EMBC Tutorial on Interpretable and Transparent Deep Learning

Wojciech Samek
(Fraunhofer HHI)

Grégoire Montavon
(TU Berlin)

Klaus-Robert Müller
(TU Berlin)

13:30 - 14:00	Introduction KRM
14:00 - 15:00	Techniques for Interpretability GM
15:00 - 15:30	Coffee Break ALL
15:30 - 16:15	Evaluating Interpretability & Applications WS
16:15 - 17:15	Applications in BME & the Sciences and Wrap-Up KRM

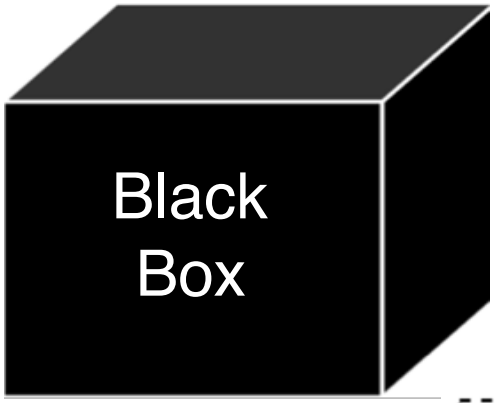
Berlin EMBConference 2019

BERLIN BIG DATA CENTER

Berliner Zentrum für MASCHINELLES LERNEN

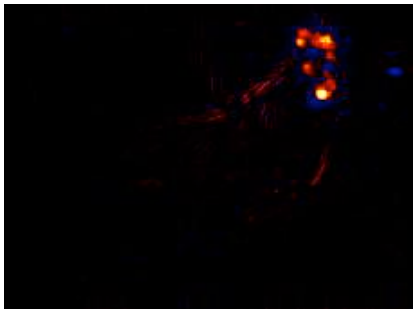
LRP Revisited

input $\boldsymbol{x}$



Rooster

prediction $f(\boldsymbol{x})$



heatmap

Explain prediction

(how much each pixel contributes to prediction)

Idea: Decompose function

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

Theoretical Interpretation
(Deep) Taylor decomposition**.**

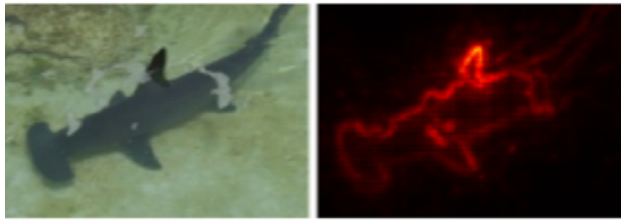
alpha-beta LRP rule (Bach et al. 2015)

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

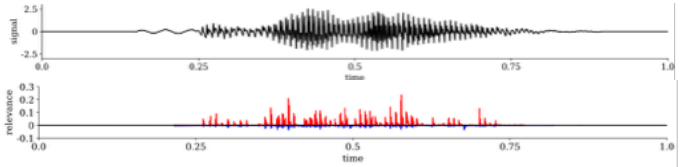
where $\alpha + \beta = 1$

LRP Works for Different Data

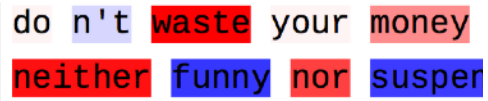
General Images (Bach' 15, Lapuschkin'16)



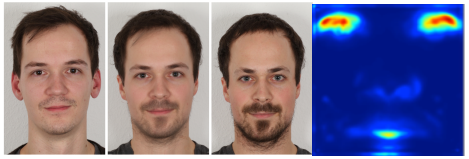
Speech (Becker'18)



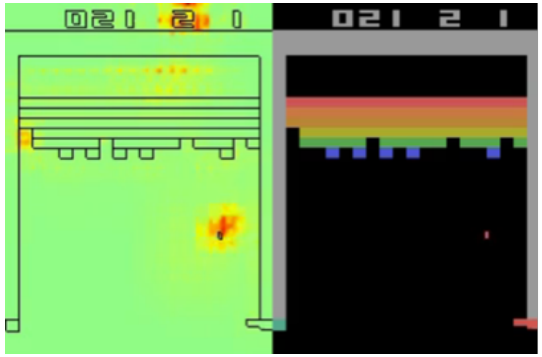
Text Analysis (Arras'16 &17)



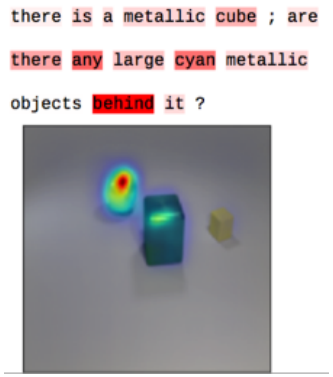
Morphing (Seibold'18)



Games (Lapuschkin'19)

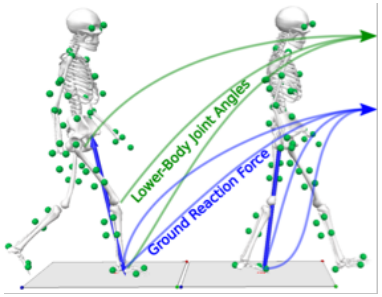


VQA (Samek'19)

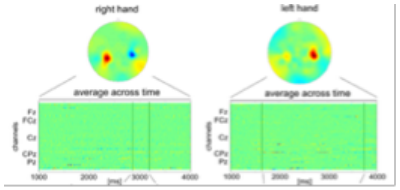


Video (Anders'18)

Gait Patterns (Horst'19)

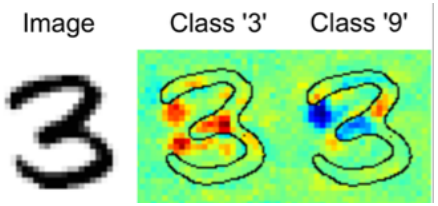


EEG (Sturm'16)

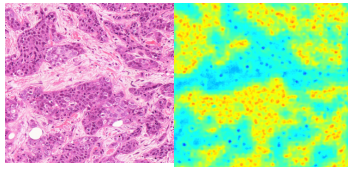


Faces (Lapuschkin'17)

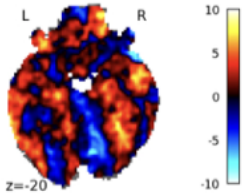
Digits (Bach' 15)



Histopathology (Binder'18)

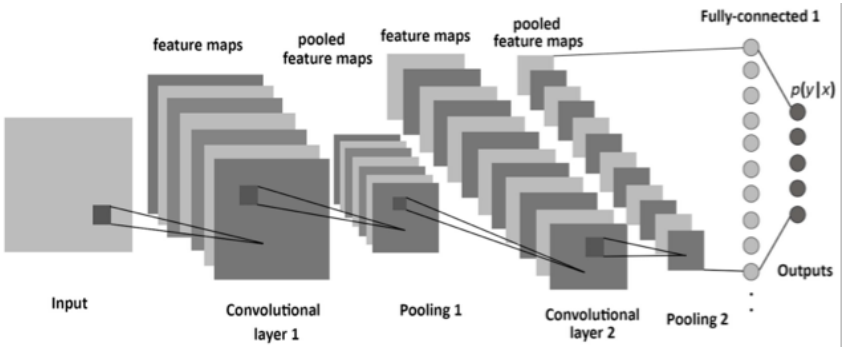


fMRI (Thomas'18)

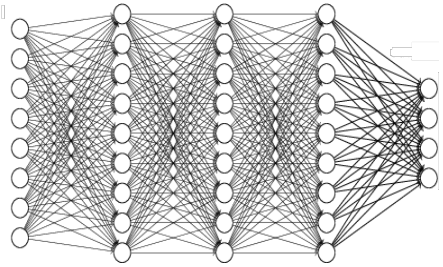


LRP Works for Different Models

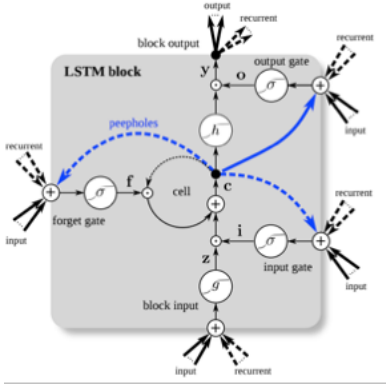
Convolutional NNs (Bach'15, Arras'17 …)



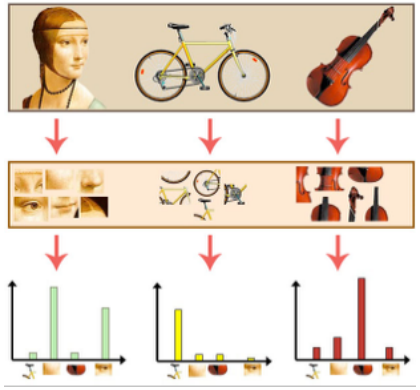
Local Renormalization Layers (Binder'16)



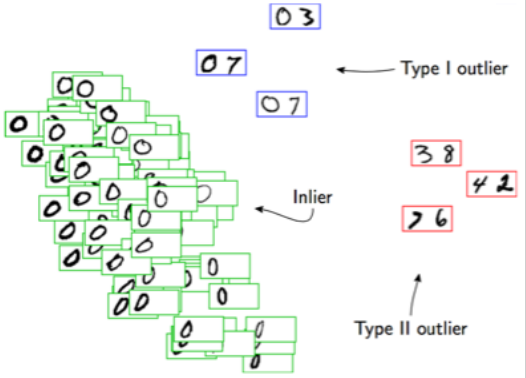
LSTM (Arras'17, Thomas'18)



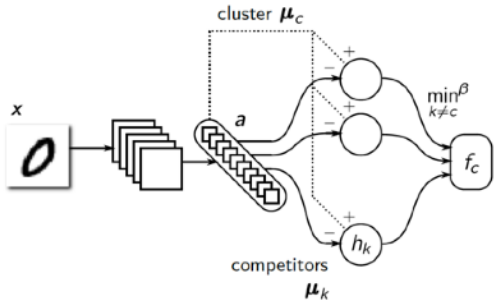
Bag-of-words / Fisher Vector models (Bach'15, Arras'16, Lapuschkin'17, Binder'18)



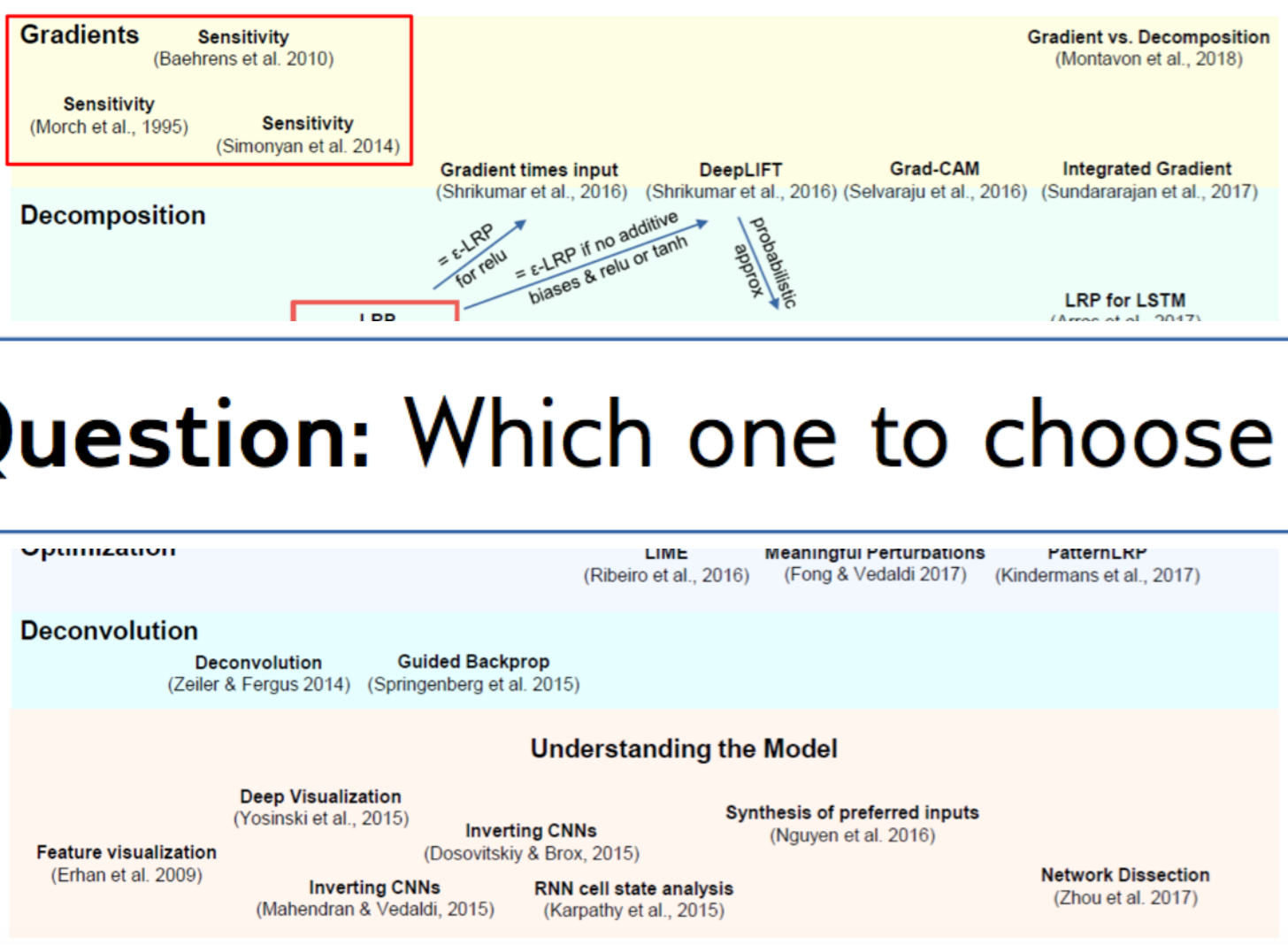
One-class SVM (Kauffmann'18)



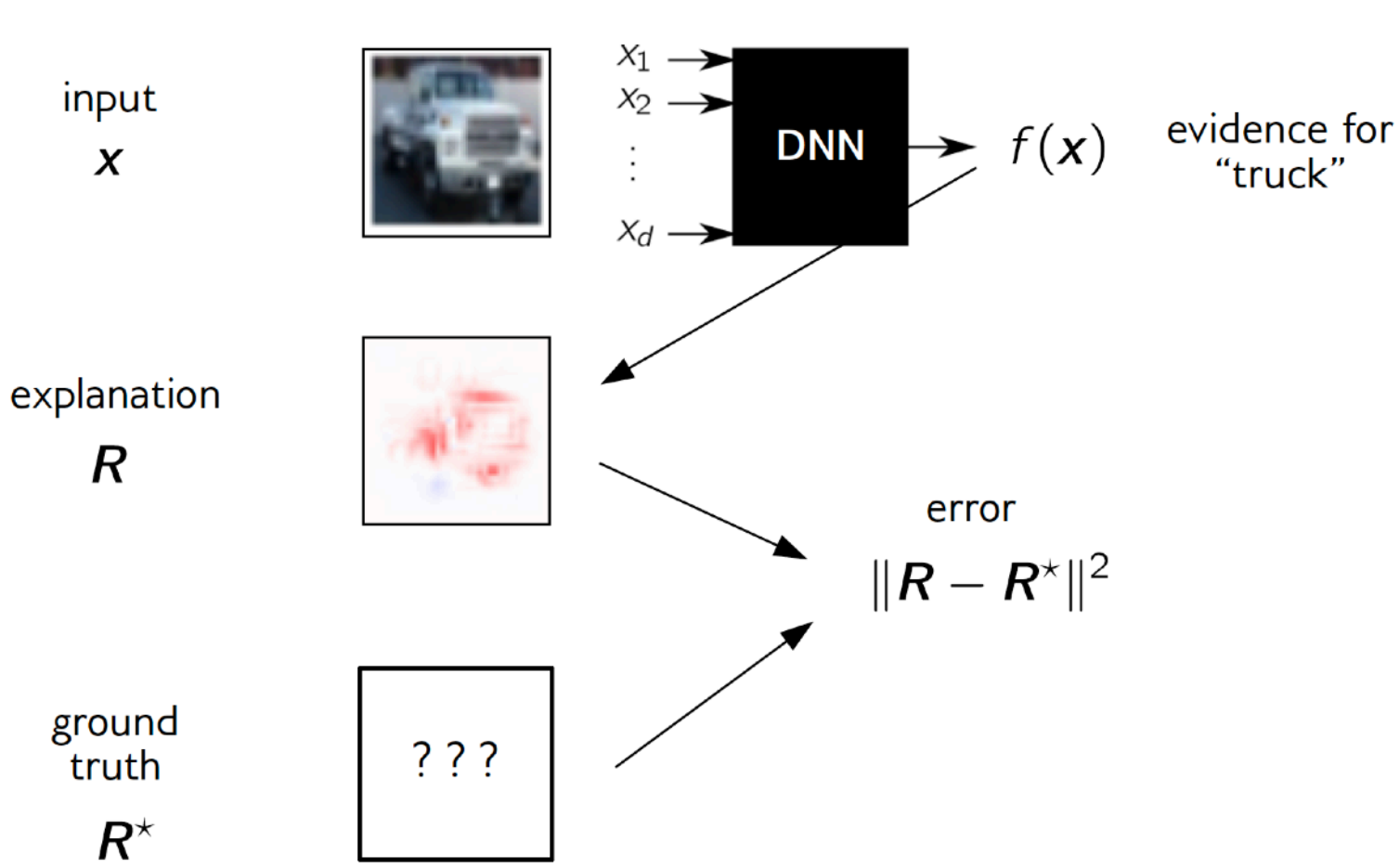
Clustering (Kauffmann'19)



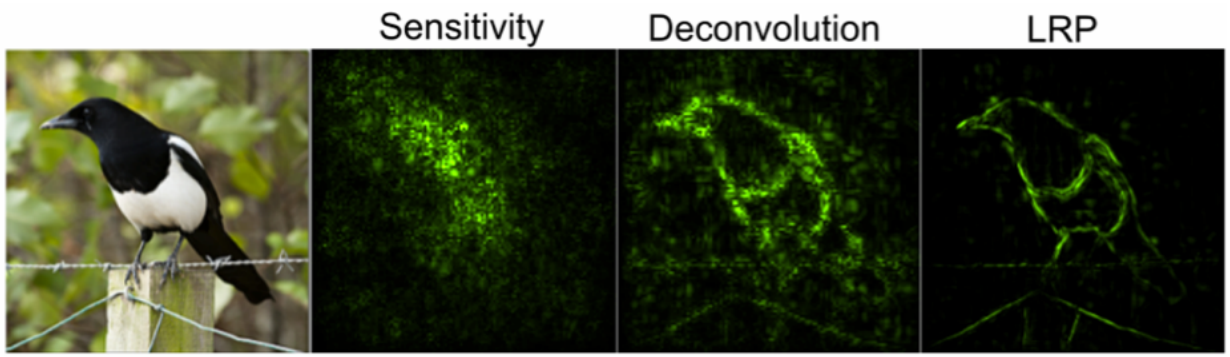
Other Explanation Methods



How to Evaluate Quality of Explanations ?



Compare Explanation Methods



Idea: Compare selectivity (Bach'15, Samek'17):

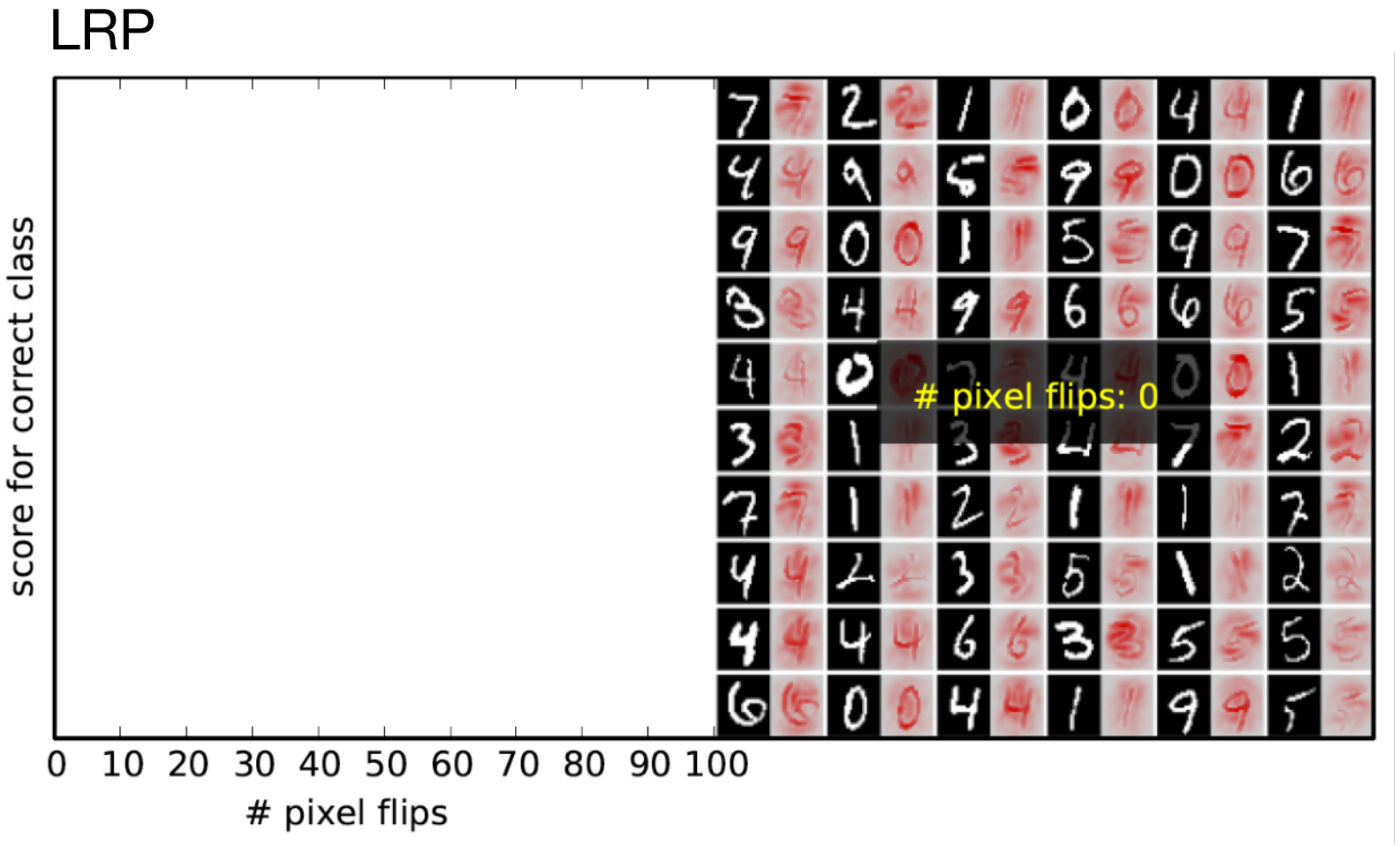
"If input features are deemed relevant, removing them should reduce evidence at the output of the network."

Algorithm ("Pixel Flipping")

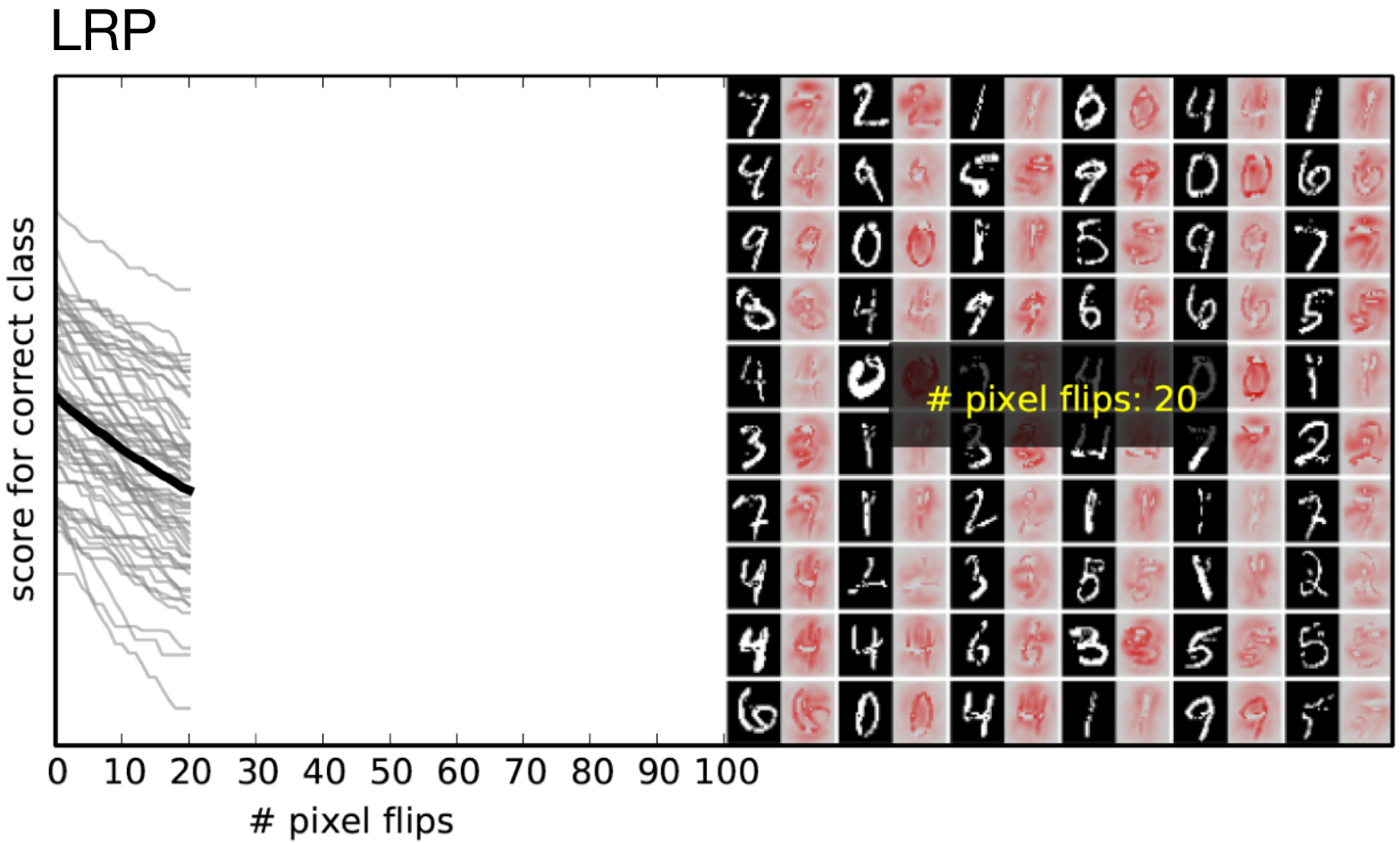
```
Sort pixels / patches by relevance
Iterate
  destroy pixel / patch
  evaluate f(x)
Measure decrease of f(x)
```

Important: Remove information in a non-specific manner (e.g. sample from uniform distribution)

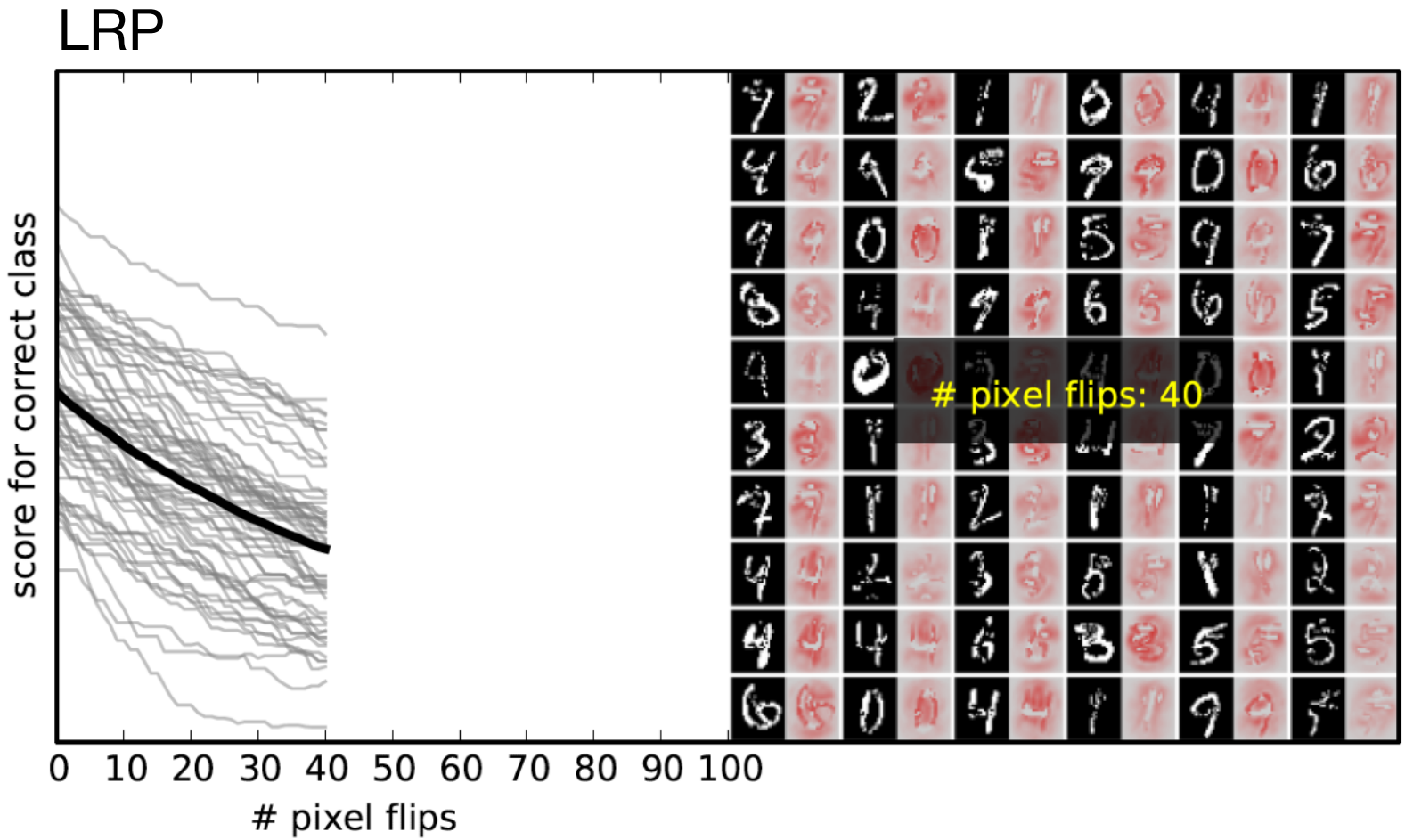
Compare Explanation Methods



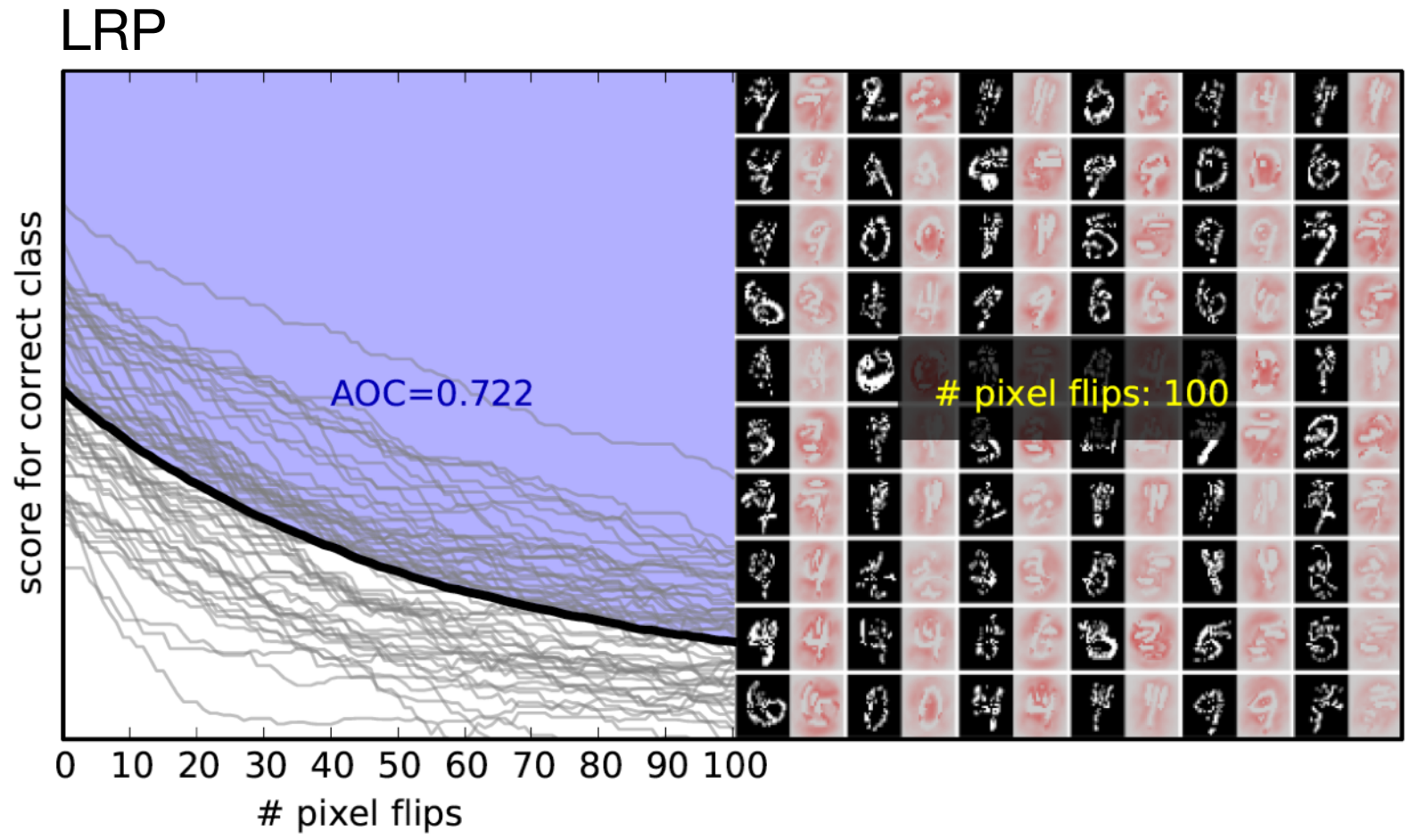
Compare Explanation Methods



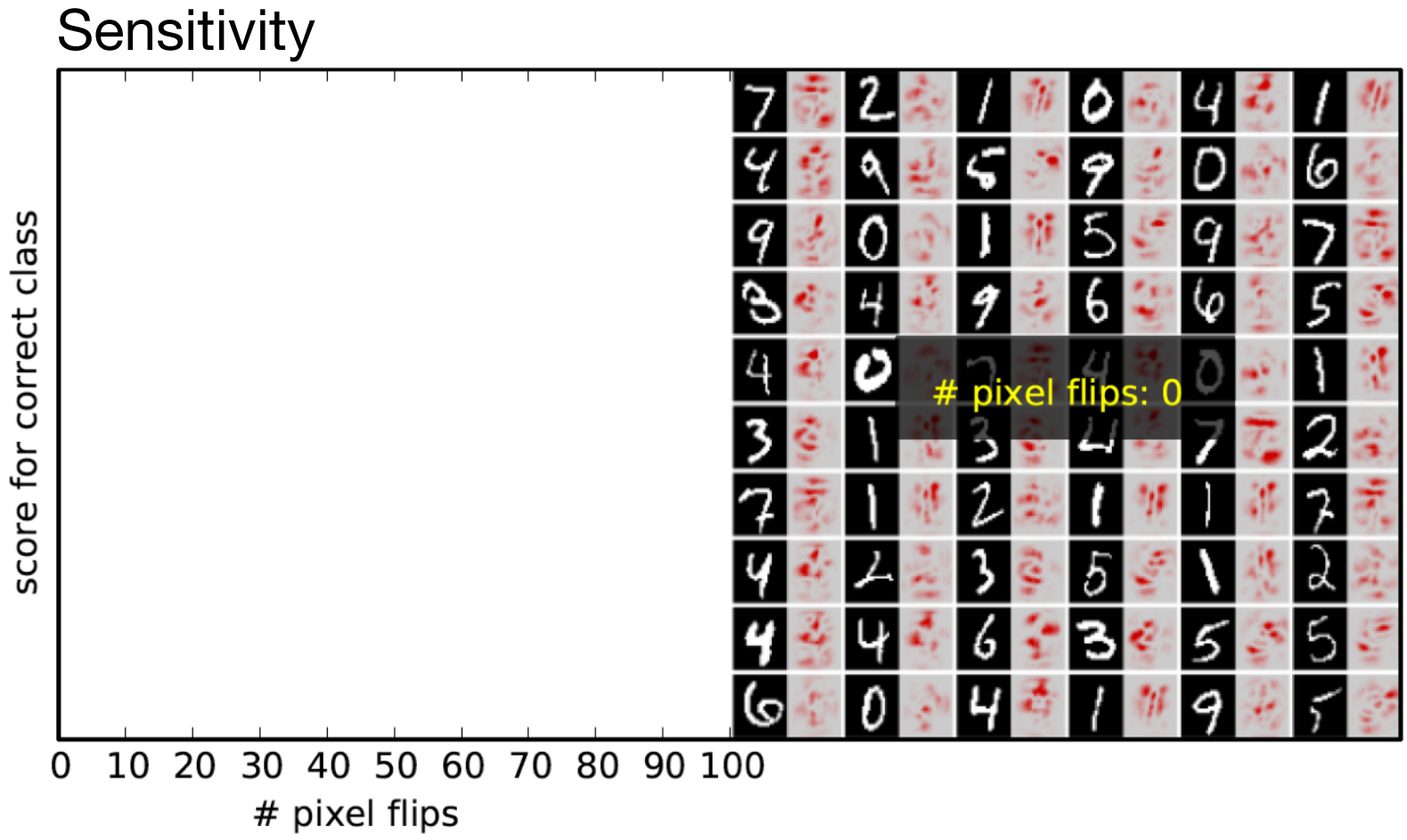
Compare Explanation Methods



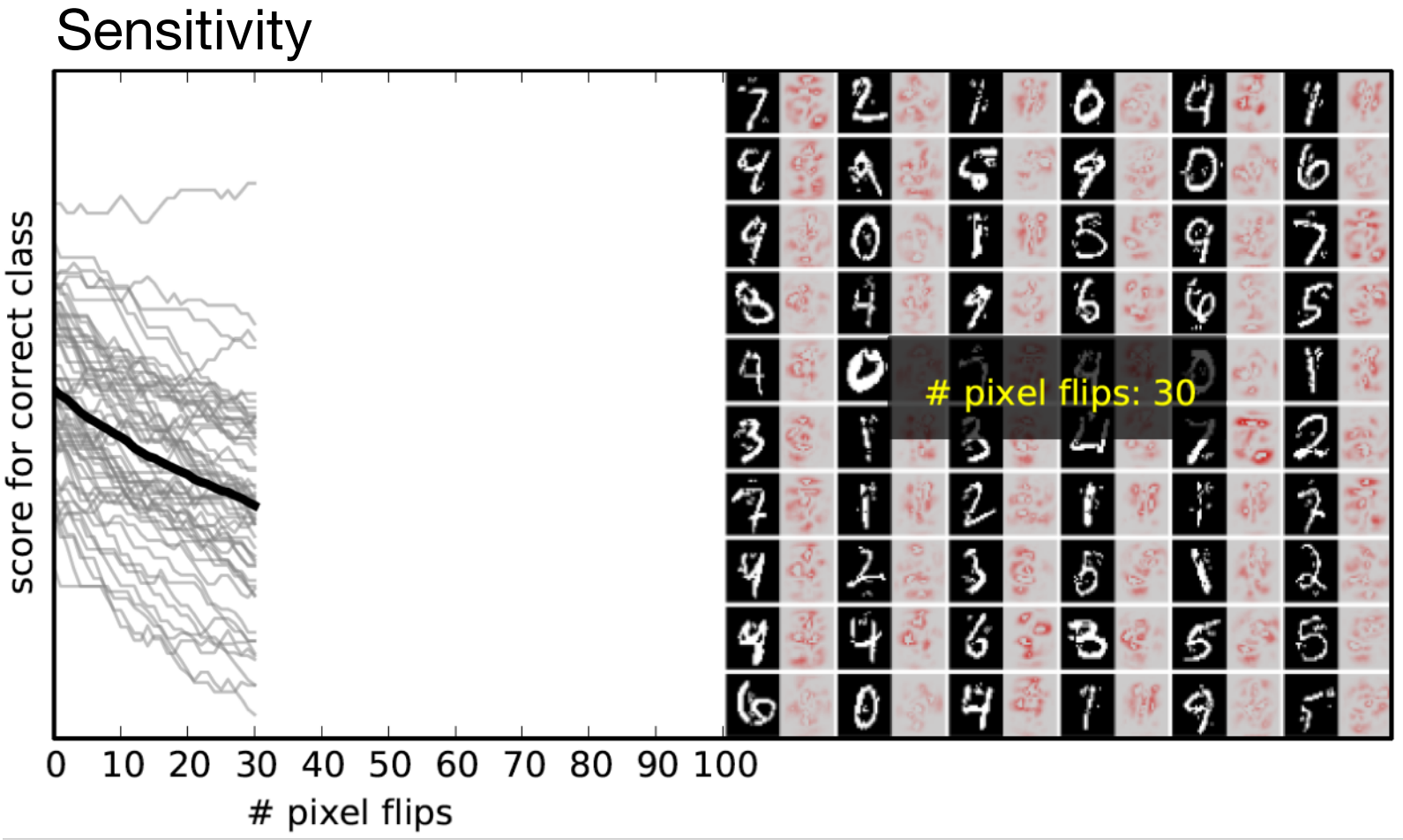
Compare Explanation Methods



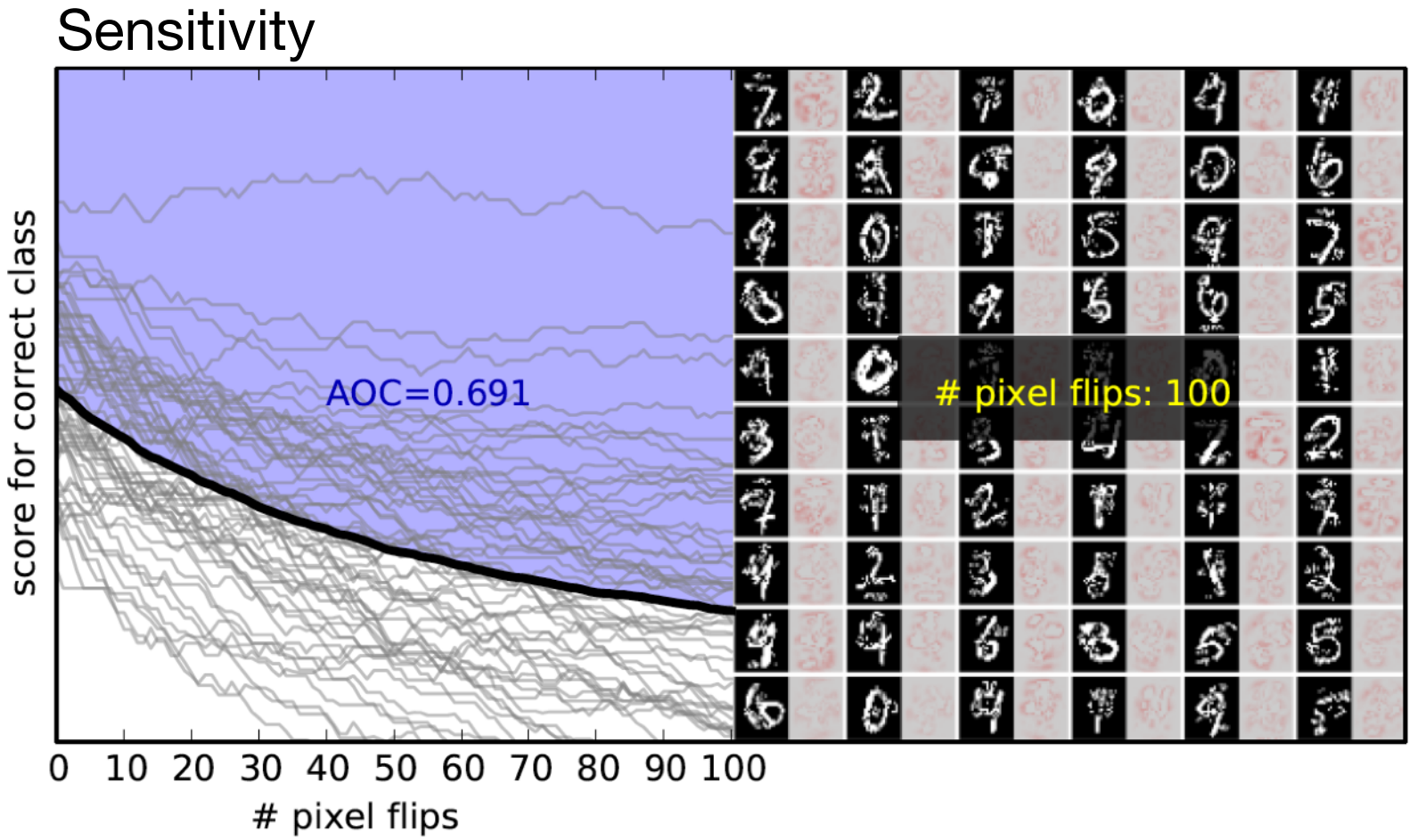
Compare Explanation Methods



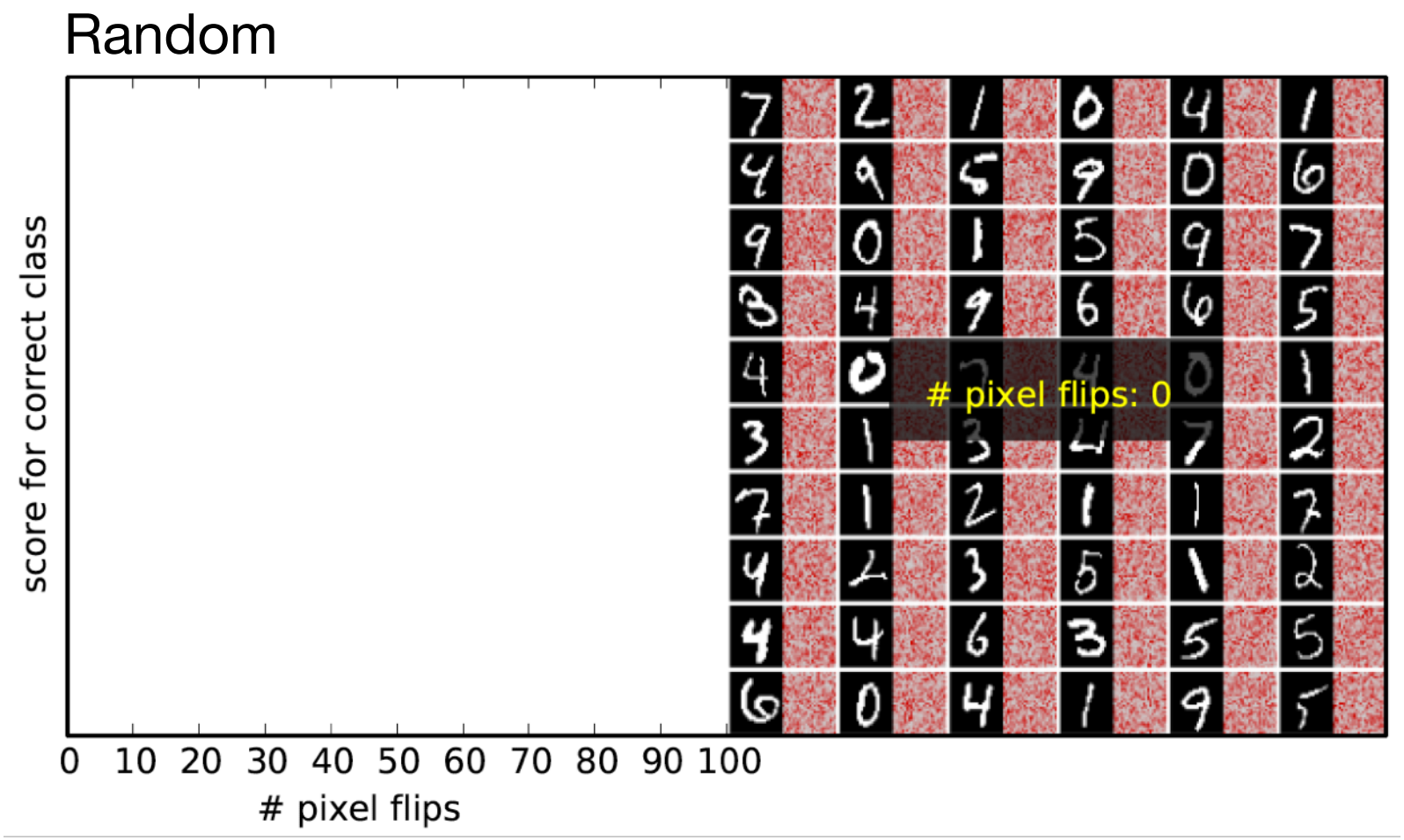
Compare Explanation Methods



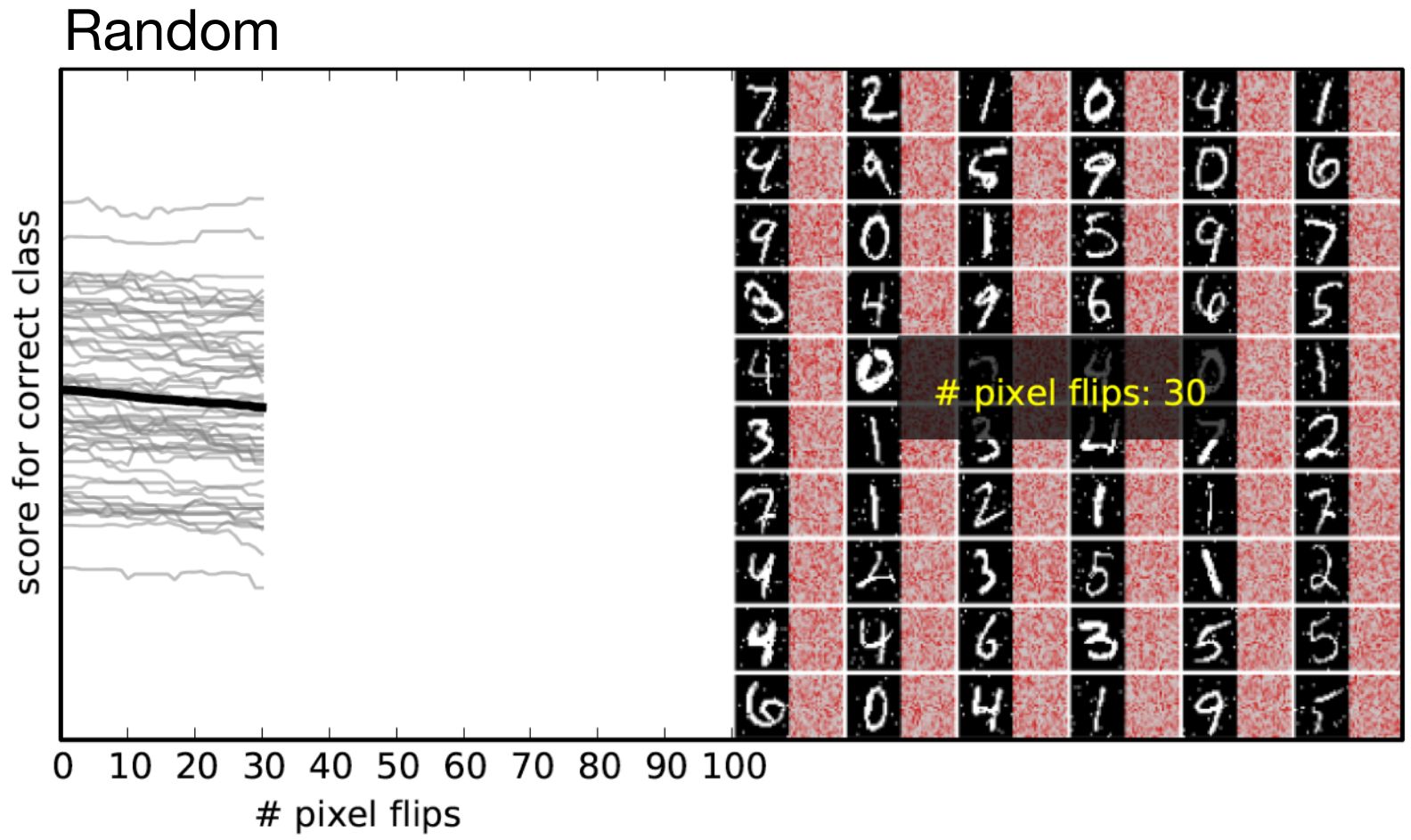
Compare Explanation Methods



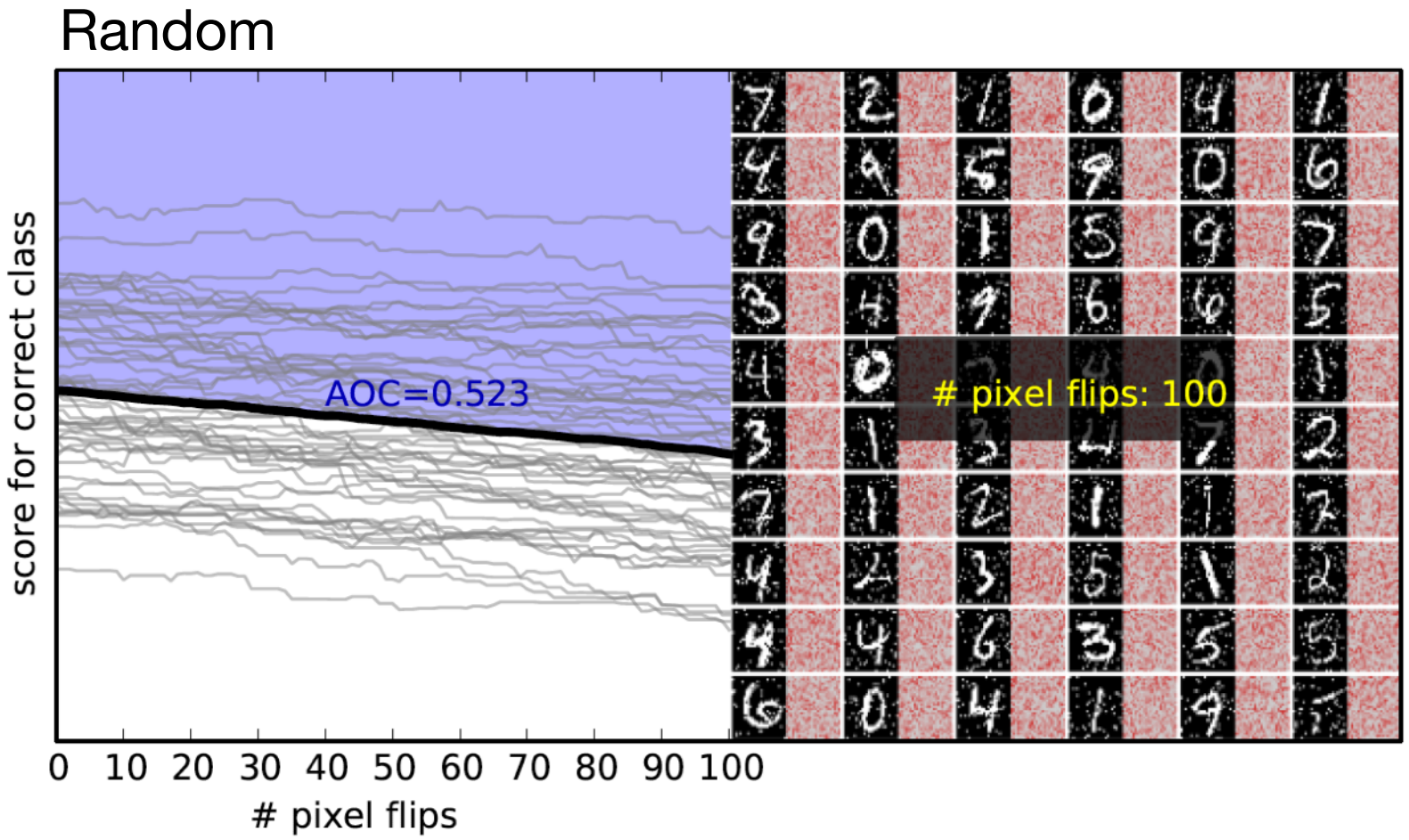
Compare Explanation Methods



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)

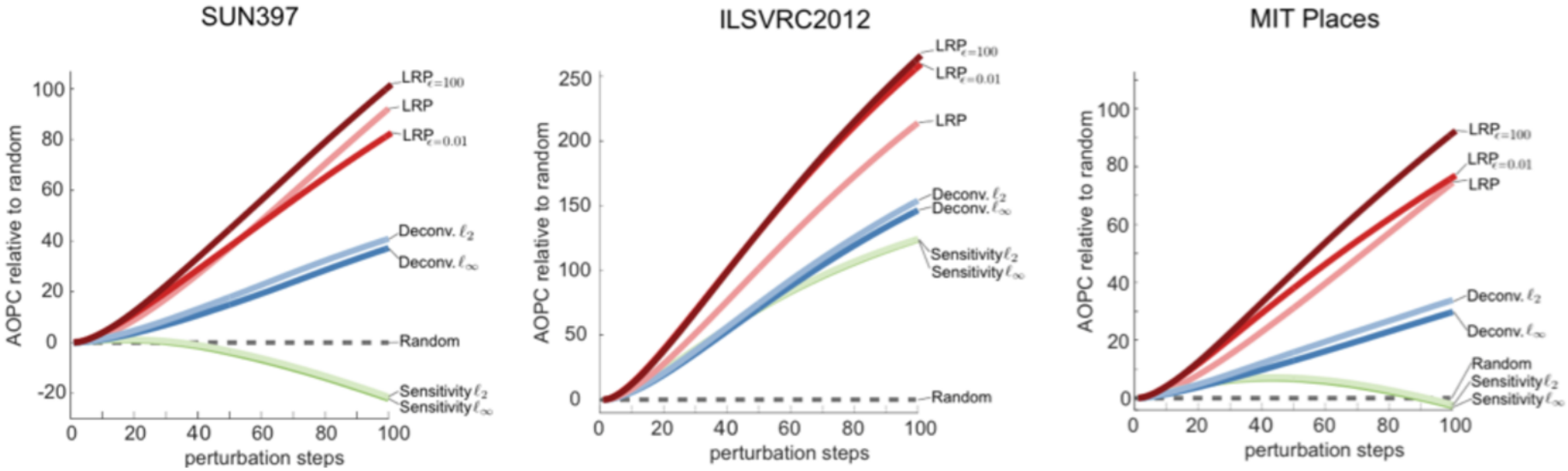
MIT Places

205 scene categories
(2.5 millions of images)

Compare Explanation Methods

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

- ImageNet: Caffe reference model
- Places & SUN: Classifier from MIT
- AOPC averages over 5040 images
- perturb 9 × 9 nonoverlapping regions
- 100 steps (15.7% of the image)
- uniform sampling in pixel space



(Samek et al. 2017)

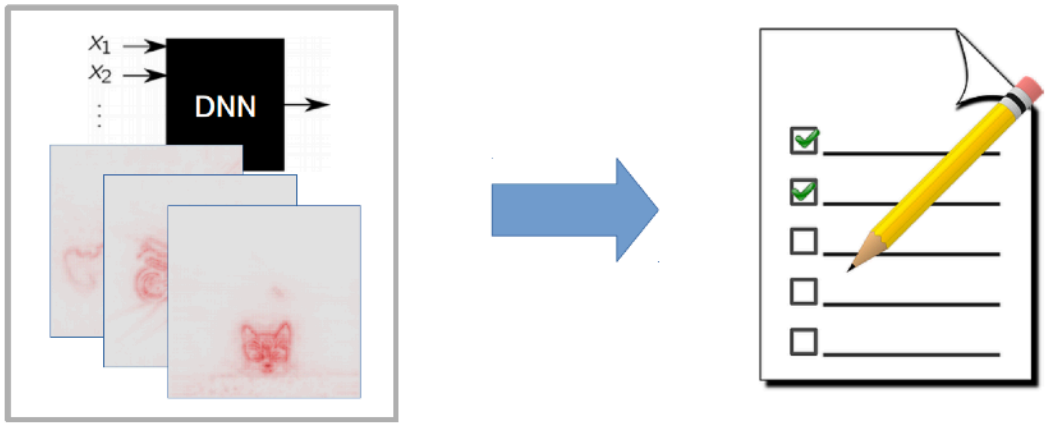
LRP produces better heatmaps

Axiomatic Approach to Interpretability

Idea: Evaluate the explanation technique axiomatically, i.e. it must pass a number of predefined "unit tests".

[Sun'11, Bach'15, Montavon'17, Samek'17, Sundarajan'17, Kindermans'17, Montavon'18].

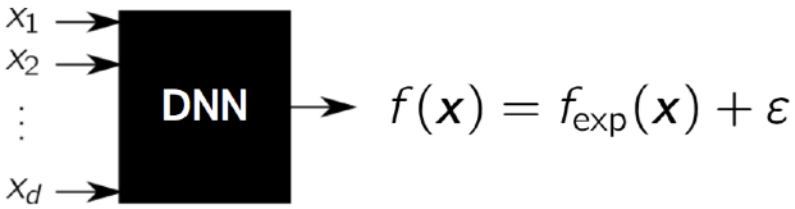
explanation technique



Axiomatic Approach to Interpretability

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).

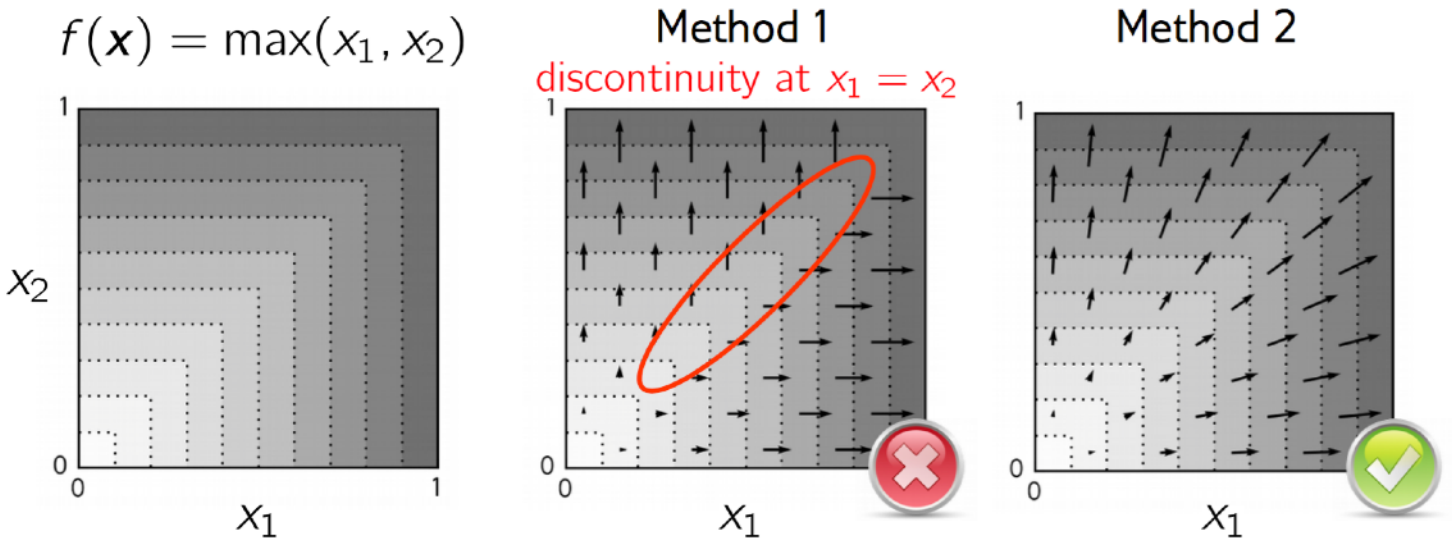
$$\forall_{p=1}^{d} : \; R_p \geq 0$$

Axiomatic Approach to Interpretability

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.

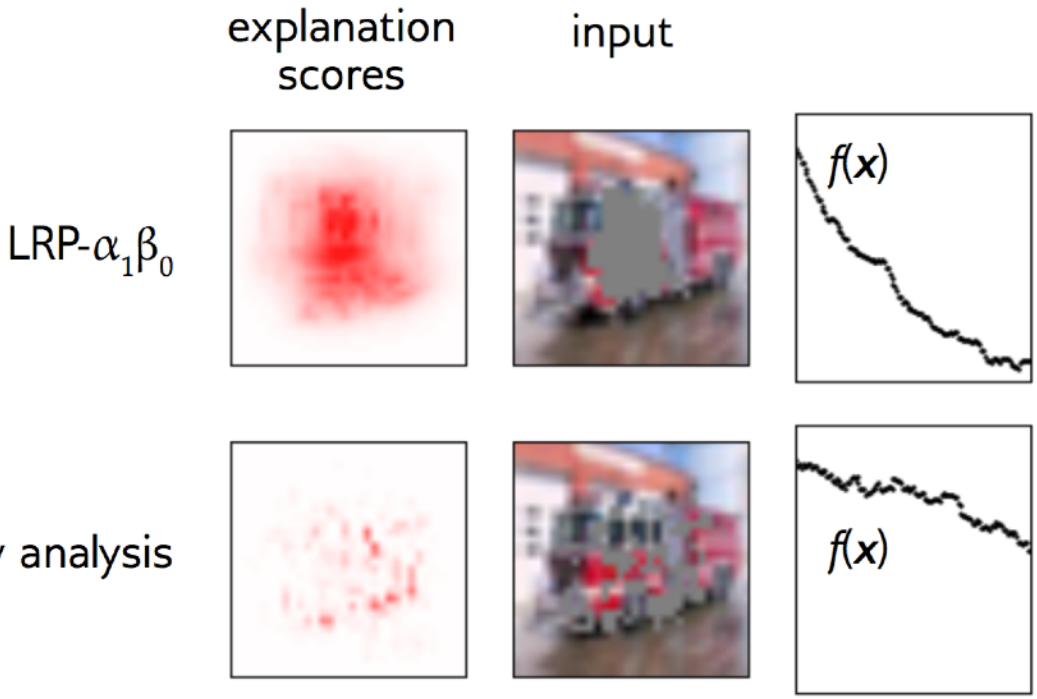
Example:



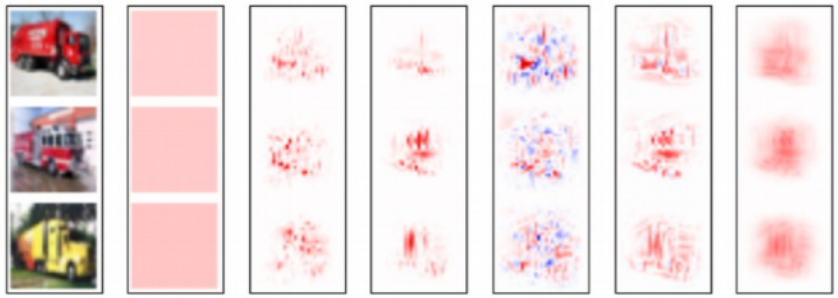
Axiomatic Approach to Interpretability

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

Evaluating with Ground Truth

Idea: Use toy model for evaluating explanations. [Arras'19]

"Explanations should only highlight the parts in the input which are actually important (by design)."

1. LSTM is trained to **add** and **subtract** numbers from first row.
2. Measure correlation with relevance values

$$\begin{pmatrix} 0 & 0 & 0 & n_a & 0 & 0 & 0 & n_b & 0 & 0 & 0 \\ n_1 & \dots & n_{a-1} & 0 & n_{a+1} & \dots & n_{b-1} & 0 & n_{b+1} & \dots & n_T \end{pmatrix}$$

Toy Task Addition	$\rho(R_{\boldsymbol{x_a}}, n_a)$ (in %)	$\rho(R_{\boldsymbol{x_b}}, n_b)$ (in %)
Occlusion	**99.990** (0.004)	**99.990** (0.004)
LRP-half [ACL best paper]	29.035 (9.478)	51.460 (19.939)
LRP-all [ours]	**99.995** (0.002)	**99.995** (0.002)
CD [ICLR oral]	**99.997** (0.002)	**99.997** (0.002)

Toy Task Subtraction	$\rho(R_{\boldsymbol{x_a}}, n_a)$ (in %)	$\rho(R_{\boldsymbol{x_b}}, n_b)$ (in %)
Occlusion	99.0 (2.0)	-69.0 (19.1)
LRP-half [ACL best paper]	7.7 (15.3)	-28.9 (6.4)
LRP-all [ours]	**98.5** (3.5)	**-99.3** (1.3)
CD [ICLR oral]	-25.9 (39.1)	-50.0 (29.2)

Summary Evaluation

LRP heatmaps are informative, truthworthy and fulfil important axioms.

Furthermore, LRP produces qualitative and quantitatively better explanations than sensitivity analysis, deconvolution, context decomposition, gradient times input, occlusion and …

Let's now see some applications!

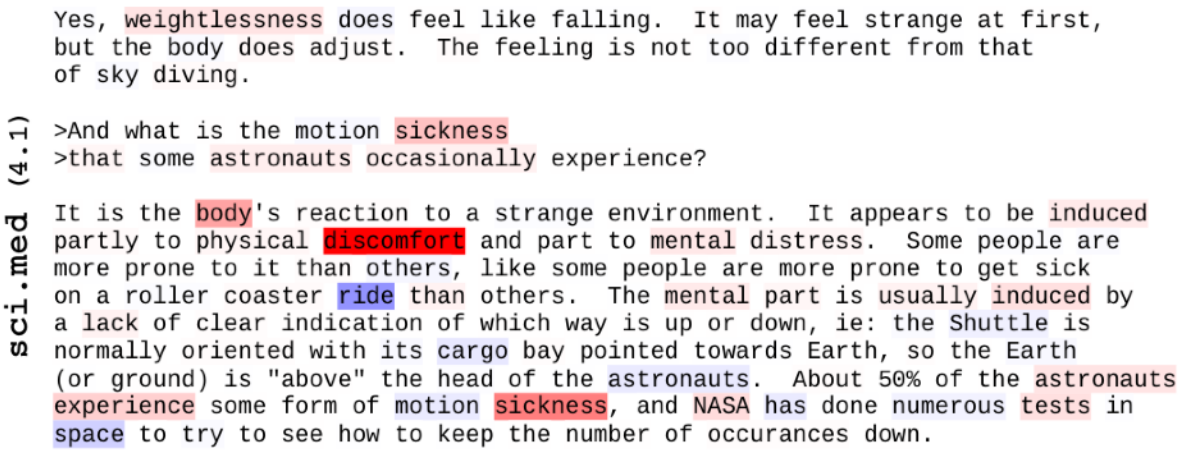
Application of LRP
Compare models

Application: Compare Classifiers

word2vec/CNN:

Performance: 80.19%

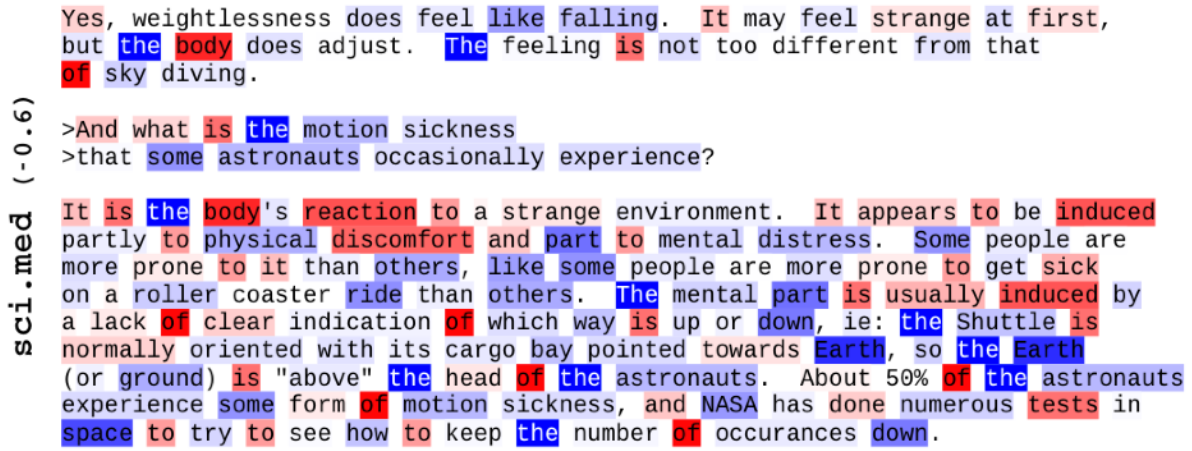
Strategy to solve the problem: identify semantically meaningful words related to the topic.



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

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 & 2017)

Application of LRP
Quantify Context Use

Application: Measure Context Use

how important is context ?

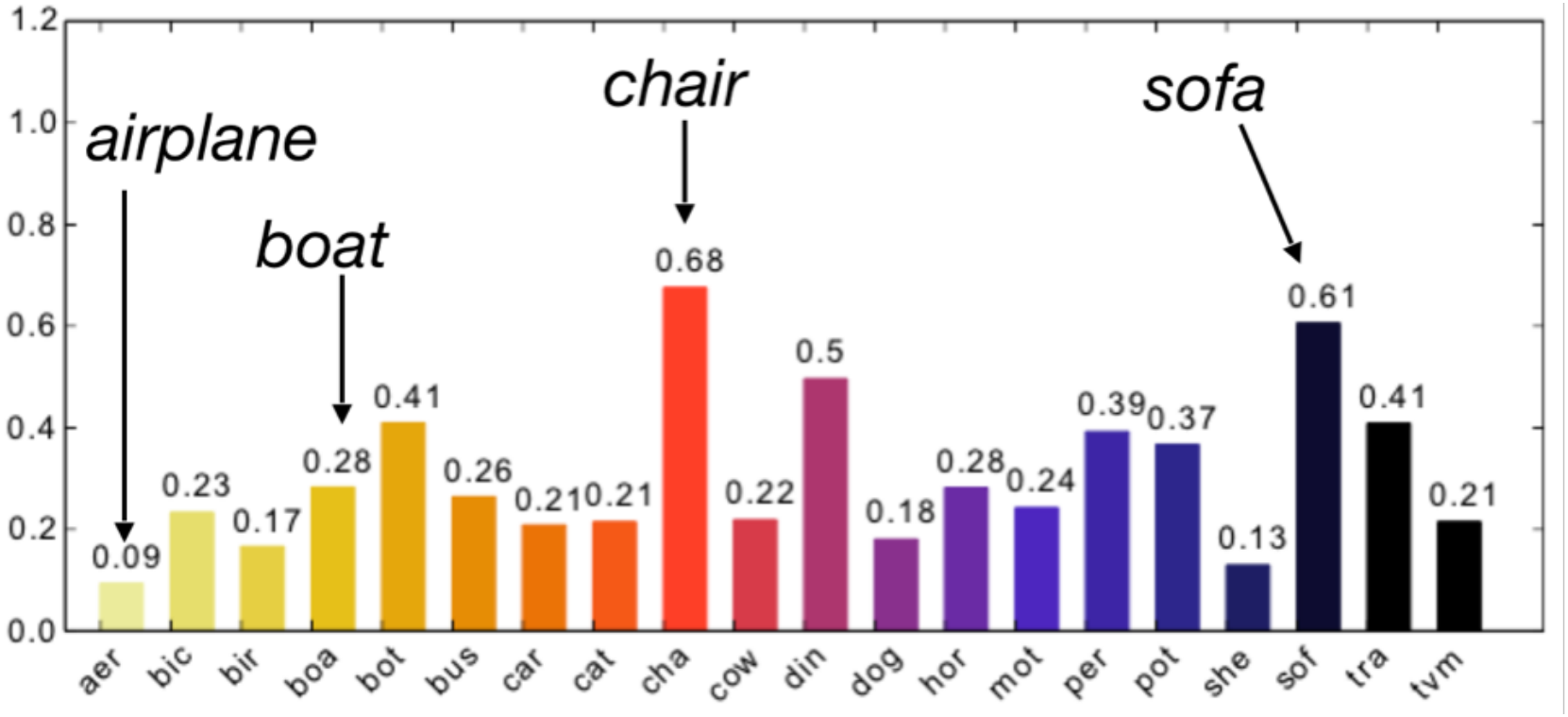
how important is context ?

classifier

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

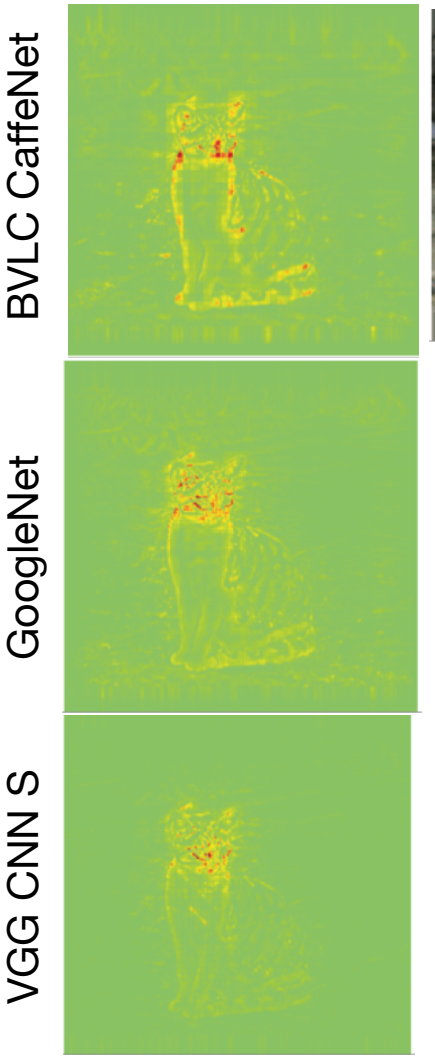
Application: Measure Context Use

- BVLC reference model + fine tuning
- PASCAL VOC 2007

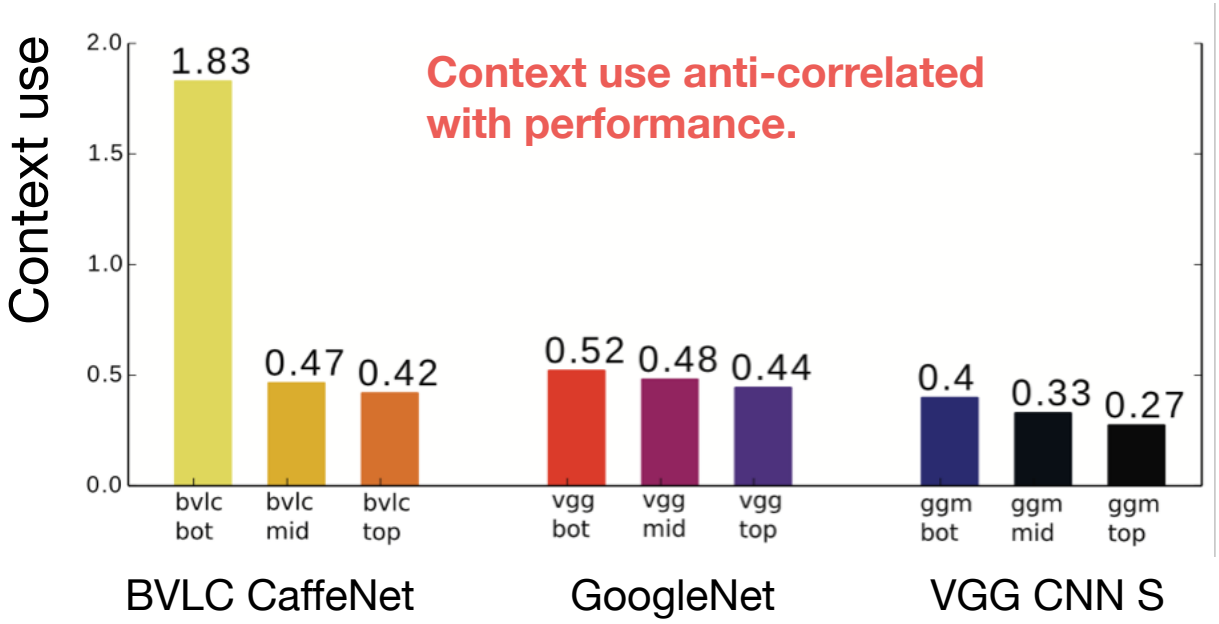


(Lapuschkin et al., 2016)

Application: Measure Context Use



- Differen models (BVLC CaffeNet, GoogleNet, VGG CNN S)
- ILSVCR 2012



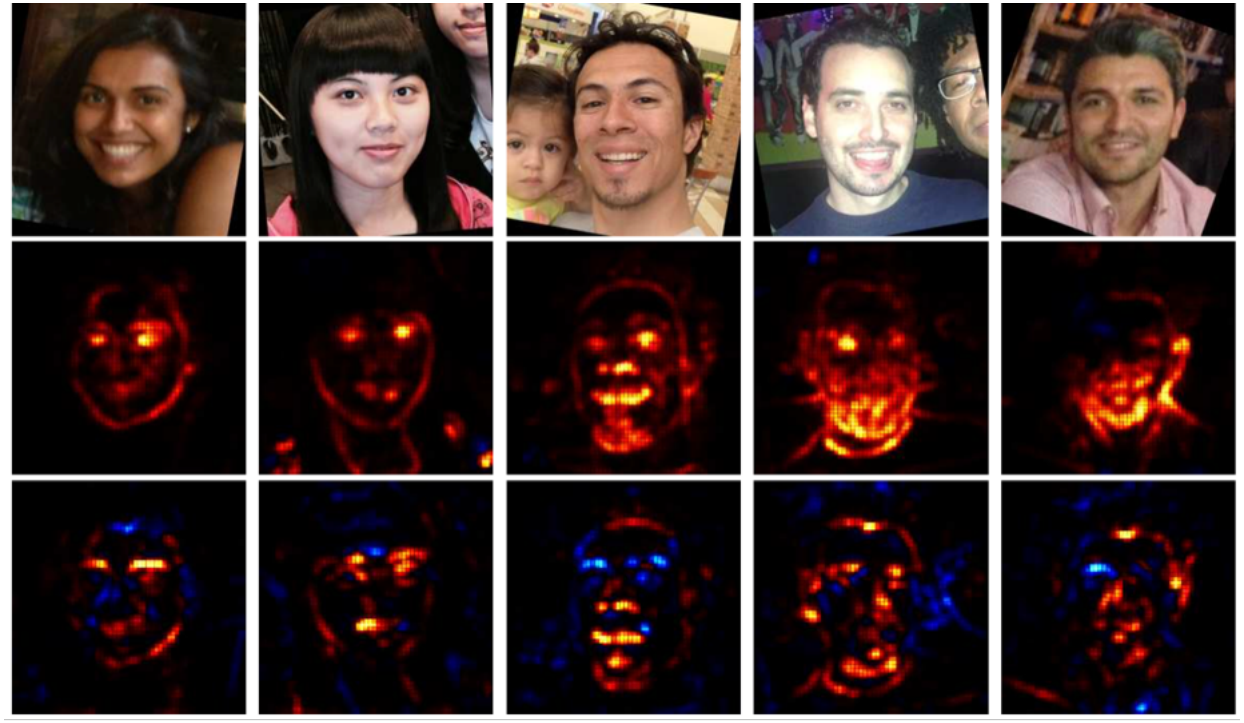
Context use anti-correlated with performance.

(Lapuschkin et al. 2016)

Application of LRP
Detect Biases

Application: Face analysis

Gender classification

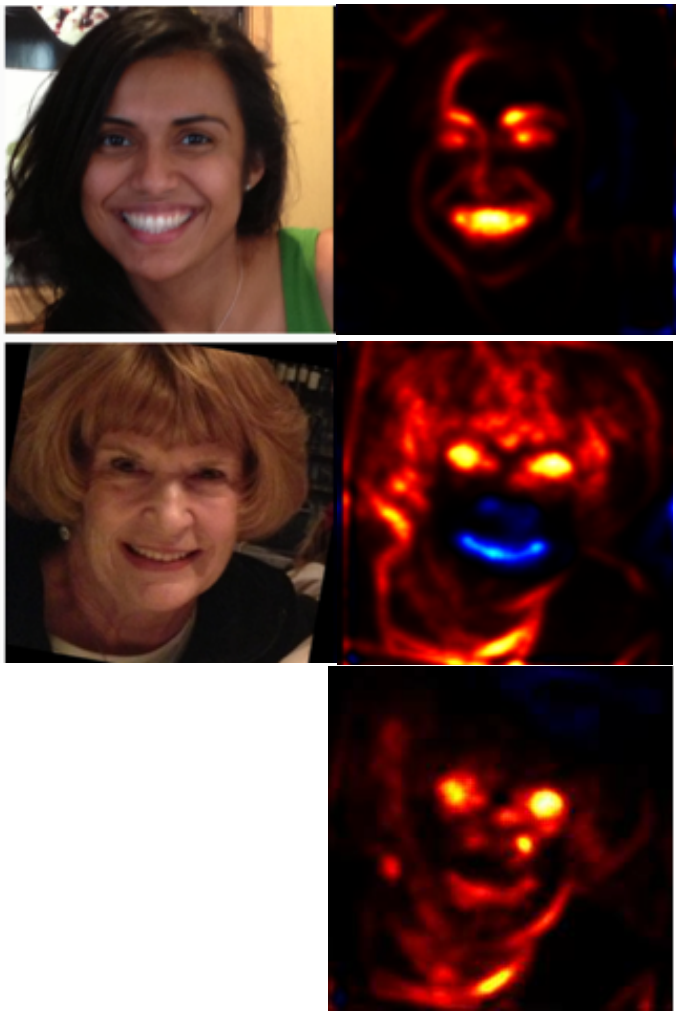


Strategy to solve the problem: Focus on chin / beard, eyes & hear,
but without pretraining the model overfits

(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

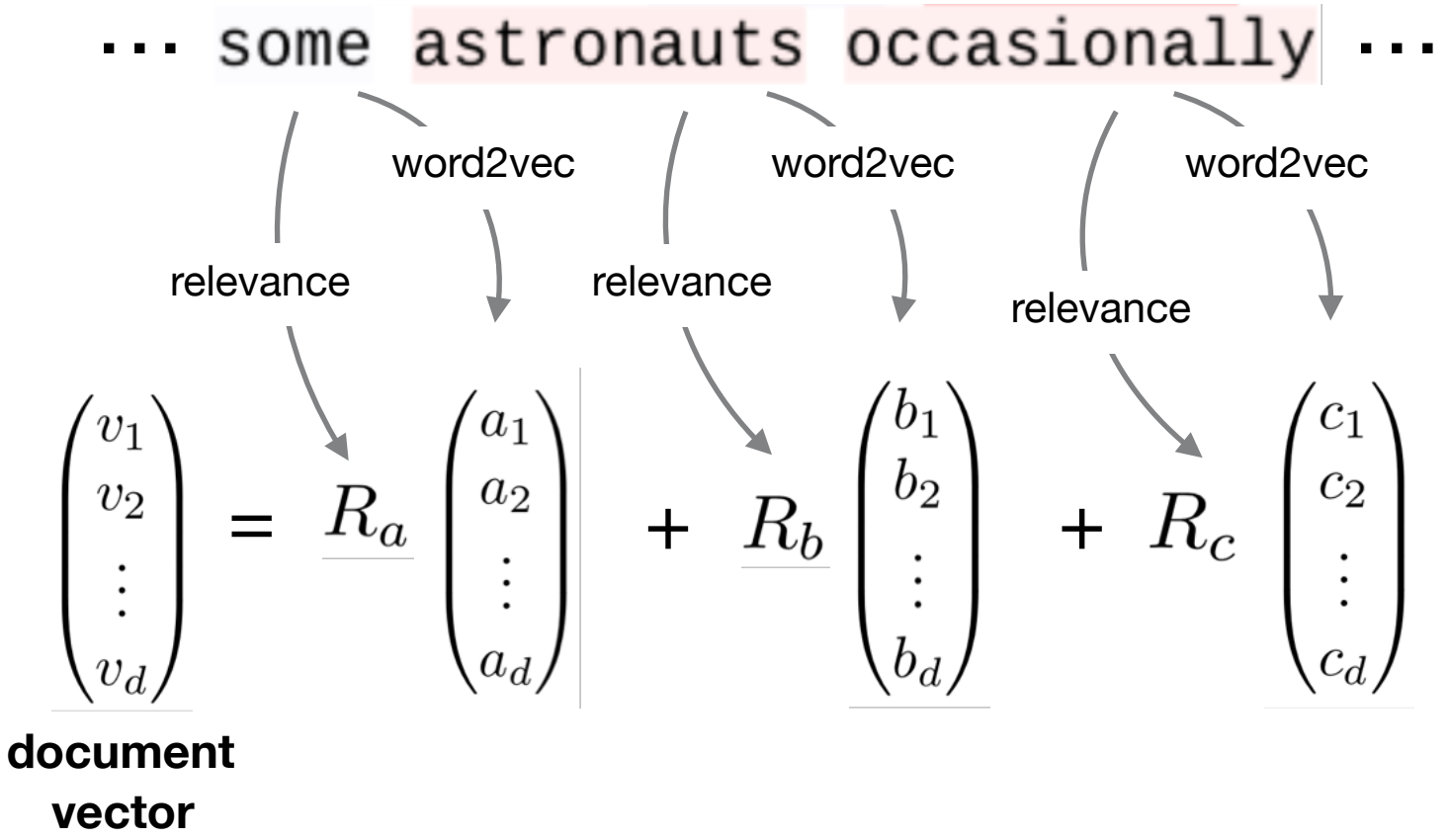
laughing speaks against 60+ (i.e., model learned that old people do not laugh)

pretraining on IMDB-WIKI

(Lapuschkin et al., 2017)

Application of LRP
Relevance-Based Filtering

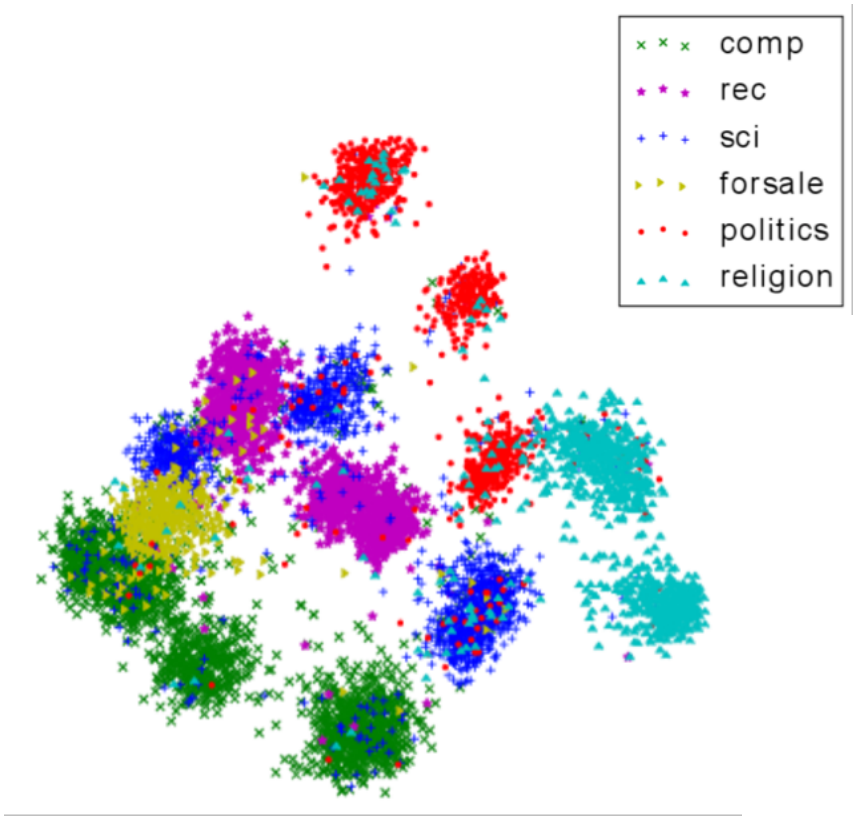
Application: Learn new Representations



(Arras et al. 2016 & 2017)

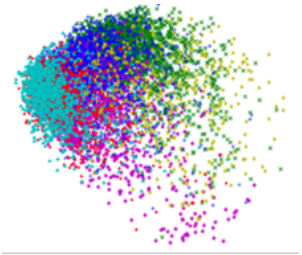
Application: Learn new Representations

2D PCA projection of document vectors



uniform

TFIDF



Document vector computation is unsupervised (given we have a classifier).

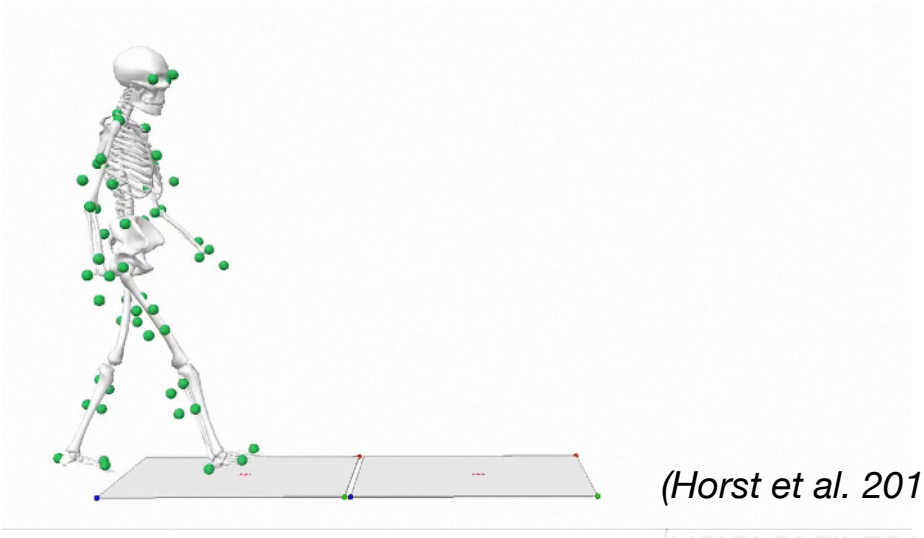
(Arras et al. 2016 & 2017)

Application of LRP
Understand Model

Application: Gait Analysis

Our approach:

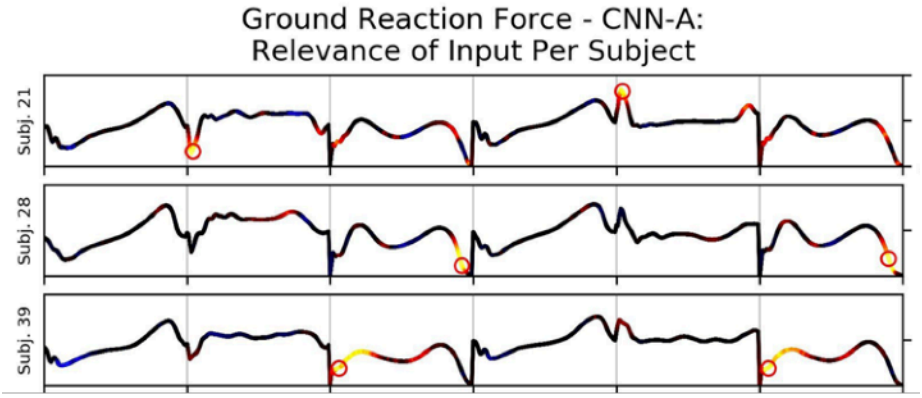
- Classify & explain individual gait patterns
- Important for understanding diseases such as Parkinson

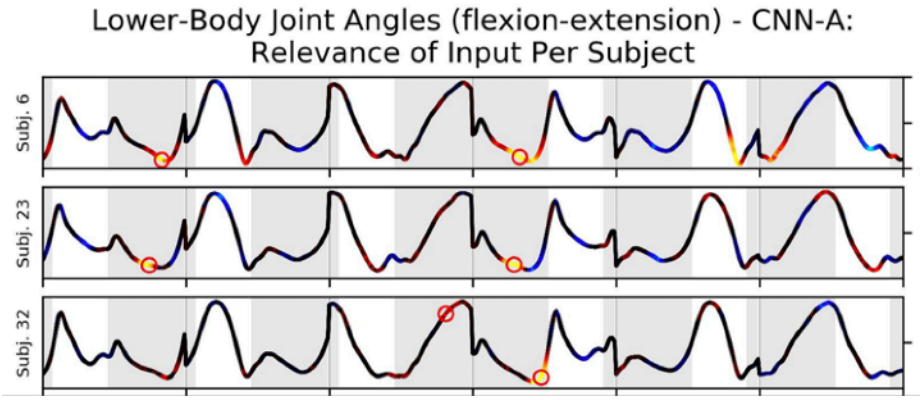


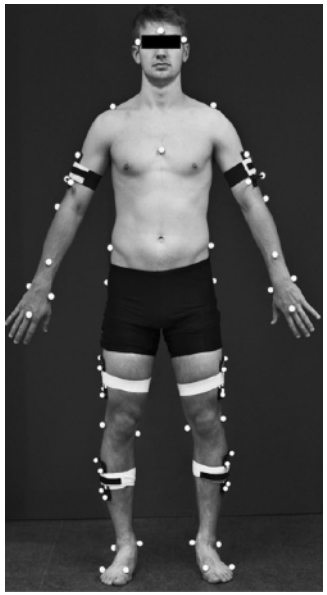
(Horst et al. 2019)

Application: Gait Analysis

Model	Ground Reaction Forces [%]	Joint Angles Full-Body [%]	Joint Angles Full-Body (flex.-ext.) [%]	Joint Angles Lower-Body [%]	Joint Angles Lower-Body (flex.-ext.) [%]
CNN-A	99.1 (0.8)	**100.0** (**0.0**)	95.6 (1.7)	99.9 (0.3)	92.0 (3.9)





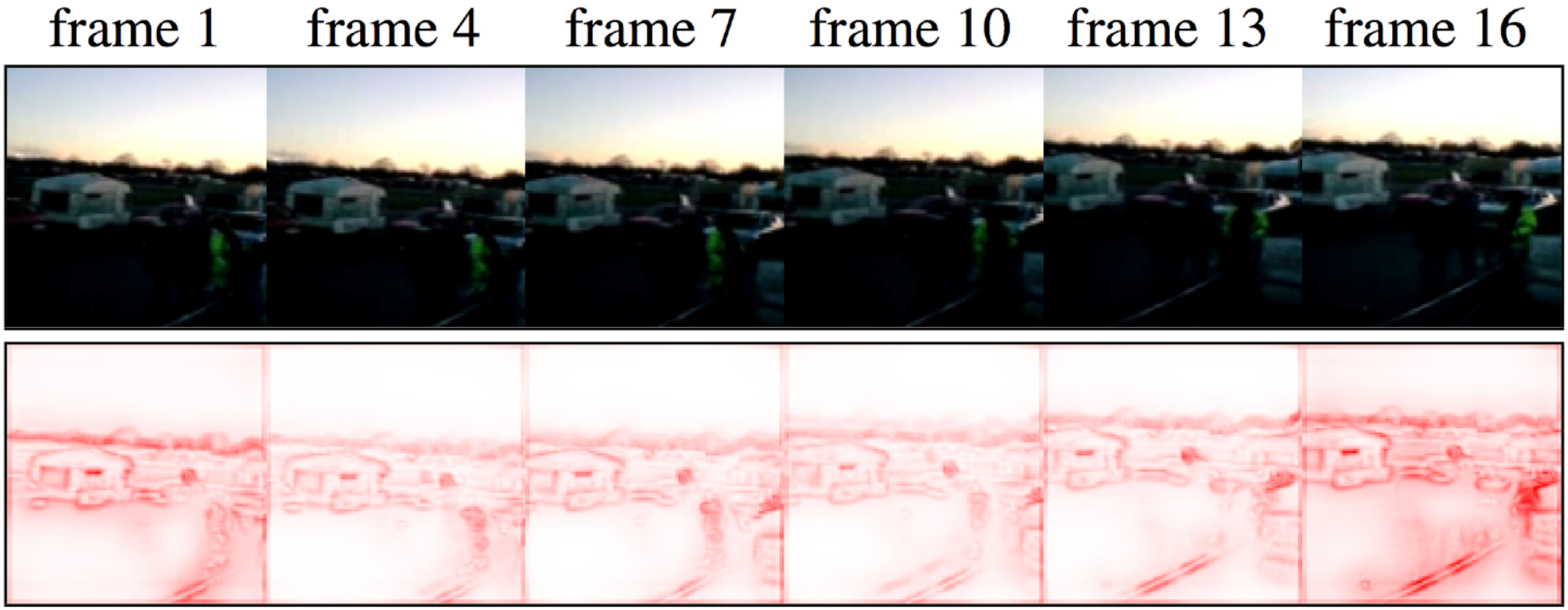


Insights for subject 6

- extension of the ankle during the terminal stance phase
- flexion of the knee and hip during the initial contact of the right leg

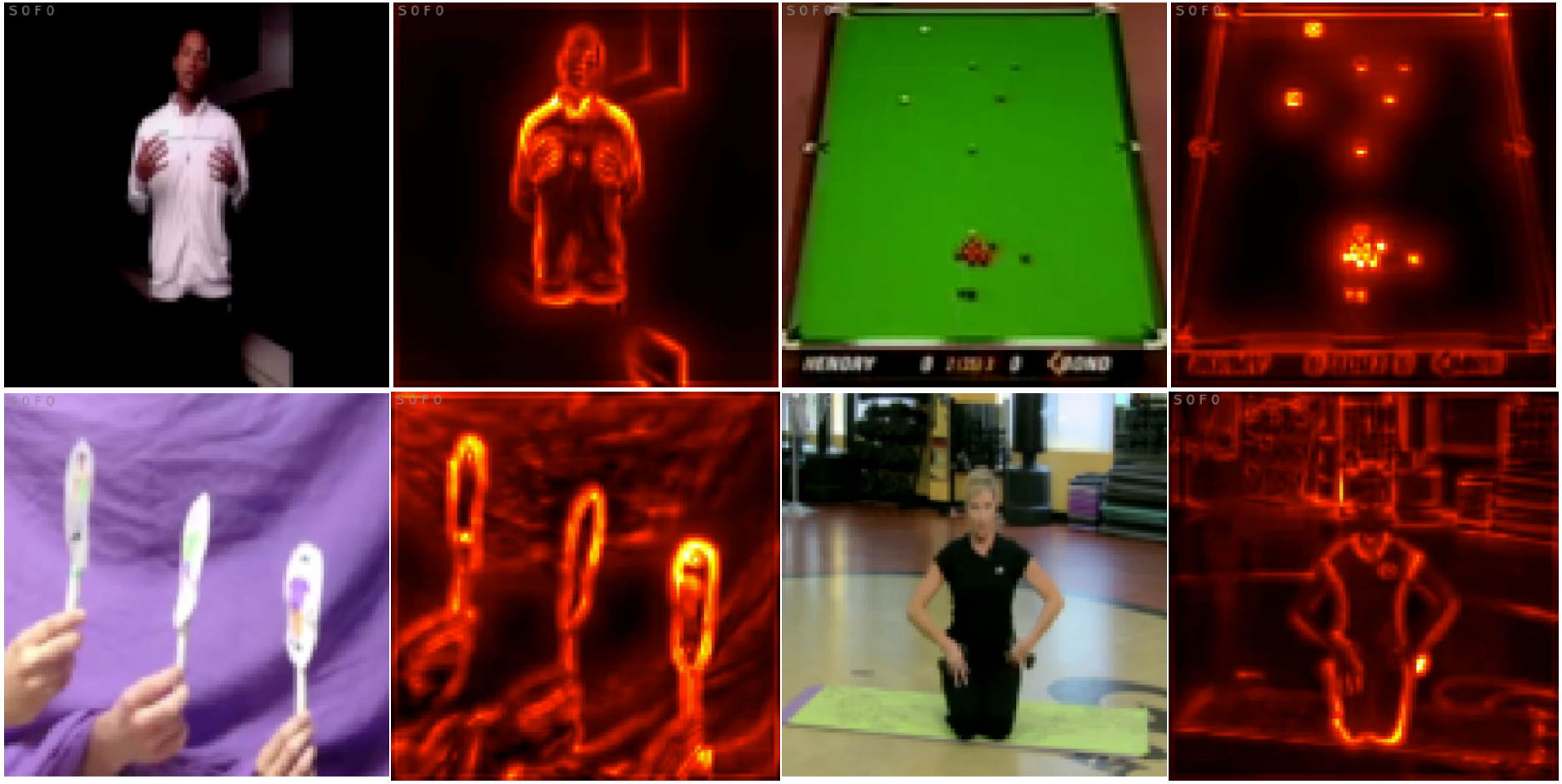
Application: Understand the model

- 3-dimensional CNN (C3D)
- trained on Sports-1M
- explain predictions for 1000 videos from the test set



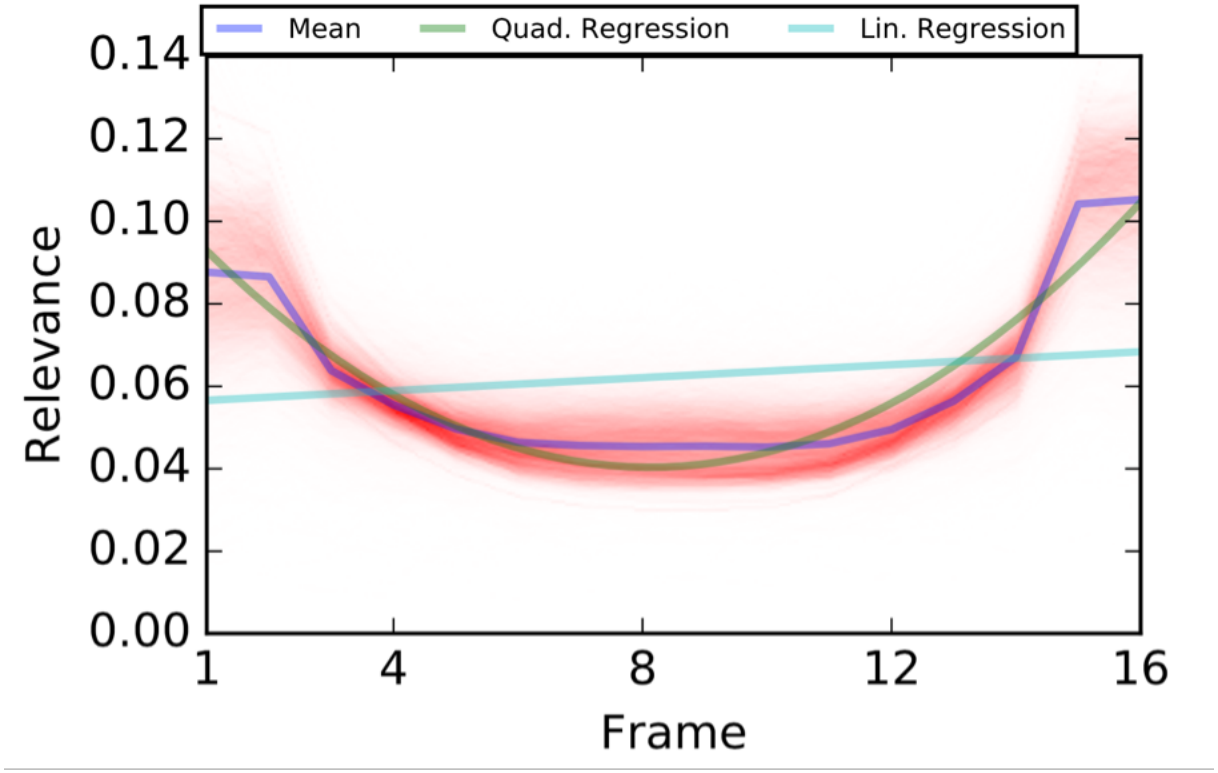
(Anders et al., 2018)

Application: Understand the model



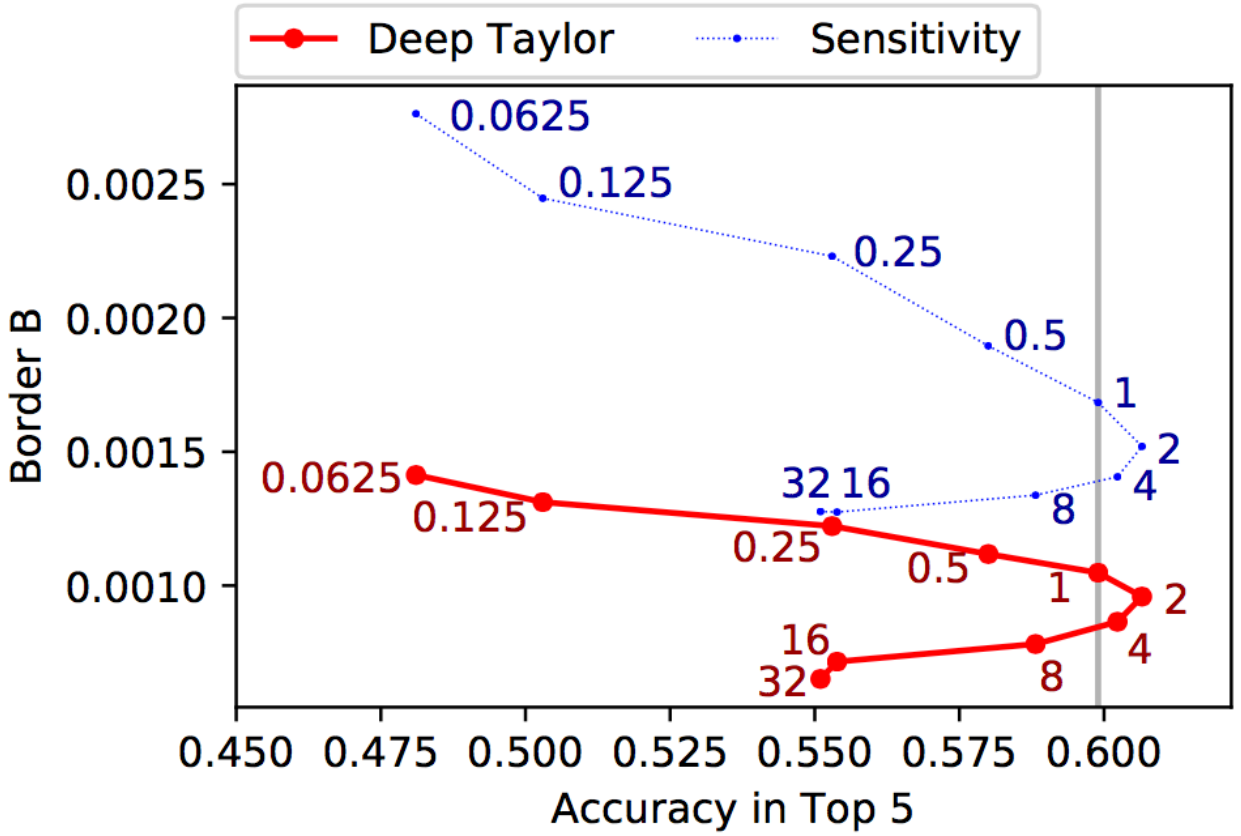
(Anders et al., 2018)

Application: Understand the model



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

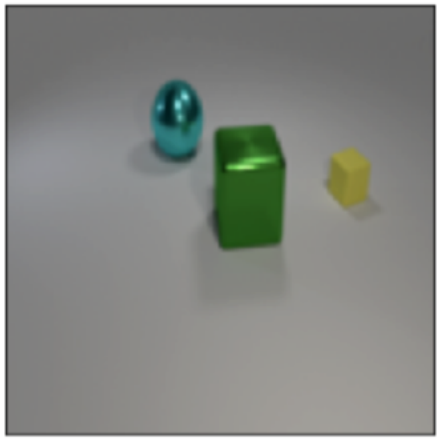
Application: Understand the model

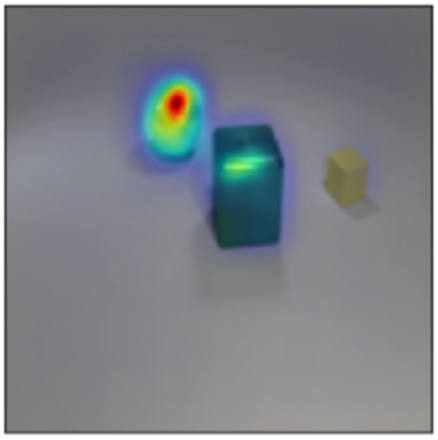


Idea: Play video in fast forward (without retraining) and then the classification accuracy improves.

Application: Understand the model

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 ?





- reimplement model of (Santoro et al., 2017)
- test accuracy of 91,0%
- CLEVR dataset

Question Type	LRP
equal_size	size 2279, the 1939, cylinder 1201 (0, 1000, 2000, 3000, 4000)
greater_than	number 2177, more 1656, than 1515 (0, 1000, 2000, 3000, 4000)
query_color	color 12819, what 6837, is 5735 (0, 10000, 20000)
count	what 10823, how 9310, many 5691 (0, 10000, 20000, 30000)

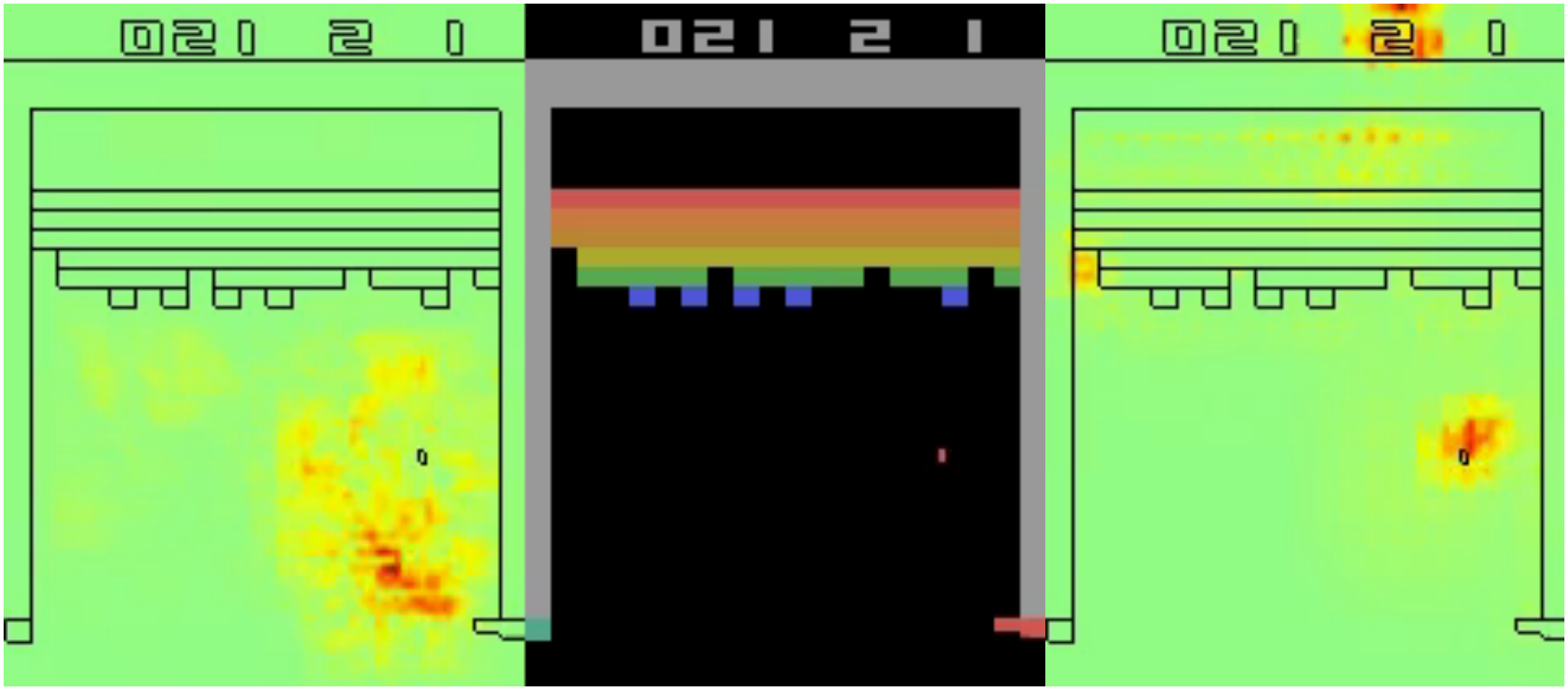
model understands the question and correctly identifies the object of interest

(in prep)

Application: Understand the model

Sensitivity Analysis

LRP



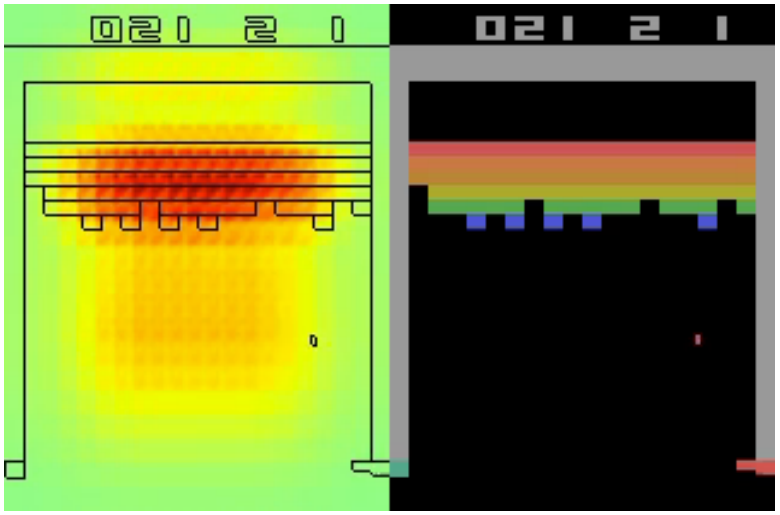
does not focus on where the ball is, but on where the ball could be in the next frame

LRP shows that that model tracks the ball

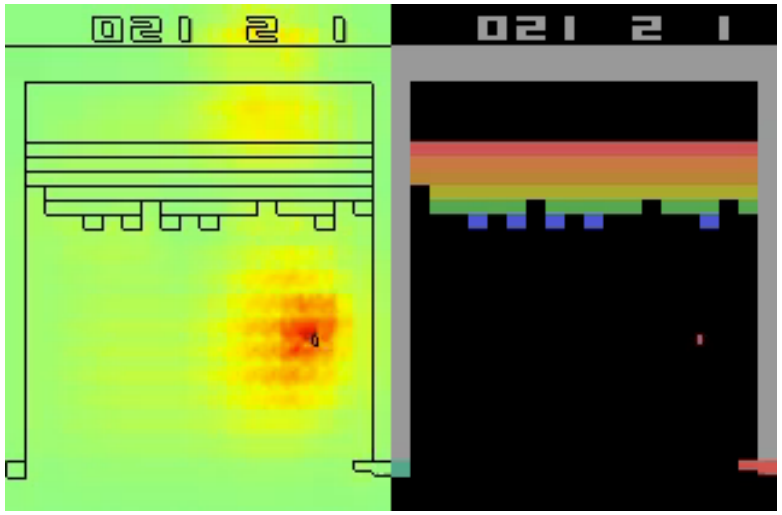
(Lapuschkin et al., 2019)

Application: Understand the model

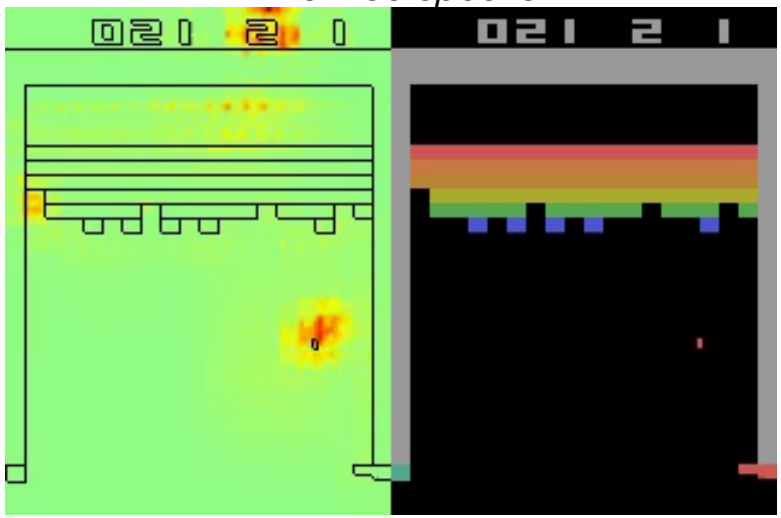
After 0 epochs



After 25 epochs

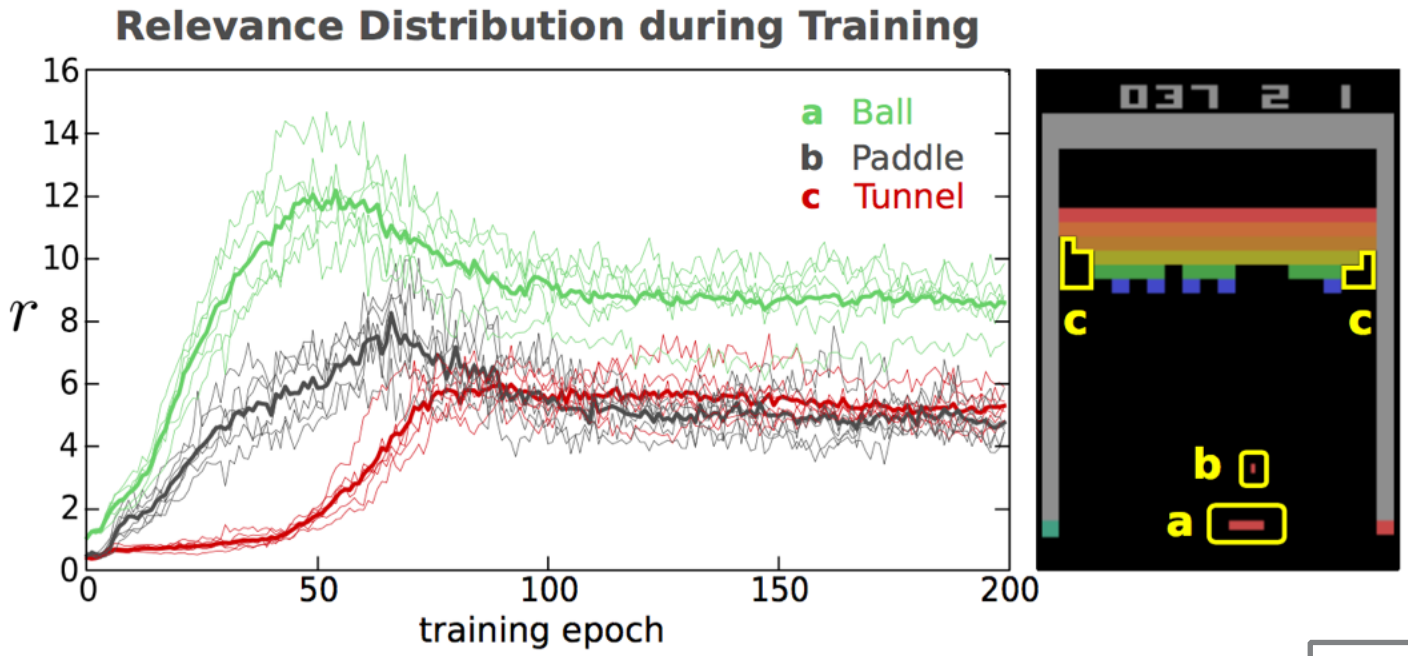


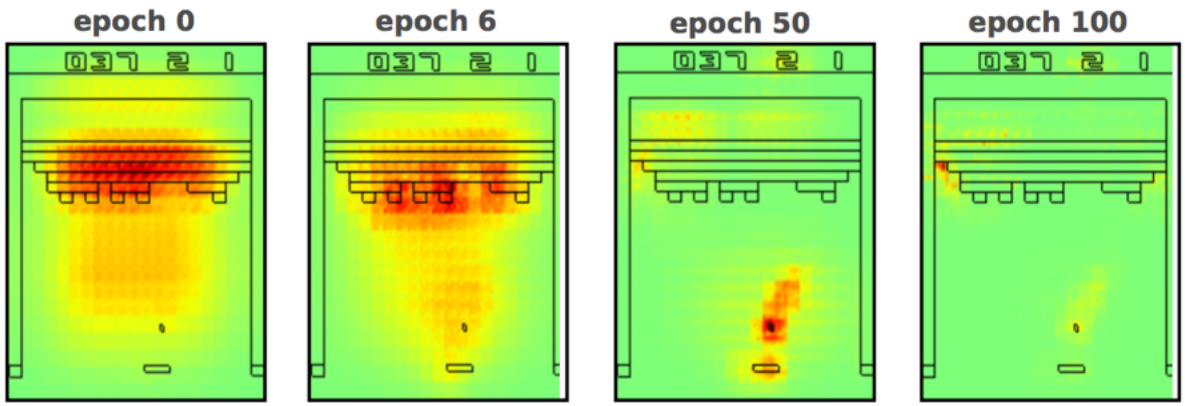
After 195 epochs



(Lapuschkin et al., 2019)

Application: Understand the model

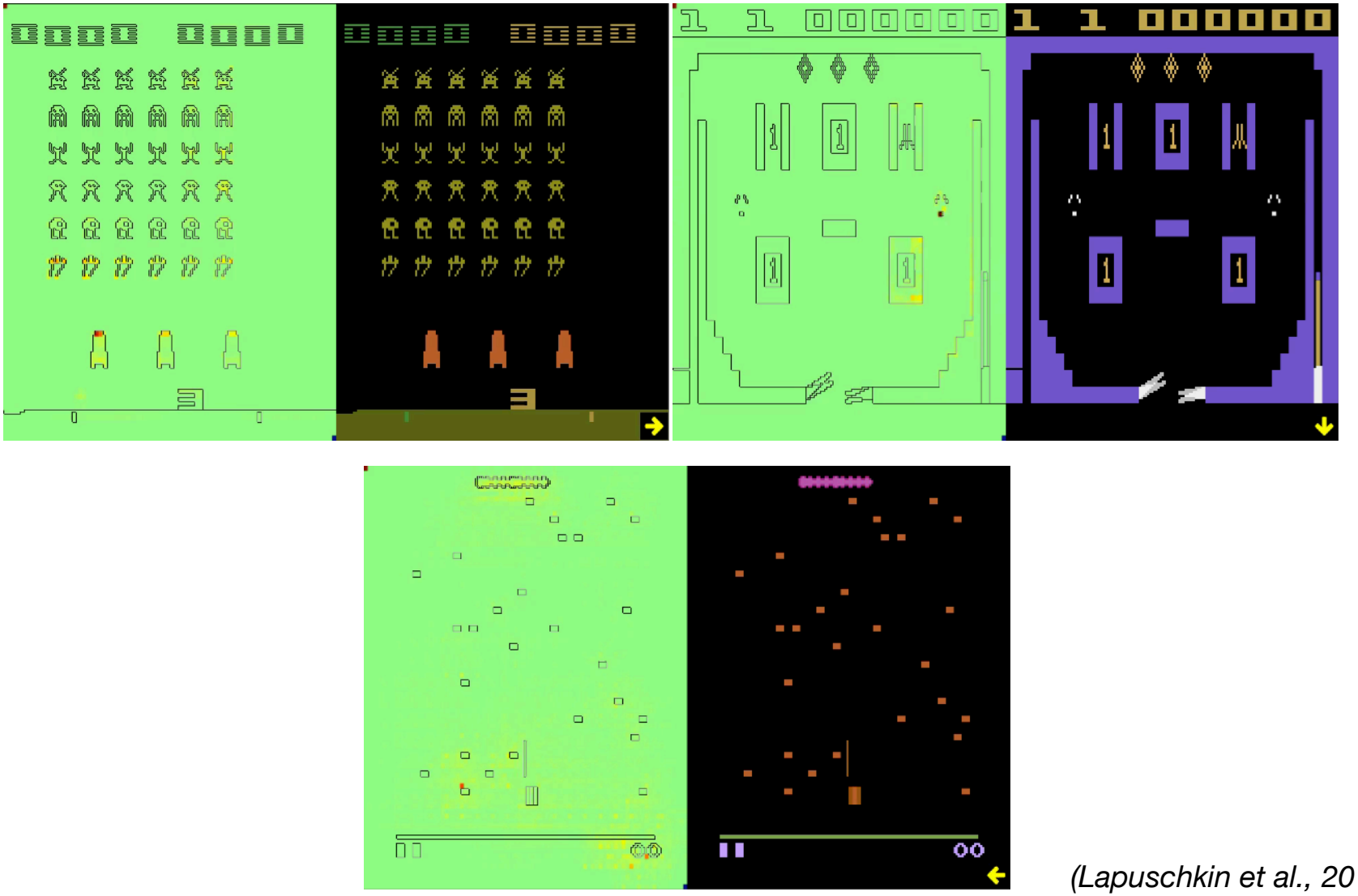




model learns
1. track the ball
2. focus on paddle
3. focus on the tunnel

(Lapuschkin et al., 2019)

Application: Understand the model



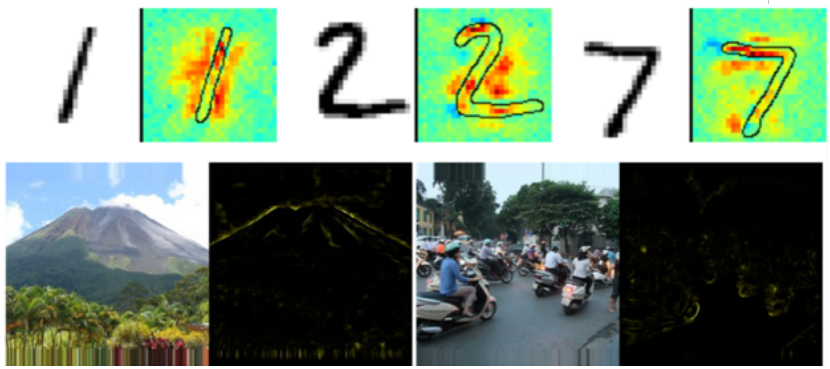
(Lapuschkin et al., 2019)

More information

Visit:

`http://www.heatmapping.org`

- Tutorials
- Software
- Online Demos



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`

References

Opinion Paper

S Lapuschkin, S Wäldchen, A Binder, G Montavon, W Samek, KR Müller. Unmasking Clever Hans Predictors and Assessing What Machines Really Learn. *Nature Communications*, 10:1096, 2019.

Tutorial / Overview Papers

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

W Samek, T Wiegand, and KR Müller, Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models, ITU Journal: ICT Discoveries - Special Issue 1 - The Impact of Artificial Intelligence (AI) on Communication Networks and Services, 1(1):39-48, 2018.

Methods Papers

S Bach, A Binder, G Montavon, F Klauschen, KR Müller, W Samek. On Pixel-wise Explanations for Non-Linear Classifier Decisions by Layer-wise Relevance Propagation. *PLOS ONE*, 10(7):e0130140, 2015.

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

L Arras, G Montavon, K-R Müller, W Samek. Explaining Recurrent Neural Network Predictions in Sentiment Analysis. *EMNLP'17 Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA)*, 159-168, 2017.

A Binder, G Montavon, S Lapuschkin, KR Müller, W Samek. Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers. *Artificial Neural Networks and Machine Learning – ICANN 2016*, Part II, Lecture Notes in Computer Science, Springer-Verlag, 9887:63-71, 2016.

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

References

Application to Text

L Arras, F Horn, G Montavon, KR Müller, W Samek. Explaining Predictions of Non-Linear Classifiers in NLP. *Workshop on Representation Learning for NLP,* Association for Computational Linguistics, 1-7, 2016.

L Arras, F Horn, G Montavon, KR Müller, W Samek. "What is Relevant in a Text Document?": An Interpretable Machine Learning Approach. *PLOS ONE, 12(8):e0181142*, 2017.

L Arras, G Montavon, K-R Müller, W Samek. Explaining Recurrent Neural Network Predictions in Sentiment Analysis. *EMNLP'17 Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA)*, 159-168, 2017.

Application to Images & Faces

S Lapuschkin, A Binder, G Montavon, KR Müller, Wojciech Samek. Analyzing Classifiers: Fisher Vectors and Deep Neural Networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2912-20, 2016.

S Bach, A Binder, KR Müller, W Samek. Controlling Explanatory Heatmap Resolution and Semantics via Decomposition Depth. *IEEE International Conference on Image Processing (ICIP)*, 2271-75, 2016.

F Arbabzadeh, G Montavon, KR Müller, W Samek. Identifying Individual Facial Expressions by Deconstructing a Neural Network. *Pattern Recognition - 38th German Conference, GCPR 2016*, Lecture Notes in Computer Science, 9796:344-54, Springer International Publishing, 2016.

S Lapuschkin, A Binder, KR Müller, W Samek. Understanding and Comparing Deep Neural Networks for Age and Gender Classification. *IIEEE International Conference on Computer Vision Workshops (ICCVW)*, 1629-38, 2017.

C Seibold, W Samek, A Hilsmann, P Eisert. Accurate and Robust Neural Networks for Security Related Applications Exampled by Face Morphing Attacks. arXiv:1806.04265, 2018.

References

Application to Video

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

V Srinivasan, S Lapuschkin, C Hellge, KR Müller, W Samek. Interpretable Human Action Recognition in Compressed Domain. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1692-96, 2017.

Application to Speech

S Becker, M Ackermann, S Lapuschkin, KR Müller, W Samek. Interpreting and Explaining Deep Neural Networks for Classification of Audio Signals. *arXiv:1807.03418,* 2018.

Application to the Sciences

F Horst, S Lapuschkin, W Samek, KR Müller, WI Schöllhorn. Explaining the Unique Nature of Individual Gait Patterns with Deep Learning, *Scientific Reports*, 9:2391, 2019.

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

A Thomas, H Heekeren, KR Müller, W Samek. Interpretable LSTMs For Whole-Brain Neuroimaging Analyses. *arXiv:1810.09945*, 2018.

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

References

Evaluation Explanations

W Samek, A Binder, G Montavon, S Lapuschkin, KR Müller. Evaluating the visualization of what a Deep Neural Network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11):2660-2673, 2017.

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

Software

M Alber, S Lapuschkin, P Seegerer, M Hägele, KT Schütt, G Montavon, W Samek, KR Müller, S Dähne, PJ Kindermans. iNNvestigate neural networks!. *Journal of Machine Learning Research,* 20:1-8, 2019.

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