

POMDPs 環境のための Deep Q-Network の有効性の検証と改善

片岡大喜 長名優子

東京工科大学 コンピュータサイエンス学部

1 はじめに

Deep Q-Network[1] は、強化学習の代表的な手法である Q Learning [2] を深層学習の代表的な手法である畳み込みニューラルネットワーク [3] を用いて実現したものであり、様々なゲームにおいて有効性が確認されている。しかし、Deep Q-Network では、部分観測マルコフ (Partially Observable Markov Processes : POMDPs) 環境では適切に学習が行えない可能性がある。POMDPs 環境とは、実際に異なる状態がエージェントにとっては同一の状態として知覚されるような不完全知覚状態を有する環境のことである。

それに対し、POMDPs 環境のための Deep Q-Network[4] が提案されている。この手法は、入力として用いる観測の長さが異なる複数の Deep Q-Network を用いる手法である。通常は 4 フレーム分の観測を入力とする Deep Q-Network で出力された行動価値に基づいて行動選択を行うが、観測が不完全知覚状態であると判断された場合には 8 フレーム分の観測を入力とする Deep Q-Network で出力された行動価値に基づいて行動選択を行う。不完全知覚状態であるかどうかの判定は観測ごとの行動の決定度と学習開始時からのステップ数を用いて行う。しかしながら、文献 [4] では、Atari 2600 のゲームを題材として学習を行っているため、学習において不完全知覚状態の判定などが適切に行われているか確認されていない。また、不完全知覚状態であると判断された場合には 8 フレーム分の観測を入力として用いているが、何フレーム分の観測を用いるのが適切であるかの検討も行われていない。

本研究では、POMDPs 環境のための Deep Q-Network において、迷路問題を題材として、不完全知覚状態の判定が正しく行われているかや不完全知覚状態であると判断された際に用いるフレーム数などについて検討を行う。

2 POMDPs 環境のための Deep Q-Network

Verification of Effectiveness and Improvement in Deep Q-Network for POMDPs Environment
Taiki Kataoka and Yuko Osana (Tokyo University of Technology, osana@stf.teu.ac.jp)

本研究で扱う POMDPs 環境のための Deep Q-Network について説明する。POMDPs 環境のための Deep Q-Network では、入力として用いる観測の長さの異なる複数の Deep Q-Network を用いる。観測が不完全知覚状態であると判断された場合は、観測の長さが複数フレームの Deep Q-Network の行動価値、不完全知覚状態であると判断されなかった場合は、観測の長さが 1 フレームの Deep Q-Network の行動価値を用いて行動選択を行う。なお、不完全知覚状態であるかの判断は、観測ごとの行動の決定度と学習の進行度を用いて行う。

2.1 不完全知覚状態の判断

不完全知覚状態の判断は、行動の決定度と学習の進行度に基づいて行う。観測 o における行動の決定度 $d(o)$ は以下のように与えられる。

$$d(o) = \frac{1}{N} \sum_{a \in C^A} \left(q_n(o, a) - \frac{1}{|C^A|} \right)^2 \quad (1)$$

ここで、 $q_n(o, a)$ は観測 o において行動 a をとることの行動価値 $q(o, a)$ を観測 o におけるすべての行動価値の和が 1 になるように正規化したものである。また、 C^A はとることのできる行動の集合であり、 $\frac{1}{|C^A|}$ はすべての行動に対する行動価値が等しい状況における各行動の行動価値を表している。なお、 $q(o, a)$ は基本となる観測の長さが 1 フレームの Deep Q-Network に観測 o を入力したときに出力される行動 a の行動価値の値を用いる。また、 N は決定度 $d(o)$ の最大値を 1 にするような正規化定数であり

$$N = 1 - \frac{1}{|C^A|} \quad (2)$$

で与えられる。

行動の決定度 $d(o)$ の値は学習の初期には大きく、適切に学習が進めば小さくなっていくと考えられるため、学習が進行しているにも関わらず行動が確率的に選択されているときに不完全知覚状態であると判断する。なお、学習の進行度には学習開始時からのステップ数 t を用いる。学習時には

$$d(o) < \theta_1 \text{ かつ } t > \theta_2 \quad (3)$$

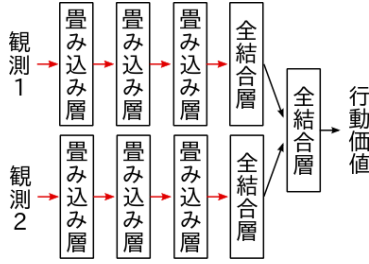


図 1: 提案手法における学習のとき、学習後には

$$d(o) < \theta_1 \quad (4)$$

であれば不完全知覚状態であると判定する。ここで、 θ_1, θ_2 は不完全知覚状態の判定に用いる決定度とステップ数のしきい値である。

なお、行動選択は ε -greedy 法を用いて行う。

2.2 学習

Deep Q-Network では、エージェントの観測を入力として、その観測におけるそれぞれの行動価値を出力するように学習を行う。出力となる行動価値は Q Learning における行動価値を用いるため、学習の際に用いられる誤差関数は以下のように与えられる。

$$E = \frac{1}{2} \left(r_t + \gamma \max_{a' \in C^A(o_{t+1})} q(o_{t+1}, a') - q(o_t, a_t) \right)^2 \quad (5)$$

ここで、 r_t は時刻 t における報酬、 $C^A(o_{t+1})$ は観測 o_{t+1} においてエージェントのとり得る行動の集合、 γ は割引率を表す。 $q(o_t, a_t)$ は観測 o_t が入力として与えられたときの行動 a_t に対応する行動価値であり、Deep Q-Network の出力の値が用いられる。また、 $r_t + \gamma \max_{a' \in C^A(o_{t+1})} q(o_{t+1}, a')$ がそれに対応する教師信号となる。

なお、学習の初期の段階では、1 フレーム分の観測を入力とする Deep Q-Network のみを用いるが、不完全知覚状態と判定された状態が出てきた段階で、複数フレーム分の観測を入力とする Deep Q-Network を生成し、学習を開始することになる。その際、Deep Q-Network の出力層の手前の層までの重みは1 フレーム分の観測を入力とする Deep Q-Network の重みと同じとし、出力層の重みのみを学習するものとする。図 1 では、観測 1 が1 フレーム分の観測、観測 1 と観測 2 が2 フレーム分の観測であるとする、赤で示した部分の重みがコピーされることになる。

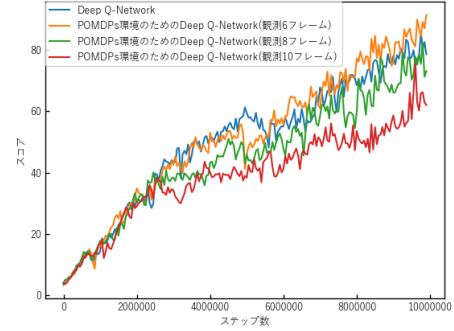


図 2: 獲得スコアの推移 (Amidar)

3 計算機実験

図 2 に Amidar というゲームにおける獲得スコアの推移を示す。Amidar は、あみだ状のフィールドに配置された敵を避けつつ、フィールド上のドットを回収するゲームである。エージェントは、ドットの回収により報酬を獲得する。いずれの学習においても 1000 万ステップまでスコアが増加しており適切な学習ができていたことが確認できる。550 万ステップ付近で観測 6 フレームのスコアが他の学習よりも高くなっており、最終的には観測 6 フレームの学習が最も高いスコアとなった。通常の Deep Q-Network よりも高いスコアを獲得していることから POMDPs 環境のための Deep Q-Network の有効性が示されている。

参考文献

- [1] V. Mnih, K. Kavukcuoglu, D. Silver, A. Grave, I. Antonoglou, D. Wierstra and M. Riedmiller : “Playing Atari with deep reinforcement learning,” NIPS Deep Learning Workshop, 2013.
- [2] C. J. C. H. Watkins and P. Dayan : “Technical Note: Q-Learning,” Machine Learning, Vol.8, pp.55–68, 1992.
- [3] V.Mnih et al. : “Human-level control through deep reinforcement learning,” Nature, No.518, pp.529–533, 2015.
- [4] 西川未来, 長名優子 : “POMDPs 環境のための Deep Q-Network,” 第 82 回情報処理学会全国大会, 2020.