Understanding ML Models

Klaus-Robert Müller et al.!!
Outline

• understanding single decisions of nonlinear learners
• Layer-wise Relevance Propagation (LRP)
• Applications in Physics, Chemistry and Medicine: towards insights
Towards Explaining: Machine Learning = black box?
Explaining single Predictions Pixel-wise

input image

Forward Propagation

$\quad x_j = \text{sign}(\sum_i x_i w_{ij})$

Relevance Propagation (Bach et al. 2015)

$\quad R_i = \sum_j R_j \sum_i x_i w_{ij}$

heatmap

Explaining single decisions is difficult!
Explaining nonlinear decisions is difficult.

- **Linear classification**
  - $S_1$: sepal width
  - $S_2$: sepal width
  - $V_1$: sepal width

- **Non-linear classification**
  - $S_1$: sepal width & length
  - $S_2$: sepal width
  - $V_1$: sepal length

---

Iris setosa (red)

Iris virginica (green)

Iris versicolor (blue)
Explaining single decisions is difficult
Explaining single Predictions Pixel-wise

input image

Forward Propagation

$\tilde{x}_j = \text{sign}(\sum_i x_i w_{ij})$

Relevance Propagation
(Bach et al. 2015)

$R_i = \sum_j R_j \frac{x_i w_{ij}}{\sum_k x_k w_{kj}}$

heat map

Goodbye Blackbox ML!
Historical remarks on Explaining Predictors

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

**Decomposition**
- LRP (Bach et al., 2015)
- Deep Taylor Decomposition (Montavon et al., 2017 (arXiv 2015))
- Gradient times input (Shrikumar et al., 2016)
- DeepLIFT (Shrikumar et al., 2016) (Selvaraju et al., 2016)
- Grad-CAM (Shrikumar et al., 2016)
- Integrated Gradient (Sundararajan et al., 2017)
- LRP for LSTM (Arras et al., 2017)
- Probabilistic Diff (Zintgraf et al., 2016)
- Excitation Backprop (Zhang et al., 2016)

**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)

**Understanding the Model**
- Deep Visualization (Yosinski et al., 2015)
- Inverting CNNs (Dosovitskiy & Brox, 2015)
- Synthesis of preferred inputs (Nguyen et al. 2016)
- Network Dissection (Zhou et al. 2017)
- TCAV (Kim et al. 2018)
Explaining Neural Network Predictions

Layer-wise relevance Propagation (LRP, Bach et al 15) first method to explain nonlinear classifiers - based on generic theory (related to Taylor decomposition – deep taylor decomposition M et al 17) - applicable to any NN with monotonous activation, BoW models, Fisher Vectors, SVMs etc.

**Explanation:** “Which pixels contribute how much to the classification” *(Bach et al 2015)*

*(what makes this image to be classified as a car)*

\[
f(x) = \sum_p h_p
\]

**Sensitivity / Saliency:** “Which pixels lead to increase/decrease of prediction score when changed” *(what makes this image to be classified more/less as a car)* *(Baehrens et al 10, Simonyan et al 14)*

\[
h_p = \left\| \frac{\partial}{\partial x_p} f(x) \right\|_\infty
\]

**Deconvolution:** “Matching input pattern for the classified object in the image” *(Zeiler & Fergus 2014)* *(relation to f(x) not specified)*

Each method solves a different problem!!!
Explaining Neural Network Predictions

Classification

\[ x_j = \sigma \left( \sum_i x_i w_{ij} + b_j \right) \]
Explaining Neural Network Predictions

Initialization

Explanation

$r_j$ = $f(x)$
Explaining Neural Network Predictions

Theoretical interpretation
Deep Taylor Decomposition

\[ r_i = x_i \sum_j \frac{w_{ij} r_j}{\sum_i x_i w_{ij}} = x_i C_i \]

\( r_i \) depends on the activations and the weights
Explaining Neural Network Predictions

Relevance Conservation Property

\[ \sum_p r_p = \ldots = \sum_i r_i = \sum_j r_j = \ldots = f(x) \]
Explaining Predictions Pixel-wise

Neural networks

Kernel methods
Some Digestion on Explaining
Sensitivity analysis is often not the question that you would like to ask!
LRP can ‘say’ positive and negative things

Positive and Negative Evidence: LRP distinguishes between positive evidence, supporting the classification decision, and negative evidence, speaking against the prediction.

LRP indicates what speaks for class ‘3’ and speaks against class ‘9’.
Measuring the Quality of Explanation (Samek et al 2017)

Is this a good explanation?

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

[Samek et al IEEE TNNLS 2017]
Measuring the Quality of Explanation

LRP outperforms Sensitivity and Deconvolution on all three datasets.
Application: Comparing Classifiers

Large values indicate importance of context

[Lapuschkin et al CVPR 2016]
Applying Explanation in Vision and Text
Application: Faces

What makes you look old?

What makes you look sad?

What makes you look attractive?
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.
LRP for LSTMs

- 2 types of operations:
  - Dot Product
  - Multiplicative

- Dot Product:
  - \( R_l = \frac{z_{ij}}{z_j + \epsilon \cdot \text{sign}(z_j)} \times R_{l+1} \)
  - \( z_{ij} = x_i w_{ij} \)
  - \( z_j = \sum_i x_i w_{ij} \)

- Multiplicative:
  - Occurrence:
    - output of gate \times source (e.g., cell state)
    - \( R_{\text{gate}} = 0 \)
    - \( R_{\text{source}} = R_{l+1} \)
  - Gate already accounted for in forward-pass (decides which information to keep)

(Arras et al., 2017)
Explaining LSTMs

**Example:** Visual question answering on the CLEVR dataset.

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

→ model understands the question and correctly identifies the object of interest

(Arras et al., in Press)
Is the Generalization Error all we need?
Application: Comparing Classifiers (Lapuschkin et al CVPR 2016)

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></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></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></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></td>
</tr>
</tbody>
</table>

**Image**  **FV**  **DNN**

![Image with horse]
Explaining problem solving strategies in scale
Spectral Relevance Analysis (SpRAy)

Lapuschkin et al. Nat Comms, March 11th 2019
**Figure 28:** Cluster label assignments for class “aeroplane” via SC for input images, FV model relevance maps and DNN relevance maps. Embedding coordinates in $\mathbb{R}^2$ for visualization have been computed on pair-wise distances derived from the weighted affinity matrix $W$ used for SC. The samples at the bottom right (square images) show DNN relevance maps and images with strong reaction of the DNN models to the border padding. FV relevance maps for the same images are shown to the left. Enlarged relevance maps and images are shown without preprocessing.
ML4 Quantum Chemistry
Machine Learning in Chemistry, Physics and Materials

Matthias Rupp, Anatole von Lilienfeld, Alexandre Tkatchenko, Klaus-Robert Müller

Ansatz:

\[ \{Z_I, R_I\} \xrightarrow{\text{ML}} E \]

instead of

\[ \hat{H}(\{Z_I, R_I\}) \xrightarrow{\Psi} E \]

\[ \hat{H}\Psi = E\Psi \]
Predicting Energy of small molecules: Results

March 2012
Rupp et al., PRL
**9.99 kcal/mol**
(kernels + eigenspectrum)

December 2012
Montavon et al., NIPS
**3.51 kcal/mol**
(Neural nets + Coulomb sets)

2015 Hansen et al 1.3 kcal/mol at 10 million times faster than the state of the art
Prediction considered chemically accurate when MAE is below **1 kcal/mol**

Now: 0.3 kcal/mol with DTNN and SchNet

Dataset available at [http://quantum-machine.org](http://quantum-machine.org)
Gaining insights for Physics
Toward Quantum Chemical Insight

Energy prediction:
\[ E = \sum_{i=1}^{n} E_i \]

Learned potential:
\[ \Omega^M_A(r) = E_{probe} \]

XAI for unsupervised learning
Support Vector Data Description (SVDD)

- Compute minimal enclosing sphere with center $c$ and radius $R$
- Anomaly score as the distance to center $c$, that is $f(x) = ||\phi(x) - c||$
- Accept data point $x$ if $f(x) \leq R$ and ...
  ...
  ... reject $x$ if $f(x) > R$
Explaining one-class

Figure 1: Illustration of the outlier detection and explanation setting. Left: Data is generated from an unknown distribution, we are for examples interested in potential outliers. Middle: Transformed machine learning techniques estimate the data generating distribution and assign an outlier score $o(x)$ to ordinary data points. Right: Our explanation method assigns a relevance score to every input variable that reflects the contribution of input variable $x_i$ to the model decision. We apply dithering to all heatmaps for printing reliability.

Figure 5: A One-Class SVM is trained on small $7 \times 7$ patches of the very image itself. Parameter $\nu = 0.1$ is set to allow at most 10% outliers. Images from a texture data set $[14]$ (row one, two and four) and PatternNet $[41]$; top image is altered by us. For every image, we show Left: input image; Middle: decomposition of one-class SVM; Right: Sobel filter for reference. All images were resized to 256 pixels width.

[Kaufmann, Müller, Montavon 2018, 2019]
Interpretable Clustering
NEON (Neuralization-Propagation)

**NEON’s idea:** When the ML model is not a neural network (e.g. a kernel machine), convert it into a neural network first (‘neuralize’ it).
Neuralizing K-means

Fig. 7. NEON analysis of images represented at different layers of a deep neural network (pretrained VGG16). K-means clustering with $K = 8$ is performed at these two layers. Each column shows the pixel-contributions for one of these clusters.
Semi-final Conclusion

• explaining & interpreting nonlinear models is essential
• orthogonal to improving DNNs and other models
• need for opening the blackbox …
• understanding nonlinear models is essential for Sciences & AI
• new theory: LRP is based on deep taylor expansion
• tool for gaining insight

www.heatmapping.org
Thank you for your attention

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


Further Reading II


Further Reading III


Further Reading IV


