Explainable AI for Fairness and Accountability

Leilani H. Gilpin, Ph.D.
Assistant Professor of Computer Science & Engineering
UC Santa Cruz
lgilpin@ucsc.edu
lgilpin.com
Talk Agenda

Brief Intro

Motivate problem: Systems are imperfect

What is explainability?

What is actually being explained?

How to evaluate explainability?

How to explain complex systems? (autonomous driving)
About Me

• B.S in Computer Science, B.S. in Mathematics at UC San Diego

• M.S. in Computational Mathematics from Stanford University (2013), Ph.D. in EECS from MIT (2020).

• Industry experience
  • Xerox PARC
  • INRIA (France)
  • Sony AI

• Research: The methodologies and technologies for complex systems to explain themselves.
Complex Systems Fail in Complex Ways

Societal Need for Explanation

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

Business Impact

An AI-Fueled Credit Formula Might Help You Get a Loan

Startup ZestFinance says it has built a machine-learning system that's smart enough to find new borrowers and keep bias out of its credit analysis.

by Nanette Byrnes   February 14, 2017
Talk Agenda

Motivate problem: Systems are imperfect

What is explainability?

What is actually being explained?

How to evaluate explainability?

How to explain complex systems? (autonomous driving)
What is Explainability?

From Darpa XAI
“Explanations...express answer to not just any questions but to questions that present the kind of intellectual difficulty...”

Sylvain Bromberger, *On What We Know We Don’t Know*
Deep Nets are Everywhere

- Self-driving Cars
- Playing Go
- Making Medical Decisions
Deep Nets are Not Understandable

Whenever correct: “whatever you did in the middle, do more.”
Whenever wrong: “whatever you did in the middle, do less.”
Review of Research in XAI

- Definitions
- Taxonomy
  - Survey: Literature review (87 papers) in computer science, artificial intelligence, and philosophy.
  - Recommendations for Evaluation
- How can explanations help (e.g. anomaly detection).
- Contributions and Future Work
Definitions

- Explainability $\neq$ Interpretability

- **Interpretability** describes the internals of a system that is *understandable* to humans.

- **Completeness** describes operation in an *accurate* way.

- An explanation needs **both**.
What we Have

Visual cues
- Interpretable, not complete

Role of individual units
- Complete, not interpretable

Attention based
- Interpretable, not complete

Q: Is this a healthy meal?
- Textual Justification: ...because it is a hot dog with a lot of toppings.
- Visual Pointing: ...because it contains a variety of vegetables on the table.

A: Yes
A: No
Why this Matters

Interpretability

- GDPR
- Liability for decision making
Why this Matters

Completeness

• Explaining the wrong thing.

• Making decisions for the wrong reasons.

From Claudia Perlich at Women in Data Science 2018.
Motivate problem: Systems are imperfect

What is explainability?

What is actually being explained?

How to evaluate explainability?

How to explain complex systems? (autonomous driving)
What is Being Explained?

- **Visual cues**
  - Explain processing

- **Role of individual units**
  - Explain representation

- **Attention based**
  - Explanation producing

- **Questions and Answers**
  - *Q: Is this a healthy meal?*
  - *A: Yes* because it contains a variety of vegetables on the table.
  - *A: No* because it is a hot dog with a lot of toppings.
# Taxonomy

<table>
<thead>
<tr>
<th>Methods</th>
<th>Processing</th>
<th>Representation</th>
<th>Explanation producing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proxy Methods</td>
<td>Decision Trees</td>
<td>Role of layers</td>
<td>Scripted conversations</td>
</tr>
<tr>
<td></td>
<td>Salience Mapping</td>
<td>Role of neurons</td>
<td>Attention based</td>
</tr>
<tr>
<td></td>
<td>Automatic-rule extraction</td>
<td>Role of vectors</td>
<td>Disentangled representations</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Methods that Explain Processing

DeepRED –
Rule Extraction from Deep Neural Networks

Jan Ruben Zilke, Eneldo Loza Mencía, and Frederik Janssen

Technische Universität Darmstadt
Knowledge Engineering Group
j.zilke@mail.de, {eneldo,janssen}@ke.tu-darmstadt.de

Extracting Rules from Artificial Neural Networks with Distributed Representations

Sebastian Thrun
University of Bonn
Department of Computer Science III
Römerstr. 164, D-53117 Bonn, Germany
E-mail: thrun@carbon.informatik.uni-bonn.de

“Why Should I Trust You?”
Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro
University of Washington
Seattle, WA 98105, USA
marcolcr@cs.uw.edu

Sameer Singh
University of Washington
Seattle, WA 98105, USA
sameer@cs.uw.edu

Carlos Guestrin
University of Washington
Seattle, WA 98105, USA
guestrin@cs.uw.edu
Examples of Processing Methods


# Taxonomy

<table>
<thead>
<tr>
<th>Methods</th>
<th>Processing</th>
<th>Representation</th>
<th>Explanation producing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proxy Methods</td>
<td>Proxy Methods</td>
<td>Role of layers</td>
<td>Scripted conversations</td>
</tr>
<tr>
<td>Decision Trees</td>
<td>Decision Trees</td>
<td>Role of neurons</td>
<td>Attention based</td>
</tr>
<tr>
<td>Salience Mapping</td>
<td>Salience Mapping</td>
<td>Role of vectors</td>
<td>Disentangled representations</td>
</tr>
<tr>
<td>Automatic-rule extraction</td>
<td>Automatic-rule extraction</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Methods that Explain Representations

Network Dissection: Quantifying Interpretability of Deep Visual Representations
David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba
CSAIL, MIT
{davidbau, bzhou, khosla, oliva, torralba}@csail.mit.edu

Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)
Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, Rory Sayres

CNN Features off-the-shelf: an Astounding Baseline for Recognition
Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, Stefan Carlsson
CVAP, KTH (Royal Institute of Technology)
Stockholm, Sweden
{razavian, azizpour, sullivan, stefanc}@csc.kth.se
Examples of Explained Representations

Network Dissection:
Quantifying Interpretability of Deep Visual Representations

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba
CSAIL, MIT
{davidbau, bzhou, khosla, oliva, torralba}@csail.mit.edu

Interpretability Beyond Feature Attribution:
Quantitative Testing with Concept Activation Vectors (TCAV)

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler,
Fernanda Viegas, Rory Sayres


# Taxonomy

<table>
<thead>
<tr>
<th>Methods</th>
<th>Processing</th>
<th>Representation</th>
<th>Explanation producing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Proxy Methods</td>
<td>Role of layers</td>
<td>Scripted conversations</td>
</tr>
<tr>
<td></td>
<td>Decision Trees</td>
<td>Role of neurons</td>
<td>Attention based</td>
</tr>
<tr>
<td></td>
<td>Salience Mapping</td>
<td>Role of vectors</td>
<td>Disentangled representations</td>
</tr>
<tr>
<td></td>
<td>Automatic-rule extraction</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Methods that Produce Explanations

Multimodal Explanations: Justifying Decisions and Pointing to the Evidence

Dong Huk Park¹, Lisa Anne Hendricks¹, Zeynep Akata²,³, Anna Rohrbach¹,³, Bernt Schiele³, Trevor Darrell¹, and Marcus Rohrbach⁴
¹EECS, UC Berkeley, ²University of Amsterdam, ³MPI for Informatics, ⁴Facebook AI Research

Hierarchical Question-Image Co-Attention for Visual Question Answering

Jiasen Lu*, Jianwei Yang*, Dhruv Batra†, Devi Parikh†
* Virginia Tech, † Georgia Institute of Technology
{jiasenlu, jw2yang, dbatra, parikh}@vt.edu

InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets

Xi Chen†, Yan Duan†, Rein Houthooft†, John Schulman†, Ilya Sutskever†, Pieter Abbeel†
† UC Berkeley, Department of Electrical Engineering and Computer Sciences
‡ OpenAI
Examples that Produce Explanations

Multimodal Explanations: Justifying Decisions and Pointing to the Evidence

Dong Huk Park¹, Lisa Anne Hendricks¹, Zeynep Akata²,³, Anna Rohrbach¹,³,
Bernt Schiele³, Trevor Darrell¹, and Marcus Rohrbach⁴

¹EECS, UC Berkeley, ²University of Amsterdam, ³MPI for Informatics, ⁴Facebook AI Research

The best option is to veer and slow down. The vehicle is traveling too fast to suddenly stop. The vision system is inconsistent, but the lidar system has provided a reasonable and strong claim to avoid the object moving across the street.


A Problem: Insides Matter
The More Complex (Deeper)  
The Deeper the Mystery

AlexNet (2012)  
8 layers; acc 84.7%

VGG (2014)  
19 layers; acc 91.5%

GoogLeNet (2015)  
22 layers; acc 92.2%

ResNet (2016)  
152 layers; acc 95.6%
Talk Agenda

Motivate problem: Systems are imperfect

What is explainability?

What is actually being explained?

How to evaluate explainability?

How to explain complex systems? (autonomous driving)
What is Being Explained?

Visual cues
- Completeness to model

Role of individual units
- Completeness on other tasks

Attention based
- Human evaluation
- Textual Justification
  - ...because it is a hot dog with a lot of toppings.
- Visual Pointing
  - ...because it contains a variety of vegetables on the table.
# Taxonomy

<table>
<thead>
<tr>
<th>Methods</th>
<th>Processing</th>
<th>Representation</th>
<th>Explanation producing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Proxy Methods</td>
<td>Role of layers</td>
<td>Scripted conversations</td>
</tr>
<tr>
<td></td>
<td>Decision Trees</td>
<td>Role of neurons</td>
<td>Attention based</td>
</tr>
<tr>
<td></td>
<td>Salience Mapping</td>
<td>Role of vectors</td>
<td>Disentangled representations</td>
</tr>
<tr>
<td></td>
<td>Automatic-rule extraction</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Proxy Methods**
- **Decision Trees**
- **Salience Mapping**
- **Automatic-rule extraction**
- **Role of layers**
- **Role of neurons**
- **Role of vectors**
- **Scripted conversations**
- **Attention based**
- **Disentangled representations**
Challenges in Explainability

- Standards and metrics for explanations
- How to evaluate explanations?
- Current metrics of evaluation are “fuzzy”
- User based evaluations are not always appropriate
- Benchmarks for safety-critical and mission-critical tasks.
A Neural Network Labels Camera Data

Label

Inception Network - Google

Label
e.g. pedestrian

RADAR SENSORS
(both sides)

CAMERAS

LIDAR UNIT

MAIN COMPUTER
(housed in trunk)

RADAR SENSOR

LIDAR UNIT

RADAR SENSOR
Problem: Neural Networks are Brittle

For self-driving, and other mission-critical, safety-critical applications, these mistakes have CONSEQUENCES.

Vision: Real World Adversarial Examples

“Realistic” Adversarial examples

Vision: Real World Adversarial Examples
Anticipatory Thinking Layer for Error Detection

The traffic lights are on top of the truck. The lights are not illuminated. The lights are moving at the same rate as the truck, therefore this is not a “regular” traffic light for slowing down and stopping at.

“Realistic” Adversarial examples
Lack of Data and Challenges for AVs

- Existing Challenges
  - Targeted as optimizing a mission or trajectory and not safety.
  - Data is hand-curated.
  - Failure data is not available
    - Unethical to get it (cannot just drive into bad situations).
  - Want the data to be realistic (usually difficult in simulation).

Data from NuScenes
Approach: Content Generation
Anticipatory Thinking Layer for Error Detection

Approach: Content Generation
Anticipatory Thinking Layer for Error Detection

Contributions

Brief Intro

Motivate problem: Systems are imperfect

What is explainability?

What is actually being explained?

How to evaluate explainability?

How to critical systems? (autonomous driving)