Part 3: Applications of XAI

Klaus-Robert Müller, Grégoire Montavon, Wojciech Samek
Utilitarian View on XAI

Explanations should provide some additional (measurable) advantage.
Riding the Clever Hans

Uncommented: The classifier predicts correctly based on an artifact in the data (aka. ‘Clever Hans’).

[ Lapuschkin et al. 2019]
Riding the Clever Hans

[Lapuschkin et al. 2019]
Riding the Clever Hans

Extending SpRAy from [4]
- Further automating spurious cluster/class discovery by analyzing $\Phi$ with FDA
- Visualizing the spectral embedding $\Phi$, instead of affinity structure

[Anders et al. 2019]
Riding the Clever Hans

The solution of FDA can be understood as directions of maximal separability between clusterings, and, when normalized and plugged into the original objective, gives scores of separability.

[Anders et al. 2019]
Riding the Clever Hans

![Prediction: binder, True Label Rank: 3]

$\mathcal{T} = 4.77 : 1st$

![Prediction: laptop, True Label Rank: 1]

![Prediction: laptop, True Label Rank: 1]

$\mathcal{T} = 0.41 : last$

![Prediction: jigsaw_puzzle, True Label Rank: 1]

![Prediction: jigsaw_puzzle, True Label Rank: 1]

![Prediction: jigsaw_puzzle, True Label Rank: 1]

$\mathcal{T} = 0.83 : 52nd$

![Prediction: tow_truck, True Label Rank: 2]

![Prediction: garbage_truck, True Label Rank: 1]

![Prediction: garbage_truck, True Label Rank: 1]

$\mathcal{T} = 0.53 : 798th$
Riding the Clever Hans

Isolate artefact, add to other/all classes, re-train model. [Anders et al. 2019]
Riding the Clever Hans

P-ClArC Projective Class Artifact Compensation

Detect problem in CAV space --> project out (no retraining)

CAV-Predictor  CAV-Predictor

Isolate artefact, add to other/all classes, re-train model. [Anders et al. 2019]
Explanation Guided Training

Cross-domain few-shot classification task (CD-FSC)

Examples of support images:
- dog
- crate
- cuirass
- lion
- vase

Q1: pred: dog
Q2: pred: lion

[Sun et al. 2021]
Explanation Guided Training

\[ w_{l_{rp}} = 1 + R(f_p) \]
\[ f_{p-l_{rp}} = w_{l_{rp}} \odot f_p \]
\[
\mathcal{L} = \xi \mathcal{L}_{ce}(y, p) + \lambda \mathcal{L}_{ce}(y, p_{l_{rp}})
\]
### Explanation Guided Training

<table>
<thead>
<tr>
<th></th>
<th>miniImagenet</th>
<th></th>
<th>mini-CUB</th>
<th></th>
<th>mini-Cars</th>
<th></th>
<th>mini-Plantae</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1-shot</td>
<td>1-shot-T</td>
<td>5-shot</td>
<td>5-shot-T</td>
<td>1-shot</td>
<td>1-shot-T</td>
<td>5-shot</td>
<td>5-shot-T</td>
</tr>
<tr>
<td>RN</td>
<td>58.31±0.47%</td>
<td>61.52±0.58%</td>
<td>72.72±0.37%</td>
<td>73.64±0.40%</td>
<td>41.98±0.41%</td>
<td>42.52±0.48%</td>
<td>58.75±0.36%</td>
<td>59.10±0.42%</td>
</tr>
<tr>
<td>LRP-RN</td>
<td>60.06±0.47%</td>
<td>62.65±0.56%</td>
<td>73.63±0.37%</td>
<td>74.67±0.39%</td>
<td>42.44±0.41%</td>
<td>42.88±0.48%</td>
<td>59.30±0.40%</td>
<td>59.22±0.42%</td>
</tr>
<tr>
<td></td>
<td>29.32±0.34%</td>
<td>28.56±0.37%</td>
<td>38.91±0.38%</td>
<td>37.45±0.40%</td>
<td>29.65±0.33%</td>
<td>29.61±0.37%</td>
<td>39.19±0.38%</td>
<td>38.31±0.39%</td>
</tr>
<tr>
<td>RN</td>
<td>33.53±0.36%</td>
<td>33.69±0.42%</td>
<td>47.40±0.36%</td>
<td>46.51±0.40%</td>
<td>34.80±0.37%</td>
<td>34.54±0.42%</td>
<td>48.09±0.35%</td>
<td>47.67±0.39%</td>
</tr>
<tr>
<td>LRP-RN</td>
<td>34.80±0.37%</td>
<td>34.54±0.42%</td>
<td>48.09±0.35%</td>
<td>47.67±0.39%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Negative explanations imply that model relies on linguistic information rather than the image information.

[A person sitting on a bench with a skateboard. A bedroom with a bed, a chair, and a television.

[Sun et al. 2021b]
Explanation Guided Training

Negative explanations imply that the model relies on linguistic information rather than the image information.

Baseline: A blond woman in a blue shirt is riding a bike in a crowd.
LRP-IFT: A blond woman in a blue tank top is sitting on a bench in a crowd.

Baseline: A man in a jean jacket is holding a cellphone in his arms.
LRP-IFT: A young boy in a green jacket is standing in front of a library.

Baseline: Two young boys are playing with toys on a floor.
LRP-IFT: A baby in a white shirt is playing with a game.

Baseline: A group of people are standing on a beach.
LRP-IFT: A group of people are standing on a boardwalk in the beach.

Baseline: A brown dog is sitting in the grass.
LRP-IFT: A brown dog is standing in the water.

Baseline: A group of people sitting around a table with a cake.
LRP-IFT: A group of people playing a video game in a living room.

[Sun et al. 2021b]
Better Insights into Model Training

(Lapuschnik et al., 2019)
Better Insights into Model Training

model learns
1. track the ball
2. focus on paddle
3. focus on the tunnel

Unmasking Clever Hans predictors and assessing what machines really learn
Better Insights into Model Training

Relevance Distribution during Training

<table>
<thead>
<tr>
<th>NIPS architecture</th>
<th>Nature architecture</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1 ((4 \times 8 \times 8) \rightarrow (16), [4 \times 4])</td>
<td>C1 ((4 \times 8 \times 8) \rightarrow (32), [4 \times 4])</td>
</tr>
<tr>
<td>C2 ((16 \times 4 \times 4) \rightarrow (32), [2 \times 2])</td>
<td>C2 ((32 \times 4 \times 4) \rightarrow (64), [2 \times 2])</td>
</tr>
<tr>
<td>F1 ((2592) \rightarrow (256))</td>
<td>F1 ((3136) \rightarrow (512))</td>
</tr>
<tr>
<td>F2 ((256) \rightarrow (4))</td>
<td>F2 ((512) \rightarrow (4))</td>
</tr>
</tbody>
</table>

Small architecture

| C1 \((4 \times 8 \times 8) \rightarrow (32), [4 \times 4]\) |
| C2 \((32 \times 4 \times 4) \rightarrow (64), [2 \times 2]\) |
| C3 \((64 \times 3 \times 3) \rightarrow (64), [1 \times 1]\) |
| F1 \((3136) \rightarrow (4)\) | (Lapuschkin et al., 2019)
XAI-Based Model Pruning

A. Forward Propagation with given image

B. Evaluation on relevance of neurons/filters using LRP

C. Iterative pruning of the irrelevant neurons/filters and fine-tuning

Relevance conservation property

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

(Yeom et al. 2021)
XAI-Based Model Pruning

Only 10 samples per class
(domain adaptation scenario)
Conclusion

Post-hoc explanations WORK very well!
(but they have to be applied correctly)

Theoretical frameworks for XAI exist (e.g. Deep Taylor), no need to apply XAI methods in empirical ad-hoc manner.

Explanations can be used beyond visualization purposes.

XAI is applicable beyond deep learning.

New book to come soon ...
References

W Samek, G Montavon, S Lapuschkin, C Anders, KR Müller

Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications

Proceedings of the IEEE, 109(3):247-278, 2021

With the broader and highly successful usage of machine learning (ML) in industry and the sciences, there has been a growing demand for explainable artificial intelligence (XAI). Interpretability and explanation methods for gaining a better understanding of the problem-solving abilities and strategies of nonlinear ML, in particular, deep neural networks, are, therefore, receiving increased attention. In this work, we aim to: 1) provide a timely overview of this active emerging field, with a focus on “post hoc” explanations, and explain its theoretical foundations; 2) put interpretability algorithms to a test both from a theory and comparative evaluation perspective using extensive simulations; 3) outline best practice aspects, i.e., how to best include interpretation methods into the standard usage of ML; and 4) demonstrate successful usage of XAI in a representative selection of application scenarios. Finally, we discuss challenges and possible future directions of this exciting foundational field of ML.
References

Tutorial / Overview Papers


References

Methods Papers

  PLOS ONE, 10(7):e0130140, 2015 [preprint, bibtex]

  Pattern Recognition, 65:211–222, 2017 [preprint, bibtex]

  Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN), 2019 [preprint, bibtex]

- A Binder, G Montavon, S Lapuschkin, KR Müller, W Samek. Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers

  Proceedings of the International Conference on Learning Representations (ICLR), 2018

References

Explaining Beyond DNN Classifiers


References

Evaluation of Explanations

- L Arras, A Osman, W Samek. Ground Truth Evaluation of Neural Network Explanations with CLEVR-XAI

- W Samek, A Binder, G Montavon, S Bach, KR Müller. Evaluating the Visualization of What a Deep Neural Network has Learned

  Proceedings of the ACL Workshop on BlackboxNLP, 113-126, 2019 [preprint, bibtex]

- G Montavon. Gradient-Based Vs. Propagation-Based Explanations: An Axiomatic Comparison
References

Detecting Model and Dataset Artefacts


References

Software Papers


References

Application to Sciences


References

Application to Text


- L Arras, G Montavon, KR Müller, W Samek. Explaining Recurrent Neural Network Predictions in Sentiment Analysis
  Proceedings of the EMNLP Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, 159-168, 2017
  [preprint, bibtex]

- L Arras, F Horn, G Montavon, KR Müller, W Samek. Explaining Predictions of Non-Linear Classifiers in NLP
  Proceedings of the ACL Workshop on Representation Learning for NLP, 1-7, 2016 [preprint, bibtex]

- F Horn, L Arras, G Montavon, KR Müller, W Samek. Exploring text datasets by visualizing relevant words
  arXiv:1707.05261, 2017
References

Application to Images & Faces

- S Lapuschkin, A Binder, KR Müller, W Samek. *Understanding and Comparing Deep Neural Networks for Age and Gender Classification* Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW), 1629-1638, 2017 [preprint, bibtex]


References

Application to Video


Application to Speech

References

Application to Neural Network Pruning


Interpretability and Causality


Model Improvement & Training Enhancement


References

Link to the book

Organization of the book
Part I Towards AI Transparency
Part II Methods for Interpreting AI Systems
Part III Explaining the Decisions of AI Systems
Part IV Evaluating Interpretability and Explanations
Part V Applications of Explainable AI
--> 22 Chapters
Thank you for your attention

http://www.heatmapping.org

- Tutorials
- Software
- Online Demos