4G2-OS-14b-4

汎用 AI 実現のための鍵となる自律性とマルチモーダル性についての考察

Autonomy and Multimodality for Constructing Artificial General Intelligence

栗原 聡[*1*2] Satoshi Kurihara　高屋 英知[*1] Eichi Takaya　高橋 良暢[*1] Yoshinobu Takahashi　芦原祐太[*1] Yuta Ashihara

[*1]電気通信大学 The University of Electro-Communications　[*2]人工知能先端研究センター Artificial Intelligence eXploration Research Center

In this paper, we will discuss how to construct AGI (Artificial General Intelligence). Autonomy and multimodality are key factors. We also show an AGI architecture consists of complex multi-layered architecture and complex network of diverse multimodal information. Behavior selection is also important function for the consciousness space for interaction with environment, and we will propose a multiagent based approach.

1. はじめに

昨年の AAAI2016 での基調講演において，Demis Hassabis 氏は「Deep Mind 社はこれから GAI (General Artificial Intelligence) 開発を本格化させる」と宣言した．そもそも共通した具体的な定義付けが難しい AI という言葉にさらに「汎用」を付けた単語であるからして，汎用 AI の定義はさらに抽象的である．著者としては，過去の学習の再利用や異なる学習同士の組み合わせにより，より適応的にかつ能動的に実環境とのインタラクションが可能であり，プランニングであればリアクティブ性・熟考と即応の両立など，「高い適応性・柔軟性」や「頑健性」といった性質を持つ AI，というとらえ方をしている．無論その最高の手本は我々人である．

そして，今後の AI の発展については，主に次の 2 つの道に分化すると考えている．1 つ目が汎用 AI であり，その活躍の場は人とのインタラクションが必須な我々の日常生活環境である．「場の空気を読む」「人との阿吽の呼吸」といった関係を構築できることが課題であり，人が AI に対して「感情」や「意識」を感じられる AI の構築が目的となる．人と同じ仕組みではなく，人が AI に対して一方的に感情や意識を感じることで十分なのかもしれないが，少なくとも，AI には何らかの意図 (行動を誘発させるエンジン) を埋め込む必要がある．重要なのが目的指向をどのように組み込むかである．

なぜ汎用性を持つ AI が必要となるのか？　力任せでよいのなら Narrow AI の寄せ集めでも実現可能という考え方もあるかもしれない．しかし，寄せ集めて適宜使い分けるメカニズムが肥大することになる．結局，複数の narrow AI をまとめたモジュール化や階層化といった展開となり，選択する部分に負荷が集中することとなる．

人の知能が汎用性を持つ理由は自明である．身体という限定されたリソースで全てに対応しなければならないからである．当然，ロボットを動かすなら，Wifi 等を使い，頭脳部分はロボット本体には搭載されていないサーバー等を利用することが可能であろう．しかし，我々生物にはそれができない．加えて一つ一つの神経細胞の動作速度はコンピュータに比べてはるかに遅く精度も低い．この状況で地球という自然環境で生き抜くために獲得したのが，脳という汎用性の高いデバイスであり，超多数かつ並列に動作する神経細胞における大規模複雑ネットワーク化により汎用性を獲得したと言えよう．つまりは汎用性の特徴は「限られたリソースを最大限に活用する効率性や省エネ」ということになる．特に前者が重要な能力であり，この能力の工学的な有用性は高い．

連絡先: 栗原 聡，電気通信大学，〒 182-8585 東京都調布市調布ヶ丘 1 － 5 － 1，042-443-5660，skurihara@uec.ac.jp

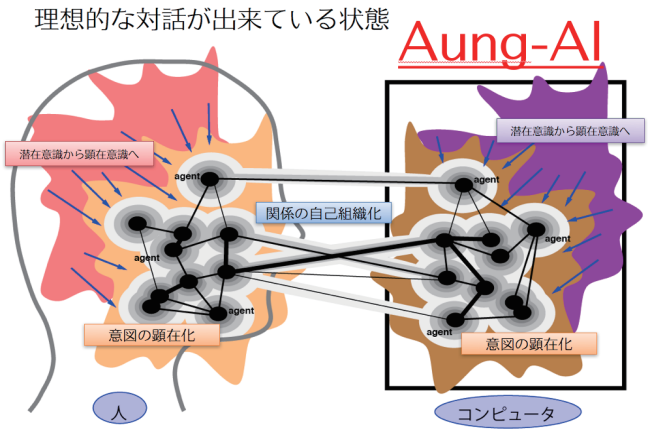

図 1: 人と AI とのインタラクション

2 つ目がビッグデータ用 AI とも呼ぶべきルートで，超ビッグデータを対象としたデータマイニング能力であり，膨大な論文から新たな関係や科学的発見の手かがりを探索するなど，ハイパフォーマンスコンピューティングも活用する必要がある．

本論では，汎用 AI 実現に向けたいくつかの可能性について議論する．

2. 注目すべき要素

唯一の手本である人，そして，その中心的モジュールである脳からの知見として，以下の 3 点が重要であると考えている．

1. 膨大な数のネットワーク化された神経細胞群が並列に動作する超多数自律分散システムであること．
2. そのネットワークが階層性とスモールワールド性を有していること．
3. 五感というマルチモーダルなセンシング入力を効果的に利用しての認知機構と，身体を使っての実環境とのインタラクション能力を有すること．

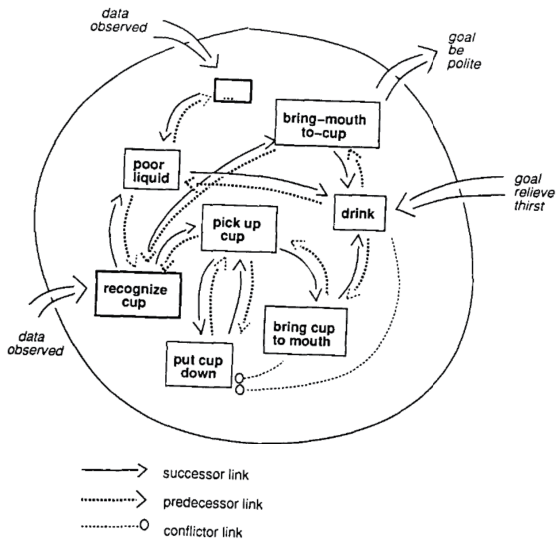

図 2: Agent Network Architecture(ANA)

2.1 AI への自律性の付与が鍵

Siri を始めとして様々な対話システムが登場しているが，これらは対話というより Q&A システムと捉えた方が自然であり，それらにおいては阿吽の呼吸的関係の構築は対象外である．人と AI との間で息の合った会話を行うためには，AI に自律性，言い換えれば目的指向性が必要となる．例えば，相手が「のどが渇いた」としゃべっても，その相手の健康への気遣いと状況によっては「自販機を探そう」と言う場合もあれば「今は我慢して」という言う場合もあろう．このような会話は過去の膨大な会話対のデータを学習しても実現できない．

ここでの目的指向性における AI が持つ目的とは，ロボットであればバッテリ残存電力が尽きないよう行動する目的や，指定された目的地まで移動する目的といった具体的な行動レベルではなく，より上位のレベル，例えば「ロボットの主人の家庭内での快適さを維持する」といった抽象的なものである．このような目的の達成は容易ではない．その目的の達成のための具体的な目的を設定する能力が必要となり，動的に変化する実環境に対応するには，複数の代替プランを用意しておく必要もあるだろう．過去のプランを参照する場合もあれば，複数の過去のプランを組み合わせることで新たな状況に適応するといった能力も必要になるであろう．まさに汎用的な行動選択能力が求められる．

この課題を解決するアーキテクチャをどのようにデザインするかであるが，1991 年の Patte Maes のモデルは今に至っても新鮮である [1]．古典的プランニング法である STRIPS の各単位プランをネットワーク化したものであり，環境側とゴール側から活性伝搬を継続的に送り込む中で活性値が閾値を超えた単位プランが実行される仕組みである (図 2)．脳はこのネットワークが大規模複雑化させたものであり，加えて記憶や優先度や達成までに要する時間が異なる超多数ゴールからの様々な活性伝搬が並列に発生すると考えられる．そして，このモデルは『意識はモニタである』という考え方 [2] とも親和性が高い．

2.2 マルチモーダル性

よく Deep Learning は，高い学習能力を有するものの，相当数の学習データを必要とすると指摘される．たしかに，我々が生涯に見る猫は多くてもせいぜい数百匹程度であろう．しかし，無論であるが我々は 2 次元的な猫の画像のみで猫の概念を学習しているのではなく，時系列的な猫の動作に加え，鳴き声や猫を見た時の情景など，膨大な情報と関連させて学習している．AI においてもマルチモーダルな時系列情報を，各モーダル同士の関連性も含めた学習が出来ることが必須である．これにより学習量の低減化も期待できる．例えるなら，Deep Learning が 1 種類のデータで 100 の量で学習可能なタスクがあるとした場合，10 種類のデータであればそれぞれ 10 の量で学習が完了すると言いたいところ，さらに少ない 7 の量で可能であるという意味である．30 の量の学習が足りないように思えるが，これがモーダル同士の関連性で補完されることになる．

マルチモーダル型の Deep Learning も提案されているが，各モーダルごとに CNN などを適用し次元を圧縮した状態で他のモーダルからの出力と併せて一つの DNN の入力とするといった方法が通例である [3]．

これに対して，本論でのマルチモーダル性はそのとらえ方が大きく異なり，マルチモーダルデータによる複雑ネットワークとして生成されるネットワークにおいて，ある時刻におけるマルチセンサーからの入力により活性化ないしは新たに生成される部分ネットワークを入力とする．マルチモーダルデータによる複雑ネットワークとは，ある時刻において入力されたマルチモーダルデータ同士がネットワーク接続され，さらに各モーダルにおいても類似するデータ同士がネットワーク接続された常に成長する複雑ネットワークである．このネットワークにマルチモーダルなセンサー群から情報が入力されると，その入力に基づきネットワークの一部分が活性化する．実際の神経細胞と同様，その活性化は徐々に沈静化するダイナミクスを有し，ネットワークにはアクチュエータなど，AI が身体を持っていれば腕や足，また音声を制御するモジュールとも接続されており，活性化したネットワークによりそれらが活性化し，外界に対してインタラクションを発動することとなる．実装する上で課題となるのが，活性化したネットワークに対してどのようなインタラクションを発動させるかのマッピングである．マルチモーダルデータにて生成されるネットワークは大規模複雑ネットワークであり，その活性化の組み合わせは膨大であり，活性化のパタンごとにマッピングするのは不可能であろう．活性化した部分ネットワークのサンプリングを行い，特徴を維持しつつネットワークのサイズを縮小する方法や，ネットワークの次元を CNN 型 DNN にて縮小し，中間層レベルの出力をマッピングに利用する方法などを利用することが妥当であろう．特に後者において，CNN の次元圧縮能力は高いものの，入力をネットワークとする CNN 法の開発が必要となる．

3. アーキテクチャ

図 3 に示すアーキテクチャを想定している．入力は上述するようにマルチモーダル型であり，画像等は CNN 型を介して中間層にて概念化された情報を入力とする．環境音なども含め CNN+LNTM 型 [4] での取り込みが適切であろう．人とのインタラクションにおいて対話が最も重要な入力であるが，音声信号から直接的に意図に変換する方法ではなく，音声から発話テキストに変換しての取り込みとする予定である[*1]．その際，音声信号から感情情報を抽出しこれも感情情報として入力する．図 3 左上段は入力されたマルチモーダル情報を時系列順にネットワークを構築させる．各ノードは入力された発話テキストであり，個々のテキストに，そのテキストが発話された状況での画像や環境情報，感情などが付加される．ノード同士は時系列発生順にネットワーク化され，その強度は発生頻度に基

∗1 究極は音声信号そのものを直接入力されることであろう

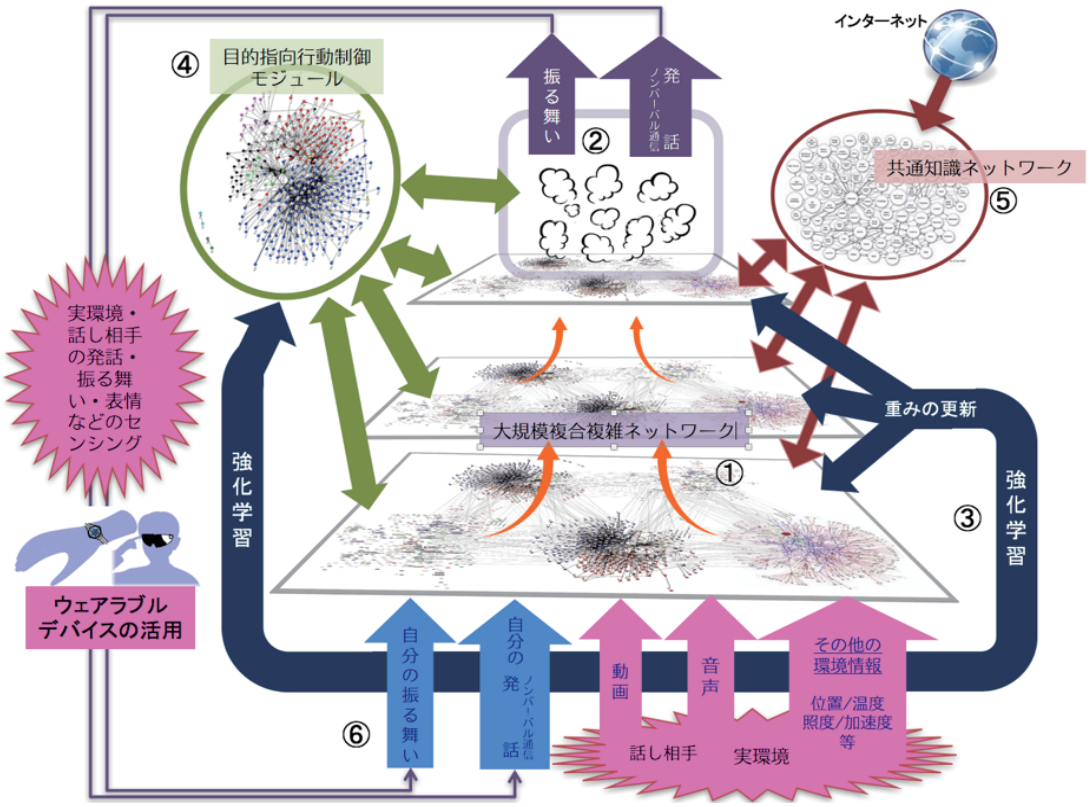

図 3: AGI architecture

づいて強化される．Ant Colony Optimization(ACO)[5] 型のネットワーク構築等が適切であろうと考えている．さらに，この複数のノードが規則的に発火する状況を，メタレベルノードとして上位階層化させ，これを多段階層とすることで，上位階層にて抽象度の高いノードを創発させる．無論，これだけはこのアーキテクチャを有するシステム自体が経験した以上の情報が獲得できないことから，図 3 右上段にて，背景知識や辞書的情報を用意する．そして重要なのがプランニングモジュールであり，この機能により自律性を実現する．無論，発動したインタラクションに対する環境フィードバックに基づき，自身のインタラクション戦略を修正するための強化学習機能も必要である．

4. まとめ

以上，汎用人工知能（AGI）を実現する上に重要となる自律性とマルチモーダル性について考察するとともに，それらを実現する認知アーキテクチャについて概観した．このアーキテクチャにおいて最も負荷が集中するのがマルチモーダルデータにて構成される複雑ネットワークの制御である．脳は現状にコンピュータに比べて，大規模層的複雑ネットワークの個々のノード全体が並列処理を行うアーキテクチャという意味で全く異なっており，このアーキテクチャが創発する能力が，個々の神経細胞の処理速度の遅さや不確実性を補っている．脳を手本とする取組において，この並列性をどのようにノイマン型アーキテクチャにて実現させるかも大きな課題となるが，現在開発が進んでいるニューロ型 CPU や FPGA など，ハードウェア技術の進展も必要である．

参考文献

[1] Pattie Maes, The agent network architecture (ANA), ACM SIGART Bulletin Homepage archive Volume 2 Issue 4, Aug. 1991 Pages 115-120

[2] 前野隆司，ロボットの心の作り方 受動意識仮説に基づく基本概念の提案 ，日本ロボット学会誌 23 巻 1 号，pp. 51-62, 2005.

[3] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y.: Multimodal deep learning, in Proceedings of the 28th international conference on machine learning (ICML), pp. 689?696 (2011)

[4] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko and, and Trevor Darrell, Long-term Recurrent Convolutional Networks for Visual Recognition and Description, CoRR, abs/1411.4389, 2014.

[5] http://iridia.ulb.ac.be/ mdorigo/ACO/ACO.html