九州大学
KYUSHU UNIVERSITY

PRESS RELEASE（2022/07/29）

AI を壊してより良いものにする

―AI が未知の要素をどのように扱うか評価する手法を開発―

ポイント

① 現在の AI は、画像認識の精度は高いが融通が利かず、その理由は現在も分かっていない
② 今回、ニューラルネットワーク(※1)が未知の要素をどのように扱うかを評価する「Raw Zero-Shot」と呼ばれる手法を開発
③ 今後、この手法が将来の AI のロバスト性(※2)向上に役立つことに期待

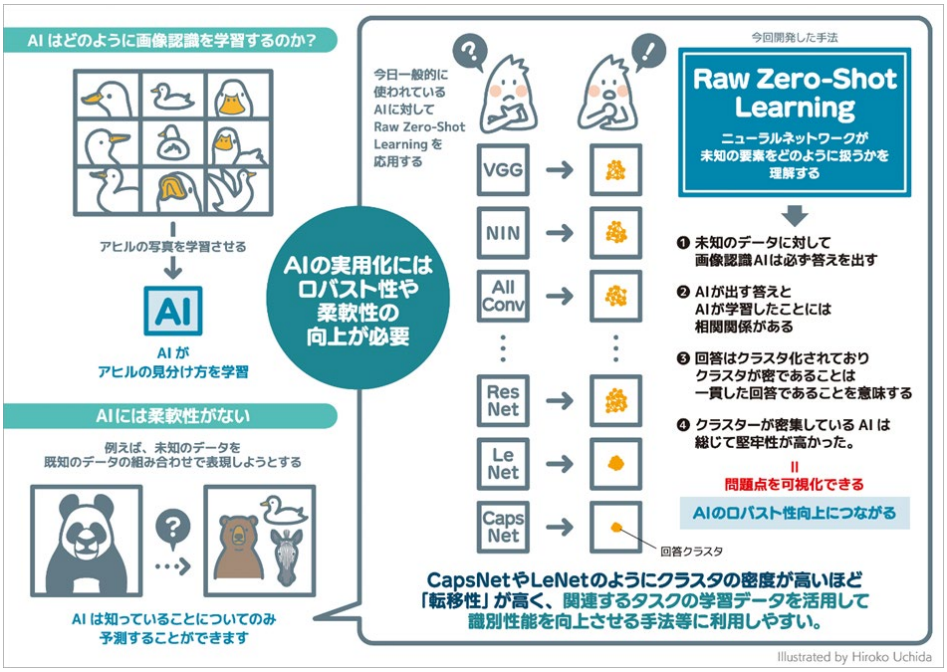
概要

現在の AI は、画像認識の精度は高いですが融通が利きません。しかし、なぜこのようなことが起こるのか、正確には謎のままです。

今回、九州大学大学院システム情報科学研究院の VASCONCELLOS VARGAS DANILO（ヴァスコンセロス　ヴァルガス ダニロ）らの研究グループは、ニューラルネットワークが未知の要素をどのように扱うかを評価する「Raw Zero-Shot」と呼ばれる手法を開発しました。

この結果は、研究者がニューラルネットワークを堅牢でなくしている共通の特徴を特定し、AI をより信頼性の高いものにする手法を開発するのに役立つ可能性があります。

本研究成果はアメリカの雑誌「PLOS ONE (2022)」に 2022 年 4 月 27 日（水）に掲載されました。



画像認識 AI は強力だが柔軟性に欠け、特定のデータで学習させないと画像を認識することができない。ロー・ゼロショット学習では、研究者が画像認識 AI に様々なデータを与え、その答えのパターンを観察する。AI が一貫した答えを出せば出すほど、少し変わった画像にも強くなる。研究チームは、この方法論が将来の AI のロバスト性向上に役立つことを期待している。

【研究の背景と経緯】

あるテーブルゲームで「ガラスの大砲」と呼ばれる、攻撃力は高いが防御力に欠けるキャラクターが登場します。つまり自身の攻撃は強いですが、攻撃を受けると崩れてしまいます。画像認識に用いられるディープラーニング(※3)のアルゴリズムである現在の人工ニューラルネットワークも、同様の欠点を持っています。人間離れした精度を誇る一方で、画像を少し修正すると簡単に壊れてしまうのです。

この「堅牢性」の欠如は、より優れた人工知能の構築を目指す研究者にとって、大きなハードルとなっています。しかし、なぜこのような現象が起こるのか、そのメカニズムはまだほとんど分かっていません。

【研究の内容と成果】

九州大学大学院システム情報科学研究院の研究者らは、こうした欠点を克服するため、ニューラルネットワークが未知の要素をどのように扱うか評価する手法を開発しました。この結果は、さまざまなニューラルネットワークに共通する、堅牢でない特徴を特定し、根本的な問題を解決する方法を開発するのに役立つことが期待されます。

画像認識のためのニューラルネットワークは、非常に強力で、ヘルスケアから自動運転車まで、数多くのアプリケーションで使用される可能性があります。しかし、AI がどんなによく訓練されていても、画像のわずかな変化で失敗することがあります。

実際には、画像認識 AI は、1 つの画像を識別するよう求められる前に、多くのサンプル画像で「訓練」されます。例えば、AI にアヒルを識別させたい場合、まず多くのアヒルの写真で学習させることになります。

しかし、どんなによく訓練された AI でも誤認識することがあります。実際画像を加工すると、人間の目にはそのままの画像に見えても、AI には正確に識別できないことが分かっています。画像に 1 ピクセルの変化があっただけでも、AI に混乱が生じることがあるのです。

研究チームはこの原因を探るため、さまざまな画像認識 AI を調査し、AI が学習していない、つまり AI にとって未知のサンプルに直面したときの挙動にパターンを見出すことを目指しました。

AI に画像を与えると、その答えが正しいかどうかに関わらず、それが何であるかを教えようとします。そこで、現在最もよく使われている 12 種類の AI を取り上げ、『Raw Zero-Shot Learning』という新しい手法を適用しました。基本的には、AI にヒントやトレーニングを与えず、一連の画像を与えたのです。私たちの仮説は、AI の回答に相関性があるというものでした。間違っても、同じように間違うはずです。

結果はまさにその通りでした。どのような場合でも、画像認識 AI は答えを出し、その答えは間違っていても一貫している、つまりクラスター化(※4)するのです。各クラスターの密度は、異なる要素に基づく知識を持つ AI が、未知の要素をどのように処理するかを示しているのです。この未知の要素を処理するために学習した知識の「伝達性」は、AI が少し加工された画像を処理する方法にもつながっている。

研究チームは CapsNet として知られる Capsule Networks が、ニューラルネットワークの中で最も密なクラスターを生成し、伝達性が高いことを確認しました。これは CapsNet が動的な性質を持つためではないかと考えています。

【今後の展開】

現在の AI は精度が高い反面、ロバスト性に欠けるため、さらなる実用化が難しい。何が問題なのか、なぜそうなっているのかを理解する必要があります。この研究では、これらの問題を解決するための可能な戦略を示しました。精度だけにこだわるのではなく、ロバスト性と柔軟性を向上させる方法を検討する必要があります。そうすれば、真の人工知能を開発できるかもしれません。

【用語解説】
(※1) ニューラルネットワーク
人工ニューラルネットワークのこと。人工ニューロンと呼ばれるユニットやノードの集合体に基づいており、生物の脳のニューロンをモデル化したものです。
(※2) ロバスト性
ロバスト性とは機械学習モデルが騙されないようにする能力のことです。例えばモデルが画像の分類を行うときに、入力の変化を行うことで望んだ分類結果とは異なる結果を得られるような敵対的アルゴリズムが存在しています。
(※3) ディープラーニング
ディープラーニングは人工ニューラルネットワークをベースとした機械学習の一種です。データから段階的に複雑な特徴を抽出するため、複数の処理層を使用します。
(※4) クラスター化
一般には細かな要素が大きな集合を形成することを指します。機械学習の文脈ではモデルの出力が複数の集合を形成していることを示しており、データのパターン化を行なっているということになります。

【謝辞】
本研究は JSPS 科研費 （JP20241216,)、JST 研究費（JP=50243）の助成を受けたものです。

【論文情報】
掲載誌：PLOS ONE (2022)
タイトル：Transferability of features for neural networks links to adversarial attacks and defences
著者名：Shashank Kotyan, Moe Matsuki, and Danilo Vasconcellos Vargas,.
D O I ： 10.1371/journal.pone.0266060

【お問合せ先】
＜研究に関すること＞
九州大学大学院システム情報科学研究院情報学部門
VASCONCELLOS VARGAS DANILO（ヴァスコンセロス　ヴァルガス ダニロ）
TEL：092-802-3599
Mail：vargas@inf.kyushu-u.ac.jp

＜報道に関すること＞
九州大学広報室
TEL：092-802-2130　FAX：092-802-2139
Mail：koho@jimu.kyushu-u.ac.jp