Explainable & Interpretable Models

Wenxi Chen

wc2580@columbia.edu

Columbia University

November 15, 2017

Outline

1. Motivation
2. Felten's Explainability Problems [5]
3. Lipton's Interpretability [9]
   - Setting
   - Objectives of Interpretability Researches
   - Properties of Interpretable Models: Transparency
   - Properties of Interpretable Models: Post-hoc Interpretability
4. Conclusion
5. Further Discussions

Outline

Motivation

We have a complex ML model to perform automatic medical treatment and recommendation.

Motivation

We have a complex ML model to perform automatic medical treatment and recommendation.

- Can you explain the model to me?

Motivation

We have a complex ML model to perform automatic medical treatment and recommendation.

- Can you explain the model to me?
- Why should we trust the decisions made by this ML model?

Motivation

We have a complex ML model to perform automatic medical treatment and recommendation.

- Can you explain the model to me?
- Why should we trust the decisions made by this ML model?
- How can we use the knowledge learned by the model?

Motivation

We have a complex ML model to perform automatic medical treatment and recommendation.

- Can you explain the model to me?
- Why should we trust the decisions made by this ML model?
- How can we use the knowledge learned by the model?
- The questions may be unclear.

Motivation

We have a complex ML model to perform automatic medical treatment and recommendation.

- Can you explain the model to me?
- Why should we trust the decisions made by this ML model?
- How can we use the knowledge learned by the model?
- The questions may be unclear.
- How should be answer this questions?

Outline

Human Brain is Worse

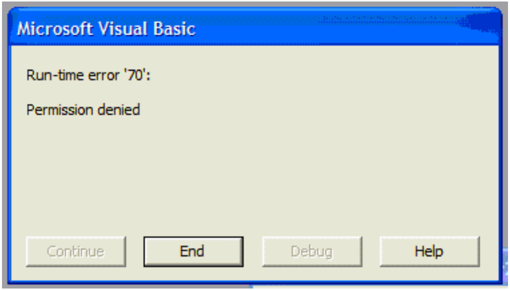
Human brain is more complex and harder to understand.

Human Brain is Worse

Human brain is more complex and harder to understand.
We say an algorithm is unexplainable, we mean 1 of the 4 problems:

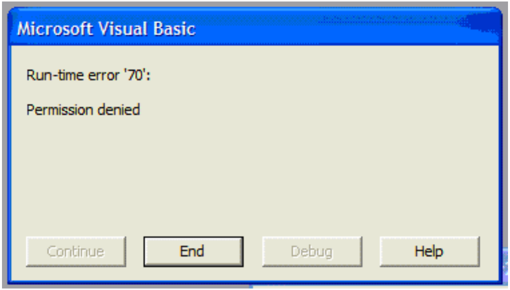
- Claim of Confidentiality
- Complexity
- Unreasonableness
- Injustice

Claim of Confidentiality



- Someone understand the algorithm, but you cannot see the code or know the details. (e.g. trade secrets)

Claim of Confidentiality



- Someone understand the algorithm, but you cannot see the code or know the details. (e.g. trade secrets)
- Nothing to do with the algorithm itself.

Complexity

- The algorithm is too complex to understand.
  (e.g. Deep Neural Networks with non-linear activation functions)

Complexity

- The algorithm is too complex to understand.
  (e.g. Deep Neural Networks with non-linear activation functions)
- Thus, we only have obstructed view of "big-picture understanding".

Complexity

- The algorithm is too complex to understand.
  (e.g. Deep Neural Networks with non-linear activation functions)
- Thus, we only have obstructed view of "big-picture understanding".
- We can still ask "what-if questions".

Unreasonableness

- We understand the algorithm and how it makes decision.

Unreasonableness

- We understand the algorithm and how it makes decision.
- But the explanation does not align with our mental model of the world. (e.g. haircut $\rightarrow$ not eat meat)

Injustice

- We understand the algorithm and how it makes decision.

Injustice

- We understand the algorithm and how it makes decision.
- But the algorithm's decision is not consistent with law or ethics.

From Felten to Lipton

Felten's four explainability problems:

- Claim of Confidentiality
- **Complexity**
- Unreasonableness
- **Injustice**

Outline

Setting: Supervised Learning

- Given labelled dataset $S = \{Z_1, \ldots, Z_n\}$
- $Z_i = (x_i, y_i) \in \mathcal{X} \times \mathcal{Y}, \forall 1 \leq i \leq n$
- Algorithms try to learn a mapping from the feature space $\mathcal{X}$ to the output space $\mathcal{Y}$
- We can compute scores to measure the performance of an algorithm A (e.g. use $err_n := \sum_{i=1}^{n} \mathbb{1}_{\{A(x_i) \neq y_i\}}$)

Objectives of Interpretability Researches

Five objectives of interpretability research:

- Trust
- Causality
- Transferability
- Informativeness
- Fair and Ethical Decision-Making

Objectives of Interpretability Researches: Trust

To build a model the users find trustworthy.

- Trust can be subjective.
- It is unclear what do people mean by saying an algorithm is trustworthy.

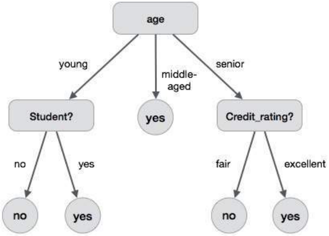
To build a model that find causal relationships.

- Association does not mean causation.
- It can be very hard to prove causality.

Objectives of Interpretability Researches: Informativeness

To extract useful information from the model.

- Sometimes the information we extract can be valuable.
- e.g. decisions in a decision tree
- e.g. sparse conditional linear regression segment



Objectives of Interpretability Researches: Fair and Ethical Decision-Making

To build a model that is fair and ethical.

Objectives of Interpretability Researches: Fair and Ethical Decision-Making

To build a model that is fair and ethical.

- This resonates with Felten's injustice problem.

Objectives of Interpretability Researches: Fair and Ethical Decision-Making

To build a model that is fair and ethical.

- This resonates with Felten's injustice problem.
- e.g. Predictive Policing: female post-secondary education

Objectives of Interpretability Researches: Fair and Ethical Decision-Making

To build a model that is fair and ethical.

- This resonates with Felten's injustice problem.
- e.g. Predictive Policing: female post-secondary education
- e.g. MIT researchers tries to learn human moral choice for driverless-vehicles [4].
  - "And if it must kill either a homeless person or a person who is not homeless, it will kill the homeless person."
  - "It makes the AI ethical or unethical in the same way that large numbers of people are ethical or unethical." - James Grimmelmann, professor at Cornell Law School

Objectives of Interpretability Researches: Transferability

To build a model that works well when domain shifts.

- e.g. use classifier trained on ImageNet on microorganism classification.

Objectives of Interpretability Researches: Transferability

To build a model that works well when domain shifts.

- e.g. use classifier trained on ImageNet on microorganism classification.
- adversarial examples that have different distribution but imperceptible to human

Objectives of Interpretability Researches: Transferability

To build a model that works well when domain shifts.

- e.g. use classifier trained on ImageNet on microorganism classification.
- adversarial examples that have different distribution but imperceptible to human

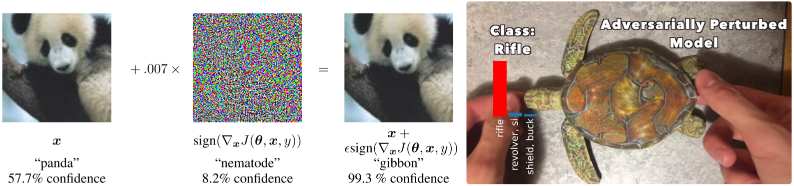


Figure: adversarial examples [6] [2]

Side Discussion: AI Safety [1] [13]

Explainability and Interpretability research is a overlap between AI capability research and AI safety research.

- Adversarial attack and defense research is part of AI Safety research.
- More focuses on Reinforcement Learning
  - Robustness to Distributional Shift (Transferability)
  - Value alignment problem is one of the hot research area in AI safety
  - Safe Exploration

Objectives and Problems

Five objectives by Lipton:

- Trust
- Causality
- Informativeness
- Fair and Ethical Decision-Making
- Transferability

Four problems by Felten:

- Claim of Confidentiality
- Complexity
- Unreasonableness
- Injustice

Properties of Interpretable Models

Transparencies:

- Simulatability
- Decomposability
- Algorithmic Transparency

Post-hoc Interpretabilities:

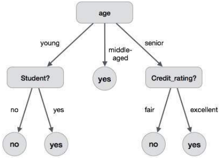
- Text Explanations
- Visualization
- Local Explanations
- Explanation by Example

Properties of Interpretable Models: Transparency

How does the model work?

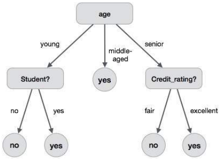
- Simulatability
- Decomposability
- Algorithmic Transparency

Transparency: Simulatability



Given an input a person can walk through the model within *reasonable* time to produce an output.

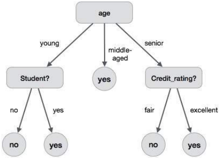
Transparency: Simulatability



Given an input a person can walk through the model within *reasonable* time to produce an output.

- This has similar meaning as Felten's "big-picture understanding".

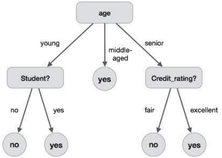
Transparency: Simulatability



Given an input a person can walk through the model within *reasonable* time to produce an output.

- This has similar meaning as Felten's "big-picture understanding".
- This is only possible when models are simple and small.

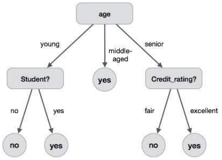
Transparency: Simulatability



Given an input a person can walk through the model within *reasonable* time to produce an output.

- This has similar meaning as Felten's "big-picture understanding".
- This is only possible when models are simple and small.
- Simple but not small: a deep decision tree

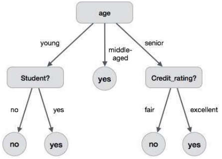
Transparency: Simulatability



Given an input a person can walk through the model within *reasonable* time to produce an output.

- This has similar meaning as Felten's "big-picture understanding".
- This is only possible when models are simple and small.
- Simple but not small: a deep decision tree
- Small but not simple: a neural network with 1 hidden layer

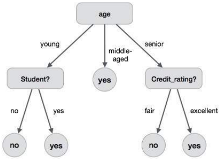
Transparency: Simulatability



Given an input a person can walk through the model within *reasonable* time to produce an output.

- This has similar meaning as Felten's "big-picture understanding".
- This is only possible when models are simple and small.
- Simple but not small: a deep decision tree
- Small but not simple: a neural network with 1 hidden layer
- Simple and small model by regularization: Lasso

Transparency: Simulatability



Given an input a person can walk through the model within *reasonable* time to produce an output.

- This has similar meaning as Felten's "big-picture understanding".
- This is only possible when models are simple and small.
- Simple but not small: a deep decision tree
- Small but not simple: a neural network with 1 hidden layer
- Simple and small model by regularization: Lasso
- Simple and small model by condition: conditional sparse linear regression

Transparency: Decomposability

Instead of the big-picture, we try to understand the small components of the model - input, parameter and calculation.

Transparency: Decomposability

Instead of the big-picture, we try to understand the small components of the model - input, parameter and calculation.

- e.g. a node in a deep decision tree

Transparency: Decomposability

Instead of the big-picture, we try to understand the small components of the model - input, parameter and calculation.

- e.g. a node in a deep decision tree
- e.g. a feature in a linear regression model

Transparency: Decomposability

Instead of the big-picture, we try to understand the small components of the model - input, parameter and calculation.

- e.g. a node in a deep decision tree
- e.g. a feature in a linear regression model
- However weights can be influenced by feature selection (e.g., collinearities in features)

Transparency: Decomposability

Instead of the big-picture, we try to understand the small components of the model - input, parameter and calculation.

- e.g. a node in a deep decision tree
- e.g. a feature in a linear regression model
- However weights can be influenced by feature selection (e.g., collinearities in features)
- Achieving meaningful decomposition by forcing the algorithm learn monotonic functions [7]

Transparency: Algorithmic Transparency

We understand the learning algorithm enough that we can theoretically bound the error rate or proof convergence.

Transparency: Algorithmic Transparency

We understand the learning algorithm enough that we can theoretically bound the error rate or proof convergence.

- What we have learned so far have algorithmic transparency.

Transparency: Algorithmic Transparency

We understand the learning algorithm enough that we can theoretically bound the error rate or proof convergence.

- What we have learned so far have algorithmic transparency.
- Neural networks and deep learning models in general fail in this regard. [page 35]

Properties of Interpretable Models: Post-hoc Interpretability

What else can the model tell us?

Properties of Interpretable Models: Post-hoc Interpretability

What else can the model tell us?
Brain is a worse black-box. This is how we interpret human brain.

Properties of Interpretable Models: Post-hoc Interpretability

What else can the model tell us?
Brain is a worse black-box. This is how we interpret human brain.
Rescue to understand deep learning.

Properties of Interpretable Models: Post-hoc Interpretability

What else can the model tell us?
Brain is a worse black-box. This is how we interpret human brain.
Rescue to understand deep learning.

- Text Explanations
  e.g. use duo models - 1 making decision and 1 explain
- Visualization
- Local Explanations
- Explanation by Example

Post-hoc Interpretability: Visualization

We try to qualitatively understand the model using visualizations.

Post-hoc Interpretability: Visualization

We try to qualitatively understand the model using visualizations.

- e.g. visualization of CNN features [11] (`https://distill.pub/2017/feature-visualization/`)

Post-hoc Interpretability: Visualization

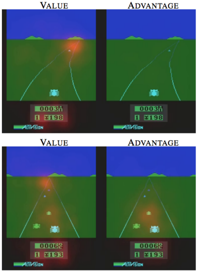
We try to qualitatively understand the model using visualizations.

- e.g. visualization of CNN features [11] (`https://distill.pub/2017/feature-visualization/`)
- e.g. finding structures in datapoints or embeddings using t-SNE [15] (`https://distill.pub/2016/misread-tsne/`)

Post-hoc Interpretability: Local Explanations

If we do not have decomposability, can we render some intuition into local dependences?

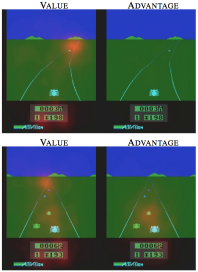
If we do not have decomposability, can we render some intuition into local dependences?



(a) Saliency map on Q-network playing Atari games [14]

Post-hoc Interpretability: Local Explanations

If we do not have decomposability, can we render some intuition into local dependences?



(a) Saliency map on Q-network playing Atari games [14]

(b) Random cell weight/excitement in CharRNN [8]

Post-hoc Interpretability: Explanation by Example

We can use examples in similar situations to explain a decision made by the model.

Table 8: *Examples of the word pair relationships, using the best word vectors from Table 4 (Skip-gram model trained on 783M words with 300 dimensionality).*

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

Figure: use examples to explain relationships in the word2vec model [10]
e.g. Paris - France + Italy = Rome

Remarks on Properties of Interpretable Models

Post-hoc Interpretability approaches can overlap

Remarks on Properties of Interpretable Models

Post-hoc Interpretability approaches can overlap

- e.g. CharRNN visualization of cell excitment
- e.g. Image captioning

Remarks on Properties of Interpretable Models

Post-hoc Interpretability approaches can overlap

- e.g. CharRNN visualization of cell excitment
- e.g. Image captioning

Felten's "what-if question": The model do not have simulatability property. How can we answer the what-if question?

Remarks on Properties of Interpretable Models

Post-hoc Interpretability approaches can overlap

- e.g. CharRNN visualization of cell excitment
- e.g. Image captioning

Felten's "what-if question": The model do not have simulatability property. How can we answer the what-if question?

1. We can use decomposability on input.

Remarks on Properties of Interpretable Models

Post-hoc Interpretability approaches can overlap

- e.g. CharRNN visualization of cell excitment
- e.g. Image captioning

Felten's "what-if question": The model do not have simulatability property. How can we answer the what-if question?

1. We can use decomposability on input.
   - However, this may not be meaningful depend on the input feature.

Remarks on Properties of Interpretable Models

Post-hoc Interpretability approaches can overlap

- e.g. CharRNN visualization of cell excitment
- e.g. Image captioning

Felten's "what-if question": The model do not have simulatability property. How can we answer the what-if question?

1. We can use decomposability on input.
   - However, this may not be meaningful depend on the input feature.
2. We can use post-hoc interpretability approaches.

Outline

Key take away

1. Linear model may not be easy to interpret.

Key take away

1. Linear model may not be easy to interpret.
2. Deep learning models can be interpretable
   - use post-hoc interpretability approaches.

Key take away

1. Linear model may not be easy to interpret.
2. Deep learning models can be interpretable
   - use post-hoc interpretability approaches.
3. Be clear about
   1) what problems you are trying to solve and
   2) what approaches you are using
   when you are talking about interpretability of a model.

Outline

Further Discussion: Algorithmic Transparency of Deep Learning Models

Two main difficulties in achieving generalization in Deep Learning Models:

1. High VC dimension of Deep Learning Models ($\geq \#$ of parameters):
   - Neural Network with ReLU: $\Omega(\frac{WL\log(W/L)}{C})$ where $W \geq CL \geq C^2$. [3]
   - No enough data to achieve meaningful error bound.
2. No theoretical proof of convergence.

References I

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané.
Concrete problems in ai safety.
arXiv preprint arXiv:1606.06565, 2016.

Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok.
Synthesizing robust adversarial examples.
CoRR, abs/1707.07397, 2017.

Maria-Florina Balcan, Travis Dick, and Ellen Vitercik.
Private and online optimization of piecewise lipschitz functions.
arXiv preprint arXiv:1711.03091, 2017.

Jon Christian.
What would the average human do?
2017.

References II

Ed Felten.
What does it mean to ask for an "explainable" algorithm?
2017.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy.
Explaining and harnessing adversarial examples.
arXiv preprint arXiv:1412.6572, 2014.

Maya Gupta, Andrew Cotter, Jan Pfeifer, Konstantin Voevodski, Kevin Canini, Alexander Mangylov, Wojciech Moczydlowski, and Alexander Van Esbroeck.
Monotonic calibrated interpolated look-up tables.
The Journal of Machine Learning Research, 17(1):3790–3836, 2016.

Andrej Karpathy.
The unreasonable effectiveness of recurrent neural networks.
2015.

References III

Zachary Chase Lipton.
The mythos of model interpretability.
CoRR, abs/1606.03490, 2016.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean.
Efficient estimation of word representations in vector space.
arXiv preprint arXiv:1301.3781, 2013.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert.
Feature visualization.
Distill, 2017.
https://distill.pub/2017/feature-visualization.

Cathy O'Neil.
Gillian tett gets it very wrong on racial profiling.
2014.

References IV

Jessica Taylor, Eliezer Yudkowsky, Patrick LaVictoire, and Andrew Critch.
Alignment for advanced machine learning systems.
Machine Intelligence Research Institute, 2016.

Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Van Hasselt, Marc Lanctot, and Nando De Freitas.
Dueling network architectures for deep reinforcement learning.
arXiv preprint arXiv:1511.06581, 2015.

Martin Wattenberg, Fernanda Viégas, and Ian Johnson.
How to use t-sne effectively.
Distill, 2016.