5M-08

畳み込みニューラルネットワークを用いたProfit Sharingの実現

中矢裕太　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

人工知能の一分野として研究されている機械学習，その中でDeep Learningの手法は，画像認識分野などの研究により有名になっている[1]．2013年にVolodymyr Mnihらが提案したDeep-Q-Network[2]と呼ばれる手法は，強化学習の一種のQ Learning[3]を，Deep Learningの一手法である畳み込みニューラルネットワーク[4]で実現したもので，様々なゲームにおいて，人間と同程度もしくはそれ以上の記録を出し，有効性が確認されている．しかし，Q Learning以外の強化学習を用いた実例は見られない．

強化学習の学習方法は，Profit Sharing[5]に代表される経験強化型の学習と，Q Learningに代表される環境同定型の学習とに大きく分けられる．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．

本研究では，Profit Shearingの一部である報酬分配を取り入れたDeep-Q-Networkを実現する．この手法は，Deep-Q-Networkにおいて伝播されない負の報酬を分配することで，学習にかかる時間をさらに短縮できる可能性がある．この手法の有効性を確認するためDeep-Q-Networkとの比較を行う．

2　畳み込みニューラルネットワーク

Deep Learningの一手法である畳み込みニューラルネットワーク[4]は，脳の視覚野の構造における知見を参考にしており，ニューロンの層間結合を局所的であることが特徴で，主に画像認識分野に応用されている．

畳み込みニューラルネットワークは，図1に示すような多層構造のネットワークである．畳み込み層では，フィルタの濃淡パターンに類似した局所的特徴を抽出する．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．

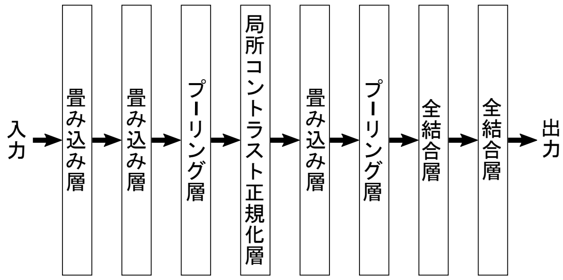

図1: 畳み込みニューラルネットワークの構造

3　Q Learning

Q Learning[3]では，エージェントの観測と行動の組をルールとし，将来もらえる報酬を計算することで学習を行う．エージェントが行動したときに獲得できる報酬と，次以降の行動の価値から，価値の更新をする．

$$q(o_x,a_x) \leftarrow q(o_x,a_x) + \alpha\left[r + \gamma \max_{a' \in C^A(o'_x)} q(o'_x,a'_x) - q(o_x,a_x)\right] \quad (1)$$

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を取るというルールの価値，rは報酬量，αは学習率$(0<\alpha\leq 1)$，γは割引率$(0\leq\gamma\leq 1)$を表している．$\max_{a'\in C^A(o'_x)} q(o'_x,a'_x)$の部分は次のステップで獲得できると予想できる最大報酬であり，割引率に応じた割合で加味して価値を更新している．

4　Profit Sharing

Profit Sharing[5]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．

$$q(o_x,a_x) \leftarrow q(o_x,a_x) + r \cdot F(x) \quad (2)$$

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

Profit Sharing using Convolutinal Neural Network
Yuta Nakaya and Yuko Osana (Tokyo University of Technology, osana@stf.teu.ac.jp)

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

$$F(x) = \frac{1}{(|C^A|+1)^{W-x}} \tag{3}$$

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．

5 Deep-Q-Network

Q Learning では，すべての状態と行動の組を学習することで最適な価値関数を得ることができるが，コンピュータゲームのように状態と行動の組み合わせが高次元になると，収束するまでにかかる時間が膨大になってしまう．そこで，高次元のデータを扱うことにたけており，画像認識などに用いられていた畳み込みニューラルネットワークを Q Learning と組み合わせた Deep-Q-Network という手法が提案されている．Deep-Q-Network では，入力に状態である画面情報をとり，出力にその状態におけるそれぞれの行動価値をとることで，学習すべき問題を回帰問題としてとらえ，学習を行う．

6 報酬分配を用いた Deep-Q-Network

提案する報酬分配を行う Deep-Q-Network では，Profit Shearing で用いられている報酬分配を追加で実装する．Q Learning では次の行動の最大の価値をもとに学習を行うので，正の報酬については時間とともに伝播されていく．しかし，負の報酬は最大の価値にはなりえないので，学習において伝播されない．そこで，負の報酬についてのみ Profit Sharing で行われている報酬分配を行い，間違った行動をとらないことをより学習しやすくする．

報酬分配は，以下のように行う．

$$r_\tau = \frac{1}{(|C^{s_\tau}|+1)^{\tau_r-\tau}} \tag{4}$$

ここで，r_τ は時刻 τ における報酬，$|C^{s_\tau}|$ は時刻 τ の状態において取ることができる行動数，τ_r は負の報酬を獲得した時刻を表す．なお，Profit Sharing では報酬の分配はエピソードが終了した後に行うが，今回は報酬の入手時に時刻を遡りながら分配を行うものとする．

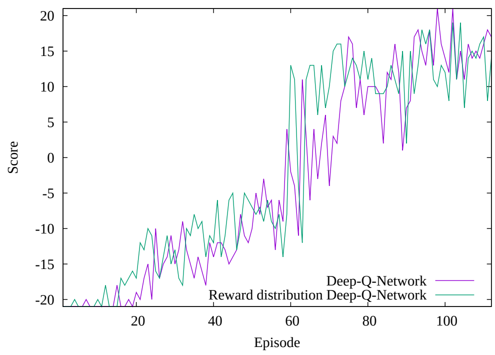

図 2: Pong のスコア推移比較

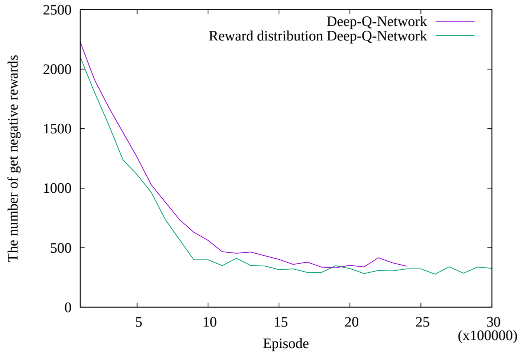

図 3: Pong の学習エピソードにおける負の報酬獲得回数

7 計算機実験

Deep-Q-Network と提案手法において Atari2600 のゲーム Pong を学習させたときのスコアの推移を図 2 に示す．これをみると両者の結果に大きな差は見られないことが分かる．図 3 は，両手法において獲得された負の報酬の遷移である．この結果より，提案手法の方が負の報酬を獲得しないように学習することができていることが分かる．

参考文献

[1] A. Krizhevsky, I. Sutsukever, and G. E. Hinton : ImageNet classification with deep convolutional neural networks. In Advances in NIPS, pp. 1097–1105, 2012.

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

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

[4] 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.

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