

生成 AI の動向と用語

7月号担当編集委員

生成 AI の現在地

衝撃的な ChatGPT の登場 (2022.11) によって、生成 AI 革命が始まり、覇権を争う各社が挙って大規模言語モデル (LLM) の構築・改良に邁進している。また、各地で AI 用データセンターの建設が始まり、GPU の争奪戦を繰り広げ、独占的な NVIDIA の時価総額が Apple と MS を抜いて一時首位となった。

(1) 前兆

今の第3次 AI ブームはディープラーニング (深層学習) が推し進めているが、前兆はビッグデータと機械学習の進化により始まった。ビッグデータ処理は2004年に Google Map Reduce の論文が発表され、翌年の Hadoop により普及が始まった。機械学習は当初、人間が特徴量を教示することにより機械自ら学習するものであったが、50 年来のブレイクスルーと言われる深層学習 (DL) が登場した。トロント大の J.Hinton 教授が2006年に深層化に成功し、2012年の画像認識コンペ ILSVRC で圧勝して一躍注目され、世界の AI 研究者が争って深層学習にのめり込んだ。深層学習はニューラルネットワーク (NN) を深化させて、情報抽出を多階層に行って高い抽象化 (メタ化) を実現する機械学習の一種である。コンピュータが不得手だったパターン認識でも人間を凌駕し始めた。

(2) 生成 AI 革命

そして、現在進行形の生成 AI 革命であるが、第4次 AI ブームだとする人もいる。大規模言語モデル (LLM) は、Google が2017年に Transformer の論文を発表、再帰型 NN (RNN) を使わず Attention 機構の採用と並列処理が可能となり注目された。翌年に、このアーキテクチャに基づく Google/BERT を公開、OpenAI も GPT-1 (パラメータ 1.17 億) を発表、以降、GPT シリーズで性能の指標となるパラメータ数を急増させ、

GPT-3.5 (パラメータ 3550 億) による ChatGPT の公表に繋がった。現在の GPT-4 のパラメータ数は非公開だが、1.8 兆とも言われている。パラメータ数を増やすとスケールリング則等が働き、格段に賢くなる。

一方、画像生成 AI は、2021 年に OpenAI が GPT-3 を応用して、テキストから画像を生成する DALL・E を発表した。翌年には stable Diffusion が注目され、Midjourney はアートコンテストで人間を出し抜いて優勝した。

(3) 何故、会話できるのか？

ChatGPT は流暢な会話ができるが、意味を理解しているのではなく、次の単語を予測しているだけとの解説もあるが、次の単語を予測するだけではなく、知識体系を構築して、理解しているように見える。理解しているかの定義は、意見の分かれるところである。ここに意識の萌芽とみる向きもあり、危険性を指摘する声もある。

単語をベクトル変換、単語同士の距離を内積で測り、意味的關係を掴み、文脈を理解しているようだが、うまく機能する理由はよく解っていない。演繹ではなく、統計的に処理する帰納法なので、何故そうなるかの説明は難しい。

また何故「知らない」と言わず嘘をつく (ハルシネーション) のか？ この原因も帰納法の必然のようだが、調整で対応していた。近頃は RAG (検索拡張生成) による低減対策も行っている。

(4) GPU

計算機には GPU が使われている。GPU は3次元グラフィックに必要な行列演算と並列処理によるアクセラレータとして発達したが、これが LLM にも向いていて、NVIDIA の躍進となっている。

半導体はムーアの法則の限界が近づき、チップレット、3D 積層、HBM、省電力化等の技術を進めている。

(5) 巨大テック企業等の動向

Microsoft は、OpenAI に出資して連合を組み、GPT-4 等を Windows 等の自社製品に Copilot として組み込み、Azure と連携して収益化している。対抗する Google は、Transformer 等の LLM 技術で先行していたが、ChatGPT に出し抜かれ、慎重姿勢を一転、コード・レッド（非常事態）を発令して追撃している。クラウド最大の Amazon は、生成 AI 開発では後追いとなっているが Q を展開。Meta もメタバースから生成 AI に主軸を移し、LLM をオープンソースで提供している。Apple は AI チップ M4 を製品に搭載し始め、生成 AI で出遅れていたが、GPT-4o と連携して Siri 等を強化、画像生成にも対応する。

OpenAI の最大ライバルとも言われる Anthropic は、Amazon や Google 等から出資を受けている。加 Cohere は Oracle や NVIDIA 等の出資を受け、RAG（検索拡張生成）に強い。

(6) 現在進行形

LLM はその危険性の対処法として、ソースの非公開派と公開派に分かれていて、オープンソースは、Meta/Llama 3、x AI/Grok、仏 Mistral AI 等が提供している。

生成 AI は言語と画像等を同時に扱うマルチモーダルに進化し、OpenAI は 23 年 3 月に GPT-4 で対応し、24 年 5 月の GPT-4o（Omni）でテキスト・画像・音声のモダリティの組み合わせを実装した。Google も多種多様な情報に対応できる Gemini を 23 年 12 月に、Anthropic は 24 年 3 月に Claude 3 を発表している。

そして、マルチモーダルをさらに発展させた「AI エージェント」は、言語モデルをベースに、AI が問題解決のための行動を決定し、必要なツールを選択して目的を達成する。このような秘書代わりに雑務を処理してくれる自律型のサービスである。こうしたアプリの他、仮想空間の AI アバターやバーチャルヒューマン、実世界のロボット等、多様な形態があり、自律性のレベルも様々である。AI エージェント開発は、生成 AI 技術をテコに加速していて、LLM をベースに LangChain 等の開発ツールを利用して作られる。そして、AI エージェントの初期段階としては、タスクを自動化する AutoGPT や AgentGPT、BabyAGI 等が登場してきている。

(7) 国産 LLM とユーザ

巨大資本による一部の汎用モデルの寡占化が進んでいるが、言語別・用途別の専用モデルが多数生まれる

余地があり、国産 LLM はそこを目指しているようだ。

国内の AI 用データセンター（DC）は、GPU の入手が困難な中、通信会社やさくらインターネット等が整備を進めている。

スタートアップの ELYZA はオープンソースの Meta/Llama 2 を利用して、世界水準の LLM を開発したが、KDDI の傘下に入った。また、著名な Google 出身者達が創業した注目の Sakana AI は、特徴的な「進化的モデルマージ」手法による開発を進めている。

大手では、ソフトバンクがパラメータ数 1 兆超の LLM を開発後に、サイズを小さくして最適化を図るとしている。産総研、NITC 等も開発に取り組んでいる。NTT や NEC は、最初から小さな日本語 LLM 開発を選んでいる。

一方、ユーザは、当初、プロンプト入力による企業秘密の漏洩やハルシネーションを懸念し、導入の躊躇がみられたが、オンプレミス運用や RAG（検索拡張生成）機能により、最新の Web データや社内データを追加し、ハルシネーション対策ともしている。

(8) 将来

東京大学の松尾豊教授は、LLM の次は「世界モデル」（外界を観測して外界構造を学習）だと予測していて、知能の身体性にも注目している。世界モデルのユースケースとしては自動運転への応用が研究されているが、これは建設機械にも必要な研究課題である。

生成 AI のフェーズが学習から推論へと移行する中で、推論特化型の AI 半導体が GPU の牙城の一部を崩す可能性が高い。

また、ポスト GPU としては、ノイマン型計算機からデータフロー型への移行が考えられる。LLM 等の DL はデータフロー型であるが、ノイマン型計算機上で動かしている。そのため、演算機近くに高速メモリを配置するが、データフロー型計算機では必要なくなる。

OpenAI は 24 年 2 月に難しいとされた動画生成 AI の Sora を公開したが、この拡散モデルを応用して、ロボットの行動生成への応用がその分野で進められている。動画を生成できるということは、世界の様々な物体の形状やダイナミクスを把握（世界モデル）していることを意味する。

また、OpenAI では謎のプロジェクト Q* が進行中で、論理的思考や数学的能力を飛躍させた AGI の開発とも噂され、これがアルトマン解任劇の背景にあったとの報道もある。

さて、AI エージェントが自律性を高め、その行動範囲を広げていく先に、AGI（汎用人工知能）という

ゴールが見えてくる。その過程で AI が意思を持ち、人間の利害に反する行動をとる危険性を危惧する人々が多い。

紙幅の関係で省いた事案も多いが、近年の動向を概観した。毎週のように革新が起き、変革の加速が急だ。手元にある 1 年半以前の多くの AI 本が古くなったように、ここで記した動向も間もなく陳腐化する運命にある。

生成 AI 関連の用語

用語解説のリクエストがあったので、初学者が遭遇しそうな用語の解説を加える。

あ

アテンション：attention：「注意機構」を参照

う

埋込みベクトル：Embedding Vector：ベクトル埋込みとは、単語や文章、画像を、それらの意味や関係性を捉える数値に変換する技術で、様々なデータタイプを多次元空間上の点で表す。

え

AI アバター：人の代わりに窓口業務等を行うアバターで、プライベート LLM を土台として開発が進む。

AI アライメント：人間の意図した目標と AI が達成する目標を一致（アライメント）させること

AI エージェント：複数の AI モデルを組み合わせ、単一のモデルでは困難な高度なタスクを自動で実行できるシステムのことである。一方、Google は、論理的な思考（Reasoning）と行動計画（Planning）、作業内容に関する長期記憶（Memory）の能力を備え、未来に関する多段階の思考が可能であり、様々なソフトウェア等を操って、ユーザーの代わりに業務を遂行できるものと定義しているが、安全性確保がその開発の課題となっている。

AI バイアス：機械学習バイアスまたはアルゴリズムバイアスとも呼ばれ、人間のバイアスによって元のトレーニングデータや AI アルゴリズムが歪められ、歪んだ出力や潜在的に有害な結果に繋がる偏った成果の発生を指す。

演繹的アプローチ：Deduction：既に知られている法則や前提から論理を積み重ねて結論を出す方法

エンベディング：「埋込みベクトル」を参照

お

オートエンコーダ：入力データを効率よく本質的な特徴まで圧縮（エンコード）し、この圧縮表現から元

の入力を再構成（デコード）するように設計されたニューラル・ネットワーク・アーキテクチャの一種である。オートエンコーダは、教師なし機械学習を用いて、入力データの潜在変数を発見するように訓練される。

オープンソース：Open Source：無償（または最低限の必要なコスト）で公開されたプログラム等のソースコードで、自由に利用や、改変、再配布ができるという特徴がある。

音声処理：speech processing：音声信号を分析して何らかの情報を表す種々の特徴パラメータを取り出し、それに基づいて合成や認識等を行うこと。主な分野としては、音声を音響特徴量へ変換する音声分析、音声から言語情報を取り出す音声認識、話者情報を取り出す話者認識、音響・言語特徴量から音声を作り出す音声合成、音声信号を圧縮して記憶・伝送する圧縮・符号化などがある。

オンプレミス：システムの稼働やインフラの構築に必要なとなるサーバやネットワーク機器等を自社で保有し運用するシステム利用形態のことである。

か

回帰分析：複数のデータ同士の関係性や影響力を調べる統計的手法。単回帰分析では最小二乗法が代表的、その他に重回帰分析、ロジスティック回帰分析がある。

過学習：overfitting：モデル規模を大きくすると、最初は性能が上がるが、その後性能が下がる現象の要因

拡散モデル：Diffusion Models：画像データの拡散過程（ノイズが加えられてデータが破壊される過程）を学習した画像生成 AI モデルの一つである。

画像生成 AI：テキストで指示を与えると画像を自動生成する AI、代表的なものに Stable Diffusion や DALL・E2、Midjourney 等がある。

き

機械学習：データを分析する方法の一つで、データから、コンピュータが自動で学習し、データの背景にあるルールやパターンを発見する方法

記号主義：symbolism：人間の知能は記号として表現された知識を操作することで実現されているという考え方。或いは、知識を記号として表現し、コンピュータ上で操作して、知能を人工的に実現できるとの考え方

記号接地問題：人工知能において、実世界にある事物の名称と、コンピュータ内部で扱う記号との結びつけに伴う問題

帰納的アプローチ：Induction：実際に観測された個別的，特殊的な事例から一般的・普遍的な規則・法則を見出そうとする論理的推論

強化学習：Reinforcement Learning：機械学習の一種で，コンピュータが動的環境と，繰り返し試行錯誤のやりとりを重ねることによってタスクを実行できるようになる手法である。この学習手法により，タスクの報酬を最大化する一連の意思決定を行うことができる。

教師あり学習：学習データに正解を与えた状態で学習させる手法，代表的なアルゴリズムには，回帰と分類がある。

教師なし学習：学習データに正解を与えない状態で学習させる学習手法で，データ内に存在する未知のパターンを見つけ出す。

く

クエリ：query：問い合わせ等の意味を持つ英単語で，ITではデータの問い合わせや要求等を一定形式で文字に表現することを意味する。

クオリア：qualia：感覚的な意識や経験のこと

クラスタリング：データセットを特定のルールに基づいていくつかのグループ（クラスタ）に分類することを指す。特にデータ間の類似度に基づいて，似たものを集めたグループに分ける手法が代表的で，クラスタ分析やクラスタ解析とも呼ばれる。

グラフニューラルネットワーク：GNN：グラフデータに深層学習の仕組みを取り入れたニューラルネットワークで，ここでいうグラフは，ノード（節点）とエッジ（線）から構成されるデータ形式である。

け

決定木：Decision Tree：木構造を用いて分類や回帰を行う機械学習の手法の一つ

言語モデル：自然言語処理に使われる技術の一つで，人間が日常的に使用する言語のパターンや言い回し，文法，意味を理解すると同時に，単語の次にどの単語が続くかを予測することができ，端的に言う「言語表現の確率分布」のこと

こ

広帯域メモリ：「HBM」を参照

勾配降下法：Gradient Descent：最適化の手法の一つで，関数の最小値や最大値を見つけるために使用され，機械学習やディープラーニングでは，モデルのパラメータを最適化するために使われる。

誤差逆伝播法：Backpropagation：出力層で誤差関数の勾配を計算し，重みとバイアスを調整した後，出力から逆方向に向かって逆伝播し，勾配を計算して

重みとバイアスを調整，これを逆伝播しながら繰り返す。

コネクショニズム：connectionism：AI研究においてNNに基づいた知能体を実現・実装する立場

コーパス：Corpus：自然言語の文章や使い方を大規模に収集し，コンピュータで検索できるように整理された膨大な量のデータベースのことで，AIが自然言語を扱う際のデータ学習に利用する。

コンテキスト：Context：文脈，前後関係

さ

再帰型ニューラルネットワーク：Recurrent NN：音声認識で一般的に使用されるNNの一種，RNNはデータの時系列特性を認識し，そのパターンを使用して次に起こり得るシナリオを予測するように設計されている。他のNNとは異なり，RNNには過去の入力を記憶できる内部メモリがあり，現在の入力と過去の入力から学んだことを考慮して意思決定を行うことができる。

し

次元削減：多次元のデータをなるべく情報を失わないように低次元のデータに落とし込むこと。データの量を節約できるので計算を高速化でき，またデータが解釈しやすくなるというメリットがある。

自己回帰モデル：データを先頭から最後まで順に生成するモデル

自己注意機構：Self-Attention：データの流れ方自体を学習し決定するような方法である。RNN向けに提案されたが，CNNなど他のNNにも利用されている。自己注意機構は強力であり機械翻訳・質問応答・画像生成等，多くの問題で最高精度を達成している。自分自身の途中の計算結果に注意し，そこから読み込むことからこの名がついている。

事前学習：既に学習済みのモデルを利用する手法であり，NNに新たな層を加え，その加えた層のパラメータだけを調整することによって別のタスクに転用する転移学習等に用いられる。

自然言語処理：NLP/Natural Language Processing：従来の計算言語学に機械学習を組み合わせ，自然言語の文脈を理解する処理

蒸留テクニック：大きなモデルでしか解けなかった問題を解く能力を小さいモデルに移せること

自律型AIエージェント：AIによる自律的なアプリケーション，指示により自動的に業務をサブタスクに分割して実行する。

深層学習：Deep Learning：大量のデータを自動で解析して，データの特徴を抽出するAI技術。機械学

習の一種で、ディープ NN (DNN) と呼ばれる NN を多層化して学習する。DNN は、人間の脳神経回路 (ニューロン) を模して作られており、入力層、中間層、出力層で構成されている。2012 年にトロント大学のヒントン教授のチームが画期的な画像認識として発表後、第 3 次 AI ブームを引き起こした。

す

スケーリング則：自然言語処理モデルのサイズ (= パラメータ数) や、データセットのサイズ、トレーニングに使用される計算量が増えるほど、より高い性能を発揮できる、という法則 (べき乗則)

せ

生成 AI : Generative AI：クリエイティブな成果物を生み出すことができる AI で、テキスト・画像・動画・音声等を生成する。

た

大規模言語モデル : LLM/Large Language Model：自然言語処理 (NLP) の分野で広く使用される機械学習モデルの一種である。LLM は大量のテキストデータを学習してそのデータのパターンや構造を理解し、人間のような文章生成や自然言語タスクの実行が可能である。LLM は、文章生成、文章の要約、質問応答、文書分類、機械翻訳などのタスクに活用される。モデルは一般的に大規模なパラメータセットを持ち、トレーニングデータの情報を学習することで、与えられた入力に対して予測や生成を行う。

畳込みニューラルネットワーク : CNN：畳込みとは、フィルターをかけた画像を特徴マップに落とし込むことであり、そのうえで一致不一致などを判断する。CNN は画像の特徴を高精度で抽出でき、画像認識の分野で活用されてきて、画像生成にも多大な影響を及ぼした。

ち

チップレット：回路を複数の小さなチップに個片化し、「インターポーザ」と呼ぶチップレット間をつなぐ基板上に乗せて大規模化してパッケージに収める技術

注意機構 : attention mechanism：AI が深層学習によってモデルを構築する際に重要な役割を果たす仕組みで、畳込み NN (CNN) や再帰型 NN (RNN) 等と並ぶアーキテクチャで、自然言語処理等に利用されている。これは「Attention Is All You Need」という論文の中で、Transformer という新たな NN の技術として紹介された。この Attention と呼ばれる仕組みにより、AI がタスクを完了するためにどの単語が重要であるかに注意を向けることが可能と

なった。

チューリングテスト : Turing test：アラン・チューリングが提案した、ある機械が「人間的」かどうかを判定するためのテスト

て

ディープラーニング：「深層学習」を参照

敵対的生成ネットワーク : GANを参照

データサイエンス：数学や統計学、機械学習、プログラミング等の理論を活用して、莫大なデータ (big data) の分析や解析を行い、有益な洞察を導き出す学問のことを指す。

データマイニング : data mining：大量のデータ (big data) を統計学や AI 等を駆使して分析を行い、何らかの知見を得るための活動のことで、「マイニング」とは「採鉱」を意味する。

転移学習 : Transfer Learning：既存の優れたモデルで捉えた特徴を新たなモデルに転用して、学習時間を短縮できる。

と

トークン：LLM におけるトークンは、テキストデータを処理する際に使用される基本的な単位である。言語モデルや自然言語処理のコンテキストでは、テキストをトークン化することで、個々の単語、文字、またはサブワードに分割される。トークン化の方法は、言語やモデルの要件によって異なる。

トロッコ (トロリー) 問題：“命の選別”をめぐる倫理的ジレンマを扱った思考実験、暴走するトロッコの軌道上に 5 人の作業員がいて、そのままでは 5 人は轢死するが、支線へのポイントを切り替えて助けると、支線側の別人が犠牲になる。どうすべきか？ AI 設計上の課題でもある。

ドメイン特化型 LLM：特定の業種や企業などに特化した大規模言語モデル

に

二重降下 : double descent：CNN, ResNet, Transformer 等の深層学習モデルでは、「Early Stopping」または「正則化」を使用しない場合、「二重降下現象」が発生する。パフォーマンスは最初に向上し、次に悪化して、モデルサイズ、データサイズや訓練時間を増加させると再び向上する W 字型曲線を示す。

ニューラルネットワーク : Neural network：人間の脳の働きを模した方法でデータを処理するようにコンピュータに教える人工知能の一手法。これは、深層学習と呼ばれる一種の機械学習プロセスであり、人間の脳に似た層状構造で相互接続されたノードやニューロンを使用する。

ニューロモーフィックデバイス：Neuromorphic device：人間の脳の仕組みを模倣したデバイス

の
ノーコード／ローコード開発ツール：プログラムのソースコードを極力記述せずに開発できるツール群である。今後、生成 AI により、より進化する。

は
パーセプトロン：Perceptron：脳神経を模することによってパターン認識を行う。F. ローゼンブラットが 1957 年に考案したが、単層パーセプトロンの限界から研究は停滞した。その後、ボルツマンマシンや、多層パーセプトロンの機械学習をある程度実用化した誤差逆伝播法等によって 1980 年代に再興隆した後、90 年代に再び停滞した。21 世紀に入って、コンピュータ能力の向上と研究の進展により、多層パーセプトロンでの「深層学習」の実用化と、その画像認識等の成果が、現在の第 3 次 AI ブームとなった。

バックプロパゲーション：「誤差逆伝播法」を参照

パラメータ数：機械学習モデルが学習中に最適化する必要のある変数の数を指し、これらの変数は、モデルが入力データを処理し、予測や生成を行う際に使用される。モデルのパラメータは、学習データを使って最適な値を見つけ出すことにより、モデルの性能を向上させるために調整される。例えば、NN の場合、重みやバイアスというパラメータがあり、ネットワークの構造によって異なるが、それぞれのニューロンが持つ重みやバイアス値が学習によって最適化され、モデルの性能を向上させる。LLM である GPT-3 のパラメータ数は 1,750 億個で、GPT-4 では 1.8 兆個のパラメータを持つと言われている。

ハルシネーション：hallucination, 幻覚：生成 AI（言語モデル）が、もっともらしい誤情報（事実とは異なる内容）を出力することを指す。

汎化：generalization：様々な異なる対象に共通する性質や、共通して適用できる法則等を見出すこと

汎用人工知能：AGI：Artificial General Intelligence：あらゆるタスクを人間と同等かそれ以上にこなすことができる、AI の発展形である。特定のタスクを処理する特化型 AI とは異なり、AGI はあらゆるタスクを柔軟かつ独立して処理する能力を有している。生成 AI の出現が大きなマイルストーンとなり、AI エージェントの先に出現か？

ひ
ビッグデータ：Big data：伝統的なデータ処理やデータ管理アプリケーションでは対処できない莫大な量の複雑なデータセットを指す。

ふ
ファインチューニング（微調整）：「事前学習」した訓練済み NN モデルの一部もしくは全体を、別のデータセットを使って再トレーニングすることで、新しいタスク向けにモデルのパラメータを微調整することを指す。

フォン・ノイマン・アーキテクチャ：ジョン・フォン・ノイマン（Neumann）が提唱したプログラム内蔵方式の一般的なコンピュータ構成を指す。ニューロコンピュータや量子コンピュータは、非ノイマン型である。

フォンノイマンボトルネック：CPU とメモリの間のデータ転送に制限があることから生じる処理能力の限界を指し、コンピュータの処理能力を制限する原因の一つとなっている。

フレーム問題：frame problem：AI が本来問題解決には必要のない背景も考慮して（枠を越えて）計算し続けてしまい、情報処理能力を超えて機能が停止する問題のことである。

プロンプトエンジニアリング：言語モデルへの質問において、課題に合わせて入力を工夫する技術

へ
ベイズの定理：確率と不確実性を扱うベイズ統計学の基盤となる公式で、観測されたデータと曖昧な情報を含む事前確率から事後確率を求めるという点で従来の統計学とは異なる考え方とされている。

ベクトル：単語をベクトル変換、単語同士の距離を内積で測り、意味的關係を掴み、文脈を理解する。

ベクトル検索：単語（キーワード）が含まれる文章を探す従来型の「キーワード検索」に対して、「ベクトル検索」はテキストを高次元の数値ベクトルで表現し、似たもの同士を示す「コサイン類似度」により、質問に適した回答を探す仕組み

ほ
ボルツマンマシン：全てのニューロンが互いに接続され、層状のニューロンではなく、接続に向きがなく、情報伝搬が双方向に確率的に行われる NN

ま
マルコフ連鎖：Markov chain：確率過程の一種であるマルコフ過程のうち、とりうる状態が離散的なものや時間が離散的なものを指すことが多い。マルコフ連鎖は、未来の挙動が現在の値だけで決定され、過去の挙動と無関係である（マルコフ性）。

マルチモーダル：テキストに画像・映像・音声等を組み合わせたモデル。Gemini, GPT-4V, PaLM-SayCan 等がある。

む

ムーアの法則：インテルの共同創業者の一人である
ゴードン・ムーアが1965年に発表した半導体技術
の進歩についての経験則で「半導体回路の集積密度
は1.5～2年で2倍となる」という法則

も

モラベックのパラドックス：人間にとって難しいこと
は機械にとって簡単であり，人間にとって簡単なこ
とが機械にとって難しいという観察事象

ら

ラッダイト運動：産業革命に沸く19世紀の大英帝国
で，失業を恐れた労働者が機械を打ち壊した騒動。
進化するAIが人の仕事を奪うというAI反対運動
に対する反面教師としてよく使われる。

ろ

ロジスティック回帰モデル：他のML手法よりも数
学的に複雑ではなく，ML初学者でも実装できる。
また，メモリや処理能力等の計算能力が低くても，
大量のデータを高速に処理できる。

A

AGI：Artificial General Intelligence：「汎用人工知能」
を参照

AI PC：AIをクラウドではなく，ローカルのCPU/
GPU/NPU等のプロセッサで処理されるPCのこと
(NPU内蔵のSoCを搭載したPC)

B

BERT：Transformerを構成するエンコーダを応用し
た言語モデルで，大規模言語モデルの元祖となる。

C

CASE：Connected, Autonomous/Automated,
Shared, Electricの頭文字を取った次世代における
自動車の概念

CNN：「畳込みニューラルネットワーク」を参照

Copilot：本来は副操縦士の意味であるが，近年，
Microsoft等が提供するアシスタント機能を有する
AI支援ツールの名称となっている。

CUDA：GPUでAIを高速に動かすためのNVIDIA
のソフト基盤，多くの開発企業がCUDAに合わせて
AIのソフトやアプリを作っていて，強固な開発
者用エコシステムを構築している。

D

Diffusion Models：「拡散モデル」を参照

DNN：Deep NN：現在のDNNはパラメータを多く持ち，
各ニューロンが非線形関数を表現できるため，複雑
なデータに対応できる表現力の高いモデルとなる。

E

EDA：Electronic Design Automation(電子設計自動化)：
半導体や電子機器の設計の自動化，そのツールやソ
フトウェアも意味し，コンピュータによる設計支援
システムの総称

G

GAN：Generative Adversarial Network：敵対的生成
ネットワーク：生成モデル(Generator)と識別器
(Discriminator)の2つのNNを意図的に競争させ
ることによって，高品質な画像の生成を可能にする
技術である。

GitHub Copilot：Microsoft傘下のギットハブ社の設計・
開発(コーディング)・テスト等を支援する生成AI
ツール群

GLUE：General Language Understanding Evaluation：
自然言語処理のベンチマークテスト

GPGPU：General-purpose computing on GPU：GPU
の演算資源を画像処理以外の目的に応用する技術

GPT：General Pre-trained Transformer の略

GPT-1：2018.6に発表，1.17億パラメータ，トラン
スフォーマーアーキテクチャと教師なし事前学習を組
み合わせ，特定のタスクにファインチューニング

GPT-2：2019.2に発表，15億パラメータ

GPT-3：1,750億パラメータ

ChatGPT：2022.11に発表，GPT-3.5にチャットの
ようなユーザーインターフェイスを付加，これが支持
され，爆発的に急速普及

GPT-4：2023.3に公開，**GPT-4o**：2024.5に発表

GPU：Graphics Processing Unit：画像処理用半導体
であるが，並列処理と行列演算に強いので，DLにも
不可欠の存在となった。コア数が数千個に及び並列処
理に優れている。また，メモリに順次アクセスし，条
件分岐のない単純計算に強く，行列演算に向いている。

grokking：あるタスクができないように見えていたモ
デルが，突然理解できるようになる場合があること
がわかった。これは想定されていた深層学習の機能
とは異なる働きで，OpenAIの研究チームはこの振
る舞いを「グロッキング」と名付けた。

H

HBM：High Bandwidth Memory(広帯域メモリ)：圧
倒的な転送速度(帯域幅)を実現したDRAM，複
数のDRAMを積層し，TSV(シリコン貫通ビア)
で接続，DL学習に必要な大量データのやり取りを
高速化する。

L

LLM：「大規模言語モデル」を参照

M

MoE : Mixture of Expert : 計算処理ごとに LLM の必要な部分だけを選んで使う手法

N

NLP : Natural Language Processing : 自然言語処理

NPU : Neural Processing Unit : AI 専用アクセラレータ, DL (深層学習) で必要な NN の演算を高速に行う専用プロセッサである。

O

OpenAI : AI の研究・開発を行う非営利団体として 2015 年に設立された。

P

Python : インタープリタ型の高水準汎用プログラミング言語であるが, 機械学習や AI の開発にもよく使われる。

R

RAG : Retrieval Augmented Generation (検索拡張生成) : ユーザの質問に応じて知識情報を参照する。社内データを追加学習させるファインチューニングと異なり, LLM そのものに手を加えないため実装しやすい。RAG の一般的な構成は, 企業などの保有する構造化データ等をベクトル変換して, 質問への回答に必要なデータを近似的ベクトル値を基に検索する。その上で, 質問に応じた回答を生成する。ハルシネーション (幻覚) 対策としても有望視されている。

Rerank model : 事前に大量の質問文とそれに相応しい回答ペアを学習して, 質問に対する関連性をランク付けする。

RISC-V : リスク ファイブはカリフォルニア大学バークレー校で開発されオープンソースで提供されている命令セットアーキテクチャ (ISA) である。

RNN : Recurrent Neural Network : 再帰型 NN (連続データ認識による自然言語処理)

RPA : ロボティック・プロセス・オートメーション

S

SDV : ソフトウェア・デファインド・ビークル (ソフトウェアが定義するクルマ)

SLAM : Simultaneous Localization and Mapping の略で, 移動ロボットにおいて, LiDAR やカメラにより環境地図と自己位置推定を交互に行い, 「場所」の概念を獲得させる手法

SoC : システム・オン・チップ

Stable Diffusion : ミュンヘン大学において開発され, 2022 年に公開された潜在拡散モデル (latent diffusion model) であり, 深層生成 NN (deep generative

neural network) の一種である。主にテキスト入力に基づく画像生成 (text-to-image) に使用される。

T

TensorFlow : Google が開発しオープンソース (2017 年正式版) の機械学習用ソフトウェアライブラリで, 数値解析, 深層学習に Python との組み合わせでよく使われる。

TPU : TensorFlow 専用半導体 (ASIC)

Transformer : Google の研究者が 2017 年 12 月に発表したモデルで, 文章に含まれる単語のような連続したデータの関係を, アテンション機構により, 同じ系内にある隔たったデータ要素間の微妙な相互影響や相互依存関係を見つけ, 文脈ひいては意味を学習する NN である。これが AI のパラダイム シフトをもたらすことになった。

Triton : OpenAI が 2021 年 7 月にリリースした GPU 用オープンソース言語, Python ライクな容易な言語で, NVIDIA/CUDA の代替を狙う。Meta や Intel 等が採用を始めた。

TSMC : 台湾積体回路製造

U

UMAP : Uniform Manifold Approximation and Projection の略で, 2018 年に提案された非線形次元削減手法である。リーマン多様体の概念に基づいており, 高次元データを低次元空間に埋め込む際に, データの位相的構造を保持することを目的としている。

X

XGBoost : Extreme Gradient Boosting の略称であるこの手法は, 決定木に基づき, ランダム フォレストや勾配ブーストなどの他の手法を改善したもので, 様々な最適化手法を用いて, 大規模で複雑なデータセット処理で有効である。

(文責 : 岡本)

JCMA

《参考文献》

- 1) 建設機械のシンギュラリティ, 岡本, JCMA, 2015.12
- 2) あたらしい人工知能技術の教科書, 我妻幸長, SE, 2021.9
- 3) AI 技術史, M.Wooldridge, インプレス, 2022.3
- 4) AI 技術の最前線, 岡野原大輔, 日経 BP, 2022.8
- 5) 半導体戦争, C.Millere, タイヤモンド社, 2023.2
- 6) GPT-4/ChatGPT/LangChain, 布留川英一, BD, 2023.6
- 7) 全容解説 GPT, S.kublik, S.Saboo, インプレス, 2023.12
- 8) その仕事 AI エージェントが, 西見公宏, 技術評論社, 2023.12
- 9) AI で世界はこう変わる, 今井翔太, SB 新書, 2024.1
- 10) AI ナビゲーター, 野村総研, 東洋経済新報社, 2024.3
- 11) Google vs Microsoft, 山本康正, 日経 BP, 2024.5
- 12) 生成 AI 真の勝者, 島津翔, 日経 BP, 2024.6
- 13) 日経 XTECH : <https://xtech.nikkei.com>