Rethinking Shapley value for Negative Interactions in Non-convex Games

Wonjoon Chang, Myeonjin Lee, Jaesik Choi

Korea Advanced Institute of Science and Technology (KAIST), South Korea

{one_jj, lmjk311, jaesik.choi}@kaist.ac.kr

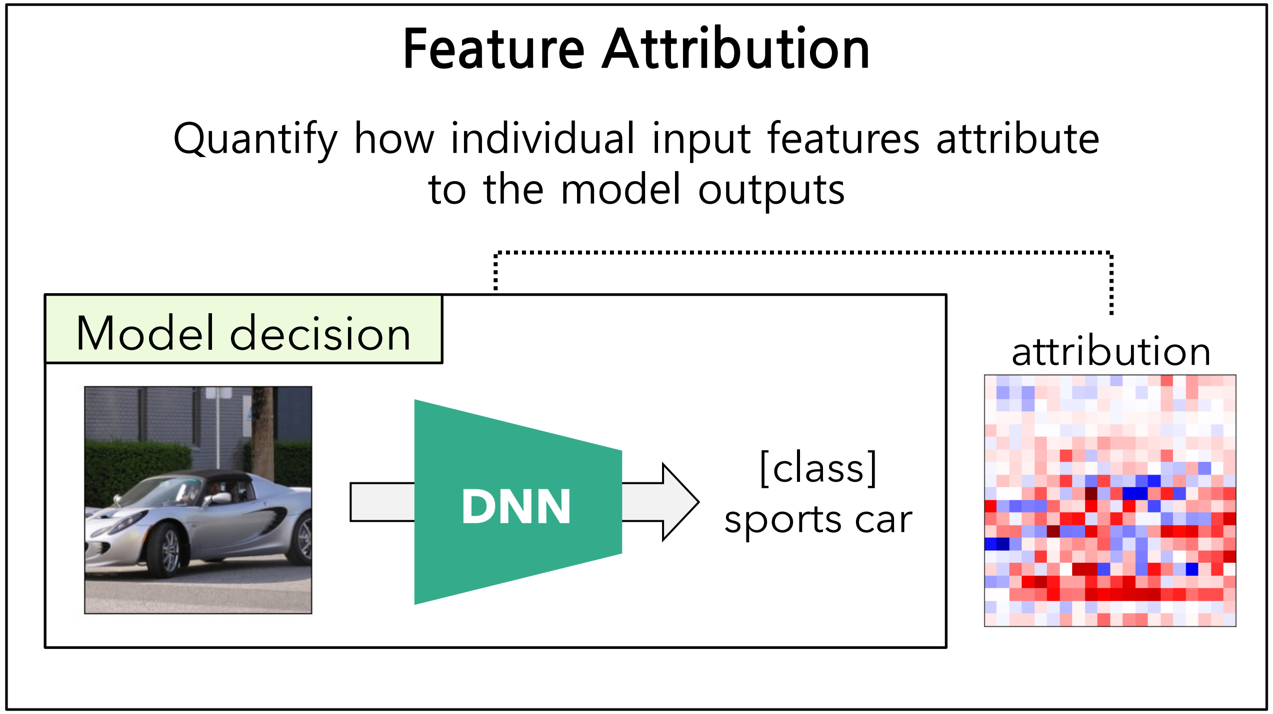
ICLR
International Conference On Learning Representations

KAIST

Motivation

Model Interpretability & Reliability

- In black-box models, it is crucial to understand the cause of the model decision.



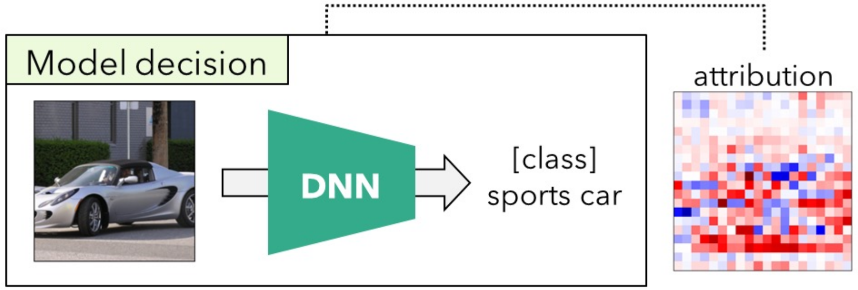
Motivation

Feature Attribution & Shapley value

- Theoretically, most feature attributions are grounded in the Shapley value.

Feature Attribution

Quantify how individual input features attribute to the model outputs



Shapley value

- An axiom-based solution in cooperative games.
- The Shapley value $\phi_i(v)$ calculates the average change in the model output $v(\cdot)$ according to the participation of the target feature i.

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{1}{n} \binom{n-1}{s}^{-1} [v(S \cup \{i\}) - v(S)]$$

- v : game (or model output)
- N : a set of the entire players (or features)
- n, s : the size of N, S

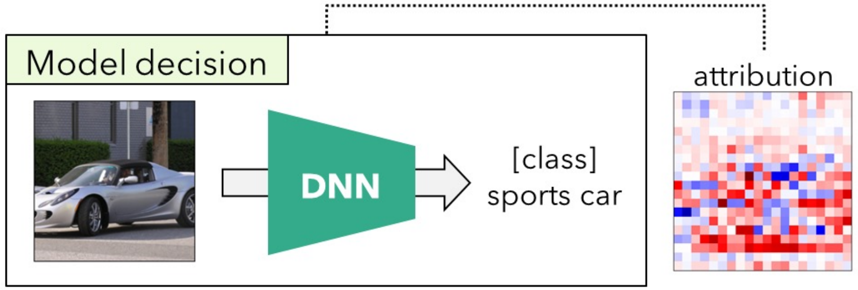
Motivation

Feature Attribution & Shapley value

- Theoretically, most feature attributions are grounded in the **Shapley value**.

Feature Attribution

Quantify how individual input features attribute to the model outputs



Shapley value

- An axiom-based solution in cooperative games.
- The Shapley value $\phi_i(v)$ calculates the average change in the model output $v(\cdot)$ according to the participation of the target feature i.

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{1}{n} \binom{n-1}{s}^{-1} [v(S \cup \{i\}) - v(S)]$$

Research Question

- Is the Shapley value suitable for evaluating feature attribution in complex black-box models?
- How does the Shapley value account for **interactions** between features?

Interactions in Shapley value

- The Shapley value can be viewed as the expectation of **causal effects**.

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{1}{n} \binom{n-1}{s}^{-1} \Delta_i v(S) = \sum_{S \subseteq N \setminus \{i\}} \frac{1}{n} \binom{n-1}{s}^{-1} [v(S \cup \{i\}) - v(S)]$$

Effect / Contribution

$$\Delta_i v(T) = v(T \cup \{i\}) - v(T)$$

Interaction

$$I_{ij}(T) = \Delta_{ij} v(T) = \Delta_i v(T \cup \{j\}) - \Delta_i v(T)$$
$$= v(T \cup \{i, j\}) - v(T \cup \{i\}) - v(T \cup \{j\}) + v(T)$$

Interactions in Shapley value

- The Shapley value can be viewed as the expectation of **causal effects**.

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{1}{n} \binom{n-1}{s}^{-1} \Delta_i v(S) = \sum_{S \subseteq N \setminus \{i\}} \frac{1}{n} \binom{n-1}{s}^{-1} [v(S \cup \{i\}) - v(S)]$$

Effect / Contribution

$$\Delta_i v(T) = v(T \cup \{i\}) - v(T)$$

Interaction

$$I_{ij}(T) = \Delta_{ij} v(T) = \Delta_i v(T \cup \{j\}) - \Delta_i v(T)$$
$$= v(T \cup \{i,j\}) - v(T \cup \{i\}) - v(T \cup \{j\}) + v(T)$$

Theorem 1. *The Shapley value is a weighted sum of interactions :*

$$\phi_i(v) = \Delta_i v(\emptyset) + \sum_{t=0}^{n-2} \frac{1}{n} \binom{n-1}{t}^{-1} \sum_{\substack{j \in N \\ j \neq i}} \sum_{\substack{T \subseteq N \setminus \{i,j\} \\ |T|=t}} I_{ij}(T)$$

Interactions in Shapley value

Theorem 1. *The Shapley value is a weighted sum of interactions :*

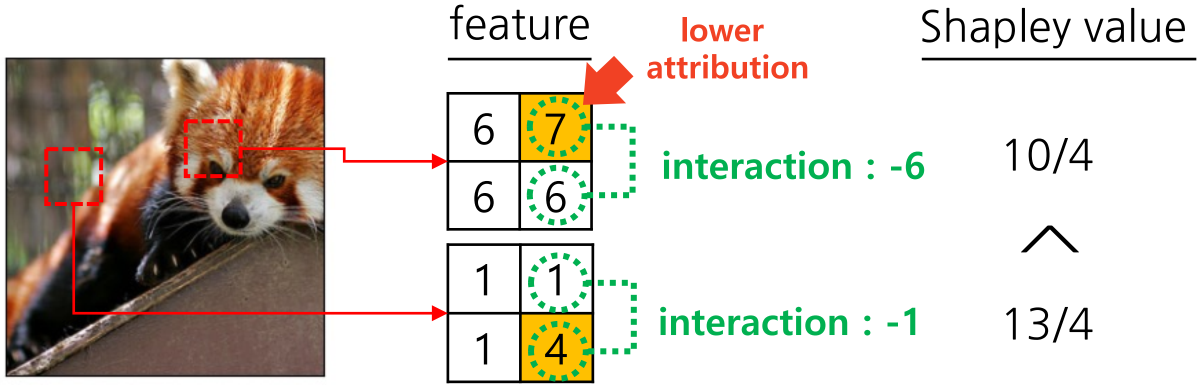
$$\phi_i(v) = \Delta_i v(\emptyset) + \sum_{t=0}^{n-2} \frac{1}{n} \binom{n-1}{t}^{-1} \sum_{\substack{j \in N \\ j \neq i}} \sum_{\substack{T \subseteq N \setminus \{i,j\} \\ |T| = t}} I_{ij}(T)$$

General Assumption in Cooperative Games

- The players cooperate with each other to maximize their payoffs.
- It does not holds in non-convex games, where negative interactions arise $(I_{ij}(T) < 0)$.

→ undervaluation issue

e.g. max pooling
output
= max(6,7,6,6) + max(1,1,1,4)
= 7 + 4 = 11



Aggregated Positive Interactions

- We aim to suggest a new solution that:
  - avoids the attribution undervaluation in non-convex games;
  - follows the Shapley value in convex games.
- **Aggregated Positive Interactions (API)**
  - We propose a new solution that decomposes each contribution into interactions and aggregates the positive parts, which represents the player's potential influence on improving the game output.

$$\phi_i(v) = \Delta_i v(\emptyset) + \sum_{t=0}^{n-2} \frac{1}{n} \binom{n-1}{t}^{-1} \sum_{\substack{j \in N \\ j \neq i}} \sum_{\substack{T \subseteq N \setminus \{i,j\} \\ |T| = t}} \max(I_{ij}(T), 0)$$

Aggregated Positive Interactions

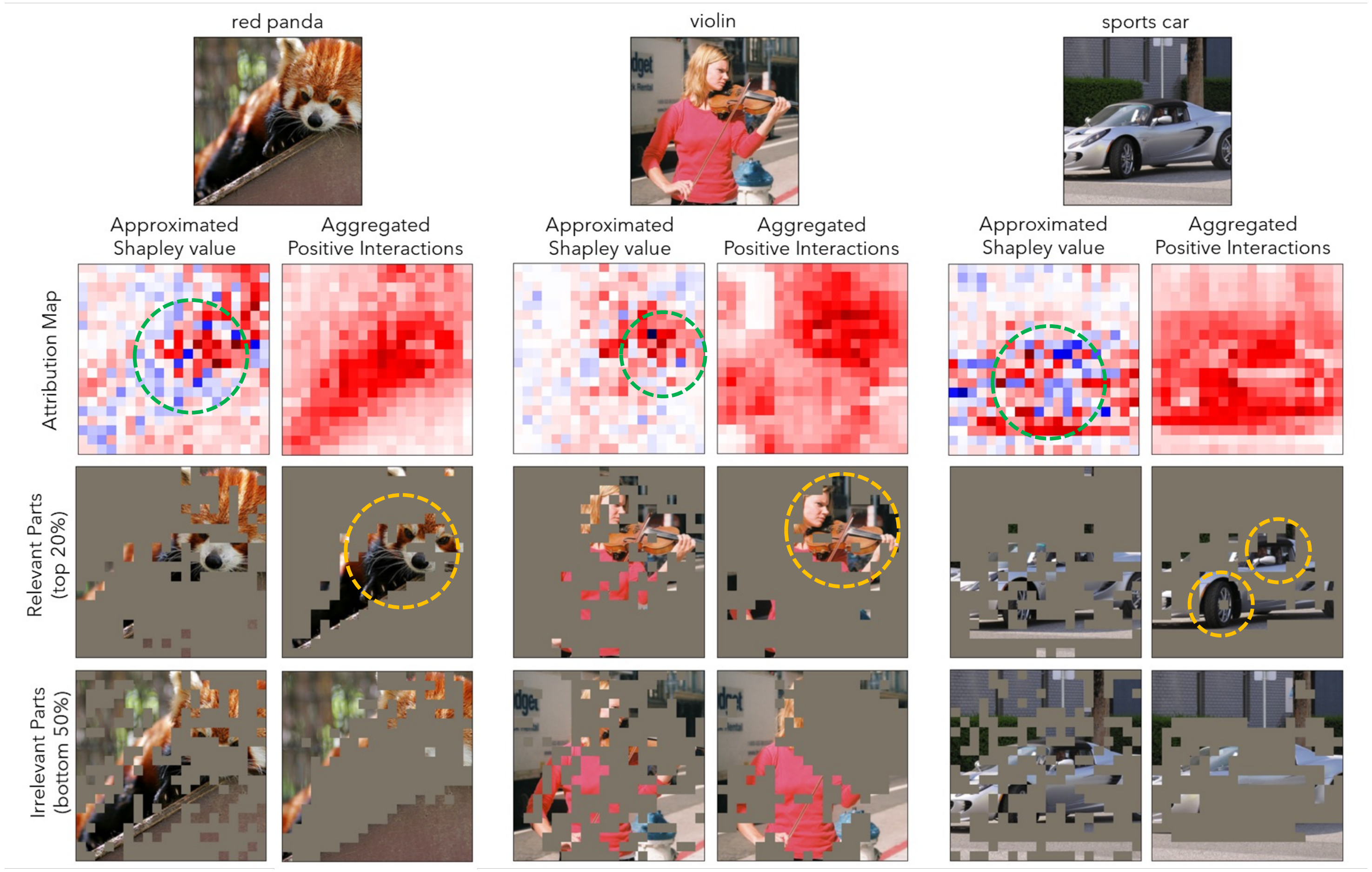
- We aim to suggest a new solution that:
  - avoids the attribution undervaluation in non-convex games;
  - follows the Shapley value in convex games.
- **Aggregated Positive Interactions (API)**
  - We propose a new solution that decomposes each contribution into interactions and aggregates the positive parts, which represents the player's potential influence on improving the game output.

$$\phi_i(v) = \Delta_i v(\emptyset) + \sum_{t=0}^{n-2} \frac{1}{n} \binom{n-1}{t}^{-1} \sum_{\substack{j \in N \\ j \neq i}} \sum_{\substack{T \subseteq N \setminus \{i,j\} \\ |T|=t}} \max(I_{ij}(T), 0)$$

- To apply it for complex black-box models, we additionally provide
  - an unbiased estimator using permutation sampling (Corollary 1);
  - an approximation algorithm based on backpropagation (Algorithm 1).

Applications

- Compare the results of summing all interactions, as done in the original Shapley value, with the API results (20x20 patches).

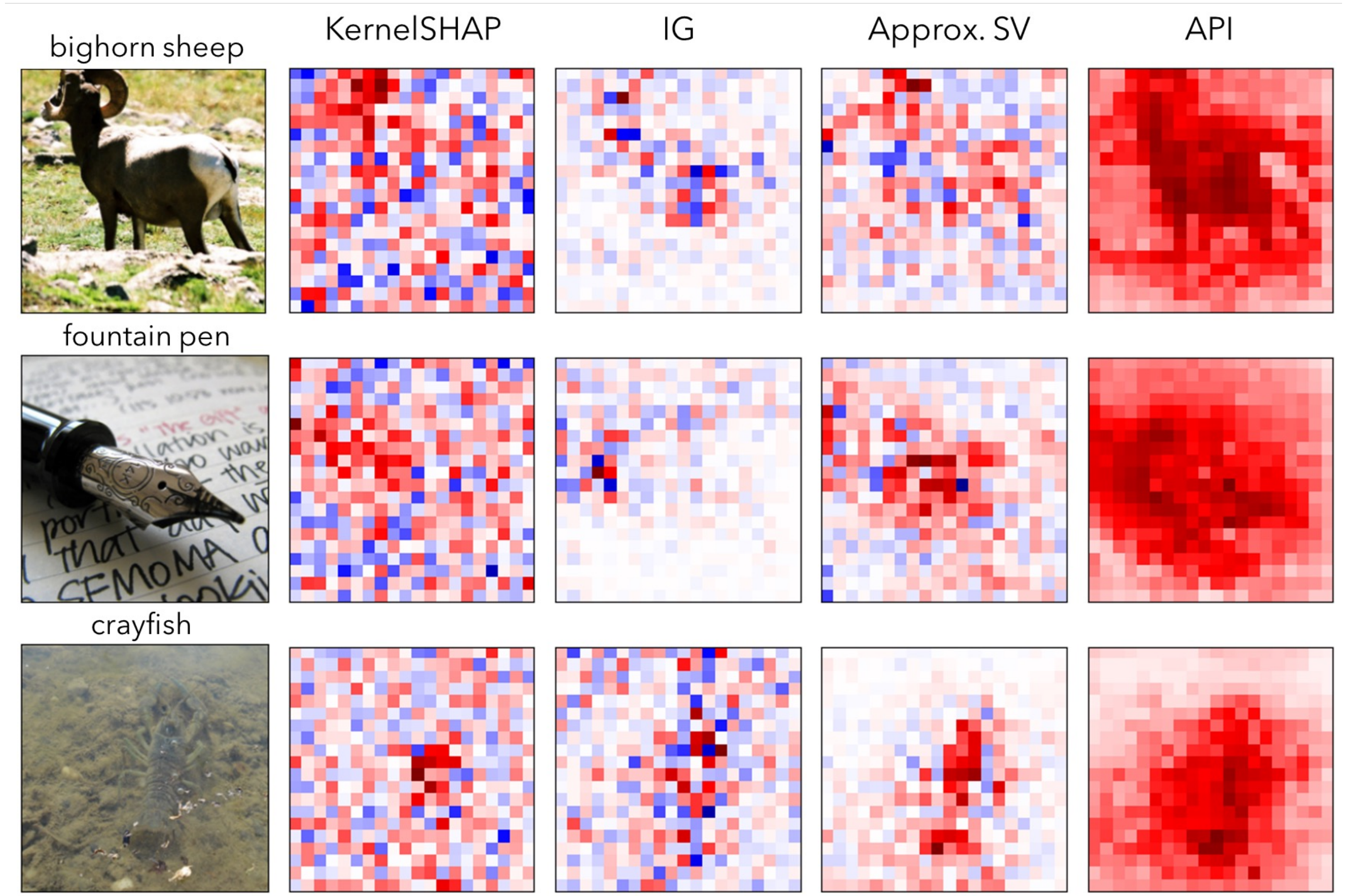


Inconsistency when negative interactions are included

more concentrated attributions by API

Applications

- Comparison with other attribution methods



Thank you!

Presenter: Wonjoon Chang

SAILab, KAIST AI

one_jj@kaist.ac.kr