Interpreting and Explaining Deep Models in Computer Vision

Fraunhofer HHI, Machine Learning Group
Wojciech Samek
“Superhuman” AI Systems

- Traffic Sign Recognition
- Skin cancer detection
- Lung cancer detection
- Poker
- Computer games
- Jeopardy
- OCR
Can we trust these black boxes?

- Huge volumes of data
- Computing power
- Deep Neural Network
- Solve task
- Information (implicit)
Can we trust these black boxes?

Is minimizing the error a guarantee for the model to work well in practice?
Can we trust these black boxes?

Is the way error is measured a satisfying specification of the problem?

$$\min_{f \in F} \int_{x,y} \| f(x) - y \|^2 dp(x, y)$$

Are we measuring the error on the true data distribution?
Can we trust these black boxes?

We need interpretability in order to:

- verify the system
- understand weaknesses
- consider legal aspects
- learn new things from data
Dimensions of Interpretability

Different dimensions of “interpretability”

Prediction

“How 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.”
Dimensions of Interpretability

model analysis

decision analysis

mechanistic understanding

functional understanding

$f : \mathbb{R}^d \rightarrow \mathbb{R}$
Approach 1: Class Prototypes

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

\[
\arg \max_x f(x) + \text{reg.}
\]
Dimensions of Interpretability

Find the input pattern that maximizes class probability.

→ Find the most likely input pattern for a given class.

\[ x_0 \rightarrow x^* \]
Images from **Nguyen et al. 2016.** “Synthesizing the preferred inputs for neurons in neural networks via deep generator networks”

**Observation:** Connecting AM to the data distribution leads to more realistic and more interpretable images.
Approach 2: Individual Explanations

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

heatmap = $LRP(x, f)$
Dimensions of Interpretability

- train interpretable model
  
  suboptimal or biased due to assumptions (linearity, sparsity ...) vs.

- train best model

  interpret it
Explain Predictions of Deep Neural Networks
Naive Approach: Sensitivity Analysis

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

$$R_i = \left( \frac{\partial f}{\partial x_i} \right)^2$$
**Naive Approach: Sensitivity Analysis**

Sensitivity analysis:

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

**Problem:** sensitivity analysis does not highlight cars

- Relevant pixels are found both on cars and on the background.
- Explains what *reduces/increases* the evidence for cars rather what *is* the evidence for cars.
Naive Approach: Sensitivity Analysis

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

Structure's view

Function's view (cartoon)

shallow  deep
Naive Approach: Sensitivity Analysis

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

Example in $[0,1]$:

- **Depth 1**: 2 linear regions
- **Depth 2**: 4 linear regions
- **Depth 3**: 8 linear regions

The number of linear regions grows exponentially with depth.
Better Approach: LRP

Layer-wise Relevance Propagation (LRP)
(Bach et al., PLOS ONE, 2015)
Better Approach: LRP

Classification

- cat
- rooster
- dog
Better Approach: LRP

What makes this image a “rooster image”? 

**Idea:** Redistribute the evidence for class rooster back to image space.
Better Approach: LRP

Theoretical interpretation
Deep Taylor Decomposition (Montavon et al., 2017)
not based on gradient!

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

\[
R_i^{(l)} = \sum_j \left( \alpha \cdot \frac{(x_i \cdot w_{ij})^+}{\sum_i (x_i \cdot w_{ij})^+} + \beta \cdot \frac{(x_i \cdot w_{ij})^-}{\sum_i (x_i \cdot w_{ij})^-} \right) R_j^{(l+1)}
\]

where $\alpha + \beta = 1$
Better Approach: LRP

Layer-wise relevance conservation

\[ \sum_i R_i = \ldots = \sum_i R^{(l)}_i = \sum_j R^{(l+1)}_j = \ldots = f(x) \]
Better Approach: LRP

Heatmap of prediction “3”

Heatmap of prediction “9”
Better Approach: LRP

Image

Sensitivity Analysis

LRP / Deep Taylor

Explains what influences prediction “cars”.

Slope decomposition

\[ \sum_i R_i = \| \nabla_x f \|^2 \]

Explains prediction “cars” as is.

Value decomposition

\[ \sum_i R_i = f(x) \]

More information (Montavon et al., 2017 & 2018)
Decomposing the Correct Quantity

\[ \sum_i R_i = \| \nabla_x f \|^2 \quad \rightarrow \quad \sum_i R_i = f(x) \]

**Candidate**: Taylor decomposition

\[
f(x) = f(\tilde{x}) + \sum_{i=1}^{d} \left. \frac{\partial f}{\partial x_i} \right|_{x=\tilde{x}} (x_i - \tilde{x}_i) + O(\varepsilon x^T x) + O(\varepsilon^2)
\]

Achievable for linear models and deep ReLU networks without biases, by choosing:

\[
\tilde{x} = \lim_{\varepsilon \to 0} \varepsilon \cdot x \approx 0.
\]
Why Simple Taylor doesn’t work?

Two Reasons:

1. Root point is hard to find or too far → includes too much information (incl. negative evidence)

2. Gradient shattering problem → gradient of deep nets has low informative value
**Idea:** Since neural network is composed of simple functions, we propose a *deep* Taylor decomposition.

**Each explanation step:**
- easy to find good root point
- no gradient shattering

(Montavon et al., 2017
Montavon et al. 2018)
Other Explanation Methods

Question: Which one to choose?

Understanding the Model

Deep Visualization (Yosinski et al., 2015)
Inverting CNNs (Dosovitskiy & Brox, 2015)
Synthesis of preferred inputs (Nguyen et al. 2016)
Network Dissection (Zhou et al. 2017)

Feature visualization (Erhan et al. 2009)
Inverting CNNs (Mahendran & Vedaldi, 2015)
RNN cell state analysis (Karpathy et al. 2015)
Axiomatic Approach to Interpretability
First Attempt: Distance to Ground Truth

\[ \| R - R^* \|^2 \]

input \( x \)

DNN

\[ f(x) \]

evidence for “truck”

explanation \( R \)

error

ground truth \( R^* \)

???
**Idea:** Evaluate the explanation technique **axiomatically**, i.e. it must pass a number of predefined “unit tests”.

[Sun’11, Bach’15, Montavon’17, Samek’17, Sundarajan’17, Kindermans’17, Montavon’18].
Axiomatic Approach to Interpretability

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.

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

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 almost the same, and the prediction is also almost the same, then the explanation should also be almost the same.

Example:

\[ f(x) = \max(x_1, x_2) \]
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.

LRP-$\alpha_1 \beta_0$

Sensitivity analysis
Axiomatic Approach to Interpretability

| Explanation techniques | Uniform | (Gradient)$^2$ | (Guided BP)$^2$ | Gradient $\times$ Input | Guided BP $\times$ Input | LRP-$\alpha_{1,8}$ | ...
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Properties</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Conservation</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>2. Positivity</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>3. Continuity</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>4. Selectivity</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Deep Taylor Decomposition
Can we express $R_k$ as a simple function of $(a_j)_j$?

Can we do a Taylor decomposition of $R_k((a_j)_j)$?
Deep Taylor Decomposition

Observe that \( R_k \approx a_k \cdot \text{const.} \)

Move to the lower-layer
**Proposition:** Relevance at each layer is a product of the activation and an approximately constant term.

\[ R_j = a_j \times c_j \]
Deep Taylor Decomposition

1. Build the Relevance Neuron

\[ R_j = a_j c_j \]
\[ = \max(0, \sum_i a_i w_{ij}) \cdot c_j \]
\[ = \max(0, \sum_i a_i w'_{ij}) \]

\[ w'_{ij} = w_{ij} c_j \]
Deep Taylor Decomposition

2. Expand the Relevance Neuron

\[ R_j((a_i)_i) = R_j((\tilde{a}_i)_i) + \sum_i \left. \frac{\partial R_j}{\partial a_i} \right|_{(\tilde{a}_i)_i} \cdot (a_i - \tilde{a}_i) + \varepsilon \]
Deep Taylor Decomposition

3

Decompose Relevance

Taylor expansion at root point:

\[ R_j(a) = R_j(\tilde{a}^{(j)}) + \sum_i \frac{\partial R_j}{\partial a_i} \bigg|_{\tilde{a}^{(j)}} \cdot (a_i - \tilde{a}_i^{(j)}) + \varepsilon \]

Relevance can now be backward propagated
Deep Taylor Decomposition

\[
R_{i \leftarrow j} = \frac{(a_i - \tilde{a}_i^{(j)})w_{ij}}{\sum_i(a_i - \tilde{a}_i^{(j)})w_{ij}} R_j
\]  
(Deep Taylor generic)

### Choice of root point

<table>
<thead>
<tr>
<th>Choice</th>
<th>( \tilde{a}^{(j)} = a - t \cdot w_j )</th>
<th>( \tilde{a}^{(j)} \in D )</th>
<th>( |a - \tilde{a}^{(j)}| )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. nearest root</td>
<td>( \tilde{a}^{(j)} = a - t \cdot w_j )</td>
<td>( \checkmark )</td>
<td></td>
</tr>
<tr>
<td>2. rescaled activation</td>
<td>( \tilde{a}^{(j)} = a - t \cdot a )</td>
<td>( \checkmark )</td>
<td></td>
</tr>
<tr>
<td>3. rescaled excitations</td>
<td>( \tilde{a}^{(j)} = a - t \cdot a \odot 1_{w_j &gt; 0} )</td>
<td>( \checkmark )</td>
<td>( \checkmark )</td>
</tr>
</tbody>
</table>

\[
R_{i \leftarrow j} = \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+} R_j
\]  
(LRP-\( \alpha_1 \beta_0 \))
### Deep Taylor Decomposition

<table>
<thead>
<tr>
<th>Input domain</th>
<th>Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>ReLU activations ((a_j \geq 0))</td>
<td>(R_j = \sum_k \frac{a_j w_{jk}^+}{\sum_j a_j w_{jk}^+} R_k)</td>
</tr>
<tr>
<td>Pixel intensities ((x_i \in [l_i, h_i], l_i \leq 0 \leq h_i))</td>
<td>(R_i = \sum_j \frac{x_i w_{ij} - l_i w_{ij}^+ - h_i w_{ij}^-}{\sum_i x_i w_{ij} - l_i w_{ij}^+ - h_i w_{ij}^-} R_j)</td>
</tr>
<tr>
<td>Real values ((x_i \in \mathbb{R}))</td>
<td>(R_i = \sum_j \frac{w_{ij}^2}{\sum_i w_{ij}^2} R_j)</td>
</tr>
</tbody>
</table>

Deep Taylor LRP rules [Montavon’17]

More refined rules can also be constructed to match the input data distribution [Kindermans’17]
Deep Taylor Decomposition

4 Pooling relevance over all outgoing neurons

\[ \sum_j \]

\[ R_i \rightarrow R_j \]
Deep Taylor Decomposition

The LRP-$\alpha_1\beta_0$ rule

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

can be seen as

a deep Taylor decomposition (DTD)

[Montavon’17]

which then yields

domain- and layer-specific rules
Applications
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)

Morphing (Seibold’18)

Games (Lapuschkin’18, in prep.)

VQA (Arras’18)

Video (Anders’18)

EEG (Sturm’16)

fMRI (Thomas’18)

Faces (Lapuschkin’17)

Digits (Bach’15)

Histopathology (Binder’18)

Gait Patterns (Horst’18, in prep.)
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)
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.

## Application: Compare 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 DeepNet</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>Fisher 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>Fisher DeepNet</td>
<td>59.92%</td>
<td>51.92%</td>
<td>47.60%</td>
<td>58.06%</td>
<td>42.28%</td>
<td>80.55%</td>
<td>81.00%</td>
</tr>
<tr>
<td>Fisher DeepNet</td>
<td>81.10%</td>
<td>51.04%</td>
<td>61.10%</td>
<td>64.62%</td>
<td>76.17%</td>
<td>69.34%</td>
<td>79.33%</td>
</tr>
<tr>
<td>Fisher DeepNet</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>Fisher 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 of a person riding a horse]

Same performance —> same strategy?  

(Lapuschkin et al. 2016)
Application: Compare Classifiers

‘horse’ images in PASCAL VOC 2007
Application: Compare Classifiers

GoogleNet focuses on faces of animal. → suppresses background noise

BVLC CaffeNet heatmaps are much more noisy.

(Binder et al. 2016)
Application: Measure Context Use

how important
is context ?

classifier

importance
of context

= relevance outside bbox

relevance inside bbox

how important
is context ?
Application: Measure Context Use

(Lapuschkin et al., 2016)
Application: Measure Context Use

Context use anti-correlated with performance.

BVLC CaffeNet

GoogleNet

VGG CNN S

Context use

BVLC CaffeNet

GoogleNet

VGG CNN S

(Lapuschkin et al. 2016)
Application: Face analysis

Gender classification

with pretraining

without pretraining

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

(Lapuschnik 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: Sentiment analysis

How to handle multiplicative interactions?

\[ z_j = z_g \cdot z_s \]

\[ R_g = 0 \quad R_s = R_j \]

gate neuron indirectly affect relevance distribution in forward pass

Negative sentiment

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

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

neither funny nor suspenseful nor particularly well-drawn.

Model understands negation!

(Arras et al., 2017)
Brain-Computer Interfacing

How brain works subject-dependent —> individual explanations

Application: EEG Analysis

Movement Imagination → Preprocessing → DNN → Movement Decoding

Feedback → LRP → explain

(Sturm et al. 2016)
Application: EEG Analysis

With LRP we can analyze what made a trial being misclassified.

(Sturm et al. 2016)
Application: fMRI Analysis

(Thomas et al. 2018)
Application: Understand the model

(frame 1) frame 4 frame 7 frame 10 frame 13 frame 16

(Anders et al., 2018)
**Observation:** Explanations focus on the bordering of the video, as if it wants to watch more of it.
Idea: Play video in fast forward (without retraining) and then the classification accuracy improves.
Application: Understand the model

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

(Becker et al., 2018)
**Application: Understand the model**

<table>
<thead>
<tr>
<th>Question</th>
<th>LRP</th>
</tr>
</thead>
<tbody>
<tr>
<td>there is a metallic cube; are any cyan</td>
<td>there is a metallic cube; are any cyan</td>
</tr>
<tr>
<td>large cyan metallic objects behind it?</td>
<td>cyan metallic objects behind it?</td>
</tr>
</tbody>
</table>

Model understands the question and correctly identifies the object of interest

(Arras et al., 2018)
Application: Understand the model

Sensitivity Analysis

LRP

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

LRP shows that that model tracks the ball

(Lapuschkin et al., in prep.)
Application: Understand the model

(Lapuschkin et al., in prep.)
Application: Understand the model

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

(Lapuschkin et al., in prep.)
Take Home Messages
Sensitivity analysis is not the question that you would like to ask!
Take Home Messages

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

Our LRP method is robust to this.
Take Home Messages

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

Desirable properties of an explanation
- positivity
- conservation
- selectivity
- continuity

“Tricks of the trade”

Underlying theory for consistency
\[ R_i = \sum_j \frac{\partial R}{\partial a_j} (a_i - \tilde{a}_i^{(U)}) \]

Deep Taylor Decomposition

LRP Explanation Framework

(small, tutorials, demos, insights, applications)

Fraunhofer
Helnrich Hertz Institute
Take Home Messages

- LRP ≠ Gradient × Input

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

High flexibility: Different LRP variants, free parameters

**Good news**: No need to reimplement LRP, check our software at [www.heatmapping.org](http://www.heatmapping.org).
Take Home Messages

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

[Samek et al. IEEE TNNLS 2017]
Take Home Messages

Explanation helps to improve models

Explaining ML, Now What?
Take Home Messages

Explanation helps to find flaws in models
Take Home Messages

Getting **new** Insights in the Sciences

**Example:** Understanding physical systems at the quantum level.

\[
\hat{H}\psi = E\psi
\]

- time-independent Schrödinger Equation
- Hamiltonian
- energy

equation describing
general physical systems

- molecular structure
- atomization energy

DNN approximation
for organic molecules

- Interpretation of the trained
DNN model

References

**Tutorial / Overview Papers**


**Methods Papers**


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


**Evaluation Explanations**

References

Application to Text


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


Application to Images & Faces


References

Application to Video


Application to Speech

Application to the Sciences


Thank you for your attention

Visit:

http://www.heatmapping.org

- Tutorials
- Software
- Online Demos

Acknowledgement
Klaus-Robert Müller (TUB)
Grégoire Montavon (TUB)
Sebastian Lapuschkin (HHI)
Leila Arras (HHI)
Alexander Binder (SUTD)
...

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/investigate