Instance-wise Causal Feature Selection for Model Interpretation

Pranoy Panda
IIT Hyderabad
cs20mtech12002@iith.ac.in

Sai Srinivas Kancheti
IIT Hyderabad
cs21resch01004@iith.ac.in

Vineeth N Balasubramanian
IIT Hyderabad
vineethnb@iith.ac.in

Abstract

We formulate a causal extension to the recently introduced paradigm of instance-wise feature selection to explain black-box visual classifiers. Our method selects a subset of input features that has the greatest causal effect on the model's output. We quantify the causal influence of a subset of features by the Relative Entropy Distance measure. Under certain assumptions this is equivalent to the conditional mutual information between the selected subset and the output variable. The resulting causal selections are sparser and cover salient objects in the scene. We show the efficacy of our approach on multiple vision datasets by measuring the post-hoc accuracy and Average Causal Effect of selected features on the model's output.

1. Introduction

Explaining the predictions of black-box classifiers is important for their integration in real world applications. There have been many efforts to understand the predictions of visual systems, by generating saliency maps that quantify the importance of each pixel to the model's output. However, such methods usually require gradient information and often suffer from insensitivity to the model and the data [1]. Researchers in the recent past have taken a different perspective by proposing the task of *instance-wise feature selection* for explaining classifiers in general, and visual classifiers in particular. Here, the aim is to select a subset of pixels or superpixels to explain a black-box model's output.

L2X[4] is the first work in this space wherein it selects a fixed number of features which maximize the mutual information w.r.t. the output variable. Its successor, INVASE[12] removes the constraint of having to fix the number of features to be selected. But, both INVASE and L2X have optimization functions which try to maximize mutual information in some form or the other. For an explanation to be correct, the selected features should be causally consistent with the model being explained. Therefore, a good instance-wise feature selection method should capture the most causal features in an instance. We hypothesize that the most sparse and class discriminative features are indeed the most causal features, and they form good visual explanations. However, existing methods(L2X and INVASE) select features that may not capture causal influence, since mutual information does not always capture causal strength[2].

In this work we take a step towards unifying causality and instance-wise feature selection to select causally important features for explaining a black box model's output. First, in order to measure causal influence of input features w.r.t the output, we choose a causal metric which satisfies properties relevant for our task and subsequently we simplify this metric(under certain assumptions) to conditional mutual information. Secondly, we derive an objective function for training our explainer using continuous subset sampling. We evaluate our explainer on 3 vision datasets and compare it with with 3 popular baseline explainability methods. For the purposes of quantitative comparison, we use two metrics, post-hoc accuracy[4] and a variant of average causal effect(ACE) which we introduce in our work. Our results show performance improvements over the baselines, especially in terms of the ACE values, which verifies our claim of selecting causal features.

1.1. Problem Formulation

Our goal is to explain a black-box classifier $\mathcal{F} : \mathcal{R}^d \rightarrow \mathcal{Y}$ by learning a explainer network $E_k : \mathcal{R}^d \rightarrow S_k$ where $S_k = \{e|e \in \{0,1\}^d, |e| = k\}$. The explainer/selector network E_k chooses a subset of features(size of subset is fixed as k) for each input that best explains the predictions made by $\mathcal{F}$.

We intend to find the subset(of cardinality k) which has the maximum causal strength. Since there is no gold standard for measuring causal influence between random variables, we choose a metric which has some good properties.

Causal Model: Assuming that the underlying architecture of the black-box model is a directed acyclic graph(DAG), it can be shown that such models can be interpreted as structural causal models [3]. This SCM(Figure 1) simply has directed edges from input layer to the output layer representing the fact that the output is only a function of the inputs. We explain explicitly what X_i and Y_j mean

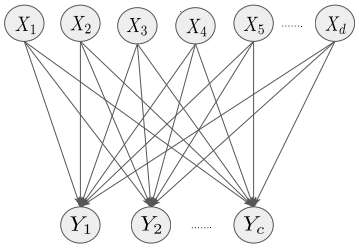

Figure 1. Black-box model as a SCM

in our context as we go along.

Causal metric: We now quantify the causal influence between two random variables in the causal graph(Figure 1) by the Relative Entropy Distance(RED)[6] metric. RED is an information theoretic causal strength measure, which is based on performing interventions on the causal graph and exposing the target of the cut/intervened edges with product of marginal distributions.

We choose RED as the metric for causal strength because it satisfies certain useful properties:

- It can capture the non-linear, complex relationship between input and output variables in a black-box model [8]. Such complex relationships are common in models such as neural networks.
- The causal strength of an edge $X_i \rightarrow Y_j$ depends on X_i and the other parents of Y_j and on the joint distribution of parents of Y_j. This locality property ensures that we do not have to take into account the causes of X_i when computing causal influence of X_i on Y_j, unless the causes are also immediate causes of Y_j[6].

As we are operating in the vision domain, instead of reasoning at the level of subset of pixels, we reason at the level of superpixels/patches in an image. These patches are disjoint i.e. there is no overlap between the patches. Now, for simplifying the RED metric for the purposes of experimenting in real-world setting, we assume local influence-pixels only depend on other pixels within a patch. That is the reason why we have no edges between X_i's in the SCM(Figure 1). Also, in the given SCM X_i refers to the i^{th} patch in the image and Y_j refers to j^{th} output node in the model. Under this setting we propose lemma 1.1, which is an extension of the findings by authors in [6].

Lemma 1.1 *The causal strength of a subset of links s going from input X to the response variable of the model Y, denoted as CS_s, is given as follows:*

$$CS_s = I(X_s; Y|X_{\overline{s}}) \tag{1}$$

Here, X_s denotes the features in set s, $X_{\overline{s}}$ denotes the features not in set s, and $X = X_s \cup X_{\overline{s}}$.

Now, we simplify the equation 1 to formulate an optimization problem.

Objective Function: The conditional mutual information between X_s and Y can be expressed as follows:

$$CS_s = I\left(X_s; Y|X_{\overline{s}}\right)$$
$$CS_s = -H(Y|X) + H(Y|X_{\overline{s}})$$

In order to solve the above objective function, we need to focus only on $H(Y|X_{\overline{s}})$ as the other term is independent of set s. Now, if we further simplify the remaining term by expanding it and viewing it as an expectation over variables Y and $X_{\overline{s}}$, we get the following equation:

$$\max_s CS_s \equiv \min_s E_{Y,X_{\overline{s}}}[\log(P(Y|X_{\overline{s}})] \tag{2}$$

$P(\ \overline{s}|X\ )$ is the explainer's output distribution, i.e. given an input X, the explainer gives its corresponding $\overline{s}$ i.e. the complement of the explanation.

2. Methodology

There are two main issues with maximizing the causal strength in equation 2. First, approximating the conditional distribution $P(Y|X_{\overline{s}})$, and second, dealing with subset sampling. We address these issues below.

Approximating $P(Y|X_{\overline{s}})$: We simply use the output of the black-box model $\mathcal{F}$ when $X_{\overline{s}}$ is given as input, to estimate $P(Y|X_{\overline{s}})$. $X_{\overline{s}}$ is represented as follows: if $s_i = 0$, $X_{\overline{s}i} = X_i$, else $X_{\overline{s}i} = 0$. [9] have used a similar approximation in their works.

Continuous subset sampling: Our objective function(2) requires sampling of subsets which is a non-differentiable operation. Similar to [4] we use the Gumbel Softmax trick [5, 7] for continuous subset sampling.

The goal of this procedure is to sample a subset s consisting of k distinct features out of the d input dimensions. Sampling set s is similar to sampling a k-hot random vector, where the length of this random vector is d. Before we perform sampling, we define a function g which maps each input feature X_i to a value which indicates its probability of being part of $\overline{s}$. In other words, $g(X)$ defines a categorical distribution from which we wish to sample from. We learn this function g via a neural network parameterized by θ. Then, we use the gumbel-softmax continuous subset sampling for sampling from this categorical distribution. The result of this sampling is a random variable Z which is a function of the neural net parameters θ and Gumbel random variables ζ. This effectively means that we can estimate $\overline{s}$ by $Z(\theta, \zeta)$.

Final optimization function: After applying continuous subset sampling, and approximation of $P(Y|X_{\overline{s}})$ we have simplified our objective to the following equation:

$$\min_{\theta} \; E_{X,Y,\zeta}[\log(\mathcal{F}(Z(\theta,\zeta)\odot X))]$$

$$= \min_{\theta} \; E_{X,\zeta}\left[\sum_{y=1}^{c} P(y|X)\log(\mathcal{F}(Z(\theta,\zeta)\odot X))\right] \quad (3)$$

$P(y|X)$ is equivalent to $\mathcal{F}(X)$. The expectation operator in the above equation does not depend on the parameter θ. So, we can learn the parameter θ by using stochastic gradient descent. We use Adam optimizer in our work.

3. Experiments and Results

In this section, we quantitatively and qualitatively compare our method with one instance-wise feature selection method L2X and, two pixel attribution methods GradCAM[10] and Saliency[11]. We use MNIST, Fashion MNIST and CIFAR as our datasets for experiments.

Our method reasons at the level of superpixels or patches of images, which means that our explainer selects patches instead of pixels. Therefore, for the baseline pixel attribution methods we consider the average attribution value for each patch, and then pick top k patches. This is done in order to fairly compare the instance-wise feature selection and existing pixel attribution methods.

Below we briefly explain the two metrics we use for quantitative evaluation:

3.1. Post-hoc accuracy[4]:

Each explainability method would return a subset of features/patches s for every instance x. Post-hoc accuracy measures how close is the predictive performance of the black-box model when it gets x_s as input w.r.t. getting the entire x as input. It is given by the following formula:

$$\frac{1}{|\mathcal{X}_{val}|}\sum_{x\in\mathcal{X}_{val}} (\arg\max(P(y|x)) == \arg\max(P(y|x_s))$$

$\mathcal{X}_{val}$ refers to the validation set.

3.2. Average Causal Effect:

First, we define individual causal effect(ICE) of a set of features s for a particular instance x as follows:

$$ICE = P(y|x_s) - P(y|x_{random})$$

Here, x_{random} represents an image in which k patches($k = |s|$) belong to x and the rest patches are null. These k patches are selected randomly from x. To compute average causal effect(ACE) we simply take the average of the ICE values over the validation set.

3.3. Results:

For the all datasets, we report the mean and standard deviation of the post-hoc accuracy and ACE values across 5 runs for L2X and our method. In the tables shown below, k denotes the number of 4×4 patches that are selected.

MNIST: This data set has 28×28 images of handwritten digits. We use a subset of the classes i.e. class 3 and 8 for experimentation. We train a simple convolutional neural net consisting of 2 layers of convolution and 1 fully connected layer, and achieve 99.74% accuracy on the test data. We parameterize the selector network of L2X and our method by a 3 layer fully convolutional net. For the L2X, we parameterize its variational approximator by the same network as that of the black-box model(for all the experiments).

Method	*k=4*	*k=6*	*k=8*
Our	**0.953± 0.006**	**0.976 ± 0.004**	**0.985± 0.005**
L2X	0.942± 0.008	0.970± 0.004	0.981± 0.003
GradCAM	0.804	0.832	0.844
Saliency	0.868	0.923	0.958

Table 1. Post-hoc accuracy(MNIST)

Method	*k=4*	*k=6*	*k=8*
Our	**0.351± 0.012**	**0.358± 0.009**	**0.353± 0.006**
L2X	0.318± 0.003	0.341± 0.012	0.343± 0.004
GradCAM	0.127	0.142	0.151
Saliency	0.242	0.277	0.308

Table 2. Average Causal Effect(MNIST)

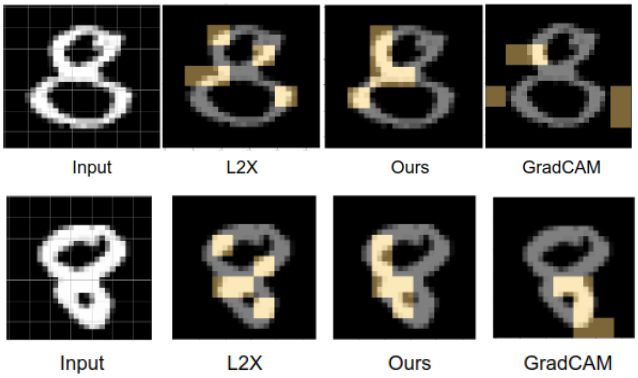

Figure 2. In the above figure, we can see that selected patches(highlighted in copper color) of our method capture class discriminative regions(i.e. regions differentiating 3 vs 8) in the query image.(Here, k=5)

FMNIST: This data set has 28×28 images of fashion items such as t-shirts, shoes, purse etc. We use a subset of the classes i.e. class 0 and 9(t-shirt and shoe) for experimentation. We train a simple convolutional neural net of same architecture as before for MNIST, and achieve 99.9% accuracy on the test data. We parameterize the selector network of L2X and our method by a 3 layer fully convolutional net.

Method	*k=4*	*k=6*	*k=8*
Our	**0.910± 0.022**	0.956± 0.014	**0.978± 0.005**
L2X	0.885± 0.026	**0.963± 0.006**	0.970± 0.013
GradCAM	0.589	0.636	0.679
Saliency	0.558	0.831	0.927

Table 3. Post-hoc accuracy(FMNIST)

Method	*k=4*	*k=6*	*k=8*
Our	**0.177± 0.009**	**0.193± 0.016**	**0.163± 0.007**
L2X	0.138± 0.027	0.173± 0.018	0.142± 0.016
GradCAM	-0.113	-0.159	-0.196
Saliency	-0.053	0.053	0.071

Table 4. Average Causal Effect(FMNIST)

CIFAR: This data set has 32×32 images of 10 different classes. We use a subset of the classes i.e. class 2 and 9(bird and truck) for experimentation. We train a 3 layer convolutional neural net, and achieve 94% accuracy on the test data. We parameterize the selector network of L2X and our method by a 3 layer fully convolutional net.

Method	*20% pixels*	*30% pixels*	*40% pixels*
Our	**0.600± 0.060**	**0.720± 0.050**	**0.780± 0.030**
L2X	0.510± 0.130	0.600± 0.010	0.660± 0.010
GradCAM	0.580	0.660	0.710
Saliency	0.551	0.570	0.610

Table 5. Post-hoc accuracy(CIFAR)

Method	*20% pixels*	*30% pixels*	*40% pixels*
Our	**0.078± 0.055**	**0.130± 0.048**	**0.153± 0.028**
L2X	-0.017± 0.12	0.080± 0.001	0.101-0.001
GradCAM	0.055	0.08	0.088
Saliency	0.028	0.033	-0.005

Table 6. Average Causal Effect(CIFAR)

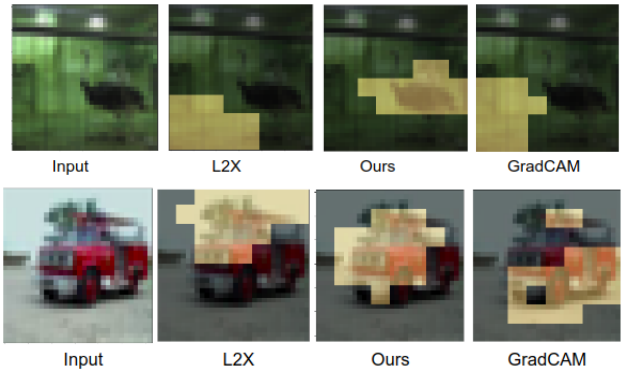

Figure 3. The first row shows 20% pixels being selected, and the second row shows result of selecting 30% pixels. Our method appears to focus well on the object in the scene, w.r.t. other methods.

4. Conclusion

In this work, we derive a causal objective from a rigorously chosen causal strength measure for the task of instance-wise feature subset selection. We also describe a training procedure for solving our causal objective function for real-world experiments. Finally, through carefully chosen metrics we evaluate the proposed method on multiple vision datasets and show its efficacy w.r.t other existing methods. When sparse explanations are required, our method often finds discriminative salient objects. We also provide the code [1] for reproducing our results.

References

[1] Julius Adebayo, J. Gilmer, Michael Muelly, I. Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *NeurIPS*, 2018.

[2] Shiyu Chang, Yang Zhang, Mo Yu, and Tommi Jaakkola. Invariant rationalization. In *International Conference on Machine Learning*, pages 1448–1458. PMLR, 2020.

[3] Aditya Chattopadhyay, Piyushi Manupriya, Anirban Sarkar, and Vineeth N Balasubramanian. Neural network attributions: A causal perspective. In *International Conference on Machine Learning*, pages 981–990. PMLR, 2019.

[4] Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. Learning to explain: An information-theoretic perspective on model interpretation. In *International Conference on Machine Learning*, pages 883–892. PMLR, 2018.

[5] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[6] Dominik Janzing, David Balduzzi, Moritz Grosse-Wentrup, Bernhard Schölkopf, et al. Quantifying causal influences. *The Annals of Statistics*, 41(5):2324–2358, 2013.

[7] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

[8] Matthew O'Shaughnessy, Gregory Canal, Marissa Connor, Mark Davenport, and Christopher Rozell. Generative causal explanations of black-box classifiers. *arXiv preprint arXiv:2006.13913*, 2020.

[9] Patrick Schwab and Walter Karlen. Cxplain: Causal explanations for model interpretation under uncertainty. In *Advances in Neural Information Processing Systems*, pages 10220–10230, 2019.

[10] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[11] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[12] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. Invase: Instance-wise variable selection using neural networks. In *International Conference on Learning Representations*, 2018.

[1] https://github.com/pranoy-panda/Causal-Feature-Subset-Selection