AI 特許紹介(49)
AI 特許を学ぶ！究める！
～内的報酬特許～

2023 年 2 月 10 日
河野特許事務所
所長弁理士　河野英仁

「AI 特許紹介」シリーズは、注目すべき AI 特許のポイントを紹介します。熾烈な競争となっている第４次産業革命下では AI 技術がキーとなり、この AI 技術・ソリューションを特許として適切に権利化しておくことが重要であることは言うまでもありません。

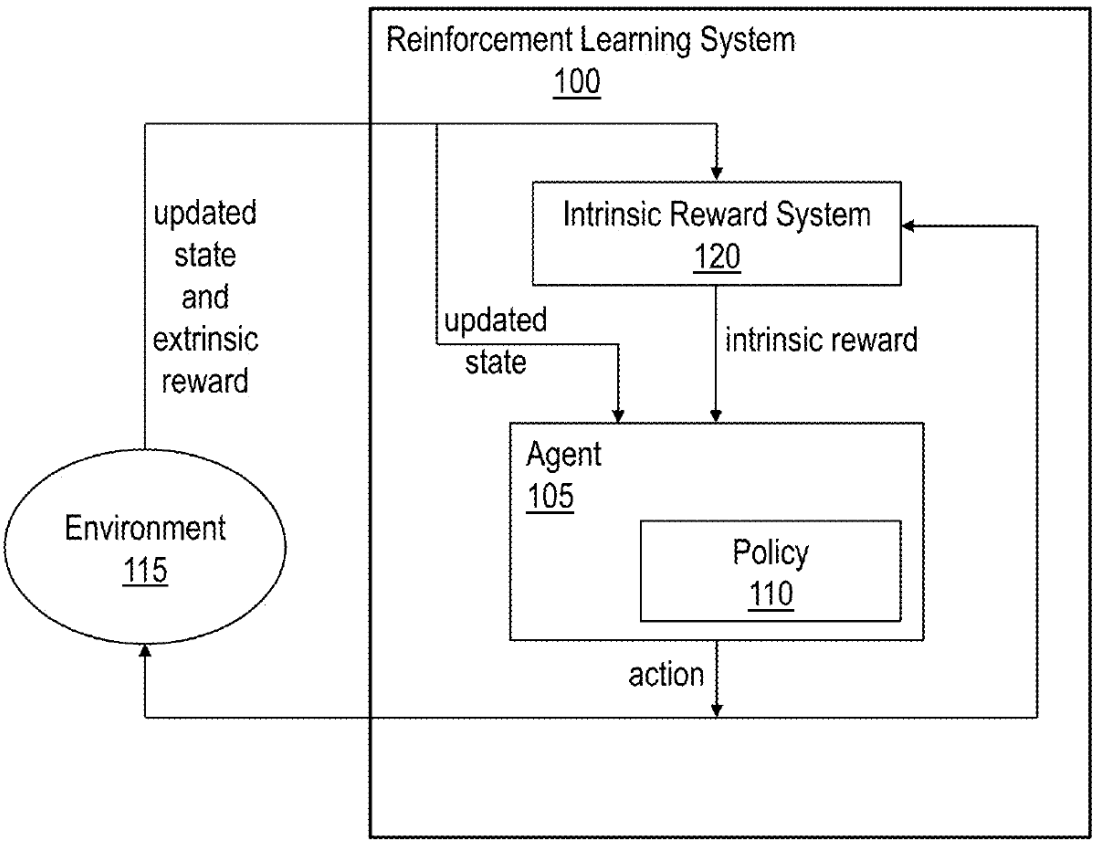
AI 技術は Google, Microsoft, Amazon を始めとした IT プラットフォーマ、研究機関及び大学から毎週のように新たな手法が提案されており、また AI 技術を活用した新たなソリューションも次々とリリースされています。

本稿では米国先進 IT 企業を中心に、これらの企業から出願された AI 特許に記載された AI テクノロジー・ソリューションのポイントをわかりやすく解説致します。

1.概要
特許出願人　DeepMind Technologies
出願日　2020 年 9 月 25 日
公開日　2021 年 3 月 25 日
公開番号 US20210089910
発明の名称　メタ学習された内的報酬を使用した強化学習

910 特許は、強化学習においてエージェントにより実行されるタスクに基づき付与される外的報酬(extrinsic reward)に加えて、エージェントによって取られたアクションに基づいて生成された内的報酬（intrinsic reward）を用いてエージェントのポリシーを更新する技術に関する。

2.特許内容の説明
下記図は、強化学習システム 100 を示す。



強化学習システム 100 は、ポリシー110 に基づいてアクションを実行するように構成されたエージェント 105 を備える。アクションが決定されるたびに、アクションはエージェント 105 が対話している環境 115 に出力される。アクションは環境 115 の状態を更新する。更新された状態は、強化学習システム 100 に、アクションに対する関連する外的報酬とともに返される。環境 115 によって生成される報酬は、外的報酬として知られている。

対照的に、内的報酬は、強化学習システム 100 によって内部的に生成される。これに関して、強化学習システム 100 は、固有報酬システム 120 をさらに備える。内的報酬システム 120 は、エージェント 105 によって取られた行動および環境 115 の更新された状態に基づいて、エージェント 105 のための内的報酬を生成する。内的報酬は、アクションおよび更新された状態に関連する環境 115 から受け取った外的報酬に基づいて生成される。

従来の強化学習システムは、外的報酬に基づいてエージェントを訓練するが、本強化学習システム 100 は、内的報酬に基づいてエージェント 105 を訓練する。エージェント 105 は、更新された状態および固有の報酬を受け取り、受け取ったデータおよびポリ

シー110 に基づいて次のアクションを決定する。

特に、エージェント 105 による特定のタスクの実行のために、生成された内的報酬を通じて、内的報酬システム 120 は、タスクを実行するために必要な情報を取得するために特定の方法で環境 115 を探索することをエージェント 105 に奨励する。必要な情報が取得されると、内的報酬システム 120 は、環境 115 によって生成される外的報酬を最大化するために、エージェント 105 が環境と対話することを奨励する。

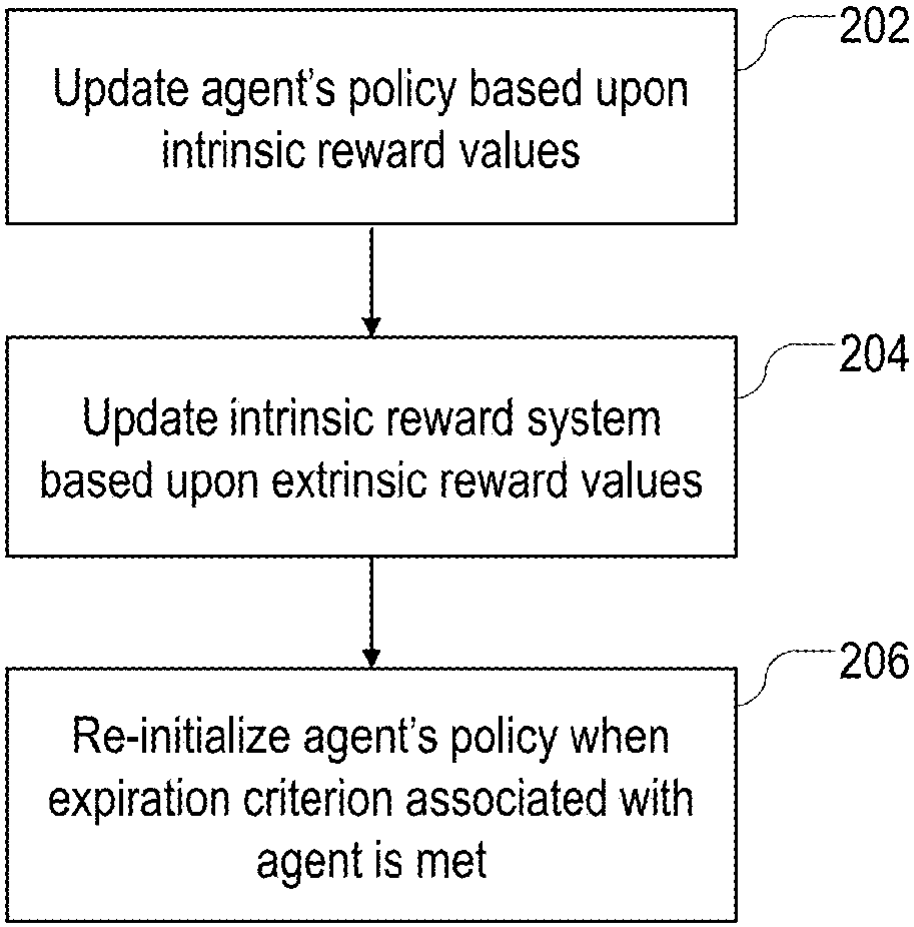
したがって、内的報酬システム 120 は、エージェント 105 に、いつ探索し、いつ利用するか、つまり何をすべきかについてのガイダンスを提供し、エージェントのポリシー110 は、そのような目標を達成するための行動、すなわち、どのようにそれを行うかを決定する。内的報酬システム 120 およびポリシー110 は、そのような知識を符号化するように訓練される。

エージェント 105 によって実行される現在のアクションに加えて、内的報酬システム 120 は、エージェント 105 の履歴に基づいて内的報酬を生成する。履歴は、環境（状態)、エージェントのアクション、取得された外的報酬、および状態が最終であるかどうか、つまり、状態がトレーニングエピソードの最終状態であるかどうかの表示のすべての以前の観察を含む。

ポリシー110 は、エージェントが環境の状態に基づいてアクションを実行する方法を定義する。エージェント 105 が従うポリシー110 は、ポリシー110 を使用して取られたアクションから期待されるリターンを改善するために、近似値関数、またはリターン関数に従ってアクションの値を評価することによって更新される。

これは通常、「リターン」と呼ばれることもある、エージェント 105 によって実行されるアクションの成功を評価するための予測と制御の組み合わせによって達成される。リターンは、特定のアクションの後に受け取った報酬（この場合は本質的な報酬）に基づいて計算される。

次に、強化学習システム 100 のトレーニングについて説明する。下記図は、強化学習システム 100 をトレーニングするためのプロセス 200 を示す。



強化学習システム 100 は、複数のタスクでトレーニングされる。ステップ 202 で、エージェントのポリシー110 は、タスクの遂行において内的報酬システム 120 によって生成された内的報酬値に基づいて更新される。

ステップ 204 で、内的報酬システム 120 は、エージェント 105 によって実行されているタスクに基づいて取得された外的報酬値に基づいて更新される。ステップ 206 で、エージェントに関連付けられた満了基準が満たされると、エージェントのポリシー110 が再初期化される。

エージェントのポリシーの再初期化には、新しいエージェントの生成が含まれる場合がある。この点で、新しいエージェントは、異なるポリシー更新方法を使用するエージェントである。例えば、第１のエージェントはアクター・クリティック技術を使用して訓練され、第２のエージェントはQ学習技術を使用して訓練される。

エージェントのポリシーが期限切れになり、再初期化される可能性がある一方で、内的報酬システム 120 は、エージェントのライフタイムにわたって持続し、エージェントの複数の世代から知識を蓄積する。

この蓄積された知識を通じて、内的報酬システム 120 は、エージェント 105 を訓練するために外的報酬のみを使用する従来の強化学習システムと比較して、特定のタスクの実行においてエージェント 105 の訓練を加速することができる。

3.クレーム

910 特許のクレーム 1 は以下の通りである。

1. ポリシーに基づいてアクションを実行するように構成されたエージェントと、エージェントによって取られたアクションに基づいてエージェントの内的報酬値を生成するように構成された内的報酬システムとを含む強化学習システムのトレーニング方法において、

複数のタスクに基づいて強化学習システムをトレーニングし、該トレーニングは、内的報酬システムによって生成された内的報酬値に基づいてエージェントのポリシーを更新することを含み、

エージェントによって実行されているタスクに基づいて取得された外的報酬値に基づいて、内的報酬システムを更新し、

エージェントに関連付けられた有効期限基準が満たされたときに、エージェントのポリシーを再初期化する。

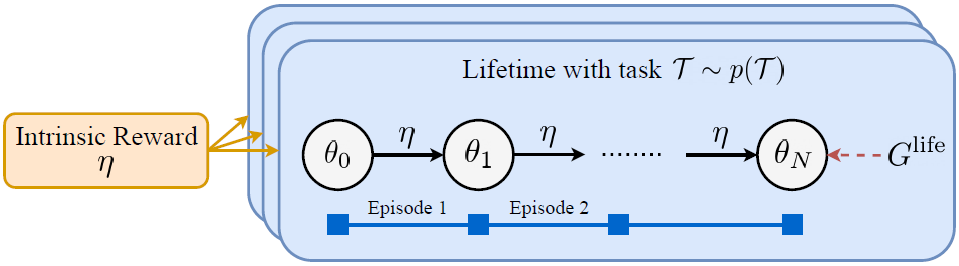
4. 本特許に関連する論文

本特許に関する論文“What Can Learned Intrinsic Rewards Capture?”[1]が、DeepMind の Zeyu Zheng 氏らにより公表されている。

強化学習エージェントの目的は、状態の適切なスカラー関数の合計である報酬を最大化するように動作することにあるところ、本論文では、報酬関数自体が学習された知識の適切な場所になり得るという命題を検討している。これを調査するために、経験の複数のライフタイムにわたって有用な固有の報酬関数を学習するためのスケーラブルな

[1] Zeyu Zheng, Junhyuk Oh, Matteo Hessel, Zhongwen Xu, Manuel Kroiss, Hado van Hasselt, David Silver, Satinder Singh “What Can Learned Intrinsic Rewards Capture?” arXiv:1912.05500v3 [cs.AI] 21 Aug 2020

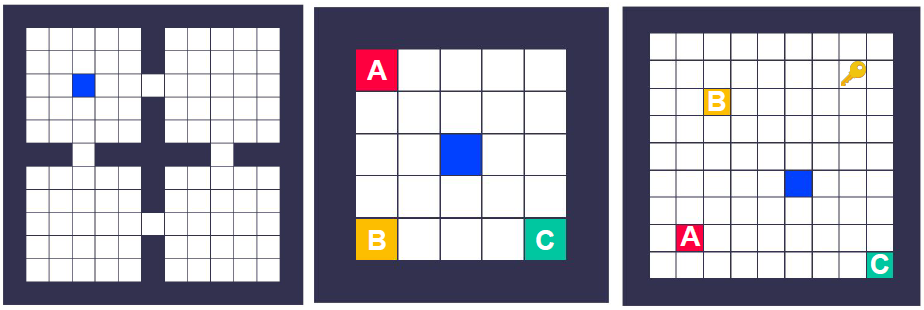
メタ勾配フレームワークを提案している。



上記図は、提案された内的報酬学習フレームワークである。内的報酬 r_η は、多くのエピソードで構成されるエージェントのライフタイムを通じて、エージェントのパラメータ θ_i を更新するために使用される。

目標は、ランダムに初期化されたエージェントと、分布 p(T) から引き出された非定常タスクが与えられた場合に、ライフタイムリターン (G^{life}) を最大化する、多くのライフタイムにわたる最適な内的報酬パラメータ η^* を見つけることである。

以下に示す概念実証実験を通じて、長期的な探査と搾取に関する知識を学習し、報酬関数に取り込むことが実現可能であることを示している。



(a) Empty Rooms　　(b) ABC　　(c) Key-Box

上記図は各タスクを示したものである

(a) Empty Rooms

エージェントは、正の報酬を与えるゴールの場所を見つける必要があるが、ゴールはエージェントには見えていない。

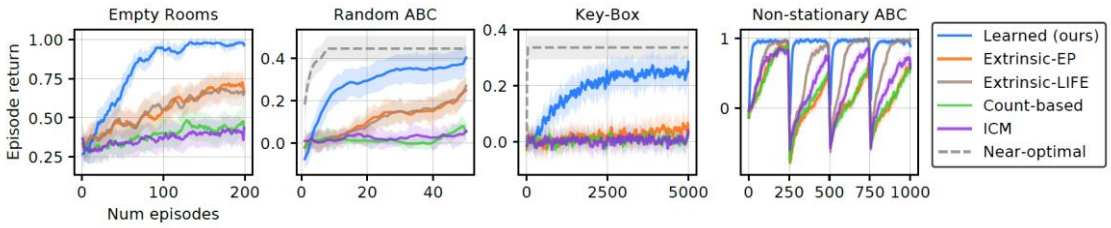
(b) ABC

各オブジェクト (A、B、および C) は報酬を与える。

(c)Key-Box

エージェントは、最初にキーを収集し、ボックス (A、B、および C) のいずれかにア

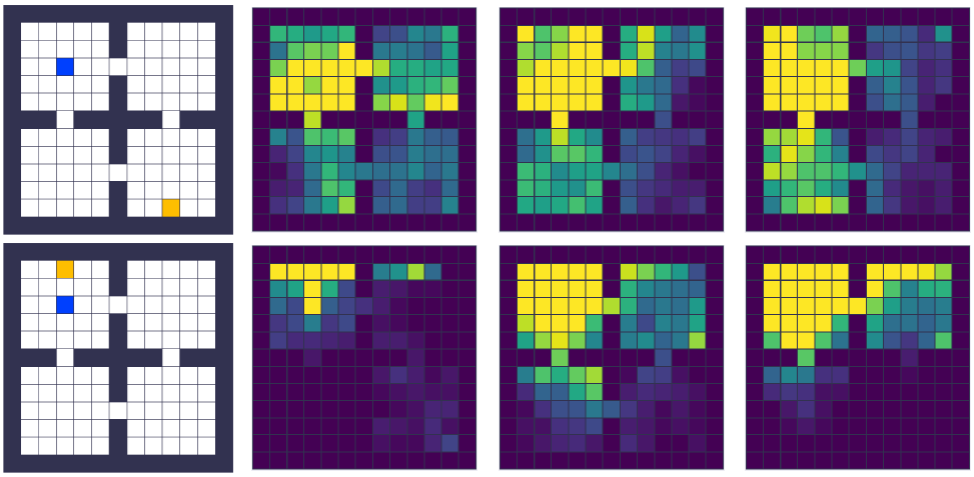
クセスして、対応する報酬を受け取る必要がある。すべてのオブジェクトは、各エピソードの前にランダムな場所に配置される。



上記図は、様々なタスクのエピソード数の関数としてのエピソードリターンのプロットを示す。

プロットは、青色の内的報酬システムを含む強化学習システム、オレンジ色のエピソードリターン目標を使用して外的報酬を使用してトレーニングされたエージェント、茶色のライフタイムリターン目標を使用して外的報酬を使用してトレーニングされたエージェント、緑色のカウントベースの探索報酬、および紫色の逆動力学モデル（ICM: inverse dynamics model）に基づく外的報酬と好奇心報酬を使用してトレーニングされたエージェントの比較を示す。破線は、手作業で設計された最適に近い探査戦略に対応する。

図から分かるように、本内的報酬システムを含む強化学習システムは、エピソード収益が増加し、またはるかに速い速度で機能する。したがって、本強化学習システムは、より最適なエージェントをトレーニングするために必要な計算リソースが少なくなる。



(a) Room (b) Intrinsic (c) Count (d) ICM

上記図は Empty Rooms でさまざまな報酬関数を使用してトレーニングされたエー

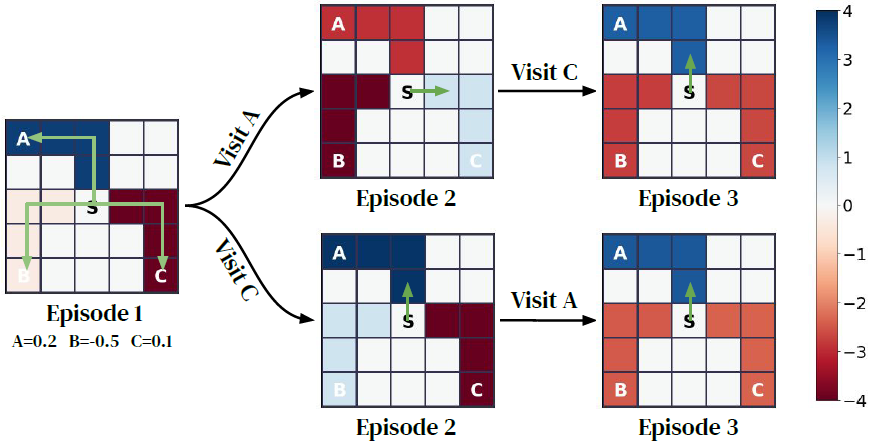
ジェントの最初の 3000 ステップを可視化したものである。

(a)青と黄色の四角は、それぞれエージェントと隠れたゴールを表している。
(b)学習した報酬は、目標が見つからない場合、エージェントが多くの場所を訪問することを奨励する（上)。 ただし、目標が早期に発見されると、内的報酬により、エージェントはそれ以上探索せずにそれを利用するようになる（下)。
(c-d) カウントベースと ICM の両方の報酬は、探索を促進する傾向があるが（上)、目標が見つかったときの搾取を妨げる（下)。

このタスクでは、エージェントの目標は、環境内の隠された目標を発見することである。目標が見つかったら、エージェントはこの知識をライフタイムにわたって活用する必要がある。つまり、(b)の上部の可視化から分かるように、内的報酬システムは、目標が見つからない場合、エージェントが多くの場所を訪れることを奨励する。(b)下部の視覚化では、エージェントが目標を早期に発見した場合、内的報酬システムは、１つの目標のみが存在することを内的報酬システムが学習したことを考えると、エージェントがそれ以上探索することなく利用することを奨励する。

比較すると、(c)および(d)に示されるカウントベースおよび ICM 方法は探索を促進するが、目標が見つかったときの利用を妨げる（下)。 このように、内的報酬システムは、いつ探索し、いつより効果的に活用するかを決定することができる。

(b) (d)



(a) (c) (e)

上記図は、「ランダム ABC」タスクの学習された内的報酬を視覚化したものである。

このタスクでは、3 つのオブジェクト A、B、および C に、ライフタイムの開始時に A[−1, 1,]、B[−0.5, 0]、および C[0, 0.5] の範囲からランダムに外的報酬が割り当てられ、ライフタイムの期間中固定される。

環境内のオブジェクト A の外的報酬は 0.2、環境内のオブジェクト B の外的報酬は -0.5、環境内のオブジェクト C の外的報酬は 0.1 である。最適な行動は、ライフタイムの初めに A と C を調べてどちらが優れているかを評価し、その後のすべてのエピソードでより優れた方にコミットすることである。内的報酬システムは、そのような挙動を示す。

より詳細には、上記図の各視覚化は、各オブジェクトへの軌道に対する内的報酬の合計を示している。図(a) は、内的報酬がエージェントに A の探索を促す最初のエピソードを示している。

第２のエピソードでは、Aが訪問された場合、内的報酬は、図(b)に示されるように、エージェントがCを探索することを奨励する。Cが訪問された場合、図(c)に示されるように、内的報酬は、エージェントがAを探索することを奨励する。

第３のエピソードでは、両方の場合において、図(d)および(e)に示されるように、エージェントは、最大の外的報酬を得るためにAを再訪するように促される。

したがって、図の視覚化は、内的報酬システムが、必要なタスク情報を取得するために探索し、情報が取得されたら利用する時期を決定できることをさらに示している。

以上

著者紹介
河野英仁
河野特許事務所、所長弁理士。立命館大学情報システム学博士前期課程修了、米国フランクリンピアースローセンター知的財産権法修士修了、中国清華大学法学院知的財産夏季セミナー修了、MIT(マサチューセッツ工科大学)コンピュータ科学・AI 研究所 AI コース修了。
AI 特許コンサルティング、医療 AI 特許コンサルティングの他、米国・中国特許の権利化・侵害訴訟を専門としている。著書に「世界のソフトウエア特許(共著)」、「FinTech 特許入門」、「AI/IoT 特許入門 3」、「ブロックチェーン 3.0(共著)」がある。