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　５ 年　 ５ 月 ２６ 日現在

機関番号：１２６０１
研究種目：基盤研究(B)（一般）
研究期間：2020～2022
課題番号：２０Ｈ０４３０１

研究課題名（和文）環境モデルと戦略の同時学習による深層強化学習

研究課題名（英文）Deep Reinforcement Learning by Simultaneous Learning of Environment Models and Strategies

研究代表者
鶴岡　慶雅（Tsuruoka, Yoshimasa）

東京大学・大学院情報理工学系研究科・教授

研究者番号：５０５６６３６２

交付決定額（研究期間全体）：（直接経費）　　12,700,000円

研究成果の概要（和文）：複数の環境モデルを活用して誤差の影響を軽減するプランニング手法および複数ステップ先の状態を直接予測するマルチステップモデルを開発し、効率的な深層強化学習の実現に成功した。また、部分観測環境における教師なし強化学習のための内発的報酬および行動類似性に基づく潜在状態表現を設計し、強化学習の汎化性能を向上させた。さらに、ローグライクゲームでの報酬設計の改良、オフポリシー強化学習でのメモリ消費の削減、階層強化学習の利用による解釈性の高い戦略の構築を実現した。

研究成果の学術的意義や社会的意義
本研究成果は、モデルベース強化学習における環境モデルのより良い活用法、内発的報酬の設計、潜在状態表現の改善などを深層強化学習に導入することで、深層強化学習の性能を改善し、より効率的で汎用性の高い学習を実現することに貢献するものである。また、社会的には、本研究の成果は、ビデオゲームだけでなく、自動運転、ロボット制御、エネルギー管理など、実世界の多様なタスクに対する深層強化学習の適用可能性を高めることに貢献する可能性がある。

研究成果の概要（英文）：We developed a planning method that leverages multiple environment models to reduce the impact of errors, and a multi-step model that directly predicts states several steps ahead, successfully achieving efficient deep reinforcement learning. We also designed an intrinsic reward and a latent state representation based on action similarity for unsupervised reinforcement learning in partially observable environments, improving the generalization performance of reinforcement learning. Furthermore, we improved the design of rewards in roguelike games, reduced memory consumption in off-policy reinforcement learning, and realized the construction of highly interpretable strategies through the use of hierarchical reinforcement learning.

研究分野：強化学習、自然言語処理、ゲームAI

キーワード：強化学習　深層学習

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

知的なエージェントを計算機で構築するためのアプローチとして、強化学習（reinforcement learning）と呼ばれる手法が研究されている。これは、エージェントが与えられた環境の中で試行錯誤を繰り返し、得られる累積報酬の期待値を最大化するような方策（行動規則）を見つけ出すように学習させるというアプローチである。近年、強化学習に深層学習（deep learning）に基づく関数近似を導入することで、ビデオゲームなど、状態空間が極めて大きなゲームのプレイヤが半自動的に構築できるようになっている。このようなアプローチは、深層強化学習（deep reinforcement learning）と呼ばれ、計算機で知的なエージェントを実現するための最有力の技術のひとつとして盛んに研究されている。

囲碁や将棋といったボードゲームに関しては、強化学習によって人間をはるかに超えるレベルのゲーム AI が実現されている。このようなゲームでは、強化学習における「環境モデル」、すなわち、ゲームの中での状態遷移や報酬に関する知識がエージェントにとって最初から全て既知であるため、現実の囲碁盤や将棋盤を利用する必要はなく、効率的な強化学習が可能となっている。また、実行時に行動（指し手）を選択する際にも、環境の「完全なモデル」を利用した先読み（lookahead）が可能であるため、先読みに計算コストを投入することでより優れた行動選択が可能となっている。それに対して、ビデオゲームやリアルタイムストラテジーゲームといったより現代的なゲームでは、学習エージェントの立場から見た場合、環境モデルは未知であるため、環境モデルを利用しないモデルフリー強化学習（model-free reinforcement learning）と呼ばれる手法がよく用いられる。Atari 2600 と呼ばれるビデオゲームで深層強化学習に成功し世界的な注目を集めた Deep Q-Network や、リアルタイムストラテジーゲームの一種である Dota 2 上で、深層強化学習によって人間のトップレベルのプレイヤと互角に戦える AI を構築した研究などは、いずれもモデルフリー強化学習によるアプローチに基づいている。

モデルフリー強化学習が現代的なゲームに対して AI を構築するうえで有力なアプローチであることは間違いないが、いくつかの深刻な問題も存在する。ひとつは、そのサンプル効率の悪さである。Atari 2600 のような比較的シンプルなビデオゲームでも、エージェントがある程度のレベルに達するためには数百万フレームを超える学習が必要とされる。エージェントが報酬を得られる機会が多くないゲーム、すなわち報酬が疎（スパース）なゲームでこの問題は特に深刻であり、現実的な時間では実質的な学習がほとんど進まないということが起こる。

もうひとつの問題は、実行時に先読みを行うための環境モデルが存在しないことである。このため、モデルフリー強化学習に基づくエージェントは、観測を入力としたニューラルネットワークから得られる出力を直接利用して行動を決定する。これは、囲碁や将棋の AI が $\alpha\beta$ 探索やモンテカルロ木探索を利用し、実行時に深い先読みを行って優れた指し手を決定していることとは対照的である。すなわち、実行時に計算コストをかけて行動選択の精度を向上させるということが実現できていない。

一方、強化学習においては、環境モデルを明示的に学習し、それを利用してエージェントの方策を効率的に学習するモデルベース強化学習（model-based reinforcement learning）と呼ばれるアプローチが存在する。モデルベース強化学習では、前述のモデルフリー強化学習の欠点を克服し、サンプル効率の大幅な向上や実行時の先読み機構の導入が実現できる可能性がある。しかし現在のところ、環境モデルの学習の難しさや、学習した環境モデルと真のモデルとのずれの問題などから、現代的なゲームの AI の構築においてモデルベース強化学習がモデルフリー強化学習にとってかわれるという状況にはなっていない。

２．研究の目的

前節で述べた学術的背景のもと、本研究プロジェクトでは、ビデオゲームやリアルタイムストラテジーゲーム、ローグライクゲームといった現代的なゲームにおいて、環境モデルを利用した効率的な深層強化学習を実現するためのアプローチの確立を目指す。予備的な研究により、画像などの高次元入力を直接扱う必要のある現代的なゲームにおいても、環境モデルを陽に学習するモデルベース強化学習が有効である可能性が明らかになっている。しかし、同時に以下のような問題も明らかになった。

ひとつは計算コストの問題である。我々が提案したアーキテクチャでは、低次元にエンコードした潜在表現を利用しているため、モデルベース強化学習の代表的な先行研究である Imagination-Augmented Agents (I2As) よりもはるかに小さい計算コストで先読みを行うことができる。しかし、それでもモデルフリー強化学習と比較した場合、より大きな計算コストが学習に必要となっている。この問題は、先読みの長さや幅を大きくした場合にはより大きな問題になることが予想されるため、計算コストをいかにして削減するかは重要な課題である。

もう一つの問題は、学習された環境モデルの精度の問題である。環境モデルは有限のサンプルから学習されるために、状態遷移も報酬予測も真のモデルとの間には誤差が存在する。そのため、学習された環境モデルを用いた先読みの結果にも少なからず誤差が存在し、先読みの深さや幅を大きくしても必ずしもエージェントの性能が向上するとは限らない。これは、コンピュータ将棋や囲碁において、実行時の先読みの長さや幅を大きくすればするほど AI の棋力が向上する状況とは対照的である。

三つ目の問題は、学習の安定性の問題である。一般に、強化学習は教師付き学習と比べて学習が不安定になりがちであることがよく知られている。これは、エージェントの方策の変化にともなって入力と出力の確率分布が変化することが大きな要因であるが、本研究プロジェクトが目指すような、環境モデルを方策と同時に学習するアプローチの場合、学習中に変化する環境モデルを利用して方策を学習するため、学習の安定性の問題はよりシビアな問題となる。

以上のような問題を抱えつつも、環境モデルを利用することによるサンプル効率の向上はモデルベース強化学習の大きなメリットであり、これらの問題点を解決することができれば、深層強化学習が適用できるゲームの範囲は大きく広がることが期待される。また、近年、環境モデルの誤差を定量的に評価することにより、強化学習の本質的な課題である探索行動（explorartion）の質を大きく改善できる可能性があることが明らかになっている。本研究プロジェクトでは、現状のモデルベース強化学習の課題を解決すると同時に、環境モデルを利用した探索行動の改善を実現し、いままで深層強化学習のみでは構築が困難であった、Minecraft やロールプレイングゲームといった、自由度の非常に高いゲームでも AI を構築可能にする深層強化学習技術の確立を目指す。

３．研究の方法

(1) 環境モデルの誤差への対処

モデルベース強化学習における課題の一つは、学習された環境モデルの精度の問題である。環境モデルは有限の大きさのサンプルから学習されるために、状態遷移も報酬予測も真のモデルとの間には誤差が存在する。そのため、学習された環境モデルを用いたプランニングの結果にも少なからず誤差が存在し、プランニングの深さや幅を大きくしても必ずしもエージェントの性能が向上するとは限らない。この問題に対処する手法として、複数の環境モデルを活用することによって誤差の影響を軽減したプランニングを可能にする手法の開発を行った。

(2) マルチステップモデル

環境モデルの誤差の問題に対するもう一つの対処法として、複数ステップ先の状態を直接予測するモデル（マルチステップモデル）を利用する手法の研究を行った。複数ステップ先の状態を直接予測することにより、予測誤差の蓄積の問題が軽減することが期待される。

(3) 部分観測環境における深層強化学習

部分観測環境における深層強化学習における報酬設計の問題に対処するため、部分観測環境における教師なし強化学習のアルゴリズムの開発を行った。

(4) モデルベース強化学習の汎化性能

強化学習の問題点の一つとして、学習時に見たことがない未知の環境においてエージェントの性能が大きく低下することが知られている。その問題に対処するため、行動類似性に基づく潜在状態表現を利用することで、モデルベース強化学習の汎化性能を向上させる手法の開発を行った。

(5) 深層強化学習における報酬設計

深層強化学習における報酬設計の問題に対処するため、これまでに様々な内発的報酬の仕組みが提案されている。本研究では、状態遷移の予測不可能性と、状態の新規性をベースにした内発的報酬を組み合わせることで、noisy-TV problem と呼ばれる、ランダムな状態遷移が継続して起きる状況にエージェントがトラップされる問題の解消を試みた。

(6) ローグライクゲームにおける内部報酬の設計

本研究プロジェクトが対象とするゲーム AI 環境のひとつに、ローグライクゲームと呼ばれるダンジョン探索型環境がある。先行研究では「好奇心」による内部報酬を用いた手法が利用されているが、探索済み状態を過剰に避けるなどの問題点が指摘されている。そこで本研究では、ローグライク環境において３種類の報酬設計で学習を行った。

(7) 深層強化学習におけるメモリ消費の改善

Off-policy 強化学習では、エージェントが環境から収集した遷移データを保持するために大量のメモリが消費される。本研究では，遷移データの学習における優先度を計算し、相対的に重

要でないと判断されたものから破棄することで、バッファによるメモリ消費を節約する手法を開発した。

(8) 不完全情報ゲームにおける隠れ情報の推定
不完全情報ゲームにおいては、対戦相手の状態など、自分からは見えない情報を、観測可能な情報から推測することが重要である。そこで本研究では、代表的な不完全情報ゲームである麻雀を題材として、相手の手牌を深層学習によって推定する手法を開発した。

(9) 階層強化学習による戦略の解釈性の向上
強化学習に基づくゲーム AI の課題のひとつに、エージェントがどのような戦略に基づいて行動を決定しているのかが人間にとってブラックボックスであるという問題がある。そこで本研究では、説明可能なゲーム AI を実現する方法として階層強化学習を用いる手法を開発した。

４．研究成果

(1) 環境モデルの誤差への対処
提案法では、各環境モデルの信頼度を、他の多数のモデルとの異なりの大きさによって定量化する。プランニングのための行動列候補の累積報酬の計算を行う際には、信頼度によって重みづけされた報酬を用いることで、信頼性の高い環境モデルによって予測された行動候補列を優先的に考慮する。Open AI Gym 環境を用いた評価実験の結果、本手法が従来のモデル予測制御による手法よりも高い性能を達成することが確認された。

(2) マルチステップモデル
モデルを用いて方策の学習を行う際に、モデルの精度に応じて学習に用いるステップ数を調整することで効率的な学習を行う。Atari のゲームを用いて評価実験を行った結果、従来手法よりも高い性能が得られる傾向にあることが確認された。

(3) 部分観測環境における深層強化学習
部分観測性に対処するための記憶機構、および相互情報量に基づいた内発的報酬を設計した。提案する内発的報酬は、観測情報が限られている状態空間を優先的に探索し、有効な記憶を学習することを可能にする。実験では、外部報酬を使用せずに、部分観測環境において有益な方策を学習することに成功した。

(4) モデルベース強化学習の汎化性能
モデルベース強化学習の潜在状態表現を Policy Similarity Embeddings と呼ばれる手法を用いて学習する手法を開発た。提案手法を、背景画像が変化する環境における連続行動空間の制御タスクに適用し、汎化性能の検証を行なったところ、一部のタスクにおいて汎化性能の向上が見られた。

(5) 深層強化学習における報酬設計
モデルベース強化学習の手法内の遷移予測モデルを利用して内発的報酬を計算する手法を開発した。Atari 環境を用いた実験の結果、探索が難しいとされるベンチマークでの性能向上を確認した。図１に、Montezuma's Revenge と呼ばれる環境での実験結果を示す。

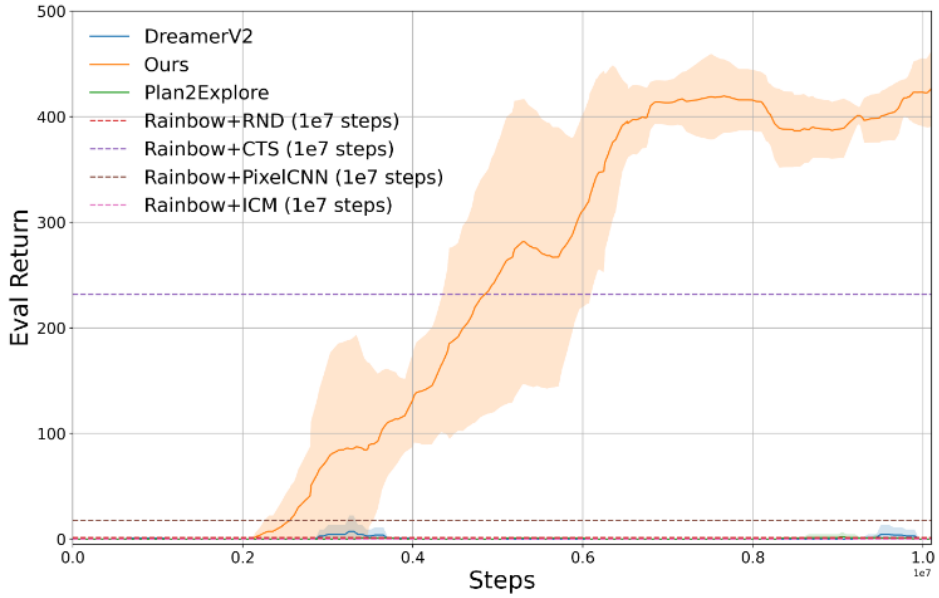

図１　Montezuma's Revenge 環境における実験結果

さらに新規性ベースの内発的報酬と組み合わせることで noisy-TV problem を緩和できることを確認した。

(6) ローグライクゲームにおける内部報酬の設計
自作したローグライク環境において評価実験を行い、提案した内発的報酬設計によって学習が促進されることを確認した。

(7) 深層強化学習におけるメモリ消費の改善
遷移データの学習における優先度を計算し，相対的に重要でないと判断されたものから破棄することで，バッファによるメモリ消費を節約する手法を開発した。

(8) 不完全情報ゲームにおける隠れ情報の推定
代表的な不完全情報ゲームである麻雀における相手手牌推定に関する実験の結果、近年注目を集めている深層学習モデルである Transformer を用いて自己回帰的に推定することで高精度の推定が可能であることが明らかになった。

(9) 階層強化学習による戦略の解釈性の向上
実験では、簡単な Atari 環境であるブロック崩しにおいて、AI の戦略を人間が容易に理解可能な形で可視化できることが明らかになった。

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計11件（うち招待講演　0件／うち国際学会　3件）

１．発表者名 岩瀬 諒
２．発表標題 選択的注意機構を用いたロバストな強化学習手法の実現
３．学会等名 第26回ゲームプログラミングワークショップ（GPW21）
４．発表年 2021年

１．発表者名 脇 聡志
２．発表標題 世界モデルによる好奇心と新規性に基づく探索
３．学会等名 第26回ゲームプログラミングワークショップ（GPW21）
４．発表年 2021年

１．発表者名 橋本 大世
２．発表標題 リセット機能を活用したシミュレータにおける効率的な方策学習
３．学会等名 第26回ゲームプログラミングワークショップ（GPW21）
４．発表年 2021年

１．発表者名 中本 光彦
２．発表標題 外部記憶を用いた部分観測環境における教師なし強化学習
３．学会等名 第26回ゲームプログラミングワークショップ（GPW21）
４．発表年 2021年

１．発表者名
Mitsuhiko Nakamoto

２．発表標題
Unsupervised Reinforcement Learning for Partially Observable Environments Using External Memory

３．学会等名
NeurIPS 2021 Workshop on Ecological Theory of Reinforcement Learning（国際学会）

４．発表年
2021年

１．発表者名
Yuhang Jiao

２．発表標題
HiRL: Dealing with Non-stationarity in Hierarchical Reinforcement Learning via High-level Relearning

３．学会等名
AAAI-22 Workshop on Reinforcement Learning in Games（国際学会）

４．発表年
2022年

１．発表者名
橋本大世、鶴岡慶雅

２．発表標題
深層強化学習における擬似的な行動による中間フレームの有効活用

３．学会等名
ゲームプログラミングワークショップ2020

４．発表年
2020年

１．発表者名
中田惇貴、鶴岡慶雅

２．発表標題
環境モデルの誤差による影響を抑える強化学習手法

３．学会等名
ゲームプログラミングワークショップ2020

４．発表年
2020年

１．発表者名
海野良介、鶴岡慶雅

２．発表標題
離散行動空間における教師なしスキルの獲得手法

３．学会等名
ゲームプログラミングワークショップ2020

４．発表年
2020年

１．発表者名
藤田航輝、鶴岡慶雅

２．発表標題
モデルベース強化学習における方策ネットワーク手法の活用

３．学会等名
ゲームプログラミングワークショップ2020

４．発表年
2020年

１．発表者名
Taisei Hashimoto, Yoshimasa Tsuruoka

２．発表標題
Utilizing Skipped Frames in Action Repeats via Pseudo-Actions

３．学会等名
NeurIPS 2020 Deep Reinforcement Learning Workshop（国際学会）

４．発表年
2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

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