Tutorial on Interpreting and Explaining Deep Models in Computer Vision

Wojciech Samek
(Fraunhofer HHI)

Grégoire Montavon
(TU Berlin)

Klaus-Robert Müller
(TU Berlin)

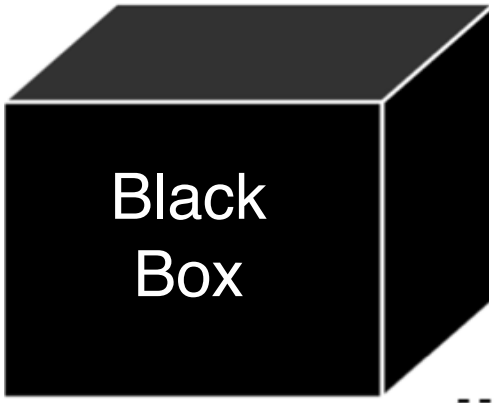
08:30 - 09:15	Introduction KRM
09:15 - 10:00	Techniques for Interpretability GM
10:00 - 10:30	Coffee Break ALL
10:30 - 11:15	Applications of Interpretability WS
11:15 - 12:00	Further Applications and Wrap-Up KRM

CVPR 2018
SALT LAKE CITY • JUNE 18-22

TU berlin
LIBERTAS JUSTITIA VERITAS
mpi max planck institut informatik
HHI
BBDC BERLIN BIG DATA CENTER

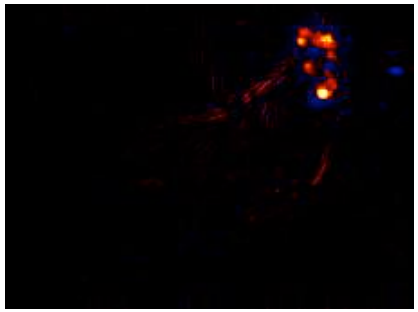
Opening the Black Box with LRP

input $\boldsymbol{x}$



Rooster

prediction $f(\boldsymbol{x})$



heatmap

Explain prediction

(how much each pixel contributes to prediction)

Idea: Decompose function

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

Simple LRP rule (Bach et al. 2015)

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

Every neuron gets its "share" of the redistributed relevance

Opening the Black Box with LRP

input $\boldsymbol{x}$

Black Box

Rooster

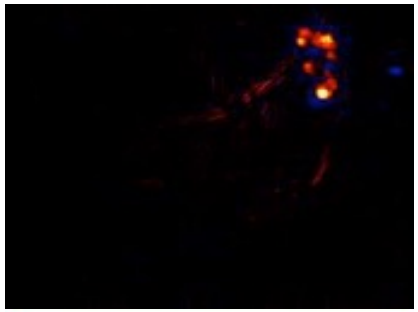
prediction $f(\boldsymbol{x})$

Idea: Decompose function

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

Explain prediction

(how much each pixel contributes to prediction)



heatmap

Theoretical Interpretation
(Deep) Taylor decomposition

Excitation Backprop (Zhang et al., 2016) is special case of LRP (α=1).

.

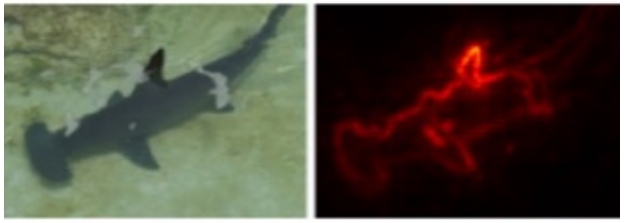
alpha-beta LRP rule (Bach et al. 2015)

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

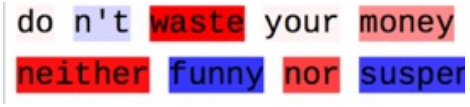
where $\alpha + \beta = 1$

LRP applied to different Data

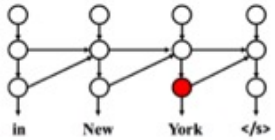
General Images (Bach' 15, Lapuschkin'16)



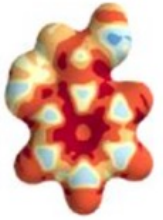
Text Analysis (Arras'16 &17)



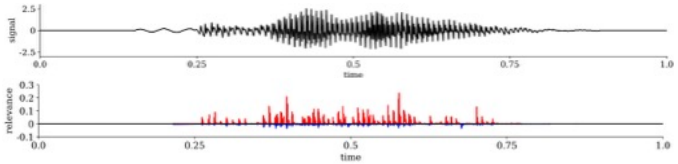
Translation (Ding'17)



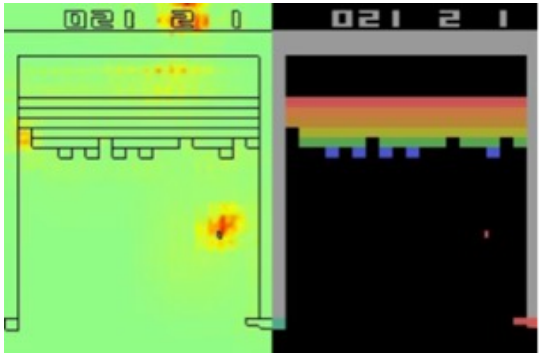
Molecules (Schütt'17)



Speech (Becker'18)



Games (Lapuschkin'18, in prep.)

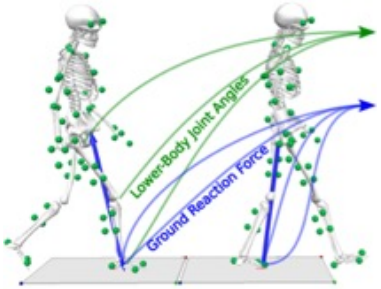


VQA (Arras'18)

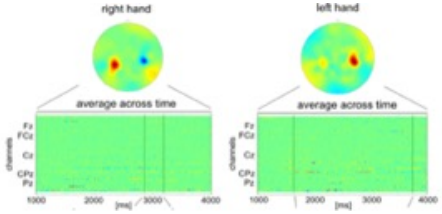
Video (Anders'18)

Morphing (Seibold'18)

Gait Patterns (Horst'18, in prep.)

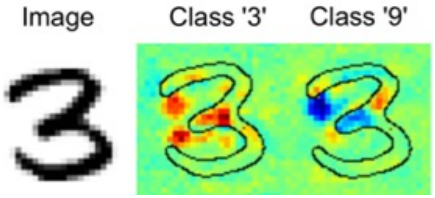


EEG (Sturm'16)

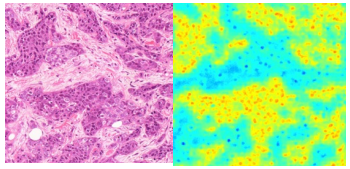


Faces (Arbabzadeh'16, Lapuschkin'17)

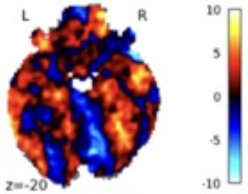
Digits (Bach' 15)



Histopathology (Binder'18)

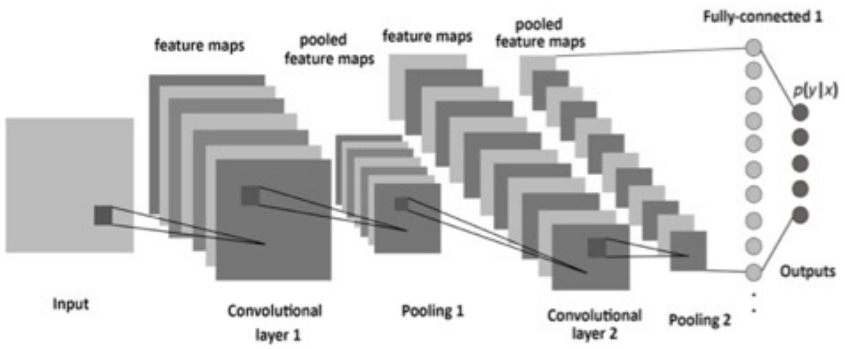


fMRI (Thomas'18)

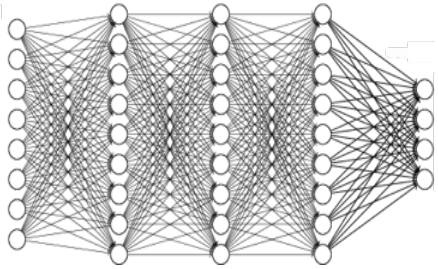


LRP applied to different Models

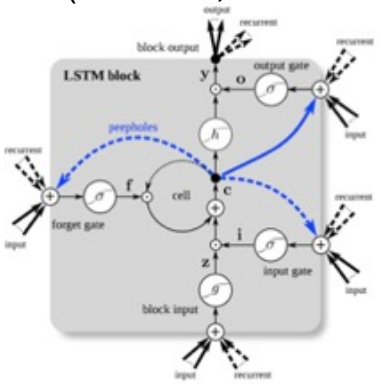
Convolutional NNs (Bach'15, Arras'17 …)



Local Renormalization Layers (Binder'16)

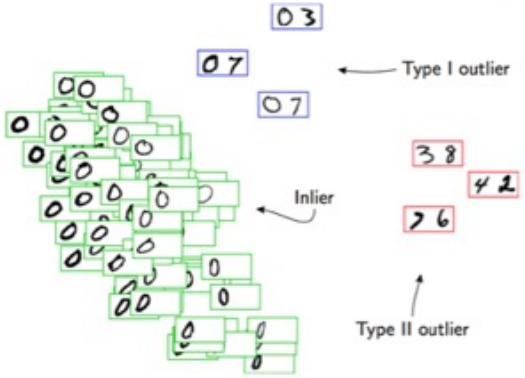


LSTM (Arras'17, Thomas'18)



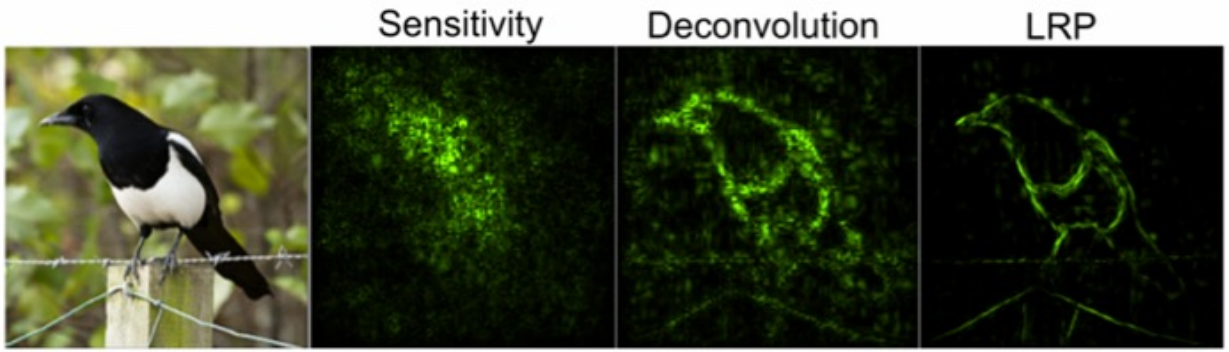
Bag-of-words / Fisher Vector models (Bach'15, Arras'16, Lapuschkin'17, Binder'18)

One-class SVM (Kauffmann'18)



Now What ?

Compare Explanation Methods



Can we objectively measure which heatmap is best ?

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

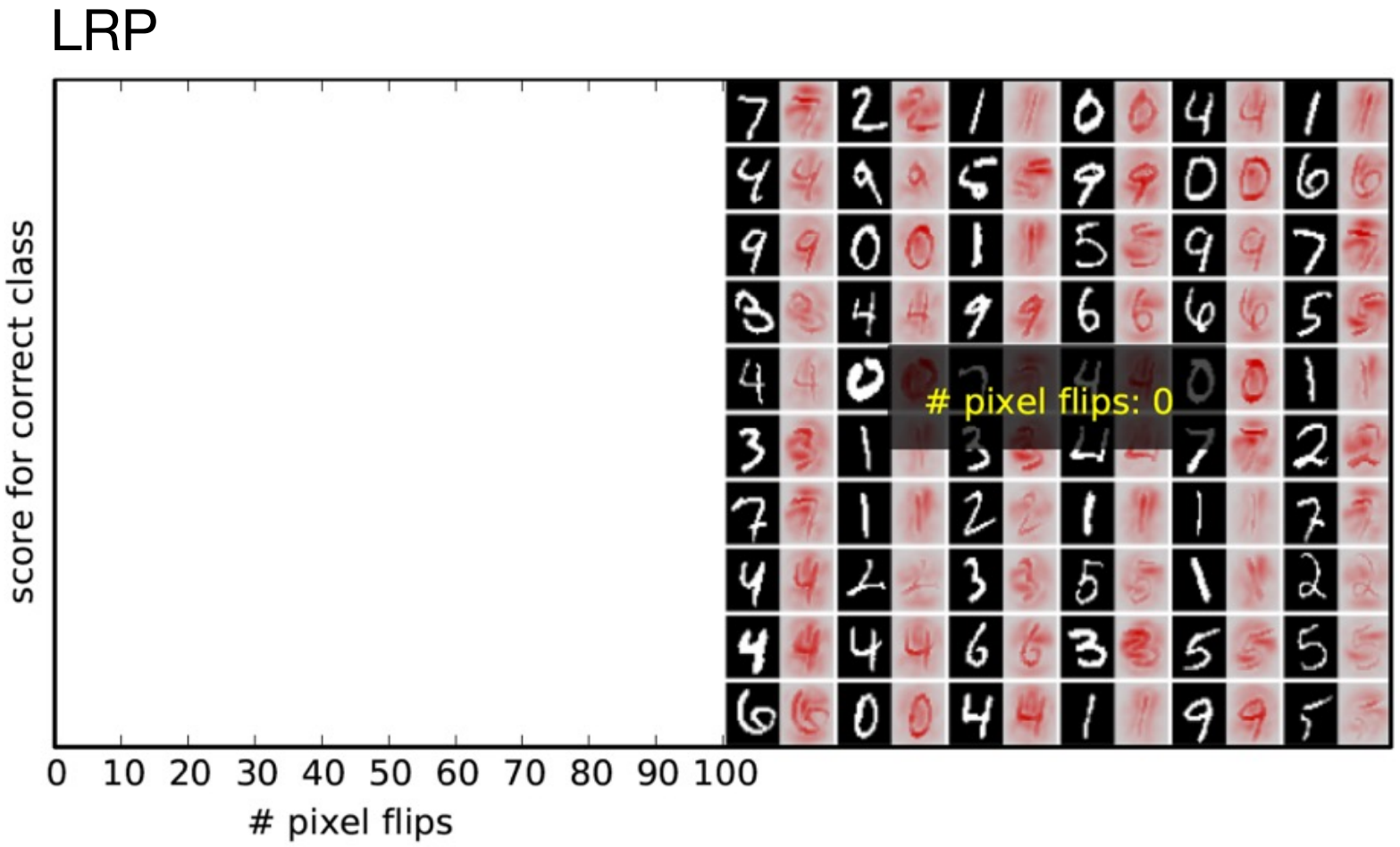
"If input features are deemed relevant, removing them should reduce evidence at the output of the network."

Algorithm ("Pixel Flipping")

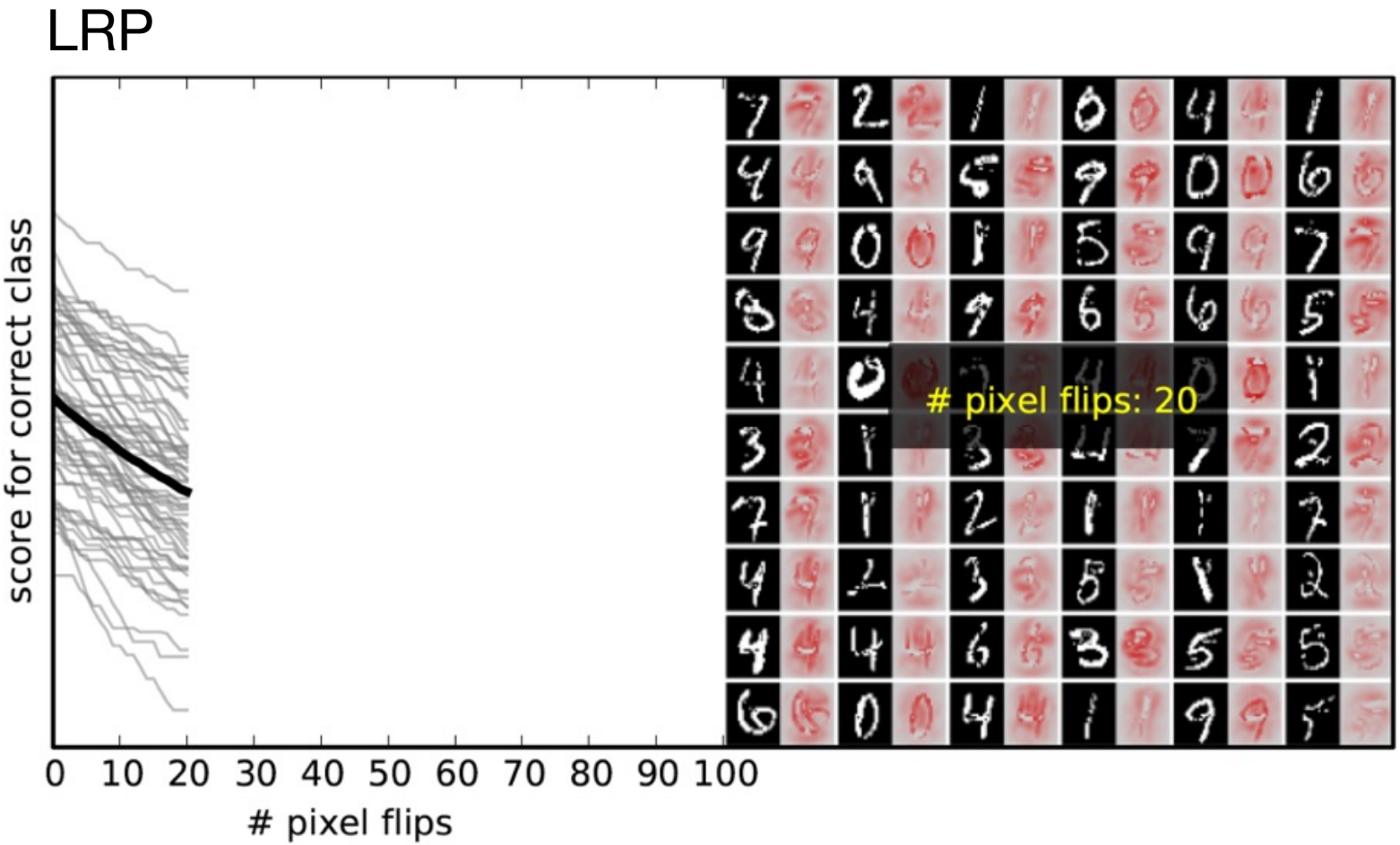
```
Sort pixels / patches by relevance
Iterate
  destroy pixel / patch
  evaluate f(x)
Measure decrease of f(x)
```

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

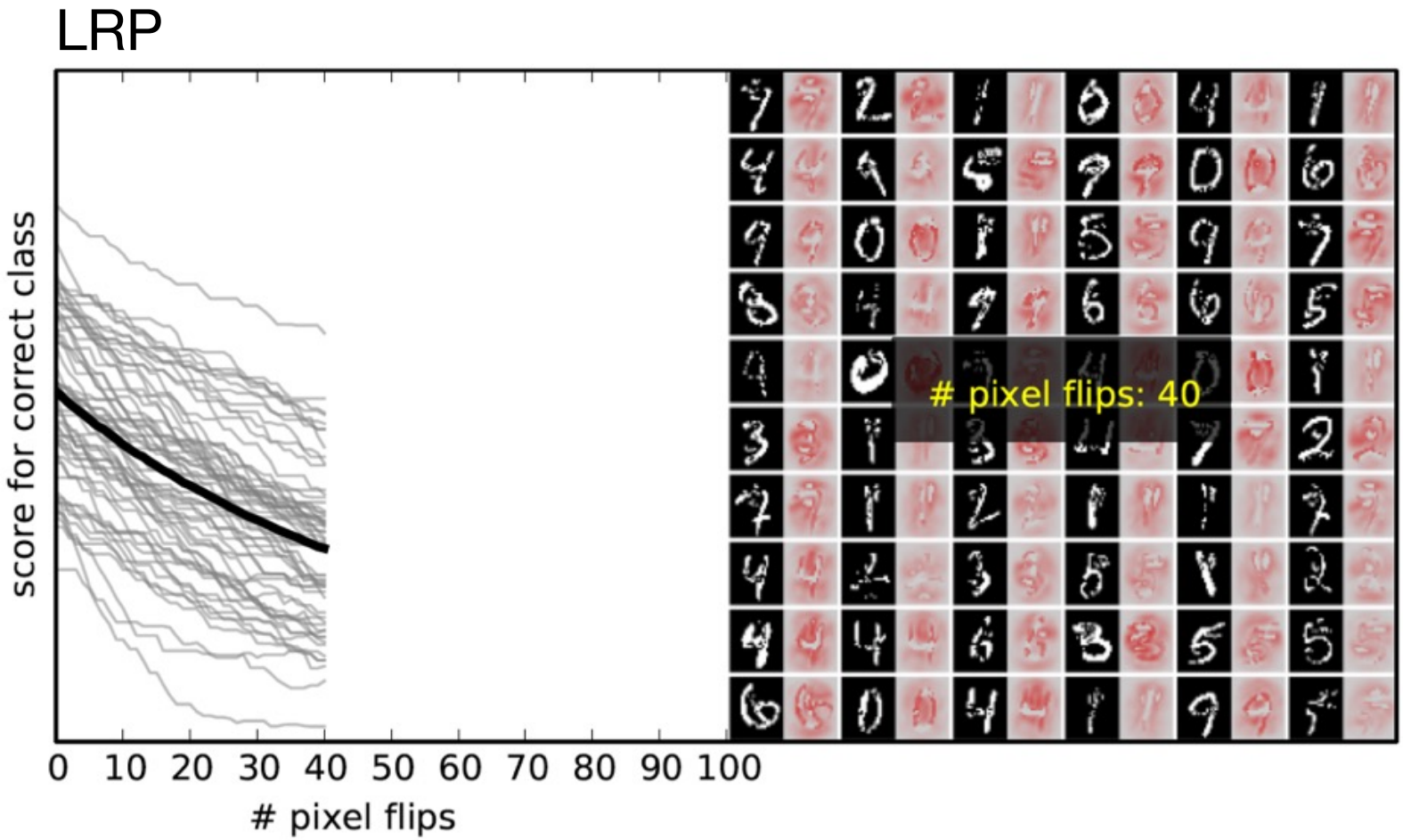
Compare Explanation Methods



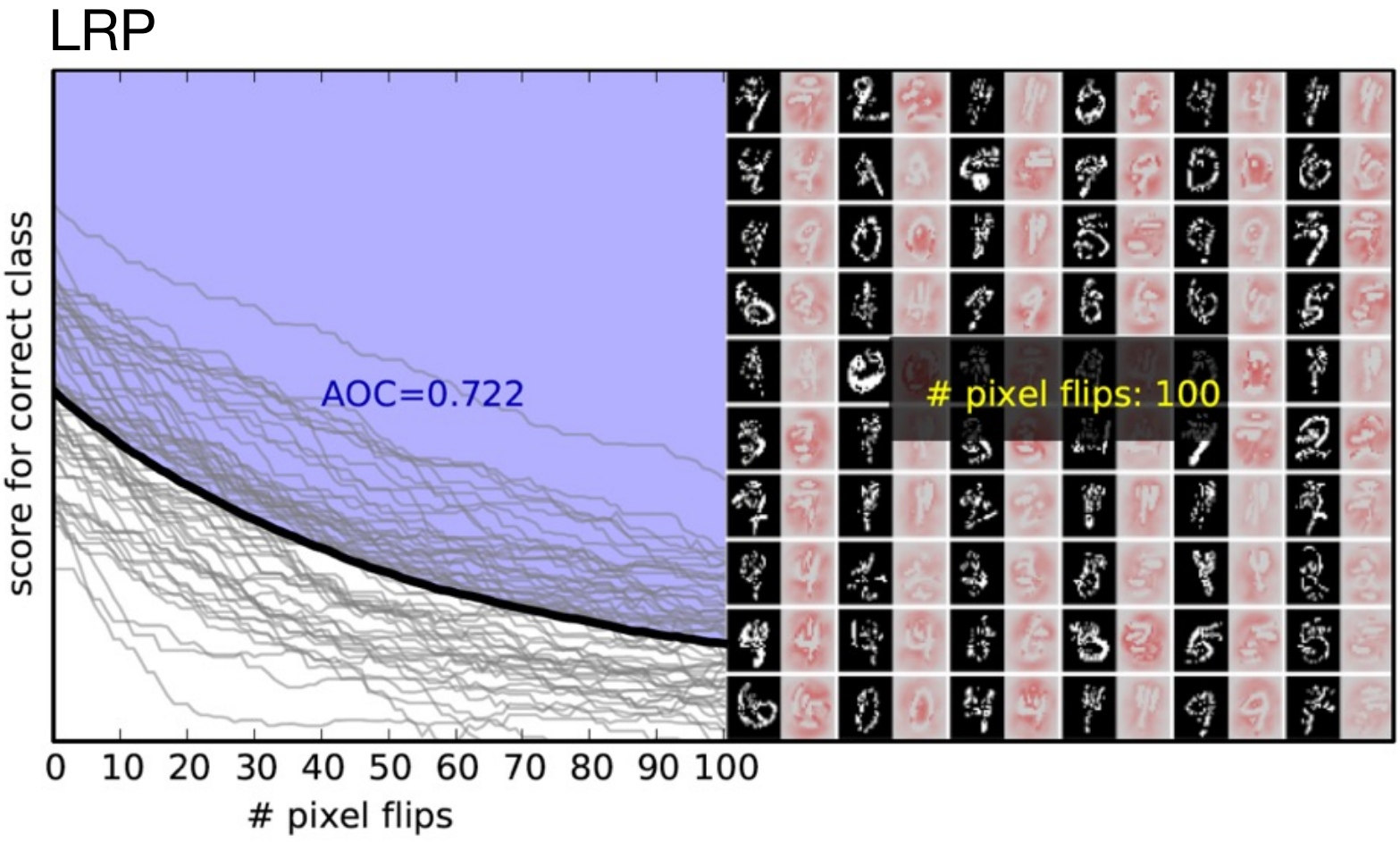
Compare Explanation Methods



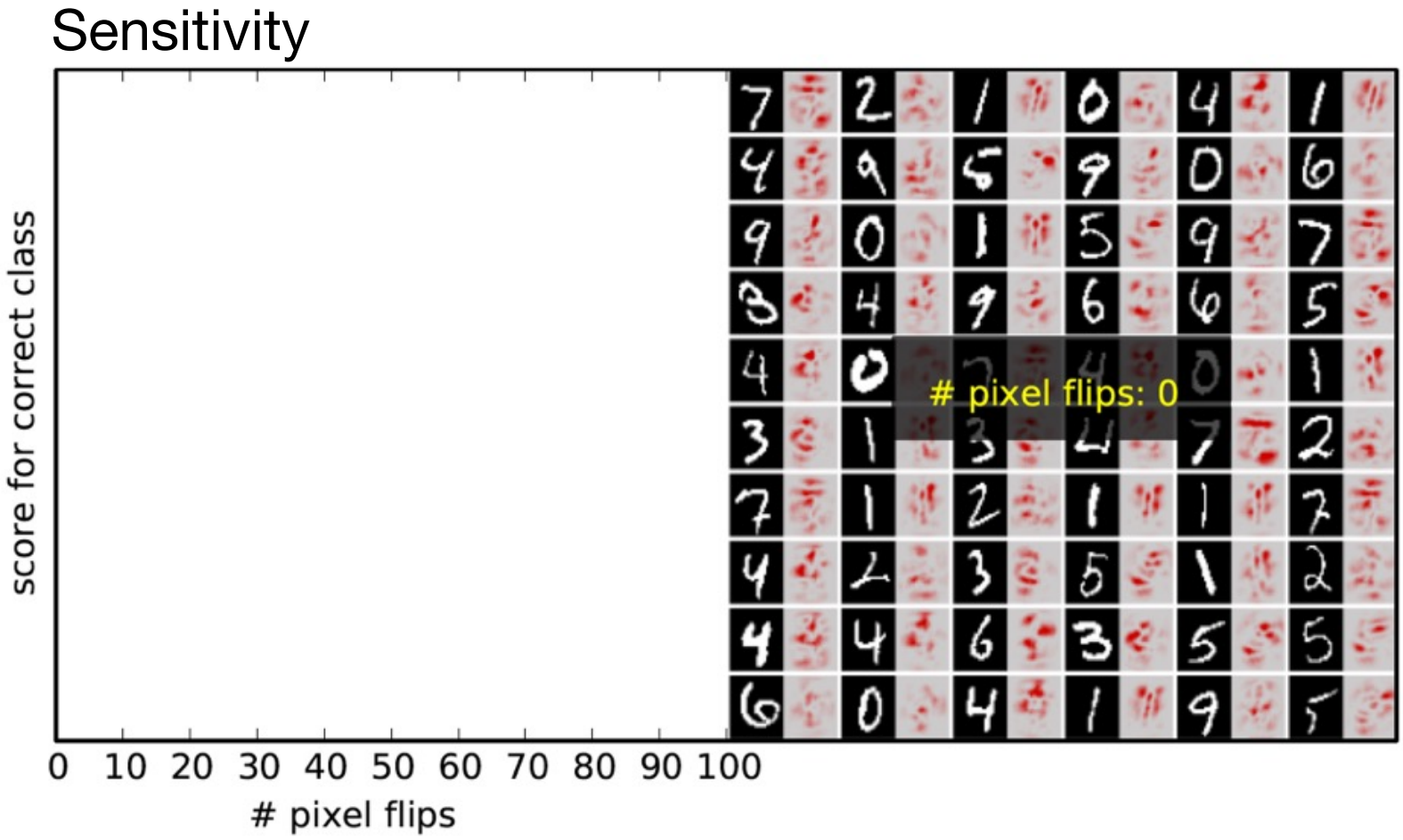
Compare Explanation Methods



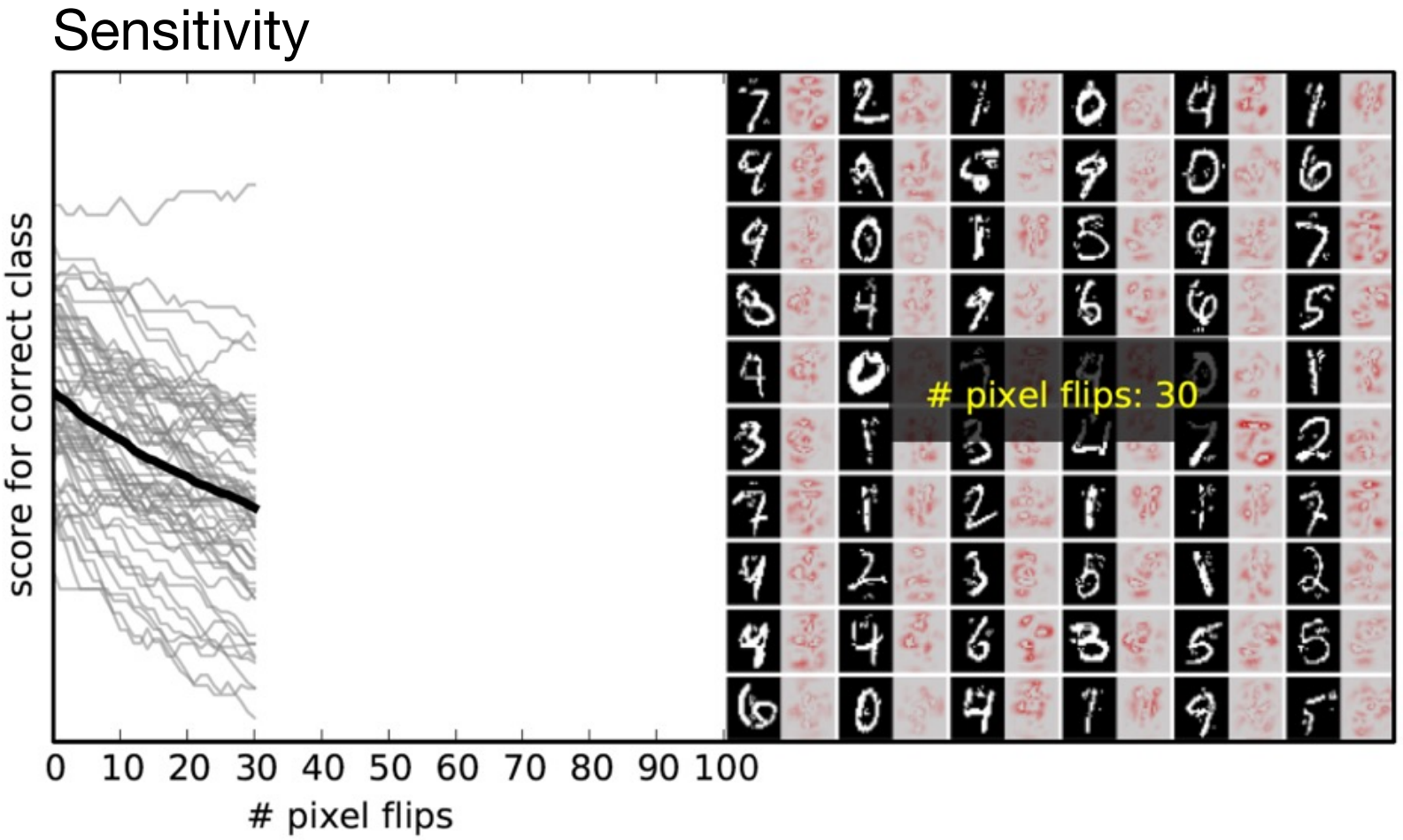
Compare Explanation Methods



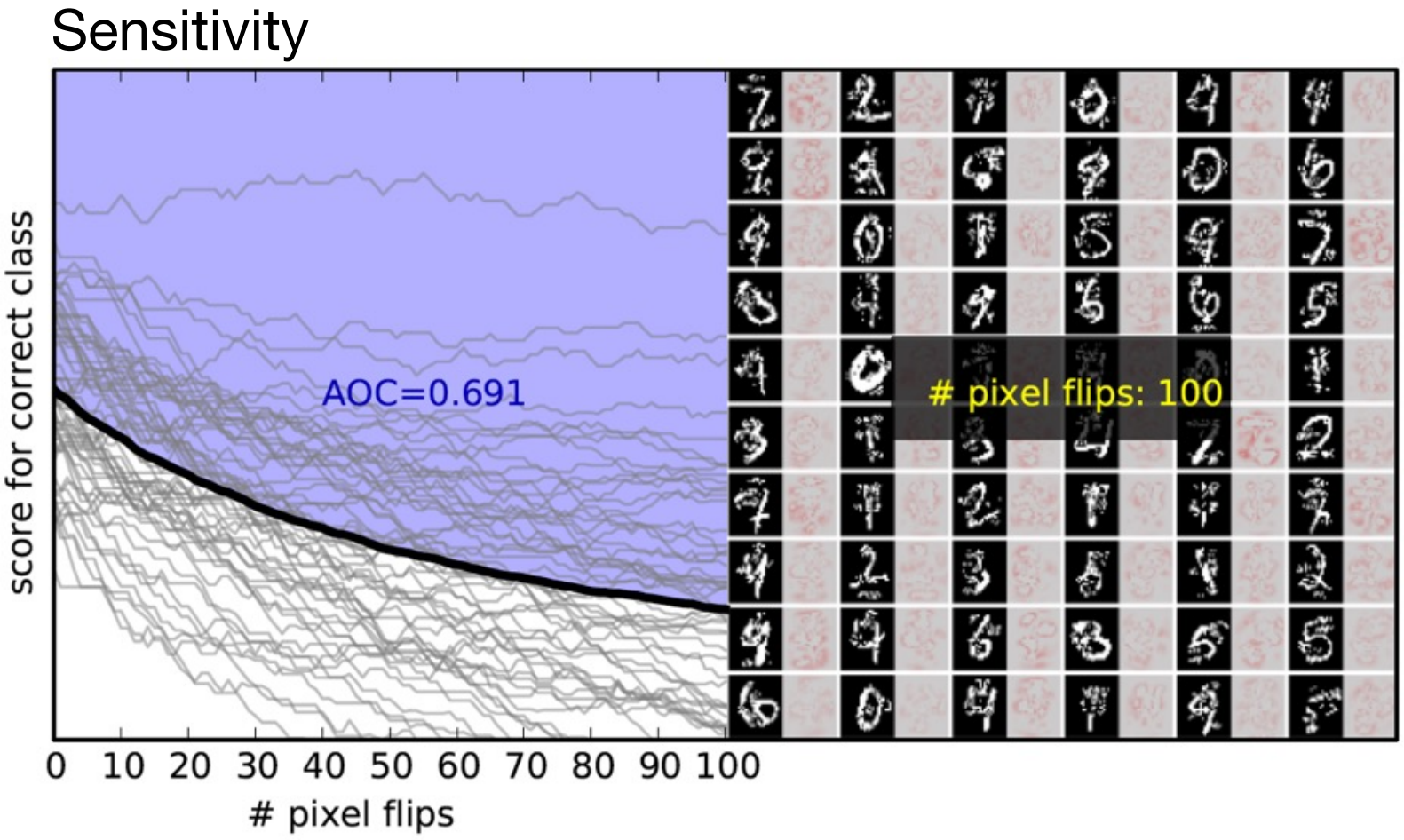
Compare Explanation Methods



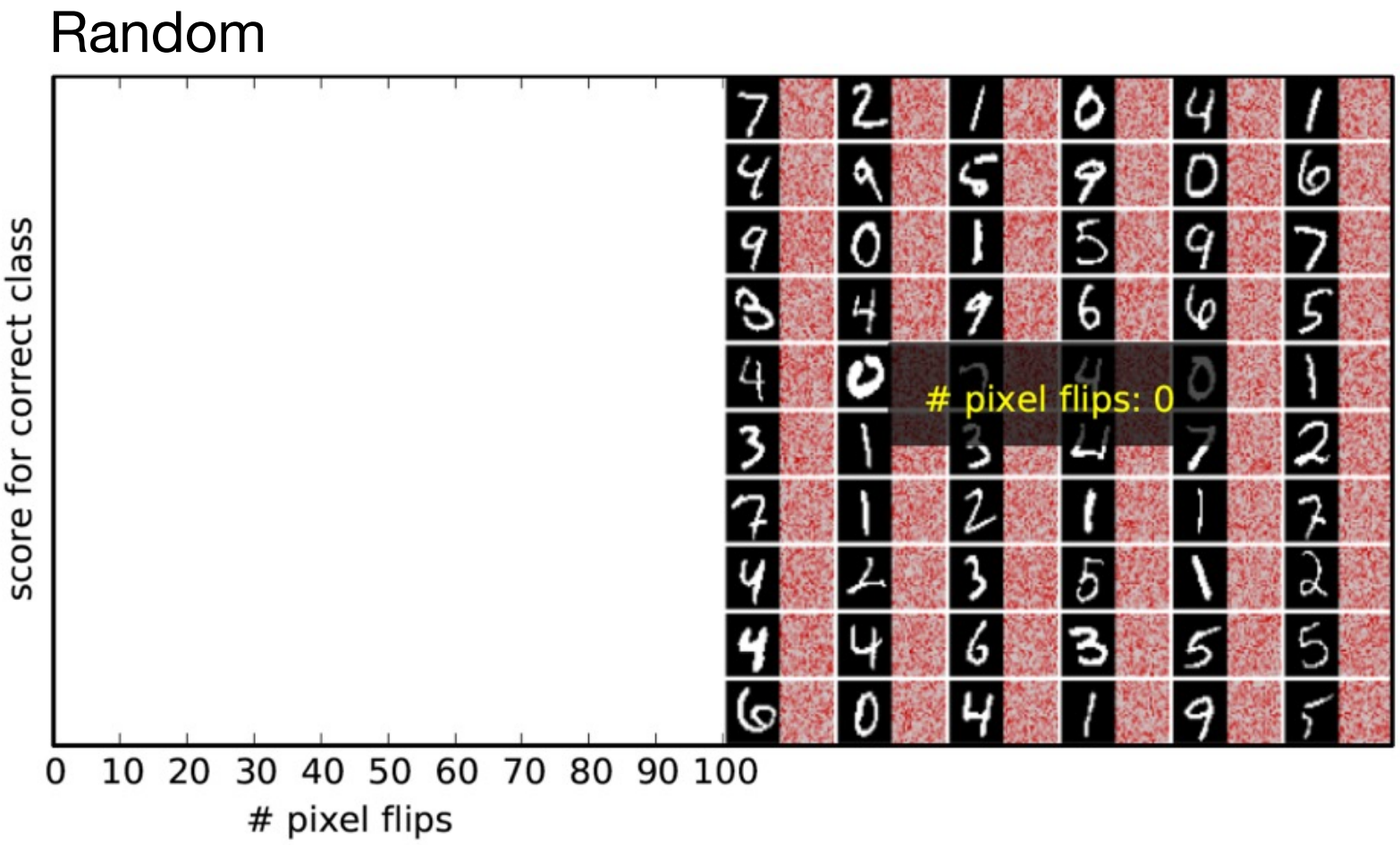
Compare Explanation Methods



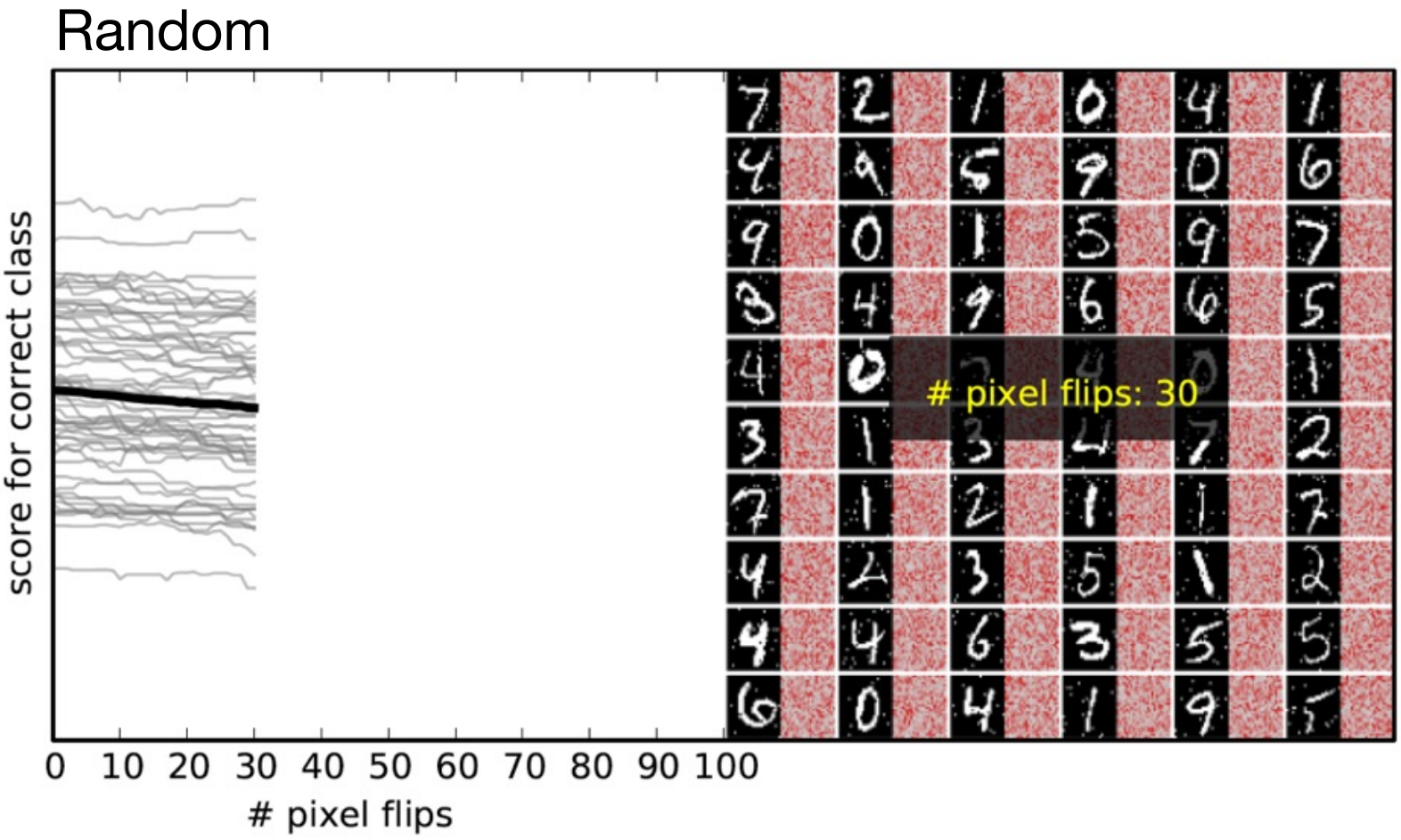
Compare Explanation Methods



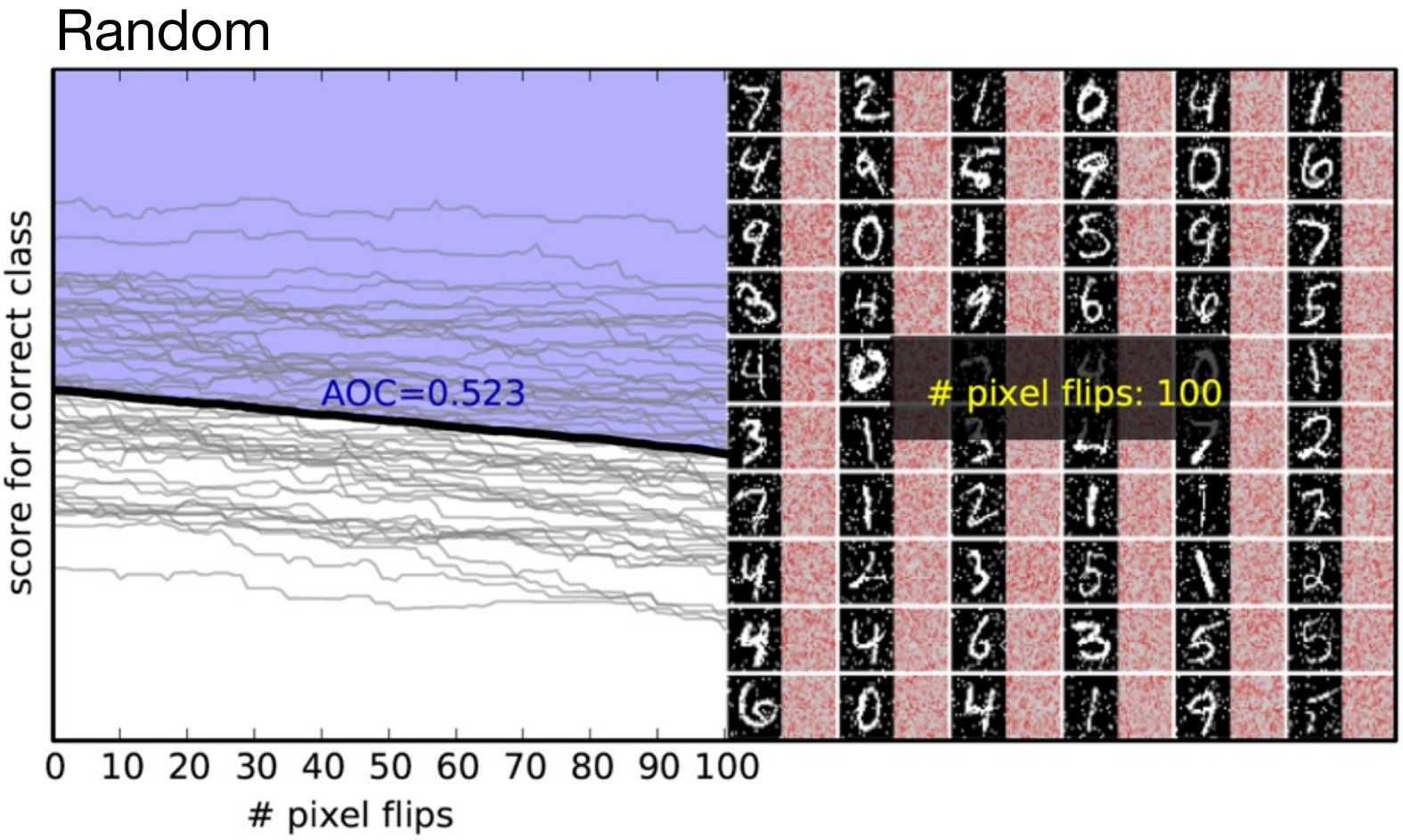
Compare Explanation Methods



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)

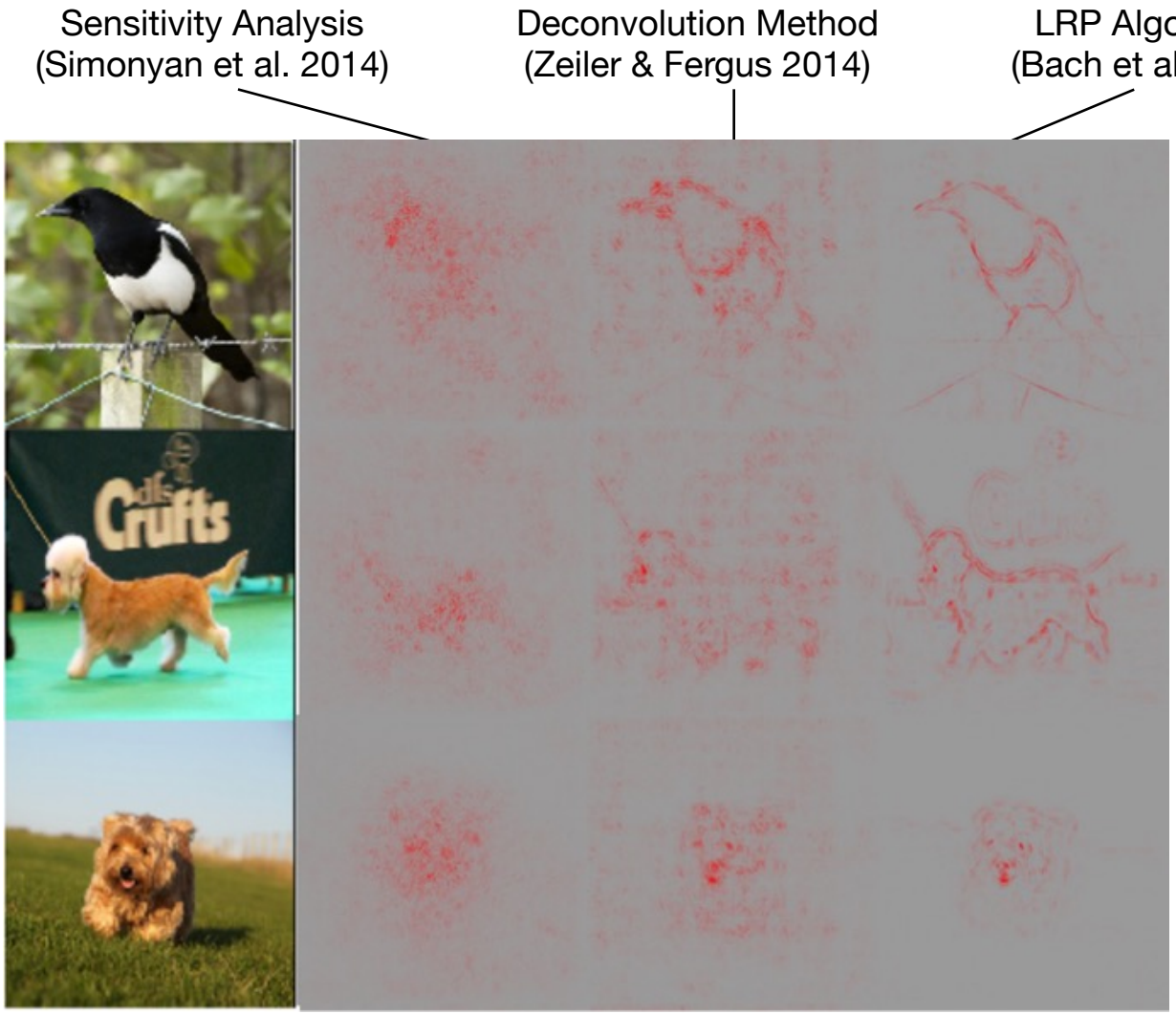
ILSVRC2012

1000 categories
(1.2 million training images)

MIT Places

205 scene categories
(2.5 millions of images)

Compare Explanation Methods

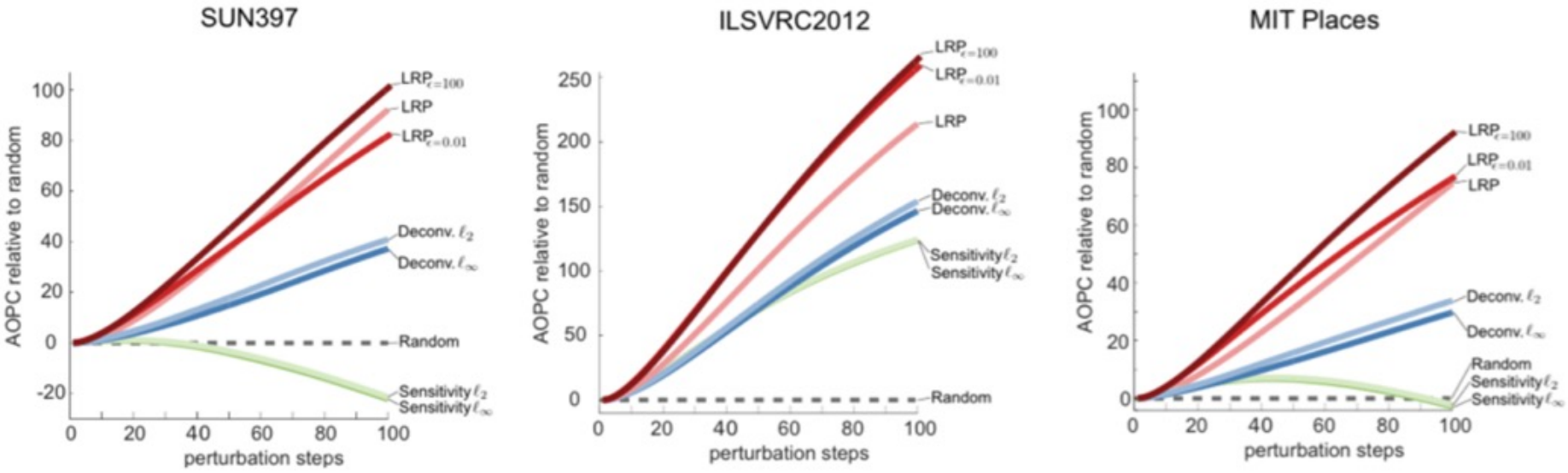


(Samek et al. 2017)

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

- Sensitivity heatmaps are noisy (gradient shuttering)
- Deconvolution and sensitivity analysis solve a different problem

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.

sci.med (4.1)

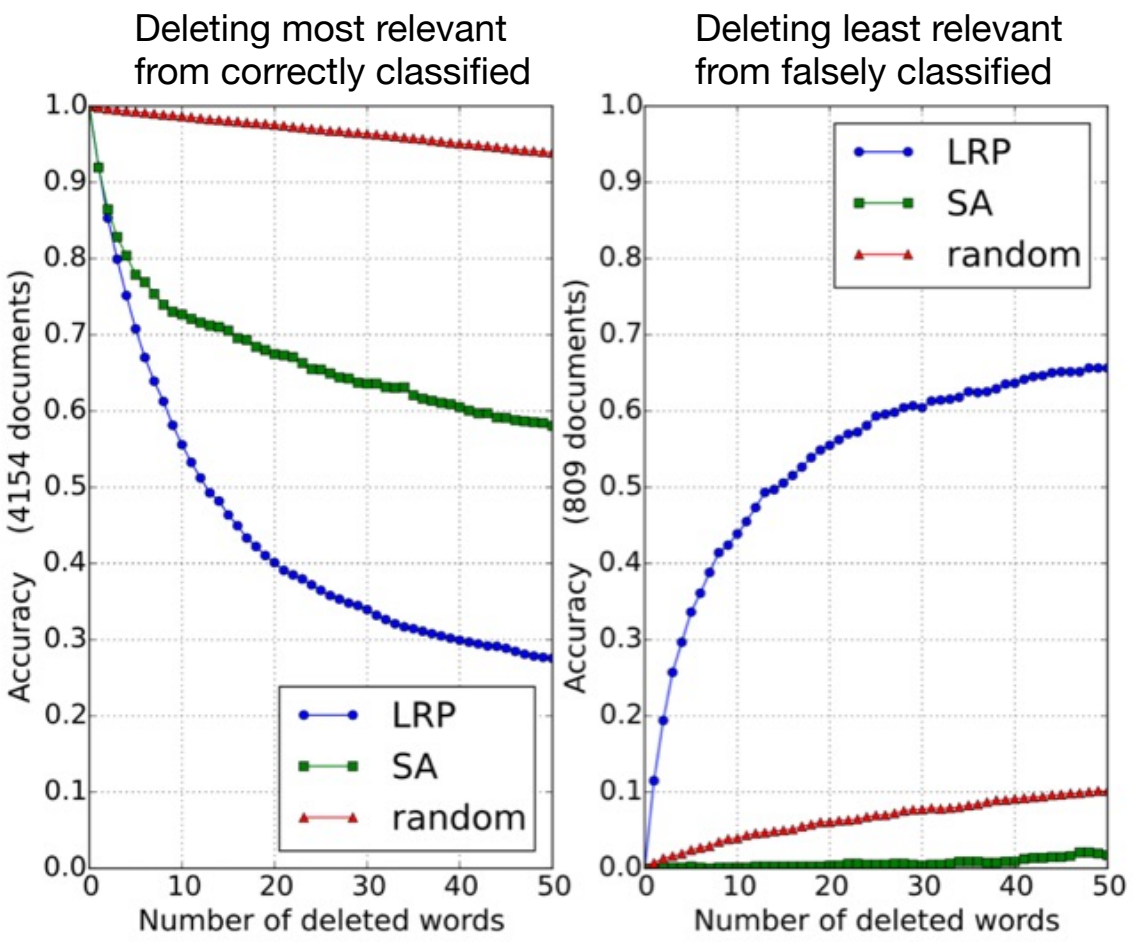
Yes, weightlessness does feel like falling. It may feel strange at first, but the body does adjust. The feeling is not too different from that of sky diving.

>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.

(Arras et al. 2017)

Compare Explanation Methods



- word2vec / CNN model
- Conv → ReLU → 1-Max-Pool → FC
- trained on 20Newsgroup Dataset
- accuracy: 80.19%

LRP better than SA

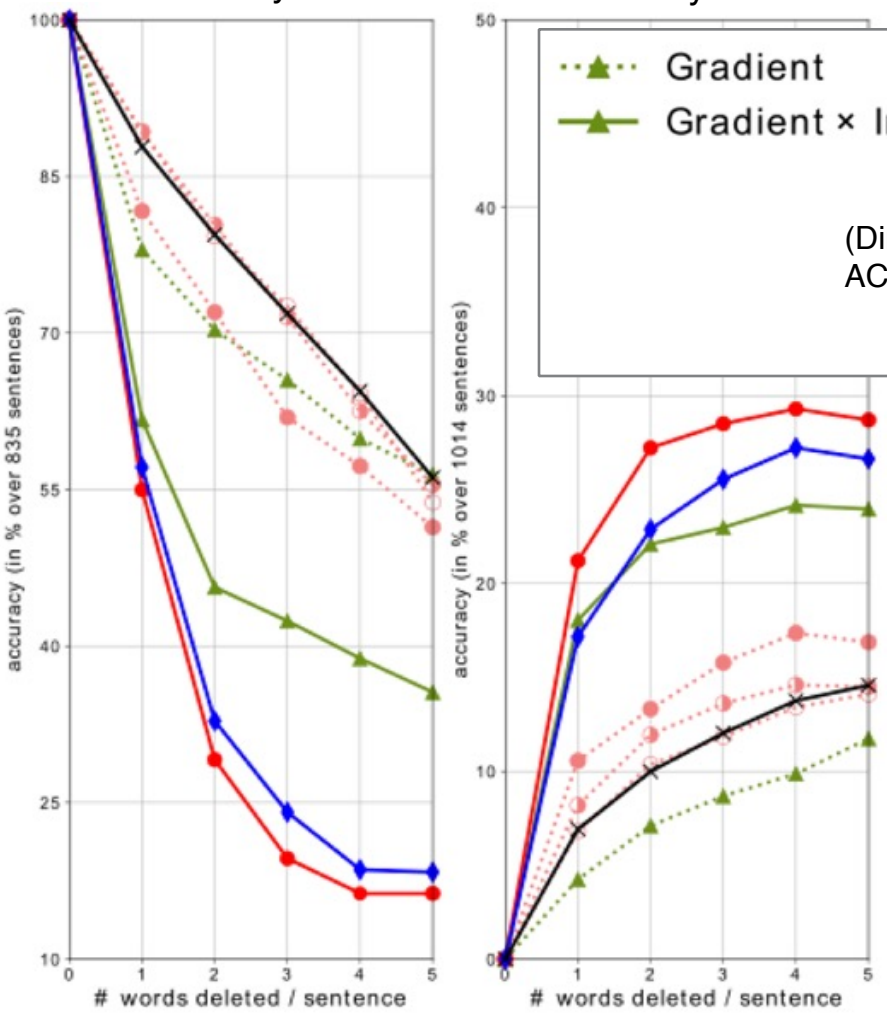
LRP distinguishes between positive and negative evidence

(Arras et al. 2016)

Compare Explanation Methods

Deleting most relevant from correctly classified

Deleting least relevant from falsely classified



Gradient

Gradient × Input

(Ding et al. ACL, 2017)

$LRP_{\odot sum}$

$LRP_{\odot abs}$

$LRP_{\odot half}$

(Murdoch et al. ICLR, 2018)

$LRP_{\odot full}$

CD

random

(Arras et al. EMNLP-WASSA, 2017)

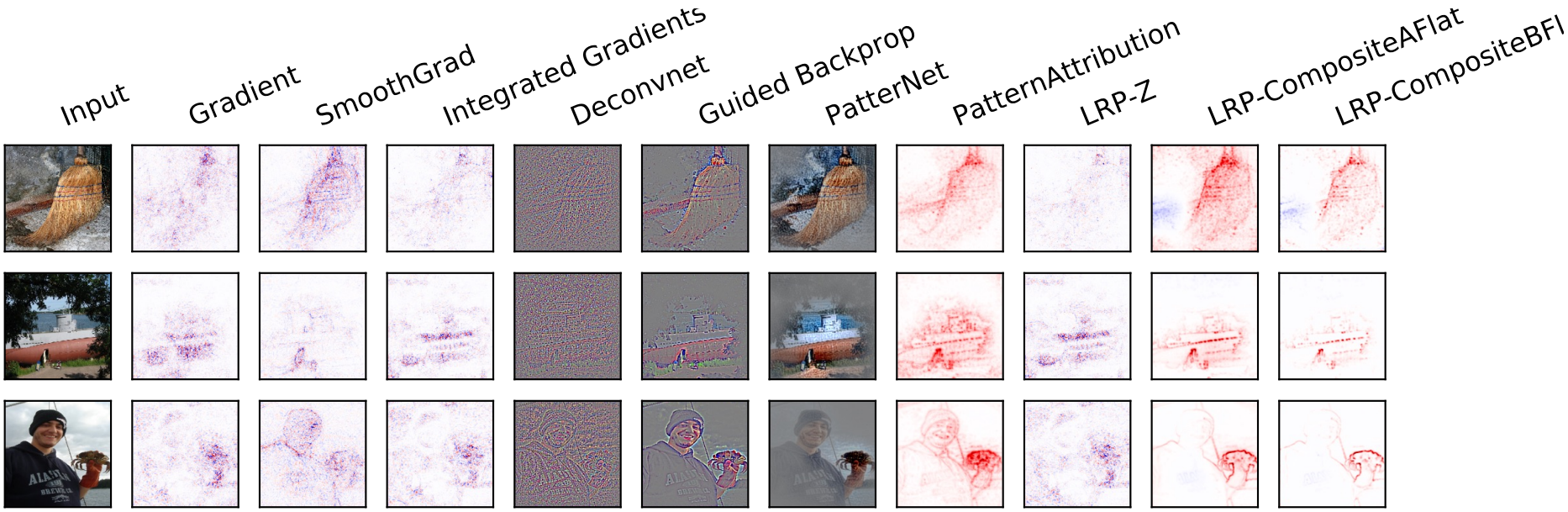
- bidirectional LSTM model (Li'16)
- Stanford Sentiment Treebank dataset
- delete up to 5 words per sentence

LRP outperforms baselines (also recently proposed contextual decomposition)

LRP ≠ Gradient x Input

(Arras et al. 2018)

Compare Explanation Methods



Highly efficient (e.g., 0.01 sec per VGG16 explanation) !

New Keras Toolbox available for explanation methods:
`https://github.com/albermax/innvestigate`

Application of LRP

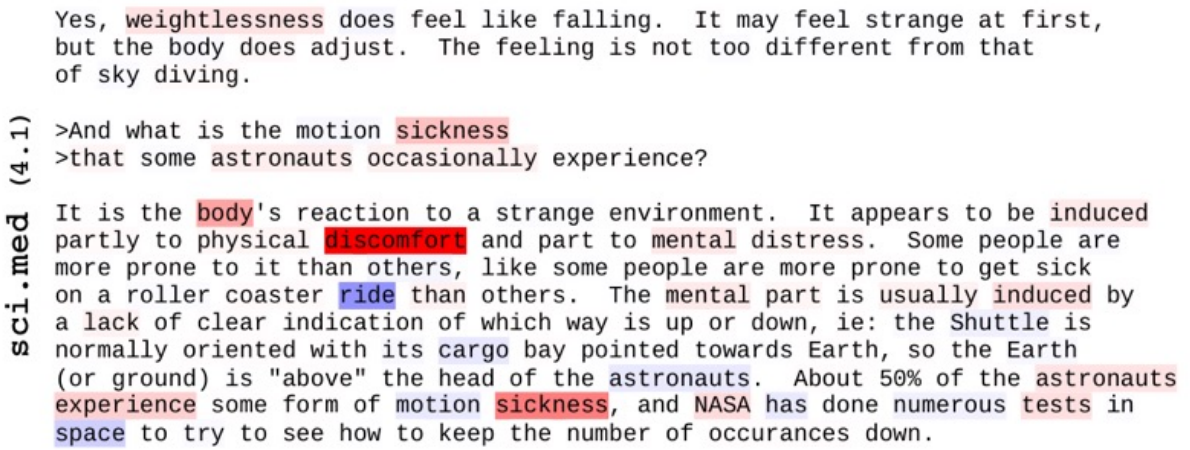
Compare models

Application: Compare Classifiers

word2vec/CNN:

Performance: 80.19%

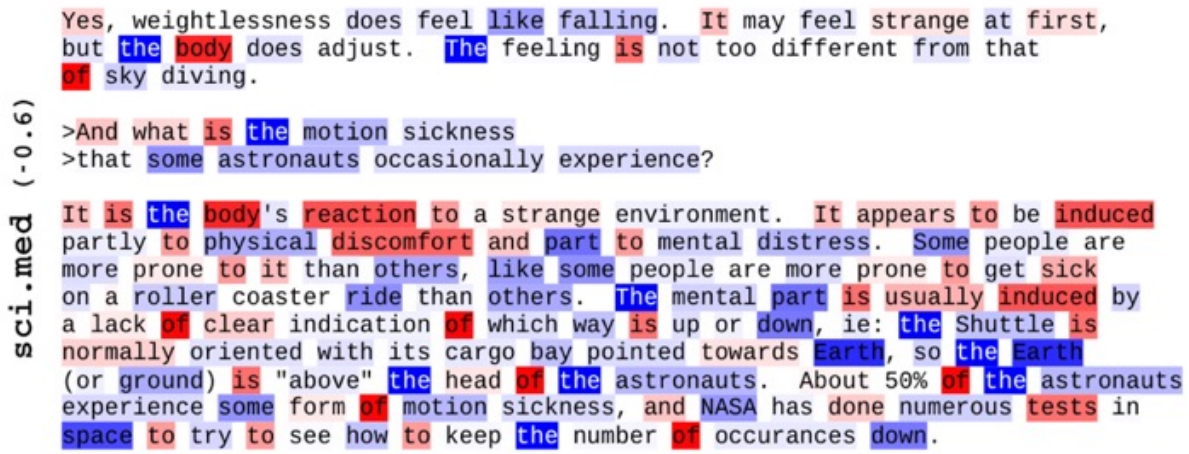
Strategy to solve the problem: identify semantically meaningful words related to the topic.



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

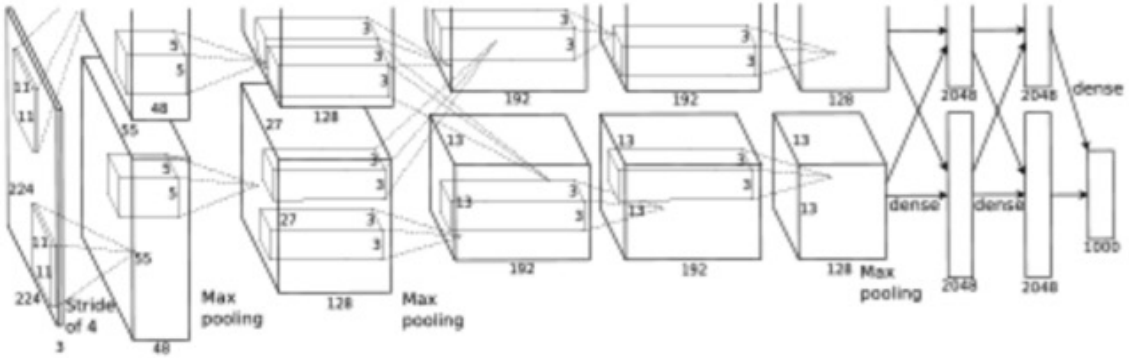
sci.med

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).

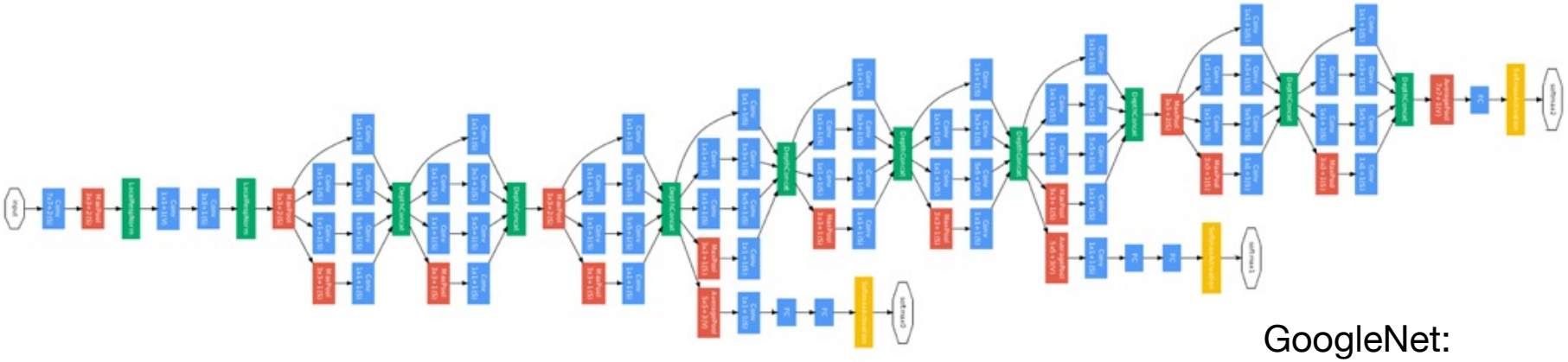
Words with maximum relevance

(Arras et al. 2016 & 2017)

Application: Compare Classifiers



BVLC:
- 8 Layers
- ILSRCV: 16.4%



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

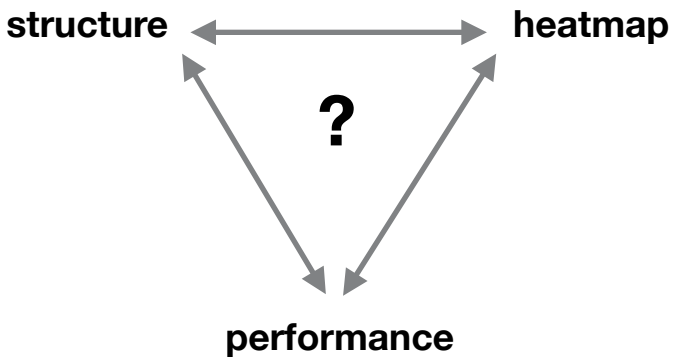
Application: Compare Classifiers

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

BVLC CaffeNet heatmaps are much more noisy.

Is it related to the architecture ?

Is it related to the performance ?



(Binder et al. 2016)

Application of LRP
Quantify Context Use

Application: Measure Context Use

how important is context ?

how important is context ?

classifier

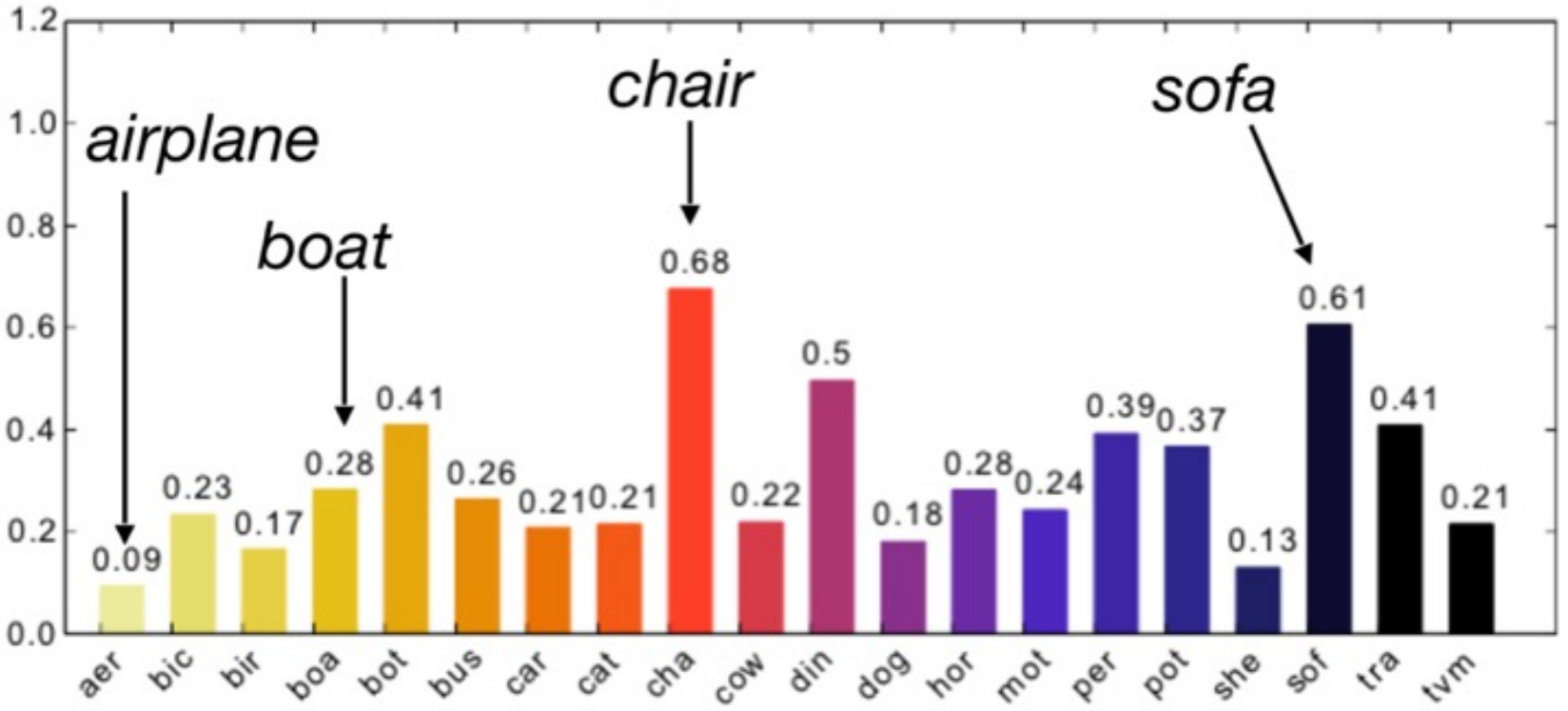
LRP decomposition allows meaningful pooling over bbox !

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

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

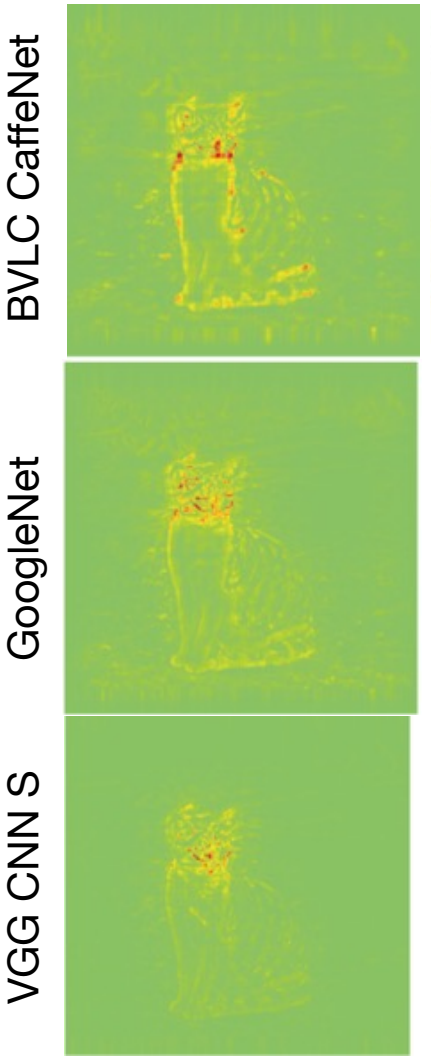
Application: Measure Context Use

- BVLC reference model + fine tuning
- PASCAL VOC 2007

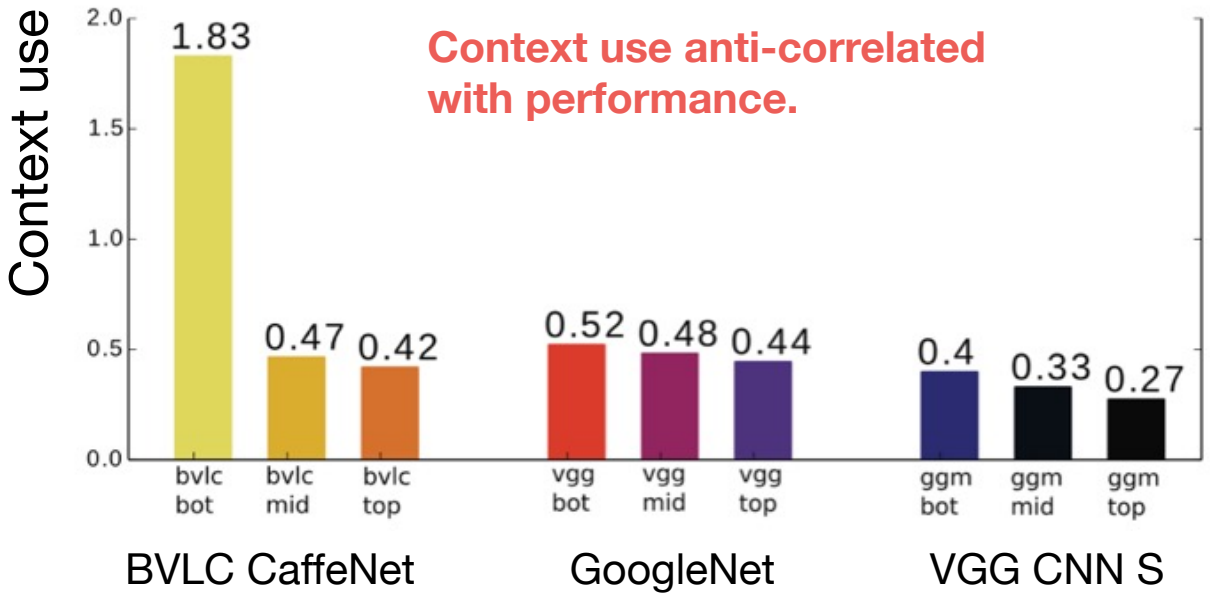


(Lapuschkin et al., 2016)

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

Compare Configuration,
Detect Biases & Improve Models

Application: Face analysis

- Compare AdienceNet, CaffeNet, GoogleNet, VGG-16
- state-of-the-art performance in age and gender classification
- Adience dataset, 26,580 images

Age classification

	A	C	G	V
[i]	51.4 87.0	52.1 87.9	54.3 89.1	–
[r]	51.9 87.4	52.3 88.9	53.3 89.9	–
[m]	53.6 88.4	54.3 89.7	56.2 90.7	–
[i,n]	–	51.6 87.4	56.2 90.9	53.6 88.2
[r,n]	–	52.1 87.0	57.4 91.9	–
[m,n]	–	52.8 88.3	58.5 92.6	56.5 90.0
[i,w]	–	–	–	59.7 94.2
[r,w]	–	–	–	–
[m,w]	–	–	–	62.8 95.8

Gender classification

	A	C	G	V
[i]	88.1	87.4	87.9	–
[r]	88.3	87.8	88.9	–
[m]	89.0	88.8	89.7	–
[i,n]	–	89.9	**91.0**	**92.0**
[r,n]	–	90.6	**91.6**	–
[m,n]	–	90.6	**91.7**	**92.6**
[i,w]	–	–	–	90.5
[r,w]	–	–	–	–
[m,w]	–	–	–	**92.2**

A = AdienceNet
C = CaffeNet
G = GoogleNet
V = VGG-16

[i] = in-place face alignment
[r] = rotation based alignment
[m] = mixing aligned images for training
[n] = initialization on Imagenet
[w] = initialization on IMDB-WIKI

(Lapuschkin et al., 2017)

Application: Face analysis

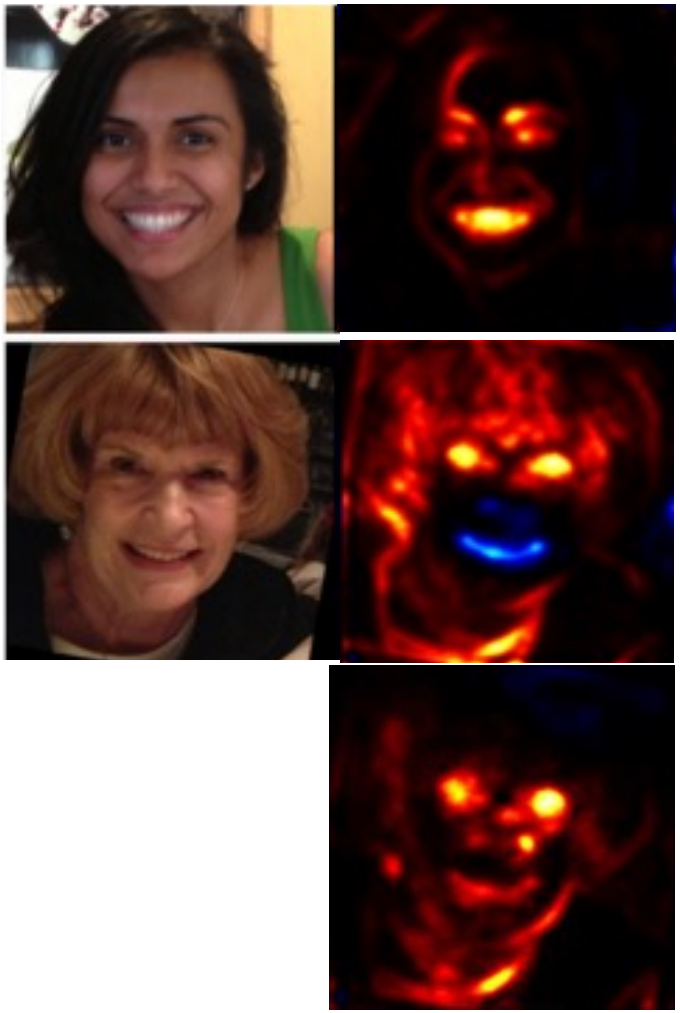
Gender classification

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

<u>Strategy to solve the problem</u>:
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: Face analysis

real person | fake person | real person

- 1,900 images of different individuals
- pretrained VGG19 model
- different ways to train the models

Different training methods

	naive	one morphed	complex morphs	multiclass
true positive	95%	90%	93%	92%
true negative	98%	95%	95%	99%
EER	3.1%	7.2%	6.1%	2.8%

50% genuine images,
50% complete morphs

50% genuine images,
10% complete morphs and
4 × 10% one region morphed

50% genuine images,
10% complete morphs,
partial morphs with 10%
one, two, three and four
region morphed

partial morphs with zero,
one, two, three or four
morphed regions,
for two class classification
last layer reinitialized

(Seibold et al., 2018)

Application: Face analysis

Semantic attack on the model

Table 4. Robustness against partial morphs.

	left eye	right eye	nose	mouth	average
naive	25%	21%	14%	13%	20%
one morphed	81%	89%	79%	71%	80%
complex morphs	78%	74%	73%	54%	70%
multiclass	86%	93%	90%	79%	87%

Black box adversarial attack on the model

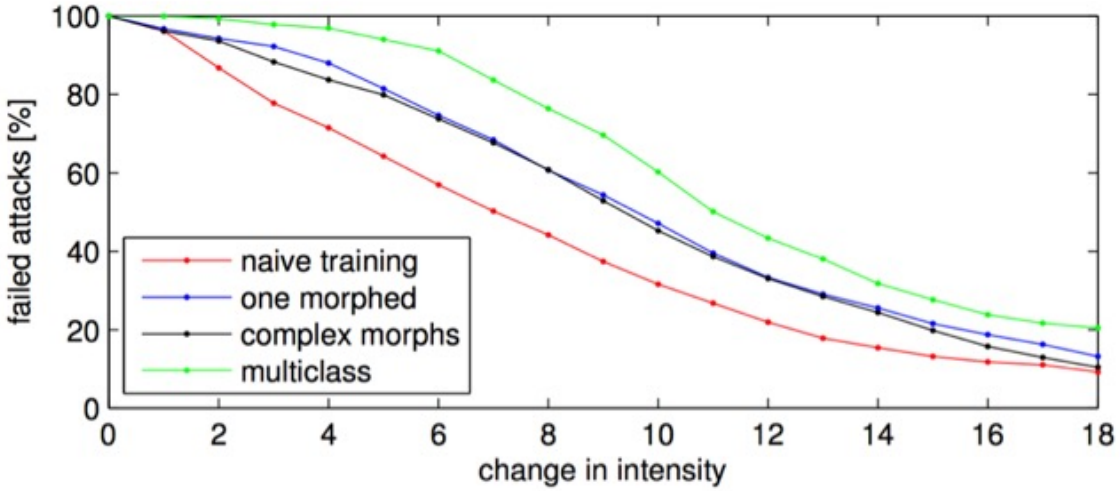


Fig. 5. Robustness against fast gradient sign attacks.

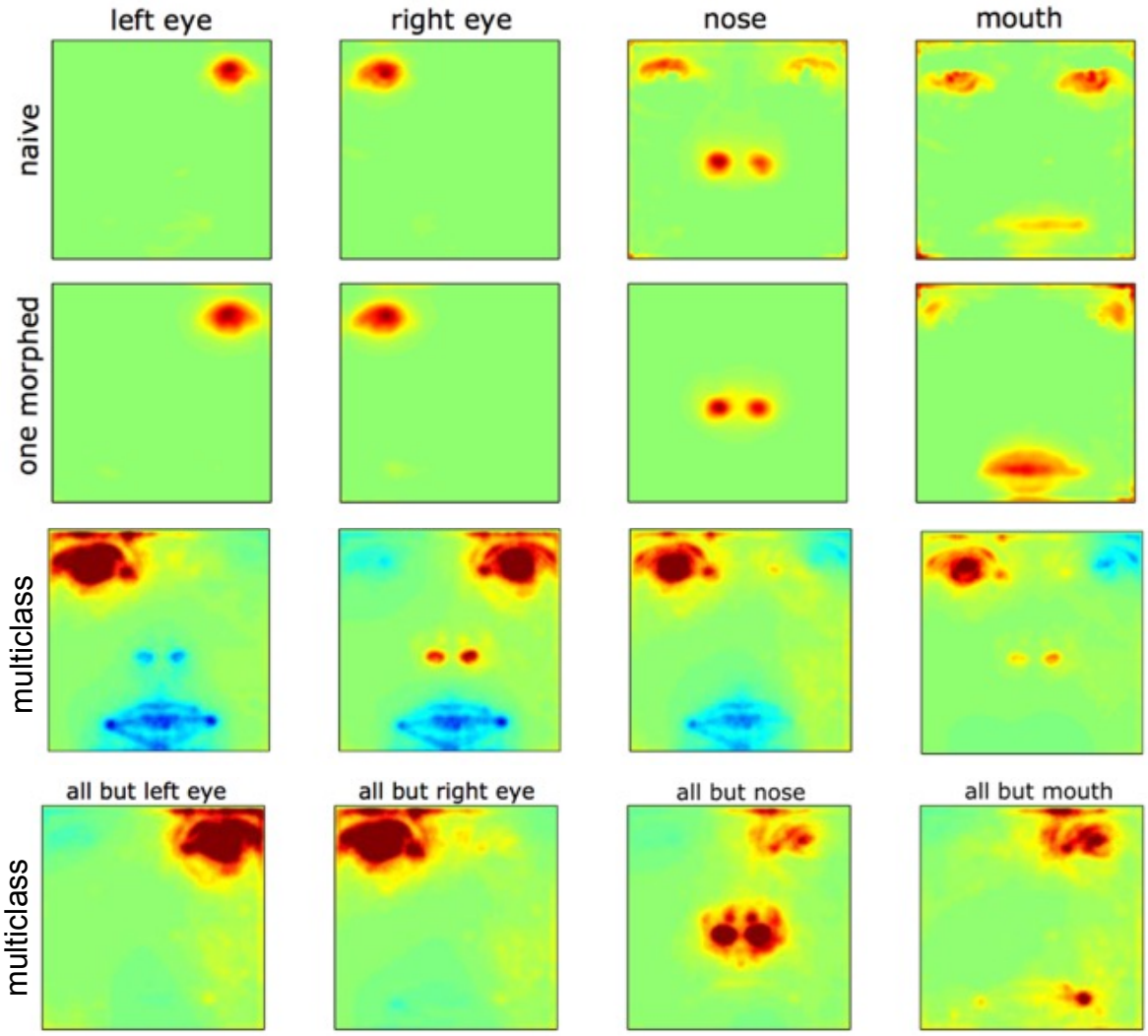
Application: Face analysis

morphed region	relative amount of relevance per region: naive				one morphed			
	left eye	right eye	nose	mouth	left eye	right eye	nose	mouth
left eye	0.84	0.00	0.02	0.14	0.96	0.00	0.01	0.04
right eye	0.00	0.91	0.05	0.05	0.00	0.92	0.01	0.07
nose	0.21	0.28	0.47	0.04	0.00	0.01	0.97	0.02
mouth	0.34	0.27	0.04	0.35	0.17	0.12	0.04	0.68

	complex morphs				multiclass			
	left eye	right eye	nose	mouth	left eye	right eye	nose	mouth
left eye	0.98	0.00	0.00	0.02	0.00	0.98	0.00	0.01
right eye	0.00	0.92	0.00	0.08	0.98	0.00	0.02	0.00
nose	0.02	0.03	0.92	0.02	0.01	0.10	0.19	0.70
mouth	0.06	0.00	0.41	0.53	0.11	0.18	0.58	0.13

(Seibold et al., 2018)

Application: Face analysis



Different models have different strategies !

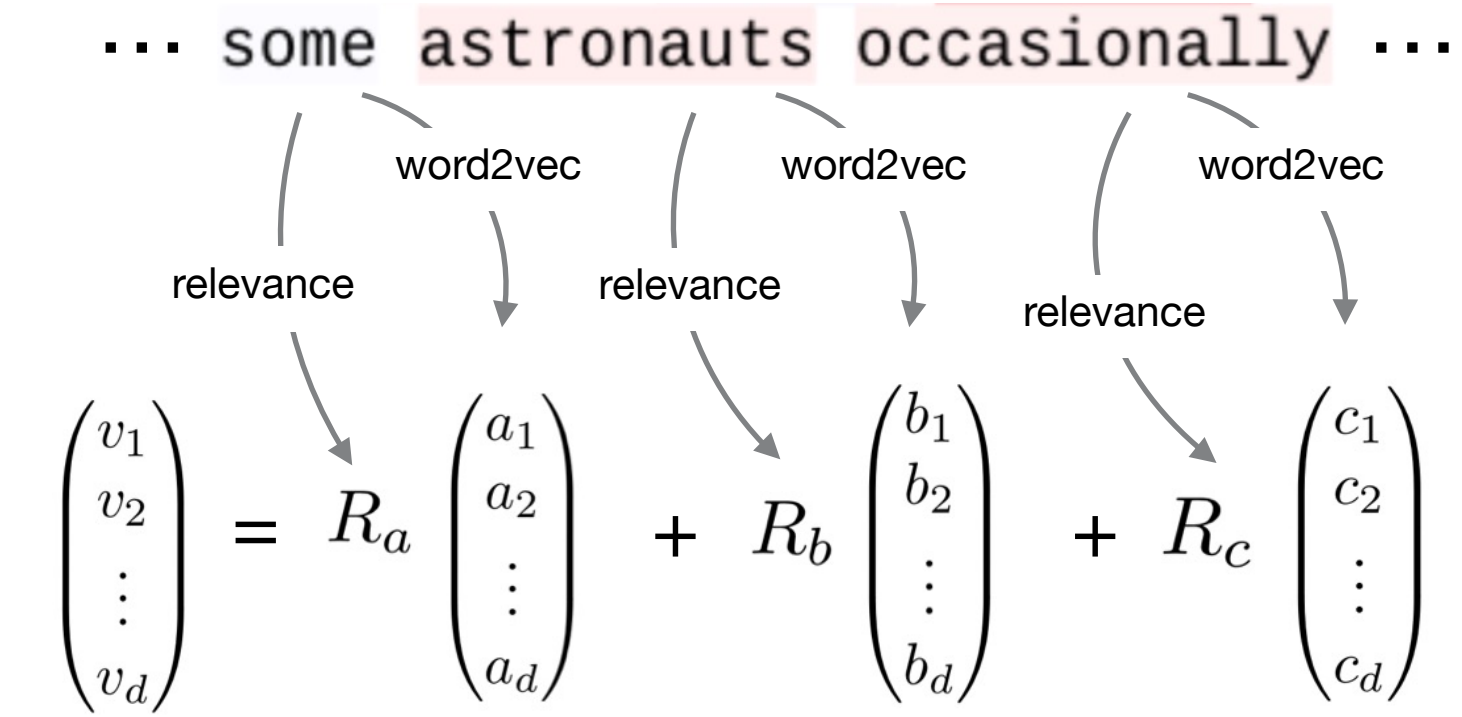
network seems to compare different structures

network seems to identify “original” parts

(Seibold et al., 2018)

Application of LRP
Learn new Representations

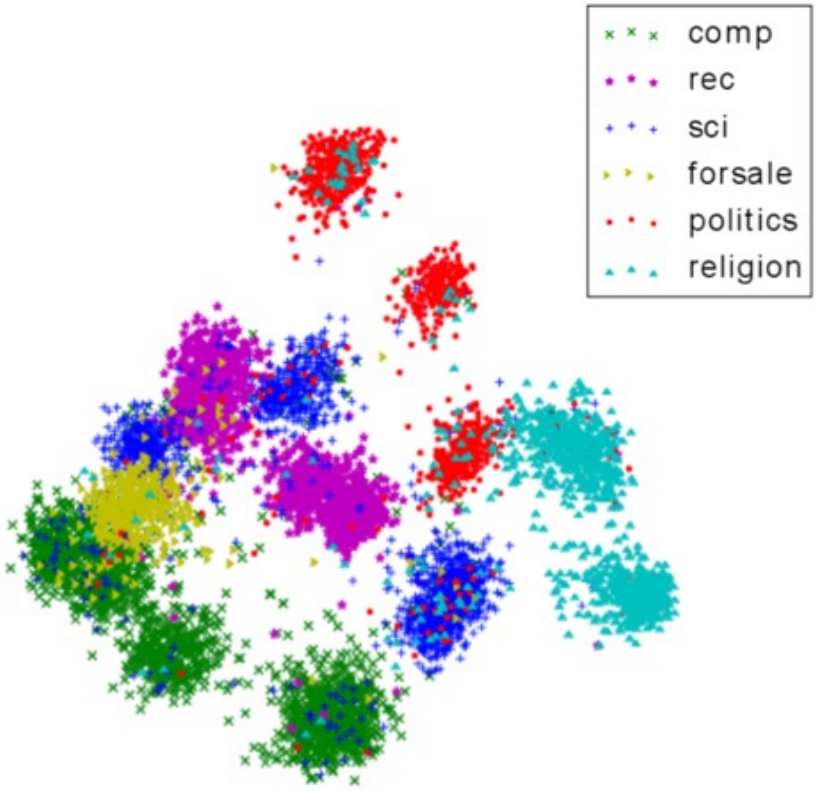
Application: Learn new Representations



(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).

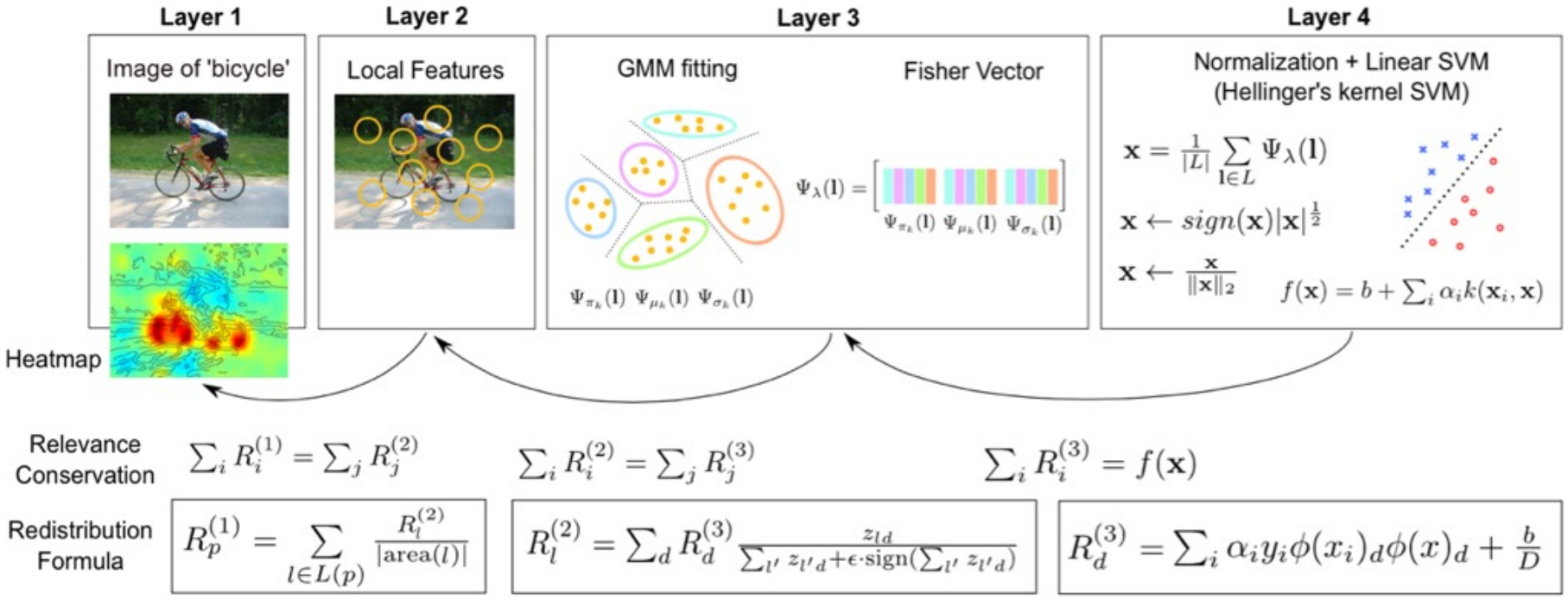
(Arras et al. 2016 & 2017)

Application of LRP

Understand Model &
Obtain new Insights

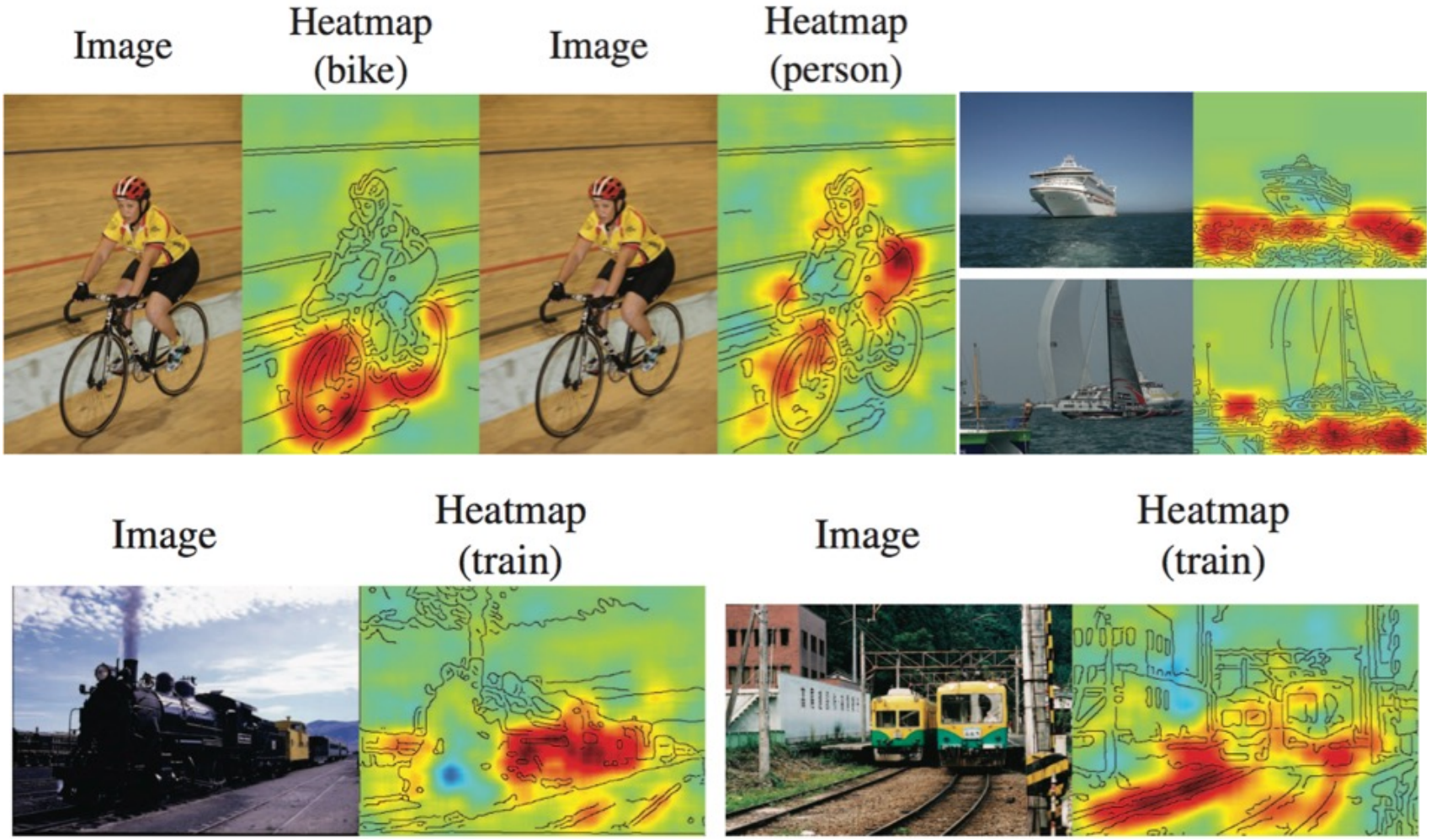
Application: Understand the model

- Fisher Vector / SVM classifier
- PASCAL VOC 2007



(Lapuschkin et al. 2016)

Application: Understand the model

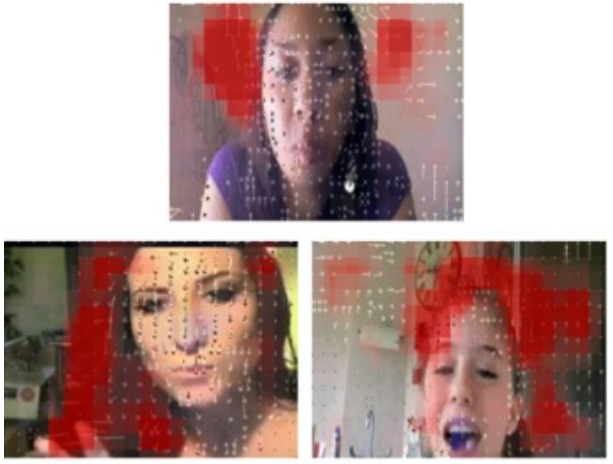


(Lapuschkin et al. 2016)

Application: Understand the model

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

- Fisher Vector / SVM classifier
- Model of Kantorov & Laptev, (CVPR'14)
- Histogram Of Flow, Motion Boundary Histogram
- HMDB51 dataset



(Srinivasan et al. 2017)

Application: Understand the model

movie review:
++, −

- bidirectional LSTM model (Li'16)
- Stanford Sentiment Treebank dataset

How to handle multiplicative interactions ?

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

gate neuron indirectly affect relevance distribution in forward pass

Negative sentiment

... too slow , too boring , and occasionally annoying .

it 's neither as romantic nor as thrilling as it should be .

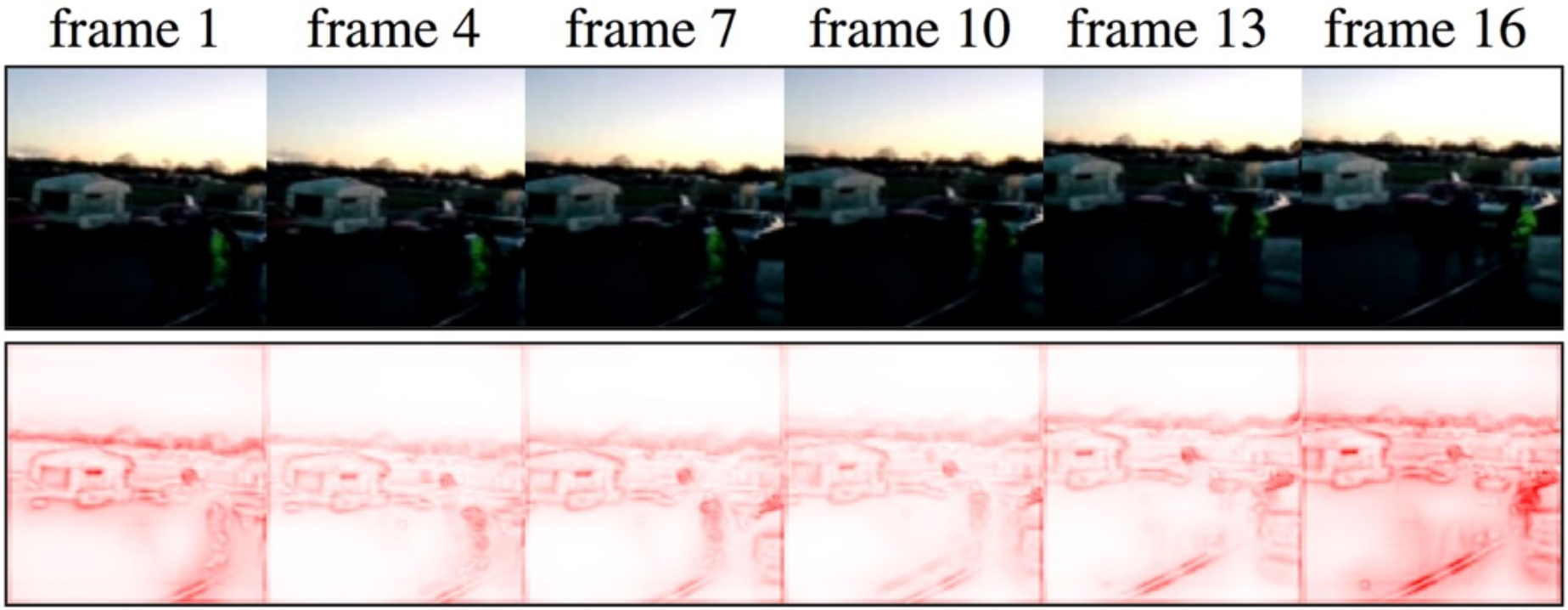
neither funny nor suspenseful nor particularly well-drawn .

Model understands negation !

(Arras et al., 2017 & 2018)

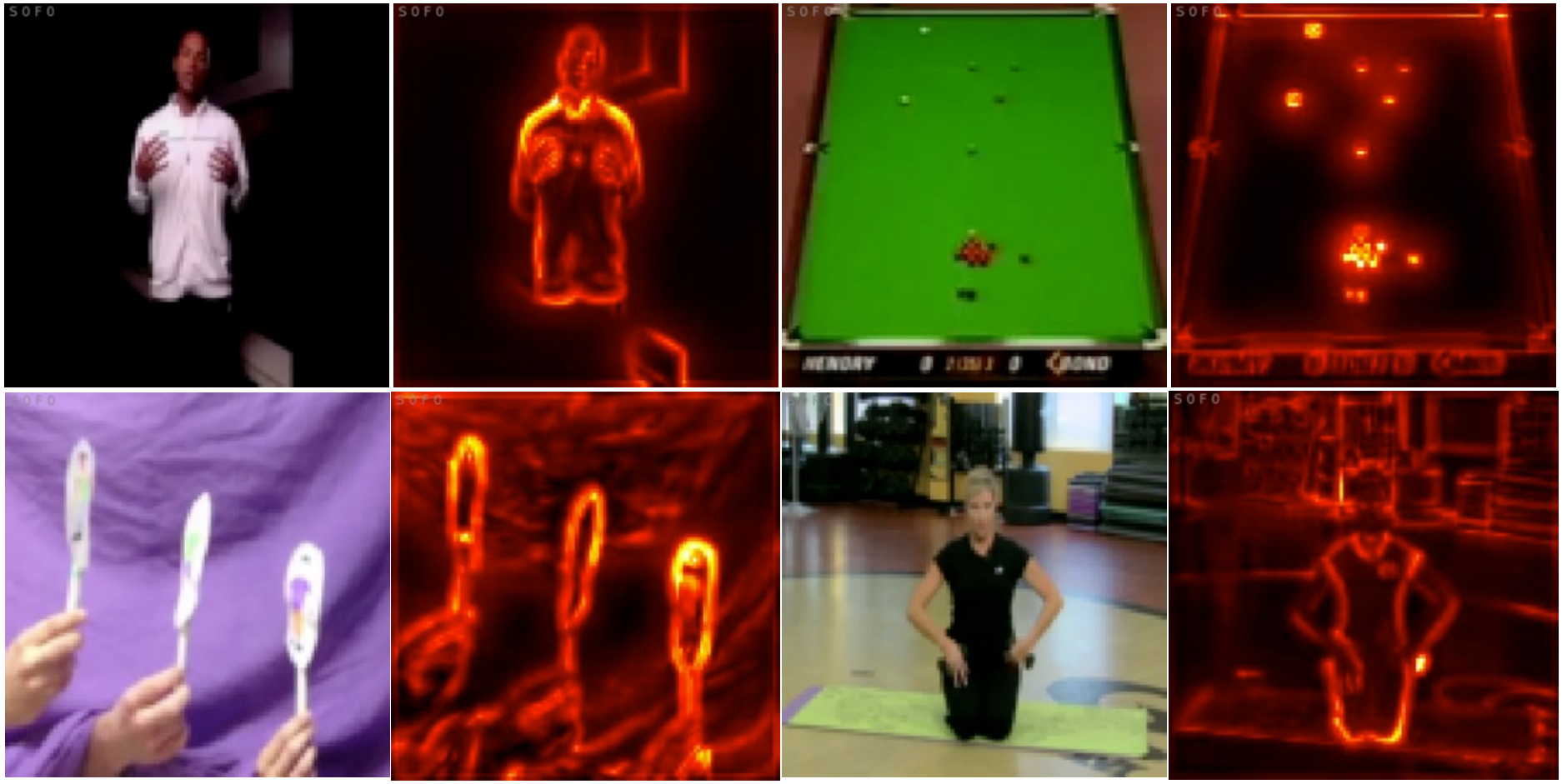
Application: Understand the model

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



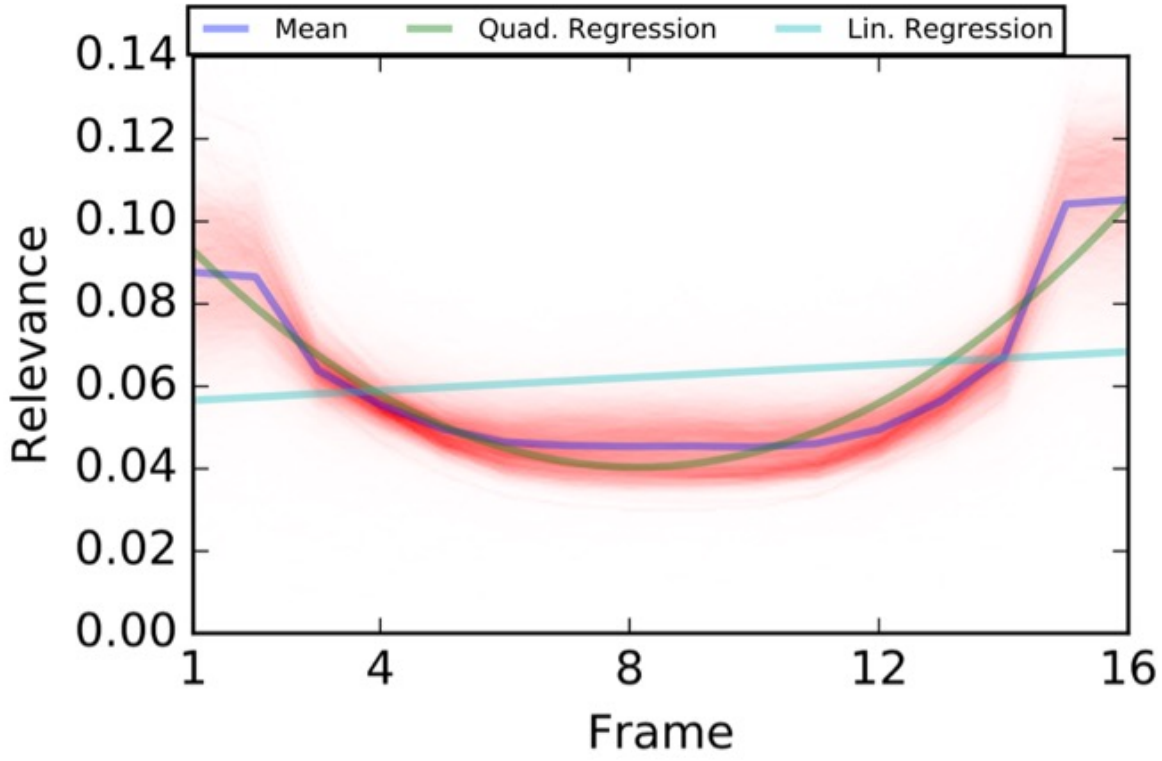
(Anders et al., 2018)

Application: Understand the model



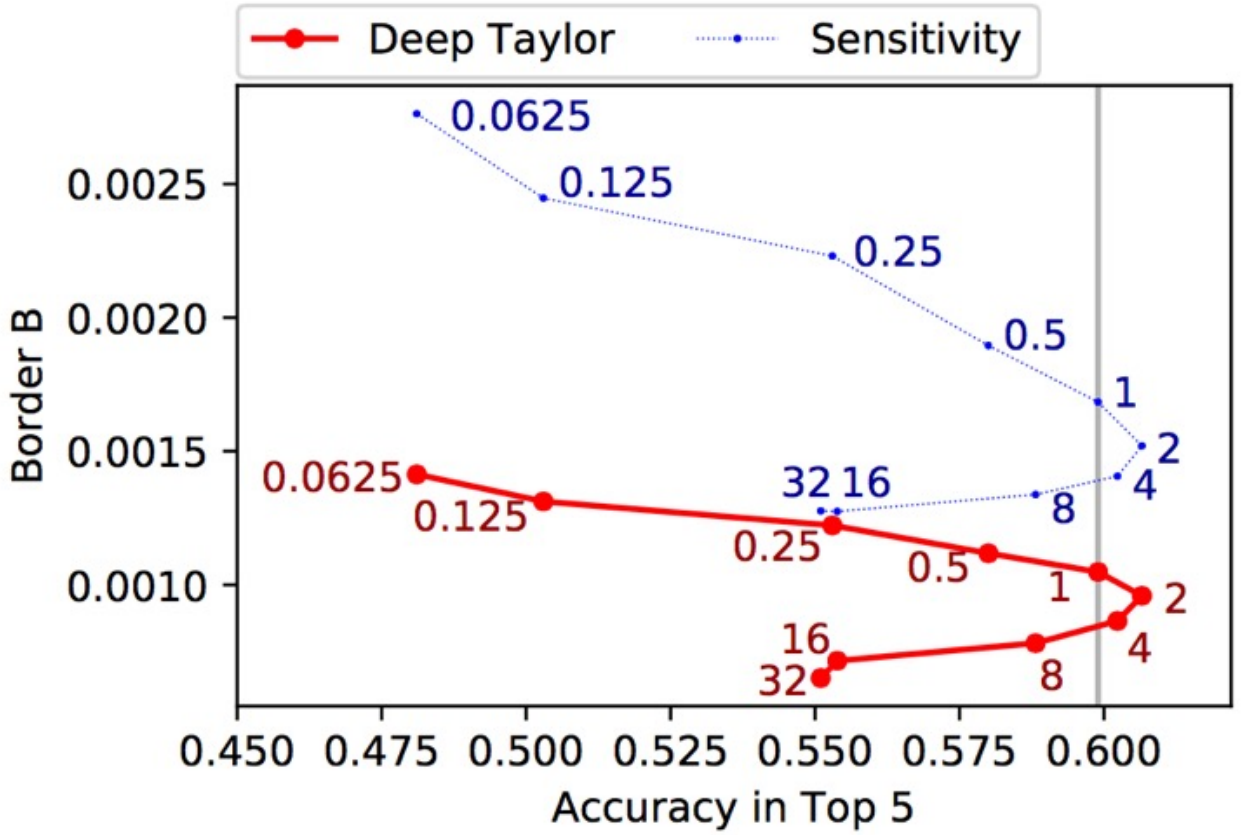
(Anders et al., 2018)

Application: Understand the model



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

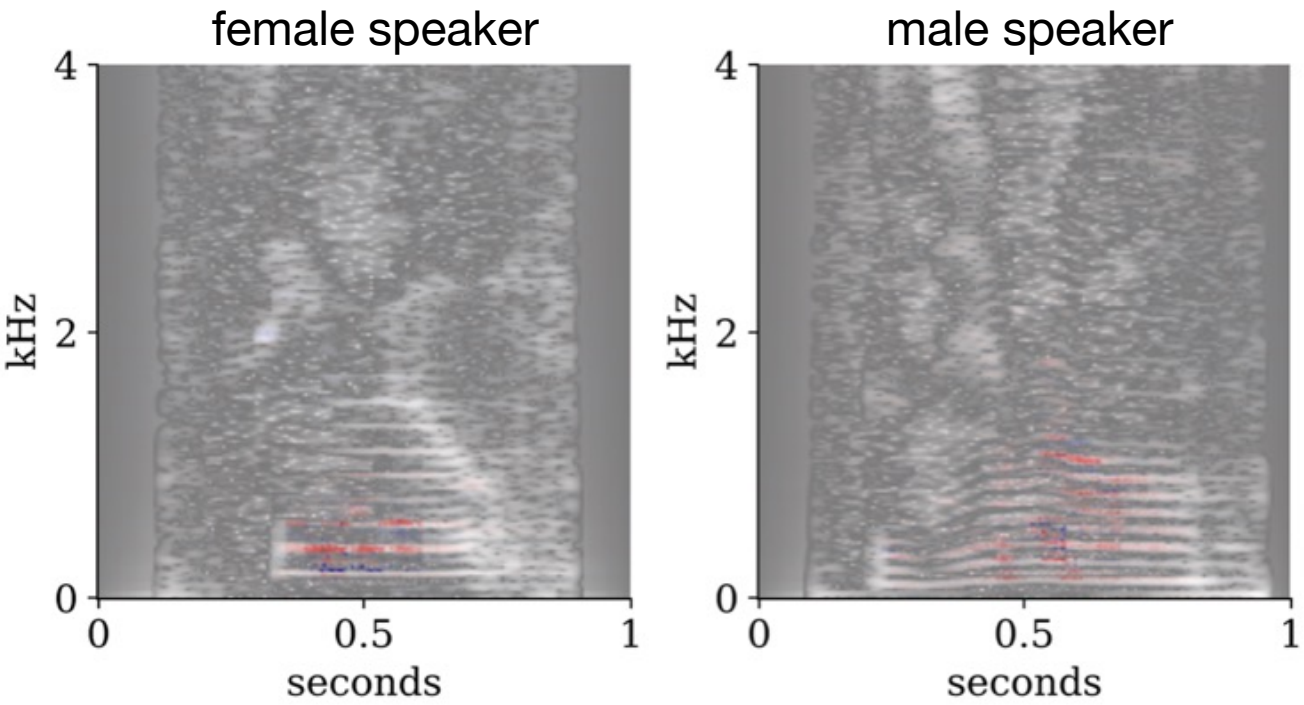
Application: Understand the model



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

Application: Understand the model

- AlexNet model
- trained on spectrograms
- spoken digits dataset (AudioMNIST)

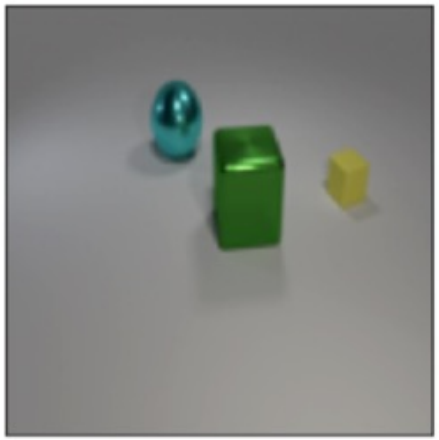


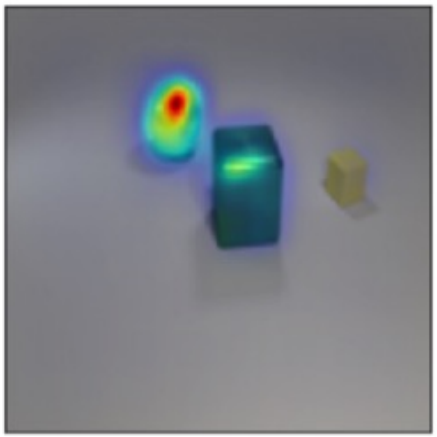
model classifies gender based on the fundamental frequency and its immediate harmonics (see also Traunmüller & Eriksson 1995)

(Becker et al., 2018)

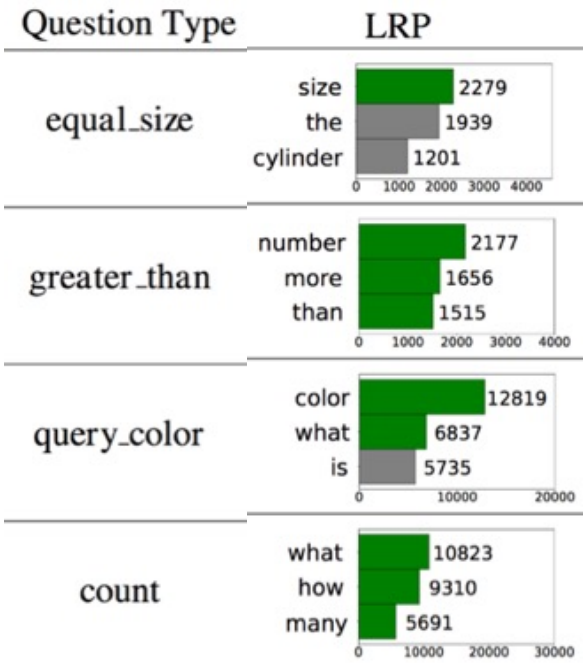
Application: Understand the model

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



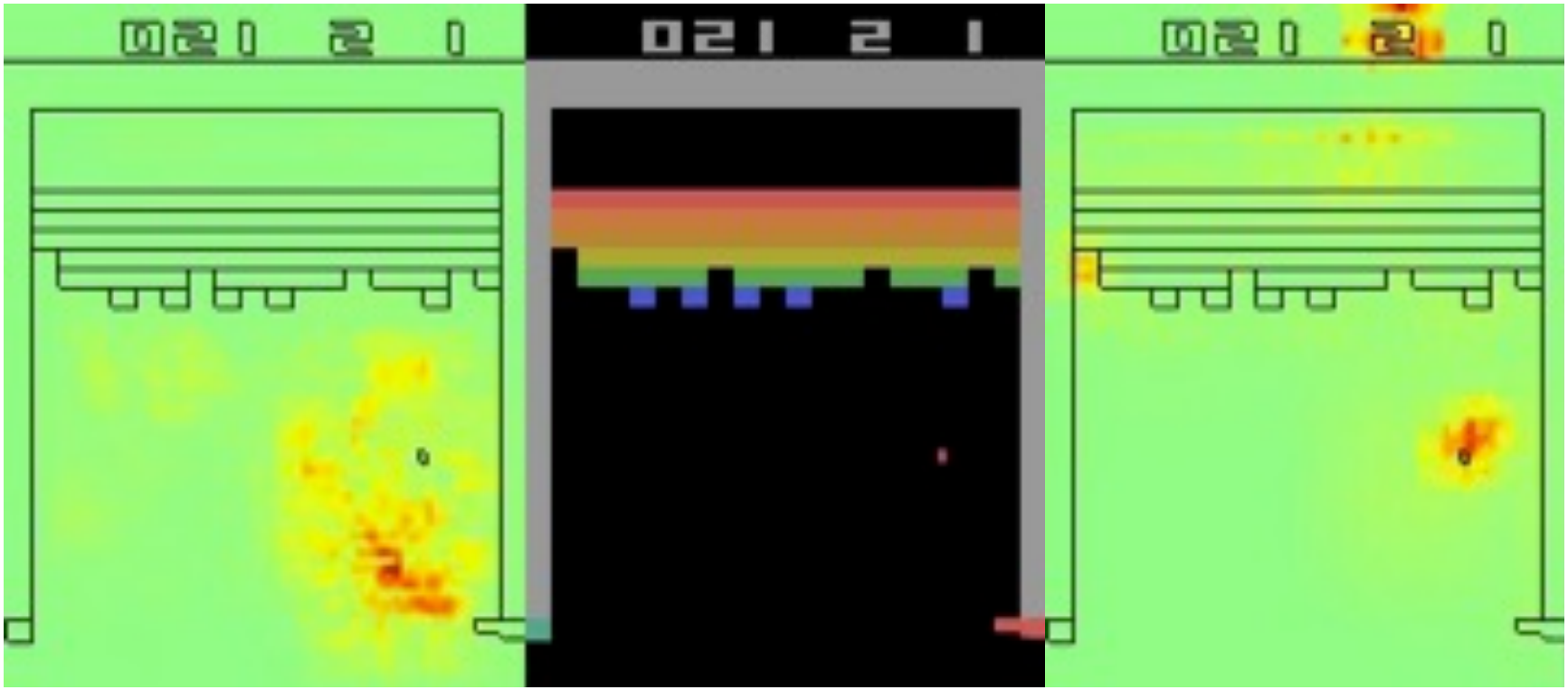
model understands the question and correctly identifies the object of interest

(Arras et al., 2018)

Application: Understand the model

Sensitivity Analysis

LRP



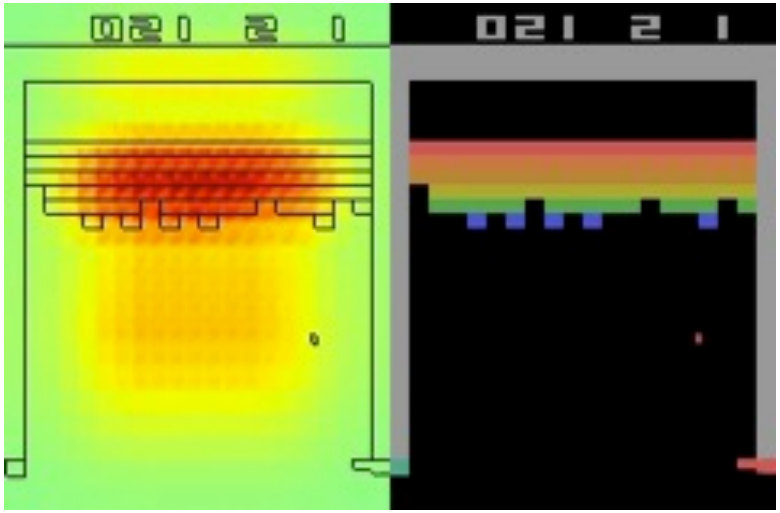
does not focus on where the ball is, but on where the ball could be in the next frame

LRP shows that that model tracks the ball

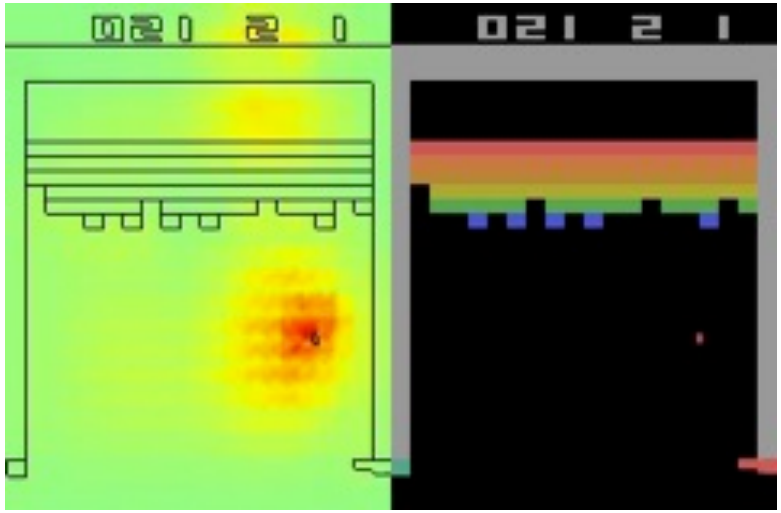
(Lapuschkin et al., in prep.)

Application: Understand the model

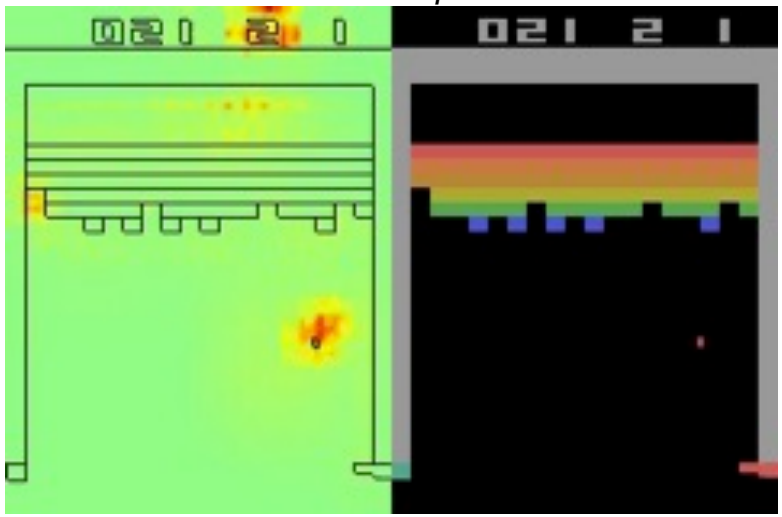
After 0 epochs



After 25 epochs

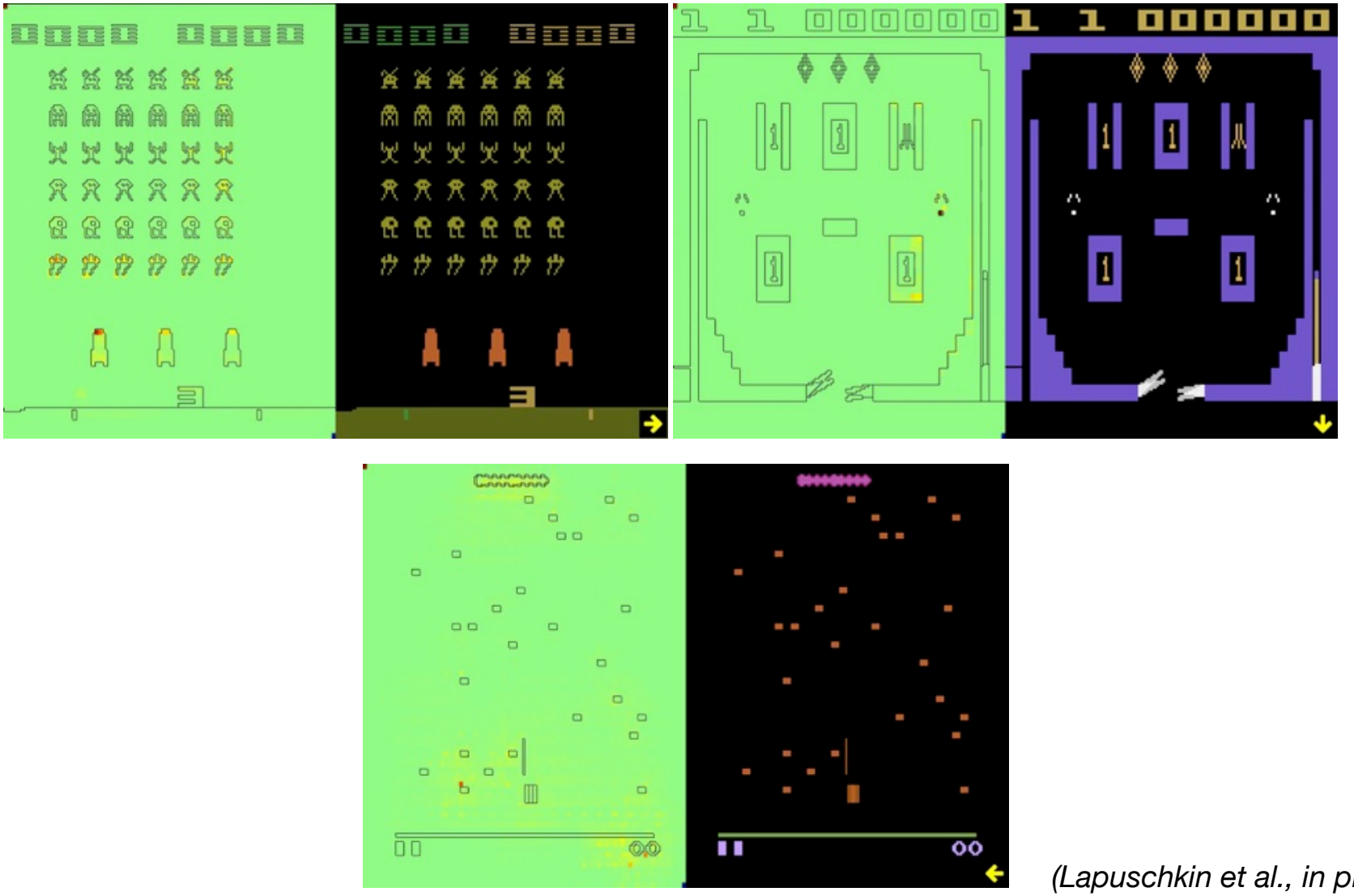


After 195 epochs



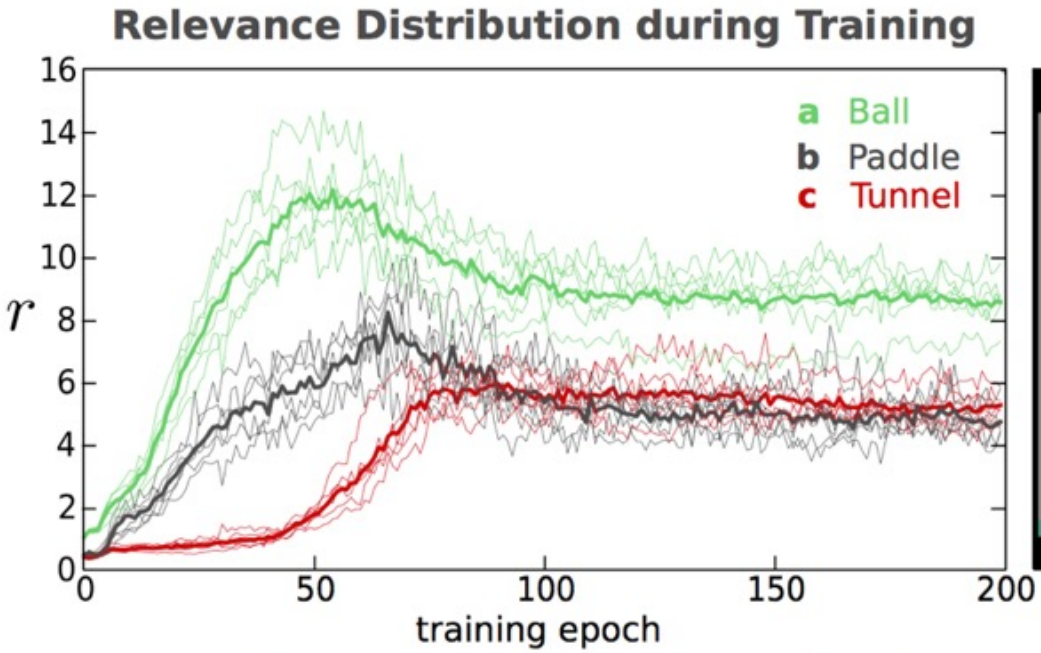
(Lapuschkin et al., in prep.)

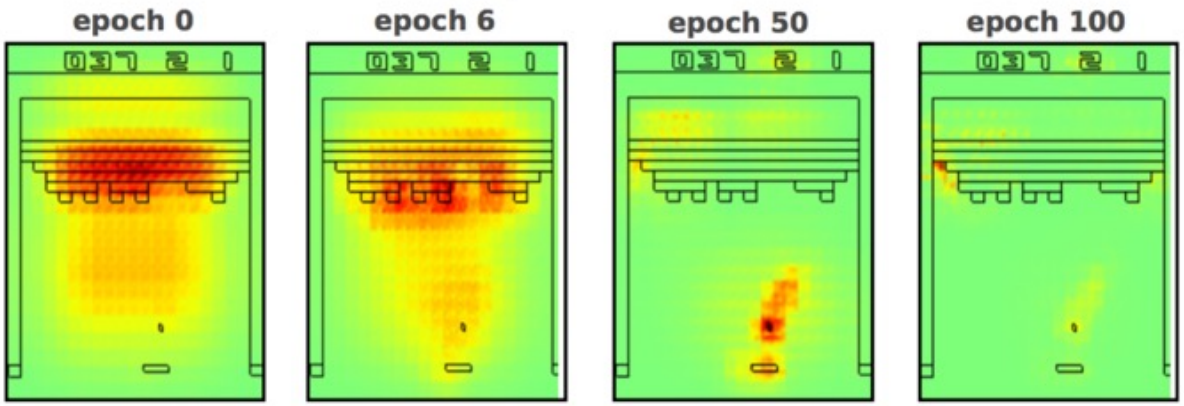
Application: Understand the model



(Lapuschkin et al., in prep.)

Application: Understand the model





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

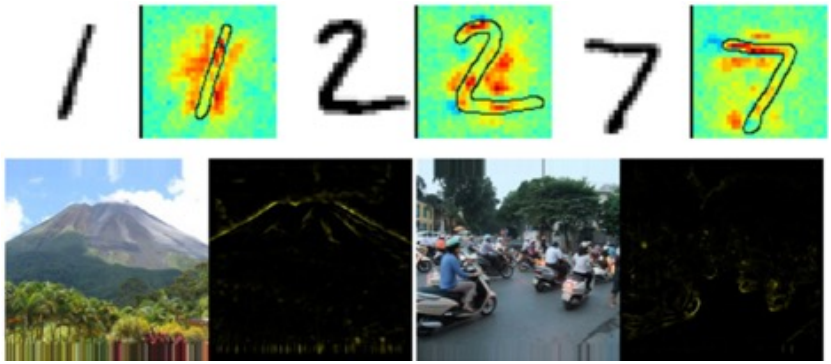
(Lapuschkin et al., in prep.)

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

Tutorial / Overview Papers

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

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

Methods Papers

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

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

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

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

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

Evaluation Explanations

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

References

Application to Text

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

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

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

L Arras, A Osman, G Montavon, KR Müller, W Samek. Evaluating and Comparing Recurrent Neural Network Explanation Methods in NLP. *arXiv*, 2018

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*, 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,* 2018.

Application to the Sciences

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

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

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

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