

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号

特許第7225292号

(P7225292)

(45)発行日 令和5年2月20日(2023.2.20)

(24)登録日 令和5年2月10日(2023.2.10)

(51)国際特許分類

F I

G 0 5 B 13/02 (2006.01)

G 0 5 B 13/02

L

G 0 6 N 20/00 (2019.01)

G 0 6 N 20/00

G 0 5 D 1/02 (2020.01)

G 0 5 D 1/02

K

G 0 8 G 1/00 (2006.01)

G 0 8 G 1/00

X

請求項の数 20 外国語出願 (全53頁)

(21)出願番号 特願2021-42252(P2021-42252)
 (22)出願日 令和3年3月16日(2021.3.16)
 (65)公開番号 特開2022-46402(P2022-46402A)
 (43)公開日 令和4年3月23日(2022.3.23)
 審査請求日 令和3年9月3日(2021.9.3)
 (31)優先権主張番号 2014285.7
 (32)優先日 令和2年9月10日(2020.9.10)
 (33)優先権主張国・地域又は機関
 英国(GB)

(73)特許権者 000003078
 株式会社東芝
 東京都港区芝浦一丁目1番1号
 (74)代理人 110003708
 弁理士法人鈴榮特許総合事務所
 (72)発明者 スティーヴン モラド
 イギリス国、 シービー4・0ジーゼツ
 ト、 ケンブリッジシャー、 ケンブリッ
 ジ、 ミルトン・ロード、 ケンブリッ
 ジ・サイエンス・パーク 208、 トーシ
 バ・リサーチ・ヨーロッパ・リミテッド
 、ケンブリッジ・リサーチ・ラボラトリ
 ー内
 (72)発明者 ロベルト メッカ
 イギリス国、 シービー4・0ジーゼツ
 最終頁に続く

(54)【発明の名称】 タスク実行エージェントシステム及び方法

(57)【特許請求の範囲】

【請求項1】

装置と前記装置の物理的環境とを備える第2のコンテキストにおいて前記装置がナビゲーションタスクを実行することを可能にするように、エンティティと前記エンティティの環境とを備える第1のコンテキストにおいて、特定のナビゲーションタスクを前記エンティティが実行するために一連の行動を前記エンティティが実行するように前記エンティティを制御するエージェントをトレーニングするためのコンピュータが実施する方法であって、前記装置は、前記装置の前記物理的環境の画像を受け取るように適応されており、前記装置の方向を制御するように適応されたステアリングデバイスを備え、前記方法は、

1つ又は複数のナビゲーションタスクを取得することと、ここで、前記取得するとは、前記エンティティの環境に基づく第1のナビゲーションタスクを生成することと、前記エンティティが前記エンティティの環境においてターゲット位置まで前記エージェントによるナビゲーションが成功する度合いを示す、ナビゲーションタスクの容易性を推定するようにトレーニングされた機械学習モデルを使用して前記第1のナビゲーションタスクをスコアリングすることと、スコアが第1の選択基準を満たす第1のナビゲーションタスクを前記1つ又は複数のナビゲーションタスクのうちの1つとして選択することとを備え、
前記エンティティが、前記エンティティの環境の画像を使用して前記1つ又は複数のナビゲーションタスクを実行しようと試みることを備える強化学習方法を使用して前記エージェントをトレーニングすることと
 を備え、

10

20

前記第1のナビゲーションタスクは、前記エンティティの初期位置とターゲット位置とを含み、前記初期位置と前記ターゲット位置とはナビゲート可能なエリア内に制限された位置上でランダムに生成され、

前記スコアリングは、前記第1のナビゲーションタスクの幾何学的特性に基づいて得られ、

前記第1の選択基準を満たすナビゲーションタスクは、前記第1の選択基準を満たさないナビゲーションタスクより、前記エンティティが前記エンティティの環境において前記ターゲット位置まで前記エージェントによってナビゲートされる成功確率が高い、方法。

【請求項2】

前記機械学習モデルは、複数のナビゲーションタスクの各々について、それぞれのナビゲーションタスクの1つ又は複数の特性を備えるトレーニングセットと、前記第1のナビゲーションタスクを実行しようとする試行がナビゲーションタスクの成功基準を満たしたかどうかを示す指示と、に基づいて、前記ナビゲーションタスクの容易性を推定するようにトレーニングされる、請求項1に記載の方法。

10

【請求項3】

前記ナビゲーションタスクの成功基準は、前記エンティティが前記エンティティの環境においてターゲット位置に前記エージェントによってナビゲートされた場合、満たされる、請求項2に記載の方法。

【請求項4】

前記機械学習モデルは、前記エージェントと共にトレーニングされる、請求項1乃至3のいずれか一項に記載の方法。

20

【請求項5】

前記スコアは、前記スコアが決定された範囲内である場合、前記第1のナビゲーションタスクが中間容易性であることを示し、前記スコアに対応する前記第1のナビゲーションタスクが前記中間容易性を示す場合、前記選択基準を満たす、請求項1乃至4のいずれか一項に記載の方法。

【請求項6】

前記スコアは、前記スコアが決定された範囲内である場合、前記第1のナビゲーションタスクが中間容易性であることを示し、前記範囲を決定することは、

複数のナビゲーションタスクを生成することと、

30

前記機械学習モデルを使用して、前記複数のナビゲーションタスクをスコアリングすることと、

前記複数のナビゲーションタスクの各々に対応する複数のスコアに基づいて前記範囲を算出することと

を備える、請求項5に記載の方法。

【請求項7】

前記1つ又は複数のナビゲーションタスクを取得することは、

前記第1のナビゲーションタスクとは異なる、ランダムに生成される第2のナビゲーションタスクを生成することと、

前記ナビゲーションタスクの容易性を推定するようにトレーニングされた前記機械学習モデルを使用して前記第2のナビゲーションタスクをスコアリングすることと、

40

前記第2のナビゲーションタスクをスコアリングした結果得られたスコアが前記第1の選択基準と同一又は前記第1の選択基準とは異なる第2の選択基準を満たす前記第2のナビゲーションタスクを前記1つ又は複数のナビゲーションタスクのうちの1つとして選択することと

を備える、請求項1乃至6のいずれか一項に記載の方法。

【請求項8】

前記第1のナビゲーションタスクとは異なる第2のナビゲーションタスクをスコアリングした結果得られたスコアが閾値スコアより大きい場合、前記第2のナビゲーションタスクが容易であることを示し、前記スコアに対応する前記第2のナビゲーションタスクが容易

50

であることを示す場合、第2のナビゲーションタスクをスコアリングした結果得られたスコアが前記第1の選択基準と同一又は前記第1の選択基準とは異なる第2の選択基準を満たす、請求項1乃至7のいずれか一項に記載の方法。

【請求項9】

前記スコアは、前記スコアが前記閾値スコアより大きい場合、前記第2のナビゲーションタスクが容易であることを示し、前記閾値スコアを決定することは、

複数のナビゲーションタスクを生成することと、

前記機械学習モデルを使用して、前記複数のナビゲーションタスクをスコアリングすることと、

前記複数のナビゲーションタスクの各々に対応する複数のスコアに基づいて前記閾値スコアを算出することと

を備える、請求項8に記載の方法。

【請求項10】

前記エンティティが、前記エンティティの環境の画像を使用して前記1つ又は複数のナビゲーションタスクを実行しようと試みることは、

1つ又は複数の行動ステップを実行すること

を備え、各行動ステップは、

前記エンティティの環境の1つ又は複数の画像を受け取ることと、

前記エージェントを使用して、前記1つ又は複数の画像に基づいて前記エンティティの環境において前記エンティティが実行可能な複数の行動のうちの1つ又は複数の行動を決定することと、

前記決定された1つ又は複数の行動を実行させることと、

前記第1のコンテキストの更新状態とナビゲーションタスクの成功基準とに基づいて、前記強化学習方法を使用して前記エージェントをトレーニングすることに対する報酬を算出することと

を備える、請求項1乃至9のいずれか一項に記載の方法。

【請求項11】

前記第1のコンテキストの更新状態が前記第1のコンテキストの以前の状態とは異なる場合に、前記算出される報酬が増やされる、請求項10に記載の方法。

【請求項12】

トレーニングのエンティティがトレーニング環境において1つ又は複数のオブジェクトと衝突したことを前記第1のコンテキストの前記更新状態が示す場合に、前記算出される報酬が減らされる、請求項11に記載の方法。

【請求項13】

前記エンティティは、シミュレートされたエンティティであり、前記エンティティの環境は、シミュレートされた環境である、請求項1乃至12のいずれか一項に記載の方法。

【請求項14】

前記エージェントは、

1つ又は複数の潜在的特徴を決定するために1つ又は複数の画像を処理するように構成されたネットワークである、1つ又は複数の特徴ネットワークと、

前記エンティティの環境において前記エンティティが実行可能な1つ又は複数の行動の1つ又は複数の行動値を決定するために前記1つ又は複数の潜在的特徴を処理するように構成されたネットワークである、ポリシーネットワークと

を備える、請求項1乃至13のいずれか一項に記載の方法。

【請求項15】

前記1つ又は複数の特徴ネットワークは、物理的環境の複数の画像を使用して事前にトレーニングされる、請求項14に記載の方法。

【請求項16】

前記1つ又は複数の特徴ネットワークのパラメータは、前記エージェントをトレーニングするとき、更新されない、請求項15に記載の方法。

10

20

30

40

50

【請求項 17】

前記1つ又は複数の特徴ネットワークは、1つ又は複数のセマンティックターゲットの特徴ネットワークを備える、請求項14乃至16のいずれか一項に記載の方法。

【請求項 18】

装置の物理的環境においてナビゲーションタスクを実行するように前記装置を制御するためのコンピュータが実施する方法であって、前記装置は、前記装置の前記物理環境の画像を受け取るように適応されており、前記装置の方向を制御するように適応されたステアリングデバイスを備え、前記方法は、

前記装置の物理環境の1つ又は複数の画像を受け取ることと、

請求項1乃至17のいずれか一項に記載の方法に従ってトレーニングされたエージェントを使用して、前記1つ又は複数の画像に基づいて、前記装置が実行可能な複数の行動のうちの1つ又は複数の行動を決定することと、

前記装置に前記決定された1つ又は複数の行動を実行させることと

を備える、方法。

【請求項 19】

請求項17に記載の方法を実行するように構成された1つ又は複数のプロセッサを備えるシステム。

【請求項 20】

前記装置を備える、請求項19に記載のシステム。

【発明の詳細な説明】

【技術分野】

【0001】

本明細書で説明する実施形態は、タスク実行エージェントシステム及び方法と、それをトレーニングするための方法及びシステムとに関する。

【背景技術】

【0002】

エージェントシステム及び方法は、実在する（例えば物理的である）か、シミュレートされたものであり得るエンティティを、それらのコンテキストの観察に基づいて、例えばエンティティの環境及び／又はエンティティの状態の観察に基づいて制御する。そのようなエージェントは、特定のタスクを実行するために一連の行動を実行するようにエンティティを制御し得る。既存のエージェントが悪戦苦闘する1つのタスクは、雑然とした現実的な屋内環境内をナビゲートすることである。特に、エージェントは、これらの環境において衝突を回避しながらゴールのオブジェクトまでナビゲートするようにエンティティを制御するのに悪戦苦闘する。衝突を回避しながらナビゲートすることができることは、オフラインのマップを利用することができない可能性が高い、オフィスの廊下を通して重い箱を運ぶこと、レストランでテーブルを片付けること、注射器を病院の手術室に届けること、及び障害回復のための自律探査などのタスクをロボットが実行することを容易にするのに重要である。従って、説明した既存のナビゲーションエージェントの制限を克服することで、説明したような、より豊富なタスクに対して及びより豊富なコンテキストにおけるロボットの使用が容易になり得る。

【0003】

既存のエージェントは、古典的なアプローチに基づき得、古典的なエージェントと称され、例えば環境における観察に応答するために固定のアルゴリズムを使用する非学習ポリシーを利用し得る。しかしながら、これらのエージェントは、実世界のコンテキストにおいて一般化することができないため、新しい状況に適応することができず、規定のコンテキスト又は予測可能なコンテキスト以外ではそれらの実用性は限られている。

【0004】

近年では、深層強化学習技法を使用してトレーニングされたエージェントは、例えば単純なアーケードゲームをプレイするといった特定のコンテキストで多大な成功を収めており、古典的なエージェントよりもかなり高度な一般化能力を示している。深層強化学習技

10

20

30

40

50

法を使用してトレーニングされたエージェントは、コンテキストの観察に基づいてエンティティが実行すべき行動を決定するために、1つ又は複数のニューラルネットワークを使用する。トレーニング中、エージェントは、エンティティに行動を実行させ、報酬が、行動を実行した後のコンテキストと、この報酬を最大にするための1つ又は複数のニューラルネットワークのパラメータとに基づいて算出される。報酬は、実行される行動が、エージェントが実行するようにトレーニングされているゴール、例えばタスクの完了、を達成する、及び／又は、そのゴールに向かって前進するかどうかに基づき得る。

【図面の簡単な説明】

【0005】

【図1A】例となる実施形態による、自律走行車両の例示である。

10

【図1B】例となる実施形態による、自律航空車両の例示である。

【図1C】例となる実施形態による、自律車両がナビゲートすることができる例となる環境の例示である。

【図1D】例となる実施形態による、環境において自律車両をナビゲートするための方法のフロー図である。

【図1E】例となる実施形態による、エージェントをトレーニングするためのシステムを例示する概略ブロック図である。

【図2】例となる実施形態による、エージェント及びタスク容易性推定関数をトレーニングするための方法のフロー図である。

【図3A】例となる実施形態による、1つ又は複数のタスクの取得に使用される第1の方法のフロー図である。

20

【図3B】例となる実施形態による、1つ又は複数のタスクの取得に使用される第2の方法のフロー図である。

【図3C】例となる実施形態による、1つ又は複数のタスクの取得に使用される第3の方法のフロー図である。

【図4】例となる実施形態による、エージェントによってタスクの実行を試行するための方法のフロー図である。

【図5A】例となる実施形態による、プロダクション環境において使用されているエージェントを例示する概略図である。

【図5B】例となる実施形態による、トレーニング中のエージェントを例示する概略図である。

30

【図6】プロダクション環境においてエンティティを制御するための方法のフロー図である。

【図7】例となる実施形態が実施され得る使用するコンピューティングハードウェアの概略図である。

【図8A】ナビゲーションタスクを実行する際の例となる実施形態のSuccess weighted by (normalized inverse) Path Length (SPL) メトリックを比較するグラフである。

【図8B】例となる実施形態を使用して実行されるナビゲーションを示す平面図である。

【図9】様々なトレーニングエポックにおける異なる幾何学的特性にわたるタスク空間を視覚化したものである。

40

【図10A】一様にランダムなタスク、NavACL-GOID、及びNavACL-Adaptiveについて、単一の視覚ナビゲーションテストトレーニング環境での3回の試行における検証エピソード成功率を比較するグラフである。

【図10B】一様にランダムなタスク、NavACL-GOID、及びNavACL-Adaptiveについて、平均的な測地タスク距離を例示するグラフである。

【図11A】例となる実施形態によるエージェントがトレーニング及び／又は評価され得る第1のシミュレート環境の写真である。

【図11B】例となる実施形態によるエージェントがトレーニング及び／又は評価され得る第2のシミュレート環境の写真である。

【図11C】例となる実施形態によるエージェントがトレーニング及び／又は評価され得

50

る第3のシミュレート環境の写真である。

【図12A】エージェントがトレーニングされるタスク及びトレーニング中の検証タスクに対するエージェントの例となる実施形態のSPLメトリックを比較するグラフである。

【図12B】検証タスクに対するエージェントの例となる実施形態の成功率と、エージェントのトレーニング中に取得されるタスクの幾何学的特性とを例示する。

【図13A】エージェントの例となる実施形態を使用して自動搬送車（AGV）が屋内物理的環境の中を誘導される様子の例示である。

【図13B】エージェントの例となる実施形態を使用して無人航空機（UAV）が屋内物理的環境の中を誘導される様子の例示である。

【発明を実施するための形態】

【0006】

第1の実施形態では、装置と装置の物理的環境とを備える第2のコンテキストにおいて装置がナビゲーションタスクを実行することを可能にするように、エンティティとエンティティの環境とを備える第1のコンテキストにおいてエージェントをトレーニングするためのコンピュータ実施方法が提供される。装置は、装置の物理的環境の画像を受け取るように適応される。装置は、装置の方向を制御するように適応されたステアリングデバイスを備える。方法は、1つ又は複数のナビゲーションタスクを取得することと、エンティティが、エンティティの環境の画像を使用して1つ又は複数のナビゲーションタスクを実行しようと試みることを備える強化学習方法を使用してエージェントをトレーニングすることとを備える。1つ又は複数のナビゲーションタスクを取得することは、ナビゲーションタスクを生成することと、タスクの容易性を推定するようにトレーニングされた機械学習モデルを使用してナビゲーションタスクをスコアリングすることと、スコアが選択基準を満たすことに応答して、ナビゲーションタスクを1つ又は複数のナビゲーションタスクのうちの1つとして選択することとを備える。

【0007】

上記実施形態は、ナビゲーションタスクを実行するか又は少なくともより良く実行するように装置を制御することができるエージェントのトレーニングを容易にする。エージェントのトレーニングの改善はまた、エージェントのトレーニングで実行される行動及び／又はパラメータ更新の数を低減し得る。従って、提供される改善は、エージェントのトレーニングにおける計算リソース使用量を低減する。ナビゲーションタスクをより良く実行するように装置を制御することができることで、エネルギーを浪費する装置の誤った動作が低減されるため、トレーニングされたエージェントは、装置のエネルギー使用量を低減する。

【0008】

装置は、自律搬送車両又は無人航空機などの車両であり得る。

【0009】

ナビゲーションタスクは、装置の開始ロケーションと装置のターゲットロケーションとによって定義され得る。

【0010】

ナビゲーションタスクの生成は、ランダムな（又は擬似ランダムな）ナビゲーションタスクの生成であり得る。

【0011】

ナビゲーションタスクのスコアリングは、ナビゲーションタスクの1つ又は複数の特性に基づき得る。ナビゲーションタスクの1つ又は複数の特性は、ナビゲーションタスクの1つ又は複数の幾何学的特性であり得るか又はそれを含み得る。ナビゲーションタスクの1つ又は複数の特性は、機械学習モデルへの入力またはある入力であり得る。

【0012】

機械学習モデルは、複数のナビゲーションタスクの各々について、それぞれのナビゲーションタスクの1つ又は複数の特性を備えるトレーニングセットと、ナビゲーションタスクを実行しようとする試行がナビゲーションタスク成功基準を満たしたかどうかを示す指

10

20

30

40

50

示とに基づいて、ナビゲーションタスクの容易性を推定するようにトレーニングされ得る。

【0013】

ナビゲーションタスク成功基準は、エンティティがエンティティの環境においてターゲットロケーションにエージェントによってナビゲートされた場合に満たされ得る。ターゲットオブジェクトは、ターゲットロケーションにあり得るか又はその最も近くにあり得る。ターゲットオブジェクトは、ターゲットセマンティッククラスのオブジェクトであり得る。

【0014】

機械学習モデルは、フィードフォワードニューラルネットワークを含み得る。フィードフォワードニューラルネットワークは、全結合ニューラルネットワークであり得る。

【0015】

例えば、畳み込みニューラルネットワーク又は再帰型ニューラルネットワークのような他のタイプの同等なニューラルネットワークよりも、フィードフォワードニューラルネットワークのトレーニングにかかる時間は少なく、利用される計算リソースは少ない。

【0016】

機械学習モデルは、機械学習モデルがエージェントの現在の能力に基づいてタスクをスコアリングすることができように、エージェントと共にトレーニングされ得る。例えば、機械学習モデルは、トレーニングの現在の反復におけるエージェントについて、成功確率のような容易性を推定し得る。

【0017】

スコアは、ナビゲーションタスクが中間容易性であることをスコアが示す場合、選択基準を満たし得る。

【0018】

中間容易性のナビゲーションタスクは、エージェントにその現在の能力で新しいタスクを解決するように教えるため、中間容易性タスクがランダムなタスクよりもポリシー改善のためのより多くの学習信号を供給することによって、中間容易性タスクがエージェントのより良好な及び／又はより効率的なトレーニングを容易にするため、エージェントのトレーニングでの使用に有益である。従って、中間容易性のナビゲーションタスクは、使用中のエージェントのナビゲーションパフォーマンスを改善し得、及び／又は、エージェントをトレーニングする際に使用される時間及び／又は計算リソースを低減し得る。

【0019】

スコアは、スコアが決定された範囲内である場合、ナビゲーションタスクが中間容易性であることを示し得る。範囲を決定することは、複数のナビゲーションタスクを生成することと、機械学習モデルを使用して、複数のナビゲーションタスクをスコアリングすることと、複数のナビゲーションタスクの各々の対応する複数のスコアに基づいて範囲を算出することとを含み得る。

【0020】

スコアは、ナビゲーションタスクが中間困難性であることをスコアが示す場合、選択基準を満たし得る。

【0021】

スコアは、スコアが決定された範囲内である場合、ナビゲーションタスクが中間困難性であることを示し得る。範囲を決定することは、複数のナビゲーションタスクを生成することと、機械学習モデルを使用して、複数のナビゲーションタスクをスコアリングすることと、複数のナビゲーションタスクの各々の対応する複数のスコアに基づいて範囲を算出することとを含み得る。

【0022】

範囲は、複数のナビゲーションタスクの各々の対応する複数のスコアについての1つ又は複数の要約統計量に基づいて算出され得る。

【0023】

1つ又は複数のナビゲーションタスクを取得することは、第2のナビゲーションタスク

10

20

30

40

50

を生成することと、タスクの容易性を推定するようにトレーニングされた機械学習モデルを使用して第2のナビゲーションタスクをスコアリングすることと、スコアが第2の選択基準を満たすことに応答して、第2のナビゲーションタスクを1つ又は複数のナビゲーションタスクのうちの1つとして選択することとを含み得る。

【0024】

第2のナビゲーションタスクの生成は、ランダムな（又は擬似ランダムな）ナビゲーションタスクの生成であり得る。

【0025】

スコアは、第2のナビゲーションタスクが容易であることをスコアが示す場合、第2の選択基準を満たし得る。

【0026】

容易なナビゲーションタスクは、破滅的忘却を軽減し得るため、エージェントのトレーニングにおいて有益である。例えば、トレーニング中、エージェントは、困難性のより高いタスクについてのみトレーニングされている場合、より容易なナビゲーションタスクがどのように実行され得るかを忘れ得る。

【0027】

スコアは、スコアが閾値スコアより大きい場合、ナビゲーションタスクが容易であることを示し得る。閾値スコアを決定することは、複数のナビゲーションタスクを生成することと、機械学習モデルを使用して、複数のナビゲーションタスクをスコアリングすることと、複数のナビゲーションタスクの各々の対応する複数のスコアに基づいて閾値スコアを算出することとを含み得る。

【0028】

閾値スコアは、複数のナビゲーションタスクの各々の対応する複数のスコアの1つ又は複数の要約統計量に基づいて算出され得る。

【0029】

1つ又は複数のナビゲーションタスクを取得することは、第3のナビゲーションタスクを生成することと、第3のナビゲーションタスクを1つ又は複数のナビゲーションタスクのうちの1つとして選択することとを含み得る。

【0030】

第3のナビゲーションタスクの生成は、ランダムな（又は擬似ランダムな）ナビゲーションタスクの生成であり得る。

【0031】

ランダムな（擬似ランダムな）ナビゲーションタスクは、エージェントが特定のタイプのナビゲーションタスクに過剰適合されることを防ぎ得るエントロピーをエージェントのトレーニングに導入するため、エージェントのトレーニングにおいて有益である。

【0032】

エンティティが、エンティティの環境の画像を使用して1つ又は複数のナビゲーションタスクを実行しようと試みることは、1つ又は複数の行動ステップを実行することを含み得る。各行動ステップは、エンティティの環境の1つ又は複数の画像を受け取ることと、エージェントを使用して、1つ又は複数の画像に基づいてエンティティの環境においてエンティティが実行可能な複数の行動のうちの1つ又は複数の行動を決定することと、1つ又は複数の行動を実行させることと、第1のコンテキストの更新状態とナビゲーションタスク成功基準とに基づいて、強化学習方法を使用してエージェントをトレーニングすることに対する報酬を算出することとを含み得る。

【0033】

第1のコンテキストの更新状態が第1のコンテキストの以前の状態とは異なることに応答して、算出される報酬が増やされ得る。

【0034】

更新状態が以前の状態とは異なることに応答して算出される報酬を増やすことで、探索を促進し、トレーニングエージェントがトレーニング中に極小に捕らえられることを防ぎ

10

20

30

40

50

得る。

【0035】

算出される報酬は、また、移動した距離に応答して増やされ得る。

【0036】

移動した距離に応答して算出される報酬を増やすことで、探査が促進され、エージェントがその現在のロケーションに留まること、及び／又は、成功基準を達成する前にスピンすることを防ぎ得る。

【0037】

算出される報酬は、トレーニングエンティティがトレーニング環境において1つ又は複数のオブジェクトと衝突したことを第1のコンテキストの更新状態が示すことに応答して減らされる。

10

【0038】

衝突に応答して算出される報酬を減らすことで、衝突なくナビゲートするように又は少なくとも他の場合に発生するよりも少ない衝突でナビゲートするように装置を制御するエージェントのトレーニングにつながる。これは、そのような衝突から生じる装置に対する損傷又は少なくとも損耗を低減し、装置が衝突したであろうオブジェクトに対する損傷を低減する。衝突があれば装置の運動量が低減したため、装置のエネルギー消費も低減され得る。

【0039】

第1のエンティティは、シミュレートエンティティであり得る。第1のエンティティの環境は、シミュレート環境であり得る。

20

【0040】

シミュレートエンティティとシミュレート環境とを使用してエージェントをトレーニングすることで、エージェントのトレーニングにおける装置の使用が不要になるか又は少なくとも低減される。結果として、トレーニングする際に上記装置によって行動が実行されないため、エージェントのトレーニングにおけるエネルギー使用量が低減する。更に、シミュレート環境においてシミュレートエンティティが行動を実行するほうが、物理的環境において装置が行動を実行するよりも速い、典型的には桁違いに速いため、シミュレート環境においてエージェントをトレーニングすることは著しく速いであろう。

【0041】

エージェントは、1つ又は複数の潜在的特徴を決定するために1つ又は複数の画像を処理するように構成された1つ又は複数の特徴ネットワークを含み得る。

30

【0042】

エージェントは、エンティティの環境においてエンティティが実行可能な1つ又は複数の行動の1つ又は複数の行動値を決定するために1つ又は複数の潜在的特徴を処理するように構成されたポリシーネットワークを更に含み得る。

【0043】

1つ又は複数の特徴ネットワークは、物理的環境の複数の画像を使用して事前トレーニングされ得る。

【0044】

物理的環境の複数の画像を使用して事前トレーニングされた特徴ネットワークを使用することで、物理的環境において装置を制御することができるエージェントをシミュレート環境においてトレーニングすることが容易になる。

40

【0045】

1つ又は複数の特徴ネットワークのパラメータは、エージェントをトレーニングするとき凍結され得る。

【0046】

トレーニング中、1つ又は複数の特徴ネットワークのパラメータを凍結することで、1つ又は複数の特徴ネットワークがシミュレート環境に過剰適合することを防ぐか又は少なくとも低減する。従って、エージェントがシミュレート環境にまったく又はあまり過剰適

50

合されないため、エージェントは、物理的環境において装置をより良好に制御することができる。

【0047】

1つ又は複数の特徴ネットワークは、1つ又は複数の空間特徴ネットワークを含み得る。

【0048】

1つ又は複数の空間特徴ネットワークは、1つ又は複数の畳み込みニューラルネットワーク層を含み得る。

【0049】

1つ又は複数の特徴ネットワークは、1つ又は複数のセマンティックターゲット特徴ネットワークを含み得る。

【0050】

1つ又は複数のセマンティックターゲット特徴ネットワークは、1つ又は複数の畳み込みニューラルネットワーク層を含み得る。

【0051】

1つ又は複数のセマンティックターゲット特徴ネットワークは、1つ又は複数の後処理層を含み得る。1つ又は複数の後処理層は、1つ又は複数のオブジェクトクラスのうちのオブジェクトクラスのためのバイナリマスクを抽出するように構成され得、このオブジェクトクラスは、設定可能なラベルに基づいている。

【0052】

設定可能なラベルに基づく1つ又は複数のオブジェクトクラスの指定は、ランタイムにオブジェクトクラスを設定し、従って、トレーニングされたエージェントを使用して、エージェントのトレーニング中に見られなかったオブジェクトクラスのターゲットオブジェクトにナビゲートするように装置を制御することを容易にし得る。

【0053】

ポリシーネットワークは、特徴圧縮セクションを含み得る。

【0054】

ポリシーネットワークは、メモリセクションを含み得る。メモリセクションは、1つ又は複数の長期メモリ（LSTM）層及び／又は1つ又は複数の勾配回帰ユニット（GRU）層を含み得る。

【0055】

第2の実施形態によれば、装置の物理的環境においてナビゲーションタスクを実行するように装置を制御するためのコンピュータ実施方法が提供される。装置は、装置の物理的環境の画像を受け取るように適応される。装置は、装置の方向を制御するように適応されたステアリングデバイスを含む。方法は、装置の物理的環境の1つ又は複数の画像を受け取ることと、第1の実施形態の方法に従ってトレーニングされたエージェントを使用して、1つ又は複数の画像に基づいて装置が実行可能な複数の行動のうちの1つ又は複数の行動を決定することと、装置に1つ又は複数の行動を実行させることとを含む。

【0056】

第3の実施形態によれば、第2の実施形態による方法を実行するように構成された1つ又は複数のプロセッサを含むシステムが提供される。

【0057】

システムは、方法によって制御される装置を含み得る。

【0058】

前述の3つの態様はナビゲーションタスクに関するものであるが、本明細書で説明する詳細が様々なタスクタイプのコンテキストで適用され得ることは理解されるべきである。

【0059】

第4の実施形態によれば、エージェントをトレーニングするためのコンピュータ実施方法が提供され、この方法は、以下を含む：1つ又は複数のタスクを取得することであって、ここにおいて、1つ又は複数のタスクの各タスクが1つ又は複数の成功基準を定義し、タスクを生成することと、タスク容易性推定関数を使用して、生成されたタスクの容易性

10

20

30

40

50

を推定することと、生成されたタスクの推定された容易性が中間容易性範囲内であると決定することと、生成されたタスクの容易性が中間容易性範囲内であると決定することに対応して、生成されたタスクを1つ又は複数のタスクのうちのタスクとして選択することを含む取得すること；トレーニングコンテキストにおいてエージェントをトレーニングすることであって、1つ又は複数のタスクの各タスクについて、エージェントによってタスクの実行を試行することと、ここにおいて、これは、タスクによって定義された1つ又は複数の成功基準のうちの少なくとも1つが満たされること又は1つ又は複数の失敗基準のうちの少なくとも1つが満たされることに応答して終了する、エージェントによる各タスクの実行の試行に基づいてエージェントのパラメータを更新することを含むトレーニングすること；1つ又は複数のタスクの各タスクについて1つ又は複数のタスク特性を決定すること；及び、1つ又は複数のタスクの各タスクについて、タスクの1つ又は複数の特性と、タスクの実行を試行した結果、1つ又は複数の成功基準のうちの少なくとも1つが満たされたかどうかを示す指示とを含むトレーニングセットに基づいてタスク容易性推定関数をトレーニングすること。

10

【0060】

タスク容易性推定関数をトレーニングすることは、損失関数を使用して、トレーニングセットに基づいて1つ又は複数の損失を算出することと、1つ又は複数の損失に基づいてタスク容易性推定関数のパラメータを更新することを含む。

【0061】

タスクを生成することは、ランダムなタスクを生成することを含む。

20

【0062】

方法は、中間容易性範囲を決定することを更に含む。中間容易性範囲を決定することは、複数のタスクを生成することと、タスク容易性推定関数を使用して、複数のランダムなタスクの各々の対応する複数の容易性を推定することと、複数のランダムなタスクの各々の対応する複数の推定された容易性に基づいて、中間容易性範囲を算出することを含む。

【0063】

1つ又は複数のタスクを取得することは、第2のタスクを生成することと、タスク容易性推定関数を使用して、第2の生成されたタスクの容易性を推定することと、第2の生成されたタスクの推定された容易性が閾値容易性より大きいと決定することと、第2の生成されたタスクの容易性が閾値容易性より大きいと決定することに対応して、第2の生成されたタスクを1つ又は複数のタスクのうちのタスクとして選択することを含む。

30

【0064】

1つ又は複数のタスクを取得することは、第3のタスクを生成することと、第3の生成されたタスクを1つ又は複数のタスクのうちのタスクとして選択することを含む。

【0065】

推定された容易性関数は、フィードフォワードニューラルネットワークを含む。

【0066】

トレーニングコンテキストは、トレーニング環境及びトレーニングエンティティを含む。

40

【0067】

1つ又は複数の成功基準は、トレーニングコンテキストの状態が、トレーニングエンティティのターゲット状態に等しいトレーニングエンティティの状態を備える場合に満たされる成功基準を含む。

【0068】

1つ又は複数の失敗基準は、タイムステップの最大数に達した場合に満たされる失敗基準を含む。

【0069】

エージェントによってタスクの実行を試行することは、1つ又は複数の行動ステップを実行することを含む。各行動ステップは、トレーニング環境の1つ又は複数の観察を

50

受け取ることと、エージェントを使用して、1つ又は複数の観察に基づいて、トレーニング環境においてトレーニングエンティティが実行可能な複数の行動のうちの1つ又は複数の行動を決定することと、トレーニング環境においてトレーニングエンティティに1つ又は複数の行動を実行させることと、トレーニングコンテキストの更新状態とタスクによって定義された1つ又は複数の成功基準とに基づいて報酬を算出することを含み得る。

【0070】

エージェントによる各タスクの実行の試行に基づいてエージェントのパラメータを更新することは、1つ又は複数のタスクの各タスクの1つ又は複数の行動ステップの各行動ステップについて、行動ステップの開始時のトレーニングエンティティの状態の表現と、選択された1つ又は複数の行動と、算出される報酬とを備えるトレーニングセットに基づいて、エージェントのパラメータを更新することを含み得る。

10

【0071】

算出される報酬は、1つ又は複数の成功基準のうちの少なくとも1つが満たされることに応答して増やされ得る。

【0072】

算出される報酬は、トレーニングコンテキストの更新状態が、トレーニングエンティティの以前の状態とは異なるトレーニングエンティティの更新状態を含むことに応答して増やされ得る。

【0073】

算出される報酬は、トレーニングエンティティがトレーニング環境において1つ又は複数のオブジェクトと衝突したことをトレーニングコンテキストの更新状態が示すことに応答して減らされ得る。

20

【0074】

エージェントは、1つ又は複数の特徴ネットワークと1つ又は複数のポリシーネットワークとを含み得る。1つ又は複数の行動を決定することは、1つ又は複数の潜在的特徴を決定するために、1つ又は複数の特徴ネットワークを使用して、1つ又は複数の観察を処理することと、トレーニング環境においてトレーニングエンティティが実行可能な複数の行動の複数の行動値を決定するために、ポリシーネットワークを使用して、1つ又は複数の潜在的特徴を処理することを含み得る。

【0075】

1つ又は複数の特徴ネットワークは、1つ又は複数の実環境の複数の観察を使用して事前トレーニングされ得る。

30

【0076】

1つ又は複数の実環境の複数の観察は、複数の実際の画像であり得る。

【0077】

1つ又は複数の特徴ネットワークの1つ又は複数のパラメータは、エージェントをトレーニングするときに凍結され得る。

【0078】

1つ又は複数の特徴ネットワークは、1つ又は複数の空間特徴ネットワークを含み得る。

【0079】

1つ又は複数の特徴ネットワークは、1つ又は複数のセマンティックターゲット特徴ネットワークを含み得る。

40

【0080】

1つ又は複数のセマンティックターゲット特徴ネットワークは、1つ又は複数のオブジェクトクラスのうちのオブジェクトクラスのためのバイナリマスクを抽象化する (abstract) ように構成された1つ又は複数の後処理層を含み得、このオブジェクトクラスは、設定可能なラベルに基づいている。

【0081】

1つ又は複数の後処理層は、バイナリマスクに基づいて、出力に対して最大プーリング動作を適用するように更に構成され得る。

50

【0082】

ポリシーネットワークは、特徴圧縮セクションを含み得る。

【0083】

ポリシーネットワークは、メモリセクションを含み得る。メモリセクションは、1つ又は複数のLSTM層及び／又は1つ又は複数のGRU層を含み得る。

【0084】

トレーニングエンティティは、シミュレートエンティティであり得る。

【0085】

トレーニング環境は、シミュレート環境であり得る。

【0086】

第5の実施形態によれば、プロダクション環境においてタスクを実行するようにエンティティを制御するためのコンピュータ実施方法が提供される。方法は、プロダクション環境の1つ又は複数の観察を受け取ることと、第4の実施形態による方法に従ってトレーニングされたエージェントを使用して、プロダクション環境の1つ又は複数の観察に基づいて、プロダクション環境においてエンティティが実行可能な複数の行動のうちの1つ又は複数の行動を決定することと、プロダクション環境においてエンティティに1つ又は複数の行動を実行させることとを含む。

【0087】

エンティティは、移動可能なエンティティであり得る。複数の行動は、1つ又は複数の移動行動(movement action)を含み得る。

【0088】

プロダクション環境においてエンティティを制御することは、プロダクション環境を通じてエンティティをナビゲートすることを含み得る。

【0089】

プロダクション環境は、物理的環境であり得る。

【0090】

エンティティは、装置であり得る。

【0091】

第6の実施形態によれば、コンピュータプログラムが提供され、これは、オプションで非一時的なコンピュータ読取可能な媒体に記憶されている。コンピュータプログラムは、プログラムがコンピューティング装置によって実行されると、コンピューティング装置に、第4の実施形態による方法を実行させる命令を含む。

【0092】

第7の実施形態によれば、第4の実施形態による方法を実行するように構成された1つ又は複数のプロセッサを含むシステムが提供される。

【0093】

システムは、エンティティを含み得る。

【0094】

第8の実施形態によれば、コンピュータ実施方法が提供され、この方法は、1つ又は複数のタスクを取得することと、ここにおいて、1つ又は複数のタスクの各タスクは、初期状態及び1つ又は複数の成功基準を定義する、1つ又は複数のタスクの各タスクについて、シミュレート環境においてシミュレートエンティティをシミュレートするように構成されたシミュレータを使用してポリシーネットワークをトレーニングすることとを含み、以下を含む：タスクによって定義された初期状態を示す指示をシミュレータに送ること、ここにおいて、タスクによって定義された初期状態を示す指示を受け取ったことに応答して、シミュレータは、初期状態を示す指示に基づいて、シミュレートエンティティの状態を設定する；エージェントによってタスクの実行を試行することであって、タスクによって定義された1つ又は複数の成功基準のうちの少なくとも1つが満たされるか又は1つ又は複数の失敗基準のうちの少なくとも1つが満たされるまで1つ又は複数の行動ステップを実行することを含み、各行動ステップが、シミュレートエンティティの状態に基づいてシ

10

20

30

40

50

ミュレート環境の1つ又は複数のビューを表す1つ又は複数の画像をシミュレータから受け取ることと、1つ又は複数の潜在的特徴を決定するために、1つ又は複数の特徴ネットワークを使用して1つ又は複数の画像を処理することと、シミュレート環境においてシミュレートエンティティが実行可能な複数の行動の複数の行動値を決定するために、ポリシーネットワークを使用して1つ又は複数の潜在的特徴を処理することと、複数の行動のうちの1つ又は複数の行動を、複数の行動値に基づいて選択することと、選択された1つ又は複数の行動を示す指示をシミュレータに送ることと、ここにおいて、選択された1つ又は複数の行動を示す指示を受け取ったことに応答して、シミュレータは、選択された1つ又は複数の行動に基づいてシミュレートエンティティの状態を更新する、シミュレートエンティティの更新状態と、タスクによって定義されたターゲット状態とに基づいて報酬を算出することを含む、試行すること；並びに、1つ又は複数のタスクのうちの各タスクの1つ又は複数の行動ステップの各行動ステップについて、行動ステップの開始時のシミュレートエンティティの状態の表現と、選択された1つ又は複数の行動と、算出される報酬との表現を含むトレーニングセットに基づいてポリシーネットワークのパラメータを更新すること。

【0095】

本明細書で説明する実施形態の理解を助けるために、具体的な例が最初に提示される。この具体的な例はこれらの実施形態の理解を助けるために提供されることが理解されるべきであり、以降の特許請求の範囲又は本明細書の範囲を限定するものと見なされるべきではない。

【0096】

自律走行車両

図1Aは、例となる実施形態による、自律走行車両1、例えば自動搬送車の例示である。

【0097】

自律走行車両1は、本体2と、車輪3と、カメラ4とを含む。自律走行車両1は、例えば、自動走行車両を自動掃除機として使用することができるような真空吸引要素、及び／又は、自律走行車両がオブジェクトを収集及び／又は運搬することができるようなフォーク及び／又は荷台など、図示されたもの以外の追加の要素を含み得る。

【0098】

本体2は、車輪3を回転させるために使用され得る1つ又は複数のモータと、ステアリングシステム、例えばラックアンドピニオンとを含み、それによって、車輪3のうちの少なくともいくつかを転向させて、所望の方向に移動することができる。本体2はまた、制御システム、例えば動作を実行させるために使用され得る1つ又は複数のコンピューティングデバイスを含む。例えば、制御システムは、車輪をモータによって回転させて、自動走行車両を前方に移動させ（そして、サポートされている場合には、車輪を反対方向に回転させて、自律車両を後方に移動させ）、ステアリングシステムに車輪を回転させて、自律車両を方向転換し得る。制御システムは、例えば図1Dに関して説明する方法を使用して、それ自体で実行されるべき動作を決定し得るか、又はネットワークハードウェア、例えばワイヤレストランシーバを使用して、外部コンピュータシステムから、実行されるべき行動を示す指示を受信し得る。本体2はまた、1つ又は複数のモータ、ステアリングシステム、及び制御システムにエネルギーを供給することができるように、電源（例えばバッテリー及び／又はソーラーパネル）又は電源への接続（例えば電源接続部）を含む。

【0099】

車輪3は、車輪3a、3bと、これらに対向する、例えば車輪3a、3bに対して本体2の反対側（図示せず）に対して配置された2つの車輪（図示せず）という4つの車輪を含み得る。4つの車輪のうちの各2つの車輪、例えば車輪3aと反対側の車輪及び車輪3bと反対側の車輪は、本体2を通して延在する車軸によって接続され得る。車軸のうちの少なくとも1つは、それに接続されたそれぞれの2つの車輪がモータによって回転され得るように、モータに接続され得る。車軸のうちの少なくとも1つは、それぞれの2つの車輪がステアリングシステムを使用して転向され得るように、ステアリングシステムに接続

され得る。

【0100】

カメラ4は、画像をキャプチャすることができる。この実施形態では、2D画像をキャプチャする単一のカメラが提供される。カメラは、本体2内の制御システムに結合される。カメラは、例えば制御システムから受け取った要求に応答して、1つ又は複数の画像をキャプチャし、1つ又は複数の画像を制御システムに送るように構成される。制御システムは、実行されるべき1つ又は複数の動作を1つ又は複数の画像に基づいて決定するためにそれ自体で1つ又は複数の画像を処理し得るか、又は、外部コンピュータシステムが、実行されるべき1つ又は複数の動作を1つ又は複数の画像に基づいて決定することができるように、ワイヤレストランシーバのようなネットワークハードウェアを使用して、1つ又は複数の画像を外部コンピュータシステムに送信し得る。

10

【0101】

自律航空車両

図1Bは、例となる実施形態による、自律航空車両5、例えば無人航空機又はドローンの例示である。

【0102】

自律航空車両5は、本体6と、プロペラ7と、カメラ8とを含む。自律航空車両5は、図示されたもの以外の追加の要素を含み得る。例えば、オブジェクトを収集及び／又は運搬するために自律航空車両を使用することができるようなペイロード保持又は取付け要素である。

20

【0103】

本体6は、それぞれのプロペラを回転させるために使用され得るプロペラ7の各々のためのそれぞれのモータを含む。本体6はまた、制御システム、例えば動作を実行させるために使用され得る1つ又は複数のコンピューティングデバイスを含む。例えば、制御システムは、モータにプロペラを回転させること、及び、モータがプロペラを回転させる速度を制御することができ、それによって、自律航空車両は、上昇及び下降すること、前方に移動すること及び／又は後方に移動すること、及び方向転換することができる。制御システムは、例えば図1Dに関して説明する方法を使用して、それ自体で実行されるべき動作を決定し得るか、又はネットワークハードウェア、例えばワイヤレストランシーバを使用して、外部コンピュータシステムから、実行されるべき行動を示す指示を受信し得る。決定及び／又は受信された動作は、自律航空車両が実行可能な動作よりも抽象度が高い可能性があり、制御システムは、これらのより抽象的な動作を、航空車両が実行可能な動作に変換し得る。例えば、より抽象的な動作は、前方への移動、左への方向転換、及び右への方向転換であり得、制御システムは、これらを、プロペラ7がモータによって回転される速度の差に変換し得る。本体6はまた、1つ又は複数のモータ及び制御システムにエネルギーを供給することができるように、電源（例えばバッテリー及び／又はソーラーパネル）又は電源への接続（例えば電源接続部）を含む。

30

【0104】

プロペラ7は、4つのプロペラ7a、7b、7c、及び7dを含み得る。説明したように、プロペラ7a～7dの各々は、本体6にあるそれぞれのモータによって回転させられ得、プロペラ7a～7dのうちの1つ又は複数のは、自律航空車両5が左又は右に方向転換したり前方に移動したりするように、プロペラ7a～7dのうちの別の1つ又は複数とは異なる速度で回転させられ得る。

40

【0105】

カメラ8は、画像をキャプチャすることができる。この実施形態では、2D画像をキャプチャする単一のカメラが提供される。カメラは、本体6内の制御システムに結合される。カメラは、例えば制御システムから受け取った要求に応答して、1つ又は複数の画像をキャプチャし、1つ又は複数の画像を制御システムに送るように構成される。制御システムは、実行されるべき1つ又は複数の動作を1つ又は複数の画像に基づいて決定するためにそれ自体で1つ又は複数の画像を処理し得るか、又は、外部コンピュータシステム

50

が、実行されるべき1つ又は複数の行動を1つ又は複数の画像に基づいて決定することができるよう、ワイヤレストランシーバのようなネットワークハードウェアを使用して、1つ又は複数の画像を外部コンピュータシステムに送信し得る。

【0106】

例となる環境

図1Cは、例となる実施形態による、エージェントを使用してナビゲートされ得る例となる環境10の平面図を例示したものである。

【0107】

例となる環境10は、自律車両11、例えば自律走行車両1又は自律航空車両5と、自律車両によってその現在位置からナビゲートされる先のターゲットオブジェクト12、例えばボール又は花瓶とを含む。例示するように、自律車両11とターゲットオブジェクト12との間に障害物が存在する。自律車両11及び障害物への損傷を回避し、障害物に引っかからないようにするために、自律車両11は、ターゲットオブジェクト12にナビゲートするときに障害物との衝突を回避する。

【0108】

例となる環境10は、雑然とした屋内環境、例えばオフィス又は自宅であり得、テーブル、椅子、棚、書棚、収納設備、箱、電子デバイス、及び電力ケーブルのような障害物を含み得る。

【0109】

例となるナビゲーション方法

図1Dは、例となる実施形態による、例となる環境、例えば例となる環境11において、自律車両、例えば自律走行車両1又は自律航空車両5をナビゲートするための方法21のフロー図である。例となる方法は、1つ又は複数のコンピューティングデバイス、例えば図7に関連して説明するコンピューティングデバイス700によって実行される1つ又は複数のコンピュータ実行可能命令として実施され得る。1つ又は複数のコンピューティングデバイスは、自律車両の内部にある、例えば本体内に位置する制御システムであり得るか、又は自律車両の外部にあり得る。

【0110】

ステップ21において、環境の1つ又は複数の画像を受け取る。1つ又は複数の画像は、自律車両のカメラ、例えばカメラ4又はカメラ8を使用してキャプチャされた1つ又は複数の画像である。方法を実行する1つ又は複数のコンピューティングデバイスが自律車両の内部にある場合、カメラと1つ又は複数のコンピューティングデバイスとの間の1つ又は複数のワイヤード接続を介して1つ又は複数の画像を受け取り得る。方法を実行する1つ又は複数のコンピューティングデバイスが自律車両の外部にある場合、1つ又は複数の画像は、自律車両のネットワークハードウェア、例えばワイヤレストランシーバを使用して1つ又は複数のコンピューティングデバイスに送信され得る。

【0111】

ステップ22において、環境において実行されるべき1つ又は複数の動作が、エージェントを使用して、1つ又は複数の画像に基づいて決定される。

【0112】

エージェントは、図2及び図3A～図3Cに関連して説明するような自動タスク選択方法を使用して取得されたタスクを使用してトレーニングされている。タスクは、エージェントが、障害物を含む環境、例えば環境10において、エンティティ、例えば自律車両11を、その制御下で、ターゲットオブジェクト、例えばターゲットオブジェクト12にナビゲートしようと試みるナビゲーションタスクである。エージェントは、これらのタスクの実行を試行し、タスクにおけるエージェントのパフォーマンス、例えばターゲットオブジェクトに成功裏に到達したかオブジェクトと衝突したかに基づいて、例えば強化学習方法を使用して、エージェントのパラメータを更新することによってトレーニングされている。自動タスク選択方法で取得されたタスクを使用してトレーニングされたエージェントは、エージェントにとって適切な困難性のタスク、例えばエージェントが実行するのに適

10

20

30

40

50

度に容易であると決定されたタスクについてトレーニングされる。適切な困難性のタスクを使用してトレーニングされたエージェントは、その制御下にあるエンティティをターゲットオブジェクトにナビゲートするのに好適な動作をより良好に決定することができ、その結果、エージェントは、エンティティをターゲットオブジェクトにより良好にナビゲートすることができ、例えば自律車両をターゲットオブジェクトにより良好にナビゲートすることができる。

【0113】

エージェントは、シミュレータを使用してトレーニングされていてもよい。例えば、エージェントは、シミュレート環境、例えばシミュレートされた雑然とした屋内環境において、シミュレートされた車両、例えばシミュレートされた走行車両を移動させることでタスクを実行しようと試みることによってトレーニングされていてもよい。しかしながら、エージェントは、シミュレートされた車両のカメラビューを表す実質的にフォトリアリスティックな画像をレンダリングすることができるシミュレータを使用してトレーニングされていてもよく、シミュレータでトレーニングされているにもかかわらず、エージェントが1つ又は複数の画像に基づいて自律車両をナビゲートするのに好適な動作を決定することができるように、実際の画像を使用して事前トレーニングされた特徴ネットワークを含んでいてもよい。

【0114】

エージェントは、指定されたセマンティックタイプのターゲットオブジェクト、例えば花瓶に自律車両をナビゲートすることができるように設定可能であり得る。この機能性は、エージェントが、コンテキスト内の日常的なオブジェクトの実際の画像について事前トレーニングされたセマンティックターゲット特徴ネットワークを含むことによって容易にされ得、これは、ターゲットとなるオブジェクトのセマンティックタイプを指定するラベルを設定することによって構成され得る。そのようなエージェントは、トレーニング中にエージェントによって試行されたタスクに含まれていなかったセマンティックタイプのターゲットオブジェクトにナビゲートすることが可能であり得る。例えば、エージェントは、ボールをターゲットオブジェクトとしてナビゲートするタスクだけを試行していたが、花瓶をターゲットとすべきであることを指定するようにラベルを設定することによって、自律車両を花瓶にナビゲートし得る。

【0115】

説明する方法を使用してトレーニングされたエージェントは、所与のエンティティ又はそのタイプで使用するためにトレーニングされたエージェントが、他のエンティティ又はそのタイプに使用されることができるよう、一般化することができる。例えば、自律走行車両をナビゲートするためにトレーニングされたエージェント、例えばシミュレートされた走行車両を移動させることでタスクの実行を試行することによってトレーニングされたエージェントを、自律航空車両のナビゲートに使用することができる。

【0116】

説明する方法を使用してトレーニングされたエージェントは、複数のカメラ及び／又はマルチモーダル入力ナビゲーションに必要とされないように、単一のカメラを使用してキャプチャされた1つ又は複数の画像に基づいて、その制御下にある自律車両をターゲットオブジェクトにナビゲートするのに好適な行動を決定することが可能であり得る。例えば、自動走行車両1の単一のカメラ4又は自動航空車両の単一のカメラ8を使用してキャプチャされた画像が、それぞれのエンティティをナビゲートするのに十分であり得る。

【0117】

ステップ23において、自律車両に1つ又は複数の動作を実行させる。方法を実行するための1つ又は複数のコンピューティングデバイスが自律車両の内部にある場合、1つ又は複数の動作の実行を引き起こすことは、自律車両の構成要素によって好適な動作を引き起こすことを含み得る。例えば、自律車両が自律走行車両である場合には、車輪を回転させて自律走行車両を前方に移動させるか、車輪を転向させて自律走行車両を方向転換させて、自律車両が自律航空車両である場合は、プロペラを異なる速度で回転させて、自律航空

10

20

30

40

50

車両を前方に移動させたり、左に方向転換させたり、右に方向転換させたりする。方法を実行するための1つ又は複数のコンピューティングデバイスが自律車両の外部にある場合、1つ又は複数の動作の実行を引き起こすことは、実行されるべき1つ又は複数の動作を示す指示を自律車両に送信することを含み得る。自律車両の制御システムは、これらの指示を受け取り、その構成要素によって好適な動作を実行させ得る。

【0118】

上記実施形態のシステム及び方法は、タスク実行エージェント及びそのトレーニングを改善しようと試みる。実行され得るタスクの例には、ナビゲーションタスク、オブジェクトピッキングタスク、例えば倉庫及び／又は小売店におけるオブジェクトピッキング、ゲームプレイタスク、製造タスク、適応制御タスク、及び電力システム制御タスクが含まれる。これらの改善により、既存エージェントのよりも良好にタスクを実行することができるタスク実行エージェント、又は、以前のエージェントが実行することができなかったタスクを実行することができるエージェントが得られる。そのようなエージェントが既存のエージェントよりも良好にタスクを実行する場合、タスクを実行するための行動の数及び／又はタスクを試行する際にエージェントが行う試行の数は低減され得る。更に、行動及び／又はタスク試行を実行することは、計算リソースを消費し、タスクが物理的環境において装置を使用して実行されている場合、例えば装置又はその構成要素を移動させるためにエネルギーを消費し、結果として、装置の損耗を発生させ得る。従って、タスク（複数を含む）をより良好に実行するという固有の利点に加えて、提供される改善は、エージェントによるタスクの実行において、計算リソースの使用量、エネルギーの使用量、及び損耗を低減する。

【0119】

タスク実行エージェントのトレーニングの改善も、エージェントのトレーニングにおいて実行される行動及び／又はパラメータ更新の数を低減し得る。行動及びパラメータ更新を実行することは、計算リソースを消費する。従って、実行される行動及び／又はパラメータ更新の数を低減することで、エージェントのトレーニング中に消費される計算リソースが低減される。更に、トレーニングが、物理的環境において装置を使用して行動及び／又はタスクの試行を実行することを伴う場合、トレーニングは、例えば装置又はその構成要素を移動させるためにエネルギーを消費し、結果として、装置の損耗を発生させ得る。従って、実行される行動及び試行の数を低減することで、エネルギー使用及び損耗も低減する。

【0120】

タスク実行エージェントのトレーニングの特定の改善はまた、例えば、シミュレート環境において少なくとも部分的に使用してエージェントをトレーニングすることによって、物理的環境において行動又はタスク試行を実行することも、物理的環境において行動の実行及び／又はタスク試行を更に低減することもなく、物理的環境において使用するためのエージェントのトレーニングを容易にし得る。その結果、装置によるエネルギー使用及び装置の損耗がなくなり得るか、又は少なくとも大幅に低減され得る。従って、タスク（複数を含む）をより良好に実行することができるエージェントをトレーニングするという固有の利点に加えて、提供される改善は、計算リソースの使用量を低減し、エージェントのトレーニングにおけるエネルギーの使用量及び装置の損耗を低減し得る。

【0121】

説明される利点のうちの少なくともいくつかは、物理的環境において装置を使用してナビゲーションタスク又はナビゲーションが構成要素である他のタスクを実行するための改善されたエージェントを提供する実施形態において特に顕著であり得る。これは、トレーニング中にナビゲーションタスクを実行しようと試みることに、パフォーマンスが不十分なエージェントが使用されることの両方が、装置と物理的環境との衝突を発生させやすいためである。そのような衝突は、装置の損耗、衝突する環境の部分への潜在的な損傷を発生させ、例えば人間との衝突など危険な場合もある。トレーニング中に物理的環境において実行される行動及び／又は試行されるタスクを低減するか又は無くすことによって及び／

又はトレーニングされたエージェントのパフォーマンスを改善することによって、衝突を回避すること又は少なくとも最小限に抑えることができる。その結果、装置の損耗、物理的環境への潜在的な損傷、及び危険な衝突が回避され得る、又は少なくとも大幅に低減され得る。

【0122】

先に示した例から明らかとなるように、及び、明らかなように、説明される実施形態の少なくともいくつかは、カリキュラム学習の変形例に関し、特に、自動カリキュラム学習に関する。カリキュラム学習の本質は、学習者、例えばトレーニングされているエージェントの所望の特徴を作り出すように、トレーニングデータを選択して順序付けることである。自動カリキュラム学習は、人間が参加せずにこのカリキュラムを生成するプロセスである。

10

【0123】

トレーニングシステム

図1Eは、例となる実施形態による、エージェントをトレーニングするためのシステム100を例示する概略ブロック図である。システム100は、1つ又は複数のコンピューティングデバイス、例えば図7に関して説明するコンピュータハードウェア700上に、又はそれを使用して実装され得る。

【0124】

システム100は、トレーニングコンテキスト110と、タスク取得部と、トレーニングエージェント130と、タスク容易性推定関数トレーナ140とを含む。システム100のこれらの特徴は、1つ又は複数のコンピュータプログラム、例えば1つ又は複数のプロセッサによって実行可能なコンピュータ読取可能な命令、実行可能命令に解釈可能なスクリプト、及び／又は実行可能命令に変換可能な中間表現、例えばバイトコードとして実装され得る。説明の目的上、別個の特徴として例示されているが、これらの態様の任意の適切な組合せが単一のコンピュータプログラムに統合され得ることは理解されるべきである。同様に、これらの特徴のうちの1つ又は複数の複数のコンピュータプログラムを使用して実装され得ることは理解されるべきである。システム100の様々な機能は、互いに、ユーザライブラリと、システムライブラリと、オペレーティングシステムと通信し、及び／又は、任意の適切な機構（複数を含む）、例えばAPIコール、リモートサービスコール、メッセージキュー、プロセス間通信（IPC）プロトコル、共有メモリ、及び／又は永続ストレージを使用して特徴内で通信し得る。

20

30

【0125】

トレーニングコンテキスト110は、トレーニングエージェント130がタスクを実行するようにトレーニングされ得るコンテキスト、例えばトレーニングエージェント130が行動を実行させ得、コンテキストの観察が行われ得るコンテキストを提供する。トレーニングコンテキスト110は、例えばシミュレータを使用して実装され得る実際のコンテキストのシミュレーションのようなシミュレートされたコンテキストや、例えばビデオゲームのような仮想コンテキストであり得るか、又は例えば物理的コンテキストのような実際のコンテキストであり得る。単一のトレーニングコンテキストが示されているが、複数のトレーニングコンテキストが使用され得ることは理解されるべきである。例えば、第1のトレーニングコンテキストは、トレーニングエージェント130の初期のトレーニングを実行するために使用され得、1つ又は複数の他のトレーニングコンテキストは、エージェントの更なるトレーニングを実行するために使用され得る。いくつかの実施形態では、第1のトレーニングコンテキストは、シミュレートされたトレーニングコンテキストであり得、1つ又は複数の他のトレーニングコンテキストのうちの少なくとも1つは、物理的なトレーニングコンテキストであり得る。トレーニングコンテキスト110は、トレーニングエージェント130が使用されることとなるコンテキストとは異なり得る。例えば、トレーニングコンテキスト110は、シミュレートされたコンテキストであり得、トレーニングエージェント130が使用されることとなるコンテキストは、実際のコンテキストであり得る。

40

50

【0126】

トレーニングコンテキスト110は、トレーニング環境112と、トレーニングエンティティ114とを含み得る。

【0127】

トレーニング環境112は、トレーニングエンティティ114がトレーニングエージェント30によって行動を実行され得る環境である。トレーニング環境112は、例えばシミュレータを使用して実装され得る実環境のシミュレーションのようなシミュレート環境、例えばビデオゲーム環境のような仮想環境、又は例えば物理的環境のような実環境であり得る。

【0128】

トレーニング環境112は、障害物を含み得、例えばトレーニング環境112は、雑然としているであろう。トレーニング環境112がナビゲートされるためには、そのような障害物を避けてナビゲートする必要がある、そのような障害物との衝突を回避することが望ましい。そのようなトレーニング環境112は、追加的又は代替的に、相互作用可能なオブジェクト及び／又はエンティティを含み得る。いくつかのそのようなオブジェクト及び／又はエンティティは、ターゲットオブジェクト及び／又は到達及び／又は相互作用することが望まれるエンティティであり得る。ターゲットオブジェクトは、例えばボール、扇風機、花瓶、食器棚、又は机のような家庭用オブジェクト、例えば浴室、寝室、書斎、専用オフィス、又はステーションナリーカップボードのような所与のタイプの部屋、例えばコピー機、ビデオ会議機材、又はファイリングカップボードのようなオフィス用オブジェクト、例えば所与のタイプの機械若しくは組立ライン又はそれ上の位置のような産業用オブジェクトクラス、及び／又は例えば所与のタイプの医薬品又は医療機器のような医療用オブジェクトクラスのいずれかであり得る。

【0129】

例えば、トレーニング環境112は、シミュレートされた又は物理的な屋内環境、例えば、自宅、オフィス、工場、若しくは倉庫、又はそれらの一部であり得る。トレーニング環境112は、代替的に、シミュレートされた又は実際の屋外環境であるか、又はそれを追加的に含み得、例えば、トレーニング環境は、シミュレートされた又は物理的な公園、スタジアム、屋内及び屋外の両方のセクションを含む倉庫又は工業団地、又は街であり得る。トレーニング環境112は、代替的に、ビデオゲーム環境、例えば、ビデオゲームレベル、オープンワールド、マルチプレーヤマップ、ダンジョン、又は建物であり得る。ビデオゲーム環境は、2次元又は3次元であり得るか、又は2次元セクション及び3次元セクションの両方を含み得る。ビデオゲーム環境は、シミュレートされた屋内及び／又は屋外環境に類似し得る。しかしながら、少なくともいくつかの事例では、それらは実質的な違いを示し得、例えば、非現実的な物理的現象を有し、及び／又はロケーション間の瞬間的な移動を可能にし得る。

【0130】

トレーニングエンティティ114は、トレーニング環境112内でトレーニングエージェント130によって行動を実行させることができるエンティティである。トレーニングエンティティ114は、例えばシミュレートエンティティのようなシミュレートエンティティ、又は例えば装置のような現実のエンティティ、又は例えばビデオゲームキャラクタ又は車両のようなビデオゲームエンティティであり得る。

【0131】

トレーニングエンティティ114は、トレーニング環境内で移動すること又は移動させられることが可能であり得る。そのようなトレーニングエンティティ114はまた、オブジェクト及び／又は他のエンティティと相互作用することが可能であり得る。トレーニングエンティティ114は、ターゲットオブジェクト及び／又はターゲットである他のエンティティに到達し及び／又はそれらと相互作用するようにトレーニングエージェント130によって制御され得る。

【0132】

例えば、トレーニングエンティティ 114 は、屋内環境のための及び／又は屋内環境内で動作するように構成されたシミュレートエンティティ又は物理的な装置、例えば、1つ又は複数の所与のタイプの装置を表すシミュレートされた可動エンティティ、自律掃除機、自動搬送車（AGV）、搬送又は運搬ロボット、人型ロボット、又は産業ロボットであり得る。トレーニングエンティティは、代替的に、ビデオゲームエンティティ、例えば、ビデオゲームでプレイ可能なキャラクタ、ノンプレイアブルキャラクター（NPC）、敵、又は車両であり得る。

【0133】

いくつかの実施形態では、トレーニングエンティティ 114 がシミュレートされたものであり得、トレーニング環境 112 が現実であり得ることに留意されたい。例えば、実環境の表現が、トレーニング環境 112 として、例えば 1つ又は複数のカメラ、ライダーデバイス、及び／又はマイクロフォンを使用して一度だけ又は連続的にキャプチャされ得、そのような環境内の装置又は人物の期待される挙動が、シミュレートエンティティ 114 を使用してシミュレートされ得る。

【0134】

タスク取得部 120 は、エージェントのトレーニング中にエージェントによって試行されるべき 1つ又は複数のタスクを取得するために使用される。

【0135】

タスク取得部 120 は、タスクジェネレータ 122 を含む。タスクジェネレータは、トレーニングコンテキスト 110 においてトレーニングエージェント 130 によって試行され得るタイプの 1つ又は複数のタスクを生成する。例えば、各タスクは、トレーニング環境 112 においてトレーニングエンティティ 114 を制御することによって、トレーニングエージェント 130 によって試行され得る。各タスクは、初期状態、例えばトレーニングコンテキスト 110 の初期状態を定義し得る。各タスクはまた、1つ複数の成功基準を定義し得る。1つ又は複数の成功基準は、トレーニングコンテキスト 110 のターゲット状態であり得るか、又はそれを含み得る。トレーニングコンテキスト 110 のターゲット状態は、トレーニングエンティティ 114 のターゲット状態及び／又はトレーニング環境 112 のターゲット状態であり得るか、又はそれらを含み得る。いくつかの事例では、各タスク又はタスクのサブセットの各々はまた、1つ又は複数のタスク固有の失敗基準を定義し得る。これらのタスク固有の失敗は、特定のタスクに対して定義され得る失敗基準であり得、1つ又は複数の一般的な失敗基準とは別のものであり得る。

【0136】

各タスクは、1つ又は複数のパラメータを含み得る。パラメータは、初期状態とタスクの 1つ又は複数の成功基準とを定義し得る。存在する場合、パラメータはまた、タスクに対して 1つ又は複数のタスク固有の失敗基準を定義し得る。これらのパラメータは、連続的な数値パラメータ、離散的な数値パラメータ、バイナリパラメータ、及び例えば複数のオプションからの選択のような選択パラメータの任意の数及び／又は任意の組合せを含み得る。数値パラメータは、好適な範囲内に制限され得る。特定の部分的な範囲もまた、可能な数値パラメータから除外され得る。タスクジェネレータ 122 は、ランダムなタスクを生成し得る。例えば、タスクパラメータの各々は、ランダムに生成され得、例えば、各パラメータは、1つ又は複数の一様分布及び／又はベルヌーイ分布又はポアソン分布のような 1つ又は複数の他の適切な確率分布に従って生成され得る。数値パラメータの場合、乱数が生成され得る。バイナリパラメータの場合、ランダムビット又はビットストリングが生成され得る。選択パラメータの場合、複数のオプションのうちのいくつか、例えば 1つ又は複数、ランダムに選択され得る。パラメータをランダムに生成するために、乱数ジェネレータ（RNG）又は擬似乱数ジェネレータ（PRNG）が使用され得る。本明細書における「ランダム」及びその派生語への言及は、擬似ランダムである、例えば数又は他のパラメータを擬似ランダムに生成する実施形態を含むと理解されるべきである。

【0137】

ナビゲーションタスクは、あるタイプのタスクの一例である。ナビゲーションタスクは

10

20

30

40

50

、初期状態をトレーニングエンティティの初期位置として定義し得、トレーニングエンティティ 114 のターゲット位置を含む 1 つ又は複数の成功基準を定義し得る。初期位置及びターゲット位置の各々は、座標、例えば 2 D 又は 3 D 座標を使用して表され得る。初期位置とターゲット位置の座標は、ランダムに生成され得る。ランダムに生成される座標は、ナビゲート可能なエリア内に制限され得、例えば、障害物が位置している座標を除外し、及び／又はトレーニング環境 112 の境界に対応する所与の座標範囲内に制限されている。ターゲット位置は、ターゲットオブジェクトが位置付けられることとなる位置を示し得る。ナビゲーションタスクは、追加的に、ターゲットオブジェクトのタイプ、例えばターゲットオブジェクトのセマンティッククラスを含み得る。ナビゲーションタスクはまた、1 つ又は複数のタスク固有の失敗基準を定義し得る。ナビゲーションタスクの 1 つ又は複数のタスク固有の失敗基準は、衝突回避基準、例えば衝突が発生した場合にタスクを実行しようとする試行が失敗する基準、最大移動基準、例えばエンティティが最大量以上移動した場合にタスクを実行しようとする試行が失敗する基準、最大エネルギー使用基準、例えば試行においてエンティティによって最大量以上のエネルギーが使用された場合にタスクを実行しようとする試行が失敗する基準、及び／又は時間制限基準、例えばタスクを実行しようとする試行の開始から最大量以上の時間が経過した場合にタスクを実行しようとする試行が失敗する基準を含み得る。

【0138】

ナビゲーションタスクの変形例は、マルチストップナビゲーションタスク、収集タスク、運搬タスク、及び収集と運搬の複合タスクを含む。マルチストップナビゲーションタスクでは、1 つ又は複数の成功基準は、タスクを成功裏に完了するために、トレーニングエンティティが複数のターゲット位置を通過することを定義する。収集タスクでは、1 つ又は複数の成功基準は、タスクを成功裏に完了するために、例えば、カメラ若しくは重量センサ、そのシミュレーション、又はシミュレータ状態を使用して、オブジェクトがトレーニングエンティティ 114 によってピックアップされたか又はそれに積まれたことを認識することによって、所与のオブジェクト（複数を含む）がトレーニングエンティティ 114 によって収集されることを定義する。収集タスクの場合、初期状態は、トレーニング環境内の、収集されるべき所与のオブジェクト（複数を含む）の位置（複数を含む）を定義し得る。運搬タスクでは、1 つ又は複数の成功基準は、タスクを成功裏に完了するために、例えば、カメラ若しくは重量センサ、そのシミュレーション、又はシミュレータ状態を使用して、オブジェクトがトレーニングエンティティ 114 によって又はそれから降ろされたことを認識することによって、所与のオブジェクト（複数を含む）がトレーニングエンティティ 114 によって運搬されることを定義する。運搬タスクの場合、初期状態は、タスクの開始時に所与のオブジェクト（複数を含む）がトレーニングエンティティ 114 に積まれている状態を定義し得る。少なくとも 1 つのロケーションからの収集及び少なくとも 1 つの他の場所への運搬を伴う、収集と運搬の複合タスクは、従って、マルチストップナビゲーションタスク、収集タスク、及び運搬タスクに関する特徴を含む。

【0139】

アセンブリタスクは、タスクのタイプの別の例である。アセンブリタスクは、アセンブルされるべき構成要素の集合体としてのトレーニングコンテキスト 114 の初期状態と、トレーニング環境 112 内でのこれらの構成要素の初期ロケーション（例えば座標を使用して表される）とを定義し得る。定義された初期状態はまた、トレーニングエンティティ 114 の初期位置を含み得るが、いくつかの実施形態では、トレーニングエンティティの位置は固定され得、例えば、トレーニングエンティティ 114 は、固定された産業用組立ロボット又はそのシミュレーションである。1 つ又は複数の成功基準は、アセンブリの完了時の構成要素の相対位置を定義し得る。

【0140】

タスク取得部 120 は、タスク容易性推定部 124、例えばタスクの容易性を推定するようにトレーニングされた機械学習モデルを含む。タスク容易性推定部 124 は、タスクをスコアリングし、例えば、タスク容易性推定部 124 は、タスク及び／又はそれに関す

10

20

30

40

50

る情報に基づいてタスクの容易性を推定する。タスク容易性は、タスク容易性値、タスク困難性値、タスク成功確率（例えば、タスクがトレーニングエージェント130によって成功裏に完了される推定確率）、及び／又はタスク失敗確率（例えば、タスクがトレーニングエージェント130によって成功裏に完了されない推定確率）のうちの1つ又は複数として定義され得るか、又はそれらに基づき得る。

【0141】

タスク容易性推定部124は、タスク特性決定部124-1を含む。タスク特性決定部124-1は、タスクのパラメータと、いくつかの実施形態では、トレーニングコンテキスト情報、例えばトレーニングコンテキスト110又はその構成要素、例えばトレーニング環境112又はトレーニングエンティティ114から検索された又は受け取った情報とに基づいて、1つ又は複数のタスク特性を決定する。1つ又は複数のタスク特性は、例えば1つ又は複数の幾何学的特性、タスクパラメータ若しくはその関数、及び／又はトレーニングエンティティ情報及びトレーニング環境情報の一方又は両方を使用して及び／又はそれに基づいて導出される1つ又は複数の特性のいずれか又は任意の組合せを含み得る。

【0142】

トレーニングエンティティの初期位置 s_0 を含む初期状態と、トレーニングエンティティのターゲット位置 s_g を含む1つ又は複数の成功基準とを定義するタスク、例えばナビゲーションタスクについて、決定され得る幾何学的特性の例は、測地的距離、すなわち s_0 から s_g までの最短経路距離；経路複雑度、すなわち s_0 、 s_g の測地的距離に対するユークリッド距離の比；方向転換の角度の正弦、ここで、方向転換の角度は、焦点平面法線と、原点から s_g へのベクトルとの間の角度、例えばトレーニングエンティティがターゲット位置を見るために初期の位置においてどれだけ方向転換しなければならないか、である；方向転換の角度の余弦；エージェントクリアランス、すなわち初期位置 s_0 から最も近い障害物までの距離；ゴールクリアランス、すなわちターゲット位置 s_g から最も近い障害物までの距離；エージェントアイランド、すなわち初期位置 s_0 において通過可能なエリアの半径；及びゴールアイランド、すなわちターゲット位置 s_g において通過可能なエリアの半径を含む。トレーニングコンテキスト110がシミュレートされたコンテキスト又は仮想コンテキストである場合、これらの特性は、トレーニングコンテキスト110から検索可能な及び／又は受け取り可能な情報を使用して容易に決定され得る。例えば、仮想トレーニング環境又はシミュレートされたトレーニング環境における障害物の位置に関する幾何学的情報は容易に取得可能であり得、説明された特性のうちの1つ又は複数を導出するために使用され得る。実環境、例えば物理的環境の場合、これらの特性を決定するための情報は、適切な感知及び／又はキャプチャデバイス、例えば1つ又は複数のカメラ、ライダーデバイス、近接センサ、及び／又は超音波センサを使用して取得され得る。代替的又は追加的に、実環境に関するそのような情報は、1つ又は複数の人間の観察に基づいて手動で入力されていてもよい。

【0143】

説明したタイプのタスクの場合、トレーニングエンティティ情報及び／又はトレーニング環境情報を使用して及び／又はそれらに基づいて導出され得る特性の例は、トレーニングエンティティの幅と、初期位置 s_0 からターゲット位置 s_g に到達するために通過することとなる最も狭い通路との間の比及び／又はそれらの差、及び／又は、トレーニングエンティティ114の加速度及び／又は最大速度の既知の特性と最短経路長とに基づいて算出され得る最短タスク完了時間を含み得る。

【0144】

収集と運搬タスクの場合、トレーニングエンティティ情報及び／又はトレーニング環境情報を使用して及び／又はそれに基づいて導出され得る特性の例は、トレーニングエンティティ114の最大積載量と運搬及び／又は収集されるべきオブジェクト（複数を含む）の質量との間の差又は比、及び／又は、トレーニングエンティティ114の最大容積と運搬及び／又は収集されるオブジェクト（複数を含む）の体積との間の差又は比を含み得る。

【0145】

アセンブリタスクの場合、トレーニングエンティティ情報及び／又はトレーニング環境情報を使用して及び／又はそれらに基づいて導出され得る特性の例は、最小構成要素の高さ、幅、及び／又は長さ、これらの最小構成要素を扱うためのトレーニングエンティティ 114 の 1 つ又は複数の操作要素、例えばロボットハンド又はペンチの高さ、幅、及び／又は長さとの間の差又は比、及び／又は、構成要素の現在の相対位置と構成要素のターゲット相対位置との間の差を含み得る。

【0146】

タスク容易性推定部 124 はまた、タスク容易性推定関数 124-2 を含む。タスク容易性推定関数 124-2 は、1 つ又は複数のタスク特性に基づいてタスクの容易性を推定する。タスク容易性推定関数 124-2 は、1 つ又は複数のトレーニング可能なパラメータを有する関数であり得る。例えば、トレーニング容易性推定関数 124-2 は、ニューラルネットワークの重みなどの複数の更新可能なパラメータを有するニューラルネットワークであり得、及び／又はそれを含み得る。しかしながら、タスク容易性推定関数 124-2 は、更新可能なパラメータを有する 1 つ又は複数の他のタイプの関数、例えば、1 つ又は複数のタスク特性のパラメータ化された線形関数若しくはそれに基づくパラメータ化された線形関数、又は 1 つ又は複数のタスク特性のパラメータ化されたロジスティック関数若しくはそれに基づくパラメータ化されたロジスティック関数、又はサポートベクターマシンであり得、及び／又はそれを含み得る。

【0147】

いくつかの実施形態では、タスク容易性推定関数 124-2 は、フィードフォワードニューラルネットワークであり、及び／又はそれを含む。フィードフォワードニューラルネットワークは、全結合ニューラルネットワークであり得る。フィードフォワードニューラルネットワークは、同じであっても異なってもよい 1 つ又は複数の隠れ層サイズ、例えば各層内のノードの数、の 1 つ又は複数の隠れ層を有し得る。ある実装形態では、フィードフォワードニューラルネットワークは 2 つの隠れ層を有し、その各々のサイズは 64 である。隠れ層の数及び／又は各隠れ層のサイズは、フィードフォワードニューラルネットワークをトレーニングするときに損失を最小にするためにグリッド検索を使用して選択され得る。フィードフォワードニューラルネットワークの損失及びトレーニングの更なる詳細は、タスク容易性推定関数に関連して説明される。

【0148】

タスク取得部 120 は、タスク容易性範囲決定部 126 を含み得る。タスク容易性範囲決定部 126 は、タスクジェネレータ 122 によって生成された複数のタスク（例えば、ランダムなタスク）の容易性に基づいて中間容易性範囲を決定し得る。中間容易性範囲は、タスクが中間容易性であり、同等に中間困難性であると見なされる範囲を定義し得る。従って、中間容易性範囲は、中間困難性範囲とも称され得る。中間容易性範囲決定部は、複数のタスクの容易性にわたって算出される 1 つ又は複数の要約統計量、例えば平均及び／又は標準偏差に基づいて決定され得る。中間容易性範囲を決定するための例となる方法は、図 3 A に関して説明される。

【0149】

タスク容易性範囲決定部 126 は、追加的又は代替的に、閾値容易性を決定し得る。閾値容易性は、それを上回るとタスクが容易であると見なされる容易性を定義し得る。閾値容易性は、1 つ又は複数の要約統計量、例えば平均及び／又は標準偏差に基づいて決定され得る。中間容易性範囲も決定されている場合には、その中間容易性範囲に基づいて容易性閾値が決定され得る。例えば、タスク容易性閾値は、中間容易性範囲の上限であり得る。閾値容易性を決定するための例となる方法は、図 3 B に関して説明される。

【0150】

タスク取得部 120 は、タスクセクタ 128 を含み得る。タスクセクタ 128 は、タスクジェネレータ 122 によって生成された複数の生成タスクから 1 つ又は複数のタスクを選択する。タスクの選択は、タスクのスコア、例えば、タスク容易性推定部 124 によって推定された複数のタスクの各々の推定された容易性に基づき得る。タスクセクタ

10

20

30

40

50

128は、スコアが、1つの選択基準、1つ又は複数の選択基準、又は1つ又は複数の選択基準のうちの少なくとも1つを満たすことに応答して、タスクを選択し得る。タスクセレクト128は、1つ又は複数のタスクを選択するために、タスク容易性範囲決定部126から中間容易性範囲及び／又は閾値容易性を受け取り得る。

【0151】

選択された1つ又は複数のタスクは、1つ又は複数の中間容易性タスクを含み得る。例えば、成功基準は、1つ又は複数のタスクが中間容易性であることをスコアが示す場合に満たされ得、1つ又は複数の中間容易性タスクは、タスク容易性範囲決定部126によって決定された中間容易性範囲内の、及び／又は静的な中間容易性範囲内のスコア、例えば容易性を有する1つ又は複数のタスクであり得、例えば容易性がタスク成功確率である場合、静的な中間容易性は0.4～0.6の範囲であり得る。中間容易性タスクは、中間困難性の又はフロンティアタスクのゴールとも称され得、少なくとも、エージェントにその現在の能力で新しいタスクを解決するように教えるため中間容易性タスクがランダムなタスクよりもポリシー改善のためのより多くの学習信号を供給することから、中間容易性タスクがトレーニングエージェント130のより良好な及び／又はより効率的なトレーニングを容易にするため、トレーニングエージェントのトレーニングでの使用に有益である。1つ又は複数の中間容易性タスクを選択するための例となる方法は、図3Aに関して説明される。

【0152】

選択された1つ又は複数のタスクは、1つ又は複数の容易なタスクを含み得る。例えば、1つ又は複数の容易なタスクは、タスク容易性範囲決定部126によって決定された閾値容易性を上回る及び／又は静的閾値容易性を上回る容易性を有する1つ又は複数のタスクであり得、例えば、容易性がタスク成功確率である場合、静的閾値容易性は0.6であり得る。容易なタスクは、少なくとも、容易なタスクがトレーニング中のトレーニングエージェント130による破滅的忘却を軽減し得るため、トレーニングエージェント130のトレーニングでの使用に有益である。例えば、トレーニング中、エージェントは、より困難なタスクについてしか後にトレーニングされないと、より容易なタスクを実行するのを忘れ得る。1つ又は複数の容易なタスクを選択するための例となる方法は、図3Bに関して説明される。

【0153】

選択された1つ又は複数のタスクは、1つ又は複数のランダムなタスク、例えば複数のタスクのうちの1つ又は複数のランダムなタスクを含み得る。ランダムなタスクは、過剰適合を防ぎ、及び／又は、トレーニングプロセスにエントロピーを注入することによって、トレーニングエージェントの一般化を可能にし得るため、トレーニング中、有益である。1つ又は複数のランダムなタスクを選択するための例となる方法は、図3Cに関して説明される。

【0154】

選択された1つ又は複数のタスクは、1つのタイプのタスク、例えば中間容易性タスクを含み得るか、又は指定されたタスクタイプの任意の組合せを含み得る。選択された1つ又は複数のタスクは、指定されたタスクの各々の1つ又は複数のタスクを含み得る。これは、指定されたタスクタイプの各々の利益、例えば、より良好な及び／又はより効率的なトレーニング、破滅的忘却の緩和、並びに過剰適合の低減及び／又はより良好な一般化がトレーニングにおいて実現されることを可能にし得るため、有益であり得る。選択された1つ又は複数のタスクにおける各タスクタイプの割合は、トレーニングエージェント130のトレーニング全体を通して変化し得る。トレーニングの初め及び早い段階では、容易なタスク及び中間容易性タスクが、選択された1つ又は複数のタスクの大部分を形成し得る。トレーニングの後ろの段階では、トレーニングエージェント130が一般化することを学習するため、ランダムなタスクがタスクの大部分を形成し得る。

【0155】

トレーニングエージェント130は、トレーニング環境112の観察及び／又はトレー

10

20

30

40

50

ニングエンティティ 114 の状態を受け取り、トレーニング環境 112 においてトレーニングエンティティ 114 によって行動を実行させることによって、1 つ又は複数のタスク、例えば選択された 1 つ又は複数のタスクの実行を試行することができるエージェントである。トレーニングエージェント 130 は、強化学習エージェント、例えば深層強化学習エージェントであり得る。トレーニングエージェントは、観察に応答してエージェントが取る行動を定義するポリシー π に従って行動を実行し得る。タスクの実行を試行することによって、トレーニングエージェントは、タスクをより良好に実行するようにトレーニングされる。トレーニングエージェント 130 の実施形態及びそれがどのようにトレーニングされ得るかの例は、後の図に関して説明される。

【0156】

トレーニングエージェント 130 が実行するように試みる各タスクについて、トレーニングエージェント 130 は、タスク容易性推定関数トレーナ 140 に成功基準満足指示を提供する。成功基準満足指示は、タスクが成功裏に実行されたか否かを示すバイナリインジケータであり得る。タスクの成功裏な完了は、1 として表され得る。タスクの失敗は、0 として表され得る。成功基準満足指示は、タスク固有の失敗基準又は一般的な失敗基準が満たされたと決定されたためにタスクの試行が終了する前に、タスクを実行しようとする試行において 1 つ又は複数の成功基準又はそれらのうちの少なくとも 1 つが満たされたかどうかを示し得る。成功基準及びタスク固有の失敗基準の例は、タスクジェネレータ 122 に関連して上で説明されている。一般的な失敗基準は、タスク固有の失敗基準について説明したものと同一タイプのもののいずれかであり得る。一般的な失敗基準の例は、タイムステップ限界、例えばタイムステップの最大数に達したときに満たされる失敗基準であり、ここで、タイムステップは、このインスタンスで行動が実行させられない場合であっても、行動が実行させられ得るインスタンスである。

【0157】

タスク容易性推定関数トレーナ 140 は、タスクの容易性をより正確に推定することができるようにタスク容易性推定関数 124-2 をトレーニングする。タスク容易性推定関数トレーナ 140 は、損失関数算出部 142 と、パラメータ更新部 144 とを有する。タスク容易性推定関数トレーナ 140 は、タスク容易性推定関数 124-2 が推定するように、トレーニングエージェント 130 と共にトレーニングされ得る。

【0158】

損失関数算出部 142 は、タスクの実際の容易性と、容易性推定関数によって推定されたタスクの期待容易性との差に基づいて損失を算出する。タスクの実際の容易性は、成功基準満足指示に基づいて決定され得る。成功裏に完了したタスクは、高い容易性を有すると決定され得、成功裏に完了しなかったタスクは、低い容易性を有すると見なされ得る。容易性が成功確率として定義される場合、成功が 1 として表され、失敗が 0 として表される成功基準満足指示は、成功裏に完了したタスクは成功確率が 1 であり、失敗したタスクは成功確率が 0 であるため、それ自体がタスクの実際の容易性であり得る。損失関数算出部は、任意の適切な損失関数、例えば交差エントロピー損失、ヒンジ損失、又はカルバック・ライブラー情報量 (Kullback-Leibler divergence) を使用して損失を算出し得る。

【0159】

パラメータ更新部 144 は、損失を最小にするか又は少なくとも低減するようにタスク容易性推定関数 124-2 のパラメータを更新し得る。タスク容易性推定関数 124-2 がニューラルネットワークであるか又はそれを含む場合、ニューラルネットワークのパラメータ、例えば重みは、逆伝搬を使用して更新され得る。パラメータ更新の量は、学習率に基づき得る。学習率は、グリッド検索を使用して損失を最小にするように選択され得る。ある実装形態では、学習率は 0.1 である。

【0160】

エージェント及びタスク容易性推定関数のトレーニング方法

図 2 は、エージェントをトレーニングするため及びタスク容易性推定関数をトレーニングするための例となる方法を例示する。エージェントは、図 5B に関して説明するトレー

10

20

30

40

50

ニングエージェント 502B 及び／又は図 1 に関して説明したトレーニングエージェント 130 であり得る。例となる方法は、1 つ又は複数のコンピューティングデバイス、例えば図 7 に関連して説明するコンピューティングデバイス 700 によって実行される 1 つ又は複数のコンピュータ実行可能命令として実施され得る。例となる方法は、システム 100 又はその一部を使用して実行され得る。オプションのステップは破線で示されている。

【0161】

ステップ 210 において、1 つ又は複数のタスクが取得される。1 つ又は複数のタスクの各タスクは、それぞれ初期状態及び 1 つ又は複数の成功基準を定義する。タスク及びどのように 1 つ又は複数のタスクの例は、図 1 のタスク取得部 120 に関して説明される。1 つ又は複数のタスクがどのように取得され得るかの例は、それぞれ、図 2A～図 2C の方法 210-A、210-B、及び 210-C に関して説明される。

10

【0162】

ステップ 220 において、エージェントは、トレーニングコンテキストにおいてトレーニングされる。例えば、エージェントは、強化学習方法を使用してトレーニングされ得る。強化学習方法は、エージェントによって選択された行動によって得られる累積報酬を最大化する目的でエージェントがトレーニングされる方法として定義され得、例えば強化学習方法は、受け取った報酬に基づいて、特定の行動が強化される、例えば所与の状態においてより頻繁に実行されるように選択される方法であり得る。エージェントは、図 5B に関連して説明するエージェントトレーナ 530 によってトレーニングされ得、エージェントがどのようにトレーニングされ得るかの例は、これに関連して含まれる。例えば、エージェントは、近接ポリシー最適化（PPO：proximal policy optimization）の変形例、例えばクリッピングを伴う PPO、を使用してトレーニングされ得る。トレーニングコンテキストの例は、図 1 のトレーニングコンテキスト 110 に関して説明される。

20

【0163】

ステップ 220 は、1 つ又は複数のタスクの各々について実行される初期化ステップ（複数を含む）222 及びタスク実行試行ステップ（複数を含む）224 を含む得る。ステップ 220 はまた、パラメータエージェント更新ステップ 222 を含む得る。

【0164】

初期化ステップ（複数を含む）222 において、トレーニングコンテキストは、タスクによって定義された初期状態に初期化される。タスクによって定義され得る初期状態の例は、タスクジェネレータ 122 に関して説明される。

30

【0165】

タスク実行試行ステップ（複数を含む）224 において、タスクの実行がエージェントによって試行される。タスク実行試行ステップ 224 の例となる実施形態は、図 4 に関して説明される。タスクの実行がどのように試行され得るかの例はまた、図 1 のトレーニングエージェント 130 に関連して説明される。

【0166】

ステップ 226 において、エージェントのパラメータは、エージェントによる各タスクの実行の試行に基づいて更新される。エージェントのパラメータは、任意の適切な方法を使用して更新され得る。例えば、エージェントが、1 つ又は複数のニューラルネットワークを含む深層強化学習エージェントである場合、ニューラルネットワークのうちの 1 つ又は複数のうちの少なくとも 1 つのニューラルネットワークのパラメータは、逆伝搬を使用して更新され得る。パラメータの更新は、報酬関数を使用して算出される報酬及び／又はアドバンテージ関数を使用して算出されたアドバンテージに基づき得る。パラメータがどのように更新され得るかの例は、図 5B のエージェントトレーナ 530 に関して説明される。

40

【0167】

ステップ 230 において、1 つ又は複数のタスクの各タスクについて 1 つ又は複数のタスク特性が決定される。1 つ又は複数のタスク特性及びそれらがどのように決定され得るかの例は、図 1 のタスク特性決定部 124-1 に関して説明される。

50

【0168】

ステップ240において、タスク容易性推定関数がトレーニングされる。タスク容易性推定関数をトレーニングすることは、1つ又は複数のタスクの各タスクについて、タスクの1つ又は複数の特性と、タスクの実行を試行した結果、1つ又は複数の成功基準のうちの少なくとも1つが満たされたかどうかを示す指示とを備えるトレーニングセットに基づき得る。タスク容易性推定関数をトレーニングすることは、損失関数を使用して、トレーニングセットに基づいて1つ又は複数の損失を算出することと、1つ又は複数の損失に基づいてタスク容易性推定関数のパラメータを更新することを含み得る。1つ又は複数の損失がどのように算出され得るかの例は、図1の損失関数算出部142に関して説明される。パラメータがどのように更新され得るかの例は、図1のパラメータ更新部144に関連して説明される。

10

【0169】

タスク取得方法

図3A～図3Cは、1つ又は複数のタスクを取得する際に使用される例となる方法を説明する。

【0170】

方法210Aは、中間容易性のタスクを選択するための方法である。オプションのステップは破線で示されている。

【0171】

ステップ310において、中間容易性範囲が決定される。中間容易性範囲は、中間困難性範囲とも称され得る。中間容易性範囲の決定は、生成ステップ312と、推定ステップ314と、算出ステップ314とを含み得る。いくつかの実施形態では、中間容易性範囲が静的であり得ること、例えば容易性が成功確率である場合には0.4～0.6であり得、そうでない場合には説明したものとは異なる方法を使用して決定されることに留意されたい。

20

【0172】

ステップ312において、複数のタスクが生成される。複数のタスクは、ランダムな又は擬似ランダムなタスクであり得る。複数のタスクがどのように生成され得るかの例は、図1のタスクジェネレータ122及びタスク容易性範囲決定部126に関連して説明される。

30

【0173】

ステップ314において、タスク容易性推定関数、例えば機械学習モデルを使用して、複数のタスクについての対応する複数の推定された容易性が推定される。複数のタスクの容易性を推定することは、複数のタスクをスコアリングすることと見なされ得るか、又はその一例であり得る。容易性推定関数は、 f_{π} と称され得、従って、所与のタスク h についての推定された容易性は、 $f_{\pi}(h)$ と称され得る。推定された容易性がタスク成功確率である場合、タスクの困難性は、推定された容易性の補数 $1 - f_{\pi}(h)$ と定義され得る。タスク容易性がどのように推定され得るかの例は、図1のタスク容易性推定部に関連して説明される。タスク容易性推定関数の例は、図1のタスク容易性推定関数124-2に関連して説明される。

40

【0174】

ステップ316において、対応する複数の推定された容易性に基づいて中間容易性範囲が算出される。中間容易性範囲を決定することは、範囲、例えばスコアの範囲を決定することであるか、又はその一例であり得る。中間容易性範囲は、対応する複数の変数の要約統計量に基づいて算出され得る。例えば、中間容易性範囲は、推定された容易性の平均 μ_f と、推定された容易性の標準偏差 σ_f とに基づいて決定され得る。中間容易性範囲は、 $\mu_f - \gamma \sigma_f$ から $\mu_f + \gamma \sigma_f$ であり得、ここで、 γ は、ハイパーパラメータである。中間容易性範囲がどのように決定され得るかの更なる詳細は、図1のタスク容易性範囲決定部126に関連して説明される。

【0175】

50

ステップ 3 2 2 において、タスクが生成される。生成されるタスクは、ランダムなタスクであり得る。タスクがどのように生成され得るかの例は、図 1 のタスクジェネレータ 1 2 2 に関連して説明される。

【0176】

ステップ 3 2 4 において、タスク容易性推定関数を使用して、生成されたタスクの容易性が推定される。タスクの容易性は、ステップ 3 1 4 に関連して説明した方法のいずれかで推定され得る。

【0177】

ステップ 3 2 6 において、タスクの推定された容易性が中間容易性範囲内であると決定される。推定された容易性が中間容易性範囲内であることは、選択基準の一例である。いくつかの実施形態では、中間容易性範囲は排他的範囲であり得、他の実施形態では、中間容易性範囲は包含的範囲であり得る。先に示した中間容易性範囲の例では、範囲が排他的範囲である場合には、推定された容易性が中間容易性範囲内であると決定することは、次のように表され得、

$$\mu_f - \gamma \sigma_f < f_{\pi}(h) < \mu_f + \gamma \sigma_f$$

範囲が包含的範囲である場合には、推定された容易性が中間容易性範囲内であると決定することは、次のように表され得る：

$$\mu_f - \gamma \sigma_f \leq f_{\pi}(h) \leq \mu_f + \gamma \sigma_f$$

更なる詳細は、図 1 のタスクセレクト 1 2 8 に関連して説明される。

【0178】

ステップ 3 2 8 において、生成されたタスクは、1 つ又は複数のタスクのうちのタスクとして選択される。更なる詳細は、図 1 のタスクセレクト 1 2 8 に関連して説明される。

【0179】

方法 2 1 0 B は、容易なタスクを選択するための方法である。オプションのステップは破線で示されている。

【0180】

ステップ 3 3 0 において、閾値容易性が決定される。閾値容易性の決定は、閾値スコアを決定することであり得るか、又はその一例であり得る。閾値容易性の決定は、生成ステップ 3 3 2 と、推定ステップ 3 3 4 と、算出ステップ 3 3 4 とを含み得る。いくつかの実施形態では、閾値容易性が静的であり得ること、例えば容易性が成功確率である場合には 0.8 であり得、そうでない場合には説明したものとは異なる方法を使用して決定され得ることに留意されたい。

【0181】

ステップ 3 3 2 において、複数のタスクが生成される。複数のタスクは、ランダムな又は擬似ランダムなタスクであり得る。複数のタスクがどのように生成され得るかの例は、図 1 のタスクジェネレータ 1 2 2 及びタスク容易性範囲決定部 1 2 6 に関連して説明される。

【0182】

ステップ 3 3 4 において、タスク容易性推定関数を使用して、複数のタスクについての対応する複数の推定された容易性が推定される。複数のタスクについての対応する複数の推定された容易性は、ステップ 3 1 4 に関連して説明した方法のいずれかで推定され得る。

【0183】

ステップ 3 3 6 において、閾値容易性が、対応する複数の推定された容易性に基づいて算出される。閾値容易性は、対応する複数の推定された容易性の要約統計量に基づいて算出され得る。例えば、閾値容易性は、推定された容易性の平均 μ_f と、推定された容易性の標準偏差 σ_f とに基づいて決定され得る。閾値容易性は、 $\mu_f + \beta \sigma_f$ であり得、ここで、 β は、ハイパーパラメータである。閾値容易性がどのように決定され得るかの更なる詳細は、図 1 のタスク容易性範囲決定部 1 2 6 に関連して説明される。

【0184】

ステップ 3 4 2 において、タスクが生成される。生成されるタスクは、ランダムなタスク

10

20

30

40

50

クであり得る。タスクがどのように生成され得るかの例は、図1のタスクジェネレータ122に関連して説明される。

【0185】

ステップ344において、タスク容易性推定関数を使用して、生成されたタスクの容易性が推定される。タスクの容易性は、ステップ314に関連して説明した方法のいずれかで推定され得る。

【0186】

ステップ346において、タスクの推定された容易性が閾値容易性より大きいと決定される。推定された容易性が閾値容易性よりも大きいことは、選択基準の一例である。先に示した閾値容易性の例では、推定された容易性が閾値容易性より大きいと決定することは、 $f_{\pi}(h) > \mu_f + \beta \sigma_f$ と表すことができる。更なる詳細は、図1のタスクセレクトタ128に関連して説明される。

10

【0187】

ステップ348において、生成されたタスクは、1つ又は複数のタスクのうちのタスクとして選択される。更なる詳細は、図1のタスクセレクトタ128に関連して説明される。

【0188】

方法210Cは、任意のタスク、例えばランダムなタスクを選択するための方法である。

【0189】

ステップ352において、タスクが生成される。生成されるタスクは、ランダムなタスクであり得る。タスクがどのように生成され得るかの例は、図1のタスクジェネレータ122に関連して説明される。

20

【0190】

ステップ354において、このタスクが、1つ又は複数のタスクのうちのタスクとして選択される。例えば、生成されたタスクがランダムなタスクである場合、ランダムなタスクが1つ又は複数のタスクのうちのタスクとして選択される。更なる詳細は、図1のタスクセレクトタ128に関連して説明される。

【0191】

1つ又は複数のタスクを取得するステップ210を実施するためにこれらの方法がどのように組み合わせられ得るかの更なる詳細は、図1のタスク取得部120及びその構成要素に関して説明される。

30

【0192】

1つ又は複数のタスクを取得することが、中間容易性範囲を決定するステップ310及び／又は閾値容易性を決定するステップ330を含む実施形態は、「適応フィルタリング」を用いるものと見なされる。

【0193】

1つ又は複数のタスクが取得され得る例となる実装形態を示すアルゴリズムの一例が以下に含まれる。

40

50

【表 1】

Algorithm 1: Training loop with f_{π} update

```

input :  $\emptyset$ 
output :  $\pi$ 
 $\pi, f_{\pi}, \mu_f, \sigma_f \leftarrow \text{Init}();$ 
for  $i \leftarrow 0$  to  $\text{numEpochs}$  do
     $\text{tasks}, \text{successes}, \text{states}, \text{actions}, \text{rewards} \leftarrow$ 
         $\text{Rollouts}(\pi, f_{\pi}, \mu_f, \sigma_f);$ 
     $\pi \leftarrow \text{PPO}(\pi, \text{states}, \text{actions}, \text{rewards});$ 
     $f_{\pi} \leftarrow \text{Train}(f_{\pi}, \text{tasks}, \text{successes});$ 
     $\text{randomTasks} \leftarrow \text{GetRandomTasks}();$ 
     $\mu_f, \sigma_f \leftarrow \text{FitNormal}(f_{\pi}, \text{randomTasks});$ 
return  $\pi;$ 

```

Algorithm 2: Rollouts

```

input :  $\pi, f_{\pi}; \mu_f; \sigma_f;$ 
output :  $\text{rollouts}$ 
for  $i \leftarrow 0$  to  $\text{batchSize}$  do
     $\text{task} \leftarrow \text{GetDynamicTask}(f_{\pi}, \mu_f, \sigma_f);$ 
     $\text{rollouts}[i] \leftarrow \text{RunEpisode}(\pi, \text{task});$ 
return  $\text{rollouts};$ 

```

Algorithm 3: GetDynamicTask

```

input : Timestep  $t; f_{\pi}; \mu_f; \sigma_f;$ 
    Hyperparameters  $\beta, \gamma$ 
output : Task  $h$ 
 $\text{taskType} \leftarrow \text{GetTaskType}(t);$ 
while true do
     $h \leftarrow \text{RandomTask}();$ 
    switch  $\text{taskType}$  do
        case easy do
            if  $f_{\pi}(h) > \mu_f + \beta \sigma_f$  then
                return  $h;$ 
        case frontier do
            if  $\mu_f - \gamma \sigma_f < f_{\pi}(h) <$ 
                 $\mu_f + \gamma \sigma_f$  then
                return  $h;$ 
        case random do
            return  $h;$ 

```

【 0 1 9 4 】

タスク実行方法

図 4 は、例となる実施形態による、エージェントによってタスクの実行を試行するための方法 2 2 4 のフロー図である。

【 0 1 9 5 】

ステップ 4 1 0 において、1 つ又は複数の行動ステップが実行される。1 つ又は複数の行動ステップの各行動ステップについて、観察受取ステップ 4 1 2、行動決定ステップ 4 1 4、行動発生ステップ 4 1 6、及び報酬算出ステップ 4 1 8 がある。行動ステップは、成功基準又は失敗基準が満たされるまで実行される。

【 0 1 9 6 】

ステップ 4 1 2 において、トレーニング環境の 1 つ又は複数の観察を受け取る。トレーニング環境が物理的環境である場合、トレーニング環境の 1 つ又は複数の観察は、トレーニング環境の、キャプチャされ得る任意の適切な観察であり得る。例えば、観察は、1 つ又は複数の画像のうちの 1 つ又は複数であり得る。1 つ又は複数の画像は、パッシブ画像及び／又はアクティブ画像であり得る。画像は、2 D 画像、及び／又は深度マップを含む 3 D 画像であり得る。パッシブ画像とは、特別な介入なしに、例えばエンティティ又はその構成要素からの出力なしに、環境に存在する波に基づく画像であり得る。例えば、パッシブ画像は、特別な介入なしに環境に存在する電磁波、例えば可視光、赤外線、及び／又は紫外線に基づく画像であり得る。これらの電磁波は、1 つ又は複数のカメラ、例えば、可視光（又は通常）カメラ、赤外線カメラ、又は紫外線カメラによってキャプチャされ得る。アクティブ画像とは、撮像及びその反射の目的で発せられる波、例えばエンティティ又はその構成要素によって出力された波に基づく画像であり得る。例えば、アクティブ画像は、エンティティの構成要素によって出力される電磁波及びその反射、例えば、レーダデバイスを使用して発せられる電波、ライダーデバイスを使用して発せられるレーザ光、X 線デバイスによって発せられる X 線、又はマイクロ波撮像デバイスを使用して発せられるマイクロ波に基づき得る。アクティブ画像は、エンティティによって出力される音波、例えば超音波撮像デバイスによって発せられる超音波に基づき得る。観察はまた、1 つ又は複数のセンサ観察は、例えば、トレーニングエンティティ上の 1 つ又は複数の熱センサからの 1 つ又は複数の熱センサ観察、又はトレーニングエンティティ上の 1 つ又は複数のマイクロフォンによってキャプチャされた 1 つ又は複数のサウンド観察であり得るか、又はそれらを含み得る。トレーニング環境がシミュレート環境である場合、観察は、上記のいずれかのシミュレーション、例えばシミュレート環境においてトレーニングエンティティの 1 つ又は複数のシミュレートされたカメラによってキャプチャされるであろう画像に

10

20

30

40

50

対応する環境の1つ又は複数のレンダリングであり得る。環境のレンダリングは、実質的にフォトリアリスティックであり得る。実質的にフォトリアリスティックなレンダリングを使用することで、シミュレート環境におけるエージェントのトレーニングが容易になり得る。より一般的には、実環境においてエンティティによってキャプチャされるであろう実際の観察の実質的に現実的なシミュレーションを使用することで、物理的環境において利用され得るエンティティをシミュレート環境においてトレーニングすることが容易になり得る。

【0197】

ある例では、観察は、シミュレート環境においてシミュレートエンティティの単一のカメラによってキャプチャされたであろう1つ又は複数の画像に対応する1つ又は複数のレンダリングであり得る。シミュレート環境においてシミュレートエンティティによってキャプチャされるであろう1つ又は複数の画像に対応するレンダリングを使用することによって、物理的環境で展開されたときに、物理的環境において利用されるために複数のカメラ及び／又はマルチモーダル入力を必要としないであろうエンティティのトレーニングが容易になり得る。

10

【0198】

トレーニング環境がビデオゲーム環境である場合、観察は、ビデオゲームのプレーヤが見るであろうもの、例えばビデオゲームの画面及び／又は出力、又は、同じくプレーヤに提供される環境状態に対応し得る。これは、エージェントが実行される行動をプレーヤと同じ情報に基づかせているため、より人間らしい挙動を提供するエージェントのトレーニングを容易にし得る。トレーニングエンティティの状態の1つ又は複数の指標

20

【0199】

ステップ414において、トレーニング環境においてトレーニングエンティティが実行可能な複数の行動のうちの1つ又は複数の行動が決定される。

【0200】

トレーニング環境においてトレーニングエンティティが実行可能な複数の行動は、1つ又は複数の移動行動、例えば、前方への移動、後方への移動、左への方向転換、及び／又は右への方向転換を含み得る。ある実装形態では、エージェントの行動は、プリミティブの外側に明示的なループクロージャがない、離散的な、位置ベースのモーションプリミティブ（すなわち、前方への移動、左への方向転換、又は右への方向転換）であり得る。トレーニング環境においてトレーニングエンティティが実行可能な複数の行動は、1つ又は複数の操作行動、例えばロボットハンド又はペンチなどの1つ又は複数の操作構成要素の移動又は作動を備え得る。実行可能な複数の行動は、トレーニング環境においてトレーニングエンティティが実行可能な1つ又は複数の器具行動、例えば掃除機構構成要素を作動させることを含み得る。

30

【0201】

複数の行動のうちの1つ又は複数の機能、例えばエージェントの1つ又は複数のニューラルネットワークを使用して、複数の行動の各々が最適な行動である確率を示す行動確率を算出することによって決定され得、ここで、これらの機能のうちの少なくとも一部、例えばエージェントの1つ又は複数のニューラルネットワークのうちの少なくとも1つはトレーニング可能である。1つ又は複数の行動は、最も高い報酬を有する確率が最も高い1つ又は複数の行動として決定、例えば選択され得る。しかしながら、1つ又は複数の行動を決定するとき、少なくともトレーニング中、1つ又は複数の行動の決定にはある程度のエントロピーが存在し得る。これは、探索を促進し得る。例えば、1つ又は複数の行動は、複数の実行可能な行動のうちの1つ又は複数の行動を、行動が最適である確率に基づく確率分布に従ってランダムに選択することによって決定され得る。探索を促進するために、最も高い報酬を有する確率が低いと決定されている行動を実行する確率を上げ得る。

40

【0202】

ステップ416において、1つ又は複数の行動が、トレーニング環境においてトレーニ

50

ングエンティティによって実行させられる。次いで、エージェントによって実行されるべき行動を示す指示を受け取る制御インターフェースを使用して実行されるべき1つ又は複数の行動は、示された行動を実行するようにエージェントを制御する。制御インターフェースを使用することで、エージェントによって示される行動をそれらの制御実行から抽象化することができるため、同一又は同様のエージェントが、シミュレート環境及び実環境においてトレーニング及び／又は使用されることを容易にし得る。

【0203】

ステップ418において、報酬が、トレーニングコンテキストの更新状態に基づいて、例えばトレーニングエンティティの更新状態及び／又はトレーニング環境の更新状態に基づいて算出される。算出される報酬は、1つ又は複数の成功基準のうちの少なくとも1つが満たされることに応答して増やされ得る。算出される報酬は、トレーニングコンテキストの更新状態がトレーニングエンティティの以前の状態とは異なるトレーニングエンティティの更新状態、例えば以前の位置とは異なる更新された位置を含むことに応答して増やされ得、これは、以前の位置のバッファ、例えばトレーニングエンティティの以前の状態を保持することによって容易にされ得る。更新状態が以前の状態と異なることに応答して算出される報酬を増やすことで、探索が促進され、トレーニング中にトレーニングエージェントが極小に捕らえられることを防ぎ得る。トレーニングエンティティ又はその構成要素が移動行動を実行するナビゲーションタスクのようなタスクの場合、算出される報酬はまた、移動した距離に応答して増やされ得る。これもまた、探索を促進し、エージェントが留まること及び／又は1つ又は複数の成功基準のうちの少なくとも1つを達成する前にスピンすることを防ぎ得る。算出される報酬は、トレーニングエンティティがトレーニング環境において1つ又は複数のオブジェクトと衝突することに応答して減らされ得る。これは、トレーニングエージェントが衝突を回避するようにする。

【0204】

ナビゲーションタスクについての報酬が算出され得る報酬関数の例は、次の通りである：

【数1】

$$r(s) = \mathbb{1}_{\text{succ}} + \delta(-\mathbb{1}_{\text{coll}} + \mathbb{1}_{\text{expl}}) + 0.01d$$

ここで、sは更新状態である。バイナリインジケータ

【数2】

$$\mathbb{1}_{\text{succ}}$$

は、1つ又は複数の成功基準のうちの少なくとも1つを満たすと、例えばターゲット位置に到達すると真であり、そうでなければ偽である。

【数3】

$$\mathbb{1}_{\text{coll}}$$

は衝突時に真であり、そうでなければ偽である。ハイパーパラメータ $0 < \delta < 1$ は、ターゲット探索挙動と比較して、学習探索及び運動技能に対するエージェントの親和性を制御する。

【数4】

$$\mathbb{1}_{\text{expl}}$$

は、探索に対する固有の報酬であり、これは、タスクの実行の試行にわたって過去のエージェント位置のバッファを保持することによって決定され、エージェントの現在位置がすべての以前の位置からある程度離れている場合に報酬を提供し得る。探索に対する固有の報酬は、衝突からの負の報酬を回避するために、エージェントがその場でスピンすることになる極大に入ることを防ぎ得る。dは、現在のステップで移動した距離であり、エージェントが可能な限り広い探索エリアをカバーしようとするべきである以前を表す。

【0205】

少なくとも実施形態では、上記による方法がエージェントによるタスクの実行の試行に使用される場合、各タスクの実行の試行に基づいてエージェントのパラメータを更新すること、例えば、方法200のステップ226は、1つ又は複数のタスクの各タスクに対する1つ又は複数の行動ステップの各行動ステップについて、行動ステップの開始時のトレーニングエンティティの状態の表現と、選択された1つ又は複数の行動と、算出される報酬とを備えるトレーニングセットに基づいて、エージェントのパラメータを更新することを含み得る。

【0206】

プロダクション環境におけるエージェント

図5Aは、例となる実施形態による、制御インターフェース503Aを使用してプロダクションエンティティ504を制御するためにプロダクション環境501において使用されているプロダクションエージェント502Aを例示する概略図500Aである。

10

【0207】

プロダクション環境501は、トレーニングされるものとは対照的にトレーニングエージェント502Aが使用されている環境を指す。プロダクション環境501は、任意の適切な環境、例えば図1のトレーニング環境112に関連して説明した環境のいずれかであり得る。いくつかの実施形態では、プロダクション環境は、実環境、例えば図1のトレーニング環境112に関して説明した物理的環境のうちの1つである。ビデオゲーム環境のコンテキストでは、プロダクション環境は、プレイ中のビデオゲーム環境、例えば、トレーニングが行われているビデオゲーム環境ではなく、プレイされているビデオゲーム環境であり得る。

20

【0208】

プロダクション環境501の1つ又は複数の観察、例えば1つ又は複数のアクティブ又はパッシブ画像が、プロダクションエージェント502Aによって受け取られ得る。ある例では、観察は、プロダクションエンティティ504の単一のカメラによってキャプチャされた1つ又は複数の画像であり得る。適切にトレーニングされたプロダクションエージェント502Aは、単一のカメラ入力に基づいてプロダクションエンティティ504Aを制御することが可能であり得、指定されたタスクを実行するために複数のカメラ及び／又はマルチモーダル入力を必要としないであろう。受け取られ得る観察の他の例は、図4のステップ412に関して説明される。

30

【0209】

プロダクションエージェント502Aは、トレーニングされているものとは対照的に、使用されているエージェントを指す。プロダクションエージェント502Aは、説明するように、1つ又は複数の観察に基づいて1つ又は複数の行動を実行させるように構成されたエージェントである。

【0210】

プロダクションエージェント502Aは、1つ又は複数の特徴ネットワーク510及びプロダクションポリシーネットワーク520Aを含み得る。

【0211】

1つ又は複数の特徴ネットワーク510は、1つ又は複数の潜在的特徴を取得するために、観察を処理する。1つ又は複数の特徴ネットワーク510のパラメータは、トレーニング環境、例えばシミュレーションへの過剰適合を防ぐために、トレーニング中、凍結されていてもよい。取得された1つ又は複数の潜在的特徴は、空間的にコヒーレントな潜在的特徴であり得る。取得された1つ又は複数の潜在的特徴は、スタックされ、ポリシーネットワーク520に提供され得る。

40

【0212】

1つ又は複数の特徴ネットワーク510は、1つ又は複数の空間特徴ネットワーク512を含み得る。1つ又は複数の空間特徴ネットワーク512は、観察から1つ又は複数の空間特徴、例えば環境の空間特徴を取得するために、1つ又は複数の観察を処理し得る。1つ又は複数の空間特徴ネットワークは、1つ又は複数の畳み込みニューラルネットワー

50

ク層 5 1 2 - 1 を含み得る。

【0 2 1 3】

1つ又は複数の観察が画像を含む場合、1つ又は複数の空間特徴ネットワークは、少なくとも、物理的環境の画像のような実際の画像について事前トレーニングされた、例えばエージェント 5 0 2 のトレーニングの前にトレーニングされた1つ又は複数の空間オートエンコーダのエンコーダであり得るか又はそれを含み得る。他のタイプの観察のための空間特徴ネットワークの場合、1つ又は複数の空間特徴ネットワークは、少なくとも、所与のタイプの実際の観察について事前トレーニングされた1つ又は複数の空間オートエンコーダのエンコーダであり得るか又はそれを含み得る。1つ又は複数の実際の画像及び／又は他の実際の観察を使用して事前トレーニングされた空間特徴ネットワークを使用することで、シミュレート環境及び／又はシミュレートエンティティを使用してトレーニングされたエージェントが、実環境において現実のエンティティで機能することが容易になり得る。

10

【0 2 1 4】

1つ又は複数の特徴ネットワークは、1つ又は複数のセマンティックターゲット特徴ネットワーク 5 1 4 を含み得る。1つ又は複数のセマンティックターゲット特徴ネットワークは、1つ又は複数のセマンティックターゲット特徴を取得するために、1つ又は複数の観察を処理し得る。

【0 2 1 5】

セマンティックターゲット特徴は、セマンティックターゲット、例えば所与のオブジェクトクラスのオブジェクトが位置している1つ又は複数の観察のエリアを示し得る。オブジェクトクラスの例には、例えばボール、扇風機、花瓶、食器棚、又は机のような家庭用オブジェクトクラス、例えば浴室、寝室、書斎、専用オフィス、又はステーションナリーカップボードのような部屋タイプ用オブジェクトクラス、例えばコピー機、ビデオ会議機材、又はファイリングカップボードのようなオフィス用オブジェクトクラス、例えば所与のタイプの機械若しくは組立ライン又はそれ上の位置のような産業用オブジェクトクラス、及び／又は例えば所与のタイプの医薬品又は医療機器のような医療用オブジェクトクラスが含まれ得る。

20

【0 2 1 6】

1つ又は複数のセマンティックターゲット特徴ネットワーク 5 1 4 は、1つ又は複数の畳み込みニューラルネットワーク層 5 1 4 - 1 を含み得る。観察、例えば画像が与えられると、畳み込みニューラルネットワーク層 5 1 4 - 1 は、複数のオブジェクトクラスの各オブジェクトクラスについてバイナリマスク M を予測し得る。1つ又は複数の畳み込みニューラルネットワーク層 5 1 4 - 1 はまた、複数のオブジェクトの各オブジェクトクラスについての予測信頼性を与え得る。1つ又は複数の畳み込みニューラルネットワーク層 5 2 2 - 1 は、物理的環境の画像のような、コンテキスト内の日常的なオブジェクトの描写を含むデータセット、例えば C O C O データセットについて事前トレーニングされ得る。1つ又は複数の畳み込みニューラルネットワーク層 5 2 2 - 1 は、M a s k - R - C N N であり得るか、又はそれを含み得る。M a s k - R - C N N は、F P N ネットワークバックボーンを有し得る。

30

【0 2 1 7】

1つ又は複数のセマンティックターゲット特徴ネットワーク 5 1 4 は、1つ又は複数の後処理層 5 1 4 - 2 を含み得る。1つ又は複数の後処理層 5 1 4 - 2 は、1つ又は複数のオブジェクトクラスのうちのオブジェクトクラスのためのバイナリマスクを抽出するように構成され得、このオブジェクトクラスは、設定可能なラベル L に基づいている。オブジェクトクラスを指定するために設定可能なラベルを使用することで、ラベルを設定することによってランタイムに異なるオブジェクトクラスを探索することが容易になる。設定可能なラベルは、オブジェクトクラスのオブジェクトがトレーニング中に実行されたタスクに含まれなかったオブジェクトクラス、例えば所与のクラスのオブジェクトがそのようなオブジェクトに対するナビゲーションタスクのターゲットではなかったオブジェクトクラ

40

50

スに設定され得る。例えば、トレーニングタスクは、ボールオブジェクトへのナビゲーションタスクと共に実行され得、設定可能なラベルは、花瓶オブジェクトに設定され得る。

【0218】

ある実装形態では、1つ又は複数の後処理層514-2は、所与のラベル l のオブジェクトクラスのための抽出されたバイナリマスクと予測信頼性とのスカラー乗算を実行して、出力 O を得ることができる：

$$O(x, y) = P(M(x, y) \text{ label} = l)$$

O のピクセルは依然として、ターゲットオブジェクトに関する形状情報（例えば、ボールマスクは円形であるが、ボックスマスクは正方形である）を含んでいる。エージェント502のトレーニング中、ポリシーネットワーク520が特定のオブジェクト形状に過剰適合することを防ぐため、及び1つ又は複数のセマンティックターゲット特徴のサイズを低減するために、最大プーリング演算が O に適用され得る。最大プーリング演算を O に適用した結果が、1つ又は複数のセマンティックターゲット特徴であり得る。

【0219】

所与のオブジェクトクラスのオブジェクトがどこに位置するかを決定する代わりに又はそれに加えて、1つ又は複数のセマンティックターゲット特徴ネットワーク514は、別のタスク、例えばより高いレベルの抽象化で定義されたタスク及び／又はナビゲーションに加えて実行されるべきタスクを実行するための所望のターゲットに関連するロケーションを決定し得る。例えば、別のタスクは、エンティティのソーラーパネルを使用してエネルギーを受け取ることであり得、セマンティックターゲット特徴ネットワークは、受け取るべきエネルギーの量を最大に近づける画像のターゲットロケーションを識別し得る。別の例として、別のタスクは、犯罪現場において裁判で重要なオブジェクトを位置特定することであり得、セマンティックターゲット特徴ネットワークは、ターゲットオブジェクトの特定のタイプ又はクラスが指定されなくても、提供される情報の量を最大に近づけるターゲットオブジェクトを識別し得る。これらの例のセマンティックターゲット特徴ネットワークは、これらのセマンティックターゲットが位置している画像においてターゲットロケーションを識別するための1つ又は複数の畳み込みニューラルネットワーク層を含み得る、及び／又は、ランタイムに所望のターゲットの異なるバリエーションを指定するための1つ又は複数の後処理層を含み得る。いくつかの例では、これらの代替の又は追加の1つ又は複数のセマンティックターゲット特徴ネットワークの（事前）トレーニングは、報酬を伴う強化学習方法を使用してこれらのセマンティックターゲット特徴ネットワークをトレーニングすることを含み得る。この報酬は、前述したナビゲーションのための算出される報酬とは異なり得、別のタスクの実行のためにセマンティックターゲット特徴ネットワークをトレーニングするのに適している報酬であり得る。

【0220】

プロダクションポリシーネットワーク520Aは、1つ又は複数の行動値、例えば1つ又は複数の行動の各々が、取るべき最適行動である推定確率を決定するために、1つ又は複数の潜在的特徴を処理し得る。

【0221】

プロダクションポリシーネットワーク520Aは、1つ又は複数の潜在的特徴を圧縮特徴表現に圧縮するための特徴圧縮セクション522を含み得る。特徴圧縮セクション522は、1つ又は複数の畳み込みニューラルネットワーク層522-1を含み得る。1つ又は複数の畳み込みニューラルネットワーク層522-1は、1つ又は複数の潜在的特徴を圧縮するために使用され得、例えば、ここで、1つ又は複数の潜在的特徴は、空間的にコヒーレントな潜在的特徴である。1つ又は複数の畳み込みニューラルネットワーク層522-1は、 $1 \times (8n + 1) \times 16 \times 16$ 行列を入力として取る三層の畳み込みニューラルネットワークであり、ここで、 n は、空間特徴の数であり、ReLUアクティベーションを伴う3つの畳み込み層を使用してそれらを圧縮する。例えば、3つの畳み込み層は、以下の表に指定されるようなものであり得る。この例では、出力は、圧縮された特徴表現であり、長さ256のベクトルである。

【表 2】

層	フィルタ数	カーネルサイズ	ストライド
1	32	4	2
2	64	4	2
3	128	2	1

【0222】

プロダクションポリシーネットワーク520Aは、メモリセクション524を含み得る。メモリセクション524は、圧縮特徴表現を受け取って処理し、メモリセクション出力を取得する。メモリセクション524は、エージェントによって実行されるタスクをマルコフ型（Markovian）に保つために使用され得る状態情報を維持すること、例えば記憶することを容易にする。例えば、ナビゲーションタスクをマルコフ型に保つためには、状態は、エージェントがどこにいたか及び以前にターゲットを見ているかどうかに関する情報を含まなければならない。メモリセクションは、1つ又は複数のメモリニューラルネットワーク層524-1を含み得る。1つ又は複数のメモリニューラルネットワーク層524-1セルのセルは、部分観測マルコフ決定過程（POMDP）の状態を表し得る。1つ又は複数のメモリニューラルネットワーク層524は、1つ又は複数の長期短期記憶（LSTM）層、1つ又は複数の勾配回帰型ユニット（GRU）層、及び／又はそれらの任意の適切な同等物であり得るか、又はそれらを含み得る。ある実装形態では、1つ又は複数のメモリニューラルネットワーク層は、サイズが256の1つ又は複数の長期短期記憶層である。

【0223】

プロダクションポリシーネットワーク520Aは、出力層526Aを含み得る。出力層526Aは、全結合出力層であり得る。出力層526Aは、少なくとも1つ又は複数の行動値を作り出す。少なくとも1つ又は複数の行動値を作り出す出力層526A又はその一部は、アクターと称され得、所与の観察された状態s、例えば1つ又は複数の受け取った観察について、ポリシー $\pi(s)$ を決定すると見なされ得る。プロダクションエージェント502Aの出力層526Aは、1つ又は複数の行動値に加えて更なる値、例えば図5Bに関連して説明する推定された状態値を出力し得るが、それを行う必然性はない。これらの他の値は、行動の選択に必要ではない。場合によっては、出力層526Aは、行動値に関係のない出力層の部分を除くことによって、より大きい出力層、例えば、トレーニングの際に使用される出力層から導出され得る。

【0224】

プロダクションエージェント502Aは、行動セクタ528Aを含み得る。行動セクタ528Aは、最も高い行動値を有する1つ又は複数の行動を選択し得る。しかしながら、1つ又は複数の行動を選択するとき、図4のステップ416に関して説明したように、1つ又は複数の行動の決定にはある程度のエントロピーが存在し得る。プロダクションにおいて、導入されるエントロピーの量は、トレーニングの時と同じであり得るか、又はエージェントを使用するときには探索ではなく開発（exploitation）が焦点となるため、エージェントをトレーニングするときに導入されるものよりも少ないであろう。しかしながら、エントロピーは、プロダクションにおいて、エージェント、ひいてはプロダクションエンティティ504が準最適な状態から抜け出せない状況を改善する際に依然として有用である。選択された1つ又は複数の行動は、プロダクションエージェント502Aから制御インターフェース503Aに行動セクタによって出力される。

【0225】

制御インターフェース503Aは、選択された1つ又は複数の行動を受け取り、プロダクションエンティティ504によって、選択された1つ又は複数の行動を実行させる。制御インターフェース503Aは、選択された1つ又は複数の行動をそれらのパフォーマンスの実施から抽象化することによって、同一又は同様のエージェントがシミュレート環境及び実環境においてトレーニング及び／又は使用されることを容易にし得る。

【0226】

プロダクションエンティティ504は、エージェントをトレーニングするために使用されているエンティティ以外の使用中のエンティティを指す。プロダクションエンティティ504は、任意の適切なエンティティ、例えば図1のトレーニングエンティティ114に関連して説明したエンティティのいずれかであり得る。いくつかの実施形態では、プロダクション環境は、現実のエンティティ、例えば図1のトレーニングエンティティ114に関して説明した物理装置のうちの1つである。例えば、プロダクションエンティティは、装置の物理的環境の画像を受け取るように構成された装置であり得、装置は、装置の方向を制御するように構成されたステアリングデバイスを含み得る。

【0227】

トレーニング中のエージェント

図5Aは、エージェントトレーナ530によってトレーニングを受けているエージェント502B、本明細書ではトレーニングエージェント502Bと呼ばれる、を例示する概略図500Bである。

【0228】

概略図500Bは、トレーニングコンテキスト110と、タスク取得部120とを含み、これらは、トレーニングエージェント502Bのトレーニングで利用される。

【0229】

トレーニングエージェント502Bがトレーニング環境112から観察を受け取る。観察は、図4のステップ412に関連して説明するタイプのいずれかであり得る。

【0230】

トレーニングエージェント502Bは、プロダクションエージェント502Aと同じ構成要素を含み、概ね同様に動作する。しかしながら、プロダクションポリシーネットワーク520Aの代わりに、トレーニングポリシーネットワークのようなポリシーネットワーク502Bの変形例が使用され得、行動セクタ528Aの代わりに、行動セクタ528Bの変形例が使用され得る。

【0231】

トレーニングポリシーネットワーク520Bは、異なる出力層526Bを利用することが、プロダクションポリシーネットワークのものとは異なり得る。行動値を推定することに加えて、出力層526Bは、行動値に加えて、状態 s についての状態値 $V^{\pi}(s)$ を推定し得る。状態値を推定する出力層526Bの部分は、クリティックと称され得る。推定された状態値 $V^{\pi}(s)$ は、エージェント502Bをトレーニングする際にエージェントトレーナ530によって利用される。出力層526Bが、プロダクションエージェント502Aでも使用され、推定された状態値が破棄されることは留意されるべきである。推定された状態値及び行動値は、エージェントトレーナ530に提供される。

【0232】

行動セクタ528Bは、エントロピーを導入するという点で又は1つ又は複数の行動の選択のエントロピーを増加させ得るという点で、行動セクタ528Aとは異なり得る。エントロピーは、トレーニング中の探索を促進するために導入及び／又は増加され得る。行動は、図4のステップ416に関して説明した方法のいずれかを使用して選択され得る。

【0233】

制御インターフェース503Bは、選択された1つ又は複数の行動を受け取り、プロダクションエンティティ504によって、選択された1つ又は複数の行動を実行させる。制御インターフェース503Bは、選択された1つ又は複数の行動をそれらのパフォーマンスの実施から抽象化することによって、同一又は同様のエージェントがシミュレート環境及び実環境においてトレーニング及び／又は使用されることを容易にし得る。

【0234】

エージェントトレーナ530は、トレーニングポリシーネットワーク520Bのパラメータ、例えば重みを更新することによって、トレーニングエージェント502Bをトレー

10

20

30

40

50

ニングする。前述したように、特徴ネットワークのパラメータは凍結され得、エージェントのトレーニング中、更新されなくてもよい。エージェントのトレーニング中、特徴ネットワークのパラメータを凍結することは、エージェントがプロダクション環境においてより良好に機能するように、トレーニング環境への過剰適合を防ぐだけでなく、エージェントのトレーニング中の収束を加速し、例えば、所与のパフォーマンスレベルにエージェントをトレーニングするために使用される時間及び／又は計算リソースを低減する。上で説明したように、トレーニングに使用されるべきタスクが取得される方法もエージェントのトレーニング中の収束を加速し、同様の効果を有する。特徴ネットワークのパラメータを凍結することと、本方法によって得られたタスクを使用することとの両方による収束の加速は、組み合わせることで、他の方法では不可能であったハードウェア上で、合理的な時間枠内でのエージェントのトレーニングを容易にし得る。例えば、1日未満、1週間未満、2週間未満、1ヶ月未満、6ヶ月未満の時間枠内で、グラフィックスプロセッシングユニットを有するデスクトップコンピュータ上でのエージェントのトレーニングが達成され得る。

【0235】

エージェントトレーナ530は、アクター-クリティック方法に従ってエージェントをトレーニングし得、それによって、パラメータは、1つ又は複数の行動に対する報酬と、1つ又は複数の推定された状態値とに基づいて最適化される。報酬は、行動を実行した後のトレーニングコンテキストの状態に基づいて算出され得る。例えば、報酬は、図4のステップ418に関して説明した方法のいずれかに従って算出され得る。報酬はタスクによって定義された1つ又は複数の成功基準に基づいて算出され得るため、報酬を算出する際にタスクが利用され得る。推定された容易性値は、出力層526Bのクリティックを使用して決定される。上記に従ってポリシーネットワーク502Bをトレーニングするための例となる方法は、クリッピングを伴う近接ポリシー最適化（PPO）である。ある実装形態では、トレーニングエージェント502Bは、以下のハイパーパラメータを有するPPOを使用してトレーニングされる：

【表3】

ミニバッチの数	1
学習率	0.005
クリッピング範囲(ϵ)	0.10
ディスカウント係数(γ)	0.99
価値関数係数(c_1)	0.5
エントロピー係数(β 又は c_2)	0.01
ポリシー更新あたりのタイムステップ	4000
ロールアウトワーカー	12
インナーループエポック	4
GAE λ	0.95

トレーニングが完了すると、トレーニングエージェント502Bは、プロダクションエージェント502Aとして使用され得る。

【0236】

エンティティ制御方法

図6は、エージェントを使用してプロダクション環境においてエンティティを制御するための例となる方法を例示する。エージェントは、図5Aに関して説明したプロダクションエージェント502Aであり得る。例えば、エンティティは、装置が物理的環境においてナビゲーションタスクを実行することができるようにエージェントによって制御される装置であり得る。この例は、1つ又は複数のコンピューティングデバイス、例えば図7に関連して説明するコンピューティングデバイス700によって実行される1つ又は複数の

コンピュータ実行可能命令として実装され得る。

【0237】

ステップ610において、プロダクション環境の1つ又は複数の観察を受け取る。1つ又は複数の観察は、プロダクション環境501に関して説明した形態のいずれかであり得る。説明した例では、プロダクション環境の1つ又は複数の観察は、装置が受け取った物理的環境の1つ又は複数の画像であり得る。

【0238】

ステップ620において、プロダクション環境の1つ又は複数の観察に基づいてプロダクション環境においてエンティティによって実行可能な複数の行動のうちの1つ又は複数の行動が、エージェントを使用して決定される。1つ又は複数の行動は、図4のステップ414に関連して及び／又は図5Aのプロダクションエージェント502Aに関して説明した方法に従って決定され得る。

【0239】

ステップ630において、1つ又は複数の行動は、プロダクション環境においてエンティティによって実行させられる。1つ又は複数の行動は、図5Aの制御インターフェース503A及びプロダクションエンティティ504に関連して説明した方法に従って実行させられ得る。

【0240】

コンピューティングデバイス

図7は、実施形態による方法を実施するために使用され得るハードウェアの概略図である。これがほんの一例に過ぎず、他の配列が使用され得ることに留意されたい。

【0241】

ハードウェアは、コンピューティングセクション700を備える。この特定の例では、このセクションの構成要素は、まとめて描かれている。しかしながら、それらが必ずしもクロケットされる必要がないことは認識されるであろう。

【0242】

コンピューティングシステム700の構成要素は、処理ユニット713（中央処理装置CPUなど）と、システムメモリ701と、システムメモリ701を含む様々なシステム構成要素を処理ユニット713に結合するシステムバス711とを含み得るが、これらに限定されるものではない。システムバス711は、メモリバス又はメモリコントローラ、周辺機器用バス、及び多様なバスアーキテクチャのいずれかを使用するローカルバスなどを含むいくつかのタイプのバス構造のいずれかであり得る。コンピューティングセクション700はまた、バス711に接続された外部メモリ715を含む。

【0243】

システムメモリ701は、読取専用メモリのような揮発性又は不揮発性メモリの形態のコンピュータ記憶媒体を含む。例えば、スタートアップ中に、コンピュータ内の要素間での情報の伝達を助けるルーチンを含む基本入力出力システム（BIOS）703は、典型的に、システムメモリ701に記憶される。加えて、システムメモリは、オペレーティングシステム705と、アプリケーションプログラム707と、CPU713によって使用されるプログラムデータ709とを含む。

【0244】

また、インターフェース725は、バス711に接続される。インターフェースは、コンピュータシステムが更なるデバイスから情報を受け取るためのネットワークインターフェースであり得る。インターフェースはまた、ユーザが特定のコマンドなどに応答することを可能にするユーザインターフェースであり得る。

【0245】

この例では、ビデオインターフェース717が提供される。ビデオインターフェース717は、グラフィックスプロセッシングメモリ721に接続されたグラフィックスプロセッシングユニット719を備える。

【0246】

10

20

30

40

50

グラフィックスプロセッシングユニット（GPU）719は、データ並列動作へのその適応により、ニューラルネットワークトレーニングのような、エージェントのトレーニングに特によく適している。従って、ある実施形態では、エージェントをトレーニングするための処理は、CPU 713とGPU 719との間で分けられ得る。

【0247】

いくつかの実施形態では、異なるハードウェアが、エージェントのトレーニング及びエージェントを使用したタスクの実行に使用され得ることに留意されたい。例えば、エージェントのトレーニングは、1つ又は複数のローカルデスクトップ又はワークステーションコンピュータ上で、又は、クラウドコンピューティングシステムのデバイス上で行われ得、これは、1つ又は複数の個別のデスクトップ又はワークステーションGPUと、1つ又は複数の個別のデスクトップ又はワークステーションCPU、例えばPC指向アーキテクチャを有するプロセッサと、かなりの量の揮発性システムメモリ、例えば16GB以上とを含み得る。一方で、例えば、エージェントによるタスクの実行は、モバイル又は埋め込まれたハードウェアを使用し得、これは、システムオンチップ（SoC）の一部としてモバイルGPUを含むかGPUを含まず、1つ又は複数のモバイル又は埋め込まれたCPU、例えばモバイル指向アーキテクチャ又はマイクロコントローラ指向アーキテクチャを有するプロセッサと、より少ない量の揮発性メモリ、例えば1GB未満とを含み得る。エージェントのトレーニングに使用されるハードウェアは、エージェントを使用してタスクを実行することに使用されるハードウェアと比べて、はるかに多くの算出能力を有することができ、例えば、毎秒より多くの演算を実行し、より多くのメモリを有することができる。例えば1つ又は複数のニューラルネットワークを使用して推論を実行することによって、エージェントを使用してタスクを実行するほうが、例えば1つ又は複数のニューラルネットワークをトレーニングすることによってエージェントをトレーニングするよりも消費するリソースが計算的に著しく少ないため、より少ないリソースを有するハードウェアを使用することが可能である。更に、エージェントを使用してタスクを実行すること、例えば1つ又は複数のニューラルネットワークを使用して推論を実行することに使用される計算リソースを低減するための技法が使用され得る。そのような技法の例には、モデルの蒸留が含まれ、ニューラルネットワークの場合には、プルーニング及び量子化のようなニューラルネットワーク圧縮技法が含まれる。

【0248】

SPLメトリックの結果

図8Aは、ナビゲーションタスクを実行する際の例となる実施形態のSuccess weighted by (normalized inverse) Path Length (SPL) メトリックを示すグラフである。

【0249】

SPLメトリックは、以下の通りであり：

【数5】

$$\frac{1}{N} \sum_{i=1}^N \mathbb{1}_{\text{succ},i} \frac{l_i}{\max(l_i, p_i)}$$

ここで、 l は、エージェントの経路の長さであり、 p は、スタートからゴールまでの最短経路の長さであり、 N は、エピソードの数である。

【0250】

グラフは、5回以上の試行にわたって単一のトレーニング／テストシーンに対する検証SPLの平均及び標準偏差を表すプロットを示している。衝突のないエージェントは、例えば、シミュレート環境においてトレーニングされたエージェントを実環境において使用するSim2Real転送にとって重要である。

【0251】

図8Bは、例となる実施形態を使用して実行されたナビゲーションを比較する平面図である。エージェントが進む経路が、衝突力学に依存せずに壁に沿ってスライドしており、

実施形態が衝突を回避していることに留意されたい。

【0252】

タスク空間視覚化

図9は、様々なトレーニングエポックEにおける、ゴールクリアランス及びエージェントクリアランスという2つの異なる幾何学的特性にわたるナビゲーションタスク空間の推定成功確率 $f_{\pi}(h)$ 推定値を視覚化したものである。時間の経過と共に、タスク分布はシフトする。適応フィルタリングは、このシフトを考慮する。

【0253】

アブレーション実験

図10A及び図10Bは、本明細書で説明するNavACL-GOID及びNavACL-Adaptiveという2つの実施形態に従ってタスクを取得することが視覚ナビゲーションに対して与える影響を、一様にランダムなタスクを取得する方法と比較する実験に関する。この実験では、エージェントのトレーニングに使用されるすべてのパラメータ、例えば近接ポリシー最適化で使用されるパラメータが一定に保たれる。500万個のサンプルの3回のナビゲーション試行が、GibsonデータセットからCoopers town環境で実行される。一様にランダムなタスクは、Habitatの内蔵タスクジェネレータを使用して生成される。ランダムなタスクを取得する方法の場合、Habitatの内蔵タスクジェネレータによって生成されたタスクが利用される。NavACL-GOIDの場合、Habitatの内蔵タスクジェネレータを使用してタスクが生成され得、推定成功確率が0.4~0.6（両端値を含む）の中間容易性範囲（すなわち中間困難性範囲）内であると推定されたものが選択される。NavACL-Adaptiveの場合、タスクは、ハイパーパラメータ $\beta=1$ 、 $\gamma=0.1$ を用いて、アルゴリズム3に従って取得される。見て分かるように、NavACL-GOID及びNavACL-Adaptiveは両方とも、一様にランダムなタスクを取得することよりもパフォーマンスが著しく優れている。NavACL-Adaptiveは、NavACL-GOIDに勝る改善を示しているため、残りの評価ではNavACL-Adaptiveが使用される。従って、以下でNavACLに言及する場合、NavACL-Adaptiveを参照するものと理解されるべきである。

【0254】

図10Aは、一様にランダムなタスク、NavACL-GOID、及びNavACL-Adaptiveについて、単一の視覚ナビゲーションテストトレーニング環境での3回の試行における検証エピソード成功率を比較するグラフである。

【0255】

図10Bは、一様にランダムなタスク、NavACL-GOID、及びNavACL-Adaptiveについて、平均的な測地タスク距離を例示するグラフである。NavACL-Adaptiveは、スタートからゴールまでの距離をゆっくりと増やした。時間が経過し、ポリシーが学習するにつれて、NavACLは、困難性を増加させる。NavACL-Adaptiveは、時間が経過するにつれて一様にランダムなタスクの距離に近づくが、NavACL-GOIDは近づくかない。

【0256】

シミュレーション実験

エージェントは、例となる実施形態に従ってトレーニングされる。エージェントは、図11A~図11Cに示されている3つの初見の(unseen)テスト環境において10回の試行で30個のエピソードにわたって評価される。図11Aは、40m²であるCoopers town環境を示す。図11Bは、60m²であるAvonia環境を示す。図11Cは、61m²であるHometown環境を示す。

【0257】

テストセットは、Habitatチャレンジデータセットと同じである一様なサンプリングを使用してランダムに生成される。ターゲットオブジェクトは、半径11cmのフットボール（サッカーボール）である。タスク実行の試行は、最大150個のタイムステッ

に制限されており、例えば150個のタイムステップのタイムステップ限界に達すると失敗する。タイムステップの最大量を150よりも多くしても、ほとんど改善されないことが分かった。複数の実行可能な行動は、±30度の回転及び0.2mの前方移動である。
【0258】

ランダムなエージェントは、パフォーマンスの下限を提供するためにランダムな行動を選択する。NavACLエージェントは、タスク取得のためにNavACL-Adaptiveを使用して、深度、再シェーディング（テクスチャ除去及びリライティング（re-lighting））、及びセマンティック特徴を使用してトレーニングされる。NavACL Zero-Shotは、NavACLと同一であるが、テスト中に、ターゲットをボールから大きな花瓶に変更して、異なる形状及びサイズの初見のターゲットへのゼロショットセマンティック一般化を評価する。パフォーマンスの上限を設けるために、様々なバックグラウンドを持つ10人のボランティアを採用した。人間のエージェントは、エージェントと類似した方法でトレーニングされ、テストされる。ボランティアは、制御に慣れるまでトレーニングセットをプレイし、エージェントと同じ観察を受け取り、エージェントと同じ行動を実行することができるようになった。以下の表及び図12A～図12Bに関連して結果を示す。表では、「ポリシー」として記述されたエージェントが参照される。表は、初見の環境での10回の試行にわたる衝突のないエージェントのシミュレーション結果を示している。

【表4】

ポリシー	シーン	成功率		SPL	
		μ	σ	μ	σ
ランダム	全て	0.03	0.03	0.02	0.03
	Cooperstown	0.07	0.00	0.06	0.00
	Avonia	0.01	0.00	0.01	0.00
	Hometown	0.00	0.00	0.00	0.00
NavACL Zero-Shot	全て	0.36	0.21	0.17	0.10
	Cooperstown	0.55	0.11	0.25	0.08
	Avonia	0.43	0.04	0.20	0.02
	Hometown	0.09	0.06	0.06	0.03
NavACL	全て	0.42	0.19	0.24	0.09
	Cooperstown	0.63	0.14	0.31	0.09
	Avonia	0.50	0.06	0.21	0.04
	Hometown	0.12	0.08	0.09	0.06
人間	全て	0.79	0.20	0.56	0.15
	Cooperstown	0.76	0.22	0.54	0.17
	Avonia	0.89	0.13	0.63	0.13
	Hometown	0.72	0.19	0.51	0.13

【0259】

図12Aは、SPLメトリックを使用したNavACLエージェントのトレーニング及び検証結果を示している。

【0260】

図12Bは、NavACLエージェントが改善するにつれて、NavACL-Adaptiveを使用してより困難なタスクが取得されることを例示する。最初は、より短い経路（測地的距離）及びより狭い通路（エージェントクリアランス）が、エージェントを誘導するのに役立つ。エージェントが改善するにつれて、エージェントは、通路誘導なしにナビゲートし、障害物の近くに配置されたゴールに到達する（ゴールクリアランス）。

【0261】

結果から分かるように、NavACLエージェント及びNavACL Zero-Shot

エージェントは、シミュレーションにおいてセマンティックに指定されたターゲットを見つけることができ、ランダムよりもはるかに良好に機能し、NavACL Zero-Shotの場合には、ゼロショットセマンティック一般化能力を明らかに示す。

【0262】

Cooperstown環境では、エージェントは、成功率が人間レベルのパフォーマンスから1標準偏差以内であり、これは、場合によっては一部の人間よりもパフォーマンスが優れていることを示唆している。

【0263】

示された結果は、示された実験において実行され得たトレーニングの量が制限されていた。モデルは、2000万個のタイムステップに対してトレーニングされたものであり、以前の実験は、6000万個のタイムステップにおいても依然として改善を示していた。

【0264】

シミュレート環境においてトレーニングされたエージェントの実環境における評価
シミュレート環境においてトレーニングされたエージェントを実環境において使用することは、Sim2Real転送と称され得る。実環境における説明したエージェントのパフォーマンスを評価するために、Turtlebot3の車輪型ロボット（AGV）及びDJIのTelloクアドロプタ（UAV）の修正なしにエージェントが転送された。

【0265】

AGVは、モーションプリミティブ（単一行動）の閉ループ制御にホイールエンコーダを使用するが、行動間のオドメトリは推定しない。AGVは、1つは初見のオブジェクトで、初見のセマンティッククラスからのものである3つのオブジェクト及び3つの環境に及ぶ7つのタスクでテストされた。AGVのSPLを測定するためにホイールオドメトリが使用される。AGVは、テスト中に走行した29mにわたって一度も衝突も経験せず、アクチュエータノイズと、地形（硬材、カーペット、及びラグ）によって引き起こされる車輪スリップとに対してロバストであった。結果を以下の表に示す。

【表5】

シーン	ターゲット	タスク距離(m)	SPL
オフィス1	フットボール	2.95	0.92
	フットボール	3.09	0.53
	フットボール	2.92	0.42
	フットボール	1.67	0.18
オフィス2	フットボール	9.10	0.65
自宅	オレンジ色のフットボール	4.19	0.92
自宅	花瓶	5.14	0.54

【0266】

UAVは、IRセンサを使用して高さを決定し、IMUを使用してモーションプリミティブ及びホバリング安定性について非常にノイズの多い位置推定値を得る。別個のエージェントは、UAVについてはトレーニングされなかった。AGVの高さ及びカメラの視野でトレーニングされたエージェントが使用された。AGVの高さ及びカメラの視野は、UAVとは大きく異なる（それぞれ0.2m対0.75～1.5m及び68度対47度）。AGVのためにトレーニングされたエージェントは、UAVに対して驚くべき有効性を示し、これは、高度の一般化を示唆している。UAVは、カメラの三脚の脚間を飛行し、出入口を通り、動いている人の周りを衝突することなく幾度も飛行することができた。この驚くべき一般化は、単一のナビゲーションモデルが、同様のモーションプリミティブを実装する多様なタイプのロボットで使用するためにトレーニングされ得ることを意味する。

【0267】

図13Aは、自動搬送車（AGV）がエージェントを使用して屋内物理的環境を通して誘導される様子の例示である。AGVは、いくつかの見たことのない障害物がターゲットまでの経路に散在している状態で、自宅シーンにおいて花瓶ターゲットまでナビゲートする。ゼロショットセマンティックナビゲーション能力を強調するために、トレーニング中に使用された以前のセマンティックターゲット（フットボール及びピンクの星のボール）が存在している。エージェントは、360度回ってそのオプションを評価し（1）、次いで、机とテントとの間の経路を進み（2）、青いテントの前の広く開けたエリアに向かって軌道を調整し（3）、360度回転した（4）。エージェントは、自転車、本棚、及び青いテントを囲むエリアを探索した（5、6、7）。（8）においてターゲット検出が行われ、AGVは、ターゲットまでの最短コースを進んだ（9、10）。

10

【0268】

図13Bは、無人航空機（UAV）がエージェントを使用して屋内物理的環境を通して誘導される様子の例示である。廊下を飛行している（1）最中に、UAVは、空いている小部屋に気付く（2）。針に糸を通すように、椅子の車輪とシートとの間を飛行する（3）。小部屋を探索した（4、5）後、その小部屋から出て、衝突することなく隣接するオープンオフィスに向かう（6）。

【0269】

変形例

特定の実施形態を説明してきたが、これらの実施形態は、例としてのみ提示されており、本発明の範囲を限定することを意図しない。実際、新規のデバイスと、本明細書で説明する方法とは、様々な形態で具現化され得、更に、本明細書で説明するデバイス、方法、及び製品の形式の様々な省略、置換、及び変更は、本発明の精神から逸脱することなく行われ得る。添付の特許請求の範囲とその同等物とは、本発明の範囲及び精神内に収まることとなるそのような形態又は変形をカバーするように意図されている。

20

30

40

50

【図面】
【図 1 A】

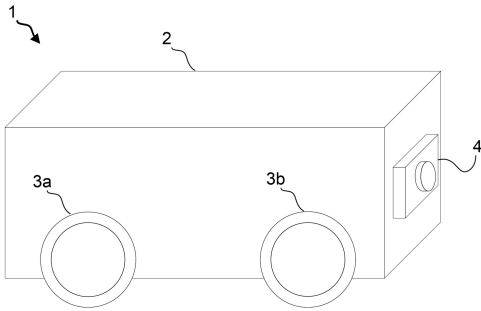


FIG. 1A

【図 1 B】

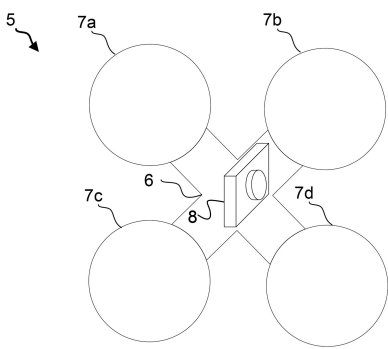


FIG. 1B

【図 1 C】

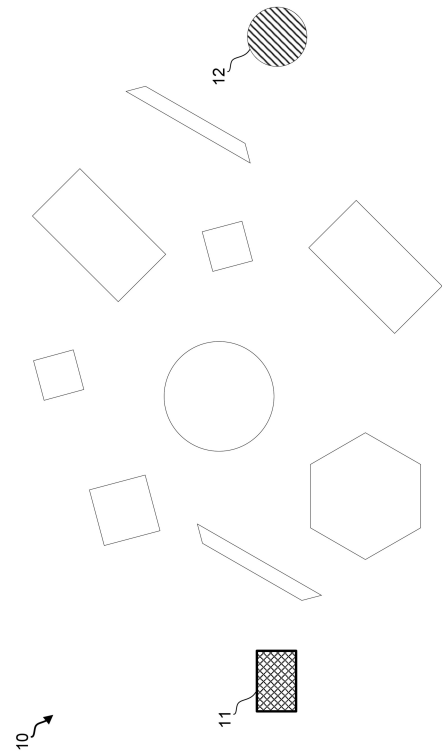


FIG. 1C

【図 1 D】

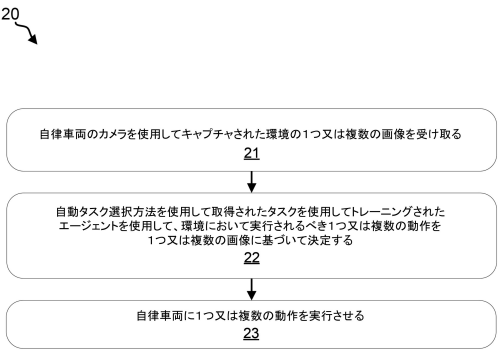
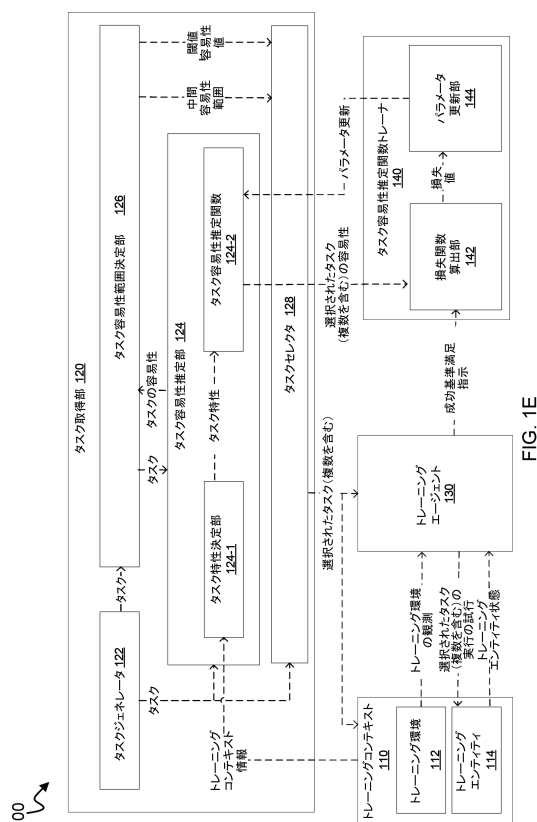


FIG. 1D

【図 1 E】



【圖 2】

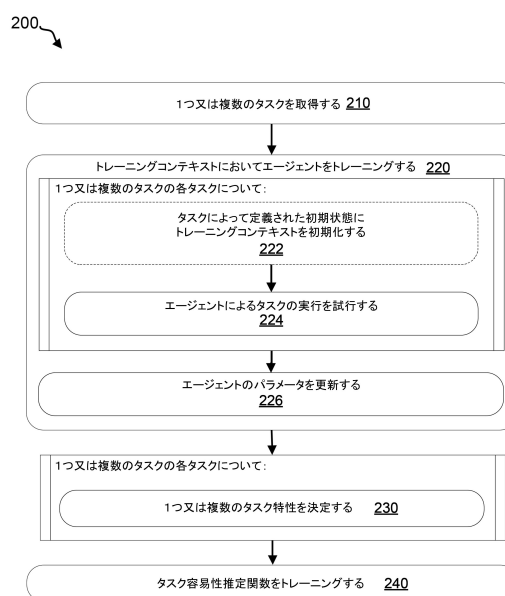


FIG. 2

【図 3 A】

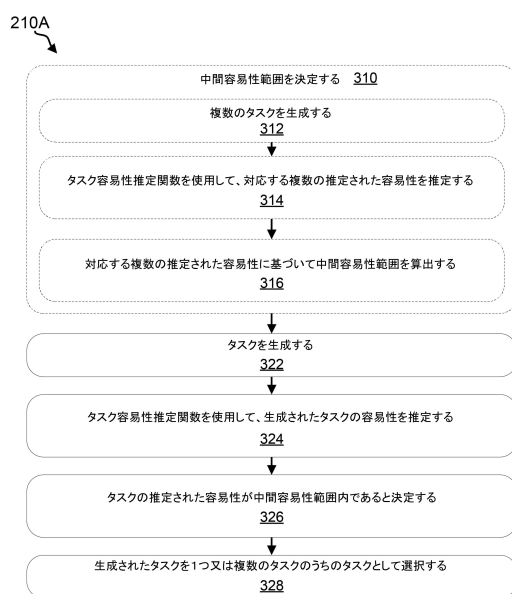


FIG. 3A

【図 3 B】

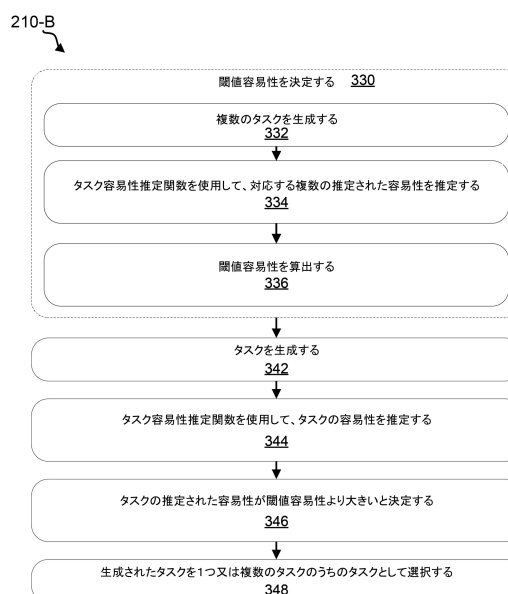


FIG. 3B

【図 3 C】

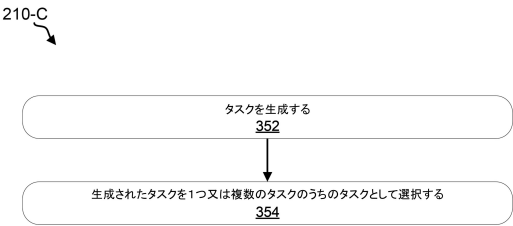


FIG. 3C

【図 4】

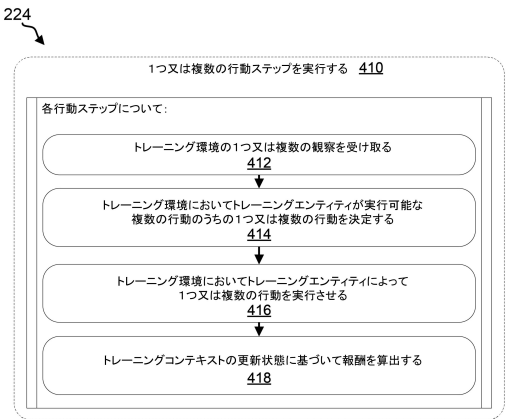


FIG. 4

【図 5 A】

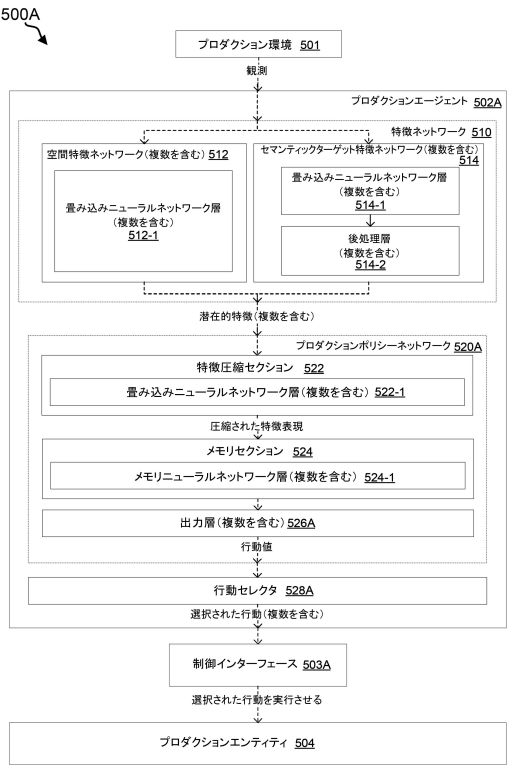


FIG. 5A

【図 5 B】

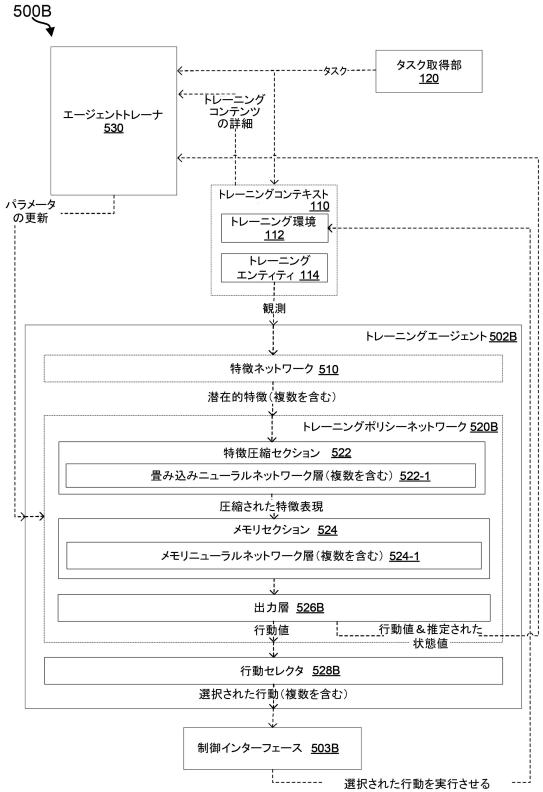
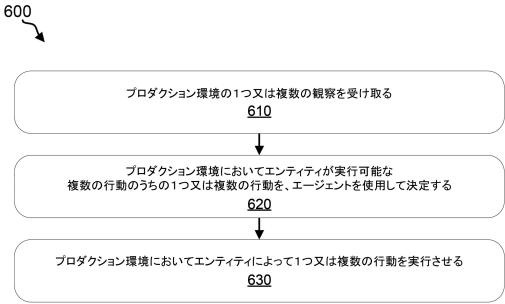
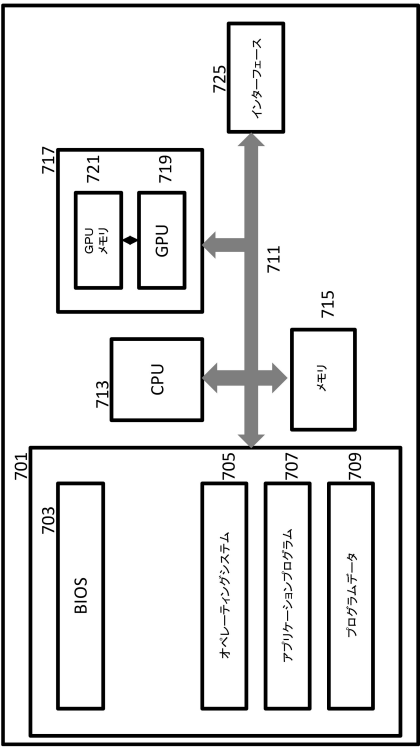


FIG. 5B

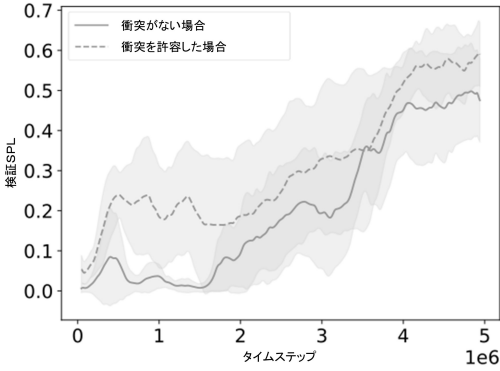
【図 6】



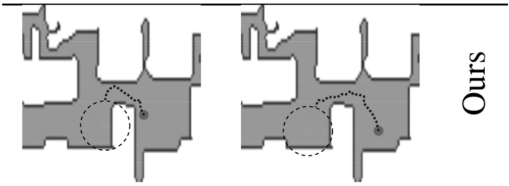
【図 7】



【図 8 A】



【図 8 B】



10

20

30

40

50

【図 9】

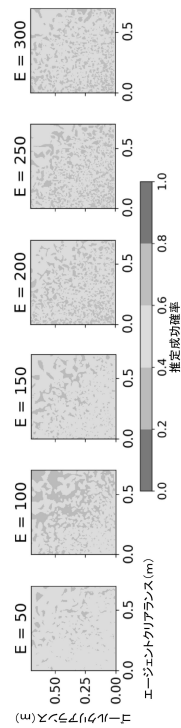


FIG. 9

【図 10 A】

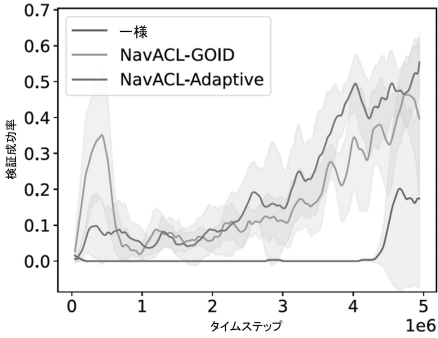


FIG. 10A

10

20

【図 10 B】

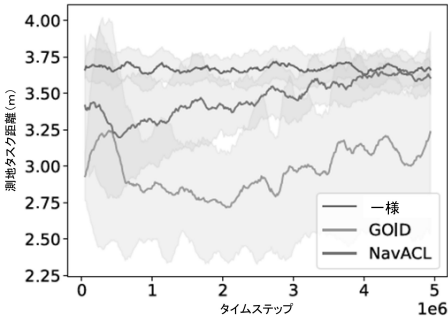


FIG. 10B

【図 11 A】



FIG. 11A

30

40

50

【図 1 1 B】

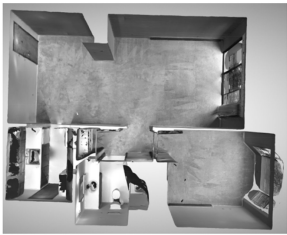


FIG. 11B

【図 1 1 C】



FIG. 11C

10

20

【図 1 2 A】

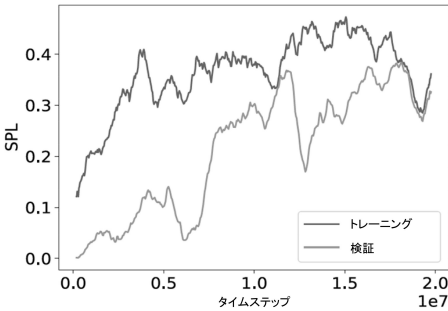


FIG. 12A

【図 1 2 B】

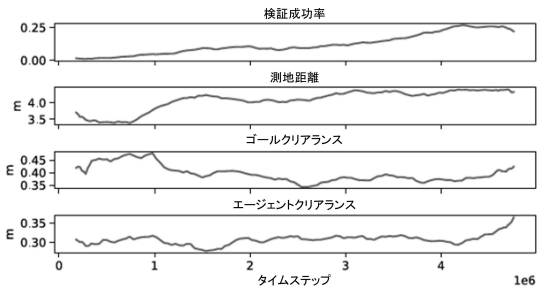


FIG. 12B

30

40

50

【図 13 A】



FIG. 13A

【図 13 B】



FIG. 13B

10

20

30

40

50

フロントページの続き

ト、ケンブリッジシャー、ケンブリッジ、ミルトン・ロード、ケンブリッジ・サイエンス・パーク 208、トシバ・リサーチ・ヨーロッパ・リミテッド、ケンブリッジ・リサーチ・ラボラトリー内

(72)発明者 ルドラ パウデル

イギリス国、シービー4・0ジーゼット、ケンブリッジシャー、ケンブリッジ、ミルトン・ロード、ケンブリッジ・サイエンス・パーク 208、トシバ・リサーチ・ヨーロッパ・リミテッド、ケンブリッジ・リサーチ・ラボラトリー内

(72)発明者 ステファン リウィッキ

イギリス国、シービー4・0ジーゼット、ケンブリッジシャー、ケンブリッジ、ミルトン・ロード、ケンブリッジ・サイエンス・パーク 208、トシバ・リサーチ・ヨーロッパ・リミテッド、ケンブリッジ・リサーチ・ラボラトリー内

(72)発明者 ロベルト シポラ

イギリス国、シービー4・0ジーゼット、ケンブリッジシャー、ケンブリッジ、ミルトン・ロード、ケンブリッジ・サイエンス・パーク 208、トシバ・リサーチ・ヨーロッパ・リミテッド、ケンブリッジ・リサーチ・ラボラトリー内

審査官 松本 泰典

(56)参考文献 国際公開第2020/136770 (WO, A1)

国際公開第2018/220829 (WO, A1)

特開2017-168029 (JP, A)

特開2019-125345 (JP, A)

特開2019-185742 (JP, A)

特表2019-509905 (JP, A)

(58)調査した分野 (Int.Cl., DB名)

G05B 13/02

G06N 20/00

G05D 1/02

G08G 1/00