7S-02

Half Field Offenseにおける好奇心駆動探索とコミュニケーションを利用した深層強化学習

田村 純一　　李 嘉誠　　能登 正人
神奈川大学大学院工学研究科工学専攻電気電子情報工学領域

1 はじめに

近年 RoboCup 2D サッカーシミュレーションのサブタスクで，強化学習環境のテストベットとして用いられている Half Field Offense (HFO) への深層強化学習を適用する研究がされている．HFO はマルチエージェント強化学習環境として課題が多く存在する．この課題の一つとして同時学習における非定常性と学習エピソードの増加が挙げられる．従来の手法では離散行動のみか連続行動のみを扱っているためパラメータ化された行動には対応しておらず HFO のような環境では利用することができない．

本研究ではこの課題を解決するために Random Network Distillation[1] とコミュニケーションを組み合わせた手法を提案する．提案手法は HFO のマルチエージェント強化学習環境において同時学習が可能となり，コミュニケーションなしの場合と比べて学習効率が改善することを実験結果より示す．

2 先行研究

2.1 Deep Deterministic Policy Gradient (DDPG)

DDPG は決定論的な方策勾配法とニューラルネットワークを用いることで連続状態，連続行動の環境でも学習することが可能な深層強化学習の一種である．Critic ネットワークを $Q(s,a|\theta^Q)$，Actor ネットワークを $\mu(s|\theta^\mu)$ とし θ^Q，θ^μ で重み付けする．また，学習の安定化のために利用されるターゲットネットワークもそれぞれ Q'，μ' とし $\theta^{Q'}$，$\theta^{\mu'}$ で重み付けする．学習には式 (2) の損失関数により Critic ネットワークを，式 (3) の勾配を利用して Actor ネットワークを更新する．

$$y_t = r_t + \gamma Q'(s_{t+1}, \mu'(s_{t+1}|\theta^{\mu'})|\theta^{Q'}) \tag{1}$$

$$L_Q(\theta^Q) = \mathbb{E}_{(s_t,a_t,r_t,s_{t+1})\sim D}[(Q(s_t,a_t|\theta^Q) - y_t)^2] \tag{2}$$

Deep Reinforcement Learning with Curiosity-driven Exploration and Communication in Half Field Offense
Junichi Tamura, Jiacheng Li and Masato Noto
Graduate School of Electrical, Electronics and Information Engineering, Kanagawa University

$$\nabla_{\theta^\mu}\mu = \mathbb{E}_{s_t\sim D}[\nabla_a Q(s_t,a|\theta^Q)\nabla_{\theta^\mu}\mu(s_t|\theta^\mu)] \tag{3}$$

ここで，式 (1) は教師あり学習のターゲットであり，D はエージェントが μ に従って行動した軌跡を保存しておくメモリである．Hausknecht ら [2] は DDPG をパラメータ化された行動空間に適応させた．

2.2 Random Network Distillation (RND)

RND はランダムに初期化された固定のターゲットネットワーク $\hat{f}(s_{t+1}|\theta^{\hat{f}})$ の出力と，エージェントによって収集されたデータで訓練する予測ネットワーク $f(s_{t+1}|\theta^f)$ の出力間との誤差である式 (4) を内部報酬 r_t^i とする．

$$r_t^i = \|f(s_{t+1}|\theta^f) - \hat{f}(s_{t+1}|\theta^{\hat{f}})\|_2^2 \tag{4}$$

実験では RND ネットワークと DDPG ネットワークの更新は同じタイミングで行う．

3 提案手法

本研究ではエージェント 2 体がそれぞれ図 1 のような RND と DDPG を使用し同時学習することを考える．タイムステップ t でのエージェント 2 が観測する状態を $s_t^{(2)}$ とし，コミュニケーションで利用するメッセージは $m_t^{(2)} = s_t^{(2)}$ となる．このメッセージを式 (4) における各ネットワークへの入力として扱うためエージェント 1 の状態 $s_t^{(1)}$ と結合するとエージェント 1 の内部報酬は式 (5) となる．

$$r_t^i = \|f(s_{t+1}^{(1)} \oplus m_{t+1}^{(2)}|\theta^f) - \hat{f}(s_{t+1}^{(1)} \oplus m_{t+1}^{(2)}|\theta^{\hat{f}})\|_2^2 \tag{5}$$

また，RND による内部報酬 r_t^i と環境から得られる外部報酬 r_t^e を足し，報酬 r_t とするため，エージェント 1 のターゲット y_t は式 (6) のようになる．

$$y_t = r_t^i + r_t^e + \gamma Q'(s_{t+1}^{(1)}, \mu'(s_{t+1}^{(1)})) \tag{6}$$

式 (5) は RND の損失関数としても扱い，ネットワークの更新は式 (6) を利用したエージェント 1 の DDPG ネットワークの更新と同じタイミングに行う．ここまでをエージェント 2 も同様とし，この提案手法を RND-Comm として実験をする．

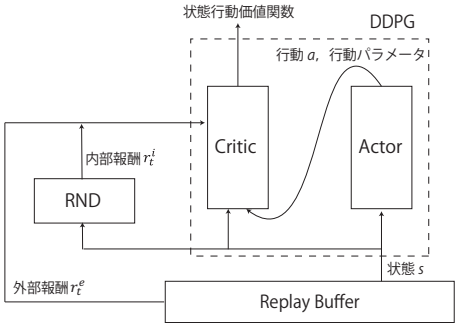

図 1: RND と DDPG を組み合わせたモデル

4 実験

実験環境は HFO を使用する．2 体のエージェントとボールはエピソード毎にランダム配置される．エピソードの終了条件はゴールをするか，ボールをフィールドの外に出すか，ボールに触らず 100 タイムステップ過ぎることである．トレーニング期間は 30 万エピソードとし，学習エピソードの終了後に 1000 タイムステップの評価を行う．提案手法（RNDComm）と DDPG のみの 2 つを学習過程での学習効率と評価エピソードで比較する．外部報酬 r_t^e は式 (7) の報酬関数によって得られる．

$$r_t^e = d_{t-1}(a,b) - d_t(a,b) + I_t^{kick} + 3(d_{t-1}(b,g) - d_t(b,g)) + 5G_t^{goal} \quad (7)$$

式 (7) はステップ毎に報酬を得られる外部報酬である．$d(a,b)$，$d(b,g)$ はそれぞれエージェントとボール，ボールとゴールの距離に比例した関数，I_t^{kick}，G_t^{goal} はそれぞれキック，ゴール時に 1 得られる報酬で G_t^{goal} のみエージェント同士で共有される．

5 結果と考察

学習過程の比較結果を図 2 に示す．学習後の 1000 エピソードの評価結果は RNDComm がゴール率 82.0% であり，DDPG のみがゴール率 0.6% の結果となった．

図 2 より RNDComm は 3 万エピソードからゴールをする行動を見つけ出し，30 万エピソードでのエージェント 1 の平均外部報酬は約 5.9，エージェント 2 は約 4.5 となった．一方，DDPG のみの場合はコミュニケーションによる自身の状態を送ることができないため，味方エージェントの状態を加味して行動を学習していくことができない．そのため，学習の効率が悪く 2 体のエージェントは 25 万エピソードから外部報酬が上昇し始める．

評価エピソードではどちらもトレーニング期間で学習させた方策のみを利用した．その結果，式 (7) の報酬関数をエピソード中に最大化するように学んだ RNDComm が評価エピソードにおいても 82.0% の高いゴール率となっている．よって，本研究での提案手法が同時学習を可能としトレーニング期間中の学習効率が良いことが分かった．

さらに，エージェント 1 とエージェント 2 の最終的な外部報酬値の差は約 1 となった．式 (7) の G_t^{goal} がエージェント同士で共有されていることにより，どちらかがゴールをするとゴールをしていないエージェントは行動せずに比較的高い報酬値を得ることになる．そのため，ゴールに直結しない行動が最適行動と学習することでエピソードの初期地点から動かない様子が見られた．

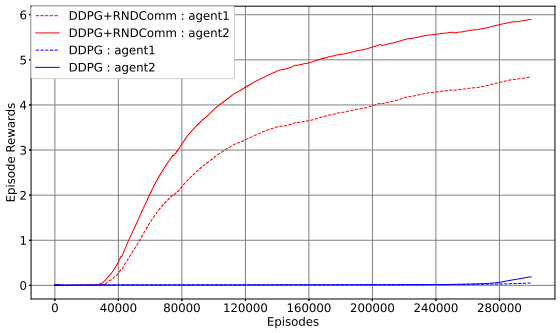

図 2: RNDComm（赤線）と DDPG のみ（青線）の学習過程の比較

6 おわりに

本研究では提案手法が HFO において同時学習を可能とし，既存手法のみに比べて学習効率が大幅に改善することを示した．今後の課題として，ある目的に最適なメッセージを見つけることである．人の手で設計するメッセージよりも目的に合ったメッセージを学習させて見つけることでより高度な戦略を獲得できると考えられる．

参考文献

[1] Burda, Y., Edwards, H., Storkey, A. and Klimov, O.: Exploration by Random Network Distillation, *Proceedings of the International Conference on Learning Representations* (2019).

[2] Hausknecht, M. and Stone, P.: Deep Reinforcement Learning in Parameterized Action Space, *Proceedings of the International Conference on Learning Representations* (2016).