2020年度 卒業論文

強化学習アルゴリズムPPOとSACの比較と考察

龍谷大学 理工学部 数理情報学科
T160023 川上 悠斗
指導教員 佐野 彰

概要

近年、機械学習の発展によって計算機でも人間のように画像認識したり、自然言語処理を行うことが出来るようになってきた。

2015年にAlpha Goという囲碁の人工知能が話題になった。その頃は機械学習についてほとんど知識がなく深く調べることはなかったが、授業を通して機械学習に興味を持つようになった。機械学習の中でも特に興味を持ったのが強化学習である。機械学習は教師あり学習、教師なし学習、強化学習の3つに分類される。強化学習はエージェントと呼ばれる学習主体が環境上で行動することによって報酬を受け取り、その報酬が最大になるようにその環境下での行動を学習する方法である。

強化学習にはいくつかの手法があるが、本論ではまず最も代表的なアルゴリズムであるQ学習について説明する。

さらに2017年に提案されたPPOと2018年に提案されたSACという2つの強化学習アルゴリズムについて述べる。強化学習にはon-policyなアルゴリズムとoff-policyなアルゴリズムの二種類があり、学習的にエージェントの行動ルールである方策を用いるか否かの違いがある。PPOはon-policyでSACはoff-policyな学習アルゴリズムである。この二つを比較することによって強化学習において、学習時の方策の利用が学習にどのように影響するのかを調べることが本研究の目標である。

本研究では強化学習環境を実装するために、統合開発プラットフォームであるUnityと強化学習用のオープンソースプラグインであるML-Agentsを用いた。はじめに玉転がしというUnityのサンプルプログラムを用いて、Windows環境でのUnityの動作を確認した。また、ML-Agentsのパッケージにあるサンプルプログラムを用いて、ML-Agentsの動作も確認した。

本研究では、PPOとSACでの学習を比較するためにUnityの3DBallとRollerBallという2つの強化学習環境上にML-Agentsを用いてそれぞれのアルゴリズムを実装し、学習過程を調べることによって、on-policyとoff-policyの違いからなぜその差が生まれたのかを考察した。

ML-Agentsを用いた実験では学習過程をグラフにした。3DBallの学習ではSACでの学習のほうが学習が早かったが、RollerBallの学習ではPPOとSACでの学習過程にほとんど差が見られなかった。このことから、報酬を得る条件や座標が変化しない環境ではoff-policyであるSACのほうが適しているのではないかと考えた。しかし、PPOのメリットが分からなかった。おそらくon-policyの特徴からNPCを相手とするような報酬を得る条件の変化が多い環境ではPPOのほうがより学習が安定するのではないかと考えた。

本研究ではPPOとSACを比較して考察することでon-policyとoff-policyの違いを明らかにしようとした。結果、変化の少ない環境であるほど、off-policyの学習が安定することを考察できた。今後も今回用いた2つの環境以外でもPPOとSACでの学習を比較することで、もっと明確な違いが明らかにしていこうと考えている。