Tutorial on Interpretable Machine Learning

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

9:00 - 9:30  Part 1a: Introduction
9:30 - 10:30 Part 1b: Making Deep Neural Networks Transparent
10:30 - 11:00 Coffee Break
11:00 - 11:30 Part 2a: Making Deep Neural Networks Transparent
11:30 - 12:30 Part 2b: Applications & Discussion
Before we start

We thank our collaborators!

Grégoire Montavon (TU Berlin)
Alexander Binder (SUTD)
Sebastian Lapuschkin (Fraunhofer HHI)
Leila Arras (Fraunhofer HHI)
...

http://interpretable-ml.org/

Please ask questions at any time!

NIPS’17 Workshop “Interpreting, Explaining and Visualizing Deep Learning - Now what?”
Tutorial on Interpretable Machine Learning

W. Samek & K.-R. Müller

Part 1a: Introduction
Recent ML Systems achieve superhuman Performance

- AlphaGo beats Go human champ
- Deep Net outperforms humans in image classification
- Autonomous search-and-rescue drones outperform humans
- DeepStack beats professional poker players
- Computer out-plays humans in "doom"
- IBM's Watson destroys humans in jeopardy
- Deep Net beats human at recognizing traffic signs
From Data to Information

Huge volumes of data

Computing power

Deep Nets / Kernel Machines / …

Solve task

Information (implicit)
From Data to Information

Huge volumes of data

Computing power

Deep Nets / Kernel Machines / …

Interpretable Information

Solve task

Information (implicit)
From Data to Information
From Data to Information

IMAGENET

AlexNet (16.4%) → Performance

Data → Information
From Data to Information

IMAGENET

AlexNet (16.4%) Clarifai (11.1%)

Data → Information

Performance
From Data to Information

<table>
<thead>
<tr>
<th>Model</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlexNet</td>
<td>16.4%</td>
</tr>
<tr>
<td>Clarifai</td>
<td>11.1%</td>
</tr>
<tr>
<td>VGG</td>
<td>7.3%</td>
</tr>
</tbody>
</table>

Data → Information
From Data to Information

**IMAGENET**

- AlexNet (16.4%)
- Clarifai (11.1%)
- VGG (7.3%)
- GoogleNet (6.7%)

**Performance**

Data → Information
From Data to Information

IMAGENET

AlexNet (16.4%)  Clarifai (11.1%)  VGG (7.3%)  GoogleNet (6.7%)  ResNet (3.57%)

Performance

Data  Information
From Data to Information

**IMAGENET**

![Diagram showing performance and interpretability of different models]

- **AlexNet** (16.4%)
- **Clarifai** (11.1%)
- **VGG** (7.3%)
- **GoogleNet** (6.7%)
- **ResNet** (3.57%)

Data → Information
From Data to Information

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

Interpretability

Data → Information → Interpretable for human

Crucial in many applications (industry, sciences …)
Interpretable vs. Powerful Models?

Linear model

Poor fit, but easily interpretable
“global explanation”

Non-linear model

Can be very complex
“individual explanation”
Interpretable vs. Powerful Models?

Linear model

Non-linear model

- Poor fit, but easily interpretable
- Can be very complex

60 million parameters
650,000 neurons
Interpretable vs. Powerful Models?

Linear model

Non-linear model

60 million parameters
650,000 neurons

We have techniques to interpret and explain such complex models!
Interpretable vs. Powerful Models?

**Ante-hoc interpretability:**

Choose a model that is readily interpretable and train it.

*Example:*

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

Is the model expressive enough to predict the data?

**Post-hoc interpretability:**

Choose a model that works well in practice, and develop a special technique to interpret it.

*Example:*

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

How to determine the contribution each input variable?
Dimensions of Interpretability

Different dimensions of “interpretability”

- Data
- Prediction
- Model
Dimensions of Interpretability

Different dimensions of “interpretability”

“Explain why a certain pattern \( x \) has been classified in a certain way \( f(x) \).”

prediction

model

data
Dimensions of Interpretability

Different dimensions of “interpretability”

- **Data**
- **Prediction**
- **Model**

**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.”
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

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.”
Why Interpretability?

1) Verify that classifier works as expected

Wrong decisions can be costly and dangerous

“Autonomous car crashes, because it wrongly recognizes …”

“AI medical diagnosis system misclassifies patient’s disease …”
Why Interpretability?

2) Improve classifier

Standard ML

Interpretable ML

Generalization error

Generalization error + human experience
Why Interpretability?

3) Learn from the learning machine

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

Old promise:
“Learn about the human brain.”
Why Interpretability?

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 → “right to explanation”

Retain human decision in order to assign responsibility.

“With interpretability we can ensure that ML models work in compliance to proposed legislation.”
Techniques of Interpretation

DNN transparency

- interpreting models
  - activation maximization
    - Berkes 2006
    - Erhan 2010
    - Simonyan 2013
    - Nguyen 2015/16
  - data generation
    - Hinton 2006
    - Goodfellow 2014
    - v. den Oord 2016
    - Nguyen 2016

- explaining decisions
  - sensitivity analysis
    - Khan 2001
    - Gevrey 2003
    - Baehrens 2010
    - Simonyan 2013
  - decomposition
    - Poulin 2006
    - Landecker 2013
    - Bach 2015
    - Montavon 2017

focus on model  focus on data
Techniques of Interpretation

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

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

better understand internal representation

crucial for many practical applications
Techniques of Interpretation

In medical context

• Population view (ensemble)
  • Which symptoms are most common for the disease
  • Which drugs are most helpful for patients

• Patient’s view (individual)
  • Which particular symptoms does the patient have
  • Which drugs does he need to take in order to recover

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

Interpreting models

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

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

Interpreting models

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

simple regularizer
(Simonyan et al. 2013)

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

Interpreting models

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

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

complex regularizer (Nguyen et al. 2016)
Techniques of Interpretation

Explaining decisions

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

data → ML blackbox → decision

it’s a shark
Techniques of Interpretation

Explaining decisions

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

\[ \text{data} \xrightarrow{\text{ML blackbox}} f(x) \xrightarrow{\text{decision}} \text{it’s a shark} \]

\[ x \xrightarrow{\text{explanation}} R(x) \]
Techniques of Interpretation

Explaining decisions

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

data \( x \) \[ \rightarrow \] ML blackbox \[ \rightarrow f(x) \] decision

\[ \text{it’s a shark} \]

explanation

- Sensitivity Analysis
- Layer-wise Relevance Propagation (LRP)

\( R(x) \)
Techniques of Interpretation

Sensitivity Analysis
(Simonyan et al. 2014)

Classifier

Bird

prediction \( f(\mathbf{x}) \)

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

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

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

input $x$

Classifier

Rooster

prediction $f(x)$

Explain prediction

(how much each pixel contributes to prediction)

Idea: Decompose function

$$\sum_i R_i = f(x)$$

heatmap

redistribute $f(x)$
Techniques of Interpretation

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

"every neuron gets it’s share of relevance depending on activation and strength of connection."

Idea: Decompose function

\[ \sum_i R_i = f(x) \]
Techniques of Interpretation

Classification

\[ x_j = \sigma(\sum_i x_i w_{ij} + b_j) \]
Techniques of Interpretation

Explanation

Idea: Backpropagate "relevances"

Initialization

$$R_j^{(l+1)} = f(x)$$
Techniques of Interpretation

Explanation

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.
Techniques of Interpretation

Explanation

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

\[ R_i^{(l)} = \sum_j (\alpha \cdot \frac{(x_i \cdot w_{ij})^+}{\sum_{i'} (x_{i'} \cdot w_{i'j})^+} + \beta \cdot \frac{(x_i \cdot w_{ij})^-}{\sum_{i'} (x_{i'} \cdot w_{i'j})^-}) R_j^{(l+1)} \]

where \( \alpha + \beta = 1 \)
Techniques of Interpretation

Explanation

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

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

where \( \alpha + \beta = 1 \)

special case
\( \alpha = 1, \beta = 0 \)
Explanation

**Techniques of Interpretation**

Equivalence to redistribution rule proposed in Excitation Backprop (Zhang et al., 2016)

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

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

where \( \alpha + \beta = 1 \)
Techniques of Interpretation

Explanation

Theoretical interpretation
Deep Taylor Decomposition
(Montavon et al., 2017)

**special case**
$$\alpha = 1, \beta = 0$$

Equivalent to redistribution rule proposed in
Excitation Backprop (Zhang et al., 2016)

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

$$R_{i}^{(l)} = \sum_{j} \left( \alpha \cdot \frac{(x_{i} \cdot w_{i,j})^+}{\sum_{i'}(x_{i'} \cdot w_{i',j})^+} + \beta \cdot \frac{(x_{i} \cdot w_{i,j})^-}{\sum_{i'}(x_{i'} \cdot w_{i',j})^-} \right) R_{j}^{(l+1)}$$

where $\alpha + \beta = 1$
Techniques of Interpretation

Explanation

Layer-wise relevance conservation

\[ \sum_i R_i = \ldots = \sum_i R_i^{(l)} = \sum_j R_j^{(l+1)} = \ldots = f(x) \]
More to come

Part 1b

- activation maximization
  - Berkes 2006
  - Erhan 2010
  - Simonyan 2013
  - Nguyen 2015/16

- data generation
  - Hinton 2006
  - Goodfellow 2014
  - v. den Oord 2016
  - Nguyen 2016

Part 2a

- explaining decisions
  - Khan 2001
  - Gevrey 2003
  - Baehrens 2010
  - Simonyan 2013

  - decomposition
    - Poulin 2006
    - Landecker 2013
    - Bach 2015
    - Montavon 2017

Part 2b

quality of explanations, applications, interpretability in the sciences, discussion

focus on model

focus on data
Tutorial on Interpretable Machine Learning

W. Samek & K.-R. Müller

Part 1b: Making Deep Neural Networks Transparent
Making Deep Neural Nets Transparent

model analysis

- visualizing filters
- max. class activation

f(\mathcal{X})

decision analysis

- include distribution (RBM, DGN, etc.)

- sensitivity analysis
- decomposition
Interpreting Classes and Outputs

Image classification:

GoogleNet

"motorbike"

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

Quantum chemical calculations:

GDB-7

Question: How to interpret “α high” in terms of molecular geometry?
The Activation Maximization (AM) Method

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

\[ \text{input pattern} \xrightarrow{} \text{deep neural network} \xrightarrow{\log p(\omega_c|\mathbf{x})} \text{class probability} \xrightarrow{f(\mathbf{x})} \text{real-valued output} \]

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

![Image of handwritten digits]

**Interpretation for $\omega_c$**

$\log p(\omega_d | x)$

class probability

**Initial solutions**

- Optimizing $\max_x p(\omega_c | x)$

**Converged solutions $x^*$**

![Graph showing convergence over iterations]
Interpreting a DNN Image Classifier


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

Activation-maximization produces class-related patterns, but they are not resembling true data points. This can lower the quality of the interpretation for the predicted class $\omega_c$.

Idea:

- Force the interpretation $x^*$ to match the data more closely.

This can be achieved by redefining the optimization problem:

Find the input pattern that maximizes class probability. $\rightarrow$ Find the most likely input pattern for a given class.
Improving Activation Maximization

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

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

Nguyen et al. 2016 introduced several enhancements for activation maximization:

- Multiplying the objective by an expert $p(x)$:
  \[
  p(x|\omega_c) \propto \underbrace{p(\omega_c|x) \cdot p(x)}_{\text{old}}
  \]

- Optimization in code space:
  \[
  \max_{z \in \mathcal{Z}} p(\omega_c|g(z)) + \lambda \|z\|^2 \quad x^* = g(z^*)
  \]

These two techniques require an unsupervised model of the data, either a density model $p(x)$ or a generator $g(z)$. 
Improving Activation Maximization

- AM + density
  - optimum has clear meaning
  - objective can be hard to optimize

- AM + generator
  - more straightforward to optimize
  - not optimizing $\log p(x|\omega_c)$

- density model
  - $\log p(x)$

- discriminative model
  - $\log p(\omega_c|x)$

- generator
  - $x = g(z)$

- generative model
  - $\log p(\omega_c|x)$
Comparison of Activation Maximization Variants

simple AM (initialized to mean)

simple AM (init. to class means)

AM-density (init. to class means)

AM-gen (init. to class means)

Observation: Connecting to the data leads to sharper prototypes.
Enhanced AM on Natural Images

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.
Summary: Interpreting models

- Deep neural networks can be interpreted by finding input patterns that maximize a certain output quantity (e.g. class probability).
- Connecting to the data (e.g. by adding a generative or density model) improves the interpretability of the solution.
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.
From Prototypes to Individual Explanations

Finding a prototype:

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

Individual explanation:

Question: Why is this example classified as a motorbike?
From Prototypes to Individual Explanations

Finding a prototype:

GDB-7

\{ \text{image} \}

Question: How to interpret “\(\alpha\) high” in terms of molecular geometry?

Individual explanation:

GDB-7

\{ \text{image} \}

Question: Why \(\alpha\) has a certain value for this molecule?
From Prototypes to Individual Explanations

Other examples where individual explanations are preferable to global interpretations:

- **Brain-computer interfaces**: Analyze input data for a *given* user at a *given* time in a *given* environment.

- **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.
From Prototypes to Individual Explanations

- visualizing filters
- max. class activation

- include distribution (RBM, DGN, etc.)

- sensitivity analysis
- decomposition
From Prototypes to Individual Explanations

**Goal:** Determine the relevance of each input variable for a given decision $f(x_1, x_2, \ldots, x_d)$, by assigning to these variables *relevance* scores $R_1, R_2, \ldots, R_d$. 
Basic Technique: Sensitivity Analysis

Consider a function $f$, a data point $\mathbf{x} = (x_1, \ldots, x_d)$, and the prediction

$$f(x_1, \ldots, x_d).$$

Sensitivity analysis measures the local variation of the function along each input dimension

$$R_i = \left( \frac{\partial f}{\partial x_i} \bigg|_{x=x} \right)^2$$

Remarks:

- Easy to implement (we only need access to the gradient of the decision function).
- But does it really explain the prediction?
Explaining by Decomposing

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

**Examples:**

- Economic activity (e.g. petroleum, cars, medicaments, ...)
- Energy production (e.g. coal, nuclear, hydraulic, ...)
- Evidence for object in an image (e.g. pixel 1, pixel 2, pixel 3, ...)
- Evidence for meaning in a text (e.g. word 1, word 2, word 3, ...)
What Does Sensitivity Analysis Decompose?

Sensitivity analysis

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

is a decomposition of the gradient norm \( \| \nabla_x f \|^2 \).

**Proof:** \[ \sum_i R_i = \| \nabla_x f \|^2 \]

Sensitivity analysis explains a *variation* of the function, not the function value itself.
What Does Sensitivity Analysis Decompose?

Example: Sensitivity for class “car”

- Relevant pixels are found both on cars and on the background.
- Explains what reduces/increases the evidence for cars rather what is the evidence for cars.
Decomposing the Correct Quantity

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

**Candidate:** Taylor decomposition

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

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

\[ \tilde{x} = \lim_{\varepsilon \to 0} \varepsilon \cdot x \approx 0. \]
Experiment on a Randomly Initialized DNN

\[
x_1 \rightarrow 500 \rightarrow 500 \rightarrow 500 \rightarrow 500 \rightarrow f(x)
\]
Decomposing the Output of the DNN

\[ R_i = \left. \frac{\partial f}{\partial x_i} \right|_{\tilde{x}} \cdot (x_i - \tilde{x}_i) \]

\[ R_1(x) = +0.21 \]
\[ R_2(x) = +0.16 \]
Decomposing the Output of the DNN

\[ R_i = \left. \frac{\partial f}{\partial x_i} \right|_{x = \tilde{x}} \cdot (x_i - \tilde{x}_i) \]

⇒ “Naive” Taylor decomposition
Decomposing the Output of the DNN

Advantages
- Decomposes the desired quantity $f(x)$ in a principled way.

Disadvantages
- Relevance functions are highly non-smooth.
- Relevance scores are sometimes negative.
- Inflexible w.r.t. the model.
Experiment on Handwritten Digits

Data to classify:

2 1 0 6 8

3-layer MLP:
Sensitivity analysis

Naive Taylor ($\widetilde{x} = 0$)

6-layer CNN:
Sensitivity analysis

Naive Taylor ($\widetilde{x} = 0$)

Observation: Both analyses produce noisy explanations of the MLP and CNN predictions.
Experiment on BVLC CaffeNet

Observation: Explanations are noisy and (over/under)represent certain regions of the image.
Explaining DNN Predictions

- Standard methods (sensitivity analysis, naive Taylor decomposition) are subject to gradient noise and do not work well on deep neural networks.

DNN predictions need more advanced explanation methods.
Tutorial on Interpretable Machine Learning
W. Samek & K.-R. Müller

Part 2a: Making Deep Neural Networks Transparent
From Prototypes to Individual Explanations

- visualizing filters
- max. class activation
- include distribution (RBM, DGN, etc.)
- sensitivity analysis
- decomposition
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
Gradient Shattering

**Structure's view**

- Input \( x \) to output \( f(x) \) through layers.

**Function's view (cartoon)**

- Shallow function representations vs. deep function representations.

\[ f(x) \]

\[ x \]
Sensitivity and Simple Taylor

Gradient-based methods usually do not work for explaining deep nets.
Deep Taylor Decomposition

**Key Idea:** If a decision is too complex to explain, break the decision function into sub-functions, and explain each sub-decision separately.

1. Decompose decision function
2. Explain subfunctions
3. Aggregate explanations
Deep Taylor Decomposition

Taylor decomposition (TD)

\[ f(x), \nabla f, \ldots \]

\[ f(x) = \nabla f \bigg|_{x=\tilde{x}}^{\top} \cdot (x - \tilde{x}) + \varepsilon \]

\[ f(x) = R_1 + R_2 + \varepsilon \]

deep Taylor decomposition (DTD)

\[ f(x) = h_1 + h_2 + h_3 \]

\[ f(x) = R_1 + R_2 \]
Deep Taylor Decomposition

1. forward computation

2. relevance propagation
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
Deep Taylor Decomposition / LRP

\[ R_j = \sum_k \left( \alpha \frac{a_j w_{jk}^+}{\sum_j a_j w_{jk}^-} - \beta \frac{a_j w_{jk}^-}{\sum_j a_j w_{jk}^-} \right) R_k \]

intuition [Bach’15]

Relevance should be redistributed to the lower-layer neurons \((a_j)_j\) in proportion to their excitatory effect on \(a_k\). “Counter-relevance” should be redistributed to the lower-layer neurons \((a_j)_j\) in proportion to their inhibitory effect on \(a_k\).

analysis [Montavon’17]

For the specific case \(\alpha = 1\), the whole LRP procedure can be seen as a *deep Taylor decomposition* of the neural network function.
Deep Taylor Decomposition / LRP

Simple Taylor

\[ R_i = \frac{\partial f}{\partial x_i} \cdot (x_i - 0) \]

Deep Taylor

\[ R_i = \frac{\partial R_k}{\partial a_j} \cdot (a_j - \tilde{a}_j) \]
Deep Taylor Decomposition / LRP

**step 1:** forward pass (linear time)

**step 2:** relevance propagation

also linear time!

Propagation rule:

\[ R_i = \sum_j q_{ij} R_j \quad \sum_i q_{ij} = 1 \]

Various rules are available for pixel layers, intermediate layers, or special layers.
Deep Taylor Decomposition / LRP

Observation: Only deep Taylor LRP focuses on cars.
Deep Taylor Decomposition / LRP

Two similar images should have similar explanations.
Technical Details

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

\[ R_j = a_j c_j \]

\[ R_i = a_i \sum_j w_{ij}^+ \frac{\max(0, \sum_i a_i w_{ij})}{\sum_i a_i w_{ij}^+} c_j \]

\[ R_i = a_i c_i \]

Relevance has product structure at all layers.
Technical Details

1

Build the Relevance Neuron

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

\[ w'_{ij} = w_{ij} c_j \]
Technical Details

2. Expand the Relevance Neuron

\[ R_j((a_i)_i) = R_j((\tilde{a}_i)_i) + \sum_i \frac{\partial R_j}{\partial a_i}_{(\tilde{a}_i)_i} \cdot (a_i - \tilde{a}_i) + \varepsilon \]
Technical Details

3 Decompose Relevance

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

Closed-form solution:

\[ R_{i \leftarrow j} = \frac{(a_i - \tilde{a}_{i(j)})w_{ij}}{\sum_i (a_i - \tilde{a}_{i(j)})w_{ij}} R_j \]
Technical Details

Closed-form solution:

$$R_{i \leftarrow j} = \frac{(a_i - \tilde{a}_i^{(j)}) w_{ij}}{\sum_i (a_i - \tilde{a}_i^{(j)}) w_{ij}} R_j$$

LRP rule [Bach15, Zhang16]

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

4. Pooling relevance over all outgoing neurons

5. Search root point along specific direction:

$$\left( a_i - \tilde{a}_i^{(j)} \right) \propto a_i 1_{w_{ij}^+ > 0}$$

But what does that mean?
Technical Details

**case 1**

- \((a_i - \tilde{a}_i^{(j)}) \propto w_{ij}'\)
- \((a_i - \tilde{a}_i^{(j)}) \propto a_i\)
- \((a_i - \tilde{a}_i^{(j)}) \propto a_i 1_{w_{ij}' > 0}\)

**case 2**

- nearest root
- origin (like simple Taylor)
- root along positive activations (LRP rule)
## Technical Details

<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>
</tbody>
</table>
| Pixel intensities \((x_i \in [l_i, h_i],\)  \\
l_i \leq 0 \leq h_i\)       | \(R_i = \sum_j \frac{x_i w_{ij} - l_i w_{ij}^+ - h_i w_{ij}^-}{\sum_i x_i w_{ij} - l_i w_{ij}^+ - h_i w_{ij}^-} R_j\) |
| Real values \((x_i \in \mathbb{R})\)   | \(R_i = \sum_j \frac{w_{ij}^2}{\sum_i w_{ij}^2} R_j\)               |

Deep Taylor LRP rules [Montavon’17]

More refined rules can also be constructed to match the input data distribution [Kindermans’17]
Tutorial on Interpretable Machine Learning

W. Samek & K.-R. Müller

Part 2b: Applications & Discussion
Recap: Layer-wise Relevance Propagation (LRP)

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

Theoretical interpretation

Deep Taylor Decomposition (Montavon et al., 2017)

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

The idea is to decompose the function as follows:

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

where \( R_i \) represents the relevance of the \( i \)-th neuron.
Explanations and now?

Data → Information → Interpretable for human

Rooster
Explanations and now?

Data → Information → Interpretable for human

How good is the explanation?
Explanations and now?

How good is the explanation?
What can we do with it?
Explanations and now?

How good is the explanation?
- Objective measure of quality
- Compare explanation methods

What can we do with it?
- Compare classifiers
- Detect biases and flaws
- Quantify use of context
- Novel representation
...
Measuring Quality of Explanations

Heatmap depends on
- classifier
- explanation method

If we want to compare classifiers or explanations methods, we need an *objective* measure of heatmap quality.
Measuring Quality of Explanations

Heatmap depends on
- classifier
- explanation method

If we want to compare classifiers or explanations methods, we need an *objective* measure of heatmap quality.

**Algorithm** (Pixel Flipping)

Sort pixel scores
Iterate
  - flip pixels
  - evaluate $f(x)$
Measure decrease of $f(x)$
Compare Explanation Methods

LRP

# 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

score for correct class

# 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

# pixel flips: 0
Compare Explanation Methods

Random

# pixel flips: 30
Compare Explanation Methods

Random

score for correct class

# pixel flips

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.
Compare Explanation Methods

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. 2016)
Compare Explanation Methods

**Red:** LRP method

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

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

(Samek et al. 2016)
Compare Explanation Methods

**Red:** LRP method

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

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

LRP produces quantitatively better heatmaps.

(Samek et al. 2016)
Compare Explanation Methods

Same idea can be applied for other domains (e.g. text document classification)

“Pixel flipping”
= “Word deleting”
Compare Explanation Methods

Same idea can be applied for other domains (e.g. text document classification)

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. 2016)
Deleting relevant words leads to a quick decrease of classifier accuracy.

The decrease is much steeper for LRP than for random word deletion and deletion according to sensitivity.

**LRP better identifies relevant words.**

*(Arras et al. 2016)*
Explanations and now?

- Objective measure of quality
- Compare explanation methods

What can we do with it?

- Compare classifiers
- Detect biases and flaws
- Quantify use of context
- Novel representation
- Application in the sciences
...
Opening the Black-Box

what speaks for / against classification as “3”

what speaks for / against classification as “9”

[number]: explanation target class
red color: evidence for prediction
blue color: evidence against prediction
Application: Compare Classifiers

### 20 Newsgroups data set

<table>
<thead>
<tr>
<th>comp.graphics</th>
<th>rec.autos</th>
<th>sci.crypt</th>
</tr>
</thead>
<tbody>
<tr>
<td>comp.os.ms-windows.misc</td>
<td>rec.motorcycles</td>
<td>sci.electronics</td>
</tr>
<tr>
<td>comp.sys.ibm.pc.hardware</td>
<td>rec.sport.baseball</td>
<td>sci.med</td>
</tr>
<tr>
<td>comp.sys.mac.hardware</td>
<td>rec.sport.hockey</td>
<td>sci.space</td>
</tr>
<tr>
<td>comp.windows.x</td>
<td>talk.politics.misc</td>
<td>talk.religion.misc</td>
</tr>
<tr>
<td>misc.forsale</td>
<td>talk.politics.guns</td>
<td>alt.atheism</td>
</tr>
<tr>
<td></td>
<td>talk.politics.mideast</td>
<td>soc.religion.christian</td>
</tr>
</tbody>
</table>

### Test set performance

word2vec / CNN model: 80.19%
Application: Compare Classifiers

20 Newsgroups data set

<table>
<thead>
<tr>
<th>comp.graphics</th>
<th>rec.autos</th>
<th>sci.crypt</th>
</tr>
</thead>
<tbody>
<tr>
<td>comp.os.ms-windows.misc</td>
<td>rec.motorcycles</td>
<td>sci.electronics</td>
</tr>
<tr>
<td>comp.sys.ibm.pc.hardware</td>
<td>rec.sport.baseball</td>
<td>sci.med</td>
</tr>
<tr>
<td>comp.sys.mac.hardware</td>
<td>rec.sport.hockey</td>
<td>sci.space</td>
</tr>
<tr>
<td>comp.windows.x</td>
<td>talk.politics.misc</td>
<td>alt.atheism</td>
</tr>
<tr>
<td>misc.forsale</td>
<td>talk.politics.guns</td>
<td>talk.religion.misc</td>
</tr>
<tr>
<td></td>
<td>talk.politics.mideast</td>
<td>soc.religion.christian</td>
</tr>
</tbody>
</table>

Test set performance
word2vec / CNN model: 80.19%
Application: Compare Classifiers

20 Newsgroups data set

<table>
<thead>
<tr>
<th>comp.graphics</th>
<th>comp.os.ms-windows.misc</th>
<th>comp.sys.ibm.pc.hardware</th>
<th>comp.sys.mac.hardware</th>
<th>comp.windows.x</th>
</tr>
</thead>
<tbody>
<tr>
<td>rec.autos</td>
<td>rec.motorcycles</td>
<td>rec.sport.baseball</td>
<td>rec.sport.hockey</td>
<td></td>
</tr>
<tr>
<td>sci.crypt</td>
<td>sci.electronics</td>
<td>sci.med</td>
<td>sci.space</td>
<td></td>
</tr>
<tr>
<td>talk.politics.misc</td>
<td>talk.politics.guns</td>
<td>talk.politics.mideast</td>
<td>talk.religion.misc</td>
<td>alt.atheism</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>soc.religion.christian</td>
</tr>
</tbody>
</table>

Test set performance
word2vec / CNN model: 80.19%

man  king  woman

\[
\begin{pmatrix}
a_1 \\ a_2 \\ \vdots \\ a_d \\
\end{pmatrix}
\begin{pmatrix}
b_1 \\ b_2 \\ \vdots \\ b_d \\
\end{pmatrix}
\begin{pmatrix}
c_1 \\ c_2 \\ \vdots \\ c_d \\
\end{pmatrix}
\]

man  woman  queen
Application: Compare Classifiers

20 Newsgroups data set

<table>
<thead>
<tr>
<th>comp.graphics</th>
<th>rec.autos</th>
<th>sci.crypt</th>
</tr>
</thead>
<tbody>
<tr>
<td>comp.os.ms-windows.misc</td>
<td>rec.motorcycles</td>
<td>sci.electronics</td>
</tr>
<tr>
<td>comp.sys.ibm.pc.hardware</td>
<td>rec.sport.baseball</td>
<td>sci.med</td>
</tr>
<tr>
<td>comp.sys.mac.hardware</td>
<td>rec.sport.hockey</td>
<td>sci.space</td>
</tr>
<tr>
<td>comp.windows.x</td>
<td>talk.politics.misc</td>
<td>talk.religion.misc</td>
</tr>
<tr>
<td>misc.forsale</td>
<td>talk.politics.guns</td>
<td>alt.atheism</td>
</tr>
<tr>
<td></td>
<td>talk.politics.mideast</td>
<td>soc.religion.christian</td>
</tr>
</tbody>
</table>

Test set performance

- word2vec / CNN model: 80.19%
- BoW/SVM model: 80.10%
Application: Compare Classifiers

20 Newsgroups data set

| comp.graphics | rec.autos | sci.crypt |
| comp.os.ms-windows.misc | rec.motorcycles | sci.electronics |
| comp.sys.ibm.pc.hardware | rec.sport.baseball | sci.med |
| comp.sys.mac.hardware | rec.sport.hockey | sci.space |
| misc.forsale | talk.politics.misc | talk.religion.misc |
| | talk.politics.guns | alt.atheism |
| | talk.politics.mideast | soc.religion.christian |

Test set performance

- word2vec / CNN model: 80.19%
- BoW/SVM model: 80.10%

same performance —> same strategy?
Application: Compare Classifiers

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. 2016)
Application: Compare Classifiers

word2vec / CNN model

<table>
<thead>
<tr>
<th>sci.med</th>
</tr>
</thead>
<tbody>
<tr>
<td>symptoms (7.3), treatments (6.6), medication (6.4), osteopathy (6.3), ulcers (6.2), sciatica (6.0), hypertension (6.0), herb (5.6), doctor (5.4), physician (5.1), Therapy (5.1), antibiotics (5.1), Asthma (5.0), renal (5.0), medicines (4.9), caffeine (4.9), infection (4.9), gastrointestinal (4.8), therapy (4.8), homeopathic (4.7), medicine (4.7), allergic (4.7), dosages (4.7), esophagitis (4.7), inflammation (4.6), arrhythmias (4.6), cancer (4.6), disease (4.6), migraine (4.6), patients (4.5).</td>
</tr>
</tbody>
</table>

BoW/SVM model

<table>
<thead>
<tr>
<th>sci.med</th>
</tr>
</thead>
<tbody>
<tr>
<td>cancer (1.4), photography (1.0), doctor (1.0), msg (0.9), disease (0.9), medical (0.8), sleep (0.8), radiologist (0.7), eye (0.7), treatment (0.7), prozac (0.7), vitamin (0.7), epilepsy (0.7), health (0.6), yeast (0.6), skin (0.6), pain (0.5), liver (0.5), physician (0.5), she (0.5), needles (0.5), dn (0.5), circumcision (0.5), syndrome (0.5), migraine (0.5), antibiotic (0.5), water (0.5), blood (0.5), fat (0.4), weight (0.4).</td>
</tr>
</tbody>
</table>

Words with maximum relevance

(Arras et al. 2016)
### Application: Compare Classifiers

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

**Words with maximum relevance**

*(Arras et al. 2016)*
Application: Compare Classifiers

Fisher Vector / SVM Classifier

Layer 1
Image of 'bicycle'

Layer 2
Local Features

Layer 3
GMM fitting

\[ \Psi'(l) = \begin{bmatrix} \Psi_{\pi_k}(l) & \Psi_{\mu_k}(l) & \Psi_{\sigma_k}(l) \end{bmatrix} \]

Layer 4
Normalization + Linear SVM
(Hellinger's kernel SVM)

\[ \mathbf{x} = \frac{1}{|L|} \sum_{l \in L} \Psi'(l) \]

\[ \mathbf{x} \leftarrow \text{sign}(\mathbf{x}) ||\mathbf{x}||^{\frac{1}{2}} \]

\[ \mathbf{x} \leftarrow \frac{\mathbf{x}}{||\mathbf{x}||^2} \]

\[ f(\mathbf{x}) = b + \sum_i \alpha_i k(\mathbf{x}_i, \mathbf{x}) \]

(Lapuschkin et al. 2016)
Application: Compare Classifiers

Fisher Vector / SVM Classifier

Layer 1
Image of 'bicycle'

Layer 2
Local Features

Layer 3
GMM fitting
\[ \Psi(1) = \begin{bmatrix} \Psi_{\pi_1}(1) & \Psi_{\mu_1}(1) & \Psi_{\sigma_1}(1) \\ \Psi_{\pi_2}(1) & \Psi_{\mu_2}(1) & \Psi_{\sigma_2}(1) \end{bmatrix} \]

Layer 4
Normalization + Linear SVM
(Hellinger’s kernel SVM)
\[ \mathbf{x} = \frac{1}{|L|} \sum_{l \in L} \Psi_l(1) \]
\[ \mathbf{x} \leftarrow \text{sign}(\mathbf{x})|\mathbf{x}|^{\frac{1}{2}} \]
\[ \mathbf{x} \leftarrow \frac{\mathbf{x}}{||\mathbf{x}||_2} \quad f(\mathbf{x}) = b + \sum_i \alpha_i k(\mathbf{x}_i, \mathbf{x}) \]

LRP general method for non-linear classifiers

(Lapuschkin et al. 2016)
Application: Compare Classifiers

Fisher Vector / SVM Classifier

Layer 1: Image of 'bicycle'
Layer 2: Local Features
Layer 3: GMM fitting
Layer 4: Normalization + Linear SVM (Hellinger's kernel SVM)

Heatmap

Relevance Conservation
\[ \sum_i R_i^{(1)} = \sum_j R_j^{(2)} \]

Redistribution Formula
\[ R_p^{(1)} = \sum_{l \in L(p)} \frac{R_l^{(2)}}{|\text{area}(l)|} \]

\[ \sum_i R_i^{(2)} = \sum_j R_j^{(3)} \]

\[ \sum_i R_i^{(3)} = f(x) \]

\[ R_d^{(3)} = \sum_i \alpha_i y_i \phi(x_i) d \phi(x) d + \frac{b}{D} \]

(Lapuschkin et al. 2016)
Application: Compare Classifiers

Fisher Vector / SVM Classifier

Layer 1
Image of 'bicycle'
Heatmap

Layer 2
Local Features

Layer 3
GMM fitting
Fisher Vector

Layer 4
Normalization + Linear SVM
(Hellinger's kernel SVM)

\[ x = \frac{1}{|L|} \sum_{l \in L} \Psi_{\lambda}(l) \]
\[ x \leftarrow \text{sign}(x)|x|^{\frac{1}{3}} \]
\[ x \leftarrow \frac{x}{\|x\|_2} f(x) = b + \sum_i \alpha_i k(x_i, x) \]

Deep Neural Network
- BVLC reference model + fine tuning.

Dataset
- PASCAL VOC 2007

(Lapuschkin et al. 2016)
Application: Compare Classifiers

Fisher Vector / SVM Classifier

- Aeroplanes
- Bicycles
- Birds
- Boats
- Bottles
- Buses
- Cars
- Cats
- Chairs
- Cows
- Dining tables
- Dogs
- Horses
- Motorbikes
- People
- Potted plants
- Sheep
- Sofas
- Trains
- TV/Monitors

Dataset
- PASCAL VOC 2007

Deep Neural Network
- BVLC reference model + fine tuning. (Lapuschkin et al. 2016)

\[ \sum_i \alpha_i k(x_i, x) \]

\[ (x)_d + \frac{b}{D} \]

2007

(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>mAP</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>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>Fisher DeepNet</td>
<td>59.92%</td>
<td>51.92%</td>
<td>47.60%</td>
<td>58.06%</td>
<td>42.28%</td>
<td><strong>80.45%</strong></td>
<td><strong>81.60%</strong></td>
</tr>
<tr>
<td>81.10%</td>
<td>51.04%</td>
<td>61.10%</td>
<td>64.62%</td>
<td>76.17%</td>
<td></td>
<td>69.34%</td>
<td>79.33%</td>
</tr>
<tr>
<td>Fisher DeepNet</td>
<td><strong>85.10%</strong></td>
<td><strong>28.62%</strong></td>
<td><strong>49.58%</strong></td>
<td>49.31%</td>
<td><strong>82.71%</strong></td>
<td><strong>54.33%</strong></td>
<td>mAP</td>
</tr>
<tr>
<td>92.43%</td>
<td><strong>49.99%</strong></td>
<td>74.04%</td>
<td>49.48%</td>
<td>87.07%</td>
<td>67.08%</td>
<td>72.12%</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 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></td>
<td>88.08%</td>
<td>79.69%</td>
<td>80.77%</td>
<td>77.20%</td>
<td>35.48%</td>
<td>72.71%</td>
<td>86.30%</td>
</tr>
<tr>
<td>Fisher DeepNet</td>
<td>59.92%</td>
<td>51.92%</td>
<td>47.60%</td>
<td>58.06%</td>
<td>42.28%</td>
<td>80.45%</td>
<td>69.34%</td>
</tr>
<tr>
<td></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 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></td>
<td>92.43%</td>
<td>49.99%</td>
<td>74.04%</td>
<td>49.48%</td>
<td>87.07%</td>
<td>67.08%</td>
<td>72.12%</td>
</tr>
</tbody>
</table>

Image

![Image](image_url)

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 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></td>
<td>88.08%</td>
<td>79.69%</td>
<td>80.77%</td>
<td>77.20%</td>
<td>35.48%</td>
<td>72.71%</td>
<td>86.30%</td>
</tr>
<tr>
<td>Fisher DeepNet</td>
<td>59.92%</td>
<td>51.92%</td>
<td>47.60%</td>
<td>58.06%</td>
<td>42.28%</td>
<td>80.45%</td>
<td>69.34%</td>
</tr>
<tr>
<td></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></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

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

how important is context?

classifier

how important is context?
Application: Measure Context Use

how important is context?

classifier

how important is context?

LRP decomposition allows meaningful pooling over bbox!

GCPR 2017 Tutorial — W. Samek & K.-R. Müller
Application: Measure Context Use

LRP decomposition allows meaningful pooling over bbox!

impotance of context = relevance outside bbox

relevance inside bbox
Application: Measure Context Use

Large values indicate importance of context

(Lapuschkin et al. 2016)
GoogleNet uses less context than BVLC model.

Context use anti-correlated with performance.

(Lapuschkin et al. 2016)
Application: Novel Representation

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

(Arras et al. 2016)
Application: Novel Representation

2D PCA projection of document vectors

(Arras et al. 2016)
Application: Novel Representation

2D PCA projection of document vectors

LRP

uniform

TFIDF

(Arras et al. 2016)
Application: Novel Representation

2D PCA projection of document vectors

Document vector computation is unsupervised.

(Arras et al. 2016)
Application: Novel Representation

2D PCA projection of document vectors

Document vector computation is unsupervised.

KNN-Performance on document vectors (Explanatory Power Index)

LRP: 0.8076
TFIDF: 0.6816
uniform: 0.6208

(Arras et al. 2016)
Application: Sentiment Analysis

How to handle multiplicative interactions?

\[ z_j = z_g \cdot z_s \]

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

gate neuron indirectly affect relevance distribution in forward pass

Negative sentiment

1. do n't waste your money .
2. neither funny nor suspenseful nor particularly well-drawn .
3. it 's not horrible , just horribly mediocre .
4. ... too slow , too boring , and occasionally annoying .
5. it 's neither as romantic nor as thrilling as it should be .

Positive sentiment

19. a worthy entry into a very difficult genre .
20. it 's a good film -- not a classic , but odd , entertaining and authentic .
21. it never fails to engage us .
Application: Sentiment Analysis

How to handle multiplicative interactions?

\[ z_j = z_g \cdot z_s \]

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

Gate neuron indirectly affect relevance distribution in forward pass

Negative sentiment

1. do n't waste your money .
2. neither funny nor suspenseful nor particularly well-drawn .
3. it 's not horrible , just horribly mediocre .
4. ... too slow , too boring , and occasionally annoying .
5. it 's neither as romantic nor as thrilling as it should be .

Positive sentiment

19. a worthy entry into a very difficult genre .
20. it 's a good film -- not a classic , but odd , entertaining and authentic .
21. it never fails to engage us .

(Arras et al., 2017)
Application: Face Analysis

Identifying age-related features

Attractivity

Emotions

(Arbabzadah et al. 2016)
Application: Face Analysis

Gender Classification

CaffeNet

GoogleNet

VGG16
Application: Video Analysis

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

(Srinivasan et al. 2017)
Application: Video Analysis

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

(Srinivasan et al. 2017)
Application: Video Analysis

Explaining prediction: "sit-up"

LRP relevances per frame

Video frame

Which features are most relevant?

Mbh

HOF

MBh

(Srinivasan et al. 2017)
Application: Semantic Boundary Detection

Pixel-wise labelling are very costly

Idea: Use image-label labels
      + apply explanation methods

(Koh et al. 2017)
Application: Semantic Boundary Detection

Pixel-wise labelling are very costly

**Idea:** Use image-label labels
+ apply explanation methods

(Koh et al. 2017)
Application: Semantic Boundary Detection

Pixel-wise labelling are very costly

**Idea**: Use image-label labels
+ apply explanation methods

![Diagram](image.png)

![Table](table.png)

(Koh et al. 2017)
Application: Interpretability in the Sciences

Brain-Computer Interfacing

How brain works subject-dependent —> individual explanations

(Sturm et al. 2016)
Application: Interpretability in the Sciences

Brain-Computer Interfacing

How brain works subject-dependent → individual explanations

Movement Imagination → Preprocessing → DNN → Movement Decoding

Feedback

LRP

(Sturm et al. 2016)
Application: Interpretability in the Sciences

With LRP we can analyze relevant activity at each time point and channel.

Allows to **spatially** & **temporally** identify important activity in EEG data.

(Sturm et al. 2016)
Application: Interpretability in the Sciences

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

(Sturm et al. 2016)
Application: Interpretability in the Sciences

Deep Tensor Network predicts molecular properties with state-of-the-art accuracy.

Effect of energy of test charge —> interpretable for human expert

(Schütt et al. 2017)
Summary

- In many problems interpretability as important as prediction.

- Explaining individual predictions is key.

- We have powerful, mathematically well-founded methods (LRP / deep Taylor) to explain individual predictions.

- How can we use interpretability for improving models (focus of NIPS’17 Workshop “Interpretability - Now what?”)

- Many interesting applications with interpretable deep nets —> more to come soon!
Thank you for your attention

Visit:

http://www.heatmaping.org

- Tutorials
- Software
- Online Demos

For more information, check out our tutorial paper:
Montavon et al. “Methods for Interpreting and Understanding Deep Neural Networks”
References


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)*, 1-10, 2017.


References


