(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6862426号
(P6862426)
(45)発行日　令和3年4月21日(2021.4.21)　(24)登録日　令和3年4月2日(2021.4.2)

(51)Int.Cl.　　F I
G06N 20/00 (2019.01)　　G06N 20/00
G06N 3/08 (2006.01)　　G06N 3/08

請求項の数 15　(全 31 頁)

(21)出願番号　特願2018-510504 (P2018-510504)
(86)(22)出願日　平成28年8月11日 (2016.8.11)
(65)公表番号　特表2018-529159 (P2018-529159A)
(43)公表日　平成30年10月4日 (2018.10.4)
(86)国際出願番号　PCT/US2016/046576
(87)国際公開番号　WO2017/034820
(87)国際公開日　平成29年3月2日 (2017.3.2)
審査請求日　令和1年7月19日 (2019.7.19)
(31)優先権主張番号　62/209,859
(32)優先日　平成27年8月25日 (2015.8.25)
(33)優先権主張国・地域又は機関
米国 (US)
(31)優先権主張番号　14/863,410
(32)優先日　平成27年9月23日 (2015.9.23)
(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】トレーニングされた機械学習モデルのパフォーマンスを改善するための方法

(57)【特許請求の範囲】
【請求項１】
トレーニングされた機械学習モデルのパフォーマンスを改善するためのコンピュータによりインプリメントされる方法であって、
第１の目的関数を有する第１の分類器に、第２の目的関数を有する第２の分類器を追加すること、前記第１の目的関数は、前記第２の目的関数とは異なり、前記第１の目的関数は、微分可能であり、前記第２の目的関数は、微分不可能であり、前記第２の目的関数は、前記第１の分類器の誤差を直接的に低減させるために使用される、
を備える、コンピュータによりインプリメントされる方法。
【請求項２】
前記第２の目的関数は、前記第１の分類器と前記第２の分類器との誤差間の差の関数である、請求項１に記載のコンピュータによりインプリメントされる方法。
【請求項３】
より高い複雑度のモデルからの確率の混合に少なくとも部分的に基づいて、前記第２の目的関数を決定することをさらに備える、請求項１に記載のコンピュータによりインプリメントされる方法。
【請求項４】
前記第１の分類器を再トレーニングすることなく、前記第２の分類器を追加することをさらに備える、請求項１に記載のコンピュータによりインプリメントされる方法。
【請求項５】

前記第１の分類器の外部に前記第２の分類器を追加することをさらに備える、請求項１に記載のコンピュータによりインプリメントされる方法。
【請求項６】
アイデンティティ値に、前記第１の分類器によってトレーニングされたモデルによって生成される特徴に対する重みを割り当てることをさらに備える、請求項１に記載のコンピュータによりインプリメントされる方法。
【請求項７】
ゼロに、高い複雑度のモデルの確率ベクトルによって生成される特徴に対する重みを割り当てることをさらに備える、請求項６に記載のコンピュータによりインプリメントされる方法。
【請求項８】
前記第２の分類器の確率ベクトルによって生成される特徴に対する重みを割り当てることをさらに備える、請求項１に記載のコンピュータによりインプリメントされる方法。
【請求項９】
ゼロに、高い複雑度のモデルの確率ベクトルによって生成される特徴に対する重みを割り当てることをさらに備える、請求項１に記載のコンピュータによりインプリメントされる方法。
【請求項１０】
前記第２の分類器の確率ベクトルによって生成される特徴に対する重みを割り当てることをさらに備える、請求項９に記載のコンピュータによりインプリメントされる方法。
【請求項１１】
固定された温度Ｔによって、より高い複雑度のモデルによって生成される確率ベクトルをスケーリングすることをさらに備える、請求項１に記載のコンピュータによりインプリメントされる方法。
【請求項１２】
トレーニングされた機械学習モデルのパフォーマンスを改善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
を備え、前記少なくとも１つのプロセッサは、
第１の目的関数を有する第１の分類器に、第２の目的関数を有する第２の分類器を追加すること、前記第１の目的関数は、微分可能であり、前記第２の目的関数は、微分不可能であり、前記第１の目的関数は、前記第２の目的関数とは異なり、前記第２の目的関数は、前記第１の分類器の誤差を直接的に低減させるために使用される、
を行うように構成される、装置。
【請求項１３】
ｉ）前記第２の目的関数は、前記第１の分類器と前記第２の分類器との誤差間の差の関数である、
ｉｉ）前記少なくとも１つのプロセッサは、より高い複雑度のモデルからの確率の混合に少なくとも部分的に基づいて、前記第２の目的関数を決定するようにさらに構成される、
ｉｉｉ）前記少なくとも１つのプロセッサは、前記第１の分類器を再トレーニングすることなく、前記第２の分類器を追加するようにさらに構成される、
ｉｖ）前記少なくとも１つのプロセッサは、前記第１の分類器の外部に前記第２の分類器を追加するようにさらに構成される、
ｖ）前記少なくとも１つのプロセッサは、アイデンティティ値に、前記第１の分類器によってトレーニングされたモデルによって生成される特徴に対する重みを割り当てるようにさらに構成される、
ｖｉ）前記少なくとも１つのプロセッサは、前記第２の分類器の確率ベクトルによって生成される特徴に対する重みを割り当てるようにさらに構成される、
ｖｉｉ前記少なくとも１つのプロセッサは、ゼロに、高い複雑度のモデルの確率ベクト

ルによって生成される特徴に対する重みを割り当てるようにさらに構成される、
ｖｉｉｉ）前記少なくとも１つのプロセッサは、固定された温度Ｔによって、より高い複雑度のモデルによって生成される確率ベクトルをスケーリングするようにさらに構成される、
のうちの１つを備える、請求項１２に記載の装置。
【請求項１４】
トレーニングされた機械学習モデルのパフォーマンスを改善するための装置であって、
第１の目的関数を有する第１の分類器に、第２の目的関数を有する第２の分類器を追加するための手段と、前記第１の目的関数は、前記第２の目的関数とは異なり、前記第１の目的関数は、微分可能であり、前記第２の目的関数は、微分不可能であり、前記第２の目的関数は、前記第１の分類器の誤差を直接的に低減させるために使用される、
前記トレーニングされた機械学習モデルを介して受信される入力に少なくとも部分的に基づいて、前記第２の分類器から特徴ベクトルを出力するための手段と
を備える装置。
【請求項１５】
トレーニングされた機械学習モデルのパフォーマンスを改善するためのプログラムコードをその上に符号化された非一時的なコンピュータ可読媒体であって、前記プログラムコードは、プロセッサによって実行され、第１の目的関数を有する第１の分類器に、第２の目的関数を有する第２の分類器を追加するためのプログラムコードを備え、前記第１の目的関数は、前記第２の目的関数とは異なり、前記第１の目的関数は、微分可能であり、前記第２の目的関数は、微分不可能であり、前記第２の目的関数は、前記第１の分類器の誤差を直接的に低減させるために使用される、非一時的なコンピュータ可読媒体。
【発明の詳細な説明】
【関連出願への相互参照】
【０００１】
[0001]本願は、２０１５年８月２５日に出願された「トレーニングされた機械学習モデルのパフォーマンスを改善するための方法（METHOD FOR IMPROVING PERFORMANCE OF A TRAINED MACHINE LEARNING MODEL）」と題する米国仮特許出願第６２／２０９，８５９号の利益を主張し、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。
【技術分野】
【０００２】
[0002]本開示のある特定の態様は、一般に機械学習に関し、さらに詳細には、トレーニングされた機械学習モデルのパフォーマンスを改善するシステムおよび方法に関する。
【背景技術】
【０００３】
[0003]相互結合された人工ニューロン（例えば、ニューロンモデル）のグループを備え得る、人工ニューラルネットワークのような機械学習モデルは、計算デバイスであるか、または計算デバイスによって実行される方法を表す。
【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（feed-forward）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かつ入力空間を集合的にタイリングする（collectively tile）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 networks）は、多数のアプリケーションを有する。特に、ＣＮＮは、パターン認識および分類（classification）の分野において幅広く使用されてきた。
【０００５】
[0005]ディープビリーフネットワーク（deep belief networks）およびディープ畳み込みネットワークのような、ディープラーニングアーキテクチャ（deep learning architectures）は、第１の層のニューロンの出力が第２の層のニューロンへの入力となり、第２の層のニューロンの出力が第３の層のニューロンへの入力となるというような、層状のニ

ューラルネットワークアーキテクチャ（layered neural networks architectures）である。ディープニューラルネットワークは、特徴の階層（hierarchy of features）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、オブジェクト認識アプリケーション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と同様に、これらのディープラーニングアーキテクチャにおける計算は、処理ノードの集団（population）にわたって分散され得、これは、１つまたは複数の計算チェーンにおいて構成され得る。これらの多層アーキテクチャは、一度に１層ずつ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（fine-tuned）され得る。

【０００６】

[0006]他のモデルもまた、オブジェクト認識のために利用可能である。例えば、サポートベクターマシン（ＳＶＭ：support vector machines）は、分類に適用されることができる学習ツールである。サポートベクターマシンは、データをカテゴリ化する（categorizes）分離超平面（例えば、決定境界）を含む。超平面は、教師あり学習（supervised learning）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大させる。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（greatest minimum distance）を有するべきである。

【０００７】

[0007]これらの解決策は、多数の（a number of）分類ベンチマークに対して優れた結果を達成するが、それらの計算複雑度は、極めて高くなり得る。加えて、モデルのトレーニングは、困難であり得る。

【発明の概要】

【０００８】

[0008]本開示のある態様では、トレーニングされた機械学習モデルのパフォーマンスを改善するための方法が提示される。方法は、第１の目的関数を有する第１の分類器に、第２の目的関数を有する第２の分類器を追加することを備える。第２の目的関数は、第１の分類器の誤差（errors）を直接的に低減させるために使用される。

【０００９】

[0009]別の態様では、トレーニングされた機械学習モデルのパフォーマンスを改善するための装置が提示される。装置は、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含む。（１つまたは複数の）プロセッサは、第１の目的関数を有する第１の分類器に、第２の目的関数を有する第２の分類器を追加するように構成される。第２の目的関数は、第１の分類器の誤差を直接的に低減させるために使用される。

【００１０】

[0010]さらに別の態様では、トレーニングされた機械学習モデルのパフォーマンスを改善するための装置が提示される。装置は、第１の目的関数を有する第１の分類器に、第２の目的関数を有する第２の分類器を追加するための手段を含む。第２の目的関数は、第１の分類器の誤差を直接的に低減させるために使用される。装置は、トレーニングされた機械学習モデルを介して受信される入力に基づいて、第２の分類器から特徴ベクトルを出力するための手段をさらに含む。

【００１１】

[0011]さらになお別の態様では、非一時的なコンピュータ可読媒体が提示される。非一時的なコンピュータ可読媒体は、トレーニングされた学習機械モデルのパフォーマンスを改善するためのプログラムコードをその上に符号化されている。プログラムコードは、プロセッサによって実行され、第１の目的関数を有する第１の分類器に、第２の目的関数を有する第２の分類器を追加するためのプログラムコードを含む。第２の目的関数は、第１の分類器の誤差を直接的に低減させるために使用される。

【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。本開示が、本開示と同じ目的を実行するための他の構造を修正（modifying）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が、当業者によって理解されるべきである。また、そのような同等

の構造が、添付された特許請求の範囲に示される本開示の教示から逸脱しないことも、当業者によって認識されるべきである。さらなる目的および利点と共に、その構成および動作の方法の両方について、本開示の特徴であると考えられる新規の特徴は、添付の図面に関連して考慮されるとき、以下の説明からより良く理解されるであろう。しかしながら、図面の各々は、例示および説明のみの目的で提供されており、本開示の限定の定義として意図されるものではないことが、明確に理解されるべきである。

【図面の簡単な説明】

【００１３】

[0013]本開示の特徴、性質、および利点は、同様の参照符号が全体を通して同様のものを指す図面と共に考慮されるとき、以下に示される詳細な説明からより明らかになるであろう。

【図１】[0014]図１は、本開示のある特定の態様による、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となるインプリメンテーションを例示する。

【図２】[0015]図２は、本開示の態様による、システムの例となるインプリメンテーションを例示する。

【図３Ａ】[0016]図３Ａは、本開示の態様による、ニューラルネットワークを例示する図である。

【図３Ｂ】[0017]図３Ｂは、本開示の態様による、例示的なディープ畳み込みネットワーク（ＤＣＮ）を例示するブロック図である。

【図４】[0018]図４は、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例示するブロック図である。

【図５】[0019]図５は、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（run-time operation）を例示するブロック図である。

【図６Ａ】[0020]図６Ａは、本開示の態様による、機械学習モデルのパフォーマンスを改善するために、第１の分類器に第２の分類器を追加するためのバリエーションを例示するブロック図である。

【図６Ｂ】[0020]図６Ｂは、本開示の態様による、機械学習モデルのパフォーマンスを改善するために、第１の分類器に第２の分類器を追加するためのバリエーションを例示するブロック図である。

【図７】[0021]図７は、本開示の態様による、トレーニングされた機械学習モデルのパフォーマンスを改善するための例示的な分類器の概略図である。

【図８】[0022]図８は、本開示の態様による、トレーニングされた機械学習モデルのパフォーマンスを改善するための方法を例示する。

【図９】[0023]図９は、本開示の態様による、トレーニングされた機械学習モデルのパフォーマンスを改善するための方法を例示するブロック図である。

【発明の詳細な説明】

【００１４】

[0024]添付された図面に関連して以下に示される詳細な説明は、様々な構成の説明として意図され、ここで説明される概念が実施され得る唯一の構成を表すようには意図されない。詳細な説明は、様々な概念の完全な理解を提供することを目的とした特定の詳細を含む。しかしながら、これらの概念が、これらの特定の詳細なしで実施され得ることは、当業者にとって明らかであろう。いくつかの事例では、周知の構造およびコンポーネントが、このような概念を曖昧にすることを避けるために、ブロック図形式で示される。

【００１５】

[0025]本教示に基づき、当業者は、本開示の範囲が、本開示のその他任意の態様と組み合わされてインプリメントされようと、あるいは独立してインプリメントされようと、本開示の任意の態様をカバーするように意図されていることを理解すべきである。例えば、示される任意の数の態様を使用して、装置がインプリメントされ得、または方法が実施され得る。加えて、本発明の範囲は、示される本開示の様々な態様に加えて、またはそれ以

外の、他の構造、機能、または構造と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ように意図される。開示される本開示の任意の態様が、請求項の１つまたは複数の要素によって具現化され得ることが理解されるべきである。

【００１６】

[0026]「例示的（exemplary）」という用語は、ここで、「例、事例、または例示を提供する」という意味で使用される。「例示的」であるとしてここで説明される任意の態様は、必ずしも他の態様よりも好ましいまたは有利であるようには解釈されるべきでない。

【００１７】

[0027]特定の態様がここで説明されるが、これらの態様の多くの変形および置換が、本開示の範囲内に含まれる。好ましい態様のいくつかの利益および利点が述べられるが、本開示の範囲は、特定の利益、用途または目的に限定されるようには意図されない。むしろ、本開示の態様は、異なる技術、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ように意図されており、そのうちのいくつかは、図面および好ましい態様の以下の説明において、例として例示される。詳細な説明および図面は、限定ではなく、本開示の単なる例示であり、本開示の範囲は、添付された特許請求の範囲およびそれらの同等物によって定義されている。

【００１８】

[0028]本開示の態様が、トレーニングされたより低い複雑度の機械学習モデルのパフォーマンスを改善することに向けられる。本開示の態様によると、モデルのパフォーマンスは、低い複雑度の分類器の分類誤差の数を直接的に最小化するまたは低減させるように構成された第２の分類器を追加することによって改善され得る。すなわち、標準的な技法（例えば、勾配降下法（gradient descent））を使用して、典型的なコスト関数（例えば、平方和（ＳＳＥ：sum of squares）、または負の対数尤度（negative log likelihood））によって与えられるような誤差の関数（function of errors）を最小化するのではなく、追加された分類器のための新しい目的関数が、誤差の数を直接的に最小化するまたは低減させるように定義される。例えば、７つの正確な分類と３つの不正確な分類と共に、分類動作（classification operations）が実行された場合、目的関数は、３つの誤差をゼロに低減させるように設計され得る。

【００１９】

[0029]加えて、本開示の態様によると、トレーニングされたより低い複雑度の機械学習モデルのパフォーマンスは、より高い複雑度のモデルの軟確率（soft probabilities）を使用してさらに改善され得る。

【００２０】

軟確率

[0030]軟確率は、確率ベクトルの未知の値（dark values）または非最大確率値（non-maximum probability values）である。多くの従来の分類システムでは、確率ベクトルは、クラスラベルを予測するために使用される。このような従来のシステムでは、クラスラベルは、確率ベクトルにおける最高または最大確率値を使用して予測される。非最大確率値または軟確率は、無視される。

【００２１】

[0031]例えば、機械学習モデルM_{λ}（W）が、Ｎ個のサンプルの入力データ$X^{tr}=[x_0, x_1, x_2, \cdots x_{N-1}]$、ここで、

【００２２】

【数１】

$$x_i \in \mathbb{R}^D$$

【００２３】

である、と、対応するＮ個のトレーニングサンプルのＣ個のラベル付けされた出力データ（C-labeled output data）$y^{tr}=[y_0, y_1, y_2, \cdots y_{N-1}]$、ここで、$y_i \in [0, C-1]$である、とから成るトレーニングデータを使用してトレーニングされる、教師あり機械学習の分類問題を考慮する。典型的に、機械学習モデル（例えば、ニュー

ラルネットワーク）のアーキテクチャを定義するパラメータλおよびこのモデルをトレーニングするための学習プロセスのパラメータは、予め決定されている。その後、トレーニングデータ｛X^{tr}，y^{tr}｝は、モデルM_{λ}の重みWを学習するために使用される。このトレーニングは、

【００２４】

【数２】

$$p_j \in \mathbb{Z}_2^C$$

【００２５】

となるように、Ｐ＝［p_0，p_1，…p_{N-1}］に、１－Ｋ符号化（1-K encoding）を使用して、ラベル付けされたデータｙ＝［y_0，y_1，…y_{N-1}］を符号化することを含み得、ここで、y_j＝ｋの場合、p_{jk}＝１であり、

【００２６】

【数３】

$$\sum_{k=0}^{C-1} p_{jk} = 1$$

【００２７】

である。

【００２８】

[0032]入力ｘを与えられると、機械学習モデルM_{λ}は、出力確率に対する推定値（estimate）を生成し、それは、

【００２９】

【数４】

$$\hat{p} = M_{\lambda}(x, W) \qquad (1)$$

【００３０】

として表され得、結果として、

【００３１】

【数５】

$$C = \sum_{i=0}^{N-1} \sum_{j=0}^{C-1} p_{ij} log(\hat{p}_{ij}). \qquad (2)$$

【００３２】

によって与えられるマルチクラス交差エントロピー関数を最小化するようにする。

【００３３】

[0033]出力クラスラベルは、次のように得られる：

【００３４】

【数６】

$$\hat{y} = \underset{j}{argmax}\,[\hat{p}] \qquad (3)$$

【００３５】

[0034]したがって、硬確率（hard-probability）と呼ばれる、ベクトル

【００３６】

【数７】

$$\hat{p}$$

【００３７】

の最大値の指標のみが、推論のために利用され、非最大値は、無視される。

【００３８】

[0035]本開示の態様は、分類パフォーマンスを改善するために軟確率を利用する。いくつかの態様では、軟確率は、温度スケーリング（temperature scaling）を使用して抽出され得る。例えば、ニューラルネットワークモデルによって生成される確率

【００３９】
【数８】

$\hat{p}$

【００４０】
は、次のようなソフトマックス関数（softmax function）を介した推定である：
【００４１】
【数９】

$$\hat{p}_k = \frac{exp\left(a_{out,k}\right)}{\sum_{j=0}^{C-1} exp\left(a_{out,j}\right)} \qquad (4)$$

【００４２】
ここで、$a_{out}=[a_{out,0}, a_{out,1}, \cdots, a_{out,C-1}]$は、ニューラルネットワークの出力ノードから出てきた活性化値（activation values）である。
【００４３】
[0036]トレーニングされた機械学習モデル（例えば、ニューラルネットワーク）によって生成される出力確率は、次のように、軟確率において隠されている情報を抽出するために、温度Ｔによってスケーリングされ得る：
【００４４】
【数１０】

$$\hat{p}_k^{Te} = \frac{exp\left(\frac{a_{out,k}}{T}\right)}{\sum_{j=0}^{C-1} exp\left(\frac{a_{out,j}}{T}\right)} \qquad (5)$$

【００４５】
[0037]１つの目的が、トレーニングされたモデルによって生成される確率ベクトル
【００４６】
【数１１】

$\hat{p}$

【００４７】
の分布を緩和する（soften）ことである。温度Ｔを介してスケーリングすることは、確率の分布を平坦化し（flattens）、それによって、軟確率における情報が活用されることを可能にする。
【００４８】
[0038]いったん抽出されると、軟確率は、分類パフォーマンスを改善するために使用され得る。例えば、b_mおよびW_mが軟確率における情報を共にプーリング（pooling together）するために使用されるバイアスおよび重みのセットを表す一例では、混合確率（mixture probability）が、
【００４９】
【数１２】

$$\tilde{p}^{Te} = \frac{1}{1+exp\left(-(W_m\hat{p}^{Te}+b_m)\right)} \qquad (6)$$

【００５０】
によって与えられ得る。
【００５１】
[0039]混合確率は、次のように、トレーニングされた機械学習モデルによって出力クラスラベルを予測するために使用され得る：
【００５２】
【数１３】

$$\tilde{y} = \underset{j}{argmax}\left[\tilde{p}^{Te}\right] \qquad (7)$$

【００５３】

[0040]トレーニングデータ｛X^{tr}，y^{tr}｝は、軟確率の混合を生成するために使用される重みおよびバイアスについての値を推定するために使用され得る。出力ラベルが硬確率のみを使用して予測されるときの、トレーニングされた機械学習モデルによって生成されるフラクショナルトレーニング誤差（fractional training error）e_d（式３）、および出力ラベルが軟確率を使用して予測されるときのフラクショナルトレーニング誤差（ｅ）（式７）が、以下によって与えられる：

【００５４】

【数１４】

$$e_d = \frac{1}{N}\sum_{j=0}^{N-1} \mathbb{I}_{\hat{y}_j \neq \hat{y}_j} \qquad (8)$$

$$e = \frac{1}{N}\sum_{j=0}^{N-1} \mathbb{I}_{\hat{y}_j \neq \hat{y}_j} \qquad (9)$$

【００５５】

[0041]コスト関数Ｃが、分類誤差を低減させるために使用され得る。すなわち、コスト関数Ｃは、軟確率の混合によって生成された出力ラベルに関する予測された値を使用したときのトレーニングデータに対する誤差（the error on the training data）が、コスト関数が正の非ゼロ値をとる（takes on）確率を使用することによって得られる誤差よりも低くなるように設計され得る。コスト関数は、次のように表され得る：

【００５６】

【数１５】

$$C = max\,(0,(e_d - e)/e_d) \qquad (10)$$

【００５７】

[0042]軟確率の混合についての改善されたまたは最適な重みおよびバイアスが、次の最適化問題を解くことによって得られ得る：

【００５８】

【数１６】

$$\{W_m^*, b_m^*\} = \underset{\{w_M, b_m\}}{argmin}\,[1 - C] \qquad (11)$$

【００５９】

[0043]式１１の最適化問題は、初期状態｛W_m（０），b_m（０）｝＝｛Ｉ，０｝で（with）勾配値を使用しない標準の制約なし最適化プロセス（standard unconstrained optimization processes）のうちの任意のものを使用して解かれ得る。いくつかの態様では、最適化技法はまた、軟確率を生成するための改善されたまたは最適な温度を決定するために用いられ得る。例えば、式１１の最適化問題は、次のように修正され得る：

【００６０】

【数１７】

$$\{T^*, W_m^*\, b_m^*\} = \underset{\{T, W_m, b_m\}}{argmin}\,[1 - C] \qquad (12)$$

【００６１】

[0044]標準の制約なし最小化プロセスを使用することは、ある解をもたらし、それは、温度の初期選択の周辺のＣについての局所極小値である（a local minima for C around the initial choice of the temperature）。収束ストラテジが、温度の初期選択の周辺の局所極小値から脱出するために使用され得る。例えば、いくつかの態様では、このストラテジは、パラメータの初期セット：｛Ｔ（０），W_m（０），b_m（０）｝で始まり、式１１を使用して、重みおよびバイアスについての最適値

【００６２】

【数１８】

$$\left\{W_m^{*,T(0)},\ b_m^{*,T(0)}\right\}$$

【００６３】

を求め得る。初期状態Ｔ′（０）から開始し、コスト関数：Ｃ＝ｍａｘ（０，（ｅ－ｅ′）／ｅ）を最適化し、ここで、ｅは、

【００６４】

【数１９】

$$\left\{T'(0), W_m^{*,T(0)}, b_m^{*,T(0)}\right\}$$

【００６５】

を用いて式１１を使用して計算され、ｅ′は、

【００６６】

【数２０】

$$\left\{T'(0), W_m^{*,T(0)}, b_m^{*,T(0)}\right\}$$

【００６７】

を用いて式１１を使用して計算される。このシーケンスは、収束するまで繰り返され得る。

【００６８】

[0045]いくつかの態様では、アンサンブル平均化（ensemble averaging）が、複数の機械学習モデルにわたって、および／または単一の機械学習モデルにおける複数のロジスティック回帰層（logistic regression layers）にわたって、インプリメントされ得る。一例では、複数の機械学習モデル（Ｍ＞１）が、Ｍ個のトレーニングされたモデルによって生成される出力確率

【００６９】

【数２１】

$$\{\hat{p}_0, \hat{p}_1 \cdots \hat{p}_{M-1}\}$$

【００７０】

を有する（with）トレーニングデータを使用してトレーニングされる。これらのモデルの各々について、軟確率の最適な混合が、上記のプロシージャ最適化技法および/または収束ストラテジを使用して生成され得る。結果として得られる混合確率

【００７１】

【数２２】

$$\{\tilde{p}_0^{Te_0},\ \tilde{p}_1^{Te_1} \cdots \tilde{p}_{M-1}^{Te_{M-1}}\}$$

【００７２】

は、

【００７３】

【数２３】

$$y^{pred} = \underset{j}{\operatorname{argmax}} \left[\sum_k w_k \tilde{p}_k^{Te_k}\right] \qquad (13)$$

【００７４】

として出力ラベルを予測するために使用され得る。

【００７５】

[0046]｛w_k｝についての１つの選択が、ｋ＝（１，２，…Ｍ－１）について、w_k＝１／Ｍである。代替として、上記の最適化技法および収束ストラテジ、または他の同様の技法が、マルチモデル確率混合重み｛w_k｝の最適なセットを推定するために使用され得る。

【００７６】
[0047]別の例では、複数のロジスティック回帰出力層を有してではあるが、単一の機械学習モデルにおいて、最適化技法、収束ストラテジ、および同様のものは、このモデルの異なるロジスティック回帰層から結果として得られる軟確率を改善するまたは最適化するために使用され得る。
【００７７】
[0048]いくつかの態様では、推論は、クラスの数が大きい（例えば、Ｃ≫１）ときに、軟確率を使用して改善され得る。軟確率の最適な混合を生成するためのパラメータの数は、C^2としてスケーリングされ、推論のための軟確率の混合を推定するときに問題になることがある。このケースでは、有用な情報を含むと考えられる各クラスについての最も高い軟確率のサブセットＰ＜＜Ｃが、分類パフォーマンスを改善するために活用され得る。順に、式１１は、推定されるべきパラメータの総数がＰ（Ｐ＋１）となるように、重みおよびバイアスを得るために解かれ得る。推論時間またはおよそ推論時間において、上位Ｐ個の軟確率（the top P soft probabilities）の指標が、追跡され、最適な重みおよびバイアスを使用して推定された混合確率を介して付加され得る。
【００７８】
[0049]図１は、本開示のある特定の態様による、汎用プロセッサ（ＣＰＵ）またはマルチコア汎用プロセッサ（ＣＰＵｓ）１０２を含み得る、システムオンチップ（ＳＯＣ）１００を使用した、前述のトレーニングされた機械学習モデルのパフォーマンスを改善する方法の例となるインプリメンテーションを例示する。変数（例えば、モデルの重み）、計算デバイスに関連付けられたシステムパラメータ（例えば、重みを有する機械学習モデル）、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に関連付けられたメモリブロック中、ＣＰＵ １０２に関連付けられたメモリブロック中、グラフィックス処理ユニット（ＧＰＵ）１０４に関連付けられたメモリブロック中、デジタルシグナルプロセッサ（ＤＳＰ）１０６に関連付けられたメモリブロック中、専用メモリブロック１１８中に記憶され得、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は、ＣＰＵ １０２に関連付けられた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。
【００７９】
[0050]ＳＯＣ １００はまた、ＧＰＵ １０４、ＤＳＰ １０６、接続性ブロック１１０、これは、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、免許不要のＷｉ－Ｆｉ接続性、ＵＳＢ接続性、Ｂｌｕｅｔｏｏｔｈ（登録商標）接続性、および同様のものを含み得る、および、例えば、ジェスチャを検出および認識し得るマルチメディアプロセッサ１１２のような、特定の機能に合わせられた（tailored to）追加の処理ブロックを含み得る。１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。ＳＯＣ １００はまた、センサプロセッサ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。
【００８０】
[0051]ＳＯＣ １００は、ＡＲＭ命令セットに基づき得る。本開示のある態様では、汎用プロセッサ１０２にロードされる命令は、第１の目的関数（例えば、コスト）を有する第１の分類器に、第２の目的関数（例えば、コスト）を有する第２の分類器を追加するためのコードを備え得る。第２の目的関数は、第１の分類器の誤差を直接的に低減させるために使用される。
【００８１】
[0052]図２は、本開示のある特定の態様による、システム２００の例となるインプリメンテーションを例示する。図２に例示されるように、システム２００は、ここで説明される方法の様々な動作を実行し得る複数のローカル処理ユニット（local processing units）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。加えて、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に例示されるように、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提供するための構成プロセッサユニット２１４と、およびローカル処理ユニット２０２間のルーティングを提供するルーティング接続処理ユニット２１６とインタフェースし得る。

【００８２】

[0053]ディープラーニングアーキテクチャは、各層において連続的により高度な抽象化レベル（successively higher levels of abstraction）で入力を表現する（represent）ことを学習することによってオブジェクト認識タスクを実行し得、それにより、入力データの有用な特徴表現を構築（building up）する。このようにして、ディープラーニングは、従来の機械学習の主要なボトルネック（major bottleneck）に対処する。ディープラーニングが出現する前は、オブジェクト認識問題に対する機械学習アプローチは、ことによるとシャロー分類器（shallow classifier）との組合せにおいて、人間によって設計された特徴（human engineered features）に依存するところが大きかった。シャロー分類器は、２クラス線形分類器であり得、例えば、そこで、特徴ベクトル成分の加重和（weighted sum）が、どのクラスに入力が属するか予測するためにしきい値と比較され得る。人間によって設計された特徴は、領域の専門知識を有するエンジニアによって、特定の問題領域に合わせられたテンプレートまたはカーネルであり得る。ディープラーニングアーキテクチャは、対照的に、トレーニングを通じてであるが、人間のエンジニアが設計し得るものと同様の特徴を表現することを学習し得る。さらに、ディープネットワークは、人間が考慮することがなかったであろう新しいタイプの特徴を表現および認識することを学習し得る。

【００８３】

[0054]ディープラーニングアーキテクチャは、特徴の階層を学習し得る。例えば、視覚データが提示された場合、第１の層は、入力ストリームにおける、エッジのような、比較的単純な特徴を認識することを学習し得る。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パワーを認識することを学習し得る。第１の層の出力を入力として受ける第２の層は、視覚データについては単純な形状または聴覚データについては音の組合せのような、特徴の組合せを認識することを学習し得る。例えば、上位層は、視覚データにおける複雑な形状または聴覚データにおける単語を表現することを学習し得る。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ことを学習し得る。

【００８４】

[0055]ディープラーニングアーキテクチャは、自然階層構造を有する問題に適用されるとき、特によく機能し得る。例えば、モーターの付いた乗り物の分類は、車輪、風防ガラス、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され得る。

【００８５】

[0056]ニューラルネットワークのような機械学習モデルは、様々な結合パターン（connectivity patterns）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、情報は、下位層から上位層へ渡されるとともに、所与の層における各ニューロンが、上位層におけるニューロンに伝達する。階層的な表現が、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した層において構築され得る。ニューラルネットワークはまた、再帰型（recurrent）結合またはフィードバック（トップダウンとも呼ばれる）結合を有し得る。再帰型結合では、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伝達され得る。再帰型アーキテクチャは、シーケンスにおいてニュー

ラルネットワークに伝達される入力データチャンクのうちの１つよりも多くにまたがる（span）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。多くのフィードバック結合を有するネットワークは、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区別することを支援し得るときに役立ち得る。

【００８６】

[0057]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の層における全てのニューロンから入力を受信するように、その出力を第２の層における全てのニューロンに伝達し得る。代替として、局所結合されたネットワーク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、局所結合され得、第２の層における各ニューロンのための入力に関連付けられた結合強度が共有されるようにさらに構成される（例えば、３０８）。より一般的には、ネットワークの局所結合された層は、層における各ニューロンが、異なる値を有し得る結合強度を持ってではあるが、同じまたは同様の結合パターンを有するように構成され得る（例えば、３１０、３１２、３１４、および３１６）。局所結合された結合パターンは、所与の領域における上位層のニューロンが、ネットワークへの総入力の制限された部分の特性（properties）にトレーニングを通じて調整される入力を受信し得るので、上位層における空間的に別個の受容野（spatially distinct receptive fields）を生じさせ得る。

【００８７】

[0058]局所結合されたニューラルネットワークは、入力の空間的ロケーションが意味をもつ問題によく適し得る。例えば、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、異なる特性を有する上位層のニューロンを、画像の下部対上部（the lower versus the upper portion）とのそれらの関連付けに依存して発達（develop）させ得る。例えば、画像の下部に関連付けられたニューロンは、車線区分線を認識することを学習し得、一方、画像の上部に関連付けられたニューロンは、交通信号、交通標識、および同様のものを認識することを学習し得る。

【００８８】

[0059]ＤＣＮは、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、ＤＣＮは、速度制限標識のクロップされた画像（cropped image）３２６のような、画像を提示され得、その後、「フォワードパス（forward pass）」が、出力３２２を生成するために計算され得る。出力３２２は、「標識」、「６０」、および「１００」のような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、例えば、トレーニングされたネットワーク３００についての出力３２２において示される「標識」および「６０」に対応するもののような、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について、高いスコアを出力することを望み得る。トレーニング前は、ＤＣＮによって生成される出力は、不正確である可能性が高く、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。その後、ＤＣＮの重みは、ＤＣＮの出力スコアがターゲットにより密接に合わせられる（aligned）ように調整され得る。

【００８９】

[0060]重みを調整するために、学習アルゴリズムが、重みについての勾配ベクトルを計算し得る。勾配は、重みがわずかに調整された場合に、誤差が増大または低減するであろう量を示し得る。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。下位層において、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。その後、重みは、誤差を低減させるように調整され得る。重みを調整するこの方法は、それがニューラルネットワークを通じた「バックワードパス（backward pass）」を伴うので、「バックプロパゲーション」と呼ばれ得る。

【００９０】

[0061]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（true error gradient）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、確率的勾配降下法（stochastic gradient descent）と呼ばれ得る。確率的勾配降下法は、システム全体の達成可能な誤差率（error rate）の低減が止まるまで、または誤差率がターゲットレベルに達するまで繰り返され得る。

【００９１】

[0062]学習後、ＤＣＮは、新しい画像３２６を提示され得、ネットワークを通じたフォワードパスが、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。

【００９２】

[0063]ディープビリーフネットワーク（ＤＢＮ）は、隠れノードの複数の層（multiple layers of hidden nodes）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的な表現（hierarchical representation）を抽出するために使用され得る。ＤＢＮは、制限付きボルツマンマシン（ＲＢＭ：Restricted Boltzmann Machines）の層を積み上げること（stacking up）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、各入力がカテゴリ化されるべきクラスについての情報がない状態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の下方のＲＢＭ（bottom RBM）は、教師なしの方法でトレーニングされ得、かつ特徴抽出器として機能し得、また、上方のＲＢＭ（top RBM）は、（ターゲットクラスおよび前の層からの入力の同時分布（joint distribution）で）教師ありの方法でトレーニングされ得、かつ分類器として機能し得る。

【００９３】

[0064]ディープ畳み込みネットワーク（ＤＣＮ）は、畳み込みネットワークのネットワークであり、追加のプーリング層および正規化層を用いて構成される。ＤＣＮは、多くのタスクについて最先端のパフォーマンスを達成している。ＤＣＮは、入力ターゲットおよび出力ターゲットの両方が、多くのエグゼンプラー（exemplars）について知られており、かつ勾配降下法の使用によってネットワークの重みを修正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ることができる。

【００９４】

[0065]ＤＣＮは、フィードフォワードネットワークであり得る。加えて、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結合および共有結合は、高速処理に活用され得る。ＤＣＮの計算負担は、例えば、再帰型結合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、はるかに少なくなり得る。

【００９５】

[0066]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射影（basis projection）であると見なされ得る。入力が、最初にカラー画像の赤、緑、および青チャネルのような複数のチャネルに分解される場合には、その入力で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の次元とを有する、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得るとともに、特徴マップ（例えば、３２０）の各要素が、複数のチャネルの各々から、および前の層（例えば、３１８）におけるある範囲のニューロン（a range of neurons）から入力を受信する。特徴マップにおける値は、正規化（rectification）、すなわちｍａｘ（０，ｘ）のような、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、さらにプーリングされ得、これは、ダウンサンプリングに対応し、追加の局所不変性および次元削減（dimensionality reduction）を提供し得る。白色化に対応する正規化がまた、特徴マップにおけるニューロン間の側抑制（lateral inhibition）を通じて適用され得る。

【００９６】
[0067]ディープラーニングアーキテクチャのパフォーマンスは、より多くのラベル付けされたデータポイントが利用可能になるにつれて、または計算能力が増大するにつれて増大し得る。現代のディープニューラルネットワークは、わずか１５年前に一般的な研究者に利用可能であったものよりも何千倍も大きいコンピューティングリソースを用いて、日常的に（routinely）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は、ディープラーニングのパフォーマンスをさらに高め得る。整流された線形ユニットは、勾配消失（vanishing gradients）として知られるトレーニング課題を低減し得る。新しいトレーニング技法は、過剰適合（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象化（abstarct）し、全体的なパフォーマンスをさらに高め得る。

【００９７】
[0068]図３Ｂは、例示的なディープ畳み込みネットワーク３５０を例示するブロック図である。ディープ畳み込みネットワーク３５０は、結合性および重みの共有に基づいて、複数の異なるタイプの層を含み得る。図３Ｂに示されるように、例示的なディープ畳み込みネットワーク３５０は、複数の畳み込みブロック（例えば、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、正規化層（ＬＮｏｒｍ）、およびプーリング層で構成され得る。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するものではなく、代わりに、設計の選好に従って、任意の数の畳み込みブロックがディープ畳み込みネットワーク３５０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。例えば、正規化層は、白色化または側抑制を提供し得る。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提供し得る。

【００９８】
[0069]ディープ畳み込みネットワークの、例えば、並列フィルタバンク（parallel filter banks）は、高いパフォーマンスおよび低い電力消費を達成するために、オプションとしてＡＲＭ命令セットに基づいて、ＳＯＣ １００のＣＰＵ １０２またはＧＰＵ １０４上に搭載（loaded on）され得る。代替的な実施形態では、並列フィルタバンクは、ＳＯＣ １００のＤＳＰ １０６またはＩＳＰ １１６上に搭載され得る。加えて、ＤＣＮは、センサ１１４およびナビゲーション１２０に専用の処理ブロックのような、ＳＯＣ上に存在し得る他の処理ブロックにアクセスし得る。

【００９９】
[0070]ディープ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（例えば、ＦＣ１およびＦＣ２）を含み得る。ディープ畳み込みネットワーク３５０は、ロジスティック回帰（ＬＲ：logistic regression）層をさらに含み得る。ディープ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（例えば、画像、音声、ビデオ、センサデータ、および／または他の入力データ）から階層的な特徴表現を学習するために、ディープ畳み込みネットワーク３５０における後続の層の入力として機能し得る。

【０１００】
[0071]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を例示するブロック図である。このアーキテクチャを使用して、ＳＯＣ ４２０の様々な処理ブロック（例えば、ＣＰＵ ４２２、ＤＳＰ ４２４、ＧＰＵ ４２６および／またはＮＰＵ ４２８）に、アプリケーション４０２のランタイム動作中にサポート計算（supporting computations）を実行させ得るアプリケーション４０２が設計され得る。

【０１０１】

[0072]ＡＩアプリケーション４０２は、例えば、そこで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もたらし（provide for）得る、ユーザ空間４０４において定義された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、例えば、認識されたシーンがオフィス、講堂、レストラン、または湖のような屋外環境であるかどうかに依存して異なるように、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提供するために、ＳｃｅｎｅＤｅｔｅｃｔアプリケーションプログラミングインタフェース（ＡＰＩ）４０６において定義されたライブラリに関連付けられた、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、例えば、ビデオおよび測位データに基づいてシーン推定を提供するように構成されたディープニューラルネットワークの出力に最終的に依拠し得る。

【０１０２】

[0073]ランタイムフレームワーク（Runtime Framework）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にとってさらにアクセス可能であり得る。ＡＩアプリケーション４０２は、例えば、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る、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、順に、ＳＯＣ ４２０上で実行中のＬｉｎｕｘ（登録商標）カーネル４１２のような、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、順に、ＣＰＵ ４２２、ＤＳＰ ４２４、ＧＰＵ ４２６、ＮＰＵ ４２８、またはこれらの何らかの組合せ上で、計算を実行させ得る。ＣＰＵ ４２２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、ＤＳＰ ４２４のための、ＧＰＵ ４２６のための、またはＮＰＵ ４２８のためのドライバ４１４～４１８のような、ドライバを通じてアクセスされ得る。例示的な例では、ディープニューラルネットワークは、ＣＰＵ ４２２およびＧＰＵ ４２６のような、処理ブロックの組合せ上で実行するように構成され得るか、または、存在する場合、ＮＰＵ ４２８上で実行され得る。

【０１０３】

[0074]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を例示するブロック図である。ＡＩアプリケーションは、画像５０６のフォーマットを変換し、その後、画像５０８をクロップおよび／またはリサイズするように（例えば、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。その後、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（例えば、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を含む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。例えば、画像は、結果として得られる画像が２２４ピクセル×２２４ピクセルとなるように、スケーリングされ、クロップされ得る。これらの次元（dimensions）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、ＳＯＣ １００の様々な処理ブロックに、ディープ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、ディープ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。その後、ディープニューラルネットワークの結果は、しきい値処理（thresholded）５２２され、分類アプリケーション５１０内の指数平滑化（exponential smoothing）ブロック５２４を通され得る。その後、平滑化された結果は、スマートフォン５０２の設定および／または表示の変更を生じ得る。

【０１０４】

[0075]一構成では、機械学習モデルが、第１の目的（例えば、コスト）関数を有する第１の分類器に、第２の目的（例えば、コスト）関数を有する第２の分類器を追加するために構成され、第２の目的関数は、第１の分類器の誤差を直接的に低減させるために使用される。機械学習モデルはまた、トレーニングされた機械学習モデルを介して受信される入

力に基づいて、第２の分類器から特徴ベクトルを出力するために構成される。機械学習モデルは、追加する手段および／または出力する手段を含む。一態様では、追加する手段および／または出力する手段は、記載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に関連付けられたプログラムメモリ、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、上述された手段は、これら上述された手段によって、記載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。

【０１０５】

[0076]本開示のある特定の態様によると、各ローカル処理ユニット２０２は、ネットワークの所望の１つまたは複数の機能的特徴に基づいて、ネットワークのパラメータを決定することと、決定されたパラメータがさらに適合、調整、および更新されるにつれて、所望の機能的特徴に向けて１つまたは複数の機能的特徴を発展させることと、を行うように構成され得る。

【０１０６】

[0077]図６Ａおよび図６Ｂは、ニューラルネットワークモデルのような機械学習モデルのパフォーマンスを改善するために、第１の分類器に第２の分類器を追加するためのバリエーションを例示するブロック図である。図６Ａおよび図６Ｂを参照すると、第２の分類器６０２が、トレーニングされた機械学習モデル６０６の第１の分類器６０４に追加され得る。いくつかの態様では、機械学習モデル６０６は、局所結合された（Ｌ－Ｃ）層を含むディープ畳み込みネットワーク（ＤＣＮ）または別の機械学習モデルを備え得る。この機械学習モデルは、複雑度が低くなり得る。いくつかの例示的な態様では、１０億回未満の積和演算（ＭＡＣ）を有する機械学習モデルが、低い複雑度のモデルであると考えられ得る。他方では、１０億回を上回る積和演算を有する機械学習モデルが、高い複雑度のモデルであると考えられ得る。むろん、他の測定基準もまた、モデルの相対的な複雑度を決定するために使用され得る（例えば、パラメータの数、ステージ（層）の数、および／またはステージのタイプ）。

【０１０７】

[0078]トレーニングされた機械学習モデル６０６は、入力（例えば、画像）（図示せず）を受信するように構成され得る。機械学習モデル６０６は、入力から特徴のセットを抽出するために、画像を処理し得る。入力に対応する特徴ベクトルが、第１の分類器６０４に供給され得る。第１の分類器６０４は、微分可能な（例えば、勾配が決定できる）目的関数を用いて構成され得、これは、分類精度（classification accuracy）を改善するために使用され得る。順に、第１の分類器６０４は、出力クラスラベルを決定するために使用され得る確率ベクトルP_cを生成し得る。

【０１０８】

[0079]第１の分類器６０４のパフォーマンスおよび精度を改善するために、第２の分類器６０２が追加され得る。第２の分類器６０２は、微分不可能な（non-differentiable）（例えば、勾配がない）目的関数を用いて構成され得る。目的関数は、第１の分類器６０４によって生成される誤差の数を直接的に低減させるように構成され得る。すなわち、第１の分類器６０４のためのコスト関数または誤差の関数を最小化するように試みるのではなく、第２の分類器６０２は、誤差の総数を低減させる。例えば、いくつかの態様では、第２の分類器６０２のための目的関数は、

【０１０９】

【数２４】

$$Objective\ function: argmax\ [max(0,(e_d - e))] \qquad (14)$$

【０１１０】

として表され得る。

【０１１１】

[0080]目的関数は、上記で説明されたような制約なし最小化技法を使用して、第２の分類器６０２のための重みおよびバイアスの項を決定するために使用され得る。したがって

、第２の分類器６０２からの出力クラスラベルは、第２の分類器６０２のみを介して生成されるよりも少ない誤差を含み得る。

【０１１２】

[0081]この構成は、分類パフォーマンスにおける改善が、予めトレーニングされた機械学習モデルを再トレーニングすることなく達成され得るので、特に有益であり得る。代わりに、パフォーマンスは、第２の分類器６０２を再トレーニングすることのみによって改善され得る。

【０１１３】

[0082]いくつかの態様では、図６Ｂに示されるように、第２の分類器６０２は、代替として、（例えば、トレーニングされた機械学習モデルからのモデルの層として）トレーニングされた機械学習モデル６０６内に提供され得る。さらに、いくつかの態様では、（図６Ａおよび図６Ｂに示される）機械学習モデル６０６のパフォーマンスは、高い複雑度のモデル６０８を介して供給される軟確率を使用してさらに改善され得る。

【０１１４】

[0083]図７は、本開示の態様による、トレーニングされた機械学習モデル（例えば、ニューラルネットワーク）のパフォーマンスを改善するための例示的な分類器７００の概略図を提示する。図７を参照すると、微分不可能な目的関数Ｏが、ニューラルネットワークの分類器（回帰）層の出力において追加される。目的関数は、所与のトレーニング（またはテスティング）データセットに対する目的関数についての最大の非ゼロ値が、トレーニング（テスティング）誤差の数が、元のトレーニングされたニューラルネットワークについて得られたそれらを下回るときのみ生じるであろうように指定（specified）され得る。

【０１１５】

[0084]入力

【０１１６】

【数２５】

$$X \in \mathbb{R}^D$$

【０１１７】

を与えられると、機械学習モデル７０２は、入力をＣ個のクラスのうちの１つに分類するように構成され得る。ワンホット符号化（one-hot encoding）のような、符号化スキームを使用して、クラスラベルは、所与のクラスラベルｌ＜Ｃについて、Ｐ＝［$p_1\ p_2 \cdots p_C$］^{T}であるように、確率ベクトル

【０１１８】

【数２６】

$$P \in \mathbb{Z}_2^C$$

【０１１９】

によって表され得、ここで、ｉ＝ｌの場合、$p_i = 1$であり、

【０１２０】

【数２７】

$$\sum_{i=2}^{C} p_i = 1$$

【０１２１】

である。トレーニングされた機械学習モデル（例えば、ニューラルネットワーク）

【０１２２】

【数２８】

$$M: X \in \mathbb{R}^D \rightarrow Z \in \mathbb{R}^C$$

【０１２３】

を与えられると、推定された確率ベクトル

【０１２４】

【数２９】

$$\hat{P}$$

【０１２５】
が、
【０１２６】
【数３０】

$$\hat{P} = \sigma(Z)$$

【０１２７】
として、Ｚから得られ得、ここで、σは、ソフトマックス非線形性（soft-max nonlinearity）である。
【０１２８】
[0085]上記で説明されたように、従来のアプローチは、
【０１２９】
【数３１】

$$\hat{l} = \operatorname{argmax}[\hat{P}]$$

【０１３０】
としてクラスラベルを予測するために、
【０１３１】
【数３２】

$$\hat{P}$$

【０１３２】
を使用する。Ｕ個のトレーニングサンプルを有する所与のデータセットについて、その後、トレーニング誤差は、
【０１３３】
【数３３】

$$e_d^{t_r} = \frac{1}{U}\sum_{i=1}^{U} \mathbb{I}_{\hat{l}\neq l}$$

【０１３４】
として得られ、Ｖ個のテストサンプルに対するテスティング誤差は、同様に、
【０１３５】
【数３４】

$$e_d^{t_s} = \frac{1}{V}\sum_{i=1}^{V} \mathbb{I}_{\hat{l}\neq l}$$

【０１３６】
として得られる。
【０１３７】
【数３５】

$$t_r^d$$

【０１３８】
および
【０１３９】
【数３６】

$$t_s^d$$

【０１４０】
についての値は、モデルＭの優良性（goodness）または精度を決定する。トレーニングモデルＭについての１つの優良性または精度の測定基準が、
【０１４１】

【数３７】

$$e_d^{t_r} = 0$$

【０１４２】

および

【０１４３】

【数３８】

$$e_d^{t_s} << 1$$

【０１４４】

である。本開示の態様は、それに対して

【０１４５】

【数３９】

$$e_d^{t_r} \neq 0$$

【０１４６】

である、トレーニングされたモデルＭのパフォーマンスを改善することを目的とする。

【０１４７】

[0086]本開示の態様によると、トレーニングされたモデル７０２を介して生成される特徴表現が、分類器７００に供給され得る。分類器７００は、特徴ベクトルＺを受信し、それは、モデルの重みW_Zと混合（mixed）されて、新しい特徴ベクトル

【０１４８】

【数４０】

$$Z_s = W_z^T Z$$

【０１４９】

を生成し得る。その後、特徴ベクトルZ_sは、確率ベクトル$P_s = \sigma(Z_s)$を推定するために使用され得る。その後、確率特徴ベクトル

【０１５０】

【数４１】

$$P_f = W_p^T P_s$$

【０１５１】

が、

【０１５２】

【数４２】

$$e^{tr} = \frac{1}{U} = \sum_{i=l}^{U} \mathbb{I}_{\hat{l}_f \neq l},$$

【０１５３】

ここで、

【０１５４】

【数４３】

$$\hat{l}_f = argmax\,[P_f]$$

【０１５５】

である、としてトレーニングセットに対する推定予測誤差を計算するために使用され得る。パラメータλ＝［W_z，W_p］は、次の目的関数上で最適化することによって推定される：

【０１５６】

【数４４】

$$O = MAX\left(0, (e_d^{tr} - e^{tr})\right) \qquad (14)$$

【０１５７】

[0087]いくつかの態様では、高い複雑度のモデル７０４が、機械学習モデル７０２に軟確率ベクトルP_Hを提供し得る。軟確率ベクトルは、モデルの重みW_hと混合され得る。順に、確率特徴ベクトル

【０１５８】

【数４５】

$$P_f = W_p^T P_s + W_h^T P_H$$

【０１５９】

は、

【０１６０】

【数４６】

$$e^{tr} = \frac{1}{U} = \sum_{i=l}^{U} \mathbb{I}_{\hat{l}_f \neq l},$$

【０１６１】

ここで、

【０１６２】

【数４７】

$$\hat{l}_f = argmax\,[P_f]$$

【０１６３】

である、としてトレーニングセットに対する推定予測誤差を計算するために使用され得る。パラメータλ＝［W_z，W_p，W_h，Ｔ］は、式１４の目的関数上で最適化することによって推定され得る。

【０１６４】

[0088]Ｏが微分不可能な関数であると仮定すると、制約なし最小化プロセスは、λ＊＝ａｒｇｍａｘ［Ｏ］として、最適なλ＊を求めるために使用され得る。Ｏに関する非ゼロ収束値は、

【０１６５】

【数４８】

$$e_d^{tr} < e^{tr}$$

【０１６６】

ということ暗示し、したがって、パラメータの追加のセットを推定するという代償を払って、元のモデルよりも良いパフォーマンスを有する、結果として得られるモデルを生成する。

【０１６７】

[0089]いくつかの態様では、λ（例えば、W_z、W_p、W_h、またはＴ）におけるパラメータのうちのいくつかは、アプリオリに設定され得る。したがって、新しいパラメータのうちのいくつかの追加による過剰適合（overfitting）の問題が、緩和または低減され得る。

【０１６８】

[0090]いくつかの態様では、様々な単純化が、設計の選好に従ってパフォーマンスを改善しながら用いられ得る。例えば、トレーニングされた学習モデルによって生成される特徴に対応する重みが、アイデンティティ値（identity value）に設定され得る。したがって、トレーニングされた機械学習モデルによって生成される特徴ベクトルの混合は、考慮されないであろう。一方、第２の例では、トレーニングされた機械学習モデルを介して生成される特徴ベクトルの混合のみが考慮され得る。

【０１６９】
[0091]第３の例では、トレーニングされた学習モデルによって生成される特徴に対応する重みが、アイデンティティ値に設定され得、高い複雑度のモデル７０４からの利用可能な軟確率情報は、無視され得る。
【０１７０】
[0092]第４の例では、高い複雑度のモデル７０４からの軟確率P_Hが、所与の温度値（例えば、Ｔ＝α，α＞１）によって再スケーリング（rescaled）され得る。
【０１７１】
[0093]図８は、トレーニングされた機械学習モデルのパフォーマンスを改善するための方法８００を例示する。ブロック８０２において、プロセスは、第１の目的関数（例えば、コスト）を有する第１の分類器に、第２の目的関数（例えば、コスト）を有する第２の分類器を追加する。第２の目的関数は、第１の分類器の誤差を直接的に低減させるために使用される。
【０１７２】
[0094]第１の目的関数は、微分可能であり、第２の目的関数は、微分不可能である。いくつかの態様では、第２の目的関数は、第１の分類器と第２の分類器との誤差間の差の関数であり得る。他の態様では、第２の目的関数は、より高い複雑度のモデルからの確率の混合に基づいて決定され得る。
【０１７３】
[0095]いくつかの態様では、第２の分類器は、第１の分類器の外部に（externally）追加され得る。代替として、第２の分類器は、第１の分類器（例えば、第１の分類器の層）の内部に組み込まれ得る。さらに、第２の分類器は、第１の分類器を再トレーニングすることなく追加され得る。
【０１７４】
[0096]ブロック８０４において、プロセスは、トレーニングされた機械学習モデルを介して受信される入力に基づいて、第２の分類器から特徴ベクトルを出力する。
【０１７５】
[0097]いくつかの態様では、プロセスは、過剰適合の問題を低減または緩和するために、様々な単純化をインプリメントし得る。例えば、プロセスは、アイデンティティ値に、第１の分類器によってトレーニングされたモデルによって生成される特徴に対する重みを割り当て（assign）得る。プロセスはまた、ゼロに、より高い複雑度のモデルの確率ベクトルによって生成される特徴に対する重みを割り当て得る。プロセスは、第２の分類器の確率ベクトルによって生成される特徴に対する重みをさらに割り当て得る。プロセスは、ゼロに、より高い複雑度のモデルの確率ベクトルによって生成される特徴に対する重みを割り当て得る。プロセスは、第２の分類器の確率ベクトルによって生成される特徴に対する重みをさらに割り当て得る。プロセスはまた、固定された温度Ｔによって、より高い複雑度のモデルによって生成される確率ベクトルをスケーリングし得る。
【０１７６】
[0098]図９は、本開示の態様による、トレーニングされた機械学習モデルのパフォーマンスを改善するための方法９００を例示するブロック図である。ブロック９０２において、プロセスは、トレーニングされた機械学習モデルを介して、機械確率ベクトルを機械学習モデル（例えば、分類器）において受信する。確率ベクトルは、トレーニングされた機械学習モデルにおいて受信される入力に対応する。ブロック９０４において、モデルの重みおよびバイアスのような、機械学習モデルのパラメータは、トレーニングされた機械学習モデルの誤差を直接的に低減させる目的関数に基づいて計算され得る。すなわち、目的関数は、トレーニングされた機械学習モデルについての誤差の関数ではなく、誤差の数を直接的に低減させるように設計される。したがって、機械学習モデルの目的関数は、微分不可能である。
【０１７７】
[0099]いくつかの態様では、高い複雑度のモデルおよび／またはトレーニングされた機

械学習モデルからの軟確率は、パラメータを計算するために使用され得る。

【０１７８】

[00100]ブロック９０６において、プロセスは、機械学習モデルのパラメータを更新し得る。その後、機械学習モデルは、ブロック９０８において、受信された確率ベクトルに対応する入力についての出力クラスラベルを生成し得る。したがって、更新の後に続く分類誤差は、同じ入力についてトレーニングされた機械学習モデルによって生成されるそれらよりも少なくなり得る。したがって、トレーニングされた機械学習モデルのパフォーマンスは、改善され得る。

【０１７９】

[00101]上記で説明された方法の様々な動作は、対応する機能を実行することが可能な任意の適切な手段によって実行され得る。これら手段は、それに限定されるものではないが、回路、特定用途向け集積回路（ＡＳＩＣ）、またはプロセッサを含む、様々なハードウェアおよび／またはソフトウェアの（１つまたは複数の）コンポーネントおよび／または（１つまたは複数の）モジュールを含み得る。一般に、図中に例示された動作がある場合、これらの動作は、同様の番号付けを有する、対応する対をなすミーンズプラスファンクションのコンポーネントを有し得る。

【０１８０】

[00102]ここで使用される場合、「決定すること」という用語は、幅広いアクションを包含する。例えば、「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（例えば、表、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）、確定すること、および同様のことを含み得る。加えて、「決定すること」は、受信すること（例えば、情報を受信すること）、アクセスすること（例えば、メモリにおけるデータにアクセスすること）、および同様のことを含み得る。さらに、「決定すること」は、解決すること、選択すること、選ぶこと、確立すること、および同様のことを含み得る。

【０１８１】

[00103]ここで使用される場合、アイテムのリスト「のうちの少なくとも１つ」を指す表現は、単一のメンバ（members）を含む、それらのアイテムの任意の組合せを指す。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をカバーするように意図される。

【０１８２】

[00104]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、デジタルシグナルプロセッサ（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ロジック、個別ハードウェアコンポーネント、あるいはここで説明された機能を実行するように設計されたこれらの任意の組合せを用いてインプリメントまたは実行され得る。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、このプロセッサは、任意の商業的に利用可能なプロセッサ、コントローラ、マイクロコントローラまたはステートマシンであり得る。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、例えば、ＤＳＰとマイクロプロセッサの組合せ、複数のマイクロプロセッサ、ＤＳＰコアと連携した１つまたは複数のマイクロプロセッサ、あるいはその他任意のこのような構成としてインプリメントされ得る。

【０１８３】

[00105]本開示に関連して説明されたアルゴリズムまたは方法のステップは、直接ハードウェアにおいて、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において、またはこれら２つの組合せにおいて、具現化され得る。ソフトウェアモジュールは、当該技術分野で知られている任意の形態の記憶媒体内に存在し得る。使用され得る記憶媒体のいくつかの例は、ランダムアクセスメモリ（ＲＡＭ）、読取専用メモリ（ＲＯＭ）、フラッシュメモリ、消去可能なプログラマブル読取専用メモリ（ＥＰＲＯＭ）、電気的に消去可

能なプログラマブル読取専用メモリ（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、等を含む。ソフトウェアモジュールは、単一の命令、または多くの命令を備え得、いくつかの異なるコードセグメントにわたって、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、プロセッサが記憶媒体から情報を読み取り、および／または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、プロセッサと一体化され得る。

【０１８４】

[00106]ここ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。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置き換えられ得る。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明記されない限り、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修正され得る。

【０１８５】

[00107]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、またはこ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、例となる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アプリケーションおよび全体的な設計制約に依存して、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、機械可読媒体、およびバスインタフェースを含む様々な回路を共にリンクし得る。バスインタフェースは、特に、バスを介してネットワークアダプタを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。ある特定の態様では、ユーザインタフェース（例えば、キーパッド、ディスプレイ、マウス、ジョイスティック、等）がまた、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、電圧レギュレータ、電力管理回路、および同様のもののような、様々な他の回路をリンクし得、これらは、当該技術分野において周知であり、したがって、これ以上は説明されない。

【０１８６】

[00108]プロセッサは、バスの管理と、機械可読媒体上に記憶されたソフトウェアの実行を含む汎用処理とを担い得る。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その他の回路を含む。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、またはその他の方法で呼ばれるかにかかわらず、命令、データ、またはこれらの任意の組合せを意味するように広く解釈されるべきである。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、読取専用メモリ（ＲＯＭ）、プログラマブル読取専用メモリ（ＰＲＯＭ）、消去可能なプログラマブル読取専用メモリ（ＥＰＲＯＭ）、電気的に消去可能なプログラマブル読取専用メモリ（ＥＥＰＲＯＭ）、レジスタ、磁気ディスク、光ディスク、ハードドライブ、またはその他任意の適切な記憶媒体、あるいはこ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具現化され得る。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。

【０１８７】

[00109]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、当業者が容易に理解するであろうように、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、伝送路（transmission line）、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、これら全ては、バスインタフェースを通じてプロセッサによってアクセスされ得る。代替として、またはこれに加えて、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

スタファイルでのケースでそうであり得るように、プロセッサに組み込まれ得る。ローカルコンポーネントのような、説明された様々なコンポーネントは、特定のロケーションを有するものとして説明されているが、それらはまた、分散型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ある特定のコンポーネントように、様々な方法で構成され得る。

【０１８８】

[00110]処理システムは、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分を提供する外部メモリとを有し、全てが外部バスアーキテクチャを通じて他のサポート回路と共にリンクされている、汎用処理システムとして構成され得る。代替として、処理システムは、ここで説明されたニューロンモデルおよびニューラルシステムのモデル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、プロセッサと、バスインタフェースと、ユーザインタフェースと、サポート回路と、単一のチップに組み込まれた機械可読媒体の少なくとも一部分とを有する特定用途向け集積回路（ＡＳＩＣ）を用いて、または、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、ステートマシン、ゲート論理、個別ハードウェアコンポーネント、またはその他任意の適切な回路、あるいは本開示全体にわたって説明された様々な機能を実行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。当業者であれば、特定のアプリケーションおよびシステム全体に課せられる全体的な設計制約に依存して、処理システムに関して説明された機能をインプリメントするのに最良の方法を認識するであろう。

【０１８９】

[00111]機械可読媒体は、多数のソフトウェアモジュールを備え得る。これら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ると、様々な機能を処理システムに実行させる命令を含む。これらソフトウェアモジュールは、送信モジュールおよび受信モジュールを含み得る。各ソフトウェアモジュールは、単一の記憶デバイス内に存在し得るか、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、ソフトウェアモジュールは、トリガリングイベントが生じたときに、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、アクセス速度を増大させるために、命令のうちのいくつかをキャッシュにロードし得る。その後、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するとき、プロセッサによってインプリメントされるということが理解されるであろう。さらに、本開示の態様が、プロセッサ、コンピュータ、機械、またはこのような態様をインプリメントする他のシステムの機能に改善をもたらすことが理解されるべきである。

【０１９０】

[00112]ソフトウェアでインプリメントされる場合、これら機能は、コンピュータ可読媒体上で、１つまたは複数の命令またはコードとして送信または記憶され得る。コンピュータ可読媒体は、１つの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容易にする任意の媒体を含む通信媒体とコンピュータ記憶媒体との両方を含む。記憶媒体は、コンピュータによってアクセスされることができ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、こ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記憶装置、磁気ディスク記憶装置またはその他の磁気記憶デバイス、あるいは、データ構造または命令の形式で所望のプログラムコードを記憶または搬送するために使用されることができ、かつ、コンピュータによってアクセスされることができるその他任意の媒体を備えることができる。また、任意の接続は、厳密にはコンピュータ可読媒体と称される。例えば、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、およびマイクロ波のようなワイヤレス技術を使用して、ウェブサイト、サーバ、また

は他の遠隔ソースから送信される場合には、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、およびマイクロ波のようなワイヤレス技術は、媒体の定義に含まれる。ここで使用される場合、ディスク（disk）およびディスク（disc）は、コンパクトディスク（ＣＤ）、レーザーディスク（登録商標）、光ディスク、デジタル多目的ディスク（ＤＶＤ）、フロッピー（登録商標）ディスク、およびブルーレイ（登録商標）ディスクを含み、ここでディスク（disks）は、通常磁気的にデータを再生し、一方ディスク（discs）は、レーザーを用いて光学的にデータを再生する。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、非一時的なコンピュータ可読媒体（例えば、有形媒体）を備え得る。加えて、他の態様では、コンピュータ可読媒体は、一時的なコンピュータ可読媒体（例えば、信号）を備え得る。上記の組合せもまた、コンピュータ可読媒体の範囲内に含まれるべきである。

【０１９１】

[00113]したがって、ある特定の態様は、ここで提示された動作を実行するためのコンピュータプログラム製品を備え得る。例えば、このようなコンピュータプログラム製品は、その上に命令が記憶された（および／または符号化された）コンピュータ可読媒体を備え得、これら命令は、ここ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。ある特定の態様では、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。

【０１９２】

[00114]さらに、ここ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、適宜、ユーザ端末および／または基地局によって、ダウンロードされ得ること、および／または、別の方法で取得され得ることが理解されるべきである。例えば、このようなデバイスは、ここで説明された方法を実行するための手段の転送を容易にするためにサーバに結合されることができる。代替として、ここ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、デバイスに記憶手段を結合または提供する際に、様々な方法を得ることができるように、記憶手段（例えば、ＲＡＭ、ＲＯＭ、コンパクトディスク（ＣＤ）またはフロッピーディスクのような物理記憶媒体、等）を介して提供されることができる。さらに、ここ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その他任意の適切な技法が、利用されることができる。

【０１９３】

[00115]特許請求の範囲は、上記に例示された厳密な構成およびコンポーネントに限定されないことが理解されるべきである。様々な修正、変更、および変形が、特許請求の範囲から逸脱することなく、上記で説明された方法および装置の配置、動作および詳細において行われ得る。

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］　トレーニングされた機械学習モデルのパフォーマンスを改善するための方法であって、

第１の目的関数を有する第１の分類器に、第２の目的関数を有する第２の分類器を追加すること、前記第２の目的関数は、前記第１の分類器の誤差を直接的に低減させるために使用される、

を備える方法。

［Ｃ２］　前記第１の目的関数は、微分可能である、Ｃ１に記載の方法。

［Ｃ３］　前記第２の目的関数は、微分不可能である、Ｃ１に記載の方法。

［Ｃ４］　前記第２の目的関数は、前記第１の分類器と前記第２の分類器との誤差間の差の関数である、Ｃ１に記載の方法。

［Ｃ５］　より高い複雑度のモデルからの確率の混合に少なくとも部分的に基づいて、前記第２の目的関数を決定することをさらに備える、Ｃ１に記載の方法。

［Ｃ６］　前記第１の分類器を再トレーニングすることなく、前記第２の分類器を追加することをさらに備える、Ｃ１に記載の方法。

［Ｃ７］　前記第１の分類器の外部に前記第２の分類器を追加することをさらに備える、

Ｃ１に記載の方法。
［Ｃ８］　アイデンティティ値に、前記第１の分類器によってトレーニングされたモデルによって生成される特徴に対する重みを割り当てることをさらに備える、Ｃ１に記載の方法。
［Ｃ９］　ゼロに、より高い複雑度のモデルの確率ベクトルによって生成される特徴に対する重みを割り当てることをさらに備える、Ｃ８に記載の方法。
［Ｃ１０］　前記第２の分類器の確率ベクトルによって生成される特徴に対する重みを割り当てることをさらに備える、Ｃ１に記載の方法。
［Ｃ１１］　ゼロに、より高い複雑度のモデルの確率ベクトルによって生成される特徴に対する重みを割り当てることをさらに備える、Ｃ１に記載の方法。
［Ｃ１２］　前記第２の分類器の確率ベクトルによって生成される特徴に対する重みを割り当てることをさらに備える、Ｃ１１に記載の方法。
［Ｃ１３］　固定された温度Ｔによって、より高い複雑度のモデルによって生成される確率ベクトルをスケーリングすることをさらに備える、Ｃ１に記載の方法。
［Ｃ１４］　トレーニングされた機械学習モデルのパフォーマンスを改善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
　を備え、前記少なくとも１つのプロセッサは、
　　第１の目的関数を有する第１の分類器に、第２の目的関数を有する第２の分類器を追加すること、前記第２の目的関数は、前記第１の分類器の誤差を直接的に低減させるために使用される、
　を行うように構成される、装置。
［Ｃ１５］　前記第１の目的関数は、微分可能である、Ｃ１４に記載の装置。
［Ｃ１６］　前記第２の目的関数は、微分不可能である、Ｃ１４に記載の装置。
［Ｃ１７］　前記第２の目的関数は、前記第１の分類器と前記第２の分類器との誤差間の差の関数である、Ｃ１４に記載の装置。
［Ｃ１８］　前記少なくとも１つのプロセッサは、より高い複雑度のモデルからの確率の混合に少なくとも部分的に基づいて、前記第２の目的関数を決定するようにさらに構成される、Ｃ１４に記載の装置。
［Ｃ１９］　前記少なくとも１つのプロセッサは、前記第１の分類器を再トレーニングすることなく、前記第２の分類器を追加するようにさらに構成される、Ｃ１４に記載の装置。
［Ｃ２０］　前記少なくとも１つのプロセッサは、前記第１の分類器の外部に前記第２の分類器を追加するようにさらに構成される、Ｃ１４に記載の装置。
［Ｃ２１］　前記少なくとも１つのプロセッサは、アイデンティティ値に、前記第１の分類器によってトレーニングされたモデルによって生成される特徴に対する重みを割り当てるようにさらに構成される、Ｃ１４に記載の装置。
［Ｃ２２］　前記少なくとも１つのプロセッサは、ゼロに、より高い複雑度のモデルの確率ベクトルによって生成される特徴に対する重みを割り当てるようにさらに構成される、Ｃ２１に記載の装置。
［Ｃ２３］　前記少なくとも１つのプロセッサは、前記第２の分類器の確率ベクトルによって生成される特徴に対する重みを割り当てるようにさらに構成される、Ｃ１４に記載の装置。
［Ｃ２４］　前記少なくとも１つのプロセッサは、ゼロに、より高い複雑度のモデルの確率ベクトルによって生成される特徴に対する重みを割り当てるようにさらに構成される、Ｃ１４に記載の装置。
［Ｃ２５］　前記少なくとも１つのプロセッサは、前記第２の分類器の確率ベクトルによって生成される特徴に対する重みを割り当てるようにさらに構成される、Ｃ２４に記載の装置。

［Ｃ２６］　前記少なくとも１つのプロセッサは、固定された温度Ｔによって、より高い複雑度のモデルによって生成される確率ベクトルをスケーリングするようにさらに構成される、Ｃ１４に記載の装置。

［Ｃ２７］　トレーニングされた機械学習モデルのパフォーマンスを改善するための装置であって、

第１の目的関数を有する第１の分類器に、第２の目的関数を有する第２の分類器を追加するための手段と、前記第２の目的関数は、前記第１の分類器の誤差を直接的に低減させるために使用される、

前記トレーニングされた機械学習モデルを介して受信される入力に少なくとも部分的に基づいて、前記第２の分類器から特徴ベクトルを出力するための手段と

を備える装置。

［Ｃ２８］　トレーニングされた機械学習モデルのパフォーマンスを改善するためのプログラムコードをその上に符号化された非一時的なコンピュータ可読媒体であって、前記プログラムコードは、プロセッサによって実行され、第１の目的関数を有する第１の分類器に、第２の目的関数を有する第２の分類器を追加するためのプログラムコードを備え、前記第２の目的関数は、前記第１の分類器の誤差を直接的に低減させるために使用される、非一時的なコンピュータ可読媒体。

【図１】

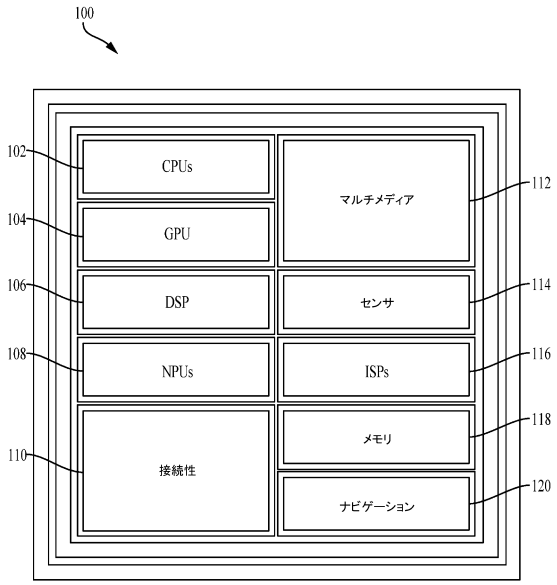

FIG. 1

【図２】

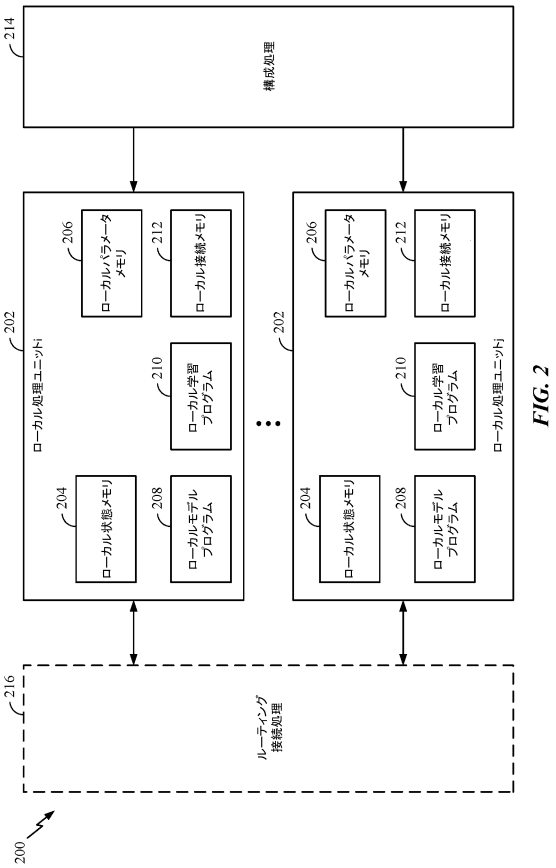

FIG. 2

【図３Ａ】

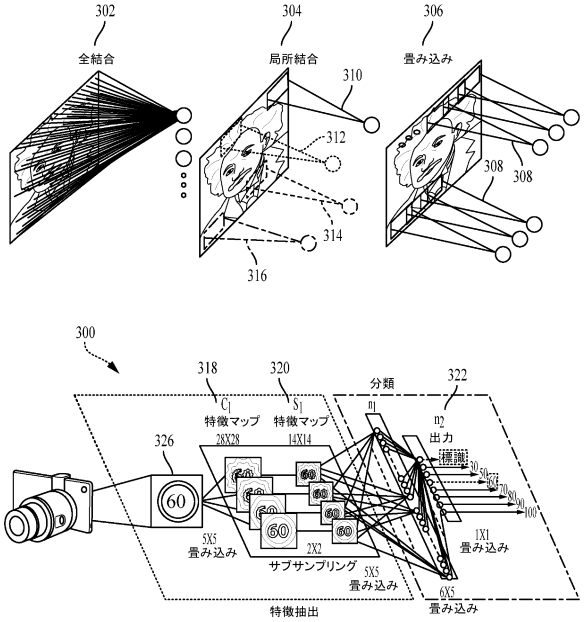

FIG. 3A

【図３Ｂ】

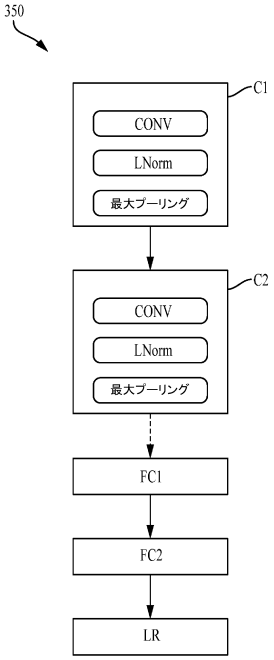

FIG. 3B

【図４】

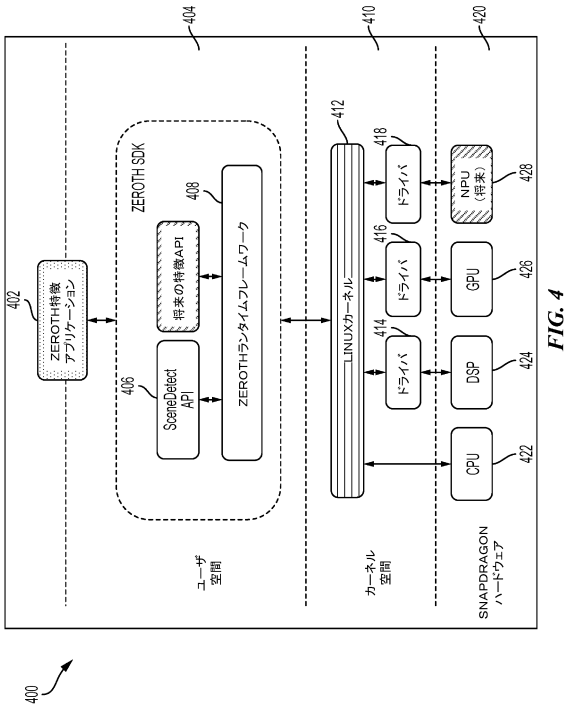

FIG. 4

【図５】

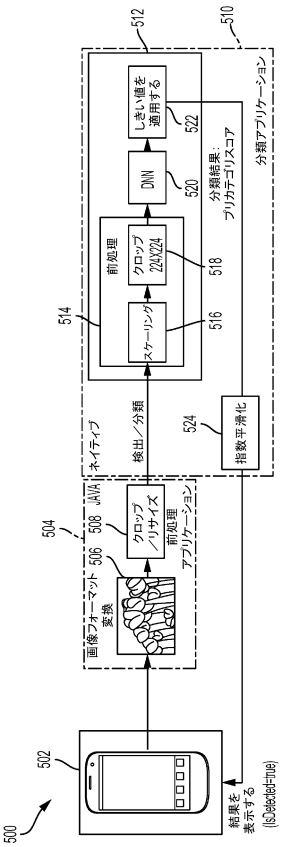

FIG. 5

【図６Ａ】

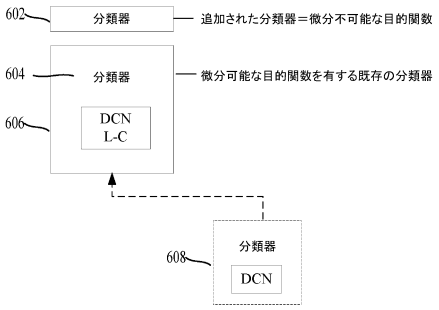

FIG. 6A

【図６Ｂ】

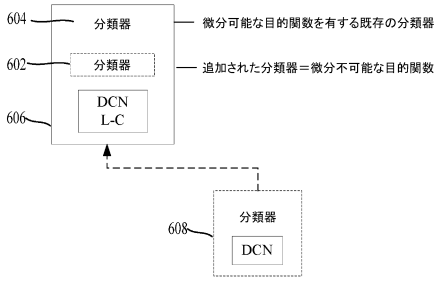

FIG. 6B

【図７】

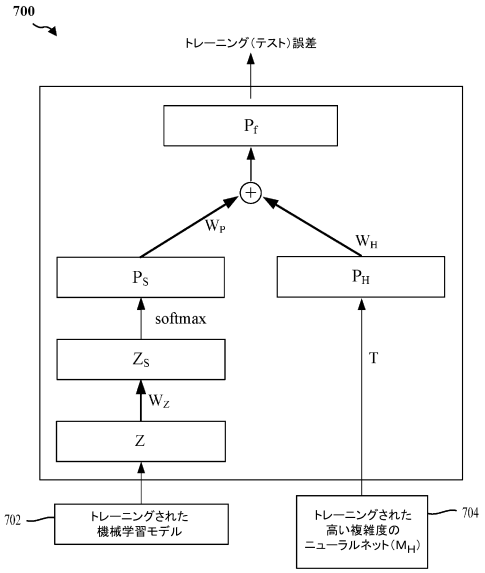

FIG. 7

【図８】

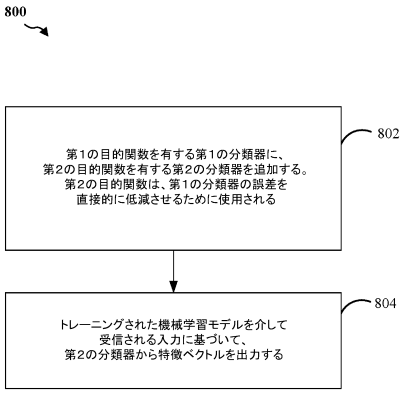

FIG. 8

【図９】

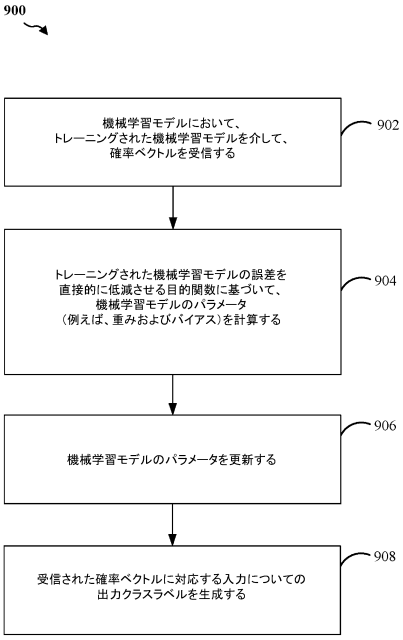

FIG. 9

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 タラティ、サチン・スバシュ
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 バルタク、アニケット
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

審査官 多賀 実

(56)参考文献 特開平０５－２９００１３（ＪＰ，Ａ）
特開平０９－１８５３９４（ＪＰ，Ａ）
特開２０１１－０４８２６２（ＪＰ，Ａ）
Geoffrey Hinton et al., "Distilling the Knowledge in a Neural Network", arXiv.org [online], Cornell University, ２０１５年 3月, arXiv:1503.02531, pp.1-9, [令和２年８月２８日検索],インターネット<URL:https://arxiv.org/pdf/1503.02531>

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