Georgia Tech Research Corporation

Technologies Available for LICENSING

OFFICE OF TECHNOLOGY LICENSING

https://licensing.research.gatech.edu | techlicensing@gtrc.gatech.edu

Generative Causal Explanations for Black-Box Classifiers

For improved transparency and trust in machine learning systems and results

This novel machine learning technique uses a generative framework to learn a rich and flexible vocabulary to explain a black-box classifier and applies this vocabulary to construct explanations using principles of causal modeling. Developed by Georgia Tech, it enables the complete capture of complex causal relationships while ensuring that resulting explanations respect the data distribution. This use of causal modeling allows learning of explanatory factors that have a causal rather than correlational relationship with the classifier.

The learning framework has two fundamental components to operationalize these causal explanations: (1) a method to represent and move within the data distribution and (2) an information-theoretic metric for causal influence of different data aspects on the classifier output. The learning procedure finds a low-dimensional set of latent factors that represent the data, partitioning the representation into a set of “noncausal” factors that are irrelevant to the classifier and a set of “causal” factors that affect the classifier’s output. Because each point in latent space maps to an in-distribution data sample, the model naturally ensures that perturbations result in valid data points.

Using this framework, a user can understand the aspects of the data that are important to the classifier at large by visualizing the effect of changing each causal factor in data space. They can also determine the aspects that dictated the classifier output for a specific input by observing an input’s corresponding latent (hidden) values. These latent factors can describe much more complex patterns and relationships in the data than explanation methods that rely on single features or masks of features in input space.

Summary Bullets

- **Increased trust**: Has the potential to improve the transparency and fairness of machine learning systems and increase the level of trust users place in their decisions
- **Flexible explanation vocabulary**: Provides a rich and flexible vocabulary for explanation that is more expressive than feature selection or saliency map–based methods
- **Deeper insights**: Offers visualizations that can be much more descriptive than saliency maps, particularly in vision applications, leading to better user understanding

Solution Advantages

- **Increased trust**: Has the potential to improve the transparency and fairness of machine learning systems and increase the level of trust users place in their decisions
- **Flexible explanation vocabulary**: Provides a rich and flexible vocabulary for explanation that is more expressive than feature selection or saliency map–based methods
- **Deeper insights**: Offers visualizations that can be much more descriptive than saliency maps, particularly in vision applications, leading to better user understanding
- **Higher levels of interaction**: Allows a user to explore causal factors by sweeping latent factors rather than inspecting a static mask or saliency map
- **Causal interpretation**: Leverages notions of causality rather than relying on ad hoc heuristics

Potential Commercial Applications

This technology is applicable where trust, bias, or fairness of a classifier is being evaluated by a practitioner who seeks interpretable insights from the classifier decisions, such as:

- Automated medical diagnoses
- Loan application approval
- Job candidate selection
- Facial recognition systems

Background and More Information

The widespread use of machine learning is placing more emphasis on transparent algorithmic decision making. While complex black boxes may have reliable results, their internal reasoning is not always apparent to end-users trying to grasp the reasoning behind forecasts. There is a growing consensus among researchers, ethicists, and the public that machine learning models deployed in sensitive applications should be able to explain their decisions. A powerful way to make these explanations mathematically precise is using the language of causality. Explanations should identify causal relationships between certain data aspects—features which may or may not be semantically meaningful—and the classifier output.

Constructing causal explanations requires reasoning about how changing different aspects of the input data affects the classifier output. However, these observed changes are only meaningful if the modified combination of aspects occurs naturally in the dataset. For example, it is not helpful to tell a loan applicant that their loan would have been approved if they had made a negative number of late payments, and a doctor can’t prescribe a treatment if their automated diagnosis system depends on a biologically implausible attribute.

A central challenge in constructing causal explanations is, therefore, the ability to change certain aspects of data samples without leaving the data distribution. Georgia Tech’s novel generative-model-based framework overcomes this challenge.

Inventors

- Matthew O'Shaughnessy
  PhD Student - Georgia Tech School of Electrical and Computer Engineering
- Gregory Canal
  PhD Candidate - Georgia Tech School of Electrical Engineering
- Marissa Connor
  Research Assistant - Georgia Institute of Technology

- Dr. Mark Davenport
  Associate Professor -Georgia Tech School of Electrical & Computer Engineering
- Dr. Christopher Rozell
  Professor - Georgia Tech School of Electrical and Computer Engineering

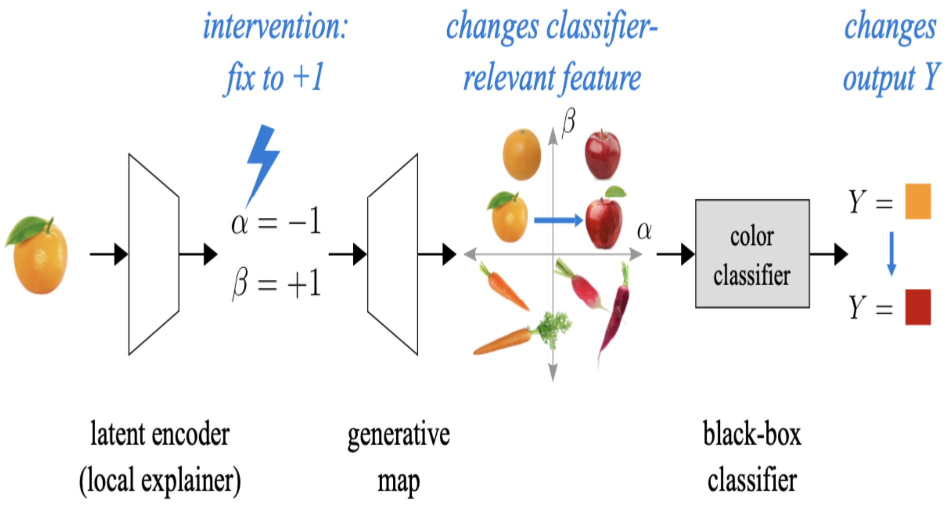
IP Status

<p>An application has been filed</p>: PCT/ US2021/03884

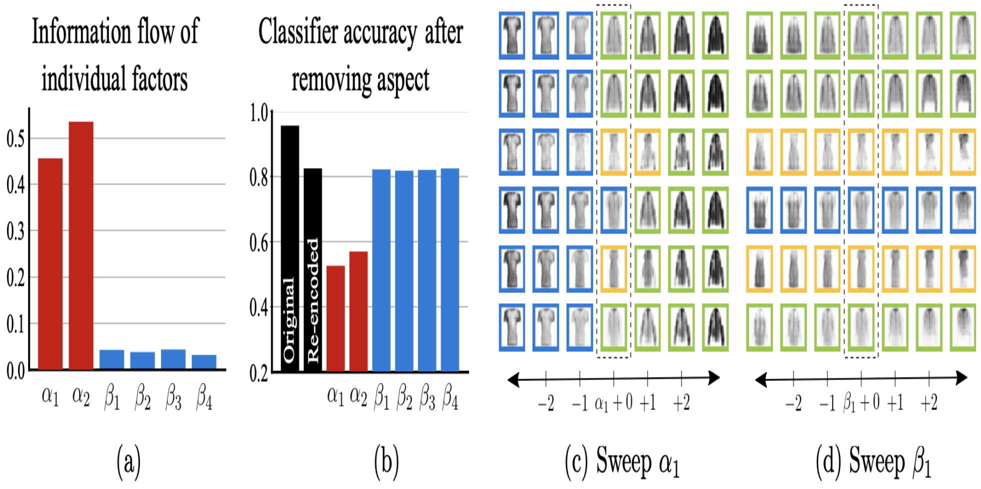
Publications

Generative causal explanations of black-box classifiers, Proc. 2020 Conference on Neural Information Processing Systems (NeurIPS) - June 11, 2020

Images



Computational architecture used to learn explanations, showing the low-dimensional representation (?, ?) learning to describe the color and shape of inputs; changing ? (color) changes the output of the classifier, which detects the color of the data sample, while changing ? (shape) does not affect the classifier output



(a)

(b)

(c) Sweep α_1

(d) Sweep β_1

(a) Information flow (causal influence) of each latent factor on the classifier output statistics; (b) Classifier accuracy when data aspects controlled by individual latent factors are removed, showing that learned causal factors—but not noncausal factors—control data aspects relevant to the classifier; (c-d) Modifying ?1 changes the classifier output, while modifying ?1 does not

Visit the Technology here:
Generative Causal Explanations for Black-Box Classifiers

https://s3.sandbox.research.gatech.edu//print/pdf/node/3285