The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

3D 仮想環境 Minecraft における深層強化学習への視線方向の影響

Effect of Viewing Directions on Deep Reinforcement Learning in 3D Virtual Environment *Minecraft*

松井　太樹　　小山　聡　　栗原　正仁
Taiju Matsui　Satoshi Oyama　Masahito Kurihara

北海道大学大学院情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University

In recent years, deep reinforcement learning has attracted interests from AI researchers. Deep reinforcement learning is a method combining a deep neural network (DNN) and reinforcement learning (RL). By approximating a function in RL with a DNN, it enables an agent to learn in complex environment represented by low-level features, such as pixels given by a 3D video game. However, learning from low-level features is sometimes problematic. For example, a small difference in input pixels results in completely different behaviors of an agent. In this study, as an example of such problems, we focus on viewing directions of an agent in 3D virtual environment Minecraft and analyze the effect of them on the efficiency of deep reinforcement learning.

1. はじめに

近年, 人工知能 (AI) 技術は深層学習 (Deep Neural Network; DNN) により, 画像などの高次元データに対する情報処理能力を飛躍的に進化させた. これによりゲーム AI の分野においても, 画像データを用いてゲームの学習を行う Visual Learning の分野が開拓されてきた. 特に最近では google DeepMind によって提案された手法 Deep Q-Network (DQN) [Mnih 13] [Mnih 15] によって映像のみを入力とした学習で Atari2600 という簡単な 2D 環境のゲームにおいて人間と同等以上のスコアを獲得するに至った. また, 同チームは 2016 年に非同期な学習方式を提案し [Mnih 16], その内の一つである Asynchronous Advantage Actor-Critic (A3C) では必要スペックや計算量, 獲得した行動において DQN を圧倒する結果を記録している.

これらのアルゴリズムは深層強化学習と呼ばれている. その名の通り, 深層学習を用いた画像処理と試行錯誤から学習を行う強化学習を組み合わせた手法である. 深層学習を用いることで環境の状態を画像で定義することが可能となり, また深層学習の出力で強化学習の出力を近似することで, 少し工夫を加えた上で強化学習の手法を用いることができる. しかし, 画像データは状態としてはかなり複雑である. RGB の画像で考えると「画像の高さ」×「画像の幅」× 3(RGB) × 256(0 ～ 255) もの表現が可能である. 実際に学習を行う環境においてはある程度のパターンに制限されることがほとんどであるが, 既存の表現に比べてこれが複雑であることに変わりはない. さらに 3D 仮想環境において一人称視点で学習を行う場合, 視界に入る情報が制限され自身の置かれた状態を完全には知ることができない. また画像を入力とした深層強化学習では, いくつかの画素の色や視点の少しの変化による入力画像の差が行動に大きな影響を与えてしまうことがよく起こる. この 3D 仮想環境における一人称視点からの行動獲得という問題を解決するためには, この複雑さを克服したアルゴリズムを考案することが必要とされる.

そこで本研究では, アルゴリズム考案の前段階として, 入力情報の差を視野の仰角の変化に限定し, 3D 仮想環境で自由度の高い環境設定と問題設定を行える Minecraft を用いて, 学習過程や学習結果にどのような変化がみられるか調査することを目的とする.

2. 関連研究

この章では, 本研究の関連研究として深層強化学習と利用環境について述べる.

2.1 深層強化学習

深層強化学習は深層学習の出力で強化学習の方策や価値関数を近似するアルゴリズムである. これにより, 深層学習と強化学習を組み合わせ, 画像のような複雑な環境を学習することが可能となった. しかし, 方策や価値関数をニューラルネットワークのような非線形近似器を用いて表現して行う強化学習は安定しないことが知られている. これは強化学習に用いる時系列に並んだデータと方策, 価値関数の間に相関があることが原因と知られている. 深層強化学習の手法である DQN と A3C はそれぞれ異なる方法でこの問題を解決している.

2.1.1 Deep Q-Network

DQN では学習の安定のために 2 つの主な工夫を採用している. まず 1 つは経験再生 (Experience Replay, ER) である. これは過去一定期間においてある時間の「状態, 行動, 報酬, 次の状態」のタプルを経験として ER メモリに保存し, 一定周期で ER メモリからランダムにサンプリングしたミニバッチデータを用いて学習を行う手法である. これによりデータとの相関を打ち消すことができるが, 過去の方策に基づいたデータも学習に用いられるため方策オン型のアルゴリズムが利用できない点, ER メモリに相当な容量が必要となる点に注意が必要である.

またもう 1 つの工夫は固定目標ネットワーク (fixed target network) である. DQN における学習は一定期間のデータ収集の度に行われるが, この際に少しの価値関数の変化により方策が大きく更新されてしまうことがあり学習が安定しない. これを回避するために一度の学習の間は最適化するパラメータ θ のコピー θ^- を用いて最適化を固定することで学習を安定化させる.

2.1.2 Asynchronous Advantage Actor-Critic

A3C ではデータの相関を打ち消すために複数のスレッド上で並列にシミュレーションを行い学習用データを収集する. これにより, DQN においてランダムにサンプリングしたデータで学習を行ったことと同様の結果を得ることができる. それぞれのスレッドの動作は次のようになる.

1. global network のパラメータ θ を自身の local network のパラメータ θ^- にコピーする.
2. 現在のパラメータに従い探索を行い, 各時間ステップについて TD 誤差により計算される誤差関数から勾配 $d\theta$ を計算し累加する.
3. 一定の探索期間の後, 勾配 $d\theta$ を global network に伝え, 勾配法に従い θ を更新する.
4. $d\theta$ をリセットし 1. に戻る.

以上の処理を非同期並列に行うことでデータの相関を打ち消して学習を行う. ここで 1 においてパラメータをコピーすることは DQN の固定目標ネットワークと同様の働きをする.

2.2 Minecraft (Project Malmo)

Minecraft は Microsoft より販売されているサンドボックスゲームである. サンドボックスゲームとはビデオゲームのジャンルの 1 つであり, ゲームにより定められた目標が無く, 自由度の高い環境で自由に目標を定められることが特徴である. 特に Minecraft は主に立方体のブロックで構成される点を除き, 現実の空間と非常に近い環境を提供することができることが魅力である. この環境は次のような特徴を有する.

- エージェントは重力の影響を受けるが, 一部を除いたブロックは重力に影響されない.
- 条件により複数種の敵性生物が自然発生する.
- 時間の概念があり, 空間の明るさが変化する.
- バイオーム (生物群系) の概念があり, バイオームごとに違った地形, ブロック, 生物が自動で配置される.

また環境内において, エージェントは次のことが可能である.

- ほぼ全てのブロックは攻撃し続けることで破壊し, 入手することができる (一部専用のアイテムが必要).
- 入手したブロックは任意のブロックに隣接したスペースに再配置することができる.
- 入手したブロックは組み合わせることで別のブロックまたはアイテムを作成することができる (クラフティング).
- 作成したアイテムを使用する.

また 2016 年, Microsoft Research は Minecraft を用いた人工知能開発を目的としたプラットフォーム Project Malmo [Johnson 16] を GitHub にて公開した. これにより Minecraft の環境と実装したエージェントとの相互作用が容易になり, 以前は困難であった Minecraft の環境を簡単に利用することが可能となった.

3. 実装手法

この章では, 本研究の実験において実装した手法について述べる.

3.1 Asynchronous Advantage Actor-Critic

Asynchronous Advantage Actor-Critic(A3C) は深層強化学習の手法であり, 前章で述べた非同期な学習方式として提案された手法の 1 つである. 当時の最先端手法であった DQN を圧倒する結果を記録した [Mnih 16]. 現在においても, 3D 環境の探索タスクで最高水準の手法である. 非同期な学習方式については前節で説明したため, ここでは Advantage Actor-Critic について述べる.

Actor-Critic 手法は方策を定める Actor と状態を評価する Critic からなる強化学習の手法である. 深層強化学習においては, これらは CNN の出力で $\pi(s,a) \approx \pi(s,a;\theta), V(s) \approx V(s;\theta_v)$ と近似的に表現される. ここで, θ, θ_v は CNN のパラメータである. これらのパラメータは REINFORCE アルゴリズム [Willia 92] の勾配法を用いて最適化されるが, この際 Advantage Actor-Critic では現在の状態推定に Advantage 関数を用いる. Advantage 関数は次の式で表される.

$$A_t = \sum_{i=0}^{k-1} \gamma^i r_{t+1} + \gamma^k V(s_{t+k};\theta_v) - V(s_t;\theta_v) \tag{1}$$

Advantage 関数を用いることで, Bellman 方程式より計算された TD 誤差よりも先のステップを考慮に入れることができるようになる. これを用いて勾配は次のように計算される.

$$d\theta \leftarrow d\theta + \alpha \nabla_\theta \log \pi(s_t,a_t;\theta) A_t + \beta \nabla_\theta H(\pi(s_t;\theta)) \tag{2}$$

$$d\theta_v \leftarrow d\theta_v + \alpha \nabla_{\theta_v} A_t^2 \tag{3}$$

ここで式 (2) の第 3 項は確率の正規化のために加算されたエントロピーである.

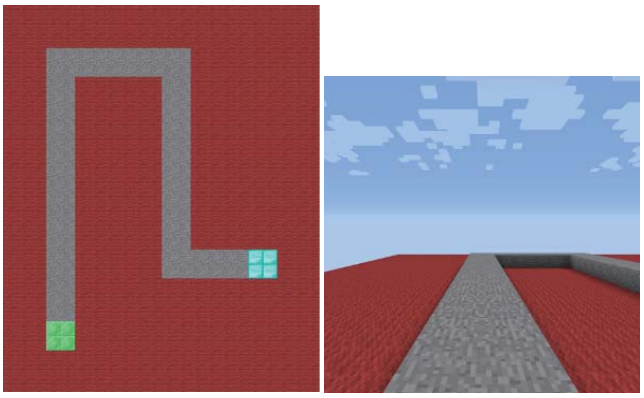
4. Minecraft による実験

4.1 実験目的

本実験では, 3D 仮想環境である Minecraft を用いた問題の学習において, 視野の仰角の変化によってどのような学習過程, 獲得行動の変化が見られるか調査を行うことを目的とする.

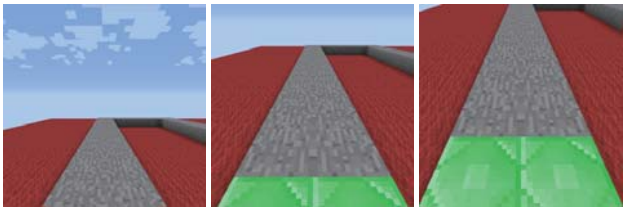
4.2 問題設定

この節では, 実際に Minecraft を用いた実験のエージェントと環境の設定について詳細を述べる. 本実験では視野の仰角の変化以外の変化を極力取り除くために簡単な問題設定として「一人称視点で一本道を落ちずにゴールまで進む」というタスクを用意した. 1 回のエピソードは, ゴールするか, 通路から落下するか, または制限時間に達することで終了し, Advantage 関数をスコアとして記録した. 環境は曲がり角を持つ一本道であり, エージェントは環境から一人称視点の RGB 画像 ($84 \times 84 \times 3$) を状態として受け取る (図 1). エージェントは状態から「前進する, 左を向く, 右を向く」を組み合わせた行動群から行動を確率的に選択し行動に対する結果として報酬を受け取る. 報酬は, ゴールした際に 1.0, 通路から落下した際に-1.0, またその場で回り続ける行動を獲得しないように, Minecraft の 1 ブロック分通路を進むたびに 0.1 の報酬を与えた. 実験ではこの問題をエージェントの視野の仰角を「$0°, -30°, -45°$」と変化させ, それぞれ 200 万ステップ (1 ステップはゲーム内時間で約 0.1 秒) 学習させた (図 2). 視点によって問題の根幹である通路について知ることのできる情報の量が変化するため, 調査内容にも変化が現れることが予想される.



(a) 環境の俯瞰図　　(b) エージェントの視点

図 1: 学習に用いた環境



(a) 仰角 0°　　(b) 仰角 −30°　　(c) 仰角 −45°

図 2: 視野の仰角を変化させた際の視点の変化

4.3 実験結果

4.3.1 学習過程への影響

まず 200 万ステップの学習において, 各 1 万ステップ毎にエピソードのスコアの平均の推移を調べた結果, 図 3 のようになった. より多くの通路に関する情報を視界に納めたエージェントほど, 早い段階で問題に適した行動を獲得していることが分かる.

4.3.2 獲得行動への影響

次に学習の結果獲得した行動について各視点のエージェントに 1000 回ずつエピソードを割り振り, 平均スコアとゴール到達を 100 とする平均進行度, ゴール到達率を調査した. 結果は図 4 のようになり, これもまた視界内における通路の情報の量に大きく影響を受けた結果となった. 実験 4.3.1 と比べてもおよそ対応した結果になっていることが見て取れる.

4.3.3 落下位置の比較

それぞれのエージェントについて落下位置についてプロットした結果, 図 5 のようになった. まず仰角 0° のエージェントは図 5 からもわかるように学習が不十分であり, 初めの直進という簡単な行動も学習する事ができていない. 仰角 −30°, −45° のエージェントに関してはゴールに到達可能なだけの行動を獲得している. しかし, これらのエージェントでも一定の割合で落下してしまっている. これは離散的かつ「する」か「しない」かの二値化された行動により, 曲がり角付近では向いている方向によって曲がり角を十分に認識できていないためであると考えられる.

4.4 考察

実験 4.3.1 より, 学習に関して視点の変化により学習の速度に大きな影響を与えることが伺える. 実験 4.3.2 に関しても獲得した行動に視点の変化による大きな影響があることがわかった. これらの結果より, 一人称視点の 3D 仮想環境の探索タスクを適切に学習するためには, 学習したい問題に関してより多くの情報が視界に含まれていることが望ましいとわかる. また実験 4.3.3 より, 視点を与えるだけでなく, 視点や行動を連続的または細かく多値化して制御することの必要性が考えられる. これらの実験結果より 3D 仮想空間における一人称視点での AI 開発のためには学習する行動に「視点の制御」を加えるべきであると考察する.

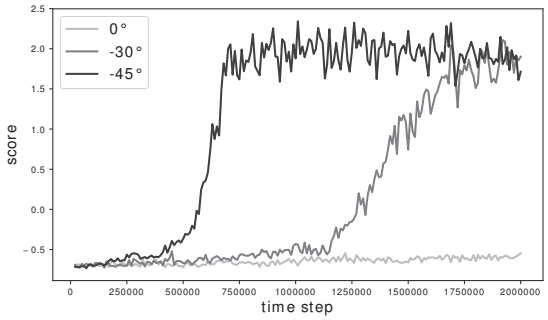


図 3: 学習ステップと平均スコアの遷移

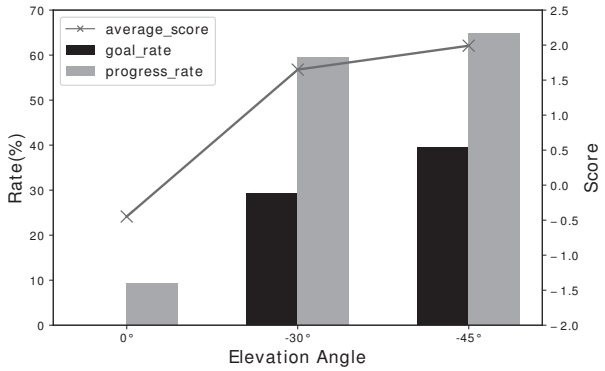


図 4: 1000 回の試行によるゴール率, 進行率, 平均スコア

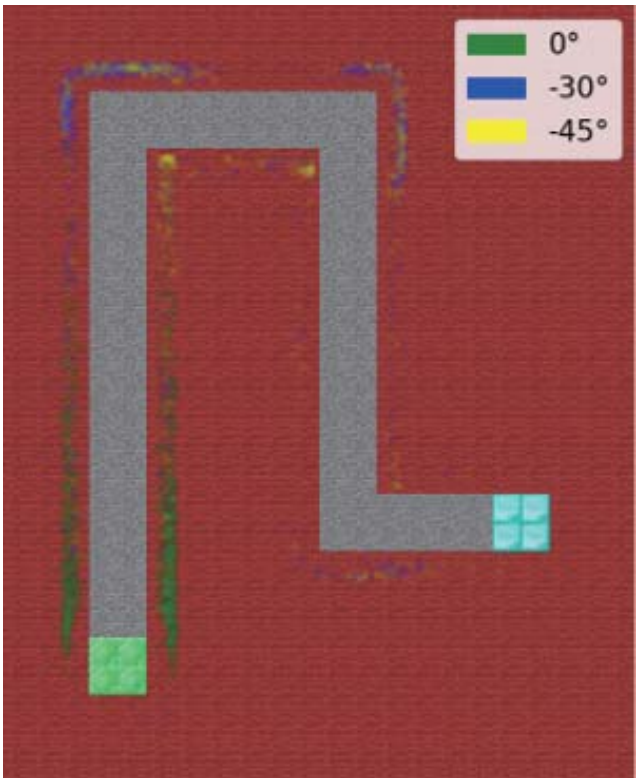


図 5: 視点の変化による落下位置の変化

5. まとめ · 今後の展望

本論分では, 強化学習を映像データ入力に対応させた深層強化学習の手法を紹介し, その手法の 1 つである A3C を用いて, 3D 仮想環境における一人称視点の行動獲得タスクの学習に視点の仰角の変化が与える影響を調査することが目的であった. その点において, 複数の視点に対して学習過程と学習結果について比較し調査することができたため, この目的は達成できた.

今後の展望として,「視点の制御」を加えた深層強化学習アルゴリズムの開発を行いたいと考えている. A3C アルゴリズムが 2016 年に発表されて以降, 様々な工夫を付け加えた手法が提案されている. その 1 つである自己教師付き予測による「好奇心」に導かれた探索を用いた研究 [Pathak 17] では, 行動の結果として環境から与えられる外部報酬とは別に, 入力である画像データから内部報酬を計算している. この内部報酬という考え方は 3D 仮想環境における深層強化学習及びその視点の制御において, 視点を評価するなど有効に働き得ると考えている. 今後はこの方法を第一に視点を制御するアルゴリズムを提案することを目標とする. 以上を今後の展望として本論分を締めくくる.

参考文献

[Johnson 16] Johnson, M., Hofmann, K., Hutton, T., and Bignell, D.: The malmo platform for artificial intelligence experimentation, in *IJCAI International Joint Conference on Artificial Intelligence*, Vol. 2016-Janua, pp. 4246–4247 (2016)

[Mnih 13] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M.: Playing Atari with Deep Reinforcement Learning, in *arXiv preprint arXiv:1312.5602*, pp. 1–9 (2013)

[Mnih 15] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D.: Human-level control through deep reinforcement learning., in *Nature*, Vol. 518, pp. 529–33, Nature Publishing Group (2015)

[Mnih 16] Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T. P., Harley, T., Silver, D., and Kavukcuoglu, K.: Asynchronous Methods for Deep Reinforcement Learning, in *International Conference on Machine Learning*, Vol. 48, pp. 1928–1937 (2016)

[Pathak 17] Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T.: Curiosity-Driven Exploration by Self-Supervised Prediction, in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Vol. 2017-July, pp. 488–489 (2017)

[Willia 92] Willia, R. J.: Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning, in *Machine Learning*, Vol. 8, pp. 229–256 (1992)