SHAP-IQ: Unified Approximation of any-order Shapley Interactions

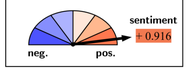
Fabian Fumagalli[1,*], Maximilian Muschalik[2,*], Patrick Kolpaczki[3],
Eyke Hüllermeier[2], and Barbara Hammer[1]

NEURAL INFORMATION PROCESSING SYSTEMS

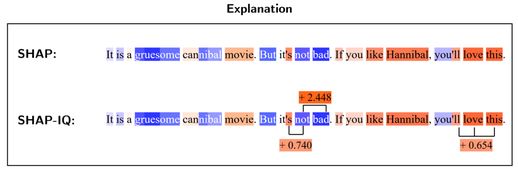
Motivation: Explaining Language Models

Sentiment Analysis Model

"It is a gruesome cannibal movie. But it's not bad. If you like Hannibal, you'll love this."



Explanation



Background

SHAP

Feature Set: $D := \{1, \dots, d\}$

Model Behavior: prediction given feature subsets

$$\nu : \mathcal{P}(D) \to \mathbb{R}$$

Marginal Contribution: Impact of single features

$$\delta_{\{i\}}(T) := \nu(T \cup \{i\}) - \nu(T)$$

Shapley Value [1]

$$I^{\text{SV}}(i) = \sum_{T \subseteq D \setminus \{i\}} \frac{(d-t-1)!t!}{d!} \delta^{\nu}_{\{i\}}(T)$$

average marginal contribution
unique attribution given axioms

Beyond SHAP: From Feature Attributions to Interactions

Discrete Derivatives

Idea: Recursively attribute residual contribution

$$\delta^{\nu}_{\{i,j\}}(T) = \nu(T \cup \{i,j\}) - \nu(T) - \delta^{\nu}_{\{i\}}(T) - \delta^{\nu}_{\{j\}}(T)$$

Marginal contributions for arbitrary groups

$$\delta^{\nu}_{S}(T) := \sum_{L \subseteq S} (-1)^{s-l} \nu(T \cup L)$$

Cardinal Interaction Index (CII)

$$I^{m}(S) := \sum_{T \subseteq D \setminus S} m_s(t) \delta^{\nu}_{S}(T)$$

average discrete derivatives
axiomatic extension for uniqueness **unclear**

Shapley Interactions

Shapley Interaction Index (SII) [2]

$$m^{\text{SII}}_s(t) := \frac{(d-t-s)!t!}{(d-s+1)!}$$

unique index with **recursive axiom does not fulfill efficiency**

Other CIIs with Efficiency

n-SII: n-Shapley Value [3] (aggregates SII values)

STI: Shapley Taylor Interaction Index [4] (efficiency and interaction distribution)

FSI: Faithful Shapley Interaction Index [5] (efficiency and faithfulness)

Approximations and Challenges

Exponential Complexity requires Approximation!

SII, STI: Permutation-based (PB) Approximation (Extension of ApproShapley [6])

FSI: Kernel-based (KB) Approximation (Extension of KernelSHAP [7])

Existing Approximations are limited!

No Unification: Approximations are index-specific!

Inefficient: PB updates estimates only selectively!

Unknown Guarantees: KB is hard to analyze!

References

[1] Shapley, L. S. (1953). A value for n-person games. In Contributions to the Theory of Games (AM-28), Volume II, pages 307–318. Princeton University Press.
[2] Fujimoto, K., Kojadinovic, I., and Marichal, J. (2006). Axiomatic characterizations of probabilistic and cardinal-probabilistic interaction indices. Games Econ. Behav., 55(1):72–99.
[3] Grabisch, M. and Roubens, M. (1999). An axiomatic approach to the concept of interaction among players in cooperative games. International Journal of Game Theory, 28(4):547–565
[4] Bordt, S. and von Luxburg, U. (2023). From shapley values to generalized additive models and back. In AISTATS'23, pp. 709–745.
[5] Sundararajan, M., Dhamdhere, K., and Agarwal, A. (2020). The shapley taylor interaction index. In ICML'20, pp. 9259–9268.
[6] Tsai, C., Yeh, C., and Ravikumar, P. (2023). Faith-shap: The faithful shapley interaction index. Journal of Machine Learning Research, 24(94):1–42.
[7] Lundberg, S. M. and Lee, S. (2017). A unified approach to interpreting model predictions. In NeurIPS'17, pp.4765–4774.
[8] Covert, I. and Lee, S.-I. (2021). Improving KernelSHAP: Practical shapley value estimation using linear regression. In AISTATS 2021, pp. 3457–3465.

Contribution

1. We consider a general form of interaction indices, known as CII [2] and establish a **novel representation**, which we utilize to construct **SHAP-IQ**, an efficient sampling-based estimator.
2. We show that SHAP-IQ is **unbiased**, **consistent** and provide a general **approximation bound** while maintaining the **efficiency** condition for n-SII [4] and STI [5].
3. For the Shapley value [1], we find a novel representation and prove that SHAP-IQ is linked to Unbiased KernelSHAP [8], greatly **simplifying its representation**.
4. We use SHAP-IQ to compute **any-order** n-SII values on different ML models and demonstrate that SHAP-IQ **outperforms** existing **baseline methods**.

SHAP-IQ: SHAPley Interaction Quantification

SHAP-IQ

Theorem 4.1: Novel Representation

$$I^m(S) = \sum_{T \subseteq D} \nu_0(T) \gamma^m_s(t, |T \cap S|)$$

with weights: $\gamma^m_s(t,k) := (-1)^{s-k} m_s(t-k)$

Definition 4.2

$$\hat{I}^m_{k_0}(S) = \text{Exact} + \text{Monte Carlo}$$

exact calculation for low- and high-cardinality subsets

sampling for remaining subsets

Theoretical Properties

Theorem 4.3

SHAP-IQ estimates are **unbiased**, **consistent** with a **finite sample deviation bound**

Theorem 4.7

SHAP-IQ estimates **maintain efficiency for n-SII and STI** and all **s-efficient indices**

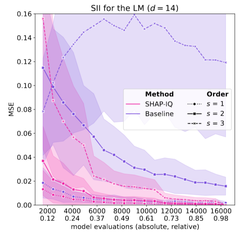
Theorems 4.4 and 4.5

Theorem 4.1 leads to a **novel Shapley value representation** and SHAP-IQ simplifies **Unbiased KernelSHAP** [8]

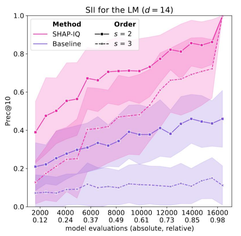
Empirical Evaluation

Approximation Quality: SHAP-IQ vs. SII Baseline

MSE



Precision at 10



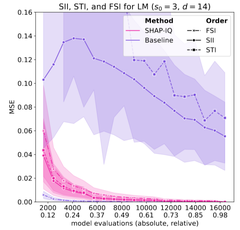
Setup

- **Task**: explanation of a transformer-based sentiment analysis model with *SII*
- **Model**: *DistilBERT* fine-tuned on *IMDB*
- **Data**: tokenized sentences with $d = 14$ words

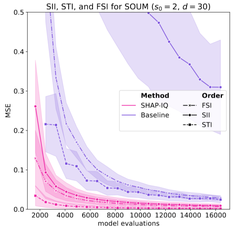
➢ SHAP-IQ substantially outperforms the permutation sampling baseline yielding higher-quality approximation results for the SII.

Approximation of different CIIs using SHAP-IQ

Language Model (LM)



Sum of Unanimity (SOUM)



Setup

- **Indices**: SII and STI are estimated with permutation sampling and FSI with a regression
- **LM**: sentiment analysis model
- **SOUM**: synthetic model after [6] with strong interactions

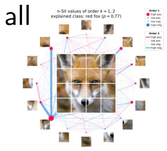
➢ SHAP-IQ efficiently estimates all types of CIIs, but the FSI regression estimator on the LM is superior to SHAP-IQ, showcasing the power of the weighted least squares representation.

Open Source Implementation

- **shapiq** is available for python via **pip**

```
pip install shapiq
```

- **compute** interactions for all CII indices with **available estimators**
- **plot interactions**



Institutions

1) Bielefeld University, Bielefeld, Germany
2) LMU Munich, Munich, Germany
3) Paderborn University, Paderborn, Germany

Funding

Ministry of Culture and Science of the State of North Rhine-Westphalia

Funded by DFG Deutsche Forschungsgemeinschaft German Research Foundation

Paper Site