ISSN 2186-5647

2-45

探索時に ES を用いたハイブリッド強化学習

日大生産工（院）　○木内　稜久　　日大生産工　山内　ゆかり

1. まえがき

近年、強化学習ではロボット制御やゲームAI、自動運転などの分野で活用されている。深層強化学習では、Deep Q-Network(DQN)[1]が代表的な手法として知られている。エージェントが未知の領域を「探索」し、そこで得た知識を「活用」することで最適な行動を選択する。探索不足だと、局所解に陥り最適解を見逃す可能性が高くなる。また、活用不足だと報酬が少なく、学習が遅くなる。そのため、探索と活用のバランスを適切にとることが必要である。従来の強化学習では、実験レベル、エピソードレベル、ステップレベルで探索と活用の切り替えが行われている。

Pislarらは”When should agents explore?”[2]では、あまり研究されてこなかったステップレベルとエピソードレベルの中間にあたるintra-episodic（エピソード内探索）を提案した。これは数ステップ継続して探索を行う手法である。また、従来の強化学習では探索と活用のトレードオフをどう設定するかという問題、つまり探索の量に関して研究されている。しかし、探索行動（ランダム、内発的動機付け、その他）をいつすべきかというタイミングに関してはあまり研究されていない。そこで、Pislarらは探索・活用モードの切り替え手法Informed switchingについて提案している。

強化学習(RL)と進化戦略(ES)[3]は相互補完な特性を持つため、RLとESを組み合わせた手法が研究されている。本研究では、RLとESを組み合わせた手法に対して、切り替え基準にInformed switchingを導入した手法を提案する。活用モードではRL、探索モードではESを用いて学習を行う。提案では局所解に陥りやすい問題の緩和による学習の安定性向上を目指す。

2. 従来研究

2.1. Deep Q Network

Deep Q Network(DQN)[1]は深層強化学習における代表的なアルゴリズムで、Q学習と深層ニューラルネットワーク(DNN)を組み合わせた手法である。深層ニューラルネットワークで行動価値関数を近似することにより連続値の状態を扱うことが可能になる。Figure 1にDNNを強化学習に適用させるネットワーク構造を示す。

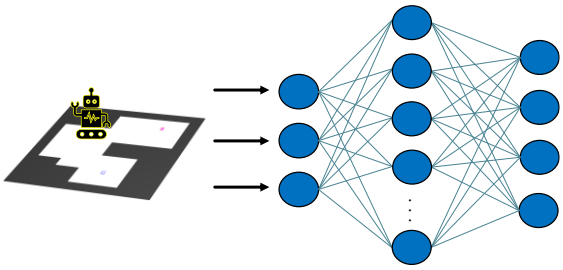

Figure 1　DNN による行動関数の近似

DQNのネットワーク構造は状態s_tを入力として、各行動a_tに対応する行動価値Q値を出力している。DQNでは、損失関数を最小化するように式(1)を教師として学習させる。

$$Q(s_t, a_t) = R_{t+1} + \gamma \max_{a'} Q(s_{t+1}, a') \qquad (1)$$

式(1)中のγは割引率、Rは報酬を示す。

2.2. Experience Replay

Experience Replay （経験再生）[1]は、エージェントの経験における相関を低減させ、学習の安定性を高めるための工夫である。Figure 2に経験再生の流れを示す。

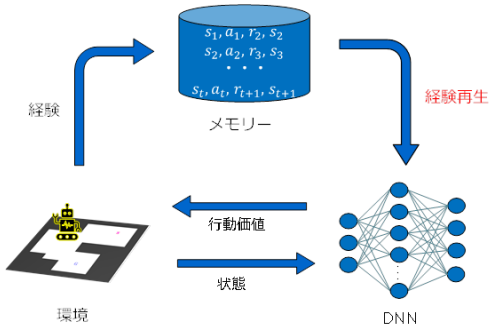

Figure 2　経験再生の流れ

エージェントが得た経験$(s_t, a_t, r_{t+1}, s_{t+1})$をメモリーに保存し、学習時にはランダムに選択された複数の経験を用いて、ミニバッチ学習を行う。

2.3. When should agents explore?

“When should agents explore?”[2]は探索と活用のトレードオフのどのタイミングでモードを切り替えるかという問題に焦点を当てた研究である。探索モードにおける一連のステップの探索期間は実験レベル(内発的動機づけ)、エピソードレベル、ステップレベル$(\varepsilon - greedy)$がある。ステップレベルとエピソードレベルの中

Hybrid Reinforcement Learning with ES during Exploration
Riku KIUCHI and Yukari YAMAUCHI

間にあたるintra-episodic（エピソード内探索）が提案された。さらに重要なのは、エピソード内を探索するタイミングである。モード切り替え手法Informed switchingは、エージェントの現在の情報を用いたトリガー信号に基づいてモード切り替えを行う。予測価値差分(value promise discrepancy)であるトリガー信号を式(2)に表す。

$$D_{promise}(t-k,t) \coloneqq \left| V(s_t-k) - \sum_{i=0}^{k-1} \gamma^i R_{t-i} - \gamma^k V(s_t) \right| \quad (2)$$

これはkステップ前の予測と実際に行動した後の実現値の差分を表している。この値が大きい場合、予測が上手くいっていないと判断することができ、活用モードから探索モードに切り替える。

2.4. 進化戦略

進化戦略(Evolutionary Strategy：ES)[3]はブラックボックス関数の最適化手法の１つである。目的関数の内部構造や勾配情報を必要とせず、目的関数の値さえ得られれば最適化が可能である。目的関数$F(\theta)$が最適となるように方策パラメータθを更新するアルゴリズムについて説明する。本文における目的関数は報酬である。

まず、$N(0,1)$に従う摂動（ノイズベクトル）$\varepsilon_{1\cdots n}$を生成する。摂動を加えた方策パラメータで環境でのエピソードを実行し、報酬F_iを得る。これを式(3)に示す。

$$F_i = F(\theta_t + \sigma\varepsilon_i) \quad (3)$$

式(4)で得た報酬から勾配を推定する。推定された勾配に基づき方策パラメータを更新する。これを式(5)に示す。

$$\nabla_i \leftarrow \frac{1}{n\sigma} \sum_{i=1}^{n} F_i\, \varepsilon_i \quad (4)$$

$$\theta_{t+1} \leftarrow \theta_t + \alpha\nabla_i \quad (5)$$

2.5. Random Network Distillation

Random Network Distillation(RND)[4]は報酬が疎な環境での学習を改善するために設計された探索手法である。RNDはTarget NetworkとPredictor Networkの2つのネットワークを使用し学習する。Target Networkはランダムに初期化され、Predictor NetworkはTarget Networkの出力を予測するように学習する。ネットワークの2乗誤差を内部報酬として扱うことでエージェントの未知の状態への探索が促進される。このとき、強化学習における報酬は式(6)で表され、外部報酬(R_e)と内部報酬(R_i)を足し合わせたものである。

$$R = R_e + R_i \quad (6)$$

3. 提案手法

提案手法では、強化学習(RL)と進化戦略(ES)を組み合わせた手法に切り替え基準としてInformed switching を加えた手法を提案する。従来研究の Informed switching は探索モードで$\varepsilon - greedy$、RND を用いた新規性を追求する内発的動機づけを行っている。ランダム方策による探索は、最適解を導き出すことも乱数次第なため学習の安定性に欠ける。そこで探索モードのときに進化戦略を用いて探索を行う。

4. 実験および検討

本研究では従来手法として、探索モードで$\varepsilon - greedy$、活用モードでDQNを使用する。探索モードでRNDを用いた新規性を追求する内発的動機づけ、ESを使用し、実験を行う。10*10の迷路を実験環境とする。実験では平均行動回数について各手法で比較を行う。

5. まとめ

本研究では強化学習と進化戦略(ES)を組み合わせた手法に切り替え基準としてInformed switchingを加えた手法を提案した。提案手法によって局所解に陥りやすい問題が緩和され、学習が安定化されると考える

参考文献

[1] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, M. Riedmiller, "Playing atari with deep reinforcement learning", arXiv preprint arXiv:1312.5602 (2013).

[2] M. Pîslar, D. Szepesvari, G. Ostrovski, D. Borsa, T. Schaul, “When should agents explore?”, arXiv preprint arXiv:2108.11811 (2021)

[3] T. Salimans, J. Ho, X. Chen, S. Sidor, I. Sutskever, “Evolution Strategies as a Scalable Alternative to Reinforcement Learning”, arXiv preprint arXiv:1703.03864 (2017)

[4] Y. Burda, H. Edwards, A. Storkey, O. Klimov, "Exploration by random network distillation", arXiv preprint arXiv:1810.12894 (2018)