200. 深層強化学習

秋葉洋哉

2024 年 7 月 15 日

1 A3C

1.1 概要

A3C(Asynchronuous Advantage Actor-Critic) とは、強化学習の学習方法の一つで、複数のエージェントが同一の環境で非同期に学習するという特徴を有している。名前の由来は

- Asynchronous：複数エージェントが非同期で並列学習する、という意
- Advantage：複数ステップ先を考慮して更新する、という意
- Actor：方策によって行動を選択する、という意
- Critic：状態価値関数を用いて方策を修正する、という意

Actor-Critic とは、方策 (Actor) を直接改善しながら、方策を評価する価値関数 (Critic) を同時に学習させるアプローチを指している。

1.2 アルゴリズム

A3C のアルゴリズムは、以下の通りである。

1. グローバルネットワークを初期化する
2. エージェントごとにローカルネットワークを初期化する
3. エージェントごとに以下の手順を繰り返す
   (a) グローバルネットワークの重みをローカルネットワーク (Worker n) に pop する
   (b) 環境 (E_n) から状態 (s_n^t) を取得する
   (c) 状態 (s_n^t) を入力として、ローカルネットワークから方策 (π_n) と価値 (V_n) を取得する
   (d) 方策 (π_n) に基づいて行動を選択し、環境 (E_n) に適用する
   (e) 環境から報酬と次の状態 ($s_n^{(t+1)}$) を取得する
   (f) 状態 (s_n^t) と次の状態 ($s_n^{(t+1)}$) を入力として、ローカルネットワークから方策と価値を取得する
   (g) 方策と価値を用いて、方策勾配と価値勾配を計算する
   (h) 方策勾配と価値勾配を用いて、ローカルネットワークの重みを更新する
   (i) ローカルネットワークの重みをグローバルネットワークに push する

一般的な Actor-Critic ネットワークでは、方策ネットワークと価値ネットワークを別々に定義し、別々の

損失関数 (方策勾配ロス/価値ロス) でネットワークを更新していた。しかし、A3C では、パラメータ共有型の Actor-Critic であり、1 つの分岐型のネットワークが、方策と価値の両方を出力し、一つの「トータルロス関数」を用いてネットワークを更新する。これにより、方策勾配と価値勾配を同時に学習することができる。トータルロス関数は、以下で表せる。

$$\text{Total Loss} = -\text{アドバンテージ方策勾配} + \alpha \cdot \text{価値関数ロス} - \beta \cdot \text{方策エントロピー} \tag{1}$$

ここで、アドバンテージ方策勾配は、方策勾配にアドバンテージを掛けたものであり、アドバンテージは、報酬と価値の差分である。また、方策エントロピーは、方策の多様性を保つための項である。

1.3 A3C のメリット

A3C のメリットは、以下の通りである。

- 複数エージェントが同時に学習するため、学習速度が速い
- 方策勾配と価値勾配を同時に学習するため、学習が安定する

2. については、強化学習長年の課題であった、経験の自己相関が引き起こす学習の不安定化を解消することができる、というメリットになる。この課題に対しては、かつては、DQN(Deep Q-Network の略、Q 学習を用いた強化学習手法) が、Experience Replay(経験再生) という手法を用いることで解決していた。Experience Replay は、バッファに蓄積した経験をランダムに取り出すことで、経験の自己相関を低減する手法である。しかし、経験再生は基本的にはオフポリシー手法 (方策とは異なる方策で学習する手法) であるため、オンポリシー手法 (方策と同じ方策で学習する手法) である A3C とは相性が悪い。そのため、A3C は、経験再生を用いずに、サンプルを集めるエージェントを並列化することで、経験の自己相関を低減する手法を採用している。

1.4 A3C の課題

A3C では、Python 言語の特性上、非同期並列処理を行うのが面倒であったり、大規模なリソースが必要になるという課題がある。そのため、A3C の改良版として、A2C という手法が発表された。A2C は、A3C の非同期処理を同期処理に変更することで、性能を大きく変えずに学習の安定化と計算リソースの削減を実現している。

■参考文献

1. 岡谷貴之/深層学習 改訂第 2 版 [機械学習プロフェッショナルシリーズ]/ 講談社サイエンティフィク/ 2022-01-17
2. https://arxiv.org/abs/1602.01783