—論文題目—

ああああを用いた
ああああ予測

Hoge Hoge Hoge

指導教授

萩原 将文 教授

慶應義塾大学 理工学部 情報工学科

令和 99 年度

学籍番号 12345678

田中 太郎

目 次

あらまし

本論文では木探索を用いた強化学習アルゴリズムに対して判断の可視化を行った。具体的にはゲームの中で重要度が高いとされる地点を検出し、その地点を親としてそこから派生する決定木中の先読みを結果に基づいてグループ化し、最終的に最も多数派となったグループの分岐をユーザーに提示することで決定木内の傾向を示した。本論文におけるタスクは connect4、使用するネットワークは alphaZero を簡易的に模した AlphaZero baseline である。ゲームの途中の地点から最終図を予測し、その精度を調査する実験とユーザーインタフェースを用いて使用感や上達の条件等を記録する実験の二種類の実験を行った。

結果として、既存手法に対して高い予測精度や、ユーザーの満足度を出すことに成功した。

第1章

はじめに

近年のAIの発展は目覚ましく, 画像分類や異常検知などの単純なルールで記述する事が困難なタスクや、更には長らく人間に固有の技術であると考えられてきた画像や文章の生成の分野においてさえ、高い性能を発揮するまでにAI技術は成長した。特に昨年のstable diffusion[1], Instruct GPT[2]の登場により人間の生産労働の在り方, 人間とAIの関係, ひいてはこの先の社会がAIとどう付き合っていくのか,AIによってどう変わっていくのかを専門家だけでなく一般の人々も含めて考えざるを得ない段階に差し掛かっていると言える。この「優れたAIに対して人間はどう接するべきか」という命題を考える際には既に人間を大きく凌駕したAIが存在する領域において手法の構築や実験を行うのが適当である。そのため、ここではconnect4と呼ばれる比較的単純なボードゲームを題材とし、2016年に当時世界有数のプレイヤーであったイ・セドルを四勝一敗で圧倒したAlphaGo[3][4]を簡易的に模したネットワークであるAlphaZero_baseline[5]を用いて本論文を執筆した。優れたAIが社会で広く実用化され、受容されるためにはAIの判断や生成物(以下単純に「出力」という表現を使用する)が生み出される過程の透明性がいずれは必要不可欠になることが予測される。実際に画像生成AIが上述の様に一度オープンソース化された2024年現在においても多くの国ではAI生成物を市販することが法律で禁止されており[5], 市販が違法と明記してはいない日本においてもAI法の議論は活発に行われている。また、そのような法的・倫理的観点によるAIの透明性への希求だけでなく「どうすれば人間もAIのような成果を生むことができるのか」という探求心や学習意欲から成る説明性へのニーズが広く湧き起こる事が予測される。特にゲームのように

「AI 対人間」と表現すべき対立構造を強く有する領域においては後者のニーズがより大きくなっていくものと予測される。現に囲碁, 将棋, チェスなどの主要なボードゲームをオンラインでプレイできるサービスにはゲーム終了後の振り返り (感想戦) において AI の判断や AI による想定図を閲覧できるサービスが数多く提供されている。研究においてもボードゲームの学習支援という形で「AI に学ぶ」試みが存在する。しかし、後述するようにそれらの研究は各ゲームのドメイン知識等を使用した、従来の人間による指導方法の自動化に近い形態のものが多い。ここでは人間の学習のサポートツールとして AI を用いるというよりは、AI の動作を人間のユーザー向けに説明する手法を構築し、人間に AI の挙動を理解してもらうことを通じて人間側のスキルを向上させる試みを行った。本論文における提案はゲームの勝敗を左右する地点を検出する指標の定義、そして適切な予測を決定木から取り出す手法の二つに大別される。また、実験も提案手法を用いてゲームの終了状態を予測する実験と、ユーザーの使用感や学習効果を調査する実験の二つを行った。

第2章

関連研究

この章ではまず、既存のボードゲーム AI について AlphaZero を中心に強化学習的枠組みからその理論を説明する。次に Alpha Zero の問題点とそれを補完する既存手法とその課題について述べる。

2.1 強化学習

強化学習はタスクを主体と環境のやり取りとして定式化する形でタスクに取り組む分野である。状態 (s) と行動 (a) が次の状態 s' と環境から与えられる報酬 r が決定されると仮定する。その仮定の下、環境から与えられる報酬の合計 (以下収益と記載) を最大化する。報酬を大きくするためには状態 s に応じて適切な (より大きな報酬をもらえる可能性が高い) 行動を選択する必要がある。ある状態である行動をとった場合の収益に対して見積もりをとり、見込まれる値が最も大きい行動を選択することでより大きな収益を獲得できると期待できる。このようなある状態である行動を取った場合の収益の見積もりを $Q(s,a)$ とした場合、

$$a = argmax_{a'}Q(s,a') \tag{2.1}$$

となる a 選択することによって収益の最大化が期待される。また、ある状態から獲得できる収益の合計の予想値 $V(s)$ は、最適な行動 a を取った場合の値として推定される。

$$V(s) = Q(s,a)(a = argmax_{a'}Q(s,a')) \tag{2.2}$$

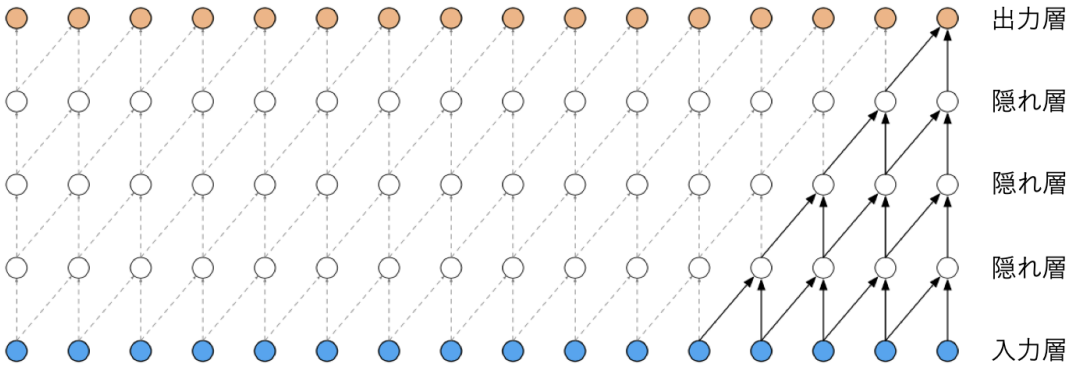

図 2.1: AlphaZero のチェスのやつと connect4 と強化学習の式を合わせていい感じに

強化学習手法によってタスクの最適化を図る際にはこの $V(s), Q(s,a')$ を正しく推定することが直接的な目標となる。$V(s), Q(s,a)$ は主体が実際に環境とやり取りを行う（タスクを実行していく）中で改善されていき、Temporary Difference 法 [5] や Monte Carlo 法 [5] 等が基本的な $V(s), Q(s,a)$ の更新則である。また、DQN[5] や Rainbow[5] 等はニューラルネットワークを使用して $V(s), Q(s,a')$ を推定することでより高い性能を発揮している。

2.1.1 ボードゲームへの応用

ボードゲームでは通例、状態 s は盤面の状況、行動はプレイヤーの選択、報酬はゲームの最後に勝敗として与えられる。状態 s(ゲームの状況) と行動 a(プレイヤーの選択) によって盤面は次の状態 s' に遷移し、次の行動 a'(他のプレイヤーによる選択) を受け付ける、というサイクルにゲームの進行を定式化して表現することができる。また、上述した強化学習における $V(s), Q(s,a)$ の推定はそれぞれ「ある盤面はプレイヤーにとって勝利に近いのか」、「ある盤面においてある選択をした場合、プレイヤーはどれ程勝利に近くなるのか」を表現していると解釈される。AlphaGo[5] やチェスのなんか強いやつ [5]、激指 [5] では状態 s は最新 N ステップの盤面である。盤面は行列に抽象化される。また、行動は次にプレイヤーが打つ箇所の座標となる。本論文で使用した alphazero baseline

における入力は最新の盤面の状態を空白を 0, 先番 (赤) の石の位置を 1, 後ろ番（黄）の位置を-1 として抽象化した 6×7 の行列となる。また、後述する connect4 のルール上の制約により次にプレイヤーが打つ箇所 (行動) は列の数と同数の 7 つに限定される。

2.1.2 AlphaZero

AlphaZero は 2016 年に登場し、元世界チャンピオンであるイ・セドルに対して四勝一敗の成績を収めた AlphaGo の汎用版である。AlphaZero は先述の $V(s), Q(s,a)$ を推定する際にニューラルネットワークとモンテカルロ木探索システムを使用する。

2.1.2.0.1 ニューラルネットワーク AlphaZero 内のニューラルネットワークに対する入力は最新 N ステップの盤面 ($\{s_{-N+1}, ..., s_0\}$,s_{-i} は i ステップ前の盤面,s_0 は現在の盤面) であり、出力は方策 $P(\{s_{-N+1}, ..., s_0\})$ と局面評価（後で確認）$V(s_0)$ の二種類である。 ネットワークの構成は 層の残差結合ネットワークである。方策は「現在の状況 s_0 から次にどこを選択すべきか」を表現しており、次に選択すべき座標を確率分布の形式で表現する。alphazero baseline における選択肢は列の数と等しい 7 であるため、1×7 の行列となる。方策内の値が大きさが AI によるその着手の評価と解釈され、成分が大きい座標を次に選択することが推奨される。例えば方策が $\{0, 0.1, 0.2, 0, 0, 0.7, 0.8\}$ であるとき、方策中の最も大きい成分は 7 番目の 0.8 であるため、プレイヤーは次に 7 列目を選択する事が推奨される。また、局面評価 $V(s_0)$ は「現在の状況 s_0 は勝利に近いのか」を表現しており、値が上限に近ければ近い程、現在の状況 s_0 が次の着手を選択するプレイヤーにとっての勝利に近いことを表している。

訓練とかどうするん？

2.1.2.0.2 モンテカルロ木探索 モンテカルロ木探索ではニューラルネットワークから得た方策 $P(s)$(以下 $s = \{s_{-N+1}, ..., s_0\}$ とおく) と局面評価 $V(s)$ をシミュレーションによって改善する。モンテカルロ木探索ではシミュレーショ

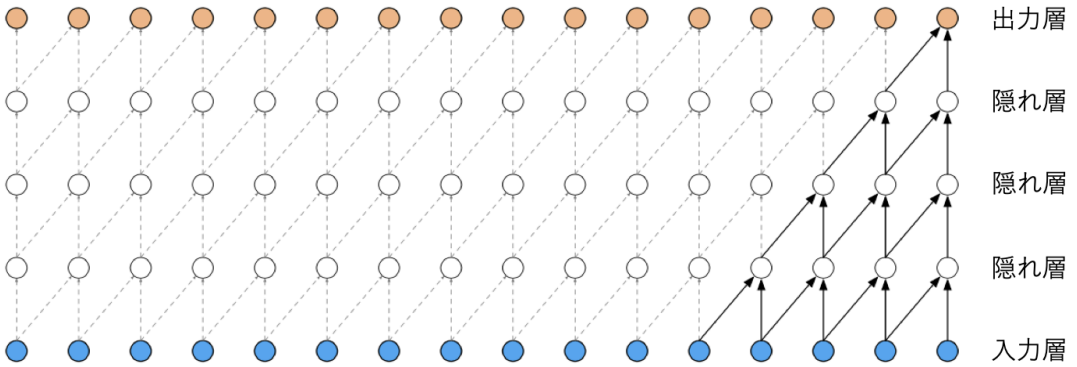

図 2.2: 方策と局面評価、あと残差結合の図も

ンによって各状態 s をノード、各行動 a を枝とした決定木を構築する。最終的に各ノード s から派生する各行動 a の分布 $\{a_0, a_1, ..., a_n\}$ が改善された方策となる。一方、局面評価もまたモンテカルロ木探索により決定木が拡張されるなかで更新される。以下に決定木と局面評 $V(s)$ の更新アルゴリズムを示す。表アルゴリズム 4.1 step1 ではまず、探索の開始地点となるノード s を決定する。次に Step2:再帰部分では以下の処理を再起的に呼び出す。

1. ノードを探索したことがない場合ニューラルネットワークから出力された方策 $P(s)$ と局面評価 $V(s)$ を返却する

2. ノードを探索したことがある場合以下の puct スコアに従い、子ノード s_c を選び、s_c に対して探索を行いその結果である $P(s_e), V(s_e)$(s_eは再帰処理の結果たどり着く決定木の端のノード) を受け取る。

$P(s_e), V(s_e)$ を用いてパラメータをこんなふうに更新する。
このようなプロセスによって構築された決定木を用いてモデルは対戦を行う。

2.1.3 AlphaZero の問題点

StockFish[5] やなんか [5] などの従来のボードゲーム AI はそのゲーム固有の知識 (ドメイン知識) に基づくものが多く、「ある条件を満たすときにある選択をする」と言ったようにその挙動をルールとして表すことが可能である。一方

でこのニューラルネットワーク+木探索の手法では人間がネットワークから得られる情報は方策と局面評価のみである。その他のパラメータも観察可能ではあるが、その方策や局面評価の根拠を得ることができない、つまり AlphsZero の問題点として説明性の欠如が挙げられるのである。説明性の欠如は AI の判断に対する責任の不在を意味し、優れた性能を持つシステムのより、ハイレベル、ハイリスクなタスクへの実用化に対する障害となる。

2.2 XAI

2.2.1 概要

XAI とは explainable AI(説明可能 AI) の略語であり、AI を人間に対して説明可能なものにする、もしくは説明可能な AI を構築する領域である。本論文は AI の判断根拠として先読みを示す意味で XAI の分野に属する研究であると言える。ここでの「説明可能性」の語は「人間に理解できる形での説明を与える能力」[5] と定義され、「いつ、どのような、どのように」説明を与えるかによってさらに細かく分類される。「いつ」、つまりどの時点で説明を与えるか、に関しては既存のネットワークに対して新たに説明を加える「事後的」説明と初めから動作の根拠を示せるようにネットワークやシステムを構築する「事前的」説明に分類できる [5]。「どのような」、つまり説明の内容についてはが存在する。また、「どのように」つまり説明を表現する形態としては saliency map, Grad-CAM[5] といった視覚的な可視化や、あとで [5] といった文章生成等が存在する。本論文において構築するシステムは「事後的」「局所的」「視覚的」説明を提供する。

Wavenet[5] は音声波形を時系列データとして自己回帰モデルで学習することによって，人間の声のような自然な音声を生成することができる．時点 t における観測値を x_t，$\boldsymbol{x} = \{x_1, ..., x_T\}$ を観測値の全体集合とする．このとき，波形の同時確率は条件付き確率の積として以下のよう表現される．

$$p(\boldsymbol{x}) = \prod_{t=1}^{T} p(x_t | x_1, ..., x_{t-1}) \tag{2.3}$$

つまり，x_t は前時点の全てにおけるサンプルに条件づけられる．また、本論文はXAI分野の中でも特に強化学習方面に対して説明を加える領域をXRL(explainable Reinforcement learning) と呼ぶ。XRL の試みは様々な強化学習システムを対象とし、

2.2.2 ボードゲームにおける XAI

グーグルのやつはチェスにおける人間の知識が AlphaZero にどれだけ反映されているかを訓練段階やネットワークの深さなどの多様な指標で調査している。[5] ディコードチェス [5] は AI の着手に対して、そのゲーム固有の知識（以下ドメイン知識と記載する）を用いて解説を生成するしかし、このようなドメイン知識が必ずしも AI の挙動と相関が無いことも指摘されている。[5] AI による画像分類の可視化手法であるサリエンシーマップ [5] や Grad_CAM[5] を強化学習に用いる例も存在する。しかしそれらのニューラルネットワークの活性を根拠とした指標は木探索部分との繋がりが弱く、最終的に決定木を用いて意思決定を行うシステムの動作根拠を直接的に説明できない。また、これらの画像分類用の手法は

- 本来ゲームに存在する時系列の要素を説明に含められない
- 時系列を無視してゲーム画面や盤面の一部を変更する必要がある

という問題点が存在する。

2.2.3 contrastive explanation

上述の問題点を解決するために、本論文では AlphaZero が構築する決定木を用いた contrastive explanation（対象説明、比較説明）を提供する。contrastive explanation は事象を説明する際の方法論の一つであり、ある事象 a が起こった際にその理由を直接説明する代わりに「他の事象 $\bar{a}$ が起こらなかった理由」を説明することで間接的にある事象 a の原因を説明するアイデアである。では言

語の何か [5] ではニューラルネットワークの判断を決定木に近似した上でニューラルネットワークの判断 a から派生する予想と別の判断 b から派生する予想を同時に提示する。ロボット [5] の、ではある時点 t で異なる行動を選んだ場合の結果の違いを文章で説明するアミア [5] では異なるシステム間の特徴をユーザーに示す目的で contrastive explanation が用いられている。

2.2.4 ボードゲーム学習支援

本論文が提案する手法は高いパフォーマンスを発揮する AI の動作を人間に理解させることを目標としており、学習支援の面を持つ。既存の AI を用いたボードゲーム学習支援システムとしてはドメインみたいなやつをいくつかが存在し、決定木の先読みを用いて文章生成や解説生成を行う。しかし「ボードゲームに対する XAI」の段での内容と同様にその多くがゲームのドメイン知識に依存しており、指導の内容も人間の知識に依存したものになってしまうという欠点がある。

第3章

提案手法

関連研究の章ではニューラルネットワークや決定木が内包する説明性の欠如という問題点と説明を加える既存手法の持つ課題について述べた。本論文ではそれらを踏まえた

- 本来ゲームに存在する時系列の要素を含む
- 評価基準が勝敗に直結する
- 人間のドメイン知識に依存しない

説明手法を提案する。

3.1 時系列予測

hoge は今まで多く行われてきた．[6] はあああ．

第4章

評価実験

4.1　実験条件の設定

4.1.1　データセット

為

4.1.2　比較手法

一

4.2　あああの予測

4.2.1　実験方法

2017年

4.2.2　実験結果

表4.1に

表 4.1: あああといいいの予測誤差

	2019		2018		2017	
モデル	ああ	いい	ああ	いい	ああ	いい
Naive	**1**	1	**1**	1	**1**	1
TCN	1.0895	0.9032	1.4791	**0.9198**	1.2888	0.8555
LSTM	1.0384	0.9295	1.4917	0.9725	1.1627	0.8541
提案手法	1.0977	**0.8698**	1.3824	0.9439	1.2061	**0.8516**

第5章

結論

本論文では
今後の課題を以下に挙げる.

- の向上

  必要がある.

- への応用

  を行いたい.

- の改善

  今後，取り組みたい．

謝辞

本研究を行うにあたり親身に相談に乗っていただき，ご指導してくださった萩原将文教授，ならびに共に問題解決，議論，相談, および実験に付き合ってくださった研究室の先輩方，同期の皆様, 実験に参加してくださった大学の友人達に深く感謝いたします. 誠にありがとうございました.

参考文献

[1] Andreas Blattmann Robin Rombach, Patrick Esser Dominik Lorenz, and Björn Ommer. “High-resolution image synthesis with latent diffusion models”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2021).

[2] Jeff Wu Long Ouyang et al. “Training language models to follow instructions with human feedback”. In: *Advances in Neural Information Processing Systems 35* (2022).

[3] et al. SILVER David. “Mastering the game of Go with deep neural networks and tree search”. In: *nature 529.7587* (2016).

[4] “AI に敗れた李九段、「アルファ碁」に教わったこと”. In: **日本経済新聞** (2016).

[5] Aaron van den Oord et al. “Wavenet: A generative model for raw audio”. In: *arXiv preprint arXiv:1609.03499* (2016).

[6] Biing Hwang Juang and Laurence R Rabiner. “Hidden Markov models for speech recognition”. In: *Technometrics* 33.3 (1991), pp. 251–272.

付録 A

AlphaZero モデルの訓練

A.1　ヒストリカルデータ

為

付録Ｂ

実験結果詳細

B.1　の予測

第

付録C

のモデル化

C.1　異なる

あ

付録D

のヒストグラム

D.1　異なる期間

図