Analysis of Trustworthiness in Machine Learning and Deep Learning

▸ **AUTHORS:** MOHAMED KENTOUR, JOAN LU

▸ **PRESENTER:** MOHAMED KENTOUR

▸ University of Huddersfield, UK

▸ Email: mohamed.kentour@hud.ac.uk
Presenter's short Bio

- MSc Software Engineering and information Processing, University of Boumerdes, Algeria (2018)

- MSc Computing, University of Huddersfield, UK (2020)

- Currently doing a PhD in Computer Science and Informatics, University of Huddersfield, UK

Website: [KentMoh/GitHub.uk](http://KentMoh/GitHub.uk)
Research Interest

- Explainable machine learning
- Transparency within neural networks
- Deep learning for sentiment analysis
- Graph theory

Mohamed Kentour & Joan Lu

University of Huddersfield / IARIA
Plan

Introduction
- Overview
- Objectives

Background
- Interpretable vs Explainable ML models
- Case of DL
- Limits

Insights
- Model decomposition
- Bring users' perceptual metrics into the learning flow

Demonstration

Legal concerns

Conclusion and future work

Mohamed Kentour & Joan Lu
University of Huddersfield / IARIA
1. Introduction

- Data deluge and decision making [1]
  - Performance vs transparency
  - Need for transparency

- Data-science life cycle [2]
  - Trustworthiness within ML and DL life cycle

- Users' behavior changing
- Users' cognitive level

---

Mohamed Kentour & Joan Lu

University of Huddersfield / IARIA
1. Introduction

1.1 Overview
- New data sampling
- New perceptual dimension
- Adjust to the users' requirements

1.2 Objectives
- Analyse literature models
- Show the impact of the perceptual metrics on models' performance
- Increase the trustworthiness of the model through a demonstration

Mohamed Kentour & Joan Lu

Figure 2. Perceptual metric's inclusion within ML life cycle.
2. Background

2.1 Interpretable ML

- Post deployed analysis

-> Challenge
  - Bridge learning theory with quantifiable metrics

-> Solutions
  - Matrix factorization (knowledge, method)
  - Fuzzy System and ontology for decision trees
  - Invoke explainable models (LIME, COVAR) to measure quantifiable metrics

Mohamed Kentour & Joan Lu

University of Huddersfield / IARIA
2. Background

2.2. Explainable ML models

**LIME / SP-LIME [3]**
- Goes beyond a single trust of a prediction (trade-off approximation/complexity)
- Sub-modular Pick LIME to study features impact on explanations

**IBM 360° [4]**
- More flexible: separating obvious explanation from black-boxes (local/global variables)

**DARPA [5]**
- Highest accuracy and lowest complexity by mapping from high-level to low-level features which is part of learning process (backpropagation)

Figure 5. Users' reactions on explainable models.
2. Background
2.2. Case of DL

- (1) Generative modeling
- (2) Post-hoc techniques
2. Background

2.3. Limits

- Although explainable models show high performance, they fail to infer missing concepts.

- Data sparsity within perceptual metrics remains an issue.

- Users may express a changing behavior regarding any explainable model.
3. Insight

- Model decomposition (Figure 6) allows a link between explainability and a learning theory.
  - Credit assignment path.
  - Abductive learning, etc.

- To justify new features (trustworthy metrics) based on their impact on the whole performance.
  - I.e., active neurons in neural networks; feature selection, etc.
4. Demonstration

Importance of handling users' changing behavior in a recommender system.

-> Does the recommender explain or support a user changing behavior?

-> Solution:
CHR (Constraint handling rules)
:- chr_constraint actor/1, actress/1.
:- chr_constraint movie/10, recommendation/3.

- movie(_,___,___,X,___) ==> actor(X); actress(X).
- recommendation(_,_,A) ==> movie(_,_,A,___,___,___).

Model's resilience against undeclared instances

Mohamed Kentour & Joan Lu
University of Huddersfield / IARIA
5. Legal concerns

- Exposure of explainable models and data privacy.
  - How far shall we explain?
  - How far shall we contextually adapt data?
- Explanability vs adversarial attacks.
  - e.g., IBM (predicted behavior of "WayBlazer app") [6]
- Fairness and the need to introduce new regulatory metrics [7].
6. Conclusion and future research

- User-centered analysis.
- Gap: explanability/interpretability.

Research perspective

- Logical reasoning for model certainty.
  - Perceptual metrics could be formalized before being trained.
  - Perceptual metrics could be typed and attributed for model exceptions.
- AI policy [8] for an easy disparate behavior deletion.
References


Thanks for listening