Explainable AI Framework

Xiang Wang
25 April 2019
• Background

• Our Explainable AI Framework
  • Three-Level Interpretability
  • Three Factors
  • Human-based Evaluation

• Summarization
Motivation: Deep Learning Achieves Great Success

AlphaGo beats Go human champ

Deep Net outperforms humans in image classification

Autonomous search-and-rescue drones outperform humans

DeepStack beats professional poker players

Computer out-plays humans in "doom"

IBM's Watson destroys humans in jeopardy

Deep Net beats human at recognizing traffic signs

[Visual Analytics for Explainable Machine Learning; Shixia Liu 2019]
Motivation: Deep Learning as a Black-box

Black-box nature of deep learning makes models opaque, non-intuitive, and difficult for people to understand.
A new age of AI applications
• Deep learning is the core technology.

The current generation of AI systems
• They offer tremendous benefits;
• But, the concerns on their black-box nature limit their effectiveness & application.

Motivation: Why Explainable AI (XAI)?

• Why did you do that? Why not something else?
• Why & When can I trust you?
• When do you succeed or fail?
Motivation: What we are trying to do?

Open the black box!

**Explainable Model:**
develop a range of new or modified machine learning techniques to produce more explainable models

**Explainable Interface:**
integrate human-computer interaction (HCI) with new principles, strategies, and techniques to generate effective explanations

**Psychology of Explanation:**
summarize, extend, and apply current psychological theories of explanation to develop a computational theory

- I understand why & why not
- I know when to trust you
- I know when you will succeed & fail.
**XAI aims to:**

- Maintain a high level of performance without sacrificing the performance (e.g., prediction accuracy).
- Enable human users to:
  - understand
  - appropriately trust
  - effectively manage/monitor the AI systems.

**Benefits from Next-Generation AI**

- **Fintech**
- **Healthcare**
- **Smart City**
Key Point: What is Explainability?

**Explainability (a.k.a., Interpretability):**

- “the degree to which a human can understand the cause of a decision” —— [T. Miller, et al. AI’2019]

- “the degree to which a human can consistently predict the model’s result” —— [B. Kim, et al. NIPS’2016]

No formal, technical, agreed upon definition!
Key Point: Dimensions of Interpretability (1)

Interpretability as Multi-Faceted Concept

**Intrinsic Interpretability** vs **Post-Hoc Interpretability**

- **It designs self-explanatory models.**
  - Tradeoff between Accuracy & Explainability

- **It constructs a second model to interpret the trained model.**
  - Lack guarantees about explanation quality.

© NExT++. Copyright Reserved.
Key Point: Dimensions of Interpretability (1)

Interpretability as Multi-Faceted Concept

Local Interpretability

**Decisions**: Explain why a certain pattern x has been classified in a certain way f(x).

Global Interpretability

**Model**: What would a pattern belonging to a certain category typically look like according to the model.
Our Focus: Three-Level Interpretability

Working Mechanism
Making the logic behind deep learning-based models transparent

Model Components
Exhibiting the relationships between units of deep learning-based models & semantic concepts

Individual Decisions
Indicating the salient features that influenced an individual decision

How to achieve such three-level interpretability?
How to establish **interpretable knowledge**?
• Background

• Our Explainable AI Framework
  • Three-Level Interpretability
    • Three Factors
    • Human-based Evaluation

• Summarization
Our Framework:
Learning to Explain

Interpretable Knowledge Learner

Mission-oriented Explainer

Intrinsic Interpretability
- Working Mechanism

Post-hoc Interpretability
- Model Components
- Individual Decisions

Human-based Evaluation

Statistical Criterial

Human-level Indicator

Robustness

Privacy

Fairness

Decision Trees

Knowledge Graph

Explanatory Graph
• **Extract interpretable knowledge proxy from raw data:**
  - Decision trees
  - Inference rules
  - Knowledge-aware paths
  - Symbolic reasoning ...

• **Integrate such proxy into deep model structures:**
  - E.g., transform them into vector representations + attention mechanism
  - Combine **interpretability** of shallow models & **strong representation ability** of deep models

• **We enhance automatic knowledge inference and reasoning mechanism.**
Our Preliminary Effort: Tree-enhanced Embedding Model

Component 1: Extract rules with decision trees.

Example rules:
<age: 20~30, country: US, ingredients: {beef chunk, yellow onion, bay leaves}, food: swedish beef stew >
<age: 30~40, country: CN, restaurant: China, ingredients: {lamp, garlic sprout} >

Component 2: Embed the rules to an interpretable neural attention network.
- Explain a user eating habits as the top-contributed rules.

Routine of Network Dissection

- Identify meaningful labels
  - e.g., semantic concepts, visual concepts

- Couple hidden neurons with known labels

- Organize the neuro-label pairs in the form of **interpretable structure**
  - Explanatory graph
  - Sensitivity Analysis
  - Decision trees ...

This allows us to disentangle neurons of deep models, particularly CNNs and GNNs, into a human-interpretable graphical structure.
Couple the underlying relations between users and items with the explicit relations of Knowledge Graph (KG):

- Each user-item interaction modeled via the translational function
- Each KG triplet modeled via the translational function.

[Cao, et al. "Unifying Knowledge Graph Learning and Recommendation: Towards a Better Understanding of User Preferences." WWW (2019).]
• Apply inspiring **information theoretic principles**:
  • Mutual Information
  • Paired-input nonlinear knockoff
  • Layer-wise Relevance Propagation...

• Identify contributions of each feature in a particular instance to the decision.
  • Salient maps
  • Sparse structures
  • Sparse features

• Such explainer can be applied to arbitrary models like CNNs, RNNs, and GNNs.
OUTLINE

• Background

• Our Explainable AI Framework
  • Three-Level Interpretability
  • Three Factors
  • Human-based Evaluation

• Summarization
Desiderata of an Interpretable Model

**Robustness**
Perform independently from small variations of the parameters or of the input data.

**Privacy**
Erase sensitive information about people.

**Fairness**
Guarantee the protection of groups against discrimination.

How to consider such three factors? How to establish **interpretable knowledge**?
• Characterize the robustness factor as the *interpretable knowledge* to distinguish adversarial regions:
  • Local intrinsic dimensionality
  • Adversarial perturbation
  • Structured perturbation...
Our Preliminary Work

Adversarial Ranking

• Perform adversarial training to enhance the robustness of a predictive model
  • We add adversarial perturbations on model parameters, in order to make the predictions robust against the perturbed inputs.

• Privacy
  • Making AI systems more transparent runs the risk of information leak, i.e., the explanations reveal sensitive information about users, such as geolocations, age, and gender

• Fairness
  • Subgroups with particular attributes, such as race, gender, and age, would benefit from or be hurt by the decisions.
  • Such issue is largely caused by the bias present in the data.

• We use constraints to identify the sensitive information with appropriate justification, and then remove them from the explanations.
OUTLINE

• Background

• Our Explainable AI Framework
  • Three-Level Interpretability
  • Three Factors
  • Human-based Evaluation

• Summarization
• **Keep human in the loop** is hence essential to measure the interpretability value.

<table>
<thead>
<tr>
<th>User Satisfaction</th>
<th>Task Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Clarity of the explanation (user rating)</td>
<td>• Does the explanation improve the user’s decision, task performance?</td>
</tr>
<tr>
<td>• Utility of the explanation (user rating)</td>
<td>• Artificial decision tasks introduced to diagnose the user’s understanding</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mental Model</th>
<th>Statistical Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Understanding individual decisions</td>
<td>• FDR (false discovery rate)</td>
</tr>
<tr>
<td>• Understanding the overall model</td>
<td>• PDR (positive discovery rate)</td>
</tr>
<tr>
<td>• Strength/weakness assessment</td>
<td>• Inconsistency rate...</td>
</tr>
<tr>
<td>• ‘What will it do’ prediction</td>
<td></td>
</tr>
<tr>
<td>• ‘How do I intervene’ prediction</td>
<td></td>
</tr>
</tbody>
</table>

© NExT++ Copyright Reserved.
• The study of the **cognitive process of human** can inspire the design of novel and effective **AI models**

- Investigating real human behavior
- Improving task performance
- Implementing computational models
• **Investigate** user’s reading behavior during relevance judgment

• **Implement** a Reading Inspired Model (RIM) with RL

• **Improve** the ranking performance of Web search

---

**Figure 1:** An example of user reading pattern during relevance judgment.

**Figure 5:** General schema of our proposed Reading Inspired Model (RIM).

(Li et al., Teach Machine How to Read: Reading Behavior Inspired Relevance Estimation. to appear in SIGIR 2019)
Investigate how human solve reading comprehension problems

- A two-stage process: 1) answer seeking; 2) verification among candidates
- Users pay more attention on the answer text

Implement an attention prediction model

Use predicted attention to improve the performance of the MRC task

Human-based Evaluation

Statistical Criteria

Heuristic Metric

Mission-oriented Explainer

Working Mechanism

Model Components

Individual Decisions

Key Factors

Robustness

Privacy

Fairness

Dependency of Model

By-design Interpretability

Post-hoc Interpretability

Scope of Interpretability

Global Interpretability

Local Interpretability

Interpretable Knowledge Learner

Prior Domain Knowledge

Information Theoretic Principles

Adversarial Algorithm

Our Research Interest: Learning to Explain Framework
THANK YOU

NExT++ research is supported by the National Research Foundation, Prime Minister’s Office, Singapore under its IRC@SG Funding Initiative.

xiangwang1223@gmail.com
Existing works in Different Fields:
NLP

Fine-grained explanations are in the form of:
- texts in a real-world dataset;
- Numerical scores

### Explainable NLP


<table>
<thead>
<tr>
<th>Truth</th>
<th>Model</th>
<th>Key words</th>
</tr>
</thead>
<tbody>
<tr>
<td>positive</td>
<td>positive</td>
<td>Ray Liotta and Tom Hulce shine in this stirring example of brotherly love and commitment. Hulce plays Dominic, (nickly a mentally handicapped young man who is putting his 12 minutes younger, twin brother, Liotta, who plays Eugene, through medical school. It is set in Baltimore and deals with the issues of sibling rivalry, the unbreakable bond of twins, child abuse and good always winning out over evil. It is captivating, and filled with laughter and tears. If you have not yet seen this film, please rent it. I promise, you'll be amazed at how such a wonderful film could go unnoticed.</td>
</tr>
<tr>
<td>negative</td>
<td>negative</td>
<td>Scary to go against the flow but I thought this film was incredible boring and way too long. I got tired of watching Gena Rowlands long arduous battle with herself and the cross she was experiencing. Maybe the film has some cinematic value or represented an important step for the director but for pure entertainment value I wish I would have skipped it.</td>
</tr>
<tr>
<td>negative</td>
<td>positive</td>
<td>This movie is a chilling reminder of Hollywood being just a parasite of Hollywood. Hollywood also tends to feed on past blockbusters for furthering its industry. Vidhu Vinod Chopra made this movie with the reasoning that a cocktail mix of despair and on the waterfront will bring home an Oscar. It turned out to be rookie mistake. Even the idea of the title is inspired from the Elia Kazan classic. In the original, Brandos is shown as raising doves as symbol of peace. Hollywood must move out of Hollywoods shadow if it needs to be taken seriously.</td>
</tr>
<tr>
<td>positive</td>
<td>negative</td>
<td>When a small town is threatened by a child killer, a lady police officer goes after him by pretending to be his friend. As she becomes more and more emotionally involved with the murderer her psyche begins to take a beating causing her to lose focus on the job of catching the criminal. Not a film of high volatile excitement, but solid police work and a good depiction of the faulty mind of a psychotic killer.</td>
</tr>
</tbody>
</table>

### LIME


GAN DISSECTION: VISUALIZING AND UNDERSTANDING GENERATIVE ADVERSARIAL NETWORKS. ICLR 2019

Visual Explanation
Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, Trevor Darrell: Generating Visual Explanations. ECCV (4) 2016: 3-19

Uncertainty Map
Alex Kendall, Yarin Gal: What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? NIPS 2017: 5580-5590

Saliency Map

Artifact Units