

Deep Learning (Parte 5)

Índice Geral

Parte 5.1

1 Interpretabilidade em redes neurais profundas	2
2 Distintas abordagens na literatura	10
3 Análise de sensibilidade e suas limitações	16
4 Explicando as previsões de redes neurais profundas	20
5 Comparação entre diferentes técnicas	39
6 Aplicações	45
7 Referências bibliográficas	53

Parte 5.2

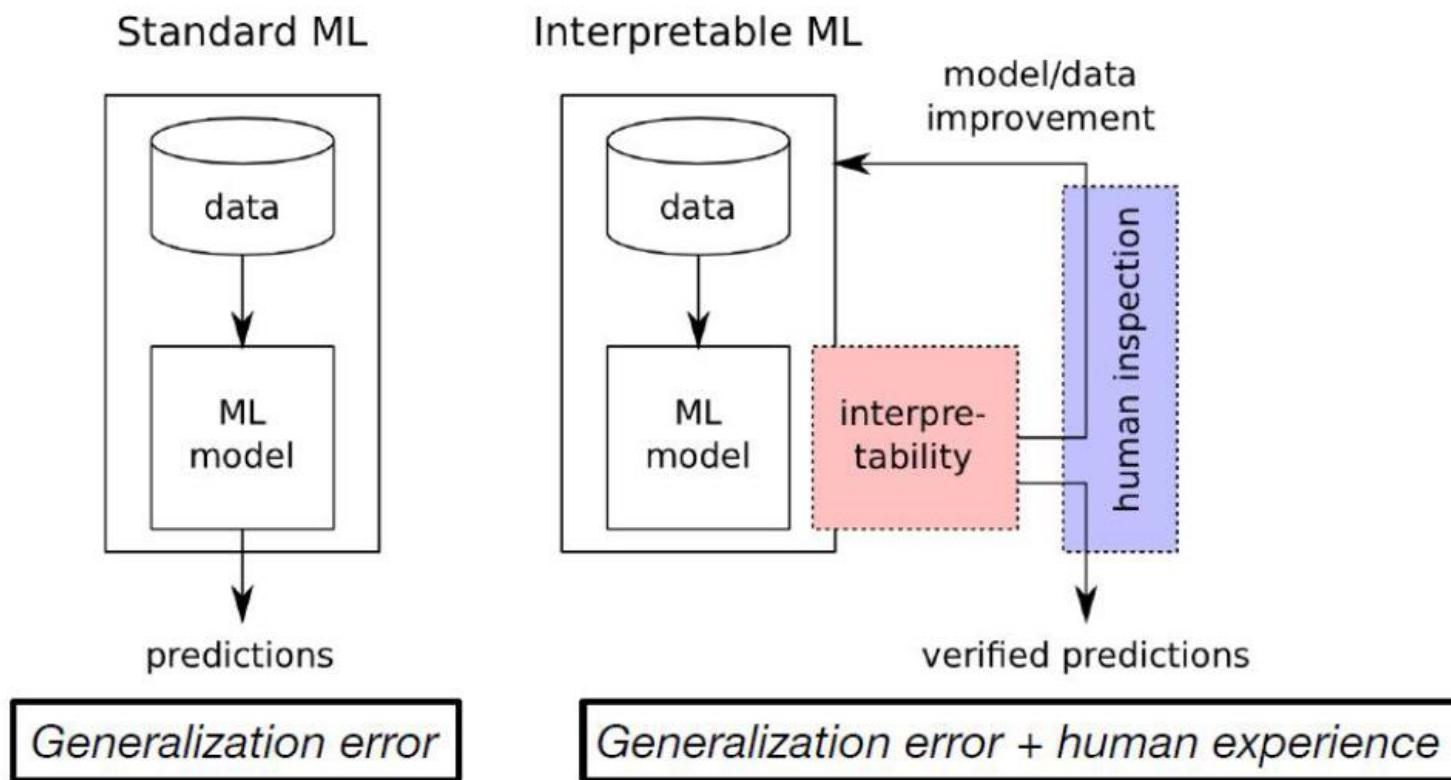
8 Generative Adversarial Networks	57
---	----

1 Interpretabilidade em redes neurais profundas

- Nota: Todo o conteúdo em inglês deste material 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/>]
 - ✓ Montavon, G. “Deep Taylor Decomposition”
https://www.youtube.com/watch?v=gy_Cb4Do_YE&list=PLMZWEhlgn-Oggo9AJdRaHA7k1-rJP5qI
- A busca por modelos de aprendizado interpretáveis está longe de ser recente, mas até recentemente 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 conseguiram 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.
- 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 previsões.
- 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.

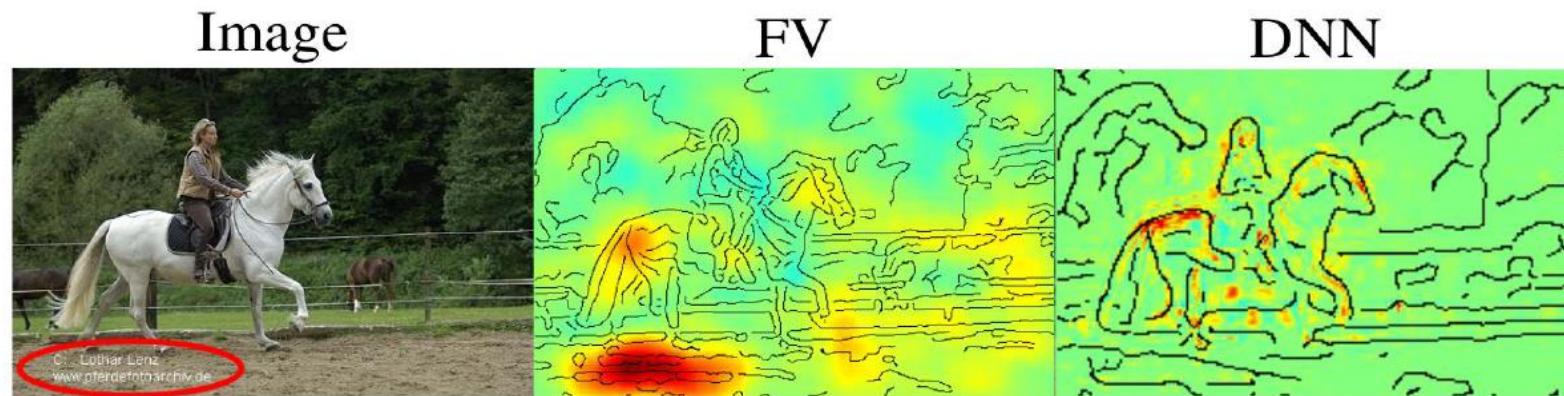
“With interpretability we can ensure that ML models work in compliance to proposed legislation.”

- 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 para que uma imagem seja classificada numa certa classe?

Application: Comparing 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%



- 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.

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.

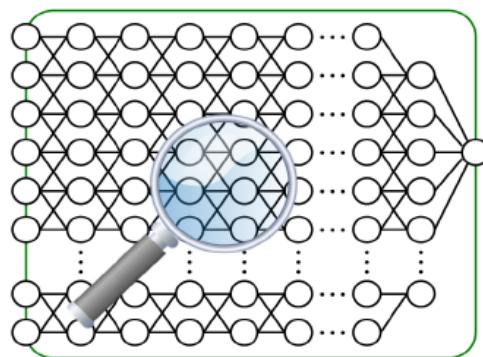
Index Terms—Data science, knowledge discovery, domain knowledge, scientific theory, physical consistency, interpretability

◆

2 Distintas abordagens na literatura

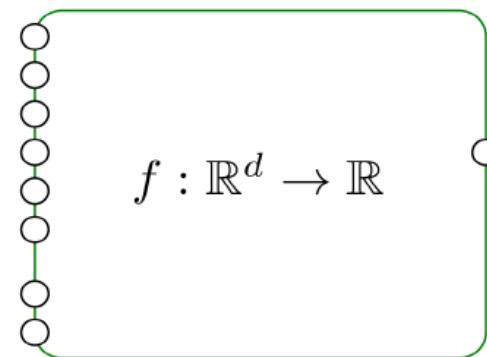
Understanding Deep Nets: Two Views

mechanistic
understanding

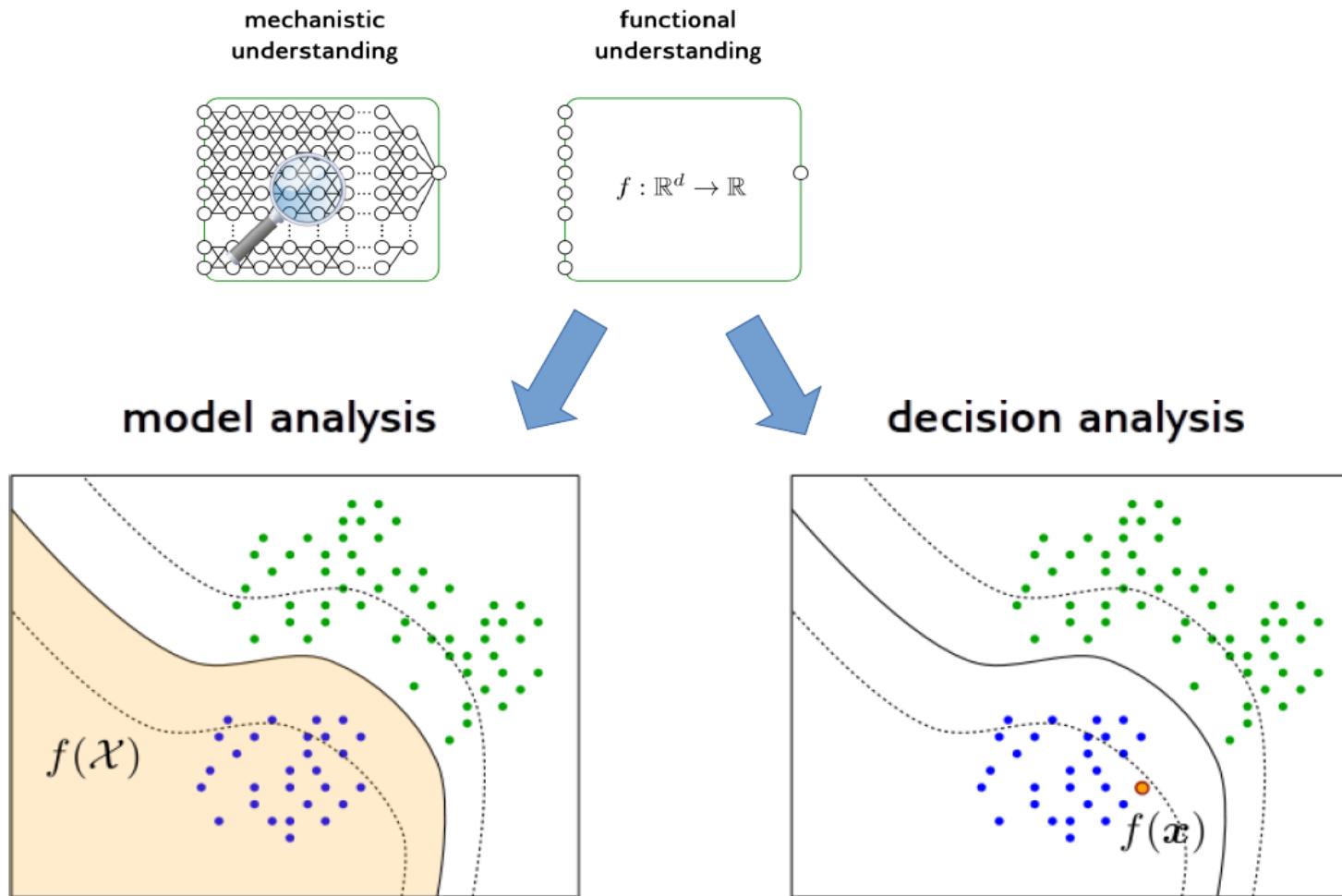


Understanding what mechanism the network uses to solve a problem or implement a function.

functional
understanding

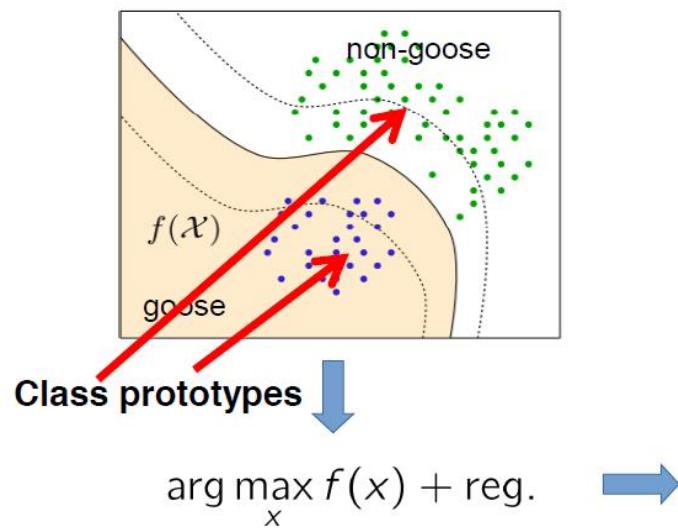


Understanding how the network relates the input to the output variables.



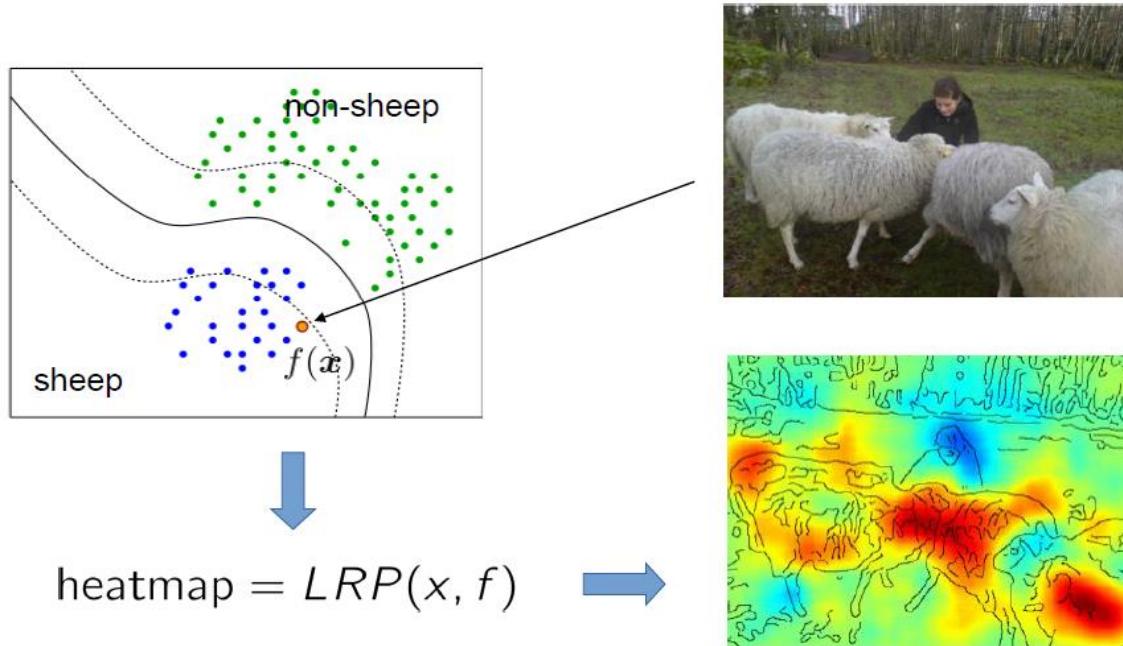
Approach 1: Class Prototypes

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



Approach 2: Individual Explanations

“Why is a given image classified as a sheep?”



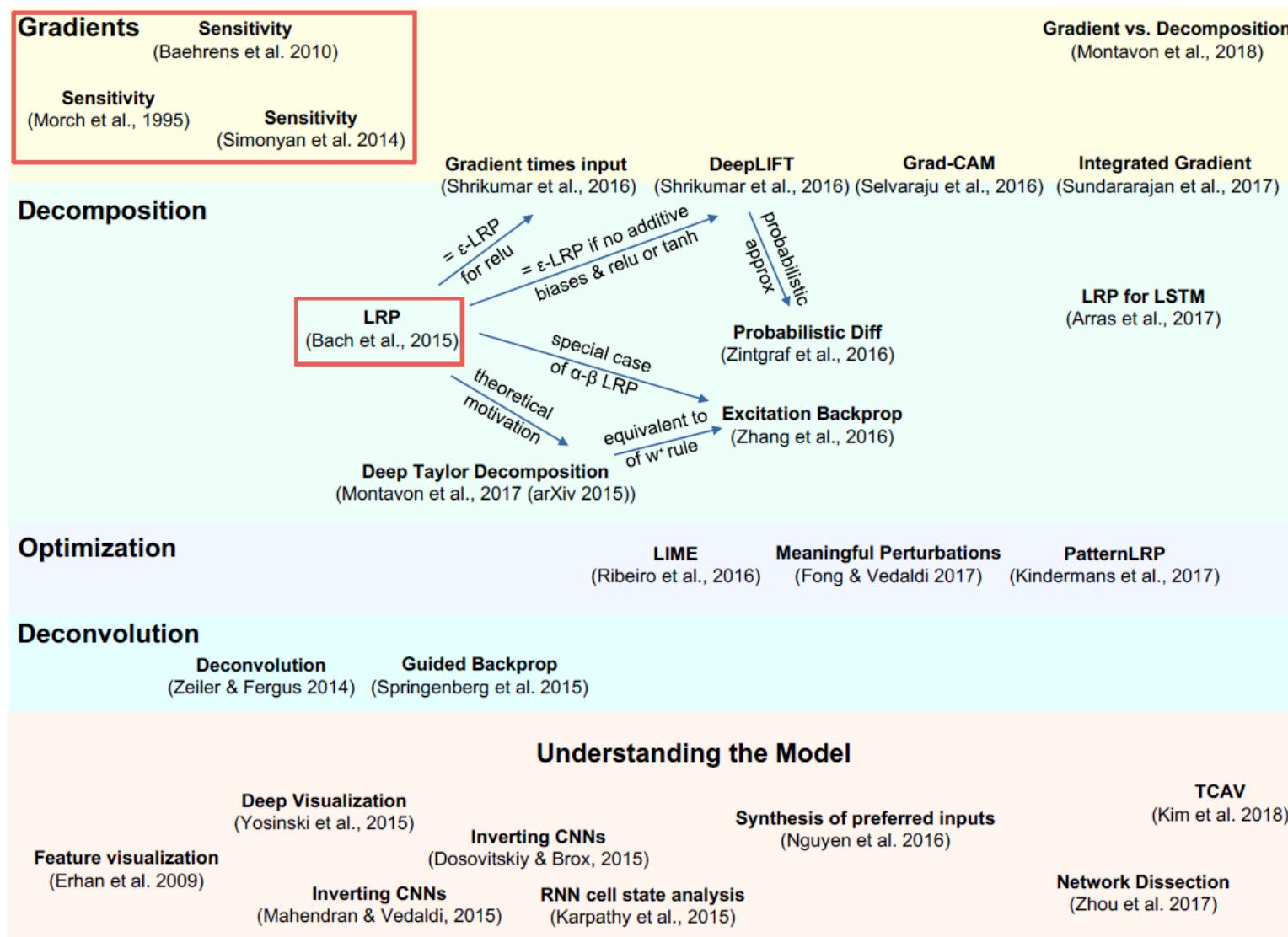
Images from Lapuschkin'16

Overview of Explanation Methods

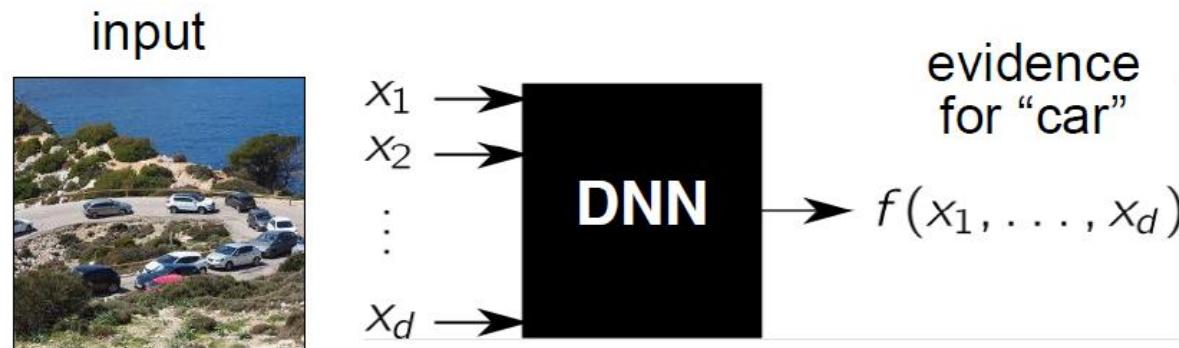
Baehrens'10 Gradient	Sundarajan'17 Int Grad	Zintgraf'17 Pred Diff	Ribeiro'16 LIME	Haufe'15 Pattern
Zurada'94 Gradient	Symonian'13 Gradient	Zeiler'14 Occlusions	Fong'17 M Perturb	Kindermans'17 PatternNet
Poulin'06 Additive	Lundberg'17 Shapley	Bazen'13 Taylor	Montavon'17 Deep Taylor	Shrikumar'17 DeepLIFT
Zeiler'14 Deconv	Landecker'13 Contrib Prop		Bach'15 LRP	Zhang'16 Excitation BP
Caruana'15 Fitted Additive	Springenberg'14 Guided BP	Zhou'16 GAP		Selvaraju'17 Grad-CAM

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:

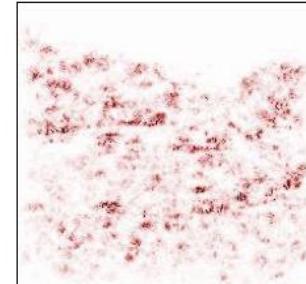
$$R_i = \left(\frac{\partial f}{\partial x_i} \right)^2$$

Understanding Sensitivity Analysis

Sensitivity analysis:



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



Problem: sensitivity analysis does not highlight cars

Observation:

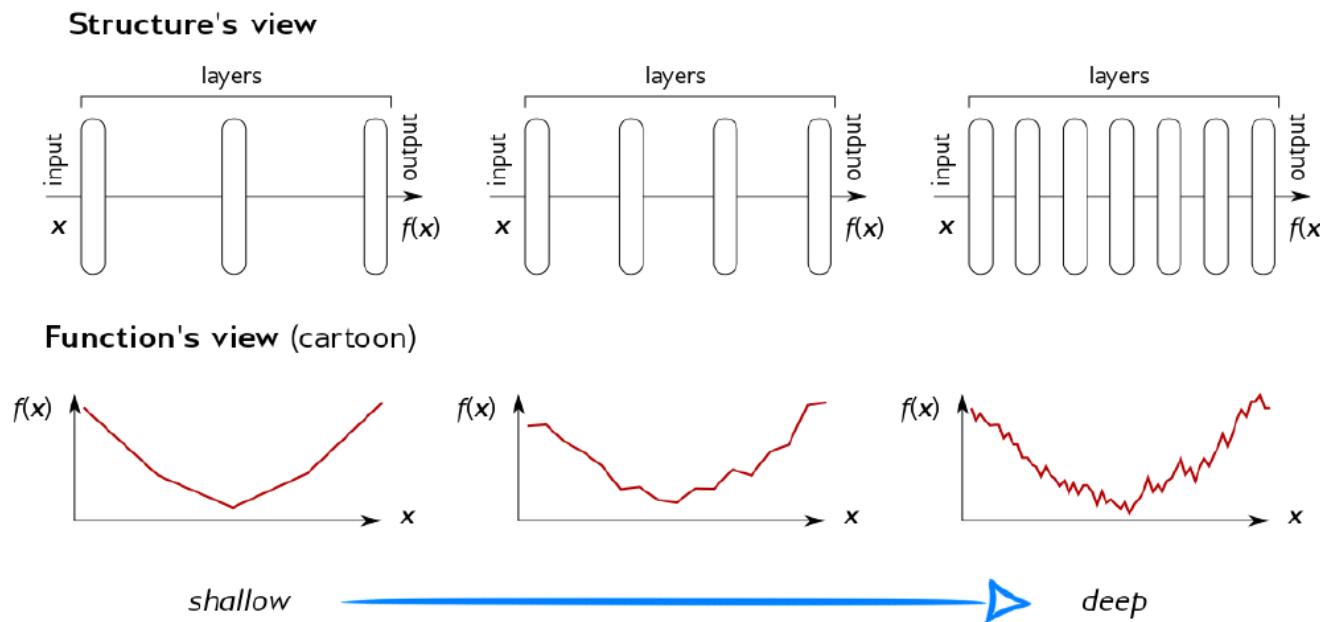
$$\sum_{i=1}^d \left(\frac{\partial f}{\partial x_i} \right)^2 = \|\nabla_x f\|^2$$

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.



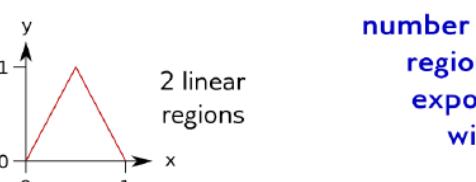
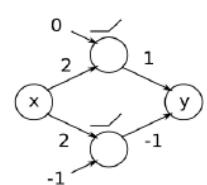
Shattered Gradients II

[Montufar'14, Balduzzi'17]

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

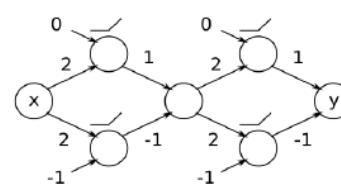
Example in $[0, 1]$:

depth 1

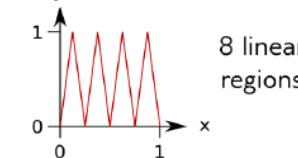
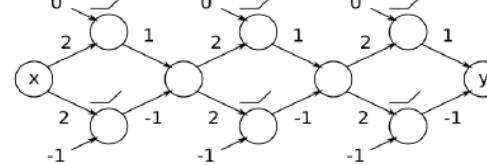


number of linear regions grows exponentially with depth

depth 2



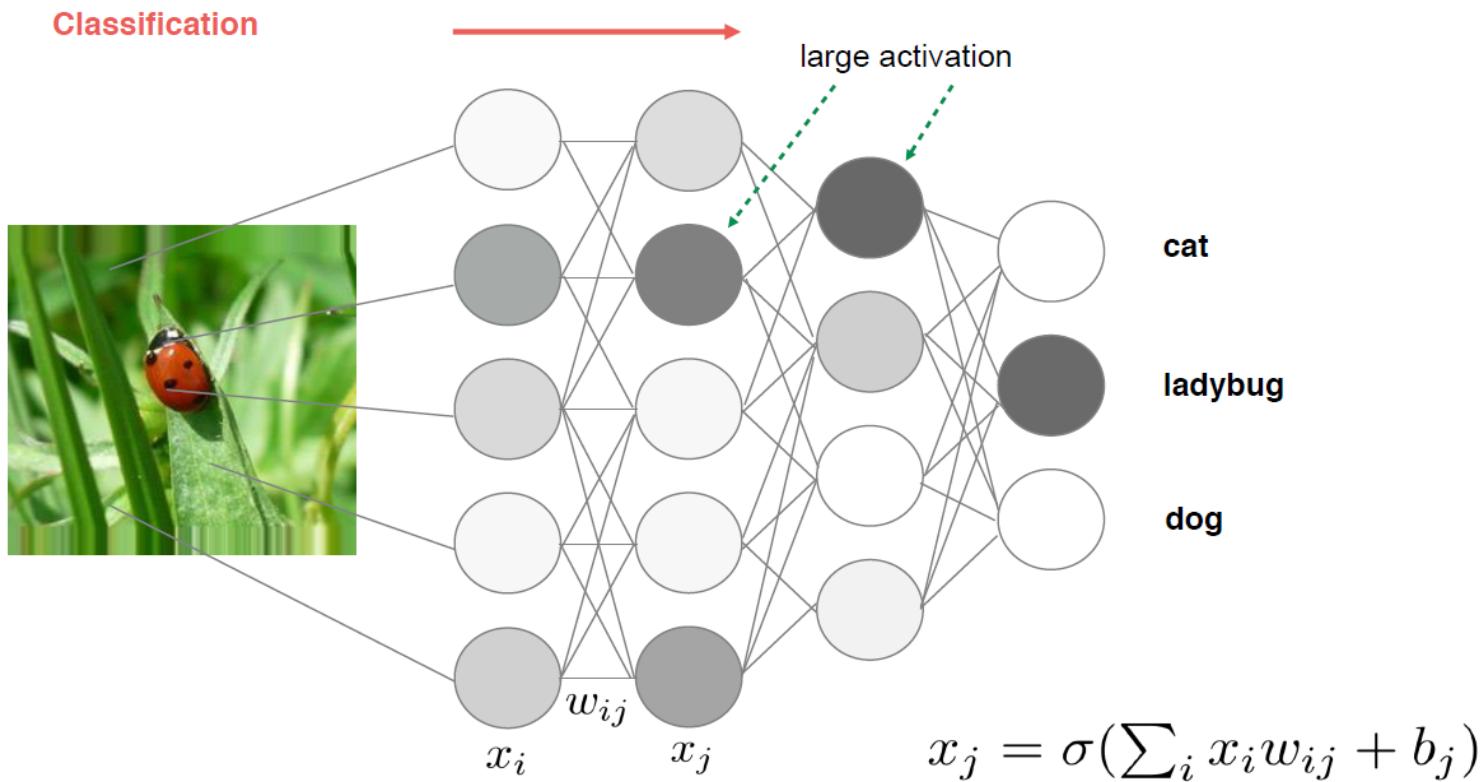
depth 3



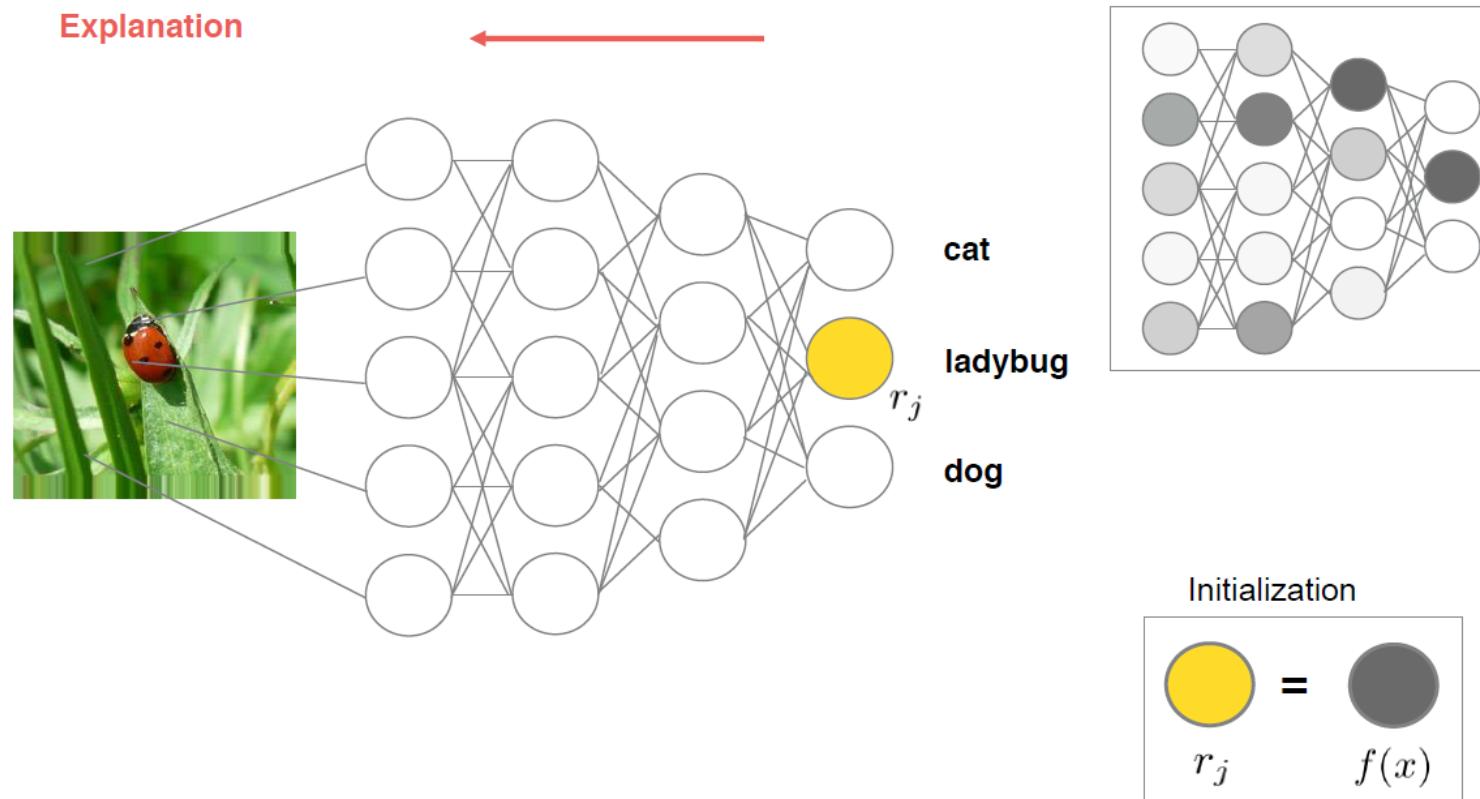
4 Explicando as predições de redes neurais profundas

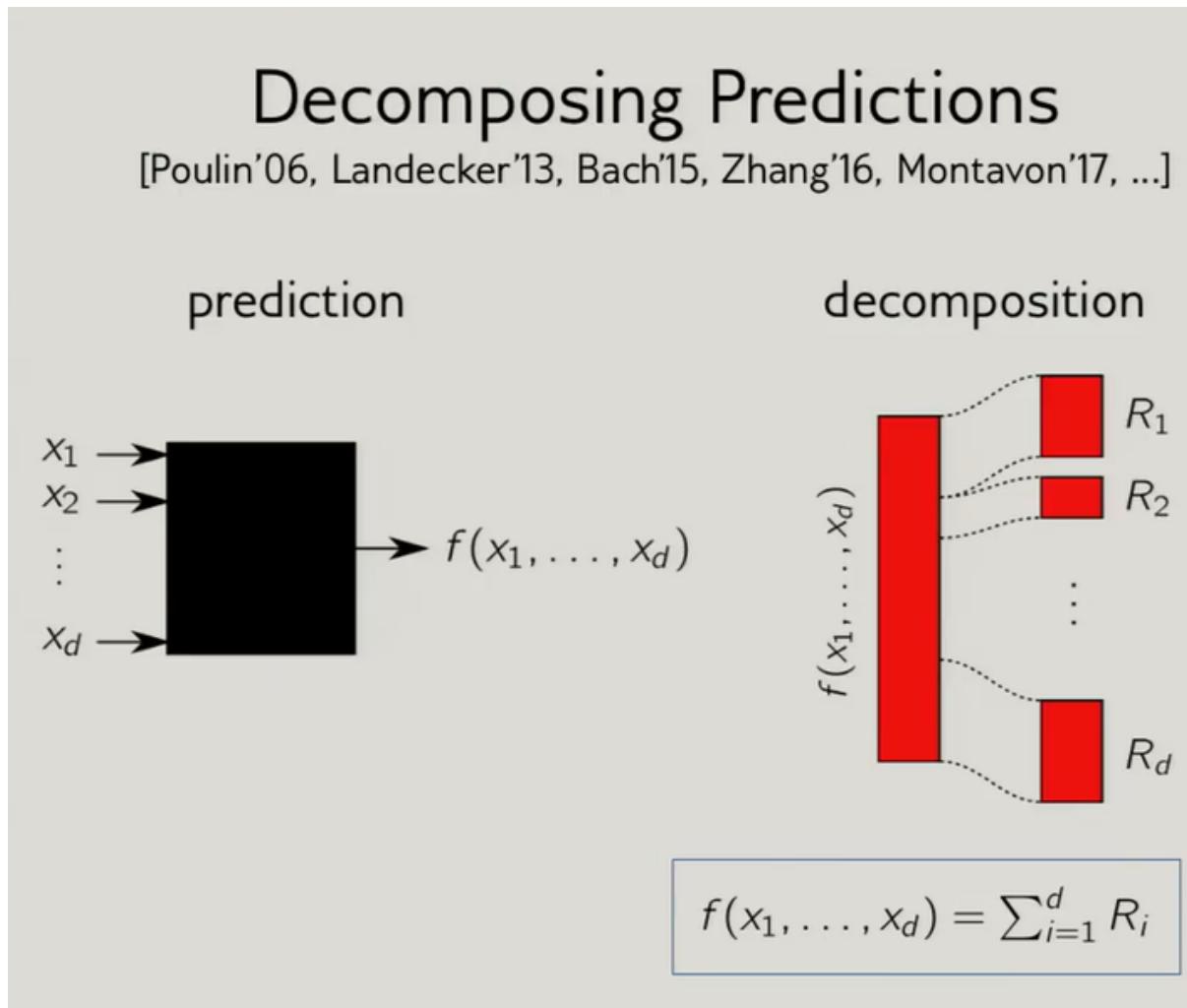
- Uma bem-sucedida iniciativa voltada para a interpretação do comportamento de redes neurais profundas é o *Heatmapping Project* (<http://www.heatmapping.org/>), que visa estudar técnicas específicas para decompor a predição em termos de contribuições de variáveis de entrada individuais, de tal modo 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 texto e fala, por exemplo.

Explaining Neural Network Predictions



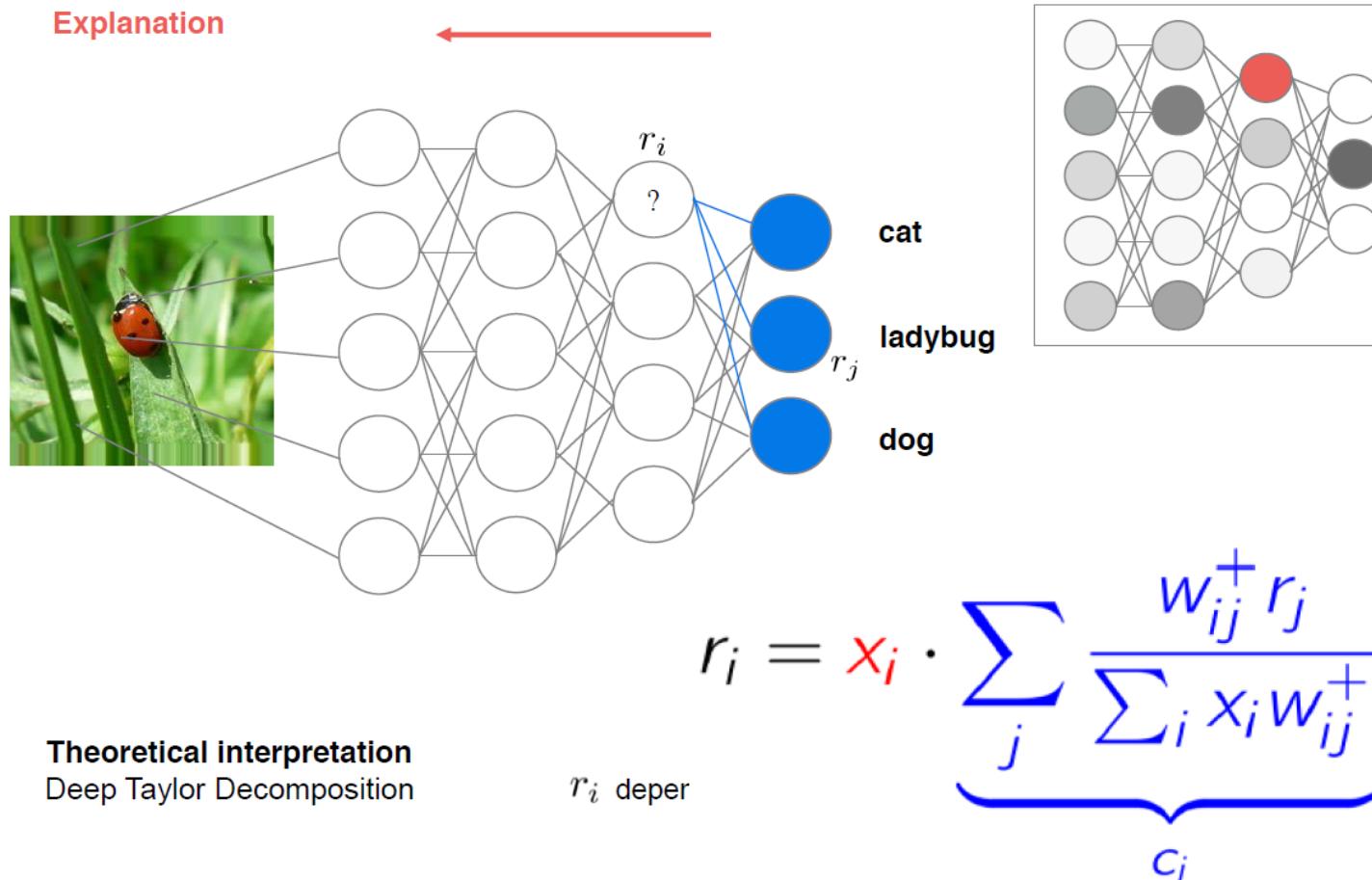
Explaining Neural Network Predictions



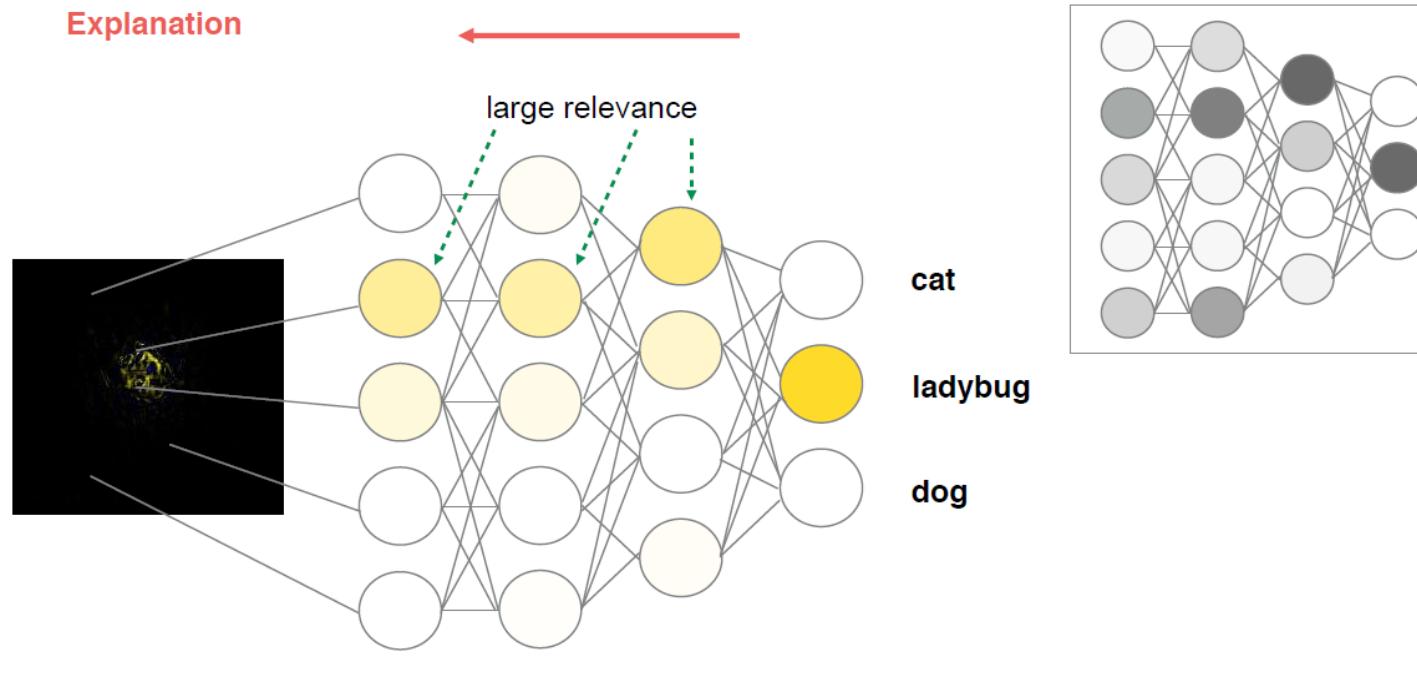


- Explora-se a propriedade de conservação de relevância, a partir da saída produzida.

Explaining Neural Network Predictions



Explaining Neural Network Predictions

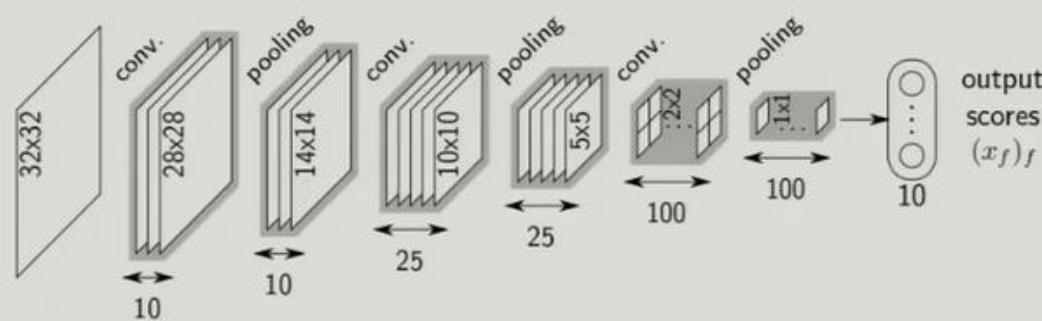
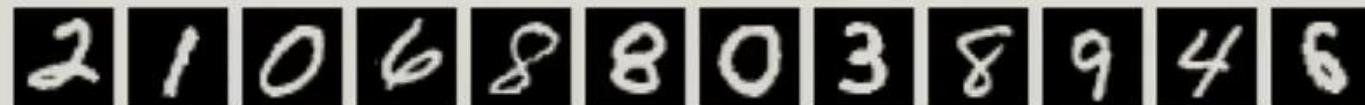


Relevance Conservation Property

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

Simple Taylor in Practice

Input Data:



Simple Taylor Decomposition:



$$R_i = \frac{\partial f}{\partial x_i} \cdot x_i$$

6/24

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

Simple Taylor vs. Relevance Propagation

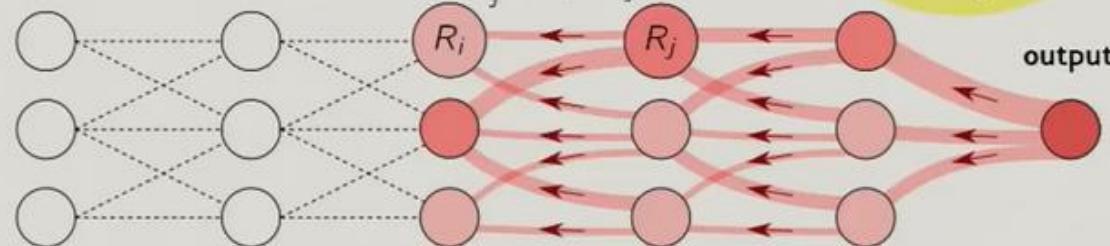
Simple Taylor Decomposition

$$R_i = \frac{\partial f}{\partial x_i} \cdot x_i$$

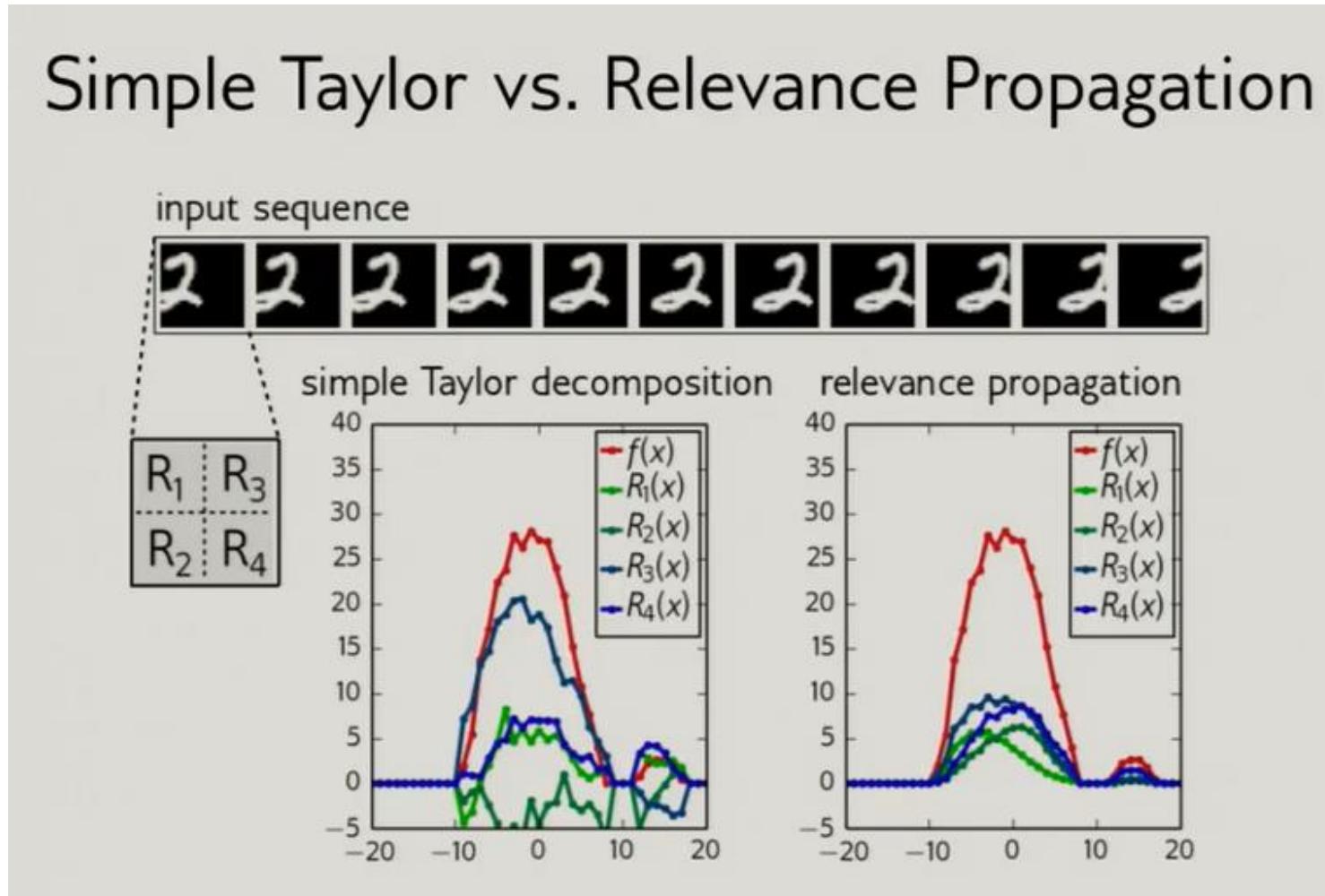
Relevance Propagation

[Landeker'13 | Bach'15 | Zhang'16 | Montavon'17]

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

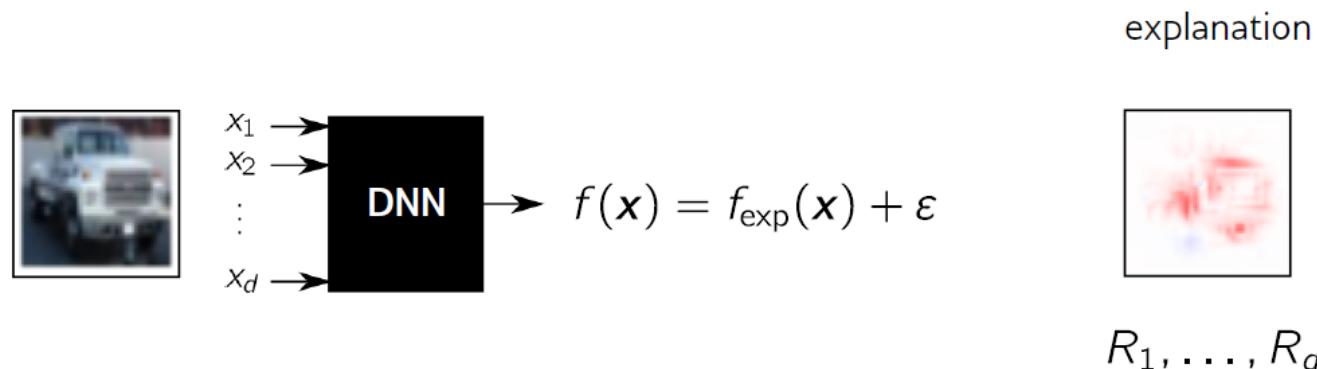


Simple Taylor vs. Relevance Propagation



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.

$$\sum_{p=1}^d R_p = f_{\text{exp}}(\mathbf{x})$$

Positivity: If the neural network is certain about its prediction, input features are either relevant (positive) or irrelevant (zero).

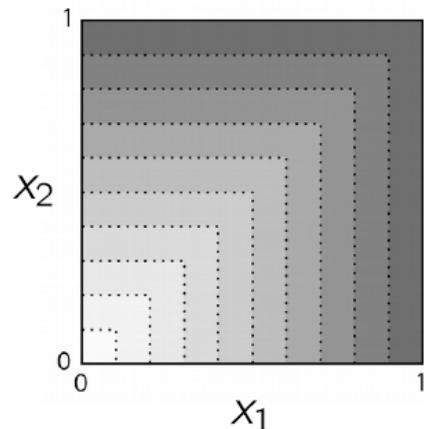
$$\forall_{p=1}^d : R_p \geq 0$$

Property 3: Continuity [Montavon'18]

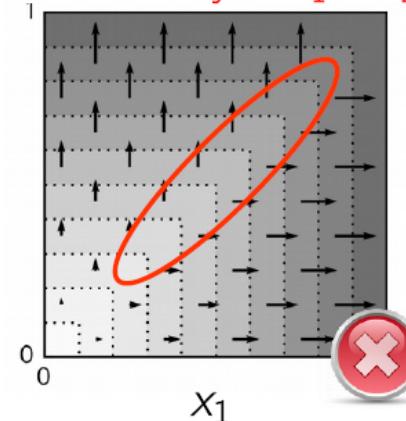
If two inputs are the almost the same, and the prediction is also almost the same, then the explanation should also be almost the same.

Example:

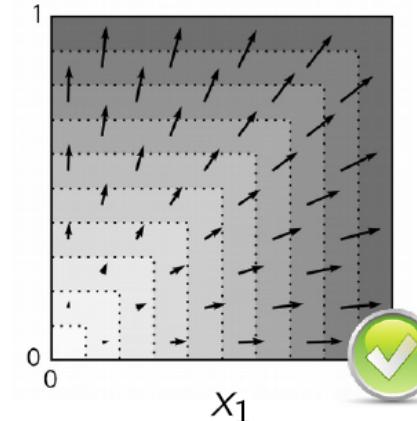
$$f(x) = \max(x_1, x_2)$$



Method 1
discontinuity at $x_1 = x_2$



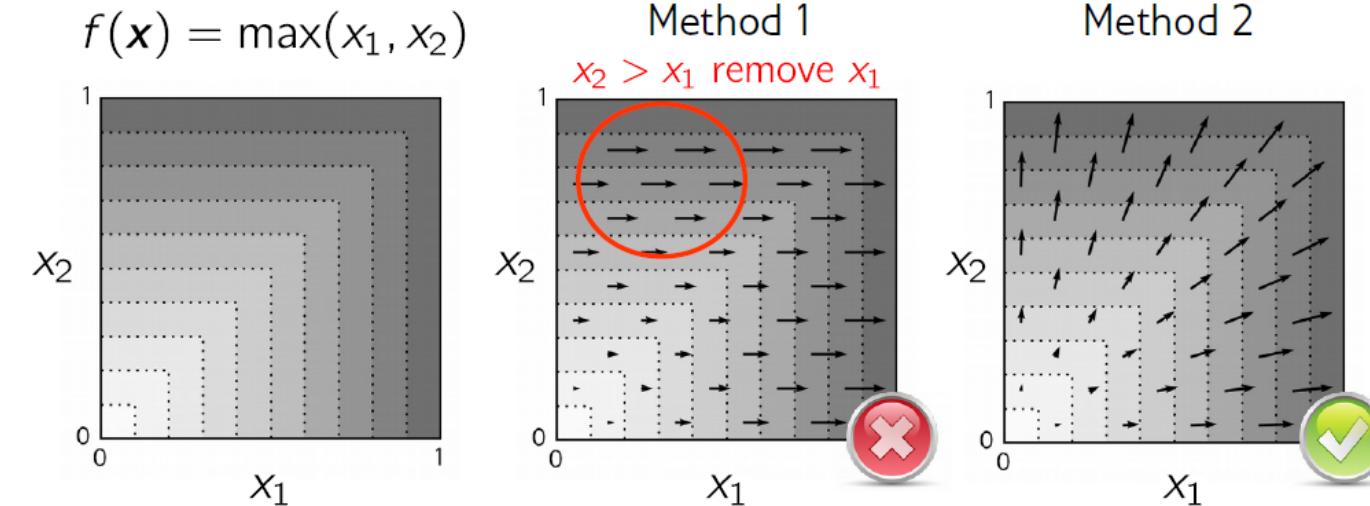
Method 2

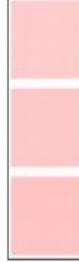


Property 4: Selectivity [Bach'15, Samek'17]

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

Example:



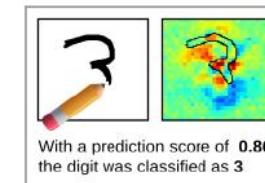
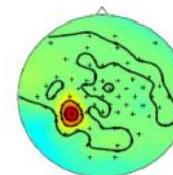
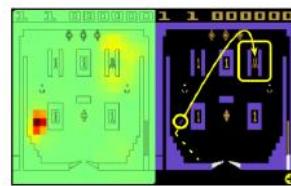
Explanation techniques	Uniform	$(\text{Gradient})^2$	$(\text{Guided BP})^2$	$\text{Gradient} \times \text{Input}$	$\text{Guided BP} \times \text{Input}$	$\text{LRP-}\alpha, \beta_0$...
Properties							
1. Conservation	✓			✓	✓	✓	
2. Positivity	✓	✓	✓		✓	✓	
3. Continuity	✓			✓	✓	✓	
4. Selectivity			✓	✓	✓	✓	
...							

Layer-Wise Relevance Propagation



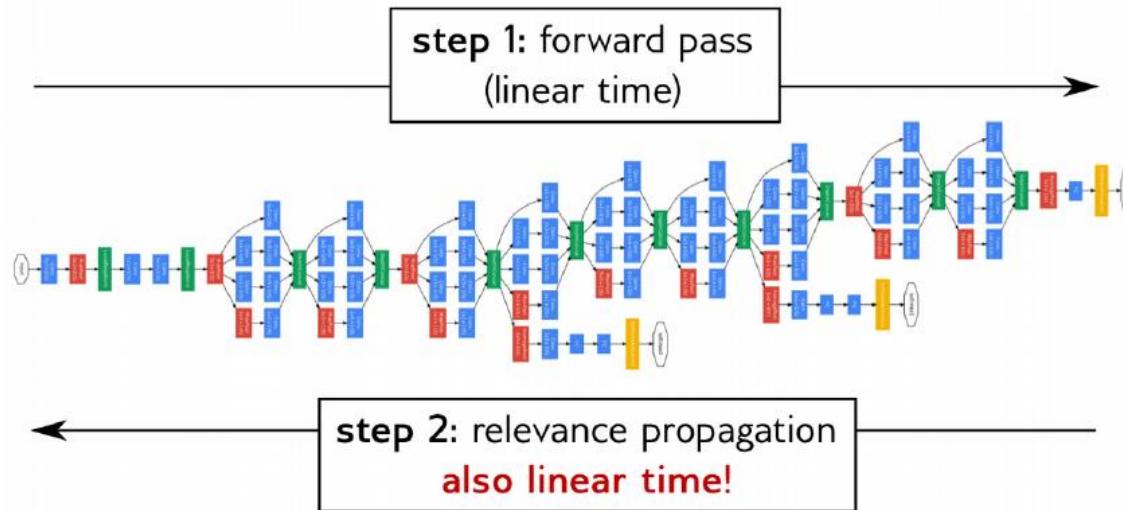
LRP Explanation Framework

e people are more prone to g
The mental part is usually
y is up or down, ie: the Shu
ointed towards Earth, so the
astronauts. About 50% of t
s, and NASA has done numerou



<http://www.heatmapping.org/>

How LRP Scales

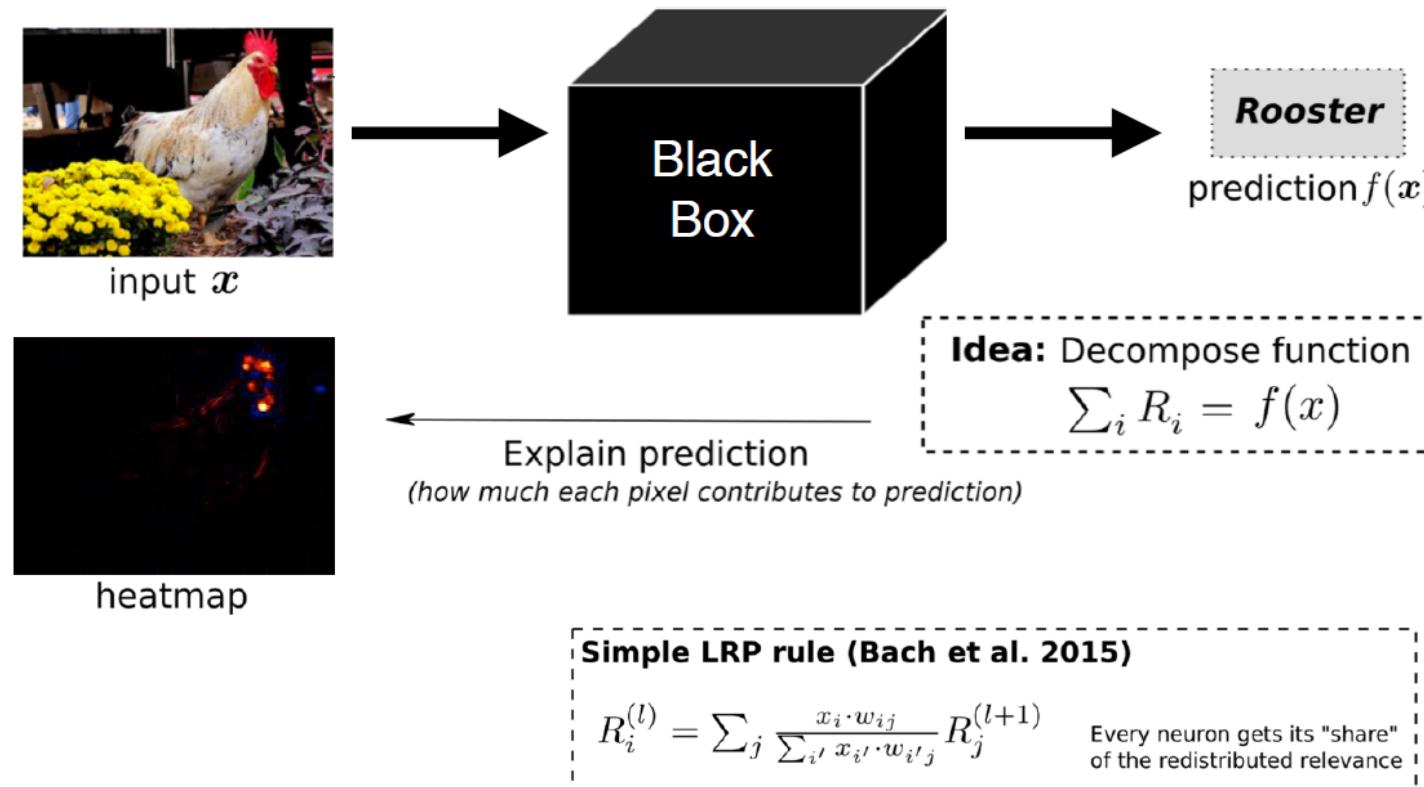


No need for much computing power. GoogleNet explanation for single image can be done on the CPU.

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



LRP applied to different Data

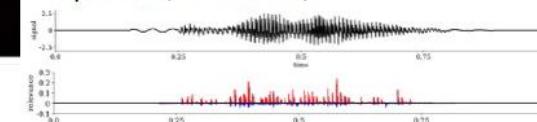
General Images (Bach' 15, Lapuschkin'16)



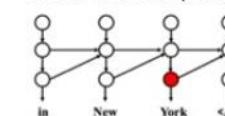
Text Analysis (Arras'16 &17)

do n't waste your money
neither funny nor susper

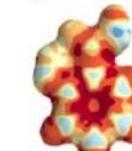
Speech (Becker'18)



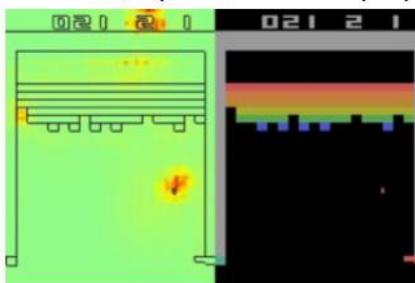
Translation (Ding'17)



Molecules (Schütt'17)



Games (Lapuschkin'18, in prep.)



VQA (Arras'18)

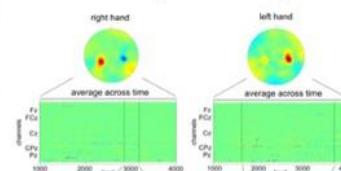
there is a metallic cube ; are
there any large cyan metallic
objects near it ?



Video (Anders'18)



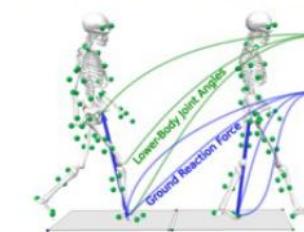
EEG (Sturm'16)



Morphing (Seibold'18)



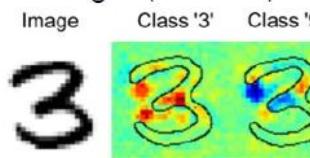
Gait Patterns (Horst'18, in prep.)



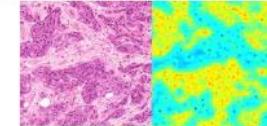
Faces (Arbabzadeh'16, Lapuschkin'17)



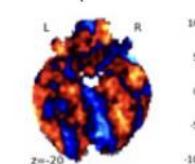
Digits (Bach' 15)



Histopathology (Binder'18)

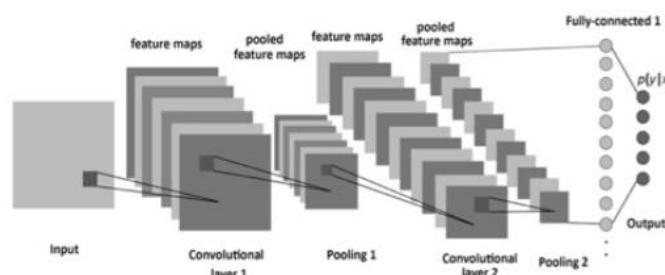


fMRI (Thomas'18)

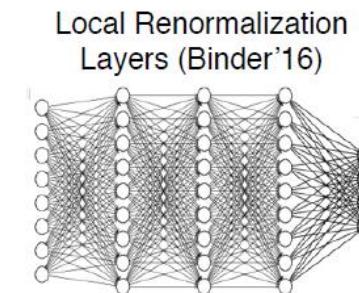


LRP applied to different Models

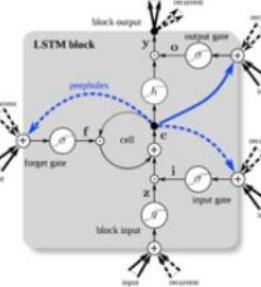
Convolutional NNs (Bach'15, Arras'17 ...)



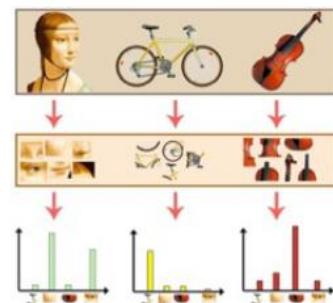
Local Renormalization Layers (Binder'16)



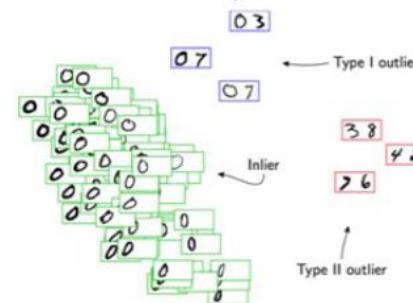
LSTM (Arras'17, Thomas'18)



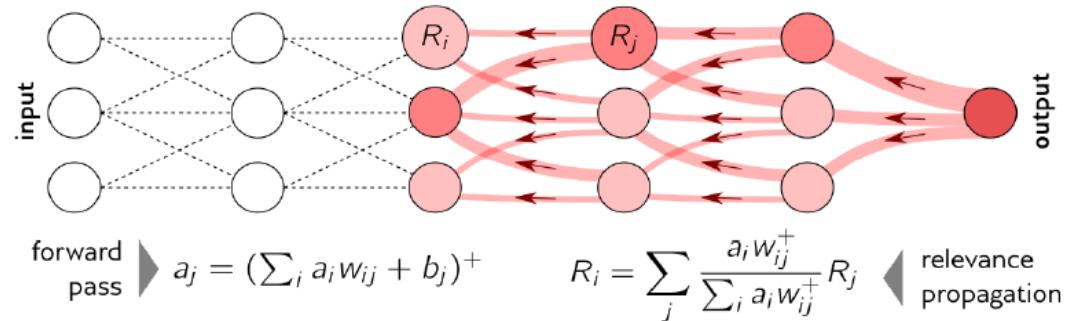
Bag-of-words / Fisher Vector models
(Bach'15, Arras'16, Lapuschkin'17, Binder'18)



One-class SVM (Kauffmann'18)

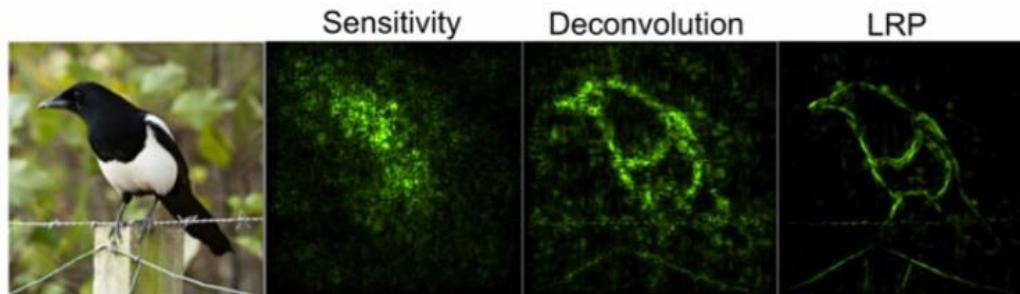


LRP works 4 all: deep models, LSTMs, kernel methods ...



5 Comparação entre diferentes técnicas

Compare Explanation Methods



Can we objectively measure which heatmap is best ?

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

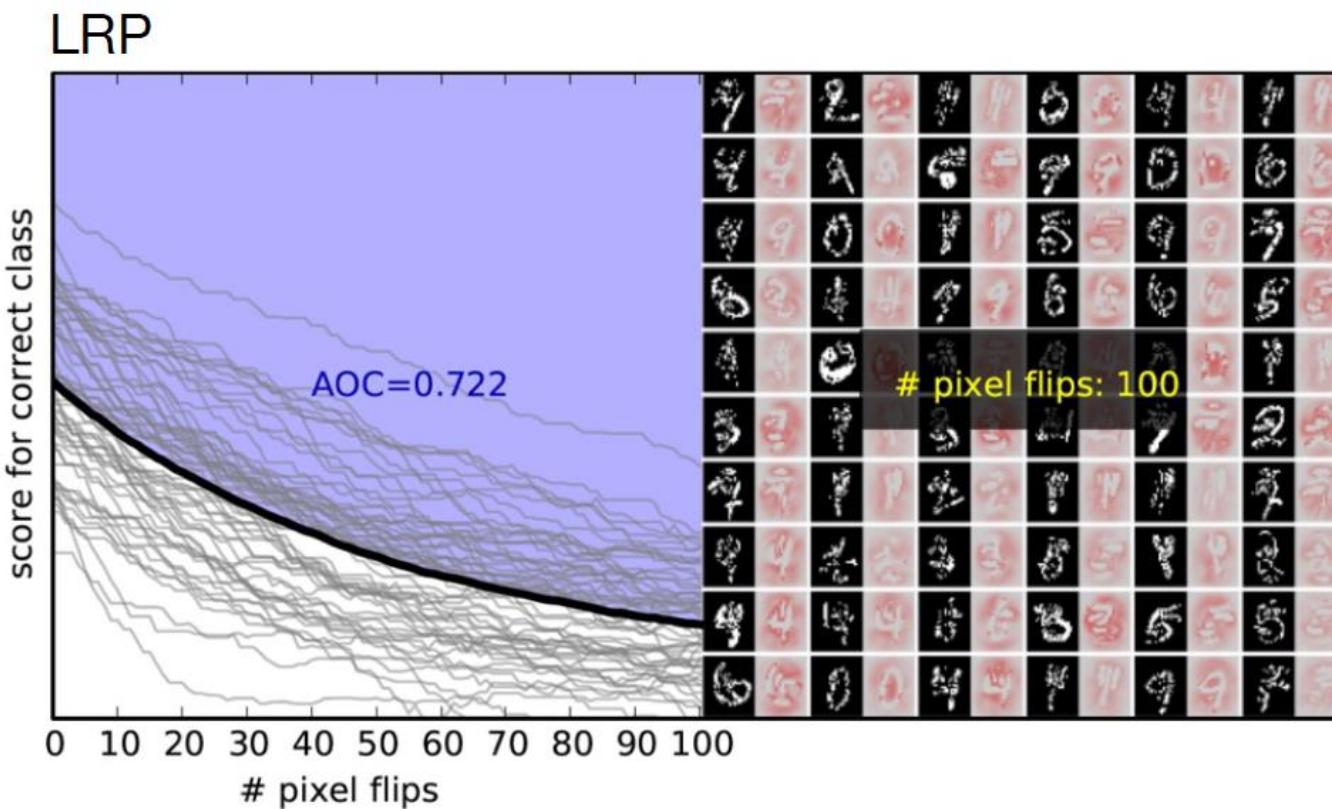
"If input features are deemed relevant, removing them should reduce evidence at the output of the network."

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

LRP:	0.722
Sensitivity:	0.691
Random:	0.523

LRP produces quantitatively better heatmaps
than sensitivity analysis and random.

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.

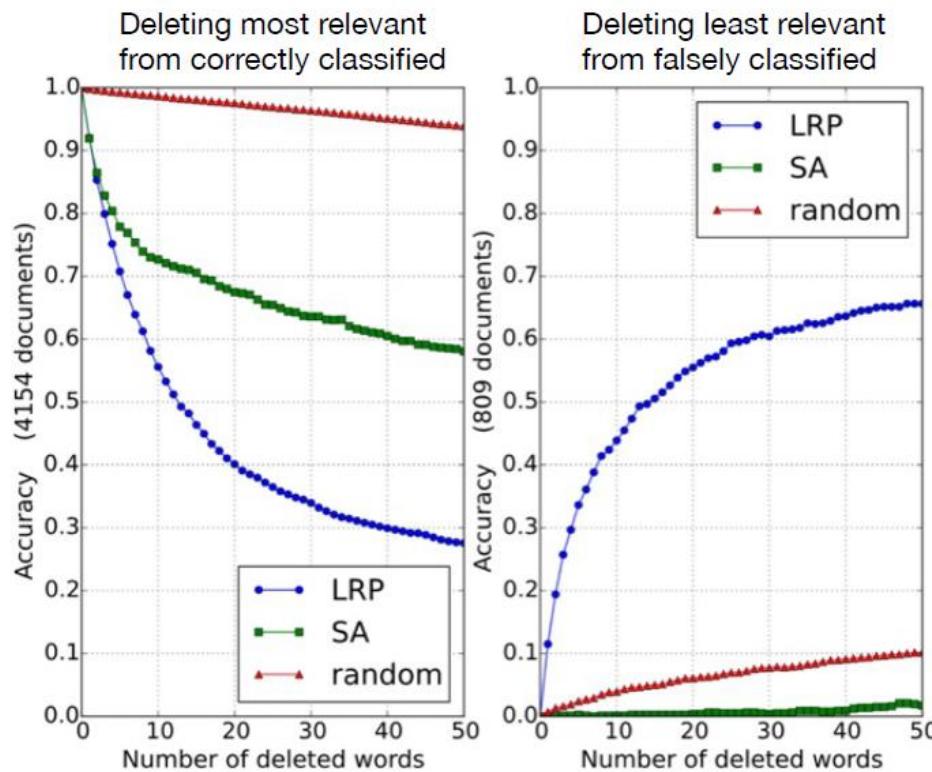
Yes, weightlessness does feel like falling. It may feel strange at first, but the body does adjust. The feeling is not too different from that of sky diving.

(4.1) sci.med
>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 occurrences down.

(Arras et al. 2017)

Compare Explanation Methods



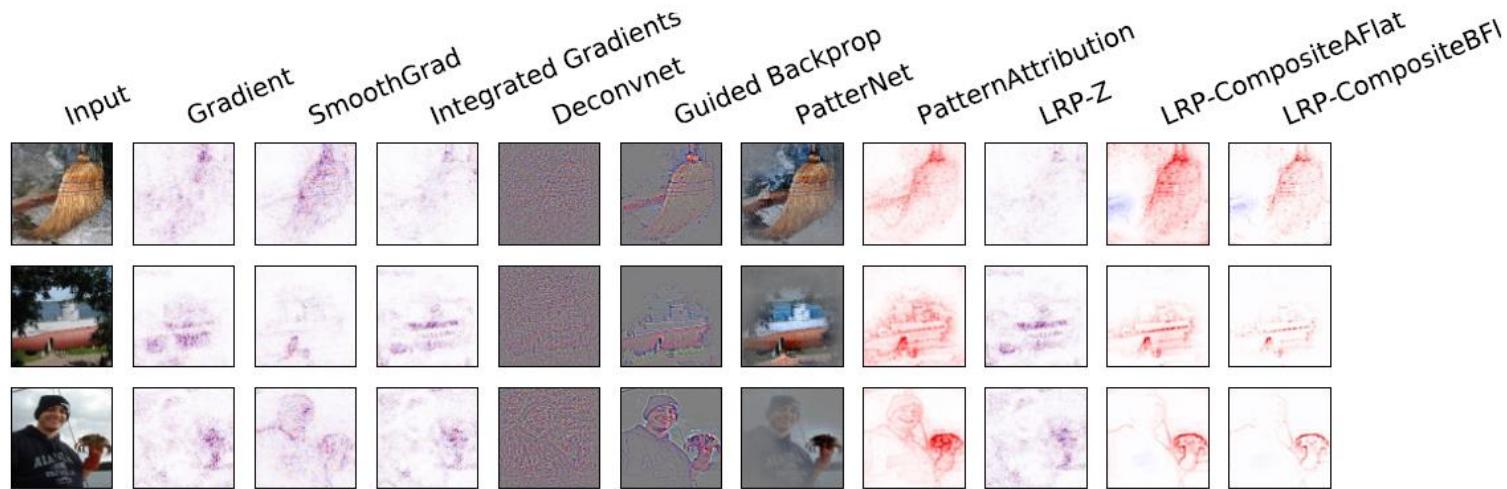
- word2vec / CNN model
- Conv → ReLU → 1-Max-Pool → FC
- trained on 20Newsgroup Dataset
- accuracy: 80.19%

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/investigate>

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

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.

sci.med (4.1)
>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 occurrences down.

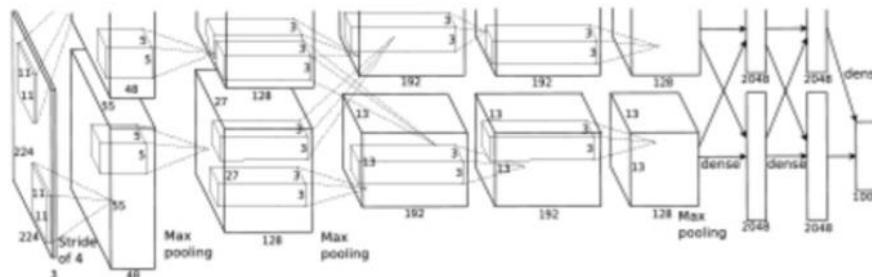
Yes, weightlessness does feel like falling. It may feel strange at first, but the body does adjust. The feeling is not too different from that of sky diving.

sci.med (-0.6)
>And what is the motion sickness
>that some astronauts occasionally experience?

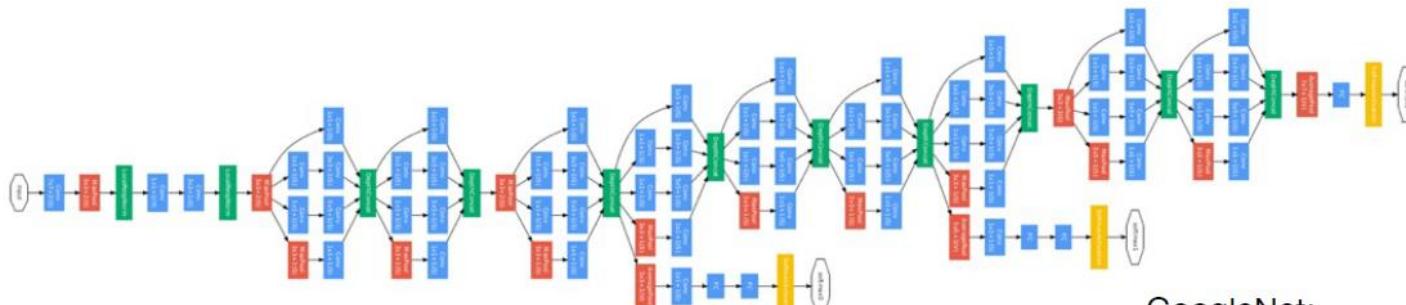
It is the body's reaction to a strange environment. It appears to be induced partly to physical discomfort and part to mental distress. Some people are more prone to it than others, like some people are more prone to get sick on a roller coaster ride than others. The mental part is usually induced by 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 occurrences down.

(Arras et al. 2016 & 2017)

Application: Compare Classifiers

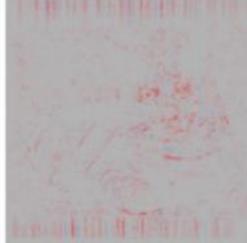
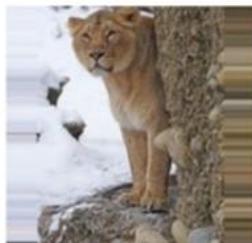
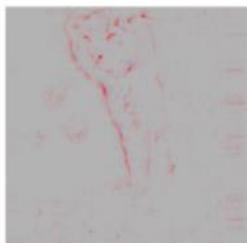


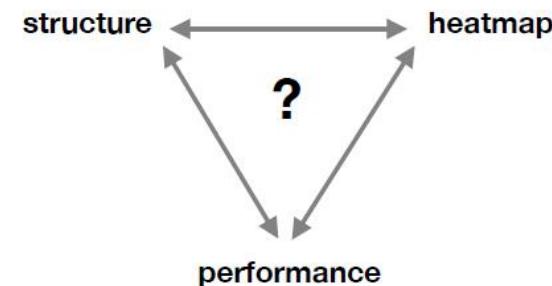
BVLC:
- 8 Layers
- ILSRCV: 16.4%



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

Application: Compare Classifiers

Image	BVLC CaffeNet	GoogleNet	
			<p>GoogleNet focuses on faces of animal. → suppresses background noise</p>
			<p>BVLC CaffeNet heatmaps are much more noisy.</p>
			<p><i>Is it related to the architecture ?</i> <i>Is it related to the performance ?</i></p>



(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

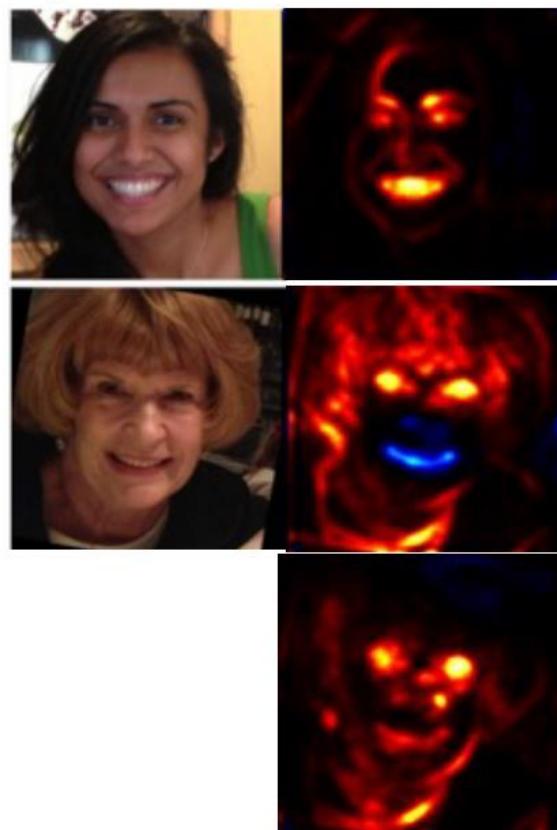


Strategy to solve the problem: Focus on chin / beard, eyes & hear,
but without pretraining the model overfits

(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

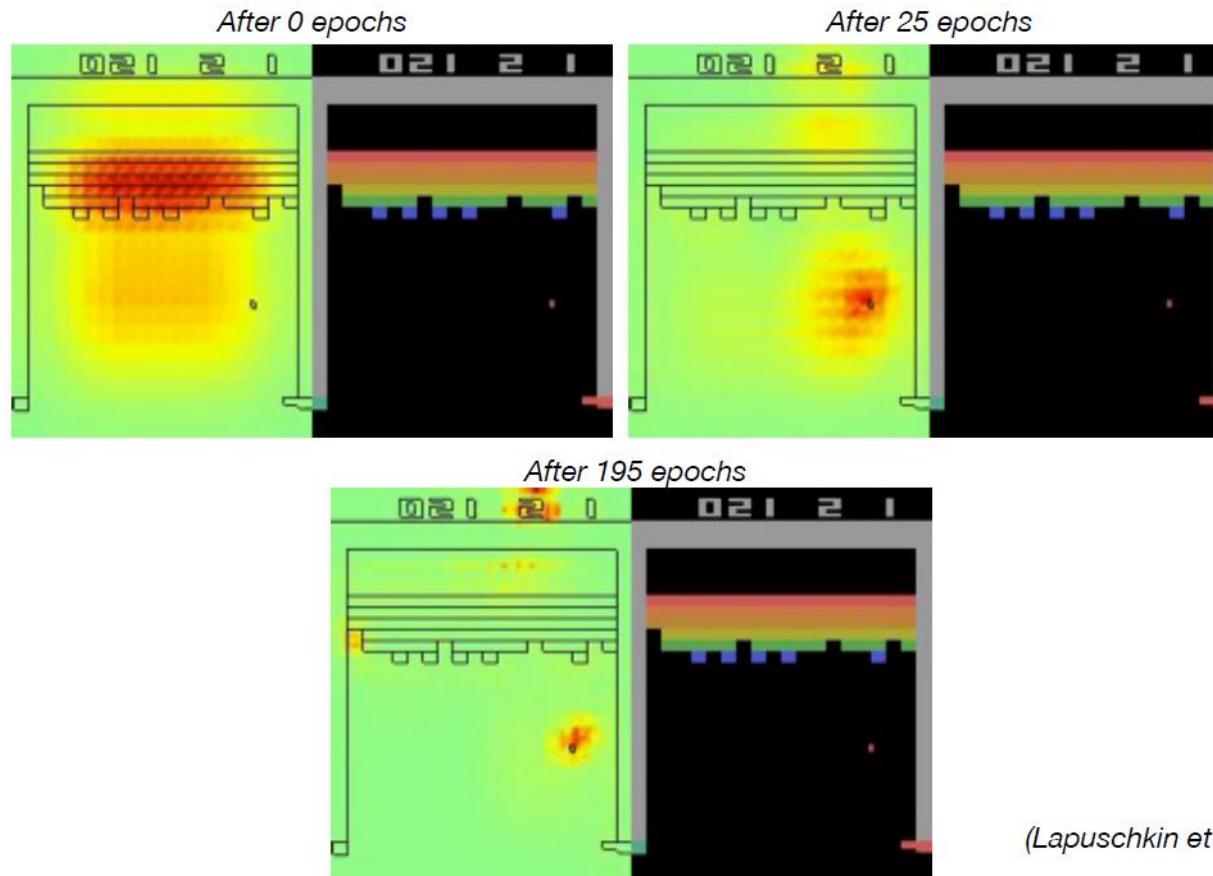
laughing speaks against 60+
(i.e., model learned that old
people do not laugh)

pretraining on
ImageNet

pretraining on
IMDB-WIKI

(Lapuschkin et al., 2017)

Application: Understand the model



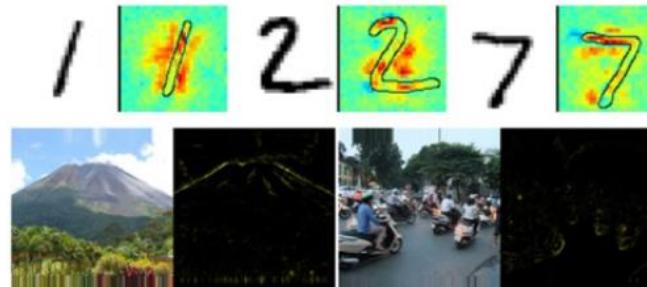
(Lapuschkin et al., in prep.)

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>

7 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.
- MONTAVON, G.; SAMEK, W.; MÜLLER, K.-R. “Methods for interpreting and understanding deep neural networks”, Digital Signal Processing, vol. 71, pp. 1-15, 2018.

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 Mikä, Gunnar Rätsch, Jason Weston, Bernhard Schölkopf, KR Müller (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

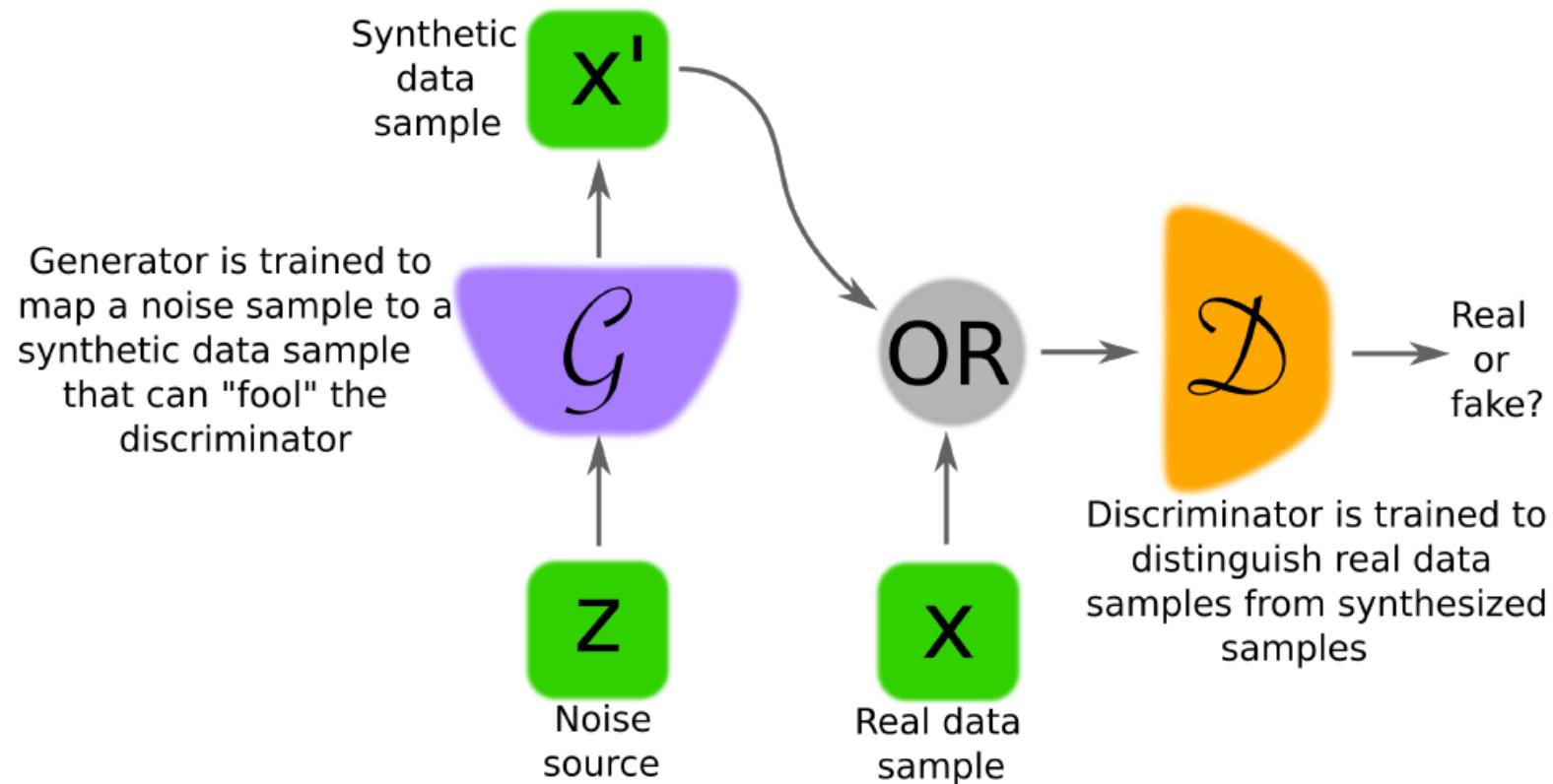
- Müller, K. R., Mika, S., Rätsch, G., Tsuda, K., & Schölkopf, B. (2001). An introduction to kernel-based learning algorithms. *Neural Networks, IEEE Transactions on*, 12(2), 181-201.
- Montavon, G., Braun, M. L., & Müller, K. R. (2011). Kernel analysis of deep networks. *The Journal of Machine Learning 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)

8 Generative Adversarial Networks

- Saímos agora do foco em interpretabilidade e entramos em uma nova estratégia bem-sucedida em *deep learning*. *Generative Adversarial Networks* (GAN) tem atraído muita atenção pelas novas possibilidades de aplicação.
- Como aplicações bem-sucedidas já é possível mencionar síntese de imagens, edição semântica de imagens, transferência de estilo e imagens de super-resolução.
- O texto sugerido para uma introdução a esta relevante área de aprendizado de máquina e de onde foi extraído o conteúdo a seguir é: Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; Bharath, A.A. “Generative Adversarial Networks: An Overview”, arXiv:1710.07035v1, 2017 (<https://arxiv.org/abs/1710.07035>).
- O foco maior está nos modelos gerativos, os quais aprendem a capturar a distribuição estatística exibida pelos dados de treinamento.



Generative Adversarial Nets

Ian J. Goodfellow, Jean Pouget-Abadie*, Mehdi Mirza, Bing Xu, David Warde-Farley,
Sherjil Ozair†, Aaron Courville, Yoshua Bengio†

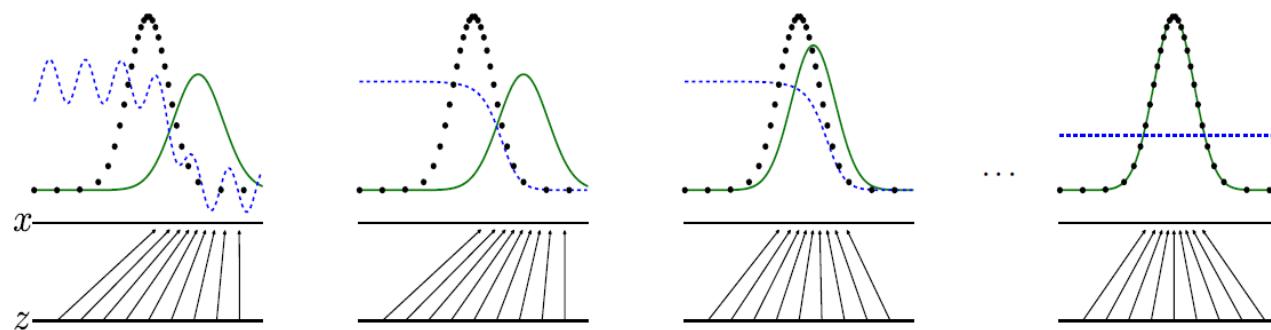
Département d'informatique et de recherche opérationnelle

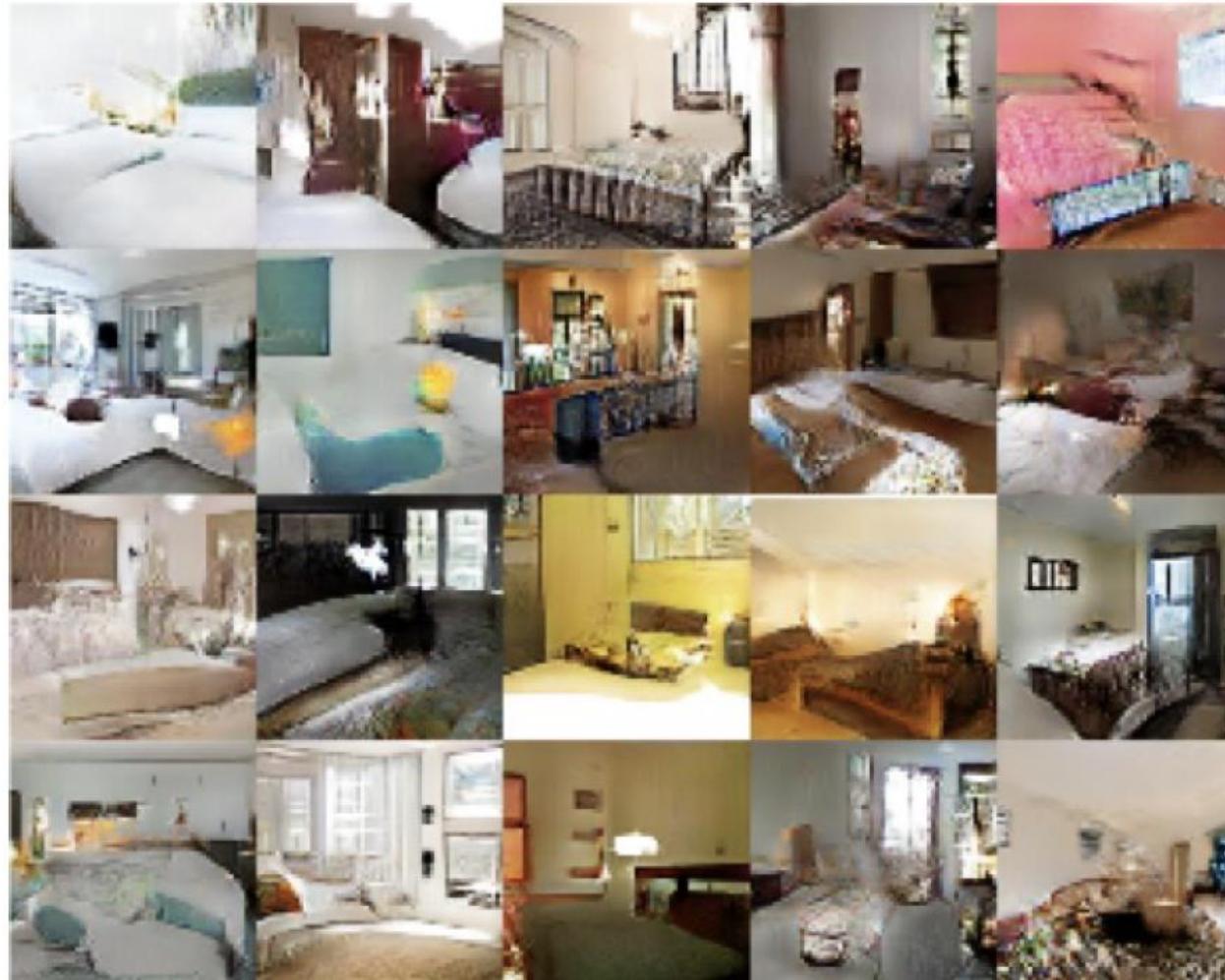
Université de Montréal
Montréal, QC H3C 3J7

v1 [stat.ML] 10 Jun 2014

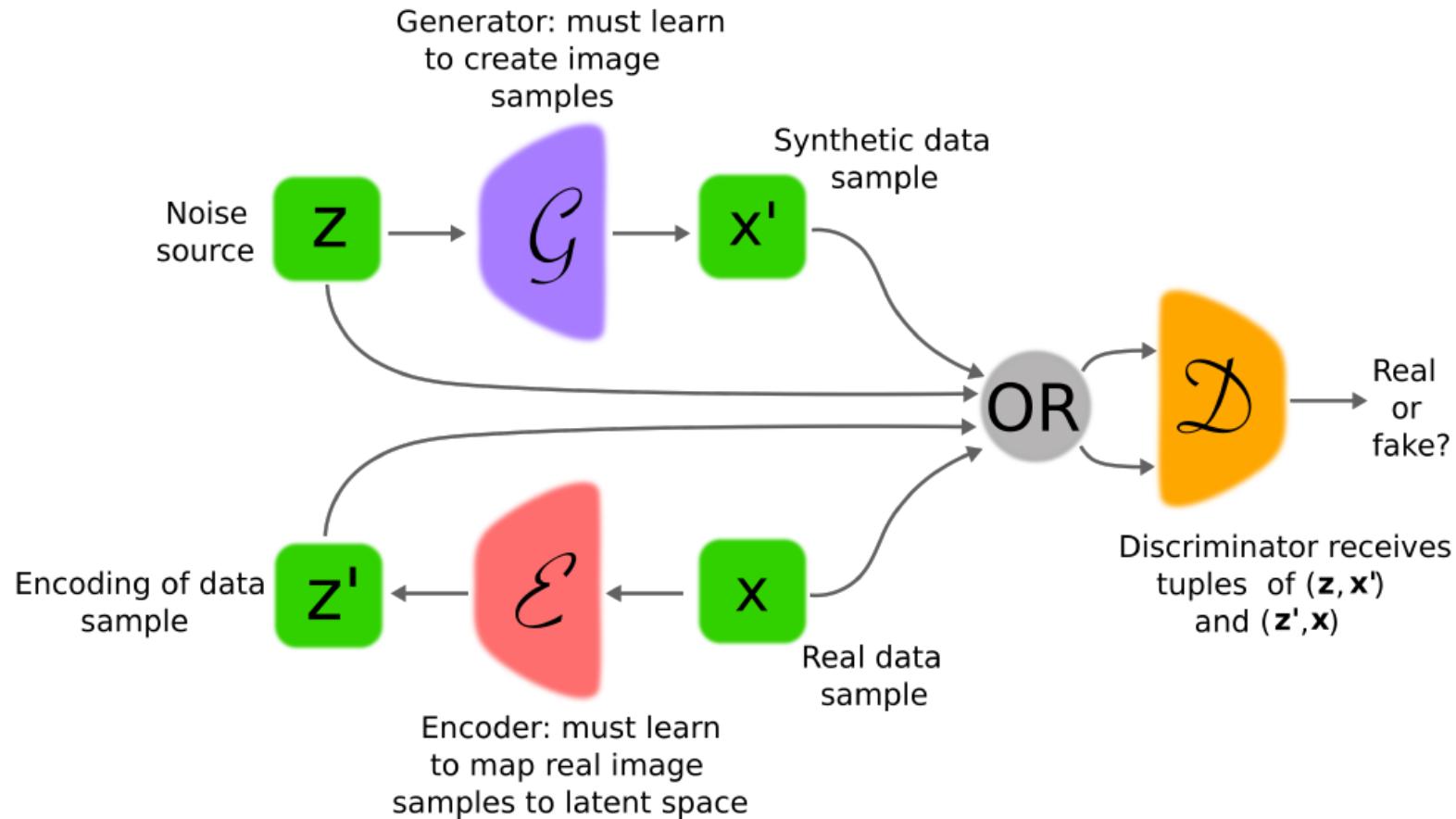
Abstract

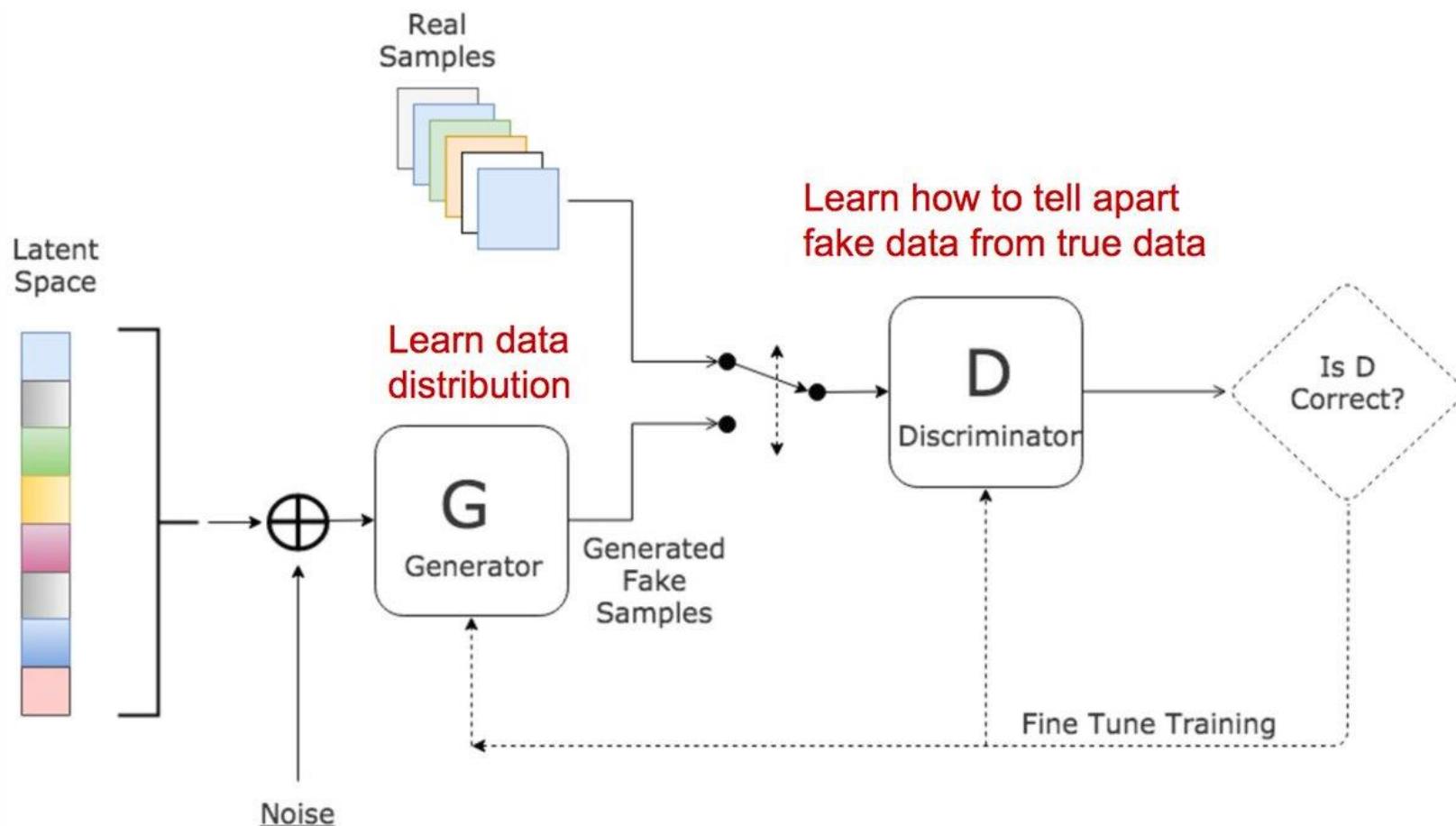
We propose a new framework for estimating generative models via an adversarial process, in which we simultaneously train two models: a generative model G that captures the data distribution, and a discriminative model D that estimates the probability that a sample came from the training data rather than G . The training procedure for G is to maximize the probability of D making a mistake. This framework corresponds to a minimax two-player game. In the space of arbitrary functions G and D , a unique solution exists, with G recovering the training data distribution and D equal to $\frac{1}{2}$ everywhere. In the case where G and D are defined by multilayer perceptrons, the entire system can be trained with backpropagation. There is no need for any Markov chains or unrolled approximate inference networks during either training or generation of samples. Experiments demonstrate the potential of the framework through qualitative and quantitative evaluation of the generated samples.



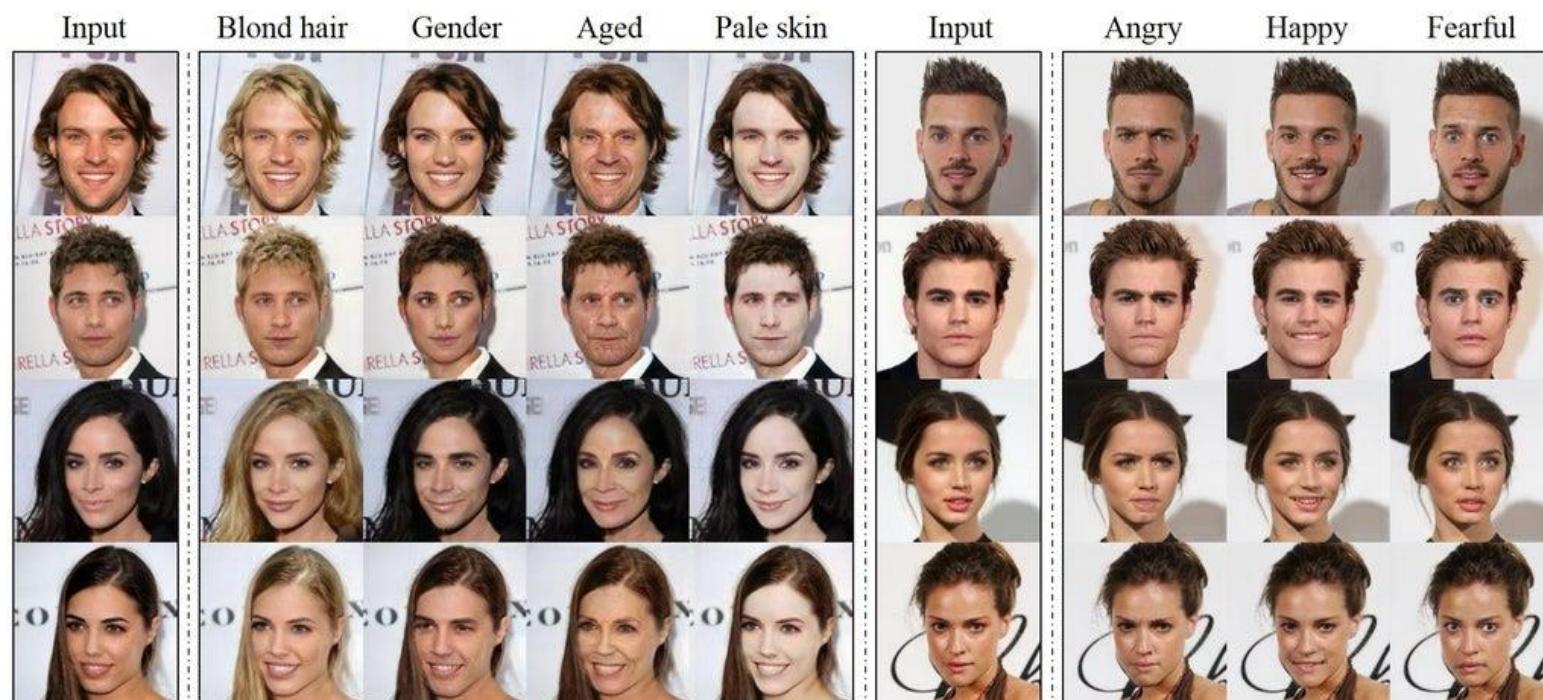


Goodfellow, I. Tutorial at NIPS'2016.

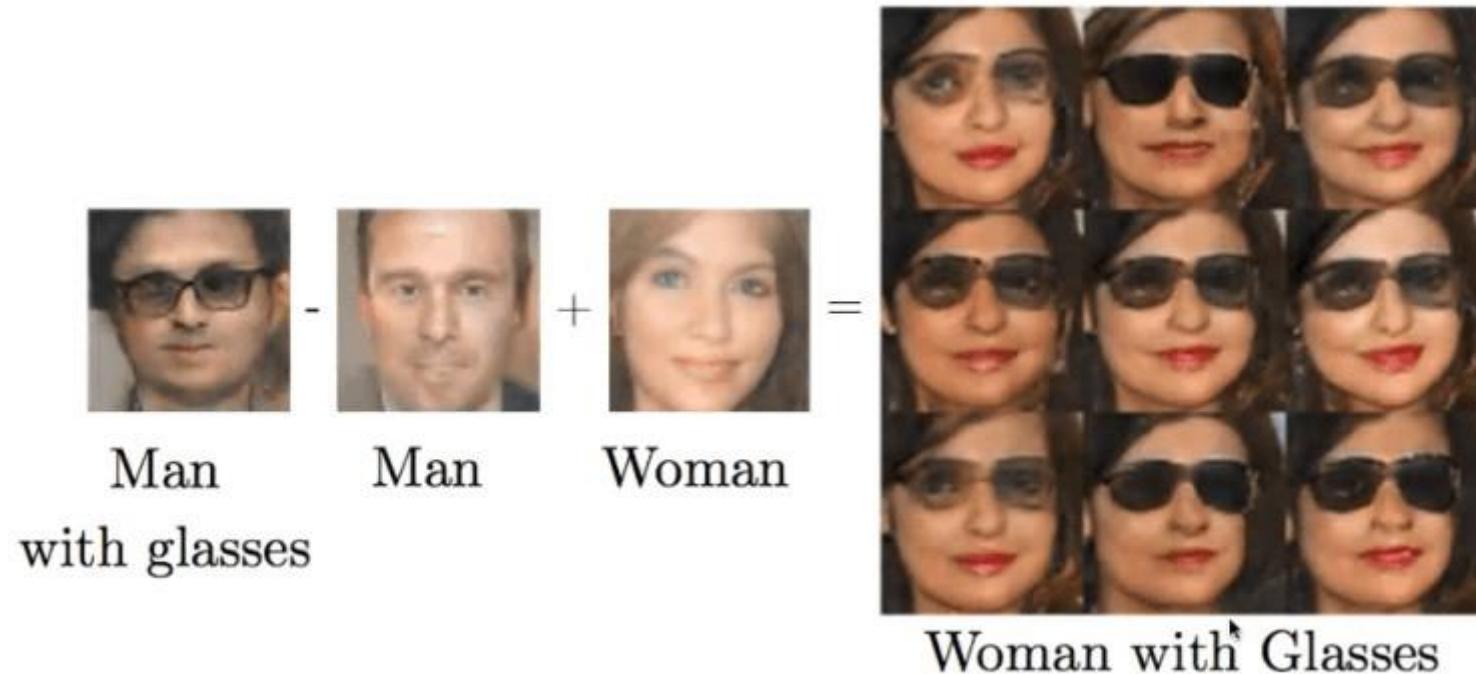




<https://lilianweng.github.io/lil-log/2017/08/20/from-GAN-to-WGAN.html>



https://www.reddit.com/r/MachineLearning/comments/7fro3g/r_stargan_unified_generative_adversarial_networks/

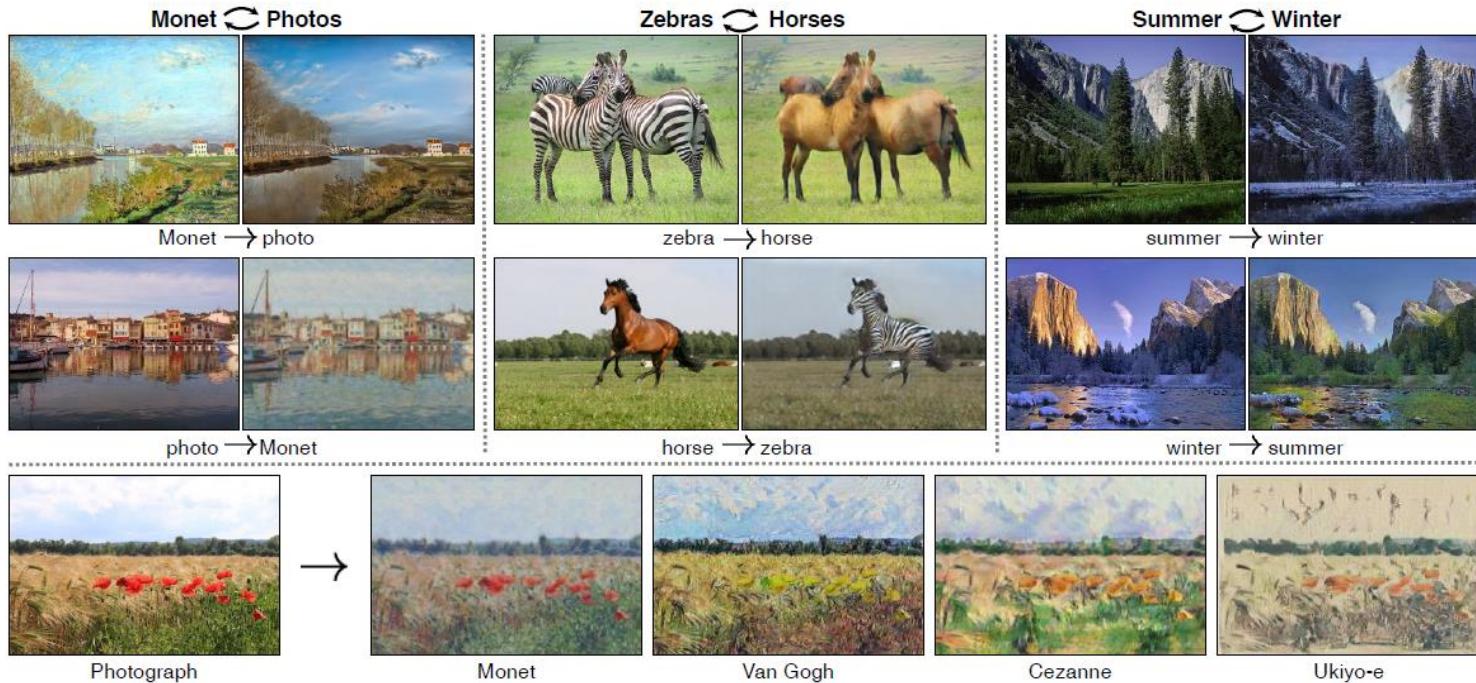


Goodfellow, I. Tutorial at NIPS'2016.

Associar com disentangled VAEs.

Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks

Jun-Yan Zhu* Taesung Park* Phillip Isola Alexei A. Efros
Berkeley AI Research (BAIR) laboratory, UC Berkeley



5 [cs.CV] 15 Nov 2018