東京大学松尾研究室
DLHacks
“The Thousand Brains Project: A New Paradigm for Sensorimotor Intelligence”
（arXiv 2412.18354v1）

Kazuyuki Yoshikawa

Cortical Messaging Protocol (CMP)

- CMP はモジュール間（センサー→LM、LM 同士、LM→モーターなど）の通信仕様。
- メッセージ内容の主な要素：
- 「特徴（feature）」：色・曲率・形など、センサや LM が検出するもの。
- 「姿勢 / 位置 (pose)」：検出された特徴が身体あるいは“参照フレーム”内でどこにあるか、どの向きか。3D または 2D。arXiv
- 「発信者 ID」「信頼度（confidence）」など。arXiv
- この CMP によって、モジュール間で異なるセンサモダリティや異なる LM が共通の言語でやり取りでき、投票・仮説共有が可能になります。

Sensor Modules (SM)

- 各センサモジュールは、生のセンサー出力（カメラパッチ／深度マップなど）を取り、それを “feature + pose” の CMP コンプライアントな形式に変換します。
- 重要なのは、「センサパッチがどこを見ているか（sensor の姿勢）」「観測された特徴が body-centric な参照フレームでどこにあるか」を明示できること。
- 特徴の例: 色・曲率・表面の向き（normal）など。姿勢の定義には主に点の法線・主曲率方向などが使われます。

Learning Modules (LM)

- これは Monty の中心モジュールで、以下のような構造・動作を持ちます：
- **入力**：feature + pose（sensor module から）あるいは lower-level LM からの object ID + pose
- **内部記憶**：
  - **バッファ**（短期記憶）：現在のエピソードで観測した feature＋pose の一時的な集まり。arXiv
  - **グラフメモリ**（長期記憶）：以前学習したオブジェクトモデルを格納。オブジェクトモデルはグラフ構造で、ノード＝観測点／特徴点、エッジ＝観測点同士の相対的変位（displacement）。arXiv
- **仮説生成と証拠 (Hypothesis & Evidence)**：
  - エピソード開始時に「この観測がどのオブジェクト／どの姿勢（pose）か」という仮説空間を設定。すべての既存オブジェクトモデルの可能な pose を考慮。arXiv
  - 各観測で仮説に「証拠（evidence）」を追加し、より可能性の高い仮説を絞っていく。arXiv
- **モデル更新**：
  - 新しいオブジェクトを見つけたらグラフメモリに保存
  - 既存オブジェクトでも、新しい視点・特徴で観測があれば追加・改善
  - 類似観測（近くの位置／類似特徴）は重複しないように整理する工夫あり（効率化のため）arXiv
- **予測・問い合わせ (Prediction / Query)**：
  - グラフメモリを使って、「この動きをしたら次はどの特徴が見られるか」など予測（forward model）をする
  - ある目標の特徴がほしい→それに向かう動きを提案する（inverse model）機能あり

モジュールの組み合わせ：階層構造 + 投票 (Voting)

- 複数の LM を並列または階層的に配置できる。
- 階層構造：下位 LM は細かい部分（センサパッチ）をモデル化し、高位 LM は複合オブジェクトや大きなスケールの構造をモデル化。これにより複雑な物体／シーンを効率よく扱える。
- 投票（Voting）：並列 LM 間での lateral connection を通じて、それぞれの LM が仮定する物体 ID と姿勢を共有／調整し、一致する答え（コンセンサス）を早く得る。これにより認識を高速化し誤認を減らす。

アクションポリシー (Policies) とモータ系

- モデルベースとモデルフリー両方のポリシーを併用。
- モデルベース：LM によって得られた仮説や内部モデルを使って「どの動きをすれば観測の不確定性が減るか」「目的地に到達するにはどう動くか」を決定
- モデルフリー：頻繁な動作・反射的動き等を高速に行うためのパターンを学習する（例：センサを表面に滑らせる、特徴追従など）
- モータシステムは LM からの「目標状態(goal state)」を受け取り、それを実行するための実際の運動命令に変換する。LM は “目標としての pose” を CMP に則って出力する。

実験と評価

特徴	内容
環境	Habitat シミュレータを使った 3D 環境。単一オブジェクトが置かれた空間、あるいは複雑なオブジェクトデータセット（YCB 等）を使用。 arXiv
センサー構成	センサパッチ（小さなズームされたカメラ視野）、時に複数パッチ、また視界全体を把握する view-finder（ただし学習には通常使わない）など。エージェントはセンサを動かして複数視点を取得。 arXiv
タスク	主に物体の識別（ID）とその姿勢（pose）の推定。新しい角度から見たときや部分的な観測だけからも認識できるかどうかを評価。 arXiv
評価法	エピソード／ステップ単位で観測と行動を繰り返し、「認識できるまで」「最大ステップ数を超えるまで」などのターミナル条件を設定。評価フェーズではモデル固定で推論のみ行う。 arXiv

成果としては：

- Monty は比較的少ない視点（観測）でも物体を識別できる仮説を生成できる。
- 複数の LM を使った投票が、単一 LM よりも認識を早く・安定させる。
- 新しい視点の観測や部分的な遮蔽にもある程度対応可能。
- ただし、現状は「物体一つ」「接触なし」「物理変化なし」など比較制約の多いタスクが中心。複雑な操作や複数物体・動的環境への応用はまだ限定的。

強み・興味深い点

- 参照フレーム (Reference Frames) の明示的使用
- 物体や特徴の位置・向きを身体基準 (body-centric) またはオブジェクト基準で扱うことにより、視点変化や身体運動があってもモデルが安定する。
- 感覚-運動 (sensorimotor)中心の学習
- 単に静的画像で学ぶのではなく、観測→動き→観測のループを前提。動き（sensor motion / efference copy）を活用して未知の部分を補う。
- モジュール性とスケーラビリティ
- Repeating Learning Module の構成、モッドュール間通信プロトコル CMP、および階層構造 + 投票方式により、「小さな部分」から「大きな全体」へ、段階的にモデルを構築可能。
- 継続学習と仮説ベースの認識
- 学びと推論が分離されておらず、常に観測で仮説を更新する。新しいオブジェクトを自律的に追加しうる構造。

限界・課題・将来の改善点

課題	内容
複雑な現実世界への拡張	現在のタスクは比較的制限された環境（静止物体・１物体・遮蔽物や重力・動きなどの物理変化が限定）であり、多物体・動的環境・物理インタラクションのある状況でどこまで通用するか未検証。 arXiv
計算コストと効率	グラフベースモデル、仮説空間、投票などは計算負荷が高くなる。特に多くの LM、複数センサ、複雑なオブジェクトでの処理コスト。速度・メモリの最適化が必要。
部分観測と遮蔽	部分しか見えない物体、あるいは遮蔽物の後ろにある物体などでrobust に認識する能力は限定的。仮説の初期化、証拠の集約がより洗練される必要あり。
モーター操作および操作タスクへの応用	物体認識・姿勢推定までは進んでいるが、「目的を持って物体を操作する『道具を使う」など複雑な動的制御タスクではまだ未踏の部分。ポリシー生成の整備が今後。
表現の内部表現のブラックボックス化	現在はグラフで明示的に記述されており可視化が可能。将来的にはもっと抽象的・神経的な (grid cell や sparse coding 等) 表現に拡張される予定。ですがそのとき可解釈性がどうなるかは課題。 arXiv

「オブジェクトモデル」はグラフ的構造

- **Thousand Brains でいう「オブジェクトモデル」とは？**
- 大脳新皮質のミニカラム（あるいはコラム）は、それぞれが「物体（object）」を **部分的に** 表すモデルを持つ。
- このモデルは「感覚入力（視覚・触覚など）」と「参照フレーム（位置・方向）」の組み合わせで成り立っています。
- たとえば「カップの取っ手の部分を、この位置・方向から見たときの感覚パターン」＝部分的な表現。

グラフ構造としての特徴

- **ノード（頂点）**：部分的な特徴やサブパーツ。
  例:「カップの取っ手」「カップの縁」「円筒の胴体」。
- **エッジ（辺）**：それぞれの特徴の間の **位置関係・空間的関係**。
  例:「取っ手は胴体の右側にある」「縁は胴体の上にある」。
- 各コラムは、部分的なノードとその周囲との関係を学習・保持する。
- たくさんのコラムが協調することで、**全体の物体グラフ**が統合されていく。
- つまり、単一のコラムが「小さな部分グラフ」を持ち、
- 多数のコラムがリンクして「大きなオブジェクトのグラフ」を構築、
- それを統合すると「脳が持つオブジェクトモデル」になる、
  というイメージです。

世界モデルとの違い

- 世界モデル（強化学習や生成モデルの文脈でいうもの）は「時間を通した状態遷移」を重視しがち（動的な予測）。
- Thousand Brains のオブジェクトモデルは、むしろ「空間的な構造・参照フレーム・関係性」に重点がある。
- ただし、複数オブジェクトをリンクし、さらに時間発展を組み込めば「世界モデル」に近い表現になる。

Welcome to the Thousand Brains Project Documentation!

- これらのコア原則に従う必要がある
- **感覚運動の学習と推論:**
- 静的な入力の代わりに、アクティブに生成された感覚入力の時間的シーケンスを使用しています。
- **モジュラー構造:**
- 簡単に拡張でき、拡張可能です。
- **皮質メッセージングプロトコル:**
- モジュールの内部動作は高度にカスタマイズ可能ですが、その入力と出力は定義されたプロトコルに準拠しているため、多くの異なるセンサーモジュール(およびモダリティ)と学習モジュールがシームレスに連携できます。
- **投票：**専門家の集まりがさまざまな情報とモデルを使用して、より迅速で堅牢かつ安定した結論に達することができるメカニズム。
- **参照フレーム:**学習したモデルには、構造化された 4D 世界(3次元空間＋時間（＝4次元）)のモデリングに自然に長けている帰納バイアスが必要です。学習したモデルは、操作、計画、これまで見たことのない世界の状態の想像、高速学習、一般化など、さまざまなタスクに使用できます。
- **学習と推論が密接に絡み合っている迅速で継続的な学習:**感覚運動の具体化と参照フレームによってサポートされ、生物学的にもっともらしい学習メカニズムにより、継続的な学習の設定下で堅牢性を維持しながら、迅速な知識の蓄積と保存された表現の更新が可能になります。学習と推論の明確な区別もありません。私たちは常に学習し、常に推論を行っています。

ここで構築しようとしているシステムと他の AI システムとの最も重要な違い

- 私たちは感覚運動システムを構築しています。世界と対話し、時間の経過とともに世界のさまざまな部分を感知することで学習します。静的データセットからは学習しません。これは、今日のほとんどの主要な AI システムとは根本的に異なる学習方法であり、(部分的に重複する) 一連の異なる問題に対処します。
- 皮質柱に匹敵する、基本的で反復可能なモデリングユニットとして学習モジュールを紹介します。ここで指摘すべき重要なことは、これらのモデリングユニットのいずれも完全な感覚入力を受け取らないということです。たとえば、視覚では、どこにも「完全な画像」はありません。各センサーは、世界の小さなパッチを感知します。これは、入力全体が単一のモデルに入力される今日の多くの AI システムとは対照的です。
- 単一のモデリングユニットで、オブジェクトの認識と操作のすべての基本的なタスクを実行できる必要があります。より多くのモデリングユニットを使用すると、システムがより高速かつ効率的になり、構成的および抽象的な表現がサポートされますが、単一の学習モジュール自体が強力なシステムです。単一モデルのシナリオでは、指の 1 本でコーヒーカップを認識するのにその表面を移動する必要があるのと同じように、推論には一連の観測値を収集するために常に動きが必要です。
- すべてのモデルは参照フレームによって構成されています。オブジェクトは単なる特徴の袋ではありません。これは、場所の機能のコレクションです。フィーチャ同士の相対的な位置は、フィーチャ自体よりも重要です。

私たちのシステムがすでに持っている機能(少なくともある程度):

- 世界内の位置や向きに関係なくオブジェクトを認識します。
- オブザーバー、または世界の別のオブジェクトに対するオブジェクトの位置と方向を決定します。
- ノイズの多い条件下で学習と推論を実行します。
- 少数のサンプルから学習します。
- 明示的な監督なしに環境との継続的な相互作用から学習し、以前に学習した表現を維持します。
- オブジェクトが他のオブジェクトによって部分的に隠されているときにオブジェクトを認識します。

現在取り組んでいるその他の機能:

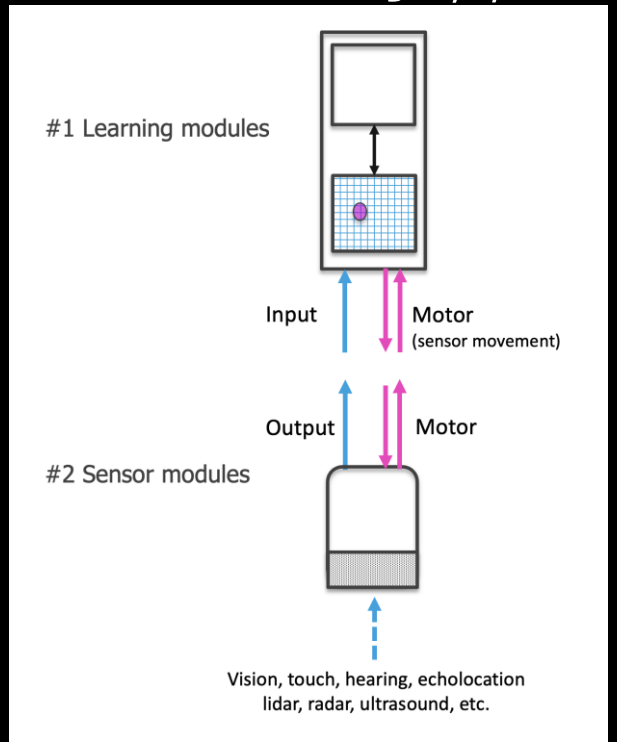
- オブジェクトのカテゴリを学習し、カテゴリの新しいインスタンスに一般化します。
- 構成オブジェクトの学習と認識、それらの部分の新しい組み合わせを含む。
- 新しい変形を受ける物体を認識する(例えば、ダリの「溶ける時計」、しわくちゃになったTシャツ、3Dで学習したが2Dで見える物体)。
- スケールとは無関係にオブジェクトを認識し、そのスケールを推定します。
- オブジェクトの状態と動作をモデル化して認識します (たとえば、ホッチキスが開いているか閉じているか、人が歩いているか走っているか、これらの条件下で時間の経過とともに体がどのように進化するか)。
- 学習したモデルを使用して世界を変え、より単純なタスクに分解する必要がある目標を含む目標を達成します。最高レベルの包括的な目標は、外部で設定できます。

システムが最終的に備えるべき機能
(これらは、以前の機能が構築されているのと同じ原則から一般化されます)。

- 具象モデルから導出された抽象概念へのモデリングの一般化
- 言語をモデル化し、それを世界のグラウンディングモデルと関連付けます。
- 他の実体のモデル化(「心の理論」)。

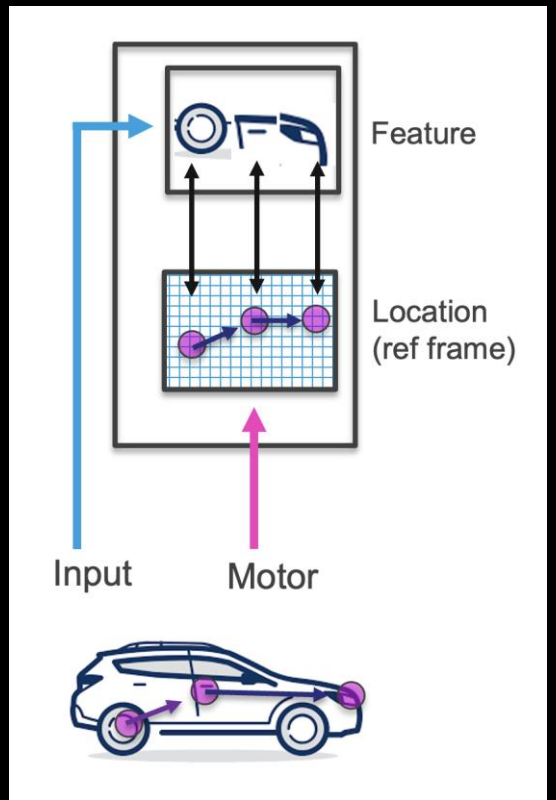
センサモジュール

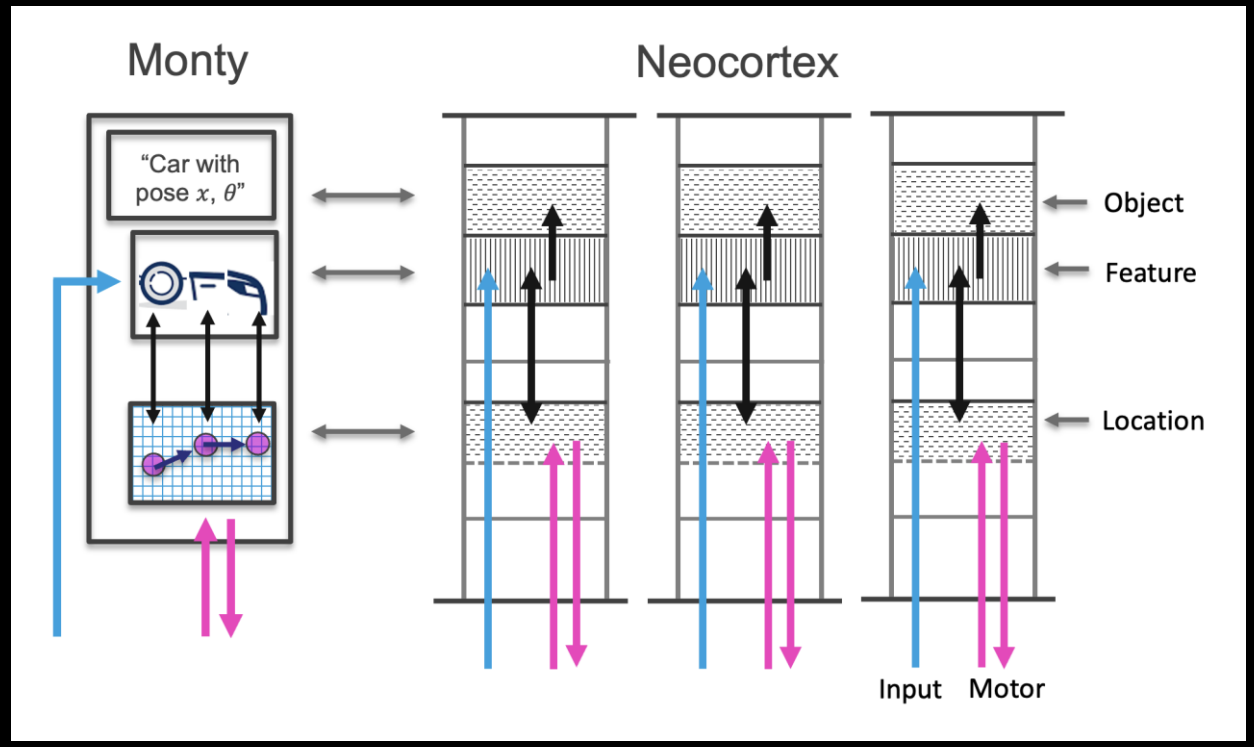
- *センサーモジュールは、生の感覚入力を受信して処理します。これは、共通のメッセージングプロトコルを介して学習モジュールに伝達され、学習モジュールはこれを使用して、環境内のあらゆるもののモデルを学習および認識します。*



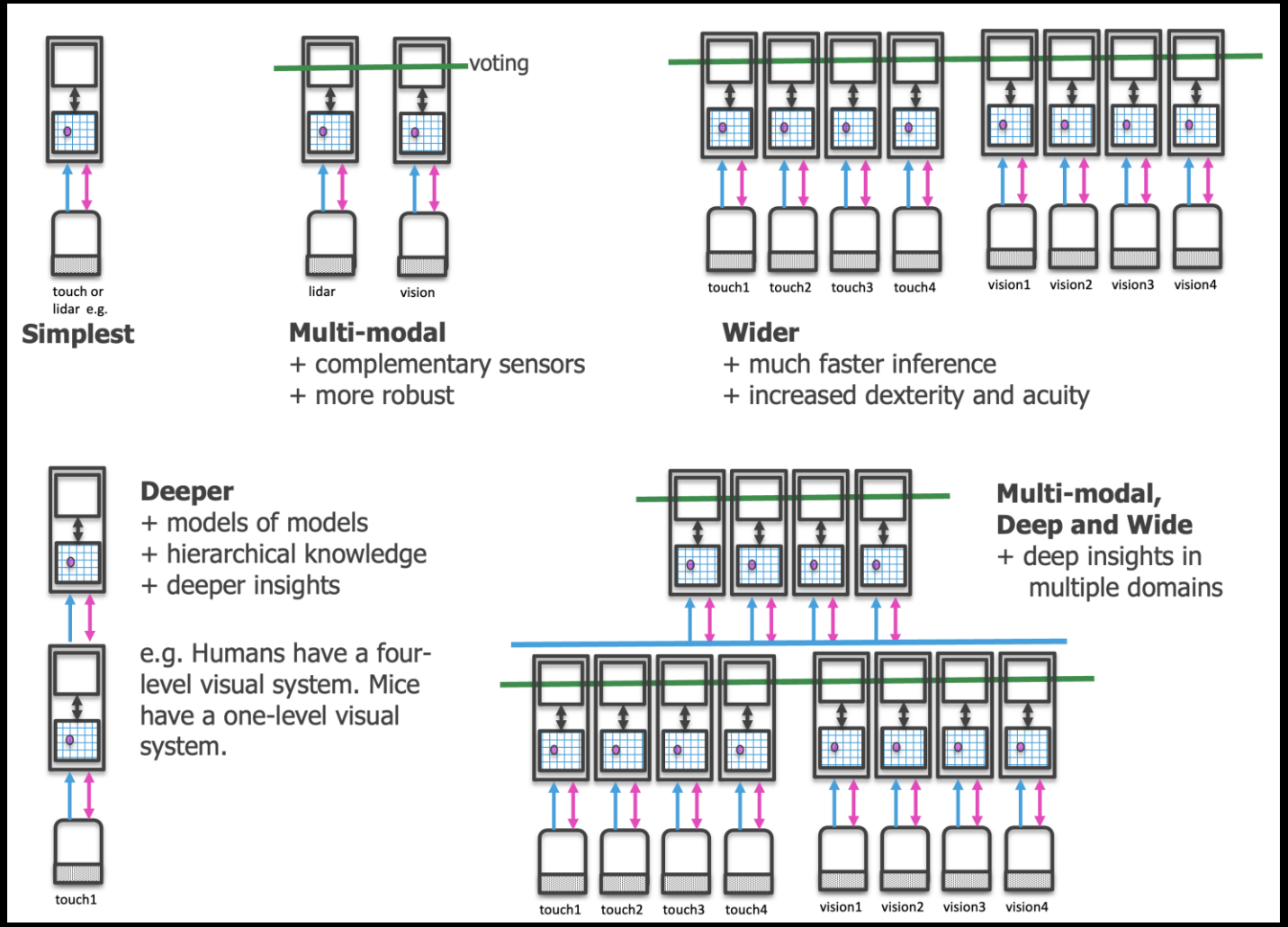
学習モジュール

- *学習モジュールは、参照フレームを使用して、感覚運動相互作用を通じて構造化モデルを学習します。入力フィーチャが空間と時間において相互に相対的にどのように配置されるかをモデル化します。*

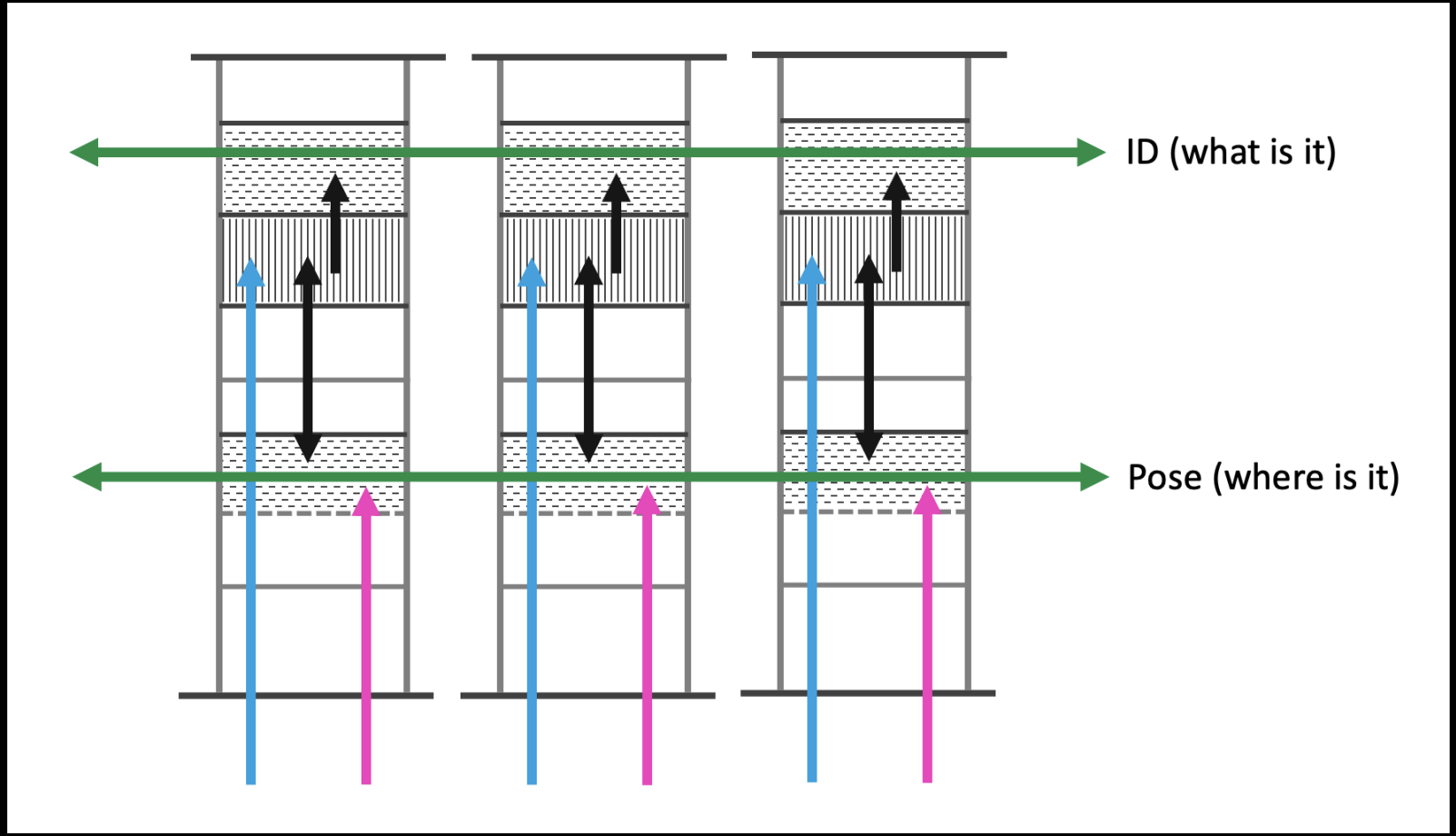




Cortical Messaging Protocol



投票/コンセンサス



- 青い線は、階層の上位の情報の主な流れを示します。紫色の線はトップダウンの接続を示し、下位レベルの学習モジュールにバイアスをかけます。
- 緑色の線は横方向の投票接続を示します。
- ピンクの線は、最終的にモーターシステムのモーターコマンドに変換される目標状態の通信を示します。
- すべてのLMには、直接モーター出力があります。実線に沿って伝達される情報は、CMPに従います(つまり、特徴とポーズが含まれています)。
- 図の中止は、線の端にドットでマークされています。
- 破線は、システムと世界および皮質下のコンピューティングユニットとのインターフェースであり、CMPに従う必要はありません。
- 青い破線は、センサーからの生の感覚入力を伝達します。
- ピンクの破線は、モーターコマンドをアクチュエーターに伝達します
- SMからモーターシステムへの直接の青い線は、高速でモデルフリーのポリシーをサポートするために感覚情報を送信します。
- 大きな半透明の青い矢印は、より大きな受容野からより高いレベルのLMに直接感覚出力を伝送する接続の例です