光リザーバコンピューティングによる強化学習の実装

Reinforcement learing based on photonic reservoir computing

埼玉大 [1]　○菅野 円隆 [1]

Saitama Univ.[1], ○Kazutaka Kanno[1]

E-mail: kkanno@mail.saitama-u.ac.jp

1. 背景: 強化学習は機械学習の一つであり，試行錯誤を繰り返しながら優れた行動方策を学習するフレームワークである．強化学習に関する近年の発展に大きく寄与した点の一つとして，深層学習を利用したことが挙げられる [1]．しかしながら深層学習は学習コストが大きく，学習速度やエネルギー効率の面で課題がある．これに対してリザーバコンピューティングを用いた強化学習が提案されている [2]．リザーバコンピューティングは深層学習に比べ学習が容易であり，学習速度やエネルギー効率の課題を解決できる可能性がある．

さらに近年，機械学習の物理実装に関する研究にも注目が集まっている [3]．その一つとして光リザーバコンピューティングが挙げられる．これはレーザーと時間遅延フィードバックループを用いて疑似的にネットワークを実装する手法であり [4]，光実装により非常に高速な情報処理が実現可能である．この光リザーバコンピューティングにより強化学習を実装することで，高速かつ低消費電力の強化学習システムが可能であると考えられる．そこで本研究では，光リザーバコンピューティングによる強化学習が達成できることを数値計算により示す．

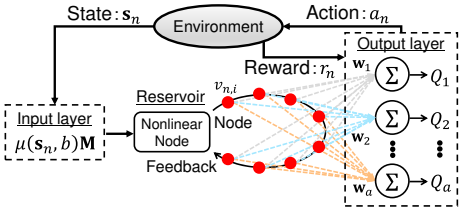

図 1: Schematic diagram of reinforcement learning based on delay-based reservoir computing.

2. 方法: 光リザーバコンピューティングを用いた強化学習の概念図を図 1 に示す．強化学習タスクの環境の状態を入力としてリザーバに入力し，リザーバの出力からノード状態を取得する．ノード状態値の重み付き線形和をリザーバコンピューティングの出力とし，この出力結果に基づいて行動を選択する．選択した行動の結果として得られる報酬に基づいて，リザーバコンピューティングの出力の重みを学習する．

光リザーバとして電気光遅延システムを使用する [4]．時間 n における環境の状態 $\mathbf{s}_n$ に対して前処理を行い，システムに入力する．システムの出力を時間方向に微小間隔で分割してノード状態を定義する．入力 $\mathbf{s}_n$ に対するノード状態を $\mathbf{v}_n$ として表す．ノード状態の重み付き線形和を出力として計算し，重みは学習により更新する．重みの学習方法として Q 学習を使用する．強化学習タスクの行動 a_n の数だけ重み $\mathbf{w}_{a_n}$ を用意し，以下を用いて重みを更新する [2].

$$\mathbf{w}_{a_n} \leftarrow \mathbf{w}_{a_n} + \alpha \left[r_{n+1} + \gamma \max_a (\mathbf{w}_{a_n} \mathbf{v}_{n+1}^T) - \mathbf{w}_{a_n} \mathbf{v}_n^T \right] \mathbf{v}_n \tag{1}$$

ここで α は更新率であり，γ は割引率を表す．

3. CartPole-v0 タスクの結果: 性能評価のために，OpenAI Gym が提供する CartPole-v0 を用いた [5]．これはカート上の棒が倒れないように，カートを左右に加速させる行動を学習するタスクである (図 2(a))．タスクの 200 ステップを 1 エピソードとし，棒が倒れないエピソードが 100 回連続して続けば学習が成功したとみなす．また報酬として各ステップで +1 が得られる．

図 2(b) に CartPole タスクの結果を示す．図の横軸はエピソードを表し，縦軸は各エピソードで得られた総報酬を示している．総報酬が 200 であれば，1 エピソードの間，棒が倒れなかったことを表す．エピソードが少ない場合，棒が途中で倒れるが，エピソードが進むにつれて棒が倒れなくなることが分かる．以上の結果から学習が成功したと言える．タスクの乱数を変えて繰り返し本タスクを実行したが，同様に学習が成功したことが確認できた．したがって光リザーバコンピューティングを用いた強化学習が実現可能であると考えられる．

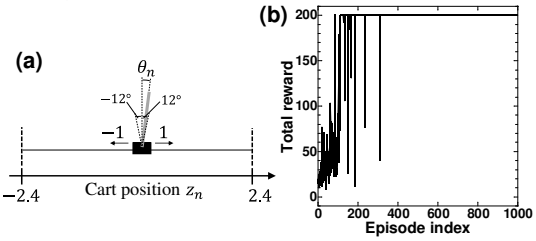

図 2: Numerical result of the CartPole-v0 task.

4. まとめ: 本研究では，数値シミュレーションにより光リザーバコンピューティングを用いた強化学習を実装した．本システムを用いて CartPole-v0 タスクを行った結果，本タスクに成功したため，正しく学習を行えていると考えられる．発表では他のタスクの結果や実験結果についても示す予定である．

謝辞: 本研究の一部は JST CREST(JPMJCR17N2), JSPS 科研費 (JP19H00868, JP20K15185), 電気通信普及財団の支援を受けたものである．

参考文献

[1] V. Mnih et al., Nature, vol. 518, pp. 529–533, 2015.
[2] I. Szita, et al., ICANN 2006, Berlin, Heidelberg, vol. 4131, pp. 830–839, 2006.
[3] G. Tanaka et al., Neural Networks, vol. 115, pp. 100–123, 2019.
[4] Y. Paquot et al., Sci. Rep., vol. 2, p. 287, 2012.
[5] G. Brockman et al., arXiv:1606.01540, 2016.