

# POMDPs 環境のための Deep Q-Network

西川未来 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

近年、畳み込みニューラルネットワーク [1] に代表される Deep Learning は画像認識などの分野において従来の手法を上回る精度で認識が行えることで注目を集めている。また、機械学習の一手法である強化学習に関する研究も盛んに行われている。2013 年に Deep Q-Network[2] が Volodymyr Mnih らによって提案されて以来、Deep Learning と強化学習とを組み合わせた深層強化学習が注目され、盛んに研究が行われている。Deep Q-Network は、Q Learning[3] を畳み込みニューラルネットワーク [1] を用いて実現したもので、様々なゲームにおいて人間と同程度もしくはそれ以上の記録を出し、有効性が確認されている。しかしながら、Deep Q-Network では行動選択に  $\epsilon$ -greedy 法が用いられているため、同一の観測であったとしても異なる行動をとる必要のある状況を含む問題では適切に学習が行えないことも分かっている。異なる状況に対してエージェントの知覚能力が制限されているために同じ観測であると判断されている状態を不完全知覚状態といい、そのような状況を含む環境は、Partially Observable Markov Decision Processes (POMDPs) 環境と呼ばれる。

それに対し、POMDPs 環境において決定的な政策を学習する手法として、POMDPs 環境のための決定的政策を学習する Profit Sharing[4] が提案されている。この手法では、観測ごとの行動の決定度と学習の進行度を用いて不完全知覚状態の判断を行い、観測が不完全知覚状態であると判断されている場合には、現在の観測だけでなく観測の履歴を考慮して行動の選択を行うことで、POMDPs 環境下でも決定的な行動選択が行えるような学習が実現されている。

本研究では、POMDPs 環境のための Deep Q-Network を提案する。提案手法は、入力として用いる観測の長さの異なる複数の Deep Q-Network を用いる手法で状態の判断を行い、不完全知覚状態であると判断された場合には、より多くのステップにおける観測を入力とすることで、適切な行動が選択できるようにする。

Deep Q-Network for POMDPs Environment  
Mirai Nishikawa and Yuko Osana (Tokyo University of Technology, osana@stf.teu.ac.jp)

にする。

## 2 Deep Q-Network

Deep Q-Network[2] は、畳み込みニューラルネットワーク [1] に基づくモデルであり、3 層の畳み込み層と 2 層の全結合層から構成されている。ゲームを学習させる場合には、ゲームのプレイ画面の情報を観測として入力し、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。出力の行動価値は Q Learning のものを用いる。ゲームのプレイ画面においては、画面上でのわずかな位置のずれも重要な意味を持つと考えられるため、プーリング層は用いない。

この手法では、多くのゲームにおいて人間と同程度もしくはそれ以上の記録を出し、一部のゲームでは熟練した人間にも勝る成績を収め、有効性が確認されている。

## 3 POMDPs 環境のための Deep Q-Network

提案する POMDPs 環境のための Deep Q-Network は入力として用いる観測の長さの異なる複数の Deep Q-Network を用いる手法である。提案手法では、観測ごとの行動の決定度と学習の進行度を用いて不完全知覚状態の判断を行い、不完全知覚状態であると判断された場合には、より多くのステップにおける観測を入力として扱えるようなネットワークに切り替え、その出力によって行動選択を行うようにすることで、適切な行動が選択できるようにする。

### 3.1 構造

提案手法では、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。入力として用いる長さとしては様々なものが考えられるが、実験では 4, 8 の 2 つネットワークを用いる。なお、基本となるネットワークにおいて考慮する観測の長さを 4 としているのは、ゲームを題材として学習を行う場合には、

文献 [2] をはじめとして多くの研究において 4 フレーム分のゲーム画面が入力として用いられているためである。また、文献 [5] の研究において、すべての観測に関して考慮する長さを長くしても性能が上がらず、多くの場合、4 フレーム分を入力とした場合に高い性能が得られることが分かっている。

### 3.2 不完全知覚状態の判断

学習時には、観測  $o$  での行動の決定度  $d(o)$  と学習の進行度  $t$  を用いて不完全知覚状態であるかどうかの判断を行う。観測  $o$  における決定度  $d(o)$  は

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

で与えられる。ここで、 $q_n(o, a)$  は観測  $o$  における行動  $a$  の行動価値  $q(o, a)$  をすべての行動価値の和が 1 となるように正規化した行動価値、 $C^A$  はとることのできる行動の集合である。また、 $N$  は  $d(o)$  の最大値が 1 になるように正規化を行うための定数である。式 (1) において、 $1/|C^A|$  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選択確率を表している。行動選択が決定的に行われている場合には、特定の行動に対する行動価値の値が大きく、それ以外の行動に対する行動価値の値が小さくなっているため、行動の決定度  $d(o)$  の値は 1 に近い値をとることになる。

提案手法では、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を行う。POMDPs 環境のための決定的政策を学習する Profit Sharing[4] では、学習の進行度として観測ごとの価値の更新回数を用いているが、提案手法では観測ごとではなく、すべての観測に対する更新回数、つまり学習開始時からのステップ数  $t$  を学習の進行度として用いるものとする。行動の決定度がしきい値以下であり、ステップ数がしきい値以上であれば、その観測は不完全知覚状態であると判断される。

### 3.3 動作

学習済みのネットワークを用いる際には、はじめに入力として考慮する観測の長さの最も短い Deep Q-Network に観測を入力し、各行動に対する行動価値を出力する。出力された行動価値に基づいて不完全知覚状態であるかどうかの判断を行い、不完全知覚状態であると判断された場合には、より長い時刻の観測を考慮したネットワークに切り替え、出力された行動価値

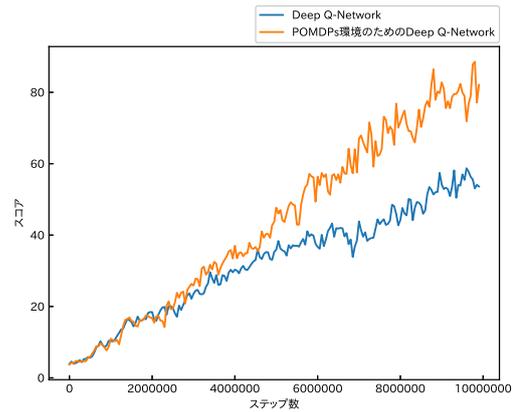


図 1: 獲得スコアの推移 (Amidar)

に基づいて行動選択を行う。行動選択には、 $\epsilon$ -greedy 法を用いる。

## 4 計算機実験

提案手法と Deep Q-Network[2] において学習を行った際の獲得スコアの推移の例を図 1 に示す。

### 参考文献

- [1] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner : “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, Vol.86, No.11, pp.2278–2324, 1998.
- [2] V. Mnih *et al.* : “Human-level control through deep reinforcement learning,” Nature, No.518, pp.529–533, 2015.
- [3] C. J. C. H. Watkins and P. Dayan : “Technical Note: Q-Learning,” Machine Learning, Vol.8, pp.55–68, 1992.
- [4] Y. Takamori and Y. Osana : “Profit sharing that can learn deterministic policy for POMDPs environments,” Proceedings of IEEE International Conference on System, Man and Cybernetic, Anchorage, 2011.
- [5] J. Niitsuma and Y. Osana : “Influence on learning of various conditions in deep Q-network,” Proceedings of IEEE International Conference on System, Man and Cybernetic, Banff, 2017.