4E-01

優先度付き経験再生を用いた深層強化学習によるカリキュラム設計法

堀田 颯平 †　　飯間 等 ‡
京都工芸繊維大学 †　京都工芸繊維大学 ‡

1. はじめに

学習者が難しい課題を解決できるようにするために，最初は簡単な課題を与えておき，徐々に難しい課題を与えて学習させる方法を取ることが多い．この時，どのような課題を順番に与えるか，すなわちカリキュラムをどのように設計するかが問題となる．一般的にこのカリキュラムの設計は人手で行うことが多いと思われるが，適切に設計することは難しい．従って，カリキュラム設計が自動的に行えるようになれば有用である．そこで，著者らは強化学習器を学習者とみなしてカリキュラムを自動的に設計する深層強化学習法を以前に提案した [1]．一方，深層強化学習をより効率的に学習させる改良として優先度付き経験再生 [2] が提案されている．そこで，本論文ではこの優先度付き経験再生を用いた深層強化学習によってカリキュラムを自動的に設計する手法を提案する．

2. カリキュラム学習

提案する強化学習法はカリキュラム学習 [3] の一種となっている．カリキュラム学習は強化学習タスクが難しくてそのままでは学習できない場合に用いられる方法である．この学習法では高難度の一つのタスクで学習させたり，様々なタスクで汎用的に学習させるために難易度や条件などの異なるタスクを複数個用意し，より簡単なタスクから学習を開始させ，徐々に難しいタスクで学習させていく．これらのタスク系列がカリキュラムとなっており，あとで行うタスクの学習は直前で行ったタスクの学習結果を引き継いで行われる．先に述べたように通常のカリキュラム学習は高難度のタスクや様々なタスクで学習させることを目的としている．これらに対して本研究はカリキュラムを設計することを目的としている．

3. 優先度付き経験再生

深層強化学習では，ネットワークの重みやバイアスをより良く最適化するために経験再生が用いられる．これはセルフプレイで経験した状態，行動，報酬，次状態を経験リストに保存しておき，そこからランダムにサンプリングした経験を学習に用いるものである．ところが，ランダムにサンプリングするために，学習の必要がない経験をサンプリングしたり，逆に学習に必要な経験をサンプリングしないことが生じて，学習が非効率となっている．これを解決する方法が優先度付き経験再生である．優先度付き経験再生では経験リストの i 番目の経験に関して，その優先度 pr_i に基づいたサンプリング確率 $P(i)$ を以下の式で計算し，この確率に従って経験をサンプリングする．

$$P(i) = \frac{pr_i^\alpha}{\sum_k pr_k^\alpha}$$

ここで，α はどの程度優先度を考慮するかを決める定数である．pr_i の計算は，TD 誤差 δ_i を用いて，$pr_i = |\delta_i| + \epsilon$ (ϵは微小な定数) とする方法と，$pr_i = \frac{1}{rank(i)}$ とする方法の 2 種類がある．ここで，$rank(i)$ は $|\delta_i|$ を降順に並べたときの順位である．

4. 提案手法

ここでは，優先度付き経験再生を導入した深層強化学習に基づくカリキュラム学習により，自動的にカリキュラムを設計する方法を提案する．本研究では，学習者も強化学習器とし，以降ではこの学習を下位の強化学習と呼ぶ．また，下位の強化学習で最終的に学習させたい目的タスクに，その難易度を変更して様々なタスクを作成できるパラメータ p があると仮定する．例えば，障害物を避けて平面内を移動するロボットの経路を学習させるタスクでは，平面サイズや障害物の個数などをパラメータ p とできる．

提案する深層強化学習法の行動，状態，報酬を説明する．時刻 $t(=1,2,\dots)$ における行動は，パラメータ p の値 p_t を決定してカリキュラムにおける t 個目のタスク T_t を作成し，T_t に下位の強化学習を適用させることとする．また，T_t に下位の強化学習を適用させた時の学習成功率を w_t とする時，時刻 t の状態は $(p_{t-2}, w_{t-2}, p_{t-1}, w_{t-1})$ とする．ここで学習成功率について，下位の強化学習を適用させたときに，あらかじめ定めたエピソード間隔 E に対して，$(n-1)E$ エピソードと nE エピソードの間 $(n=1,2,\dots)$ でタスク T_t の目的を達成できた回

Curriculum design method using deep reinforcement learning with prioritized experience replay
† Sohei Hotta, Kyoto Institute of Technology
‡ Hitoshi Iima, Kyoto Institute of Technology

数 $success\_count_n$ を求め，$r_n = success\_count_n/E$ とし，その最大値を $w_t = \max_n r_n$ とする．また，最低限達成すべき目的達成率を w_{lower_limit} とすると，$w_t \geq w_{lower\_limit}$ の時，下位の強化学習で学習に成功したと判定することとする．

報酬に関しては，目的タスクに対する下位の強化学習が学習に成功した場合に $+1$ の報酬を与える．目的タスク以外のタスクに対する学習に成功したときには -1 の報酬を与える．いずれのタスクにおいても学習に失敗した場合は -20 の報酬を与える．目的タスクでの学習に成功したとき，あるいはいずれかのタスクでの学習に失敗したときの次状態は終端状態とし，そこで 1 つのエピソードを終了させ初期状態に戻る．

以上に述べた行動，状態，報酬を用い，深層強化学習には Deep Q-network [5] を使用する．ネットワークの中間層は 1 つとし，そのニューロン数は 10 個とする．また，前節で述べた優先度付き経験再生を用いる．各タスクに対して適用する下位の強化学習では状態などを適切に設定して学習パラメータを学習させる．この強化学習で 2 つ目以降のタスク $T_t(t \geq 2)$ で学習するときの初期パラメータの値は，直前のタスク T_{t-1} で学習したパラメータの値を引き継ぐ．

Deep Q-network による学習を完了させた後，各時刻 t で最大 Q 値をもつ行動を常に選択してタスク T_t を決定することを繰り返してカリキュラムを作成する．しかしこのとき，カリキュラムを作成しようとする度に，下位の強化学習の成功率 w_t が変化して次状態が変化し，従って作成されるカリキュラムも変化する可能性がある．また，下位の強化学習が失敗する可能性もある．そこで，カリキュラム作成手続きを 100 回行って得られたカリキュラムの中で，最も多く作成できたカリキュラムを，設計されたカリキュラムと定める．

5. 数値実験

提案手法を Keylock Markov Decision Process [4] というタスクに適用する実験を行い，その性能を確認する．このタスクは 10 個の状態 $s_1, s_2, \ldots, s_{10}$ と最終状態 s_T を有し，初期状態 s_{k_0} ($k_0 \in \{1, 2, \ldots, 10\}$) から s_T に到達することを目的とするタスクである．各状態 s_k では 2 つの行動 a, b を取る事ができ，行動 a を取ると状態 s_{k-1} に移り，行動 b を取ると状態 s_{10} に移る．また，s_1 において行動 a を取ることで最終状態 s_T へ到達する．従って，最適行動は行動 a を取り続けることであり，k_0 が大きいほどタスクが難しくなる．そこで，値を変更することでタスクの難易度を変更できるパラメータは $p = k_0$ とし，目的タスクは $p = 10$ のタスクとする．

下位の強化学習は Q 学習とした．優先度付き経験再生に用いる優先度は $pr_i = |\delta_i| + \epsilon$ の方法で計算し，$\epsilon = 0.00001$，$\alpha = 1$ とした．学習成功率 w_t に関して，$E = 100$，$w_{lower\_limit} = 0.8$ とした．提案手法を 100 回実行して設計されたカリキュラムのタスク数と，そのカリキュラムで目的タスクの学習に成功した割合 (成功率と呼ぶ) を，優先度付き経験再生を用いた場合と用いなかった場合で比較して提案手法を評価する．

提案手法で設計した 100 個のカリキュラムを，タスク数と成功率で分類した場合のカリキュラム数を表 1 に，優先度付き経験再生を用いなかった場合のものを表 2 に示す．ただし，これらの表には，タスク数が 4 以上と多い場合，及びカリキュラムの設計に失敗した場合のカリキュラム数は表示していない．表 1, 2 より，成功率が 0.95 以上であるカリキュラムの数は，提案手法を用いた場合は 71 個であり，優先度付き経験再生を用いなかった場合は 49 個であるので，優先度付き経験再生を用いることで，目的タスクで学習できるカリキュラムをより安定的に設計できている．

表 1　タスク数と成功率で分類したカリキュラム数 (提案手法)

成功率＼タスク数	2	3
0〜0.9	18	5
0.9〜0.95	0	1
0.95〜1	33	39

表 2　タスク数と成功率で分類したカリキュラム数 (優先度付き経験再生なし)

成功率＼タスク数	2	3
0〜0.9	26	15
0.9〜0.95	3	1
0.95〜1	35	15

参考文献

[1] 堀田颯平，飯間等：令和 3 年電気学会全国大会講演論文集，pp.129–130 (2021).

[2] Schaul,T., Quan,J., Antonoglou,I., and Silver,D.: Prioritized experience replay, *Proceedings of International Conference on Learning Representations* (2016).

[3] Bengio,Y., Louradour,J., Collobert,R., and Weston,J.: Curriculum learning, *Proceedings of the 26th Annual International Conference on Machine Learning*, pp.41–48 (2009).

[4] Matiisen,T., Oliver,A., Cohen,T., and Schulman,J.: Teacher–student curriculum learning, *IEEE Transactions on Neural Networks and Learning Systems*, Vol.31, No.9, pp.3732–3740 (2019).

[5] Mnih,V., Kavukcuoglu,K., et al.: Human–level control through deep reinforcementlearning, *Nature*, Vol.518, pp.529–533 (2015).