人工知能学会研究会資料
SIG-AGI-009-02

Hierarchical Multi Objective Deep Reinforcement Learning with Accumulator Based Arbitration Mode

滝口 啓介 [1*]　大澤 正彦 [1,2]　今井 倫太 [1]
Keisuke Takiguchi[1], Masahiko Osawa[1], Michita Imai[1]

[1] 慶應義塾大学大学院理工学研究科
[1] Faculty of Science and Technology, Keio University
[2] 日本学術振興会 特別研究員 (DC1)
[2] Japan Society for the Promotion of Science, Research Fellow (DC1)

Abstract: In reinforcement learning, it is difficult to solve high-dimensional multi-objective decision problem. In this paper, we propose a novel hierarchical architecture of deep reinforcement learning with Accumulator Based Arbitration Mode(ABAM). This architecture has some groups which have deep Q-network (DQN) modules respectively and layers containing these modules to solve high-dimensional multi-objective decision problem. And ABAM arbitrates the outputs of modules in one layer by using the outputs of modules in the above layer which has more abstractive objective. By using this architecture, it is able to solve difficult maze task which simple DQN cannot solve.

1　はじめに

深層強化学習によって，ゲームタスクやロボット制御などの様々なタスクを解決できるようになってきている．しかし，実世界におけるタスクは高次元の多目的意思決定問題であることが多く，単純な深層強化学習のアーキテクチャでは複数の目的を持つことができないために，そのような問題を解くことは困難である．

Roijers ら [5] は，複数の目的を同時に最適化する際，複数の目的の最適化を行う事と特定の目的の最適化を行う事にトレードオフが発生するという問題があると述べている．

Vezhnevets[3] らの研究では，上位層と下位層で構成される FeUdal Networks という階層型の深層強化学習アーキテクチャを提案している．FeUdal Networks では，上位のモジュールがより抽象的な目的を学習することで，ネットワーク内部に複数の目的を獲得することができるということを示している．この手法では end-to-end に学習が行える一方で，学習する内容を事前に設定できず，また学習済みのモデルに対して機能を部分的に変更するということが難しい．

また，Tajmajer[4] らの研究では，異なる目的ごとに学習を行なった DQN モジュールを並列に並べ，全体の出力をそれぞれのモジュールの出力と，同時に出力する信頼値に応じて制御するという手法を提案している．この手法では，モジュール間のやりとりが発生しないので，全てのモジュールのもつ目的が同じ抽象度のものに限定されてしまう．

本研究では，Deep Q-Network (DQN)[1] のようなモジュールを複数利用することで，高次元の多目的意思決定問題の解決を可能とする階層型アーキテクチャを提案する．提案するアーキテクチャを用いることで，単純な深層強化学習のアーキテクチャでは学習が困難なタスクを解くことが可能となると考えられる．提案するアーキテクチャの内部には目的の抽象度ごとに分けられた層構造があり，またそれぞれの層には複数のモジュールが存在する．それぞれの層におけるモジュール群の出力は，Accumulator Based Arbitration Mode（ABAM）[2] という手法を利用した ABAM Connection に接続されている．ABAM Connection という接続手法により，より抽象的な目的を持つ上位層のモジュールの出力を利用して，接続されている層のモジュール群の出力を調停する．

提案するアーキテクチャでは，異なる目的を個別のモジュールによって管理し，モジュールごとに学習を行うことが出来るため，Roijers らが述べたような学習時の問題が発生しにくくなると考えられる．さらに，提案するアーキテクチャでは，目的ごとに異なるモジュールを設置しているため，機能を部分的に変更するということを比較的容易に行うことが出来ると考えれられる．

本論文では,2 章で本研究に関する関連研究を挙げ，3

*連絡先：慶應義塾大学大学院理工学研究科
神奈川県横浜市港北区日吉 3-14-1 26-203
E-mail: takiguchi@ailab.ics.keio.ac.jp

章で提案手法についての説明を述べる．4 章の実験では，時系列要素のある迷路探索タスクの学習について説明する．5 章の考察では実験結果から提案手法の是非について述べる．

2　関連研究

ABAM[2] は，前頭前野の知見に基づいたアンサンブル学習の手法であり，接続された複数のモジュールの出力を信頼度に基づいて調停することを可能とする．ABAM では，時刻 t における各モジュールの出力に対して，累積証拠 A_t を割り当てる．A_t は，以下の式にしたがって毎ステップ更新される．

$$A_t = \lambda A_{t-1} + p_t \tag{1}$$

ここで，p_t は，時刻 t におけるモジュールの選択確率である．また，λ は割引定数である．A_t があらかじめ設定された閾値を超えた場合，対応するモジュールが発火したとみなし，発火したモジュールの出力を ABAM の出力とする．

3　提案する階層型アーキテクチャ

本研究では，複数の学習済みモジュールを階層的に配置することで，高次元の多目的意思決定問題の解決を可能とする階層型アーキテクチャを提案する．ここで，提案するアーキテクチャの概要を図 1 に示す．この章では，提案するアーキテクチャの構造についてと，それぞれの層の出力を制御する ABAM Connection についてを説明する．

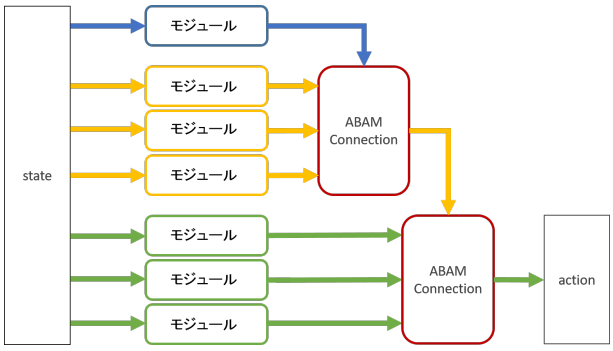

図 1: 提案するアーキテクチャの概略図．ABAM Connection には，接続されたモジュール群の出力と上位の層の出力が入力として与えらえる．

3.1　階層構造

提案するアーキテクチャでは，目的の抽象度ごとに異なる層を定義する．また，それぞれの層には，その層に定義された目的の抽象度に適した目的に対して学習を行なったモジュールを設置する．それぞれの層にあるモジュール群の出力は，後述する ABAM Connection という手法で制御する．目的の抽象度ごとに階層を作り，層ごとのモジュール群の出力を上位の出力を利用して制御することにより，より複雑な行動決定が可能となる．また，複数のモジュールを利用することにより，タスクを複数の簡単なタスクへと細分化することが可能となり，個々のモジュールの学習を容易に行うことが可能となる．

3.2　ABAM Connection

ABAM Connection は，より抽象的な目的を持つ上位層のモジュールの出力を利用して，接続されている層のモジュール群の出力を調停する．ABAM Connection には入力として上位層のモジュール群の出力と，下位層のモジュール群の出力が与えられる．初めに，受け取った上位層のモジュール群の出力を利用して，下位層のモジュール群に接続している ABAM の閾値を制御する．その後，下位層のモジュール群の出力を接続している ABAM に入力として与え，発火したモジュールの出力を ABAM Connection の出力とする．閾値を制御する方法は，ニューラルネットワークを用いて学習を行う方法や，条件分岐によって制御する方法，事前に固定値に定めておく方法などがあげられる．ABAM Connection は，従来までの ABAM[2] が同じ抽象度の行動しか調停できていなかったのに対して，新たに抽象度が異なるものも調停できるように拡張している．

4　実験

提案するアーキテクチャの実用性を示すために，時系列要素のある迷路探索タスクを用意した．

4.1　提案するアーキテクチャの設定

このタスクを解くために，提案するアーキテクチャの内部に二つの層を作成した．上位に配置する層には 1 つの抽象的な目的を持つモジュールを設置し，下位に配置する層には 5 つの移動に関するモジュールを設置した．上位の層に配置するモジュールは，事前に順序に関するタスクを学習したモジュールを設置した．下位の層には，それぞれ迷路の右上，右下，左上，左下，中央へと移動するという目的に対して学習を行なった

モジュールを設置した．また，下位の層のモジュール群の出力は ABAM Connection に接続され，上位の層のモジュールの出力に応じて制御される．今回の実験では，ABAM Connection による閾値の制御は学習によって獲得するのではなく，事前に条件分岐によって設定したものを使用した．

4.2 迷路探索タスクの設定

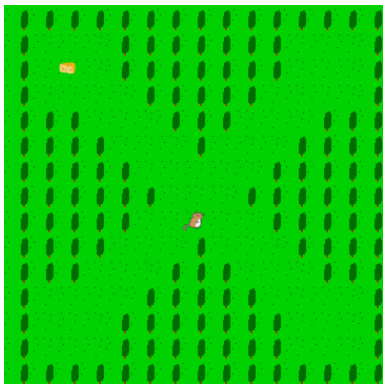

図 2: 実験で用意した迷路探索タスクの環境．

図 2 に，この迷路探索タスクの環境を示す．このタスクでは，ネズミを模した Agent が，迷路内に出現するチーズを獲得することで報酬が与えられる．ネズミは，各ステップで自分がいるマスの上下左右のマスのどこかに移動することができるが，木が描かれているマスへの移動はできない．チーズはネズミに獲得されると迷路内の別の場所へと移動する．チーズの移動には規則があり，右上，左下，右下，左上という順番で，それぞれ決められた場所へと移動する．このタスクでは経路探索と時系列要素の記憶を同時に学習しなければならないため，単純な深層強化学習のアーキテクチャでは学習を行うのが困難であると考えられる．

4.3 事前学習の設定

抽象的な目的を持たせるため，上位の層に設置するモジュールに対して事前に順序に関するタスクに関して学習を行った．図 3 に，順序に関するタスクにおける環境を示す．順序に関するタスクでは，迷路探索タスクと同様にネズミを模した Agent が環境内を動き回り，配置されたチーズを獲得することで報酬が与えられる．チーズは，迷路探索タスクと同様の移動規則を持つ．このタスクでは，ネズミは右上，左下，右下，左上，中心の 5 つの場所へ直接移動することが出来る．また，右上，左下，右下，左上へと移動する際には，中心を通らなければならない．このタスクを学習するこ

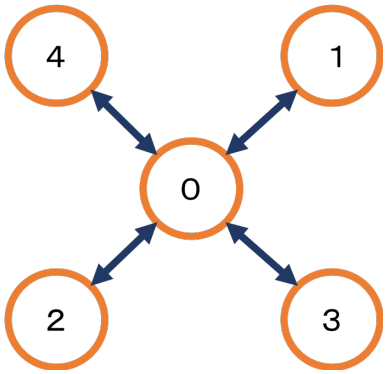

図 3: 上位モジュールが学習を行なったタスクの環境．

とによって，このモジュールはチーズの出現順序を推測することが可能となる．学習は 1 回のエピソード中に 10000 ステップを実行し，チーズを獲得するたびに報酬を獲得する．

下位の層に設置するモジュールは，それぞれ事前に迷路の右上，右下，左上，左下，中央へと移動するという目的に対して学習を行なった．モジュールには DQN を使用し，入力としてこれらのタスクを学習したモジュールを設置することにより，方向というより高次元な移動目的のもと行動を選択することが可能となる．

4.4 実験結果

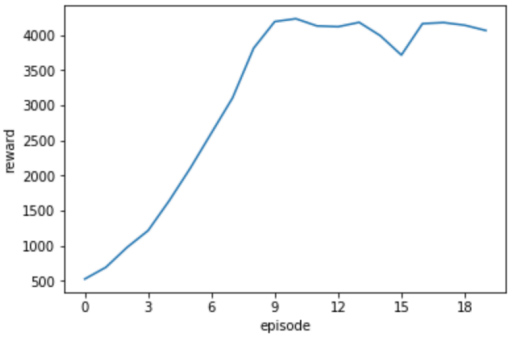

図 4: 順序に関するタスクを学習した際の結果．

表 4 に，順序に関するタスクを学習した際の結果を示す．タスクをより学習が容易なタスクへと分解したことで，学習が行えていることが確認できる．

また，すべてのモジュールをそれぞれ学習させ，提案するアーキテクチャに配置し，ABAM Connection に適切な制御条件を設定した結果，Agent が効率よくチーズを獲得する行動をすることが確認できた．この実験において，ABAM Connection による閾値の更新は 20 ステップごとに行なった．

5　考察

Tajmajer[4] らの手法では，結果的にモジュールは環境から直接行動を推測しなければならず，高次元の空間では学習が困難になることが予想される．一方で提案したアーキテクチャでは，層ごとに抽象的な目的を解いているため，それぞれの層でタスクが簡易化されていると考えることが出来る．また，個々のモジュールが学習するタスクはより簡易なものとなっている．そのため，全体的に学習のコストを下げることができていると考えられる．以下の図 5 と図 6 はそれぞれ，下位層のモジュールによって簡易化されたタスクの環境のイメージと，アーキテクチャ全体によって簡易化されたタスクの環境のイメージである．

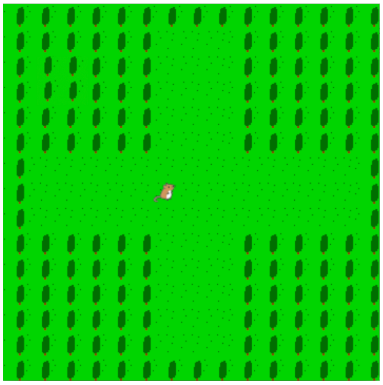

図 5: 下位層のモジュールによって簡易化されたタスクの環境のイメージ．

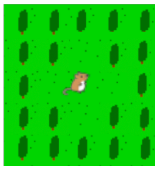

図 6: アーキテクチャ全体によって簡易化されたタスクの環境のイメージ．

実世界におけるタスクはそのほとんどが高次元の多目的意思決定問題であるため，タスクを直接解くというのは現実的ではない．そのような際，提案したアーキテクチャのように，タスクを分離し効率的に複数のモジュールを使用するというアーキテクチャを用いることでタスクの解決が期待できるのではないかと考えられる．この実験のタスクでは，提案したアーキテクチャがタスクを分離し効率的に複数のモジュールを使用する事が可能であるということを示している．

また，本手法では層と層の接続において ABAM Connectin という手法を用たが，重み部分の効率的な学習の手法については提案していない．さらに，それぞれのモジュールの学習における報酬設計についても，全て人間が設計している．今後の課題として，ABAM Connectin の閾値制御に関する重み部分の効率的な学習の手法や，それぞれのモジュールの学習をさらに効率よく行う方法を考える必要がある．

6　おわりに

本研究では，複数のモジュールを階層的に組み合わせた階層型アーキテクチャを提案した．実験から，提案したアーキテクチャを用いることで，単純な深層強化学習のアーキテクチャでは解決が困難な迷路探索タスクを解決できるということを示した．

参考文献

[1] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al.: Human-level control through deep reinforcement learning, *Nature*, Vol. 518, No. 7540, pp. 529-533 (2015)

[2] Masahiko Osawa, Yuta Ashihara, Takuma Seno, Michita Imai, and Satoshi Kurihara.: Accumulator based arbitration model for both supervised and reinforcement learning inspired by prefrontal cortex, *ICONIP*, (2017)

[3] Alexander Sasha Vezhnevets, Simon Osindero, Tom Schaul, Nicolas Heess, Max Jaderberg, David Silver, and Koray Kavukcuoglu.: FeUdal Networks for Hierarchical Reinforcement Learning, *ICML*, (2017)

[4] Tomasz Tajmajer.: Modular Multi-Objective Deep Reinforcement Learning with Decision Values, arXiv preprint arXiv:1704.06676, 2017.

[5] Diederik M. Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley.: A survey of multi-objective sequential decision-making, *Journal of Artificial Intelligence Research*, Vol. 48, pp. 67–113 (2013)