Machine Learning and AI for the sciences – Towards Understanding

Klaus-Robert Müller et al.!!
Outline

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

Explaining single decisions is difficult!
Explaining single Predictions Pixel-wise

Goodbye Blackbox ML!
Historical remarks on Explaining Predictors

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

**Deep Taylor Decomposition** (Montavon et al., 2017 (arXiv 2015))
- Probabilistic Diff (Zintgraf et al., 2016)
- Meaningsful Perturbations (Fong & Vedaldi 2017)
- PatternLRP (Kindermans et al., 2017)

**Optimization**
- LIME (Ribeiro et al., 2016)
- Gradient vs. Decomposition (Montavon et al., 2018)
- Gradient times input (Shrikumar et al., 2016)
- DeepLIFT (Shrikumar et al., 2016)
- Grad-CAM (Selvaraju et al., 2016)
- Integrated Gradient (Sundararajan et al., 2017)

**Deconvolution**
- Deconvolution (Zeiler & Fergus 2014)
- Guided Backprop (Springenberg et al. 2015)
- Deep Visualization (Yosinski et al., 2015)
- Inverting CNNs (Dosovitskiy & Brox, 2015)
- Synthesis of preferred inputs (Nguyen et al. 2016)
- TCAV (Kim et al. 2018)
- Inverting CNNs (Mahendran & Vedaldi, 2015)
- RNN cell state analysis (Karpathy et al., 2015)
- Network Dissection (Zhou et al. 2017)
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

large activation

$x_j = \sigma(\sum_i x_i w_{ij} + b_j)$
Explaining Neural Network Predictions

Explanation

 Initialization

\[ r_j = f(x) \]
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!
Advantages of LRP over both Sensitivity and Deconvolution

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’

The sign of Sensitivity and Deconvolution does not have this interpretation.

-> taking norm gives unsigned visualizations
Applying Explanation in Vision and Text
Application: Faces

What makes you look old?

What makes you look attractive?

What makes you look sad?
Male or Female?

[Image showing relevance maps labeled 'male' and 'female']

http://interpretable-ml.org
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.
Understanding learning models for complex gaming scenarios
Analysing Breakout: LRP vs. Sensitivity
Machine Learning in the Sciences
Machine Learning in Neuroscience
Brain Computer Interfacing: ‘Brain Pong‘

Berlin Brain Computer Interface

• ML reduces patient training from 300h -> 5min

Applications

• help/hope for patients (ALS, stroke…)

• neuroscience

• neurotechnology (video coding, gaming, monitoring driving)

Leitmotiv: ›let the machines learn‹
ML4 Quantum Chemistry
Machine Learning in Chemistry, Physics and Materials

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

Machine Learning for chemical compound space

Ansatz:

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

instead of

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

\[ \hat{H}\Psi = E\Psi \]

[from von Lilienfeld]
Coulomb representation of molecules

\[ M_{ii} = Z_i^{2.4} \]
\[ M_{ij} = \frac{Z_i Z_j}{\| R_i - R_j \|} \]

+ phantom atoms

\[ \{0, R_{21}\} \quad \{0, R_{22}\} \quad \{0, R_{23}\} \]

Coulomb Matrix (Rupp, Müller et al 2012, PRL)

\[ d(M, M') = \sqrt{\sum_{IJ} |M_{IJ} - M'_{IJ}|^2} \]
Kernel ridge regression

Distances between $\mathbf{M}$ define Gaussian kernel matrix $\mathbf{K}$

$$k(\mathbf{M}, \mathbf{M}') = \exp\left(-\frac{d(\mathbf{M}, \mathbf{M}')^2}{2\sigma^2}\right)$$

Predict energy as sum over weighted Gaussians

$$E^{est}(\mathbf{M}) = \sum_i \alpha_i k(\mathbf{M}, \mathbf{M}_i) + b$$

using weights that minimize error in training set

$$\min_{\alpha} \sum_i \left(E^{est}(\mathbf{M}_i) - E_i^{ref}\right)^2 + \lambda \sum_i \alpha_i^2$$

Exact solution

$$\alpha = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{E}^{ref}$$

As many parameters as molecules + 2 global parameters, characteristic length-scale or $kT$ of system ($\sigma$), and noise-level ($\lambda$)

[from von Lilienfeld]
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 1.3kcal/mol at **10 million** times faster than the state of the art

Prediction considered chemically accurate when MAE is below **1 kcal/mol**

Dataset available at [http://quantum-machine.org](http://quantum-machine.org)
Is the Generalization Error all we need?
Application: Comparing Classifiers (Lapuschkin et al CVPR 2016)

Test error for various classes:

<table>
<thead>
<tr>
<th>Class</th>
<th>Fisher 79.08%</th>
<th>DeepNet 88.08%</th>
<th>Fisher 59.92%</th>
<th>DeepNet 81.10%</th>
<th>Fisher 85.10%</th>
<th>DeepNet 92.43%</th>
</tr>
</thead>
<tbody>
<tr>
<td>aeroplane</td>
<td>79.08%</td>
<td>66.44%</td>
<td>51.92%</td>
<td>51.04%</td>
<td>28.62%</td>
<td>49.99%</td>
</tr>
<tr>
<td>bicycle</td>
<td>45.90%</td>
<td>80.77%</td>
<td>47.60%</td>
<td>61.10%</td>
<td>49.58%</td>
<td>74.04%</td>
</tr>
<tr>
<td>bird</td>
<td>70.88%</td>
<td>77.20%</td>
<td>58.06%</td>
<td>64.62%</td>
<td>49.31%</td>
<td>49.48%</td>
</tr>
<tr>
<td>boat</td>
<td>27.64%</td>
<td>35.48%</td>
<td>42.28%</td>
<td>76.17%</td>
<td>82.71%</td>
<td>87.07%</td>
</tr>
<tr>
<td>bottle</td>
<td>69.67%</td>
<td>72.71%</td>
<td>80.45%</td>
<td>81.60%</td>
<td>54.33%</td>
<td>67.08%</td>
</tr>
<tr>
<td>bus</td>
<td>80.96%</td>
<td>86.30%</td>
<td>69.34%</td>
<td>79.33%</td>
<td>59.99%</td>
<td>72.12%</td>
</tr>
<tr>
<td>car</td>
<td>80.96%</td>
<td>86.30%</td>
<td>69.34%</td>
<td>79.33%</td>
<td>59.99%</td>
<td>72.12%</td>
</tr>
</tbody>
</table>

Image

FV

DNN
Learning Atomistic Representations with Deep Tensor Neural Networks

Kristof Schütt, Farhad Arbabzadah, Stefan Chmiela, Alexandre Tkatchenko, Klaus-Robert Müller

Deep Tensor Neural Network (DTNN) for representing molecules

Input: Atomic numbers and interatomic distances

\[ Z = \begin{bmatrix} Z_1 & Z_2 & \cdots & Z_n \end{bmatrix} \]
\[ D = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\
    d_{21} & d_{22} & \cdots & d_{2n} \\
    \vdots & \vdots & \ddots & \vdots \\
    d_{n1} & d_{n2} & \cdots & d_{nn} \end{bmatrix} \]

Embedding of based on atom types

\[ x_i^{(0)} = x_{Z_i} \in \mathbb{R}^d \]

Add interaction with environment using \( t = 1 \ldots T \) sequential refinements \( v_i^{(t)} \)

\[ x_i^{(t+1)} = x_i^{(t)} + v_i^{(t)} \left( x_1^{(t)}, \ldots, x_n^{(t)} \right) \]

Prediction via atom-wise contributions:

\[ \hat{E} = \sum_{i=1}^{n_{\text{atoms}}} f_{\text{out}}(x_i^{(T)}) \]

Schütt, Arbabsazadah, Chmiela, Müller, Tkatchenko, Nature Communications 8, 13890 (2017)
Gaining insights for Physics
Toward Quantum Chemical Insights: supervised

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

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

Machine Learning for morpho-molecular Integration

Alexander Binder¹,⁶, Michael Bockmayr²,¹⁰, Miriam Hägele¹, Stephan Wienert², Daniel Heim², Katharina Hellweg³, Albrecht Stenzinger⁴, Laura Parlow², Jan Budczies², Benjamin Goeppert⁴, Denise Treue², Manato Kotani⁵, Masaru Ishii⁵, Manfred Dietel², Andreas Hocke³, Carsten Denkert²,⁷, Klaus-Robert Müller¹,⁸,⁹,*, and Frederick Klauschen²,⁷,*
Interpretable ML

Bach et al., PLoS1 2015
Klauschen et al., US Patent #9558550
Binder et al., in revision
Machine learning based integration of morphological and molecular tumor profiles

- **MICROSCOPIC AND MOLECULAR DATA**
- **TRAINING**
- **PREDICTION**
- **INTEGRATION/INTERPRETATION**

**Histo-morphological features**
- Div. image modalities: brightfield, confocal
- In-house database
- Cell types: cancer, lymphocytes, stroma

**TCGA database**

**Molecular profiles**

**Computational fluorescence microscopy**
- Red: carcinoma
- Green: TILs
- Blue: molecular property

"computational fluorescence microscopy"
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
- compare the right thing

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 I


Brockherde, F., Vogt, L., Li, L., Tuckerman, M., Burke, K., Müller, K. R., By-passing the Kohn-Sham Equations with machine learning, Nature Communications, 8:872 (2017)


Further Reading II


Further Reading III


