2023 年（令和 5 年）3 月 17 日

画像識別 AI の誤識別リスクを効果的・効率的に低減する技術を開発
～自動運転システムにおける安全性ベンチマークにて効果を検証～

情報・システム研究機構 国立情報学研究所（ＮＩＩ、所長：喜連川 優、東京都千代田区）のアーキテクチャ科学研究系 准教授 石川 冬樹らの研究チームは、九州大学（九大、総長：石橋 達朗、福岡県福岡市）大学院 システム情報科学研究院 情報知能工学部門 准教授 馬 雷らの研究チームとともに、画像識別 AI の誤識別に対するリスクを効果的・効率的に低減する技術を開発しました。本研究成果は、科学技術振興機構（ＪＳＴ、理事長：橋本 和仁、東京都千代田区）の未来社会創造事業 Engineerable AI プロジェクト(*1)（通称 eAI プロジェクト、研究開発代表者：NII アーキテクチャ科学研究系 准教授 石川 冬樹）によるものです。

深層ニューラルネットワーク（以下 DNN: Deep Neural Network）では、多数のパラメーターが異なる物体の識別結果に対して複雑に影響するため、ある誤識別を改善するための修正が、他の識別結果に意図しない低下（デグレ）を発生させる問題があります。

本プロジェクトでは、役割の異なる複数の DNN 修正技術を組み合わせ、画像識別用 DNN を狙い通りに修正する研究開発を進めてきました。具体的には、様々な誤識別を分類し、タイプごとの原因と修正方法を発見する技術（NII）、パラメーター修正と誤識別改善の履歴情報を利用することで修正による低下を抑制する技術（NII）、パラメーター値だけでなく DNN の基本構造自体も修正する技術（九大）などに取り組んできました。

自動運転 AI 向けの実験では、自動車企業などを交えて定めた安全性ベンチマークで評価を行い、多数の安全要求を満たした上で狙い通りの修正が可能であり、効果的・効率的にリスクを低減できることを確認しました。

今後、開発した修正技術をフレームワークとして統合するとともに、自動運転のあり方に関するビジョンやポリシー、走行データの特性など企業ごとに異なるニーズに応じた産業実証に取り組んでいきます。

NII、九大の研究成果は、それぞれ、ソフトウェアテストに関するフラッグシップ国際会議 ICST 2023(*2) で 2023 年 4 月（アイルランド時間）、ソフトウェア解析に関するフラッグシップ国際会議 SANER 2023(*3) で 2023 年 3 月 23 日（マカオ時間）、ソフトウェア工学に関するフラッグシップ雑誌 TOSEM(*4) で 2023 年内に発表されます。

【背景】

自動運転や高度運転アシストの技術では、歩行者などの障害物や、標識、走行レーンなどの検出と分類において、画像識別 AI が大きな役割を果たします。画像識別 AI を実現する技術として、深層学習（ディープラーニング）がよく用いられます。数百万あるいはそれ以上の数のパラメーターを持つ深層

大学共同利用機関法人 情報・システム研究機構 国立情報学研究所
国立大学法人 九州大学
国立研究開発法人 科学技術振興機構（JST）

ニューラルネットワーク（DNN）と呼ばれる計算モデルに対し、用意した正解データを用いた訓練を行うことで、それらパラメーターの値を自動的に設定し、画像内の物体識別を実現しています。

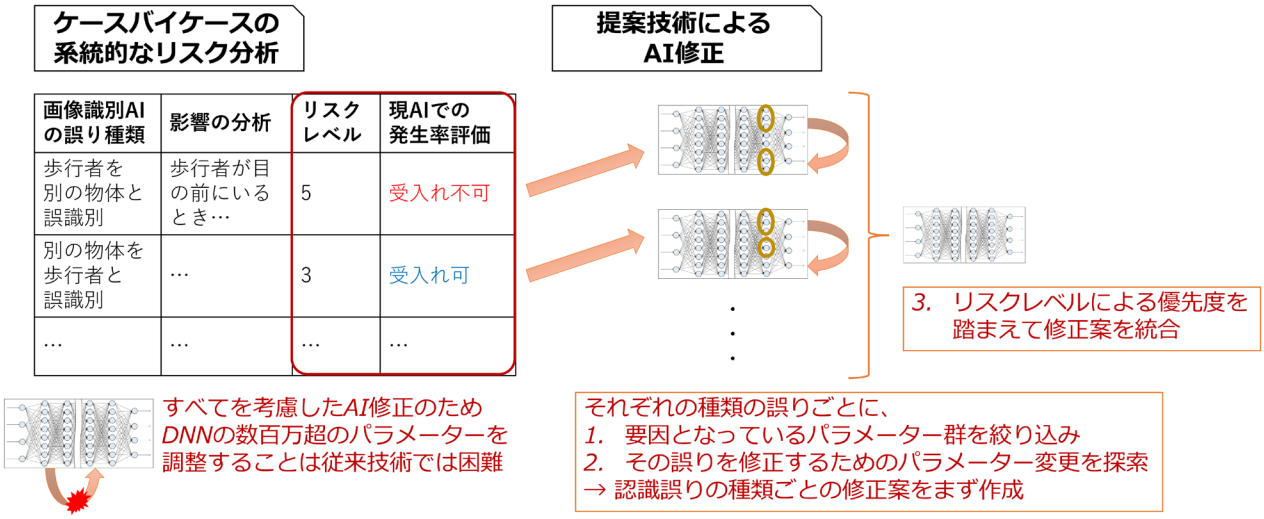
一方、自動運転のように安全性が重要となるシステムにおいては、様々な誤りや故障のリスクが充分に小さいことを示すことが必要です。このために、異なるタイプの誤りや故障ごとに、それぞれがどのような状況下でどれだけの危険につながるかケースバイケースで系統的に分析し、リスクを低減するシステムの設計や設定を追及します。つまり、様々な識別対象や環境条件に対し、AI の誤識別による事故のリスクを評価し、低減していくことが必要となります。このような考え方は、近年発行された AI システムの品質に関するガイドライン（AIQM(*5)、QA4AI(*6)）や、自律動作の安全性に関する標準（ISO 21448(*7)、ANSI/UL 4600(*8)）などでも示されており、画像識別 AI 開発上の共通認識になっていくと考えられます。

しかし、従来の DNN 技術には、誤識別に対する修正が難しいという課題がありました。例として、訓練により得られた DNN において、「歩行者を別の物体と間違える」タイプの誤識別の発生率が高く、開発者がそのリスクを低減したいという状況を考えてみます。この場合、一般的にはその誤識別を正せるような学習データを収集し、DNN に追加して再訓練を行います。しかし、再訓練の結果、DNN の数百万を超えるパラメーター値が再設定されるため、識別性能が期待通りに変化しないことが多くあります。そのため、狙い通りの修正を行うためには、訓練方法の試行錯誤が必要になります。また、ある誤識別を改善するための修正が、他の識別結果に意図しない低下（デグレ）を発生させ、今まで識別できていたものが識別できなくなることもあります。このように、従来の DNN 技術には、複数タイプの致命的な誤識別に起因するリスクを狙い通りに低減できないという課題がありました。

【NII の技術開発】

本研究開発では、様々な誤識別を分類し、タイプごとに要因となるパラメーター群を探索することで、これらの問題を解決する DNN 修正技術を開発しました。

本技術では、以下の方法で、様々な誤識別に対する効果的な修正候補を効率よく見つけ出します。最初に、「歩行者をバイク搭乗者と間違える」、「電車をバスと間違える」といったそれぞれの誤識別タイプに対し、その要因となっているパラメーター群を欠陥局所化（Fault Localization）という技術をもとにした方法で絞り込みます。次に、その誤りを修正するためのパラメーター変更のパターンを探索します。最後に、異なる誤識別のタイプそれぞれに対して効果的な修正候補を見いだした上で、それぞれの誤識別のリスクの大きさを踏まえてそれらの修正候補を統合し（図 1）、従来の DNN 修正技術では困難であったリスクの効果的・効率的な低減を実現します。



<図 1>修正パラメーター探索とリスクレベルを踏まえた統合の仕組み

本技術によるリスク低減効果は、自動車企業などの実務者を交えたワーキンググループにおいて定めた安全性ベンチマークのプロトタイプによって評価しました。本ベンチマークでは、自動運転における誤識別のタイプごとの影響を分析することで、12 種類の誤識別を 3 段階のリスクレベルで分類し、総合リスクスコアを評価しました。その結果、従来では困難だった複数タイプの認識別を踏まえた修正が可能となることが実証されました。この実証結果については、2023 年 4 月にソフトウェアテストに関するフラッグシップ国際会議 ICST 2023(*2)にて発表予定です。

また、NII の研究チームは富士通株式会社（富士通、代表取締役社長：時田 隆仁、東京都港区）と協働して、重要な識別対象に対する意図せぬ識別性能の低下を抑制する技術を開発しました（2022 年 3 月発表の成果(*9)）。この技術は、パラメーター修正と誤識別改善の履歴情報を利用することで、重要な対象に対する識別性能の低下を抑制します。2022 年秋には、富士通を中心とした社会セキュリティ AI 向けの実験で、性能低下を抑制した DNN 修正の効果が実証されました。この実証結果については、2023 年 3 月にソフトウェア解析に関するフラッグシップ国際会議 SANER 2023(*3)の産業応用トラックにて発表予定です。

【九大の技術開発と実証評価】

九大では、馬雷准教授、趙建軍教授を中心として、複数の異なるアプローチで AI 修正の技術に取り組みました。2021 年に発表された技術(*10)では、運用時に検出された未知のノイズパターンに対応するために、スタイル移転と呼ばれる技術を用い、失敗を引き起こした画像の特徴を訓練データに付与します。この手法により、エンジニアが明示的に言葉にできない失敗の傾向や、開発時の想定が困難な運用時ノイズ分布に対応する修正を実現しました。2022 年度に取り組んだ技術では、DNN のパラメーター値だけでなく、DNN の基本構造自体を修正することで、より幅広い修正を実現しました。この技術は、

NII や富士通によるものも含め、他の DNN 修正技術と組み合わることで、その修正の効果・効率を高めることが期待されます。この技術成果については、2023 年にソフトウェア工学におけるフラッグシップ雑誌である ACM TOSEM(*4)に掲載予定です。

【eAI プロジェクトでの成果と展望】

eAI プロジェクトでは、NII、九大、富士通が中心となり、産業界における異なるユースケース・要求を踏まえた「AI 修正ツール」の開発を進めてきました。これらの技術に対し、自動車企業や安全性の専門家らとともに定めた安全性ベンチマークを反復的に実施することで、産業界のニーズに応える技術の開発・評価に取り組んでいます。

今後はこれまでに得られた、役割の異なる複数の AI 修正技術を統合し、自動運転の品質・安全性・信頼性に対する多様なニーズを満たすべく、細やかで複合的な安全要求に適合した画像識別 AI に関する本格的な実証実験を行っていきます。具体的には、今後、自動運転のあり方に関するビジョンやポリシー、走行データの特性など企業ごとに異なるニーズに応じた産業実証に取り組んでいきます。

eAI プロジェクトではもう一つの大きな軸として、東京工業大学科学技術創成研究院 鈴木 賢治（すずき けんじ） 教授を中心に、データの量が限られている場合でもニーズを踏まえ信頼できる AI を構築できる技術の研究開発、および医療分野でのその実証にも取り組んでいます。さらに、早稲田大学理工学術院基幹理工学部 鷲崎 弘宜（わしざき ひろのり） 教授を中心に、細やかな要求やリスクの分析から AI の構築・修正、運用管理までを一気通貫で扱う開発支援フレームワークの構築と実証を行っています（計算機科学のフラッグシップ雑誌である IEEE Computer にて発表 (*11)）。そして、これらの取り組みにより、産業界・社会の細やかなニーズに応える AI のための工学的方法論を確立していきます。

石川 冬樹 eAI プロジェクトリーダー（NII 准教授）のコメント：

「eAI プロジェクトは、機械学習・深層学習を用いた AI システムの品質やソフトウェア工学技術へのニーズの高まりを受けて立ち上げたプロジェクトです。NII、九州大学、富士通では、産業界の方々との議論を通じて、「それなりに動く AI」ができた先にある、AI 修正という重要な課題に注力してきました。従来ソフトウェアシステムでは、修正、あるいはデバッグ、その際のデグレ抑制といったタスクは、最も問題が多発しコストがかかる部分でした。AI に対しても、リスクを綿密に分析し低減していくことが重要となる応用領域では、AI 修正は避けて通れない課題です。

自動運転領域に関して NII では、蓮尾 一郎教授らの ERATO-MMSD プロジェクトにおいて、自動運転の計画・制御の安全性について多数の研究開発を行ってきました(*12)。eAI プロジェクトではこれと補完・連動する形で、不確実性が高い画像識別 AI に主眼をおいて取り組んでいます。今後自動車企業とのさらなる議論・実証を通し、安全性改善のために DNN による画像識別 AI を修正する技術を高めていきます。また画像識別 AI を補完するような周りの技術とも組み合わせ、自動運転の安全性全体を論証・改善していくための包括的な枠組みに取り組んでいきます。」

大学共同利用機関法人 情報・システム研究機構 国立情報学研究所
国立大学法人 九州大学
国立研究開発法人 科学技術振興機構（JST）

【研究プロジェクトについて】

本研究開発は科学技術振興機構 未来社会創造事業 サイバー世界とフィジカル世界を結ぶモデリングと AI　超スマート社会の実現領域 「機械学習を用いたシステムの高品質化・実用化を加速する"Engineerable AI"技術の開発」プロジェクト（JPMJMI20B8）の一環で行われました。

【論文タイトルと著者（NII)】

タイトル： Distributed Repair of Deep Neural Networks
著　　者： Davide Li Calsi, Matias Duran, Xiao-Yi Zhang, Paolo Arcaini, Fuyuki Ishikawa
発表会議： The 16th IEEE International Conference on Software Testing, Verification and Validation (ICST 2023)
発 表 日： 2023 年 4 月 16～20 日（調整中）　口頭発表予定（アイルランド時間）

【論文タイトルと著者（九大)】

タイトル： ArchRepair: Block-Level Architecture-Oriented Repairing for Deep Neural Networks
著　　者： Hua Qi, Zhijie Wang, Qing Guo, Jianlang Chen, Felix Juefei-Xu, Fuyuan Zhang, Lei Ma, Jianjun Zhao
掲 載 誌： ACM Transactions on Software Engineering and Methodology (TOSEM)
発 表 日： 2023 年予定

【論文タイトルと著者（NII、富士通)】

タイトル： An Experience Report on Regression-Free Repair of Deep Neural Network Model
著　　者： Takao Nakagawa, Susumu Tokumoto, Shogo Tokui, Fuyuki Ishikawa
発表会議： The 30th IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER 2023, Industry Track)
発 表 日： 2023 年 3 月 23 日（木）口頭発表予定（マカオ時間）

〈メディアの皆様からのお問い合わせ先〉

大学共同利用機関法人 情報・システム研究機構 国立情報学研究所
総務部企画課 広報チーム
TEL：03-4212-2164　E-mail：media@nii.ac.jp

国立大学法人 九州大学
広報室
TEL：092-802-2130　E-mail：koho@jimu.kyushu-u.ac.jp

国立研究開発法人 科学技術振興機構（JST）
広報課
TEL：03-5214-8404　E-mail：jstkoho@jst.go.jp

〈JST の事業に関すること〉

国立研究開発法人 科学技術振興機構（JST）
未来創造研究開発推進部　小泉 輝武
TEL：03- 6272-4004　E-mail：kaikaku_mirai@jst.go.jp

(*1) Engineerable AI プロジェクト：JST における「未来社会創造事業　サイバー世界とフィジカル世界を結ぶモデリングと AI　超スマート社会の実現領域」に採択されている研究プロジェクトで、自動運転をはじめとして深層学習技術を用いた AI システムの安全性・信頼性確保・向上のため、細やかなニーズに応える AI の構築や修正が可能な技術と、医療・交通の二領域における概念実証に取り組む。正式名称は「機械学習を用いたシステムの高品質化・実用化を加速する"Engineerable AI"技術の開発」、略称は eAI プロジェクト。 https://engineerable.ai/

(*2) ICST 2023 : The 16th IEEE International Conference on Software Testing, Verification and Validation。CORE と呼ばれる計算機科学系の国際会議ランキングにて A ランク。

(*3) SANER 2023 : The 30th IEEE International Conference on Software Analysis, Evolution and Reengineering。CORE と呼ばれる計算機科学系の国際会議ランキングにて A ランク。

(*4) TOSEM : ACM Transactions on Software Engineering and Methodology。CORE と呼ばれる計算機科学系の雑誌ランキングにて A*ランク。

(*5) AIQM : 機械学習品質マネジメントガイドライン。産業総合研究所が中心となってまとめた AI システムの品質に関するガイドライン。考慮すべき品質の種類（品質特性）やそれらのレベル分けなど概念が規範的に整理されているのが特徴。 https://www.digiarc.aist.go.jp/publication/aiqm/

(*6) QA4AI : AI プロダクト品質保証ガイドライン。ソフトウェア品質やテスト技術の専門家が合同でまとめた AI システムの品質に関するガイドライン。具体的なテスト計画などエンジニア視点での指針が整理されているのが特徴。 https://www.qa4ai.jp/

(*7) ISO 21448 : 自動車の安全のうち、故障の考慮ではなく、センサー・アクチュエーターを活用した自律的な機能の安全性を主眼においた標準。通称 SOTIF (safety of the intended functionality)。

(*8) ANSI/UL 4600 : 機械学習技術などを用い自律動作を行うプロダクトにおいて、セーフティケースと呼ばれる安全性論証の枠組みを示した標準。

(*9) Tokui ら, NeuRecover: Regression-Controlled Repair of Deep Neural Networks with Training History, The 29th IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER 2022)

(*10) Yu ら, DeepRepair: Style-Guided Repairing for Deep Neural Networks in the Real-World Operational Environment, IEEE Transactions on Reliability, Vol. 71 No. 4, 2021

(*11) Washizaki ら, Software-Engineering Design Patterns for Machine Learning Applications, IEEE Computer, Vol. 55 No. 3, 2022

(*12) ERATO 蓮尾メタ数理システムデザインプロジェクト：JST「戦略的創造研究推進事業 ERATO」における研究プロジェクトで、自動運転システムを中心として、物理世界で動作する自律システムの検証技術に、多様な学術分野の協働と統合を通して取り組んできた。
https://www.jst.go.jp/erato/hasuo/ja/
これまでの成果として以下の NII プレスリリースがある。
- 「自動運転車の安全性に数学的証明を与える新手法を開発、～論理的安全ルールの効率的導出により自動運転の社会受容を加速～」、2022 年 7 月
https://www.nii.ac.jp/news/release/2022/0707.html
- 「自動運転における重大な問題をシミュレーションで検出する技術を開発～問題が発生するかを探り、起こりうる問題だけを効率的に自動探索～」、2021 年 11 月
https://www.nii.ac.jp/news/release/2021/1115.html