Prediction Via Shapley Value Regression

Amr Alkhatib, Roman Bresson, Henrik Boström, Michalis Vazirgiannis

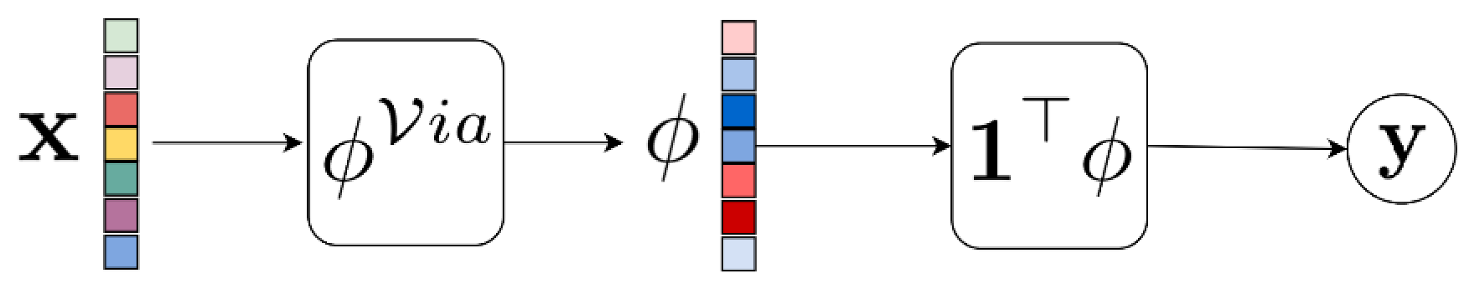
KTH Royal Institute of Technology

Department of Computer Science

Abstract

Shapley values have several desirable properties for explaining black-box model predictions, which come with strong theoretical support. Traditionally, Shapley values are computed post-hoc, leading to additional computational cost at inference time. To overcome this, we introduce ViaSHAP, a novel approach that learns a function to compute Shapley values, from which the predictions can be derived directly by summation. We explore two learning approaches based on the universal approximation theorem and the Kolmogorov-Arnold representation theorem. Results from a large-scale empirical investigation are presented, in which the predictive performance of ViaSHAP is compared to state-of-the-art algorithms for tabular data, where the implementation using Kolmogorov-Arnold Networks showed a superior performance. It is also demonstrated that the explanations of ViaSHAP are accurate, and that the accuracy is controllable through the hyperparameters.

ViaSHAP



- ✓ The Shapley values are not computed in a post-hoc setup
- ✓ The learning of Shapley values is integrated into the training of the predictive model
- ✓ The Shapley values are used directly to generate predictions

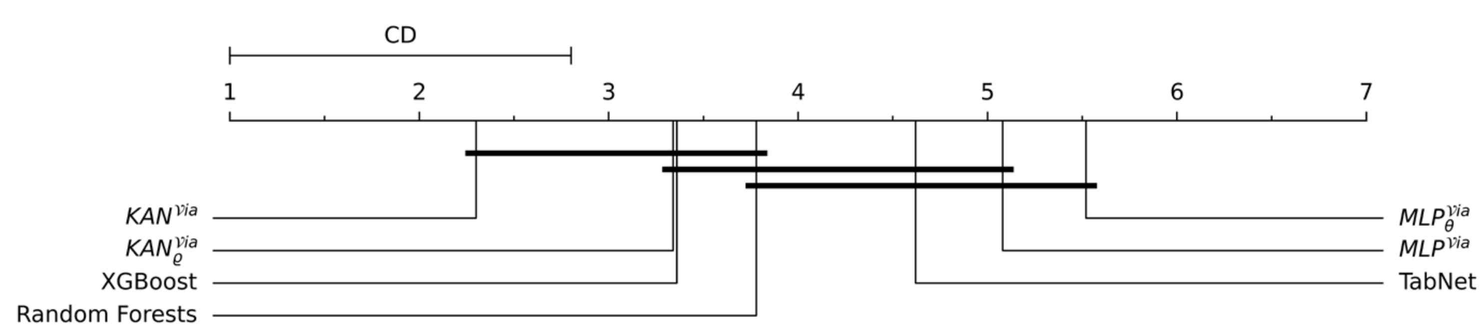
The Optimization of ViaSHAP

Algorithm 1: $\mathcal{V}ia^{SHAP}$

Data: training data X, labels Y, scalar β
Result: model parameters θ
Initialize $\mathcal{V} : \mathcal{V}ia^{SHAP}(\phi^{\mathcal{V}ia}(\mathbf{x};\theta))$
while *not converged* **do**
  $\mathcal{L} \leftarrow 0$
  for *each* $\mathbf{x} \in X$ *and* $\mathbf{y} \in Y$ **do**
    sample $S \sim p(S)$
    $\mathbf{y}' \leftarrow \mathcal{V}(\mathbf{x})$
    $\mathcal{L}_{pred} \leftarrow$ *prediction loss*$(\mathbf{y}', \mathbf{y})$
    $\mathcal{L}_{\phi} \leftarrow \left(\mathcal{V}_{\mathbf{y}}(\mathbf{x}^S) - \mathcal{V}_{\mathbf{y}}(\mathbf{0}) - \mathbf{1}_S^{\top}\phi_{\mathbf{y}}^{\mathcal{V}ia}(\mathbf{x};\theta)\right)^2$
    $\mathcal{L} \xleftarrow{+} \mathcal{L}_{pred} + \beta \cdot \mathcal{L}_{\phi}$
  end
  Compute gradients $\nabla_{\theta}\mathcal{L}$
  Update $\theta \leftarrow \theta - \nabla_{\theta}\mathcal{L}$
end

Empirical Evaluation Investigation

- ✓ We experimented with four different implementations of ViaSHAP, using Kolmogorov–Arnold Networks (KANs) [1] and feedforward neural networks
- ✓ The evaluation of predictive performance has been conducted using 25 public datasets. ViaSHAP is compared to the following baselines: Random Forests, XGBoost, and TabNet

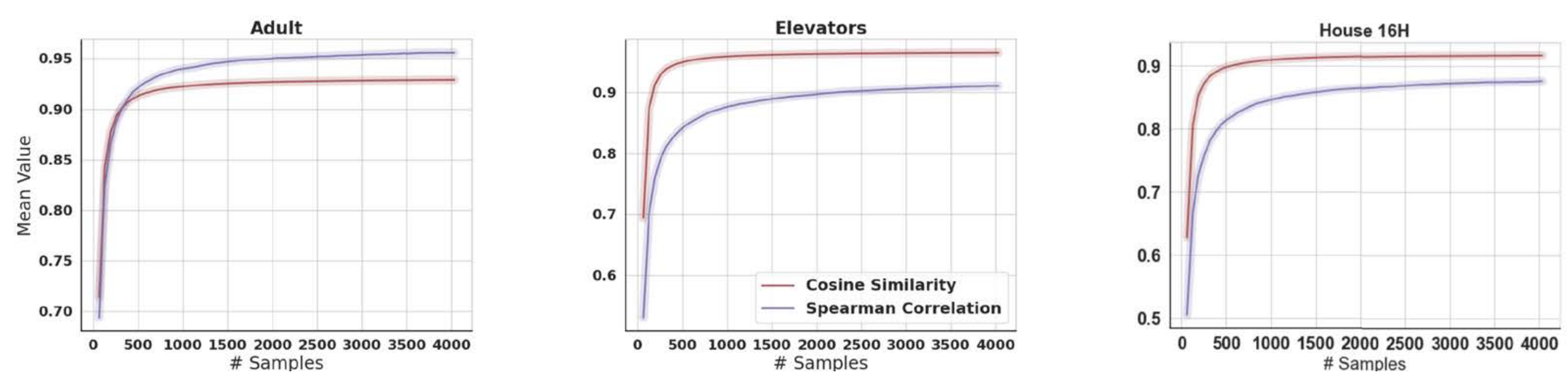


The average rank of the 7 predictors on the 25 datasets with respect to the AUC (the lower rank is better). The critical difference (CD) is the largest statistically insignificant difference.

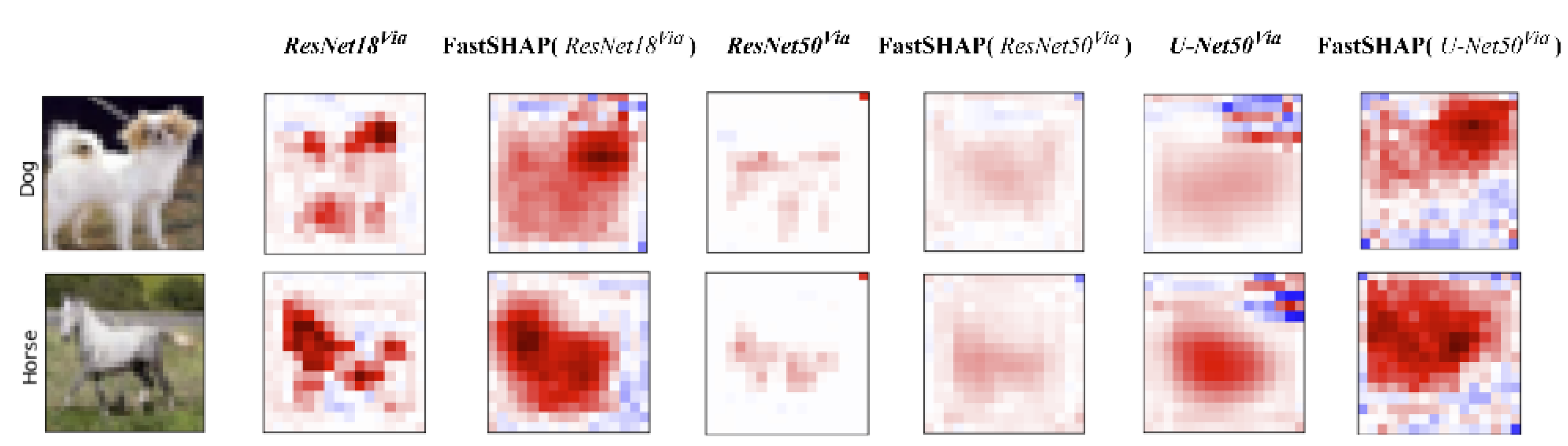
- ✓ ViaSHAP is evaluated on CIFAR-10 dataset using 3 implementations based on ResNet-18, ResNet-50, and UNet.

	AUC	0.95 Confidence Interval
$U\text{-}Net^{\mathcal{V}ia}$	**0.983**	**(0.981, 0.986)**
$ResNet18^{\mathcal{V}ia}$	0.968	(0.964, 0.971)
$ResNet50^{\mathcal{V}ia}$	0.96	(0.956, 0.964)

Evaluation of Explanations



As KernelSHAP refines its approximations with more samples, the similarity to ViaSHAP's values grows



The explanation of the predicted class using two random images from the CIFAR-10

References

1. Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljacic,Thomas Y. Hou, and Max Tegmark. Kan: Kolmogorov-arnold networks, 2024. URL https://arxiv.org/abs/2404.19756.

2. Neil Jethani, Mukund Sudarshan, Ian Connick Covert, Su-In Lee, and Rajesh Ranganath. FastSHAP: Real-time shapley value estimation. In International Conference on Learning Representations, 2022.