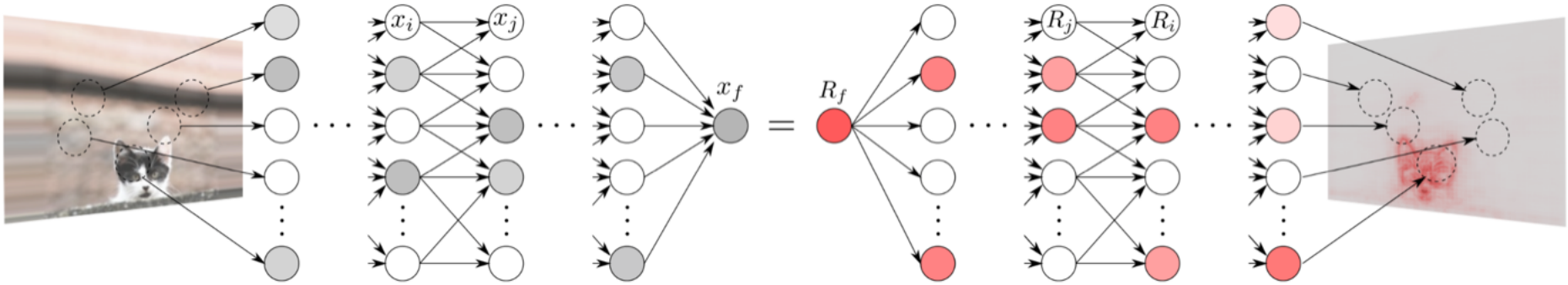
Towards explainable Deep Learning

Wojciech Samek
Machine Learning Group
Fraunhofer HHI



Before we start

Acknowledgements

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

Slides, code and demos

`http://interpretable-ml.org/`

Overview papers

Montavon et al., "Methods for Interpreting and Understanding Deep Neural Networks", *Digital Signal Processing*, 73:1-15, 2017

Samek et al., "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models", ITU Journal: ICT Discoveries, 1:1-10, 2017

NIPS Workshop “Interpreting, Explaining and Visualizing Deep Learning - Now what ?”, 9th Dec, Long Beach, CA

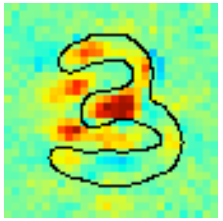
Recap

Dimensions of Interpretation

Different dimensions of "interpretability"

prediction

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

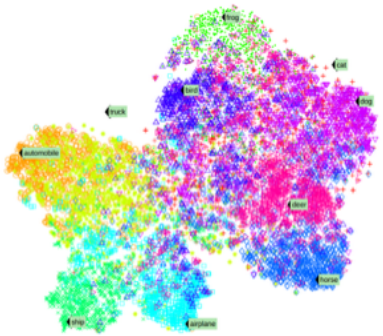


model

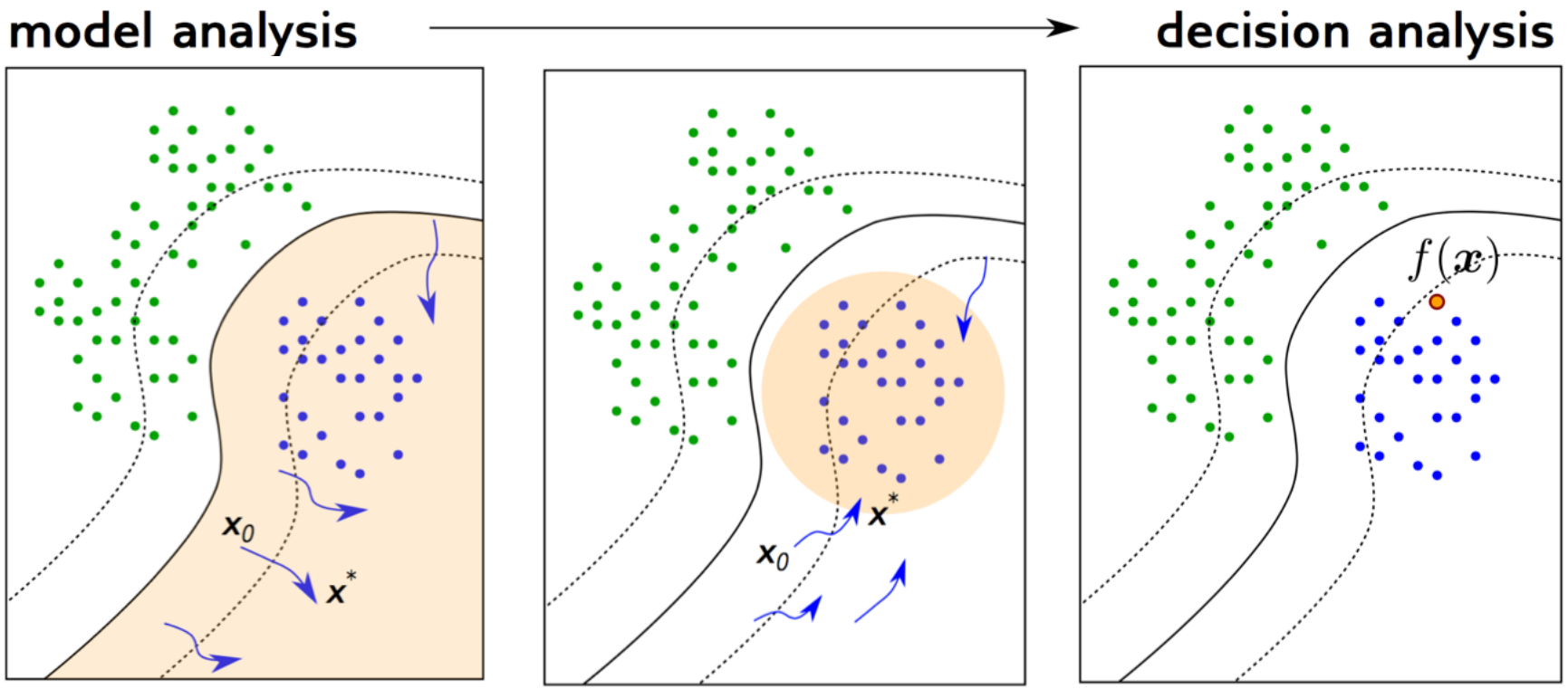
"What would a pattern belonging to a certain category typically look like according to the model."

data

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



Dimensions of Interpretation



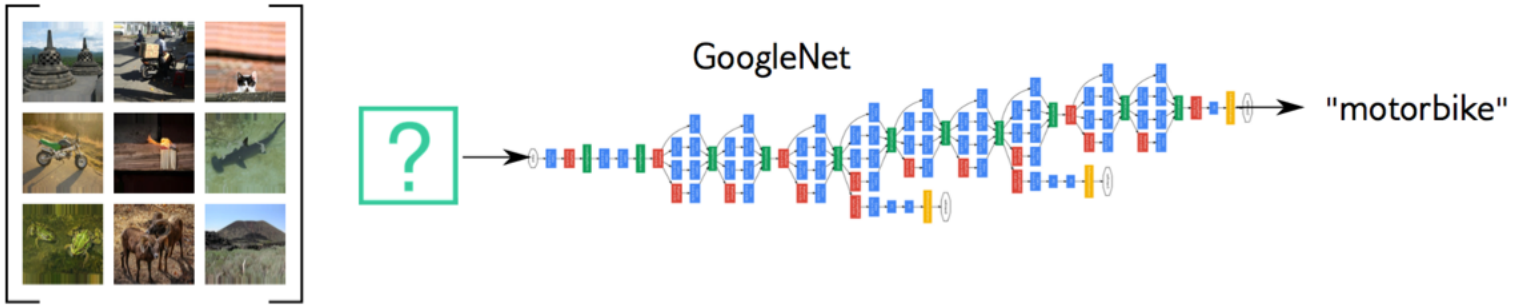
Find the input pattern that maximizes class probability.

Find the most likely input pattern for a given class.

Explain individual prediction.

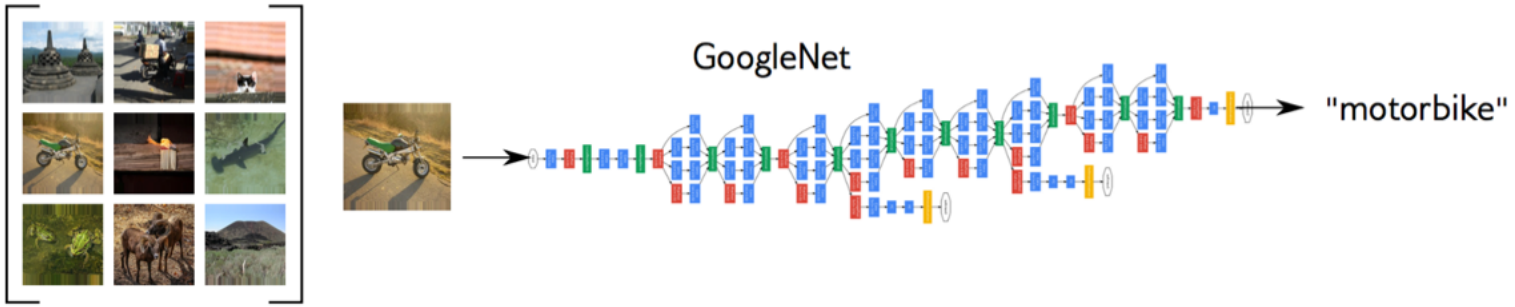
Dimensions of Interpretation

Finding a prototype:



Question: How does a "motorbike" typically look like?

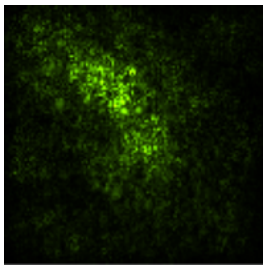
Individual explanation:



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

Techniques of Interpretation

Sensitivity Analysis
(Simonyan et al. 2014)



Classifier

Bird

prediction $f(x)$

Explain prediction

(which pixels lead to decrease of prediction score when changed)

$$\left|\left| \frac{\partial}{\partial x_p} f(x) \right|\right|$$

Techniques of Interpretation

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

input $\boldsymbol{x}$

Classifier

Rooster

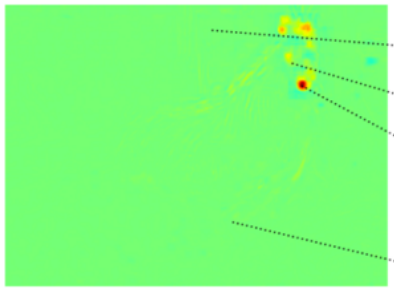
prediction $f(\boldsymbol{x})$

Explain prediction

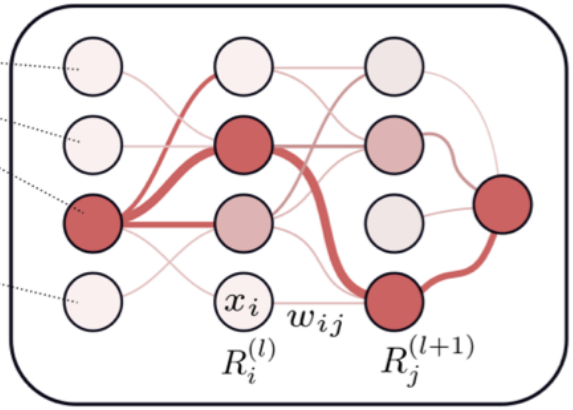
(how much each pixel contributes to prediction)

Idea: Decompose function

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



heatmap

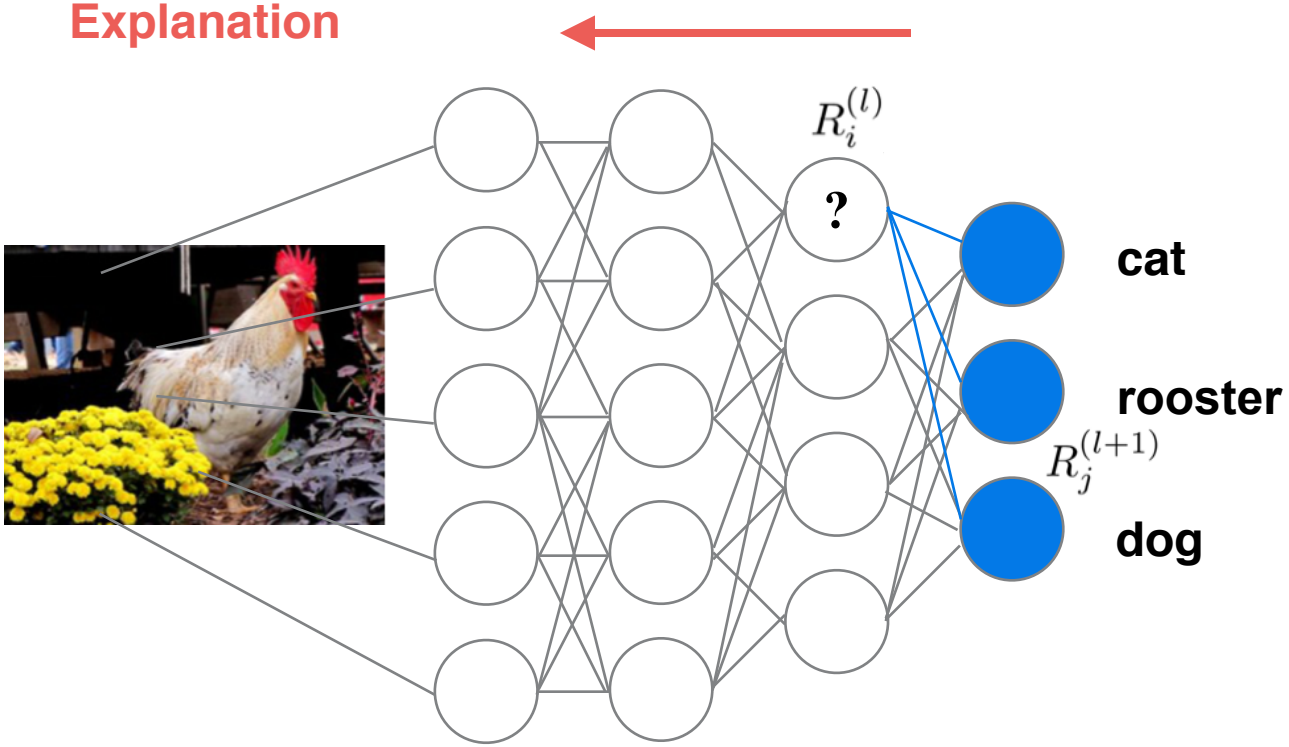


"every neuron gets it's share of relevance depending on activation and strength of connection."

redistribute $f(\boldsymbol{x})$

Techniques of Interpretation

Explanation



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

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

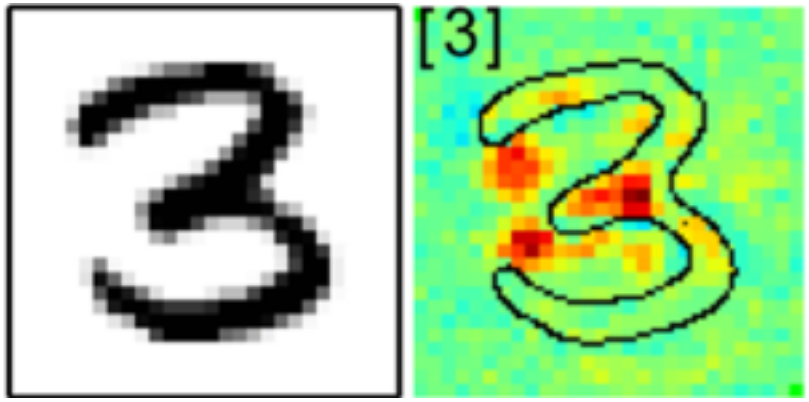
$$R_i^{(l)} = \sum_j \left(\alpha \cdot \frac{(x_i \cdot w_{ij})^+}{\sum_{i'} (x_{i'} \cdot w_{i'j})^+} + \beta \cdot \frac{(x_i \cdot w_{ij})^-}{\sum_{i'} (x_{i'} \cdot w_{i'j})^-}\right) R_j^{(l+1)}$$

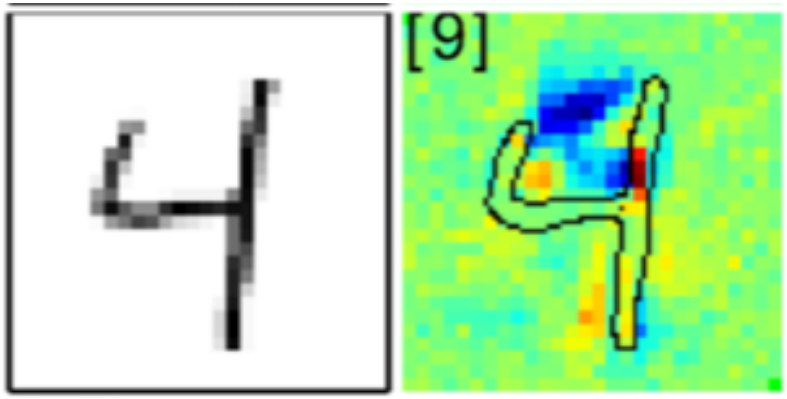
where $\alpha + \beta = 1$

Techniques of Interpretation

what speaks for / against classification as “3”

what speaks for / against classification as “9”





[number]: explanation target class

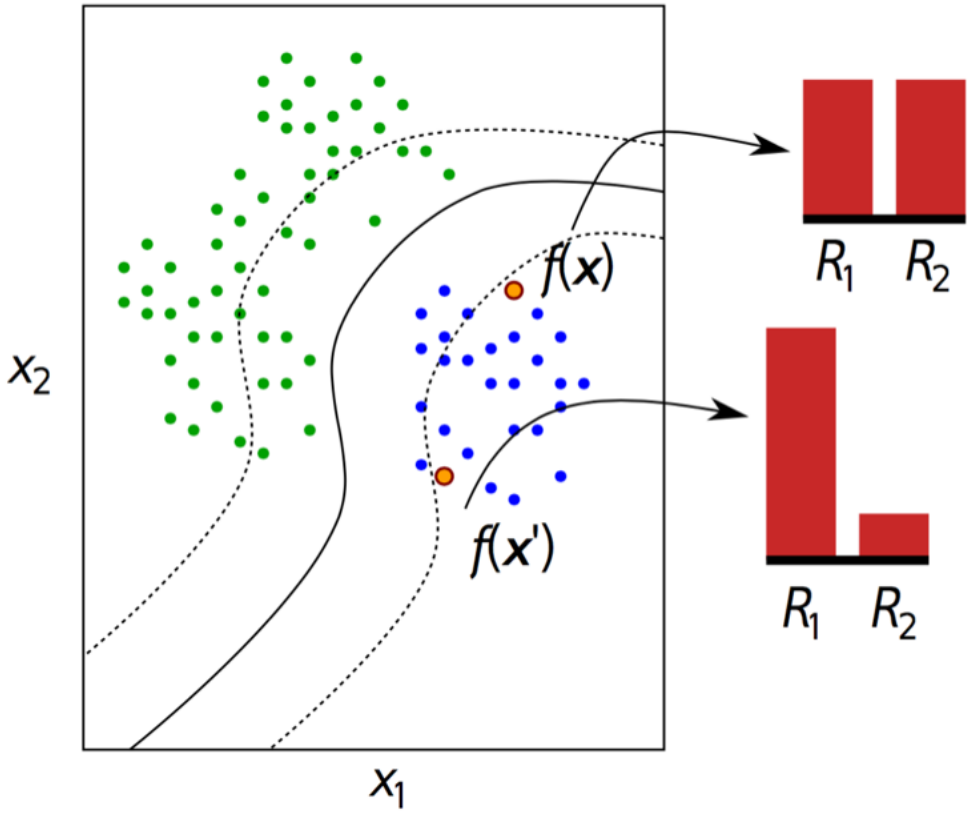
red color: evidence for prediction

blue color: evidence against prediction

Interpretation by decomposition

Explaining Individual Decisions

Goal: Determine the relevance of each input variable for a given decision $f(x_1, x_2, \ldots, x_d)$, by assigning to these variables *relevance* scores $R_1, R_2, \ldots, R_d$.



Basic Technique: Sensitivity Analysis

Consider a function f, a data point $\boldsymbol{x} = (x_1, \ldots, x_d)$, and the prediction

$$f(x_1, \ldots, x_d).$$

Sensitivity analysis measures the local variation of the function along each input dimension

$$R_i = \left(\frac{\partial f}{\partial x_i} \Big|_{\boldsymbol{x}=\boldsymbol{x}} \right)^2$$

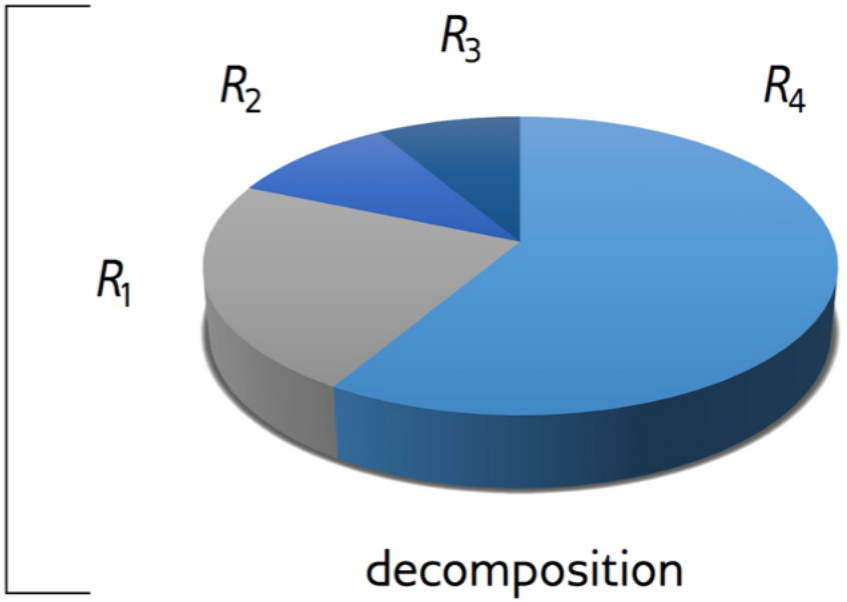
Remarks:

- Easy to implement (we only need access to the gradient of the decision function).
- But does it really explain the prediction?

Explaining by Decomposing

aggregate quantity $f(\boldsymbol{x}) =$

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



Examples:

- Economic activity (e.g. petroleum, cars, medicaments, ...)
- Energy production (e.g. coal, nuclear, hydraulic, ...)
- Evidence for object in an image (e.g. pixel 1, pixel 2, pixel 3, ...)
- Evidence for meaning in a text (e.g. word 1, word 2, word 3, ...)

What Does Sensitivity Analysis Decompose?

Sensitivity analysis

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

is a decomposition of the gradient norm $\|\nabla_x f\|^2$.

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

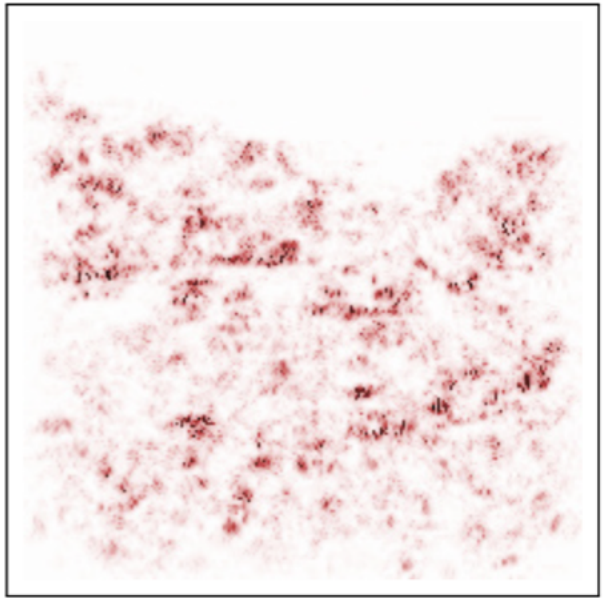
Sensitivity analysis explains a *variation* of the function, not the function value itself.

What Does Sensitivity Analysis Decompose?

Example: Sensitivity for class "car"

input image

sensitivity



- Relevant pixels are found both on cars and on the background.
- Explains what *reduces/increases* the evidence for cars rather what *is* the evidence for cars.

Decomposing the Correct Quantity

slope decomposition

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

$\rightarrow$

value decomposition

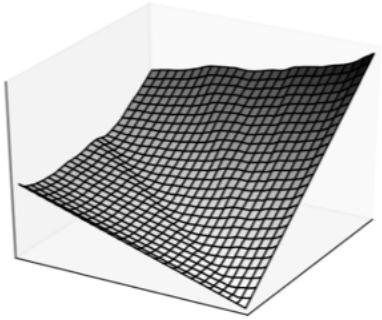
$$\sum_i R_i = f(\boldsymbol{x})$$

Candidate: Taylor decomposition

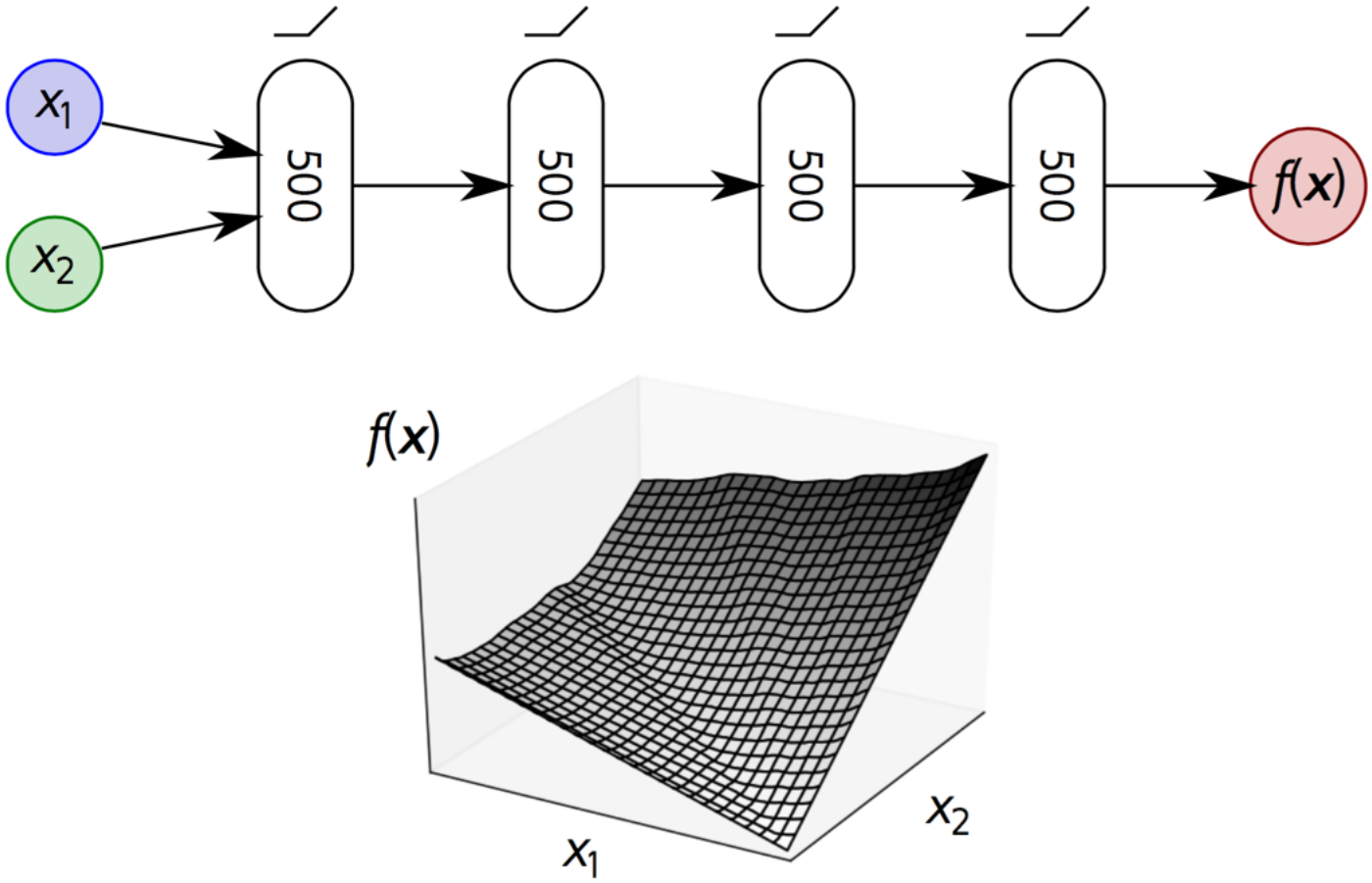
$$f(\boldsymbol{x}) = \underbrace{f(\widetilde{\boldsymbol{x}})}_{0} + \sum_{i=1}^{d} \underbrace{\frac{\partial f}{\partial x_i}\Big|_{\boldsymbol{x}=\widetilde{\boldsymbol{x}}} (x_i - \widetilde{x}_i)}_{R_i} + \underbrace{O(\boldsymbol{x}\boldsymbol{x}^\top)}_{0}$$

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

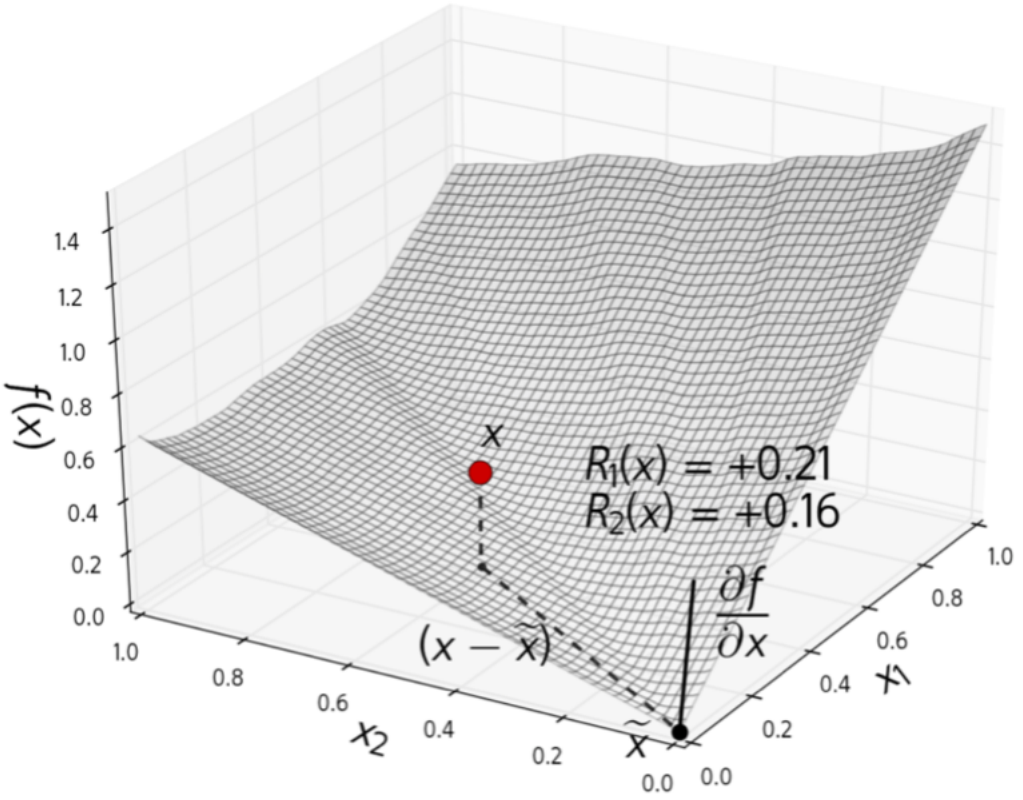
$$\widetilde{\boldsymbol{x}} = \lim_{\varepsilon \to 0} \varepsilon \cdot \boldsymbol{x} \approx \boldsymbol{0}.$$



Experiment on a Randomly Initialized DNN

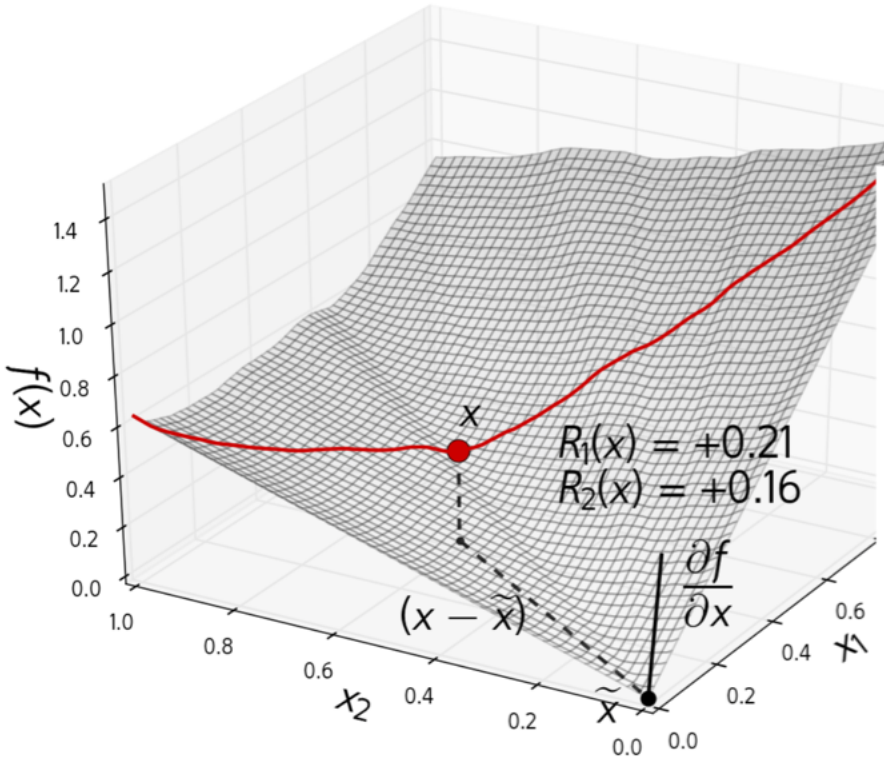


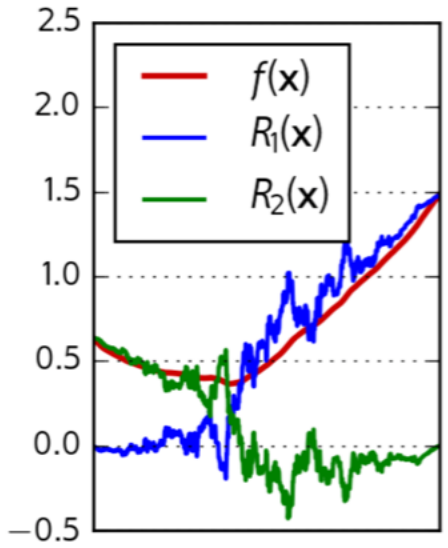
Decomposing the Output of the DNN



$$R_i = \frac{\partial f}{\partial x_i}\Big|_{\boldsymbol{x}=\tilde{\boldsymbol{x}}} \cdot (x_i - \tilde{x}_i)$$

Decomposing the Output of the DNN

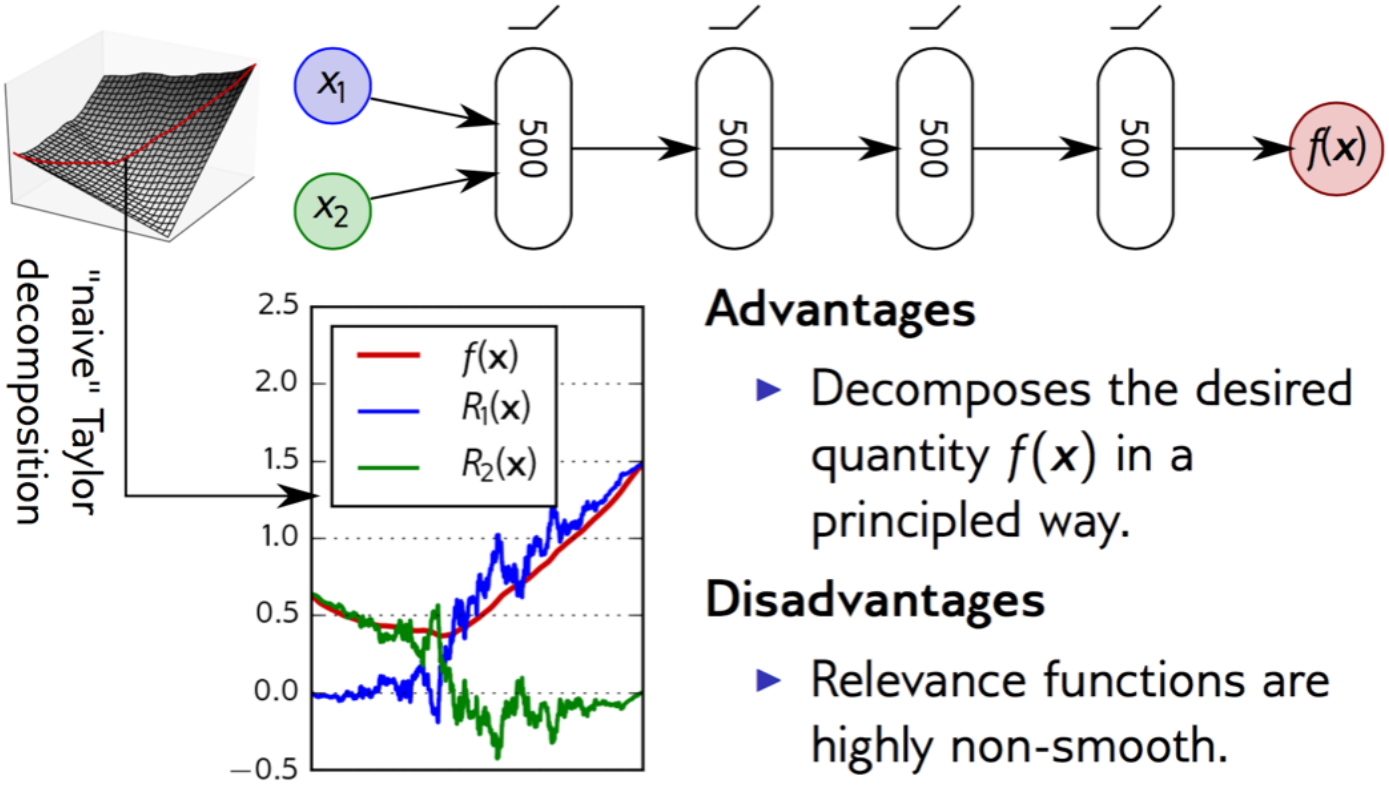




$$R_i = \frac{\partial f}{\partial x_i}\Big|_{\boldsymbol{x}=\tilde{\boldsymbol{x}}} \cdot (x_i - \tilde{x}_i)$$

⇒ "Naive" Taylor decomposition

Decomposing the Output of the DNN



Advantages

- Decomposes the desired quantity $f(\boldsymbol{x})$ in a principled way.

Disadvantages

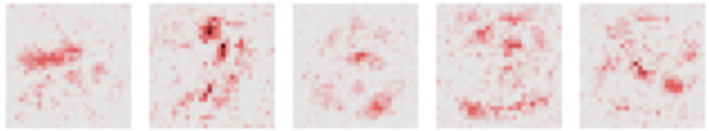
- Relevance functions are highly non-smooth.
- Relevance scores are sometimes negative.
- Inflexible w.r.t. the model.

Experiment on Handwritten Digits

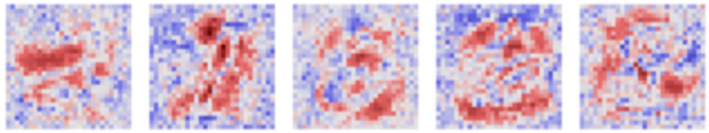
Data to classify:

3-layer MLP:

Sensitivity analysis



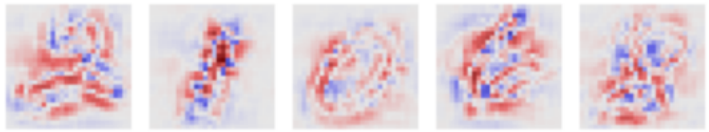
Naive Taylor ($\widetilde{\boldsymbol{x}} = 0$)



6-layer CNN:

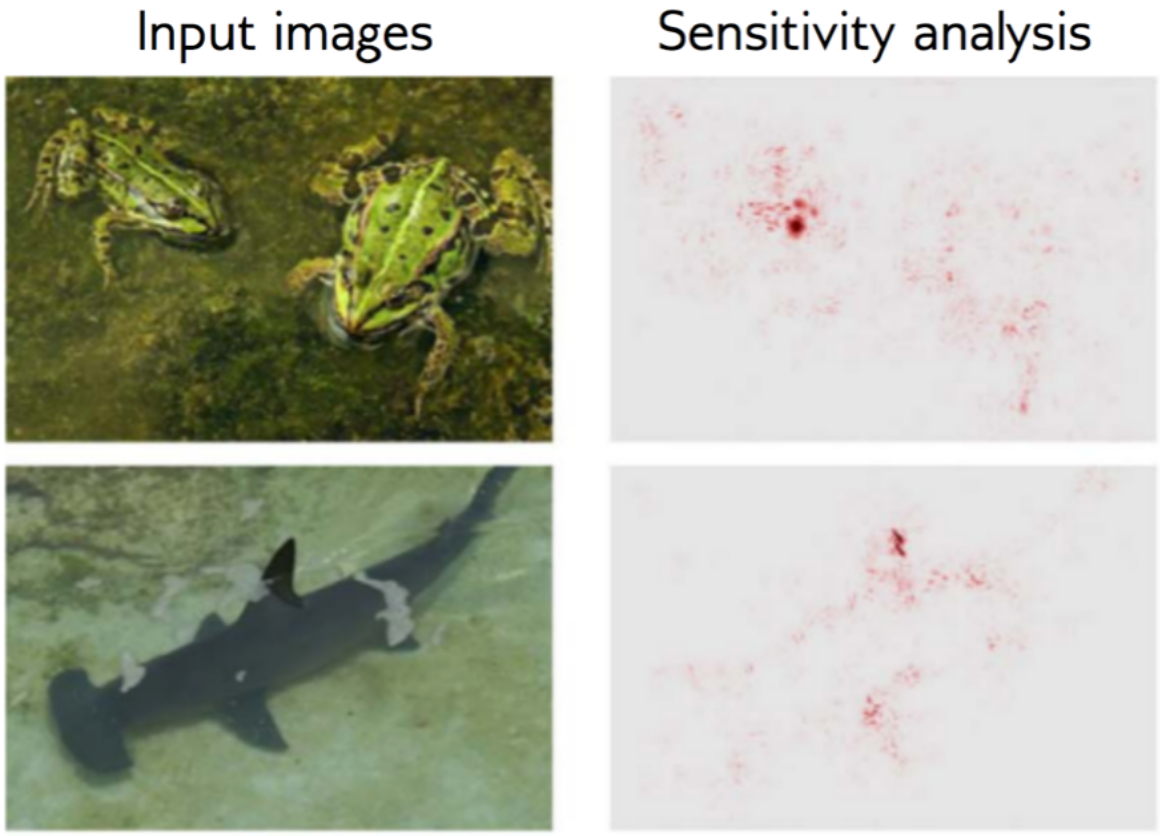
Sensitivity analysis

Naive Taylor ($\widetilde{\boldsymbol{x}} = 0$)



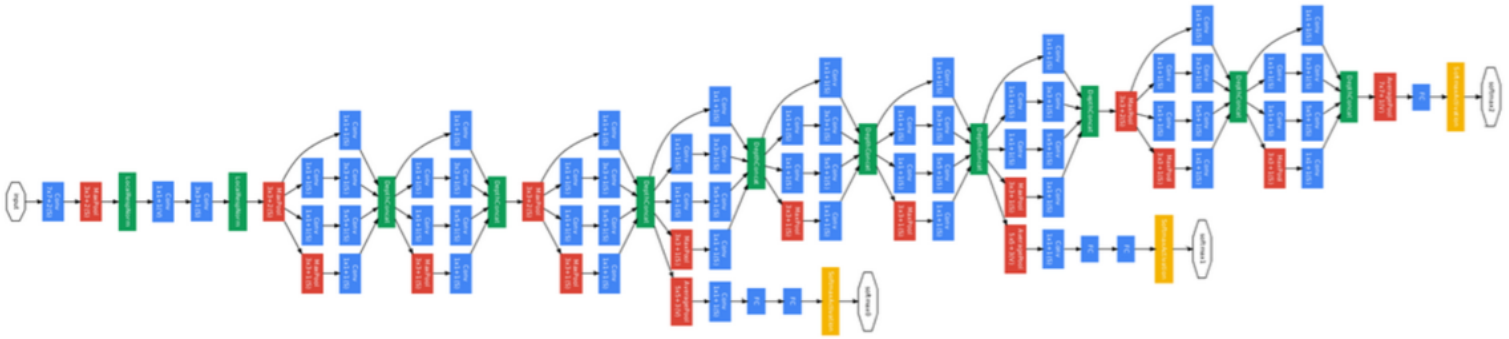
Observation: Both analyses produce noisy explanations of the MLP and CNN predictions.

Experiment on BVLC CaffeNet



Observation: Explanations are noisy and (over/under)represent certain regions of the image.

Explaining DNN Predictions



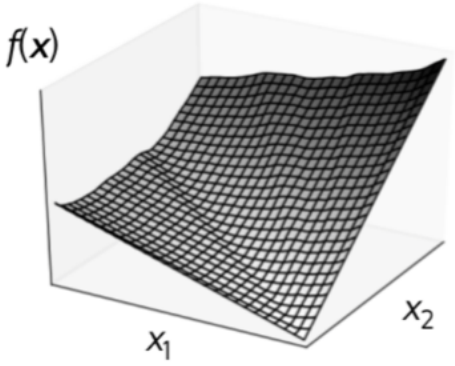
- Standard methods (sensitivity analysis, naive Taylor decomposition) are subject to gradient noise and do not work well on deep neural networks.

DNN predictions need more advanced explanation methods.

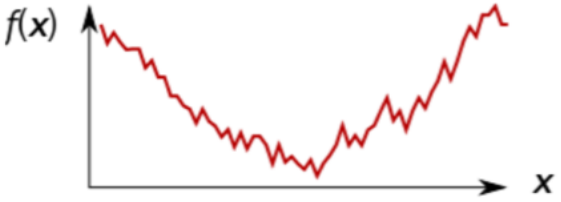
Why Simple Taylor doesn't work?

Two Reasons:

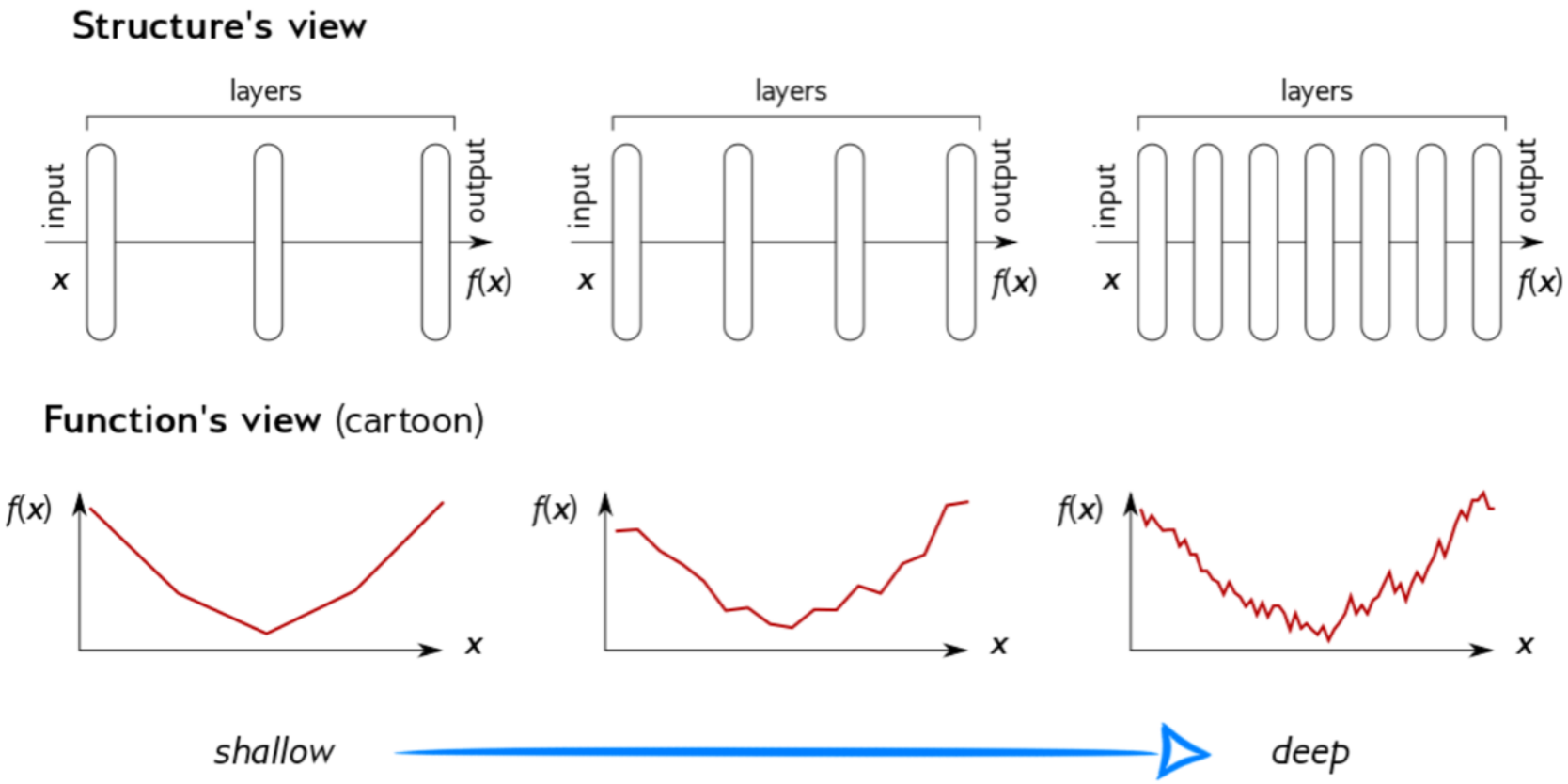
Root point is hard to find or too far → includes too much information (incl. negative evidence)



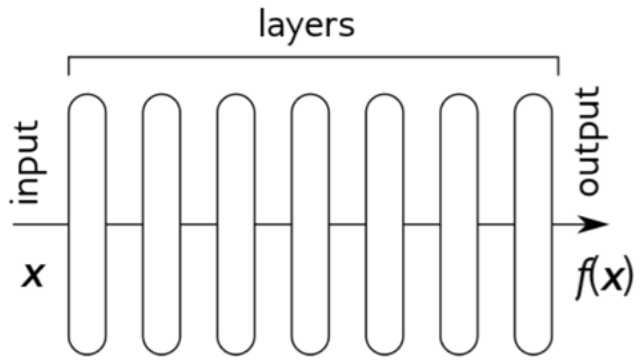
Gradient shattering problem → gradient of deep nets has low informative value

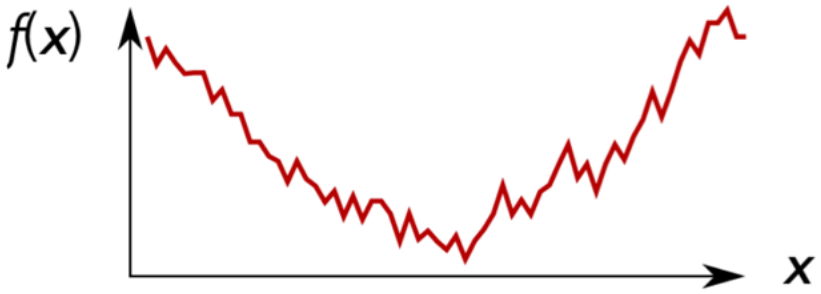


Gradient Shattering



Sensitivity and Simple Taylor

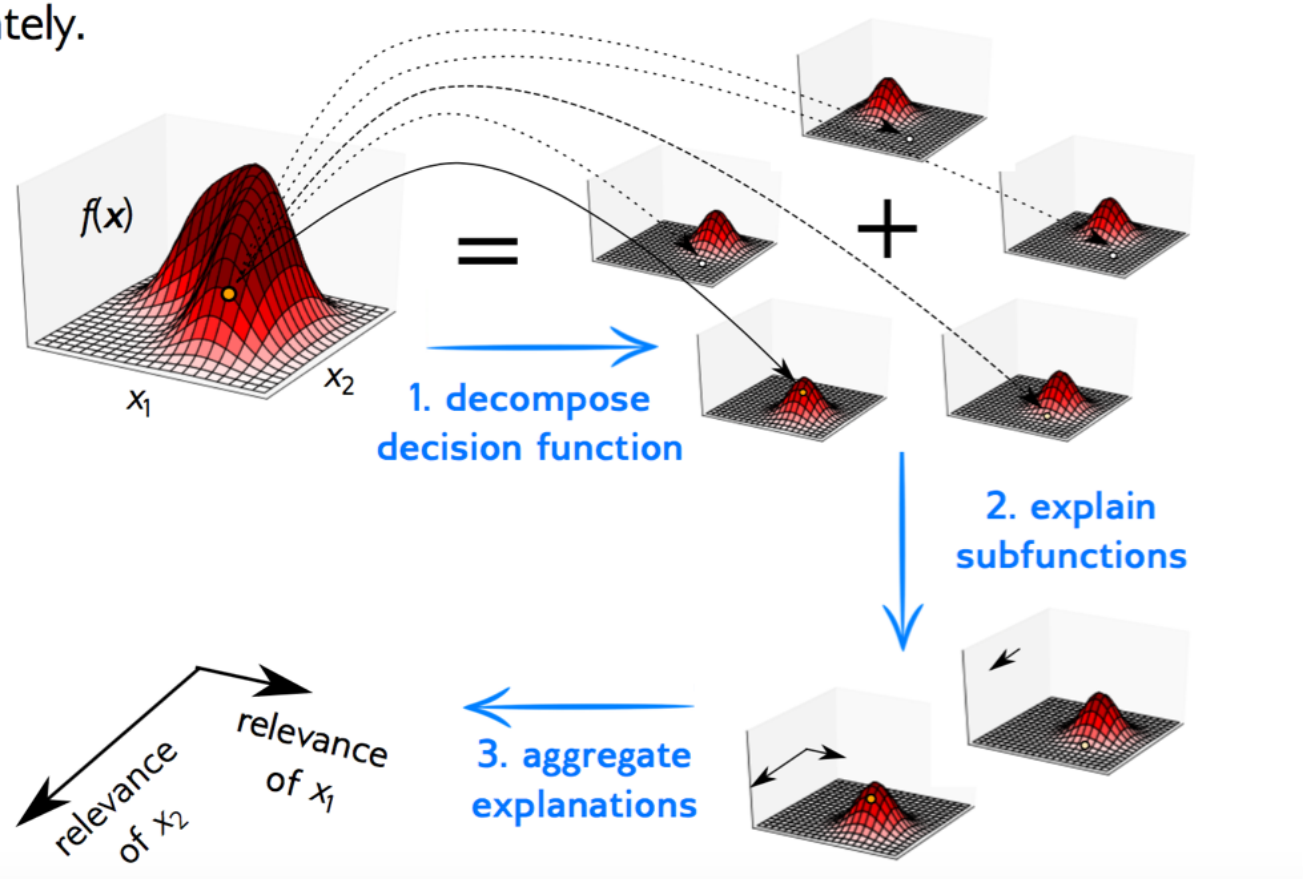




Gradient-based methods usually do not work for explaining deep nets.

Deep Taylor Decomposition

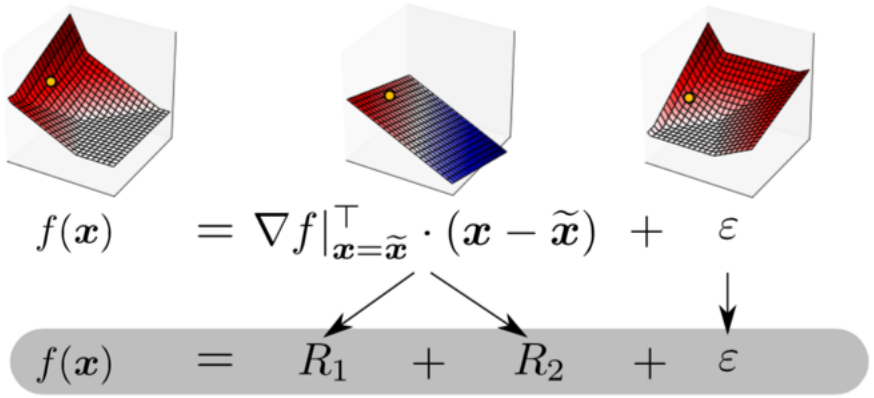
Key Idea: If a decision is too complex to explain, break the decision function into sub-functions, and explain each sub-decision separately.



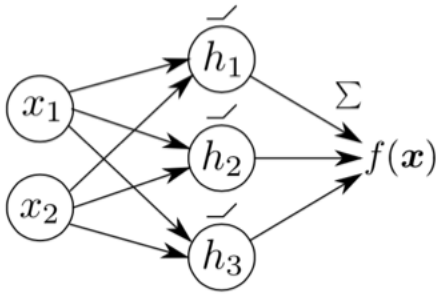
Deep Taylor Decomposition

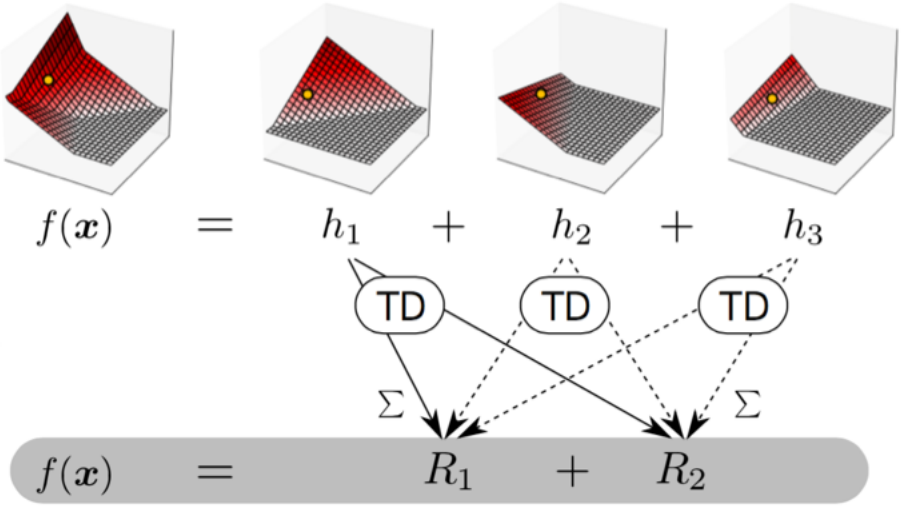
Taylor decomposition (TD)

$$f(\boldsymbol{x}), \nabla f, \ldots$$

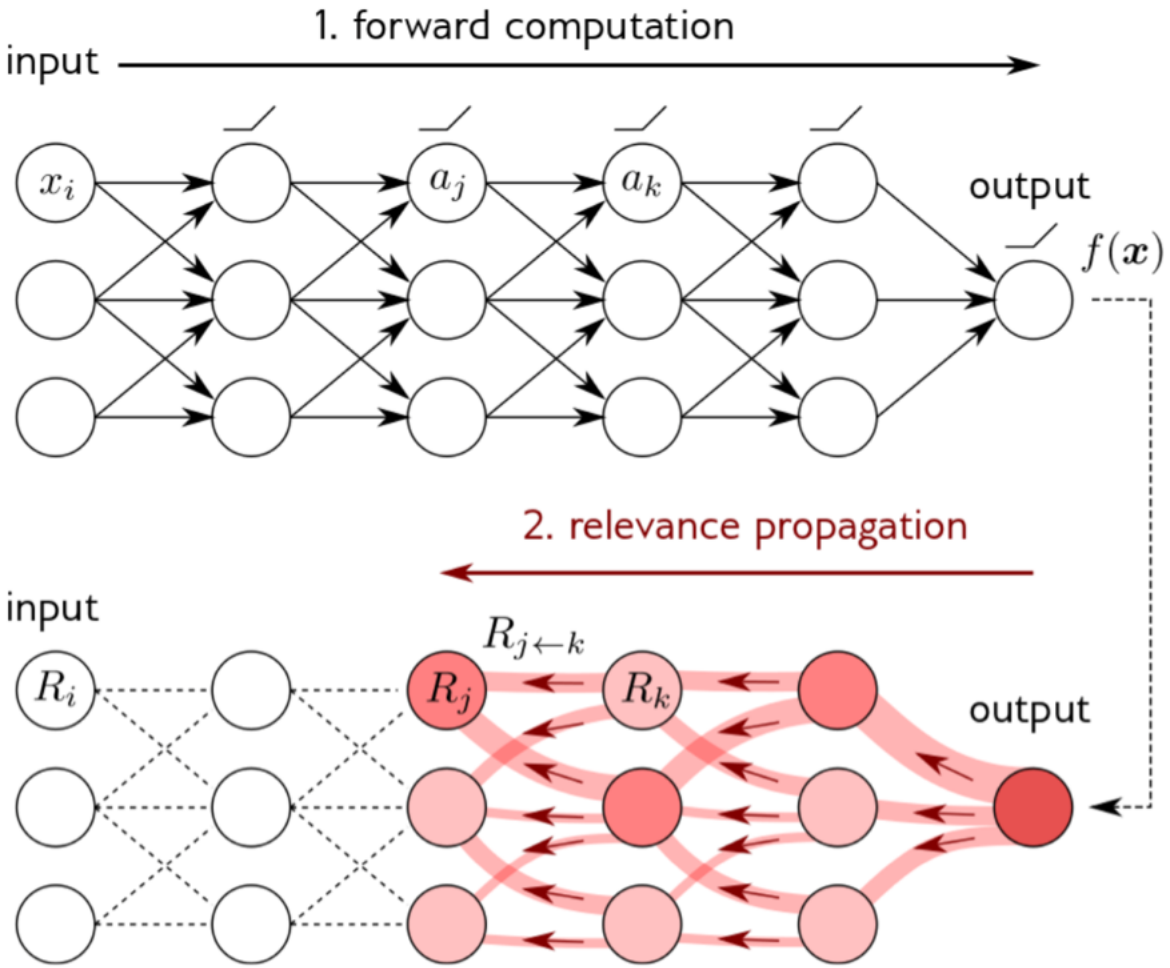


deep Taylor decomposition (DTD)

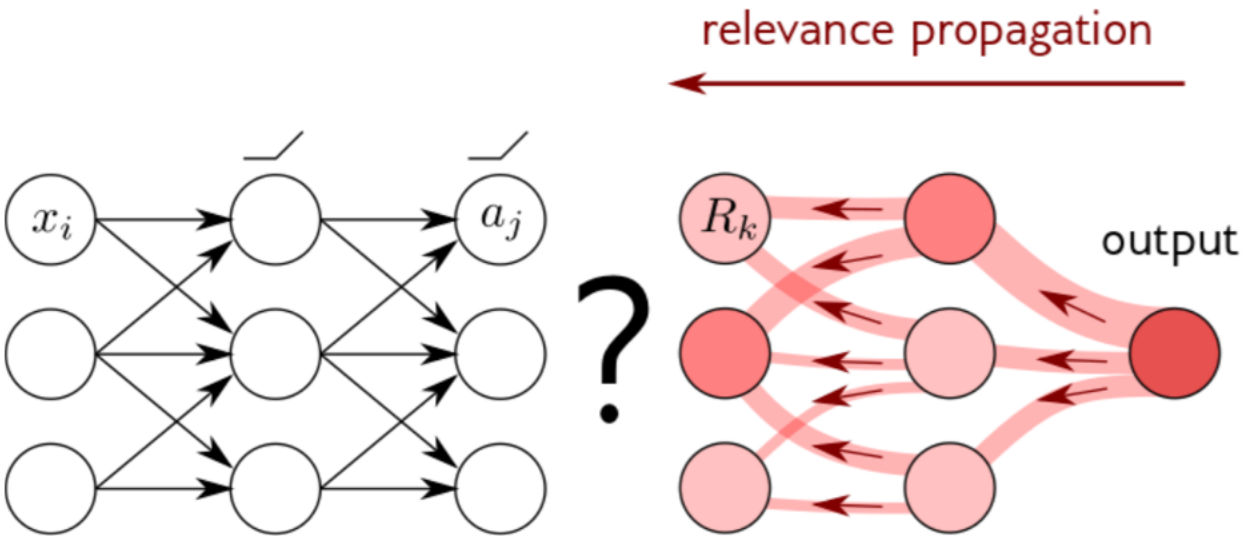




Deep Taylor Decomposition



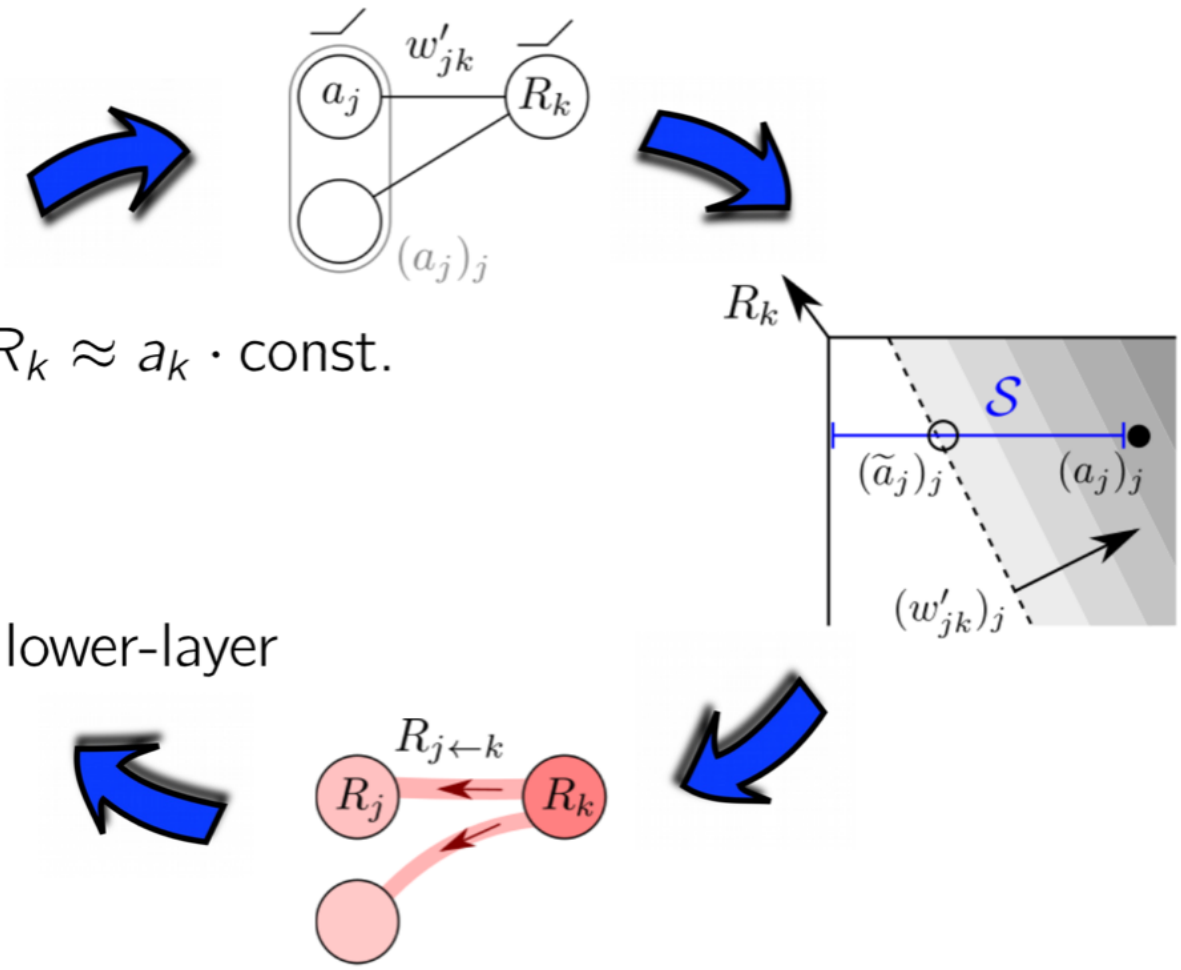
Deep Taylor Decomposition



Can we express R_k as a simple function of $(a_j)_j$?

Can we do a Taylor decomposition of $R_k((a_j)_j)$?

Deep Taylor Decomposition



Deep Taylor Decomposition / LRP

$$R_j = \sum_k \left(\alpha \frac{a_j w_{jk}^+}{\sum_j a_j w_{jk}^+} - \beta \frac{a_j w_{jk}^-}{\sum_j a_j w_{jk}^-} \right) R_k$$

intuition

[Bach'15]

analysis

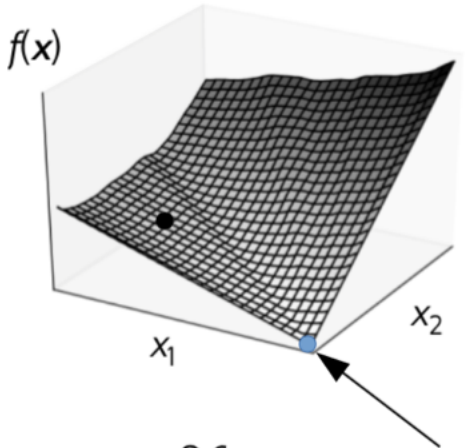
[Montavon'17]

Relevance should be redistributed to the lower-layer neurons $(a_j)_j$ in proportion to their excitatory effect on a_k. "Counter-relevance" should be redistributed to the lower-layer neurons $(a_j)_j$ in proportion to their inhibitory effect on a_k.

For the specific case $\alpha = 1$, the whole LRP procedure can be seen as a *deep Taylor decomposition* of the neural network function.

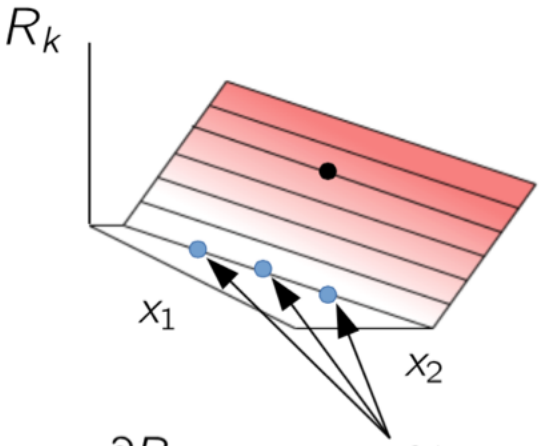
Deep Taylor Decomposition / LRP

Simple Taylor



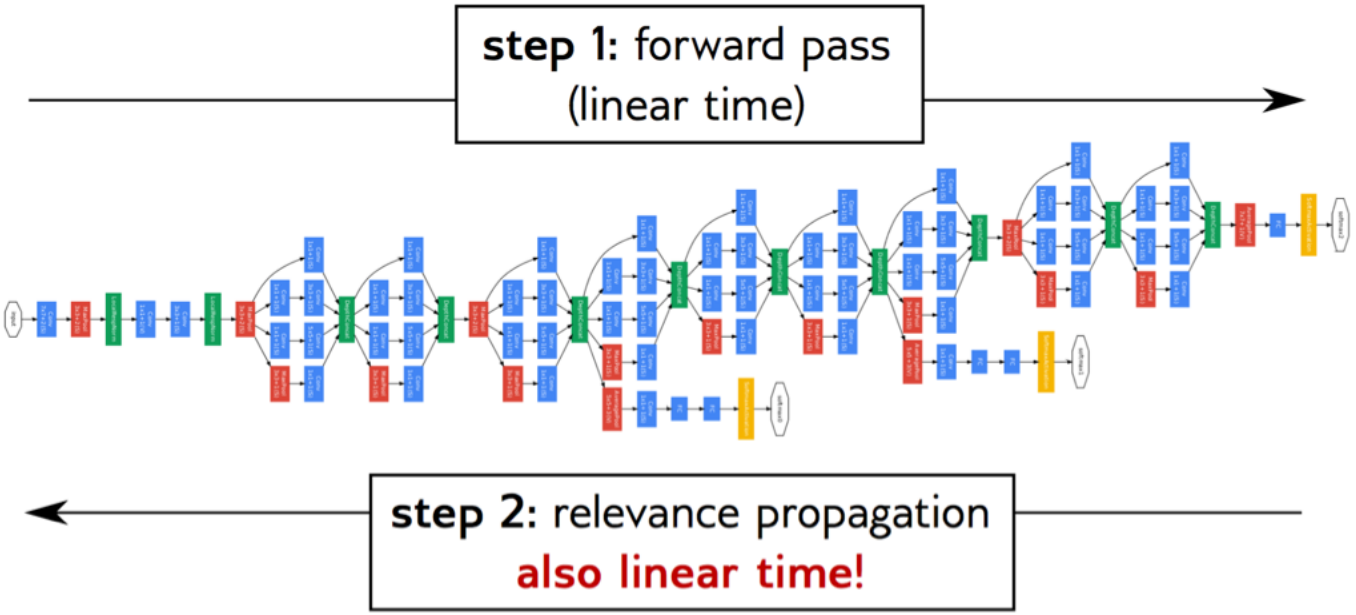
$$R_i = \frac{\partial f}{\partial x_i} \cdot (x_i - 0)$$

Deep Taylor



$$R_i = \frac{\partial R_k}{\partial a_j} \cdot (a_j - \widetilde{a}_j)$$

Deep Taylor Decomposition / LRP

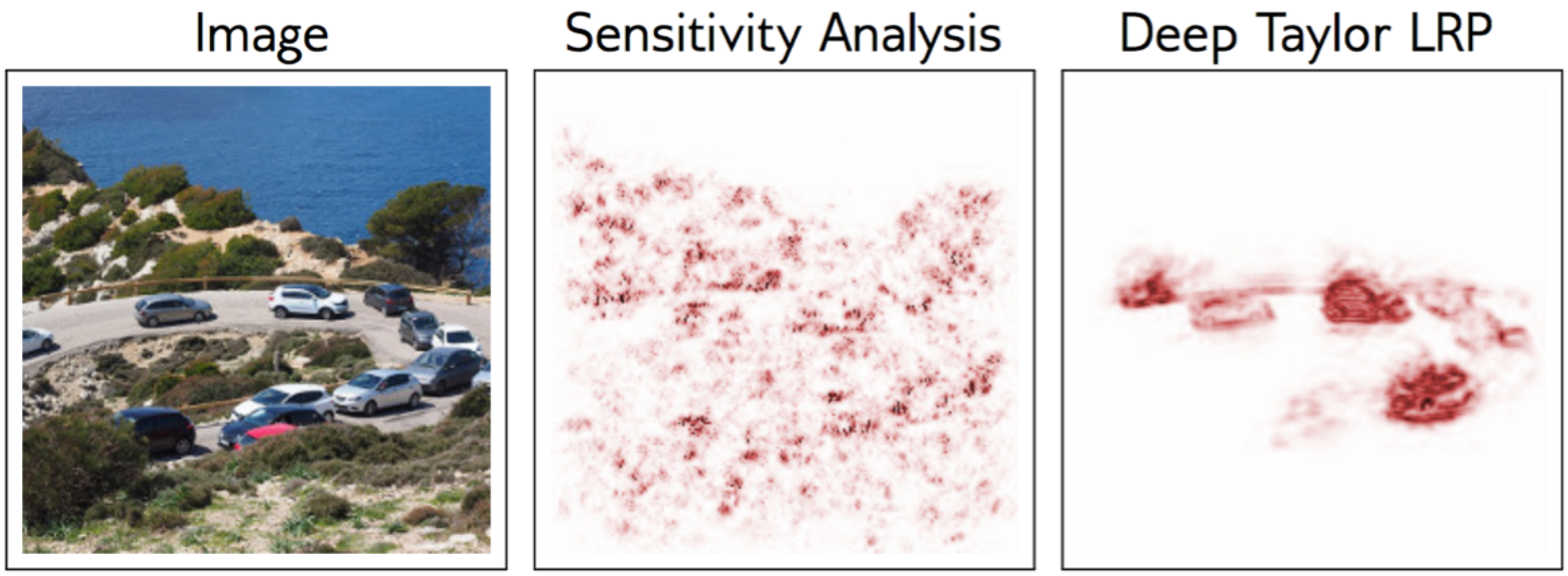


Propagation rule:

$$R_i = \sum_j q_{ij} R_j \qquad \sum_i q_{ij} = 1$$

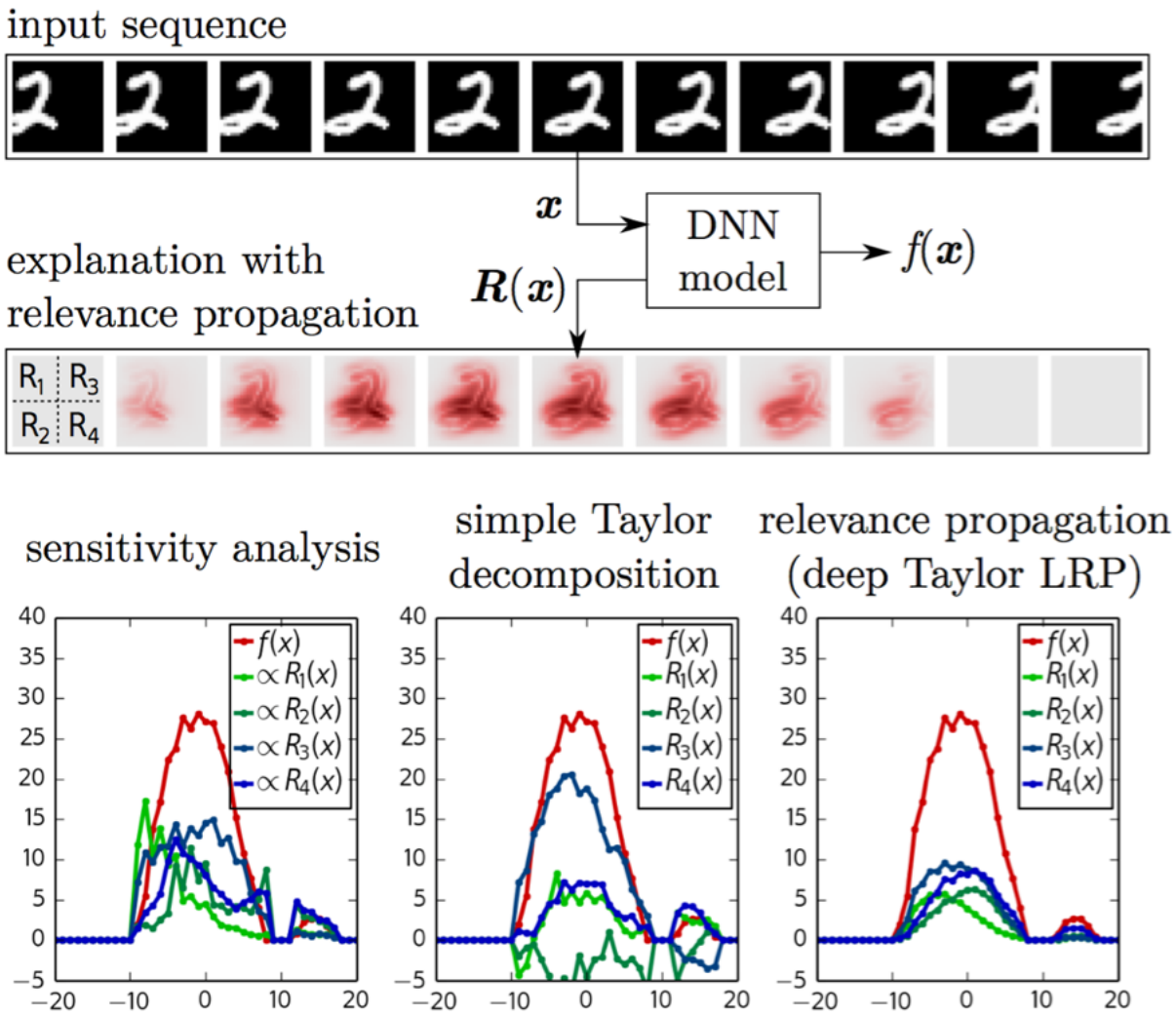
Various rules are available for pixel layers, intermediate layers, or special layers.

Deep Taylor Decomposition / LRP



Observation: Only deep Taylor LRP focuses on cars.

Deep Taylor Decomposition / LRP

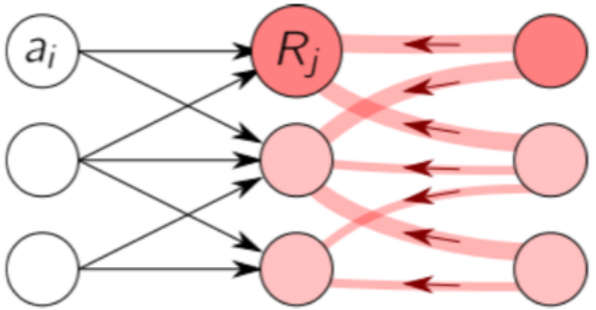


Two similar images should have similar explanations.

Technical Details

$$R_i = \sum_j \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+} R_j$$

$$R_j = a_j c_j$$



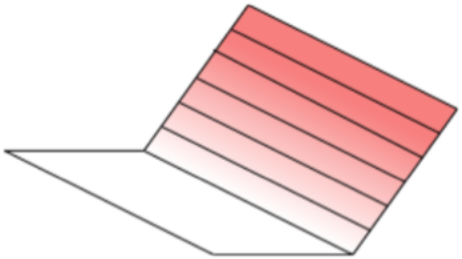
$$R_i = a_i \sum_j w_{ij}^+ \frac{\max(0, \sum_i a_i w_{ij})}{\sum_i a_i w_{ij}^+} c_j$$

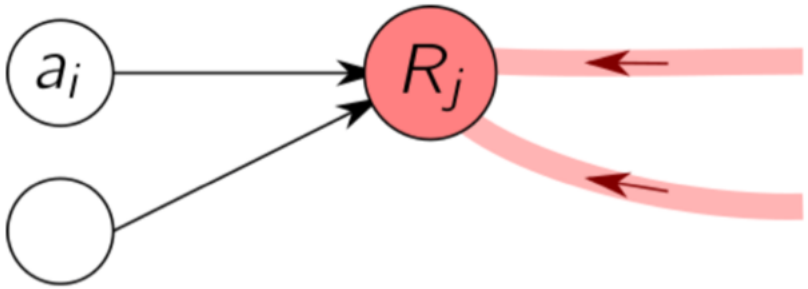
$$R_i = a_i c_i$$

Relevance has product structure at all layers.

Technical Details

Build the Relevance Neuron



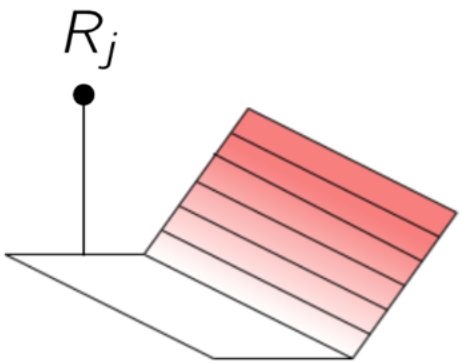


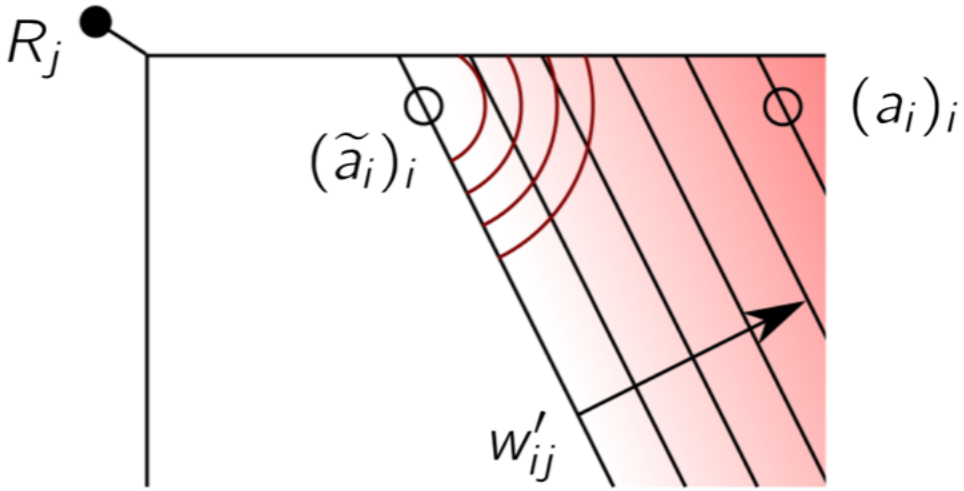
$$R_j = a_j c_j$$
$$= \max(0, \sum_i a_i w_{ij}) \cdot c_j$$
$$= \max(0, \sum_i a_i w'_{ij}) \qquad w'_{ij} = w_{ij} c_j$$

Technical Details

Expand the Relevance Neuron

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

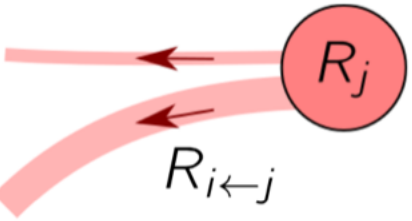


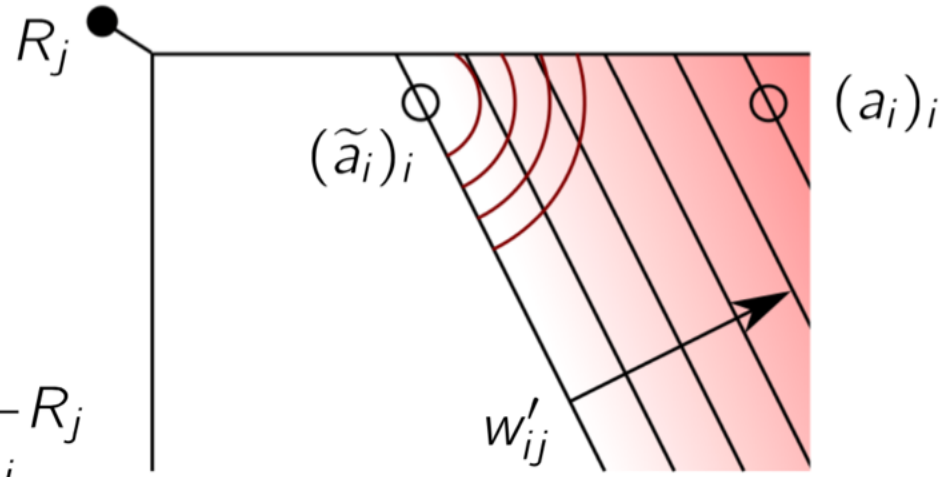


Technical Details

Decompose Relevance

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





Closed-form solution:

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

Technical Details

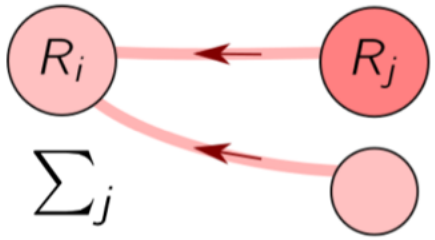
Closed-form solution:

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

LRP rule [Bach15, Zhang16]

$$R_i = \sum_j \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+} R_j$$

4 Pooling relevance over all outgoing neurons



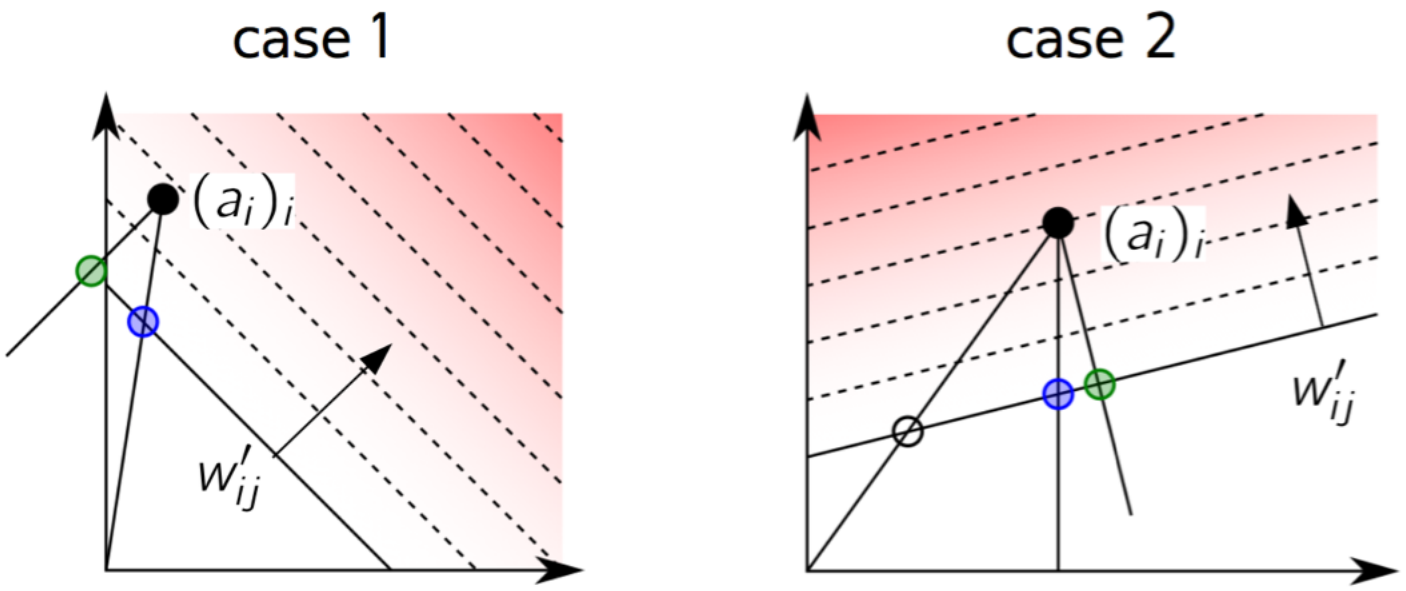
5 Search root point along specific direction:

$$(a_i - \widetilde{a}_i^{(j)}) \propto a_i 1_{w'_{ij} > 0}$$

But what does that mean?

18/22

Technical Details



- $(a_i - \widetilde{a}_i^{(j)}) \propto w'_{ij}$ — nearest root
- $(a_i - \widetilde{a}_i^{(j)}) \propto a_i$ — origin (like simple Taylor)
- $(a_i - \widetilde{a}_i^{(j)}) \propto a_i 1_{w'_{ij}>0}$ — root along positive activations (LRP rule)

19/22

Technical Details

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 $(x_i \in \mathbb{R})$	$R_i = \sum_j \frac{w_{ij}^2}{\sum_i w_{ij}^2} R_j$

Deep Taylor LRP rules [Montavon'17]

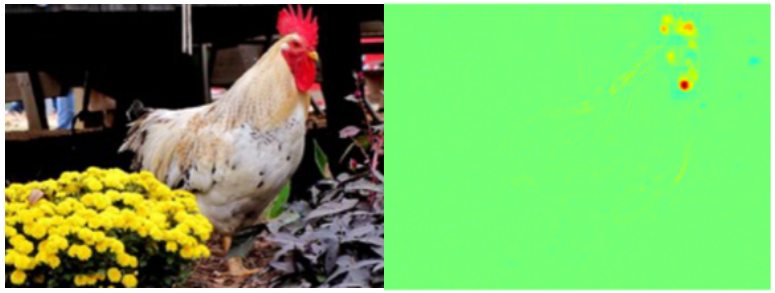
More refined rules can also be constructed to match the input data distribution [Kindermans'17]

How to measure the
quality of explanations ?

Measuring Quality of Explanations

Heatmap depends on

- classifier
- explanation method

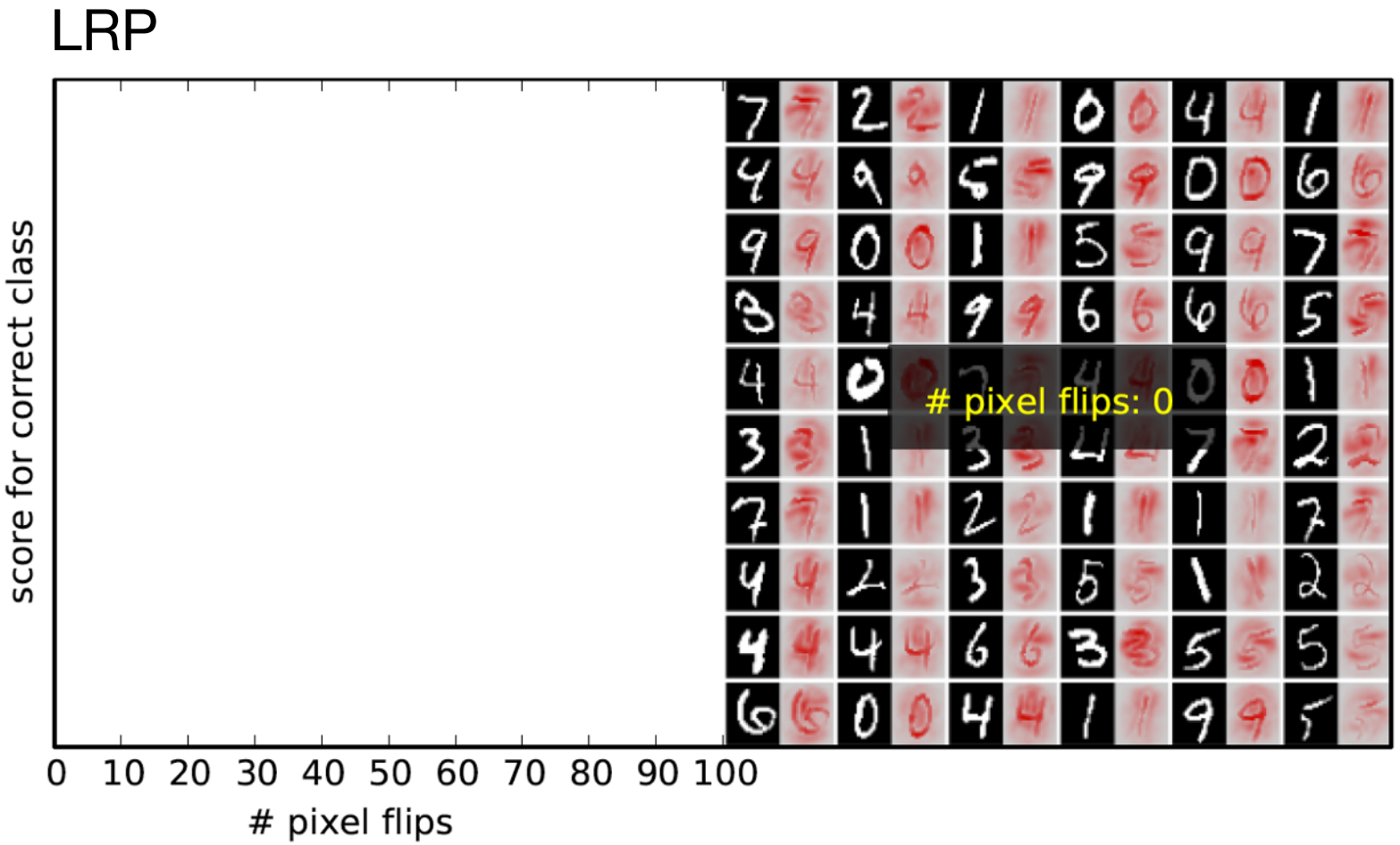


If we want to compare classifiers or explanations methods,
we need an *objective* measure of heatmap quality.

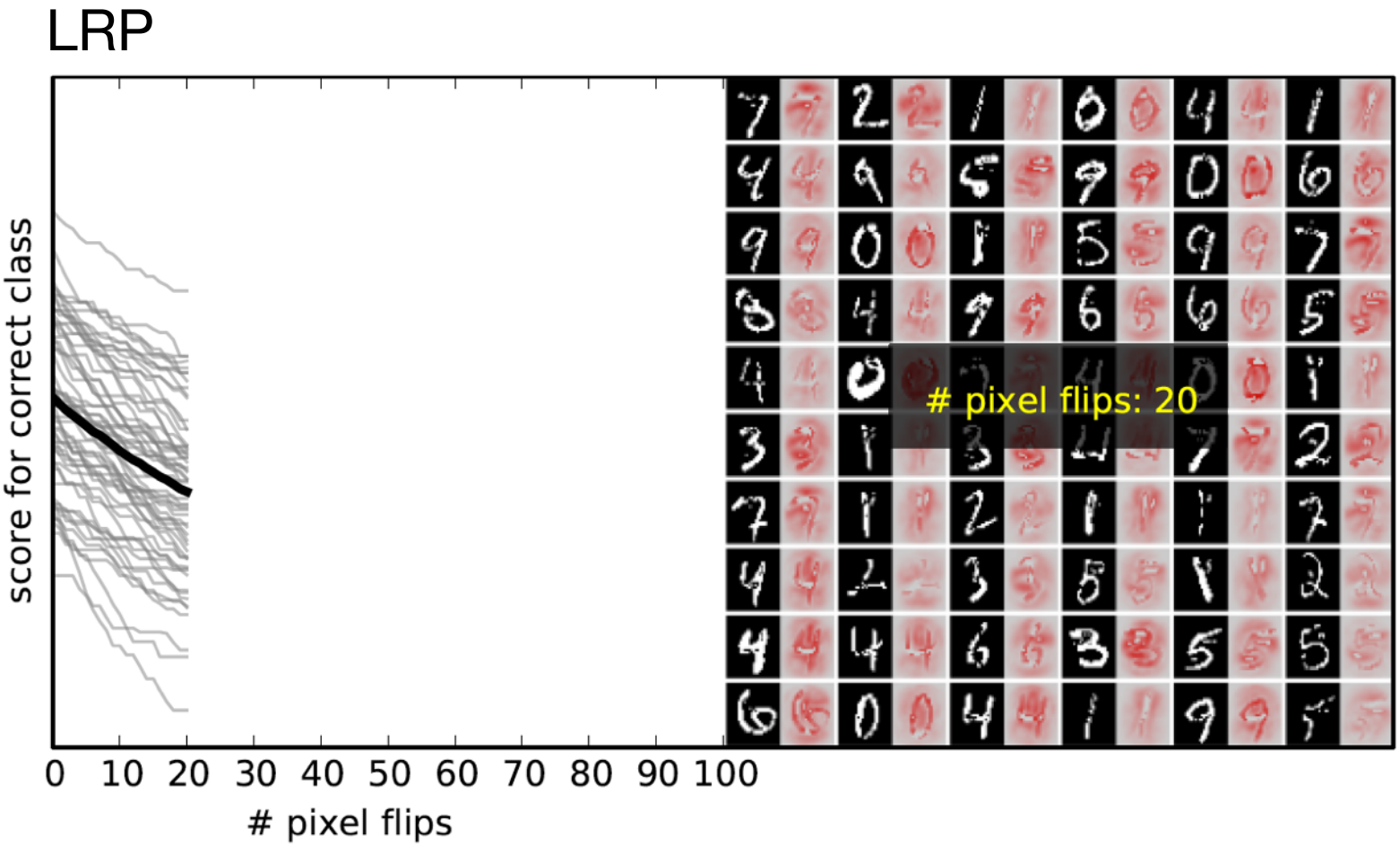
Algorithm (Pixel Flipping)

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

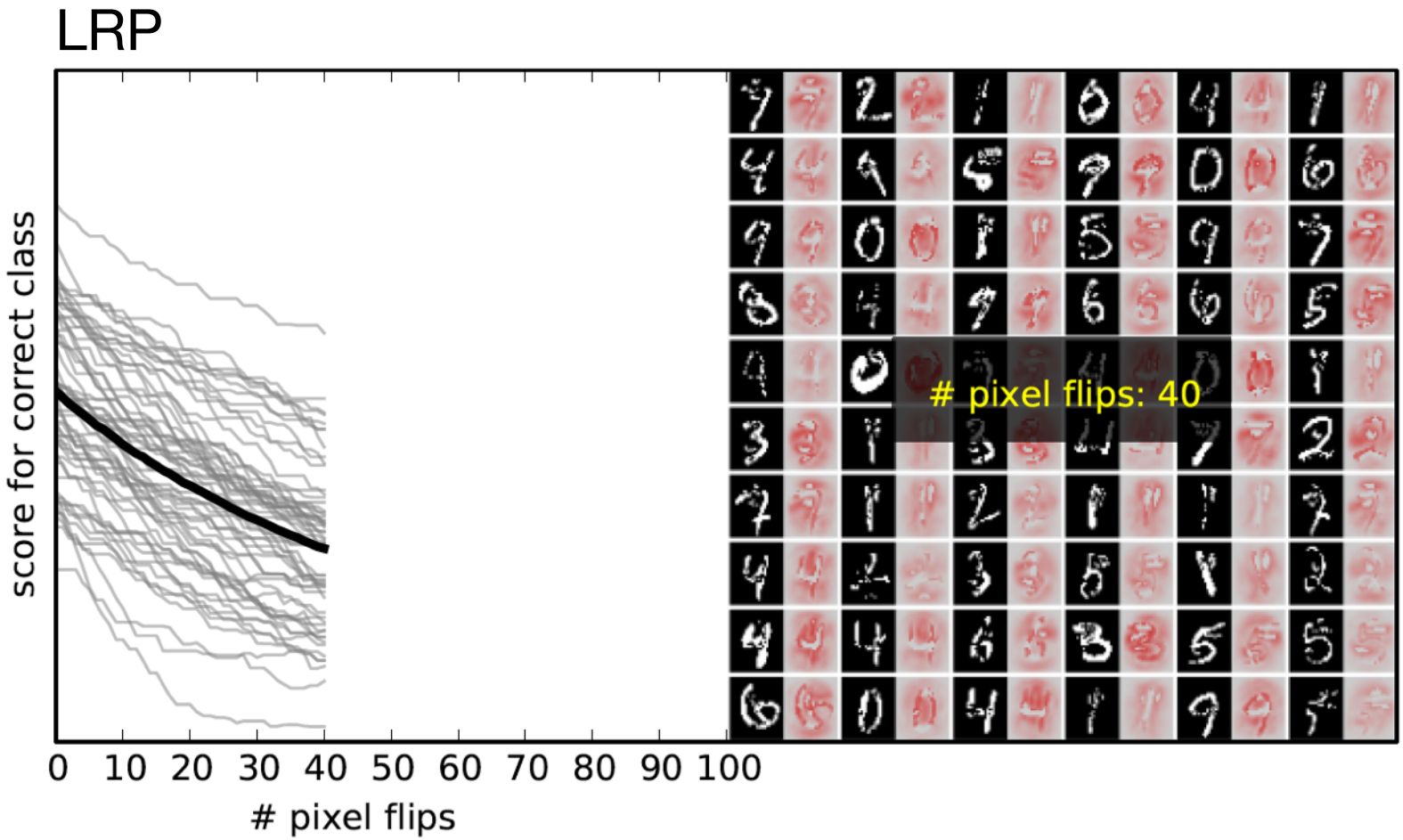
Compare Explanation Methods



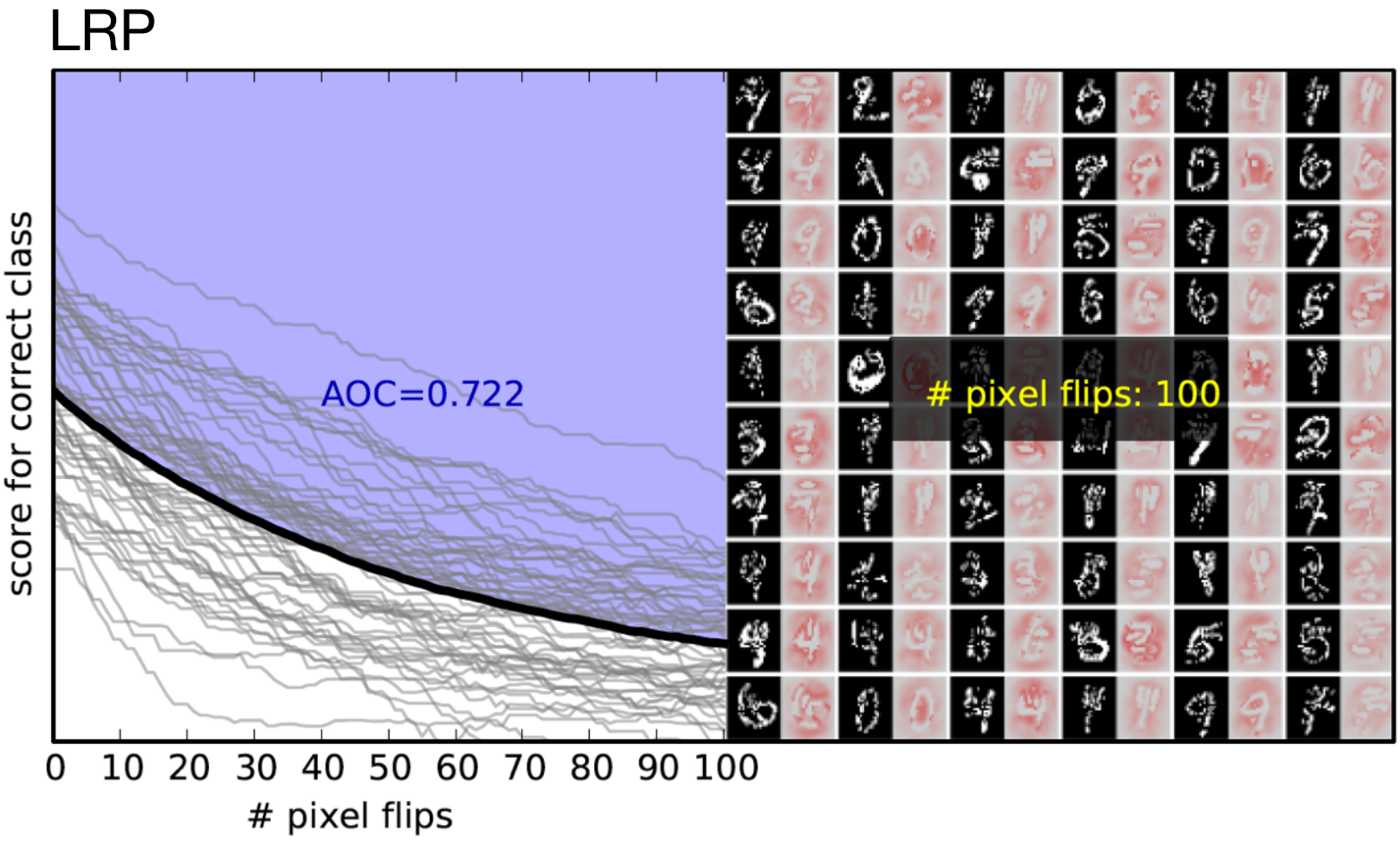
Compare Explanation Methods



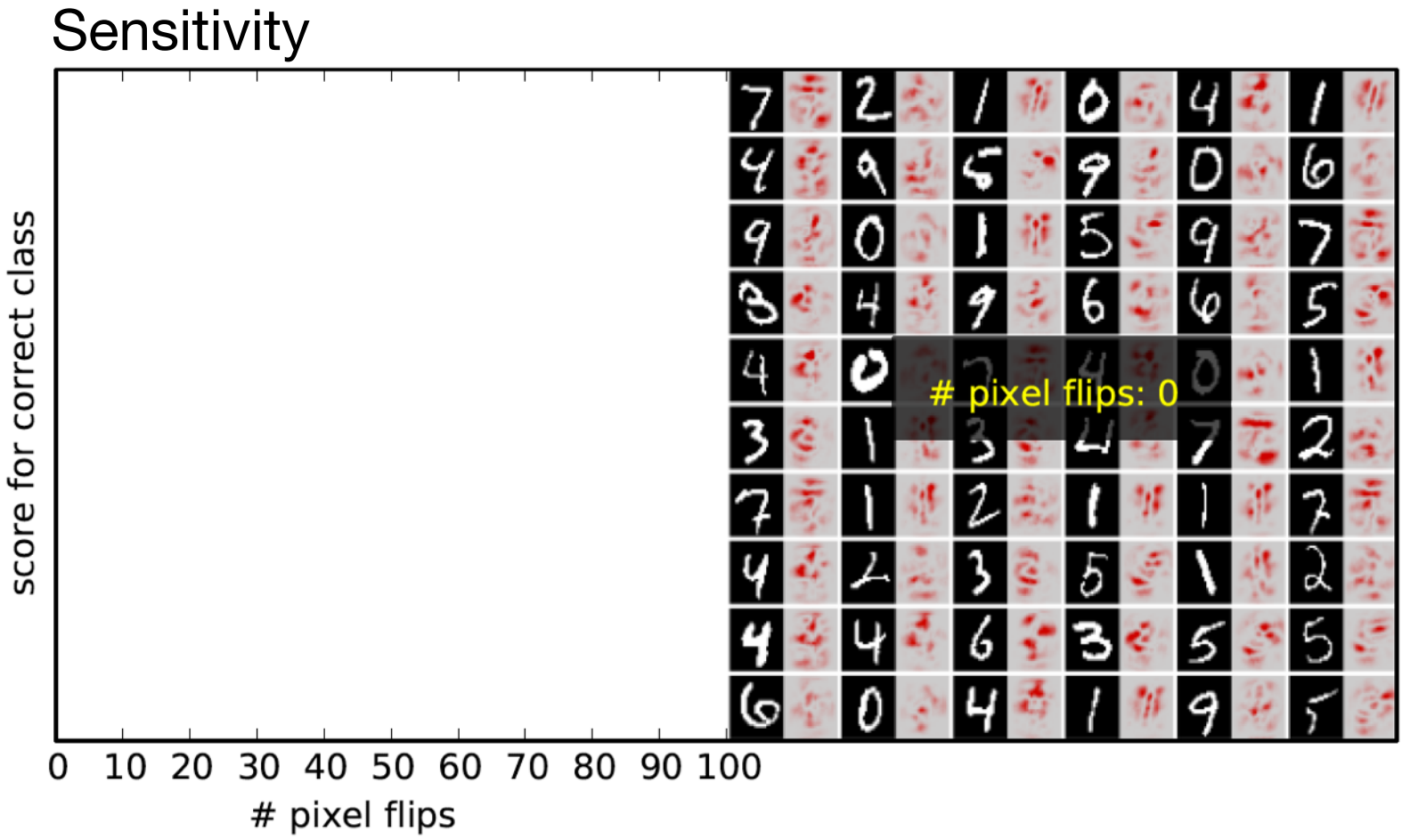
Compare Explanation Methods



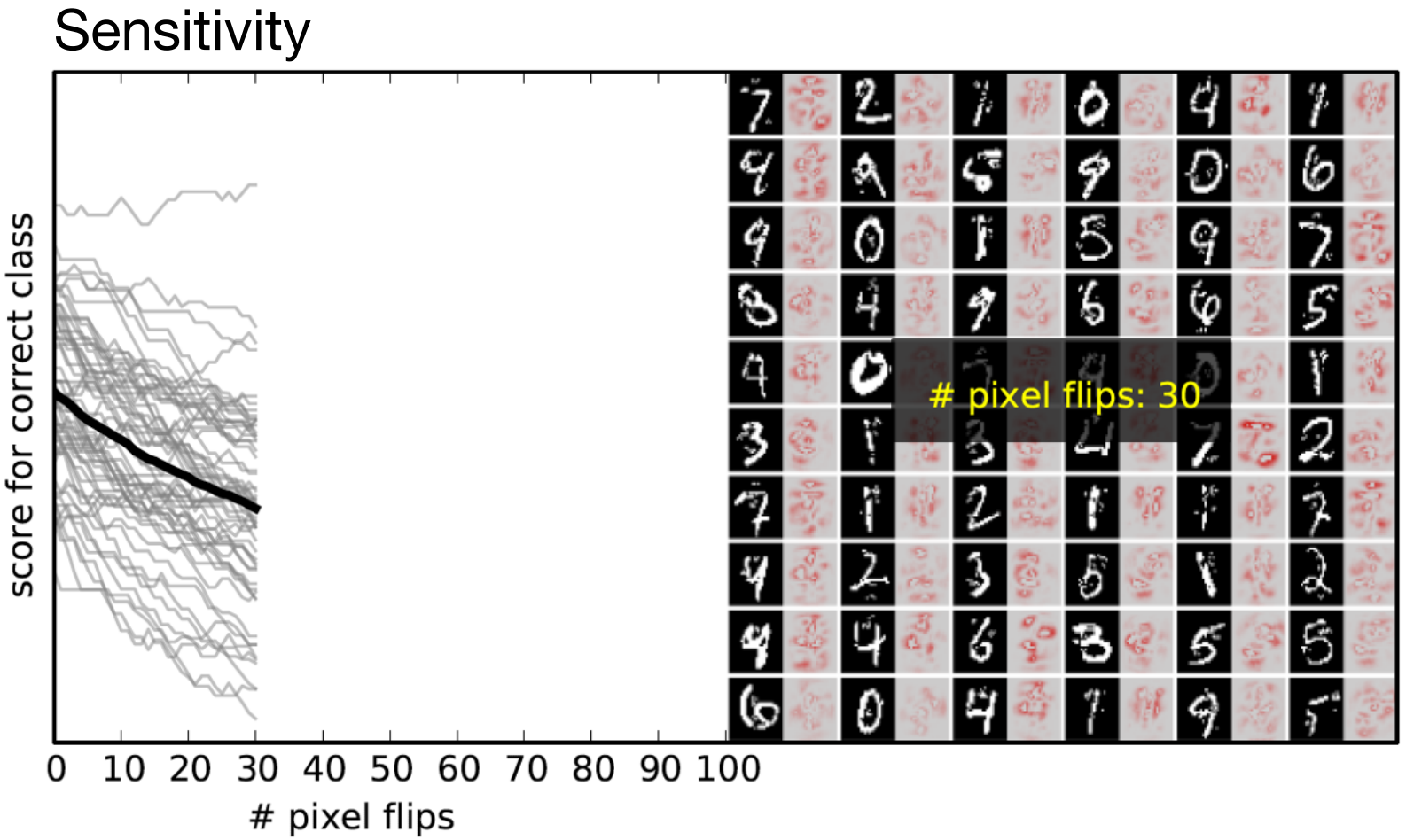
Compare Explanation Methods



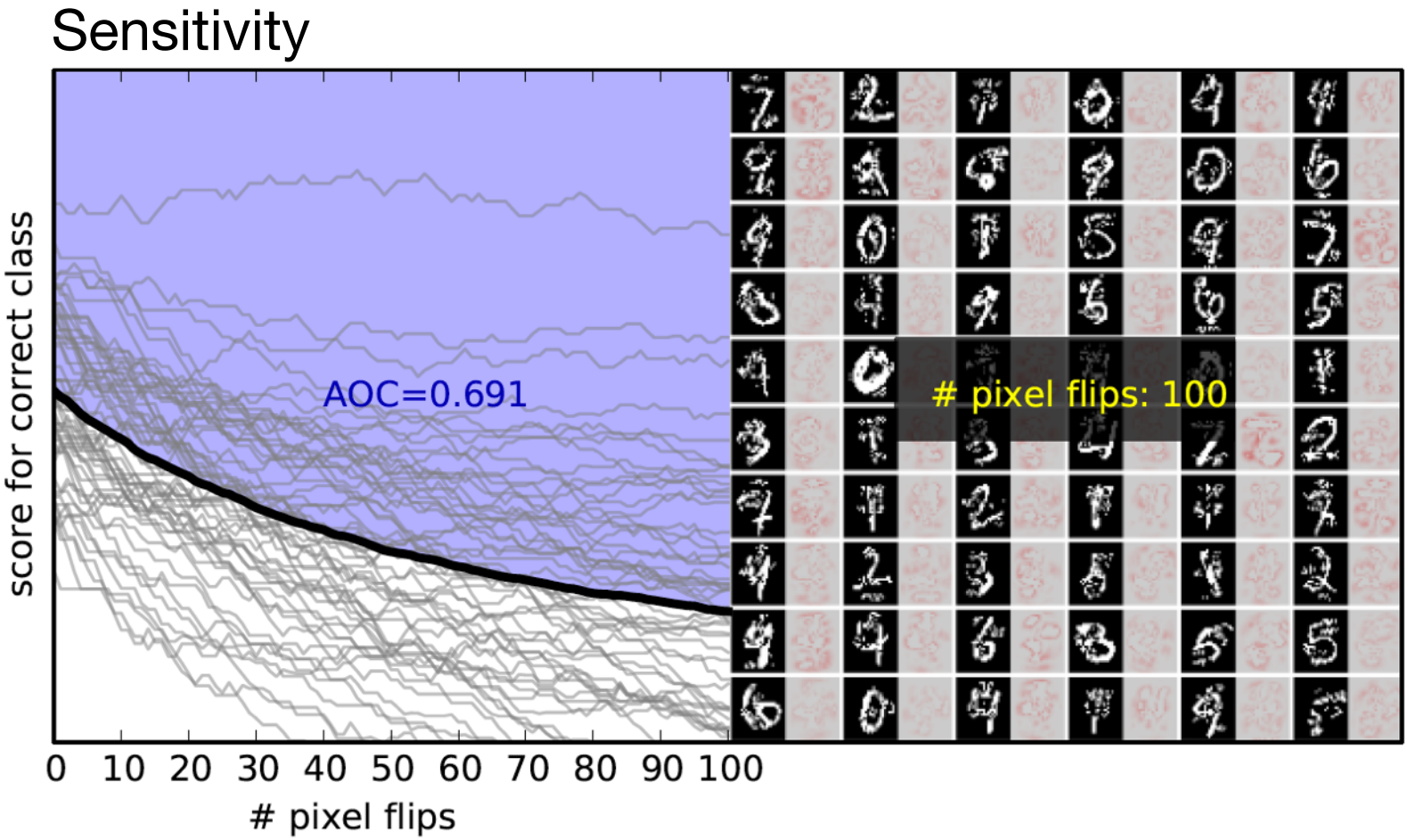
Compare Explanation Methods



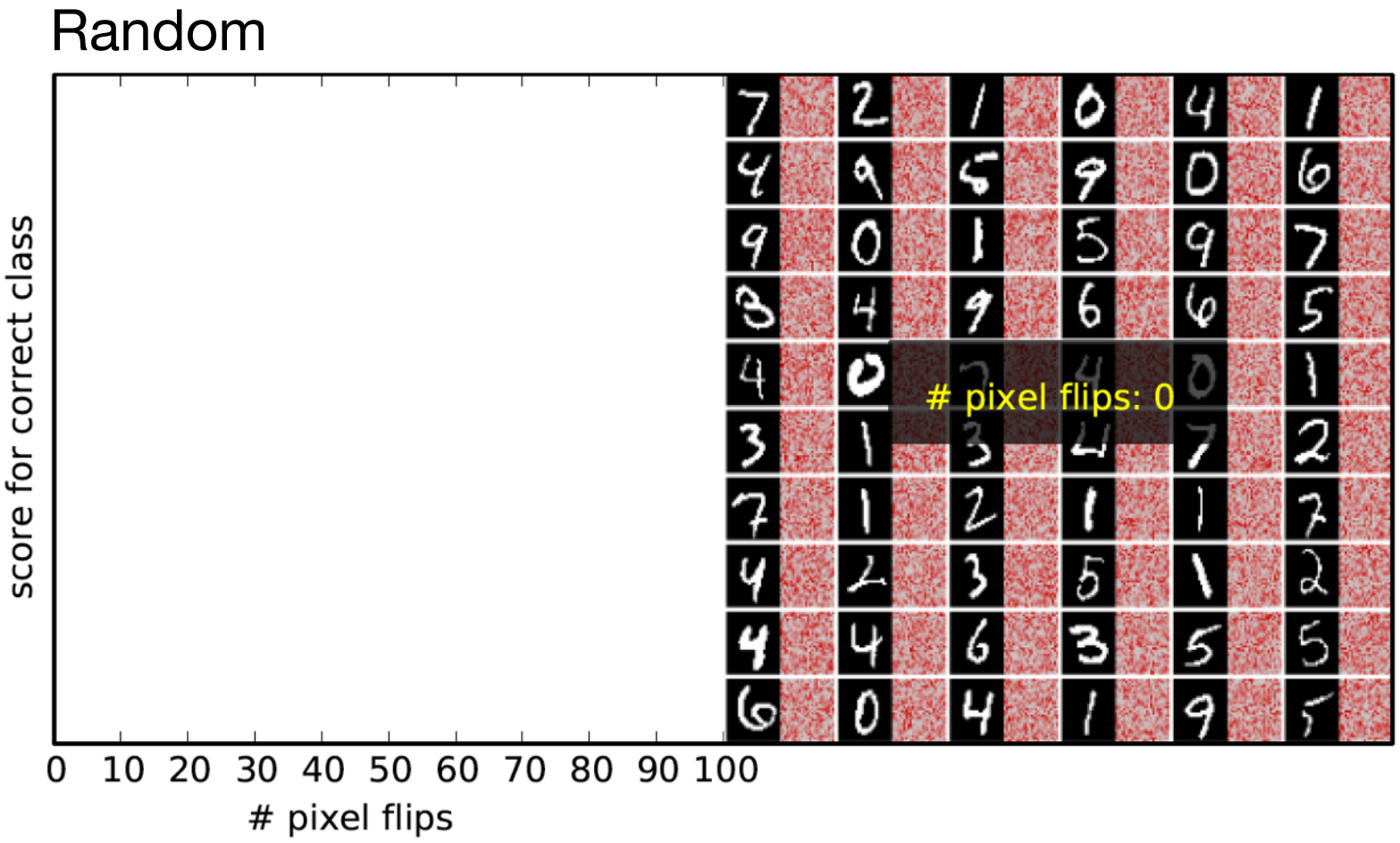
Compare Explanation Methods



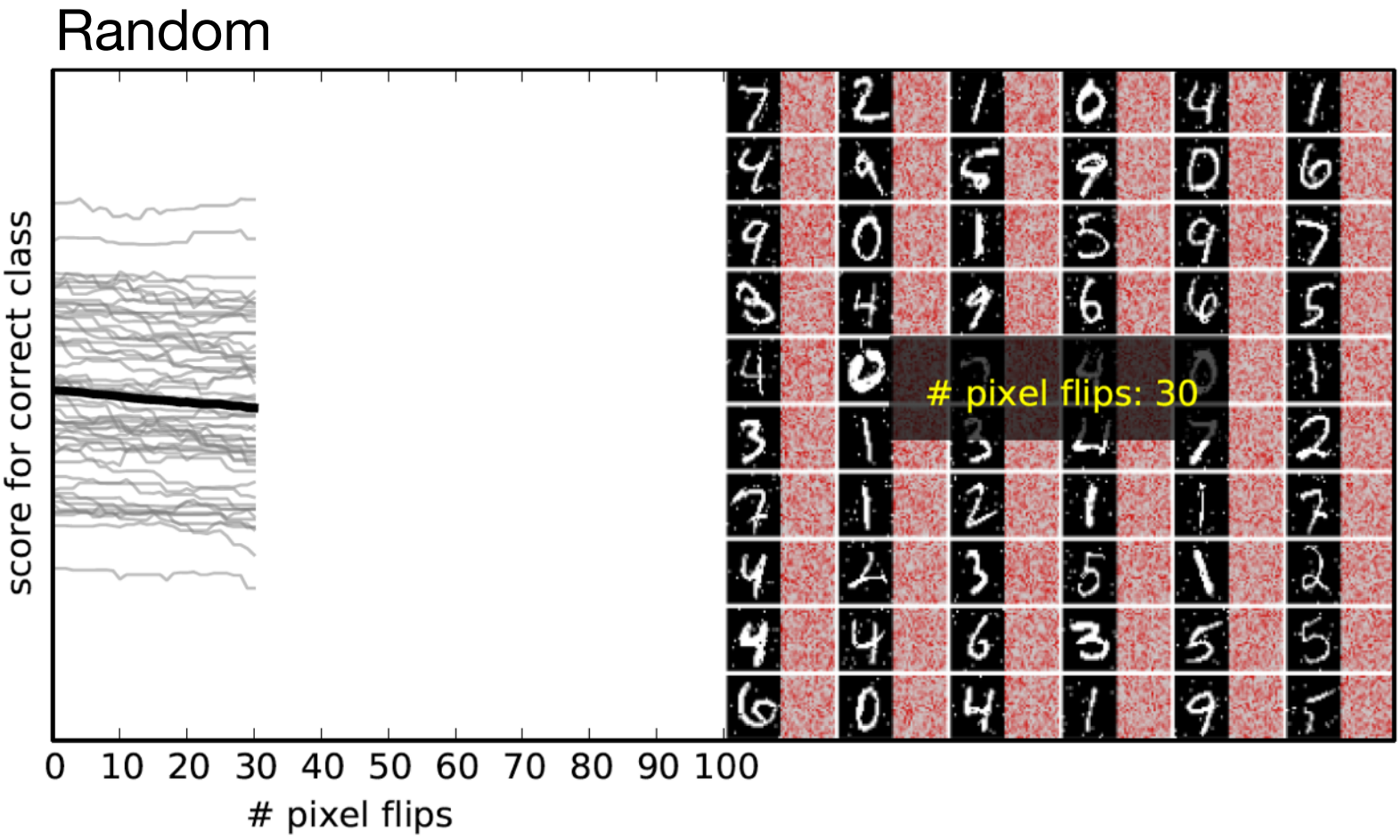
Compare Explanation Methods



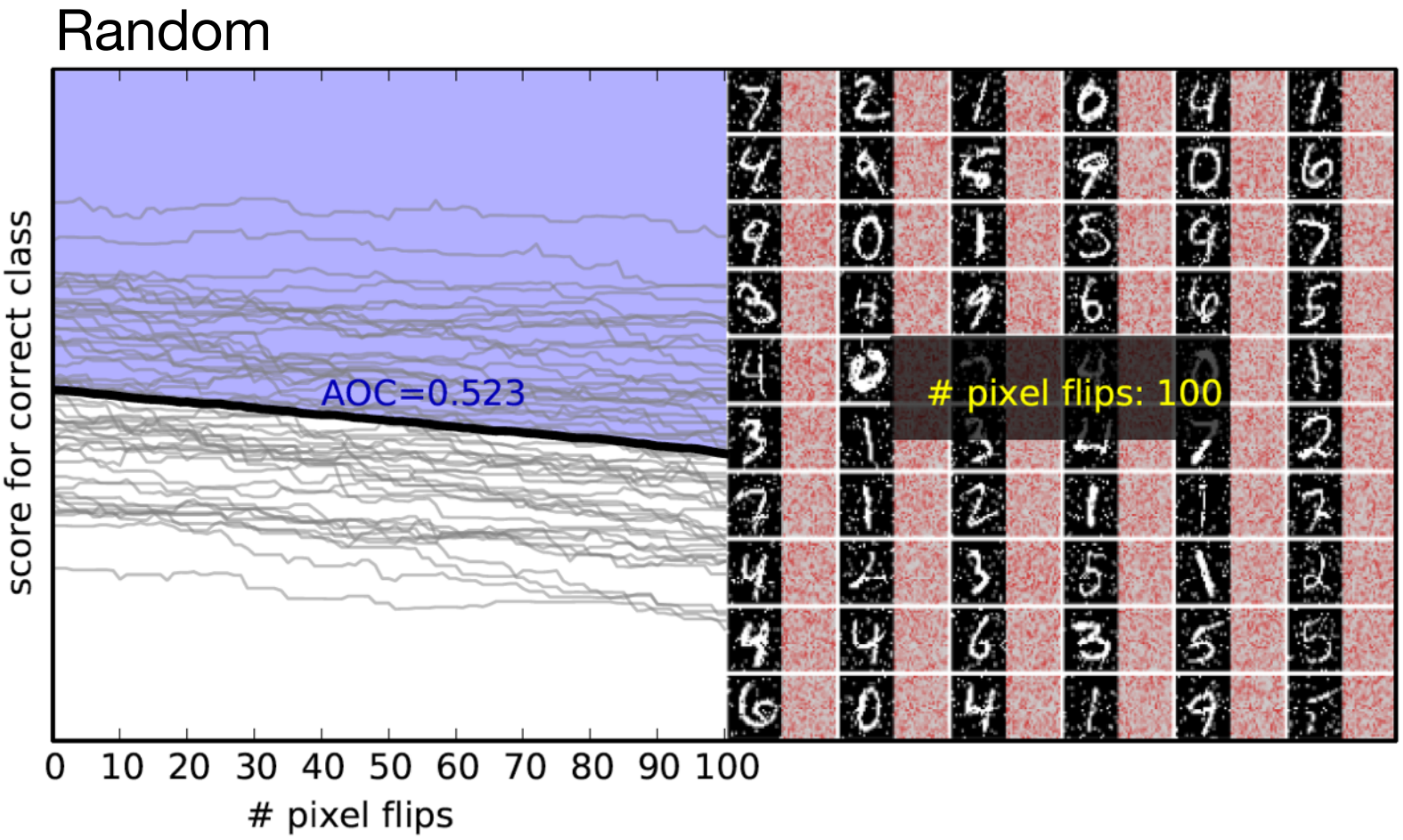
Compare Explanation Methods



Compare Explanation Methods



Compare Explanation Methods



Compare Explanation Methods

LRP:	**0.722**
Sensitivity:	0.691
Random:	0.523

LRP produces quantitatively better heatmaps than sensitivity analysis and random.

What about more complex datasets ?

SUN397

397 scene categories
(108,754 images in total)

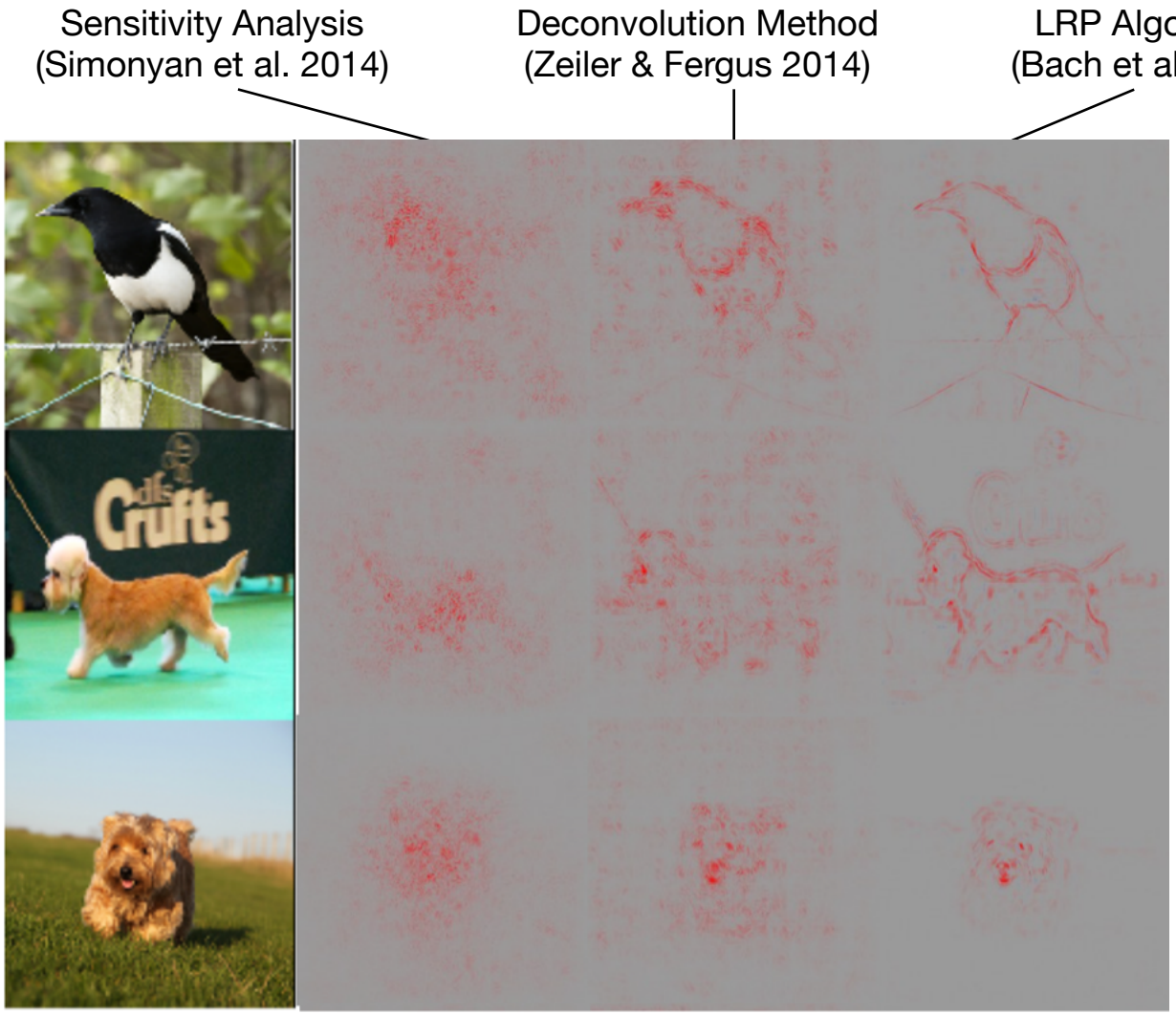
ILSVRC2012

1000 categories
(1.2 million training images)

MIT Places

205 scene categories
(2.5 millions of images)

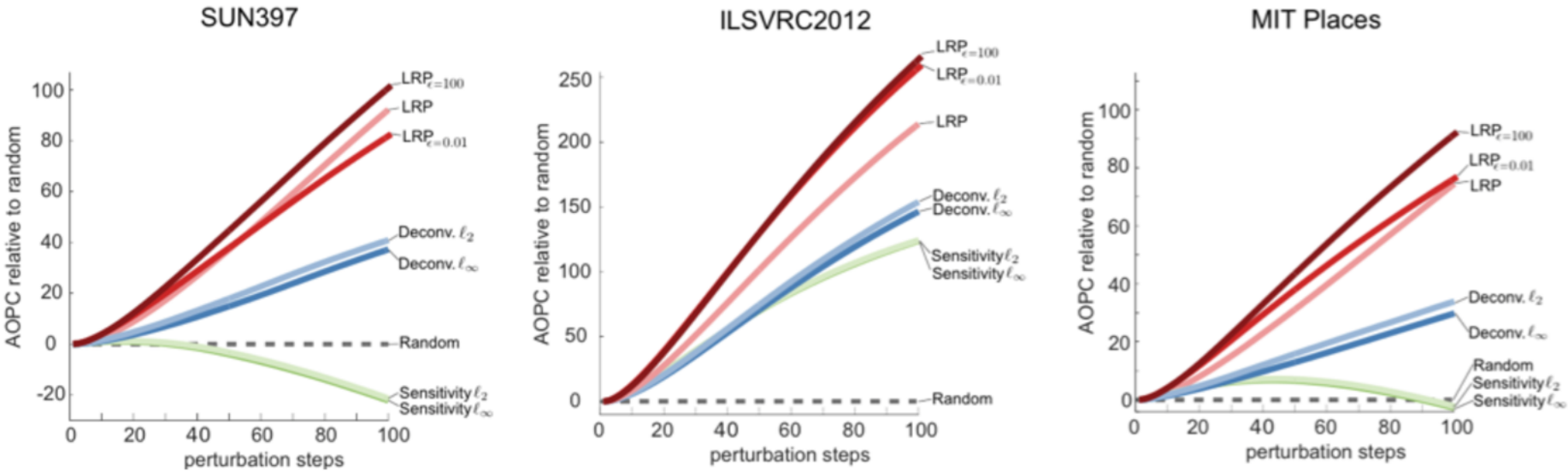
Compare Explanation Methods



(Samek et al. 2017)

Compare Explanation Methods

Red: LRP method
Blue: Deconvolution method (Zeiler & Fergus, 2014)
Green: Sensitivity method (Simonyan et al., 2014)



LRP produces quantitatively better heatmaps.

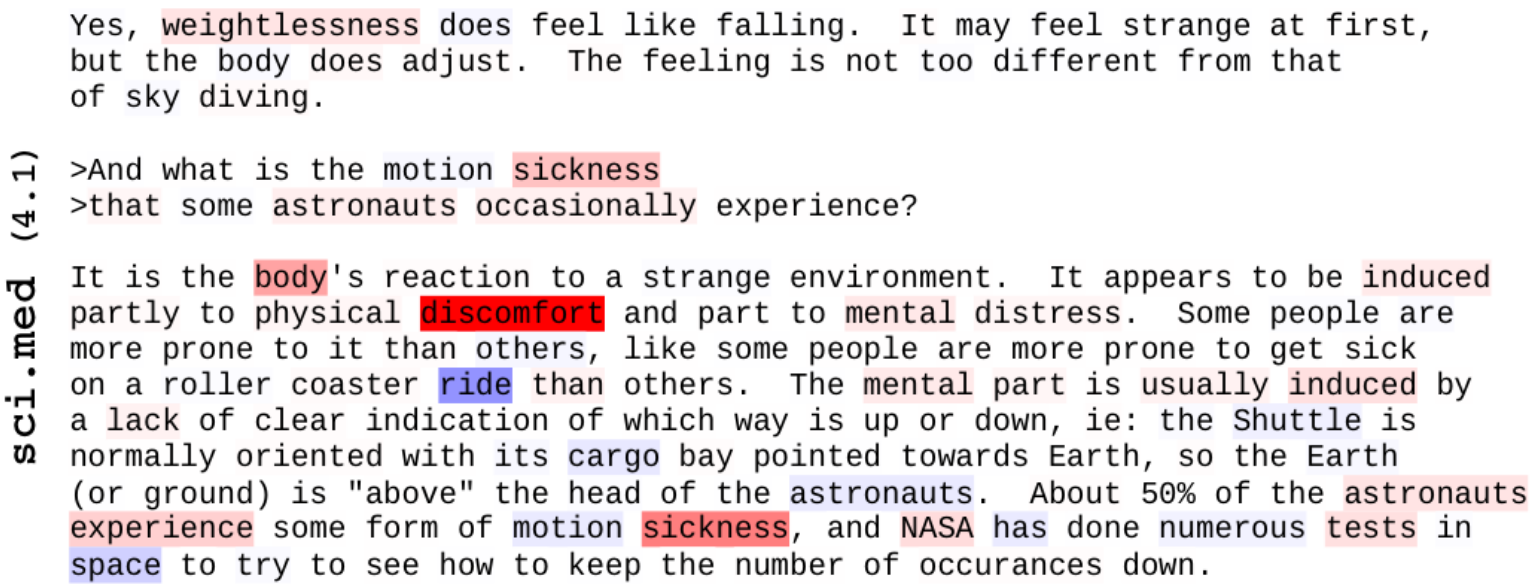
(Samek et al. 2017)

Compare Explanation Methods

Same idea can be applied for other domains (e.g. text document classification)

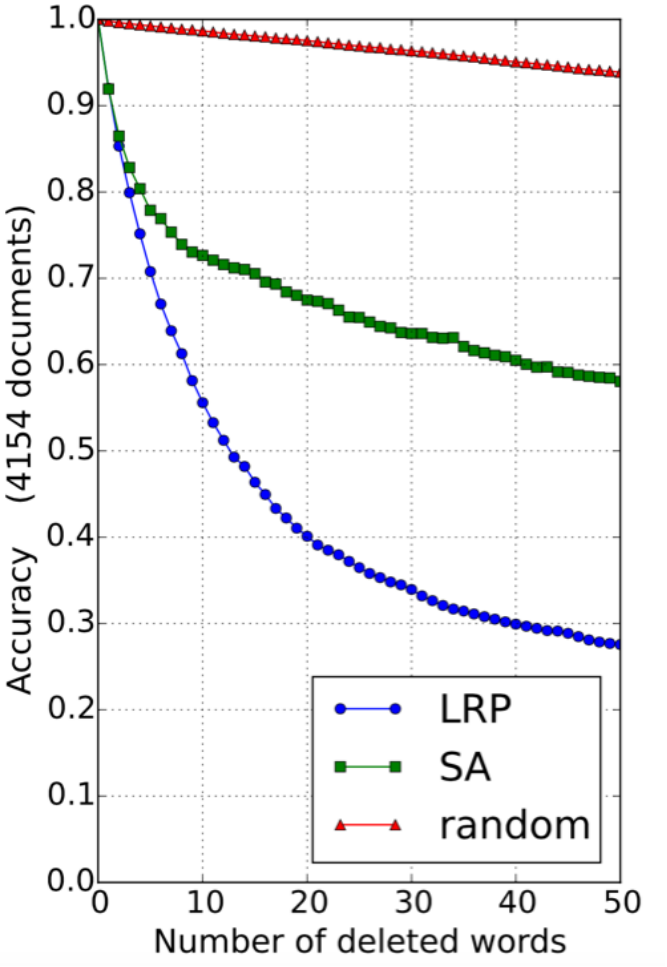
"Pixel flipping"
=
"Word deleting"

Text classified as "sci.med" —> LRP identifies most relevant words.



(Arras et al. 2016)

Compare Explanation Methods



Deleting relevant words leads to a quick decrease of classifier accuracy.

The decrease is much steeper for LRP than for random word deletion and deletion according to sensitivity.

LRP better identifies relevant words.

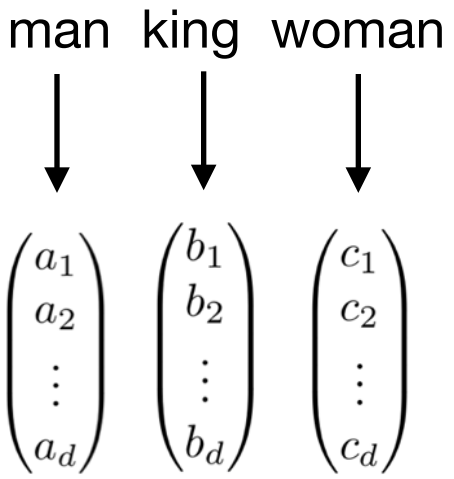
(Arras et al. 2016)

Applications

Application: Compare Classifiers

20 Newsgroups data set

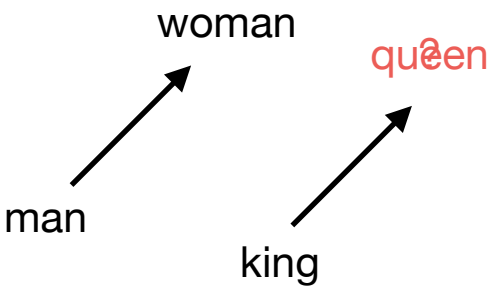
comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism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



Test set performance

word2vec / CNN model: 80.19%

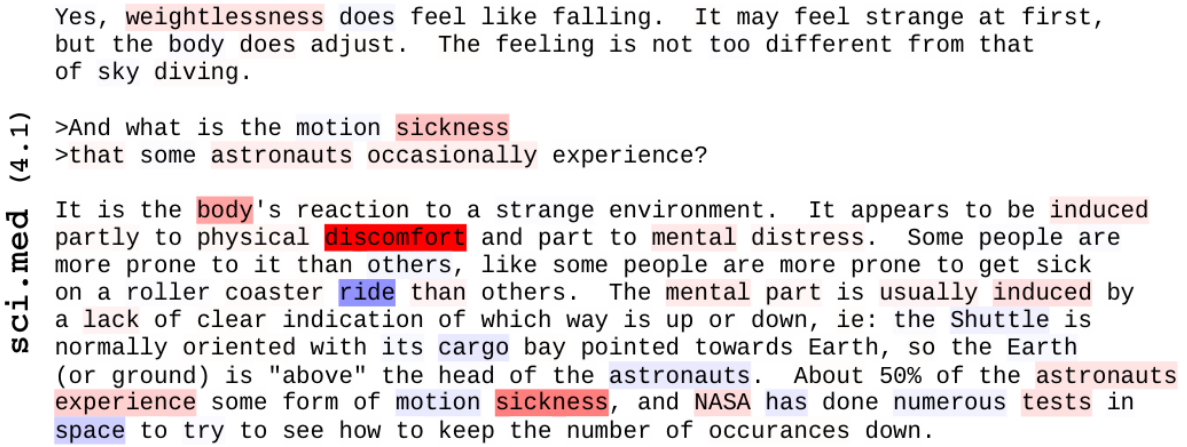
BoW/SVM model: 80.10%



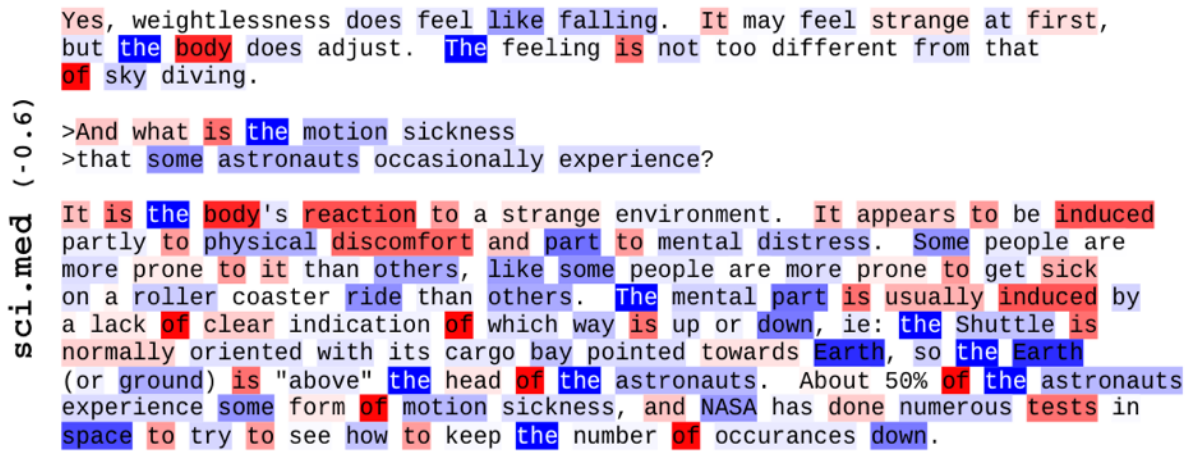
same performance —> same strategy ?

Application: Compare Classifiers

word2vec / CNN model



BoW/SVM model



(Arras et al. 2016)

Application: Compare Classifiers

word2vec / CNN model

sci.med

symptoms (7.3), treatments (6.6), medication (6.4), osteopathy (6.3), ulcers (6.2), sciatica (6.0), hypertension (6.0), herb (5.6), doctor (5.4), physician (5.1), Therapy (5.1), antibiotics (5.1), Asthma (5.0), renal (5.0), medicines (4.9), caffeine (4.9), infection (4.9), gastrointestinal (4.8), therapy (4.8), homeopathic (4.7), medicine (4.7), allergic (4.7), dosages (4.7), esophagitis (4.7), inflammation (4.6), arrhythmias (4.6), cancer (4.6), disease (4.6), migraine (4.6), patients (4.5).

BoW/SVM model

sci.med

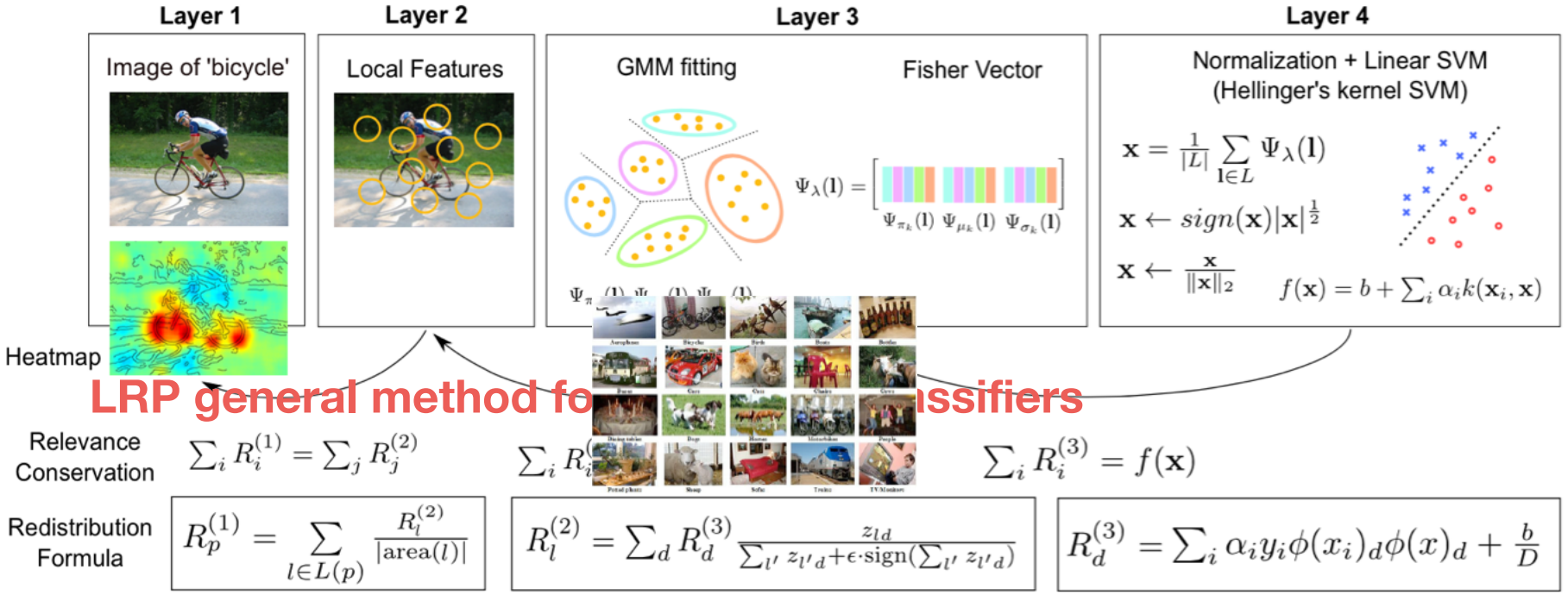
cancer (1.4), photography (1.0), doctor (1.0), msg (0.9), disease (0.9), medical (0.8), sleep (0.8), radiologist (0.7), eye (0.7), treatment (0.7), prozac (0.7), vitamin (0.7), epilepsy (0.7), health (0.6), yeast (0.6), skin (0.6), pain (0.5), liver (0.5), physician (0.5), she (0.5), needles (0.5), dn (0.5), circumcision (0.5), syndrome (0.5), migraine (0.5), antibiotic (0.5), water (0.5), blood (0.5), fat (0.4), weight (0.4).

Words with maximum relevance

(Arras et al. 2016)

Application: Compare Classifiers

Fisher Vector / SVM Classifier



Deep Neural Network

- BVLC reference model + fine tuning.

Dataset

- PASCAL VOC 2007

(Lapuschkin et al. 2016)

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%

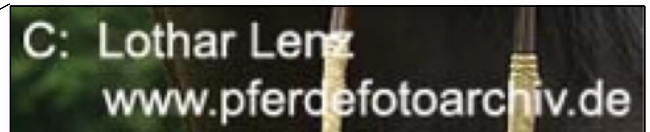
Image

same performance —> same strategy ?

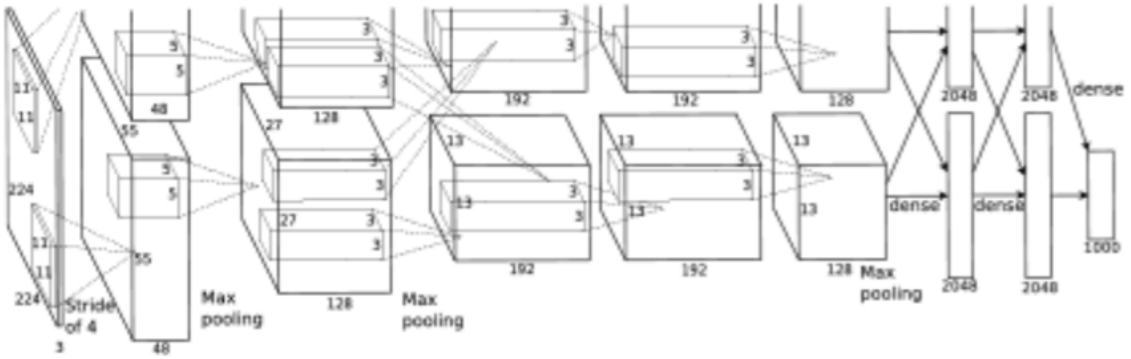
(Lapuschkin et al. 2016)

Application: Compare Classifiers

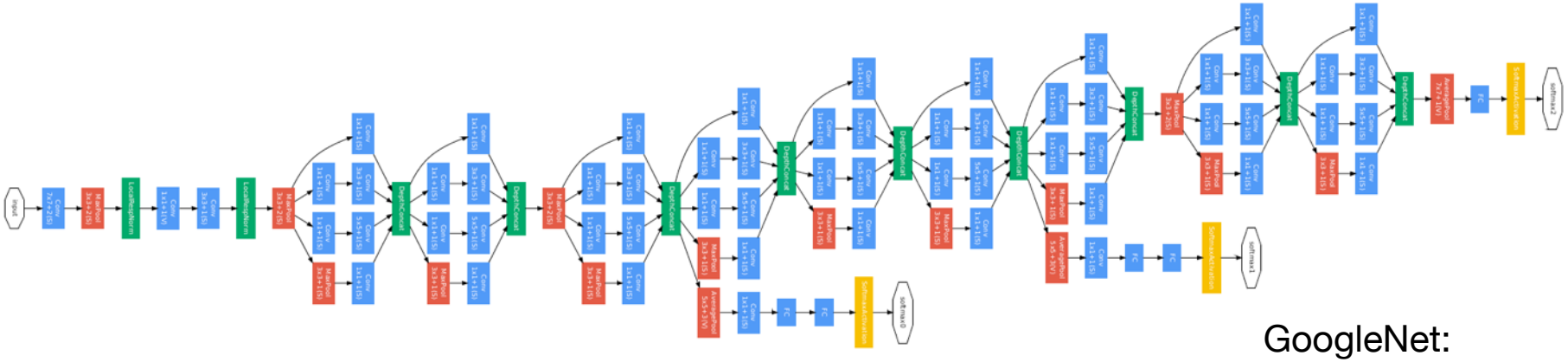
‘horse’ images in PASCAL VOC 2007



Application: Compare Classifiers



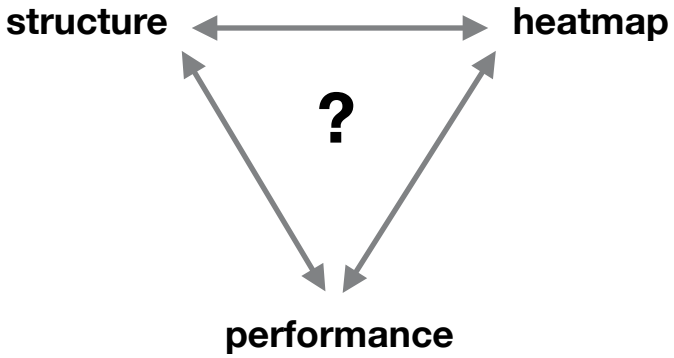
BVLC:
- 8 Layers
- ILSRCV: 16.4%



GoogleNet:
- 22 Layers
- ILSRCV: 6.7%
- Inception layers

Application: Compare Classifiers

GoogleNet focuses on
faces of animal.
—> suppresses background noise



(Binder et al. 2016)

Application: Measure Context Use

how important
is context ?

how important
is context ?

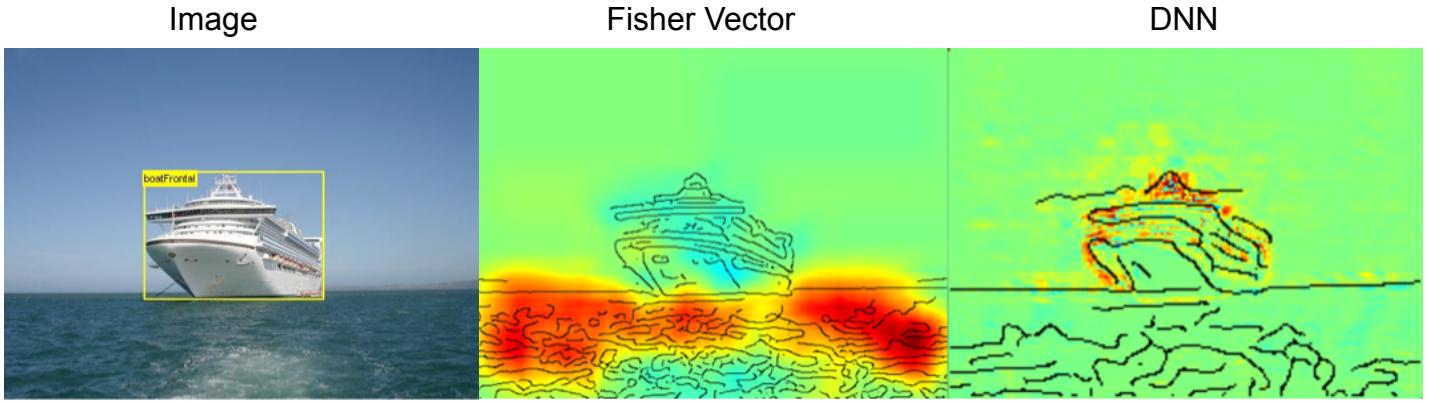
classifier

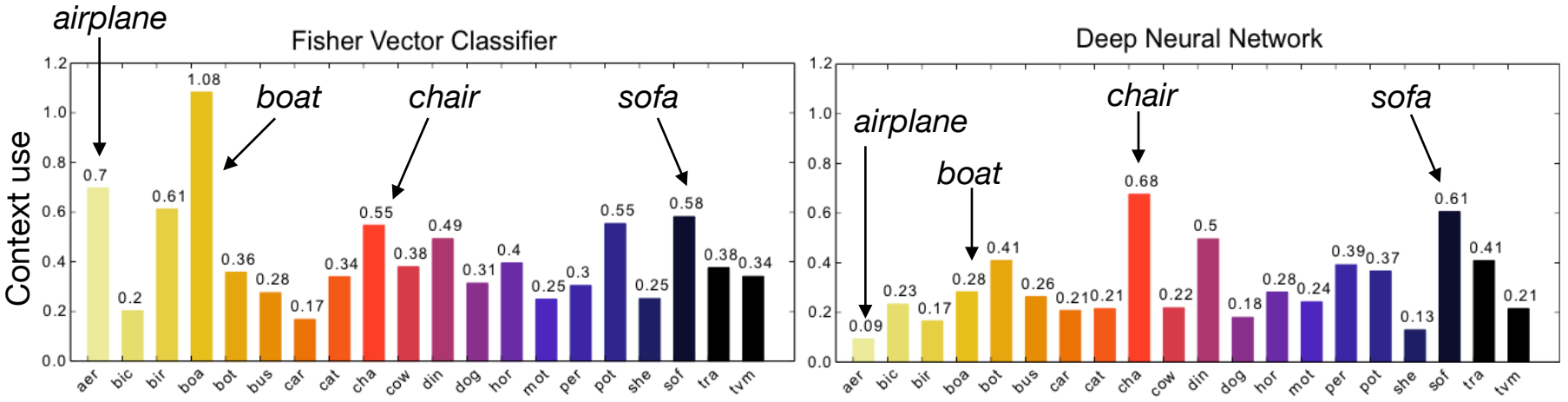
LRP decomposition allows meaningful pooling over bbox !

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

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

Application: Measure Context Use





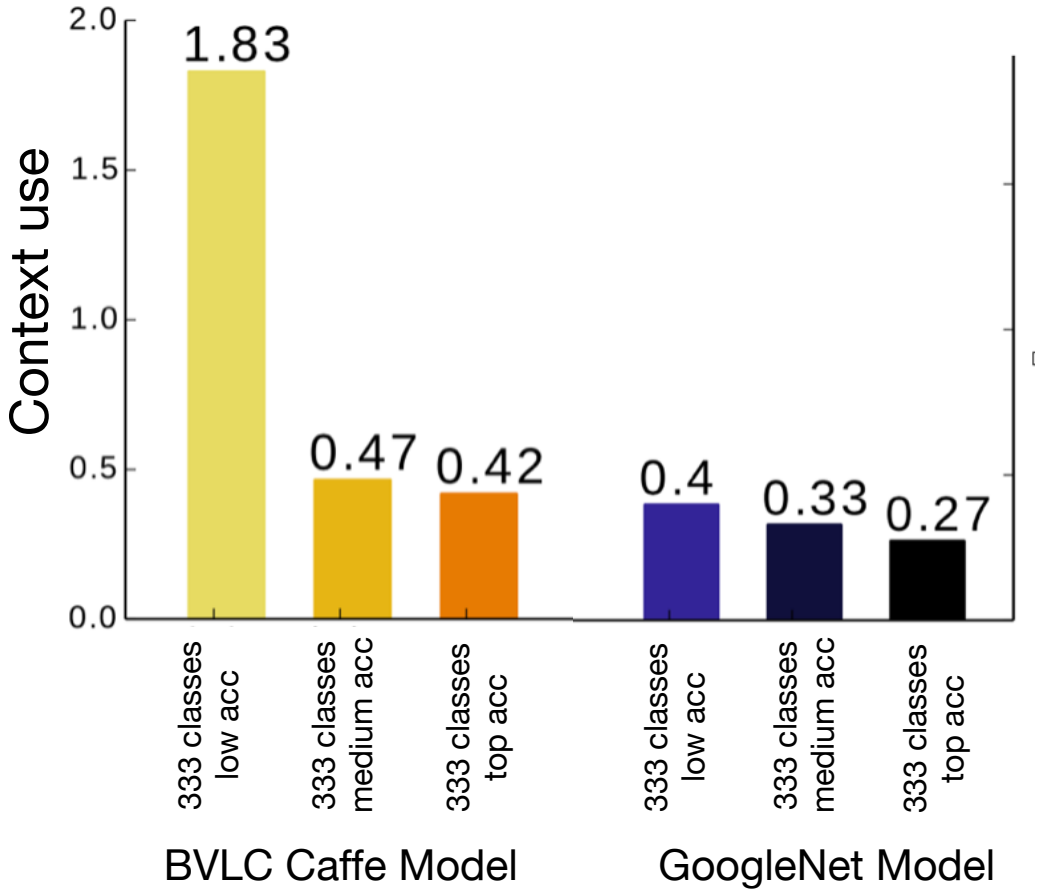
Large values indicate importance of context

(Lapuschkin et al. 2016)

Application: Measure Context Use

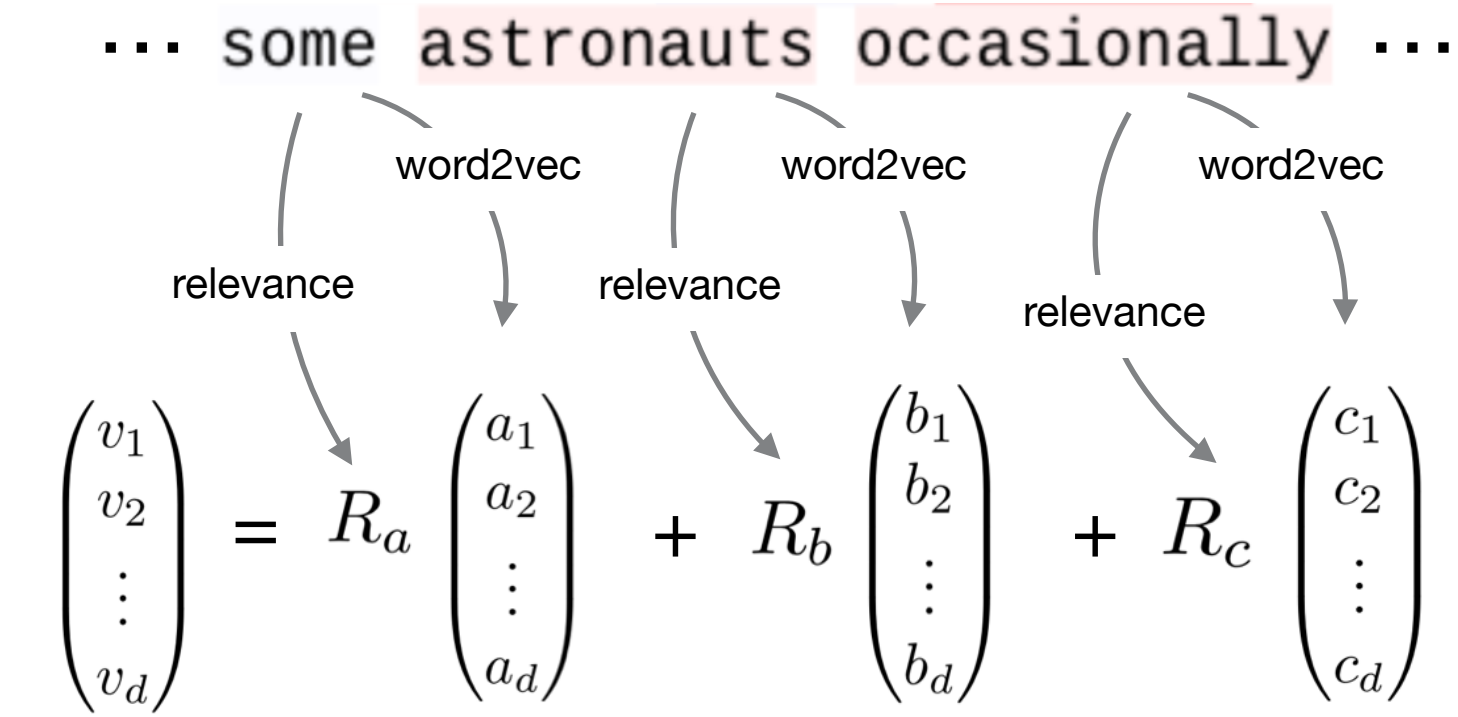
GoogleNet uses less context than BVLC model.

Context use anti-correlated with performance.



(Lapuschkin et al. 2016)

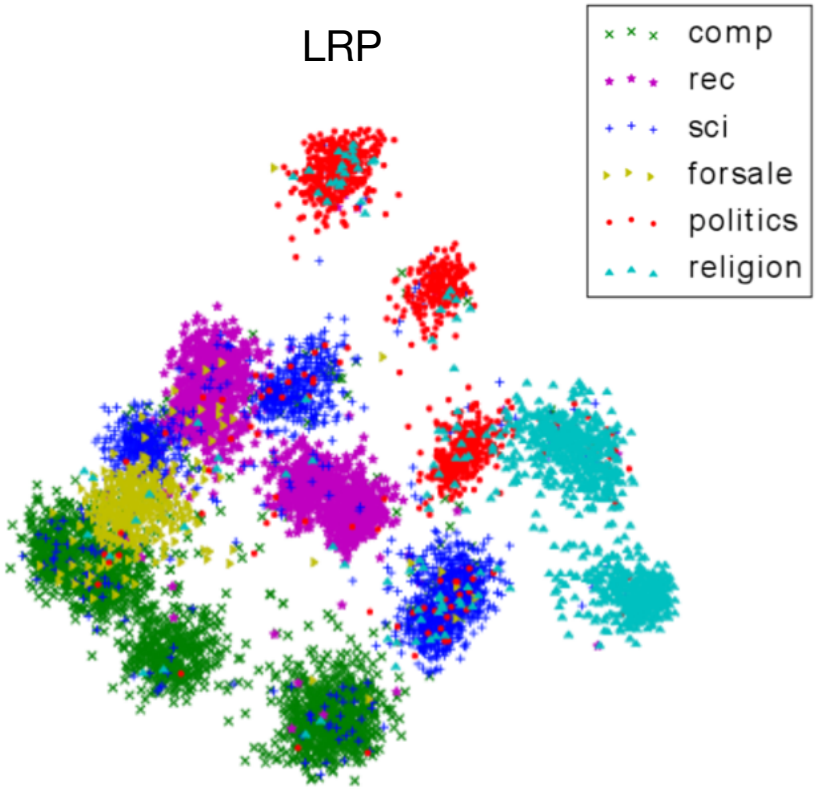
Application: Novel Representation



(Arras et al. 2016)

Application: Novel Representation

2D PCA projection of document vectors



uniform

TFIDF

Document vector computation is unsupervised.

KNN-Performance on document vectors (Explanatory Power Index)

LRP: 0.8076
TFIDF: 0.6816
uniform: 0.6208

(Arras et al. 2016)

Application: Sentiment Analysis

Model: word-based bidirectional LSTM

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

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

[model released by Li et al. 2016]

Task: five-class sentiment prediction

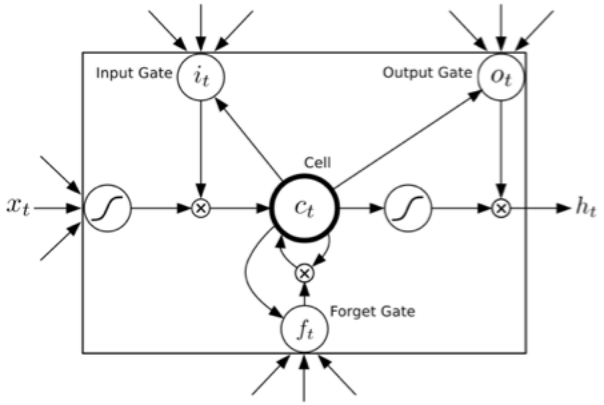
training on phrases and sentences of the Stanford Sentiment Treebank

[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]



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

Application: Sentiment Analysis

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 .

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: Sentiment Analysis

e.g. word deleting
by decreasing LRP relevance:

original sentence the cat sat on the mat

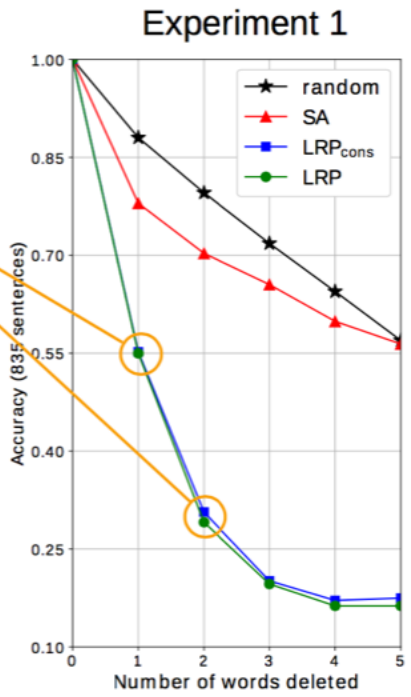
1 deleted word the ◯ sat on the mat

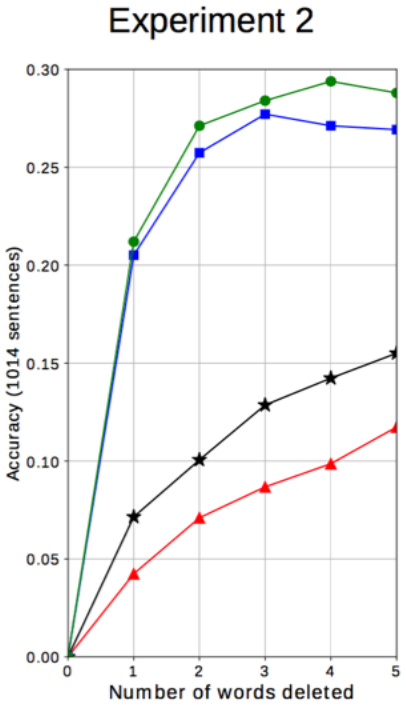
2 deleted words the ◯ ◯ on the mat

3 deleted words ◯ ◯ ◯ on the mat

⋮ ⋮

deletion =
word-embedding set to zero
in the input sentence

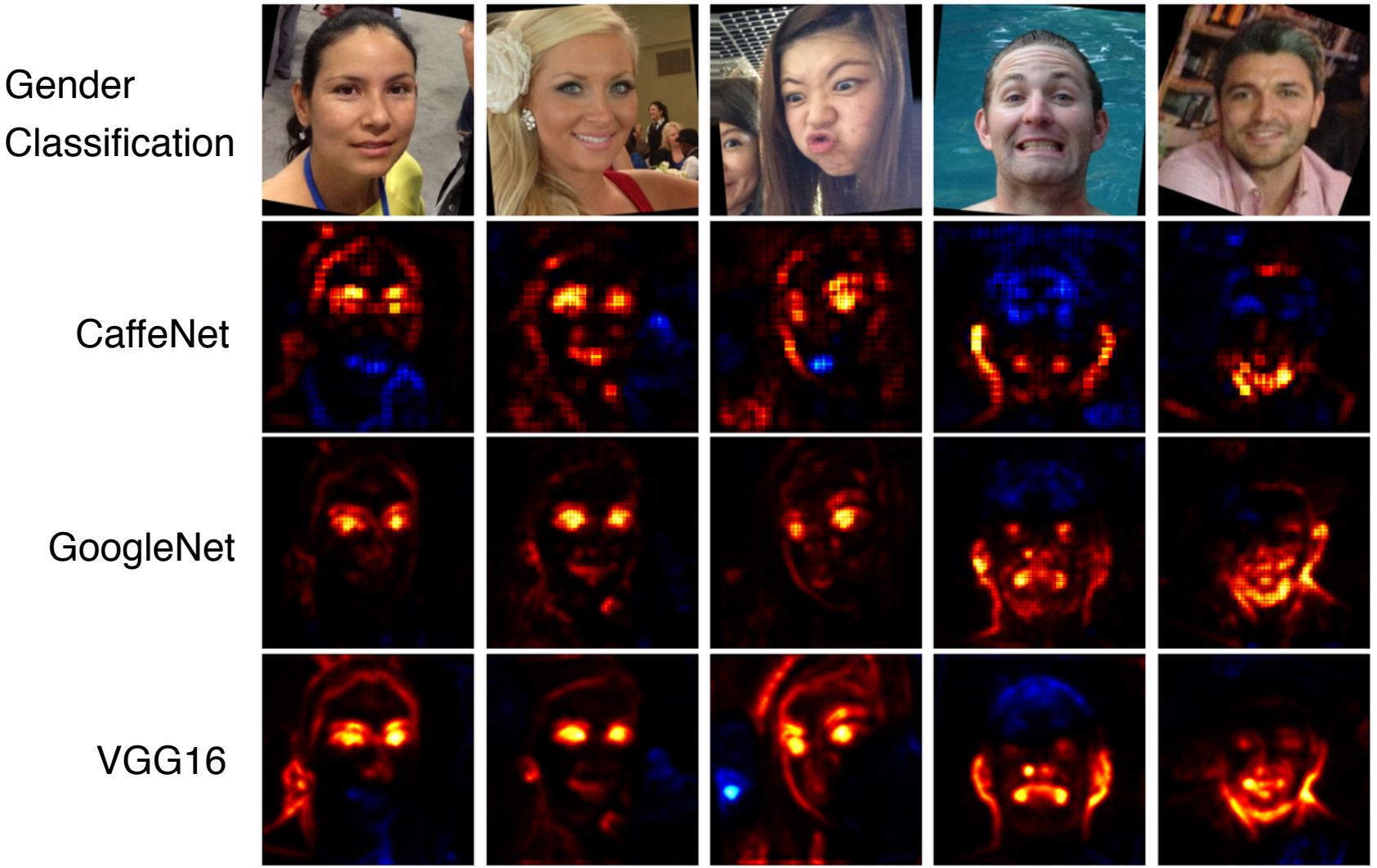




start with sentences initially:	correctly classified	falsely classified
deletion order:	decreasing relevance	increasing relevance

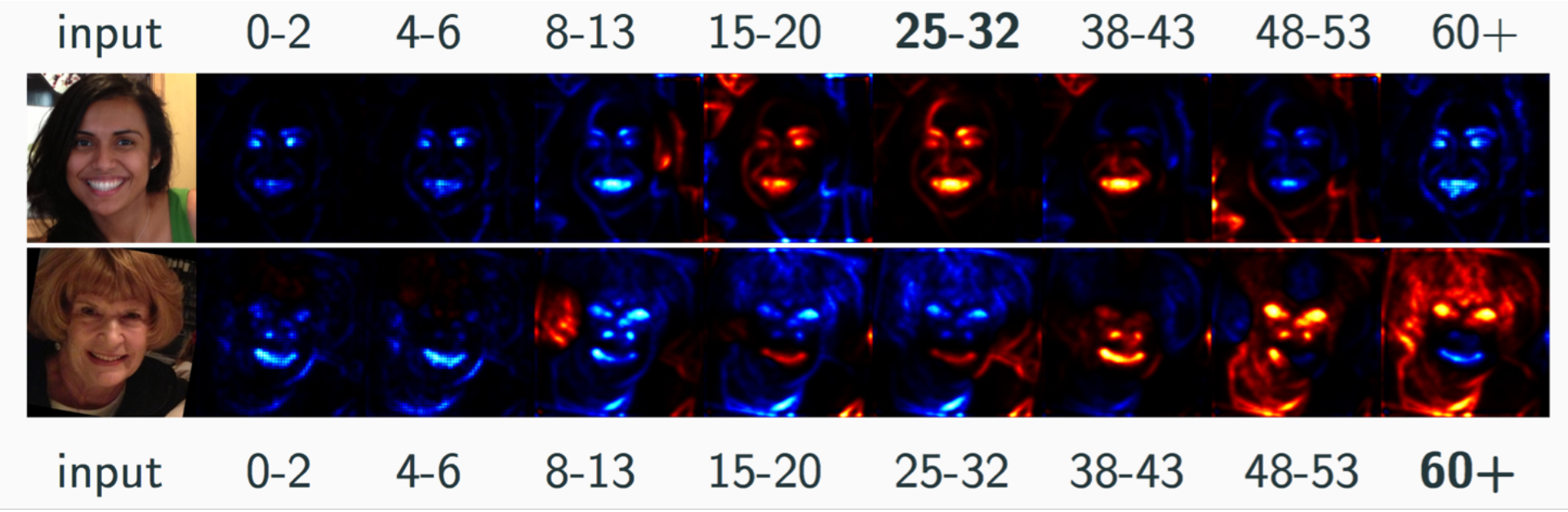
(Arras et al., 2017)

Application: Face Analysis



(Lapuschkin et al. 2017)

Application: Face analysis

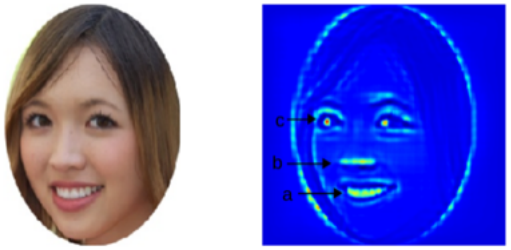


Why image classified as young ?
- smile

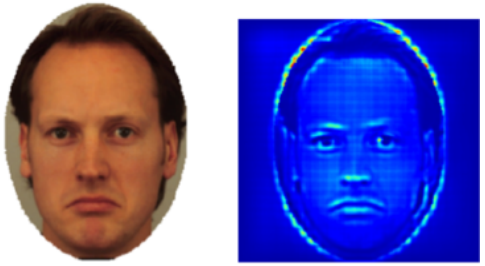
Why image classified as old ?
- eyes, wrinkles

(Lapuschkin et al. 2017)

Attractivity



Emotions

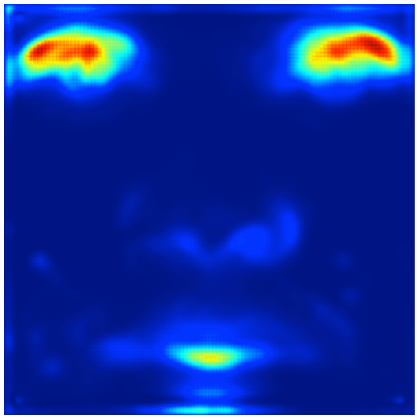


(Arbabzadah et al. 2016)

Application: Face analysis

real person | fake person | real person

fake persons have different eyes



	AlexNet		GoogLeNet		VGG19	
	FRR	FAR	FRR	FAR	FRR	FAR
From Scratch	16.2%	1.9%	10.0%	1.8%	10.9%	2.2%
Pretrained	11.4%	0.9%	5.6%	1.2%	3.5%	0.8%

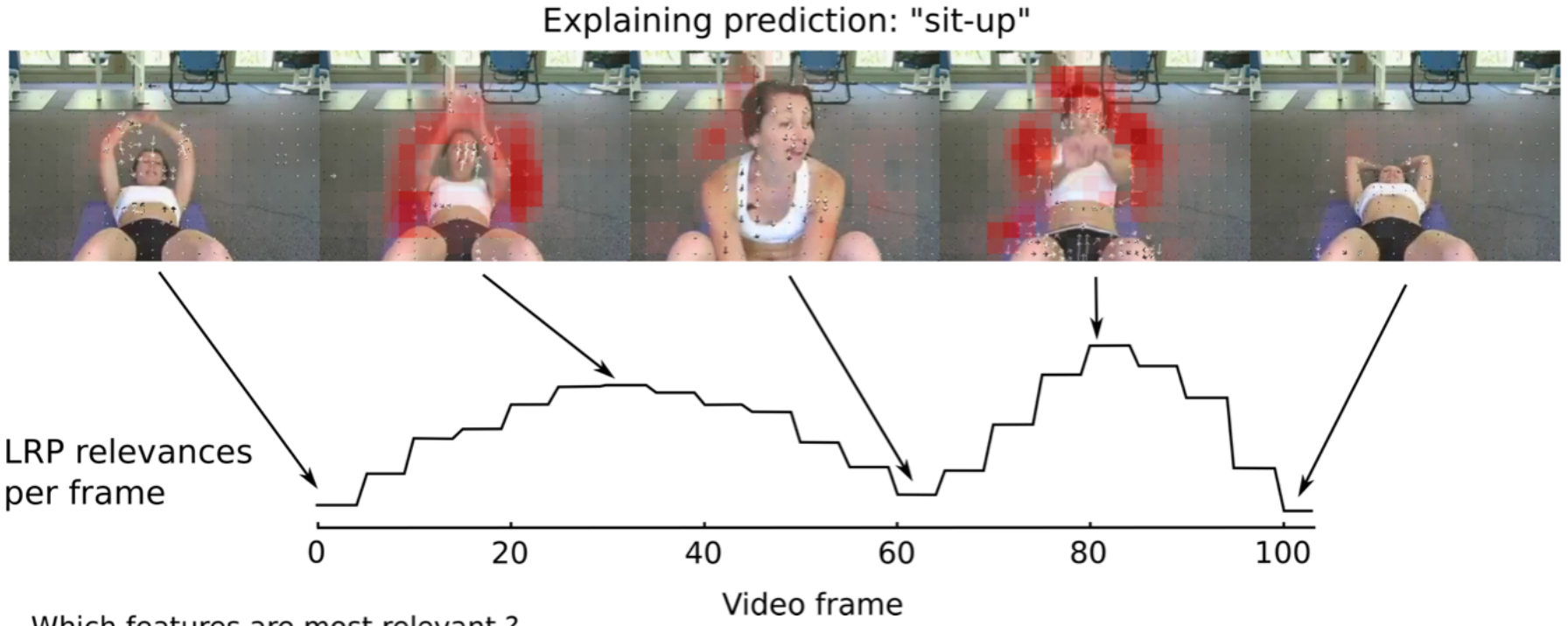
(Seibold et al., IWDW, 2017)

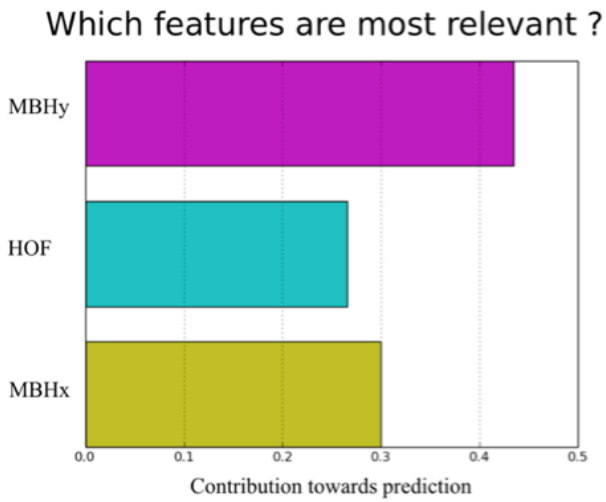
Application: Video Analysis

Motion vectors can be extracted from the compressed video
-> allows very efficient analysis

(Srinivasan et al. 2017)

Application: Video Analysis

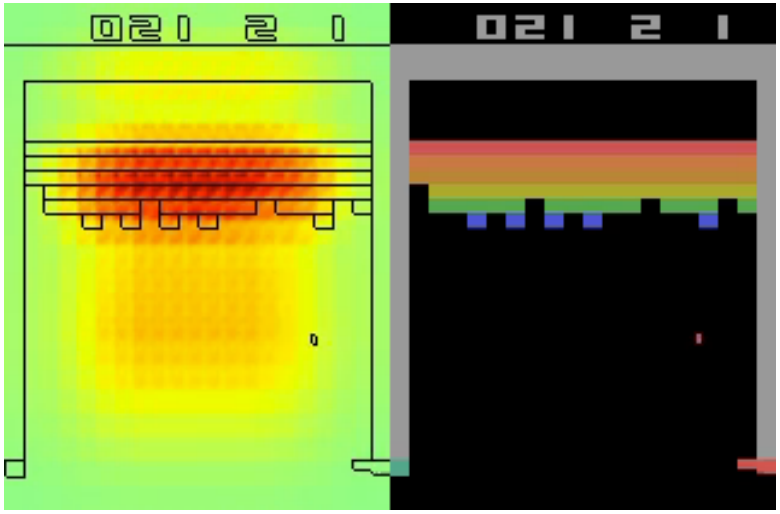




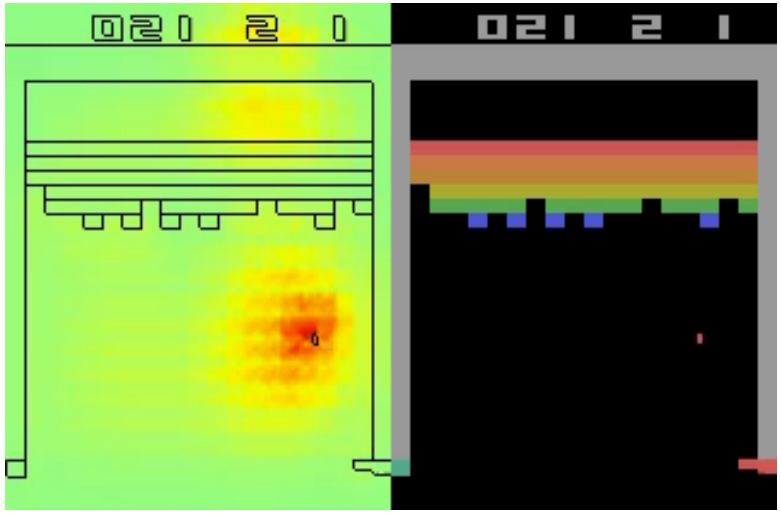
(Srinivasan et al. 2017)

Application: Machines playing Games

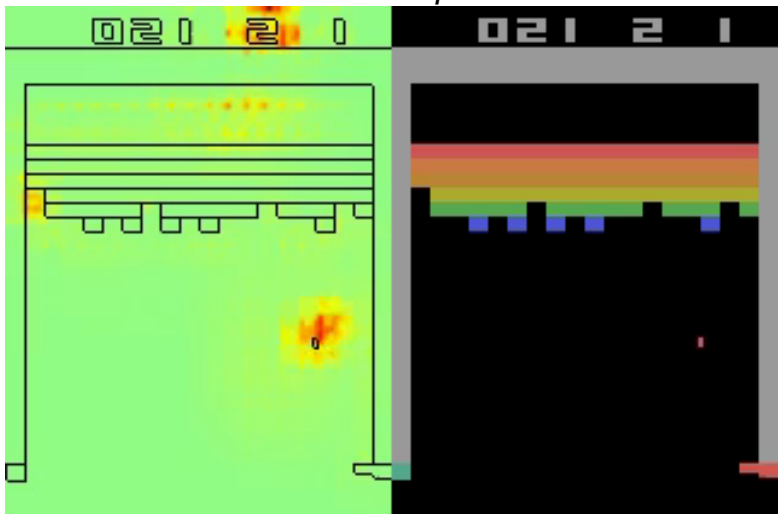
After 0 epochs



After 25 epochs

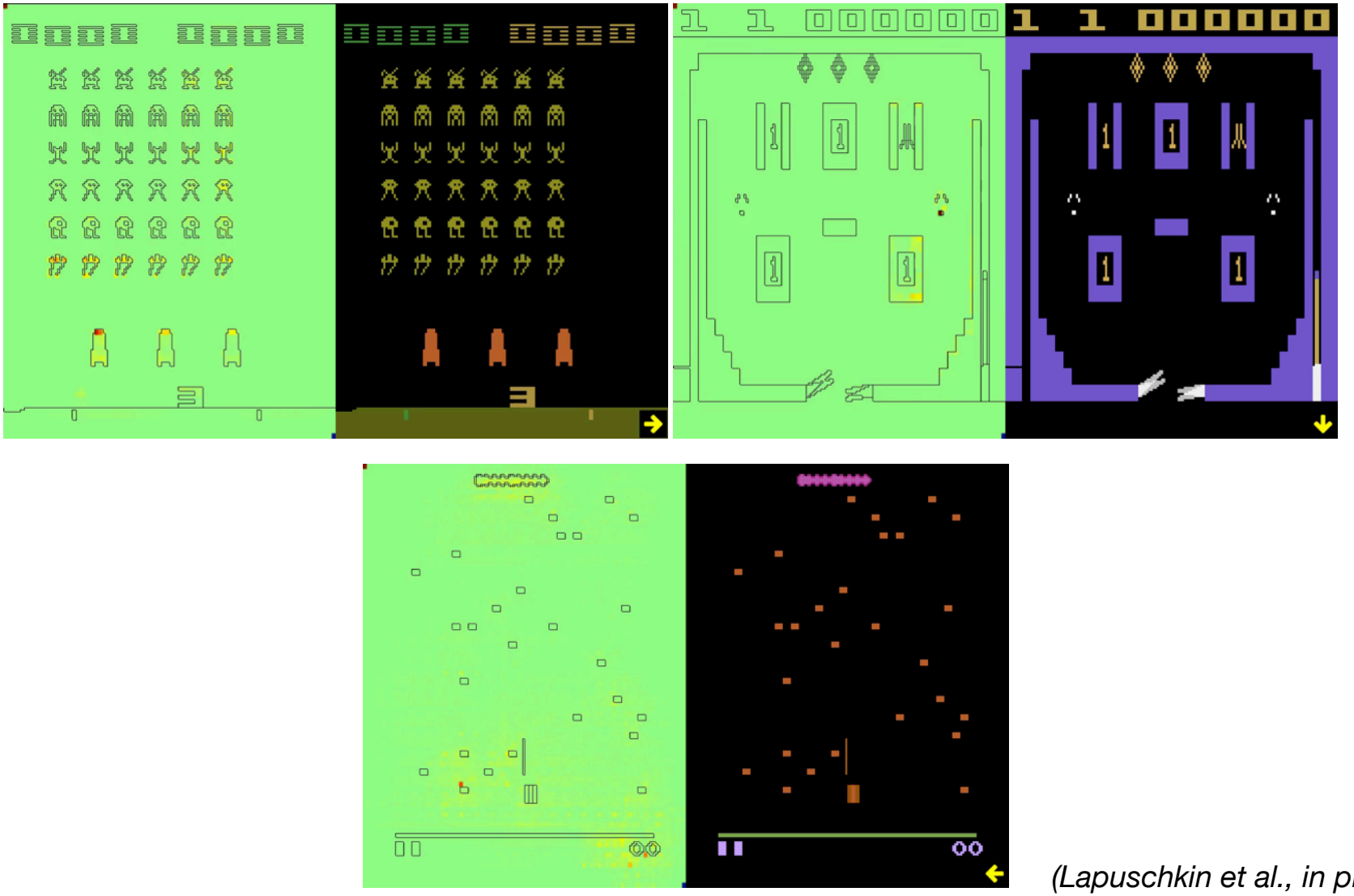


After 195 epochs



(Lapuschkin et al., in prep.)

Application: Machines playing Games



(Lapuschkin et al., in prep.)

Summary

- In many problems interpretability as important as prediction.

- Explaining individual predictions is key.

- We have powerful, mathematically well-founded methods (LRP / deep Taylor) to explain individual predictions.

- More research needed on how to compare and evaluation all the different aspects of interpretability.

- Many interesting applications with interpretable deep nets
  —> more to come soon !

Our 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.

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.

A Binder, W Samek, G Montavon, S Bach, KR Müller. Analyzing and Validating Neural Networks Predictions. *ICML Workshop on Visualization for Deep Learning*, 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.

Our References

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. *ICCV'17 Workshop on Analysis and Modeling of Faces and Gestures (AMFG)*, 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. DeepTaylor Decomposition of Neural Networks. *ICML Workshop on Visualization for Deep Learning*, 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.

G Montavon, W Samek, KR Müller. Methods for Interpreting and Understanding Deep Neural Networks. *Digital Signal Processing, 73:1-15*, 2017.

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*, 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.