(19)日本国特許庁(JP)

(12)特許公報(B1)

(11)特許番号
特許第7112802号
(P7112802)

(45)発行日　令和4年8月4日(2022.8.4)

(24)登録日　令和4年7月27日(2022.7.27)

(51)国際特許分類
Ｇ０６Ｎ　20/00　(2019.01)
Ｇ０６Ｎ　3/04　(2006.01)

ＦＩ
Ｇ０６Ｎ　20/00　１３０
Ｇ０６Ｎ　3/04

請求項の数　9　（全20頁）

(21)出願番号　特願2022-73380(P2022-73380)
(22)出願日　令和4年4月27日(2022.4.27)
審査請求日　令和4年4月27日(2022.4.27)
早期審査対象出願

(73)特許権者　520040304
窪田 望
東京都港区港南１丁目９番３６号 ＮＴＴ ＤＡＴＡ 品川ビル１３階 株式会社クリエイターズネクスト内
(74)代理人　100079108
弁理士　稲葉 良幸
(74)代理人　100109346
弁理士　大貫 敏史
(74)代理人　100117189
弁理士　江口 昭彦
(74)代理人　100134120
弁理士　内藤 和彦
(72)発明者　窪田 望
東京都港区港南１丁目９番３６号 ＮＴＴ
最終頁に続く

(54)【発明の名称】　学習モデルの軽量化

(57)【特許請求の範囲】
【請求項１】
情報処理装置に含まれる１又は複数のプロセッサが、
所定の学習データを取得すること、
ニューラルネットワークを用いる所定の学習モデルに対して、蒸留処理された第１学習モデル、枝刈り処理された第２学習モデル、及び量子化処理された第３学習モデルの少なくとも２つのモデルを含む各モデルがそれぞれ重み付けされた重み学習モデルに、所定のデータを入力して機械学習を行うこと、
前記各モデルそれぞれの重みが変更された重み学習モデルごとに、前記所定の学習データを入力して前記機械学習が行われた場合の学習結果を取得すること、
変更された各重みが重み付けされた各重み学習モデルと、前記各重み学習モデルで学習されたときの各学習結果とを含む学習データを用いて、教師あり学習を行うこと、
前記教師あり学習により、任意の学習データを入力する場合に、各重みの集合ごとに学習結果を予測する予測モデルを生成すること、
を実行する、情報処理方法。
【請求項２】
前記１又は複数のプロセッサは、
任意の学習データを前記予測モデルに入力し、前記各重みの集合ごとに、前記重み学習モデルを実行した場合の学習結果を予測することを実行する、請求項１に記載の情報処理方法。

【請求項３】
前記１又は複数のプロセッサが、
前記任意の学習データを前記所定の学習モデルに入力した場合の学習結果と、前記予測モデルにより予測された学習結果とが、軽量化に関する所定条件を満たすか否かを判定すること、
前記所定条件の判定結果に基づいて、前記各重みの有効性を判定すること、
をさらに実行する請求項２に記載の情報処理方法。
【請求項４】
前記１又は複数のプロセッサが、
前記軽量化に関する所定条件に関するユーザ操作を受け付けること、
前記ユーザ操作に基づいて、前記軽量化に関する所定条件を設定すること、
をさらに実行する請求項３に記載の情報処理方法。
【請求項５】
前記プロセッサは、
前記学習結果に含まれる学習精度を第１変数、前記学習結果に含まれるモデルサイズに関する値を第２変数とし、前記第１変数及び前記第２変数と、前記各重みとを対応付ける関係情報を生成すること、
を実行する請求項１に記載の情報処理方法。
【請求項６】
前記プロセッサは、
前記第１変数の第１値及び前記第２変数の第２値を取得すること、
前記関係情報に基づいて、前記第１値及び前記第２値に対応する各重みを特定すること、
を実行する請求項５に記載の情報処理方法。
【請求項７】
前記重み学習モデルは、前記第１学習モデル、前記第２学習モデル、及び前記第３学習モデルそれぞれに重みが付与されて線形結合されたモデルを含む、請求項１に記載の情報処理方法。
【請求項８】
メモリと、１又は複数のプロセッサとを備える情報処理装置であって、
前記メモリは、
ニューラルネットワークを用いる所定の学習モデルと、
前記所定の学習モデルに対して、蒸留処理された第１学習モデル、枝刈り処理された第２学習モデル、及び量子化処理された第３学習モデルの少なくとも２つのモデルを含む各モデルがそれぞれ重み付けされた重み学習モデルと、を記憶し、
前記１又は複数のプロセッサは、
所定の学習データを取得すること、
前記重み学習モデルに、前記所定の学習データを入力して機械学習を行うこと、
前記各モデルそれぞれの重みが変更された重み学習モデルごとに、前記所定の学習データを入力して前記機械学習が行われた場合の学習結果を取得すること、
変更された各重みが重み付けされた各重み学習モデルと、前記各重み学習モデルで学習されたときの各学習結果とを含む学習データを用いて、教師あり学習を行うこと、
前記教師あり学習により、任意の学習データを入力する場合に、各重みの集合ごとに学習結果を予測する予測モデルを生成すること、
を実行する、情報処理装置。
【請求項９】
情報処理装置に含まれる１又は複数のプロセッサに、
所定の学習データを取得すること、
ニューラルネットワークを用いる所定の学習モデルに対して、蒸留処理された第１学習モデル、枝刈り処理された第２学習モデル、及び量子化処理された第３学習モデルの少なくとも２つのモデルを含む各モデルがそれぞれ重み付けされた重み学習モデルに、所定の

学習データを入力して機械学習を行うこと、
前記各モデルそれぞれの重みが変更された重み学習モデルごとに、前記所定の学習データを入力して前記機械学習が行われた場合の学習結果を取得すること、
変更された各重みが重み付けされた各重み学習モデルと、前記各重み学習モデルで学習されたときの各学習結果とを含む学習データを用いて、教師あり学習を行うこと、
前記教師あり学習により、任意の学習データを入力する場合に、各重みの集合ごとに学習結果を予測する予測モデルを生成すること、
を実行させる、プログラムを記録したコンピュータ読み取り可能な非一時的な記憶媒体。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、学習モデルの軽量化に関する情報処理方法、プログラム及び情報処理装置に関する。
【背景技術】
【０００２】
近年、学習モデルを軽量化する研究が行われている。例えば、下記特許文献１には、パラメータの量子化を用いて学習モデルの軽量化を行う技術が記載されている。
【先行技術文献】
【特許文献】
【０００３】
【文献】特開２０１９－１３３６２８号公報
【発明の概要】
【発明が解決しようとする課題】
【０００４】
ここで、学習モデルを軽量化するには、枝刈り（Pruning）、量子化（Quantization）、蒸留（Distillation）などの手法がある。これらの軽量化手法の少なくとも１つを学習モデルに適用し、エンジニアが適宜パラメータ調整等を行って軽量化が行われる。
【０００５】
しかしながら、学習データや学習モデルなどの様々な条件に応じて適切な軽量化手法は異なるものと思われるが、エンジニアが適宜パラメータを調整して決定された軽量化手法が最適なものであるとは限らなかった。
【０００６】
そこで、本発明の目的の１つは、学習モデルに対する軽量化手法を、より適切にすることを可能にする情報処理方法、プログラム及び情報処理装置を提供する。
【課題を解決するための手段】
【０００７】
本発明の一態様に係る情報処理方法は、情報処理装置に含まれる１又は複数のプロセッサが、所定の学習データを取得すること、ニューラルネットワークを用いる所定の学習モデルに対して、蒸留処理された第１学習モデル、枝刈り処理された第２学習モデル、及び量子化処理された第３学習モデルの少なくとも２つのモデルを含む各モデルがそれぞれ重み付けされた重み学習モデルに、所定のデータを入力して機械学習を行うこと、前記各モデルそれぞれの重みが変更された重み学習モデルごとに、前記所定の学習データを入力して前記機械学習が行われた場合の学習結果を取得すること、変更された各重みが重み付けされた各重み学習モデルと、前記各重み学習モデルで学習されたときの各学習結果とを含む学習データを用いて、教師あり学習を行うこと、前記教師あり学習により、任意の学習データを入力する場合に、各重みの集合ごとに学習結果を予測する予測モデルを生成すること、を実行する。
【発明の効果】
【０００８】
本発明によれば、学習モデルに対する軽量化手法を、より適切にすることを可能にする

情報処理方法、プログラム及び情報処理装置を提供することができる。
【図面の簡単な説明】
【０００９】
【図１】実施形態に係るシステム構成の一例を示す図である。
【図２】実施形態に係る情報処理装置の物理的構成の一例を示す図である。
【図３】実施形態に係る情報処理装置の処理ブロックの一例を示す図である。
【図４】学習済みモデルの蒸留を説明するための図である。
【図５】学習済みモデルの枝刈りを説明するための図である。
【図６】実施形態に係る情報処理装置の処理ブロックの一例を示す図である。
【図７】実施形態に係る関係情報の一例を示す図である。
【図８】実施形態に係る関係情報の表示例を示す図である。
【図９】実施形態に係る予測モデルの生成に関する処理の一例を示すフローチャートである。
【図１０】実施形態に係るユーザが利用する情報処理装置における処理の一例を示すフローチャートである。
【発明を実施するための形態】
【００１０】
添付図面を参照して、本発明の実施形態について説明する。なお、各図において、同一の符号を付したものは、同一又は同様の構成を有する。
【００１１】
［実施形態］
＜システム構成＞
図１は、実施形態に係るシステム構成の一例を示す図である。図１に示す例では、サーバ１０と、各情報処理装置２０Ａ、２０Ｂ、２０Ｃ、２０Ｄとが、ネットワークを介してデータ送受信可能なように接続される。情報処理装置を個別に区別しない場合は情報処理装置２０とも表記する。
【００１２】
サーバ１０は、データを収集、分析可能な情報処理装置であり、１つ又は複数の情報処理装置から構成されてもよい。情報処理装置２０は、スマートフォン、パーソナルコンピュータ、タブレット端末、サーバ、コネクティッドカーなど、機械学習を実行可能な情報処理装置である。なお、情報処理装置２０は、脳波をセンシングする侵襲型又は非侵襲型の電極に直接的又は間接的に接続されており、脳波データを解析、送受信可能な装置でもよい。
【００１３】
図１に示すシステムでは、サーバ１０は、例えば、所定の学習データにおいて学習済みの学習モデルに、様々な軽量化手法（軽量化アルゴリズム）を適用する。様々な軽量化手法は、既存の軽量化手法１つを適用したり、任意の軽量化手法の組み合わせを適用したりすることを含む。このとき、サーバ１０は、所定のデータセット、所定の学習モデル及び所定の軽量化手法のときの学習結果を関連付けて記憶する。
【００１４】
次に、サーバ１０は、任意のデータセットと、任意の軽量化手法と、これらの学習結果（例えば学習精度）とを訓練データとして、学習結果が適切な軽量化手法を特定する予測モデルを学習して生成する。学習結果の適切さは、例えば、学習精度や、モデルサイズの圧縮率などにより決定される。
【００１５】
これにより、学習済みの学習モデルに対し、軽量化をより適切に行うことが可能になる。また、サーバ１０は、各軽量化手法を重み付けして線形結合したモデルを用いて、各軽量化手法の適用割合を定める各重みを適切に調整してもよい。
【００１６】
＜ハードウェア構成＞

図２は、実施形態に係る情報処理装置１０の物理的構成の一例を示す図である。情報処理装置１０は、演算部に相当する１又は複数のＣＰＵ（Central Processing Unit）１０ａと、記憶部に相当するＲＡＭ（Random Access Memory）１０ｂと、記憶部に相当するＲＯＭ（Read only Memory）１０ｃと、通信部１０ｄと、入力部１０ｅと、表示部１０ｆと、を有する。これらの各構成は、バスを介して相互にデータ送受信可能に接続される。

【００１７】

実施形態では、情報処理装置１０が一台のコンピュータで構成される場合について説明するが、情報処理装置１０は、複数のコンピュータ又は複数の演算部が組み合わされて実現されてもよい。また、図２で示す構成は一例であり、情報処理装置１０はこれら以外の構成を有してもよいし、これらの構成のうち一部を有さなくてもよい。

【００１８】

ＣＰＵ１０ａは、ＲＡＭ１０ｂ又はＲＯＭ１０ｃに記憶されたプログラムの実行に関する制御やデータの演算、加工を行う制御部である。ＣＰＵ１０ａは、より適切な軽量化手法を調べるための学習モデルを用いて学習を行うプログラム（学習プログラム）や、任意のデータを入力したときに適切な軽量化手法を出力する予測モデルを生成するための学習を行うプログラム（予測プログラム）を実行する演算部である。ＣＰＵ１０ａは、入力部１０ｅや通信部１０ｄから種々のデータを受け取り、データの演算結果を表示部１０ｆに表示したり、ＲＡＭ１０ｂに格納したりする。

【００１９】

ＲＡＭ１０ｂは、記憶部のうちデータの書き換えが可能なものであり、例えば半導体記憶素子で構成されてよい。ＲＡＭ１０ｂは、ＣＰＵ１０ａが実行するプログラム、様々な軽量化手法に関する軽量化データ（例えば軽量化アルゴリズム）、適切な軽量化手法を予測する予測モデル、学習対象のデータに関する情報と、このデータに対応する適切な軽量化手法との対応関係を示す関係情報などのデータを記憶してもよい。なお、これらは例示であって、ＲＡＭ１０ｂには、これら以外のデータが記憶されていてもよいし、これらの一部が記憶されていなくてもよい。

【００２０】

ＲＯＭ１０ｃは、記憶部のうちデータの読み出しが可能なものであり、例えば半導体記憶素子で構成されてよい。ＲＯＭ１０ｃは、例えば学習プログラムや、書き換えが行われないデータを記憶してよい。

【００２１】

通信部１０ｄは、情報処理装置１０を他の機器に接続するインターフェースである。通信部１０ｄは、インターネット等の通信ネットワークに接続されてよい。

【００２２】

入力部１０ｅは、ユーザからデータの入力を受け付けるものであり、例えば、キーボード及びタッチパネルを含んでよい。

【００２３】

表示部１０ｆは、ＣＰＵ１０ａによる演算結果を視覚的に表示するものであり、例えば、ＬＣＤ（Liquid Crystal Display）により構成されてよい。表示部１０ｆが演算結果を表示することは、ＸＡＩ（eXplainable AI：説明可能なＡＩ）に貢献し得る。表示部１０ｆは、例えば、学習結果や、学習モデルに関する情報を表示してもよい。

【００２４】

学習プログラムは、ＲＡＭ１０ｂやＲＯＭ１０ｃ等のコンピュータによって読み取り可能な記憶媒体に記憶されて提供されてもよいし、通信部１０ｄにより接続される通信ネットワークを介して提供されてもよい。情報処理装置１０では、ＣＰＵ１０ａが学習プログラムを実行することにより、後述する図３を用いて説明する様々な動作が実現される。なお、これらの物理的な構成は例示であって、必ずしも独立した構成でなくてもよい。例えば、情報処理装置１０は、ＣＰＵ１０ａとＲＡＭ１０ｂやＲＯＭ１０ｃが一体化したＬＳＩ（Large-Scale Integration）を備えていてもよい。また、情報処理装置１０は、ＧＰ

Ｕ（Graphical Processing Unit）やＡＳＩＣ（Application Specific Integrated Circuit）を備えていてもよい。

【００２５】

なお、情報処理装置２０の構成は、図２に示す情報処理装置１０の構成と同様であるため、その説明を省略する。また、情報処理装置１０と情報処理装置２０とは、データ処理を行う基本的な構成であるＣＰＵ１０ａやＲＡＭ１０ｂ等を有していればよく、入力部１０ｅや表示部１０ｆは設けられなくてもよい。また、入力部１０ｅや表示部１０ｆは、外部からインターフェースを用いて接続されてもよい。

【００２６】

＜処理構成＞

図３は、実施形態に係る情報処理装置１０の処理ブロックの一例を示す図である。情報処理装置１０は、取得部１０１、第１学習部１０２、変更部１０３、第２学習部１０４、予測部１０５、判定部１０６、設定部１０７、関連付け部１０８、特定部１０９、表示制御部１１０、出力部１１１、及び記憶部１１２を備える。例えば、図３に示す第１学習部１０２、変更部１０３、第２学習部１０４、予測部１０５、判定部１０６、設定部１０７、関連付け部１０８、特定部１０９、表示制御部１１０は、例えばＣＰＵ１０ａなどにより実行されて実現され、取得部１０１及び出力部１１１は、例えば通信部１０ｄなどにより実現され、記憶部１１２は、ＲＡＭ１０ｂ及び／又はＲＯＭ１０ｃなどにより実現され得る。

【００２７】

取得部１０１は、所定の学習データを取得する。例えば、取得部１０１は、所定の学習データとして、画像データ、系列データ、テキストデータなどの公知のデータセットを取得してもよい。なお、取得部１０１は、記憶部１１２に記憶されたデータを取得してもよいし、他の情報処理装置により送信されたデータを取得してもよい。

【００２８】

第１学習部１０２は、所定の問題を解くため、ニューラルネットワークを用いる所定の学習モデル１０２ａに対して、蒸留処理された第１学習モデル、枝刈り処理された第２学習モデル、及び量子化処理された第３学習モデルの少なくとも２つのモデルを含む各モデルがそれぞれ重み付けされた重み学習モデルに、所定の学習データを入力して機械学習を行う。

【００２９】

ここで、学習済みの学習モデル１０２ａの軽量化手法の例として、蒸留（Distillation）、枝刈り（Pruning）、及び量子化（Quantization）の各アルゴリズムについて、以下に簡単に説明する。

【００３０】

図４は、学習済みモデルの蒸留を説明するための図である。図４に示す蒸留は、学習済みモデルＭ１１の予測結果を教師データとして、より小さいモデルＭ１２を学習することで軽量化を行う。このとき、この小さいモデルＭ１２は、大きいモデルＭ１１と同程度の精度を持つ場合がある。

【００３１】

例えば、蒸留において、学習済みモデルＭ１１は、Ｔｅａｃｈｅｒモデル、小さいモデルＭ１２は、Ｓｔｕｄｅｎｔモデルと呼ばれる。Ｓｔｕｄｅｎｔモデルは、エンジニアが適宜設計する。

【００３２】

図４に示す例では、分類器を例とした学習データについて説明する。モデルＭ１１のＴｅａｃｈｅｒモデルは、０と１とで表現され、１が正解である教師データを用いて学習を行なう。これに対し、モデルＭ１２のＳｔｕｄｅｎｔモデルは、Ｔｅａｃｈｅｒモデルが出力した値（例：Ａ＝０．７、Ｂ＝０．３）を教師データとして学習する。実施形態では、１つの学習モデルＭ１１に対して複数の異なる蒸留後のモデルＭ１２が用意されてもよい。

【００３３】
図５は、学習済みモデルの枝刈りを説明するための図である。図５に示す枝刈りは、学習済みモデルＭ２１の重みやノードを削除することで、軽量化が行われたモデルＭ２２が生成される。これにより計算回数、メモリ使用量の削減を行うことが可能になる。
【００３４】
枝刈りの手法は、ノード間の接続において重みの小さいところを対象に削除が行われてもよい。例えば、枝刈りは、蒸留と違い別途モデルを設計する必要はないが、パラメータの削除が行われるため、再学習を行い、学習精度を維持するとよい。例えば、学習への影響が小さい枝（エッジ）、例えば重みが所定値以下の枝をカットし、軽量化が行われてもよい。
【００３５】
量子化は、モデルに含まれるパラメータを少ないビット数で表現する。これにより、ネットワークの構造を変えずにモデルを小さくすることが可能になる。例えば、重みパラメータを６個持つ簡単なネットワークを例にした場合、３２ビット精度の場合は合計１９２ビットを必要とするが、８ビット精度の制約にすると合計４８ビットで表現することとになり、軽量化が行われていることになる。
【００３６】
図３に戻り、例えば、第１学習部１０２は、学習済みの学習モデル１０２ａに対して、第１モデル、第２モデル、及び第３モデルのうちの少なくとも２つの軽量化モデルが選択され、各モデルに付与される重みとして、デフォルトの重みを設定する。
【００３７】
第１モデル、第２モデル、及び第３モデルは、学習済みのモデルのカテゴリごとに予め設定されていてもよいし、学習済みモデルごとに、所定の基準に従って自動で生成されてもよい。例えば、第１学習部１０２は、蒸留の場合、学習済みモデルに適した蒸留後のモデルを機械学習により決定してもよく、枝刈りの場合、重みが所定値以下の枝をカットして枝刈り後のモデルを生成してもよく、量子化の場合、所定値ビット精度の制約（量子化）にしてもよい。また、１つの学習済みモデルに対し、複数の第１モデル、複数の第２モデル、複数の第３モデルが設定され、それぞれのモデルに重みが付与されてもよい。
【００３８】
所定の問題は、例えば画像データ、系列データ及びテキストデータの少なくともいずれかについて、分類、生成及び最適化の少なくともいずれかを行う問題を含む。ここで、画像データは、静止画のデータと、動画のデータとを含む。系列データは、音声データや株価のデータを含む。
【００３９】
また、所定の学習モデル１０２ａは、ニューラルネットワークを含む学習済みの学習モデルであり、例えば、画像認識モデル、系列データ解析モデル、ロボットの制御モデル、強化学習モデル、音声認識モデル、音声生成モデル、画像生成モデル、自然言語処理モデル等の少なくとも１つを含む。また、具体例としては、所定の学習モデル１０２ａは、ＣＮＮ（Convolutional Neural Network）、ＲＮＮ（Recurrent Neural Network）、ＤＮＮ（Deep Neural Network）、ＬＳＴＭ（Long Short-Term Memory）、双方向ＬＳＴＭ、ＤＱＮ（Deep Q-Network）、ＶＡＥ（Variational AutoEncoder）、ＧＡＮｓ（Generative Adversarial Networks）、ｆｌｏｗ－ｂａｓｅｄ生成モデル等のいずれかでもよい。
【００４０】
変更部１０３は、所定の学習データ及び／又は重み学習モデルの各重みを変更する。例えば、変更部１０３は、複数の学習データの中から、第１学習部１２に入力される所定の学習データを１つずつ順に変更する。また、変更部１０３は、ある１つの重み学習モデルに対して全ての所定の学習データが入力されて学習が行われた場合、重み学習モデルの別の各重みを利用するため、複数の各重みの集合（セット）の中から１つのセットを選択し、用意された全てのセットで学習を行い、学習結果を取得してもよい。

【００４１】
　また、第１学習部１０２は、所定の学習データを重み学習モデルに入力し、適切な学習結果が出力されるように、重み学習モデルのハイパーパラメータ等の学習を行う。このとき、第１学習部１０２は、ハイパーパラメータが更新（調整）される際に、重み学習モデルの各モデルに付与される各重みも所定の方法により調整する。
【００４２】
　例えば、各重みの調整については、あらかじめ設定される初期値から逐次的に各重みが調整されるとよい。このとき、各重みが全て加算して１になるように調整され、以前に行った調整と異なる調整が行われればいずれの調整方法が用いられてもよい。例えば、第１学習部１０２は、各重みを順に所定値ずつ変更していき、全ての組み合わせについて変更する。例えば、第１学習部１０２は、重みw_kに対して初期値から所定値ずつ減算し、重みw_{k+1}に対して初期値から所定値ずつ加算し、どちらかの重みが０以下になると、ｋに１を加算して、各初期値からの変更を繰り返す。また、各重みが全て加算して１になる条件は設けなくてもよく、この場合、Ｓｏｆｔｍａｘ関数などを用いて、各重みを加算して１になるように最後に調整されればよい。
【００４３】
　これにより、所定の学習データと所定の各重みのセットとの任意の組み合わせに対して学習させることが可能になる。例えば、変更部１０３は、所定の学習データと所定の各重みのセットとの全ての組み合わせが学習されるように、所定の学習データ及び／又は所定の各重みのセットを１つずつ順に変更してもよいし、所定の条件が満たされるまで所定の学習データ及び／又は所定の各重みのセットを１つずつ順に変更してもよい。所定の条件は、例えば、学習精度やモデルサイズの圧縮率などにより設定されてもよい。
【００４４】
　取得部１０１又は第１学習部１０２は、各モデルそれぞれの重みが変更された重み学習モデルごとに、所定の学習データを入力して機械学習が行われた場合の学習結果を取得する。例えば、取得部１０１又は第１学習部１０２は、様々な組み合わせの所定の学習データ及び／又は所定の各重みのセットを用いて学習された学習結果を取得する。
【００４５】
　ここで、重み学習モデルについて具体例を用いて説明する。例えば、第１学習部１０２は、第１モデル、第２モデル、第３モデルにそれぞれ重みw_1，w_2，w_3を付与して線形結合した重み学習モデルを利用してもよい。この場合の重み学習の関数Ｍ（ｘ）として、式（１）が挙げられるが、一例にすぎない。

$$M_1(x)=w_1m_1(x)+w_2m_2(x)+w_3m_3(x) \quad \cdots\text{式（１）}$$

w_n：重み（各重みの集合（セット）をＷとも表記する）
m_n（ｘ）：第ｎモデル
ｘ：学習データ
【００４６】
　変更部１０３は、例えば、$w_1+w_2+w_3=1$となるように、各重みを所定基準に従って、１つずつ順に変更する。第１学習部１０２は、変更後の各重みに対する学習結果を取得し、各重みのセットに対して学習結果を関連付けておく。学習結果は、学習精度と、軽量化による効果を示すモデルサイズの圧縮率である。モデルサイズの圧縮率とは、例えば、軽量化後の学習済みモデルのパラメータ数の、軽量化前の学習済みモデルのパラメータ数に対する割合である。
【００４７】
　また、変更部１０３が所定の学習データを変更すると、第１学習部１０２は、変更後の学習データに対して、上述されたように各重みのセットでの重み学習モデルを学習し、学習結果を取得する。これにより、任意の学習データ、任意の各重みのセット、これらの場合の学習結果を含む訓練データが生成される。
【００４８】
　第２学習部１０４は、変更された各重みが重み付けされた各重み学習モデルと、各重み

学習モデルで学習されたときの各学習結果とを含む学習データを用いて、教師あり学習を行う。例えば、第２学習部１０４は、任意の学習データ及び任意の各重みのセットを用いて学習された際の学習結果（例えば学習性能及び／又はモデルサイズの圧縮率）を正解ラベルとする訓練データを用いて、教師あり学習を行う。

【００４９】

また、第２学習部１０４は、教師あり学習により、任意の学習データを入力する場合に、各重みのセットごとに学習結果を予測する予測モデル１０４ａを生成する。例えば、第２学習部１０４は、任意の学習データを入力すると、この学習データに対する各軽量化手法の各重みのセットごとに、学習精度やモデルサイズの圧縮率を出力する予測モデルを生成する。

【００５０】

以上の構成により、様々な学習データや、様々な軽量化手法により軽量化した各学習モデルを用いた学習結果を訓練データとして教師あり学習を行うことにより、各重みのセットごとに、学習結果を予測する予測モデルを生成することができる。その結果、第２学習部１０４により生成された予測モデルを用いることで、軽量化手法をより適切にすることが可能になる。

【００５１】

予測部１０５は、任意の学習データを予測モデル１０４ａに入力し、各モデルそれぞれの重みのセットごとに、重み学習モデルを実行した場合の学習結果を予測する。例えば、予測部１０５は、学習データとして画像のデータセットを入力した場合、特定の重みセットＷｎ（ｗ１ｎ，ｗ２ｎ，ｗ３ｎ）ごとに、学習精度とモデルサイズに関する値（例えば圧縮率）とを予測する。

【００５２】

これにより、任意のデータ（例、データセット）に対して、各軽量化手法をどれくらい適用するかの各重みのセットごとに、学習結果が予測されるため、この学習結果に基づいて、より適切な各重みを選択することなどが可能になる。

【００５３】

判定部１０６は、任意の学習データを所定の学習モデル１０２ａに入力した場合の学習結果と、予測モデル１０４ａにより予測された学習結果とが、軽量化に関する所定条件を満たすか否かを判定する。例えば、判定部１０６は、軽量化前の学習済みの学習モデル１０２ａに学習データＡを入力したときの学習精度Ａ１と、予測モデル１０４ａにより予測された学習精度Ｂ１との第１差分値が、第１閾値内であるか否かを判定する。この第１差分値が小さければと小さいほど、学習モデルを軽量化後でも学習精度を維持することができており、学習精度Ｂ１のときの各重みは適切な軽量化手法となる。

【００５４】

また、判定部１０６は、軽量化前の学習済みの学習モデル１０２ａの圧縮率Ａ２（＝１）と、予測モデル１０４ａにより予測された圧縮率Ｂ２との第２差分値が、第２閾値以上であるか否かを判定する。この第２差分値が大きければ大きいほど、学習モデルがより軽量化できていることを示す。

【００５５】

判定部１０６は、軽量化に関する判定結果に基づいて、各重みの有効性を判定する。例えば、判定部１０６は、第１差分値と第２差分値とに基づいて、圧縮率Ｂ２が大きく、学習精度Ｂ１が軽量化前の精度を維持できている各重みに対し、有効な軽量化手法であると判定する。具体例としては、判定部１０６は、第１差分値が第１閾値以下、第２差分値が第２閾値以上の各重みを有効な軽量化手法、それ以外の各重みを有効ではない軽量化手法と判定してもよい。

【００５６】

これにより、モデルサイズに関する値（例えば圧縮率）と学習精度とに基づき、それぞれの予測値を参考に、適切な各重みを選定することができる。例えば、判定部１０６は、最も学習精度が高い各重みを選定してもよいし、圧縮率が第２閾値以上のもので、学習精

度が最も高い各重みを選定してもよい。

【００５７】

設定部１０７は、軽量化に関する所定条件に関するユーザ操作を受け付ける。例えば、設定部１０７は、表示部１０ｆに表示された条件入力画面から、ユーザが入力部１０ｅを操作して軽量化に関する所定条件を入力した場合、この入力操作を受け付ける。

【００５８】

設定部１０７は、受け付けたユーザ操作に基づいて、軽量化に関する所定条件を判定部１０６の判定条件に設定する。例えば、設定部１０７は、ユーザの入力操作に基づいて、学習性能に関する第１閾値、及び／又は、モデルサイズに関する第２閾値を設定することを可能としてもよい。

【００５９】

これにより、ユーザが所望する条件を用いて、有効な軽量化手法を特定することができるようになる。

【００６０】

関連付け部１０８は、学習結果に含まれる学習精度を第１変数、学習結果に含まれるモデルサイズに関する値（例えば圧縮率）を第２変数とし、第１変数及び第２変数と、各重みとを対応付ける関係情報を生成する。例えば、関連付け部１０８は、縦軸を第１変数、横軸を第２変数とする場合に、それぞれの変数の交点に各重みＷを対応付けたマトリックスを生成してもよい。また、関連付け部１０８は、各情報処理装置２０から取得された学習精度や圧縮率に基づいて、第１変数及び第２変数と、各重みＷとを対応付ける関係情報（実測関係情報）を生成してもよい。

【００６１】

以上の処理により、第１変数又は第２変数が変更された場合に、対応する各重みＷを迅速に特定することが可能になる。また、第１変数と第２変数とは、適宜変更されてもよい。例えば、第１変数として学習精度、第２変数として各重みＷを適用し、特定される情報モデルサイズに関する値でもよい。

【００６２】

また、取得部１０１は、第１変数の第１値及び第２変数の第２値を取得してもよい。例えば、取得部１０１は、ユーザから指定される第１変数の第１値及び第２変数の第２値を取得する。第１値又は第２値はユーザにより適宜指定される。

【００６３】

この場合、特定部１０９は、関連付け部１０８により生成された関係情報に基づいて、第１変数の第１値及び第２変数の第２値に対応する各重みＷを特定する。例えば、特定部１０９は、関係情報を用いて、変更される第１変数の値、又は第２変数の値に対応する各重みＷを特定する。

【００６４】

表示制御部１１０は、特定部１０９により特定された各重みＷを表示装置（表示部１０ｆ）に表示制御する。また、表示制御部１１０は、第１変数及び第２変数を変更可能にしたマトリックスをＧＵＩ（Graphical User Interface）で表してもよい（例えば、後述する図８等）。

【００６５】

以上の処理により、ユーザにより指定された第１変数又は第２変数に応じて特定される各重みＷを、ユーザに対して可視化することが可能になる。ユーザは、第１変数又は第２変数を変更することで、所望の各重みＷを特定し、学習済みモデルの軽量化に適用することができる。

【００６６】

出力部１１１は、第２学習部１０４により予測された各重みＷを、他の情報処理装置２０に出力してもよい。例えば、出力部１１１は、所定の学習データを送信した情報処理装置２０であって、適切な各重みＷの取得を要求した情報処理装置２０に対し、所定の学習データに対応する適切な各重みＷを出力してもよい。また、出力部１１１は、予測された

各重みＷを記憶部１１２に出力してもよい。

【００６７】

記憶部１１２は、学習に関するデータを記憶する。記憶部１１２は、所定のデータセット１１２ａや、軽量化手法１１２ｂに関するデータ、上述した関係情報１１２ｃ、訓練データ、学習途中のデータ、学習結果に関する情報などを記憶する。

【００６８】

図６は、実施形態に係る情報処理装置２０の処理ブロックの一例を示す図である。情報処理装置２０は、取得部２０１、学習部２０２、出力部２０３、及び記憶部２０４を備える。情報処理装置２０は、汎用のコンピュータで構成されてもよい。

【００６９】

取得部２０１は、他の情報処理装置（例えばサーバ１０）により、分散学習の指示とともに、所定の重み学習モデルに関する情報や所定のデータセットに関する情報を取得してもよい。所定の重み学習モデルに関する情報は、各重みを示す情報や、重み学習モデル自体を示す情報でもよい。所定のデータセットに関する情報は、データセット自体でもよく、所定のデータセットが格納された格納先を示す情報でもよい。

【００７０】

学習部２０２は、所定の重み学習モデル２０２ａに学習対象の所定のデータセットを入力して学習を行う。学習部２０２は、学習後の学習結果をサーバ１０にフィードバックするように制御する。学習結果は、例えば、学習性能などを含み、モデルサイズに関する情報をさらに含んでもよい。学習部２０２は、学習対象のデータセットの種類、及び／又は、解くべき問題に応じて、学習モデル２０２ａを選択してもよい。

【００７１】

また、所定の重み学習モデル２０２ａは、ニューラルネットワークを含む学習モデルであり、例えば、画像認識モデル、系列データ解析モデル、ロボットの制御モデル、強化学習モデル、音声認識モデル、音声生成モデル、画像生成モデル、自然言語処理モデル等の少なくとも１つをベースに、各軽量化手法が重み付けされたモデルを含む。また、具体例としては、所定の重み学習モデル２０２ａのベースは、ＣＮＮ（Convolutional Neural Network）、ＲＮＮ（Recurrent Neural Network）、ＤＮＮ（Deep Neural Network）、ＬＳＴＭ（Long Short-Term Memory）、双方向ＬＳＴＭ、ＤＱＮ（Deep Q-Network）、ＶＡＥ（Variational AutoEncoder）、ＧＡＮｓ（Generative Adversarial Networks）、ｆｌｏｗ－ｂａｓｅｄ生成モデル等のいずれかでもよい。

【００７２】

出力部２０３は、分散学習の学習結果に関する情報を他の情報処理装置に出力する。例えば、出力部２０３は、学習部２０２による学習結果に関する情報をサーバ１０に出力する。例えば、分散学習の学習結果に関する情報は、上述したように、学習性能を含み、モデルサイズに関する情報をさらに含んでもよい。

【００７３】

記憶部２０４は、学習部２０２に関するデータを記憶する。記憶部２０４は、所定のデータセット２０４ａや、サーバ１０から取得したデータ、学習途中のデータ、学習結果に関する情報などを記憶する。

【００７４】

これにより、情報処理装置２０は、他の情報処理装置（例えばサーバ１０）からの指示により、所定のデータセットに対して、所定の重み学習モデルを適用した分散学習を実行し、学習結果をサーバ１０にフィードバックすることが可能になる。

【００７５】

また、出力部２０３は、所定のデータに関する情報を他の情報処理装置（例えばサーバ１０）に出力する。出力部２０３は、所定のデータ（例えば学習対象のデータセット）を出力してもよいし、所定のデータの特徴情報を出力してもよい。

【００７６】

取得部２０１は、他の情報処理装置から、所定のデータに対応する各重みＷを取得して

もよい。取得される各重みWは、他の情報処理装置が予測モデルを利用して予測した、所定のデータに適切な各重みである。

【００７７】

学習部２０２は、取得された各重みを重み学習モデル２０２ａに適用する。このとき、重み学習モデル２０２ａは、上述した学習に用いられた重み学習モデル２２ａに各重みを適用してもよい。また、重み学習モデル２０２ａは、他の情報処理装置１０から取得される学習モデルでもよいし、自装置で管理する学習モデルでもよい。

【００７８】

学習部２０２は、各重みが適用された重み学習モデル２０２ａに、所定のデータを入力して学習結果を取得する。この学習結果は、所定のデータに適した各重みを用いて学習した結果である。学習部２０２は、学習性能を保ちつつ適切に軽量化された学習モデルを使用することができる。

＜データ例＞

図７は、実施形態に係る関係情報の一例を示す図である。図７に示す例では、関係情報は、各第１変数（例、Ｐ１１）及び各第２変数（例、Ｐ２１）に対応する各重み（例、Ｗ１）を含む。第１変数Ｐ１ｎは、例えば学習精度であり、第２変数Ｐ２ｎは例えばモデルサイズの圧縮率であり、変数としてはいずれかの変数だけでもよい。各重みＷ（Ｐ１ｎ，Ｐ２ｍ）は、第第１変数Ｐ１ｎ及び第２変数Ｐ２ｎの場合の重みである。

【００７９】

図７に示す関係情報について、サーバ１０は、所定の分散インスタンス数とハイパーパラメータの組み合わせで分散学習を行わせた情報処理装置２０から、又は、自装置の教師あり学習の結果から、学習精度（第１変数）と、圧縮率（第２変数）とを取得する。サーバ１０は、取得された学習精度と圧縮率に、各重みWを対応付ける。サーバ１０は、教師あり学習により実測された学習精度と圧縮率とを取得するたびに行うことで、図７に示す関係情報を生成することが可能になる。また、関係情報は、予測部１０５により予測された結果に基づいて、任意のデータセットに対する予測関係情報が生成されてもよい。

【００８０】

＜ユーザインタフェースの例＞

図８は、実施形態に係る関係情報の表示例を示す図である。図８に示す例では、関係情報に含まれる第１変数と第２変数とをスライドバーを用いて変更可能にする。ユーザが第１変数又は第２変数に対してスライドバーを用いて移動させることで、例えば、移動後の第１変数（Ｐ１ｎ）又は第２変数（Ｐ２ｍ）に対応する各重みWのセットＷ（Ｐ１ｎ，Ｐ２ｍ）が、対応する点に関連付けて表示される。

【００８１】

また、ユーザは、第１変数及び第２変数の二次元のグラフ上に所定の点を指定することで、指定された点に対応する学習精度と、圧縮率との組み合わせが表示されるようにしてもよい。

【００８２】

これにより、サーバ１０は、第１変数と第２変数との組み合わせに対応する、適切な各重みWを表示可能になる。また、視覚的に対応関係をユーザに示しながら、これから分散学習が行われる任意のデータセットに対して適切な分散インスタンス数やハイパーパラメータを選択させるユーザインタフェースを提供することが可能になる。

【００８３】

＜動作＞

図９は、実施形態に係る予測モデルの生成に関する処理の一例を示すフローチャートである。図９に示す処理は、情報処理装置１０により実行される。

【００８４】

ステップＳ１０２において、情報処理装置１０の取得部１０１は、所定の学習データを取得する。所定の学習データは、記憶部１１２のデータセット１１２ａから選択されてもよいし、他の装置からネットワークを介して受信された所定のデータでもよいし、ユーザ

操作に応じて入力された所定のデータを取得してもよい。
【００８５】
ステップＳ１０４において、情報処理装置１０の第１学習部１０２は、ニューラルネットワークを用いる所定の学習モデルに対して、蒸留処理された第１学習モデル、枝刈り処理された第２学習モデル、及び量子化処理された第３学習モデルの少なくとも２つのモデルを含む各モデルがそれぞれ重み付けされた重み学習モデルに、所定のデータを入力して機械学習を行う。
【００８６】
ステップＳ１０６において、情報処理装置１０の第２学習部１０４は、各モデルそれぞれの重みが変更された重み学習モデルごとに、所定の学習データを入力して機械学習が行われた場合の学習結果を取得する。
【００８７】
ステップＳ１０８において、情報処理装置１０の第２学習部１０４は、変更された各重みが重み付けされた各重み学習モデルと、各重み学習モデルで学習されたときの各学習結果とを含む学習データを用いて、教師あり学習を行う。
【００８８】
ステップＳ１１０において、情報処理装置１０の第２学習部１０４は、教師あり学習により、任意の学習データを入力する場合に、各重みの組み合わせごとに学習結果を予測する予測モデルを生成する。
【００８９】
以上の処理により、生成された予測モデルを利用することで、ニューラルネットワークを用いる学習済みモデルを、学習精度を保ちつつ、より適切に軽量化を行うことを可能にする。
【００９０】
図１０は、実施形態に係るユーザが利用する情報処理装置２０における処理の一例を示すフローチャートである。ステップＳ２０２において、情報処理装置２０の出力部２０３は、学習対象の所定の学習データに関する情報を他の情報処理装置（例えばサーバ１０）に出力する。
【００９１】
ステップＳ２０４において、情報処理装置２０の取得部２０１は、他の情報処理装置（例えばサーバ１０）から、所定の学習データに対応する各重みを示す情報を取得する。
【００９２】
ステップＳ２０６において、情報処理装置２０の学習部２０２は、取得された各重みを所定の重み学習モデル２０２ａに適用する。
【００９３】
ステップＳ２０８において、情報処理装置２０の学習部２０２は、各重みが適用された学習モデル２０２ａに、所定の学習データを入力して学習結果を取得する。
【００９４】
これにより、エッジ側の情報処理装置であっても、学習対象のデータに対して、適切な軽量化を行った学習モデルを用いて学習を行うことで、学習精度を保つことができる。
【００９５】
以上説明した実施形態は、本発明の理解を容易にするためのものであり、本発明を限定して解釈するためのものではない。実施形態が備える各要素並びにその配置、材料、条件、形状及びサイズ等は、例示したものに限定されるわけではなく適宜変更することができる。また、第１学習部１０２を備える装置と、第２学習部１０４を備える装置とは別のコンピュータでもよい。この場合、生成された第１学習部１０２により学習された学習結果が、ネットワークを介して、第２学習部１０４を備える装置に送信されてもよい。
【００９６】
また、情報処理装置１０は、変更部１０３を必ずしも設けなくてもよい。例えば、情報処理装置１０は、任意の学習対象のデータと任意の重みのセットとの組の各学習性能を取

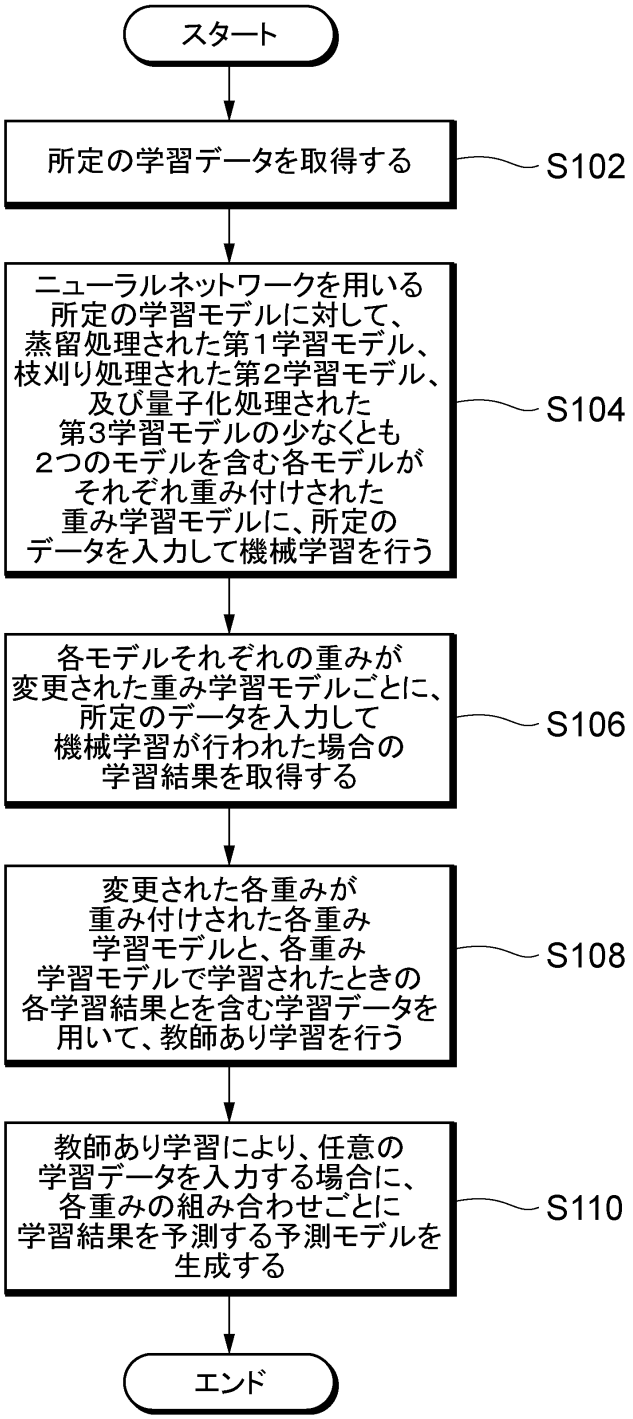
得して第２学習部１０４による学習を行ってもよい。

【符号の説明】

【００９７】

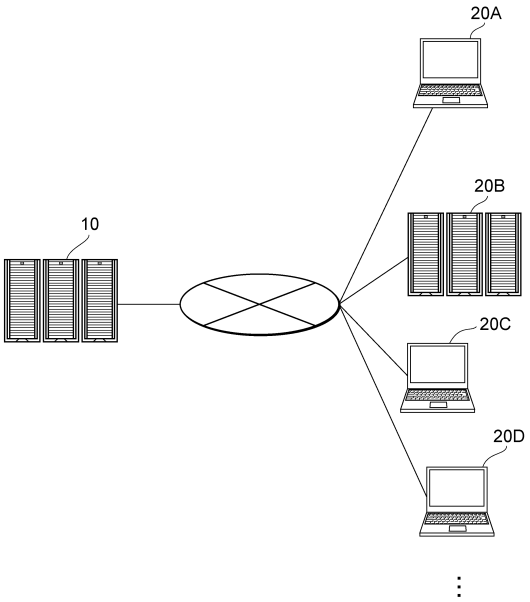
１０…情報処理装置、１０ａ…ＣＰＵ、１０ｂ…ＲＡＭ、１０ｃ…ＲＯＭ、１０ｄ…通信部、１０ｅ…入力部、１０ｆ…表示部、１０１…取得部、１０２…第１学習部、１０２ａ…学習モデル、１０３…変更部、１０４…第２学習部、１０４ａ…予測モデル、１０５…予測部、１０６…判定部、１０７…設定部、１０８…関連付け部、１０９…特定部、１１０…表示制御部、１１１…出力部、１１２…記憶部、１１２ａ…データセット、１１２ｂ…軽量化手法、１１２ｃ…関係情報、２０１…取得部、２０２…学習部、２０２ａ…学習モデル、２０３…出力部、２０４…記憶部、２０４ａ…データセット

【要約】
【課題】学習モデルに対する軽量化手法を、より適切にする。
【解決手段】情報処理方法は、情報処理装置に含まれる１又は複数のプロセッサが、所定の学習データを取得すること、ニューラルネットワークを用いる所定の学習モデルに対して、各軽量化モデルの少なくとも２つのモデルを含む各モデルがそれぞれ重み付けされた重み学習モデルに、所定のデータを入力して機械学習を行うこと、重み学習モデルごとに、所定の学習データを入力して機械学習が行われた場合の学習結果を取得すること、各重み学習モデルと、各重み学習モデルで学習されたときの各学習結果とを含む学習データを用いて、教師あり学習を行うこと、教師あり学習により、任意の学習データを入力する場合に、各重みの集合ごとに学習結果を予測する予測モデルを生成すること、を実行する。
【選択図】図９

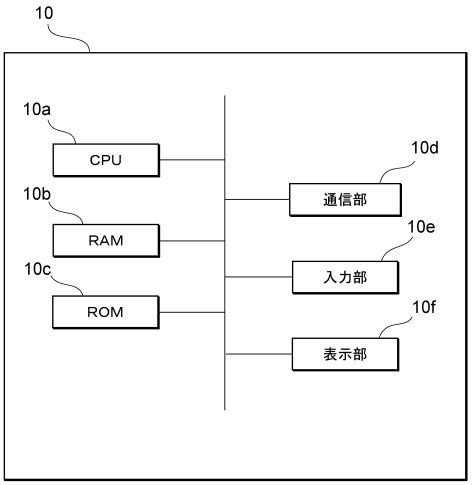

スタート
所定の学習データを取得する
S102
ニューラルネットワークを用いる
所定の学習モデルに対して、
蒸留処理された第1学習モデル、
枝刈り処理された第2学習モデル、
及び量子化処理された
第3学習モデルの少なくとも
2つのモデルを含む各モデルが
それぞれ重み付けされた
重み学習モデルに、所定の
データを入力して機械学習を行う
S104
各モデルそれぞれの重みが
変更された重み学習モデルごとに、
所定のデータを入力して
機械学習が行われた場合の
学習結果を取得する
S106
変更された各重みが
重み付けされた各重み
学習モデルと、各重み
学習モデルで学習されたときの
各学習結果とを含む学習データを
用いて、教師あり学習を行う
S108
教師あり学習により、任意の
学習データを入力する場合に、
各重みの組み合わせごとに
学習結果を予測する予測モデルを
生成する
S110
エンド

【図面】

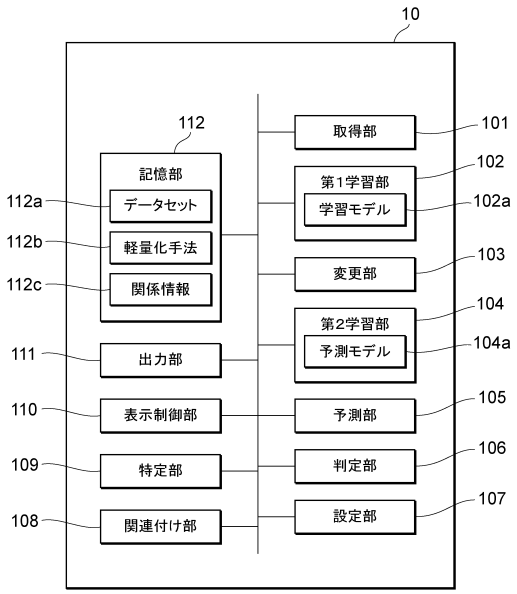
【図１】


20A
20B
10
20C
20D

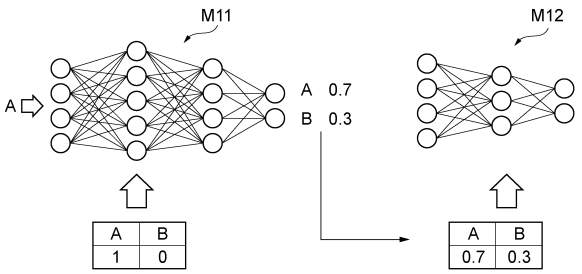
【図２】


10
10a
CPU
10b
RAM
10c
ROM
10d
通信部
10e
入力部
10f
表示部

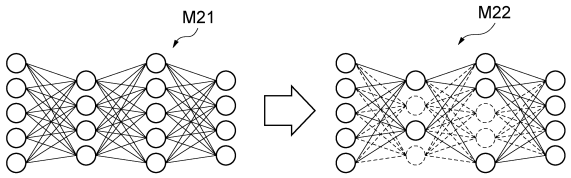
【図３】


10
112
記憶部
112a
データセット
112b
軽量化手法
112c
関係情報
111
出力部
110
表示制御部
109
特定部
108
関連付け部
取得部
101
第1学習部
学習モデル
102
102a
変更部
103
第2学習部
予測モデル
104
104a
予測部
105
判定部
106
設定部
107

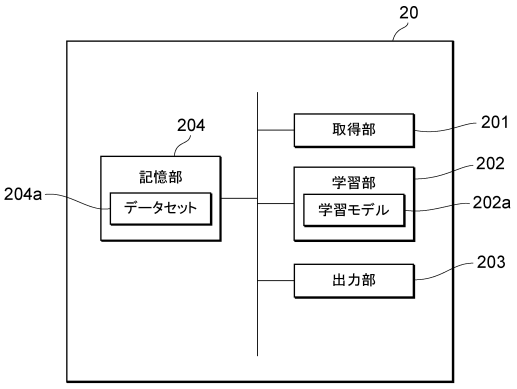
【図４】


M11
M12
A
A 0.7
B 0.3
A B
1 0
A B
0.7 0.3

【図５】



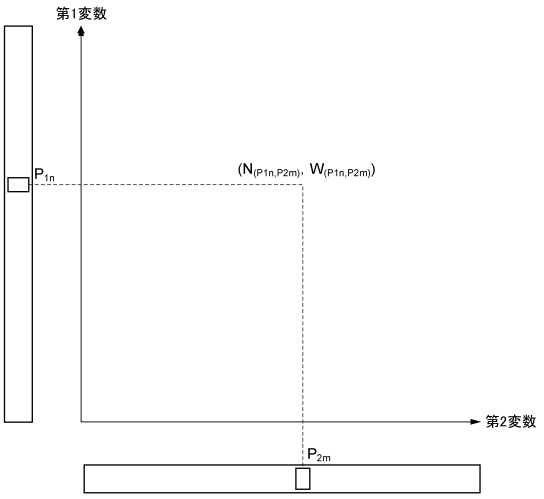
【図６】



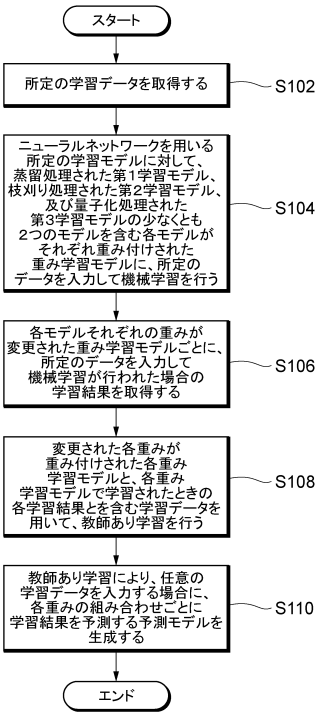
【図７】

関係情報		
第１変数	第２変数	重み
P_{11}	P_{21}	$W_1(w_1, w_2, w_3)$
・・・	・・・	・・・

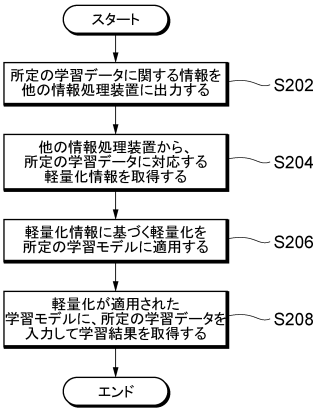
【図８】



【図９】


スタート
所定の学習データを取得する
S102
ニューラルネットワークを用いる
所定の学習モデルに対して、
蒸留処理された第1学習モデル、
枝刈り処理された第2学習モデル、
及び量子化処理された
第3学習モデルの少なくとも
2つのモデルを含む各モデルが
それぞれ重み付けされた
重み学習モデルに、所定の
データを入力して機械学習を行う
S104
各モデルそれぞれの重みが
変更された重み学習モデルごとに、
所定のデータを入力して
機械学習が行われた場合の
学習結果を取得する
S106
変更された各重みが
重み付けされた各重み
学習モデルと、各重み
学習モデルで学習されたときの
各学習結果とを含む学習データを
用いて、教師あり学習を行う
S108
教師あり学習により、任意の
学習データを入力する場合に、
各重みの組み合わせごとに
学習結果を予測する予測モデルを
生成する
S110
エンド

【図１０】


スタート
所定の学習データに関する情報を
他の情報処理装置に出力する
S202
他の情報処理装置から、
所定の学習データに対応する
軽量化情報を取得する
S204
軽量化情報に基づく軽量化を
所定の学習モデルに適用する
S206
軽量化が適用された
学習モデルに、所定の学習データを
入力して学習結果を取得する
S208
エンド

フロントページの続き

ＤＡＴＡ 品川ビル１３階 株式会社クリエイターズネクスト内

審査官　中村 信也

(56)参考文献　国際公開第２０２２／０２３０２２（WO，Ａ１）
米国特許出願公開第２０２０／０１２５９５６（ＵＳ，Ａ１）
米国特許出願公開第２０２０／０３１１５５２（ＵＳ，Ａ１）

(58)調査した分野　(Int.Cl.，ＤＢ名)

Ｇ０６Ｎ　３／００－３／１２
Ｇ０６Ｎ　７／０８－９９／００
Ｇ０６Ｎ　５／００－７／０６