CXPlain: Causal Explanations for Model Interpretation under Uncertainty

Patrick Schwab and Walter Karlen

Presented by Chuan Lu and Chengyue Huang

Apr 27, 2020

Introduction and Related Work

Methodology

Experiments

Conclusion

Introduction and Related Work

Methodology

Experiments

Conclusion

Introduction

- Explanation methods for machine learning models are important, but complex models are difficult to interpret
- Wide variety of machine learning models make it challenging to develop a fast, accurate, unified and optimized approach to importance attribution for any machine learning models
- State-of-the-art explanation methods typically have significant uncertainty, difficult to judge if the explanation results are accurate

Drawbacks of Existing Methods

Computational cost and Applicability

Table 1: Comparison of CXPlain to several representative methods for feature importance estimation.

	CXPlain	SG [8] / IG [10]	DeepSHAP [1, 6]	LIME [16]	SHAP [6]
Accuracy	high	moderate	high	high	high
Model-agnostic	✓	×	×	✓	✓
Uncertainty estimates	✓	×	×	×	×
Computation time	fast	fast	fast	slow	slow

Drawbacks of Existing Methods

Uncertainty in explanation

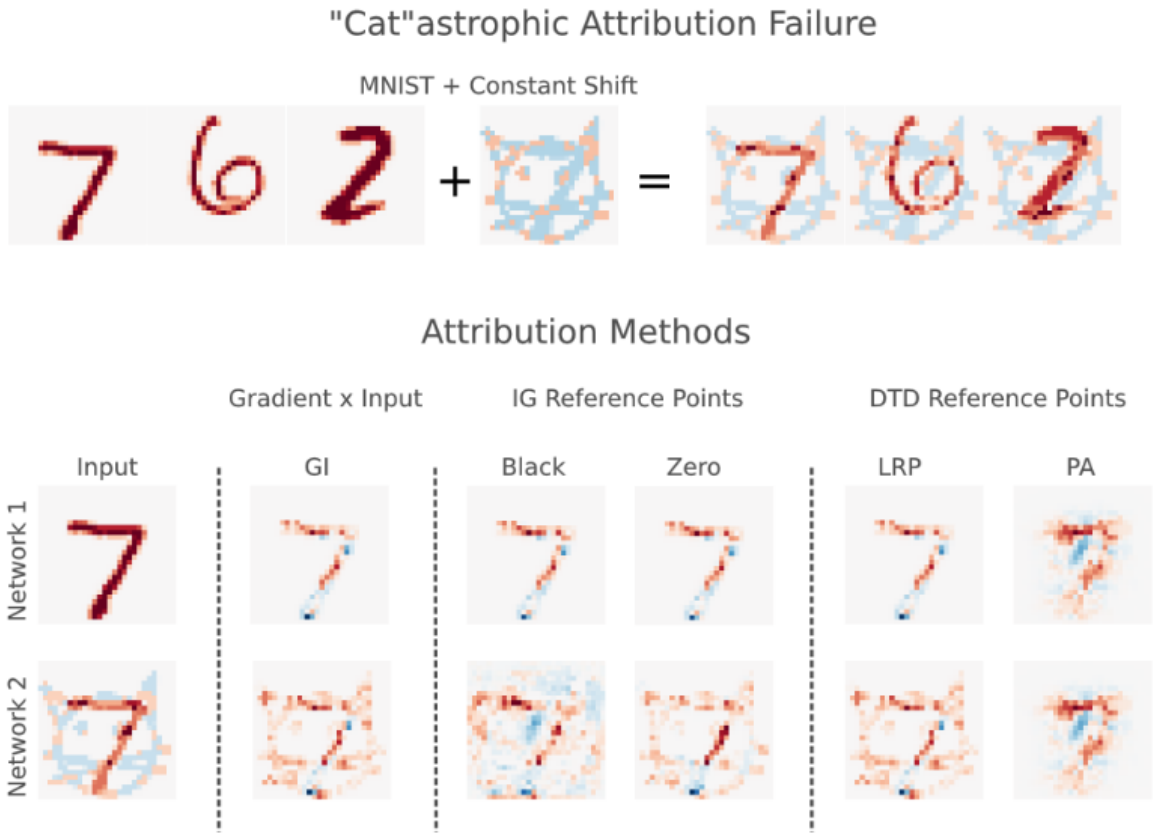


Figure 6: Evaluation of attribution method sensitivity using MNIST. Gradient x Input, IG with both a black and zero reference point and DTD with a LRP reference point, do not satisfy input invariance and produce different attribution for each network. DTD with a PA reference point is not sensitive to the transformation of the input.

1

[1] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un) reliability of saliency methods. arXiv preprint arXiv:1711.00867, 2017.

Drawbacks of Existing Methods

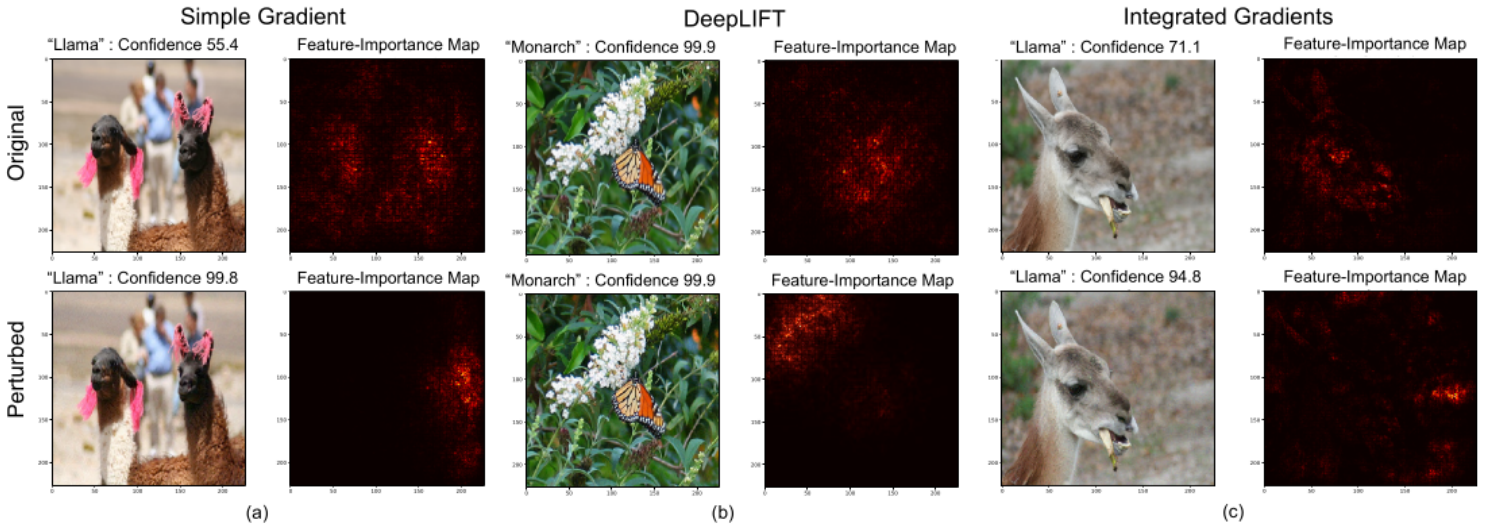


Figure 1: **Adversarial attack against feature-importance maps.** We generate feature-importance scores, also called saliency maps, using three popular interpretation methods: (a) simple gradients, (b) DeepLIFT, and (c) integrated gradients. The **top row** shows the the original images and their saliency maps and the **bottom row** shows the perturbed images (using the center attack with $\epsilon = 8$, as described in Section 2) and corresponding saliency maps. In all three images, the predicted label does not change from the perturbation; however, the saliency maps of the perturbed images shifts dramatically to features that would not be considered salient by human perception.

2

[2] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. AAAI Conference on Artificial Intelligence, 2019.

CXPlain

- Causal Explanation Model
- Estimate feature importance for **any** machine learning model
- Significantly **more accurate** and **faster** in evaluation than existing model-agnostic methods
- Provide uncertainty estimates for feature importance by bootstrap resampling
- Uncertainty estimates are strongly correlated with the accuracy of the explanation model

Idea: train a seperate explanation model to produce the importance scores for each input features.

Questions?

Introduction and Related Work

Methodology

Experiments

Conclusion

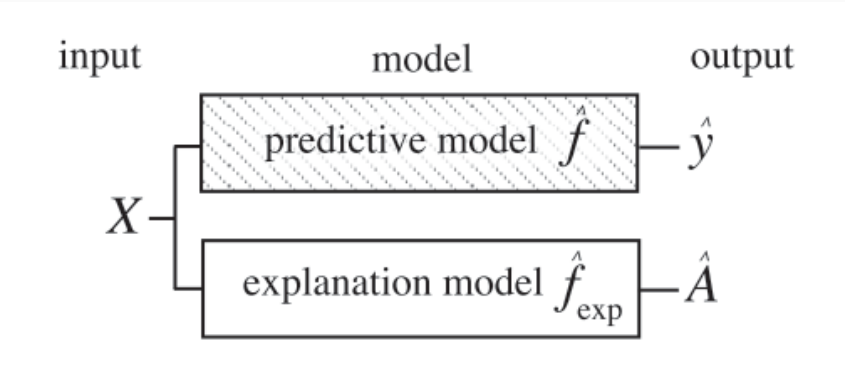
Problem Setup

Predictive model

- Suppose we have a predictive model $\hat{f} : X \rightarrow \hat{y} \in \mathbb{R}^k$, where X consists of p input features $\{x_i\}_{i=1}^{p}$
- The predictive model is scored by an objective function $\mathcal{L}$
- No other restrictions on $\hat{f}$, no need for knowledge on how $\hat{f}$ produces its output

Explanation model

- Goal: train an explanation model $\hat{f}_{exp} : X \rightarrow \hat{A} \in \mathbb{R}^p$, where each element $\hat{a}_i$ corresponds to the importance score of input feature x_i.



Causal Objective

Given input features X, we compute the outputs of the predictive model $\hat{f}$ with and without the ith input feature x_i:

$$\hat{y}_{X \setminus \{i\}} = \hat{f}(X \setminus \{i\}), \quad \hat{y}_X = \hat{f}(X) \tag{1}$$

Then we compute the errors of the predictive model $\hat{f}$ with and without x_i using the loss function $\mathcal{L}$ of $\hat{f}$:

$$\varepsilon_{X \setminus \{i\}} = \mathcal{L}(y, \hat{y}_{X \setminus \{i\}}), \quad \varepsilon_X = \mathcal{L}(y, \hat{y}_X) \tag{2}$$

The **causality** of input feature x_i to model output y is defined as the **marginal decrease** in the predictive error:

$$\Delta\varepsilon_{X,i} = \varepsilon_{X \setminus \{i\}} - \varepsilon_X \tag{3}$$

Causal Objective

We get the **importance scores** by normalizing the causalities:

$$\omega_i(X) = \frac{\Delta\varepsilon_{X,i}}{\sum_{j=1}^{p} \Delta\varepsilon_{X,j}} \tag{4}$$

Finally, we define the **causal objective** as the Kullback-Leibler divergence between the target importance distribution Ω and the one computed by the explanation model:

$$\mathcal{L}_{\text{causal}} = \frac{1}{N} \sum_{l=1}^{N} \text{KL}(\Omega_{X_l}, \hat{A}_{X_l}), \tag{5}$$

where $\Omega(i) = \omega_i(X)$, and $\hat{A}(i) = \hat{a}_i$ are the output of $\hat{f}_{exp}$.

Remarks

Mask input features

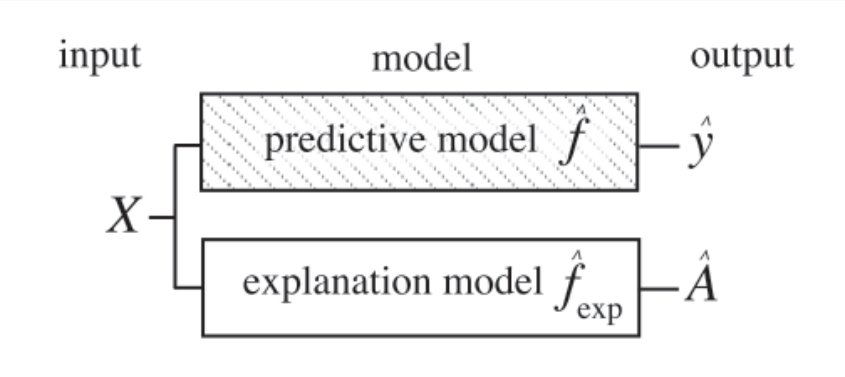
- Replace x_i with zeros when the zero value has no special meaning
- Replace x_i with mean value across the entire dataset
- Advanced schemes considering the distribution of the features can be more principled alternatives

Precompute importance scores

- To precompute the importance score for one training sample X with p features, $p + 1$ evaluations of the predictive model are required
- For high-dimensional images, we can group non-overlapping regions of adjacent pixels into feature groups

Remarks

CXPlain models



$$\hat{f}_{\text{exp}} : X = (x_1, x_2, \ldots, x_p) \rightarrow (\omega_1(X), \omega_2(X), \ldots, \omega_p(X)) \qquad (6)$$

- Any supervised machine learning model that can be trained with a custom objective can be used as explanation models
- In this work, the authors use DNNs, specifically MLPs and U-nets

Uncertainty Quantification

The uncertainty associated in feature importance estimate $\hat{a}$ can be quantified via bootstrap ensemble methods.

- Randomly draw N training samples X with repeats from the original training set (of size N), and train an explanation model on the selected set.
- Repeat the process M times to obtain a bootstrap ensemble of M explanation models.
- Use the median of the outputs as the assigned output, and use the $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ quantiles to form the confidence intervals $\mathrm{CI}_\gamma = [c_{\frac{\alpha}{2}}, c_{1-\frac{\alpha}{2}}]$ at confidence level $\gamma = 1-\alpha$.
- The width of CI_γ can be used to quantify the uncertainty of $\hat{a}$.

Questions?

Introduction and Related Work

Methodology

Experiments

Conclusion

Questions to answer

- How does the feature importance estimation performance of CXPlain compare to that of the state-of-the-art methods? (Accuracy)
- How does the computational performance of CXPlain compare to existing model-agnostic and model-specific methods for feature importance estimation? (Computational cost)
- Are uncertainty estimates of CXPlain models qualitatively and quantitatively correlated with their ability to accurately determine feature importance? (Quality of uncertainty quantification)

Tasks

Binary classification tasks:

- 8 vs. 3 MNIST benchmark (model accuracy: 99.85%)
- Gorilla vs. Zebra ImageNet benchmark (model accuracy: 96.73%)
- The test set contains 100 unseen images

Metrics

Accuracy of important features

- To quantify the accuracy of important features predicted, we measure the change in the classification models' confidence after masking the top 10 and 30% of the most importanct pixels for MNIST and ImageNet test images respectively:

$$\Delta\text{log-odds} = \text{log-odds}(p_{\text{original}}) - \text{log-odds}(p_{\text{masked}}) \tag{7}$$

- $\text{log-odds}(p) = \log\left(\frac{p}{1-p}\right)$
- Higher Δlog-odds means higher accuracy in important feature estimation
- Mann-Whitney-Wilcoxon (MWW) tests are used to calculate p-values for comparison between distributions.

Metrics

Uncertainty Quantification

- We want to test whether the uncertainty estimates u_i of input feature x_i are correlated with the errors in feature importance estimation on test set. That is to say, whether a small uncertainty provided by the ensemble models leads to a high accuracy in the estimation of importance scores.
- The authors state that typically the per-feature ground-truth values of importance scores are not available, so the Rank Error

$$\text{RE}_i = |\text{rank}_{\Delta\text{log-odds}}(i) - \text{rank}_{\hat{f}_{exp}}(i)|$$

defined as the difference in the rank of feature x_i implied by Δlog-odds and the rank implied by $\hat{f}_{exp}$ is used as the error in importance estimation.

Metrics

Uncertainty Quantification

- Pearson correlation coefficient $\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$ is used to measure the correlation between RE_i and u_i
- The coefficient ρ is computed on the top 2.5% of pixels by Δlog-odds across the test set of $N = 100$ images.
- Fisher z-transform is applied to the correlation scores to correct the skew in the distribution of sample correlation.

Results

Introduction and Related Work

Methodology

Experiments

Conclusion