Explaining Models by Propagating Shapley Values of Local Components

Hugh Chen[1], Scott Lundberg[2], Su-In Lee[1]

[1]Paul G. Allen School of Computer Science and Engineering, University of Washington
[2]Microsoft Research

UNIVERSITY of WASHINGTON

Introduction

Neural networks and ensembles of models are widely used across many domains. In medicine, explainable AI (XAI) is important for scientific discovery, transparency, and much more [1]. One popular XAI method is a per-sample feature attribution that assigns an importance value to each feature in a given prediction.

Approach

In this paper, we focus on SHAP values [2] - Shapley values with a conditional expectation of the model prediction as the set function. In order to approximate SHAP values for neural networks, DeepSHAP builds upon a previous method named DeepLIFT. In this section we explain how DeepLIFT's rules connect to SHAP values. This has been briefly touched upon in [2] and [3], but here we explicitly define the relationship by looking at a few scenarios:

a. In the linear case (a), the chain rule gives us a perfect approximation to the SHAP values.
b. Next, both rules proposed in DeepLIFT (b-c) can be viewed as an approximation to the SHAP values.
   - The Rescale rule (b) makes sure that attributions sum up appropriately for the smallest possible components.
   - The RevealCancel rule (c) gives a better approximation by grouping positive and negative features.
c. We can explain neural networks fed into trees (d) by combining DeepSHAP with methods to obtain SHAP values for trees.

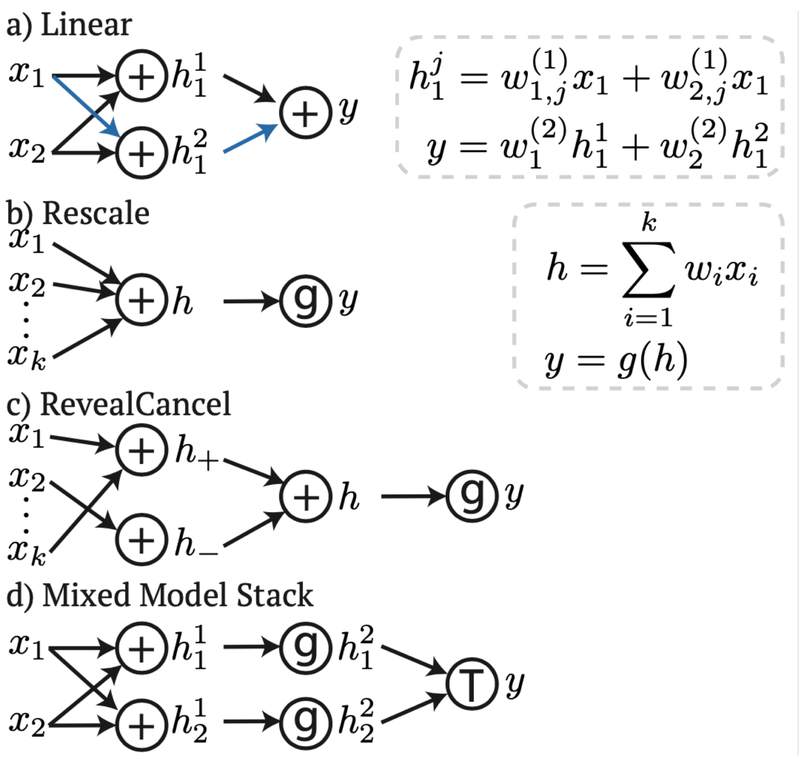

Fig. 1 Visualization of cases for to explain DeepLIFT's connection to SHAP values. T is a non-differentiable tree model and g is a non-linear function.

Approach (cont'd)

Finally, we prove that to obtain the SHAP value for a foreground sample with a uniform background distribution, we can simply average all the SHAP values for each sample in a background set. (Theorem in the paper)

Experiments

Background distributions avoid bias

Using a single black reference (DeepLIFT) has a bias that results in no attributions for darker pixels. For DeepSHAP, having many references solves this problem and we see attributions in sensical dark pixels (Fig. 2).

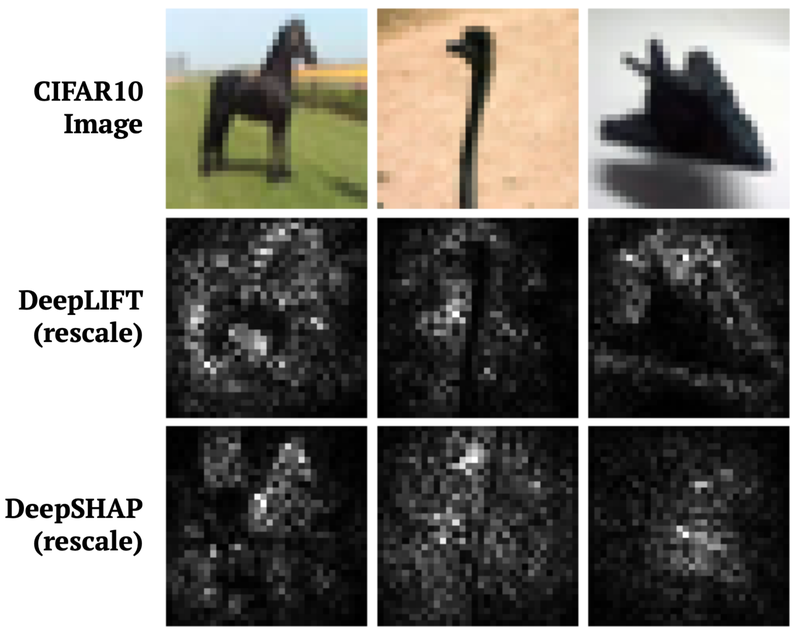

Fig. 2 Explaining black images on CIFAR10.

Explaining Mortality Prediction

We explain an MLP predicting 15 year mortality (82.6% test accuracy) based on NHANES I Epidemiologic Followup Study [4].

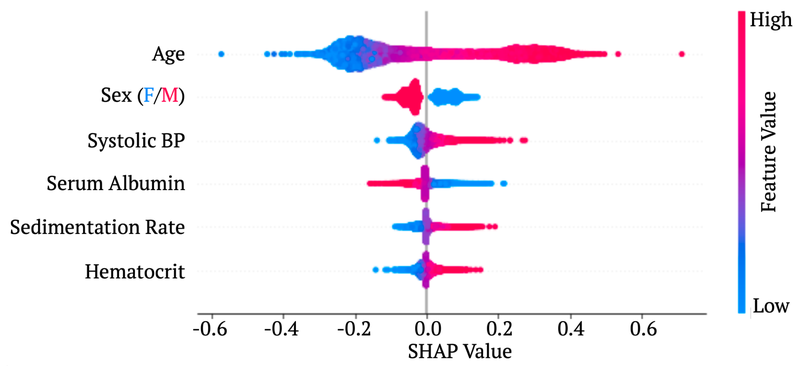

Fig. 3 Summary plot. Points are the feature attribution value colored by feature value.

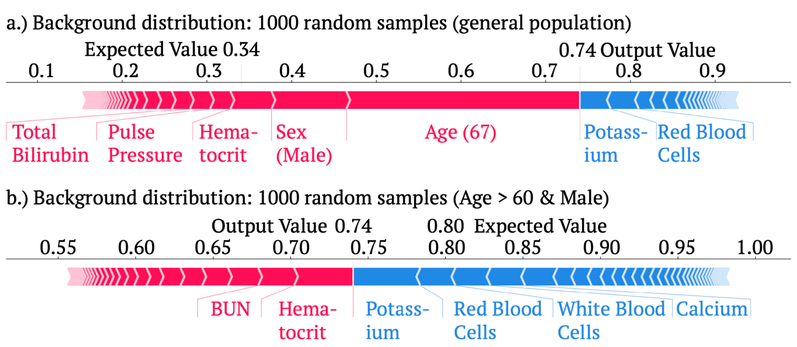

Fig. 4 Individual Explanation. Explaining an individual's mortality prediction for different background distributions.

Interpreting a stack of mixed model types

We evaluate the efficacy of DeepSHAP for a neural network feature extractor fed into a tree model on a simulated dataset with tightly correlated features. We use Independent Tree SHAP [5] to explain the tree model, which provides the same SHAP values DeepSHAP approximates.

We evaluate DeepSHAP with an ablation metric (*keep absolute (mask)*): 1) Obtain attributions for test samples 2) Mean impute all features (mask) 3) Introduce one feature at a time (unmask) from largest absolute attribution to smallest and measure R^2 (should initially increase rapidly, since we introduce the most important features first).

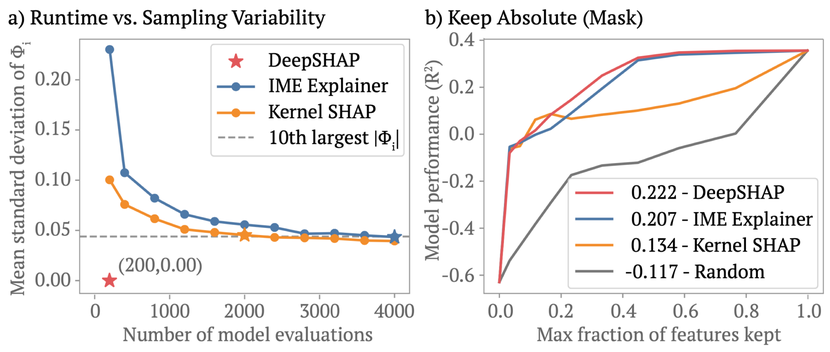

Fig. 5 Ablation test for explaining an LSTM feature extractor fed into an XGB model. [a.] Convergence of methods for a single explanation. [b.] Model performance versus # features kept.

Improving the RevealCancel rule

In this section, we propose a simple way to improve to the RevealCancel rule by grouping features by features that are larger or smaller than the mean rather than by positive and negative features.

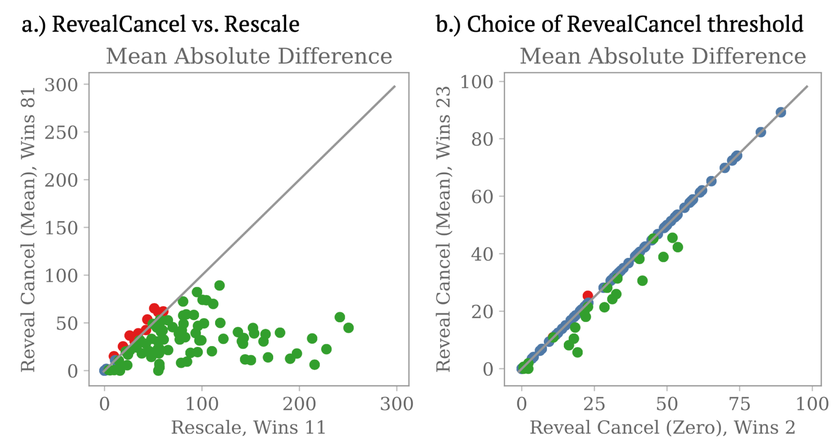

Fig. 6 Comparison of new RevealCancel rule for estimating SHAP values on a toy example. Axes correspond to mean absolute difference from SHAP values. Green means we win, and red means we lose.

References

1. Holzinger, Andreas, et al. "What do we need to build explainable AI systems for the medical domain?." *arXiv preprint arXiv:1712.09923* (2017).
2. Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." *Advances in neural information processing systems*. 2017.
3. Shrikumar, Avanti, Peyton Greenside, and Anshul Kundaje. "Learning important features through propagating activation differences." *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017.
4. Cox, Christine S. *Plan and operation of the NHANES I Epidemiologic Followup Study, 1992*. No. 35. National Ctr for Health Statistics, 1998.
5. Lundberg, Scott M., et al. "From local explanations to global understanding with explainable AI for trees." Nature Machine Intelligence, 2020.