(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7521010号
(P7521010)

(45)発行日　令和6年7月23日(2024.7.23)

(24)登録日　令和6年7月12日(2024.7.12)

(51)国際特許分類		FI	
G06N	*3/045(2023.01)*	G06N	3/045
G06N	*3/0455(2023.01)*	G06N	3/0455
G06N	*3/0464(2023.01)*	G06N	3/0464
G06N	*3/08 (2023.01)*	G06N	3/08
G06F	*21/55 (2013.01)*	G06F	21/55

請求項の数　16　（全13頁）

(21)出願番号　特願2022-570235(P2022-570235)
(86)(22)出願日　令和3年5月14日(2021.5.14)
(65)公表番号　特表2023-526809(P2023-526809A)
(43)公表日　令和5年6月23日(2023.6.23)
(86)国際出願番号　PCT/FR2021/050842
(87)国際公開番号　WO2021/234252
(87)国際公開日　令和3年11月25日(2021.11.25)
審査請求日　令和6年5月14日(2024.5.14)
(31)優先権主張番号　2004945
(32)優先日　令和2年5月18日(2020.5.18)
(33)優先権主張国・地域又は機関
フランス(FR)
早期審査対象出願

(73)特許権者　519110386
アイデミア・アイデンティティ・アンド・セキュリティー・フランス
フランス国、９２４００・クルブボア、プラス・サミュエル・ドゥ・シャンプラン・２
(74)代理人　110002077
園田・小林弁理士法人
(72)発明者　シャバンヌ， エアベ
フランス国 ９２４００ クルブボア，プラス サミュエル ドゥ シャンプラン ２， シー／オー アイデミア アイデンティティ アンド セキュリティー フランス
(72)発明者　グイガ， リンダ

最終頁に続く

(54)【発明の名称】　入力データに対して第１のニューラルネットワークを安全に使用するための方法及び第２のニューラルネットワークのパラメータを学習するための方法

(57)【特許請求の範囲】

【請求項１】

入力データに対して第１のニューラルネットワークを安全に使用するための方法において、以下のステップ：

（ａ）前記第１のニューラルネットワークの層の中から前記第１のニューラルネットワークの標的層を選択し、且つ恒等関数を近似する少なくとも１つの畳み込みニューラルネットワークが前記第１のニューラルネットワークの前記標的層の入力において挿入される前記第１のニューラルネットワークに対応する第２のニューラルネットワークを構築するステップ、

（ｂ）前記入力データに対して前記第２のニューラルネットワークを使用するステップを端末（２）のデータ処理手段（２１）によって実装することを含むことを特徴とする方法。

【請求項２】

前記畳み込みニューラルネットワークは、前記標的層の入力サイズよりも小さい出力サイズを有して、前記標的層の特定の入力チャネルのみを近似する、請求項１に記載の方法。

【請求項３】

ステップ（ａ）は、前記標的層の入力チャネルの全ての中から、近似される前記標的層の入力チャネルを選択するステップを含む、請求項１又は２に記載の方法。

【請求項４】

前記恒等関数を近似する前記少なくとも１つの畳み込みニューラルネットワークは、２

つの整数の積に等しい出力サイズを有する、請求項１～３の何れか一項に記載の方法。
【請求項５】
前記第１のニューラルネットワーク及び前記恒等関数を近似する前記少なくとも１つの畳み込みニューラルネットワークのパラメータを得る予備ステップ（ａ０）を含む、請求項１～４の何れか一項に記載の方法。
【請求項６】
ステップ（ａ０）は、前記恒等関数を近似する畳み込みニューラルネットワークの組のパラメータを得るステップを含み、ステップ（ａ）は、挿入される前記恒等関数を近似する少なくとも１つの畳み込みニューラルネットワークを前記組から選択するステップを含む、請求項５に記載の方法。
【請求項７】
ステップ（ａ）は、前記恒等関数を近似する選択された畳み込みニューラルネットワークごとに、前記第１のニューラルネットワークの線形層の中から前記第１のニューラルネットワークの前記標的層を選択するステップ及び／又は前記標的層の入力チャネルの全ての中から、近似される前記標的層の前記入力チャネルを選択するステップを含む、請求項６に記載の方法。
【請求項８】
ステップ（ａ）は、選択される前記組の前記恒等関数を近似する畳み込みニューラルネットワークの数を事前に選択するステップを更に含む、請求項６又は７に記載の方法。
【請求項９】
ステップ（ａ０）は、前記第１のニューラルネットワーク及び前記恒等関数を近似する前記少なくとも１つの畳み込みニューラルネットワークの前記パラメータを少なくとも１つの学習データベースから学習する、サーバ（１）のデータ処理手段（１１）によって実装されるステップである、請求項５～８の何れか一項に記載の方法。
【請求項１０】
前記第１のニューラルネットワーク及び前記恒等関数を近似する前記少なくとも１つの畳み込みニューラルネットワークは、ＲｅＬＵ関数等の活性化関数を有する交互の線形層及び非線形層を含む、請求項１～９の何れか一項に記載の方法。
【請求項１１】
前記標的層は、線形層である、請求項１０に記載の方法。
【請求項１２】
前記恒等関数を近似する前記少なくとも１つの畳み込みニューラルネットワークは、フィルタ畳み込み層である２つ又は３つの線形層を含む、請求項１０又は１１に記載の方法。
【請求項１３】
第２のニューラルネットワークのパラメータを学習するための方法において、以下のステップ：
（ａ）第１のニューラルネットワークの層の中から前記第１のニューラルネットワークの標的層を選択し、且つ恒等関数を近似する少なくとも１つの畳み込みニューラルネットワークが前記第１のニューラルネットワークの前記標的層の入力において挿入される前記第１のニューラルネットワークに対応する前記第２のニューラルネットワークを構築するステップ、
（ａ１）前記第２のニューラルネットワークの前記パラメータを学習データベースから学習するステップ
をサーバ（１）のデータ処理手段（１１）によって実装することを含むことを特徴とする方法。
【請求項１４】
入力データに対して第１のニューラルネットワークを安全に使用するための方法であって、請求項１３に記載の方法に従って第２のニューラルネットワークのパラメータを学習するステップと、前記入力データに対して前記第２のニューラルネットワークを使用するステップ（ｂ）を端末（２）のデータ処理手段（２１）によって実装するステップとを含

む方法。
【請求項１５】
コンピュータプログラムであって、前記プログラムがコンピュータによって実行されるとき、第２のニューラルネットワークのパラメータを学習するために、又は入力データに対して第１のニューラルネットワークを安全に使用するために、請求項１～１４の何れか一項に記載の方法を実行するためのコード命令を含むコンピュータプログラム。
【請求項１６】
コンピュータ機器によって読み出され得る記憶媒体であって、第２のニューラルネットワークのパラメータを学習するために、又は入力データに対して第１のニューラルネットワークを安全に使用するために、請求項１～１４の何れか一項に記載の方法を実行するためのコード命令を含むコンピュータプログラムを記憶した、記憶媒体。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、人工知能の分野に関し、具体的には入力データに対して第１のニューラルネットワークを安全に使用するための方法に関する。
【背景技術】
【０００２】
ニューラルネットワーク（又はＮＮ）は、データを分類するために広く使用されている。
【０００３】
（概して教師付きである、即ち既に分類されているデータの参照データベースに基づく）機械学習の段階後、ニューラルネットワークは、「学習」し、未知のデータに対して同じ分類を自然に適用できるようになる。より正確には、ＮＮの重み及びパラメータの値は、目標とするタスクを実装可能になるまで漸進的に修正される。
【０００４】
ニューラルネットワークのアーキテクチャ及び学習技法（具体的にはディープラーニング）の両方について、更に学習ベース（そのサイズ及び品質）について著しい進歩が近年見られ、過去に不可能であると考えられていたタスクは、今日では、卓越した信頼性と共にニューラルネットワークによって実行されている。
【０００５】
これらの全ては、高性能のニューラルネットワーク及びその学習ベースが、今日、高い商業的価値を有し、保護すべき「企業秘密」として扱われていることを意味する。加えて、多くのデータベースは、秘密に保たれなければならない潜在的に個人的なデータ（例えば、指紋）を含む。
【０００６】
不都合なことに、Ｃｒｙｐｔａｎａｌｙｔｉｃ Ｅｘｔｒａｃｔｉｏｎ ｏｆ Ｎｅｕｒａｌ Ｎｅｔｗｏｒｋ Ｍｏｄｅｌｓ，Ｎｉｃｈｏｌａｓ Ｃａｒｌｉｎｉ，Ｍａｔｔｈｅｗ Ｊａｇｉｅｌｓｋｉ，Ｉｌｙａ Ｍｉｒｏｎｏｖ ｈｔｔｐｓ：／／ａｒｘｉｖ．ｏｒｇ／ｐｄｆ／２００３．０４８８４ｖ１．ｐｄｆの文献に記載されているように、任意のニューラルネットワークに十分な精選された要求を提出できることを条件として、任意のニューラルネットワークのパラメータ及びモデルを攻撃者が抽出することを可能にする「リバースエンジニアリング」技法が最近編み出されている。従って、（例えば、ウェブクライアントによる）入力及び出力に対してのみアクセスがある「ブラックボックス」動作においてさえ、ネットワークの内部が回復され得る。
【０００７】
この概念は、ニューラルネットワークが、ＲｅＬＵ等の活性化関数を実装する交互の線形層及び非線形層を含むことを観察するものである。この非線形性は、勾配の跳ね上がりを伴う「臨界点」をもたらし、従って出力が臨界点にあるように、ネットワークの入力空間の超平面をニューロンごとに幾何学的に定めることができる。第２の層の超平面は、第１の層の超平面によって「フォールド」等される。

【０００８】
攻撃者は、超平面の交点及び漸進的に全ニューラルネットワークを回復するために探索を使用し得る。
【０００９】
ニューラルネットワークが直面する更なる課題は、「敵対的外乱」、即ちニューラルネットワークの入力に適用されるときに出力を著しく変更する、感知できない変更の存在である。Ａｄｉ Ｓｈａｍｉｒ、Ｉｔａｙ Ｓａｆｒａｎ、Ｅｙａｌ Ｒｏｎｅｎ及びＯｒｒ Ｄｕｎｋｅｌｍａｎによる文献Ａ Ｓｉｍｐｌｅ Ｅｘｐｌａｎａｔｉｏｎ ｆｏｒ ｔｈｅ Ｅｘｉｓｔｅｎｃｅ ｏｆ Ａｄｖｅｒｓａｒｉａｌ Ｅｘａｍｐｌｅｓ ｗｉｔｈ Ｓｍａｌｌ Ｈａｍｍｉｎｇ Ｄｉｓｔａｎｃｅ，ｈｔｔｐｓ：／／ａｒｘｉｖ．ｏｒｇ／ｐｄｆ／１９０１．１０８６１ｖ１．ｐｄｆは、例えば、猫の画像に適用された敵対的外乱が、どのようにグアカモーレの画像としてそれが誤分類されることを引き起こし得るかを開示している。
【００１０】
より正確には、攻撃者が、上記で説明した超平面への分割を識別することに成功すると、その攻撃者は、入力空間内の点から超平面を交差可能にする、従って出力を修正できるようにするベクトルを決定することができる。
【００１１】
従って、ニューラルネットワークを安全にすることに成功することは、必須であることが理解されるであろう。
【００１２】
第１の手法は、攻撃者のタスクを複雑化するために、ネットワークのサイズ、層の数及びパラメータの数を増やすものである。この手法が機能する場合、この手法が行うのは、一方では攻撃者の速度を落とすことのみであり、ニューラルネットワークが不要に煩雑になり、訓練しにくくなるため、この手法は、何よりも性能を悪化させる。
【００１３】
第２の手法は、ニューラルネットワークに提出可能な入力の数を制限するか、又は少なくとも疑わしい入力シーケンスを検出するものである。しかし、例えば、無制限アクセスの代金を支払っている攻撃者は、ニューラルネットワークに合法的にアクセスし得るため、この手法は、常に適用可能であるわけではない。
【発明の概要】
【発明が解決しようとする課題】
【００１４】
従って、この状況は、更に改善され得る。
【課題を解決するための手段】
【００１５】
第１の態様によれば、本発明は、入力データに対して第１のニューラルネットワークを安全に使用するための方法において、以下のステップ：
（ａ）恒等関数を近似する少なくとも１つの畳み込みニューラルネットワークが挿入される第１のニューラルネットワークに対応する第２のニューラルネットワークを構築するステップ、
（ｂ）前記入力データに対して第２のニューラルネットワークを使用するステップ
を端末のデータ処理手段によって実装することを含むことを特徴とする方法に関する。
【００１６】
他の有利な且つ非限定的な特徴によれば、前記畳み込みニューラルネットワークは、第１のニューラルネットワークの標的層の入力において挿入される。
【００１７】
前記標的層は、第１のニューラルネットワーク内の層である。
【００１８】
前記畳み込みニューラルネットワークは、前記標的層の入力サイズよりも小さい出力サ

イズを有して、この標的層の特定の入力チャネルのみを近似する。
【００１９】
ステップ（ａ）は、前記第１のニューラルネットワークの層の中から第１のニューラルネットワークの前記標的層を選択するステップを含む。
【００２０】
ステップ（ａ）は、標的層の入力チャネルの全ての中から、近似される前記標的層の入力チャネルを選択するステップを含む。
【００２１】
恒等関数を近似する少なくとも１つの畳み込みニューラルネットワークは、２つの整数の積に等しい出力サイズを有する。
【００２２】
この方法は、第１のニューラルネットワーク及び恒等関数を近似する少なくとも１つの畳み込みニューラルネットワークのパラメータを得る予備ステップ（ａ０）を含む。
【００２３】
ステップ（ａ０）は、恒等関数を近似する畳み込みニューラルネットワークの組のパラメータを得るステップを含む。
【００２４】
ステップ（ａ）は、挿入される恒等関数を近似する少なくとも１つの畳み込みニューラルネットワークを前記組から選択するステップを含む。
【００２５】
ステップ（ａ）は、恒等関数を近似する選択された畳み込みニューラルネットワークごとに、前記第１のニューラルネットワークの層の中から第１のニューラルネットワークの前記標的層を選択する前記ステップ及び／又は標的層の入力チャネルの全ての中から、近似される前記標的層の入力チャネルを選択するステップを含む。
【００２６】
ステップ（ａ）は、選択される前記組の恒等関数を近似する畳み込みニューラルネットワークの数を事前に選択するステップを更に含む。
【００２７】
ステップ（ａ０）は、第１のニューラルネットワーク及び恒等関数を近似する少なくとも１つの畳み込みニューラルネットワークのパラメータを少なくとも１つの学習データベースから学習する、サーバのデータ処理手段によって実装されるステップである。
【００２８】
第１のニューラルネットワーク及び恒等関数を近似する１つ又は複数の畳み込みニューラルネットワークは、活性化関数を有する交互の線形層及び非線形層を含む。
【００２９】
前記活性化関数は、ＲｅＬＵ関数である。
【００３０】
前記標的層は、第１のニューラルネットワークの線形層である。
【００３１】
恒等関数を近似する少なくとも１つの畳み込みニューラルネットワークは、２つ又は３つの線形層を含む。
【００３２】
畳み込みニューラルネットワークの線形層は、例えば、サイズ５×５のフィルタ畳み込み層である。
【００３３】
第２の態様によれば、提案されるものは、第２のニューラルネットワークのパラメータを学習するための方法において、以下のステップ：
（ａ）恒等関数を近似する少なくとも１つの畳み込みニューラルネットワークが挿入される第１のニューラルネットワークに対応する第２のニューラルネットワークを構築するステップ、

（ａ１）第２のニューラルネットワークのパラメータを学習データベースから学習するステップ
をサーバのデータ処理手段によって実装することを含むことを特徴とする方法である。
【００３４】
第３の態様によれば、提案されるものは、入力データに対して第１のニューラルネットワークを安全に使用するための方法であって、第２の態様による方法に従って第２のニューラルネットワークのパラメータを学習するステップと、前記入力データに対して第２のニューラルネットワークを使用するステップ（ｂ）を端末のデータ処理手段によって実装するステップとを含む方法である。
【００３５】
第４の態様及び第５の態様によれば、本発明は、コンピュータプログラム製品であって、入力データに対して第１のニューラルネットワークを安全に使用するために、第１の態様若しくは第３の態様による方法を、又は第２のニューラルネットワークのパラメータを学習するために、第２の態様による方法を実行するためのコード命令を含むコンピュータプログラム製品と、コンピュータ機器によって読み出され得る記憶手段であって、コンピュータプログラム製品は、入力データに対して第１のニューラルネットワークを安全に使用するために、第１の態様若しくは第３の態様による方法を、又は第２のニューラルネットワークのパラメータを学習するために、第２の態様による方法を実行するためのコード命令を含む、記憶手段とに関する。
【００３６】
１つの好ましい実施形態についての以下の説明を読むことで、本発明の他の特徴及び利点が明らかになる。この説明は、添付図面に関して与えられる。
【図面の簡単な説明】
【００３７】
【図１】本発明による方法を実装するためのアーキテクチャの図である。
【図２ａ】本発明による、入力データに対して第１のニューラルネットワークを安全に使用するための方法の第１の実施形態のステップを概略的に示す。
【図２ｂ】本発明による、入力データに対して第１のニューラルネットワークを安全に使用するための方法の第２の実施形態のステップを概略的に示す。
【図３】本発明による方法の実装において遭遇する第２のニューラルネットワークのアーキテクチャの一例を概略的に示す。
【発明を実施するための形態】
【００３８】
アーキテクチャ
本発明の２つの補完的態様によれば、提案されるものは、
－　第１のニューラルネットワーク（第１のＮＮ）を安全に使用するための方法、
－　第２のニューラルネットワーク（第２のＮＮ）のパラメータを学習するための方法である。
【００３９】
これらの２種類の方法は、少なくとも１つのサーバ１及び１つの端末２により、［図１］によって示すアーキテクチャ内で実装される。サーバ１は、（第２の方法を実装する）学習機器であり、端末２は、（第１の方法を実装する）ユーザ機器である。前記使用方法は、入力データに対して実装され、例えば分類ＮＮである場合、複数のクラスの中からの入力データの分類である（しかし、分類が最も慣例的であるが、このタスクは、必ずしも分類ではない）。
【００４０】
任意の種類のＮＮが、典型的には、σ（ｘ）＝ｍａｘ（０，ｘ）に等しいＲｅＬＵ（正規化線形ユニット）活性化関数を有する交互の線形層及び非線形層を含むとしても、任意の種類のＮＮに特に制限を課さない。従って、線形層の出力がゼロに等しいように、各超平面が入力空間の点集合に対応することが理解されるであろう。かかるニューラルネット

ワークを示すために「ＲｅＬＵ　ＮＮ」を使用する。
【００４１】
何れにせよ、各機器１、２は、典型的には、データをやり取りするためにインターネット等の広域ネットワーク１０に接続されるリモートコンピュータ機器である。それぞれがプロセッサ型のデータ処理手段１１、２１及びコンピュータメモリ、例えばハードディスク等のデータ記憶手段１２、２２を含む。
【００４２】
サーバ１は、例えば、既に分類されているデータの学習データベース、即ち（処理することが明確に望まれる入力データとして知られるものとは対照的に）関連する出力が既に知られているデータの組を記憶する。これは、秘密を保つことが求められる高い商業的価値を有する学習ベースであり得る。
【００４３】
機器１及び２が同じ機器であることが依然として可能であるか、又は学習ベースが更にパブリックベースであり得ることが理解されるであろう。
【００４４】
本方法は、１種類のＮＮ、従って１つの特定の種類のデータに限定されず、入力データ又は学習データは、画像、音声等を表し得ることに留意すべきである。第１のＮＮは、ＣＮＮであることが大いに考えられ得るが、本方法に関連して使用される専用ＣＮＮについて以下に記載する。
【００４５】
１つの好ましい実施形態では、バイオメトリクデータが関与し、入力データ又は学習データは、典型的には、画像又は更に直接的にバイオメトリク特徴（顔、指紋、虹彩等）の画像若しくは直接的にバイオメトリク特徴から生じる前処理されたデータ（例えば、指紋の場合には特徴点の位置）を表す。
【００４６】
原理
本発明は、人工超平面を使用して、ＮＮを複雑にすることなく、攻撃者のタスクを複雑化することを提案する。換言すれば、ＮＮをさもなければより煩雑にし、その性能を悪化させることなく、著しくよりロバストにすることにより、ＮＮが安全にされる。
【００４７】
便宜上、保護される元のＮＮを「第１のニューラルネットワーク」と呼び、修正され、従って安全にされるＮＮを「第２のニューラルネットワーク」と呼ぶ。以下で認められるように、第１のＮＮは、事後的に（学習されると）安全にすることができるか、又は最初から安全にされ得る（即ちＮＮの安全なバージョンが直接学習される）。
【００４８】
より詳細には、第１のネットワークを第２のネットワークとして安全にすることは、恒等関数を近似する少なくとも１つの畳み込みニューラルネットワーク（ＣＮＮ）（これを便宜上「恒等ＣＮＮ」と呼ぶ）をそのアーキテクチャに組み込むことで構成される。
【００４９】
この「寄生的」なＣＮＮは、その出力がその入力とほぼ等しいため、ＮＮの動作を修正しない。他方では、かかるＣＮＮは、元の超平面構造を壊す。
【００５０】
恒等関数を近似する概念は、実現するのが困難な不自然なタスクであるため、ＣＮＮにとって極めて斬新である。別の言い方をすれば、ＣＮＮが意味的に複雑な処理演算（例えば、画像分割等）を行い、自らの入力を再現するほど自明なタスクを決して行わないことが常に望ましい。
【００５１】
加えて及び以下で認められるように、必要に応じて全て動的に且つ無作為に選択される特定のチャネルを含む様々な位置において、複数の恒等ＣＮＮを第１のＮＮ内に挿入し、それにより攻撃者に機会を与えないことができる（人工超平面の下では、元の第１のＮＮ

を回復するために第２のＮＮに送信される想像を絶する数の要求が必要になる）。
【００５２】
方法
第１の態様によれば、［図２ａ］に関して提案されるものは、端末２のデータ処理手段２１によって実装される、入力データに対して第１のＮＮを安全に使用するための方法の第１の実施形態である。
【００５３】
この方法は、第１のＮＮ及び少なくとも１つの恒等ＣＮＮのパラメータ、可能な場合には多様な恒等ＣＮＮの組を定めるように、具体的には様々なアーキテクチャの、様々な入力及び出力サイズの、様々なベースに基づいて訓練等される複数の恒等ＣＮＮのパラメータを得る「準備」ステップ（ａ０）から始まり、このステップは、後により詳細に認められる。
【００５４】
このステップ（ａ０）は、専用学習ベースに基づいてネットワークのそれぞれ、具体的には第１のＮＮを訓練するステップであり得、好ましくはサーバ１のデータ処理手段１１によってこの目的で実装されるが、ネットワーク（具体的には恒等ＣＮＮ）は、既存でそのまま取られ得ることが理解されるであろう。何れにせよ、１つ又は複数の恒等ＣＮＮは、具体的には、任意のパブリック画像ベースに基づいて又は更に乱データに基づいて訓練され得る（場合により、雑音を別として、入力は、予期される出力でもあると想定されるため、それらのデータが注釈付けされる必要はない。以下を参照されたい）。そのようなステップ（ａ０）がない１つの代替的実施形態を以下で認めることができる。
【００５５】
主たるステップ（ａ）は、恒等関数を近似する少なくとも１つの畳み込みニューラルネットワーク、具体的には１つ又は複数の選択された恒等ＣＮＮが挿入される第１のＮＮに対応する前記第２のＮＮを構築することを含む。換言すれば、ステップ（ａ）は、第１のＮＮ内に１つ又は複数の恒等ＣＮＮを挿入するステップである。複数の選択された恒等ＣＮＮがある場合、それらは、順に挿入され得る。
【００５６】
そのために、ステップ（ａ）は、前記恒等ＣＮＮの組の中から１つ又は複数の恒等ＣＮＮを例えば無作為に選択することを有利に含む。他の「挿入パラメータ」、具体的には第１のＮＮ内の位置（標的層）及び／又は第１のＮＮの標的層のチャネルが選択され得、以下を参照されたい。何れにせよ、恒等ＣＮＮの組が１つのＣＮＮのみを含むことが依然として可能であり、そのため、選択する必要がなく、更に恒等ＣＮＮをオンザフライで訓練する必要がない。
【００５７】
挿入は、第１のＮＮの「標的」層の上流に恒等ＣＮＮの層を追加することを意味することが理解され、そのため、この層の入力は、少なくとも部分的には恒等ＣＮＮの出力である。換言すれば、恒等ＣＮＮは、標的層の入力の全て又は一部を自らの出力と置換するために「インタセプト」する。
【００５８】
恒等ＣＮＮは、恒等関数を近似するため、その出力は、その入力とほぼ同一であり、そのため、標的層によって受信されるデータは、インタセプトされるデータとほぼ同一であることが理解されるであろう。
【００５９】
標的層は、好ましくは、（例えば、活性化関数を有する非線形層ではなく）線形層であり、そのため、恒等ＣＮＮは、第１のＮＮの線形層の入力において挿入される。
【００６０】
標的層は、好ましくは、第１のＮＮ内の層、即ち１番目以外の（２番目の層と最後の層との間の）層である。特に好ましくは、標的層は、従って、第１のＮＮ内の線形層である。
【００６１】

有利には、恒等ＣＮＮは、前記線形層の入力サイズよりも小さい出力サイズを有して、この線形層の特定の入力チャネルのみ（即ち入力チャネルの全てではない）を近似する。入力／出力サイズは、入力／出力チャネルの数を意味することが理解されるであろう。

【００６２】

これは、［図３］の例で認められる内容であり、［図３］は、（中央の隠れ層を含む）３つの線形層を有する第１のＮＮを示し、この中で、恒等ＣＮＮは、第２の層の入力において配置される。

【００６３】

第１の層は、８つの入力チャネル（サイズ８）を有するのに対して、恒等ＣＮＮは、４つの入力／出力チャネルのみを有することが見て取れる（定義上、恒等関数を近似するＣＮＮは、同じ入力及び出力の次元を有する）。従って、第１の線形層の８つの入力チャネルのうちの４つのみが近似され、残りの４つは、そのままである。チャネルの一部にのみ（即ちニューロンの全てではなく）作用することには、以下の３つの利点があり、その利点とは、即ち、ＣＮＮがより小さくなり得、従って実行時により少ない計算を伴うこと、層に対して部分的にのみ作用することが攻撃者にとって驚くべき混乱を発生させること及び更に層の下に複数のＣＮＮを配置し、従って攻撃者にとっての超平面の混乱を更に高めることができることである。

【００６４】

ステップ（ａ）は、説明したように、（該当する場合には事前に選択される）恒等ＣＮＮによって作用される標的層及び／又は標的層の入力チャネルを選択することも含み得る。例えば、図３では、これらは、チャネル１、２、３及び４であるが、８つのチャネルのうちの任意の４チャネル、例えばチャネル１、３、５及び７の組を取ることもできる。

【００６５】

この選択は、ここでも、無作為に且つ動的に行うことができ、即ち、新たなチャネルは、第１のＮＮを使用するための新たな要求ごとに、しかし、同様に順に又はさもなければコンテキストデータ、具体的には入力データに基づいて得られる。潜在的な攻撃者にとってのタスクを一層複雑化するためにエントロピーを追加するための複数の可能性の中の有効な選択肢がある限り、本発明は、とりわけ標的層／近似されるチャネル／１つ又は複数の恒等ＣＮＮを選択するいかなる方法にも限定されない。

【００６６】

実際には、選択は、以下のプロトコルに従って行うことができる（各ステップは、任意選択的であり、各選択は、無作為又は既定であり得る）：

１．挿入される恒等ＣＮＮの数を選択すること、

２．この数と同数の恒等ＣＮＮを（戻すことを伴って又は伴わずに）恒等ＣＮＮの組から得ること、

３．得た恒等ＣＮＮごとに、作用される（即ちその入力においてＣＮＮが挿入される）標的層を第１のＮＮの層（具体的には線形及び／又は中間層）の中から選択すること、

４．得られた恒等ＣＮＮごとに、この恒等ＣＮＮの入力／出力チャネルの数と同数の関連標的層の入力チャネルを選択すること。

【００６７】

点３に関して、同じ標的層に作用するように２つの恒等ＣＮＮが選択され得ることに留意すべきであり、即ち対象のチャネルが別々であり、問題がないか、又は少なくとも１つのチャネルが重複し、その場合、かかる重複を不所望であると判断する（得ることをやり直す）か、又は一方の恒等ＣＮＮが他方の上流にあることが受け入れられると判断することができ、従って標的層の入力に到達する前にチャネルが２回連続して近似され得る。

【００６８】

点４に関して、恒等ＣＮＮは、典型的には、画像、即ち２次元オブジェクト（「矩形」）を扱うネットワークであり、従って２つの整数の積、即ちａ＊ｂの形式に等しい入力／出力チャネルの数を有し、ａ及びｂは、それぞれ２以上の整数であり、更に優先的には次元の「２乗」a^2であることに留意すべきである。３次元オブジェクトを扱う、従って３

つの整数の積、即ちａ＊ｂ＊ｃの形式等に等しい入力／出力チャネルの数を有するＣＮＮを使用することを想像することは、完全に可能である。図３の例では、２×２の画像を扱う、従って４つの入力／出力チャネルを有する恒等ＣＮＮがある。

【００６９】

最後に、選択及び構築のアクションは、部分的にネスト（従って同時に実装する）ことができ、即ち挿入される複数の恒等ＣＮＮがある場合、最初のものの挿入パラメータを決定し、それを挿入し、２番目のものの挿入パラメータを決定し、それを挿入すること等ができる。加えて、説明したように、標的層及び／又はチャネルは、ステップ（ａ）においてオンザフライで選択することができる。

【００７０】

ステップ（ａ）の終了時に第２のＮＮが構築されていると想定される。次いで、ステップ（ｂ）において、前記入力データに対してこの第２のＮＮを使用することができ、即ち第２のＮＮが入力データに適用され、これにより第１のＮＮに戻ることが可能であるいかなるリスクもなしに、端末２のユーザに与えられ得る出力データが得られる。

【００７１】

恒等ＣＮＮ

ＲｅＬＵ等の活性化関数を有する交互の畳み込み層（線形層）及び非線形層のみで構成される小さいＣＮＮは、恒等の近似の品質及び第１のＮＮを別様により煩雑にすることなく、超平面を複雑化することの両方の点で非常に優れた結果を与える。

【００７２】

例えば、恒等ＣＮＮは、（いかなるアーキテクチャに制限を課さなくても）２つ又は３つの畳み込み層のみを含む場合がある。

【００７３】

１つの特に好ましい実施形態によれば、サイズ５×５の２つ又は３つのフィルタ畳み込み層を有する１６×１６までのサイズの平方入力／出力恒等ＣＮＮを取ることができる。

【００７４】

各恒等ＣＮＮは、場合により、少量の雑音を別として、恒等を最良に近似することを目指すことに留意すべきである。そのために、予期される出力として入力を「そのまま」使用する代替策として、入力に雑音、好ましくは平均０のガウス雑音が加わったものを学習中に使用することができる。本出願人は、これが特に良好に機能すること（Ａ Ｐｒｏｔｅｃｔｉｏｎ ａｇａｉｎｓｔ ｔｈｅ Ｅｘｔｒａｃｔｉｏｎ ｏｆ Ｎｅｕｒａｌ Ｎｅｔｗｏｒｋ Ｍｏｄｅｌｓ，Ｈｅｒｖｅ Ｃｈａｂａｎｎｅ，Ｖｉｎｃｅｎｔ Ｄｅｓｐｉｅｇｅｌ，Ｌｉｎｄａ Ｇｕｉｇａ，ｈｔｔｐｓ：／／ａｒｘｉｖ．ｏｒｇ／ｐｄｆ／２００５．１２７８２．ｐｄｆの文献を参照されたい）及び様々な雑音を追加することにより、恒等ＣＮＮの組に更に多くの変動性を追加することも加えて可能であることを見出した。

【００７５】

当然のことながら、恒等ＣＮＮの組の全て又は一部においていかなる雑音も追加しないこと、即ち恒等関数を最良に近似することが可能である。

【００７６】

試験

全結合ネットワーク（ＦＣＮ）タイプの３つの隠れ層を有するＲｅＬＵ ＮＮを第１のＮＮとして取って試験を行い、隠れ層は、５１２個、５１２個及び３２個の入力チャネルをそれぞれ有し、このＦＣＮは、手書き数字を認識するため（任意のサイズの入力画像の分類）に使用されている。この第１のＮＮは、ＭＮＩＳＴ（Ｍｉｘｅｄ Ｎａｔｉｏｎａｌ Ｉｎｓｔｉｔｕｔｅ ｏｆ Ｓｔａｎｄａｒｄｓ ａｎｄ Ｔｅｃｈｎｏｌｏｇｙ）学習ベースに基づいてこのタスクのために訓練することができ、９７．９％の正分類率を示す。

【００７７】

入力サイズ１６×１６（２５６チャネル）の上述の恒等ＣＮＮは、１００００個の無作為画像に基づいて訓練することができ、かかる訓練は、入力と出力との間の０．０９１３

％の平均絶対誤差を与える。
【００７８】
第１のＮＮの第１又は第２の隠れ層の５１２個のうちの２５６個の入力チャネル上にこの恒等ＣＮＮを挿入することは、第２のＮＮの正分類率の低下を示さない。
【００７９】
事後学習
第１のＮＮ及び１つ又は複数の恒等ＣＮＮのパラメータを事前に得る代替策として、該当する場合には第２のＮＮのアーキテクチャを決定するための上述の選択を実装することにより、第１のＮＮ及び恒等ＣＮＮのモデルから第２のＮＮを構築するステップ（ａ）から直接開始し、その後にのみ第１のＮＮの学習ベース（例えば、上述のＮＩＳＴベース）に基づいて第２のＮＮのパラメータを学習することができる。これは、［図２ｂ］によって示されている実施形態であり、今回は構築及び学習がサーバ１側で実装されていることが理解されるであろう。
【００８０】
この形態は、恒等ＣＮＮのパラメータを別々に学習しなければならないことを防ぎ、それは、残りのＮＮのパラメータと同時に自らのパラメータが自動で学習されるからである。
【００８１】
結果は、均等であり、唯一の不都合は、学習を毎回再実行するのに時間がかかり過ぎるため、要求ごとに第２のＮＮを動的に「再構築」できないことである。
【００８２】
従って、第２の態様によれば、本発明は、恒等関数を近似する少なくとも１つの畳み込みニューラルネットワークが挿入される第１のニューラルネットワークに対応する第２のニューラルネットワークを構築するステップ（ａ）、次いで第２のニューラルネットワークのパラメータをパブリック学習データベースから学習するステップ（ａ１）をやはり含む、サーバ１のデータ処理手段１１によって実装される第２のニューラルネットワークのための学習方法に関する。
【００８３】
本発明の第３の態様では、前記入力データに対して第２のＮＮを使用する同じステップ（ｂ）を追加すること（今回は端末１のデータ処理手段２１によって実装される）により、（第１の態様による方法と同様に）入力データに対して第１のニューラルネットワークを安全に使用するための方法の一部として、この学習方法を使用することができ、即ち第２のＮＮが入力データに適用され、これにより第１のＮＮに戻ることが可能であるいかなるリスクもなしに、端末２のユーザに与えられ得る出力データが得られる。
【００８４】
コンピュータプログラム製品
第４の態様及び第５の態様によれば、本発明は、入力データに対して第１のニューラルネットワークを安全に使用するために、本発明の第１の態様若しくは第３の態様による方法を、又は第２のニューラルネットワークのパラメータを学習するために、本発明の第２の態様による方法を（具体的にはサーバ１又は端末２のデータ処理手段１１、２１上で）実行するためのコード命令を含むコンピュータプログラム製品と、また、このコンピュータプログラム製品が含まれる、コンピュータ機器によって読み出され得る記憶手段（サーバ１又は端末２のメモリ１２、２２）とに関する。

【図面】

【図１】

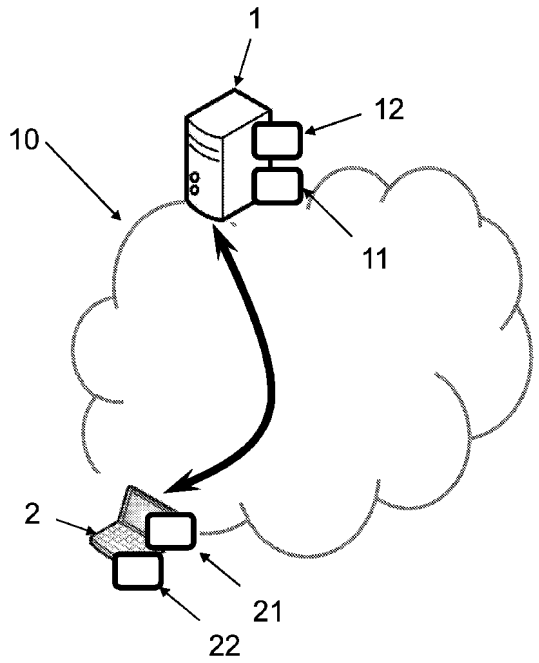


FIG. 1

【図２ａ】

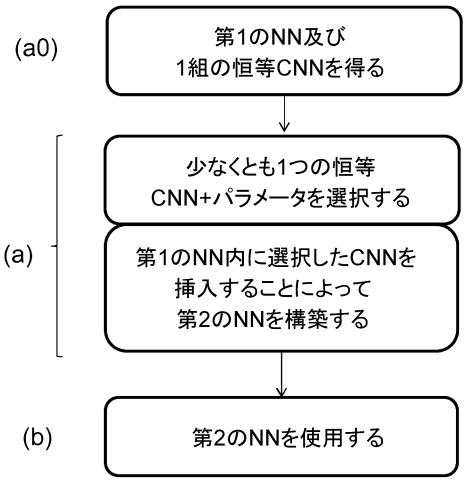


図 2a

【図２ｂ】

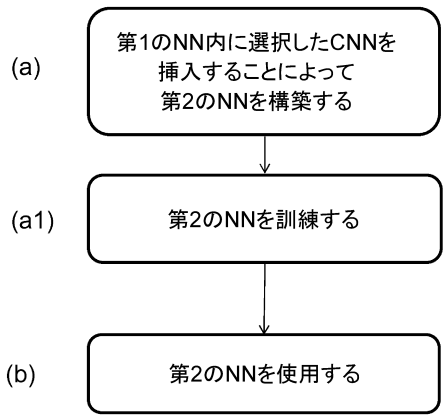


図 2b

【図３】

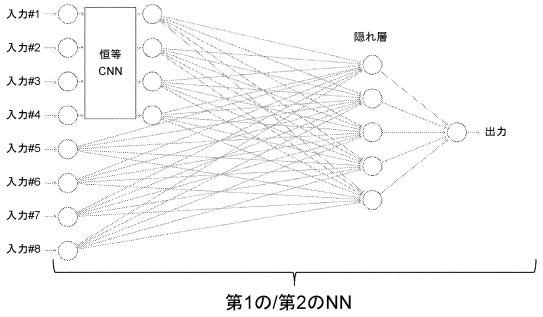


図 3

フロントページの続き

フランス国 ９２４００ クルブボア， プラス サミュエル ドゥ シャンプラン ２， シー／オー アイデミア アイデンティティ アンド セキュリティー フランス

審査官 多賀 実

(56)参考文献

米国特許出願公開第２０１９／００９５６２９（ＵＳ，Ａ１）
中国特許出願公開第１０９８８６２１０（ＣＮ，Ａ）
Mathias Lecuyer et al.，"Certified Robustness to Adversarial Examples with Differential Privacy"，arXiv.org [online]，arXiv:1802.03471v4，米国，Cornell University，2019年，pp.1-18，［令和６年６月３日検索］,インターネット:〈URL: https://arxiv.org/abs/1802.03471v4〉

(58)調査した分野 (Int.Cl.，ＤＢ名)

Ｇ０６Ｎ ３／００－９９／００
Ｇ０６Ｆ ２１／００－２１／８８