Vol.2022-GI-47 No.17
2022/3/19

方策勾配法による 協力型不完全情報ゲーム Hanabi の戦略学習

比企野 純一[†1,a)]　鶴岡 慶雅[†1,b)]

概要：近年，AI に関する研究分野では複雑な環境を主な対象として研究が進み，その中でもマルチエージェントシステムを対象とする強化学習分野は最近注目されている．交通システムや経済状況，選挙投票など様々な状況がマルチエージェントの相互作用によって形成されており，これらの問題の解決方法の一端を強化学習によって担う事を期待されてマルチエージェント強化学習の研究が推し進められている．マルチエージェント強化学習の中でも協力型マルチエージェント強化学習のベンチマークとして最近整備されているボードゲームが不完全情報ゲーム Hanabi である．Hanabi の研究では一般的に部分観測かつマルチエージェントでありサンプル効率が悪いことを理由として，通常の設定のオンポリシー手法では学習がうまくいかず，一般的にはオフポリシー手法が用いられてきたが，近年設定を変更することによりオンポリシー手法においてもサンプル効率が落ちずに同程度の結果を残す研究も現れた．ここで重要な部分がパラメータ調整によって学習の効率を向上させたことである．そこで本研究は協力型不完全情報ゲーム Hanabi に対してさらに詳しく学習の鍵となっている設定を分析し，確認した．

Reinforcement Learning of cooperative incomplete information game, Hanabi by policy gradient method

Abstract: In recent years, research in the field of AI has focused on complex environments, and the field of reinforcement learning for multi-agent systems has recently attracted much attention. Various situations, such as traffic systems, economic situations, and election voting, are formed by the interaction of multiple agents, and research on multi-agent reinforcement learning is being promoted with the expectation that reinforcement learning will play a part in solving these problems. A board game called Hanabi, recently developed as a benchmark for cooperative multi-agent reinforcement learning, is an imperfect information game. In Hanabi's research, on-policy methods in the usual setting do not work well due to the low sample efficiency of partial observations and multiple agents, and off-policy methods have generally been used.However, in recent years, some studies have shown that on-policy methods can achieve the same level of results without losing sample efficiency by changing the settings. The important part here is that the learning efficiency is improved by adjusting the parameters. In this study, we analyzed and confirmed the key learning settings for Hanabi, a cooperative imperfect information game, in more detail.

1. はじめに

近年，強化学習は自動運転や自動制御ロボット，トッププロに対してボードゲームで勝利を収めるゲーム AI など様々な分野で活躍している [17] [13] 技術である．強化学習とは機械学習の一種であり，学習の対象となる環境と自律的に行動するエージェントとの相互作用の中で，エージェントは自分の行動に付随して環境から与えられる報酬を基に最適な行動方策を学習するような枠組みである．強化学習の利点は，環境の状態遷移規則が未知で，人間による教師データやドメイン知識などの事前知識が無いような場合においても，エージェントは報酬を最大化するような行動方策を学習できるという点にある．特にその強化学習分野の中でも現実世界の複雑な問題を解決するためにマルチエージェント強化学習に興味が集まっている．[14] 現実世界においては経済市場や交通状況から選挙投票など様々な状況がマルチエージェントの相互作用によって形成されており，これらの問題の解決方法の一端を担うことを期待さ

[1] 情報処理学会
IPSJ, Chiyoda, Tokyo 101–0062, Japan
[†1] 現在，東京大学大学院
Presently with The Uniersity of Tokyo Graduate School
[a)] hikino@logos.t.u-tokyo.ac.jp
[b)] tsuruoka@logos.t.u-tokyo.ac.jp

れてマルチエージェント強化学習の研究が推し進められている．

マルチエージェント強化学習の最新の研究としては不完全情報ゲームの麻雀においてで，最終局面を予測して方策決定のサポートを行う機構と学習時に不完全情報の部分の情報を観測できる機構と更新されていく情報の更新を行う機構の 3 種類を兼ね備えて 99．99%のプレイヤーを上回る戦績を残した Microsoft の Super Phoenix [13] やテキサスホールデムにおいて人間のプレイヤーを圧倒した研究 [5] などが挙げられる．

そのような中現在のマルチエージェント強化学習の研究においてはベンチマークとして用いられているゲームが Hanabi である．Hanabi は協力型不完全情報ゲームであり，観測が部分観測となる．インディアンポーカーに近い自分の手札は見えず他のプレイヤーの手札が見えるような部分観測のもとで各エージェントは協力しあって全体の報酬を最大化させることを目指す．情報の明示的な共有はルールに則るもの以外は禁止であり，ルールに則って他のエージェントから与えられたヒントとそのエージェントが行動をしたという事実から推察される情報から自分の手札や取るべき行動を推測し，方策を学習する．Hanabi に対してはマルチエージェント性と部分的観測の側面からサンプル効率の側面から主にオフポリシー手法が用いられて研究されてきて，ルールベースエージェントを超えるような成果もあげられた．[4] [8] [9] またオンポリシー手法においても学習を行う際に通常の設定では学習がうまく進まないが，設定を調整することで学習をうまく進めることができるようになるということが指摘された．[20] この設定に伴う問題解決のために実際にオンポリシー手法である PPO を基にしたアルゴリズムを自作の Hanabi の環境下で学習を進める環境を構築して，各パラメータへの分析を行い結果から各パラメータが学習に寄与する影響を考察した．

2. 関連研究

一般的なマルチエージェント強化学習は完全情報ゲームの分野において成果を上げている．チェス [6]，チェッカー [15]，囲碁 [17, 18]，バックギャモン [19] など 2 人用の完全情報ゲームに対してエージェントが人類を上回る結果を残してる．一方で不完全情報ゲーム分野においても完全情報ゲームの分野の進展からやや遅れつつも研究が盛んに行われている．[5] [13]

不完全情報ゲームである Hanabi は近年注目されている協力型マルチエージェント強化学習のベンチマークである．[4]Hanabi はマルチエージェントタスクの中でも他のエージェントの方策や意図を察することで結果の向上を見込むことのできるマルチエージェントタスクとして位置付けられており，共同作業において欠かせない相手の意図を汲み取る能力を測る研究のベンチマークとして用いられる．エージェントたちは互いにコミュニケーションを禁じられた状態で，ヒントのやりとりのみから自分の持っているカードを推察して場札を積み重ねることを目指す．ヒントから得られる情報だけではなく，前述した暗黙のコミュニケーションから得られる情報も用いて学習を進める．Bard らは [4] は Hanabi をマルチエージェント強化学習としてのベンチマークとして整備した．Hanabi はマルチエージェントタスクであり，不完全情報ゲームであり，各エージェントは持つ環境の状態についてそれぞれ異なる観測を持っている．さらに協力的な目標を持っているところが既存の囲碁などの研究と大きく異なっているとして新しいベンチマークとして提唱される．またコミュニケーションが限られているため，他のプレイヤーの行動から自分の理想の行動のヒントを得る必要がある．この研究においてオープンソースの Hanbi Learning Environment と呼ばれるオープンソースの強化学習環境がリリースされた．[1] いくつかのルールベースエージェントと強化学習エージェントを比較し最新の強化学習アルゴリズムを評価し，セルフプレイ環境で評価した場合，現在のルールベースのエージェントを凌駕するにはほとんど不十分であることを示した．また協調的かつ不完全情報ゲームであることから Hanabi はマルチエージェント環境における機械学習技術の研究課題となっていることが提示された．Foerster ら [7] は複雑かつ部分的に観測可能なマルチエージェント環境においてベイズ行動デコーダー (Bayesian Action Decoder:BAD) を用いることによって暗黙の慣習の学習をうまく行えることを示した．この研究によって特異な慣習を共通の学習を行ったエージェント間で設けることで Hanabi において好スコアを出せるエージェントが発表された．BAD をさらに改良した研究が Simplified action decoder (SAD) [8] である．Hu らは BAD はシンプルさと一般性にかけていると指摘し，暗黙の慣習に依存しない場合でも最終的な方策が分散実行と互換性を持たせ，集中トレーニング中に全てのエージェント間で自由に情報交換ができるように学習を行うことで，強化学習エージェントによって Hanabi で高スコアを出すことを可能にした．また Hanabi だけではなく協調的なマルチエージェントタスクには未知の相手との協調問題，ゼロショット調整問題が課題として挙げられている．ゼロショット調整問題にセルフプレイをそのまま適用すると，エージェントは高度に専門化された方策を確立して，共に訓練されていない他のエージェントとはうまく結果を残すことができない．そこで Hu ら [10] は Other-Play (OP) 新しい学習アルゴリズムを導入し，対称性をうまく用いることで異なる方法で対称性を破るパートナーに対して最大のロバスト性を持つ戦略を学習させることに成功した．また強化学習エージェントと人間との間の協調性を実験し，SAD のみを用いたエージェントよりも OP を組み合わせたエージェントの方が人間と協調的な結果を残せることを示した．Yu ら [20]

はマルチエージェント設定に特化した PPO [16] の変種であるマルチエージェント PPO（MAPPO）を用いても Hanabi において強化学習エージェントが高スコアを残せることを示した．近接型方策最適化アルゴリズム (PPO [16]) は一般的なオンポリシー強化学習アルゴリズムであるが，マルチエージェント環境ではオフポリシー学習アルゴリズムに比べて利用がかなり少ない．これは，マルチエージェント問題において，オンポリシー手法はオフポリシー手法に比べサンプル効率が著しく低いという考えによるものである．しかし，パラメータの調整を施すことにより，オンポリシーでありながらよく知られているオフポリシー手法と同等の結果を残し，サンプル効率の面でも匹敵した．

3. 背景知識

3.1 強化学習

強化学習とは機械学習の一種である．環境においてエージェントが報酬を最大化されるような方策を得ることを目的としている．一般的に環境は有限なマルコフ決定過程 (Markov Decision Process:MDP) として定式化される．MDP は S, A, T, r で定義される．

- 状態空間：$S = \{s_1, s_2, ..., s_m\}$
- 行動空間：$A = \{a_1, a_2, ..., a_m\}$
- 遷移関数：状態 $s_t \in S$ で行動 $a_t \in A$ を選択した時に状態 $s_{t+1} \in S$ に遷移する確率 $T(s_{t+1}|s_t, a_t)$
- 報酬関数：状態 $s_t \in S$ で行動 $a_t \in A$ を選択し状態 $s_{t+1} \in S$ に遷移したエージェントに与えられる報酬の期待値 $r(s_t, a_t, s_{t+1})$

基本的な強化学習では離散的な時間ステップ t における環境と環境の中で行動し学習を行うエージェントの相互作用を繰り返すことでエージェントの方策を学習していく．ある時刻 t におけるある状態 s_t でとる行動 a_t は方策 (policy)$\pi(a|s)$ に基づいて選択される．$\pi(a|s)$ は状態 s において行動 a が選択される確率である．この時エージェントは遷移した状態 s_{t+1} に基づいた報酬 $r_{t+1} = r(s_t, a_t, s_{t+1})$ を受け取る．

強化学習の目的は最終的に受け取る割引された累積報酬の和 R_t を最大化させるような方策を最大化する方策を学習することである．ここで

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + ... = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \qquad (1)$$

ただし $\gamma \in [0, 1)$ は割引率とする．

割引率とは将来の報酬が現在においてどれだけ価値があるかを示したものである．

3.1.1 Proximal Policy Optimization (PPO)

方策勾配に基づいた深層強化学習の手法の一つに，Proximal Policy Optimization (PPO) [16] がある．

PPO では，行動を決める Actor を直接改善しながら，その方策を評価する Critic も同時に学習させる，Actor-Critic のアプローチが用いられている．具体的には，エージェントが観測した状態 s を入力とするニューラルネットワーク（パラメータ θ）の出力として，Actor としての方策 π_θ と，Critic としての状態価値関数 $V^{\pi_\theta}(s)$ を同時に計算し，これを学習するものである．

方策勾配に基づいた Actor-Critic 系の手法には，非同期の分散学習を行う Asynchronous Advantage Actor-Critic (A3C) [**?**] という手法や，A3C の学習を同期的に行えるように改良した Advantage Actor-Critic (A2C) [**?**] という手法などがあった．これらの手法に共通して，マルチエージェントによる分散学習をする方式が採用されており，より効率的に環境を学習できることが期待される．しかしながら，学習中に行われるニューラルネットワークのパラメータ更新が起因して，1 度でも方策関数が大きく劣化してしまうと，その後は報酬が得られにくくなり方策関数の改善も困難になるなど，学習が不安定である問題が指摘されていた．

PPO では，学習の安定性を向上させるために，方策関数が大きく更新されるのを防ぐようなアルゴリズムが提案されている．まず，パラメータ更新前後でのネットワークの方策関数の出力の比 $r(\theta)$ は

$$r(\theta) = \frac{\pi_\theta(a \mid s)}{\pi_{\theta_{\text{old}}}(a \mid s)} \qquad (2)$$

と表せる．

その時点でのアドバンテージ関数の推定値を $\hat{A}_t$ とすると，PPO における方策の損失関数 $L^{\text{CLIP}(\theta}$ は次の式で表される．

$$L_t^{\text{CLIP}}(\theta) = \hat{\mathbb{E}}_t \left[\min \left(r(\theta)\hat{A}_t, \text{clip}(r(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t \right) \right] \qquad (3)$$

更新前後の方策関数の比 $r(\theta)$ を clip 関数を用いて $[1-\epsilon, 1+\epsilon]$ の範囲に抑えた値で方策勾配の計算を行う事で，方策関数の大きな更新を起こりにくくしている．具体的には，$\hat{A}_t$ の時には $r(\theta)$ が $1+\epsilon$ を超えないように，$\hat{A}_t < 0$ の時は $r(\theta)$ が $1-\epsilon$ を下回らないように計算が行われている．なお上式中で用いられている clip 関数とは以下の式を指す．

$$\text{clip}(x, a, b) = \begin{cases} a & (x < a) \\ b & (x > b) \\ x & (\text{otherwise}) \end{cases} \qquad (4)$$

上式のように $\text{clip}(x, a, b)(a < b)$ という関数は，x の値を $[a, b]$ の範囲におさめるという操作を意味している．

PPO ではこの方策の損失関数 $L_t^{\text{CLIP}}(\theta)$ に，アドバンテージ関数を推定させるための状態価値の損失関数 $L_t^{\text{VF}}(\theta) = (V_\theta(st) - V_t^{\text{targ}})^2$(ただし，$V_t^{\text{targ}} = \hat{A}_t + V_{\theta_{\text{old}}}(s_t)$) と，学習を安定化させるためのエントロピー項 $S[\pi_\theta](s_t)$ を加えたものを損失関数として，これを最大化させる学習

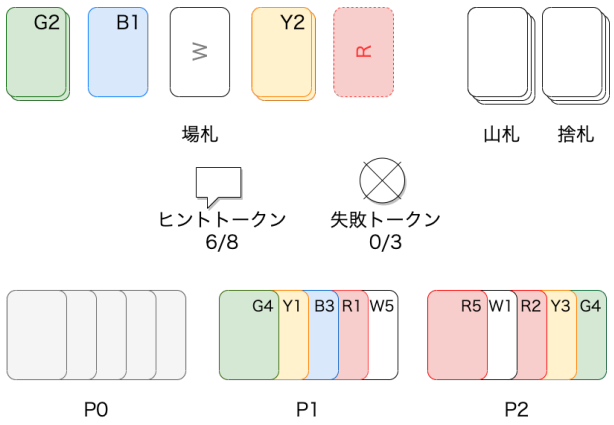

図 1 Hanabi のプレイ盤面
Fig. 1 Play board of Hanabi

を行う．

$$L_t^{\mathrm{CLIP+VF+S}}(\theta) = \hat{\mathbb{E}}_t\left[L_t^{\mathrm{CLIP}}(\theta) - c_1 L_t^{\mathrm{VF}}(\theta) + c_2 S\left[\pi_\theta\right](s_t)\right] \tag{5}$$

ここで c_1, c_2 は定数である．

またアドバンテージ関数を近似的に推定するために Generalized Advantage Estimator (GAE) [**?**] という手法が PPO では用いられている．通常のアドバンテージ関数は，T ステップの経験の軌跡から次のように計算される．

$$\hat{A}_t = -V(s_t) + r_t + \gamma r_{t+1} + \cdots + \gamma^{T-t+1} r_{T-1} + \gamma^{T-t} V(s_T) \tag{6}$$

この計算を一般化したものが GAE である．GAE の具体的な計算方法は次の式のようになる．

$$\begin{aligned} &\hat{A}_t = \delta_t + (\gamma\lambda)\delta_{t+1} + \cdots + \cdots + (\gamma\lambda)^{T-t+1}\delta_{T-1} \\ &\text{where } \delta_t = r_t + \gamma V(s_{t+1}) - V(s_t) \end{aligned} \tag{7}$$

$\lambda = 1$ の時，GAE で得られるアドバンテージ関数は，$\lambda = 0$ の時のアドバンテージを考慮しない計算になり，通常のアドバンテージ関数と同じものになる．つまり，λ を変更することでアドバンテージの考慮具合を調整することができる．

3.2 Hanabi

3.2.1 ルール

Hanabi は 2-5 人用の協力型不完全情報ボードゲームで，インディアンポーカーなどと同じく自分の手札がわからない状態でゲームを進行する．図 1 カードの種類は緑青白黄赤の 5 色と 1 から 5 までの 5 種類のカードで構成されており，各色 1 が 3 枚，2-4 が 2 枚，5 が 1 枚の合計 50 枚である．各プレイヤーは 4 枚 (プレイヤー人数が 2 人または 3 人の場合は 5 枚) のカードを手札として持ち，ソリティアのように色毎に 1 から 5 の順番でカードを場に揃えて出すことを目的としている．そして最終的に場に出せた 5 色の数の合計がゲームの点数となる．つまり最高スコアは 25 点である．このゲームの特徴的なルールとしてプレイ中に言葉や仕草などでコミュニケーションをとることは禁止されている．

プレイヤーが自分のターンにできることは以下の三つである．

・ヒントトークンを消費して他プレイヤーにヒントを出す．
・カードを捨ててヒントトークンを一つ増やす．
・場にカードを出す．

3.2.1.1 ヒント

ここでヒントトークンは上限が 8 個であり，ゲームの開始時点で 8 個全体に与えられている．そして一つ消費することでヒントを一つ他のプレイヤーに教えることができる．ヒントトークンが残っていない場合はカード捨ててトークンを増やすか，カードを場に出すかのどちらかしか選ぶことができない．ここでヒントを出す際には，手番のプレイヤーがヒントを出す数か色を選び，1 人のプレイヤーの持つ手札の中でその数や色と一致する手札を示すことができる．ヒントを与えることができるのはそのプレイヤーの手札に持つカードのみに限られる．与えられたヒントをどのように生かすかは各プレイヤーの判断に委ねられている．例えば図 1 の例ではプレイヤー 0 は次のようなヒントを出すことができる．

「プレイヤー 2 の持っている一番左と真ん中のカードは赤である．」

しかし実際にはプレイヤー 2 の持っている赤のカードは現在プレイできないためこのヒントを出すメリットが存在しない．実際に行われるヒントの例としてはプレイヤー 1 に右から二番目のカードが赤であることを伝えるなどがある．これを受け取ったプレイヤー 1 は次の手番で何故自分が赤のカードを持っていることを教えられたかを考えて，一番右のカードが赤の 1 であることを推論する．このような形でゲームは進行する．

3.2.1.2 カードのプレイ

手番のプレイヤーは手札の中から一枚カードを選び場に出すことができる．場に出した後，出すことのできるカードであれば場札が更新されて次のプレイヤーに手番が移る．またこのプレイによりいずれかの色の 5 のカードが場札として出た場合ヒントトークンを一つ得る．出すことのできないカードがプレイされた時はカードは捨てられて失敗トークンが場に一つ追加される．また，自分の手札についての情報が少ない場合やそうすることが有効であると判断した場合は手札の任意のカードを一枚捨てることでヒントトークンを一枚増やすことができる．いずれの場合も手番終了前に山札から手札に一枚カードを加える．

3.2.1.3 ゲームの終了条件

ゲームの終了条件は 3 種類存在する．

・全ての色の場札を 5 まで揃えること．
・失敗トークンが 3 つになること．

・プレイヤーが引くカードが山札からなくなってから手番が一周すること．

一つ目の全ての色の場札を 5 まで揃えることができた場合はゲームクリアとなり 25 点満点となる．二つ目の失敗トークンが三つ溜まった場合はその場でゲームが終了されて，その時点での点数が計算されそのままスコアとなる．三つ目の山札が尽きた場合も各々の手番が終わった後にその時点での点数が計算されてスコアとなる．

3.2.2 基本戦略

ヒントによって明確に伝えられた根拠のある情報だけで 25 種類のカードを完全にプレイするにはヒントトークンの数が足りていない．そのために複数のカードに対して情報を与えられるヒントを出すなどの手が有るが，場札の進行度合いによって直近でプレイ可能な少数のカードに対してヒントを出すか，プレイが可能になる場面は遠いが情報量の多いヒントを出すことを優先するかなどの手から取るべきである最適な手が変化する．またプレイ可能なカードに対するヒントとは別に捨てても良いカードを示したり，捨ててはいけないカードを示したりなどヒントが示唆する内容は多岐にわたるため，プレイヤーは現在の場札と捨て札と他のプレイヤーの手札から出されたヒントの意味合いを推察してプレイを進めていく必要がある．

3.2.2.1 暗黙のコミュニケーション

Hanabi での明示的なコミュニケーションは先述の通りヒントアクションによって行われるものに限定されているが，プレイ中の全てのアクションが全てのプレイヤーから観察可能であり，この観察から暗黙的に情報の伝達が行われる場合がある．この暗黙のコミュニケーションとでもいうような情報はある行動が環境に与える影響ではなくて，他のプレイヤーがその行動をとることを決めた事実によって伝えられる．これはプレイヤーが他のプレイヤーの立場に立ち，様々な状況下での行動の選択肢を考え行動から状況の逆算を行い推察をすすめることで成立する．

例えば図 1 の例では，プレイヤー 0 の手番でありプレイヤー 1 に右から 2 枚目のカードは赤であるとヒントを出しプレイヤー 1 が赤の 1 をプレイし終わった際のことを考える．ここでプレイヤー 2 が次の手番のプレイヤー 0 に対してヒントを出さずにプレイヤー 1 に対して何らかのヒントを出した場合，プレイヤー 0 は以下のように推察できる．

「私がヒントをもらえなかったということは私が次に出せるカードがないということなので，どのカードを捨てても良い.」よって手札の中でまだ一つもヒントを得ていないカードのうちいずれかを安心して捨てることができる．このような形でヒントの有無だけでも情報を推察して得ることができる．

4. 実験結果および分析

本研究の目的はマルチエージェント強化学習の中でも分散型部分観測可能マルコフ決定過程に属する協力型不完全情報ゲーム Hanabi に対して，オンポリシー手法により学習を効率よく進めるための設定の分析を行う．

前項で述べたように Hanabi は協力型マルチエージェント強化学習のベンチマークとして研究が盛んに行われている．その中で Q 学習ベースの R2D2 [11] などのオフポリシー手法を用いて研究が進められてきた [4] が，MAPPO [20] によって従来のオンポリシー手法によるアプローチでも Hanabi のような協力型不完全情報ゲームでも学習を行えることがわかった．ただし通常の設定では Hanabi の観測が部分観測かつ学習がマルチエージェントであることを理由に学習が難しく，細やかなハイパーパラメータの調整が必要となる．そこで MAPPO の流れを汲み，より詳しくどのような設定の調整を行うと Hanabi の学習が効率よく行えるかの分析，研究を行う．

4.1 実験設定

実験では Actor-Critic の一種である PPO アルゴリズムを用いる．実行環境は Hanabi を Gym 環境に合わせて自作したものを使用する．エージェントの観測の初期設定としては対戦相手の人数，手札枚数，ヒントトークンの最大数 (初期数)，失敗トークンの許容数，色の種類数，各色の最高ランク，各カードのデッキ中の枚数が与えられている．またゲームのプレイ中は残り山札の枚数がスカラー量で，他のプレイヤーの手札の内容がランクと色の配列で，場札の状況が二次元配列で，捨て札の状況が二次元配列で，ヒントトークンの数がスカラー量で，失敗トークンの数がスカラー量で，現在の自分の手札でわかっている情報が配列で，それぞれ特徴量として与えられている．これらの情報から出すべきカード，捨てるべきカードの判断を行う方策を学習する．今回の報酬関数は場札にカードを出した時，それが出せるカードであった場合に+1 の報酬を受け取るような形になっている．用いている Actor-Critic アルゴリズムにおいて，Actor と Critic には違う入力が与えられている．具体的には Actor には観測そのままの入力が与えられているが，Critic にはそれに加えて本来は見えない自分のカード情報を与えている．この方法は MAPPO で有効と指摘されており，同様に取り入れた．

実験対象としたパラメータとその設定の変更が与える影響の仮説は以下のようになる．

- 割引率 γ： 人間が考える情報は数ターン先までのみになるため，同様にエージェントも γ を小さくなると分散が抑えられ結果が向上すると予想される．よく使われる設定値は 0.99 である．
- GAE の λ (式 7) : 報酬の偏りや分散を制御するパラメータであるが，マルチエージェントにおいては分散は大きくなる傾向にあるため，通常より λ を小さくすることで分散を抑えると結果が

向上することが予想される．よく使われる設定値は 0.95 である．

- エントロピー係数 c (式 5)：λ と同様マルチエージェントの関係上報酬の分散が大きくなりがちなので，エントロピー係数を小さめにした方が結果が良くなると予想される．よく使われる設定値は 0.01 である．
- PPO の clip パラメーター ϵ (式 5)：マルチエージェントの関係上パラメータの変化がデータ分布に大きく影響するため，小さめの値の方が結果が良くなると予想される．よく使われる設定値は 0.2 である．
- モデル構造 Gated Recurrent Unit (GRU)：Recurrent Neural Network (RNN) の一種であり，これが存在することによって過去の観測を使って方策を決定できる．用いないと暗黙の慣習を学習することができず結果が低くなると予想される．

本実験において，学習にかかる計算量の側面から Hanabi のルールに変更を加えた．従来では 2 人プレイの場合はカードが 5 色 25 種類 50 枚のところを 3 色 15 種類 30 枚に変更して学習を行った．

4.2 結果と分析

4.2.1 割引率 γ

割引率 γ を変化させた結果が図 2 である．横軸は学習のステップ数であり，縦軸は得られた報酬である．報酬は最大で 15 点となる．γ を 0.93 から 0.99 まで 0.02 刻みで動かしたグラフを示した．結果としては 0.93 では全く学習を行わなかった．また 0.95 の際も学習は途中まで行えていなかった．また 0.97 と 0.99 の場合ではそこまで大きな差は見られず正常に学習を行っている．

この結果は仮説の「人間が考える情報は数ターン先までのみになるため，同様にエージェントも γ を小さくしても結果には影響しない.」とは反するものである．この理由としては割引率の低下がモデルの表現力の低下に繋がり，学習効率の低下に繋がった可能性がある．割引率は正則化を行う働きがある．[3] この割引率が一定以下に下がった結果正則化の働きが上がり，モデルの表現力の低下を招いたことにより学習が阻害されたのであると考えられる．[12]

4.2.2 GAE の λ

GAE の λ を変化させた結果が図 3 である．横軸は学習のステップ数であり，縦軸は得られた報酬である．λ を 0.6 から 0.95 まで動かしたグラフを示した．結果としては λ が 0.7 時に学習が最も効率よく進みそれより下げると学習結果が悪くなった．

この結果は仮説の「マルチエージェントにおいては分散は大きくなりがちなため，通常より λ を小さくすることで分散を抑えると結果が向上する.」とは部分的に一致するものである．一定の値より小さくすると偏りが大きくなる形で学習が進み，学習に悪影響をもたらすためであると考えられる．

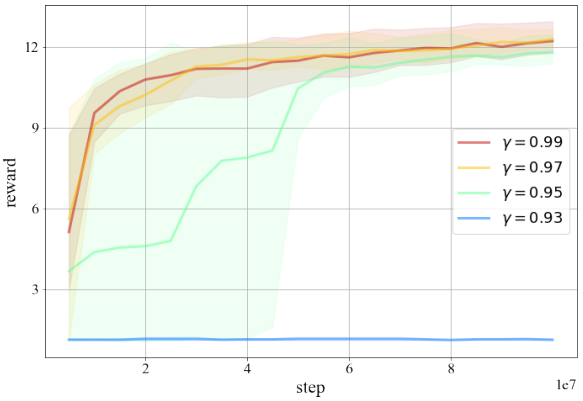

図 2 割引率 γ のを変化させた学習結果の比較
Fig. 2 Comparison of learning results with varying discount rate, γ

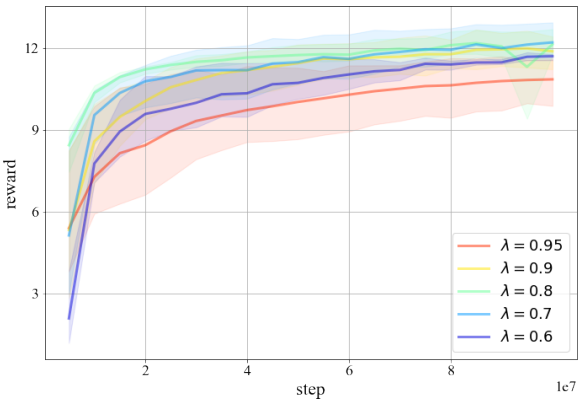

図 3 GAE の λ を変化させた学習結果の比較
Fig. 3 Comparison of learning results with varying λ of GAE

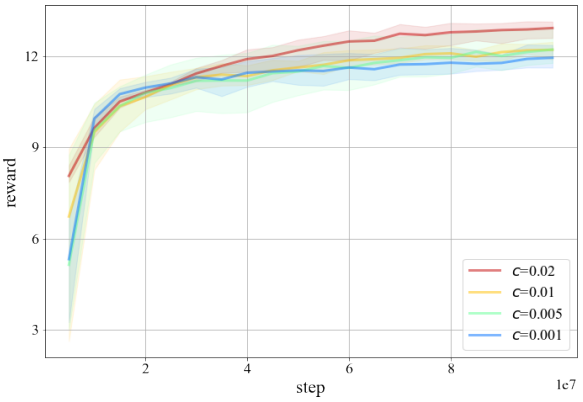

図 4 エントロピー係数 c を変化させた学習結果の比較
Fig. 4 Comparison of learning results with varying entropy factor, c

4.2.3 エントロピー係数 c

エントロピー係数 c を変化させた結果が図 4 である．横軸は学習のステップ数であり，縦軸は得られた報酬である．c を 0.001 から 0.02 までの間で動かしたグラフを示した．結果としては大きければ大きいほど結果が良くなった．

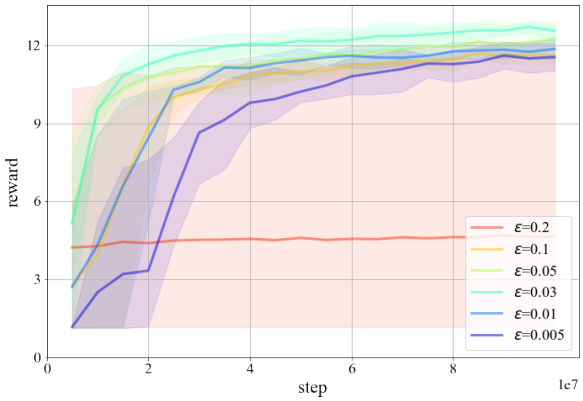

図 5　clip パラメーター ϵ を変化させた学習結果の比較
Fig. 5　Comparison of learning results with varying clip parameter, ϵ

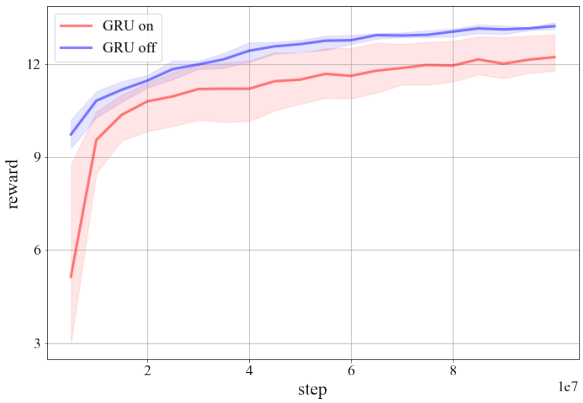

図 6　GRU の有無による学習結果の比較
Fig. 6　Comparison of learning results with or without GRU

この結果は仮説の「マルチエージェントの関係上報酬の分散が大きくなりがちなので，エントロピー係数を小さめにした方が結果が良くなる.」と異なる．エントロピー係数は小さいと分散が抑えられるが，探索が進みにくくなる．さらにエントロピーによる正則化は探索だけでなくパラメータの最適化をしやすくしているとも指摘されている [2] ため，エントロピーを下げるメリットが薄かった可能性があると考えられる．

4.2.4　PPO の clip パラメーター ϵ

PPO の clip パラメーター ϵ を変化させた結果が図 5 である．横軸は学習のステップ数であり，縦軸は得られた報酬である．ϵ を 0.005 から 0.2 までの間で動かしたグラフを示した．結果としては数字が小さい方が学習をうまく進めることに貢献しているが，数字が小さすぎると学習の速度が落ちてしまう結果となった．

この結果は仮説の「マルチエージェントの関係上パラメータの変化がデータ分布に大きく影響するため，小さめの値の方が結果が良くなる.」に沿った結果となる．MAPPO で指摘されたように大きいと学習自体が進みにくい一方で，小さすぎると学習が遅くなる結果となった．

4.2.5　Gated Recurrent Unit (GRU)

GRU の有無を変化させた結果が図 6 である．横軸は学習のステップ数であり，縦軸は得られた報酬である．GRU を用いている場合と用いていない場合を示した．結果としては GRU がない方が学習の立ち上がりが早く，よく学習するという結果になった．

これは仮説の「GRU を用いない場合暗黙の慣習を学習することができず結果が低くなる.」とは異なる結果となる．GRU を使用すると計算量が大きくなるため，一定の計算量内で性能の向上を目指す本実験においては，用いない方が良い結果に繋がったと考えられる．両グラフともに学習が進んでいるが，GRU がない方が学習が早く進むことがわかる結果となった．さらに学習ステップを増やした場合は GRU を用いた場合の方が結果が良くなることもまた考えられる．

5.　おわりに

本稿では，協力型不完全情報ゲーム Hanabi をマルチエージェント強化学習の研究対象として捉え，オフポリシー手法において Hanabi を研究し，人間のプレイを基に模倣学習を行ったエージェントやルールベースのエージェントに匹敵する結果を残した研究や，それらのエージェントを超えるスコアを獲得した研究を紹介した．またマルチエージェントであることと Hanabi が部分観測であることから，オンポリシー手法において学習を行う際に通常の設定では学習がうまく進まず設定を調整することで学習を進めることができるようになるということを指摘した．この設定に伴う問題解決のために実際にオンポリシー手法である PPO を基にしたアルゴリズムを自作の Hanabi の環境下で学習を進める環境を構築して，各パラメータへの分析を行い結果から各パラメータが学習に寄与する影響を考察した．

今後の課題としては，まず本研究では計算量の観点から限られた計算量での学習しか行えず，学習が正常に行えているかどうかや，学習の早さの観点からの評価できたものの，学習を完全に終えたエージェントのスコアでは評価を行えなかった．そこでより長い時間の学習を行うことで限られた時間での結果だけではなく学習の最終着地スコアも加味したパラメータの評価を行いたいという点が挙げられる．この観点からの評価も行うことでより最適な Hanabi の学習を行う設定の解明に繋がるのではないかと期待する．また今回の実験では試さなかったモデルのニューラルネットワークの数の変化などがどのような影響を学習に与えるかの研究も必要であると考える．これまで述べたように Hanabi での学習は，現実世界でのマルチエージェントタスクをエージェントが効率よく学習することにつながるため，その目標に一歩近づくことができるだろう．

謝辞　本研究を遂行するにあたり，大変多くの方にご指導ご協力をいただきました．

指導教員である東京大学 鶴岡慶雅教授には研究内容や研究生活に関して様々なご指導をいただきました．博士課程の李凌寒さんには親身にご指導いただきました．博士課程の安井豪さんには研究のアドバイスをいただくとともに研究に対する姿勢や心構えなどを教授いただきました．同期の橋本大世くんには研究の基礎的な部分を教えてもらいました．綿引隼人くんには日常的な会話の中で気づきや教えをいただきました．他にも様々な先輩や同期，後輩のおかげで研究を進めることができました．感謝申し上げます．また最後に常に心身の健康を気遣い，気にかけてくれた家族や友人の皆様に心より感謝を申し上げます．

参考文献

[1] : Hanabi Learning Environment, https://github.com/deepmind/hanabi-learning-environment.
[2] Ahmed, Z., L. R. N. N. M. and Schuurmans, D.: Understanding the impact of entropy on policy optimization, International Conference on Machine Learning (pp. 151-160). PMLR (2019).
[3] Amit, Ron, R. M. and Ciosek, K.: Discount factor as a regularizer in reinforcement learning, International conference on machine learning. PMLR, 2020 (2020).
[4] Bard, N., Foerster, J. N., Chandar, S., Burch, N., Lanctot, M., Song, H. F., Parisotto, E., Dumoulin, V., Moitra, S., Hughes, E. et al.: The hanabi challenge: A new frontier for ai research, Artificial Intelligence (2020).
[5] Brown, N. and Sandholm, T.: Superhuman AI for heads-up no-limit poker: Libratus beats top professionals, Science (2018).
[6] Campbell, M., Hoane Jr, A. J. and Hsu, F.-h.: Deep blue, Artificial intelligence (2002).
[7] Foerster, J., Song, F., Hughes, E., Burch, N., Dunning, I., Whiteson, S., Botvinick, M. and Bowling, M.: Bayesian action decoder for deep multi-agent reinforcement learning, International Conference on Machine Learning (2019).
[8] Hu, H. and Foerster, J. N.: Simplified action decoder for deep multi-agent reinforcement learning, arXiv preprint arXiv:1912.02288 (2019).
[9] Hu, H., Lerer, A., Cui, B., Pineda, L., Wu, D., Brown, N. and Foerster, J.: Off-Belief Learning, arXiv preprint arXiv:2103.04000 (2021).
[10] Hu, H., Lerer, A., Peysakhovich, A. and Foerster, J.: “Other-Play” for Zero-Shot Coordination，International Conference on Machine Learning (2020).
[11] Kapturowski, S., Ostrovski, G., Quan, J., Munos, R. and Dabney, W.: Recurrent experience replay in distributed reinforcement learning, International conference on learning representations (2018).
[12] Kumar, A., A. R. G. D. and Levine, S.: Implicit under-parameterization inhibits data-efficient deep reinforcement learning, arXiv preprint arXiv:2010.14498 (2020).
[13] Li, J., Koyamada, S., Ye, Q., Liu, G., Wang, C., Yang, R., Zhao, L., Qin, T., Liu, T.-Y. and Hon, H.-W.: Suphx: Mastering mahjong with deep reinforcement learning, arXiv preprint arXiv:2003.13590 (2020).
[14] Panait, L. and Luke, S.: Cooperative multi-agent learning: The state of the art, Autonomous agents and multi-agent systems (2005).
[15] Schaeffer, J., Lake, R., Lu, P. and Bryant, M.: Chinook the world man-machine checkers champion, AI Magazine (1996).
[16] Schulman, J., Wolski, F., Dhariwal, P., Radford, A. and Klimov, O.: Proximal policy optimization algorithms, arXiv preprint arXiv:1707.06347 (2017).
[17] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T. and Hassabis, D.: Mastering the Game of Go with Deep Neural Networks and Tree Search, Nature (2016).
[18] Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A. et al.: Mastering the game of go without human knowledge, nature (2017).
[19] Tesauro, G. et al.: Temporal difference learning and TD-Gammon, Communications of the ACM, Vol. 38, No. 3, pp. 58–68 (1995).
[20] Yu, C., Velu, A., Vinitsky, E., Wang, Y., Bayen, A. and Wu, Y.: The surprising effectiveness of mappo in cooperative, multi-agent games, arXiv preprint arXiv:2103.01955 (2021).