The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

強化学習における報酬なしスキル獲得の階層化

Hierarchization of Skill Acquisition without a Reward Function in Reinforcement Learning

皿海孝典 Takanori Saragai　狩野泉実 Izumi Karino　國吉康夫 Yasuo Kuniyoshi

東京大学
The University of Tokyo

In recent years, many researches on pre-training have been done. For the pre-training in which a target task is unknown, some methods enable an agent to acquire skills without a reward function. These methods have a limitation that learning becomes unstable as the number of skills increases. In this research, we propose a model to acquire more skills while keeping diversity by hierarchizing skill acquisition. As a result, we achieved learning in a short time compared to existing methods for the same number of skills. Examination of multiple indicators showed that the diversity of skills is preserved. In addition, we confirmed that features of an agent state tend to be passed down through the hierarchical skills.

1. はじめに

近年, 強化学習はビデオゲーム [Mnih 13] やロボットコントロール [Schulman 15] など様々な分野に応用されている. これらは単一のタスクにおいて極めて高い性能を示すものの, それぞれのタスクにおいて長い学習時間が必要となる. この問題を解消するために, 転移学習 [Pan 10] やメタラーニング [Finn 2017] のように目的タスクを解くために別のタスクで事前学習した知識を転用する手法が提案されている.

事前学習において, 目的タスクが既知であるかどうかによって方針が大きく分かれる. 目的タスクが既知であれば, 目的タスクに対して活用可能な知識を得られるタスクを選び学習することで直接目的タスクを学習するより速く学習が進むことが知られている [Svetlik 17].

一方, 目的タスクが未知であれば, 外部からの報酬がわからないため内部報酬によって事前学習を行う必要があり, エージェントが多様な動き方を獲得することを内部報酬とする手法が提案されている [Eysenbach 18, Gregor 16]. この動き方をスキルと呼び, 目的タスクを与えられたときに獲得済みスキルの中で最もタスクに適したスキルを使う, あるいはそのスキルを方策の良い初期値として用いることで目的タスクでの学習を高速化することが可能である. このとき獲得しておくスキルは, 多数かつ多様であることで広い範囲のタスクに対して適用可能であると考えられるが, スキルの数が増えると学習が不安定になることが指摘されている [Gregor 16]. また, スキルの数が増えるに従って, 必要なスキルを検索する時間も増える.

本稿では多数かつ多様なスキルの安定的な獲得と, 獲得したスキルの検索性の向上のために, スキルの獲得を階層化する手法を提案する.

2. 関連研究

目的タスクが未知の事前学習において獲得するスキルを多様化する. スキルが多様であるとは, 多数のスキルに対してスキルごとにエージェントの状態が異なるということを意味する. これを定式化するためにエージェントの状態とスキルの相互情報量に注目する.

連絡先: 皿海孝典, 東京大学, saragai@isi.imi.i.u-tokyo.ac.jp
國吉康夫, 東京大学, kuniyosh@isi.imi.i.u-tokyo.ac.jp

相互情報量 MI は, 2つの確率変数 X, Y によって以下のように定義される.

$$MI(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{1}$$

一般に相互情報量は非負であり, X と Y が完全従属かつ X のエントロピーが最大のときに最大値をとる.

スキルと状態の相互情報量を考える. スキルと状態の相互情報量はスキルと状態が従属であり, かつ多様な状態を取るときに最大化する. これは上で述べた多様なスキルの条件を満たすため, スキルと状態の相互情報量を内部報酬とする研究がなされている.

Variational Intrinsic Control は内部報酬をスキルと終端状態の相互情報量とする手法である [Gregor 16]. 初期状態から終端状態に到達するまでの途中経過を無視しているため, 1 エピソードあたり 1 ステップ分の経験しか得られないため学習の効率が悪い. また, 獲得するスキル数を増やすと不安定になることが指摘されている [Gregor 16].

Diversity is All You Need (DIAYN) は終端状態に限らずに, 全ての時点でのスキルと状態の相互情報量を内部報酬とする手法である [Eysenbach 18]. さらに DIAYN では安定化のために方策のエントロピー最大化強化学習を用いて定式化している. 具体的にはスキルを z, 状態を s, 行動を a, エントロピーを $\mathcal{H}[\cdot]$ として,

$$\mathcal{F}(\theta) \equiv MI(s,z) + \mathcal{H}[a|s] - MI(a,z|s) \tag{2}$$

を最大化する. 式の第一項はスキルによってエージェントの到達する状態を操作できるようにするために, 第二項はすべてのスキルを合わせたものを一つの方策としてみたときにエントロピー最大になるようにするために, 第三項はスキルを多様にするのは行動ではなく到達する状態であるようにするために, それぞれの項が設計されている. この結果, 状態空間と行動空間が連続となってもスキルを獲得することが可能となった.

しかし, いまだに獲得するスキル数を増やしたときに安定して学習を行うことはできていない. そこで本手法はスキルを階層化することによって安定してスキル数を増やす方法を提案する.

3. 目的と手法

本研究の目的は安定的に多くのスキルを獲得すること, タスクごとに最適なスキルの高速な検索を可能することであり, スキルの学習を階層化することによって目的を達成する.

図 1 に示すモデルによってスキルの階層化を行う. 基本的な構造は DIAYN のものを用いる. DIAYN には方策と分類器が存在し, 各エピソードで無作為に 1 つのスキルが方策に与えられる. 分類器は各ステップの行動後の状態から方策に与えられたスキルを予測する.

式 2 を変形する.

$$\mathcal{F}(\theta) \geq \mathcal{H}[a|s,z] + \mathbb{E}[\log q_\theta(z|s) - \log p(z)] \tag{3}$$

ここで, q_θ はパラメータ θ で表される分類器を表し, 状態 s のときに方策に与えられたスキルが z である確率の予測値を出力する関数となっている.

本研究ではスキルの獲得を階層化するために, この分類器を用いることを提案する.

まず DIAYN によりスキルを獲得し, 獲得したスキル集合 Z^0 のうち i 番目のものを選択する. これをスキル z_i^0 とする. このとき, DIAYN の学習において分類器 $q^0(z|s)$ を学習したとする. ここで学習が収束し, 完全な分類器ができたとすると, z_i^0 についての方策である $\pi_\theta(a|s,z_i^0)$ は報酬

$$r_i^0(s) \equiv \log q^0(z_i^0|s) - \log p(z_i^0) \tag{4}$$

の累積値を最大化する方策となっている. このとき, $q^0(z|s)$ と z_i^0 を固定することで報酬を s のみの関数とみなしている.

次に, 同じ環境においてスキルを獲得する際, DIAYN の報酬に $r_i^0(s)$ を加える. その結果スキル z_i^0 が遷移していた状態に対する指向性を持ちつつ, 多様な状態へスキルを分けようとするため, z_i^0 のスキルの特徴を持ちつつ, 様々なバリエーションを持つようなスキル集合 Z_i^1 が生まれることが予想される.

一般に, DIAYN の n 回目の適用時に獲得された i 番目のスキルに対して分類器 $q^n(z|s)$ が存在し, 同じ環境において

$$\begin{aligned} r^n(s,a) =& \mathcal{H}[a|s,z] + \mathbb{E}[\log q_\phi^n(z|s) - \log p(z)] \\ &+ \mathbb{E}[\log q^{n-1}(z_i^{n-1}|s) - \log p(z_i^{n-1})] \end{aligned} \tag{5}$$

の最大化を行うことでスキル z_i^{n-1} に判別されるという制限を満たしながら多様なスキル集合 Z_i^n を獲得することを期待する.

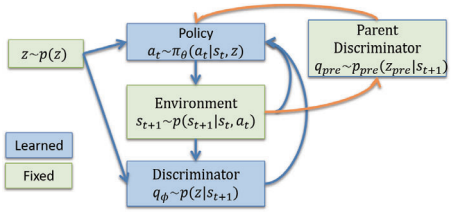


図 1: 提案手法のモデル.

4. 実験

4.1 実験環境と条件

実験環境は OpenAI Gym[Brockman 16] の Mujoco 物理シュミレータ [Todorov 12] において, 図 2(a) に示す Hopper を用いた. Hopper は状態が 11 次元, 行動が 3 次元の 3DoF ロボットである.

方策の学習手法には Soft Actor-Critic(SAC)[Haarnoja 18] を使用した. 隠れ層の大きさは全てのネットワークで (300, 300), 学習率は全て 0.0003, 割引率を 0.99, バッチサイズを 128, エントロピー最大化強化学習における温度計数を 0.1, SAC の状態価値関数の更新率は 0.01, SAC の方策の Gaussian Mixture Model の混合数を 4, リプレイバッファのサイズを 10^6, 最大エピソード長を 1000 ステップとした.

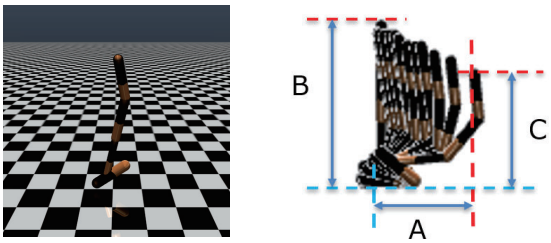
4.2 実験方法

まず DIAYN において 5 個のスキルを獲得させ, これを階層化してそれぞれ 5 個のスキルを持つ子エージェントを 5 個生成して, 計 25 個の子スキルを獲得させる. 比較のため, DIAYN において 25 個のスキルを獲得させる. ここでスキルの獲得における学習の収束時間を比較し, 提案手法の安定性を確認する.

次に獲得したスキルに対してエージェントの軌跡を取って複数の指標を測り, 既存手法と比較したときのスキルの多様性と階層構造における親スキルと子スキルの関係を調べる. 指標として,

- エピソード終了時の位置
- エピソード中の最高到達高さ
- エピソード中の最低到達高さ

の 3 つを測定する. 図 2(b) にそれぞれの指標を図示した.



(a) Hopper 環境　　(b) 指標

図 2: 実験環境の概観. (a)Hopper のシミュレーション環境. (b) 各指標を Hopper の連続写真を用いて図示した. (A) エピソード終了時の位置. (B) エピソード中の最高到達高さ. (C) エピソード中の最低到達高さ.

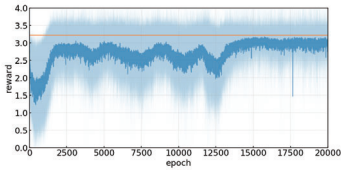
5. 結果

5.1 収束速度の比較

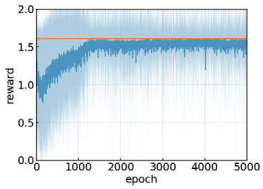
分類器の学習曲線について図 3 に示した. 図 3(a) と図 3(b) を見ると, スキル数が 5 個の場合はおよそ 1500 エポックで収束しているのに対し, スキル数が 25 個の場合は収束におよそ 15000 エポック費やしている. これは, スキル数が増えると学習が安定しないという先行研究の結果を再現している.

図 3(c) は提案手法である階層化スキル獲得に図 3(b) の 3000 エポックまで学習した分類器を用いたときの学習曲線である. 5 つの図はそれぞれ 5 個の親スキルに対応している. これらは 2000 エポックから 4000 エポック程度で十分理想値に近づき収束していることがわかる.

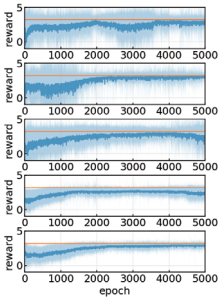
親スキルの獲得に用いた 3000 エポックを加えても 7000 エポック程度で収束をしており, 既存手法が同数のスキル獲得に費やした 15000 エポックと比較して十分早く学習ができている.



(a) DIAYN, スキル数 25 個



(b) DIAYN, スキル数 5 個



(c) 提案手法の下階層

図 3: 分類器の学習曲線. 薄い色は標準偏差, 橙色の直線は報酬の理論値を表す. (a)DIAYN において 25 個のスキルを獲得する際の学習曲線. (b)DIAYN において 5 個のスキルを獲得する際の学習曲線. (c) 提案手法において (b) の 3000 エポックにおける分類器を用いて, 子スキルを学習する際の報酬曲線.

ここで獲得したスキルを図 4 に示す. 図 4(a) の 5 個のスキルを親として, 図 4(b) に示すようにそれぞれ 5 個, 計 25 個の子スキルを獲得した. 前進するスキルを親とすると子スキルも前進するスキルとなり, 子スキル内ではそれぞれが区別可能なスキルが獲得された.

(a) 親スキル

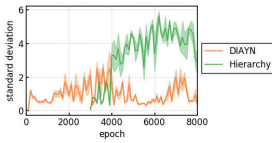
(b) 子スキル

図 4: 獲得されたスキルの 20step ごとの連続写真. (a)3000 エポック学習後の親エージェントの全 5 個のスキル. (b)4000 エポック学習後の子エージェントの全 25 個のスキル. 各列の 5 個のスキルは (a) の同じ列のスキルを親とする子スキルである.

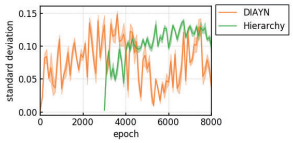
5.2 スキルの多様性

各指標について, 学習回数とスキルの分散について図 5 に示した. 提案手法である階層的なスキルについては, 親スキルの学習に 3000 エポックを使ったため, それだけ右にずらしたものとなっている. エピソード終了時の位置と最高到達高さに関しては階層的なスキルが大きな分散を持っており, 最低到達高さに関しては大きな違いは見られなかった.

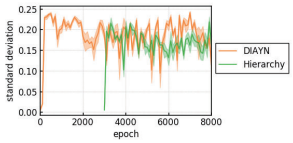
このことから, 計測した範囲内では階層化してもスキルの多様性は保たれていることがわかる.



(a) 終了時の位置



(b) 最高到達高さ



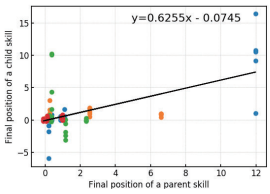
(c) 最低到達高さ

図 5: 25 個のスキルの各指標に対する標準偏差. 緑色は親の 5 スキルそれぞれから獲得された 5 × 5 スキルの標準偏差, 橙色は一つのエージェントの 25 スキルの標準偏差を示す. 100 エポックごとに各スキルで 5 個の軌跡をとった. 実線は標準偏差の平均値, 薄い色は標準偏差の標準偏差を示している.

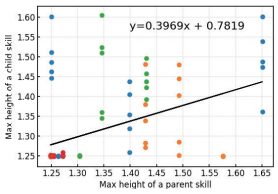
5.3 親と子の相関

同じく複数指標について, 横軸は親スキルの得た値, 縦軸は子スキルの得た値として全ての子スキルに対して散布図を取って最小二乗法によってフィッティングした結果を図 6 に示し, 親スキルと子スキルの相関係数と p 値を表 1 に示した. ここで, 分布に正規性を仮定できないため, Spearman の順位相関係数を用いた.

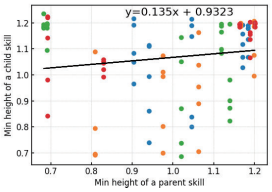
有意水準を 0.05 とすると, 全ての指標において親の値と子の値で有意に正の相関が見られた. 相関係数の大きさに関して, 最低到達低さに関しては弱い相関, それ以外では 0.4 を少し上回る値で中程度の相関であった.



(a) 終了時の位置



(b) 最高到達高さ



(c) 最低到達高さ

図 6: 各指標に対する親スキルと子スキルの値の関係. 同色の点は同じ親エージェントを持つ子スキルを表す. 黒い実線は最小二乗法によってフィッティングしたもので, その式は図中に示している.

表 1: 各指標に対する親スキルと子スキルの Spearman の順位相関係数と p 値.

	相関係数	p 値
エピソード終了時の位置	0.4279	0.000
エピソード中の最高到達高さ	0.4126	0.000
エピソード中の最低到達高さ	0.2360	0.018

6. 結論

スキルの獲得を階層化することにより, 25 個のスキル獲得における学習の収束時間は既存手法の約 15000 エポックから約 7000 エポックと短縮され, スキル獲得が安定化した.

また, 実験した指標において, 階層化をしてもスキルの多様性は損なわれなかった. 加えて, 階層化して生まれた子スキルは実験した指標において親スキルの性質を受け継いでいることがわかった. これは, 必要なスキルを検索する際にまず親スキルを見ることによって子スキルの傾向がある程度推測できるという利点がある.

さらに, 提案手法はスキルを獲得した後でスキル数を増やすことが可能であるため, 例えば獲得したスキルを確認して前進するスキルに関して追加で種類を増やすというような操作が可能であることが示唆される.

7. 展望

親スキルの性質が子スキルに一部受け継がれるという性質が本稿で示された. これを利用して親スキルの傾向から子スキルの傾向を予測し, 目的タスクに有効なスキルの検索を高速化する手法が今後考えられる.

また, 本稿の制限として, 今回実験した範囲では階層は 2 つとして学習を行ったが, 定式化は階層をどれだけでも深くできるようになっている. そのため, 実際に階層を深くするとスキルの獲得にどのような影響があるのかを調べることは今後の課題であると考えている.

深い階層のスキルが獲得可能になれば, 複雑な構造を持つヒューマノイドロボットなどに提案手法を適用して大量のスキルを獲得させることが考えられる. その際に手や表情といった細やかな表現が必要となる部位に関するスキルに限って深く学習をすることにより, 再学習なしでもヒトの表現に足るスキルセットの獲得につながると期待する.

8. 謝辞

本研究は NEDO の支援を受けたものです. また, 田中一敏特任研究員には多くの助言を受けました. 心より感謝申し上げます.

参考文献

[Brockman 16] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, Wojciech Zeremba: Openai gym, arXiv:1606.01540v1 (2016).

[Eysenbach 18] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, Sergey Levine: Diversity is all you need: Learning skills without a reward function, arXiv:1802.06070v4 (2018).

[Finn 2017] Chelsea Finn, Pieter Abbeel, Sergey Levine: Model-agnostic meta-learning for fast adaptation of deep networks, ICML, 1126-1135 (2017).

[Gregor 16] Karol Gregor, Danilo Jimenez Rezende, Dann Wierstra: Variational intrinsic control, arXiv:1611.07507v1 (2016).

[Haarnoja 18] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, Sergey Levine: Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor, NIPS Deep Reinforcement Learning Symposium (2017).

[Mnih 13] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, Martin Riedmiller: Playing atari with deep reinforcement learning, NIPS Deep Learning Workshop (2013).

[Pan 10] Sinno Jialin Pan, Qiang Yang, others: A survey on transfer learning, IEEE TKDE, 22(10), 1345-1359, 10.1109/TKDE.2009.191 (2010).

[Schulman 15] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, Philipp Moritz: Trust region policy optimization, ICML, 1889-1897 (2015).

[Svetlik 17] Maxwell Svetlik, Matteo Leonetti, Jivko Sinapov, Rishi Shah, Nick Walker Peter Stone: Automatic curriculum graph generation for reinforcement learning agents, AAAI, 2590-2596 (2017).

[Todorov 12] Emanuel Todorov, Tom Erez, Yuval Tassa: Mujoco: A physics engine for model-based control, IROS, 5026-5033, 10.1109/IROS.2012.6386109 (2012).