4U-08

畳み込みニューラルネットワークを用いたProfit Sharingの学習能力に関する検討

片井遼太　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

強化学習は機械学習の一種であり、現在、盛んに研究が行われている [1]。強化学習は環境との相互作用によって適切な政策を獲得する学習方法であり、データとそれに対応する教師データを用いて学習を行う教師あり学習や大量のデータのみから学習を行う教師なし学習とは大きく異なる。強化学習の手法は、Profit Sharing[2] に代表される経験強化型の学習と Q Learning[3] に代表される環境同定型の 2 つに大きく分けることができる。経験強化型の学習では、報酬を獲得する上で経験したルールに関する価値を強化することで学習を行う。それに対し、環境同定型の学習では、環境を同定し、最適な行動を行うための政策を獲得することを目的として学習が行われる。近年、強化学習とディープラーニングとを組み合わせた深層強化学習に関する研究も行われている。これまでに、畳み込みニューラルネットワークを用いた Profit Sharing では Atari2600 のいくつかのゲームを題材として実験が行われており [4][5]、Deep Q-Network を上回るスコアが獲得できることが分かっている。しかし、まだ、学習能力に対する十分な検討が行われているとは言えない。

本研究では、畳み込みニューラルネットワークを用いた Profit Sharing において Atari2600 のゲームを題材として学習能力に関する検討を行った。

2 畳み込みニューラルネットワーク

畳み込みニューラルネットワーク [6] は、全結合層のみから構成される一般的な順やプーリング層などの層から構成される構造を持つ。

畳み込み層では、畳み込み演算による特徴の抽出を行う。畳み込み演算はフィルタに類似した濃淡パターンを入力画像から検出する働きがある。畳み込み層の後、もしくは畳み込み層が複数回繰り返された後に、プーリング層が配置される。畳み込み層とプーリング層の組み合わせが繰り返される中で、局所コントラスト正規化層を挟むこともある。局所コントラスト層では、前の層の出力の明るさやコントラストを正規化している。最後に全結合層を通り、最終的な出力が出力される。

3 Profit Sharing

Profit Sharing[2] では、エージェントの観測と行動の組をルールとし、報酬をもとにルールの価値を更新することで学習を行う。エージェントが報酬を獲得したときに、初期状態から報酬を得るまでの一連のルール (エピソード) に報酬を以下のように分配する。

$$q(o_x, a_x) \leftarrow q(o_x, a_x) + \alpha[r \cdot F(x) - q(o_x, a_x)] \quad (1)$$

ここで、$q(o_x, a_x)$ は時刻 x における観測 o_x のときに行動 a_x をとるというルールの価値、r は報酬、α は競り値を表し、以前のルールの価値に報酬分配関数 $F(x)$ に基づいて分配された報酬を加算することで価値を更新している。報酬分配関数 $F(x)$ は

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

で与えられる。ここで、$|C^A|$ はエージェントのとり得る行動の種類の数、W はエピソードの長さ、x は時刻を表す。この関数は、報酬獲得の直前のルールに最も多く報酬が分配されることで、報酬獲得時の時刻から離れるほど分配される報酬の量が減るようになっている。行動選択には ε グリーディ法を用いる。

4 畳み込みニューラルネットワークを用いたProfit Sharing

本研究では、畳み込みニューラルネットワークを用いた Profit Sharing[4] の学習能力の検討を行う。このモデルでは Deep Q-Network と同様、ゲーム画面の

Study of Learning Ability of Profit Sharing using Convolutional Neural Network
Ryota Katai and Yuko Osana (Tokyo University of Technology, osana@stf.teu.ac.jp)

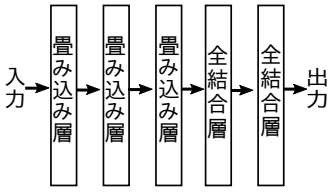

図 1: 本研究で用いるネットワークの構造

画像を観測として入力し、各行動の行動価値を出力とするように学習を行う。Deep Q-Network では教師信号として Q Learning の行動価値が使用されていたが、畳み込みニューラルネットワークを用いた Profit Sharing では教師信号として Profit Sharing の行動価値を使用する。

4.1 構造

本研究ではゲームの学習を行うため、特徴の位置やコントラストはとても重要なものとなる。プーリング層や局所コントラスト正規化層でこれらを調整してしまうと、適切な状況把握が行えなくなってしまうと考えられるため、プーリング層と局所コントラスト正規化層を使用しない図 1 の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。

4.2 学習

畳み込みニューラルネットワークを用いた Profit Sharing では、観測であるゲームの画面の情報を入力とし、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、回帰問題として学習を行う。出力の行動価値は Profiit Sharing のものを用いるため、学習の際に用いられる誤差関数は

$$E = \frac{1}{2}(rF(x) - q(o_x, a_x))^2 \qquad (3)$$

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 はそのエピソード内で得た報酬、$F(x)$ 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。ゲームの画面 o_x が入力されると、観測 o_x におけるすべての行動の価値が出力層から出力され、その行動価値を用いて ε グリーディ法で行動を決定する。

また、本研究では Deep Q-Network の学習で用いられている Reward Clipping の考え方の導入に関する検討も行う。ただし、報酬の値を ± 1 にクリッピングしてしまうと Profit Sharing において報酬分配を行う際に分配される報酬量が小さくなりすぎる可能性あるため、クリッピングする値を変えるなどの検討も行う。

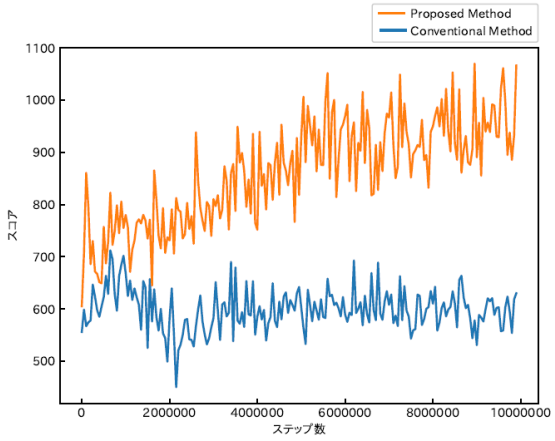

図 2: 獲得スコアの推移 (Centipede)

5 計算機実験

本研究では、畳み込みニューラルネットワークを用いた Profit Sharing において Atari2600 のゲームを題材として学習能力の検討を行った。その結果、いくつかのゲームにおいて Deep Q-Network を上回るスコアが獲得できることを確認した。図 2 に Centipede を畳み込みニューラルネットワークを用いた Profit Sharing(Proposed Method) と Deep Q-Network(Conventional Method) で学習した際のスコアの推移を示す。

参考文献

[1] R. S. Sutton and A. G. Barto : Reinforcement Learning : An Introduction, The MIT Press, 1998.

[2] J. J. Grefenstette : “Credit assignment in rule discovery systems based on genetic algorithms,” Machine Learning, Vol.3, pp.225–245, 1988.

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

[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] K. Hashiba and Y. Osana : “Study of learning ability in profit sharing using convolutional neural network,” Proceedings of IEEE International Conference on Artificial Intelligence and Soft Computing, Zakopnae, 2019.

[6] 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.

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