5M-06

畳み込みニューラルネットワークを用いた過去の履歴を考慮した強化学習

新妻純　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

教師信号を用いずに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 [1]。代表的な強化学習の手法である Q Learning[2] では観測され得るすべての状態において可能な行動の価値を評価する必要があり，状態数と行動数が大きくなると実用的ではない。そのためこの価値を関数近似するなどして応用することが多く，状態行動空間をどのように表現するかが問題となる。

また，一方で，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として Deep Learning[3] が注目されている。従来の画像認識や音声認識では，有効な特徴量を事前に抽出し，それに対して分類を行っていたが，Deep Learning ではデータから有効な特徴量を自動的に抽出することができる。

そのような中で，Deep Learning と Q Learning を組み合わせた手法を用いて学習を行う Deep Q-Network[4] が提案されている。Deep Q-Network は複数のアーケードゲームに対してそれぞれのゲームのための調整をせずに適用され，ゲームによっては人間よりも高いスコアを獲得している。しかし，4 時刻分の画面のデータを入力としているので，過去の履歴を考慮して行動を決定しなければならないゲームは苦手であるという問題がある。

本研究では，Deep Q-Network への入力として過去の履歴も含めたパターンを利用することにより，過去の履歴を考慮した学習について検討を行う。

2　強化学習

強化学習は取るべき行動を決定する問題を扱う機械学習の一種であり，教師信号を用いず，環境から報酬を得て学習が進行し，確率的な行動規則の獲得が行えるという特徴がある。

2.1　Q Learning

Q Learning[2] では，エージェントの観測と行動の組をルールとし，将来もらえる期待報酬に基づいて学習する。観測 o_x において行動 a_x をとるというルールの価値 $q(o_x, a_x)$ は以下のように更新される。

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

ここで，$C^A(o'_x)$ は観測 o'_x においてエージェントのとり得る行動の集合，r は報酬，α は学習率，γ は割引率を表す。学習率は $0 < \alpha \leq 1$ の範囲で設定し，値が小さいほどそれまでに学習されたルールの価値を重視することを意味する。割引率は $0 \leq \gamma \leq 1$ の範囲で設定し，値が大きいほど将来獲得予定の報酬を重視しながら価値の更新を行うことを意味する。

2.2　行動選択

(1) ε グリーディ選択

確率 $1-\varepsilon$ で最も価値の高いルールの行動を選択するグリーディ選択により行動を選択する。確率 ε $(0 \leq \varepsilon \leq 1)$ でランダムに行動を選択する。Deep Q-Network では ε グリーディ選択により，行動の選択を行っている。

(2) ボルツマン選択

ボルツマン選択では，観測 o のときに行動 a をとる確率 $P(o, a)$ は

$$P(o, a) = \frac{\exp\left(q(o, a)/T\right)}{\sum_{b \in C^A} \exp\left(q(o, b)/T\right)} \quad (2)$$

で与えられる。ここで，T は温度パラメータであり，学習の進行に伴い 0 に近づけていく。また，C^A はエージェントのとり得る行動の集合を表している。T の値は学習の開始直後では大きな値に設定されるため，行動はほぼランダムに選択される。学習が進むにつれて T の値は 0 に近づくため，価値の高いルールが高確率で選択されるようになる。

Reinforcement Learning by Convolutional Neural Network considerg History

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

3 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，Deep Learning と呼ばれる多層構造を持つ階層型のニューラルネットワークの代表的なモデルであり，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として注目されている。

3.1 構造

畳み込みニューラルネットワークは，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。複数の畳み込み層の後にプーリング層がくることもある。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，全結合層を通して，最終的な出力が出力される。なお，畳み込み層へ入力される信号は一般に複数のチャネルから構成されている。

3.2 畳み込み層

畳み込み層では，前の層の出力に対して畳み込み演算を行うことで，特徴抽出を行う。畳み込み層のニューロンは前の層のニューロンの一部のみと結合しており，結合している範囲を受容野と呼ぶ。受容野内のニューロンの出力に重みで表されるフィルタと類似したパターンが含まれているとき，その受容野と結合した畳み込み層のニューロンが反応し，活性化する。

第 l 層のフィルタ m に関する重みと結合したニューロン (i,j) の出力 $x_{ijm}^{(l)}$ は

$$x_{ijm}^{(l)}=f\left(\sum_{k=0}^{K-1}\sum_{p=0}^{H-1}\sum_{q=0}^{H-1}x_{i+p,j+q,k}^{(l-1)}h_{pqkm}^{(l)}+b_{ijm}^{(l)}\right) \quad (3)$$

で与えられる。ここで，K はチャネル数，H はフィルタの 1 辺のサイズ，$x_{i+p,j+q,k}^{(l-1)}$ は第 $l-1$ 層のチャネル k に関するニューロン $(i+p,j+q)$ の出力，$h_{pqkm}^{(l)}$ は第 $l-1$ 層のチャネル k の信号に対する受容野内の (p,q) の位置にあるニューロンから第 l 層のニューロンへのフィルタ m に関する重みである。また，$b_{ijm}^{(l)}$ は第 l 層のフィルタ m に関する重みと結合したニューロン (i,j) のしきい値を表している。$f(\cdot)$ は出力関数であり，正規化線形関数などが用いられる。

3.3 プーリング層

プーリング層では，畳み込み層からの出力を受け取り，プーリングを行う。畳み込み層の出力は，その前の層の出力の信号に含まれる特徴的なパターンを抽出したものとなっており，そのパターンが含まれる位置に対応するニューロンが活性化している。プーリング層のニューロンは，直前の畳み込み層の一部のニューロンのみと結合しており，受容野内に発火しているニューロンがあれば，発火するようになっており，位置のずれに対するロバスト性がある。

4 Deep Q-Network

Deep Q-Network[4] では，ゲームの 4 時刻分のプレイ画面を縮小してグレースケール化したものを観測として，畳み込みニューラルネットワークに入力として与え，Q Learning における行動価値を出力するように学習を行う。なお，行動選択には，ε グリーディ選択を用いている。Deep Q-Network では，Atari2600 のゲーム（ブロック崩し・シューティングゲームなど）の学習に適用し，有効性が確認されている。

5 過去の履歴を考慮した強化学習に関する検討

Deep Q-Network は 4 時刻分の画面のデータのみを入力としているため，過去の履歴を考慮して行動を決定しなければならないようなゲームの学習は苦手である。そこで Deep Q-Network への入力として過去の履歴も含めたパターンを利用するなどの方法で，過去の履歴を考慮した学習を実現する。本研究では，過去の履歴の考慮方法を変えて検討を行った。

参考文献

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

[2] C. J. C. H. Watkins and P. Dayan : “Technical note: Q-Learning,” Machine Learning, Vol.8, pp.279–292, 1992.

[3] 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.

[4] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra and M. Riedmiller : “Playing Atari with deep reinforcement learning,” NIPS Deep Learning Workshop, 2013.

[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, pp.2278–2324, 1998.

[6] S. Hochreiter and J. Schmidhuber : “Long short-term memory,” Neural Computation, Vol.9, No.8, pp.1735–1780, 1997.