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

Part 1: Introduction to XAI

Wojciech Samek, Grégoire Montavon

September 18, 2020

Fraunhofer HHI
TU berlin
BIFOLD

ML Models are Black Boxes

Report No. 85-460-1

THE PERCEPTRON

A PERCEIVING AND RECOGNIZING AUTOMATON

(PROJECT PARA)

January, 1957

Prepared by: Frank Rosenblatt

Frank Rosenblatt,
Project Engineer

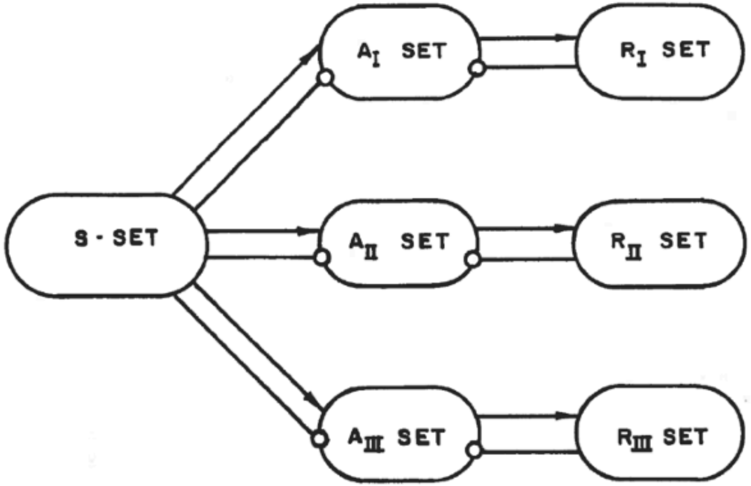


FIGURE 2

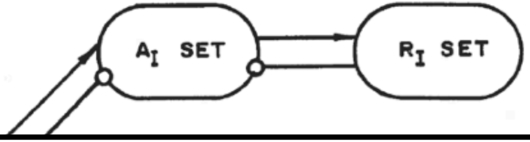
ORGANIZATION OF A PERCEPTRON WITH THREE INDEPENDENT OUTPUT-SETS

ML Models are Black Boxes

II. GENERAL DESCRIPTION OF A PHOTOPERCEPTRON

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

Frank Rosenblatt,
Project Engineer



ORGANIZATION OF A PERCEPTRON WITH THREE INDEPENDENT OUTPUT-SETS

ML Models are Black Boxes

1957

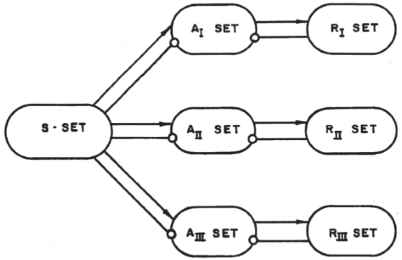
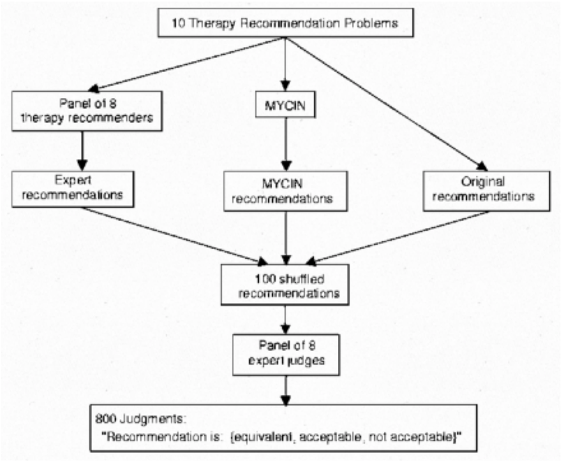


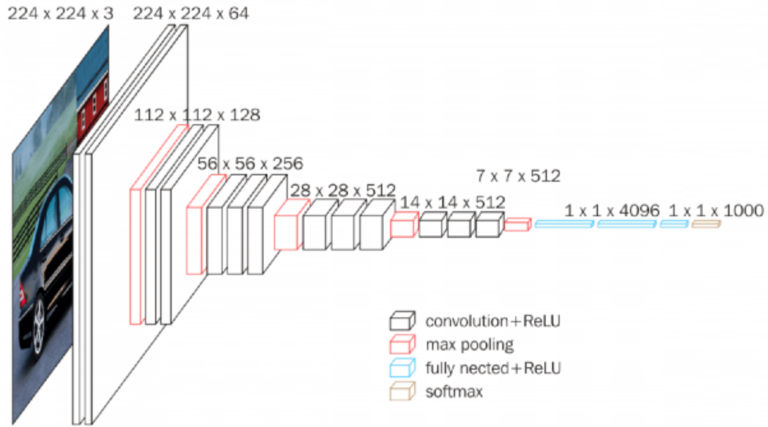
FIGURE 2

ORGANIZATION OF A PERCEPTRON WITH THREE INDEPENDENT OUTPUT-SETS

1972



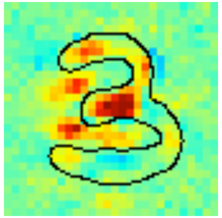
> 2012



What do we want to explain ?

prediction

“Explain why a certain pattern x has been classified in a certain way f(x).”



model

“What concept does a particular neural encode?"

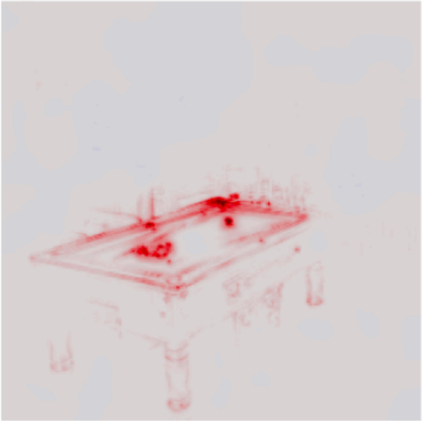
data

“Which dimensions of the data are most relevant for the task.”

Explaining Predictions

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

some pool table



why it is classified as a pool table

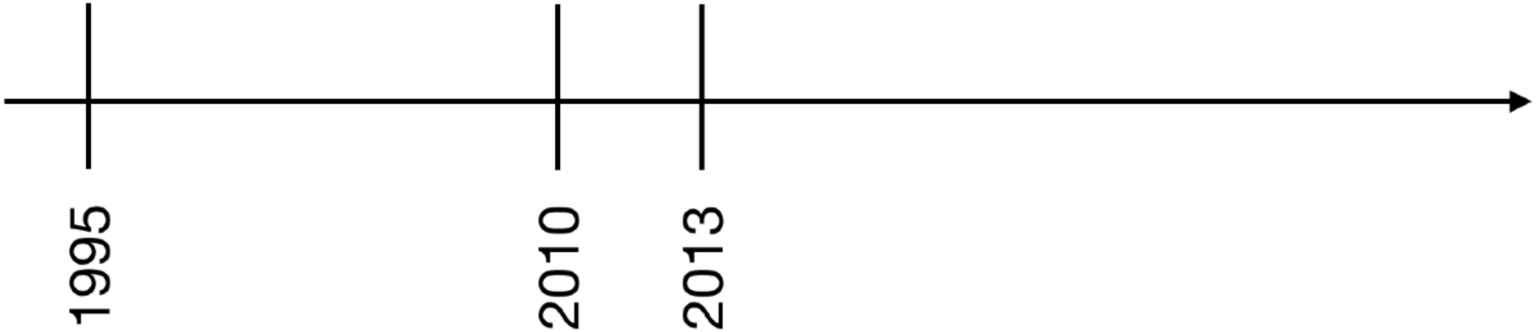
Brief History

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

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

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

sensitivity analysis



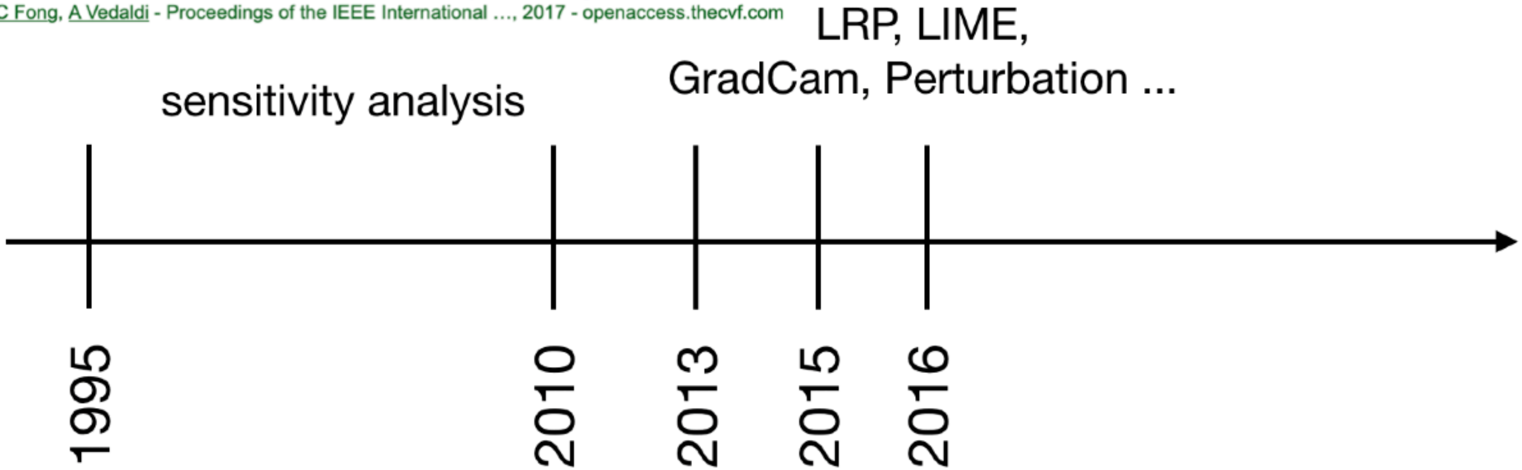
Brief History

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

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

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

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



Brief History

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

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

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

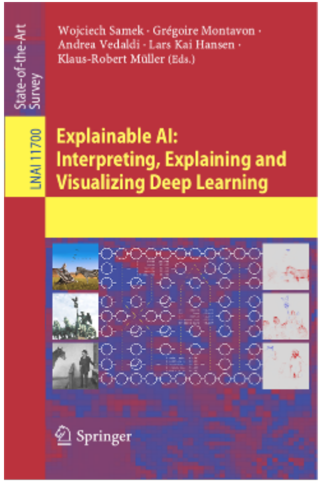
XAI for LSTMs

Theoretical frameworks for XAI

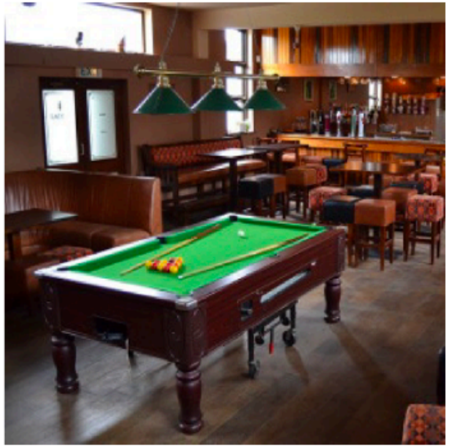
LRP, LIME, GradCam, Perturbation …

sensitivity analysis

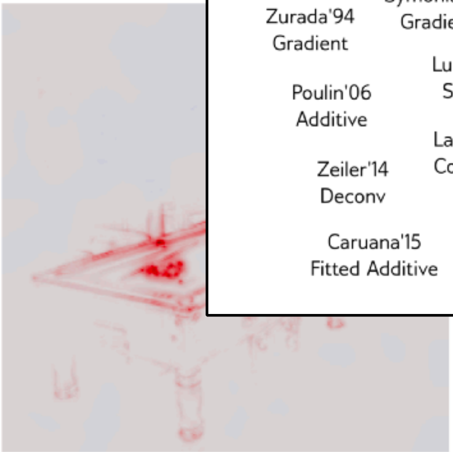
1995 2010 2013 2015 2016 2017 2018 2019 …



Explaining Predictions



some pool table



why it is classified as a pool table

Which XAI method 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

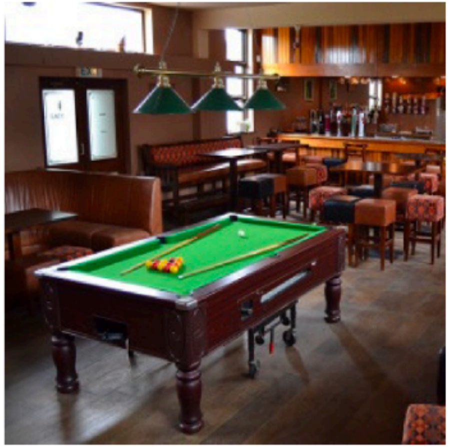
Selvaraju'17 Grad-CAM

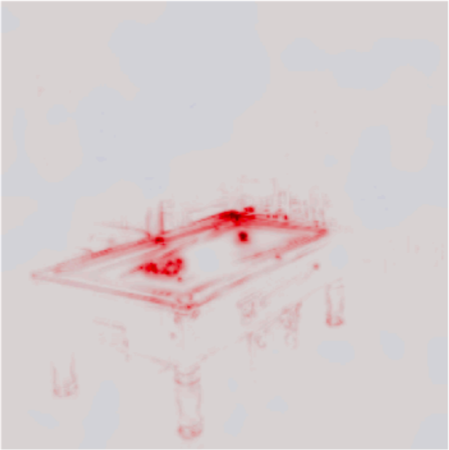
Explaining Predictions

What can we do with it ?

Explaining Predictions

Explaining more than classifiers





Tutorial

Part 1
- Why to explain ?
- Types of XAI methods
- What is a "good" explanation
- Example

Part 2
- Formalizing explanations
- Overview of XAI techniques

Part 3
- Empirical comparison
- Implementation
- Theoretical embedding
- Extensions

Part 4
- Aggregating explanations
- Applications
- Future topics

Why to explain?

“Superhuman” AI Systems

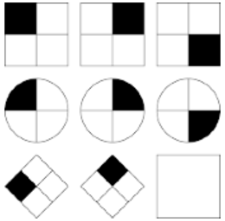
Game GO

Texas Hold’em Poker

Image classification

Traffic sign recognition

IQ Test



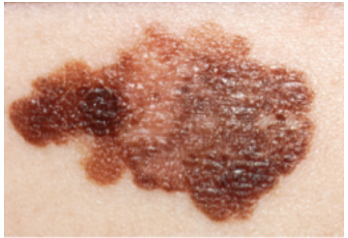
Computer games

Jeopardy

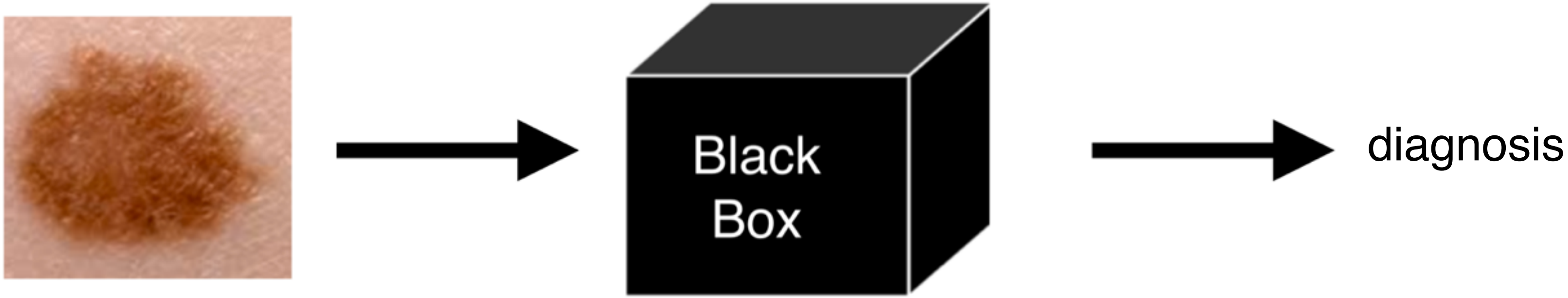
Drone control

Lung cancer detection

Skin cancer detection



Can we trust these black boxes ?



Is minimizing the error a guarantee for the model to work well in practice?

Can we trust these black boxes ?

We need interpretability in order to:

trust & verification

legal aspects

learn from the system

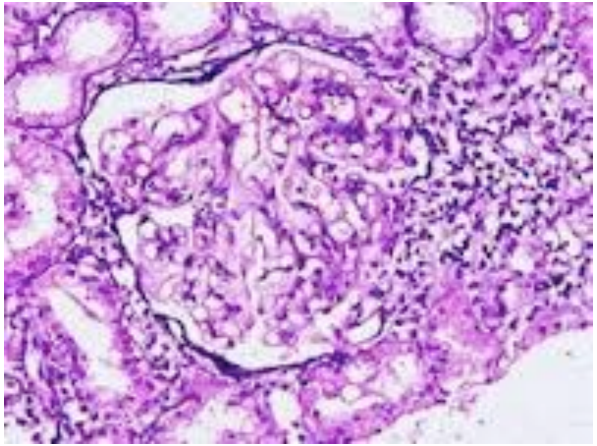
improve system

Can we trust these black boxes ?

We need interpretab

trust & verification

“AI medical diagnosis system misclassifies patient’s disease …”



I

as

Can we trust these black boxes ?

We need interpretability in orde

trust & verification

legal aspects

"right to explanation"

avoid discrimination

Retain human decision in order to assign responsibility.

Can we trust these black boxes ?

"It's not a human move. I've never seen a human play this move." (Fan Hui)

r to:

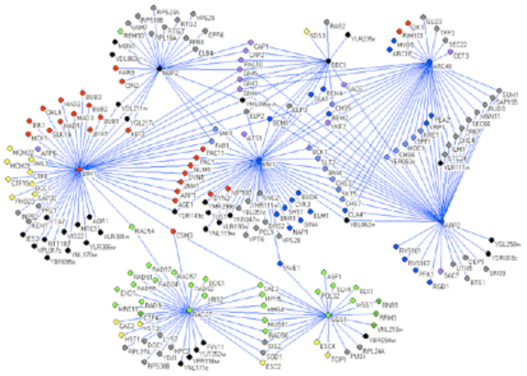
improve system

learn from the system

Can we trust these black boxes ?

r to:

Learn about the physical / biological / chemical mechanisms. (e.g. find genes linked to cancer)

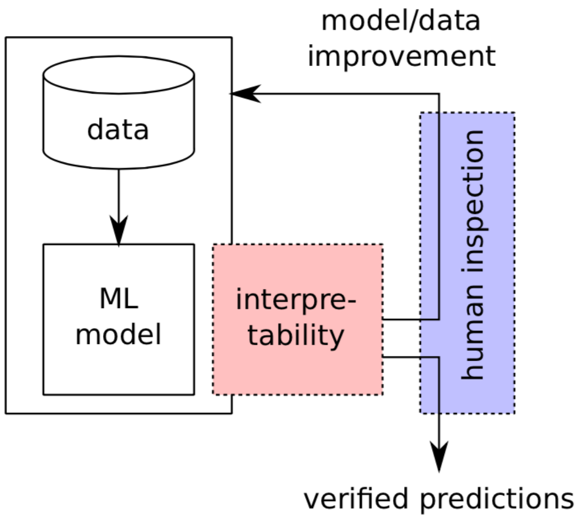


improve system

learn from the system

Can we trust these black boxes ?

W



Generalization error + human experience

):

improve system

learn from the system

Why to explain?

Interpretability as a gateway between ML and society

- Make complex models acceptable for certain applications.
- Retain human decision in order to assign responsibility.
- “Right to explanation”

Why to explain?

Interpretability as powerful engineering tool

- Optimize models / architectures
- Detect flaws / biases in the data
- Gain new insights about the problem
- Make sure that ML models behave "correctly"

Types of XAI 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)

…

Explanation Methods

Perturbation-Based

1. perturb
2. measure reaction
3. identify relevant information

Function-Based

1. treat the NN as function
2. compute simple quantities on it
3. construct explanation

Surrogate- / Sampling-Based

1. locally approximate prediction using simple function
2. explain simple function

Structure-Based

1. if the decision is too complex to explain, break the function into subfunctions.
2. explain each subfunction
3. meaningfully aggregate

Explanation Methods

Perturbation-Based

1. perturb
2. measure reaction
3. identify relevant information

Function-Based

1. treat the NN as function
2. compute simple quantities on it
3. construct explanation

Surrogate- / Sampling-Based

1. locally approximate prediction using simple function
2. explain simple function

Structure-Based

1. if the decision is too complex to explain, break the function into subfunctions.
2. explain each subfunction
3. meaningfully aggregate

Explanation Methods

Perturbation-Based

1. perturb
2. measure reaction
3. identify relevant information

Function-Based

1. treat the NN as function
2. compute simple quantities on it
3. construct explanation

Surrogate- / Sampling-Based

1. locally approximate prediction using simple function
2. explain simple function

Structure-Based

1. if the decision is too complex to explain, break the function into subfunctions.
2. explain each subfunction
3. meaningfully aggregate

Explanation Methods

Perturbation-Based

1. perturb
2. measure reaction
3. identify relevant information

Function-Based

1. treat the NN as function
2. compute simple quantities on it
3. construct explanation

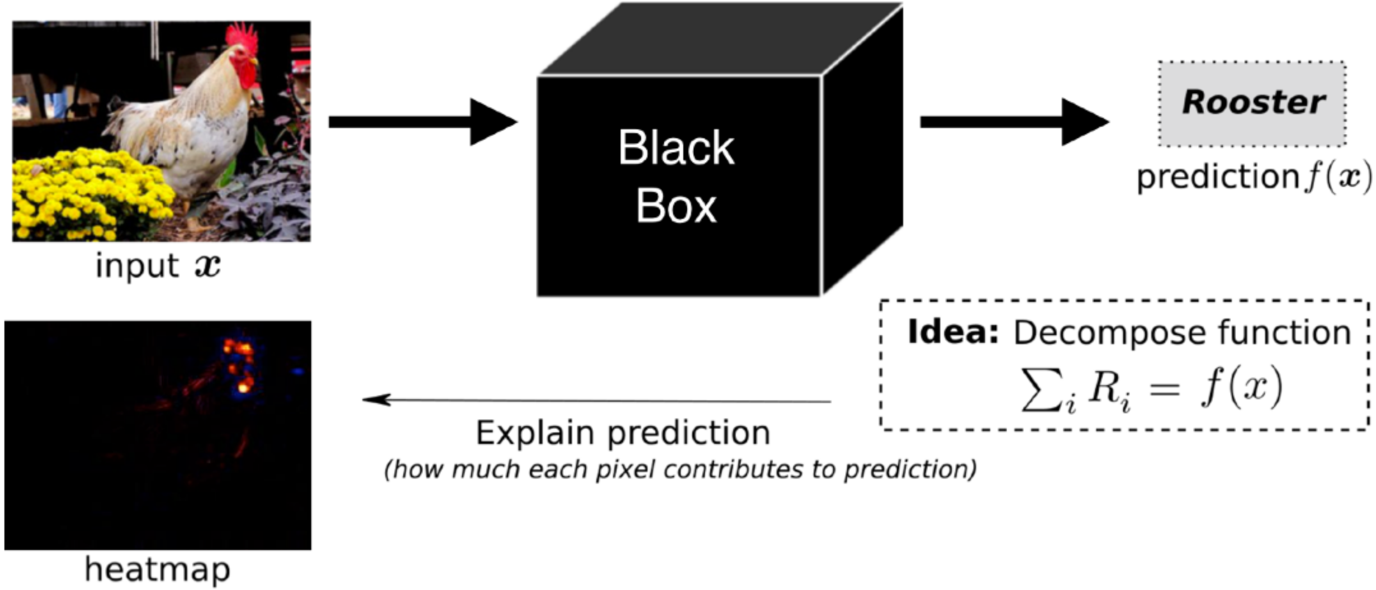
Surrogate- / Sampling-Based

1. locally approximate prediction using simple function
2. explain simple function

Structure-Based

1. if the decision is too complex to explain, break the function into subfunctions.
2. explain each subfunction
3. meaningfully aggregate

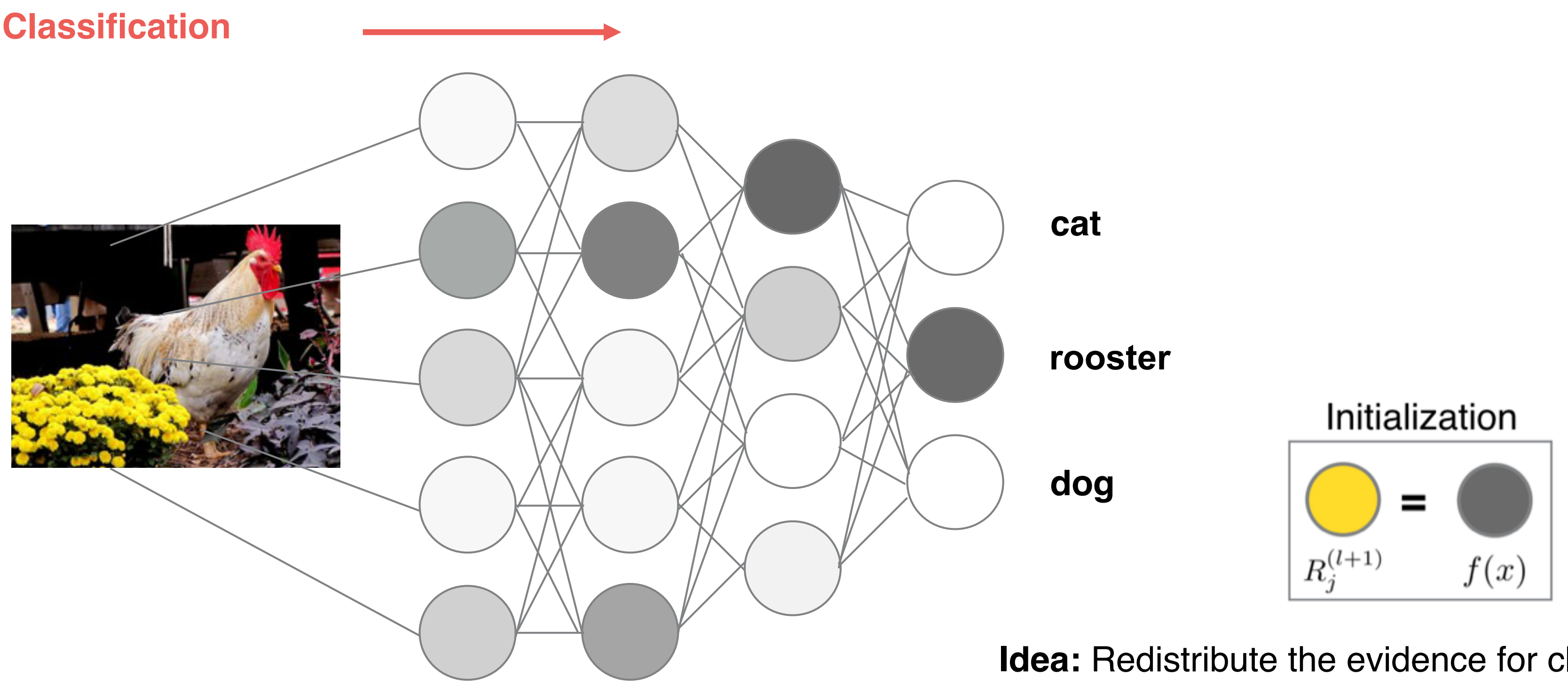
Layer-wise Relevance Propagation



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

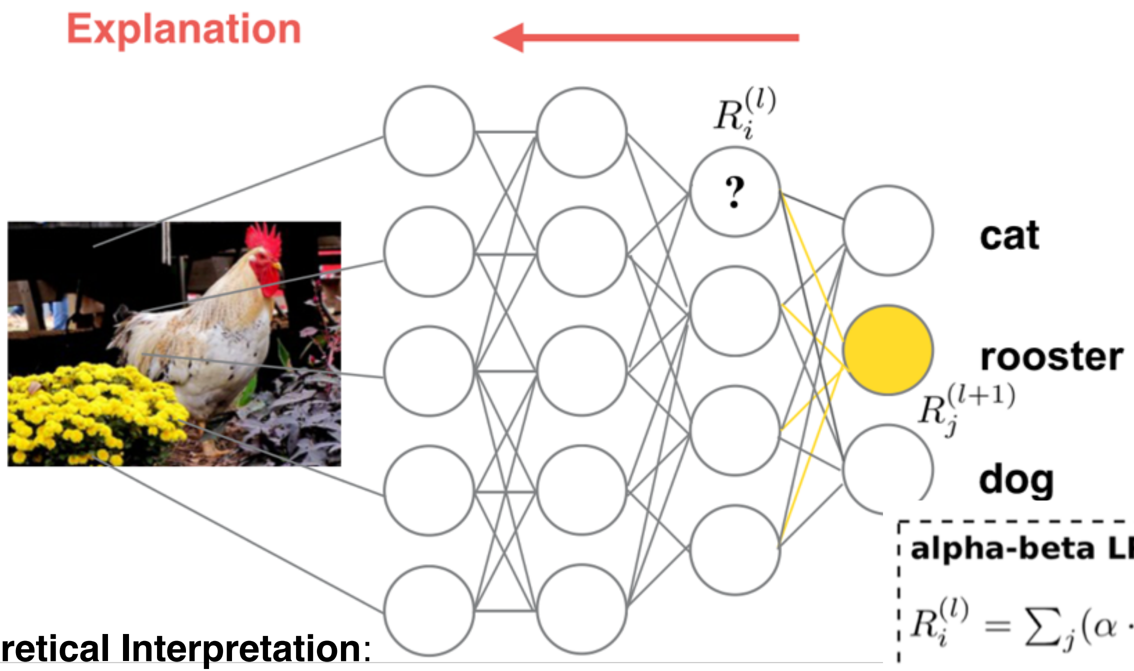
(Bach et al., PLOS ONE, 2015)

Layer-wise Relevance Propagation



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

Layer-wise Relevance Propagation



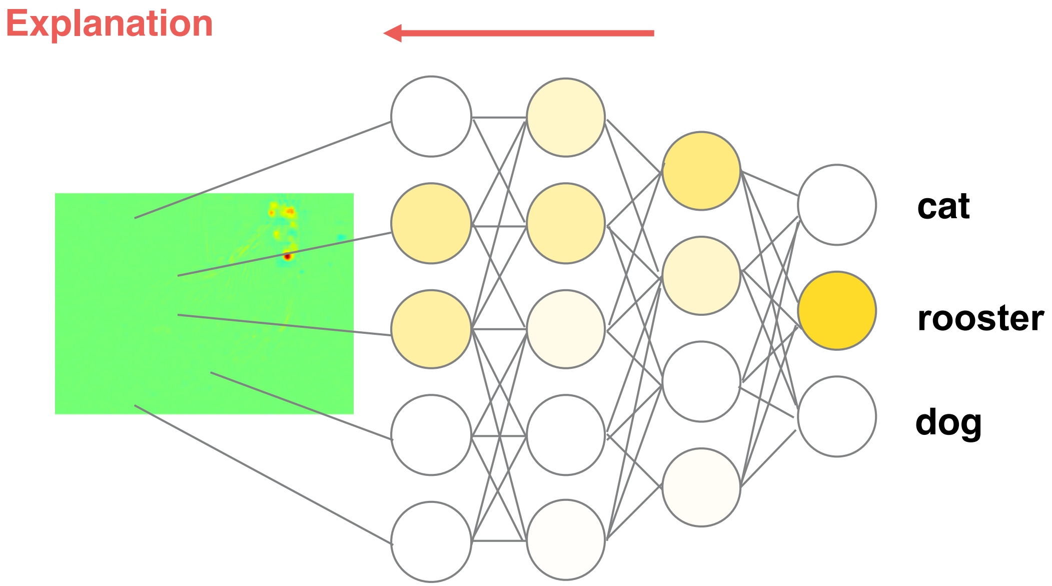
alpha-beta LRP rule (Bach et al. 2015)

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

where $\alpha + \beta = 1$

Theoretical Interpretation:
Deep Taylor Decomposition

Layer-wise Relevance Propagation

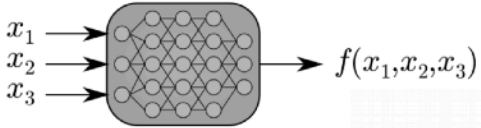


Layer-wise relevance conservation

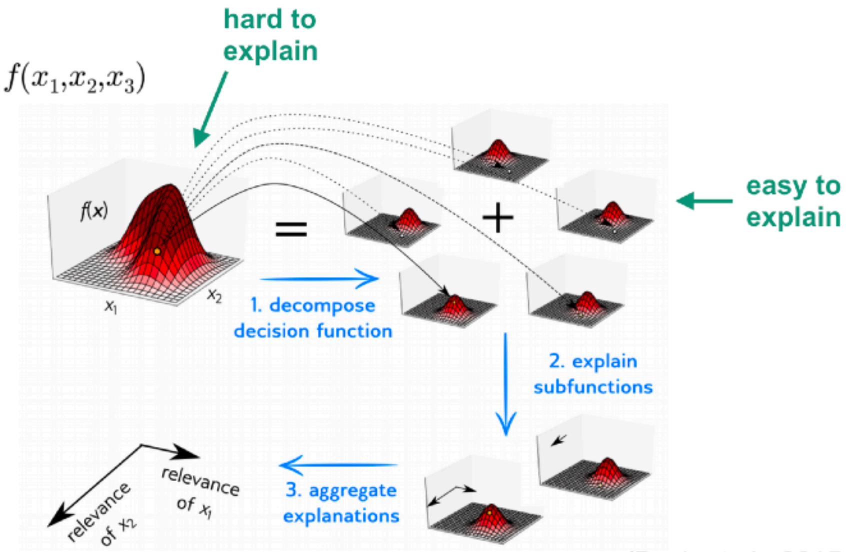
$$\sum_i R_i = \ldots = \sum_i R_i^{(l)} = \sum_j R_j^{(l+1)} = \ldots = f(x)$$

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)

Best Practice for LRP

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

Evaluating Explanations

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

Pointing Game
[Zhang’16]

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

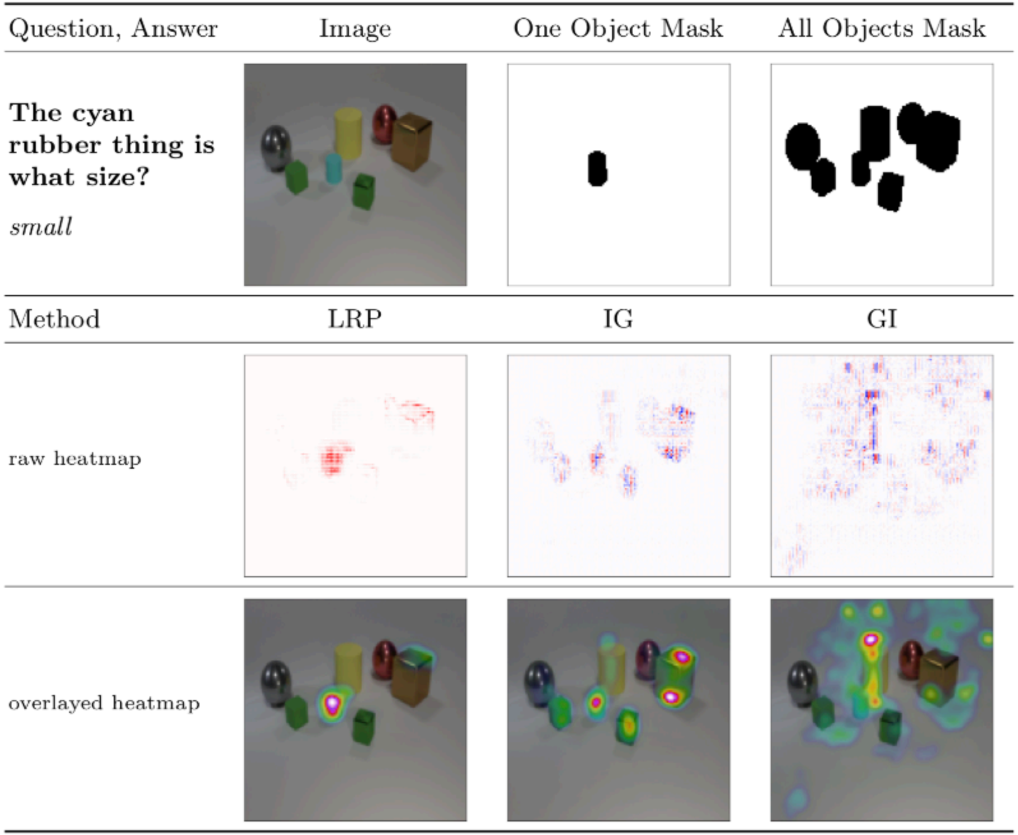
Task Specific Evaluation
[Poerner’18]

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

Using Ground Truth
[Arras’19]

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

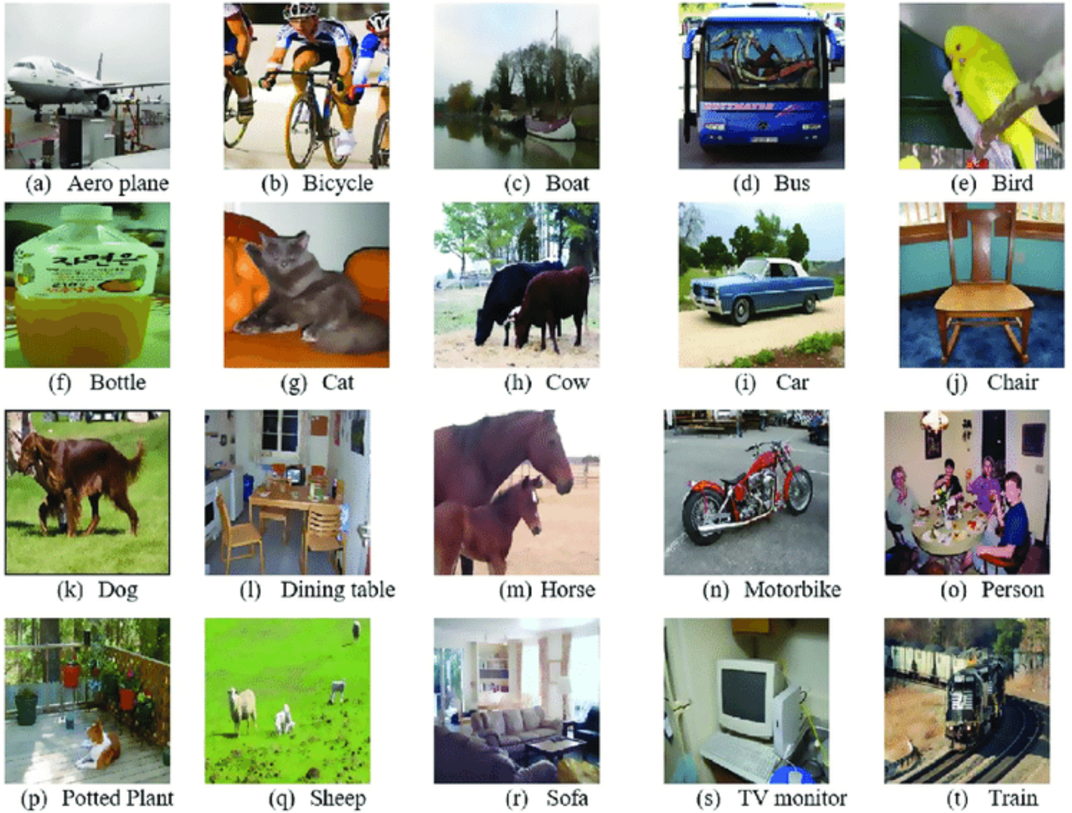
Evaluating Explanations



Using Ground Truth
[Arras'19]

Example

PASCAL VOC Challenge (2005 - 2012)



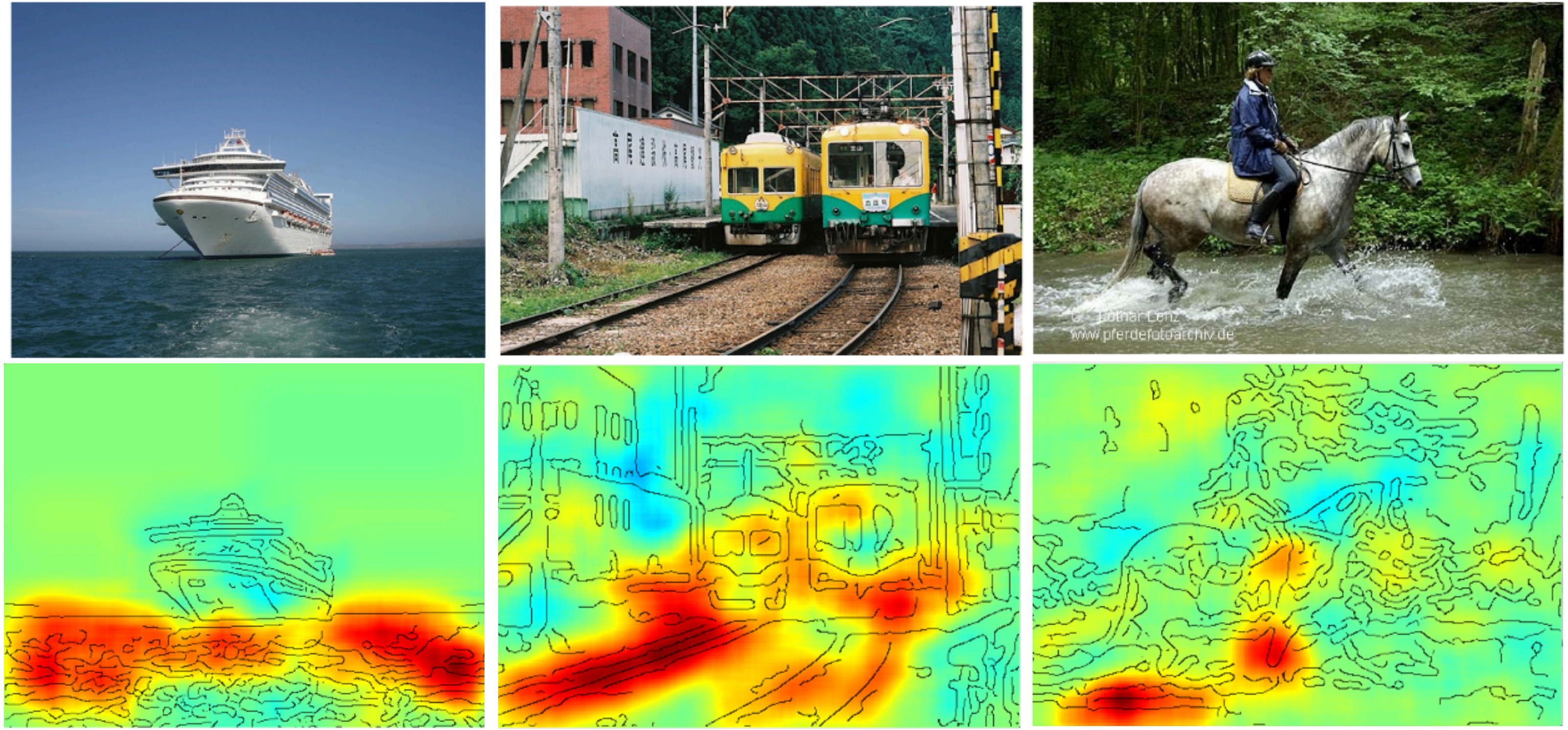
	mean
SRN+ [?]	88.8
SFA_NET [?]	87.5
SE [?]	86.5
LIG_DCNN_FEAT_ALL [?]	85.4
S&P_OverFeast_Fast_Bayes [?]	82.8
NUSPSL_CTX_GPM_SCM [?]	82.2
BCE_loss [?]	82.1
Resnet [?]	80.7
CNN_SIGMOID [?]	79.7
NUSPSL_CTX_GPM [?]	78.6
NUS_Context_SVM [?]	78.3
NLPR_PLS_SSVW [?]	78.3
Semi-Semantic Visual Words & Partial 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

Unmasking Clever Hans Predictors

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

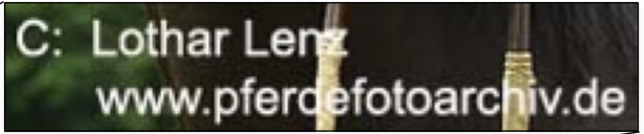
Unmasking Clever Hans Predictors

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

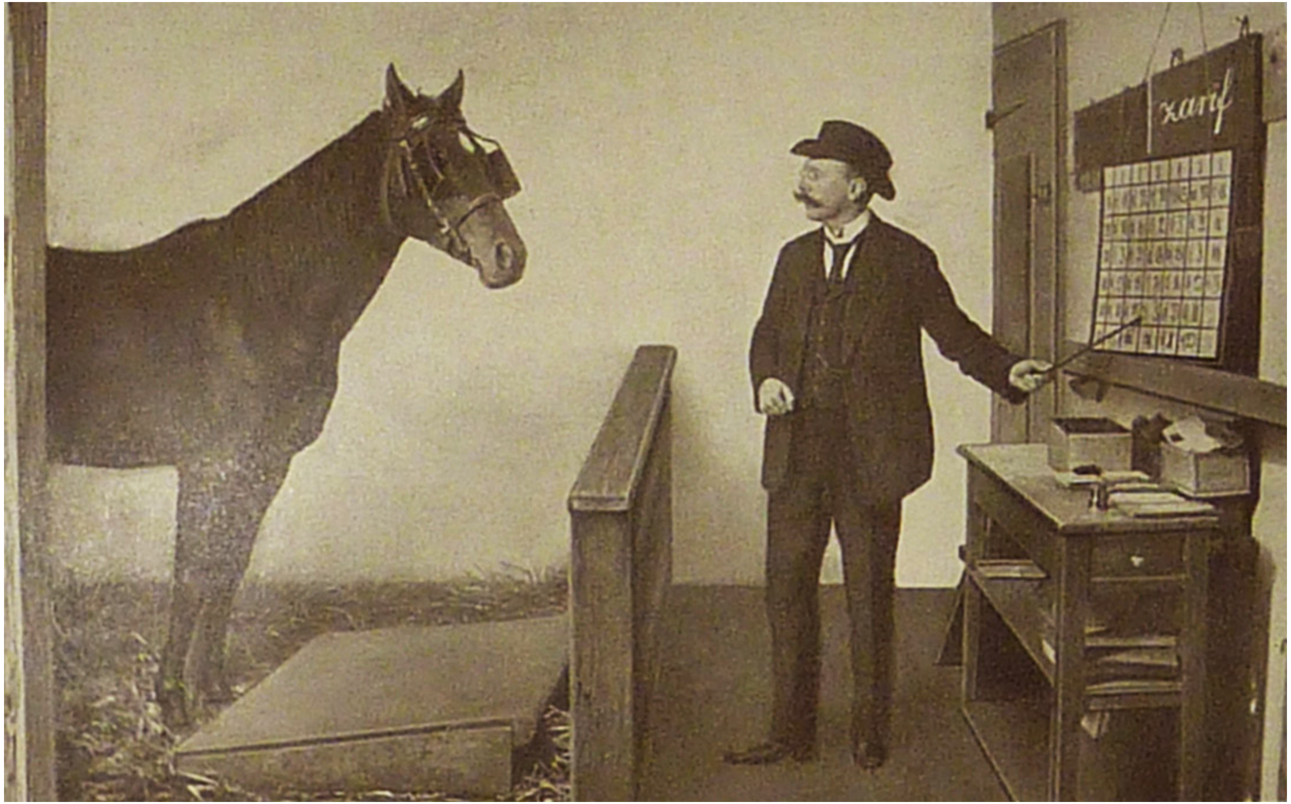


Unmasking Clever Hans Predictors

‘horse’ images in PASCAL VOC 2007



Unmasking Clever Hans Predictors



We need to ensure that models are right for the right reason!

Summary

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

Many good reasons for "explaining"

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

Explainability can help to unmask Clever Hans predictors (and much more)

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

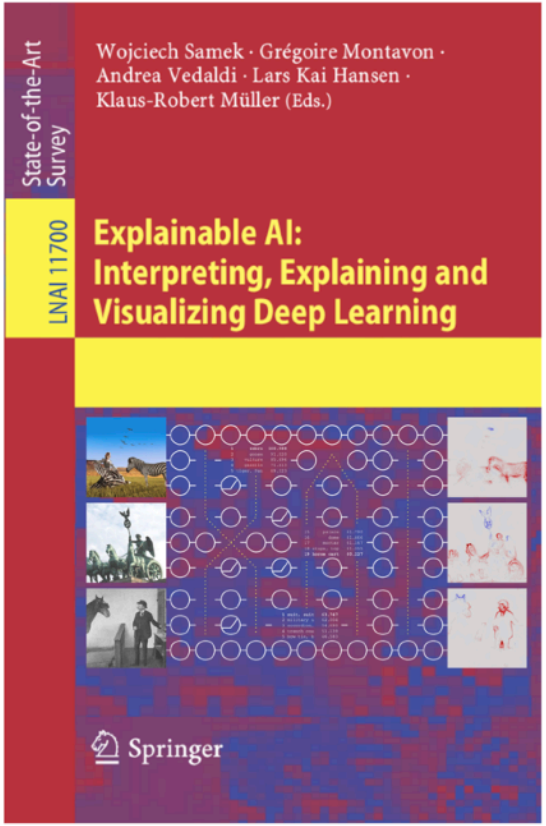
Application to Neural Network Pruning

- 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