(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2021-64387
(P2021-64387A)
(43)公開日　令和3年4月22日(2021.4.22)

(51)Int.Cl.		F I		テーマコード（参考）
G06N 3/08	(2006.01)	G06N	3/08	140
G06N 20/00	(2019.01)	G06N	20/00	

審査請求 有　請求項の数 20　ＯＬ　外国語出願　（全 24 頁）

(21)出願番号　特願2020-213556 (P2020-213556)
(22)出願日　令和2年12月23日 (2020.12.23)
(62)分割の表示　特願2019-546227 (P2019-546227)の分割
原出願日　平成30年2月26日 (2018.2.26)
(31)優先権主張番号　62/463,558
(32)優先日　平成29年2月24日 (2017.2.24)
(33)優先権主張国・地域又は機関
米国 (US)

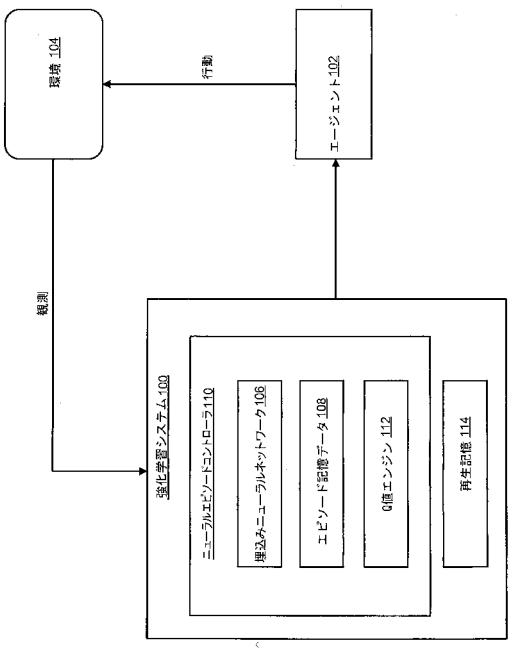
（特許庁注：以下のものは登録商標）
１．ＴＥＮＳＯＲＦＬＯＷ

(71)出願人　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　ベニグノ・ウリアーマルティネス
イギリス・Ｎ１Ｃ・４ＡＧ・ロンドン・パンクラス・スクエア・６

最終頁に続く

(54)【発明の名称】ニューラルエピソード制御

(57)【要約】
【課題】強化学習に関する。
【解決手段】方法は、複数の行動のそれぞれのための個々のエピソード記憶データを維持することと、エージェントによって対話される環境の現在の状態を特徴付ける現在の観測を受け取ることと、現在の観測に関する現在のキー埋込みを生成するために、埋込みニューラルネットワークのパラメータの現在の値に応じて埋込みニューラルネットワークを使用して現在の観測を処理することと、複数の行動の各行動について、距離測定に応じて、現在のキー埋込みに対する、行動のためのエピソード記憶データ内のp近傍キー埋込みを判断することと、および行動のためのエピソード記憶データ内のp近傍キー埋込みによってマッピングされるリターン推定から、行動に関するQ値を判断することと、行動に関するQ値を使用して、エージェントによって行われることになる行動として行動を複数の行動から選択することと、を含む。
【選択図】図1



【特許請求の範囲】
【請求項１】
1つまたは複数のコンピュータによって、複数の行動の各行動のための個々のエピソード記憶データを維持するステップであって、各行動のための前記エピソード記憶データが、個々の複数のキー埋込みのそれぞれを個々のリターン推定にマッピングする、ステップと、
前記1つまたは複数のコンピュータによって、エージェントによって対話される環境の現在の状態を特徴付ける現在の観測を受け取るステップと、
前記1つまたは複数のコンピュータによって、前記現在の観測に関する現在のキー埋込みを生成するために、埋込みニューラルネットワークのパラメータの現在の値に応じて前記埋込みニューラルネットワークを使用して前記現在の観測を処理するステップと、
前記複数の行動の各行動について、
1つまたは複数のコンピュータによって、距離測定に応じて、前記現在のキー埋込みに対する、前記行動のための前記エピソード記憶データ内のp近傍キー埋込みを判断するステップ、および
1つまたは複数のコンピュータによって、前記行動のための前記エピソード記憶データ内の前記p近傍キー埋込みによってマッピングされる前記リターン推定から、前記行動に関するQ値を判断するステップであって、前記行動に関する前記Q値が、前記現在の観測に応答して前記行動を行う前記エージェントから生じることになる予測されるリターンである、ステップと、
前記1つまたは複数のコンピュータによって、また前記行動に関する前記Q値を使用して、前記現在の観測に応答して前記エージェントによって行われることになる行動として前記行動を前記複数の行動から選択するステップとを含む、
方法。
【請求項２】
行われることになる前記行動を選択するステップが、
行われることになる前記行動として最も高いQ値を有する前記行動を選択するステップを含む、
請求項1に記載の方法。
【請求項３】
行われることになる前記行動を選択するステップが、
確率εで行われることになる前記行動として最も高いQ値を有する前記行動を選択するステップ、および確率1-εで行動の所定のセットからランダム行動を選択するステップを含む、
請求項1に記載の方法。
【請求項４】
前記行動のための前記エピソード記憶データ内の前記p近傍キー埋込みによってマッピングされる前記リターン推定から、前記行動に関するQ値を判断するステップが、
前記距離測定に応じて、前記p近傍キー埋込みと、前記現在のキー埋込みの間の距離から、前記行動のための前記エピソード記憶データ内の前記p近傍キー埋込みのそれぞれに対する個々の重みを判断するステップと、
前記行動のための前記エピソード記憶データ内の前記p近傍キー埋込みのそれぞれに対して、個々の重み付き推定リターンを判断するために、前記キー埋込みに対する前記重みによって、前記キー埋込みにマッピングされた前記推定リターンに重み付けするステップとを含む、
請求項1から3のいずれか一項に記載の方法。
【請求項５】
前記行動に関する前記Q値を判断するステップが、
前記行動に対する前記重み付き推定リターンを合計するステップと、
前記合計された重み付き推定リターンを前記Q値として使用するステップとを含む、

請求項4に記載の方法。
【請求項６】
前記行動に関する前記Q値を判断するステップが、
前記行動に対する前記重み付き推定リターンを合計するステップと、
前記Q値を生成するために、リターンニューラルネットワークのパラメータの現在の値に応じて、前記リターンニューラルネットワークを通じて、前記合計された重み付き推定リターンを含むネットワーク入力を処理するステップとを含む、
請求項4に記載の方法。
【請求項７】
前記行動に関する前記Q値を判断するステップが、
前記Q値を生成するために、リターンニューラルネットワークのパラメータの現在の値に応じて、前記リターンニューラルネットワークを通じて、前記重み付き推定リターンを含むネットワーク入力を処理するステップを含む、
請求項4に記載の方法。
【請求項８】
前記ネットワーク入力が、前記現在の観測をさらに含む、
請求項6または7に記載の方法。
【請求項９】
前記埋込みニューラルネットワークが、畳み込みニューラルネットワークである、
請求項1から8のいずれか一項に記載の方法。
【請求項１０】
前記埋込みニューラルネットワークが、1つまたは複数の回帰型ニューラルネットワーク層を含む、
請求項1から9のいずれか一項に記載の方法。
【請求項１１】
前記現在の観測に応答して、選択された前記行動を行う前記エージェントから生じる現在のリターンを判断するステップをさらに含む、
請求項1から10のいずれか一項に記載の方法。
【請求項１２】
前記行動のための前記エピソード記憶データ内の前記キー埋込みのいずれかに前記現在のキー埋込みがマッチするかどうかを判断するステップと、
前記行動のための前記エピソード記憶データ内の前記キー埋込みのいずれにも前記現在のキー埋込みがマッチしないとき、
前記現在のリターンに前記現在のキー埋込みをマッピングするデータを前記行動のための前記エピソード記憶データに追加するステップとをさらに含む、
請求項11に記載の方法。
【請求項１３】
前記行動のための前記エピソード記憶データ内のキー埋込みに前記現在のキー埋込みがマッチするとき、
前記現在のリターンと、前記マッチするキー埋込みによって現在マッピングされる前記リターン推定との重み付き平均に、前記マッチするキー埋込みをマッピングするために前記エピソード記憶データを更新するステップをさらに含む、
請求項12に記載の方法。
【請求項１４】
訓練の観測、訓練の選択済み行動、および訓練のリターンを含む訓練タプルを再生記憶からサンプリングするステップと、
前記訓練の観測から前記訓練の選択済み行動に関するQ値を判断するステップと、
前記キー埋込み、推定リターン、および前記埋込みニューラルネットワークの前記パラメータの前記現在の値を更新するために、前記訓練の選択済み行動に関する前記Q値と前記訓練のリターンとの間のエラーの勾配を逆伝搬させるステップとをさらに含む、

請求項11から13のいずれか一項に記載の方法。
【請求項１５】
前記現在の観測、前記選択済み行動、およびブートストラップリターンを含む新しいタプルを生成するステップと、
前記新しいタプルを前記再生記憶に追加するステップとをさらに含む、
請求項14に記載の方法。
【請求項１６】
1つまたは複数のコンピュータ、および前記1つまたは複数のコンピュータによって実行されると、請求項1から15のいずれか一項に記載の方法を行うことを前記1つまたは複数のコンピュータに行わせるように動作可能な命令を格納する1つまたは複数のストレージデバイスを備えるシステム。
【請求項１７】
1つまたは複数のコンピュータによって実行されると、請求項1から15のいずれか一項に記載の方法を行うことを前記1つまたは複数のコンピュータに行わせる命令でエンコードされたコンピュータ可読記憶媒体。
【発明の詳細な説明】
【技術分野】
【０００１】
関連出願への相互参照
本出願は、2017年2月24日に出願の米国仮出願第62/463,558号の優先権を主張する。先行出願の開示は、本出願の一部と考えられ、本出願の開示に参照により組み込まれる。
【０００２】
本明細書は、強化学習に関する。
【背景技術】
【０００３】
強化学習システムにおいて、エージェントは、環境の現在の状態を特徴付ける観測を受け取ることに応答して強化学習システムによって選択される行動を行うことによって環境と対話する。
【０００４】
強化学習システムの中には、ニューラルネットワークの出力に応じて所与の観測を受け取ることに応答してエージェントによって行われることになる行動を選択するものもある。
【０００５】
ニューラルネットワークは、非線形ユニットの1つまたは複数の層を用いて、受け取られた入力に対する出力を予測する機械学習モデルである。ニューラルネットワークの中には、出力層の他に1つまたは複数の隠れ層を含むディープニューラルネットワークであるものもある。各隠れ層の出力は、ネットワーク内の次の層、すなわち、次の隠れ層または出力層への入力として使用される。ネットワークの各層は、パラメータの個々のセットの現在の値に応じて、受け取られた入力から出力を生成する。
【先行技術文献】
【非特許文献】
【０００６】
【非特許文献１】Mnih et al.「Human-level control through deep reinforcement learning.」、Nature、518(7540):529-533、2015年
【非特許文献２】Bentley、Jon Louis.「Multidimensional binary search trees used for associative searching.」Commun. ACM、18(9):509-517、1975年9月
【発明の概要】
【課題を解決するための手段】
【０００７】
本明細書は、1つまたは複数の場所における1つまたは複数のコンピュータ上のコンピュ

ータプログラムとして実行されるシステムが、システムによって維持されるエピソード記憶データを使用して、環境と対話するエージェントによって行われることになる行動を行動のセットから、どのように選択することができるかを説明する。

【０００８】

全体的に、本明細書において説明される主題の1つの革新的な態様は、1つまたは複数のコンピュータによって、複数の行動の各行動のための個々のエピソード記憶データを維持することを含むコンピュータ実行方法の中で具体化されることが可能である。各行動のためのエピソード記憶データは、個々の複数のキー埋込み(key embedding)のそれぞれを個々のリターン推定にマッピングする。方法は、1つまたは複数のコンピュータによって、エージェントによって対話される環境の現在の状態を特徴付ける現在の観測を受け取ることを含む。方法は、1つまたは複数のコンピュータによって、現在の観測に関する現在のキー埋込みを生成するために、埋込みニューラルネットワークのパラメータの現在の値に応じて埋込みニューラルネットワークを使用して現在の観測を処理することを含む。複数の行動の各行動について、方法は、1つまたは複数のコンピュータによって、距離測定に応じて、現在のキー埋込みに対する、行動のためのエピソード記憶データ内のp近傍キー埋込み(p nearest key embedding)を判断することと、1つまたは複数のコンピュータによって、行動のためのエピソード記憶データ内のp近傍キー埋込みによってマッピングされるリターン推定から、行動に関するQ値を判断することと、を含む。行動に関するQ値は、現在の観測に応答して行動を行うエージェントから生じることになる予測されるリターンである。方法は、1つまたは複数のコンピュータによって、また行動に関するQ値を使用して、現在の観測に応答してエージェントによって行われることになる行動として行動を複数の行動から選択することをさらに含む。

【０００９】

前述の実施形態および他の実施形態はそれぞれ任意選択で、以下の特徴の1つまたは複数を単独で、または組み合わせて含むことができる。

【００１０】

行われることになる行動を選択することは、行われることになる行動として最も高いQ値を有する行動を選択することを含むことができる。行われることになる行動を選択することは、確率εで行われることになる行動として最も高いQ値を有する行動を選択すること、および確率1-εで行動の所定のセットからランダム行動を選択することを含むことができる。

【００１１】

行動のためのエピソード記憶データ内のp近傍キー埋込みによってマッピングされるリターン推定から、行動に関するQ値を判断することは、距離測定に応じて、p近傍キー埋込みと、現在のキー埋込みの間の距離から、行動のためのエピソード記憶データ内のp近傍キー埋込みのそれぞれに対する個々の重みを判断することと、行動のためのエピソード記憶データ内のp近傍キー埋込みのそれぞれに対して、個々の重み付き推定リターンを判断するために、キー埋込みに対する重みによって、キー埋込みにマッピングされた推定リターンに重み付けすることと、を含むことができる。

【００１２】

行動に関するQ値を判断することは、行動に対する重み付き推定リターンを合計することと、合計された重み付き推定リターンをQ値として使用することと、を含むことができる。

【００１３】

行動に関するQ値を判断することは、行動に対する重み付き推定リターンを合計することと、Q値を生成するために、リターンニューラルネットワークのパラメータの現在の値に応じて、リターンニューラルネットワークを通じて、合計された重み付き推定リターンを含むネットワーク入力を処理することと、を含むことができる。

【００１４】

行動に関するQ値を判断することは、Q値を生成するために、リターンニューラルネット

ワークのパラメータの現在の値に応じて、リターンニューラルネットワークを通じて、重み付き推定リターンを含むネットワーク入力を処理することを含むことができる。

【００１５】

ネットワーク入力は、現在の観測を含むことができる。埋込みニューラルネットワークは、畳み込みニューラルネットワークであることが可能である。埋込みニューラルネットワークは、1つまたは複数の回帰型ニューラルネットワーク層を含むことができる。

【００１６】

方法は、現在の観測に応答して、選択済み行動を行うエージェントから生じる現在のリターンを判断することをさらに含むことができる。この現在のリターンは、エピソード記憶に格納されたデータに依存するQ値を判断するために使用されるので、現在のブートストラップリターン(bootstrapped return)と後で呼ばれ、このQ値はこれ自体が、特にキー埋込みおよび対応するリターン推定といった、エピソード記憶に格納されたデータを更新するために使用される。

【００１７】

方法は、行動のためのエピソード記憶データ内のキー埋込みのいずれかに現在のキー埋込みがマッチするかどうかを判断することを含むことができる。行動のためのエピソード記憶データ内のキー埋込みのいずれにも現在のキー埋込みがマッチしないとき、方法は、現在の(自力)リターンに現在のキー埋込みをマッピングするデータを行動のためのエピソード記憶データに追加することを含むことができる。行動のためのエピソード記憶データ内のキー埋込みに現在のキー埋込みがマッチするとき、方法は、現在の(自力)リターンと、マッチするキー埋込みによって現在マッピングされるリターン推定との重み付き平均に、マッチするキー埋込みをマッピングするためにエピソード記憶データを更新することを含むことができる。

【００１８】

方法は、現在の観測、選択済み行動、および現在の(自力)リターンを含む新しいタプルを生成することと、新しいタプルを再生記憶(replay memory)に追加することと、をさらに含むことができる。方法は、訓練の観測、訓練の選択済み行動、および訓練のリターンを含む訓練タプルを再生記憶からサンプリングすることと、上述の個々の方法を使用して、訓練の観測から訓練の選択済み行動に関するQ値を判断することと、キー埋込み、推定リターン、および埋込みニューラルネットワークのパラメータの現在の値を更新するために、訓練の選択済み行動に関するQ値と訓練のリターンとの間のエラーの勾配を逆伝搬させることと、をさらに含むことができる。現在のリターンと同様に、訓練のリターンは、訓練のブートストラップリターンと後で呼ばれる。

【００１９】

本明細書において説明される主題は、以下の長所の1つまたは複数を実現するように、特定の実施形態において実行されることが可能である。本明細書において説明されるように使用される記憶、すなわち(行動のそれぞれのための個々のエピソード記憶モジュールを含む)エピソード記憶データを組み込むことによって、システムは、様々な強化学習タスクをより良く行うことができる。これは、本明細書において説明されるエピソード記憶アーキテクチャが、最近の経験(状態表現および対応するリターン値の推定)を迅速に統合できるからであり、したがって記憶は、エピソード的であるとして適切に説明される。エピソード記憶は、強化学習エージェントの将来の挙動にこの情報が迅速に統合されることを可能にし、将来の挙動を改善する。したがって、いくつかの実装形態において、システムは、以前のシステムより桁違いに少ない環境との対話からタスクを学習することができる。エピソード記憶アーキテクチャは、より小さいニューラルネットワークをさらに可能にすることができ、すなわちシステムは、より少ない数のネットワークパラメータを有することができる一方で、依然として、同種の性能またはより良い性能を実現する。さらに、システムの構成要素は完全に区別可能であり、エピソード記憶が読み書きされる手法のために、構成要素は、より少ない時間で訓練されることが可能であり、所与の強化学習タスクに関する高レベルの性能にシステムが到達するために必要な、例えば記憶の数および

処理能力の大きさといった、コンピューティングリソースの数を減らす。
【００２０】
特に、本明細書において説明されるエピソード記憶は、キー埋込みをQ値にマッピングし、キー埋込みに対するコンテキストベースの検索を使用して、行動選択処理中に有用なQ値を取得する。一般に、所与の行動について、キー埋込みは、エージェントによって所与の行動が行われたことに応答した、観測についてのエピソード記憶データ内の埋込みである。埋込みニューラルネットワークによって提供される穏やかに変化する安定した表現を有することによって、エピソード記憶に格納されたキー埋込みは、比較的安定したままであり、したがって安定した訓練を可能にする。強化学習のための他のニューラル記憶アーキテクチャとは対照的に、エピソード記憶から取得されるQ値は、さらに速く更新されることが可能であり、これは、ネットワーク全体にあてはまる確率勾配の降下の典型的に穏やかな重み更新を緩和するのに役立つ。例えば、訓練中に区別可能なエピソード記憶に格納されたキー埋込みとリターン推定のペアを更新するための学習率より低い学習率が埋込みニューラルネットワークを更新するために使用されることが可能である。いくつかの実装形態において、システムは、モンテカルロ値の推定とバックアップされたオフポリシの推定との間に挿入するNステップのQ値関数を使用してQ値を推定し、これは、報酬伝搬の速さと最適性との間のトレードオフをもたらす。より一般に、リターン推定は、キー埋込みによって表されるような環境の状態と関連付けられた価値関数の推定を含むことができる。訓練中、訓練エピソードの各時間ステップにおいて、このようなキーと値のペアがエピソード記憶に挿入されることが可能である。
【００２１】
さらに、LSTMなどの他の記憶アーキテクチャとは異なり、本明細書において説明されるエピソード記憶は、これが、ゆっくり学習し、かなりの量の時間をかける可能性があるので、いつ記憶に書き込むべきかを学習しようとしない。むしろ、システムは、(記憶の最大容量に到達するまで)全ての経験をエピソード記憶に書き込むことを選び、記憶を既存の記憶アーキテクチャに比べて非常に大きく成長させる。システムは、例えばkdツリーに基づく最近傍アルゴリズムといった、高速近似最近傍アルゴリズムを使用することによって、この大きい記憶から効率的に読み込むことができる。エピソード記憶に格納されたデータは、訓練エピソードが、タスクの終わりまたは完了まで訓練することを含むことができる複数の訓練エピソードにわたって持続することができる。
【００２２】
本明細書の主題の1つまたは複数の実施形態の詳細は、添付の図面および下記の説明において示される。主題の他の特徴、態様、および長所は、説明、図面、および特許請求の範囲から明らかになるであろう。
【図面の簡単な説明】
【００２３】
【図１】強化学習システムの例を示す図である。
【図２】行動に関するQ値を判断するための処理の例のブロック図である。
【図３】エピソード記憶データを更新するため、および新しいタプルを再生記憶に追加するための処理の例の流れ図である。
【図４】再生記憶からの訓練タプルを使用して、ニューラルエピソードコントローラを訓練するための処理の例の流れ図である。
【発明を実施するための形態】
【００２４】
様々な図面における同様の参照番号および名称は同様の要素を示す。
【００２５】
本明細書は一般に、ニューラルエピソードコントローラを使用して、行動のセットから、環境と対話する強化学習エージェントによって行われることになる行動を選択する強化学習システムを説明する。
【００２６】

環境と対話するために、エージェントは、環境の現在の状態を特徴付けるデータを受け取り、受け取られたデータに応答して行動空間から行動を行う。環境の状態を特徴付けるデータは、本明細書において観測と呼ばれることになる。

【００２７】

いくつかの実装形態において、環境はシミュレート環境であり、エージェントは、シミュレート環境と対話する1つまたは複数のコンピュータとして実行される。例えば、シミュレート環境はビデオゲームであることが可能であり、エージェントは、ビデオゲームをプレイするシミュレートされたユーザであることが可能である。

【００２８】

いくつかの他の実装形態において、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する機械エージェントである。例えば、エージェントは、特定のタスクを行うために環境と対話するロボットであることが可能である。別の例として、エージェントは、環境の中を進む自律走行車両または半自律走行車両であることが可能である。これらのケースにおいて、観測は、機械エージェントが環境と対話するときに、例えば、カメラ、ライダーセンサ、温度センサ、などといった、機械エージェントの1つまたは複数のセンサによってキャプチャされたデータであることが可能である。

【００２９】

いくつかの他の実装形態において、環境は現実世界環境であり、エージェントは、設備(例えばデータセンタ)の運営を管理する。これらの実装形態において、行動は、設備の運営における様々な設定への変更を生じる行動(例えば、一定の構成要素をオン/オフすること、および様々な構成要素の設定を調節すること)である可能性がある。

【００３０】

いくつかの他の実装形態において、環境は現実世界環境であり、エージェントは、(例えば、モバイルデバイス上の、またはデータセンタにおける)コンピューティングリソースにわたるタスクの分散を管理する。これらの実装形態において、行動は、特定のコンピューティングリソースにタスクを割り当てることである可能性がある。

【００３１】

図1は、強化学習システム100の例を示す。強化学習システム100は、下記で説明されるシステム、構成要素、および技法が実行される1つまたは複数の場所において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行されるシステムの例である。

【００３２】

強化学習システム100は、環境104と対話する強化学習エージェント102によって行われることになる行動を選択する。すなわち、強化学習システム100は、各観測が環境104の個々の状態を特徴付ける観測を受け取り、各観測に応答して、観測に応答して強化学習エージェント102によって行われることになる行動を行動空間から選択する。エージェント102が選択済み行動を行った後、環境104は新しい状態に遷移し、システム100は、環境104の次の状態を特徴付ける別の観測および報酬を受け取る。報酬は、エージェント102が選択済み行動を行った結果として環境104からシステム100またはエージェント102によって受け取られる数値であることが可能である。

【００３３】

強化学習システム100は、ニューラルエピソードコントローラ110を使用して、観測に応答してエージェント102によって行われることになる行動を選択する。ニューラルエピソードコントローラ110は、埋込みニューラルネットワーク106、エピソード記憶データ108、およびQ値エンジン112を含む。

【００３４】

ニューラルエピソードコントローラ110を訓練するために、システム100は、再生記憶114を維持する。再生記憶114は訓練タプルを格納する。各訓練タプルは、訓練の観測、訓練の選択済み行動、および訓練のブートストラップリターンを含む。ブートストラップリターンは、図3を参照しながら下記でさらに詳細に説明される。

【００３５】

ニューラルエピソードコントローラ110は、観測に応答してエージェント102によって行われる可能性のある、可能な行動のセットAにおける各行動aのための個々のエピソード記憶モジュールを含むエピソード記憶データ108を維持する。

【００３６】

一般に、所与の行動について、所与の行動のためのエピソード記憶モジュールは、キー埋込みに対する個々のリターン推定に複数のキー埋込みのそれぞれをマッピングする。所与の行動のためのエピソード記憶モジュール内のキー埋込みは、エージェントによって所与の行動が行われたことに応答した観測のキー埋込みである。所与の観測のキー埋込みによってマッピングされる個々のリターン推定は、観測に応答して所与の行動をエージェント102が行った後に、システム100またはエージェント102によって受け取られた報酬の、例えば合計または平均といった、時間を無視した(time-discounted)組合せの推定であることが可能である。

【００３７】

具体的には、各行動$a \in A$について、個々のエピソード記憶モジュールは$M_a=(K_a, V_a)$と表されることが可能であり、ここで、K_aおよびV_aは、それぞれが同じ数のベクトルを含むベクトルの配列である。配列K_aおよび配列V_aの大きさは動的に変化させることができる。エピソード記憶モジュールは、配列k_aのi番目の要素であるキー埋込みh_iを、配列V_aのi番目の要素である対応するリターン推定Q_iにマッピングする。下記でさらに詳細に説明されることになる、検索および書込みという、エピソード記憶モジュール上で可能な2つの動作が存在する。

【００３８】

エージェント102によって対話されている環境104の現在の状態を特徴付ける現在の観測sをシステム100が受け取ると、ニューラルエピソードコントローラ110は、現在の観測sに対する現在のキー埋込みhを生成するために、埋込みニューラルネットワーク106を使用して現在の観測sを処理する。現在の観測sは、埋込みニューラルネットワーク106のパラメータの現在の値に応じて処理される。いくつかの実装形態において、埋込みニューラルネットワーク106は、畳み込みニューラルネットワークである。いくつかの実装形態において、埋込みニューラルネットワーク106は、1つまたは複数の回帰型ニューラルネットワーク層を含む。

【００３９】

いくつかの実装形態において、埋込みニューラルネットワーク106は、1つまたは複数の完全に接続された層が後に続く1つまたは複数の畳み込みニューラルネットワーク層を含む。このようなニューラルネットワークのアーキテクチャの例は、Mnih et al.「Human-level control through deep reinforcement learning.」、Nature、518(7540):529-533、2015年において説明される。

【００４０】

ニューラルエピソードコントローラ110は次に、現在のキー埋込みhを使用して、個々のエピソード記憶モジュール上で各行動$a \in A$を検索する。特に、ニューラルエピソードコントローラ110は、各行動$a \in A$について、例えば核関数を使用して計算された距離といった距離測定に応じて、現在のキー埋込みhに対する個々のエピソード記憶モジュール$M_a=(K_a, V_a)$内のp近傍キー埋込みを判断する。pは、例えば、10、25、50、または100といった、1より大きい整数であることが可能である。いくつかのケースにおいて、pは固定されることが可能である。他のいくつかのケースにおいて、pは、(例えば、観測ごとに、またはいくつかの観測ごとに)変更されることが可能である。

【００４１】

いくつかの実装形態において、拡張可能な非常に大きいエピソード記憶モジュールに照会するために、ニューラルエピソードコントローラ110は、例えばkdツリーに基づく最近傍アルゴリズムといった、近似最近傍アルゴリズムを使用して検索を行うことができる。Kdツリーに基づく最近傍アルゴリズムは、Bentley、Jon Louis.「Multidimensional binary search trees used for associative searching.」Commun. ACM、18(9):509-517、197

5年9月において説明される。エピソード記憶モジュールは、エピソード記憶モジュールが効率的にアクセスされることが可能であることを保証するために、例えばp=25またはp=50といった小さい数に、p近傍埋込みの数をさらに減らすことができる。

【００４２】

いくつかの実装形態において、エピソード記憶モジュールM_aが検索された後、ニューラルエピソードコントローラ110は、新しいキー埋込みおよびその関連付けられたリターン推定をエピソード記憶モジュールM_aに書き込む。書き込まれることになるキー埋込みは、検索されたキー埋込みhに対応する。書き込まれることになる関連付けられたリターン推定は、図3を参照しながら下記で詳細に説明される処理に基づいて判断されることが可能な現在のブートストラップリターンである。

【００４３】

ニューラルエピソードコントローラ110は、新しいキー埋込みおよび関連付けられたリターン推定を配列K_aおよび配列V_aの終わりにそれぞれ付け加えることによって、これらをエピソード記憶モジュールに書き込む。エピソード記憶モジュール内にキー埋込みが既に存在する場合、この関連付けられたリターン推定は、エピソード記憶モジュール内で重複されるのではなく更新される。エピソード記憶モジュールを更新することは、図3を参照しながら下記でさらに詳細に説明される。

【００４４】

行動aのための個々のエピソード記憶モジュールM_aにおけるp近傍キー埋込みを判断した後、ニューラルエピソードコントローラ110は、Q値エンジン112を使用して、行動に関するQ値を判断する。Q値は、現在の観測sに応答して行動aをエージェント102が行うことから生じるであろう予測されるリターンである。Q値エンジン112は、行動aのためのエピソード記憶モジュールM_a内のp近傍キー埋込みによってマッピングされるリターン推定から、行動aに関するQ値を判断するように構成される。行動に関するQ値を判断することは、図2を参照しながら下記でさらに詳細に説明される。

【００４５】

可能な行動Aのセット内の行動に関するQ値が判断された後、ニューラルエピソードコントローラ110は、判断されたQ値を使用して、現在の観測sに応答してエージェント102によって行われることになる行動としての行動を可能な行動Aのセットから選択する。

【００４６】

いくつかの実装形態において、選択済み行動をエージェント102が行った後、システム100は、現在の観測sに応答して選択済み行動をエージェント102が行うことから生じる現在のブートストラップリターンを判断する。システム100は次に、現在の観測、選択済み行動、およびブートストラップリターンを含む新しいタプルを生成する。システムは、新しいタプルを再生記憶114に追加する。現在のブートストラップリターンを判断し、再生記憶を更新するための処理は、図3を参照しながら下記でさらに詳細に説明される。

【００４７】

システム100は、再生記憶114からサンプリングされた訓練タプルを使用して、埋込みニューラルネットワーク106のパラメータの値を更新するように、埋込みニューラルネットワーク106を訓練することができる。特に、訓練タプルは、訓練のためのミニバッチを形成するようにサンプリングされることが可能である。システム100は次に、所与の訓練行動に関する予測されるQ値と、再生記憶114からのミニバッチに対するブートストラップリターン推定との間の損失を最小化するように、埋込みニューラルネットワーク106のパラメータの値を更新することができる。ニューラルエピソードコントローラのアーキテクチャは完全に区別可能であるので、システム100は、勾配降下法を使用することによって損失を最小化することができる。埋込みニューラルネットワーク106を訓練することは、図4を参照しながら下記でさらに詳細に説明される。

【００４８】

図2は、例えば図1のニューラルエピソードコントローラ110といったニューラルエピソードコントローラが行動に関するQ値をどのように判断するかを示すブロック図である。

【００４９】
ニューラルエピソードコントローラ110の埋込みニューラルネットワーク106は、環境の現在の状態を特徴付ける観測s(202)を受け取る。例えば、観測202は、ロボットエージェントによってキャプチャされた画像である。埋込みニューラルネットワークは次に、画像に対するキー埋込みhを生成するために、埋込みニューラルネットワークのパラメータの現在の値に応じて画像を処理する。
【００５０】
各行動a∈Aについて、ニューラルエピソードコントローラ110は、現在のキー埋込みhを使用して、個々の、行動aのためのエピソード記憶モジュール204上で検索を行う。行動aのためのエピソード記憶モジュール204は、$M_a=(K_a, V_a)$と表されることが可能であり、ここで、k_aおよびV_aはベクトルの配列であり、それぞれが同じ数のベクトルを含む。配列K_aおよび配列V_aの大きさは動的に変化させることができる。エピソード記憶モジュール204は、配列k_aのi番目の要素であるキー埋込みh_iを、配列V_aのi番目の要素である対応するリターン推定Q_iにマッピングする。
【００５１】
特に、ニューラルエピソードコントローラ110は、例えば核関数といった距離測定に応じて、現在のキー埋込みhに対するエピソード記憶モジュール204内のp近傍キー埋込みを判断する。例えば、ニューラルエピソードコントローラ110は、以下の核関数を使用して、現在のキー埋込みhからエピソード記憶モジュール204内のキー埋込みh_iのそれぞれまでの距離を計算することができる。
【００５２】
【数１】

$$k(h, h_i) = \frac{1}{\|h - h_i\|_2^2 + \delta}$$

【００５３】
ここで、δは、例えば$\delta=10^{-3}$といった所定の定数である。ニューラルエピソードコントローラ110は、核関数に応じて現在のキー埋込みhへの最短距離を有するエピソード記憶モジュール204内のpのキー埋込みを選択することができる。
【００５４】
行動aのためのエピソード記憶モジュール204内のp近傍キー埋込みが判断された後、Q値エンジン112は、上記の距離測定に応じて、p近傍キー埋込みと、現在のキー埋込みhの間の距離から、p近傍キー埋込みのそれぞれに対する個々の重みを判断するように構成される。具体的には、p近傍キー埋込みのセット内の各キー埋込みh_iに対する個々の重みw_iは、以下のように判断されることが可能である。
【００５５】
【数２】

$$w_i = \frac{k(h, h_i)}{\sum_j k(h, h_j)}$$

【００５６】
p近傍キー埋込みの各キー埋込みh_iについて、Q値エンジンは、キー埋込みh_iにマッピングされた推定リターンQ_iに、キー埋込みh_iに対する重みw_iを掛けて、個々の重み付き推定リターン$Y_i=w_iQ_i$を得るように構成される。
【００５７】
いくつかの実装形態において、Q値エンジン112は、以下のように、行動aに対する重み付き推定リターンを合計することと、現在の観測sおよび行動aに関するQ値として、合計された重み付き推定リターンを使用することと、を行うように構成される。

【００５８】
【数３】

$$Q(s,a) = \sum_i w_i Q_i$$

【００５９】
いくつかの他の実装形態において、ニューラルエピソードコントローラ110は、Q値を生成するために、例えば1つまたは複数の完全に接続された層を含むネットワークといった、リターンニューラルネットワークを使用して、p近傍埋込みおよび推定リターンを処理することができる。
【００６０】
例えば、いくつかのケースにおいて、ニューラルエピソードコントローラ110は、Q値を生成するために、リターンニューラルネットワークのパラメータの現在の値に応じて、リターンニューラルネットワークを通じて、重み付き推定リターンの合計を含むネットワーク入力を処理する。他のいくつかのケースにおいて、ニューラルエピソードコントローラ110は、Q値を生成するために、リターンニューラルネットワークのパラメータの現在の値に応じて、リターンニューラルネットワークを通じて、重み付き推定リターンを含むネットワーク入力を処理する。
【００６１】
図3は、エピソード記憶データを更新するため、および新しいタプルを再生記憶に追加するための処理300の例の流れ図である。便宜上、処理300は、1つまたは複数の場所に置かれた1つまたは複数のコンピュータのシステムによって行われるものとして説明されることになる。例えば、本明細書による適切にプログラムされた、例えば図1の強化学習システム100といった、強化学習システムは、処理300を行うことができる。
【００６２】
システムは、時間ステップtにおける現在の観測s_tに応答して、選択済み行動aを行うエージェントから生じる現在のブートストラップリターンを判断する(ステップ302)。システムは、現在の時間ステップiに続くN-1の時間ステップにおいてシステムによって受け取られたN-1の時間を無視した報酬を合計することによって第1の期間を判断する。システムは、時間ステップt+Nにおいて観測s_{t+N}に対する時間を無視した最大Q値である第2の期間を判断する。観測s_{t+N}に対する最大Q値を見つけるために、システムは、各行動aのための全てのエピソード記憶モジュールM_aを照会し、最も高い推定Q値を選択する。システムは、現在のブートストラップリターンを判断するために第1の期間と第2の期間を合計する。
【００６３】
特に、現在のブートストラップリターンは、以下のNステップのQ値推定方程式を使用して計算されることが可能である。
【００６４】
【数４】

$$Q^{(N)}(s_t,a) = \sum_{j=0}^{N-1} \gamma^j r_{t+j} + \gamma^N \max_{a'} Q(s_{t+N},a')$$

【００６５】
ここで、γ^jおよびγ^Nは無視する率であり、a'は、時間ステップt+Nにおける観測s_{t+N}に対する最も高い推定Q値を生じる行動である。Nは所定の整数であり、例えばNは、50、100、または200であることが可能である。
【００６６】
システムは、行動のためのエピソード記憶データ(すなわちエピソード記憶モジュール)内のキー埋込みのいずれかに、現在のキー埋込みがマッチするかどうかを判断する(ステ

ップ304)。

【００６７】

行動のためのエピソード記憶データ内のキー埋込みに現在のキー埋込みがマッチするとき、システムは、現在のブートストラップリターンの重み付き平均、およびマッチするキー埋込みによって現在マッピングされるリターン推定に、マッチするキー埋込みをマッピングするようにエピソード記憶データを更新する(ステップ306)。

【００６８】

特に、現在のキー埋込みhがキー埋込みh_iにマッチすると仮定すると、システムは、

$$Q_i + \alpha (Q^{(N)}(s, a) - Q_i)$$

に、マッチするキー埋込みh_iをマッピングするようにエピソード記憶データを更新し、ここで、$Q^N(s, a)$は、現在のブートストラップリターンであり、Q_iは、マッチするキー埋込みh_iによって現在マッピングされるリターン推定であり、αは、Q更新の学習率である。学習率αは変更されることが可能である。例えば、いくつかのケースにおいて、学習率は、安定した表現で繰り返し訪れる状態がこれらのQ値推定を迅速に更新できるようにするために、高い値を持つことができる。

【００６９】

行動のためのエピソード記憶データ内のキー埋込みのいずれにも現在のキー埋込みがマッチしないとき、システムは、現在のブートストラップリターンに現在のキー埋込みをマッピングするデータを、行動のためのエピソード記憶データに追加する(ステップ308)。

【００７０】

任意選択で、システムは、現在の観測、選択済み行動、およびブートストラップリターンを含む新しいタプルを生成することができる(ステップ310)。

【００７１】

システムは次に、新しいタプルを再生記憶に追加することができる(ステップ312)。

【００７２】

図4は、再生記憶からの訓練タプルを使用して、ニューラルエピソードコントローラを訓練するための処理400の例の流れ図である。便宜上、処理400は、1つまたは複数の場所に置かれた1つまたは複数のコンピュータのシステムによって行われるものとして説明されることになる。例えば、本明細書による適切にプログラムされた、例えば図1の強化学習システム100といった、強化学習システム、は、処理400を行うことができる。

【００７３】

システムは、訓練タプルを再生記憶からサンプリングする(ステップ402)。

【００７４】

いくつかのケースにおいて、システムは、ニューラルエピソードコントローラをより効果的に訓練するために、再生記憶から訓練タプルのミニバッチをランダムにサンプリングする。

【００７５】

各訓練タプルは、訓練の観測、訓練の選択済み行動、および訓練のブートストラップリターンを含む。

【００７６】

各訓練タプルについて、システムは、訓練の観測から訓練の選択済み行動に関するQ値を判断する(ステップ404)。システムは、図2を参照しながら上記で詳細に説明された処理に従うことによって訓練の選択済み行動に関するQ値を判断する。

【００７７】

システムは、(i)エピソード記憶データ内のキー埋込みおよび推定リターンを更新するため、ならびに(ii)埋込みニューラルネットワークのパラメータの現在の値を更新するために、訓練の選択済み行動に関するQ値と、訓練のブートストラップリターンとの間のエラーの勾配を逆伝搬させる(ステップ406)。訓練中、システムは、エピソード記憶データに照会した後にエピソード記憶データを更新するために使用される学習率より低い学習率を使用することができる。

【００７８】
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に関連して用語「構成された」を使用する。1つまたは複数のコンピュータのシステムにとって、特定の動作または行動を行うように構成されることは、動作中に、動作または行動をシステムに行わせるソフトウェア、ファームウェア、ハードウェア、またはこれらの組合せをシステムがシステム上にインストールしたということを意味する。1つまたは複数のコンピュータプログラムにとって、特定の動作または行動を行うように構成されることは、データ処理装置によって実行されると、動作または行動を装置に行わせる命令を1つまたは複数のプログラムが含むということを意味する。
【００７９】
本明細書において説明される主題および機能的な動作の実施形態は、デジタル電子回路機器の中、現実に具体化されたコンピュータソフトウェアもしくはファームウェアの中、本明細書において開示された構造およびこれらの構造上の均等物を含むコンピュータハードウェアの中、またはこれらの1つもしくは複数の組合せの中で実行されることが可能である。本明細書において説明される主題の実施形態は、1つまたは複数のコンピュータプログラム、すなわち、データ処理装置による実行のため、またはデータ処理装置の動作を制御するための、有形の非一時的ストレージ媒体上にエンコード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行されることが可能である。コンピュータストレージ媒体は、機械可読ストレージデバイス、機械可読ストレージ基板、ランダムもしくはシリアルアクセスメモリデバイス、またはこれらの1つもしくは複数の組合せであることが可能である。一方またはさらに、プログラム命令は、データ処理装置による実行のために、適切な受信装置への送信のための情報をエンコードするために生成される、例えば、機械生成された電気信号、光信号、または電磁気信号といった、人工的に生成された伝搬信号に対してエンコードされることが可能である。
【００８０】
用語「データ処理装置」はデータ処理ハードウェアを指し、例として、プログラマブルプロセッサ、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装置、デバイス、および機械の全ての種類を包含する。装置は、同様に、例えば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)といった、特殊用途のロジック回路機器であることも、これらをさらに含むことも可能である。装置は、ハードウェアの他に、例えば、プロセッサのファームウェア、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、またはこれらの1つもしくは複数の組合せを構成するコードといった、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を任意選択で含むことができる。
【００８１】
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こと、または説明される可能性もあるコンピュータプログラムは、コンパイラ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むプログラミング言語の任意の形式で書かれることが可能であり、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、もしくはコンピューティング環境において使用するのに適した他のユニットとして含む任意の形式で導入されることが可能である。プログラムは、ファイルシステムにおけるファイルに対応することが可能であるが、対応する必要はない。プログラムは、例えば、マークアップ言語文書の中、当該のプログラムに専用の単一ファイルの中、または例えば、1つもしくは複数のモジュール、サブプログラム、もしくはコードの一部を格納するファイルといった複数の協調ファイル(coordinated file)の中に格納された1つまたは複数のスクリプトといった、他のプログラムまたはデータを保持するファイルの一部に格納されることが可能である。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つのサイトに置かれるか、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導入される

ことが可能である。

【００８２】

本明細書において説明される処理および論理フローは、1つまたは複数のコンピュータプログラムを実行して、入力データに対して演算すること、および出力を生成することによって機能を行う、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行われることが可能である。処理および論理フローは、例えばFPGAもしくはASICといった特殊用途のロジック回路機器によって、または特殊用途のロジック回路機器と1つまたは複数のプログラムされたコンピュータとの組合せによって行われることも可能である。

【００８３】

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、汎用もしくは特殊用途のマイクロプロセッサまたは両方、あるいは他の任意の種類の中央処理装置に基づくことが可能である。一般に、中央処理装置は、リードオンリ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け取ることになる。コンピュータの本質的な要素は、命令を行うため、または実行するための中央処理装置、ならびに命令およびデータを格納するための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、特殊用途のロジック回路機器によって補われること、または特殊用途のロジック回路機器に組み込まれることが可能である。一般に、コンピュータは、同様に、例えば、磁気ディスク、光磁気ディスク、または光ディスクといった、データを格納するための1つまたは複数の大容量ストレージデバイスを含むか、1つまたは複数の大容量ストレージデバイスからデータを受け取ること、データを移送すること、または両方を行うように動作連結されることになる。しかし、コンピュータは、このようなデバイスを有する必要はない。さらに、コンピュータは、例えば、ほんの数例を挙げれば、携帯電話、パーソナルデジタルアシスタント(PDA)、モバイルオーディオプレーヤもしくはモバイルビデオプレーヤ、ゲーム機、全地球測位システム(GPS)受信機、または例えば、ユニバーサルシリアルバス(USB)フラッシュドライブといった携帯型ストレージデバイスといった、別のデバイスに組み込まれることが可能である。

【００８４】

コンピュータプログラム命令およびデータを格納するのに適したコンピュータ可読媒体は、例として、例えば、EPROM、EEPROM、およびフラッシュメモリデバイスといった半導体メモリデバイス、例えば内部ハードディスクまたは取外し可能ディスクといった磁気ディスク、光磁気ディスク、ならびにCD-ROMディスクおよびDVD-ROMディスクを含む、不揮発性のメモリ、媒体、およびメモリデバイスの全ての形式を含む。

【００８５】

ユーザとの対話を行うために、本明細書において説明される主題の実施形態は、ユーザに情報を表示するための、例えばCRT(陰極線管)モニタまたはLCD(液晶ディスプレイ)モニタといった表示デバイス、ならびに、ユーザがコンピュータに入力することができる、例えばマウスまたはトラックボールといったキーボードおよびポインティングデバイスを有するコンピュータ上で実行されることが可能である。同様にユーザとの対話を行うためにデバイスの他の種類が使用されることが可能であり、例えば、ユーザに提供されるフィードバックは、例えば、視覚のフィードバック、聴覚のフィードバック、または触覚のフィードバックといった、知覚によるフィードバックのいずれかの形式であることが可能であり、ユーザからの入力は、音響の入力、会話の入力、または触知可能な入力を含む任意の形式で受け取られることが可能である。さらに、コンピュータは、例えば、ウェブブラウザから受け取られたリクエストに応答して、ユーザのデバイス上のウェブブラウザにウェブページを送ることなど、ユーザによって使用されるデバイスに文書を送ること、およびデバイスから文書を受け取ることによってユーザと対話することができる。同様に、コンピュータは、メッセージアプリケーションを実行している、例えばスマートフォンなどの個人デバイスにテキストメッセージまたは他の形式のメッセージを送ること、および返信として応答メッセージをユーザから受け取ることによって、ユーザと対話することができる。

【００８６】
機械学習モデルを実行するためのデータ処理装置は、例えば、機械学習訓練または制作の共通かつ計算負荷の大きい部分、すなわち、推測、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むこともできる。
【００８７】
機械学習モデルは、例えば、TensorFlowフレームワーク、Microsoft Cognitive Toolkitフレームワーク、Apache Singaフレームワーク、またはApache MXNetフレームワークといった、機械学習フレームワークを使用して実行されること、および導入されることが可能である。
【００８８】
本明細書において説明される主題の実施形態は、例えばデータサーバのようなバックエンド構成要素を含むコンピューティングシステム、または、例えばアプリケーションサーバといったミドルウェア構成要素を含むコンピューティングシステム、または、例えば、グラフィカルユーザインターフェース、ウェブブラウザ、もしくはユーザが本明細書において説明される主題の実装形態と対話できるアプリを有するクライアントコンピュータといったフロントエンド構成要素を含むコンピューティングシステム、または、1つもしくは複数のこのようなバックエンド構成要素、ミドルウェア構成要素、もしく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において実行されることが可能である。システムの構成要素は、例えば通信ネットワークといったデジタルデータ通信の任意の形式または媒体によって相互接続されることが可能である。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)、および例えばインターネットといった広域ネットワーク(WAN)を含む。
【００８９】
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めることができる。クライアントおよびサーバは一般に、互いに離れており、典型的には通信ネットワークを通じて対話する。クライアントとサーバの関係は、個々のコンピュータ上で実行し、互いにクライアントとサーバの関係を有するコンピュータプログラムによって生じる。いくつかの実施形態において、サーバは、例えば、クライアントとして振る舞うデバイスと対話するユーザにデータを表示するため、およびユーザからユーザ入力を受け取るために、例えばHTMLページといったデータをユーザデバイスに送信する。例えばユーザ対話の結果といったユーザデバイスにおいて生成されたデータは、デバイスからサーバにおいて受け取られることが可能である。
【００９０】
本明細書は、多くの特定の実装形態の詳細を収めるが、これらは、任意の発明の範囲、または請求される可能性のあるものの範囲に対する限定としてではなく、むしろ、特定の発明の特定の実施形態に固有である可能性のある特徴の説明として解釈されるべきである。個別の実施形態の背景において、本明細書において説明される一定の特徴は、単一の実施形態の中で組み合わせて実行されることも可能である。逆に、単一の実施形態の背景において説明される様々な特徴は、複数の実施形態の中で別々に、または任意の適切な小結合の中で実行されることも可能である。さらに、一定の組合せの中で振る舞うものとして、また、このように最初に請求されるものとして特徴がなお上述されたが、請求される組合せからの1つまたは複数の特徴は、いくつかのケースにおいて、組合せから削除されることが可能であり、請求される組合せは、小結合または小結合の変形物を対象とすることが可能である。
【００９１】
同様に、動作は、特定の順序で、図面の中に描写され、特許請求の範囲の中で列挙されるが、これは、望ましい結果を実現するために、示された特定の順序で、もしくは順番に、このような動作が行われること、または全ての示された動作が行われることを必要とするものとして理解されるべきではない。一定の状況において、マルチタスク処理および並列処理が有利なことがある。さらに、上述の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、全ての実施形態においてこのような分離を必要とするものとして理解されるべきではなく、説明されるプログラム構成要素およびシステムは全体的に、単一のソフトウェア製品の中で相互に統合されること、または複数のソフトウェア製品の中にパッケージ化されることが可能であるということを理解されたい。

【００９２】

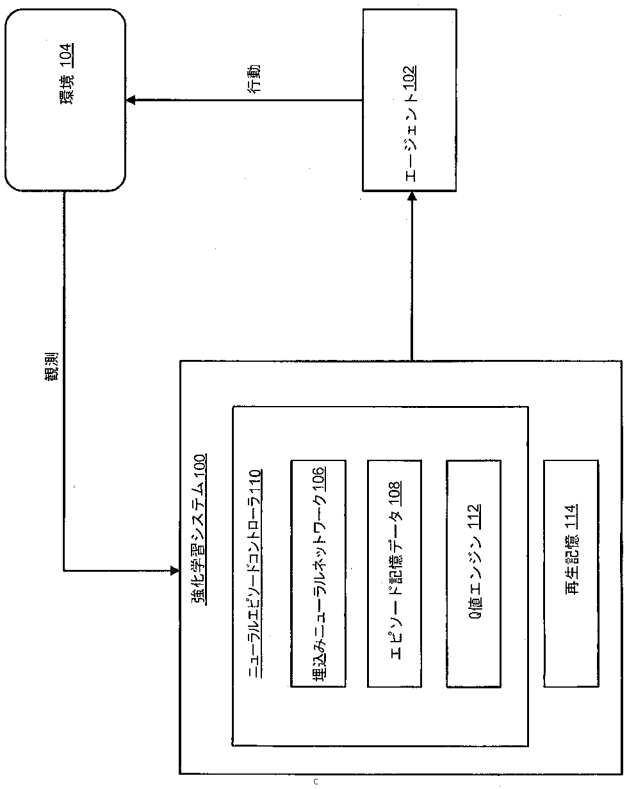
主題の特定の実施形態が説明された。他の実施形態は、以下の特許請求の範囲の範囲に含まれる。例えば、特許請求の範囲の中で列挙される行動は、異なる順序で行われることが可能であるが、それでも望ましい結果を実現する。1つの例として、付随する図において描写された処理は、望ましい結果を実現するために、示された特定の順序、または連続した順序を必ずしも必要としない。いくつかのケースにおいて、マルチタスク処理および並列処理が有利なことがある。

【符号の説明】

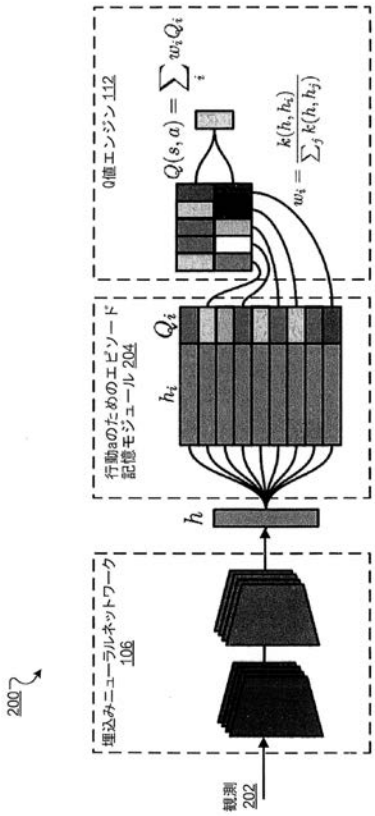
【００９３】

100　強化学習システム、システム
102　エージェント、強化学習エージェント
104　環境
106　埋込みニューラルネットワーク
108　エピソード記憶データ
110　ニューラルエピソードコントローラ
112　Q値エンジン
114　再生記憶
202　観測
204　行動aのためのエピソード記憶モジュール、エピソード記憶モジュール
300　処理
400　処理

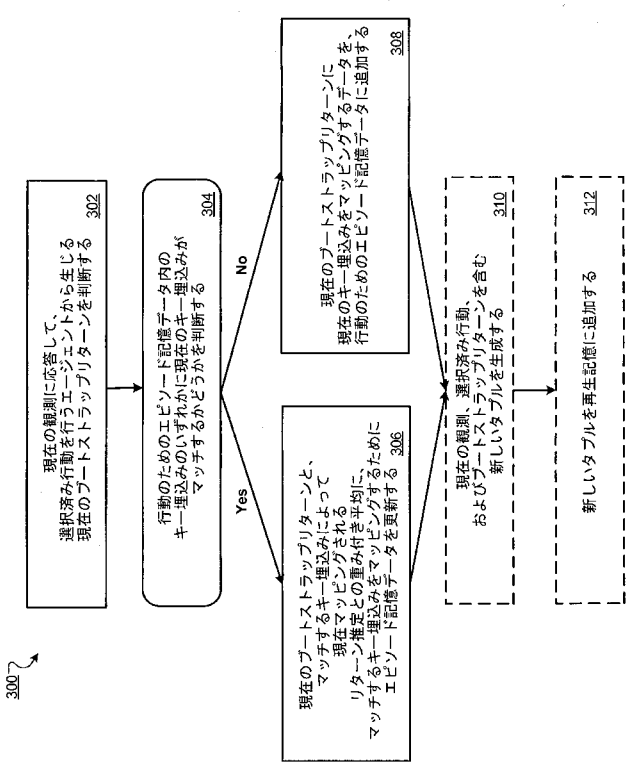
【図１】



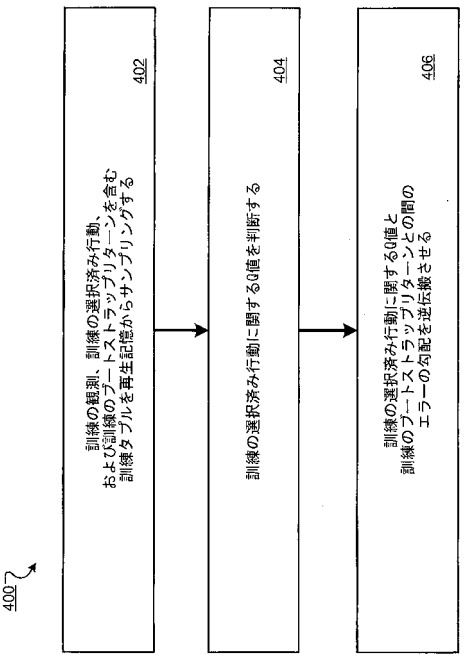
【図２】



【図３】



【図４】



【手続補正書】
【提出日】令和3年1月21日(2021.1.21)
【手続補正１】
【補正対象書類名】特許請求の範囲
【補正対象項目名】全文
【補正方法】変更
【補正の内容】
【特許請求の範囲】
【請求項１】
埋込みニューラルネットワークを含むニューラルエピソードコントローラを訓練するためのコンピュータにより実行される方法であって、
前記ニューラルエピソードコントローラが、観測に応答して、エージェントによって実行される複数の行動の各行動のための個々のエピソード記憶モジュールを含むエピソード記憶データを維持し、
各行動のための前記エピソード記憶モジュールが、個々の複数のキー埋込みのそれぞれを個々のリターン推定にマッピングし、
前記方法が、
1つまたは複数のコンピュータによって、訓練タプルを再生記憶からサンプリングするステップであって、
前記訓練タプルが、訓練の観測、訓練の選択済み行動、および訓練のリターンを含む、ステップと、
前記訓練の観測のための訓練キー埋込みを生成するために、前記埋込みニューラルネットワークのパラメータの現在の値に応じて前記埋込みニューラルネットワークを使用して、前記訓練の観測を処理するステップと、
前記訓練の選択済み行動のための前記エピソード記憶モジュール内の前記キー埋込みの

いずれかに前記訓練の観測に関連付けられた前記訓練キー埋込みがマッチするかどうかを判断するステップと、
前記訓練の選択済み行動のための前記エピソード記憶モジュール内の前記キー埋込みのいずれかに前記訓練キー埋込みがマッチするとき、(i)前記訓練のリターン、(ii)前記マッチするキー埋込みによって現在マッピングされる前記リターン推定、および(iii)エピソード記憶学習率に基づいて計算される新しいリターン推定に、前記マッチするキー埋込みをマッピングするために前記エピソード記憶モジュールを更新するステップと、
前記1つまたは複数のコンピュータによって、前記訓練の観測から前記訓練の選択済み行動のためのQ値を判断するステップであって、
前記選択済み行動のための前記Q値は、前記訓練の観測に応答して前記訓練の選択済み行動を行う前記エージェントから生じることになる予測されるリターンである、ステップと、
前記1つまたは複数のコンピュータによって、前記エピソード記憶学習率より小さい埋込みニューラルネットワーク学習率を使用して、前記埋込みニューラルネットワークのパラメータの現在の値を更新するために、前記訓練の選択済み行動に関する前記Q値と前記訓練のリターンとの間のエラーの勾配を逆伝搬させるステップと
を含む方法。
【請求項２】
前記訓練の選択済み行動に関する前記Q値を判断するステップが、
距離測定に応じて、前記訓練キー埋込みに対する、前記訓練の選択済み行動のための前記エピソード記憶データ内のp近傍キー埋込みを判断するステップと、
前記訓練の選択済み行動のための前記エピソード記憶データ内の前記p近傍キー埋込みによってマッピングされる前記リターン推定から、前記訓練の選択済み行動に関するQ値を判断するステップと
を含む、請求項1に記載の方法。
【請求項３】
前記訓練の選択済み行動のための前記エピソード記憶データ内の前記p近傍キー埋込みによってマッピングされる前記リターン推定から、前記訓練の選択済み行動に関する前記Q値を判断するステップが、
前記距離測定に応じて、前記p近傍キー埋込みと、前記訓練キー埋込みの間の距離から、前記訓練の選択済み行動のための前記エピソード記憶データ内の前記p近傍キー埋込みのそれぞれに対する個々の重みを判断するステップと、
前記訓練の選択済み行動のための前記エピソード記憶データ内の前記p近傍キー埋込みのそれぞれに対して、個々の重み付きリターン推定を判断するために、前記訓練キー埋込みに対する前記重みによって、前記訓練キー埋込みにマッピングされた前記リターン推定に重み付けするステップと
を含む、請求項2に記載の方法。
【請求項４】
前記訓練の選択済み行動に関する前記Q値を判断するステップが、
前記訓練の選択済み行動に対する前記重み付きリターン推定を合計するステップと、
前記合計された重み付きリターン推定を前記Q値として使用するステップと
を含む、請求項3に記載の方法。
【請求項５】
前記訓練の選択済み行動に関する前記Q値を判断するステップが、
前記訓練の選択済み行動に対する前記重み付きリターン推定を合計するステップと、
前記Q値を生成するために、リターンニューラルネットワークのパラメータの現在の値に応じて、前記リターンニューラルネットワークを通じて、前記合計された重み付きリターン推定を含むネットワーク入力を処理するステップと
を含む、請求項3に記載の方法。
【請求項６】

前記訓練の選択済み行動に関する前記Q値を判断するステップが、
前記Q値を生成するために、リターンニューラルネットワークのパラメータの現在の値に応じて、前記リターンニューラルネットワークを通じて、前記重み付きリターン推定を含むネットワーク入力を処理するステップを含む、請求項3に記載の方法。
【請求項７】
前記埋込みニューラルネットワークが、畳み込みニューラルネットワークである、請求項1に記載の方法。
【請求項８】
前記埋込みニューラルネットワークが、1つまたは複数の回帰型ニューラルネットワーク層を含む、請求項1に記載の方法。
【請求項９】
前記訓練の選択済み行動のための前記エピソード記憶データ内の前記キー埋込みのいずれにも前記訓練キー埋込みがマッチしないとき、訓練リターンに前記現在のキー埋込みをマッピングするデータを前記訓練の選択済み行動のための前記エピソード記憶モジュールに追加するステップ
をさらに含む、請求項1に記載の方法。
【請求項１０】
命令を記憶した1つまたは複数のコンピュータ可読記録媒体であって、
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、埋込みニューラルネットワークを含むニューラルエピソードコントローラを訓練するための動作を実行させ、
前記ニューラルエピソードコントローラが、観測に応答して、エージェントによって実行される複数の行動の各行動のための個々のエピソード記憶モジュールを含むエピソード記憶データを維持し、
各行動のための前記エピソード記憶モジュールが、個々の複数のキー埋込みのそれぞれを個々のリターン推定にマッピングし、
前記動作が、
1つまたは複数のコンピュータによって、訓練タプルを再生記憶からサンプリングするステップであって、
前記訓練タプルが、訓練の観測、訓練の選択済み行動、および訓練のリターンを含む、ステップと、
前記訓練の観測のための訓練キー埋込みを生成するために、前記埋込みニューラルネットワークのパラメータの現在の値に応じて前記埋込みニューラルネットワークを使用して、前記訓練の観測を処理するステップと、
前記訓練の選択済み行動のための前記エピソード記憶モジュール内の前記キー埋込みのいずれかに前記訓練の観測に関連付けられた前記訓練キー埋込みがマッチするかどうかを判断するステップと、
前記訓練の選択済み行動のための前記エピソード記憶モジュール内の前記キー埋込みのいずれかに前記訓練キー埋込みがマッチするとき、(i)前記訓練のリターン、(ii)前記マッチするキー埋込みによって現在マッピングされる前記リターン推定、および(iii)エピソード記憶学習率に基づいて計算される新しいリターン推定に、前記マッチするキー埋込みをマッピングするために前記エピソード記憶モジュールを更新するステップと、
前記1つまたは複数のコンピュータによって、前記訓練の観測から前記訓練の選択済み行動のためのQ値を判断するステップであって、
前記選択済み行動のための前記Q値は、前記訓練の観測に応答して前記訓練の選択済み行動を行う前記エージェントから生じることになる予測されるリターンである、ステップと、
前記1つまたは複数のコンピュータによって、前記エピソード記憶学習率より小さい埋込みニューラルネットワーク学習率を使用して、前記埋込みニューラルネットワークのパラメータの現在の値を更新するために、前記訓練の選択済み行動に関する前記Q値と前記

訓練のリターンとの間のエラーの勾配を逆伝搬させるステップと
を含む1つまたは複数のコンピュータ可読記録媒体。
【請求項１１】
前記訓練の選択済み行動に関する前記Q値を判断するステップのための前記動作が、
距離測定に応じて、前記訓練キー埋込みに対する、前記訓練の選択済み行動のための前記エピソード記憶データ内のp近傍キー埋込みを判断するステップと、
前記訓練の選択済み行動のための前記エピソード記憶データ内の前記p近傍キー埋込みによってマッピングされる前記リターン推定から、前記訓練の選択済み行動に関するQ値を判断するステップと
を含む、請求項10に記載の1つまたは複数のコンピュータ可読記録媒体。
【請求項１２】
前記訓練の選択済み行動のための前記エピソード記憶データ内の前記p近傍キー埋込みによってマッピングされる前記リターン推定から、前記訓練の選択済み行動に関する前記Q値を判断するステップのための前記動作が、
前記距離測定に応じて、前記p近傍キー埋込みと、前記訓練キー埋込みの間の距離から、前記訓練の選択済み行動のための前記エピソード記憶データ内の前記p近傍キー埋込みのそれぞれに対する個々の重みを判断するステップと、
前記訓練の選択済み行動のための前記エピソード記憶データ内の前記p近傍キー埋込みのそれぞれに対して、個々の重み付きリターン推定を判断するために、前記訓練キー埋込みに対する前記重みによって、前記訓練キー埋込みにマッピングされた前記リターン推定に重み付けするステップと
を含む、請求項11に記載の1つまたは複数のコンピュータ可読記録媒体。
【請求項１３】
前記訓練の選択済み行動に関する前記Q値を判断するステップのための前記動作が、
前記訓練の選択済み行動に対する前記重み付きリターン推定を合計するステップと、
前記合計された重み付きリターン推定を前記Q値として使用するステップと
を含む、請求項12に記載の1つまたは複数のコンピュータ可読記録媒体。
【請求項１４】
前記訓練の選択済み行動に関する前記Q値を判断するステップのための前記動作が、
前記訓練の選択済み行動に対する前記重み付きリターン推定を合計するステップと、
前記Q値を生成するために、リターンニューラルネットワークのパラメータの現在の値に応じて、前記リターンニューラルネットワークを通じて、前記合計された重み付きリターン推定を含むネットワーク入力を処理するステップと
を含む、請求項12に記載の1つまたは複数のコンピュータ可読記録媒体。
【請求項１５】
前記訓練の選択済み行動に関する前記Q値を判断するステップのための前記動作が、
前記Q値を生成するために、リターンニューラルネットワークのパラメータの現在の値に応じて、前記リターンニューラルネットワークを通じて、前記重み付きリターン推定を含むネットワーク入力を処理するステップを含む、請求項12に記載の1つまたは複数のコンピュータ可読記録媒体。
【請求項１６】
前記埋込みニューラルネットワークが、畳み込みニューラルネットワークである、請求項10に記載の1つまたは複数のコンピュータ可読記録媒体。
【請求項１７】
前記埋込みニューラルネットワークが、1つまたは複数の回帰型ニューラルネットワーク層を含む、請求項10に記載の1つまたは複数のコンピュータ可読記録媒体。
【請求項１８】
前記動作が、前記訓練の選択済み行動のための前記エピソード記憶データ内の前記キー埋込みのいずれにも前記訓練キー埋込みがマッチしないとき、訓練リターンに前記現在のキー埋込みをマッピングするデータを前記訓練の選択済み行動のための前記エピソード記

憶モジュールに追加するステップ
をさらに含む、請求項10に記載の1つまたは複数のコンピュータ可読記録媒体。
【請求項１９】
1つまたは複数のコンピュータと、
命令を記憶した1つまたは複数のコンピュータ可読記録媒体と
を含むシステムであって、
前記命令は、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、埋込みニューラルネットワークを含むニューラルエピソードコントローラを訓練するための動作を実行させ、
前記ニューラルエピソードコントローラが、観測に応答して、エージェントによって実行される複数の行動の各行動のための個々のエピソード記憶モジュールを含むエピソード記憶データを維持し、
各行動のための前記エピソード記憶モジュールが、個々の複数のキー埋込みのそれぞれを個々のリターン推定にマッピングし、
前記動作が、
1つまたは複数のコンピュータによって、訓練タプルを再生記憶からサンプリングするステップであって、
前記訓練タプルが、訓練の観測、訓練の選択済み行動、および訓練のリターンを含む、ステップと、
前記訓練の観測のための訓練キー埋込みを生成するために、前記埋込みニューラルネットワークのパラメータの現在の値に応じて前記埋込みニューラルネットワークを使用して、前記訓練の観測を処理するステップと、
前記訓練の選択済み行動のための前記エピソード記憶モジュール内の前記キー埋込みのいずれかに前記訓練の観測に関連付けられた前記訓練キー埋込みがマッチするかどうかを判断するステップと、
前記訓練の選択済み行動のための前記エピソード記憶モジュール内の前記キー埋込みのいずれかに前記訓練キー埋込みがマッチするとき、(i)前記訓練のリターン、(ii)前記マッチするキー埋込みによって現在マッピングされる前記リターン推定、および(iii)エピソード記憶学習率に基づいて計算される新しいリターン推定に、前記マッチするキー埋込みをマッピングするために前記エピソード記憶モジュールを更新するステップと、
前記1つまたは複数のコンピュータによって、前記訓練の観測から前記訓練の選択済み行動のためのQ値を判断するステップであって、
前記選択済み行動のための前記Q値は、前記訓練の観測に応答して前記訓練の選択済み行動を行う前記エージェントから生じることになる予測されるリターンである、ステップと、
前記1つまたは複数のコンピュータによって、前記エピソード記憶学習率より小さい埋込みニューラルネットワーク学習率を使用して、前記埋込みニューラルネットワークのパラメータの現在の値を更新するために、前記訓練の選択済み行動に関する前記Q値と前記訓練のリターンとの間のエラーの勾配を逆伝搬させるステップと
を含むシステム。
【請求項２０】
前記動作が、前記訓練の選択済み行動のための前記エピソード記憶データ内の前記キー埋込みのいずれにも前記訓練キー埋込みがマッチしないとき、訓練リターンに前記現在のキー埋込みをマッピングするデータを前記訓練の選択済み行動のための前記エピソード記憶モジュールに追加するステップ
をさらに含む、請求項19に記載のシステム。

フロントページの続き

(72)発明者　アレクサンダー・プリッツェル
イギリス・Ｎ１Ｃ・４ＡＧ・ロンドン・パンクラス・スクエア・６
(72)発明者　チャールズ・ブランデル
イギリス・Ｎ１Ｃ・４ＡＧ・ロンドン・パンクラス・スクエア・６
(72)発明者　アドリア・ピュイグドメネク・バディア
イギリス・Ｎ１Ｃ・４ＡＧ・ロンドン・パンクラス・スクエア・６

【外国語明細書】
2021064387000001.pdf