第 14 講　AI と深層学習の基礎と応用

【学習到達目標】

・AI の発展と深層学習の基本的な概念と仕組みを説明できる.
・深層学習の代表的なモデルの特徴を理解し、適用例を説明できる.
・LLM・VLM・VLA の関係と役割を整理して説明できる。

１．AI の歴史とブームの変遷

人工知能（Artificial Intelligence; AI）とは、人間が行っている知的な活動を計算機上で実現しようとする試みの総称である。推論、学習、認識、計画、対話など、その対象は多岐にわたる。近年の AI の急速な発展を支えている中心技術が「深層学習（ディープラーニング）」であり、多数のパラメータをもつニューラルネットワークを用いて、大量のデータから自動的に特徴を抽出し、さまざまなタスクを高精度に実現する学習手法である。現在の AI ブームを正しく理解するためには、AI の歴史を俯瞰し、ルールベースの時代から機械学習、深層学習、さらに大規模基盤モデルへと至る流れを押さえておくことが重要である。

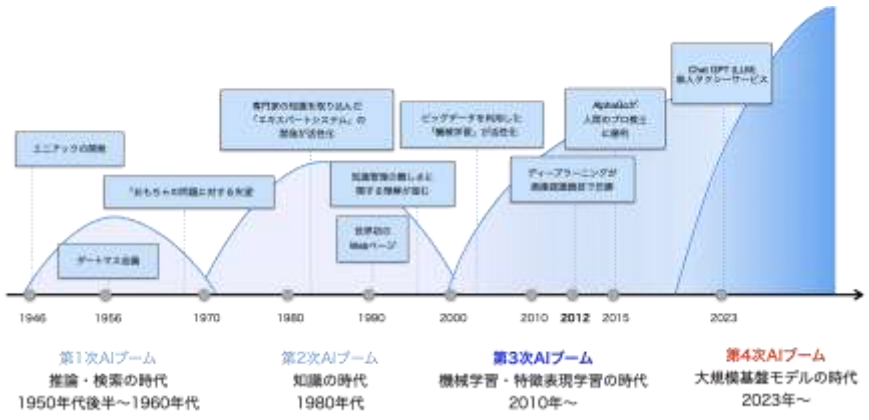

図１　AI ブームの変遷

(1) 第 1 次 AI ブーム：記号処理と推論の時代

1950～1960 年代にかけての第 1 次 AI ブームでは、「記号処理」に基づ

く AI が研究の中心であった。知識やルールを記号（シンボル）として表現し、論理式を用いて推論を行うことで、人間のような思考を再現しようとしたのである。この時期には「人工知能（Artificial Intelligence）」という用語が定義され、推論プログラムや定理証明システムなどが開発されたが、現実世界の複雑な問題に対しては性能が伸び悩み、期待ほどの成果は得られなかった。計算機資源やデータの制約も大きく、やがて第 1 次ブームは収束することになった。

(2) 第 2 次 AI ブーム：エキスパートシステムと知識表現

1980 年代に起こった第 2 次 AI ブームでは、「エキスパートシステム」が大きな注目を集めた。エキスパートシステムとは、医師や技術者などの専門家が持つ知識を多数の IF-THEN ルールとしてシステムに組み込み、そのルールに基づいて診断や助言を行うシステムである。このアプローチは、ある程度限定された領域では高い性能を発揮したものの、次のような問題を抱えていた。

- 専門家の知識をルールとして大量に書き出す「知識獲得」に多大なコストがかかること
- ルールが増えるほど、ルール同士の矛盾や抜け漏れの管理が難しくなること

これらがボトルネックとなり、汎用的かつ拡張性の高い人工知能を実現するには限界があることが明らかになった。

(3) 第 3 次 AI ブーム：機械学習と深層学習

2010 年頃から本格化した第 3 次 AI ブームでは、人手でルールを書くのではなく、「データから法則を学習する」機械学習が主役となった。その中でも特に重要なのが「深層学習」である。深層学習を用いた画像認識モデルは、2012 年の国際コンテスト（ImageNet Large Scale Visual Recognition Challenge）において従来手法に大差をつけて優勝し、その有効性を世界に示した。それ以降、画像認識、音声認識、自然言語処理など、多くの分野で深層学習が従来手法を凌駕する精度を達成し、第 3 次 AI ブームを牽引する存在となった。この流れの象徴的な出来事が、囲碁 AI「AlphaGo」による世界トップ棋士への勝利である。

(4) 第 4 次 AI ブーム：大規模基盤モデルと生成 AI

2025 年の現在では、深層学習をさらに大規模化し、多様なタスクに対応可能な「大規模基盤モデル（Foundation Model）」が登場した第 4 次 AI ブームの段階にあると考えられている。膨大なテキスト、画像、音声、動画などを事前学習したモデルは、テキスト生成、翻訳、要約、プログラム生成、画像・音声の生成など、多様な「生成タスク」を高い品質で実行できるようになった。これらは総称して「生成 AI」と呼ばれ、社会・産業・教育など、さまざまな領域にインパクトを与えつつある。

2. 囲碁 AI AlphaGo と深層学習・強化学習

囲碁 AI「AlphaGo（アルファ碁）」は、第 3 次 AI ブームを象徴する存在として、人工知能の歴史に大きな足跡を残したシステムである。2016 年に世界トップクラスのプロ棋士を破ったニュースは、単に「コンピュータが囲碁に勝った」という話題にとどまらず、「人間の直感や経験に頼ってきた領域にも、深層学習と強化学習が入り込んできた」という象徴的な出来事として受け止められた。そもそも、囲碁は長らく「ゲーム AI の最後の砦」と言われてきた。オセロやチェスのようなボードゲームでも AI は強力であったが、それらは計算資源の増加と探索アルゴリズムの工夫により、かなり早い段階で人間を凌駕していた。しかし囲碁の場合、盤面は 19×19 の交差点で構成され、碁石を置く場所の選択肢が非常に多い。そのため、可能な局面や手順の数、すなわち探索空間は天文学的な大きさになるとされている。この難題に対して AlphaGo は、「深層学習（畳み込みニューラルネットワーク）」と「強化学習」を組み合わせるというアプローチを取った点である。

(1) 畳み込みニューラルネットワークによる教師あり学習

まず AlphaGo は、過去の棋譜データから「上級者がどのような手を打ってきたか」を学習する。インターネット上には、プロや高段者の対局記録、すなわち棋譜が多数公開されている。AlphaGo はそれらを大量に集め、盤面の状態を画像のように入力し、「その局面で実際に人間が打った一手はどこか」を正解として学習したのである。このとき使われたのが、畳み込みニューラルネットワーク（CNN）と呼ばれる深層学習モデルである。盤面の状態を入力とすると、ネッ

トワークは 19×19 の全ての交差点に対して「ここに打つ確率がどれくらい高いか」を出力する。学習の初期段階では、重みはランダムに近いため出力もでたらめであるが、多数の棋譜に対して誤差を小さくするようにパラメータを更新していくことで、次第に「人間の上級者が選びそうな手」を打つことができる。

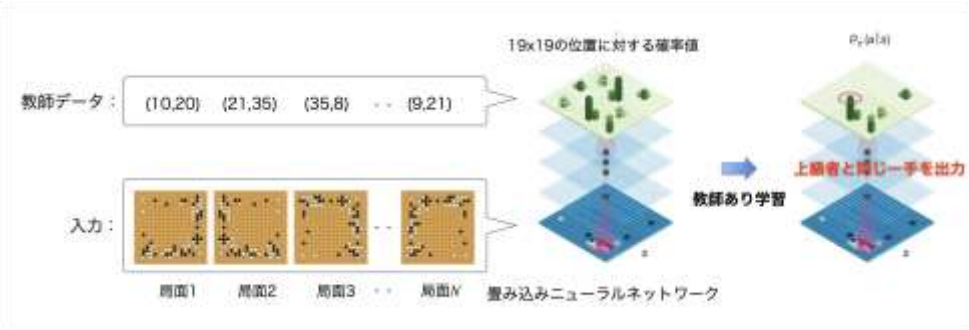

図 2　CNN の教師あり学習

(2) 経験から学ぶ強化学習

しかし、上級者のまねをするだけでは世界チャンピオンに勝つことはできない。プロ棋士同士の対局記録は確かに質の高いデータであるが、それは「過去に人間が経験した局面」に限られている。未知の局面や、まったく新しい戦略については、過去の棋譜に頼るだけでは対応しきれない。そこで AlphaGo は、人間の棋譜から学んだモデルを基礎としつつ、その後は自分自身と対局を繰り返す「自己対戦」によって、さらなる実力向上を図っている。ここで用いられる枠組みが「強化学習」である。強化学習では、個々の手に対する「正解」が明示的に与えられるわけではない。代わりに、自己対戦で勝利した側にはプラスの報酬が、敗北した側にはマイナスの報酬が与えられる。そして、その結果に基づいて、対局中に選んだ手の確率が調整される。勝ったときによく出現した手は、今後選ばれやすくなるように確率が高められ、負けたときによく出現した手は、確率が下げられていくのである。このような自己対戦は、一回や二回の対局では意味がない。AlphaGo は多数のコンピュータを用いて、1 日に膨大な数の対局を繰り返したと報告されている。人間が一生をかけても経験できないほどの対局経験を、短期間に積み重ねることができるわけである。その結果、「人間の棋譜に基づく上級者レベルの打ち方」を土台としながらも、人間が思いつかなかったような新しい打ち方や、長期的な勝ちやすさを意識した戦略を自ら発見していくことになる。

AlphaGo は、人間の棋譜という既存の知識を起点にしつつ、自身との対局から新しい知識を獲得しており、データから学び、経験からさらに強くなる AI と言える。

3．大規模基盤モデルと生成ＡＩ

「第 4 次 AI ブーム」とも呼べる潮流が生まれ、その中心にあるのが大規模基盤モデル（Foundation Model）である。従来の AI は、「タスク専用のモデル」を一つずつ作るのが基本であった。顔認識をしたければ顔認識用のモデル、翻訳をしたければ翻訳用のモデル、音声認識なら音声認識用のモデル、といった具合である。ところが大規模基盤モデルは、その発想を大きく変える。膨大なテキスト、画像、音声、動画などをまとめて学習しておき、その一つの巨大なモデルをさまざまなタスクに“流用”するのである。

こうしたモデルは、数十億から数兆といった桁のパラメータを持ち、高性能な GPU や専用チップを用いて長時間かけて学習される。ここで重要になるのが「生成 AI（Generative AI）」という概念である。生成 AI とは、与えられた入力や文脈に基づいて、新しいテキスト、画像、音声、プログラムコードなどのコンテンツを自動生成する AI モデルの総称であり、大規模基盤モデルはその代表的な実装形態である。いったん学習が終われば、プロンプトの工夫や少量の追加学習によって、文章生成、翻訳、要約、質問応答、プログラム生成、画像の説明、さらには画像生成や音声合成といった多様な生成タスクをこなすことができる。

このように、一つの大規模基盤モデルが生成 AI として多目的に働き、「何でも相談できる AI アシスタント」のような振る舞いを見せることこそが、第 4 次 AI ブームを支える原動力となっているのである。

(1) Transformer と大規模言語モデル

言語モデルは、「次に来る単語の確率」を学習したモデルであり、「英国の首都は」の後に「東京」が続く確率、「パリ」が続く確率、「ロンドン」が続く確率を計算し、最も確率の高い単語を選び出すのである。言語モデルとして用いられる Transformer は、「自己注意機構（Self-Attention）」と呼ばれる仕組みを用いて、文章中のすべての単語同士が互いに直接参照し合えるようにした。

「この単語は、あの単語と強く関係している」といった類似度を計算し、それをもとに重要な単語に多く“注意”を向けることで、文脈全体を効率的に捉える。これにより、長い文章や複雑な依存関係を持つ文でも、比較的安定して扱うことができるようになった。この Transformer を用いて、Web 上の膨大なテキストを読み込ませたものが、大規模言語モデル(LLM)である。LLM は、「次に来る単語」を予測するという単純な作業を延々と繰り返しながら、文法や語彙だけでなく、世界の一般常識や、さまざまな専門分野の断片的な知識まで取り込んでいく。こうして訓練されたモデルは、ただの「次の一語予測器」でありながら、結果として文章生成、翻訳、要約、質問応答、プログラム生成などを実現した。GPT（Generative Pre-trained Transformer）は、その代表例である。

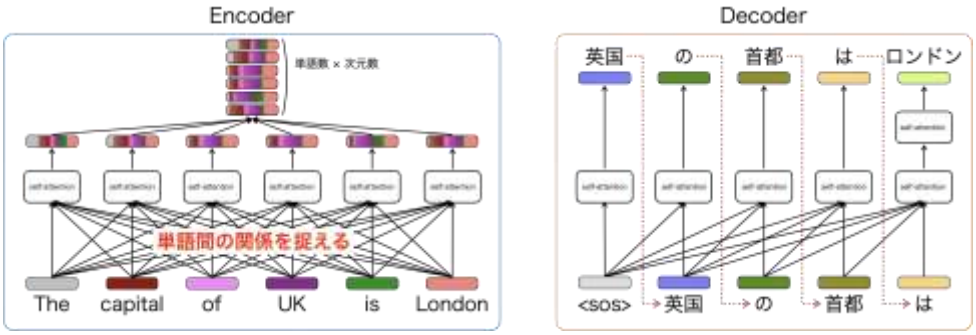

図 3　Transformer

(2) プロンプトと Chain-of-Thought

プロンプトとは、モデルに与える入力文、すなわち指示文のことである。「この文章を要約してください」と、「大学 1 年生にも分かるように、300 字程度で要約してください」とプロンプトにより、出力される文章は大きく変わる。このプロンプトの興味深いテクニックが「Chain-of-Thought（思考の連鎖）」である。これは、「途中の考え方も含めて説明して」と促すプロンプトの書き方である。例えば、ある算数の文章題に対して、「まず問題の条件を整理し、そのあと式を立て、最後に計算して答えを出してください」とすると、モデルは、1.条件を箇条書きに整理し、2. 必要な式を導き、3. 計算を行い、4. 最終的な答えを示す、という“解き方”を文章として出力するようになる。このプロセスを経由することで、モデルが複雑な問題を解きやすくなるのである。

(3) 大規模言語モデルの限界と外部知識の利用

LLM は事前学習のときに大量のテキストを読み込むが、その後に世界で起きた出来事に関するテキストを学習しているわけではない。そのため、ある年までのニュースや論文で学習したモデルは、「その年までの世界」については詳しいが、それ以降の出来事については回答できない。もう一つの限界は、計算や厳密な論理である。言語モデルは「次に来そうなトークン」を予測する仕組みにすぎないので、複雑な数値計算や形式的な証明を内部で厳密に行っているわけではない。そのため、桁の多い掛け算や入り組んだ計算問題に対して、もっともらしいが誤った答えを返すことがある。

こうした弱点を補うために登場したのが、RAG（Retrieval-Augmented Generation）のような枠組みである。RAG は、ユーザーの質問を一度検索モジュールに渡し、企業内文書や Web ページ、論文などから関連する情報を取得する。その検索結果の一部をプロンプトに含めてモデルに入力することで、モデルが本来獲得していない最新の知識や専門的な情報を取り入れて回答することができる。

4. 画像と言語を結びつける VLM（Vision-Language Model）

大規模言語モデルと言語以外のモダリティ（画像や音声など）を組み合わせたモデルが盛んに研究されている。その一つが、画像とテキストを統合的に扱う「VLM（Vision-Language Model）」である。

(2) 画像と言語のアライメントとコントラスト学習

VLM を構築するうえで鍵となるのが、画像とテキストの対応関係（アライメント）を学習することである。典型的な枠組みとして、

- 画像と、その画像を説明するテキスト（キャプション）のペアを大量に用意する。
- 画像から特徴ベクトルを抽出する「画像エンコーダ」と、テキストから特徴ベクトルを抽出する「テキストエンコーダ」を用意する。

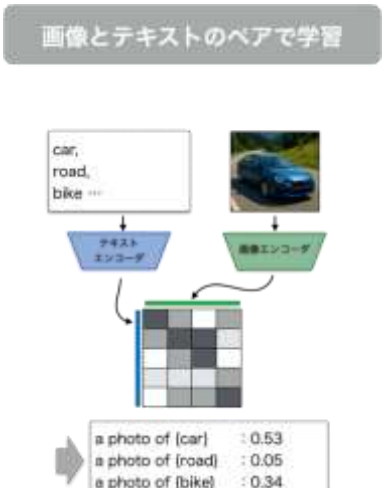

図 4　画像と言語のアライメント

- 同じ内容を表す画像とテキストの特徴が近くなり、異なる内容を表すペアが離れるように、コントラスト学習（対照学習）を行う。

といった手法が用いられる。

これにより、「車の画像」と「car」というテキストがベクトル空間上で近接し、画像だけを見ても「これは車である」という言語的概念にアクセスできるようになる。画像エンコーダには、画像をパッチに分割して Transformer で処理する Vision Transformer（ViT）、テキストエンコーダには、Transformer 型の言語モデルが利用される。

(3) VLM（Vision-Language Model）

こうして学習した画像エンコーダと大規模言語モデル(LLM)を組み合わせることで、VLM を構成することができる。画像特徴を LLM の入力トークンと同じ次元に射影し、LLM の入力列に「視覚トークン」として埋め込むことで、画像と言語を統一的に処理する。このような VLM を用いることで、以下のようなタスクが実現可能となる。

- 画像キャプション生成：画像の内容を自然な文章で説明する。「公園で子どもがボール遊びをしている」など。
- 視覚質問応答（Visual Question Answering; VQA）：画像と質問文を入力し、「この画像の中で赤い物体は何か」「右側にいる人物は何をしているか」といった問いに答える。
- 画像内の関係性推論：単に物体のラベルを認識するだけでなく、「この缶は潰れており、飲み終わった後のゴミである」「このボトルはキャップが閉まっており、中身が残っているのでまだ飲める」といった属性や状態、関係性を推論する。

図 5　VLM

(4) 画像・言語・行動を統合する VLA（Vision-Language-Action Model）

VLM は画像と言語を統合するモデルであるが、近年ではこれに「アクション（行動）」の生成能力を加えた「VLA（Vision-Language-Action Model）」が登場している。VLA は、ロボット制御やエージェント制御への応用を念頭に置いた枠組みであり、「見て理解し、言葉で指示を受け、その指示に基づいて行動する」システムの実現を目指している。VLA の利点の一つは、言語モデルが持つ豊富な世界知識により、学習時に見たことのない物体やタスクの組み合わせにも柔軟に対応しうる点である。

図 6　VLA

例えば、「飲み終わったコーラの缶をゴミ箱に捨ててください」という指示が与えられたとき、VLA（Vision-Language-Action）モデルを使ったロボットは、まずカメラの画像からテーブル上の物体を認識する。「金属の缶が 2 つある」「そのうち 1 つは潰れていて、飲み口が開いている」「もう 1 つは形がきれいで、まだ開封されていない」といった状態を画像から読み取る。同時に、言語モデルは「飲み終わったコーラの缶」という日本語の意味を解釈する。「コーラ缶の中身がもうない」「通常はゴミとして扱うもの」といった世界知識が呼び出される。その結果、「潰れていて飲み口が開いている缶こそが、指示されている対象だ」と判断することができる。続いて、ロボットは潰れた缶の位置を特定し、アームを伸ばして適切な力加減でつかみ、ゴミ箱の位置まで運び、中に落とす動作を生成する。

このように、VLA によって、ロボットは「決められた動きだけをする機械」から、「見て・聞いて・考えて・動く汎用エージェント」へと進化する。カメラで周囲の状況を認識し、人の自然な言葉による指示を理解し、その場その場で行動の手順を自分で組み立てて実行するロボットである。これにより、工場のような決められた環境だけでなく、家庭やオフィス、サービス現場といった予測しにくい状況の中でも、ある程度柔軟に対応できるロボットが現れていく。また、大規模な VLA を共通の「頭脳」として共有し、クラウド経由で新しいスキルや知識をアップデートしていく方向性も強まるだろう。

課　題

1．囲碁 AI である AlphaGo の仕組みを説明しなさい.

2．大規模言語モデルの限界と RAG の役割を説明しなさい.

3．VLM とは何か、どのような応用が可能かを説明しなさい.

4．VLA を用いたロボットは、何かできるかを説明しなさい.