インタラクションを考慮した複数ブランチネットワークによるマルチエージェント深層強化学習

EP16050 五藤強志　　指導教授：山下隆義

1. はじめに

マルチエージェント強化学習 [1] は，複数のエージェントを同時に学習する強化学習手法である．マルチエージェント強化学習では，各エージェントが自己の利益を優先することでデッドロックが発生し，学習が停滞するという問題がある．そのため，デッドロックを回避する機構が不可欠である．本研究では，デッドロックを回避することを目的とし，単一ネットワークにおいてエージェントごとにブランチを分け，同時学習を行う手法を提案する．これにより，他のエージェントの学習を考慮し，デッドロックの回避が可能となる．

2. A3C

Asynchronous Advantage Actor Critic (A3C)[2] は Actor Critic 法をベースとする強化学習手法で，数ステップ先までの報酬を考慮する Advantage と複数の worker により経験を収集する分散学習を組み合わせた手法である．また，各 worker のパラメータ更新は非同期的に行う．これにより，経験の獲得の高速化と安定した学習を可能としている．図 1 に A3C の構造を示す．

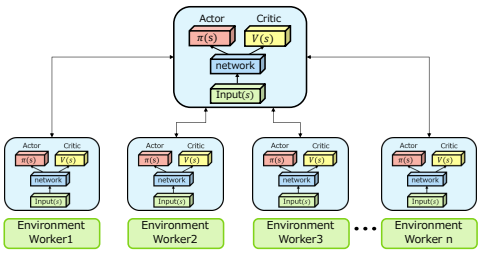

図 1：A3C の構造

3. 提案手法

独立したネットワークを複数用いて学習を行う場合，他エージェントの学習を反映する機構がないと自己の学習のみを考慮して行動するため，デッドロックの回避行動獲得が遅れるという問題がある．そこで，単一ネットワークでエージェントごとにブランチを導入し，同時学習を行う手法を提案する．全てのエージェントが畳み込み層を共有するネットワーク構造とする．これにより，他エージェントの学習を自分の学習に反映し，他エージェントとのインタラクションを考慮することが可能となる．デッドロックが発生した場合にデッドロックを回避する行動獲得の促進を図る．図 2 にネットワーク構造を示す．ネットワークには車両の鳥瞰画像，速度と曲率を入力する．各エージェントはそれぞれ対応したブランチからの出力を制御値とする．また，強化学習手法には A3C を用いる．提案手法を図 1 で示した A3C の各 worker のネットワークとし，学習を行う．

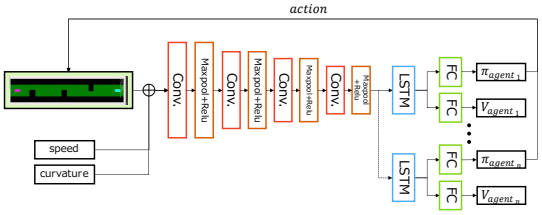

図 2：ネットワーク構造

4. 評価実験

自動運転環境において，提案手法の有効性を確認する．

4.1. 実験概要

環境には，図 3 のような 2 つの自動車エージェントが画面の端のゴールに向かって走行する自動運転環境を用いる．各自動車は，エージェントとして走行するレーンは固定とする．障害物は車両が 2 台以上通れない程度の間隔を空けて設置する．レーンにはランダムに障害物を配置する．

表 1：正の報酬

	報酬値
ゴール報酬	+10
車両の前進	+0.5
左側通行	+1.5

表 2：負の報酬

	報酬値
衝突ペナルティ	-10
速度変化によるペナルティ	-0.5
車間距離によるペナルティ	-3

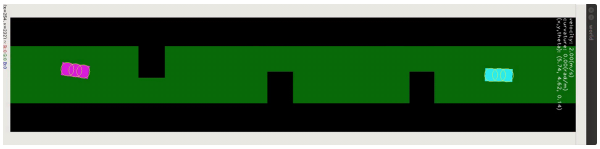

図 3：自動運転環境

エージェントの報酬設計を，表 1，2 に示す．また，エピソードの終了条件は，エージェントが道路外，障害物及び相手車両に衝突した場合と一定時間の経過で終了とした．各エージェントの行動は速度増減，変化なし曲率増減，変化なしの組み合わせで合計 9 通りであり，最大速度は 3m/s，最低速度が 0m/s，最大曲率が 0.25，最低曲率が-0.25 とした．

エージェント毎に独立したネットワークで学習する (共有なし) と提案手法 (共有あり) の 2 通りを比較する．学習は 1.0×10^7 ステップまで行う．評価では，ステップ数ごとのステップ及び図 5 で示すような道路の長さを制限したデッドロックを誘発する環境での 1000 試行における衝突回数で比較する．

4.2. 実験結果

ステップ数毎のスコア推移を図 4 に示す．図 4 から，共有ありは共有なしと比較してより高いスコアを獲得できている．このことから，提案手法を用いることでデッドロック状態を解消する学習が可能である．また，共有なしのエージェントはどちらも同じスコア推移であるのに対して，共有ありのエージェント同士を比較すると agent2 が agent1 より，スコアが低下している．これは agent2 が他エージェントを考慮して学習しているため，スコアに差が生まれたと考えられる．デッドロックを誘発する環境における衝突回数を表 3 に示す．表 3 から，共有なしと比較し共有ありの衝突回数が 395 回減少した．このことから，提案手法を用いることでデッドロックを回避することが可能である．以上より，提案手法を用いることでデッドロックが発生した場合における回避行動の獲得を促進できた．

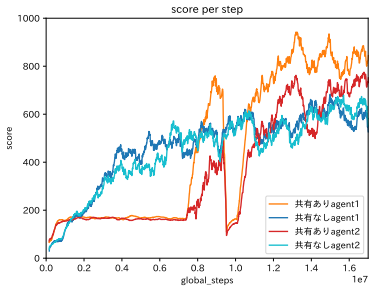

図 4：スコア推移

表 3：衝突回数

	衝突回数
共有なし	997
共有あり	602

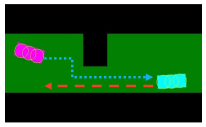

図 5：デッドロック環境

5. おわりに

本研究では，他エージェントの学習を考慮してデッドロックを回避する行動を獲得する手法を提案し，有効性を確認した．今後は，他環境における提案手法の有効性調査や，報酬以外を考慮するモデル模索を行う．

参考文献

[1] 三上貞芳, “Reinforcement Learning for Multi-Agent Systems”, 人工知能学会誌, 1997.

[2] V. Mnih , *et al.*, “Asynchronous methods for deep reinforcement learning”, ICML, 2016.