F-030

RoboCup サッカーの Keepaway タスクへの深層強化学習の適用
Application of Deep Reinforcement Learning to the RoboCup Soccer Keepaway Task

田村 啓朗† 相馬 隆郎†
Hiroaki Tamura Takao Soma

1. はじめに

近年，強化学習はディープラーニングの技術と融合し，ロボットの制御やゲーム AI など様々な分野での応用が研究されている．複数のエージェントが存在する環境においては，単一エージェントでは達成不可能な課題や，複数のエージェントが協力することでより高い成果を得られるよう学習する，マルチエージェントの強化学習が用いられる．

RoboCup Soccer Simulation 2D (以下 RCSS)は自律移動型エージェントによるサッカーの 2D シミュレーション環境である．本研究では RCSS のサブタスクであり強化学習の実験に適した Keepaway タスクで、マルチエージェントシステムにおける Deep Q-Network の適用を試みた．

2. Keepaway

2.1 問題設定

Keepaway は Stone ら[1]によって提案されたタスクである．Keepaway は正方形のフィールド内に Keeper チーム及び Taker チームのエージェント，そしてボール 1 つが配置され，Taker チームにボールを取られないよう Keeper チームがパス回しを学習するタスクである．Taker チームがボールを取るか，ボールがフィールド外に出るとエピソードは終了となる．

Keeper チームのエージェントはボールを持っているステップにおいて行動選択の学習を行う．

2.2 状態表現

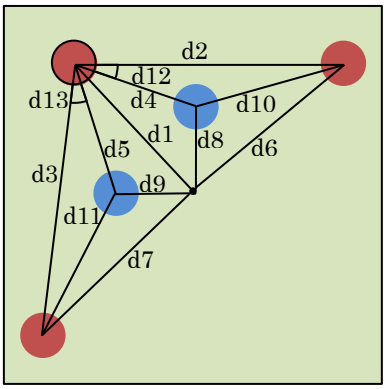

図 1　3vs2 Keepaway と各状態変数

Keepaway ではボールを保持している Keeper は状態を表す情報として表 1 に示す情報が実数で与えられる．ただし要素数の項目は Keeper チームが K 台，Taker チームが T 台の場合の情報の次元を表す．標準的な Keepaway タスクではフィールドの一辺の長さは 20[m]であり，Keeper チームは 3 台，Taker チームは 2 台で構成される．このとき状態変数の次元は 13 となる．本研究でもこの条件を使用した．

表 1　Keepaway の状態変数

	要素数
自身のフィールド中心からの距離	1
他の各 Keeper からの距離	K-1
各 Taker からの距離	T
他の各 Keeper の中心からの距離	K-1
各 Taker の中心からの距離	T
他の各 Keeper の最近接 Taker からの距離	K-1
他の各 Keeper について各 Taker となす角の最小値	K-1

2.3 従来研究

Stone ら[1]の研究ではボールを保持している Keeper の行動選択肢は，i (1≦i≦K-1)番目に近い他の Keeper にパスを出す，または保持を続けるという K 通りであった．

また伊佐野ら[2]の研究ではフィールド中心方向に対する法線方向への条件付きドリブルが導入された．敵が接近し，他の Keeper へのパスコースが塞がれ，ドリブル方向に Keeper も Taker も存在しないという条件が全て満たされた場合にドリブル行動を取る手法が提案されている．

また Kurek[3]の研究では学習に Deep Q-Network が用いられ，他の学習アルゴリズムと比較して最も高い性能が得られている．

3. ドリブル行動の導入

本研究では前述の Stone ら[1]によって定義された保持及びパス行動に加えて，Keepaway エージェントのマクロ行動より，フィールド中心方向から固定された角度± θ の方向へ固定された速度 V でドリブルするという行動を，ボールを持つ Keeper の行動選択肢として追加した．

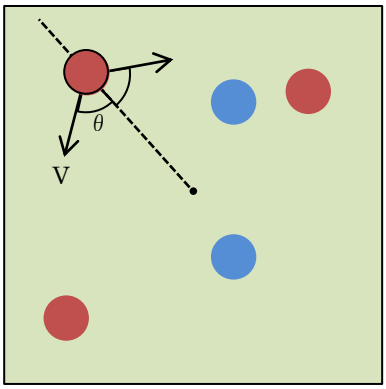

図 2　ドリブル行動

4. 学習方法

本研究では Deep Q-Network を用いた．ニューラルネットワークの構造及び学習におけるハイパーパラメータは Kurek[3]の研究を参考に決定した．ニューラルネットワークの構造を表 2 に示す．ただし，出力層のニューロン数は行動選択肢の数によって決定するので，Keeper が 3 台のとき，ドリブル無しの場合は 3，ドリブル有り場合は 5 となる．方策はのボルツマン選択とし，ミニバッチのサイズは 128 とした．また，Stone ら[1]と同様に，1 ステップごとに 1 の報酬を与えた．

表 2　ニューラルネットワークの構造

	ニューロン数	活性化関数
入力層	13	(なし)
中間層 1	200	ReLU
中間層 2	30	ReLU
出力層	3 または 5	(なし)

5. 実験

本稿ではドリブルの速度は FAST,SLOW, WITHBALL の 3 段階，ドリブル角度 θ は 70 度, 80 度, 90 度の 3 種類ついて，これらの全ての組み合わせで実験を行った．また，比較のためドリブルをしない場合も実験した．ドリブルを追加しなかった場合の学習曲線を図 3 に，ドリブル速度 FAST,SLOW, WITHBALL の場合の各角度における学習曲線をそれぞれ図 4，図 5，図 6 に示す．また，ドリブル有りとドリブル無しの場合の 30000 エピソード学習後におけるエピソード継続時間について，手コード，ランダム行動，パスなしとの比較を図 7 に示す．

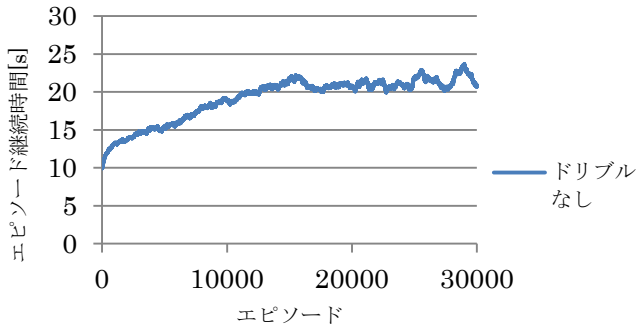

図 3　ドリブル無しの場合の学習曲線

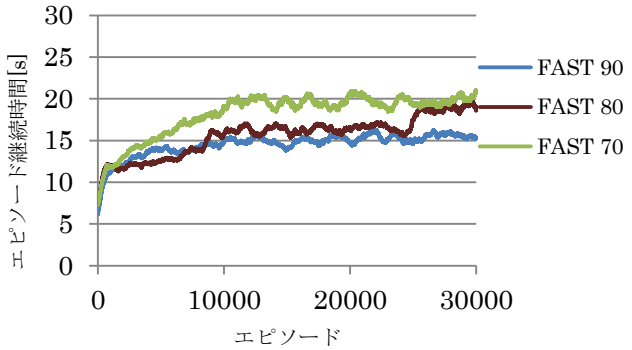

図 4　ドリブル速度 FAST の場合の学習曲線

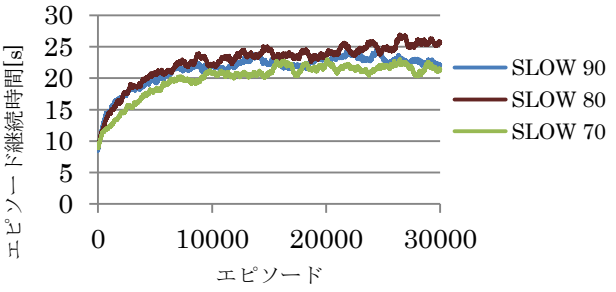

図 5　ドリブル速度 SLOW の場合の学習曲線

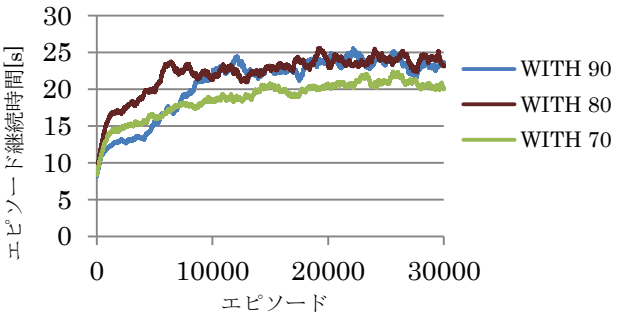

図 6　ドリブル速度 WITH BALL の場合の学習曲線

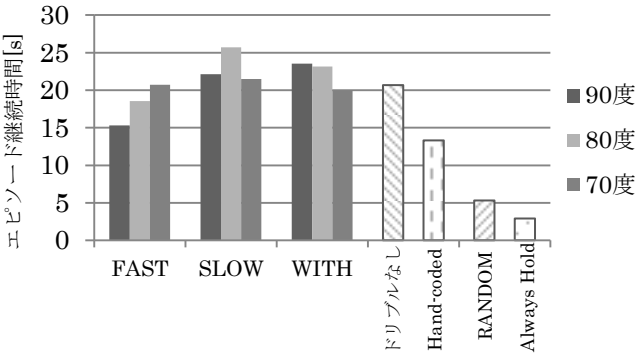

図 7　エピソード継続時間の比較

6. おわりに

本研究では Keepaway タスクに Deep Q-Network を適用し，さらに Keeper の行動選択肢にドリブル行動を追加する実験を行った．ドリブルの角度や速度を変えて試行を重ねた結果，速度を SLOW，角度をフィールド中心方向に対して 80 度としたドリブルが最も優れていることがわかった．

参考文献

[1] Stone, P. and Sutton, R. S. “Reinforcement Learning toward RoboCup Soccer” in Proceedings of 18th International Conference on Machine Learning, pp.537-544, (2001).
[2] 伊佐野勝人, 片上大輔, 新田克己, “ロボカップサッカーシミュレーションの Keepaway における協調行動の学習”, 論文誌名, Vol.n, No.n (2008).
[3] Mateusz Kurek, “Deep Reinforcement Learning in Keepaway Soccer”, 論文誌名, Vol.n, No.n (2015).