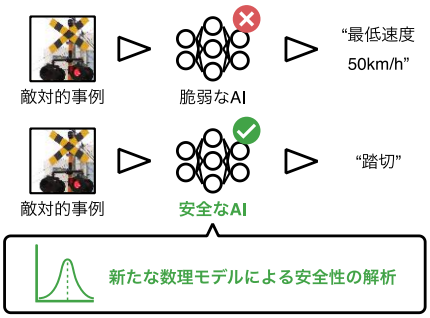
PRESS RELEASE

東京大学

AI を悪意から守れるか
――平均場理論を応用した敵対的訓練の解析――

発表のポイント

◆AI を悪意ある攻撃から守る手法、敵対的訓練に対して、数学的解析を行いました。
◆新たな平均場理論を用いて、AI を安全に機能させるためには、ニューラルネットワークの「幅」が重要であることを明らかにしました。
◆AI の特性を多面的に理解することは、社会と AI の安全な共存にむけた大きな前進です。また、考案した平均場理論は、AI 解析の新たな理論的枠組みとなることが期待されます。



新たな数理モデルによって敵対的訓練を解析することで、より安全な AI の開発を促進

概要

東京大学大学院情報理工学系研究科の熊野創一郎大学院生と山崎俊彦教授を中心とする研究チームは、平均場理論（注 1）を基にした新たな数理モデルを用いて、敵対的事例（注 2）から AI を守る防御手法「敵対的訓練」（注 3）の様々な特性を明らかにしました。

特に興味深い結果として、安全な AI を実現するためにはニューラルネットワーク（注 4）の“幅”構造を広くすることが重要であるという知見を得ました。この特性は多くのニューラルネットワークに対して成り立ち、私たちの解析の幅広い適用可能性を示唆しています。

本研究成果は、2023 年 12 月 10 日から 16 日に米国で開催される、機械学習の分野で最も著名な国際会議 Conference on Neural Information Processing Systems (NeurIPS)にて発表されます。

発表内容

研究の背景

近年、AI は社会に広く浸透しつつあり、時に熟達した人間をも上回る能力を誇ります。しかし、現在の AI 技術にはいくつかの課題が存在しています。「敵対的事例」はそのような課題の一つであり、AI に誤った認識を引き起こさせるように悪意をもって作られます。これは特に人命が関わる事象にとって大きな脅威です。たとえば、自動運転車が敵対的に改ざんされた踏切を誤認識してしまうと、重大な事故を引き起こしてしまうかもしれません（図 1 左図）。

このような問題が起こらないよう、つまり敵対的事例に対しても正しい認識ができるように AI を訓練することを「敵対的訓練」と呼びます。現在、この手法は多くの敵対的事例に一定の耐性を示しており、今後の発展が期待されています。しかしながら、敵対的訓練のメカニズムについては理論的な理解がほとんどなされておらず、未だに「失敗する可能性がある」というリスクを抱えています。このような安全性の不確かさは、特に自動運転のような人命に関わる

システムにとって深刻です。AI の安全性を確かなものにするため、敵対的訓練を理論的に理解し、その効果を数学的に保証することが求められています。本研究では、平均場理論という数理的アプローチを用いた敵対的訓練の解析を目的としています（図 1 右図）。

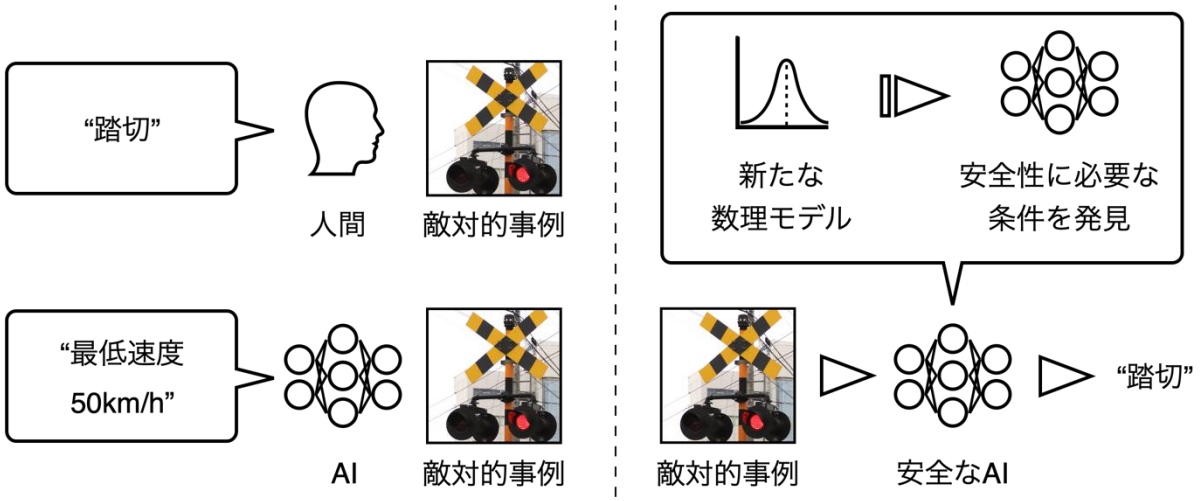


図 1：（左図）敵対的事例の一例。人間には踏切に見えるが、AI は最低速度 50km/h と誤解してしまう可能性がある。（右図）新たな数理モデルによって AI の特性を解析し、安全な AI に求められる条件を明らかにした。

研究内容と手法

本研究では、AI 技術の中心であるニューラルネットワーク上で行われる敵対的訓練を数学的に解析しました。解析ツールとして、これまで敵対的訓練の解析では注目されてこなかった平均場理論に着目しています。平均場理論は、どのような場合に AI が訓練可能になるのか、またどのようにすると高い性能が得られるようになるかなど、ニューラルネットワークの様々な特性を明らかにすることができます。ただし、これまでの平均場理論のままではネットワークのごく狭い範囲しか見通せず、敵対的訓練の解析に用いることはできませんでした。

本研究ではまず、この制限を無くした新たな平均場理論を提案しました。これにより、ニューラルネットワーク全体の情報伝達を簡単な式で表現でき、敵対的訓練によるニューラルネットワークの変化が解析可能となりました。

次にこの理論を利用し、敵対的訓練の様々な特性を明らかにしました。特に興味深い結果として、敵対的訓練が「上手く機能し、高い性能を出す」、すなわち安全な AI を実現するためにはニューラルネットワークの“幅”という構造（図 2）を広くすることが重要であることを導き出しました。近年の AI 技術は別名「深層学習」と呼ばれる通り、できるだけ層を深く、すなわち何段もニューラルネットワークを直列に繋げていくことで優れた性能を発揮します。これに対して、敵対的攻撃への耐性を獲得するにはその深さ方向ではなく、一つ一つのニューラルネットワークの層における入力の並列数を大きくすることが重要であることがわかりました。この特性は多くのニューラルネットワークに成り立つものであり、この解析手法が幅広く適用できる可能性を示唆しています。

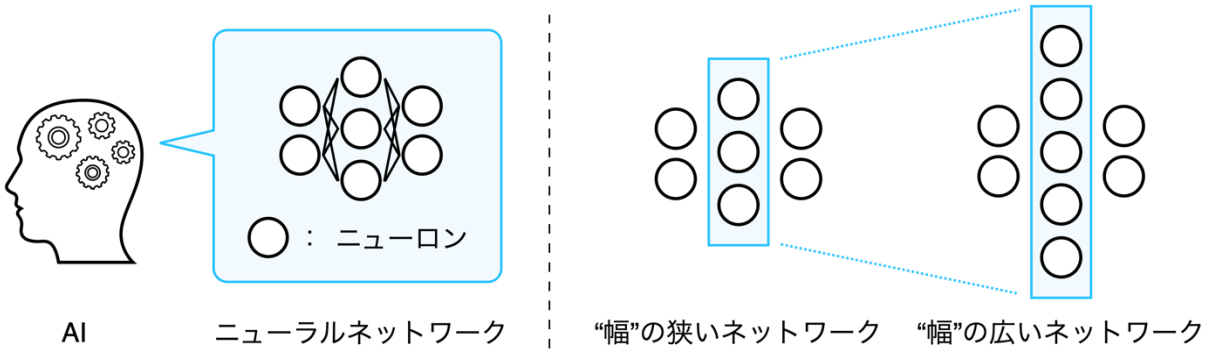


図 2：ニューラルネットワークは近年の AI の中核を担う。本研究ではニューラルネットワークの“幅”が、高い品質の敵対的訓練に必要であることを明らかにした。

社会的意義・今後の予定

敵対的事例は AI の普及が進みつつある私たちの社会にとって大きな脅威であり、この脅威に対する有効な防御策として敵対的訓練が期待されています。本研究の成果は、敵対的訓練に対する包括的な分析結果を提供しており、より安全な AI 開発を進める上での重要な示唆といえます。今後は実際のシステムに組み込める手法の提案など、実装を含めた研究の発展を検討しています。

また、解析ツールとして提案された新たな平均場理論は、敵対的訓練の解析にとどまらず、幅広い深層学習手法の解析に応用可能です。これは停滞しがちな深層学習手法の理論的解析を新たな視点から開拓する一手であり、より進化した AI 開発へと繋がることが期待されます。

発表者・研究者等情報

東京大学大学院情報理工学系研究科
山崎　俊彦　教授
熊野　創一郎　博士課程（日本学術振興会特別研究員）

学会情報

学会名：Conference on Neural Information Processing Systems, 2023
題　名：Adversarial Training from Mean Field Perspective
著者名：Soichiro Kumano*, Hiroshi Kera, and Toshihiko Yamasaki
URL: https://nips.cc/

研究助成

本研究は、科研費「敵対的画像に関する統一的存在定理の確立とそれを利用した敵対的攻撃・防御手法の提案（課題番号：JP23KJ0789）」、「深層ゼロ関数学習の計算と理論（課題番号：JP22K17962）」、JST ACT-X「拡張平均場理論を用いた敵対的訓練の理論的解析（課題番号：JPMJAX23C7）」、Microsoft Research Asia の支援により実施されました。

用語解説

（注1）　平均場理論：深層学習の文脈では、ニューラルネットワークの各パラメータがある法則に従うという仮定の下で構築される理論体系。
（注2）　敵対的事例：明らかにおかしな挙動を AI に引き起こすオブジェクト。
（注3）　敵対的訓練：敵対的事例から AI を守る防御手法の一つ。
（注4）　ニューラルネットワーク：近年の AI において中心となる技術。