2R-04

畳み込みニューラルネットワークを用いたProfit Sharingによるゲーム学習の実現

村上魁一　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手法よりも優れた性能を示すものとしてDeep Learning[1] が注目されている．また，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]．強化学習は，Profit Sharing[3] に代表される経験強化型の学習とQ Learning[4] に代表される環境同定型の学習とに分けられる．そのような中で，Deep Learning の手法の1つである畳み込みニューラルネットワーク [5] とQ Learning を組み合わせた手法を用いて学習を行うDeep Q Network[6] が提案されている．Deep Q Networkでは，複数のゲームに対して人間と同等もしくは人間よりも高いスコアを獲得できるような学習が実現されている．この研究では，強化学習の手法としてQ Learning が用いられているが，Profit Sharingなどの他の手法と畳み込みニューラルネットワークを組み合わせることも可能であると考えられる．

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を提案し，2Dアクションゲームの学習を実現する．

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 [5] は，多層構造を持つ階層型のニューラルネットワークであり，畳み込み層とプーリング層と呼ばれる2種類の層のペアを複数重ねた構造を持つ．また，畳み込み層とプーリング層のペアの後に局所コントラスト正規化層が挿入されることもある．局所コントラスト正規化層では，1つ前の層の出力のコントラストの正規化を行う．最後に，最終的な出力が全結合層を通して出力される．

3　Profit Sharing

Profot Sharing[3] では，エージェントの観測と行動の組をルールとし，報酬を基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．

$$q(o_x, a_x) \leftarrow q(o_x, a_x) + r \cdot F(x) \tag{1}$$

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 を取るというルールの価値，r は報酬量を表し，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

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

で与えられる．ここで，$|C^A|$ はエージェントの取りうる行動の数，W はエピソードの長さ，x は時刻を表す．報酬獲得の直前のルールに最も多く報酬が分配され，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．

4　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

4.1　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 [5] を用いて，観測を入力として，それに対応する各行動の価値を出力するように学習を行う．なお，ここでは，Profit Sharing[3] における価値関数を用いる．

4.2　学習させるゲーム

本研究では，スタート地点からゴール地点までを目指す2Dアクションゲームにおける学習を行う．このゲームでは，スタートからゴールまで移動する間は，敵を倒す，またはパワーアップアイテムをとることにより，スコアが獲得される．また，ゴールまで到達した場合には，ゲームクリアまでに要した時間に応じてスコアが加算される．なお，敵にぶつかったり，崖から落ちることでゲームオーバーとなる．プレイヤーは，歩く (左右)，走る (左右)，ジャンプ (強弱)，止まるという操作ができる．また，複数の操作を同時に行うこ

Learning in Action Game by Profit Sharing using Convolutional Neural Network
Kaichi Murakami and Yuko Osana (Tokyo University of Technology, osana@stf.teu.ac.jp)

ともできるため，歩く (左右)+ジャンプ (強弱)，走る (左右)+ジャンプ (強弱) のような操作をすることもあり得る．

4.3 2D アクションゲームの学習

2D アクションゲームを畳み込みニューラルネットワークに学習させる際には，ゲーム画面を観測とし，各行動に対する行動価値を出力するように学習を行う．ここでは，Profit Sharing では，行動価値に基づいて 1 つの行動を選択するため，歩く (左右)+ジャンプ (強弱) のように，複数の操作からなるような行動をそのままでは表現することができない．そこで，本研究では，歩く (左右)，走る (左右)，ジャンプ (強弱)，止まるという基本的な行動に加え，歩く (左右)+ジャンプ (強弱)，走る (左右)+ジャンプ (強弱) のように 2 つの行動を組み合わせたものも行動として扱うものとする．また，Profit Sharing では，エピソードごとに報酬に基づいて行動価値の更新を行うが，スタートからゴールまでを 1 つのエピソードとしてしまうと，どの行動が報酬獲得において重要な役割を果たしているかが分からなくなってしまうため，スタートからゴールまでを複数のエピソードに分割して扱うものとする．スタートからゴールまでの間は，敵を倒すかパワーアップアイテムをとらない限りスコアは獲得されず，敵を避けながら進めば，ゴールに到達するまでの間にはまったくスコアが獲得されないこともあり得る．そこで，基本的には 5 秒ごとにエピソードを区切るものとする．ただし，スコアが獲得された場合には，そこをエピソードの区切りとし，エピソードの長さが 5 秒以上 10 秒未満になるように直前の 5 秒のエピソードも含めたものを 1 つのエピソードとして扱う．

4.4 エピソード e に対する報酬

エピソード e に対する報酬 $r(e)$ は以下のように与えられる．

$$r^{(e)} = r_{s1}^{(e)} + \frac{t^{(e)}}{5} r_{s2}^{(e)} + \frac{t^{(e)}}{5} r_n^{(e)} + \frac{t^{(e)}}{5} r_c \qquad (3)$$

ここで，$r_{s1}^{(e)}$ はエピソード e において獲得されたスコアに関する正の報酬を表し，$r_{s1}^{(e)}$

$$r_{s1}^{(e)} = s(e) \qquad (4)$$

のように与えられる．$s(e)$ はエピソード e 内で獲得されたスコアを表す．また，r_{s2} はゲームクリア時のスコアに関する正の報酬を表し，r_{s2} は

$$r_{s2} = s \qquad (5)$$

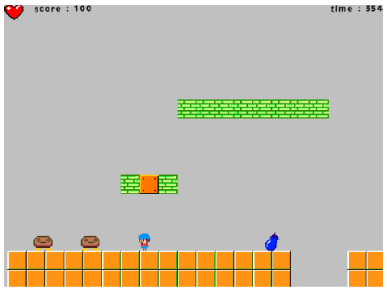

図 1: ゲームの実行画面

のように与えられる．s はゲームクリア時のゲーム全体のスコアを表す．$r_n^{(e)}$ はゲームオーバーによる負の報酬を表しており

$$r_n^{(e)} = \begin{cases} r_n, & \text{エピソード } e \text{ においてゲームオーバーになったとき} \\ 0, & \text{それ以外} \end{cases} \qquad (6)$$

のように与えられる．エピソード e においてエージェントがゲームオーバーになったときだけ，負の報酬を与える．また，r_c は各エピソードに与える一定の正の報酬を表す．エージェントがエピソード e においてスコアが更新できず報酬が与えられなかった場合でもゴールに近づく行動選択をしている可能性があるため，すべてのエピソードに対して一定の報酬を与えるものとしている．式 (3) において，r_{s2} と r_c はエピソード内の特定の事象に対して与えられる報酬ではなく，エピソード全体に対して与えられる報酬であるため，エピソードの時間に応じた値になるように $t^{(e)}/5$ を掛けることで調整を行っている．ここで，$t^{(e)}$ はエピソード e の時間であり，r_{s2} と r_c はエピソードの長さが 5 秒であるときの値としている．

5 計算機実験

提案手法を用いて図 1 のような 2D アクションゲームの学習を行い，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] J. J. Grefenstette : “Credit assignment in rule discovery systems based on genetic algorithms,” Machine Learning, Vol.3, pp.225–245, 1988.

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

[5] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner : Gradient-based learning applied to document recognition. Proceedings of the IEEE, Vol.86, No.11, 2278–2324, 1998.

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