Tutorials

Interpretable Deep Learning: Towards Understanding & Explaining DNNs

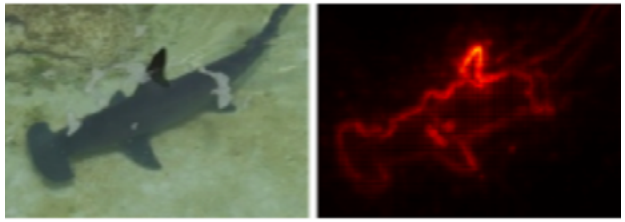
Part 4: Applications

Wojciech Samek, Grégoire Montavon, Klaus-Robert Müller

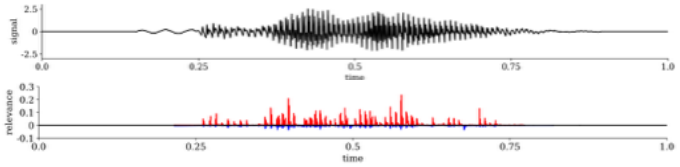
Technische Universität Berlin
berlin
Fraunhofer Heinrich Hertz Institute
Berliner Zentrum für MASCHINELLES LERNEN

LRP revisited

General Images (Bach' 15, Lapuschkin'16)



Speech (Becker'18)

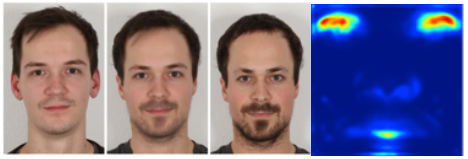


Text Analysis (Arras'16 &17)

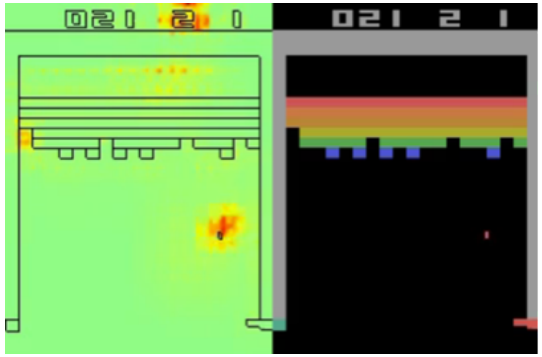
do n't waste your money

neither funny nor susper

Morphing (Seibold'18)



Games (Lapuschkin'18)

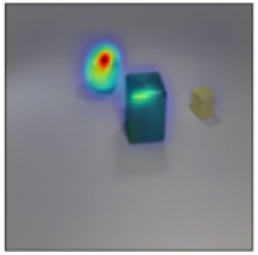


VQA (Arras'18)

there is a metallic cube ; are

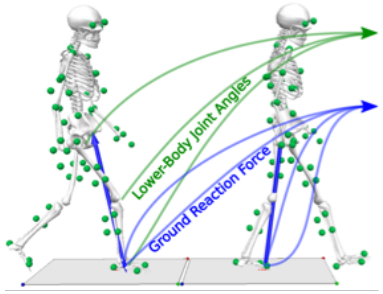
there any large cyan metallic

objects behind it ?

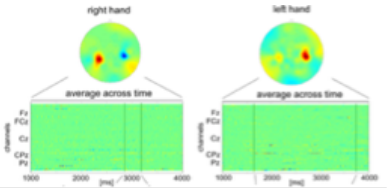


Video (Anders'18)

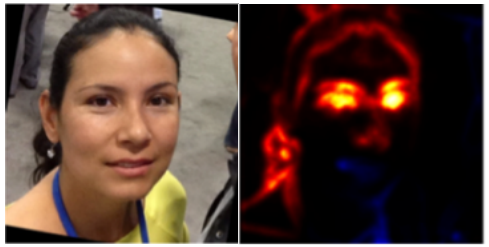
Gait Patterns (Horst'18)



EEG (Sturm'16)

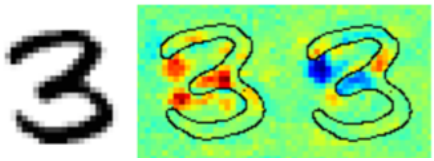


Faces (Lapuschkin'17)

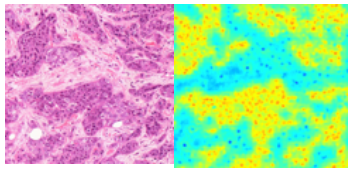


Digits (Bach' 15)

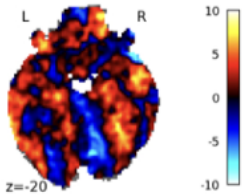
Image Class '3' Class '9'



Histopathology (Binder'18)

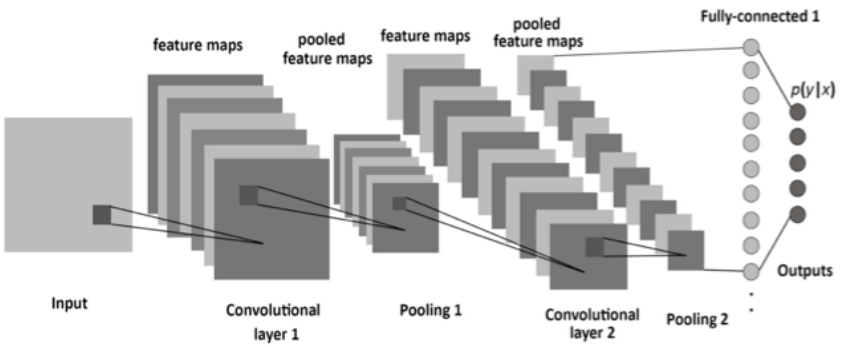


fMRI (Thomas'18)

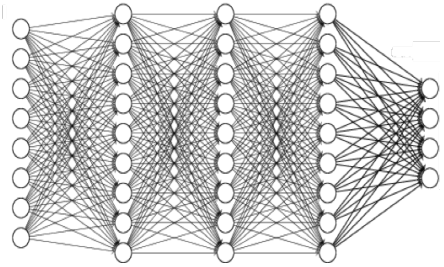


LRP revisited

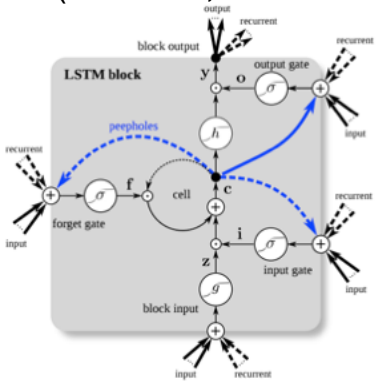
Convolutional NNs (Bach'15, Arras'17 …)



Local Renormalization Layers (Binder'16)

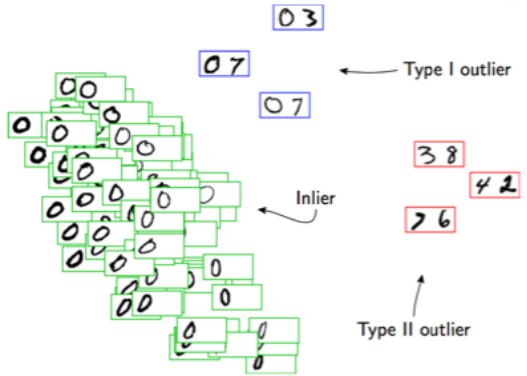


LSTM (Arras'17, Thomas'18)



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

One-class SVM (Kauffmann'18)



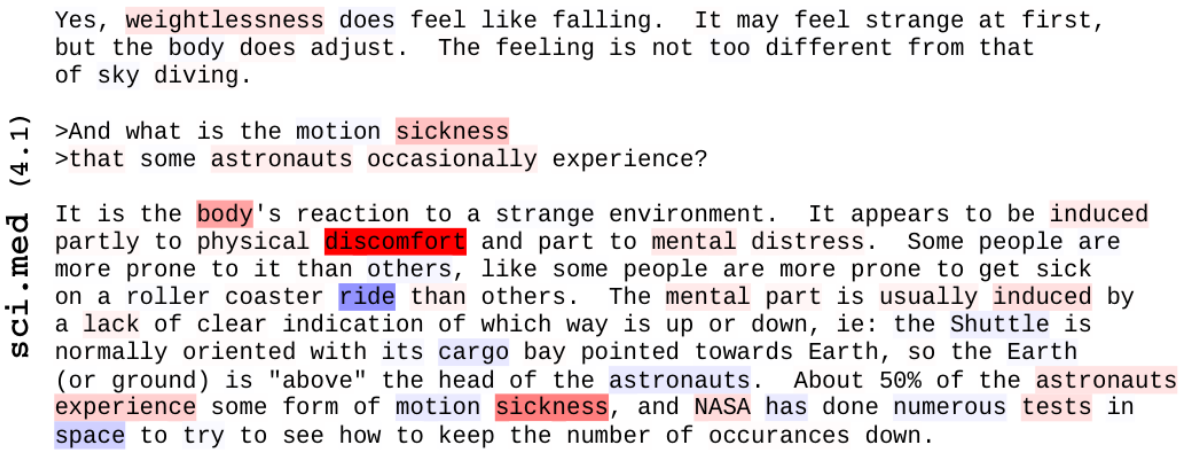
Application of LRP
Compare models

Application: Compare Classifiers

word2vec/CNN:

Performance: 80.19%

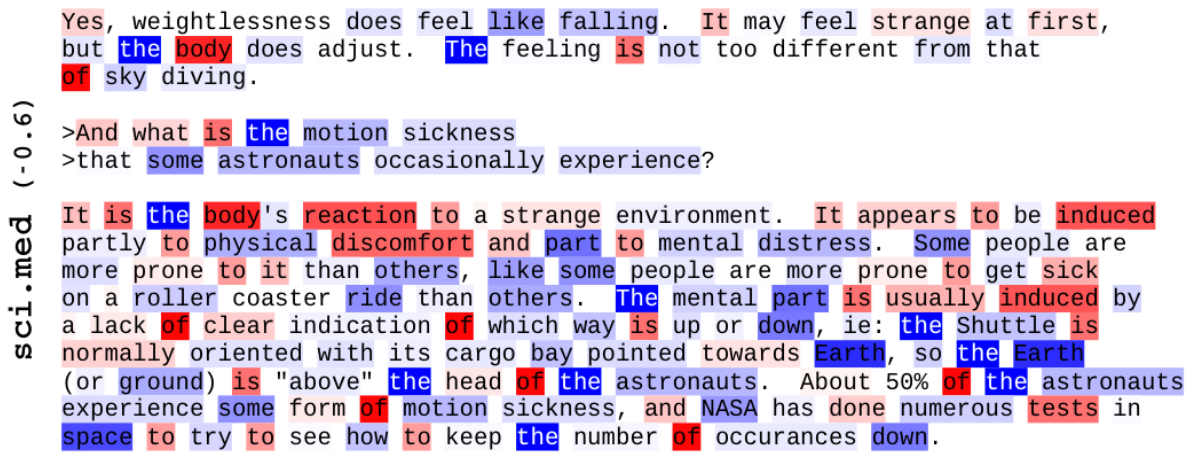
Strategy to solve the problem: identify semantically meaningful words related to the topic.



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)

LRP in Practice

Visual Object Classes Challenge: 2005 - 2012

	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	diningtable	dog	horse	motorbike	person	pottedplant	sheep	sofa	train	tvmonitor
INRIA_Flat	*74.8*	*62.5*	*51.2*	*69.4*	*29.2*	*60.4*	*76.3*	*57.6*	*53.1*	*41.1*	*54.0*	*42.8*	*76.5*	*62.3*	*84.5*	*35.3*	*41.3*	*50.1*	*77.6*	*49.3*
INRIA_Genetic	**77.5**	**63.6**	**56.1**	**71.9**	**33.1**	**60.6**	**78.0**	**58.8**	**53.5**	**42.6**	**54.9**	**45.8**	**77.5**	**64.0**	**85.9**	**36.3**	**44.7**	*50.6*	**79.2**	**53.2**
INRIA_Larlus	62.6	54.0	32.8	47.5	17.8	46.4	69.6	44.2	44.6	26.0	38.1	34.0	66.0	55.1	77.2	13.1	29.1	36.7	62.7	43.3
MPI_BOW	58.9	46.0	31.3	59.0	16.9	40.5	67.2	40.2	44.3	28.3	31.9	34.4	63.6	53.5	75.7	22.3	26.6	35.4	60.6	40.6
PRIPUVA	48.6	20.9	21.3	17.2	6.4	14.2	45.0	31.4	27.4	12.3	14.3	23.7	30.1	13.3	62.0	10.0	12.4	13.3	26.7	26.2
QMUL_HSLS	70.6	54.8	35.7	64.5	27.8	51.1	71.4	54.0	46.6	36.6	34.4	39.9	71.5	55.4	80.6	15.8	35.8	41.5	73.1	45.5
QMUL_LSPCH	71.6	55.0	41.1	65.5	27.2	51.1	72.2	*55.1*	47.4	35.9	37.4	41.5	71.5	57.9	80.8	15.6	33.3	41.9	*76.5*	45.9
TKK	71.4	51.7	48.5	63.4	27.3	49.9	70.1	51.2	51.7	32.3	46.3	41.5	72.6	*60.2*	82.2	31.7	30.1	39.2	71.1	41.0
ToshCam_rdf	59.9	36.8	29.9	40.0	23.6	33.3	60.2	33.0	41.0	17.8	33.2	33.7	63.9	53.1	77.9	29.0	27.3	31.2	50.1	37.6
ToshCam_svm	54.0	27.1	30.3	35.6	17.0	22.3	58.0	34.6	38.0	19.0	27.5	32.4	48.0	40.7	78.1	23.4	21.8	28.0	45.5	31.8
Tsinghua	62.9	42.4	33.9	49.7	23.7	40.7	62.0	35.2	42.7	21.0	38.9	34.7	65.0	48.1	76.9	16.9	30.8	32.8	58.9	33.1
UVA_Bigrams	61.2	33.2	29.4	45.0	16.5	37.6	54.6	31.3	39.9	17.2	31.4	30.6	61.6	42.4	74.6	14.5	20.9	23.5	49.9	30.0
UVA_FuseAll	67.1	48.1	43.3	58.1	19.9	46.3	61.8	41.9	48.4	27.8	41.9	38.5	69.8	51.4	79.4	32.5	31.9	36.0	66.2	40.3
UVA_MCIP	66.5	47.9	41.0	58.0	16.8	44.0	61.2	40.5	48.5	27.8	41.7	37.1	66.4	50.1	78.6	31.2	32.3	31.9	66.6	40.3
UVA_SFS	66.3	49.7	43.5	60.7	18.8	44.9	64.8	41.9	46.8	24.9	42.3	33.9	71.5	53.4	80.4	29.7	31.2	31.8	67.4	43.5
UVA_WGT	59.7	33.7	34.9	44.5	22.2	32.9	55.9	36.3	36.8	20.6	25.2	34.7	65.1	40.1	74.2	26.4	26.9	25.1	50.7	29.7
XRCE	*72.3*	*57.5*	*53.2*	*68.9*	*28.5*	*57.5*	*75.4*	50.3	*52.2*	*39.0*	*46.8*	*45.3*	*75.7*	58.5	*84.0*	*32.6*	*39.7*	**50.9**	75.1	*49.5*

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%

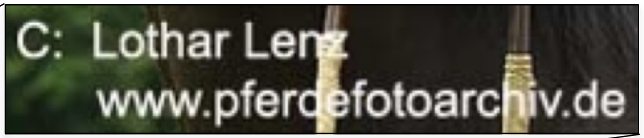
Image

same performance —> same strategy ?

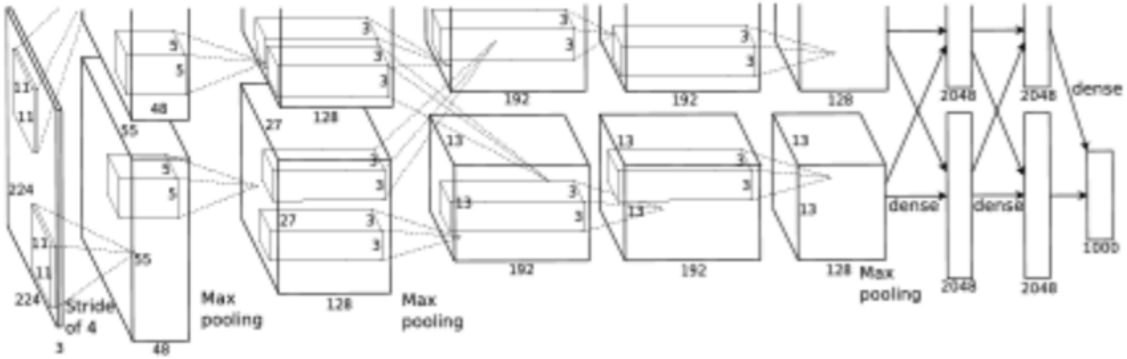
(Lapuschkin et al. 2016)

Application: Compare Classifiers

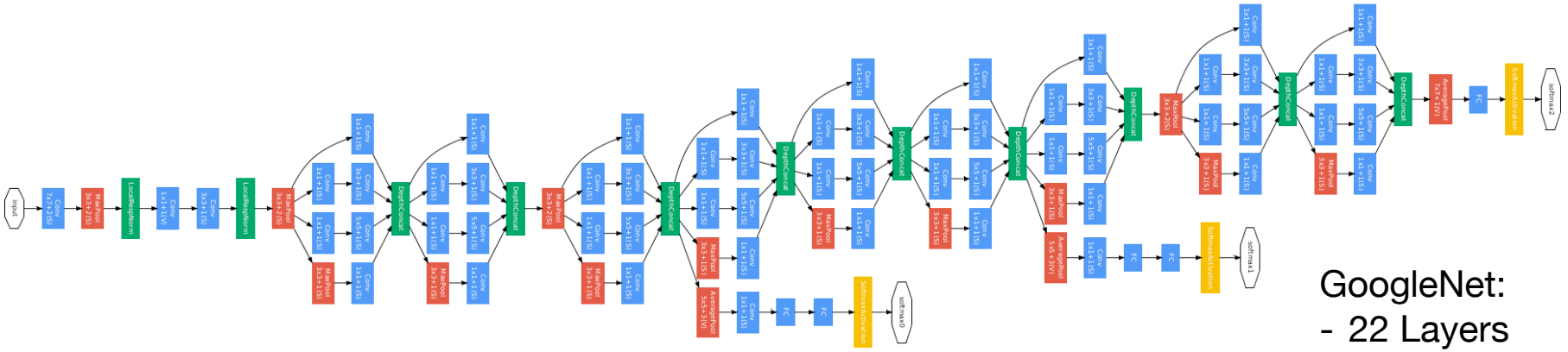
‘horse’ images in PASCAL VOC 2007



Application: Compare Classifiers



BVLC:
- 8 Layers
- ILSRCV: 16.4%



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

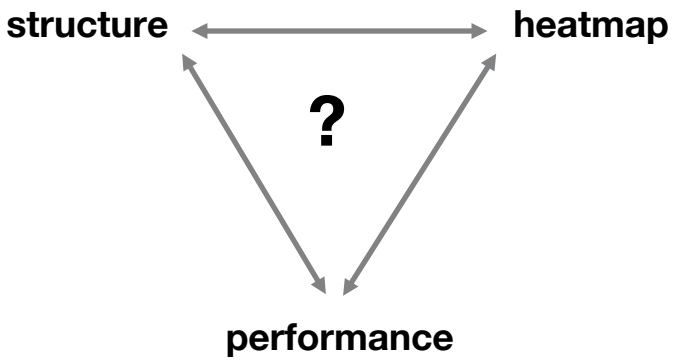
Application: Compare Classifiers

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

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

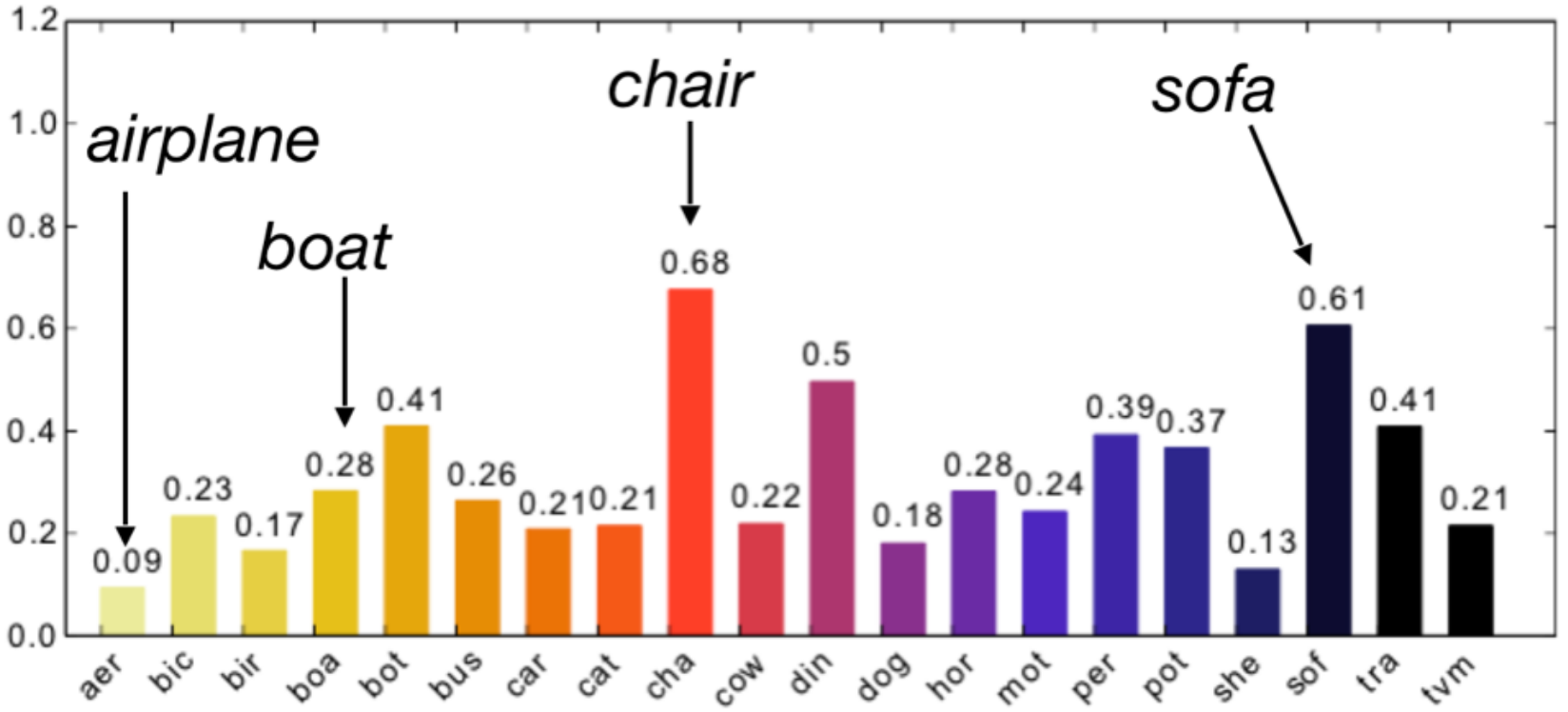
LRP decomposition allows meaningful pooling over bbox !

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

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

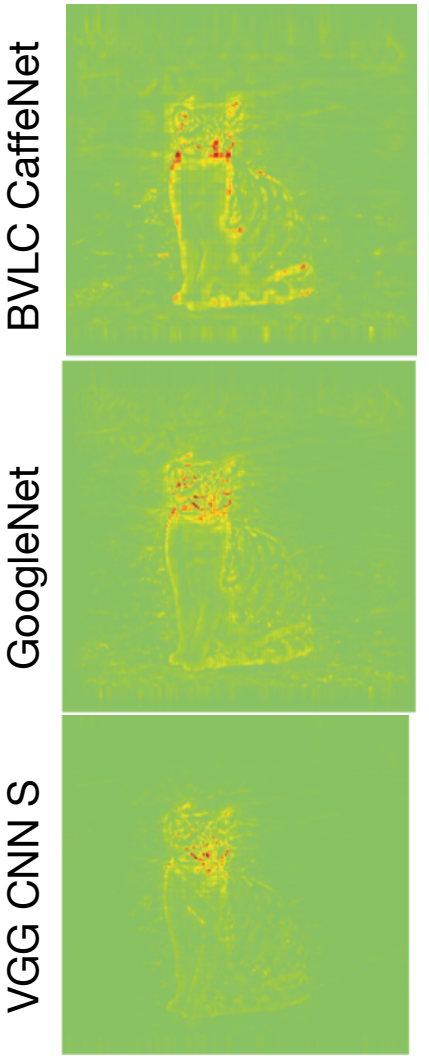
Application: Measure Context Use

- BVLC reference model + fine tuning
- PASCAL VOC 2007

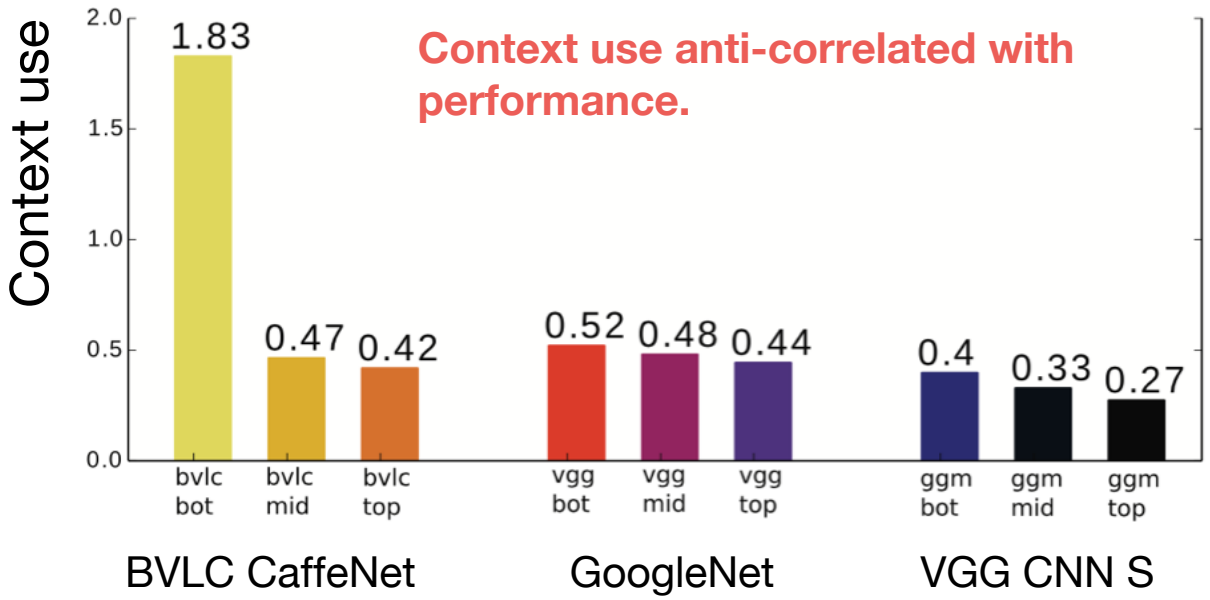


(Lapuschkin et al., 2016)

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 & Improve Models

Application: Face analysis

- Compare AdienceNet, CaffeNet, GoogleNet, VGG-16
- 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

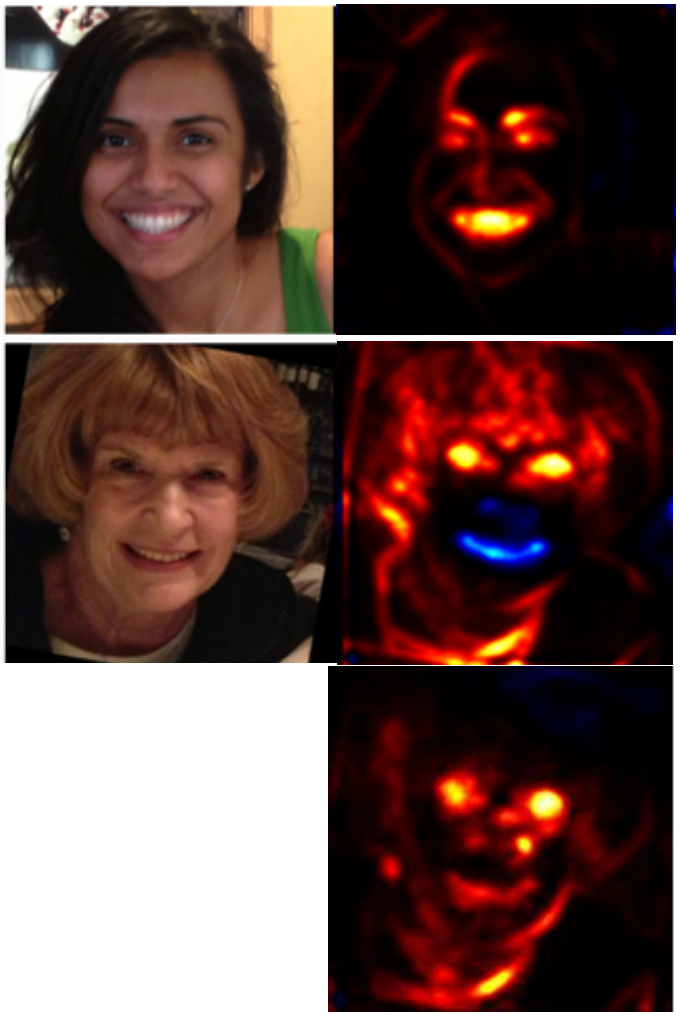
Gender classification

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

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: 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

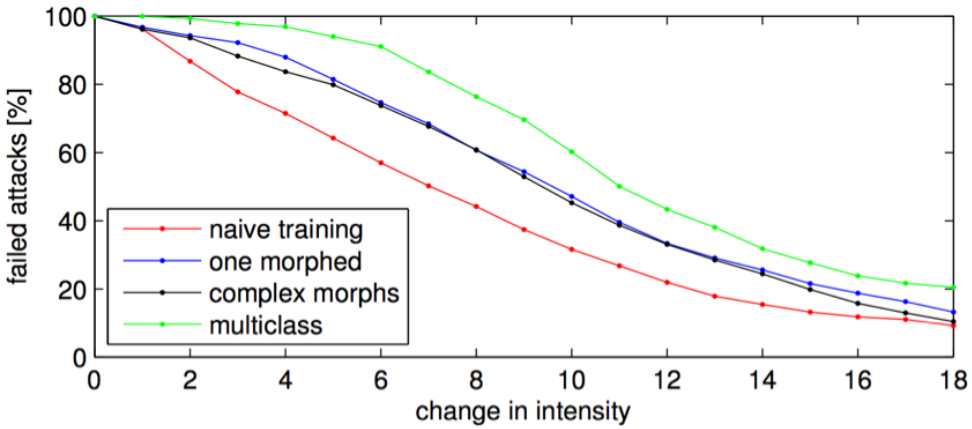


Fig. 5. Robustness against fast gradient sign attacks.

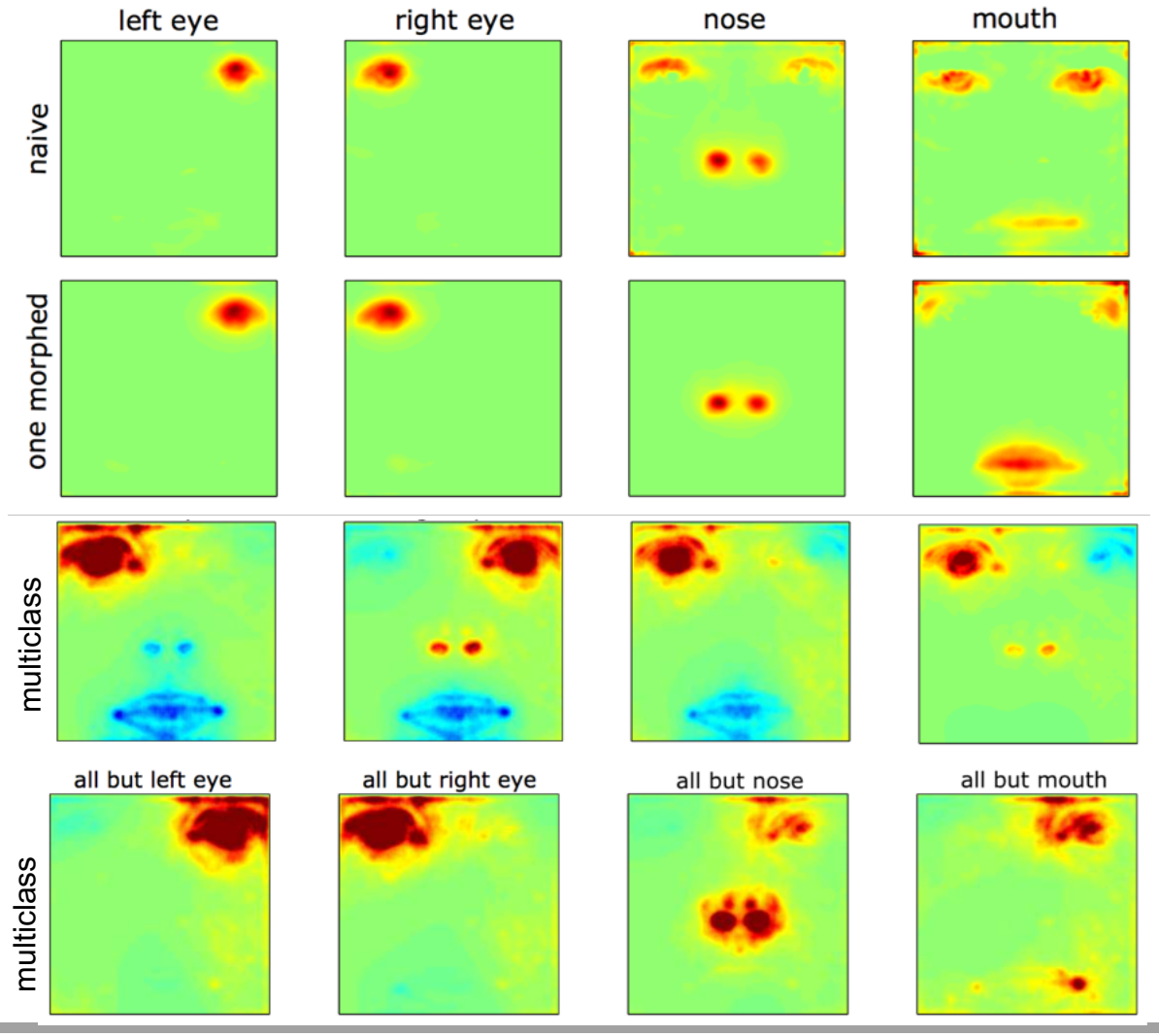
Application: Face analysis

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 !

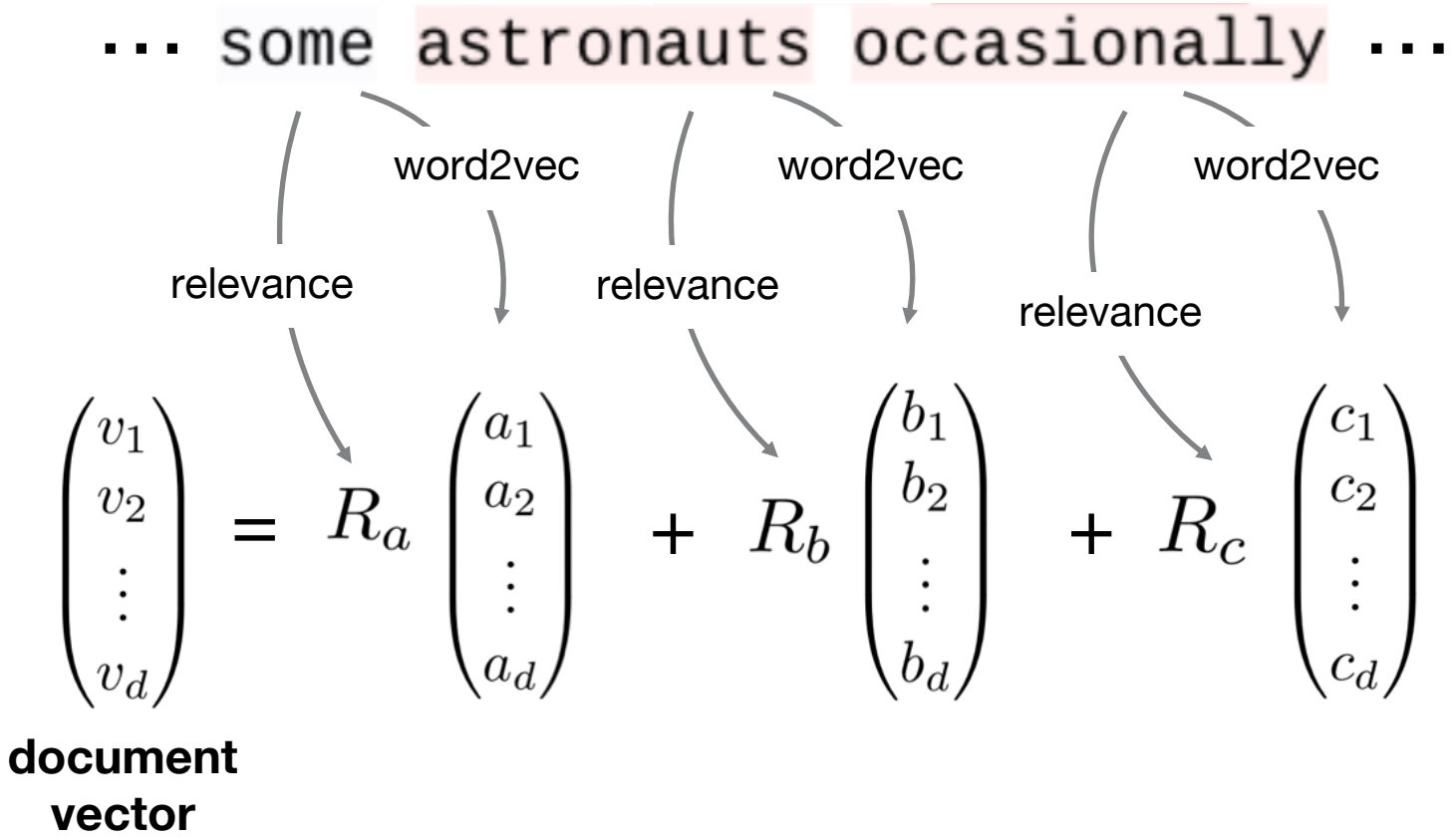
network seems to compare different structures

network seems to identify “original” parts

(Seibold et al., 2018)

Application of LRP
Learn new Representations

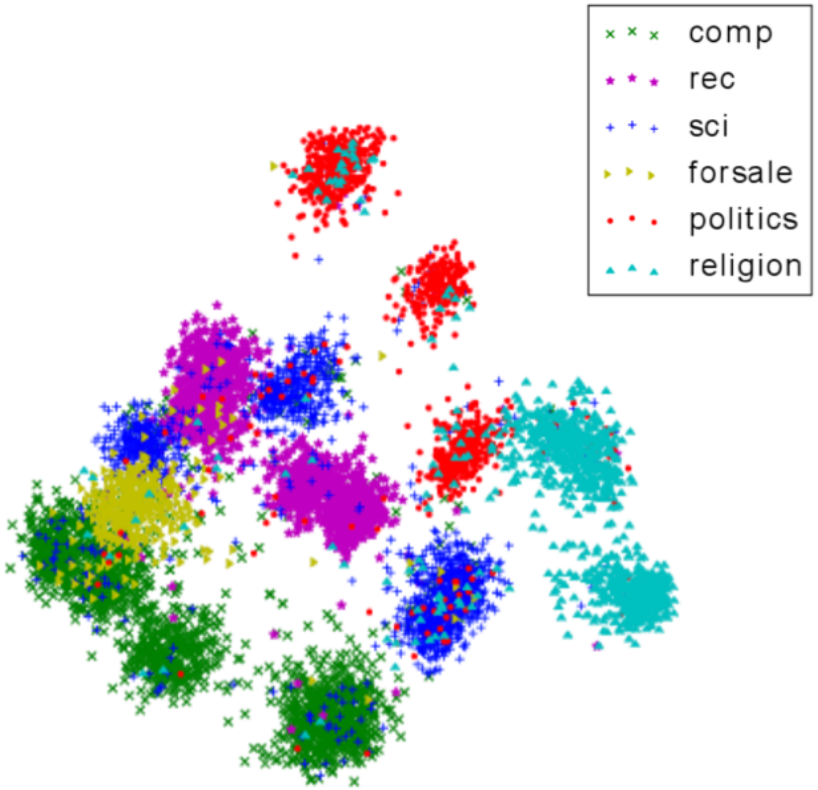
Application: Learn new Representations



(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).

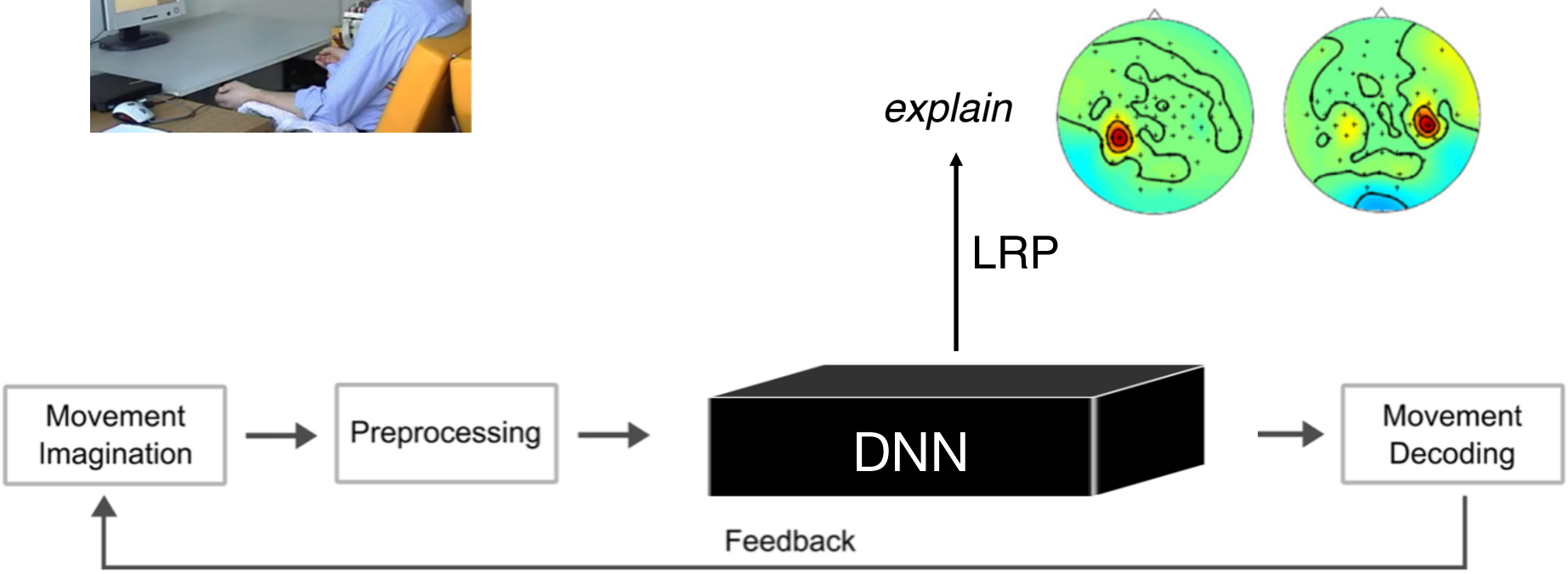
(Arras et al. 2016 & 2017)

Application of LRP
Interpreting Scientific Data

Application: EEG Analysis

Brain-Computer Interfacing

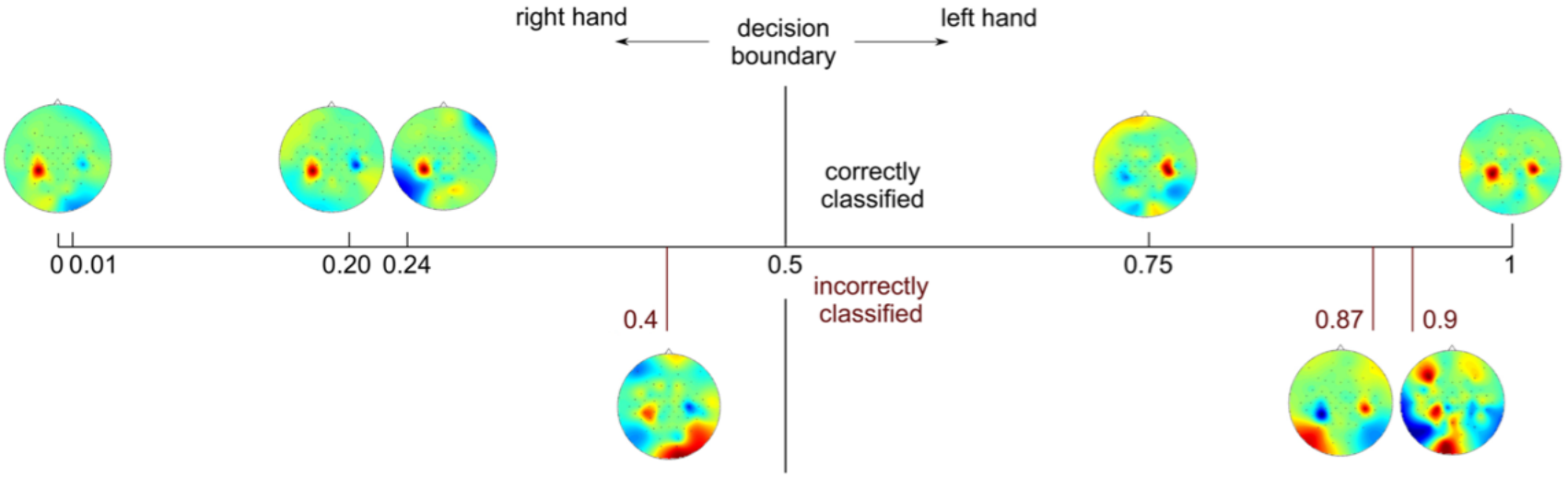
Neural network learns that:
Left hand movement imagination leads to desynchronization over right sensorimotor cortext (and vice versa).



(Sturm et al. 2016)

Application: EEG Analysis

Our neural networks are interpretable:
We can see for every trial "why" it is classified the way it is.

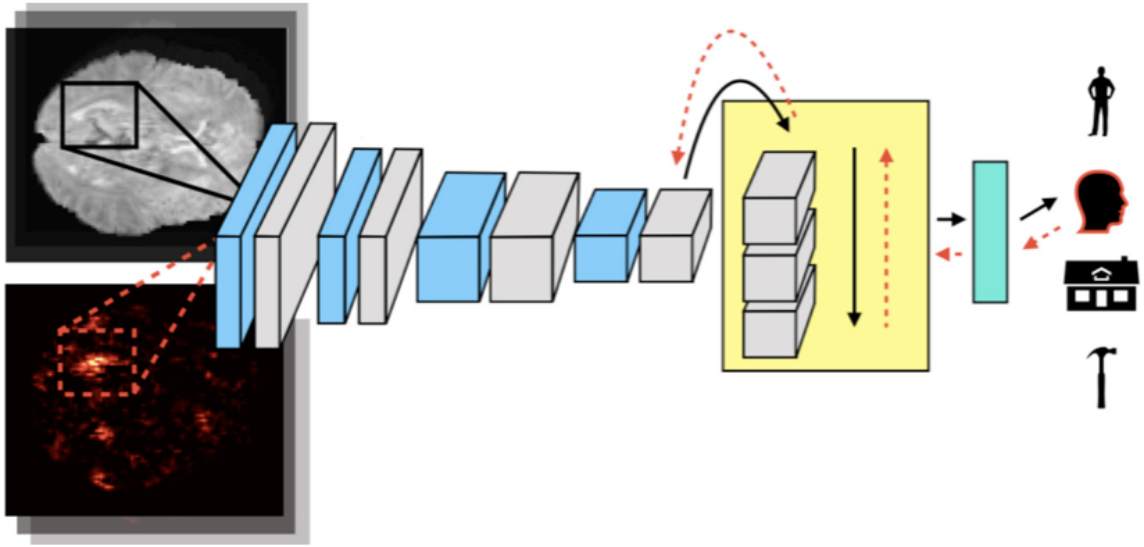


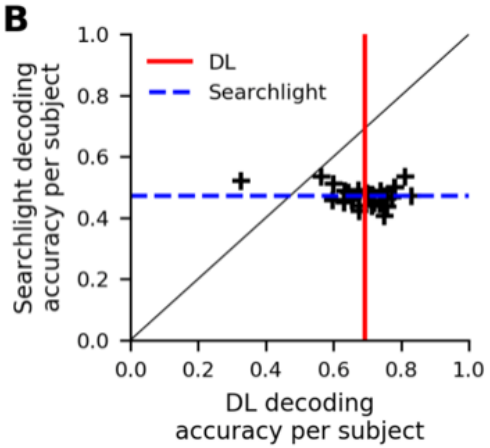
(Sturm et al. 2016)

Application: fMRI Analysis

Difficulty to apply deep learning to fMRI :
- high dimensional data (100 000 voxels), but only few subjects
- results must be interpretable (key in neuroscience)

Our approach:
- Recurrent neural networks (CNN + LSTM) for whole-brain analysis
- LRP allows to interpret the results

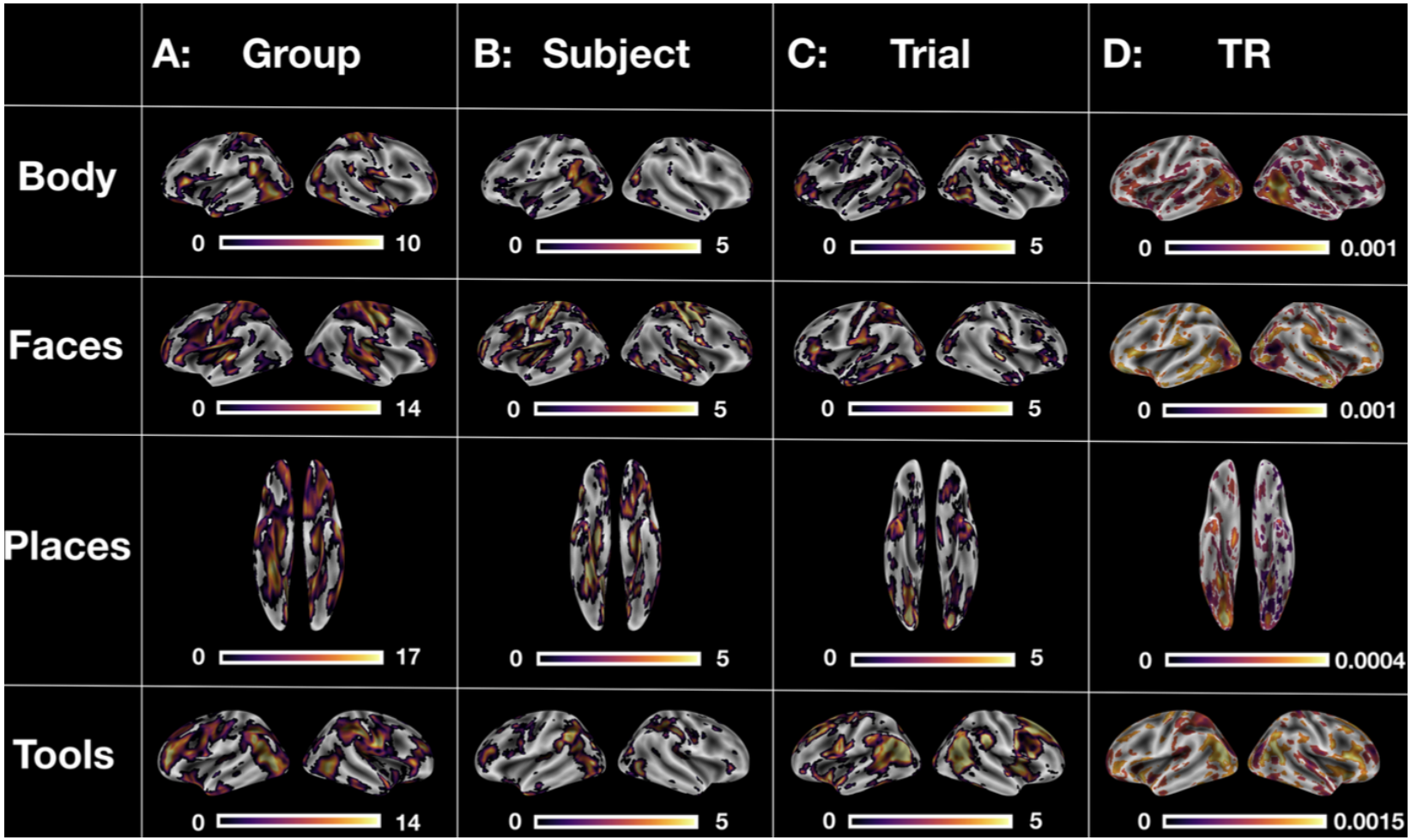




Dataset:
- 100 subjects from Human Connectome Project
- N-back task (faces, places, tools and body parts)

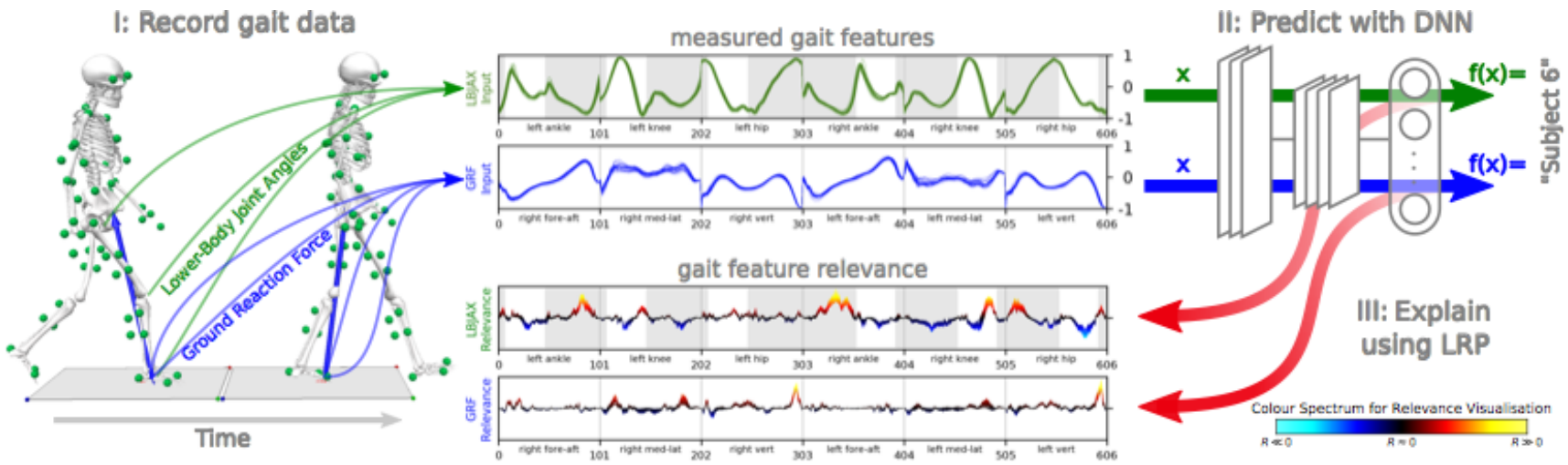
(Thomas et al. 2018)

Application: fMRI Analysis



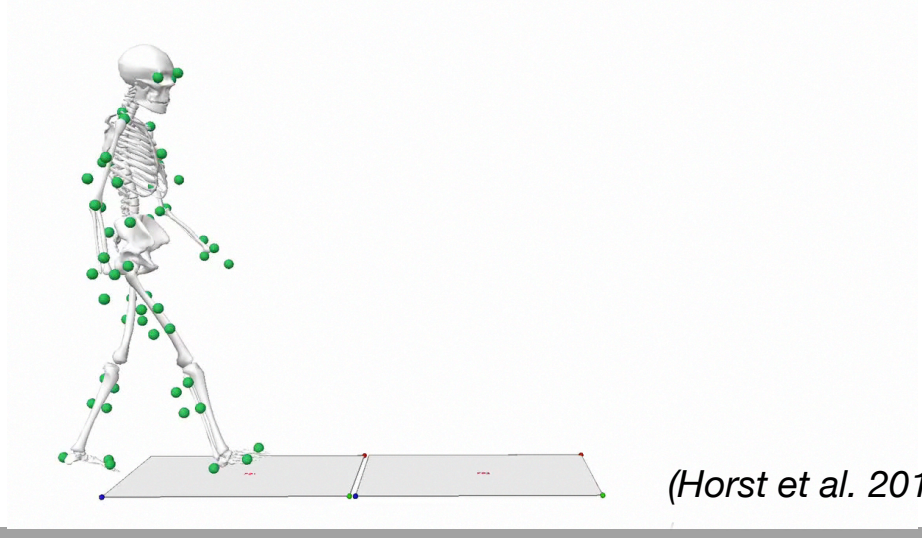
(Thomas et al. 2018)

Application: Gait Analysis



Our approach:

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

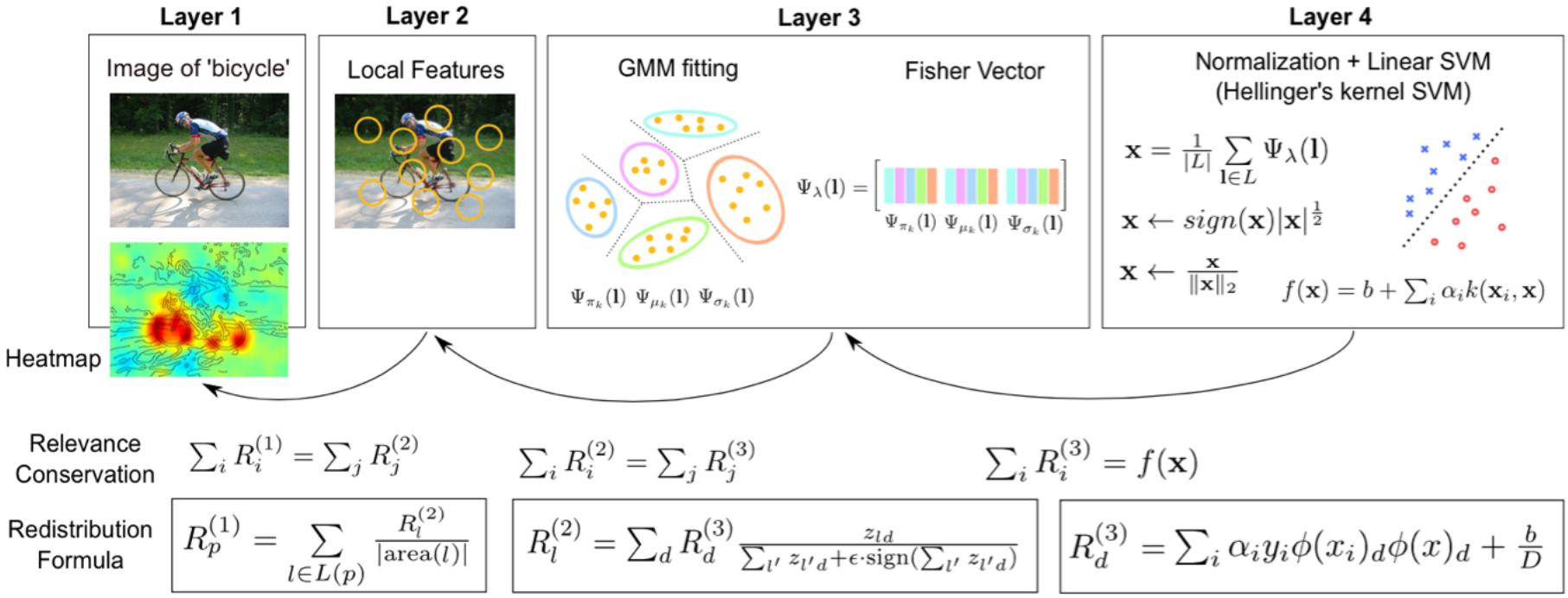


(Horst et al. 2018)

Application of LRP
Understand Model & Obtain new Insights

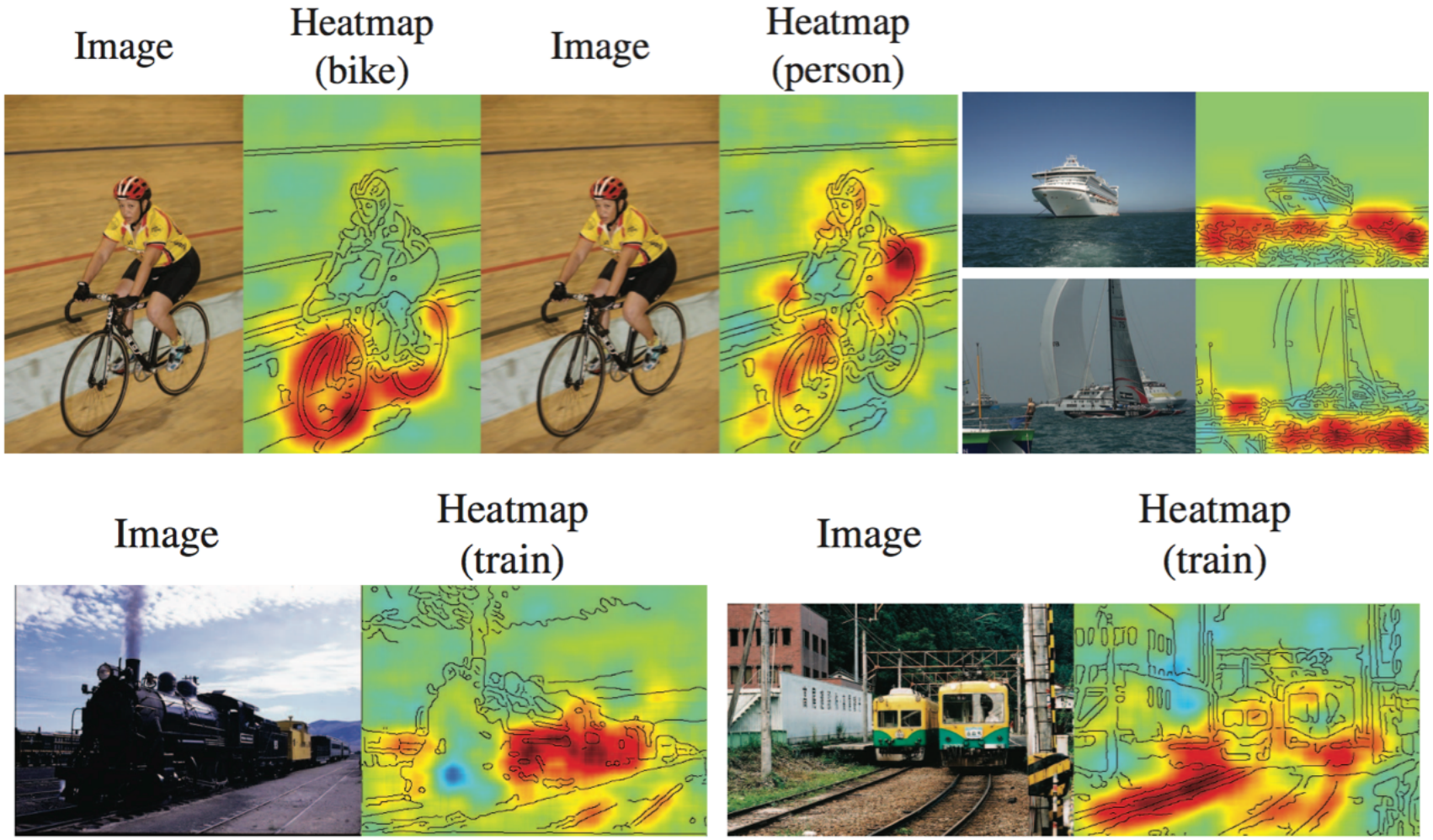
Application: Understand the model

- Fisher Vector / SVM classifier
- PASCAL VOC 2007



(Lapuschkin et al. 2016)

Application: Understand the model

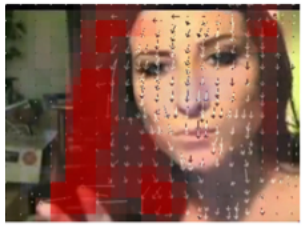


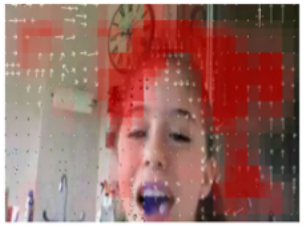
(Lapuschkin et al. 2016)

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 .

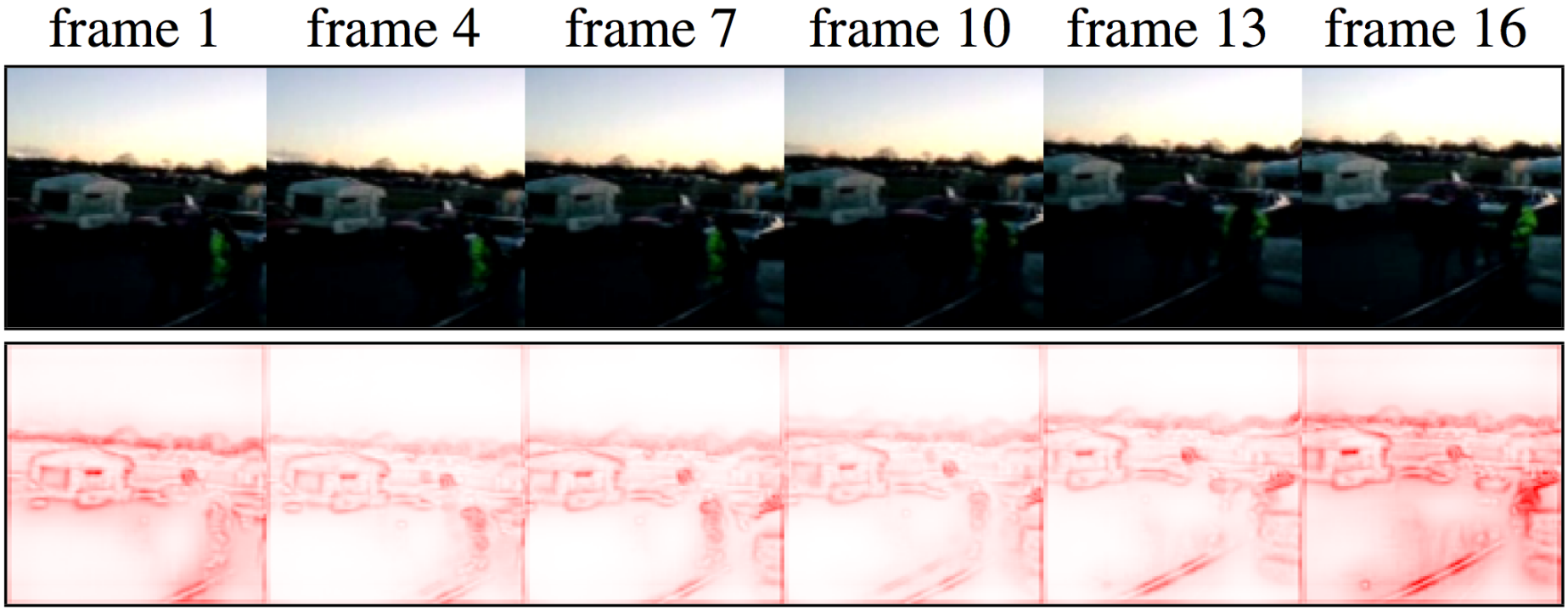
neither funny nor suspenseful nor particularly well-drawn .

Model understands negation !

(Arras et al., 2017 & 2018)

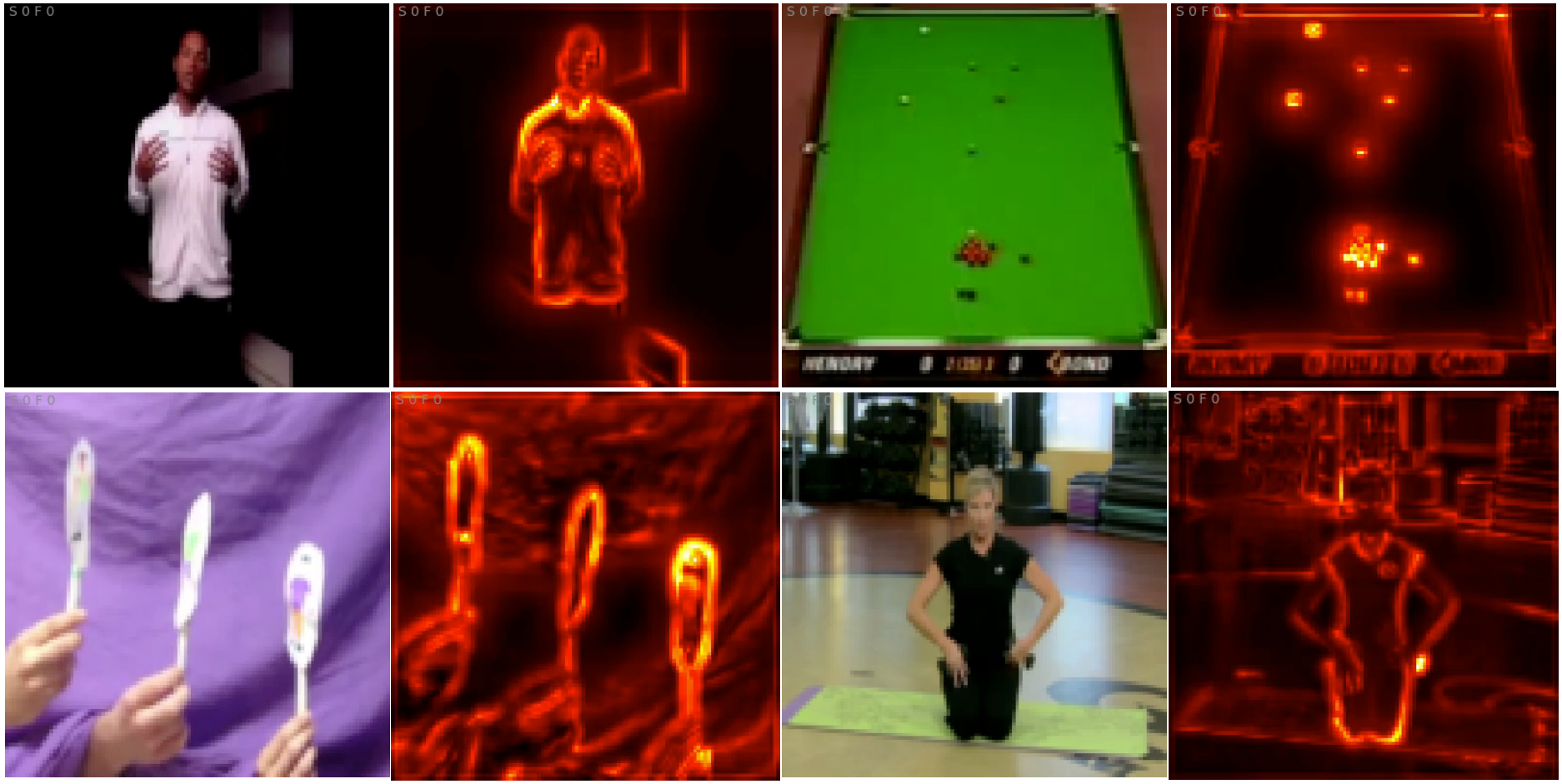
Application: Understand the model

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



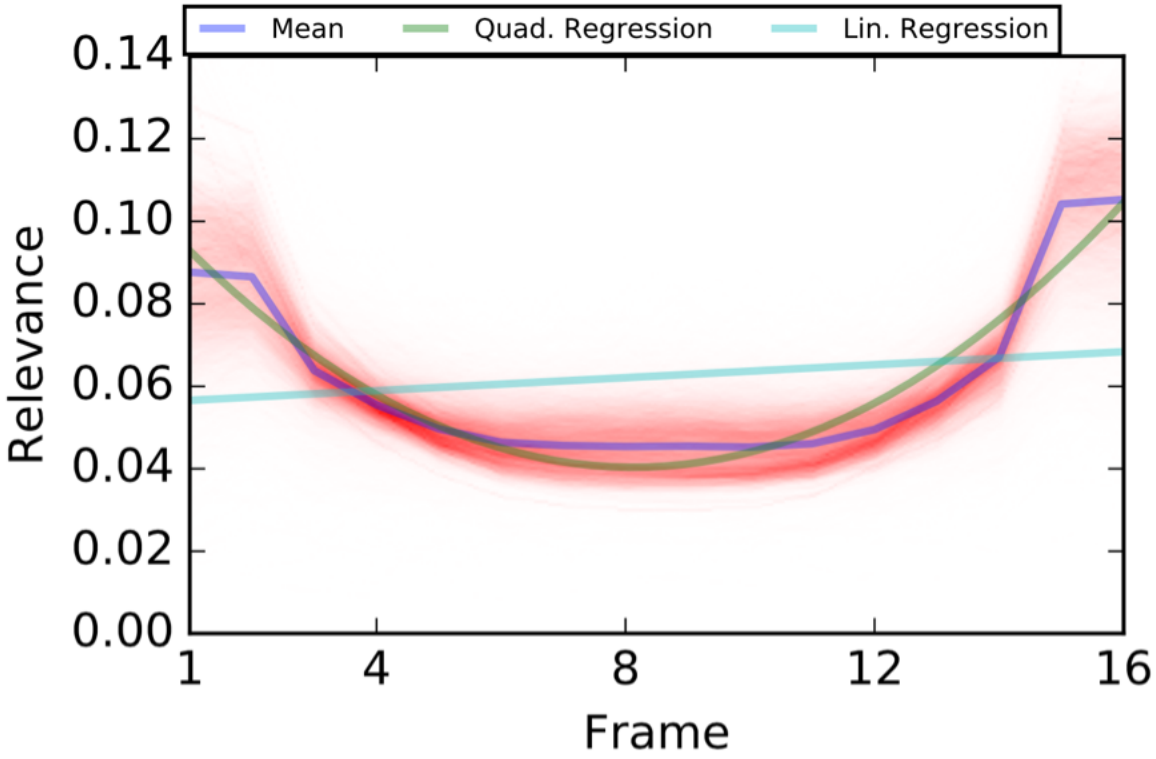
(Anders et al., 2018)

Application: Understand the model



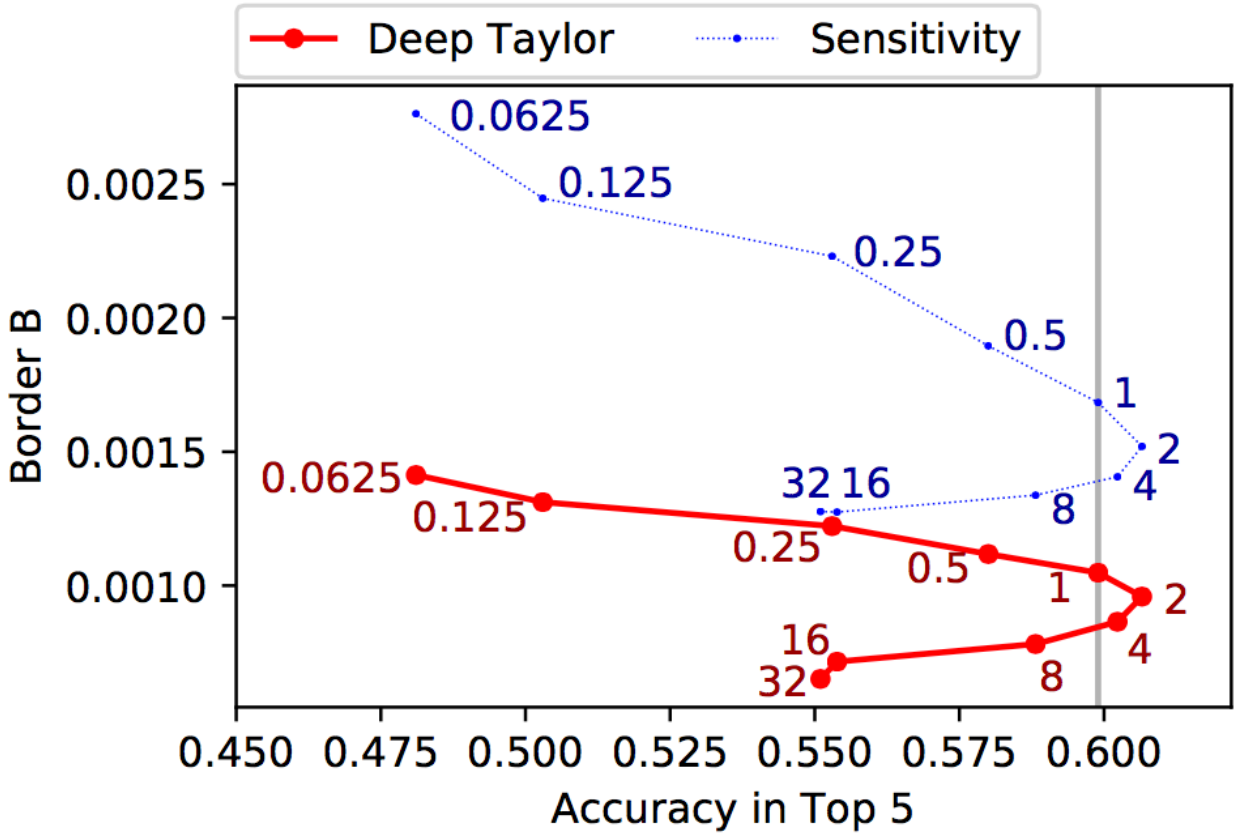
(Anders et al., 2018)

Application: Understand the model



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

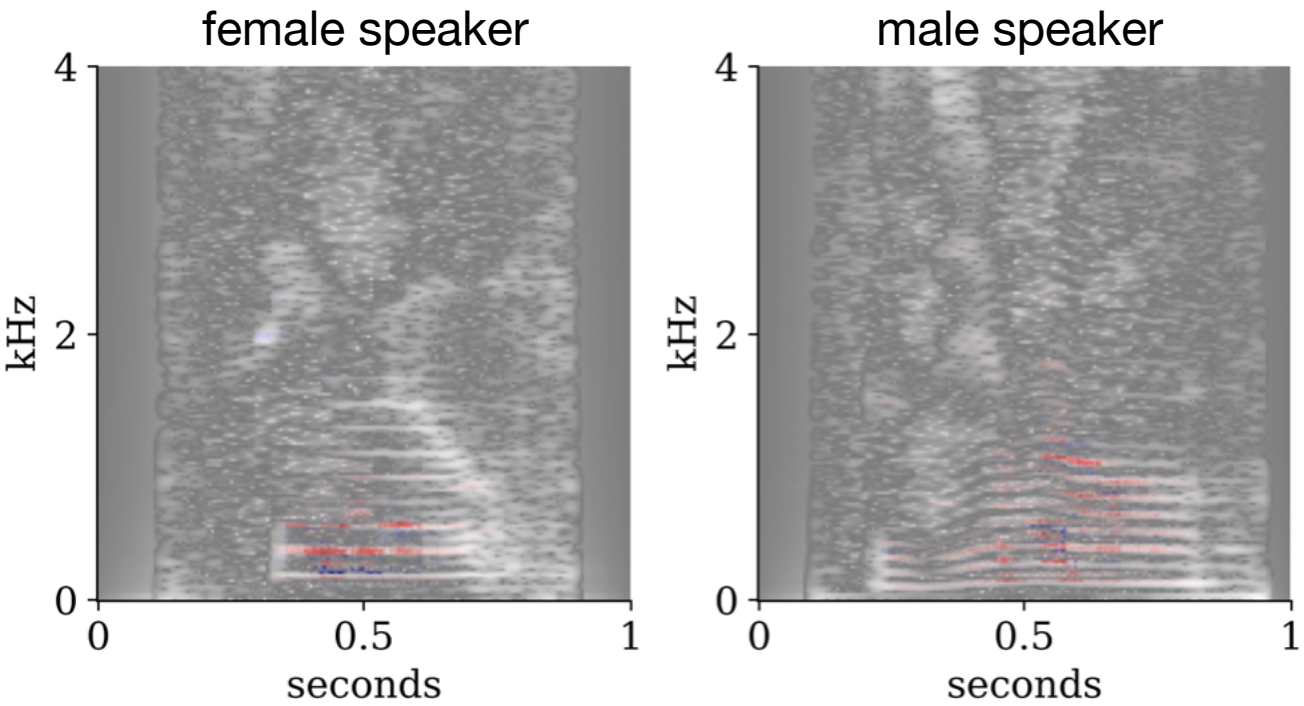
Application: Understand the model



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

Application: Understand the model

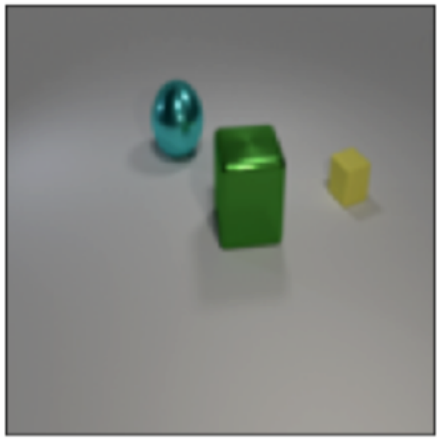
- AlexNet model
- trained on spectrograms
- spoken digits dataset (AudioMNIST)

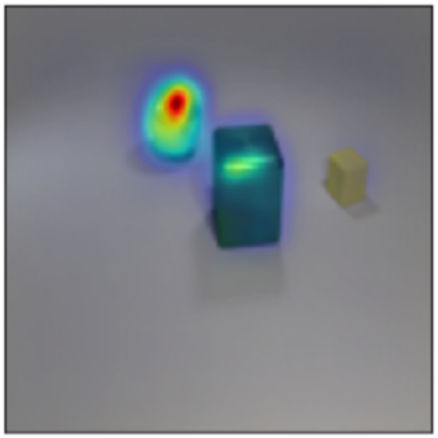


model classifies gender based on the fundamental frequency and its immediate harmonics (see also Traunmüller & Eriksson 1995) *(Becker et al., 2018)*

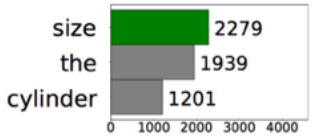
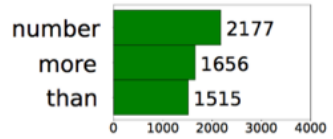
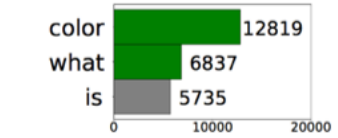
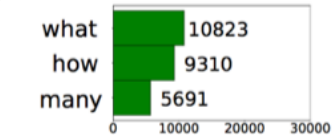
Application: Understand the model

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

Question Type	LRP
equal_size	
greater_than	
query_color	
count	

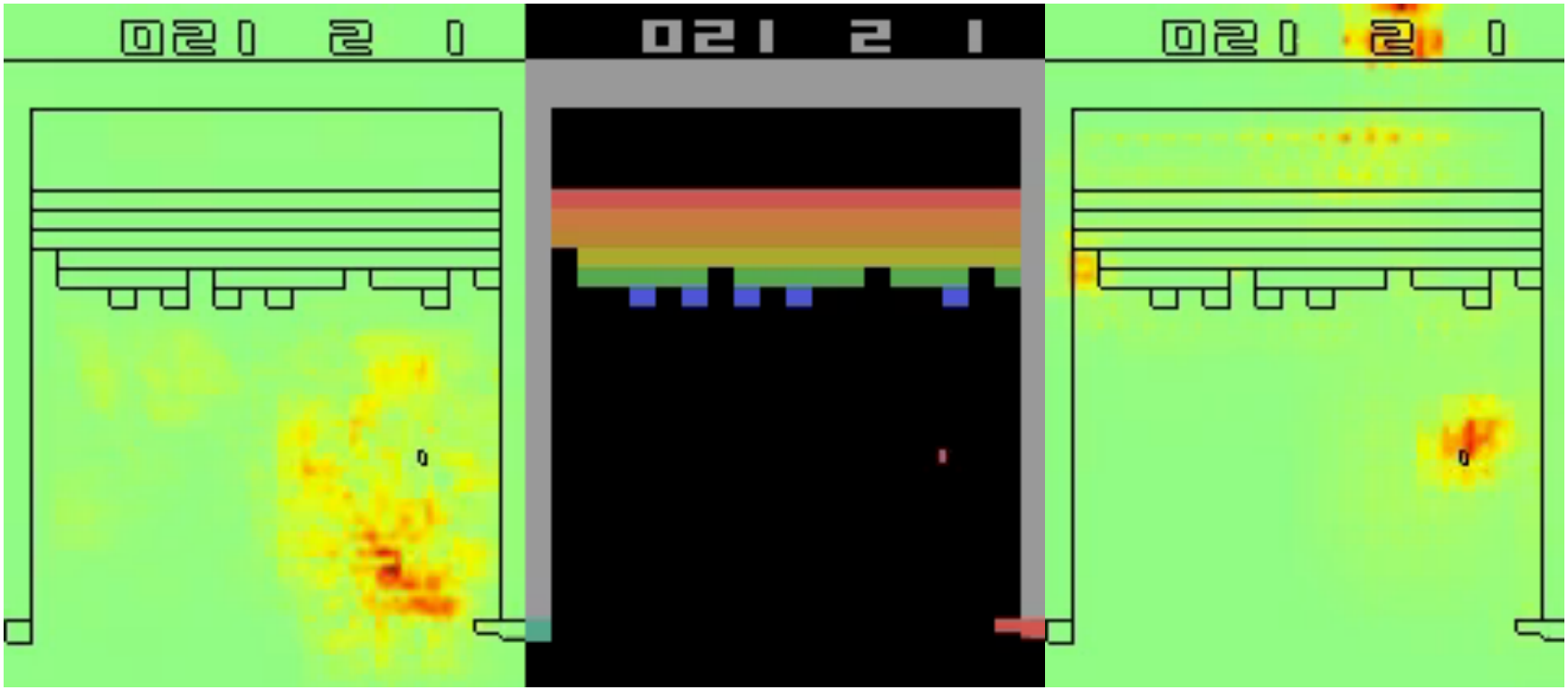
model understands the question and correctly identifies the object of interest

(Arras et al., 2018)

Application: Understand the model

Sensitivity Analysis

LRP

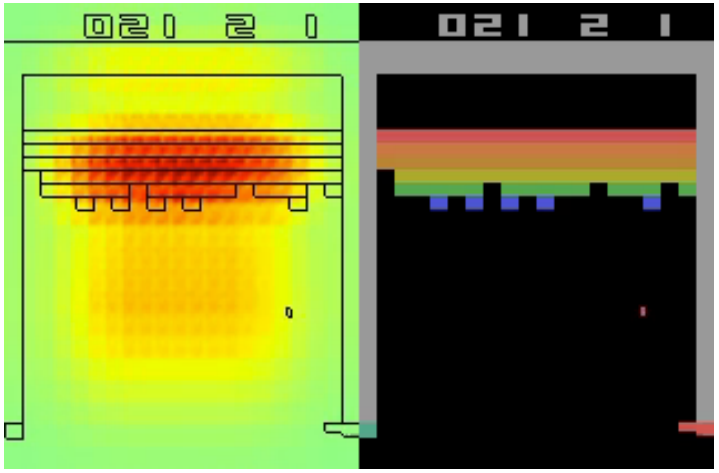


does not focus on where the ball is, but on where the ball could be in the next frame

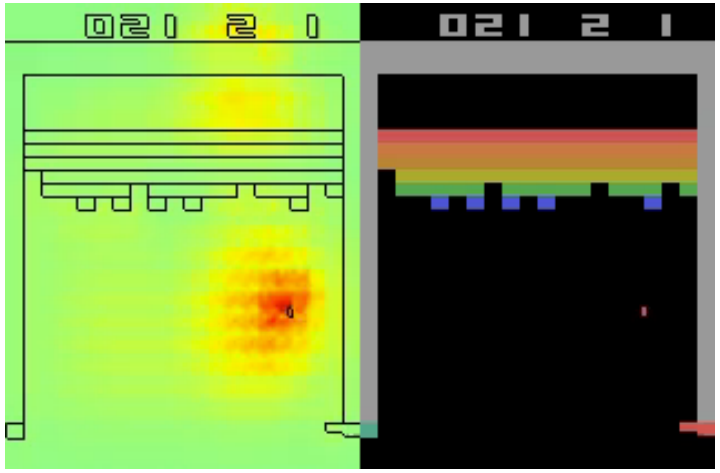
LRP shows that that model tracks the ball
(Lapuschkin et al., in prep.)

Application: Understand the model

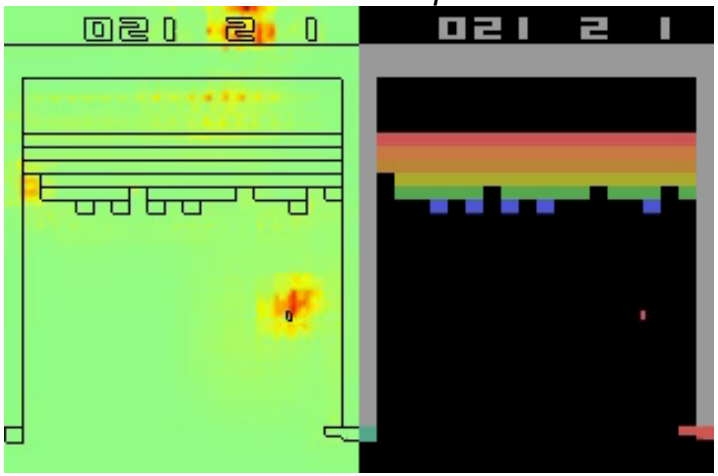
After 0 epochs



After 25 epochs

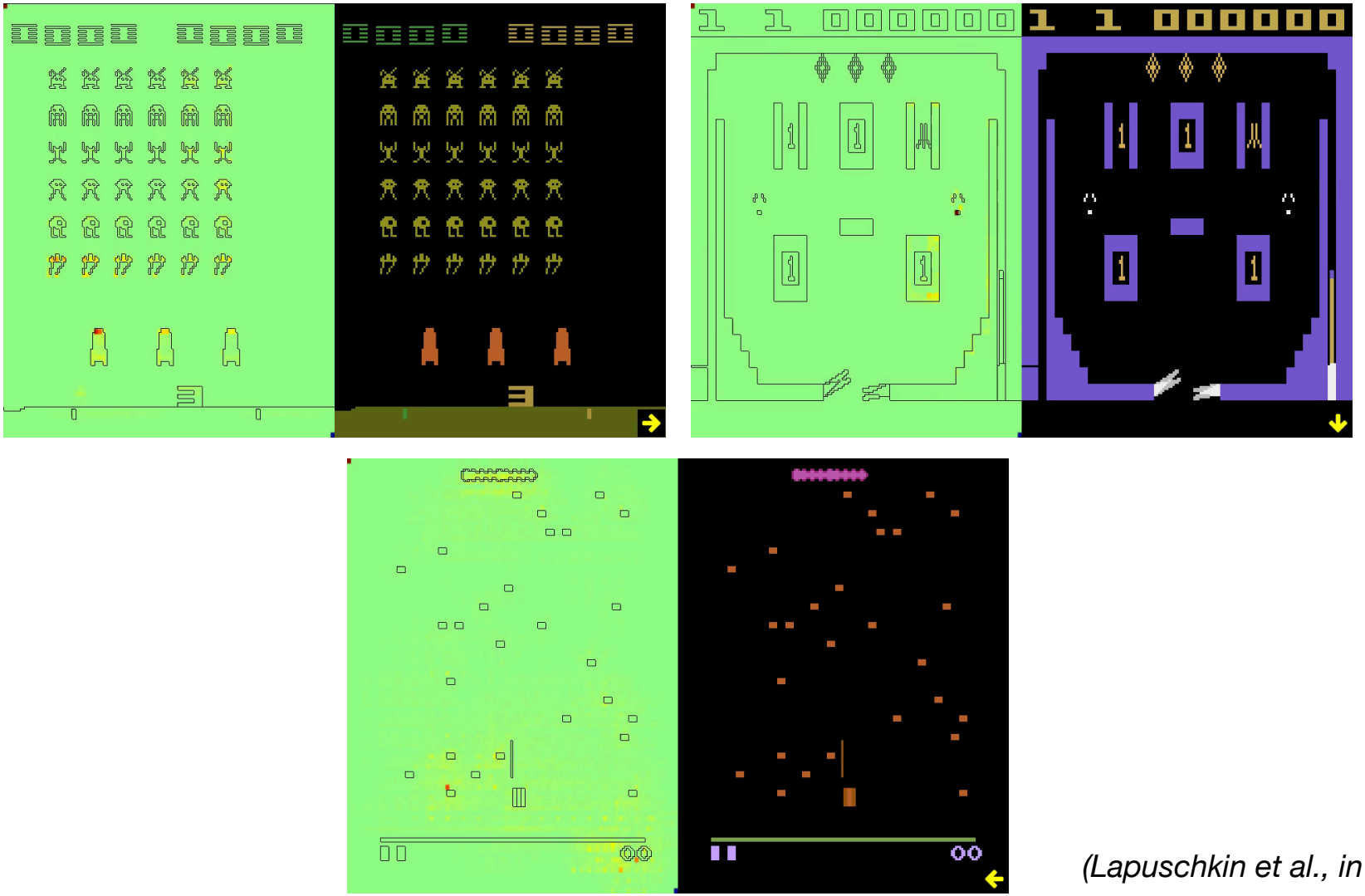


After 195 epochs



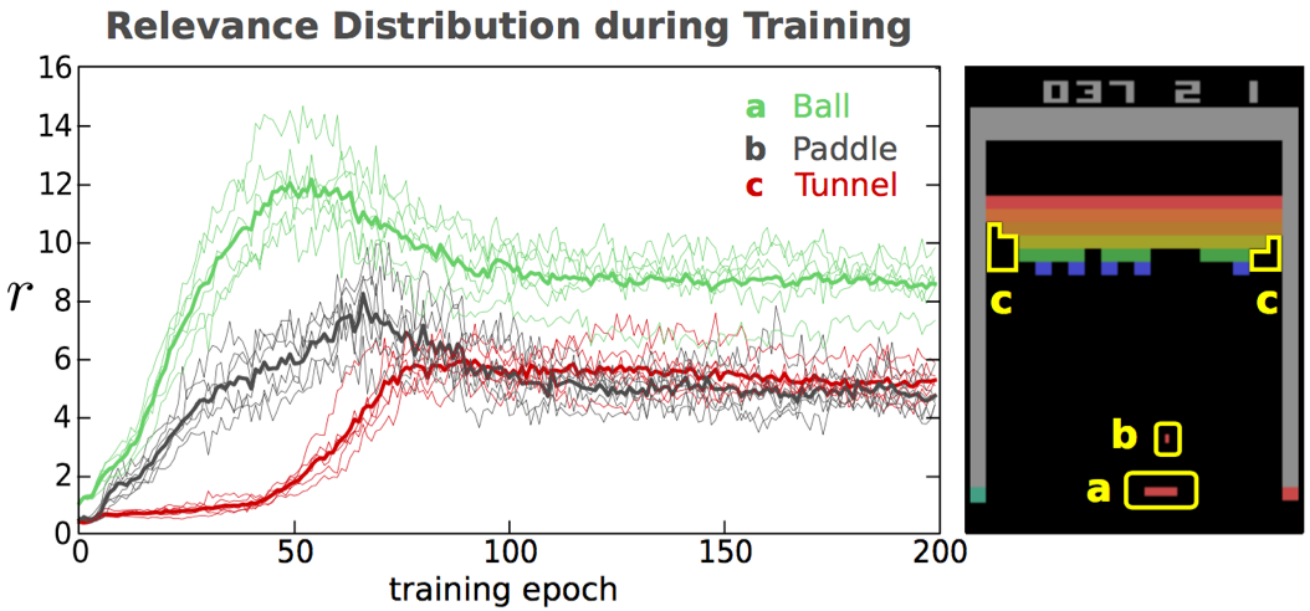
(Lapuschkin et al., in prep.)

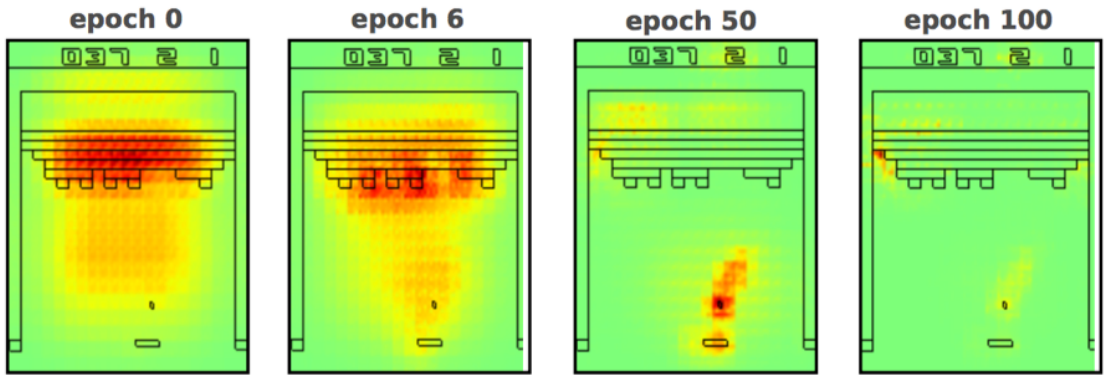
Application: Understand the model



(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.)

Tutorials

Interpretable Deep Learning: Towards Understanding & Explaining DNNs

Wrap-up

Wojciech Samek, Grégoire Montavon, Klaus-Robert Müller

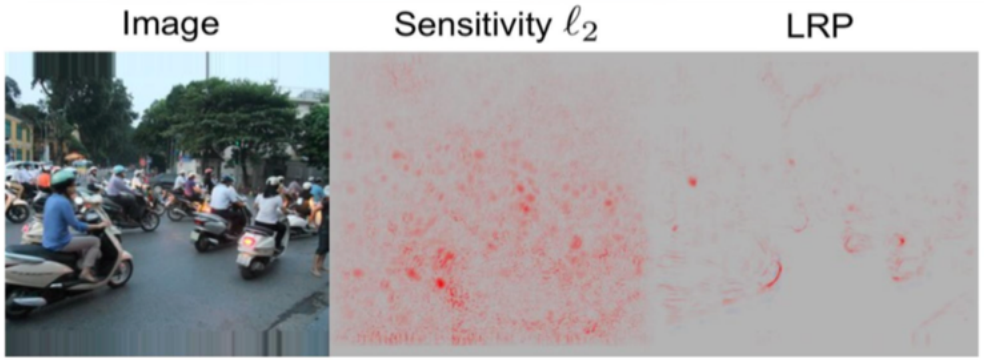
Technische Universität Berlin

Fraunhofer Heinrich Hertz Institute

Berliner Zentrum für MASCHINELLES LERNEN

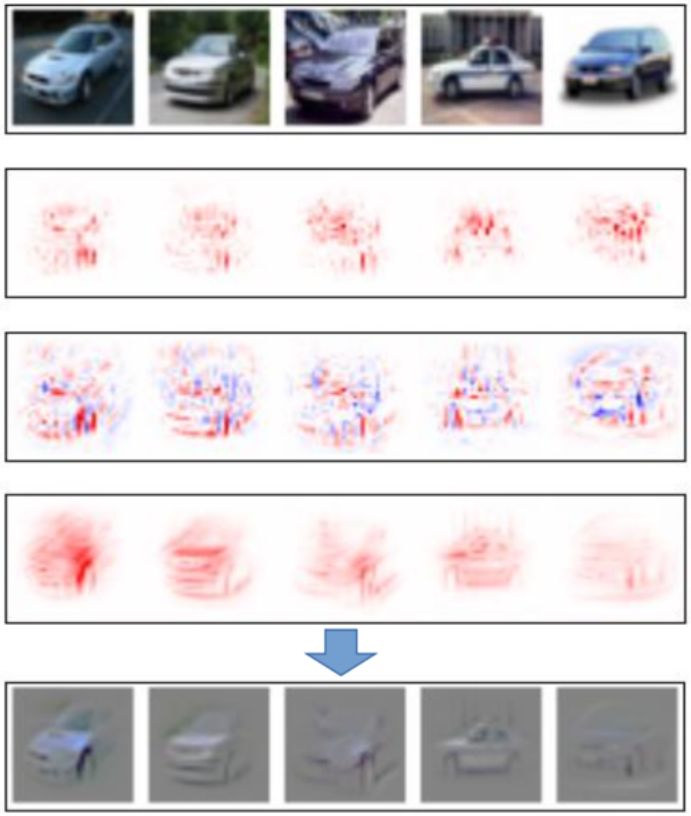
Take Home Messages

Sensitivity analysis is not the question that you would like to ask!



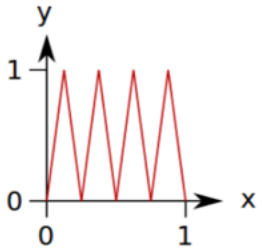
Take Home Messages

What works for simple models doesn't work for deep models.



gradient-based methods

vulnerable to shattered gradients



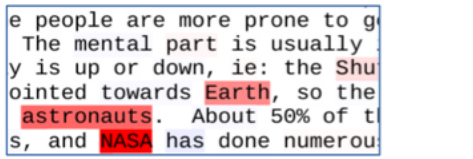
Our LRP method is robust to this.

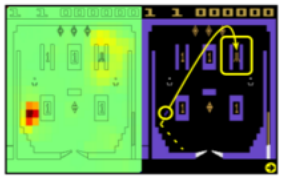
Take Home Messages

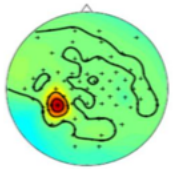
LRP works 4 all: deep models, LSTMs, kernel methods …

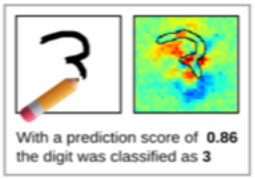
LRP Explanation Framework

(software, tutorials, demos, insights, applications)









Take Home Messages

$$\text{LRP} \neq \text{Gradient} \times \text{Input}$$

… except for special cases. LRP was developed among others because gradient-based methods aren't satisfying.

High flexibility: Different LRP variants, free parameters

Good news: No need to reimplement LRP, check our software at www.heatmapping.org.

Take Home Messages

train interpretable model

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

vs.

train best model → interpret it

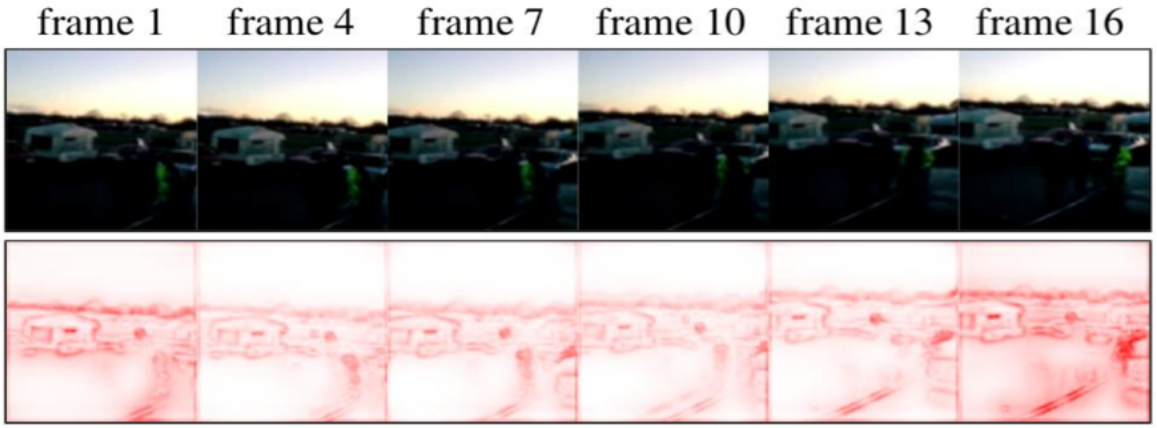
Take Home Messages

Explanations can be evaluated:
Pixel flipping (model agnostic)
And beyond LRP and DTD

[Samek et al. IEEE TNNLS 2017]

Take Home Messages

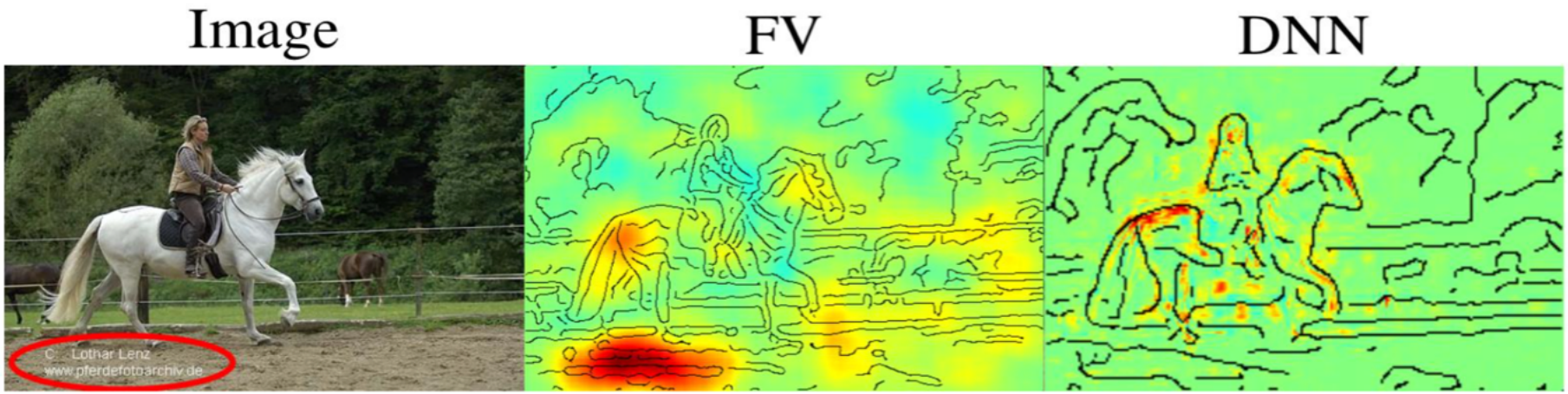
Explanation helps to improve models



Explaining ML, Now What?

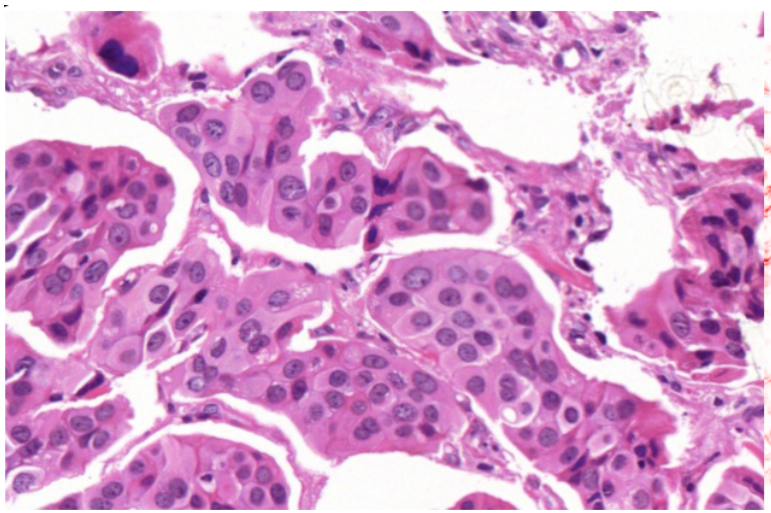
Take Home Messages

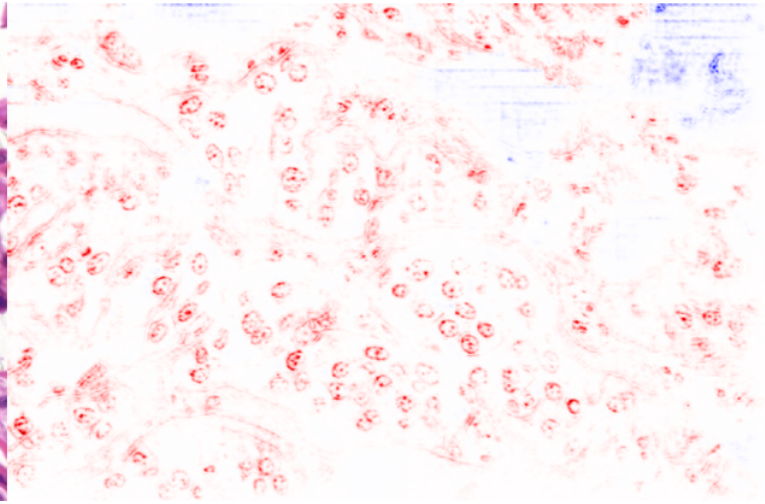
Explanation helps to find flaws in models

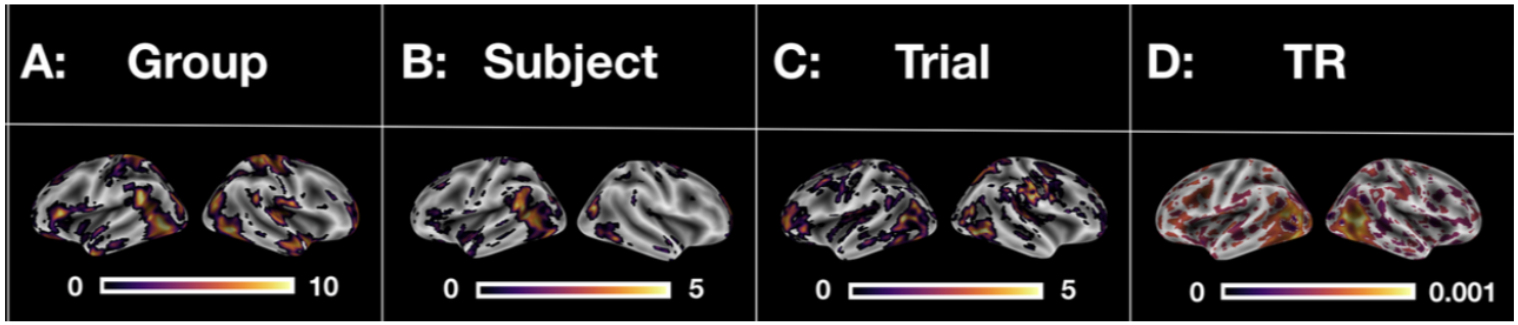


Take Home Messages

Getting new Insights in the Sciences





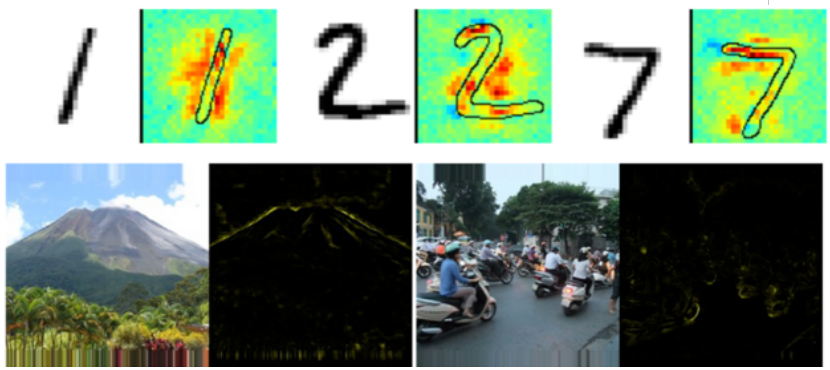


More information

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. Highly Cited Paper

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: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 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.

F Horst, S Lapuschkin, W Samek, KR Müller, WI Schöllhorn. What is Unique in Individual Gait Patterns? Understanding and Interpreting Deep Learning in Gait Analysis. *arXiv:1808.04308*, 2018

References

Software

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

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