3K1-OS-06a-5

全脳アーキテクチャに必要な新皮質マスターアルゴリズムの検討

Towards the cortical master algorithm for whole brain architecture

山川 宏 *1, *3
Hiroshi Yamakawa

荒川 直哉 *1, *3
Naoya Arakawa

高橋 恒一 *2, *3
Koichi Takahashi

*1 ㈱ドワンゴ
Dwaongo Co., Ltd.

*2 理化学研究所
RIKEN

*3 全脳アーキテクチャ・イニシアティブ
The Whole Brain Architecture Initiative

The local circuit of the cerebral cortex that acquires various functions with a uniform mechanism is called a canonical cortical circuit, which could be drew on as the foundation of general intelligence. While a whole brain architecture requires a canonical cortical circuit to implement a master algorithm (MA), its required specifications have not been clarified. In this paper, we present a MA framework, defining the interface including input/output semantics and its functions, while integrating models in neuroscience. In addition, candidate artificial neural network models for the MA are evaluated with the specifications.

1. はじめに

全脳アーキテクチャは脳全体のアーキテクチャに学びヒトのような汎用人工知能の工学的な創造をめざす研究アプローチである. この研究では脳の主な器官である, 大脳新皮質, 大脳基底核, 海馬, 扁桃核, 視床さらには小脳などを機械学習器として実装し, それらをメゾスコピックレベルのコネクトームなどの神経科学知見に基づき結合してゆく.

全脳アーキテクチャ研究は経験からの学習を通じて多様な問題に対する解決能力を獲得する人工知能である汎用人工知能の構築が目標である. 大脳新皮質の局所神経回路は一様な構造を持ちつつも入出力に与えられる情報に応じて多様な機能を発揮できるため, 大脳新皮質はこの汎用性を支える中心的な脳器官であろう. 神経科学においても新皮質の標準モデルである canonical cortical circuit の検討は度々行われてきている [Harris 13]. 全脳アーキテクチャの完成に必要な新皮質の局所回路に対応する標準的な機械学習アルゴリズムを, 新皮質マスターアルゴリズム (MA) と呼ぶことにする [Domingos 15].

本稿では既に蓄積されてきた神経科学知見と機械学習の議論をベースに, 新皮質 MA に必須となる入出力を特定し, 新皮質 MA に近い機能を持つ機械学習アルゴリズムを列挙したうえで, その入出力および必要とされる機能について整理する.

なお本議論では、人や動物の脳のように実時間で動作する認知アーキテクチャを想定する.

2. 新皮質 MA とその入出力

近年の深層学習を中心とした人工ニューラルネットワーク (ANN)の発展により, 十分にデータがあればヒトの様々な認識能力をデータからの学習により実現できる例が増えている. たとえば, convolutional neural network (CNN) に含まれる情報圧縮やプーリングの仕組み, LSTM で利用されるゲーティングの仕組みや強化学習を組み込む仕組みなどは, 新皮質 MA に至るために必要な機能セットの相当部分をカバーしているだろう.

2.1 新皮質 MA のインタフェースを検討する

新皮質はヒトの場合に約 140 億個の神経細胞より構成され, 第一次視覚野 (V1), 補足運動野 (SMA) などのような機能的な部位に分解される. 新皮質 MA はそうした部位を単位として実装されると想定され, その入出力信号の種類を特定した上で標準的なセマンティクス(意味付け)を決定することが重要な意味を持つ.

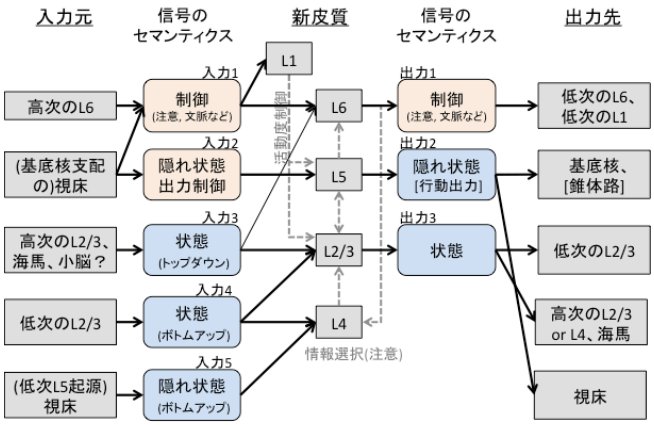

図 1: 新皮質 MA の入出力
信号のセマンティクスに対応した形で入力 1〜5 および出力 1〜3 を定義している。詳しくは 2.2 節を参照。

解剖学的には新皮質は 6 層構造を形成するが, ここでは情報処理の単位として5種類の層{L1, L2/3, L4, L5, L6}に分けて考える. 各層には高々数種類 (サブタイプ) の興奮性細胞 (主に錐体細胞) が存在する.

新皮質 MA の入出力信号のセマンティクスは以下の様に決定する. まず特定のサブタイプの神経細胞は, 特定のセマンティクスを持つ信号を出力すると考えられるため, 出力のセマンティクスは基本的には層ごとに決定される. 一方で, 入力信号については情報を発信する脳器官の性質によって分類する必要があるが, この際に新皮質内で異なるセマンティクスの情報が統合される可能性を考慮する必要がある.

我々は[船水 15]のベイジアンフィルタ仮説, AGI.io [Kowadlo 15, Kowadlo 16]の標準新皮質回路, [Solari 11]の Cognitive Consilience, また [Numenta 10] の新皮質の層間の関係を主に参考として, それぞれの知見からみて整合性の高い形で入出力をまとめた(図1).

連絡先:山川宏, ㈱ドワンゴ ドワンゴ人工知能研究所, 〒113-0033 東京都文京区本郷 5 丁目 24-5

2.2 新皮質 MA の入出力とセマンティクス

本論では実時間で変化する外界から得られるセンサ情報に対応して現在の状況に対する信念を表す信号を「状態」信号と呼ぶ. 信号のセマンティクスは, この状態信号を代表とするものと, 制御に関わる制御信号に大別される.

状態信号は L2/3 から出力し(出力 3), 他領域の L2/3, 高次領野の L4, 海馬に送信する. また,「隠れ状態」信号は, ある状態の履歴や未来への予測を含む時間的情報を持つ信号で, L5 から出力し(出力 2), 基底核, 視床に送信される. なお L5 の異なるサブタイプから出力する「行動出力」信号は錐体路に送信される.

入力信号のうち状態に関わるものは, 低次領野から受け取る「状態 (ボトムアップ)」信号(入力 3)と高次領野から受け取る「状態 (トップダウン)」信号(入力 4)に分類される. 高次領野の L2/3 と海馬などから得られる状態(トップダウン)信号は, 主に L2/3, さらに L6 に入力する. 低次領野の L2/3 から得られる状態 (ボトムアップ) 信号(入力 4)は, L4 もしくは L2/3 に入力する. 低次領野の L5 を起源とする隠れ状態信号(入力 5)は視床を介して投射(伝達)されており L4 に入力する.

「制御」信号は注意や文脈に関わる信号であり, L6 から出力され(出力 1), 低次の L6 や低次の L1 に送信される. 高次領野の L6 もしくは基底核支配で視床から得られる制御信号は L1 もしくは L6 に入力される(入力 1). 基底核支配で視床から得られる制御信号は,「隠れ状態」制御出力として L5 に入力される(入力 2).

本稿では新皮質の動作機序として以下のようなモデルを仮定している. L4 は視覚情報処理におけるプーリングなどの情報選択を行う場合にのみ利用する. 外界のモデルは基本的に L2/3 および L5 で構成されるリカレント結合に保持される. L5 は時間的な広がりをもつ隠れ状態である. L6 が扱う制御信号は L1 を経由して L5 と L2/3 の活動度を制御する.

全般的にみて, 認識の経路である L2/3 については神経科学知見が多く, これを実装した計算モデルは多い. 強化学習を支える L5 を起点とする基底核ループについては, 神経科学的知見は増えてきている(下記「PBWM モデル」の項参照). L6 の入出力については神経科学の知見が未だ少なくその機能についての仮説も計算モデルも見当たらない.

近年, 神経修飾因子の大域的な投射による制御(学習係数の制御などを含む)の可能性も示唆されている [Pi 13] ため, 新皮質全般の学習パラメータ制御の仕組みの考慮も必要である.

3. 新皮質 MA の機能

3.1 機能一覧

ここでは理想的な新皮質 MA に想定される機能を列挙する.

階層性: 学習器が同一あるいは互換性のある計算単位(層)で構成され, それぞれの学習器の入出力に下位上位の方向付けがあり, 任意の数の層を積み上げてより表現力の高い大きな学習器を構成できること(入出力定義に関しては「新皮質 MA とその入出力」の項参照).

次元削減: 高次元の状態空間を, より低次元の状態空間でできるだけ情報を失わないように近似的に表現する機能.

例えば強化学習では現実的に生起する状態のみを含む空間で探索をおこなうことで効率化できる.

教師なし学習: 教師信号もしくは報酬信号を用いずに状態の系列から内部パラメータを変更する能力. その内部では教師あり学習を利用してもよい. 典型例としては後述する時系列予測, Disentangle/直交化, カテゴリー化などがある.

新皮質内の膨大なパラメータの詳細は経験からの学習で決定されるが, 教師／報酬信号は現実世界においてスパースすぎるため, 本機能が必須となる.

時系列予測: 時系列を入力とし, 次の入力を予測したり時系列の背後にある隠れ状態を推定したりする機能. 従来技法としては隠れ状態マルコフモデル(HMM)やカルマンフィルターが用いられる.

予測することはさまざまな問題解決やいきものの生存のために必要な機能である. 脳の予測能力に関しては予測符号化と呼ばれる仮説があり, 脳の中で予測誤差信号が重要な役割を果たしているとされる [Bastos 12] .

Disentangle/直交化: 一般に高次元の状態空間においては複数の因子が絡みあった形で存在する. 特に次元削減を行った場合にはこの性質が顕著であると想定される. Disentangle や直交化はそうした因子を解きほぐして独立性の高い因子として抽出する機能である. 例えば ATARI ゲームタスクでは, ボールの位置(上下/左右)や点数などが独立した変数として取り出せる.

独立した因子が取り出せると, その組み合わせにより未経験の状態に対する推論が可能になる.

カテゴリー化: 連続値の状態空間を離散化された有限集合に分類する機能. 脳では, 分類個数が可変(ノンパラメトリック)なカテゴリー化として教師なし学習が行われていると考えるのが適切であろう.

カテゴリー化された要素は記号との対応付けが可能となり, 記号的な概念の生成や記号の使用の基盤となる. またシステムをオートマトンとみなすために必要な離散化を提供するとともに, 行動選択の基盤を与える. さらに否定的概念や存在論のある性質を支えるためにも必須である [山川 14] .

スパース表現: 状態毎の記述を少数個の変数に制限した多くの変数の値が0であるような情報表現. 新皮質においては抑制細胞を組合せた k-WTA (Winners Take All) のような実装が想定される. 工学的には正則化項を導入することが多い.

この表現では, 見かけ上の変数の次元数は大きくなるが, 各状態の近傍が減るため実質的な次元数は小さくなるためカテゴリー化が容易になる [山川 14] .

内部状態: 多様な認知機能を実現するためには, 操作可能な形で何らかの(主に離散的な)情報を一時的に保持する作業記憶が必要である. 特に生体の神経回路では, 個々のニューロンもしくはニューロン群においてユニークに表現される意味(例えば外部世界における「猫」)を保ちつつ情報処理を行うためには状態保持の機能が必須となる. 状態保持は局所リカレント回路やニューロンごとのフィードバックによる活性化状態状態の保持, もしくは速いシナプス可塑性などにより実現されると考えられている.

単にある状態を保持しつづける機能を「単純状態保持」と呼ぶことにする(この機能は必ずしも離散化された状態である必要はない). 一方, 離散的な状態を内部に保持しつつ時系列を処理する系はオートマトンと呼ばれ, 認識できる形式言語の種類により正規文法を認識する有限状態機械や入れ子構造を持つ文脈自由文法を認識するプッシュダウン・オートマトン(PDA)などに分類される. なお, 機能的な観点からは入出力が離散的な時系列であればオートマトンの内部状態は必ずしも離散値でなくてもよい.

表1: 既存モデルの入出力と機能の一覧

		入出力セマンティクス／機能	理想の新皮質MA	AGI.io	ベイジアンフィルタ仮説	Cognitive consillience	PredNet	HTM	BESOM	Stacked Auto Encoder	Ladder Net	CNN	Frontal cortex for PBWM
入出力		階層性	○	○	○	N/A	○	△	○	○	○	○	N/A
	入力1	制御	○	○	○	○	×	×	×	×	×	×	○
	入力2	隠れ状態出力制御	○	○	×	○	×	×	×	×	×	×	○
	入力3	状態(トップダウン)	○	○	×	○	○	○	○	○	○	×	×
	入力4	状態(ボトムアップ)	○	○	○	○	○	○	○	○	○	○	○
	入力5	隠れ状態(ボトムアップ)	○	○	○	○	×	○	×	×	×	×	?
	出力1	制御	○	○	○	○	×	×	×	×	×	×	×
	出力2	隠れ状態	○	○	○	○	×	○	×	×	×	×	○
	出力3	状態	○	○	×	○	○	○	○	○	○	×	×
機能	教師なし学習	時系列予測	○	N/A	N/A	N/A	○	○	×	×	×	×	N/A
		Disentangle/直交化	○				×	×	○	×	?	×	
		カテゴリー化	○				×	△	○	○	×	×	
		次元削減	○				○	×	○	○	○	○	
		スパース表現	○				○	○	○	△	○	×	
	内部状態	単純状態保持	○				○	×	×	×	×	×	
		有限状態機械	○				○	○	×	×	×	×	
		プッシュダウン・オートマトン	○				?	×	×	×	×	×	

脳にオートマトンの能力を想定する事例としては，例えば鳥やクジラの「歌」が正規文法で表現可能なことや，ヒトの言語の文法の多くの部分が文脈自由型文法で記述可能なことなどが挙げられる．脳がこうした時系列パターンを生成したり認識したりするためには，少なくともそれぞれの文法に対応する種類のオートマトンをエミュレートする能力がなければならない．また，作業記憶を扱う能力もオートマトンとしてモデル化できる．

リカレントニューラルネットワーク（RNN）は，任意のオートマトンをエミュレート可能なチューリングマシンと理論的に等価であることが証明されている [Siegelmann 95]．実用においても RNN は有限状態機械や PDA をエミュレートすることが知られている [Tino 98][Gers 01]．LSTM や GRU といったゲート付きの RNN は，ゲート付回路により状態を保持することから PDA のにおけるスタックを実現しているとみなすことができる．LSTM の PDA 的な挙動の一例として，XML のネストしたタグの関係を認識，生成させる実験がある [Graves 13]．

表には盛り込まなかった重要な項目として，以下のようなものがある．例えば，トップダウン方向に状態信号の入出力を持てば，生成モデル機能（トップダウン情報を変換して，運動も含むセンサ情報を再構築する能力）を持つと想定できるが，これは新皮質 MA が保持すべき重要な機能の一つである．また，制御信号入力を持つ場合は，活動度制御を通じたトップダウンの注意を実現するための回路的な要件を満たしていると考えられる．このほかの，新皮質 MA が工学的に有用であるために満たすべき機能的な要件の一つとしてモデル並列化可能性がある．これは1つの大きな学習器を N 個の並列に動作する計算単位に分割して実行した際に，N が増えるに従って1ステップの計算の入力が与えられてから出力を行うまでの遅延時間を減少させられるという計算手法の特性である．脳と同等の規模の実時間処理を実現しようとするならばモデル並列性を持つことが有利であるが，現状の手法の多くはネットワーク全体を誤差逆伝播で学習するものが多く，同期ボトルネックが存在する．

3.2 既存モデルの評価

以下、新皮質 MA の候補となりうる既存モデルを紹介する．このうち PredNet や HTM, BESOM は，脳の神経回路への準拠の度合いが大きいが，Stacked Auto Encoder や Ladder Network，CNN は工学的な発想から構築されている．

Deep Predictive Coding Network (Deep PredNet): Deep PredNet はトップダウン信号を予測値，ボトムアップ信号を予測誤差とする階層型のニューラルネットワークである．予測符号化の考え [Rao 99] にもとづき，David Cox らにより開発された [Lotter16]．CNN と層内結合を表現する LSTM を組み合わせることで時系列予測による教師なし学習を行う [Lotter 16]．基本的に新皮質の 2/3 層の機能を再現している．

Hierarchical Temporal Memory (HTM): HTM は，時系列予測，スパースコーディング，階層性を特徴とする，大脳新皮質の構造にヒントを得た機械学習システムである．大脳皮質との関連については [George 09] に詳しい．実装されたアルゴリズムと大脳皮質の関係についてはホワイトペーパー [Numenta 10] を日本語で読むことができる．多階層モデルの実装は未確認であるが，クラスタリングの実装については [Balasubramaniam 15] を参照されたい．

BESOM: BESOM(BidirEctional Self-Organizing Map)は新皮質のような階層性を持つベイジアンネットワークに，自己組織化マップによるカテゴリー化機能と独立成分分析による直交化機能を付与した大脳新皮質計算モデルである．[一杉 08]

Stacked Auto Encoder: Auto-Encoder（自己符号化器）型ニューラルネットワークでは入力層と出力層のニューロンを同じ数とし，その間の隠れ層のニューロンの数をそれらよりも少なくすることで，入力データが一度圧縮され元に戻るような学習がなされ，一種の次元削減器として機能する．自己符号化器を積み上げて多層化したものが，Stacked Auto-Encoder である．低次層から順次学習を行なった後に全体の学習を行うことが多い．

Ladder Network: 教師ありの学習のモデルに対して，教師なし学習である Auto-Encoder モデルを混合させた半教師あり学習モデル．比較的少ない教師データからでも良い分類性能を得ることができる[Rasmus 15].

Convolutional Neural Network (CNN): 次元削減のための畳み込み層と情報選択のためのプーリング層を交互に積み重ねた階層型ニューラルネットワークで，各層はフィルタと呼ばれる検出器を複数持つ．概ね，畳み込み層を L2/3 に，プーリング層を L4 にそれぞれ対応づけることができる．主に画像処理に利用され，入力に近い層では狭い領域を扱い，段階的に広い領域をカバーする．センサから入力された信号をボトムアップに処理し，ネットワークにループは存在しない．

以下の二つは実装がなかったり，新皮質のモデルでなかったりするが，参考のために記しておく．

Canonical Microcircuits for Predictive Coding（CMPC）:

[Bastos 12] では，神経系における情報処理の自由エネルギー原理や能動推論による定式化を提唱するカール・フリストンらが大脳皮質の構造を予測符号化の観点から分析している．ここでの分析では，ボトムアップの情報は L2/3 層から L4 層への予測誤差であり，トップダウンの情報は L5 から L2/3 層への予測情報とされている．

PBWM モデル: PBWM (Prefrontal cortex and Basal ganglia Working Memory) モデルは，[O'Reilly 06] によって提唱された作業記憶のモデルで，大脳皮質と大脳基底核，視床を含むマクロな回路により，内部状態の保持を必要とする作業が実行可能になるとするものである．前頭野・視床ループで保持される作業記憶の更新のタイミングは，大脳基底核における強化学習によって決定されるとされる．

4. おわりに

本稿では，主に神経科学分野の計算モデル研究を参考に，標準新皮質回路に対応する機械学習モデルである新皮質マスターアルゴリズム (MA) の要求仕様を提案した．MA の入出力定義とそのセマンティクス，および MA がその内部で実現すべき計算機能を述べた．新皮質モデルの候補となりうるいくつかの人工ニューラルネットワークモデルについて評価を行なったところ，現状の ANN のほとんどが状態信号のトップダウン・ボトムアップでのやり取りに留まっており，注意などに関わる制御信号や時間に関わる隠れ状態信号を十分にモデル化していないことがわかった．

今後は，本論文で示した新皮質 MA をたたき台として，ANN などの機械学習として実装する研究を進めたい．

また，本論考で，新皮質 MA に資する神経科学的知見がまだまだ不十分であることが明らかになり，神経科学者との協業を一層強化する必要性を認識した．

参考文献

[Balasubramaniam 15] J. Balasubramaniam, C.B.G. Krishnaa, F. Zhu: "Enhancement of Classifiers in HTM-CLA Using Similarity Evaluation Methods," *Procedia Computer Science*, Vol. 60, pp.1516-1523 (2015)

[Bastos 12] Bastos, Andre M., et al. "Canonical microcircuits for predictive coding." *Neuron* 76.4, 695-711 (2012)

[Domingos 15] Pedro Domingos: *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, Basic Books (2015)

[George 09] D. George, J. Hawkins: "Towards a mathematical theory of cortical micro-circuits," *PLoS Comput. Biol.*, 5, p. e1000532 (2009)

[Gers 01] Felix A. Gers and Jürgen Schmidhuber: "LSTM Recurrent Networks Learn Simple Context Free and Context Sensitive Languages," IEEE TNN 12(6):1333–1340 (2001)

[Graves 13] Alex Graves: "Generating Sequences With Recurrent Neural Networks," arXiv: 1308.0850 (2013)

[Harris 13] Kenneth D. Harris & Thomas D. Mrsic-Flogel: "Cortical connectivity and sensory coding," *Nature* 503, 51–58, (2013)

[Kowadlo 15] Gideon Kowadlo & David Rawlinson: How to build a General Intelligence: Circuits and Pathways, Project AGI (Blog), http://agi.io/ (2015)

[Kowadlo 16] Gideon Kowadlo & David Rawlinson: How to build a General Intelligence: An interpretation of the biology, Project AGI (Blog), http://agi.io/ (2016)

[Lotter 16] William Lotter, et.al.: "Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning," https://arxiv.org/abs/1605.08104 (2016)

[Numenta 10] Numenta, Inc.: HIERARCHICAL TEMPORAL MEMORY including HTM Cortical Learning Algorithms (2010)

[O'Reilly 06] O'Reilly, R.C & Frank, M.J.: "Making Working Memory Work: A Computational Model of Learning in the Frontal Cortex and Basal Ganglia," *Neural Computation*. 18 (2): 283–328. doi:10.1162/089976606775093909 (2006)

[Pi 13] Hyun-Jae Pi, et.al.: "Cortical interneurons that specialize in disinhibitory control," *Nature* 503, 521–524, (2013)

[Rao 99] R. P. N. Rao and D. H. Ballard: "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects," *Nature Neuroscience* (1999)

[Rasmus 15] A. Rasmus, H. Valpola, M. Honkala, M. Berglund, and T. Raiko: "Semisupervised learning with ladder network," CoRR, abs/1507.02672, 2015.

[Siegelmann 95] H.T. Siegelmann and E.D. Sontag: "On the Computational Power of Neural Nets," *J. Comput. Syst. Sci.* 50, 1, 132-150. DOI=http://dx.doi.org/10.1006/jcss. (1995)

[Solari 11] Solari SVH, Stoner R.: "Cognitive Consilience: Primate Non-Primary Neuroanatomical Circuits Underlying Cognition," *Frontiers in Neuroanatomy* (2011)

[Tino 98] Tino, Peter, Bill G. Horne, and C. Lee Giles: "Finite state machines and recurrent neural networks--automata and dynamical systems approaches" (1998)

[一杉 08]一杉裕志, "脳の情報処理原理の解明状況, 産業技術総合研究所テクニカルレポート," AIST07-J00012 (2008)

[山川 14] 山川宏: "記号を接地しうる表現としての大脳新皮質カラム構造の検討," JSAI2014, 2C4-OS-22a-4 (2014)

[船水 15] 船水章大, 銅谷賢治: "予測―大脳新皮質のベイジアンフィルタ仮説," 生体の科学. 66 (1), pp.33-37 (2015)