

負の報酬を獲得する状況を重視した Deep Q-Network

浅沼駿哉 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、画像認識や音声認識の分野で従来の手法よりも優れた性能を示すとして Deep Learning[1] が注目されている。また一方で、教師信号を用いずに環境との相互作用により適切な行動系列を獲得するための学習手法として、強化学習に関する様々な研究が行われている [2]。そのような中で、Deep Learning と強化学習を組み合わせた手法を用いて学習を行う Deep Q-Network[3] が提案されている。Deep Q-Network は複数のゲームに適用され、ゲームによっては人間よりも高いスコアを獲得するほどの学習が実現されている。

Q Learning[4] では最も価値の高いルールを重視してルールの価値を更新していくため、この手法では正の報酬に対してのみ学習し、負の報酬を獲得しないことが重要になる課題に対して効率的な学習ができなくなってしまう。負の報酬を重視して学習を行う手法としては罰を受ける状況に関する情報の抽象化と強化学習効率化への利用 [5] や報酬分配を用いた Deep Q-Network の実現 [6]、負の報酬を獲得する状況を重視した畳み込みニューラルネットワークを用いた Profit Sharing[7] などが提案されている。

本研究では、負の報酬を獲得する状況を重視した Deep Q-Network を提案する。

2 Q Learning

Q Learning[4] では、エージェントの観測と行動の組をルールとし、将来もらえる期待報酬を考慮して学習する。観測 o_x において行動 a_x をとるというルールの価値 $q(o_x, a_x)$ は以下のように更新される。

$$q(o_\tau, a_\tau) \leftarrow q(o_\tau, a_\tau) + \alpha \left[r + \gamma \max_{a' \in C^A(o_{\tau+1})} q(o_{\tau+1}, a_{\tau+1}) - q(o_\tau, a_\tau) \right] \quad (1)$$

ここで、 $C^A(o_{\tau+1})$ は観測 $o_{\tau+1}$ においてエージェントのとり得る行動の集合、 r は報酬、 α は学習率、 γ は割

引率を表す。学習率は $0 < \alpha \leq 1$ の範囲の値をとり、値が小さいほど今までの行動価値の推定値を重視して価値の更新が行われることになる。割引率は $0 \leq \gamma \leq 1$ の範囲の値をとり、値が大きいほど将来獲得予定の報酬を重視しながら価値の更新が行われることになる。

3 Deep Q-Network

Deep Q-Network[3] では、ゲームのプレイ画面を観測として畳み込みニューラルネットワーク [8] に入力として与え、Q Learning における行動価値を出力するように学習を行う。Deep Q-Network は、ブロック崩し・シューティングゲームなどの様々なゲームにおける学習において、有効性が確認されている。

4 負の報酬を獲得する状況を重視した Deep Q-Network

ここでは、提案する負の報酬を獲得する状況を重視した Deep Q-Network について説明する。

4.1 構造

負の報酬を獲得する状況を重視した Deep Q-Network では、従来の Deep Q-Network と同様に 3 層の畳み込み層と 2 層の全結合層から構成される図?? に示すような構造を持つ畳み込みニューラルネットワークを用いる。

畳み込みニューラルネットワークへは観測が入力として与えられる。出力層は、(1) 負の報酬を獲得する可能性があるかを状況判断に対応する部分、(2) 負の報酬を獲得する可能性がある状況における行動価値、(3) それ以外における行動価値を表す 3 つの部分から構成されている。

4.2 学習

畳み込みニューラルネットワークを用いた Q Learning では観測を入力として、その観測におけるそれぞ

Deep Q-Network with Emphasis on Situation of Acquiring Negative Reward
Syunya Asanuma and Yuko Osana (Tokyo University of Technology, osana@stf.teu.ac.jp)

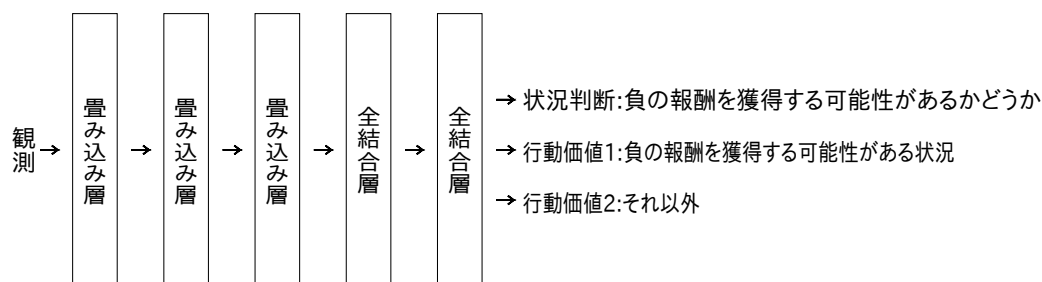


図 1: 負の報酬を獲得する状況を重視した Deep Q-Network の構造

れの行動価値と負の報酬を獲得する可能性があるかの状況を出力するように回帰問題として学習を行う。状況判断を表す部分ではニューロンの出力と教師信号の2乗誤差を用いて学習を行う。状況を表す部分の教師信号は学習の初期は0にして置き、障害物に衝突したときに教師信号の値を0から1に変更することでニューロンの出力が負の報酬を獲得する可能性があることを表すようにする。ニューロンの出力が表す行動価値はQ Learningのものを用いるため、学習の際に用いられる誤差関数は

$$E = \frac{1}{2} \left(r_\tau + \gamma \max_{a' \in C_{o_{\tau+1}}^A} q(o_{\tau+1}, a') - q(o_\tau, a_\tau) \right)^2$$

のような2乗誤差の式で与えられることになる。ここで、 r_τ は時刻 τ における報酬、 γ は割引率、 $C_{o_{\tau+1}}^A$ は観測 $o_{\tau+1}$ においてエージェントのとり得る行動の集合、 $q(o_\tau, a_\tau)$ は観測 o_τ において行動 a_τ をとることの価値を表す。

4.3 行動選択

負の報酬を獲得する可能性がある状況であるかを表すニューロンの出力の値により入力された観測の状況を判断し、行動選択を行う。状況判断を表す出力が負の報酬を獲得する可能性があることを表す1であれば、負の報酬を獲得する可能性がある状況での行動価値を用いて ε -greedy法により行動を選択する。状況判断を表す出力が0であれば、それ以外の状況での行動価値を用いて ε -greedy法により行動を選択する。

5 計算機実験

障害物回避を学習する課題として用いて実験を行った。エージェントは格子状に区切られたフィールドをスタートからゴールまで障害物を避けながら移動する。観測としては、エージェント視点の画像を用い、前進、右への方向転換、左への方向転換の3種類の行

動をとることができるものとしている。提案モデルを用いて実験を行い、学習が行えることを確認した。

参考文献

- [1] 岡谷貴之：機械学習プロフェッショナルシリーズ 深層学習，講談社，2015.
- [2] R. S. Sutton and A. G. Barto：Reinforcement Learning：An Introduction, The MIT Press, 1998.
- [3] 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.
- [4] C. J. C. H. Watkins and P. Dayan："Technical Note: Q-Learning," Machine Learning, Vol.8, pp.55-68, 1992.
- [5] 坂下裕太、村田 純一："罰を受ける状況に関する情報の抽象化と強化学習効率化への利用," 計測自動制御学会 第12回コンピュータショナル・インテリジェンス研究回, 2017.
- [6] 中矢裕太、長名優子："報酬分配を用いた Deep Q-Network の実現," 情報処理学会 第79回全国大会, 2017.
- [7] 志村成章、長名優子："負の報酬を獲得する状況を重視した畳み込みニューラルネットワークを用いた Profit Sharing による学習," 電子情報通信学会総合大会, 2019.
- [8] 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.