RICE UNIVERSITY
School of Engineering
Department of Computer Science

Accelerating Shapley Explanation via Contributive Cooperator Selection

Guanchu Wang*[1], Yu-Neng Chuang*[1], Mengnan Du[2], Fan Yang[1], Quan Zhou[3], Pushkar Tripathi[3], Xuanting Cai[3], and Xia Hu[1]

[1]Department of Computer Science, Rice University
[2]Department of Computer Science and Engineering, Texas A&M University.
[3]Meta Platforms, Inc.

Problem Statement

The Shapley value is a game theory-based interpretation for black-box model inference.

Problem of Calculating the Shapley value: NP-hard problem

$$\phi_i(f_v, \mathcal{U}) = \frac{1}{M} \sum_{\boldsymbol{S} \subseteq \mathcal{U} \setminus \{i\}} \binom{M-1}{|\boldsymbol{S}|}^{-1} \left[f_v(\{i\} \cup \boldsymbol{S}) - f_v(\boldsymbol{S}) \right].$$

2^M times of model evaluation

We focus on Low complexity estimation (acceleration) of the Shapley values.

[1] Wang, R. et. al. Shapley explanation networks. arXiv preprint arXiv:2104.02297, 2021.
[2] Lundberg, S. M. et. al. A unified approach to interpreting model predictions. NIPS, 2017.

High-level Idea of the Acceleration

Reduce the number of cooperators:

$$\phi_i(f_v, \mathcal{U}) = \frac{1}{M} \sum_{\boldsymbol{S} \subseteq \mathcal{U} \setminus \{i\}} \binom{M-1}{|\boldsymbol{S}|}^{-1} \left[f_v(\{i\} \cup \boldsymbol{S}) - f_v(\boldsymbol{S}) \right].$$

Brute force calculation regard all remainig features as the cooperators

Reduce cooperators

$$\hat{\phi}_i = \frac{1}{|\mathcal{S}_i|} \sum_{\boldsymbol{S} \subseteq \mathcal{S}_i} \binom{|\mathcal{S}_i|}{|\boldsymbol{S}|}^{-1} \left[f_v(\{i\} \cup \boldsymbol{S}) - f_v(\boldsymbol{S}) \right].$$

$|\mathcal{S}_i| = \log_2(N/2)$ such that ϕ_i needs N times of model evaluations

S_i denotes the contributive cooperators

Questions:

- How much error caused by reducing the coopertors?
- How to select the contributive cooperators to minimize the aboslute error.

$$\mathcal{S}_i = \arg \min_{|\mathcal{S}_i| = \log_2(N/2)} \sum_{i=1}^{M} |\hat{\phi}_i - \phi_i(f_v, \mathcal{U})|$$

Answer to Q1: The Shapley Chain Rule

Q1: The Estimation Error.

Shapley Chain Rule:

$$\phi_i(f_v, \mathcal{U}) = \phi_i(f_v, \mathcal{U} \setminus \{j\}) + \Delta_{i,j} + o_{i,j}$$

Contribution to $f_v(\mathcal{U})$

Contribution to $f_v(\mathcal{U} \setminus \{j\})$

Estimation error

Infinitesimal (ignored)

where $\Delta_{i,j} = (x_i - \bar{x}_i)(x_j - \bar{x}_j) \sum_{\boldsymbol{S} \subseteq \mathcal{U} \setminus \{i,j\}} \frac{\nabla^2_{i,j} f_v(\boldsymbol{S} \cup \{i,j\}) + \nabla^2_{j,i} f_v(\boldsymbol{S} \cup \{i,j\})}{2(M - |\boldsymbol{S}| - 1)\binom{M}{|\boldsymbol{S}|+1}}$

$\nabla^2_{i,j} f_v(\boldsymbol{S}) = \frac{\partial^2 f(\boldsymbol{x}_{\boldsymbol{S}}, \bar{\boldsymbol{x}}_{\mathcal{U} \setminus \boldsymbol{S}})}{\partial x_i \partial x_j}$ Cross gradient f features i and j

Answer to Q2: Contributive Cooperator Selection

Q2: Contributive cooperators selection: Minimize the upper bound of aboslute error

$$\mathcal{S}_i = \underset{\mathcal{S}\subseteq\mathcal{U}\setminus\{i\}}{\arg\min} |\phi_i(f_v, \mathcal{U}) - \phi_i(f_v, \mathcal{S}\cup\{i\})|,$$

$$= \underset{\substack{\mathcal{S}\subseteq\mathcal{U}\setminus\{i\} \\ |\mathcal{S}|=\log_2(N/2)}}{\arg\max} \sum_{j\in\mathcal{S}} \hat{\epsilon}_{i,j}|x_i - \bar{x}_i||x_j - \bar{x}_j|.$$

Contributive cooperators

$|\mathcal{S}_i| = \log_2(N/2)$ such that ϕ_i needs N times of model evaluations

where $\hat{\epsilon}_{i,j} = \frac{1}{4}|\nabla^2_{i,j} f_v(\mathcal{U}) + \nabla^2_{j,i} f_v(\mathcal{U})|$

The estimation of Shapley values:

$$\hat{\phi}_i = \frac{1}{|\mathcal{S}_i|+1} \sum_{\boldsymbol{S}\subseteq\mathcal{S}_i} \binom{|\mathcal{S}_i|}{|\boldsymbol{S}|}^{-1} \left[f_v(\{i\}\cup\boldsymbol{S}) - f_v(\boldsymbol{S})\right].$$

Estimated contribution of feature i

Only involve contributive cooperators

Experiment Results on the Cretio Dataset

Baseline methods:

- Kernel-SHAP and its improved version.
- Permutation Sampling and its improved version.

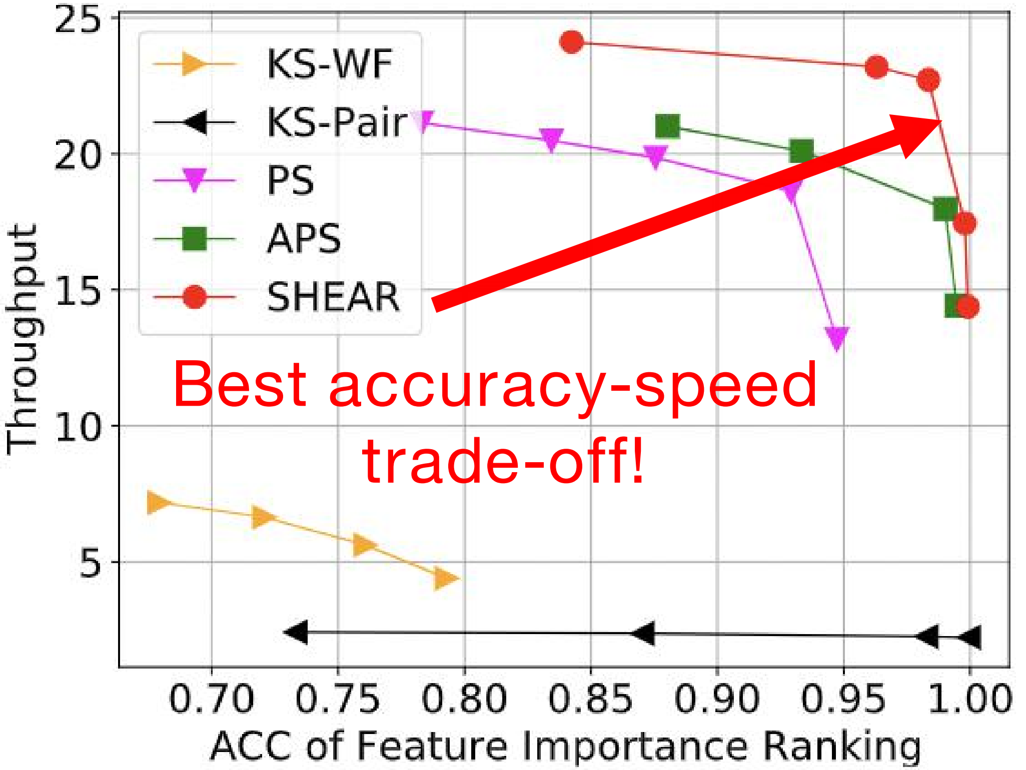
Evaulation metrics

- Accuracy of feature importance ranking:

$$\mathrm{ACC} = \frac{\sum_{m=1}^{M} \frac{\mathbf{1}_{\hat{r}_m = r_m}}{m}}{\sum_{m=1}^{M} \frac{1}{m}}$$

- Algorithmic throughput:

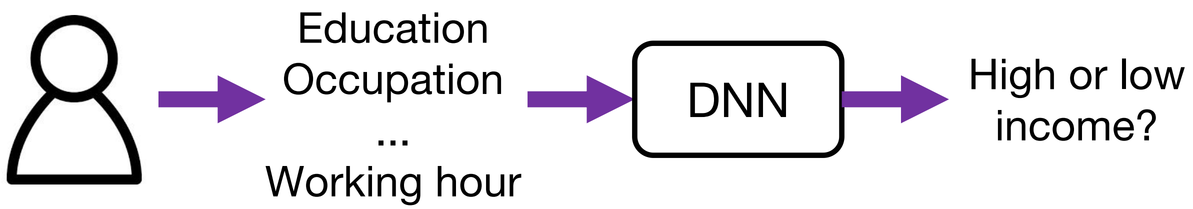
$$\mathrm{Throughput} = \frac{N_{\mathrm{test}}}{t_{\mathrm{total}}}$$



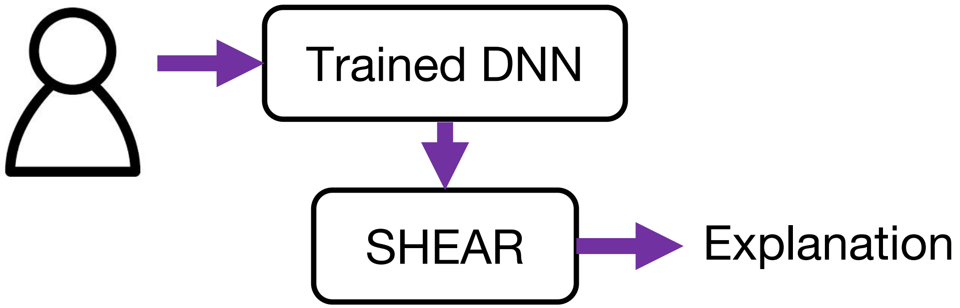
[1] Lundberg, S. M. et. al. A unified approach to interpreting model predictions. NIPS, 2017.
[2] Covert, I. et. al. Improving kernelshap: Practical shapley value estimation using linear regression. ICML, 2021
[3] Mitchell, R., et. al. Sampling permutations for shapley value estimation. arXiv preprint arXiv:2104.12199, 2021

Application of SHEAR

Application of SHEAR:



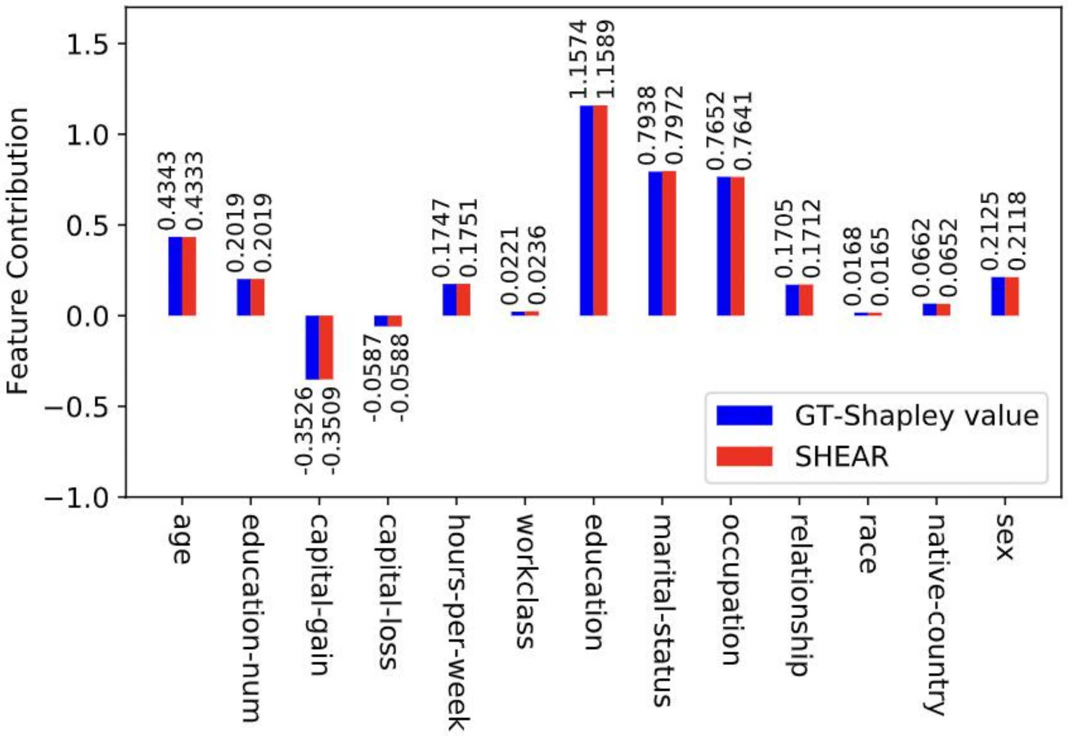
Train a DNN on the Census Income dataset.



Generate explanation of prediction for a sample.

Observations

- SHEAR gives very closed estimation to the ground-truth Shapley values.
- Top-three key features for the income prediction: education, martial-status and occupation.



Conclusion

Our work

- targets algorithmic acceleration of Shapley explanation (SHEAR).
- theoretically minimizes the upper bound of estimation error.
- experimentally demonstrate the effectiveness in terms of the accuacy and speed.

More

- Github: https://github.com/guanchuwang/SHEAR