国立研究開発法人
国立成育医療研究センター
National Center for Child Health and Development

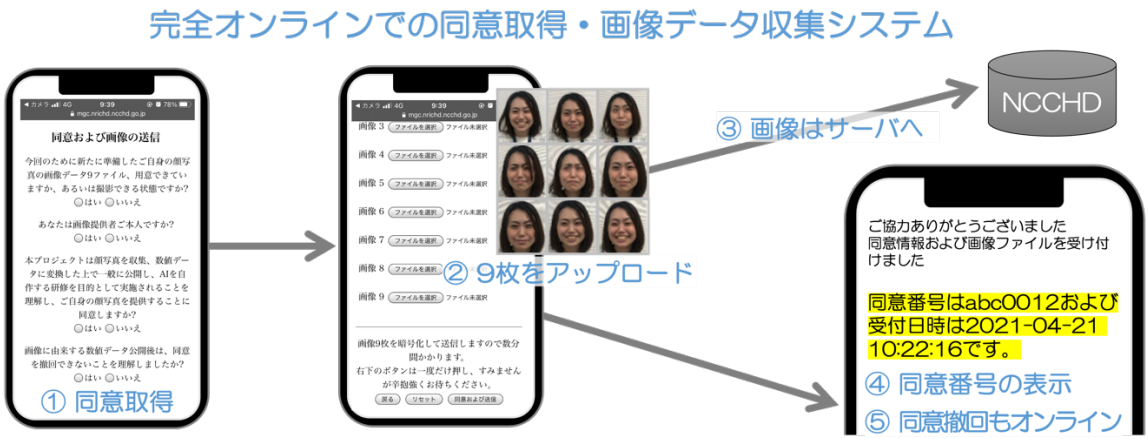
NEWS RELEASE

報道関係者各位

2022 年 03 月 9 日
国立成育医療研究センター

同意を得た顔写真収集でディープラーニングのためのデータセット作成
AI 構築に必要なデータ量の検討に活用、小児の先天性疾患への応用期待

国立成育医療研究センター（所在地：東京都世田谷区、理事長：五十嵐隆）は、内閣府の戦略的イノベーション創造プログラム（SIP）の AI（人工知能）ホスピタルによる高度診断・治療システム事業への採択を機に、研究者、医師だけでなく全ての職員を対象としたデータサイエンスの啓蒙、教育活動を行っています。その研修における取り組みの一つとして、教師あり機械学習[1]を用いて AI を構築するための顔写真データセットを作成しました。インターネットを利用し顔画像は比較的容易に集めることができますが、医学的および生物学的な応用を考えると姿勢は固定されていることが望ましく、さらに個人情報も保護する必要がありました。しかし、これまでにない規模の性別や笑顔などのラベル付きビッグデータ（2429 枚、277 名分）となり、データ拡張、転移学習[2]を活用した畳み込みニューラルネットワーク[3]により、性別は 98.2%、笑顔は 93.0%の認識精度を示すモデルを提示することができました。



小児の先天性疾患は特徴的顔貌を示すことが多く、画像認識による診断に期待が寄せられています。しかし、それらは希少疾患のため対象人数が少なく、大量データを必要とするディープラーニングによる実装を阻んでいます。今回の結果は AI 構築に必要とされるデータ量の検討に活用され、今後の AI 開発に指針を与えるものです。

---

[1] 教師あり機械学習：データにそのデータの分類結果などの正解を添え、コンピュータで実行させるプログラムに学習させる手法。

[2] 転移学習：あるデータで得られた既存の結果を、初期値などとして関心のある別な問題の学習に利用する手法。

[3] 畳み込みニューラルネットワーク：画像などに存在するデータの空間的な位置関係を、データ処理の流れに組み込んで学習を行うためのモデル。

【プレスリリースのポイント】

- ウェブスクレーピング[4]ではなく、個人情報に配慮し、同意を得た上で、姿勢が固定された顔画像を収集する仕組みを作り、実施しました。
- データはラベルとともに教育用として公開しています。
- AI モデルは全て無償ソフトウェアを利用して構築し、ソースプログラムを教育用に公開しています。
- 教師あり機械学習に必要となる性別や笑顔などのラベル情報は、提供者からは収集せず、研究メンバーの主観による多数決で準備しました。
- AI 構築に必要とされるデータ量の検討に活用され、小児の先天性疾患への応用が期待されます。

【背景・目的】

医療における AI 開発が注目される中、従事者の多くはハードウェアの性能やソフトウェア開発に気を取られ、教師あり学習におけるデータ自体の重要性や、データ収集、整理、アノテーション[5]の過程で多くの案件がつまずいていることを理解できていない状況にありました。国立成育医療研究センターのデータサイエンス研修では、特にデータ収集からアノテーションまでを体験することを企画し、特徴的顔貌を有する小児疾患の画像診断に役立てることを目指し、独自に顔画像を収集し解析しました。

【今後の展望・発表者のコメント】

医療や医学に限らず、顔写真の収集と解析は社会的に大きな意義がありますが、それには個人情報の保護への配慮が必要になります。今回の顔写真収集は、ある程度の成功を収めましたが、写真提供を拒まれる場面も多くあり、この手法のまま規模を拡大することは困難です。研究チームでは、準同型暗号[6]による解析を利用したブロックチェーンにより、個人情報が完全に守られた顔画像の社会での利用方法を検討しています。

【発表論文情報】

著者: 青砥早希, 半谷まゆみ, 上野瞳, 植田亜季, 五十嵐麻希, 伊藤愛主, 塚本元子, 神野智子, 坂本美佳, 岡﨑有香, 長谷川冬雪, 緒方広子, 名村彩季, 小島一晃, 菊谷昌央, 松原圭子, 谷口公介, 岡村浩司*
所属: 国立成育医療研究センター データサイエンス研修
タイトル: Collection of 2429 constrained headshots of 277 volunteers for deep learning
掲載誌: *Scientific Reports* (2022)
URL: https://www.nature.com/articles/s41598-022-07560-2
DOI: https://doi.org/10.1038/s41598-022-07560-2

---

[4] ウェブスクレーピング：ウェブサイトに保存および公開されているデータを、目的に従って自動的に取得するプログラミング等の技術。
[5] アノテーション：データに対して、教師あり機械学習のための正解となる注釈を付ける作業。
[6] 準同型暗号：暗号化したデータを複号せず、暗号化したままデータ処理を行うことができるデータ方式

国立研究開発法人
国立成育医療研究センター
National Center for Child Health and Development

NEWS RELEASE

【特記事項】

本研究は内閣府の戦略的イノベーション創造プログラム（SIP）「AI（人工知能）ホスピタルによる高度診断・治療システム」プロジェクトの支援を受けて行われました。性の判別については武田科学振興財団の特定領域研究「習学的アプローチによるヒトの性の多様性の解明」プロジェクトからも支援を受けました。

【問い合わせ先】

国立研究開発法人　国立成育医療研究センター

企画戦略局 広報企画室　近藤・村上

電話：０３－３４１６－０１８１（代表）

E-mail:koho@ncchd.go.jp