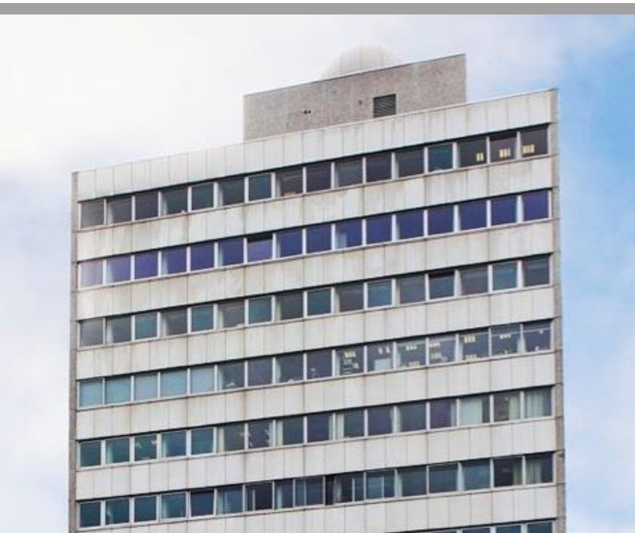
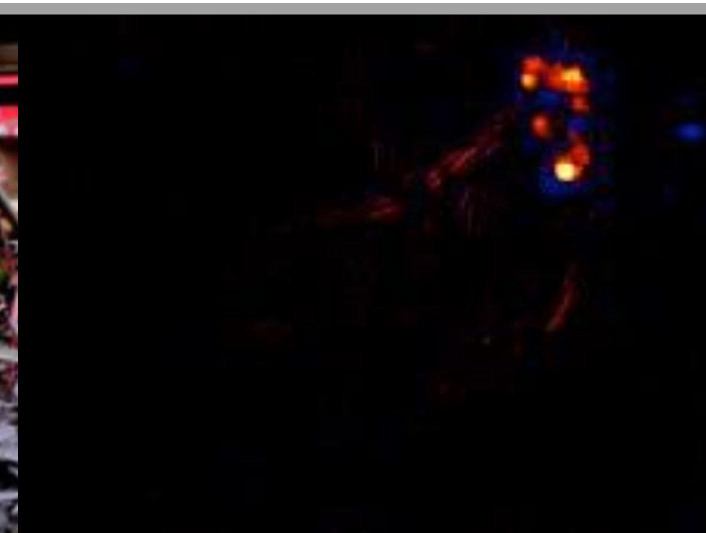
Fraunhofer
Heinrich Hertz Institute

Meta-Explanations, Interpretable Clustering & Other Recent Developments

Fraunhofer HHI, Machine Learning Group

Wojciech Samek





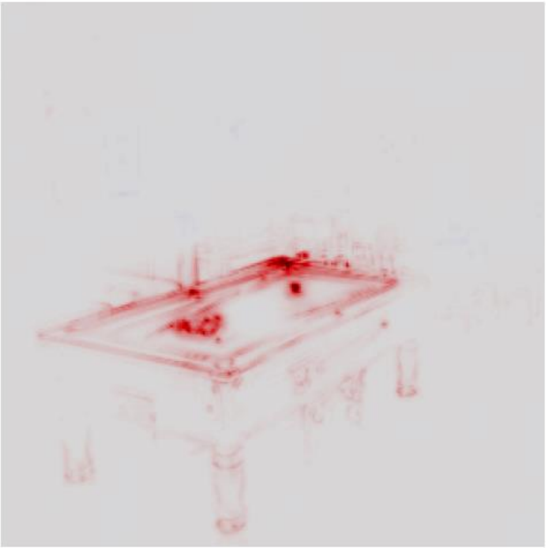
ICCV 2019 Visual XAI Workshop

Seoul, Korea, 2th November 2019

Explaining Predictions

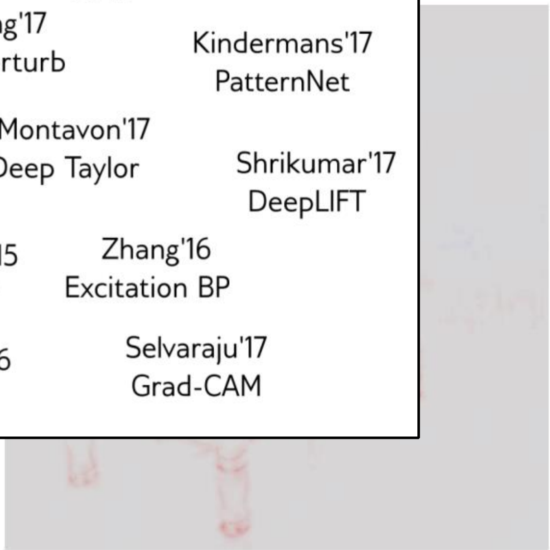
"why a given image is classified as a pool table"

some pool table



why it is classified as a pool table

Today’s Talk



Which one to choose ?

Baehrens'10 Gradient

Sundarajan'17 Int Grad

Zintgraf'17 Pred Diff

Ribeiro'16 LIME

Haufe'15 Pattern

Zurada'94 Gradient

Symonian'13 Gradient

Zeiler'14 Occlusions

Fong'17 M Perturb

Kindermans'17 PatternNet

Poulin'06 Additive

Lundberg'17 Shapley

Bazen'13 Taylor

Montavon'17 Deep Taylor

Shrikumar'17 DeepLIFT

Zeiler'14 Deconv

Landecker'13 Contrib Prop

Bach'15 LRP

Zhang'16 Excitation BP

Caruana'15 Fitted Additive

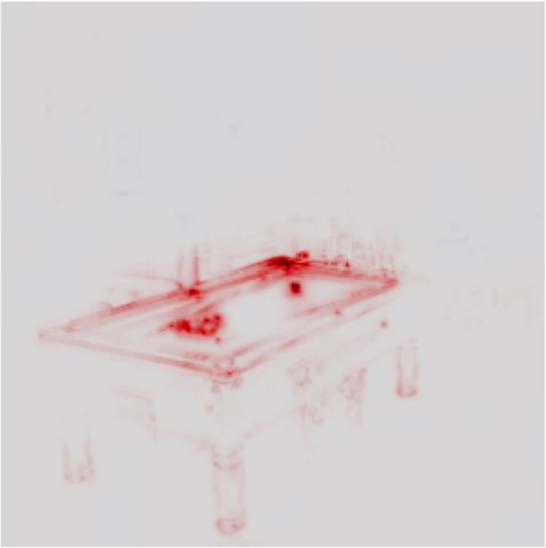
Springenberg'14 Guided BP

Zhou'16 GAP

Selvaraju'17 Grad-CAM

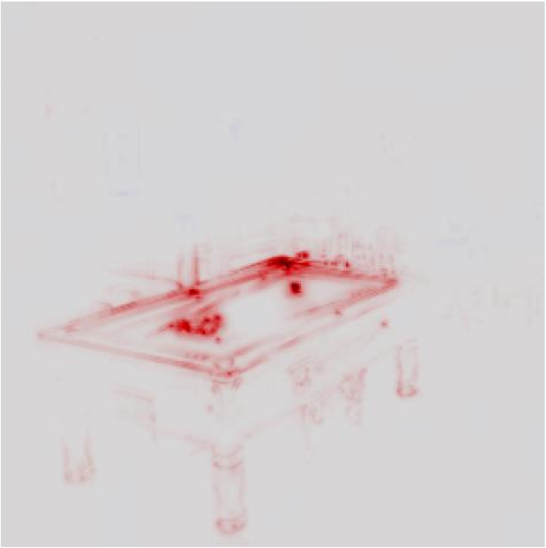
Today's Talk

From individual explanations to common prediction strategies



Today's Talk

What can we do with it ?



Today's Talk

Explaining more than classifiers

Explanation Methods

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)

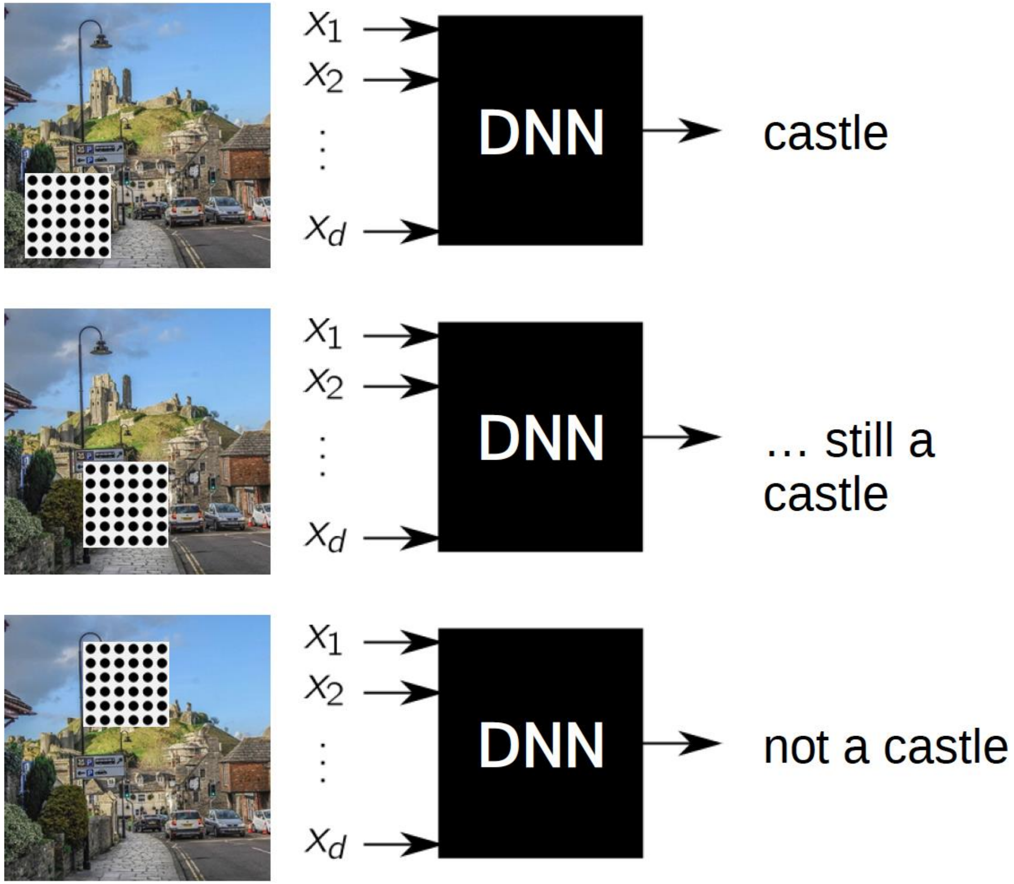
Deep Taylor Decomposition (Montavon et al. 17)

Excitation Backprop (Zhang et al. 16)

…

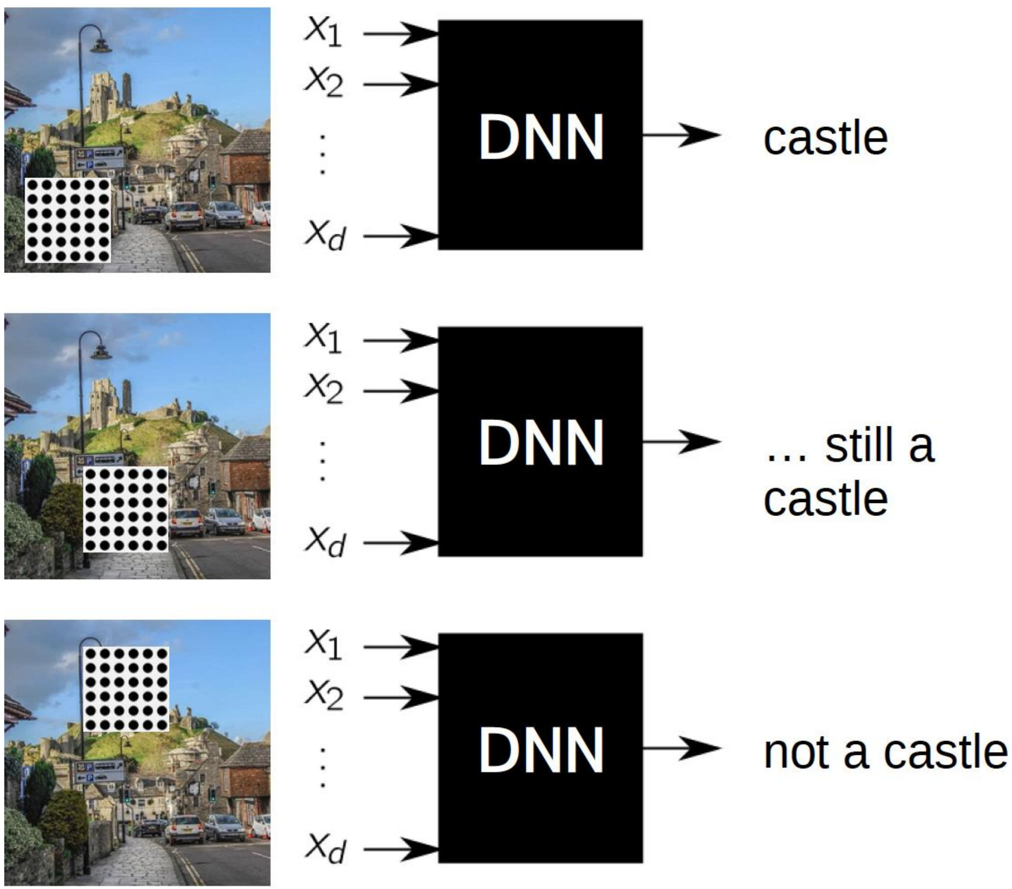
Approach 1: Perturbation

Idea: Assess features relevance by testing the model response to their removal or perturbation.



Approach 1: Perturbation

Idea: Assess features relevance by testing the model response to their removal or perturbation.



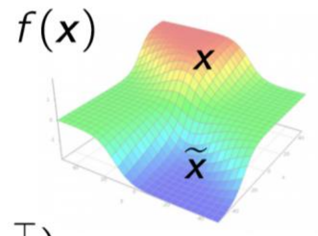
Disadvantages

- slow
- assumes locality
- perturbation may introduce artefacts

—> unreliable

Approach 2: (Simple) Taylor Expansions

Idea: identify the contribution of input features as the first-order terms of a Taylor expansion

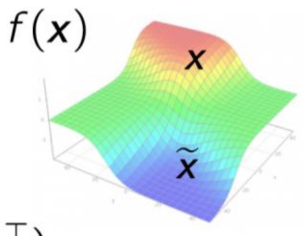


Taylor Expansion

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

Approach 2: (Simple) Taylor Expansions

Idea: identify the contribution of input features as the first-order terms of a Taylor expansion



Taylor Expansion

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

Advantages

- Can be applied to *any* (differentiable and mildly nonlinear) ML model.

Limitations

- Need to find a meaningful root point where to perform the expansion.

Approach 3: Gradient x Input

Motivation

- Compute an explanation in a single pass without having to optimize or search for a root point.

Gradient x Input

$$\forall_i : \; R_i = [\nabla f(\boldsymbol{x})]_i \cdot x_i$$

$$\boldsymbol{R} = \nabla f(\boldsymbol{x}) \odot \boldsymbol{x}$$

Observation: Complex analyses reduce to gradient x input for simple cases.

Approach 3: Gradient x Input

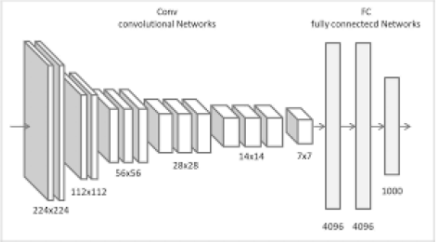
Input

x

Prediction

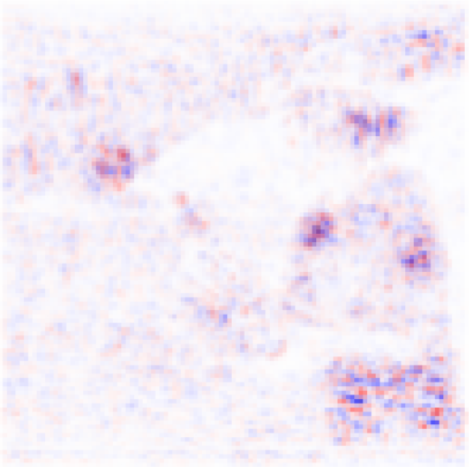
(class: baseball)

x



$f(x)$

Explanation



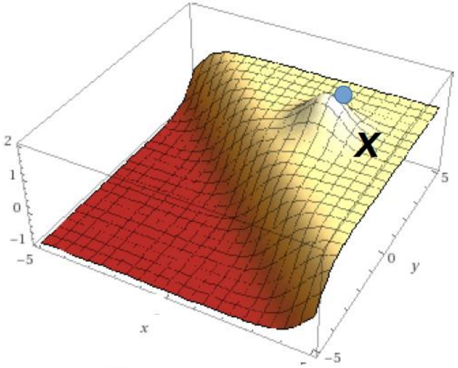
$$R = \nabla f(x) \odot x$$

Observation: Explanations are noise

Two reasons why gradient-based explanation are noisy

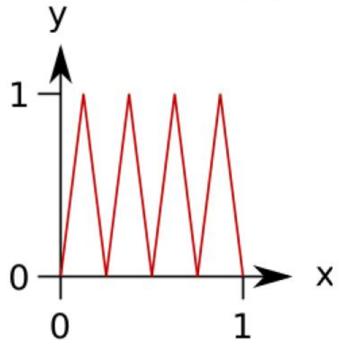
1. **Local vs. global variations**

   Global effects are not visible when looking at the function f(x) locally.



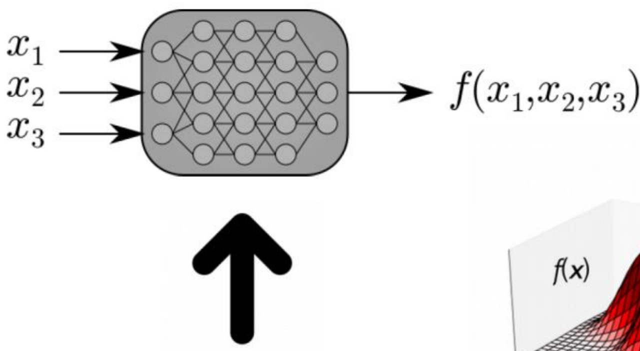
2. **Shattered gradients**

   Function local variations grows exponentially with depth.

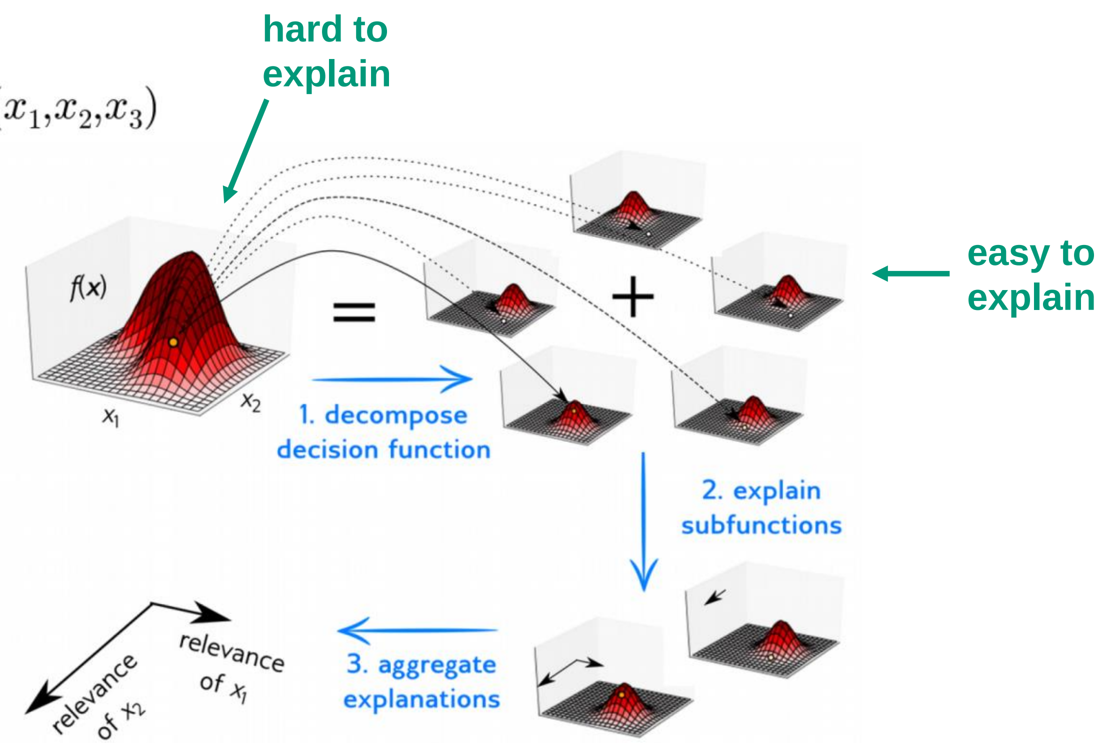


Layer-wise Relevance Propagation

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

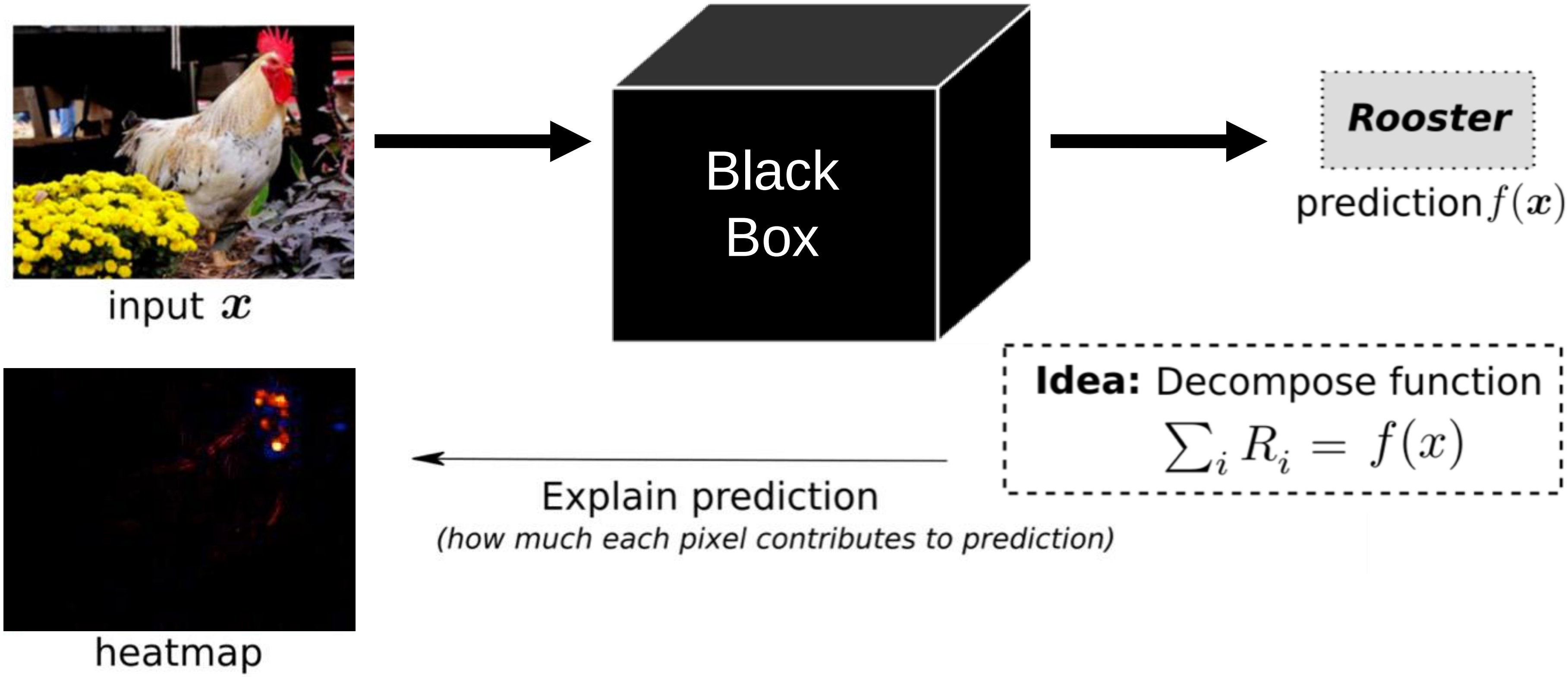


model is a composition of neurons. This can be exploited to make explanation easier.



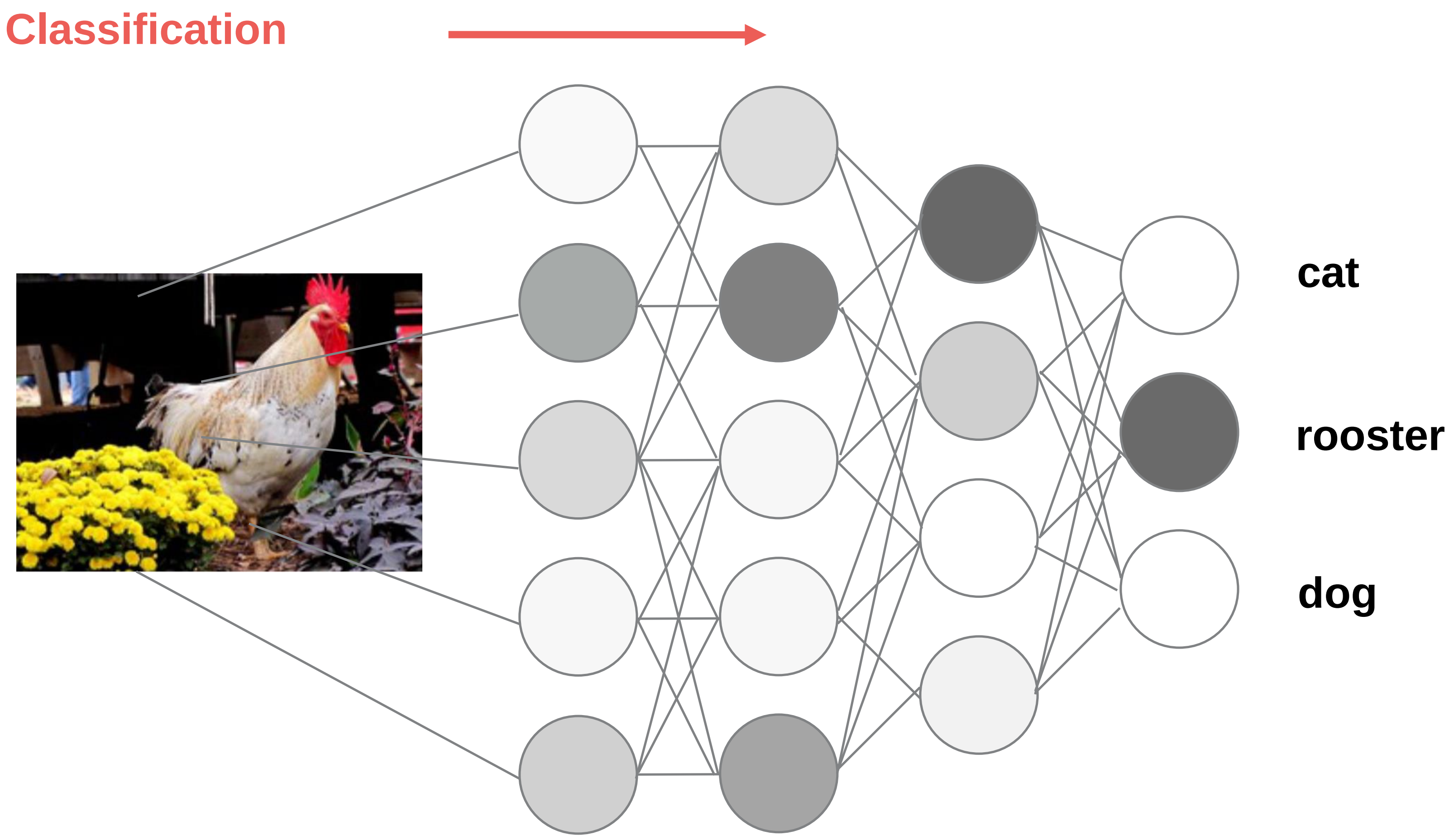
(Bach et al., 2015
Montavon et al. 2017)

Layer-wise Relevance Propagation

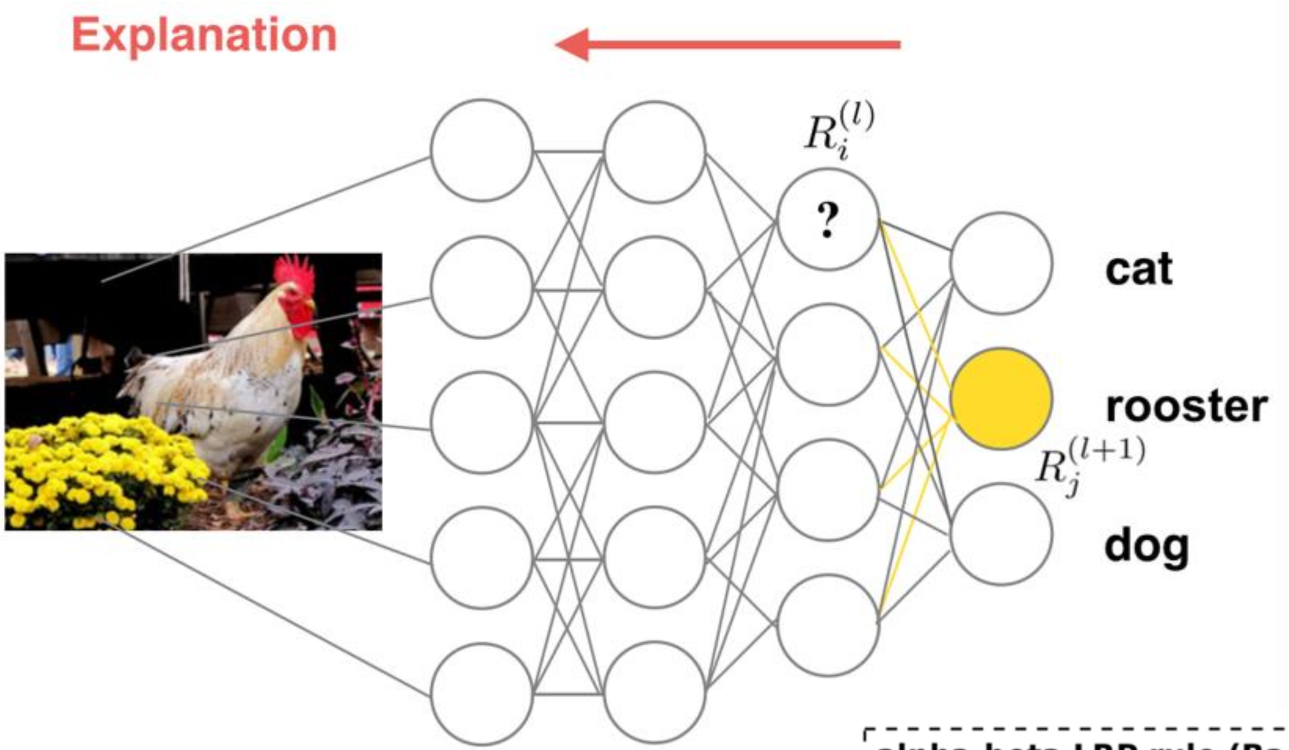


Layer-wise Relevance Propagation (LRP)
(Bach et al., PLOS ONE, 2015)

Layer-wise Relevance Propagation



Layer-wise Relevance Propagation



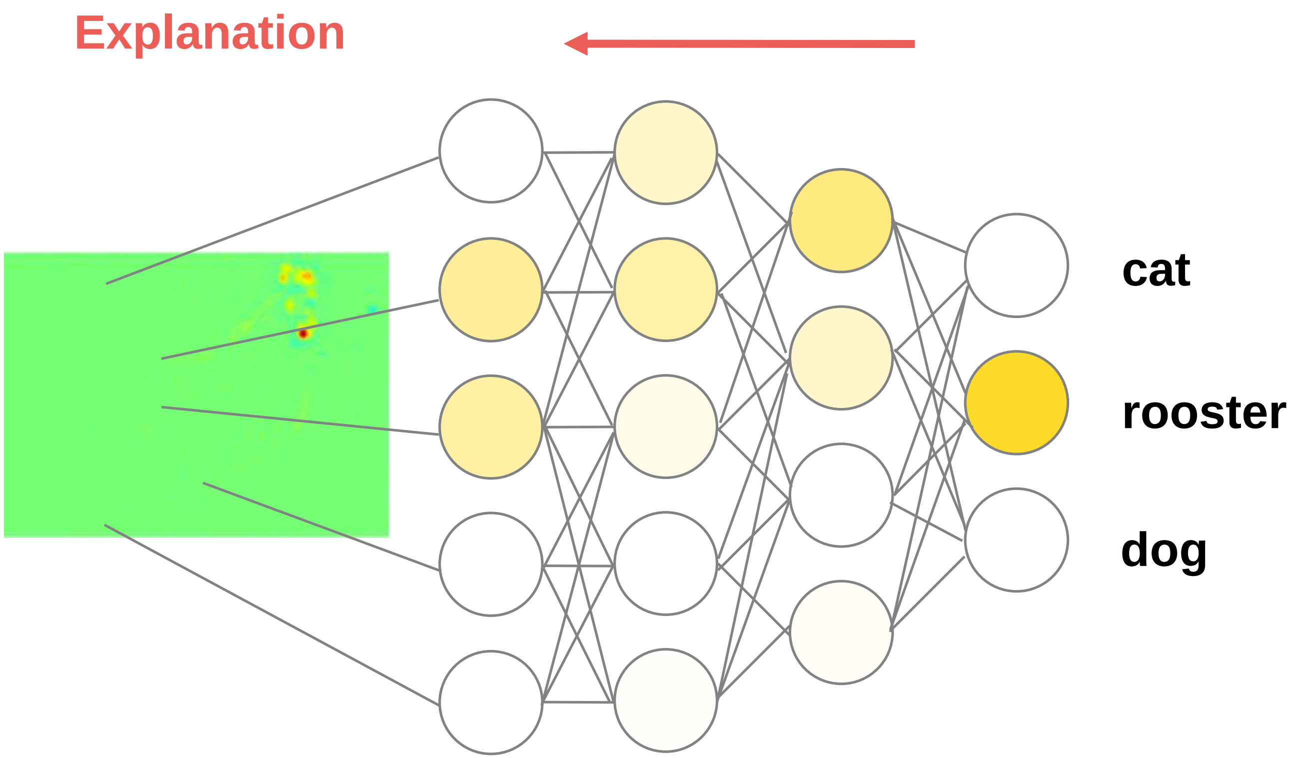
alpha-beta LRP rule (Bach et al. 2015)

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

where $\alpha + \beta = 1$

Theoretical interpretation
Deep Taylor Decomposition
(Montavon et al., 2017)

Layer-wise Relevance Propagation



Layer-wise relevance conservation

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

Equivalence

LRP-$\alpha_1\beta_0$

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

Layer-wise Relevance Propagation
(Bach'15)

=

$\mathbf{DTD{-}z}^+$

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

Deep Taylor Decomposition
(Montavon'17, arXiv in 2015)

Marginal Winning Probability

$$P(a_i) = \sum_{a_j \in \mathcal{P}_i} P(a_i|a_j)P(a_j) \qquad P(a_i|a_j) = \begin{cases} Z_j \widehat{a}_i w_{ij} & \text{if } w_{ij} \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

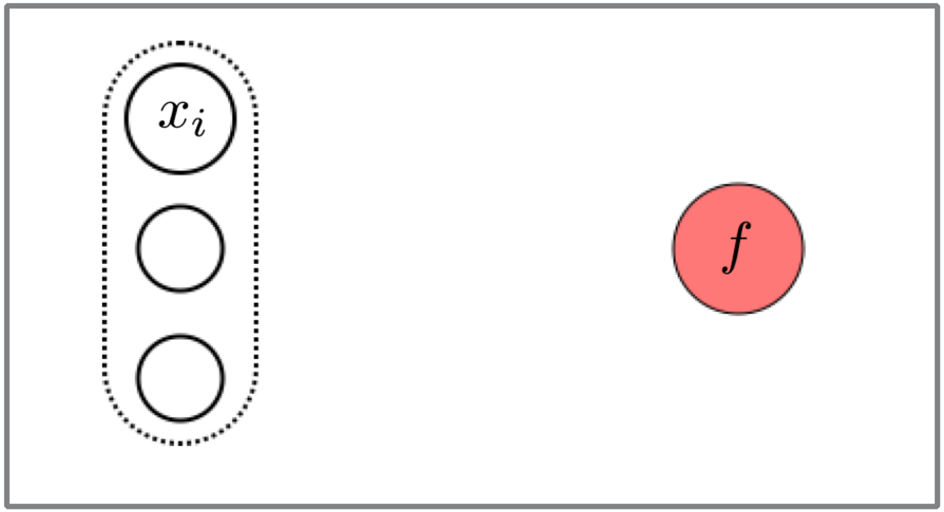
A1 activations non-negative

$$Z_j = 1/\sum_{i:w_{ij}\geq 0} \widehat{a}_i w_{ij}$$

Excitation Backprop
(Zhang'16)

Simple Taylor Decomposition

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



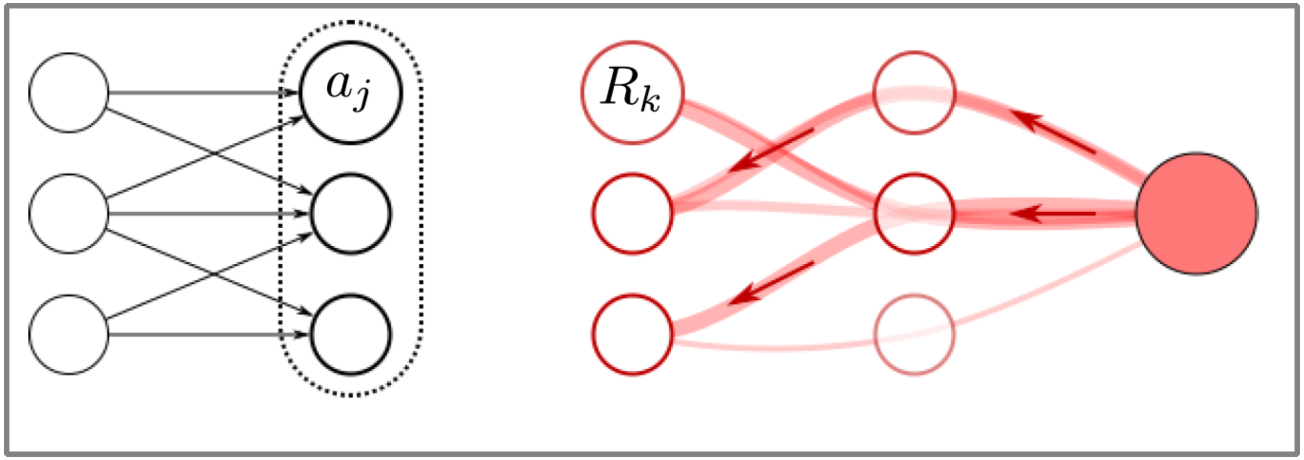
$$f(\boldsymbol{x}) = f(\widetilde{\boldsymbol{x}}) + \sum_{i=1}^{d} [\nabla f(\widetilde{\boldsymbol{x}})]_i \cdot (x_i - \widetilde{x}_i) + \mathcal{O}(\boldsymbol{x}\boldsymbol{x}^\top)$$

Idea: Use Taylor expansion to redistributed relevance <u>from output to input</u>

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

Deep Taylor Decomposition

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

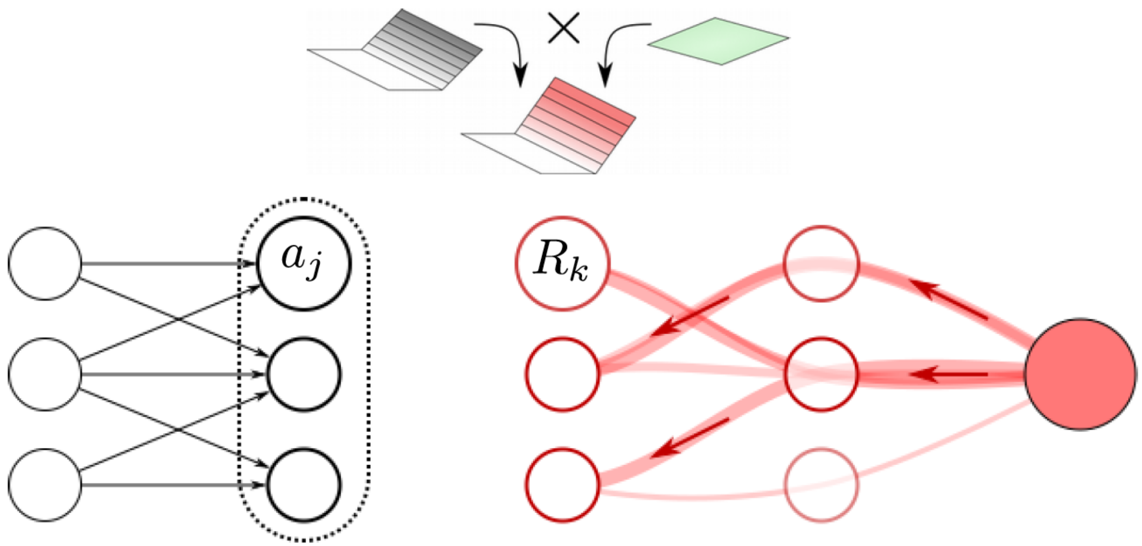


$$R_k(\boldsymbol{a}) = R_k(\widetilde{\boldsymbol{a}}) + \textstyle\sum_j [\nabla R_k(\widetilde{\boldsymbol{a}})]_j \cdot (a_j - \widetilde{a}_j) + \mathcal{O}(\boldsymbol{a}\boldsymbol{a}^\top)$$

Idea: Use Taylor expansion to redistributed relevance <u>from one layer to another</u>

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

Deep Taylor Decomposition



Key Idea: Use a “relevance model” that is easy to analyze

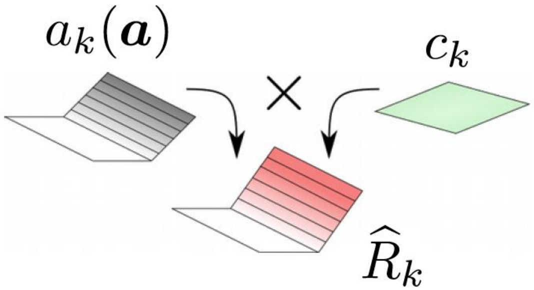
$$\widehat{R}_k(\boldsymbol{a}) = \max(0, \textstyle\sum_j a_j\, w_{jk})\, c_k$$

(Montavon et al., 2017)

Deep Taylor Decomposition

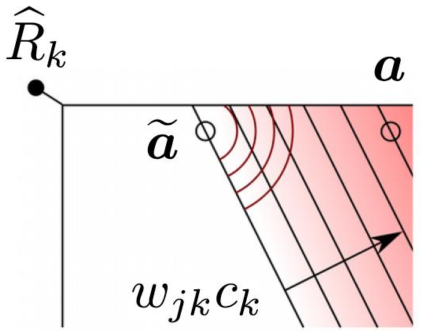
1. Relevance model

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



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

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

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

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

(Montavon et al., 2017)

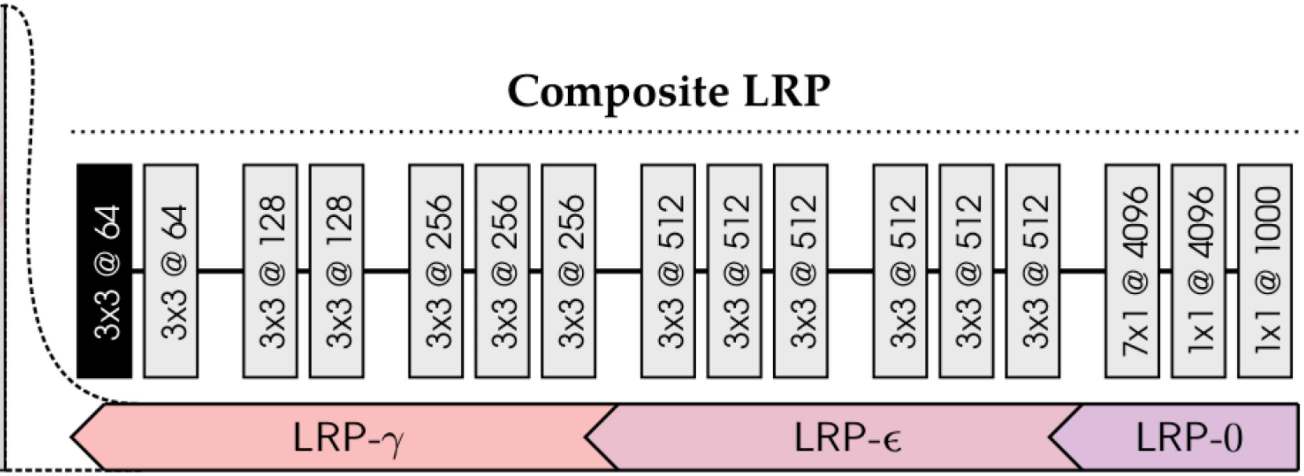
Various LRP Rules

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)	✓

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

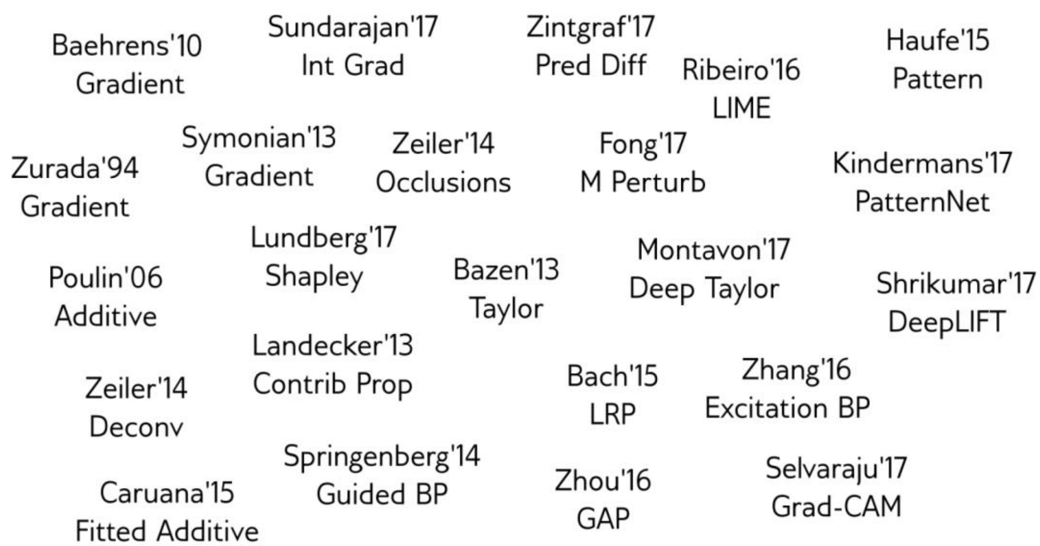
Best Practice for LRP

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)

Which one to choose ?



Evaluating Explanations

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

Pointing Game
[Zhang’16]

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

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

Task Specific Evaluation
[Poerner’18]

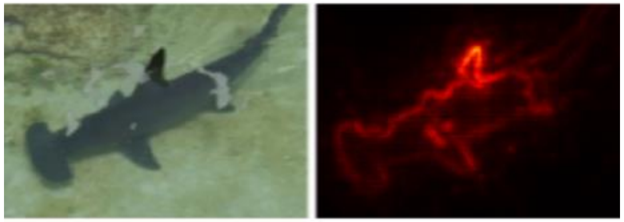
Using Ground Truth
[Arras’19]

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

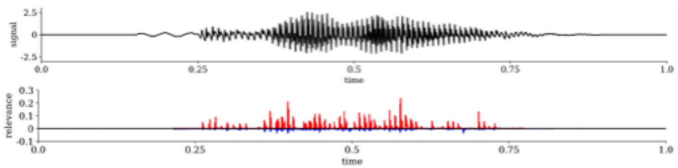
Applications of XAI

LRP Applied to Different Problems

General Images (Bach' 15, Lapuschkin'16)



Speech (Becker'18)



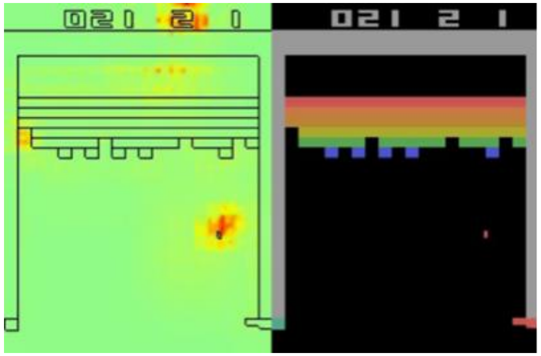
Text Analysis (Arras'16 &17)

do n't waste your money

neither funny nor susper

Morphing Attacks (Seibold'18)

Games (Lapuschkin'19)

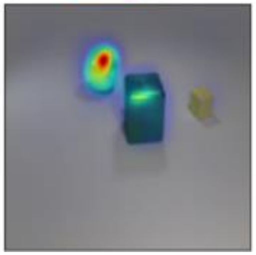


VQA (Samek'19)

there is a metallic cube ; are

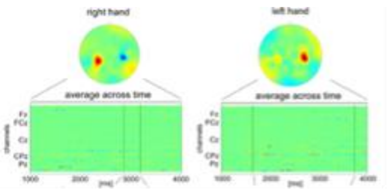
there any large cyan metallic

objects behind it ?

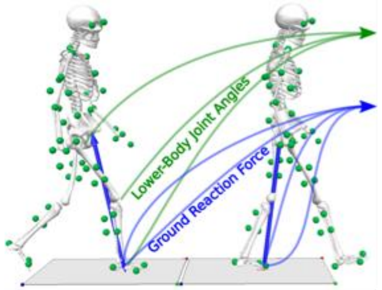


Video (Anders'19)

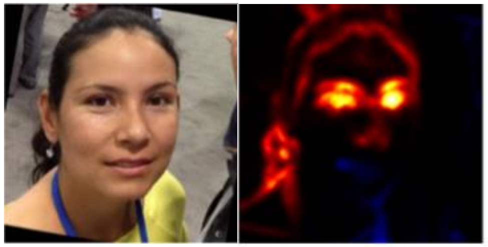
EEG (Sturm'16)



Gait Patterns (Horst'19)

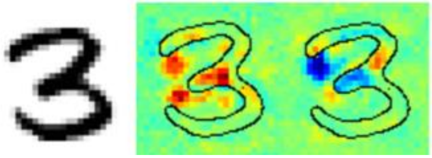


Faces (Lapuschkin'17)

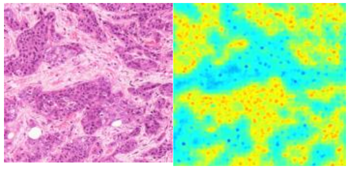


Digits (Bach' 15)

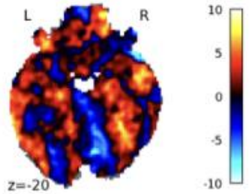
Image Class '3' Class '9'



Histopathology (Hägele'19)

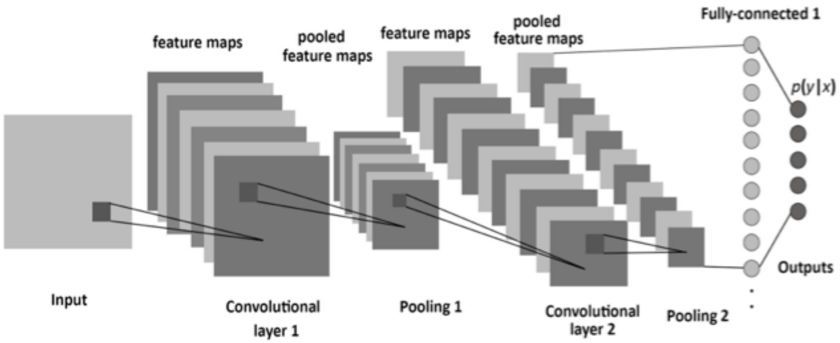


fMRI (Thomas'18)

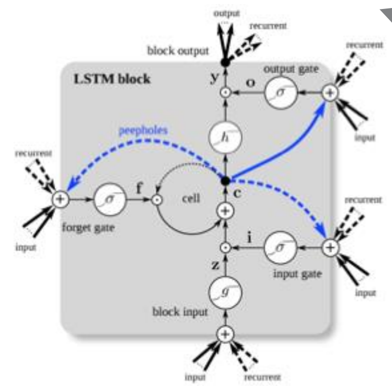


LRP Applied to Different Models

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

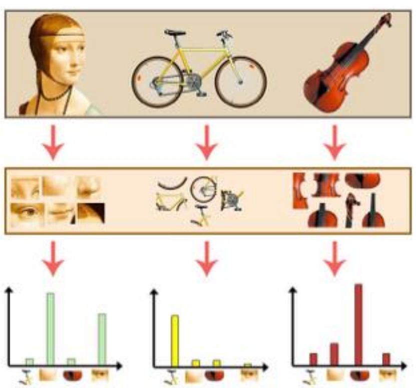


LSTM (Arras'17, Arras'19)

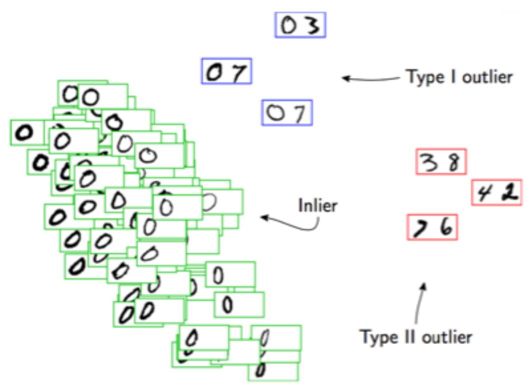


"Explaining and Interpreting LSTMs" (with S. Hochreiter)

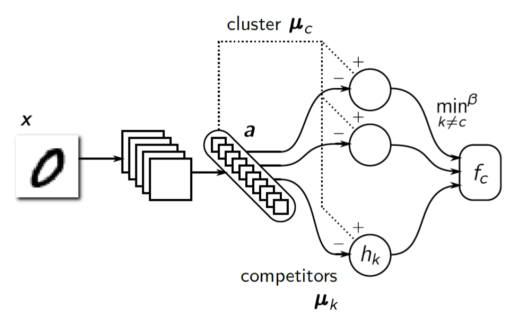
BoW / Fisher Vector models (Bach'15, Arras'16, Lapuschkin'16 …)



One-class SVM (Kauffmann'18)



Clustering (Kauffmann'19)

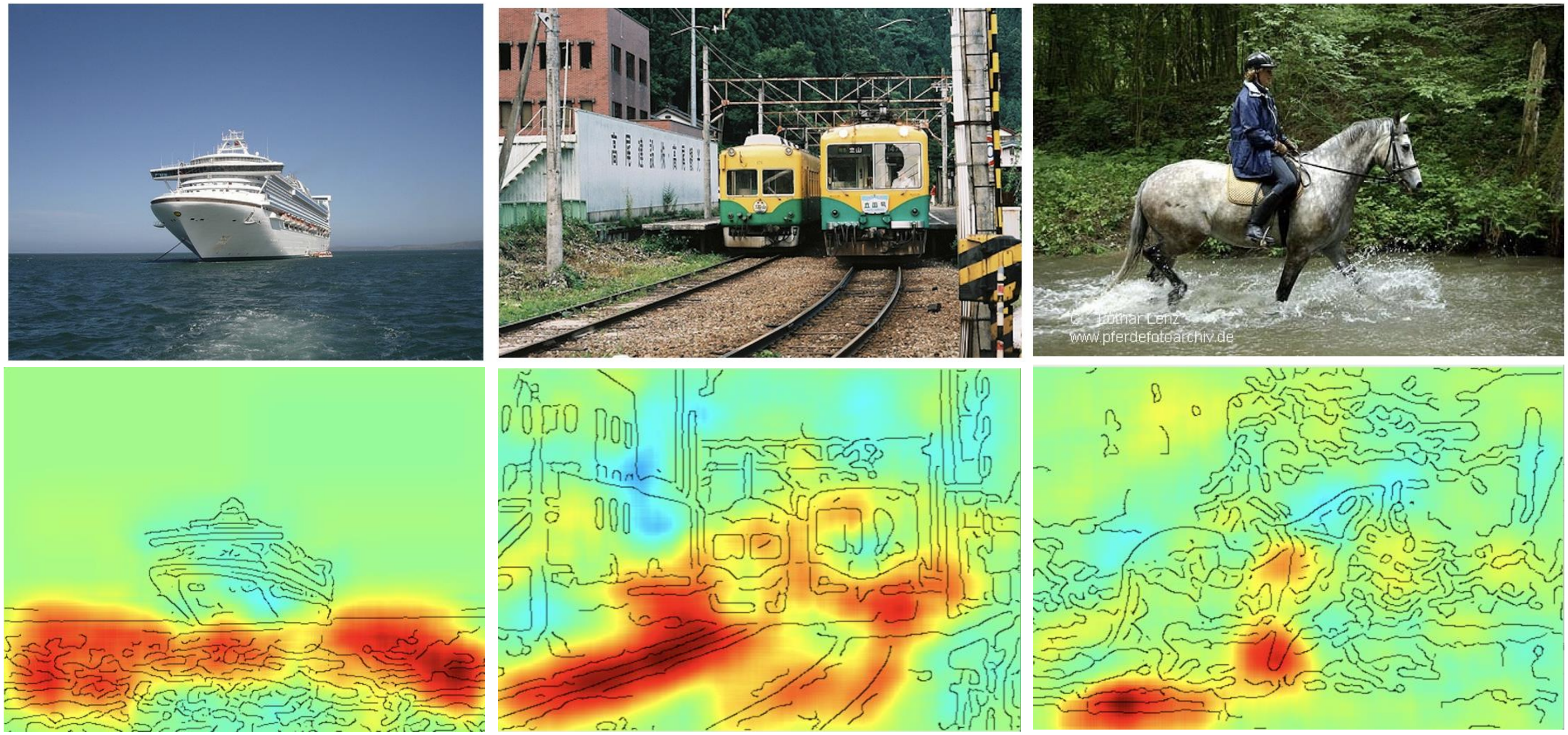


Unmasking Clever Hans Predictors

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

Unmasking Clever Hans Predictors

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



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

Unmasking Clever Hans Predictors

‘horse’ images in PASCAL VOC 2007

Identifying Biases

Smiling as a contradictor of age

Predictions

25-32 years old

<u>Strategy to solve the problem</u>:
Focus on the laughing …

60+ years old

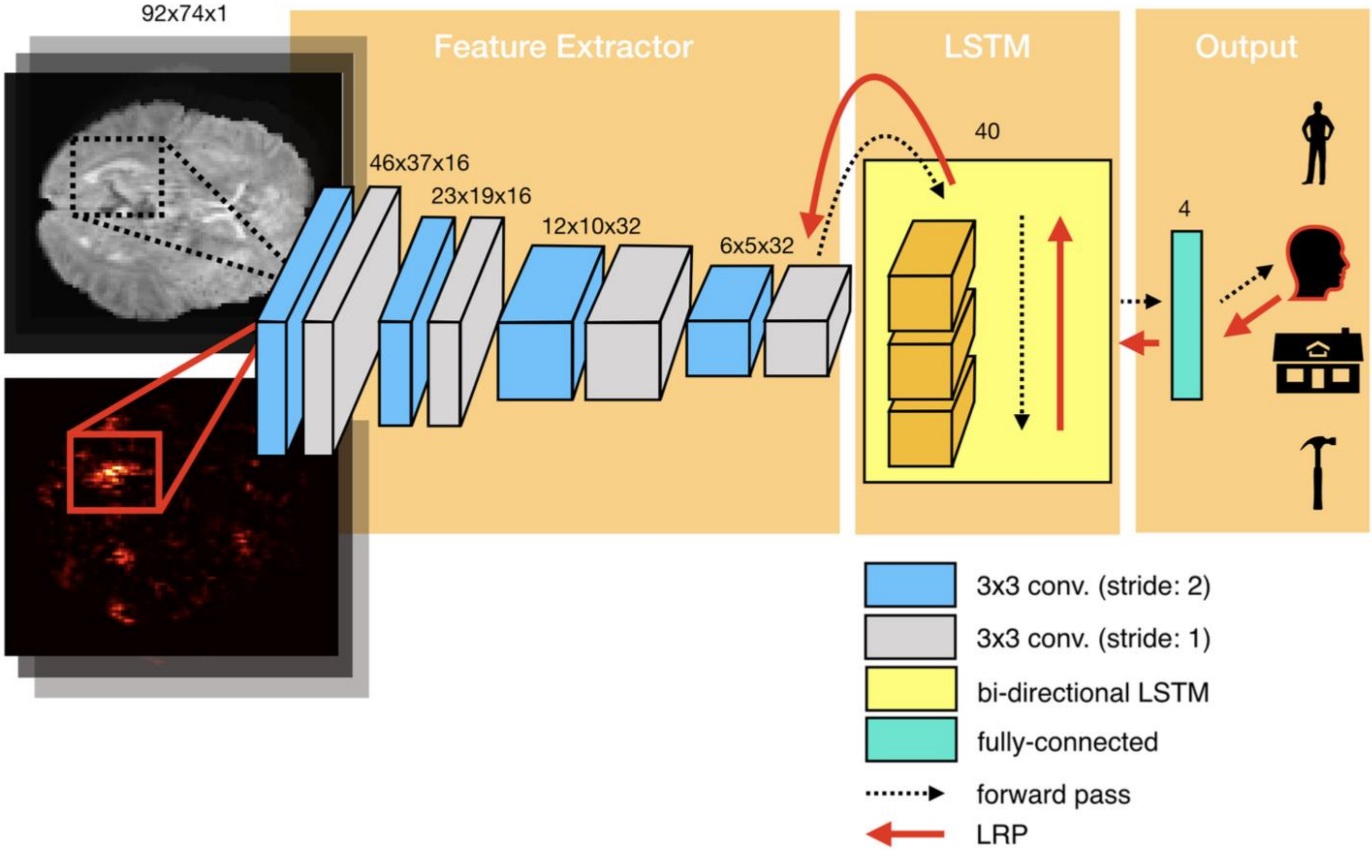
pretraining on ImageNet

laughing speaks against 60+ (i.e., model learned that old people do not laugh)

State-of-the-art DNN model, Adience Dataset (26k faces)

(Lapuschkin et al. 2017)

Scientific Insights

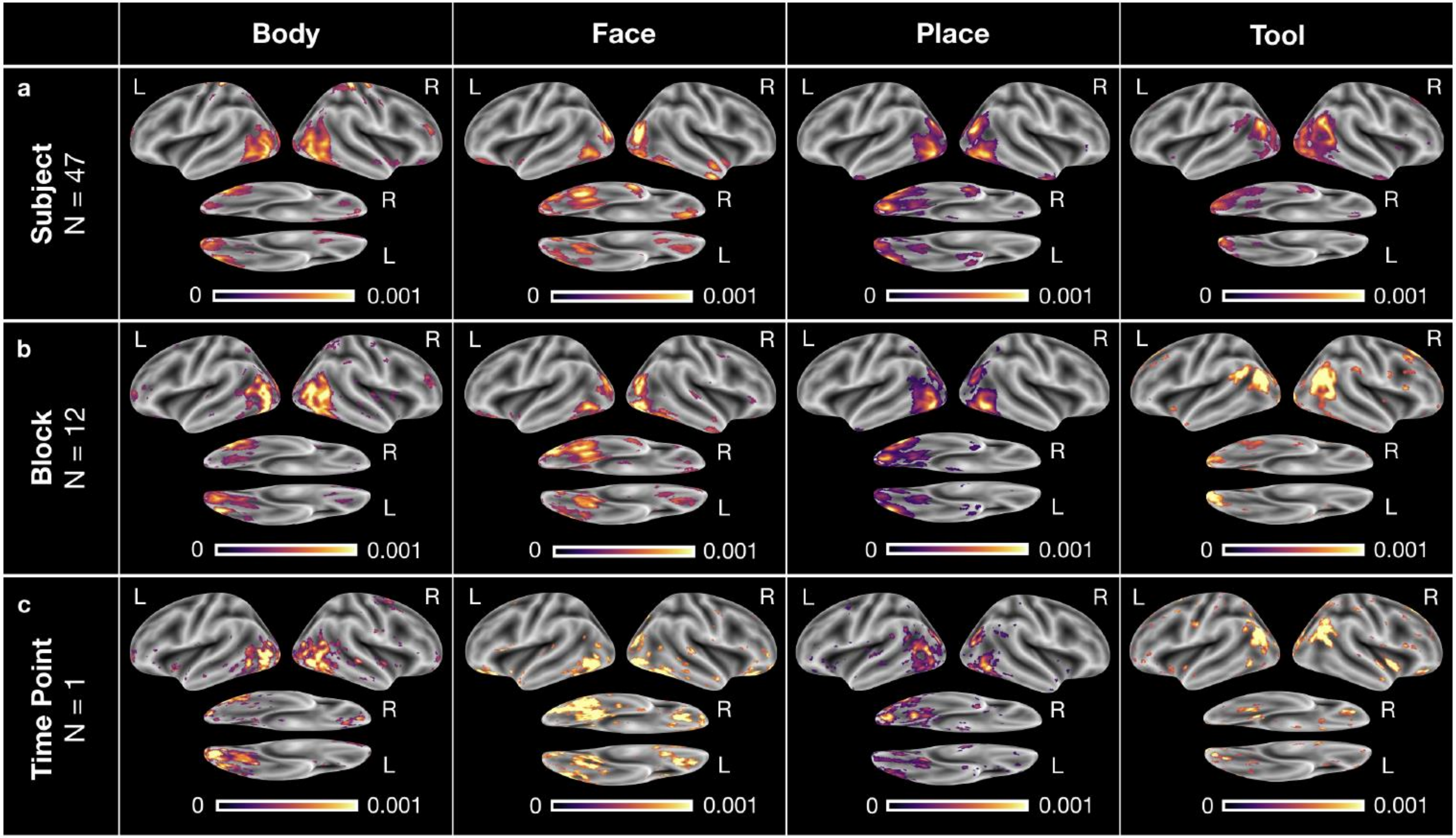


Our approach:

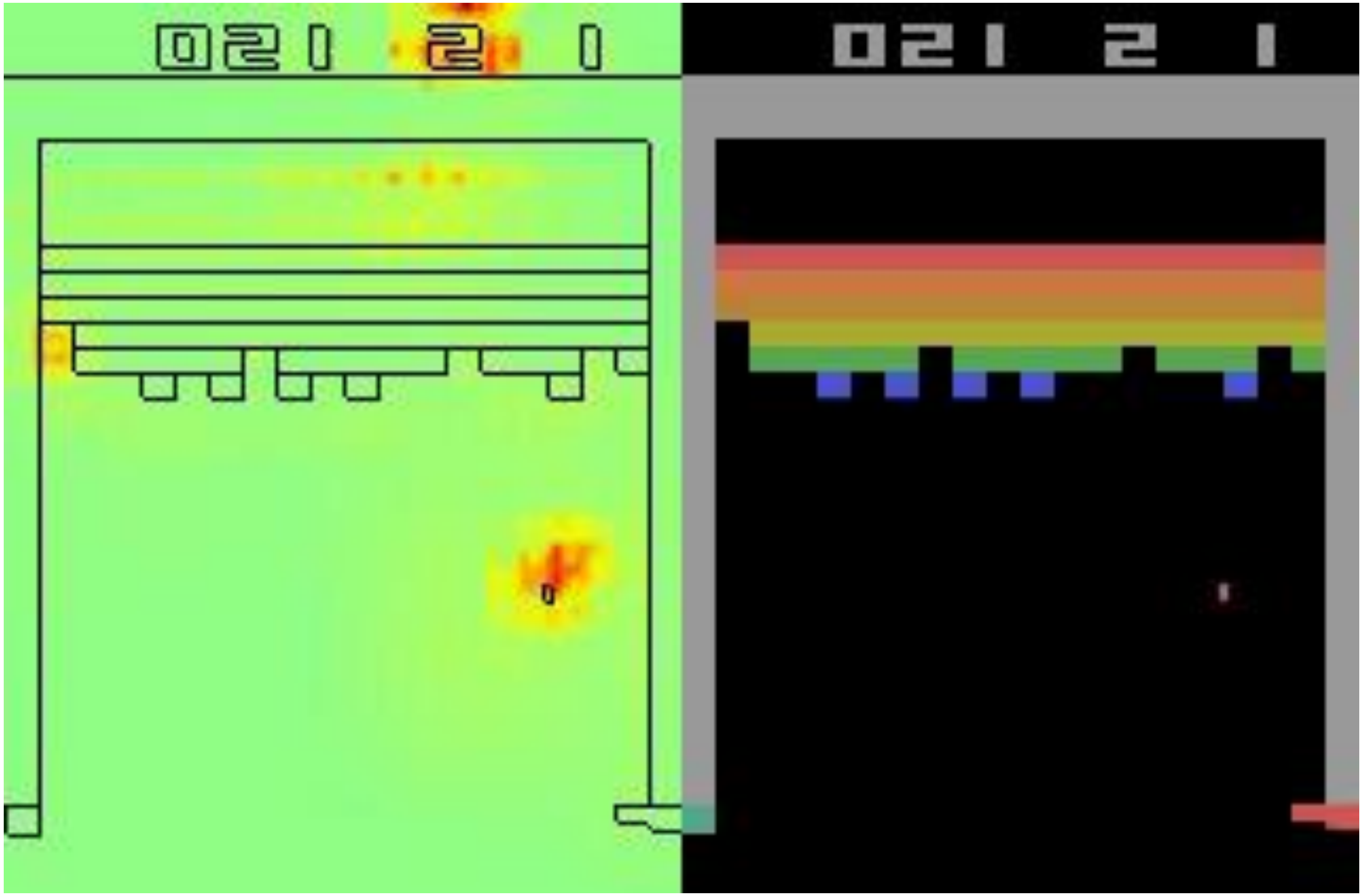
- Recurrent neural networks (CNN + LSTM) for whole-brain analysis
- LRP allows to interpret the results

(Thomas et al. 2018)

Scientific Insights

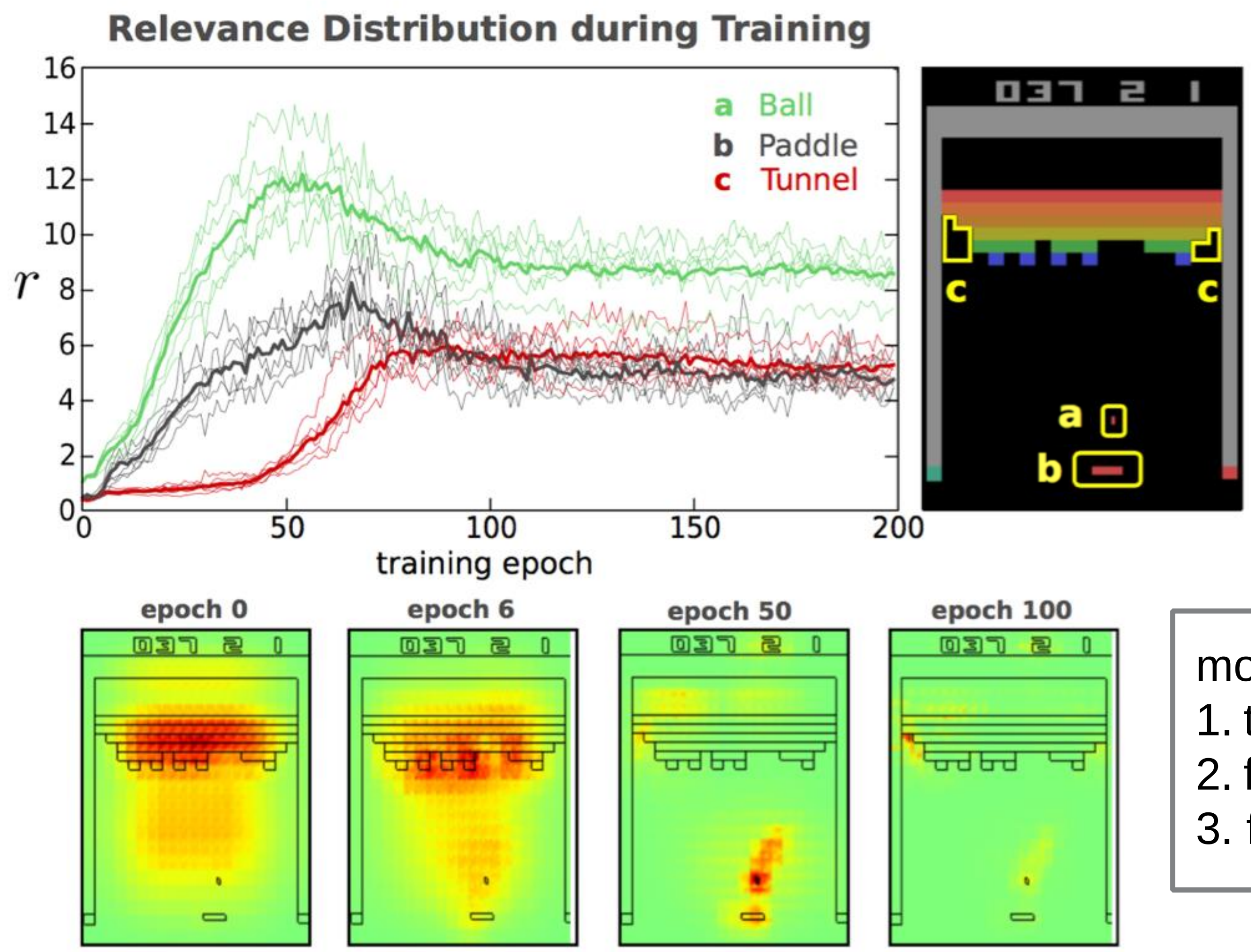


Understanding Learning Behaviour



(Lapuschkin et al., 2019)

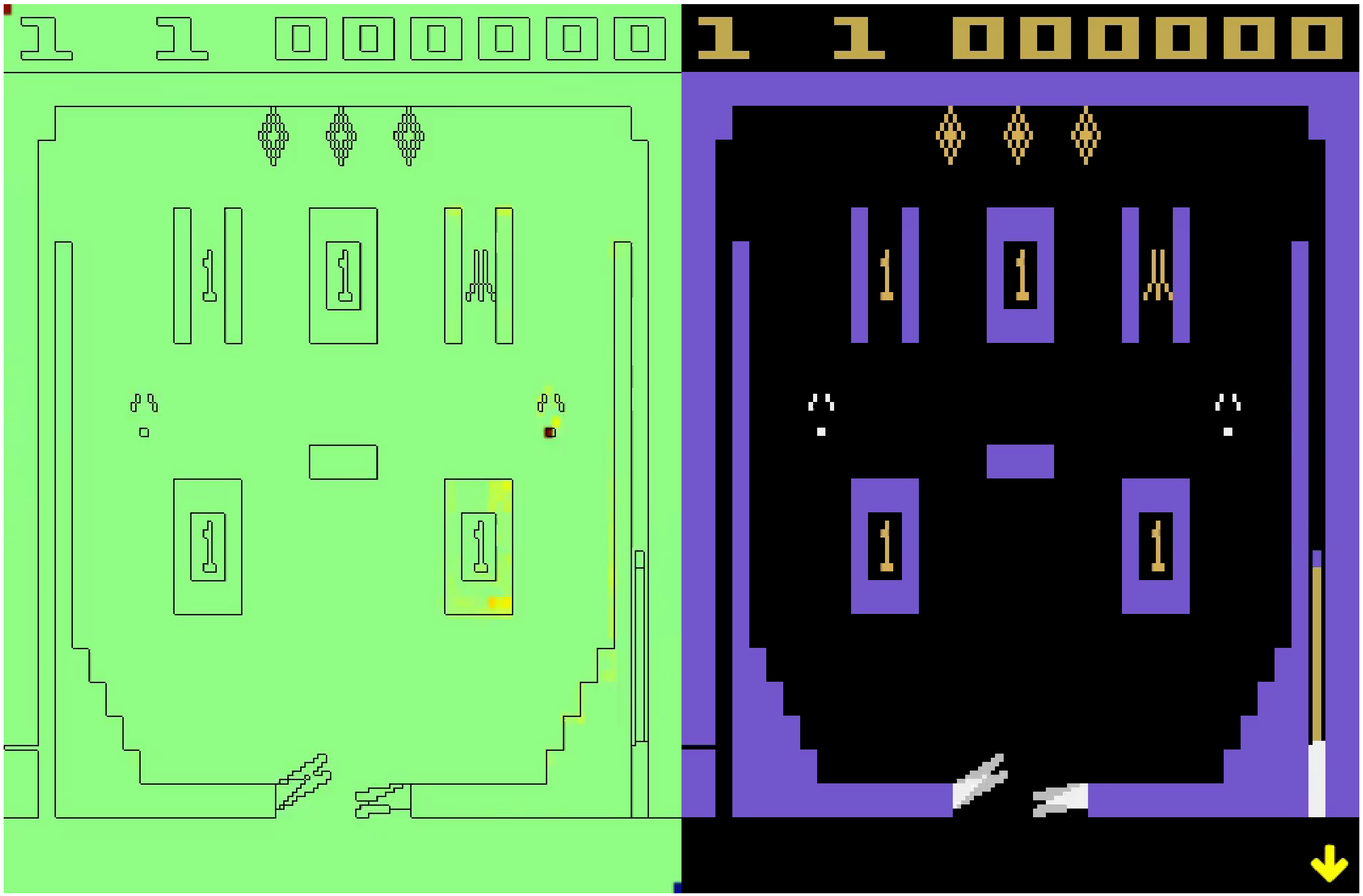
Understanding Learning Behaviour



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

(Lapuschkin et al., 2019)

Understanding Learning Behaviour

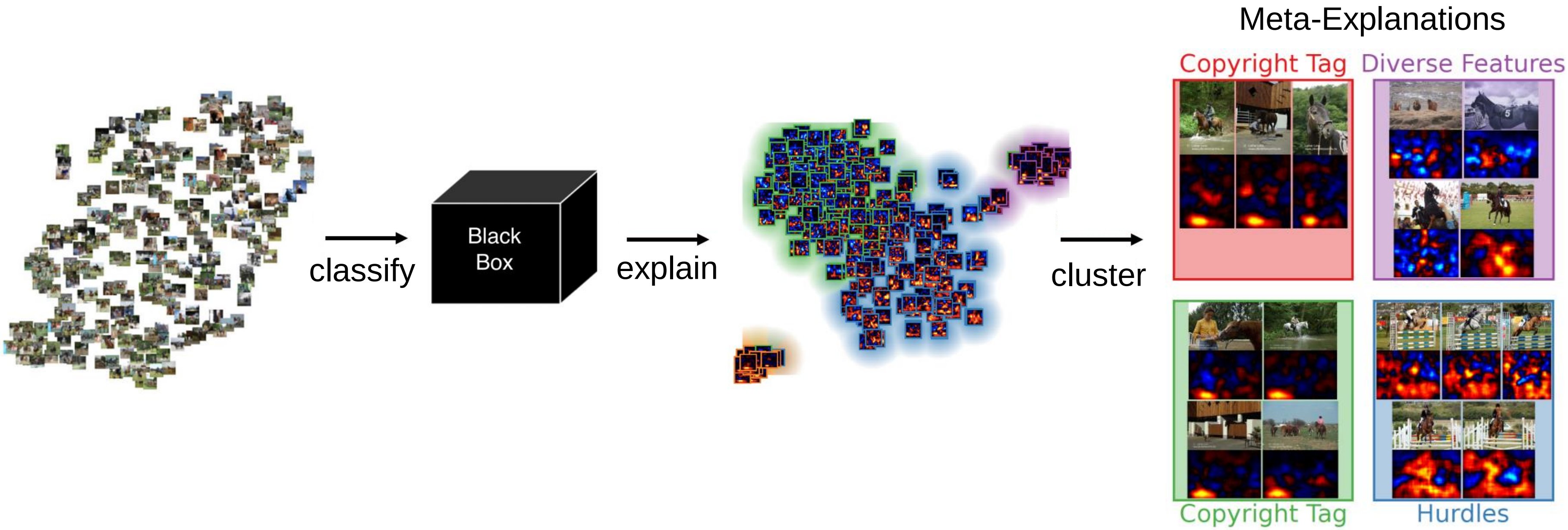


(Lapuschkin et al., 2019)

Meta-Explanations

Meta-Explanations

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

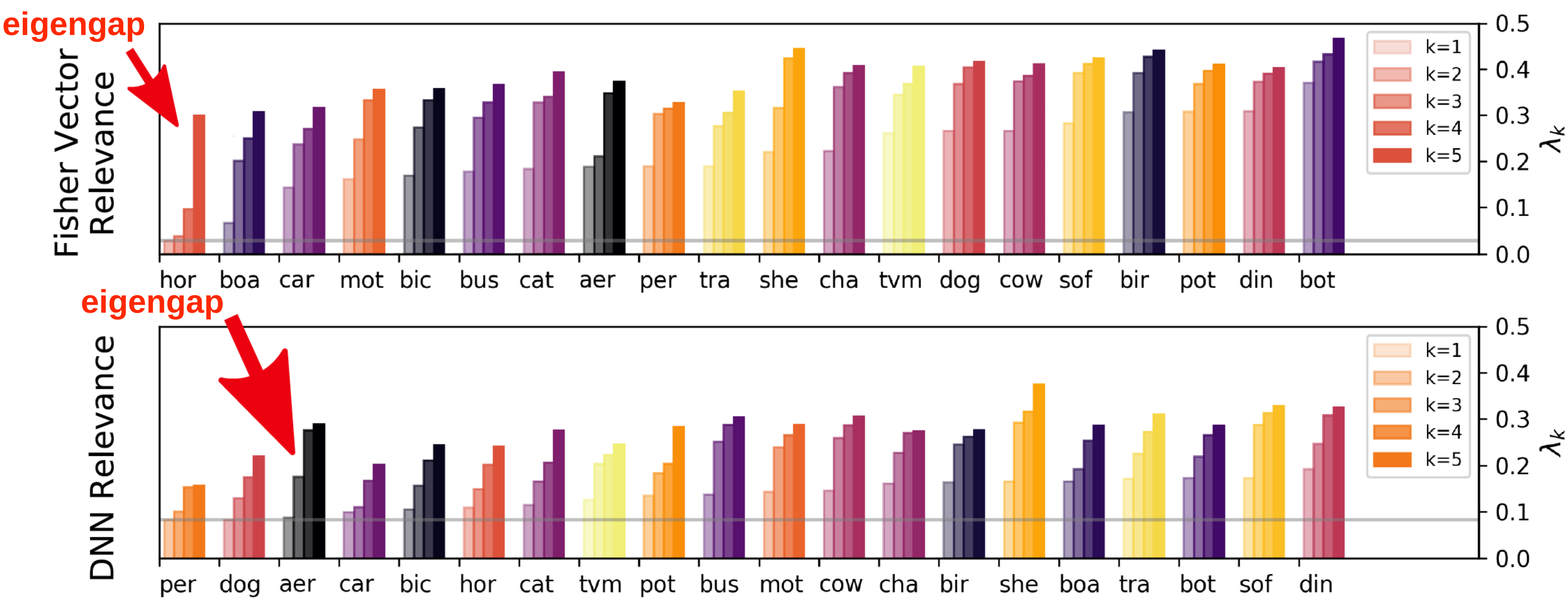


nature COMMUNICATIONS Unmasking Clever Hans predictors and assessing what machines really learn

(Lapuschkin et al., 2019)

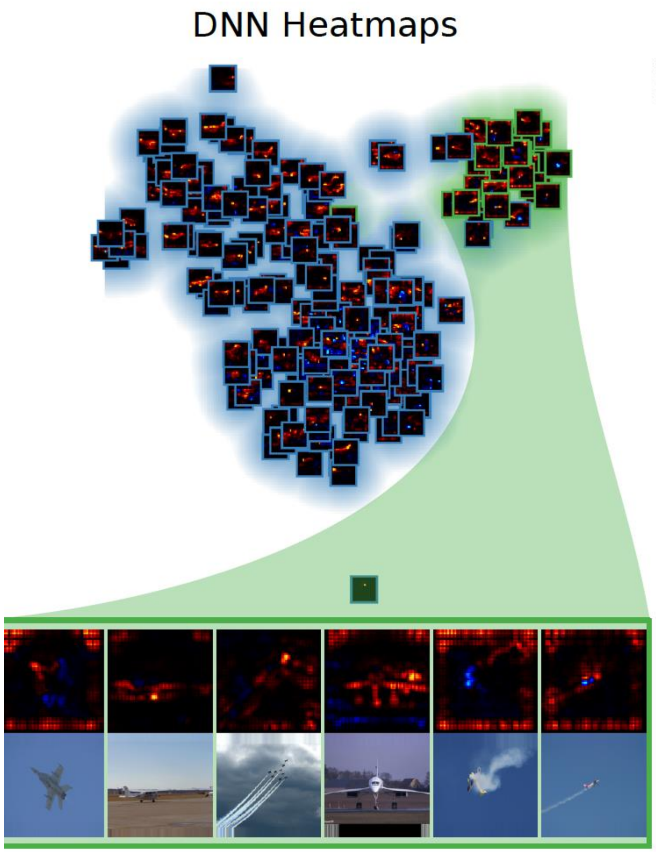
Spectral Relevance Analysis (SpRAy)

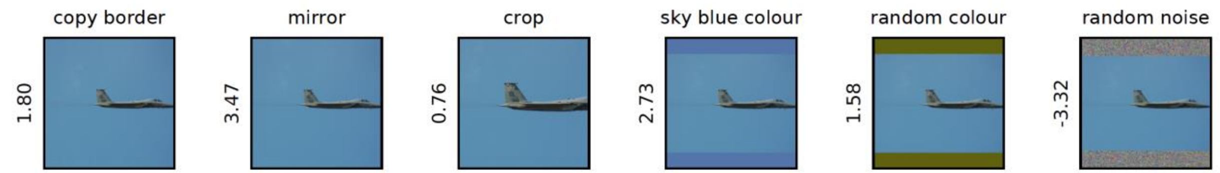
SpRAy for Fisher Vector and DNN classifiers on PASCAL VOC 2017.

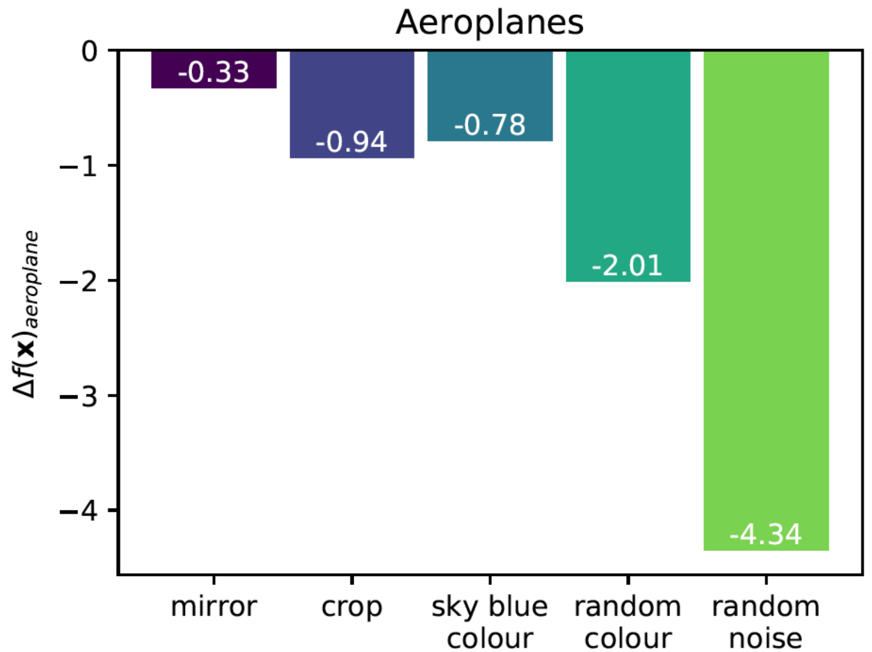


(Lapuschkin et al., 2019)

Spectral Relevance Analysis (SpRAy)







border of the image seems important

Beyond Explaining Classifiers

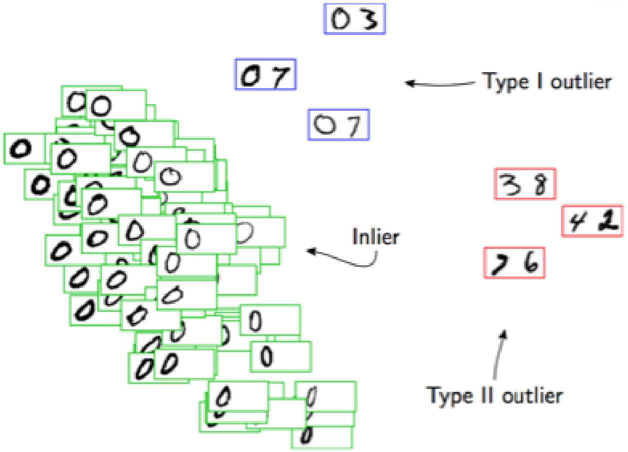
The “Neuralization” Trick

NEON (Neuralization-Propagation)

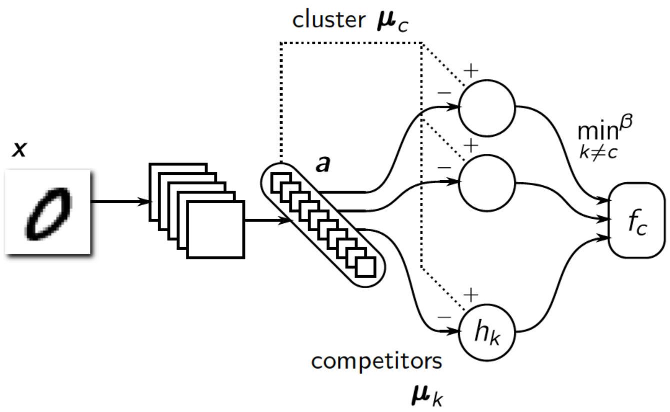
Explain ML algorithm (e.g., SVM, k-Means) in two steps:

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

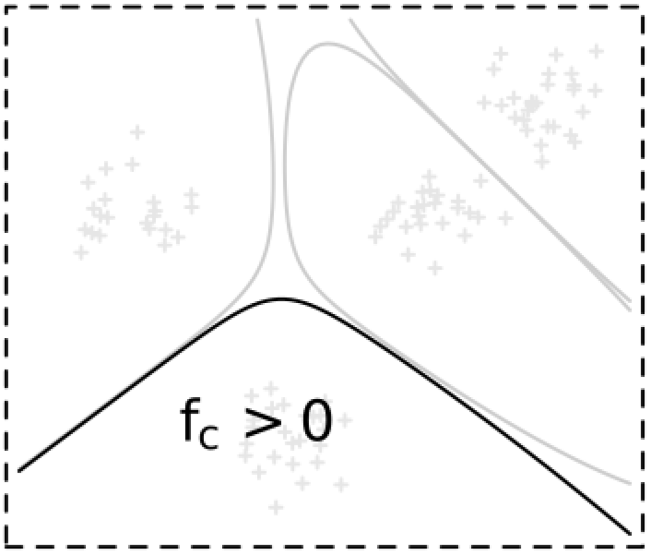
One-class SVM (Kauffmann’18)



Clustering (Kauffmann’19)



Neuralizing K-means



Represent evidence for cluster membership using logit

$$f_c(\boldsymbol{x}) = \log\Big(\frac{P(\omega_c\,|\,\boldsymbol{x})}{1 - P(\omega_c\,|\,\boldsymbol{x})}\Big)$$

with

$$P(\omega_c\,|\,\boldsymbol{x}) = \frac{\exp(-\beta \cdot o_c(\boldsymbol{x}))}{\sum_k \exp(-\beta \cdot o_k(\boldsymbol{x}))}$$

$$o_k(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{\mu}_k\|^2$$

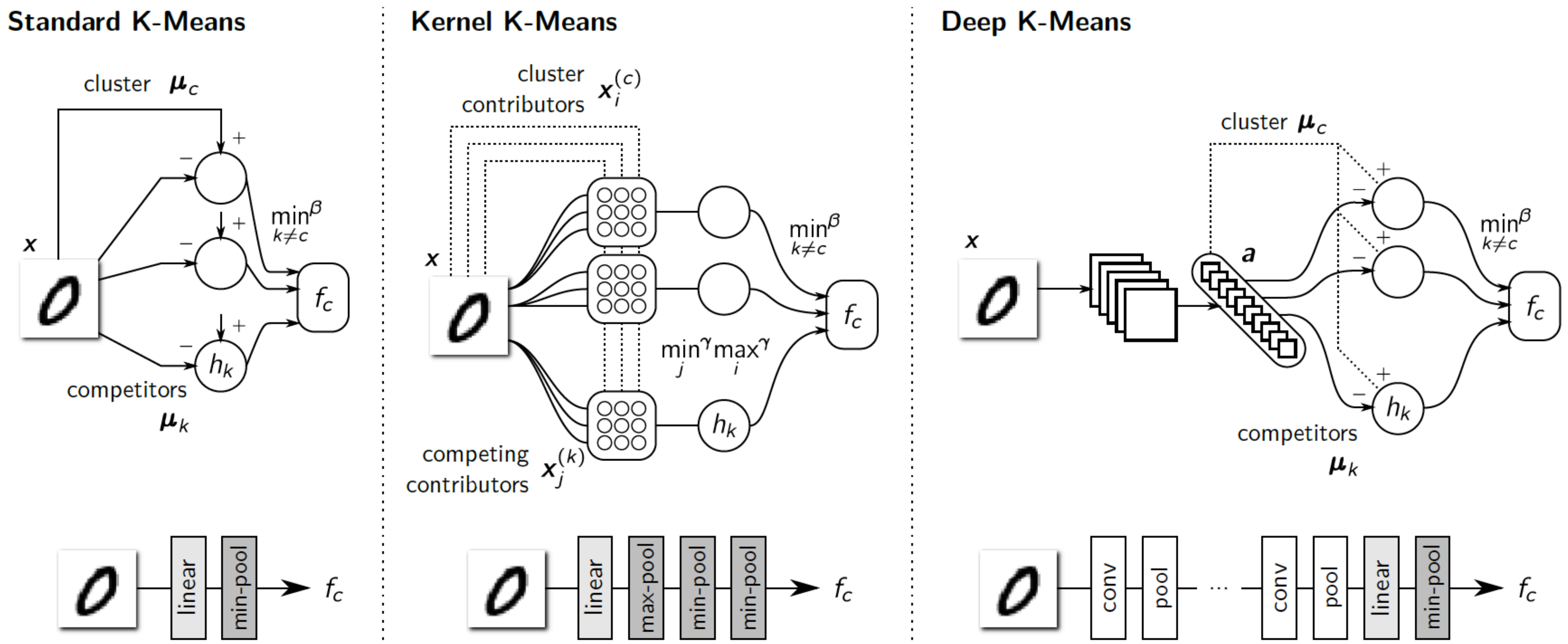
Proposition 1. *The logit that quantifies cluster membership can be written as a soft min-pooling layer*

$$f_c(\boldsymbol{x}) = \beta \cdot \min_{k \neq c}{}^{\beta}\{o_k(\boldsymbol{x}) - o_c(\boldsymbol{x})\},$$

where we define $\min^{\beta}\{\cdot\} = -\beta^{-1} \log \sum \exp(-\beta(\cdot))$.

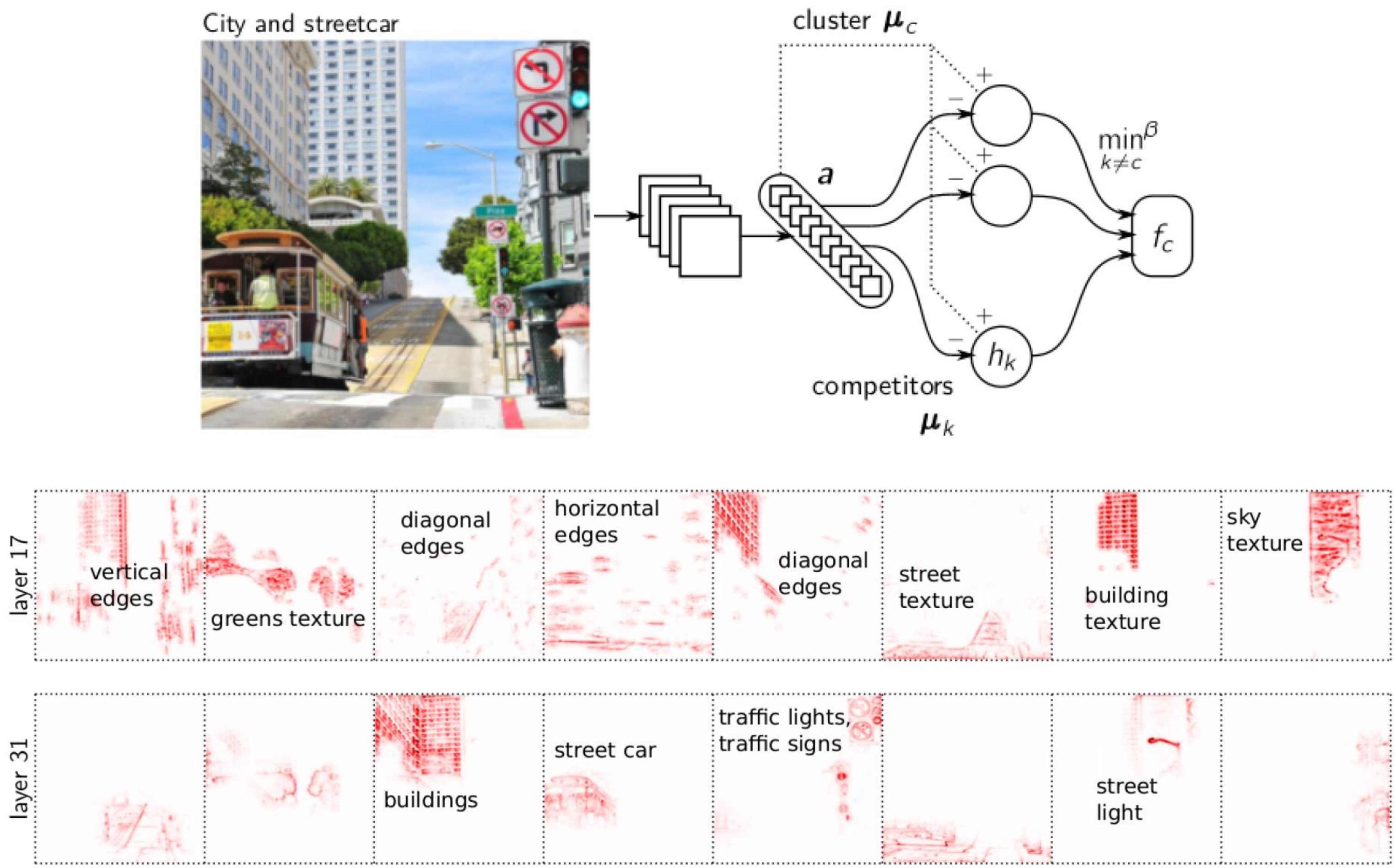
(Kauffmann et al. 2019)

Neuralizing K-means



(Kauffmann et al. 2019)

K-Means on VGG-16 Features



(Kauffmann et al. 2019)

Summary

Decisions functions of ML models are often complex, and analyzing them directly can be difficult.

Levering the model's structure largely simplifies the explanation problem.

Layer type dependent redistribution rules exist and should be used.

Explanations and Meta-Explanations can be used for various purposes.

Common ML models (e.g. OC-SVM, k-means) can often be decomposed as a sequence of explainable layers ("neuralization").

References

Opinion Paper

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.

Tutorial & Overview Papers

G Montavon, W Samek, KR Müller. Methods for Interpreting and Understanding Deep Neural Networks. *Digital Signal Processing, 73:1-15*, 2018.

W Samek, T Wiegand, and KR Müller, Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models, *ITU Journal: ICT Discoveries - Special Issue 1 - The Impact of Artificial Intelligence (AI) on Communication Networks and Services*, 1(1):39-48, 2018.

W Samek and KR Müller, Towards Explainable Artificial Intelligence. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, LNCS, Springer*, 11700:5-22, 2019.

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.

G Montavon, S Bach, A Binder, W Samek, KR Müller. Explaining NonLinear Classification Decisions with Deep Taylor Decomposition. *Pattern Recognition*, 65:211–222, 2017

G Montavon, A Binder, S Lapuschkin, W Samek, KR Müller: Layer-Wise Relevance Propagation: An Overview. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, LNCS, Springer*, 11700:193-209, 2019.

L Arras, J Arjona-Medina, M Widrich, G Montavon, M Gillhofer, K-R Müller, S Hochreiter, W Samek, Explaining and Interpreting LSTMs. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, LNCS, Springer*, 11700:193-209, 2019.

References

Further Methods Papers

J Kauffmann, M Esders, G Montavon, W Samek, KR Müller. From Clustering to Cluster Explanations via Neural Networks. arXiv:1906.07633, 2019.

L Arras, G Montavon, K-R Müller, W Samek. Explaining Recurrent Neural Network Predictions in Sentiment Analysis. *EMNLP'17 Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA)*, 159-168, 2017.

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.

Application to Images & Faces

S Lapuschkin, A Binder, G Montavon, KR Müller, Wojciech Samek. Analyzing Classifiers: Fisher Vectors and Deep Neural Networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2912-20, 2016.

S Bach, A Binder, KR Müller, W Samek. Controlling Explanatory Heatmap Resolution and Semantics via Decomposition Depth. *IEEE International Conference on Image Processing (ICIP)*, 2271-75, 2016.

F Arbabzadeh, 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, 9796:344-54, Springer International Publishing, 2016.

S Lapuschkin, A Binder, KR Müller, W Samek. Understanding and Comparing Deep Neural Networks for Age and Gender Classification. *IIEEE International Conference on Computer Vision Workshops (ICCVW)*, 1629-38, 2017.

C Seibold, W Samek, A Hilsmann, P Eisert. Accurate and Robust Neural Networks for Security Related Applications Exampled by Face Morphing Attacks. arXiv:1806.04265, 2018.

References

Application to NLP

L Arras, F Horn, G Montavon, KR Müller, W Samek. Explaining Predictions of Non-Linear Classifiers in NLP. *Workshop on Representation Learning for NLP,* Association for Computational Linguistics, 1-7, 2016.

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.

L Arras, G Montavon, K-R Müller, W Samek. Explaining Recurrent Neural Network Predictions in Sentiment Analysis. *EMNLP'17 Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA)*, 159-168, 2017.

Application to Video

C Anders, G Montavon, W Samek, KR Müller. Understanding Patch-Based Learning of Video Data by Explaining Predictions. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, LNCS, Springer*, 11700:297-309, 2019

V Srinivasan, S Lapuschkin, C Hellge, KR Müller, W Samek. Interpretable Human Action Recognition in Compressed Domain. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1692-96, 2017.

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 the Sciences

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.

I Sturm, S Lapuschkin, W Samek, KR Müller. Interpretable Deep Neural Networks for Single-Trial EEG Classification. *Journal of Neuroscience Methods*, 274:141–145, 2016.

A Thomas, H Heekeren, KR Müller, W Samek. Interpretable LSTMs For Whole-Brain Neuroimaging Analyses. *arXiv:1810.09945*, 2018.

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. *arXiv:1908.06943*, 2019.

Evaluation Explanations

W Samek, A Binder, G Montavon, S Lapuschkin, 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.

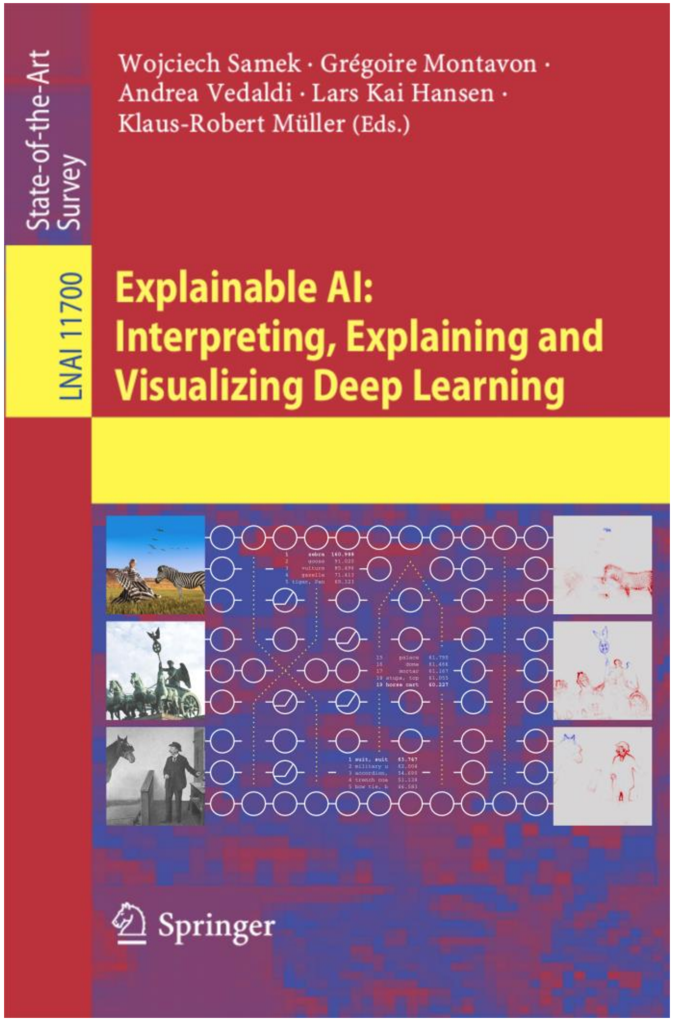
L Arras, A Osman, KR Müller, W Samek. Evaluating Recurrent Neural Network Explanations. *Proceedings of the ACL'19 Workshop on BlackboxNLP,* Association for Computational Linguistics, 113-126, 2019.

Software

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,* 20:1-8, 2019.

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.

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

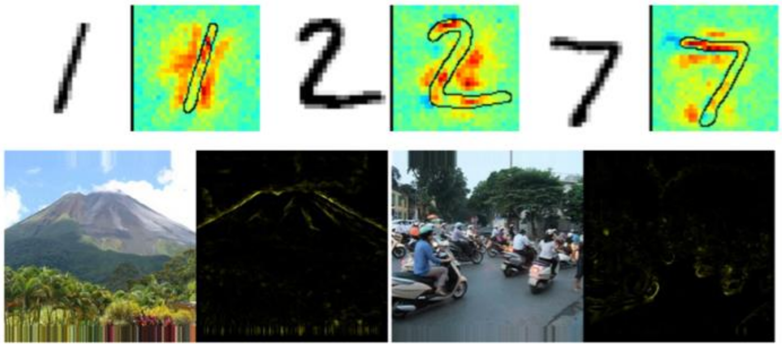
—> 22 Chapters

Thank you for your attention

Visit:

`http://www.heatmapping.org`

- Tutorials
- Software
- Online Demos



Acknowledgement
Klaus-Robert Müller (TUB)
Grégoire Montavon (TUB)
Sebastian Lapuschkin (HHI)
Leila Arras (HHI)
Alexander Binder (SUTD)
…

Tutorial Paper
Montavon et al., "Methods for interpreting and understanding deep neural networks", Digital Signal Processing, 73:1-5, 2018

Keras Explanation Toolbox
`https://github.com/albermax/innvestigate`