Interpretable Approaches to Detect Bias in Black-Box Models

Sarah Tan
Cornell University
ht395@cornell.edu

ABSTRACT

My dissertation research is grounded in the field of interpretability. I aim to develop methods to explain and interpret predictions from black-box machine learning models to help creators, as well as users, of machine learning models increase their trust and understanding of the models. In this doctoral consortium paper, I summarize my previous and current research projects in interpretability, and describe my future plans for research in this area.

CCS CONCEPTS

• **Computing methodologies** → *Model verification and validation*; *Classification and regression trees*; • **Applied computing** → *Law, social and behavioral sciences*;

KEYWORDS

Interpretability; Transparency; Black-box Models; Model Distillation; Algorithmic Fairness

ACM Reference Format:
Sarah Tan. 2018. Interpretable Approaches to Detect Bias in Black-Box Models. In *2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18), February 2–3, 2018, New Orleans, LA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3278721.3278802

1 INTRODUCTION

Black-box machine learning models permeate our lives and are increasingly being deployed for high stakes decisions, such as credit scoring [13], judicial bail decisions [2], and hospital admissions. More complicated models are being trained for the promise of an increase in accuracy, sometimes at the expense of transparency or interpretability. Yet these models are typically proprietary and opaque, and do not lend themselves to easy inspection or validation.

My dissertation research is grounded in the field of interpretability. I aim to develop methods to explain and interpret predictions from black-box machine learning models to help creators, as well as users, of machine learning models increase their trust and understanding of the models. In the field of algorithmic fairness, interpretability may be especially valuable for bias detection when specific biases are not *a priori* known (Section 4).

2 PREVIOUS RESEARCH

My past work has focused on interpreting predictions from tree-based black-box models (black-box in the sense of complexity), including random forests [5] and gradient boosted trees [9]. One line of work in interpretability [8] centers on developing models that are sparse in features or model elements. Examples include training regression models with regularization to select less features [20], or post-training pruning of the weights of a neural network to reduce model complexity. I have been exploring sparsity in *observation* methods [4]. The canonical example of this class of methods is prototype selection, where representative observations of a class are selected for presentation to a user. [11].

One output from training a random forest that has received less attention is the proximity matrix [5], a n-by-n matrix (n is the number of observations) describing the proportion of trees in the forest where a pair of observations end up in the same terminal node. This similarity metric between observations is locally adaptive in tree space [21] and reflects how the forest makes its predictions based on the observations' features. I utilized this similarity metric to develop a prototype selection method [19], presenting an alternative to other tree ensemble interpretability methods such as seeking one tree that best represents the ensemble [3] or feature importance methods [5].

3 CURRENT RESEARCH

Besides tree ensembles, I am interested in developing methods to interpret fully-connected neural networks without convolutions, an area of less research yet no less important than interpretability for convolutional neural networks (CNNs). CNNs have been applied with great success to structured data such as images [12], text [22], and speech [14]. Correspondingly there has been much interest in interpreting the outputs of CNNs, with saliency maps being a particularly well-studied method (see [15] or [1] for a review). However, data arising from critical domains such as healthcare is typically in the form of column-based features such as demographic variables, health information, etc., and if no spatial, temporal, or otherwise structured relationships are present[1], may be better modeled using fully-connected neural nets without convolutions.

3.1 Interpreting Neural Nets Using Model Distillation

Model distillation was originally introduced to distill knowledge from a large, complex model (the "teacher") to a simpler, faster model (the "student") [10]. Perhaps the first to explore the idea of model distillation for understanding were Craven and Shavit who distilled a fully-connected neural net into a decision tree [7]. I am interested in whether modern neural networks that are deeper, have more complex architectures, and trained using modern techniques, including dropout, batch normalization, weight decay, etc. can still be distilled into model classes typically considered as transparent,

[1]Long Short-Term Memory networks, a type of recurrent neural network, have been compared to CNNs on longitudinal healthcare data. See [16] for a summary.

AIES '18, February 2–3, 2018, New Orleans, LA, USA

ACM ISBN 978-1-4503-6012-8/18/02.
https://doi.org/10.1145/3278721.3278802

such as decision trees, sparse regression models, etc. This approach is called transparent model distillation. Preliminary results (as of January 2018) suggest that shallow fully connected neural net teachers on smaller data sets and classification tasks can be distilled into student models such as gradient boosted trees and tree-based generalized additive models [6]. I am working on determining if the method works on larger data sets and regression tasks. A preprint can be found at [17].

3.2 Bias Detection Using Model Distillation

I am also applying the idea of transparent model distillation to black-box risk scoring models, and I will be presenting the paper "Detecting Bias in Black-Box Models Using Transparent Model Distillation" as an oral in the main track of the AI, Ethics, and Society conference [18] in January 2018. To summarize the approach, black-box risk scoring model is treated as the teacher and distilled into a transparent student model in which each feature and its relationship to the risk score can be examined. We also train another model on the true outcome that the risk score is supposed to predict (i.e. default on a loan, for a credit score) which we use to compare against the student model of black-box risk score, to increase confidence that the student model is an accurate representation of the teacher model.

The risk score and the true outcome are closely related, since the true outcome is exactly what the black-box risk scoring model was meant to predict in the first place. Hence, feature regions for which the two models differ are of special interest, indicating that the black-box model is possibly missing information to model these feature regions accurately. On the COMPAS risk score, the approach finds significant differences between these two models for younger (age 18 and 19) and older (age above 70) age groups, as well as gender. On the Chicago Police Department's (CPD) "Strategic Subject" risk score[2], the approach picks up the eight features that CPD claims were used to construct the risk score, and none of the other features the CPD claims were not used. As part of the approach, I also proposed a statistical test to detect if data sets are missing key features used to train the black-box risk scoring model, and found that the ProPublica data is likely missing key features used in COMPAS [18].

4 FUTURE PLANS

The project on detecting bias using transparent model distillation has piqued my interest in exploring interpretability for bias detection. One compelling reason to investigate the use of transparent and interpretable models for bias detection is that specific biases need not be *a priori* known. Instead, a transparent model that reveals its inner workings could suggest areas of potential bias that did not previously come to mind but warrant more investigation.

For example, in my "Detecting Bias..." paper, the transparent model distillation approach suggested that COMPAS predicted recidivism risk for younger and older age groups (feature regions that we had not suspected of bias) to be significantly different than that for true recidivism outcomes. This then allowed us to go back to the data and attempt to generate possible explanations for this discrepancy that we could then further investigate. When deploying this approach initially on the UCI German credit data[3], after training a transparent student model on the true outcome, we found our error bars for the effect for native Germans much larger than that for foreign nationals. A quick examination of the data revealed that the data comprises mostly foreign nationals, with only a handful of German nationals, suggesting that this data is drawn from a very specific population that likely is not representative of the population one wishes to study when investigating possible bias in issuing loans.

Hence, interpretable methods for bias detection could be particularly valuable when there are likely many sources of biases – as is likely in modern data sets, with their size and complexity – that may be *a priori* not known. This motivates my dissertation research: to develop methods to explain and interpret predictions from black-box machine learning models.

REFERENCES

[1] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. 2018. Towards better understanding of gradient-based attribution methods for Deep Neural Networks. In *ICLR*.

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing Accessed May 26, 2017.

[3] Mousumi Banerjee, Ying Ding, and Anne-Michelle Noone. 2012. Identifying representative trees from ensembles. *Statistics in Medicine* (2012).

[4] Jacob Bien and Robert Tibshirani. 2011. Prototype selection for interpretable classification. *The Annals of Applied Statistics* (2011).

[5] Leo Breiman. 2001. Random forests. *Machine Learning* (2001).

[6] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In *KDD*.

[7] Mark W. Craven and Jude W. Shavlik. 1995. Extracting Tree-structured Representations of Trained Networks. In *NIPS*.

[8] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv preprint arXiv:1702.08608* (2017).

[9] Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of Statistics* (2001).

[10] Geoff Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *NIPS Deep Learning Workshop* (2015).

[11] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. 2016. Examples are not enough, learn to criticize! criticism for interpretability. In *NIPS*.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS*.

[13] Francisco Louzada, Anderson Ara, and Guilherme B Fernandes. 2016. Classification methods applied to credit scoring: Systematic review and overall comparison. *Surveys in Operations Research and Management Science* (2016).

[14] Abdel-rahman Mohamed, George E Dahl, and Geoffrey Hinton. 2012. Acoustic modeling using deep belief networks. *IEEE Transactions on Audio, Speech, and Language Processing* (2012).

[15] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2017. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing* (2017).

[16] Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. 2017. Clinical Intervention Prediction and Understanding with Deep Neural Networks. In *MLHC*.

[17] Sarah Tan, Rich Caruana, Giles Hooker, and Albert Gordo. 2018. Transparent Model Distillation. *arXiv preprint arXiv:1801.08640* (2018).

[18] Sarah Tan, Rich Caruana, Giles Hooker, and Yin. Lou. 2018. Detecting Bias in Black-Box Models Using Transparent Model Distillation. In *AIES*.

[19] Sarah Tan, Giles Hooker, and Martin T. Wells. 2016. Tree Space Prototypes: Another Look At Making Tree Ensembles Interpretable. *NIPS Interpretability Workshop* (2016).

[20] Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* (1996).

[21] Stefan Wager and Susan Athey. 2017. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *J. Amer. Statist. Assoc.* (2017).

[22] Xin Zhang and Yann LeCun. 2015. Text Understanding from Scratch. In *NIPS*.

[2] https://data.cityofchicago.org/Public-Safety/Strategic-Subject-List/4aki-r3np

[3] https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)