Fraunhofer
Heinrich Hertz Institute

Interpretable & Transparent Deep Learning

Fraunhofer HHI, Machine Learning Group
Wojciech Samek

Record Performances with ML

Research projects

Deep Net outperforms humans in image classification

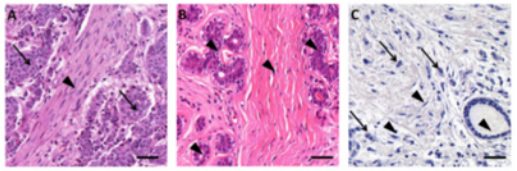
IMAGENET

AlphaGo beats Go human champ

Visual Reasoning

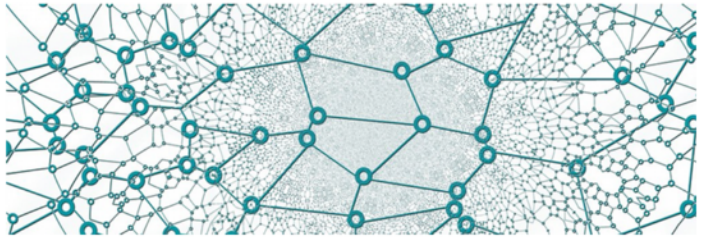
What size is the cylinder that is left of the brown metal thing that is left of the big sphere?

Medical Diagnosis



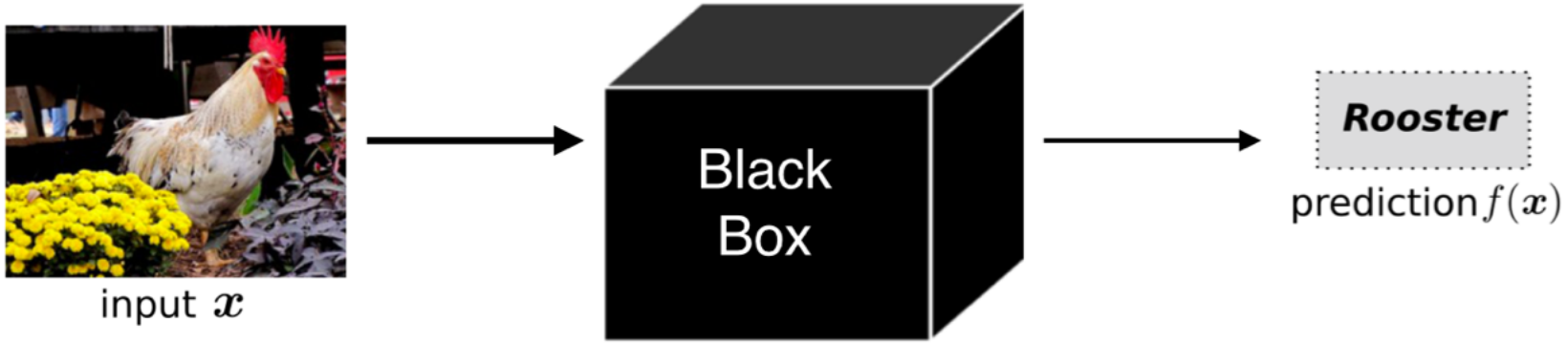
Autonomous Driving

Networks (smart grids, etc.)



Safety critical applications

Need for Interpretability



verify system

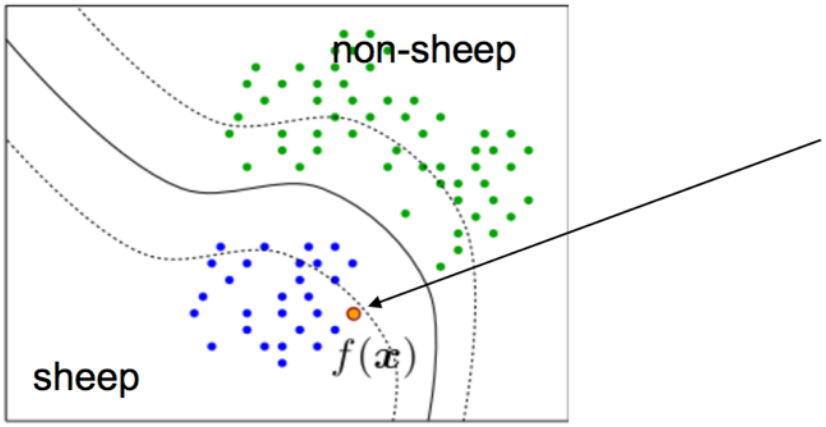
legal aspects

learn new strategies

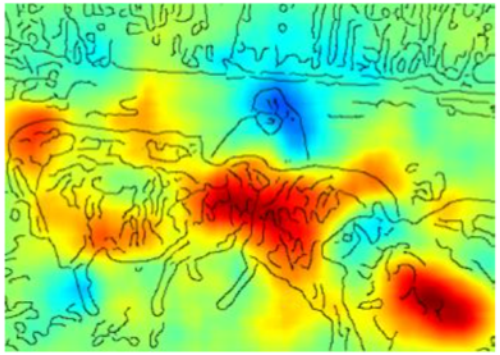
understand weaknesses

Need for Interpretability

"Why is a given image classified as a sheep?"



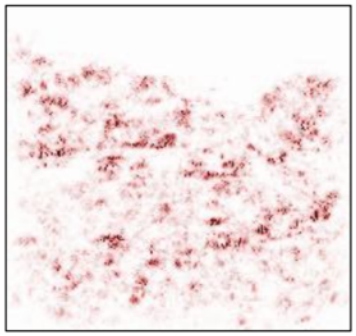
$$\text{heatmap} = LRP(x, f)$$



Naive Approach: Sensitivity Analysis

Sensitivity analysis:

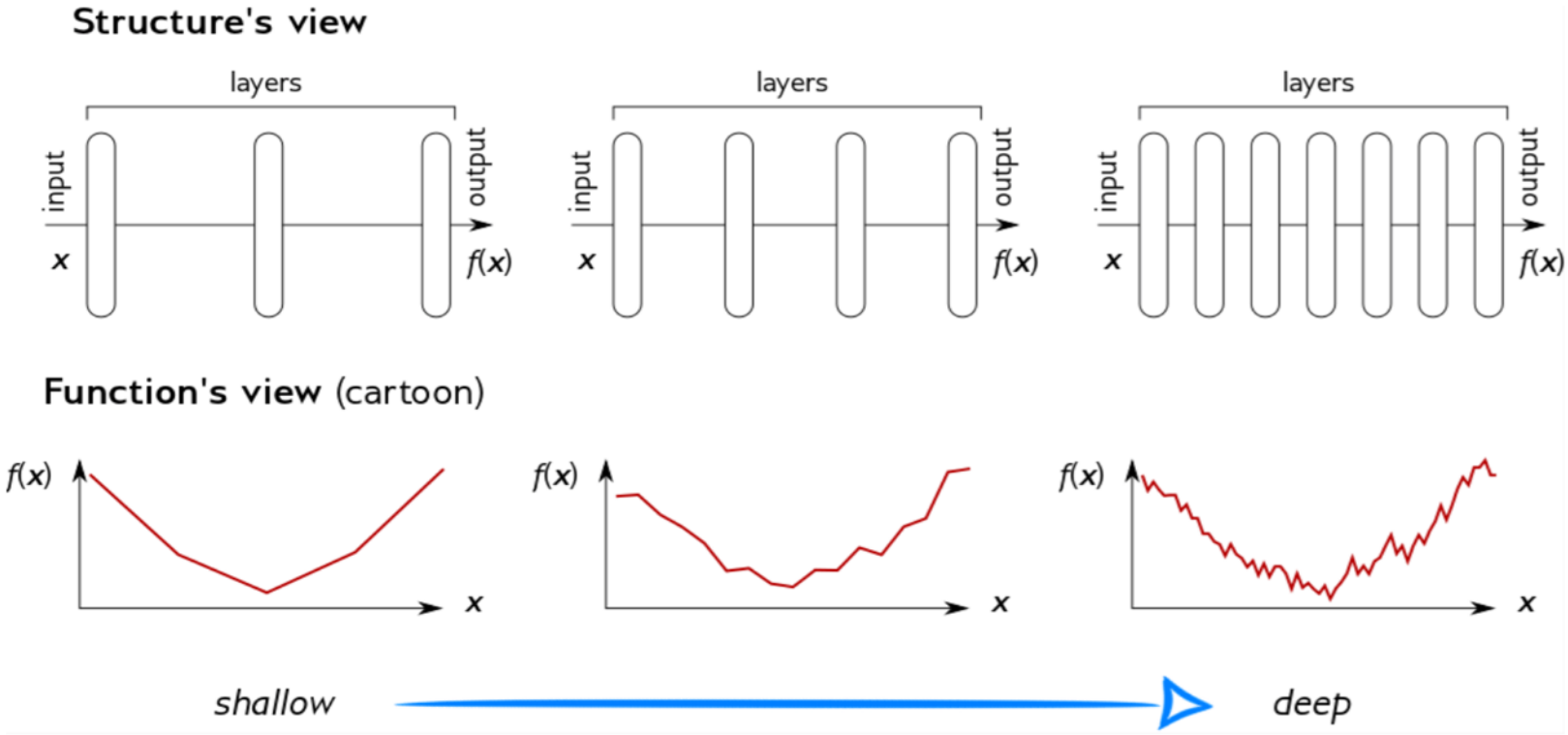
$$R_i = \left(\frac{\partial f}{\partial x_i}\right)^2$$



Problem: sensitivity analysis does not highlight cars

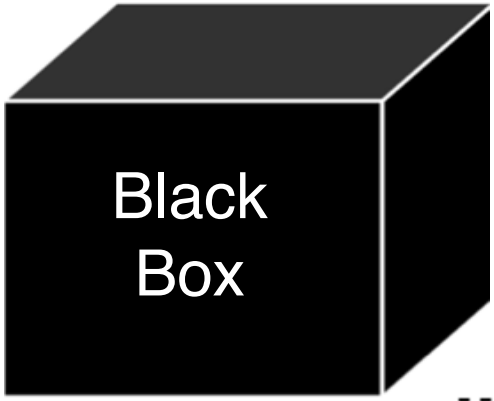
Naive Approach: Sensitivity Analysis

Input gradient (on which sensitivity analysis is based), becomes increasingly highly varying and unreliable with neural network depth.



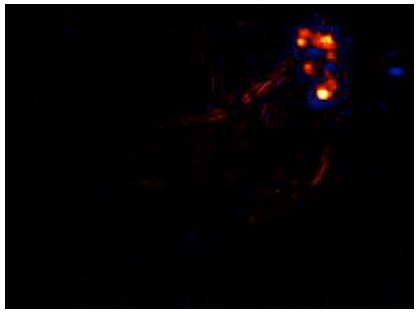
Better Approach: LRP

input x



Rooster

prediction $f(x)$



heatmap

Explain prediction

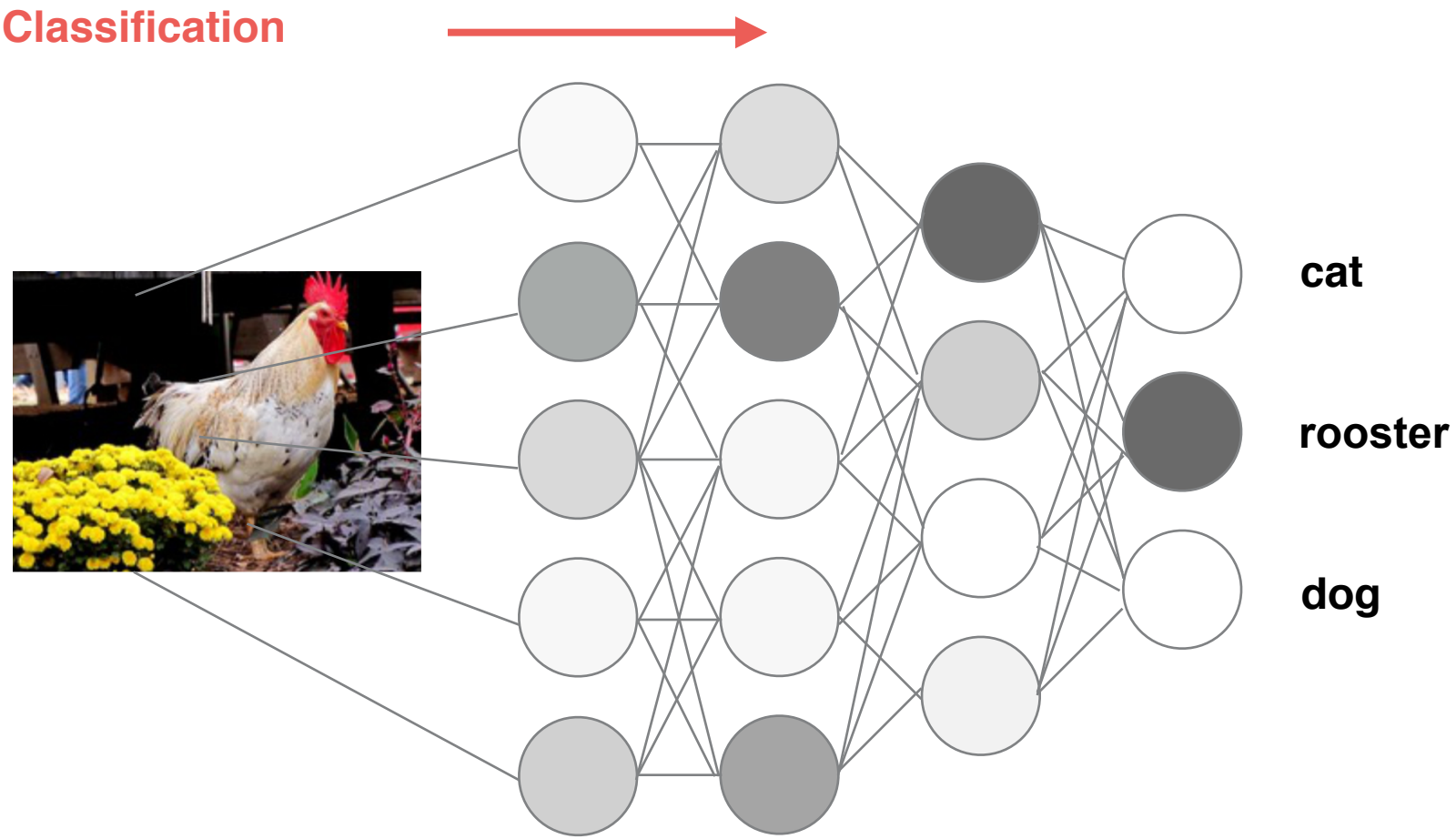
(how much each pixel contributes to prediction)

Idea: Decompose function

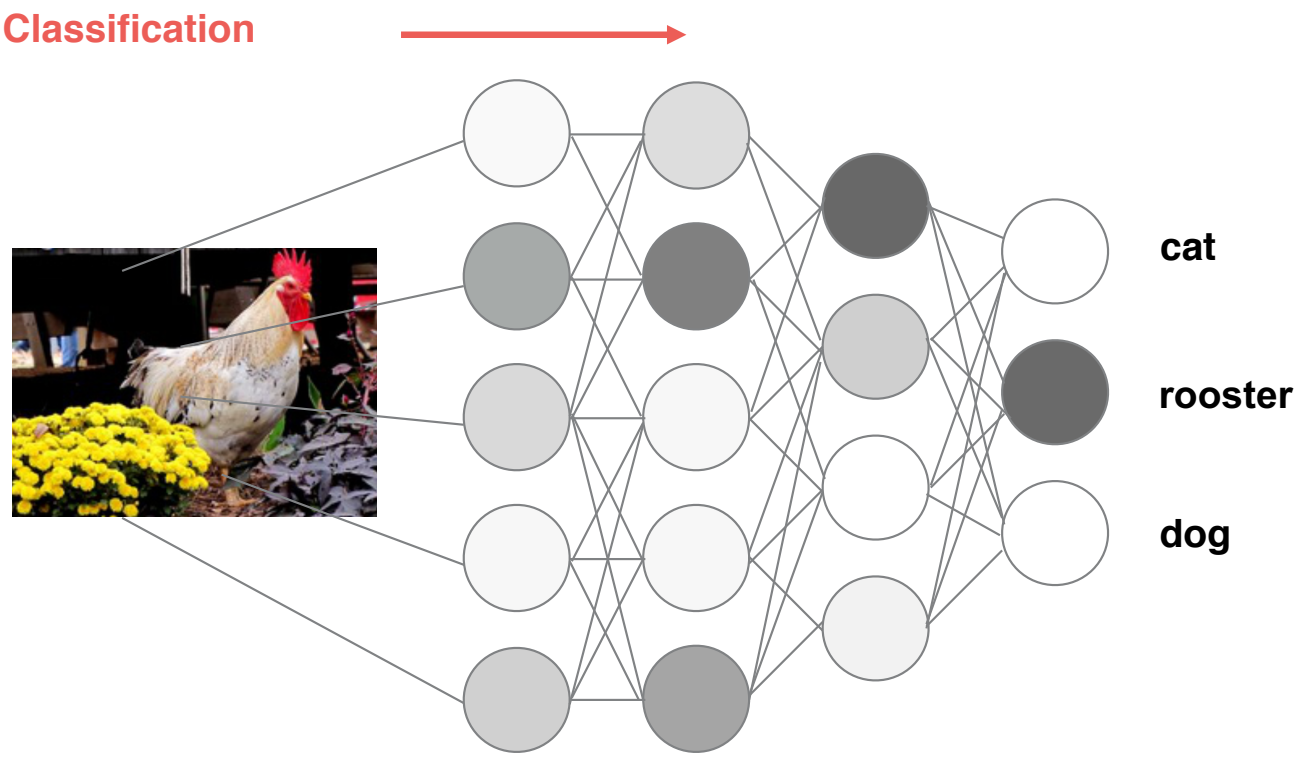
$$\sum_i R_i = f(x)$$

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

Better Approach: LRP



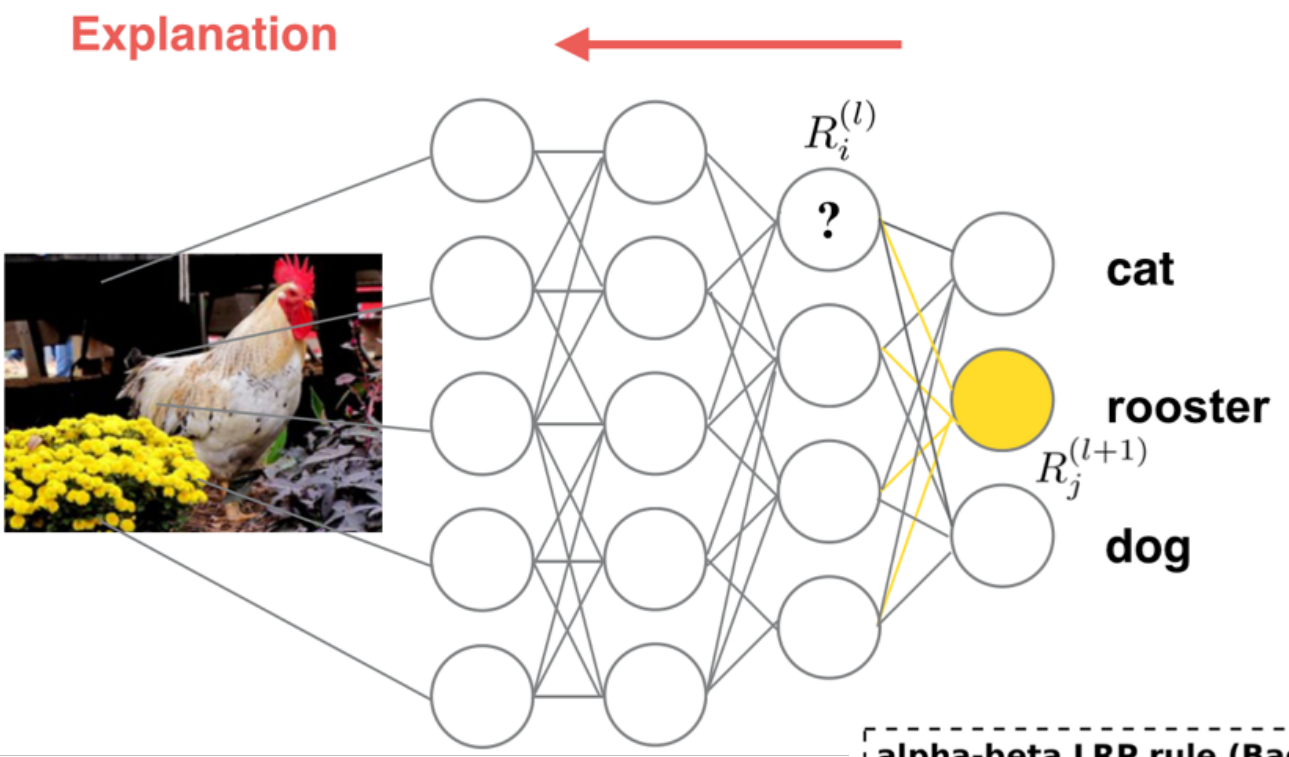
Better Approach: LRP



What makes this image a “rooster image” ?

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

Better Approach: LRP



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

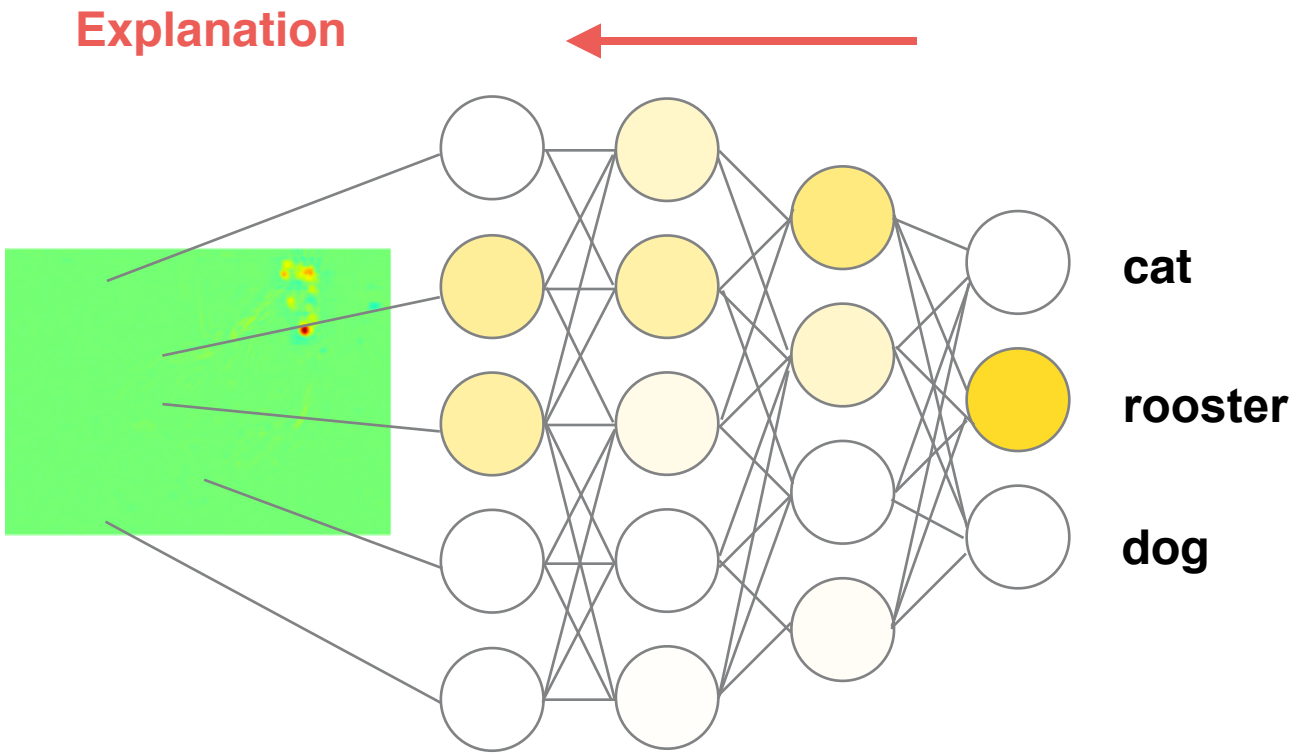
not based on gradient !

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 $\alpha + \beta = 1$

Better Approach: LRP

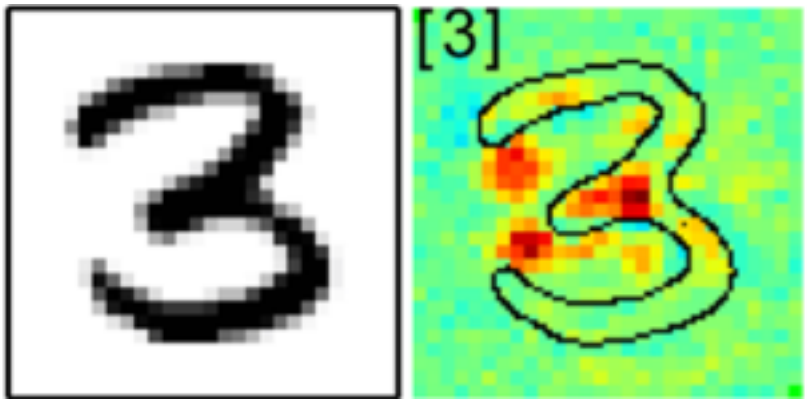


Layer-wise relevance conservation

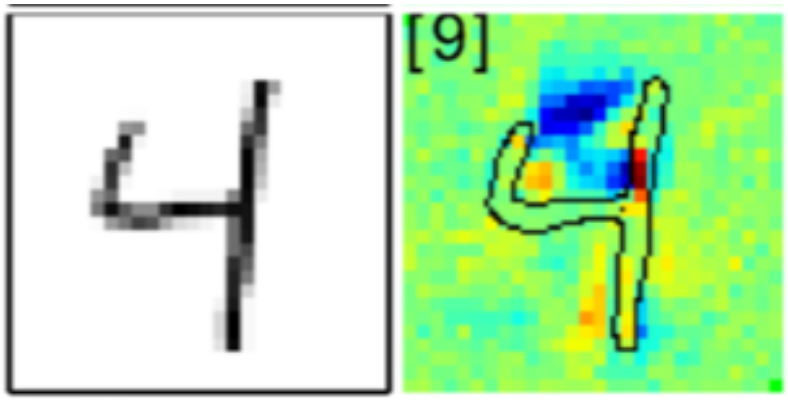
$$\sum_i R_i = \ldots = \sum_i R_i^{(l)} = \sum_j R_j^{(l+1)} = \ldots = f(x)$$

Better Approach: LRP

Heatmap of prediction “3”



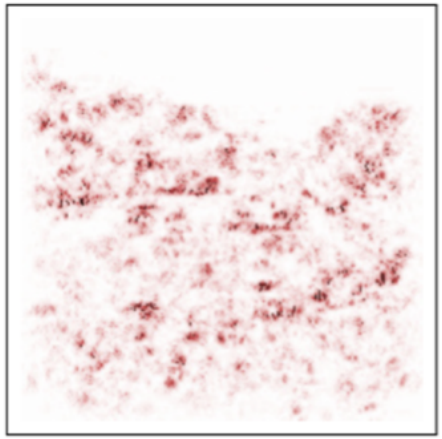
Heatmap of prediction “9”



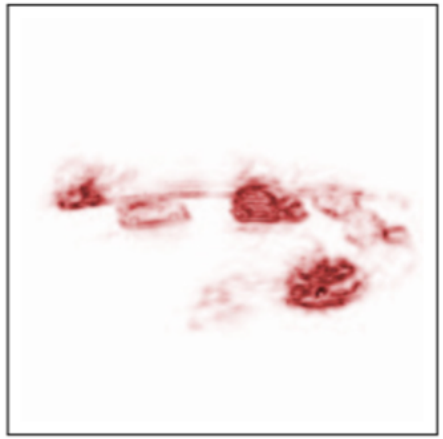
Better Approach: LRP

Image

Sensitivity Analysis



LRP / Deep Taylor



Explains what influences prediction “cars”.

Explains prediction “cars” as is.

Slope decomposition

$$\sum_i R_i = \|\nabla_x f\|^2$$

Value decomposition

$$\sum_i R_i = f(\boldsymbol{x})$$

More information (Montavon et al., 2017 & 2018)

Decomposing the Correct Quantity

slope decomposition

$$\sum_i R_i = \|\nabla_x f\|^2$$

$\rightarrow$

value decomposition

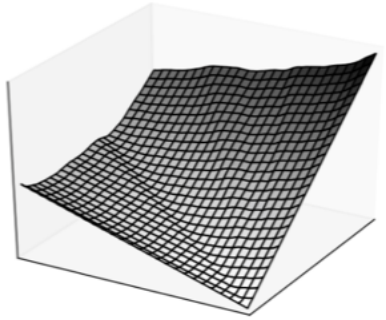
$$\sum_i R_i = f(\boldsymbol{x})$$

Candidate: Taylor decomposition

$$f(\boldsymbol{x}) = \underbrace{f(\widetilde{\boldsymbol{x}})}_{0} + \sum_{i=1}^{d} \underbrace{\frac{\partial f}{\partial x_i}\Big|_{\boldsymbol{x}=\widetilde{\boldsymbol{x}}} (x_i - \widetilde{x}_i)}_{R_i} + \underbrace{O(\boldsymbol{x}\boldsymbol{x}^\top)}_{0}$$

- Achievable for linear models and deep ReLU networks without biases, by choosing:

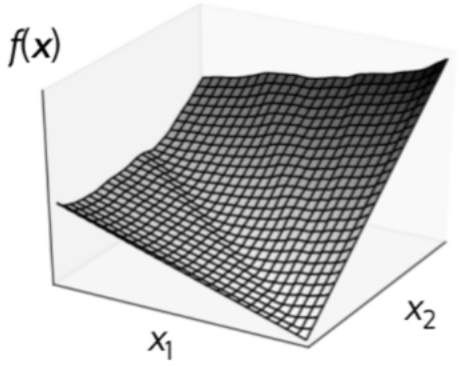
$$\widetilde{\boldsymbol{x}} = \lim_{\varepsilon \to 0} \varepsilon \cdot \boldsymbol{x} \approx \mathbf{0}.$$



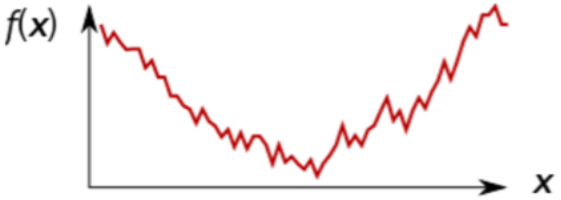
Why Simple Taylor doesn't work?

Two Reasons:

1. Root point is hard to find or too far → includes too much information (incl. negative evidence)

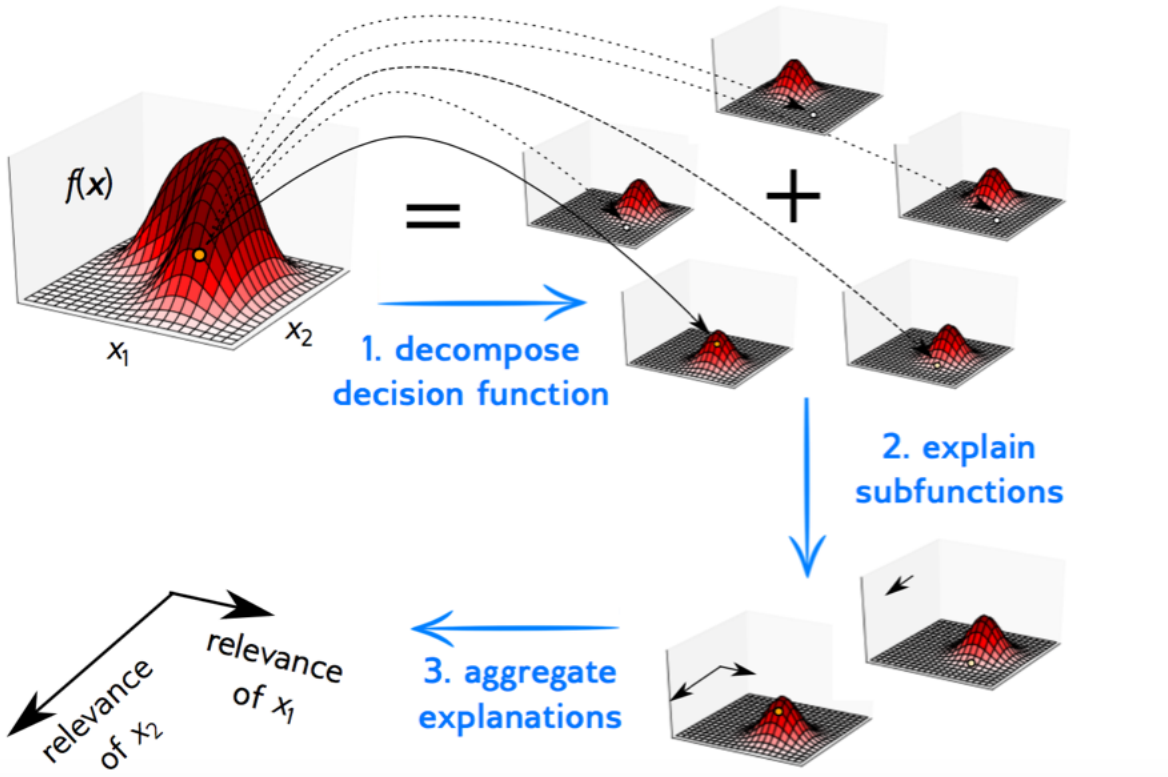


2. Gradient shattering problem → gradient of deep nets has low informative value



Deep Taylor Decomposition

Idea: Since neural network is composed of simple functions, we propose a *deep* Taylor decomposition.



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

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

Deep Taylor Decomposition

Output relevance
$f(x) = R_j = a_j * 1$

$R_j = a_j * \text{const}$

Taylor decomposition

$$R_j((a_i)_i) = R_j((\widetilde{a}_i)_i) + \sum_i \underbrace{\frac{\partial R_j}{\partial a_i}\Big|_{(\widetilde{a}_i)_i} \cdot (a_i - \widetilde{a}_i)}_{R_{i \leftarrow j}} + \varepsilon$$

redistribute to lower layer i

$$R_{i \leftarrow j} = \frac{(a_i - \widetilde{a}_i^{(j)}) w_{ij}}{\sum_i (a_i - \widetilde{a}_i^{(j)}) w_{ij}} R_j$$

compute R_i

$R_i = \sum_j R_{i \leftarrow j}$

$R_i \approx a_i * \text{const}$

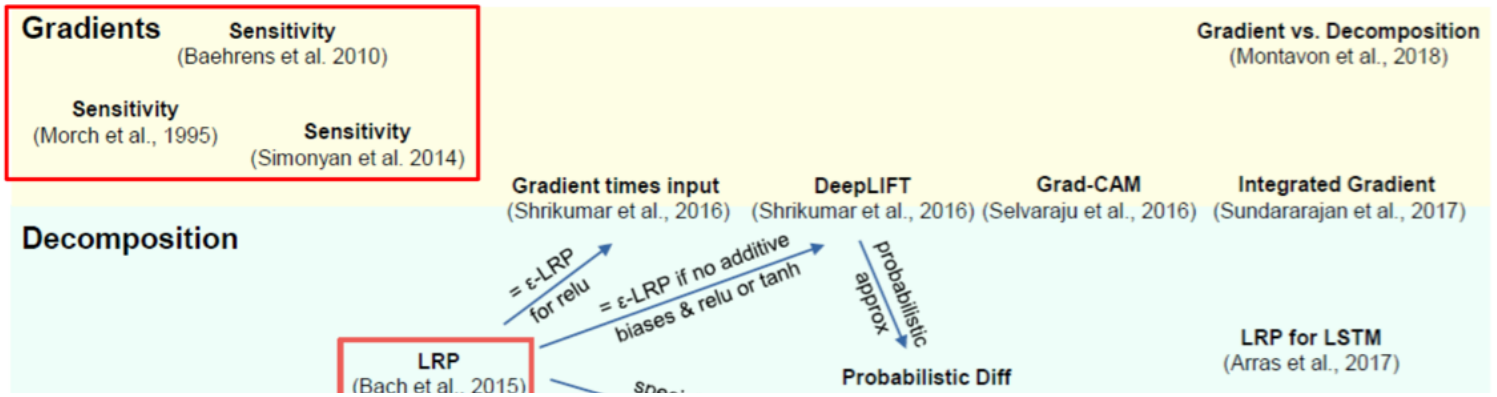
Deep Taylor Decomposition

how to choose the root point $\widetilde{a}_i^{(j)}$?

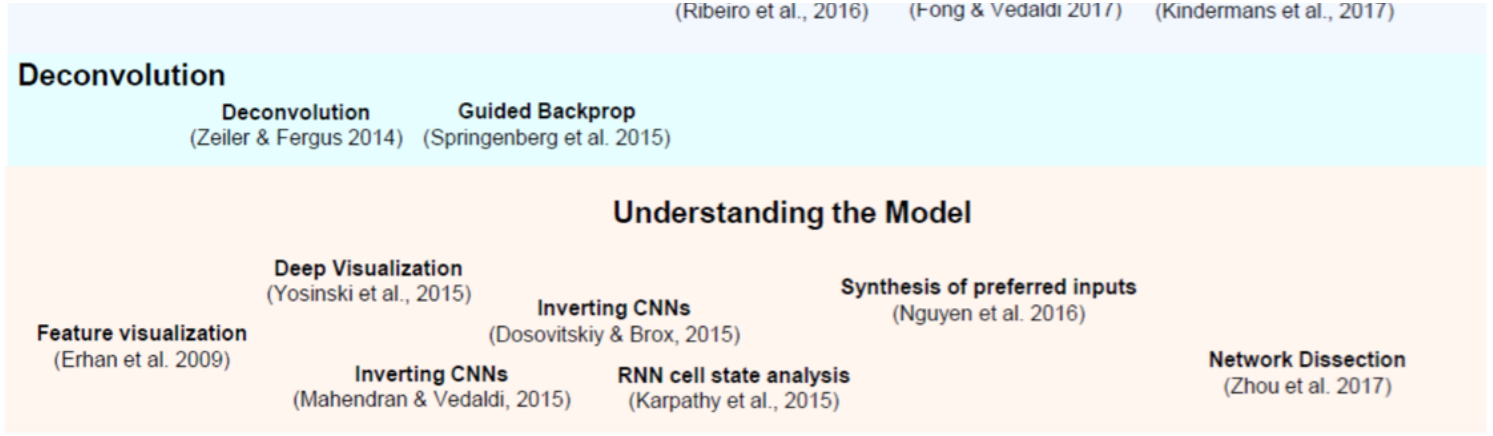
$$R_{i \leftarrow j} = \frac{(a_i - \widetilde{a}_i^{(j)}) w_{ij}}{\sum_i (a_i - \widetilde{a}_i^{(j)}) w_{ij}} R_j$$

Input domain	Rule
ReLU activations $(a_j \geq 0)$	$R_j = \sum_k \frac{a_j w_{jk}^+}{\sum_j a_j w_{jk}^+} R_k$
Pixel intensities $(x_i \in [l_i, h_i], l_i \leq 0 \leq h_i)$	$R_i = \sum_j \frac{x_i w_{ij} - l_i w_{ij}^+ - h_i w_{ij}^-}{\sum_i x_i w_{ij} - l_i w_{ij}^+ - h_i w_{ij}^-} R_j$
Real values $(x_i \in \mathbb{R})$	$R_i = \sum_j \frac{w_{ij}^2}{\sum_i w_{ij}^2} R_j$

Other Explanation Methods

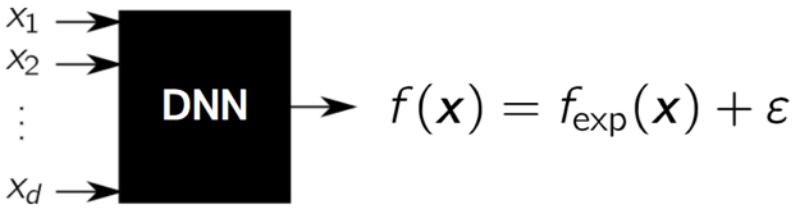


Question: Which one to choose ?



Properties 1-2: Conservation and Positivity

[Montavon'17, see also Sun'11, Landecker'13, Bach'15]



explanation

$R_1, \ldots, R_d$

Conservation: Total attribution on the input features should be proportional to the amount of (explainable) evidence at the output.

$$\sum_{p=1}^{d} R_p = f_{\text{exp}}(\boldsymbol{x})$$

Positivity: If the neural network is certain about its prediction, input features are either relevant (positive) or irrelevant (zero).

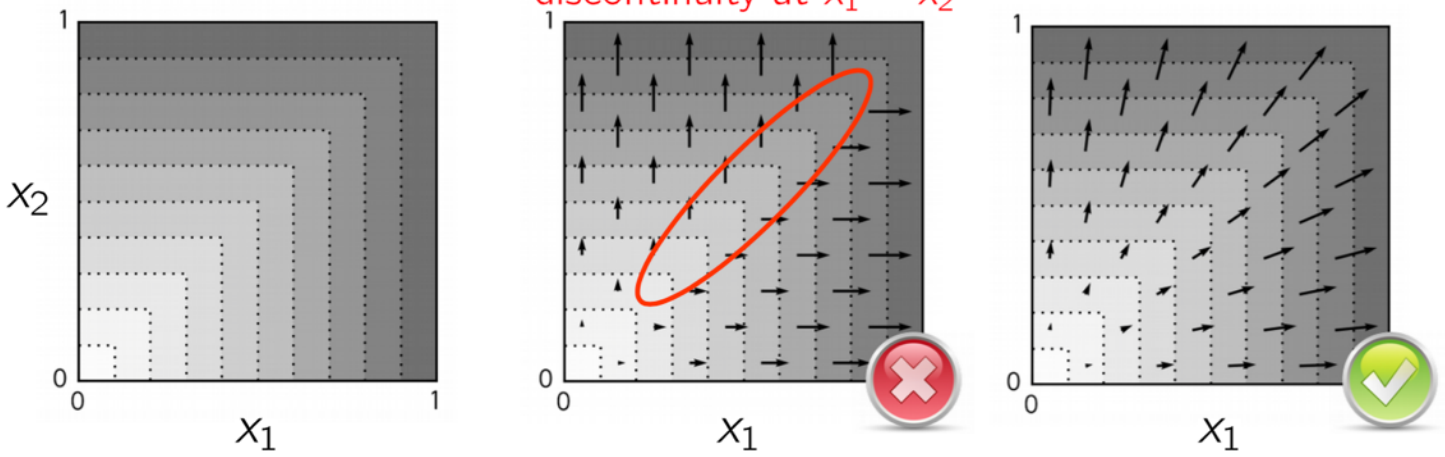
$$\forall_{p=1}^{d} : \; R_p \geq 0$$

Property 3: Continuity [Montavon'18]

If two inputs are the almost the same, and the prediction is also almost the same, then the explanation should also be almost the same.

Example:

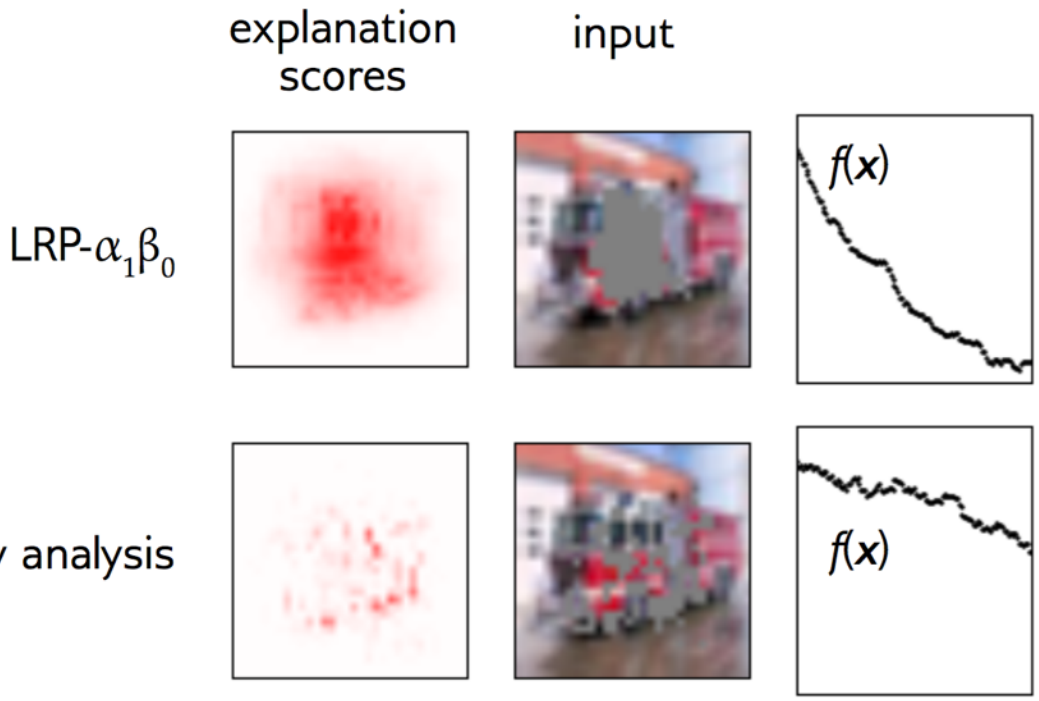
$f(\boldsymbol{x}) = \max(x_1, x_2)$



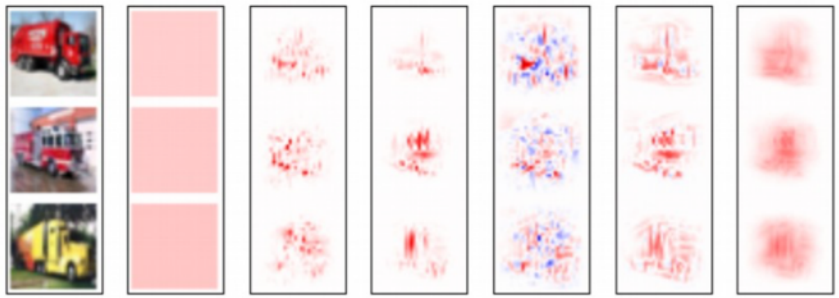
Axiomatic Approach to Interpretability

Property 4: Selectivity [Bach'15, Samek'17]

Model must agree with the explanation: If input features are attributed relevance, removing them should reduce evidence at the output.

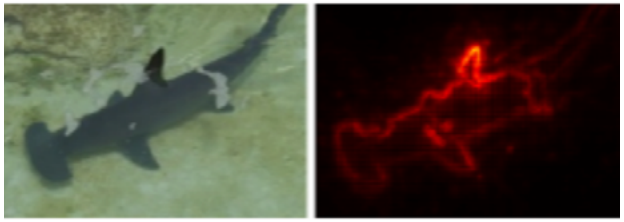


Axiomatic Approach to Interpretability

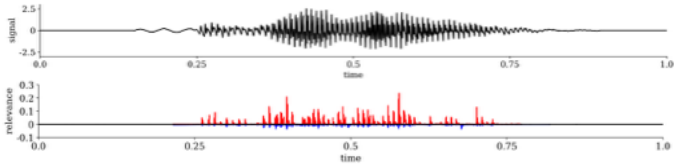
Explanation techniques / Properties		Uniform	$(\text{Gradient})^2$	$(\text{Guided BP})^2$	Gradient x Input	Guided BP x Input	LRP-$\alpha_1\beta_0$	...
1. Conservation		✔			✔	✔	✔	
2. Positivity		✔	✔	✔		✔	✔	
3. Continuity		✔		✔		✔	✔	
4. Selectivity			✔	✔	✔	✔	✔	
...								

LRP revisited

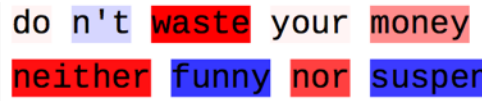
General Images (Bach' 15, Lapuschkin'16)



Speech (Becker'18)

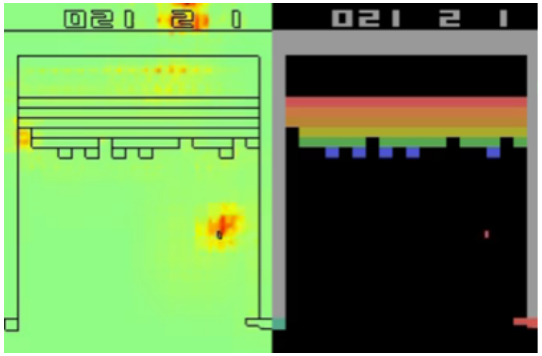


Text Analysis (Arras'16 &17)

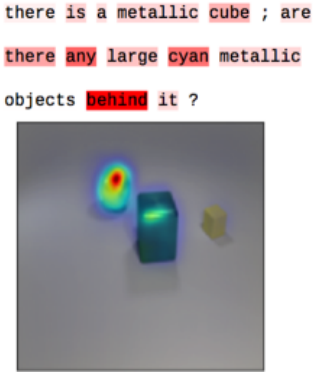


Morphing (Seibold'18)

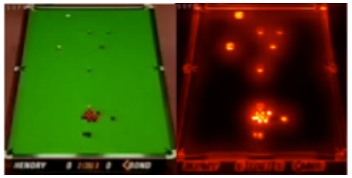
Games (Lapuschkin'19)



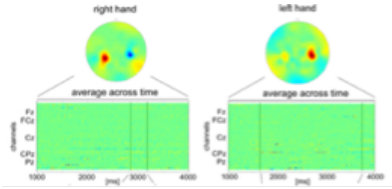
VQA (Arras'18)



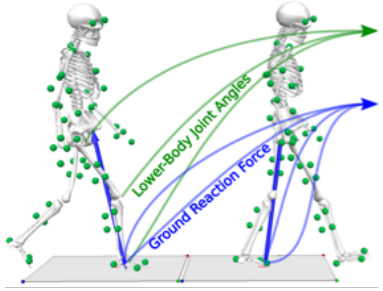
Video (Anders'18)



EEG (Sturm'16)

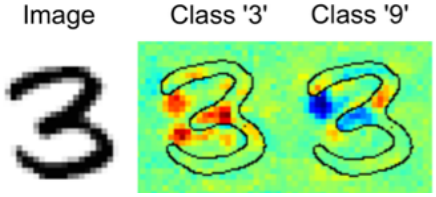


Gait Patterns (Horst'19)

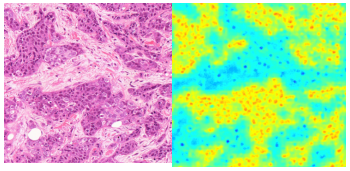


Faces (Lapuschkin'17)

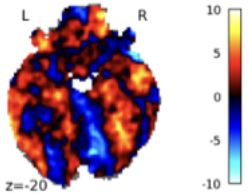
Digits (Bach' 15)



Histopathology (Binder'18)

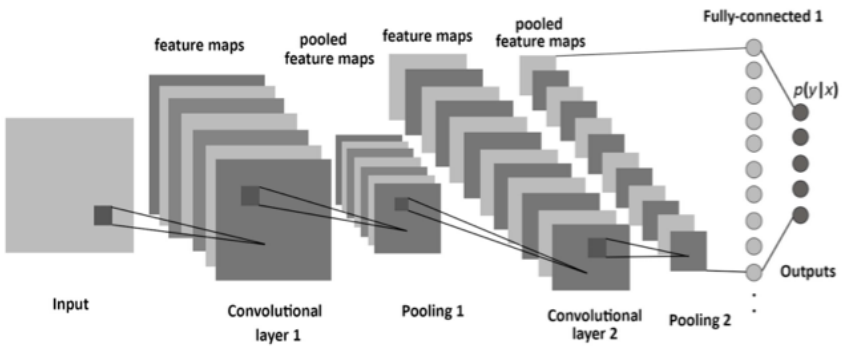


fMRI (Thomas'18)

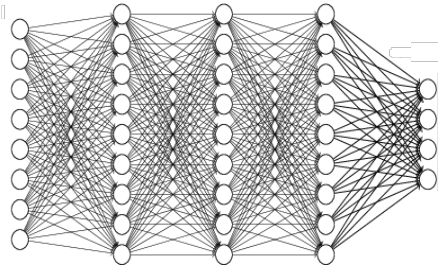


LRP applied to different Models

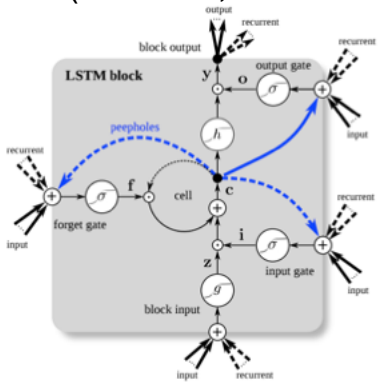
Convolutional NNs (Bach'15, Arras'17 …)



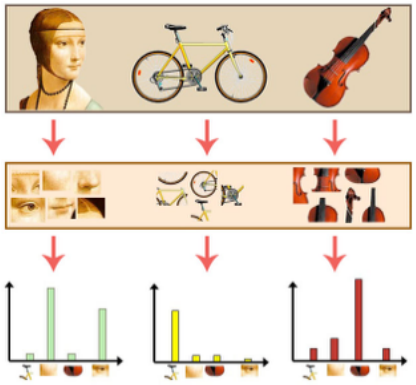
Local Renormalization Layers (Binder'16)



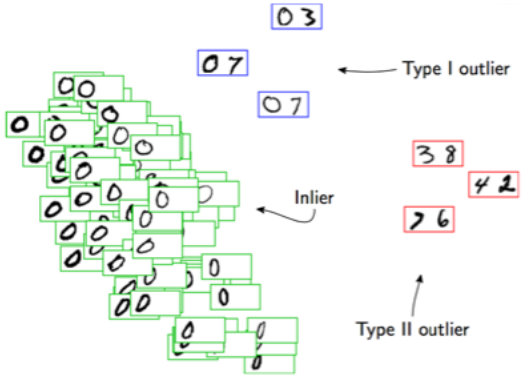
LSTM (Arras'17, Thomas'18)



Bag-of-words / Fisher Vector models (Bach'15, Arras'16, Lapuschkin'17, Binder'18)



One-class SVM (Kauffmann'18)

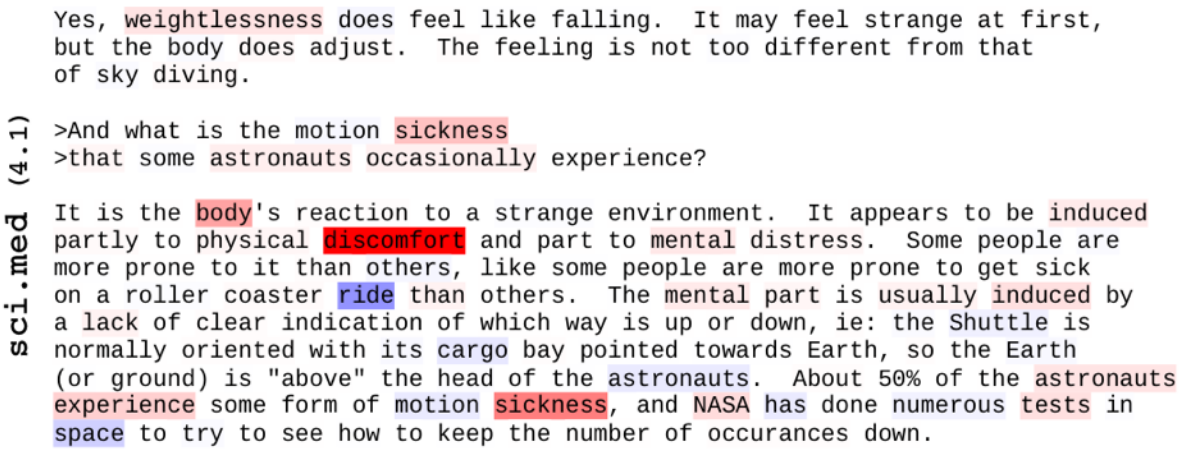


Application: Compare Classifiers

word2vec/CNN:

Performance: 80.19%

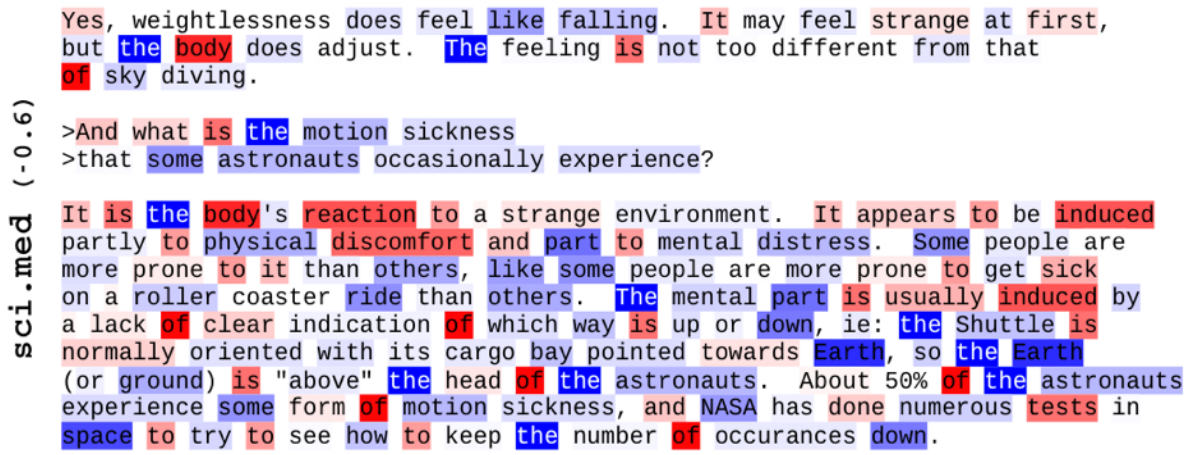
Strategy to solve the problem: identify semantically meaningful words related to the topic.



BoW/SVM:

Performance: 80.10%

Strategy to solve the problem: identify statistical patterns, i.e., use word statistics



(Arras et al. 2016 & 2017)

Application: Compare Classifiers

Test error for various classes:

	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%

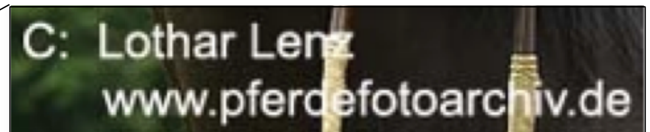
Image

same performance —> same strategy ?

(Lapuschkin et al. 2016)

Application: Compare Classifiers

'horse' images in PASCAL VOC 2007



Application: Measure Context Use

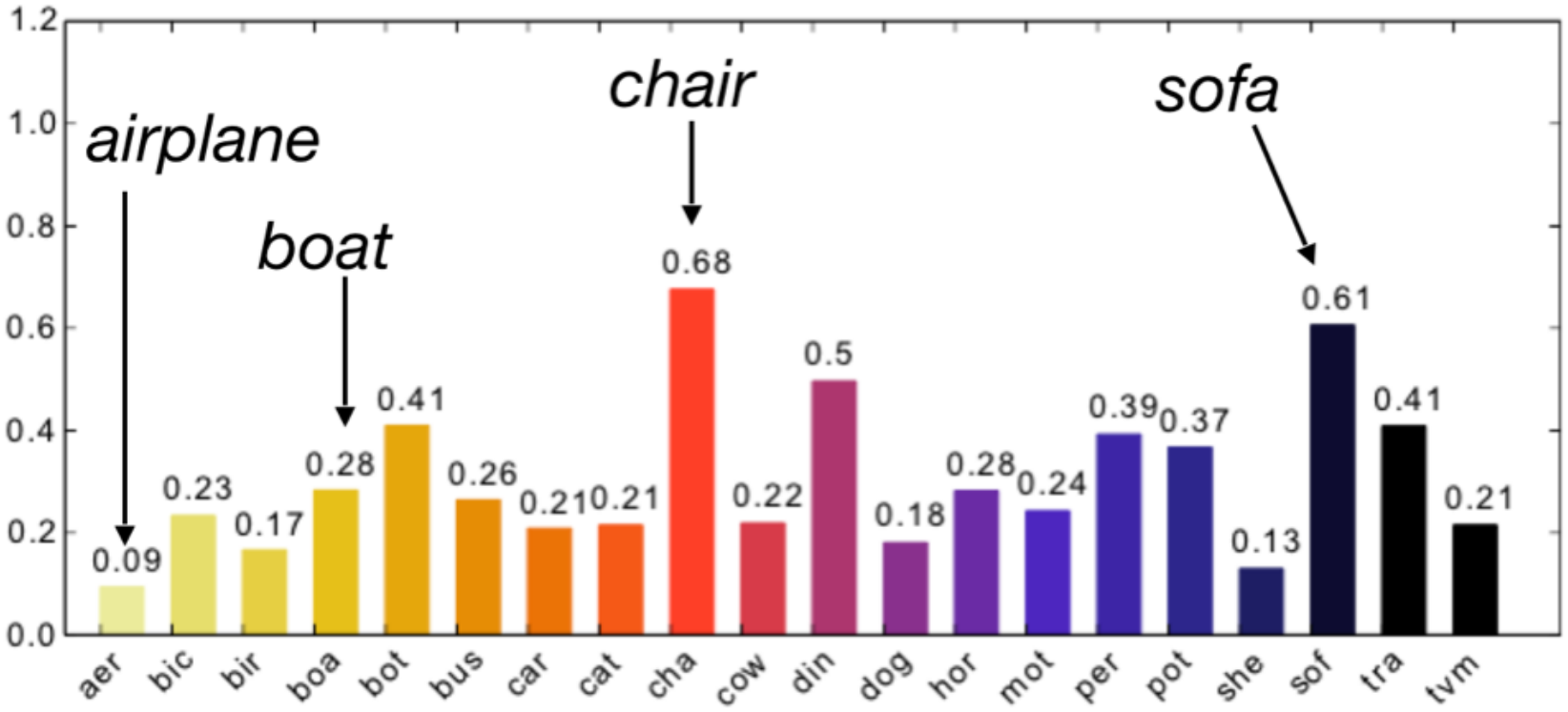
how important
is context ?

how important
is context ?

classifier

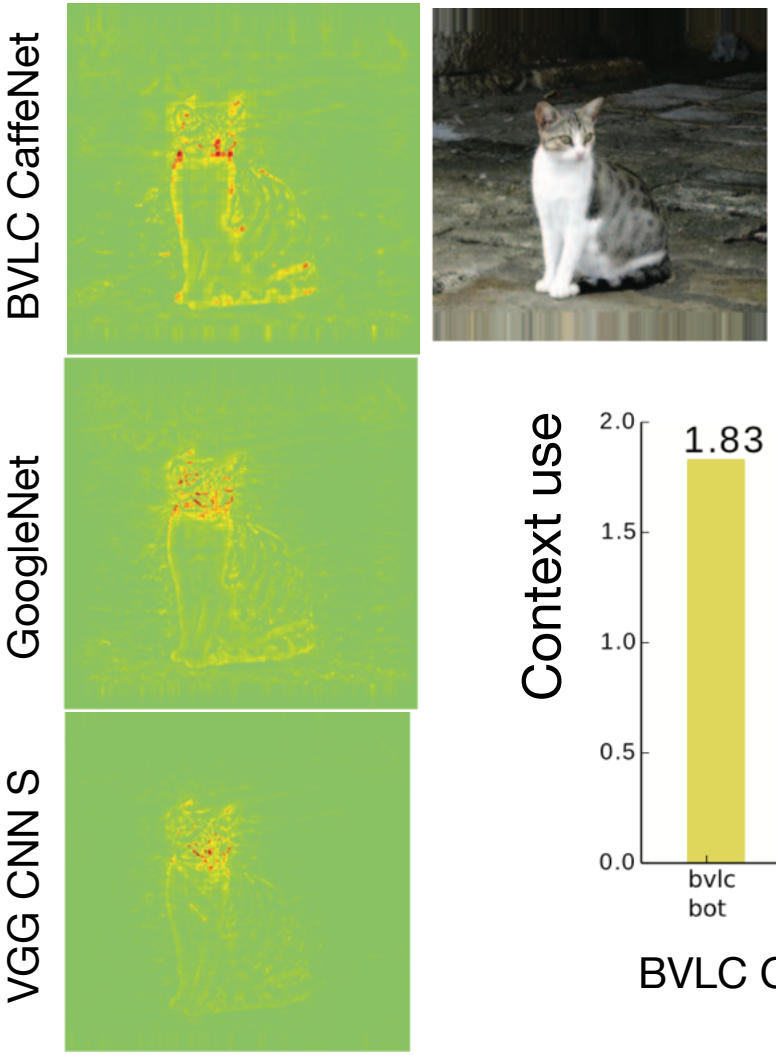
$$\text{importance of context} = \frac{\text{relevance outside bbox}}{\text{relevance inside bbox}}$$

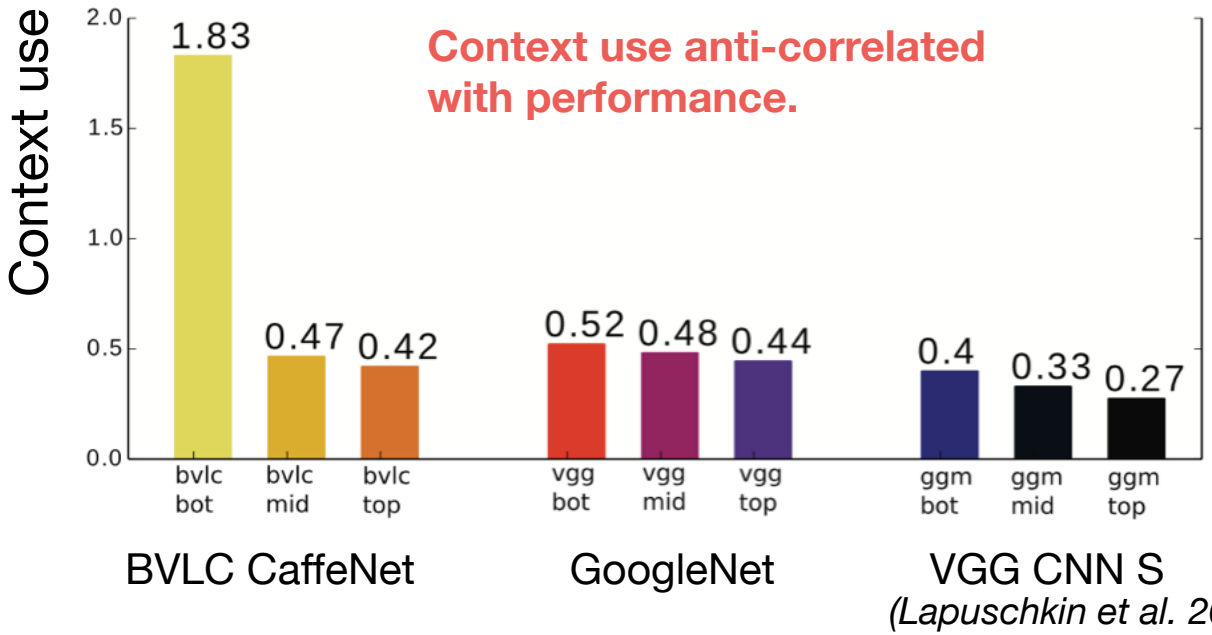
Application: Measure Context Use



(Lapuschkin et al., 2016)

Application: Measure Context Use





Context use anti-correlated with performance.

(Lapuschkin et al. 2016)

Application: Face analysis

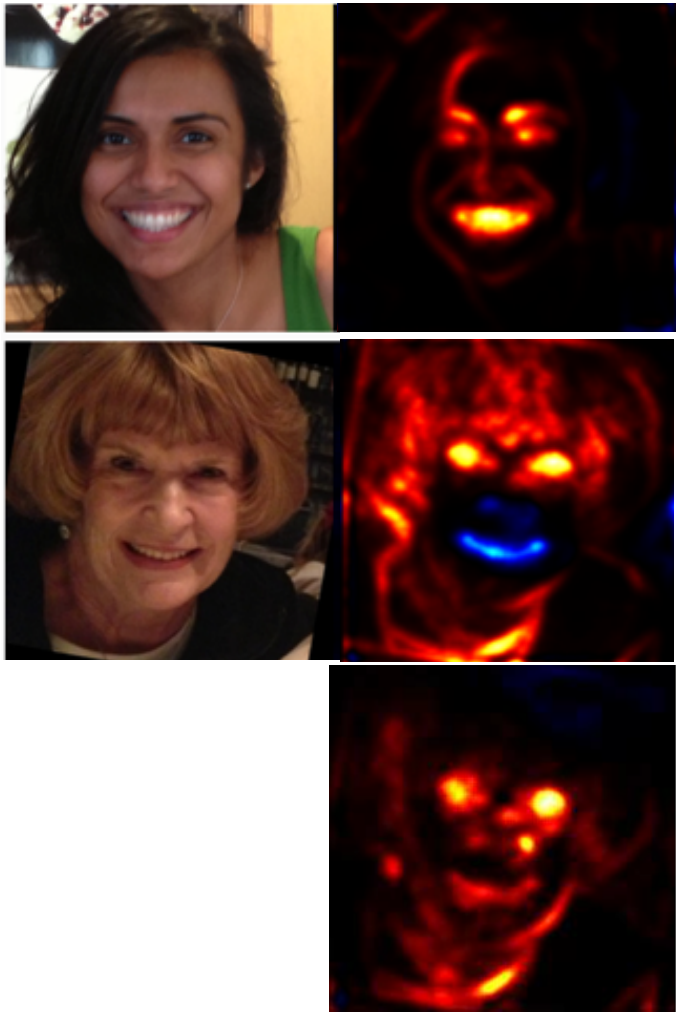
Gender classification

Strategy to solve the problem: Focus on chin / beard, eyes & hear, but without pretraining the model overfits

(Lapuschkin et al., 2017)

Application: Face analysis

Age classification



Predictions

25-32 years old

Strategy to solve the problem:
Focus on the laughing …

60+ years old

pretraining on ImageNet

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

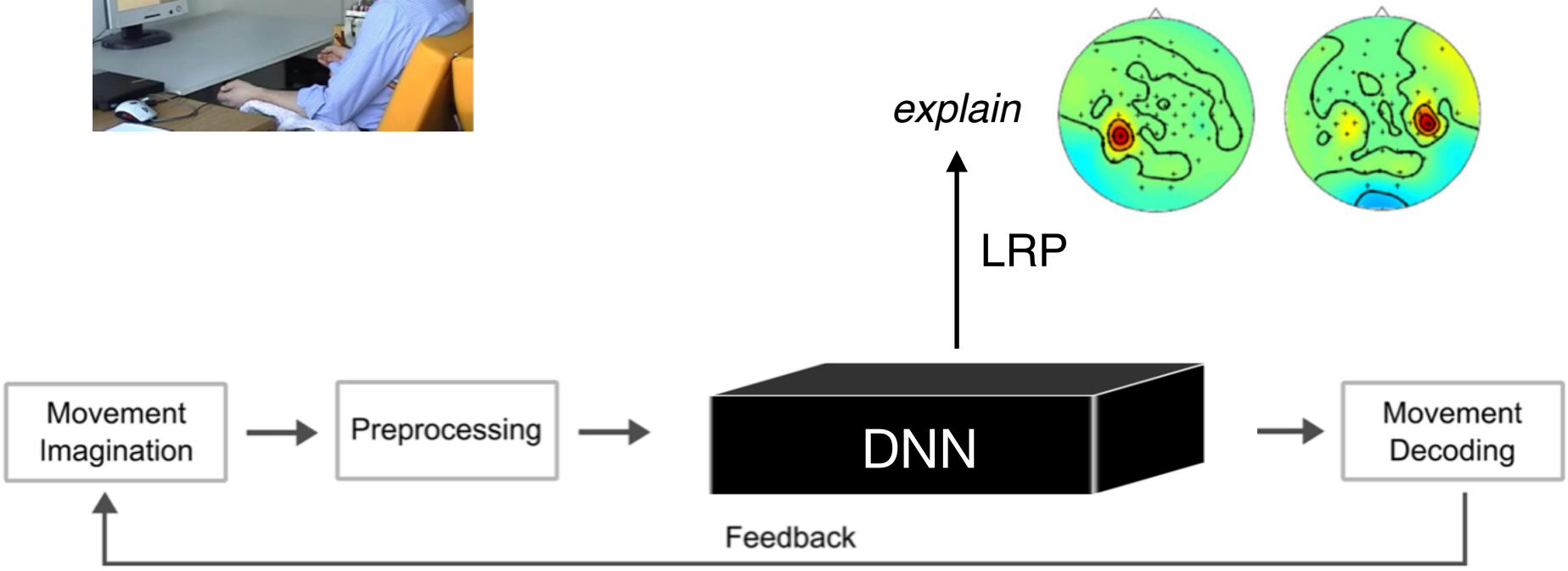
pretraining on IMDB-WIKI

(Lapuschkin et al., 2017)

Application: EEG Analysis

Brain-Computer Interfacing

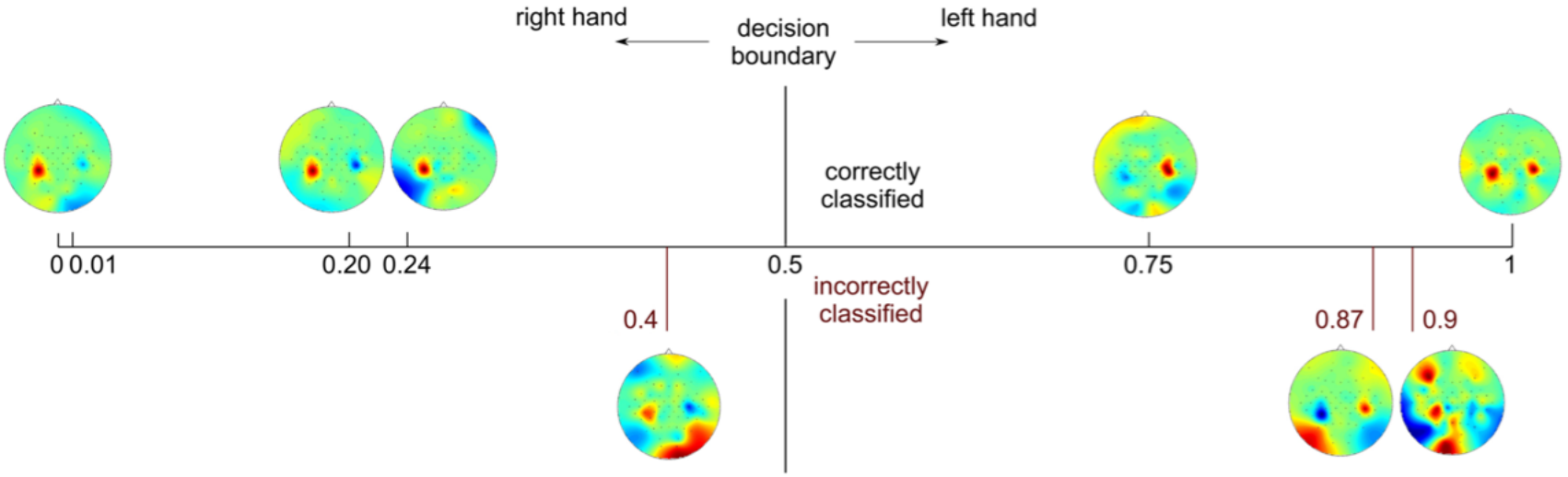
How brain works subject-dependent —> individual explanations



(Sturm et al. 2016)

Application: EEG Analysis

With LRP we can analyze what made a trial being misclassified.



(Sturm et al. 2016)

Application: Sentiment analysis

How to handle multiplicative interactions ?

$$z_j = z_g \cdot z_s \qquad R_g = 0 \quad R_s = R_j$$

gate neuron indirectly affect relevance distribution in forward pass

Negative sentiment

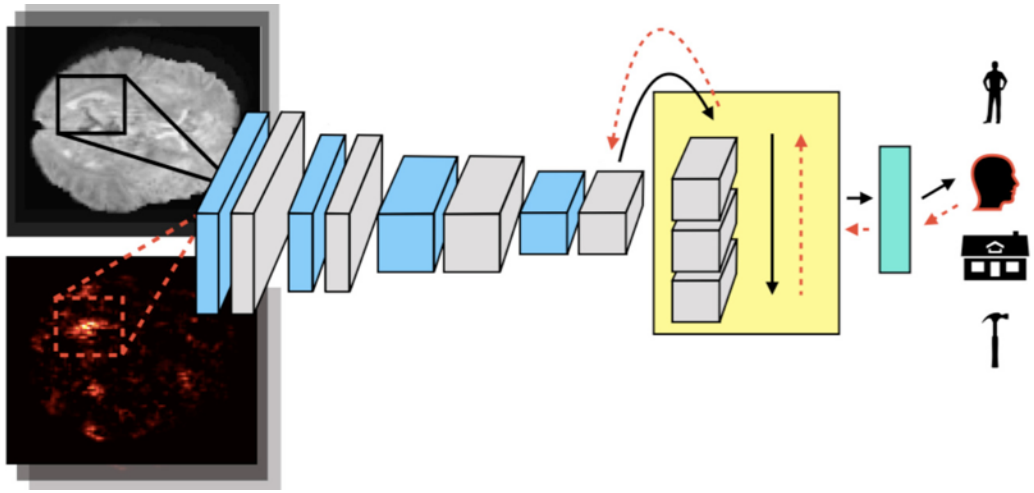
--	--	1.	do n't waste your money .
		2.	neither funny nor suspenseful nor particularly well-drawn .
		3.	it 's not horrible , just horribly mediocre .
		4.	... too slow , too boring , and occasionally annoying .
		5.	it 's neither as romantic nor as thrilling as it should be .

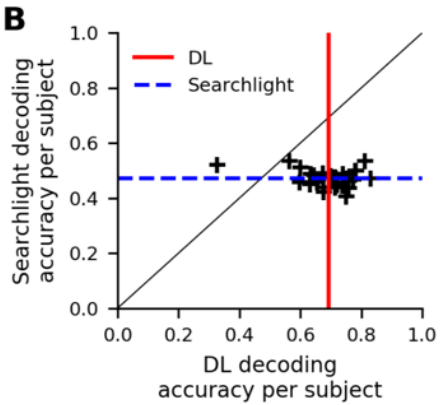
Positive sentiment

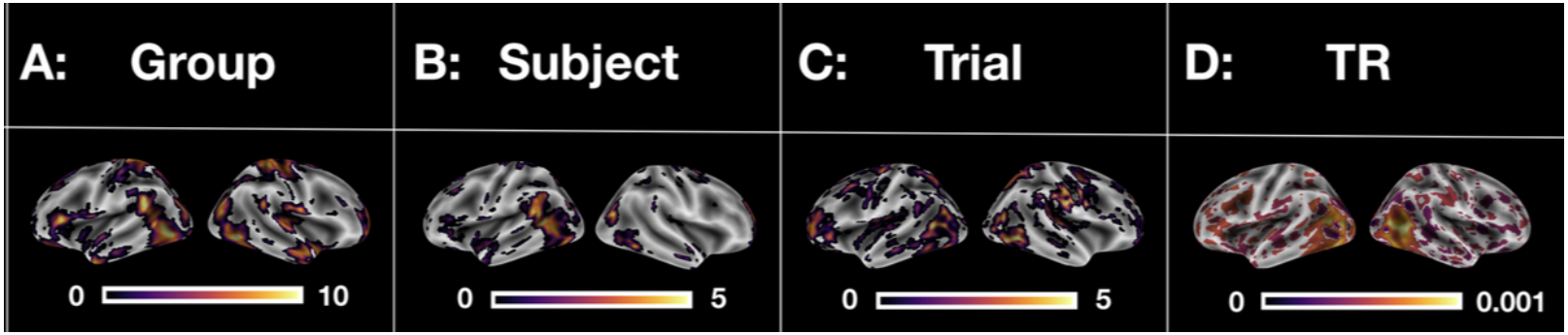
++	++	19.	a worthy entry into a very difficult genre .
		20.	it 's a good film -- not a classic , but odd , entertaining and authentic .
	--	21.	it never fails to engage us .

(Arras et al., 2017)

Application: fMRI Analysis

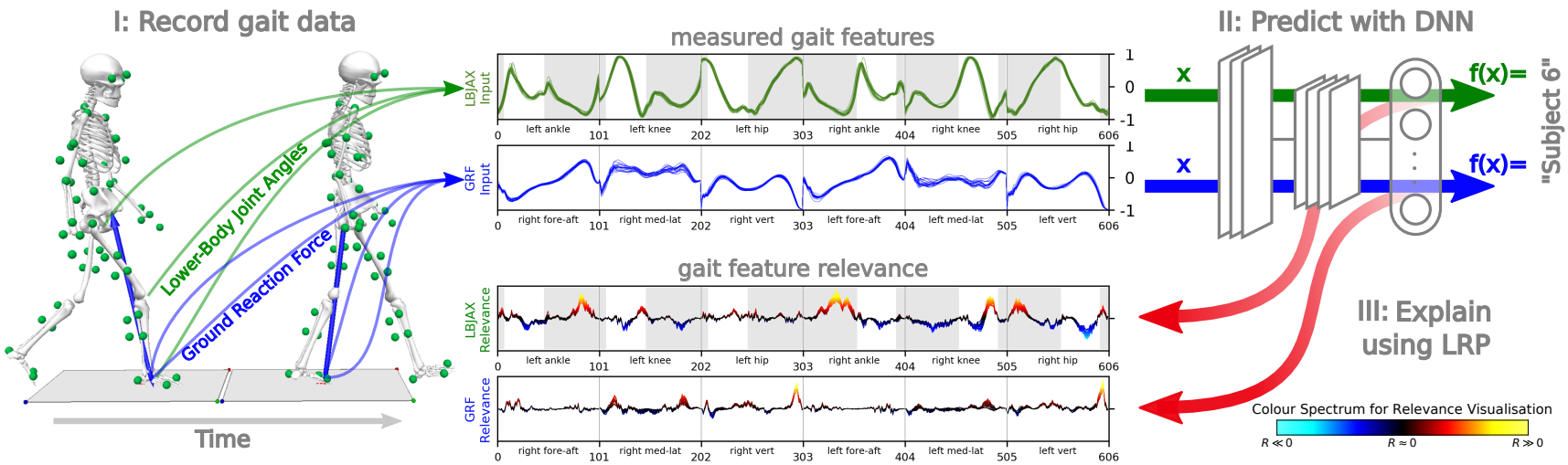






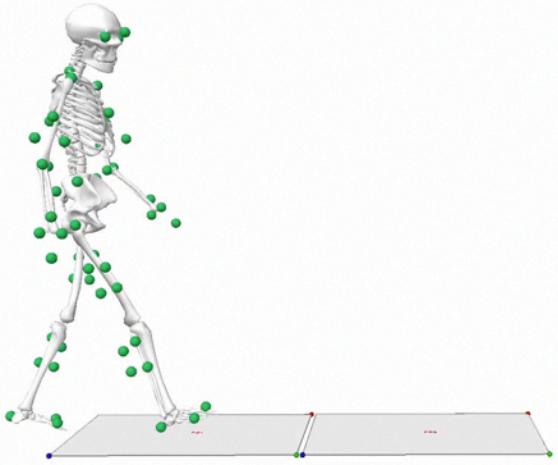
(Thomas et al. 2018)

Application: Gait Analysis



Our approach:

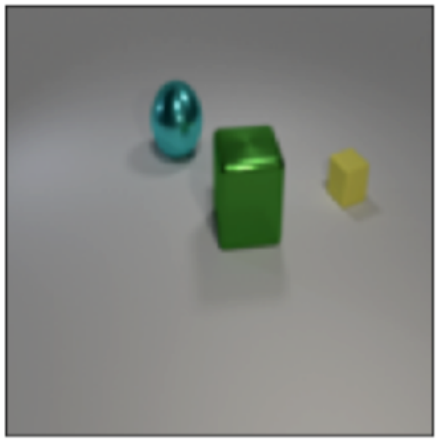
- Classify & explain individual gait patterns
- Important for understanding diseases such as Parkinson

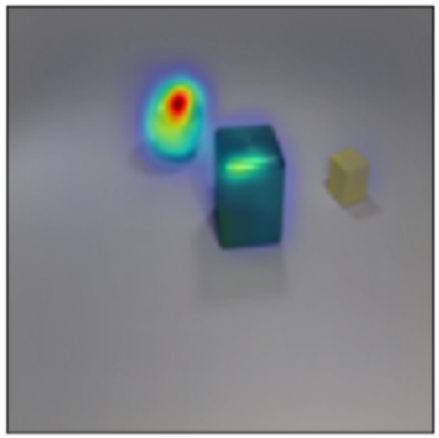


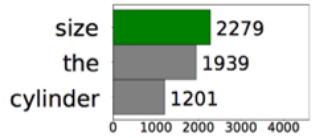
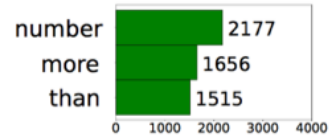
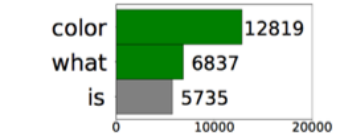
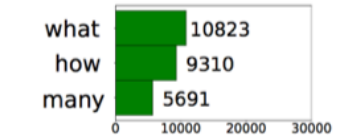
(Horst et al. 2018)

Application: Understand the model

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



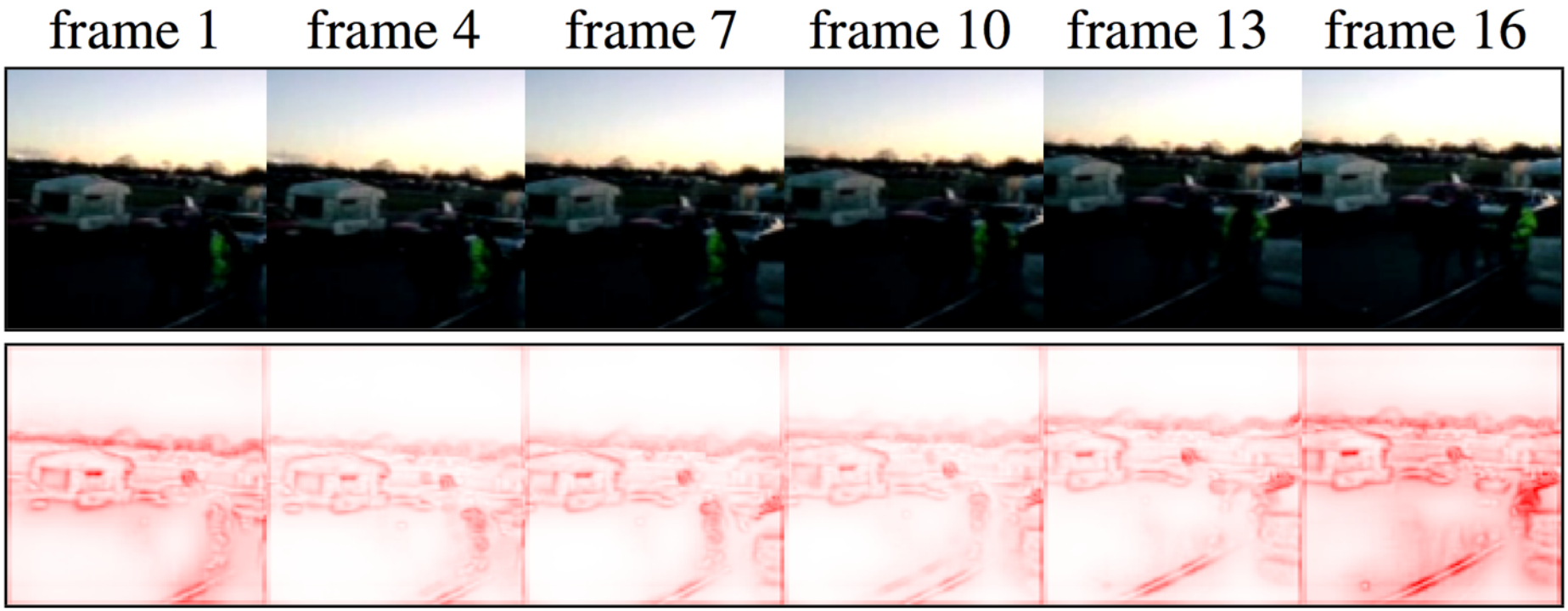


Question Type	LRP
equal_size	
greater_than	
query_color	
count	

model understands the question and correctly identifies the object of interest

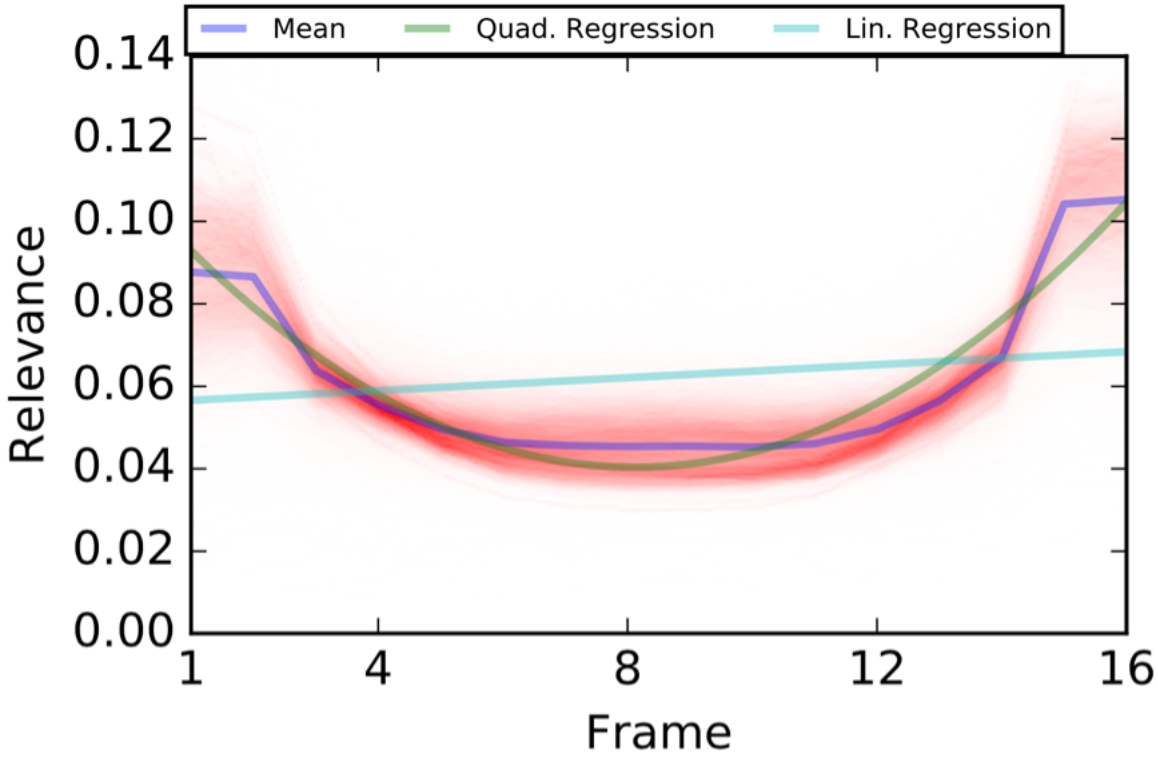
(Arras et al., 2018)

Application: Understand the model



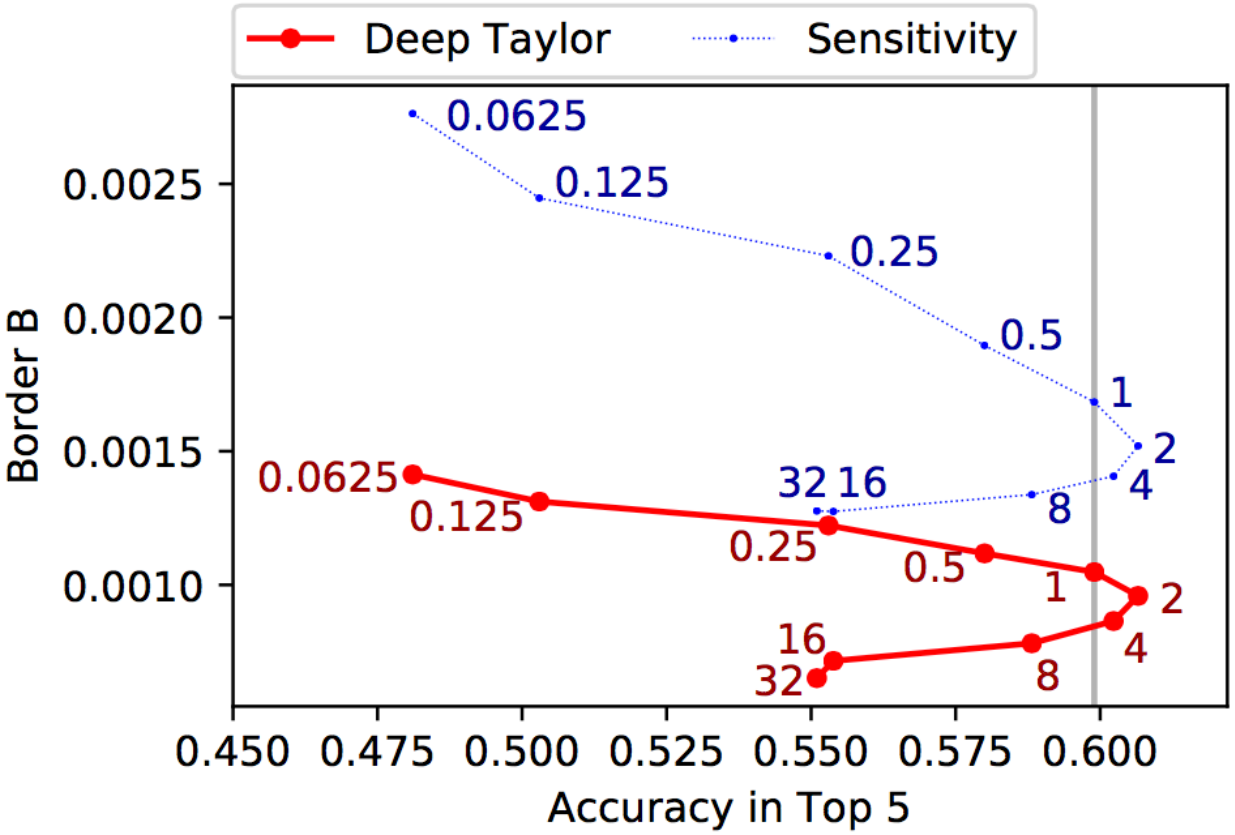
(Anders et al., 2018)

Application: Understand the model



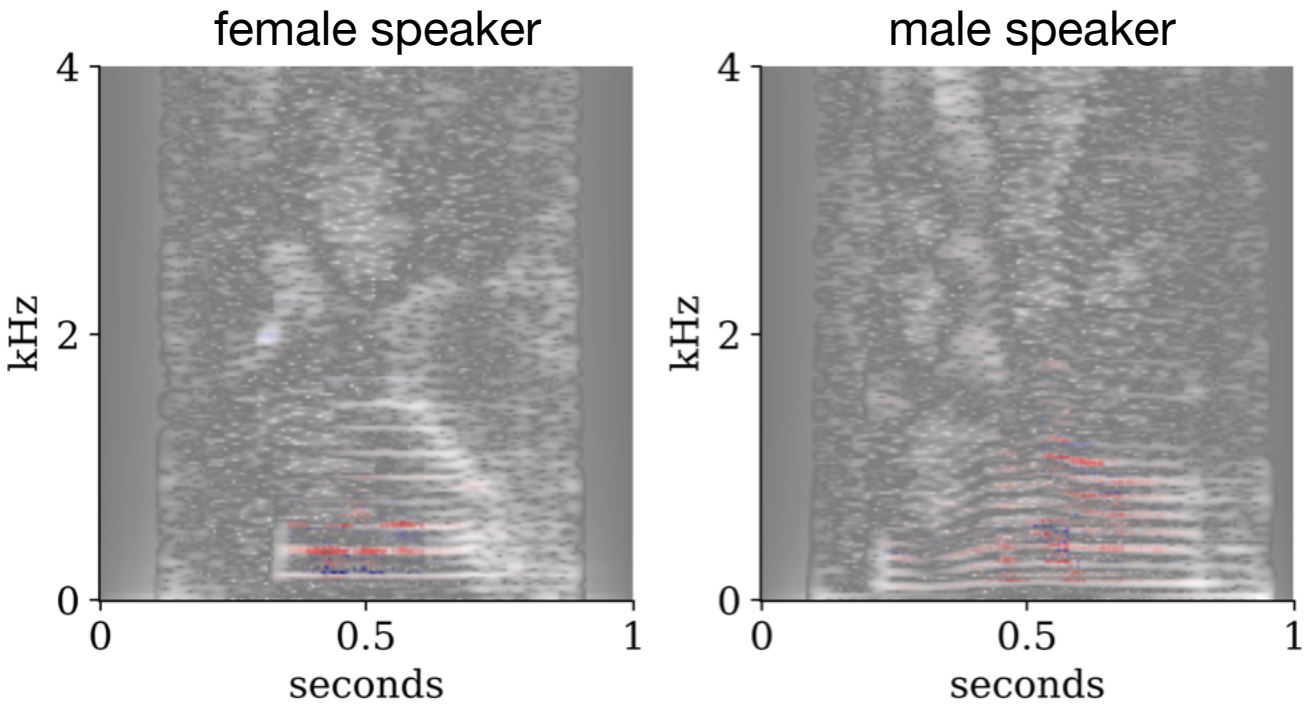
Observation: Explanations focus on the bordering of the video, as if it wants to watch more of it.

Application: Understand the model



Idea: Play video in fast forward (without retraining) and then the classification accuracy improves.

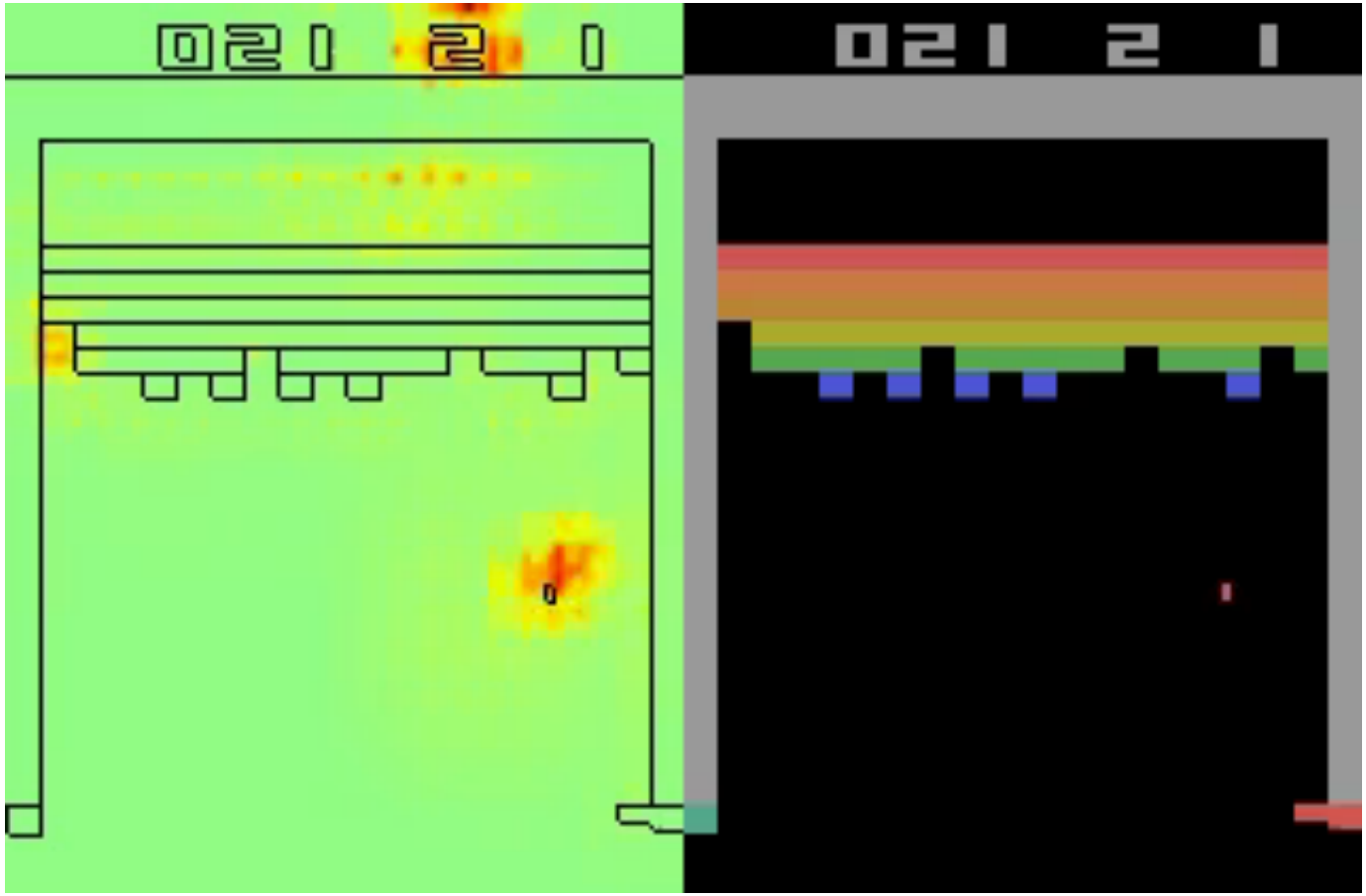
Application: Understand the model



model classifies gender based on the fundamental frequency and its immediate harmonics (see also Traunmüller & Eriksson 1995)

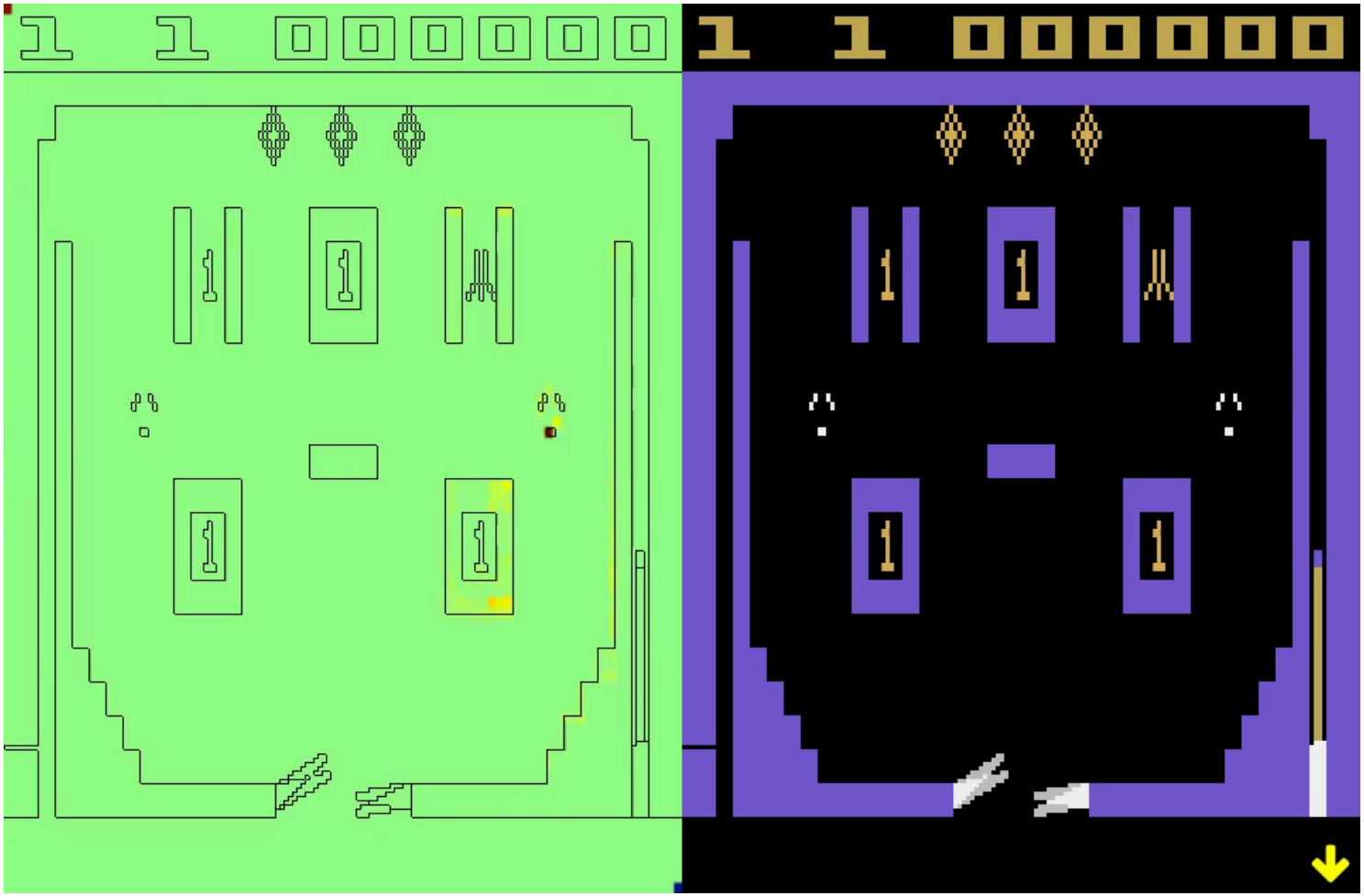
(Becker et al., 2018)

Application: Understand the model



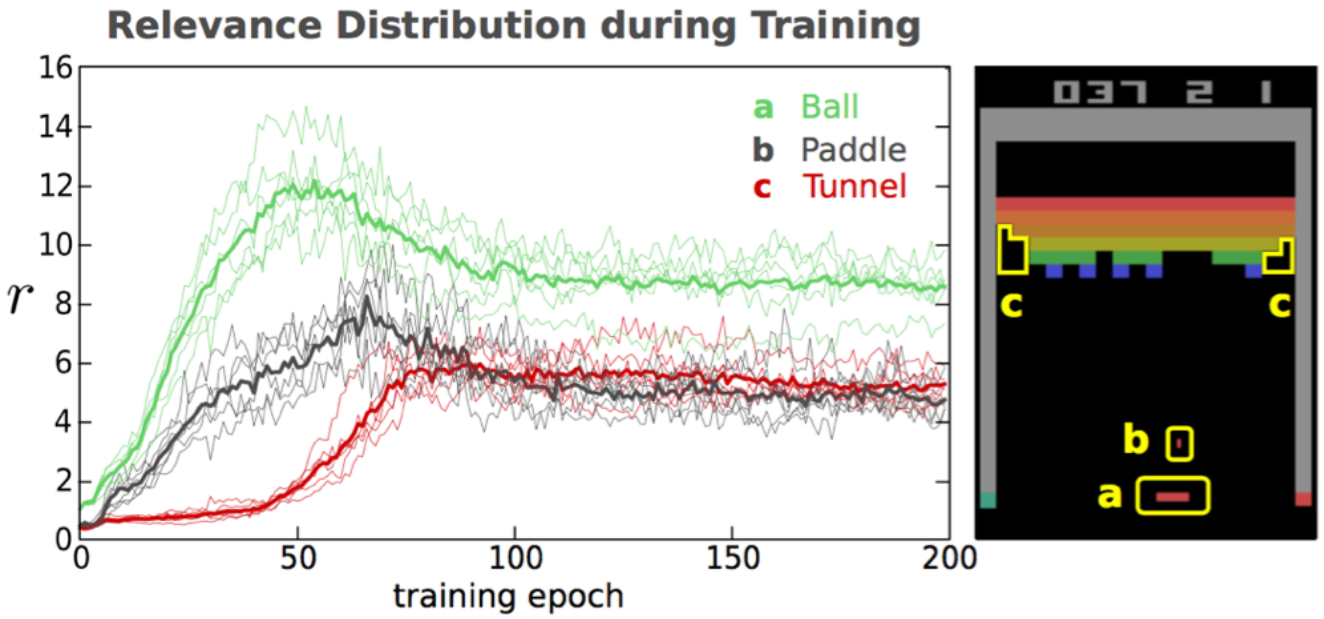
(Lapuschkin et al., in prep.)

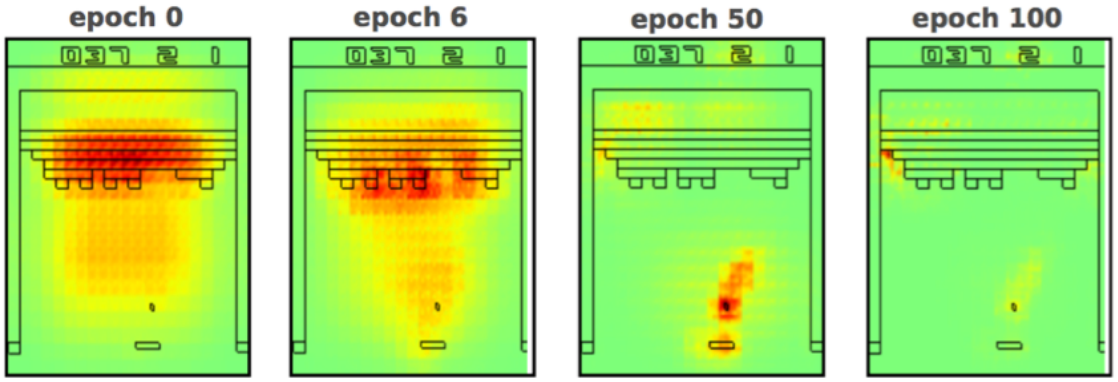
Application: Understand the model



(Lapuschkin et al., in prep.)

Application: Understand the model





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

(Lapuschkin et al., in prep.)

References

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, KR Müller, G Montavon. Towards Explaining Anomalies: A Deep Taylor Decomposition of One-Class Models. arXiv:1805.06230, 2018.

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.

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

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.

KT Schütt, F. Arbabzadah, S Chmiela, KR Müller, A Tkatchenko. Quantum-chemical insights from deep tensor neural networks. Nature Communications, 8, 13890, 2017.

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

F Horst, S Lapuschkin, W Samek, KR Müller, WI Schöllhorn. Explaining the Unique Nature of Individual Gait Patterns with Deep Learning. Scientific Reports, 2019

References

Software

M Alber, S Lapuschkin, P Seegerer, M Hägele, KT Schütt, G Montavon, W Samek, KR Müller, S Dähne, PJ Kindermans. iNNvestigate neural networks!. *arXiv:1808.04260*, 2018.

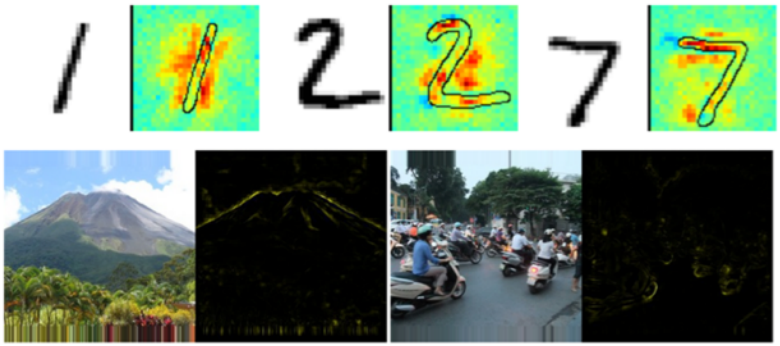
S Lapuschkin, A Binder, G Montavon, KR Müller, W Samek. The Layer-wise Relevance Propagation Toolbox for Artificial Neural Networks. *Journal of Machine Learning Research*, 17(114):1-5, 2016.

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`