TRR 318 CONSTRUCTING EXPLAINABILITY

UNIVERSITÄT BIELEFELD

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

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]

✉ ffumagalli@techfak.uni-bielefeld.de
✉ maximilian.muschalik@lmu.de

[1] Bielefeld University, [2] LMU Munich,
[3] Paderborn University, * denotes equal contribution

NEURAL INFORMATION PROCESSING SYSTEMS

Collaboration

Meet us at the conference!

Fabian [1,*]
Fumagalli

Maximilian [2,*]
Muschalik

Patrick [3]
Kolpaczki

Eyke [2]
Hüllermeier

Barbara [1]
Hammer

(1) UNIVERSITÄT BIELEFELD

(2) LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

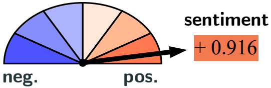
(3) UNIVERSITÄT PADERBORN

* denotes equal contribution

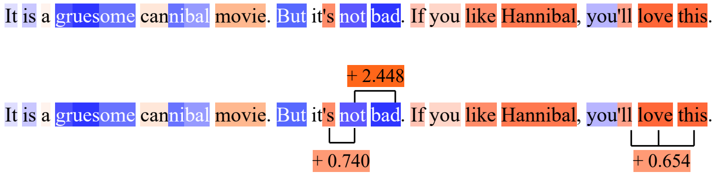
Motivation: Explaining Language Models

Sentiment Analysis Model

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



Explanation



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

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

Background – Shapley Interactions

Shapley interactions are defined as different indices

- **Shapley Interaction Index (SII)** (Grabisch and Roubens, 1999)
- **n-Shapley Values (n-SII)** (Bordt and von Luxburg, 2023)
- **Shapley Taylor Interaction Index (STI)** (Sundararajan et al., 2020)
- **Faithful Shapley Interaction Index (FSI)** (Tsai et al., 2023)

Cardinal Interaction Index (CII) subsumes all indices above (Grabisch and Roubens, 1999)

A broad class of interaction indices, including **all indices** that satisfy the (generalized) *linearity*, *symmetry* and *dummy* axioms:

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

weight depending on subset size

discrete derivatives: marginal interaction of S in the presence of T (Grabisch 2000)

Background – Existing Approximations

Problem: Existing Approximations are limited!

- **No unification:** Methods are index-specific
  - **SII and STI:** Permutation-based (PB) extends ApproShapley (Castro et al., 2009)
  - **FSI:** Kernel-based (KB) estimation extends KernelSHAP (Lundberg and Lee, 2017)
- **Inefficient:** PB approximation updates estimates only selectively
- **Unknown Guarantees:** Analyzing KB approximation remains challenging

Solution: **SHAP-IQ** as a universal approximator of general interaction indices!

- Based on the broad class of CIIs
- Updates estimates efficiently
- Supported by theoretical guarantees

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

- We provide a **novel representation of CIIs** which does **not** depend on δ_S^ν:

Theorem 4.1 (Novel Representation)

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

weight depends on
subset sizes and **used CII**

value function:
output of the game
(e.g. model) provided
only subset T

- We construct **SHAP-IQ**, an efficient **sampling-based estimator**:

Definition 4.2 (SHAP-IQ: Shapley Interaction Quantification)

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

exact calculation for low- and high-cardinality subsets

sampling for
remaining subsets

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

SHAP-IQ estimates

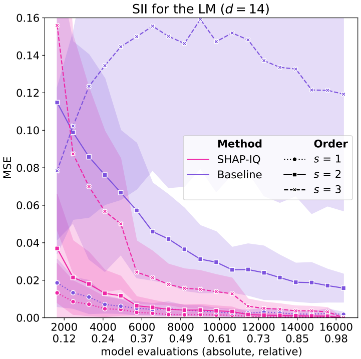
- are **unbiased and consistent** (Theorem 4.3)
- satisfy a **finite sample deviation bound** (Theorem 4.3)
- maintain **efficiency** for **n-SII and STI** (Theorem 4.7)

SHAP-IQ yields implications for the Shapley Value:

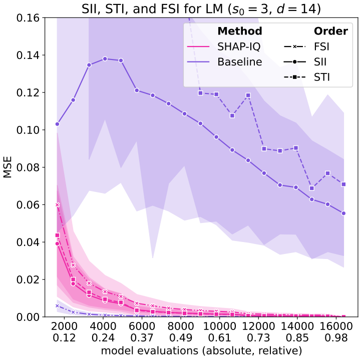
- A novel representation of the Shapley value (Theorem 4.4)
- SHAP-IQ is Unbiased KernelSHAP (Covert and Lee, 2021) (Theorem 4.5)
  - A greatly simplified representation of Unbiased KernelSHAP

Approximation Quality of SHAP-IQ compared to Baselines

SII



CIIs: SII, STI, FSI



Setup

Task: explanation of a transformer-based sentiment analysis model with the *CIIs*

Model: *DistilBERT* fine-tuned on *IMDB*

Data: tokenized sentences with $d = 14$ tokens

- SHAP-IQ efficiently and consistently estimates all types of CIIs and substantially outperforms the permutation sampling baseline for SII and STI.

The Road Ahead and Open Source Implementation

Interpretation of Shapley Interactions

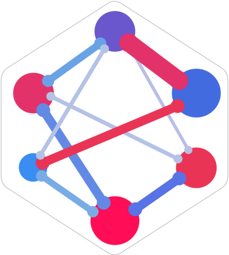
- An interaction is the joint effect of a group of features
- SHAP-IQ estimates are the (average) contribution of the interaction to the prediction.

Get in touch with us!

📍 Great Hall & Hall B1+B2

📅 **Wednesday** 12/13/2023
5:45 p.m. – 7:45 p.m.

Implementation



SHAP-IQ

- **Install**: pip install shapiq
- **Design**: if you are familiar with shap you should feel right at home
- ▶ **Join**: looking for collaborations!

NEURAL INFORMATION PROCESSING SYSTEMS

References

- Bordt, S., & von Luxburg, U. (2023).From shapley values to generalized additive models and back. *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics, (AISTATS 2023)*, *206*, 709–745.
- Castro, J., Gómez, D., & Tejada, J. (2009).Polynomial calculation of the shapley value based on sampling. *Computers & Operations Research*, *36*(5), 1726–1730. https://doi.org/10.1016/j.cor.2008.04.004
- Covert, I., & Lee, S.-I. (2021).Improving kernelshap: Practical shapley value estimation using linear regression. *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics, (AISTATS 2021)*, *130*, 3457–3465.
- Grabisch, M., Marichal, J., & Roubens, M. (2000).Equivalent representations of set functions. *Mathematics of Operations Research*, *25*(2), 157–178. https://doi.org/10.1287/moor.25.2.157.12225

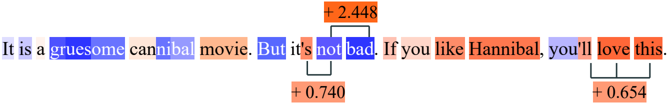
References

- Grabisch, M., & 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. https://doi.org/10.1007/s001820050125
- Lundberg, S. M., & Lee, S. (2017).A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, (NeurIPS 2017)*, 4765–4774.
- Sundararajan, M., Dhamdhere, K., & Agarwal, A. (2020).The shapley taylor interaction index. *Proceedings of the 37th International Conference on Machine Learning, (ICML 2020)*, *119*, 9259–9268.
- Tsai, C., Yeh, C., & Ravikumar, P. (2023).Faith-shap: The faithful shapley interaction index. *Journal of Machine Learning Research*, *24*(94), 1–42.

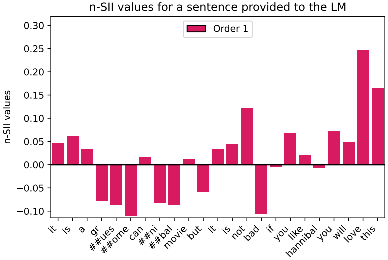
Example Use Case: Estimation of n-SII Values

id: 7567
$d = 23$

+ 0.916
(positive sentiment)

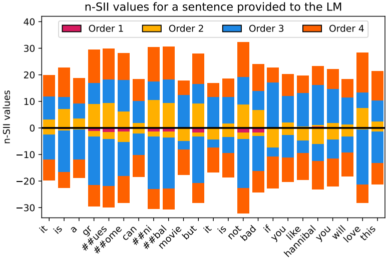


It is a gruesome cannibal movie. But it's not bad. If you like Hannibal, you'll love this.



Shapley Value
order 1

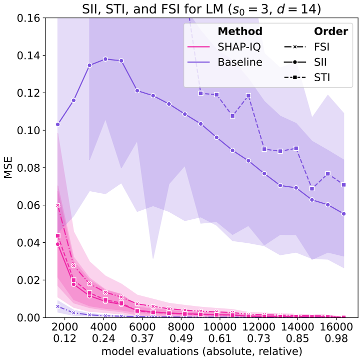
...



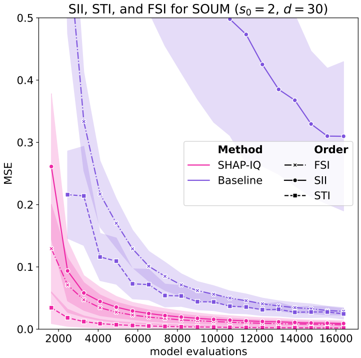
n-SII
up to order 4

Approximation of different CIIs using SHAP-IQ

Language Model (LM)



Sum of Unanimity (SOUM)



Setup

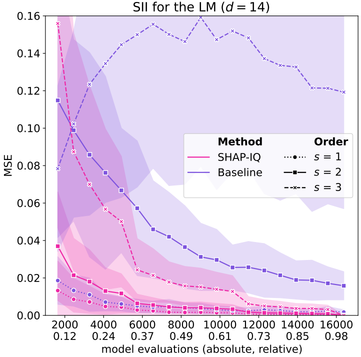
- **Indices**: SII and STI are estimated with permutation sampling and FSI with a regression
- **LM**: sentiment analysis model
- **SOUM**: synthetic model with strong interactions

- SHAP-IQ efficiently and consistently estimates all types of CIIs.
- The FSI regression estimator on the LM is superior to SHAP-IQ.

Approximation Quality of SHAP-IQ and the SII Baseline

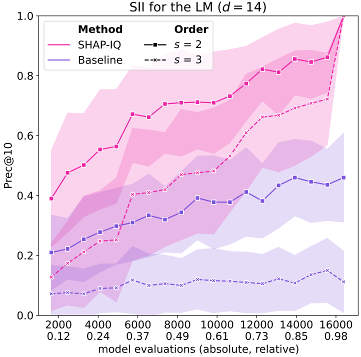
MSE

(lower is better)



Precision at 10

(higher is better)



Setup

- **Task**: explanation of a transformer-based sentiment analysis model with the *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.