PRESS RELEASE

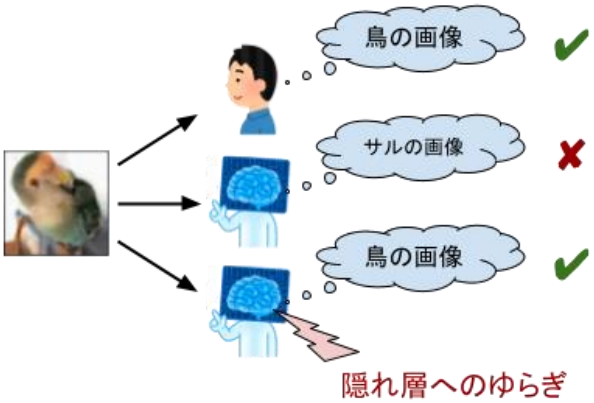
東京大学

脳のゆらぎを取り入れて AI を安全にする

――深層ニューラルネットワークの隠れ層にゆらぎを導入し脆弱性を軽減――

発表のポイント

◆人工知能の主要なモデルである深層ニューラルネットワークには、人間とは明らかに異なった間違いをするような脆弱性があることが知られており、人工知能を社会実装する上で重要な課題の一つである。
◆脳の神経細胞を模したゆらぎを深層ニューラルネットワークに導入することで、特定のタイプの脆弱性を軽減できることを発見した。
◆本研究は、脳の神経細胞のゆらぎの役割に関する新しい仮説を提唱するだけでなく、より人間の振る舞いに近い安全な人工知能を作成する上での示唆を提供する。



画像認識 AI の隠れ層にゆらぎを導入することで正しく画像を認識できるようになった。

発表概要

東京大学大学院医学系研究科 機能生物学専攻 統合生理学分野の大木研一教授と浮田純平大学院生（研究当時）の研究チームは、深層ニューラルネットワーク（注 1）に脳の神経細胞を模したゆらぎ（注 2）を導入することで、深層ニューラルネットワークが持つ脆弱性の一部が軽減できることを明らかにしました。

現在、人工知能（AI）の進化が加速度的に進んでいますが、その基礎となる構造は深層ニューラルネットワークに基づいています。しかし深層ニューラルネットワークは、敵対的攻撃と呼ばれる悪意のある攻撃によって、人間とは明らかに異なる出力をするように騙されてしまうことが知られています。例えば自動運転車に搭載された画像認識 AI は、「止まれ」の道路標識を正しく「止まれ」と認識して車が停止する必要があります。しかし敵対的攻撃によって生成された「止まれ」の道路標識は、人間が見ると明らかに「止まれ」の標識であっても、画像認識 AI は正しく認識できません。結果、車が停止できず、交通事故につながる恐れがあります。このように、AI を社会実装する上で、敵対的攻撃に対する脆弱性は大きな課題の一つです。

人間など動物の脳の性質を AI に取り入れることで、このような脆弱性を克服できる可能性があります。本研究チームは、脳の神経細胞が持つゆらぎを参考に深層ニューラルネットワークにゆらぎを導入することで、特定のタイプの脆弱性が軽減できることを明らかにしました。この方

法を用いることで、より人間などの動物の振る舞いに近い AI が作成できる可能性が高くなると考えられます。

本研究は、Beyond AI 研究推進機構、日本医療研究開発機構（AMED）「革新的技術による脳機能ネットワークの全容解明プロジェクト」、文部科学省科学研究費助成事業、CREST-JST などの支援を受けて行われました。本研究の成果は Neural Networks 誌（9 月 15 日オンライン版）に掲載されました。

発表内容

【研究背景】

現在、様々なウェブサービスや医療画像診断など、多くの場面で人工知能（AI）が用いられています。画像認識や自然言語処理における主なモデルは深層ニューラルネットワークですが、深層ニューラルネットワークは非常に多くのパラメータからなり、内部がブラックボックスに近いことから、不可解な振る舞いをすることが知られています。

そのような振る舞いの一つに敵対的サンプル（注 3）の存在があります。画像認識モデルを例に考えたときに、例えば人間が見ると明らかにパンダに見えるが、深層ニューラルネットワークはサルだと認識してしまうような画像を人工的に生成できることが知られています。このような画像は敵対的サンプルと呼ばれ、敵対的サンプルを生成するプロセスは敵対的攻撃と呼ばれています。このような敵対的サンプルが存在することで、深層ニューラルネットワークを社会実装する際に問題になる危険性をはらんでいます。例えば、人間には明らかに「止まれ」に見えるが AI には「直進」に見えるような道路標識を作ったり、顔認証システムの検知を逃れたりすることが可能になってしまいます。したがって、敵対的攻撃に対する脆弱性を軽減することは重要な課題となっています。

人間など動物における情報処理の性質を AI に取り入れることで、このような脆弱性を克服できる可能性があります。本研究チームは、動物の神経細胞において知られている性質の中でも特に、神経細胞が持つゆらぎに着目しました。ゆらぎとは、感覚入力などの外部条件が同じにもかかわらず、神経細胞の活動が同じとは限らない性質のことを指します。動物の視覚情報処理においては、情報処理の入力部に位置する網膜だけでなく、その後段に位置する脳においてもゆらぎが存在します（図 1）。情報処理の入力部におけるゆらぎは、外界の情報が持つばらつきにも対応できる回路を学習する上で重要と考えられます。実際、深層ニューラルネットワークにおいても、入力層にゆらぎを加えることで敵対的攻撃に対する脆弱性を克服する研究は多く行われています。しかし動物の脳のように、後段の隠れ層においてゆらぎを導入することの意義は明確ではありませんでした。

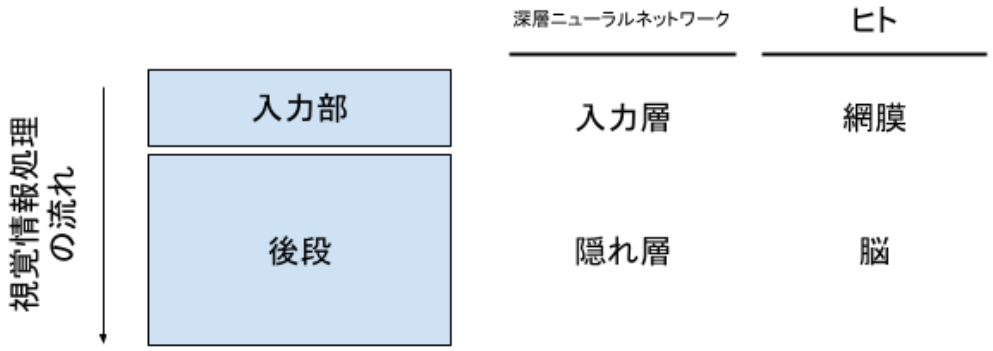


図 1: 深層ニューラルネットワークとヒトにおける視覚情報処理の流れの対応。

【研究手法と結果】

本研究チームはまず、敵対的攻撃を行う新しい手法を考案しました。この手法によって生成された敵対的サンプルの例を図 2 に示します。この手法で生成された敵対的サンプルは、深層ニューラルネットワークが誤認識してしまうだけではなく、以下の 2 つの性質を持ちます。1 つ目は、

深層ニューラルネットワークの入力層においては、敵対的サンプルと通常の画像（図 2 の例では、深層ニューラルネットワークが正しく鳥と判定できるような鳥の画像）の距離が遠くなるという特徴です。2 つ目は、深層ニューラルネットワークの隠れ層においては、敵対的サンプルと通常の画像の距離が近くなるという特徴です。

図 2: 本研究において生成された敵対的サンプルの例。画像認識モデルは、この画像に写っているものが鳥ではなくサルだと誤って認識した。

本研究では、様々な画像データセットを用いて、この手法によって敵対的サンプルを生成しました。その後、深層ニューラルネットワークの入力層、隠れ層の 2 通りの箇所にそれぞれゆらぎを導入し、敵対的サンプルが正しいカテゴリー（図 2 の例では鳥）に分類されるか、もしくは敵対的サンプルと同じく誤って（図 2 の例ではサルに）分類されるかどうかを調べました。結果、図 3 のように、入力層にゆらぎを導入した時よりも、後段の隠れ層にゆらぎを導入した時の方がはるかに、敵対的サンプルが正しいカテゴリーに分類される割合が増えました。すなわち、入力層ではなく隠れ層にゆらぎを導入することで、今回の敵対的攻撃に対する脆弱性を軽減することに成功しました。

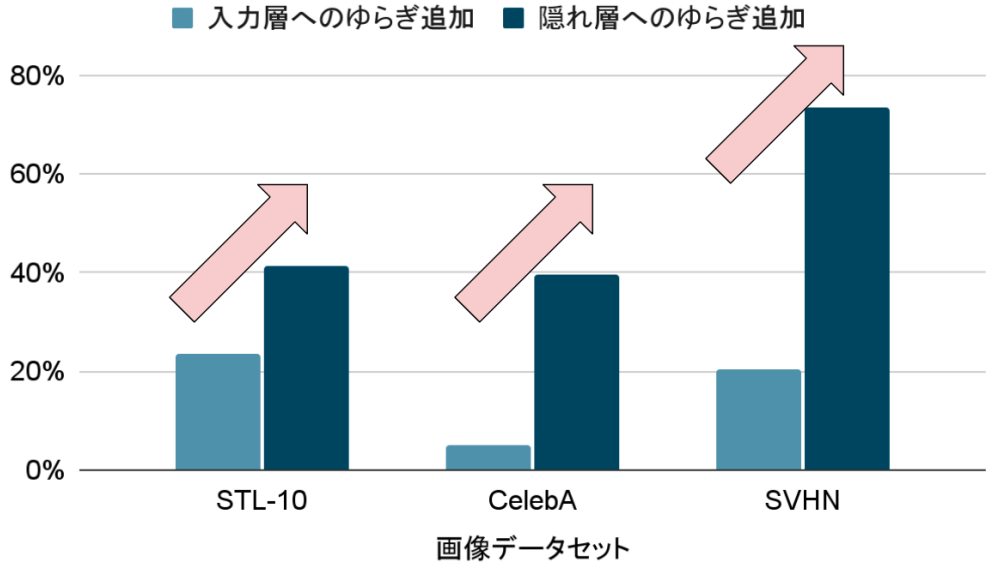


図 3: 生成した敵対的サンプルに対して、深層ニューラルネットワークの入力層、隠れ層それぞれにゆらぎを導入した時に、正しく認識できるようになった割合。値が 100%に近いほど、正しく認識できるようになった、すなわち敵対的攻撃に対する脆弱性が減ったことを意味する。

【研究結果から期待されること】

本研究により、深層ニューラルネットワークの情報処理の後段に位置する隠れ層にゆらぎを導入することで、特定の敵対的攻撃に対する脆弱性が軽減できることが判明しました。この方法を用いることで、より人間などの動物の振る舞いに近い AI が作成できる可能性が高くなると考えられます。

また神経科学分野においても、脳の神経細胞が持つゆらぎの意義に関しては様々な仮説が立てられ、検証されています。脳の神経細胞の活動がゆらぐことで、本研究で生成したような敵対的

サンプルに対する脆弱性が減り、不可解な振る舞いが起きにくくなっている可能性が考えられます。

発表者

東京大学大学院医学系研究科機能生物学専攻
大木 研一（教授）〈兼：ニューロインテリジェンス国際研究機構（WPI-IRCN） 副機構長／Beyond AI 研究推進機構 教授〉
浮田 純平（研究当時：博士課程）

論文情報

〈雑誌〉 Neural Networks
〈題名〉 Adversarial attacks and defenses using feature-space stochasticity
〈著者〉 Jumpei Ukita and Kenichi Ohki
〈DOI〉 10.1016/j.neunet.2023.08.022 （2023 年 9 月 15 日追記）

研究助成

本研究は、Beyond AI 連携研究機構、日本医療研究開発機構（AMED）「革新的技術による脳機能ネットワークの全容解明プロジェクト（課題番号：14533320、JP16dm0207034、JP20dm0207048）」、科研費「大脳皮質の領野間相互作用を担う神経回路の細胞・シナプスレベルでの機能解明（課題番号：25221001）」、「多階層光遺伝学による大脳皮質の認知・学習機構の解明（課題番号：19H05642）」、「臨界期における大脳皮質神経回路の多様性形成メカニズムの解明（課題番号：20H05917）」、CREST-JST「多感覚の統一的知覚を担う座標変換回路の解明（課題番号：JPMJCR22P1）」の支援により実施されました。

用語解説

（注 1）深層ニューラルネットワーク
多数の層からなる人工ニューラルネットワークのこと。近年、画像認識など幅広い分野で使われ、大きな進展を遂げた。

（注 2）ゆらぎ
ランダムな変動（ノイズ）のこと。深層ニューラルネットワークの素子にゆらぎがあることで、その素子は入力に対して毎回ランダムに少しだけ異なる出力を返すようになる。

（注 3）敵対的サンプル
深層ニューラルネットワークが実際とは異なる出力をするように意図的に生成されたサンプル（画像認識モデルの場合は画像）。例えば、人間が見ると明らかにパンダであり深層ニューラルネットワークもパンダと認識している画像に対して、人為的にその画像を少しだけ改変することで、人間が見ると改変前の画像との違いは分からないが深層ニューラルネットワークは誤認識させられることが知られている。このような改変された画像を敵対的サンプル（adversarial example）と呼び、Szegedy et al., ICLR, 2014 で始めて報告された。どんな画像に対しても人為的な改変を加えて敵対的サンプルを生成することが可能だが、人為的な改変を加えていない撮影したままの画像を誤認識する確率は小さい。なお、多くの敵対的サンプルに関する論文では、実際に存在する画像に微小な摂動を加えることで敵対的サンプルを生成しているが、本研究では、より広い定義である unrestricted adversarial example [Song et al., NeurIPS, 2018] に基づく。

問合せ先

〈研究に関する問合せ〉
東京大学大学院医学系研究科
教授　大木 研一（おおき けんいち）
Tel：03-5841-3459　E-mail：kohki@m.u-tokyo.ac.jp

〈報道に関する問合せ〉
東京大学大学院医学系研究科 総務チーム
Tel：03-5841-3304　E-mail：ishomu@m.u-tokyo.ac.jp

東京大学国際高等研究所 ニューロインテリジェンス国際研究機構（WPI-IRCN） 広報担当
E-mail：pr.ircn@gs.mail.u-tokyo.ac.jp

東京大学産学協創部（Beyond AI 研究推進機構 広報担当）
E-mail：kyoso-info.adm@gs.mail.u-tokyo.ac.jp
※メールの件名の冒頭に【脳のゆらぎ】と記載ください。