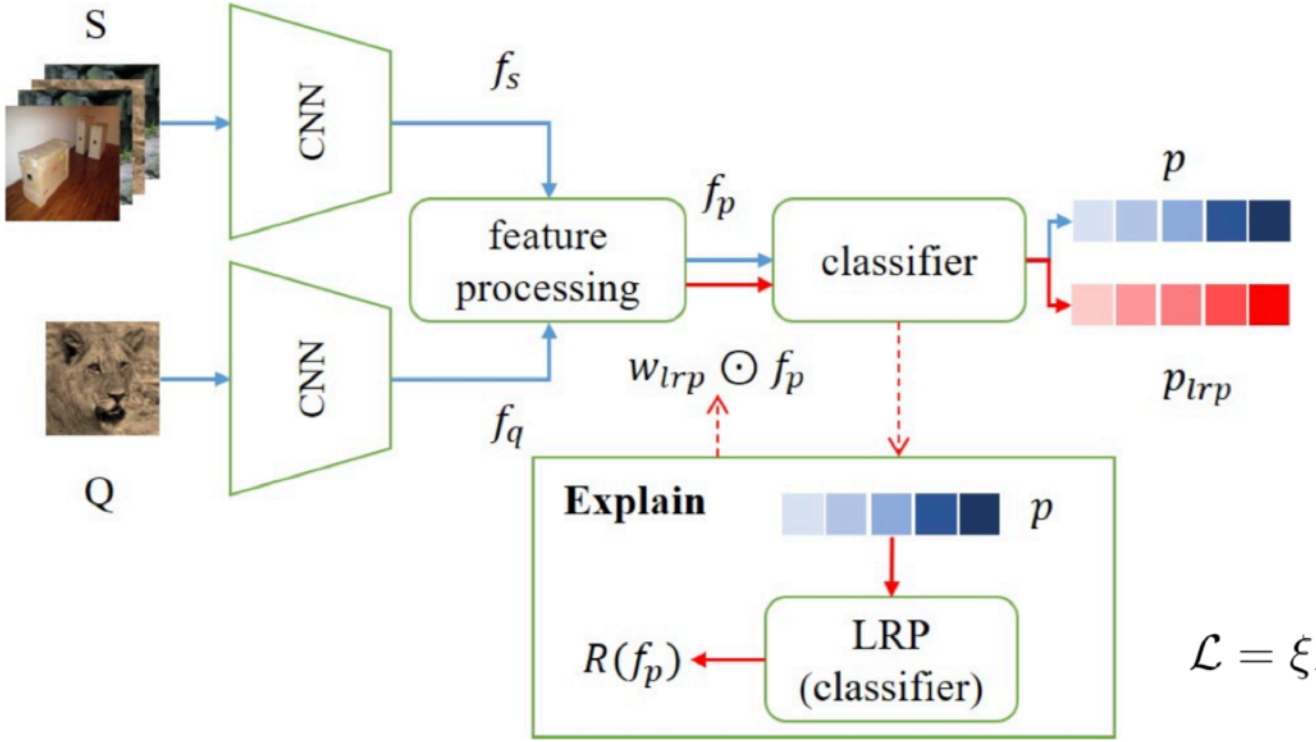
Explanation-Guided Training

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



(Sun et al. 2020)

Explanation-Guided Training



$$w_{lrp} = 1 + R(f_p)$$

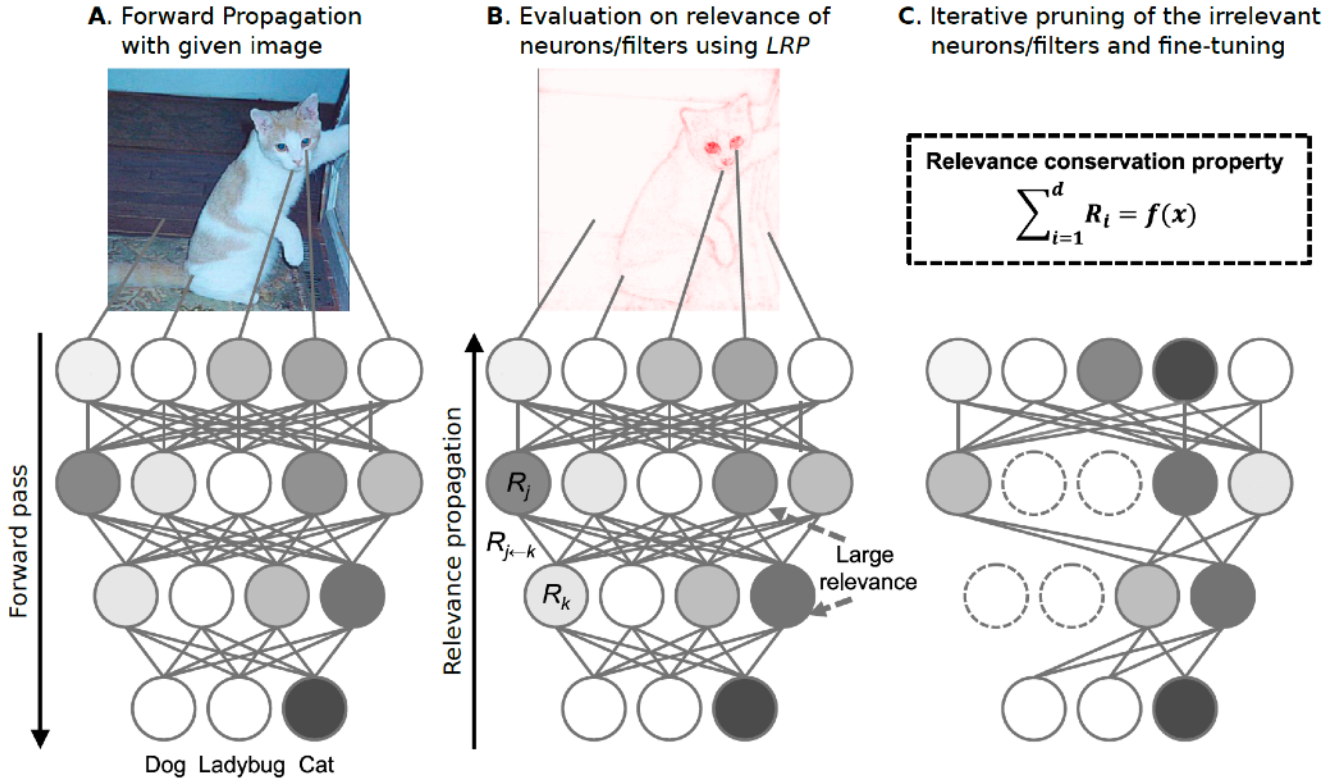
$$f_{p-lrp} = w_{lrp} \odot f_p$$

$$\mathcal{L} = \xi \mathcal{L}_{ce}(y, p) + \lambda \mathcal{L}_{ce}(y, p_{lrp})$$

Explanation-Guided Training

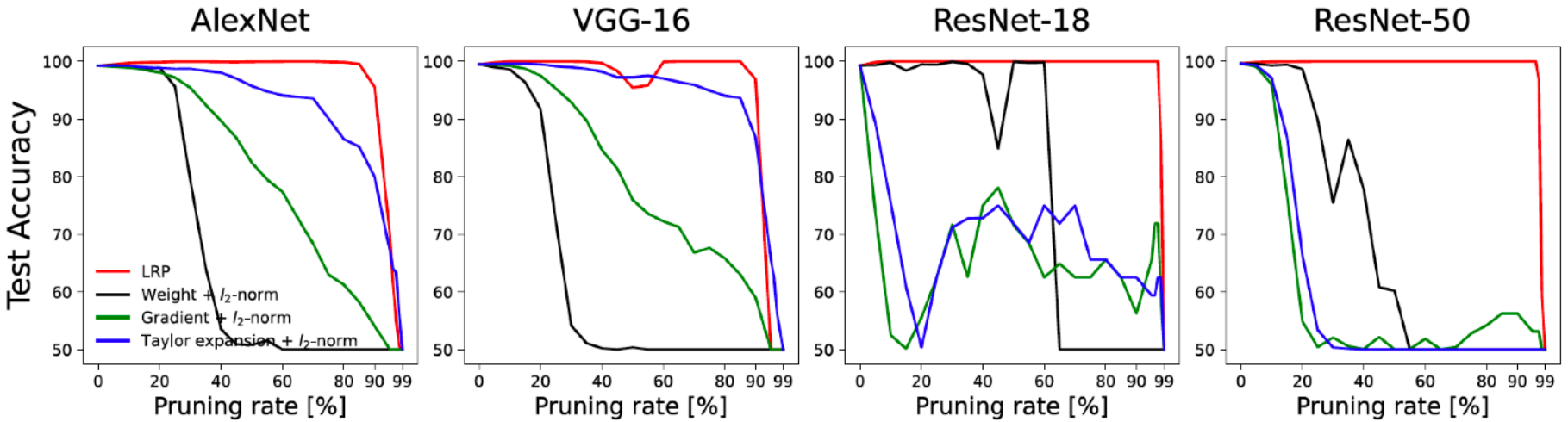
5-way 1-shot	Cars	Places	CUB	Plantae
RN	29.40±0.33%	48.05±0.46%	44.33±0.43%	34.57±0.38%
FT-RN	30.09±0.36%	48.12±0.45%	44.87±0.44%	35.53±0.39%
LRP-RN	30.00±0.32%	48.74±0.45%	45.64±0.42%	36.04±0.38%
LFT-RN	30.27±0.34%	48.07±0.46%	47.35±0.44%	35.54±0.38%
LFT-LRP-RN	**30.68±0.34%**	**50.19±0.47%**	**47.78±0.43%**	**36.58±0.40%**

XAI-Based Pruning



(Yeom et al. 2019)

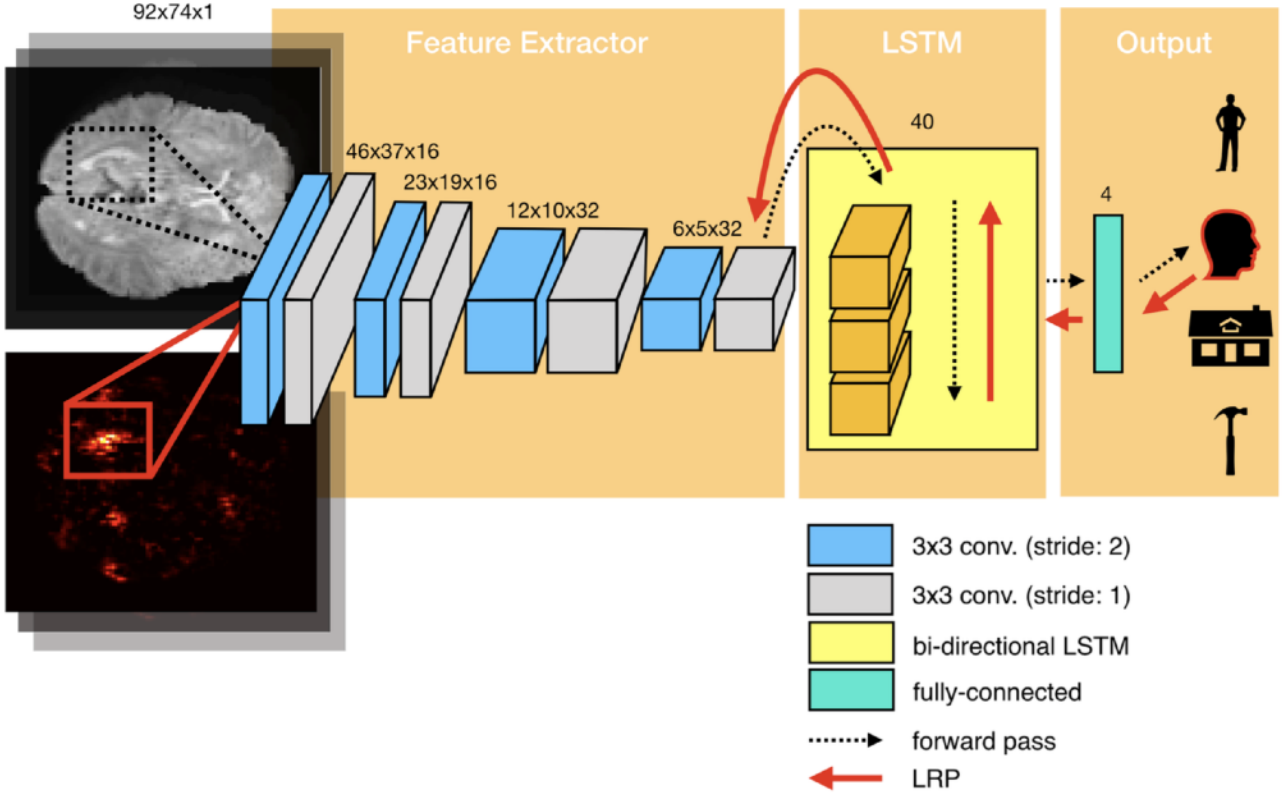
XAI-Based Pruning



No fine-tuning

only 10 samples per class
(domain adaptation scenario)

XAI in the Sciences

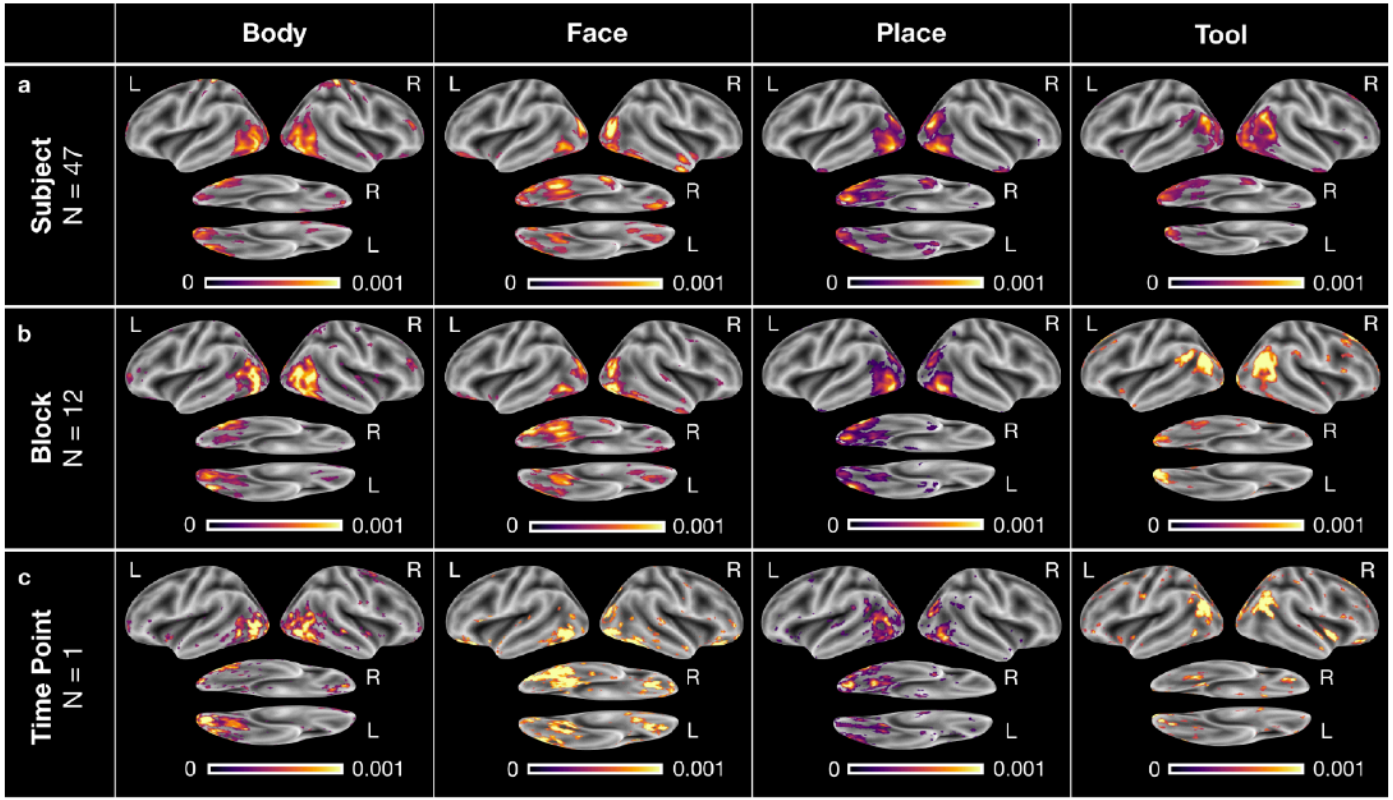


Our approach:

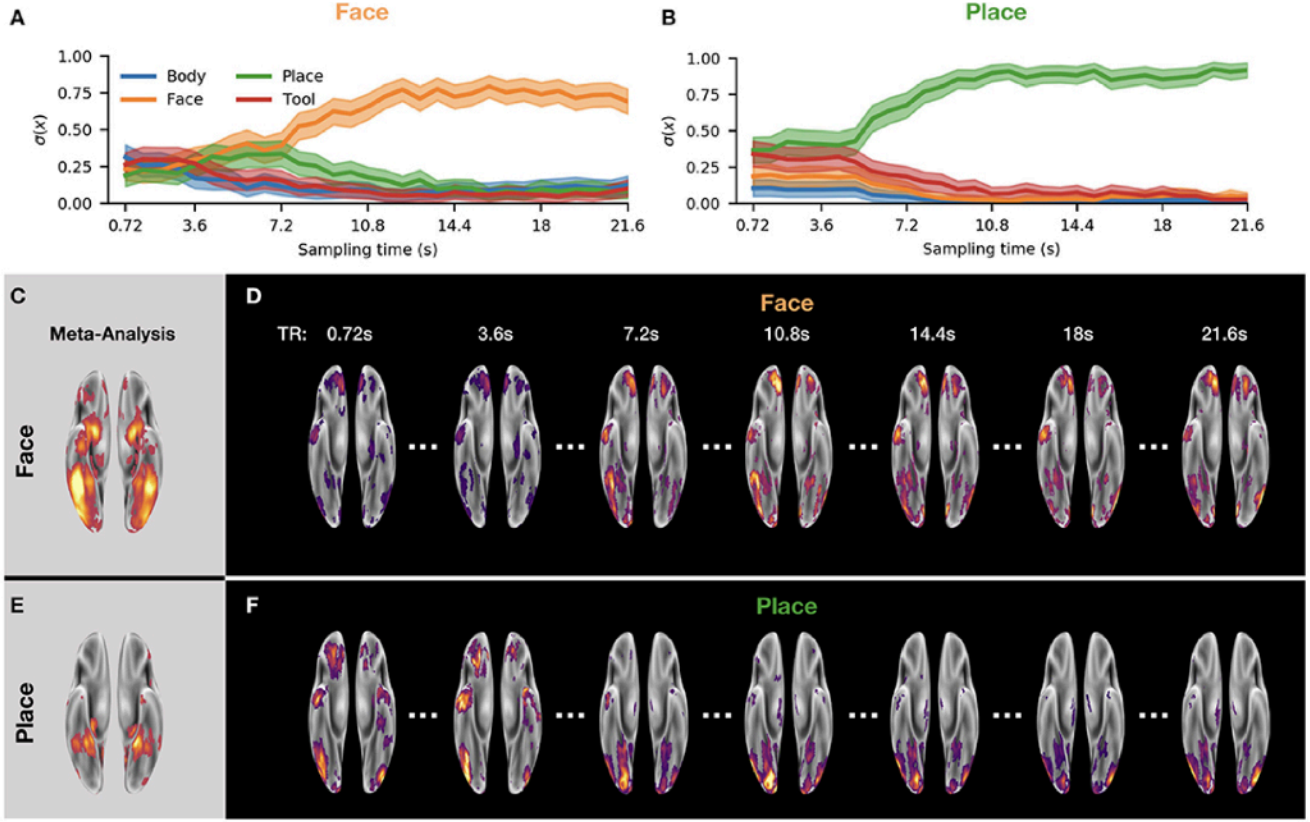
- Recurrent neural networks (CNN + LSTM) for whole-brain analysis
- LRP allows to interpret the results

(Thomas et al. 2019)

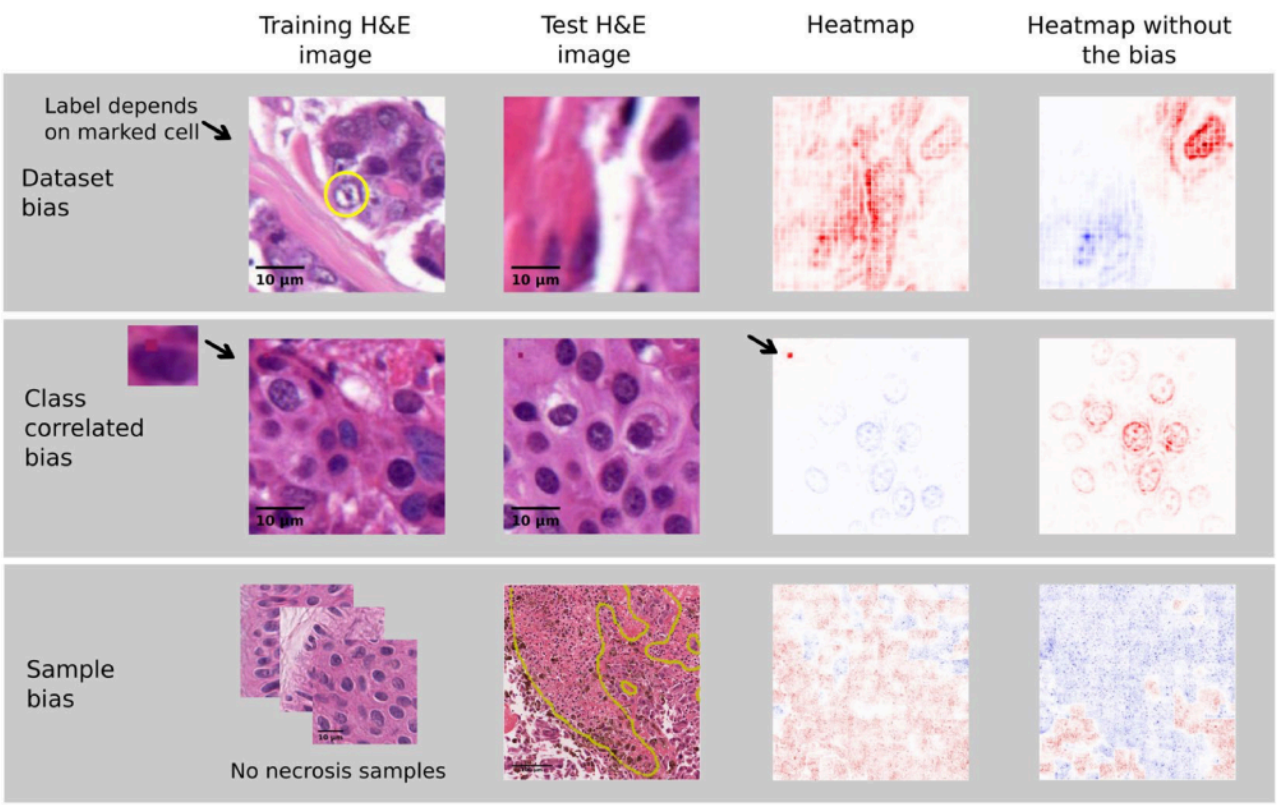
XAI in the Sciences



XAI in the Sciences



XAI in the Sciences



determining the label solely from the patch's centre cell (yellow mark)

small artificial corruption

training a classifier on a dataset lacking examples of necrosis

(Hägele et al., 2020)

Conclusion

Conclusion

XXAI: Extending Explainable AI Beyond Deep Models and Classifiers

ICML 2020 Workshop

Explanations can be used beyond visualization purposed

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

Large interested of XAI in scientific communities

References

Tutorial / Overview Papers

- W Samek, G Montavon, S Lapuschkin, C Anders, KR Müller. Toward Interpretable Machine Learning: Transparent Deep Neural Networks and Beyond
  arXiv:2003.07631, 2020
- 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 - Special Issue 1 - The Impact of AI on Communication Networks and Services, 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]

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

- A Osman, L Arras, W Samek. Towards Ground Truth Evaluation of Visual Explanations 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

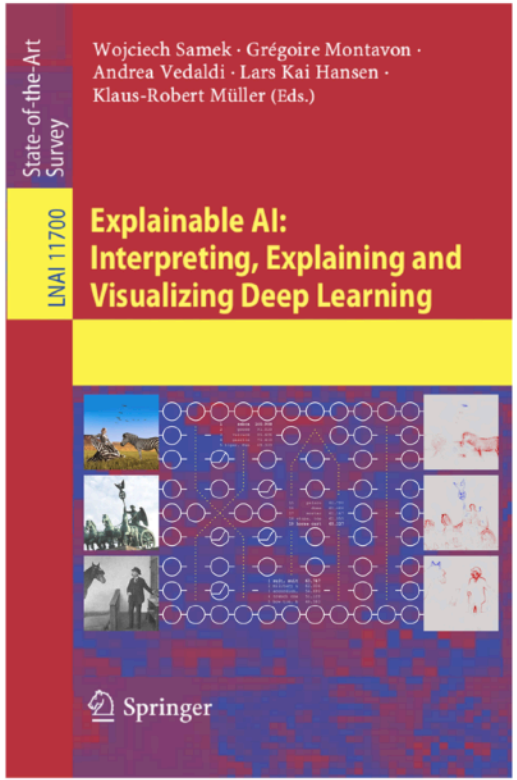
Application to Neural Network Pruning

- S Yeom, P Seegerer, S Lapuschkin, S Wiedemann, KR Müller, W Samek. Pruning by Explaining: A Novel Criterion for Deep Neural Network Pruning
  arXiv:1912.08881, 2019

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
  arXiv:2007.08790, 2020

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