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　６ 年　 ５ 月　 ２ 日現在

機関番号：３３９１８
研究種目：基盤研究(C)（一般）
研究期間：2020～2023
課題番号：２０Ｋ１１９７７

研究課題名（和文）進化計算を用いたAdversarial Examplesの生成における複数解探索

研究課題名（英文）Multimodal Optimization in Generating Adversarial Examples Using Evolutionary Algorithm

研究代表者
串田　淳一（KUSHIDA, Jun-ichi）

日本福祉大学・健康科学部・教授

研究者番号：１０５５８５９７

交付決定額（研究期間全体）：（直接経費）　　2,800,000円

研究成果の概要（和文）：Adversarial Examples (AE)は攻撃者によって意図的に機械学習モデルが誤認識を起こすように設計された入力であり，その対策は機械学習モデルを実用化するうえで非常に重要な問題となっている．AEに対する頑健な防御手段を構築するためには，実際に起こりうるような様々な攻撃を徹底的にテストしておく必要がある．そのため，多様なAEパターンを生成するような攻撃方法の開発も重要となる．本課題では進化計算を用いAEを探索する方法に着目し，一度の探索で複数のAEを同時に探索できる攻撃手法を開発した．

研究成果の学術的意義や社会的意義
本課題では，差分進化（DE）を用いた攻撃であるOne pixel attackを基にした攻撃方法を開発した．開発した手法は，AE探索問題における目的関数の多峰性に注目し，目的関数に動的にペナルティを追加することにより，順番に異なるAEを探索することが可能である．この攻撃結果（発見された解の数や種類など）を解析することにより，機械学習モデルのAEに対する防御能力を多角的に評価できる．その結果，より堅牢な機械学習モデルの設計に繋がる重要な知見を得ることが期待される．

研究成果の概要（英文）：Adversarial Examples (AE) are inputs intentionally designed by attackers to cause machine learning models to misidentify them. Developing countermeasures against AEs is an extremely important issue in the real-world applications of machine learning models. To construct robust defenses against AEs, it is necessary to thoroughly test the models against a variety of realistic adversarial attack scenarios. Consequently, the development of attack methods that can generate diverse AE patterns becomes crucial. This research focuses on the use of evolutionary computation to search for AEs and has developed an attack method capable of simultaneously searching for multiple AEs in a single search.

研究分野：進化計算

キーワード：Adversarial Examples　機械学習　進化計算　Differential Evolution　複数解探索

１．研究開始当初の背景

近年，深層学習は，画像認識，自然言語処理，音声認識などの幅広い分野に応用され，目覚ましい発展を遂げている．その一方で，深層学習に基づく画像認識や手書き文字認識に対し，誤った判定結果が出力される入力データを設計する Adversarial examples (AE) に関する研究が数多く報告されている[1]. AE は攻撃者によって機械学習モデルが誤認識を起こすように設計された入力であり，その対策は機械学習モデルを実用化するうえで非常に重要な問題となっている．AE に対する頑健性を向上させるためには，様々な攻撃を想定した防御方法を用意する必要がある．そのためにも，現実的なシナリオの下で多様な AE を生成できる攻撃方法が必須となる．

２．研究の目的

本課題では AE の生成手法（Adversarial attack）として進化計算に着目する．進化計算を用いた AE の生成方法として Differential Evolution (DE) を用いた画像分類モデルに対する攻撃方法（One pixel attack）が提案されている[2]．この手法では画像に対する摂動を DE の解候補として進化させることで，少ない pixel 数の変化で誤分類を起こす AE を生成することができる．

Adversarial attack において攻撃対象が多クラス分類器の場合，その攻撃方針は Targeted attack と Non-targeted attack に分類される．Targeted attack はあらかじめ誤分類させたいターゲットクラスを指定し，そのクラスに属する確率（確信度）が上昇するような AE を生成する．一方，Non-targeted attack では入力データの元のクラスの確信度を減少させることのみを目的とし AE を生成する．そのため，攻撃を行う際にどのクラスに誤分類させるかは指定できない．また，攻撃対象が K クラス分類器であれば，正解クラスを除く $K-1$ 種類の AE が存在することになる．通常の One pixel attack であれば，これらの複数の AE を獲得するためには，ターゲットクラスを変更しながら Targeted attack を繰り返し実行する必要がある．しかしながら，複数試行が必要となるため評価回数（モデルへのアクセス回数）はクラス数に比例して増大する．

そこで本課題では One pixel attack を改良し，一度の DE の実行で複数の AE を獲得する手法を提案する．また，One pixel attack は画像の 1 ピクセルのみを変更することによって機械学習モデルの予測を誤らせる攻撃であるが，高解像度画像に対しては 1 つのピクセルを変更しても，その影響が画像全体に及ぶことは少ない．つまり，モデルがその変更を検出しにくいため One pixel attack の効果は低下する．そのため，高解像度画像を対象とした One pixel attack の改良も目的とした．

３．研究の方法

本研究では，Non-targeted attack における目的関数を，複数の解が存在する多峰性の景観と考えた．このような景観において，1 つの有望解に収束するメカニズムである通常の DE は，集団全体がある 1 つの谷に収束するとそこから抜け出すことは難しい．つまり探索が成功したとしても，1 種類の AE しか獲得することができない．そのため，提案手法では，この多峰性関数において複数解を順番に探索できるように，目的関数にペナルティを動的に追加するという変更を加えた．この手法では，探索中にある AE を発見した場合，以後はそのクラスの確信度にペナルティを加える．つまり，発見済の谷の領域内ではペナルティを受けるため，それ以外の谷へと進化が誘導される．この振る舞いが解空間における未知の領域の探索を促進することになり，図 1 に示すように順番に異なる谷の AE を探索できると考えた．

また，提案手法による複数 AE 探索の性能向上を図るために，DE の改良手法である Rank-based DE (RDE)[4]を導入した．RDE は個体群の適応度によって決定されるランク情報に基づき，親個体ごとに異なる制御パラメータの値を割り当てる手法である．RDE は標準的な DE や改良を加えた DE と比較し，優れた探索性能を持つため，AE 探索問題においても有効に機能すると考えた．

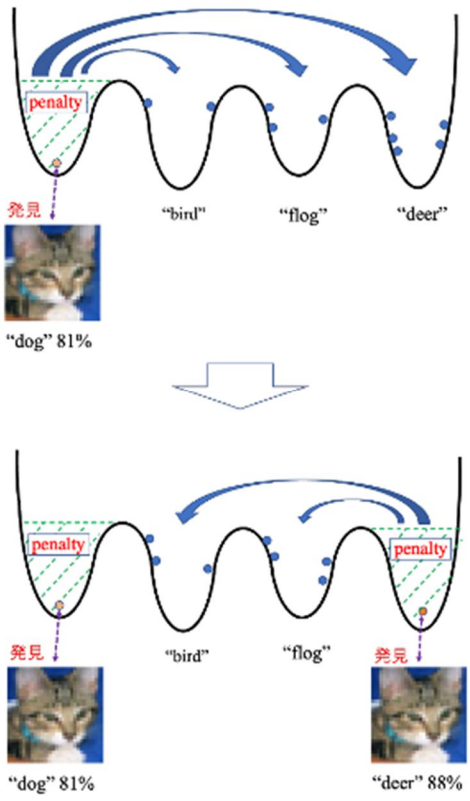

図 1: 提案手法における探索のイメージ

４．研究成果

(1) 複数 AE 探索手法の有効性の検証

提案した複数 AE 探索手法の有効性を確認するため，学習済の画像分類器を対象とした実験を行った．攻撃対象のモデルは画像の 10 クラス分類問題である CIFAR-10 を学習した ResNet, DenseNet, WideResNet とした．表 1 と表 2 は通常の DE による攻撃（ターゲットクラスを変えながら Targeted attack を繰り返す）と提案手法による攻撃の結果を示している．

表 1：複数 AE を発見できた試行数（DE）

	ResNet	DenseNet	WideResnet
Targeted/DE	186	137	145
Proposed/DE	161	121	103

表 2：複数 AE を発見できた試行数（RDE）

	ResNet	DenseNet	WideResnet
Targeted/RDE	205	156	146
Proposed/RDE	188	143	135

これらの結果より，提案手法は通常の Targeted と比べると発見した AE の数は若干少ない．ただし，Targeted は正解クラス以外の 9 回分実行するため，評価回数は約 70,000 回程度となる．一方，提案手法は 一度の実行で 10,000 評価回数であるため, 少ない評価回数で効率的に複数の AE を探索できているといえる．

次に，図 2 に ResNet に対し提案手法で得られた AE を示す．この試行では，dog → deer → horse → cat の順に AE が発見されている．先行研究[4]より ResNet では動物に誤分類する AE が発見されやすいことが分かっており，提案手法においてもそのような AE を順番に発見する試行が多いことが確認できた．

なお提案手法は, 通常の DE 以外のアルゴリズムと組み合わせることができるが, 図 3 に示すように RDE を用いることで DE よりも複数 AE 探索の性能が向上することが確認できた．これは RDE の多様性保持のメカニズムが複数解を連続して探索する際にも有効に機能するためと考えられる．また，提案手法は比較的少ない個体数で実行できるため, この点は多くの個体数を要する niching 手法よりも効率的と考えられる．

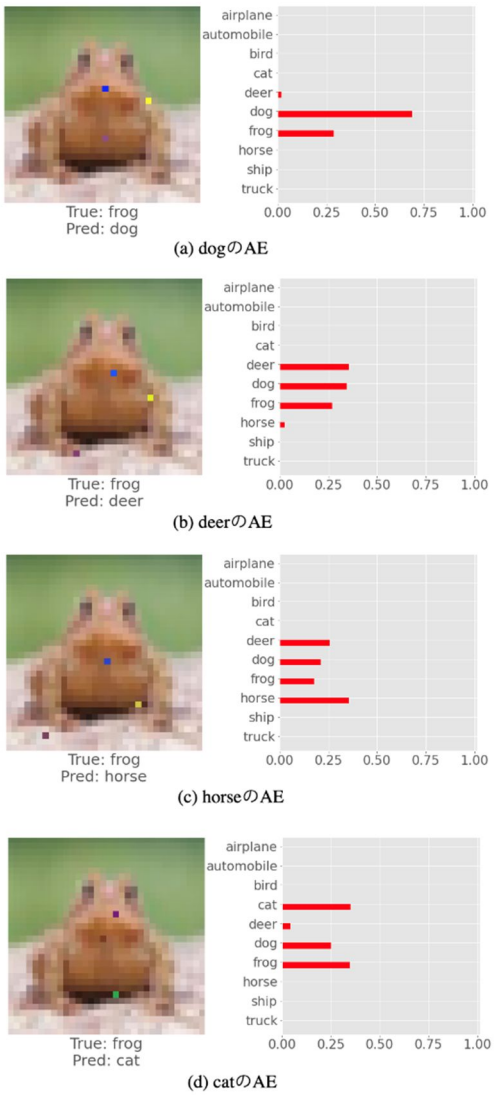

図 2: dog → deer → horse → cat の順に獲得した試行の AE

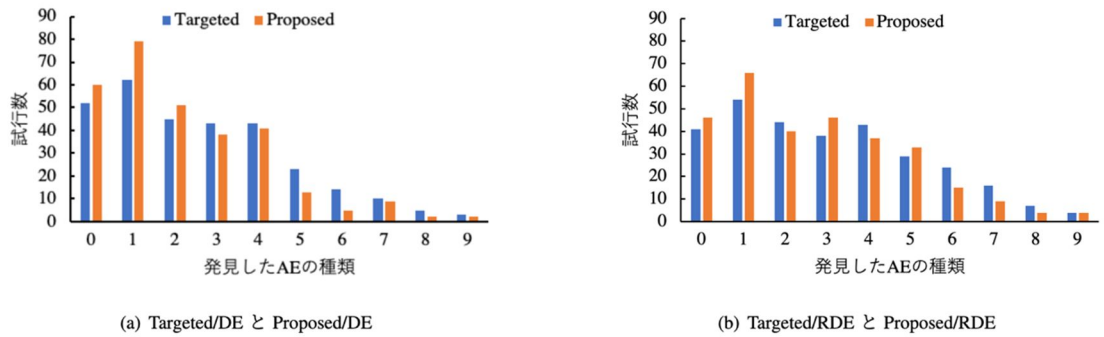

図 3: ResNet に対する攻撃結果

(2) 高解像度画像に対する RDE を用いた AE 探索手法の有効性の検証

(1)で行った実験では，CIFAR-10 による 32×32 ピクセルの画像を対象としたものであり，次の段階の研究として，より高解像度な画像に対しても AE 探索が可能であるかを検証した．ここでもベースとなる DE アルゴリズムに RDE を用い，複数の差分ベクトルを使った multivector mutation[5]を導入した．実験では，ImageNet で学習した VGG16 モデルを対象とし，高解像度の画像(224×224)に対する 20 ピクセルの攻撃（探索空間は 100 次元）を行った．実験結果として，図 4 に各手法の適応度（正解クラスの確信度）の変化を示す．ここでは，通常の DE，RDE，multivector mutation を適用した RDE(MMi と MMd)を比較している．実験結果より，突然変異における差分ベクトルの数をベースベクトルのランクが良いほど大きくする RDE/MMd が最も良い性能を示すことが確認できた．また図 5 に RDE/MMd で生成された AE を示す．この画像は dowitcher という鳥の画像であり，生成された AE は weevil という昆虫に誤分類されている．他にも cab が school_bus として誤分類される結果などがあり，ImageNet においても誤分類が起こりやすいクラスに特徴が見られることが確認できた．また，これらの実験を通し，RDE は複数解探索や複数ピクセルの摂動の探索（高次元の問題）においても，有効に機能することが示された．

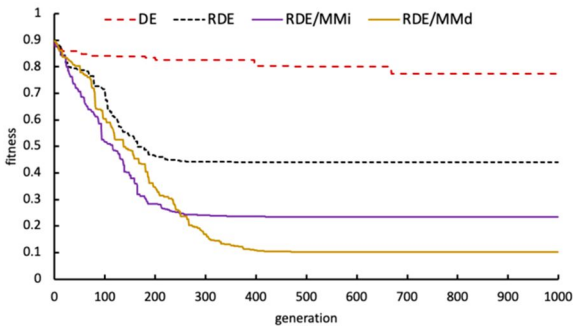

図 4: 各手法の適応度の変化

図 5:RDE/MMd で生成された AE

参考文献

[1] Goodfellow, I., Shlens, J., and Szegedy, C.: Explaining and harnessing adversarial examples, in International Conference on Learning Representations (2015)

[2] Su, J., Vargas, D. V., and Sakurai, K.: One pixel attack for fooling deep neural networks, IEEE Transactions on Evolutionary Computation, Vol. 23, No. 5, pp. 828-841 (2019)

[3] 高濱 徹行，阪井 節子，原 章：RDE: 探索点のランク情報を利用した効率的な Differential Evolution の提案，電子情報通信学会論文誌 D，Vol. 95，No. 5，pp. 1196-1205 (2012)

[4] 串田 淳一，原 章，高濱 徹行：ε 制約 Differential Evolution による摂動量の制約を考慮した Adversarial Examples の生成，進化計算学会論文誌，Vol. 11，No. 3，pp. 55-65 (2020)

[5] J. Kushida, A. Hara, and T. Takahama: Improving the search performance of rank-based differential evolution with multivector mutation, International Journal of Innovative Computing, Information and Control (ICIC), vol. 14, pp. 171-180 (2018)

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Jun-ichi Kushida	４．巻 18(4)
２．論文標題 High-Resolution Adversarial Example Generation Using Rank-Based Differential Evolution	５．発行年 2024年
３．雑誌名 ICIC Express Letters	６．最初と最後の頁 375-384
掲載論文のDOI（デジタルオブジェクト識別子） 10.24507/icicel.18.04.375	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 串田 淳一	４．巻 14
２．論文標題 Differential Evolutionを用いたAdversarial Examplesの生成における複数解探索	５．発行年 2023年
３．雑誌名 進化計算学会論文誌	６．最初と最後の頁 1～11
掲載論文のDOI（デジタルオブジェクト識別子） 10.11394/tjpnsec.14.1	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 －

１．著者名 串田 淳一, 原 章, 高濱 徹行	４．巻 11(3)
２．論文標題 ε 制約Differential Evolutionによる摂動量の制約を考慮したAdversarial Examplesの生成	５．発行年 2020年
３．雑誌名 進化計算学会論文誌	６．最初と最後の頁 55-65
掲載論文のDOI（デジタルオブジェクト識別子） 10.11394/tjpnsec.11.55	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計6件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 串田淳一
２．発表標題 遺伝的アルゴリズムを用いた電気自動車の充電設備の最適化のための基礎検討
３．学会等名 第25回進化計算学会研究会
４．発表年 2024年

１．発表者名
串田淳一

２．発表標題
ε制約遺伝的アルゴリズムを用いたデータの匿名化に関する基礎検討

３．学会等名
第23回進化計算学会研究会

４．発表年
2023年

１．発表者名
串田淳一

２．発表標題
ニッチング手法を用いたDifferential EvolutionによるAdversarial Examplesの生成に関する基礎検討

３．学会等名
第21回進化計算学会研究会

４．発表年
2022年

１．発表者名
串田淳一

２．発表標題
進化計算の実問題への応用進化計算の実問題への応用

３．学会等名
第26回日本知能情報ファジィ学会中国・四国支部大会

４．発表年
2022年

１．発表者名
串田淳一，原章，高濱徹行

２．発表標題
風車最適化問題に対するε制約Differential Evolutionの適用

３．学会等名
第18回進化計算学会研究会

４．発表年
2020年

１．発表者名
串田淳一，原章，高濱徹行

２．発表標題
Differential Evolutionを用いたAdversarial Examplesの生成における摂動のコード化の検討

３．学会等名
進化計算シンポジウム2020

４．発表年
2020年

〔図書〕　計0件

〔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
研究分担者	高濱　徹行 (Tetsuyuki Takahama) (80197194)	広島市立大学・情報科学研究科・教授 (25403)	

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