Explaining and Interpreting Deep Neural Networks

Klaus-Robert Müller, Wojciech Samek, Gregoire Montavon, Sebastian Lapuschkin, Leila Arras et al.
Tutorial on Methods for Interpreting and Understanding Deep Neural Networks

Wojciech Samek (Fraunhofer HHI)
Grégoire Montavon (TU Berlin)
Klaus-Robert Müller (TU Berlin)
Acknowledgements

We thank our collaborators!

Alexander Binder (SUTD)

Sebastian Lapuschkin (Fraunhofer HHI)

Lecture notes will be online soon at:
http://www.heatmaping.org
Outline: tutorial

- Introduction to explaining and interpreting of nonlinear learners, Layer-wise Relevance Propagation (LRP) (9:30-10:45: L1 KRM)
- The LRP math: Deep Taylor, Evaluation (10:45-11:15: L2 WS)
- Exercise 1: playing with LRP toolbox and MNIST (11:30-12:00 WS)
- LRP Applications I - Neuroscience and Physics (13:00-13:50: L3 KRM)
- LRP Applications II – Image, Videos, Text, Age (15:00-15:45 L4 WS)
- Exercise 2: applying LRP to Text (16:10-17:00 WS)
Outline: L1

• general remarks, also on explaining and interpreting
• understanding single decisions of nonlinear learners
• Layer-wise Relevance Propagation (LRP)
• Applications in Neuroscience and Physics
Recent ML systems reach superhuman performance

AlphaGo beats Go human champ

Deep Net outperforms humans in image classification

DeepStack beats professional poker players

Autonomous search-and-rescue drones outperform humans

ML in the sciences

Computer out-plays humans in "doom"

IBM's Watson destroys humans in jeopardy

Deep Net beats human at recognizing traffic signs
From Data to Information

<table>
<thead>
<tr>
<th>Model</th>
<th>Interpretable (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlexNet</td>
<td>16.4%</td>
</tr>
<tr>
<td>Clarifai</td>
<td>11.1%</td>
</tr>
<tr>
<td>VGG</td>
<td>7.3%</td>
</tr>
<tr>
<td>GoogleNet</td>
<td>6.7%</td>
</tr>
<tr>
<td>ResNet</td>
<td>3.57%</td>
</tr>
</tbody>
</table>

Data → Information → Interpretable for human

Crucial in many applications (industry, sciences ...)

IMAGENET
Interpretable vs. powerful models?

Linear model

Non-linear model

Poor fit, but easily interpretable
“global explanation”

Can be very complex
“individual explanation”
Interpretable vs. powerful models?!

60 million parameters
650,000 neurons

We have techniques to interpret and explain such complex models!

Kernel machines
Interpretable vs. powerful models?

- Train best model $\rightarrow$ interpret it
- Train interpretable model
  
  suboptimal or biased due to assumptions (linearity, sparsity ...)

TU Berlin
Different dimensions of interpretability

- **Prediction**: “Explain why a certain pattern x has been classified in a certain way f(x).”

- **Model**: “What would a pattern belonging to a certain category typically look like according to the model.”

- **Data**: “Which dimensions of the data are most relevant for the task.”

---

Different dimensions of “interpretability”
Why interpretability?

1) Verify that classifier works as expected

Wrong decisions can be costly and dangerous

“Autonomous car crashes, because it wrongly recognizes ...”

“AI medical diagnosis system misclassifies patient’s disease ...”
Why interpretability?

2) Improve classifier

Standard ML

- data
- ML model
- predictions

**Generalization error**

Interpretable ML

- data
- ML model
- interpretability
- verified predictions

**Generalization error + human experience**

model/data improvement

human inspection
Why interpretability?

3) Learn from the learning machine

“It's not a human move. I've never seen a human play this move.” (Fan Hui)

Old promise:
“Learn about the human brain.”

[Image of Go game and brain diagram]
4) Interpretability in the sciences

Learn about the physical / biological / chemical mechanisms. (e.g. find genes linked to cancer, identify binding sites ... )
5) Compliance to legislation

European Union’s new General Data Protection Regulation $\rightarrow$ “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.”
Why interpretability?

**Interpretability as a gateway between ML and society**
- Make complex models acceptable for certain applications.
- Retain human decision in order to assign responsibility.
- “Right to explanation”

**Interpretability as powerful engineering tool**
- Optimize models / architectures
- Detect flaws / biases in the data
- Gain new insights about the problem
- Make sure that ML models behave “correctly”
Techniques of Interpretation

DNN transparency

Interpreting models
- Activation maximization
  - Berkes 2006
  - Erhan 2010
  - Simonyan 2013
  - Nguyen 2015/16
- Data generation
  - Hinton 2006
  - Goodfellow 2014
  - v. den Oord 2016
  - Nguyen 2016

Explaining decisions
- Sensitivity analysis
  - Khan 2001
  - Gevrey 2003
  - Baehrens 2010
  - Simonyan 2013
- Decomposition
  - Poulin 2006
  - Landecker 2013
  - Bach 2015
  - Montavon 2017

Focus on model  
Focus on data
Techniques of Interpretation

Interpreting models (ensemble)
- find prototypical example of a category
- find pattern maximizing activity of a neuron

Explaining decisions (individual)
- “why” does the model arrive at this particular prediction
- verify that model behaves as expected

better understand internal representation

crucial for many practical applications
Techniques of Interpretation

In medical context

- Population view (ensemble)
  - Which symptoms are most common for the disease
  - Which drugs are most helpful for patients
- Patient’s view (individual)
  - Which particular symptoms does the patient have
  - Which drugs does he need to take in order to recover

Both aspects can be important depending on who you are (FDA, doctor, patient).
Techniques of Interpretation

Interpreting models

- find prototypical example of a category
- find pattern maximizing activity of a neuron

\[
\max_{x \in X} p_{\theta}(\omega_c | x) + \lambda \Omega(x)
\]
Techniques of Interpretation

Explaining decisions
- “why” does the model arrive at a certain prediction
- verify that model behaves as expected

data → ML blackbox → decision

it’s a shark
Techniques of Interpretation

Explaining decisions

- “why” does the model arrive at a certain prediction
- verify that model behaves as expected

Data $x$ → ML blackbox $f(x)$ → Decision: it’s a shark

Explanation $R(x)$

- Sensitivity Analysis
- Layer-wise Relevance Propagation (LRP)
Techniques of Interpretation

Sensitivity Analysis
(Simonyan et al. 2014)

Explain prediction
(which pixels lead to decrease of prediction score when changed)

\[ \left| \frac{\partial}{\partial x_p} f(x) \right| \]
Techniques of Interpretation

Layer-wise Relevance Propagation (LRP) (Bach et al. 2015)

Explain prediction
(how much each pixel contributes to prediction)

Idea: Decompose function
\[ \sum_i R_i = f(x) \]

"every neuron gets its share of relevance depending on activation and strength of connection."

redistribute \( f(x) \)

Theoretical interpretation
Deep Taylor Decomposition (Montavon et al., 2017)
Interpreting models
Interpreting with class prototypes

Machine learning “blackbox” (e.g. deep neural network)

Input (e.g. RGB image)

\[ x \]

Output (e.g. probability that image has class car)

\[ p(\omega_c | x) \]

**Goal:** Understand what the model considers as a “typical car”.

**Activation maximization framework:**

\[
\max_{x \in \mathcal{X}} \ log p(\omega_c | x) - \lambda \|x\|^2
\]

class prototype
Examples of Class Prototypes

Activation maximization: \[ \max_{x \in \mathcal{X}} \log p(\omega_c | x) - \lambda \|x\|^2 \]

Symonian’13: Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps
Building more natural prototypes

\[ \max_{x \in X} \log p(\omega_c \mid x) - \lambda \|x\|^2 \]

- find input pattern that produces the strongest class activation

replace l2-norm by a data-dependent regularizer

\[ \max_{x \in X} \log p(\omega_c \mid x) + \log p(x) \]

- apply the Bayes theorem

\[ \max_{x \in X} \log p(x \mid \omega_c) \]

- find most likely input pattern for a given class

**Insight:** a good prototype must also depend on the data distribution.

Montavon, Samek, Müller arxiv 2017
Building Prototypes using a generator

 Activation maximization in code space:

$$\max_{z \in \mathcal{Z}} \log p(\omega_c \mid g(z)) - \lambda \|z\|^2$$
Building Prototypes using a generator

Nguyen'16: Synthesizing the preferred inputs for neurons in neural networks via deep generator networks.
Types of Interpretation

**Global Interpretation:**
Understanding how a lamp typically looks like.

*model’s prototypical lamp*

Nguyen’16: Synthesizing the preferred inputs for neurons in neural networks via deep generator networks

**Local Interpretation:**
Understanding why *this* image of a lamp contains a lamp.

*some image of a lamp*

Approaches to interpretability

**Ante-hoc interpretability:**
Choose a model that is readily interpretable and train it.

*Example:*

\[ f(x) = \sum_{i=1}^{d} g_i(x_i) \]

Is the model expressive enough to predict the data?

**Post-hoc interpretability:**
Choose a model that works well in practice, and develop a special technique to interpret it.

*Example:*

\[ f(x) = \text{DeepNet}(x) \]

How to determine the contribution each input variable?
Explaining models
Explaining Neural Network Predictions

Layer-wise relevance Propagation (LRP, Bach et al 15) first method to explain nonlinear classifiers - based on generic theory (related to Taylor decomposition – deep taylor decomposition M et al 16) - applicable to any NN with monotonous activation, BoW models, Fisher Vectors, SVMs etc.

**Explanation:** “Which pixels contribute how much to the classification” (Bach et al 2015)

\[ f(x) = \sum_p h_p \]

**Sensitivity / Saliency:** “Which pixels lead to increase/decrease of prediction score when changed” (what makes this image to be classified more/less as a car) (Baehrens et al 10, Simonyan et al 14)

\[ h_p = \left\| \frac{\partial}{\partial x_p} f(x) \right\|_\infty \]

Cf. Deconvolution: “Matching input pattern for the classified object in the image” (Zeiler & Fergus 2014), (relation to f(x) not specified) **Activation Maximization**

Each method solves a different problem!!!
Explaining Neural Network Predictions

Classification

large activation

$x_j = \sigma(\sum_i x_i w_{ij} + b_j)$
Explaining Neural Network Predictions

Initialization

$r_j \quad = \quad f(x)$
Explaining Neural Network Predictions

Theoretical interpretation
Deep Taylor Decomposition

\[ r_i = x_i \sum_j \frac{w_{ij} r_j}{\sum_i x_i w_{ij}} = x_i C_i \]

\( r_i \) depends on the activations and the weights
Explaining Neural Network Predictions

Relevance Conservation Property

$$\sum_p r_p = \ldots = \sum_i r_i = \sum_j r_j = \ldots = f(x)$$
Advantages of LRP over Sensitivity

1. Global explanations: What makes a car a car and not what makes a car less / more a car.

2. No discontinuities: small variations do not result in large changes of the relevance.
Advantages of LRP over both Sensitivity and Deconvolution

Image specific explanations: LRP takes into account the activations.

LRP provides different explanations for different input images.

For NNs without pooling layers Sensitivity and Deconvolution provides the **same** explanations for **different** samples.
Advantages of LRP over both Sensitivity and Deconvolution

Positive and Negative Evidence: LRP distinguishes between positive evidence, supporting the classification decision, and negative evidence, speaking against the prediction.

LRP indicates what speaks for class ‘3’ and speaks against class ‘9’.

The sign of Sensitivity and Deconvolution does not have this interpretation.

-> taking norm gives unsigned visualizations.
Advantages of LRP over both Sensitivity and Deconvolution

Aggregation of Relevance: LRP explanations are normalized (conservation of relevance). This allows to meaningfully aggregate relevance over datasets or regions in an image.
Explaining Neural Network Predictions

<table>
<thead>
<tr>
<th>Sensitivity</th>
<th>Deconvolution</th>
<th>LRP</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Sensitivity" /></td>
<td><img src="image2.png" alt="Deconvolution" /></td>
<td><img src="image3.png" alt="LRP" /></td>
</tr>
<tr>
<td><img src="image4.png" alt="Sensitivity" /></td>
<td><img src="image5.png" alt="Deconvolution" /></td>
<td><img src="image6.png" alt="LRP" /></td>
</tr>
<tr>
<td><img src="image7.png" alt="Sensitivity" /></td>
<td><img src="image8.png" alt="Deconvolution" /></td>
<td><img src="image9.png" alt="LRP" /></td>
</tr>
</tbody>
</table>
Application: understanding different DNN Architectures

GoogleNet focuses on the animal faces and only few pixels.

BVLC CaffeNet is less sparse.
Explaining Predictions Pixel-wise

Neural networks

Kernel methods
Perspectives
Is the Generalization Error all we need?
Application: Comparing Classifiers

Test error for various classes:

<table>
<thead>
<tr>
<th></th>
<th>aeroplane</th>
<th>bicycle</th>
<th>bird</th>
<th>boat</th>
<th>bottle</th>
<th>bus</th>
<th>car</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fisher</td>
<td>79.08%</td>
<td>66.44%</td>
<td>45.90%</td>
<td>70.88%</td>
<td>27.64%</td>
<td>69.67%</td>
<td>80.96%</td>
</tr>
<tr>
<td>DeepNet</td>
<td>88.08%</td>
<td>79.69%</td>
<td>80.77%</td>
<td>77.20%</td>
<td>35.48%</td>
<td>72.71%</td>
<td>86.30%</td>
</tr>
<tr>
<td></td>
<td>cat</td>
<td>chair</td>
<td>cow</td>
<td>diningtable</td>
<td>dog</td>
<td>horse</td>
<td>motorbike</td>
</tr>
<tr>
<td>Fisher</td>
<td>59.92%</td>
<td>51.92%</td>
<td>47.60%</td>
<td>58.06%</td>
<td>42.28%</td>
<td>80.45%</td>
<td>69.34%</td>
</tr>
<tr>
<td>DeepNet</td>
<td>81.10%</td>
<td>51.04%</td>
<td>61.10%</td>
<td>64.62%</td>
<td>76.17%</td>
<td>81.60%</td>
<td>79.33%</td>
</tr>
<tr>
<td></td>
<td>person</td>
<td>pottedplant</td>
<td>sheep</td>
<td>sofa</td>
<td>train</td>
<td>tvmonitor</td>
<td>mAP</td>
</tr>
<tr>
<td>Fisher</td>
<td>85.10%</td>
<td>28.62%</td>
<td>49.58%</td>
<td>49.31%</td>
<td>82.71%</td>
<td>54.33%</td>
<td>59.99%</td>
</tr>
<tr>
<td>DeepNet</td>
<td>92.43%</td>
<td>49.99%</td>
<td>74.04%</td>
<td>49.48%</td>
<td>87.07%</td>
<td>67.08%</td>
<td>72.12%</td>
</tr>
</tbody>
</table>

![Image](image.png) ![FV](fv.png) ![DNN](dnn.png)
Understanding Models is only possible if we explain
Conclusion

• explaining & interpreting nonlinear models is essential
• orthogonal to improving DNNs and other models
• need for opening the blackbox …
• understanding nonlinear models is essential for Sciences & AI
• new **theory**: LRP is based on deep taylor expansion NEXT LECTURE
Further Reading I


Further Reading II


Further Reading III


