Tutorial on Interpreting and Explaining Deep Models in Computer Vision

Wojciech Samek  
(Fraunhofer HHI)

Grégoire Montavon  
(TU Berlin)

Klaus-Robert Müller  
(TU Berlin)

08:30 - 09:15  
Introduction KRM

09:15 - 10:00  
Techniques for Interpretability GM

10:00 - 10:30  
Coffee Break ALL

10:30 - 11:15  
Applications of Interpretability WS

11:15 - 12:00  
Further Applications and Wrap-Up KRM
Opening the Black Box with LRP

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

Explain prediction
(how much each pixel contributes to prediction)

Simple LRP rule (Bach et al. 2015)
\[ R_i^{(l)} = \sum_j \frac{x_i \cdot w_{ij}}{\sum_i x_i \cdot w_{ij}} R_j^{(l+1)} \]

Every neuron gets its "share" of the redistributed relevance
Opening the Black Box with LRP

Theoretical Interpretation

(Deep) Taylor decomposition

Excitation Backprop (Zhang et al., 2016) is special case of LRP (α=1).

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

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

where \( \alpha + \beta = 1 \)
LRP applied to different Data

General Images (Bach’15, Lapuschkin’16)

Text Analysis (Arras’16 &’17)

Translation (Ding’17)

Molecules (Schütt’17)

Speech (Becker’18)

Gait Patterns (Horst’18, in prep.)

VQA (Arras’18)

Video (Anders’18)

EEG (Sturm’16)

Morphing (Seibold’18)

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)
Now What?
Compare Explanation Methods

Algorithm ("Pixel Flipping")
Sort pixels / patches by relevance
Iterate
  destroy pixel / patch
  evaluate $f(x)$
Measure decrease of $f(x)$

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)
Compare Explanation Methods

LRP

(score for correct class)

( # pixel flips)

# pixel flips: 0
Compare Explanation Methods

LRP

# pixel flips: 20
Compare Explanation Methods

LRP

# pixel flips: 40
Compare Explanation Methods

LRP

AOC=0.722

# pixel flips: 100
Compare Explanation Methods

Sensitivity

![Graph showing sensitivity with a heatmap of pixel flips]

# pixel flips: 0
Compare Explanation Methods

Sensitivity

# pixel flips: 30
Compare Explanation Methods

Sensitivity

AOC = 0.691

# pixel flips: 100
Compare Explanation Methods

Random

# pixel flips: 0

score for correct class

# pixel flips

0 10 20 30 40 50 60 70 80 90 100
Compare Explanation Methods

Random

# pixel flips: 30

# pixel flips: 30

# 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?

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUN397</td>
<td>397 scene categories (108,754 images in total)</td>
</tr>
<tr>
<td>ILSVRC2012</td>
<td>1000 categories (1.2 million training images)</td>
</tr>
<tr>
<td>MIT Places</td>
<td>205 scene categories (2.5 millions of images)</td>
</tr>
</tbody>
</table>
Compare Explanation Methods

- Sensitivity Analysis (Simonyan et al. 2014)
- Deconvolution Method (Zeiler & Fergus 2014)
- LRP Algorithm (Bach et al. 2015)

(Samek et al. 2017)
Compare Explanation Methods

**Red:** LRP method

**Blue:** Deconvolution method (Zeiler & Fergus, 2014)

**Green:** Sensitivity method (Simonyan et al., 2014)

---

LRP produces better heatmaps

- Sensitivity heatmaps are noisy (gradient shuttering)
- Deconvolution and sensitivity analysis solve a different problem

---

(Samek et al. 2017)

- 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
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.

>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.

(Arras et al. 2017)
Compare Explanation Methods

Deleting most relevant from correctly classified

Deleting least relevant from falsely classified

LRP better than SA

LRP distinguishes between positive and negative evidence

(Arras et al. 2016)

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

Deleting most relevant from correctly classified
Deleting least relevant from falsely classified

- bidirectional LSTM model (Li’16)
- Stanford Sentiment Treebank dataset
- delete up to 5 words per sentence

LRP outperforms baselines (also recently proposed contextual decomposition)

LRP ≠ Gradient x Input

(Ding et al. ACL, 2017)
(Murdoch et al. ICLR, 2018)
(Arras et al. EMNLP-WASSA, 2017)
Compare Explanation Methods

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

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

---

Application: Compare Classifiers

word2vec / CNN model

\[ \text{sci.med} \]

- 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).

BoW/SVM model

\[ \text{sci.med} \]

- cancer (1.4), photography (1.0), doctor (1.0), \text{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).

Words with maximum relevance

Application: Compare Classifiers

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

BVLC:
- 8 Layers
- ILSRCV: 16.4%
Application: Compare Classifiers

GoogleNet focuses on faces of animal.  
→suppresses background noise

BVLC CaffeNet heatmaps are much more noisy.

*Is it related to the architecture?*

*Is it related to the performance?*

<table>
<thead>
<tr>
<th>Image</th>
<th>BVLC CaffeNet</th>
<th>GoogleNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Image of a cat]</td>
<td>[Heatmap of a cat's face]</td>
<td>[Heatmap of a cat's face]</td>
</tr>
<tr>
<td>[Image of a cat]</td>
<td>[Heatmap of a cat's face]</td>
<td>[Heatmap of a cat's face]</td>
</tr>
<tr>
<td>[Image of a lion]</td>
<td>[Heatmap of a lion's face]</td>
<td>[Heatmap of a lion's face]</td>
</tr>
</tbody>
</table>

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

LRP decomposition allows meaningful pooling over bbox!

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

importance of context = relevance outside bbox

relevance inside bbox
Application: Measure Context Use

- BVLC reference model + fine tuning
- PASCAL VOC 2007

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

Context use anti-correlated with performance.

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

(Lapuschkin et al. 2016)
Application of LRP
Compare Configuration,
Detect Biases & Improve Models
Application: Face analysis

- Compare AdienceNet, CaffeNet, GoogleNet, VGG-16
- State-of-the-art performance in age and gender classification
- Adience dataset, 26,580 images

### Age classification

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>[i]</td>
<td>51.4</td>
<td>87.0</td>
<td>52.1</td>
<td>87.9</td>
</tr>
<tr>
<td>[r]</td>
<td>51.9</td>
<td>87.4</td>
<td>52.3</td>
<td>88.9</td>
</tr>
<tr>
<td>[m]</td>
<td>53.6</td>
<td>88.4</td>
<td>54.3</td>
<td>89.7</td>
</tr>
<tr>
<td>[i,n]</td>
<td>–</td>
<td>51.6</td>
<td>87.4</td>
<td>56.2</td>
</tr>
<tr>
<td>[r,n]</td>
<td>–</td>
<td>52.1</td>
<td>87.0</td>
<td>57.4</td>
</tr>
<tr>
<td>[m,n]</td>
<td>–</td>
<td>52.8</td>
<td>88.3</td>
<td>58.5</td>
</tr>
<tr>
<td>[i,w]</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>[r,w]</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>[m,w]</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

### Gender classification

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>[i]</td>
<td>88.1</td>
<td>87.4</td>
<td>87.9</td>
<td>–</td>
</tr>
<tr>
<td>[r]</td>
<td>88.3</td>
<td>87.8</td>
<td>88.9</td>
<td>–</td>
</tr>
<tr>
<td>[m]</td>
<td>89.0</td>
<td>88.8</td>
<td>89.7</td>
<td>–</td>
</tr>
<tr>
<td>[i,n]</td>
<td>–</td>
<td>89.9</td>
<td>–</td>
<td>91.0</td>
</tr>
<tr>
<td>[r,n]</td>
<td>–</td>
<td>90.6</td>
<td>–</td>
<td>91.6</td>
</tr>
<tr>
<td>[m,n]</td>
<td>–</td>
<td>90.6</td>
<td>–</td>
<td>91.7</td>
</tr>
<tr>
<td>[i,w]</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>[r,w]</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>[m,w]</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

A = AdienceNet  
C = CaffeNet  
G = GoogleNet  
V = VGG-16  
[i] = in-place face alignment  
[r] = rotation based alignment  
[m] = mixing aligned images for training  
[n] = initialization on Imagenet  
[w] = initialization on IMDB-WIKI  

*(Lapuschkin et al., 2017)*
Application: Face analysis

Gender classification

Strategy to solve the problem: Focus on chin / beard, eyes & ear, 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: Face analysis

<table>
<thead>
<tr>
<th>real person</th>
<th>fake person</th>
<th>real person</th>
</tr>
</thead>
</table>

Different training methods

<table>
<thead>
<tr>
<th></th>
<th>naive</th>
<th>one morphed</th>
<th>complex morphs</th>
<th>multiclass</th>
</tr>
</thead>
<tbody>
<tr>
<td>true positive</td>
<td>95%</td>
<td>90%</td>
<td>93%</td>
<td>92%</td>
</tr>
<tr>
<td>true negative</td>
<td>98%</td>
<td>95%</td>
<td>95%</td>
<td>99%</td>
</tr>
<tr>
<td>EER</td>
<td>3.1%</td>
<td>7.2%</td>
<td>6.1%</td>
<td>2.8%</td>
</tr>
</tbody>
</table>

- 1,900 images of different individuals
- pretrained VGG19 model
- different ways to train the models

50% genuine images, 50% complete morphs

50% genuine images, 10% complete morphs and 4 x 10% one region morphed

50% genuine images, 10% complete morphs, partial morphs with 10% one, two, three and four region morphed

partial morphs with zero, one, two, three or four morphed regions, for two class classification last layer reinitialized

(Seibold et al., 2018)
Application: Face analysis

Semantic attack on the model

<table>
<thead>
<tr>
<th></th>
<th>left eye</th>
<th>right eye</th>
<th>nose</th>
<th>mouth</th>
<th>average</th>
</tr>
</thead>
<tbody>
<tr>
<td>naive</td>
<td>25%</td>
<td>21%</td>
<td>14%</td>
<td>13%</td>
<td>20%</td>
</tr>
<tr>
<td>one morphed</td>
<td>81%</td>
<td>89%</td>
<td>79%</td>
<td>71%</td>
<td>80%</td>
</tr>
<tr>
<td>complex morphs</td>
<td>78%</td>
<td>74%</td>
<td>73%</td>
<td>54%</td>
<td>70%</td>
</tr>
<tr>
<td>multiclass</td>
<td>86%</td>
<td>93%</td>
<td>90%</td>
<td>79%</td>
<td>87%</td>
</tr>
</tbody>
</table>

Table 4. Robustness against partial morphs.

Black box adversarial attack on the model

Fig. 5. Robustness against fast gradient sign attacks.
### Application: Face analysis

<table>
<thead>
<tr>
<th>morphed region</th>
<th>relative amount of relevance per region</th>
<th>one morphed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>naive</td>
<td>left eye</td>
</tr>
<tr>
<td>left eye</td>
<td>0.84</td>
<td>0.96</td>
</tr>
<tr>
<td>right eye</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>nose</td>
<td>0.21</td>
<td>0.00</td>
</tr>
<tr>
<td>mouth</td>
<td>0.34</td>
<td>0.17</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>morphed region</th>
<th>complex morphs</th>
<th>multiclass</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>left eye</td>
<td>right eye</td>
</tr>
<tr>
<td>left eye</td>
<td>0.98</td>
<td>0.00</td>
</tr>
<tr>
<td>right eye</td>
<td>0.00</td>
<td>0.92</td>
</tr>
<tr>
<td>nose</td>
<td>0.02</td>
<td>0.03</td>
</tr>
<tr>
<td>mouth</td>
<td>0.06</td>
<td>0.00</td>
</tr>
</tbody>
</table>

(Seibold et al., 2018)
Application: Face analysis

Different models have different strategies!

network seems to compare different structures

network seems to identify “original” parts

(Seibold et al., 2018)
Application of LRP
Learn new Representations
Application: Learn new Representations

\[ \text{document vector} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_d \end{pmatrix} = R_a \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_d \end{pmatrix} + R_b \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_d \end{pmatrix} + R_c \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_d \end{pmatrix} \]

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 & Obtain new Insights
Application: Understand the model

- Fisher Vector / SVM classifier
- PASCAL VOC 2007

(Lapuschkin et al. 2016)
Application: Understand the model

(Lapuschkin et al. 2016)
Application: Understand the model

Motion vectors can be extracted from the compressed video -> allows very efficient analysis

- Fisher Vector / SVM classifier
- Model of Kantorov & Laptev, (CVPR’14)
- Histogram Of Flow, Motion Boundary Histogram
- HMDB51 dataset

(Srinivasan et al. 2017)
Application: Understand the model

- bidirectional LSTM model (Li’16)
- Stanford Sentiment Treebank dataset

movie review: ++, −

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 & 2018)
### Application: Understand the model

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

<table>
<thead>
<tr>
<th>frame 1</th>
<th>frame 4</th>
<th>frame 7</th>
<th>frame 10</th>
<th>frame 13</th>
<th>frame 16</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Frame 1" /></td>
<td><img src="image4.png" alt="Frame 4" /></td>
<td><img src="image7.png" alt="Frame 7" /></td>
<td><img src="image10.png" alt="Frame 10" /></td>
<td><img src="image13.png" alt="Frame 13" /></td>
<td><img src="image16.png" alt="Frame 16" /></td>
</tr>
</tbody>
</table>

*(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.
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 there any large cyan metallic objects behind it?</td>
<td>there is a metallic cube; are there 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

(Arras et al., 2018)
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

After 0 epochs

After 25 epochs

After 195 epochs

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

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

Relevance Distribution during Training

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

(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
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

