解釈が容易な特徴量を用いたFeature Attributionに基づく将棋AIの指し手の解釈可能性向上手法

廣瀬 雄一　和賀 正樹　末永 幸平

将棋 AI の発展は著しく，近年ではプロ棋士の将棋研究にも用いられるようになっている．一方で，将棋 AI が選ぶ指し手の解釈は人間，特に初心者にとって難しいという問題がある．将棋 AI の指し手の解釈可能性が向上することによって，将棋 AI の将棋研究への活用がより容易になることが期待される．本研究では，将棋 AI の指し手の feature attribution に基づく解釈可能性向上手法を提案する．提案手法では，各マスにおける駒の存在・非存在や利きの数といった，解釈が容易な盤面情報を特徴量として説明に用いる．提案手法は，選ばれた指し手に対する各特徴量の貢献度を計算することによって，指し手にとって重要な特徴量を求める．貢献度の計算には，特異度と関連度という二つの指標から特徴量の貢献度を算出する SARFA と呼ばれる手法を利用する．提案手法は対象とするモデルに依存しないため，評価関数を持つ任意の将棋 AI に適用可能である．本手法をいくつかの盤面に適用した．本論文ではその有用性と改善点について議論する．

1　はじめに

将棋 AI は，トッププロ相手に勝利を収めるという結果を残すなど棋力の著しい向上を見せている．また，将棋研究や対局の解説といった，対局以外の用途へ活用されることも増えてきている．

一般的な将棋 AI は，局面を評価する評価関数を用いたゲーム木探索を行うことで指し手を選択しており，ゲーム木探索の効率や評価関数の精度が将棋 AI の棋力に大きく影響する．現在では，大量の対局データから学習されたニューラルネットワークモデルが評価関数として用いられるようになっており，評価関数の表現力や精度が向上している．また，モンテカルロ木探索 [1] などの効率的な探索アルゴリズムの登場やコンピュータの性能向上によって，ゲーム木探索でより多くの局面を探索することが可能になっている．

ニューラルネットワークは高い表現力を持つ関数を学習することができる一方で，その推論結果については，なぜその結果が得られたのかという判断根拠を説明することが困難であるという問題がある．これは，将棋 AI が将棋研究や解説といった，対局以外の目的で使用される場合に特に問題となる．将棋 AI に限らず，画像処理や自然言語処理などの他のタスクでも機械学習モデルの解釈可能性は課題になっており，解釈可能性を向上させるための様々な手法 [6] [4] [8] [7] [5] [10] が提案されている．

解釈可能性向上手法でよく用いられているアプローチの 1 つに *feature attribution (FA)* がある．FA は，入力データの各特徴量が推論結果に与える影響の大きさ（**貢献度**）を数値化することによって重要な特徴量を特定する手法である．例えば，画像処理では入力画像のピクセルごとに貢献度を計算し，その結果をヒートマップとして提示するという手法が多く研究されている [6] [10].

将棋 AI やチェス AI に FA 手法を適用した先行研究はいくつかあるが，入力局面において将棋 AI が選んだ指し手に対する貢献度を駒/マスごとに算出し，それをヒートマップとして盤面に重ねて表示しているものが多い [3] [11]．ヒートマップによる説明は有用

Improving Interpretability of Shogi AI's Moves Based on Feature Attribution Using Easy-to-Interpret Features.

Yuichi Hirose, Masaki Waga, Kohei Suenaga, 京都大学大学院情報学研究科, Graduate School of Informatics, Kyoto University.

であるが，高い貢献度が与えられた駒/マスが指し手にとってどのような意味で重要なのかまでは分からないため，特に初心者にとっては出力されたヒートマップの解釈が困難な場合がある．我々は，将棋 AI の 1 つである dlshogi [†1] が用いている入力特徴量についてそれぞれの貢献度を求めることで，単にヒートマップを提示するよりも分かりやすい説明を生成することを目的とした手法を提案した [12]．一方で，この手法には dlshogi 以外の将棋 AI に適用することが困難であるという欠点がある．

本研究では，将棋 AI の種類に依存しない，解釈が容易な特徴量を用いた解釈可能性向上手法を提案する．本研究を水匠 5[†2] の評価関数に適用した例を図 1 に示す．この局面は△ 3 九馬，▲ 1 八王，△ 2 八馬という 3 手詰めが存在する局面である．本研究の手法は，△ 3 九馬という指し手にとって重要な特徴量を列挙している．さらに，指し手にとって重要な駒/マスをヒートマップとして盤面に重ねて表示している．列挙されている特徴量について，例えば「3 九に先手の 1 個以上の利きがない」ことはそれにより馬が先手に取られないという意味で重要であり,「2 八に駒がない」ことはそれにより 3 九馬が王手になるという意味で重要であるといえる．実行例では， dlshogi の特徴量に新たにいくつかの種類の特徴量を加えたものを説明に使用している.「○○に駒がある/ない」という特徴量はその 1 つである．このように，本研究の手法は dlshogi 以外にも適用可能であり，用いる特徴量も変更することができる．

本研究の手法は，局面 s，指し手 a を指定すると，将棋 AI による指し手の評価値 $Q(s,a)$ に対する s に含まれる特徴量ごとの貢献度を算出する．貢献度を求める特徴量の集合は，対象とする将棋 AI に依存せず任意に設定することができる．また，本研究の手法は $Q(s,a)$ の値を得ることができる任意の将棋 AI に対して適用することができる．

提案手法は，[3] や [12] と同じく摂動に基づく FA と呼ばれる手法を用いている．摂動に基づく FA では，推論結果に対する特徴量の貢献度を計算するために，元の入力データに摂動を加えた新たな入力データを生成する．摂動を加えたことにより変動するモデルの推論結果を基にして，摂動後に元の入力から変化した特徴量の貢献度を計算する．例えば SARFA [3] では，盤面から駒を 1 つずつ取り除くという摂動によって駒ごとの貢献度を計算している．また，[12] で提案された手法 (以降 DL-SARFA と呼ぶ) は dlshogi の特徴量の値の 1 つを反転させることによって特徴量ごとの貢献度を計算している．

DL-SARFA は，この摂動の方法により dlshogi 以外の将棋 AI に適用することが困難になっている．DL-SARFA における摂動後の入力は，例えば，あるマスに歩があるのにその 1 つ前のマスにその歩の利きがない，というような異常な入力となっているからである．dlshogi とは異なる入力特徴量を用いる他の将棋 AI はこのような入力を受け取ることができないため，手法を適用することができない．一方で，提案手法では入力局面に対して駒の移動や除去といった操作を行うことで摂動を加える．特定の将棋 AI の特徴量に対して摂動を加えるわけではないため，対象とする将棋 AI が採用している入力特徴量の種類にかかわらず手法を適用することが可能となっている．

提案手法は入力局面を摂動して新たなデータ群を生成した後，摂動された局面それぞれに対して貢献度を計算する．貢献度計算は SARFA を拡張した DL-SARFA における計算式に従って行う．SARFA は指し手にとって真に重要な特徴量のみで貢献度が高くなることを目的に設計された手法である．貢献度計算に特異度と関連度という 2 つの指標を用いることによって，説明したい指し手に直接関係ない特徴量や，説明したい指し手だけでなく他の指し手の評価値にも影響を与えるような特徴量の貢献度を低く見積もるようにしている．SARFA には特異度が正の場合のみしか貢献度を計算できないという欠点があったが，DL-SARFA では特異度が負の場合にも貢献度を計算できるように拡張されている．

SARFA やその拡張では，一度の摂動によって値が変化する特徴量は 1 つだけだった．ゆえに，摂動された局面に対して計算した貢献度を，摂動により変

†1 https://github.com/TadaoYamaoka/DeepLearningShogi

†2 https://drive.google.com/file/d/1T-Go2KImMfKD_4m_j4fQFXrEfaGgAcS_

貢献度	特徴量
0.675	3 九に先手の 1 個以上の利きがない
0.603	2 八に先手の 2 個以上の利きがない
0.584	2 八に駒がない
0.541	2 九に先手の 1 個以上の利きがない
0.506	3 八に先手の 2 個以上の利きがない
0.444	3 七に後手の銀がある

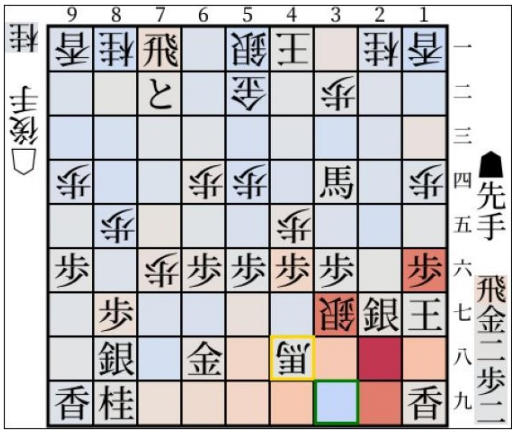

図 1　提案手法の実行例．3 手詰めの局面で後手が 3 九馬を指す場面である．左の表では指し手に対する貢献度の高い特徴量が列挙されている．右の図は各駒/マスの貢献度を表すヒートマップである．

化した特徴量の貢献度とすることができた．一方で，提案手法では摂動によって複数の特徴量の値が変化するため，摂動された局面から計算した貢献度の値をそのまま 1 つの特徴量の貢献度として用いることはできない．そこで，各特徴量の貢献度を，その特徴量の値が変化した局面と変化しなかった局面における貢献度の平均の差として定義する．この計算方法は PN-RISE [4] に基づいており，得られる値は特徴量に摂動を加えたことによる貢献度の変化量を意味する．

本稿では，提案手法を複数の将棋 AI を対象として局面に適用した結果を報告する．また，既存手法との定量的な比較を行い，提案手法の持つ有用性と課題について議論した．

本論文は以下のように構成されている．2 節 で 関連研究について述べ，3 節 で背景知識として解釈可能性向上手法の 1 つである DL-SARFA について説明する．4 節 で提案手法についての説明を行い，5 節では既存手法と比較する実験を行った結果について述べ，6 節で結論を述べる．

2　関連研究

機械学習モデルの解釈可能性向上手法には様々なアプローチが存在する．入力の各特徴量が予測に対してどれだけ貢献したか（貢献度）を数値化する feature attribution (FA) はその 1 つである．特徴量の貢献度を求めるための方法もいくつかあるが，その 1 つが摂動に基づく FA と呼ばれる手法である．摂動に基づく FA では，入力データに摂動を加えることで新たな入力データ群を生成し，それらをモデルの入力として得られる出力の，元の入力に対する出力からの変化を基に貢献度を計算する．SHAP [5] と LIME [7] は摂動に基づく FA に属する代表的な手法である．これらの手法では，入力データと一対一対応する，解釈可能な特徴ベクトルを設定し，その特徴ベクトルに摂動を加えることで重要な特徴量を求める．これらの手法と比較して本研究は，入力データと相互変換が可能でない特徴量を説明に利用しようとしている点で異なる．

ゲーム AI などの強化学習モデルを対象とした摂動に基づく FA の手法として SARFA [3] などがある．SARFA は強化学習モデルにおける Q 関数の値に基づいて貢献度を計算する．論文内で手法をチェスに対して適用する際には，駒を 1 つ盤面から取り除くという摂動を加えることによって各駒の貢献度を計算していた．Fritz ら [2] は，SARFA のいくつかの改善点を指摘しており，その 1 つとして駒が置かれていない空きマスの貢献度を計算できないという点を挙げている．本研究では，より複雑な摂動を加えることによって，空きマスについての貢献度も計算することが可能になっている．

将棋 AI に解釈可能性向上手法を適用した先行研究もいくつか存在する．中屋敷ら [11] は，将棋 AI のおける局面評価値に対して，ニューラルネットワークの

勾配を用いる手法 [9][10] を適用し比較している．また，廣瀬ら [12] は将棋 AI の 1 つである dlshogi を対象として，Q 関数に対する dlshogi の入力特徴量それぞれの貢献度を計算する手法を提案した．前者で用いられている手法は適用可能な範囲が畳み込みニューラルネットワークの構造を持つモデルに限られており，後者で提案されている手法はモデルの入力特徴量に依存しているが，本研究ではモデルの構造に依存しない手法を提案する．

3 背景知識: DL-SARFA

廣瀬らは， dlshogi を対象とした解釈可能性向上手法を提案した [12]．この手法 (DL-SARFA と呼ぶ) は dlshogi の入力特徴量それぞれの指し手に対する貢献度を計算する．

図 2 に示すように，この手法は選ばれた指し手に対する貢献度が高い特徴量のリストを表形式で表示し，また，特徴量ごとの貢献度を基に各マスの貢献度を計算してヒートマップとして表示している．重要な特徴量のリストを表示する際には，特徴量のフィルタリングとグループ化という処理を行うことでより有用な説明を生成することを試みている．例えば,「2 八に駒がない」ことは,「2 八に後手の歩がない」ことや,「2 八に先手の銀がない」ことなどの特徴量を 1 つにまとめたものである．

DL-SARFA では，特徴量 f の貢献度を求めるために入力局面 s に対して f の値を反転させた入力 $s|_f$ を生成する．貢献度の計算手法には SARFA [3] を拡張した手法を用いている．SARFA は，説明の対象とする指し手のみにとって重要な特徴量を抽出することで精度の高い説明を生成することを目的にした手法であり，特異度と関連度という 2 つの指標を組み合わせて貢献度を求める．

3.1 特異度

特異度は，局面 s における特徴量 f が指し手 a の評価値のみに貢献し，その他の指し手 a' には影響を与えていないことを測る指標である．例えば，飛車や角行といった大駒と呼ばれる駒に関する特徴量に摂動を加えるとする．このような特徴量の値の変化は局面自体に対して大きな影響を与えるため，指し手全体の評価値が大きく変動する．説明したい指し手の評価値が大きく変動したとしても，指し手全体に同じように大きな影響を与えるような特徴量は説明したい指し手のみにとって重要であるとはいえない．SARFA では，このような特徴量の貢献度が高くならないように，摂動前後での評価値の絶対的な変動ではなく相対的な変化量を特異度として定義し，貢献度の指標としている．具体的には，まず相対的な評価値を表す $P_A(s,a)$ を

$$P_A(s,a) := \frac{\exp(Q(s,a))}{\sum_{a' \in A} \exp(Q(s,a'))}$$

で定義する．$P_A(s,a)$ は，指し手の集合 A の中で softmax 関数によって正規化された，局面 s における a の評価値を表す．その上で，局面に摂動が加えられたことによる指し手の相対的な評価値の変動を表す $\Delta p(s,a,s')$ を

$$\Delta p(s,a,s') = P_{A_{s,s'}}(s,a) - P_{A_{s,s'}}(s',a)$$

とする．ここで，$A_{s,s'}$ は局面 s と s' に共通する合法手の集合である．この $\Delta p(s,a,s|_f)$ を f の特異度とする．

3.2 関連度

関連度は，特徴量 f に摂動を加えたことによる，a を除いた指し手の評価値の変動を測る指標である．特徴量 f に対する摂動によって指し手の評価値の分布が大きく変動する場合，指し手 a とは直接関係がなくても特異度が高くなる場合がある．例えば， f に摂動を加えたときに指し手 a の評価値は変化せずそれ以外の特定の指し手の評価値が大きく増加した場合，a の相対的な評価値は低下するため特異度は高くなる．一方で，指し手 a の評価値は変化していないため f が a に関係があるわけではない．関連度は，このような特徴量の貢献度を低く見積もるための指標である．具体的には，まず $P_{\mathrm{rem}}(s,a,s')(a')$ を

$$P_{\mathrm{rem}}(s,a,s')(a') = P_{A_{s,s'}-\{a\}}(s,a')$$

で定義する．$P_{\mathrm{rem}}(s,a,s')(a')$ は s と s' の合法手の共通部分から a を除いた指し手集合の中で a' の評価値を softmax 関数によって正規化した値である．したがって，$P_{\mathrm{rem}}(s,a,s')$ は a 以外の合法手に対する評価値の確率分布と見ることができる．次に $P_{\mathrm{rem}}(s,a,s')$

貢献度	特徴量
0.602	3 九に先手の利きがない
0.543	2 八に駒がない
0.439	1 七に先手の王がある
0.306	2 七に先手の銀がある
0.273	3 七に後手の銀がある
0.233	2 八に後手の 1 個以上の利きがある
0.231	2 六に先手の 2 個以上の利きがある
0.228	2 六に後手の 2 個以上の利きがない

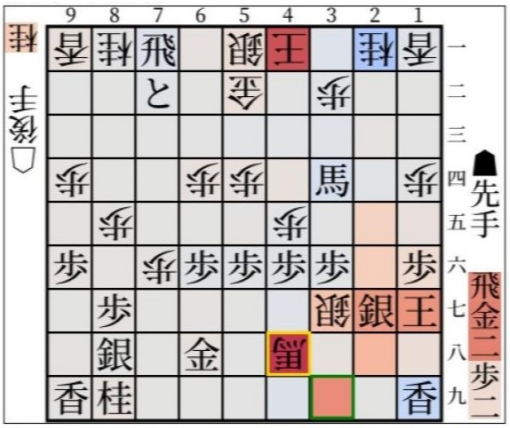

図 2　DL-SARFA の適用例．△ 3 九馬に対する説明である．左の表では指し手に対する貢献度の高い dlshogi の特徴量が列挙されている．右図は特徴量ごとの貢献度から計算した，各マスの貢献度を表すヒートマップである．

と $P_{\mathrm{rem}}(s', a, s)$ という 2 つの分布の間の差異を測る．2 つの分布間の KL ダイバージェンスを $D_{\mathrm{KL}}(s, a, s')$ とし，

$$K(s, a, s') = \frac{1}{1 + D_{\mathrm{KL}}(s, a, s')}$$

とする．$D_{\mathrm{KL}}(s, a, s')$ は $[0, \infty)$ の範囲の値を取り，2 つの分布が同一であれば 0 となり，差異が大きくなるにつれて大きな値を取る．したがって K の取る範囲は $(0, 1]$ であり，2 つの分布が異なるほど小さい値を取る．$K(s, a, s|_f)$ を s における a に対する f の関連度と定義する．

3.3　貢献度

貢献度は，特異度と関連度の調和平均として計算する．このとき，単純に調和平均を計算しようとすると特異度が負の場合に値を計算できない場合がある．そこで，DL-SARFA では特異度の絶対値と関連度の調和平均を計算した後に特異度の符号を掛けている．すなわち，

$$S[s, a, s'] = \mathrm{sgn}(\Delta p(s, a, s')) \cdot \frac{2K(s, a, s')|\Delta p(s, a, s')|}{K(s, a, s') + |\Delta p(s, a, s')|}$$

である．上式より計算された $S[s, a, s|_f]$ を s における a に対する f の貢献度と定義する．

3.4　改善点

DL-SARFA には (1) 説明の対象とするモデルに対応する特徴量しか説明に用いることができない，(2) 摂動によって違法な入力が生成されるという 2 つの欠点がある．(1) は，dlshogi とは異なる入力特徴量を採用していているモデルに対して手法を適用するのが困難であるという点で問題である．(2) は明らかな分布外データを用いている点で問題である．例えば，dlshogi の特徴空間では同じマスに後手の歩と先手の角が同時に存在するようなありえない局面も表現することができる．そのような異常な入力に対する推論結果の精度は保証されない．DL-SARFA において生成される $s|_f$ は合法な盤面から特徴量を 1 つだけ反転させた違法な入力であり，推論結果 $Q(s|_f, a)$ を用いて得られる貢献度も妥当な値であることは保証されない．

4　提案手法

本研究では，特定のモデルに依存せず任意に設計した特徴量に対して，各特徴量のある局面における指し手に対する貢献度を計算する手法を提案する．特徴量ごとの貢献度を計算する疑似コードをソースコード 1 に示す．

3.4 節で述べた改善点を解決するために，本研究では盤面 s に対する説明を生成する際に (1) DL-SARFA のように特徴量を直接摂動させるのではなく，盤面 s を合法手の範囲内で摂動させた場面 $s_1, \ldots, s_{N_s}$ を作り，摂動 i ごとの貢献度を I_i 求め，(2) 各特徴量 f に対して，f の値が s と変化した摂動の集合 $M^+_{s,f}$ と変化しなかった摂動の集合 $M^-_{s,f}$ を求め，(3) $M^+_{s,f}$ に属

ソースコード 1　各特徴量の貢献度を計算する擬似コード

```
def explain(q_func, state, hand,
    convert_to_features):
  # q_func: 局面と指し手を受け取り，指し手の評価値
    (Q値)を計算する関数．
  # state: 盤面を表すデータ
  # hand: 指し手を表すデータ
  # convert_to_features: 局面を説明に用いる特徴量の
    リストに変換する関数．返り値は，サイズ N_f で
    要素が{0, 1}の配列．

  # perturbated_states:
    stateに摂動を加えた局面の集合．サイズを N_p と
    する．
  perturbated_states = perturbate(state)

  # sarfa_n: 元の局面，指し手，摂動された局面からそ
    の摂動の貢献度を計算する関数
  # importances_of_states: サイズ N_p の配列．
  importances_of_states = map(lambda ps: sarfa_n(
    q_func, state, hand, ps), perturbated_states)

  # 局面を説明に用いる特徴量に変換．
  # original_features: サイズ N_f の配列．
  original_features = convert_to_features(state)
  # perturbated_features_array: サイズ N_p * N_f の
    配列．
  perturbated_features_array = map(lambda ps:
    convert_to_features(ps), perturbated_states)

  # importances_of_features:サイズ N_f の配列．
  # 各要素は説明の用いる特徴量ごとに計算した貢献度
    .
  importances_of_features = zero_array(N_f)

  for i from 0 to N_f:
    feature_i = original_features[i]
    # masked: 摂動前後で値が変化した特徴量のインデ
      ックスのリスト
    masked = list of j where
      (perturbated_features_array[j][i] !=
        feature_i)
    # not_masked: 摂動前後で値が変化しなかった特徴
      量のインデックスのリスト
    not_masked = list of j where
      (perturbated_features_array[j][i] ==
        feature_i)
    importances_of_features[i] =
        mean(importances_of_states[masked])
        - mean(importances_of_states[not_masked])

  return importances_of_features
```

する摂動の貢献度の平均値と $M^{-}_{s,f}$ に属する摂動の貢献度の平均値の差を f の貢献度とする．このように盤面の摂動と，それにより変化した特徴量の貢献度の計算を分離することで，(1) 任意のモデルと任意の特徴量の集合に対して手法を適用できるようになり，(2) 違法な入力を生成することなく貢献度を計算できるようになる．以下では，局面の摂動と貢献度の計算について，より具体的に述べる．

4.1　局面の摂動

提案手法では，入力に摂動を加えて新たな局面を生成する際には以下に挙げる操作のいずれかを行う．

(a)　盤上の駒や持ち駒を局面から 1 つ取り除く．
(b)　1 枚の持ち駒を盤上に配置する．持ち駒を置くマスに駒があった場合その駒は取り除く．
(c)　マスを 2 箇所選び，駒を入れ替える．
(d)　複数のマス/持ち駒をランダムに選択し，存在する駒をシャッフルして再び配置する．

SARFA では操作 (a) のみを行っていたが，空きマスに摂動を加えることができないという問題点があった．空きマスの貢献度を計算することを可能にするために，本手法では空きマスにも駒が配置されるような摂動を加える．単に駒を新しく追加して空きマスに配置するという操作を行うと，駒の枚数が規定の枚数より多くなり違法な入力となってしまう．そこで，操作 (b)-(d) のような駒の配置を入れ替える操作を行う．摂動を行う回数について，(a)-(c) については，考えられる全てのパターンを試す．(d) については，組み合わせの総数が莫大であるため生成する局面の数を指定するようにする．

以上の方法で生成された局面それぞれについて，合法な局面であるかどうかを判定する．局面が合法であるかどうかの判定には以下の条件を用いた．

- 駒が規定の枚数より多くない
- 王と玉が盤上に存在する
- 王手が放置されていない
- 二歩がない
- 移動先がない歩，桂，香がない

条件を満たしていると判定された局面のみを実際の将棋 AI の入力とする．

4.2 貢献度計算

摂動により生成された局面の集合を S' と書く．全ての摂動された局面 $s' \in S'$ に対して 3.3 節で示した手法に従って貢献度 $S[s,a,s']$ を求める．これらの値を基に，特徴量 f の貢献度 $S_{\mathrm{feat}}[s,a,f]$ を以下のように計算する．

$$S_{\mathrm{feat}}[s,a,f] = E_{S'}[S[s,a,s']|f_s \neq f_{s'}] - E_{S'}[S[s,a,s']|f_s = f_{s'}].$$

ここで f_s は局面 s における特徴量 f の値である．$E_{S'}[S[s,a,s']|f_s \neq f_{s'}]$ は f が摂動された場合における貢献度の期待値である．また，$E_{S'}[S[s,a,s']|f_s = f_{s'}]$ は f が摂動されない場合における貢献度の期待値である．この項は，指し手に無関係な特徴量の貢献度を 0 に近づけるための項であり PN-RISE [4] における貢献度計算の手法に基づく．したがって，$S_{\mathrm{feat}}[s,a,f]$ は f を摂動した場合と摂動しなかった場合の貢献度の期待値の差を表す．

5 実験

5.1 定性的評価

図 3 は水匠の評価関数 †3 を，図 4 は dlshogi を対象として同じ局面，指し手に対して提案手法を適用した例である．ここで，水匠の評価関数の出力は局面の評価値 $V(s)$ であるため，局面 s における指し手 a の評価値は s で a を指した後の局面 s_a における評価値との差分，$V(s_a) - V(s)$ によって定義した．実行例の局面においては，水匠と dlshogi で貢献度が上位となる特徴量がほとんど一致していることが分かる．このように複数の種類の将棋 AI に対して解釈容易な特徴量を用いた説明を生成し比較することは，DL-SARFA では行えず，提案手法によって可能になった点である．

一方で，提示されている特徴量のほとんどが 5 八の竜王に関するものであり，説明としては冗長になってしまっている．よりよい特徴量の設計や提示の方法の工夫は今後の課題である．

5.2 定量的評価

5.2.1 評価指標

評価指標として，insertion metric [6] を用いた．insertion metric は FA に属する解釈可能性向上手法を評価するための指標である．入力データ s，指し手 a と，FA 手法によって得られた各特徴量の貢献度を入力とする．まず，開始状態の入力 s_0 を設定する．例えば入力データが画像だとすると，開始状態として全ピクセルを黒く塗りつぶした画像を用いることが考えられる．開始状態の入力に，貢献度の高い順に特徴量を追加していく．k 個の特徴量を追加したときの入力データを s_k とする．モデルが入力 s において指し手 a を選択する確率を $P(s,a)$ とすると，insertion の値 (AUC: Area Under Curve) は以下のように計算される．

$$\mathrm{AUC} = \frac{1}{n} \cdot \sum_{k=1}^{n} P(s_k, a)$$

$P(s,a)$ は Q 関数 $Q(s,a)$ を合法手の中で正規化した値として定義する．つまり，

$$P(s,a) = \frac{\exp(Q(s,a))}{\sum_{a' \in A_s} \exp(Q(s,a'))}$$

である．AUC は 0 から 1 の値を取り，1 に近いほど良い値である．FA 手法が指し手にとって真に重要な特徴量に対して高い貢献度を与えている場合，k の値を 1 から増やしていくと $P(s_k,a)$ の値が急激に高くなるため，AUC は大きい値を取る．

今回の実験では dlshogi の入力特徴量について貢献度を計算し，insertion metric を計算する．dlshogi の入力特徴量を用いて insertion metric を計算するとき，開始状態は全要素が 0 の入力とし，元の入力で値が 1 を取る特徴量の中で貢献度が高いものから順に 1 に変更していく．

5.2.2 ベースライン

ベースラインとして DL-SARFA [12] を実装し評価した．DL-SARFA は dlshogi の入力特徴量を 1 つずつ摂動し，各特徴量の貢献度を計算する手法である．

5.2.3 実験設定

実験では dlshogi のモデルとして第 2 回世界将棋 AI 電竜戦エキシビジョンバージョン †4 を使用した．実

†3 https://drive.google.com/file/d/1T-Go2KImMfKD_4m_j4fQFXrEfaGgAcS_

†4 https://github.com/TadaoYamaoka/DeepLearningShogi/releases/tag/dr2_exhi

貢献度	特徴量
0.673	8 八に後手の 1 個以上の利きがある
0.673	7 八に後手の 2 個以上の利きがある
0.628	6 八に後手の 1 個以上の利きがある
0.619	4 八に後手の 1 個以上の利きがある
0.611	5 七に後手の 1 個以上の利きがある
0.608	5 九に後手の 1 個以上の利きがある
0.606	4 九に後手の 1 個以上の利きがある

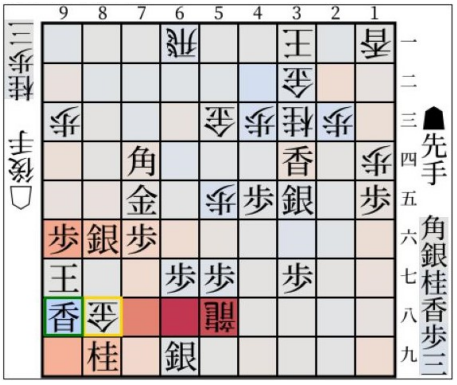

図 3　提案手法を水匠に適用した実行例．3 手詰めの局面であり，後手が 9 八金を指している．表で提示している特徴量は dlshogi のものである．

貢献度	特徴量
0.362	8 八に後手の 1 個以上の利きがある
0.348	7 八に後手の 2 個以上の利きがある
0.333	6 八に後手の 1 個以上の利きがある
0.327	4 八に後手の 1 個以上の利きがある
0.321	5 七に後手の 1 個以上の利きがある
0.320	5 九に後手の 1 個以上の利きがある
0.320	5 八に後手の竜がある

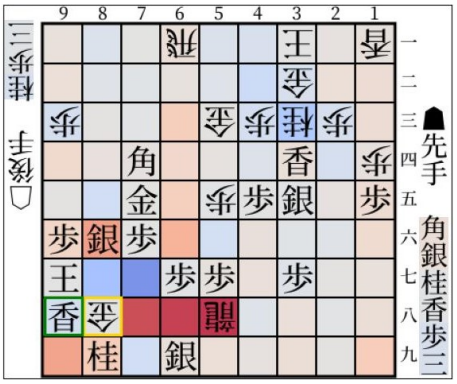

図 4　図 3 と同じ局面，指し手について提案手法を dlshogi に適用した実行例．

表 1　AUC と実行時間の平均値の比較

	AUC	Time (sec)
提案手法	0.256	6.61
DL-SARFA	0.663	9.45

験に使用した PC は Intel Xeon Gold 6226 2.7GHz CPU, 754GB RAM であり，Quadro RTX 8000 を持つ．実行環境としては Python 3.9.15，PyTorch 1.12.1 を使用した．また，実験データとして将棋 DB2[†5] から，実行時点 (2023 年 7 月 26 日) にて最新のプロ棋士の対局 135 局を取得し，そこから 500 個の局面を無作為に抽出した．抽出の際に局面の重複は取り除いた．

5. 2. 4　実験結果

表 1 にそれぞれの手法の AUC と実行時間を示す．提案手法は DL-SARFA と比較して AUC が大幅に低くなっていることが分かる．

AUC に大きな差が生まれる原因の 1 つに，提案手法が dlshogi の特徴量への摂動ではなく局面上の駒への摂動を通じて特徴量に貢献度を割り振っていることがあると考えられる．このことが問題になる場合として，例えば「2 八に先手に歩がある」と「2 七に先手の歩の利きがある」という dlshogi の 2 つの特徴量について，前者は指し手に対してあまり重要ではないが後者は重要であるとする．DL-SARFA は特徴量を 1 つずつ摂動することができ，入力局面から 1 つの特徴量の値を反転させた入力を用いて貢献度を計算する．そのため，前者のみに高い貢献度を与える

†5　https://shogidb2.com/

ことができる．一方で，提案手法では局面上で駒を動かすという摂動を加えるため，摂動によってこの 2 つの特徴量のうち片方だけの値が変化することはなく，必ず両方の値が変化する．そのため，両方の特徴量に高い貢献度が与えられることになる．この場合，提案手法は無関係な特徴量に高い貢献度を割り振ることになるため，AUC は低くなると考えられる．このように，提案手法は相関関係，依存関係がある特徴量に対しては正しく貢献度を計算できないことがある．

6 結論

本研究では，将棋 AI が選ぶ指し手の解釈可能性を向上させるために，指し手に対する各特徴量の貢献度を計算する手法を提案した．提案手法は，我々が以前提案した DL-SARFA [12] と同様に，盤面上の重要な駒を示すヒートマップと局面で重要な特徴量の表を説明として提示する．DL-SARFA では dlshogi のモデルを対象として，dlshogi の入力特徴量の指し手に対する貢献度を計算して提示しており，他の将棋 AI に適用することが困難だった．それに対し，提案手法は将棋 AI モデルや入力特徴量の種類に依らずに適用できる手法になっている．

提案手法と DL-SARFA を比較する実験を行ったところ，提案手法の説明としての精度は DL-SARFA に比べて非常に低くなっているという結果が得られた．摂動の方法，貢献度計算の方法については大きな改良の余地があると考えられ，説明の精度改善は今後の課題である．

謝辞 本研究の一部は，JST CREST JPMJCR2012 の支援を受けて行われたものです．

参 考 文 献

[1] Coulom, R.: Efficient Selectivity and Backup Operators in Monte-Carlo Tree Search, *Computers and Games*, van den Herik, H. J., Ciancarini, P., and Donkers, H. H. L. M. J.(eds.), Berlin, Heidelberg, Springer Berlin Heidelberg, 2007, pp. 72–83.

[2] Fritz, J. and Fürnkranz, J.: Some Chess-Specific Improvements for Perturbation-Based Saliency Maps, *2021 IEEE Conference on Games (CoG)*, 2021, pp. 01–08.

[3] Gupta, P., Puri, N., Verma, S., Kayastha, D., Deshmukh, S., Krishnamurthy, B., and Singh, S.: Explain Your Move: Understanding Agent Actions Using Specific and Relevant Feature Attribution, *International Conference on Learning Representations (ICLR)*.

[4] Hatakeyama, Y., Sakuma, H., Konishi, Y., and Suenaga, K.: Visualizing Color-wise Saliency of Black-Box Image Classification Models, *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020.

[5] Lundberg, S. M. and Lee, S.-I.: A Unified Approach to Interpreting Model Predictions, *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, Red Hook, NY, USA, Curran Associates Inc., 2017, pp. 4768–4777.

[6] Petsiuk, V., Das, A., and Saenko, K.: RISE: Randomized Input Sampling for Explanation of Black-box Models, *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[7] Ribeiro, M. T., Singh, S., and Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, New York, NY, USA, Association for Computing Machinery, 2016, pp. 1135–1144.

[8] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D.: Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization, *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[9] Simonyan, K., Vedaldi, A., and Zisserman, A.: Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, *Workshop at International Conference on Learning Representations*, 2014.

[10] Smilkov, D., Thorat, N., Kim, B., Viégas, F. B., and Wattenberg, M.: SmoothGrad: removing noise by adding noise, *CoRR*, Vol. abs/1706.03825(2017).

[11] 中屋敷太一, 金子知適: 将棋用ニューラルネットワークへの顕著性抽出手法の適用, ゲームプログラミングワークショップ *2018* 論文集, Vol. 2018, nov 2018, pp. 1–8.

[12] 廣瀬雄一, 和賀正樹, 末永幸平: Feature Attribution を用いた dlshogi の指し手の解釈可能性向上手法, 研究報告ゲーム情報学（*GI*）, Vol. 2023-GI-49, No. 13, March 2023, pp. 1–8.