シンクイット Think IT
Think IT White Paper

エンジニアなら知っておきたいAIのキホン

― いま、AIの技術はどこまで来ているのか ―

著＝梅田 弘之（うめだ ひろゆき）

CONTENTS

impress

Think IT
White Paper
02

[第1回]

ガートナー社の「ハイプ・サイクル」に見るAIの「過度な期待」のピーク期と幻滅期

はじめに

2012年に始まった第3のAIブームは、2015年頃から一気に加熱し、2017年頃に「過度な期待のピーク」を迎えました。当時、Think ITで「ビジネスに活用するためのAIを学ぶ」という連載記事を2017年10月～2018年6月にわたって掲載し、また、筆者の会社のHPではAIの技術に関するブログ「AI技術をぱっと理解する(基礎編)」を2018年1月～2018年6月まで掲載しました。そして、これらの内容をベースとした書籍「エンジニアなら知っておきたいAIのキホン」をインプレス社から2019年1月に出版しています。

それから3年が経過した2021年、AIはどこまで実用化され、どこで壁にぶち当たっているのでしょうか。過度な期待はすっかり落ち着きましたが、AIは着実に社会のあちこちに浸透しています。一方で、本当はAIとは呼べない従来型のロジック処理を宣伝のためにAIと謳っている例も多く、実際にどのくらいAIが活用されているのかが見えにくくもなっています。

そこで本記事では「AIのキホン」の続編的なスタンスで、2021年現在のAI技術の実状を解説していきます。

ハイプ・サイクルが示すAIの社会浸透度

「AIがどこまで社会に浸透しているのか」を判断するのに便利なツールがアメリカの調査会社ガートナー社が毎年発表している「ハイプ・サイクル」です。これは、ある技術が登場したときにどのような関心を持たれて社会に浸透するかを表した曲線グラフで、その技術が今どのようなう状態にあるのかを「黎明期」→「過度な期待のピーク期」→「幻滅期」→「啓発期」→「生産性の安定期」という5つの段階に分けて表しています(図1)。

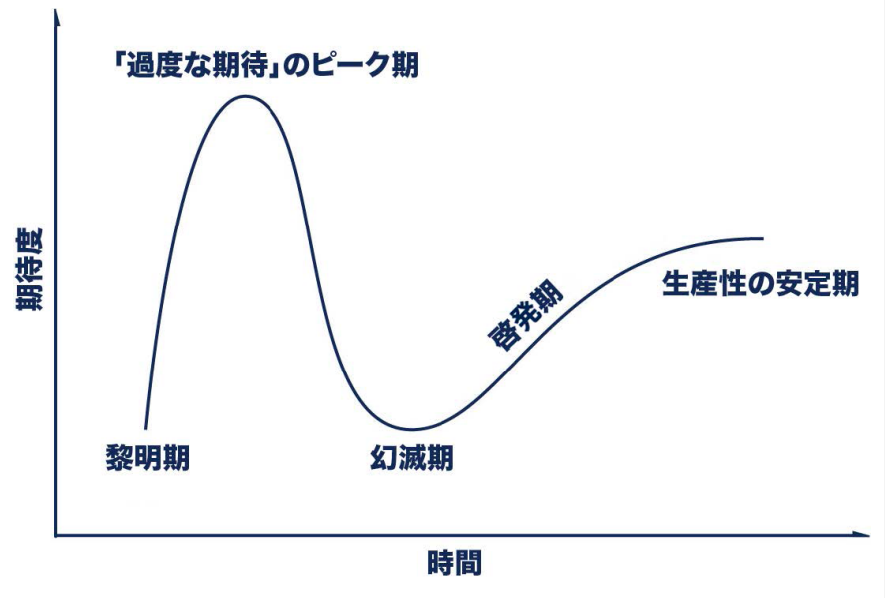

図1：ガートナーのハイプ・サイクル（出典：ガートナージャパン）

ハイプ・サイクルの5段階は、ガートナージャパン社のホームページで次のように定義されていてます。これによると、技術がビジネスに花開くには、いったん過度な期待から幻滅期を経て、ようやく理解が広まっていくことになるようです。

フェーズ	状況
黎明期	潜在的技術革新によって幕が開く。初期の概念実証（(POC)にまつわる話やメディア報道により、大きな注目が集まる。多くの場合、使用可能な製品は存在せず、実用化の可能性は証明されていない
「過度な期待」のピーク期	初期の宣伝では、数多くのサクセスストーリーが紹介されるが、失敗を伴うものも少なくない。行動を起こす企業もあるが、多くはない
幻滅期	実験や実装で成果が出ないため関心は薄れる。テクノロジの創造者らは再編されるか失敗する。生き残ったプロバイダーが早期採用者の満足のいくように自社製品を改善した場合に限り、投資は継続する
啓発期	テクノロジが企業にどのようなメリットをもたらすのかを示す具体的な事例が増え始め、理解が広まる。第2世代と第3世代の製品がテクノロジ・プロバイダーから登場する。パイロットに資金提供する企業は増えるが、保守的な企業は慎重なまま
生産性の安定期	主流採用が始まる。プロバイダーの実行存続性を評価する基準がより明確に定義される。テクノロジの適用可能な範囲と関連性が広がり、投資は確実に回収されつつある

表1：ハイプ・サイクルの5つのフェーズ（出典：ガートナー・ジャパン）

世界と日本のハイプ・サイクル

図2は2020年8月にガートナージャパン社が発表した「先進テクノロジのハイプ・サイクル」です。これはアメリカのガートナー社が発表した世界のトレンドですが、ガートナージャパン社は図3の「日本における未来志向型インフラ・テクノロジのハイプ・サイクル」も発表しています。日本の方はインフラ関連ということで対象が少し違いますが、この2つを見比べると興味深いです。

プロットの記号にも意味があります。安定的に使われるまでに必要な期間として、白丸は2年以内、水色

は2～5年、青色は5～10年、黄色三角は10年以上要する技術であることを示しています。

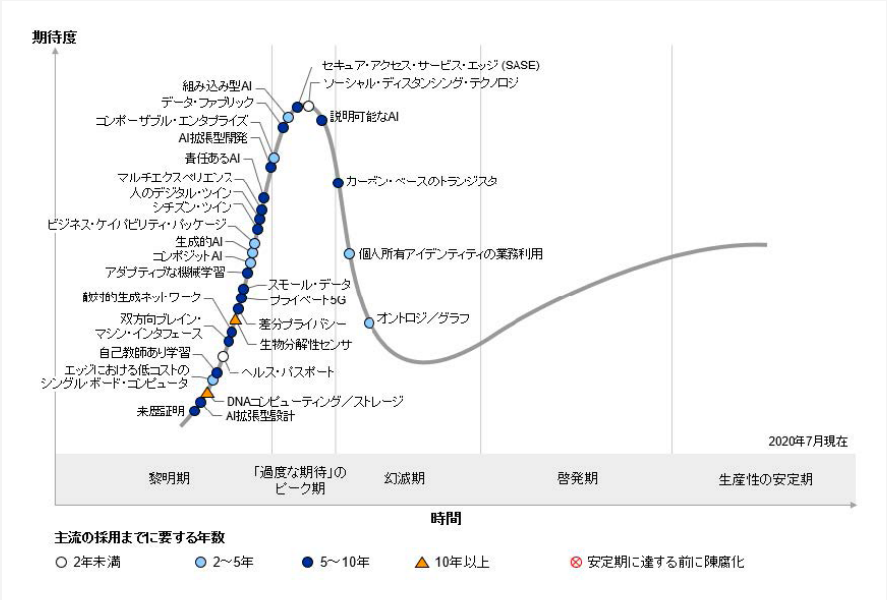

図2：先進テクノロジのハイプ・サイクル2020（出典：ガートナージャパン）

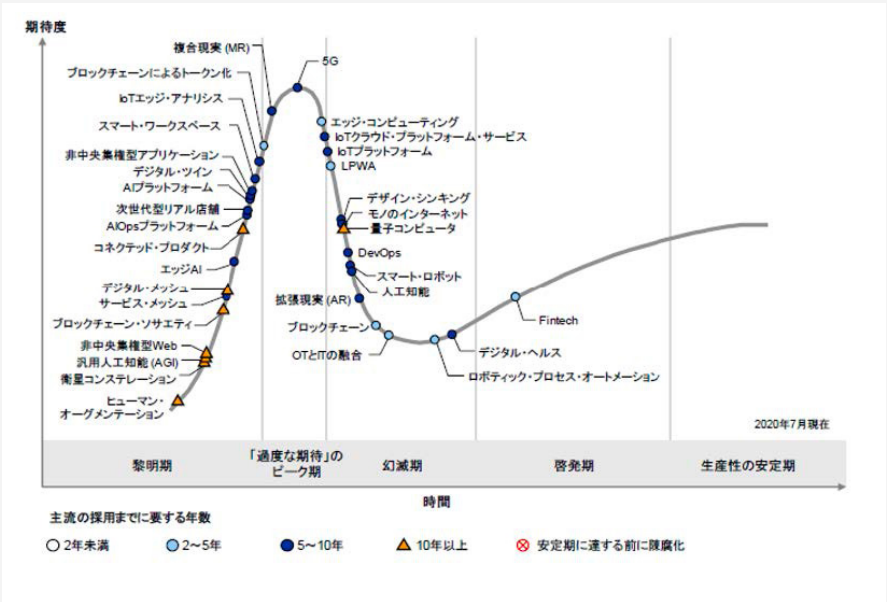

図3：日本における未来志向型インフラ・テクノロジのハイプ・サイクル（出典：ガートナージャパン）

ハイプサイクルに登場するAI技術

2つのハイプ・サイクルの中からAIに関係するものをピックアップすると表2のようになります。登場している技術は「AIのキホン」でも取り上げたものが多く、これらは2017年の段階で少なくとも概念はありました。つまり、2020年のハイプ・サイクルに載るAI技術は、全く新しいものというわけではなく、AIが浸透するにつれ概念が形になってきたものが多いのです。そして、それらの多くが未だに黎明期にとどまっている状況が見て取れます。

	世界	日本
黎明期	自己教師あり学習(Self-Supervised Learning) 敵対的生成ネットワーク(Generative Adversarial Networks) アダプティブな機械学習(Adaptive ML) コンポジットAI(Composite AI) 生成的AI(Generative AI) 責任あるAI(Responsible AI)	汎用人工知能 AIOpsプラットフォーム エッジAI AIプラットフォーム
過度な期待のピーク期	組込型AI(Embedded AI) 説明可能なAI(Explainable AI)	エッジ・コンピューティング
幻滅期		人工知能

表2：世界と日本のハイプ・サイクル(2020年のハイプ・サイクルよりAI関連を抜粋)

日本の方は“インフラ・テクノロジ”と謳っているだけあってプラットフォームやエッジなどインフラ系技術が中心ですが、人工知能という看板キーワードが幻滅期に入っているのが少し気になりますね。

4年間のハイプ・サイクルの変遷

AI技術の停滞感が少し気になるので、本当にそうなのか別の角度で見てみましょう。2017年から2020年までの4年間の「先進テクノロジのハイプ・サイクル」を比較し、この中からAIに関する技術のみピックアップして表3にまとめてみました。

技術要素	黎明期				過度な期待のピーク期			
	2017	2018	2019	2020	2017	2018	2019	2020
汎用人工知能 (Artificial General intelligence)	○	○						
深層強化学習 (Deep Reinforcement Learning)	○							
エッジコンピューティング(Edge Computing)	○							
仮想アシスタント (Virtual Assistants)					○	○		
ディープラーニング (Deep Learning)					○	○		
機械学習(Machine Learning)					○			
コグニティブコンピューティング (Cognitive Computing)					○			
転移学習(Transfer Learning)			○					
会話型AIプラットフォーム (Conversation AI Platform)		○						
説明可能なAI(Explainable AI)							○	○
エッジAI(Edge AI)		○					○	
AI PaaS		○					○	
感情AI(Emotional AI)			○					
敵対的生成ネットワーク(GAN)			○	○				

生成的AI(Generative AI)				○				
AI拡張型設計(AI-Assisted Design)				○				
AI拡張型開発 (AI Augmented Development)				○				
アダプティブな機械学習 (Adaptive ML)			○	○				
自己教師あり学習 (Self-Supervised Learning)				○				
コンポジットAI(Composite AI)				○				
責任あるAI(Responsible AI)				○				
組み込みAI(Embedded AI)								○

表3：4年間のハイプ・サイクルからAI関連を抜粋

この表からは、以下の3つのことが読み取れます。

- a. ハイプ・サイクルにはその時点のトレンド技術が登場
  前年登場した技術の多くは消えており、丁寧にトレンドを追跡するものではない
- b. AI関連技術は、未だに幻滅期に到達していない
  幻滅期を経て啓発期(普及期)になるのだが、まだそこまで到達できていない
- c. 未だに黎明期に登場する技術が引きも切らず
  未だに黎明期に新ピックアップが登場しており、まだまだ活性状態と言える

まとめると、まだまだ「黎明期」「過度な期待のピーク期」を抜け出せていないが、相変わらず注目度は高く勢いは衰えていないという状況が伺えます。

おわりに

日本では(たぶん、世界でも)、なんでもかんでも“AIによる”という枕詞を付けて「すごい技術」を強調する風潮があります。特にTVドラマなどでは顕著で、思わず微笑んでしまいます。しかし、AIビジネスに携わっている立場からすると、これらの多くはAI技術を使っておらず、ロジック処理で行っているか、架空の虚言である実態が透けて見えます。

では、AIは行き詰っているのでしょうか。確かに日本ではそんなムードも漂い始めています。しかし、世界に目を向けると全くそんなことはなく、相変わらずすごいスピードで膨張し続けています。日本ももっと食らいつかなければ差がますます開く、そんな危機感から再び連載を執筆することにしました。よろしくお願いします!

[第2回]

ハイプ・サイクルに登場する技術①
― エッジと組み込み型AI

はじめに

前回は、ガートナー社が発表している「ハイプ・サイクル」を題材に、AI関連技術が今、どのような状況なのかを紹介しました。ハイプ・サイクル上にいろいろなAI技術が乗っかっていましたが、1つ1つが何なのかを知らずに流すのはちょっともったいないです。そこで、今回からは、これらの技術を取り上げて解説していきます。

ただし、これらの技術を断片的に解説するのではなく、その技術が注目されている背景や関連技術も説明し、AIや機械学習の概要を把握してもらいます。まだ、AIについてよくわかっていない人も、この機会にAIのキホンを理解しましょう。

ハイプ・サイクルに登場する主な技術

それでは、早速ハイプ・サイクルに登場する技術を順番に解説していきましょう。まずはAIのキホンを理解するために2017年、2018年の黎明期に登場した技術からです。

(1)AI PaaSとエッジAI(2018黎明期)

2017のハイプ・サイクルにエッジコンピューティングが登場しましたが、2018年にはAI PaaSとエッジAIがそろって黎明期に入っています。これらの技術については、AIの提供形態の変遷の中で理解することにしましょう。

図1は、AIの提供形態が多様化していることを表したものです。今回は、①研究室から②のクラウドサービスを経て③AI PaaSというサービス形態が確立したところまで説明します。AIが社会に浸透するにつれて注目されるようになった④エッジ端末から⑤組み込み型AI、⑥AIチップに関しては次回に解説します。

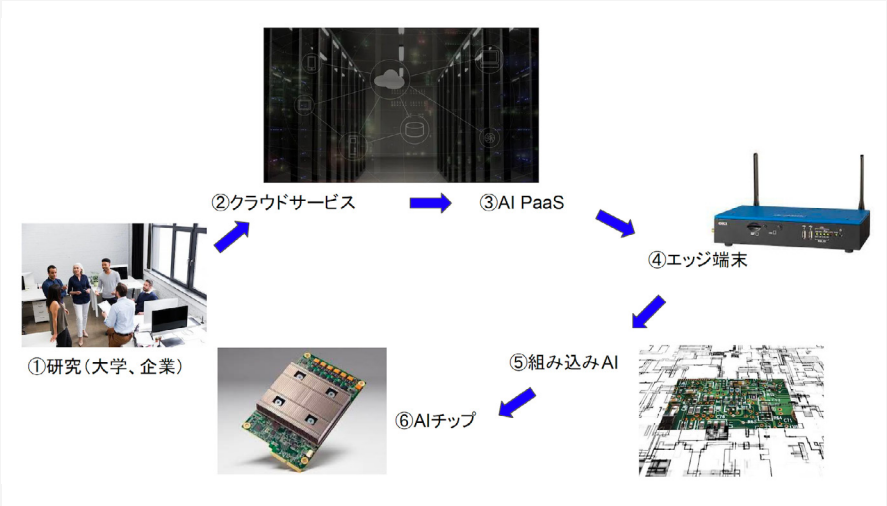

図1：AIの提供形態の多様化（画像出典 エッジ端末：OKI）

①研究（大学、企業）

今回のAIブーム（第3次AIブーム）は、最初は大学や企業の研究部門が主役でした。2012年にトロント大学がディープラーニングを使って画像認識コンテスト（ILSVRC）で圧勝すれば、モントリオール大学がAIライブラリTheano（テアノ）を開発、翌年にはバークレー大学が画像認識用ライブラリCaffe（カフェ）を開発という具合に、技術革新が一気に進みました。

＜＜Note＞＞ILSVRC

ILSVRCは「ImageNet Large Scale Visual Recognition Challenge」の略で、画像認識技術の進歩を促すという目的で行われた世界的な画像認識コンテストです。2010年から始まったのですが、2012年にディープラーニング技術が1位を取ってからはディープラーニング技術同士の戦いになっていきました。毎年、ディープラーニングの階層が深くなったライブラリがチャンピオンを取るようになり、もう役目は終わったということで2017年を最後に終了しています。

企業も、大学教授をスカウトするなどしながら本格参入してきました。トップはGoogleで、2012年に教師なし学習で猫を認識できるAIを発表したのち、2015年にはAIライブラリTensorFlow（テンソーフロー）を公開し、これが現在一番シェアの高いライブラリになっています。このころはライブラリ競争が激しく、FacebookがTorch（トーチ）、日本のPreferred NetworksがChainer（チェイナー）、これらのラッパーとしてKeras（ケラス）などが次々と公開されました。続いてMicrosoftがCognitive Toolkit（CNTK）、FacebookがPytorch（パイトーチ）とCaffe2、中国の百度がPaddle Paddle（パドルパドル）、AmazonがMXNetというライブラリを公開しています。

＜＜Note＞＞AIライブラリ

「AIライブラリ」とは、ディープラーニングを行うために必要な機能を取り揃えたソフトウェアツールです。コンピュータに置き換えると、コンピュータを操作するさまざまな基本機能を持っているOSのような役目です。モデルや階層、重み、閾値、バイアスなどをパラメータ設定するだけで学習できる仕組みを持っているので、ユーザーはこれを使うことで、簡単にディープラーニングを学習できます。

②クラウドサービス

今回のAIブームは、ちょうどクラウド時代に花開いたことが特徴です。いろいろな分野でディープラーニングの成果が出てくると、大手クラウドベンダー各社はクラウドに機械学習環境(プラットフォーム)を提供するようになりました。

まず、2015年にMicrosoftがMicrosoft Azure Machine Learning、AmazonがAmazon Machine Learningというクラウド上で機械学習を行う環境を提供しました。2016年にはGoogleもGoogle Cloud Machine Learningを発表し、OracleはIntelligent Applications、SalesforceはEinsteinというクラウドのプラットフォームを発表しています。さらに2017年にはNVIDIAがNVIDIA GPU Cloud、テンセントが智能雲を発表しています。

同時に、これらのクラウドはAIライブラリで学習したAI(学習済モデル)をサービスとして提供するようになりました。その1つがGoogle翻訳です。かねてからGoogle翻訳は一定の評価を得ていましたが、2016年にGoogleがディープラーニング技術に切り換えたとたんに飛躍的に翻訳精度が向上して驚嘆されました。

③AI PaaS

AIサービスが増えるにつれ、単発ではなく総合的なAIプラットホーム(AI PaaS)としてサービス提供するようになりました。これはAIサービスだけというよりも、クラウドサービスの中にAI関連のサービスが増えていった結果、1ジャンルとしてAI関連サービス群ができた形態です。

例えば、Google Cloud Platform(GCP)には、Cloud SQL(データベース)やBigQuery(データウェアハウス)、Dataflow(ストリーム処理)など100以上のサービスがあります。この中からAIや機械学習に関連したものだけをピックアップし、それをAI Building BlocksとVertex AI(旧AI Platform)に分類し、さらにAI Building Blocksを学習済AIサービスとカスタム学習(AUTOML)に分けています(図2)。

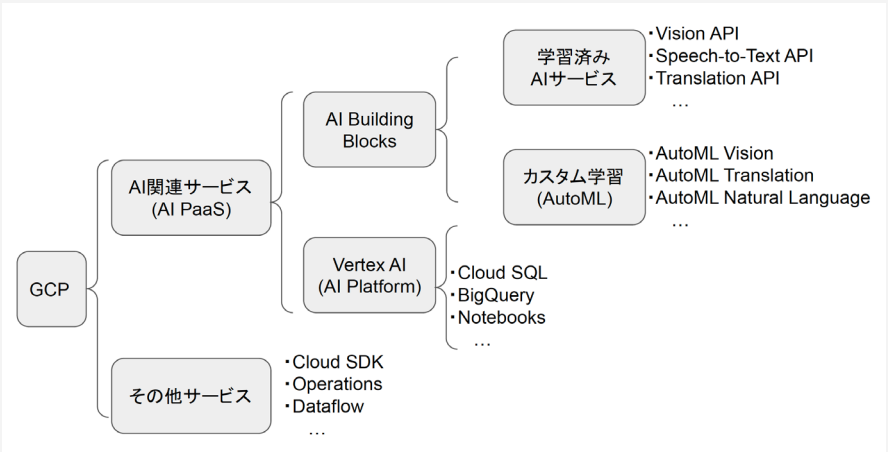

図2:Google Cloud Pratform(GCP)のAI関連サービス

表1に現在(執筆時点)のAI Building Blocksの各サービスを示します。学習済みAIサービスはAPIの形式で提供され、ユーザーはこれらのAPIから画像認識や翻訳などのサービスをそのまま利用します。3年前にビジネスに活用するためのAIを学ぶを執筆した際はこのようなサービス提供が主流だったのですが、現在ではユーザーが用途に合わせて学習できる仕組みを提供するAutoML(カスタム機械学

習）サービスが拡充されてきました。

1. 学習済みAIのサービス	
画像/動画分析	・VIsion API（画像認識、オブジェクト検出） ・Vision Intelligence API（動画処理・分析）
音声認識/音声合成	・Cloud Speech-to-Text API（音声認識と音声文字変換） ・Cloud Text-to-Speech（音声合成）
機械翻訳	・Translation API（機械翻訳） ・Media Translation API（ファイルやストリームの翻訳）β
自然言語処理	・Natural Language API（自然言語処理） ・Natural Language API（医療向け自然言語処理） ・Dialogflow（チャットBot作成）
構造化データ	・Cloud Inference API（時系列データの相関分析）α ・Recommendations AI（レコメンデーション）
2. AutoML（カスタム機械学習）	
画像/動画分析	・AutoML Vision（画像認識モデルをトレーニング） ・AutoML Vision Intelligence API（動画処理モデルをトレーニング）β
機械翻訳	・AutoML Translation（翻訳モデルをトレーニング）
自然言語処理	・AutoML Natural Language（言語処理モデルをトレーニング）
構造化データ	・AutoML Tables（表形式データからMLモデルを自動的に構築）β

表1：AI Building Blocksのサービス
[注]α、βはアルファ版、ベータ版提供を示す

一方、Vertex AIは、さらに本格的にトレーニングするために必要なリソースを取り揃えた学習環境です。もともとはAI Platformという名前で提供されていたのですが、2021年5月に新しい機械学習プラットフォームとしてVertex AIを発表し、AI Platformはその中に組み込まれる形となりました。

Vertex AIは、AIモデルの作成作業をData Readiness(データの準備)からModel Management(AIモデル管理)までのユースケースに整理して、各プロセスにおいて必要となるAIサービスを総合的に提供します(図3、表2)。

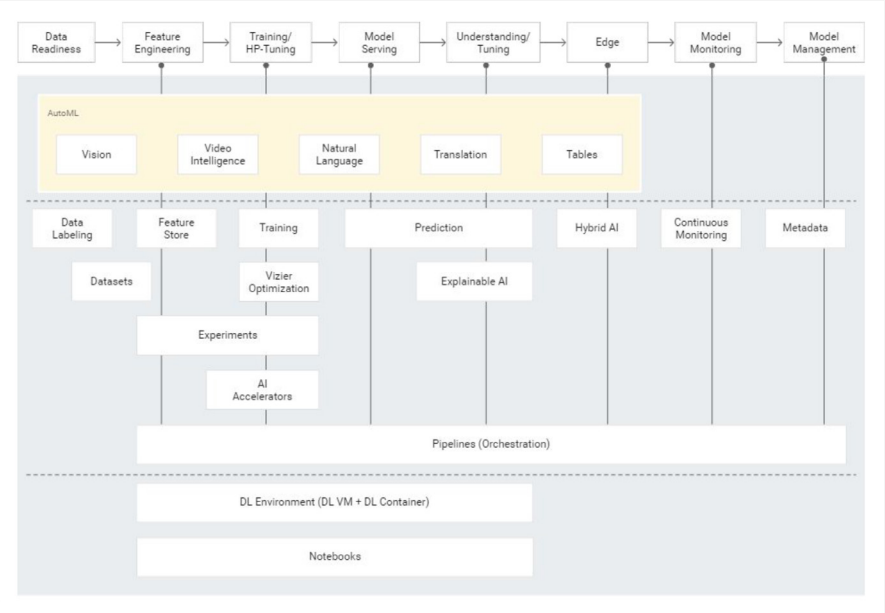

図3：Vertex（出典：GoogleのHP）

ユースケース	サービス
Data Reediness データの準備	BigQueryやCloud Storageからデータを取り込み、Data Labelingを活用してアノテーション（ラベル付け）を行えます。
Feature Engineering特徴量エンジニアリング	Feature Engineering（特徴量エンジニアリング）とは、データから機械学習モデルに役立つ変数（特徴量）を作成する作業のこと。Feature Storeは、その作業に必要な機能を提供するリポジトリで、Experimentsを使って効率的にモデルの選択を行えます。
Training/HP-Tuningトレーニングとハイパーパラメータ調整	Notebooksは、Pythonなどの言語を使ってAIモデルを構築できる総合開発環境。Trainingを使ってAIモデルのトレーニングを行う際に、Vizierを利用してハイパーパラメータを最適化します。
Model Servingモデルの提供	Predictionを使用すると、HTTP経由でオンライン サービス提供したり、AIモデルを本番環境にデプロイできます。
Understanding/Tuningモデルの調整と理解	Explainable AI（説明可能なAI）は、AIモデルがどうしてその結論を出したかを理解するためのサービスで、各特徴量が予測にとってどの程度重要としたかを見える化できます。
Edgeエッジ	Edge Managerは、トレーニング済のAIモデルをエッジコンピュータにデプロイしたり、監視したりできるAPIを用意しています。
Model Monitoringモデルのモニタリング	デプロイされたAIモデルは、継続的にモニタリングされ、アラートや偏差の原因診断、トレーニング データの収集などが行えます。
Model Managementモデルの管理	Pipelines は、機械学習ワークフローを管理するためのツール。Metadataを使用すると、Pipelines内のすべてのコンポーネントの入出力が追跡され、監査とガバナンスが容易になります。

表2：Vertex AIのユースケースとサービス内容

おわりに

今回は、Google Cloud Platformを題材に、AI PaaSでどのようなAIサービスを提供しているかをイメージしてもらいました。3年前は画像認識や自然言語処理など用途別の学習済モデルの提供が中

心でしたが、AIをお試しではなく、本格的に利用するためにはユーザー自身が目的に合わせて学習する必要があります。そうしたニーズに対応して、ユーザーが簡単に機械学習できるAutoMLが、まだβ版が多いもののラインナップされてきています。

一方で、AutoMLは未だ進化の途中であるため、ユーザーはやはりAIライブラリを使って自身で機械学習してモデルを作成する必要があります。そのための環境を体系的にまとめたVertex AIにより、ユーザーは高度なAIモデルを効率的にトレーニングしてデプロイできるようになりました。

さて、このように学習環境も充実してきたわけですが、学習されたAIがどのような形で社会に利用・浸透されつつあるのか、次回のエッジ、組込みAI、AIチップでそのあたりを解説します。

[第3回]

ハイプ・サイクルに登場する技術②
― エッジAIや組み込みAI、AIチップ

はじめに

前回は、AI PaaSの代表としてGoogle Cloud Platform(GCP)を取り上げ、AIモデルの学習環境と学習済モデルをクラウドサービスとして提供する仕組みを共有しました。データの準備からAIモデルの学習(トレーニング)、モデルの評価といったAIモデル作成を支援する機能だけでなく、学習済モデルをEdge(エッジ)コンピュータにデプロイしたり、監視・モニタリングするなどの管理機能まで充実しているクラウドサービスの状況が理解できたと思います。

今回は、クラウド側ではなく、身近な機器側でAIを利用する形態として「エッジAI」「組み込みAI」「AIチップ」について学びましょう。

エッジコンピューティング / エッジAI

ハイプ・サイクルでは、2017年にエッジコンピューティングが黎明期に登場しています。そして2018年にエッジAIが黎明期に入り、2019年に「過度な期待のピーク期」に移行しています。ここで使われるエッジ(Edge)とは端(はし)や縁(ふち)の意味です。これは、クラウドを中心に発達してきたAIを、クラウドから見て末端にあるIoT機器やスマートデバイス内に置いて利用することからのネーミングです。人間から見ると一番近い側なのでちょっと違和感もありますが、まあ、日本もアメリカから見ればFar East(極東)なので、何事も起点によるということですね。

AI PaaSによるクラウドコンピューティングが発展する一方で、なぜ、エッジコンピュータが脚光を浴びつつ広まったのでしょうか。その原因の1つはIoTの普及です。人間が日本語の文章を送って、クラウドで英語に翻訳して返すというような使い方であれば、AI PaaSでも問題なく処理できます。ところがIoT端末のように時々刻々とデータを送信し続けるような場合、同じようにクラウドで処理して結果を返す方法だといろいろと支障が出てきます。

工場で流れてくる製品の外観検査(異常検知)を例に説明しましょう(図1)。上部のクラウドサービスは、前回取り上げたGCPのようなAI PaaSを想定してもらえば良いでしょう。左の図はAIモデルの機械学習を行うだけでなく判定処理(推論)までクラウドで行うモデルです。マシンビジョンカメラで連続撮影されている製品の外観画像が時々刻々クラウドに送られて異常の有無を判定し、異常と判定した場合に現場にアラームを出すようなイメージです。

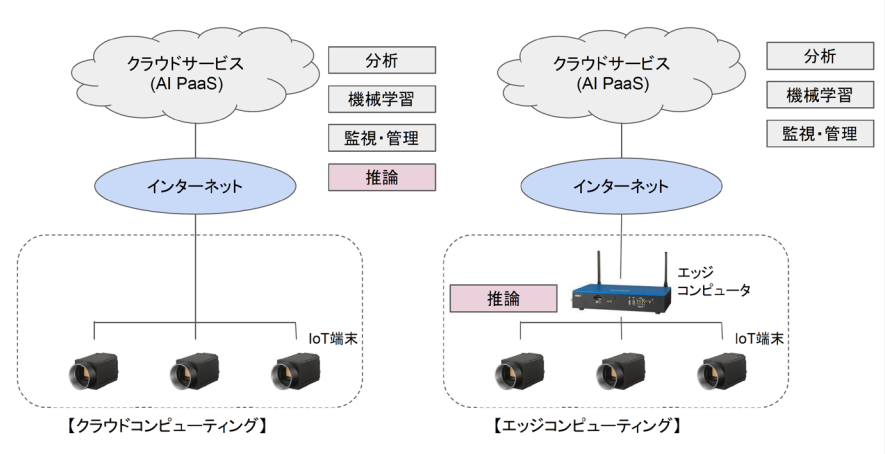

図1：クラウドコンピューティングとエッジコンピューティング（画像出典 エッジコンピュータ：OKI IoT端末：Vision Systems Design）

この方法はインターネットを経由してクラウド上で判定処理するため、結果のフィードバックが遅延して異常製品の除去が間に合わない恐れが出てきます。また、インターネットを流れるトラフィック（通信量）やクラウドの処理量が膨大になりコストもかさみます。さらにインターネットは不安定なため、万一ネットが止まったら製造に支障をきたしてしまいます。製品の異常箇所や状態は極秘情報なのでクラウドにあげたくないという企業もあるでしょう。

整理すると、次のような不都合があるわけです。

- インターネット経由で処理されるため、リアルタイム性が劣る
- 通信量や処理量が膨大になり、通信量やクラウド利用料が大幅に増える
- インターネットの接続が不安定になると処理がストップする
- セキュリティの理由で、インターネット上に異常画像を送りたくない

これらの問題を解決するのが図1右のエッジコンピューティングです。こちらのモデルでは、クラウドで学習したAIモデルを現場側のエッジコンピュータに格納（デプロイ）し、ここで送られてくる画像を判定処理（推論）します。このモデルであれば、インターネット上に大量データを流す必要がなく、現場で判定処理するためリアルタイム性も確保できます。

エッジコンピューティングの場合でも、クラウドにはクラウドの利点もあるため、通常はクラウドとエッジとで役割を分担したシステム構成とします。例えば、私の会社ではディープラーニングを使った異常検知システム「AISIA-AD」という製品を出していますが、AIモデルの学習と分析、監視・管理はクラウド（Microsoft Azure Cloud）で行い、学習されたモデルはAzure IoT Edgeでエッジコンピュータにデプロイしています。

Embedded AI（組込みAI）

異常検知などで使うエッジ端末は、CPUやGPUを搭載したコンピュータに学習済AIをデプロイしたものです。ここからさらに一歩進んで、デバイスレベルにAIを搭載したものが組込みAIで、2020年のハイプ・サイクルで「過度な期待のピーク期」に登場しています。組込みAIを使うとAIの分析や処理はデバイスレベルで実行され、実行結果をもとに直接デバイス付近でアクションできます。

ここ数年、AIとIoTが発展・普及する中で、AIの処理はクラウドから現場(エッジ)やデバイスに領域を拡大しています。AIとIoTを組み合わせたThe Artificial Intelligence of Things(AIoT)という言葉も生まれており、デバイスにAIが入るのが当たり前になりつつあります。例えば、自動運転車はAIでリアルタイムに画像処理をして運転していますし、ロボットもAIでリアルタイムに判断して動作します。もっと身近なところではスマホの顔認証やスマートスピーカー、ドローン、セキュリティカメラなど、さまざまなモノに組込まれたAIが脚光を浴びているのです。

組込みAIが「過度な期待のピーク期」に登場したのは、Edge AIと同じく次のようなメリットがあるからです。

- インターネットや中央コンピュータを介さずローカルで処理するため、自動運転車や安全システムのような応答性の必要な処理に利用できる
- センサーやデバイスで収集したデータをインターネットや中央コンピュータに送信しないため、ネットワークトラフィックを削減できる
- インターネットやネットワークが不安定になっても動作を継続でき、ネットが再接続された際に必要なデータを送信できる
- 顔認証などの個人データや重要なデータをクラウドや中央コンピュータに送信する必要がないため、セキュリティが強化され、安全性が高まる
- デバイス自身の障害を自己診断したり、消費電力を抑制するなど制御を改善したりできる。また、音声認識や自然言語処理、顔認証、モーション(ジェスチャー)認識を組み込んだりして、新しい機能を実現して価値を高められる

いつの間にか身の回りのいろいろなデバイスにAIが入っている、私たちはそんな時代に直面しています。そして、この組み込みAIのコアとなる部分がAIチップ化されているのです。

AIチップ

AIチップとは、AIに特化した半導体チップのことで、一般にAIアクセラレータ(Accelerator)とも呼ばれています。

GPU

AIで使われる半導体チップと言って、まず思い浮かぶのがGPU(Graphics Process Unit)ですね。NVIDIA(エヌビディア)社は、以前からグラフィックボードなどで使われていたGPUをディープラーニングの画像処理に応用することに成功し、一気に世界的なAIチップメーカーとして名を馳せています。

AIの画像処理は、処理内容は単純なものの処理量が膨大という特徴があります。これを汎用的な処理ができるCPUで行うこともできますが、シンプルな処理を大量に行えるGPUを使った方が効率良く実行できるため一気に利用が広まりました(表1)。

	CPU	GPU
主な用途	中央演算処理装置	画像や動画処理、3D、CAD、AI
処理内容	汎用な処理向き(if ～ else ～が得意)	シンプルな処理向き(for ～ loopが得意)
並列処理	1個当たり数コア	1個当たり数十～数千コア
利用形態	単独で利用できる	CPUと一緒に利用

表1:CPUとGPU

FPGA

AIチップはGPUだけではありません。エッジコンピュータやロボット、自動運転車などに組み込んで処理(推論)できるAIチップには、GPU以外にFPGAとASICやSoCなどがあります。

FPGA(Field-Programmable Gate Array)は、“現場でプログラミングできるゲートアレイ”という名称の通り、目的に合わせてIC(集積回路)の内部ロジックを作りこめるカスタムICです。ディープラーニングのフレームワークは進化を続けているため、FPGAのような再構成可能なデバイスを推論に用いることで柔軟に対応できるメリットがあります。

ASIC

一方、ASIC(Application Specific Integrated Circuit)は、“特定用途向け集積回路”という名称の通り、特定用途向けに製造されたカスタムチップです。FPGAは汎用品なため1種類でいろいろな用途に使えますが、ASICは1つの用途専用に作られたオーダーメイド品です。専用チップのため何倍もの効率性が得られ、消費電力も小さいため、GoogleやFacebook、Amazonなどさまざまな企業が独自モデルの設計が確定したときにASICを大量生産しています(表2)。

	FPGA	ASIC
ロジックの書き換え	ロジックの再プログラム可能	ロジックの再設定不可
コスト	ASICより高い	大量生産により低コスト
ロジックの開発コスト	汎用品なので初期コスト低い	設計やテストなどの負荷高い
リスク	再プログラミング可能なので仕様変更や設計ミスに強い	仕様変更や設計ミスが発覚すると作り直しとなる
消費電力	GPUより低い	FPGAより低くできる
生産ロット	小ロット	大量ロット

表2:FPGAとASIC

Googleが囲碁AIのAlphaGoやストリートビューなどに利用したTPU(Tensor Procesing Unit)というチップは、TensorFlowライブラリ用に特別に設計されたASICです。2016年に発表されて以来、2017年、2018年、2021年に改良版が出されており、2018にはエッジでの推論に向けて小型化、省電力化したEdge TPUも出されています。

機械学習の2つのプロセス

AIチップは機械学習の2つのプロセスで利用されています。1つが学習プロセスで、AIモデルが十分な能力を発揮できるまでトレーニングします。もう1つが推論プロセスで、合格レベルに達した学習済モデルを現場に用いて目的の用途に利用します。

図1右の異常検知のケースで説明しましょう。まずクラウドで大量の画像を用いて機械学習(トレーニング)を行います。そして目標とする精度以上に見分けられるようになったら、その学習済モデルをエッジコンピュータにデプロイして、製造ラインを流れてくる製品の正常/異常を推論するわけです。

ところで、この2つのプロセスのうち、学習プロセスと推論プロセスのどちらが負荷が高いでしょうか。答えは学習プロセスです。何千、何万ものデータを使って繰返しトレーニングする学習プロセスに対して、学習済プロセスが推論する負荷ははるかに軽い処理です。英語の勉強に置き換えてみてください。英語の学習に費やす膨大な時間と比べて、英語ができるようになった人が英文を読み解く方がずっと簡

単だと思い当たるでしょう。

AIチップの２つのプロセス

AIチップには学習用と推論用の２種類あります。クラウドやデータセンターで学習に使われるAIチップは主に高性能なGPUが使われており、NVIDIAが強い領域です。一方、推論に使われるAIチップ(推論チップ)は、学習で使われるチップと違って小型軽量化、省電力化、低コスト化が可能で、GPU、FPGA、ASIC、SoCなどいろいろなAIチップが主導権を争っています。

スマホやカメラなど端末デバイスに組み込まれる推論チップはSmall Edgeとも呼ばれ、サイズや消費電力を小さくすることが目的です。アップルが2017年に発表したiPhoneXで顔認証が採用されましたが、ここでもパッと認証するために自社開発したAIチップを導入しています。また、スマホのカメラが画像を良い感じに処理する用途にもAIチップが使われています。

一方、自動運転車やロボットなどに組込まれるチップは、リアルタイム性やコンプライアンス(命令に応じる)などを目的に利用されています。NVIDIAやMobileeye(モービルアイ)、AImotive(AIモーティブ)など、さまざまな企業が自動運転車用AIチップを開発していますが、テスラのように自動車メーカーが独自でAIチップを開発しているケースもあります。

SoC

スマホに搭載されているのは、スマホ用SoC(System on Chip)です。GPUやFPGA、ASICがアクセラレータとしてCPUと一緒に使われるのに対し、SoCは必要な要素がすべて単体で組み込まれているチップでCPUやGPU、メモリなどもワンチップに含まれています。

例えば、NVIDIAは2015年に自動運転車のためのGPUアーキテクチャを使ったNVIDIA DRIVE PXというAIチップを発表していましたが、今ではSoCとしてNVIDIA DRIVE AGX Orinというチップを提供しています。テスラが独自開発したFSDチップもSoCで、CPUやGPU、AIアクセラレータなどを1チップに集積しています。

AIチップの利用イメージ

AIチップの利用イメージをエッジコンピューティングによる異常検知をモデルに考えてみましょう。

最も汎用的なのはCPUベースのエッジコンピュータで、クラウドで学習した学習済みモデルを搭載して正常/異常を推論します。CPUだけでなくGPUという高速処理できるアクセラレータを搭載すると、さらに高パフォーマンスで推論できます。

GPUの代わりにFPGAを搭載すれば、推論モデルの変化に現場で対応できます。PGAは画像処理用のGPUと違ってAI専用に作られているため、省電力で効率的に推論処理が行えます。そして、AIモデルが確立したならば、それを専用チップのASICとして生産することで、さらに省電力で効率化を実現できます。

ただし、ASICの設計・開発は大変なので、検査対象製品が変わる度にチップを作るのは手間がかかります。これを補うために、例えばGyrfalcon Technology(ジルファコン テクノロジー)社のようなAIチップ開発ツールを提供する会社も出てきています(ソフトウェアの世界のSDKのようなものですね)。

おわりに

社会のいたるところにAIが浸透している未来を思い浮かべてみてください。その光景のいたるところでAIチップが使われていることになります。この膨大な市場のため、世界中のビッグカンパニー、スタートアップ、ユーザー企業がAIチップの分野に参入して最先端技術を競い合っているのです。

日本でも大手やベンチャーなど、さまざまな企業がAIチップに参入しています。また、経済産業省も「AIチップ・次世代コンピューティングの技術開発事業」を打ち出し、NEDO(新エネルギー産業技術総合開発機構)が「AIチップ開発加速のためのイノベーション推進事業」という助成金を出しています。AIでは世界に遅れをとってしまいましたが、AIチップで巻き返しできるでしょうか。う〜ん、期待したいところです。

[第4回]

ハイプ・サイクルに登場する技術③ ― ディープラーニングの基礎技術

はじめに

前回まで、2回に渡ってAIの利用形態が研究室からAI PaaS、そしてエッジAI、組込みAI、AIチップへと、より身近なところに広がり、AIがいたるところに存在する社会になりつつある状況をお伝えしました。今回は、引き続きハイプ・サイクルに登場するAI技術とその関連技術を取り上げて、AIの概要と変遷を理解していきましょう。

汎用人工知能 (AGI：Artificial General intelligence)

汎用人工知能は、2017年と2018年にハイプ・サイクルの黎明期に登場したキーワードです。この頃はAIが将来、人間を超えて脅威の存在になるという論調も多く聞かれており、「強いAI」と「弱いAI」という言葉もよく使われていました。囲碁や自動運転車など特定の分野にのみ力を発揮する特化型AIを「弱いA」と称し、人間のできるあらゆる知的作業をこなせる万能型AIが「強いAI」と呼ばれるものです。

この「強いAI」が汎用人工知能(AGI)です。私は「さすがにそんなの無理!」と思っていましたが、実際、AIの実用化模索が進む中で「過度の期待」が現実レベルに落ちたため、最近ではあまり聞かれないキーワードになってきました。ただ、個々の「弱いAI」が次々と実用レベルになるにつれ、それらを統集合して結果的に「強いAI」に見えるAIが登場する日は来るかも知れないと感じています。

深層強化学習 (Deep Reinforcement Learning)

2017年の黎明期に取り上げられた強化学習は、エージェントが報酬を最大化するために試行錯誤を繰り返して学習する手法で、ディープラーニング技術を使ったものが深層強化学習です。2016年に囲碁で韓国のトップ棋士に勝って一躍有名になったAIエージェント(ALPHA GO)は、深層強化学習によって鍛えられたものです。

深層強化学習は、試行錯誤により正しい行動を見つけ出すアプローチで、何百万回と繰り返し学習できるケースに向いたアルゴリズムです。例えば囲碁などではAI同士を果てしなく戦わせて、どの局面でどういう手を打ったら勝利(報酬)を得られるかを学びます。自動運転車のドライビング技術習得にも使われており、ソフトシミュレータで死ぬほど(AIは不死身ですが)運転を繰り返し、無事故(報酬)で運転する技術を身につけます。

転移学習(Transfer Learning)と ファイン・チューニング(Fine Tuning)

2019年の黎明期に登場した転移学習は、少ないデータで学習する方法の1つです。これを説明する前に、そもそもディープラーニングがどういうもので、なぜ学習することによって賢くなってゆくかという原点から順を追って説明しましょう。

(1) ニューロンとシナプス

人間の脳は、図1のようにニューロンと呼ばれる神経細胞が無数にあり、ニューロンからシナプス(樹状突起)を介して次のニューロンにインパルス(電気信号)が伝わります。ニューロンからニューロンへ情報が伝達することをスパイク(発火)と呼びますが、伝達のしやすさ(結合強度)は1つずつ異なっています。

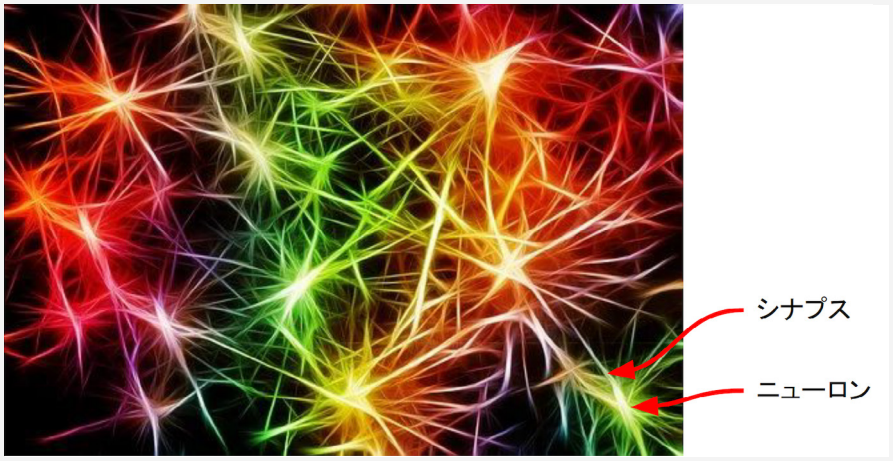

図1:ニューロンとシナプス

(2) ニューラルネットワーク

この脳の構造をモデル化したものがニューラルネットワークです。図2のようにニューロンを模した人工ニューロンを行列演算しやすいように層に並べて配置した構造で、層を数十から数百と増やしたものを深層学習と呼びます。一般に層が増えれば増えるほど推論精度は高まりますが、その分、学習に時間がかかります。シナプスと同じように人工ニューロンから人工ニューロンへ信号が伝わりやすさは1つずつ異なっていて、これをWeight(重み)と呼んでいます(図では線の太さで表しています)。

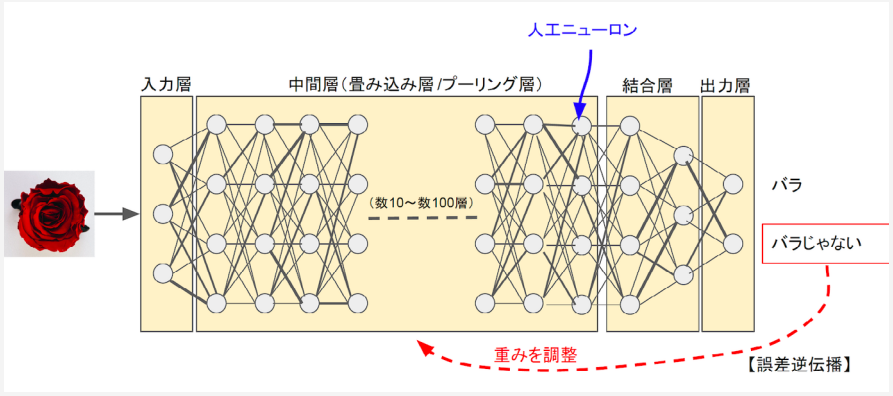

図2:ニューラルネットワークと誤差逆伝播

(3) 誤差逆伝播

ところで、人工知能はどうやって頭が良くなっていくのでしょう。その基本原理が誤差逆伝播(ごさぎゃくでんぱ)です。例えば、図2は出力が2つの分類モデルで、バラかバラじゃないかを判定するだけのAIです。学習データとしていろいろな花の写真を5000枚用意し、その中にバラの写真が1000枚入っているとしましょう。ランダムに100枚ずつ抽出してAIを繰り返しトレーニングするのですが、最初のうちAIはバラをバラじゃないと判定したり、バラじゃない花をバラと判定したりします。

このトレーニングにおいて、間違った場合には重みを調整します。間違っては重みを調整、間違っては重みを調整、という処理をずっと繰り返すことにより、AIはバラを判定するのに最適な重み配置となります。これがAIが頭が良くなる仕組みです。間違い(誤差)を逆方法に伝播するので、誤差逆伝播(Backpropagation)と呼びます。

バラの代わりに製品の正常画像と異常画像を用意して学習し、異常か異常じゃないかを判定できるモデルを作るのが異常検知AIです。出力は2値とは限りません。例えば出力を500に増やしてバラや牡丹、菊など500種類の花を学習させれば、もっと多くの花を判別できるAIになります。異常検知においても異常を「キズ」「サイズ規格外」「色ムラ」「異物混入」など状態ごとに分類すれば異常発生原因の特定に役立ちますが、その分、学習の難易度は高くなります。

(4) 転移学習

このように、ディープラーニングでは大量データを繰り返し学習することでニューラルネットワークの重み調整を行い賢くなります。しかし、大量データを用意して、繰り返し学習するのは容易ではありません。そこで、すでに別のところで訓練した学習済AIを持ってきて、大部分の重みバランスを凍結した状態で追加学習することで、少ないデータで短期間に学習できる転移学習(Transfer Learning)というテクニックが登場しました。

図3は、バラを判別できる学習済AIモデルを使って牡丹を見分けるAIを作るものです。通常の機械学習がイチから学習するのに対し、ここでは中間層の大部分の重みを凍結し、残りの少ない層のみで学習します。もともとバラを判別するのに適した重みとなっているので、似たような判別ならその重みバランスを利用して、少ないデータ、少ない回数で学習できます。

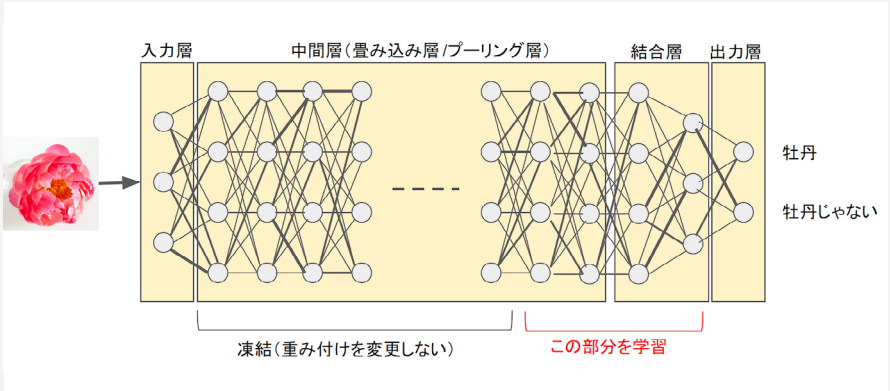

図3:転移学習

(4) ファインチューニング

転移学習は学習済モデルの重みを固定して学習していますが、類似性がそれほど高くないモデルを作る場合は逆にうまくいかない場合があります(Negative Transfer:負の転移)。そこで登場したのが

ファインチューニング(Fine Tuning)です。これは学習済モデルの重みを初期値として使い、そこから再学習をして重みを調整していく方法です。

(5) 蒸留

知識の蒸留(Knowledge Distillation)についても説明しておきましょう。ディープラーニングは階層が深くてパラメータ数も多いモデルのほうが精度が高くなります。例えば、2010年から毎年行われた画像認識コンテスト(ILSVRC)の優勝モデルは、2012年は8層のCNN(畳み込みニューラルネットワーク)でしたが、2014年は22層、2015年152層、2016年は200超という具合に層が深くなっています。

ただし、階層が深くなればなるほど計算コストや消費電力が大きくなります。コンテストのような特殊目的なら良いのですが、実用には不向きであり、ILSVRCも本来の目的は達成したとの判断で2017年を持って終了しています。

前回、機械学習には学習プロセスと推論プロセスがあり、スマホなどに搭載する推論チップは小型軽量化、省電力化、低コスト化が必要と述べました。そのようなニーズに対応するために、層の深いモデルで得られた知識を軽量モデルの学習に利用することで、軽量なのに高い精度を得られる蒸留というテクニックが考えられました。

蒸留は、図4のようにTeacherが学習して得られた知識(Knowledge)をStudentの学習に使います。知識をそのまま適用して完成とする方法もありますが、通常は知識を適用した後で追加学習を行います。何を知識とするかはいろいろなモデルが次々と考えられていて、AIの実用化を背景に蒸留はホットなテーマの1つです。転移学習やファインチューニングが同じモデルなのに対し、蒸留は重量モデルから軽量モデルに適用する技の総称で、具体的な方法には多数があると覚えておいてください。

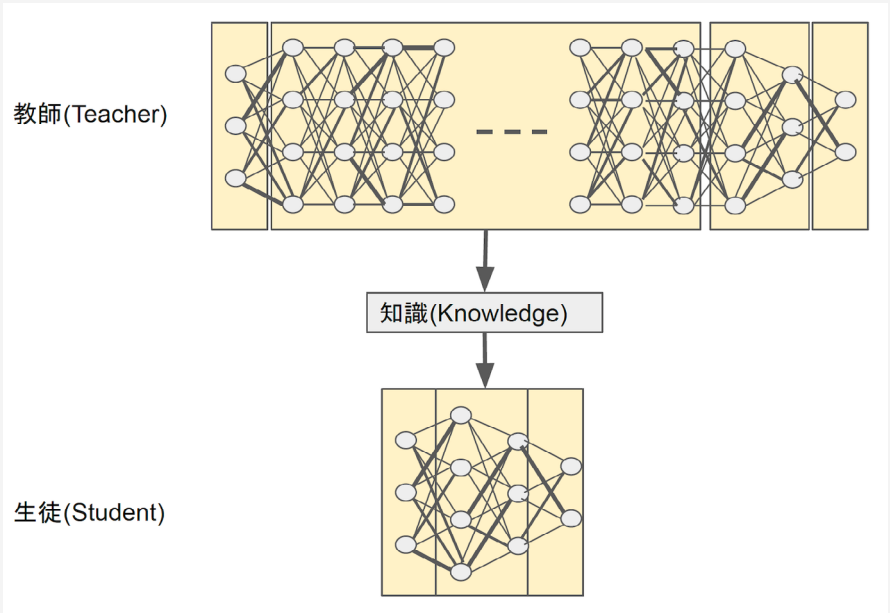

図4：蒸留(Knowledge Distillation)

説明可能なAI(Explainable AI)

説明可能なAIは、2019年と2020年に過度な期待のピーク期としてハイプ・サイクルに登場しています。統計的手法による従来型の機械学習モデルは、どのような計算やしきい値でAIが結論を出したかがわかります。一方、ディープラーニングは基本的にブラックボックスです。子どもがどのように考えて犬と猫を見分けているか脳の中がわからないように、犬と猫を見分けるトレーニングしたAIがどうして判別したのか普通はわかりません。

しかし、AIは幅広く社会に使われるようになってきましたが、ブラックボックスのままでは重要な役割を任せるわけにはいかないと考える分野もあります。例えば医療でなぜAIがそのような判断をしたかわからないまま治療を行い、もし悪い結果となった場合に責任が取れないわけです。この壁を乗り越えるために、AIがどのようなデータをどのように判断したかを、できる限り見える化する仕組みを用意したものが説明可能なAI(XAI)です。

例えば、グーグルは2019年にGCP(Google Cloud Platform)のサービスとして説明可能なAI(Explainable AI)を提供開始しました(β版)。例えば犬と猫を見分ける画像認識AIであれば、説明可能AIは画像のどの特徴量をどれくらい判断に使ったか(スコア)をヒートマップや数値で示してくれます。異常検知であれば、AIがどの部分を見て異常と判断したかをヒートマップで示すことで、人間が「本当に異常かどうか」を最終確認するのに役立ちます。

おわりに

今回はハイプ・サイクルに登場する技術をきっかけに、ディープラーニングの基本について説明しました。基本的なところは3年前と大きくは変わっていませんが、AIが研究室レベルから身近で使うものとなるにつれて注目度が高まっているファインチューニングや蒸留、説明可能AIなどの技術は覚えておいてください。

[第5回]

ハイプ・サイクルに登場する技術④ ― ディープラーニングの基礎技術

はじめに

前回は、ハイプ・サイクルに登場するキーワードの中から汎用人工知能や深層強化学習、転移学習、説明可能なAIなどを取り上げました。人間の脳のニューロンとシナプス、それをモデルとしたニューラルネットワーク、頭が良くなっていく仕組みとして誤差逆伝播などディープラーニングの基礎も解説したので、”いまさら聞けない”という方も楽しめたと思います。今回はハイプ・サイクルに登場する新しめのキーワードの中からアダプティブな機械学習、自己教師あり学習などを解説します。

アダプティブな機械学習 (Adaptive Machien Learning)

ハイプ・サイクルの2019年と2020年の黎明期にアダプティブ(適応型)な機械学習が登場しています。これは、これまでの機械学習とどこが違うのでしょうか。

第3回で解説したように、機械学習は学習プロセス(トレーニング)と推論プロセス(判定や予測、意思決定)の2つのパイプラインで実用化されます。学習プロセスでは大量データを使ってバッチでトレーニングを行い、実用に耐えるレベルに到達させます。そして合格点を得た学習済のモデルを判定や予測、意思決定などの目的に利用するのが推論プロセスです。

このプロセスを深層ニューラルネットワークを使うことで高い実行能力が発揮されて急激に脚光を浴びたのですが、実運用に当たっては、環境(データ)が変化した場合に学習プロセスからやり直しになる不便さが浮き彫りになってきました。当社が取り組んでいる異常検知で説明しましょう。最初に、ある製品の正常・異常をトレーニングして実運用に適用するのですが、設備の経年変化で当初想定してなかった異常が発生する場合があります。また、製品のマイナーチェンジのたびに現場で再トレーニングを行うのは大変です。

そこで求められたのがアダプティブな機械学習です。教師あり学習や教師なし学習ではなく、自ら自律的に学べる強化学習を使い、バッチではなくインストリーム分析(ISA)でオンライン学習します。少量のデータで微調整しながら適応し、AIの学習コストや運用コストを削減して永続的に利用できる理想のAIがアダプティブML(AML)なのです。

教師あり学習と教師なし学習

教師あり学習(Supervised Learning)と教師なし学習(Unsupervised Learning)についてもおさらいしておきましょう。図1は花の画像を分類する手法として、教師あり学習の分類(Classification)と

教師なし学習のクラスタリング(Clustering)を比較した例です。

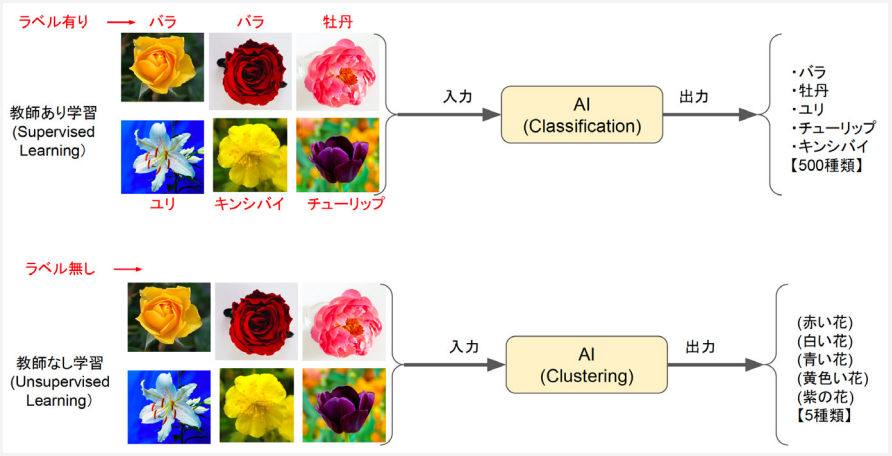

図1：教師あり学習と教師なし学習

教師あり学習は、画像にラベルを付けてから学習する手法で、ここでは花の名前500種類を出力としてトレーニングしています。AIは学習によって500種類の花を見分けられるようになりますが、人が「バラだと思うけど牡丹にも似ているなぁ」と思案するように「バラ85%、牡丹14%、ラナンキュラス1%」というようにスコアを付けて推論します。

うっかりするのですが、出力にない花が解答に選ばれることがありません。例えばシャクナゲを正解に含めずにトレーニングした場合は、シャクナゲの画像を似たような花(バラや牡丹)とスコア低めで判断することになります。何を出力(目的変数)としてトレーニングするかは重要です。例えば花の種類ではなく、花の色を分類するAIを作る場合は花の色を出力とし、赤や黄色といった色をラベルとしてトレーニングします。

一方、教師なし学習はラベルを付ける必要がありません。出力に目的変数(バラや赤いなど分類したいターゲット)を指定することもなく、ただ分類したい数だけを指定します。例えば「花を5つの色に分ける」という命題で処理すると、読み込んだ画像を5種類の色にグルーピングします。人間が分類結果を見ると赤い花の集団、白い花の集団などと見えますが、目的変数はないのでグループに色の名前が付いているわけではありません。意図した色に分類してくれるとは限りませんが、予想外の結果が得られて気付かされることもあります(表)。

学習方法	アルゴリズム	目的変数	特徴
教師あり	分類(Classification)	有り	・精度が高い ・目的に合った分類ができる
教師なし	クラスタリング(Clustering)	無し (分類数のみ指定)	・学習データが不要 ・ラベル付けが不要 ・学習の手間が不要 ・予想外の結果が得られる

表：師あり学習と教師なし学習(分類の例)

この2つをどのように使い分けるのかを、今度はeコマースの例で説明しましょう。例えばアパレルサイトでは、トップス/ジャケット/パンツ/スカートというようなカテゴリーで商品が検索できます。これに、こだわり検索条件として、無地/ボーダー/花柄/チェックなどの柄や、ホワイト/レッド/ブルーなどの色も指定できればさらに便利になります。しかし、膨大な商品全てにこうした細かい分類をるのは

大変です。そこで写真を見て柄や色を自動タグ付けしてくれるAIが求められるのです。このようなタグ付けAIは、検索条件にしたい柄や色を目的変数として学習するので「教師あり学習」になります。

一方、eコマースの顧客を嗜好や購買パターンが似ているグループごとに分類することを顧客セグメンテーションと言います。購買情報をもとに好きなブランド、柄、色、価格帯などの属性で顧客を分類できれば、顧客が興味ありそうなキャンペーンや新商品などの案内を送るターゲティングメールの効果が増します。このような顧客セグメンテーションAIは、ショップが一人ひとりにラベルを付けてトレーニングするわけではないので「教師なし学習」になります。

半教師あり学習(Semi-supervised Learning)

教師あり学習では、大量データにラベルを付け、バッチでトレーニングを行います。これをパッシブ機械学習(Passive Machine Learning)とも呼びます。このやり方の課題は、大量データにラベルを付けるアノテーション作業が大変な点です。また、作業負担の軽減だけでなく、物理的に既知のデータが少なく大部分のデータが未知であるという状況もよくあります。そこで登場したのが半教師あり学習(Semi-Supervised Learning)です。これは一部にだけラベルが付いている状態から、残りのデータにシステムが自動的にラベルを付けて学習するアプローチです。

ラベル無しデータにラベルを付ける処理の基本原理を説明します。例えば犬と猫の画像が10000枚あって、500枚にのみラベルが付いているとしましょう。バッチ処理の1回目で9500枚の未知データの中から既知(500枚)の画像に最も似ている(ベクトル空間の近い)ものを100枚選んで自動でラベルを付けます(合計600枚のラベルが付く)。2回目も同様にラベル付き600枚に似ている画像に100枚ラベルを付けます。この処理を25回繰り返すと3000枚の画像にラベルが付くことになりますね。半教師あり学習は、いかに未知のデータを正しくラベル付けするかがポイントで、自己学習(Self-Training)や共訓練(Co-Training)、PNU Learningなどいろいろな手法があります。

アクティブラーニング(Active Learning)

半教師あり学習は、最初に少量のラベル付きデータを用意することで、残りのラベル無しデータを一定のアルゴリズムで自動的にラベル付けをする方法です。いわばアノテーション(ラベル付け)の自動化を中心にしたもので、最初のラベル付け以外は人間が介在しません。

アクティブラーニングもラベル付きデータとラベル無しデータがある点は一緒ですが、こちらはラベル付け作業はあくまでも人間が行います。自動ラベル付けではなく自動データ抽出、すなわちラベル付け効果の高いデータだけを抽出(Query)するアルゴリズムがポイントです。全てのデータを人間がアノテーションするのがPassive Learningなので、それに対比してActive Learningという名前なのです。

同じ抽出するにしても、ランダムサンプリングだと学習に不向きなデータも混じります。アクティブラーニングは「これとこれが学習に向いてるよ」というデータを選んで渡してくれるわけです。図2のように、選んでもらったデータをアノテーションし、そのデータで学習したAIモデルを本番運用に使用します。そして、本番運用で発生する大量のラベル無しデータをプールし、それを再びQueryしてアノテーションするというサイクルを繰り返し、AIモデルの精度を運用しながら高めて行きます。

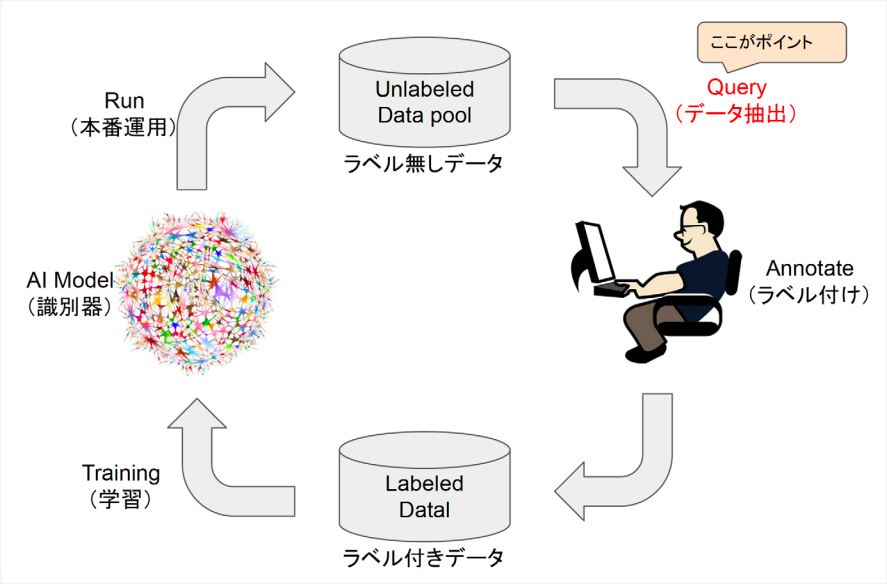

図2：アクティブラーニング

自己教師あり学習(Self-supervised Learning)

自己教師あり学習(SSL)は、早い段階から自然言語処理(Natural Language Processing)の分野で活躍しています。人間の会話では単語が抜け落ちたり、文法が正しくなくても、だいたい何が言いたいか通じます。これは人間の脳が、抜け落ちている単語を予測して補填しているからです。同じように自己教師あり学習を使って「こういう場合はこういう単語が出現する確率が高い」といった予測学習をするわけです。

例えばGoogleは2018年にBERTという言語処理モデルを発表し、文脈を読む能力を格段にアップさせました。学習データの単語の一部を別の単語に置き換えて元の単語を推測させたり、2つの関連ある文章のうち1つを別の文章に置き換えて文章の関連性をスコアするなどして、文脈を理解する力や曖昧な文章を読み解く力を学習したのです。BERTは2019年にGoogleの検索エンジンに導入され、2021年にはAIスピーカーのGoogleアシスタント(英語)にも取り入れられています。

画像処理でも、自己教師あり学習で不足部分を推定するトレーニングを行うことはできます。シンプルな例としては、画像を3D回転、色付け、深度補完などをして特徴点を学習したり、画像データの一部をマスクした上でそこに何があるか予測したり、画像を分割して個々の分割画像が元画像のどこに位置するか当てるなどです。

画像処理(Computer Vision)分野への適用は、変数が離散的な自然言語処理に比べて欠落した部分が連続分布の変数であるため、難易度が高いとされていました。しかし、現代ではさまざまな手法が次から次へと現れ、とてもホットな技術分野になっています。

＜＜Note＞＞半教師あり学習 Vs. 自己教師あり学習

半教師あり学習も自己教師あり学習も、ラベル付きデータが十分用意できない環境でも学習できる手法です。ただし、半教師あり学習は少量のラベル付きデータが必要ですが、自己教師あり学習はデータの基礎構造の不足を補う予測学習であり、教師データが必須ではありません。

ゲームでいえば、半教師あり学習が似たものを次々と一味にしてゆくロールプレイングゲームなのに対し、自己教師あり学習は今年９月に終了する長寿クイズ番組「アタック25」です。若い人はこの番組を知らないかもしれないので説明しておくと、勝ち抜いた人が最後に挑戦する旅行チャレンジVTRクイズで、25枚のうち勝ち取ったパネルにだけ映る映像を見て何が写っているか当てます。このとき、挑戦者の脳は写っている映像だけでなく、伏せられたパネル部分を経験や常識で補完して見ています。自己教師あり学習は、このようにAIに常識や経験を与えて不足している情報を推測できる能力を授ける予測学習なのです。

おわりに

AIはさまざまな分野で実用化されつつあり、それによって要求されるレベルも高度になってきています。最初は教師あり学習で高い推論精度を出すのに汲々としていたのに、説明可能なAI(XAI)やアダプティブな機械学習(AML)、自己教師あり学習(SSL)など、運用を楽にするために必要な技術が次々と登場しています。これらの新しい技術を取り入れていかないと競争に打ち勝てない状況になっていますので、留まらずに積極的にキャッチアップして行きましょう。

Think IT White Paper

29

[第6回]

ハイプ・サイクルに登場する技術④ ― 新たに「ハイプ・サイクル2021」で発表されたAI技術

はじめに

今回はハイプ・サイクル2021の技術を解説します。ハイプ・サイクルには、この先に花開くかもしれないトレンド技術が次々登場します。前年に掲載された技術の多くは消えていますが、それらを丁寧に追跡するものではありません。そのあたりを踏まえて技術動向を読み取っていきましょう。

ガートナーのハイプ・サイクル2021

2021年8月24日に、ガートナーより先進テクノロジのハイプ・サイクル2021が発表されました(図1)。この中からAIに関連するものは次の6つです。

- 生成的AI(Generative AI)
- AI拡張型設計(AI-Augmented Design)
- AI拡張型ソフトウェアエンジニアリング(AI-Augmented SE)
- 量子機械学習(Quantum ML)
- AI主導のイノベーション(AI-Driven Innovation)
- 物理学に基づくAI(Physics-Informed AI)

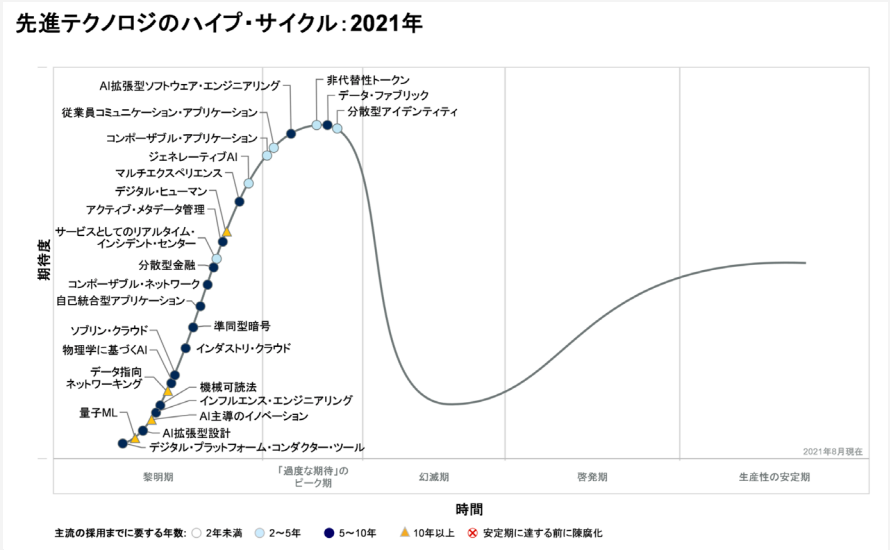

図1:先進テクノロジのハイプ・サイクル2021(出典:ガートナージャパン)

これまでの技術との繋がりで見てみましょう。図2は過去5年のハイプ・サイクルからAI関連技術をピックアップしたものです。昨年に引き続いて掲載されているのは、生成的AIとAI拡張型設計、そして名称が変わったAI拡張型ソフトウェアエンジニアリングの3つ。新しく黎明期に登場したものが量子機械学習、AI主導のイノベーション、物理学に基づくAIの3つです。今回はこれらの技術を中心に解説しましょう。

技術要素	黎明期					過度な期待のピーク期				
	2017	2018	2019	2020	2021	2017	2018	2019	2020	2021
汎用人工知能(Artificial General intelligence)	○	○								
深層強化学習(Deep Reinforcement Learning)	○									
エッジコンピューティング(Edge Computing)	○									
仮想アシスタント(Virtual Assistants)						○	○			
ディープラーニング(Deep Learning)						○	○			
機械学習(Machine Learning)						○				
コグニティブコンピューティング(Cognitive Computing)						○				
転移学習(Transfer Learning)			○							
会話型AIプラットフォーム(Conversation AI Platform)		○								
説明可能なAI(Explainable AI)								○	○	
エッジAI(Edge AI)		○						○		
AI PaaS		○						○		
感情AI(Emotional AI)			○							
敵対的生成ネットワーク(GAN)			○	○						
生成的AI(Generative AI)				○	○					
AI拡張型設計(AI-Assisted Design)				○	○					
AI拡張型開発(AI Augmented Development)				○						
AI拡張型ソフトウェアエンジニアリング(AI-Augmented SE)										○
アダプティブな機械学習(Adaptive ML)			○	○						
自己教師あり学習(Self-Supervised Learning)				○						
コンポジットAI(Composite AI)				○						
責任あるAI(Responsible AI)				○						
量子機械学習(Quantum ML)					○					
AI主導のイノベーション(AI-Driven Innovation)					○					
物理学に基づくAI(Physics-Informed AI)					○					

図2：5年間のハイプ・サイクル(AI関連)

AI拡張型設計(AI-Augmented Design)

Augmentedは「増やす」「拡大する」という英語です。まずは、この聞き慣れない言葉の意味を説明しましょう。「AIは何の略ですか？」はい、これは誰でもわかる質問で、答えは「Artificial Intelligence」、日本語で人工知能ですね。ただ、AIの実態を知るにつけ、ちょっと違和感を覚える言葉でもあるのです。

Artificial Intelligenceは“人工で作った知能”ですから、人間と切り離されて単独で機能する知能というニュアンスが感じられます。しかし、実際は人間に寄り添い、人間の能力を拡大するお手伝いをする役割なので「同じAIでもAugmented Intelligenceと呼ぼう」。そんな考えで拡張知能という言葉が提唱されたのです(図3)。

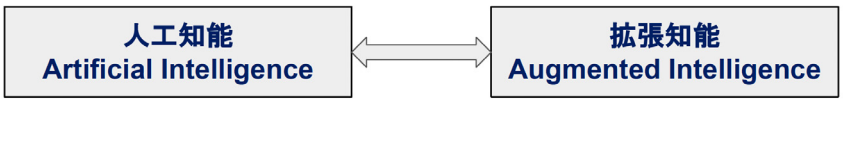

図3：人工知能と拡張知能

その背景を知ると、AI拡張型設計(AI-Augmented Design)は「設計者の能力をAIにより拡張する」「設計作業をお手伝いするAI」という意味だとわかります。

ここで言う設計とは、ソフトウェアの設計だけではありません。自動車、スマホ、建築、エンジニアリングなど、さまざまな分野が対象となります。図4は一般的な設計作業のモデルです。エンジニアは、豊富な経験やノウハウ、知識を駆使して、法規や規格などに則りながら、CAD、CAE、CAMなどのツールを使いこなして設計作業を進めて行きます。円滑に進めるために膨大な資料、図面、Knowledgeなどがデータベース化されており、これらの情報を活用して効率よく設計作業を行います。

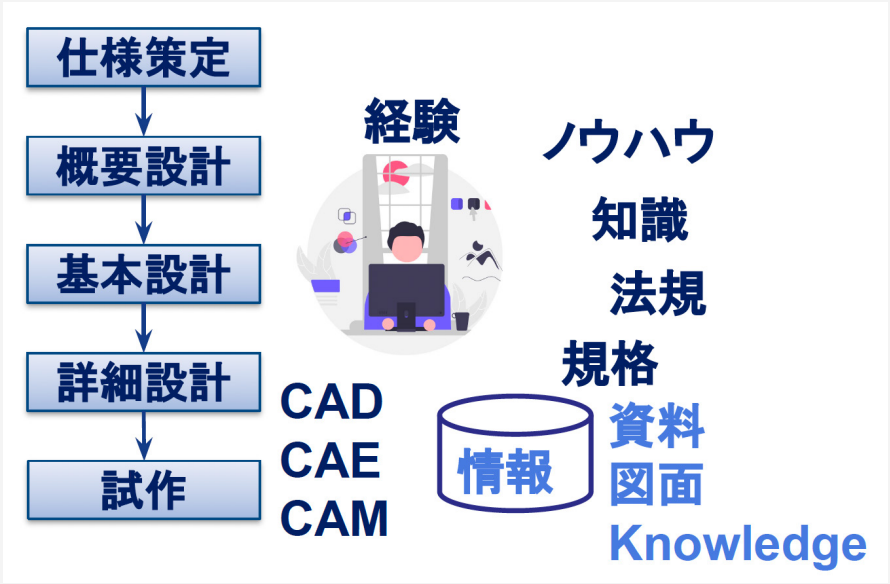

図4：エンジニアの設計作業

この設計作業をお手伝いするAIがAI拡張型設計です。お手伝いの方法はいろいろあります。膨大な知識の検索を支援するナレッジ&リサーチAIや設計レビューのコラボレーションを行うAI、論理および物理シミュレーションを支援するAI、熟練エンジニアの経験やノウハウを習得して一般エンジニアを支援するAIなど、さまざまな分野で研究・開発が行われています。

AI拡張型ソフトウェアエンジニアリング (AI-Augmented SE)

AI-Augmented Software Engneeringは、2020年のハイプ・サイクルのAI拡張型開発(AI-Augmented Software Development)が名称変更したものです。こちらは対象を“ソフトウェア”に限定した開発、つまりプログラミングやテストを支援するAIということになります。代表的なものが自動プログラミング、すなわちAIを使ったコード生成です。ここではツールを3つ紹介しましょう。

OpenAI Codex

この分野で注目する技術に、人工知能を研究する非営利団体OpenAIが開発したGPT-3という文章生成言語モデルがあります。GPT-3は、簡単な指示を与えると、それをもとに自然な文章を作成してくれます。これがもっと進化して完成度が高くなれば、骨子やあらすじを示すだけでブログや記事、小説などをAIが書いてくれる時代が来る、そんなふうに期待されているのです。

これを応用して文章の代わりにソースコードを生成するようにしたものがOpenAI Codexです。2021年8月10日にβ版のAPIが提供開始され、OpenAIのホームページで簡単なデモ動画も見ることがで

きます。宇宙船のデモはまだまだ稚拙ですが、何事も最初の一歩はこんなもの、どこまで進化できるか期待したいです。

GitHub copilot

GitHubがCodexを搭載して開発し、2021年6月に発表されたのがCopilot(副操縦士の意味)です。このサービスはGitHub上の数百万のオープンソースでトレーニングされていて、人間が書いたコメントやコードをもとに新しいコードを自動生成して提案するAIペアプロです。現在、Microsoftのコードエディタ VSCode(Visual Studio Code)のテクニカルプレビュー(β版)として公開されています。こちらもGitHub copilotのホームページで簡単なデモ動画を見ることができます。CodexはAIを使ったテスト支援もターゲットにしており、テストコード自動生成の機能も開発されています。

Microsoft Power Apps

実は、2020年9月にMicrosoftはGPT-3の独占ライセンスを取得しました。そして2021年5月にMicrosoftはGPT-3をPower Appsに統合し、ローコードプログラミング言語Power Fxを生成するサービスをプレビュー提供しています。

<<Note>>Power AppsとPower Fx

Microsoft Power Appsは、Microsoft 365(旧Office365)やDynamics365で利用できるアプリケーション作成ツール基盤です。プログラミングをしなくても簡単にビジネスアプリケーションを作成できるという最近のノーコード/ローコード開発の流れを汲んで2016年にリリースされています。

そしてPower Fxは、Microsoftが2021年3月に発表したローコード開発向けのプログラミング言語です。Excelベースで作られた言語なのでExcelと同じような操作で使えますし、Excelでコードを書いていた経験も活かせる上に、SQLでデータを操作したり行コメントを書いたりもできます。ここにGPT-3自然言語モデルを組み込んで、話し言葉でPower Fxのコードが生成されるようになるわけです。

なお、このPower AppsとPower BI(BIサービス)、Power Automate(RPAとワークフロー)の3つのサービスを合わせたものがMicrosoft Power Platformです。

量子機械学習(Quantum Machine Learning)

2019年にGoogleが“量子超越性の実証実験”で「スパコンでは1万年かかる計算をゲート式の量子コンピュータで200秒で実行できることを実証した」と発表して話題になりました。

この量子コンピュータの超高速計算能力を機械学習に取り入れたら、どんなことができるだろうと期待が高まりますね。例えばAIが期待されている分野の1つに創薬があります。創薬AIは基本的に“手当り次第に試して効き目のある組み合わせを発見する”ような処理なのですが、膨大な組み合わせを量子コンピュータで処理したら新型コロナの薬もできるかもしれません。

量子コンピュータは量子ビット(Qubit)を使っています。古典的なコンピュータのビットと違い、0と1だけでなくその重ね合わせやもつれ(entanglement)の状態も取れる特徴があります。また、量子コンピュータで指数関数的な高速計算を得るには、その特徴を活かした量子アルゴリズムが必要になります。しかしながら、現在発見されている実用的な量子アルゴリズムはまだ少なく、量子データもエラー耐性が無く重ね合わせ状態が崩れるなどの課題を抱えています。

そのため、現在は図5の組み合わせのうち③のハイブリッド型の研究が多く、量子畳み込みニューラルネットワーク(QCNN)、量子敵対的生成ネットワーク(QGAN)、量子サポートベクターマシーン(QSVM)、量子強化学習(QRL)などのモデルが次々と発表されています。

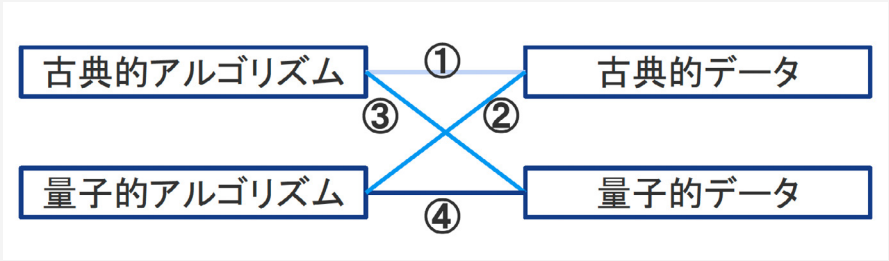

図5：量子コンピュータの組み合わせ

物理学に基づくAI(Physics-Informed AI)

ディープラーニングは基本的にブラックボックスです。子どもが犬を見て“わんわん”と指差したときに、親はこの子がどうして犬と判断したかわからないのと一緒です。そして子どもがとんでもないもの(例えばカラス)を見て“わんわん”と言ったとき、親は首を傾げて驚くしかありません。

ディープラーニングは統計をベースに判断する手法なので、学習されていないデータに対して非現実な結果を生成する可能性があります。また、膨大なデータで学習する必要があり、なぜそう推論したか説明できないという課題があります。

この課題を解消するのが「物理学に基づくAI」です。全くイチからトレーニングするのではなく、物理学でわかっている範囲を当てはめ、追加データのみで学習します。この方法なら物理学で制約するため非現実な間違いを回避でき、学習データも少なく済むのです。先程の例で言えば「耳が2つ立ったり寝たりしていて」「口が尖っていて」「足が前と後ろに2本ずつある」とキャップを当てはめれば、カラスをワンワンと呼ぶことはないのです。

IBMが煙が拡散するモデルの実験を行ったレポートでも、物理的な制約なしでニューラルネットワークをトレーニングした結果より、移流拡散法則という物理法則を課してトレーニングした方が精度が高くなったと報告されています。

生成的AI(Generative AI)

この数年、生成モデルが脚光を浴びていて、2019年と2020年に敵対的生成ネットワーク(GAN)、2020年と2021年に生成的AI(Generative AI)が黎明期にランクインしています。GANは画像や映

像などを作る才能が豊かなAIの手法で、これを使ったDeepfakeが話題になっていますが、文章作成や作曲などさまざまな生成技術が進化する中で汎用的な生成的AIというキーワードに切り替わりました。

「生成」は「認識」の逆動作です。例えば図6の朝顔の写真を見て、ここから色が青、大きさが10cm、花びらの数が6枚などの特徴量を取り出し、n次元の潜在変数で表すのが認識モデルです。これを使えば、例えば1000枚の花の写真の中から「青い花」や「花びらが6枚」という条件で検索できるようになります。

一方、このような潜在変数をもとに朝顔の画像を作り出すのが生成モデルです。n次元の潜在変数とは、Zipファイルのようなものです。画像を圧縮してZipファイルにした場合、Zipファイルさえあれば画像を復元できますね。それと同じです。

次元を少なくして考えてみましょう。例えば朝顔の代わりに赤い丸だとします。この場合の潜在変数は「色が赤(#F15B5B)」「形状が丸」「半径5cm」「線の太さは2ピクセル」という4次元でしょうか。そして、この4次元の変数さえ与えれば誰でも全く同じ赤丸を書くことができるわけです。

図6：生成モデル(Generative AI)

敵対的生成ネットワーク(GAN)

GANについても説明しておきましょう。GANは生成モデルのジェネレーター(生成器)と識別モデルのディスクリミネーター(識別器)が敵対しながら学習するモデルです。敵対という言葉になっていますが、私には生成器(ジェネ君)と識別器(ディス君)という兄弟が切磋琢磨しながら共に成長してゆくドラマに見えます。

図7がGANの基本構造です。ここで実物サンプルが先程の朝顔の画像だとしましょう。ジェネ君はせっせと朝顔に似せた画像を生成し、本物とランダムに切り替えてディス君に見せ、ディス君はそれが

本物かどうかを判定します。偽物とバレた場合は誤差逆伝播によりジェネ君が調教され、間違えた場合はディス君が調教されます。最初のうちはふたりとも下手っぴいなのですが、どちらも上達するにつれジェネ君はかなり本物に近い画像を生成できるようになります。こうして贋作づくりの名人となったジェネ君を取り出して生成AIとして利用するのです。主役はジェネ君でディス君はそのトレーニングパートナーなのですが、面白いことにジェネ君は一度も本物の画像を見たことがなく、ただディス君を騙そうと必死にやっているうちに本物そっくりを作れるようになったのです。

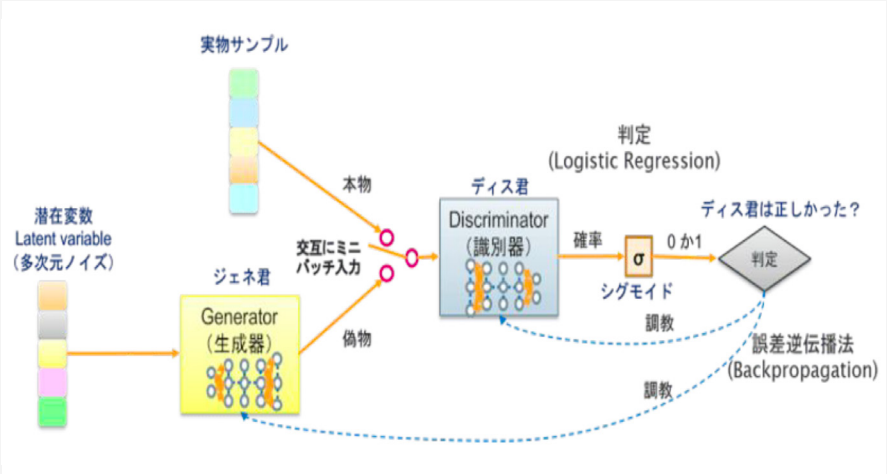

図7：敵対的生成ネットワーク(GAN)

コンポジットAI(Composite AI)

最後に2020年の黎明期に登場したComposite AIを説明します。Compositeとは複合の意味で、最良の結果を達成するために複数のAI技術を組み合わせることです。一般にAI技術は、自然言語処理(NLP)、音声認識・合成、コンテキスト分析、ナレッジグラフ、感情分析、機械翻訳、画像認識、画像生成など数えきれないほどあって、それぞれが実用的になっています。そんな状況になったので、ある目的を果たすために複数のAI技術を組み合わせることが当たり前になりつつあり、それをコンポジットAIと名付けたわけです。

自動運転車で考えてみても、ざっと「車両位置特定技術」「検知・認識技術」「AI運転操作」「予測技術」「走行ルート計画」「運転手監視」「通信技術」など数多くの技術が必要です。これらがすべてAI技術というわけではありませんが、複数のAI技術が組み合わさるのが普通の時代となっているのです。

おわりに

ここまで、ハイプ・サイクルに登場したAI技術を中心に最近の動向を解説しました。第1部の技術解説編は今回で終了です。次回からは、第2部としてこれらの技術が現在どのように社会で実用化されているのかというビジネス編をお届けします。

特別編集号
2022 winter
◎発行人 中村 照明
◎編集 Think IT編集部
◎AD・表紙デザイン・制作 岡田 章志（岡田デザイン事務所）

シンクイット™

Think IT

thinkit.co.jp

エンジニアのための オープンソース実践活用メディア

“オープンソース技術の実践活用メディア”をスローガンに、インプレスグループが運営するエンジニアのための技術解説サイト。開発の現場で役立つノウハウ記事を毎日公開しています。

2004 年の開設当初からOSS（オープンソースソフトウェア）に着目、近年は特にクラウドを取り巻く技術動向に注力し、ビジネスシーンで OSS を有効活用するための情報発信を続けています。OSS に特化したビジネスセミナーの開催や、Web 連載記事の書籍化など、Web サイトにとどまらない統合的なメディア展開に挑戦しています。また、エンジニアを含むクリエイターの独立・起業、フリーランスなどの多様化する「働き方」や「IT で社会課題を解決する」等をテーマに、世の中のさまざまな取り組みにも注目し、解説記事や取材記事も積極的に公開しています。

ご利用のお客様へ

impress
株式会社インプレス
法人営業局
customer@impressbm.co.jp
〒101-0051　東京都千代田区神田神保町1-105　神保町三井ビルディング