(19)日本国特許庁(JP)

# (12)特許公報(B2)

(11)特許番号
特許第7646870号
(P7646870)

(45)発行日　令和7年3月17日(2025.3.17)

(24)登録日　令和7年3月7日(2025.3.7)

(51)国際特許分類
*Ｇ０６Ｎ　3/092(2023.01)*

ＦＩ
Ｇ０６Ｎ　3/092

請求項の数　20　（全31頁）

(21)出願番号　特願2023-562576(P2023-562576)
(86)(22)出願日　令和4年5月27日(2022.5.27)
(65)公表番号　特表2024-519271(P2024-519271A)
(43)公表日　令和6年5月10日(2024.5.10)
(86)国際出願番号　PCT/EP2022/064499
(87)国際公開番号　WO2022/248725
(87)国際公開日　令和4年12月1日(2022.12.1)
審査請求日　令和5年11月10日(2023.11.10)
(31)優先権主張番号　63/194,798
(32)優先日　令和3年5月28日(2021.5.28)
(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　スティーブン・ステンバーグ・ハンセン
イギリス・Ｎ１Ｃ・４ＡＧ・ロンドン・パンクラス・スクエア・６
(72)発明者　ダニエル・ジョセフ・ストラウス
最終頁に続く

(54)【発明の名称】　弁別器モデルの集合を使用した強化学習

(57)【特許請求の範囲】
【請求項１】
1つまたは複数のデータ処理装置によって実施される方法であって、
可能な潜在性のセットからの潜在性をサンプリングするステップと、
前記サンプリングされた潜在性を条件とするアクション選択ニューラルネットワークを使用して、時間ステップのシーケンスにわたって環境と対話するためにエージェントによって実施されるべきアクションを選択するステップと、
時間ステップの前記シーケンスの各時間ステップに対して受けるそれぞれの報酬を決定するステップであって、1つまたは複数の時間ステップの各々に対して、
前記時間ステップにおける前記環境の状態を表す観測を複数の弁別器モデルの集合内の各弁別器モデルに提供するステップであって、各弁別器モデルが、前記環境を前記観測によって特徴付けられる前記状態に入らせるために前記アクション選択ニューラルネットワークが可能な潜在性の前記セットからのどの潜在性を条件としたかを予測するそれぞれの予測出力を生成するために、前記観測を処理する、提供するステップと、
複数の弁別器モデルの前記集合によって生成される前記予測出力同士の間の不一致の尺度に少なくとも部分的に基づいて、前記時間ステップに対する前記報酬を決定するステップと
を含む、決定するステップと、
強化学習技法を使用して、前記報酬に基づいて、前記アクション選択ニューラルネットワークをトレーニングするステップと

を含む、方法。
【請求項２】
前記1つまたは複数の時間ステップの各々に対して、前記時間ステップに対して各弁別器モデルによって生成される前記予測出力が、可能な潜在性の前記セット内の各潜在性に対するそれぞれのスコアを定義する、可能な潜在性の前記セットにわたるそれぞれのスコア分布を含む、請求項1に記載の方法。
【請求項３】
前記1つまたは複数の時間ステップの各々に対して、複数の弁別器モデルの前記集合によって生成される前記予測出力同士の間の不一致の前記尺度を決定するステップであって、
前記時間ステップに対して各弁別器モデルによって生成される前記それぞれのスコア分布を組み合わせることによって、前記時間ステップに対して可能な潜在性の前記セットにわたって組み合わされたスコア分布を決定するステップと、
(i)前記時間ステップに対して前記組み合わされたスコア分布、および(ii)前記時間ステップに対して各弁別器モデルによって生成される前記それぞれのスコア分布に基づいて、不一致の前記尺度を決定するステップと
を含む、決定するステップをさらに含む、請求項2に記載の方法。
【請求項４】
前記時間ステップに対して可能な潜在性の前記セットにわたって前記組み合わされたスコア分布を決定するステップが、前記時間ステップに対して前記弁別器モデルによって生成される前記スコア分布を平均化するステップを含む、請求項3に記載の方法。
【請求項５】
(i)前記時間ステップに対して前記組み合わされたスコア分布、および(ii)前記時間ステップに対して各弁別器モデルによって生成される前記それぞれのスコア分布に基づいて、不一致の前記尺度を決定するステップが、
前記時間ステップに対して前記組み合わされたスコア分布のエントロピーを決定するステップと、
前記時間ステップに対して各弁別器モデルによって生成される前記それぞれのスコア分布のそれぞれのエントロピーを決定するステップと、
前記時間ステップに対して前記組み合わされたスコア分布、および前記時間ステップに対して各弁別器モデルによって生成される前記それぞれのスコア分布の前記それぞれのエントロピーに基づいて、不一致の前記尺度を決定するステップと
を含む、請求項3または4に記載の方法。
【請求項６】
前記時間ステップに対して前記組み合わされたスコア分布、および前記時間ステップに対して各弁別器モデルによって生成される前記それぞれのスコア分布の前記それぞれのエントロピーに基づいて、不一致の前記尺度を決定するステップが、
(i)前記時間ステップに対して前記組み合わされたスコア分布の前記エントロピーと(ii)前記時間ステップに対して各弁別器モデルによって生成される前記スコア分布の前記エントロピーの平均との間の差異として、不一致の前記尺度を決定するステップ
を含む、請求項5に記載の方法。
【請求項７】
前記1つまたは複数の時間ステップの各々に対して、前記時間ステップに対する前記報酬を決定するステップが、
前記時間ステップに対して各弁別器モデルによって生成される前記予測出力のそれぞれの精度を決定するステップと、
前記時間ステップに対して前記弁別器モデルによって生成される前記予測出力の前記精度に少なくとも部分的に基づいて、前記時間ステップに対する前記報酬を決定するステップと
を含む、請求項1から4のいずれか一項に記載の方法。
【請求項８】

前記時間ステップに対して前記弁別器モデルによって生成される前記予測出力の前記精度に少なくとも部分的に基づいて、前記時間ステップに対する前記報酬を決定するステップが、
前記時間ステップに対して前記弁別器モデルによって生成される前記予測出力の前記精度の平均に少なくとも部分的に基づいて、前記報酬を決定するステップ
を含む、請求項7に記載の方法。
【請求項９】
各弁別器モデルが弁別器モデルパラメータのそれぞれのセットを有し、弁別器モデルパラメータの前記それぞれのセットのそれぞれの値が、各弁別器モデルに対して異なる、請求項1から4のいずれか一項に記載の方法。
【請求項１０】
各弁別器モデルが、リプレイメモリからのトレーニング例の単独でサンプリングされたバッチに基づいてトレーニングされる、請求項9に記載の方法。
【請求項１１】
前記リプレイメモリ内の各トレーニング例が、(i)前記環境との前記エージェントの前の対話中の前記環境の状態を特徴付けるトレーニング観測と、(ii)前記環境を前記観測によって特徴付けられる前記状態に入らせるために前記アクション選択ニューラルネットワークが条件とした潜在性を定義するターゲット出力とを含む、請求項10に記載の方法。
【請求項１２】
トレーニングされる前に、各弁別器モデルの弁別器モデルパラメータの前記セットの前記それぞれの値が、各他の弁別器モデルの弁別器モデルパラメータの前記セットとは異なるように初期化される、請求項11に記載の方法。
【請求項１３】
時間ステップの前記シーケンスの各時間ステップに対して、
前記環境におけるタスクの達成における前記エージェントの進歩を測定するタスク報酬に少なくとも部分的に基づいて、前記時間ステップに対する報酬を決定するステップ
をさらに含む、請求項1から4のいずれか一項に記載の方法。
【請求項１４】
可能な潜在性の前記セットが、有限に多くの可能な潜在性のみを含む、請求項1から4のいずれか一項に記載の方法。
【請求項１５】
前記サンプリングされた潜在性を条件とする前記アクション選択ニューラルネットワークを使用して、時間ステップの前記シーケンスにわたって前記環境と対話するために前記エージェントによって実施されるべきアクションを選択するステップが、各時間ステップに対して、
アクション選択出力を生成するために、前記アクション選択ニューラルネットワークを使用して、前記時間ステップにおける前記環境の状態を特徴付ける観測および前記サンプリングされた潜在性を処理するステップと、
前記アクション選択出力に基づいて、前記時間ステップにおいて実施されるべき前記アクションを選択するステップと
を含む、請求項1から4のいずれか一項に記載の方法。
【請求項１６】
各時間ステップに対して、前記アクション選択出力が、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スコアを含む、請求項15に記載の方法。
【請求項１７】
各時間ステップに対して、前記アクション選択出力に基づいて、前記時間ステップにおいて実施されるべき前記アクションを選択するステップが、
前記アクション選択出力に従って最高スコアを有する前記アクションを選択するステップを含む、請求項16に記載の方法。
【請求項１８】

前記アクション選択ニューラルネットワークをトレーニングした後、現実世界環境と対話している現実世界エージェントによって実施されるべきアクションを選択するために、前記アクション選択ニューラルネットワークを使用するステップをさらに含む、請求項1から4のいずれか一項に記載の方法。

【請求項１９】

システムであって、

1つまたは複数のコンピュータと、

前記1つまたは複数のコンピュータに通信可能に結合された1つまたは複数の記憶デバイスと

を含み、前記1つまたは複数の記憶デバイスが、前記1つまたは複数のコンピュータによって実行されると、前記1つまたは複数のコンピュータに請求項1から4のいずれか一項に記載の方法を実施させる命令を記憶する

システム。

【請求項２０】

命令を記憶した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ると、前記1つまたは複数のコンピュータに請求項1から4のいずれか一項に記載の方法を実施させる、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】

関連出願の相互参照

本出願は、その全体が参照により本明細書に組み込まれている、2021年5月28日に出願した「REINFORCEMENT LEARNING USING AN ENSEMBLE OF DISCRIMINATOR MODELS」に対する米国仮特許出願第63/194,798号の出願日の利益を主張するものである。

【０００２】

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】

【０００３】

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、予測される出力、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。

【０００４】

いくつかの機械学習モデルは、受信された入力に対して出力を生成するために多層モデルを採用し得る深層モデルである。たとえば、深層ニューラルネットワークは、出力層と、各々が受信された入力に非線形変換を適用して出力を生成する1つまたは複数の隠れ層とを含む、深層機械学習モデルである。

【発明の概要】

【課題を解決するための手段】

【０００５】

本明細書は、概して、環境と対話しているエージェントによって実施されるべきアクションを選択するために使用されるアクション選択ニューラルネットワークをトレーニングする、1つまたは複数のロケーションにおいて1つまたは複数のコンピュータ上でコンピュータプログラムとして実装されるシステムについて説明する。

【０００６】

本明細書で説明するシステムは、「タスク報酬」、たとえば、環境におけるタスクの達成に向けたエージェントの進歩を特徴付ける外的報酬、がない場合ですら、アクション選択ニューラルネットワークをトレーニングし得る。

【０００７】

たとえば、システムは最初に、弁別器モデルの集合によって生成される「教師なし」報酬のみに基づく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をトレーニングし得る。教師なし報酬を使用してアクション選択ニューラルネットワークを事前トレーニングした後、システムは、タスク報酬に基づいて、またはタスク報酬と教師なし報酬の組合せに基づいて、アクション選択ニューラルネットワークをトレーニングし得る。

【０００８】

システムは、任意の適切な強化学習技法、たとえば、Q学習技法またはポリシー勾配法技法(policy gradient technique)を使用して、アクション選択ニューラルネットワークをトレーニングし得る。

【０００９】

第1の態様によれば、1つまたは複数のデータ処理装置によって実施される方法であって、可能な潜在性のセットからの潜在性(latent)をサンプリングするステップと、サンプリングされた潜在性を条件とするアクション選択ニューラルネットワークを使用して、時間ステップのシーケンスにわたって環境と対話するためにエージェントによって実施されるべきアクションを選択するステップと、時間ステップのシーケンスの各時間ステップに対して受けるそれぞれの報酬を決定するステップと、強化学習技法を使用して、報酬に基づいて、アクション選択ニューラルネットワークをトレーニングするステップとを含む、方法が提供される。

【００１０】

時間ステップのシーケンスの各時間ステップに対して受けるそれぞれの報酬を決定するステップは、1つまたは複数の時間ステップの各々に対して、時間ステップにおける環境の状態を表す観測を複数の弁別器モデルの集合内の各弁別器モデルに提供するステップであって、各弁別器モデルが、環境をその観測によって特徴付けられる状態に入らせるためにアクション選択ニューラルネットワークが可能な潜在性のセットからのどの潜在性を条件としたかを予測するそれぞれの予測出力を生成するために、その観測を処理する、提供するステップと、複数の弁別器モデルの集合によって生成される予測出力同士の間の不一致の尺度に少なくとも部分的に基づいて、時間ステップに対する報酬を決定するステップとを含み得る。

【００１１】

いくつかの実装形態では、1つまたは複数の時間ステップの各々に対して、時間ステップに対して各弁別器モデルによって生成される予測出力は、可能な潜在性のセット内の各潜在性に対するそれぞれのスコアを定義する、可能な潜在性のセットにわたるそれぞれのスコア分布を含む。

【００１２】

いくつかの実装形態では、方法はさらに、1つまたは複数の時間ステップの各々に対して、複数の弁別器モデルの集合によって生成される予測出力同士の間の不一致の尺度を決定するステップであって、時間ステップに対して各弁別器モデルによって生成されるそれぞれのスコア分布を組み合わせることによって、時間ステップに対して可能な潜在性のセットにわたって組み合わされたスコア分布を決定するステップと、(i)時間ステップに対して組み合わされたスコア分布、および(ii)時間ステップに対して各弁別器モデルによって生成されるそれぞれのスコア分布に基づいて、不一致の尺度を決定するステップと含む、決定するステップを含む。

【００１３】

いくつかの実装形態では、時間ステップに対して可能な潜在性のセットにわたって組み合わされたスコア分布を決定するステップは、時間ステップに対して弁別器モデルによって生成されるスコア分布を平均化するステップを含む。

【００１４】

いくつかの実装形態では、(i)時間ステップに対して組み合わされたスコア分布、および(ii)時間ステップに対して各弁別器モデルによって生成されるそれぞれのスコア分布に基づいて、不一致の尺度を決定するステップは、時間ステップに対して組み合わされたスコア

分布のエントロピーを決定するステップと、時間ステップに対して各弁別器モデルによって生成されるそれぞれのスコア分布のそれぞれのエントロピーを決定するステップと、時間ステップに対して組み合わされたスコア分布、および時間ステップに対して各弁別器モデルによって生成されるそれぞれのスコア分布のそれぞれのエントロピーに基づいて、不一致の尺度を決定するステップとを含む。

【００１５】

いくつかの実装形態では、時間ステップに対して組み合わされたスコア分布、および時間ステップに対して各弁別器モデルによって生成されるそれぞれのスコア分布のそれぞれのエントロピーに基づいて、不一致の尺度を決定するステップは、(i)時間ステップに対して組み合わされたスコア分布のエントロピーと(ii)時間ステップに対して各弁別器モデルによって生成されるスコア分布のエントロピーの平均との間の差異として不一致の尺度を決定するステップを含む。

【００１６】

いくつかの実装形態では、1つまたは複数の時間ステップの各々に対して、時間ステップに対する報酬を決定するステップは、時間ステップに対して各弁別器モデルによって生成される予測出力のそれぞれの精度を決定するステップと、時間ステップに対して弁別器モデルによって生成される予測出力の精度に少なくとも部分的に基づいて、時間ステップに対する報酬を決定するステップとを含む。

【００１７】

いくつかの実装形態では、時間ステップに対して弁別器モデルによって生成される予測出力の精度に少なくとも部分的に基づいて、時間ステップに対する報酬を決定するステップは、時間ステップに対して弁別器モデルによって生成される予測出力の精度の平均に少なくとも部分的に基づいて、報酬を決定するステップを含む。

【００１８】

いくつかの実装形態では、各弁別器モデルは弁別器モデルパラメータのそれぞれのセットを有し、弁別器モデルパラメータのそれぞれのセットのそれぞれの値は、各弁別器モデルに対して異なる。

【００１９】

いくつかの実装形態では、各弁別器モデルは、リプレイメモリからのトレーニング例の単独でサンプリングされたバッチに基づいてトレーニングされる。

【００２０】

いくつかの実装形態では、リプレイメモリ内の各トレーニング例は、(i)環境とのエージェントの前の対話中の環境の状態を特徴付けるトレーニング観測と、(ii)環境を観測によって特徴付けられる状態に入らせるためにアクション選択ニューラルネットワークが条件とした潜在性を定義するターゲット出力とを含む。

【００２１】

いくつかの実装形態では、トレーニングされる前に、各弁別器モデルの弁別器モデルパラメータのセットのそれぞれの値は、各他の弁別器モデルの弁別器モデルパラメータのセットとは異なるように初期化される。

【００２２】

いくつかの実装形態では、方法はさらに、時間ステップのシーケンスの各時間ステップに対して、環境におけるタスクの達成におけるエージェントの進歩を測定するタスク報酬に少なくとも部分的に基づいて、時間ステップに対する報酬を決定するステップを含む。

【００２３】

いくつかの実装形態では、可能な潜在性のセットは、有限に多くの可能な潜在性のみを含む。

【００２４】

いくつかの実装形態では、サンプリングされた潜在性を条件とするアクション選択ニューラルネットワークを使用して、時間ステップのシーケンスにわたって環境と対話するためにエージェントによって実施されるべきアクションを選択するステップは、各時間ステ

ップに対して、アクション選択出力を生成するために、アクション選択ニューラルネットワークを使用して、時間ステップにおける環境の状態を特徴付ける観測およびサンプリングされた潜在性を処理するステップと、アクション選択出力に基づいて、時間ステップにおいて実施されるべきアクションを選択するステップとを含む。

【００２５】

いくつかの実装形態では、各時間ステップに対して、アクション選択出力は、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スコアを含む。

【００２６】

いくつかの実装形態では、各時間ステップに対して、アクション選択出力に基づいて、時間ステップにおいて実施されるべきアクションを選択するステップは、アクション選択出力に従って最高スコアを有するアクションを選択するステップを含む。

【００２７】

いくつかの実装形態では、方法はさらに、アクション選択ニューラルネットワークをトレーニングした後、現実世界環境と対話している現実世界エージェントによって実施されるべきアクションを選択するために、アクション選択ニューラルネットワークを使用するステップを含む。

【００２８】

第2の態様によれば、1つまたは複数のコンピュータと、1つまたは複数のコンピュータに通信可能に結合された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、1つまたは複数の記憶デバイスが、1つまたは複数のコンピュータによって実行されると、1つまたは複数のコンピュータに任意の前述の態様のそれぞれの方法の動作を実施させる命令を記憶する、システムが提供される。

【００２９】

第3の態様によれば、命令を記憶した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、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ると、1つまたは複数のコンピュータに、任意の前述の態様のそれぞれの方法の動作を実施させる、1つまたは複数の非一時的コンピュータ記憶媒体が提供される。

【００３０】

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

【００３１】

本明細書で説明するシステムは、「タスク」報酬を使用してアクション選択ニューラルネットワークをトレーニングする代替案としてまたはそれと組み合わせて、弁別器モデルの集合を使用して生成される「教師なし」報酬を使用してアクション選択ニューラルネットワークをトレーニングし得る。教師なし報酬を使用してアクション選択ニューラルネットワークをトレーニングするために、システムは、潜在性をサンプリングし、次いで、アクション選択ニューラルネットワークがサンプリングされた潜在性を条件とすると同時に、時間ステップのシーケンスにわたって環境と対話するためにエージェントによって実施されるべきアクションを選択し得る。教師なし報酬は、たとえば、外部から受けないこと(たとえば、環境から受けないこと)によって、外的(たとえば、タスク)報酬とは異なり得る。たとえば、教師なし(または固有の)報酬は、代わりに、アクションが実施された後に環境の観測に基づいてシステムによって決定される報酬であってよい。教師なし報酬はまた、アクションの選択を条件付けるためにシステムによって使用される潜在性に基づいてもよい。

【００３２】

システムは、弁別器モデルが、環境とのエージェントの対話中にその環境が遷移する状態に基づいて、アクション選択ニューラルネットワークが条件とする潜在性をどの程度正確に予測し得るかに部分的に基づいて、教師なし報酬を決定し得る。これらの教師なし報酬に基づいてアクション選択ニューラルネットワークをトレーニングすることは、アクション選択ニューラルネットワークが、一貫した認識可能な方法で環境の状態を改変させる

「スキル」を学習することを可能にし得る。スキルは、アクション選択ニューラルネットワークを潜在性で条件付けすることによって定義されるアクション選択ポリシーを指す。

【００３３】

アクション選択ニューラルネットワークは、たとえば、より少ないトレーニングデータを使用して、より少ないトレーニング反復にわたって、より迅速に環境においてタスクを実施することを学習するために教師なし報酬を使用して学習される、一貫した認識可能なスキルの多様なセットを活用し得る。しかしながら、弁別器モデルが、アクション選択ニューラルネットワークが条件とする潜在性をどの程度正確に予測し得るかに基づいて教師なし報酬を決定することは、環境の探求を妨げることがある。特に、エージェントが環境の新しい部分を探求するとき、弁別器モデルの予測精度を低下させる可能性があり(環境の新しい部分からの観測に基づいてトレーニングされていないので)、それによって弁別器モデルの予測精度に基づいた報酬が下がる可能性がある。エージェントによって環境の探求を妨げることは、環境において、どのようにして効果的にタスクを実行するかを学習するためにエージェントの能力を制限することがある。

【００３４】

探求を奨励するために、本明細書において説明されるシステムは、たとえば、トレーニング例の単独でサンプリングされたバッチに基づいてそれぞれトレーニングされる、弁別器モデルの集合を使用し、弁別器モデルによって生成される予測同士の間の不一致の尺度に少なくとも部分的に基づいて、教師なし報酬を生成する。エージェントが環境の新しい部分を探求するとき、弁別器モデルによって生成される予測は一致しない傾向があり、それにより、教師なし報酬の価値を増大させる傾向があり、探求を奨励する。弁別器モデル同士の間の不一致の尺度に基づいて教師なし報酬を決定することによって、システムは、エージェントが環境を探求するように奨励し、それにより、エージェントが、たとえば、より少ないトレーニングデータを使用して、より少ないトレーニング反復にわたって、より効果的に環境においてタスクを実施することを学習することを可能にし得る。対照的に、単一の弁別器モデルを使用するシステムは、探求を奨励するための機構として弁別器モデルの不一致を活用することができず、したがって、これらのシステムの実行効果は、本明細書で説明するシステムよりも低い可能性がある。

【００３５】

本明細書の主題の1つまたは複数の実施形態の詳細は、添付の図面および以下の説明に記載される。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らかになろう。

【図面の簡単な説明】

【００３６】

【図１】弁別器モデルの集合を含む例示的なアクション選択システムのブロック図である。

【図２】弁別器モデルの集合によって生成される予測出力同士の間の不一致の尺度を決定するための例示的な技法を示す図である。

【図３】アクション選択ニューラルネットワークをトレーニングするための例示的なプロセスの流れ図である。

【図４Ａ】アクション選択システムを使用して達成される例示的な実験結果を示す図である。

【図４Ｂ】アクション選択システムを使用して達成される例示的な実験結果を示す図である。

【発明を実施するための形態】

【００３７】

同様の数字および記号は、様々な図面において同様の要素を示す。

【００３８】

図1は、環境106と対話しているエージェント104を制御し得る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いて1つまたは複数のコ

ンピュータ上で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。

【００３９】

システム100は、目標を達成するために時間ステップのシーケンスにわたって環境106と対話するためにエージェント104によって実施されるべきアクション102を選択し得る。各時間ステップにおいて、システム100は、環境106の現状を特徴付けるデータを受信し、受信されたデータに応答して、エージェント104によって実施されるべきアクション102を選択し得る。環境106の状態を特徴付けるデータは、本明細書において観測110と呼ばれ、たとえば、画像、または任意の他の適切なデータを含み得る。場合によっては、エージェント104は、たとえば、ロボットであってよく、観測110は、たとえば、ジョイント位置、速度およびトルク、または以下でより詳細に説明するような任意の他の適切なデータを含み得る。各時間ステップにおいて、時間ステップにおける環境106の状態(観測110によって特徴付けられる)は、前の時間ステップにおける環境106の状態、および前の時間ステップにおいてエージェント104によって実施されたアクション102に依存する。

【００４０】

システム100は、潜在変数118を条件とするアクション選択ニューラルネットワーク120を使用して、時間ステップのシーケンスにわたって環境106と対話するためにエージェント104によって実施させるべきアクション102を選択し得る。システム100は、可能な潜在変数のセットからの潜在変数118を(たとえば、ランダムに)サンプリングし、アクション選択ニューラルネットワーク120を潜在変数118で条件付け得る。次いで、システム100は、時間ステップのシーケンスにわたってエージェント104によって実施されるべきアクション102を選択するために、潜在変数118を条件とするアクション選択ニューラルネットワーク120を使用し得る。概して、アクション選択ニューラルネットワーク120を潜在変数118で「条件付けること」は、入力として潜在変数118をアクション選択ニューラルネットワーク120に提供することを指すことがある。いくつかの実装形態では、可能な潜在変数のセットは、有限に多くの可能な潜在変数、たとえば、10、100、1000、10,000、または任意の他の適切な数の潜在変数のみを含み得る。いくつかの実装形態では、可能な潜在変数のセットは、無限に多くの可能な潜在変数を含んでよく、たとえば、可能な潜在変数のセットは、連続範囲[0,1]であってよい。

【００４１】

各潜在変数は、「スキル」、たとえば、時間ステップのシーケンスにわたるエージェント104の挙動を特徴付けるアクションのセット、を表し得る。すなわち、各潜在変数は、環境106の状態を一貫した認識可能な方法で時間ステップのシーケンスにわたって改変させるように環境106と対話するようにエージェント104を促し得る。言い換えれば、「スキル」は、アクション選択ニューラルネットワーク120を潜在変数118で条件付けることによって定義されるアクション選択ポリシーを指すことがある。たとえば、スキルは、アクションaにわたって状態sおよび潜在変数zを分布にマッピングするアクション選択ポリシー$\pi_\theta(a \mid s,z)$を指すことがあり、ここで、θは、アクション選択ニューラルネットワーク120のパラメータのセットである。以下でより詳細に説明するように、教師なし報酬109に基づいてアクション選択ニューラルネットワーク120をトレーニングすることによって、システム100は、アクション選択ニューラルネットワーク120が別個の認識可能なスキルを学習することを可能にし得る。

【００４２】

特定の例として、エージェント104がロボットアームを有する物理的なロボットである場合、第1の潜在変数は、前方に、後方に、左に、および右に移動する動作を含む、ロボットの可能なアクションの第1のセットを定義し得る。第2の潜在変数は、ロボットアームを前方に、後方に、左に、および右に移動するアクションを含むロボットの可能なアクションの第2のセットを定義し得る。この場合、第1の潜在変数および第2の潜在変数は各々、ロボットの別個の認識可能なスキルを表し得る。

【００４３】

アクション選択ニューラルネットワーク120を潜在変数118で条件付けた後、システム

100は、時間ステップのシーケンスにわたって環境106と対話するためにエージェント104によって実施されるべきアクション102を選択するために、アクション選択ニューラルネットワーク120を使用し得る。たとえば、各時間ステップにおいて、システム100は、アクション選択出力122を生成するために、アクション選択ニューラルネットワーク120を使用して、時間ステップにおいて環境106の現状を特徴付ける観測110および潜在変数118を処理し得る。いくつかの実装形態では、時間ステップにおける環境106の現状を特徴付ける観測110を処理することに加えて、システム100は、各々が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付ける1つまたは複数の観測をやはり処理し得る。

【００４４】

アクション選択出力122は、エージェント104によって実施され得る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スコアを含み得る。いくつかの実装形態では、システム100は、時間ステップにおいてエージェント104によって実施されるべきアクションとして、アクション選択出力122に従って、最高スコアを有するアクションを選択し得る。いくつかの実装形態では、システム100は、探求戦略に従って、エージェント104によって実施されるべきアクション102を選択する。たとえば、システム100は、ε-グリーディー探索戦略を使用し得る。この例では、システム100は、確率1-εを有する(アクション選択出力122による)最高スコアを選択し、確率εを有するアクションをランダムに選択し得、ここで、εは0と1との間の数である。

【００４５】

システム100が時間ステップにおいてエージェント104によって実施されるべきアクション102を選択した後、エージェント104は、アクション102を実施することによって環境106と対話し、システム100は、対話に基づいて報酬、たとえば、タスク報酬108、教師なし報酬109、または両方、を受け得る。

【００４６】

たとえば、各時間ステップにおいて、システム100は、環境106の現状および時間ステップにおけるエージェント104のアクション102に基づいて、タスク報酬108を受け得る。概して、タスク報酬108は、数値として表されてよい。タスク報酬108は、環境106内の何らかのイベントまたは側面に基づき得る。たとえば、タスク報酬108は、エージェント104がタスク(たとえば、環境において物体を所望の方法で物理的に操作した)を達成したかどうか、またはタスクの達成に向けたエージェント104の進歩を示し得る。

【００４７】

追加または代替として、システム100は、教師なし報酬109を受け得る。システム100は、複数の弁別器モデル150の集合を使用して、1つまたは複数の時間ステップの各々に対してそれぞれの教師なし報酬109を決定し得る。集合内の各弁別器モデル150は、環境106の現状を特徴付ける観測110を処理し、環境106を観測110によって特徴付けられる状態に入らせるためにアクション選択ニューラルネットワーク120がどの潜在変数118を条件としたかを予測するそれぞれの予測出力を生成するように構成され得る。言い換えれば、各弁別器モデルは、アクション選択出力122を生成するためにアクション選択ニューラルネットワーク120が使用したスキルに関する予測を生成し得る。

【００４８】

弁別器モデルの予測出力は、可能な潜在変数のセット内の各潜在変数のそれぞれのスコアを定義する、可能な潜在変数のセットにわたるスコア分布を含み得る。たとえば、時間ステップtにおいて、パラメータ$\phi_i$のセットを有する弁別器モデルiは、可能な潜在変数$z_t$のセットにわたってスコア分布

【００４９】

【数１】

$$q_{\varphi_i}(z_t \mid o_t)$$

【００５０】

を生成するために、環境106の現状を特徴付ける観測$o_t$を処理し得る。

【００５１】

概して、時間ステップにおいて弁別器モデルによって処理される入力は、(i)時間ステップにおける観測、また随意に(ii)1つまたは複数の前の時間ステップに対する観測の表現を含み得る。場合によっては、弁別器モデルは、観測を直接処理し得る。他の場合には、弁別器モデルは、たとえば、オートエンコーダニューラルネットワークのエンコーダサブネットワークを使用して観測を処理することによって生成される観測の特徴表現を処理し得る。

【００５２】

図2を参照しながら以下でより詳細に説明するように、システム100は、時間ステップに対する教師なし報酬109を決定するために、弁別器モデル150の予測出力を使用し得る。たとえば、1つまたは複数の時間ステップの各々において、システム100は、複数の弁別器モデル150の集合によって生成される予測出力同士の間の不一致の尺度を決定し、不一致の尺度に基づいて教師なし報酬109を決定し得る。(弁別器モデルの集合によって生成される予測出力同士の間の「不一致の尺度」は、弁別器モデルの集合によって生成される予測出力同士の間の相違点の尺度を指す。不一致の尺度のいくつか例については以下でより詳細に説明する)。いくつかの実装形態では、システム100は、加えて、弁別器モデルによって生成される予測出力の精度に基づいて、教師なし報酬を決定し得る。精度は、たとえば、アクション選択ニューラルネットワーク120が時間ステップにおいて学習したスキルの有効数を反映し得る。特定の例として、時間ステップにおける教師なし報酬rは、

$$r=r_{skill}+\lambda r_{DIS} \quad (1)$$

として表されてよく、式中、$r_{skill}$は、時間ステップにおいて弁別器モデルの集合によって生成される予測出力の精度を特徴付ける報酬条件であり、$r_{DIS}$は、時間ステップにおいて弁別器モデル150の集合によって生成される予測出力同士の間の不一致の尺度に基づく報酬条件であり、λは、同調可能重み係数である。報酬条件$r_{skill}$は、たとえば、

$$r_{skill}=\log q_\phi(z_t \mid \varrho)-\log p(z_t) \quad (2)$$

の形をとってよく、式中、$q_\phi(z_t \mid \varrho)$は、弁別器モデルの集合が、アクション選択ニューラルネットワーク120が条件とした潜在変数118(たとえば、正確な潜在変数)に対して予測する平均スコアである。言い換えれば、$q_\phi(z_t \mid \varrho)$は、弁別器モデル150の集合によって生成される予測出力の精度を表す(特徴付ける)。

【００５３】

式(2)において、$p(z_t)$は、時間ステップにおける可能な潜在変数$z_t$のセットにわたる前のスコア分布である。いくつかの実装形態では、システム100は、前のスコア分布$p(z_t)$に従って、潜在変数118をサンプリングし得る。たとえば、潜在変数のセットにわたる前のスコア分布が一様である場合、システム100は、等しい確率で潜在変数をサンプリングし得る。

【００５４】

いくつかの実装形態では、潜在変数$p(z_t)$のセットにわたる前のスコア分布は、合計N個のスキルで一様になるように固定される。そのような場合、弁別器モデルが、アクション選択ニューラルネットワーク120が条件とした潜在性を一様に予測する予測出力を生成する場合、$\log q_\phi(z_t \mid \varrho)=-\log N$であり、$r_{skill}=0$になる。これは、アクション選択ニューラルネットワーク120が学習したスキルの推定数が低いこと、またはごく少ないことを示し得、これは、低い報酬、またはごく少数の報酬に対応する。他方で、弁別器モデルが、アクション選択ニューラルネットワーク120が条件とした潜在性を正確に予測する場合、$\log q_\phi(z_t \mid \varrho)=0$であり、$r_{skill}=\log N$になる。これは、学習されたスキルの推定数が比較的高いことを示し得、高い報酬をもたらす。

【００５５】

上記で説明したように、システム100は、たとえば、上記の式(1)において$r_{DIS}$を決定することによって、弁別器モデル150の集合によって生成される予測出力同士の間の不一致の尺度にやはり基づいて、教師なし報酬を生成し得る。このプロセスについては、図2を

参照しながら以下でより詳細に説明する。

【００５６】

概して、弁別器モデル150によって生成される予測同士の間の高い尺度の不一致は、エージェント104が時間ステップにおける環境106、たとえば、前に訪問された可能性が高い環境106の一部、の「新しい」状態に遭遇したことを示し得る。言い換えれば、高い尺度の不一致は、弁別器モデルが、環境の状態を時間ステップにおける観測110によって特徴付けられる状態に改変させるためにアクション選択ニューラルネットワーク120がどの潜在変数118を条件としたかに関して実質的に一致しないことを示し得る。環境106の異なる部分の探求を奨励するために、システム100は、それに応じて、上記の式(1)において増大した$r_{DIS}$値をもたらすアクションを選択し、それにより、環境106の新しい状態に遭遇するようにアクション選択ニューラルネットワーク120に報酬を与えることができる。

【００５７】

対照的に、低い尺度の不一致は、弁別器モデル150のうちのいくつか、またはすべてがその予測について一致することを示し、それにより、エージェント104が時間ステップにおいて観測110によって表される環境106の部分を前に訪問した比較的高い可能性が存在することを示し得る。そのような場合、システム100は、上記の式(1)において比較的低い$r_{DIS}$値を生成し得る。式(1)に従って教師なし報酬109を生成し、教師なし報酬109に基づいてアクション選択ニューラルネットワーク120をトレーニングすることによって、システム100は、より区別可能な、認識可能なスキルを学習し、環境106の新しい状態を探求するようにアクション選択ニューラルネットワーク120を奨励し得る。

【００５８】

上記で説明したように、教師なし報酬109を受けることに加えて、システムはまた、目標の達成に向けたエージェント104の進歩を特徴付けるタスク報酬108を受け得る。たとえば、エージェント104が時間ステップにおいてタスクを達成した場合、タスク報酬は正の値(たとえば、1)であってよい。さもなければ、報酬はゼロであってよい。いくつかの実装形態では、各時間ステップにおいて、アクション選択システム100は、タスク報酬108および教師なし報酬109に基づいて、報酬全体をさらに決定し得る。時間ステップに対する報酬全体は、たとえば、タスク報酬108と教師なし報酬109の線形結合であってよい。いくつかの実装形態では、システム100は、時間ステップのシーケンスの各時間ステップまたは時間ステップのシーケンスの1つまたは複数の時間ステップにおいてのみ教師なし報酬109を生成し得る。たとえば、システム100は、最終的な時間ステップにおいてのみ、たとえば、エピソードの終結においてのみ、教師なし報酬109を生成し得、ここで、「エピソード」は、アクション選択ニューラルネットワーク120が同じ潜在変数118を条件とする時間ステップのシーケンスを指す。

【００５９】

アクション選択ニューラルネットワーク120は、アクション選択ニューラルネットワーク120がその説明する機能、たとえば、観測110に応答して、エージェント103によって実施されるべきアクション102を特徴付けるアクション選択出力122を生成するために、潜在性118および観測110を処理すること、を実施すること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有し得る。たとえば、アクション選択ニューラルネットワーク120は、任意の適切なタイプの(たとえば、全結合層、畳み込み層、アテンション層、変換層、など)の任意の適切な構成で接続された(たとえば、層の線形シーケンスとして)任意の適切な数の層(たとえば、5個の層、10個の層、または25個の層)を含み得る。

【００６０】

弁別器モデル150の集合内の各弁別器モデルは、任意の適切な機械学習モデルアーキテクチャ、たとえば、ニューラルネットワークアーキテクチャ、サポートベクター機械アーキテクチャ、またはランダムフォレストアーキテクチャを有し得る。いくつかの実装形態では、各弁別器モデルは、同じアーキテクチャ(たとえば、同じニューラルネットワークアーキテクチャ)を共有するが、たとえば、トレーニング例の単独でサンプリングされたバッ

チに基づいてトレーニングされている結果として、または異なって初期化されたパラメータ値を有する結果として、弁別器モデルパラメータのそれぞれの値は、各弁別器モデルに対して異なる。複数の弁別器モデル150の集合は、任意の適切な数の弁別器モデル、たとえば、2個、10個、100個、1000個、または任意の他の適切な数の弁別器モデルを含み得る。

【００６１】

トレーニングエンジン112は、強化学習技法を使用して、式(1)を参照しながら上記で説明した教師なし報酬109に基づいて、アクション選択ニューラルネットワーク120をトレーニングに得る。トレーニングエンジン112は、アクション選択ニューラルネットワーク120のパラメータを反復的に調整することによって、アクション選択ニューラルネットワーク120をトレーニングする。トレーニングエンジン112は、アクション選択ニューラルネットワーク120を通じて強化学習目的関数(たとえば、Q学習目的関数、ポリシー勾配目的関数、または任意の他の適切な強化学習目的関数)の勾配を反復的に逆伝搬することによって、アクション選択ニューラルネットワーク120のパラメータを調整し得る。アクション選択ニューラルネットワーク120をトレーニングすることによって、トレーニングエンジン112は、アクション選択ニューラルネットワーク120にアクション選択ニューラルネットワーク120によって受ける教師なし報酬109(たとえば、長期時間割引累積的全報酬)の累積尺度を増大させるアクションを選択させ得る。いくつかの実装形態では、トレーニングエンジン112は、上記で説明した強化学習技法を使用して、報酬全体、たとえば、教師なし報酬109とタスク報酬108の組合せに基づいて、アクション選択ニューラルネットワーク120をトレーニングし得る。

【００６２】

さらに、トレーニングエンジン112は、任意の適切なトレーニング技法、たとえば、教師あり学習技法を使用して、複数の弁別器モデル150の集合をトレーニングし得る。複数の弁別器モデル150の集合内の各弁別器モデルは、弁別器モデルパラメータのそれぞれのセットを有し得る。いくつかの実装形態では、トレーニングエンジン112は、集合内の各弁別器モデル150に対して異なるように、弁別器モデルパラメータのそれぞれのセットのそれぞれの値を初期化し得る。たとえば、トレーニングエンジン112は、分布、たとえば、通常の正規分布、から各パラメータをランダムにサンプリングすることによって、パラメータを初期化し得る。

【００６３】

いくつかの実装形態では、トレーニングエンジン112は、リプレイメモリ114内に記憶されたトレーニング例を使用して、弁別器モデル150の集合をトレーニングし得る。リプレイメモリ114は、たとえば、論理データ記憶エリアまたは物理データ記憶デバイスとして実装され得る。リプレイメモリ114は、複数の前の時間ステップの各々に対応するそれぞれの「経験タプル」を記憶し得る(たとえば、メモリ114は、現在の時間ステップの前の各時間ステップに対するそれぞれの経験タプルを記憶し得る)。時間ステップに対する経験タプルは、前の時間ステップにおける環境106とのエージェント104の対話を特徴付けるデータを指す。たとえば、前の時間ステップに対する経験タプルは、(i)環境106とのエージェント104の前の対話中の環境106の状態を特徴付けるトレーニング観測110と、(ii)環境106を観測110によって特徴付けられる状態に入らせるためにアクション選択ニューラルネットワーク120が条件とした潜在変数118を定義するターゲット出力とを含み得る。

【００６４】

各トレーニング反復において、トレーニングエンジン112は、リプレイメモリ114からのトレーニング例のバッチをサンプリングし、弁別器モデル150の集合を使用して、トレーニング例によって指定されたトレーニング入力を処理し得る。いくつかの実装形態では、トレーニングエンジン112は、トレーニング例の単独でサンプリングされたバッチに基づいて集合内の各弁別器モデル150をトレーニングし得、たとえば、トレーニングエンジン112は、リプレイメモリ114からのトレーニング例のそれぞれのバッチを単独でサンプ

リングし、各バッチによって指定されたトレーニング入力をそれぞれの弁別器モデルに提供し得る。

【００６５】

各弁別器モデルは、可能な潜在変数のセットにわたってそれぞれのスコア分布を含む、対応する予測出力を生成するために、弁別器モデルパラメータのそれぞれのセットに従って、それぞれのトレーニング例によって指定されたそれぞれのトレーニング観測を処理し得る。

【００６６】

各弁別器モデルに対して、トレーニングエンジン112は、それぞれの弁別器モデルパラメータに関する目的関数の勾配を決定し得る。目的関数は、(i)予測出力によって定義されたスコア分布と(ii)アクション選択ニューラルネットワーク120が条件とした潜在変数118を定義するターゲット出力との間のエラー、たとえば、クロスエントロピーエラー、を測定し得る。特定の例として、トレーニングエンジン112は、目的関数を最小化することによって、各弁別器モデルの弁別器モデルパラメータφをトレーニングし得る:

【００６７】

【数２】

$$L(\phi) = -\mathbb{E}[log q_{\varphi}(z \mid o(\tau)] \qquad (3)$$

式中、$q_{\varphi}(z \mid o(\tau))$

【００６８】

は、弁別器モデルによって生成される可能な潜在変数のセットにわたるスコア分布である。各弁別器モデルに対して、トレーニングエンジン112は、たとえば、逆伝搬技法を使用して、弁別器モデルパラメータに関する目的関数の勾配を決定し得る。トレーニングエンジン112は、勾配降下最適化アルゴリズム、たとえば、AdamまたはRMSpropの更新規則を使用して、弁別器モデルパラメータを更新するために勾配を使用し得る。

【００６９】

アクション選択ニューラルネットワーク120、弁別器モデル150の集合、または両方をトレーニングした後、システム100は、環境106と対話するためにエージェント104によって実施されるべきアクション102を選択するために、アクション選択ニューラルネットワーク120を使用し得る。

【００７０】

可能なエージェント、環境、およびタスクの例について、次により詳細に説明する。

【００７１】

いくつかの実装形態では、環境は現実世界環境であり、エージェントは、現実世界環境と対話している機械エージェントである。たとえば、エージェントは、目標(またはタスク)を達成するために、たとえば、環境内の関心の物体を位置特定するために、関心の物体を環境内の指定されたロケーションに移動させるために、関心の物体を環境内で指定された方法で物理的に操作するために、または環境内の指定された目的地にナビゲートするために、環境と対話しているロボットであってよく、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じてナビゲートしている自律型または半自律型の陸上車両、飛行機、または船舶であってもよい。

【００７２】

これらの実装形態では、観測は、たとえば、エージェントが環境と対話するときの観測をキャプチャするための画像、物体の位置データ、およびセンサーデータのうちの1つまたは複数、たとえば、画像、距離、もしくは位置センサーまたはアクチュエータからのセンサーデータを含み得る。

【００７３】

たとえば、ロボットの場合、観測は、ロボットの現状を特徴付けるデータ、たとえば、ロボットによって保持されている部材のジョイント位置、ジョイント速度、ジョイント力、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、および全体的または相対的なポーズのうちの1つまたは複数を含み得る。

【００７４】

ロボットまたは他の機械エージェントもしくは車両の場合、観測は、エージェントの1つまたは複数の部分の位置、線形速度または角速度、力、トルクまたは加速度、および全体的または相対的なポーズのうちの1つまたは複数を同様に含み得る。観測は、1次元、2次元、または3次元で定義されてよく、絶対的および/または相対的観測であってよい。

【００７５】

観測はまた、たとえば、現実世界環境を感知する1つまたは複数のセンサーデバイスによって取得されるデータ、たとえば、モーター電流または温度信号などの感知された電子信号、および/または、たとえば、カメラまたはLIDARセンサーからの画像またはビデオデータ、たとえば、エージェントのセンサーからのデータまたは環境内でエージェントから離れて別に位置するセンサーからのデータを含み得る。

【００７６】

電子エージェントの場合、観測は、電流、電圧、電力、温度、および電子機器および/または機械機器の機能を表す他のセンサーおよび/または電子信号など、工場またはサービス施設の部分を監視する1つまたは複数のセンサーからのデータを含み得る。電子エージェントは、たとえば、産業設備内の機械、またはデータセンター内の冷却システムを制御するアクションを生成し得る。

【００７７】

アクションは、ロボットまたは他の機械エージェントを制御するための制御信号、たとえば、ロボットのジョイントもしくはより高いレベルの制御コマンド、または自律型または半自律型の陸上車両もしくは飛行機または船舶用のトルク、たとえば、車両の制御面もしくは他の制御要素またはより高いレベルの制御コマンドに対するトルクであってよい。

【００７８】

制御信号は、たとえば、ロボットの1つまたは複数のジョイントまたは別の機械エージェントの部分に関する位置、速度、または力/トルク/加速度データを含み得る。制御信号は、さらにまたはその代わりに、モーター制御データなどの電子制御データ、またはより一般的に、その制御が環境の観測状態に影響を与える、環境内で1つまたは複数の電子デバイスを制御することに関するデータを含み得る。たとえば、自律的または半自律的な陸上車両、飛行機、または船舶の場合、制御信号は、ナビゲーションを制御するアクション、たとえば、車両のステアリング、および移動、たとえば、制動および/または加速を定義し得る。

【００７９】

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対話している1つまたは複数のコンピュータとして実装される。いくつかの実装形態では、環境は、上記の現実世界環境のシミュレーションである。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ってよく、強化学習システムは、シミュレーションに基づいてトレーニングされ、次いで、トレーニングされると、現実世界で使用され得る。

【００８０】

たとえば、シミュレートされた環境は、動きシミュレーション環境、たとえば、運転シミュレーションまたは飛行シミュレーションであってよく、エージェントは、動きシミュレーションを通じてナビゲートしている、シミュレートされた車両であってよい。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ってよい。

【００８１】

別の例では、シミュレートされた環境は、ビデオゲームであってよく、エージェントは

、ビデオゲームをプレイしている、シミュレートされたユーザであってよい。
【００８２】
いくつかの実装形態では、環境は、化学製品、生物学的製品、もしくは機械製品、または食料品など、製品を製造するための現実世界製造環境である。本明細書で使用する、製品を「製造する」は、製品を作成するために開始材料を精製すること、または、清潔なまたは再生された製品を生成するために、たとえば、汚染物質を除去するために、開始材料を処理することを含む。製造工場は、化学物質または生物物質用の容器など、複数の製造ユニット、または固体材料または他の材料を処理するための機械、たとえばロボット、を含み得る。製造ユニットは、製品の中間バージョンまたは構成要素が製品の製造中に、たとえば、パイプまたは機械運搬を介して、製造ユニット間で移動可能であるように構成される。本明細書で使用する、製品の製造はまた、台所ロボットによる食品の製造を含む。
【００８３】
エージェントは、製造ユニットを制御するように構成された電子エージェント、または、製品を製造するために動作する、ロボットなどの機械を含み得る。すなわち、エージェントは、化学製品、生物学的製品、または機械製品の製造を制御するように構成された制御システムを含み得る。たとえば、制御システムは、製造ユニットまたは製造機械のうちの1つまたは複数を制御するように、または製造ユニットまたは製造機械同士の間の製品の中間バージョンまたは構成要素の移動を制御するように、構成され得る。
【００８４】
一例として、エージェントによって実施されるタスクは、製品、またはその中間バージョンまたは構成要素を製造するためのタスクを含み得る。別の例として、エージェントによって実施されるタスクは、電力消費、もしくは水の消費、または製造プロセスにおいて使用される任意の材料または消耗品の消費を制御するタスクなど、リソースの使用を制御する、たとえば、最小限に抑える、ためのタスクを含み得る。
【００８５】
アクションは、製品、またはその中間生成物または構成要素を製造するために固体材料または液体材料を処理するための機械または製造ユニットの使用を制御するための、または、たとえば、製造ユニットまたは機械同士の間で、製造環境内で製品の中間バージョンまたは構成要素の移動を制御するための制御アクションを含み得る。概して、アクションは、環境の観測される状態に影響を及ぼす任意のアクション、たとえば、以下で説明する感知されたパラメータのうちのいずれかを調整するように構成されたアクション、であってよい。これらは、製造ユニットの物理的状態または化学的状態を調整するためのアクション、または機械の機械部分またはロボットのジョイントの移動を制御するためのアクションを含み得る。アクションは、製造ユニットまたは製造機械に動作条件を課すアクション、または製造ユニットまたは製造機械の動作を調整するため、制御するため、またはオンまたはオフに切り替えるための設定を変更させるアクションを含み得る。
【００８６】
報酬またはリターンは、タスクのパフォーマンスのメトリックに関係し得る。たとえば、製品を製造するためのタスクの場合、メトリックは、製造されている製品の量、製品の品質、製品の製造速度、または製造タスクを実施する物理的コストのメトリック、たとえば、タスクを実施するために使用されるエネルギー、材料、または他のリソースの量のメトリックを含み得る。リソースの使用を制御するためのタスクの場合、メトリックは、リソースの使用量の任意のメトリックを含み得る。
【００８７】
概して、環境の状態の観測は、電子機器および/または機械機器の機能を表す任意の電子信号を含み得る。たとえば、環境の状態の表現は、製造環境の状態を感知するセンサー、たとえば、製造ユニットまたは製造機械の状態または構成を感知するセンサー、または製造ユニットまたは製造機械同士の間の材料の移動を感知するセンサーによって行われる観測から導出され得る。
【００８８】

いくつかの例として、そのようなセンサーは、機械的な移動または力、圧力、温度;電流、電圧、周波数、インピーダンスなどの電気条件;1つまたは複数の材料の量、レベル、フロー/移動レートまたはフロー/移動経路;物理的または化学的条件、たとえば、物理的状態、形状もしくは構成、またはpHなどの科学的状態;ユニットまたは機械の機械的構成などのユニットまたは機械の構成、またはバルブ構成;製造ユニットのまたは機械の画像観測またはビデオ観測、または移動をキャプチャするための画像センサーまたはビデオセンサー;または任意の他の適切なタイプのセンサーを感知するように構成され得る。ロボットなどの機械の場合、センサーからの観測は、位置、線形速度もしくは角速度、力、トルクもしくは加速度、または機械の1つまたは複数の部分のポーズの観測、たとえば、機械またはロボットの、または機械またはロボットによって保持されるかまたは処理される物品の現状を特徴付けるデータを含み得る。観測はまた、たとえば、モーター電流または温度信号など、感知された電子信号、またはたとえば、カメラまたはLIDARセンサーからの画像データまたはビデオデータを含み得る。これらのようなセンサーは、環境内のエージェントの部分であってよく、またはそこから離れてエージェントとは別に位置してもよい。

【００８９】

いくつかの他の適用例では、エージェントは、たとえば、データセンターまたはグリッド主電源もしくは水分配システム内のまたは製造工場またはサービス施設内の機器を含む現実世界環境内のアクションを制御し得る。観測は、その場合、工場または施設の動作に関係し得る。たとえば、観測は、機器による電力または水の使用量の観測、もしくは電力の生成または分配制御の観測、またはリソースの使用量のまたは廃棄物の観測を含み得る。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、効率増大の目標を達成するために、および/または、たとえば、廃棄物を削減することによって、環境における動作の環境的な影響を低減するために、環境においてアクションを制御し得る。アクションは、工場/施設の機器に対する動作条件を制御するまたは課すアクション、および/または、たとえば、工場/施設の構成要素を調整するためまたはオン/オフにするために工場/施設の動作の設定を変更させるアクションを含み得る。

【００９０】

いくつかの実装形態では、環境は、サーバファームまたはデータセンター、たとえば、電気通信データセンター、もしくはデータを記憶および処理するためのコンピュータデータセンター、または任意のサービス施設など、複数の電子機器を備えたサービス施設の現実世界環境である。サービス施設はまた、機器の動作環境を制御する補助制御機器、たとえば、温度制御などの環境制御機器、たとえば、冷却機器、または空気流機器もしくは空調機器を含み得る。タスクは、電力消費または水消費を制御するためのタスクなど、リソースの使用を制御するための、たとえば、最小限に抑えるための、タスクを含み得る。エージェントは、機器の動作を制御するように、または補助、たとえば、環境、制御機器の動作を制御するように構成された電子エージェントを含み得る。

【００９１】

概して、アクションは、環境の観測される状態に対して影響を及ぼす任意のアクション、たとえば、以下で説明する感知されたパラメータのうちのいずれかを調整するように構成されたアクションであってよい。これらは、機器または補助制御機器を制御するための、またはそれに動作条件を課すためのアクション、たとえば、機器または補助制御機器の動作を調整する、制御する、またはオンまたはオフに切り替えるための設定を変更させるアクションを含み得る。

【００９２】

概して、環境の状態の観測は、施設の機能または施設の機器を表す任意の電子信号を含み得る。たとえば、環境の状態の表現は、施設の物理的環境の状態を感知している任意のセンサーによって行われる観測、または1つまたは複数の機器もしくは1つまたは複数の補助制御機器の状態を感知している任意のセンサーによって行われる観測から導出され得る。これらは、電流、電圧、電力またはエネルギーなど、電気的条件;施設の温度; 施設内のまたは施設の冷却システム内の液体フロー、温度または圧力;またはベントが開いているか

否かなど、物理的な施設構成を感知するように構成されたセンサーを含む。
【００９３】
報酬またはリターンは、タスクのパフォーマンスのメトリックに関係し得る。たとえば、電力または水の使用を制御するためのタスクなど、リソースの使用を制御するための、たとえば、最小限に抑えるためのタスクの場合、メトリックは、リソースの使用の任意のメトリックを含み得る。
【００９４】
いくつかの実装形態では、環境は、発電施設の現実世界環境、たとえば、太陽光発電所または風力発電所など、再生可能発電施設である。タスクは、たとえば、需要を満たすために、もしくはグリッドの素子同士の間のミスマッチのリスクを低減するために、または施設によって生成される電力を最大化するために、施設によって生成される電力を制御するための、たとえば、配電グリッドへの電力の供給を制御するための、制御タスクを含み得る。エージェントは、施設による電力の生成または生成された電力のグリッドへの結合を制御するように構成された電子エージェントを含み得る。
【００９５】
アクションは、たとえば、風力タービンのまたは1つまたは複数のソーラーパネルもしくはミラーの構成を制御するための、1つまたは複数の再生可能発電素子の電気的構成または機械的構成、または回転発電機械の電気的構成または機械的構成など、発電機の電気的構成または機械的構成を制御するためのアクションを含み得る。機械的制御アクションは、たとえば、エネルギー入力から電気エネルギー出力への変換、たとえば、エネルギー入力の電気エネルギー出力への変換またはその結合程度の効率性を制御するアクションを含み得る。電気的制御アクションは、たとえば、生成された電力の電圧、電流、周波数、または位相のうちの1つまたは複数を制御するアクションを含み得る。
【００９６】
報酬またはリターンは、タスクのパフォーマンスのメトリックに関係し得る。たとえば、配電グリッドへの電力の供給を制御するためのタスクの場合、基準は、伝達される電力の尺度、もしくは電圧、電流、周波数、または位相のミスマッチなど、発電施設とグリッドとの間の電力的ミスマッチの尺度、または発電施設内の電力またはエネルギー損失の尺度に関係し得る。配電グリッドへの電力の供給を最小限に抑えるためのタスクの場合、メトリックは、グリッドに伝達される電力またはエネルギーの尺度、または発電施設内の電力またはエネルギー損失の尺度に関係し得る。
【００９７】
概して、環境の状態の観測は、発電施設内の発電機器の電気的機能または機械的機能を表す任意の電気信号を含み得る。たとえば、環境の状態の表現は、電力を生成している発電施設内の機器の物理的状態または電気的状態、もしくはそのような機器の物理的環境、または発電機器をサポートしている補助機器の状態を感知している任意のセンサーによって行われる観測から導出され得る。そのようなセンサーは、電流、電圧、電力、またはエネルギーなどの機器の電気的状態;物理的環境の温度または冷却;液体フロー;または機器の物理的構成;および、たとえば、ローカルセンサーまたはリモートセンサーからのグリッドの電気的状態の観測を感知するように構成されたセンサーを含み得る。環境の状態の観測はまた、将来の風力レベルまたは太陽放射照度の予測またはグリッドの将来の電気的状態の予測など、発電機器の動作の将来の状態に関する1つまたは複数の予測を含み得る。
【００９８】
別の例として、環境は、各状態がタンパク質鎖のまたは1つまたは複数の中間生成物または前駆体化学物質のそれぞれの状態であるように、化学合成または環境を包み込むタンパク質であってよく、エージェントは、タンパク質鎖をどのように包み込むかまたは化学物質をどのように合成するかを決定するためのコンピュータシステムである。この例では、アクションは、タンパク質鎖を包み込むための可能な包み込みアクションまたは前駆体化学物質/中間生成物を集めるためのアクションであり、達成すべき結果は、たとえば、タンパク質が安定するように、またそれが特定の生物学的機能を達成するように、タンパク

質を包み込むこと、または化学物質に有効な合成ルートを提供することを含み得る。別の例として、エージェントは、人間の対話なしに自動的にシステムによって選択されるタンパク質包み込みアクションまたは化学合成ステップを実施または制御する機械的エージェントであってよい。観測は、タンパク質または化学的/中間生成物/前駆体の状態の直接的または間接的な観測を含んでよく、かつ/またはシミュレーションから導出されてもよい。

【００９９】

同様の方法で、環境は、各状態が潜在的な薬剤のそれぞれの状態であり、エージェントが医薬品の要素および/または医薬品のための合成経路を決定するためのコンピュータシステムであるように、医薬品設計環境であってよい。薬物/合成は、たとえば、シミュレーションにおいて、薬物に対するターゲットから導出される報酬に基づいて設計され得る。別の例として、エージェントは、薬物の合成を実施または制御する機械的エージェントであってよい。

【０１００】

いくつかのさらなる適用例では、環境は現実世界環境であり、エージェントは、たとえば、モバイルデバイス上のかつ/またはデータセンター内のコンピューティングリソースにわたってタスクの分布を管理する。これらの実装形態では、アクションは、タスクを特定のコンピューティングリソースに割り当てることを含んでよく、達成すべき目標は、指定されたコンピューティングリソースを使用してタスクのセットを完了するために必要とされる時間を最小限に抑えることを含み得る。

【０１０１】

さらなる例として、アクションは、広告を提示することを含んでよく、観測は、広告表示回数またはクリックスルーカウントまたはレートを含んでよく、報酬は、1人または複数のユーザによってとられたアイテムまたはコンテンツの前の選択を特徴付け得る。この例では、達成すべき目標は、1人または複数のユーザによるアイテムまたはコンテンツの選択を最大化することを含んでよい。

【０１０２】

場合によっては、観測は、第三者(たとえば、エージェントのオペレータ)によってエージェントに提供されるテキスト命令または発話命令を含み得る。たとえば、エージェントは自律車両であってよく、自律的車両のユーザは、(たとえば、特定のロケーションにナビゲートするために)テキスト命令または発話命令をエージェントに提供し得る。

【０１０３】

別の例として、環境は、電気的、機械的、または電気機械的な設計環境、たとえば、電気的、機械的または電気機械的なエンティティの設計がシミュレートされる環境であってよい。シミュレートされる環境は、エンティティが作業することが意図される現実世界環境のシミュレーションであってよい。タスクは、エンティティを設計することであってよい。観測は、エンティティを特徴付ける観測、すなわち、機械的形状の、あるいはエンティティの電気的、機械的、または電気機械的の構成の観測、またはエンティティのパラメータまたは属性の観測を含み得る。

【０１０４】

アクションは、エンティティを修正する、たとえば、観測のうちの1つまたは複数を修正するアクションを含み得る。報酬またはリターンは、エンティティの設計のパフォーマンスの1つまたは複数のメトリックを含み得る。たとえば、報酬またはリターンは、重み、もしくは強度など、エンティティの1つまたは複数の物理的特徴に、またはエンティティが設計された特定の機能を実施することにおける効率の尺度など、エンティティのまたは1つまたは複数の電気的特徴に関係し得る。設計プロセスは、製造用の設計を、たとえば、エンティティを製造するためのコンピュータ実行可能命令の形で出力することを含み得る。プロセスは、設計に従ってエンティティを作成することを含み得る。したがって、エンティティの設計は、たとえば、強化学習、次いで、たとえば、コンピュータ実行可能命令として、エンティティを製造するための最適化された設計出力によって最適化され得、次いで、最適化された設計を備えたエンティティが製造され得る。

【０１０５】
前に説明したように、環境はシミュレートされた環境であってよい。概して、シミュレートされた環境の場合、観測は、前に説明した観測または観測のタイプのうちの1つまたは複数のシミュレートされたバージョンを含み得、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み得る。たとえば、シミュレートされた環境は、動きシミュレーション環境、たとえば、運転シミュレーションまたは飛行シミュレーションであってよく、エージェントは、動きシミュレーションを通じてナビゲートする、シミュレートされた車両であってよい。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ってよい。概して、エージェントは、シミュレートされた環境と対話している1つまたは複数のコンピュータとして実装され得る。

【０１０６】
シミュレートされた環境は、特定の現実世界環境およびエージェントのシミュレーションであってよい。たとえば、システムは、システムのトレーニングまたは評価中に、シミュレートされた環境においてアクションを選択するために使用され得、トレーニング、もしくは評価、または両方が完了した後、シミュレーションの対象であった特定の現実世界環境において現実世界エージェントを制御するために展開され得る。これは、現実世界環境または現実世界エージェントに対する不要な摩耗と損傷および破損を回避し得、まれに生じる、または現実世界環境において再生することが困難であるかまたは安全ではない状況において制御ニューラルネットワークがトレーニングされ評価されることを可能にし得る。たとえば、システムは、特定の現実世界環境のシミュレーションにおいて機械的エージェントのシミュレーションを使用して部分的にトレーニングされ、後で特定の現実世界環境において実際の機械的エージェントを制御するために展開され得る。したがって、そのような場合、シミュレートされた環境の観測は現実世界環境に関係し、シミュレートされた環境において選択されるアクションは、現実世界環境における機械的エージェントによって実施されるべきアクションに関係する。

【０１０７】
随意に、上記の実装のうち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は、環境を特徴付ける際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施されたアクション、前の時間ステップにおいて受けた報酬、などを含み得る。

【０１０８】
弁別器モデルの1つの例示的な集合について、次により詳細に説明する。

【０１０９】
図2は、図1を参照しながら上記で説明したアクション選択システム100内に含まれる弁別器モデルの集合によって生成される予測出力同士の間の不一致の尺度240を決定するための例示的な技法を示す。

【０１１０】
システム100は、時間ステップのシーケンスの1つまたは複数の時間ステップの各々において教師なし報酬を決定するために弁別器モデルを使用し得る。システム100は、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120をトレーニングするために、教師なし報酬を使用し得る。トレーニングの後、システム100は、環境と対話しているエージェントによって実施されるべきアクションを選択するために、アクション選択ニューラルネットワーク120を使用し得る。

【０１１１】
各弁別器モデルは、時間ステップにおける環境の状態を表す観測210を処理し、環境を観測210によって特徴付けられる状態に入らせるためにアクション選択ニューラルネットワーク120がどの潜在性を条件としたかを予測するそれぞれの予測出力を生成するように構成され得る。図1を参照しながら上記で説明したように、予測出力は、可能な潜在性のセット内の各潜在性に対するそれぞれのスコアを定義する、可能な潜在性のセットにわた

るスコア分布を含み得る。

【０１１２】

時間ステップのシーケンスの1つまたは複数の時間ステップの各々における教師なし報酬、たとえば、式(1)によって上記で定義した報酬、を決定するために、システム100は、複数の弁別器モデルの集合によって生成される予測出力同士の間の不一致の尺度240を決定し得る。システム100は、不一致の尺度240に基づいて、式(1)において$r_{DIS}$を決定することによって、教師なし報酬を決定し得る。

【０１１３】

いくつかの実装形態では、不一致の尺度240を決定するために、システム100は、時間ステップに対して集合内の各弁別器モデルによって生成されるそれぞれのスコア分布220を組み合わせる(たとえば、平均化する)ことによって、時間ステップに対して組み合わされたスコア分布230を決定し得る。特定の例として、システム100は、次のように、組み合わされたスコア分布$q_\phi(Z \mid O)$を決定し得る:

【０１１４】

【数３】

$$q_\phi(Z \mid o_t) = \frac{1}{N}\sum_{N=i}^{N} q_{\phi_i}(Z \mid o_t) \qquad (4)$$

【０１１５】

式中、iは、第iの弁別器モデルを示し、和はN個を超える弁別器モデルである。たとえば、式(4)を使用することによって、組み合わされたスコア分布を決定した後、システム100は、時間ステップに対して組み合わされたスコア分布230および時間ステップに対して各弁別器モデルによって生成されるそれぞれのスコア分布220に基づいて、不一致の尺度を決定し得る。

【０１１６】

いくつかの実装形態では、システム100は、時間ステップに対して各スコア分布のエントロピーを決定し得る。たとえば、システム100は、たとえば、上記の式(4)によって定義された、組み合わされたスコア分布230のエントロピー(H)を次のように決定し得る:

【０１１７】

【数４】

$$H\left[\frac{1}{N}\sum_{N=i}^{N} q_{\phi_i}(Z \mid o_t)\right] \qquad (5)$$

【０１１８】

システム100は、時間ステップに対して各弁別器モデルによって生成されるそれぞれのスコア分布220のそれぞれのエントロピーを次のようにさらに決定し得る:

【０１１９】

【数５】

$$H\left[q_{\phi_i}(Z \mid o_t)\right] \qquad (6)$$

【０１２０】

システムは、上記の式(5)および式(6)を使用して、時間ステップに対して組み合わされたスコア分布のエントロピーと時間ステップに対して各弁別器モデルによって生成されるスコア分布のエントロピーの平均との間の差異として、時間ステップに対する不一致の尺度240を決定し得る。不一致の尺度240に基づいて、システム100は、ステップに対する報酬$r_{DIS}$を次のように決定し得る:

【０１２１】

【数６】

$$r_{DIS} = H\left[\frac{1}{N}\sum_{N=i}^{N} q_{\phi_i}(Z \mid o_t)\right] - \frac{1}{N}\sum_{N=i}^{N} H\left[q_{\phi_i}(Z \mid o_t)\right] \quad (7)$$

【０１２２】

集合内のいくつかの、またはすべての弁別器モデルによって生成される潜在変数のセットにわたるスコア分布がほぼ等しい場合、式(7)における2項もほぼ等しく、小さな、またはごく少ない報酬$r_{DIS}$をもたらす。対照的に、いくつかの、またはすべての弁別器モデルが潜在変数のセットにわたって異なるスコア分布を生成する場合、報酬$r_{DIS}$は、かなり高い。したがって、弁別器モデルによって生成される予測出力同士の間の不一致の尺度240に基づいて教師なし報酬を決定することによって、システム100は、新しい状態の探求を奨励し得る。

【０１２３】

いくつかの実装形態では、システム100は、集合内の弁別器モデルの各対に対する発散尺度(divergence measure)を決定することによって、複数の弁別器モデルの集合によって生成される予測出力同士の間の不一致の尺度240を決定し得る。たとえば、システム100は、弁別器モデルの各対によって生成されるスコア分布同士の間の(すなわち、可能な潜在性のセットにわたる)それぞれの発散(たとえば、Kullback-Leibler発散)を決定することによって、発散尺度を決定し得る。システム100は、たとえば、発散尺度の平均、または中央値として、発散尺度に基づいて不一致の尺度240を決定し得る。次いで、システム100は、たとえば、上記で説明したのと同様の方法で、不一致の尺度240に基づいて教師なし報酬を決定し得る。

【０１２４】

いくつかの実装形態では、システム100は、たとえば、式(2)によって例示されるように、集合内の各弁別器モデルによって生成される予測出力の精度の尺度にやはり基づいて教師なし報酬rを決定し得る。別の例では、システム100は、予測出力の精度の平均として、教師なし報酬を決定し得る。

【０１２５】

アクション選択ニューラルネットワークおよび弁別器モデルの集合をトレーニングするための例示的なプロセスについて、次により詳細に説明する。

【０１２６】

図3は、アクション選択ニューラルネットワーク(たとえば、図1のアクション選択ニューラルネットワーク120)をトレーニングするための例示的なプロセスの流れ図である。便宜上、プロセス300は、1つまたは複数のロケーションに位置する1つまたは複数のコンピュータのシステムによって実施されているとして説明される。たとえば、本明細書に従って適切にプログラムされたアクション選択システム、たとえば、図1のアクション選択システム100、がプロセス300を実行し得る。

【０１２７】

システムは、可能な潜在性のセットからの潜在性をサンプリングする(302)。いくつかの実装形態では、可能な潜在性のセットは、有限に多くの可能な潜在性のみを含み得る。システムは、たとえば、可能な潜在性のセットからの潜在性をランダムにサンプリングし得る。

【０１２８】

システムは、サンプリングされた潜在性を条件とするアクション選択ニューラルネットワークを使用して、時間ステップのシーケンスにわたって環境と対話するためにエージェントによって実施されるべきアクションを選択する(304)。たとえば、システムは、アクション選択出力を生成するために、アクション選択ニューラルネットワークを使用して、時間ステップにおける環境の状態を特徴付ける観測およびサンプリングされた潜在性を処理し得る。アクション選択出力は、たとえば、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スコアを含み得る。次いで、システムは、アクション選択出力に基づいて、時間ステップにおいて実施されるべきアクションを選択し得、たとえば、システ

ムは、最高スコアを有するアクションを選択し得る。

【０１２９】

システムは、時間ステップのシーケンスの各時間ステップに対して受けるそれぞれの報酬を決定する(306)。1つまたは複数の時間ステップのそれぞれに対して、システムは、時間ステップにおける環境の状態を表す測定値を複数の弁別器モデルの集合内の各弁別器モデルに提供する。次いで、1つまたは複数の時間ステップの各々に対して、システムは、複数の弁別器モデルの集合によって生成される予測出力同士の間の不一致の尺度に少なくとも部分的に基づいて、時間ステップに対する報酬を決定する。

【０１３０】

各弁別器モデルは、環境を観測によって特徴付けられる状態に入らせるために、アクション選択ニューラルネットワークが可能な潜在性のセットからのどの潜在性を条件としたかを予測するそれぞれの予測出力を生成するために観測を処理するように構成され得る。予測出力は、たとえば、可能な潜在性のセット内の各潜在性に対するそれぞれのスコアを定義する、可能な潜在性のセットにわたるスコア分布を含み得る。

【０１３１】

システムは、強化学習技法を使用して、報酬に基づいて、アクション選択ニューラルネットワークをトレーニングする(308)。いくつかの実装形態では、システムは、報酬全体、たとえば、目標の達成に向けたエージェントの進歩を特徴付けるタスク報酬、および弁別器モデルによって生成される予測出力同士の間の不一致の尺度に基づく報酬、に基づいて、アクション選択ニューラルネットワークをトレーニングし得る。

【０１３２】

いくつかの実装形態では、システムは、可能な潜在性のセットにわたって組み合わされたスコア分布に基づいて、複数の弁別器モデルの集合によって生成される予測出力同士の間の不一致の尺度を決定し得る。たとえば、システムは、時間ステップに対して各弁別器モデルによって生成されるそれぞれのスコア分布を組み合わせること(たとえば、平均化すること)によって、時間ステップに対して組み合わされたスコア分布を決定し得る。

【０１３３】

時間ステップに対して組み合わされたスコア分布を決定した後、システムは、(i)組み合わされたスコア分布、および(ii)時間ステップに対して各弁別器モデルによって生成されるそれぞれのスコア分布に基づいて、不一致の尺度を決定し得る。たとえば、システムは、組み合わされたスコア分布のエントロピーを決定することによって、時間ステップに対する不一致の尺度を決定し得る。さらに、システムは、各弁別器モデルによって生成されるそれぞれのスコア分布のそれぞれのエントロピーを決定し得る。

【０１３４】

次いで、システムは、これらのエントロピーに基づいて、不一致の尺度を決定し得る。特定の例として、システムは、(i)時間ステップに対して組み合わされたスコア分布のエントロピーと(ii)時間ステップに対して各弁別器モデルによって生成されるスコア分布のエントロピーの平均との間の差異を決定し得る。

【０１３５】

いくつかの実装形態では、システムは、予測出力の精度に基づいて、時間ステップに対する報酬を決定し得る。たとえば、システムは、時間ステップに対して各弁別器によって生成される予測出力のそれぞれの精度を決定し得る。さらに、システムは、予測出力の精度に少なくとも部分的に基づいて、たとえば、予測出力の精度の平均に少なくとも部分的に基づいて、時間ステップに対する報酬を決定し得る。いくつかの実装形態では、システムは、環境におけるタスクの達成におけるエージェントの進歩を測定するタスク報酬に少なくとも部分的に基づいて、時間ステップに対する報酬を決定し得る。

【０１３６】

図1を参照しながら上記で説明したように、システムは、弁別器モデルの集合をトレーニングし得る。各弁別器モデルは、弁別器モデルパラメータのそれぞれのセットを有し得る。いくつかの実装形態では、弁別器モデルパラメータのそれぞれのセットのそれぞれの

値は、各弁別器モデルに対して異なり得る。弁別器モデルをトレーニングする前に、システムは、各他の弁別器モデルの弁別器モデルパラメータのセットとは異なるように各弁別器モデルの弁別器モデルパラメータのセットのそれぞれの値を初期化し得る。

【０１３７】

システムは、リプレイメモリからのトレーニング例の単独でサンプリングされたバッチに基づいて各弁別器モデルをトレーニングし得る。リプレイメモリ内の各トレーニング例は、たとえば、(i)環境とのエージェントの前の対話中の環境の状態を特徴付けるトレーニング観測と、(ii)環境を観測によって特徴付けられる状態に入らせるためにアクション選択ニューラルネットワークが条件とした潜在性を定義するターゲット出力とを含み得る。

【０１３８】

いくつかの実装形態では、アクション選択ニューラルネットワークをトレーニングした後、システムは、現実世界環境と対話している現実世界エージェントによって実施されるべきアクションを選択するためにアクション選択ニューラルネットワークを使用し得る。たとえば、エージェントは、物理的環境における物理的ロボットであってよく、システムは、特定のタスクを実施するための、たとえば、ロボットを物理的環境において第1のロケーションから第2のロケーションに移動させるための、アクションを選択し得る。

【０１３９】

例示的な実験結果について、次により詳細に説明する。

【０１４０】

図4Aおよび図4Bは、アクション選択システム100を使用して達成された例示的な実験結果400を示す。図4Aおよび図4Bの各パネルは、4つの空間を有するグリッドワールドを示す。各グリッドワールドは、合計で104個の状態(たとえば、グリッド内の方形)を有することになる。システム100は、左上隅の方形として環境の第1の状態を初期化する。

【０１４１】

第1の状態を初期化した後、システム100は、可能な潜在性のセットからの潜在性をサンプリングし、時間ステップのシーケンス(たとえば、20時間ステップ)にわたって、サンプリングされた潜在性を条件とするアクション選択ニューラルネットワーク120を使用して、グリッドワールド内でエージェントによって実施されるべきアクションを選択し得る。時間ステップのシーケンスの長さは、1個の状態以外のすべてに達するために十分であり、合計で103個の区別可能なスキルを可能にする。アクションは、たとえば、左に移動する、右に移動する、上に移動する、下に移動する、および現状に留まる、を含み得る。潜在性のセットは、128個の潜在変数を含む。

【０１４２】

図4Aは、弁別器モデルの集合によって生成される予測出力同士の間の不一致の尺度に基づいて生成された教師なし報酬がない場合のシステム100のパフォーマンスを示す。示すように、エージェントは4個の空間のうちの3つにほとんど入らず、それにより、環境を効果的に探求できない。この場合、エージェントは、合計で30個の別個のスキルをどうにか学習する。

【０１４３】

図4Bは、教師なし報酬を用いたシステム100のパフォーマンスを示す。示すように、エージェントは、すべての4個の空間に入ることによって環境を効果的にどうにか探求する。さらに、この場合、エージェントは、3倍の数のスキル、たとえば、およそ90個の別個のスキルを学習する。したがって、本明細書で説明するシステム100は、エージェントが、より多くのスキルを学習し、環境の新しい部分を探求することを可能にする。

【０１４４】

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ともに「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施するように構成された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、システムにそれらの動作またはアクションを実施させる、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その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

ョンを実施するように構成された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、データ処理装置によって実行されると、装置にそれらの動作またはアクションを実施させる命令を含むことを意味する。

【０１４５】

本明細書で説明した主題および機能的動作の実施形態は、デジタル電子回路で、有形に実施されたコンピュータソフトウェアまたはコンピュータファームウェアで、本明細書で開示する構造およびそれらの構造的均等物を含めて、コンピュータハードウェアで、またはそれらのうちの1つまたは複数の組合せで実装され得る。

【０１４６】

本明細書で説明した主題の実施形態は、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、またはそ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、ランダムアクセスメモリデバイス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ってよい。代替または追加として、プログラム命令は、人工的に生成された伝搬信号、たとえば、データ処理装置による実行のため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、機械生成された電気信号、光信号、または電磁信号、の上で符号化され得る。

【０１４７】

「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、1つのプログラマブルプロセッサ、1つのコンピュータ、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、および機械を含む。装置は、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イ)またはASIC(特定用途向け集積回路)を同じく、またはさらに含んでよい。装置は、ハードウェアに加えて、コンピュータプログラムのための実行環境を作り出すコード、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つまたは複数の組合せをなすコード、を随意に含んでよい。

【０１４８】

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それらとして説明されることもあるコンピュータプログラムは、コンパイル型言語もしくはインタープリタ型言語、または宣言型言語もしくは手続き型言語を含めて、いかなる形態のプログラミング言語で書き込まれてもよく、コンピュータプログラムは、スタンドアロンプログラムとして、またはモジュール、構成要素、サブルーチン、またはコンピューティング環境で使用するのに適した他のユニットとして、を含めて、任意の形態で展開されてよい。プログラムは、そうでなくてもよいが、ファイルシステム内のファイルに対応し得る。プログラムは、他のプログラムまたはデータ、たとえば、マークアップ言語文書で記録された1つまたは複数のスクリプト、を保持するファイルの一部分の中に、当該プログラム専用の単一ファイルの中に、または複数の協調ファイル、たとえば、1つまたは複数のモジュール、サブプログラム、またはコードの部分を記憶するファイル、の中に記憶されてよい。コンピュータプログラムは、1つのコンピュータ上、または1つのサイトに位置するか、または複数のサイトにわたって分散され、データ通信ネットワークによって相互接続された、複数のコンピュータ上で実行されるように展開され得る。

【０１４９】

本明細書で、「エンジン」という用語は、ソフトウェアベースのシステム、サブシステム、または1つまたは複数の特定の機能を実施するよう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。概して、エンジンは、1つまたは複数のロケーションにおいて1つまたは複数のコンピュータ上にインストールされた、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ことになる。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専用になり、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インストールされ、実行していることがある。

【０１５０】

本明細書で説明した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、出力を生成することによって機能を実施するための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マブルコンピュータによって実施され得る。プロセスおよび論理フローはまた、専用論理回路、たとえば、FPGAまたはASIC、によって、または専用論理回路と1つまたは複数のプログラムされたコンピュータの組合せによって実施されてもよい。

【０１５１】

コンピュータプログラムの実行に適したコンピュータは、汎用マイクロプロセッサもしくは専用マイクロプロセッサ、または両方、あるいは任意の他の種類の中央処理装置に基づいてよい。概して、中央処理装置は、読取り専用メモリもしくはランダムアクセスメモリまたは両方から命令およびデータを受信する。コンピュータの必須要素は、命令を実施または実行するための中央処理装置、および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助されてよく、またはその中に組み込まれてもよい。概して、コンピュータはまた、データを記憶するための1つまたは複数の大容量記憶デバイス、たとえば、磁気ディスク、光磁気ディスク、または光ディスク、を含むか、あるいはそれらからデータを受信もしくはそれらにデータを転送、またはその両方を行うために動作可能に結合される。しかしながら、コンピュータはそのようなデバイスを有さなくてもよい。さらに、コンピュータは、別のデバイス、たとえば、ほんのいくつかを上げると、モバイル電話、携帯情報端末(PDA)、モバイルオーディオプレーヤーまたはビデオプレーヤー、ゲーム機、全地球測位システム(GPS)受信機、またはポータブル記憶デバイス、たとえば、ユニバーサルシリアルバス(USB)フラッシュデバイスの中に埋め込まれてもよい。

【０１５２】

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適したコンピュータ可読媒体は、例として、半導体メモリデバイス、たとえば、EPROM、EEPROM、およびフラッシュメモリデバイス;磁気ディスク、たとえば、内部ハードディスクまたはリムーバブルディスク;光磁気ディスク;ならびにCD ROMおよびDVD-ROMディスクを含めて、すべての形態の不揮発性メモリ、媒体、およびメモリデバイスを含む。

【０１５３】

ユーザとの対話を実現するために、本明細書で説明した主題の実施形態は、情報をユーザに表示するためのディスプレイデバイス、たとえば、CRT(陰極線管)またはLCD(液晶ディスプレイ)モニター、およびそれによりユーザが入力をコンピュータに提供し得るキーボードおよびポインティングデバイス、たとえば、マウスまたはトラックボール、を有するコンピュータ上で実装され得る。ユーザとの対話を提供するために他の種類のデバイスが同様に使用されてもよい。たとえば、ユーザに提供されるフィードバックは、任意の形態の感覚フィードバック、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、であってよく、ユーザからの入力は、音響入力、音声入力、または触覚入力を含めて、任意の形態で受信されてよい。加えて、コンピュータは、たとえば、ユーザのデバイス上のウェブブラウザから受信された要求に応答して、ウェブページをそのウェブブラウザに送ることによって、ユーザによって使用されるデバイスに文書を送り、そこから文書を受信することによってユーザと対話し得る。また、コンピュータは、テキストメッセージまたは他の形態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り、返事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、ユーザと対話し得る。

【０１５４】

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、機械学習トレーニングまたは生産の一般的な計算集約的な部分、すなわち、推論、作業負荷、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み得る。

【０１５５】

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、を使用して実装および展開され得る。

【０１５６】

本明細書で説明した主題の実施形態は、バックエンド構成要素、たとえば、データサーバ、を含む、もしくはミドルウェア構成要素、たとえば、アプリケーションサーバ、を含む、またはフロントエンド構成要素、たとえば、それを通してユーザが本明細書で説明する主題の実装形態と対話し得る、グラフィカルユーザインターフェース、ウェブブラウザ、またはアプリを有するクライアントコンピュータを含む、または、1つまたは複数のそのようなバックエンド構成要素、ミドルウェア構成要素、またはフロントエンド構成要素の任意の組合せを含む、コンピューティングシステムで実装され得る。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は、ローカルエリアネットワーク(LAN)および広域ネットワーク(WAN)、たとえば、インターネットを含む。

【０１５７】

コンピューティングシステムは、クライアントおよびサーバを含み得る。クライアントおよびサーバは、概して、互いと離れており、一般に、通信ネットワークを通して対話する。クライアントとサーバの関係は、それぞれのコンピュータ上で実行し、互いに対してクライアント-サーバ関係を有するコンピュータプログラムにより生じる。いくつかの実施形態では、サーバは、たとえば、クライアントとして動作するデバイスと対話しているユーザにデータを表示し、そこからユーザ入力を受信するために、データ、たとえば、HTMLページ、をユーザデバイスに送信する。ユーザデバイスにおいて生成されるデータ、たとえば、ユーザ対話の結果、は、デバイスからサーバにおいて受信され得る。

【０１５８】

本明細書は多くの特定の実装詳細を含むが、これらは、いずれの発明の範囲に対するまたは特許請求され得る範囲に対する限定と解釈されるべきではなく、むしろ、特定の発明の特定の実施形態に固有であり得る特徴の説明と解釈されるべきである。本明細書において個別の実施形態の文脈で説明したいくつかの特徴は、単一の実施形態で組み合わせて実装されてもよい。逆に、単一の実施形態の文脈で説明された様々な特徴は、複数の実施形態で別々にまたは任意の好適な部分組合せで実装されてもよい。さらに、特徴は、一定の組合せで作用するとして上記で説明され、当初そういうものとして特許請求されることすらあるが、特許請求される組合せからの1つまたは複数の特徴は、場合によっては、組合せから削除されてよく、特許請求される組合せは部分組合せまたは部分組合せの変種を対象とすることがある。

【０１５９】

同様に、動作は図面において特定の順序で示され、また特許請求の範囲において特定の順序で記載されているが、これは、所望の結果を達成するために、そのような動作が示された特定の順序でまたは順番で実行されること、またはすべての示された動作が実行されることを必要とすると理解すべきではない。いくつかの状況では、マルチタスキングおよび並列処理が有利であり得る。さらに、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、そのような分離がすべての実施形態で必要とされると理解すべきではなく、説明したプログラム構成要素およびシステムは、概して、単一のソフトウェア製品内に一緒に統合されてよいか、または複数のソフトウェア製品にパッケージングされてよいと理解されたい。

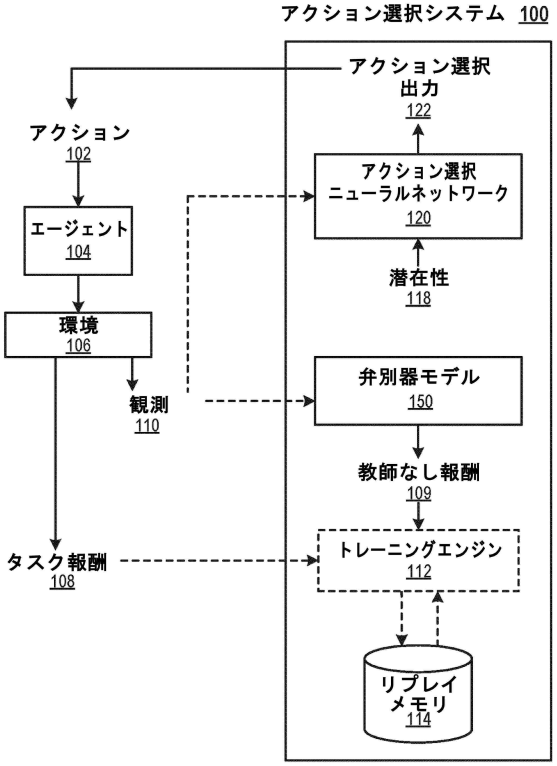
【０１６０】

主題の特定の実施形態について説明してきた。他の実施形態は、以下の特許請求の範囲内である。たとえば、特許請求の範囲に記載されるアクションは、異なる順序で実行されてよく、依然として所望の結果を達成する。一例として、添付の図面に示されたプロセスは、所望の結果を達成するために、必ずしも示された特定の順序、または順番を必要とす

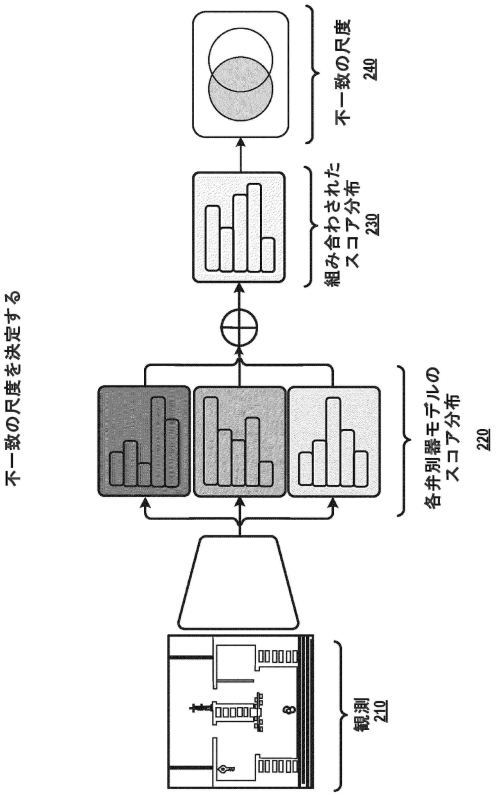
るとは限らない。場合によっては、マルチタスキングおよび並列処理が有利であり得る。
【符号の説明】
【０１６１】
100 アクション選択システム、システム
102 アクション
104 エージェント
106 環境
108 タスク報酬
109 教師なし報酬
110 観測
112 トレーニングエンジン
114 リプレイメモリ、メモリ
118 潜在性、潜在変数
120 アクション選択ニューラルネットワーク
122 アクション選択出力
150 弁別器モデル
210 観測
220 スコア分布
230 組み合わされたスコア分布
240 不一致の尺度
300 プロセス
400 例示的な実験結果

【図面】

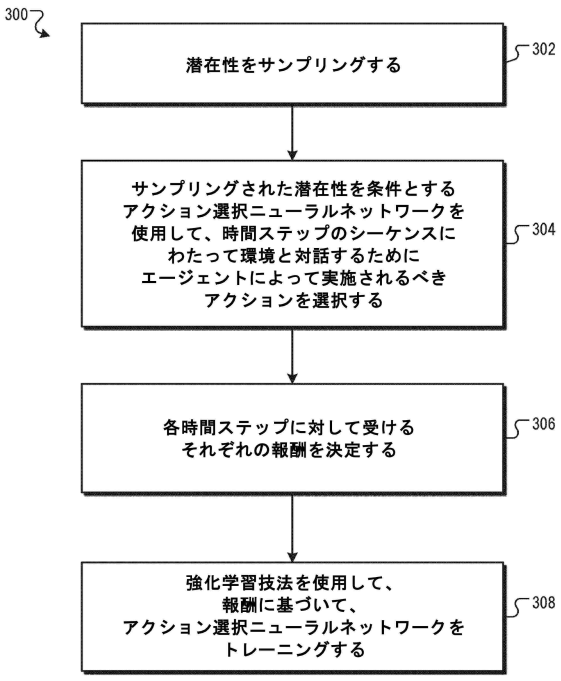
【図１】



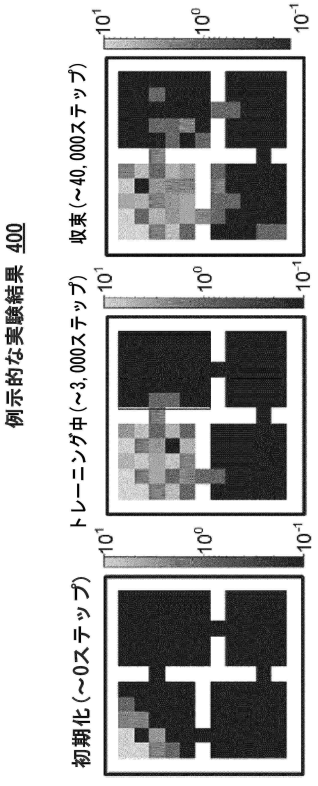

【図２】



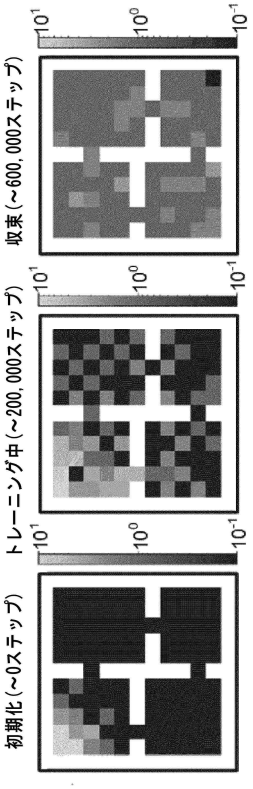

【図３】




【図４Ａ】



【図４Ｂ】



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

審査官　福西 章人

(56)参考文献
特表２０１９－５３４５１７（ＪＰ，Ａ）
国際公開第２０１８／０８３６７１（WO，Ａ１）
CHEN, Annie S. et al.，Batch Exploration with Examples for Scalable Robotic Reinforcement Learning，arXiv [online]，2021年04月23日，pp.1-11，[検索日 2024.10.30]、インターネット:〈URL:https://arxiv.org/pdf/2010.11917v2〉

(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ０６Ｎ　３／００－９９／００
Ｇ０６Ｆ　１８／００－１８／４０