Explaining deep learning for identifying structures and biases in computer vision

A Talk at: Interpretable ML in Vision@ICCV 2019.

Joint work with W. Samek, S. Lapuschkin (nee Bach), G. Montavon, K.-R. Müller, and deserving others
Alexander Binder

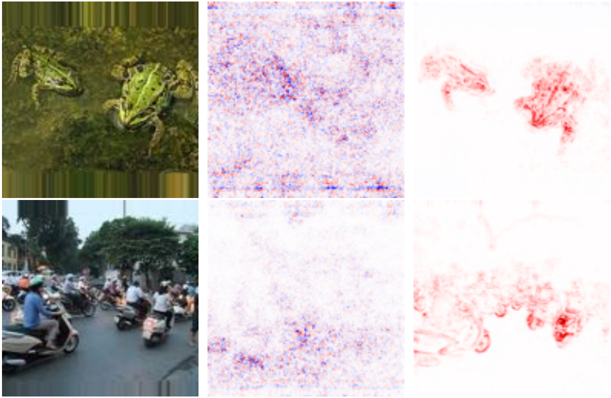
October 28, 2019

Fraunhofer Heinrich Hertz Institute
SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN
berlin

What is a possible explanation of a prediction? for images: (Densenet121, Keras+innvestigate, 2019)

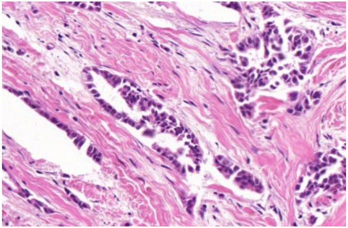
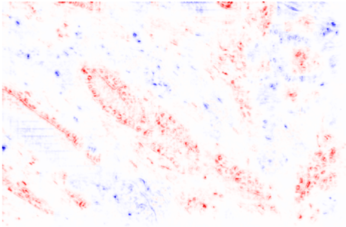
- case of images: compute a score for every pixel

image | gradient | LRP-α-β



What is a possible explanation of a prediction?

- case of images: compute a score for every pixel
  - patch-wise classification: label $= 1$ if patch contains breast cancer
  - pixel-wise explanation
- general case: score for every dim of an input sample $x = (x_1, \ldots, x_d, \ldots, x_D)$

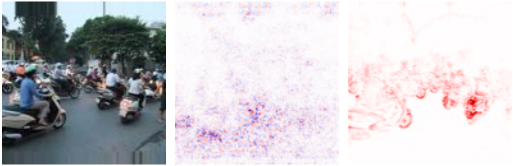



What is LRP as explanation? (Densenet121, Keras+innvestigate, 2019)

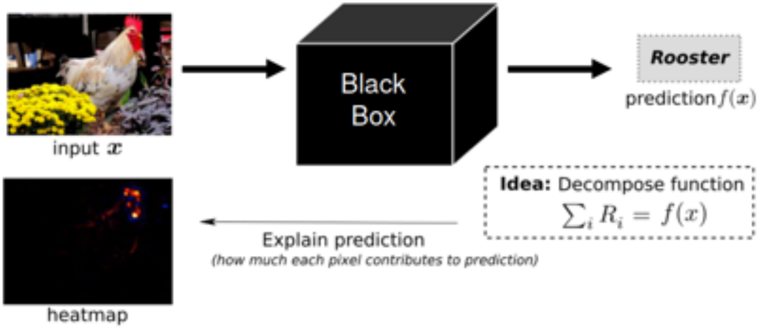
- given: A. trained model f, B. a prediction $f(x)$ for input $x = (x_1, \ldots, x_d, \ldots, x_D)$.
- general case: LRP computes a relevance score $r_d(x)$ for every input dimension x_d of input x explaining the prediction $f(x)$, such that approximately:

$$f(x) \approx \sum_{d=1}^{D} r_d(x) \leftarrow \text{decomposition with constraints} \tag{1}$$

image gradient LRP-α-β

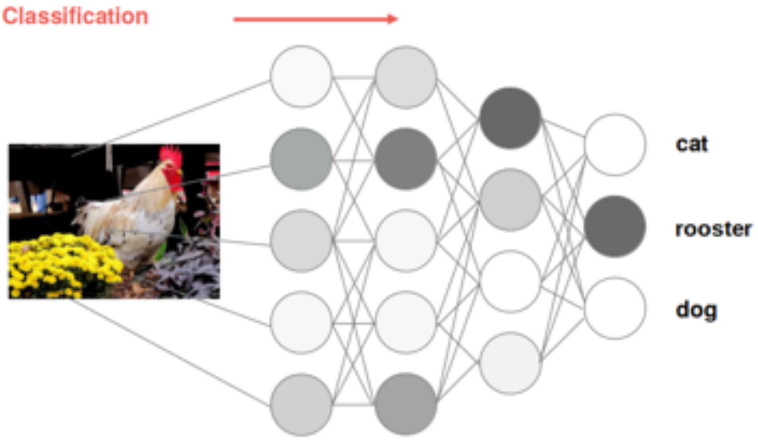


What is a possible explanation of a prediction?

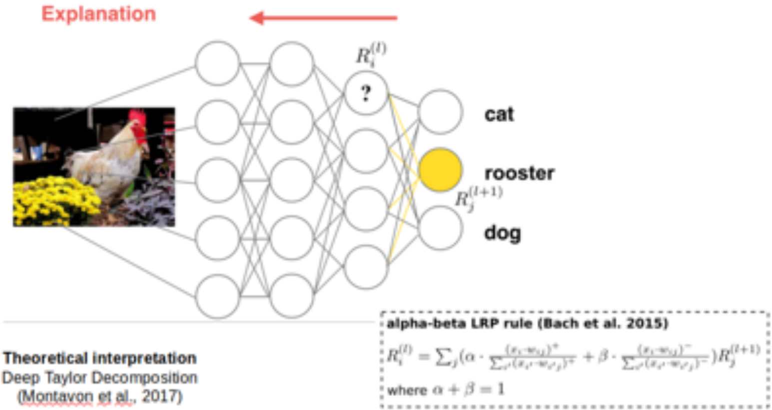


Layer-wise Relevance Propagation (LRP)
(Bach et al., PLOS ONE, 2015)

What is a possible explanation of a prediction?



What is a possible explanation of a prediction?



Trivial rules

Given $f(x)$, can obtain desired decomposition

$$f(\mathbf{x}) = \sum_{d=1}^{D} r_d(\mathbf{x}) \text{ by e.g.} \tag{2}$$

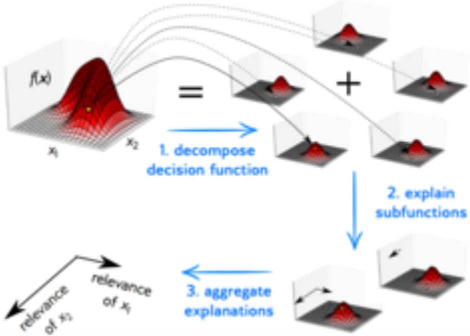
$$r_d(\mathbf{x}) = f(\mathbf{x})/D \tag{3}$$

$$r_d(\mathbf{x}) = \begin{cases} f(\mathbf{x}) & d = 1 \\ 0 & \text{else} \end{cases} \tag{4}$$

- underdetermined, many non-plausible decompositions
- need additional constraints
- theoretical foundation yielding constraints: Deep Taylor framework
  - Taylor decomposition of every single neuron with customized root points.

Deep Taylor Decomposition

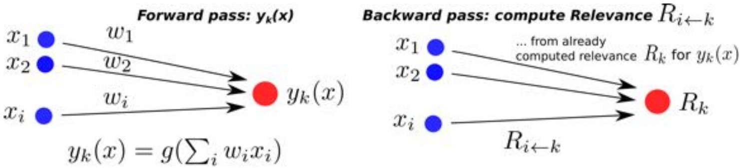
LRP's idea: To robustly explain a model, leverage the neural network structure of the decision function.



Each explanation step:
- easy to find good root point
- no gradient shattering

(Montavon et al., 2017
Montavon et al. 2018)

Relevance distribution for one neuron: example ϵ-rule



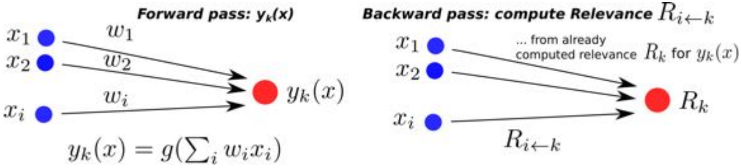
ϵ-rule:

$$R_{i\leftarrow k}(\mathbf{x}) \propto R_k h(w_i x_i) \tag{5}$$

$$R_{i\leftarrow k}(\mathbf{x}) = R_k \frac{w_i x_i}{\sum_{i'} w_{i'} x_{i'} + b + \epsilon \cdot \mathrm{sign}} \tag{6}$$

- ϵ – dampening factor, numerical stabilization
- recommended for fully connected layers and good for LSTMs (cf. Leila Arras et al.)
- NOT recommended for conv layers

Relevance distribution for one neuron: example α-β-rule



β-rule:

$$R_{i \leftarrow k}(\mathbf{x}) \propto R_k h(w_i x_i) \tag{7}$$

$$R_{i \leftarrow k}(\mathbf{x}) = R_k \left((1+\beta) \frac{(w_i x_i)_+}{\sum_{i'} (w_{i'} x_{i'})_+ + b_+} - \beta \frac{(w_i x_i)_-}{\sum_{i'} (w_{i'} x_{i'})_- + b_-} \right) \tag{8}$$

- β – controls ratio of negative to positive evidence.
- $\beta = 0$ only positive evidence (analogous to e.g. guided backprop)
- suitable for conv layers (with modifications: batchnorm layers)

Gradient $\times$ Input?

Motivation

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

Gradient x Input

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

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

$\mathbf{x}$  $f(\mathbf{x})$

Gradient $\times$ Input?

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

Perturbation Analysis

$$f(\mathbf{x}) = \sum_{i=1}^{d} x_i w_i + b$$

Taylor Expansions

$$\forall_{\mathbf{x}, t \geq 0} : \; f(t\mathbf{x}) = t f(\mathbf{x})$$

Gradient x Input

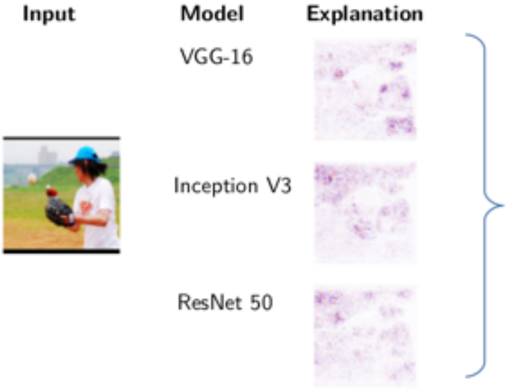
$$\forall_i : \; R_i = [\nabla f(\mathbf{x})]_i \cdot x_i$$

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

$\mathbf{x}$  $f(\mathbf{x})$

Question: Does it work in practice?

Gradient $\times$ Input?

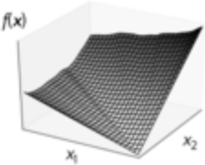


Observation:
Explanations are noisy.

Gradient $\times$ Input?

Two reasons why explanations are noisy:

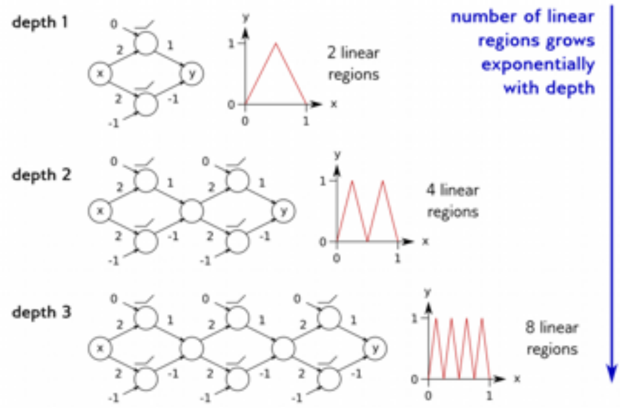
1. Not local enough. Too much context introduced when multiplying by the input.



2. Shattered gradient problem $\rightarrow$ gradient of deep nets has low informative value

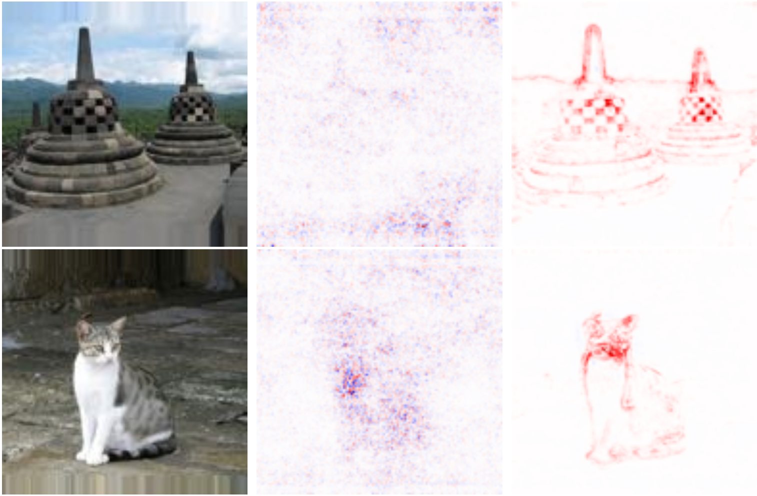
Gradient $\times$ Input?

The Shattered gradients problem [Montufar'14, Balduzzi'17]



Examples (Densenet121, Keras, 2019)

img gradient α-β

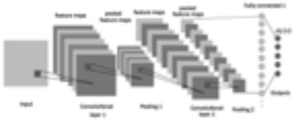


hybrid rule: $\beta = 0$ for conv layers, $\epsilon = 0.01$ for fc layer

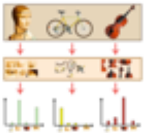
Tell them something interesting!

LRP Applied to Variety of Models

Convolutional NNs
(Bach'15, Binder'16, Arras'17 ...)

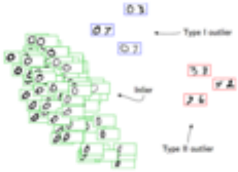


BoW models
(Bach'15, Lapuschkin'17 ...)

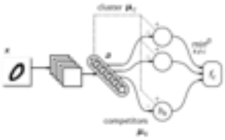


LSTM
(Arras'17, Arras ...
Hochreiter et al. 2019)

One-class SVM
(Kauffmann'18)

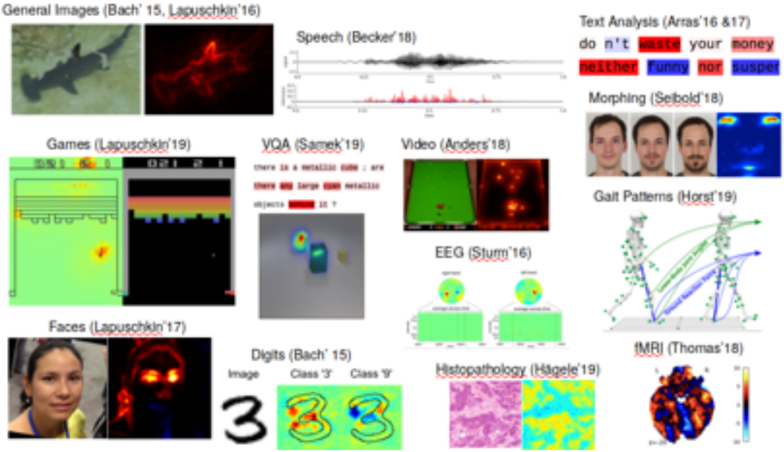


Clustering
(Kauffmann'19)



explaining unsupervised learning

LRP Applied to Variety of Tasks



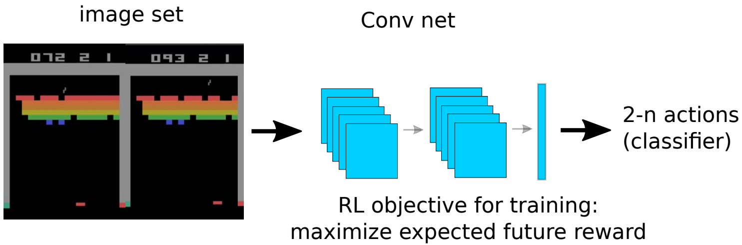
The value of explanations

A. application case: identify action strategies in reinforcement learning predictors
B. general: Identify Biases in Train+Test data (where labels do not help you at all)
C. medical imaging: Identify Fail Cases *without labelling efforts* $\longrightarrow$ Iterative Dataset Design
D. application case: LRP in neuroscience

LRP: DNN and Atari Breakout

A. application case: identify action strategies in reinforcement learning predictors

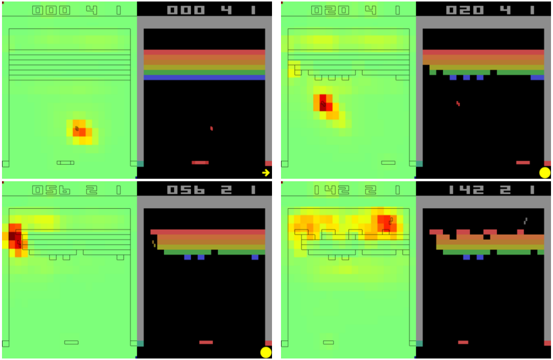
Trained a reinforcement learning classifier according to Mnih et al's Nature 2016 paper:
Volodymyr Mnih et al. Human-level control through deep reinforcement learning,
Nature 518, pages 529533, 2015



LRP: DNN and Atari Breakout

Trained a reinforcement learning classifier according to Mnih et al's Nature 2016 paper.

Explain a test game. LRP helps to discover strategies: building a tunnel.



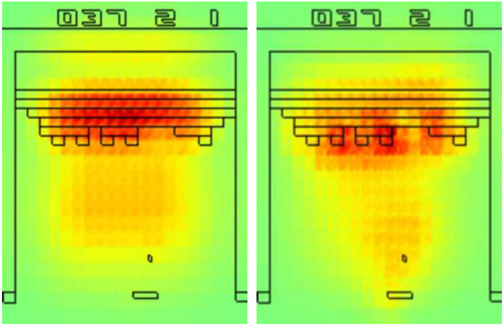
Lapuschkin et al., Unmasking Clever Hans predictors and assessing what machines really learn,

Nature Communications, 2019

LRP: DNN and Atari Breakout

Trained a reinforcement learning classifier according to Mnih et al's Nature 2016 paper.

LRP can help to discover strategies: building a tunnel - evolution of focus during training



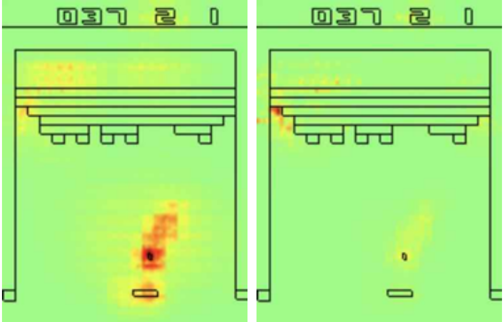
epoch 0 and 6

Lapuschkin et al., Unmasking Clever Hans predictors and assessing what machines really learn, Nature Communications, 2019

LRP: DNN and Atari Breakout

Trained a reinforcement learning classifier according to Mnih et al's Nature 2016 paper.

LRP can help to discover strategies: building a tunnel - evolution during training



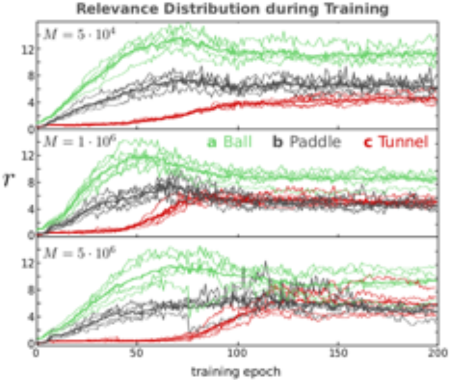
epoch 50 and 100

Lapuschkin et al., Unmasking Clever Hans predictors and assessing what machines really learn,

Nature Communications, 2019

LRP: DNN and Atari Breakout

LRP can help to find parameters for fast learning of known strategies. Here: impact of M = replay memory size



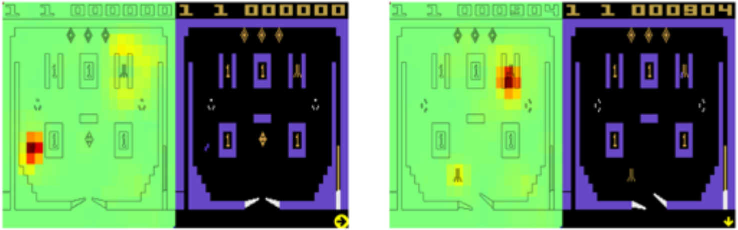
Lapuschkin et al., Unmasking Clever Hans predictors and assessing what machines really learn,

Nature Communications, 2019

LRP in reinforcement learning

Interpretability methods (here: LRP) can uncover complex relationships

Atari Pinball:



move ball 4 times over switch to activate a score multiplier.

.. if there are any

Lapuschkin et al., Unmasking Clever Hans predictors and assessing what machines really learn,

Nature Communications, 2019

Identify Biases in Train+Test data (where labels do not help you at all)

C. general: Identify Biases in Train+Test data (where labels do not help you at all)

At first: general images ... less careful about biases

Identify Biases in Train+Test data (where labels do not help you at all)

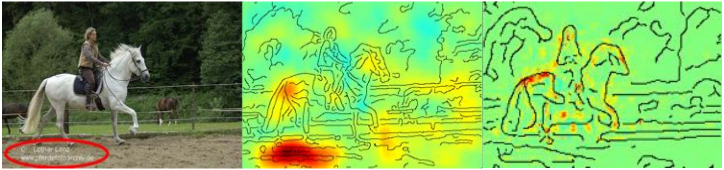
	aeroplane	bicycle	bird	boat	bottle	bus	car
Fisher	79.08%	66.44%	45.90%	70.88%	27.64%	69.67%	80.96%
DeepNet	88.08%	79.69%	80.77%	77.20%	35.48%	72.71%	86.30%
	cat	chair	cow	diningtable	dog	horse	motorbike
Fisher	59.92%	51.92%	47.60%	58.06%	42.28%	80.45%	69.34%
DeepNet	81.10%	51.04%	61.10%	64.62%	76.17%	81.60%	79.33%
	person	pottedplant	sheep	sofa	train	tvmonitor	mAP
Fisher	85.10%	28.62%	49.58%	49.31%	82.71%	54.33%	59.99%
DeepNet	92.43%	49.99%	74.04%	49.48%	87.07%	67.08%	72.12%

Analyzing Classifiers: Fisher Vectors and Deep Neural Networks, Lapuschkin et al., CVPR 2016

Identify Biases in Train+Test data (where labels do not help you at all)

	aeroplane	bicycle	bird	boat	bottle	bus	car
Fisher	79.08%	66.44%	45.90%	70.88%	27.64%	69.67%	80.96%
DeepNet	88.08%	79.69%	80.77%	77.20%	35.48%	72.71%	86.30%
	cat	chair	cow	diningtable	dog	horse	motorbike
Fisher	59.92%	51.92%	47.60%	58.06%	42.28%	80.45%	69.34%
DeepNet	81.10%	51.04%	61.10%	64.62%	76.17%	81.60%	79.33%
	person	pottedplant	sheep	sofa	train	tvmonitor	mAP
Fisher	85.10%	28.62%	49.58%	49.31%	82.71%	54.33%	59.99%
DeepNet	92.43%	49.99%	74.04%	49.48%	87.07%	67.08%	72.12%

Image Fisher Vector Deep Neural Net



Analyzing Classifiers: Fisher Vectors and Deep Neural Networks, Lapuschkin et al., CVPR 2016

SpRAy: semi-automatic discovery of correlations

Lapuschkin et al. Nature Communications 2019:

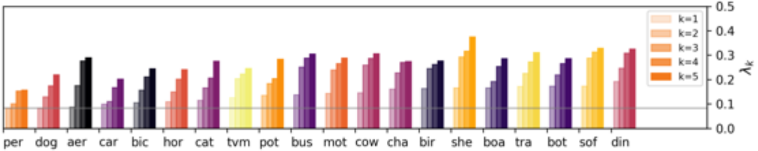
Principle

- compute heatmaps, pool them into a uniform low resolution 20×20
- compute binarized similarity w_{ij} between heatmaps of samples i and j using $k = \log$ sample size

$$w_{ij} = \begin{cases} 1 & \text{if } i \text{ is among the } k\text{-nearest neighbors of } j \end{cases} \tag{9}$$

- symmetrize $W = (w_{ij})_{i,j} \mapsto \max(w_{ij}, w_{ji})$
- compute eigenvalue/vectors of Laplacian $L = I - D^{-1/2} W D^{-1/2}$
- inspect eigenvalue gaps

SpRAy: DNN and Pascal VOC Aeroplane class

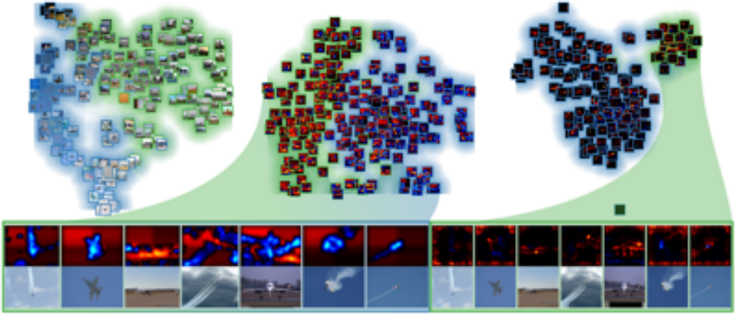


SpRAy: Two Large gaps in low eigenvalues for aeroplane – conspicuous.

Lapuschkin et al., Unmasking Clever Hans predictors and assessing what machines really learn,

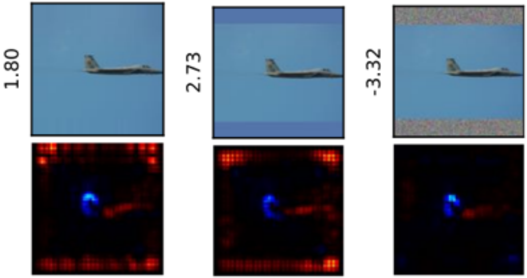
Nature Communications, 2019

SpRAy: DNN and Pascal VOC Aeroplane class



- t-sne shows one cluster where aeroplanes have strong evidence on edges due to data preparation artefact combined with frequency of blue sky.
- Did not wanted to use center crops: avoid cutting off object parts. So edges were padded with border pixels. This is used in one part of the aeroplane images as cue.

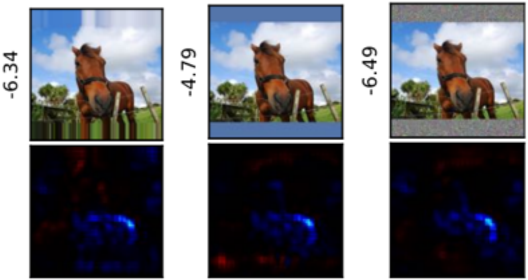
SpRAy: DNN and Pascal VOC Aeroplane class



Confirm that paddings are a cue:

- images with aeroplane predicted: changing borders to random noise destroys aeroplane scores
- images with no aeroplane predicted: changing borders to sky blue color improved aeroplane score, even random but constant color helps.

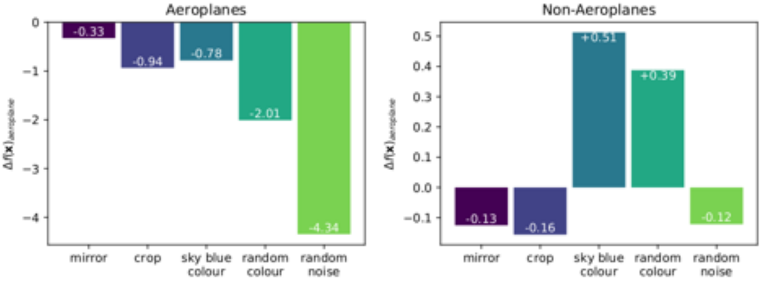
SpRAy: DNN and Pascal VOC Aeroplane class



Confirm that paddings are a cue:

- images with aeroplane predicted: changing borders to random noise destroys aeroplane scores
- images with no aeroplane predicted: changing borders to sky blue color improved aeroplane score, even random but constant color helps.

SpRAy: DNN and Pascal VOC Aeroplane class



Result show:

- identified another bias by inspecting heatmaps – this one is hard to see for humans: at borders (psychologically suppressed as irrelevant!) plus constant color in one class

Lapuschkin et al., Unmasking Clever Hans predictors and assessing what machines really learn,

Nature Communications, 2019

Identify Biases in Train+Test data (where labels do not help you at all)

C. general: Identify Biases in Train+Test data (where labels do not help you at all)

Identify Biases in Train+Test data (where labels do not help you at all)

C. general: Identify Biases in Train+Test data (where labels do not help you at all)

and now to something more relevant please!

Medical datasets

Identify Biases in Train+Test data (where labels do not help you at all)

Haegele et al., Resolving challenges in deep learning-based analyses of histopathological images using explanation methods, arxiv 2019:

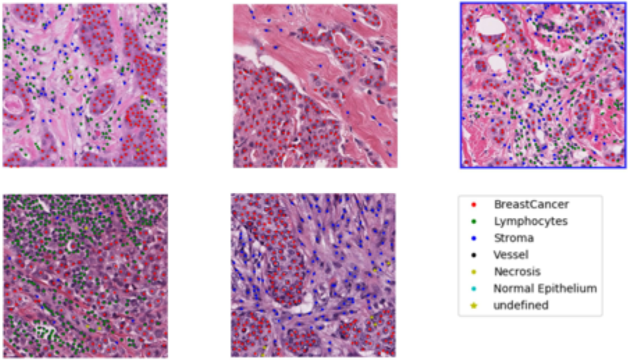
?– Are heatmaps of patch-level classifiers *quantifiably* meaningful in terms of resolution at cell nucleus level ? Do they consider nuclei as evidence? How good are heatmaps in terms of measured localization accuracy?

?– Are heatmaps useful to resolve biases in histopathology?

- systematic biases
- class-correlation biases
- sampling biases
- LRP for evaluating the impact of class sampling ratios

Quantifying heatmaps on cell level

Three datasets: Annotate nuclei densely.

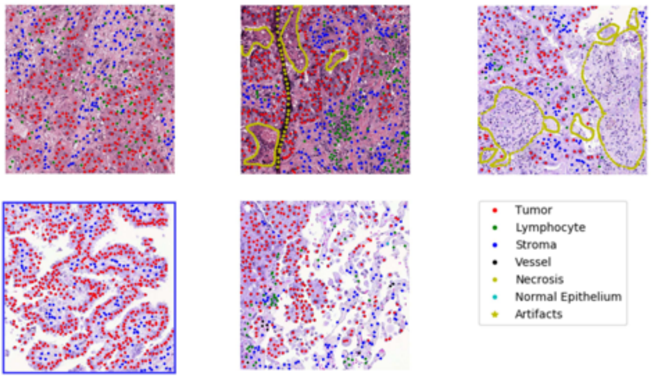


BRCA

Haegele et al., Resolving challenges in deep learning-based analyses of histopathological images using explanation methods, arxiv 2019

Quantifying heatmaps on cell level

Three datasets: Annotate nuclei densely.

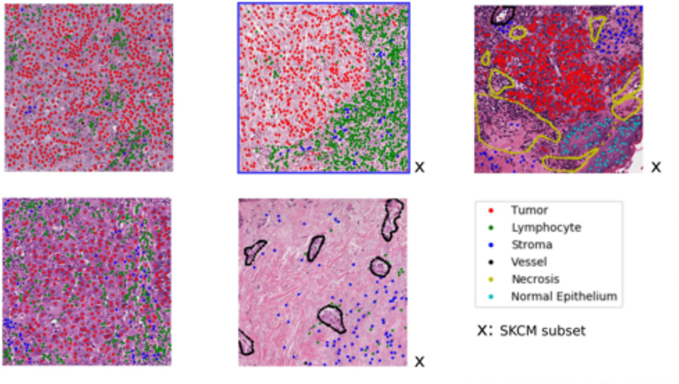


LUAD (lung)

Haegele et al., Resolving challenges in deep learning-based analyses of histopathological images using explanation methods, arxiv 2019

Quantifying heatmaps on cell level

Three datasets: Annotate nuclei densely.



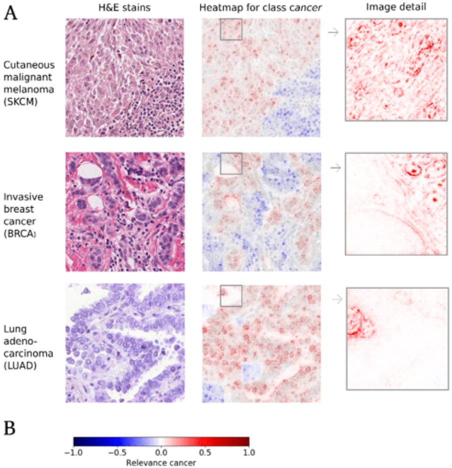
SKCM (Melanoma)

Haegele et al., Resolving challenges in deep learning-based analyses of histopathological images using explanation methods, arxiv 2019

Quantifying heatmaps on cell level

Train patch classifier, compute heatmaps.

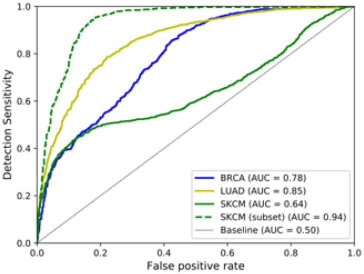
Tumor entity	Number of cases	Total number of patches	Number of tumor patches	F_1
SKCM	38	26,746	19,139 (71.6 %)	**91.5%**
BRCA	72	2,748	2,308 (84.0 %)	**92.1%**
LUAD	39	13,165	4,805 (36.5 %)	**94.6%**



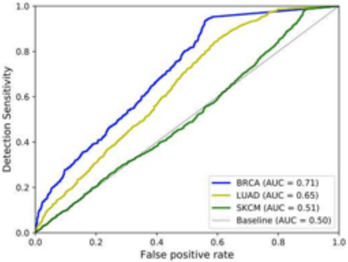
Haegele et al., Resolving challenges in deep learning-based analyses of histopathological images using explanation methods, arxiv 2019

Quantifying heatmaps on cell level

Do we need high res methods like LRP or guided BP ? (a lil bit bashing please be forgiven)



LRP



GradCAM

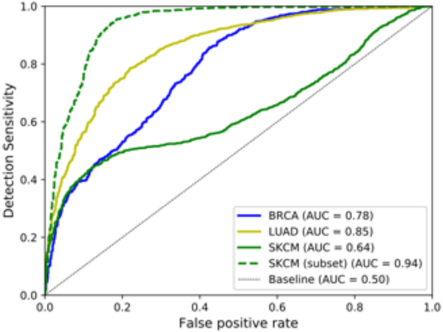
Haegele et al., Resolving challenges in deep learning-based analyses of histopathological images using explanation methods, arxiv 2019

Quantifying heatmaps on cell level

Evaluation Data on nucleus level

Overview of the available annotations for ROC curves.

Tumor entity	Total number of cells	Number of cancer cells
BRCA	1,803	820
SKCM	3,961	2,247
LUAD	2,722	1,650

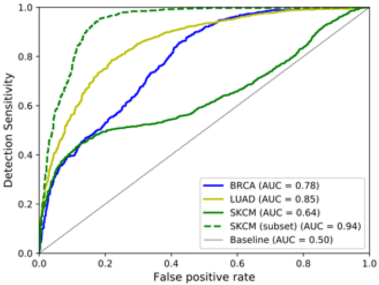


Haegele et al., Resolving challenges in deep learning-based analyses of histopathological images using explanation methods, arxiv 2019

Quantifying heatmaps on cell level

Evaluation Data on the level of nuclei:

- Poor sensitivity on mid ranges for SKCM and BRCA.
- Inspecting heatmaps for SKCM reveals two slides with dense tissue invading lymphocytes – receiving moderately positive scores.
- Points at insufficient sampling of patches with TiLs in training :) .

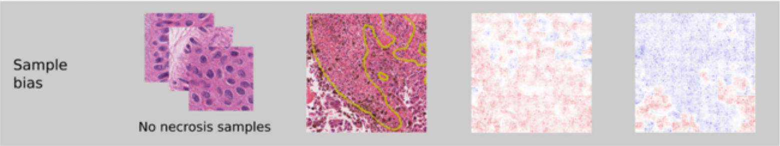


Haegele et al., Resolving challenges in deep learning-based analyses of histopathological images using explanation methods, arxiv 2019

Sampling bias

- left heatmap: false positive scores on unlabeled subclass.
- right heatmap: after augmenting training dataset with necrosis samples (negative labeled)

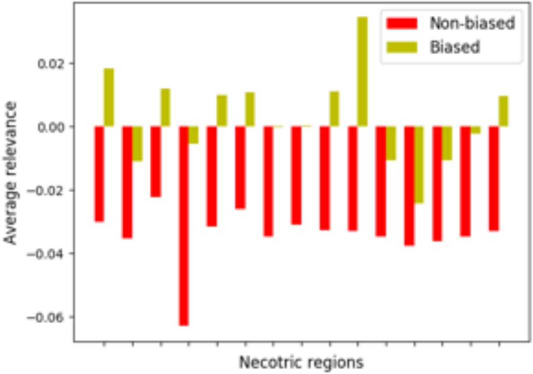
training image	test image	heatmap with bias	heatmap w/o bias



Haegele et al., Resolving challenges in deep learning-based analyses of histopathological images using explanation methods, arxiv 2019

Sampling bias

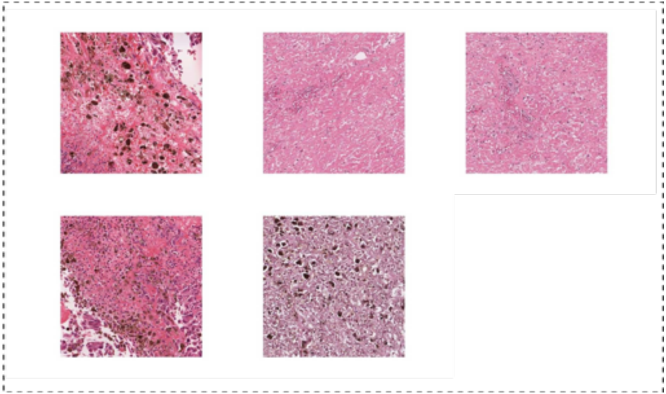
Retraining has statistically visible effect.



Haegele et al., Resolving challenges in deep learning-based analyses of histopathological images using explanation methods, arxiv 2019

Sampling bias

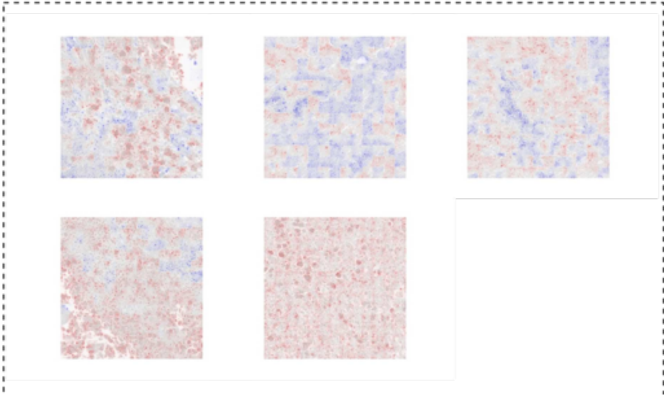
Retraining has a visually visible effect, too.



Haegele et al., Resolving challenges in deep learning-based analyses of histopathological images using explanation methods, arxiv 2019

Sampling bias

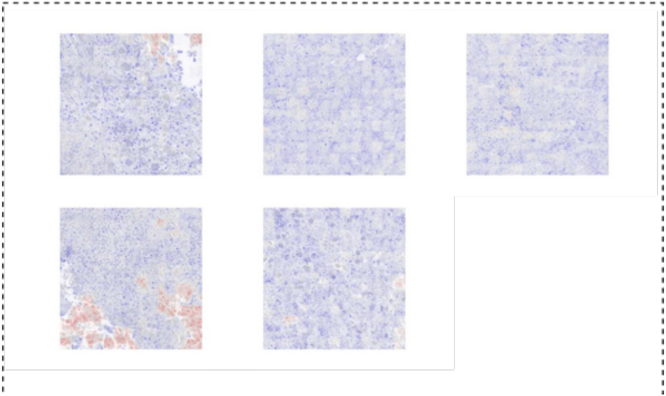
Here: *without* necrosis samples.



Haegele et al., Resolving challenges in deep learning-based analyses of histopathological images using explanation methods, arxiv 2019

Sampling bias

Here: *with* necrosis samples.

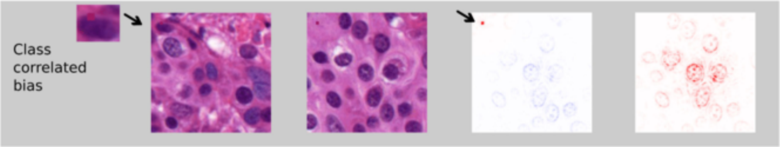


your version1 labels and test set error cannot discover it

Haegele et al., Resolving challenges in deep learning-based analyses of histopathological images using explanation methods, arxiv 2019

Class-correlation bias

training image	test image	heatmap with bias	heatmap w/o bias



- biases are identifiable
- test set labels are of no help (!) for discovery
- debiasing improves explanations

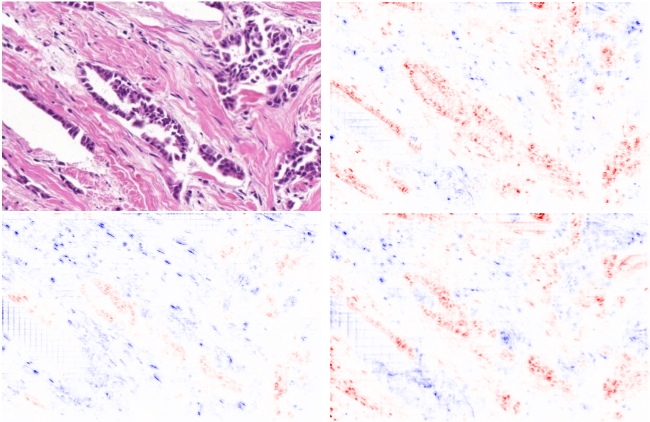
Haegele et al., Resolving challenges in deep learning-based analyses of histopathological images using explanation methods, arxiv 2019

Identify Fail Cases *without labelling efforts*: Evaluate Impact of data augmentation

Image scaling ?

orig

80%



100%

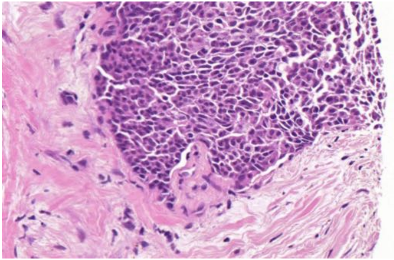
66%

Medical Data: Identify Fail Cases *without labelling efforts*

C. medical imaging: Identify Fail Cases *without labelling efforts*
$\longrightarrow$ Iterative Dataset Design

Why not just using test error ?

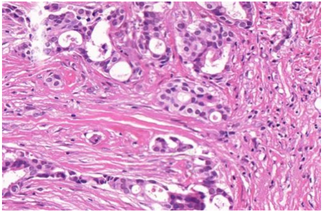
- some problems: labels very costly, unlabeled data abundant

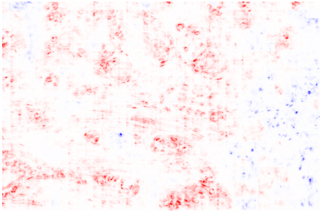


Identify Fail Cases *without labelling efforts*

More Importantly:

- decide what unlabeled data to add into next iteration of train and test set



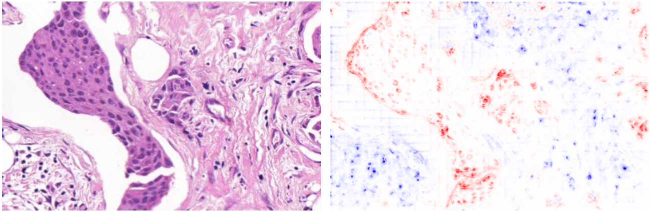


- Interpretability for efficiency in the selection step before labelling!

Identify Fail Cases *without labelling efforts*

More Importantly:

- decide what unlabeled data to add into next iteration of train and test set – precursor to labelling.

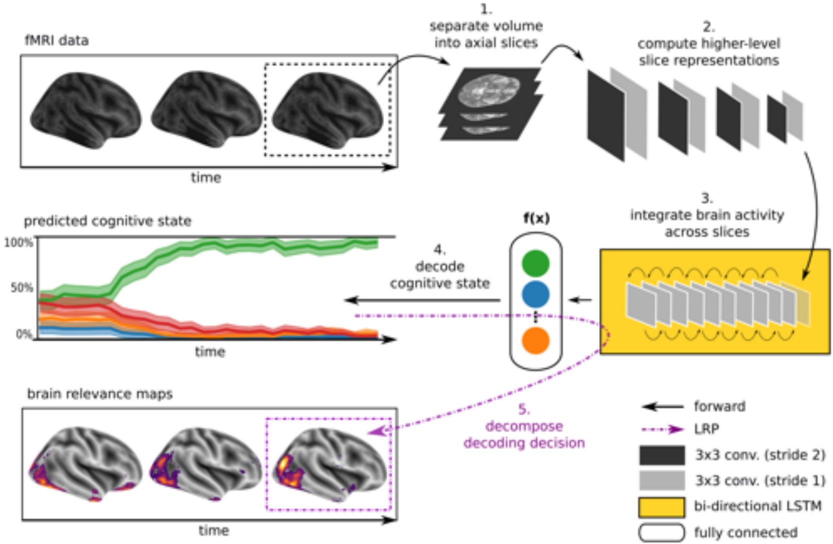


- Interpretability for efficiency in the selection step before labelling!

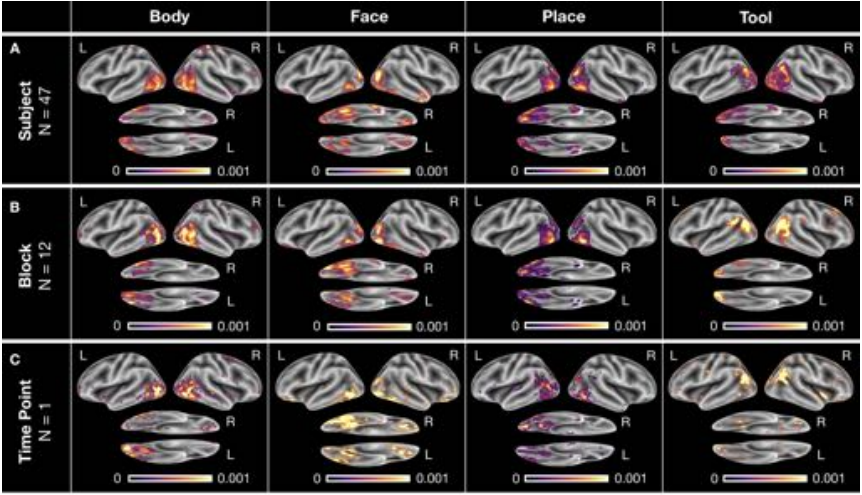
LRP in Neuroscience

Thomas et al.
Analyzing Neuroimaging Data Through Recurrent Deep Learning Models, arxiv 2019

LRP in Neuroscience



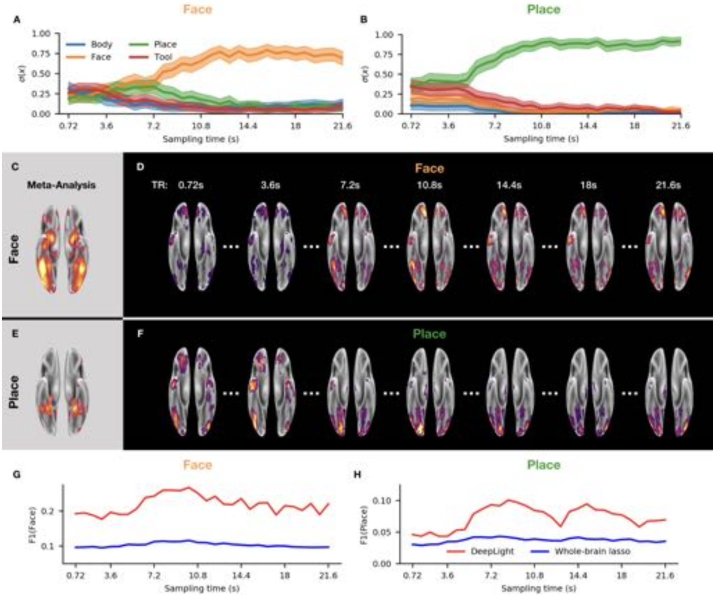
LRP in Neuroscience



Thomas et al.

Analyzing Neuroimaging Data Through Recurrent Deep Learning Models, arxiv 2019

LRP in Neuroscience



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.

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

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.

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

References

Application to Text

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 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 Video

C Anders, G Montavon, W Samek, KR Müller. Understanding Patch-Based Learning by Explaining Predictions. *arXiv:1806.06926*, 2018.

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.

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.

A Binder, M Bockmayr, M Hägele and others. Towards computational fluorescence microscopy: Machine learning-based integrated prediction of morphological and molecular tumor profiles. *arXiv:1805.11178*, 2018

References

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.

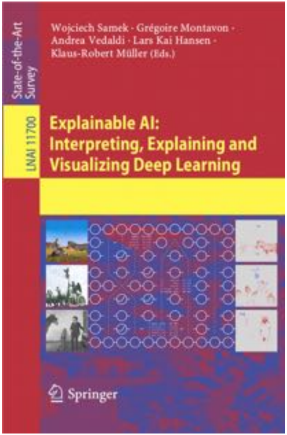
L Arras, A Osman, KR Müller, W Samek, Evaluating Recurrent Neural Network Explanations. *Proceedings of the ACL'19 Workshop on BlackboxNLP,* Association for Computational Linguistics, 113-126, 2019.

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.

New book 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

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`

papers, demos, ice cream at: `www.explain-ai.org`

Questions?!