人工知能学会研究会資料
SIG-AGI-006-09

好奇心で動機付けされた強化学習の実験

Experiments for Reinforcement Learning with Curiosity-driven Exploration

疋田 聡[1]

Satoshi Hikida[1]

[1]株式会社リコー
[1]Ricoh Company, LTD.

Abstract: In recent years, reinforcement learning motivated by curiosity has attracted attention. By learning features related to agents' behavior, they can focus on changes in things that are interesting from images, without being distracted by screen noise or meaningless changes. In addition, it has been reported that even with the learning result based only on the internal compensation, it has the generalization ability to be effective even at the stage other than the stage used for learning. However, in the reference paper, it was only applied to 2 stages in the "Super Mario Bros." Therefore, to further investigate the generalization ability of reinforcement learning motivated by curiosity, I experimented on more stages. As a result, I confirmed the generalization ability of reinforcement learning motivated by curiosity.

1. 背景

エージェントが自ら行動を決定し学習を進めていく強化学習は、汎用人工知能を実現するための有力な方法として盛んに研究されている。しかし、DQN[4][5]のように確率 ε でランダムに行動することにより新たな探索を行う方法では、報酬までの距離が遠い場合に報酬が得られない期間が長くなるため勾配が 0 に近くなってしまい、うまく学習が進まないという問題がある。

このような問題に対応する一つの方法として、外部報酬がなかなか得られない場合にも、好奇心のような内部報酬によって動機付けをした強化学習を行うことが提案されている。例えば、[8]では、環境モデルによる推定と実際の観測の差を内部報酬として誤差が少なくなるように探索している。[9]では、オートエンコーダで学習して今までの状態遷移モデルからの推定と異なる状態に移動することに内部報酬を与えている。[6]では、動画予測を用いて長期予測を作成して最近訪れていない状態へ移動することに内部報酬を与えている。[1]では、密度モデルから算出した状態訪問回数の推定値を用いて内部報酬を与えている。[10]では、状態をハッシュコード化して訪問回数をカウントし、内部報酬を算出している。

さらに、[7]では、自分の行動に関係した特徴量を学習するようにしたことにより、画面のノイズや無意味な変化に惑わされることなく、画像から興味のあるものの変化だけに注目することができるようにしている。これにより、外部報酬がなかなか得られない長い迷路でも学習が進むようになったことや、外部報酬を一切与えなくても好奇心による内部報酬のみで探索が進むようになったことが報告されている。また、内部報酬のみによる学習結果をそのまま用いても、学習に用いたステージ以外でも探索が進むという汎化能力を持つことも報告されている。以下に、この[7]方式についてより詳しく説明する。

2. 自分の行動に関係した好奇心による強化学習

文献[7]に報告されている、自分の行動に関係した好奇心で動機付けされた強化学習について、より詳しく説明する。

この手法では、画像から興味のあるものの変化だけに注目することで、画面のノイズや無意味な変化に惑わされることなく行動が選択できるようにしている。そのため、図 1 に示されているように、Intrinsic Curiosity Module (ICM)の中に順モデルと逆モデルを持っている。

逆モデルは、ある時刻の状態の特徴量と次の時刻の状態の特徴量からある時刻での行動の評価値を推定し、順モデルは、逆モデルで求めたある時刻の状態の特徴量と行動の評価値から次の時刻の状態の特徴量を推定している。また、内部報酬は、順モデルの誤差に定数項を掛けた形になっている。

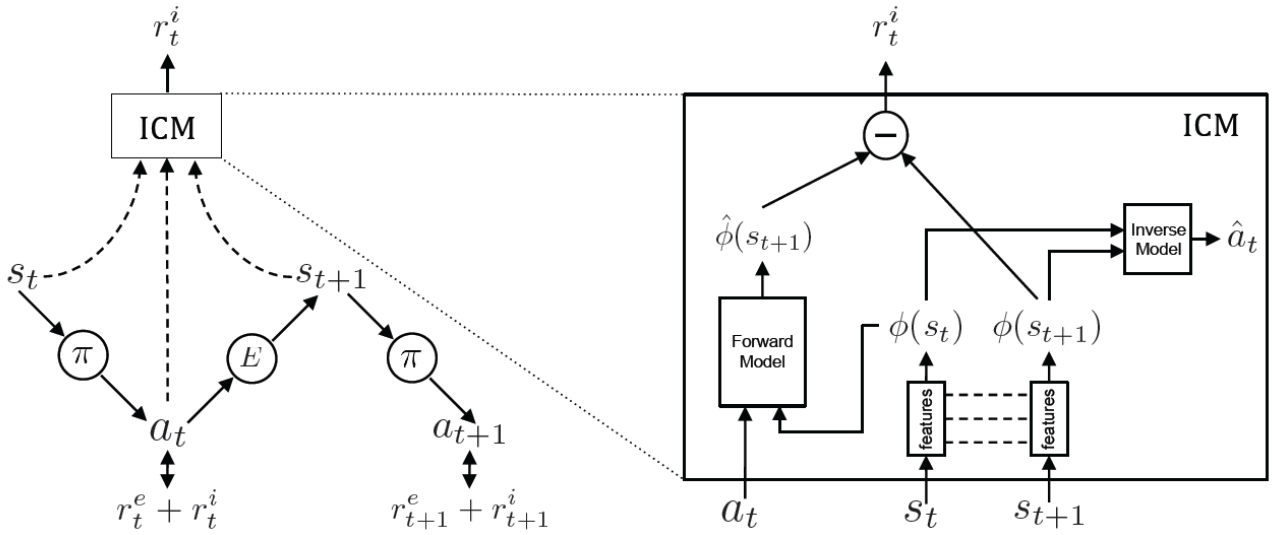

図 1：自分の行動に関係した好奇心で動機付けされた強化学習の原理図([7]の Figure 2 より引用)

Level Ids	Level-1	Level-2				Level-3			
Accuracy	Scratch	Run as is	Fine-tuned	Scratch	Scratch	Run as is	Fine-tuned	Scratch	Scratch
Iterations	1.5M	0	1.5M	1.5M	3.5M	0	1.5M	1.5M	5.0M
Mean ± stderr	711 ± 59.3	31.9 ± 4.2	466 ± 37.9	399.7 ± 22.5	455.5 ± 33.4	319.3 ± 9.7	97.5 ± 17.4	11.8 ± 3.3	42.2 ± 6.4
% distance > 200	50.0 ± 0.0	0	64.2 ± 5.6	88.2 ± 3.3	69.6 ± 5.7	50.0 ± 0.0	1.5 ± 1.4	0	0
% distance > 400	35.0 ± 4.1	0	63.6 ± 6.6	33.2 ± 7.1	51.9 ± 5.7	8.4 ± 2.8	0	0	0
% distance > 600	35.8 ± 4.5	0	42.6 ± 6.1	14.9 ± 4.4	28.1 ± 5.4	0	0	0	0

表 1：自分の行動に関係した好奇心で動機付けされた強化学習の汎化性能([7]の Table 1 より引用)

強化学習は、式 1 に示す全体損失を最小化する形で行われており、方策には A3C[3]が用いられている。

$$\min_{\theta_P,\theta_I,\theta_F}\left[-\lambda\mathbb{E}_{\pi(s_t;\theta_P)}[\Sigma_t r_t]+(1-\beta)L_I+\beta L_F\right]$$

内部報酬　順モデルの誤差　逆モデルの誤差　方策による期待報酬

式 1：全体損失の式（[7]の式(7)より引用）

この手法により、外部報酬がなかなか得られない長い迷路でも学習が進むようになったこと、外部報酬を一切与えなくても好奇心による内部報酬のみで探索が進むこと、内部報酬のみによる学習結果をそのまま用いても、学習に用いたステージ以外でも探索が進むという汎化能力を持つことなどが報告されており、大変興味深い結果となっている。

しかしながら、この文献で報告されている汎化能力の実験では、表 1 にあるように、「スーパーマリオブラザーズ」に関しては 2 ステージに適用しているだけであった。(なお、表 1 の Level-1、Level-2、Level-3 は、本報告書ではステージ 1-1、ステージ 1-2、ステージ 1-3 と表記。)

そこで、好奇心で動機付けされた強化学習の汎化能力についてさらに調べるため、もっと多くのステージに対して、ステージ 1-1 で学習した結果をそのまま用いて、どの程度新しいシナリオに対応できるかの実験を行った。

3. 実験方法

OpenAI Gym[2]上で、[7]のアルゴリズムにより「スーパーマリオブラザーズ」のステージ 1-1 で学習し、ステージ 1-1 で学習した状態をそのまま用いてステージ 1-1 からステージ 3-4 を実行して、どの程度新しいシナリオに対応できるかを評価した。

以下に実験方法の詳細を記載する。

① OpenAI Gym 上で「スーパーマリオブラザーズ」のステージ 1-1 を 1.5M ステップ学習（8 プロセス並行実行）

② ステージ 1-1 で学習した状態をそのまま用いて、ステージ 1-1 からステージ 3-4 の各ステージで 100 エピソードずつ実行

③ 各エピソードにおける最遠到達距離(distance)を評価値として保存

4. 実験結果

図 2 に、「スーパーマリオブラザーズ」のステージ 1-1 で学習し、ステージ 1-1 で学習した状態をそのまま用いて、ステージ 1-1 からステージ 3-4 の各ステージで 100 エピソードずつ実行して得た、各ステージの 100 エピソードの最遠到達距離(distance)の平均値を示す。この結果を見ると、学習に用いたステージ 1-1 の評価値と比較して、ステージ 1-1 で学習した状態をそのまま用いた他のステージでも、平均 distance が 200 以上あるものが 6/11 と 54%あり、汎化性能があることが確認できる。

なお、ステージ 2-2 の平均 distance が特に悪くなっているが、この理由については、5 章で考察する。

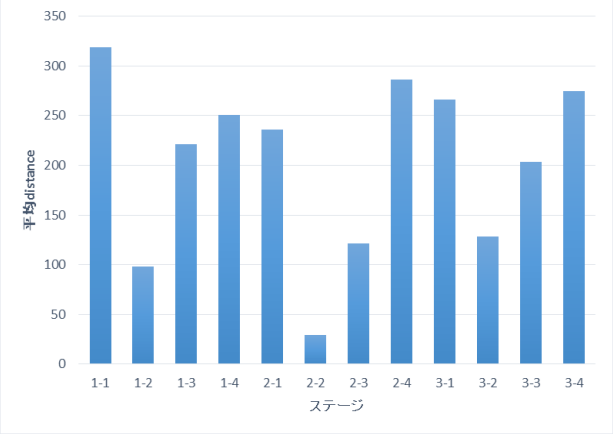

図 2：各ステージの平均 distance

ステージ	% distance > 200	% distance > 400	% distance > 600
1-1	75.0	27.0	1.0
1-2	20.0	3.0	0.0
1-3	67.0	1.0	0.0
1-4	98.0	12.0	0.0
2-1	74.0	2.0	0.0
2-2	0.0	0.0	0.0
2-3	4.0	0.0	0.0
2-4	100.0	2.0	0.0
3-1	77.0	7.0	0.0
3-2	28.0	8.0	2.0
3-3	67.0	1.0	0.0
3-4	100.0	2.0	0.0

表 2：各ステージの到達割合

次に、表 2 に、上記と同じ条件で最遠到達距離(distance)が 200、400、600 以上に達したエピソードが各ステージで 100 エピソード中何%あったかを示す。 参考文献の結果の表 1 と比較すると、ステージ 1-1 の成績が distance > 200 は 75.0%で表 1 の 50.0%より良いが、distance > 600 の割合は 1.0%と表 1 の 35.8%より悪くなっている。また、ステージ 1-2 は distance > 200 が 20.0%と表 1 の 0.0%より成績が良くなっている。これらの理由については、5 章で考察する。

最後に、図 3 から図 6 に、ステージ 1-1 から 1-4 までの各ステージの 100 エピソードの最遠到達距離(distance)のヒストグラムを示す。

学習に用いたステージ 1-1 の図 3 では最遠 distance が幅広く分布しているが、学習していない新しいシナリオである図 4 から図 6 では、少数の最遠 distance に偏る傾向がみられる。この理由についても、5 章で考察する。

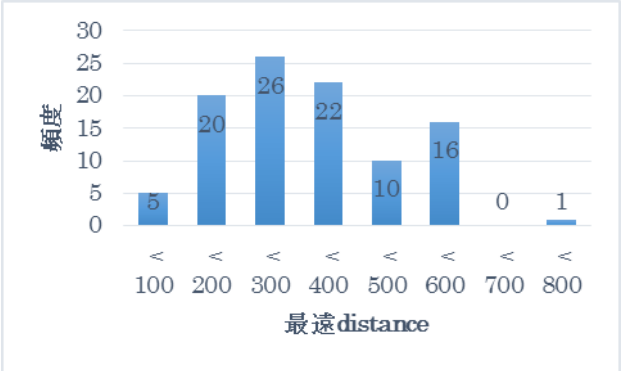

図 3：ステージ 1-1 の distance ヒストグラム

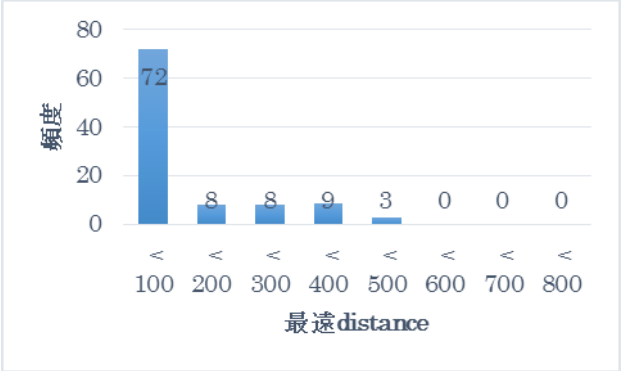

図 4：ステージ 1-2 の distance ヒストグラム

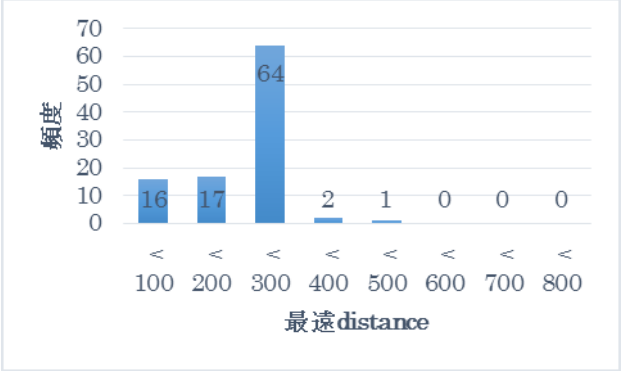

図 5：ステージ 1-3 の distance ヒストグラム

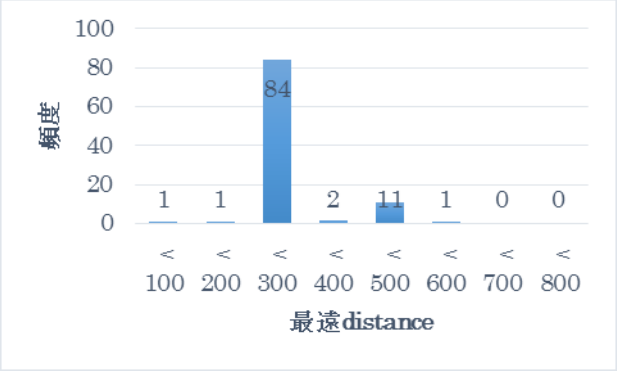

図 6：ステージ 1-4 の distance ヒストグラム

5. まとめと考察

学習に用いたステージ 1-1 を除く 11 ステージでの実験結果より、好奇心で動機付けされた強化学習の汎化能力が確認できた。

ステージ 2-2 の成績が良くない理由は、ステージ 2-2 が水中のステージで、上にジャンプするのではなく、泳いで下に潜るという、学習したステージ 1-1 とは全く異なるアクションが要求されるため、成績が良くないと考えられる。

ステージ 1-1 の成績が参考文献と比較すると、distance > 200 は良いが、distance > 600 が悪くなっている理由としては、「スーパーマリオブラザーズ」の特有の事情として、学習試行時にステージの後半まで行くと、後半から開始されるという点が挙げられ、[7]の著者の環境では 20 プロセスで学習していたと思われるが、本実験ではマシンの能力の関係上 8 プロセスで学習したため、早い段階で大部分のプロセスが後半に移行してしまい、前半の学習が不足していることが原因として考えられる。逆に、ステージ 1-2 が参考文献と比較して成績が良いのは、割合の多くなった後半の学習がロングジャンプ等を多く含み、成績に良い影響を与えていると推測される。

学習に用いたステージ 1-1 では最遠 distance が幅広く分布しているが、学習していない新しいシナリオでは少数の最遠 distance に偏る傾向がみられた理由としては、学習に用いたステージでは進むのが難しい状況があっても多数回の学習試行の中に進む方策を見つけ出すことができる可能性が高いため、そこで引っかかって進まなくなってしまう可能性が下がるが、学習していない新しいシナリオでは、以前に学習した方策の汎化能力を超える難しい状況に出会ったときに、そこで引っかかって進まなくなってしまう可能性が高いためと推測される。

以上のように、好奇心で動機付けされた強化学習の汎化能力を確認したが、現実の環境でも外部報酬がなかなか得られない場合が多く存在するので、好奇心のような内部報酬で動機付けされた強化学習は今後も注目されていくと思われる。

参考文献

[1] Bellemare, M., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., & Munos, R.: Unifying count-based exploration and intrinsic motivation, In Advances in Neural Information Processing Systems (pp. 1471-1479), (2016)

[2] Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., & Zaremba, W.: OpenAI gym, arXiv preprint arXiv:1606.01540, (2016)

[3] Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., ... & Kavukcuoglu, K.: Asynchronous methods for deep reinforcement learning, In International Conference on Machine Learning (pp. 1928-1937), (2016)

[4] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M.: Playing atari with deep reinforcement learning, arXiv preprint arXiv:1312.5602, (2013)

[5] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Petersen, S.: Human-level control through deep reinforcement learning, Nature 518(7540) 529-533, (2015)

[6] Oh, J., Guo, X., Lee, H., Lewis, R. L., & Singh, S.: Action-conditional video prediction using deep networks in atari games, In Advances in Neural Information Processing Systems (pp. 2863-2871), (2015)

[7] Pathak, D., Agrawal, P., Efros, A. A., & Darrell, T.: Curiosity-driven exploration by self-supervised prediction, arXiv preprint arXiv:1705.05363, (2017)

[8] Schmidhuber, J.: A possibility for implementing curiosity and boredom in model-building neural controllers, In From animals to animats: proceedings of the first international conference on simulation of adaptive behavior (SAB90), (1991)

[9] Stadie, B. C., Levine, S., & Abbeel, P.: Incentivizing exploration in reinforcement learning with deep predictive models, arXiv preprint arXiv:1507.00814, (2015)

[10] Tang, H., Houthooft, R., Foote, D., Stooke, A., Chen, X., Duan, Y., ... & Abbeel, P.: #Exploration: A Study of Count-Based Exploration for Deep Reinforcement Learning, arXiv preprint arXiv:1611.04717, (2016)