AI 特許紹介(34)
AI 特許を学ぶ！究める！
～メタ学習特許～

2021 年 11 月 10 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　Google
出願日　2020 年 1 月 23 日
公開日　2020 年 7 月 30 日
公開番号 WO2020154542
発明の名称　メタ模倣学習とメタ強化学習に基づくメタ学習を使用した新しいタスクへのロボット制御ポリシーの効率的な適応

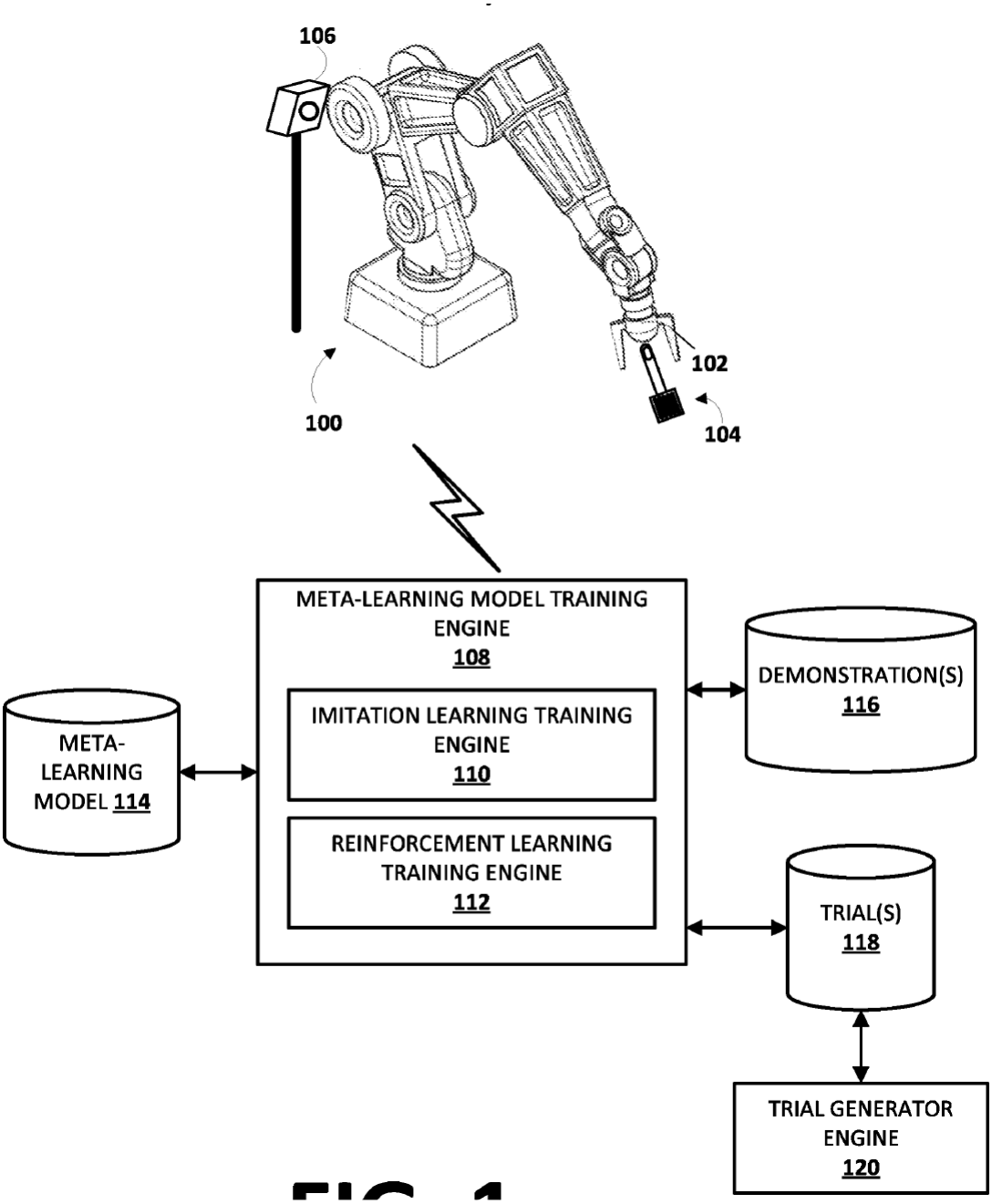
542 特許は、模倣学習の要素と試行錯誤の強化学習を組み込んだ新しいメタ学習アルゴリズムに関する。

2.特許内容の説明
模倣学習により、エージェントはデモンストレーションから複雑な行動を学ぶことができる。しかしながら、複雑なビジョンベースのタスクを学習するには、非現実的な数のデモンストレーションが必要となる。

メタ模倣学習は、エージェントが同様のタスクの学習からの経験を活用することにより、複数のデモンストレーションから新しいタスクを学習できる。しかしながら、タスクのあいまいさや観察されないダイナミクスが存在する場合、デモンストレーションだけでは十分な情報が得られない。

そこで、本発明では、報酬フィードバックを使用して、デモンストレーションと試行錯誤の両方の経験から学習できるメタ学習アルゴリズムを提供している。

図１は本発明の実装例を示す。



ロボット 100 は、複数の潜在的な経路のいずれかに沿って把持エンドエフェクタ 102 を所望の位置に配置する。ロボット 100 はさらに、その把持エンドエフェクタ 102 の 2 つの対向する「爪」を制御して開位置と閉位置との間で爪を作動させる。

図 1 では、ビジョンコンポーネント 106 は、ロボット 100 のベースまたは他の静止基準点に対して固定された姿勢で取り付けられている。オブジェクト 104 は、へら、ホッチキス、および鉛筆を含む。

ロボット 100 からのデータ（例えば、状態データ）は、メタ学習モデルトレーニングエンジン 108 を使用してメタ学習モデル 114 をトレーニングするために利用される。例えば、メタ学習モデルトレーニングエンジン 108 は、メタ学習を使用して、メタ学習モデル 114 の試行ポリシーおよび適合試行ポリシーをトレーニングする。

メタ学習モデルトレーニングエンジン 108 は、模倣学習トレーニングエンジン 110、及び、強化学習トレーニングエンジン 112 を含む。模倣学習トレーニングエンジン 110 は、人間がガイドするデモンストレーション 116 を使用してメタ学習モデル 114 をトレーニングする。

例えば、メタ学習モデル 114 の試行ポリシーは、模倣学習エンジン 110 によって、模倣学習を使用して訓練することができる。強化学習訓練エンジン 112 は、タスクを実行するロボット 100 の試行 118 に基づいてメタ学習モデル 114 を訓練する。試行ジェネレータエンジン 120 は、メタ学習モデル 114 の試行ポリシーを使用して、試行 118 を生成する。

いくつかの実装形態では、試行ポリシーと適合試行ポリシーが単一のモデルに統合され、試行ポリシーを更新すると、適合試行ポリシーも変更される。同様に、適合試行ポリシーを変更すると、試行ポリシーが変更される。

いくつかのそのような実装形態では、試行ジェネレータエンジン 120 は、試行ポリシーが時間とともに変化するときに、試行ポリシーに基づいて試行を継続的に生成する。次に、これらの試行を強化学習トレーニングエンジン 112 が使用して、メタ学習モデル 114 の適合ポリシーをトレーニングし、これにより、メタ学習モデル 114 の試行ポリシーも更新される。

次に、この更新された試行ポリシーは、試行ジェネレータエンジン 120 が追加の試行を生成する際に使用することができる。これらの追加の試行は、その後、強化学習トレ

ーニングエンジン 112 によって使用され、適合試行ポリシーを更新することができ、これにより、試行ポリシーが更新される。

適応試行ポリシーのトレーニングに基づいて試行ポリシーを更新し、更新された試行ポリシーを使用して追加の試行を生成し、追加の試行を使用して適応試行ポリシーを更新するというこのサイクルは、メタ学習モデルのトレーニングが完了するまで繰り返される。

さらに、強化学習トレーニングエンジン 112 は、適合試行ポリシーをトレーニングするときに、ロボットが試行 118 において新しいタスクを正常に完了したかどうかを示す報酬を利用することができる。この報酬はスパースな(まばらな)報酬信号である。たとえば、人間は、クライアントデバイスでのユーザーインターフェイス入力を介して、成功または失敗のバイナリ報酬表示を提供できる。これは、ロボットが試行錯誤の試行でタスクを正常に完了したかどうかを示す。

図 2 Aは、メタ学習モデル 200 を示す。

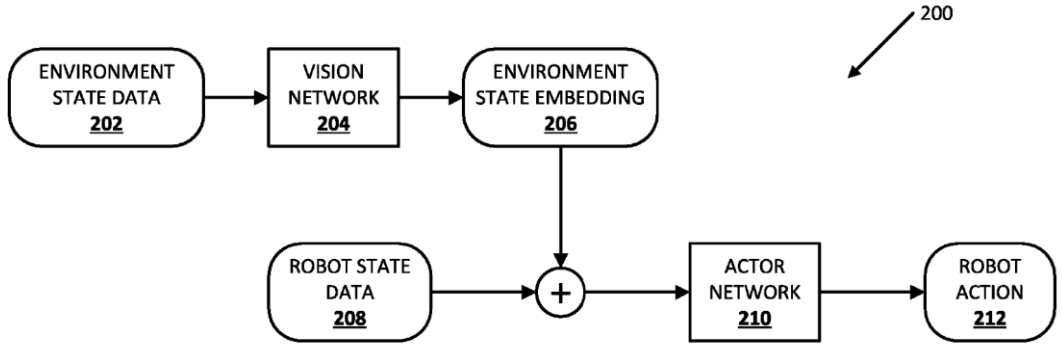


FIG. 2A

環境状態データ 202 は、ロボットの現在の環境に関する情報を収集する。環境状態データは、図 1 に示されるビジョンコンポーネント 106 などのビジョンコンポーネントを使用してキャプチャすることができる。環境状態データ 202 は、ビジョンネットワーク 204 を使用して処理され、環境状態埋め込み 206 を生成する。環境状態埋め込み 206 は、ロボットの環境の視覚的特徴を表す。

環境状態埋め込みは、ロボット状態データ 208 と組み合わせることができる。例えば、環境状態埋め込み 206 は、ロボット状態データ 208 と連結することができる。ロ

ボット状態データ 208 は、現在のエンドエフェクタポーズ、現在のエンドエフェクタ角度、現在のエンドエフェクタ速度、またはロボットの現在の位置およびロボットの１つまたは複数の構成要素に関する追加情報の表現を含む。

ロボットの状態データには、たとえば、エンドエフェクタの X、Y、Z の位置、および エンドエフェクタの方向を示す６次元ポーズ等、タスク空間でのエンドエフェクタのポーズの表現を含む。

アクターネットワーク 210 は、環境およびロボットの現在の状態に基づいてロボットタスクを実行するための１つまたは複数のロボットアクション 212 を生成するために使用することができる訓練されたメタ学習モデルである。

図２Ｂは、メタ学習モデル 250 を示す。

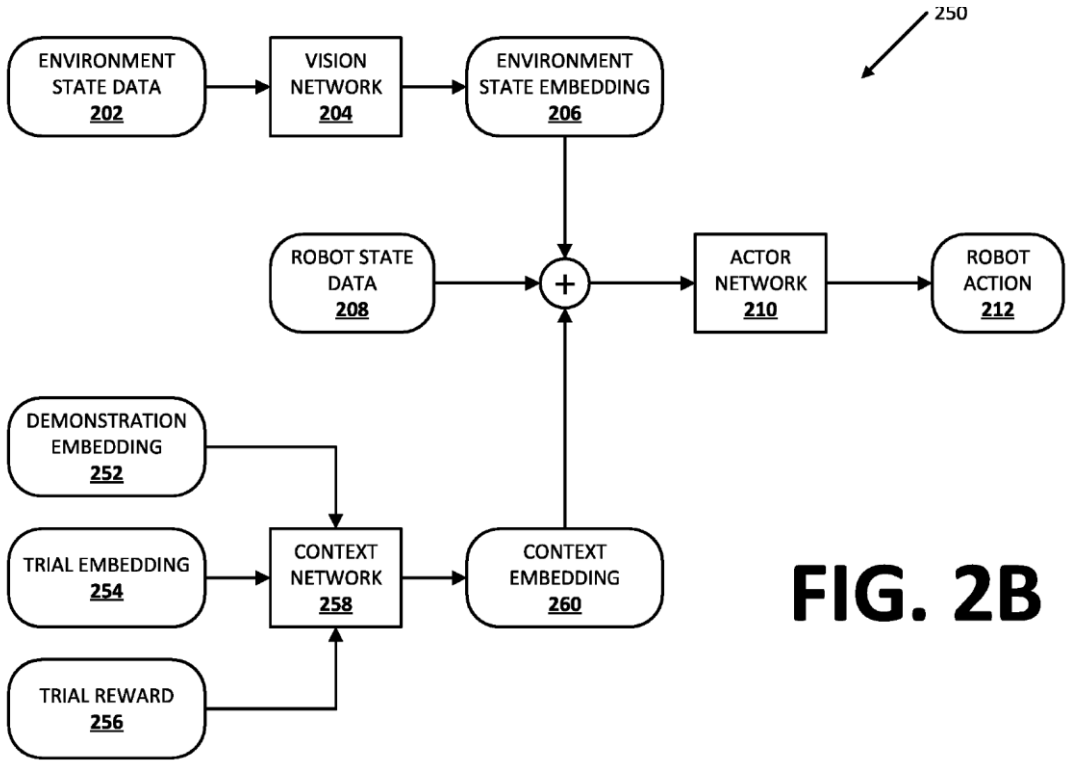


FIG. 2B

環境状態データ 202 は、ビジョンネットワーク 204 を使用して処理され、環境状態埋め込み 206 を生成する。環境状態埋め込み 206 は、ロボット状態データ 208 と組み合わせることができる。図示の例では、環境状態埋め込み 206 およびロボット状態データ 208 は、コンテキスト埋め込み 260 とさらに組み合わされる。例えば、環境状態埋め込み 206 は、ロボット状態データ 208 およびコンテキスト埋め込み 260 と連結することができる。

コンテキスト埋め込み 260 は、新しいタスクの表現を提供する。デモンストレーション埋め込み 252 は、タスクの人間によるガイド付きデモンストレーションの機能をキャプチャする。たとえば、デモンストレーションデータには、タスクを実行するロボットの人間によるガイド付きデモンストレーションのビデオを含む。

試行埋め込み 254 は、タスクを実行する試行錯誤の試みの特徴を取り込む。試行報酬 256 は、試行埋め込み 254 でキャプチャされた試行がタスクの実行に成功したかどうかを示す。

デモンストレーション埋め込み 252、試行埋め込み 254、試行報酬 256 を組み合わせることができる。例えば、デモンストレーション埋め込み 252 は、試行埋め込み 254 および試行報酬 256 と連結することができる。この組み合わせをコンテキストネットワーク 258 に提供して、コンテキスト埋め込み 260 を生成する。

環境状態埋め込み 206、ロボット状態データ 208、およびコンテキスト埋め込み 260 の組み合わせをアクターネットワーク 210 に提供して、ロボットタスクを実行するためのロボットアクション 212 を生成する。

図 3 は、複数のトレーニングタスクを使用してメタ学習モデルをトレーニングし、トレーニングされたメタ学習モデルを新しいタスクのためにトレーニングするプロセス 300 の例を示すフローチャートである。

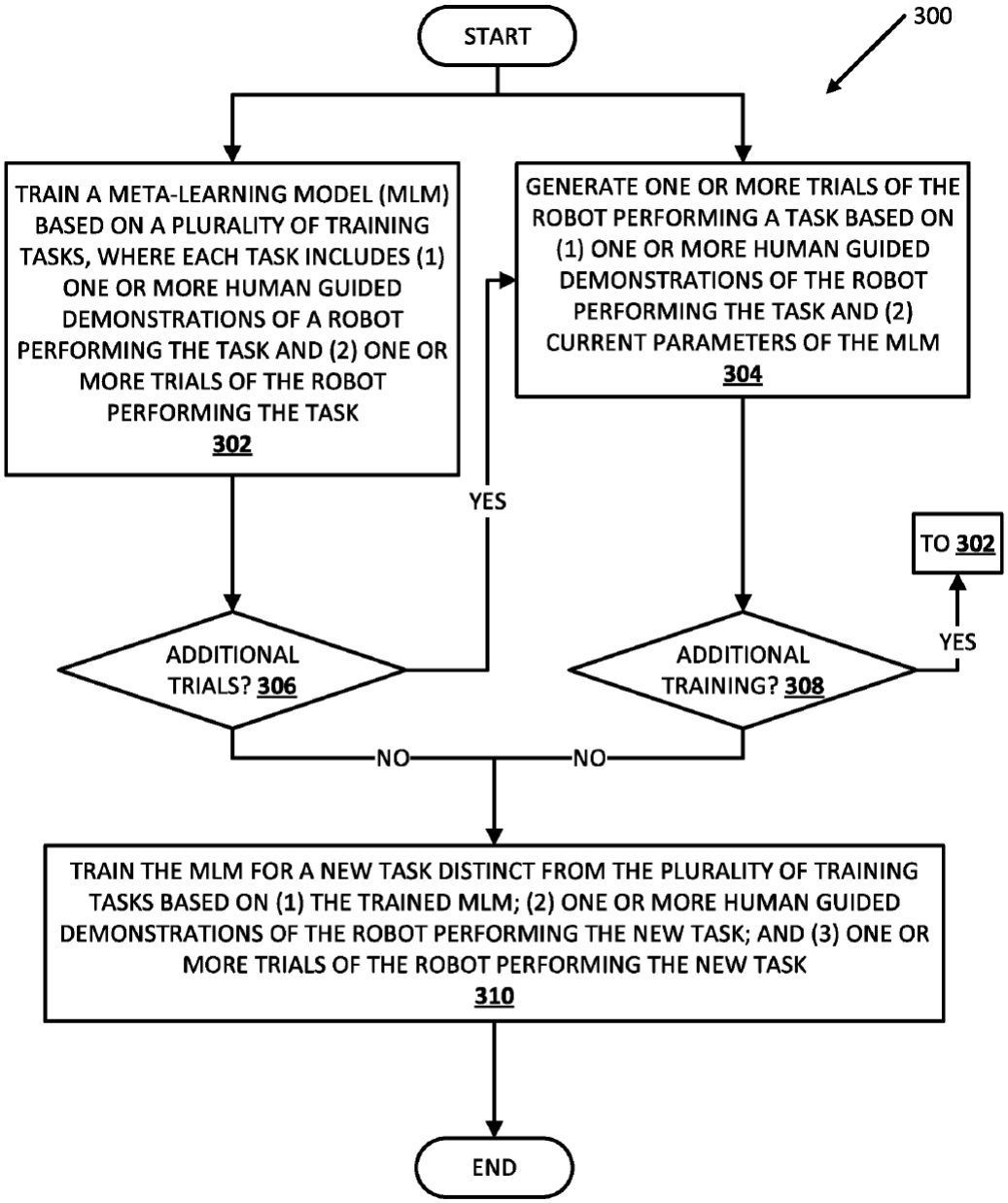


FIG. 3

ブロック 302 で、システムは、複数のトレーニングタスクに基づいてメタ学習モデルをトレーニングし、各タスクは、（1）ロボットの１つまたは複数の人間によるガイド付きデモンストレーション、および（2）タスクを実行するロボットの１つまたは複数の試行を含む。

ブロック 304 で、システムは、（1）タスクを実行するロボットの１つまたは複数の

人間によるガイド付きデモンストレーション、および（２）メタ学習モデルの試行ポリシーに基づいて、タスクを実行するロボットの１つまたは複数の試行を生成する。

たとえば、生成された試行には、一連のロボットアクションと対応するロボット状態が含まれ、ロボットはアクションを実行して、現在の状態から次の状態に遷移する。状態には、ロボットの現在の環境をキャプチャする環境状態データ（図 2A,2B の環境状態データ 202)、およびロボットのコンポーネントの現在の位置及びその他の機能をキャプチャするロボット状態データ(図 2A,2B のロボット状態データ 208)が含まれる。
メタ学習モデルの試行ポリシーは、ブロック 302 で生成される。

ブロック 306 で、システムは、タスクを実行するロボットの追加の試行を生成するかどうかを決定する。その場合、システムはブロック 304 に進み、メタ学習モデルの試行ポリシーに基づいて１つまたは複数の追加の試行を生成する。

そうでない場合、システムはブロック 310 に進む。例えば、システムは、システムがメタ学習モデルのトレーニングを完了したときに、追加の試行を生成しないことを決定する。

ブロック 308 で、システムは、メタ学習モデルの追加のトレーニングを実行するかどうかを決定する。そうである場合、システムはブロック 302 に進み、生成された試行のうちの１つまたは複数を使用して、メタ学習モデルの適合試行ポリシーを訓練し続ける。そうでない場合、システムはブロック 310 に進む。

ブロック 310 で、システムは、複数のトレーニングタスクとは異なる新しいタスクのためにトレーニングされたメタ学習モデルをトレーニングし（１）トレーニングされたメタ学習モデル、（２）新しいタスクを実行するロボットの１つまたは複数の人間によるガイド付きデモンストレーション、または（３）新しいタスクを実行するロボットの1回以上の試行に基づいて、トレーニングを実行する。

3.クレーム

542 特許のクレーム１は以下の通りである。

1. 一または複数のプロセッサにより実装される方法において、

新しいタスクを実行するロボットの人間によるガイド付きデモンストレーションに基づいて、新しいタスクを実行するロボットの制御に使用するための、トレーニングされたメタ学習モデルの適応ポリシーネットワークを生成し、前記メタ学習モデルは、複

数の異なるタスクを使用してトレーニングされ、新しいタスクについてはトレーニングされておらず、ここで、適応ポリシーネットワークの生成には以下が含まれ、

人間によるガイド付きデモンストレーションとメタラーニングモデルの試行ポリシーネットワークに基づいて、ポリシーネットワークの初期適応を生成し、

ポリシーネットワークの初期適応を使用して、ロボットアクションの初期シーケンスと、新しいタスクを実行するロボットの対応する状態を生成し、

ロボットにロボットアクションの初期シーケンスと対応するロボット状態を実行させ、

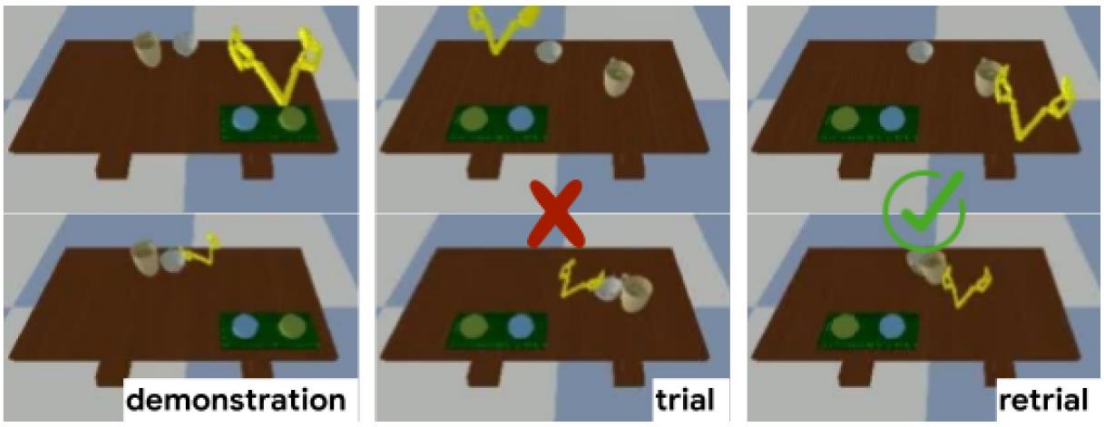
ロボットアクションの初期シーケンスと対応するロボット状態が新しいタスクを正常に完了したかどうかを判断し、

ロボットアクションの初期シーケンスと対応するロボット状態が新しいタスクを正常に完了したかどうかの判断に基づいて、適合ポリシーネットワークを生成する。

4. 本特許に関する論文

本特許に関連する論文[1]「WATCH, TRY, LEARN: META-LEARNING FROM DEMONSTRATIONS AND REWARDS」が Allan Zhou 氏らにより発表されている。

論文では、模倣学習の要素と試行錯誤の強化学習を組み込んだ新しいメタ学習アルゴリズムを解説している。



上記図の各列には、エピソードの最初と最後のフレームが上下に示されている。1つ

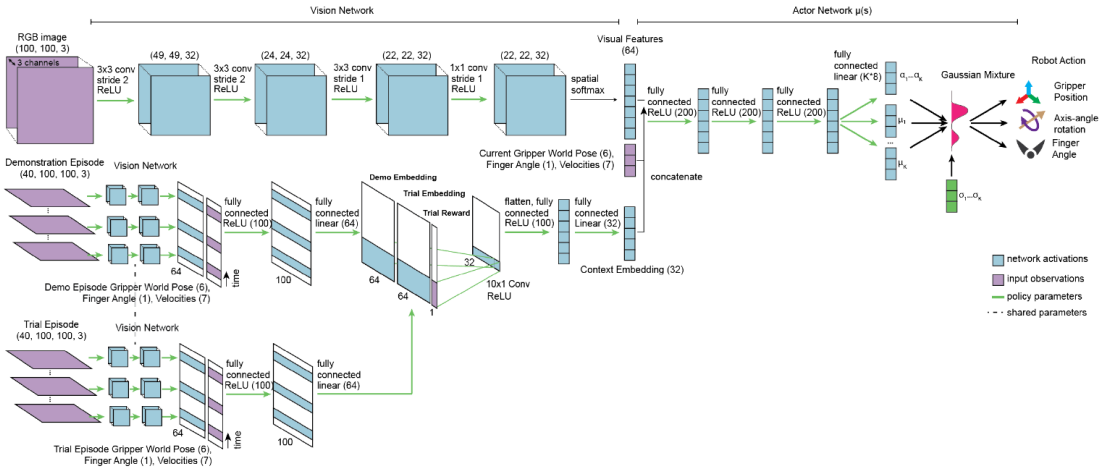
[1] Allan Zhou, Eric Jang, Daniel Kappler, Alex Herzog, Mohi Khansari, Paul Wohlart, Yunfei Bai & Mrinal Kalakrishnan, Sergey Levine, Chelsea Finn “WATCH, TRY, LEARN: META-LEARNING FROM DEMONSTRATIONS AND REWARDS” arXiv:1906.03352v4 [cs.LG] 30 Jan 2020

のデモンストレーション（左）を見た後、シーンが再配置される。1 つの試行エピソード（中央）で、本方法は、デモと試行錯誤の両方の経験を活用することによって、タスクの解決を学ぶことができる（右）。

下記 URL に動画が掲載されている。

https://sites.google.com/view/watch-try-learn-project

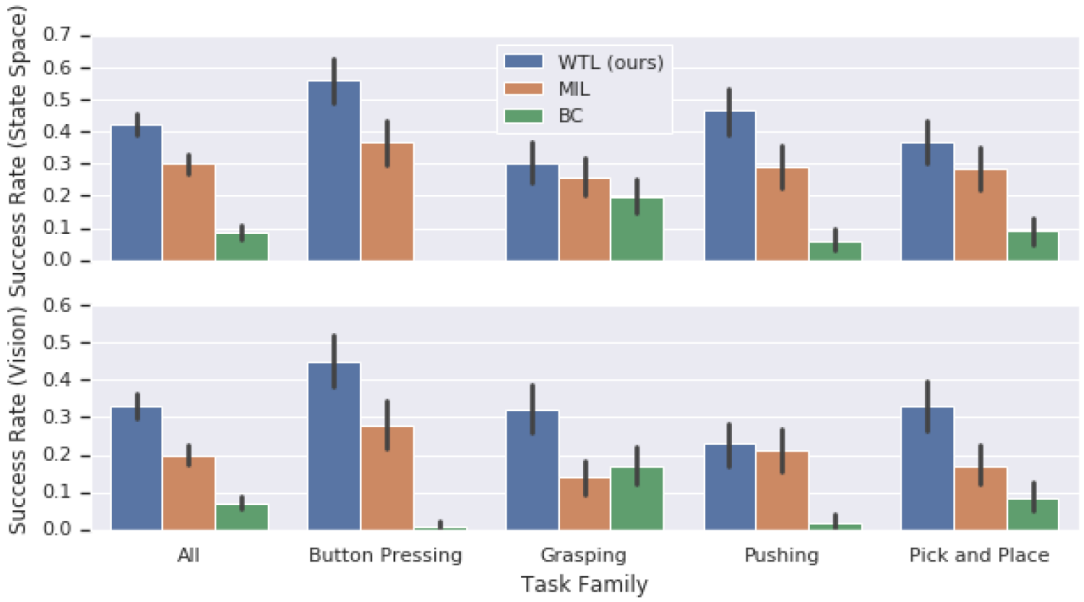
下記図はネットワーク構成図である。



左上図：各タイムステップの RGB 観測を、ReLU アクティベーションとレイヤー正規化を使用した 4 レイヤーCNN に渡し、その後に 2D キーポイントを抽出する空間ソフトマックスレイヤーに渡す。出力キーポイントをフラット化し、現在のグリッパーポーズ、グリッパー速度、およびコンテキスト埋め込みと連結する。

右上図: 結果のベクトルをアクターネットワークに渡す。アクターネットワークは、コマンドされたエンドエフェクタの位置、軸角度の方向、および指の角度でガウス混合のパラメータを予測する。

左下図: コンテキスト埋め込みを生成するために、埋め込みネットワークは、デモおよび試行の軌跡からランダムにサンプリングされた 40 の順序付けられた観測にビジョンネットワークを適用する。デモ及び試行の出力を、埋め込み機能ディメンションに沿った試行エピソードの報酬と連結してから、時間ディメンション全体に 10x1 の畳み込みを適用し、フラット化して MLP を適用し、最終的なコンテキスト埋め込みを生成する。



上記図は、状態空間（非ビジョン）とビジョンベースのポリシーの両方について、グリッパー制御環境での各種方法における平均成功率を示す。左端の列には、すべてのタスクファミリの集計結果を示している。本論文における Watch-Try-Learn（WTL）メソッドは、メタイミテーション（MIL）ベースライン及び動作クローニング（BC）ベースラインを大幅に上回っている。

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 2.0」、「ブロックチェーン 3.0(共著)」がある。