(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7232599号
(P7232599)

(45)発行日　令和5年3月3日(2023.3.3)

(24)登録日　令和5年2月22日(2023.2.22)

(51)国際特許分類
Ｇ０６Ｎ　3/082(2023.01)

ＦＩ
Ｇ０６Ｎ　3/082

請求項の数　11　外国語出願　（全11頁）

(21)出願番号　特願2018-165782(P2018-165782)
(22)出願日　平成30年9月5日(2018.9.5)
(65)公開番号　特開2019-49977(P2019-49977A)
(43)公開日　平成31年3月28日(2019.3.28)
審査請求日　令和3年8月25日(2021.8.25)
(31)優先権主張番号　15/699,438
(32)優先日　平成29年9月8日(2017.9.8)
(33)優先権主張国・地域又は機関
米国(US)

(73)特許権者　507186492
ビバンテ コーポレーション
アメリカ合衆国，カリフォルニア州 ９５００２，サン ノゼ，ゴールド ストリート ２１５０，スイート ２００
(74)代理人　100121083
弁理士　青木 宏義
(74)代理人　100138391
弁理士　天田 昌行
(74)代理人　100074099
弁理士　大菅 義之
(74)代理人　100106851
弁理士　野村 泰久
(72)発明者　シン ウォン
アメリカ合衆国，カリフォルニア州 ９
最終頁に続く

(54)【発明の名称】　畳み込みニューラルネットワークのための刈り込みと再学習法

(57)【特許請求の範囲】
【請求項１】
（ａ）コンピューティングデバイスによって、ニューラルネットワークを学習し、
（ｂ）前記コンピューティングデバイスによって、前記ニューラルネットワークを刈り込み、
（ｃ）前記コンピューティングデバイスによって、前記（ｂ）を実行した後に、再学習の期間中に前記（ｂ）の刈り込みに従って制約を課すことなく前記ニューラルネットワークを前記再学習し、
（ｄ）前記コンピューティングデバイスによって、前記（ｃ）の再学習が、後続繰り返しにおいて、前の繰り返しでの前記（ｂ）の刈り込みに従う制約なしに実行されるように、前記（ｂ）および前記（ｃ）を１回以上繰り返す、
ことを含む方法であって、
前記（ｂ）が、前記コンピューティングデバイスによって、残りの結合の数が、前記ニューラルネットワーク内の可能な結合の数に関する刈り込み率となるように、除去されていない前記ニューラルネットワークの前記残りの結合よりも低い重みを有する前記ニューラルネットワーク内の結合を除去することによって、前記ニューラルネットワークを刈り込むことを含み、
前記刈り込み率は所定の値であり、
前記方法が、
（ｅ）前記コンピューティングデバイスによって、前記（ｄ）の各繰り返しの前に前記刈

り込み率を増加させる、
ことを更に含む、
方法。
【請求項２】
目標率と等しい前記刈り込み率で前記（ｄ）が実行されるまで、前記（ｄ）及び前記（ｅ）を実行することを、更に含む、請求項１に記載の方法。
【請求項３】
（ｆ）前記（ｄ）および前記（ｅ）の実行の後に、前記（ｄ）が前記目標率に等しい前記刈り込み率で実行されるまで、前記コンピューティングデバイスによって、前記（ｂ）の最後の繰り返しでゼロに重み付けするように制約されて刈り込まれた結合を用いて前記ニューラルネットワークを再学習し、
前記コンピューティングデバイスによって、前記（ａ）から前記（ｆ）の各繰り返しで前記目標率を増加しながら、前記（ａ）から前記（ｆ）を１回以上繰り返す、
ことを更に含む、請求項２に記載の方法。
【請求項４】
前記（ｅ）の後で前記（ｆ）を実行する前に、前記ニューラルネットワークの精度が安定な閾値条件を達成するまで、前記刈り込み率を前記目標率に固定して前記（ｄ）を繰り返すことを、更に含む請求項３に記載の方法。
【請求項５】
前記（ｄ）を実行することは、前記ニューラルネットワークの精度が前記（ａ）を実行することによって達成される前記ニューラルネットワークの精度よりも小さく、閾値量の範囲内になるまで、前記（ｂ）および前記（ｃ）を実行することを含む、請求項３に記載の方法。
【請求項６】
前記ニューラルネットワークは、畳み込みニューラルネットワーク（ＣＮＮ）である、請求項１に記載の方法。
【請求項７】
１つ以上の処理装置と１つ以上のメモリデバイスとを含むシステムであって、前記１つ以上のメモリデバイスは、
（ａ）畳み込みニューラルネットワーク（ＣＮＮ）であるニューラルネットワークを学習し、
（ｂ）前記ニューラルネットワークを刈り込み、および、
（ｃ）前記（ｂ）を実行した後で、再学習の期間中に前記（ｂ）の刈り込みに従って前記ニューラルネットワークに制約を課すことなく、前記ニューラルネットワークを前記再学習し、
（ｄ）前記（ｃ）の再学習が、次の繰り返しにおいて前の繰り返しにおける前記（ｂ）の刈り込みに従って制約されることなしに実行されるように、前記（ｂ）および前記（ｃ）を１回以上繰り返す、
処理を、前記１つ以上の処理装置に実行させるのに有効である実行可能コードを格納し、
前記実行可能コードは、
残りの結合の数が、前記ニューラルネットワーク内の可能な結合の数に関する刈り込み率となるように、除去されていない前記ニューラルネットワークの前記残りの結合よりも低い重みを有する前記ニューラルネットワーク内の結合を除去することによって、前記（ｂ）において刈り込みを実行する、処理を、前記１つ以上の処理装置に実行させるのに更に有効であり、
前記刈り込み率は所定の値であり、
前記実行可能コードは、
（ｅ）前記刈り込み率を前記（ｄ）の各繰り返しの後に増加させる、
処理を、前記１つ以上の処理装置に実行させるのに更に有効である、
システム。

【請求項８】
前記実行可能コードは、目標率と等しい前記刈り込み率で前記（ｄ）が実行されるまで、前記１つ以上の処理装置に前記（ｄ）および前記（ｅ）を実行させるのに更に有効である、請求項７に記載のシステム。
【請求項９】
前記実行可能コードは、
（ｆ）前記（ｄ）および前記（ｅ）の実行の後に、前記（ｄ）が前記目標率に等しい前記刈り込み率で実行されるまで、前記（ｂ）の最後の繰り返しでゼロに重み付けするように制約されて刈り込まれた結合を用いて前記ニューラルネットワークを再学習し、
前記ニューラルネットワークの精度が安定な閾値条件を満たすまで、前記（ａ）から前記（ｆ）の各繰り返しで前記目標率を増加しながら、１回以上前記（ａ）から前記（ｆ）を繰り返す、
処理を、前記１つ以上の処理装置に実行させるのに更に有効である、請求項８に記載のシステム。
【請求項１０】
前記実行可能コードは、
前記（ｅ）の後で前記（ｆ）を実行する前に、前記ニューラルネットワークの精度が安定な閾値条件を達成するまで、前記刈り込み率を前記目標率に固定して前記（ｄ）を繰り返す、
処理を、前記１つ以上の処理装置に実行させるのに更に有効である、請求項９に記載のシステム。
【請求項１１】
前記実行可能コードは、
前記ニューラルネットワークの精度が前記（ａ）を実行することによって達成される前記ニューラルネットワークの精度よりも小さく、閾値量の範囲内になるまで、前記（ｂ）および前記（ｃ）を実行することによって、前記（ｄ）を実行する、
処理を、前記１つ以上の処理装置に実行させるのに更に有効である、請求項９に記載のシステム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、畳み込みニューラルネットワークなどのニューラルネットワークの学習を実行するシステムおよび方法に関する。
【背景技術】
【０００２】
畳み込みニューラルネットワーク（ＣＮＮ）を使用する多くのマシンラーニングアプリケーションは、非常に高い計算およびメモリ帯域幅を必要とする。計算負荷を軽減する１つの方法は、係数をゼロに刈り込み、係数がゼロである場合に計算をスキップすることである。様々な既存のソフトウェアおよびハードウェアの最適化技術は、ゼロに評価された係数の利点を利用する。１つの例は、疎畳み込みニューラルネットワーク（Baoyuan Liu、Min Wang1、Hassan Foroosh1、Marshall Tappen、およびMarianna Penksy）およびディープニューラルネットワーク圧縮および効率的推論エンジン（Song Han CVA group、Stanford University）に記載される疎行列乗算技術であり、これらの文献の両方は、その全体が参照により本明細書に組み込まれる。
【０００３】
ＣＮＮは、閾値よりも弱いいくつかの結合荷重をゼロにクランプすることで刈り込むことができる。精度は、刈り込みによって大きく影響される。刈り込まれた結合を維持しながら、刈り込まれたモデル精度を復元するための特別な再学習方法が必要である。刈り込みの従来のアプローチでは、再学習の間、刈り込まれた結合荷重に無効化マスク（disabling mask）が適用される。刈り込みと再学習の複数回の繰り返しは、一般に、精度の更な

る向上に役立てることができる。しかし、各繰り返しに対して、再学習する前に無効化マスクが固定される。

【０００４】

本明細書に開示されたシステムおよび方法は、高いレベルの精度を達成しながらもゼロ係数の数を増加させるためにＣＮＮを刈り込みする改善されたアプローチを提供する。

【図面の簡単な説明】

【０００５】

本発明の利点が容易に理解されるように、添付図面に示される特定の実施形態を参照することによって、上で簡潔に説明された本発明のより詳細な説明が提供されるだろう。これらの図面は、本発明の典型的な実施形態のみを示すものであり、したがって、その範囲を限定するものとみなされるべきではないことを理解されたく、本発明は、添付の図面の使用を通じてさらなる具体性および詳細とともに記載および説明されるだろう。

【図１】本発明の一実施形態に係るニューラルネットワークを刈り込む方法の処理フロー図である。

【図２】本発明の一実施形態に係る図２の方法を繰り返し実行するための方法の処理フロー図である。

【図３】図１の刈り込み方法を示すプロットである。

【図４】本発明の方法を実施するためのコンピューティングデバイスの概略ブロック図である。

【発明を実施するための形態】

【０００６】

本明細書の図面に概して記述および図示されているような本発明の構成要素は、多種多様な異なる構成で配置および設計できることが容易に理解されるであろう。したがって、図面に表された本発明の実施形態の以下のより詳細な説明は、請求されるような本発明の範囲を限定するものではなく、単に本発明による現在考えられる実施形態の特定の例を表すに過ぎない。現在記述されている実施形態は、同様の部材が全体を通じて同様の数字で示されている図面を参照することで最もよく理解されるであろう。

【０００７】

本発明に係る実施形態は、装置、方法、またはコンピュータプログラム製品として具体化され得る。したがって、本発明は、完全にハードウェアの実施形態、完全にソフトウェアの実施形態（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）、または、本明細書ではすべて「モジュール」または「システム」として参照され得るソフトウェアおよびハードウェアの態様を組み合わせた実施形態の形態をとってよい。さらに、本発明は、媒体に埋め込まれたコンピュータ使用可能プログラムコードを有する表現の任意の有形の媒体に具体化されたコンピュータプログラム製品の形態を取ってよい。

【０００８】

非一時的媒体を含む、１以上のコンピュータ使用可能媒体または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。例えば、コンピュータ可読媒体は、ポータブル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）デバイス、読み出し専用メモリ（ＲＯＭ）デバイス、消去可能プログラマブル読み出し専用メモリ（ＥＰＲＯＭまたはフラッシュメモリ）デバイス、ポータブルコンパクトディスク読み取り専用メモリ（ＣＤＲＯＭ）、光記憶装置、および磁気記憶装置の１つ以上を含む。選ばれた実施形態では、コンピュータ可読媒体は、命令実行システム、装置、またはデバイスによって使用される、又は関連するプログラムを、含む、格納する、通信する、伝播する、または転送することができる任意の非一時的媒体を含み得る。

【０００９】

本発明の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（登録商標）、Ｓｍａｌｌｔａｌｋ、またはＣ＋＋などのオブジェクト指向プログラミング言語、および"Ｃ"プログラミング言語または同様のプログラミング言語などの従来の手続き型プログラミング言語を含む、１つ以上のプログラミング言語の任意の組み合わせで記載され得

る。プログラムコードは、全体がスタンドアローンのソフトウェアパッケージとしてコンピュータシステム上で、全体がスタンドアローンのハードウェアユニット上で、一部がコンピュータからある程度の距離離れたリモートコンピュータ上で、または全体がリモートコンピュータ或いはサーバ上で実行され得る。後者のシナリオでは、リモートコンピュータは、ローカルエリアネットワーク（ＬＡＮ）またはワイドエリアネットワーク（ＷＡＮ）を含む任意のタイプのネットワークを介してコンピュータに接続されてもよく、または、（例えば、インターネットサービスプロバイダを使用してインターネットを介して）外部コンピュータに接続されてもよい。

【００１０】

本発明は、本発明の実施形態に係る方法、装置（システム）およびコンピュータプログラム製品のフローチャート図および／またはブロック図を参照して以下に記述される。フローチャート図および／またはブロック図の各ブロック、ならびにフローチャート図および／またはブロック図のブロックの組み合わせは、コンピュータプログラム命令またはコードによって実装できることが理解されよう。これらのコンピュータプログラム命令を、汎用コンピュータ、専用コンピュータ、または他のプログラム可能データ処理装置のプロセッサに提供して、マシンを製造することもでき、こうすることで、その命令が、コンピュータまたは他のプログラム可能データ処理装置のプロセッサを介して実行され、フローチャートおよび／またはブロック図のブロックまたは複数のブロックで指定された機能／動作を実装するための手段を作成する。

【００１１】

これらのコンピュータプログラム命令はまた、コンピュータ可読媒体に格納された命令が、フローチャート並びに／或いはブロック図のブロックまたは複数のブロックで指定された機能／動作を実装する命令手段を含む製造品を製造するように、コンピュータまたは他のプログラム可能データ処理装置に特定の仕方で機能するように指示することができる非一時的コンピュータ可読媒体に格納されてもよい。

【００１２】

コンピュータプログラム命令はまた、コンピュータまたは他のプログラム可能装置上で実行される命令が、フローチャート並びに／或いはブロック図のブロックまたは複数のブロックで指定される機能／動作を実装するためのプロセスを提供するように、コンピュータまたは他のプログラム可能データ処理装置にロードされて、コンピュータまたは他のプログラム可能装置上で一連の動作ステップを実行させて、コンピュータ実装プロセスを生成してもよい。

【００１３】

図１を参照すると、本明細書で開示されるシステムおよび方法は、高精度を達成しつつも、ゼロ係数の数を増加させるために畳み込みニューラルネットワーク（ＣＮＮ）の学習に刈り込みを組み込む改良されたアプローチを提供する。

【００１４】

ＣＮＮの応用例は、２０１６年８月１１日に出願された「ZERO COEFFICIENT SKIPPING CONVOLUTION NEURAL NETWORK ENGINE」と題する米国特許出願第６２／３７３，５１８号、２０１７年８月８日に出願された「ZERO COEFFICIENT SKIPPING CONVOLUTION NEURAL NETWORK ENGINE」と題する米国特許出願第１５／６７１，８２９号、および、２０１７年８月８日に出願された「ZERO COEFFICIENT SKIPPING CONVOLUTION NEURAL NETWORK ENGINE」と題する米国特許出願第１５／６７１，８６０号に記載されており、これらの全ては、その全体が参照により本明細書に組み込まれる。

【００１５】

図示された方法１００は、ゼロ係数の数を増やすためにＣＮＮの結合を刈り込みもするが、一方で、方法１００は、ＣＮＮ、他の種類のニューラルネットワーク、または他の種類のマシンラーニングモデルの学習のために使用されてよい。以下の記載では、ＣＮＮを参照する。しかし、開示された方法は、任意の種類のニューラルネットワークでの使用に

適用可能であると理解されるべきである。

【００１６】

方法１００は、ＣＮＮモデルのフル学習（full training）を実行する（１０２）ことを含み得る。フル学習は、当技術分野で知られているＣＮＮのための任意の学習方法を含み得る。当技術分野で知られているように、学習（１０２）は、ＣＮＮモデルの結合の重みを設定するために、学習データを入力として取り、学習アルゴリズムを適用することを含み得る。フル学習の期間中に、ＣＮＮモデルの任意の結合の重みは、学習データに基づいて所与の入力データに対して正確な出力を生成するように変更され得る。当技術分野で知られているように、学習データは、エントリを含み、各エントリは、１つ以上の入力と１つ以上の所望の出力とを含む。学習アルゴリズムは、１つ以上の入力を受信したときに、モデルが１つ以上の所望の出力を与えるように、ＣＮＮの結合の重みを選択する。ステップ１０２は、例えば、前の繰り返しに対して閾値の割合よりも小さく変化する（例えば、０．０１～２パーセントの間の値で、例えば１パーセントが許容可能な値である）といったように、ＣＮＮモデルの精度が安定するまで学習アルゴリズムを繰り返すことによってなど、１以上のフル再学習の繰り返しを含み得る。

【００１７】

方法１００は、ＣＮＮモデルを刈り込む（１０４）ことを含み得る。これは、ゼロに近い結合の重みをゼロにクランプすることを含み得る。刈り込むステップ１０４は、所定の刈り込み率で実行され得る。刈り込み率は、ゼロに設定する結合の割合である。刈り込み率Ｐは、初期値（例えば５０％から６０％の間の値）に設定され得る。したがって、刈り込みステップ１０４は、ステップ１０２の後で最も低い重みを有するＰパーセントの結合について重みをゼロにクランプすることを含み得る。言い換えると、保持される１－Ｐパーセントの結合は、ゼロにクランプされるＰパーセントの結合のすべての刈り込み前の重みよりも、高い重みをそれぞれが有することになる。

【００１８】

方法１００は、刈り込みの後にＣＮＮモデルのフル再学習を実行する（１０６）ことを含み得る。フル再学習は、刈り込みステップに従ってＣＮＮモデルを制約しない。しかしながら、ステップ１０６の開始時のＣＮＮモデルは、ステップ１０４由来の刈り込まれたモデルである。これにより、刈り込みされた状態にとどまる制約なしに、刈り込みされたモデルに向かってバイアスが与えられる。したがって、刈り込みされた結合の重みは、再学習（１０６）の期間中に増加することが許容されることになる。

【００１９】

ステップ１０６は、ＣＮＮモデルの精度が復元されるべきと分かる（１０８）まで、すなわち、ステップ１０２に続くＣＮＮモデルの精度の或る閾値パーセンテージ（Δ１）内の精度を達成された状態になるまで、繰り返され得る。

【００２０】

方法１００は、刈り込み率Ｐが目標の刈り込み率P_Tに等しいかどうかを評価することを含み得る。目標の刈り込み率の例は、５０％から９０％の範囲であってよい。しかしながら、何が妥当な目標の刈り込み率であるかは、用途に大きく依存する。そうでない場合、刈り込み率Ｐは、Ｐ＋ΔＰに増加されるか、又は、P_Tに設定され得（いずれか小さい方）、ここでΔＰはインクリメント量であるとする。インクリメントΔＰは、定数であってよく、又は、非線形関数に従ってステップ１１２の各繰り返しごとに変化してもよい。方法１００は、その後、ステップ１１２で設定された値に等しいＰで、ＣＮＮモデルを再び刈り込むことによって、ステップ１０４に進み得る。次いで、方法は、上記のステップ１０４から進み得る。

【００２１】

ＣＮＮモデルが、いったん目標刈り込み率P_Tに刈り込み１０４されており、閾値精度内に学習１０６されると、方法は、ステップ１１４、１１６、および１１８に進み得、ここで、ＣＮＮモデルは、その精度が安定した状態にあると分かる（１１６）まで、１回以上、再学習（１１４）され、および刈り込まれる（１１８）。安定性は、ステップ１１４

の１回の繰り返し後の精度が、前の繰り返しから閾値差（例えば０．０１％～２％の間の値、例えば１％の値が許容されることが多い）内である場合に、達成されたと判定され得る。

【００２２】

方法１００は、その後、ＣＮＮモデルのマスクされた再学習を実行する（１２０）ことを含み得る。マスクされた再学習では、非ゼロ結合のみが値を変更することを許可される。換言すれば、ステップ１２０では、選択された結合、すなわち、最後の刈り込み（ステップ１０４または１１８）の期間中に刈り込まれなかった結合のみが変更可能である。

【００２３】

マスクされた再学習１２０は、当技術分野で知られている任意のマスクされたＣＮＮ再学習アプローチによる無効化マスクを用いた学習を含み得る。ステップ１２０は、ＣＮＮモデルの精度が安定な状態にあると分かる（１２２）まで、１回以上繰り返され得る。安定性は、ステップ１１６と同じ方法で決定され得る。

【００２４】

その後、方法１００は終了し得、それからＣＮＮモデルは、生産データを処理するために使用され得る。

【００２５】

図２を参照すると、いくつかの実施形態では、方法１００（本明細書ではＰＦＰ－Ｓ：刈り込み（prune）、フル再学習（full retrain）、刈り込み（prune）、および選択された結合の再学習（selected connection retraining）として参照される）は、図示された方法２００の内容の範囲内で繰り返され得る。

【００２６】

方法２００は、ＣＮＮモデルが安定化された状態にあることが分かる（２０４）まで（ステップ１１６参照）、ＰＦＰ－Ｓ方法１００を実行する（２０２）ことを含み得る。ＰＦＰ－Ｓ方法１００が繰り返される前に、刈り込み目標P_Tは、例えば、固定のまたは可変の減少量分など、増加され得る。刈り込み率Ｐの初期値は、ＰＦＰ－Ｓ方法１００の以前の実行由来のP_Tの値に、またはそれよりも小さい値に設定され得る。

【００２７】

その後、方法２００の繰り返し由来のＣＮＮモデルが使用され得る。これは、方法２００の最後の繰り返しによって生成されるＣＮＮモデルであり得る。刈り込み目標P_Tの増加が、許容できない量でＣＮＮモデルの精度を低下させる場合、ＣＮＮモデルの最新バージョンではなく、ＣＮＮモデルの前のバージョンが、ステップ２０８で復元され得、生産データを処理するために使用され得る。

【００２８】

図３は、方法１００がどのように動作し得るかの例を示す。横軸は刈り込み率Ｐを表す。縦軸はＣＮＮモデルの精度を表す。点３００は、ピーク精度（“ピーク％”）を有する方法１００の初期学習ステップ１０２に続くＣＮＮモデルの状態を表す。図３に示すように、ＣＮＮモデルは刈り込みされ、その結果、刈り込み率が増加し、また、割合（Ｐｅａｋ－Δ０）分精度が減少する。ＣＮＮモデルは、その後、各学習の後で精度がＰｅａｋ－Δ１に達するように、繰り返し学習され（１０６）、および刈り込まれる（１０４）。学習ステップ１０６の期間中、刈り込みされた結合が非ゼロ値に変更されることが許されるので、刈り込み率は減少することに留意されたい。

【００２９】

点３０２は、Ｐが刈り込み目標P_Tに到達した後のＣＮＮモデルの状態を示す。その後、ステップ１１８の最後の繰り返しの後の点３０４によって示される状態にＣＮＮモデルがなる点において、ＣＮＮモデルの精度が安定すると分かる（１１６）まで、再学習（１１４）及び刈り込み（１１８）が実行され得る。その後、ＣＮＮモデルはマスクを用いて再学習され（１２０）、刈り込み目標P_Tにおける刈り込み率を維持しながらモデルの精度を向上させる。その後、ＣＮＮモデルは点３０６になる。その後、方法１００は、方法２００に従って、異なる刈り込み目標P_Tで繰り返されてもよく、または、ＣＮＮモデル

が生産データを処理するために使用されてもよい。
【００３０】
図４は、例示的なコンピューティングデバイス４００を示すブロック図である。コンピューティングデバイス４００は、本明細書で論じられるものなどの様々な手順を実行するために使用され得る。コンピューティングデバイス４００は、デスクトップコンピュータ、ノートブックコンピュータ、サーバコンピュータ、ハンドヘルドコンピュータ、タブレットコンピュータなどの多種多様なコンピューティングデバイスのいずれかとすることができる。
【００３１】
コンピューティングデバイス４００は、１つ以上のプロセッサ４０２、１つ以上のメモリデバイス４０４、１つ以上のインタフェース４０６、１つ以上の大容量記憶デバイス４０８、１つ以上の入力／出力（Ｉ／Ｏ）デバイス４１０、および表示装置４３０を含み、これらのすべてがバス４１２に接続される。１つの（または複数の）プロセッサ４０２は、１つ以上のプロセッサまたはコントローラを含み、これらは１つの（または複数の）メモリデバイス４０４および／または１つの（または複数の）大容量記憶デバイス４０８に記憶された命令を実行する。１つの（または複数の）プロセッサ４０２はまた、キャッシュメモリなどの様々な種類のコンピュータ可読媒体を含み得る。
【００３２】
１つの（または複数の）メモリデバイス４０４は、揮発性メモリ（例えば、ランダムアクセスメモリ（ＲＡＭ）４１４）および／または不揮発性メモリ（例えば、読み出し専用メモリ（ＲＯＭ）４１６）などの様々なコンピュータ可読媒体を含む。メモリデバイス４０４はまた、フラッシュメモリなどの書き換え可能なＲＯＭを含み得る。
【００３３】
１つの（または複数の）大容量記憶デバイス４０８は、磁気テープ、磁気ディスク、光ディスク、ソリッドステートメモリ（例えば、フラッシュメモリ）などの様々なコンピュータ可読媒体を含む。図４に示すように、具体的な大容量記憶装置はハードディスクドライブ４２４である。１つの（または複数の）大容量記憶デバイス４０８には様々なドライブが含まれ、様々なコンピュータ可読媒体からの読み取りおよび／または書き込みを可能にしてもよい。１つの（または複数の）大容量記憶デバイス４０８は、リムーバブルメディア４２６および／または非リムーバブルメディアを含む。
【００３４】
１つの（または複数の）Ｉ／Ｏデバイス４１０は、データおよび／または他の情報がコンピューティングデバイス４００に入力されるか、またはコンピューティングデバイス４００から取り出されることを可能にする様々なデバイスを含む。例示的な１つの（または複数の）Ｉ／Ｏデバイス４１０は、カーソル制御デバイス、キーボード、キーパッド、マイク、モニタまたは他の表示装置、スピーカ、プリンタ、ネットワークインタフェースカード、モデム、レンズ、およびＣＣＤまたは他の画像キャプチャ装置などを含む。
【００３５】
表示装置４３０は、コンピューティングデバイス４００の１人または複数のユーザに情報を表示することができる任意の種類のデバイスを含む。表示装置４３０の例は、モニタ、表示端末、およびビデオ投影装置などを含む。
【００３６】
グラフィックスプロセッシングユニット（ＧＰＵ）４３２は、１つの（または複数の）プロセッサ４０２および／または表示装置４３０に接続されてよい。ＧＰＵは、コンピュータ生成画像をレンダリングし、他のグラフィカル処理を実行するように動作可能であってよい。ＧＰＵは、１つの（または複数の）プロセッサ４０２などの汎用プロセッサの機能の一部または全部を含み得る。ＧＰＵは、グラフィックス処理に特有の追加の機能を含んでもよい。ＧＰＵは、座標変換、シェーディング、テクスチャリング、ラスタライズ、およびコンピュータ生成画像のレンダリングに役立つ他の機能に関連するハードコードされたグラフィックス機能、および／またはハードワイヤードのグラフィックス機能を含み

得る。

【００３７】

１つの（または複数の）インタフェース４０６は、コンピューティングデバイス４００が他のシステム、デバイス、またはコンピューティング環境と情報をやり取りすることを可能にする様々なインタフェースを含む。例示的な１つの（または複数の）インタフェース４０６は、ローカルエリアネットワーク（ＬＡＮ）、ワイドエリアネットワーク（ＷＡＮ）、無線ネットワーク、およびインターネットへのインタフェースなどの任意の数の様々なネットワークインタフェース４２０を含む。他の１つ（または複数の）インタフェースは、ユーザインタフェース４１８と周辺デバイスインタフェース４２２とを含む。１つの（または複数の）インタフェース４０６はまた、１つ以上のユーザインタフェース素子４１８を含んでもよい。１つの（または複数の）インタフェース４０６はまた、例えば、プリンタ、ポインティングデバイス（マウス、トラックパッドなど）、およびキーボードなどのインタフェースのような、１つ以上の周辺インタフェースを含んでもよい。

【００３８】

バス４１２は、１つの（または複数の）プロセッサ４０２、１つの（または複数の）メモリデバイス４０４、１つの（または複数の）インタフェース４０６、１つの（または複数の）大容量記憶デバイス４０８、および１つの（または複数の）Ｉ／Ｏデバイス４１０が相互に通信することを可能にし、バス４１２に接続された他のデバイスまたは構成要素との通信も可能にする。バス４１２は、システムバス、ＰＣＩバス、ＩＥＥＥ１３９４バス、およびＵＳＢバスなどのいくつかの種類のバス構造の１つ以上を表す。

【００３９】

説明のために、プログラムおよび他の実行可能なプログラム構成要素は、本明細書では個別のブロックとして示されているが、そのようなプログラムおよび構成要素は、様々な時間で、コンピューティングデバイス４００の異なる記憶構成要素内に存在し得、１つの（または複数の）プロセッサ４０２によって実行され得ることが理解される。代わりに、本明細書に記載のシステムおよび手順は、ハードウェアで、またはハードウェア、ソフトウェア、および／またはファームウェアの組み合わせで実装されることができる。例えば、１つ以上の特定用途向け集積回路（ＡＳＩＣ）は、本明細書で記載されるシステムおよび手順の１つ以上を実行するようにプログラムすることができる。

【００４０】

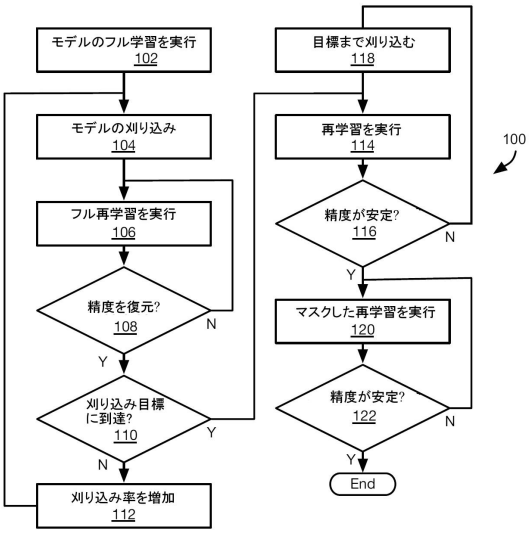
本発明は、その趣旨または本質的な特徴から逸脱することなく、他の特定の形態で実施することができる。記載された実施形態は、すべての点において、例示的なものであり、限定的なものではないとみなされるべきである。したがって、本発明の範囲は、前述の説明によってではなく、添付の特許請求の範囲によって示される。特許請求の範囲と均等の意味および範囲内に入るすべての変更は、その範囲内に含まれるべきである。

【００４１】

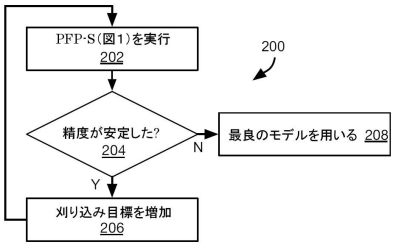
請求されているものは以下に列挙される。

【図面】

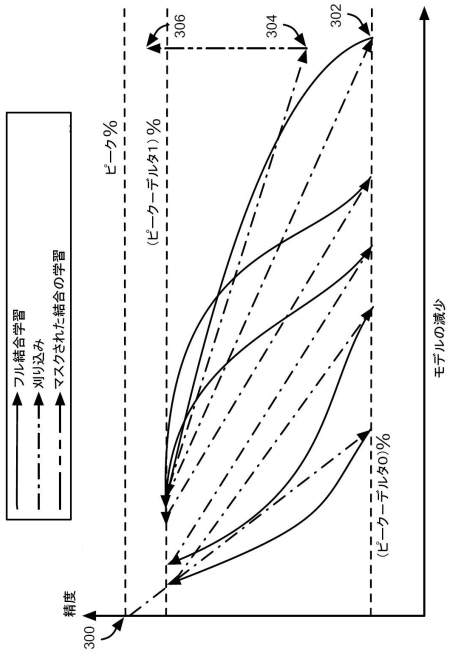
【図１】


モデルのフル学習を実行
102
モデルの刈り込み
104
フル再学習を実行
106
精度を復元?
108
N
Y
刈り込み目標
に到達?
110
Y
N
刈り込み率を増加
112
目標まで刈り込む
118
再学習を実行
114
精度が安定?
116
N
Y
マスクした再学習を実行
120
精度が安定?
122
N
Y
End
100

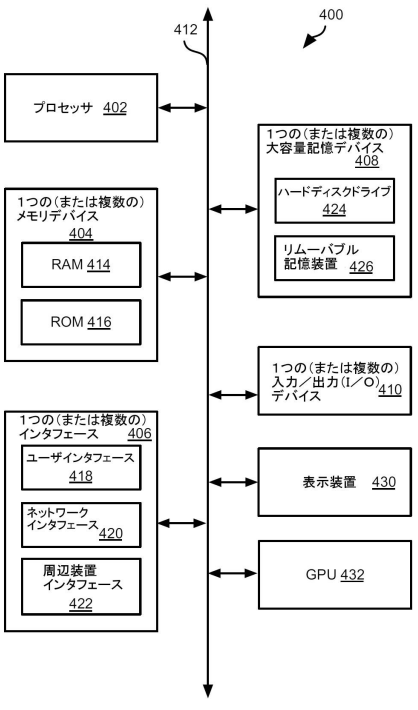
【図２】


PFP-S(図1)を実行
202
200
精度が安定した?
204
N
最良のモデルを用いる 208
Y
刈り込み目標を増加
206

【図３】


フル結合学習
刈り込み
マスクされた結合の学習
ピーク%
(ピークデルタ1)%
306
304
302
モデルの減少
精度
300
(ピークデルタ0)%

【図４】


400
412
プロセッサ 402
1つの(または複数の)
メモリデバイス
404
RAM 414
ROM 416
1つの(または複数の)
インタフェース 406
ユーザインタフェース
418
ネットワーク
インタフェース 420
周辺装置
インタフェース
422
1つの(または複数の)
大容量記憶デバイス
408
ハードディスクドライブ
424
リムーバブル
記憶装置 426
1つの(または複数の)
入力/出力(I/O)
デバイス 410
表示装置 430
GPU 432

フロントページの続き

５００２，サン ノゼ，ゴールド ストリート ２１５０，スイート ２００
(72)発明者 シャン‐ハン リン
アメリカ合衆国，カリフォルニア州 ９５００２，サン ノゼ，ゴールド ストリート ２１５０，スイート ２００

審査官 石川 亮

(56)参考文献
特開２０１８－１５２０００（ＪＰ，Ａ）
米国特許出願公開第２０１６／０３０７０９８（ＵＳ，Ａ１）
特開２０１１－０５４２００（ＪＰ，Ａ）
HU Hengyuan et al.，Network Trimming: A Data-Driven Neuron Pruning Approach towards Efficient Deep Architectures，arXiv.org [online]，2016年，pp.1-9，[2022年7月28日検索]，インターネット〈URL: https://arxiv.org/abs/1607.03250〉

(58)調査した分野 (Int.Cl.，ＤＢ名)
Ｇ０６Ｎ ３／０２－３／１０