# EMBC Tutorial on Interpretable and Transparent Deep Learning

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

<table>
<thead>
<tr>
<th>Time</th>
<th>Session Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>13:30 - 14:00</td>
<td>Introduction KRM</td>
</tr>
<tr>
<td>14:00 - 15:00</td>
<td>Techniques for Interpretability GM</td>
</tr>
<tr>
<td>15:00 - 15:30</td>
<td>Coffee Break ALL</td>
</tr>
<tr>
<td>15:30 - 16:15</td>
<td>Evaluating Interpretability &amp; Applications WS</td>
</tr>
<tr>
<td>16:15 - 17:15</td>
<td>Applications in BME &amp; the Sciences and Wrap-Up KRM</td>
</tr>
</tbody>
</table>
LRP Revisited

Theoretical Interpretation
(Deep) Taylor decomposition.

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

**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 \( \alpha + \beta = 1 \)
LRP Works for Different Data

General Images (Bach’15, Lapuschkin’16)

Speech (Becker’18)

Text Analysis (Arras’16 & 17)
do n’t waste your money
neither funny nor susper

Morphing (Seibold’18)

Games (Lapuschkin’19)

VQA (Samek’19)

Video (Anders’18)

Gait Patterns (Horst’19)

Faces (Lapuschkin’17)

Digits (Bach’15)

Histopathology (Binder’18)

EEG (Sturm’16)

fMRI (Thomas’18)
LRP Works for 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)

Clustering (Kauffmann’19)
Other Explanation Methods

Question: Which one to choose?
How to Evaluate Quality of Explanations?

input $x$

explanation $R$

ground truth $R^*$

$\text{DNN}$

$f(x)$

evidence for “truck”

$\| R - R^* \|^2$

error
Compare Explanation Methods

Idea: Compare selectivity (Bach’15, Samek’17):
“If input features are deemed relevant, removing them should reduce evidence at the output of the network.”

Important: Remove information in a non-specific manner (e.g. sample from uniform distribution)

Algorithm (“Pixel Flipping”)
Sort pixels / patches by relevance
Iterate
  destroy pixel / patch
  evaluate f(x)
Measure decrease of f(x)
Compare Explanation Methods

LRP

# pixel flips: 0

score for correct class

# pixel flips
Compare Explanation Methods

LRP

score for correct class

# pixel flips

# pixel flips: 20
Compare Explanation Methods

LRP

# pixel flips: 40
Compare Explanation Methods

LRP

AOC=0.722

# pixel flips: 100
Compare Explanation Methods

Sensitivity

# pixel flips: 0
Compare Explanation Methods

Sensitivity

score for correct class

# pixel flips

# pixel flips: 30
Compare Explanation Methods

Sensitivity

AOC = 0.691

# pixel flips: 100
Compare Explanation Methods

Random

(score for correct class)

# pixel flips

# pixel flips: 0
Compare Explanation Methods

Random

# pixel flips: 30
Compare Explanation Methods

Random

AOC=0.523

# pixel flips: 100
Compare Explanation Methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>LRP</td>
<td>0.722</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.691</td>
</tr>
<tr>
<td>Random</td>
<td>0.523</td>
</tr>
</tbody>
</table>

LRP produces quantitatively better heatmaps than sensitivity analysis and random.

What about more complex datasets?

- **SUN397**: 397 scene categories (108,754 images in total)
- **ILSVRC2012**: 1000 categories (1.2 million training images)
- **MIT Places**: 205 scene categories (2.5 millions of images)
Compare Explanation Methods

Red: LRP method
Blue: Deconvolution method (Zeiler & Fergus, 2014)
Green: Sensitivity method (Simonyan et al., 2014)

- ImageNet: Caffe reference model
- Places & SUN: Classifier from MIT
- AOPC averages over 5040 images
- perturb $9 \times 9$ nonoverlapping regions
- 100 steps (15.7% of the image)
- uniform sampling in pixel space

LRP produces better heatmaps

(Samek et al. 2017)
Axiomatic Approach to Interpretability

Idea: Evaluate the explanation technique axiologically, 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.

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

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

\[ \forall_{p=1}^{d} : R_p \geq 0 \]
Axiomatic Approach to Interpretability

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:

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

![Diagram showing LRP-α₁β₀ and sensitivity analysis](image)

- LRP-α₁β₀
- Sensitivity analysis

![Function graphs](image)
# Axiomatic Approach to Interpretability

| Explanation techniques | Uniform | (Gradient)$^2$ | (Guided BP)$^2$ | Gradient x Input | Guided BP x Input | LRP-$\alpha,\beta_0$ | ...
<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>
The Idea: Use toy model for evaluating explanations. [Arras’19]

"Explanations should only highlight the parts in the input which are actually important (by design)."

1. LSTM is trained to add and subtract numbers from first row.
2. Measure correlation with relevance values

\[
\begin{pmatrix}
0 & 0 & 0 & n_a & 0 & 0 & 0 & n_b & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & n_a & 0 & 0 & 0 & n_b & 0 & 0 & 0 \\
... & ... & ... & ... & ... & ... & ... & ... & ... & ... & ... & ...
\end{pmatrix}
\]

<table>
<thead>
<tr>
<th>Toy Task Addition</th>
<th>(\rho(R_{xα}, n_α)) (in %)</th>
<th>(\rho(R_{xβ}, n_β)) (in %)</th>
<th>Toy Task Subtraction</th>
<th>(\rho(R_{xα}, n_α)) (in %)</th>
<th>(\rho(R_{xβ}, n_β)) (in %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Occlusion</td>
<td>99.990 (0.004)</td>
<td>99.990 (0.004)</td>
<td>Occlusion</td>
<td>99.0 (2.0)</td>
<td>-69.0 (19.1)</td>
</tr>
<tr>
<td>LRP-half [ACL best paper]</td>
<td>29.035 (9.478)</td>
<td>51.460 (19.939)</td>
<td>LRP-half [ACL best paper]</td>
<td>7.7 (15.3)</td>
<td>-28.9 (6.4)</td>
</tr>
<tr>
<td><strong>LRP-all [ours]</strong></td>
<td><strong>99.995 (0.002)</strong></td>
<td><strong>99.995 (0.002)</strong></td>
<td><strong>LRP-all [ours]</strong></td>
<td><strong>98.5 (3.5)</strong></td>
<td><strong>-99.3 (1.3)</strong></td>
</tr>
<tr>
<td>CD [ICLR oral]</td>
<td><strong>99.997 (0.002)</strong></td>
<td><strong>99.997 (0.002)</strong></td>
<td>CD [ICLR oral]</td>
<td>-25.9 (39.1)</td>
<td>-50.0 (29.2)</td>
</tr>
</tbody>
</table>
Summary Evaluation

LRP heatmaps are informative, truthworthy and fulfil important axioms.

Furthermore, LRP produces qualitative and quantitatively better explanations than sensitivity analysis, deconvolution, context decomposition, gradient times input, occlusion and …

Let’s now see some applications!
Application of LRP
Compare models
Application: Compare Classifiers

**word2vec/CNN:**

Performance: 80.19%

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

**BoW/SVM:**

Performance: 80.10%

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

---

>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, i.e., 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.

---

## Application: Compare Classifiers

<table>
<thead>
<tr>
<th>word2vec / CNN model</th>
<th>BoW/SVM model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>sci.med</strong></td>
<td><strong>sci.med</strong></td>
</tr>
<tr>
<td>symptoms (7.3), treatments (6.6), medication (6.4), osteopathy (6.3), ulcers (6.2), sciatica (6.0), hypertension (6.0), herb (5.6), doctor (5.4), physician (5.1), Therapy (5.1), antibiotics (5.1), Asthma (5.0), renal (5.0), medicines (4.9), caffeine (4.9), infection (4.9), gastrointestinal (4.8), therapy (4.8), homeopathic (4.7), medicine (4.7), allergic (4.7), dosages (4.7), esophagitis (4.7), inflammation (4.6), arrhythmias (4.6), cancer (4.6), disease (4.6), migraine (4.6), patients (4.5).</td>
<td>cancer (1.4), photography (1.0), doctor (1.0), msg (0.9), disease (0.9), medical (0.8), sleep (0.8), radiologist (0.7), eye (0.7), treatment (0.7), prozac (0.7), vitamin (0.7), epilepsy (0.7), health (0.6), yeast (0.6), skin (0.6), pain (0.5), liver (0.5), physician (0.5), she (0.5), needles (0.5), dn (0.5), circumcision (0.5), syndrome (0.5), migraine (0.5), antibiotic (0.5), water (0.5), blood (0.5), fat (0.4), weight (0.4).</td>
</tr>
</tbody>
</table>

**Words with maximum relevance**

Application of LRP
Quantify Context Use
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

- BVLC reference model + fine tuning
- PASCAL VOC 2007

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

- Different models (BVLC CaffeNet, GoogleNet, VGG CNN S)
- ILSVCR 2012

Context use anti-correlated with performance.

(Lapuschkin et al. 2016)
Application of LRP
Detect Biases
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 of LRP
Relevance-Based Filtering
Application: Learn new Representations

\[ \text{document vector} = R_a \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_d \end{pmatrix} + R_b \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_d \end{pmatrix} + R_c \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_d \end{pmatrix} \]

Application: Learn new Representations

2D PCA projection of document vectors

Document vector computation is **unsupervised** (given we have a classifier).

Application of LRP
Understand Model
Application: Gait Analysis

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

(Horst et al. 2019)
Application: Gait Analysis

<table>
<thead>
<tr>
<th>Model</th>
<th>Ground Reaction Forces [%]</th>
<th>Joint Angles Full-Body [%]</th>
<th>Joint Angles Full-Body (flex.-ext.) [%]</th>
<th>Joint Angles Lower-Body [%]</th>
<th>Joint Angles Lower-Body (flex.-ext.) [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN-A</td>
<td>99.1 (0.8)</td>
<td>100.0 (0.0)</td>
<td>95.6 (1.7)</td>
<td>99.9 (0.3)</td>
<td>92.0 (3.9)</td>
</tr>
</tbody>
</table>

- **Ground Reaction Force - CNN-A:** Relevance of Input Per Subject
- **Lower-Body Joint Angles (flexion-extension) - CNN-A:** Relevance of Input Per Subject

**Insights for subject 6**
- extension of the ankle during the terminal stance phase
- flexion of the knee and hip during the initial contact of the right leg
Application: Understand the model

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

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

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

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

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

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

- reimplement model of (Santoro et al., 2017)
- test accuracy of 91.0%
- CLEVR dataset

model understands the question and correctly identifies the object of interest

(in prep)
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., 2019)
Application: Understand the model

After 0 epochs

After 25 epochs

After 195 epochs

(Lapuschkin et al., 2019)
Application: Understand the model

Relevance Distribution during Training

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

(Lapuschkin et al., 2019)
Application: Understand the model

(Lapuschkin et al., 2019)
More information

Visit:

http://www.heatmapping.org

- Tutorials
- Software
- Online Demos

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

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

Opinion Paper

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.


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**


References

Evaluation Explanations


Software
