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 １０ 日現在

機関番号：１２５０１
研究種目：研究活動スタート支援
研究期間：2020～2021
課題番号：２０Ｋ２３３４１

研究課題名（和文）計算代数幾何に基づく深層学習モデルの理論と応用

研究課題名（英文）Theory and Application of deep learning models through the lens of computational algebraic geometry

研究代表者
計良　宥志（Kera, Hiroshi）

千葉大学・大学院工学研究院・助教

研究者番号：００８８７７０５

交付決定額（研究期間全体）：（直接経費）　　2,200,000円

研究成果の概要（和文）：深層学習モデルはその高い表現能力で知られるが，入力に微小な変動をうまく加えることでその振る舞いを大きく変化するため，信頼性の観点で大きな課題を残す．この微小変動が，複数の深層学習モデルに対し悪影響を与える課題に対し，主に二つの成果を得た．一つは深層学習モデルの構造を進化計算で探索し，いくつかモデルに悪影響を与える微小変動に対して頑健な関数を学習できるような構造が存在することを示した．またこの微小変動を利用し，あるドメインでの学習結果を別のドメインの学習に活用できるという新たな応用を物体認識タスクにおいて示した．両研究とも国際論文誌IEEE Accessに採択されている．

研究成果の学術的意義や社会的意義
本研究では，敵対的攻撃から深層学習モデルを守るメカニズムをモデル構造という新たな観点から分析し，また敵対的攻撃をドメイン適応タスクでの精度向上へと繋げる新たな応用を示した．前者は学術的には深層学習で学習される関数の特性に関わり，統計的・幾何的理解が求められており，また産業的にも人工知能システムの信頼性に深く関わる問題である．後者は敵対的攻撃の手法を頑健性向上でなく精度向上へ活用している．従来は頑健性と精度にはある種のトレードオフが存在していたが，ドメイン適応というタスクでこれを回避できた点が興味深い．これらの研究を通して，深層学習の関数特性に関する基礎的・応用的貢献が行えたと考える．

研究成果の概要（英文）：Deep learning models are known for their high expressive power. However, their behavior can be significantly altered by small perturbations in the input, which poses severe concerns in reliability. We have achieved two main results that mitigate and handle such small malicious perturbations. First, we showed the existence of architecture of deep learning models that is robust against malicious perturbations that adversely affect other models. We also showed a new application of the malicious perturbations in the domain adaptation of object recognition. Both papers have been accepted by an international journal, IEEE Access.

研究分野：近似計算機代数

キーワード：深層学習　敵対的攻撃　敵対的転移性　ドメイン適応

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属します。

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

機械学習分野における近年の躍進は，Deep Neural Networks に代表される新たな学習モデル（深層学習モデル）によって支えられている．深層学習モデルはそれ以前のモデルと異なり，線形演算と非線形演算の対（レイヤー・層などと呼ばれる）を何層にも重ねて構成された「深い非線形モデル」となっている．深層学習モデルは，それ以前は難しかった様々な応用（一般物体認識や画像生成など）で非常に高い性能を示しているが，しかしその一方で，深層学習モデルに関する理論は未だ発展途上である．例えば，なぜ深層学習モデルが有効なのか，何をどこまで学習できるのか，この複雑なモデルがなぜうまく最適化できてしまうのか，高い精度を達成するためにはどの程度の質・量のデータがいるのかといった疑問に対しての明確な理論は未だ存在しない．

２．研究の目的

当初の目的は，計算代数幾何を活用し，線形演算と非線形演算の繰り返しによる深いモデルがなぜ，どの様に有効なのかという問いに取り組み，有限次元のモデルの構造とその表現能力・学習能力の関係を定式化することを目指すというものであった．しかし，深層学習モデルの奇妙な振る舞いとして敵対的摂動に対する脆弱性があることが調査でわかり，これが深層学習モデルの特性を強く反映する対象であることを鑑みて調査対象を移した．新たな目的を，深層学習モデルの敵対的摂動に対する頑健性の解明と活用と再設定した．特に次の二点に焦点を絞った．

(1)　深層学習モデルの構造と，敵対的頑健性及び敵対的転移性の関連を明らかにする．
(2) 敵対的摂動に頑健になるよう学習された深層学習モデルの特性を明らかにし，それを画像処理タスクにおける精度向上へと活用する．

(1)の背景．二つの異なる深層学習モデルがあった時，片方を騙すことができる敵対的摂動は，もう一方も騙すことができることが大きい．これは二つのモデルのパラメータや構造が異なる場合でも見られる傾向である．この現象は敵対的転移性と呼ばれるが，一見確率的に学習が進み多様な関数に到達すると見られる深層学習モデル間に共通する性質を反映するものである．さまざまな研究が，二つのモデル間のパラメータの関係からこの敵対的転移性をコントロールすることに取り組んでいたが，モデルの構造から取り組んだ研究はほとんどない．

(2)の背景．学習時に敵対的摂動を伴って学習を行ったモデル（敵対的訓練モデル）は，敵対的摂動に対して頑健になる．しかしその一方で，敵対的摂動を伴わない入力に対する（例えば）分類精度が低下することが知られている．これは，敵対的訓練モデルはより関数をなめらかにし，入力の変動に対し鈍感になるように学習されるためである．つまり，通常の学習をおこなったモデルと，敵対的訓練モデルは入力から抽出する特徴が異なると考えられる．どのように異なるのか，またこの性質を単なる防御以外に活用することができるのかという点は大きな関心を集めている．

３．研究の方法

(1) 進化計算アルゴリズムを用いて，モデルを多様化かつ適応的に進化させる．特に二つのモデル間の入力勾配の直交性（敵対的転移性の抑制につながる）に関するペナルティ項を与えつつ，どのような構造がこの直交性のペナルティ項と画像分類精度の両立を可能にするのかを明らかにする．

(2) タスクをドメイン適応に絞る．ドメイン適応では，教師データのあるデータセット（ソースドメイン）で事前に学習したモデルを，教師データのない別のデータセット（ターゲットドメイン）に適応させることを目的とする．この際，元のソースドメインでの精度は求められない．つまり，ソースドメインにおいて敵対的訓練を行い，通常とは異なる特徴を得つつ，その精度低下を不問にすることができる．そのモデルを新たなドメインに適用した時の精度を調べる．

４．研究成果

(1) 実験の結果，モデルパラメータだけではなく，構造も敵対的転移性の抑制に大きな影響を与えることが明らかになった．具体的には，ある学習済みモデル R（ここでは参照ネットワークと呼ぶ）を準備し，そのモデルと同じ構造をしたモデル R’を，R と入力勾配が直交するようなペナルティをかけつつ学習したものとする．さらに，R と入力勾配が直交するようなペナルティをかけつつ構造も進化せさたもの E を準備した時に，R から R’への敵対的転移性と R から E へのそれを比較した時，10%以上の開きがあった．これにより，モデル構造を適切に設計させると，

より敵対的転移性を抑えられることが明らかになった．既存の異なるモデル間での敵対的転移性は，それらの構造がある程度似通っていたためだと考えられる．図(1)は，ある画像に対してモデルの注目領域を可視化したものである．右上が R，左下が R’，右下が E に対応する．R と R’，R と E がそれぞれ異なるピクセルに注目を置いていることがわかる．

(2)ソースドメインで通常の学習を行なったものより，敵対的訓練を行なったモデルの方がドメイン適応のタスクにおいて多くの場合で優れていることが明らかになった．特に PASCAL VOC データセットから Clipart1k データセットへのドメイン適応では，mean accuracy precision (mAP) が 6 程度向上し，さらに両ドメインでの特徴のアラインメントをとる工夫
を追加することで，その差が 12 程度まで大きく広がることが確認できた．しかしソースドメインとターゲットドメインの組み合わせによって，この差が大きい場合とそうでない場合，あるいは悪化する場合が存在することがわかった．この原因を明らかにするために，ドメイン間の距離を計測した結果，大きくドメイン距離が離れているものほど敵対的訓練を行う効果が大きいことがわかった．これは既に知られている敵対的訓練の特性に合致したもので，学習したドメインでの精度自体は低下しているため，ソースドメインとターゲットドメインの距離が近い場合はこの影響があるからだと考えられる．しかし，ソースドメインとターゲットドメインの多くの場合の距離では，提案手法の有効性が観察された．ソースドメインでの敵対的訓練が効果的な理由は，敵対的訓練が「非ロバストな特徴」を排除し，擾乱の影響を受けづらいロバストな特徴を中心に抽出するからだと考えられる．言い換えれば，ドメインによらない特徴を抽出できているといえ，これがドメイン適応のタスクで効果的な理由であると考えられる．図 2 はロバストな特徴と非ロバストな特徴を可視化したものである．

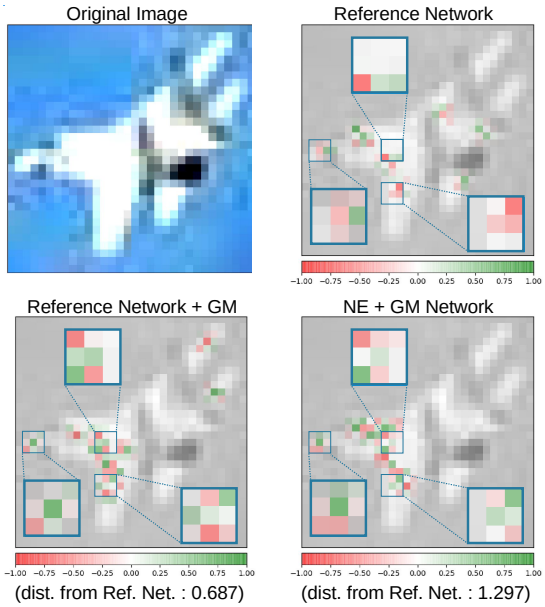

図 1：同一の画像に対するモデルの注目領域を可視化したもの．

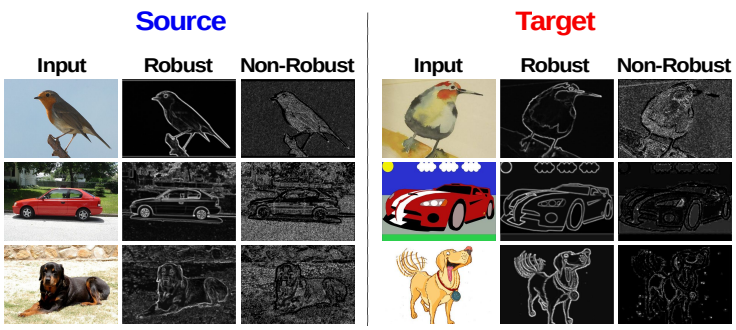

図 2：ロバストな特徴と非ロバストな特徴の例

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　2件／うち国際共著　1件／うちオープンアクセス　2件）

１．著者名 Operiano Kevin Richard G.、Pora Wanchalerm、Iba Hitoshi、Kera Hiroshi	４．巻 9
２．論文標題 Evolving Architectures With Gradient Misalignment Toward Low Adversarial Transferability	５．発行年 2021年
３．雑誌名 IEEE Access	６．最初と最後の頁 164379～164393
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ACCESS.2021.3134840	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 Kazuma Fujii、Hiroshi Kera、Kazuhiko Kawamoto	４．巻 10
２．論文標題 Adversarially Trained Object Detector for Unsupervised Domain Adaptation	５．発行年 2022年
３．雑誌名 IEEE Access	６．最初と最後の頁 5953～59543
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ACCESS.2022.3180344	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 －

〔学会発表〕　計2件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 藤井 一磨, 計良 宥志, 川本一彦
２．発表標題 敵対的訓練を用いたドメイン不変な特徴抽出
３．学会等名 情報処理学会
４．発表年 2022年

１．発表者名 Kevin Richard Operiano, Wanchalerm Pora, 伊庭斉志, 計良宥志
２．発表標題 Reducing Transferability using Neuroevolution with Gradient Misalignment
３．学会等名 進化計算シンポジウム2021
４．発表年 2021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

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			
タイ	チュラロンコーン大学			