6M-07

ニューラルネットワークを用いた強化学習における環境状態の分布を考慮した擬似リハーサルの導入

辺見 航平† 服部 元信‡
山梨大学 大学院医工農学総合教育部† 山梨大学 大学院総合研究部‡

1. はじめに

近年の強化学習研究では,ニューラルネットワーク(NN)を用いた手法が主流となっている.中でも深層 NN を用いた Deep Q Network[1]の登場により,強化学習の性能が大幅に向上している.しかし,強化学習はその試行錯誤的な学習方法から,最適な方策を獲得するまでに多くの学習回数を要する.それに加え NN 特有の問題が存在し,学習効率の悪化が起こっていると考えられている.この問題を解決するために Marochko らは擬似リハーサル[2]と呼ばれる手法を強化学習に導入し,その有効性を示した[3].本研究では学習効率の更なる向上を期待し,擬似リハーサルにおける擬似入力生成方法の改良を行い,計算機実験によってその有効性を検証した.

2. ニューラルネットワークにおける破局的忘却とその抑制方法

ある学習済み NN が存在したとき,その NN に新たな情報を追加するという状況において,それまでに学習した情報を保持し続けることが必要となる.追加学習データのみを用いて NN の典型的な学習方法である誤差逆伝播学習法を行うと,学習済み NN のニューロン間重みが追加学習データに対する出力と教師信号との誤差を最小化するように更新される.これによりニューロン間重みは上書き更新され,それまでに学習した情報が破壊されてしまう可能性が非常に高くなる.このような情報の忘却現象を破局的忘却と呼ぶ.

この問題点の抑制方法として擬似リハーサルが提案された[2].この方法では新しい学習データを追加学習させる際,それまでに学習した情報を保持するための学習パターンとして擬似パターンを生成する.擬似パターンとは擬似入力ベクトルと擬似出力ベクトルの組のことを指している.擬似入力ベクトルは NN の入力層のそれぞれの次元に対してランダムな値を与えることで生成される.そして,擬似入力ベクトルを入力として追加学習前の NN へフォワード処理を行い,その出力値を擬似出力ベクトルとして保持する.また,生成した複数の擬似パターンをバッファに格納する.このバッファから1 epoch 毎擬似パターンを選択し,追加学習パターンと合わせて誤差逆伝播法によって学習することにより,それまでに学習した情報を NN に極力保持しながら追加学習を行うことができる.

3. 強化学習への擬似リハーサルの導入

Introduction of pseudorehearsal considering the distribution of environmental state for the reinforcement learning using neural network
†Kohei Henmi, University of Yamanashi
‡Motonobu Hattori, University of Yamanashi

本章ではまず従来手法の説明を行ったのちに本研究の研究目的及び本研究の提案手法についての説明を行う.

3.1. 従来手法

評価値の更新を1 step 毎に行う NN を用いた強化学習では,以前学習した評価値への影響を及ぼす可能性が非常に高い.これは一般的な NN における破局的忘却と非常に類似した現象であると考えられる.NN を用いた強化学習の学習効率の向上を目的とし,Marochko らは NN を用いた強化学習へ擬似リハーサルを導入した[3].この手法では,入力行列を定義し,1 列目には実際に観測した特徴ベクトルを,残りの列には生成した擬似入力群を格納する.また教師信号行列を定義し,1 列目には観測した特徴ベクトルに対応する教師信号ベクトルを,残りの列には擬似出力群を格納する.そして,NN の入力群を特徴行列,その教師信号群を目標行列とし,誤差逆伝播学習法によるバッチ学習を行う.この手法を用いることで強化学習における破局的忘却の抑制及び学習の収束の高速化が確認された.

3.2. 研究目的と提案手法

ここで,従来手法の擬似入力生成方法に着目すると,擬似入力生成確率分布に関する議論がなされていなかった.そのため,環境状態の発生確率を基に偏りを持った確率分布を設定し,擬似入力を生成する手法を提案し,これによる学習効率の改善を本研究の目的とした.

我々は以前学習した情報を擬似パターンによって有効に保持するために,学習時における環境状態の発生確率に着目した.入出力関数における発生確率の高い環境状態に対応する箇所は,ある程度学習が進んでおり,その箇所の情報の価値は他の箇所と比べて高いものであると考えた.そのため入出力関数において発生確率の高い環境状態に対応する箇所をより優先的に保持することが重要であると考えた.このことから擬似入力の生成に関して生成確率に偏りを持たせることによってこれが実現できると考え,擬似入力の生成方法として2つの確率分布を設定した.

1つ目は環境状態の各要素の発生確率が平均値に対して左右対称な釣鐘型であると仮定したガウス分布である.ある環境状態ベクトルの要素dにおけるガウス分布の確率密度関数を式(1)に示す.

$$G(x;\mu_d,\sigma_d^2)=\frac{1}{\sqrt{2\pi\sigma_d^2}}\exp\left(-\frac{(x-\mu_d)^2}{\sigma_d^2}\right) \quad (1)$$

ここでμ_dとは環境状態ベクトルの要素dの学習開始時からの平均値であり,σ_d^2 はその分散を示している.

2つ目は切断ガウス分布である.ガウス分布を用いた場合,ある時刻までの環境状態のあるひとつの要素に着目すると,観測値の最大値より大きい値若しくは最小値より小さい値を取る確率が存在する.NN は最小値-最大値間以外の範囲は学習範囲外であり,その範囲の入出力関数を保持することは無駄であると考えた.そのため最小値-最大値間という有限な範囲を取るガウス分布である切断ガウス分布を設定した.切断ガウス分布の確率密度関数は式(2)を用いて算出される.

$$TGauss(x;\mu_d,\sigma_d^2) = \frac{G(x;\mu_d,\sigma_d^2)}{\int_{min_d}^{max_d} G(x;\mu_d,\sigma_d^2)dx} \quad (2)$$

ここでmin_d,max_dは環境状態ベクトルの要素dのこれまでに観測した最小値と最大値を示している.設定した2つの確率分布は環境状態ベクトルの要素数分設定し,それぞれが独立している.

4. 計算機実験

強化学習への擬似リハーサルの導入及び擬似入力生成方法の提案の有効性を検証するために計算機実験を行った.

4.1. タスク環境と実験条件

実験タスクは Cart-Pole タスクを設定した.このタスクでは左右に移動可能な Cart の上部に Pole の片方が一点のみで設置されており,エージェントが Cart を左右に移動させることで Pole が倒れないようにバランスを保つタスクである.環境状態の要素は,Pole の角度と角速度,Cart の位置と速度の 4 つである.また,エージェントが選択できる行動は「Cart の左から 10N の力を加える」,「Cart の右から 10N の力を加える」の 2 つである.episode 終了条件は,「200step 経過」「Cart が±2.4m の範囲外に出たとき」「Pole が鉛直上向きから±12°の範囲を出たとき」のいずれかである.また,報酬は Pole が鉛直上向きから±12°の範囲内に存在するときであり,1step 毎 1.0 の報酬を得る.

テストは学習 1episode 毎に 100episode 行い,学習完了したとみなす条件は,テスト時の平均獲得報酬が 195.0 以上となったときであり,それまでの学習 episode 数を比較した.ある一定の方策を獲得するまでに要した学習 episode を比較することによって,各学習手法の学習効率を確認することができる.episode の上限を 1000episode と設定し,それまでに学習ができなかった場合,学習失敗とみなした.また学習失敗の場合は平均学習 episode 数の算出に用いなかった.また発生確率の高い環境状態に対応する箇所を優先的に保持しない方法と比較するために,これまでに観測した最小値min_dと最大値max_d間の値から一様分布を用いて擬似入力を生成する方法も併せて実験を行った.擬似リハーサルに用いるパラメータである擬似パターン数は pr={1,2,4,6,8,10,12,20,50,80}の10種類を設定,バッファサイズは 100 とした.各学習手法と擬似リハーサル方法と擬似パターン数の組み合わせをそれぞれ 20simulation ずつ行った.

4.2. 計算機実験結果

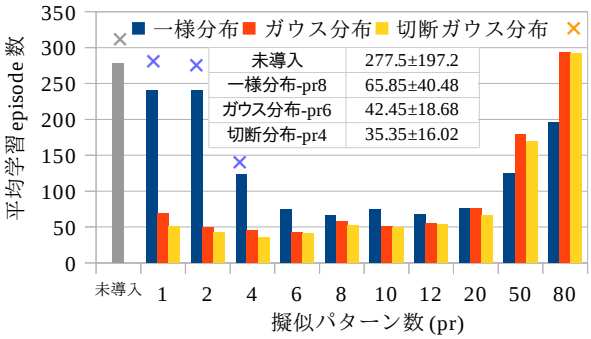

図 1: 学習完了までの episode 数と各分布の最小値と標準偏差

Cart-Pole タスクにおける実験結果を図1に示す.横軸は一番左は未導入,つまり擬似リハーサルを導入していない場合の結果を示し,それ以降は擬似リハーサルを導入した場合のパラメータである擬似パターン数を変化させたときの結果を示す.縦軸は学習が完了するまでに費やした平均学習 episode 数である.また,結果のプロット付近の×マークはそのパラメータにおいて学習中に NN の出力の発散が見られた,または学習失敗が 20simulation 中 1 つでも存在した場合を示している.

各分布の平均学習 episode 数の最小値を取った 4 つ(図 1 内の表)の有意差を t 検定によって検定したところ,未導入と他 3 分布間,一様分布-pr8 と切断ガウス分布-pr4 間に有意水準 1%で有意差が認められた,また,一様分布-pr8 とガウス分布-pr6 間に有意水準 5%で有意差が認められた.この結果から,偏りを持った分布を用いることで入出力関数を擬似パターンによってうまく保持することができ,破局的忘却をより抑えることに成功していると考えられる.

さらに,偏りを持った分布では,未導入や一様分布と比較して標準偏差を大幅に抑えることができている.このことから学習 episode 数にばらつきが少なく,安定して学習ができていることが分かる.この結果からも破局的忘却を抑えて強化学習に成功していると考えられる.

ガウス分布と切断ガウス分布に大きな差が見られないことに関しては,最大値-最小値外を取る確率が結果に影響が及ぼさないほど小さかったためと考えられる.

5. まとめ

本研究では,環境状態の発生確率を基に偏りを持った確率分布を設定し,擬似入力を生成する手法を提案した.そして計算機実験の結果から,Cart-Pole タスクにおいて NN を用いた強化学習における破局的忘却を抑え,学習効率を改善することが確認された.

参考文献

[1] Mnih, V., Kavukcuoglu, K., Silver, D., et al. “Human-level control through deep reinforcement learning,” Nature 518(7540), pp.529-533 (2015).
[2] Robins, A., “Catastrophic forgetting, rehearsal and pseudorehearsal,” Connection Sci. 7(2), pp.123–146 (1995).
[3] Marochko, V., Johard, L., Mazzara, M., “Pseudorehearsal in Value function approximation,” KES International Symposium on Agent and Multi-Agent Systems: Technologies and Applications, vol74, pp.178-189 (2017).