EMBC Tutorial on Interpretable and Transparent Deep Learning

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

13:30 - 14:00  Introduction  KRM
14:00 - 15:00  Techniques for Interpretability  GM
15:00 - 15:30  Coffee Break  ALL
15:30 - 16:15  Evaluating Interpretability & Applications  WS
16:15 - 17:15  Applications in BME & the Sciences and Wrap-Up  KRM
Perspectives
Is the Generalization Error all we need?
Application: Comparing Classifiers (Lapuschkin CVPR’16)

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></td>
<td>cat</td>
<td>chair</td>
<td>cow</td>
<td>diningtable</td>
<td>dog</td>
<td>horse</td>
<td>motorbike</td>
</tr>
<tr>
<td>Fisher</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></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</td>
<td>85.10%</td>
<td>28.62%</td>
<td>49.58%</td>
<td>49.31%</td>
<td>82.71%</td>
<td>54.33%</td>
<td>59.99%</td>
</tr>
<tr>
<td>DeepNet</td>
<td>92.43%</td>
<td>49.99%</td>
<td>74.04%</td>
<td>49.48%</td>
<td>87.07%</td>
<td>67.08%</td>
<td>72.12%</td>
</tr>
</tbody>
</table>

Image

FV

DNN
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.
Machine Learning in the Sciences
Machine Learning in Neuroscience
BBCI Set-up: *Let the machines learn*

Multi-channel EEG

- FFT based low-pass filter
- Band-pass 4-40 Hz -> AR coeffs.
- Subject-specific band-pass filter, e.g. 7-14Hz, -> multi-class CSP

Artifact removal

- Multiple feature extraction

Classifier

- Feature combiner 'PROB'

Continuous feedback

\[
\min_{w, b, \xi} \frac{1}{2} \|w\|_2^2 + \frac{C}{K} \|\xi\|_2^2
\]

subject to \( y_k(w^T x_k + b) = 1 - \xi_k \) for \( k = 1, \ldots, K \)

Brain Computer Interfacing: 'Brain Pong'

Berlin Brain Computer Interface

- ML reduces patient training from 300h to 5min

Applications

- help/hope for patients (ALS, stroke…)
- neuroscience
- neurotechnology (video coding, gaming, monitoring driving)

Leitmotiv: ’let the machines learn’
DNN Explanation Motor Imagery BCI

Note: Explanation available for single Trial (Sturm et al 2016)
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]
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]
The data

GDB-13 database of all organic molecules (within stability & synthetic constraints) of 13 heavy atoms or less: 0.9B compounds

Table 1. Structure Generation Statistics for GDB-13

<table>
<thead>
<tr>
<th>nodes</th>
<th>graphs</th>
<th>GDB</th>
<th>Cl/S</th>
<th>CPU time (h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>12</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>43</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>155</td>
<td>3</td>
<td>0.01</td>
</tr>
<tr>
<td>6</td>
<td>20</td>
<td>934</td>
<td>19</td>
<td>0.02</td>
</tr>
<tr>
<td>7</td>
<td>57</td>
<td>5726</td>
<td>315</td>
<td>0.05</td>
</tr>
<tr>
<td>8</td>
<td>194</td>
<td>37151</td>
<td>2438</td>
<td>0.33</td>
</tr>
<tr>
<td>9</td>
<td>706</td>
<td>255542</td>
<td>17056</td>
<td>2.68</td>
</tr>
<tr>
<td>10</td>
<td>2831</td>
<td>1784626</td>
<td>130465</td>
<td>25.26</td>
</tr>
<tr>
<td>11</td>
<td>12011</td>
<td>12961686</td>
<td>938704</td>
<td>223.49</td>
</tr>
<tr>
<td>12</td>
<td>53789</td>
<td>99821343</td>
<td>7240108</td>
<td>3023.79</td>
</tr>
<tr>
<td>13</td>
<td>250268</td>
<td>795244451</td>
<td>59027533</td>
<td>36606.45</td>
</tr>
<tr>
<td>Total</td>
<td>319892</td>
<td>910111673</td>
<td>67356641</td>
<td>39882.08</td>
</tr>
</tbody>
</table>

Blum & Reymond, JACS (2009)
Coulomb representation of molecules

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

\[ M \in ^{23 \times 23} \]

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

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

Distances between \( M \) define Gaussian kernel matrix \( K \)

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

Predict energy as sum over weighted Gaussians

\[
E^{est}(M) = \sum_i \alpha_i k(M, M_i) + b
\]

using weights that minimize error in training set

\[
\min_{\alpha} \sum_i \left( E^{est}(M_i) - E^{ref}_i \right)^2 + \lambda \sum_i \alpha_i^2
\]

Exact solution

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

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

Dataset available at http://quantum-machine.org
Learning Atomistic Representations with Deep Tensor Neural Networks

Kristof Schütt, Farhad Arbabsadah, 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** \( v_i^{(t)} \)

\[
x_i^{(t+1)} = x_i^{(t)} + v_i^{(t)} \left( x_1^{(t)}, \ldots, x_{n_{\text{atoms}}}^{(t)}, d_{i1}, \ldots, d_{i_{\text{in atoms}}} \right)
\]

**Prediction via atom-wise contributions:**

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

**Schütt, Arbabzadah, Chmiela, Müller, Tkatchenko, Nature Communications 8, 13890 (2017)**
\[
\tanh \left( W^f \left( c_j + b^f \circ \left( W^d d_{ij} + b^d \right) \right) \right)
\]
Gaining insights for Physics
Toward Quantum Chemical Insights: supervised

Quantum chemical insights

Inferred chemical potentials

Inferred stable and unstable carbon ring classification ‘aromaticity’
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**

**TCGA data base**

**in-house data base**

- **molecular features**
- **histo-morphological features**

**molecular profiles**

*Available image modalities: brightfield, confocal*

- **cell types**
  - cancer
  - lymphocytes
  - stroma

- **TiLs**

*Red: carcinoma, green: TiLs, blue: molecular property*

**computational fluorescence microscopy**
Heterogeneity of E-Cadherin Expression

Morphology: HE+omics
ML-Prediction: HE+omics
Validation: IHC

Binder et al., in revision.
Take Home messages
Sensitivity analysis is not the question that you would like to ask!
Explanation for simple models does not necessarily work for deep models.
What works for simple models doesn’t work for deep models.

Our LRP method is robust to this.
Layer-Wise Relevance Propagation

Desirable properties of an explanation:
- positivity
- conservation
- selectivity
- continuity

“Tricks of the trade”

Underlying theory for consistency

\[ R_i = \sum_j \frac{\partial R_j}{\partial a_i} \cdot (a_i - \tilde{a}_i^{(j)}) \]

Deep Taylor Decomposition

LRP Explanation Framework

(software, tutorials, demos, insights, applications)
LRP works 4 all: deep models, LSTMs, kernel methods ...
A Clarification on LRP

\[ \text{LRP} \neq \text{Gradient} \times \text{Input} \]

… except for special cases. LRP was developed among others because gradient-based methods aren’t satisfying.

When comparing with LRP, please use appropriate LRP parameters (Like when comparing different ML techniques).

Good news: No need to reimplement LRP, check our software at www.heatmapping.org.
Explanations can be evaluated: Pixel flipping (model agnostic) And beyond LRP and DTD

[Samek et al. IEEE TNNLS 2017]
Explanation helps to improve models

Explaining ML, Now What?
Explanation helps to find flaws in models

[Lapuschnik et al CVPR 2016]
Getting **new** Insights in the Sciences

**Example:** Understanding physical systems at the quantum level.

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

- Hamiltonian
- Energy
- Time-independent Schrödinger Equation
- Equation describing general physical systems

DNN approximation for organic molecules

Interpretation of the trained DNN model

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 use for example interested in potential outliers. Middle: Transferring machine learning techniques estimate the data generating distribution and assign an outlier score \( o(x) \) to unlabeled 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 [11] (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]
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
Explainable AI: Interpreting, Explaining and Visualizing Deep Learning
Further Reading I


Further Reading II


Further Reading III


Further Reading IV


