Introduction to Explainability in Deep Learning

Markus Hartinger
2020-01-15
Strasbourg Deep Learning Meetup

Aim of this Presentation

Explore the currently available tools and methods for explainability in deep learning and driving needs and look on upcoming developments

Explainable AI (XAI) refers to methods and techniques in the application of artificial intelligence technology (AI) such that the results of the solution can be understood by human experts. [1]

Keywords

Explainable, interpretable, transparent, understandable, intelligible, comprehensible, trusted

Deep learning, deep neural networks, machine learning (ML), artificial intelligence (AI)

Contents

- Overview: Explainable Artificial Intelligence
  - Motivation
  - Definition of terms
  - Driving needs by audience and goals
  - Approaches
  - Future Developments
- Available XIA Tools

Overview: Explainable Artificial Intelligence

Most of the content shown in this section is a summary of Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI *[2]*

Motivation

- Ensure impartiality in decision-making, i.e. to detect, and consequently, correct from **bias** in the training dataset [2]
- Robustness by highlighting potential **adversarial perturbations** that could change the prediction [2]
- Only meaningful variables infer the output, i.e., guaranteeing that an underlying truthful **causality** exists in the model reasoning. [2]

Definition of Terms (no consensus achieved yet)

- **Understandability / Intelligibility:** characteristic of a model to make a human understand its **function** (why a decision was made) without understanding its internal structure. [2]
- **Comprehensibility:** refers to the ability of a learning algorithm to represent its learned **knowledge** in a human understandable fashion [2]
- **Interpretability / Explainability**: provide the meaning in human understandable form (more general than above) [2]
- **Transparency:** The degree of a model having understandability [2]

Audience - Driving Needs [2]

Given an audience, an explainable Artificial Intelligence (XAI) is one that produces details or reasons to make its functioning clear or easy to understand. [2]

Audience	**Drivers for explainable AI**
Data scientists, developers, product owners	Ensure/improve product efficiency, research, new functionalities
Domain experts/users of the model (e.g. medical doctors, insurance agents)	Trust the model itself, gain scientific knowledge
Managers and executive board members	Assess regulatory compliance, understand corporate AI applications
Users affected by model decisions	Understand their situation, verify fair/correct decisions
Regulatory entities/agencies	Certify model compliance with the legislation in force, audits

Goals of Explainable AI (XAI)[2]

Goal	Audience according to publications
Trustworthiness	Domain experts, users of the model affected by decisions
Causality	Domain experts, managers and executive board members, regulatory entities/agencies
Transferability	Domain experts, data scientists
Informativeness	All, get support in decision making
Confidence	Domain experts, developers, managers, regulatory, entities/agencies
Fairness	Users affected by model decisions, regulatory entities/agencies
Accessibility	Product owners, managers, users affected by model decisions
Interactivity	Domain experts, users affected by model decisions
Privacy awareness	Users affected by model decisions, regulatory entities/agencies

Approaches for XAI [2]

- Two main branches: transparent models and post-hoc explainability

Transparent Models
- Algorithmic transparency
- Decomposability
- Simulatability

Post-hoc Explainability
- Text explanations
- Visualizations
- Local explanations
- Explanations by example
- Explanations by simplification
- Feature relevance

Transparent Models [2]

Algorithmic transparency
Ability of the user to understand the process followed by the model to produce any given output from its input data

Decomposability
Ability to explain each of the parts of a model (input, parameter and calculation)

Simulatability
Ability of a model of being simulated or thought about strictly by a human, hence complexity takes a dominant place in this class

Also diagram on page 11 of Explainable Artificial Intelligence (XAI)

Transparent Models - Examples [2]

- Models which are to some degree transparent:
  Linear/logistic regression, decision trees, K-nearest neighbors, rule based learners, general additive models, bayesian models
  (see also page 14 & 15 of [2])

- Models which are considered not transparent (at least in unmodified form):
  Tree ensembles, support vector machines, multi-layer neural network, convolutional neural network recurrent neural network

Post-hoc Explainability [2]

Text explanations
Generate text explanations that help explaining the results from the model

Visual explanation
Visualizing the model's behavior. Many methods come with dimensionality reduction

Local explanations
Segmenting the solution space and giving explanations to less complex subspaces

Explanations by example
Extracting representative examples that grasp the inner relationships and correlations found

Explanations by simplification
a whole new, simpler system is rebuilt based on the trained model to be explained (interesting to engineering)

Feature relevance explanation
clarify the inner functioning of a model by computing a relevance score for its variables

Also diagram on page 13 of Explainable Artificial Intelligence (XAI)

Post-hoc Explainability - Analysis Types [2]

Applicability
- Model agnostic
- Model specific

Text explanations
- Caption generation

Visual explanation
- Conditional / Dependence / Shapley plots
- Filter / Activation
- Sensitivity / Saliency

Local explanations
- Local application of transparent models

Explanations by example
- Examples extraction
- Activation clusters

Feature relevance explanation
- Influence (SHAP)
- Sensitivity
- Game theory inspired
- Saliency
- Interaction based
- Feature importance
- Activations
- Feature Extraction
- Activation propagation

Architecture modification
- Layer modification
- Model combination
- Attention networks
- Loss modification

Explanations by simplification
- Transparent models
- Probabilistic
- Local models (LIME)
- rule extraction (G-REX)

Also diagram on page 19 of Explainable Artificial Intelligence (XAI)

Post-hoc Explainability for Deep Learning [2]

Multi-layer Neural Networks

- DeepRED (rule extraction via decision trees) [4]
- DeepLift (feature relevance [5])

Recurrent Neural Networks

Not much worked on [2]

- Feature relevance
- Local explanations

Convolutional Neural Networks

- Saliency maps of input image (p.25 [2])
- Text description (p. 2 [6])
- Image reconstruction from encoded information [p. 6 [7]
- Neuron / channel / layer visualization (p.26 of [2])
- Input perturbation (LIME, p.26 [2])

Hybrid Transparent and Black-box Methods

- Knowledge base / semantic understanding with deep learning p. 28 [9]
- Probabilistic learning with deep learning, DeepProgLog [10]
- Datafusion with deep learning [2]
- Deep Nearest Neighbors DkNN [11], p28. of [2]

Also diagrams on pages 25, 26 & 28 of Explainable Artificial Intelligence (XAI)

Future Developments [2]

- Tradeoff between Interpretability and Performance: improve simultaneously
- Agree on vocabulary
- Transfer probabilistic into causal explanations
- Incorporate existing knowledge bases
- Hybrid approaches (deep learning together with transparent models)
- Go towards responsible AI (see page 46 of [2])

Overview: Explainable Artificial Intelligence

End of this section

Available XIA Tools - Cloud Service Providers

IBM: AI Explainability 360 Open Source Toolkit [12, 20], python accessible

- Rule extraction (linear)
- Examples extraction
- Text explanations
- Feature attributions
- Layer attributions

Microsoft: Model interpretability in Azure Machine Learning [15], cloud only

- Shap (tree, deep, linear, kernel)
- Simplification (mimic, lime)
- Feature attributions
- Han text explainer

Google Explainable AI [13, 14], python accessible

- Feature attributions (calculated & visualized)
- Feature influences (what-if)

Amazon AWS

No information without account

Available XIA Tools - AI Frameworks

Deep Learning Frameworks (ranked by ArXiv Articles [15])

- Tensorflow: see previous slide Google Explainable AI [13, 14] and
  - probabilistic extensions, adversarial enhancement
- Pytorch (Facebook)
  - similar to Tensorflow as it is using Tensorboard [21]
  - Includes libraries for bayesian, probabilistic models
- Keras (frontend for Tensorflow, Theano, CNTK): moved to Tensorflow 2 [22, 24]
- Caffe(2): merged with other projects
- Theano: dead [23]
- MXNET: not popular enough
- Chainer: dead [23]
- CNTK: not popular enough

Scikit-learn, python

- All the transparent models

⇒ Feasible frameworks tools for explainable AI

- Tensorboard with Tensorflow or Pytorch
- IBM AIX 360 for additional analysis
- Scikit-learn support with transparent models
- Before mentioned hybrid models not yet commercially available

Tensorboard

Introduction video Tensorboard 2017 - 20 min

Introduction videos what-if-tool: 1/3 - 3 min, 2/3 - 3 min

Github for People+AI Research (develops what-if tool)

- Saliency maps (notebook with examples)

IBM AIX 360 (stared in 2019)

General Overview (presentation, p12. overview of competition)

- Seldon alibi (github with overview)
- Oracle Skater (github with overview presentation)
- H2O (github, complete AI framework, not popular enough)
- Microsoft Interpret (github)
- Ethical ML (github)

Overview of available algorithms (github with diagram)

MNIST Example for contrastive explanation (notebook)

Bibliography (1)

1. https://en.wikipedia.org/wiki/Explainable_artificial_intelligence
2. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI
3. Explainable Machine Learning for Scientific Insights and Discoveries
4. J. R. Zilke, E. L. Menc´ıa, F. Janssen, Deepred–rule extraction from deep neural networks, in: International Conference on Discovery Science, Springer, 2016, pp. 457–473.
5. A. Shrikumar, P. Greenside, A. Shcherbina, A. Kundaje, Not just a black box: Learning important features through propagating activation differences (2016). arXiv:1605.01713
6. K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: International Conference on Machine Learning, 2015, pp. 2048–2057.
7. A. Mahendran, A. Vedaldi, Understanding deep image representations by inverting them, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 5188–5196.
8. M. T. Ribeiro, S. Singh, C. Guestrin, Model-agnostic interpretability of machine learning (2016). arXiv:1606.05386.
9. I. Donadello, Semantic image interpretation-integration of numerical data and logical knowledge for cognitive vision, Ph.D. thesis, University of Trento (2018).
10. R. Manhaeve, S. Dumancic, A. Kimmig, T. Demeester, L. De Raedt, DeepProbLog: Neural probabilistic logic programming, in: Advances in Neural Information Processing Systems 31, 2018, pp. 3749–3759.

Bibliography (2)

11. N. Papernot, P. McDaniel, Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning (2018). arXiv:1803.04765.
12. IBM: AI Explainability 360 Open Source Toolkit
13. Google: Explainable AI
14. Google: AI Explainability Whitepaper
15. Microsoft: Model interpretability in Azure Machine Learning
16. https://www.kaggle.com/discdiver/deep-learning-framework-power-scores-2018
17. Explainable Machine Learning for Scientific Insights and Discoveries
18. Explanation in Artificial Intelligence: Insights from the Social Sciences
19. https://github.com/slundberg/shap
20. IBM: One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques
21. https://pytorch.org/
22. https://keras.io/
23. https://en.wikipedia.org/wiki/Comparison_of_deep-learning_software
24. https://www.pyimagesearch.com/2019/10/21/keras-vs-tf-keras-whats-the-difference-in-tensorflow-2-0/