5M-07

Deep-Q-Networkを用いた迷路の学習

市川椋太　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

近年，画像認識と音声認識の分野で従来の手法よりも優れた性能を示すとしてDeep Learning[1][2]が注目されている．畳み込みニューラルネットワーク[3]をはじめとするDeep Learningの手法では，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ことが知られている．また一方で，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[4]．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[5]が提案されている．Deep-Q-Networkは複数のゲームに適用され，ゲームによっては人間よりも高いスコアを獲得するほどの学習が実現されている．

本研究では，Deep-Q-Networkを用いた迷路の学習を行う．3D迷路において，観測としてエージェント視点で見た前方の迷路の画像を与え，適切な行動を選択できるように学習を行う．迷路の学習は，Deep-Q-Networkが得意とするゲームの学習であるが，将来的にはロボットによる経路探索問題などにも容易に拡張することのできる問題であると考えられ，ゲーム以外の問題への適用にもつながると考えられる．

2 Q Learning

Q Learning[6]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を取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．

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

ここで，$C^A(o'_x)$は観測o'_xにおいてエージェントの取りうる行動の集合，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，値が小さいほど今までの行動価値の推定値を重視しながら価値の更新を行うことを意味する．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．

行動選択にはボルツマン選択を用いる．観測oのときに行動aを取る確率$P(o, a)$は

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

で与えられる．ここで，Tは温度パラメータであり，時間経過とともに0に近づけていく．また，C^Aはエージェントが取りうる行動の集合である．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，価値の高いルールの行動が高確率で選択されるようになる．

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[3]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．図1に畳み込みニューラルネットワークの構成を示す．

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

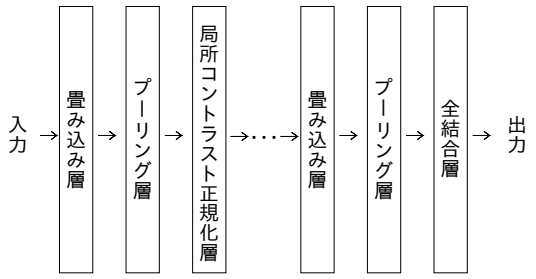

図 1: 畳み込みネットワーク

Learning in Maze Problem by Deep-Q-Network
Ryota Ichikawa and Yuko Osana (Tokyo University of Technology, osana@stf.teu.ac.jp)

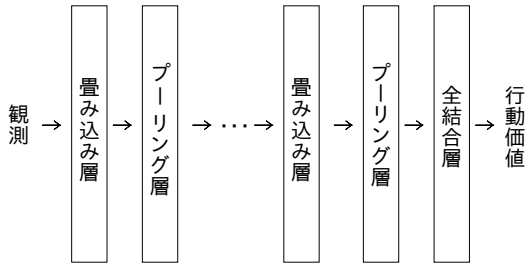

図 2: Deep-Q-Network の構造

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．

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．

畳み込みニューラルネットワークでは，勾配降下法を用いて入力に対する出力と教師信号の誤差が小さくなるように重みを学習していく．畳み込みニューラルネットワークでは，畳み込み層のフィルタに相当する重みは，同じフィルタに対応する重みが同じ値となる必要がある．これを重み共有といい，畳み込みニューラルネットワークでは重み共有の制約を考慮した上で勾配降下法により学習を行うことになる．

4 Deep-Q-Network

Deep-Q-Network[5] では，図 2 に示すようにゲームのプレイ画面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．Deep-Q-Network は，ブロック崩し・シューティングゲームなどの様々なゲームにおける学習において，有効性が確認されている．

5 Deep-Q-Network を用いた迷路の学習

本研究では，Deep-Q-Network を用いて 3D 迷路の学習を行う．図 3 に示すようなエージェント視点で見た前方の迷路の画像を観測とし，前進，左への方向転換，右への方向転換の 3 種類の行動の中から 1 つを選択する．報酬は，ゴールに到達することで与えられるものとし，報酬の値はゴールに到達するまでに要するステップ数が少ないほど大きく設定するものとする．

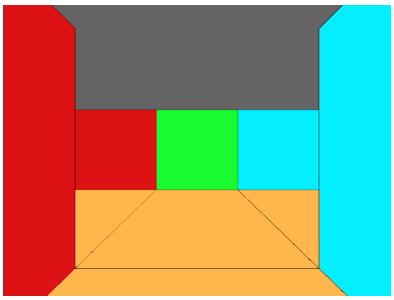

図 3: エージェントの視点からみた迷路

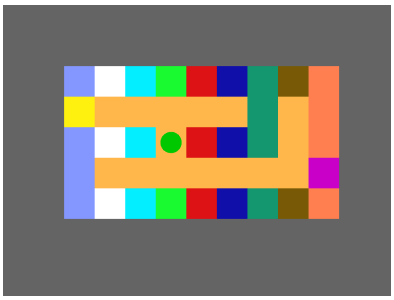

図 4: 学習に用いた迷路の例

6 計算機実験

エージェント視点で見た前方の迷路の画像を観測として学習を行う場合には，壁の色が均一であるとすると非常に多くの不完全知覚状態が存在するような環境となり，非常に難しい課題となってしまう．そのため，本研究では，迷路の壁の色が均一ではない図 4 のような迷路環境を用いることで，不完全知覚状態が多く存在しはするものの，その数が少なくなるようにしている．図 4 に示すような迷路において Deep-Q-Network を用いて学習を行い，学習が行えることを確認した．

参考文献

[1] 山下隆義：イラストで学ぶディープラーニング, 講談社, 2016.

[2] 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.

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

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

[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.

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