(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7581358号
(P7581358)

(45)発行日　令和6年11月12日(2024.11.12)

(24)登録日　令和6年11月1日(2024.11.1)

(51)国際特許分類

		FI	
G06N	*3/092(2023.01)*	G06N	3/092
G06N	*20/00 (2019.01)*	G06N	20/00

請求項の数　15　（全28頁）

(21)出願番号　特願2022-548005(P2022-548005)
(86)(22)出願日　令和3年2月8日(2021.2.8)
(65)公表番号　特表2023-512722(P2023-512722A)
(43)公表日　令和5年3月28日(2023.3.28)
(86)国際出願番号　PCT/EP2021/052988
(87)国際公開番号　WO2021/156518
(87)国際公開日　令和3年8月12日(2021.8.12)
審査請求日　令和4年10月5日(2022.10.5)
(31)優先権主張番号　62/971,890
(32)優先日　令和2年2月7日(2020.2.7)
(33)優先権主張国・地域又は機関
米国(US)

(73)特許権者　517030789
ディープマインド テクノロジーズ リミテッド
イギリス・ＥＣ４Ａ・３ＴＷ・ロンドン・ニュー・ストリート・スクエア・５
(74)代理人　100108453
弁理士　村山 靖彦
(74)代理人　100110364
弁理士　実広 信哉
(74)代理人　100133400
弁理士　阿部 達彦
(72)発明者　アドリア・プイドメネチ・バディア
イギリス・Ｎ１Ｃ・４ＡＧ・ロンドン・パンクラス・スクエア・６
(72)発明者　ビラル・ピオット

最終頁に続く

(54)【発明の名称】　適応リターン計算方式を用いた強化学習

(57)【特許請求の範囲】

【請求項１】

タスクのエピソードを実行するために環境と相互作用するエージェントを制御するために1つまたは複数のコンピュータにより実行される方法であって、

複数の異なるリターン計算方式の間で選択するためのポリシーを規定するデータを維持するステップであって、各リターン計算方式が、異なる重要度を、前記タスクの前記エピソードを実行しながら前記環境を探査することに対して割り当て、前記ポリシーは、前記リターン計算方式の各々にそれぞれの報酬スコアを割り当てる、ステップと、

前記ポリシーを使用して、前記複数の異なるリターン計算方式からリターン計算方式を選択するステップと、

前記選択されたリターン計算方式に従って計算されたリターンを最大化するように前記タスクの前記エピソードを実行するために前記エージェントを制御するステップと、

前記エージェントが前記タスクの前記エピソードを実行した結果として生成された報酬を識別するステップと、

前記識別された報酬を使用して、複数の異なるリターン計算方式の間で選択するために前記ポリシーを更新するステップと

を含む、方法。

【請求項２】

前記複数の異なるリターン計算方式が、少なくとも、リターンを生成するために報酬を組み合わせることにおいて使用されるそれぞれのディスカウントファクタをそれぞれ規定

する、請求項1に記載の方法。
【請求項３】
前記複数の異なるリターン計算方式が、リターンを生成するときに前記環境から受信された外発的報酬に対する内発的報酬の重要度を定義する少なくとも1つのそれぞれの内発的報酬スケーリングファクタをそれぞれ規定する、請求項1または2に記載の方法。
【請求項４】
前記選択されたリターン計算方式に従って計算されたリターンを最大化するように前記タスクの前記エピソードを実行するために前記エージェントを制御するステップが、以下の、
前記環境の現在の状態を特徴づける観測値を受信するステップと、
アクション選択出力を生成するために1つまたは複数のアクション選択ニューラルネットワークを使用して、前記観測値と前記選択されたリターン計算方式を規定するデータとを処理するステップと、
前記アクション選択出力を使用して前記エージェントによって実行されるアクションを選択するステップと
を反復して実行するステップを含む、請求項1から3のいずれか一項に記載の方法。
【請求項５】
前記環境は実世界の環境であり、各観測値は前記環境を感知するように構成された少なくとも1つのセンサの出力であり、前記エージェントは前記環境と相互作用する機械的エージェントである、請求項4に記載の方法。
【請求項６】
タスクのエピソードを実行するために環境と相互作用するエージェントを制御するために1つまたは複数のコンピュータにより実行される方法であって、
複数の異なるリターン計算方式の間で選択するためのポリシーを規定するデータを維持するステップであって、各リターン計算方式が、異なる重要度を、前記タスクの前記エピソードを実行しながら前記環境を探査することに対して割り当てる、ステップと、
前記ポリシーを使用して、前記複数の異なるリターン計算方式からリターン計算方式を選択するステップと、
前記選択されたリターン計算方式に従って計算されたリターンを最大化するように前記タスクの前記エピソードを実行するために前記エージェントを制御するステップであって、
前記環境の現在の状態を特徴づける観測値を受信するステップと、
アクション選択出力を生成するために1つまたは複数のアクション選択ニューラルネットワークを使用して、前記観測値と、前記選択されたリターン計算方式を規定するデータとを処理するステップと、
前記アクション選択出力を使用して前記エージェントによって実行されるアクションを選択するステップと
を反復して実行するステップを含む、制御するステップと、
前記エージェントが前記タスクの前記エピソードを実行した結果として生成された報酬を識別するステップと、
前記識別された報酬を使用して、複数の異なるリターン計算方式の間で選択するために前記ポリシーを更新するステップと
を含み、
前記1つまたは複数のアクション選択ニューラルネットワークが、
前記環境との相互作用の間に受信された観測値に基づいて内発的報酬システムによって生成された内発的報酬のみから計算された内発的リターンを推定する内発的報酬アクション選択ニューラルネットワークと、
前記環境との相互作用の結果として前記環境から受信された外発的報酬のみから計算された外発的リターンを推定する外発的報酬アクション選択ニューラルネットワークとを含む、方法。
【請求項７】

アクション選択出力を生成するために1つまたは複数のアクション選択ニューラルネットワークを使用して、前記観測値と前記選択されたリターン計算方式を規定するデータとを処理するステップが、アクションのセット内の各アクションに対して、
前記エージェントが前記観測値に応答して前記アクションを実行する場合に受信される推定された内発的リターンを生成するために、前記内発的報酬アクション選択ニューラルネットワークを使用して、前記観測値と、前記アクションと、前記選択されたリターン計算方式を規定する前記データと、を処理するステップと、
前記エージェントが前記観測値に応答して前記アクションを実行する場合に受信される推定された外発的リターンを生成するために、前記外発的報酬アクション選択ニューラルネットワークを使用して、前記観測値と、前記アクションと、前記選択されたリターン計算方式を規定する前記データと、を処理するステップと、
前記推定された内発的報酬および前記推定された外発的報酬から最終リターン推定を決定するステップとを含む、請求項6に記載の方法。
【請求項８】
前記アクション選択出力を使用して前記エージェントによって実行されるアクションを選択するステップが、
前記アクションのセット内の各アクションに対して決定した最終リターン推定のうち、最高の最終リターン推定に対応するアクションを確率1-εで選択し、確率εでアクションの前記セットからランダムアクションを選択するステップとを含む、請求項7に記載の方法。
【請求項９】
前記2つのアクション選択ニューラルネットワークは、同じアーキテクチャを有するが異なるパラメータ値を有する、請求項6から8のいずれか一項に記載の方法。
【請求項１０】
前記タスクのエピソードの実行から訓練データを生成するステップと、
強化学習を通して前記訓練データ上で前記1つまたは複数のアクション選択ニューラルネットワークを訓練するステップとをさらに含む、請求項6から9のいずれか一項に記載の方法。
【請求項１１】
前記訓練データ上で前記1つまたは複数のアクション選択ニューラルネットワークを訓練するステップが、
前記タスクのエピソードの前記実行の結果として生成された内発的報酬のみを使用して前記内発的報酬アクション選択ニューラルネットワークを訓練するステップと、
前記タスクのエピソードの前記実行の間に受信された外発的報酬のみを使用して前記外発的報酬アクション選択ニューラルネットワークを訓練するステップとを含む、請求項10に記載の方法。
【請求項１２】
タスクのエピソードを実行するために環境と相互作用するエージェントを制御するために1つまたは複数のコンピュータにより実行される方法であって、
複数の異なるリターン計算方式の間で選択するためのポリシーを規定するデータを維持するステップであって、各リターン計算方式が、異なる重要度を、前記タスクの前記エピソードを実行しながら前記環境を探査することに対して割り当てる、ステップと、
前記ポリシーを使用して、前記複数の異なるリターン計算方式からリターン計算方式を選択するステップと、
前記選択されたリターン計算方式に従って計算されたリターンを最大化するように前記タスクの前記エピソードを実行するために前記エージェントを制御するステップと、
前記エージェントが前記タスクの前記エピソードを実行した結果として生成された報酬を識別するステップと、
前記識別された報酬を使用して、複数の異なるリターン計算方式の間で選択するために前記ポリシーを更新するステップであって、前記ポリシーは、前記リターン計算方式の各

々に対応するそれぞれの腕を有する非定常多腕バンディットアルゴリズムを使用して更新される、ステップと
を含む、方法。
【請求項１３】
前記識別された報酬を使用して、複数の異なるリターン計算方式の間で選択するために前記ポリシーを更新するステップが、
前記タスクのエピソードの実行の間に受信された外発的報酬からディスカウントされない外発的リターンを決定するステップと、
前記ディスカウントされない外発的報酬を前記非定常多腕バンディットアルゴリズムに対する報酬信号として使用することによって前記ポリシーを更新するステップとを含む、請求項12に記載の方法。
【請求項１４】
1つまたは複数のコンピュータと、命令を記憶する1つまたは複数の記憶デバイスとを含むシステムであって、前記命令が、前記1つまたは複数のコンピュータによって実行されたとき、請求項1から13のいずれか一項に記載の方法の動作を前記1つまたは複数のコンピュータに実行させる、システム。
【請求項１５】
命令を記憶する1つまたは複数の非一時的コンピュータ記憶媒体であって、前記命令が、1つまたは複数のコンピュータによって実行されたとき、請求項1から13のいずれか一項に記載の方法の動作を前記1つまたは複数のコンピュータに実行させる、1つまたは複数の非一時的コンピュータ記憶媒体。
【発明の詳細な説明】
【技術分野】
【０００１】
本明細書は、機械学習モデルを使用してデータを処理することに関する。
【背景技術】
【０００２】
機械学習モデルは、入力を受信し、受信された入力に基づいて出力、たとえば予測出力を生成する。いくつかの機械学習モデルは、パラメトリックモデルであり、受信された入力およびモデルのパラメータの値に基づいて出力を生成する。
【０００３】
いくつかの機械学習モデルは、受信された入力に対する出力を生成するために複数層のモデルを採用するディープモデルである。たとえば、ディープニューラルネットワークは、出力層と、出力を生成するために受信された入力に非線形変換をそれぞれ適用する1つまたは複数の隠れ層とを含むディープ機械学習モデルである。
【先行技術文献】
【非特許文献】
【０００４】
【文献】T. Schulら、「Prioritized experience replay」、arXiv:1511.05952v4 (2016年)
【文献】Y. Burdaら、「Exploration by random network distillation」、arXiv:1810.12894v1、(2018年)
【発明の概要】
【課題を解決するための手段】
【０００５】
本明細書は、環境と相互作用しているエージェントを制御するために、1つまたは複数のアクション選択ニューラルネットワークのセットを訓練するための、1つまたは複数のロケーションにおける1つまたは複数のコンピュータ上のコンピュータプログラムとして実装されるシステムを説明する。
【０００６】

本明細書で説明する主題の特定の実施形態は、以下の利点のうちの1つまたは複数を実現するために実装され得る。

【０００７】

本明細書で説明するシステムは、タスクを実行するために環境と相互作用しているエージェントを制御するために使用される1つまたは複数のアクション選択ニューラルネットワークのセットの訓練中に実行される任意の所与のタスクエピソードのために使用するための最適リターン計算方式を選択するためにシステムが使用するポリシーを維持する。各可能なリターン計算方式は、エージェントが環境と相互作用する間に、環境の探査に対して異なる重要度を割り当てる。特に、各可能なリターン補償方式は、リターンの計算に使用されるディスカウントファクタ、総報酬の計算に使用される内発的報酬スケーリングファクタ、または両方を規定する。より具体的には、システムは、適応機構を使用して訓練プロセスを通してポリシーを調整し、異なるリターン計算方式が、訓練中に異なるポイントにおいて選択される可能性がより大きくなることをもたらす。この適応機構を使用することで、システムが、訓練中の任意の所与の時間において、リターンを計算するために最も適切な計画対象期間、最も適切な探査の程度、または両方を効果的に選択することが可能になる。これは、様々なタスクのうちのいずれかを実行するためにエージェントを制御するとき、改善された性能を示すことができる訓練されたニューラルネットワークをもたらす。

【０００８】

加えて、いくつかの実装形態では、システムは、任意の所与のアクションの「値」を推定するための2つの別々のニューラルネットワーク、すなわち外発的報酬だけを使用して訓練される外発的リターンアクション選択ニューラルネットワークと、内発的報酬だけを使用して訓練される内発的リターンアクション選択ニューラルネットワークとを使用する。アーキテクチャのこの新規のパラメータ化は、より一貫して安定した学習を可能にし、訓練に必要な計算リソースの訓練の時間および量を低減し、ならびに得られた、訓練されたニューラルネットワークの性能を改善する。特に、これらの2つのニューラルネットワークは、同じ軌道上で(しかし、軌道と異なるタイプの報酬を使用して)訓練され得、それゆえ、データ効率の低下を伴うことなく、向上した一貫性と安定性とを達成することができる。

【０００９】

従来のシステムと比較すると、本明細書で説明するシステムは、アクション選択ニューラルネットワークを訓練することによってより少ない計算リソース(たとえば、メモリおよび計算能力)を消費して、より少ない訓練の反復に対して許容レベルの性能を達成し得る。その上、本明細書で説明するシステムによって訓練された1つまたは複数のアクション選択ニューラルネットワークのセットは、エージェントが、代替システムによって訓練されたアクション選択ニューラルネットワークより効果的に(たとえば、より速やかに)タスクを完遂することを可能にするアクションを選択することができる。

【００１０】

本明細書の主題の1つまたは複数の実施形態の詳細が、添付の図面および以下の説明において説明される。主題の他の特徴、態様、および利点は、説明、図面および特許請求の範囲から明白となろう。

【図面の簡単な説明】

【００１１】

【図１】例示的なアクション選択システムを示す図である。

【図２】例示的な訓練システムを示す図である。

【図３】例示的な内発的報酬システムを示す図である。

【図４】タスクエピソードを実行するためにエージェントを制御するための、およびリターン計算方式選択ポリシーを更新するための例示的なプロセスの流れ図である。

【発明を実施するための形態】

【００１２】

様々な図における同様の参照番号および記号は、同様の要素を示す。

【００１３】

図1は、例示的なアクション選択システム100を示す。アクション選択システム100は、以下で説明するシステム、構成要素および技法が実装される1つまたは複数のロケーションにおける、1つまたは複数のコンピュータ上にコンピュータプログラムとして実装されるシステムの一例である。

【００１４】

アクション選択システム100は、1つまたは複数のアクション選択ニューラルネットワーク102およびポリシーデータ120を使用して、タスクのエピソードの実行中に複数の時間ステップの各々においてエージェント104によって実行されるアクション108を選択することによって、タスクを完遂するために環境106と相互作用するエージェント104を制御する。

【００１５】

タスクの「エピソード」は反復のシーケンスであり、その間に、エージェントは、環境のいくつかの開始状態から開始するタスクのインスタンスを実行することを試みる。言い換えれば、各タスクエピソードは、初期状態、たとえば固定された初期状態またはランダムに選択された初期状態にある環境で開始し、エージェントがタスクを完了することに成功したとき、またはいくつかの終了基準が満たされたとき、たとえば環境が終端状態として指定された状態に入るか、もしくはエージェントがタスクの完了に成功せずにしきい値の数のアクションを実行したときに終了する。

【００１６】

任意の所与のタスクエピソードの間の各時間ステップにおいて、システム100は、時間ステップにおける環境106の現在の状態を特徴づける観測値110を受信し、それに応じて、その時間ステップにおいてエージェント104によって実行されるアクション108を選択する。エージェントがアクション108を実行した後、環境106は新しい状態に移行し、システム100は環境106から外発的報酬130を受信する。

【００１７】

一般に、外発的報酬130は、スカラーの数値であり、タスクの完了に向けたエージェント104の進歩を特徴づける。

【００１８】

特定の例として、外発的報酬130は、アクションが実行された結果として、タスクが完了に成功しなかった場合にゼロであり、タスクが完了に成功した場合に1である、スパースな2値報酬であり得る。

【００１９】

別の特定の例として、外発的報酬130は、タスクの実行を試みるエピソードの間に受信された個々の観測値に基づいて決定されたときに、タスクの完了に向けたエージェントの進歩を測定する密な報酬であり得、すなわち、それにより、タスクが完了に成功する前に非ゼロ報酬が頻繁に受信され得る。

【００２０】

任意の所与のタスクエピソードを実行する間、システム100は、タスクエピソードにわたって受信されたリターンを最大化することを試みるためにアクションを選択する。

【００２１】

すなわち、エピソードの間の各時間ステップにおいて、システム100は、時間ステップから開始するタスクエピソードの残りに対して受信されるリターンを最大化することを試みるアクションを選択する。

【００２２】

一般に、任意の所与の時間ステップにおいて、受信されるリターンは、エピソード内の所与の時間ステップの後の時間ステップにおいて受信される報酬の組合せである。

【００２３】

たとえば、時間ステップtにおいて、リターンは、

$\Sigma_i \gamma^{i-t-1} r_i$、

を満足する。ここでiは、エピソード内のtの後の時間ステップのすべてにわたって、またはエピソード内のtの後の時間ステップのいくつかの固定数に対して変動し、γはディスカウントファクタであり、r_iは時間ステップiにおける総報酬である。上式から分かるように、ディスカウントファクタの値が高いほど、リターン計算に対してより長い計画対象期間をもたらし、すなわち、時間ステップtから時間的により離れた時間ステップからの報酬が、リターン計算においてより大きい重みを与えられることをもたらす。

【００２４】

いくつかの実装形態では、所与の時間ステップに対する総報酬は、時間ステップにおいて受信された、すなわち前の時間ステップにおいて実行されたアクションの結果として受信された、外発的報酬に等しい。

【００２５】

いくつかの他の実装形態では、システム100はまた、少なくとも時間ステップにおいて受信された観測値から内発的報酬132を取得、すなわち受信または生成する。内発的報酬132は、実行されるタスクとは無関係の方法で、すなわち外発的報酬130のように、時間ステップの時点でタスクの完了に向けたエージェントの進歩を特徴づけるのではなく、時間ステップの時点で環境の探査に向けたエージェントの進歩を特徴づける。すなわち、内発的報酬132は、タスクに対するエージェントの成績を測定するのではなく、環境の探査を測定する。たとえば、内発的報酬は、全環境および/または環境内の物体の可能な構成についての情報を観測値が提供する程度を示す値であり得、たとえば、環境が実世界の環境であり、観測値が環境の対応部分に関する画像(または他のセンサデータ)である場合、内発的報酬は、どれほど多くの環境が画像のうちの少なくとも1つの中に現れたかを示す値であり得る。時間ステップに対する内発的報酬の計算について、図2および図3を参照して以下でより詳細に説明する。

【００２６】

これらの実装形態では、システム100は、少なくとも(i)時間ステップに対する外発的報酬、(ii)時間ステップに対する内発的報酬、および(iii)内発的報酬スケーリングファクタに基づいて、時間ステップにおいてエージェントによって受信された「総」報酬を決定することができる。

【００２７】

特定の例として、システム100は、時間ステップtに対して総報酬γ_tを、たとえば、

【数１】

$$r_t = r_t^{task} + \beta \cdot r_t^{exploration}$$

として生成することができ、ここで、

【数２】

$$r_t^{task}$$

は時間ステップに対する外発的報酬を示し、

【数３】

$$r_t^{exploration}$$

は時間ステップに対する内発的報酬を示し、βは内発的報酬スケーリングファクタを示す

。内発的報酬スケーリングファクタの値は、たとえば内発的報酬スケーリングファクタの値が高いほど総報酬に対する内発的報酬の貢献度が高くなるように、総報酬に対する外発的報酬および内発的報酬の相対的重要度を制御することが諒解され得る。外発的報酬、内発的報酬、および内発的報酬スケーリングファクタから総報酬を決定するための方法であって、内発的報酬スケーリングファクタの値が総報酬に対する外発的報酬および内発的報酬の相対的重要度を制御する、他の方法が可能であり、上式は、単に例示的な目的で提供される。

【００２８】

ポリシーデータ120は、リターン計算方式のセットからの複数の異なるリターン計算方式の間で選択するためのポリシーを規定するデータである。

【００２９】

セット内の各リターン計算方式は、タスクのエピソードを実行しながら環境を探査することに対して異なる重要度を割り当てる。言い換えれば、セット内のいくつかのリターン計算方式は、環境を探査すること、すなわち環境について新しい情報を収集することに対してより大きい重要度を割り当てる一方で、セット内の他のリターン計算方式は、環境を利用すること、すなわち環境についての現在の知識を利用することに対してより大きい重要度を割り当てる。

【００３０】

特定の一例として、各リターン計算方式は、リターンを生成するために報酬を組み合わせることにおいて使用される少なくとも1つのそれぞれのディスカウントファクタγを規定することができる。言い換えれば、いくつかのリターン計算方式が、比較的より大きいディスカウントファクタを規定することができ、すなわち今後の時間ステップにおいて報酬をもたらすディスカウントファクタが、現在の時間ステップに対するリターン計算において、他のリターン計算方式より比較的より重く重みづけられる。

【００３１】

別の特定の例として、各リターン計算方式は、リターンを生成するときに環境から受信された外発的報酬に対する内発的報酬の重要度を定義する少なくとも1つのそれぞれの内発的報酬スケーリングファクタβを規定することができる。

【００３２】

特に、上記で説明したように、内発的報酬スケーリングファクタは、所与の時間ステップに対する内発的報酬が、時間ステップに対する総報酬を生成するために所与の時間ステップに対する外発的報酬に加えられる前に、どの程度にスケーリングされるかを定義する。言い換えれば、いくつかのリターン計算方式は、比較的大きい内発的報酬スケーリングファクタ、すなわち時間ステップにおける内発的報酬が、時間ステップにおける総リターンの計算において、他のリターン計算方式よりも比較的大きい重みを割り当てられることをもたらすスケーリングファクタを規定することができる。

【００３３】

別の特定の例として、各リターン計算方式は、それぞれのディスカウントファクタ-内発的報酬スケーリングファクタペア、すなわちγ-βペアを規定することができ、それにより、セット内の各方式は、方式内のそれぞれの他のセットからのディスカウントファクタおよびスケーリングファクタに対する値の異なる組合せを規定する。

【００３４】

エピソードを実行する前に、システム100は、ポリシーデータ120によって現在規定されているポリシーを使用してセット内の複数の異なる方式からリターン計算方式を選択する。たとえば、システム100は、ポリシーによって方式に割り当てられた報酬スコアに基づいて方式を選択することができる。方式を選択することについて、図4を参考して以下でより詳細に説明する。以下でより詳細に説明するように、ポリシーは、異なる方式が訓練中に異なる時間に選択される可能性がより高くなるように、訓練中に適応的に修正される。

【００３５】

次いで、システム100は、選択された方式に従ってタスクエピソードを実行するように、すなわち、1つまたは複数のアクション選択ニューラルネットワーク102を使用して、選択された方式を使用して計算されたリターンを最大化するように、エージェント104を制御する。

【００３６】

そうするために、エピソード内の各時間ステップにおいて、システム100は、アクション選択ニューラルネットワーク102を使用して、(i)時間ステップにおいて環境の現在の状態を特徴づける観測値110と、(ii)アクションスコア114を生成するために選択されたリターン計算方式112を規定する選択された方式データとを含む入力を処理する。選択された方式データは、任意の適切な方法において、たとえば、学習済み埋め込みとして、または選択された方式に対応する位置において「1」の値を有し、セット内のすべての他の方式に対応する位置において「0」の値を有するワンホットベクトルとして、選択される方式112を規定することができる。

【００３７】

アクションスコア114(「リターン推定」とも呼ばれる)は、可能なアクションのセット内の各アクションに対するそれぞれの数値を含むことができ、時間ステップにおいてエージェント104によって実行されるアクション108を選択するためにシステム100によって使用される。

【００３８】

アクション選択ニューラルネットワーク102は、セット内の可能なリターン計算方式によって索引を付けられたアクション選択ポリシーのファミリーを実装することとして理解され得る。特に、訓練システム200(図2を参照してより詳細に説明する)は、選択されたリターン計算方式が、対応するアクション選択ポリシーが「探査的である」度合を特徴づけるように、すなわち、エージェントに環境を探査させるアクションを選択するように、アクション選択ニューラルネットワーク102を訓練することができる。言い換えれば、訓練システム200は、選択された方式上のニューラルネットワークを調整することが、どの方式が選択されたかに応じて環境の探査対環境の利用に対して、より多いまたはより少ない重点を置くアクション選択ポリシーを定義する出力をネットワークに生成させるように、アクション選択ニューラルネットワーク102を訓練する。

【００３９】

いくつかの実装形態では、1つまたは複数のアクション選択ニューラルネットワーク102は、観測値110と、選択されたリターン計算方式112を規定するデータとを含む入力を受信し、アクションスコア114を出力として生成する単一のニューラルネットワークである。すなわち、これらの実装形態では、アクションスコア114は、単一のアクション選択ニューラルネットワーク102の出力である。

【００４０】

アクション選択ニューラルネットワーク102は、その説明された機能の実行を可能にする、任意の適切なニューラルネットワークアーキテクチャを用いて実装され得る。一例では、アクション選択ニューラルネットワーク102は、「埋め込み」サブネットワークと、「コア」サブネットワークと、「選択」サブネットワークとを含み得る。ニューラルネットワークのサブネットワークは、ニューラルネットワーク内の1つまたは複数のニューラルネットワーク層のグループを指す。観測値が画像であるとき、埋め込みサブネットワークは、畳み込みサブネットワークであり得、すなわち、時間ステップに対する観測値を処理するように構成された1つまたは複数の畳み込みニューラルネットワーク層を含む。観測値がより低次元のデータであるとき、埋め込みサブネットワークは、完全に接続されたサブネットワークであり得る。コアサブネットワークは、たとえば、1つまたは複数の長・短期記憶(LSTM:long short-term memory)ニューラルネットワーク層を含む回帰サブネットワークであり得、LSTMニューラルネットワーク層は、(i)埋め込みサブネットワークの出力、(ii)選択されたリターン計算方式を表す選択された方式データ、ならびに随意に(iii)直近に受信された外発的報酬(および随意に内発的報酬)および/または直近に実行され

たアクションを規定するデータを処理するように構成される。選択サブネットワークは、アクションスコア114を生成するためにコアサブネットワークの出力を処理するように構成され得る。

【００４１】

いくつかの他の実装形態では、1つまたは複数のアクション選択ニューラルネットワーク102は、2つの別々のニューラルネットワーク、すなわち(i)環境との相互作用の間に受信された観測値に基づいて内発的報酬システムによって生成された内発的報酬のみから計算された内発的リターンを推定する内発的報酬アクション選択ニューラルネットワーク、および(ii)環境との相互作用の結果として環境から受信された外発的報酬のみから計算された外発的リターンを推定する外発的報酬アクション選択ニューラルネットワークである。

【００４２】

これらの実装形態では、2つの別々のニューラルネットワークは同じアーキテクチャを有することができるが、異なる量を推定するように訓練された結果として、すなわち、内発的報酬アクション選択ニューラルネットワークが内発的リターンを推定するように訓練されかつ外発的報酬アクション選択ニューラルネットワークが外発的リターンを推定するように訓練された結果として、異なるパラメータ値を有することができる。

【００４３】

これらの実装形態では、システム100は、各アクションに対してそれぞれの内発的アクションスコア(「推定された内発的リターン」)を生成するために内発的報酬アクション選択ニューラルネットワークを使用して入力を処理し、各アクションに対してそれぞれの外発的アクションスコア(「推定された外発的リターン」)を生成するために外発的報酬アクション選択ニューラルネットワークを使用して入力を処理する。

【００４４】

次いで、システム100は、アクションに対する最終アクションスコア(「最終リターン推定」)を生成するために内発的報酬スケーリングファクタに従って各アクションに対する内発的アクションスコアと外発的アクションスコアとを組み合わせることができる。

【００４５】

一例として、j番目の方式が選択されたと仮定すると、観測値xに応答するアクションaに対する最終アクションスコア$Q(x, a, j; \theta)$は、

$$Q(x, a, j; \theta)=Q(x, a, j; \theta^{e})+\beta_j \cdot Q(x, a, j; \theta^{i})$$

を満たすことができ、ここで$Q(x, a, j; \theta^{e})$はアクションaに対する外発的アクションスコアであり、$Q(x, a, j; \theta^{i})$はアクションaに対する内発的アクションスコアであり、β_jはj番目の方式内のスケーリングファクタである。

【００４６】

別の例として、最終アクションスコア$Q(x, a, j; \theta)$は、

$$Q(x, a, j; \theta)=h(h^{-1}(Q(x, a, j; \theta^{e}))+\beta_j h^{-1}(Q(x, a, j; \theta^{i})))$$

を満たすことができ、ここでhは、ニューラルネットワークのために概算することを容易にするために、状態-アクション値関数、すなわち外発的および内発的報酬関数をスケーリングする単調増加の反転可能なスカッシング関数である。

【００４７】

したがって、リターン方式におけるβ_jの異なる値は、最終アクションスコアを計算するときに、内発的アクション選択ニューラルネットワークの予測が異なって重みを加えられることを生じさせる。

【００４８】

システム100は、時間ステップにおいてエージェント104によって実行されるアクション108を選択するためにアクションスコア114を使用することができる。たとえば、システム100は、可能なアクションのセットにわたる確率分布を生成するためにアクションスコア114を処理し、次いで、確率分布に従ってアクションをサンプリングすることによって、エージェント104によって実行されるアクション108を選択し得る。システム100は、たとえば、ソフトマックス関数(soft-max function)を使用してアクションスコア114

を処理することによって、可能なアクションのセットにわたって確率分布を生成することができる。別の例として、システム100は、最高のアクションスコア114に関連付けられた可能なアクションとしてエージェント104によって実行されるアクション108を選択し得る。随意に、システム100は、探査ポリシー、たとえば、システム100が確率1-εの最高の最終リターン推定を有するアクションを選択し、確率εのアクションのセットからランダムアクションを選択するε-greedy探査ポリシーを使用して時間ステップにおいてエージェント104によって実行されるアクション108を選択し得る。

【００４９】

タスクエピソードが実行されると、すなわち、エージェントがタスクの実行に成功すると、またはタスクエピソードに対するいくつかの終了基準が満たされると、システム100は、(i)ポリシーデータ120によって現在規定されているリターン計算方式の間で選択するためのポリシーを更新するために、(ii)アクション選択ニューラルネットワーク102を訓練するために、または両方のために、タスクエピソードの結果を使用することができる。

【００５０】

より一般的には、ポリシーデータ120によって規定されたポリシーと、アクション選択ニューラルネットワーク102のパラメータ値の両方が、エージェント104の環境との相互作用の結果として生成された軌道に基づいて訓練中に更新される。

【００５１】

特に、訓練中、システム100は、最近生成された軌道、たとえば直近に生成された軌道の固定サイズのスライディングウィンドウ内で生成された軌道を使用してリターン方式選択ポリシーを更新する。この適応機構を使用して訓練プロセスを通してリターン計算方式選択ポリシーを調整することによって、異なるリターン計算方式が、アクション選択ニューラルネットワークの訓練中に異なるポイントにおいて選択される可能性がより高い。この適応機構を使用することで、システム100が、訓練中の任意の所与の時間において、最も適切な計画対象期間、最も適切な探査の程度、または両方を効果的に選択することが可能になる。これは、様々なタスクのうちのいずれかを実行するためにエージェントを制御するとき、改善された性能を示すことができる訓練されたニューラルネットワークをもたらす。

【００５２】

リターン方式選択ポリシーを更新することについて、図4を参照して以下でより詳細に説明する。

【００５３】

システム100は、システム100によって生成され、リプレーバッファと呼ばれるデータストアに追加される軌道を使用してアクション選択ニューラルネットワーク102を訓練する。言い換えれば、訓練中に規定された間隔において、システム100は、リプレーバッファから軌道をサンプリングし、この軌道を軌道に使用してニューラルネットワーク102を訓練する。いくつかの実装形態では、リターン方式選択ポリシーを更新するために使用される軌道は、同じく、ニューラルネットワーク102の訓練に後で使用するためにリプレーバッファに追加される。他の実装形態では、リターン方式選択ポリシーを更新するために使用される軌道は、リプレーバッファに追加されず、ポリシーを更新するためにだけ使用される。たとえば、システム100は、ポリシーを更新するために使用されるタスクエピソードを実行することと、リプレーバッファに追加されるタスクエピソードを実行することとの間を交互に入れ替わり得る。

【００５４】

アクション選択ニューラルネットワークを訓練することについて、図2を参照して以下で説明する。

【００５５】

いくつかの実装形態では、環境は実世界の環境であり、エージェントは、たとえば、(環境の中の並進および/もしくは回転によって、ならびに/またはその構成を変更することによって)実世界の環境の中を移動して、および/もしくは実世界の環境を修正して、実世界

の環境と相互作用する機械的エージェントである。たとえば、エージェントは、たとえば、関心のあるオブジェクトを環境の中に位置付けること、関心のあるオブジェクトを環境の中の特定のロケーションに移動させること、関心のあるオブジェクトを環境の中で物理的に操作すること、もしくは環境の中の特定の目的地にナビゲートすることを行うために環境と相互作用するロボットであり得るか、またはエージェントは、環境内の指定された目的地に向けて環境を通り抜ける自律的もしくは半自律的な、地上、空中もしくは海上(海中)の車両であり得る。

【００５６】

これらの実装形態では、観測値は、たとえば、画像、オブジェクト位置データ、およびエージェントが環境と相互作用するときに観測値を捕捉するためのセンサデータ、たとえば画像、距離もしくは位置のセンサからのまたはアクチュエータからのセンサデータのうちの1つまたは複数を含み得る。

【００５７】

たとえば、ロボットの場合、観測値は、ロボットの現在の状態を特徴づけるデータ、たとえば、ジョイント位置、ジョイント速度、ジョイントの力、トルクまたは加速度、たとえば重力補償トルクフィードバック、およびロボットによって把持される品目の大域的または相対的姿勢(global or relative pose)のうちの1つまたは複数を含み得る。

【００５８】

ロボットまたは他の機械的エージェントもしくは車両の場合、観測値は、同様に、エージェントの1つまたは複数の部分の位置、線速度または角速度、力、トルクまたは加速度、および大域的または相対的姿勢のうちの1つまたは複数を含み得る。観測値は、1次元、2次元または3次元において規定され得、絶対的および/または相対的観測値であり得る。

【００５９】

観測値は、たとえば、実世界の環境を感知する1つもしくは複数のセンサデバイスによって取得されたデータ、たとえば、モータ電流もしくは温度信号などの感知された電気信号、および/または、たとえばカメラもしくはLIDARセンサからの画像もしくはビデオデータ、たとえばエージェントのセンサからのデータもしくは環境内のエージェントから離れて位置するセンサからのデータも含み得る。

【００６０】

電子的エージェントの場合、観測値は、電流、電圧、電力、温度などの工場またはサービス設備の部分をモニタする1つまたは複数のセンサからの、ならびに機器の電子的および/または機械的品目の機能を表す他のセンサおよび/または電子信号からのデータを含み得る。

【００６１】

アクションは、ロボットを制御するための制御入力、たとえばロボットのジョイントに対するトルクまたはより高いレベルの制御コマンド、あるいは自律的もしくは半自律的な地上もしくは空中もしくは海上(海中)の車両、たとえば車両の制御面もしくは他の制御要素に対するトルクまたはより高いレベルの制御コマンドであり得る。

【００６２】

言い換えれば、アクションは、たとえば、ロボットの1つもしくは複数のジョイントまたは別の機械的エージェントの部分に対する位置、速度または力/トルク/加速度のデータを含むことができる。アクションは、追加または代替として、モータ制御データなどの電子制御データ、またはより一般的には、環境内の1つもしくは複数の電子デバイスを制御するためのデータを含んでよく、それらの制御は、観測された環境の状態に影響を及ぼす。たとえば、自律的または半自律的な地上、空中または海上(海中)の車両の場合、アクションは、ナビゲーション、たとえば操舵、および運動、たとえば車両の制動および/または加速を制御するためのアクションを含み得る。

【００６３】

いくつかの実装形態では、環境はシミュレートされた環境であり、エージェントは、シミュレートされた環境と相互作用する1つまたは複数のコンピュータとして実装される。

【００６４】
たとえば、シミュレートされた環境は、ロボットまたは車両のシミュレーションであり得、アクション選択ネットワークは、シミュレーション上で訓練され得る。たとえば、シミュレートされた環境は、モーションシミュレーション環境、たとえばドライビングシミュレーションまたはフライトシミュレーションであり得、エージェントはモーションシミュレーションを介してナビゲートするシミュレートされた車両であり得る。これらの実装形態では、アクションは、シミュレートされたユーザまたはシミュレートされた車両を制御するための制御入力であり得る。
【００６５】
別の例では、シミュレートされた環境はビデオゲームであり得、エージェントは、ビデオゲームをプレイするシミュレートされたユーザであり得る。
【００６６】
さらなる例では、環境は、各状態がタンパク質鎖のそれぞれの状態であり、エージェントは、タンパク質鎖をいかにして折り畳むかを決定するためのコンピュータシステムであるような、タンパク質フォールディング環境であり得る。この例では、アクションは、タンパク質鎖を折り畳むための可能なフォールディングアクションであり、達成されるべき結果は、たとえば、タンパク質が安定しているように、およびタンパク質が特定の生物学的機能を達成するようにタンパク質を折り畳むことを含み得る。別の例として、エージェントは、人による相互作用なしに、システムによって自動的に選択されたタンパク質フォールディングアクションを実行または制御する機械的エージェントであり得る。観測値は、タンパク質の状態の直接的もしくは間接的観測値を含んでよく、および/またはシミュレーションから導出され得る。
【００６７】
一般に、シミュレートされた環境の場合、観測値は、前に説明した観測値または観測値のタイプのうちの1つまたは複数のシミュレートされたバージョンを含んでよく、アクションは、前に説明したアクションまたはアクションのタイプのうちの1つまたは複数のシミュレートされたバージョンを含んでよい。
【００６８】
シミュレートされた環境内のエージェントを訓練することは、エージェントが大量のシミュレートされた訓練データから学習しながら、実世界の環境の中でエージェントを訓練することに関連するリスク、たとえば不十分に選択されたアクションの実行に起因するエージェントへの損傷を回避することを可能にし得る。シミュレートされた環境内で訓練されるエージェントは、その後、実世界の環境の中に配置され得る。
【００６９】
いくつかの他のアプリケーションでは、エージェントは、たとえば、データセンター、または配管網から引き込んだ電力もしくは水の分配システムの中、あるいは製造工場もしくはサービス設備の中の機器の品目を含む、実世界の環境内のアクションを制御し得る。次いで、観測値は、工場または設備の動作に関連し得る。たとえば、観測値は、機器による電力もしくは水の使用量の観測値、または電力の生成もしくは分配の制御の観測値、またはリソースの使用量もしくは廃棄物産出の観測値を含み得る。エージェントは、たとえばリソース使用量を低減することによって効率を向上させるため、および/または、たとえば廃棄物を低減することによって環境内の動作の環境影響を低減するために、環境内のアクションを制御し得る。アクションは、工場/設備の機器の品目に対する動作条件を制御するかもしくは課するアクション、および/または、たとえば工場/設備の構成要素を調整もしくはオン/オフするために工場/設備の動作における設定に変更をもたらすアクションを含み得る。
【００７０】
随意に、上記の実装形態のいずれかにおいて、任意の所与の時間ステップにおける観測値は、環境を特徴づけるのに有益であり得る、前の時間ステップからのデータ、たとえば、前の時間ステップにおいて実行されたアクション、前の時間ステップにおいて実行され

たことに応答して受信された報酬などを含み得る。

【００７１】

訓練システム200は、各時間ステップにおいてエージェント104によって受信された報酬を決定することができ、エージェントによって受信された報酬の累積測度(cumulative measure)を最適化するために強化学習技法を使用してアクション選択ニューラルネットワーク102を訓練し得る。エージェントによって受信された報酬は、たとえば、スカラーの数値によって表され得る。訓練システム200は、時間ステップにおいてアクション選択ニューラルネットワーク102によって処理された内発的報酬スケーリングファクタ112に少なくとも部分的に基づいて各時間ステップにおいてエージェントによって受信された報酬を決定することができる。特に、内発的報酬スケーリングファクタ112の値は、環境106の探査がエージェントによって受信された報酬に貢献する程度を決定することができる。このようにして、訓練システム200は、内発的報酬スケーリングファクタ112のより高い値に対して、アクション選択ニューラルネットワーク102が、エージェントに環境をより速やかに探査させるアクションを選択するように、アクション選択ニューラルネットワーク102を訓練し得る。

【００７２】

図2は、例示的な訓練システム200を示す。訓練システム200は、以下で説明するシステム、構成要素および技法が実装される1つまたは複数のロケーションにおける、1つまたは複数のコンピュータ上にコンピュータプログラムとして実装されるシステムの一例である。

【００７３】

訓練システム200は、アクション選択ニューラルネットワーク102を使用して選択されたアクションを実行することによってエージェントによって受信された総報酬の累積測度を最適化するために、(図1を参照して説明したように)アクション選択ニューラルネットワーク102を訓練するように構成される。

【００７４】

上記で説明したように、訓練システム200は、少なくとも(i)時間ステップに対する「外発的」報酬204、(ii)時間ステップに対する「探査」報酬206、および(iii)時間ステップが属するエピソードに対してサンプリングされたリターン計算方式210によって規定される内発的報酬スケーリングファクタに基づいて、時間ステップにおいてエージェントによって受信された「総」報酬202を決定することができる。

【００７５】

上記で説明したように、内発的報酬206は、時間ステップにおいて環境を探査することに向けたエージェントの進歩を特徴づけ得る。たとえば、訓練システム200は、(i)時間ステップにおける環境の状態を特徴づける観測値212の埋め込みと、(ii)それぞれの前の時間ステップにおける環境の状態を特徴づける1つまたは複数の前の観測値の埋め込みとの間の類似性測度に基づいて、時間ステップに対する内発的報酬206を決定することができる。特に、時間ステップにおける観測値の埋め込みと前の時間ステップにおける観測値の埋め込みとの間のより低い類似性は、エージェントが、前に見ていない環境の様相を探査しており、それゆえ、より高い内発的報酬206をもたらすことを示し得る。訓練システム200は、図3を参照してより詳細に説明する内発的報酬システム300を使用して、時間ステップにおける環境の状態を特徴づける観測値212を処理することによって、時間ステップに対する内発的報酬206を生成することができる。

【００７６】

アクション選択ニューラルネットワーク102を訓練するために、訓練システム200は、タスクエピソードの間に1つまたは複数の(連続する)時間ステップにわたってエージェントの環境との相互作用を特徴づける「軌道」を取得する。特に、軌道は、各時間ステップに対して、(i)時間ステップにおける環境の状態を特徴づける観測値212、(ii)時間ステップに対する内発的報酬202、および(iii)時間ステップに対する外発的報酬を規定し得る。軌道はまた、軌道に対応する、すなわちタスクエピソードの間にエージェントによって実行

されるアクションを選択するために使用された、リターン計算方式を規定する。
【００７７】
単一のアクション選択ニューラルネットワークが存在するとき、訓練エンジン208は、その後、時間ステップに対する内発的報酬から、時間ステップに対する外発的報酬から、および上記で説明した内発的報酬スケーリングファクタから、すなわち一定の内発的報酬スケーリングファクタかまたは異なるリターン計算方式が異なる内発的報酬スケーリングファクタを規定する場合は軌道に対応するリターン計算方式によって規定された内発的報酬スケーリングファクタのいずれかから、各時間ステップに対するそれぞれの総報酬を計算することによってアクション選択ニューラルネットワーク102を訓練することができる。
【００７８】
次いで、訓練エンジン208は、強化学習技法を使用して軌道上のアクション選択ニューラルネットワークを訓練することができる。強化学習技法は、たとえばQ学習技法、たとえばリトレースQ学習技法、または変換されたベルマン演算子を有するリトレースQ学習技法であり得、それにより、アクション選択ニューラルネットワークはQニューラルネットワークであり、アクションスコアは、対応するアクションがエージェントによって実行される場合に受信される、予期されるリターンを推定するQ値である。
【００７９】
一般的に、強化学習技法は、アクション選択ニューラルネットワークを訓練するための目標出力を計算するために、軌道内で今後の時間ステップにおいて受信される報酬に対するディスカウントファクタを使用すること、任意の所与の時間ステップの後に受信される今後のリターンを推定すること、または両方を行う。軌道上で訓練するとき、システム200は、軌道に対応するリターン計算方式内のディスカウントファクタを使用する。これは、アクション選択ニューラルネットワークが、異なるディスカウントファクタ上で調整されたときに、今後の報酬を異なって重みづけるアクションスコアを生成するように訓練されることをもたらす。
【００８０】
2つのアクション選択ニューラルネットワーク102が存在するとき、訓練エンジン208は、同じ軌道上で別々に、すなわち軌道内の時間ステップに対する総報酬を計算することなく、2つのアクション選択ニューラルネットワーク102を訓練することができる。より具体的には、訓練エンジン208は、強化学習技法を使用して、しかし軌道内の時間ステップに対する内発的報酬だけを使用して、内発的報酬アクション選択ニューラルネットワークを訓練することができる。訓練エンジン208はまた、強化学習技法を使用して、しかし軌道内の時間ステップに対する外発的報酬だけを使用して、外発的報酬アクション選択ニューラルネットワークを訓練することができる。
【００８１】
両ニューラルネットワークを訓練するために、訓練エンジン208は、訓練に対して同じ「目標」ポリシー、すなわち、対応するリターン計算方式210内のスケーリングファクタを使用して計算された総報酬を最大化するアクションを選択するポリシーを使用することができる。同様に、上記で説明したように、軌道上で訓練するとき、システム200は、軌道に対応するリターン計算方式におけるディスカウントファクタを使用する。これは、アクション選択ニューラルネットワークが、異なるディスカウントファクタ上で調整されたときに今後の報酬を異なって重みづけるアクションスコアを生成するように訓練されることをもたらす。
【００８２】
そうすることにおいて、システム200は、対応するアクションがエージェントによって実行される場合に受信される内発的リターンの推定である内発的アクションスコア(「推定された内発的リターン」)を生成するように内発的報酬アクション選択ニューラルネットワークを訓練しながら、対応するアクションがエージェントによって実行される場合に受信される外発的リターンの推定である外発的アクションスコア(「推定された外発的リターン

」)を生成するように外発的報酬アクション選択ニューラルネットワークを訓練する。

【００８３】

いくつかの実装形態では、訓練中、システム100は、複数のアクターコンピューティングユニット(actor computing unit)を使用してアクション選択ニューラルネットワーク102を訓練することにおいて訓練システム200によって使用するための軌道を生成することができる。これらの実装形態のいくつかでは、各アクターコンピューティングユニットは、複数の異なるリターン計算方式の間で選択するためのポリシーを維持しかつ別々に更新する。これは、異なるアクターコンピューティングユニットがε-greedy制御におけるεの異なる値を使用するか、または場合によっては、エージェントを異なって制御するときに有利であり得る。いくつかの他の実装形態では、システムは、アクターコンピューティングユニットのすべてに対して同じである中央ポリシーを維持することができる。

【００８４】

複数のアクターコンピューティングユニットが使用されるとき、各アクターコンピューティングユニットは、タスクエピソードを実行し、タスクエピソードの間のエージェントの相互作用の結果を使用してリターン計算方式選択ポリシー、すなわち中央ポリシーかまたはアクターコンピューティングユニットによって別々に維持されているポリシーのいずれかを更新するためにエージェントのインスタンスを制御することと、アクション選択ニューラルネットワークを訓練するための訓練データを生成することとを行うために、図1を参照して上記で説明した演算を反復して実行することができる。

【００８５】

コンピューティングユニットは、たとえば、コンピュータ、複数のコアを有するコンピュータ内のコア、または他のハードウェアもしくはソフトウェア、たとえば、演算を別々に実行することができるコンピュータ内の専用スレッドであり得る。コンピューティングユニットは、プロセッサコア、プロセッサ、マイクロプロセッサ、専用論理回路、たとえばFPGA(フィールドプログラマブルゲートアレー)もしくはASIC(特定用途向け集積回路)、または任意の他の適切なコンピューティングユニットを含み得る。いくつかの例では、コンピューティングユニットは、すべて同じタイプのコンピューティングユニットである。他の例では、コンピューティングユニットは、異なるタイプのコンピューティングユニットであり得る。たとえば、1つのコンピューティングユニットはCPUであり得る一方で、他のコンピューティングユニットはGPUであり得る。

【００８６】

訓練システム200は、リプレーバッファと呼ばれるデータストア内の各アクターコンピューティングユニットによって生成された軌道を記憶し、複数の訓練反復の各々において、アクション選択ニューラルネットワーク102の訓練に使用するためにリプレーバッファから一群の軌道をサンプリングする。訓練システム200は、たとえば、それぞれのスコアを各記憶された軌道に割り当て、そのスコアに関連付けられた軌道をサンプリングすることによって、優先順位をつけられた経験リプレーアルゴリズム(prioritized experience replay algorithm)に従ってリプレーバッファから軌道をサンプリングすることができる。例示的な優先順位をつけられた経験リプレーアルゴリズムは、T. Schulら、「Prioritized experience replay」、arXiv:1511.05952v4 (2016年)において記載されている。

【００８７】

セット内のリターン計算方式内に含まれる、可能な内発的報酬スケーリングファクタ

【数４】

$$\{\beta_i\}_{i=0}^{N-1}$$

(すなわち、ここでNは可能な内発的報酬スケーリングファクタの数である)のセットは、内発的報酬と無関係の総報酬を表す「ベースライン」内発的報酬スケーリングファクタ(実質的にゼロ)を含むことができる。

【００８８】
他の可能な内発的報酬スケーリングファクタは、それぞれの正数(一般的にすべてが他の各々と異なる)であり得、それぞれの「探査」アクション選択ポリシーをアクション選択ニューラルネットワークに実施させると見なされ得る。探査アクション選択ポリシーは、対応する内発的報酬スケーリングファクタによって規定される程度に、そのタスクを解決することだけでなく環境を探査することも行うようにエージェントを促す。
【００８９】
アクション選択ニューラルネットワーク102は、探査アクション選択ポリシーによって提供される情報を使用して、より効果的な利用アクション選択ポリシーを学習することができる。探査ポリシーによって提供される情報は、たとえば、アクション選択ニューラルネットワークの共有される重みの中に記憶される情報を含み得る。アクション選択ポリシーの範囲を一緒に学習することによって、訓練システム200は、アクション選択ニューラルネットワーク102が、各個々のアクション選択ポリシーをより効率的に、たとえば、より少ない訓練反復にわたって学習することを可能にし得る。その上、探査ポリシーを学習することは、外発的報酬がスパースである、たとえばまれに非ゼロである場合でも、システムがアクション選択ニューラルネットワークを継続的に訓練することを可能にする。
【００９０】
アクション選択ニューラルネットワークの訓練が完了した後、システム100は、方式選択ポリシーを更新して上記で説明したように方式を選択することを継続すること、または方式選択ポリシーを固定して、方式選択ポリシーが最高の報酬スコアを示す方式を、貪欲に選択することによってエージェントを制御することのいずれかを行うことができる。
【００９１】
一般に、システムは、内発的報酬としてタスクの進歩ではなく探査の進歩を特徴づける、任意の種類の報酬を使用することができる。内発的報酬および内発的報酬を生成するシステムの説明の特定の一例について、図3を参照して以下で説明する。
【００９２】
図3は、例示的な内発的報酬システム300を示す。内発的報酬システム300は、以下で説明するシステム、構成要素および技法が実装される1つまたは複数のロケーションにおける1つまたは複数のコンピュータ上にコンピュータプログラムとして実装されるシステムの一例である。
【００９３】
内発的報酬システム300は、環境を探査することにおいてエージェントの進歩を特徴づける内発的報酬206を生成するために環境の現在の状態を特徴づける現在の観測値212を処理するように構成される。システム300によって生成された内発的報酬206は、たとえば、図2を参照して説明した訓練システム200によって使用され得る。
【００９４】
システム300は、埋め込みニューラルネットワーク302と、外部メモリ304と、比較エンジン306とを含み、それらの各々について、次により詳細に説明する。
【００９５】
埋め込みニューラルネットワーク302は、「可制御性表現」308(または「埋め込まれた可制御性表現」)と呼ばれる現在の観測値の埋め込みを生成するために現在の観測値212を処理するように構成される。現在の観測値212の可制御性表現308は、数値の順序づけられた収集、たとえば、数値の配列として表され得る。埋め込みニューラルネットワーク302は、複数の層を有するニューラルネットワークとして実装され得、層のうちの1つまたは複数は、埋め込みニューラルネットワーク302の訓練中に修正される重みによって規定される関数を実行する。場合によっては、特に現在の観測値が少なくとも1つの画像の形であるとき、埋め込みニューラルネットワークの層のうちの1つまたは複数、たとえば少なくとも第1の層は、畳み込み層として実装され得る。
【００９６】
システム300は、エージェントによって制御可能である環境の状態の態様を特徴づける

観測値の可制御性表現を生成するために埋め込みニューラルネットワーク302を訓練し得る。環境の状態の一態様は、それがエージェントによって実行されたアクションによって(少なくとも部分的に)決定される場合、エージェントによって制御可能であると呼ばれ得る。たとえば、ロボットエージェントのアクチュエータによって把持される物体の位置は、エージェントによって制御可能であり得る一方で、周囲の照明条件または環境内の他のエージェントの動きは、エージェントによって制御可能ではない。埋め込みニューラルネットワーク302を訓練するための例示的な技法について、以下でより詳細に説明する。

【００９７】

外部メモリ304は、前の時間ステップにおける環境の状態を特徴づける前の観測値の可制御性表現を記憶する。

【００９８】

比較エンジン306は、現在の観測値212の可制御性表現308と外部メモリ304内に記憶されている前の観測値の可制御性表現とを比較することによって内発的報酬206を生成するように構成される。一般的に、比較エンジン306は、現在の観測値212の可制御性表現308が、外部メモリ内に記憶されている前の観測値の可制御性表現にあまり似ていない場合、より高い内発的報酬206を生成することができる。

【００９９】

たとえば、比較エンジン306は、内発的報酬γ_tを、

【数５】

$$r_t = \left(\sqrt{\sum_{f_i \in N_k} K(f(x_t), f_i)} + c \right)^{-1} \qquad (2)$$

として生成することができ、ここで

【数６】

$$N_k = \{f_i\}_{i=1}^{k}$$

は現在の観測値212の可制御性表現308に対する最高の類似性(たとえば、ユークリッド類似性測度(Euclidean similarity measure)による)を有する外部メモリ304内のk個の可制御性表現f_iのセットを示し(ここでkは一般的に1より大きい予測される正の整数値である)、$f(x_t)$はx_tで示される現在の観測値212の可制御性表現308を示し、K(・,・)は「カーネル」関数であり、cは数値安定性を促進するために使用される予測される一定値(たとえば、c=0.001)である。カーネル関数K(・,・)は、たとえば、

【数７】

$$K(f(x_t), f_i) = \frac{\epsilon}{\frac{d^2(f(x_t), f_i)}{d_m^2} + \epsilon} \qquad (3)$$

によって与えられ得、ここで$d(f(x_t), f_i)$は可制御性表現$f(x_t)$とf_iとの間のユークリッド距離を示し、εは数値安定性を促進するために使用される予測される一定値を示し、

【数８】

d_m^2

は(i)時間ステップにおける観測値の可制御性表現と、(ii)外部メモリからのk個の最も類似する可制御性表現のうちの可制御性表現との間のユークリッド距離の二乗の平均の移動平均(すなわち、固定された複数の時間ステップなど、複数の時間ステップにわたる)を示す。現在の観測値212の可制御性表現308が、外部メモリ内に記憶されている前の観測値の可制御性表現にあまり似ていない場合、より高い内発的報酬206をもたらす内発的報酬206を生成するための他の技法が可能であり、式(2)～(3)は、例示のためだけに提供される。

【０１００】

環境の状態の制御可能な態様を特徴づける可制御性表現に基づいて内発的報酬206を決定することは、アクション選択ニューラルネットワークのより効果的な訓練を可能にし得る。たとえば、環境の状態は、たとえば、照明および妨害物体の存在における変動を有する実世界の環境の場合、エージェントによって実行されるアクションとは無関係に変動する場合がある。特に、環境の現在の状態を特徴づける観測値は、エージェントが時間ステップの間に介在するアクションを実行していない場合でも、環境の前の状態を特徴づける観測値と実質的に異なる場合がある。それゆえ、環境の状態を特徴づける観測値を直接比較することによって決定された内発的報酬を最大化するように訓練されたエージェントは、たとえば、そのエージェントが、アクションを実行しないのに正の内発的報酬を受信する場合があるので、有意な環境の探査を実行しない場合がある。対照的に、システム300は、環境の制御可能な態様の有意な探査を達成することをエージェントに動機づける内発的報酬を生成する。

【０１０１】

現在の時間ステップに対して内発的報酬206を生成するために現在の観測値212の可制御性表現308を使用することに加えて、システム300は、外部メモリ304に現在の観測値212の可制御性表現308を記憶し得る。

【０１０２】

いくつかの実装形態では、外部メモリ304は、「一時的」メモリであり得、すなわち、それにより、システム300は、メモリリセット基準が満たされるたびに、外部メモリを(たとえば、その内容を消去することによって)「リセットする」。たとえば、システム300は、現在の時間ステップの前にメモリリセット基準が時間ステップの所定の数N〉1を最後に満たした場合、またはエージェントが現在の時間ステップにおけるそのタスクを完遂する場合、メモリリセット基準が現在の時間ステップにおいて満たされたと決定することができる。外部メモリ304が一時的メモリである実装形態では、比較エンジン306によって生成された内発的報酬206は、「一時的」内発的報酬と呼ばれることがある。一時的内発的報酬は、環境の状態を各可能な状態に反復的に遷移させるアクションを実行することによって、環境を継続的に探査するようにエージェントを促し得る。

【０１０３】

一時的内発的報酬を決定することに加えて、システム300はまた、「非一時的」内発的報酬、すなわち、最後に一時的メモリがリセットされて以来のそれらの時間ステップだけではなく、前の時間ステップごとの環境の状態に応じた内発的報酬、を決定し得る。非一時的内発的報酬は、たとえば、Y. Burdaら、「Exploration by random network distillation」、arXiv:1810.12894v1、(2018年)に関して記載されるランダムネットワーク蒸留(RND:random network distillation)報酬であり得る。非一時的内発的報酬は、エージェントが環境を探査しながら時間がたつにつれて減少する場合があり、環境のすべての可能な状態を反復して再訪するようにエージェントを促さない。

【０１０４】

随意に、システム300は、一時的報酬と非一時的報酬の両方に基づいて現在の時間ステップに対する内発的報酬206を生成することができる。たとえば、システム300は、時間

ステップに対する内発的報酬R_tを、
【数９】

$$R_t = r_t^{episodic} \cdot \min\{\max\{r_t^{non-episodic}, 1\}, L\} \qquad (4)$$

として生成することができ、ここで
【数１０】

$$r_t^{episodic}$$

はたとえば一時的外部メモリ304を使用して比較エンジン306によって生成された一時的報酬を示し、
【数１１】

$$r_t^{non-episodic}$$

は非一時的報酬、たとえばランダムネットワーク蒸留(RND)報酬を示し、ここで非一時的報酬の値は所定の範囲[1、L]にクリップされ、ここでL〉1である。
【０１０５】
埋め込みニューラルネットワーク302を訓練するためのいくつかの例示的な技法について、次により詳細に説明する。
【０１０６】
一例では、システム300は、アクション予測ニューラルネットワークを用いて埋め込みニューラルネットワーク302を一緒に訓練することができる。アクション予測ニューラルネットワークは、(i)第1の時間ステップにおいて環境の状態を特徴づける第1の観測値と、(ii)次の時間ステップにおいて環境の状態を特徴づける第2の観測値とのそれぞれの可制御性表現(埋め込みニューラルネットワークによって生成された)を含む入力を受信するように構成され得る。アクション予測ニューラルネットワークは、環境の状態を第1の観測値から第2の観測値に遷移させたエージェントによって実行されるアクションに対する予測を生成するために入力を処理し得る。システム300は、(i)アクション予測ニューラルネットワークによって生成された予測されたアクションと、(ii)エージェントによって実際に実行された「目標」アクションとの間の誤差を測定する目的関数を最適化するために、埋め込みニューラルネットワーク302とアクション予測ニューラルネットワークとを訓練し得る。特に、システム300は、目的関数の勾配を、アクション予測ニューラルネットワークを介して複数の訓練反復の各々において埋め込みニューラルネットワーク302に逆伝播し得る。目的関数は、たとえば、交差エントロピー目的関数であり得る。この方式で埋め込みニューラルネットワークを訓練することは、エージェントのアクションによって影響される環境についての情報、すなわち、環境の制御可能な態様を符号化するように可制御性表現を促す。
【０１０７】
別の例では、システム300は、状態予測ニューラルネットワークを用いて埋め込みニューラルネットワーク302を一緒に訓練することができる。状態予測ニューラルネットワークは、(i)時間ステップにおいて環境の状態を特徴づける観測値の可制御性表現(埋め込みニューラルネットワーク302によって生成される)と、(ii)時間ステップにおいてエージェ

ントによって実行されるアクションの表現とを含む入力を処理するように構成され得る。状態予測ニューラルネットワークは、次のステップにおいて、すなわちエージェントがアクションを実行した後に入力を処理して、予測される環境の状態を特徴づける出力を生成し得る。出力は、たとえば、次の時間ステップにおいて予測される環境の状態を特徴づける、予測される可制御性表現を含み得る。システム300は、(i)状態予測ニューラルネットワークによって生成された予測される可制御性表現と、(ii)次の時間ステップにおいて環境の実際の状態を特徴づける「目標」可制御性表現との間の誤差を測定する目的関数を最適化するために、埋め込みニューラルネットワーク302と状態予測ニューラルネットワークとを一緒に訓練することができる。「目標」可制御性表現は、次の時間ステップにおける環境の実際の状態を特徴づける観測値に基づいて埋め込みニューラルネットワークによって生成され得る。特に、システム300は、目的関数の勾配を、状態予測ニューラルネットワークを介して複数の訓練反復の各々において埋め込みニューラルネットワーク302に逆伝播させ得る。目的関数は、たとえば、二乗誤差目的関数であり得る。この方式で埋め込みニューラルネットワークを訓練することは、エージェントのアクションによって影響される環境についての情報、すなわち、環境の制御可能な態様を符号化するように可制御性表現を促す。

【０１０８】

図4は、タスクエピソードを実行するため、およびリターン計算方式選択ポリシーを更新するための例示的なプロセス400の流れ図である。便宜上、プロセス400は、1つまたは複数のロケーション内に配置された1つまたは複数のコンピュータのシステムによって実行されているものとして説明する。たとえば、アクション選択システム、たとえば、本明細書に従って適切にプログラムされた図1のアクション選択システム100は、プロセス500を実行することができる。

【０１０９】

システムは、複数の異なるリターン計算方式の間で選択するためのポリシーを規定するポリシーデータを維持し、各リターン計算方式は、異なる重要度を、タスクのエピソードを実行しながら環境を探査することに対して割り当てる(ステップ402)。

【０１１０】

システムは、ポリシーを使用して、複数の異なるリターン計算方式からリターン計算方式を選択する(ステップ404)。たとえば、ポリシーは、方式が1つのエピソードに対してエージェントを制御するために使用される場合、受信される報酬信号の現在の推定を表す各方式に、それぞれの報酬スコアを割り当てることができる。一例として、ポリシーを選択するために、システムは、最高の報酬スコアを有する方式を選択することができる。別の例として、確率εを用いて、システムは、方式のセットからランダム方式を選択することができ、確率1-εを用いて、システムは、ポリシーによって規定された最高の報酬スコアを有する方式を選択することができる。別の例として、ポリシーを選択するために、システムは、リターン計算方式のセットにわたって報酬スコアを確率分布に対してマッピングし、次いで、確率分布から方式をサンプリングすることができる。

【０１１１】

システムは、選択されたリターン計算方式に従って計算されたリターンを最大化するようにタスクのエピソードを実行するためにエージェントを制御する(ステップ406)。すなわち、タスクエピソードの間に、システムは、選択されたリターン計算方式を識別するデータ上でアクション選択ニューラルネットワークを調整する。

【０１１２】

システムは、エージェントがタスクのエピソードを実行した結果として生成された報酬を識別する(ステップ408)。特定の例として、システムは、外発的報酬、すなわちタスクエピソード内の時間ステップの各々において受信された、タスク上の進歩を測定する報酬を識別することができる。

【０１１３】

システムは、識別された報酬を使用して、複数の異なるリターン計算方式の間で選択す

るためにポリシーを更新する(ステップ410)。

【０１１４】

一般に、システムは、リターン計算方式の各々に対応するそれぞれの腕を有する非定常多腕バンディットアルゴリズムを使用してポリシーを更新する。

【０１１５】

より具体的には、システムは、識別された報酬からバンディットアルゴリズムに対する報酬信号を生成し、次いで、報酬信号を使用してポリシーを更新することができる。報酬信号は、タスクエピソードの間に受信された外発的報酬の組合せ、たとえば、受信された報酬のディスカウントされない合計であるディスカウントされない外発的報酬であり得る。

【０１１６】

システムは、様々な非定常多腕バンディットアルゴリズムのいずれかを使用して更新を実行することができる。

【０１１７】

特定の例として、システムは、各方式に対して、現在のエピソードのいくつかの固定数のタスクエピソード内、すなわち固定長の直近の計画期間(horizon)内のエピソードに対して受信された報酬信号の経験的平均(empirical mean)を計算することができる。次いで、システムは、方式に対する経験的平均から各方式に対する報酬スコアを計算することができる。たとえば、システムは、信頼限界ボーナス(confidence bound bonus)を所与の方式に対する経験的平均に追加することによって所与の方式に対する報酬スコアを計算することができる。信頼限界ボーナスは、すなわちより少ない時間を選択された方式が、より大きいボーナスを割り当てられるように、どれほど多くの回数を、所与の方式が最近の計画期間内に選択されたかに基づいて決定され得る。特定の例として、k番目のタスクエピソードの後に計算された所与の方式aに対するボーナスは、

【数１２】

$$\beta\sqrt{\frac{1}{N_{k-1}(a,\tau)}}\ \text{or}\ \beta\sqrt{\frac{\log(k-\min(k-1,\tau))}{N_{k-1}(a,\tau)}},$$

を満たすことができ、ここでβは固定された重みであり、$N_{k-1}(a,\tau)$は所与の方式aが最近の計画期間内に選択された回数であり、τは計画期間の長さである。

【０１１８】

したがって、システムは、アクション選択ニューラルネットワークの訓練中のリターン計算方式の間で選択するためのポリシーを適応的に修正し、異なるリターン計算方式が、訓練中の異なる時間においてポリシーによって好まれる(それゆえ、選択される可能性がより高くなる)ことをもたらす。ポリシーは、訓練中の任意の所与のポイントにおいて異なる方式に対して(外発的報酬に基づく)予期される報酬信号に基づくので、システムは、より高い外発的報酬がタスクエピソードにわたって収集されることをもたらす可能性がより高い方式を選択する可能性がより高く、より高い品質の訓練データがアクション選択ニューラルネットワークに対して生成されることをもたらす。

【０１１９】

本明細書は、システムおよびコンピュータプログラム構成要素との接続において「構成される」という用語を使用する。特定の動作またはアクションを実行するように構成されるべき1つまたは複数のコンピュータのシステムは、システムが、動作中に動作またはアクションをシステムに実行させるソフトウェア、ファームウェア、ハードウェアまたはそれらの組合せを、システム上にインストールしていることを意味する。特定の動作またはアクションを実行するように構成されるべき1つまたは複数のコンピュータプログラムは、1つまたは複数のプログラムが命令を含み、その命令は、データ処理装置によって実行されたとき、装置に動作またはアクションを実行させることを意味する。

【０１２０】

本明細書で説明する主題および関数演算の実施形態は、本明細書で開示する構造およびそれらの構造的等価物を含めて、ディジタル電子回路、明確に具現化されたコンピュータソフトウェアもしくはファームウェア、コンピュータハードウェア、またはそれらのうちの1つまたは複数の組合せの中に実装され得る。本明細書で説明する主題の実施形態は、1つまたは複数のコンピュータプログラム、すなわち、データ処理装置による実行のためにまたはデータ処理装置の動作を制御するために、有形の非一時的記憶媒体上に符号化されたコンピュータプログラム命令の1つまたは複数のモジュールとして実装され得る。コンピュータ記憶媒体は、機械可読記憶デバイス、機械可読記憶基板、ランダムもしくはシリアルアクセスメモリデバイス、またはそれらのうちの1つまたは複数の組合せであり得る。代替または追加として、プログラム命令は、データ処理装置による実行に好適な受信機装置に送信するための情報を符号化するために生成される、人工的に生成された伝播信号、たとえば機械的に生成された電気、光、または電磁の信号上に符号化され得る。

【０１２１】

「データ処理装置」という用語はデータ処理ハードウェアを指し、例としてプログラム可能な一プロセッサ、一コンピュータ、または複数のプロセッサもしくはコンピュータを含む、データを処理するためのあらゆる種類の装置、デバイス、および機械を包含する。装置はまた、専用論理回路、たとえばFPGA(フィールドプログラマブルゲートアレー)またはASIC(特定用途向け集積回路)であり得、またはそれをさらに含むことができる。装置は、随意に、ハードウェアに加えて、コンピュータプログラムに対する実行環境を生成するコード、たとえばプロセッサファームウェア、プロトコルスタック、データベース管理システム、オペレーティングシステム、またはそれらのうちの1つもしくは複数の組合せを構成するコード含むことができる。

【０１２２】

プログラム、ソフトウェア、ソフトウェアアプリケーション、アプリ、モジュール、ソフトウェアモジュール、スクリプト、またはコードと呼ばれるかまたは記述されることがあるコンピュータプログラムは、コンパイラ型もしくはインタプリタ型の言語、または宣言型もしくは手続き型の言語を含む任意の形態のプログラミング言語で書かれてもよく、コンピュータプログラムは、スタンドアローンプログラムとして、またはモジュール、コンポーネント、サブルーチン、もしくはコンピューティング環境で使用するのに好適な他のユニットとして含む、任意の形態で配布され得る。プログラムは、ファイルシステム内のファイルに相当する場合があるが、必ずしも必要ではない。プログラムは、他のプログラムもしくはデータを保持するファイルの一部、たとえば、マークアップ言語文書内に記憶された1つもしくは複数のスクリプトの中、当該のプログラムに専用の単一のファイルの中、または複数の協調的ファイル、たとえば、1つもしくは複数のモジュール、サブプログラムもしくはコードの部分を記憶するファイルの中に記憶され得る。コンピュータプログラムは、1つのサイトに配置されるかまたは複数のサイトにわたって分配されてデータ通信ネットワークで相互接続される、1つのコンピュータまたは複数のコンピュータ上で実行されるように配備され得る。

【０１２３】

本明細書では、「エンジン」という用語が、ソフトウェアベースシステム、サブシステム、または1つもしくは複数の特定の機能を実行するためにプログラムされたプロセスを指すために広く使用される。一般に、エンジンは、1つまたは複数のコンピュータ上の1つまたは複数のロケーションに設置される1つまたは複数のソフトウェアモジュールまたは構成要素として実装される。場合によっては、1つまたは複数のコンピュータが特定のエンジンに専用であり、他の場合には、複数のエンジンが同じ1つまたは複数のコンピュータ上に設置されて実行することができる。

【０１２４】

本明細書で説明するプロセスおよび論理フローは、入力データに対して動作して出力を生成することによって関数を実行するために、1つまたは複数のコンピュータプログラムを実行する1つまたは複数のプログラム可能なコンピュータによって実行され得る。プロ

セスおよび論理フローもまた、専用の論理回路、たとえばFPGAもしくはASICによって、または専用の論理回路と1つもしくは複数のプログラムされたコンピュータとの組合せによって実行され得る。

【０１２５】

コンピュータプログラムの実行に好適なコンピュータは、汎用もしくは専用のマイクロプロセッサ、または汎用および専用のマイクロプロセッサ、あるいは任意の他の種類の中央処理装置に基づくことができる。一般的に、中央処理装置は、リードオンリーメモリもしくはランダムアクセスメモリ、または両メモリから命令およびデータを受信することになる。コンピュータの必須要素は、命令を実行または実施するための中央処理装置と、命令およびデータを記憶するための1つまたは複数のメモリデバイスとである。中央処理装置およびメモリは、専用論理回路によって補完され得るか、または専用論理回路内に組み込まれ得る。一般に、コンピュータはまた、データを記憶するための1つもしくは複数の大容量記憶デバイス、たとえば磁気ディスク、磁気光ディスクもしくは光ディスクを含むか、またはその記憶デバイスからデータを受信するかその記憶デバイスにデータを伝達するかもしくはその両方を行うように動作可能に結合される。しかしながら、コンピュータは、必ずしもそのようなデバイスを有する必要があるとは限らない。その上、コンピュータは、別のデバイス、たとえば数例を挙げると、モバイル電話、携帯情報端末(PDA)、モバイルオーディオもしくはビデオプレーヤ、ゲームコンソール、全地球測位システム(GPS)受信機、またはユニバーサルシリアルバス(USB)フラッシュドライブなどのポータブルストレージデバイスの中に組み込まれ得る。

【０１２６】

コンピュータプログラム命令およびデータを記憶するのに好適なコンピュータ可読媒体は、例として、EPROM、EEPROMおよびフラッシュメモリデバイスなどの半導体メモリデバイスと、内部ハードディスクまたはリムーバブルディスクなどの磁気ディスクと、磁気光ディスクと、CD-ROMおよびDVD-ROMディスクとを含む、あらゆる形態の不揮発性メモリ、メディアおよびメモリデバイスを含む。

【０１２７】

ユーザとの相互作用を提供するために、本明細書で説明する主題の実施形態は、ユーザに情報を表示するのための表示デバイス、たとえばCRT(陰極線管)もしくはLCD(液晶ディスプレイ)のモニタと、ユーザがコンピュータに入力を与えることができるキーボードおよびポインティングデバイス、たとえばマウスもしくはトラックボールとを有するコンピュータ上に実装され得る。他の種類のデバイスは、同様にユーザとの相互作用を提供するために使用され得、たとえば、ユーザに与えられるフィードバックは、任意の形態の知覚フィードバック、たとえば視覚フィードバック、聴覚フィードバックまたは触覚フィードバックであり得、ユーザからの入力は、音響、音声または触覚の入力を含む任意の形態で受信され得る。加えて、コンピュータは、ユーザによって使用されるデバイスに文書を送信すること、およびそのデバイスから文書を受信することによって、たとえば、ウェブブラウザから受信された要求に応答してユーザのデバイス上のウェブブラウザにウェブページを送信することによって、ユーザと相互作用することができる。同じく、コンピュータは、テキストメッセージまたは他の形式のメッセージをパーソナルデバイス、たとえばメッセージングアプリケーションを実行しているスマートフォンに送信し、返報としてユーザから応答メッセージを受信することによってユーザと相互作用することができる。

【０１２８】

機械学習モデルを実装するためのデータ処理装置はまた、たとえば、機械学習訓練または製作の共通の計算集約型の部分、すなわち、推論、仕事量を処理するための専用ハードウェアアクセラレータユニットを含むことができる。

【０１２９】

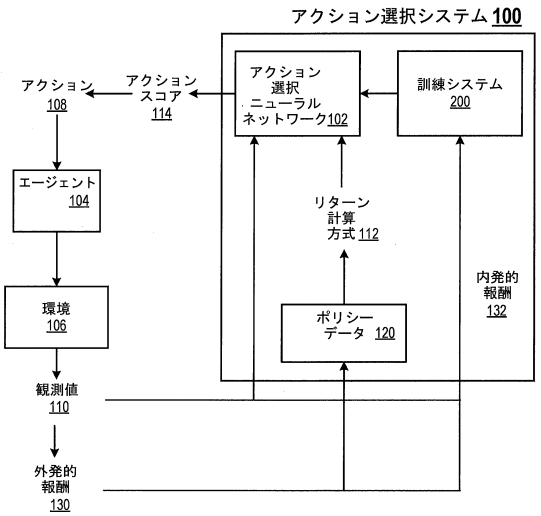
機械学習モデルは、機械学習フレームワーク、たとえば、TensorFlowフレームワーク、マイクロソフトコグニティブツールキットフレームワーク、Apache Singaフレームワーク、またはApache MXNetフレームワークを使用して実装され、配備され得る。

【０１３０】
本明細書で説明する主題の実施形態は、たとえばデータサーバとしてバックエンド構成要素を含むか、またはアプリケーションサーバなどのミドルウェア構成要素を含むか、またはユーザが、本明細書で説明する主題の実装形態と相互作用し得るグラフィカルユーザインターフェース、ウェブブラウザ、もしくはアプリを有するクライアントコンピュータなどのフロントエンド構成要素を含むコンピューティングシステムの中、あるいは1つまたは複数のそのようなバックエンド、ミドルウェアまたはフロントエンドの構成要素の任意の組合せの中に実装され得る。システムの構成要素は、任意の形態または媒体のディジタルデータ通信、たとえば通信ネットワークによって相互接続され得る。通信ネットワークの例には、ローカルエリアネットワーク(LAN)と、ワイドエリアネットワーク(WAN)、たとえばインターネットとが含まれる。
【０１３１】
コンピューティングシステムは、クライアントとサーバとを含むことができる。クライアントおよびサーバは、一般に互いに離れており、通常通信ネットワークを介して相互作用する。クライアントとサーバとの関係は、それぞれのコンピュータ上で動作し、互いにクライアント-サーバ関係を有するコンピュータプログラムによって生じる。いくつかの実施形態では、サーバは、データ、たとえば、HTMLページをユーザデバイスに、たとえば、クライアントとして働くデバイスと相互作用するユーザにデータを表示し、そのユーザからユーザ入力を受信するために送信する。ユーザデバイスにおいて生成されたデータ、たとえば、ユーザとの相互作用の結果は、サーバにおいてデバイスから受信され得る。
【０１３２】
本明細書は多くの特定の実装形態の詳細を含むが、これらは、発明の範囲または請求されるものの範囲を限定するものと解釈されるべきではなく、特定の発明の特定の実施形態に特有の特徴を説明するものと解釈されるべきである。個別の実施形態の文脈において本明細書で説明されるいくつかの特徴はまた、単一の実施形態の中で組み合わせて実装され得る。反対に、単一の実施形態の文脈において説明される様々な特徴はまた、複数の実施形態において個別に、または任意の適切な部分的組合せにおいて実装され得る。その上、特徴は、特定の組合せにおいて働くように上記で説明され、最初にそのようなものとして請求されるが、いくつかの場合には、請求される組合せからの1つまたは複数の特徴は、その組合せから削除されてもよく、請求される組合せは、部分的組合せ、または部分的組合せの変形態に移されてもよい。
【０１３３】
同様に、動作は特定の順序で図に示され、請求項に記載されるが、これは、望ましい結果を達成するために、そのような動作が図示の特定の順序で、もしくは一連の順序で実行されること、または図示の動作のすべてが実行されることを必要とするものと理解されるべきではない。いくつかの状況では、多重タスク処理および並列処理が有利であり得る。その上、上記で説明した実施形態における様々なシステムモジュールおよび構成要素の分離は、すべての実施形態においてそのような分離を必要とするものと理解されるべきではなく、説明したプログラム構成要素およびシステムは、一般に、単一のソフトウェア製品中に一緒に一体化されてもよく、または複数のソフトウェア製品中にパッケージ化されてもよいものと理解されるべきである。
【０１３４】
主題の特定の実施形態が説明された。他の実施形態は、以下の特許請求の範囲の中にある。たとえば、特許請求の範囲に記載される行動は、異なる順序で実行されても、依然として望ましい結果を達成することができる。一例として、添付の図に示すプロセスは、望ましい結果を達成するために、必ずしも図示の特定の順序または一連の順序を必要とするとは限らない。場合によっては、多重タスク処理および並列処理が有利であり得る。
【符号の説明】
【０１３５】
100　アクション選択システム

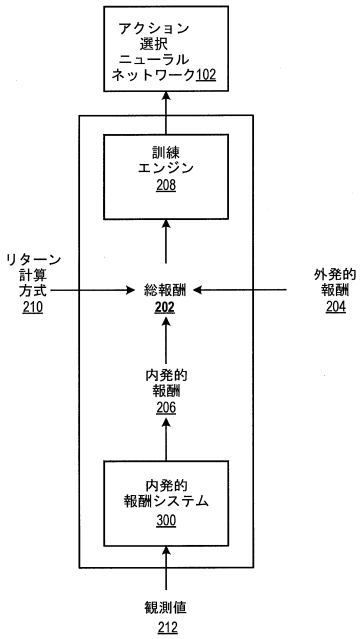
102 アクション選択ニューラルネットワーク
104 エージェント
106 環境
108 アクション
110 観測値
112 リターン計算方式
114 アクションスコア
120 ポリシーデータ
130 外発的報酬
132 内発的報酬
200 訓練システム
202 総報酬
204 外発的報酬
206 内発的報酬、探査報酬
208 訓練エンジン
210 リターン計算方式
212 観測値
300 内発的報酬システム
302 埋め込みニューラルネットワーク
304 外部メモリ
306 比較エンジン
308 可制御性表現

【図面】

【図１】

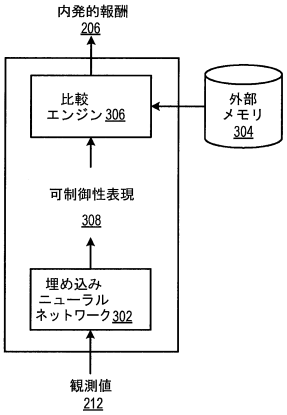


【図２】



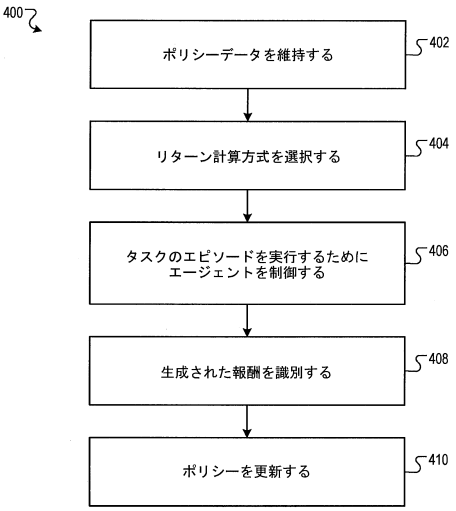
訓練システム 200

【図３】



内発的報酬システム300

【図４】



フロントページの続き

イギリス・Ｎ１Ｃ・４ＡＧ・ロンドン・パンクラス・スクエア・６
(72)発明者　パブロ・スプレッヒマン
イギリス・Ｎ１Ｃ・４ＡＧ・ロンドン・パンクラス・スクエア・６
(72)発明者　スティーヴン・ジェームズ・カプチュロヴスキ
イギリス・Ｎ１Ｃ・４ＡＧ・ロンドン・パンクラス・スクエア・６
(72)発明者　アレックス・ヴィトヴィツキイ
イギリス・Ｎ１Ｃ・４ＡＧ・ロンドン・パンクラス・スクエア・６
(72)発明者　ジャオハン・グオ
イギリス・Ｎ１Ｃ・４ＡＧ・ロンドン・パンクラス・スクエア・６
(72)発明者　チャールズ・ブランデル
イギリス・Ｎ１Ｃ・４ＡＧ・ロンドン・パンクラス・スクエア・６

審査官　渡辺 順哉
(56)参考文献　特表２０１９－５３７１３６（ＪＰ，Ａ）
ZHANG, Jingwei ほか，Scheduled Intrinsic Drive: A Hierarchical Take on Intrinsically Motivated Exploration，arXiv[online]，2019年06月21日，[retrieved on 2023.08.17]，Retrieved from the Internet: 〈URL: https://arxiv.org/pdf/1903.07400v2.pdf〉
(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ０６Ｎ　３／００－９９／００