zestfinance

Explaining Machine Learning Models

Armen Donigian
Director of Data Science Engineering

Roadmap

+ Definition of Interpretability
+ The Need for Interpretability
+ Role of Interpretability in Data Science Process
+ Relevant Application Domains
+ Barriers to Adoption
+ Conveying Interpretations
+ Research Directions

Working Definition of Interpretability

zest finance

"The ability to explain or to present in understandable terms to a human."

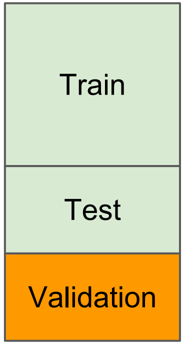
[Paper titled "Towards A Rigorous Science of Interpretable Machine Learning"]

The Need for Interpretability

In Supervised ML, we learn a model to accomplish a specific goal by minimizing a loss function.

Purpose is to ***trust*** & ***understand*** how the model uses inputs to make predictions.

Validation loss is ***Not Enough***! Can’t encode needs below into single loss function...



Bias: Non-stationarity

Fairness: Overlook gender-biased word embeddings (or other protected classes)

(Refer paper titled: “Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings”)

Safety: Infeasible to test all failure scenarios

Regulatory compliance: Adverse Action & Disparate Impact

Mismatched Objectives

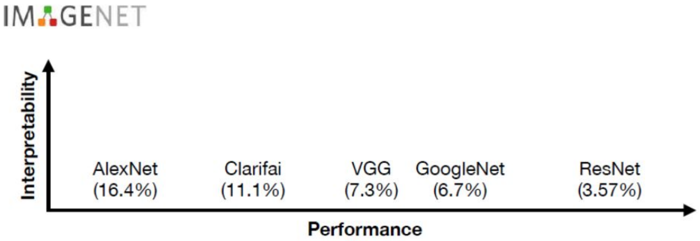
- **Single-Objective:** Overly associates wolves with snow
- **Multi-Objective trade-off:** Privacy vs Prediction Quality

Security: Is model vulnerable to an adversarial user?

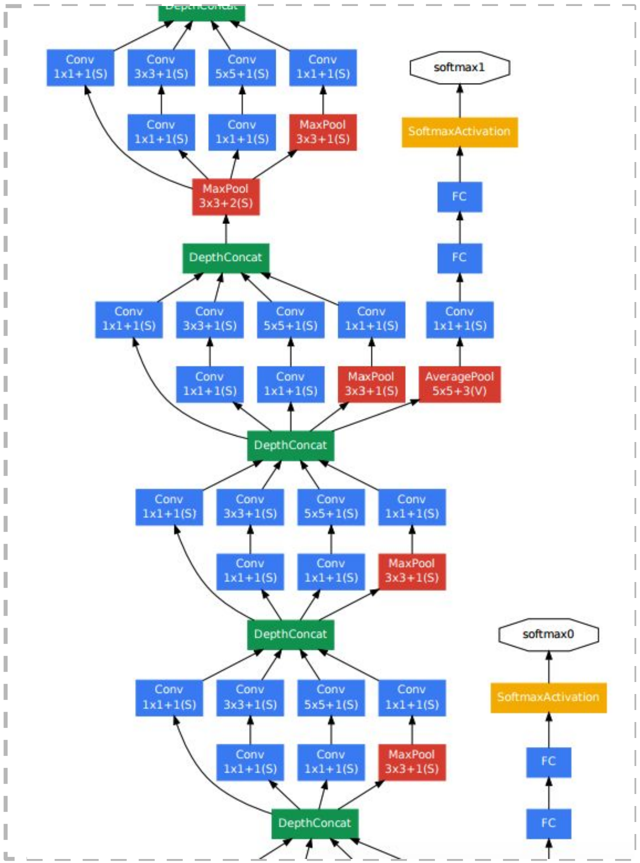
(a) Husky classified as wolf (b) Explanation

Interpretability: The Need to Keep Up

As our methods to learn patterns from data become more complex...

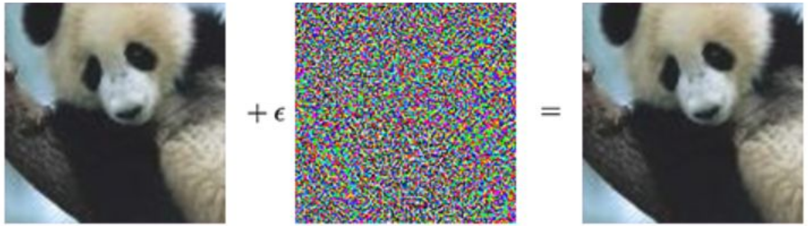


GoogLeNet



Failure Modes: Adversarial examples

(more complex model can have less intuitive failure modes)



"panda"
57.7% confidence

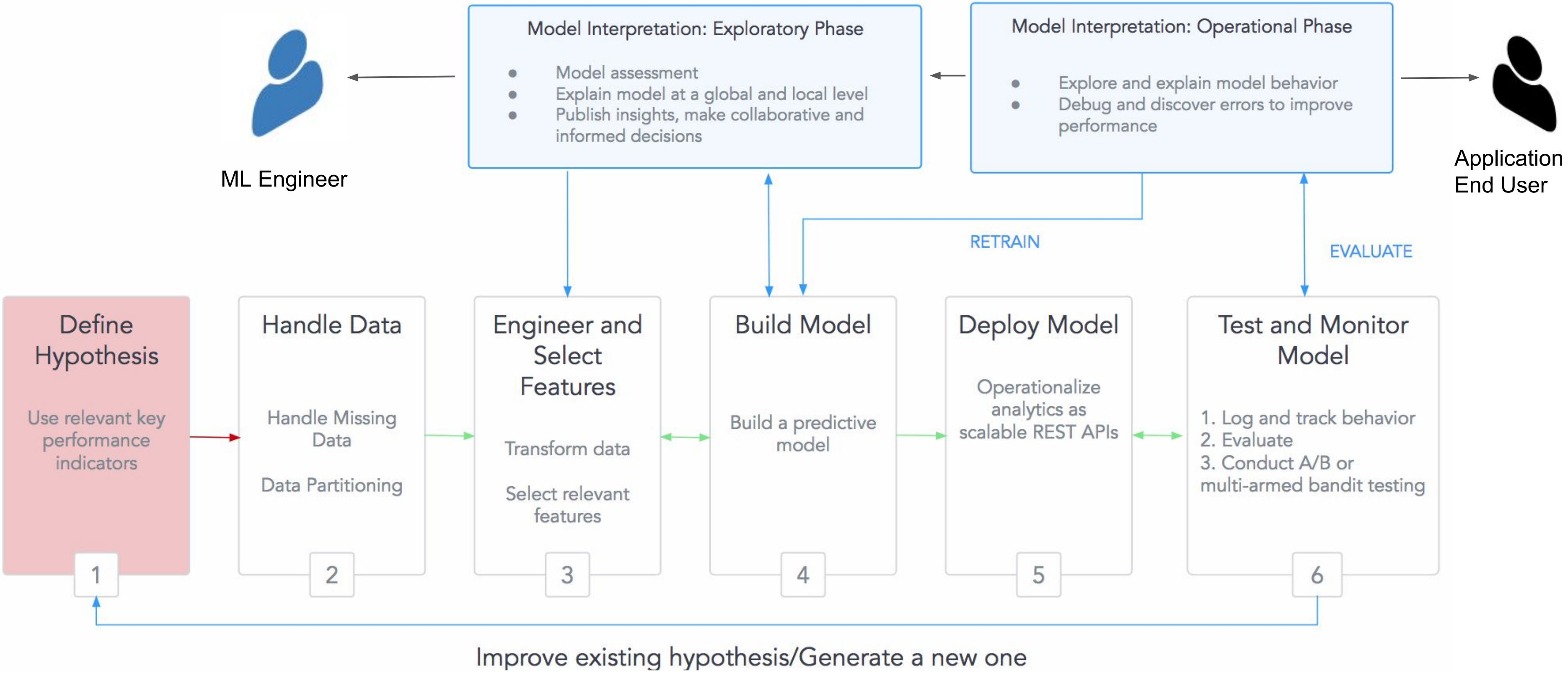
Small carefully constructed noise

"gibbon"
99.3% confidence

Figure 1

[Paper titled "Explaining and Harnessing Adversarial Examples"](#)

Role of Interpretability: Data Science Process



[Reference]

Figure 1

Application Domains for Interpretability

Credit UW (Equal Credit Opportunity Act)

- Adverse Action
- Disparate Impact

Neural machine translation

- Bridge translation gap between source & target languages
- Large corpus, unwanted co-occurrences of words which bias the model

Medical diagnoses

- Show physician regions where lesions appear in retina

Autonomous driving

- Saliency map of what model used to predict orientation & direction of steering

Scientific discoveries

- Show how molecules interact w/ enzymes, potential to learn causal relationships

Think of the cost of an incorrect prediction!

Figure 1

Figure 2

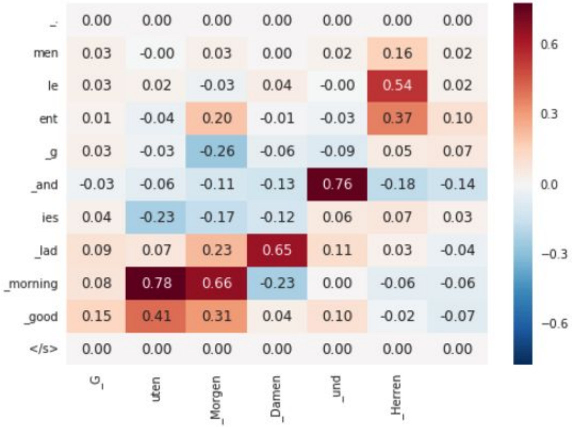


Figure 3

Barrier to Adoption in Underwriting

The Explainable Machine Learning Challenge

FICO teams up with Google, UC Berkeley, Oxford, Imperial, MIT and UC Irvine to sponsor a contest to generate new research in the area of algorithmic explainability

- Home Equity Line of Credit (HELOC) dataset
- Lines of credit $5,000 to $150,000

The black box nature of machine learning algorithms means that they are currently neither interpretable nor explainable… Without explanations, these algorithms cannot meet regulatory requirements, and thus cannot be adopted by financial institutions.

- FICO blog

Catalogue Methods by Output

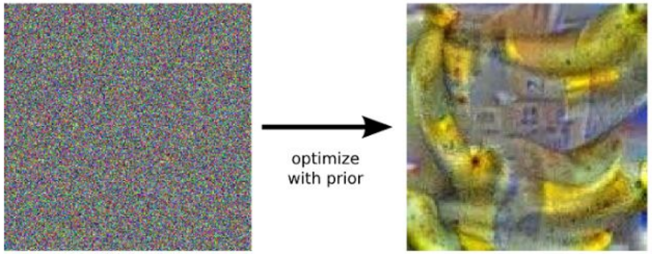
Visualizations (Intuitive)

Partial Dependence Plots, Correlations, Dim Reduction, Clustering

Text

For image captioning, we can use stochastic neighborhood embedding using n-dims to find relative neighborhoods

DeepDream



Asked to find bananas, DeepDream finds bananas in noise

Figure 1

Examples

Find most influential training samples by unweighing different samples & observe sensitivity

t-SNE

Figure 2

Influence Functions

Figure 3

Ways to Convey Interpretability (Feat Level)

Naturally Interpretable Models

$$f(x) = a_1 x_1 + a_2 x_2 + b$$

Sensitivity Analysis: "What makes the shark less/more a shark?"

- Measure sensitivity of output to changes made in the input features
- Randomly shuffle feature values one column at a time and measure change on performance
- Saliency map of what model was looking for when it made decision
  - Which pixels lead to increase/decrease of prediction score when changed?

Approach: Permutation Impact

Decomposition: "What makes the shark a shark?"

- Breaks down relevance of each feature to the prediction as a whole
- Done with respect to some reference
- Feature attributions must add up to the whole prediction (normalizing factor)

Approach: Backprop

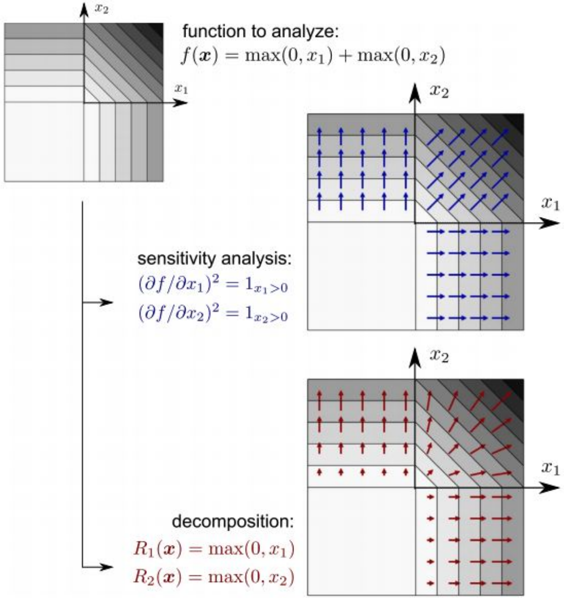


Figure 1

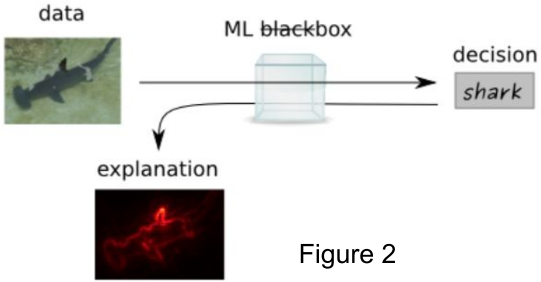


Figure 2

Naturally Interpretable Models

Linear Models

$$f(x) = a_1x_1 + a_2x_2 + b$$

$$\boxed{contrib(x_i) = a_ix_i}$$

$$f(x) - f(x_0) = \Sigma_i contrib(x_i)$$

baseline:

$x_0 = (0,0)$

Decomposition: assigns blame to causes (some reference cause)

Sensitivity: Take gradient of this model w/ respect to input, coefficients remain.

Decision Trees

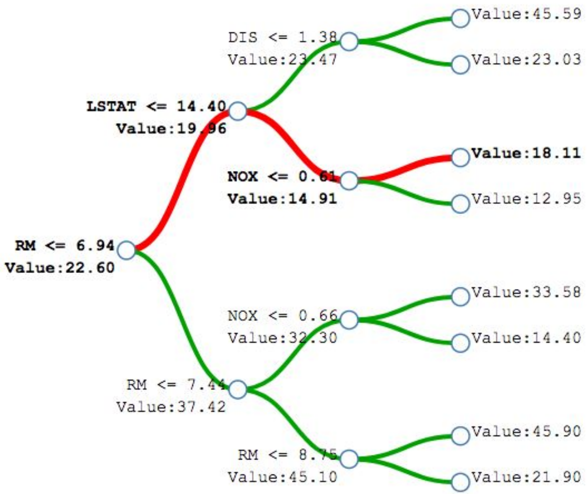


Figure 1

Boston housing prices dataset

Trace path of each decision & observe how it changes the regression value.

- Feature importances. How often a feature is used to make a decision? Check out SHapley Additive exPlanations, treeinterpreter

Permutation Feature Importance

Permutation feature importance

Randomly shuffle feature values one column at a time and measure change on performance

Pros

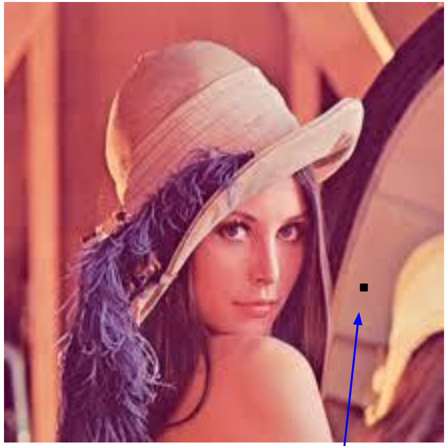
Simple implementation

Model agnostic

Cons

No variable interaction

Computationally expensive



Works when few features are important & operate independently

Single pixel perturbation does not change prediction

Surrogate Models (LIME)

Local Interpretable Model-Agnostic Explanations

Learn a simple interpretable model about the test point using proximity weighted samples

Figure 1

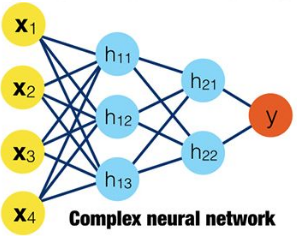
Top 3 predicted classes

(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 2

BAD	CUSTOMER_DTI	LOAN_PURPOSE	CHANNEL
0	0.18	MORT	7
1	0.42	HELOC	10
0	0.11	MORT	10
0	0.21	MORT	1

1. Train a complex machine learning model



Complex neural network

Pros

Model-agnostic

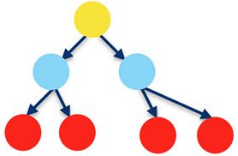
Cons

Computationally Expensive

Figure 3

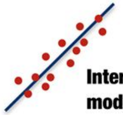
BAD	PREDICTED_BAD	CUSTOMER_DTI	LOAN_PURPOSE	CHANNEL
0	0.47	0.18	MORT	7
1	0.82	0.42	HELOC	10
0	0.18	0.11	MORT	10
0	0.12	0.21	MORT	1

2. Train an interpretable model on the original inputs and the predicted target values of the complex model



Interpretable decision tree

Or



Interpretable linear model

Backpropagation Based Approaches

Gradients (saliency map)

- Start w/ particular output
- Assign importance scores to neurons in layer below depending on function connecting those 2 layers
- Repeat process until you reach input
- With a single backward prop, you get importance scores for all features in the input

$$S_c(I) \approx w^T I + b,$$

$$w = \left. \frac{\partial S_c}{\partial I} \right|_{I_0}$$

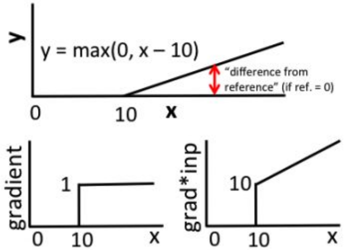
Gradient w/ respect to inputs gives us feature attributions

Pros

Simple and efficient GPU-optimized implementation

Cons

Fails in flat regions (e.g. ReLU)...gives 0 when contribution isn't zero



Backprop Approaches

Improving gradients

Dealing with absence of signal

Towards decomposition

Define a set of axioms:

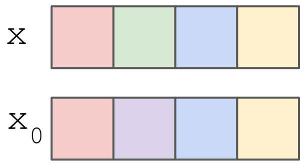
Sensitivity

Implementation invariance

Completeness/additivity

Linearity

if $f(x) \neq f(x_0)$
attribution$(x_2) > 0$



If 2 feature vectors differ only on a single feature but have different predictions then the differing feature attributions should be non-zero attribution.

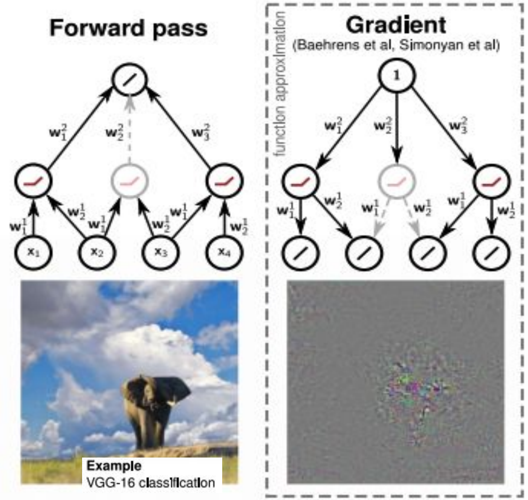
$$\frac{\partial f}{\partial g} = \frac{\partial f}{\partial h} \cdot \frac{\partial h}{\partial g}$$

$$f(x) - f(x_0) = \Sigma_i \text{attr}(x_i)$$

Backprop Approaches

Better way to backprop thru RELUs

- **DeconvNet**
  Equivalent to gradients, but ReLU in backwards direction
- **Guided Backprop**
  Gradients, but ReLU in both directions
- **PatternNet/Attribution**
  Correct gradient for correlated distracting noise
- **Layerwise Relevance Propagation**
  Equivalent to input-scaled gradients



Some other interesting approaches...

- **Integrated Gradients**
  Path integral of gradients from baseline
- **DeepLIFT**
  True decomposition relative to baseline with discrete jump
- **Deep Taylor Decomposition**
  Taylor approximation about a baseline for each neuron

Integrated Gradients

- Pick starting value, scale up linearly from reference to actual value, compute gradients along the way
- Positive & negative contribution scores

DeepLift

- Compare activation of each neuron to its reference activation
- Assign contribution scores based on difference
- Positive & negative contribution scores
- Generalizes to all activations
- Importance is propagated even when gradient is 0

Evaluating Interpretability Methods

If we have a set of feature contributions...

Spearman’s Rank-Order Correlation

What % of Top-K intersect

Experimental Evaluation Approaches

Assign a user (domain expert) tasks based on the produced feature attributions

- Show saliency maps, ask user to choose which classifier generalizes better
- Show attributions & ask user to perform feature selection to improve the model
- Ask user to identify classifier failure modes

Adversarial Examples

Interpretability can suffer from adversarial attacks independently of prediction

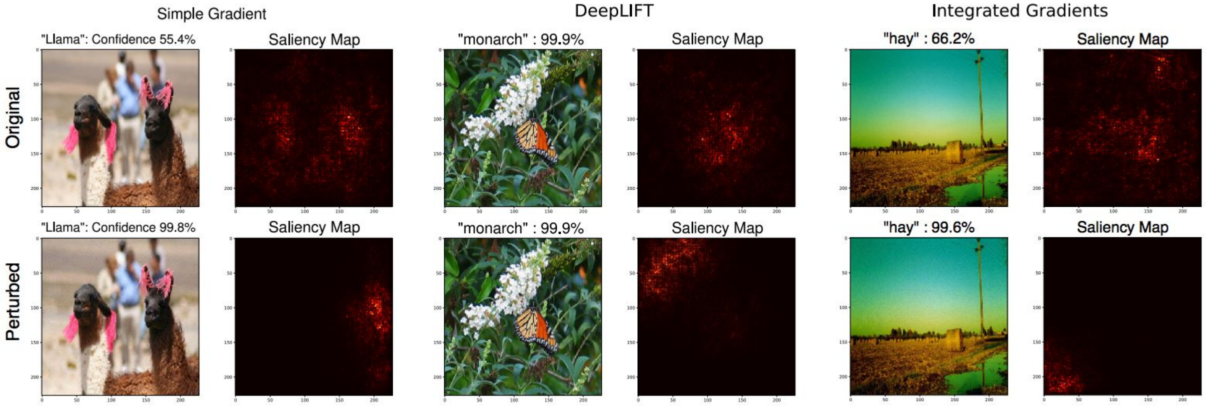


Figure 1 [Paper titled "Interpretation of Neural Networks is Fragile"]

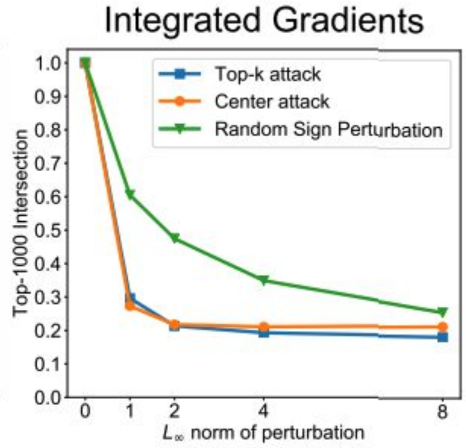


Figure 2

Attack types

Top-k attack

Take top 5 features, create distortion which drops their rank

Center attack

Take center of mass, try to move it as far as it can with some constrained distortion, goal to move the center of mass of the saliency map

Research Directions

Better loss functions for interpretability

Understand what makes certain models more ***interpretable*** and how interpretability ***fails***

Explain models in unsupervised learning, sequence learning (RNNs), and reinforcement learning

- E.g. generating text explanations of the actions of a reinforcement learning agent

Develop interpretability techniques into tools for model diagnostics, security, and compliance

Meet ZestFinance

zest finance

Mission: **Make fair and transparent credit available to everyone**

Founded in 2009 **by Douglas Merrill, former CIO and VP, Engineering, Google**

Located in Los Angeles, CA

100+ Employees primarily comprised of Data Scientists, Engineering and Business Analysts from top US institutions

Investors

matrix PARTNERS

upfront VENTURES

FLYBRIDGE CAPITAL PARTNERS

LIGHTSPEED VENTURE PARTNERS

What others are saying

"Synchrony is looking at making adjustments to its underwriting approaches... It is testing technology from vendors including ZestFinance",

THE WALL STREET JOURNAL. WSJ

"we worked with ZestFinance to harness the capability of machine learning to analyze more data and to analyze our data differently"- Ford Credit CRO, Joy Falotico

Ford

Machine-learning is also good at automating financial decisions,…Zest Finance has been in the business of automated credit-scoring since its founding in 2009."

The Economist

ZestFinance is one of the five most promising financial artificial intelligence companies in the world.

CBINSIGHTS

Backup Slides

Two Related Concepts

Transparency - understand the inner workings of a model

Attention - model learns regions of input to focus on

(a) Great crested flycatcher

(b) Yellow-breast chat