Accurate and Robust Feature Importance Estimation under Distribution Shifts

Jayaraman J. Thiagarajan[1*], **Vivek Narayanaswamy**[2], **Rushil Anirudh**[1], **Peer-Timo Bremer**[1], **Andreas Spanias**[2]

[1]Lawrence Livermore National Labs
[2]Arizona State University
jjayaram@llnl.gov, vnaray29@asu.edu, anirudh1@llnl.gov, bremer5@llnl.gov, spanias@asu.edu

Abstract

With increasing reliance on the outcomes of black-box models in critical applications, post-hoc explainability tools that do not require access to the model internals are often used to enable humans understand and trust these models. In particular, we focus on the class of methods that can reveal the influence of input features on the predicted outputs. Despite their wide-spread adoption, existing methods are known to suffer from one or more of the following challenges: computational complexities, large uncertainties and most importantly, inability to handle real-world domain shifts. In this paper, we propose PRoFILE (Producing Robust Feature Importances using Loss Estimates), a novel feature importance estimation method that addresses all these challenges. Through the use of a loss estimator jointly trained with the predictive model and a causal objective, PRoFILE can accurately estimate the feature importance scores even under complex distribution shifts, without any additional re-training. To this end, we also develop learning strategies for training the loss estimator, namely contrastive and dropout calibration, and find that it can effectively detect distribution shifts. Using empirical studies on several benchmark image and non-image data, we show significant improvements over state-of-the-art approaches, both in terms of fidelity and robustness.

Introduction

With the increased adoption of machine learning (ML) models in critical decision-making, post-hoc interpretability techniques are often required to enable decision-makers understand and trust these models. The *black-box* nature of ML models in most real-world settings (either due to their high complexity or proprietary nature) makes it challenging to interrogate their functioning. Consequently, attribution methods, which estimate the influence of different input features on the model output, are commonly utilized to explain decisions of such black-box models. Existing approaches for attribution, or more popularly feature importance estimation, range from sensitivity analysis (Ribeiro, Singh, and Guestrin 2016; Lundberg and Lee 2017), studying change in model confidences through input feature masking (Schwab and Karlen 2019) to constructing simpler explanation models (e.g. linear, tree- or rule-based) that mimic a black-box model (Schwab and Hlavacs 2015; Lakkaraju et al. 2019).

Though sensitivity analysis techniques such as LIME (Ribeiro, Singh, and Guestrin 2016) and SHAP (Lundberg and Lee 2017) are routinely used to explain individual predictions of any black-box classifier, they are computationally expensive. This challenge is typically handled in practice by constructing a *global* set of explanations using a sub-modular pick procedure (Ribeiro, Singh, and Guestrin 2016). On the other hand, despite being scalable, methods that construct simpler explanation models (Lakkaraju et al. 2019) are not guaranteed to match the behavior of the original model. While the recently proposed CXPlain (Schwab and Karlen 2019) addresses the scalability issue of feature masking methods, they are specific to the type of masking (e.g., zero masking) and the explainer needs to be re-trained if that changes (e.g., mean masking). Finally, and most importantly, it has been well documented that current approaches are highly sensitive to distribution shifts (Lakkaraju, Arsov, and Bastani 2020) and vulnerable to even small perturbations. Recently, Lakkaraju *et al.* (Lakkaraju, Arsov, and Bastani 2020) formalized this problem for the case of model mimicking approaches, and showed how adversarial training can be used to produce consistent explanations. In CXPlain, Schwab *et al.* proposed an ensembling strategy to effectively augment explanations with uncertainty estimates to better understand the explanation quality. However, they did not study the consistency of inferred explanations under distribution shifts.

In this work, we propose PRoFILE (Producing Robust Feature Importances using Loss Estimates), a novel feature importance estimation method that is highly accurate, computationally efficient, consistent with the black-box model being explained and robust under distribution shifts. The key idea of our approach is to jointly train a loss estimator while building the predictive model, and generate post-hoc explanations by measuring the influence of input features on the model output using a causal objective defined on the loss estimates. Furthermore, we introduce two different learning objectives to optimize the loss estimator, namely contrastive training and dropout calibration. Note that, once trained, the loss estimator can also be treated as a black-box. Interestingly, we find that the loss estimator is easier

*This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

to train than obtaining calibrated uncertainty estimates, yet produces higher fidelity explanations. Further, unlike existing approaches, PRoFILE requires no re-training at explanation time and natively supports arbitrary masking strategies. Finally, using a variety of benchmarks, we show that the resulting explanations are robust under regimes of distribution shifts where the black-box generalizes to. In summary, our contributions are:

- A computational efficient feature masking-based explainability method that is agnostic to the type of masking;
- Our approach is applicable to any data modality, deep architecture, or task;
- Learning objectives to train a loss estimator alongside the classifier;
- Experiments on a wide variety of both synthetic and real world data demonstrating the efficacy of PRoFILE under distribution shifts.

Related Work

Post-hoc explanation methods are the *modus-operandi* in interpreting the decisions of a black box model. Broadly, these approaches can be categorized as methods that generate explanations based on (a) sensitivity analysis; (b) gradients between the output and the input features; (c) change in model confidence through input feature masking; and (d) constructing simpler explanation models that can well approximate the black box predictor. LIME (Ribeiro, Singh, and Guestrin 2016) and SHAP (Lundberg and Lee 2017) are two popular sensitivity analysis methods, and they produce sample-wise, local explanations based on regression models by measuring the sensitivity of the black-box to perturbations in the input features. However, these methods are known to involve significant computational overheads. On the other hand, Saliency Maps (Simonyan, Vedaldi, and Zisserman 2013), Integrated Gradients (Sundararajan, Taly, and Yan 2017), Grad-CAM (Selvaraju et al. 2017), DeepLIFT (Shrikumar, Greenside, and Kundaje 2017) and a gradient based version of SHAP - DeepSHAP (Lundberg and Lee 2017), are examples of gradient-based methods which are computationally effective. More recently, Schwab *et al.* proposed CXPlain (Schwab and Karlen 2019) and Attentive Mixture of Experts (Schwab, Miladinovic, and Karlen 2019), which are popular examples for methods that estimate model confidences through feature masking. Trained using a Granger causality-based objective (Granger 1969), these methods produce attention scores reflective of the feature importances, at a significantly lower computational cost. Finally, global explanation methods rely on mimicking the black-box using simpler explainer functions. For instance, ROPE (Lakkaraju, Arsov, and Bastani 2020) and MUSE (Lakkaraju et al. 2019) construct scalable, simple linear models and decision sets, to emulate black-box models. An inherent challenge of this class of approaches is that the simple explainers are not guaranteed to match the behavior of the original model.

While these classes of methods vary in terms of their fidelity and complexity, a common limitation that has come to light recently is that explanations from most existing methods are associated with large uncertainties (Zhang et al. 2019) and are not robust under distribution shifts. Recently, Lakkaraju *et al.* (Lakkaraju, Arsov, and Bastani 2020) explored the use of adversarial minmax training to ensure that the mimicking explainer model is consistent with the black-box under adversarial perturbations. In contrast, we find that, without any adversarial training, PRoFILE estimates feature importances robustly under distribution shifts, is computationally scalable compared to existing local explanation methods, and produces higher fidelity explanations.

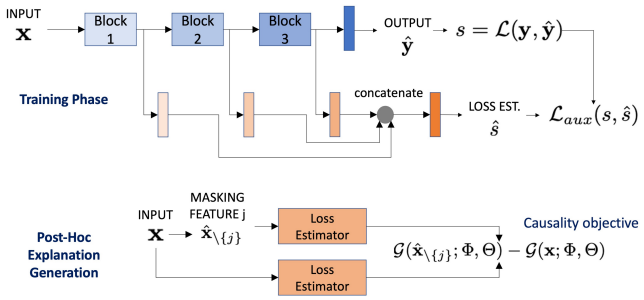


Figure 1: An illustration of the proposed approach, PRoFILE, for feature importance estimation. (top) During the training phase, we train a loss estimator along with the predictive model; (bottom) We use a Granger causality-based objective to generate post-hoc explanations using the loss estimates with no re-training.

Proposed Approach

Predictive Model Design with Loss Estimation. We consider the setup where we build a predictive model $\mathcal{F}(\Theta)$ which takes as input a sample $\mathbf{x} \in \mathbb{R}^d$ with d features and produces the output $\hat{\mathbf{y}} \in \mathbb{R}^k$ of dimensionality k. Note that this setup is deliberately unspecific in terms of both model architecture and data modality as PRoFILE is agnostic to either. Given a training set $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$, we optimize for the parameters Θ using the loss function $\mathcal{L} : \mathbf{y} \times \hat{\mathbf{y}} \rightarrow s$, where $s \in \mathbb{R}$. In other words, $\mathcal{L}$ measures the discrepancy between the true and predicted outputs using a pre-specified error metric. Examples include categorical cross-entropy for classification or mean-squared error for regression.

While our approach does not need access to the training data or specifics of the training procedure while generating explanations, similar to any post-hoc interpretability method, our approach requires the training of an auxiliary network $\mathcal{G}(\Phi; \Theta)$ that takes the same input $\mathbf{x}$ and produces the output $\hat{s} \approx \mathcal{L}(\mathbf{y}, \mathcal{F}(\mathbf{x}))$. The objective of this network is to directly estimate the fidelity for the prediction that $\mathcal{F}$ makes for $\mathbf{x}$, which we will use in order to construct our post-hoc explanations without any additional re-training. Note that, the loss estimates implicitly provide information about the inherent uncertainties; for example, in (Ash et al. 2020), the gradients of loss estimates have been used to capture the model uncertainties. We define the auxiliary objective $\mathcal{L}_{aux} : s \times \hat{s} \rightarrow \mathbb{R}$, in order to train the parameters Φ of model $\mathcal{G}$. As showed in Figure 1(top), the loss esti-

mator $\mathcal{G}$ uses the latent representations from different stages of $\mathcal{F}$ (e.g., every layer in the case of an FCN or every convolutional block in a CNN) to estimate $\hat{s}$. We use a linear layer along with non-linear activation (ReLU in our experiments) to transform each of the latent representations from $\mathcal{F}$ and they are finally concatenated to predict the loss. During training, the gradients from both the losses are used to update the parameters Θ of model $\mathcal{F}$.

Learning Objectives. Since the proposed feature estimation strategy relies directly on the quality of the loss estimator, the choice of the loss function $\mathcal{L}_{aux}$ is crucial. In particular, our approach (see Figure 1(bottom)) is based on ranking input features using the loss values obtained by masking those features. Consequently, we expect the loss estimator to preserve the ordering of samples (based on their losses), even if the original scale is discarded. Here, we explore the use of two different objectives to minimize the discrepancy between true and estimated losses.

(a) Contrastive Training: This is a widely adopted strategy when relative ordering of samples needs to be preserved. Given the loss values $\{s_i, s_j\}$ for a pair of samples $\{\mathbf{x}_i, \mathbf{x}_j\}$ in a mini-batch, we adopt an objective similar to (Yoo and Kweon 2019), which ensures that the sign of the difference $(s_i - s_j)$ is preserved in the corresponding loss estimates $(\hat{s}_i - \hat{s}_j)$. Formally, we use the following contrastive loss:

$$\mathcal{L}^{C}_{aux} = \sum_{(i,j)} \max\Big(0, -\mathbb{I}(s_i, s_j).(\hat{s}_i - \hat{s}_j) + \gamma\Big), \quad (1)$$

$$\text{where } \mathbb{I}(s_i, s_j) = \begin{cases} 1 & \text{if } s_i > s_j, \\ -1 & \text{otherwise.} \end{cases}$$

Note, when the sign of $s_i - s_j$ is positive, we assign a non-zero penalty if the estimates $\hat{s}_j > \hat{s}_i$, i.e., there is a disagreement in the ranking of samples. Here, γ is an optional margin hyper-parameter.

(b) Dropout Calibration: In this formulation, we utilize prediction intervals from the model $\mathcal{F}$ and adjust the loss estimates from $\mathcal{G}$ using an interval calibration objective. The notion of interval calibration comes from the uncertainty quantification literature and is used to evaluate uncertainty estimates in continuous-valued regression problems (Thiagarajan et al. 2020). In particular, we consider the epistemic uncertainties estimated using Monte Carlo dropout (Gal and Ghahramani 2016) to define the prediction interval $[\mu_{s_i} - \sigma_{s_i}, \mu_{s_i} + \sigma_{s_i}]$ for a sample $\mathbf{x}_i$. More specifically, we perform T independent forward passes with $\mathcal{F}$ to compute the mean μ_{s_i} and standard deviation σ_{s_i}. For the loss estimator $\mathcal{G}$, we use the latent representations averaged across T passes (for every block in Figure 1(top)) to obtain the estimate $\hat{s}_i$. Finally, we use a hinge loss objective to calibrate the estimates:

$$\mathcal{L}^{DC}_{aux} = \sum_{i} \max\Big(0, \hat{s}_i - (\mu_{s_i} + \sigma_{s_i}) + \xi\Big) \quad (2)$$

$$+ \max\Big(0, (\mu_{s_i} - \sigma_{s_i}) - \hat{s}_i + \xi\Big) \quad (3)$$

Here, ξ is the optional margin parameter and the objective encourages the estimates $\hat{s}_i$ to lie in the prediction interval for s from the model $\mathcal{F}$.

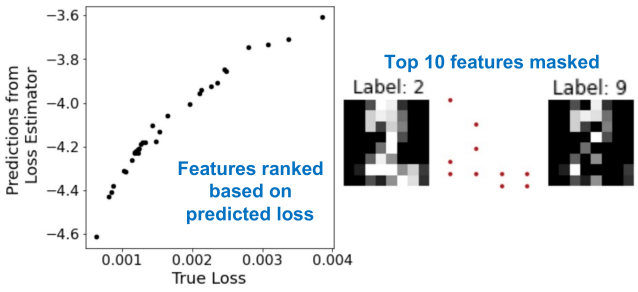


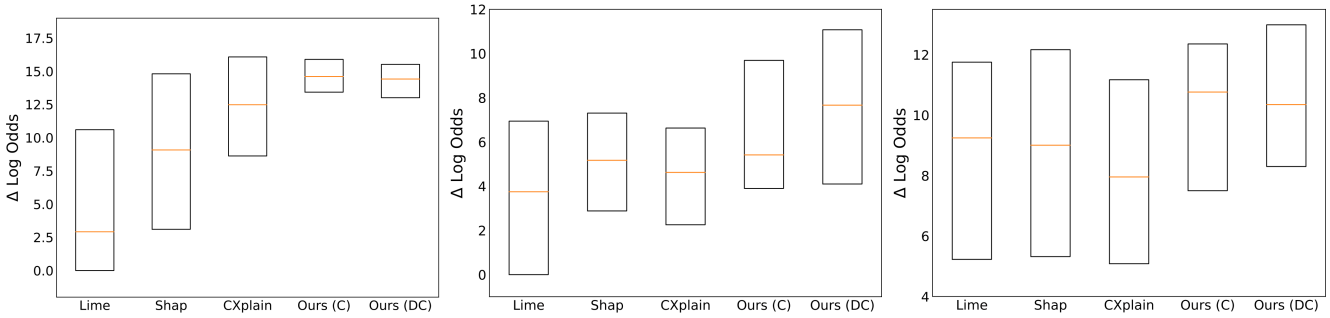
Figure 2: Demonstration of PRoFILE using the UCI handwritten digits dataset. Here, we show an example where the loss estimator was trained using the contrastive loss. For this test sample, the ranking obtained using the estimated loss agrees with that from the true loss (known ground truth). When we mask the top 10 features from PRoFILE and as expected, there is a change in the model prediction.

Feature Importance Estimation. Given the loss estimator $\mathcal{G}$, we estimate the feature importance using a Granger causality-based objective, similar to (Schwab and Karlen 2019). The Humean definition of causality adopted by Granger (Granger 1969) postulates that a causal relationship exists between random variables x_j and y, i.e., $x_j \rightarrow y$, if we can better predict using all available information than the case where the variable x_j was excluded. This definition is directly applicable to our setting since it satisfies the key assumptions of Granger causality analysis, our data sample $\mathbf{x}$ contains all relevant variables required to predict the target and $\mathbf{x}$ temporally precedes $\mathbf{y}$. Mathematically,

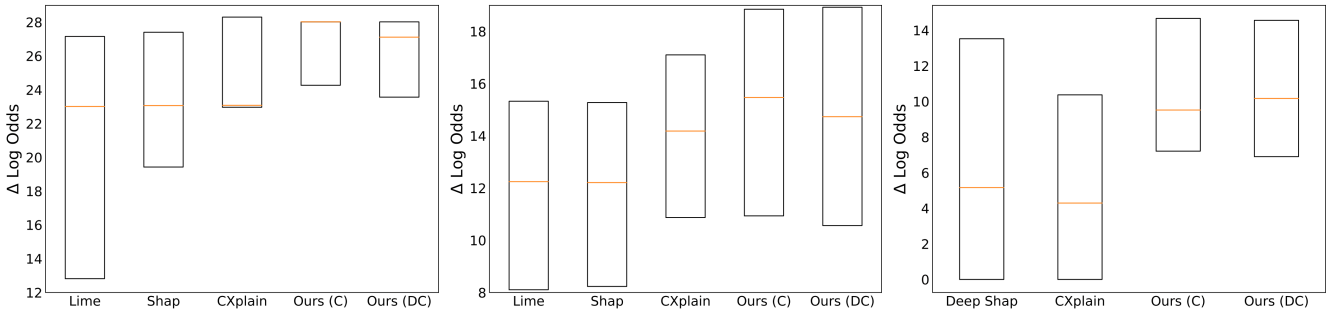
$$\Delta\epsilon_{\mathbf{x},j} = \epsilon_{\mathbf{x}\setminus\{j\}} - \epsilon_{\mathbf{x}}, \quad (4)$$

where ϵ denotes the model error. For a sample $\mathbf{x}$, we can compute this objective for each feature j to construct the explanation. As showed in Figure 1(bottom), we use the loss estimator to measure the predictive model's error in the presence and absence of a variable x_j to check if x_j causes the predicted output. There are a variety of strategies that can be adopted to construct $\mathbf{x} \setminus \{j\}$. In the simplest form, we can mask the chosen feature by replacing it with zero or a pre-specified constant. However, in practice, one can also adopt more sophisticated masking strategies that take into account the underlying data distribution (Janzing et al. 2013; Štrumbelj, Kononenko, and Šikonja 2009). Interestingly, our approach is agnostic to the masking strategy and the loss estimator can be used to compute the causal objective in Eqn.(4) for any type of masking. In contrast, existing approaches such as CXPlain requires re-training of the explanation model for the new masking strategy.

Since the loss estimator is jointly trained with the main predictor, our approach does not require any additional adversarial training as done in (Lakkaraju, Arsov, and Bastani 2020) to ensure that the explanations are consistent with the black-box. It must be noted that adversarial training for improving the robustness of the black box is independent of the design of PRoFILE. As the black box becomes more robust to arbitrary shifts/adversarial perturbations, we expect PRoFILE explanations to still be consistent.



(a) UCI Digits ($k = 15\%$) (b) Kropt ($k = 15\%$) (c) RBF ($k = 15\%$)



(d) Poker ($k = 15\%$) (e) Handwritten Letters ($k = 15\%$) (f) Cifar-10 ($k = 15\%$)

Figure 3: Comparing the fidelity of feature importances inferred using different methods. We use the Δlog-odds score (higher the better) obtained by masking the most influential input features. For each of the datasets, the ratio of features masked is also included in parentheses. Across all benchmarks, the proposed approach is consistently superior over the baselines.

Existing works in the active learning literature have also found that the loss function (Yoo and Kweon 2019) or its gradients (Ash et al. 2020) effectively capture the inherent uncertainties in a model and hence can be used for selecting informative samples. Using a similar argument, we show that even though our causal objective is similar to CXPlain, our approach more effectively generalizes to even complex distribution shifts where CXPlain fails.

Demonstration. For demonstrating the behavior of our approach, we consider the UCI handwritten digits dataset (Dua and Graff 2017) comprised of 8×8 grayscale images. In Figure 2, we show predictions from our loss estimator (contrastive training) for a test image, when each of the 64 pixels were masked (replaced with zero). We find that, though the scale of the loss function is discarded, the ordering of the features is well preserved. We also illustrate the explanation obtained by masking the top 10 features identified using the causal objective in Eqn.(4). The observed changes in the prediction (from class 2 to class 9) is intuitive and demonstrates the effectiveness of our approach.

Empirical Results

In this section, we present empirical studies to compare PRoFILE against popular baselines using both non-image and image benchmark datasets. More importantly, we evaluate the fidelity of the inferred explanations under challenging distribution shifts and demonstrate the effectiveness of PRoFILE. Before we present our findings, we will discuss the datasets, baselines and metrics used in our study.

Datasets. We consider a suite of synthetic and real-world datasets to evaluate the fidelity and robustness of our approach. For the fidelity comparison study under standard testing conditions, we use the: (a) UCI Handwritten Digits dataset, (b) OpenML benchmarks (Vanschoren et al. 2013), Kropt, Letter Image Recognition, Pokerhand and RBF datasets and (c) Cifar10 image classification dataset (Krizhevsky and Hinton 2009). The dimensionality of the input data ranges from 10 to 64 and the total number of examples varies between 1797 and 13750 for different benchmarks. For each of the UCI and OpenML datasets, we utilized $\sim 90\%$ of the data while for Cifar10, we used the prescribed dataset of 50K RGB images of size 32×32 for training our proposed model.

For the robustness study, we used the following datasets: (a) *Synthetic dataset*: In order to study the impact of distribution shifts on explanation fidelity, we constructed synthetic data based on correlation and variance shifts to the data generation process defined using a multi-variate normal distribution. More specifically, we generated multiple synthetic datasets of 5K samples, where the number of covariates was randomly varied between 10 and 50. In each case, the samples were drawn from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\mu_{ii} = \alpha, \Sigma_{ii} = 1$ and $\Sigma_{ij} = \beta$ and the values α, β (the correlation be-

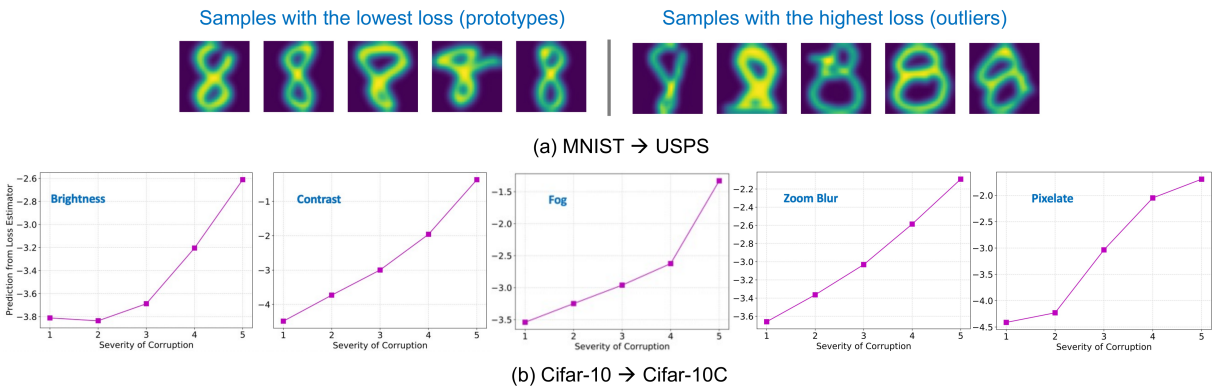


Figure 4: Effectiveness of our loss estimator $\mathcal{G}$ in detecting distribution shifts, even though the shifts are not known during training. In the MNIST-USPS case, it attributes non-typical writing styles from the USPS dataset, that are not found in MNIST, with high loss values. Similarly, in the case of Cifar-10C, the loss estimates from $\mathcal{G}$, averaged across 500 test samples, monotonically grows as the severity of the corruption increases.

tween any two variables) were randomly chosen from the uniform intervals $[-2, 2]$ and $[-1, 1]$ respectively. The label for each sample was generated using their corresponding quantiles (i.e., defining classes separated by nested concentric multi-dimensional spheres). To generate correlation shifts, we created new datasets following the same procedure, but using a different correlation $\bar{\beta} = \beta + \delta_\beta$. Here, δ_β was randomly drawn from the uniform interval $[-0.2, 0.2]$. Next, we created a third dataset to emulate variance shifts, wherein we changed the variance $\Sigma_{ii} = \Sigma_{ii} + \kappa$ and κ was drawn from the uniform interval $[0.25, 0.75]$. While the predictive model was trained only using the original dataset, the explanations were evaluated using both correlation- and variance-shifted datasets. We generated 10 different realizations with this process and report the explanation fidelity metrics averaged across the 10 trials; (b) *Cifar10 to Cifar10-C* (Hendrycks and Dietterich 2019): This is a popular benchmark for distribution-shift studies, wherein we train the predictive model and loss estimator using the standard Cifar10 dataset and generate explanations for images from the Cifar10-C dataset containing wide-variety of natural image corruptions; and (c) *MNIST-USPS*: In this case, we train the predictive model using only the MNIST handwritten digits dataset (LeCun, Cortes, and Burges 2010) and evaluate the explanations on the USPS dataset (Hull 1994) at test time.

Baselines. We compared PRoFILE against the following baseline methods that are commonly adopted to produce sample-level explanations. All baseline methods considered belong to the class of post-hoc explanation strategies which aim to construct interpretable models that can approximate the functionality of any black-box predictor.
(i) LIME (Ribeiro, Singh, and Guestrin 2016)[1]: LIME constructs linear models, which can locally approximate a black box predictor, by fitting a weighted regression model around the sample to be explained based on variants of the sample

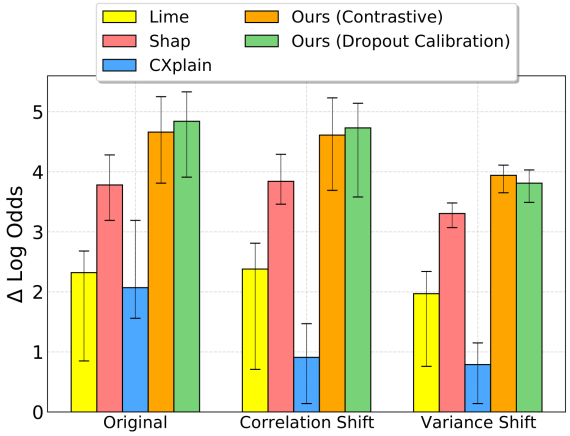


Figure 5: Using a synthetic dataset to study the robustness of explanations obtained using different approaches, under correlation and variance shifts. We mask the top 25% of features in the data to obtain the Δlog-odds scores.

obtained by perturbing or zero-masking the input features. The intuition is that the post-hoc regression model obtained is reflective of the sensitivity of the black-box predictor to the modifications in the input features. The coefficients of the obtained post-hoc model serve as attribution scores for each feature in the given sample.
(ii) Shap (Lundberg and Lee 2017)[2]: SHAP determines the feature attribution scores for a sample by marginalizing the individual contributions of every feature towards a prediction. SHAP, more specifically KernelSHAP, fits a local regression model around the sample to be explained using multiple realizations of the sample by zero masking single or groups of features. A fundamental difference between LIME and SHAP lies in the SHAP kernel used, which is a function of the cardinality of the features present in a group. The co-

[1]code: https://github.com/marcotcr/lime

[2]code: https://github.com/slundberg/shap

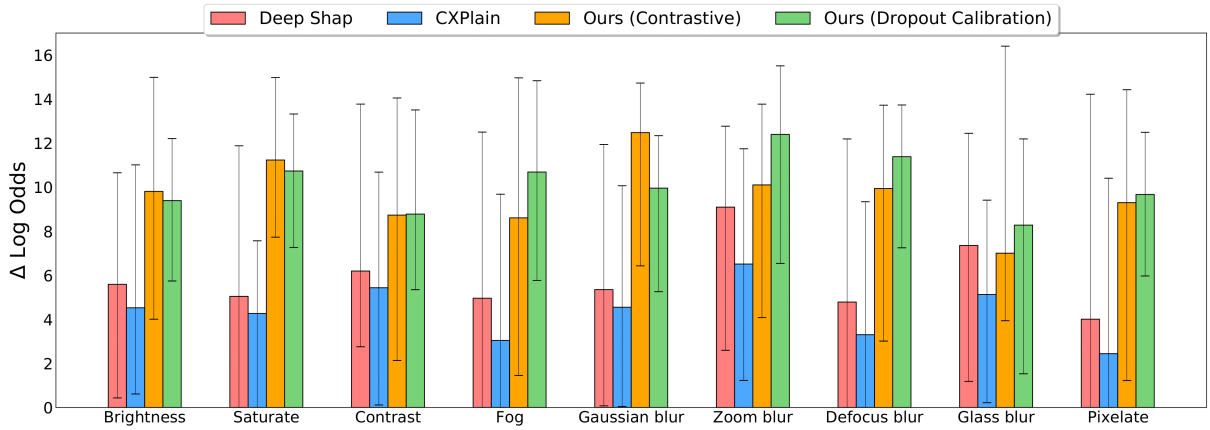


Figure 6: *Cifar-10C dataset*: We study the fidelity of explanations generated on different types of corrupted images using the loss estimator trained on the original Cifar-10 data.

efficients of the obtained model are the SHAPLey attribution scores for every feature in the given sample.
(iii) CXPlain (Schwab and Karlen 2019)[3]: This determines feature attribution scores by training a post-hoc model that learns to approximate the distribution of Granger causal errors (Granger 1969), i.e., the difference between the black-box prediction loss when no feature is masked and the loss when features are zero-masked one at a time. The feature attribution scores obtained from the model are thus reflective of the global distribution of the causality based error metric. Similar to (Schwab and Karlen 2019), we use an MLP and a U-Net model as the post-hoc explainer for the non-image and the image datasets respectively in our experiments.
(iv) Deep Shap (Lundberg and Lee 2017) DeepSHAP is a fast and scalable approximation of SHAP and also closely related to the DeepLIFT algorithm. We utilize this baseline on datasets where LIME and SHAP were expensive to run.

Evaluation Metric. To evaluate the explanation fidelity, we utilize the commonly used difference in log-odds metric, which is a measure of change in prediction when $k\%$ of the most relevant features in the input data are masked.

$$\Delta\text{log-odds} = \text{log-odds}(\text{p}_{\text{ref}}) - \text{log-odds}(\text{p}_{\text{masked}}) \quad (5)$$

Here $\text{log-odds}(\text{p}) = \log(\frac{\text{p}}{1-\text{p}})$ and p_{ref} is the reference prediction probability of the original data and p_{masked} refers to the prediction probability when a subset of features are masked. A higher value for Δlog-odds implies higher fidelity of the feature importance estimation. More specifically, for: (a) *Non-Image Datasets*. We sort the feature attribution scores obtained from the explainability method (PRoFILE and baselines) and zero mask the top $k\%$ important features in the input sample to evaluate the metric, and (b) *Image Datasets*. We use the SLIC (Achanta et al. 2012) segmentation algorithm to generate superpixels, which are then used to compute the feature importance scores. For CXPlain and DeepSHAP, we aggregate the pixel-level feature importance scores to estimate attributions for each superpixel.

[3]code: https://github.com/d909b/cxplain

Hyperparameters. For all non-imaging datasets, the black-box model was a 5 layer MLP with ReLU activations, each fully-connected (FC) layer in the loss estimator contained 16 units. In the case of Cifar-10, we used the standard ResNet-18 architecture, and the loss estimator used outputs from each residual blocks (with fully connected layers containing 128 hidden units). Finally, for the MNIST-USPS experiment, we used a $3-$layer CNN with 2 FC layers. The loss estimator was designed to access outputs from the first 4 layers of the network and utilized FC layers with 16 units each. All networks were trained using the ADAM optimizer with a learning rate 0.001 and batch size 128.

Findings

PRoFILE Produces Higher Fidelity Explanations Figure 3 illustrates the Δlog-odds obtained using PRoFILE with both the proposed learning strategies (Ours(C) and Ours(DC)) in comparison to the baselines. Note that, for the UCI and OpenML datasets, we used the held-out test set for our evaluation (90-10 split), while for Cifar-10, we used 50 randomly chosen test images for computing the fidelity metric. While PRoFILE and CXPlain are scalable to larger test sizes, the small subset of test samples was used to tractably run the other baselines. For each dataset, we show the median (orange), along with the 25^{th} and the 75^{th} percentiles, of the Δlog-odds scores across the test samples. We find that PRoFILE consistently outperforms the existing baselines on all benchmarks. In particular, both contrastive training and dropout calibration strategies are effective and perform similarly in all cases. The improved fidelity can be attributed directly to the efficacy of the loss estimator and the causal objective used for inferring the feature attribution. In comparison, both LIME and SHAP produce lower fidelity explanations, while also being computationally inefficient. Interestingly, though CXPlain also uses a causal objective similar to us, the resulting explanations are of significantly lower fidelity. In terms of computational complexity for generating post-hoc explanations, PRoFILE which requires p evaluations (number of features that need to be masked and can

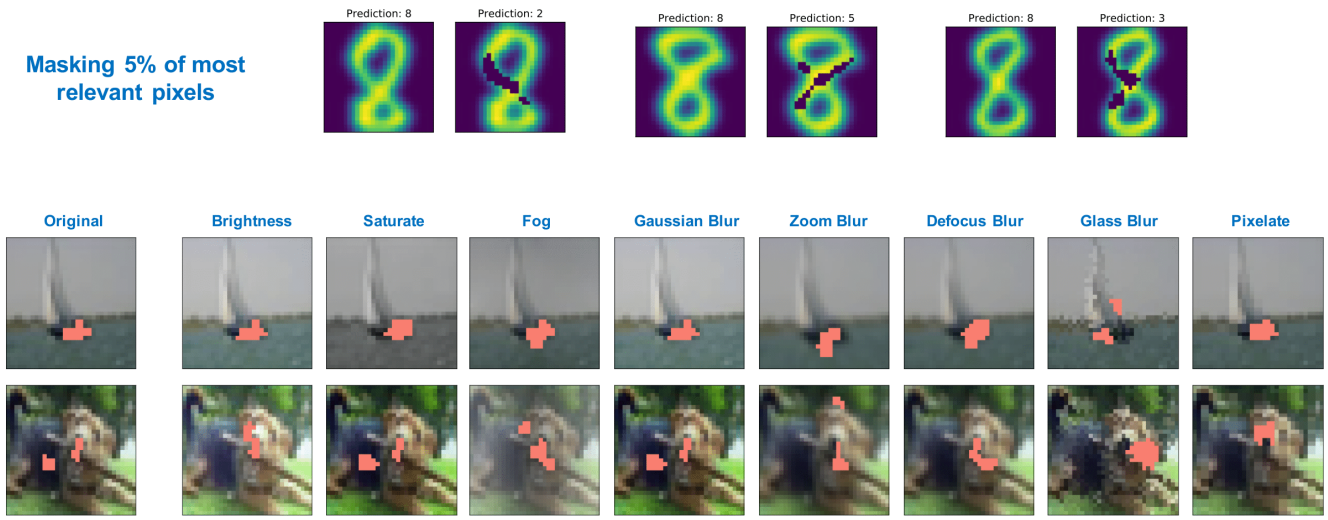


Figure 7: Examples of explanations generated using the proposed approach (with dropout calibration) on USPS and Cifar-10C datasets using models trained with MNIST and Cifar-10 respectively.

be parallelized) of the loss estimator and is only marginally more expensive than CXPlain.

Loss Estimator Detects Distribution Shifts At the core of our approach is the pre-trained loss estimator, which enables us to utilize the Granger causality objective to generate high-quality explanations. Consequently, the robustness of PRoFILE directly relies on how well the loss estimator can generalize under distribution shifts. We investigate the empirical behavior of the loss estimator using (i) MNIST-USPS and (ii) Cifar10 to Cifar10-C benchmarks. In both cases, we train the predictor and loss estimator using the original data (MNIST, Cifar10) and evaluate on the shifted data. In Figure 4(a), we show USPS images from class 8 with the lowest (in-distribution) and highest (out-distribution) loss estimates. While the former resemble the prototypical examples, the latter contains uncommon writing styles not found in the MNIST dataset. In case of Cifar10-C, we show the loss estimates for 5 different natural image corruptions (averaged across 500 examples). We observe a monotonic increase in the average loss estimates as the severity of the corruptions grow, thus demonstrating the ability of the loss estimator to detect distribution shifts.

PRoFILE Explanations are More Robust Following our observations on the behavior of the loss estimator, we now evaluate the fidelity of PRoFILE explanations in those scenarios. Figure 5 illustrates the median Δlog-odds and error bars obtained by masking the top 25% of features on 10 realizations of the synthetic dataset. In particular, we show the results for the held-out correlation and variance shifted data, while the models were trained only using the original synthetic data. We find that by utilizing a pre-trained loss estimator, PRoFILE significantly outperforms the baselines, even under complex shifts, indicating the robustness of our approach. Similar to the findings in (Lakkaraju, Arsov, and Bastani 2020), we note that the widely-adopted baselines are not immune to shifts. Figure 6 shows a detailed comparison of Δlog-odds for the Cifar10-C dataset. Note, we show the median, 25^{th} and 75^{th} percentiles. We find that PRoFILE consistently achieves superior fidelity, when compared to existing baselines, except in the case of *glass blur* where the scores are comparable.

Figure 7 shows examples of explanations obtained using PRoFILE on the USPS and Cifar10-C datasets. We observe from Figure 7 (top) that our method adapts well across domains to identify critical pixels that characterize class-specific decision regions. Interestingly, these are examples where digit 8 is suitably masked by PRoFILE (only 5% of pixels) to be predicted as one of the other classes sharing the decision boundary. It can also be seen from Figure 7 (bottom) that PRoFILE explanations obtained under different domain shifts are consistent. In all cases except *glass blur*, it identifies the hull of the boat and the mouth of the dog as critical features. These observations strongly corroborate with the performance improvements in Figure 6.

Conclusions

In this paper, we proposed PRoFILE, a novel post-hoc feature importance estimation method applicable to any data modality or architecture. In particular, PRoFILE trains an auxiliary estimator to estimate the expected loss, for a given sample, from the primary predictor model. To this end, we introduced two learning objectives, contrastive training and dropout calibration. Using the pre-trained loss estimator along with a causality based objective, PRoFILE can accurately estimate feature importance scores that are immune to a wide variety of distribution shifts. Through extensive experimental studies on different data modalities, we demonstrate that PRoFILE provides higher fidelity explanations, is robust under real-world distribution shifts and is computational effective when compared to commonly adopted feature importance estimation methods.

References

Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; and Süsstrunk, S. 2012. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence* 34(11): 2274–2282.

Ash, J. T.; Zhang, C.; Krishnamurthy, A.; Langford, J.; and Agarwal, A. 2020. Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds. In *International Conference on Learning Representations*.

Dua, D.; and Graff, C. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml, Last accessed on 08/01/2020.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, 1050–1059.

Granger, C. W. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society* 424–438.

Hendrycks, D.; and Dietterich, T. 2019. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. *Proceedings of the International Conference on Learning Representations* .

Hull, J. J. 1994. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 16(5): 550–554.

Janzing, D.; Balduzzi, D.; Grosse-Wentrup, M.; Schölkopf, B.; et al. 2013. Quantifying causal influences. *The Annals of Statistics* 41(5): 2324–2358.

Krizhevsky, A.; and Hinton, G. 2009. Learning multiple layers of features from tiny images. *Citeseer* .

Lakkaraju, H.; Arsov, N.; and Bastani, O. 2020. Robust Black Box Explanations Under Distribution Shift. *International Conference on Machine Learning (ICML)* .

Lakkaraju, H.; Kamar, E.; Caruana, R.; and Leskovec, J. 2019. Faithful and customizable explanations of black box models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 131–138.

LeCun, Y.; Cortes, C.; and Burges, C. 2010. MNIST handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist* 2.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, 4765–4774.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Schwab, P.; and Hlavacs, H. 2015. Capturing the Essence: Towards the Automated Generation of Transparent Behavior Models. In *AIIDE*, 184–190.

Schwab, P.; and Karlen, W. 2019. CXPlain: Causal explanations for model interpretation under uncertainty. In *Advances in Neural Information Processing Systems*, 10220–10230.

Schwab, P.; Miladinovic, D.; and Karlen, W. 2019. Granger-causal attentive mixtures of experts: Learning important features with neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 4846–4853.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Shrikumar, A.; Greenside, P.; and Kundaje, A. 2017. Learning Important Features Through Propagating Activation Differences. In *Proceedings of Machine Learning Research*, volume 70, 3145–3153.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* .

Štrumbelj, E.; Kononenko, I.; and Šikonja, M. R. 2009. Explaining instance classifications with interactions of subsets of feature values. *Data & Knowledge Engineering* 68(10): 886–904.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. *arXiv preprint arXiv:1703.01365* .

Thiagarajan, J. J.; Venkatesh, B.; Sattigeri, P.; and Bremer, P.-T. 2020. Building Calibrated Deep Models via Uncertainty Matching with Auxiliary Interval Predictors. In *AAAI*, 6005–6012.

Vanschoren, J.; van Rijn, J. N.; Bischl, B.; and Torgo, L. 2013. OpenML: Networked Science in Machine Learning. *SIGKDD Explorations* 15(2): 49–60. doi:10.1145/2641190.2641198. URL http://doi.acm.org/10.1145/2641190.2641198.

Yoo, D.; and Kweon, I. S. 2019. Learning loss for active learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 93–102.

Zhang, Y.; Song, K.; Sun, Y.; Tan, S.; and Udell, M. 2019. "Why Should You Trust My Explanation" Understanding Uncertainty in LIME Explanations. *arXiv preprint arXiv:1904.12991* .