公開

（事後評価・研究者用）

# 研　究　報　告　書

## 「頑強なハイブリッド深層学習モデルの自動探索システム」

研究期間：2018 年 10 月～2020 年 3 月
研究者番号：50166
研 究 者：Danilo Vasconcellos Vargas （ダニロ　ヴァスコンセロス　ヴァルガス）

### 1．研究のねらい

本研究の代表者は、これまでの研究では、一つのピクセルを変えることでニューラルネットワークを誤魔化すことが可能と紹介した。その発見は畳みこみニューラルネットワークの脆弱性を表すとともに、画像を知的に理解していないことを実証している。この脆弱性の原因は畳みこみニューラルネットワークのモデルである。しかし、モデルの種類とパラメーターは複数あり、一番適切なモデルとパラメーターを見つけることは非常に時間がかかる。更に、深層学習のモデルはその問題を解決できない可能性もある。従って、本研究は最適化を利用し、自動的に頑強なハイブリッド深層学習を探索し、ロバスト性が高い深層学習の構造を発見する。その目的に当たって、本研究は今までニューラルネットワークへの脆弱性を見つけるため利用された攻撃を評価関数のように利用し最適化を行うことによって成り立っている。

### 2．研究成果

#### (1)概要

代表者の研究では始めて一つのピクセルだけで畳み込みニューラルネットワークという広く使われている人工知能のアルゴリズムを誤魔化すことが可能と紹介した[研究業績リスト（1） 1]（図１）。その脆弱性はセキュリティの問題を表すだけではなく、人工知能のアルゴリズムが実際に何を理解しているかも説明している（BBC News に公開された申請者の研究　http://www.bbc.com/news/technology-41845878）。代表者として29 年度に「数理・データサイエンスに関する教育・研究支援プログラム」という助成金に採択された。

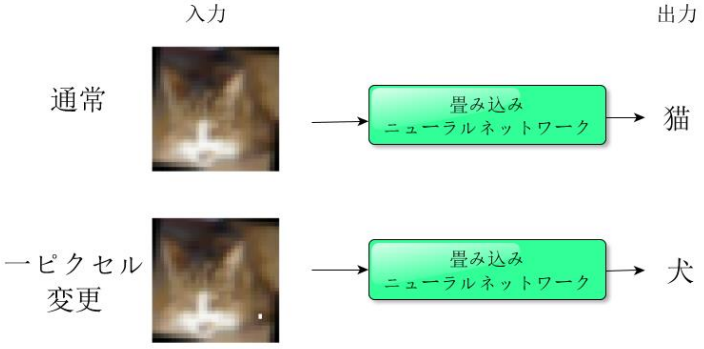


図 1　代表者の研究における一つの Pixel だけで畳み込みニューラルネットワークをごまかす例。

しかし、この研究では畳みこみニューラルネットワークだけの脆弱性を紹介した。どうすれば、その問題を解決できるかはまだ大きな課題になってい

る。セキュリティの問題だけではなく、人工知能の基礎的な問題にも繋がっている。

本研究の目的は、頑強なハイブリッド深層学習システムの探索と開発である。なお、本研究のコンテクストでは「頑強な」深層学習システムとは脆弱性が少ないことである。

**(2)詳細**

本研究は三つのステージに分かれている。

ステージ(a)．**深層学習の脆弱性の評価。**本課題の目標は、深層学習の脆弱性を評価できるシステムの開発である。その評価システムは過去研究で既に作成した一つのピクセルの脆弱性を見つけるアルゴリズムとともに画像全体にノイズを加えることで脆弱性を調査するアルゴリズムから成り立っている。

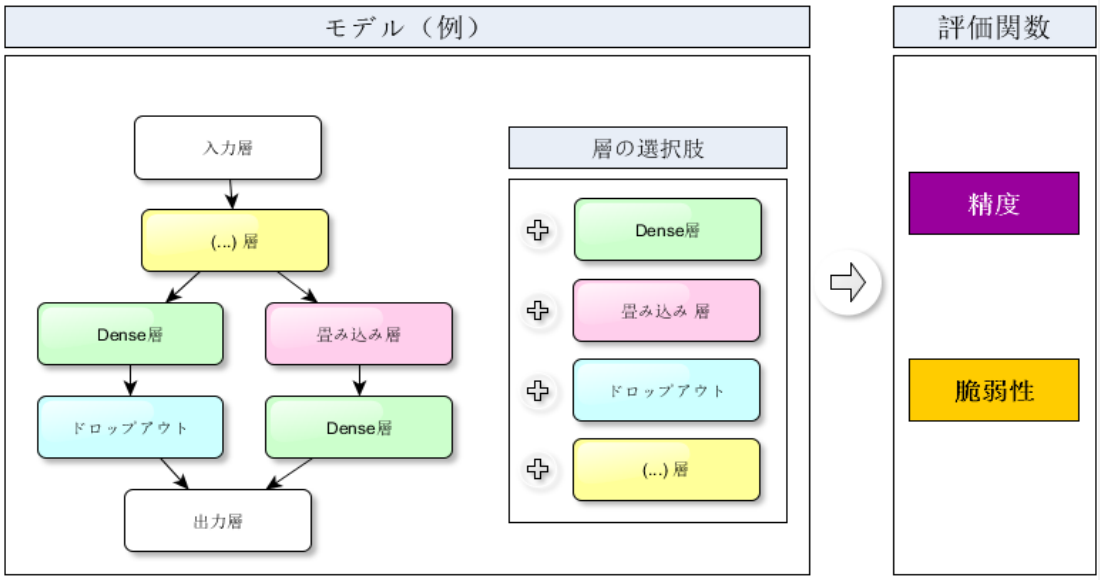


図 2　開発した脆弱性の少ない深層学習を探索する手法。

ステージ(b)．**多種類の深層学習を統一するモデルの開発。**本課題の目標は、多種類の層（ドロップアウト、Batch Normalize する層、異なる活性化関数の層、など)、異なる構造（層の数、層の繋がり、など）と異なる学習手法の深層学習を統一するモデルの開発である。そのモデルの特徴はパラメーターによって変更できるように作成する。つまり、モデルのパラメーターを変える事で層の数、種類、接続などが変えられる。

ステージ(c)．**多種類の深層学習を統一するモデルを探索できる最適化アルゴリズムの開発と実行**。本課題の目標は、(a) 課題で作成した目的関数と普通の精度の目的関数を同時に利用するとともに、(b) 課題で作成したモデルを利用し、そのモデルのパラメーターで探索可能な最適化アルゴリズムの開発である（図２）。

上記に記載している手法を開発し、脆弱性の少ない構造を持つ深層学習を探索できるようになった。その構造は特別な学習なしで、高いロバスト性を用いる深層学習を生成

した　[2]。さらに、様々な脆弱性の少ない構造を調べることで、ロバスト性に関係がある特徴が明らかになった。

追加に、上記の研究を進みながら、他にも二つの調査研究が行った：（ア）脆弱性を理解する研究と（エ）探索空間が加えていない脆弱性の少ない手法の調査である［1], [3]。その調査研究によって、ロバスト性が深層学習の表現力に関連していると紹介し、より深く脆弱性の問題を理解でき、今後の研究の最も重要な項目であった。

## 3.　今後の展開

今後の研究は探索手法の探索空間に含めていない人工知能を調べる予定である。特に、調査研究で理解した重要な特徴（より良い表現力を持つ人工知能など）をもつ珍しい学習の仕方、もしくは、新たな学習パラダイムを試す。

上記の研究に取り組むことで、以下のことが期待されます：

- 社会 5.0 の大事な一歩　－　本研究が脆弱性の少ない深層学習を作成することで、より精度を向上させるとともに安全な応用ができるようになる。それは社会 5.0 にある自動運転、医療、ロボット工学などのような応用には重要な事である。
- 機械学習の様々な分野への貢献　―　ACT-I の研究成果にも紹介したように、脆弱性を解決すると他のハイレベルなコンセプトの質に関する問題（Meta Learning、Zero-Shot Learning, Transfer Learning, など）も同時に解決される確率が高いである。
- 次世代人工知能の発見可能　－　ハイブリッド機械学習や新パラダイムを試すことで、最先端を超える次世代人工知能を見つける可能性がある。

## 4.　自己評価

・研究目的の達成状況

脆弱性の少ない構造を探索するシステムを開発完了し、そのシステムの改善にも取り組みました。更に、（ア）探索空間に加えていない脆弱性の少ない手法を探索し、（エ）脆弱性の理解に関して研究も進んでた。

・研究課題の独創性・挑戦性

上記の成果では、以下のことが初めて明らかにしました：

- ロバスト構造　－　初めて、他の訓練をせずにロバスト性を向上する構造が可能と紹介した。
- 文字列の Universal 脆弱性の発見　―　どのサンプルでもルールによって、探索をせずに、Adversarial Sample に変更することが可能と紹介した。

- ロバスト性を持つための特徴　－　様々な研究成果が特別な特徴（Feedback, Dynamic Routing, など）を持つことでロバスト性が上がることを示した。
- Adversarial Training にバイアスがあるため、訓練を変えることで解決できない可能性が高いと紹介した。

## 5.　主な研究成果リスト

### （1）論文（原著論文）発表

| |
|---|
| 1. Li, D., Vargas, D. V., & Kouichi, S. (2019, June). Universal Rules for Fooling Deep Neural Networks based Text Classification. In 2019 IEEE Congress on Evolutionary Computation (CEC) (pp. 2221-2228). IEEE |
| 2 Danilo V. Vargas, Shashank Kotyan (2020) Towards the Evolution of Neural Architectures Robust against Adversarial Samples. In Proceedings of the GECCO (Companion) Accepted |
| 3. De Melo, V. V., Vargas, D. V., & Banzhaf, W. (2019, July). Batch tournament selection for genetic programming: the quality of lexicase, the speed of tournament. In Proceedings of the Genetic and Evolutionary Computation Conference (pp. 994-1002). |

### （2）特許出願

研究期間累積件数：0　件

### （3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

Van Uytsel, S., & Vargas, D. V. (2020). Adversarial Machine Learning: A Blow to the Transportation Sharing Economy. In *Legal Tech and the New Sharing Economy* (pp. 179-208). Springer, Singapore.

CVPR 2019 の Workshop を含め、招待講演 7 件

CEATEC での研究発表, 2019 年