RNN を用いた POMDP 環境での倒立振り子問題

3 年 244 番　学籍番号 1852130904　清水翔仁　（指導教員：山根智）

1. 背景と目的

強化学習は，試行錯誤から学習する点で様々な問題に適用しやすいため，近年盛んに研究されている．DRQN は，DQN に循環神経網の一種である LSTM を組み合わせることで過去の情報を扱う．さらに LSTM はより長期的な情報を考慮することで，現実で起こりうる不完全な情報にも対応できるため，POMDP の環境でより効果を発揮することができる．

しかし，DRQN は学習の時に LSTM の初期状態をゼロベクトルで初期化するため，学習時に用いた情報のタイムステップより長い期間に対して学習することが難しいという欠点がある．そこで本研究では，LSTM の初期状態を与える方法を複数用意し，それぞれの実行結果についての考察を行う．

2. 原理と手法

ここでは，研究目的を達成するために使用した手法について述べる．

Bootstrapped Random Updates

エピソードからランダムに軌跡を取り出し，LSTM の初期状態を zero 入力して学習を行う．これを Bootstrapped Random Updates という．シンプルかつランダムにサンプリングできるメリットはあるが，初期状態を zero 入力するため LSTM が適切な表現を獲得できない可能性がある．

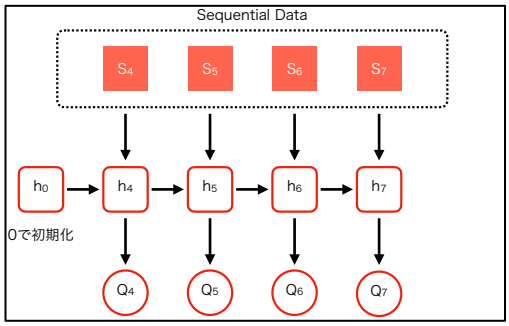

図 1：Bootstrapped Random Updates の模式図

Bootstrapped Sequential Updates

1 エピソード全部のデータを使って学習を行う．これを Bootstrapped Sequential Updates という．LSTM 初期化で困ることはないが，バリアンスが大きい，エピソード長が可変などの問題がある．

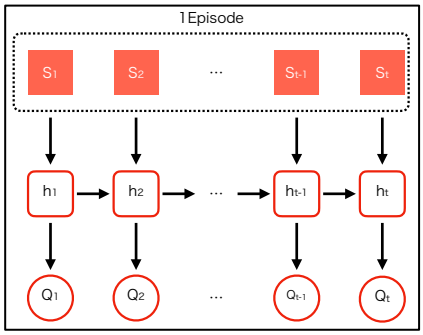

図 2：Bootstrapped Sequential Updates の模式図

Burn-in

Burn-in とは，最初は学習を行わずデータだけを流し，今のネットワークの重みに慣れさせるという手法である．これにより正確な hidden state を復元することができる．

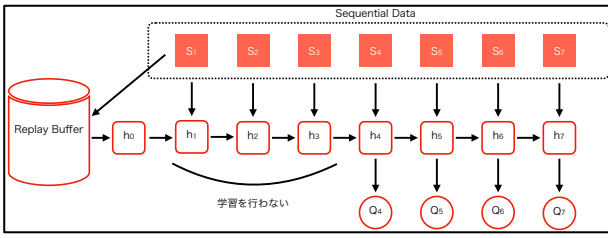

図 3：Burn-in の模式図

3. 実験

ここでは，実験内容と結果，およびそれに基づく考察を述べる．

今回は，通常の DRQN と，最も結果の良かった R2-D2 というプログラムの比較のみを示す．R2-D2 では Bootstrapped Random Updates と Replay Buffer，Burn-in を組み合わせた手法を採用している．

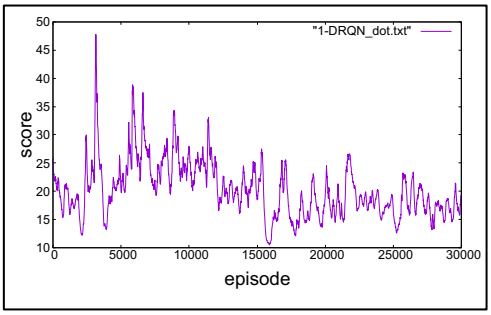

図 4：DRQN

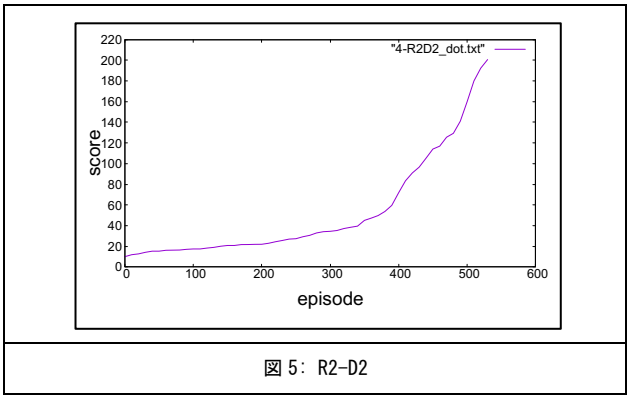

図 5：R2-D2

図 4 より，通常の DRQN では score が伸びず収束しないまま終了していることがわかる．図 5 より，R2-D2 では episode 数に比例して score が伸び続け，episode 数が 550 程度で収束していることがわかる．このことから R2-D2 では，LSTM の初期状態問題が改善されていることがわかる．

4. まとめ

本研究の目標である，LSTM の初期状態問題の検証を行うことができた．また今回使用したプログラムでは，結果が収束したため初期状態問題は改善されていると考えられる．