卒業研究概要

提出年月日　2021 年 1 月 29 日

卒業研究課題　迷路の最短経路問題を用いた強化学習パラメータの比較	
学生番号 B17028　　氏名 三田　大晃	
概要（１０００字程度）	指導教員　　印

本研究では Q-learning と迷路の最短経路学習を題材にして，Deep Q Network と比較した．とくにパラメータの変更が学習にどのような影響を与えるのかを検証した．図 1 のように正方形のマス目を s から g まで最短で結ぶことを目的とした．

Q-learning で学習にランダム性を取り込む ε-greedy 法を採用し、迷路を進むときの位置 s にて行動 a の持つ指標 Q(s,a)を学習ごとに更新した．Q 値の総量と得られた経路の長さ、学習の安定さなどを用いて，以下の 5 つの検証をした．①学習が収束するのにどれだけの学習回数が必要か．②迷路サイズの違いは必要な学習回数に影響するか．③ ε の値（ランダム行動をする割合）を変えることで学習にどのような影響を与えるか．④報酬の値を変えることで学習にどのような影響を与えるか．⑤学習率を与えることで学習にどのような影響を与えるか．(s,a)から(s’,a’)へと進む学習の過程で，Q 値の更新式は Q(s,a)→Q(s,a)+α(r(s,a)+γmaxQ(s’,a))となる．(α は学習率,r(s,a)は報酬,γ は割引率である)．

Q-learning について次のことがわかった．

① 学習の収束に 2 万回ほどの試行が必要.
② 迷路サイズが 10×10 より 5×5 の方が学習の収束にかかる時間が少ない.（図 2 ）
③ε（ランダム行動する割合）が 0.3 の時もっとも学習が上手くいく．
④ ゴールしなかったときの報酬 r(s,a)が-0.04 より 0 のときの方が学習の収束にかかる時間が少ない．
⑤ 学習率 α が 0.1 の時，収束に少し時間がかかるが学習の収束後 Q 値総量が安定する．

Deep Q Network でを使った経路学習では，何回か学習事例を蓄積して用いる Experience Replay により学習の偏りを抑え学習を安定させた．学習回数は Q-learning より必要なことがわかった．重みの初期化などのパラメータと必要学習回数に関しては卒業研究本文で解説する．

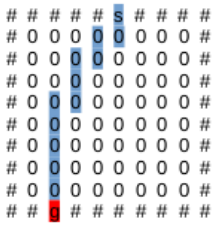

図 1:10×10 の迷路（s から g への最短経路を探す）

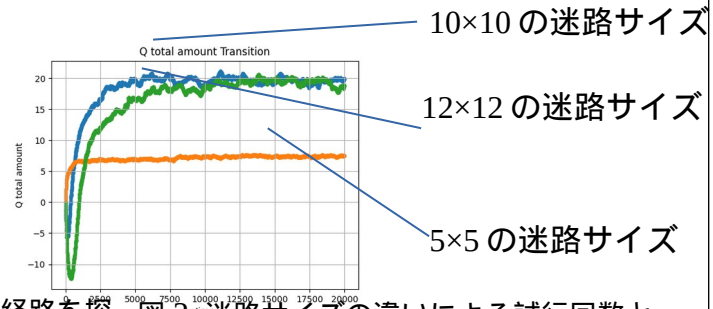

図 2:迷路サイズの違いによる試行回数と Q 値総量の違い