Selected Methods of explainable AI

Contact: andreas.holzinger AT tuwien.ac.at
LV 185.A83 Machine Learning for Health Informatics (Class of 2020)

Andreas HOLZINGER, Marcus BLOICE, Florian ENDEL, Anna SARANTI

Study Code: 066 936 Master program Medical Informatics

https://tiss.tuwien.ac.at/curriculum/public/curriculum.html?docid=9468&did=255&key=5669&semester=S20X

Semester hours: 2.0 h; ECTS-Credits: 3.0; Type: VU Lecture and Exercises with Python

ECTS-Breakdown (sum=75 h, corresponds with 3 ECTS, where 1 ECTS = 25 h workload):

<table>
<thead>
<tr>
<th>Presence during lecture</th>
<th>8 * 3 h</th>
<th>24 h</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preparation before and after lecture</td>
<td>8 * 1 h</td>
<td>08 h</td>
</tr>
<tr>
<td>Preparation of assignments and presentation</td>
<td>28 h + 2 h</td>
<td>30 h</td>
</tr>
<tr>
<td>Written exam including preparation</td>
<td>1 h + 12 h</td>
<td>13 h</td>
</tr>
<tr>
<td>TOTAL students' workload</td>
<td></td>
<td>75 h</td>
</tr>
</tbody>
</table>

Class URL: https://human-centered.ai/machine-learning-for-health-informatics-class-2020
Class Schedule for 2020 (subject to change: please check class URL for any changes):

<table>
<thead>
<tr>
<th>Nr</th>
<th>Week</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>12</td>
<td>Introduction and overview: From health informatics to ethical responsible medical AI</td>
</tr>
<tr>
<td>02</td>
<td>13</td>
<td>Data for machine learning, Probabilistic information and entropy: On data quality, data integration, data augmentation, information theory</td>
</tr>
<tr>
<td>03</td>
<td>14</td>
<td>Tutorial T01 and Python assignment A01 (Data augmentation) Tutor: Marcus BLOICE</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Happy Easter</td>
</tr>
<tr>
<td>04</td>
<td>17</td>
<td>Probabilistic graphical models: From knowledge representation to graph model learning</td>
</tr>
<tr>
<td>05</td>
<td>18</td>
<td>Tutorial T02 and Python assignment A02 (Probabilistic programming) Tutor: Florian ENDEL</td>
</tr>
<tr>
<td>06</td>
<td>19</td>
<td>Selected methods of explainable AI: LIME, BETA, LRP, Deep Taylor Decomposition, PDA, TCAV etc.</td>
</tr>
<tr>
<td>07</td>
<td>20</td>
<td>Tutorial T03 and Python assignment A03 (LRP) Tutor: Anna SARANTI</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Finalization of assignments and exam preparation</td>
</tr>
<tr>
<td>08</td>
<td>24</td>
<td>Course finalization and grading (detailed information will be given in due course)</td>
</tr>
</tbody>
</table>
- 00 Reflection – follow-up from last lecture
- 01 Explainability, Interpretability, Causability
- 02 Is xAI new?
- 03 Examples for Ante-hoc models (explainable models, interpretable machine learning)
- 04 Examples for Post-hoc models (making the “black-box” model interpretable)
  - 04a LIME, 04b BETA, 04c LRP, 04d Taylor, 04e Prediction Difference Analysis, 04f TCAV
00 Reflection
Warm-up Quiz

1. Deduction

2. Induction

3. Abduction

4. Mathematical expression:

\[ p(\theta, \mathcal{D}) = p(\theta) \prod_{i=1}^{N} p(x_i | \theta) \]
01 Explainability, Interpretability, Causability, ...
Why is explainability so important?

- Explainability = motivated by the opaqueness of so-called “black-box” ML approaches
- It is the ability to provide an explanation on why a machine decision has been reached (e.g. why is it a cat what the deep network recognized).
- Note: Finding an appropriate explanation is difficult, because this needs understanding the context and providing a description of causality and consequences of a given fact.
- German: Erklärbarkeit; siehe auch: Verstehbarkeit, Nachvollziehbarkeit, Zurückverfolgbarkeit, Transparenz
Interpretability := ability to explain or to provide the meaning in understandable terms to a human

Understandability (intelligibility) := characteristic of a model to make a human understand its function – how the model works (without any need for explaining its internal structure).

Comprehensibility := ability of a learning algorithm to represent its learned knowledge in a human understandable fashion entities.

Why is the question of why so important?

Judea Pearl & Dana Mackenzie

http://bayes.cs.ucla.edu/LECTURE/lecture_sec1.htm

https://www.youtube.com/watch?v=pEBI0vF45ic

http://bayes.cs.ucla.edu/WHY
Causality:
The art and science of cause and effect

Causability:
Mapping machine explanations with human understanding

What is the difference between explainability and causability?

- **Explainability** = in a technical sense highlights decision-relevant parts of the used representations of the algorithms and active parts in the algorithmic model, that either contribute to the model accuracy on the training set, or to a specific prediction for one particular observation. **It does not refer to an explicit human model.**

- **Causability** = as the extent to which an explanation of a statement to a human expert achieves a specified level of **causal understanding with effectiveness, efficiency and satisfaction** in a specified context of use.

How can we build successful future Human-AI interfaces?

- Explainability := a property of a system (Computer)
- Causability := a property of a person (Human)

Measuring the quality of Explanations: The Systems Causability Scale

What problems do we face in the real (medical) world?

1) ground truth is not always well defined, especially when making a medical diagnosis;

2) although human (scientific) models are often based on understanding causal mechanisms, today’s successful machine models or algorithms are typically based on correlation or related concepts of similarity and distance!
Why is it important to know to whom to explain something?

**what - to whom - how**

Why is the naïve assumption of the DARPA XAI program wrong?

This is far too naïve: Explainability (better: interpretability !) does not correlate with performance !!
How is Explainability/Interpretability contrasted to performance?

- Explanation is a **reasoning process**
- Open questions:
  - What is a good explanation?
  - When is it enough (degree of saturation)?
  - Context dependent (Emergency vs. research)
  - How can we measure the degree of comprehensibility of a given explanation -> (System Causability Scale, SCS)
  - (obviously the explanation was good when it has been understood by the human)
  - What can the system learn from the human?
  - What can the human learn from the system?
  - Measuring explanation effectiveness!

What are the main expectations to xAI?

- **Causality** - inferring causal relationships from pure observational data has been extensively studied (Pearl, 2009), however it relies strongly on prior knowledge.

- **Transferability** – humans have a much higher capacity to generalize, and can transfer learned skills to completely new situations; compare this with e.g. susceptibility of CNNs to adversarial data (please remember that we rarely have iid data in real world).

- **Informativeness** - for example, a diagnosis model might provide intuition to a human decision-maker by pointing to similar cases in support of a diagnostic decision.

- **Fairness and Ethical decision making** – interpretations for the purpose of assessing whether decisions produced automatically by algorithms conform to ethical standards.

- **Trust AI** – interpretability as prerequisite for trust (as propagated by Ribeiro et al (2016)); how is trust defined? Confidence?

Interpretable Models vs. Interpreting Models

- **Interpretable Glass-Box Models**, the model itself is already interpretable, e.g.
  - Regression
  - Naïve Bayes
  - Random Forests
  - Decision Trees/Graphs
  - ...

- **Interpreting Black-Box Models** (the model is not interpretable and needs a post-hoc interpretability method, e.g.:
  - Decomposition
  - LIME/BETA
  - LRP
  - ...

Advantages/Disadvantages of Rule-based vs. Deep Learning?

- **Rule-Based Models (e.g. decision trees):**
  - Easy to interpret, the rules provide clear explanations
  - Can learn even from little data sets
  - Problems with high-dimensional data, with noise, and with images (ambiguity)

- **Neuro-Symbolic Models (e.g. CNN):**
  - Not easy or even impossible to interpret
  - Needs a lot of top-quality training data
  - Can well generalize even from high-dimensional data, with noise and good for images
So, is xAI a good name for it?

02 Is xAI new?

Explainability was the most requested feature of early medical decision support systems!

How did MYCIN enable explainability?

IF: 1) THE STAIN OF THE ORGANISM IS GRAMNEG, AND
2) THE MORPHOLOGY OF THE ORGANISM IS ROD, AND
3) THE AEROBICITY OF THE ORGANISM ISANAEROBIC
THEN: THERE IS SUGGESTIVE EVIDENCE (.6) THAT THE IDENTITY
OF THE ORGANISM IS BACTEROIDES

How did MYCIN support WHY questions?

** WHAT IS THE IDENTITY OF ORGANISM-1? 
THE IDENTITY OF ORGANISM-1 IS E. COLI (0.30) PROTEUS (0.26) PSEUDOMONAS (0.24) KLEBSIELLA (0.21)

** HOW DID YOU DECIDE THAT ORGANISM-1 MIGHT BE A PSEUDOMONAS? 
I USED RULE085 TO CONCLUDE THAT THE IDENTITY OF ORGANISM-1 IS PSEUDOMONAS. THIS GAVE A CUMULATIVE CERTAINTY FACTOR OF 0.24.

IF: 1) THE STAIN OF THE ORGANISM IS GRAMNEG, AND 
2) THE MORPHOLOGY OF THE ORGANISM IS ROD, AND 
3) THE AEROBICITY OF THE ORGANISM IS ANAEROBIC
THEN: THERE IS SUGGESTIVE EVIDENCE (.6) THAT THE IDENTITY OF THE ORGANISM IS BACTEROIDES

** DO YOU EVER PRESCRIBE CARBENICILLIN FOR PSEUDOMONAS INFECTIONS? 
RELEVANT RULES: RULE064 RULE119 RULE120 WHICH ONES DO YOU WISH TO SEE? 
** RULE064 
IF: 1) IT IS KNOWN WITH CERTAINTY THAT THE IDENTITY OF THE ORGANISM IS PSEUDOMONAS, AND 
2) THE DRUG UNDER CONSIDERATION IS GENTAMICIN
THEN: RECORD THE FOLLOWING AS A MORE APPROPRIATE THERAPY: GENTAMICIN-AND-CARBENICILLIN
The Question-Answering Module

** Is either chloramphenicol or kanamycin okay for treatment of salmonella infections?
YES.

RULE083

IF: 
THE IDENTITY OF THE ORGANISM IS SALMONELLA

THEN:
I RECOMMEND (IN ORDER OF PREFERENCE):
1-CHLORAMPHENICOL
2-AMPICILLIN
3-CEPHALOTHIN

** How do you know if an organism is a bacteroides?

RULE030

IF: 
THE GRAMSTAIN OF THE ORGANISM IS GRAMNEG, AND THE
MORPHOLOGY OF THE ORGANISM IS ROD, AND THE
AEROBICYCITY OF THE ORGANISM IS ANAEROBIC

THEN:
CONCLUDE THAT THE IDENTITY OF THE ORGANISM IS
BACTEROIDES (MODIFIER: THE CERTAINTY TALLY FOR
THE PREMISE TIMES .6)
What caused the end of the hype in early AI DSS?

- The evaluation demonstrated that the present form of the program is not sufficiently reliable for clinical applications.
- Specific deficiencies that must be overcome include the program's inability to reason, its inability to construct differential diagnoses spanning multiple areas, its occasional attribution of findings to improper causes, and its inability to explain its "thinking" ...

Digression: History of DSS = History of AI
A ultrashort history of Early AI – the golden Years


- **1959** John McCarthy: Programs with common sense. Mechanization of thought processes (Advice Taker)


- **1978** Bellman, R. Can Computers Think? Automation of Thinking, problem solving, decision-making ...
Towards the renaissance of AI

- **1997** Deep Blue beats Geri Kasparov
- **2009** Successful autonomous driving
- **2011** IBM Watson in Jeopardy
What about the history of Health Informatics?

- 1960+ Medical Informatics (Classic AI Hype)
  - Focus on data acquisition, storage, accounting, Expert Systems
  - The term was first used in 1968 and the first course was set up in 1978!

- 1985+ Health Telematics (AI winter)
  - Health care networks, Telemedicine, CPOE-Systems, ...

- 1995+ Web Era (AI is “forgotten”)
  - Web based applications, Services, EPR, distributed systems, ...

- 2005+ Success statistical learning (AI renaissance)
  - Pervasive, ubiquitous Computing, Internet of things, ...

- 2010+ Data Era – Big Data (super for AI)
  - Massive increase of data – data integration, mapping, ...

- 2020+ Explanation Era – (towards explainable AI)
  - Re-traceability, replicability, reenactment, explainability, interpretability, sensemaking, disentangling the underlying concepts, causality, causability, human-AI interfaces, ethical responsible machine learning, trust-AI...
Why was MYCIN central for explainable AI in medicine?

How did the human-AI interaction work?


Find an emulation and a Jupyter notebook here: http://user.medunigraz.at/marcus.bloice/seminars/dss/g3/g3.htm
What was static knowledge versus dynamic knowledge?

The information available in medicine is often imperfect – imprecise - uncertain.

**Human experts** can cope with deficiencies.

Classical logic permits only **exact reasoning**:

- IF A is true THEN A is non-false and
- IF B is false THEN B is non-true

Most real-world problems do not provide this exact information, mostly it is inexact, incomplete, uncertain and/or **un-measurable**!
To every rule and every entry a certainty factor (CF) is assigned, which is between 0 and 1.

Two measures are derived:
- **MB**: measure of belief
- **MD**: measure of disbelief

Certainty factor – CF of an element is calculated by:
\[ CF[h] = MB[h] - MD[h] \]

CF is positive, if more evidence is given for a hypothesis, otherwise CF is negative.
- \( CF[h] = +1 \rightarrow h \) is 100% true
- \( CF[h] = -1 \rightarrow h \) is 100% false

How did MYCIN cope with uncertainties?
h₁ = The identity of ORGANISM-1 is streptococcus
h₂ = PATIENT-1 is febrile
h₃ = The name of PATIENT-1 is John Jones

CF[h₁,E] = .8 : There is strongly suggestive evidence (.8) that the identity of ORGANISM-1 is streptococcus

CF[h₂,E] = −.3 : There is weakly suggestive evidence (.3) that PATIENT-1 is not febrile

CF[h₃,E] = +1 : It is definite (1) that the name of PATIENT-1 is John Jones

Why was MYCIN *not* a success in the clinical routine?
So, what accelerated the AI renaissance?

Why are large neural networks a problem?


Won the 2012 ImageNet LSVRC. 60 Million parameters, 832M MAC ops

<table>
<thead>
<tr>
<th>Layer Type</th>
<th>Parameters</th>
<th>MAC Ops</th>
</tr>
</thead>
<tbody>
<tr>
<td>FULL CONNECT</td>
<td>4M</td>
<td>4Mflop</td>
</tr>
<tr>
<td>FULL 4096/ReLU</td>
<td>16M</td>
<td>16M</td>
</tr>
<tr>
<td>FULL 4096/ReLU</td>
<td>37M</td>
<td>37M</td>
</tr>
<tr>
<td>MAX POOLING</td>
<td>442K</td>
<td>74M</td>
</tr>
<tr>
<td>CONV 3x3/ReLU 256fm</td>
<td>1.3M</td>
<td>224M</td>
</tr>
<tr>
<td>CONV 3x3/ReLU 384fm</td>
<td>884K</td>
<td>149M</td>
</tr>
<tr>
<td>MAX POOLING 2x2sub</td>
<td>307K</td>
<td>223M</td>
</tr>
<tr>
<td>LOCAL CONTRAST NORM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CONV 11x11/ReLU 256fm</td>
<td>35K</td>
<td>105M</td>
</tr>
<tr>
<td>LOCAL CONTRAST NORM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CONV 11x11/ReLU 96fm</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Image credit to Yann LeCun, ICML 2013 Deep Learning Tutorial
Why are such black-box models now a problem?
Success in deep learning *) resulted in “deep problems” (e.g. complex and exploding gradients)

*) Note: “DL” methods are representation learning methods with multiple layers of representations (see LeCun, Bengio & Hinton (2015), Nature 521, 7553)

Problem in our society: “Secret algorithms” make important decisions about individuals

Black box Type 1 = too complicated for a human to understand

Black box Type 2 = proprietary = “secret algorithm”

Which two top-level explainable AI methodologies do we have?

- **Post-Hoc** (latin) = after-this (event), i.e. such approaches provide an explanation for a specific solution of a “black-box” approach, e.g. LIME, BETA, LRP, ...

- **Ante-hoc** (latin) = before-this (event), i.e. such methods can be (human) interpreted immanently in the system, i.e. they are transparent by nature (glass box), similar to the "interactive machine Learning" (iML) model.

Combination of Deep Learning with Ontologies

(1) Explaining the reasons for judgment

Deep Tensor

Output both inference result and reasons (inference factors)

Input

Inference factor identification

Output

Inference result

Inference factors

Knowledge Graph

Knowledge graph generates a logical path from input to the inference result

Basis formation

(2) Explaining the basis (evidence) for judgment

Explainable AI with Deep Tensor and Knowledge Graph

03 Examples for Ante Hoc Models (interpretable Machine Learning)
Are ante-hoc approaches new?

- **Post-Hoc** (Latin) = after-this (event), i.e. such approaches provide an explanation for a specific solution of a “black-box” approach, e.g. LIME, BETA, LRP, ... (see module 5)

- **Ante-hoc** (Latin) = before-this (event), i.e. such methods can be (human) interpreted immanently in the system, i.e. they are transparent by nature (glass box), similar to the "interactive machine Learning" (iML) model.

- Note: Many ante-hoc approaches appear to the new student particularly novel, but these have a long tradition and were used since the early beginning of AI and applied in expert systems; typical methods decision trees, linear regression, Random Forests, ...

How does an action Influence graph support explainability?

State variables:
- W - Worker number
- S - Supply depot number
- B - barracks number
- E - enemay location
- A_n - Ally unit number
- A_h - Ally unit health
- A_l - Ally unit location
- D_u - Destroyed units
- D_b - Destroyed buildings

Actions:
- A_s - build supply depot
- A_b - build barracks
- A_m - train offensive unit
- A_a - attack

What are Stochastic AND-OR graphs?

Zhangzhang Si & Song-Chun Zhu 2013. Learning and-or templates for object recognition and detection. IEEE transactions on pattern analysis and machine intelligence, 35, (9), 2189-2205, doi:10.1109/TPAMI.2013.35.
Example: Bayesian Rule Lists

04 Examples for Post Hoc Models (e.g. LIME, BETA, LRP)
**Differences: Post-hoc versus Ante-hoc**

- **Post-Hoc** (latin) = after-this (event), i.e. such approaches provide an explanation for a specific solution of a “black-box” approach, e.g. LIME, BETA, LRP, ...

- **Ante-hoc** (latin) = before-this (event), i.e. such methods can be (human) interpreted immanently in the system, i.e. they are transparent by nature (glass box), similar to the "interactive machine Learning" (iML) model.

- Note: Many ante-hoc approaches appear to the new student particularly novel, but these have a long tradition and were used since the early beginning of AI and applied in expert systems (see module 3); typical methods decision trees, linear regression, and Random Forests.

Caveat – Post hoc explanation can be misleading!

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

Cynthia Rudin

Black box machine learning models are currently being used for high-stakes decision making throughout society, causing problems in healthcare, criminal justice and other domains. Some people hope that creating methods for explaining these black box models will alleviate some of the problems, but trying to explain black box models, rather than creating models that are interpretable in the first place, is likely to perpetuate bad practice and can potentially cause great harm to society. The way forward is to design models that are inherently interpretable. This Perspective clarifies the chasm between explaining black boxes and using inherently interpretable models, outlines several key reasons why explainable black boxes should be avoided in high-stakes decisions, identifies challenges to interpretable machine learning, and provides several example applications where interpretable models could potentially replace black box models in criminal justice, healthcare and computer vision.
What are typical post-hoc approaches

- 1) Gradients
- 2) Sensitivity Analysis
- 4) Optimization (Local-IME – model agnostic, BETA transparent approximation, ...)
- 5) Deconvolution and Guided Backpropagation
- 6) Model Understanding
  - Feature visualization, Inverting CNN
  - Qualitative Testing with Concept Activation Vectors TCAV
  - Network Dissection
04a LIME – Local Interpretable Model Agnostic Explanations
What is the general principle of LIME?

What are explanations in the sense of LIME?

- Explanation := local linear approximation of the model's behaviour. While the model may be very complex globally, it is easier to approximate it around the vicinity of a particular instance.

Algorithm 1 Sparse Linear Explanations using LIME

```
Require: Classifier $f$, Number of samples $N$
Require: Instance $x$, and its interpretable version $x'$
Require: Similarity kernel $\pi_x$, Length of explanation $K$

$\mathcal{Z} \leftarrow \{\}$

for $i \in \{1, 2, 3, ..., N\}$ do
  $z'_i \leftarrow \text{sample\_around}(x')$
  $Z \leftarrow Z \cup (z'_i, f(z_i), \pi_x(z_i))$
end for

$w \leftarrow K\text{-Lasso}(Z, K)$ \triangleright with $z'_i$ as features, $f(z)$ as target

return $w$
```

How is the explanation produced by LIME?

\[ \xi(x) = \arg \min_{g \in G} L(f, g, \pi_x) + \Omega(g) \]

- \( L(f, g, \pi_x) \): Fidelity score (for local fidelity)
- \( \Omega(g) \): Complexity score (for interpretability)

\[ \pi_x(Z) \]: Distance metric (in feature space)
How does a typical LIME example look like?

Here we will take a sample from the test set (in this case the sample at index 76) and create an explainer instance for this sample. This will let us see why the algorithm made its prediction visually.

As you can see, the random forest algorithm has predicted with a probability of 0.64 that the sample at index 76 in the test set is malignant.

When using the explainer, we set the `num_features` parameter to 4, meaning the explainer shows the top 4 features that contributed to the prediction probabilities.

We chose 76 as it was a borderline decision. For example sample 86 is much more clear (this will we will set the `num_features` parameter to include all features so that we see each feature's contribution to the probability):
LIME Example

Original Image
P(tree frog) = 0.54

Perturbed Instances | P(tree frog)
--- | ---
| 0.85
| 0.00001
| 0.52

Locally weighted regression
Explanation

LIME Example

P( ) = 0.54

P( ) = 0.07

P( ) = 0.05

What are the LIME Pros and Cons?

+ very popular,
+ many applications and contributors
+ model agnostic

- local model behaviour can be unrealistic
- unclear coverage
- ambiguity (how to select the kernel width?)
Remember: there are myriads of classifiers ...

What is the difference to the Follow-up: Anchor

What is GraphLIME?

Algorithm 1 Locally nonlinear Explanation: GraphLIME

| Input: GNN classifier $f$, Number of explanation features $K$ |
| Input: the graph $G$, the node $x$ being explained |
| Output: $K$ explanation features |

1: $X_n = \text{N-hop-neighbor-sample}(x)$
2: $Z \leftarrow \{\}$
3: for all $x_i \in X_n$ do
4: \quad $y_i = f(x_i)$
5: \quad $Z \leftarrow Z \cup (x_i, y_i)$
6: end for
7: $\beta \leftarrow \text{HSIC Lasso}(Z)$ \ implies with $x_i$ as features, $y_i$ as label
8: Select top-$K$ features as explanations based on $\beta$
04b BETA (Black Box Explanation through Transparent Approximation)
BETA is a model agnostic approach to explain the behaviour of an (arbitrary) black box classifier (i.e. a function that maps a feature space to a set of classes) by simultaneously optimizing the accuracy of the original model and interpretability of the explanation for a human.

Note: Interpretability and accuracy at the same time are difficult to achieve.

Consequently, users are interactively integrated into the model and can thus explore the areas of black box models that interest them (most).
If $\text{Age} < 50$ and $\text{Male} = \text{Yes}$:

- If $\text{Past-Depression} = \text{Yes}$ and $\text{Insomnia} = \text{No}$ and $\text{Melancholy} = \text{No}$, then $\text{Healthy}$
- If $\text{Past-Depression} = \text{Yes}$ and $\text{Insomnia} = \text{Yes}$ and $\text{Melancholy} = \text{Yes}$ and $\text{Tiredness} = \text{Yes}$, then $\text{Depression}$

If $\text{Age} \geq 50$ and $\text{Male} = \text{No}$:

- If $\text{Family-Depression} = \text{Yes}$ and $\text{Insomnia} = \text{No}$ and $\text{Melancholy} = \text{Yes}$ and $\text{Tiredness} = \text{Yes}$, then $\text{Depression}$
- If $\text{Family-Depression} = \text{No}$ and $\text{Insomnia} = \text{No}$ and $\text{Melancholy} = \text{No}$ and $\text{Tiredness} = \text{No}$, then $\text{Healthy}$

Default:

- If $\text{Past-Depression} = \text{Yes}$ and $\text{Tiredness} = \text{No}$ and $\text{Exercise} = \text{No}$ and $\text{Insomnia} = \text{Yes}$, then $\text{Depression}$
- If $\text{Past-Depression} = \text{No}$ and $\text{Weight-Gain} = \text{Yes}$ and $\text{Tiredness} = \text{Yes}$ and $\text{Melancholy} = \text{Yes}$, then $\text{Depression}$
- If $\text{Family-Depression} = \text{Yes}$ and $\text{Insomnia} = \text{Yes}$ and $\text{Melancholy} = \text{Yes}$ and $\text{Tiredness} = \text{Yes}$, then $\text{Depression}$

How does the Optimization Process generally work?

\[
\arg \max_{\mathcal{R} \subseteq \mathcal{D} \times \mathcal{D} \times \mathcal{C}} \sum_{i=1}^{3} \lambda_i f_i(\mathcal{R})
\]

s.t. \text{size}(\mathcal{R}) \leq \epsilon_1, \text{maxwidth}(\mathcal{R}) \leq \epsilon_2, \text{numdsets}(\mathcal{R}) \leq \epsilon_3

Algorithm 1 Optimization Procedure [5]

1: \textbf{Input:} Objective \( f \), domain \( \mathcal{D} \times \mathcal{D} \times \mathcal{C} \), parameter \( \delta \), number of constraints \( k \)
2: \( V_1 = \mathcal{D} \times \mathcal{D} \times \mathcal{C} \)
3: for \( i \in \{1, 2 \cdots k + 1\} \) do \quad \triangleright \text{Approximation local search procedure}
4: \quad X = V_i; n = |X|; S_i = \emptyset
5: \quad \text{Let } \nu \text{ be the element with the maximum value for } f \text{ and set } S_i = \nu
6: \text{while there exists a delete/update operation which increases the value of } S_i \text{ by a factor of}
\text{at least } (1 + \frac{\delta}{n^4}) \text{ do}
7: \quad \text{Delete Operation: If } e \in S_i \text{ such that } f(S_i \backslash \{e\}) \geq (1 + \frac{\delta}{n^4}) f(S_i), \text{ then } S_i = S_i \backslash e
8: \text{end while}
9: \quad \text{Exchange Operation} \text{ If } d \in X \backslash S_i \text{ and } e_j \in S_i \text{ (for } 1 \leq j \leq k) \text{ such that}
10: \quad (S_i \backslash \{e_j\} \cup \{d\} \text{ (for } 1 \leq j \leq k) \text{ satisfies all the } k \text{ constraints and}
11: \quad f(S_i \backslash \{e_1, e_2, \cdots e_k\} \cup \{d\}) \geq (1 + \frac{\delta}{n^4}) f(S_i), \text{ then } S_i = S_i \backslash \{e_1, e_2, \cdots e_k\} \cup \{d\}
12: \text{end for}
13: V_{i+1} = V_i \backslash S_i
14: \text{end for}
15: \text{return} \text{ the solution corresponding to } \max \{f(S_1), f(S_2), \cdots f(S_{k+1})\}

What are the Measures of Fidelity, Interpretability, Unambiguity?

| Fidelity | \(disagreement(\mathcal{R}) = \sum_{i=1}^{M} |\{x \mid x \in \mathcal{D}, x \text{ satisfies } q_i \land s_i, B(x) \neq c_i\}|\) |
|---|---|
| Unambiguity | \(\text{ruleoverlap}(\mathcal{R}) = \sum_{i=1}^{M} \sum_{j=1, i \neq j} \text{overlap}(q_i \land s_i, q_j \land s_j)\) |
| | \(\text{cover}(\mathcal{R}) = |\{x \mid x \in \mathcal{D}, x \text{ satisfies } q_i \land s_i \text{ where } i \in \{1 \ldots M\}\}|\) |
| | \(\text{size}(\mathcal{R}): \text{number of rules (triples of the form } (q, s, c) \text{) in } \mathcal{R}\) |
| Interpretable | \(\text{maxwidth}(\mathcal{R}) = \max_{e \in \bigcup_{i=1}^{M} (q_i \cup s_i)} \text{width}(e)\) |
| | \(\text{numpreds}(\mathcal{R}) = \sum_{i=1}^{M} \text{width}(s_i) + \text{width}(q_i)\) |
| | \(\text{numdsets}(\mathcal{R}) = |\text{dset}(\mathcal{R})| \text{ where } \text{dset}(\mathcal{R}) = \bigcup_{i=1}^{M} q_i\) |
| | \(\text{featureoverlap}(\mathcal{R}) = \sum_{q \in \text{dset}(\mathcal{R})} \sum_{i=1}^{M} \text{featureoverlap}(q, s_i)\) |
**BETA: Example of interpretable Decision set**

If Respiratory-Illness = Yes and Smoker = Yes and Age ≥ 50 then Lung Cancer
If Risk-LungCancer = Yes and Blood-Pressure ≥ 0.3 then Lung Cancer
If Risk-Depression = Yes and Past-Depression = Yes then Depression
If BMI ≥ 0.3 and Insurance = None and Blood-Pressure ≥ 0.2 then Depression
If Smoker = Yes and BMI ≥ 0.2 and Age ≥ 60 then Diabetes
If Risk-Diabetes = Yes and BMI ≥ 0.4 and Prob-Infections ≥ 0.2 then Diabetes
If Doctor-Visits ≥ 0.4 and Childhood-Obesity = Yes then Diabetes

---

**Notation**

<table>
<thead>
<tr>
<th>Notation</th>
<th>Definition</th>
<th>Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mathcal{D} )</td>
<td>Input set of data points ( { (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N) } )</td>
<td>Dataset</td>
</tr>
<tr>
<td>( \mathbf{x} )</td>
<td>Observed attribute values of a data point</td>
<td></td>
</tr>
<tr>
<td>( y )</td>
<td>Class label of a data point</td>
<td></td>
</tr>
<tr>
<td>( \mathcal{C} )</td>
<td>Set of class labels in ( \mathcal{D} )</td>
<td></td>
</tr>
<tr>
<td>( p )</td>
<td>Predicate ( (\text{attribute}, \text{operator}, \text{value}) ) ( e.g., \text{Age} \geq 50 )</td>
<td></td>
</tr>
<tr>
<td>( s )</td>
<td>Conjunction of one or more predicates, ( e.g., \text{Age} \geq 50 ) and ( \text{Gender} = \text{Female} )</td>
<td>Itemset</td>
</tr>
<tr>
<td>( \mathcal{S} )</td>
<td>Input set of itemsets</td>
<td></td>
</tr>
<tr>
<td>( r )</td>
<td>Itemset-class pair ( (s, c) )</td>
<td>Rule</td>
</tr>
<tr>
<td>( \mathcal{R} )</td>
<td>Set of rules ( { (s_1, c_1), \ldots, (s_k, c_k) } )</td>
<td>Decision set</td>
</tr>
</tbody>
</table>

---

https://himalakkaraju.github.io

What are for BETA Pros and Cons?

+ model agnostic
+ learns a compact two-level decision set
+ unambiguously

- not so popular
- unclear coverage
- needs care
04c LRP (Layer-wise Relevance Propagation)
How can we describe LRP at a glance?

- LRP is a general solution for understanding classification decisions by pixel-by-pixel (or layer-by-layer) decomposition of nonlinear classifiers (hence the name).
- In a highly simplified way, LRP allows the "thinking processes" of neural networks to run backwards.
- Thereby it becomes comprehensible (for a human) which input had which influence on the respective result,
- e.g. in individual cases how the neural network came to a classification result, i.e. which input contributed most to the gained output.
- Example: If genetic data is entered into a network, it is not only possible to analyze the probability of a patient having a certain genetic disease, but with LRP also the characteristics of the decision.
- Such an approach is a step towards personalised medicine (remember the concept of PM - to provide an individual cancer therapy that is tailored to the particular patient).
How does LRP work in principle?

\[
a_{ij}^{(l+1)} = \sigma \left( \sum_i a_{i}^{(l)} w_{ij} + b_{j}^{(l+1)} \right)
\]

\[
R_i^{(l)} = \sum_j \frac{z_{ij}}{\sum_i z_{ij}} R_j^{(l+1)}
\]

Forward propagation:
\[
R_i = \left\| \frac{\partial}{\partial x_i} f(x) \right\|
\]

\[
\sum_i R_i = \ldots = \sum_j R_j = \sum_k R_k = \ldots = f(x)
\]
How does a NN-classifier during prediction time look like?

\[ f(x) = \cdots = \sum_{d \in l+1} R_{d}^{(l+1)} = \sum_{d \in l} R_{d}^{(l)} = \cdots = \sum_{d} R_{d}^{(1)} \]

What is the output of the LRP?

Pixel-wise decomposition for bag-of-words features

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller & Wojciech Samek
What is relevant in a text document?

Example: What is relevant in a text document?

**CNN2**

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 and see how to keep the number of occurrences down.

**SVM**

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 and see how to keep the number of occurrences down.

PCA-Projection of the summary vectors

doi:10.1371/journal.pone.0181142.

human-centered.ai (Holzinger Group)
04d  Deep Taylor Decomposition
What are Taylor series?

https://www.youtube.com/watch?v=3d6DsJlBzJ4

\[ f(x) = f(a) + \left( \frac{df}{dx} \right)_{x=a} \cdot (x - a) + \epsilon = 0 + \sum_p \frac{d^p f}{dx^p} \bigg|_{x=a} \cdot (x - a)^p + \epsilon, \]

\[ R(x) = \frac{d^p f}{dx^p} \bigg|_{x=a} \]

https://en.wikipedia.org/wiki/Brook_Taylor

Brook Taylor (1685-1731)

- **Born**: 18 August 1685, Edmonton, Middlesex, England
- **Died**: 29 December 1731 (aged 46), London, England
- **Residence**: England
- **Nationality**: English
- **Alma mater**: St John's College, Cambridge
- **Known for**: Taylor's theorem, Taylor series
What is Taylor decomposition at a glance?

http://www.heatmapping.org/deeptaylor
**Definition 1.** A heatmapping $R(x)$ is *conservative* if the sum of assigned relevances in the pixel space corresponds to the total relevance detected by the model:

$$\forall x: f(x) = \sum_{p} R_p(x).$$

**Definition 2.** A heatmapping $R(x)$ is *positive* if all values forming the heatmap are greater or equal to zero, that is:

$$\forall x, p: R_p(x) \geq 0$$

**Definition 3.** A heatmapping $R(x)$ is *consistent* if it is conservative and positive. That is, it is consistent if it complies with Definitions 1 and 2.

What about Sensitivity Analysis vs. Decomposition?

\[ \sum_p (\frac{\partial f}{\partial x_p})^2 = \| \nabla_x f(x) \|^2 \]

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

function to analyze:
\[ f(x) = \max(0, x_1) + \max(0, x_2) \]

sensitivity analysis:
\[ (\frac{\partial f}{\partial x_1})^2 = 1_{x_1 > 0} \]
\[ (\frac{\partial f}{\partial x_2})^2 = 1_{x_2 > 0} \]

decomposition:
\[ R_1(x) = \max(0, x_1) \]
\[ R_2(x) = \max(0, x_2) \]
How is the computational flow of deep Taylor decompression in detail?

How does the Relevance Redistribution work?

Example 1: Comparison
Example 2 Histopathology


Example: Deep Learning from histopatho images explainable

04e Prediction Difference Analysis
How can the relevance of a feature be measured?

\[ p(c|x_i) = \sum_{x_i} p(x_i|x_i)p(c|x_i, x_i) \]

\[ p(c|x_i) \approx \sum_{x_i} p(x_i)p(c|x_i, x_i) \]

\[ WE_i(c|x) = \log_2 \left( \text{odds}(c|x) \right) - \log_2 \left( \text{odds}(c|x_i) \right) \]


https://github.com/lmzintgraf/DeepVis-PredDiff/blob/master/README.md

https://openreview.net/forum?id=BJ5UeU9xx
How to evaluate the prediction difference?

Algorithm 1 Evaluating the prediction difference using conditional and multivariate sampling

**Input:** classifier with outputs $p(c|x)$, input image $x$ of size $n \times n$, inner patch size $k$, outer patch size $l > k$, class of interest $c$, probabilistic model over patches of size $l \times l$, number of samples $S$

**Initialization:** $WE = \text{zeros}(n^2)$, $\text{counts} = \text{zeros}(n^2)$

**for** every patch $x_w$ of size $k \times k$ in $x$ **do**

$x' = \text{copy}(x)$

$\text{sum}_w = 0$

define patch $\hat{x}_w$ of size $l \times l$ that contains $x_w$

**for** $s = 1$ to $S$ **do**

$x'_w \leftarrow x_w$ sampled from $p(x_w|\hat{x}_w\setminus x_w)$

$\text{sum}_w += p(c|x')$

**end for**

$p(c|x\setminus x_w) := \text{sum}_w / S$

$WE[\text{coordinates of } x_w] += \log_2(\text{odds}(c|x)) - \log_2(\text{odds}(c|x\setminus x_w))$

$\text{counts}[\text{coordinates of } x_w] += 1$

**end for**

**Output:** $WE / \text{counts}$

▷ point-wise division
What is Superpixel-based prediction difference analysis?

Digression: Visualizing CNN with Deconvolution
How to get insight into a deep neural network?

Is the world compositional?

04f Testing with Concept Activation Vectors (TCAV)
“It’s not enough to know if a model works, we need to know how it works”

... if Sundar Pichai is saying this ...
When does an image belong to the class concept of doctors?

https://www.youtube.com/watch?v=lyRPyRKHO8M&t=3408s
ML models work on **low-level features** (edges, dots, lines, pixel, circles, ...)

Humans are working on **high-level concepts** (shape, size, color, Gestalt-principles, ...)

Every pixel of an image is a input feature and are just numbers, which do not make sense to humans.

TCAV enables to provide an explanation that is generally true for a class of interest, beyond one image (global explanation).

The goal of TCAV is to learn ‘concepts’ from examples.
How does a Concept Activation Vector (CAV) work?

Humans work in another vector space which is spanned by implicit knowledge vectors corresponding to an unknown set of human interpretable concepts.

\[
\frac{\partial h_k(x)}{\partial x_{a,b}}
\]

\[
S_{C,k,l}(x) = \lim_{\epsilon \to 0} \frac{h_{l,k}(f_l(x) + \epsilon v_C^l) - h_{l,k}(f_l(x))}{\epsilon} = \nabla h_{l,k}(x)
\]


3.3. Directional Derivative

Interpretability methods of logit values with respect to pixels, and compute

\[
\nabla h_{l,k}(x) = \lim_{\epsilon \to 0} \frac{h_{l,k}(f_l(x) + \epsilon v_C^l) - h_{l,k}(f_l(x))}{\epsilon}
\]

where \(h_k(x)\) is the logit of class \(k\), \(x_{a,b}\) is a pixel at position \(a, b\), \(v_C^l\) is a unit vector and \(f_l(x)\) the activation of class \(l\). The derivative measures the sensitivity of \(h\) towards the direction of \(v_C^l\). By using CAVs and directional derivatives, one can measure the sensitivity of \(h\) towards the direction of \(v_C^l\). If \(v_C^l \in \mathbb{R}^m\) is a unit vector, then \(f_l(x)\) the activation of class \(l\) and \(S_{C,k,l}(x)\) the "directional sensitivity" of class \(k\) to the direction of \(v_C^l\).

\[
S_{C,k,l}(x) = \lim_{\epsilon \to 0} \frac{h_{l,k}(f_l(x) + \epsilon v_C^l) - h_{l,k}(f_l(x))}{\epsilon}
\]

where \(h_{l,k} : \mathbb{R}^m \to \mathbb{R}\) is the logit of class \(k\) and \(f_l(x)\) the activation of class \(l\). The directional derivative measures the sensitivity of \(h\) towards the direction of \(v_C^l\). By using CAVs and directional derivatives, one can measure the sensitivity of \(h\) towards the direction of \(v_C^l\). If \(v_C^l \in \mathbb{R}^m\) is a unit vector, then \(f_l(x)\) the activation of class \(l\) and \(S_{C,k,l}(x)\) the "directional sensitivity" of class \(k\) to the direction of \(v_C^l\).
https://github.com/tensorflow/tcav
How can TCAV applied in the medical domain?

Digression: Sensitivity Analysis
Sensitivity analysis (SA) is a classic, versatile and broad field with long tradition and can be used for a variety of different purposes, including:

- Robustness testing (very important for ML)
- Understanding the relationship between input and output
- Studying and reducing uncertainty

Remember: NN = nonlinear function approximators using gradient descent to minimize the error in such a function approximation.

To students this seems to be “new” – but it has a long history:

- Chain rule = back-propagation was invented by Leibniz (1676) and L’Hopital (1696).
- Calculus and Algebra have long been used to solve optimization problems and gradient descent was introduced by Cauchy (1847).
- This was used to fuel machine learning in the 1940ies > perceptron – but was limited to linear functions, therefore.
- Learning nonlinear functions required the development of a multilayer perceptron and methods to compute the gradient through such a model.
- This was elaborated by LeCun (1985), Parker (1985), Rumelhart (1986) and Hinton (1986).
What are Saliency Maps?

What are Saliency Maps?


What is the main principle of Sensitivity Analysis?

- Let us consider a function \( f \),
- a data point \( x = (x_1, \ldots, x_d) \) and the prediction \( f(x_1, \ldots, x_d) \).
- Now, SA measures the local variation of the function along each input dimension:
  \[
  Ri = \left( \frac{\partial f}{\partial x_i} \big|_{x = x} \right)^2
  \]
- With other words, SA produces local explanations for the prediction of a differentiable function \( f \) using the squared norm of its gradient w.r.t. the inputs \( x : S(x) / k \| \nabla_x f \| ^2 \).
- The saliency map \( S \) produced with this method describes the extent to which variations in the input would produce a change in the output \( S(x) \propto \| \nabla_x f \|^2 \).

Given an image classification (ConvNet), we aim to answer two questions:
- What does a class model look like?
- What makes an image belong to a class?

To this end, we visualise:
- Canonical image of a class
- Class saliency map for a given image and class

Both visualisations are based on the class score derivative w.r.t. the input image (computed using back-prop)
Thank you!
Glossar (1/3)

- Ante-hoc Explainability (AHE) = such models are interpretable by design, e.g. glass-box approaches; typical examples include linear regression, decision trees/lists, random forests, Naive Bayes and fuzzy inference systems; or GAMs, Stochastic AOGs, and deep symbolic networks; they have a long tradition and can be designed from expert knowledge or from data and are useful as framework for the interaction between human knowledge and hidden knowledge in the data.

- BETA = Black Box Explanation through Transparent Approximation, developed by Lakkarju, Bach & Leskovec (2016) it learns two-level decision sets, where each rule explains the model behaviour; this is an increasing problem in daily use of AI/ML, see e.g. [http://news.mit.edu/2019/better-fact-checking-fake-news-1017](http://news.mit.edu/2019/better-fact-checking-fake-news-1017)

- Bias = inability for a ML method to represent the true relationship; High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting);

- Causability = is a property of a human (natural intelligence) and a measurement for the degree of human understanding; we have developed a causability measurement scale (SCS).

- Decomposition = process of resolving relationships into the constituent components (hopefully representing the relevant interest). Highly theoretical, because in real-world this is hard due to the complexity (e.g. noise) and untraceable imponderabilities on our observations.

- Deduction = deriving of a conclusion by reasoning

- Explainability = motivated by the opaqueness of so called “black-box” approaches it is the ability to provide an explanation on why a machine decision has been reached (e.g. why is it a cat what the deep network recognized). Finding an appropriate explanation is difficult, because this needs understanding the context and providing a description of causality and consequences of a given fact. (German: Erklärbarkeit; siehe auch: Verstehbarkeit, Nachvollziehbarkeit, Zurückverfolgbarkeit, Transparenz)
Explanation = set of statements to describe a given set of facts to clarify causality, context and consequences thereof and is a core topic of knowledge discovery involving “why” questionss (“Why is this a cat?”). (German: Erklärung, Begründung)

Explanatory power = is the ability of a set hypothesis to effectively explain the subject matter it pertains to (opposite: explanatory impotence).

Explicit Knowledge = you can easy explain it by articulating it via natural language etc. and share it with others.

European General Data Protection Regulation (EU GDPR) = Regulation EU 2016/679 – see the EUR-Lex 32016R0679, will make black-box approaches difficult to use, because they often are not able to explain why a decision has been made (see explainable AI).

Gaussian Process (GP) = collection of stochastic variables indexed by time or space so that each of them constitute a multidimensional Gaussian distribution; provides a probabilistic approach to learning in kernel machines (See: Carl Edward Rasmussen & Christopher K.I. Williams 2006. Gaussian processes for machine learning, Cambridge (MA), MIT Press); this can be used for explanations. (see also: Visual Exploration Gaussian)

Gradient = a vector providing the direction of maximum rate of change.

Ground truth = generally information provided by direct observation (i.e. empirical evidence) instead of provided by inference. For us it is the gold standard, i.e. the ideal expected result (100% true);

Inverse Probability = an older term for the probability distribution of an unobserved variable, and was described by De Morgan 1837, in reference to Laplace’s (1774) method of probability.

Implicit Knowledge = very hard to articulate, we do it but cannot explain it (also tacit knowledge).

Kernel = class of algorithms for pattern analysis e.g. support vector machine (SVM); very useful for explainable AI

Kernel trick = transforming data into another dimension that has a clear dividing margin between the classes


Post-hoc Explainability (PHE) = such models are designed for interpreting black-box models and provide local explanations for a specific decision and re-enact on request, typical examples include LIME, BETA, LRP, or Local Gradient Explanation Vectors, prediction decomposition or simply feature selection.


Saliency map = image showing in a different representation (usually easier for human perception) each pixel’s quality.

Tacit Knowledge = Knowledge gained from personal experience that is even more difficult to express than implicit knowledge.

Transfer Learning (TL) = The ability of an algorithm to recognize and apply knowledge and skills learned in previous tasks to novel tasks or new domains, which share some commonality. Central question: Given a target task, how do we identify the commonality between the task and previous tasks, and transfer the knowledge from the previous tasks to the target one? Pan, S. J. & Yang, Q. 2010. A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering, 22, (10), 1345-1359, doi:10.1109/TKDE.2009.191.
LRP on GitHub

Computer Science > Machine Learning

**iNNvestigate neural networks!**

Maximilian Alber, Sebastian Lapuschkin, Philipp Seegerer, Miriam Hägele, Kristof T. Schütt, Grégoire Montavon, Wojciech Samek, Klaus-Robert Müller, Sven Dähne, Pieter-Jan Kindermans

(Submitted on 13 Aug 2018)

In recent years, deep neural networks have revolutionized many application domains of machine learning and are key components of many critical decision or predictive processes. Therefore, it is crucial that domain specialists can understand and analyze actions and predictions, even of the most complex neural network architectures. Despite these arguments neural networks are often treated as black boxes. In an attempt to alleviate this shortfall, many analysis methods were proposed, yet the lack of reference implementations often makes a systematic comparison between the methods a major effort. The presented library iNNvestigate addresses this by providing a common interface and out-of-the-box implementation for many analysis methods, including the reference implementation for PatternNet and PatternAttribution as well as for LRP-methods. To demonstrate the versatility of iNNvestigate, we provide an analysis of image classifications for various state-of-the-art neural network architectures.

Subjects: Machine Learning (cs.LG); Machine Learning (stat.ML)

Cite as: arXiv:1808.04260 [cs.LG]
(or arXiv:1808.04260v1 [cs.LG] for this version)

Bibliographic data

Select data provider: Semantic Scholar | Prophy | [Disable Bibex(What is Bibex?)]

References (28)  Citations (20)

https://github.com/albermax/investigate

https://github.com/sebastian-lapuschkin/lrp_toolbox

https://github.com/ArrasL/LRP_for_LSTM

Also Explore:
Theorem 2 (Shapley kernel) Under Definition 1, the specific forms of $\pi_{x'}, L, \text{ and } \Omega$ that make solutions of Equation 2 consistent with Properties 1 through 3 are:

$$\Omega(g) = 0,$$

$$\pi_{x'}(z') = \frac{(M - 1)}{(M \text{ choose } |z'|)|z'|(M - |z'|)},$$

$$L(f, g, \pi_{x'}) = \sum_{z' \in Z} |f(h_x(z')) - g(z')|^2 \pi_{x'}(z'),$$

where $|z'|$ is the number of non-zero elements in $z'$.


https://github.com/OpenXAIProject/PyConKorea2019-Tutorials