Overview

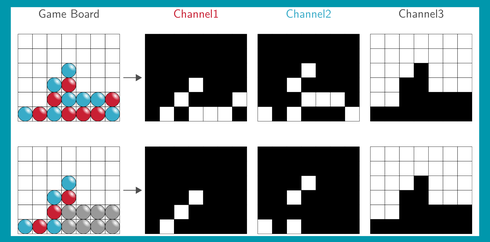
Goal: Evaluate saliency attribution methods in XAI: Which identify the important parts of a classifier input?

Problem: Hiding the relevant part of the input with baseline values or noise means evaluating the classifier off-manifold.

Solution: Train neural agents on abstract games with the same hidden information that is used in the evaluation.

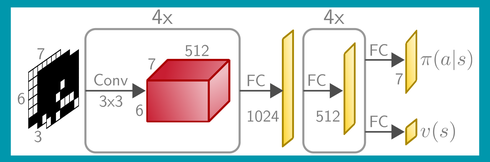
The Setup

Connect Four with randomly hidden colour information.



The Model

DNN with policy and value head for Proximal Policy Optimisation.



Advantages

- It is clear how to realise hidden information
- No off-manifold evaluation is necessary
- We can let XAI-methods play against each other

Training Characteristic Functions with Reinforcement Learning: XAI-methods play Connect Four

Stephan Wäldchen
waeldchen@zib.de

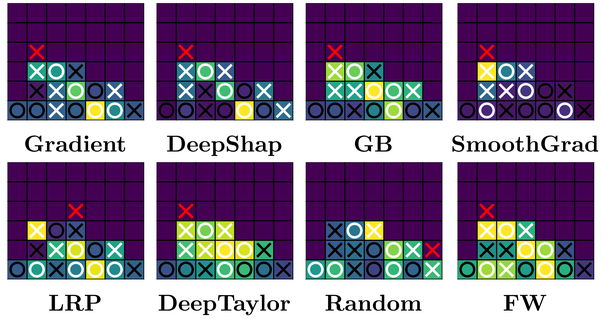
Felix Huber
huber@zib.de

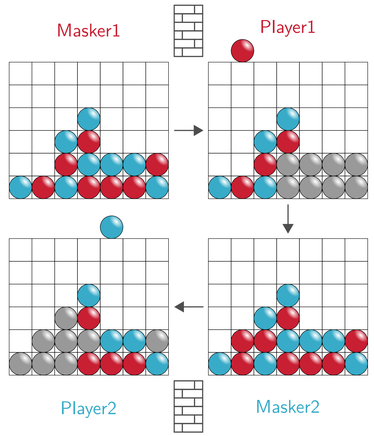
Sebastian Pokutta
pokutta@zib.de

ZIB
ZUSE INSTITUTE BERLIN

We let Explainable AI (XAI)-methods play against each other in Connect Four to compare them!

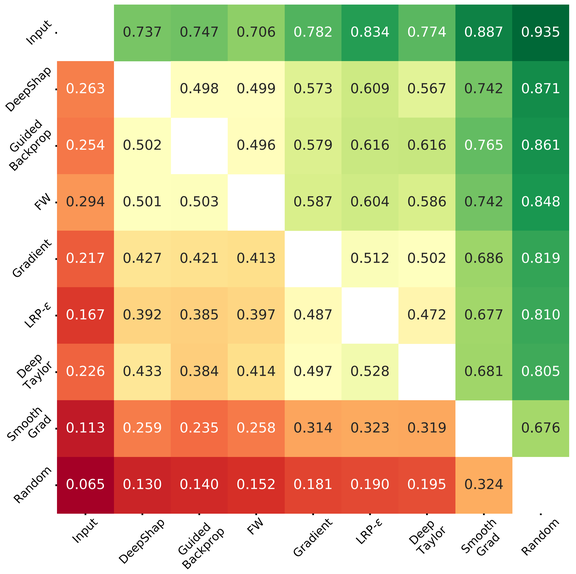
1. XAI-methods explain the output of the policy network with a heatmap of importance values.





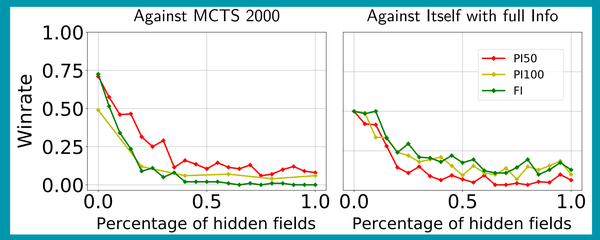
2. Select the 50% most important colour features and hide the rest for the agents next move.

3. Measure win rates of players supported by each XAI-method.



Randomly Hiding Colour Features

Every turn $t \in [42]$ we sample $p_h \sim \mathcal{U}([0, p_h^{\max}])$ and show colour features of $\lfloor p_h t \rfloor$ random game pieces each turn. We train three agents with varying $p_h^{\max}$.



The PI-50 agent with $p_h = 0.5$ performs well for all levels of hidden information and was chosen for our tournament.

Characteristic Functions

A concept from cooperative game theory. We can define

$$\nu^{\text{pol}}(S) := \pi(a^{\max}; \mathbf{x}^{(S)}) \quad \text{and} \quad \nu^{\text{val}}(S) := \nu(\mathbf{x}^{(S)}),$$
$$\text{where } \mathbf{x} \in \{0,1\}^{3\times6\times7} \text{ and } \mathbf{x}^{(S)} := [\mathbf{x}^1_S, \mathbf{x}^2_S, \mathbf{x}^3],$$

and calculate Shapley Values (SV)

$$\text{SV:} \quad \phi_{\nu,i} = \sum_{S\subseteq[d]\setminus\{i\}} \binom{d-1}{|S|}^{-1} (\nu(S\cup\{i\}) - \nu(S)),$$

and Prime Implicant Explanations (PIE)

$$\text{PIE:} \quad S^* = \operatorname*{argmin}_S |S| \quad \text{subject to} \quad \nu(S) = \nu([t]).$$

PIE with Frank-Wolfe

We solve a convex relaxation of the PIE-problem

$$\mathbf{s}^* := \operatorname*{argmin}_{\mathbf{s}} (\pi(a^*|\mathbf{x}^{(\mathbf{s})}) - \pi(a^*|\mathbf{x}))^2 \quad \text{subject to} \quad \|\mathbf{s}\|_1 \le k,$$
$$\text{where } \mathbf{s} \in [0,1]^{3\times6\times7} \text{ and } \mathbf{x}^{(\mathbf{s})} := [\mathbf{s}\odot\mathbf{x}^1, \mathbf{s}\odot\mathbf{x}^2, \mathbf{x}^3].$$

with a Frank-Wolfe solver and define the XAI-method **FW**.

Gradient [Simonyan et al. 2013]
SmoothGrad [Smikov et al. 2017]
GuidedBackprop [Springenberg et al. 2015]
DeepShap [Lundberg and Lee 2017]
LRP [Bach et al. 2015]
DeepTaylor [Montavon et al. 2018]