ECML/PKDD 2020, Tutorial: Explainable AI for Deep Network: Basics and Extensions

Part 4: Applications

Wojciech Samek, Grégoire Montavon

September 18, 2020

Outline

Walk-through examples

Meta-Explanations

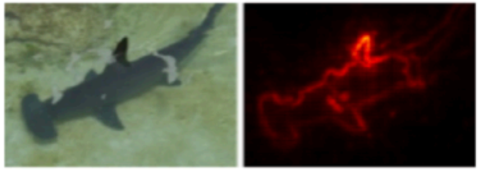
Explanation beyond visualization

XAI in the Sciences

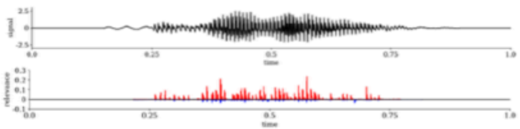
Outlook

LRP Applied to Different Problems

General Images (Bach' 15, Lapuschkin'16)



Speech (Becker'18)

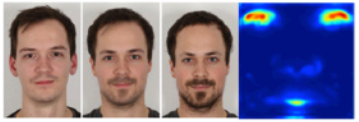


Text Analysis (Arras'16 &17)

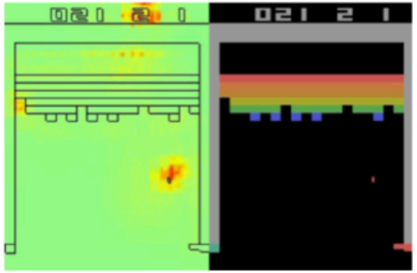
do n't waste your money

neither funny nor susper

Morphing Attacks (Seibold'18)

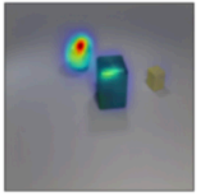


Games (Lapuschkin'19)

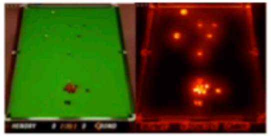


VQA (Samek'19)

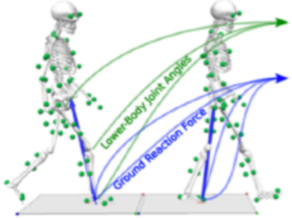
there is a metallic cube ; are there any large cyan metallic objects behind it ?



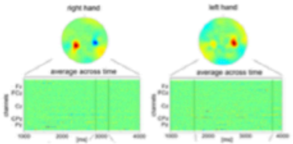
Video (Anders'19)



Gait Patterns (Horst'19)



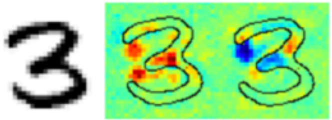
EEG (Sturm'16)



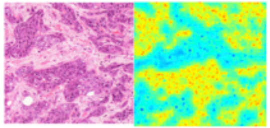
Faces (Lapuschkin'17)

Digits (Bach' 15)

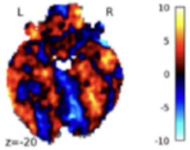
Image Class '3' Class '9'



Histopathology (Hägele'19)

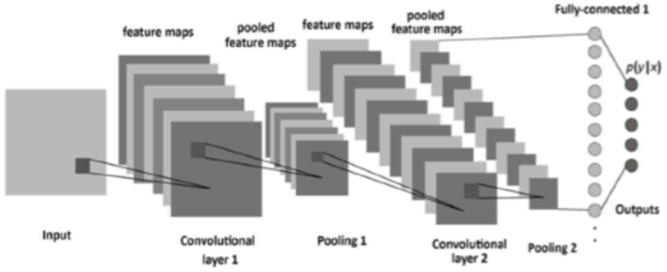


fMRI (Thomas'18)

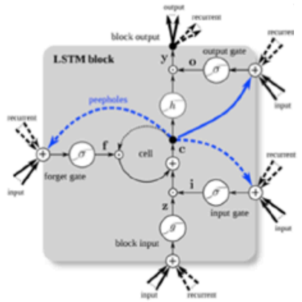


LRP Applied to Different Problems

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

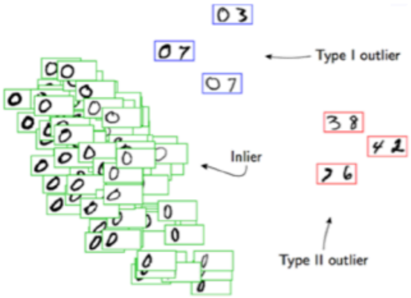


LSTM (Arras’17, Arras’19)

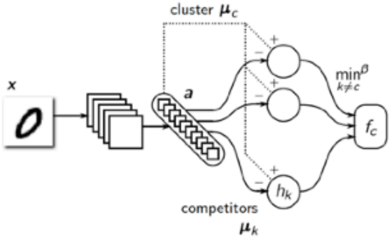


BoW / Fisher Vector models
(Bach’15, Arras’16, Lapuschkin’16 …)

One-class SVM (Kauffmann’18)



Clustering (Kauffmann’19)



Walk-Through Examples

Validating a Face Classifier

Faces in the wild (from Flickr)

#images: 26,580

Task: Predict gender & age (range)

(0-2), (4-6), (8-13), (15-20), (25-32), (38-43), (48-53), (60+)

Validating a Face Classifier

	A	C	G	V
[i]	51.4 87.0	52.1 87.9	54.3 89.1	–
[r]	51.9 87.4	52.3 88.9	53.3 89.9	–
[m]	53.6 88.4	54.3 89.7	56.2 90.7	–
[i,n]	–	51.6 87.4	56.2 90.9	53.6 88.2
[r,n]	–	52.1 87.0	57.4 91.9	–
[m,n]	–	52.8 88.3	58.5 92.6	56.5 90.0
[i,w]	–	–	–	59.7 94.2
[r,w]	–	–	–	–
[m,w]	–	–	–	62.8 95.8

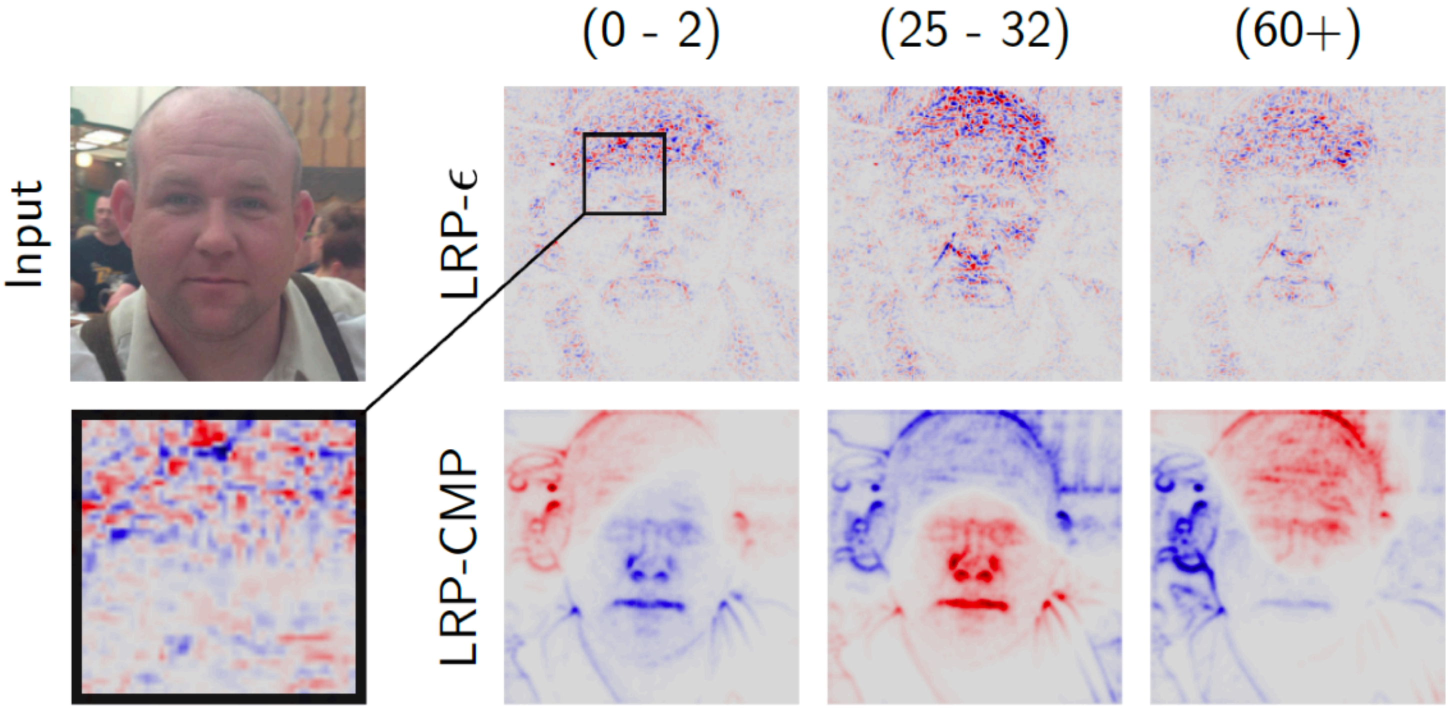
	A	C	G	V
[i]	88.1	87.4	87.9	–
[r]	88.3	87.8	88.9	–
[m]	89.0	88.8	89.7	–
[i,n]	–	89.9	**91.0**	**92.0**
[r,n]	–	90.6	**91.6**	–
[m,n]	–	90.6	**91.7**	**92.6**
[i,w]	–	–	–	90.5
[r,w]	–	–	–	–
[m,w]	–	–	–	**92.2**

A = AdienceNet
C = CaffeNet
G = GoogleNet
V = VGG-16

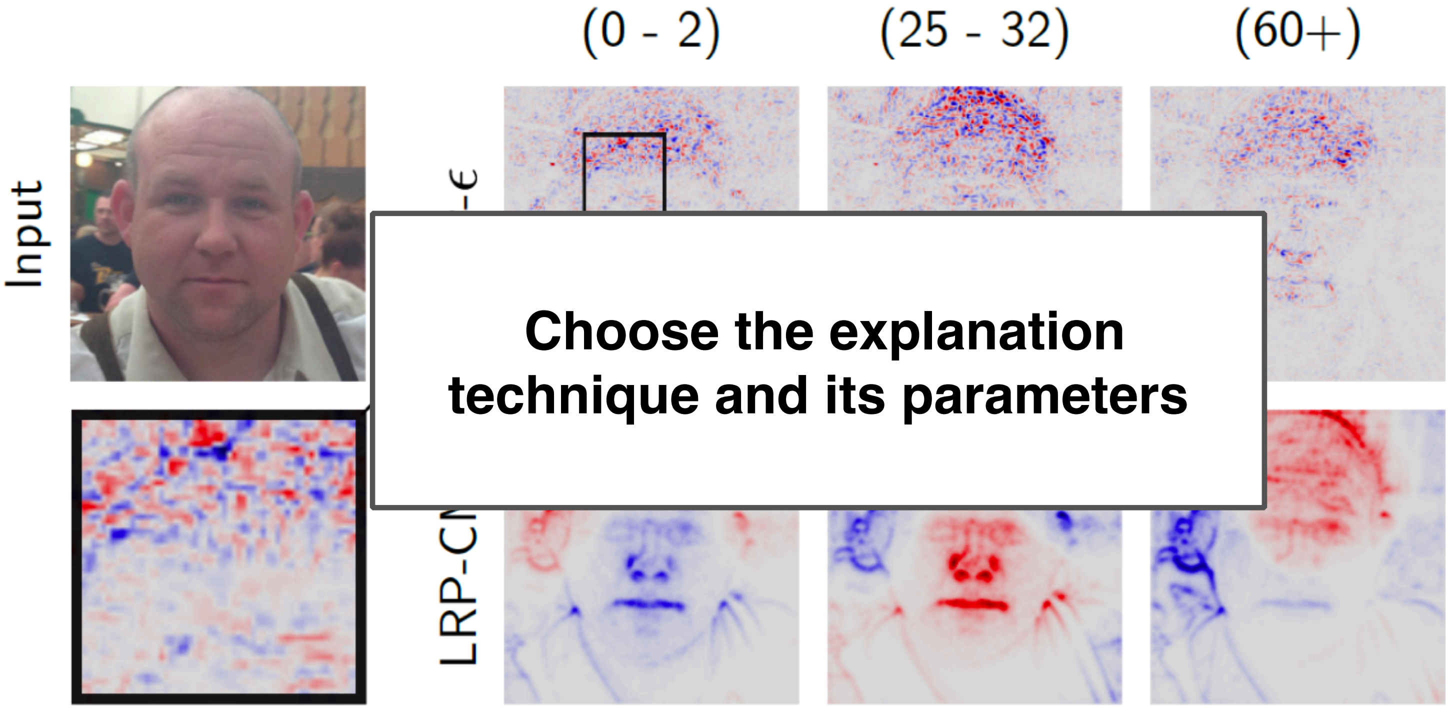
[i] = in-place face alignment
[r] = rotation based alignment
[m] = mixing aligned images for training
[n] = initialization on Imagenet
[w] = initialization on IMDB-WIKI

(Lapuschkin et al., 2017)

Validating a Face Classifier



Validating a Face Classifier

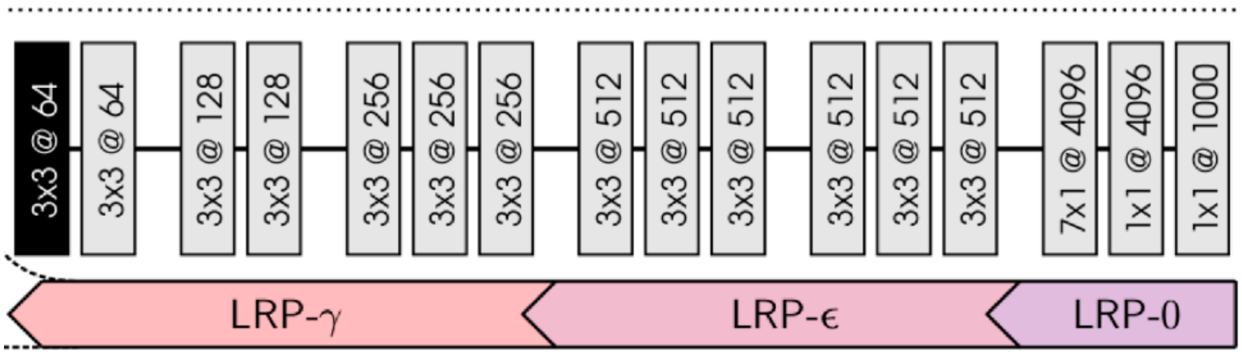


Choose the explanation technique and its parameters

Validating a Face Classifier

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

Composite LRP



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

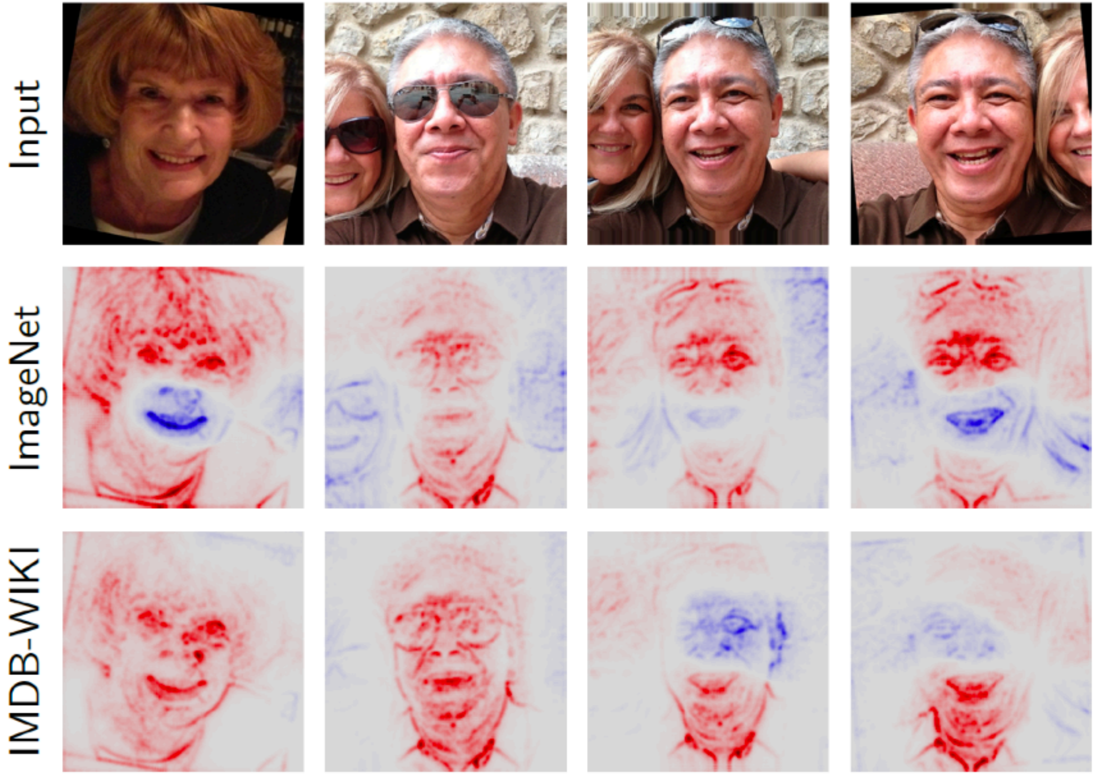
Validating a Face Classifier

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)	✓

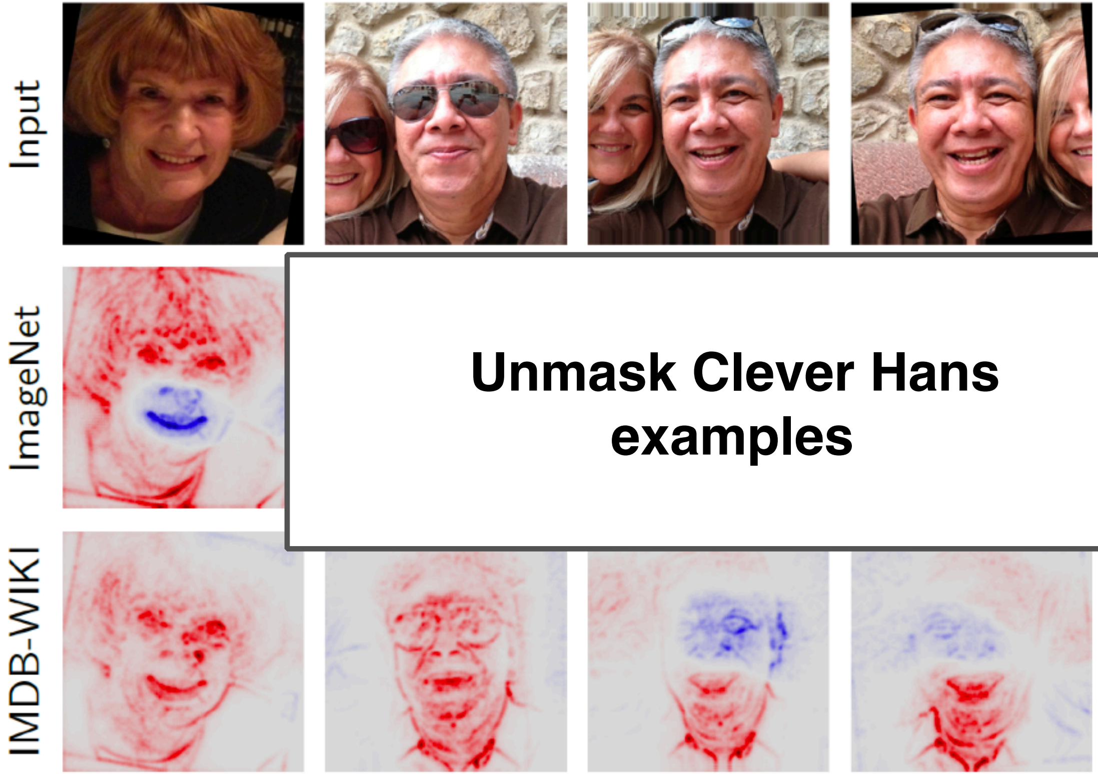
(* DTD interpretation only for the case $\alpha = 1, \beta = 0$.)

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

Validating a Face Classifier



Validating a Face Classifier



Unmask Clever Hans examples

Validating a Face Classifier

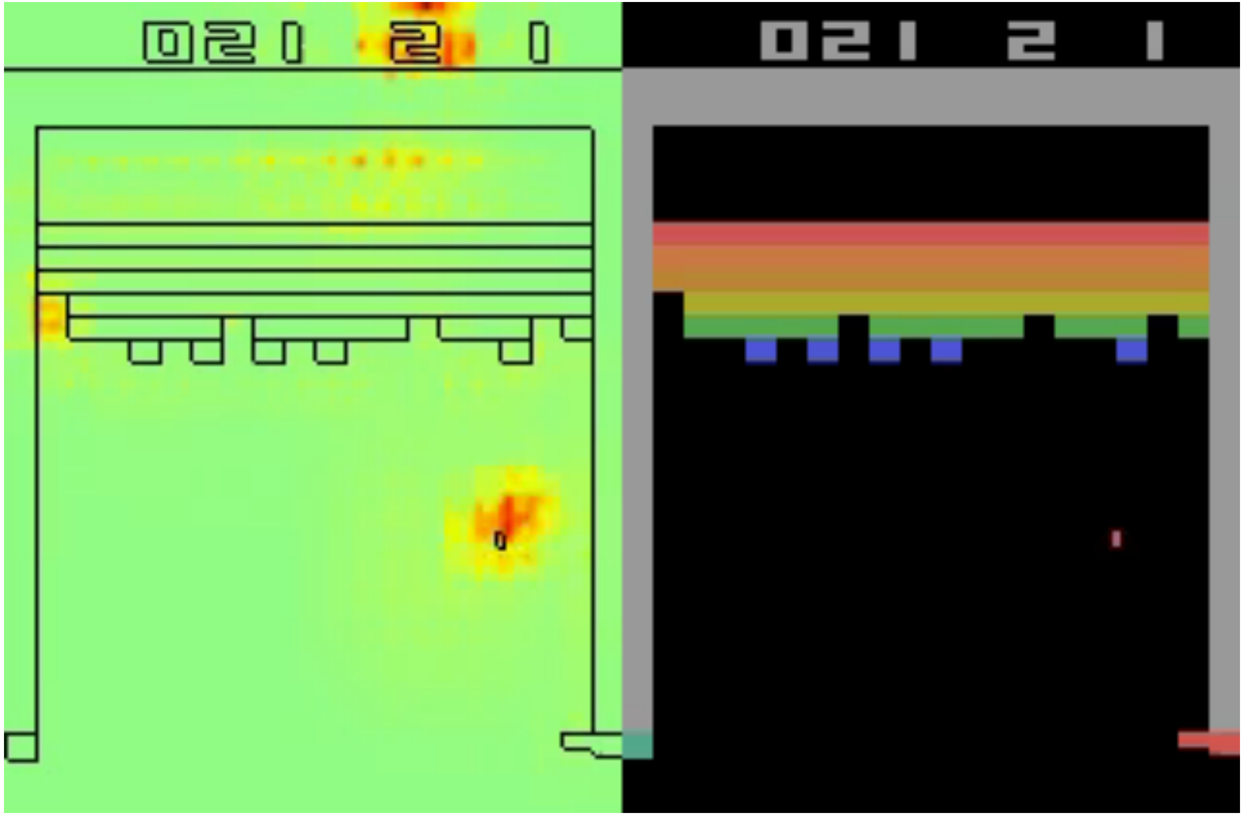
	accuracy	1-off
ImageNet pretrained	56.5	90.0
IMDB-WIKI pretrained	63.0	96.0

Validating a Face Classifier

	1-off
ImageN	90.0
IMDB-	96.0

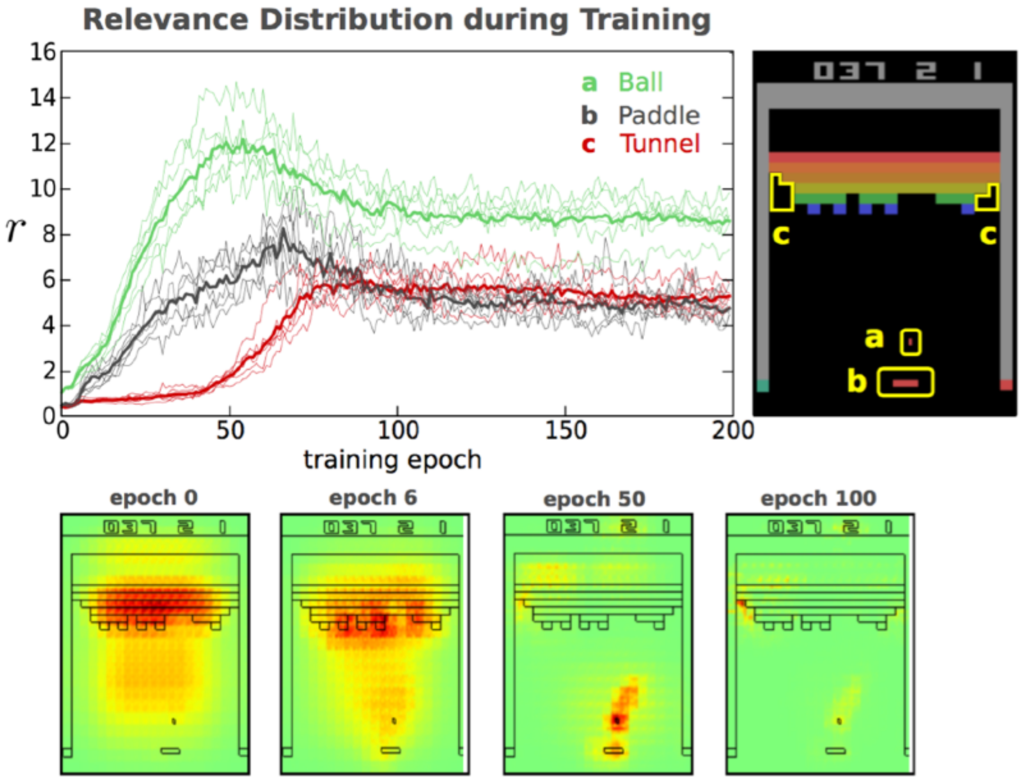
Iteratively validate and improve the model

Understanding Learning Behaviour



(Lapuschkin et al., 2019)

Understanding Learning Behaviour

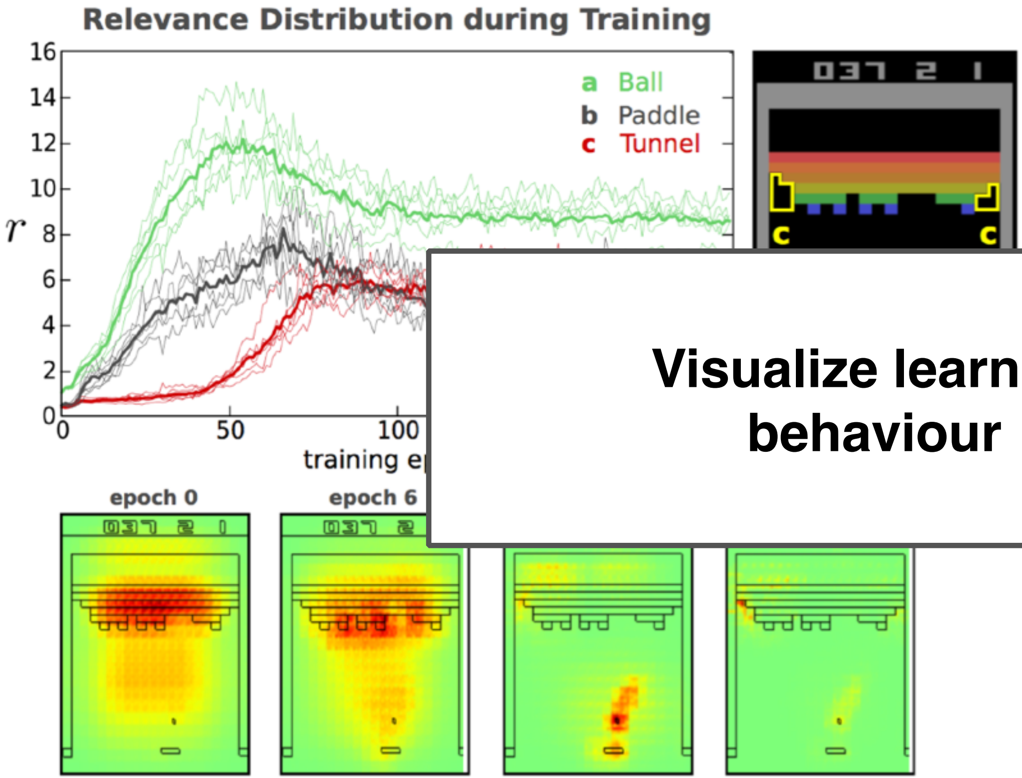


model learns
1. track the ball
2. focus on paddle
3. focus on the tunnel

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

(Lapuschkin et al., 2019)

Understanding Learning Behaviour



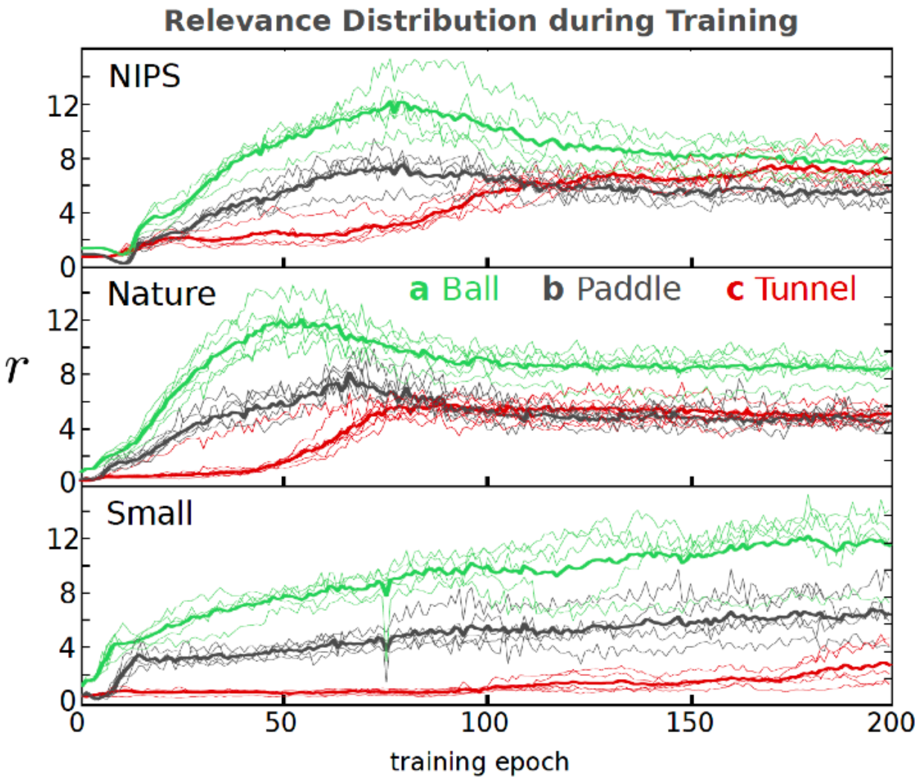
Visualize learning behaviour

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

(Lapuschkin et al., 2019)

Understanding Learning Behaviour

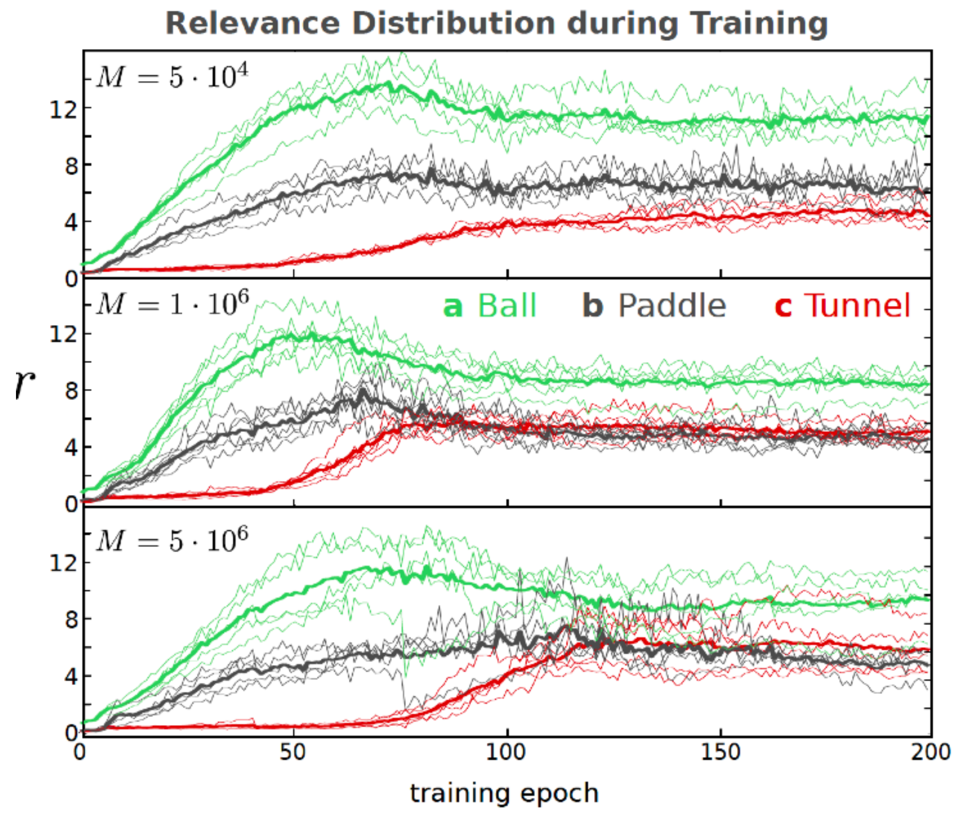


NIPS architecture	Nature architecture
C1 (4×8×8)→(16), [4×4]	C1 (4×8×8)→(32), [4×4]
C2 (16×4×4)→(32), [2×2]	C2 (32×4×4)→(64), [2×2]
	C3 (64×3×3)→(64), [1×1]
F1 (2592)→(256)	F1 (3136)→(512)
F2 (256)→(4)	F2 (512)→(4)

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)

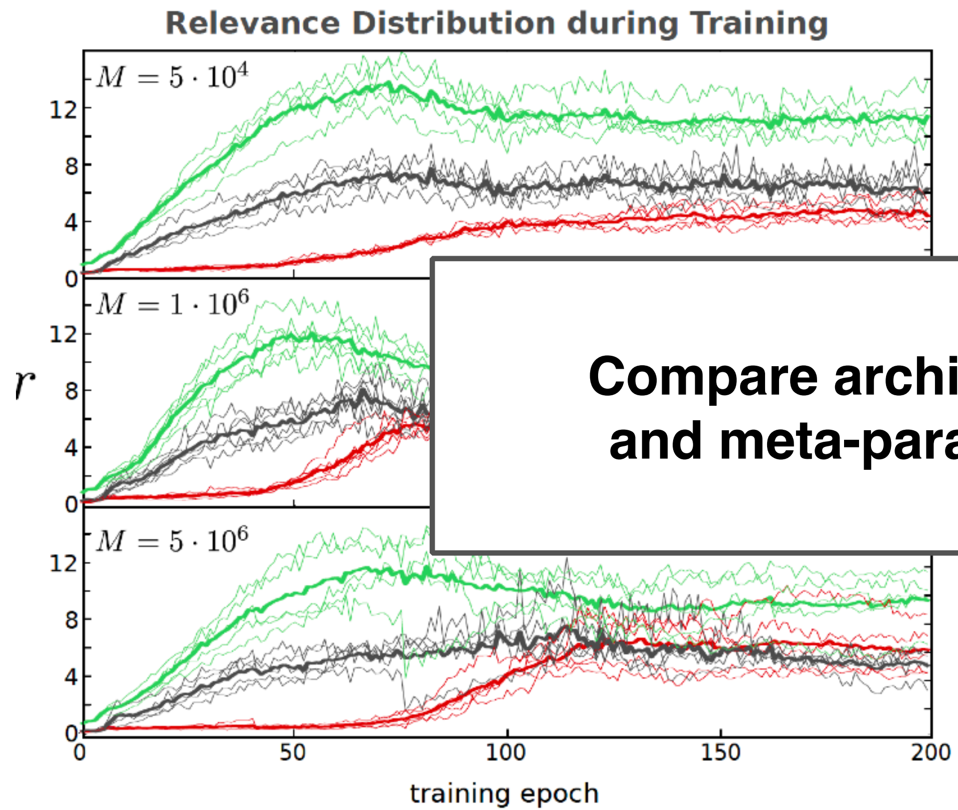
Understanding Learning Behaviour



Varying size of replay memory:
(state, action, reward, next state)

(Lapuschkin et al., 2019)

Understanding Learning Behaviour



Varying size of replay memory:
(state, action, reward, next state)

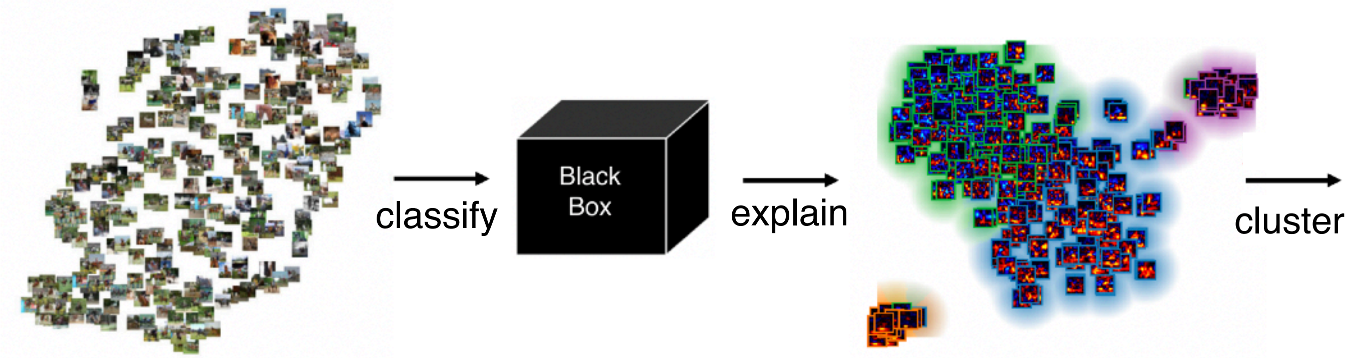
Compare architectures and meta-parameters

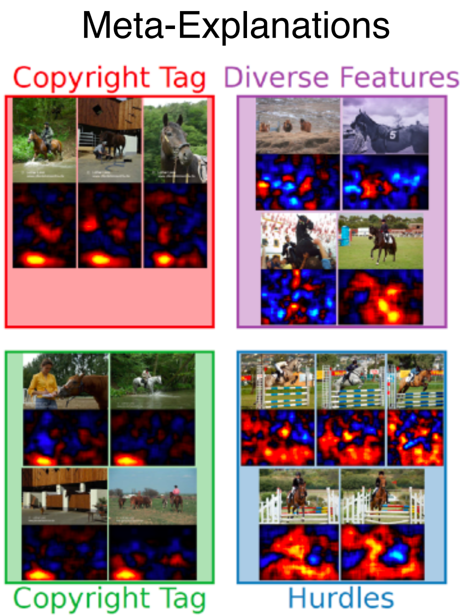
(Lapuschkin et al., 2019)

Meta-Explanations

Meta-Explanations

SpRAy's idea: Explain *whole dataset* decisions of a ML model by systematically analyzing distributions of LRP heatmaps.



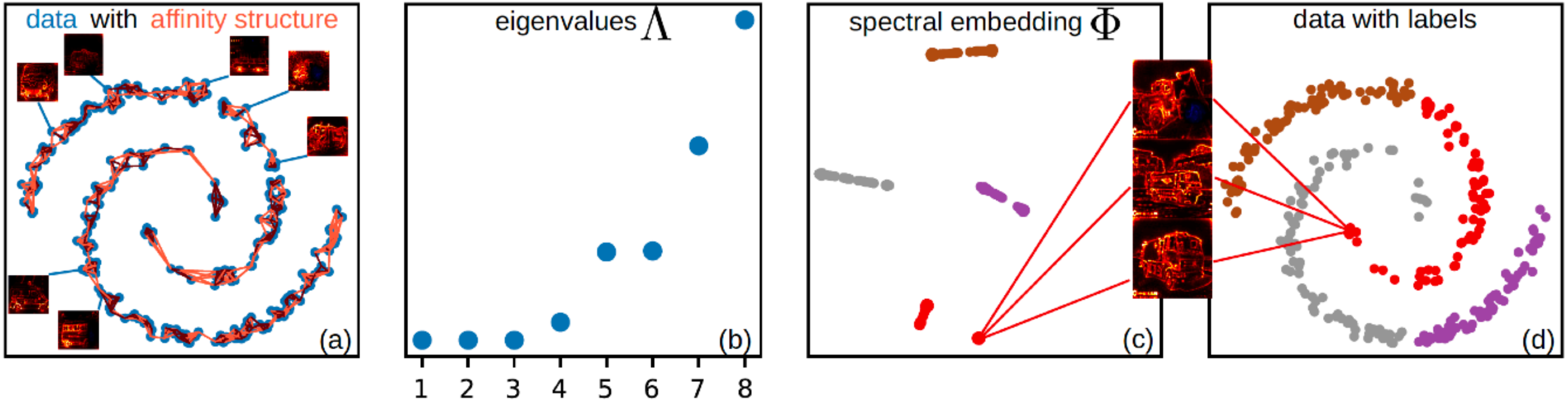


nature COMMUNICATIONS

Unmasking Clever Hans predictors and assessing what machines really learn

(Lapuschkin et al., 2019)

Spectral Relevance Analysis (SpRAy)

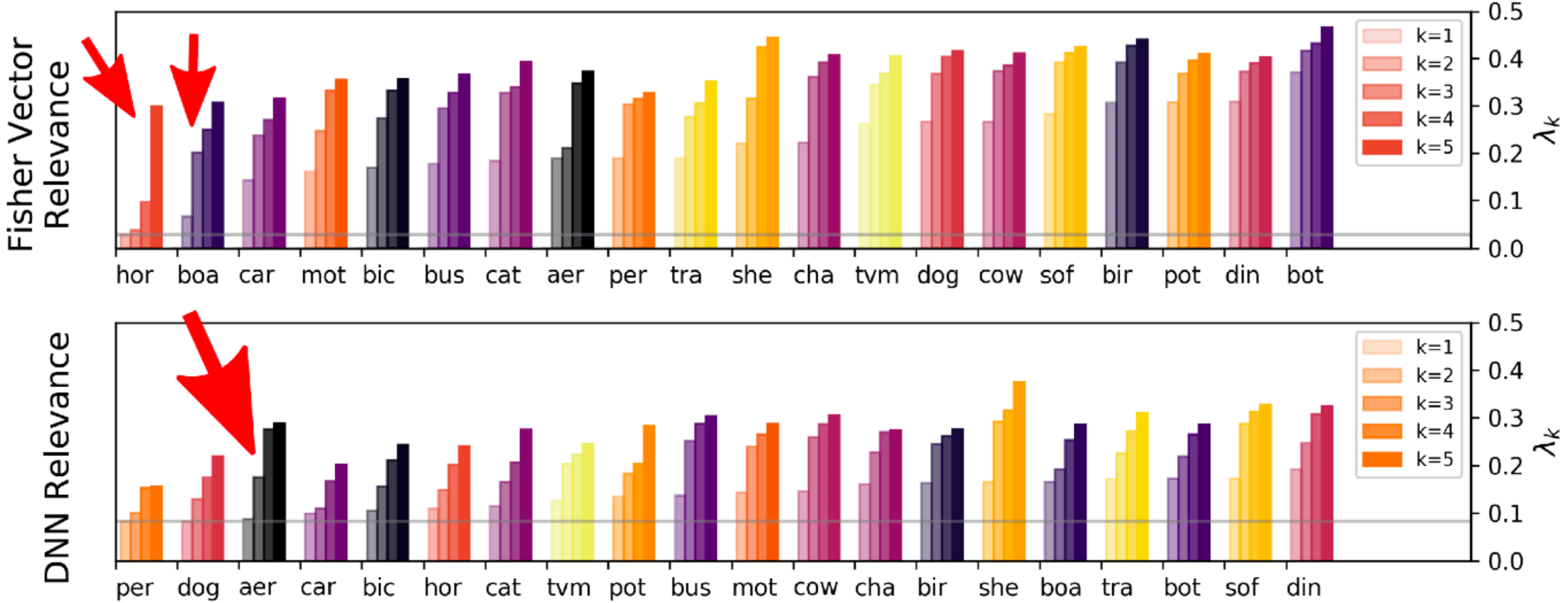


Analyze the data, *from the model's point of view*, via attribution maps[4] and Spectral Clustering[56]

(Lapuschkin et al., 2019)

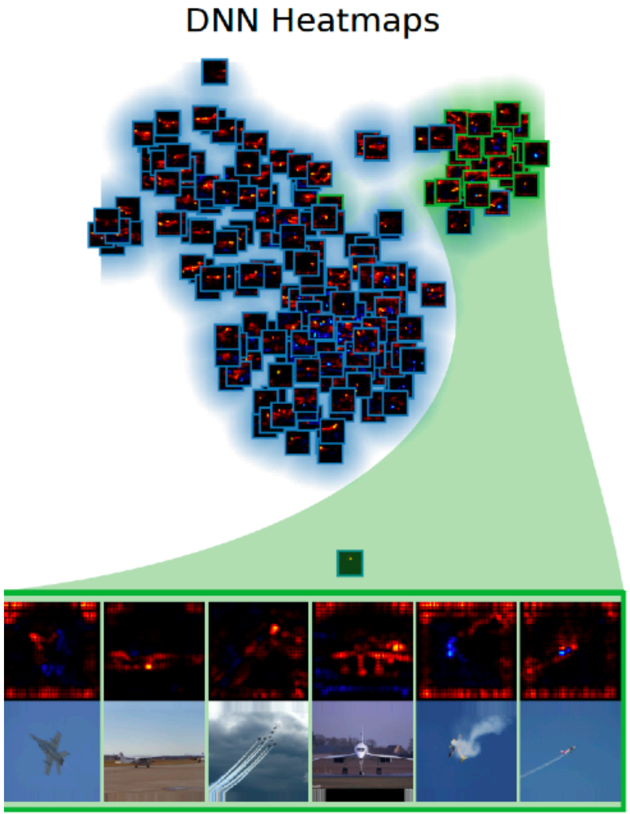
Spectral Relevance Analysis (SpRAy)

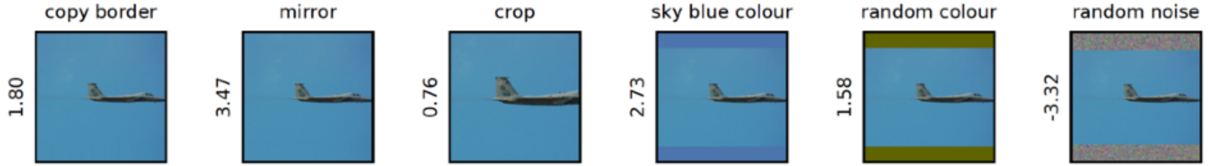
SpRAy for Fisher Vector and DNN classifiers on PASCAL VOC 2017.

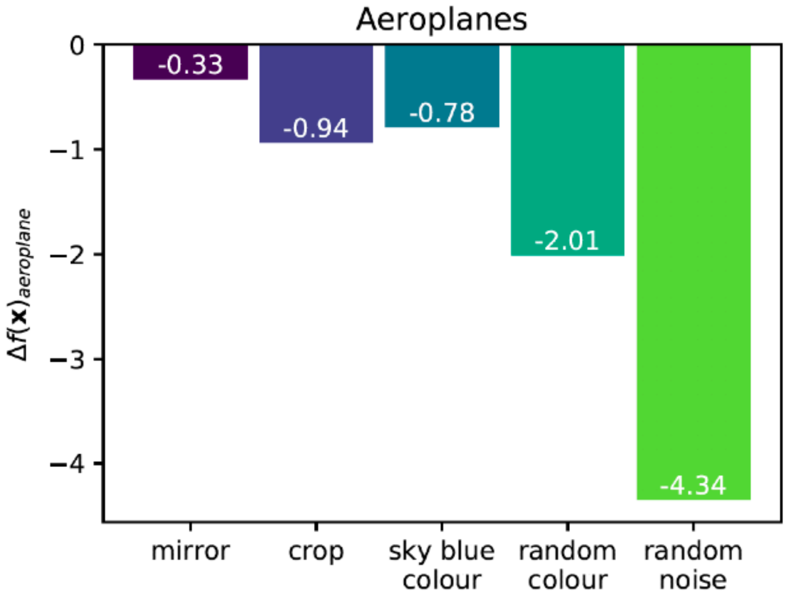


(Lapuschkin et al., 2019)

Spectral Relevance Analysis (SpRAy)



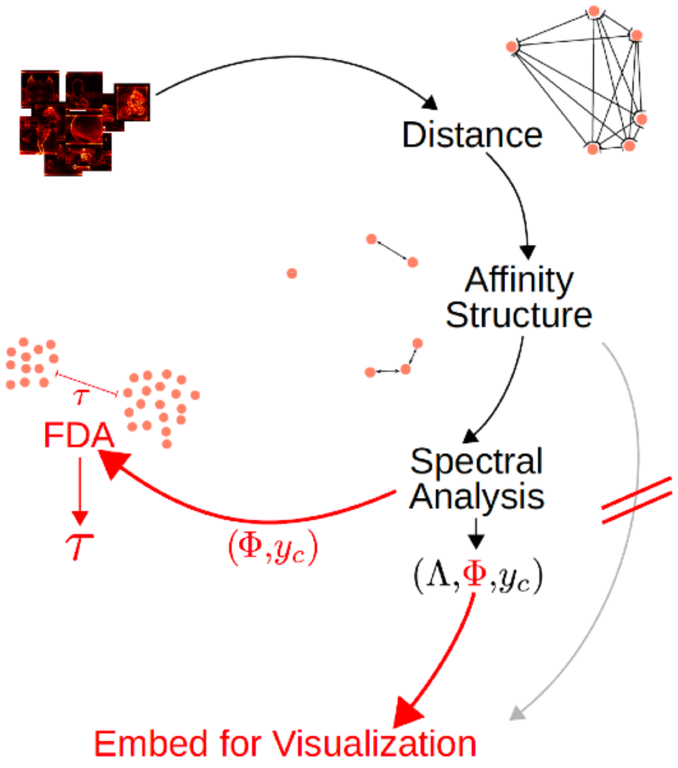




Explanation beyond visualization (Unhansing Datasets)

Anders et al. 2019

Automating Clever Hans Detection

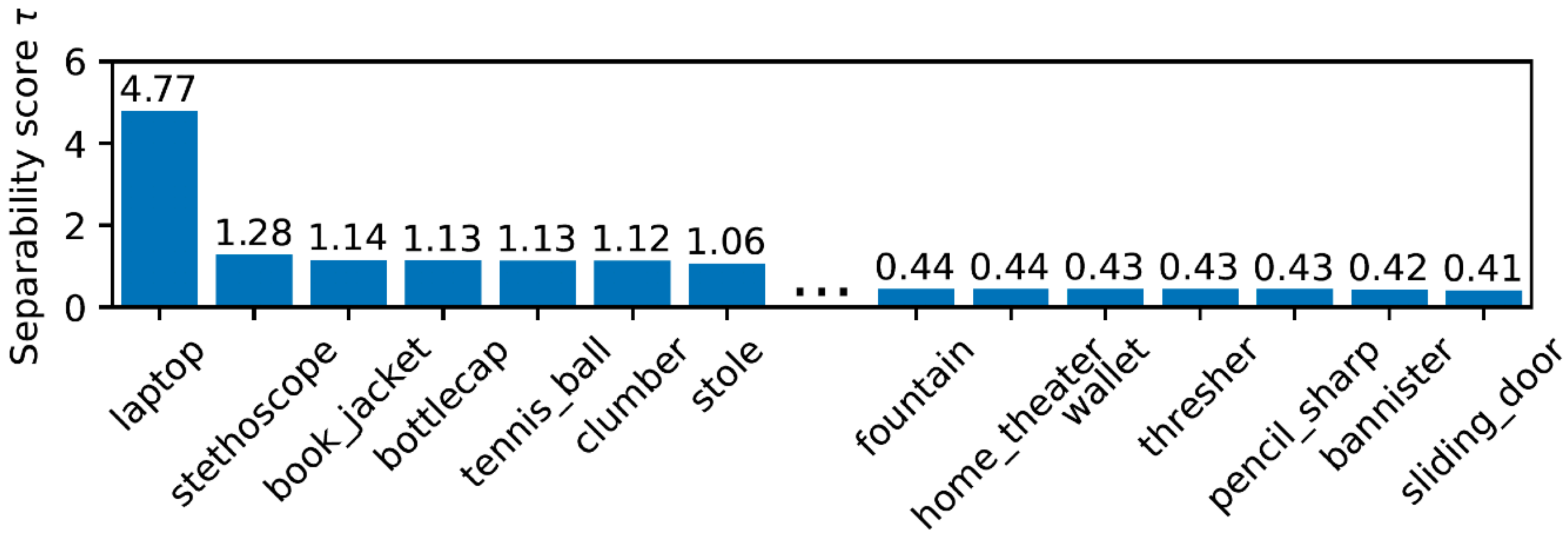


Extending SpRAy from [4]

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

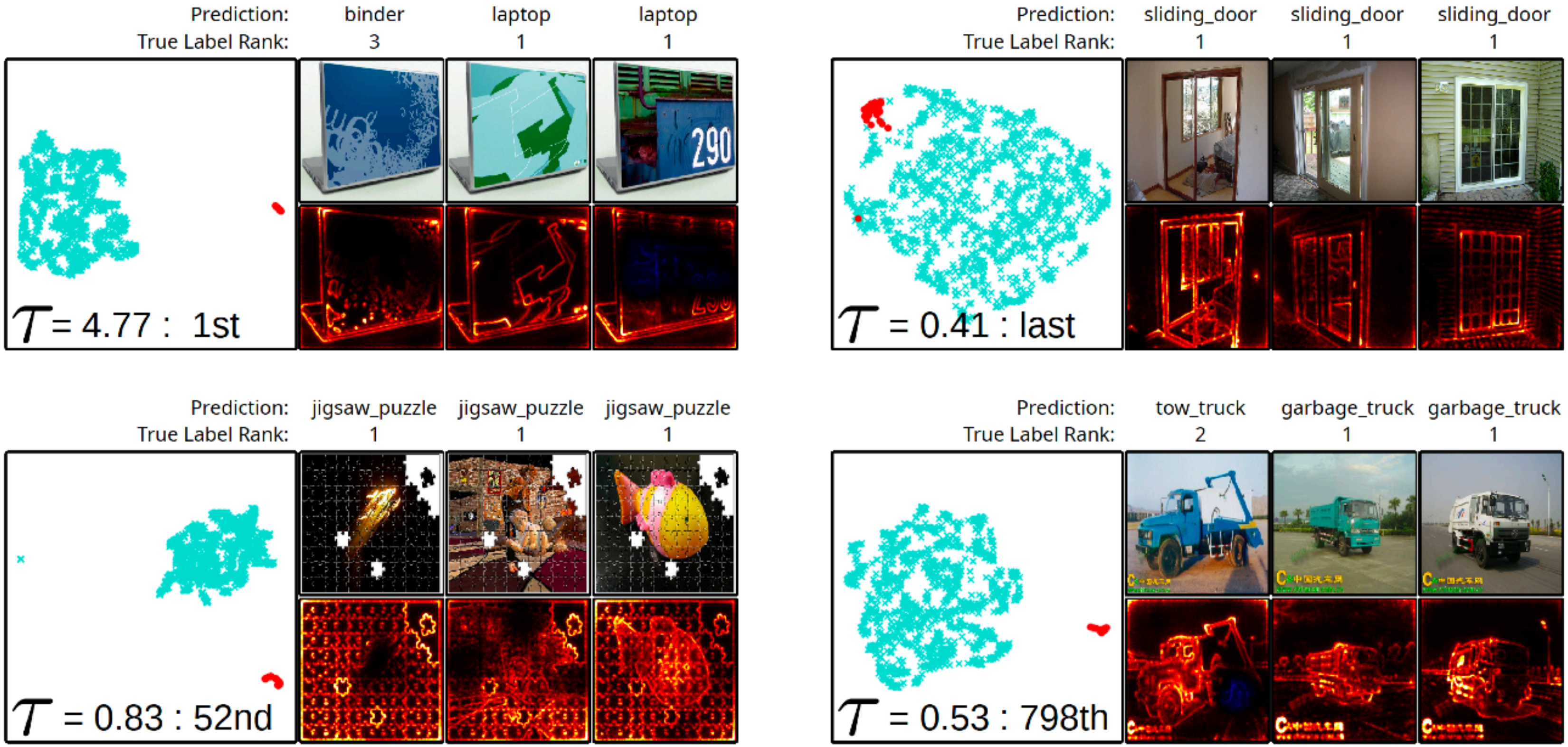
$$J(w) = \frac{w^\top S_b w}{w^\top S_w w}$$

Automating Clever Hans Detection

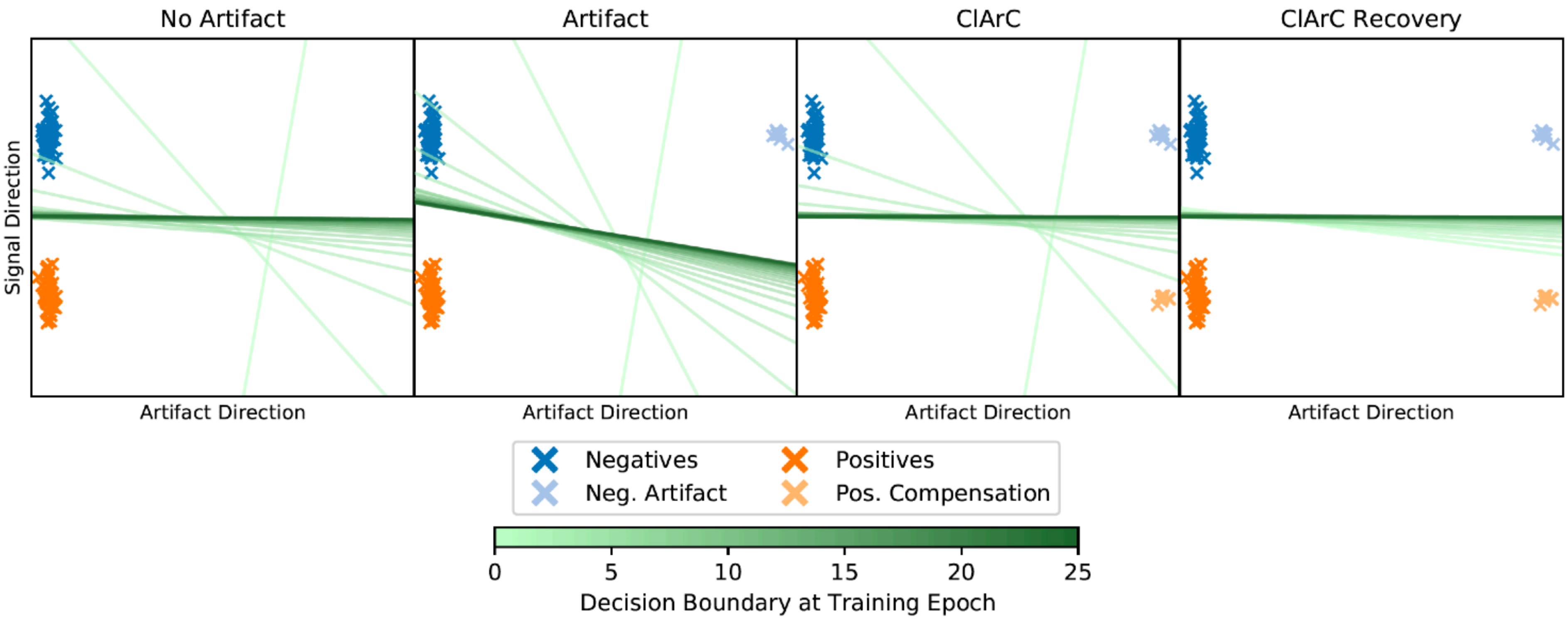


The solution of FDA can be understood as directions of maximal separability between clusterings, and, when normalized and plugged into the original objective, gives scores of separability.

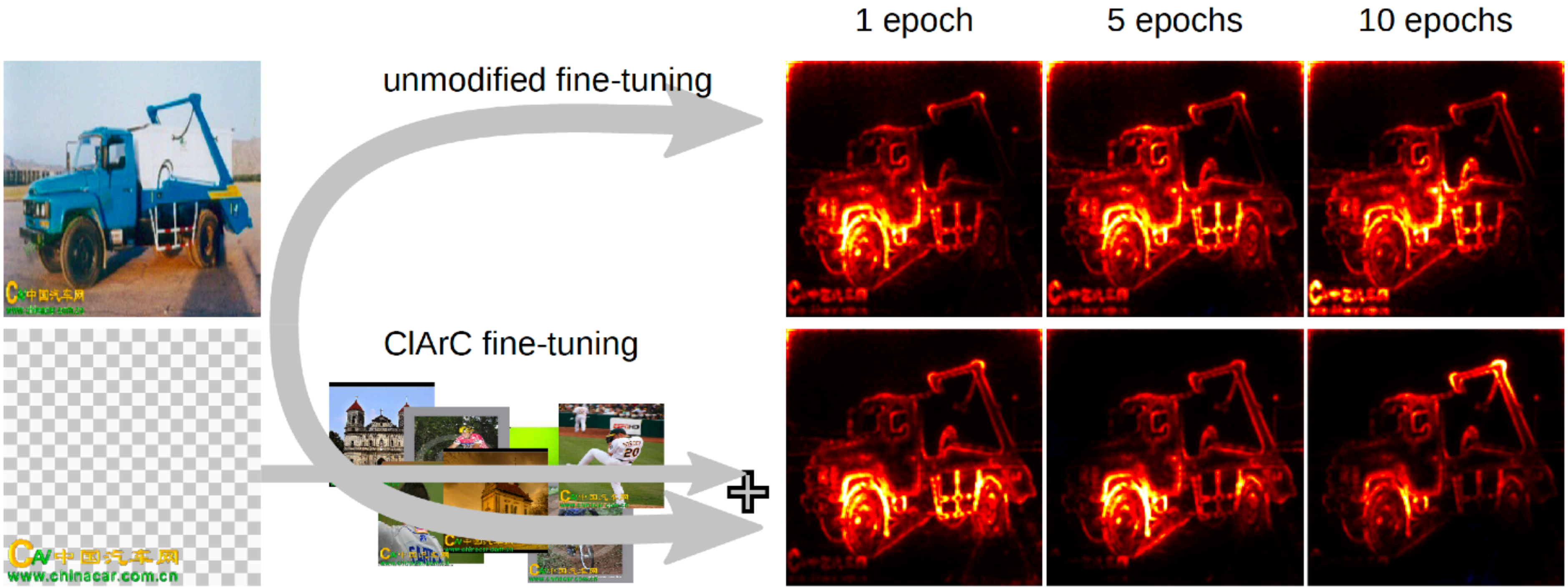
Automating Clever Hans Detection



Unhansing

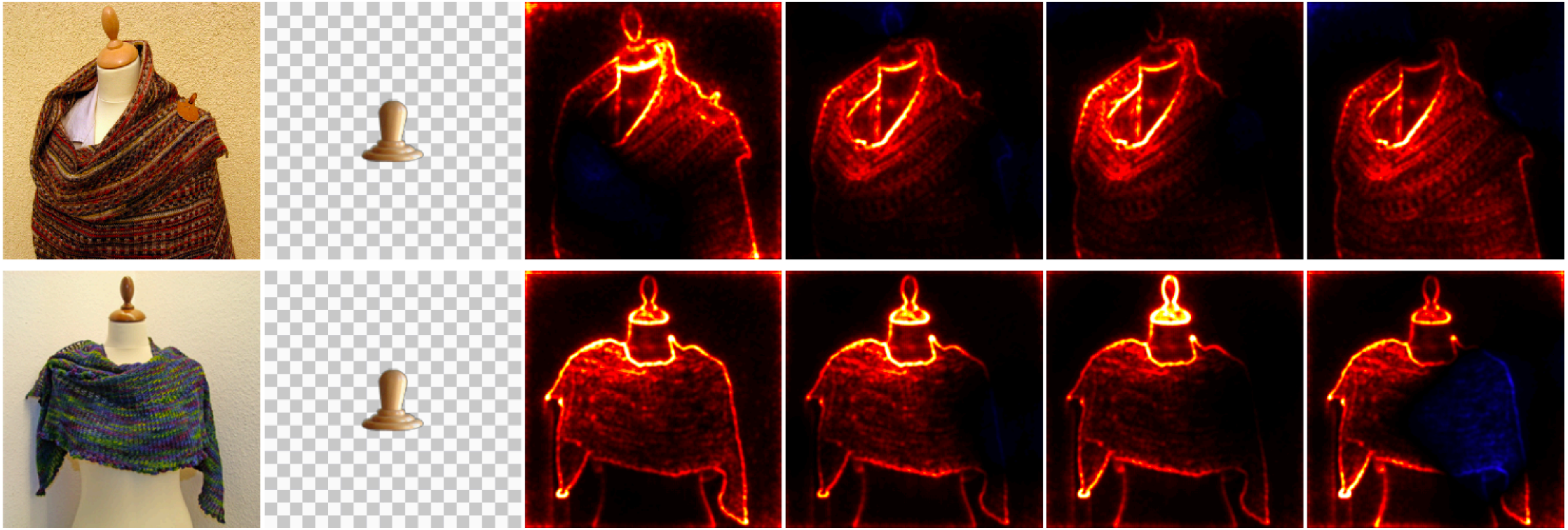

No Artifact
Artifact
ClArC
ClArC Recovery
Signal Direction
Artifact Direction
Artifact Direction
Artifact Direction
Artifact Direction
Negatives
Positives
Neg. Artifact
Pos. Compensation
0
5
10
15
20
25
Decision Boundary at Training Epoch

Unhansing

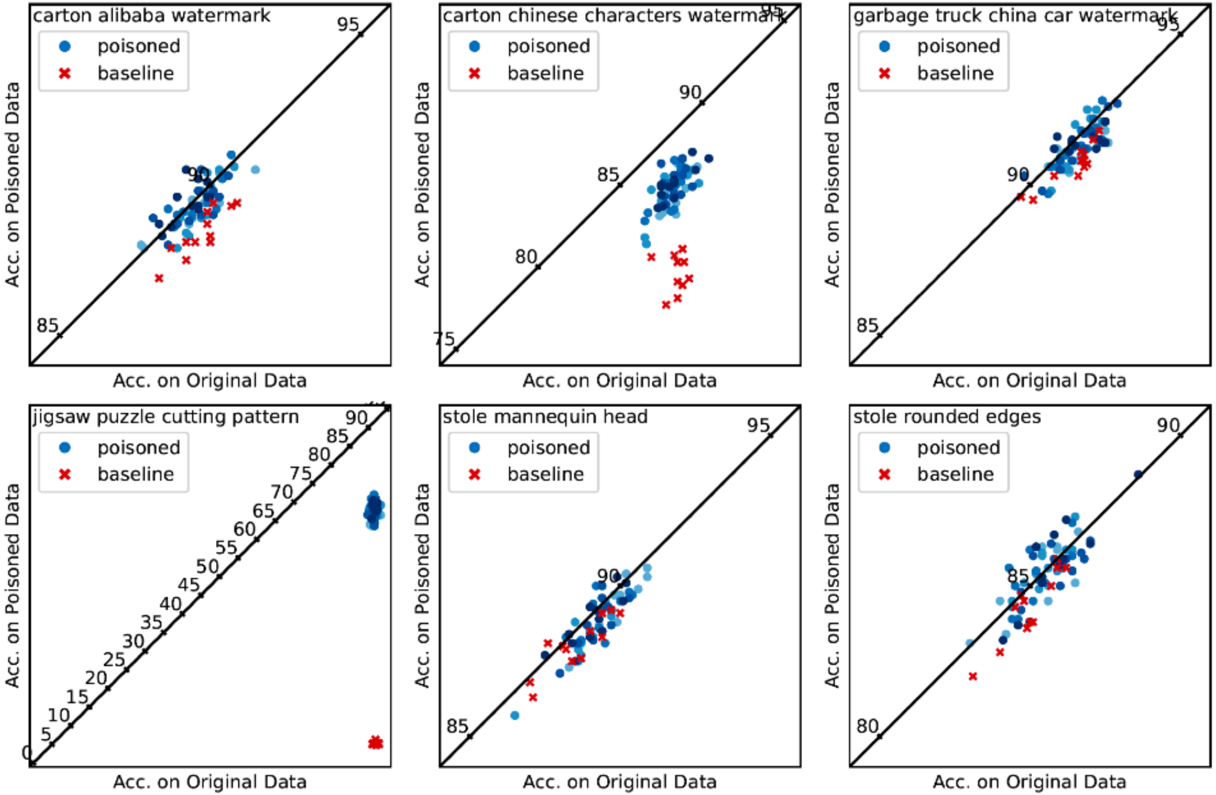


Isolate artefact, add to *other/all* classes, re-train model.

Unhansing



Unhansing



addition of artifact candidates degrades the model performance, thus validating their Clever-Hans property.

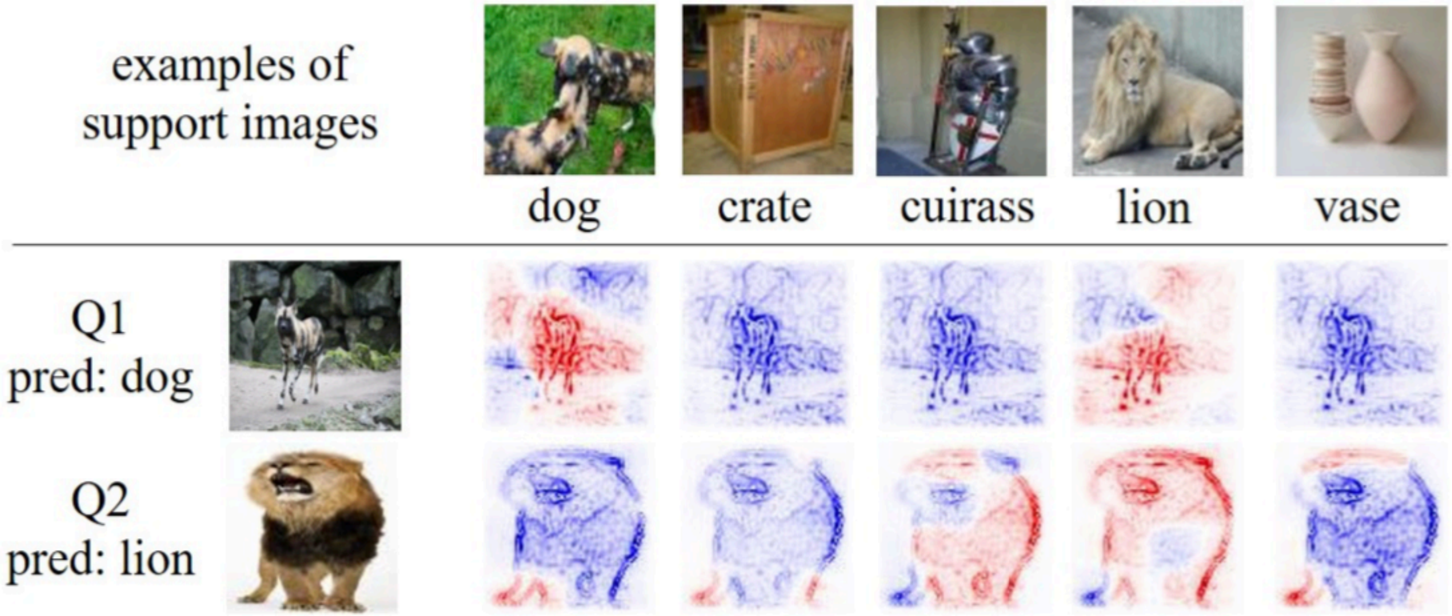
ClArC'ed models (blue) show better performance on the poisoned validation set, implying increased robustness against Clever-Hans artifacts.

Explanation beyond visualization (Explanation-Guided Training)

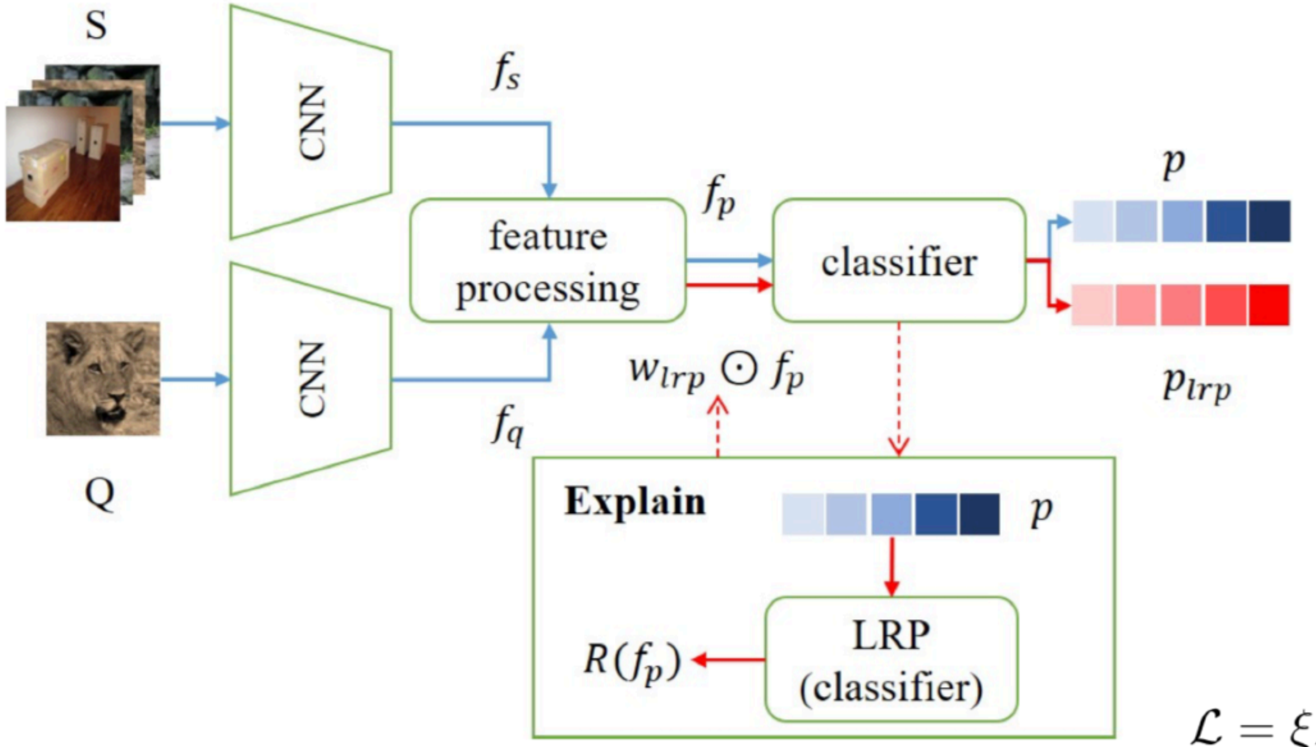
Sun et al. 2020

Explanation-Guided Training

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



Explanation-Guided Training



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

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

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

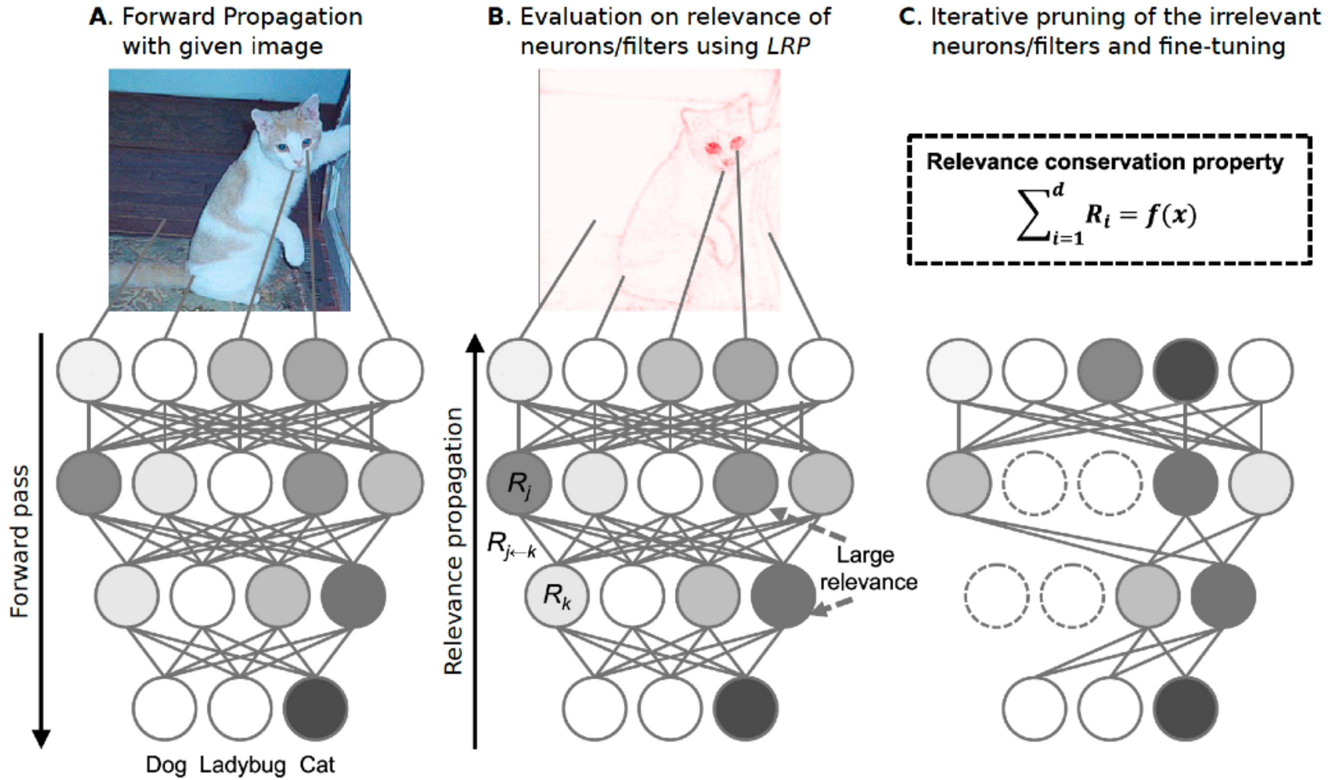
Explanation-Guided Training

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%**

Explanation beyond visualization (XAI-Based Pruning)

Yeom et al. 2019

XAI-Based Pruning

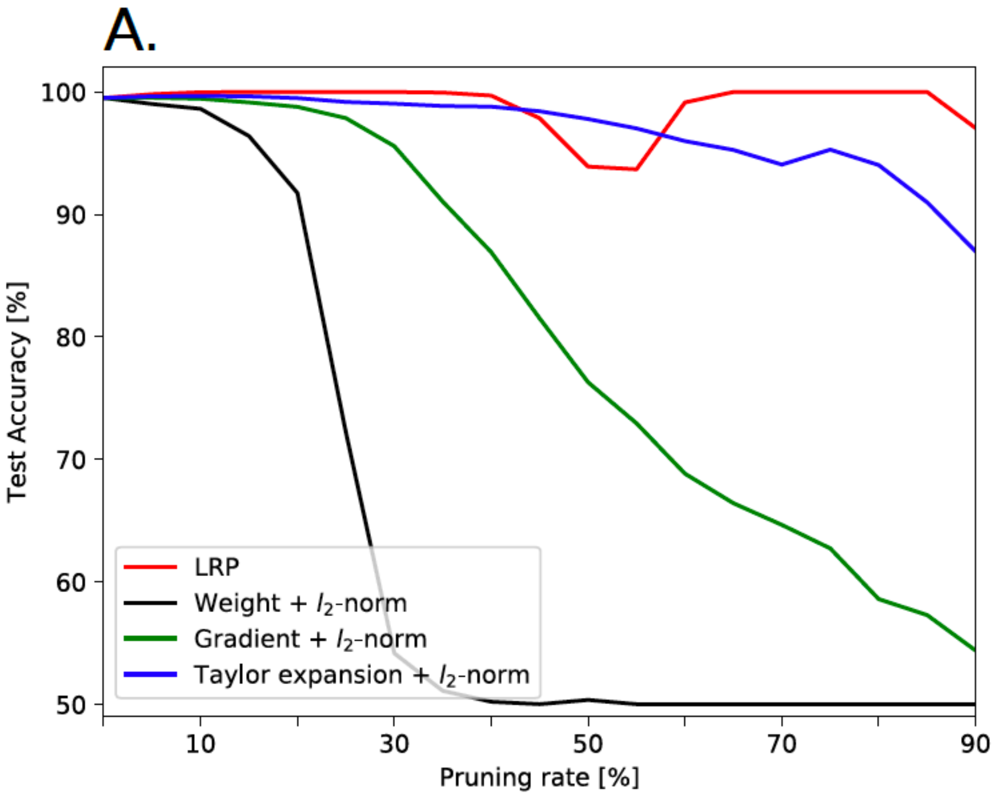


XAI-Based Pruning

Cats and Dogs	VGG-16	0.0019	99.36	60 %	119.55	15.50
	Weight	0.0050	97.90		47.47	7.02
	Taylor expansion	0.0051	97.54		51.19	3.86
	Gradient	0.0057	97.19		57.27	**3.68**
	LRP	**0.0044**	**98.24**		**43.75**	6.49
Oxford Flower 102	VGG-16	0.0369	82.26	70 %	119.96	15.50
	Weigh	0.0383	71.84		39.34	5.48
	Taylor expansion	0.0327	72.11		41.37	**2.38**
	Gradient	0.0354	70.53		42.68	2.45
	LRP	**0.0296**	**74.59**		**37.54**	4.50
Cifar 10	VGG-16	0.0157	91.04	30 %	119.59	15.50
	Weight	0.0183	93.36		**74.55**	11.70
	Taylor expansion	0.0176	93.29		97.30	**8.14**
	Gradient	0.0180	93.05		97.33	8.24
	LRP	**0.0171**	**93.42**		89.20	9.93

With fine-tuning

XAI-Based Pruning

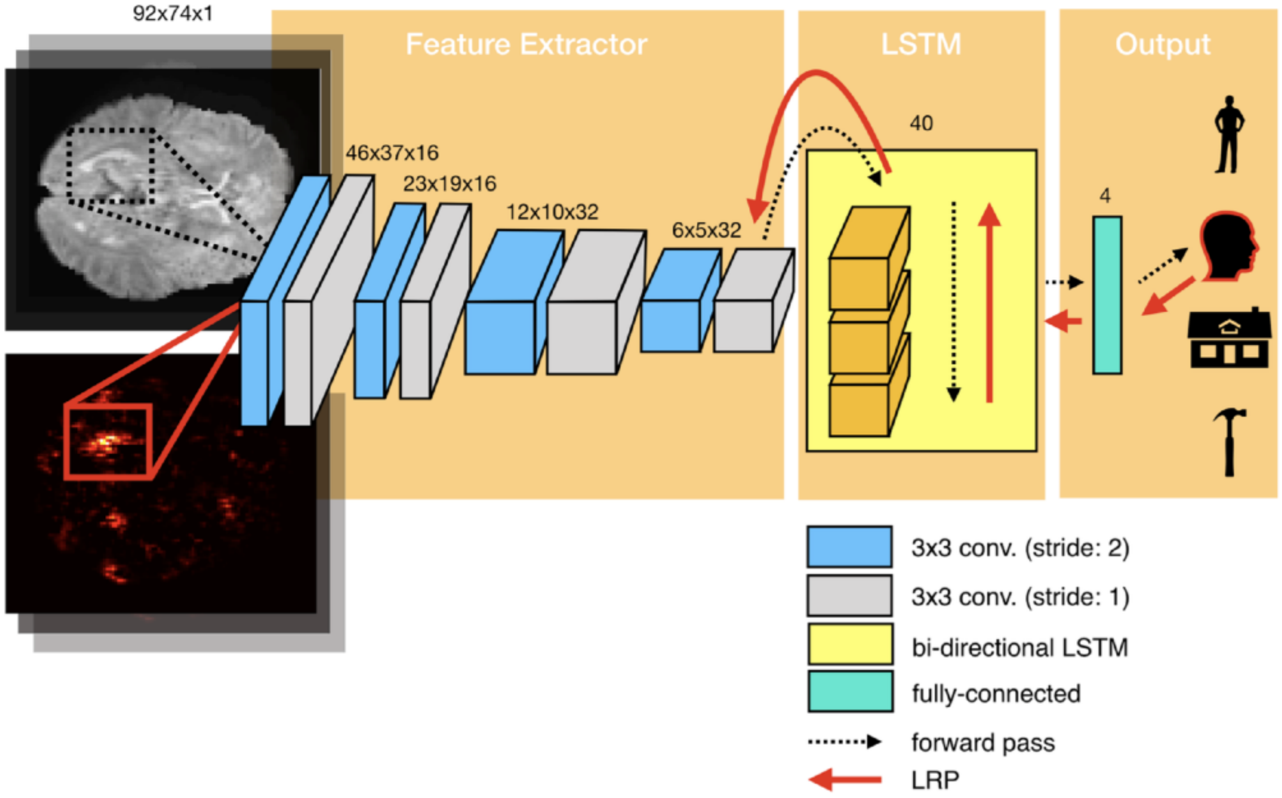


No fine-tuning

only 10 samples per class
(domain adaptation scenario)

XAI in the Sciences

XAI in the Sciences

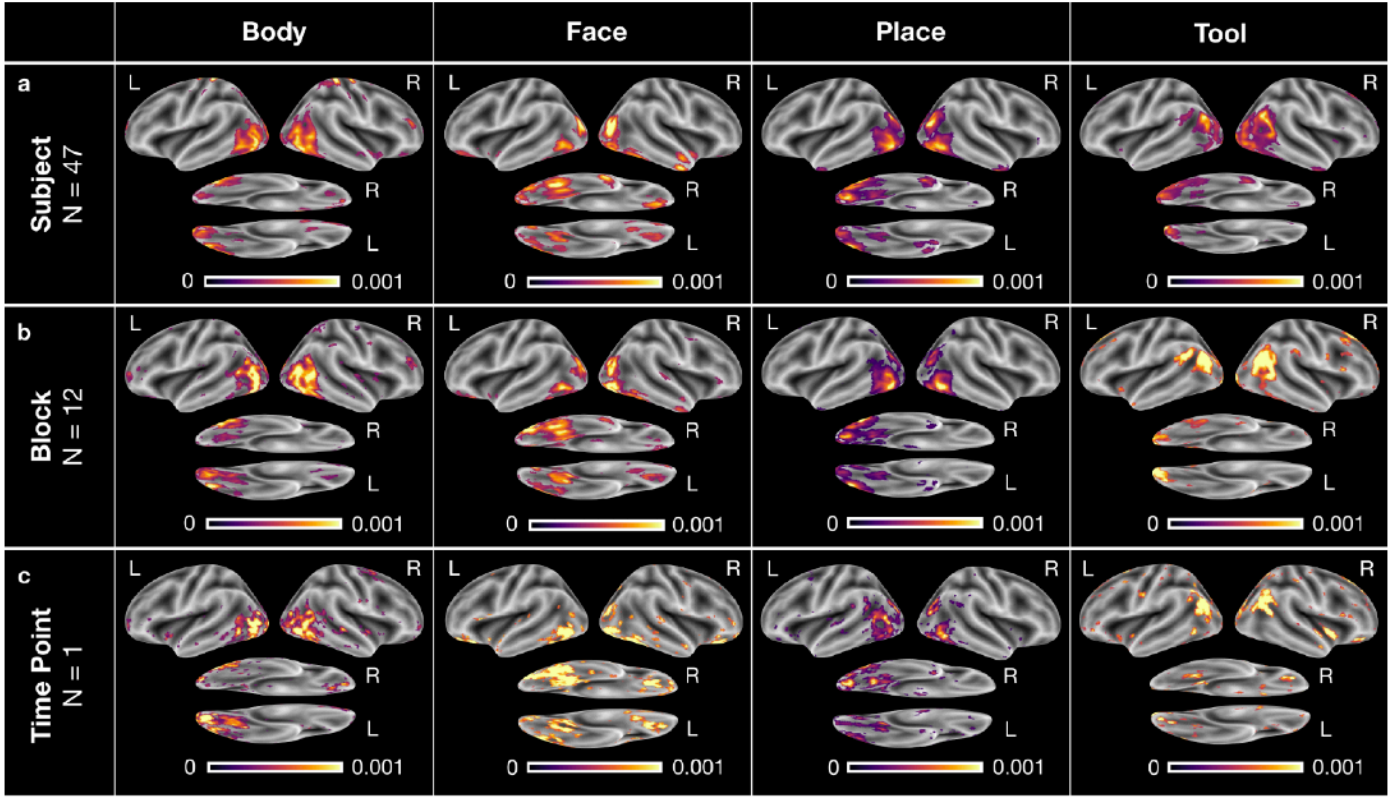


Our approach:

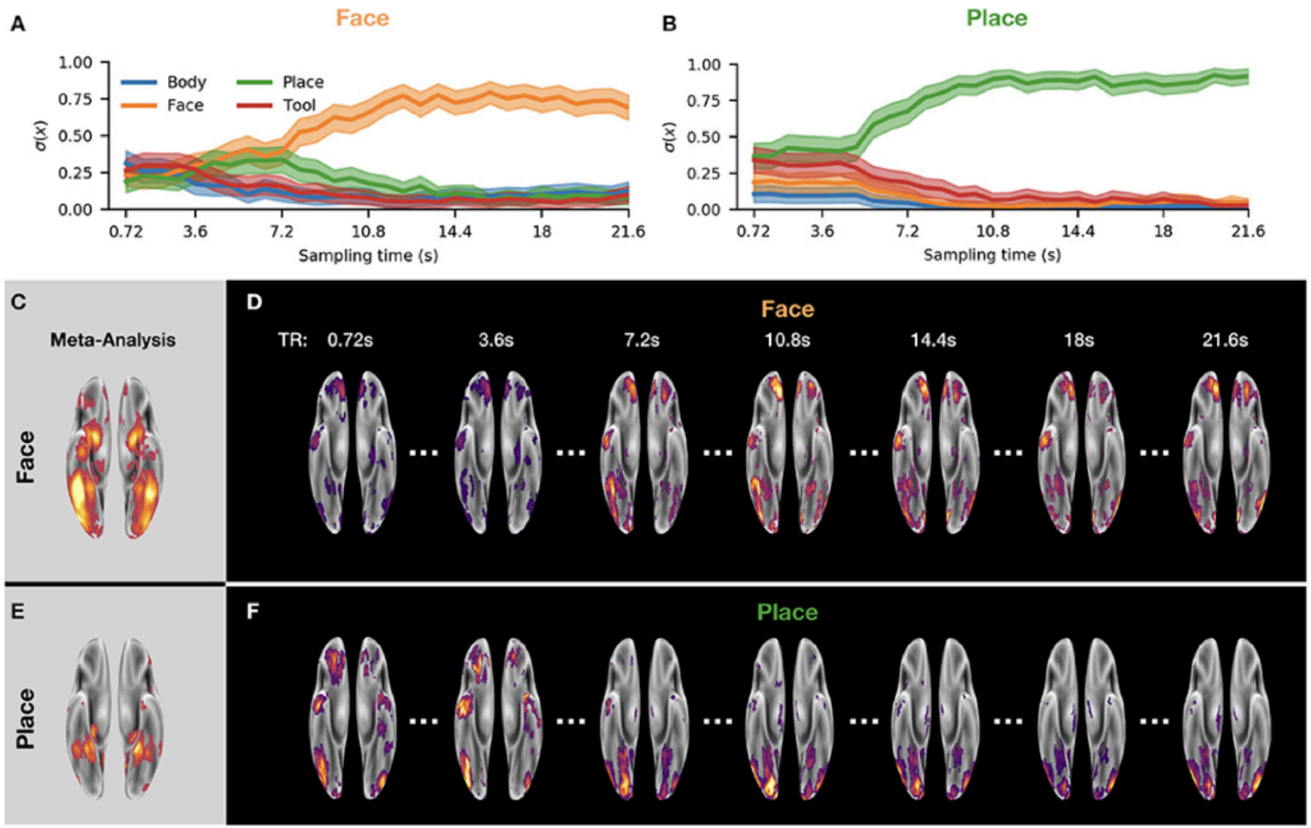
- Recurrent neural networks (CNN + LSTM) for whole-brain analysis
- LRP allows to interpret the results

(Thomas et al. 2019)

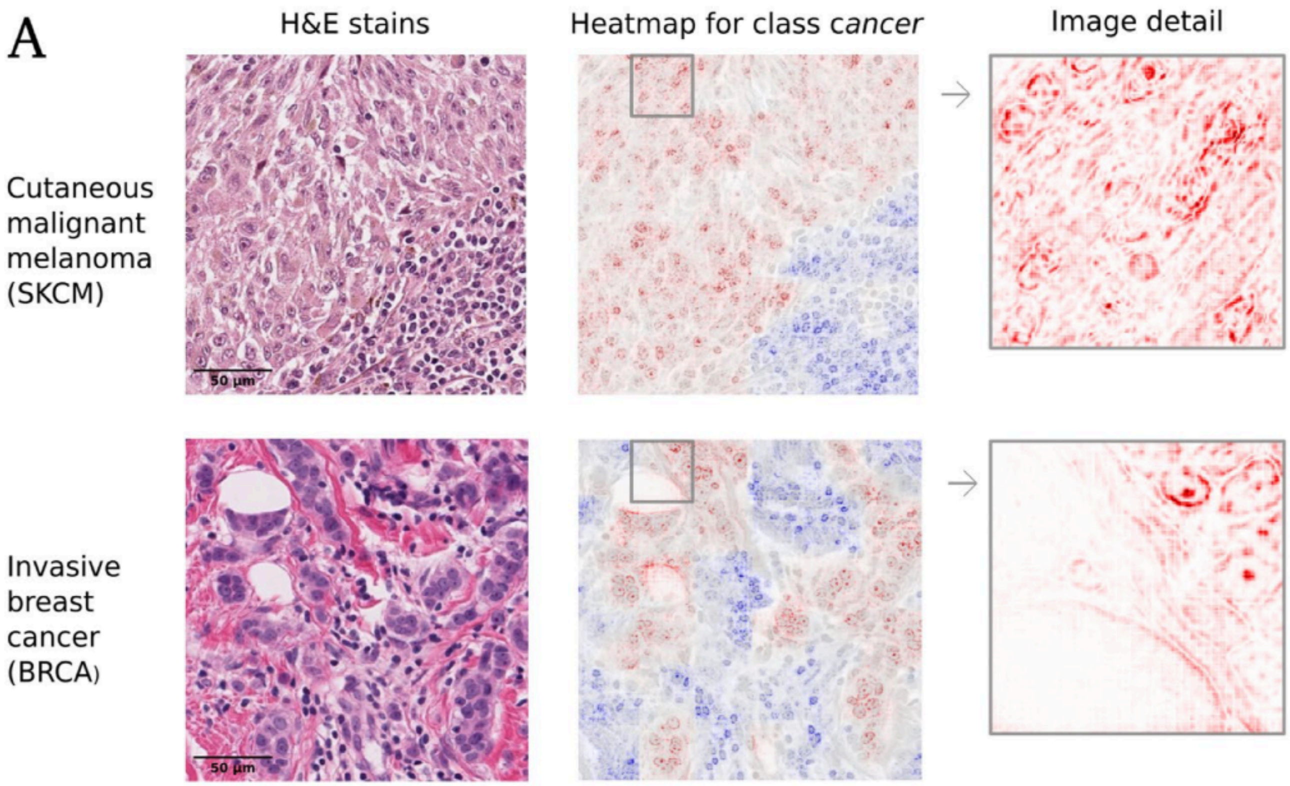
XAI in the Sciences



XAI in the Sciences



XAI in the Sciences

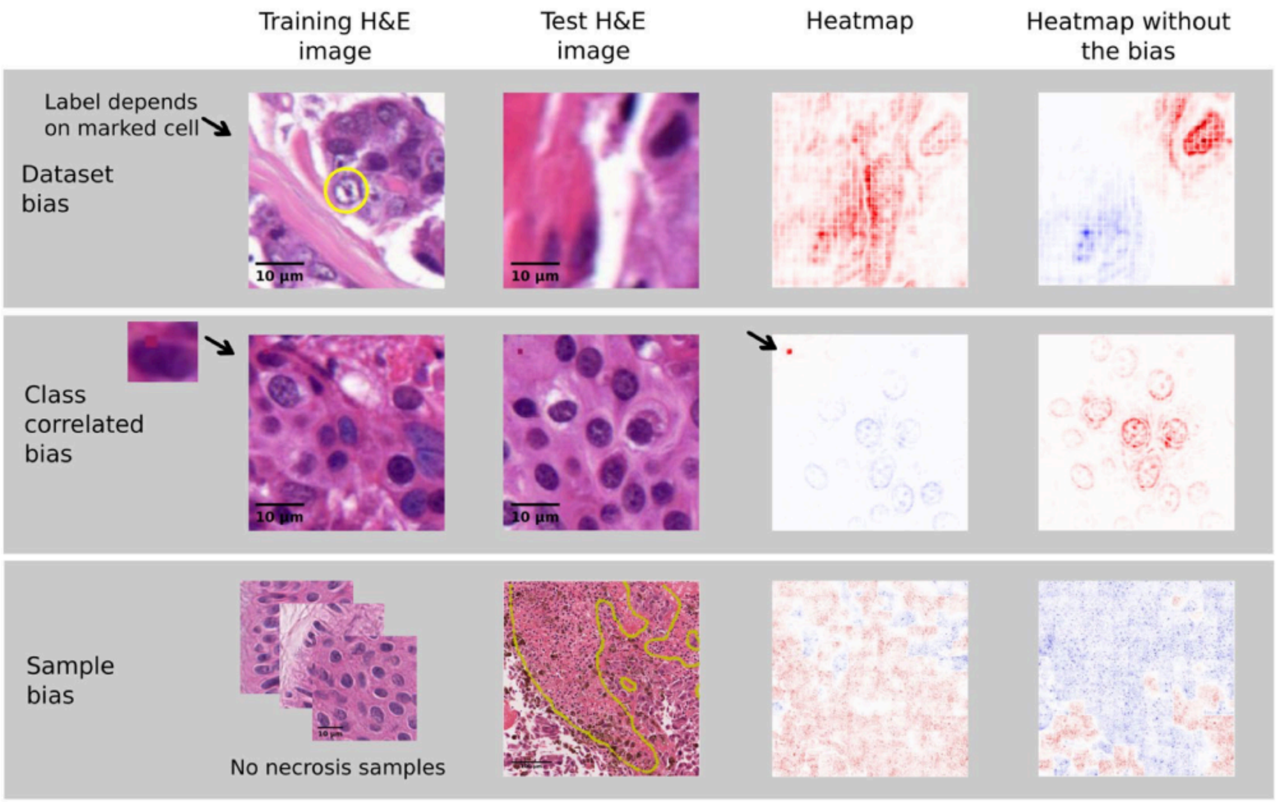


Hägele et al., 2020

XAI in the Sciences

Experiments	Description	Heatmaps	Benefit of visual explanation	Lifecycle phase
Feature visualisation	Tumour classification in various entities (BRCA, SKCM & LUAD)	Fig. 1	Visual and quantitative verification of learned features on cell level	Deployment phase (e.g. computer-aided diagnosis systems)
Class sampling ratio	Different class sampling ratios in mini-batches	Fig. 3	Deliberate manipulation for different application use cases (contrary effects of recall and precision)	Deployment phase
Dataset bias	Label bias affecting entire datasets	Fig. 4 (top)	Bias detectable on a single sample, no additional held-out data necessary	Development phase
"Class correlated" bias	Artificial corruption correlated with one class label	Fig. 4 (middle)	Bias detectable on a single sample, possible to detect very small artefacts	Development phase
Sample bias	Exclusion of a tissue type in the training data (here: necrosis)	Fig. 4 (bottom)	Bias detectable on few samples of the missing tissue type, small regions of the missing tissue type also precisely detectable	Development phase (i. e. iterative process to create comprehensive dataset)

XAI in the Sciences

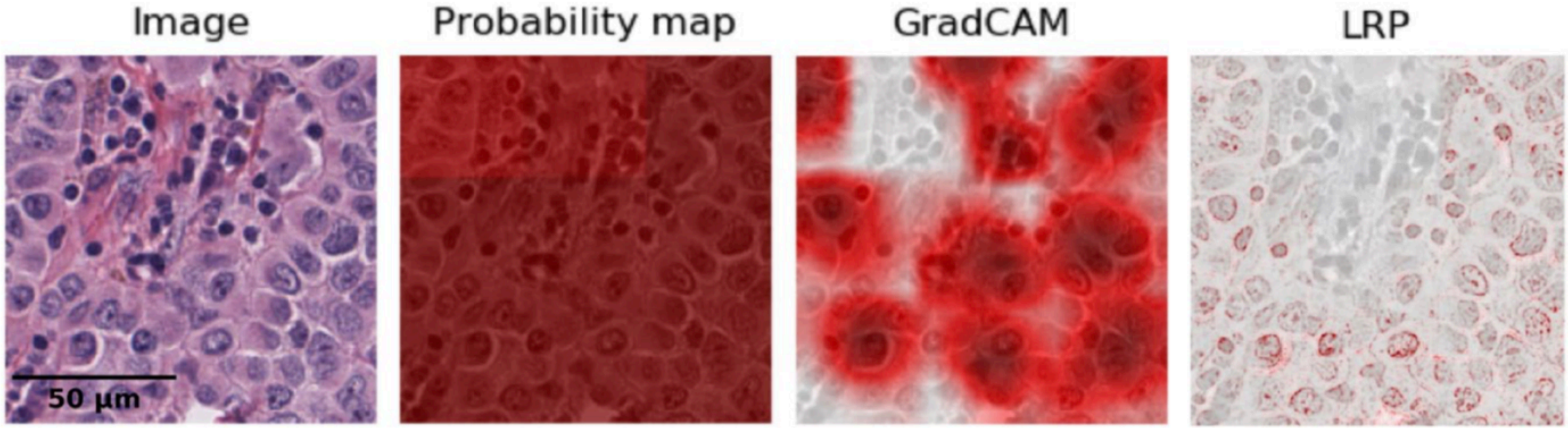


determining the label solely from the patch’s centre cell (yellow mark)

small artificial corruption

training a classifier on a dataset lacking examples of necrosis

XAI in the Sciences



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

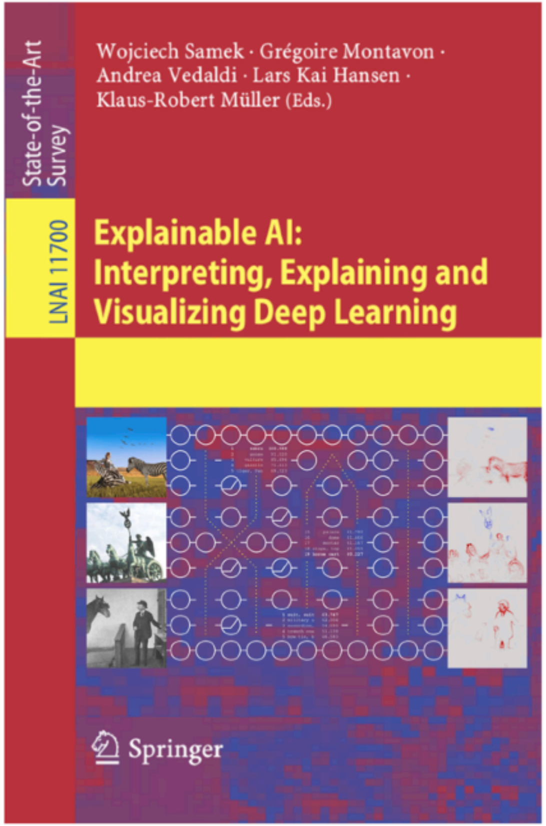
Application to Neural Network Pruning

- 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