7S-08

負の報酬を獲得する状況を考慮した畳み込みニューラルネットワークを用いた Profit Sharing への Experience Replay の導入

平間拓也　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

強化学習と深層学習とを組み合わせた深層強化学習に関する様々な研究が行われているが、それらの手法では一般的には正の報酬を得ることを重視してエージェントが得られる報酬を最大化するように学習が行われる。

そのような手法の 1 つとして畳み込みニューラルネットワーク [1] と Profit Sharing[2] を組み合わせた手法が提案されている [3]。しかしながら、障害物回避問題など、問題によっては負の報酬を獲得する状況を重視した学習が重要な場合もある。それに対し、負の報酬を獲得する状況を考慮した畳み込みニューラルネットワークを用いた Profit Sharing が提案されている [4]。この手法では、障害物回避問題を例題として実験を行い、負の報酬を獲得する状況におけるルールを異なる環境下で利用できる可能性があることが示されている。しかし、この手法では、学習する際に時間的に連続したデータを用いているため、データ間の相関が高くなってしまい偏りが生じてしまうという問題点がある。

この問題に対し、Experience Replay という手法が提案されている。これは、深層強化学習の代表的な手法である Deep Q-Network[5] で用いられている学習の工夫のひとつである。Experience Replay では、過去の経験の組み合わせをメモリに蓄積しておき、ランダムにサンプリングしてミニバッチを作成し、確率的勾配降下法による学習に利用する。また、その際に、教師信号も蓄積したメモリから取り出した値を用いて生成する。ランダムにサンプリングしてミニバッチを作成することは、時間的に連続するデータ間の相関の高さを軽減する効果もある。

本研究では、負の報酬を獲得する状況を考慮した畳み込みニューラルネットワークを用いた Profit Sharin への Experience Replay の導入を提案する。

Introduction of Experience Replay to Deep Q-Network with Emphasis on Situation of Acquiring Negative Reward

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

2 Profit Sharing

Profit Sharing[2] では、報酬をエピソード内のルールに分配することでルールの価値を更新する。エピソードとは、初期状態から報酬を獲得するまでの一連のルールのことである。

時刻 t において観測 o_t の場合に行動 a_t をとるルールの価値 $q(o_t, a_t)$ は

$$q(o_t, a_t) \leftarrow q(o_t, a_t) + \alpha \left(rF(t) - q(o_t, a_t) \right) \quad (1)$$

のように更新される。ここで、α は学習率、r は報酬を表す。また、$F(t)$ は時刻 t における報酬分配関数であり

$$F(t) = \frac{1}{(|C^A| + 1)^{W-t}} \quad (2)$$

で与えられる。ここで、C^A はとりうる行動の集合、W はエピソードの長さを表す。報酬を獲得した時刻に近いルールほど多くの報酬が分配されることになる。

3 負の報酬を獲得する状況を考慮した畳み込みニューラルネットワークを用いた Profit Sharing への Experience Replay の導入

ここでは、提案する負の報酬を獲得する状況を考慮した畳み込みニューラルネットワークを用いた Profit Sharing への Experience Replay の導入について説明する。提案手法では、畳み込みニューラルネットワークを用いて、負の報酬を獲得する可能性のある状況とそれ以外の状況を区別して Profit Sharing における行動価値を学習する。学習時には Experience Replay を用いることで連続するデータ間の相関が高くならないようにする。

3.1 構造

提案手法では、図 1 のような 3 層の畳み込み層と 2 層の全結合層から構成される 5 層の畳み込みニューラルネットワークを用いる。観測を入力として与えると、

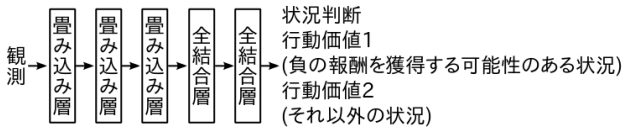

図 1: 提案手法で用いる畳み込みニューラルネットワークの構造

状況判断、行動価値が出力層から出力される。出力層は負の報酬を獲得する可能性があるかの状況判断、負の報酬を獲得する可能性のある状況での行動価値、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。

3.2 学習

負の報酬を獲得する可能性があるかの状況判断とそれぞれの状況での Proifit Sharing の行動価値を出力するように回帰問題として学習を行う。状況判断をするニューロンの教師信号は初期状態では 0 とし、障害物に衝突した場合に 1 へと変更する。そうすることによって状況判断をするニューロンは、負の報酬を獲得する可能性がある状況は 1、それ以外の状況では 0 となり負の報酬を獲得する可能性があるかの状況判断を行えるようになる。

学習に用いる誤差関数は

$$F = \frac{1}{2}\left(rF(t) - q(o_t, a_t)\right)^2 + \frac{1}{2}(d_t - x_t)^2 \quad (3)$$

で与えられる。ここで、r は報酬、$F(t)$ は時刻 t における報酬分配関数、$q(o_t, a_t)$ は時刻 t における観測 o_t の場合に行動 a_t をとるルールの価値、d_t は時刻 t における状況判断の教師信号、x_t は時刻 t における状況判断の出力を表している。また誤差関数の前半部分は行動価値に関する誤差、後半部分は状況判断に関する誤差を表している。

また、提案手法では学習において Experience Replay[5] を導入する。Deep Q-Network[5] で用いられる場合には、観測、行動、報酬、次の時刻の観測の 4 つをメモリに保存しておくが、提案手法では誤差関数で教師信号として用いる報酬分配量 $rF(t)$ の値を報酬と次の時刻の観測の代わりに保存しておく。また、学習時には状況判断に関する教師信号も必要となるため、その情報も合わせてメモリに保存しておく。つまり、観測、行動、報酬分配量、状況判断をメモリに保存しておき、メモリからランダムにサンプリングしてミニバッチを作成し、学習に利用することになる。誤差関数の報酬分配量と状況判断の部分の教師信号には保存しておいた値が利用される。

3.3 行動選択

行動選択を行う際には、まずはじめに負の報酬を獲得する可能性があるかの状況判断を表すニューロンの出力に基づいて負の報酬を獲得する可能性がある状況の行動価値とそれ以外の状況の行動価値のどちらを用いるかを決定する。決定された方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。ε グリーディ法とは ε $(0 \leq \varepsilon \leq 1)$ の確率でランダムに行動を選択し、$1 - \varepsilon$ の確率でルールの価値が最も高い行動を選択する方法である。ε の値は学習開始時には 1 に近い値に設定しておき、学習が進むにつれて小さくすることでルールの価値に基づいた行動選択を可能にする。

4 計算機実験

障害物回避問題を例題として実験を行った。エージェント視点の画像を観測とし、行動としては前進、右に 90 度回転、左に 90 度回転の 3 つをとることができる。報酬はゴールに到達したときに正の報酬、壁や障害物と接触したときに負の報酬を獲得するものとする。エージェントはスタートから壁や障害物にぶつからずにゴールに到達することを学習する。提案手法において、従来の手法と同様に学習が行えることを確認した。

参考文献

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

[2] J. J. Grefensette : "Credit assignment in rule discovery systems based on genetic algorithms," Machine Learning, Vol.3, pp225–245, 1988.

[3] 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.

[4] 名取俊輝, 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 におけるルールの再利用," 情報処理学会第 82 回全国大会, 2020.

[5] 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.