Interpreting Deep Neural Networks and their Predictions

Wojciech Samek
ML Group, Fraunhofer HHI

(joint work with S. Lapuschkin, A. Binder, G. Montavon, K.-R. Müller)
“Superhuman” AI Systems

- Game GO
- Texas Hold’em Poker
- Image classification
- Traffic sign recognition
- IQ Test

Computer games
Jeopardy
Drone control
Lung cancer detection
Skin cancer detection

Fraunhofer Heinrich Hertz Institute

Wojciech Samek: Interpreting Deep Neural Networks and their Predictions
Can we trust these black boxes?

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

Is the way error is measured a satisfying specification of the problem? Are we measuring the error on the true data distribution?

$$\min_{f \in \mathcal{F}} \int_{x,y} \| f(x) - y \|^2 d\rho(x, y)$$
Can we trust these black boxes?

We need interpretability in order to:

- verify system
- understand weaknesses
- legal aspects
- learn new things from data
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

Finding a prototype:

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

Individual explanation:

Question: Why is this example classified as a motorbike?
Dimensions of Interpretability

Find the input pattern that maximizes class probability.

Find the most likely input pattern for a given class.

Explain individual prediction.
Dimensions of Interpretability

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

\[
\max_{x \in X} p_{\theta}(\omega_c | x) + \lambda \Omega(x)
\]
Dimensions of Interpretability

train interpretable model vs. train best model

suboptimal or biased due to assumptions (linearity, sparsity ...)

interpret it
Opening the Black Box with Layer-wise Relevance Propagation
We developed a *general* method to explain *individual* classification decisions.

Main idea: \[ \sum_p r_p = f(x) \]

“measure how much each pixel contributes to the overall prediction”

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

What makes this image a “rooster image”?

Idea: Redistribute the evidence for class rooster back to image space.
Opening the black box

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

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 \( \alpha + \beta = 1 \)
Opening the black box

Layer-wise relevance conservation

\[ \sum_i R_i = \ldots = \sum_i R_i^{(l)} = \sum_j R_j^{(l+1)} = \ldots = f(x) \]
Opening the black box

LRP Image

Class '3'

Class '9'
A more Principled Approach to Explanation by Decomposition
Explaination by Decomposition

\[ f(x_1, \ldots, x_d) \]

\[ R(x) \approx f(x') \]

\[ \sum_p R_p(x) + \varepsilon = f(x) \text{ with } \varepsilon \text{ small.} \]

\[ \forall_p: R_p(x) \geq 0 \]

(Montavon et al., 2017
Montavon et al. 2018)
Explanation by Decomposition

Taylor expansion:
\[ f(x) = f(\bar{x}) + \sum_{p=1}^{d} \frac{\partial f}{\partial x_p} \bigg|_{\bar{x}} \cdot (x_p - \bar{x}_p) + \varepsilon \]

\[ R_p(x) \]

→ Taylor decomposition:
\[ R_1 + R_2 + \ldots + R_d + \varepsilon \]
Explanation by Decomposition

\[ f(x) = f(\tilde{x}) + \sum_{p=1}^{d} \frac{\partial f}{\partial x_p} \bigg|_{\tilde{x}} \cdot (x_p - \tilde{x}_p) + \epsilon \]

What did we gain?

How the explanation should be built?  \[ \xrightarrow{\text{How to choose the root point?}} \]

\[ \tilde{x} \]

Wojciech Samek: Interpreting Deep Neural Networks and their Predictions
Explanation by Decomposition

For deep ReLU networks without biases we can find a root point at the origin.

\[ \tilde{x} = \lim_{\varepsilon \to 0} \varepsilon \cdot x \approx 0. \]

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

\[ f(x) = \sum_{i=1}^{d} R_i \]
Explanation by Decomposition

Let’s try Taylor Decomposition on a ConvNet.
Explanation by Decomposition

Taylor Expansion:

\[ f(x) = f(\tilde{x}) + \sum_{p=1}^{d} \frac{\partial f}{\partial x_p} \bigg|_{\tilde{x}} \cdot (x_p - \tilde{x}_p) + \varepsilon \]

Explanations are noisy and uninformative.

Question: what went wrong?
Explaination by Decomposition

“Naive” Taylor decomposition of neural network does not give satisfactory results.

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
Key Idea: If a decision is too complex to explain, break the decision function into sub-functions, and explain each sub-decision separately.
Explanation by 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
**Insight:** In order to explain, decompose the decision function and explain sub functions.
Comparison with Other Explanation Methods
Other Explanation Methods

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

Decomposition
- LRP (Bach et al., 2015)
- Excitation Backprop (Zhang et al., 2016)
- Deep Taylor Decomposition (Montavon et al., 2017 (arXiv 2015))

Optimization
- LIME (Ribeiro et al., 2016)
- Meaningful Perturbations (Fong & Vedaldi, 2017)
- PatternLRP (Kindermans et al., 2017)

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

DeepLIFT (Shrikumar et al., 2016)
Grad-CAM (Selvaraju et al., 2016)
Integrated Gradient (Sundararajan et al., 2017)
Other Explanation Methods

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

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

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Example 1</th>
<th>Example 2</th>
<th>Example 3</th>
<th>Example 4</th>
<th>Example 5</th>
<th>Example 6</th>
<th>Example 7</th>
<th>Example 8</th>
<th>Example 9</th>
<th>Example 10</th>
<th>Example 11</th>
<th>Example 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td><img src="image1" alt="Input" /></td>
<td><img src="image2" alt="Input" /></td>
<td><img src="image3" alt="Input" /></td>
<td><img src="image4" alt="Input" /></td>
<td><img src="image5" alt="Input" /></td>
<td><img src="image6" alt="Input" /></td>
<td><img src="image7" alt="Input" /></td>
<td><img src="image8" alt="Input" /></td>
<td><img src="image9" alt="Input" /></td>
<td><img src="image10" alt="Input" /></td>
<td><img src="image11" alt="Input" /></td>
<td><img src="image12" alt="Input" /></td>
</tr>
<tr>
<td>Gradient</td>
<td><img src="image13" alt="Gradient" /></td>
<td><img src="image14" alt="Gradient" /></td>
<td><img src="image15" alt="Gradient" /></td>
<td><img src="image16" alt="Gradient" /></td>
<td><img src="image17" alt="Gradient" /></td>
<td><img src="image18" alt="Gradient" /></td>
<td><img src="image19" alt="Gradient" /></td>
<td><img src="image20" alt="Gradient" /></td>
<td><img src="image21" alt="Gradient" /></td>
<td><img src="image22" alt="Gradient" /></td>
<td><img src="image23" alt="Gradient" /></td>
<td><img src="image24" alt="Gradient" /></td>
</tr>
<tr>
<td>SmoothGrad</td>
<td><img src="image25" alt="SmoothGrad" /></td>
<td><img src="image26" alt="SmoothGrad" /></td>
<td><img src="image27" alt="SmoothGrad" /></td>
<td><img src="image28" alt="SmoothGrad" /></td>
<td><img src="image29" alt="SmoothGrad" /></td>
<td><img src="image30" alt="SmoothGrad" /></td>
<td><img src="image31" alt="SmoothGrad" /></td>
<td><img src="image32" alt="SmoothGrad" /></td>
<td><img src="image33" alt="SmoothGrad" /></td>
<td><img src="image34" alt="SmoothGrad" /></td>
<td><img src="image35" alt="SmoothGrad" /></td>
<td><img src="image36" alt="SmoothGrad" /></td>
</tr>
<tr>
<td>Integrated Gradients</td>
<td><img src="image37" alt="Integrated Gradients" /></td>
<td><img src="image38" alt="Integrated Gradients" /></td>
<td><img src="image39" alt="Integrated Gradients" /></td>
<td><img src="image40" alt="Integrated Gradients" /></td>
<td><img src="image41" alt="Integrated Gradients" /></td>
<td><img src="image42" alt="Integrated Gradients" /></td>
<td><img src="image43" alt="Integrated Gradients" /></td>
<td><img src="image44" alt="Integrated Gradients" /></td>
<td><img src="image45" alt="Integrated Gradients" /></td>
<td><img src="image46" alt="Integrated Gradients" /></td>
<td><img src="image47" alt="Integrated Gradients" /></td>
<td><img src="image48" alt="Integrated Gradients" /></td>
</tr>
<tr>
<td>Deconvnet</td>
<td><img src="image49" alt="Deconvnet" /></td>
<td><img src="image50" alt="Deconvnet" /></td>
<td><img src="image51" alt="Deconvnet" /></td>
<td><img src="image52" alt="Deconvnet" /></td>
<td><img src="image53" alt="Deconvnet" /></td>
<td><img src="image54" alt="Deconvnet" /></td>
<td><img src="image55" alt="Deconvnet" /></td>
<td><img src="image56" alt="Deconvnet" /></td>
<td><img src="image57" alt="Deconvnet" /></td>
<td><img src="image58" alt="Deconvnet" /></td>
<td><img src="image59" alt="Deconvnet" /></td>
<td><img src="image60" alt="Deconvnet" /></td>
</tr>
<tr>
<td>Guided Backprop</td>
<td><img src="image61" alt="Guided Backprop" /></td>
<td><img src="image62" alt="Guided Backprop" /></td>
<td><img src="image63" alt="Guided Backprop" /></td>
<td><img src="image64" alt="Guided Backprop" /></td>
<td><img src="image65" alt="Guided Backprop" /></td>
<td><img src="image66" alt="Guided Backprop" /></td>
<td><img src="image67" alt="Guided Backprop" /></td>
<td><img src="image68" alt="Guided Backprop" /></td>
<td><img src="image69" alt="Guided Backprop" /></td>
<td><img src="image70" alt="Guided Backprop" /></td>
<td><img src="image71" alt="Guided Backprop" /></td>
<td><img src="image72" alt="Guided Backprop" /></td>
</tr>
<tr>
<td>PatterNet</td>
<td><img src="image73" alt="PatterNet" /></td>
<td><img src="image74" alt="PatterNet" /></td>
<td><img src="image75" alt="PatterNet" /></td>
<td><img src="image76" alt="PatterNet" /></td>
<td><img src="image77" alt="PatterNet" /></td>
<td><img src="image78" alt="PatterNet" /></td>
<td><img src="image79" alt="PatterNet" /></td>
<td><img src="image80" alt="PatterNet" /></td>
<td><img src="image81" alt="PatterNet" /></td>
<td><img src="image82" alt="PatterNet" /></td>
<td><img src="image83" alt="PatterNet" /></td>
<td><img src="image84" alt="PatterNet" /></td>
</tr>
<tr>
<td>PatternAttribution</td>
<td><img src="image85" alt="PatternAttribution" /></td>
<td><img src="image86" alt="PatternAttribution" /></td>
<td><img src="image87" alt="PatternAttribution" /></td>
<td><img src="image88" alt="PatternAttribution" /></td>
<td><img src="image89" alt="PatternAttribution" /></td>
<td><img src="image90" alt="PatternAttribution" /></td>
<td><img src="image91" alt="PatternAttribution" /></td>
<td><img src="image92" alt="PatternAttribution" /></td>
<td><img src="image93" alt="PatternAttribution" /></td>
<td><img src="image94" alt="PatternAttribution" /></td>
<td><img src="image95" alt="PatternAttribution" /></td>
<td><img src="image96" alt="PatternAttribution" /></td>
</tr>
<tr>
<td>LRP-Z</td>
<td><img src="image97" alt="LRP-Z" /></td>
<td><img src="image98" alt="LRP-Z" /></td>
<td><img src="image99" alt="LRP-Z" /></td>
<td><img src="image100" alt="LRP-Z" /></td>
<td><img src="image101" alt="LRP-Z" /></td>
<td><img src="image102" alt="LRP-Z" /></td>
<td><img src="image103" alt="LRP-Z" /></td>
<td><img src="image104" alt="LRP-Z" /></td>
<td><img src="image105" alt="LRP-Z" /></td>
<td><img src="image106" alt="LRP-Z" /></td>
<td><img src="image107" alt="LRP-Z" /></td>
<td><img src="image108" alt="LRP-Z" /></td>
</tr>
<tr>
<td>LRP-CompositeAFlat</td>
<td><img src="image109" alt="LRP-CompositeAFlat" /></td>
<td><img src="image110" alt="LRP-CompositeAFlat" /></td>
<td><img src="image111" alt="LRP-CompositeAFlat" /></td>
<td><img src="image112" alt="LRP-CompositeAFlat" /></td>
<td><img src="image113" alt="LRP-CompositeAFlat" /></td>
<td><img src="image114" alt="LRP-CompositeAFlat" /></td>
<td><img src="image115" alt="LRP-CompositeAFlat" /></td>
<td><img src="image116" alt="LRP-CompositeAFlat" /></td>
<td><img src="image117" alt="LRP-CompositeAFlat" /></td>
<td><img src="image118" alt="LRP-CompositeAFlat" /></td>
<td><img src="image119" alt="LRP-CompositeAFlat" /></td>
<td><img src="image120" alt="LRP-CompositeAFlat" /></td>
</tr>
<tr>
<td>LRP-CompositeBFI</td>
<td><img src="image121" alt="LRP-CompositeBFI" /></td>
<td><img src="image122" alt="LRP-CompositeBFI" /></td>
<td><img src="image123" alt="LRP-CompositeBFI" /></td>
<td><img src="image124" alt="LRP-CompositeBFI" /></td>
<td><img src="image125" alt="LRP-CompositeBFI" /></td>
<td><img src="image126" alt="LRP-CompositeBFI" /></td>
<td><img src="image127" alt="LRP-CompositeBFI" /></td>
<td><img src="image128" alt="LRP-CompositeBFI" /></td>
<td><img src="image129" alt="LRP-CompositeBFI" /></td>
<td><img src="image130" alt="LRP-CompositeBFI" /></td>
<td><img src="image131" alt="LRP-CompositeBFI" /></td>
<td><img src="image132" alt="LRP-CompositeBFI" /></td>
</tr>
</tbody>
</table>
Measuring Quality of Explanations

Can we objectively measure which heatmap is best?

**Algorithm** (Pixel Flipping)

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

(Samek et al., 2017)
Measuring Quality of Explanations

LRP outperforms other methods on MNIST.
Measuring Quality of Explanations

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)
Interpretability in Practice
Application: Compare Classifiers

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>

Two classifiers
- similar classification accuracy on horse class
- but do they solve the problem similarly?
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><strong>Fisher DeepNet</strong></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><strong>Fisher DeepNet</strong></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><strong>Fisher DeepNet</strong></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

(Lapuschkin et al., 2016)
Application: Compare Classifiers

‘horse’ images in PASCAL VOC 2007
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></td>
<td></td>
</tr>
<tr>
<td>misc.forsale</td>
<td>talk.politics.misc</td>
<td>talk.religion.misc</td>
</tr>
<tr>
<td></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%

same performance —> same strategy?
Application: Compare Classifiers

**word2vec/CNN:** identifies semantically meaningful words

**BoW/SVM:** identifies statistical patterns (word statistics)

---

Yes, weightlessness does feel like falling. It may feel strange at first, but the body does adjust. The feeling is not too different from that of sky diving.

> And what is the motion **sickness**
> that some astronauts occasionally experience?

It is the body's reaction to a strange environment. It appears to be induced partly to physical **discomfort** and part to mental distress. Some people are more prone to it than others, like some people are more prone to get sick on a roller coaster ride than others. The mental part is usually induced by a lack of clear indication of which way is up or down, i.e. the Shuttle is normally oriented with its cargo bay pointed towards Earth, so the Earth (or ground) is "above" the head of the astronauts. About 50% of the astronauts experience some form of motion **sickness**, and NASA has done numerous tests in space to try to see how to keep the number of occurrences down.

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: Context Use

How important is context?

\[
\text{importance of context} = \frac{\text{relevance outside bbox}}{\text{relevance inside bbox}}
\]
Application: Context Use

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

Context use anti-correlated with performance.

(Lapuschnik et al., 2016)
Application: Recurrent Networks

**Model:** word-based bidirectional LSTM
word embeddings of dimension 60, one hidden layer of size 60
takes as input a sequence of word embeddings \((x_1, x_2, \ldots, x_T)\)
[model released by Li et al. 2016]

**Task:** five-class sentiment prediction
training on phrases and sentences of the Stanford Sentiment Treebank
[dataset released by Socher et al. 2013]

How to handle multiplicative interactions?

\[
\begin{align*}
\bar{z}_j &= z_g \cdot z_s \\
R_g &= 0 \\
R_s &= R_j
\end{align*}
\]

(Arras et al., 2017)
Application: Recurrent Networks

movie review: ++, —

Negative sentiment

1. do not 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.

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

Faces in the wild (from Flickr)
#images: 26,580

Task: Predict gender & age (range)
Application: Face Analysis

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

Predictions

25-32 years old

60+ years old

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

- Real person
- Fake person
- Real person

Fake persons have different eyes

(Seibold et al., 2017)
Application: Video Analysis
Application: Machines Playing Games
Application: Biomedical Engineering

Brain-Computer Interfacing

Movement Imagination → Preprocessing → Movement Decoding

Feedback

explain

LRP

(Sturm et al., 2016)
Application: Biomedical Engineering
Summary

In many problems interpretability as important as prediction (trusting a black-box system may not be an option).

Use in practice
- verify predictions, detect biases and flaws, debug models
- compare and select architectures, understand and improve models
- extract additional information, perform further tasks

We have a powerful, mathematically well-founded method to explain individual predictions of complex machine learning models.
Upcoming Tutorials

CVPR 2018

MICCAI 2018

ICIP 2018

Wojciech Samek: Interpreting Deep Neural Networks and their Predictions
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)*, 159-168, 2017.


References


References


Thank you for your attention

Visit:

http://www.heatmapping.org

Tutorial Papers:
Montavon et al. “Methods for Interpreting and Understanding Deep Neural Networks”, 2018

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