2022 年 3 月

SoC1290

AI Brittleness

By Katerie Whitman（Send us feedback）

AI の脆弱性

ディープラーニングは変革的な人工知能技術であり、幅広い製品、サービス、業界で活用されている。しかし、ディープラーニングシステムは予測不可能な不具合が発生しやすく、新規の状況に適応する能力は限られている。ディープラーニングの脆弱な特性は、自動運転車、自動医療診断システム、およびその他の画期的な AI アプリケーションの商業化に対して、解決が困難な制約とされてきた。一方で、AI の脆弱性の問題が広く認識されたことで新しい AI 手法の研究が促進されている。この研究は、AI の能力に破壊的なブレークスルーをもたらす可能性がある。

ディープラーニングは、多層人工ニューラルネットワークを活用してデータから特徴を抽出する機械学習技術の一端をなす。2012 年、AI 業界でディープラーニングによる全面的な変革が始まり、コンピュータービジョン、音声認識、自然言語処理、創薬、不正検出、レコメンデーションエンジンなどの AI アプリケーションが劇的に改善された。大がかりで急速な改善が行われたため、アナリストは、AI が程なく幅広い業界にわたる数千万人の労働者を追い出しにかかるだろうと予測し始めた。自動運転車開発企業の Waymo などディープラーニングの最前線にいた企業は投資家から非常に高い評価を得た。

ディープラーニングシステムは予測不可能な不具合が発生しやすく、新規の状況に適応する能力は限られている。

しかし、2018 年後半までには、ディープラーニングは評判ほどのものではないことが明らかになった。ディープラーニング技術が、重要なアプリケーションにおいて人間の判断に代わるだけの信頼性も適応性もなかったためである。投資家の期待ははこれに応じて変化し始めた。2020 年 3 月までに、Waymo の市場評価はピーク時から 85%近く低下し、その後主要幹部が同社を去った。他の野心的な AI 企業も同様に挫折した。例えば、IBM は 2022 年に、Watson Health サービスのコアデータ資産を売却した。このサービスは医師や医学研究者に高度な AI に基づくレコメンデーションを提供していた。ディープラーニングは引き続き広く使用されているが、その見通しと一部の既存用途への適合性について懐疑的な見方が広まっている。それにもかかわらず、ディープラーニングの脆弱性問題や他の多くの欠点を解決するための研究は非常に活発であり、業界や政府から多額の資金を集めている。

AI の脆弱性問題を克服するための取り組みは、短期的には成功する可能性が低そうだ。現在研究中のメソッドは開発の非常に初期段階にある。自動運転車などの画期的な AI アプリケーションの開発者は、引き続きブルートフォース方式に関する脆弱性の問題に対処し、コスト効率の高い方法でアプリケーションを拡張することに問題を抱えていくだろう。しかし、将来は不確実であり、状況の変化は別の結果を引き起こす可能性がある。AI の脆弱性の将来を変化させる可能性のある事象の例を以下に示す。

- **ディープラーニングとシンボリック AI の融合の成功**
  根本的に、ディープラーニングはデータセット内のパターンと相関関係を特定する手段に過ぎない。ディープラーニングシステムは、データパターンの意味を根本的に理解できず、データ内の因果関係や概念上の関係を本質的に特定することはできない。研究者は、ディープラーニングとシンボリック AI（概念とそれらの概念間の論理的な関係を記号化する AI メソッド）との融合を模索してきた。ニューロシンボリック AI 手法は、シンボリック AI とディープラーニングのそれぞれ

の強みを組み合わせる一方でそれぞれの弱点を排除することが可能で理想的である。しかし、ニューロシンボリック手法はまだ開発の初期段階にある。さらに、ニューロシンボリック AI は必ずしもディープラーニングよりも堅牢であるとは限らない。

◆ **有用な常識の定義の出現**

ディープラーニングの脆弱性の問題に対処する多くの取り組みは、ディープラーニングシステムに常識（世界がどのように機能しているかに対する日常的な理解）を与える方法を見出すことに重点を置いている。しかし、AI システムの文脈において常識を構成するものの適切な定義は現在のところ存在しない。開発者はこの問題に取り組んでいるが、AI にとってこれは目新しい問題ではない。実際、1950 年代から AI 研究の一部として同様の取り組みが行われてきた。ということは、ディープラーニングに役立つ解決策が得られるのはかなり先になるかもしれない。

◆ **AI がデータから複雑な因果関係や概念上の関係を学習できるメソッドの達成**

一部の AI 研究者は、ディープラーニングシステムがデータから直接因果関係を学習できるようにする手法を模索している。すでにさまざまな機械学習システムでデータ内の因果関係を発見することが可能になっているが、極めて初期の研究段階にある一層高度な手法は、最終的に、システムが動作する世界の概念的理解と言えるようなものを備えたディープラーニングシステムをもたらす可能性がある。理論的には、そのようなシステムは今日のシステムよりもはるかに脆弱性が少ないかもしれない。

◆ **新しいハードウェアを AI に適用する手法**

今日のディープラーニングシステムは、AI 処理タスクを高速化する機能を備えた従来のコンピュータハードウェア上で動作する。その他に従来のコンピュータハードウェアとは根本的に異なる種類の AI ハードウェアが研究されている。例えば生物学的神経組織の構造的および機能的側面を模倣する神経形態学的システムである。理論的には、このようなハードウェアは生物学的脳のように機能する画期的な AI 技術をもたらし、欠点なしにディープラーニングの主要な利点のいくつかを実現できる可能性がある。しかし、これまでのところ神経形態学的ハードウェアには現実的な明るい見込みがほとんどない。さらに、脳がどのように機能するかに関する研究者の理解は依然として非常に限られており、それが脳機能を模倣することを目的とした新しい種類のコンピューターシステムを設計する取り組みの展望を限られたものにしている。

◆ **小さなブレークスルーの蓄積**

すべての AI 専門家が、ディープラーニング技術の性能に大きなブレークスルーがなければ、ディープラーニングが重要なアプリケーションで人間の判断に取って代わることは永久にできないと認めているわけではない。ディープラーニングシステムが脆弱なままであっても、十分なデータ、演算能力、アルゴリズムの改良があれば、不具合が発生する率を許容できるレベルまで下げられるかもしれない。自動運転車のフリートを商業化するための現在の取り組みは、こうした漸進的な手法を、システムの動作領域を慎重な制限することや人間による監視の組み込みを併用しながら、ディープラーニングの限界を克服しようとしている。たとえ不完全ではあっても、十分な時間があれば力づくの対処法によって、ディープラーニングの不完全性を補い、実用的となる程度にまで演算コストを低下させられるかもしれない。

◆ **害をもたらす敵対的攻撃**

ディープラーニングシステムは、理論的には、特別なデータパターンを使用してシステムの誤動作を引き起こす悪意ある攻撃に対して脆弱である。このような敵対的な攻撃はまだ現実に重大な問題を引き起こしていないが、セーフティクリティカルなシステムが悪用され害がもたらされた場合、ディープラーニング技術に対する信頼がさらに失われる可能性がある。

SoC1290

本トピックスに関連する Signals of Change		関連する Patterns	
SoC1249	AI革新にあわせた規模拡張	P1716	メタ学習
SoC1119	AIをトレーニングするAI	P1667	AIシステムの制限の緩和
SoC985	人工知能は異質なインテリジェンス	P1551	AIとその幻滅期