様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　４ 年　 ６ 月　 １ 日現在

機関番号：１３３０２
研究種目：若手研究
研究期間：2020～2021
課題番号：２０Ｋ１９９４６

研究課題名（和文）AlphaZeroによる完全情報ゲームの理論値と最適戦略の学習

研究課題名（英文）AlphaZero toward Theoretical Values and Optimal Plays of Perfect Information Games

研究代表者
ＨＳＵＥＨ　ＣｈｕＨｓｕａｎ（HSUEH, Chu Hsuan）

北陸先端科学技術大学院大学・先端科学技術研究科・助教

研究者番号：３０８４７４９７

交付決定額（研究期間全体）：（直接経費）　　2,000,000円

研究成果の概要（和文）：AlphaZero はゲームのルールのみを知識として用い，自己対戦によってゼロからプロ棋士を上回る強さを持つことができる．しかし，最適戦略や理論値を学習できるのかどうかについては十分調べられていない．さらに，不確定要素を含むゲームへの適用も殆どない．
本研究ではまず，各局面の最適戦略と理論値がわかることができるような規模の小さいゲームを対象とした．AlphaZero の学習は，いろんな設定では最適戦略や理論値に収束できたことを示した．そして，規模がより大きい・不確定要素を含むゲームへ適用した AlphaZero が大会準優勝レベルの強さを持つことも確認した．

研究成果の学術的意義や社会的意義
AlphaZero のパラメータを丁寧に調べ，学習結果への影響を明らかにしたことは学術的意義があった．
AlphaZero を適用する研究者には，パラメータに関する試行錯誤のコストが減ることを期待する．また，サイコロを振るような不確定要素を含むゲームにおいても，AlphaZero の適用に成功したことの示しに貢献した．
さらに，AlphaZero で学習した戦略と局面評価の質がいいことを示したことで，それらの戦略や局面評価の参考価値をより深めた．人間プレイヤ（特に強いプレイヤ）の上達に利用できることを考える．利用価値を深めたことは学術的意義にも社会的意義にも貢献したと考える．

研究成果の概要（英文）：AlphaZero outperformed professionals by learning from scratch based on self-play games, which only needed to know game rules. However, it is unclear whether AlphaZero can learn the optimal policies or theoretical values. In addition, there are only a few applications to games involving uncertainty.
This research targeted games on small scales at first, where each position's optimal policy and theoretical value can be obtained. The results showed that the learning of AlphaZero under many settings could converge to the optimal policies or theoretical values. In addition, for a game on a larger scale and involving uncertainty, it was also confirmed that the program based on AlphaZero was strong enough to obtain the silver medal in a tournament.

研究分野：ゲーム情報学

キーワード：AlphaZero　ゲームの解析　最適戦略　理論値　Tabular　ニューラルネットワーク　完全情報ゲーム　確率的なゲーム

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

(1)　AlphaZero

囲碁における AlphaGo Zero の成功を踏まえ，DeepMind 社はそれを汎用化した強化学習アルゴリズム AlphaZero を開発し，囲碁・将棋・チェスに適用した[1]．AlphaZero は各ゲームのルールのみを知識として用い，自己対戦によってゼロからプロ棋士を上回る強さを持つことができる点で，領域知識や棋譜を用いる従来の方法と大きく異なる．そのため，AlphaZero はさまざまなゲームで強いプログラムを作ることができると期待されている．

一方で，AlphaZero が人間よりも強いプレイヤを作れることはすでに示されているが，学習した戦略（policy：次に打つべき手の確率分布）がどの程度最適に近いのか，および局面評価値（value：その局面の勝率）がどの程度正確なのか，については十分調べられていない．

(2)　対象：完全情報ゲーム

ゲームの分類法がさまざまあるなかで，最も重要なものとしては，「決定的か確率的か」および「完全情報か不完全情報か」があるだろう（表 1）．囲碁・将棋・チェスは全て，状態遷移が決定的であり，また隠れた情報がない完全情報ゲームである．一方で，カードを配る・さいころを振るなどの要素が入れば確率的ゲームになるし，ポーカーや麻雀のようにお互いの手札が非公開な場合は不完全情報ゲームになる．ゲームに限らず，現実世界の問題は不確定要素を含むことが多く，不確定要素を含むゲームの研究は重要だと考える．本研究では，不完全情報ゲームまでも将来に見据えたうえで，申請期間内では完全情報のゲームを対象とし，決定的なゲームだけではなく確率的なゲームも対象とする．

	完全情報	不完全情報
決定的	囲碁，将棋，チェス，オセロ	ガイスター，軍人将棋
確率的	バックギャモン	麻雀，ポーカー，人狼

表 1　ゲームの分類と代表例

(3)　ゲームの解析

ゲームの解析（solve）は情報学の大きな目標の１つである．どこまで解明したかによってさらに３つのレベルに分けることができる：strongly solved（完全解析），weakly solved，ultra weakly solved[2]．完全解析では，初期局面から至ることのできる局面についての理論値と最適戦略が分かる必要がある．Weakly solved では初期局面の理論値と最適戦略が分かると言う．Ultra weakly solved では初期局面の理論値のみが分かると言う（つまりどう打てばよいかわからない）．2007 年に checkers というボードゲームが weakly solved（最善でプレイしたら引き分けに至る）であると証明されること[3]は話題になっていた．より複雑なチェスや囲碁，将棋などの解析も期待されつつある．

２．研究の目的

本研究の主たる目的は，AlphaZero の枠組みを用いて，完全情報ゲームの理論値と最適戦略を学習できるアルゴリズムにすることである．単に強いプレイヤを作るだけではなく，学習した戦略が真の最適戦略に収束し，また学習した局面評価値が真の理論値に収束することが示せれば，それはゲームの解析につながると考える．

盛んに研究されている，完全情報で決定的なゲームに加え，本研究では不確定要素を含む，確率的なゲームも対象とする．決定的なゲームとの最大の違いは理論値のとりうる範囲である．決定的なゲームの局面の理論値は勝ち，負け，又は引き分けの 3 値に過ぎないが，確率的なゲームの場合は 0 から 1 までの連続的な期待勝率を求めなければならない．確率的なゲームにおいて理論値と最適戦略を完璧に学習できるかについての解明も本研究の目的である．

３．研究の方法

(1)　全体像

研究方法の全体像を表 2 に示す．表 1 の分類とは別に，ゲームはその規模によって「完全解析が可能なもの」「困難なもの」に分けることができる．前者では全ての局面において最適戦略と理論値を得ることができるため，AlphaZero の学習によって得た戦略と局面評価値が正確かを厳密に検証できる．また AlphaZero にはさまざまなパラメータが含まれる．異なるパラメータ設定が学習への影響を調べていく．

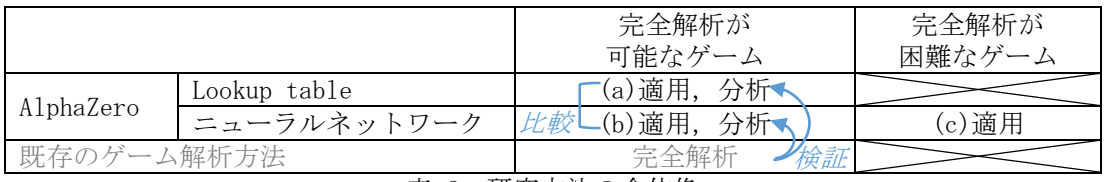

表 2　研究方法の全体像

(2)　研究の流れ

本研究ではまず，(a) 完全解析が可能なゲームを対象に，lookup table を用いた AlphaZero の性能評価を行い，異なるパラメータ設定が学習への影響を調べた．小規模なゲームでも，パラメータを任意に設定した AlphaZero が最適戦略と理論値を学習できる保証はない．学習された結果と完全解析の結果との差を分析し，パラメータが AlphaZero 学習へ与える影響，および各パラメータの適切な値の範囲を明らかにした．

続いて全ての局面を列挙せず，(b) ニューラルネットワークを特徴抽出モデルとして用い，汎化を行って学習した．特徴抽出モデルによる汎化を行う場合に，学習した局面評価や戦略は理論値や最適戦略との差がどの程度になるのかを明らかにした．

最後に (c) AlphaZero を完全解析が困難なゲームに適用し，確率的なゲームにおいても，大会で準優勝レベルの強さを持つことを示した．申請時点では df-pn(depth-first proof-number) 探索やαβ法などの手法を用いて一部の終局に近い局面を解析し，AlphaZero の学習結果と比較する予定があったが，(a) の分析は予想より深く工夫したので，この部分を割愛した．

(3)　対象ゲーム

本研究では多様な特徴を持つゲームを横断的に扱った．中心となったのは，台湾で人気がある完全情報で確率的なゲーム，Chinese dark chess（通常遊ぶ場合は 4×8 盤面）である．小規模化された 2×4 の変種は完全解析されている[4]．もう一つの確率的なゲームの EinStein Würfelt Nicht!（通常遊ぶ場合は 5×5 盤面 6 駒，4×4 盤面 6 駒は完全解析できる）[5]および決定的ゲームの NoGo（囲碁の変種，通常は 9×9 盤面，4×5 は完全解析できる）[6]も対象とした．また AlphaZero の適用はまだであったが，ニューラルネットワークの性能調査に関して TUBSTAP というターン制戦略ゲーム[7]も調べた．

４．研究成果

本研究の成果は，大まかに「完全解析可能な規模の小さいゲームでの分析」と，「規模の大きいゲームへの適用や関連する調査」に分けることができる．中では前者に多くの時間を割いた．

(1)　規模の小さいゲームでの分析

①　Lookup Table

AlphaZero にはたくさんのパラメータを持ち，設定によって学習の結果も変わっていく．パラメータ設定が学習に与える影響を調べるために，まずは先行研究[8]に引き続き，局面を全列挙する lookup table を用いた AlphaZero の分析を行った．先行研究の実験と分析には自作の枠組みを用い，その枠組みは 2×4 Chinese dark chess（以下 CDC で略す）に特化したものであった．本研究では，その枠組みをほかのゲームにも使えるような改良を施した．各ゲームにとって共通の要素（モンテカルロ木探索や自己対戦など）を抽出し，別のゲームに取り替える際には，該当ゲームに関する部分（ルールなど）を実装すれば良いように設計した．2×4 CDC のほか，小盤面の EinStein Würfelt Nicht!（以下 EWN で略す）と NoGo を実装した．また 2×4 CDC の完全解析結果についてはすでに持っていたが，EWN と NoGo にはなかったため，この２つのゲームにおいて，いくつかの盤面サイズの設定も完全解析した．分析したゲームの設定は以下の通りである：2×4 CDC の「PPPP 対 pppp」「KPPP 対 kppp」「GGCC 対 ggcc」という駒組み合わせ，EWN の「3×3 盤面 3 駒」「3×3 盤面 4 駒」「3×4 盤面 3 駒」という設定，NoGo の「1×12」「2×6」「3×4」「1×13」「1×14」という盤面サイズ．

AlphaZero のパラメータの中に，モンテカルロ木探索に関連するものを中心に調べた．AlphaZero のモンテカルロ木探索は「選択」・「展開」・「更新」という３つのステップを繰り返し，探索木を構築するアルゴリズムである．その繰り返しの回数を「シミュレーション回数」と言う．またステップ１の「選択」では探索木の葉節点まで PUCT 点数 $W(s,a)/N(s,a)+c_{puct}\times P(s,a)\times\sqrt{N(s)}/(1+N(s,a))$ が最も大きい着手を選ぶ．各関数は以下の通りである：W(s, a) は状態 s で着手 a を取った際の総報酬，N(s, a) は状態 s で着手 a を選んだ回数，c_{puct} は定数，P(s, a) は lookup table やニューラルネットワークから得た，状態 s で着手 a を選ぶ確率，N(s) は状態 s の訪問回数．W(s, a) と N(s, a) はゼロに初期化されるが，0/0 は未定義な数値なので，特別の扱いが必要であり，AlphaZero の論文では 0（負け）として計算された．学習中に各着手に試される機会を与えるため，ルート節点の $P(s_{ルート}, a)$ も工夫された．具体的には $P(s_{ルート}, a)=(1-\varepsilon)\times p_a+\varepsilon\times\eta(\alpha)$ に変更された．$\eta(\alpha)$ は分散度αの対称 Dirichlet 分布からサンプリングしたノイズ，εはノイズの重みを決める定数，p_a は lookup table やニューラルネットワークから得た，状態 $s_{ルート}$ で着手 a を選ぶ確率である．本研究では (i) c_{puct}，(ii) Dirichlet α，(iii) Dirichlet ε，(iv) W/N 初期値，(v) シミュレーション回数，という５つのパラメータを調べた．各パラメータの適切な範囲を見つけるため，設定値の範囲は広く設計した．

(i) の c_{puct} は探索中に exploration の程度を決めるパラメータである．言い換えれば，c_{puct} が小さい場合には期待報酬 W(s, a)/N(s, a) の高い着手が選ばれやすく，逆に c_{puct} が大きい場合には訪問回数 N(s, a) の小さい着手が選ばれやすい．分析の結果からは，良さそうな c_{puct} の範囲 (0.5~4) がある程度大きいであることがわかった．c_{puct} が小さすぎる場合には lookup table のランダムな初期値に偏り，うまく学習できないことが起こりうることを示した．一方

で，c_{puct} が大きすぎる場合には，自己対戦ゲームはランダムに着手を選ぶようなゲームになってしまい，ランダムプレイに近い戦略や局面評価を学習してしまうことも示した．

（ii）の Dirichlet α は AlphaZero の論文で，適用したゲーム（チェス・将棋・囲碁）によって唯一の特別に調整したパラメータであった（それぞれが 0.3，0.15，0.03）．しかし本研究での，規模の小さいゲームの実験では 0.03~12 という幅広い範囲であっても，学習結果への影響は非常に限られた．可能な理由としては，平均合法手が少ない場合には Dirichlet α は着手の選択への影響が小さいということを考えられる．

（iii）の Dirichlet ε は Dirichlet ノイズの重みを決める 0~1 の実数値定数である．分析の結果では，「3×3 盤面 3 駒の EWN」のような小さいゲーム以外に，Dirichlet ノイズを使うこと(つまり $\varepsilon > 0$)の有効性を示した．$\varepsilon \in [0, 0.5]$ の設定なら，ε が大きくなると exploration 程度が大きくなったが，c_{puct} の結果に比べて学習への影響が小さかったことがわかった．

（iv）の W/N 初期値は，一般的なモンテカルロ木探索には重要な課題であるが，AlphaZero に関する研究ではあまり議論されていない．本研究では AlphaZero 論文で使われた 0（負け）のほか，0.5（引分），1（勝ち），∞（選ばれていない着手が優先）という３つも調べた．分析の結果では 0.5 のほうが良さそうであることがわかった．なお，並列化モンテカルロ木探索を用いた AlphaZero 論文と異なり，本研究ではシングルスレッドのモンテカルロ木探索を用いることで差が出たと推測している．

（v）のシミュレーション回数は，単に強くプレイしたい場合に，シミュレーションが多いほど良いというのがモンテカルロ木探索の本質である．しかし，モンテカルロ木探索でプレイした自己対戦ゲームによって学習する AlphaZero には，シミュレーションが多いほど良いとは限らない．少なくとも，シミュレーションが多いほど計算資源がかかるという点は良くないとも言える．本研究ではシミュレーションが数回程度（6 や 12）から数万回程度（12,800 や 25,600）ある設定を調べた．分析の結果では，シミュレーション回数がある程度十分であれば，うまく学習できることを示した．例えば「3×3 盤面 3 駒の EWN」のような小さいゲームにはシミュレーション 12 回は十分であった．シミュレーションをこれ以上多くすることは，性能が少し良くなることもあったが，差が非常に限られた．計算資源が無駄になる可能性が高いとも言える．

② ニューラルネットワーク

Lookup table の結果を踏まえ，次には元の AlphaZero と同様のニューラルネットワークを用いることにした．この部分には，台湾国立陽明交通大学情報工学系呉毅成研究室で開発された CLAP という枠組み[9]を用いた．その枠組みは最初は決定的なゲーム向けのものなので，まずは確率的なゲームにも使えるように拡張した．

2×4 CDC の「PPPP 対 pppp」において lookup table と同様に分析した．ニューラルネットワークは特徴抽出を行うため，似たような局面をまとめて学習できるという利点を持つので，学習に必要な自己対戦ゲームの数は lookup table の場合より少ない．分析の結果では，ニューラルネットワークは lookup table とだいたい似たような傾向が見られた．つまり，適切な c_{puct} の範囲はある程度大きいことや，W/N 初期値は 0.5 のほうが良いことなどはだいたい同じであった．

(2) 規模の大きいゲームでの試み

2×4 CDC の CLAP 枠組みでの実装に基づき，通常のサイズの 4×8 CDC の学習を試した．Intel (R) Xeon (R) Silver 4210R CPU @ 2.40GHz と NVIDIA Quadro RTX 8000 GPU 搭載のサーバで，おおよそ１週間の学習を行った．学習がうまくいったかを確認するため，手元に持つ，強いプログラムである DarkKnight（モンテカルロ木探索系）と対戦させた．２百程度のゲームでの勝率が 58%くらいあったので，DarkKnight を上回ったことを示した．また，2021 年 8 月開催の Computer Olympiad CDC 大会と 2021 年 11 月開催の TAAI Cup CDC 大会に，CLAP_CDC というプログラム名で参加し，それぞれの大会で銅メダルと銀メダルを獲得した．

TUBSTAP というターン制戦略ゲームにおいて，局面評価用のニューラルネットワークを学習する試みがあった．AlphaZero 方式の自己対戦ゲームによる強化学習ではなく，事前に用意した棋譜を学習する教師あり学習であったが，この部分の主な目的は，異なる種類のニューラルネットワークの比較であった．具体的には，従来の畳み込みニューラルネットワークと，比較的新しいグラフニューラルネットワークを比較し，後者のほうが優れることを示した．AlphaZero の学習においても，畳み込みニューラルネットワークの代わりに，グラフニューラルネットワークを用いることで，良い性能が出ることを期待している．

ガイスターという不完全情報ゲームを完全情報化し，終局に近い局面を解析した．この部分の主な目的は，詰めガイスターというパズルを作成し，人間プレイヤを楽しませる・勉強させることであったが，用いた解析アルゴリズムの df-pn を把握したことは，本研究の成果として挙げられると考える．

(3) 位置づけと今後の展望

Haque ら[10]はチェスにおいて，AlphaZero の学習結果を，終局データベースの結果と比較・分析した．彼らの結果では本研究と同様に，学習した戦略は最適である場合が多かったことを示した．学習した戦略と局面評価だけを分析した本研究と異なり，彼らの研究では，さらにモンテ

カルロ木探索と併用した場合の戦略を分析した．ニューラルネットワークで学習した戦略（つまり木探索なし）の場合が最善手を選んだのに，木探索では逆に最善手を選ばないようになった局面もあったという結果は興味深い．また本研究では決定的なゲームだけではなく，確率的なゲームも対象とした点は，決定的なゲームのみを分析した彼らの研究と異なる．

AlphaZero の改良版がいくつかあるなかで，Danihelka ら[11]の Gumbel AlphaZero は少ないシミュレーションでもある程度うまく学習できることを示した（例えば９路盤囲碁に対してシミュレーション２回）．今後の研究課題の１つとして Gumbel AlphaZero の分析を考える．

ゲームの解析について，AlphaZero による完全解析は困難であることがわかった．理由としては，AlphaZero の学習では重要ではない局面はだんだん無視され，それらの局面をうまく学習できないからである．Weakly solved （初期局面の理論値と最適戦略がわかること）ならまだ可能だと考える．Gao ら[12]はニューラルネットワークで学習した戦略と局面評価を，df-pn と組み合わせてゲームを解析するという発想があった．しかし，彼らのニューラルネットワークは AlphaZero で学習したのではなく，教師あり学習によるものであった．Wu ら[13]は AlphaZero を，ゲームの解析をするように改良した．Gao らと Wu らの手法は決定的なゲーム向けのもので，確率的なゲームの解析には適用できないため，改良あるいは新しい手法の提案を今後の研究課題の１つとして考える．

＜引用文献＞

[1] Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., . . . Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 352 (6419), 1140-1144.

[2] Van Den Herik, H. J., Uiterwijk, J. W., & Van Rijswijck, J. (2002). Games solved: Now and in the future. *Artificial Intelligence*, 134(1-2), 277-311.

[3] Schaeffer, J., Burch, N., Bjornsson, Y., Kishimoto, A., Muller, M., Lake, R., ... & Sutphen, S. (2007). Checkers is solved. *Science*, 317(5844), 1518-1522.

[4] Chang, H. J., Chen, J. C., Hsueh, C. W., & Hsu, T. S. (2018). Analysis and efficient solutions for 2×4 Chinese dark chess. *ICGA Journal*, 40(2), 61-76.

[5] Bonnet, F., & Viennot, S. (2017). Toward solving “EinStein würfelt nicht!”. In *Advances in Computer Games* (pp. 13-25). Springer, Cham.

[6] She, P. (2013). The designed and study of NoGo program (Master's thesis, National Chiao Tung University, Hsinchu, Taiwan). Retrieved from: http://hdl.handle.net/11536/73344

[7] 佐藤 直之, 藤木 翼, & 池田 心. (2016). 戦術的ターン制ストラテジーゲームにおける AI 構成のための諸課題とそのアプローチ. *情報処理学会論文誌*, 57(11), 2337-2353.

[8] Hsueh, C. H., Wu, I. C., Chen, J. C., & Hsu, T. S. (2018). AlphaZero for a non-deterministic game. In *2018 Conference on Technologies and Applications of Artificial Intelligence* (*TAAI*) (pp. 116-121). IEEE.

[9] Chen, Y. H. (2020). A high-performance, distributed AlphaZero framework (Master's thesis, National Chiao Tung University, Hsinchu, Taiwan). Retrieved from https://hdl.handle.net/11296/64c8e8

[10] Haque, R., Wei, T. H., & Müller, M. (2021). On the road to perfection? Evaluating Leela Chess Zero against endgame tablebases. In *the 17th Conference on Advances in Computer and Games* (*ACG*).

[11] Danihelka, I., Guez, A., Schrittwieser, J., & Silver, D. (2021). Policy improvement by planning with Gumbel. In *International Conference on Learning Representations* (*ICLR*).

[12] Gao, C., Müller, M., & Hayward, R. (2017). Focused depth-first proof number search using convolutional neural networks for the game of hex. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence* (pp. 3668-3674).

[13] Wu, T. R., Shih, C. C., Wei, T. H., Tsai, M. Y., Hsu, W. Y., & Wu, I. C. (2022). AlphaZero-based proof cost network to aid game solving. In *International Conference on Learning Representations* (*ICLR*).

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計2件（うち招待講演　0件／うち国際学会　2件）

１．発表者名 Wanxiang Li, Houkuan He, Chu-Hsuan Hsueh, and Kokolo Ikeda
２．発表標題 Graph Convolutional Networks for Turn-Based Strategy Games
３．学会等名 The 14th International Conference on Agents and Artificial Intelligence（国際学会）
４．発表年 2022年

１．発表者名 Chu-Hsuan Hsueh, Kokolo Ikeda, Sang-Gyu Nam, and I-Chen Wu
２．発表標題 Analyses of Tabular AlphaZero on NoGo
３．学会等名 2020 International Conference on Technologies and Applications of Artificial Intelligence（国際学会）
４．発表年 2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

2021年8月開催された第24回国際競技会（Computer Olympiad）Chinese dark chess 大会で六つのチームの中に銅メダルを獲得した. 2021年11月開催された TAAI2021 Chinese dark chess 大会で五つのチームの中に準優勝を獲得した.

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関			
その他の国・地域　台湾	国立陽明交通大学	中央研究院	国立台北大学	