# Tutorial on Interpretable Machine Learning

**Wojciech Samek**  
(Fraunhofer HHI)  

**Alexander Binder**  
(SUTD)

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>15:00 - 15:15</td>
<td>Introduction &amp; Motivation WS</td>
</tr>
<tr>
<td>15:15 - 16:30</td>
<td>Techniques for Interpretability WS</td>
</tr>
<tr>
<td>16:30 - 17:00</td>
<td>Coffee Break ALL</td>
</tr>
<tr>
<td>17:00 - 18:00</td>
<td>Applications of Interpretability WS</td>
</tr>
<tr>
<td>18:00 - 18:50</td>
<td>Case Study: Interpretable ML in Histopathology AB</td>
</tr>
<tr>
<td>18:50 - 19:00</td>
<td>Wrap-up WS</td>
</tr>
</tbody>
</table>
Before we start

Joint work with many people

Klaus-Robert Müller (TU Berlin)
Grégoire Montavon (TU Berlin)
Sebastian Lapuschkin (Fraunhofer HHI)
Leila Arras (Fraunhofer HHI)
Frederick Klauschen (Charite)
...

http://interpretable-ml.org/miccai2018tutorial/

Please ask questions at any time!
Tutorial on Interpretable Machine Learning

Part 1: Introduction & Motivation
Record Performances with ML

Game GO  Traffic Sign Recognition  Skin cancer detection  Lung cancer detection

Poker  Computer games  Jeopardy  OCR
Black Box Models

Huge volumes of data

Computing power

Deep Neural Network

Solve task

Information (implicit)
Black Box Models

Is minimizing the error a guarantee for the model to work well in practice?
Why interpretability?
Why Interpretability?

We need interpretability in order to:

- verify system
- understand weaknesses
- legal aspects
- learn new things from data
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) Understand weaknesses & improve classifier

Generalization error

Generalization error + human experience
3) Learn new things 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.”
4) Interpretability in the sciences

Learn about the physical / biological / chemical mechanisms. (e.g. find genes linked to cancer, identify binding sites …)
Why Interpretability?

5) Compliance to legislation

European Union’s new General Data Protection Regulation \rightarrow \text{“right to explanation”}

Retain human decision in order to assign responsibility.

\text{“With interpretability we can ensure that ML models work in compliance to proposed legislation.”}
ITU/WHO Focus Group on AI4Health

Focus Group on “Artificial Intelligence for Health” established by

ITU Workshop on Artificial Intelligence for Health
Geneva, Switzerland, 25 September 2018

More information about the group:
https://www.itu.int/en/ITU-T/focusgroups/ai4h
Dimensions of Interpretability

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.”
Dimensions of Interpretability

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

- train best model
  - interpret it
Dimensions of Interpretability

Question: Which one to choose?
Tutorial on Interpretable Machine Learning

Part 2: Techniques of Interpretability
Techniques of Interpretation

**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.
Techniques of Interpretation

![Diagram of Techniques of Interpretation]

- **Functional Understanding**: The relationship $f: \mathbb{R}^d \rightarrow \mathbb{R}$
- **Model Analysis**
- **Decision Analysis**
Model analysis
Interpreting the Model

Approach 1: Class Prototypes

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

\[ \text{arg max}_x f(x) + \text{reg.} \]
Interpreting the Model

Activation Maximization

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

\[
\max_{x \in \mathcal{X}} p_{\theta}(\omega_c \mid x) + \lambda \Omega(x)
\]
Interpreting the Model

Activation Maximization

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

\[
\max_{x \in \mathcal{X}} p_{\theta}(\omega_c \mid x) + \lambda \Omega(x)
\]
Interpreting the Model

Activation Maximization

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

complex regularizer (Nguyen et al. 2016)

$$\max_{x \in \mathcal{X}} p_{\theta}(\omega_c \mid x) + \lambda \Omega(x)$$
Interpreting the Model

Activation Maximization

Let us interpret a concept predicted by a deep neural net (e.g. a class, or a real-valued quantity):

input pattern
\[ x \]

deep neural network
\[ f(x) \]

\[ \log p(\omega_c|x) \]

class probability
real-valued output

Examples:
- Creating a class prototype: \( \max_{x \in \mathcal{X}} \log p(\omega_c|x) \).
- Synthesizing an extreme case: \( \max_{x \in \mathcal{X}} f(x) \).
Interpreting the Model


Observations:
- AM builds typical patterns for these classes (e.g. beaks, legs).
- Unrelated background objects are not present in the image.
Enhancing Activation Maximization

Find the input pattern that maximizes class probability. → Find the most likely input pattern for a given class.
Enhancing Activation Maximization

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.
Finding a prototype:

GDB-7

{ }

Question: How does a molecule with properties XYZ look like?
Limitations of Global Interpretations

Question: Below are some images of motorbikes. What would be the best prototype to interpret the class “motorbike”?

Observations:

- Summarizing a concept or category like “motorbike” into a single image can be difficult (e.g. different views or colors).
- A good interpretation would grow as large as the diversity of the concept to interpret.
Need for Individual Explanations

Finding a prototype:

Question: How does a “motorbike” typically look like?

Individual explanation:

Question: Why is this example classified as a motorbike?
**Need for Individual Explanations**

**Personalized medicine:** Extracting the relevant information about a medical condition for a *given* patient at a *given* time.

Each case is unique and needs its own explanation.
Need for Individual Explanations

**Personalized medicine:** Extracting the relevant information about a medical condition for a *given* patient at a *given* time.

> Each case is unique and needs its own explanation.

**Population view:** Which symptoms are most common for the disease

Both aspects can be important depending on who you are (FDA, doctor, patient).
Making Deep Neural Nets Transparent

- visualizing filters
- max. class activation

- include distribution (RBM, DGN, etc.)

- sensitivity analysis
- decomposition
Decision analysis
Decision Analysis: Sensitivity Analysis

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

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

Sensitivity analysis:

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

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

highlights parts, which (when changed) increase or decrease the prediction for “car”.
Decision Analysis: Sensitivity Analysis

**Sensitivity analysis:**

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

**Observation:**

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

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

highlights parts, which (when changed) increase or decrease the prediction for “car”.

**Sensitivity analysis explains a variation of the function, not the function value itself.**
Decision Analysis: Sensitivity Analysis

Shattered Gradient Problem

Input gradient (on which sensitivity analysis is based), becomes increasingly highly varying and unreliable with neural network depth.
Layer-wise Relevance Propagation (LRP)
Decision Analysis: LRP

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

Explain prediction itself (not the change)
Decision Analysis: LRP

Classification

cat
rooster
dog
What makes this image a “rooster image”? 

Idea: Redistribute the evidence for class rooster back to image space.
Decision Analysis: LRP

![Diagram showing a neural network with nodes labeled as 'cat', 'rooster', and 'dog'.]

**Simple LRP rule (Bach et al. 2015)**

\[ R_i^{(l)} = \sum_{j} \frac{x_i \cdot w_{i,j}}{\sum_{i'} x_{i'} \cdot w_{i',j}} R_j^{(l+1)} \]

Every neuron gets its "share" of the redistributed relevance.
Decision Analysis: LRP

Theoretical interpretation
Deep Taylor Decomposition
(Montavon et al., 2017)
(no gradient shattering)

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 \( \alpha + \beta = 1 \)
Decision Analysis: LRP

Layer-wise relevance conservation

$$\sum_i R_i = \ldots = \sum_i R_i^{(l)} = \sum_j R_j^{(l+1)} = \ldots = f(x)$$
Decision Analysis: LRP

Heatmap of prediction “3”

Heatmap of prediction “9”
Decision Analysis: LRP

Image  | Sensitivity Analysis  | LRP / Deep Taylor

- Explains what influences prediction “cars”.
- Explains prediction “cars” as is.

Slope decomposition:
\[ \sum_i R_i = \| \nabla_x f \|^2 \]

Value decomposition:
\[ \sum_i R_i = f(x) \]

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

**Gradients**
- Sensitivity (Baehrens et al. 2010)
- Sensitivity (Morch et al., 1995)
- Sensitivity (Simonyan et al. 2014)

**Decomposition**
- Gradient times input (Shrikumar et al., 2016)
- DeepLIFT (Shrikumar et al., 2016)
- Grad-CAM (Selvaraju et al., 2016)
- Integrated Gradient (Sundararajan et al., 2017)

- LRP (Bach et al., 2015)

**Question:** Which one to choose?

**Deconvolution**
- Deconvolution (Zeiler & Fergus 2014)
- Guided Backprop (Springenberg et al. 2015)

**Understanding the Model**
- Deep Visualization (Yosinski et al., 2015)
- Inverting CNNs (Dosovitskiy & Brox, 2015)
- Synthesis of preferred inputs (Nguyen et al. 2016)
- Inverting CNNs (Mahendran & Vedaldi, 2015)
- RNN cell state analysis (Karpathy et al., 2015)
- Network Dissection (Zhou et al. 2017)
Axiomatic approach to interpretability
First Attempt: Distance to Ground Truth

input $x$

explanation $R$

ground truth $R^*$

DNN $f(x)$

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

evidence for “truck”
Axiomatic Approach to Interpretability

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]

\[ x_1 \rightarrow DNN \rightarrow f(x) = f_{exp}(x) + \varepsilon \]

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).

\[ \forall_{p=1}^{d} R_p \geq 0 \]
Property 3: Continuity \[\text{[Montavon'18]}\]

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:

\[f(x) = \max(x_1, x_2)\]
Axiomatic Approach to Interpretability

Testing Continuity

LRP-\alpha_1\beta_0

Sensitivity analysis
Property 4: Selectivity \([\text{Bach'15, Samek'17}]\)

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

\[ \text{LRP-} \alpha_1 \beta_0 \]

Sensitivity analysis
# Axiomatic Approach to Interpretability

<table>
<thead>
<tr>
<th>Explanation techniques</th>
<th>Uniform</th>
<th>(Gradient)$^2$</th>
<th>(Guided BP)$^2$</th>
<th>Gradient x Input</th>
<th>Guided BP x Input</th>
<th>LRP-α,β,...</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
<td><img src="image5.png" alt="Image" /></td>
<td><img src="image6.png" alt="Image" /></td>
<td><img src="image7.png" alt="Image" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Properties</th>
<th>Uniform</th>
<th>(Gradient)$^2$</th>
<th>(Guided BP)$^2$</th>
<th>Gradient x Input</th>
<th>Guided BP x Input</th>
<th>LRP-α,β,...</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Conservation</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>
</tr>
<tr>
<td>3. Continuity</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>
</tr>
</tbody>
</table>

...
Summary LRP

General Images (Bach’15, Lapuschkin’16)

Speech (Becker’18)

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

Morphing (Seibold’18)

Games (Lapuschkin’18)

VQA (Arras’18)

Gait Patterns (Horst’18)

Video (Anders’18)

EEG (Sturm’16)

fMRI (Thomas’18)

Faces (Lapuschkin’17)

Digits (Bach’15)

Histopathology (Binder’18)
Summary LRP

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)
Summary LRP

1. LRP solves the “correct” explanation problem
2. It has a theoretical interpretation (Deep Taylor Decomposition)
3. It can be applied to various data and models (not only deep nets)
4. It fulfills various criteria (axiomatic approach)
5. It is flexible (many explanation methods are special cases of LRP)
6. In general: $\text{LRP} \neq \text{Gradient} \times \text{Input}$

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
From LRP to Deep Taylor Decomposition
Decomposing the Correct Quantity

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

**Candidate:** Taylor decomposition

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

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

\[
\tilde{x} = \lim_{\epsilon \to 0} \epsilon \cdot x \approx 0.
\]
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
Deep Taylor Decomposition

**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)
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

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

Expand the Relevance Neuron

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

Decompose Relevance

Taylor expansion at root point:

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

\[ \begin{align*}
0 \\
\frac{(a_i - \tilde{a}_i^{(j)})w_{ij}}{\sum_i(a_i - \tilde{a}_i^{(j)})w_{ij}} R_j \\
0
\end{align*} \]

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)} \in \mathcal{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>✔️</td>
</tr>
<tr>
<td>2. rescaled activation</td>
<td>( \tilde{a}^{(j)} = a - t \cdot a )</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>✔️</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

Pooling relevance over all outgoing neurons

\[ \sum_j \]
Deep Taylor Decomposition

The LRP-α₁β₀ 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
Coffee Break
Tutorial on Interpretable Machine Learning

Part 3: Applications of Interpretability
LRP revisited

Idea: Decompose function

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

Simple LRP rule (Bach et al. 2015)

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

Every neuron gets its "share" of the redistributed relevance
LRP revisited

Theoretical Interpretation
(Deep) Taylor decomposition

Input $x$

Black Box

Rooster prediction $f(x)$

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

Explain prediction
(how much each pixel contributes to prediction)

heatmap

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 revisited

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’18)

VQA (Arras’18)

Video (Anders’18)

Gait Patterns (Horst’18)

Faces (Lapuschkin’17)

Digits (Bach’15)

Histopathology (Binder’18)

EEG (Sturm’16)

fMRI (Thomas’18)
LRP revisited

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 & Others
Evaluating Heatmap Quality
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

LRP

# pixel flips: 0

score for correct class

# pixel flips

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

LRP

# pixel flips: 20
Compare Explanation Methods

LRP

(score for correct class)

# pixel flips: 40
Compare Explanation Methods

LRP

AOC=0.722

# pixel flips: 100
Compare Explanation Methods

Sensitivity

(score for correct class vs. # pixel flips)

# pixel flips: 0
Compare Explanation Methods

Sensitivity

(score for correct class vs. # pixel flips)

# pixel flips: 30
Compare Explanation Methods

Sensitivity

\begin{itemize}
  \item AOC = 0.691
  \item \# pixel flips: 100
\end{itemize}
Compare Explanation Methods

Random

(score for correct class)

# pixel flips

# pixel flips: 0
Compare Explanation Methods

Random

(score for correct class)

# pixel flips

# pixel flips: 30
Compare Explanation Methods

Random

score for correct class vs. # pixel flips

AOC=0.523

# pixel flips: 100
LRP: 0.722
Sensitivity: 0.691
Random: 0.523

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

- 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

(From Samek et al. 2017)

- ImageNet: Caffe reference model
- Places & SUN: Classifier from MIT
- AOPC averages over 5040 images
- Perturb 9 x 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, 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

Deleting most relevant from correctly classified

Deleting least relevant from falsely classified

- 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

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

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)

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

Fraunhofer Heinrich Hertz Institute

MICCAI’18 Tutorial on Interpretable Machine Learning 93
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

| 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

| sci.med | 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). |

Words with maximum relevance

*(Arras et al. 2016 & 2017)*
## LRP in Practice

### Visual Object Classes Challenge: 2005 - 2012

<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>
<th>chair</th>
<th>cow</th>
<th>diningtable</th>
<th>dog</th>
<th>horse</th>
<th>motorbike</th>
<th>person</th>
<th>pottedplant</th>
<th>sheep</th>
<th>sofa</th>
<th>train</th>
<th>tvmonitor</th>
</tr>
</thead>
<tbody>
<tr>
<td>INRIA.Flat</td>
<td>74.8</td>
<td>62.5</td>
<td>51.2</td>
<td>69.4</td>
<td>29.2</td>
<td>60.4</td>
<td>76.3</td>
<td>57.6</td>
<td>53.1</td>
<td>41.1</td>
<td>54.0</td>
<td>42.8</td>
<td>76.5</td>
<td>62.3</td>
<td>84.5</td>
<td>35.3</td>
<td>41.3</td>
<td>50.1</td>
<td>77.6</td>
</tr>
<tr>
<td>INRIA.Genetic</td>
<td><strong>77.5</strong></td>
<td><strong>63.6</strong></td>
<td><strong>56.1</strong></td>
<td><strong>71.9</strong></td>
<td><strong>33.1</strong></td>
<td><strong>60.6</strong></td>
<td><strong>78.0</strong></td>
<td><strong>58.8</strong></td>
<td><strong>53.5</strong></td>
<td><strong>42.6</strong></td>
<td><strong>54.9</strong></td>
<td><strong>45.8</strong></td>
<td><strong>77.5</strong></td>
<td><strong>64.0</strong></td>
<td><strong>85.9</strong></td>
<td><strong>36.3</strong></td>
<td><strong>44.7</strong></td>
<td><strong>30.6</strong></td>
<td><strong>79.2</strong></td>
</tr>
<tr>
<td>INRIA.Larlus</td>
<td>62.6</td>
<td>54.0</td>
<td>32.8</td>
<td>47.5</td>
<td>17.8</td>
<td>46.4</td>
<td>69.6</td>
<td>44.2</td>
<td>44.6</td>
<td>26.0</td>
<td>38.1</td>
<td>34.0</td>
<td>66.0</td>
<td>55.1</td>
<td>77.2</td>
<td>13.1</td>
<td>29.1</td>
<td>36.7</td>
<td>62.7</td>
</tr>
<tr>
<td>MPI_BOW</td>
<td>58.9</td>
<td>46.0</td>
<td>31.3</td>
<td>59.0</td>
<td>16.9</td>
<td>40.5</td>
<td>67.2</td>
<td>40.2</td>
<td>44.3</td>
<td>28.3</td>
<td>31.9</td>
<td>34.4</td>
<td>63.6</td>
<td>53.5</td>
<td>75.7</td>
<td>22.3</td>
<td>26.6</td>
<td>35.4</td>
<td>60.6</td>
</tr>
<tr>
<td>PIPUVJA</td>
<td>48.6</td>
<td>20.9</td>
<td>21.3</td>
<td>17.2</td>
<td>6.4</td>
<td>14.2</td>
<td>45.0</td>
<td>31.4</td>
<td>27.4</td>
<td>12.3</td>
<td>14.3</td>
<td>23.7</td>
<td>30.1</td>
<td>13.3</td>
<td>62.0</td>
<td>10.0</td>
<td>12.4</td>
<td>13.3</td>
<td>26.7</td>
</tr>
<tr>
<td>QMUL_HSLS</td>
<td>70.6</td>
<td>54.8</td>
<td>35.7</td>
<td>64.5</td>
<td>27.8</td>
<td>51.1</td>
<td>71.4</td>
<td>54.0</td>
<td>46.6</td>
<td>36.6</td>
<td>34.4</td>
<td>39.9</td>
<td>71.5</td>
<td>55.4</td>
<td>80.6</td>
<td>15.8</td>
<td>35.8</td>
<td>41.5</td>
<td>73.1</td>
</tr>
<tr>
<td>QMUL_LSPCH</td>
<td>71.6</td>
<td>55.0</td>
<td>41.1</td>
<td>65.5</td>
<td>27.2</td>
<td>51.1</td>
<td>72.2</td>
<td>55.1</td>
<td>47.4</td>
<td>35.9</td>
<td>37.4</td>
<td>41.5</td>
<td>71.5</td>
<td>57.9</td>
<td>80.8</td>
<td>15.6</td>
<td>33.3</td>
<td>41.9</td>
<td>76.5</td>
</tr>
<tr>
<td>TKK</td>
<td>71.4</td>
<td>51.7</td>
<td>48.5</td>
<td>63.4</td>
<td>27.3</td>
<td>49.9</td>
<td>70.1</td>
<td>51.2</td>
<td>51.7</td>
<td>32.3</td>
<td>46.3</td>
<td>41.5</td>
<td>72.6</td>
<td>60.2</td>
<td>82.2</td>
<td>31.7</td>
<td>30.1</td>
<td>39.2</td>
<td>71.1</td>
</tr>
<tr>
<td>ToshCam_rdf</td>
<td>59.9</td>
<td>36.8</td>
<td>29.9</td>
<td>40.0</td>
<td>23.6</td>
<td>33.3</td>
<td>60.2</td>
<td>33.0</td>
<td>41.0</td>
<td>17.8</td>
<td>33.2</td>
<td>33.7</td>
<td>63.9</td>
<td>53.1</td>
<td>77.9</td>
<td>29.0</td>
<td>27.3</td>
<td>31.2</td>
<td>50.1</td>
</tr>
<tr>
<td>ToshCam_svm</td>
<td>54.0</td>
<td>27.1</td>
<td>30.3</td>
<td>35.6</td>
<td>17.0</td>
<td>22.3</td>
<td>58.0</td>
<td>34.6</td>
<td>38.0</td>
<td>19.0</td>
<td>27.5</td>
<td>32.4</td>
<td>48.0</td>
<td>40.7</td>
<td>78.1</td>
<td>23.4</td>
<td>21.8</td>
<td>28.0</td>
<td>45.5</td>
</tr>
<tr>
<td>Tsinghua</td>
<td>62.9</td>
<td>42.4</td>
<td>33.9</td>
<td>49.7</td>
<td>23.7</td>
<td>40.7</td>
<td>62.0</td>
<td>35.2</td>
<td>42.7</td>
<td>21.0</td>
<td>38.9</td>
<td>34.7</td>
<td>65.0</td>
<td>48.1</td>
<td>76.9</td>
<td>16.9</td>
<td>30.8</td>
<td>32.8</td>
<td>58.9</td>
</tr>
<tr>
<td>UVA_Bigrams</td>
<td>61.2</td>
<td>33.2</td>
<td>29.4</td>
<td>45.0</td>
<td>16.5</td>
<td>37.6</td>
<td>54.6</td>
<td>31.3</td>
<td>39.9</td>
<td>17.2</td>
<td>31.4</td>
<td>30.6</td>
<td>61.6</td>
<td>42.4</td>
<td>74.6</td>
<td>14.5</td>
<td>20.9</td>
<td>23.5</td>
<td>49.9</td>
</tr>
<tr>
<td>UVA_FuseAll</td>
<td>67.1</td>
<td>48.1</td>
<td>43.3</td>
<td>58.1</td>
<td>19.9</td>
<td>46.3</td>
<td>61.8</td>
<td>41.9</td>
<td>48.4</td>
<td>27.8</td>
<td>41.9</td>
<td>38.5</td>
<td>69.8</td>
<td>51.4</td>
<td>79.4</td>
<td>32.5</td>
<td>31.9</td>
<td>36.0</td>
<td>66.2</td>
</tr>
<tr>
<td>UVA_MCIP</td>
<td>66.5</td>
<td>47.9</td>
<td>41.0</td>
<td>58.0</td>
<td>16.8</td>
<td>44.0</td>
<td>61.2</td>
<td>40.5</td>
<td>48.5</td>
<td>27.8</td>
<td>41.7</td>
<td>37.1</td>
<td>66.4</td>
<td>50.1</td>
<td>78.6</td>
<td>31.2</td>
<td>32.3</td>
<td>31.9</td>
<td>66.6</td>
</tr>
<tr>
<td>UVA_SFS</td>
<td>66.3</td>
<td>49.7</td>
<td>43.5</td>
<td>60.7</td>
<td>18.8</td>
<td>44.9</td>
<td>64.8</td>
<td>41.9</td>
<td>46.8</td>
<td>24.9</td>
<td>42.3</td>
<td>33.9</td>
<td>71.5</td>
<td>53.4</td>
<td>80.4</td>
<td>29.7</td>
<td>31.2</td>
<td>31.8</td>
<td>67.4</td>
</tr>
<tr>
<td>UVA_WGT</td>
<td>59.7</td>
<td>33.7</td>
<td>34.9</td>
<td>44.5</td>
<td>22.2</td>
<td>32.9</td>
<td>55.9</td>
<td>36.3</td>
<td>36.8</td>
<td>20.6</td>
<td>25.2</td>
<td>34.7</td>
<td>65.1</td>
<td>40.1</td>
<td>74.2</td>
<td>26.4</td>
<td>26.9</td>
<td>25.1</td>
<td>50.7</td>
</tr>
<tr>
<td>XRCE</td>
<td>72.3</td>
<td>57.5</td>
<td>53.2</td>
<td>68.9</td>
<td>28.5</td>
<td>57.5</td>
<td>75.4</td>
<td>50.3</td>
<td>52.2</td>
<td>39.0</td>
<td>46.8</td>
<td>45.3</td>
<td>75.7</td>
<td>58.5</td>
<td>84.0</td>
<td>32.6</td>
<td>39.7</td>
<td><strong>50.9</strong></td>
<td>75.1</td>
</tr>
</tbody>
</table>

---

[Image]: Fraunhofer Heinrich Hertz Institute

MICCAI’18 Tutorial on Interpretable Machine Learning
## 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</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>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>cat</th>
<th>chair</th>
<th>cow</th>
<th>diningtable</th>
<th>dog</th>
<th>horse</th>
<th>motorbike</th>
</tr>
</thead>
<tbody>
<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>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>person</th>
<th>pottedplant</th>
<th>sheep</th>
<th>sofa</th>
<th>train</th>
<th>tvmonitor</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<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>

(Lapuschkin et al. 2016)
**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></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 DeepNet</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>Fisher 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 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>

*same performance —> same strategy?*  
*Lapuschkin et al. 2016*
## Application: Compare Classifiers

### Test error for various classes:

<table>
<thead>
<tr>
<th>Class</th>
<th>Fisher DeepNet</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></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>
<td></td>
</tr>
<tr>
<td></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>
<td></td>
</tr>
<tr>
<td>cat</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>81.60%</td>
<td></td>
</tr>
<tr>
<td>chair</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>
<td></td>
</tr>
<tr>
<td>cow</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></td>
<td></td>
</tr>
<tr>
<td>diningtable</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></td>
<td></td>
</tr>
<tr>
<td>dog</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>horse</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>motorbike</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>person</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pottedplant</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sheep</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sofa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>train</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tvmonitor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Same performance —> same strategy?**

*(Lapuschkin et al. 2016)*
### 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</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>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>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**

**FV**

**DNN**

same performance —> same strategy?  

(Lapuschkin et al. 2016)
Application: Compare Classifiers

‘horse’ images in PASCAL VOC 2007
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?

(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

- 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 & Improve Models
### Application: Face analysis

#### 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>52.1</td>
<td>54.3</td>
<td>89.1</td>
</tr>
<tr>
<td>[r]</td>
<td>51.9</td>
<td>52.3</td>
<td>53.3</td>
<td>89.9</td>
</tr>
<tr>
<td>[m]</td>
<td>53.6</td>
<td>54.3</td>
<td>56.2</td>
<td>90.7</td>
</tr>
<tr>
<td>[i,n]</td>
<td>–</td>
<td>51.6</td>
<td>56.2</td>
<td>90.9</td>
</tr>
<tr>
<td>[r,n]</td>
<td>–</td>
<td>52.1</td>
<td>57.4</td>
<td>91.9</td>
</tr>
<tr>
<td>[m,n]</td>
<td>–</td>
<td>52.8</td>
<td>58.5</td>
<td>92.6</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><strong>91.0</strong></td>
<td><strong>92.0</strong></td>
</tr>
<tr>
<td>[r,n]</td>
<td>–</td>
<td>90.6</td>
<td><strong>91.6</strong></td>
<td>–</td>
</tr>
<tr>
<td>[m,n]</td>
<td>–</td>
<td>90.6</td>
<td><strong>91.7</strong></td>
<td><strong>92.6</strong></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 & 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)

(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

- 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 × 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>

Fig. 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>complex morphs</th>
<th>multiclass</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>naive</td>
<td>left eye</td>
<td>right eye</td>
</tr>
<tr>
<td>left eye</td>
<td>0.84</td>
<td>0.00</td>
<td>0.02</td>
</tr>
<tr>
<td>right eye</td>
<td>0.00</td>
<td>0.91</td>
<td>0.05</td>
</tr>
<tr>
<td>nose</td>
<td>0.21</td>
<td>0.28</td>
<td>0.47</td>
</tr>
<tr>
<td>mouth</td>
<td>0.34</td>
<td>0.27</td>
<td>0.04</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

... some astronauts occasionally ...

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

Application: Learn new Representations

2D PCA projection of document vectors

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

Application of LRP
Interpreting Scientific Data
Application: EEG Analysis

Brain-Computer Interfacing

Neural network learns that:
Left hand movement imagination leads to desynchronization over right sensorimotor cortex (and vice versa).

(Sturm et al. 2016)
Application: EEG Analysis

Our neural networks are interpretable:
We can see for every trial “why” it is classified the way it is.

(Sturm et al. 2016)
Difficulty to apply deep learning to fMRI:
- high dimensional data (100,000 voxels), but only few subjects
- results must be interpretable (key in neuroscience)

Our approach:
- Recurrent neural networks (CNN + LSTM) for whole-brain analysis
- LRP allows to interpret the results

Dataset:
- 100 subjects from Human Connectome Project
- N-back task (faces, places, tools and body parts)

(Thomas et al. 2018)
### Application: fMRI Analysis

<table>
<thead>
<tr>
<th></th>
<th>A: Group</th>
<th>B: Subject</th>
<th>C: Trial</th>
<th>D: TR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Body</td>
<td><img src="image1" alt="Brain Images" /></td>
<td><img src="image2" alt="Brain Images" /></td>
<td><img src="image3" alt="Brain Images" /></td>
<td><img src="image4" alt="Brain Images" /></td>
</tr>
<tr>
<td>Faces</td>
<td><img src="image5" alt="Brain Images" /></td>
<td><img src="image6" alt="Brain Images" /></td>
<td><img src="image7" alt="Brain Images" /></td>
<td><img src="image8" alt="Brain Images" /></td>
</tr>
<tr>
<td>Places</td>
<td><img src="image9" alt="Brain Images" /></td>
<td><img src="image10" alt="Brain Images" /></td>
<td><img src="image11" alt="Brain Images" /></td>
<td><img src="image12" alt="Brain Images" /></td>
</tr>
<tr>
<td>Tools</td>
<td><img src="image13" alt="Brain Images" /></td>
<td><img src="image14" alt="Brain Images" /></td>
<td><img src="image15" alt="Brain Images" /></td>
<td><img src="image16" alt="Brain Images" /></td>
</tr>
</tbody>
</table>

*(Thomas et al. 2018)*
Application: Gait Analysis

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

(Horst et al. 2018)
Our approach:
- Classify & explain individual gait patterns
- Important for understanding diseases such as Parkinson

(Horst et al. 2018)
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)
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

Application: Understand the model

(Srinivasan et al. 2017)
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

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

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

(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

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

**Question**
- there is a metallic cube; are there any large cyan metallic objects behind it?

**LRP**
- there is a metallic cube; are there any large cyan metallic objects behind it?

---

The model understands the question and correctly identifies the object of interest.

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

<table>
<thead>
<tr>
<th>Question Type</th>
<th>LRP</th>
</tr>
</thead>
<tbody>
<tr>
<td>equal_size</td>
<td>size: 2279, the cylinder: 1939</td>
</tr>
<tr>
<td>greater_than</td>
<td>number: 2177, more than: 1656, 1515</td>
</tr>
<tr>
<td>query_color</td>
<td>color: 12819, what: 6837, is: 5735</td>
</tr>
<tr>
<td>count</td>
<td>what: 10823, how many: 9310, 5691</td>
</tr>
</tbody>
</table>

*(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 the model tracks the ball*

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

Sensitivity Analysis

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

LRP

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

After 0 epochs

After 25 epochs

After 195 epochs

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

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

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

(Lapuschnik et al., in prep.)

model learns
1. track the ball
2. focus on paddle
3. focus on the tunnel
Tutorial on Interpretable Machine Learning

Part 4: Case Study: Interpretable ML in Histopathology
Heatmapping - a quick case study in histopathology


Alexander Binder
Joint work with F. Klauschen, S. Lapuschkin (Bach), G. Montavon, K.-R. Müller, W. Samek

ISTD Pillar, Singapore University of Technology and Design (SUTD)

September 15, 2018
Deep Neural networks and (near-)human performance

Lipnet beats humans at lip reading

Human performance inGeneric classification

Human performance in low-res (!) traffic sign recognition

DeepStack outplays Humans in poker

Computer outplays Humans in DOOM

Mimicking art styles: https://deepart.io

Deep Learning tops human average on a constrained (!) reading comprehension task (SQuAD Dataset)
Human-like performance ≠ Human-like reasoning

Adversarial attacks against deep neural networks are easy.
Can explanation be a useful tool beyond mere curiosity?
Can explanation be a useful tool beyond mere curiosity?
Application Idea: Finding Biases in Your Training Data
Application Idea: Finding Biases in Your Training Data
Application Idea: Finding Biases in Your Training Data
Can explanation be a useful tool beyond mere curiosity?

- **BoW**: heatmapping for cancer evidence
- **BoW**: heatmapping for molecular expression evidence
- **Deep Learning**: heatmapping for looking for biases.
The next slides show authors own research. Views might be positively biased ;) . Nope I have not solved all interpretability problems with it.
BoW: Heatmapping for cancer evidence
BoW: heatmapping for cancer evidence

Why do we still talk about BoW?

- Good performance for small sample sizes (samples per class $< 10^3$).
- Stable against small changes in data augmentation / choices of negative sampling.
- Heatmapping slower compared to DNN+GPU+investigate
BoW: heatmapping for cancer evidence

Useful for ?
BoW: heatmapping for cancer evidence

Useful for ?
Shows cases where heatmapping works well

No stain normalization was used here – stability.

*Towards computational fluorescence microscopy: Machine learning-based integrated prediction of morphological and molecular tumor profiles*, Binder et al., arxiv 2018
Useful for?
Shows cases where heatmapping works well

Towards computational fluorescence microscopy: Machine learning-based integrated prediction of morphological and molecular tumor profiles, Binder et al., arxiv 2018
BoW: heatmapping for cancer evidence

Useful for?
Shows cases where heatmapping works well

Towards computational fluorescence microscopy: Machine learning-based integrated prediction of morphological and molecular tumor profiles, Binder et al., arxiv 2018
BoW: heatmapping for cancer evidence

Useful for ?
Shows cases where heatmapping works well
BoW: heatmapping for cancer evidence

Useful for?
Shows cases where heatmapping fails

Too similar to dense clusters of TiLs
BoW: heatmapping for cancer evidence

Useful for?

compare to:

Too similar to dense clusters of TiLs
BoW: heatmapping for cancer evidence

Useful for?
Shows cases where heatmapping fails

Too similar to lymphocytes, BoW feature does not capture that their distribution is untypical for lymphos
BoW: heatmapproing for cancer evidence

Useful for?
Shows cases where heatmapproing fails

Too similar to epithelial cells?? (patch-wise kernel similarity matrix may reveal this)
BoW: heatmapping for cancer evidence

Useful for?

Finding subtypes that are not recognized well, for example because undersampled in the training+testing set.

Potential solutions:

- improved sampling
- feature engineering (BoW)
- data augmentation engineering (deep learning)
Cancer is obvious. How about molecular properties?

Example: measurement of RNA in a biopsy sample. Can we localize evidence for the expression of the corresponding protein?

Example p53, a tumor suppressor molecule

No ad-hoc localization existent.

- forward pass: predict predict concentration from HE stain as a classification problem
- backward pass: find evidence localized to pixels
BoW: heatmapping for p53


Towards computational fluorescence microscopy: Machine learning-based integrated prediction of morphological and molecular tumor profiles, Binder et al., arxiv 2018
BoW: heatmapping for p53

Idea is not limited to heatmapping of evidence for cellular structures.

Accuracy is high on large subsets of patients:
BoW: heatmapping for X - How?

Forward pass is kernel machine over BoW feature.
Backward pass to obtain scores per pixel:

• Backpropagate from output of SVM $f(x)$ to kernel input dimension $x(d)$ of $x$
• Backpropagate from kernel input dimensions $x(d)$ to local features $l$ aggregated into the BoW feature
• Backpropagate from local feature $l$ to pixel $q$
BoW: heatmapping for X - How?

Backpropagate from output to kernel input dimension. Kernel is given as:

\[ f(x) = b + \sum_i a_i y_i k(z_i, x) \]  \hspace{1cm} (1)

goal: \( f(x) \approx \sum_d R_d^{(3)}(x) \), where

\( R_d^{(3)}(x) \) is the contribution of dimension \( d \) of the test feature \( x = (x_1, \ldots, x_D) \) to \( f(x) \).
BoW: heatmapping for X - How?

*Backpropagate from output to kernel input dimension.* Kernel is given as:

\[
f(x) = b + \sum_i a_i y_i k(z_i, x) \quad (3)
\]

**goal:** \[f(x) \approx \sum_d R_d^{(3)}(x) \quad (4)\]

In case of dimension-wise separable kernels, such as the HIK-kernel,

\[
k(z, x) = \sum_d \min(z(d), x(d)) \quad (5)
\]

\[
k(z, x) = \sum_d k_d(z(d), x(d)) \quad (6)
\]

\[
f(x) = b + \sum_i a_i y_i k(z_i, x) \quad (7)
\]

\[
= b + \sum_d \sum_i a_i y_i k_d(z(d), x(d)) \quad (8)
\]

\[
R_d^{(3)}(x) = \frac{b}{D} + \sum_i a_i y_i k_d(z(d), x(d)) \quad (9)
\]
BoW: heatmapping for X - How?

Backpropagate from output to kernel input dimension. Kernel is given as:

\[ f(x) = b + \sum_i a_i y_i k(z_i, x) \] (10)

goal: \[ f(x) \approx \sum_d R_d^{(3)}(x) \] (11)

In case of differentiable kernels, such as the $\chi^2$-kernel,

\[ k(z, x) = \exp(-\gamma \sum_{d: z(d) + x(d) > 0} \frac{(z(d) - x(d))^2}{z(d) + x(d)}) \] (12)

Taylor decomposition around a root \( f(x_0) = 0 \) is a way:

\[ f(x) \approx 0 + \sum_d (x(d) - x_{0,(d)}) \sum_i a_i y_i \frac{\partial k(z_i, x_0)}{\partial x_{0,(d)}} \] (13)
BoW: heatmapping for X - How?

Backpropagate from output to kernel input dimension. Kernel is given as:

\[ f(x) = b + \sum_i a_i y_i k(z_i, x) \]  \hspace{1cm} (14)

\[ \text{goal: } f(x) \approx \sum_d R_d^{(3)}(x) \]  \hspace{1cm} (15)

Taylor decomposition around a root \( f(x_0) = 0 \) is a way:

\[ f(x) \approx 0 + \sum_d (x(d) - x_0(d)) \sum_i a_i y_i \frac{\partial k(z_i, x_0)}{\partial x_0(d)} \]  \hspace{1cm} (16)

\[ R_d^{(3)}(x) = (x(d) - x_0(d)) \sum_i a_i y_i \frac{\partial k(z_i, x_0)}{\partial x_0(d)} \]  \hspace{1cm} (17)
Backpropagate from kernel input dimension to local feature. The Bow feature is a normalized sum of mappings $m_d(l)$ of local features $l$ onto visual word dimensions:

$$x_d = c \sum_l m_d(l)$$

(18)

One example mapping is the hard assignment onto the nearest visual word among the set of all visual words $\{w_d\}$:

$$m_d(l) = 1[d = \arg\min_{d'} \|l - w_d\|]$$

(19)

Note: not differentiable in $l$, so cannot use Taylor approximation again.
YOU GOT MATH PROBLEMS?

LET ME PUT ON MY THINKING CAT.
BoW: heatmapping for X - How?

*Backpropagate from kernel input dimension to local feature.* The Bow feature is a normalized sum of mappings $m_d(l)$ of local features $l$ onto visual word dimensions:

$$x_d = c \sum_l m_d(l) \tag{20}$$

Don't care how $m_d(l)$ looks like, apply special case of LRP-$\epsilon$-rule. That would look like:

$$R^{(2)}(l) = \sum_d R^{(3)}_d \frac{m_d(l)}{\sum_{l'} m_d(l')} \tag{21}$$

Have to take care for those dimensions $d$ without any weights:

$$\{d \mid \sum_l m_d(l) = 0\}$$
BoW: heatmapping for X - How?

Backpropagate from kernel input dimension to local feature.
The Bow feature is a normalized sum of mappings $m_d(l)$ of local features $l$ onto visual word dimensions:

$$x_d = c \sum_l m_d(l)$$  \hspace{1cm} (22)

Apply special case of LRP-$\varepsilon$-rule.
Have to take care for those dimensions $d$ without any weights:
$$\{ d \mid \sum_l m_d(l) = 0 \}$$

$$Z(x) = \{ d \mid \sum_l m_d(l) = 0 \}$$  \hspace{1cm} (23)

$$R^{(2)}(l) = \sum_{d \notin Z(x)} R^{(3)} \frac{m_d(l)}{\sum_{l'} m_d(l')} + \sum_{d \in Z(x)} R^{(3)} \frac{1}{\sum_{l'} 1}$$  \hspace{1cm} (24)
BoW: heatmapping for X - How?

Backpropagate from local feature to pixel

Simple idea: distribute relevance $R^{(2)}(l)$ of a local feature $l$ equally over the support pixels $q$, used to compute the same local feature.

$$LF(q) = \{ l \mid q \in supp(l) \} : \text{local features which touch } q$$ (25)

$$R^{(1)}(q) = \sum_{l \in LF(q)} \frac{R^{(2)}(l)}{|supp(l)|}$$ (26)
Changes in the work flow due to deep learning

- Feature engineering is dead. Learn your features from data.
- No need for tuning of features by hand.
Changes in the work flow due to deep learning

- Feature engineering is dead. Learn your features from data.
- No need for tuning of features by hand.
- Long live data augmentation engineering
- Long live hyperparameter search over grids of 37 different parameters.
Data Augmentation Engineering

- brightness
- contrast
- color
- distortion

This are 6 parameters here already for hyperparameter search here.
Data Augmentation Engineering

Many hyperparameters +

- batch size
- minibatch structure
- initial learning rate
- learning rate decay
- optimizer (SGD, momentum, ADAM)

Hyperparameter search in 20-dim space
How to sample elements in minibatches?
Histopathology in research phase: the inevitable problem of biases

Given a prediction target - example: find evidence for cancer cells.

- If a subclass is undersampled, poor performance on it cannot be detected, because it is not represented in the test set.
- Extreme high variability of prediction target and of background. What are relevant subclasses?
- A relevant subclass from positive or negative labeled structures possibly undersampled, and we dont know it!
Histopathology in research phase: the inevitable problem of biases

If a subclass is undersampled, poor performance on it cannot be detected, because it is not represented in the test set.

Leads to a design problem:
- how to sample positive regions for annotation? (how much of certain structures need to be sampled?)
- how to sample negative regions for annotation?

Hypothesis: Heatmapping over large test slides may reveal undersampled structures in a qualitative manner and help in the iterative solution of the design problem.
setup:

- HE stain, breast cancer
- positive annotations: positions of cancer nuclei
- negative annotations: ???
Batch composition

orig

1:1

1.5:1

2:1

62.5 %
Batch composition

orig

1.5:1

64.1%

1:1

2:1
Batch composition

orig 1:1

1.5:1 2:1

65.6 %
Batch composition

orig 1:1

1.5:1

2:1

67.2 %
Batch composition

orig 1:1

1.5:1

2:1

68.8 %
Batch composition

orig 1:1

1.5:1 2:1

70.3%
Batch composition

orig 1:1

1.5:1

2:1

71.9 %
Batch composition

orig

1.5:1

2:1

75%
Batch composition

observation: different bias learned, inconsistent to ratio
setup:

- HE stain, breast cancer
- positive annotations: positions of cancer nuclei
- negative annotations: (overlapping) windows without cancer nuclei
- preprocessing: shrink image to 80%, patchsize 120, grid stride 20
- Densenet 121, batchsize 8
- LRP-$\epsilon$ for FC layers, LRP-$\beta = 0$ for all others
- innvestigate toolbox with neuron selection index for cancer
Impact of Scaling

orig 80%

100% 66%
Impact of Scaling

orig 80% 100% 66%

100% 66%
Impact of Scaling

orig 80%

100% 66%
Impact of Scaling
Impact of Scaling

orig 80%

100% 66%
Impact of Scaling

orig 80%

100% 66%
Impact of Scaling

observation: 100% scaling: nuclei are too large for the fixed kernel sizes, difficult to recognize cancer, too faint heatmaps
Histopathology in application phase: heatmapping for acceptance

A classifier that simply tells a clinician: "its grade 3"
Histopathology in application phase: heatmapping for acceptance

A classifier that simply tells a clinician: "its grade 3"

Problem: in case of doubt clinician cannot validate the prediction. Classifier mistaken or clinician overlooked something?

Heatmapping allows to point the clinician to relevant regions.
A classifier that simply tells a clinician: ”its grade 3”

Problem: in case of doubt clinician cannot validate the prediction. Classifier mistaken or clinician overlooked something?

Heatmapping allows to point the clinician to relevant regions. Heatmapping allows to identify nonsensical predictions on outlier samples.
Applications III: How well does LRP scale?

**DenseNet-121**

Made with Keras:

https://github.com/albermax/innvestigate

<table>
<thead>
<tr>
<th>Input</th>
<th>Gradient</th>
<th>SmoothGrad</th>
<th>Integrated Gradients</th>
<th>Deconvnet</th>
<th>Guided Backprop</th>
<th>PatternNet</th>
<th>Gradient*Input</th>
<th>LRP-Z</th>
<th>LRP-Epsilon</th>
<th>LRP-PresetAFlat</th>
<th>LRP-PresetBFlat</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="broom" /></td>
<td><img src="image2" alt="gradient" /></td>
<td><img src="image3" alt="smoothgrad" /></td>
<td><img src="image4" alt="integrated_gradients" /></td>
<td><img src="image5" alt="deconvnet" /></td>
<td><img src="image6" alt="guided_backprop" /></td>
<td><img src="image7" alt="patternnet" /></td>
<td><img src="image8" alt="gradient_input" /></td>
<td><img src="image9" alt="lrp_z" /></td>
<td><img src="image10" alt="lrp_epsilon" /></td>
<td><img src="image11" alt="lrp_preset_aflat" /></td>
<td><img src="image12" alt="lrp_preset_bflat" /></td>
</tr>
<tr>
<td>label: broom</td>
<td>pred: broom</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>logit: 20.55</td>
<td>prob: 1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image13" alt="submarine" /></td>
<td><img src="image14" alt="gradient" /></td>
<td><img src="image15" alt="smoothgrad" /></td>
<td><img src="image16" alt="integrated_gradients" /></td>
<td><img src="image17" alt="deconvnet" /></td>
<td><img src="image18" alt="guided_backprop" /></td>
<td><img src="image19" alt="patternnet" /></td>
<td><img src="image20" alt="gradient_input" /></td>
<td><img src="image21" alt="lrp_z" /></td>
<td><img src="image22" alt="lrp_epsilon" /></td>
<td><img src="image23" alt="lrp_preset_aflat" /></td>
<td><img src="image24" alt="lrp_preset_bflat" /></td>
</tr>
<tr>
<td>label: submarine</td>
<td>pred: submarine</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>logit: 26.18</td>
<td>prob: 1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image25" alt="Dungeness crab" /></td>
<td><img src="image26" alt="gradient" /></td>
<td><img src="image27" alt="smoothgrad" /></td>
<td><img src="image28" alt="integrated_gradients" /></td>
<td><img src="image29" alt="deconvnet" /></td>
<td><img src="image30" alt="guided_backprop" /></td>
<td><img src="image31" alt="patternnet" /></td>
<td><img src="image32" alt="gradient_input" /></td>
<td><img src="image33" alt="lrp_z" /></td>
<td><img src="image34" alt="lrp_epsilon" /></td>
<td><img src="image35" alt="lrp_preset_aflat" /></td>
<td><img src="image36" alt="lrp_preset_bflat" /></td>
</tr>
<tr>
<td>label: Dungeness crab</td>
<td>pred: king crab</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>logit: 20.99</td>
<td>prob: 0.49</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image37" alt="ice lolly" /></td>
<td><img src="image38" alt="gradient" /></td>
<td><img src="image39" alt="smoothgrad" /></td>
<td><img src="image40" alt="integrated_gradients" /></td>
<td><img src="image41" alt="deconvnet" /></td>
<td><img src="image42" alt="guided_backprop" /></td>
<td><img src="image43" alt="patternnet" /></td>
<td><img src="image44" alt="gradient_input" /></td>
<td><img src="image45" alt="lrp_z" /></td>
<td><img src="image46" alt="lrp_epsilon" /></td>
<td><img src="image47" alt="lrp_preset_aflat" /></td>
<td><img src="image48" alt="lrp_preset_bflat" /></td>
</tr>
<tr>
<td>label: ice lolly</td>
<td>pred: ice lolly</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>logit: 14.99</td>
<td>prob: 0.53</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image49" alt="abaya" /></td>
<td><img src="image50" alt="gradient" /></td>
<td><img src="image51" alt="smoothgrad" /></td>
<td><img src="image52" alt="integrated_gradients" /></td>
<td><img src="image53" alt="deconvnet" /></td>
<td><img src="image54" alt="guided_backprop" /></td>
<td><img src="image55" alt="patternnet" /></td>
<td><img src="image56" alt="gradient_input" /></td>
<td><img src="image57" alt="lrp_z" /></td>
<td><img src="image58" alt="lrp_epsilon" /></td>
<td><img src="image59" alt="lrp_preset_aflat" /></td>
<td><img src="image60" alt="lrp_preset_bflat" /></td>
</tr>
<tr>
<td>label: abaya</td>
<td>pred: abaya</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>logit: 20.37</td>
<td>prob: 0.99</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image61" alt="baseball" /></td>
<td><img src="image62" alt="gradient" /></td>
<td><img src="image63" alt="smoothgrad" /></td>
<td><img src="image64" alt="integrated_gradients" /></td>
<td><img src="image65" alt="deconvnet" /></td>
<td><img src="image66" alt="guided_backprop" /></td>
<td><img src="image67" alt="patternnet" /></td>
<td><img src="image68" alt="gradient_input" /></td>
<td><img src="image69" alt="lrp_z" /></td>
<td><img src="image70" alt="lrp_epsilon" /></td>
<td><img src="image71" alt="lrp_preset_aflat" /></td>
<td><img src="image72" alt="lrp_preset_bflat" /></td>
</tr>
<tr>
<td>label: baseball</td>
<td>pred: baseball</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>logit: 18.82</td>
<td>prob: 0.96</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image73" alt="bell pepper" /></td>
<td><img src="image74" alt="gradient" /></td>
<td><img src="image75" alt="smoothgrad" /></td>
<td><img src="image76" alt="integrated_gradients" /></td>
<td><img src="image77" alt="deconvnet" /></td>
<td><img src="image78" alt="guided_backprop" /></td>
<td><img src="image79" alt="patternnet" /></td>
<td><img src="image80" alt="gradient_input" /></td>
<td><img src="image81" alt="lrp_z" /></td>
<td><img src="image82" alt="lrp_epsilon" /></td>
<td><img src="image83" alt="lrp_preset_aflat" /></td>
<td><img src="image84" alt="lrp_preset_bflat" /></td>
</tr>
<tr>
<td>label: bell pepper</td>
<td>pred: bell pepper</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>logit: 24.35</td>
<td>prob: 0.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Thank you!

Links (LRP for LSTM for example):
http://www.heatmapping.org/
Tutorial: http://www.heatmapping.org/tutorial/
for Keras: https://github.com/albermax/innvestigate
LRPToolbox:
https://github.com/sebastian-lapuschkin/lrp_toolbox
Experimental MXnet integration:
https://github.com/sebastian-lapuschkin/lrp_toolbox/tree/python-wip/python
Demos: https://lrpserver.hhi.fraunhofer.de/
Tutorial on Interpretable Machine Learning

Wrap-up
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.

gradient-based methods

vulnerable to shattered gradients

Our LRP method is robust to this.
Take Home Messages

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

<table>
<thead>
<tr>
<th>A: Group</th>
<th>B: Subject</th>
<th>C: Trial</th>
<th>D: TR</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image A" /></td>
<td><img src="image2.png" alt="Image B" /></td>
<td><img src="image3.png" alt="Image C" /></td>
<td><img src="image4.png" alt="Image D" /></td>
</tr>
</tbody>
</table>

Fraunhofer Heinrich Hertz Institute
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 Sciences**


**Software**
