Leveraging Symbolic AI for XAI Purposes

Pierre Marquis

Univ. Artois, CNRS, CRIL
Institut Universitaire de France

Séminaire IA Hybride, association Aristote, Ecole Polytechnique, Palaiseau, 18 janvier 2024

UNIVERSITÉ D'ARTOIS

institut
universitaire
de France

Overview

Overview

AI Today

- AI, and notably ML, is **now all around us in everyday life**
- Symbolic AI and ML exist since the very beginning of (the modern era of) AI in the fifties
- Deep ML (alias subsymbolic AI) was a **starting point of the AI revolution** for more than 10 years
- Made possible by the availability of massive data and specific computing devices (GPU)

AI Today

- AI, and notably ML, is **now all around us in everyday life**
- Symbolic AI and ML exist since the very beginning of (the modern era of) AI in the fifties
- Deep ML (alias subsymbolic AI) was a **starting point of the AI revolution** for more than 10 years
- Made possible by the availability of massive data and specific computing devices (GPU)

- Deep ML is not the same as ML
- Symbolic ML techniques exist for decades and symbolic ML still is an active research area

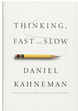
Symbolic AI Today

- Symbolic AI is de facto no longer the same as AI
- Despite the physical symbol system hypothesis by Herbert Simon and Alan Newell:

  "a physical system exhibits an intelligent behaviour if and only if it is a physical symbol system, i.e., a device which generates some time-evolving symbolic structures"

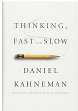
  symbolic AI is not effective enough to tackle every AI task (especially, those involving perceptions)

No One-Size-Fits-All



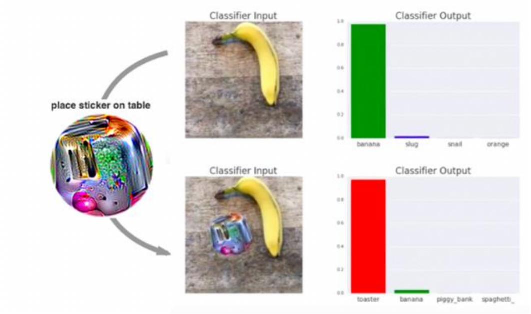
- Symbolic AI has **limitations**
  - Symbolic AI is good at reasoning, but not so good at perception tasks
  - Lack of knowledge bases (the Cyc project)
  - Scalability (complexity issues)

No One-Size-Fits-All



- Symbolic AI has **limitations**
  - Symbolic AI is good at reasoning, but not so good at perception tasks
  - Lack of knowledge bases (the Cyc project)
  - Scalability (complexity issues)
- ML has **limitations**
  - Deep ML is good at perceiving (recognizing, classifying ...) , but not so good for reasoning tasks or for generating transferable knowledge
  - Ensuring 100% correct predictions: No way!
  - Sensitivity to data (quality, quantity), garbage in, garbage out...
  - Deep models are **black boxes (opacity)**
  - Lack of common-sense

ML Models Can Be Easily Fooled



[Brown et al., NeurIPS'17]

Towards Hybrid AI

- **Taking the best of both worlds**
- Looking for synergies
  - Integrating learning and reasoning abilities to get improved AI systems (statistical relational learning, probabilistic inductive logic programming, neurosymbolic AI, concept-based NNs, etc.)
  - Reasoning to better learn
  - Learning to better reason

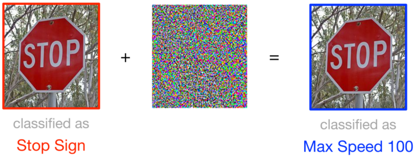
Towards Hybrid AI

- **Taking the best of both worlds**
- Looking for synergies
  - Integrating learning and reasoning abilities to get improved AI systems (statistical relational learning, probabilistic inductive logic programming, neurosymbolic AI, concept-based NNs, etc.)
  - Reasoning to better learn
  - Learning to better reason
  - **Developing trustable AI systems: trustworthy AI**

Overview

Towards Trustworthy AI

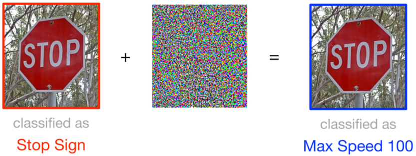
- **Trustworthy AI is mandatory for high-risk or safety-critical applications**



[Chen et al., NeurIPS'19]

Towards Trustworthy AI

- **Trustworthy AI is mandatory for high-risk or safety-critical applications**

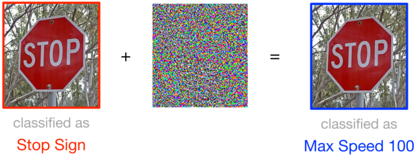


[Chen et al., NeurIPS'19]

- Trustworthy AI has a number of facets (**interpretability**, **explainability**, transparency, confidentiality, fairness, reliability, safety, etc.)

Towards Trustworthy AI

- **Trustworthy AI is mandatory for high-risk or safety-critical applications**



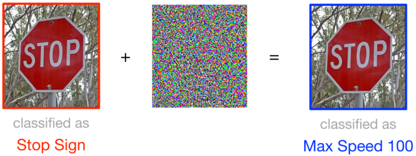
[Chen et al., NeurIPS'19]

- Trustworthy AI has a number of facets (**interpretability**, **explainability**, transparency, confidentiality, fairness, reliability, safety, etc.)

- Explaining the decisions made became a legal issue in a number of countries, especially in Europe (General Data Protection Regulation – GDPR – since May 2018, European AI Act since December 2023, etc.)

Towards Trustworthy AI

- **Trustworthy AI is mandatory for high-risk or safety-critical applications**



[Chen et al., NeurIPS'19]

- Trustworthy AI has a number of facets (**interpretability**, **explainability**, transparency, confidentiality, fairness, reliability, safety, etc.)
- Explaining the decisions made became a legal issue in a number of countries, especially in Europe (General Data Protection Regulation – GDPR – since May 2018, European AI Act since December 2023, etc.)
- Trustworthy AI has been a key topic in AI for a couple of years

eXplainable AI (XAI)

XAI is the part of Trustworthy AI focusing on the **interpretability and explainability** issues

DARPA, at the origin of the buzz word "XAI", pointed out the following purpose for XAI in 2019:

"to provide users with **explanations** that enable them to understand the system's overall strengths and weaknesses, convey an **understanding** of how it will behave in future or different situations, and perhaps permit users to **correct** the system's mistakes"

eXplainable AI (XAI)

XAI is the part of Trustworthy AI focusing on the **interpretability and explainability** issues

DARPA, at the origin of the buzz word "XAI", pointed out the following purpose for XAI in 2019:

"to provide users with **explanations** that enable them to understand the system's overall strengths and weaknesses, convey an **understanding** of how it will behave in future or different situations, and perhaps permit users to **correct** the system's mistakes"

As human beings, a truly intelligent system **should not persist in error**

XAI Issues

Two families of tasks:

- Reasoning: deriving useful information from the model (e.g., addressing explanation queries or inspection/verification queries) so that the user may decide to trust or not to trust the model or the predictions made
- Decision making: when the model or the prediction is deemed not trustworthy enough, decide what to do with them (reject the prediction, learn a new model, correct the model, etc.)

XAI Issues

Two families of tasks:

- Reasoning: deriving useful information from the model (e.g., addressing explanation queries or inspection/verification queries) so that the user may decide to trust or not to trust the model or the predictions made
- Decision making: when the model or the prediction is deemed not trustworthy enough, decide what to do with them (reject the prediction, learn a new model, correct the model, etc.)

The tasks (reasoning and decision making) can be more or less automated depending on the model under consideration

XAI Issues

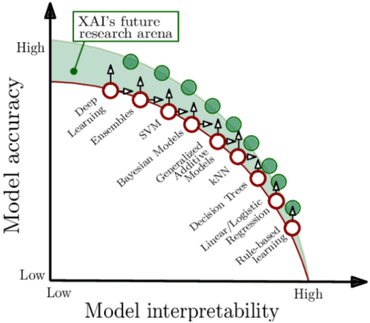
Two families of tasks:

- Reasoning: deriving useful information from the model (e.g., addressing explanation queries or inspection/verification queries) so that the user may decide to trust or not to trust the model or the predictions made
- Decision making: when the model or the prediction is deemed not trustworthy enough, decide what to do with them (reject the prediction, learn a new model, correct the model, etc.)

The tasks (reasoning and decision making) can be more or less automated depending on the model under consideration

When they can be automated (at least partly), **the connection to symbolic AI is clear**

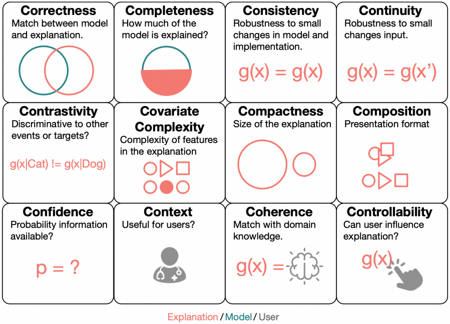
How to Evaluate the Intelligibility of an ML Model?



[Barredo Arrieta et al., Information Fusion 2020]

What is a Good Explanation?

Co-12 Properties



[Nauta et al., ACM Computing Survey 2023]

The Relevance of Data Types

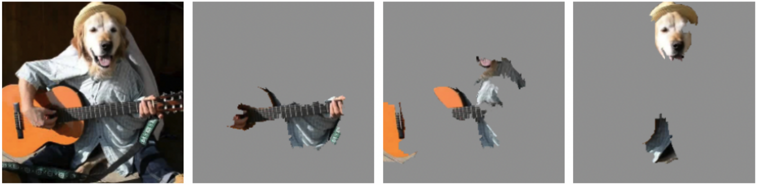
The types of the data that are processed by the ML model have **a huge impact** on the XAI techniques that can be leveraged since explanations are typically based on descriptors of the same nature as those in the data to be explained

- Subsymbolic data: for instance, pixels in a picture
- Symbolic data: for instance, tabular data, attribute/value pairs, logical formulae (pieces of knowledge as opposed to raw data)

Subsymbolic Data

- **No concepts** used in the description of the instances (in general), no intrinsic semantics
- Explanations of the predictions made are subsymbolic as well (feature attribution techniques)
- The user (aka explainee) is in charge of their interpretations

Explaining How a Picture is Classified



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

[Ribeiro et al., ACM SIGKDD'16]

- Feature importance can be displayed as saliency maps when dealing with images
- The interpretation of the explanation is achieved by the explainee
- No concepts (e.g., fretboard) are involved in the explanations!
- No formal guarantees (one cannot reason from such subsymbolic explanations)

Symbolic Data

- Instances are described using conditions, that refer to concepts
- They have a clear, formal semantics
- Formal explanations can be defined
- A two-step process
  - Representing the ML model
  - Reasoning and decision making from the representation

Overview

Formal XAI

- When dealing with high-risk applications, **correctness is paramount**

Formal XAI

- When dealing with high-risk applications, **correctness is paramount**
- Associating a circuit equivalent to the ML model in terms of inputs/outputs
- Delegating XAI queries to the circuit
- Paves the way for symbolic AI to the rescue!
  - In terms of techniques and methods used
  - In terms of approaches that are followed

The Computational Intelligibility of an ML Model

Viewing families of ML models as representations languages
[Audemard et al., KR’20]

Looking for trade-offs (reminiscent to Levesque / Brachman)
[Computational Intelligence, 1987]

- Identifying XAI queries (explanation and verification) of interest
- Such XAI queries are user-dependent
- Determining those queries that are tractable
- Choose an ML model accordingly
  (taking into account its accuracy as well)

The Computational Intelligibility of an ML Model

Viewing families of ML models as representations languages [Audemard et al., KR'20]

Looking for trade-offs (reminiscent to Levesque / Brachman) [Computational Intelligence, 1987]

- Identifying XAI queries (explanation and verification) of interest
- Such XAI queries are user-dependent
- Determining those queries that are tractable
- Choose an ML model accordingly
  (taking into account its accuracy as well)
- The case of decision trees [Audemard et al., KR'21]

Explaining Predictions

- A **hard problem**, due to the many (antagonistic) criteria to be satisfied
- Looking for trade-offs
- Human in the loop
- Several types of explanations
  - Abductive explanations
  - Contrastive explanations
- The correctness criterion
- Intractability (most of the time)
- Using heuristics and leveraging automated reasoning techniques and dedicated solvers

Computing Explanations

- Computing preferred sufficient reasons for decision trees (and preferred abductive explanations for random forests)
  [Audemard et al., AAAI'22]

- Computing abductive explanations for boosted trees
  [Audemard et al., AISTATS'23]
- Computing abductive explanations when dealing with regression problems
  [Audemard et al., IJCAI'23]
- Computing contrastive explanations for random forests
  [Audemard et al., ECAI'23]

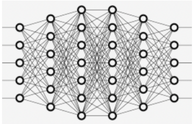
When ML Goes Wrong: The Correction Issue

- **How to change a predictor so that its predictions do not conflict with pieces of expert knowledge?**

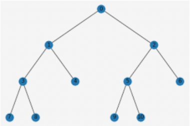
When ML Goes Wrong: The Correction Issue

- **How to change a predictor so that its predictions do not conflict with pieces of expert knowledge?**
- A KR&R issue!
- Connected to belief revision but not equivalent to it: rectification [Coste-Marquis & M., IJCAI'21]
- A principled approach to correcting an ML model
- Feasible in polynomial time for tree-based models (DT, RF, BT) when the piece of expert knowledge used takes the form of a classification rule [Coste-Marquis & M., ECAI'23]

Distilling Opaque Models

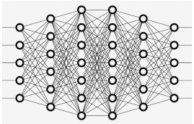


$\longrightarrow$

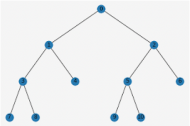


- Is it possible to do it? To which extent?

Distilling Opaque Models

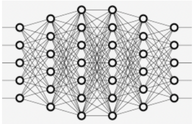


$\longrightarrow$

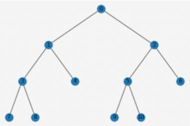


- Is it possible to do it? To which extent?
- Taking advantage of concepts coming from symbolic AI
- Succinctness of a representation language (alias spatial efficiency)

Distilling Opaque Models



$\longrightarrow$



- Is it possible to do it? To which extent?

- Taking advantage of concepts coming from symbolic AI
- Succinctness of a representation language (alias spatial efficiency)

- No polynomial-space translation from MLP to DT (or RF)
  [de Colnet & M., IJCAI'23]

For More Details on Formal XAI

See the EXPEKCTATION web page
`www.cril.univ-artois.fr/expekctation/` for additional resources
(including the open-source library PyXAI `www.cril.univ-artois.fr/pyxai/`)

- EXPEKCTATION is an acronym for "EXPlainable artificial intelligence: a KnowlEdge CompilaTion FoundATION"
- It is the name of a research and teaching chair in AI (ANR-19-CHIA-0005-01), funded by ANR, the French Agency for Research (2020-2025)
- The objective is the the development of approaches to eXplainable AI for interpretable and robust machine learning, using constraint-based automated reasoning methods, in particular knowledge compilation

Leveraging Symbolic AI for XAI Purposes

Pierre Marquis

Univ. Artois, CNRS, CRIL
Institut Universitaire de France

Séminaire IA Hybride, association Aristote, Ecole Polytechnique, Palaiseau,
18 janvier 2024