様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　４ 年　 ６ 月　 ７ 日現在

機関番号：１７１０２
研究種目：若手研究
研究期間：2020～2021
課題番号：２０Ｋ１９８２４

研究課題名（和文）Learning Internal Representations Robust against Adversarial Attacks

研究課題名（英文）Learning Internal Representations Robust against Adversarial Attacks

研究代表者
ＶＡＲＧＡＳ　ＤＡＮＩＬＯ（Vargas, Danilo）

九州大学・システム情報科学研究院・准教授

研究者番号：００７９５５３６

交付決定額（研究期間全体）：（直接経費）　　1,900,000円

研究成果の概要（和文）：本研究では、DNNが学習した内部表現を評価・改善することで、DNNの頑健性に取り組むことを提案した。DNNの内部表現の評価については、特徴の伝達性が敵対的攻撃への頑健性につながることを発見した。つまり、特徴の伝達性が高いほど、敵対的な攻撃に対する頑健性が高いことを発見した。また、複数層のDNNをまとめてグラフで評価・可視化し、多次元空間での形状を容易に点検できるK-spectrumを提案した。DNNの内部表現の改善に関しては、提案にあるように、ネットワークの頑健性を向上させるGANを用いたシステムを開発した。
この研究の成果は、雑誌やプロシーディングスに掲載され、合計で13以上の論文がある。

研究成果の学術的意義や社会的意義
Critical systems such as autonomous driving and medical applications require robust machine learning algorithms. This research paves the way to better algorithms that will allow for such applications to become a reality.

研究成果の概要（英文）：Here, I proposed to tackle the robustness of DNNs by evaluating and improving the internal representation learned by DNNs. Regarding the evaluation of the internal representation of DNNs, we discovered that the transferability of features links to robustness to adversarial attacks. In other words, the better the transfer of features the better the robustness to adversarial attacks. We also proposed K-spectrum which can evaluate and visualize multiple layers of DNNs together in a graph, allowing for easy inspection of how their shapes are in multi-dimensional space. Regarding the improvement of the internal representation of DNNs, we have developed as described in the proposition a GAN based system to improve the network robustness. The system outperformed the state-of-the-art and is being submitted to a journal now.
Results of this research were published in journals and proceedings, more than 13 articles in total.

研究分野：Artificial Intelligence

キーワード：Robust AI　Robust Machine Learning

様　式　C－１９、F－１９－１、Z－１９（共通）

１．研究開始当初の背景

Current Deep Neural Networks (DNN) are known to possess many vulnerabilities which make their application to many fields unsafe. In my past work, the One-Pixel Attack, it was revealed that even very small changes are able to change the classification of DNNs. Many defenses have been proposed, including Adversarial Training, however, all of them have the same vulnerabilities.
In our last investigations, it was understood that the problem lies in the fact that DNNs focus on the texture rather than the shape in their representation. Generative Adversarial Networks (GAN), however, learn to encode, decode as well as transform images and are known to learn internally complex models of the input that goes beyond texture.

２．研究の目的

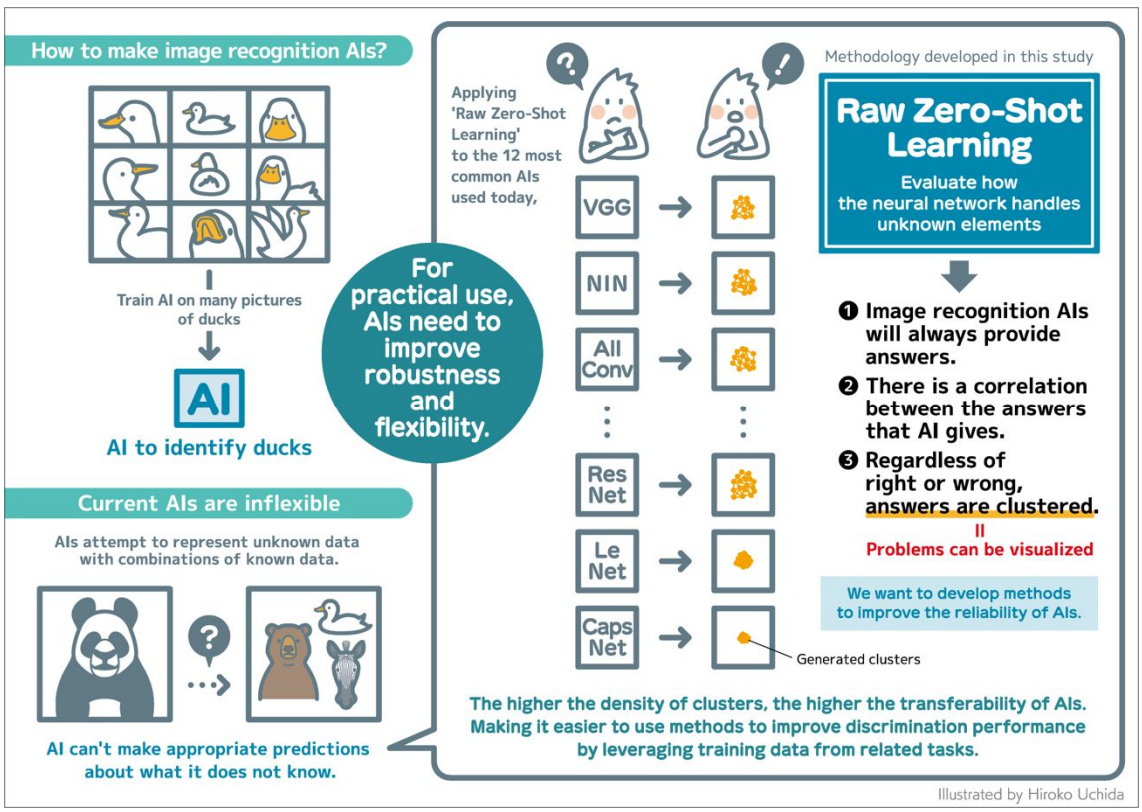
Here, I proposed to tackle the robustness of DNNs by evaluating and improving the internal representation learned by DNNs.

３．研究の方法

Investigation of DNN's features and representation as well as improving the representation of DNNs with GAN based systems, aiming to increase their robustness.

４．研究成果

Regarding the evaluation of the internal representation of DNNs, we discovered that the transferability of features links to robustness to adversarial attacks. In other words, the better the transfer of features the better the robustness to adversarial attacks. We also proposed K-spectrum which can evaluate and visualize multiple layers of DNNs together in a graph, allowing for easy inspection of how their shapes are in multi-dimensional space.



Above figure illustrates one of the key results.

Regarding the improvement of the internal representation of DNNs, we have developed as

described in the proposition a GAN based system to improve the network robustness. The system outperformed the state-of-the-art and is being submitted to a journal now. An automatic search mechanism for robust networks were also implemented to give insight into what constitutes robust DNNs and check if we can find robust ones.

Moreover, we also find out a novel paradigm (SyncMap) which outperformed all methods of the state of the art and defines learning by the equilibrium of dynamical systems. The system was shown to be adaptive and robust to all problems tested.

５．主な発表論文等

〔雑誌論文〕　計6件（うち査読付論文　6件／うち国際共著　6件／うちオープンアクセス　6件）

１．著者名 D. V. Vargas and T. Asabuki	４．巻 in press
２．論文標題 Continual General Chunking Problem and SyncMap	５．発行年 2021年
３．雑誌名 Proceedings of the AAAI21	６．最初と最後の頁 in press
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Tenorio Raul Horacio Valencia、Sham Chiu Wing、Vargas Danilo Vasconcellos	４．巻 1
２．論文標題 Preliminary study of applied binary neural networks for neural cryptography	５．発行年 2020年
３．雑誌名 Proceedings of the GECCO 2020 Companion	６．最初と最後の頁 291-292
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3377929.3389933	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Kotyan, S. and D. V. Vargas	４．巻 1
２．論文標題 Towards Evolving Robust Neural Architectures to Defend from Adversarial Attacks	５．発行年 2020年
３．雑誌名 Proceedings of the GECCO 2020 Companion	６．最初と最後の頁 290-291
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Anh Duc Ta and D. V. Vargas	４．巻 1
２．論文標題 Towards improvement of SUNA in Multiplexers with preliminary results of	５．発行年 2020年
３．雑誌名 Proceedings of the GECCO 2020 Companion	６．最初と最後の頁 289-290
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

1．著者名 D. V. Vargas and Su	4．巻 1
2．論文標題 Understanding the one-pixel attack: Propagation maps and locality analysis	5．発行年 2020年
3．雑誌名 CEUR Workshop Proceedings	6．最初と最後の頁 1-8
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

1．著者名 Kotyan, S. and D. V. Vargas	4．巻 1
2．論文標題 Evolving Robust Neural Architectures to Defend from Adversarial Attacks	5．発行年 2020年
3．雑誌名 CEUR Workshop Proceedings	6．最初と最後の頁 1-8
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

〔学会発表〕　計4件（うち招待講演　1件／うち国際学会　2件）

1．発表者名 D. V. Vargas
2．発表標題 一ピクセルで誤魔化される人工知能が人間を超えた？
3．学会等名 第7回AI Optics研究会（招待講演）
4．発表年 2020年

1．発表者名 D. V. Vargas
2．発表標題 On The Deeper Secrets of Deep Learning
3．学会等名 IJCAI20（国際学会）
4．発表年 2020年

１．発表者名 D. V. Vargas
２．発表標題 On The Deeper Secrets of Deep Learning
３．学会等名 WCCI20（国際学会）
４．発表年 2020年

１．発表者名 Kotyan, S. and D. V. Vargas
２．発表標題 Is Neural Architecture Search A Way Forward to Develop Robust Neural Networks?
３．学会等名 JSAI2020
４．発表年 2020年

〔図書〕　計1件

１．著者名 Van Uytsel, S. and D. V. Vargas	４．発行年 2021年
２．出版社 Springer	５．総ページ数 228
３．書名 Autonomous Vehicles: Business, Technology and Law	

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関