# SHAP values: a game theory tool towards model interpretability

Cristian Curaba

February 16, 2024

# Structure of the presentation

Cooperative game theory and **Shapley values**;

Unified approach for model explanation: **additive feature attribution method**

**SHAP VALUES**

Examples:

- Linear CNN
- CNN with ReLU

# Cooperative Game Theory Definition

Let $v\colon \mathcal{P}(N) \to \mathbb{R}$, with $v(\emptyset) = 0$ be the **coalitional game**, where $v(S)$ is the expected payoff sum with members of $S$ cooperation. The **Shapley value** of a player in a coalition game is defined as follows:

$$\varphi_i(v) = \frac{1}{n} \sum_{S : S \subseteq N \setminus \{i\}} \binom{n-1}{|S|}^{-1} (v(S \cup \{i\}) - v(S)),$$

where $n$ is the number of players.

# Example: glove game.

Goal: create the maximum number of paired gloves.
Simple case: let's consider three players $N = \{1, 2, 3\}$ with one right glove for 1 and 2 and a left glove for player 3.

# Example: glove game.

Goal: create the maximum number of paired gloves.
Simple case: let's consider three players $N = \{1, 2, 3\}$ with one right glove for 1 and 2 and a left glove for player 3.

$$v(S) = \begin{cases} 1 \text{ if } S \in \{\{1,3\}; \{2,3\}; \{1,2,3\}\} \\ 0 \text{ otherwise.} \end{cases}$$

| $N \setminus \{1\}$ | $v(S \cup \{1\})$ |
|---|---|
| $\emptyset$ | 0 |
| $\{2\}$ | 0 |
| $\{3\}$ | 1 |
| $\{2,3\}$ | 1 |

Table: Marginals contribution of 1

| $N \setminus \{3\}$ | $v(S \cup \{3\})$ |
|---|---|
| $\emptyset$ | 0 |
| $\{1\}$ | 1 |
| $\{2\}$ | 1 |
| $\{1,2\}$ | 1 |

Table: Marginals contribution of 3

# Glove game: Shapley values

| $N \setminus \{1\}$ | $v(S \cup \{1\})$ |
|---|---|
| $\emptyset$ | 0 |
| $\{2\}$ | 0 |
| $\{3\}$ | 1 |
| $\{2, 3\}$ | 1 |

Table: Marginals contribution of 1

| $N \setminus \{3\}$ | $v(S \cup \{3\})$ |
|---|---|
| $\emptyset$ | 0 |
| $\{1\}$ | 1 |
| $\{2\}$ | 1 |
| $\{1, 2\}$ | 1 |

Table: Marginals contribution of 3

$$\varphi_i(v) = \frac{1}{n} \sum_{S : S \subseteq N \setminus \{i\}} \frac{v(S \cup \{i\}) - v(S)}{\binom{n-1}{|S|}},$$

# Glove game: Shapley values

| $N \setminus \{1\}$ | $v(S \cup \{1\})$ |
|---|---|
| $\emptyset$ | 0 |
| $\{2\}$ | 0 |
| $\{3\}$ | 1 |
| $\{2,3\}$ | 1 |

Table: Marginals contribution of 1

| $N \setminus \{3\}$ | $v(S \cup \{3\})$ |
|---|---|
| $\emptyset$ | 0 |
| $\{1\}$ | 1 |
| $\{2\}$ | 1 |
| $\{1,2\}$ | 1 |

Table: Marginals contribution of 3

$$\varphi_i(v) = \frac{1}{n} \sum_{S : S \subseteq N \setminus \{i\}} \frac{v(S \cup \{i\}) - v(S)}{\binom{n-1}{|S|}},$$

$$\varphi_1(v) = \frac{1}{3}\left(0 + 0 + \frac{1}{\binom{2}{1}} + \frac{1-1}{\binom{2}{2}}\right) = \frac{1}{6} = \varphi_2(v)$$

$$\varphi_3(v) = \frac{1}{3}\left(0 + \frac{1}{\binom{2}{1}} + \frac{1}{\binom{2}{1}} + \frac{1}{\binom{2}{2}}\right) = \frac{2}{3}$$

# Proprieties

- **Efficiency:**

$$\sum_{i=1}^{n} \varphi_i(v) = v(N).$$

- **Symmetry:**

If $\forall S \subseteq N \setminus \{i,j\}\ v(S \cup \{i\}) = v(S \cup \{j\})$ then $\varphi_i(v) = \varphi_j(v)$.

- **Dummy Player (Null Player):**

$$\text{If } v(S \cup \{i\}) = v(S) \text{ for all } S \subseteq N \setminus \{i\}, \text{ then } \varphi_i(v) = 0.$$

- **Linearity:**

$$\text{If } v = \alpha v_1 + \beta v_2, \text{ then } \varphi_i(v) = \alpha\varphi_i(v_1) + \beta\varphi_i(v_2).$$

# Explanation model: Additive feature attribution method

We focus on **local methods** designed to explain a prediction $f(x)$ based on a single input $x \in \mathbb{X}$. Let $\tilde{\mathbb{X}}$ be the **feature set** of $\mathbb{X}$.

- $f : \mathbb{X} \to \mathbb{Y}$ be the model prediction;
- $g : B(\tilde{x}, \varepsilon) \to \mathbb{Y}$ the local explanation model;
- $h_x : \tilde{\mathbb{X}} \to \mathbb{X}$ with $h_x(\tilde{x}) = x$ maps features into data.
- Desirable propriety: $g(\tilde{z}) \approx f(h_x(\tilde{z}))$ whenever $\tilde{z} \in B(\tilde{x}, \epsilon)$.

# Explanation model: Additive feature attribution method

We focus on **local methods** designed to explain a prediction $f(x)$ based on a single input $x \in \mathbb{X}$. Let $\tilde{\mathbb{X}}$ be the **feature set** of $\mathbb{X}$.

- $f : \mathbb{X} \to \mathbb{Y}$ be the model prediction;
- $g : B(\tilde{x}, \varepsilon) \to \mathbb{Y}$ the local explanation model;
- $h_x : \tilde{\mathbb{X}} \to \mathbb{X}$ with $h_x(\tilde{x}) = x$ maps features into data.
- Desirable propriety: $g(\tilde{z}) \approx f(h_x(\tilde{z}))$ whenever $\tilde{z} \in B(\tilde{x}, \epsilon)$.

## Definition

**Additive feature attribution methods** have an **explanation model** that is a linear function of binary variables:

$$g(\tilde{z}) = \phi_0 + \sum_{i=1}^{M} \phi_i \tilde{z}_i$$

where $\tilde{z} \in \{0, 1\}^M$, $M$ is the number of features and $\phi_i \in \mathbb{R}$.

# Example of additive feature attribution: LIME

Local Interpretable Model-agnostic Explanations:

- Select an Instance;
- Generate Perturbations;
- Prediction: each perturbed instance is passed through the black-box model;
- Build a Local Surrogate Linear Model to approximate the black-box model behaviour;
- Interpreting the surrogate model.

# Desirable proprieties of Additive Feature Attributions

- **Local accuracy**: $f(x) = g(\tilde{x}) = \phi_0 + \sum_{i=1}^{M} \phi_i \tilde{x}_i$.
- **Missingness**: features where $\tilde{x}_i = 0$ have no attributed impact:

$$\tilde{x}_i = 0 \Longrightarrow \phi_i = 0$$

- **Consistency**: Let $f_x(\tilde{z}) := f(h_x(\tilde{z}))$ with $\tilde{z} \in B(\tilde{x}, \varepsilon)$ and $\tilde{z}\backslash i$ denote setting $\tilde{z}_i = 0$. For any two models $f$ and $f'$, if

$$\forall \tilde{z} \in \{0,1\}^M \quad f'_x(\tilde{z}) - f'_x(\tilde{z}\backslash i) \geq f_x(\tilde{z}) - f_x(\tilde{z}\backslash i),$$

then

$$\phi_i(f', x) \geq \phi_i(f, x).$$

# Desirable proprieties of Additive Feature Attributions

- **Local accuracy**: $f(x) = g(\tilde{x}) = \phi_0 + \sum_{i=1}^{M} \phi_i \tilde{x}_i$.
- **Missingness**: features where $\tilde{x}_i = 0$ have no attributed impact:
$$\tilde{x}_i = 0 \Longrightarrow \phi_i = 0$$
- **Consistency**: Let $f_x(\tilde{z}) := f(h_x(\tilde{z}))$ with $\tilde{z} \in B(\tilde{x}, \varepsilon)$ and $\tilde{z}\backslash i$ denote setting $\tilde{z}_i = 0$. For any two models $f$ and $f'$, if
$$\forall \tilde{z} \in \{0,1\}^M \quad f'_x(\tilde{z}) - f'_x(\tilde{z}\backslash i) \geq f_x(\tilde{z}) - f_x(\tilde{z}\backslash i),$$
then
$$\phi_i(f', x) \geq \phi_i(f, x).$$

Good news: the only possible additive feature attribution method with an explanation model satisfying the above proprieties is given by Shapley values!

# From Shapley to SHAP values

Shapley values are adapted for **feature attribution**.

| **Shapley Value** | **SHAP value** |
|---|---|
| A coalition game | model prediction $f(x)$, with $x$ fixed |
| A Player | An entry of input $x$ (data feature) |
| Player contribution | Feature contribution over a prediction |

# SHAP (SHapley Additive exPlanation) Values

SHAP summarized: **In additive feature attribution, the explanation model is constructed with Shapley values. Marginal contributions are conditional expectations.**

# SHAP (SHapley Additive exPlanation) Values

SHAP summarized: **In additive feature attribution, the explanation model is constructed with Shapley values. Marginal contributions are conditional expectations.**
Let $f$ be the prediction model, we want to find the explanation $f(x) = g(\tilde{x})$ where

$$g(\tilde{z}) = \phi_0^f + \sum_{i=1}^{M} \phi_i^f \tilde{z}_i, \qquad \tilde{z} \in \{0,1\}^M.$$

We approximate $f(h_x(\tilde{z}))$ with $\mathbb{E}[f(z)|z_S]$ where $z_S$ has missing values for features not in $S$.

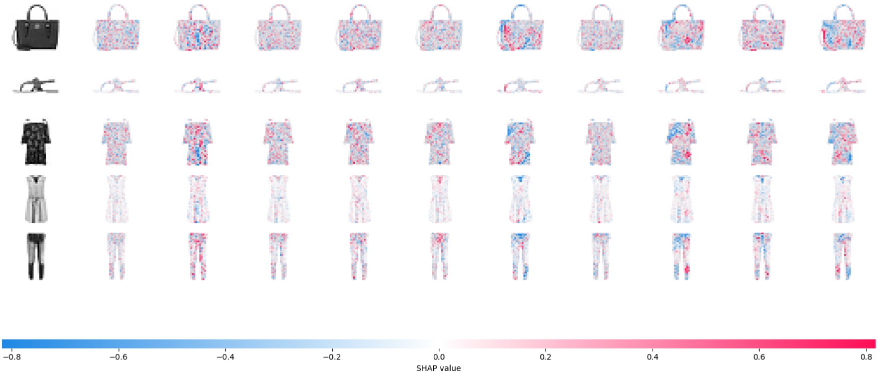
$$\phi_i^f(x) = \sum_{\tilde{z} \subseteq \tilde{x}} \frac{|\tilde{z}|!(M - |\tilde{z}| - 1)!}{M!} (\mathbb{E}[f(z)|z_S] - \mathbb{E}[f(z)|z_{S \setminus i}]),$$

where $S$ is the set of non-zero indexes in $z'$.

# Approximate $\mathbb{E}[f(z)|z_S]$: Deep SHAP method

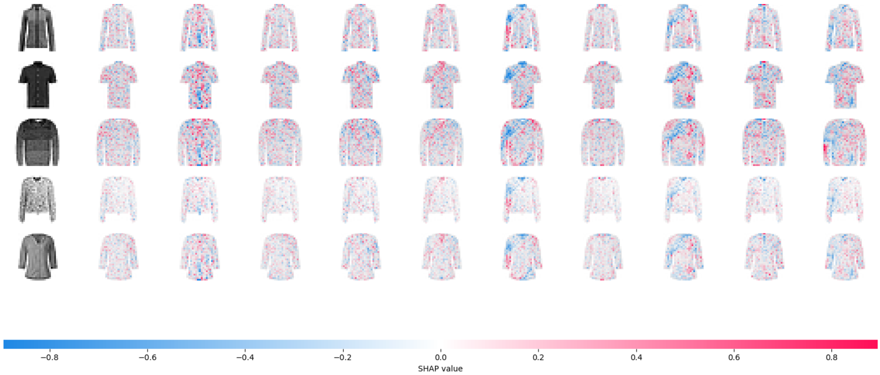
- Approximate the conditional expectations of SHAP values using a selection of **background samples**.
- It exploits the compositional nature of deep networks.

# CNN linear - Correctly Classified - White Background

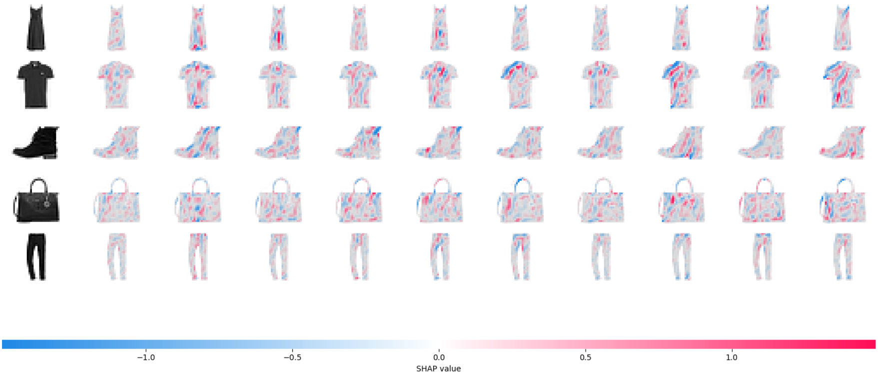



Explanation of each output class. From left to right: 'T-shirt/top', 'Trouser', 'Pullover', 'Dress', 'Coat', 'Sandal',' Shirt',' Sneaker', 'Bag',' Ankle Boot'.

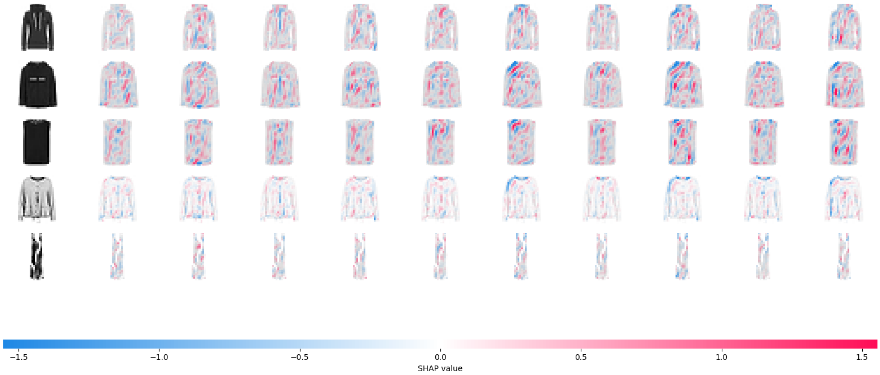
# CNN linear - Incorrectly Classified - White Background



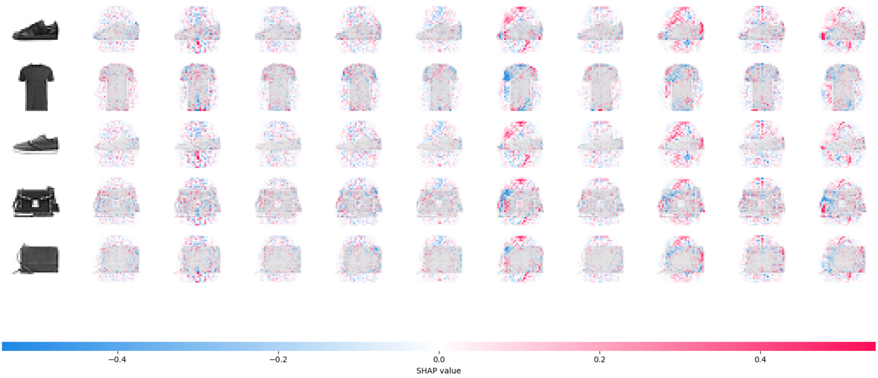
# CNN ReLU - Correctly Classified - White Background



# CNN ReLU - Correctly Classified - White Background

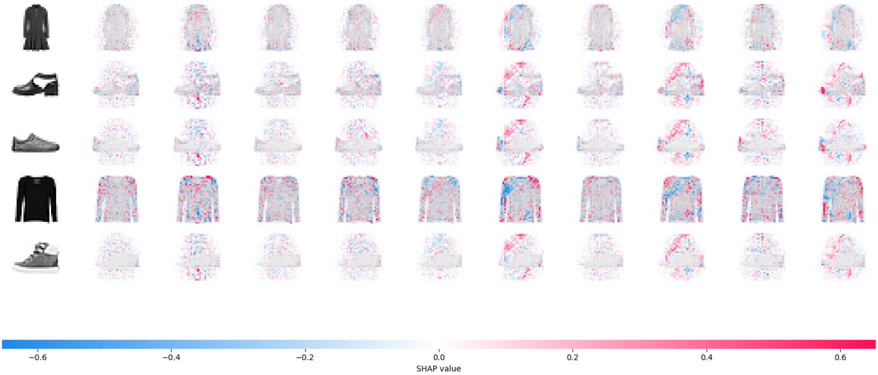


# CNN linear - Incorrectly Classified - Average Background

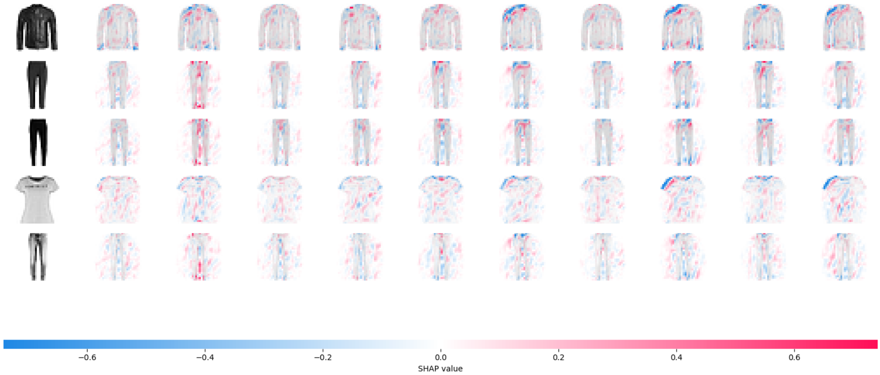



Explanation of each output class. From left to right: 'T-shirt/top', 'Trouser', 'Pullover', 'Dress', 'Coat', 'Sandal',' Shirt',' Sneaker', 'Bag',' Ankle Boot'.

# CNN linear - Incorrectly Classified - Average Background



# CNN ReLU - Correctly Classified - Average Background



# CNN ReLU - Correctly Classified - Average Background