Prof. Fernando J. Von Zuben
DCA/FEEC/Unicamp

Deep Learning – Parte 6
Interpretabilidade em Redes Neurais Profundas

Índice Geral

1 O compromisso entre acurácia e interpretabilidade em modelos de aprendizado 2

2 Distintas abordagens na literatura..... 12

3 Análise de sensibilidade e suas limitações 18

4 Explicando as predições de redes neurais profundas..... 22

5 Comparação entre diferentes técnicas..... 41

6 Aplicações..... 45

7 Spectral Relevance Analysis (SpRAy)..... 53

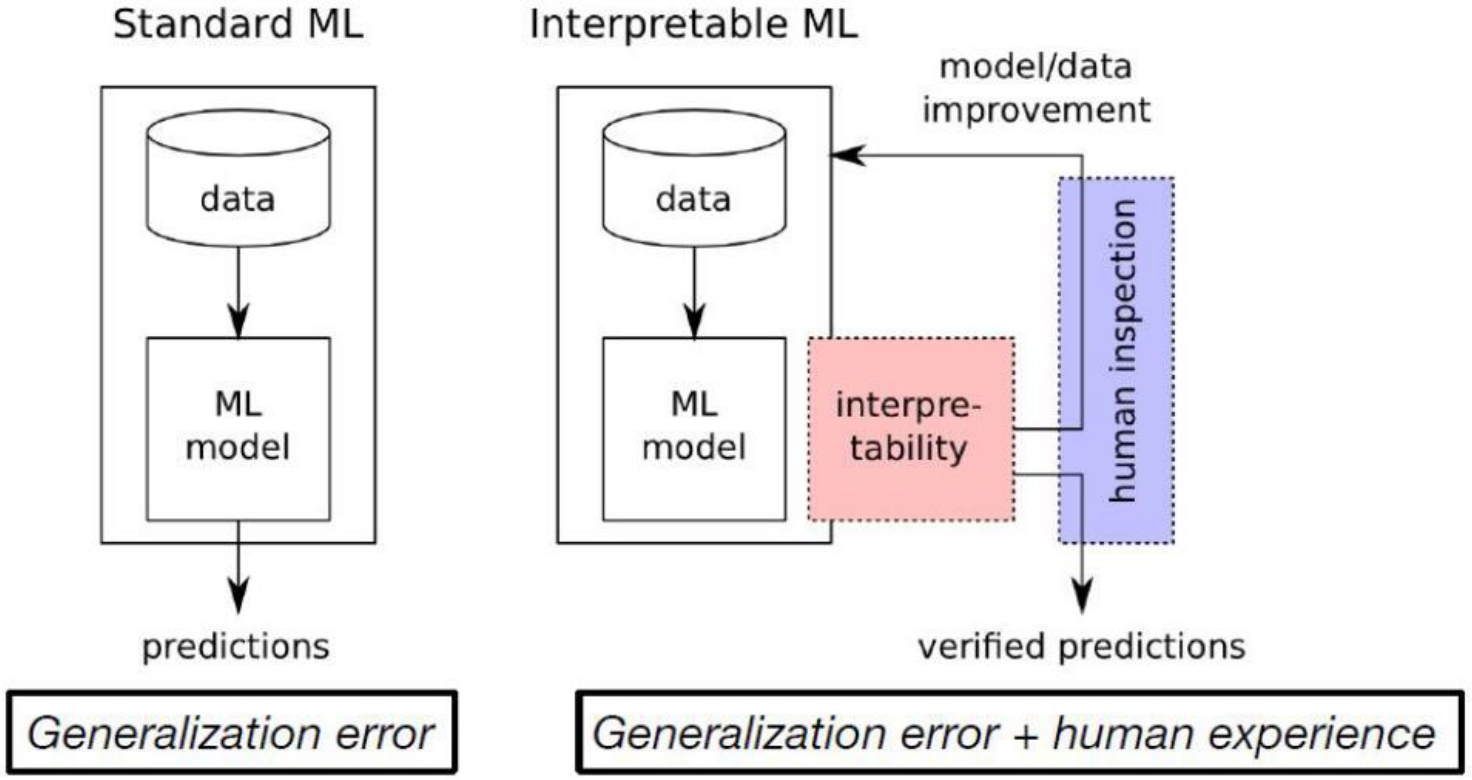
8 Referências bibliográficas..... 57

1 O compromisso entre acurácia e interpretabilidade em modelos de aprendizado

- Nota: O conteúdo em inglês deste material, a menos que citado de outra forma, foi extraído de:
  - ✓ Samek, W.; Montavon, G.; Müller, K.-R. "Tutorial on Interpreting and Explaining Deep Models in Computer Vision", CVPR'2018. [disponível em http://www.heatmapping.org/]
  - ✓ https://www.youtube.com/watch?v=gy_Cb4Do_YE
  - ✓ https://www.youtube.com/watch?v=xkN3WyNeuU0
- A busca por modelos de aprendizado interpretáveis está longe de ser recente, mas até pouco tempo atrás havia uma predominância de foco em acurácia e se defendia a existência de um compromisso entre acurácia e interpretabilidade, de modo que a busca por mais acurácia tenderia a gerar modelos menos interpretáveis, da mesma forma que modelos mais interpretáveis não conseguiriam atingir os mesmos níveis de acurácia que modelos "caixa-preta".

- As últimas conquistas voltadas para interpretabilidade em aprendizado de máquina, no entanto, demonstram que acurácia e interpretabilidade podem caminhar juntas (SAMEK et al., 2019).
- De fato, os últimos anos têm verificado a proposição de novos métodos voltados ao entendimento de como modelos de aprendizado de máquina de última geração realizam suas predições.
- A existência de uma cascata de filtros não-lineares em *deep learning* torna técnicas convencionais de extração de informação (visando interpretabilidade) pouco efetivas, o que cria uma demanda por novas soluções.
- Os modelos de aprendizado de máquina, em particular redes neurais profundas, são caracterizados por um poder preditivo muito elevado, mas, em muitos casos, não são facilmente interpretáveis por um ser humano.

- A interpretação de um classificador não-linear, por exemplo, é importante para ganhar confiança na predição realizada (o classificador está operando de acordo com o esperado?) e para identificar possíveis vieses ou artefatos que guiaram o processo, além de atender a outros requisitos específicos em certas aplicações.



Interpreting Deep Neural Networks

- Emerging research area (dedicated workshops at ICML'16 and NIPS'16)
- Interpretability techniques work well (e.g. deconvnet [Zeiler'14], guided backprop [Springenberg'14], LRP [Bach'15]) but are still mainly heuristic.

Why interpretability?

5) Compliance to legislation

European Union's new General Data Protection Regulation → "right to explanation"

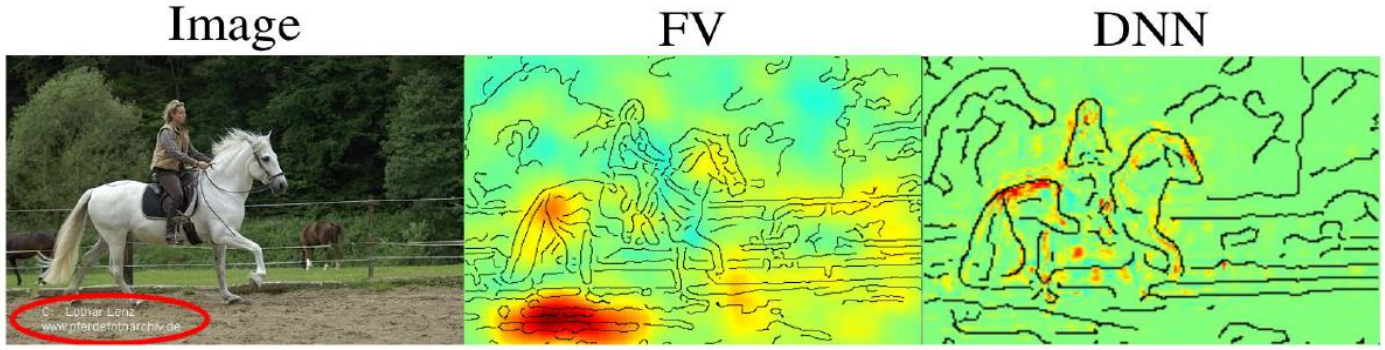
Retain human decision in order to assign responsibility.

"With interpretability we can ensure that ML models work in compliance to proposed legislation."

- Vão ser contrastadas aqui duas abordagens:
  1. Análise de sensibilidade: O que mais aumenta ou diminui a probabilidade de uma imagem ser classificada numa certa classe?
  2. *Layer-wise Relevance Propagation* (LRP): O que é relevante (positiva ou negativamente) para que uma imagem seja classificada numa certa classe?

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%



- Lapuschkin, S.; Binder, A.; Montavon, G.; Müller, K.-R.; Samek, W. (2016). Analyzing Classifiers: Fisher Vectors and Deep Neural Networks. 2912-2920. 10.1109/CVPR.2016.318.
- *Fisher vector enconding*: descreve o quanto a distribuição de atributos de uma imagem difere da distribuição de todas as imagens consideradas.
- Não basta focar apenas na capacidade de generalização.

- www.fatml.org

Bringing together a growing community of researchers and practitioners concerned with fairness, accountability, and transparency in machine learning

The past few years have seen growing recognition that machine learning raises novel challenges for ensuring non-discrimination, due process, and understandability in decision-making. In particular, policymakers, regulators, and advocates have expressed fears about the potentially discriminatory impact of machine learning, with many calling for further technical research into the dangers of inadvertently encoding bias into automated decisions.

At the same time, there is increasing alarm that the complexity of machine learning may reduce the justification for consequential decisions to “the algorithm made me do it.”

The annual event provides researchers with a venue to explore how to characterize and address these issues with computationally rigorous methods.

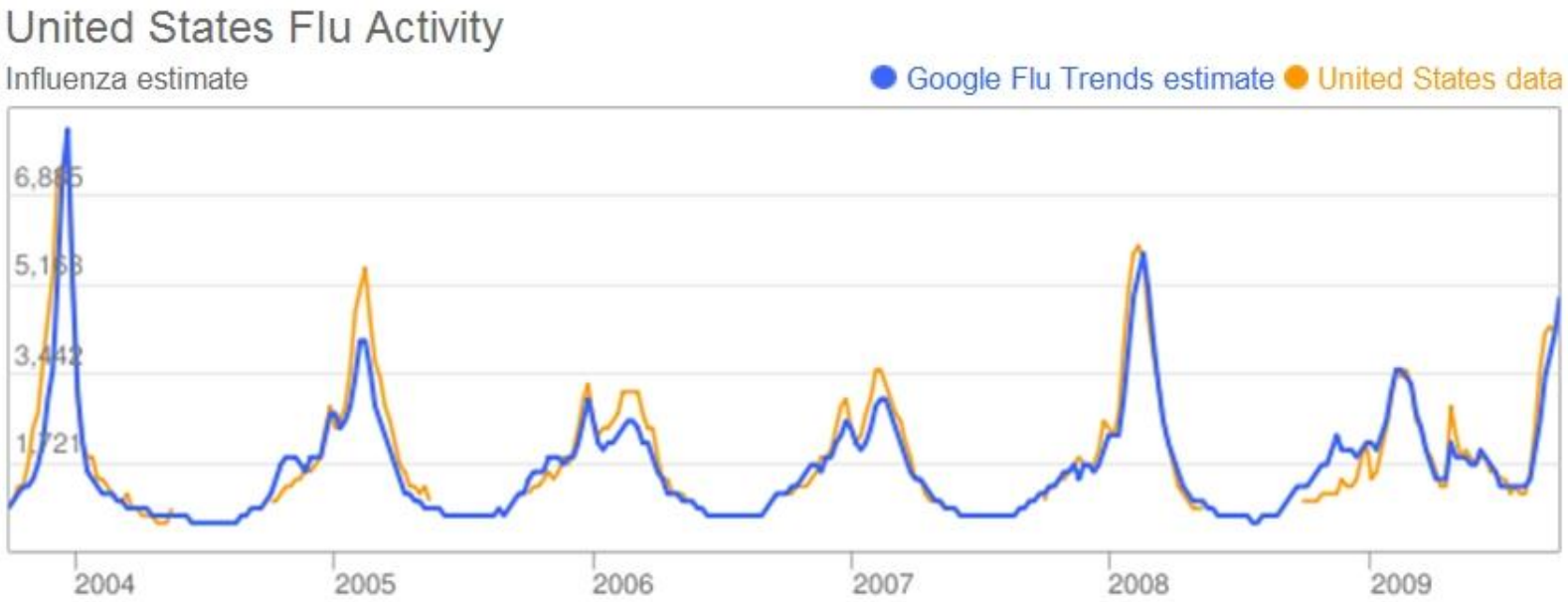
FATE: Fairness, Accountability, Transparency, and Ethics in AI

We study the complex social implications of AI, machine learning, data science, large-scale experimentation, and increasing automation. Our aim is to facilitate computational techniques that are both innovative and ethical, while drawing on the deeper context surrounding these issues from sociology, history, and science and technology studies.

The types of questions we are exploring are:

- How can AI assist users and offer enhanced insights, while avoiding exposing them to discrimination in health, housing, law enforcement, and employment?
- How can we balance the need for efficiency and exploration with fairness and sensitivity to users?
- As our world moves toward relying on intelligent agents, how can we create a system that individuals and communities can trust?

https://www.microsoft.com/en-us/research/theme/fate/



- *Certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate current flu activity around the world in near real-time.*
- *Detecting influenza epidemics using search engine query data.*
- **Só que (ainda) não …**
- *It is still early days for nowcasting and similar tools for understanding the spread of diseases like flu and dengue*.

https://www.google.org/flutrends/about/

2318 IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 29, NO. 10, OCTOBER 2017

Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data

Anuj Karpatne, Gowtham Atluri, James H. Faghmous, Michael Steinbach, Arindam Banerjee, Auroop Ganguly, Shashi Shekhar, Nagiza Samatova, and Vipin Kumar

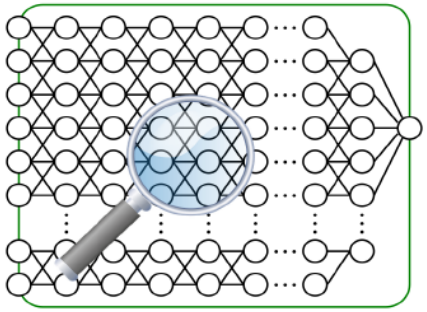
Abstract—Data science models, although successful in a number of commercial domains, have had limited applicability in scientific problems involving complex physical phenomena. Theory-guided data science (TGDS) is an emerging paradigm that aims to leverage the wealth of scientific knowledge for improving the effectiveness of data science models in enabling scientific discovery. The overarching vision of TGDS is to introduce scientific consistency as an essential component for learning generalizable models. Further, by producing scientifically interpretable models, TGDS aims to advance our scientific understanding by discovering novel domain insights. Indeed, the paradigm of TGDS has started to gain prominence in a number of scientific disciplines such as turbulence modeling, material discovery, quantum chemistry, bio-medical science, bio-marker discovery, climate science, and hydrology. In this paper, we formally conceptualize the paradigm of TGDS and present a taxonomy of research themes in TGDS. We describe several approaches for integrating domain knowledge in different research themes using illustrative examples from different disciplines. We also highlight some of the promising avenues of novel research for realizing the full potential of theory-guided data science.

Index Terms—Data science, knowledge discovery, domain knowledge, scientific theory, physical consistency, interpretability

2 Distintas abordagens na literatura

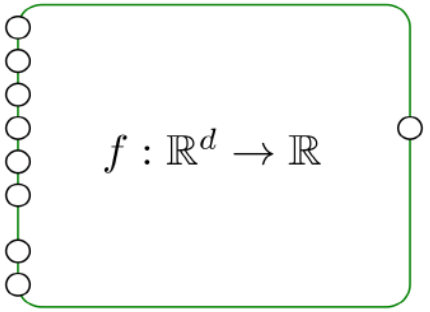
Understanding Deep Nets: Two Views

mechanistic understanding



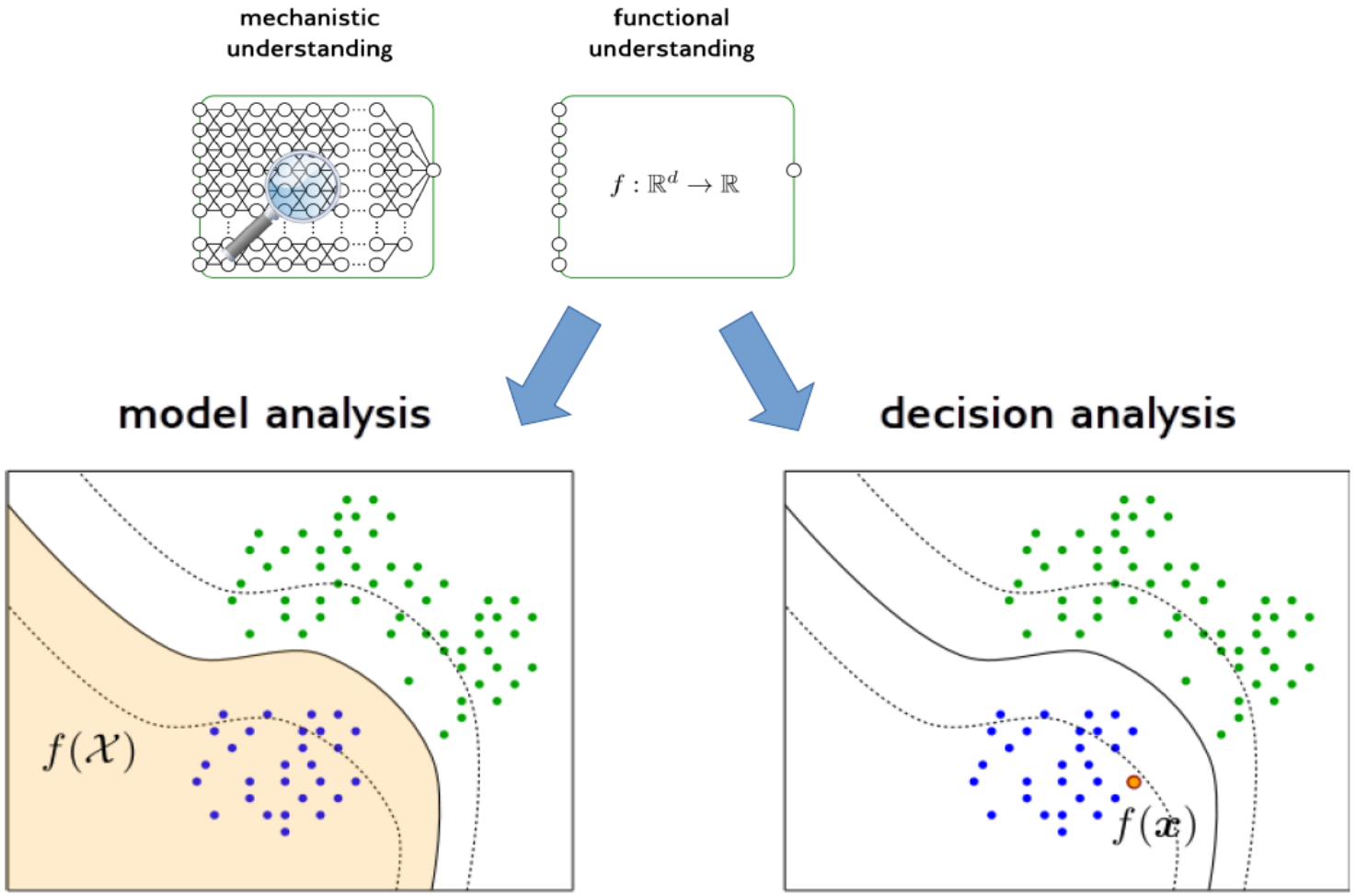
Understanding what mechanism the network uses to solve a problem or implement a function.

functional understanding



Understanding how the network relates the input to the output variables.

- A abordagem mecanicista foca nos processos envolvidos e nas relações causais entre as partes envolvidas.


mechanistic understanding
functional understanding
$f : \mathbb{R}^d \rightarrow \mathbb{R}$
model analysis
decision analysis
$f(\mathcal{X})$
$f(\boldsymbol{x})$

Approach 1: Class Prototypes

"How does a goose typically look like according to the neural network?"

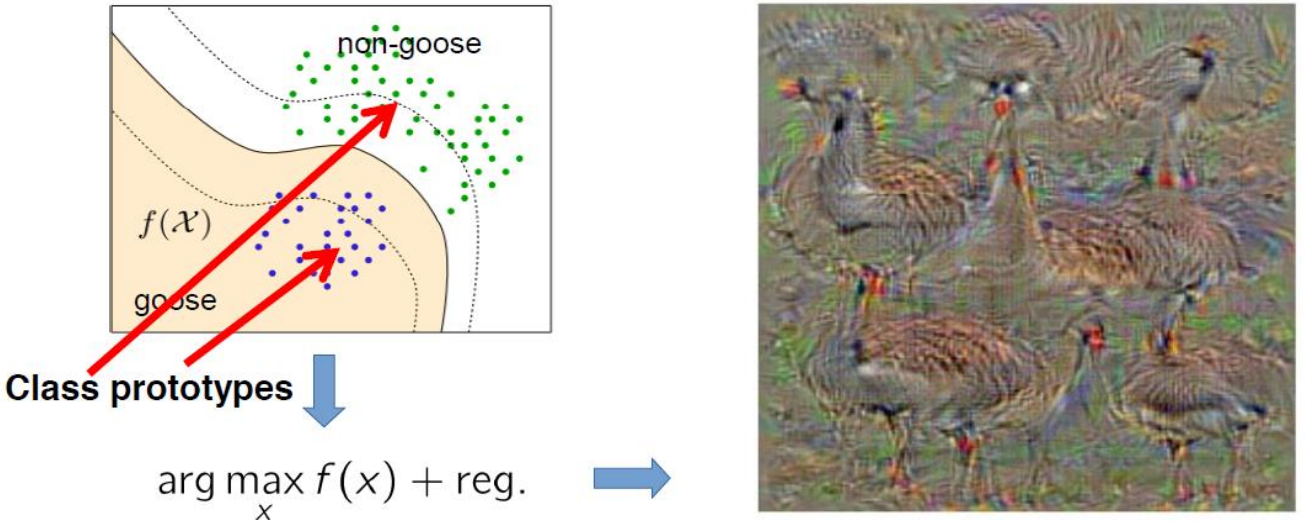
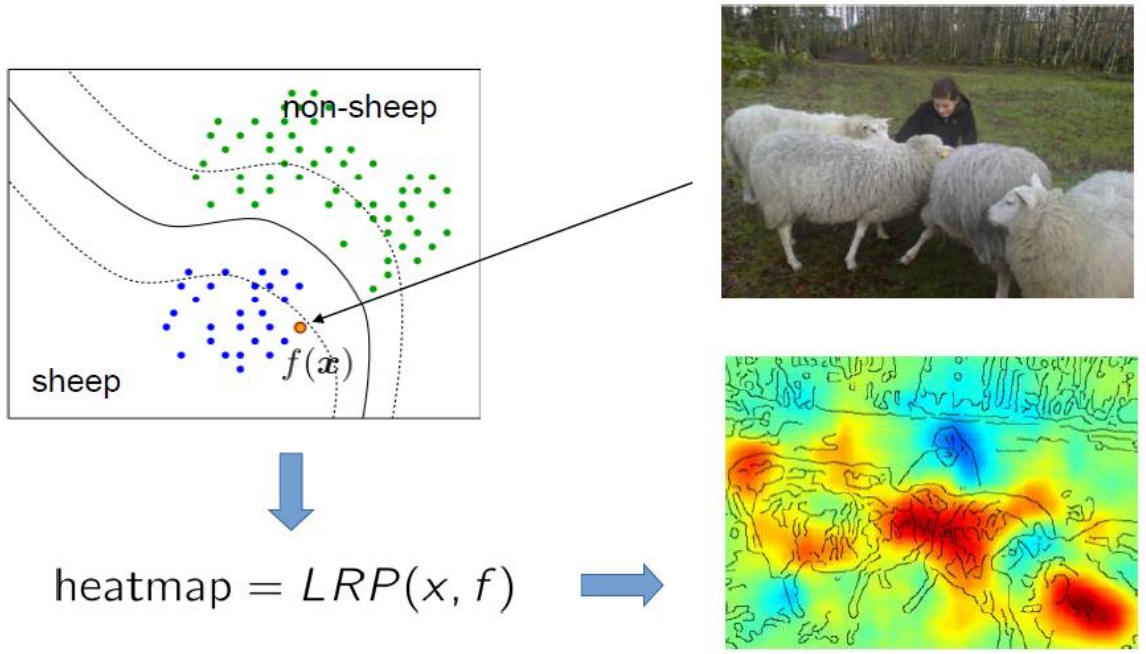


Image from **Symonian'13**

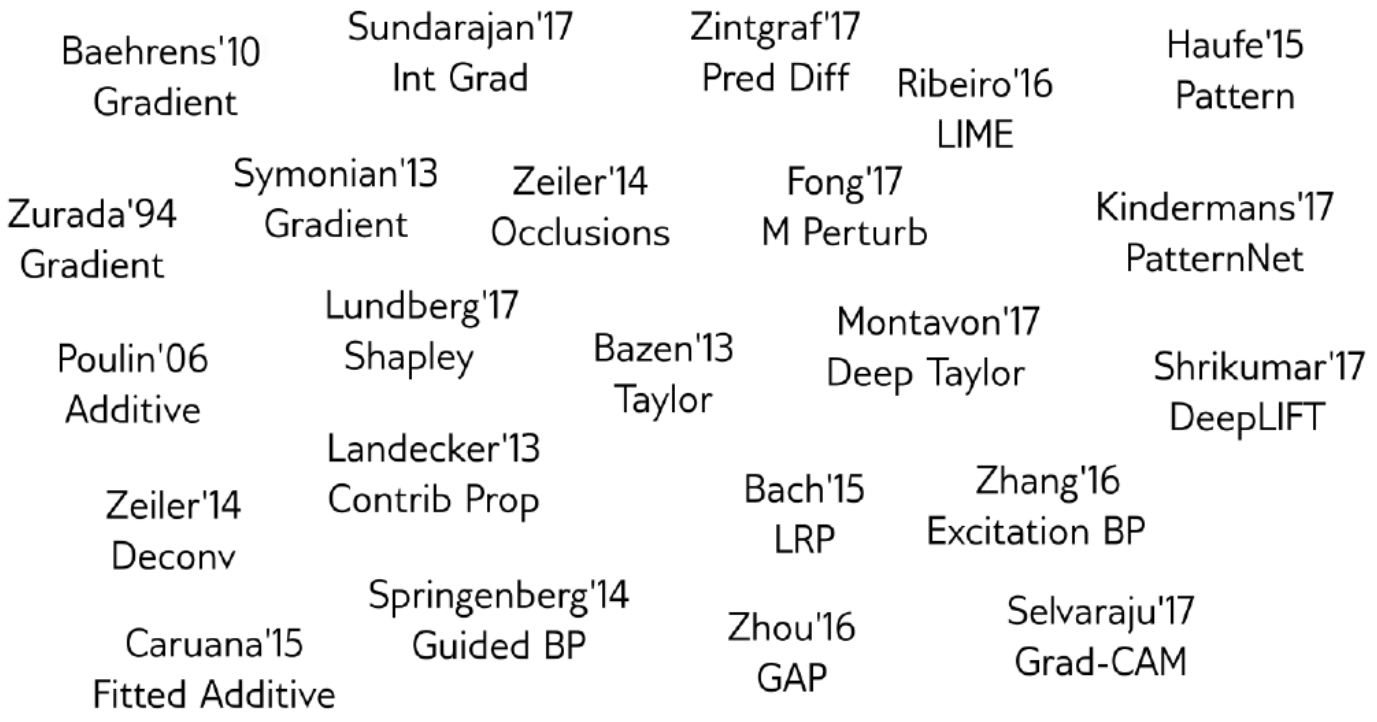
Approach 2: Individual Explanations

"Why is a given image classified as a sheep?"



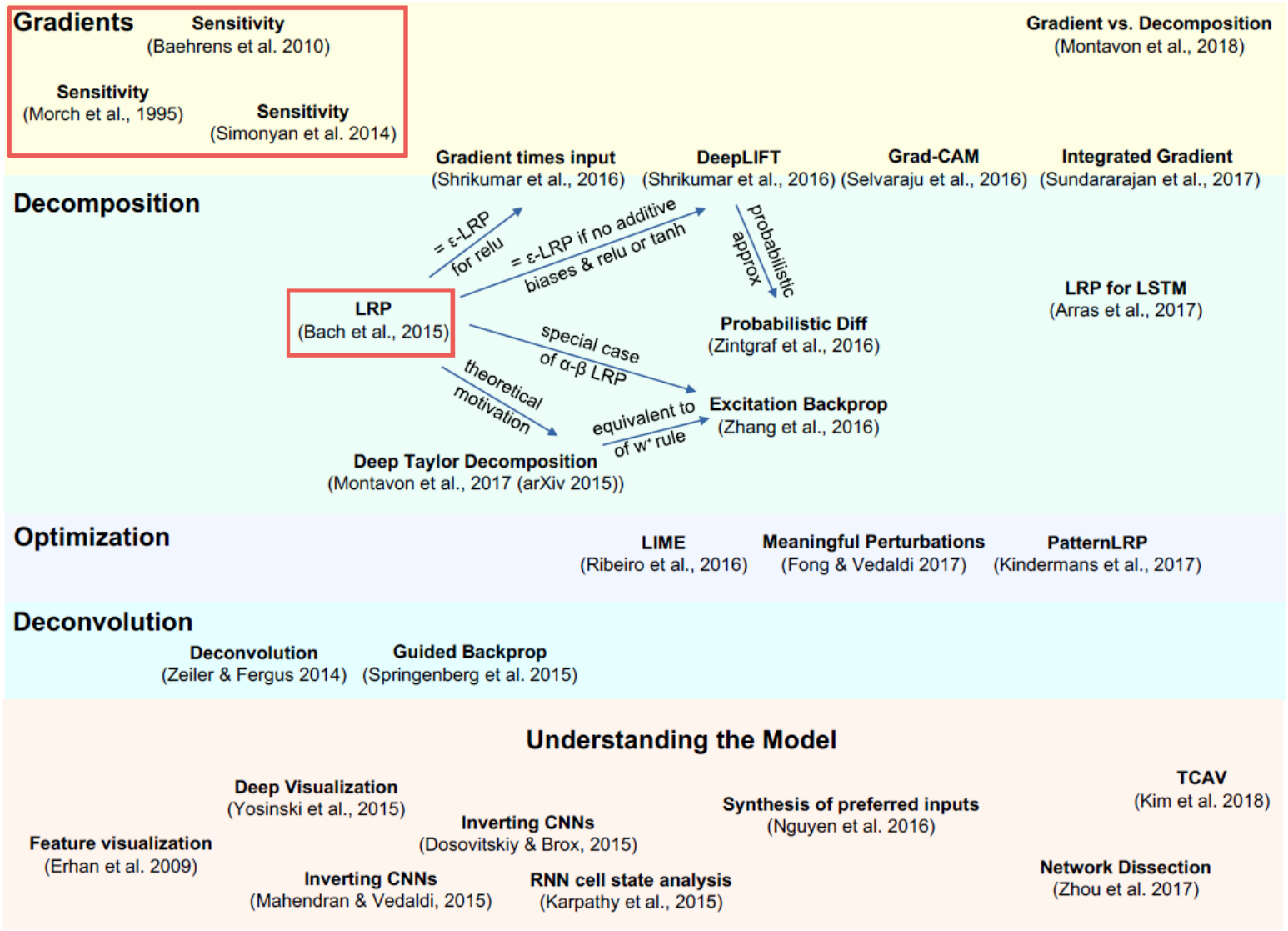
Images from **Lapuschkin'16**

Overview of Explanation Methods

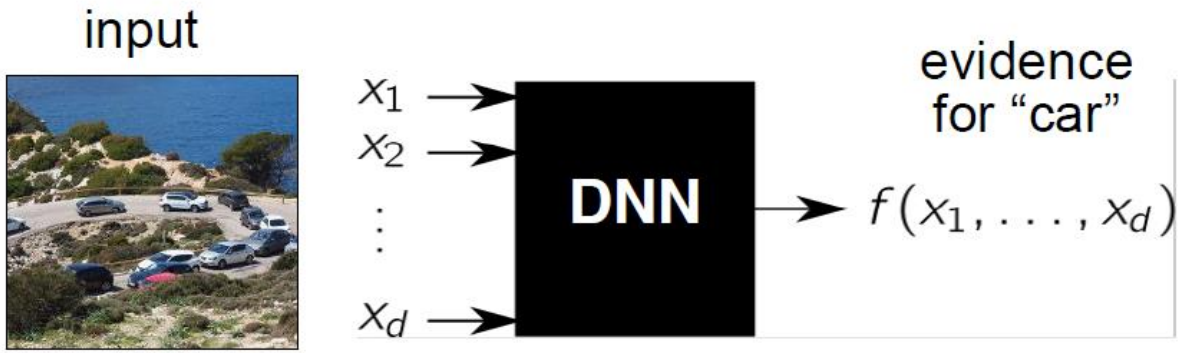


Question: Which one to choose ?

Historical remarks on Explaining Predictors



3 Análise de sensibilidade e suas limitações



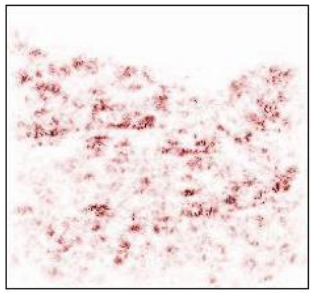
Sensitivity analysis: The relevance of input feature *i* is given by the squared partial derivative:

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

Understanding Sensitivity Analysis

Sensitivity analysis:

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



Problem: sensitivity analysis does not highlight cars

Observation:

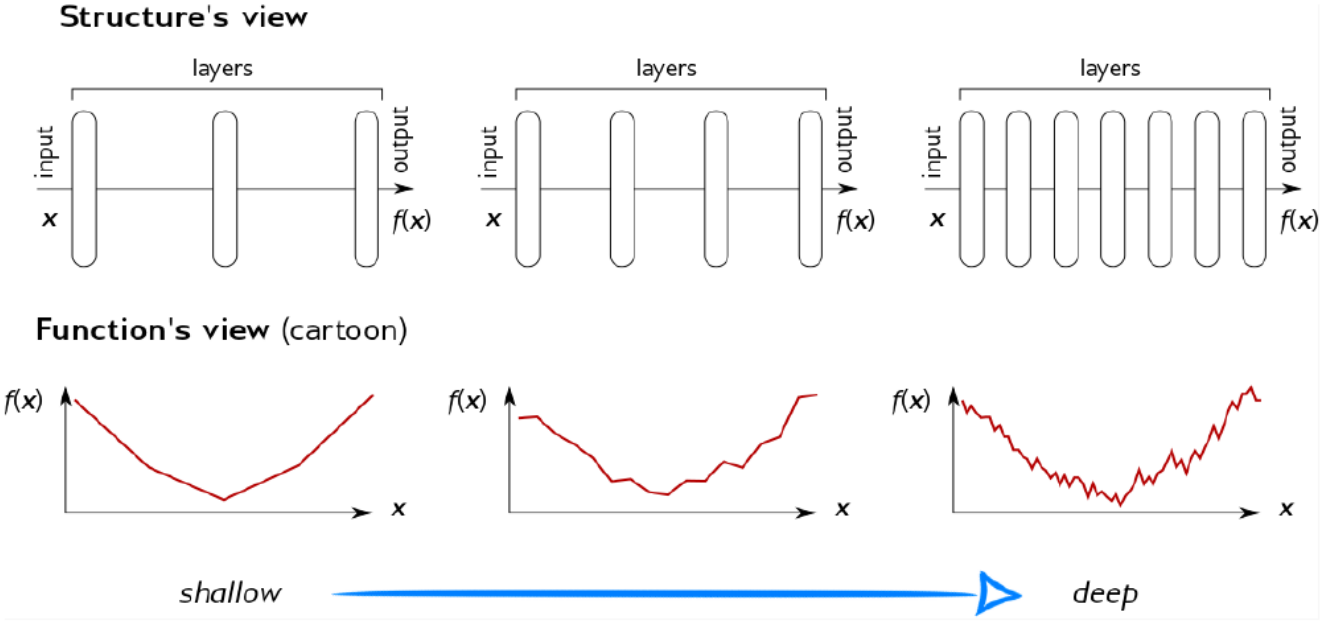
$$\sum_{i=1}^{d} \left(\frac{\partial f}{\partial x_i}\right)^2 = \|\nabla_x f\|^2$$

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

Sensitivity Analysis Problem: Shattered Gradients

[Montufar'14, Balduzzi'17]

Input gradient (on which sensitivity analysis is based), becomes increasingly highly varying and unreliable with neural network depth.

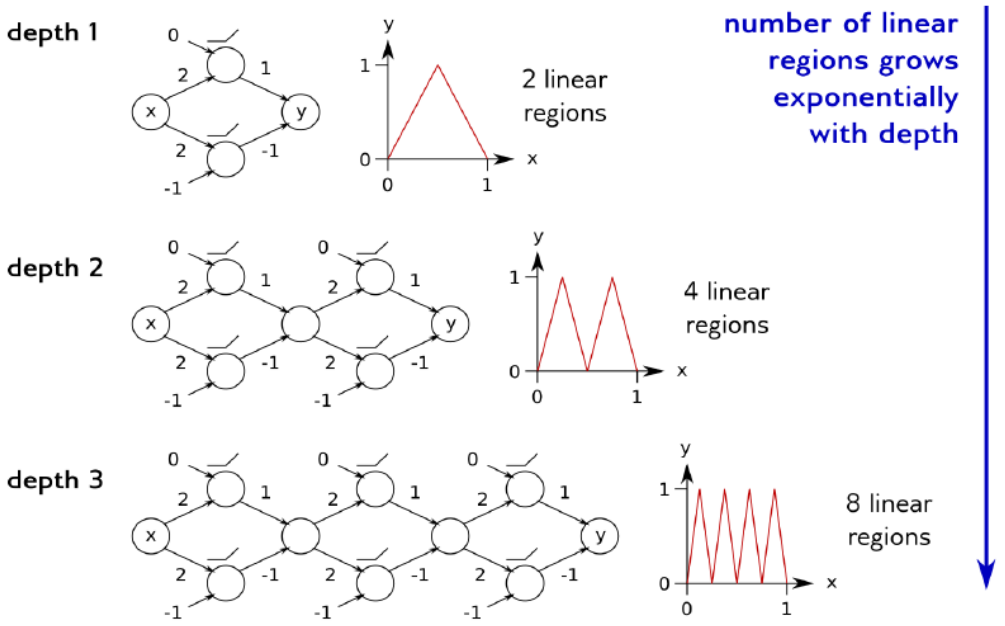


Shattered Gradients II

[Montufar'14, Balduzzi'17]

Input gradient (on which sensitivity analysis is based), becomes increasingly highly varying and unreliable with neural network depth.

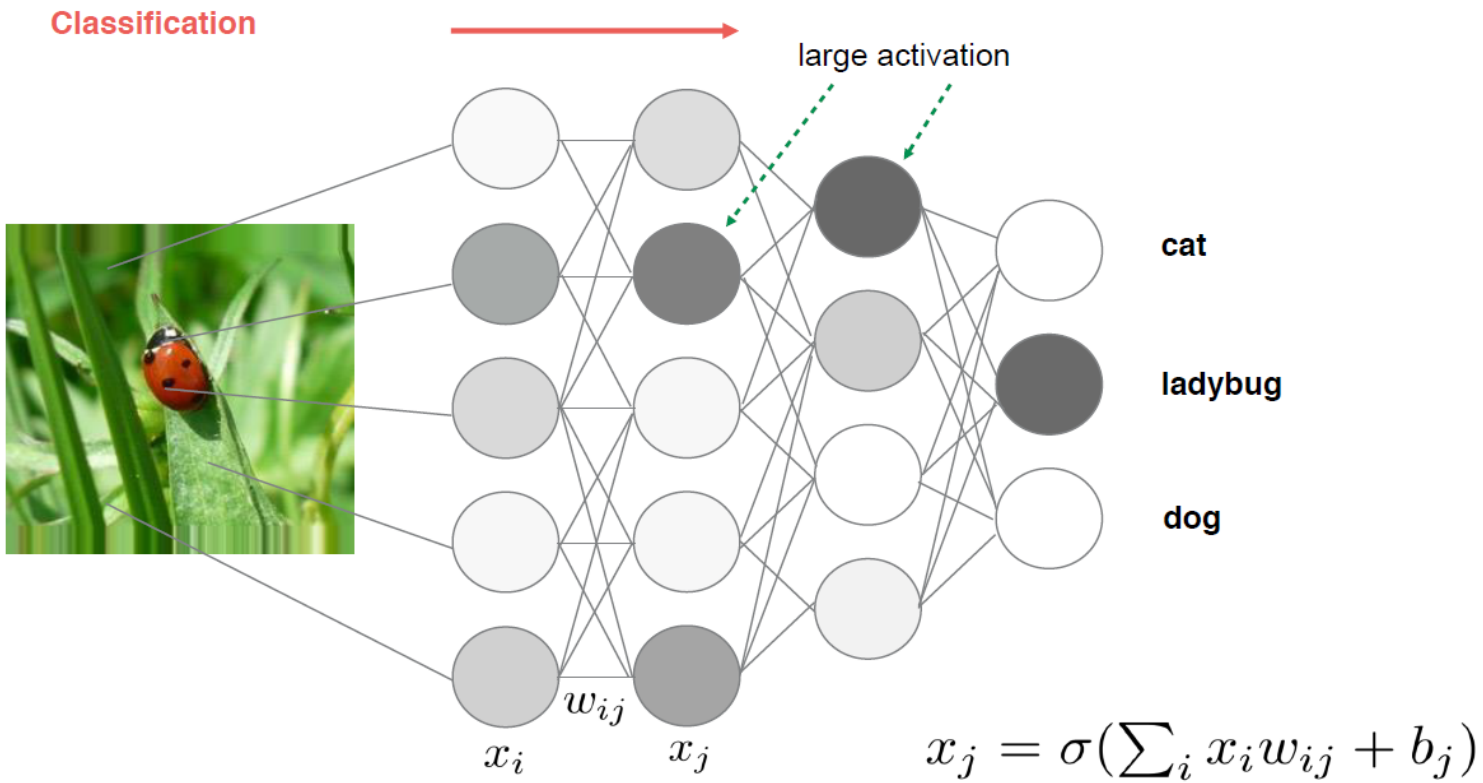
Example in [0,1]:



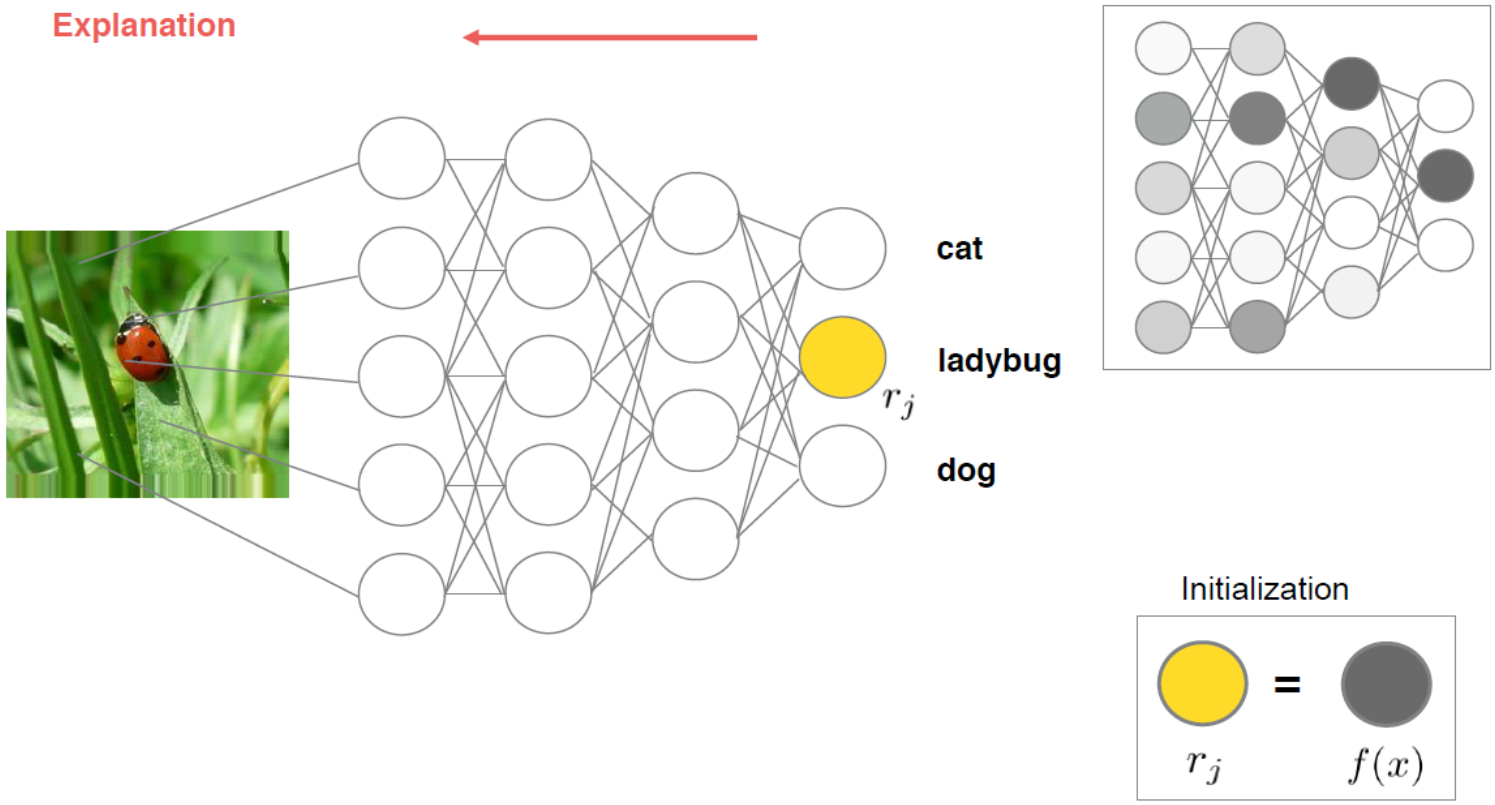
4 Explicando as predições de redes neurais profundas

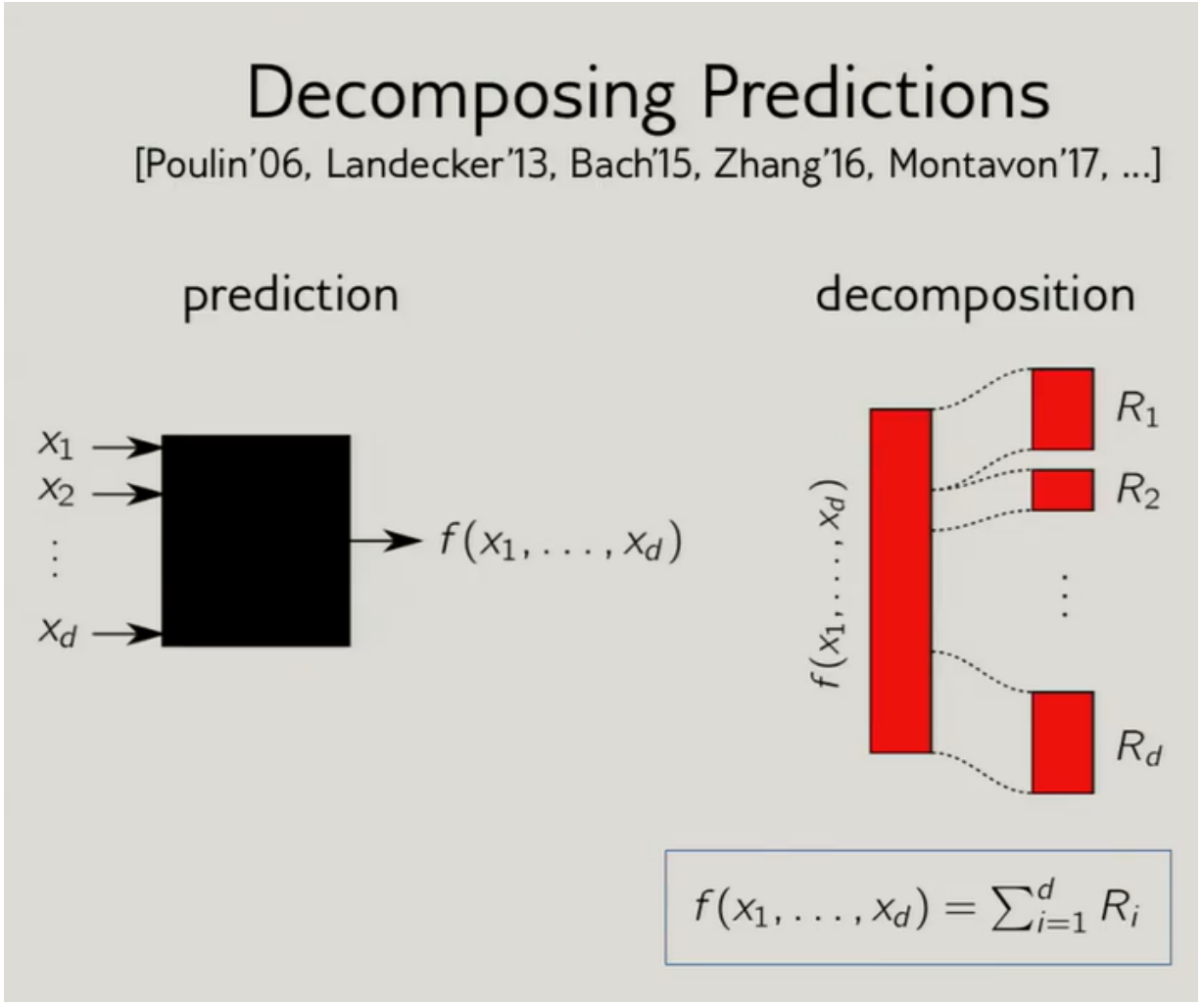
- Uma bem-sucedida iniciativa voltada para a interpretação do comportamento de redes neurais profundas é o *Heatmapping Project* (http://www.heatmapping.org/), que visa estudar técnicas específicas para decompor a predição em termos de contribuições de variáveis de entrada individuais, de tal modo que a explicação produzida possa ser visualizada da mesma forma que os dados de entrada.
- Os resultados estão fundamentados na técnica denominada *Layer-wise Relevance Propagation* (LRP), proposta em BACH et al. (2015).
- O foco continua a ser no tratamento de imagens, pela imediata possibilidade de visualização do resultado, mas a técnica é diretamente extensível a qualquer outro tipo de informação de entrada, como texto e fala.

Explaining Neural Network Predictions



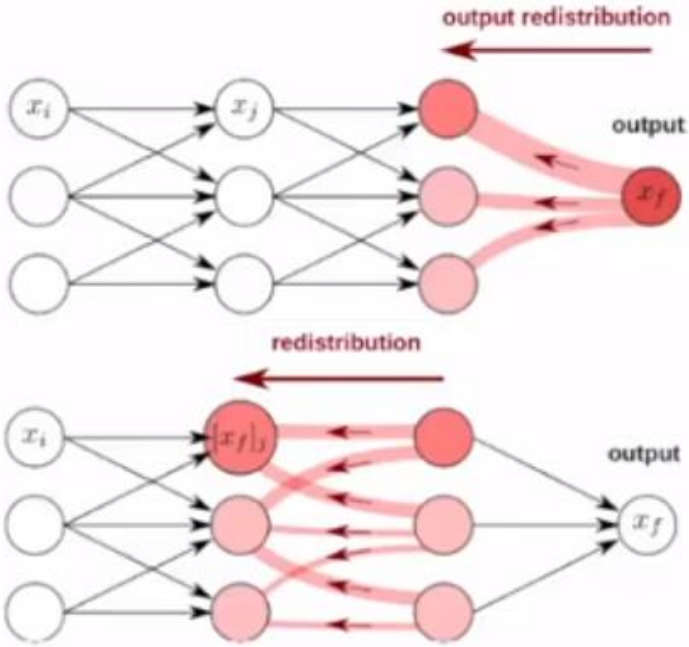
Explaining Neural Network Predictions



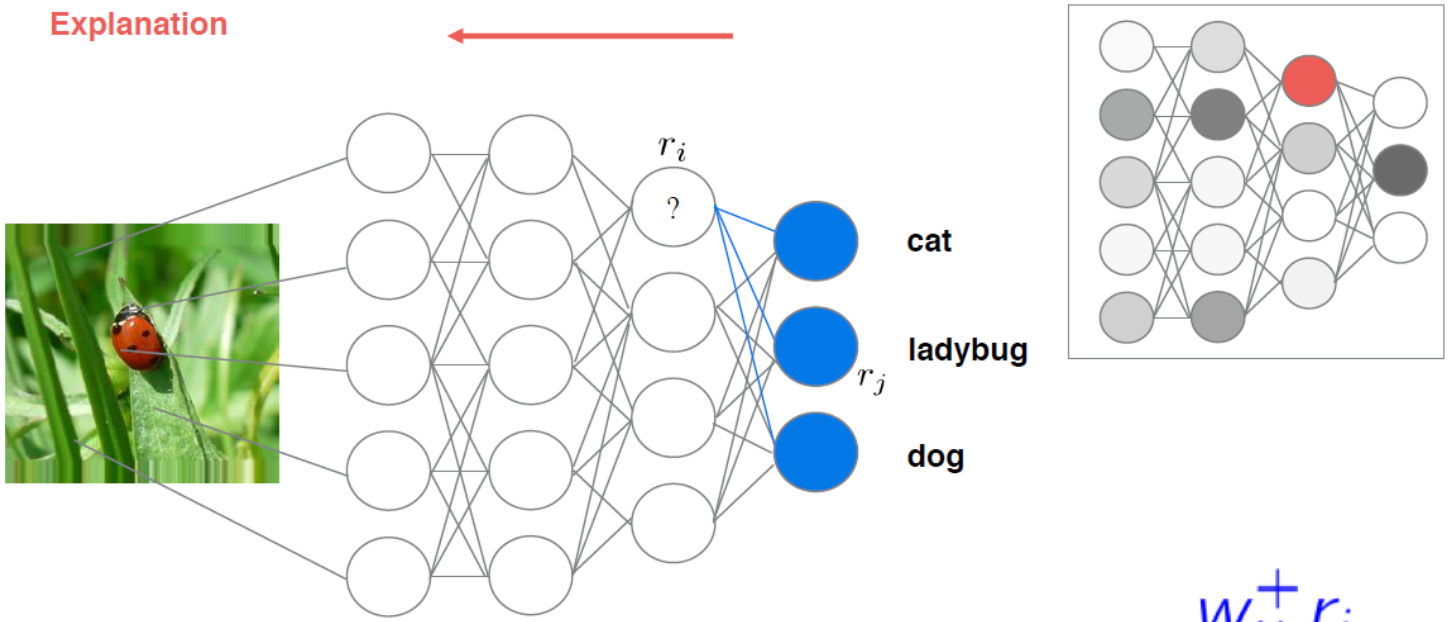


- Explora-se a propriedade de conservação de relevância, a partir da saída produzida.

- each neuron has a certain relevance
- redistribute relevance top-down: from neuron to its inputs
- relevance should be conserved when distributed (see modeling bias terms later)
- think of a neuron:
  - inputs can stimulate it to fire – positive relevance!
  - inputs can inhibit it – ???
  - negative relevance ? close to zero relevance?



Explaining Neural Network Predictions

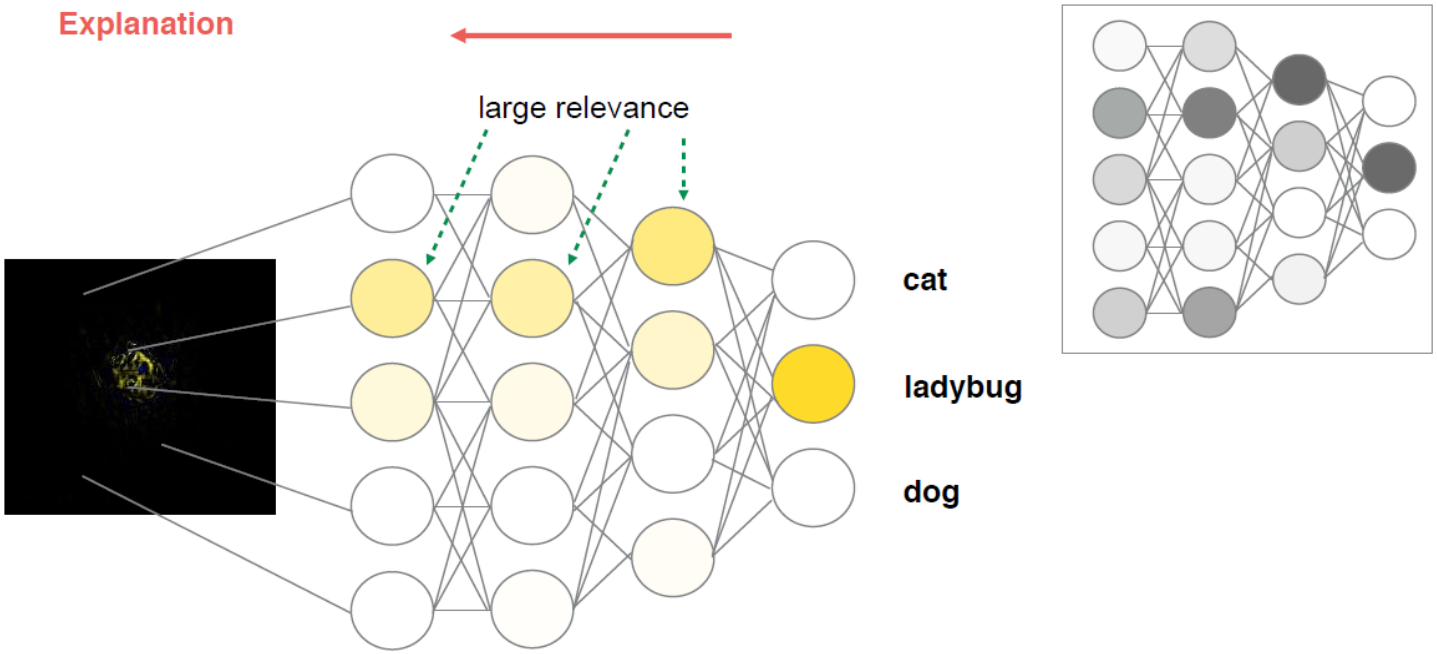


$$r_i = x_i \cdot \underbrace{\sum_j \frac{w_{ij}^+ r_j}{\sum_i x_i w_{ij}^+}}_{c_i}$$

Theoretical interpretation
Deep Taylor Decomposition

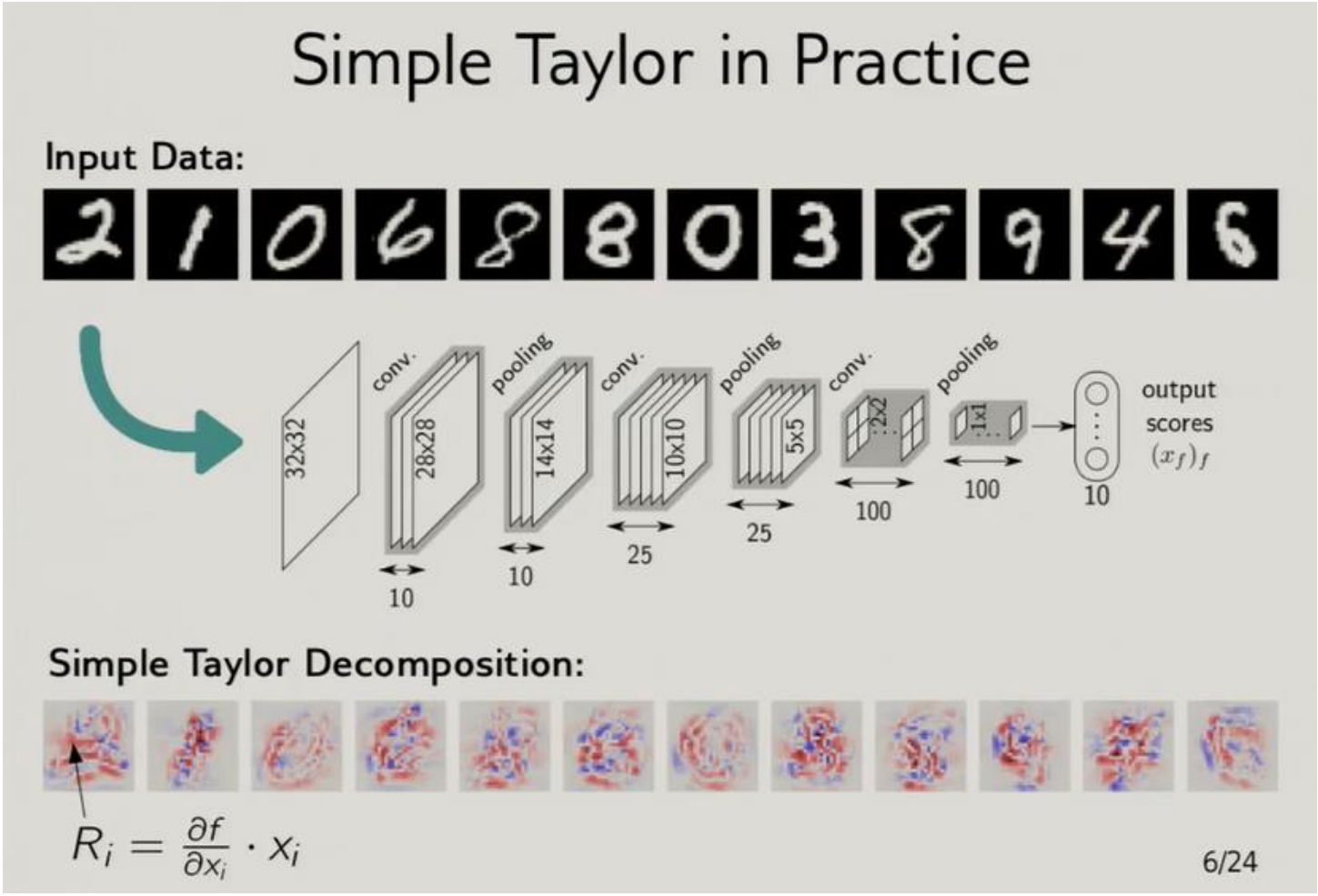
r_i deper

Explaining Neural Network Predictions

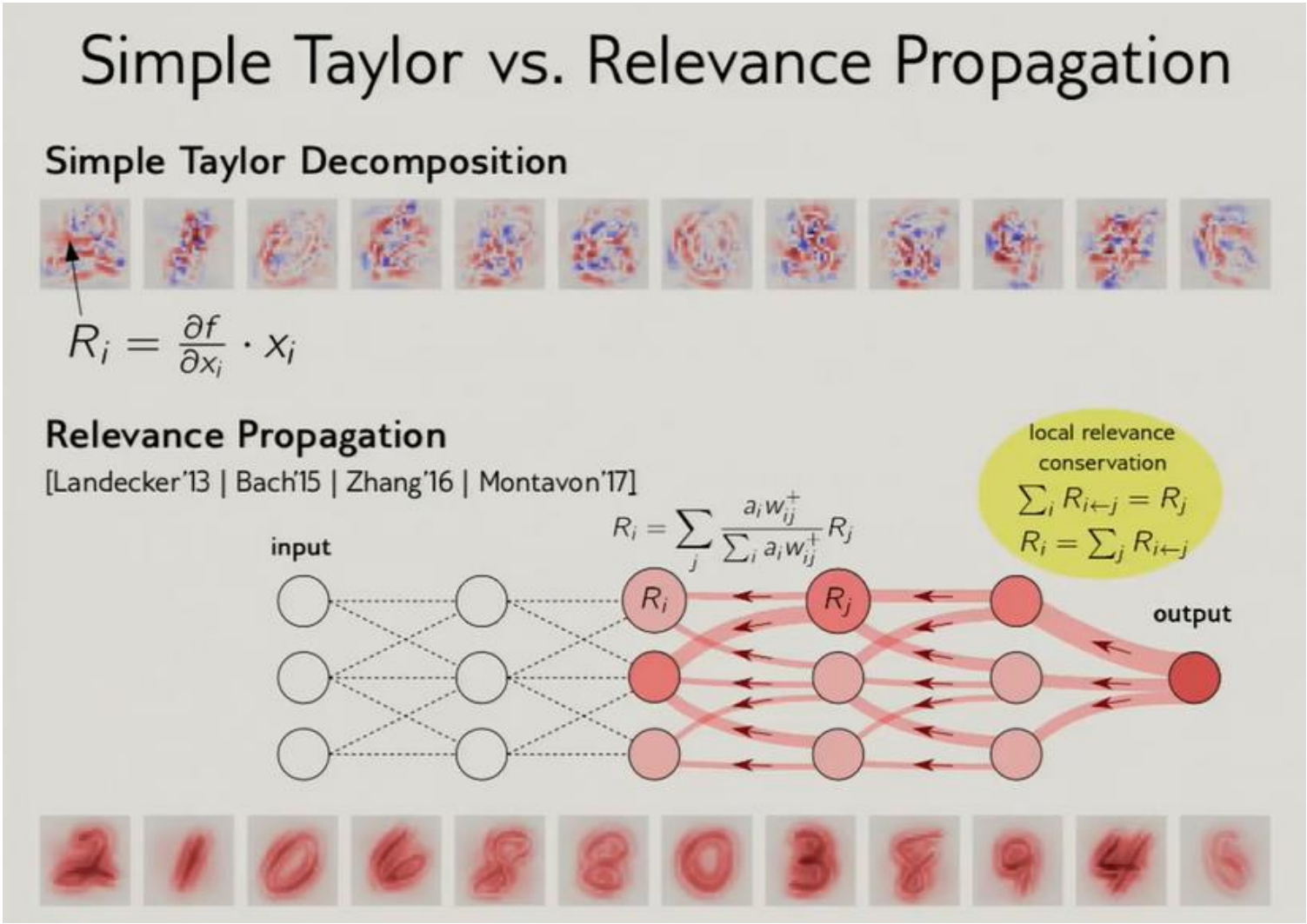


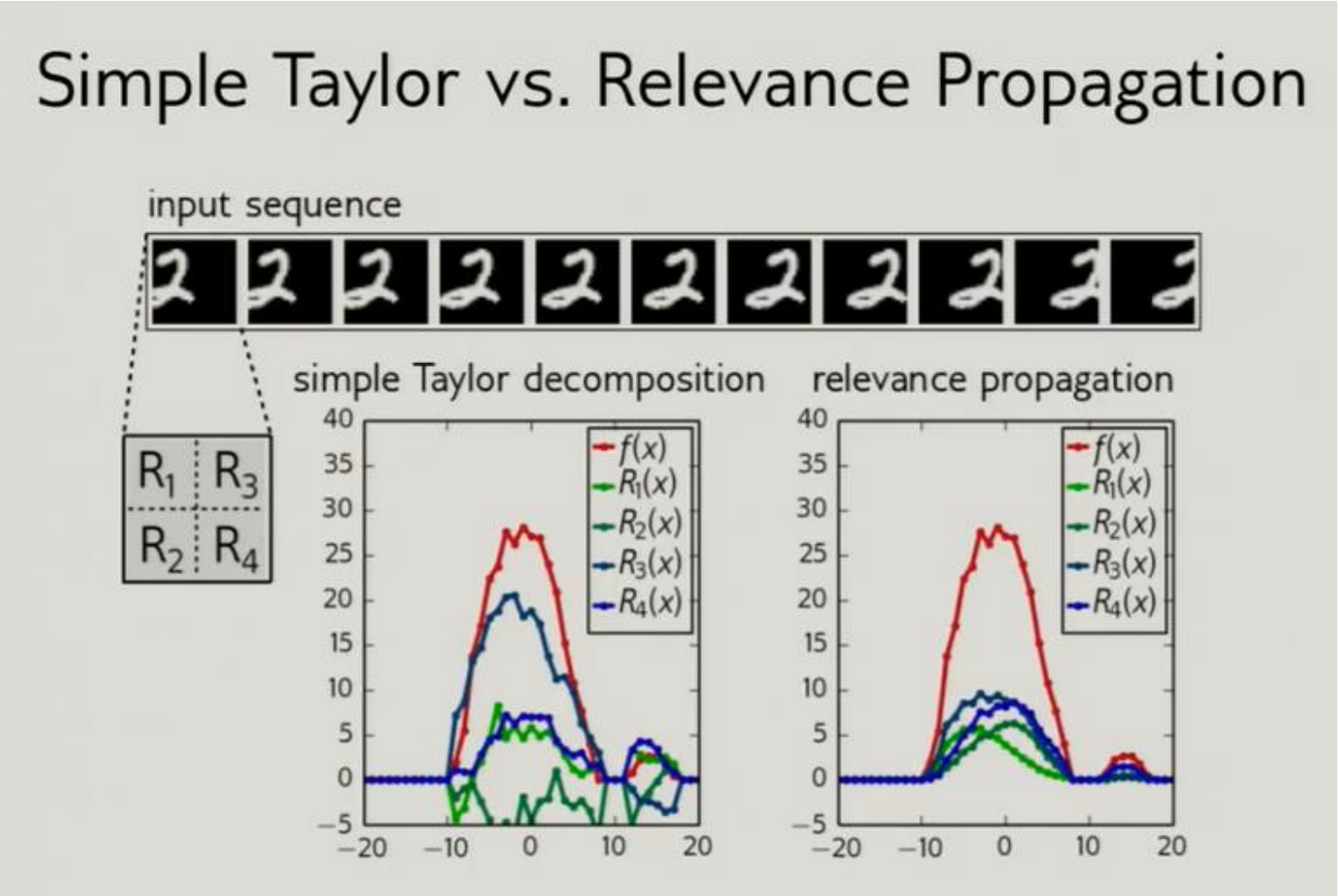
Relevance Conservation Property

$$\sum_p r_p = \ldots = \sum_i r_i = \sum_j r_j = \ldots = f(x)$$



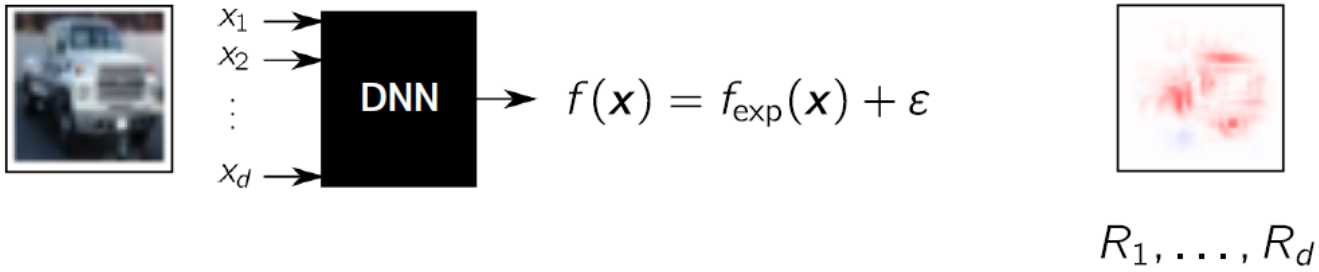
- A técnica *Simple Taylor* é equivalente à análise de sensibilidade. O exemplo acima se refere aos dados MNIST.


Simple Taylor vs. Relevance Propagation
Simple Taylor Decomposition
$R_i = \frac{\partial f}{\partial x_i} \cdot x_i$
Relevance Propagation
[Landecker'13 | Bach'15 | Zhang'16 | Montavon'17]
$R_i = \sum_j \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+} R_j$
local relevance conservation
$\sum_i R_{i \leftarrow j} = R_j$
$R_i = \sum_j R_{i \leftarrow j}$
input
output
R_i
R_j


Simple Taylor vs. Relevance Propagation
input sequence
simple Taylor decomposition
relevance propagation
R1 R3
R2 R4
f(x)
R1(x)
R2(x)
R3(x)
R4(x)
40
35
30
25
20
15
10
5
0
−5
−20
−10
0
10
20

Properties 1-2: Conservation and Positivity

[Montavon'17, see also Sun'11, Landecker'13, Bach'15]



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).

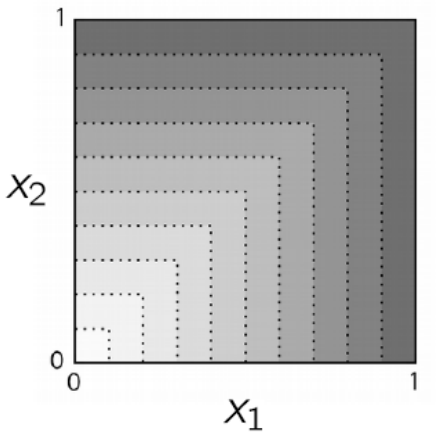
$$\forall_{p=1}^{d} : \; R_p \geq 0$$

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.

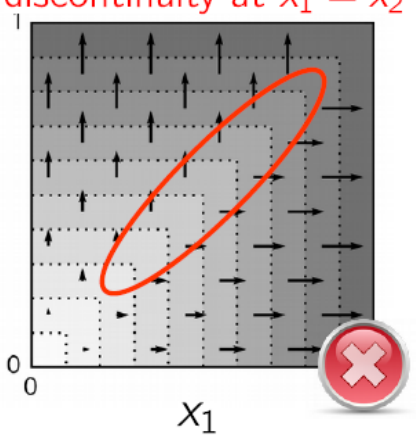
Example:

$f(\boldsymbol{x}) = \max(x_1, x_2)$

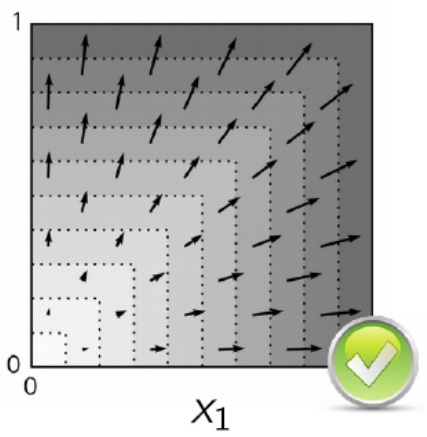


Method 1

discontinuity at $x_1 = x_2$



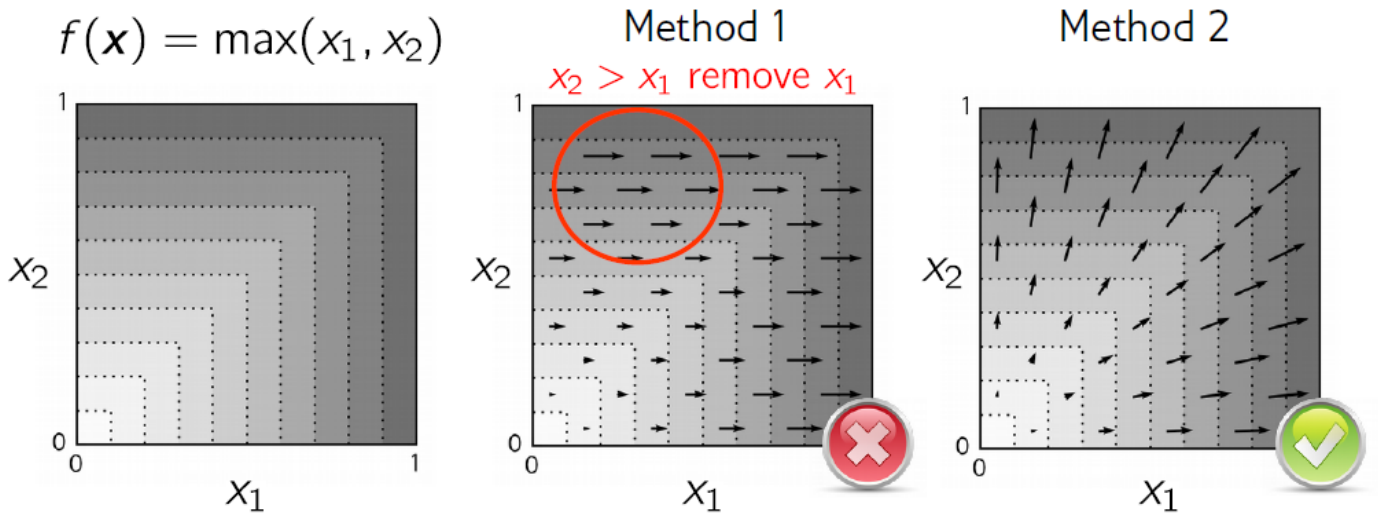
Method 2



Property 4: Selectivity [Bach'15, Samek'17]

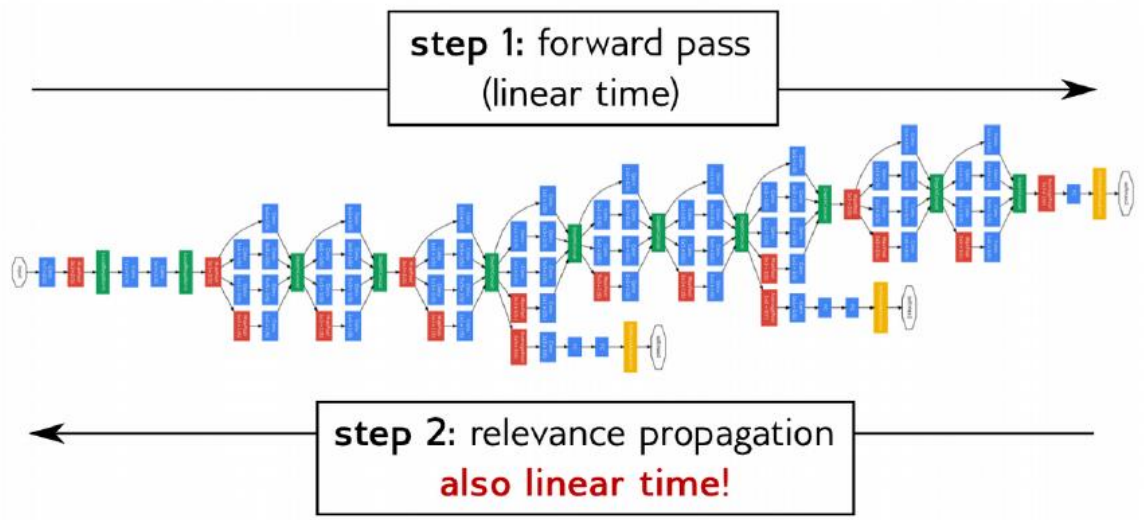
Model must agree with the explanation: If input features are attributed relevance, removing them should reduce evidence at the output.

Example:



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			✔	✔	✔	✔	✔	
...								

How LRP Scales



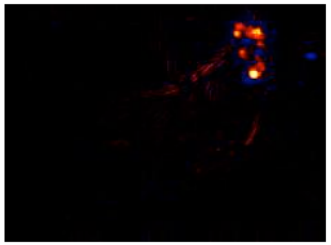
No need for much computing power. GoogleNet explanation for single image can be done on the CPU.

Linear time scaling allows to use LRP for real-time processing, or as part of training.

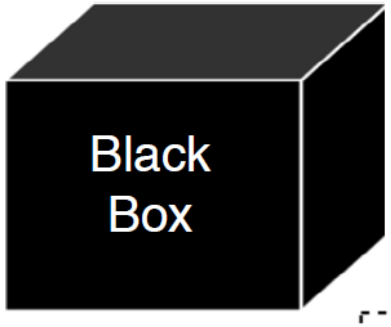
- A GoogleNet venceu a ImageNet Large-Scale Visual Recognition Challenge 2014 (ILSVRC14).

Opening the Black Box with LRP

input $\boldsymbol{x}$



heatmap



Rooster

prediction $f(\boldsymbol{x})$

Idea: Decompose function

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

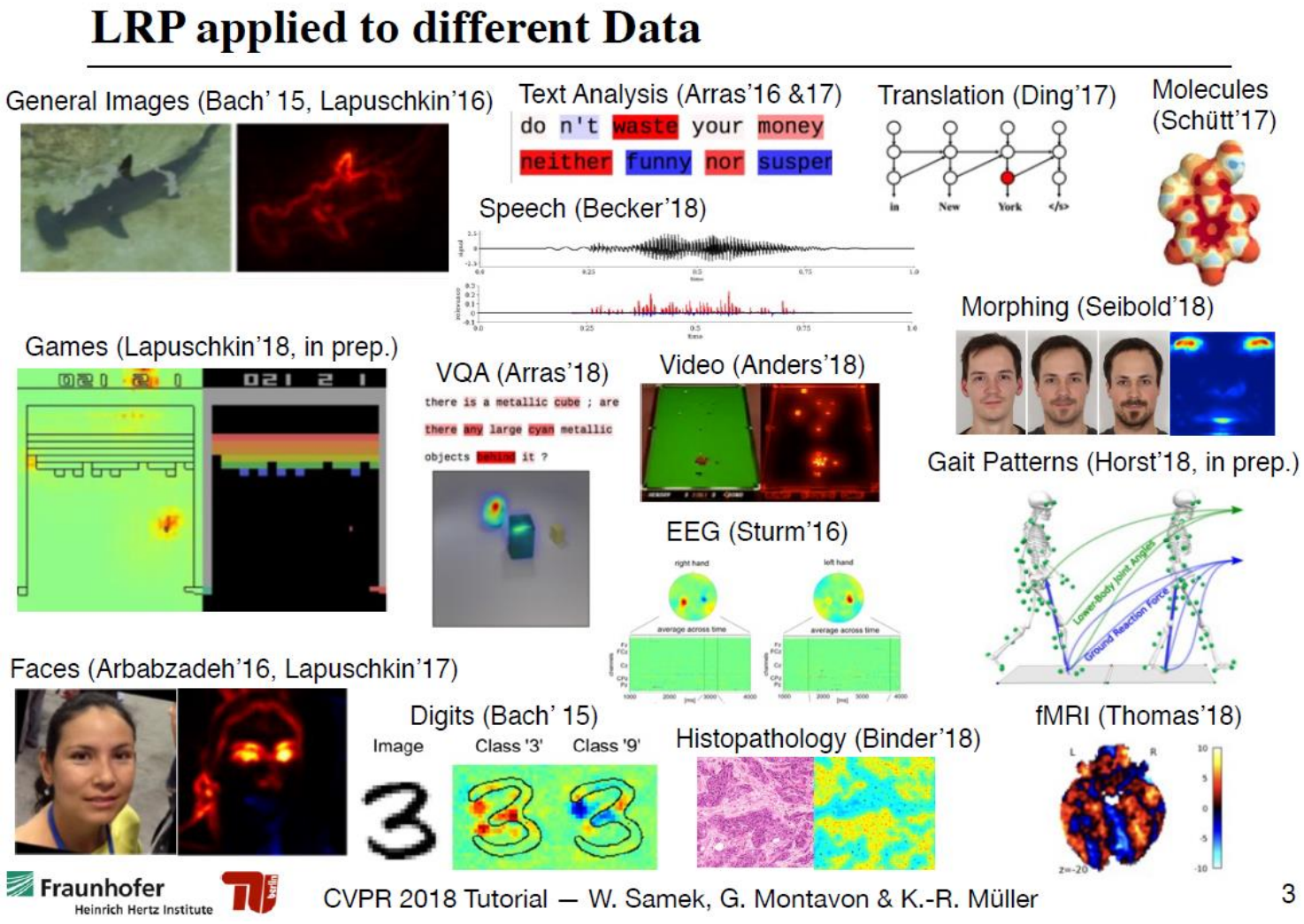
Explain prediction

(how much each pixel contributes to prediction)

Simple LRP rule (Bach et al. 2015)

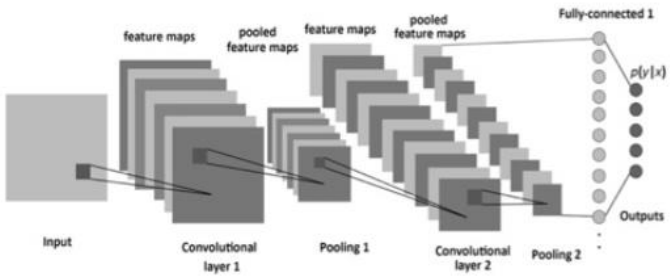
$$R_i^{(l)} = \sum_j \frac{x_i \cdot w_{ij}}{\sum_{i'} x_{i'} \cdot w_{i'j}} R_j^{(l+1)}$$

Every neuron gets its "share" of the redistributed relevance

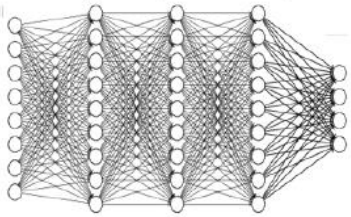

LRP applied to different Data
General Images (Bach' 15, Lapuschkin'16)
Text Analysis (Arras'16 &17)
do n't waste your money
neither funny nor susper
Translation (Ding'17)
in New York </s>
Molecules (Schütt'17)
Speech (Becker'18)
Morphing (Seibold'18)
Games (Lapuschkin'18, in prep.)
VQA (Arras'18)
there is a metallic cube ; are
there any large cyan metallic
objects behind it ?
Video (Anders'18)
Gait Patterns (Horst'18, in prep.)
Lower-Body Joint Angles
Ground Reaction Force
EEG (Sturm'16)
right hand
left hand
average across time
Faces (Arbabzadeh'16, Lapuschkin'17)
Digits (Bach' 15)
Image
Class '3'
Class '9'
Histopathology (Binder'18)
fMRI (Thomas'18)
Fraunhofer
Heinrich Hertz Institute
CVPR 2018 Tutorial — W. Samek, G. Montavon & K.-R. Müller
3

LRP applied to different Models

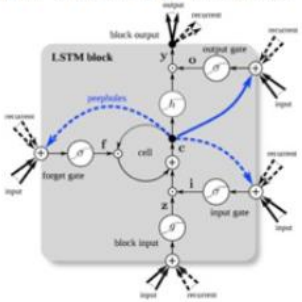
Convolutional NNs (Bach'15, Arras'17 …)



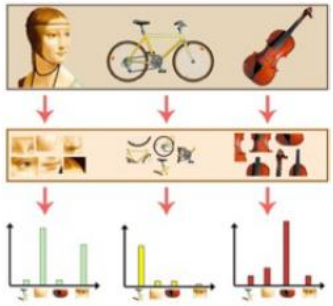
Local Renormalization Layers (Binder'16)



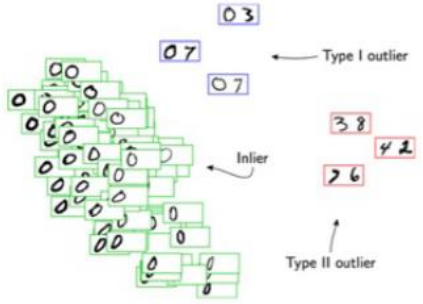
LSTM (Arras'17, Thomas'18)



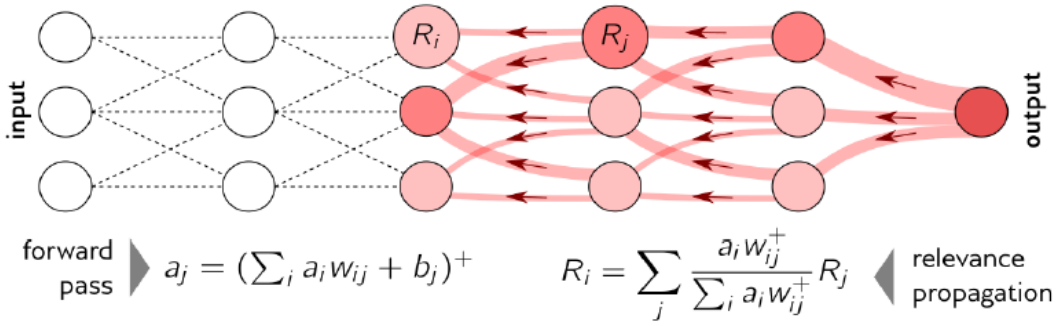
Bag-of-words / Fisher Vector models (Bach'15, Arras'16, Lapuschkin'17, Binder'18)



One-class SVM (Kauffmann'18)

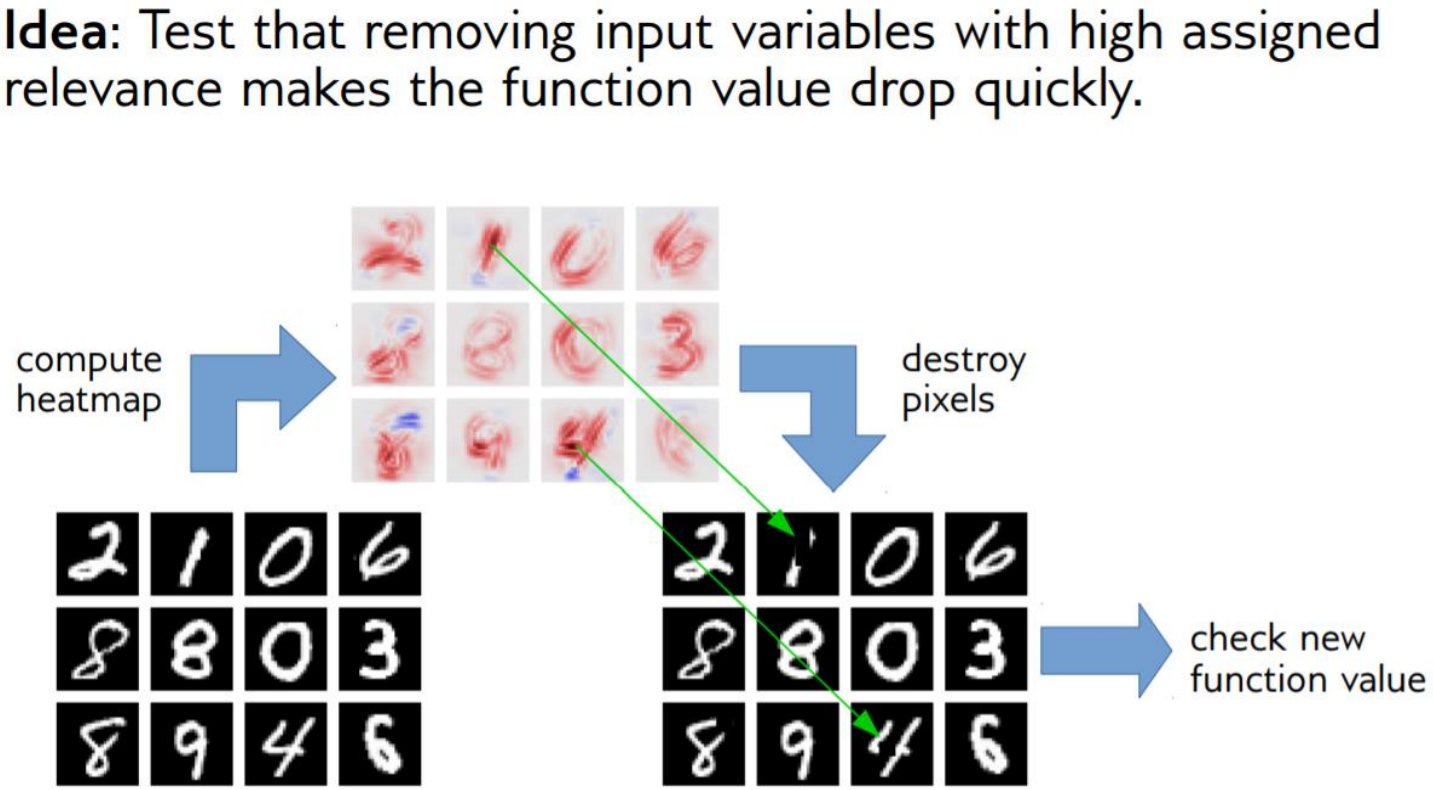


LRP works 4 all: deep models, LSTMs, kernel methods …



5 Comparação entre diferentes técnicas

- *Pixel flipping*: permite comparar métodos de interpretação alternativos



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)

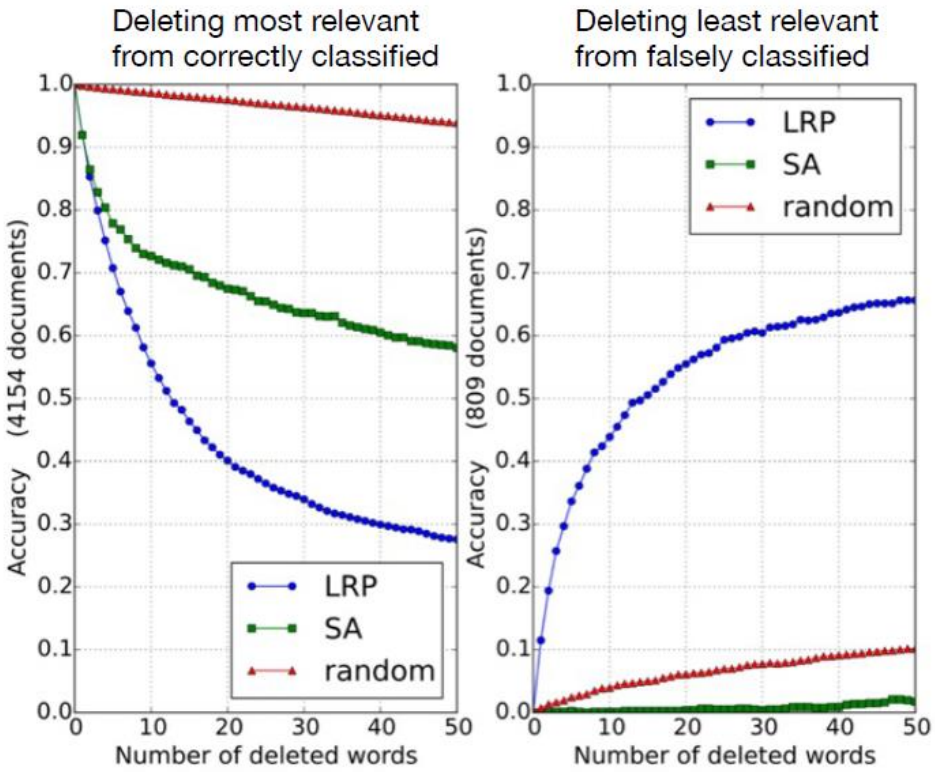
Yes, weightlessness does feel like falling. It may feel strange at first, but the body does adjust. The feeling is not too different from that of sky diving.

>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



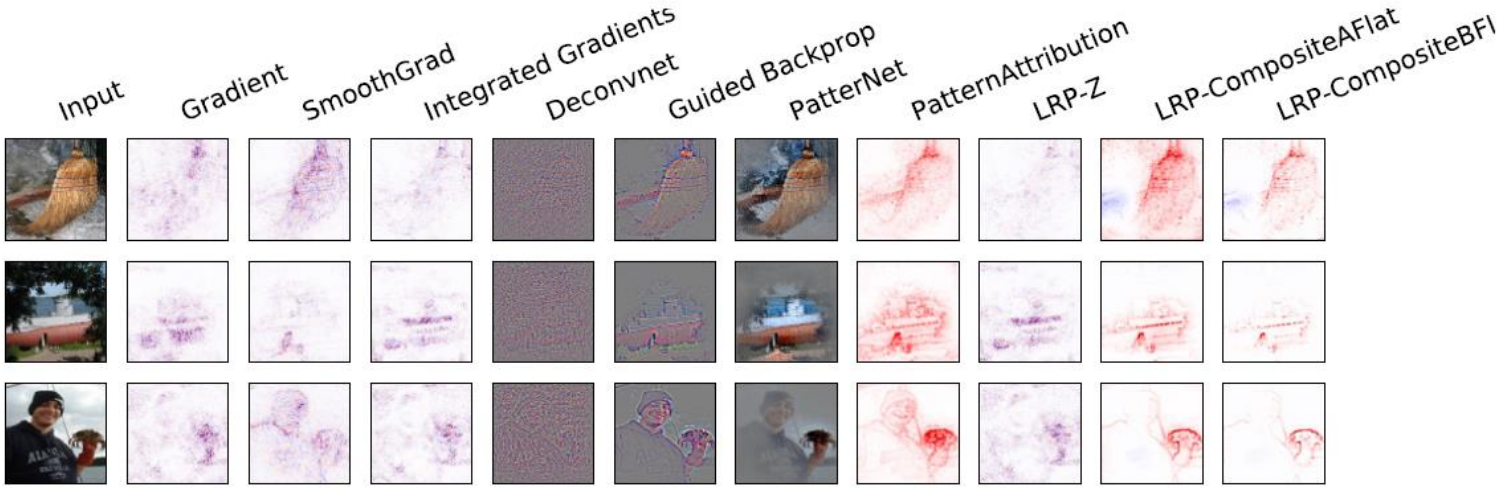
- word2vec / CNN model
- Conv → ReLU → 1-Max-Pool → FC
- trained on 20Newsgroup Dataset
- accuracy: 80.19%

LRP better than SA

LRP distinguishes between positive and negative evidence

(Arras et al. 2016)

Compare Explanation Methods



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

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

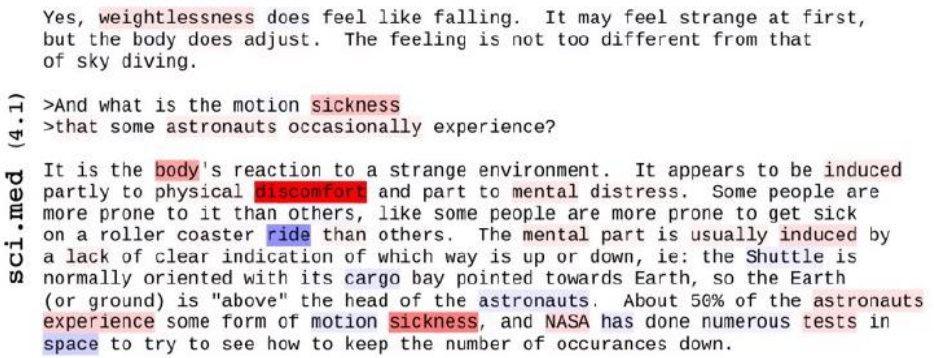
6 Aplicações

Application: Compare Classifiers

word2vec/CNN:

Performance: 80.19%

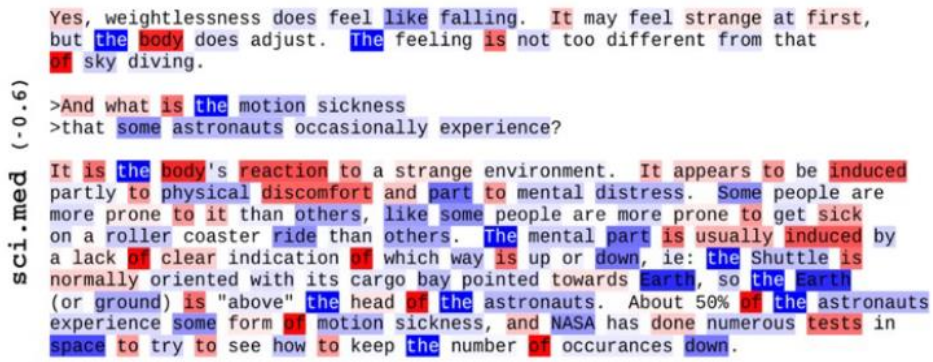
Strategy to solve the problem: identify semantically meaningful words related to the topic.



BoW/SVM:

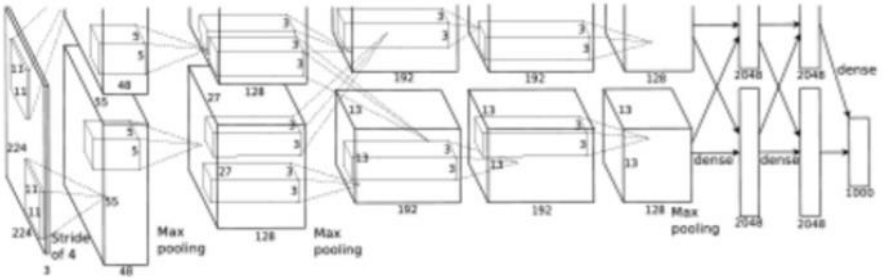
Performance: 80.10%

Strategy to solve the problem: identify statistical patterns, i.e., use word statistics

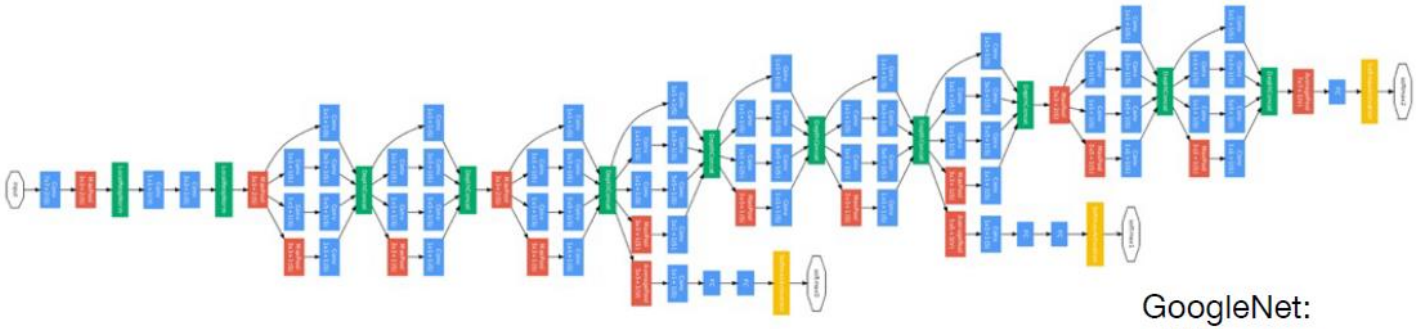


(Arras et al. 2016 & 2017)

Application: Compare Classifiers



BVLC:
- 8 Layers
- ILSRCV: 16.4%



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

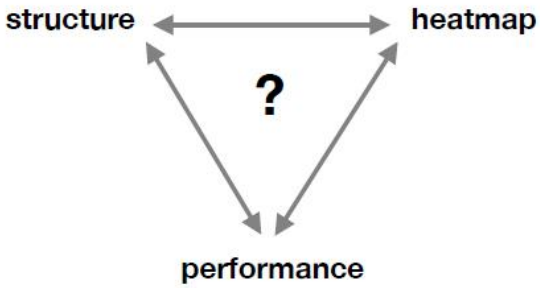
Application: Compare Classifiers

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: Measure Context Use

how important is context ?

how important is context ?

classifier

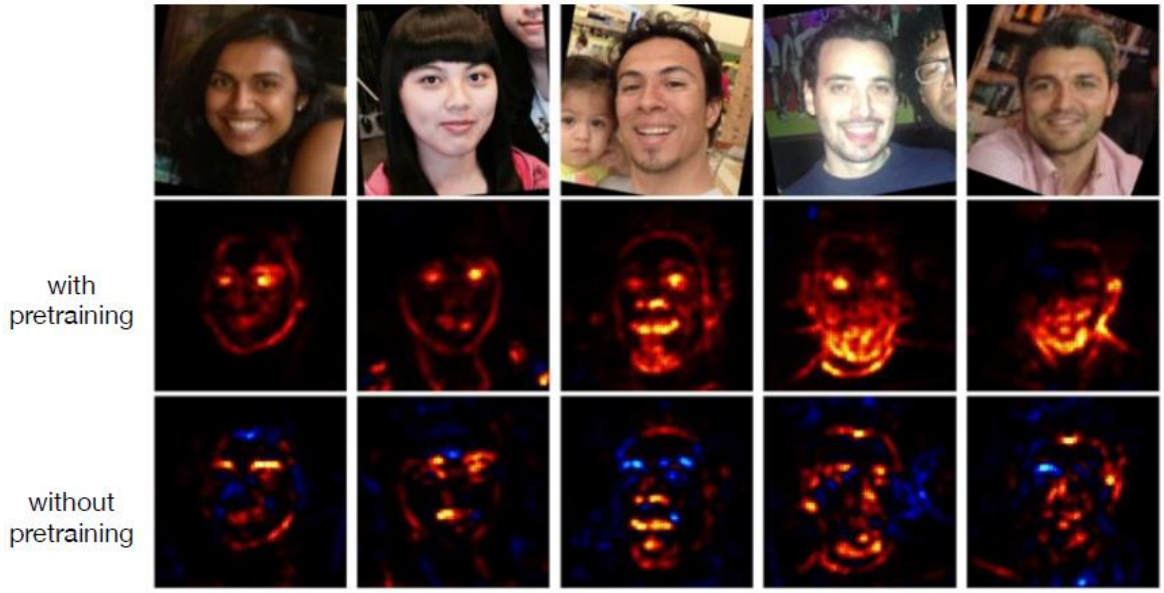
LRP decomposition allows meaningful pooling over bbox !

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

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

Application: 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

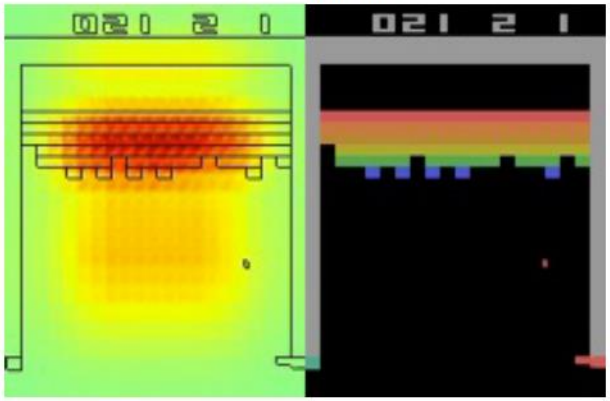
laughing speaks against 60+ (i.e., model learned that old people do not laugh)

pretraining on IMDB-WIKI

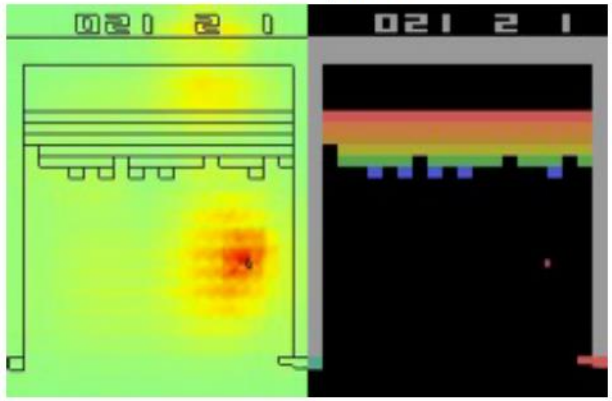
(Lapuschkin et al., 2017)

Application: Understand the model

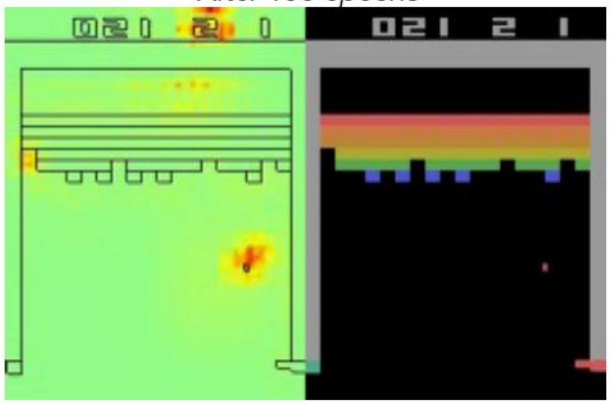
After 0 epochs



After 25 epochs



After 195 epochs



(Lapuschkin et al., in prep.)

Fraunhofer Heinrich Hertz Institute

CVPR 2018 Tutorial — W. Samek, G. Montavon & K.-R. Müller

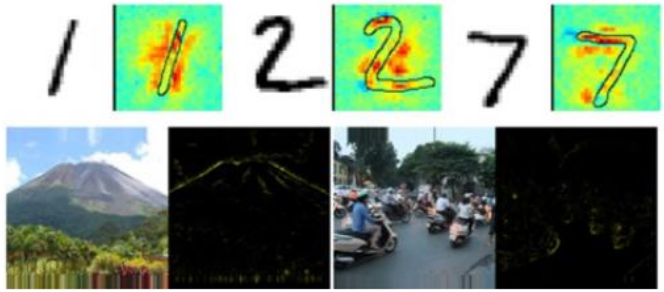
56

More information

Visit:

`http://www.heatmapping.org`

- Tutorials
- Software
- Online Demos



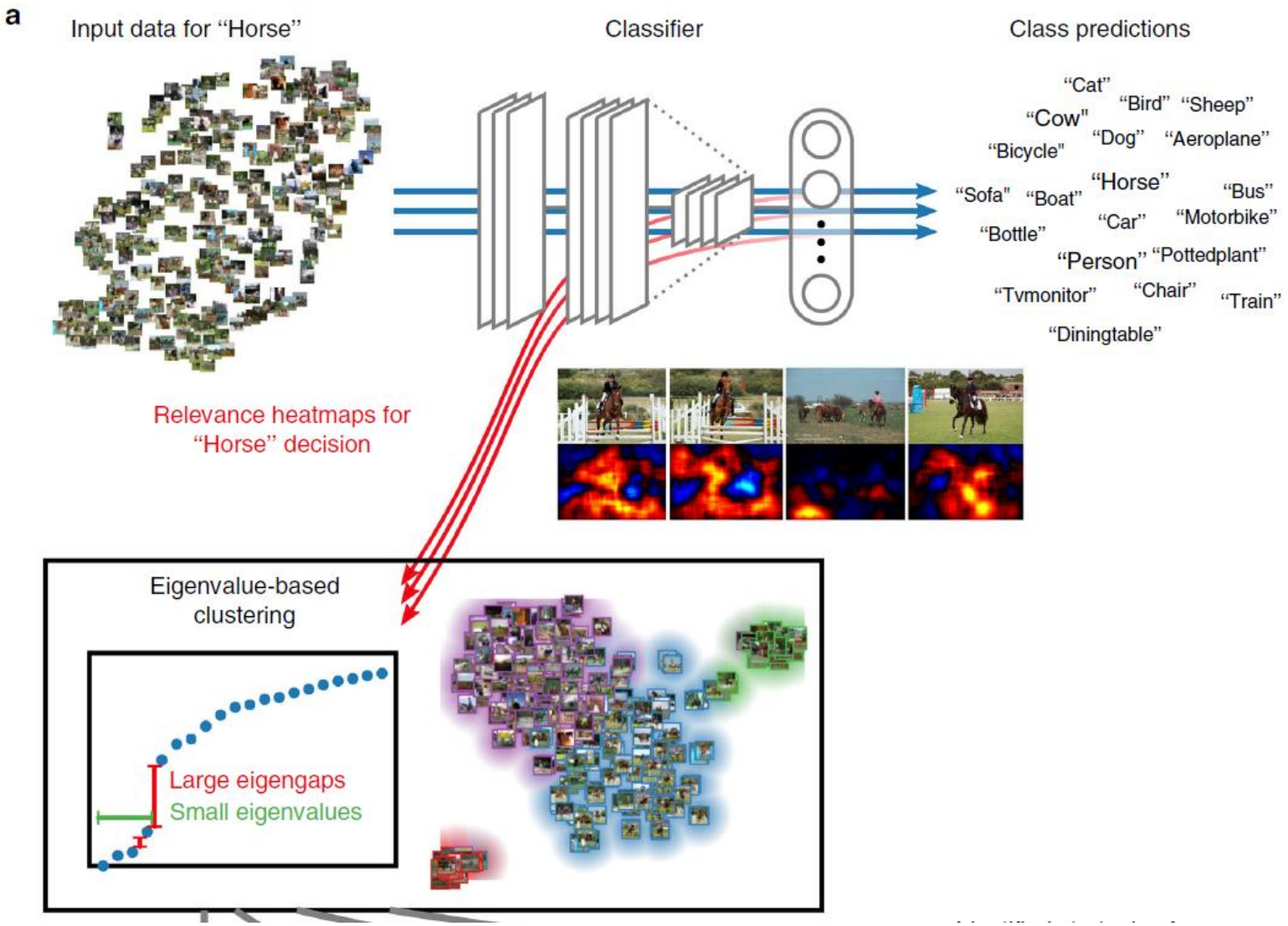
Tutorial Paper
Montavon et al., "Methods for interpreting and understanding deep neural networks", Digital Signal Processing, 73:1-5, 2018

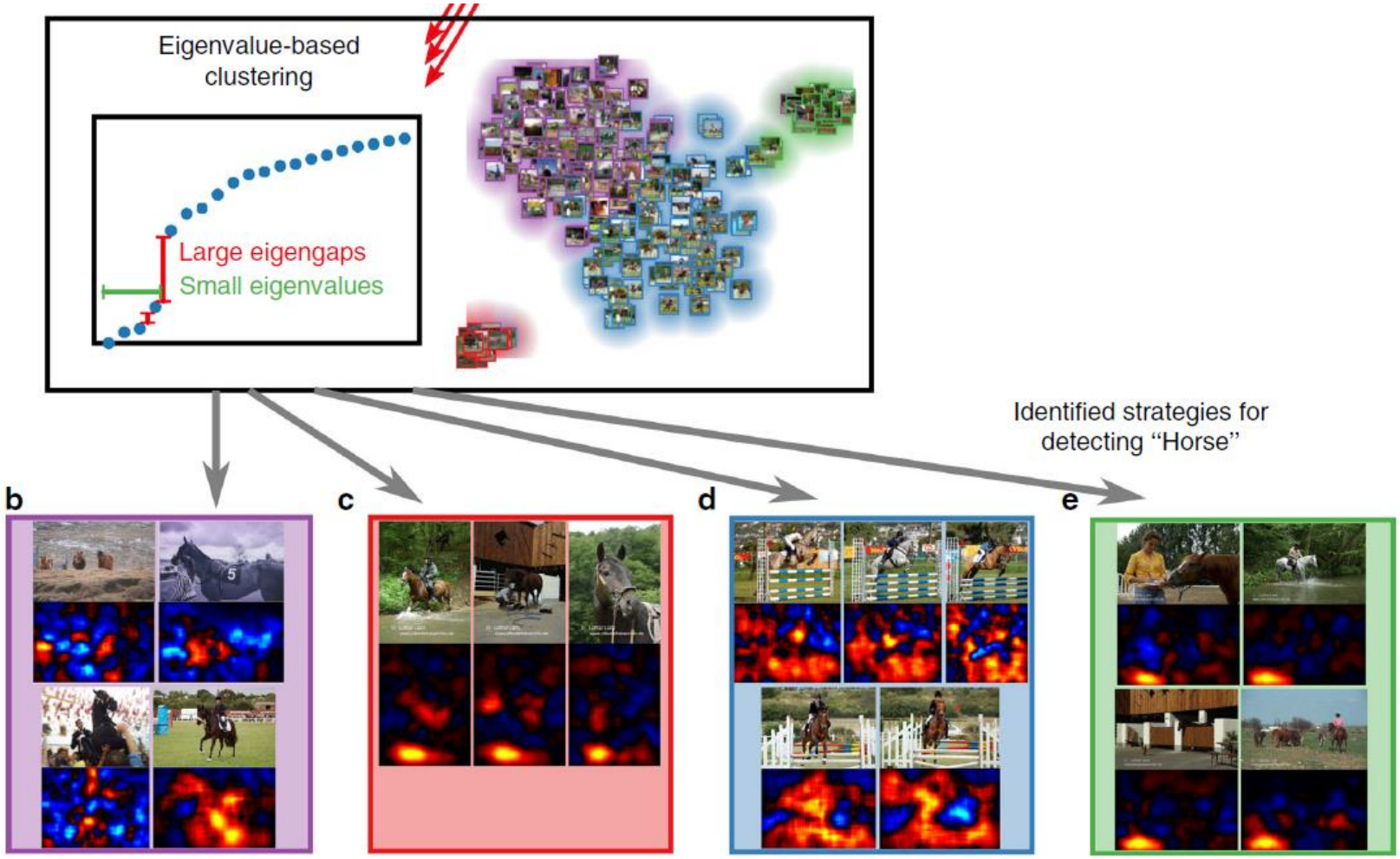
Keras Explanation Toolbox
`https://github.com/albermax/innvestigate`

7 Spectral Relevance Analysis (SpRAy)

- SpRAy (LAPUSCHKIN et al., 2019) é uma técnica que faz uma varredura nos resultados de LRP para um amplo conjunto de amostras, efetivamente capturando o comportamento do modelo de aprendizado para grandes bases de dados.
- Os textos e as figuras a seguir foram extraídos de LAPUSCHKIN et al. (2019).
- *With the SpRAy method we have proposed a tool to systematize the investigation of classifier behavior and identify the* ***broad spectrum of prediction strategies****. The SpRAy analysis is scalable and can be applied to large datasets in a semi-automated manner.*
- *We believe that such analysis is a first step towards confirming important desiderata of AI systems, such as* ***trustworthiness, fairness, and accountability*** *in the future, e.g. in context of regulations concerning models and methods of artificial intelligence, as via the general data protection regulation (GDPR).*

- *Our contribution may also add a further perspective that could in the future enrich the ongoing discussion,* ***whether machines are truly "intelligent"****.*
- *Finally, in this paper we posit that the ability to explain decisions made by learning machines allows us to judge and gain a deeper understanding of whether or not* ***the machine is embodying a particular strategic decision making****. Without this understanding we can merely monitor behavior and apply performance measures without possibility to reason deeper about the underlying learned representation. The insights obtained in this pursuit may be highly useful when striving for better learning machines and insights when applying ML in the sciences.*
- *SpRAy applies spectral clustering53 on a dataset of LRP explanations in order to identify typical, as well as atypical decision behaviors of the machine-learning model, and presents them to the user in a concise and interpretable manner.*


a
Input data for "Horse"
Classifier
Class predictions
"Cat"
"Bird"
"Sheep"
"Cow"
"Dog"
"Aeroplane"
"Bicycle"
"Horse"
"Sofa"
"Boat"
"Bus"
"Car"
"Motorbike"
"Bottle"
"Person"
"Pottedplant"
"Tvmonitor"
"Chair"
"Train"
"Diningtable"
Relevance heatmaps for
"Horse" decision
Eigenvalue-based
clustering
Large eigengaps
Small eigenvalues


Eigenvalue-based clustering
Large eigengaps
Small eigenvalues
Identified strategies for detecting "Horse"
b
c
d
e

8 Referências bibliográficas

BACH, S.; BINDER, A.; MONTAVON, G.; KLAUSCHEN, F.; MÜLLER, K.-R.; SAMEK, W. "On Pixel-wise Explanations for Non-Linear Classifier Decisions by Layer-wise Relevance Propagation", Plos One, 10(7):e0130140, 2015.

LAPUSCHKIN, S.; WÄLDCHEN, S.; BINDER, A.; MONTAVON, G.; SAMEK, W.; MÜLLER, K.-R. "Unmasking Clever Hans predictors and assessing what machines really learn", Nature Communications, 10: 1096, 2019.

MONTAVON, G.; SAMEK, W.; MÜLLER, K.-R. "Methods for interpreting and understanding deep neural networks", Digital Signal Processing, vol. 71, pp. 1-15, 2018.

SAMEK, W.; MONTAVON, G.; LAPUSCHKIN, S.; ANDERS, C.J.; MÜLLER, K.-R. "Toward Interpretable Machine Learning: Transparent Deep Neural Networks and Beyond", arXiv:2003.07631v1, 2020.

SAMEK, W.; MONTAVON, G.; VEDALDI, A.; HANSEN, L.K.; MÜLLER, K.-R. (eds.) "Explainable AI: Interpreting, Explaining and Visualizing Deep Learning", Lecture Notes on Artificial Intelligence, State-of-the-Art Survey, vol. 11700, Springer, 2019.

Further Reading I

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10, e0130140 (7).

Bach, S., Binder, A., Montavon, G., Müller, K.-R. & Samek, W. (2016). Analyzing Classifiers: Fisher Vectors and Deep Neural Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2912-2920 (2016).

Binder et al. Machine Learning for morpho-molecular Integration, *arXiv:1805.11178* (2018)

Blum, L. C., & Reymond, J. L. (2009). 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. Journal of the American Chemical Society, 131(25), 8732-8733.

Chmiela, S., Tkatchenko, A., Sauceda, H. E., Poltavsky, I., Schütt, K. T., & Müller, K. R. (2017). Machine learning of accurate energy-conserving molecular force fields. Science Advances, 3(5), e1603015.

Hansen, K., Montavon, G., Biegler, F., Fazli, S., Rupp, M., Scheffler, M., von Lilienfeld, A.O., Tkatchenko, A., and Müller, K.-R. "Assessment and validation of machine learning methods for predicting molecular atomization energies." *Journal of Chemical Theory and Computation* 9, no. 8 (2013): 3404-3419.

Hansen, K., Biegler, F., Ramakrishnan, R., Pronobis, W., von Lilienfeld, O. A., Müller, K. R., & Tkatchenko, A. (2015). Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space, J. Phys. Chem. Lett. 6, 2326−2331.

Harmeling, S., Ziehe, A., Kawanabe, M., & Müller, K. R. (2003). Kernel-based nonlinear blind source separation. *Neural Computation*, *15*(5), 1089-1124.

Sebastian Mika, Gunnar Ratsch, Jason Weston, Bernhard Scholkopf, KR Muller (1999), Fisher discriminant analysis with kernels, Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop, 41-48.

Further Reading II

Müller, K. R., Mika, S., Rätsch, G., Tsuda, K., & Schölkopf, B. (2001). An introduction to kernel-based learning algorithms. *Neural Networks, IEEE Transactions on*, *12*(2), 181-201.

Montavon, G., Braun, M. L., & Müller, K. R. (2011). Kernel analysis of deep networks. *The Journal of Machine Learning Research*, *12*, 2563-2581.

Montavon, Grégoire, Katja Hansen, Siamac Fazli, Matthias Rupp, Franziska Biegler, Andreas Ziehe, Alexandre Tkatchenko, Anatole V. Lilienfeld, and Klaus-Robert Müller. "Learning invariant representations of molecules for atomization energy prediction." In *Advances in Neural Information Processing Systems*, pp. 440-448 . (2012).

Montavon, G., Braun, M., Krueger, T., & Muller, K. R. (2013). Analyzing local structure in kernel-based learning: Explanation, complexity, and reliability assessment. *IEEE Signal Processing Magazine, 30*(4), 62-74.

Montavon, G., Orr, G. & Müller, K. R. (2012). Neural Networks: Tricks of the Trade, Springer LNCS 7700. Berlin Heidelberg.

Montavon, Grégoire, Matthias Rupp, Vivekanand Gobre, Alvaro Vazquez-Mayagoitia, Katja Hansen, Alexandre Tkatchenko, Klaus-Robert Müller, and O. Anatole von Lilienfeld. "Machine learning of molecular electronic properties in chemical compound space." New Journal of Physics 15, no. 9 (2013): 095003.

Montavon, G., Lapuschkin, S., Binder, A., Samek, W. and Müller, K.R., Explaining nonlinear classification decisions with deep taylor decomposition. Pattern Recognition, 65, 211-222 (2017)

Montavon, G., Samek, W., & Müller, K. R., Methods for interpreting and understanding deep neural networks, Digital Signal Processing, 73:1-5, (2018).

Further Reading III

Rupp, M., Tkatchenko, A., Müller, K. R., & von Lilienfeld, O. A. (2012). Fast and accurate modeling of molecular atomization energies with machine learning. Physical review letters, 108(5), 058301.

K. T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. R. Müller, and E. K. U. Gross, How to represent crystal structures for machine learning: Towards fast prediction of electronic properties Phys. Rev. B 89, 205118 (2014)

K.T. Schütt, F Arbabzadah, S Chmiela, KR Müller, A Tkatchenko, Quantum-chemical insights from deep tensor neural networks, Nature Communications 8, 13890 (2017)

K.T. Schütt, H.E. Sauceda, , P.J. Kindermans, , A. Tkatchenko and K.R. Müller, SchNet–A deep learning architecture for molecules and materials. The Journal of Chemical Physics, 148(24), p.241722. (2018)

Samek, W., Binder, A., Montavon, G., Lapuschkin, S. and Müller, K.R., Evaluating the visualization of what a deep neural network has learned. IEEE transactions on neural networks and learning systems, 28(11), pp.2660-2673 (2017)