4U-06

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におけるルールの再利用

名取俊輝　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

深層学習と強化学習とを組み合わせた手法は深層強化学習と呼ばれる。Q Learning[1] を Deep Learning 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 [2] を用いて実現した Deep Q-Network[3] は、多くのゲームで人間や従来の手法を上回るスコアを獲得できることが示され、注目されている。また、Q Learning の代わりに Profit Sharing を畳み込みニューラルネットワークを用いて実現する手法 [4] も提案されている。この手法では、いくつかのゲームにおいて Deep Q-Network よりも高いスコアが獲得できるように学習が行えることが示されている [4][5]。Deep Q-Network[3] は Q Learning に基づいた手法であるため、多くの報酬を得られるような政策を獲得するように学習が行われることになる。しかし、課題によっては負の報酬を獲得しないことを学習することが重要な場合もある。また、障害物回避問題などにおいては、負の報酬を獲得する状況におけるルールは環境が変わっても再利用できる可能性がある。

負の報酬を獲得する状況を重視した学習を行う手法として、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[5] が提案されている。この手法では、負の報酬を獲得する可能性がある状況とそれ以外の状況とを区別して行動価値の学習を行っている。文献 [5] では障害物回避問題において実験を行い、畳み込みニューラルネットワークを用いた Profit Sharing[4] よりも短時間で学習が行える可能性があることが示されているが、複数の環境において実験が行われておらず、負の報酬を獲得する状況を重視して学習を行うことの有効性が十分に確認できているとは言えない。また、負の報酬が得られる状況におけるルールが環境が変わっても再利用できるかどうかの検証も行われていない。

本研究では、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[5] において複数の障害物回避問題において実験を行い、負の報酬を獲得する状況を重視することの有効性を確認する。

Reusing Rules in Profit Sharing using Convolutional Neural Network with Emphasis on Situation of Acquiring Negative Reward

Toshiki Natori and Yuko Osana (Tokyo University of Technology, osana@stf.teu.ac.jp)

2 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

ここでは、本研究で扱う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[6] について説明する。

2.1 構造

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 では図 1 に示すような 3 層の畳み込み層、2 層の全結合層から構成される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の状況判断を表す部分、(2) 負の報酬を獲得する可能性がある状況での行動価値、(3) それ以外の状況での行動価値を表す 3 つの部分から構成されている。

2.2 学習

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 乗誤差が学習の際に用いられる。この部分の教師信号は学習の初期は 0 にしておき、障害物に衝突した場合に教師信号の値を 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 Profit

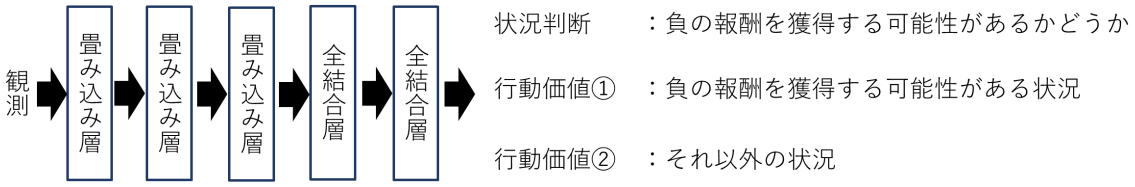

図 1: 負の報酬を獲得する状況を重視した畳み込みニューラルネットワークを用いた Profit Sharing の構造

Sharing のものを用いるため、学習の際に用いられる誤差関数の行動価値に関する部分は

$$E = \frac{1}{2}(rF(x) - q(o_x, a_x))^2 \tag{1}$$

で与えられることになる。ここで、r はエピソード内で得た報酬の合計、$F(x)$ は時刻 x における報酬分配関数の値、$q(o_x, a_x)$ は観測 o_x において行動 a_x をとることの価値を表す。

2.3 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 の場合には、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 グリーディ法により行動を選択する。状況判断を表す出力が 0 の場合には、それ以外の行動価値を用いて ε グリーディ法により行動を選択する。

3 計算機実験

負の報酬を獲得するような状況におけるルールが意味を持つような課題において実験を行うため、障害物回避問題を学習する課題として用いて実験を行った。エージェントはグリッド状に区切られたフィールドをスタートからゴールまで移動する。フィールド上には障害物が配置されており、その障害物にぶつからないように移動することを目指す。観測としては、エージェント視点で見た画像を用いる。行動としては、前進、右回り、左回りの 3 種類を考える。右回り、左回りともに曲がる角度は 90 度とする。エージェントはゴールに到達したときに正の報酬、障害物や壁などに衝突したときに負の報酬を得る。スタート地点、もしくは前回報酬を獲得した時点から次に報酬を獲得するまでを 1 エピソードとして扱う。

また、負の報酬が得られる状況におけるルールが環境が変わっても再利用できるかについて検証を行う。環境 A において学習を行った後に環境 B において学習を行った場合と、環境 B のみの学習を行った場合とで比較を行い、ルールの再利用が行えているかを検証する。環境 A において学習を行ったネットワークを用いて環境 B で学習を行う場合には、入力層から出力層の手前の全結合層までの重みと、出力層の負の報酬を獲得する可能性がある状況での行動価値に対応するニューロンと直前の全結合層のすべてのニューロンとの間の重みの値を初期値として用いた。

複数の環境において実験を行い、学習が行えること、負の報酬が得られる状況におけるルールが再利用できる可能性があることなどを確認した。

参考文献

[1] C. J. C. H. Watkins and P. Dayan : “Technical Note: Q-Learning,” Machine Learning, Vol.8, pp.55–68, 1992.

[2] 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.

[3] V.Mnih et al. : “Human-level control through deep reinforcement learning,” Nature, No.518, pp.529–533, 2015.

[4] 蓮池伸彬, 長名優子 : “畳み込みニューラルネットワークを用いた Profit Sharing によるゲームの学習,” 情報処理学会 第 80 回全国大会, 2018.

[5] 櫨場一貴, 長名優子 : “畳み込みニューラルネットワークを用いた Profit Sharing によるゲームの学習能力の検討,” 電子情報通信学会総合大会, 2019.

[6] 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.