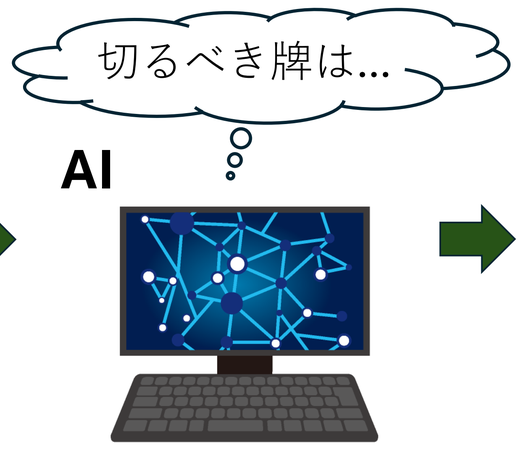
摂動付与による麻雀AIの解釈可能性向上

大倉功士, 和賀正樹, 池渕未来, 末永幸平　京都大学

https://softwarefoundationgroupatkyotou.github.io/mahjong-visualizer/

予備知識：麻雀AIのプロセス

入力:盤面情報
- 捨て牌
- 手牌
- 点棒状況
- ドラ表示牌
- …



手牌

[-5.01, -10.3, … , -7.15, -0.08]

Q値の一番高い牌が推論結果

本研究

麻雀AIの解釈可能性向上手法の提案

課題：推論結果の根拠が分かりにくい（解釈性低い）

AIが出した推論結果の根拠を、人間が説明・解釈できること

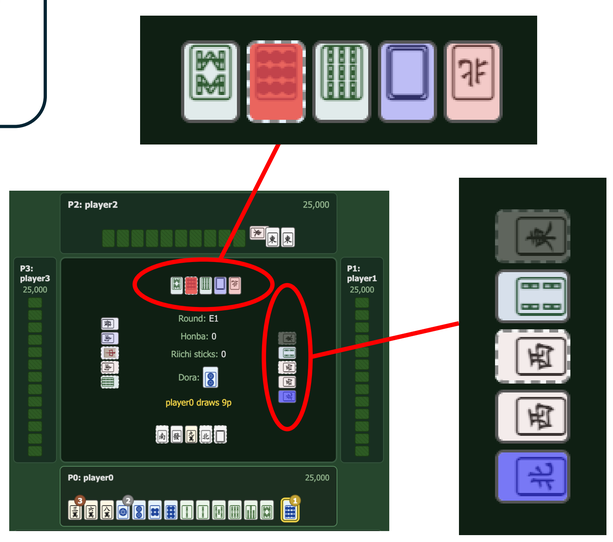
目標：出力への貢献度が高い入力要素を特定し、推論結果の根拠を推定できるように

解析したい盤面

＋

＋

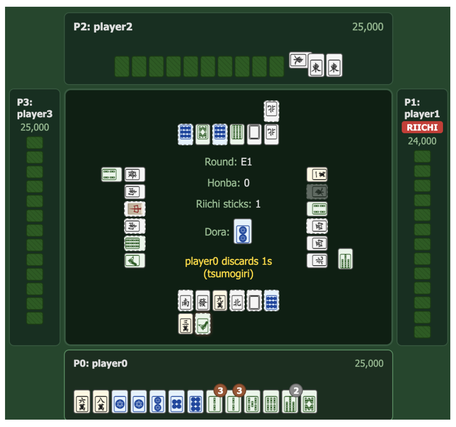
提案手法



各入力要素の貢献度を算出
赤：正の貢献度
青：負の貢献度
貢献度をどう定義？

提案手法：入力盤面を摂動させ、出力(Q値)の変化率を観察

通常時

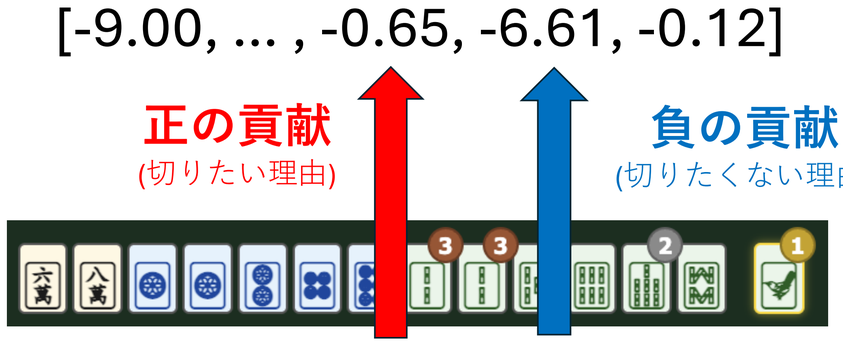


Q値の変化率が大きい＝ 推論結果に影響大

[-9.00, … , -0.65, -6.61, -0.12]

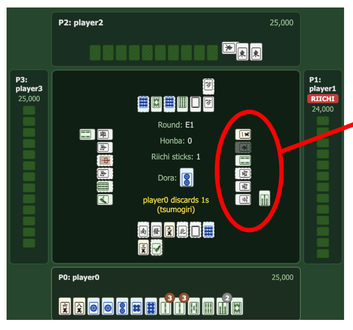
正の貢献
(切りたい理由)

負の貢献
(切りたくない理由)



[-6.83, … , -6.68, -0.79, 0.02]

摂動付与時



入力の一部(今回は捨て牌)をマスク

※現状、貢献度の高い捨て牌を基に人が根拠を判断

今後の展望

- 捨て牌以外の情報もマスクし、出力結果を観察
  ex) 手牌, ドラ表示牌, 点棒状況, …
- 貢献度の高い入力要素をもとに打牌根拠をLLMに出力させる

関連研究

本研究に似た手法
LIME : 説明したい入力の特徴量を少し変えたデータを多数作成し、それらの予測値を線形モデルで近似
SHAP : ゲーム理論の「シャープレイ値」が由来。各特徴量に関して、予測への貢献度を算出
廣瀬 et al… : dlshogiにて、入力盤面を摂動させて解釈性を向上