(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6859332号
(P6859332)
(45)発行日　令和3年4月14日(2021.4.14)　(24)登録日　令和3年3月29日(2021.3.29)

(51)Int.Cl.　F I
G06N 3/08 (2006.01)　G06N 3/08 140

請求項の数 10 (全 23 頁)

(21)出願番号	特願2018-515936(P2018-515936)
(86)(22)出願日	平成28年9月7日(2016.9.7)
(65)公表番号	特表2018-533138(P2018-533138A)
(43)公表日	平成30年11月8日(2018.11.8)
(86)国際出願番号	PCT/US2016/050539
(87)国際公開番号	WO2017/058479
(87)国際公開日	平成29年4月6日(2017.4.6)
審査請求日	令和1年8月14日(2019.8.14)
(31)優先権主張番号	62/234,559
(32)優先日	平成27年9月29日(2015.9.29)
(33)優先権主張国・地域又は機関	米国(US)
(31)優先権主張番号	15/081,780
(32)優先日	平成28年3月25日(2016.3.25)
(33)優先権主張国・地域又は機関	米国(US)

(73)特許権者　595020643
クゥアルコム・インコーポレイテッド
ＱＵＡＬＣＯＭＭ　ＩＮＣＯＲＰＯＲＡＴＥＤ
アメリカ合衆国、カリフォルニア州　９２１２１－１７１４、サン・ディエゴ、モアハウス・ドライブ　５７７５
(74)代理人　100108855
弁理士　蔵田　昌俊
(74)代理人　100109830
弁理士　福原　淑弘
(74)代理人　100158805
弁理士　井関　守三
(74)代理人　100112807
弁理士　岡田　貴志

最終頁に続く

(54)【発明の名称】選択的バックプロパゲーション

(57)【特許請求の範囲】
【請求項１】
機械学習モデルのためのクラス間のトレーニングデータの平衡を変更する方法であって、
最も少数のメンバーをもつクラスの例の数と現在のクラスの例の数との比からファクタを決定することと、
前記現在のクラスで前記機械学習モデルをトレーニングする間、前記決定されたファクタに基づいて、前記現在のクラスに関連する、バックプロパゲーションプロセスの勾配を変更することと、ここにおいて、前記変更することが、前記最も少数のメンバーをもつ前記クラスの前記例のサンプリングに少なくとも部分的に基づいて前記勾配を選択的に適用することを備え、サンプリング確率は、前記決定されたファクタに基づいて決定される、を備える、方法。
【請求項２】
前記決定されたファクタに基づいて前記変更することが、前記決定されたファクタで前記勾配をスケーリングすることを備える、請求項１に記載の方法。
【請求項３】
前記最も少数のメンバーをもつ前記クラスの前記サンプリングが、各トレーニングエポックから固定数の例を選択することによって行われる、請求項１に記載の方法。
【請求項４】
前記サンプリングが、トレーニングエポック中の例の交換なしに行われる、請求項１に

記載の方法。
【請求項５】
機械学習モデルのためのクラス間のトレーニングデータの平衡を変更するための装置であって、
最も少数のメンバーをもつクラスの例の数と現在のクラスの例の数との比からファクタを決定するための手段と、
前記現在のクラスで前記機械学習モデルをトレーニングする間、前記決定されたファクタに基づいて、前記現在のクラスに関連する、バックプロパゲーションプロセスの勾配を変更するための手段と、ここにおいて、前記変更するための手段が、前記最も少数のメンバーをもつ前記クラスの前記例のサンプリングに少なくとも部分的に基づいて前記勾配を選択的に適用するための手段を備え、サンプリング確率は、前記決定されたファクタに基づいて決定される、を備える、装置。
【請求項６】
前記決定されたファクタに基づいて前記変更するための手段が、前記決定されたファクタで前記勾配をスケーリングするための手段を備える、請求項５に記載の装置。
【請求項７】
前記クラスの前記サンプリングが、各トレーニングエポックから固定数の例を選択することによって行われる、請求項５に記載の装置。
【請求項８】
前記サンプリングが、トレーニングエポック中の例の交換なしに行われる、請求項５に記載の装置。
【請求項９】
メモリをさらに備え、
前記決定するための手段および前記変更するための手段が、前記メモリに結合された少なくとも１つのプロセッサを備える、請求項５に記載の装置。
【請求項１０】
機械学習モデルのためのクラス間のトレーニングデータの平衡を変更するための非一時的コンピュータ可読媒体であって、前記非一時的コンピュータ可読媒体がそれに記録されたプログラムコードを有し、前記プログラムコードが、実行されると請求項１～４のいずれか一項に記載の方法を実施する、非一時的コンピュータ可読媒体。
【発明の詳細な説明】
【技術分野】
【０００１】
関連出願の相互参照
[0001]本出願は、その開示全体が参照により本明細書に明確に組み込まれる、２０１５年９月２９日に出願された、「SELECTIVE BACKPROPAGATION」と題する米国仮特許出願第６２／２３４，５５９号の利益を主張する。
【背景技術】
【０００２】
[0002]本開示のいくつかの態様は、一般に機械学習に関し、より詳細には、機械学習モデルのためのクラス間のトレーニングデータの平衡を変更することに関する。
【０００３】
[0003]人工ニューロン（たとえば、ニューロンモデル）の相互結合されたグループを備え得る人工ニューラルネットワークは、計算デバイスであるか、または計算デバイスによって実施されるべき方法を表す。
【０００４】
[0004]畳み込みニューラルネットワークは、フィードフォワード人工ニューラルネットワークのタイプである。畳み込みニューラルネットワークは、各々が受容野を有し、入力空間を集合的にタイリングするニューロンの集合を含み得る。畳み込みニューラルネットワーク（ＣＮＮ：convolutional neural network）は多数の適用例を有する。特に、ＣＮ

Nは、パターン認識および分類の領域内で広く使用されている。

【０００５】

[0005]深層信念ネットワークおよび深層畳み込みネットワーク（deep convolutional network）など、深層学習アーキテクチャは、層状（layered）ニューラルネットワークアーキテクチャであり、ニューロンの第１の層の出力はニューロンの第２の層への入力になり、ニューロンの第２の層の出力はニューロンの第３の層になり、入力し、以下同様である。深層ニューラルネットワークは、特徴の階層（hierarchy）を認識するようにトレーニングされ得、したがって、それらはオブジェクト認識適用例においてますます使用されている。畳み込みニューラルネットワークのように、これらの深層学習アーキテクチャにおける計算は、１つまたは複数の計算チェーンにおいて構成され得る処理ノードの集団にわたって分散され得る。これらの多層アーキテクチャは、一度に１つの層をトレーニングされ得、バックプロパゲーション（back propagation）を使用して微調整され得る。

【０００６】

[0006]他のモデルも、オブジェクト認識のために利用可能である。たとえば、サポートベクターマシン（ＳＶＭ）は、分類のために適用され得る学習ツールである。サポートベクターマシンは、データをカテゴリー分類する分離超平面（separating hyperplane）（たとえば、決定境界（decision boundary））を含む。超平面は、教師あり学習によって定義される。所望の超平面は、トレーニングデータのマージンを増加させる。言い換えれば、超平面は、トレーニング例との最大の最小距離を有するべきである。

【０００７】

[0007]これらのソリューションは、いくつかの分類ベンチマーク上で優れた結果を達成するが、それらの計算複雑さは極めて高いことがある。さらに、モデルのトレーニングが難しいことがある。

【発明の概要】

【０００８】

[0008]一態様では、機械学習モデルのためのクラス間のトレーニングデータの平衡を変更する方法が開示される。本方法は、最も少数のメンバーをもつクラスの例の数と現在のクラスの例の数との比に基づいて、モデルをトレーニングする間、バックプロパゲーションプロセスの勾配を変更することを含む。

【０００９】

[0009]別の態様は、機械学習モデルのためのクラス間のトレーニングデータの平衡を変更するための装置を開示する。本装置は、最も少数のメンバーをもつクラスの例の数と現在のクラスの例の数との比に基づいて、勾配を変更するためのファクタを決定するための手段を含む。本装置はまた、決定されたファクタに基づいて、現在のクラスに関連する勾配を変更するための手段を含む。

【００１０】

[0010]別の態様は、メモリと、メモリに結合された少なくとも１つのプロセッサとを有するワイヤレス通信を開示する。（１つまたは複数の）プロセッサは、最も少数のメンバーをもつクラスの例の数と現在のクラスの例の数との比に基づいて、モデルをトレーニングする間、バックプロパゲーションプロセスの勾配を変更するように構成される。

【００１１】

[0011]別の態様は、それに記録された非一時的プログラムコードを有する非一時的コンピュータ可読媒体であって、（１つまたは複数の）プロセッサによって実行されたとき、（１つまたは複数の）プロセッサに、最も少数のメンバーをもつクラスの例の数と現在のクラスの例の数との比に少なくとも部分的に基づいて、モデルをトレーニングする間、バックプロパゲーションプロセスの勾配を変更する動作を実行させる、非一時的コンピュータ可読媒体を開示する。

【００１２】

[0012]本開示の追加の特徴および利点が、以下で説明される。本開示は、本開示の同じ目的を実行するための他の構造を変更または設計するための基礎として容易に利用され得

ることを、当業者は諒解されたい。また、そのような等価な構成が、添付の特許請求の範囲に記載の本開示の教示から逸脱しないことを、当業者は了解されたい。さらなる目的および利点とともに、本開示の編成と動作の方法の両方に関して、本開示を特徴づけると考えられる新規の特徴は、添付の図に関連して以下の説明を検討するとより良く理解されよう。ただし、図の各々は、例示および説明のみの目的で与えられたものであり、本開示の限界を定めるものではないことを明確に理解されたい。

【図面の簡単な説明】

【００１３】

[0013]本開示の特徴、特性、および利点は、全体を通じて同様の参照符号が同様のものを指す図面とともに、以下に記載される発明を実施するための形態を読めばより明らかになろう。

【図１】[0014]本開示のいくつかの態様による、汎用プロセッサを含むシステムオンチップ（ＳＯＣ）を使用してニューラルネットワークを設計する例示的なインプリメンテーションを示す図。

【図２】[0015]本開示の態様による、システムの例示的なインプリメンテーションを示す図。

【図３Ａ】[0016]本開示の態様による、ニューラルネットワークを示す図。

【図３Ｂ】[0017]本開示の態様による、例示的な深層畳み込みネットワーク（ＤＣＮ）を示すブロック図。

【図４】[0018]本開示の態様による、人工知能（ＡＩ）機能をモジュール化し得る例示的なソフトウェアアーキテクチャを示すブロック図。

【図５】[0019]本開示の態様による、スマートフォン上のＡＩアプリケーションのランタイム動作を示すブロック図。

【図６】[0020]本開示の態様による、トレーニングデータを平衡させるための方法を示す図。

【図７】[0021]本開示の態様による、トレーニングデータを平衡させるための全体的例を示す図。

【図８】[0022]本開示の態様による、トレーニングデータを平衡させるための方法を示す図。

【発明を実施するための形態】

【００１４】

[0023]添付の図面に関して以下に記載される発明を実施するための形態は、様々な構成を説明するものであり、本明細書で説明される概念が実施され得る構成のみを表すものではない。発明を実施するための形態は、様々な概念の完全な理解を与えるための具体的な詳細を含む。ただし、これらの概念はこれらの具体的な詳細なしに実施され得ることが当業者には明らかであろう。いくつかの事例では、そのような概念を不明瞭にしないように、よく知られている構造および構成要素がブロック図の形式で示される。

【００１５】

[0024]これらの教示に基づいて、本開示の範囲は、本開示の他の態様とは無関係にインプリメントされるにせよ、本開示の他の態様と組み合わせてインプリメントされるにせよ、本開示のいかなる態様をもカバーするものであることを、当業者なら諒解されたい。たとえば、記載された態様をいくつ使用しても、装置はインプリメントされ得るか、または方法は実施され得る。さらに、本開示の範囲は、記載された本開示の様々な態様に加えてまたはそれらの態様以外に、他の構造、機能、または構造および機能を使用して実施されるそのような装置または方法をカバーするものとする。開示される本開示のいずれの態様も、請求項の１つまたは複数の要素によって実施され得ることを理解されたい。

【００１６】

[0025]「例示的」という単語は、本明細書では「例、事例、または例示の働きをすること」を意味するために使用される。「例示的」として本明細書で説明されるいかなる態様も、必ずしも他の態様よりも好適または有利であると解釈されるべきであるとは限らない

。

【００１７】

[0026]本明細書では特定の態様が説明されるが、これらの態様の多くの変形および置換は本開示の範囲内に入る。好適な態様のいくつかの利益および利点が説明されるが、本開示の範囲は特定の利益、使用、または目的に限定されるものではない。むしろ、本開示の態様は、様々な技術、システム構成、ネットワーク、およびプロトコルに広く適用可能であるものとし、それらのいくつかが、例として、図および好適な態様についての以下の説明において示される。発明を実施するための形態および図面は、本開示を限定するものではなく説明するものにすぎず、本開示の範囲は添付の特許請求の範囲およびそれの均等物によって定義される。

選択的バックプロパゲーション

[0027]本開示の態様は、機械学習モデルにおいてクラス間のトレーニングデータの平衡を変更することを対象とする。特に、入力段においてトレーニングデータを操作し、各クラスについての例の数を調節するのではなく、本開示の態様は、勾配段における調節を対象とする。本開示の様々な態様では、データセット中のクラス例頻度に基づいて勾配を調節するかまたは選択的に適用するために、コスト関数を変更するために、選択的バックプロパゲーションが利用される。特に、勾配は、各クラスについての例の実際のまたは予想される頻度に基づいて調節され得る。

【００１８】

[0028]図１に、本開示のいくつかの態様による、汎用プロセッサ（ＣＰＵ）またはマルチコア汎用プロセッサ（ＣＰＵ）１０２など、少なくとも１つのプロセッサを含み得るシステムオンチップ（ＳＯＣ）１００を使用する、上述の選択的バックプロパゲーションの例示的なインプリメンテーションを示す。変数（たとえば、ニューラル信号およびシナプス荷重）、計算デバイスに関連するシステムパラメータ（たとえば、重みをもつニューラルネットワーク）、遅延、周波数ビン情報、およびタスク情報が、ニューラル処理ユニット（ＮＰＵ）１０８に関連するメモリブロックに記憶されるか、ＣＰＵ１０２に関連するメモリブロックに記憶されるか、グラフィックス処理ユニット（ＧＰＵ）１０４に関連するメモリブロックに記憶されるか、デジタル信号プロセッサ（ＤＳＰ）１０６に関連するメモリブロックに記憶されるか、専用メモリブロック１１８に記憶され得るか、または複数のブロックにわたって分散され得る。汎用プロセッサ１０２において実行される命令が、ＣＰＵ１０２に関連するプログラムメモリからロードされ得るか、または専用メモリブロック１１８からロードされ得る。

【００１９】

[0029]ＳＯＣ１００はまた、ＧＰＵ１０４、ＤＳＰ１０６など、特定の機能に適合された追加の処理ブロックと、第４世代ロングタームエボリューション（４Ｇ　ＬＴＥ（登録商標））接続性、無認可Ｗｉ－Ｆｉ（登録商標）接続性、ＵＳＢ接続性、Ｂｌｕｅｔｏｏｔｈ（登録商標）接続性などを含み得る接続性ブロック１１０と、たとえば、ジェスチャーを検出および認識し得るマルチメディアプロセッサ１１２とを含み得る。一つのインプリメンテーションでは、ＮＰＵは、ＣＰＵ、ＤＳＰ、および／またはＧＰＵにおいてインプリメントされる。ＳＯＣ１００はまた、センサープロセッサ１１４、画像信号プロセッサ（ＩＳＰ）、および／または全地球測位システムを含み得るナビゲーション１２０を含み得る。

【００２０】

[0030]ＳＯＣ１００はＡＲＭ命令セットに基づき得る。本開示の別の態様では、汎用プロセッサ１０２にロードされる命令は、機械学習モデルをトレーニングする間、バックプロパゲーションプロセスの勾配を変更するためのコードを備え得る。変更することは、最も少数のメンバーをもつクラスの例の数と現在のクラスの例の数との比に基づく。変更することは、現在のクラスに関連する勾配に適用される。

【００２１】

[0031]図２に、本開示のいくつかの態様による、システム２００の例示的なインプリメ

ンテーションを示す。図２に示されているように、システム２００は、本明細書で説明される方法の様々な動作を実施し得る複数のローカル処理ユニット２０２を有し得る。各ローカル処理ユニット２０２は、ローカル状態メモリ２０４と、ニューラルネットワークのパラメータを記憶し得るローカルパラメータメモリ２０６とを備え得る。さらに、ローカル処理ユニット２０２は、ローカルモデルプログラムを記憶するためのローカル（ニューロン）モデルプログラム（ＬＭＰ）メモリ２０８と、ローカル学習プログラムを記憶するためのローカル学習プログラム（ＬＬＰ）メモリ２１０と、ローカル接続メモリ２１２とを有し得る。さらに、図２に示されているように、各ローカル処理ユニット２０２は、ローカル処理ユニットのローカルメモリのための構成を与えるための構成プロセッサユニット２１４、およびローカル処理ユニット２０２間のルーティングを与えるルーティング接続処理ユニット２１６とインターフェースし得る。

【００２２】

[0032]深層学習アーキテクチャは、各層において連続的により高い抽象レベルで入力を表現するように学習し、それにより、入力データの有用な特徴表現を蓄積することによって、オブジェクト認識タスクを実施し得る。このようにして、深層学習は、旧来の機械学習の主要なボトルネックに対処する。深層学習の出現より前に、オブジェクト認識問題に対する機械学習手法は、場合によっては浅い分類器（shallow classifier）と組み合わせて、人的に設計された特徴に大きく依拠していることがある。浅い分類器は、たとえば、入力がどのクラスに属するかを予測するために、特徴ベクトル成分の重み付き和がしきい値と比較され得る２クラス線形分類器であり得る。人的に設計された特徴は、領域の専門知識をもつ技術者によって特定の問題領域に適合されたテンプレートまたはカーネルであり得る。対照的に、深層学習アーキテクチャは、人間の技術者が設計し得るものと同様である特徴を表現するように学習するが、トレーニングを通してそれを行い得る。さらに、深層ネットワークは、人間が考慮していないことがある新しいタイプの特徴を表現し、認識するように学習し得る。

【００２３】

[0033]深層学習アーキテクチャは特徴の階層を学習し得る。たとえば、視覚データが提示された場合、第１の層は、エッジなど、入力ストリーム中の比較的単純な特徴を認識するように学習し得る。別の例では、聴覚データが提示された場合、第１の層は、特定の周波数におけるスペクトル電力を認識するように学習し得る。第１の層の出力を入力として取る第２の層は、視覚データの場合の単純な形状、または聴覚データの場合の音の組合せなど、特徴の組合せを認識するように学習し得る。たとえば、上位層は、視覚データ中の複雑な形状、または聴覚データ中の単語を表現するように学習し得る。さらに上位の層は、共通の視覚オブジェクトまたは発話フレーズを認識するように学習し得る。

【００２４】

[0034]深層学習アーキテクチャは、自然階層構造を有する問題に適用されたとき、特にうまく機能し得る。たとえば、原動機付き車両の分類は、ホイール、フロントガラス、および他の特徴を認識するための第１の学習から恩恵を受け得る。これらの特徴は、車、トラック、および飛行機を認識するために、異なる方法で、上位層において組み合わせられ得る。

【００２５】

[0035]ニューラルネットワークは、様々な結合性パターンを用いて設計され得る。フィードフォワードネットワークでは、情報が下位層から上位層に受け渡され、所与の層における各ニューロンは、上位層におけるニューロンに通信する。上記で説明されたように、フィードフォワードネットワークの連続する層において、階層表現が蓄積され得る。ニューラルネットワークはまた、リカレントまたは（トップダウンとも呼ばれる）フィードバック結合を有し得る。リカレント結合では、所与の層におけるニューロンからの出力は、同じ層における別のニューロンに通信され得る。カレントアーキテクチャは、ニューラルネットワークに順次配信される入力データチャンクのうちの２つ以上にわたるパターンを認識するのに役立ち得る。所与の層におけるニューロンから下位層におけるニューロンへ

の結合は、フィードバック（またはトップダウン）結合と呼ばれる。高レベルの概念の認識が、入力の特定の低レベルの特徴を弁別することを助け得るとき、多くのフィードバック結合をもつネットワークが役立ち得る。

【００２６】

[0036]図３Ａを参照すると、ニューラルネットワークの層間の結合は全結合３０２または局所結合３０４であり得る。全結合ネットワーク３０２では、第１の層におけるニューロンは、第２の層における各ニューロンが第１の層におけるあらゆるニューロンから入力を受信するように、それの出力を第２の層におけるあらゆるニューロンに通信し得る。代替的に、局所結合ネットワーク３０４では、第１の層におけるニューロンは、第２の層における限られた数のニューロンに結合され得る。畳み込みネットワーク３０６は、局所結合であり得、第２の層における各ニューロンのための入力に関連する結合強度が共有されるようにさらに構成される（たとえば、３０８）。より一般的には、ネットワークの局所結合層は、層における各ニューロンが同じまたは同様の結合性パターンを有するように構成されるが、異なる値を有し得る結合強度で構成され得る（たとえば、３１０、３１２、３１４、および３１６）。局所結合の結合性パターンは、所与の領域中の上位層ニューロンが、ネットワークへの総入力のうちの制限された部分のプロパティにトレーニングを通して調整された入力を受信し得るので、上位層において空間的に別個の受容野を生じ得る。

【００２７】

[0037]局所結合ニューラルネットワークは、入力の空間ロケーションが有意味である問題に好適であり得る。たとえば、車載カメラからの視覚特徴を認識するように設計されたネットワーク３００は、画像の下側部分対上側部分とのそれらの関連付けに依存して、異なるプロパティをもつ上位層ニューロンを発達させ得る。画像の下側部分に関連するニューロンは、たとえば、車線区分線を認識するように学習し得るが、画像の上側部分に関連するニューロンは、交通信号、交通標識などを認識するように学習し得る。

【００２８】

[0038]深層畳み込みネットワーク（ＤＣＮ）が、教師あり学習を用いてトレーニングされ得る。トレーニング中に、ＤＣＮは、速度制限標識のクロップされた画像３２６など、画像を提示され得、次いで、出力３２２を生成するために、「フォワードパス」が計算され得る。出力３２２は、「標識」、「６０」、および「１００」など、特徴に対応する値のベクトルであり得る。ネットワーク設計者は、ＤＣＮが、出力特徴ベクトルにおけるニューロンのうちのいくつか、たとえば、トレーニングされたネットワーク３００のための出力３２２に示されているように「標識」および「６０」に対応するニューロンについて、高いスコアを出力することを希望し得る。トレーニングの前に、ＤＣＮによって生成された出力は不正確である可能性があり、したがって、実際の出力とターゲット出力との間で誤差が計算され得る。次いで、ＤＣＮの重みは、ＤＣＮの出力スコアがターゲットとより密接に整合されるように調節され得る。

【００２９】

[0039]重みを調節するために、学習アルゴリズムは、重みのための勾配ベクトルを計算し得る。勾配は、重みがわずかに調節された場合に、誤差が増加または減少する量を示し得る。最上層において、勾配は、最後から２番目の層における活性化されたニューロンと出力層におけるニューロンとを結合する重みの値に直接対応し得る。下位層では、勾配は、重みの値と、上位層の計算された誤差勾配とに依存し得る。次いで、重みは、誤差を低減するように調節され得る。重みを調節するこの様式は、それがニューラルネットワークを通して「バックワードパス」を伴うので、「バックプロパゲーション」と呼ばれることがある。

【００３０】

[0040]実際には、重みの誤差勾配は、計算された勾配が真の誤差勾配を近似するように、少数の例にわたって計算され得る。この近似方法は、確率的勾配降下（stochastic gradient descent）と呼ばれることがある。システム全体の達成可能な誤差レートが減少し

なくなるまで、または誤差レートがターゲットレベルに達するまで、確率的勾配降下が繰り返され得る。

【００３１】

[0041]学習の後に、ＤＣＮは新しい画像３２６を提示され得、ネットワークを通したフォワードパスは、ＤＣＮの推論または予測と見なされ得る出力３２２をもたらし得る。

【００３２】

[0042]深層信念ネットワーク（ＤＢＮ：deep belief network）は、隠れノードの複数の層を備える確率モデルである。ＤＢＮは、トレーニングデータセットの階層表現を抽出するために使用され得る。ＤＢＮは、制限ボルツマンマシン（ＲＢＭ：Restricted Boltzmann Machine）の層を積層することによって取得され得る。ＲＢＭは、入力のセットにわたる確率分布を学習することができる人工ニューラルネットワークのタイプである。ＲＢＭは、各入力がそれにカテゴリー分類されるべきクラスに関する情報の不在下で確率分布を学習することができるので、ＲＢＭは、教師なし学習においてしばしば使用される。ハイブリッド教師なしおよび教師ありパラダイムを使用して、ＤＢＮの下部ＲＢＭは、教師なし様式でトレーニングされ得、特徴抽出器として働き得、上部ＲＢＭは、（前の層からの入力とターゲットクラスとの同時分布上で）教師あり様式でトレーニングされ得、分類器として働き得る。

【００３３】

[0043]深層畳み込みネットワーク（ＤＣＮ）は、追加のプーリング層および正規化層で構成された、畳み込みネットワークのネットワークである。ＤＣＮは、多くのタスクに関して最先端の性能を達成している。ＤＣＮは、入力と出力ターゲットの両方が、多くの標本について知られており、勾配降下方法の使用によってネットワークの重みを変更するために使用される、教師あり学習を使用してトレーニングされ得る。

【００３４】

[0044]ＤＣＮは、フィードフォワードネットワークであり得る。さらに、上記で説明されたように、ＤＣＮの第１の層におけるニューロンから次の上位層におけるニューロンのグループへの結合は、第１の層におけるニューロンにわたって共有される。ＤＣＮのフィードフォワードおよび共有結合は、高速処理のために活用され得る。ＤＣＮの計算負担は、たとえば、リカレントまたはフィードバック結合を備える同様のサイズのニューラルネットワークのそれよりもはるかに少ないことがある。

【００３５】

[0045]畳み込みネットワークの各層の処理は、空間的に不変のテンプレートまたは基底投射と見なされ得る。入力が、カラー画像の赤色、緑色、および青色チャネルなど、複数のチャネルに最初に分解された場合、その入力に関してトレーニングされた畳み込みネットワークは、画像の軸に沿った２つの空間次元と、色情報をキャプチャする第３の次元とをもつ、３次元であると見なされ得る。畳み込み結合の出力は、後続の層３１８および３２０において特徴マップを形成すると考えられ、特徴マップ（たとえば、３２０）の各要素が、前の層（たとえば、３１８）における様々なニューロンから、および複数のチャネルの各々から入力を受信し得る。特徴マップにおける値は、整流（rectification）、ｍａｘ（０，ｘ）など、非線形性を用いてさらに処理され得る。隣接するニューロンからの値は、さらにプールされ得、これは、ダウンサンプリングに対応し、さらなる局所不変性と次元削減とを与え得る。白色化に対応する正規化はまた、特徴マップにおけるニューロン間のラテラル抑制によって適用され得る。

【００３６】

[0046]深層学習アーキテクチャの性能は、より多くのラベリングされたデータポイントが利用可能となるにつれて、または計算能力が増加するにつれて、向上し得る。現代の深層ニューラルネットワークは、ほんの１５年前に一般的な研究者にとって利用可能であったものより数千倍も大きいコンピューティングリソースを用いて、ルーチン的にトレーニングされる。新しいアーキテクチャおよびトレーニングパラダイムが、深層学習の性能をさらに高め得る。整流された線形ユニット（rectified linear unit）は、勾配消失（van

ishing gradients）として知られるトレーニング問題を低減し得る。新しいトレーニング技法は、過学習（over-fitting）を低減し、したがって、より大きいモデルがより良い汎化を達成することを可能にし得る。カプセル化技法は、所与の受容野においてデータを抽出し、全体的性能をさらに高め得る。

【００３７】

[0047]図３Ｂは、例示的な深層畳み込みネットワーク３５０を示すブロック図である。深層畳み込みネットワーク３５０は、結合性および重み共有に基づく、複数の異なるタイプの層を含み得る。図３Ｂに示されているように、例示的な深層畳み込みネットワーク３５０は、複数の畳み込みブロック（たとえば、Ｃ１およびＣ２）を含む。畳み込みブロックの各々は、畳み込み層と、正規化層（ＬＮｏｒｍ）と、プーリング層とで構成され得る。畳み込み層は、１つまたは複数の畳み込みフィルタを含み得、これは、特徴マップを生成するために入力データに適用され得る。２つの畳み込みブロックのみが示されているが、本開示はそのように限定しておらず、代わりに、設計選好に従って、任意の数の畳み込みブロックが深層畳み込みネットワーク３５０中に含まれ得る。正規化層は、畳み込みフィルタの出力を正規化するために使用され得る。たとえば、正規化層は、白色化またはラテラル抑制を行い得る。プーリング層は、局所不変性および次元削減のために、空間にわたってダウンサンプリングアグリゲーションを行い得る。

【００３８】

[0048]たとえば、深層畳み込みネットワークの並列フィルタバンクは、高性能および低電力消費を達成するために、随意にＡＲＭ命令セットに基づいて、ＳＯＣ１００のＣＰＵ１０２またはＧＰＵ１０４にロードされ得る。代替実施形態では、並列フィルタバンクは、ＳＯＣ１００のＤＳＰ１０６またはＩＳＰ１１６にロードされ得る。さらに、ＤＣＮは、センサー１１４およびナビゲーション１２０に専用の処理ブロックなど、ＳＯＣ上に存在し得る他の処理ブロックにアクセスし得る。

【００３９】

[0049]深層畳み込みネットワーク３５０はまた、１つまたは複数の全結合層（たとえば、ＦＣ１およびＦＣ２）を含み得る。深層畳み込みネットワーク３５０は、ロジスティック回帰（ＬＲ）層をさらに含み得る。深層畳み込みネットワーク３５０の各層の間には、更新されるべき重み（図示せず）がある。各層の出力は、第１の畳み込みブロックＣ１において供給された入力データ（たとえば、画像、オーディオ、ビデオ、センサーデータおよび／または他の入力データ）から階層特徴表現を学習するために、深層畳み込みネットワーク３５０中の後続の層の入力として働き得る。

【００４０】

[0050]図４は、人工知能（ＡＩ）機能をモジュール化し得る例示的なソフトウェアアーキテクチャ４００を示すブロック図である。アーキテクチャを使用して、ＳＯＣ４２０の様々な処理ブロック（たとえば、ＣＰＵ４２２、ＤＳＰ４２４、ＧＰＵ４２６および／またはＮＰＵ４２８）に、アプリケーション４０２のランタイム動作中に計算をサポートすることを実施させ得るアプリケーション４０２が設計され得る。

【００４１】

[0051]ＡＩアプリケーション４０２は、たとえば、デバイスが現在動作するロケーションを示すシーンの検出および認識を与え得る、ユーザ空間４０４において定義されている機能を呼び出すように構成され得る。ＡＩアプリケーション４０２は、たとえば、認識されたシーンがオフィス、講堂、レストラン、または湖などの屋外環境であるかどうかに応じて別様に、マイクロフォンおよびカメラを構成し得る。ＡＩアプリケーション４０２は、現在のシーンの推定を与えるために、ＳｃｅｎｅＤｅｔｅｃｔアプリケーションプログラミングインターフェース（ＡＰＩ）４０６において定義されているライブラリに関連するコンパイルされたプログラムコードへの要求を行い得る。この要求は、たとえば、ビデオおよび測位データに基づくシーン推定を与えるように構成された深層ニューラルネットワークの出力に最終的に依拠し得る。

【００４２】

[0052]さらに、ランタイムフレームワークのコンパイルされたコードであり得るランタイムエンジン４０８が、ＡＩアプリケーション４０２にとってアクセス可能であり得る。ＡＩアプリケーション４０２は、たとえば、ランタイムエンジンに、特定の時間間隔における、またはアプリケーションのユーザインターフェースによって検出されたイベントによってトリガされた、シーン推定を要求させ得る。シーンを推定させられたとき、ランタイムエンジンは、ＳＯＣ４２０上で実行している、Ｌｉｎｕｘ（登録商標）カーネル４１２など、オペレーティングシステム４１０に信号を送り得る。オペレーティングシステム４１０は、ＣＰＵ４２２、ＤＳＰ４２４、ＧＰＵ４２６、ＮＰＵ４２８、またはそれらの何らかの組合せ上で、計算を実施させ得る。ＣＰＵ４２２は、オペレーティングシステムによって直接アクセスされ得、他の処理ブロックは、ＤＳＰ４２４のための、ＧＰＵ４２６のための、またはＮＰＵ４２８のためのドライバ４１４～４１８など、ドライバを通してアクセスされ得る。例示的な例では、深層ニューラルネットワークは、ＣＰＵ４２２およびＧＰＵ４２６など、処理ブロックの組合せ上で動作するように構成され得るか、または存在する場合、ＮＰＵ４２８上で動作させられ得る。

【００４３】

[0053]図５は、スマートフォン５０２上のＡＩアプリケーションのランタイム動作５００を示すブロック図である。ＡＩアプリケーションは、画像５０６のフォーマットを変換し、次いで画像５０８をクロップおよび／またはリサイズするように（たとえば、ＪＡＶＡ（登録商標）プログラミング言語を使用して）構成され得る前処理モジュール５０４を含み得る。次いで、前処理された画像は、視覚入力に基づいてシーンを検出および分類するように（たとえば、Ｃプログラミング言語を使用して）構成され得るＳｃｅｎｅＤｅｔｅｃｔバックエンドエンジン５１２を含んでいる分類アプリケーション５１０に通信され得る。ＳｃｅｎｅＤｅｔｅｃｔバックエンドエンジン５１２は、スケーリング５１６およびクロッピング５１８によって、画像をさらに前処理５１４するように構成され得る。たとえば、画像は、得られた画像が２２４ピクセル×２２４ピクセルであるように、スケーリングされ、クロップされ得る。これらの次元は、ニューラルネットワークの入力次元にマッピングし得る。ニューラルネットワークは、ＳＯＣ１００の様々な処理ブロックに、深層ニューラルネットワークを用いて画像ピクセルをさらに処理させるように、深層ニューラルネットワークブロック５２０によって構成され得る。次いで、深層ニューラルネットワークの結果は、しきい値処理５２２され、分類アプリケーション５１０中の指数平滑化ブロック５２４を通され得る。次いで、平滑化された結果は、スマートフォン５０２の設定および／またはディスプレイの変更を生じ得る。

【００４４】

[0054]一構成では、機械学習モデルは、機械学習モデルをトレーニングする間、バックプロパゲーションプロセスの勾配を変更するために構成される。モデルは、変更手段のための手段、および／または決定するための手段を含む。一態様では、変更手段、および／または決定手段は、具陳された機能を実行するように構成された、汎用プロセッサ１０２、汎用プロセッサ１０２に関連するプログラムメモリ、メモリブロック１１８、ローカル処理ユニット２０２、およびまたはルーティング接続処理ユニット２１６であり得る。別の構成では、上述の手段は、上述の手段によって具陳された機能を実行するように構成された任意のモジュールまたは任意の装置であり得る。

【００４５】

[0055]別の態様では、変更手段は、勾配をスケーリングするための手段を含み得る。随意に、変更手段は、勾配を選択的に適用するための手段を含み得る。

【００４６】

[0056]本開示のいくつかの態様によれば、各ローカル処理ユニット２０２は、モデルの所望の１つまたは複数の機能的特徴に基づいてモデルのパラメータを決定し、決定されたパラメータがさらに適合、調整および更新されるように、１つまたは複数の機能的特徴を所望の機能的特徴のほうへ発達させるように構成され得る。

【００４７】

[0057]多くの機械学習プロセスでは、学習された分類関数の出力と所望の出力との間の誤差を定量化するために、コスト関数が使用される。機械学習プロセスの目的は、このコスト関数を最小限に抑えるように、学習された分類関数のパラメータを変えることである。分類問題では、コスト関数は、しばしば、何らかの入力に関連する実際のクラスラベルと、その入力に関数を適用することによって達成される予測されたクラスラベルとのログ確率ペナルティ関数である。トレーニングは、学習された分類関数のパラメータを変更するプロセスである。トレーニング中に、例示的な入力とそれらの関連するラベルとが、機械学習プロセスに提示される。プロセスは、現在の学習された分類関数パラメータが与えられれば予測されたラベルを見つけ、コスト関数を評価し、学習された分類関数のパラメータを何らかの更新学習則に従って変える。

【００４８】

[0058]トレーニングプロセス中に、不平衡トレーニングデータの使用が、（１つまたは複数の）分類器をバイアスし得る。各クラスラベルのほぼ等しい数の例があるように、「学習則」など、ルールが、トレーニングデータを平衡させるための試みとして利用され得る。トレーニングデータが、あるクラスの多数の例と別のクラスの少数の例とを含んでいる場合、分類関数のパラメータは、より多数の例をもつクラスのほうへバイアスされる方法で、よりしばしば更新される。極端に言うと、第１のクラスの１００万個の例と第２のクラスの１つのみの例とを用いてバイナリ分類器をトレーニングしている場合、分類器は、単に常に第１のクラスを予測することによって極めてうまく機能する。別の例では、犬認識器がトレーニングされている。この例では、トレーニングデータは、合計１０００個の例を含み、ここで、例のうちの９９０個は犬であり、例のうちの１０個は猫である。分類器は、画像を犬として分類するために学習し得、これは、トレーニングセットに対して高精度で高い再現率を生じることになる。しかしながら、分類器が何も学ばなかった可能性が高い。

【００４９】

[0059]一般に、クラス間のトレーニングデータの「平衡化」は、各クラスについてのトレーニング例の相対頻度（relative frequency）が、トレーニング中に使用されない新しい例に分類器を適用するときに遭遇すると予想される相対頻度に一致することを保証することによって対処される。しかしながら、この手法は、いくつかの欠点を有する。第１に、それは、将来のデータセット中のクラス例の相対頻度が知られていると仮定する。しかしながら、これは、決定することが常に容易であるとは限らない。第２に、トレーニングデータは、各クラスのあまりに多くの例またはあまりに少数の例を含んでいることがある。トレーニング例を平衡させるために、データは、捨てられるかまたは繰り返される。データを捨てることによって、いくつかのクラスについて有益なトレーニングデータが除外され得、これは、分類器がそのクラスに関連する入力変動を十分に表すのを妨げ得る。簡単な方法でデータを繰り返すことによって、データを段階に分けるためにより多くのディスクスペースが使用される。特に、目的が、データのすべてを使用することである場合、あらゆるクラスが、完全な平衡のために最小公倍数まで繰り返されるであろう。さらに、各例が２つまたはそれ以上のラベルについて正としてラベリングされ得るマルチラベルデータの場合、すべてのラベルにわたる平衡は、複雑なスケジューリング訓練になり、単純な繰返しは十分でないことがある。

【００５０】

[0060]本開示の態様は、機械学習モデルにおいてクラス間のトレーニングデータを平衡させることを対象とする。特に、入力段においてトレーニングデータを操作し、各クラスについての例の数を調節するのではなく、本開示の態様は、勾配段における調節を対象とする。

【００５１】

[0061]後方への誤差伝播とも呼ばれるバックプロパゲーションが、コスト関数の勾配を計算するために利用され得る。特に、バックプロパゲーションは、誤差を０のより近くに低減するために重み値をどのように調節するかを決定することを含む。本開示の様々な態

様では、選択的バックプロパゲーションは、データセット中のクラス例頻度に基づいて勾配を調節するかまたは選択的に適用するための、何らかの所与のコスト関数への変更である。画像が入力され、勾配が、バックプロパゲーションを実行するために適用されようとしている後に、勾配は、各クラスについての例の頻度に基づいて調節され得る。

【００５２】

[0062]一態様では、調節は、トレーニングデータセット中の例の最小数（

【００５３】

【数１】

$$\min_{\mathbb{C}} N_c$$

【００５４】

）とトレーニングデータセット中のすべての例の数と（Ｎc、たとえば、最も少数のメンバーをもつクラスの例の数と現在のクラスの例の数と）の比である、相対クラス頻度ｆｃに関係する。（頻度ファクタ（frequency factor）とも呼ばれる）相対クラス頻度は、次のように表され得る。

【００５５】

【数２】

$$f_c = \frac{\min_{\mathbb{C}} N_c}{N_c} \quad \mathbb{C} \ni \text{すべての概念} \qquad (1)$$

【００５６】

[0063]例の最小数は、実際のまたは予想される数に基づき得る。さらに、トレーニングデータセット中のすべての例の数は、予想される数の例の実際の数に基づき得る。再び、犬認識器がトレーニングされている猫／犬の例を参照すると、犬の９９０個の例と猫の１０個の例とがある。犬のための各クラスについての頻度ファクタは、１０／９９０であり、ここで、１０は例の最小数であり、９９０は対象のクラスについての例の数である。猫のための各クラスについてのファクタは、１０／１０である。調節ファクタ（たとえば、相対クラス頻度）は、例の最小数を有するクラスについて値「１」であり、すべての他のクラスについて１よりも小さくなり得る。

【００５７】

[0064]頻度ファクタが決定されると、バックプロパゲーション勾配が変更される。変更は、各クラスについて勾配をスケーリングすることを含み得る。スケーリングは、次のように表され得る。

【００５８】

【数３】

$$\text{スケーリング}: \frac{dE_{applied}}{dx} = f_c \frac{dE}{dx} \qquad (2)$$

【００５９】

[0065]スケーリングインプリメンテーションでは、勾配は、頻度ファクタ（たとえば、相対クラス頻度）によって乗算され得る。勾配は、特定のパラメータに関する誤差の導関数である。あるクラスの多くの例がある一例では、そのクラスの過剰学習を防ぐために、勾配の分数のみが毎回適用される。連続する、犬の１０個の例がある犬／猫の例では、勾配の１０分の１のみが適用される。目的は、モデルが猫よりも犬のはるかに多い例を参照したので、モデルがすべての画像を過剰学習し、犬としてラベリングするのを防ぐことである。スケーリングは、特定のクラスのすべての重み中のすべての勾配に等しく適用される。

【００６０】

[0066]変更は、画像からサンプリングするためのファクタを使用することをも含み得る。サンプリングは、次のように表され得る。

【００６１】

【数４】

$$\text{サンプリング：}\frac{dE_{applied}}{dx}=\begin{cases}0, & \text{s＝0の場合}\\ \dfrac{dE}{dx}, & \text{s＝1の場合}\end{cases} \qquad (3)$$

【００６２】

[0067]ここで、勾配は、クラス例のサンプリングに基づいて選択的に適用される。一例では、サンプリングはランダムに適用される。スケーリングファクタの値は、サンプルがそれから引き出されるベルヌーイ分布の確率パラメータとして使用され得る。この分布からのサンプリングは０または１を生成し、１をサンプリングすることの確率が、第１の方法において説明されたスケーリングファクタに等しい。例の最小数をもつクラスの場合、サンプリングは、１を生成する。コインフリップ（coin flip）が１を生成するとき、そのクラスについての誤差勾配がバックプロパゲートされる。コインフリップが０を生成するとき、そのクラスについての勾配がバックプロパゲートされない場合が、事実上、０に設定される。言い換えれば、画像は、多くの例があるときに単に時々勾配を返送するために、勾配段においてサンプリングされる。最小数の例があるとき、勾配は毎回返送される。これは、入力を調節するのではなく勾配を調節することによって分類器がそれから学習している例の等化を与える。一態様では、画像を前方伝搬する前に、それは、クラスが、現在のエポックのためにその画像を使用するように設定されるかどうかを検査される。各エポックについて、セットは再シャッフルされ得る。

【００６３】

[0068]サンプリングは、個々ベース、エポックベース、またはトレーニングコーパスベースで適用され得る。上記で提示されたように、個々ベースでは、トレーニングエポック中に提示される他の画像に依存しない各画像のためのランダム結果がベルヌーイ分布から生成される。いくつかのエポックは、サンプリングのランダム性質により、各クラスについて、所望の数よりも多いまたは少ない例を参照し得る。

【００６４】

[0069]エポックベースの場合、スケールファクタは、すべてのクラス例から各クラスについてランダムに選択される。各エポック中に各クラスについて固定数の例が使用される。たとえば、１０個の例が各クラスから選択され得る。それらの例のみが、特定のエポック中にバックプロパゲートされる。

【００６５】

[0070]トレーニングコーパスベースの場合、頻度ファクタは、分類器にまだ提示されていないものから各クラスについて各エポックのためにランダムに選択される。例は、交換なしにサンプリングされる。以下の例示的な例では、１０００個の犬の例があり、各エポックでは、１０個のサンプルがランダムに選択される。第１のエポックでは、１０個の例が、合計１０００個の例から選択される。次のエポックでは、前の１０個の選択された例が除去され、１０個の例が、残りの９９０個の例から選択される。これは、例のすべてが使い尽くされるまで続き、各エポック中に各クラスについて同数の例が使用され、トレーニングの過程にわたってすべての利用可能な例が使用されることを保証する。次回、データを巡回するとき、同じ順序が維持され得るか、または代替的に、異なる順序が使用され得る。別の構成では、例は、交換を用いてサンプリングされる。

【００６６】

[0071]多くの場合、トレーニングの開始の前にトレーニングコーパス全体が利用可能であり、ｆｃファクタは、トレーニングセッションにわたって静的であり、トレーニングが始まる前に各クラスについて計算され得る。しかしながら、トレーニングが始まった後にクラスが追加されるか、またはトレーニング中にトレーニング例がアドホックに供給される場合、ｆｃファクタは、時間とともに変化しているかまたはトレーニングの開始時に未知であることがある。この状況では、各例が提示された後に、各クラスについての例の数（Ｎｃ）のランニングカウントが、保たれ、更新され得る。ｆｃファクタは、次いで、特定のクラス（ｃ）についてＮｃの各更新の後にオンザフライで計算される。

【００６７】

[0072]別の態様では、各クラスについてネットワークの変化の量を等化するために、および各クラスが分類器によって比較的同様に推測される可能性があることを保証するために、クラスの相対頻度（たとえば、頻度ファクタ）が利用される。相対頻度クラスは、データセット中のクラスの一様分布を促進する。他のクラスよりもいくつかのクラスのより多くがあるという知られている予想がある場合、頻度ファクタは調節され得る。たとえば、実世界において犬よりも多くの猫がいることが知られているが、トレーニングデータが犬の１０００個の例と猫の１０個の例とを含む場合、頻度ファクタは、実世界予想を考慮するように調節され得る。実世界において犬よりも猫を見る可能性が１０倍高いと知られている場合、頻度ファクタは、猫についてファクタ１０で乗算され、犬についてファクタ１で乗算され得る。本質的に、頻度ファクタ（Ｆｃ）は、実世界において存在するものの均一な予想をターゲットにするように学習段において操作され得る。頻度ファクタは、次のように調節され得る。

【００６８】

【数５】

$$f_c \frac{\min_c p(c)}{p(c)} \frac{\min_c N_c}{N_c}, \qquad (4)$$

【００６９】

ここで、ｐ（ｃ）は、実世界（または「野生」）における特定のクラスを観測する予想される確率である。

【００７０】

[0073]図６は、機械学習モデルのためのクラス間のトレーニングデータを平衡させるための方法６００を示す。ブロック６０２において、プロセスは、最も少数のメンバーをもつクラスの例の数と現在のクラスの例の数との比に基づいて、勾配を変更するためのファクタを決定する。最も少数のメンバーは、実際のまたは予想されるメンバーの数に基づき得る。同様に、現在のクラスの例の数は、例の実際のまたは予想される数に基づき得る。ブロック６０４において、プロセスは、決定されたファクタに基づいて、現在のクラスに関連する勾配を変更する。

【００７１】

[0074]図７は、機械学習モデルのためのクラス間のトレーニングデータを平衡させるための全体的方法７００を示す。ブロック７０２において、トレーニングデータを評価する。ブロック７０４において、クラス中の例の頻度を決定する。ブロック７０６において、決定された頻度に基づいて勾配を更新する。更新は、ブロック７１０において、各クラスについて勾配にスケーリングファクタを適用することによって実行され得る。代替的に、更新は、ブロック７０８において、クラス例のサンプルに基づいて勾配を選択的に適用することによって実行され得る。選択的サンプリング更新は、ブロック７１２において個々ベースで、ブロック７１４においてエポックベースで、またはブロック７１６においてトレーニングコーパスベースで実行され得る。

【００７２】

[0075]図８は、本開示の態様による、トレーニングデータを平衡させるための方法８００を示す。ブロック８０２において、プロセスは、モデルをトレーニングする間、バックプロパゲーションプロセスの勾配を変更する。変更は、最も少数のメンバーをもつクラスの例の数と現在のクラスの例の数との比に基づく。

【００７３】

[0076]いくつかの態様では、方法６００、７００、および８００は、ＳＯＣ１００（図１）またはシステム２００（図２）によって実行され得る。すなわち、方法１１００および１２００の要素の各々は、たとえば、限定はしないが、ＳＯＣ１００またはシステム２００または１つまたは複数のプロセッサ（たとえば、ＣＰＵ１０２およびローカル処理ユニット２０２）および／あるいは本明細書中に含まれる他の構成要素によって実行され得

る。いくつかの態様では、方法６００および７００は、ＳＯＣ４２０（図４）または１つまたは複数のプロセッサ（たとえば、ＣＰＵ４２２）および／あるいは本明細書中に含まれる他の構成要素によって実行され得る。

【００７４】

[0077]上記で説明された方法の様々な動作は、対応する機能を実施することが可能な任意の好適な手段によって実施され得る。それらの手段は、限定はしないが、回路、特定用途向け集積回路（ＡＳＩＣ）、またはプロセッサを含む、様々な（１つまたは複数の）ハードウェアおよび／またはソフトウェア構成要素および／またはモジュールを含み得る。概して、図に示されている動作がある場合、それらの動作は、同様の番号をもつ対応するカウンターパートのミーンズプラスファンクション構成要素を有し得る。

【００７５】

[0078]本明細書で使用される「決定すること」という用語は、多種多様なアクションを包含する。たとえば、「決定すること」は、計算すること（calculating）、計算すること（computing）、処理すること、導出すること、調査すること、ルックアップすること（たとえば、テーブル、データベースまたは別のデータ構造においてルックアップすること）、確認することなどを含み得る。さらに、「決定すること」は、受信すること（たとえば、情報を受信すること）、アクセスすること（たとえば、メモリ中のデータにアクセスすること）などを含み得る。さらに、「決定すること」は、解決すること、選択すること、選定すること、確立することなどを含み得る。

【００７６】

[0079]本明細書で使用される、項目のリスト「のうちの少なくとも１つ」を指す句は、単一のメンバーを含む、それらの項目の任意の組合せを指す。一例として、「ａ、ｂ、またはｃのうちの少なくとも１つ」は、ａ、ｂ、ｃ、ａ－ｂ、ａ－ｃ、ｂ－ｃ、およびａ－ｂ－ｃを包含するものとする。

【００７７】

[0080]本開示に関連して説明された様々な例示的な論理ブロック、モジュールおよび回路は、汎用プロセッサ、デジタル信号プロセッサ（ＤＳＰ）、特定用途向け集積回路（ＡＳＩＣ）、フィールドプログラマブルゲートアレイ信号（ＦＰＧＡ）または他のプログラマブル論理デバイス（ＰＬＤ）、個別ゲートまたはトランジスタ論理、個別ハードウェア構成要素、あるいは本明細書で説明された機能を実施するように設計されたそれらの任意の組合せを用いてインプリメントまたは実施され得る。汎用プロセッサはマイクロプロセッサであり得るが、代替として、プロセッサは、任意の市販のプロセッサ、コントローラ、マイクロコントローラ、または状態機械であり得る。プロセッサはまた、コンピューティングデバイスの組合せ、たとえば、ＤＳＰとマイクロプロセッサとの組合せ、複数のマイクロプロセッサ、ＤＳＰコアと連携する１つまたは複数のマイクロプロセッサ、あるいは任意の他のそのような構成としてインプリメントされ得る。

【００７８】

[0081]本開示に関連して説明された方法またはアルゴリズムのステップは、ハードウェアで直接実施されるか、プロセッサによって実行されるソフトウェアモジュールで実施されるか、またはその２つの組合せで実施され得る。ソフトウェアモジュールは、当技術分野で知られている任意の形態の記憶媒体中に常駐し得る。使用され得る記憶媒体のいくつかの例としては、ランダムアクセスメモリ（ＲＡＭ）、読取り専用メモリ（ＲＯＭ）、フラッシュメモリ、消去可能プログラマブル読取り専用メモリ（ＥＰＲＯＭ）、電気消去可能プログラマブル読取り専用メモリ（ＥＥＰＲＯＭ（登録商標））、レジスタ、ハードディスク、リムーバブルディスク、ＣＤ－ＲＯＭなどがある。ソフトウェアモジュールは、単一の命令、または多数の命令を備え得、いくつかの異なるコードセグメント上で、異なるプログラム間で、および複数の記憶媒体にわたって分散され得る。記憶媒体は、プロセッサがその記憶媒体から情報を読み取ることができ、その記憶媒体に情報を書き込むことができるように、プロセッサに結合され得る。代替として、記憶媒体はプロセッサと一体であり得る。

【００７９】
[0082]本明細書で開示された方法は、説明された方法を達成するための１つまたは複数のステップまたはアクションを備える。本方法のステップおよび／またはアクションは、特許請求の範囲から逸脱することなく、互いに交換され得る。言い換えれば、ステップまたはアクションの特定の順序が指定されない限り、特定のステップおよび／またはアクションの順序および／または使用は特許請求の範囲から逸脱することなく変更され得る。
【００８０】
[0083]説明された機能は、ハードウェア、ソフトウェア、ファームウェア、またはそれらの任意の組合せでインプリメントされ得る。ハードウェアでインプリメントされる場合、例示的なハードウェア構成はデバイス中に処理システムを備え得る。処理システムは、バスアーキテクチャを用いてインプリメントされ得る。バスは、処理システムの特定の適用例および全体的な設計制約に応じて、任意の数の相互接続バスおよびブリッジを含み得る。バスは、プロセッサと、機械可読媒体と、バスインターフェースとを含む様々な回路を互いにリンクし得る。バスインターフェースは、ネットワークアダプタを、特に、バスを介して処理システムに接続するために使用され得る。ネットワークアダプタは、信号処理機能をインプリメントするために使用され得る。いくつかの態様では、ユーザインターフェース（たとえば、キーパッド、ディスプレイ、マウス、ジョイスティックなど）もバスに接続され得る。バスはまた、タイミングソース、周辺機器、電圧調整器、電力管理回路など、様々な他の回路をリンクし得るが、それらは当技術分野でよく知られており、したがってこれ以上説明されない。
【００８１】
[0084]プロセッサは、機械可読媒体に記憶されたソフトウェアの実行を含む、バスおよび一般的な処理を管理することを担当し得る。プロセッサは、１つまたは複数の汎用および／または専用プロセッサを用いてインプリメントされ得る。例としては、マイクロプロセッサ、マイクロコントローラ、ＤＳＰプロセッサ、およびソフトウェアを実行することができる他の回路がある。ソフトウェアは、ソフトウェア、ファームウェア、ミドルウェア、マイクロコード、ハードウェア記述言語などの名称にかかわらず、命令、データ、またはそれらの任意の組合せを意味すると広く解釈されたい。機械可読媒体は、例として、ランダムアクセスメモリ（ＲＡＭ）、フラッシュメモリ、読取り専用メモリ（ＲＯＭ）、プログラマブル読取り専用メモリ（ＰＲＯＭ）、消去可能プログラマブル読取り専用メモリ（ＥＰＲＯＭ）、電気消去可能プログラマブル読取り専用メモリ（ＥＥＰＲＯＭ）、レジスタ、磁気ディスク、光ディスク、ハードドライブ、または他の好適な記憶媒体、あるいはそれらの任意の組合せを含み得る。機械可読媒体はコンピュータプログラム製品において実施され得る。コンピュータプログラム製品はパッケージング材料を備え得る。
【００８２】
[0085]ハードウェアインプリメンテーションでは、機械可読媒体は、プロセッサとは別個の処理システムの一部であり得る。しかしながら、当業者なら容易に理解するように、機械可読媒体またはその任意の部分は処理システムの外部にあり得る。例として、機械可読媒体は、すべてバスインターフェースを介してプロセッサによってアクセスされ得る、伝送線路、データによって変調された搬送波、および／またはデバイスとは別個のコンピュータ製品を含み得る。代替的に、または追加として、機械可読媒体またはその任意の部分は、キャッシュおよび／または汎用レジスタファイルがそうであり得るように、プロセッサに統合され得る。局所構成要素など、説明された様々な構成要素は、特定のロケーションを有するものとして説明され得るが、それらはまた、分散コンピューティングシステムの一部として構成されているいくつかの構成要素など、様々な方法で構成され得る。
【００８３】
[0086]処理システムは、すべて外部バスアーキテクチャを介して他のサポート回路と互いにリンクされる、プロセッサ機能を提供する１つまたは複数のマイクロプロセッサと、機械可読媒体の少なくとも一部を提供する外部メモリとをもつ汎用処理システムとして構成され得る。代替的に、処理システムは、本明細書で説明されたニューロンモデルとニュ

ーラルシステムのモデルとをインプリメントするための１つまたは複数の神経形態学的プロセッサを備え得る。別の代替として、処理システムは、プロセッサをもつ特定用途向け集積回路（ＡＳＩＣ）と、バスインターフェースと、ユーザインターフェースと、サポート回路と、単一のチップに統合された機械可読媒体の少なくとも一部分とを用いて、あるいは１つまたは複数のフィールドプログラマブルゲートアレイ（ＦＰＧＡ）、プログラマブル論理デバイス（ＰＬＤ）、コントローラ、状態機械、ゲート論理、個別ハードウェア構成要素、もしくは他の好適な回路、または本開示全体にわたって説明された様々な機能を実施することができる回路の任意の組合せを用いて、インプリメントされ得る。当業者は、特定の適用例と、全体的なシステムに課される全体的な設計制約とに応じて、どのようにしたら処理システムについて説明された機能を最も良くインプリメントし得るかを理解されよう。

【００８４】

[0087]機械可読媒体はいくつかのソフトウェアモジュールを備え得る。ソフトウェアモジュールは、プロセッサによって実行されたときに、処理システムに様々な機能を実施させる命令を含む。ソフトウェアモジュールは、送信モジュールと受信モジュールとを含み得る。各ソフトウェアモジュールは、単一の記憶デバイス中に常駐するか、または複数の記憶デバイスにわたって分散され得る。例として、トリガイベントが発生したとき、ソフトウェアモジュールがハードドライブからＲＡＭにロードされ得る。ソフトウェアモジュールの実行中、プロセッサは、アクセス速度を高めるために、命令のいくつかをキャッシュにロードし得る。次いで、１つまたは複数のキャッシュラインが、プロセッサによる実行のために汎用レジスタファイルにロードされ得る。以下でソフトウェアモジュールの機能に言及する場合、そのような機能は、そのソフトウェアモジュールからの命令を実行したときにプロセッサによってインプリメントされることが理解されよう。さらに、本開示の態様が、そのような態様をインプリメントするプロセッサ、コンピュータ、機械、または他のシステムの機能に改善を生じることを諒解されたい。

【００８５】

[0088]ソフトウェアでインプリメントされる場合、機能は、１つまたは複数の命令またはコードとしてコンピュータ可読媒体上に記憶されるか、あるいはコンピュータ可読媒体を介して送信され得る。コンピュータ可読媒体は、ある場所から別の場所へのコンピュータプログラムの転送を可能にする任意の媒体を含む、コンピュータ記憶媒体と通信媒体の両方を含む。記憶媒体は、コンピュータによってアクセスされ得る任意の利用可能な媒体であり得る。限定ではなく例として、そのようなコンピュータ可読媒体は、ＲＡＭ、ＲＯＭ、ＥＥＰＲＯＭ、ＣＤ－ＲＯＭまたは他の光ディスクストレージ、磁気ディスクストレージまたは他の磁気ストレージデバイス、あるいは命令またはデータ構造の形態の所望のプログラムコードを搬送または記憶するために使用され得、コンピュータによってアクセスされ得る、任意の他の媒体を備えることができる。さらに、いかなる接続もコンピュータ可読媒体と適切に呼ばれる。たとえば、ソフトウェアが、同軸ケーブル、光ファイバーケーブル、ツイストペア、デジタル加入者回線（ＤＳＬ）、または赤外線（ＩＲ）、無線、およびマイクロ波などのワイヤレス技術を使用して、ウェブサイト、サーバ、または他のリモートソースから送信される場合、同軸ケーブル、光ファイバーケーブル、ツイストペア、ＤＳＬ、または赤外線、無線、およびマイクロ波などのワイヤレス技術は、媒体の定義に含まれる。本明細書で使用されるディスク（disk）およびディスク（disc）は、コンパクトディスク（disc）（ＣＤ）、レーザーディスク（登録商標）（disc）、光ディスク（disc）、デジタル多用途ディスク（disc）（ＤＶＤ）、フロッピー（登録商標）ディスク（disk）、およびＢｌｕ－ｒａｙ（登録商標）ディスク（disc）を含み、ディスク（disk）は、通常、データを磁気的に再生し、ディスク（disc）は、データをレーザーで光学的に再生する。したがって、いくつかの態様では、コンピュータ可読媒体は非一時的コンピュータ可読媒体（たとえば、有形媒体）を備え得る。さらに、他の態様では、コンピュータ可読媒体は一時的コンピュータ可読媒体（たとえば、信号）を備え得る。上記の組合せもコンピュータ可読媒体の範囲内に含まれるべきである。

【００８６】
[0089]したがって、いくつかの態様は、本明細書で提示された動作を実施するためのコンピュータプログラム製品を備え得る。たとえば、そのようなコンピュータプログラム製品は、本明細書で説明された動作を実行するために１つまたは複数のプロセッサによって実行可能である命令をその上に記憶した（および／または符号化した）コンピュータ可読媒体を備え得る。いくつかの態様では、コンピュータプログラム製品はパッケージング材料を含み得る。
【００８７】
[0090]さらに、本明細書で説明された方法および技法を実行するためのモジュールおよび／または他の適切な手段は、適用可能な場合にユーザ端末および／または基地局によってダウンロードされ、および／または他の方法で取得され得ることを諒解されたい。たとえば、そのようなデバイスは、本明細書で説明された方法を実行するための手段の転送を可能にするためにサーバに結合され得る。代替的に、本明細書で説明された様々な方法は、ユーザ端末および／または基地局が記憶手段（たとえば、ＲＡＭ、ＲＯＭ、コンパクトディスク（ＣＤ）またはフロッピーディスクなどの物理記憶媒体など）をデバイスに結合するかまたは与えると様々な方法を得ることができるように、記憶手段によって提供され得る。その上、本明細書で説明された方法および技法をデバイスに提供するための任意の他の好適な技法が利用され得る。
【００８８】
[0091]特許請求の範囲は、上記で示された厳密な構成および構成要素に限定されないことを理解されたい。上記で説明された方法および装置の構成、動作および詳細において、特許請求の範囲から逸脱することなく、様々な改変、変更および変形が行われ得る。
以下に本願の出願当初の特許請求の範囲に記載された発明を付記する。
［Ｃ１］
機械学習モデルのためのクラス間のトレーニングデータの平衡を変更する方法であって、
最も少数のメンバーをもつクラスの例の数と現在のクラスの例の数との比に少なくとも部分的に基づいて、前記モデルをトレーニングする間、バックプロパゲーションプロセスの勾配を変更することを備える、方法。
［Ｃ２］
前記変更することが、前記勾配をスケーリングすることを備える、Ｃ１に記載の方法。
［Ｃ３］
前記変更することが、前記クラス例のサンプリングに少なくとも部分的に基づいて前記勾配を選択的に適用することを備える、Ｃ１に記載の方法。
［Ｃ４］
前記クラスの前記サンプリングが、各トレーニングエポックから固定数の例を選択することによって行われる、Ｃ３に記載の方法。
［Ｃ５］
前記サンプリングが、トレーニングエポック中の例の交換なしに行われる、Ｃ１に記載の方法。
［Ｃ６］
機械学習モデルのためのクラス間のトレーニングデータの平衡を変更するための装置であって、
最も少数のメンバーをもつクラスの例の数と現在のクラスの例の数との比に少なくとも部分的に基づいて、勾配を変更するためのファクタを決定するための手段と、
前記決定されたファクタに基づいて、前記現在のクラスに関連する前記勾配を変更するための手段とを備える、装置。
［Ｃ７］
前記変更手段が、前記勾配をスケーリングするための手段を備える、Ｃ６に記載の装置。

［Ｃ８］
前記変更手段が、前記クラス例のサンプリングに少なくとも部分的に基づいて前記勾配を選択的に適用するための手段を備える、Ｃ６に記載の装置。
［Ｃ９］
前記クラスの前記サンプリングが、各トレーニングエポックから固定数の例を選択することによって行われる、Ｃ８に記載の装置。
［Ｃ１０］
前記サンプリングが、トレーニングエポック中の例の交換なしに行われる、Ｃ６に記載の装置。
［Ｃ１１］
機械学習モデルのためのクラス間のトレーニングデータの平衡を変更するための装置であって、
メモリと、
前記メモリに結合された少なくとも１つのプロセッサと、前記少なくとも１つのプロセッサが、最も少数のメンバーをもつクラスの例の数と現在のクラスの例の数との比に少なくとも部分的に基づいて、前記モデルをトレーニングする間、バックプロパゲーションプロセスの勾配を変更するように構成された、を備える、装置。
［Ｃ１２］
前記少なくとも１つのプロセッサが、前記勾配をスケーリングすることによって変更するように構成された、Ｃ１１に記載の装置。
［Ｃ１３］
前記少なくとも１つのプロセッサが、前記クラス例のサンプリングに少なくとも部分的に基づいて、前記勾配を選択的に適用することによって変更するように構成された、Ｃ１１に記載の装置。
［Ｃ１４］
前記クラスの前記サンプリングが、各トレーニングエポックから固定数の例を選択することによって行われる、Ｃ１３に記載の装置。
［Ｃ１５］
前記サンプリングが、トレーニングエポック中の例の交換なしに行われる、Ｃ１１に記載の装置。
［Ｃ１６］
機械学習モデルのためのクラス間のトレーニングデータの平衡を変更するための非一時的コンピュータ可読媒体であって、前記非一時的コンピュータ可読媒体がそれに記録されたプログラムコードを有し、前記プログラムコードが、
最も少数のメンバーをもつクラスの例の数と現在のクラスの例の数との比に少なくとも部分的に基づいて、前記モデルをトレーニングする間、バックプロパゲーションプロセスの勾配を変更するためのプログラムコードを備える、非一時的コンピュータ可読媒体。
［Ｃ１７］
変更するための前記プログラムコードが、前記勾配をスケーリングするためのプログラムコードを備える、Ｃ１６に記載の非一時的コンピュータ可読媒体。
［Ｃ１８］
変更するための前記プログラムコードが、前記クラス例のサンプリングに少なくとも部分的に基づいて前記勾配を選択的に適用するためのプログラムコードを備える、Ｃ１６に記載の非一時的コンピュータ可読媒体。
［Ｃ１９］
前記クラスの前記サンプリングが、各トレーニングエポックから固定数の例を選択することによって行われる、Ｃ１８に記載の非一時的コンピュータ可読媒体。
［Ｃ２０］
前記サンプリングが、トレーニングエポック中の例の交換なしに行われる、Ｃ１６に記載の非一時的コンピュータ可読媒体。

【図１】

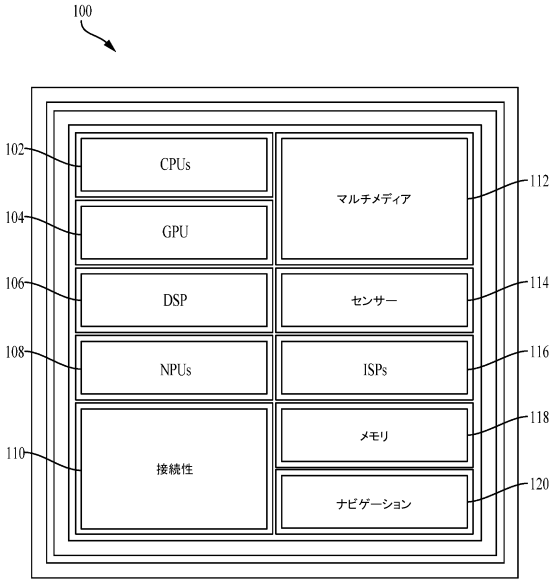

FIG. 1

【図２】

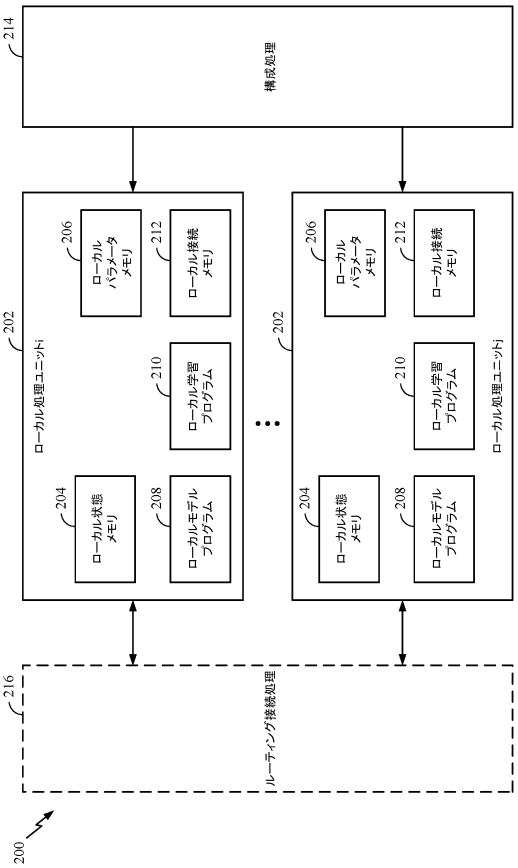

FIG. 2

【図３Ａ】

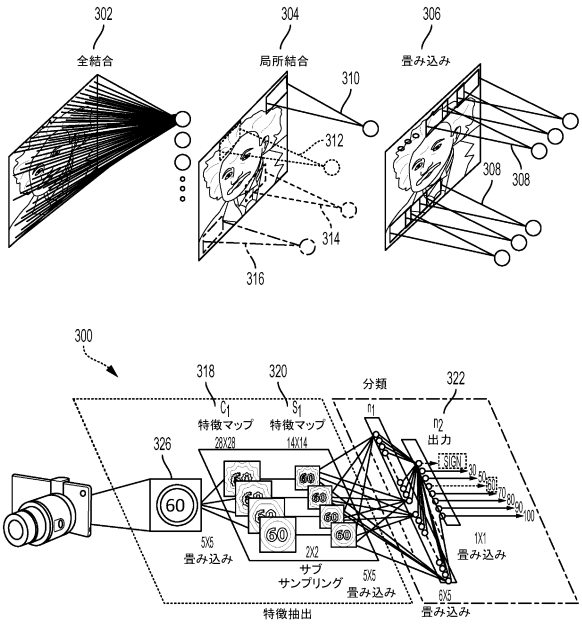

FIG. 3A

【図３Ｂ】

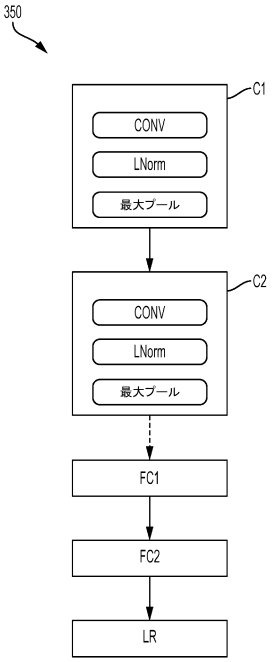

FIG. 3B

【図４】

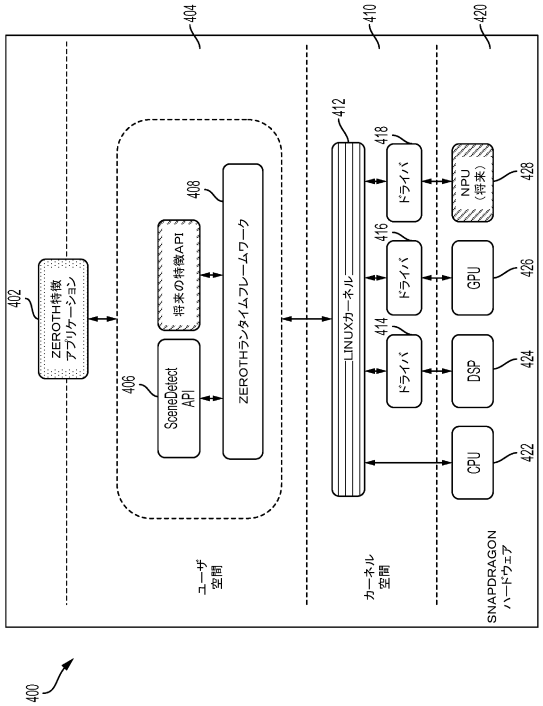

FIG. 4

【図５】

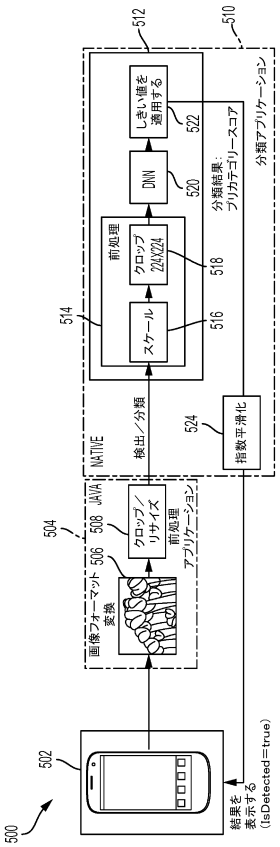

FIG. 5

【図６】

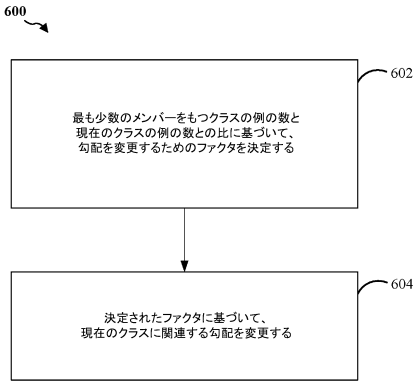

FIG. 6

【図７】

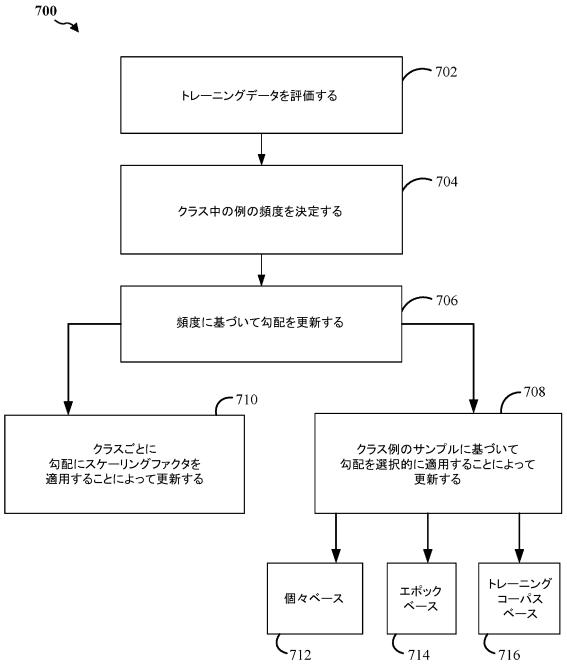

FIG. 7

【図８】

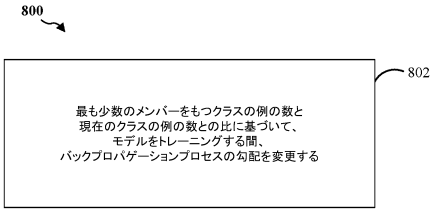
800
802
最も少数のメンバーをもつクラスの例の数と
現在のクラスの例の数との比に基づいて、
モデルをトレーニングする間、
バックプロパゲーションプロセスの勾配を変更する

FIG. 8

フロントページの続き

(74)代理人 100184332
弁理士 中丸 慶洋
(72)発明者 トワル、レーガン・ブライス
アメリカ合衆国、カリフォルニア州 ９２１２１－１７１４、サン・ディエゴ、モアハウス・ドライブ ５７７５
(72)発明者 ジュリアン、デイビッド・ジョナサン
アメリカ合衆国、カリフォルニア州 ９２１０１、サン・ディエゴ、ファースト・アベニュー ５１０、ユニット ５０４

審査官 杉浦 孝光

(56)参考文献 特開２００９－１２２８５１（ＪＰ，Ａ）
OH, Sang-Hoon , Error back-propagation algorithm for classification of imbalanced data , Neurocomputing, ２０１１年, vol.74, Issue.6, pp.1058-1061, ＵＲＬ, https://www.sciencedirect.com/science/article/abs/pii/S0925231210005084

(58)調査した分野(Int.Cl., ＤＢ名)
Ｇ０６Ｎ ３／００－９９／００