AI 特許紹介(56)

AI 特許を学ぶ！究める！
～マニピュレータ非依存表現（MIR)特許～

2023 年 9 月 8 日

河野特許事務所
所長弁理士　河野英仁

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

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

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

1.概要

特許出願人　DeepMind Technologies
出願日　2021 年 10 月 1 日
公開日　2022 年 4 月 7 日
公開番号 WO2022069732
発明の名称　目標条件付きポリシーを使用したクロスドメイン模倣学習

732 特許は、ドメインの異なるデモンストレーションの観測とエージェントの観測とをマニピュレータ非依存表現(MIR: Manipulator-Independent Representations)に埋め込み、同一の MIR 空間上で、デモンストレーション画像を用いた模倣学習を行う技術に関する。

2.特許内容の説明

模倣学習は、強化学習（RL）の報酬を指定することが不可能な場合、または探索問題が特に難しい場合のロボット学習タスクに効果的なツールである。行動クローニングまたは逆強化学習等の模倣学習は、一人称のアクション状態の軌跡のコレクションからポ

リシーを導き出す。

しかしながら、これは、人間や他の動物が模倣する方法とは正反対である。人間は、たとえ他の種の行動であっても、その行動を観察し、その行動が環境の状態に及ぼす影響を理解し、同様の結果を達成するために自分の体がどのような行動を実行できるかを理解する。

732 特許はドメインの異なるデモンストレーションの観測とエージェントの観測とを MIR に埋め込み、同一の MIR 空間上で、デモンストレーション画像を用いた模倣学習を行う。

下記図は、ニューラルネットワークトレーニングシステム 100 を示す。

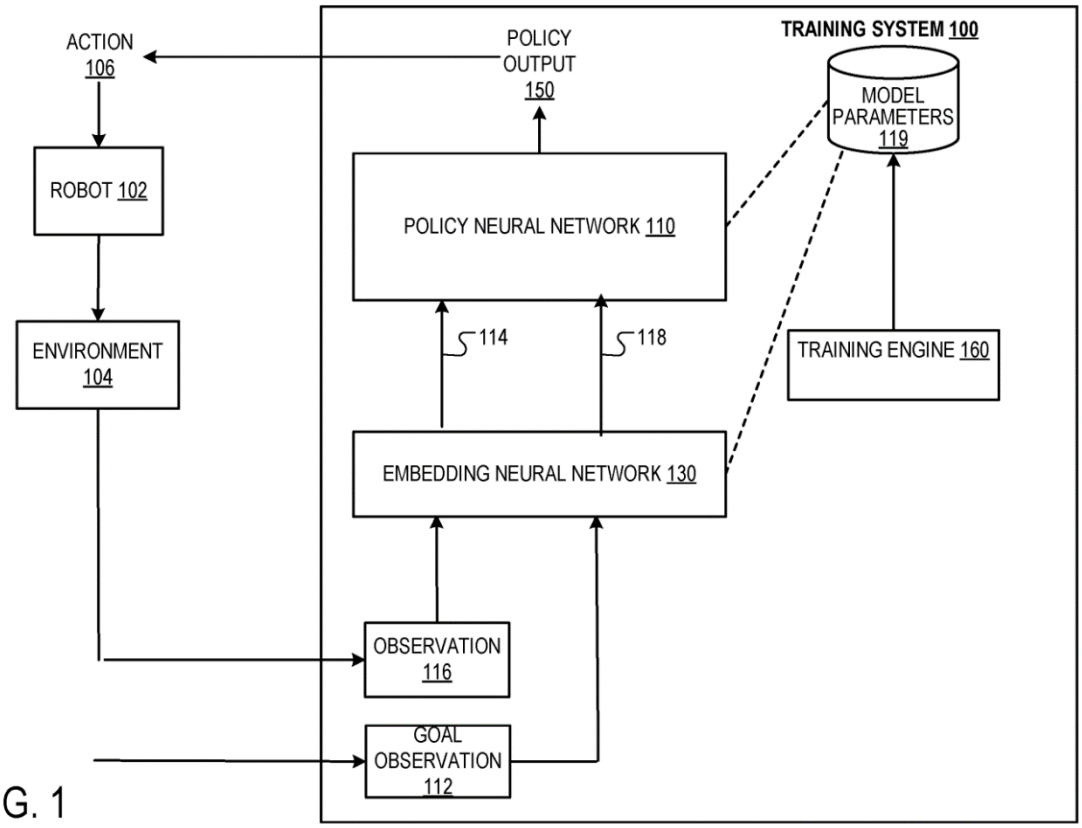


FIG. 1

ニューラルネットワークトレーニングシステム 100 は、ロボット 102 が環境 104 と対話している間にロボット 102 によって実行されるアクション 106 を選択するために、ロボット 102 に環境 104 内で特定のタスクを実行させるべく、模倣学習を通じて使用されるポリシーニューラルネットワーク 110 を訓練する。

ポリシーニューラルネットワーク 110 は、ポリシー入力を受信し、そのポリシー入力を処理して、ロボット 102 によって実行されるアクションを定義するポリシー出力 150 を生成する。

任意の時点でロボット 102 によって実行されるタスクは、環境 106 のゴール状態を特徴付けるゴール観察 112 によって指定される。例えば、ゴール観察 112 は、環境 104 がゴール状態にあるときの環境 104 の画像である。

タスクには、ロボット 102 を環境 104 内の異なる位置にナビゲートさせること、ロボットに異なるオブジェクトの位置を特定させること、ロボットに異なる物体を拾わせたり、異なる物体を 1 つ以上の指定された場所に移動させたりすることが含まれる。

特に、任意の所与のタイムステップでロボット 102 によって実行されるアクション 106 を選択するために、システム 100 は、所与のタイムステップで環境 104 が置かれている現在の状態を特徴付ける現在の観察 116 を受け取る。観察 116 は、画像生成ユニット（例えば、カメラおよび／または Lidar センサ）によってキャプチャされる。

現在の観測値 116 は、ゴール観測値 112 とは異なる視点からのものである。例えば、現在の観察 116 は、環境の 1 つ以上の一人称の自己中心画像、すなわちロボットの 1 つ以上のカメラ（または他の画像生成ユニット）によってキャプチャされた画像である。ゴール観察 112 は、環境がゴール状態にあるときのエージェント、例えばロボットまたはデモンストレーションエージェントの 1 つまたは複数の三人称画像である。

システム 100 は、現在の観測 116 の埋め込み 114 とゴール観測 112 の埋め込み 118 を生成する。埋め込みは、数値、例えばベクトルの順序付けされた集合であり、一般に対応する観察よりも次元が低い。システム 100 は、埋め込みニューラルネットワーク 130 を使用して対応する観察を処理することによって埋め込みを生成する。

すなわち、システム 100 は、埋め込みニューラルネットワーク 130 を使用して現在の観測 116 を処理して現在の観測 116 の埋め込み 114 を生成し、埋め込みニューラルネットワーク 130 を使用してゴール観測 112 を処理し、ゴール観測 112 の埋め込み 118 を生成する。

埋め込みニューラルネットワーク 130 は、観察を埋め込みにマッピングできるようにする任意の適切なアーキテクチャを有することができる。 例えば、ニューラルネットワーク 130 は畳み込みニューラルネットワークである。

システム 100 は、(i)環境 104 が所与のタイムステップにおいてある現在の状態を特徴付ける現在の観測 116 の埋め込み 115 と、(ii) ポリシーニューラルネットワーク 110

を使用してゴールを特徴付けるゴール観測 112 の埋め込み 118 とを含むポリシー入力を処理し、現在の観察 116 に応答してロボット 102 によって実行されるアクション 106 を定義するポリシー出力 150 を生成する。

したがって、任意の所与のタイムステップにおいて、ポリシーニューラルネットワーク 110 は、そのタイムステップにおける現在の状態を特徴付ける現在の観測 116 だけでなく、ゴール状態を特徴付けるゴール観測 112 にも条件付けされる。
したがって、ポリシーニューラルネットワーク 110 は、「ゴール条件付きポリシーニューラルネットワーク」と呼ぶこともできる。

埋め込みニューラルネットワーク 130 は、ニューラルネットワーク 130 が 2 つの埋め込みをポリシー出力にマッピングできるようにする任意の適切なアーキテクチャを有する。

次に、システム 100 は、ポリシー出力 150 を使用して、現在の観察 116 に応答してロボット 102 によって実行されるアクション 106 を選択する。ポリシー出力 150 は、アクションのセット内の各アクションのそれぞれのQ値を含む。システム 100 は、Q値を処理して、アクションを選択するために使用できる各アクションのそれぞれの確率値を生成することができ、または最も高いQ値を有するアクションを選択することができる。

アクションの Q 値は、現在の観察に応じてエージェントがアクションを実行し、その後ポリシーニューラルネットワークパラメーターの現在の値に従ってエージェントによって実行される将来のアクションを選択することによって得られる「リターン」の推定値である。

リターンとは、エージェントが受け取った「報酬」の累積的な尺度、たとえば、時間割引された報酬の合計を指す。トレーニング中、システム 100 は、各タイムステップでそれぞれの報酬を生成することができ、報酬はスカラー数値によって指定され、例えば、割り当てられたタスクの完了に向けたエージェントの進捗を特徴付ける。

システム 100 がポリシーニューラルネットワーク 110 を訓練する方法のため、ポリシー出力 150 によって定義されるアクション 106 は、ロボット 102 を、ポリシー入力によって表されるゴール観察 112 によって指定されるゴールの達成(つまりタスク完了)に近づけるアクションである。

特に、システム 100 は、トレーニングデータに基づいてポリシーニューラルネットワーク 110、および埋め込みニューラルネットワーク 130 をトレーニングするトレーニングエンジン 160 を含む。換言すれば、トレーニングエンジン 160 は、ポリシーニューラルネットワーク 110 および埋め込みニューラルネットワーク 130 を訓練して、ポリシーニューラルネットワーク 110 およびニューラルネットワーク 130 のモデルパラメータ 119 の訓練された値を決定する。

すなわち、埋め込みニューラルネットワーク 130 は、例えば、異なるポリシーニューラルネットワークと共同して、または 1 つ以上の教師なし学習タスクに関して、別のシステムによって事前訓練され、その後、埋め込みニューラルネットワーク 130 のモデルパラメータ 119 が固定される一方、トレーニングエンジン 160 はポリシーニューラルネットワーク 110 をトレーニングする。

下記図は、実証観察シーケンス上でポリシーニューラルネットワークを訓練する際に使用するための軌道の生成を示す。

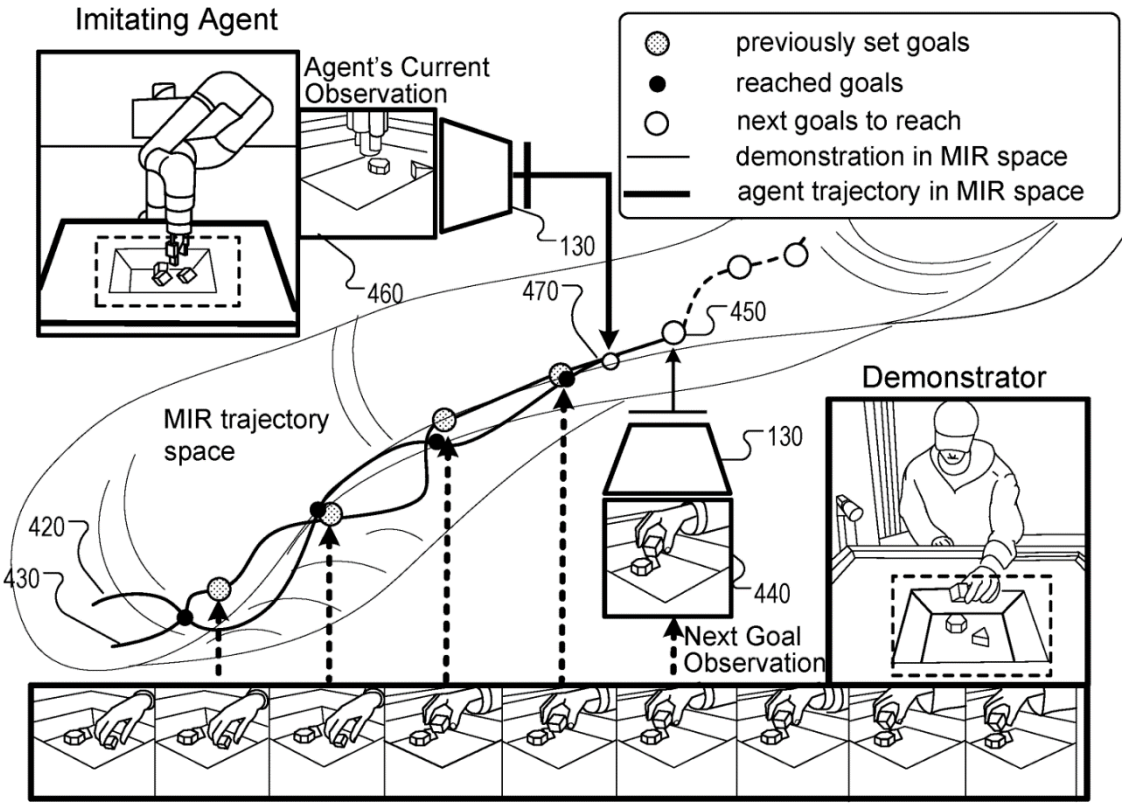


FIG. 4 410

図に示すように、システムは、デモンストレーションシーケンスからゴールデモンス

トレーションシーケンス 410 を生成しており、上述したように、ゴールデモンストレーションシーケンス 410 内のゴールデモンストレーションのそれぞれについてそれぞれの軌道 420 を生成するようにエージェントを制御する過程にある。

特に、図の例は、「MIR 軌道空間」、すなわち、埋め込みニューラルネットワークによって生成された埋め込みの埋め込み空間を通るそれぞれの経路として、デモンストレーションシーケンスによって定義された軌道 420 および元の軌道 430 を示す。図に示すようにシステムは、現在、ゴールデモンストレーション観測 440 の軌道を生成しており、したがって、軌道生成中の各タイムステップで、現在のゴールの埋め込みとして、埋め込みニューラルネットワーク 130 によって生成されるゴールデモンストレーション観測 440 の埋め込み 450 をポリシーニューラルネットワーク 110 に条件付けしている。

現在の時間状態において、システムはまた、エージェントのカメラセンサによってキャプチャされた時間ステップにおける環境の現在の状態の現在の観察 460 の埋め込みニューラルネットワーク 130 によって生成された埋め込み 470 を、ポリシーニューラルネットワーク 110 への入力として提供する。

図の例から分かるように、ゴールデモンストレーション観察シーケンス 410 におけるゴールデモンストレーション観察は、「デモンストレータ（デモンストレーションエージェント)」の三人称視点からキャプチャされた画像である。一方、現在の観察 460 は、エージェントに対する環境の一人称の自己中心的な視点からキャプチャされる。

したがって、各タイムステップで、ポリシーニューラルネットワーク 110 は、(i)一人称観察の埋め込み、および(ii)異なるエージェントの三人称観察の埋め込みを入力として受け取る。それにもかかわらず、本技術を使用することによって、ポリシーニューラルネットワーク 110 をそのようなデータに基づいて効果的に訓練することができる。

3.クレーム

732 特許のクレーム 1 は以下の通りである。

1.1 台以上のコンピュータによって実行される方法において、

複数のデモンストレーションシーケンスを含むデモンストレーションデータを取得し、各デモンストレーションシーケンスは、デモンストレーションエージェントが環境と対話する間に、環境の状態を特徴付ける複数のデモンストレーション観察を含み、

強化学習を通じてデモンストレーションデータ上でゴール条件付きポリシーニュー

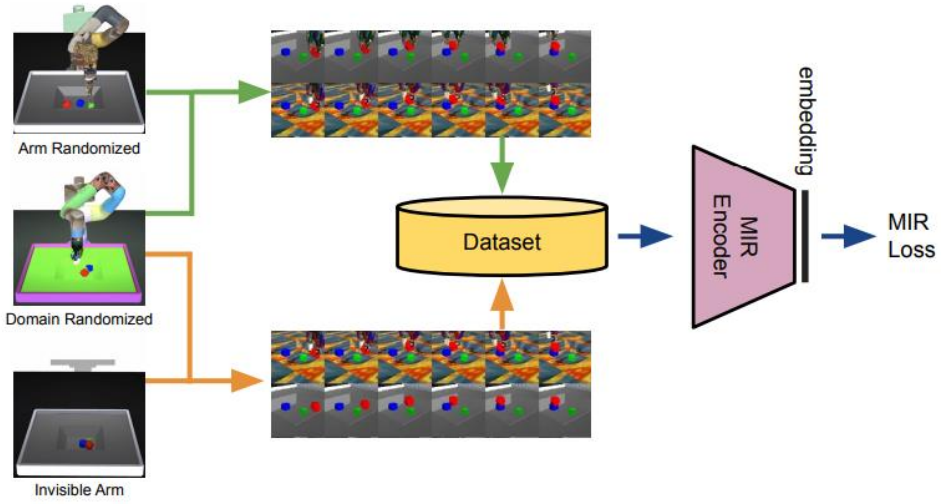
ラルネットワークをトレーニングし、前記ゴール条件付きポリシーニューラルネットワークは以下のように構成され、
環境の現在の状態を特徴付ける現在の観測値の埋め込みと、環境のゴール状態を特徴付けるゴール観測値の埋め込みとを含むポリシー入力を受信し、
現在の観察に応じてエージェントによって実行されるアクションを定義するポリシー出力を生成するために、ポリシーパラメータに従ってポリシー入力を処理し、
ここで、前記トレーニングは、複数のデモンストレーションシーケンスのそれぞれについて、以下を含む：

デモンストレーションシーケンス内のデモンストレーション観測の適切なサブセットをゴールデモンストレーション観測として選択することによって、一連のゴールデモンストレーション観測を生成し、
各ゴールデモンストレーションの観察では、一連のゴールデモンストレーションの最初のゴールデモンストレーションから始まり、一連のゴールデモンストレーションの最後のゴールデモンストレーションまで継続し、
ゴール条件付きポリシーニューラルネットワークによって生成されたポリシー出力を使用してエージェントを制御することにより、ゴールデモンストレーション観察のためのトレーニング観察の軌道を生成し、ゴール条件付きポリシーニューラルネットワークは、それぞれにゴールデモンストレーション観察の埋め込みを含むポリシー入力に条件付けされ、
トレーニング観察の埋め込みとゴールデモンストレーション観察の埋め込みとの間の類似性に基づいて、トレーニング観察のそれぞれに対してそれぞれの報酬を生成し、
強化学習を通じて、軌跡内のトレーニング観察に対するそれぞれの報酬に基づいて、ゴール条件付きポリシーニューラルネットワークをトレーニングする。

4. 本特許に関連する論文

本特許に関する論文“Manipulator-Independent Representations for Visual Imitation”[1]が、DeepMind の Yuxiang Zhou 氏らにより公表されている。下記図は、マニピュレータ非依存表現（MIR）空間の学習処理を示すブロック図である。

[1] Yuxiang Zhou et al. “Manipulator-Independent Representations for Visual Imitation” arXiv:2103.09016v2 [cs.RO] 18 Mar 2021



MIR は、(a) ドメインランダム化および「非表示アーム」環境、(b) ドメインランダム化およびアームのみランダム化環境の 2 つの環境ペアを使用して生成されたデータセットでトレーニングされる。下記テーブルは横断的模倣の表現方法の定量的比較を示す。

Method	Lifting Success					Stacking Success				
	invisible	**jaco**	**robot**	**stick**	**hand**	**invisible**	**jaco**	**robot**	**stick**	**hand**
TDC [2]	67%	38%	10%	21%	31%	0%	2%	0%	6%	0%
GCP	**100%**	50%	39%	13%	10%	0%	0%	0%	0%	0%
TCN [31]	79%	85%	27%	12%	0%	0%	0%	0%	0%	0%
CMC [2]	80%	**100%**	80%	26%	0%	20%	66%	0%	0%	0%
MIR (ours)	**100%**	**100%**	**100%**	**44%**	**50%**	**38%**	**81%**	**29%**	**17%**	**11%**

評価は、インビジブルアーム、ジャコハンド、本物のロボット、ピックアップスティック、人間の手の 5 つの領域で実証された 10 個のスタッキング軌道のそれぞれについて 100 回の試行で実行される。測定基準は、持ち上げたり積み上げたりする動作を模倣した成功率である。いずれのケースにおいても MIR が高い成功率を示している。

以上

著者紹介

河野英仁

河野特許事務所、所長弁理士。立命館大学情報システム学博士前期課程修了、米国フランクリンピアースローセンター知的財産権法修士修了、中国清華大学法学院知的財産夏季セミナー修了、MIT(マサチューセッツ工科大学)コンピュータ科学・AI 研究所 AI コ

ース修了。

AI 特許コンサルティング、医療 AI 特許コンサルティングの他、米国・中国特許の権利化・侵害訴訟を専門としている。著書に「世界のソフトウエア特許(共著)」、「FinTech 特許入門」、「AI/IoT 特許入門 3」、「ブロックチェーン 3.0(共著)」がある。