(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6755849号
(P6755849)
(45)発行日　令和2年9月16日(2020.9.16)　(24)登録日　令和2年8月28日(2020.8.28)

(51)Int.Cl.　F I
G06N 3/08 (2006.01)　G06N 3/08 120

請求項の数 7　外国語出願　(全 11 頁)

(21)出願番号　特願2017-231460(P2017-231460)
(22)出願日　平成29年12月1日(2017.12.1)
(65)公開番号　特開2018-129033(P2018-129033A)
(43)公開日　平成30年8月16日(2018.8.16)
審査請求日　令和2年4月3日(2020.4.3)
(31)優先権主張番号　16205831.7
(32)優先日　平成28年12月21日(2016.12.21)
(33)優先権主張国・地域又は機関
欧州特許庁(EP)

早期審査対象出願

(73)特許権者　502208205
アクシス　アーベー
スウェーデン国　２２３　６９　ルンド,
エンダラヴェーイェン　１４
(74)代理人　110002077
園田・小林特許業務法人
(72)発明者　セイボルド,　ロビン
スウェーデン国　２２６４６　ルンド,
ケムネルスヴェーゲン　３９エフ
(72)発明者　チェン,　チエンタン
スウェーデン国　２２６　５７　ルンド,
ヨルダバルクスヴェーゲン　５７

最終頁に続く

(54)【発明の名称】人工ニューラルネットワークのクラスに基づく枝刈り

(57)【特許請求の範囲】
【請求項１】
埋め込み装置に特定の監視状況のための人工ニューラルネットワークを構成するための方法であって、
前記人工ニューラルネットワークが訓練される全てのオブジェクトクラスの部分集合を形成する、前記監視状況に特有の幾つかのオブジェクトクラスを選択するステップ、
前記人工ニューラルネットワーク内のニューロンに対する起動頻度の値を含むデータベースにアクセスするステップであって、前記起動頻度の値が各オブジェクトクラスに関する少なくとも１つの起動頻度の値を含む、アクセスするステップ、及び
前記人工ニューラルネットワークから、前記選択されたオブジェクトクラスの前記部分集合に対する起動頻度閾値未満である起動頻度の値を有する前記ニューロンを除去するステップであって、前記除去されるニューロンの数は、前記人工ニューラルネットワークからの出力を計算するために必要な計算資源における低減が、前記除去されたニューロンについての記録を維持するために必要とされる計算資源における増加を上回るようなものである、除去するステップ、を含む、方法。
【請求項２】
前記起動頻度閾値が、前記人工ニューラルネットワークの訓練／試験手順中に各クラス毎に決定された静的値である、請求項１に記載の方法。
【請求項３】
前記起動頻度閾値が、前記人工ニューラルネットワークの性能に基づいて各クラスに対

して適合的である、請求項１に記載の方法。
【請求項４】
前記監視状況に特有の前記オブジェクトクラスの選択を、一定期間にわたって収集されたデータに基づいて精緻化すること、並びに
前記オブジェクトクラスの前記精緻化された選択のために、前記選択するステップ、前記アクセスするステップ、及び前記除去するステップを繰り返すこと、を更に含む、請求項１から３のいずれか一項に記載の方法。
【請求項５】
前記選択するステップ、前記アクセスするステップ、及び前記除去するステップのうちの少なくとも一部が、ユーザインターフェースを介して受信されたユーザ入力に応じて実行される、請求項１から４のいずれか一項に記載の方法。
【請求項６】
完全に接続された人工ニューラルネットワークにおいて、ニューロンを除去することが、前記人工ニューラルネットワークの個々の層の数学的表現を形成する行列から行及び列を除去することに対応し、それによって、前記人工ニューラルネットワークからの前記出力を計算することにおいて必要とされる計算の回数を低減させる、請求項１から５のいずれか一項に記載の方法。
【請求項７】
処理システムにおいて実行されたときに、前記処理システムに、特定の監視状況のための埋め込み装置に用いられる人工ニューラルネットワークを構成するための方法を実行させる指示を含む非一時的なコンピュータ可読記憶媒体であって、前記方法が、
前記人工ニューラルネットワークが訓練される全てのオブジェクトクラスの部分集合を形成する、前記監視状況に特有の幾つかのオブジェクトクラスを選択するステップ、
前記人工ニューラルネットワーク内のニューロンに対する起動頻度の値を含むデータベースにアクセスするステップであって、前記起動頻度の値が各オブジェクトクラスに関する少なくとも１つの起動頻度の値を含む、アクセスするステップ、及び
前記人工ニューラルネットワークから、前記選択されたオブジェクトクラスの前記部分集合に対する起動頻度閾値未満である起動頻度の値を有する前記ニューロンを除去するステップであって、前記除去されるニューロンの数は、前記人工ニューラルネットワークからの出力を計算するために必要な計算資源における低減が、前記除去されたニューロンについての記録を維持するために必要とされる計算資源における増加を上回るようなものである、除去するステップ、を含む、非一時的なコンピュータ可読記憶媒体。
【発明の詳細な説明】
【背景技術】
【０００１】
本発明は、人工ニューラルネットワークに関し、特に、オブジェクト認識及び検証のために使用される人工ニューラルネットワークに関する。近年、ネットワーク拡張及び訓練データの増加のおかげで、そのような人工ニューラルネットワークが、大幅に改良されてきた。しかし、複雑な人工ニューラルネットワークのアーキテクチャは、しばしば、何千万又は何億のパラメータを含む。そのような人工ニューラルネットワークは精度に優れるが、大量のパラメータは、殊に、しばしば、計算能力を制限してきた埋め込みシステムにおいて、展開を不可能にし得る。人工ニューラルネットワークのサイズを増加した結果として、改良を維持するために、計算能力を制限してきたシステム内に実装されることを可能にすると同時に、人工ニューラルネットワークを圧縮することに関心が高まっている。
【０００２】
上述したように、人工ニューラルネットワークは、通常、非常に大きいので、それらは、しばしば、「過剰にパラメータ化」され得る。これは、人工ニューラルネットワークの性能に大きな影響を与えることなく、重み及びバイアスなどのパラメータ、又は全体のニューロンを除去することを可能にする。この手順は、通常、人工ニューラルネットワークの「枝刈り」と称される。ニューロンが除去されるときに、そのニューロンに対する計算

をバックトレース（back-trace）することが可能である。その後、そのニューロンをもたらす全ての重みが、安全に除去され得ることが見られ得る。ニューロン出力を追跡し、そのニューロンから行く（going from）重みを除去することも可能である。しかし、枝刈りプロセスにおいてどのニューロンを除去するかを特定すること、及び、性能が得られるようなやり方で枝刈りプロセスを実施することは、尋常なことではない。

【０００３】

枝刈りは、訓練可能なパラメータを含む層、従来から完全に接続されている層、及び畳み込み層（convolutional layer）に適用可能である。これは、計算を単純化し速度を上げる助けとなる。例えば、完全に接続されている層からニューロンを除去することは、行列の行とベクトルとの間のドット積（dot product）をスキップすることに等しい。結果として、行列は、より小さくなる。畳み込み層からニューロンを除去することは、１つの行列の行と１つの行列の列との間のドット積をスキップすることを意味し、それは、１つの畳み込みをスキップすることと同じである。ニューロンの除去は、以下の詳細な説明で更に詳細に説明されることとなる。人工ニューラルネットワークの精度に重大な影響を与えることなく除去され得るニューロンを決定することは、訓練／試験段階の間に、どのニューロンが「死んで」いるか、すなわち、どのニューロンが、ゼロではない出力をほとんど又は全く生み出さないかを特定する結果としてのデータから、ニューロンを解析することによって行われ得る。ニューロンが、死んでいるものとして規定されないために、ゼロではない出力を何回生み出さなければいけないかを決定することは、種々の閾値を使用して、性能を比較することによって行われ得る。そして、ニューロンが除去された後で、性能を改良するために再訓練が行われ得る。これは、繰り返し行われ得る。

【０００４】

枝刈りのための別の１つのアプローチは、既に訓練ステップ中に一部の重みをゼロに強制するために、閾値を使用するか又はノルム（norm）を用いた正規化を使用して、重みを除去することに重点を置いている。正規化は、例えば、（幾つかの値をゼロに強制する）希薄さ（sparsity）又は滑らかさなどの条件を実施するために使用される、当業者に周知の数学的／統計的方法である。枝刈りのための正規化についての更なる詳細は、次の論文を参照すべし。「Memory bounded deep convolutional networks」arXiv CoRR 2014 - Section 3: Regularization Updates available online at https://arxiv.org/abs/1412.1442.

【０００５】

これらのやり方で人工ニューラルネットワークを十分に枝刈りすることによって、実施は除去されたパラメータに対する計算を避けることができるので、人工ニューラルネットワークを実行するための計算はフルネットワークの場合よりも低くなる。

【発明の概要】

【０００６】

本発明の目的は、上述の問題を全体的又は部分的に除去することである。この目的は、請求項１に記載の方法、請求項９に記載のコンピュータプログラム製品、請求項１５に記載の記憶媒体によって実現される。

【０００７】

第１の態様によれば、本発明は、コンピュータシステムにおいて、特定の監視状況のための人口ニューラルネットワークを構成するための方法に関する。該方法は、以下のステップによって特徴付けられる。すなわち、

人口ニューラルネットワークが訓練される全てのオブジェクトクラスの部分集合を形成する、監視状況に特有の幾つかのオブジェクトクラスを選択するステップ、

人工ニューラルネットワーク内のニューロンに対する起動頻度の値を含むデータベースにアクセスするステップであって、起動頻度の値がオブジェクトクラスの関数である、アクセスするステップ、及び

選択されたオブジェクトクラスの部分集合に対する閾値未満である起動頻度の値を有するそれらのニューロンを、人工ニューラルネットワークから除去するステップである。

【０００８】
これは、負担が軽く、したがって、カメラ又は携帯デバイスなどの埋め込みデバイスに対して適切な、人工ニューラルネットワークを生成するやり方を提供する。特に、幾つかのクラス（すなわち、全ての可能なクラスの部分集合）のみが、ニューロンの使用頻度に基づいて、予測され、それらのクラスに人工ニューラルネットワークを枝刈りし得ることを知ることによって、大規模な人工ニューラルネットワークを必要に応じて異なる前提条件に適合させることが可能であり、従来のように人工ニューラルネットワークを再訓練する必要はない。人工ニューラルネットワークを再訓練することは、通常、かなりの時間を必要とし、且つ、再訓練のための新しいデータを必要とするが、人工ニューラルネットワークを特定の組のクラスに枝刈りすることは、より速い作業である。更に、この方法は（例えば、閾値処理又は正規化を使用して重みをゼロにする）他の従来の方法と組み合わせて使用されて、更により多数のニューロンを除去することができ、埋め込みデバイスに人工ニューラルネットワークを展開するために、計算性能を更に低下させることができる。
【０００９】
一実施形態によれば、閾値は、人工ニューラルネットワークのための訓練手順中に各クラス毎に決定された静的値（static value）である。異なる関連するクラスに対して異なる閾値を可能にすることによって、関連するクラス毎に同じ閾値が全体的に使用された場合と比較して、システムの精度を向上させることが可能になる。
【００１０】
一実施形態によれば、閾値は、システム性能に基づいて各クラスに対して適合的（adaptive）である。システム性能に基づく適合的閾値を有することによって、システムが展開される実際の監視状況から得られたデータに基づいて、システムを更に強化することができる。
【００１１】
一実施形態によれば、監視状況に特有のオブジェクトクラスの選択は、一定期間にわたって収集されたデータに基づいて精緻化することができ、オブジェクトクラスの精緻化された選択のために、選択するステップ、アクセスするステップ、及び除去するステップを繰り返すことができる。これにより、実際の観測データとシステム性能の評価に基づいて、システムの更なる精緻化と微調整が可能となる。
【００１２】
一実施形態によれば、選択するステップ、アクセスするステップ、及び除去するステップのうちの少なくとも一部は、ユーザインターフェースを介して受信されたユーザ入力に応じて実行される。ユーザインターフェースを有することにより、ユーザは、枝刈りを、ユーザからの入力に基づくか、どの程度自動化すべきかを決定することができる。例えば、ユーザは、枝刈りされた人工ニューラルネットワークの機能を更に改善するために、関連するクラスの追加又は削除、個々のクラスに対する閾値の変更などを決定することができる。
【００１３】
一実施形態によれば、完全に接続された人工ニューラルネットワークにおけるニューロンを除去することは、人工ニューラルネットワークの個々の層の数学的表現を形成する行列から行及び列を除去することに対応する。重みを除去するだけではなく、ニューロン全体を除去することによって、人工ニューラルネットワークからの出力を計算するのに必要な計算回数を大幅に削減することができる。
【００１４】
一実施形態によれば、人工ニューラルネットワークからの出力を計算するために必要な計算資源の削減が、除去されたニューロンに関する記録を維持するのに必要な計算資源の増加を上回るように、人工ニューラルネットワークからニューロンを除去することは、十分な数のニューロンを除去することを含む。すなわち、人工ニューラルネットワークの枝刈りが、枝刈りされていないネットワークに比べて必要な計算資源が非常に少なく、したがって、人工ニューラルネットワークの性能が高いままであると同時に、埋め込みデバイ

スへの展開を適切にするように、損益分岐点が見つけられ得る。
【００１５】
一実施形態によれば、人工ニューラルネットワークは、埋め込みデバイスへ展開される。例えば、埋め込みデバイスは、監視カメラ又は携帯電話であり得る。人工ニューラルネットワークをこの種類の埋め込みデバイスで使用し得るようにすることによって、埋め込みデバイスの動作が大幅に高められ得る。
【００１６】
第２の態様によれば、本発明は、特定の監視状況のための人工ニューラルネットワークを構成するためのコンピュータプログラムに関する。コンピュータプログラムは、以下のステップに対応する指示命令を含む。すなわち、
人工ニューラルネットワークが訓練される全てのオブジェクトクラスの部分集合を形成する、監視状況に特有の幾つかのオブジェクトクラスを選択するステップ、
人工ニューラルネットワーク内のニューロンに対する起動頻度の値を含むデータベースにアクセスするステップであって、起動頻度の値がオブジェクトクラスの関数である、アクセスするステップ、及び
選択されたオブジェクトクラスの部分集合に対する閾値未満の起動頻度の値を有するそれらのニューロンを、人工ニューラルネットワークから除去するステップである。
【００１７】
第３の態様によれば、本発明は、そのようなプログラムを含むデジタル記憶媒体に関する。コンピュータプログラムと記憶媒体は、その方法のものと対応する利点を含み、同様に変更され得る。
【００１８】
本発明の１以上の実施形態の詳細が、添付の図面及び以下の詳細な説明において説明される。本発明の他の特徴と利点は、詳細な説明、図面、及び特許請求の範囲から明らかである。
【図面の簡単な説明】
【００１９】
【図１】一実施形態による、人工ニューラルネットワークを枝刈りするためのプロセス１００のフローチャートを示す。
【図２】一実施形態による、人工ニューラルネットワーク２００の概略図を示す。
【図３】一実施形態による、枝刈り前の図２のニューラルネットワークの完全な計算のための方程式３００を示す。
【図４】一実施形態による、枝刈り後の図２のニューラルネットワークの計算のための方程式４００を示す。
【発明を実施するための形態】
【００２０】
様々な図面内の類似の参照記号は、類似の要素を示す。
【００２１】
概観
上述されたように、本発明の１つの目的は、人工ニューラルネットワークを、カメラ及び携帯デバイスなどの埋め込みデバイス内で使用可能とするために、人工ニューラルネットワークを枝刈りすることである。本明細書で説明される様々な実施形態は、特定のクラスの画像のためのニューロンの使用頻度に基づいて、オブジェクトの各クラスに対して別々に重みを枝刈りすることによって、これを達成する。これは、従来技術と同様のやり方で、例えば、試験段階中の人工ニューラルネットワークのニューロンを解析することによって、どのニューロンがそれほど頻繁に使用されないかを特定する、そのような解析の結果から行うことができる。しかし、本明細書で説明される実施形態によれば、あるクラスに属する画像のみが、人工ニューラルネットワークを介して供給され、低性能ニューロンが特定される。
【００２２】

当業者によって認められるように、本発明の態様は、システム、方法、又はコンピュータプログラム製品として具現化が可能である。したがって、本発明の態様は、全体的にハードウェアの実施形態、全体的に（ファームウェア、常駐ソフトウェア、マイクロコードなどを含む）ソフトウェアの実施形態、又はソフトウェアとハードウェアを組み合わせた（本明細書で、概して、「回路」、「モジュール」、又は「システム」と称され得る）態様の実施形態の形を採り得る。更に、本発明の態様は、そこで具現化されたコンピュータ可読プログラムコードを有する（１以上の）コンピュータ可読媒体に具現化されたコンピュータプログラム製品の形態を採り得る。

【００２３】

１以上のコンピュータ可読媒体の任意の組み合わせが利用され得る。コンピュータ可読媒体は、コンピュータ可読信号媒体又はコンピュータ可読記憶媒体であり得る。例えば、コンピュータ可読記憶媒体は、電気的、磁気的、光の、電磁気的、赤外線の、又は半導体のシステム、装置、若しくはデバイス、又はそれらの任意の適切な組み合わせであり得るが、それらに限定されるものではない。コンピュータ可読記憶媒体のより具体的な例（網羅的でないリスト）は、以下のものを含み得る。すなわち、１以上のワイヤーを有する電気的な接続、携帯型コンピュータディスケット、ハードディスク、ランダムアクセスメモリ（ＲＡＭ）、リードオンリーメモリ（ＲＯＭ）、消去可能プログラマブルリードオンリーメモリ（ＥＰＲＯＭ又はフラッシュメモリ）、光ファイバー、携帯型コンパクトディスクリードオンリーメモリ（ＣＤ‐ＲＯＭ）、光記憶デバイス、磁気記憶デバイス、又はそれらの任意の適切な組み合わせである。この文書の文脈では、コンピュータ可読記憶媒体は、指示命令実行システム、装置、若しくはデバイスによって又はそれらとの関連で使用されるプログラムを含み又は記憶することができる、任意の有形の媒体であり得る。

【００２４】

コンピュータ可読信号媒体は、内部で具現化されるコンピュータ可読プログラムコードを有する、例えば、ベースバンド内で又は搬送波の部分として伝搬されるデータ信号を含み得る。そのような伝搬される信号は、電磁気的、光の、又はそれらの任意の適切な組み合わせを含む、任意の様々な形態を採り得るが、それらに限定されるものではない。コンピュータ可読信号媒体は、コンピュータ可読記憶媒体ではなく、指示命令実行システム、装置、若しくはデバイスによって又はそれらとの関連で使用されるプログラムを通信、伝搬、又は移送することができる、任意のコンピュータ媒体であり得る。

【００２５】

コンピュータ可読媒体で具現化されるプログラムコードは、無線、有線、光ファイバーケーブル、ＲＦなど、又はそれらの任意の適切な組み合わせを含む、任意の適切な媒体を使用して送信され得る。本発明の態様のために動作を実行するためのコンピュータプログラムコードは、Ｊａｖａ、Ｓｍａｌｌｔａｌｋ、Ｃ＋＋などのオブジェクト指向のプログラミング言語、及び、「Ｃ」プログラミング言語又は類似のプログラミング言語などの従来の手続き型のプログラミング言語を含む、１以上のプログラミング言語の任意の組み合わせにおいて書かれ得る。プログラムコードは、ユーザのコンピュータで全体的に、ユーザのコンピュータで部分的に、独立型ソフトウェアパッケージとして、ユーザのコンピュータで部分的に、遠隔コンピュータで部分的に、又は遠隔コンピュータ若しくはサーバで全体的に実行され得る。後者のシナリオでは、遠隔コンピュータが、ローカルエリアネットワーク（ＬＡＮ）若しくはワイドエリアネットワーク（ＷＡＮ）を含む、任意の種類のネットワークを通じてユーザのコンピュータに接続され得る。または、（例えば、インターネットサービスプロバイダーを使用して、インターネットを通じて）外部コンピュータに対する接続が行われ得る。

【００２６】

本発明の態様は、本発明の実施形態による、方法、装置（システム）、並びにコンピュータプログラム製品のフローチャート図解及び／又はブロック図を参照して、以下で説明される。フローチャート図解及び／又はブロック図の各ブロック、並びにフローチャート図解及び／又はブロック図内のブロックの組み合わせは、コンピュータプログラム指示命

令によって実施され得ることが理解される。コンピュータ又は他のプログラマブルデータ処理装置を介して実行される指示命令が、フローチャート及び／又はブロック図の１以上のブロック内で特定される機能／動作を実施する、ための手段を生成するように機械を生産するために、汎用コンピュータ、専用コンピュータ、又は他のプログラマブルデータ処理装置のプロセッサに、これらのプログラム指示命令が提供され得る。

【００２７】

これらのコンピュータプログラム指示命令は、特定のやり方で機能するようにコンピュータ、他のプログラマブルデータ処理装置、又は他のデバイスに指示することができるコンピュータ可読媒体にも記憶され得る。それによって、コンピュータ可読媒体に記憶された指示命令は、フローチャート及び／又はブロック図の１以上のブロック内で特定される機能／動作を実施する指示命令を含む製品を製造する。

【００２８】

コンピュータプログラム指示命令は、コンピュータ、他のプログラマブルデータ処理装置、又は他のデバイスにもロードされ、一連の動作ステップがコンピュータ、他のプログラマグルデータ処理装置、又は他のデバイスで実行されて、コンピュータ実施プロセスを生み出すことをもたらし得る。それによって、コンピュータ又は他のプログラマブル装置で実行される指示命令は、フローチャート及び／又はブロック図の１以上のブロック内で特定された機能／動作を実施するためのプロセスを提供する。

【００２９】

人工ニューラルネットワークの枝刈り

次に、本発明の様々な実施形態による技術が、実施例によって図１～図４を参照しながら説明される。この実施例では、１０００のクラスに対して訓練が実行されたところの、人工ニューラルネットワークが存在すると想定する。しかし、画像が監視カメラによって記録される状況では、例えば、ユーザは、１０００のクラスのうちの１２しか関心がない。

【００３０】

図１は、本発明の一実施形態による、人工ニューラルネットワークを枝刈りするためのプロセス１００を示しているフローチャートである。図１で示され得るように、これらのクラスに対してどのニューロンが低性能であるかを理解するために、ステップ１０２でデータベースがアクセスされる。データベースは、人工ニューラルネットワーク内のニューロンに対する起動頻度の値を、オブジェクトクラスの関数として含む。

【００３１】

次に、ステップ１０４で、人工ニューラルネットワークから安全に除去され得る低性能のニューロンが特定される。これは、例えば、１２のクラスの各々に対するニューロンのための起動頻度の値を調べることによって、及び、どの起動頻度の値が低性能のニューロンを構成するかを規定する閾値を使用することによって、行われ得る。閾値は、人工ニューラルネットワークの訓練手順の間に種々のクラスに従って予め決定され、又は予測性能に従って推論手順の間に適合的（adaptive）であり得る。すなわち、閾値は、システム性能に適合され得る。

【００３２】

最後に、ステップ１０６で、低性能のニューロンが人工ニューラルネットワークから除去される。これによりプロセス１００が終了し、ユーザの必要に応じてそのアーキテクチャを調整することができる、「適合的人工ニューラルネットワーク」をもたらす。

【００３３】

ある実施形態では、このプロセスが、ユーザインターフェース（ＵＩ）を介したユーザからの特定の入力を必要とし、プロセスは、望まれるように繰り返され得る。例えば、一実施形態では、特定の監視状況の完全な評価の期間（例えば、一週間）が存在し得る。その後に、その期間に特定されたクラスを選択及び枝刈りすることが行われる。

【００３４】

自動化の様々な程度が、ユーザがクラスを選択する助けとなるように使用され得る。例

えば、ある実施形態では、クラスの選択が完全に自動で行われ得る。他の実施形態では、ユーザに、自動的に選択されたクラスの部分集合が提示され得る。それらから、ユーザが手動で選択を行い得る。更に他の実施形態では、ユーザが、全てのクラスの中から自由に選択し得る。ある実施形態は、ユーザが一組の選択されたクラスに対する追加を行うことも可能にする。それは、ユーザが、そのような追加が目前の特定の状況に対して有益であり得ると判定した場合である。特定の状況に基づいて、当業者は、多くの変形例を想起し得る。

【００３５】

図２は、小さい人工ニューラルネットワーク２００の概略図を示している。それは、適合的人工ニューラルネットワークを生成するために、上述の技術を使用して、（破線で示されている）特定のニューロンが除去されたところの、完全に接続された層のみから成る。図３は、ニューロンが実際に除去される前の、人工ニューラルネットワークに対する完全な計算３００を示している。そして、図４は、除去されたニューロンに対応する計算を除去した後の計算４００、すなわち、結果としての適合的人工ニューラルネットワークによって実行される計算を示している。

【００３６】

それぞれ、図３と図４の方程式を比較することによって見られ得るように、低性能のニューロンを除去した後で、計算量が大幅に低減されている。データを訓練することにおける変動性が低減され、それによって、より薄い人工ニューラルネットワークをもたらすので、通常、枝刈りがより少ないクラスに基づくならば（例えば、ユーザが１２の代わりに６のクラスのみに関心があったならば）、より多くのニューロン及び重みが枝刈りされ得る。

【００３７】

疎行列の表現

人工ニューラルネットワークの枝刈りの従来のやり方は、独立して重み値を見ること、及び特定の閾値未満である重みを除去することによって、行われる。これにより、層に希薄さが導入されるが、その希薄さは構造化されてない。その代わりに、枝刈りがニューロンに基づくならば、枝刈りされたニューロンに貢献する全ての重みは除去され、構造化された希薄さをもたらし得る。完全に接続された層を有する人工ニューラルネットワークの場合では、これは、出力計算における全体の行と列が除去され得ることを意味する。

【００３８】

重み行列で必要とされる唯一の演算は乗算であるため、枝刈りされたニューロンに対応する行と列は、結果に影響を与えることなしに除去され得る。この技術を使用すると、従来の枝刈り方法でも一般的であるように、行列のサイズが小さくなるが、疎行列を表現するために使用される指標（indices）は、行列毎には記憶されず、最終的な完全なサイズの出力を再構築するためにのみ、結果としての行列について記憶される。これは、更なる利益ももたらし得る。例えば、結果としての疎行列は、以下で更に詳細に説明されるように、「ブックキーピング」を扱うためのより少ない計算能力を必要とする構造を有し得る。

【００３９】

かなりな数の枝刈りされたニューロンが存在する限り、結果としての疎行列の乗算は、完全な行列の乗算よりも速い。それは、より多くのブックキーピング（すなわち、枝刈りされたニューロンが位置する指標を追跡すること）を必要とするが、より速い少ない乗算を必要とする。それは、従来の方法と比較して記憶及びメモリ空間も節約し、通常、疎行列では不可能である単一命令多重データ（ＳＩＭＤ）などのベクトル演算を容易にし、ＳＩＭＤは演算をより速くし得る。正確にどれが「かなりな数の」枝刈りされたニューロンを構成するかは、ケースバイケースで変動する。しかし、どの場合でも、より少ない計算における利得が記録を維持することにおける損失を上回る、損益分岐点が存在するだろう。

【００４０】

更に、疎行列は、数学的に多くの異なるやり方で表現され得る。それは、特定の構造に対する正しい表現を使用することによって、この損益分岐点に到達するための閾値を更に低減させ得る可能性が高い。

【００４１】

最後に、当業者が理解するように、本明細書で示されている計算から行列の全体の行と列が除去されたときに、結果として生じる疎行列の次元は変化することとなる。結論として、ニューラルネットワークの最後の層が枝刈りされるならば、特定の出力を特定のクラスに関連付けることを可能とするために、クラスラベルファイルを更新する必要がある。別の一代替例は、最後の層を枝刈りしない。その場合、元々のクラスラベルファイルが、そのまま使用され得る。

【００４２】

本発明の様々な実施形態による、システム、方法、及びコンピュータプログラム製品の可能な実施態様のアーキテクチャ、機能、及び動作を示す図面のフローチャートとブロック図が、開示された。これに関して、フローチャート又はブロック図内の各ブロックは、（１以上の）特定の論理機能を実施するための１以上の実行可能な指示命令を含む、モジュール、セグメント、又はコードの部分を表し得る。ある代替的な実施態様では、ブロック内で記された機能が、図面で記されたのとは異なる順序で生じ得ることも留意されるべきである。例えば、連続して示された２つのブロックは、実際、実質的に同時に実行され得る。または、ブロックは、時々、含まれる機能に応じて、逆順に実行され得る。ブロック図及び／又はフローチャート図解の各ブロック、並びにブロック図及び／又はフローチャート図解内のブロックの組み合わせが、特定の機能若しくは動作を実行する、専用ハードウェアベースシステム、又は専用ハードウェアとコンピュータ指示命令の組み合わせによって実施され得る。

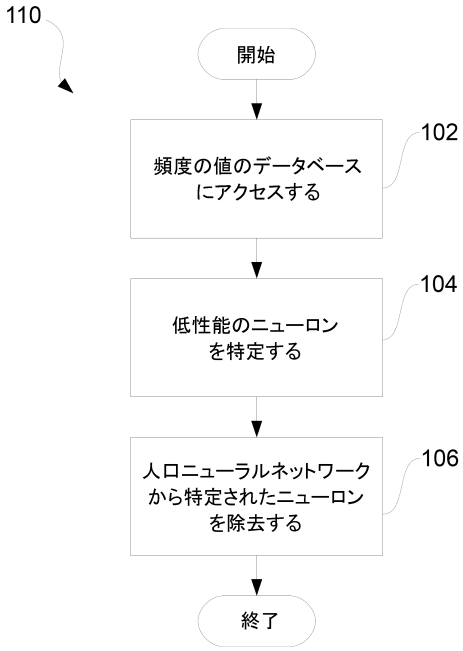
【００４３】

本明細書で使用される用語は、特定の実施形態を説明することのみを目的としており、本発明を限定することを意図していない。本明細書で使用される際に、単一形「ａ」、「ａｎ」、及び「ｔｈｅ」は、文脈が明確にその逆を示している場合を除き、複数形も含むことが意図されている。「含む、備える（comprise）」及び／又は「含む、備える（comprising）」という用語は、この明細書で使用されるときに、述べられている特徴、完全体、ステップ、動作、要素、及び／又は構成要素の存在を特定するが、１以上の他の特徴、完全体、ステップ、動作、要素、構成要素、及び／又はそれらの群の存在又は追加を排除するものではないことが、更に理解される。

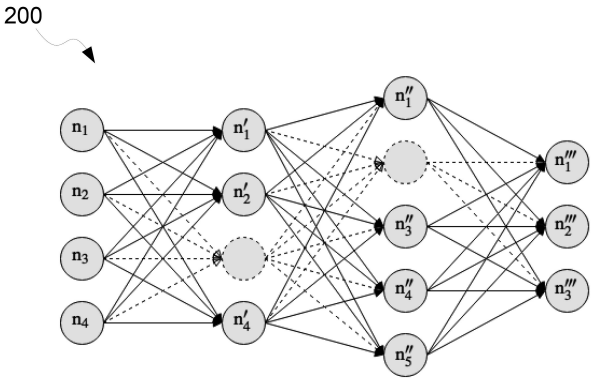
【００４４】

以下の特許請求の範囲内の全て手段又はステップと機能要素に対応する構造、材料、動作、及び等価物は、具体的に請求された他の請求された要素との組み合わせにおいて機能を実行するための任意の構造、材料、又は動作を含むことが意図されている。本発明の説明が、例示及び説明を目的として提示されてきたが、それは、網羅的なものであること、又は本発明を開示された形態に限定することを意図するものではない。本発明の範囲及び精神から逸脱することなく、多くの修正例及び変形例が、当業者には明らかであろう。例えば、本明細書で説明された方法は、独立した方法としてのみ実行され得るわけではなく、他の既知の方法とも組み合わされて、人工ニューラルネットワークの枝刈りを向上させる。本発明の原理及び実施の用途を最も優れて説明するため、他の当業者が、熟考された特定の使用に適した様々な変形例を有する様々な実施形態について、本発明を理解することを可能にするために、特定の実施形態が選ばれ説明された。

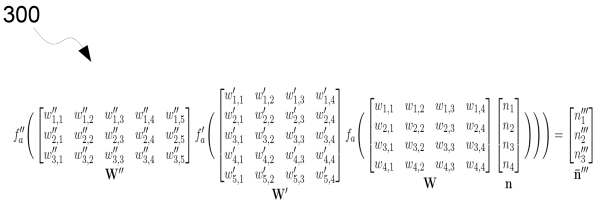
【図１】



【図２】

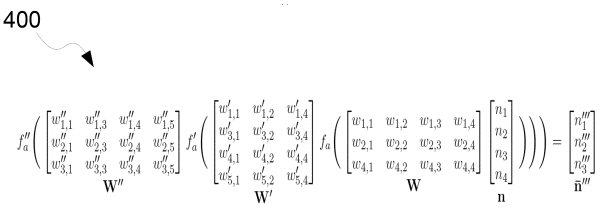


【図３】



$$f''_a\left(\begin{bmatrix} w''_{1,1} & w''_{1,2} & w''_{1,3} & w''_{1,4} & w''_{1,5} \\ w''_{2,1} & w''_{2,2} & w''_{2,3} & w''_{2,4} & w''_{2,5} \\ w''_{3,1} & w''_{3,2} & w''_{3,3} & w''_{3,4} & w''_{3,5} \end{bmatrix}_{\mathrm{W''}} f'_a\left(\begin{bmatrix} w'_{1,1} & w'_{1,2} & w'_{1,3} & w'_{1,4} \\ w'_{2,1} & w'_{2,2} & w'_{2,3} & w'_{2,4} \\ w'_{3,1} & w'_{3,2} & w'_{3,3} & w'_{3,4} \\ w'_{4,1} & w'_{4,2} & w'_{4,3} & w'_{4,4} \\ w'_{5,1} & w'_{5,2} & w'_{5,3} & w'_{5,4} \end{bmatrix}_{\mathrm{W'}} f_a\left(\begin{bmatrix} w_{1,1} & w_{1,2} & w_{1,3} & w_{1,4} \\ w_{2,1} & w_{2,2} & w_{2,3} & w_{2,4} \\ w_{3,1} & w_{3,2} & w_{3,3} & w_{3,4} \\ w_{4,1} & w_{4,2} & w_{4,3} & w_{4,4} \end{bmatrix}_{\mathrm{W}} \begin{bmatrix} n_1 \\ n_2 \\ n_3 \\ n_4 \end{bmatrix}_{\mathrm{n}}\right)\right)\right) = \begin{bmatrix} n'''_1 \\ n'''_2 \\ n'''_3 \end{bmatrix}_{\tilde{\mathrm{n}}'''}$$

【図４】



$$f''_a\left(\begin{bmatrix} w''_{1,1} & w''_{1,3} & w''_{1,4} & w''_{1,5} \\ w''_{2,1} & w''_{2,3} & w''_{2,4} & w''_{2,5} \\ w''_{3,1} & w''_{3,3} & w''_{3,4} & w''_{3,5} \end{bmatrix}_{\mathrm{W''}} f'_a\left(\begin{bmatrix} w'_{1,1} & w'_{1,2} & w'_{1,4} \\ w'_{3,1} & w'_{3,2} & w'_{3,4} \\ w'_{4,1} & w'_{4,2} & w'_{4,4} \\ w'_{5,1} & w'_{5,2} & w'_{5,4} \end{bmatrix}_{\mathrm{W'}} f_a\left(\begin{bmatrix} w_{1,1} & w_{1,2} & w_{1,3} & w_{1,4} \\ w_{2,1} & w_{2,2} & w_{2,3} & w_{2,4} \\ w_{4,1} & w_{4,2} & w_{4,3} & w_{4,4} \end{bmatrix}_{\mathrm{W}} \begin{bmatrix} n_1 \\ n_2 \\ n_3 \\ n_4 \end{bmatrix}_{\mathrm{n}}\right)\right)\right) = \begin{bmatrix} n'''_1 \\ n'''_2 \\ n'''_3 \end{bmatrix}_{\tilde{\mathrm{n}}'''}$$

フロントページの続き

(72)発明者　ビュオグヴィンスドッティル，　ハンナ
スウェーデン国　２２３　６９　ルンド，　エンダラヴェーゲン　１４，　シー／オー　アクシス　コミュニケーションズ　アーベー
(72)発明者　ユングクヴィスト，　マルティン
スウェーデン国　２２６　４９　ルンド，　クルグレンデン　３　デー

審査官　今城　朋彬

(56)参考文献　特開２００６－０１１８４９（ＪＰ，Ａ）
米国特許出願公開第２００７／０２３３６２３（ＵＳ，Ａ１）
特開２０００－２９８６６１（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｎ　　３／０８