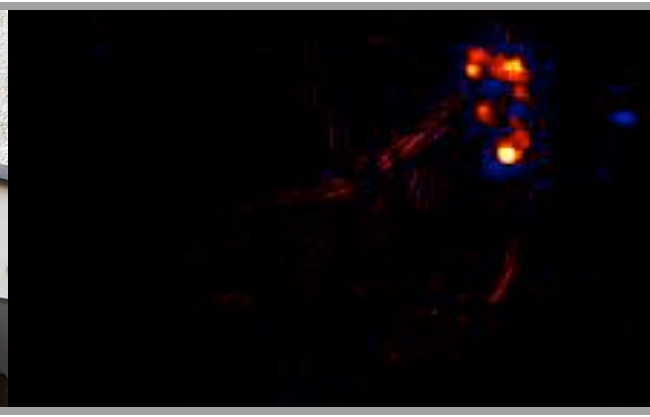
Fraunhofer
Heinrich Hertz Institute

Interpreting Deep Neural Networks and their Predictions

Wojciech Samek
ML Group, Fraunhofer HHI
(joint work with S. Lapuschkin, A. Binder, G. Montavon, K.-R. Müller)



“Superhuman” AI Systems

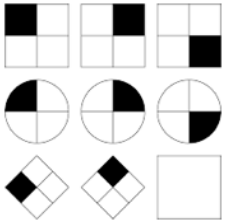
Game GO

Texas Hold’em Poker

Image classification

Traffic sign recognition

IQ Test



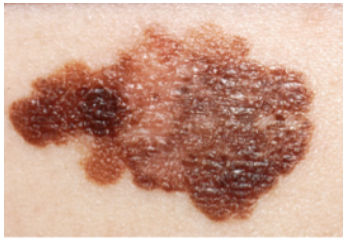
Computer games

Jeopardy

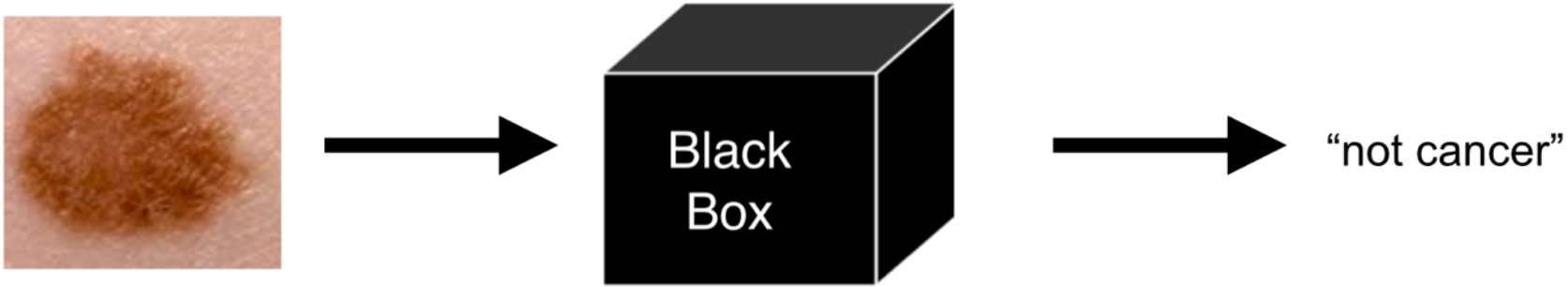
Drone control

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 ?

Is the way error is measured a satisfying specification of the problem?

Are we measuring the error on the true data distribution?

$$\min_{f \in \mathcal{F}} \int_{\mathbf{x},y} \|f(\mathbf{x}) - y\|^2 dp(\mathbf{x}, y)$$

Can we trust these black boxes ?

We need interpretability in order to:

verify system

legal aspects

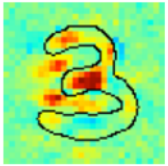
learn new things from data

understand weaknesses

Dimensions of Interpretability

prediction

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



model

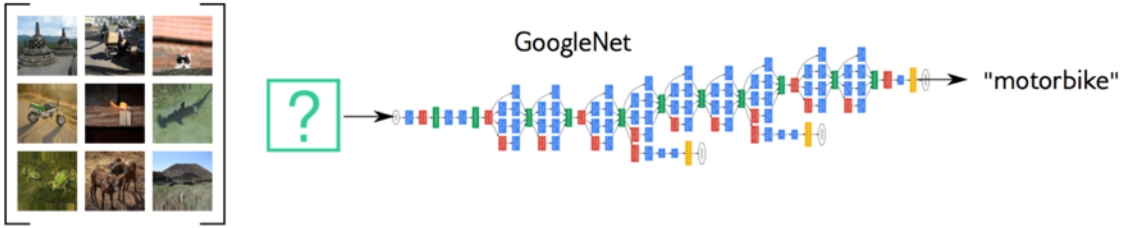
"What would a pattern belonging to a certain category typically look like according to the model."

data

"Which dimensions of the data are most relevant for the task."

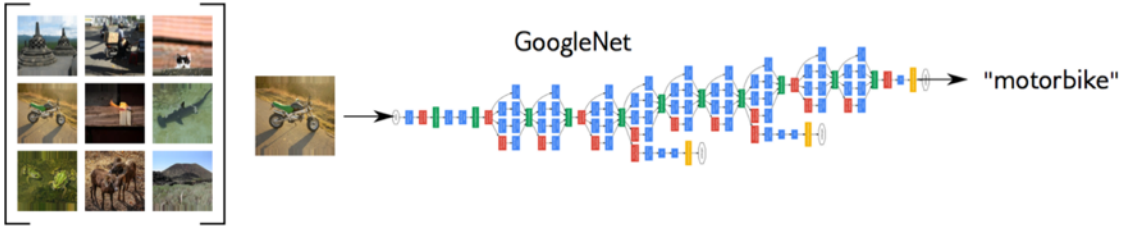
Dimensions of Interpretability

Finding a prototype:



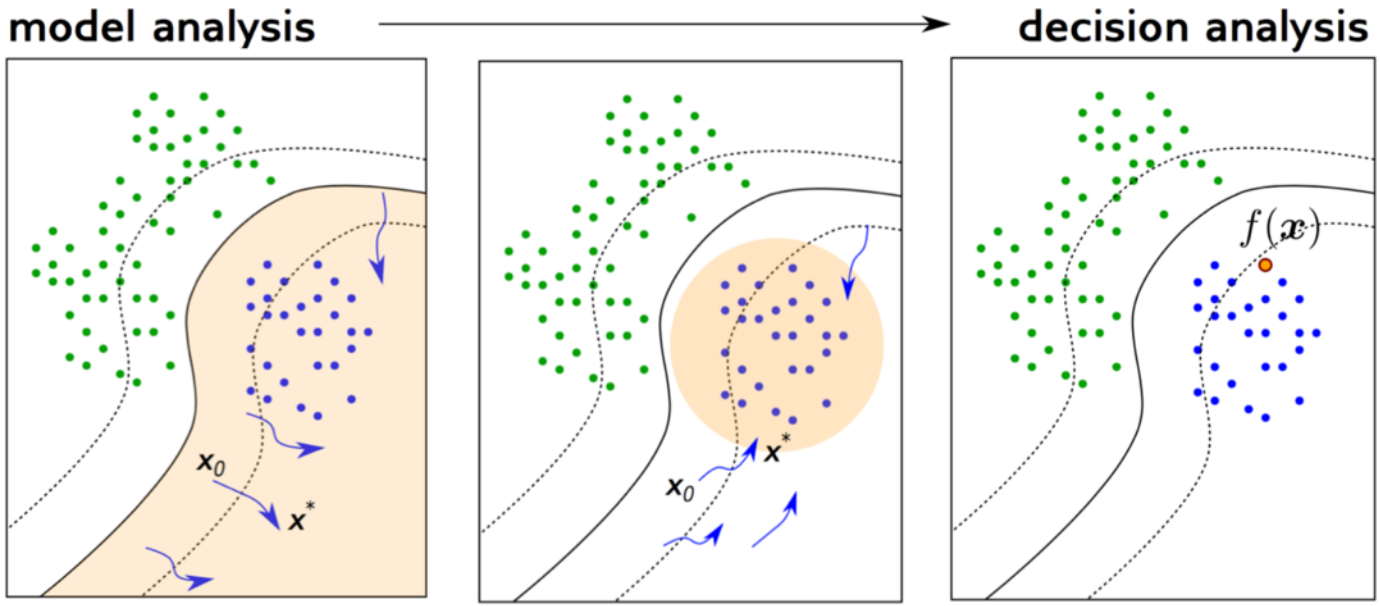
Question: How does a “motorbike” typically look like?

Individual explanation:



Question: Why is *this* example classified as a motorbike?

Dimensions of Interpretability



Find the input pattern that maximizes class probability.

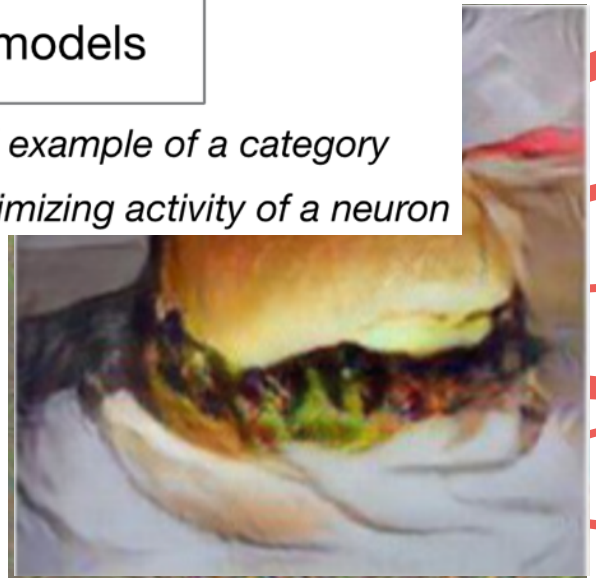
Find the most likely input pattern for a given class.

Explain individual prediction.

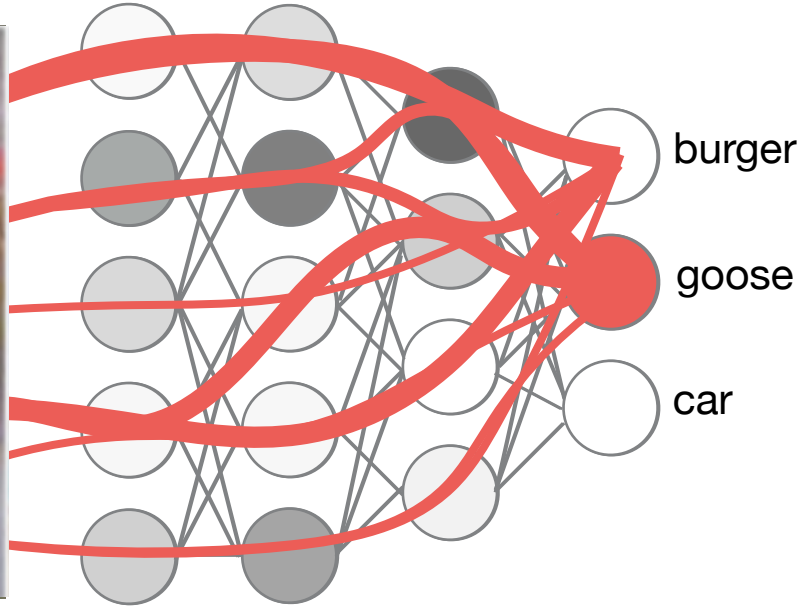
Dimensions of Interpretability

Interpreting models

- *find prototypical example of a category*
- *find pattern maximizing activity of a neuron*



simple regularizer
(Nguyen et al 2016)



$$\max_{x \in \mathcal{X}} p_\theta(\omega_c \mid x) + \lambda \Omega(x)$$

Dimensions of Interpretability

train interpretable model

suboptimal or biased due to assumptions (linearity, sparsity …)

vs.

train best model → interpret it

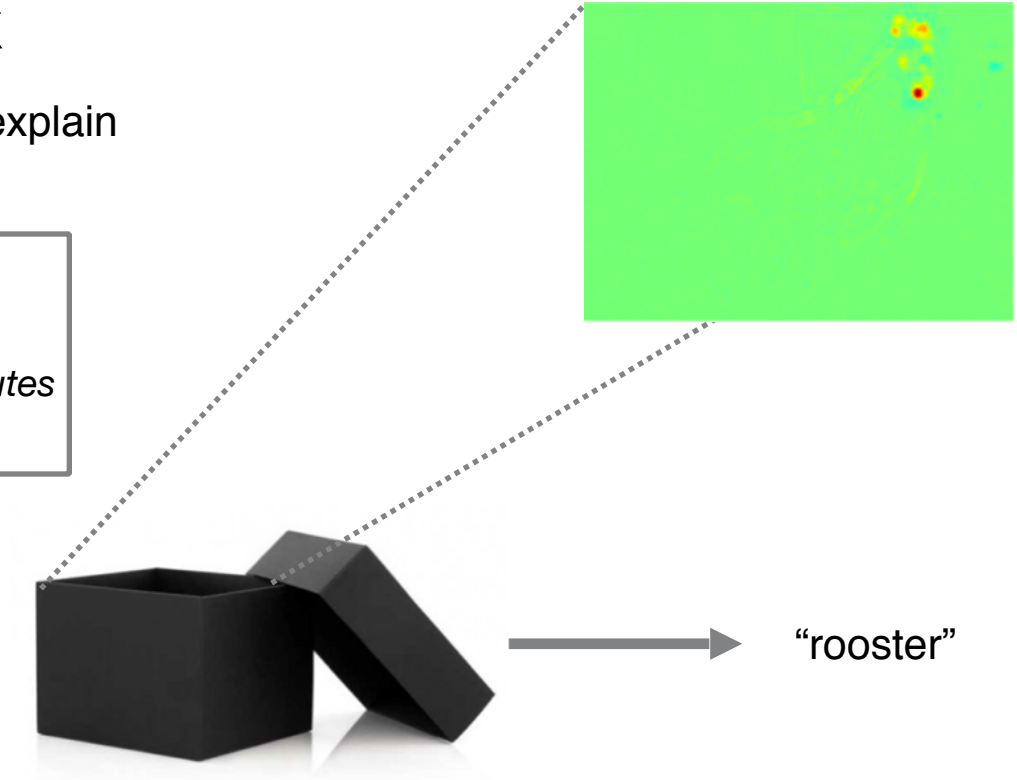
Opening the Black Box with Layer-wise Relevance Propagation

Opening the black box

We developed a *general* method to explain *individual* classification decisions.

Main idea: $\sum_p r_p = f(x)$

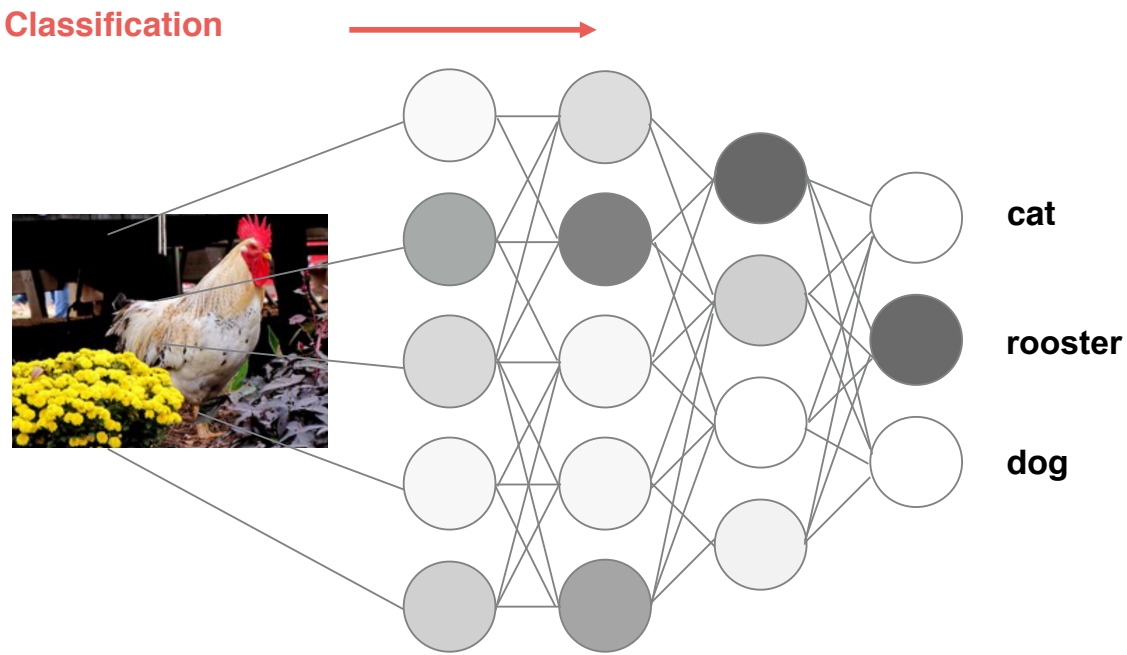
"measure how much each pixel contributes to the overall prediction"

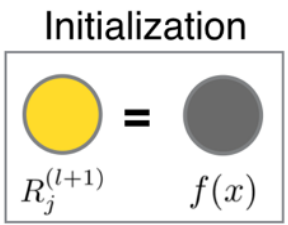


"rooster"

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

Opening the black box

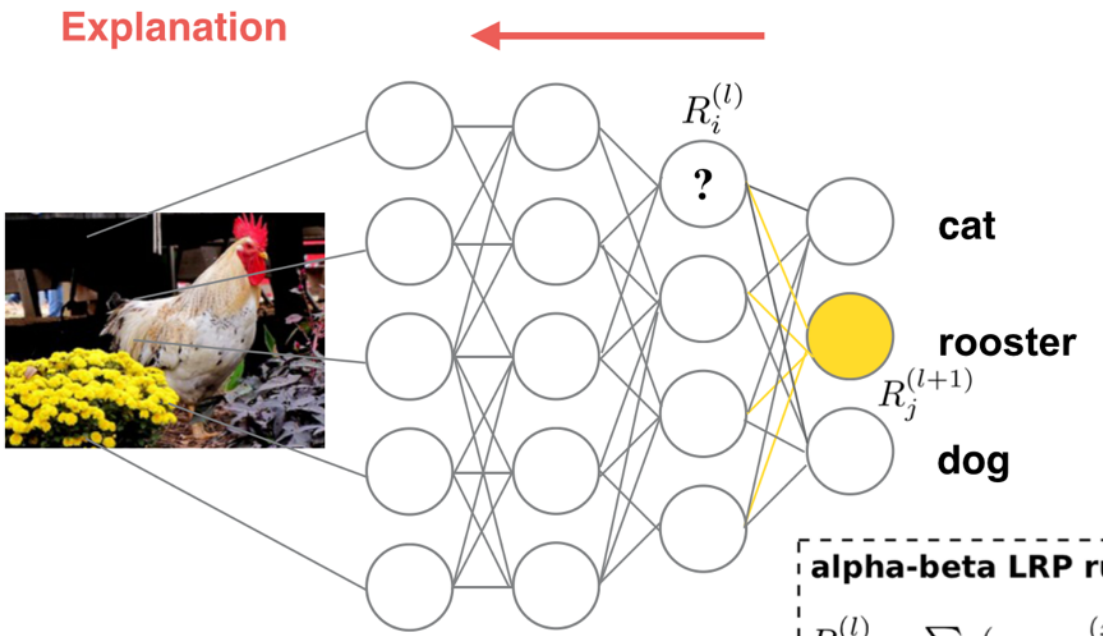




Idea: Redistribute the evidence for class rooster back to image space.

What makes this image a “rooster image” ?

Opening the black box



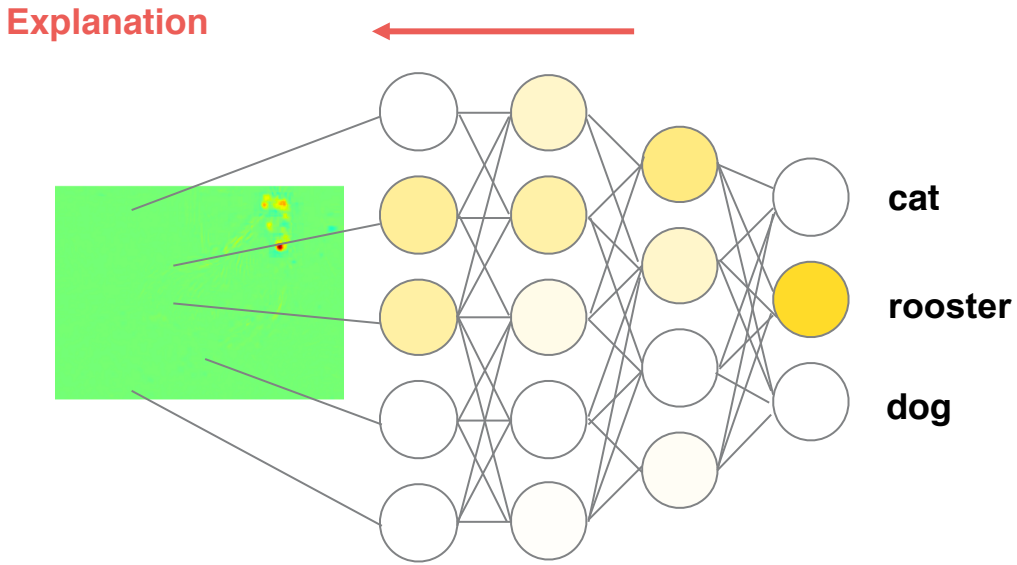
alpha-beta LRP rule (Bach et al. 2015)

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

where $\alpha + \beta = 1$

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

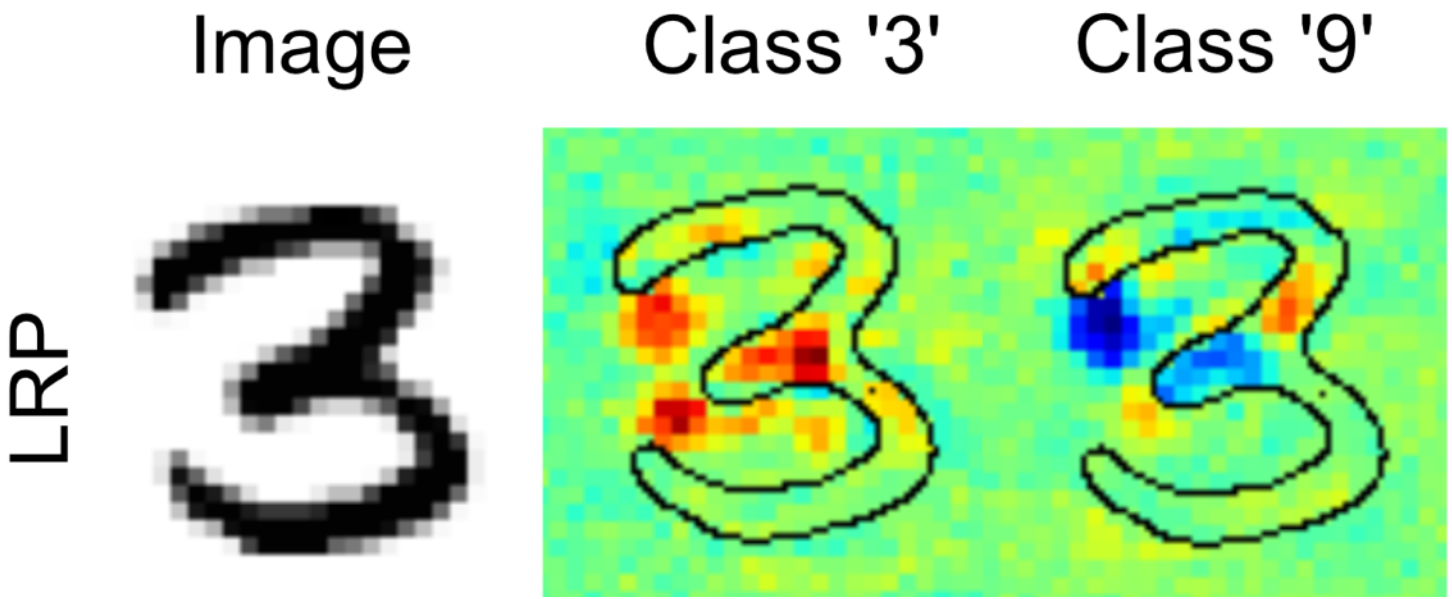
Opening the black box



Layer-wise relevance conservation

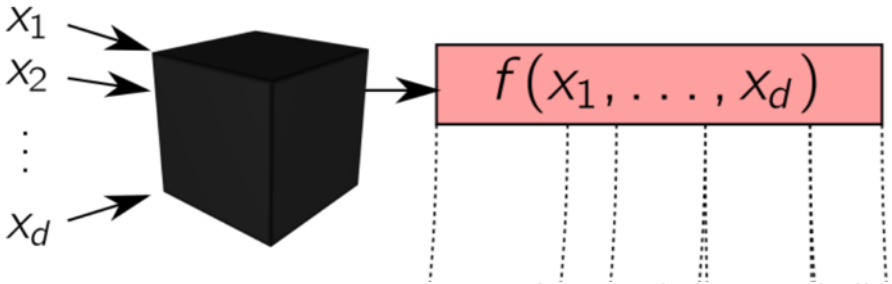
$$\sum_i R_i = \ldots = \sum_i R_i^{(l)} = \sum_j R_j^{(l+1)} = \ldots = f(x)$$

Opening the black box

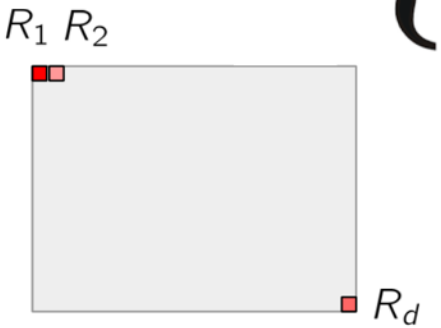


A more Principled Approach to Explanation by Decomposition

Explanation by Decomposition



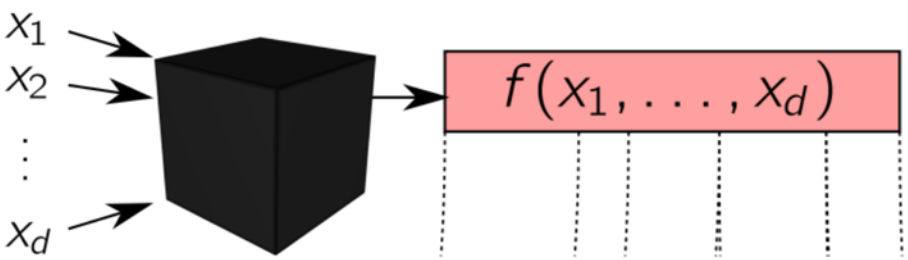
$$\left\{ \begin{array}{ll} \text{Continuity:} & \boldsymbol{x} \approx \boldsymbol{x}' \Rightarrow \boldsymbol{R}(\boldsymbol{x}) \approx \boldsymbol{R}(\boldsymbol{x}') \\ \text{Conservation:} & \sum_p R_p(\boldsymbol{x}) + \varepsilon = f(\boldsymbol{x}) \quad \text{with} \quad \varepsilon \quad \text{small.} \\ \text{Positivity:} & \forall_p : \; R_p(\boldsymbol{x}) \gtrsim 0 \end{array} \right\}$$



$$R_1 + R_2 + \ldots + R_d + \varepsilon$$

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

Explanation by Decomposition



Taylor expansion:

$$f(\boldsymbol{x}) = f(\tilde{\boldsymbol{x}}) + \sum_{p=1}^{d} \underbrace{\frac{\partial f}{\partial x_p}\Big|_{\tilde{\boldsymbol{x}}} \cdot (x_p - \tilde{x}_p)}_{R_p(\boldsymbol{x})} + \varepsilon$$

→ Taylor decomposition:

$$R_1 + R_2 + \ldots + R_d + \varepsilon$$

Explanation by Decomposition

$$f(\mathbf{x}) = f(\tilde{\mathbf{x}}) + \sum_{p=1}^{d} \frac{\partial f}{\partial x_p}\Big|_{\tilde{\mathbf{x}}} \cdot (x_p - \tilde{x}_p) + \varepsilon$$

What did we gain?

How the explanation should be built? → *How to choose the root point?* $\tilde{x}$

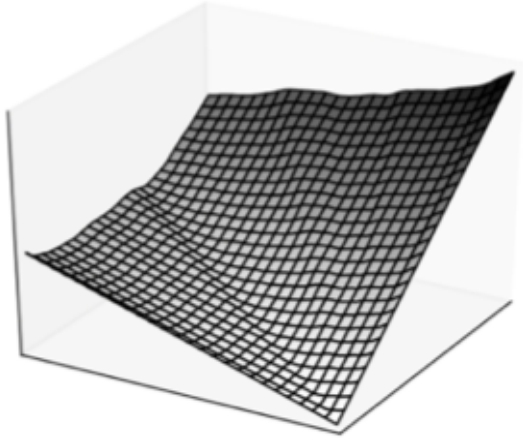
Explanation by Decomposition

For deep ReLU networks without biases we can find a root point at the origin.

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

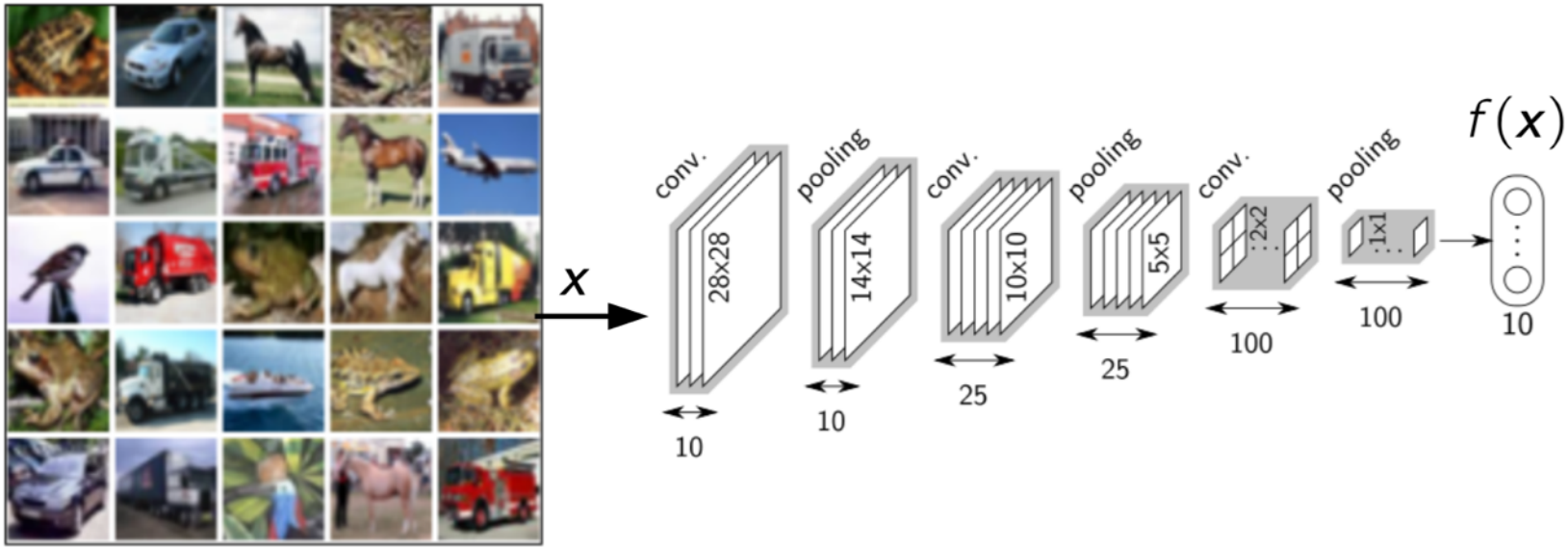
$$f(\mathbf{x}) = \underbrace{f(\tilde{\mathbf{x}})}_{0} + \sum_{i=1}^{d} \underbrace{\frac{\partial f}{\partial x_i}\Big|_{\mathbf{x}=\tilde{\mathbf{x}}} (x_i - \tilde{x}_i)}_{R_i} + \underbrace{O(\mathbf{x}\mathbf{x}^\top)}_{0}$$

$$\boxed{f(\mathbf{x}) = \sum_{i=1}^{d} R_i}$$

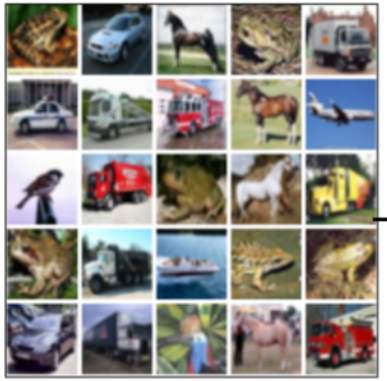


Explanation by Decomposition

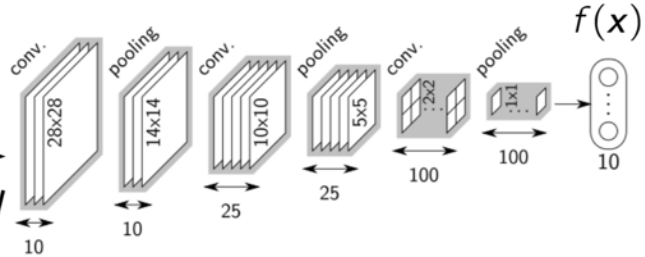
Let's try Taylor Decomposition on a ConvNet.

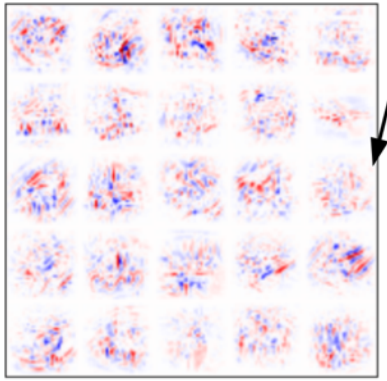


Explanation by Decomposition



x





Taylor Expansion:

$$f(\boldsymbol{x}) = f(\widetilde{\boldsymbol{x}}) + \underbrace{\sum_{p=1}^{d} \frac{\partial f}{\partial x_p}\Big|_{\widetilde{x}} \cdot (x_p - \widetilde{x}_p)}_{R_p(\boldsymbol{x})} + \varepsilon$$

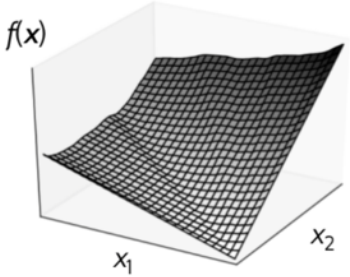
Explanations are noisy and uninformative.
Question: what went wrong?

Explanation by Decomposition

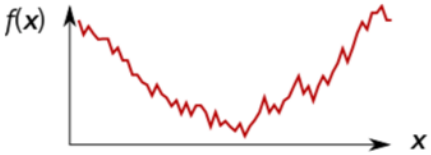
"Naive" Taylor decomposition of neural network does not give satisfactory results.

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

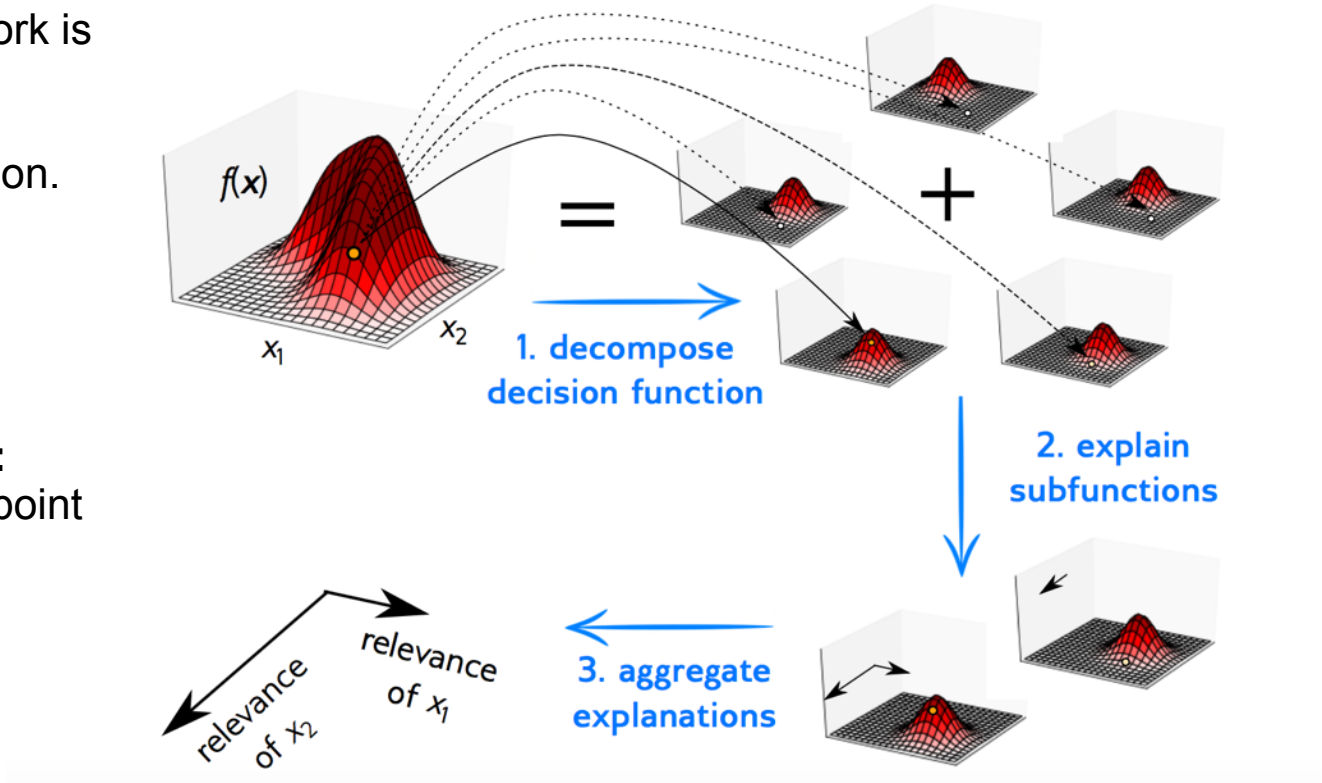


Explanation by Decomposition

Key Idea: If a decision is too complex to explain, break the decision function into sub-functions, and explain each sub-decision separately.

Explanation by Decomposition

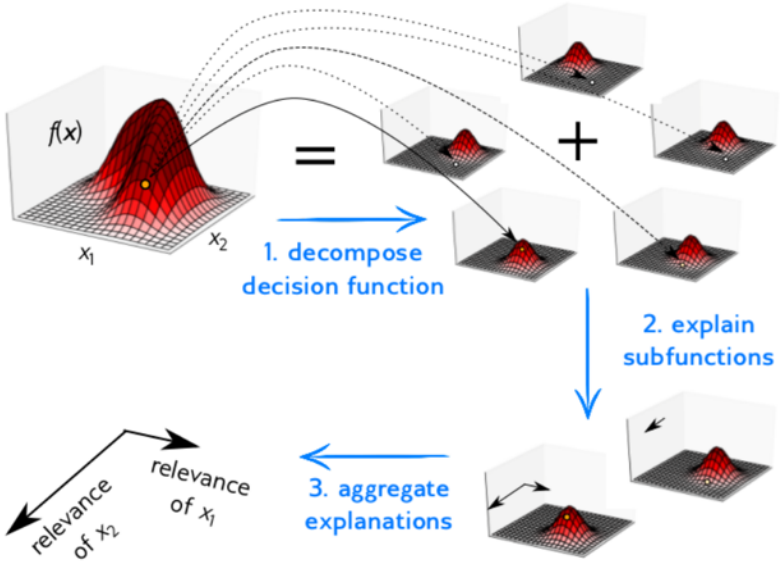
Idea: Since neural network is composed of simple functions, we propose a *deep* Taylor decomposition.



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

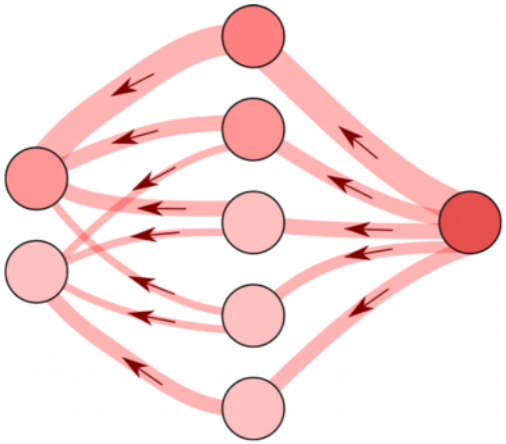
Explanation by Decomposition

Function view



Insight: In order to explain, decompose the decision function and explain sub functions.

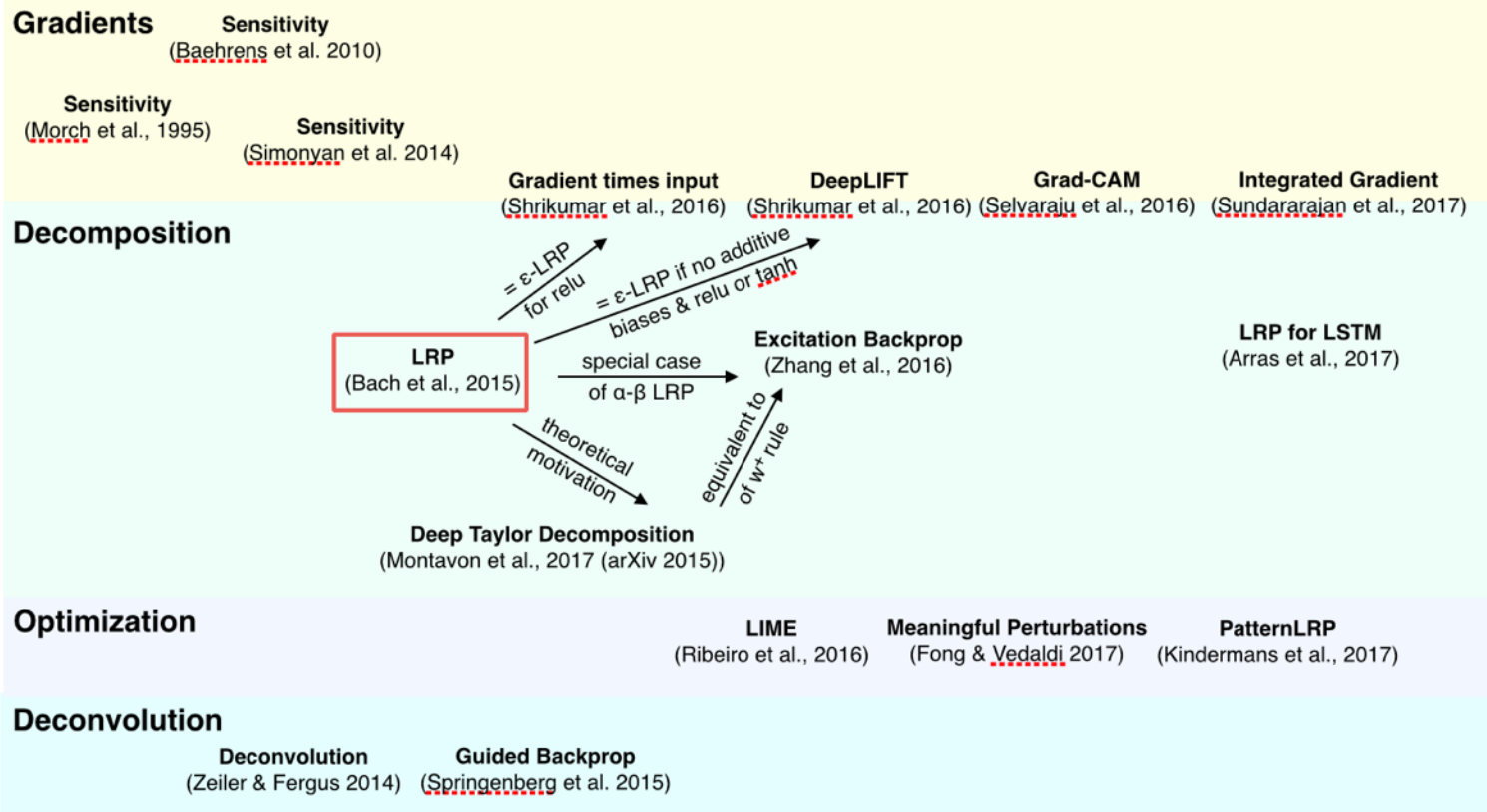
Graph propagation view



Insight: In order to explain, reuse the structure of the neural network.

Comparison with Other Explanation Methods

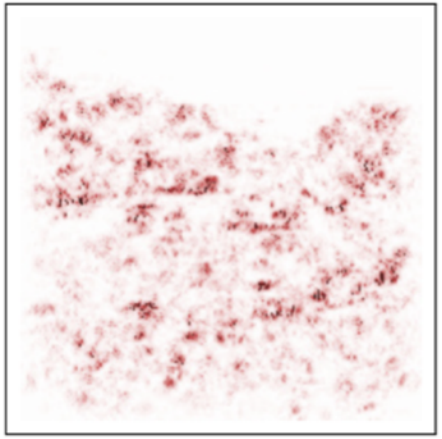
Other Explanation Methods



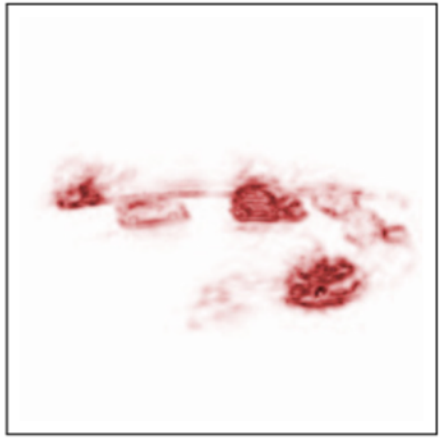
Other Explanation Methods

Image

Sensitivity Analysis



LRP / Deep Taylor



Explains what influences prediction “cars”.

Explains prediction “cars” as is.

Slope decomposition

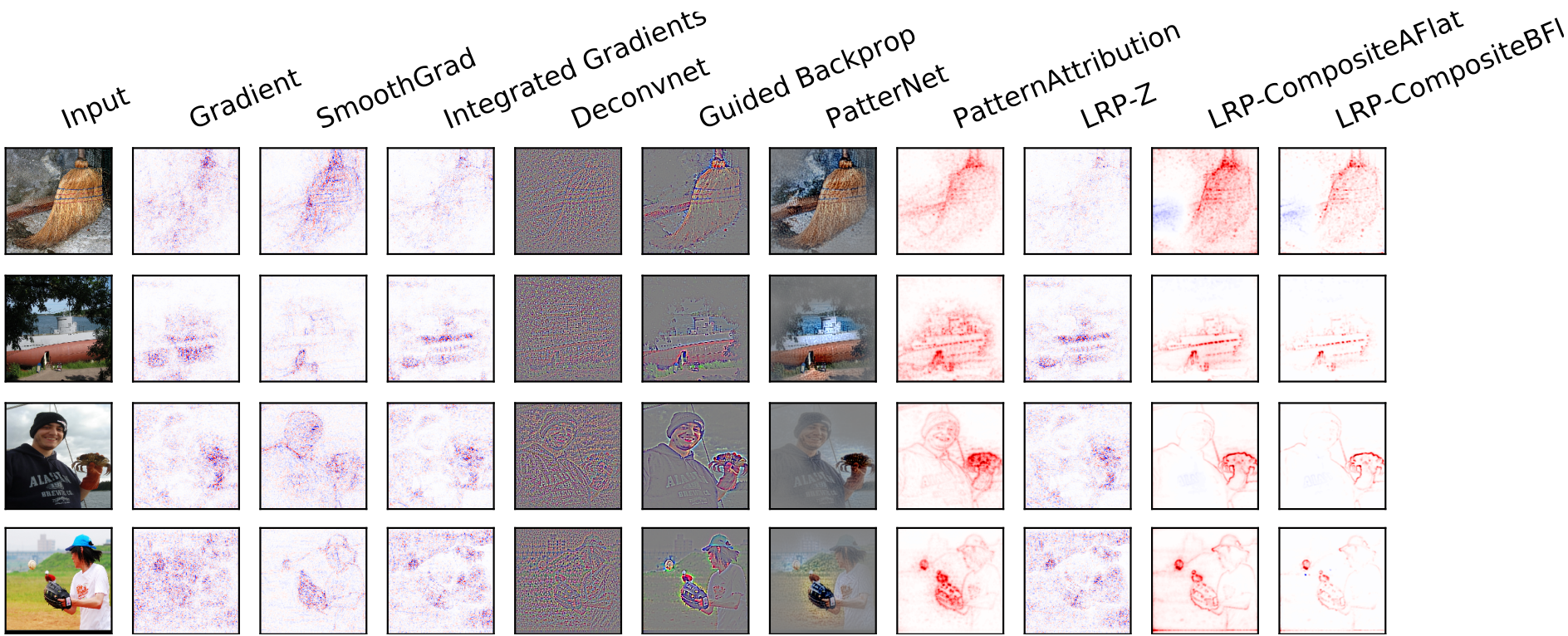
$$\sum_i R_i = \|\nabla_x f\|^2$$

Value decomposition

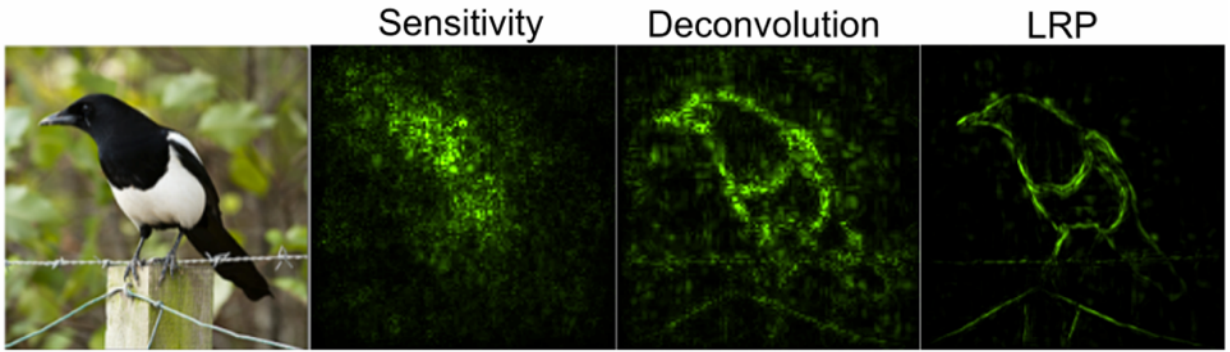
$$\sum_i R_i = f(\mathbf{x})$$

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

Other Explanation Methods



Measuring Quality of Explanations



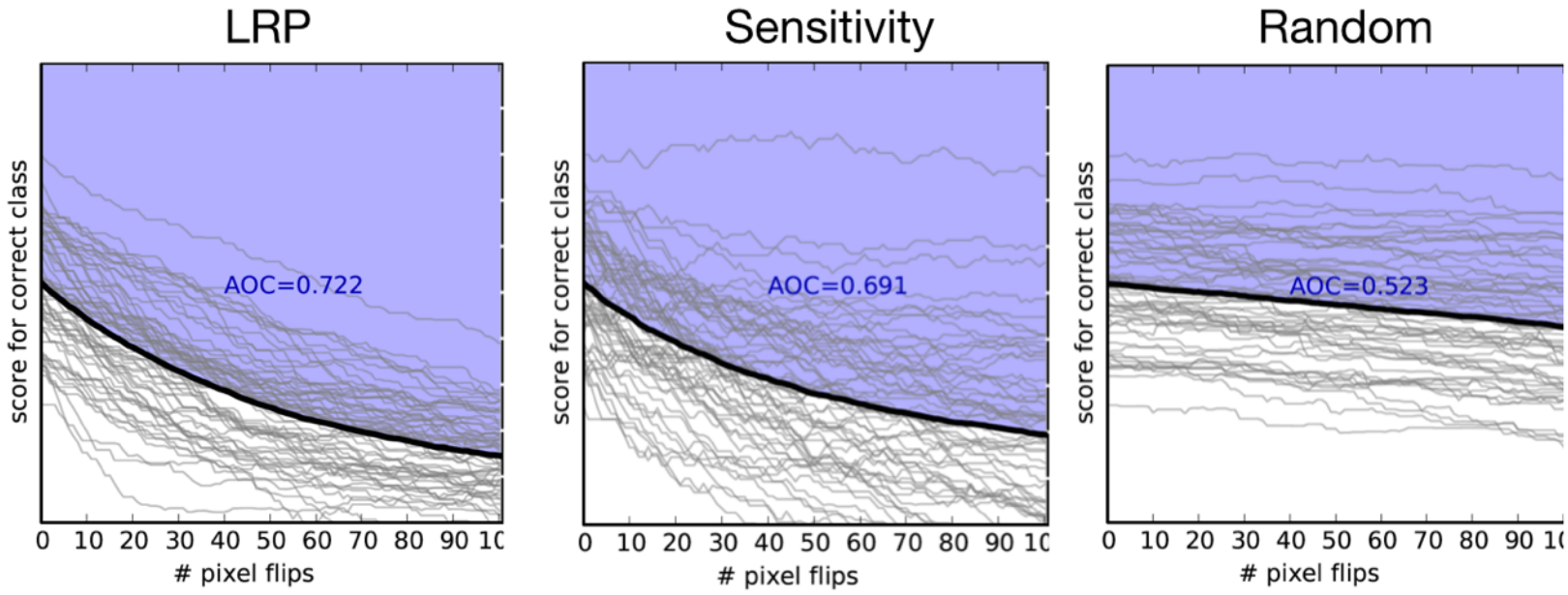
Can we objectively measure which heatmap is best ?

Algorithm (Pixel Flipping)

```
Sort pixel scores
Iterate
   flip pixels
   evaluate f(x)
Measure decrease of f(x)
```

(Samek et al., 2017)

Measuring Quality of Explanations



LRP outperforms other methods on MNIST.

Measuring Quality of Explanations

What about more complex datasets ?

SUN397

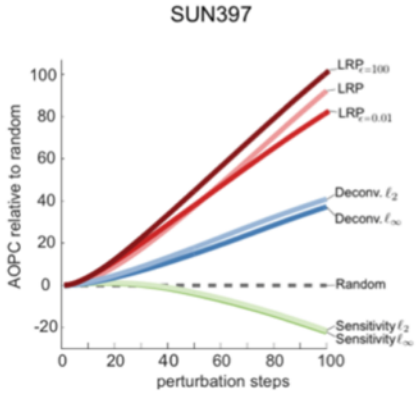
397 scene categories
(108,754 images in total)

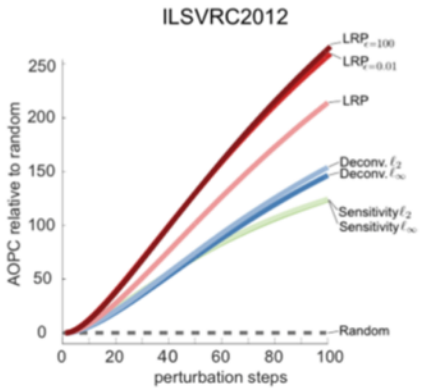
ILSVRC2012

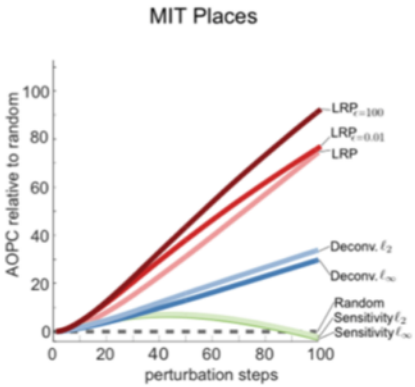
1000 categories
(1.2 million training images)

MIT Places

205 scene categories
(2.5 millions of images)







Interpretability in Practice

Application: Compare Classifiers

Visual Object Classes Challenge: 2005 - 2012

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%

Two classifiers

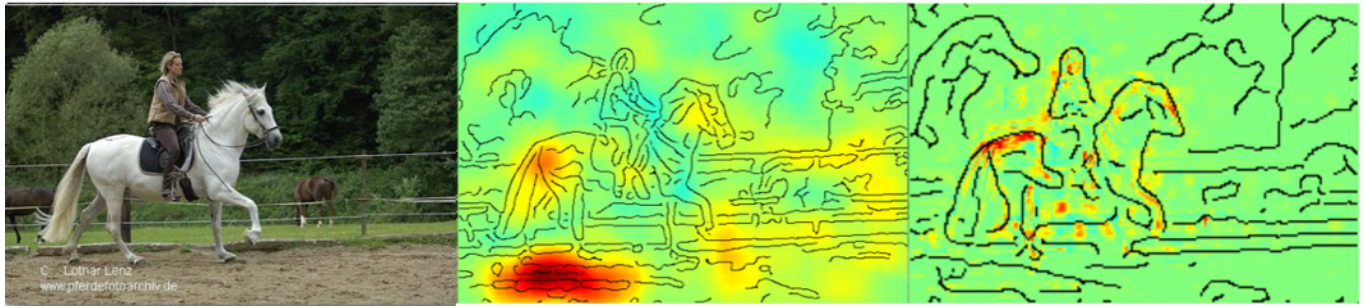
- **similar classification accuracy on horse class**
- **but do they solve the problem similarly ?**

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 FV DNN



(Lapuschkin et al., 2016)

Application: Compare Classifiers

Application: Compare Classifiers

20 Newsgroups data set

comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

Test set performance

word2vec / CNN model: 80.19%

BoW/SVM model: 80.10%

same performance —> same strategy ?

Application: Compare Classifiers

word2vec/CNN:
identifies semantically meaningful words

sci.med (4.1)

Yes, weightlessness does feel like falling. It may feel strange at first, but the body does adjust. The feeling is not too different from that of sky diving.

>And what is the motion sickness
>that some astronauts occasionally experience?

It is the body's reaction to a strange environment. It appears to be induced partly to physical discomfort and part to mental distress. Some people are more prone to it than others, like some people are more prone to get sick on a roller coaster ride than others. The mental part is usually induced by a lack of clear indication of which way is up or down, ie: the Shuttle is normally oriented with its cargo bay pointed towards Earth, so the Earth (or ground) is "above" the head of the astronauts. About 50% of the astronauts experience some form of motion sickness, and NASA has done numerous tests in space to try to see how to keep the number of occurances down.

BoW/SVM:
identifies statistical patterns (word statistics)

sci.med (-0.6)

Yes, weightlessness does feel like falling. It may feel strange at first, but the body does adjust. The feeling is not too different from that of sky diving.

>And what is the motion sickness
>that some astronauts occasionally experience?

It is the body's reaction to a strange environment. It appears to be induced partly to physical discomfort and part to mental distress. Some people are more prone to it than others, like some people are more prone to get sick on a roller coaster ride than others. The mental part is usually induced by a lack of clear indication of which way is up or down, ie: the Shuttle is normally oriented with its cargo bay pointed towards Earth, so the Earth (or ground) is "above" the head of the astronauts. About 50% of the astronauts experience some form of motion sickness, and NASA has done numerous tests in space to try to see how to keep the number of occurances down.

(Arras et al., 2016)

Application: Context Use

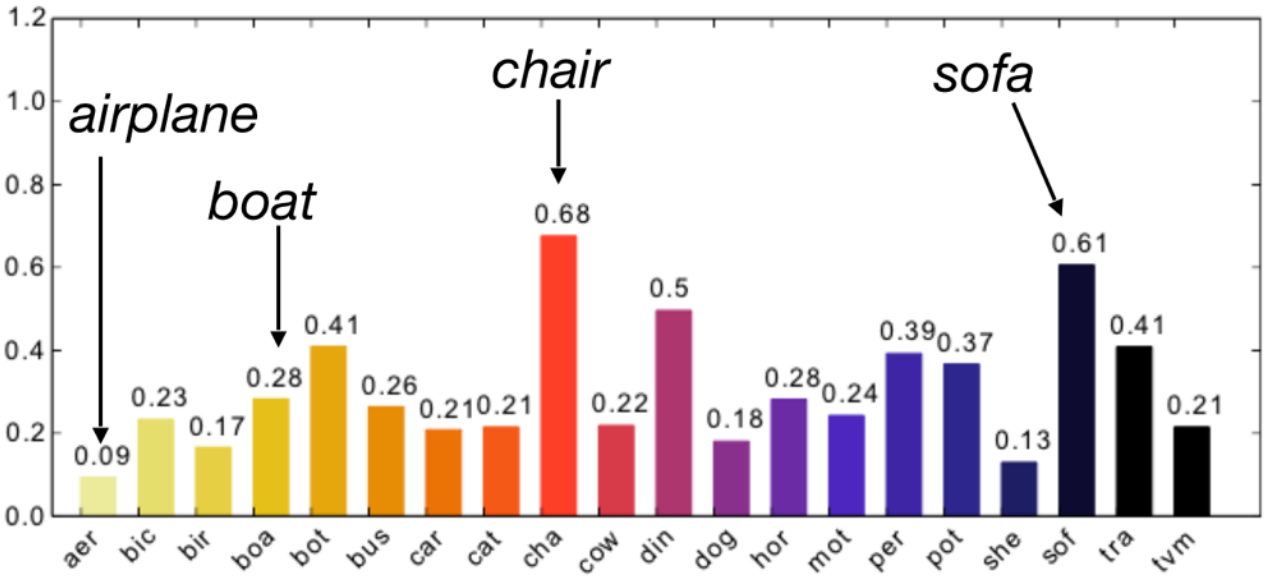
how important is context ?

how important is context ?

classifier

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

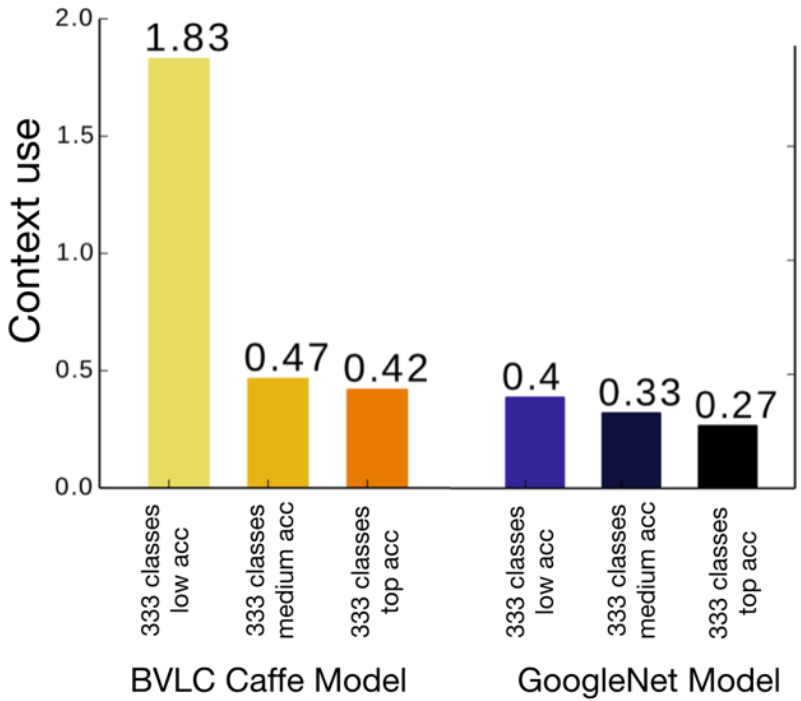
Application: Context Use

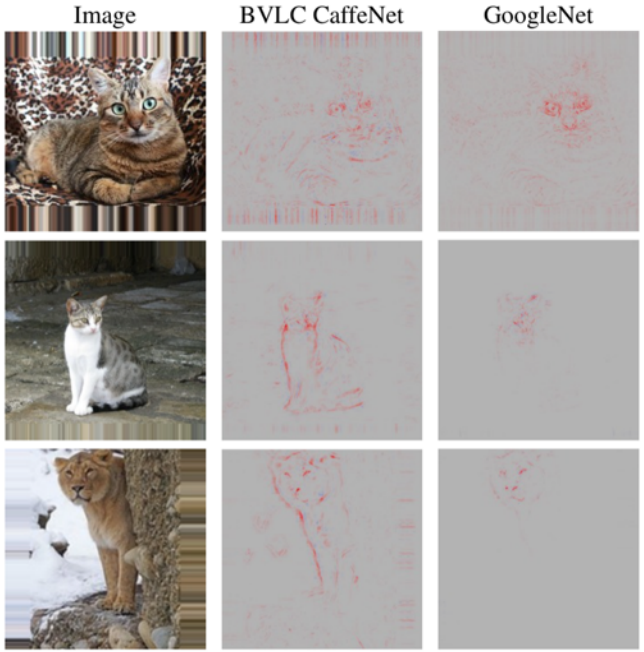


(Lapuschkin et al., 2016)

Application: Context Use

Context use anti-correlated with performance.





(Lapuschkin et al., 2016)

Application: Recurrent Networks

Model: word-based bidirectional LSTM

word embeddings of dimension 60, one hidden layer of size 60

takes as input a sequence of word embeddings $(x_1, x_2, \ldots, x_T)$

[model released by Li et al. 2016]

Task: five-class sentiment prediction

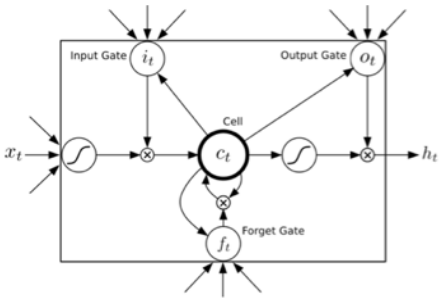
training on phrases and sentences of the Stanford Sentiment Treebank

[dataset released by Socher et al. 2013]

reccurence of the form:

$$i_t = \texttt{sigm}\left(W_i\, h_{t-1} + U_i\, x_t + b_i\right)$$
$$f_t = \texttt{sigm}\left(W_f\, h_{t-1} + U_f\, x_t + b_f\right)$$
$$o_t = \texttt{sigm}\left(W_o\, h_{t-1} + U_o\, x_t + b_o\right)$$
$$g_t = \texttt{tanh}\left(W_g\, h_{t-1} + U_g\, x_t + b_g\right)$$
$$c_t = f_t \odot c_{t-1} \; + \; i_t \odot g_t$$
$$h_t = o_t \odot \texttt{tanh}(c_t)$$

[Hochreiter&Schmidhuber 1997, Gers et al. 2000]



(Arras et al., 2017)

How to handle multiplicative interactions ?

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

Application: Recurrent Networks

movie review:
++, –

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 .

(Arras et al., 2017)

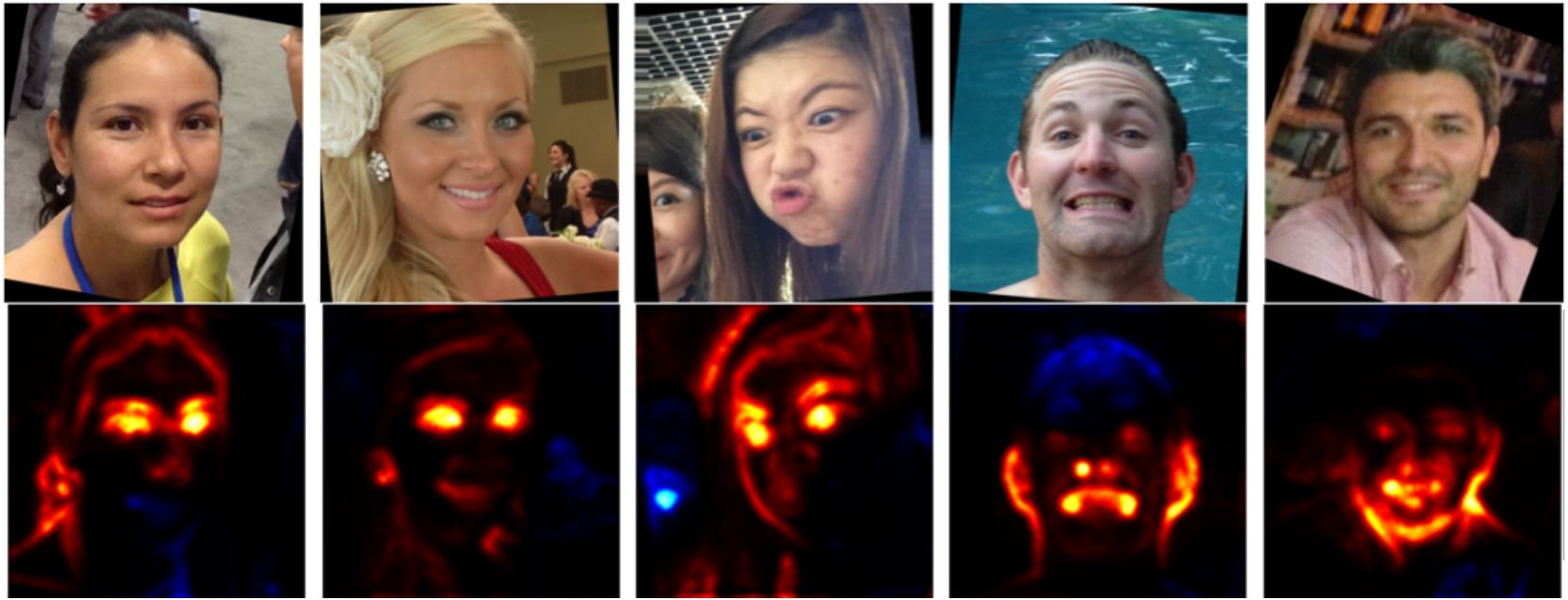
Application: Face Analysis

Faces in the wild
(from Flickr)

#images: 26,580

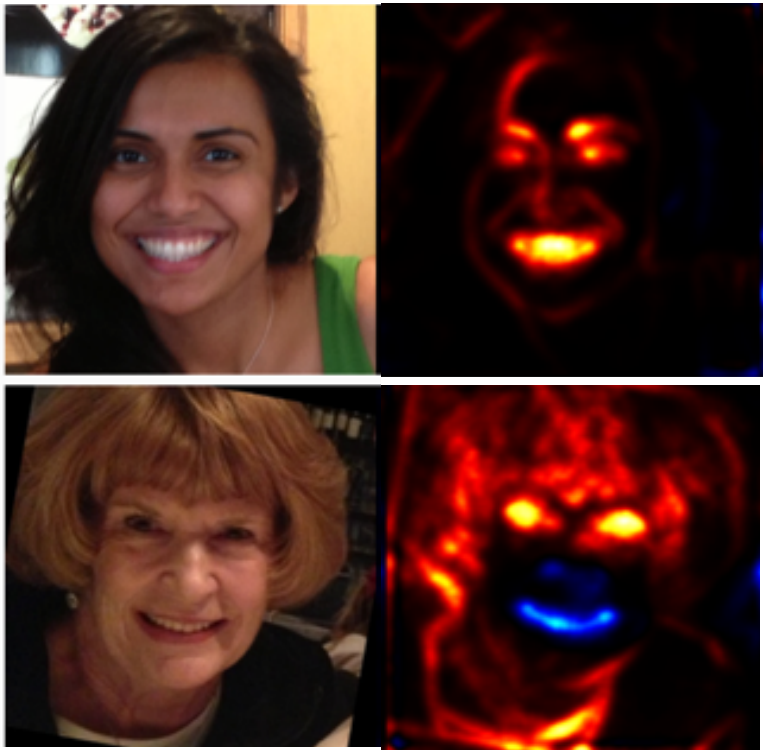
Task: Predict gender & age (range)

Application: Face Analysis



(Lapuschkin et al., 2017)

Application: Face Analysis



Predictions

25-32 years old

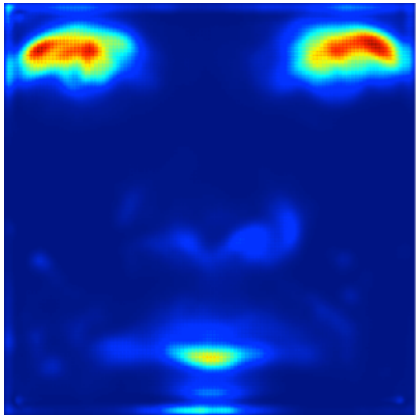
60+ years old

(Lapuschkin et al., 2017)

Application: Face Analysis

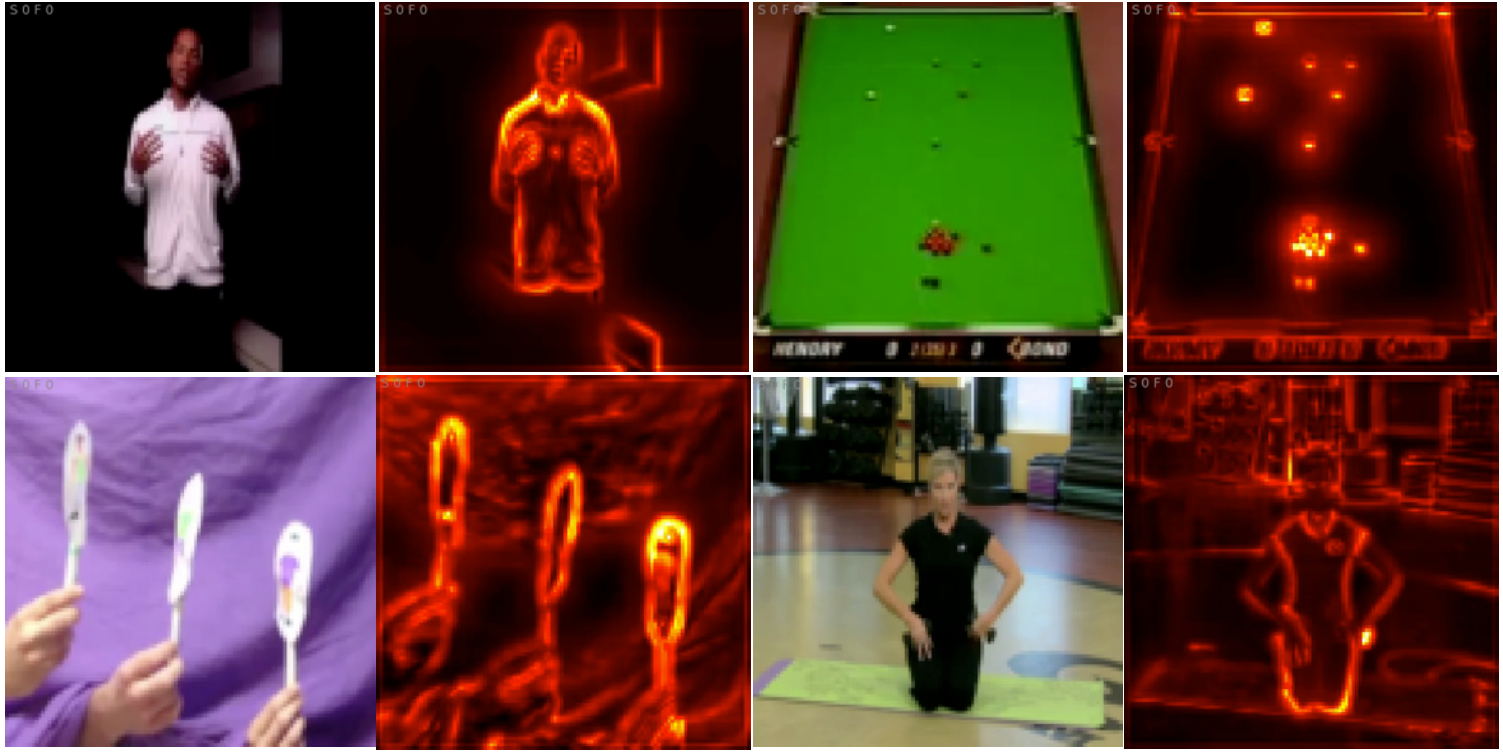
real person | fake person | real person

fake persons have different eyes

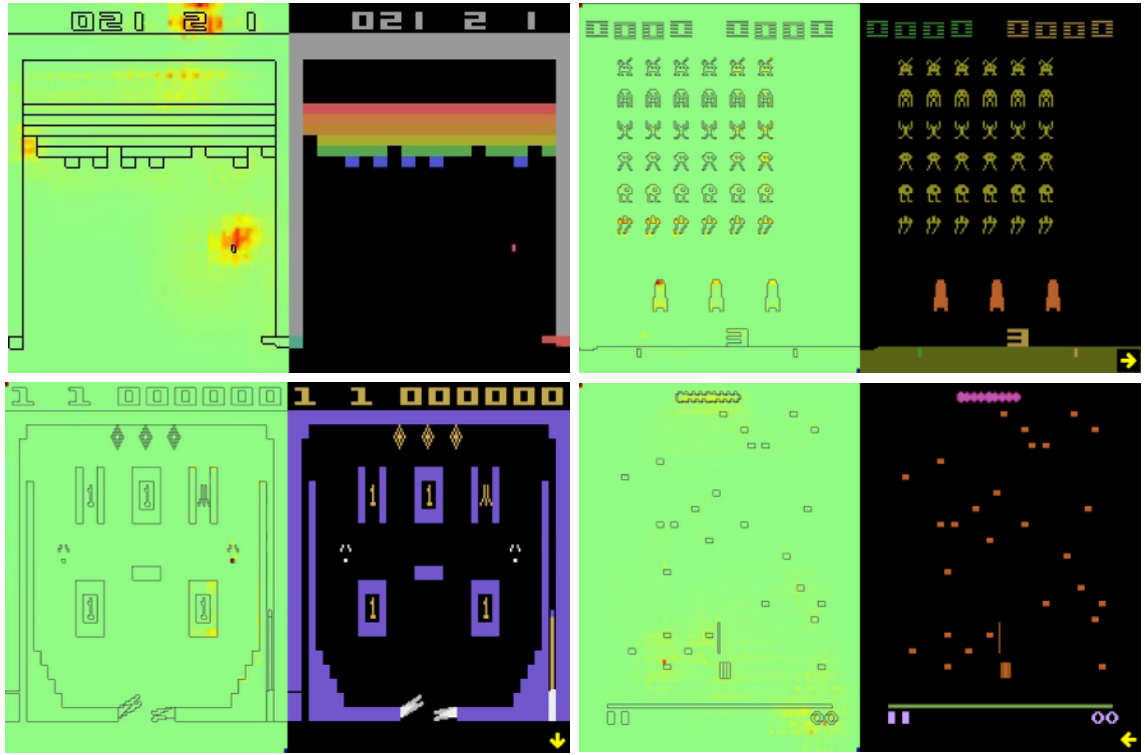


(Seibold et al., 2017)

Application: Video Analysis

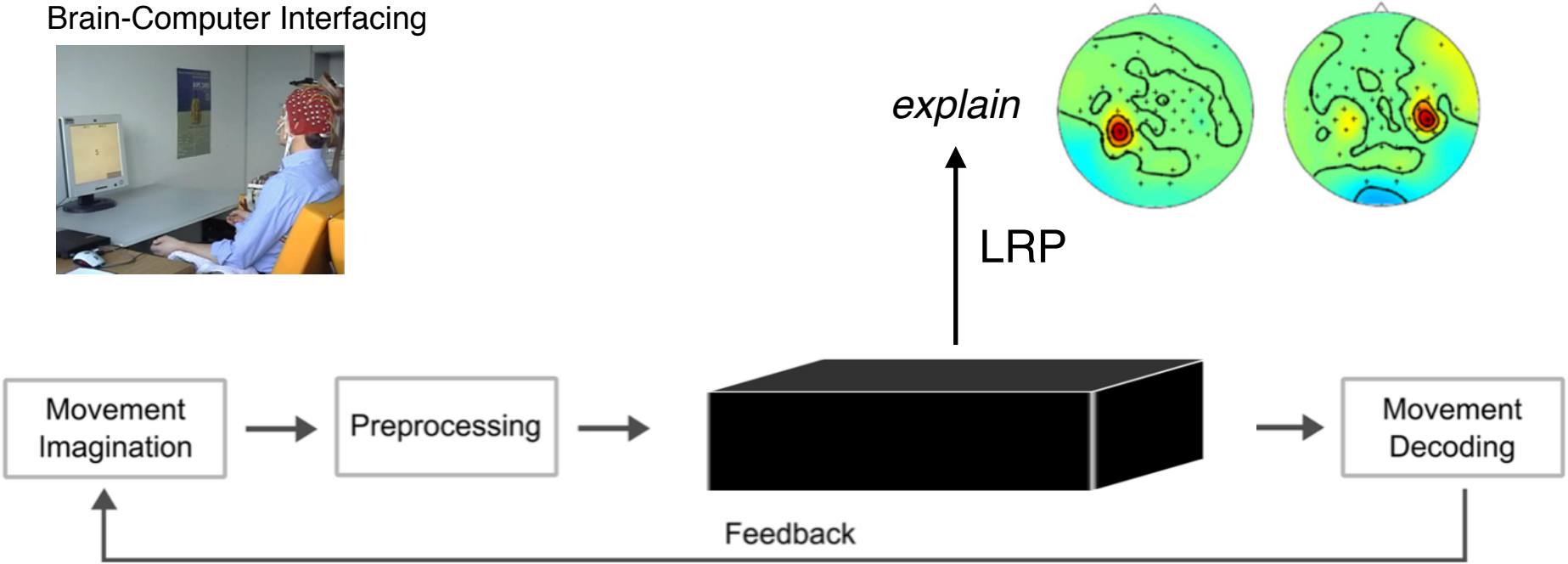


Application: Machines Playing Games



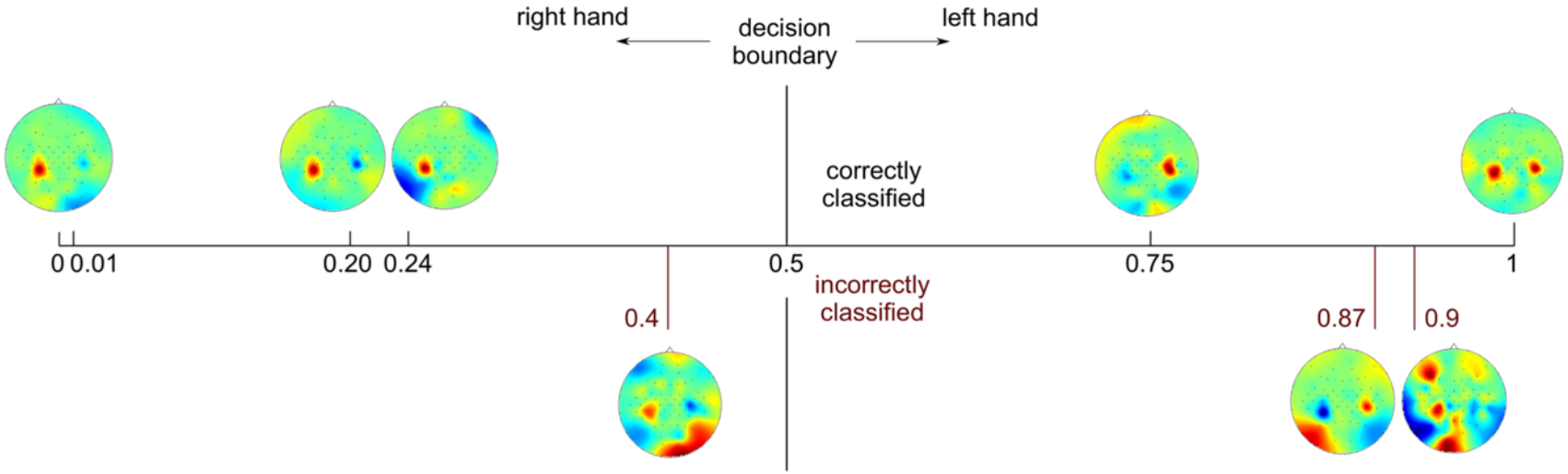
Application: Biomedical Engineering

Brain-Computer Interfacing



(Sturm et al., 2016)

Application: Biomedical Engineering



Summary

In many problems interpretability as important as prediction
(trusting a black-box system may not be an option).

Use in practice
- verify predictions, detect biases and flaws, debug models
- compare and select architectures, understand and improve models
- extract additional information, perform further tasks

We have a powerful, mathematically well-founded method to explain individual predictions of complex machine learning models.

Upcoming Tutorials

CVPR 2018

MICCAI 2018 Granada SPAIN

ICIP 2018

References

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.

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.

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.

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.

References

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 Y Koh, W Samek, KR Müller, A Binder. "Object Boundary Detection and Classification with Image-level Labels" *Pattern Recognition - 39th German Conference, GCPR 2017*, Lecture Notes in Computer Science, Springer International Publishing, 2017.

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 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.

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.

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

References

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

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.

W Samek, G Montavon, A Binder, S Lapuschkin, KR Müller. Interpreting the Predictions of Complex ML Models by Layer-wise Relevance Propagation. *NIPS'16 Workshop on Interpretable ML for Complex Systems*, 1-5, 2016

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, 1:1-10, 2017

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)*, 2017.

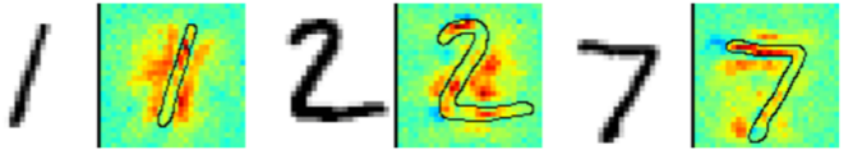
I Sturm, S Lapuschkin, W Samek, KR Müller. Interpretable Deep Neural Networks for Single-Trial EEG Classification. *Journal of Neuroscience Methods*, 274:141–145, 2016.

Thank you for your attention

Visit:

http://www.heatmapping.org

- Tutorials
- Software
- Online Demos



Tutorial Papers:

Montavon et al. “Methods for Interpreting and Understanding Deep Neural Networks”, 2018

Samek et al. “Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models”, 2018

Acknowledgement

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