(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2020年8月6日(06.08.2020)

WIPO | PCT

(10) 国際公開番号
WO 2020/158058 A1

(51) 国際特許分類:
G06N 3/08 (2006.01)　*G06N 3/04* (2006.01)

(21) 国際出願番号：　PCT/JP2019/040272

(22) 国際出願日：　2019年10月11日(11.10.2019)

(25) 国際出願の言語：　日本語

(26) 国際公開の言語：　日本語

(30) 優先権データ：
特願 2019-016217　2019年1月31日(31.01.2019)　JP

(71) 出願人: 日立オートモティブシステムズ株式会社(**HITACHI AUTOMOTIVE SYSTEMS, LTD.**) [JP/JP]; 〒3128503 茨城県ひたちなか市高場２５２０番地 Ibaraki (JP).

(72) 発明者: 村田 大智(**MURATA Daichi**); 〒1008280 東京都千代田区丸の内一丁目６番６号 株式会社日立製作所内 Tokyo (JP).

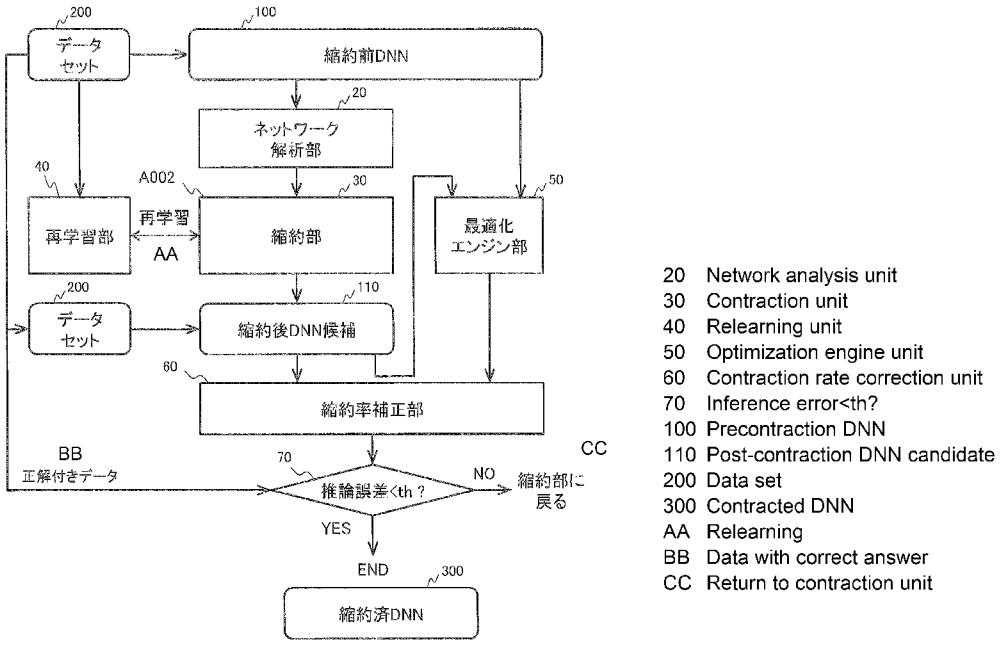
(74) 代理人: 特許業務法人藤央特許事務所 (**TOU-OU PATENT FIRM**); 〒1050001 東京都港区虎ノ門一丁目１６番４号アーバン虎ノ門ビル Tokyo (JP).

(81) 指定国(表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY,

(54) **Title:** COMPUTING DEVICE

(54) 発明の名称： 演算装置

[図2]



(57) **Abstract:** Provided is a computing device having a neural network which performs computation by using input data and a weight coefficient. The computing device has: a network analysis unit for calculating an neuronal firing state of neurons of the neural network by the input data; and a contraction unit for narrowing candidates of contraction patterns from a plurality of contraction patterns to which a contraction rate of the neural network is set on the basis of the neuronal firing state of the neurons, and performing contraction of the neural network on the basis of the narrowed candidates of

WO 2020/158058 A1

[続葉有]

MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) 指定国(表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

添付公開書類:

一 国際調査報告 (条約第21条(3))

the contraction patterns to generate a neural network after the contraction.

(57) 要約:入力データと、重み係数を用いて演算を行うニューラルネットワークを有する演算装置であって、前記入力データによる前記ニューラルネットワークのニューロンの発火状態を算出するネットワーク解析部と、前記ニューロンの発火状態に基づいて、前記ニューラルネットワークの縮約率が設定された複数の縮約パターンから縮約パターンの候補を絞り込み、前記絞り込んだ縮約パターンの候補に基づいて前記ニューラルネットワークの縮約を実行して縮約後のニューラルネットワークを生成する縮約部と、を有する。

明　細　書

発明の名称 ： 演算装置

参照による取り込み

[0001] 本出願は、平成３１年（２０１９年）１月３１日に出願された日本出願である特願２０１９－０１６２１７の優先権を主張し、その内容を参照することにより、本出願に取り込む。

技術分野

[0002] 本発明は、ニューラルネットワークを利用する演算装置に関する。

背景技術

[0003] 物体の認識や、行動の予測を自動的に行う技術として、ＤＮＮ（Ｄｅｅｐ　Ｎｅｕｒａｌ　Ｎｅｔｗｏｒｋ）を用いた機械学習が知られている。ＤＮＮを自動運転車両へ適用する場合には、車載装置の演算能力を考慮してＤＮＮの演算量を削減する必要が生じる。ＤＮＮの演算量を削減する技術としては、例えば、特許文献１が知られている。

[0004] 特許文献１には、ニューラルネットワークの重み係数の閾値を変化させて、認識精度の大きな劣化が発生する直前の閾値を決定し、当該閾値より認識精度の絶対値が小さいニューロンをプルーニングし、ＤＮＮを縮約する技術が開示されている。

先行技術文献

特許文献

[0005] 特許文献1：米国特許出願公開第２０１８／００９６２４９号明細書

発明の概要

発明が解決しようとする課題

[0006] しかしながら、上記従来技術では、再学習と推論を繰り返すことでＤＮＮの縮約（または最適化）を実施するため、自動運転車両用のＤＮＮのように大規模なニューラルネットワークに適用した場合、探索対象の組み合わせが膨大となって、処理が完了するまでに膨大な時間を要する、という問題があ

った。

[0007] また、上記従来技術では、重み係数によってニューラルネットワークの縮約を実施しているため、適用先のアプリケーションに応じて縮約を実施することが難しい、という問題があった。

[0008] そこで本発明は、上記問題点に鑑みてなされたもので、縮約の際の演算量を削減して短時間で処理を完了させることを目的とする。

課題を解決するための手段

[0009] 本発明は、入力データと、重み係数を用いて演算を行うニューラルネットワークを有する演算装置であって、前記入力データによる前記ニューラルネットワークのニューロンの発火状態を算出するネットワーク解析部と、前記ニューロンの発火状態に基づいて、前記ニューラルネットワークの縮約率が設定された複数の縮約パターンから縮約パターンの候補を絞り込み、前記絞り込んだ縮約パターンの候補に基づいて前記ニューラルネットワークの縮約を実行して縮約後のニューラルネットワークを生成する縮約部と、を有する。

発明の効果

[0010] したがって、本発明は、ニューロンの発火状態に基づいて縮約を実行できるため、縮約の際の演算量を削減して短時間で縮約処理を完了させることが可能となる。また、適用先のアプリケーション（または装置）に応じたニューラルネットワーク（ＤＮＮ）を生成することが可能となる。

[0011] 本明細書において開示される主題の、少なくとも一つの実施の詳細は、添付されている図面と以下の記述の中で述べられる。開示される主題のその他の特徴、態様、効果は、以下の開示、図面、請求項により明らかにされる。

図面の簡単な説明

[0012] [図1]本発明の実施例１を示し、ＤＮＮ縮約自動化装置の一例を示すブロック図である。

[図2]本発明の実施例１を示し、ＤＮＮ縮約自動化装置で行われる処理の一例を示す図である。

[図3]本発明の実施例１を示し、縮約パターンと縮約率と認識精度への感度の関係を示す図である。

[図4]本発明の実施例１を示し、設計期間と縮約率の関係を示すグラフである。

[図5]本発明の実施例２を示し、ＤＮＮ縮約自動化装置を車両に搭載する例を示す車両制御システムのブロック図である。

[図6]本発明の実施例３を示し、ＤＮＮ縮約自動化装置で行われる処理の一例を示す図である。

[図7]本発明の実施例４を示し、ＤＮＮ縮約自動化装置で行われる処理の一例を示す図である。

発明を実施するための形態

[0013] 以下、本発明の実施形態を添付図面に基づいて説明する。

実施例 1

[0014] 図１は、本発明の実施例１を示し、ＤＮＮ（Ｄｅｅｐ Ｎｅｕｒａｌ Ｎｅｔｗｏｒｋ）縮約自動化装置１の一例を示すブロック図である。

[0015] ＤＮＮ縮約自動化装置１は、縮約（または最適化）対象のＤＮＮ１００と、ＤＮＮ１００へ入力するデータセット２００を格納するストレージ９０と、中間データなどを保持するメモリ１０と、ネットワーク解析部２０と、縮約部３０と、再学習部４０と、最適化エンジン部５０と、縮約率補正部６０と、精度判定部７０と、ネットワーク解析部２０～精度判定部７０の各機能部を制御するスケジューラ８０と、各部を接続するインターコネクト６を含む演算装置である。なお、インターコネクト６としては、例えば、ＡＸｉ（Ａｄｖａｎｃｅｄ ｅＸｔｅｎｓｉｂｌｅ Ｉｎｔｅｒｆａｃｅ）を採用することができる。

[0016] また、メモリ１０、ネットワーク解析部２０～精度判定部７０はスレーブとして機能し、スケジューラ８０は、上記スレーブを制御するマスタとして機能する。

[0017] 本実施例１のＤＮＮ縮約自動化装置１は、ネットワーク解析部２０～精度

判定部７０の各機能部及びスケジューラ８０がハードウェアで実装される。ＤＮＮ縮約自動化装置１は、例えば、計算機の拡張スロットに装着されてデータの授受を行うことができる。なお、ハードウェアとしては、ＡＳＩＣ（Ａｐｐｌｉｃａｔｉｏｎ Ｓｐｅｃｉｆｉｃ Ｉｎｔｅｇｒａｔｅｄ Ｃｉｒｃｕｉｔ）等を採用することができる。

[0018] また、本実施例１では、ハードウェアで各機能部を構成する例を示すが、これに限定されるものではなない。例えば、ネットワーク解析部２０～スケジューラ８０の一部又は全てをソフトウェアで実装することも可能である。また、以下の説明では、ＤＮＮの各層をニューラルネットワークとして説明する。

[0019] ストレージ９０に格納される縮約前のＤＮＮ１００は、ニューラルネットワークと重み係数及びバイアスを含む。また、データセット２００は、ＤＮＮ１００の適用先のアプリケーション（または装置）に応じたデータで、正解付きのデータと、ニューラルネットワークの発火（活性化）状態を検出するためのデータが含まれる。縮約済みのＤＮＮ３００は、ネットワーク解析部２０～精度判定部７０で縮約処理を実行した結果である。

[0020] スケジューラ８０は、縮約前のＤＮＮ１００とデータセット２００を受け付けると、予め設定された順序で上記各機能部を制御してニューラルネットワーク（ニューロン）の縮約処理を実行し、縮約済みのＤＮＮ３００を生成させる。

[0021] 本実施例１のＤＮＮ縮約自動化装置１では、入力された縮約前のＤＮＮ１００と適用先のアプリケーションに応じたデータセット２００から、最適な縮約率を自動的に算出し、縮約済みのＤＮＮ３００の縮約に要する設計期間の短縮を実現する。

[0022] 本実施例１では、縮約率を、縮約済みのＤＮＮ３００の演算量／縮約前のＤＮＮ１００の演算量で表す。ただし、演算量は、単位時間当たりの処理量数（Ｏｐｅｒａｔｉｏｎ ｐｅｒ ｓｅｃｏｎｄ）を用いることができる。なお、縮約率としては、上記の他に、縮約済みのＤＮＮ３００のニューロン

数／縮約前のＤＮＮ１００のニューロン数で表すことができ、あるいは、縮約済みのＤＮＮ３００のノード数／縮約前のＤＮＮ１００のノード数で表すことができる。

[0023] 以下、ＤＮＮ縮約自動化装置１で行われる処理の概要を説明した後に、各機能部の詳細について説明する。

[0024] ＜処理の概要＞

まず、スケジューラ８０は、縮約前のＤＮＮ１００をネットワーク解析部２０へ入力する。スケジューラ８０は、データセット２００から適用先アプリケーションに応じたデータをネットワーク解析部２０へ入力して、ＤＮＮ１００の特徴量を算出させる。

[0025] ネットワーク解析部２０では、データセット２００のデータをＤＮＮ１００へ入力し、ニューラルネットワークのニューロンの発火状態から特徴量を算出する。そして、スケジューラ８０は、ネットワーク解析部２０で算出された特徴量を縮約部３０へ入力して、有望な縮約率の組合せ候補について絞り込みを実施させる。

[0026] 縮約部３０では、特徴量からニューラルネットワークの認識精度への感度を算出し、感度が低い部分については縮約率を高く設定し、感度が高い部分については縮約率を低く設定する。

[0027] 縮約部３０は、ＤＮＮ１００の各層のニューラルネットワークについて上記縮約率を設定し、縮約率の組合せ候補を複数生成し、これらの候補の中から縮約率と認識精度（への感度）の条件を満たす候補を絞り込む。なお、以下の説明では、縮約率の組合せ候補を縮約パターンとする。そして、縮約部３０は、絞り込んだ縮約パターンについてＤＮＮ１００の縮約を実行し、縮約後のＤＮＮの候補（ＤＮＮ候補１１０）として出力する。

[0028] スケジューラ８０は、縮約部３０が縮約を実行中に、再学習部４０によってＤＮＮの再学習を反復的に実行させる。再学習部４０は、再学習によって縮約に対して頑健なＤＮＮ候補１１０を構築する。次に、スケジューラ８０は、縮約部３０から出力された縮約後のＤＮＮ候補１１０と、ＤＮＮ１００

を最適化エンジン部５０へ入力して、最適化を実施させる。

[0029] 最適化エンジン部５０は、縮約後のＤＮＮ候補１１０に対して縮約率の最適化や、縮約方法の選択などを実施し、縮約に必要なパラメータ（例えば、重み係数など）の補正値を決定する。最適化エンジン部５０は、例えば、縮約後のＤＮＮ候補１１０の推論誤差からベイズ推論に基づく最適化アルゴリズムを用いて、最適な縮約パターンやパラメータを推定し、ニューラルネットワーク毎の縮約率の補正値を決定する。

[0030] 最適化エンジン部５０は、算出された縮約パターンやパラメータを縮約率補正部６０へ出力する。縮約率補正部６０では、縮約後のＤＮＮ候補１１０に上記縮約率とパラメータを適用して縮約率を補正し、縮約後のＤＮＮ候補１１０を構築する。スケジューラ８０は、縮約率補正部６０で構築された縮約後のＤＮＮ候補１１０を精度判定部７０へ入力し、推論を実行させる。

[0031] 精度判定部７０は、データセット２００から正解付きデータを取得して、縮約後のＤＮＮ候補１１０へ入力して推論を実施する。精度判定部７０は、推論の結果と正解から縮約後のＤＮＮ候補１１０の推論誤差（または推論精度）を判定し、推論誤差が所定の閾値ｔｈ未満となるまで上記処理を繰り返す。なお、推論誤差は、例えば、ＤＮＮ候補１１０の推論結果の正解率の逆数に基づく統計値（平均値など）を用いればよい。

[0032] そして、精度判定部７０は、縮約部３０で絞り込まれた縮約パターンのうち、推論誤差が所定の閾値ｔｈ未満のＤＮＮ候補１１０を、最適化が完了した縮約済みのＤＮＮ３００として出力する。

[0033] 以上のように、ＤＮＮ縮約自動化装置１は、ネットワーク解析部２０によるＤＮＮ１００の解析（１）と、縮約部３０による複数の縮約率の組み合わせ（縮約パターン）の候補の絞り込み及び縮約の実行（２）と、再学習部４０による縮約対象のＤＮＮの再学習（３）と、最適化エンジン部５０によるパラメータの最適化及び縮約率補正部６０での縮約後のＤＮＮ候補１１０の再構築（４）と、精度判定部７０による縮約後のＤＮＮ候補１１０の推論誤差の判定（５）を実施し、複数の縮約パターンの中から推論誤差が閾値ｔｈ

未満のＤＮＮ３００を自動的に出力することができる。

[0034] ＤＮＮ縮約自動化装置１は、縮約前のＤＮＮ１００を解析して、上記（１）～（５）の処理を推論誤差が所定の閾値ｔｈ未満となるまで繰り返すことで、複数の縮約パターンの中から縮約率と推論精度（認識精度）に優れた縮約済みのＤＮＮ３００を、ＤＮＮ３００の適用先のアプリケーション（または装置）に応じて自動的に生成することが可能となる。

[0035] ＤＮＮ縮約自動化装置１は、ＤＮＮ３００の適用先のアプリケーションに応じたデータセットで、縮約前のＤＮＮ１００のニューラルネットワークを解析して、特徴量（発火状態）を算出しておくことで、有望な縮約率の組み合わせを絞り込んでから探索することが可能となり、縮約の際の演算量を削減して短時間で処理を完了させることが可能となる。

[0036] さらに、ＤＮＮ縮約自動化装置１は、縮約パターンの候補の絞り込みに加えて、ベイズ推論による確率的探索を組み合わせることで、閾値ｔｈを満足する範囲内で認識精度の低下を最小とする縮約済みのＤＮＮ３００を出力することができる。

[0037] ＜機能部の詳細＞

まず、ネットワーク解析部２０は、縮約による認識精度への感度を解析し、縮約前のＤＮＮ１００のニューラルネットワーク毎に特徴量を算出する。ネットワーク解析部２０は、縮約済みのＤＮＮ３００の適用先のアプリケーションに応じた複数のデータをデータセット２００から読み込んで、縮約前のＤＮＮ１００へ順次入力し、ＤＮＮ１００のニューラルネットワーク毎に発火状態を推定（数値化）して特徴量とする。

[0038] また、ネットワーク解析部２０は、ニューラルネットワークのニューロンの発火状態をヒートマップとして算出し、このヒートマップを特徴量としてもよい。また、ネットワーク解析部２０が算出する特徴量は、ニューラルネットワーク毎に限定されるものではなく、例えば、ニューロン毎に算出してもよい。

[0039] 各ニューロンの発火状態を推定して数値化する技術としては、公知または

周知の技術を適用することができ、例えば、国際公開第２０１１／００７５６９号に開示される技術を適用すればよい。

[0040] 本実施例１では、適用先のデータの特徴に応じて発火するニューロンと、発火しないニューロンの分布が異なる点に着目し、ＤＮＮ３００の適用先のアプリケーションに応じたデータセット２００によるニューロンの発火状態を特徴量とする。なお、特徴量は、複数のデータをＤＮＮ１００へ順次入力した場合の統計値としてもよい。また、特徴量は、縮約済みＤＮＮ３００の適用先に特有の特徴を包含した解析結果として出力される。

[0041] ネットワーク解析部２０は、入力されたデータに対して頻繁に発火するニューロン（またはニューラルネットワーク）を認識精度への感度が大きいと判定することができ、逆に、発火する頻度の低いニューロン（またはニューラルネットワーク）を認識精度への感度が低いと判定することができる。

[0042] 縮約部３０は、ネットワーク解析部２０から、ニューラルネットワーク（またはニューロン）の発火状態に基づく特徴量を受け付けて、縮約率の組み合わせ（縮約パターン）の候補を絞り込んで、縮約を実行する。縮約部３０は、ニューラルネットワーク毎の特徴量に基づいて、複数の縮約パターンの候補から絞り込みを実施し、絞り込んだ複数の縮約パターンについて縮約を実行し、縮約後のＤＮＮ候補１１０を生成する。

[0043] 図３は、縮約パターンと縮約率と認識精度への感度の関係を示す図である。図３の例では、ＤＮＮ１００をｎ層のニューラルネットワークで構成し、各層毎に縮約率を設定する例を示す。図示の例では、１層目が入力層、２層目～ｎ－１層目が隠れ層（中間層）、ｎ層目が出力層となる。

[0044] 本実施例１において、１つの縮約パターンは、各層の縮約率をそれぞれ有する。換言すれば、縮約パターンは、各層毎の縮約率の組み合わせで構成される。

[0045] 縮約パターン１～３は、各層（ニューラルネットワーク）毎に縮約率が異なる組み合わせで設定される。縮約パターンは、予め設定されたパターンを使用してもよいし、予め設定された縮約率の組み合わせから縮約部３０が生

成してもよい。また、縮約パターンの数は、３つに限定されるものではなく、ＤＮＮ１００の規模に応じて適宜変更することができる。

[0046] 縮約部３０は、上述したように、ＤＮＮ３００の適用先に対応するデータの認識精度に対する感度が高いニューラルネットワークについては、縮約率を低く設定する。これにより、上記感度が高い領域では、ニューロンの数が必要以上に減少して認識（推定）精度が低下するのを抑制する。

[0047] 一方、ＤＮＮ３００の適用先に対応するデータの認識精度に対する感度が低いニューラルネットワークについては、縮約率を高く設定する。これにより、感度が低い領域では、ニューロンの数を大幅に削減しても認識精度が低下するのを抑制でき、かつ、演算量も削減できる。

[0048] 感度と縮約率の関係は、例えば、認識精度に対する感度が７０％のニューラルネットワークの縮約率を３０％とし、認識精度に対する感度が３０％のニューラルネットワークの縮約率を７０％とする。

[0049] 縮約率を増大させていくと、削減可能なニューロンが連鎖的に増大するため、演算量を大幅に削減することが可能となる。一方で、認識精度に対する感度に関わらず縮約率を増大させると、認識精度が低下（推定誤差の拡大）するという問題が発生するが、本実施例１のように、ニューラルネットワークの特徴量を認識精度に対する感度と関連させることで、縮約率と認識精度の最適解を探索することが可能となる。

[0050] なお、上記ではニューラルネットワーク毎に縮約率を設定する例を示したが、これに限定されるものではない。例えば、各層毎の縮約率を維持しながら、ニューラルネットワーク内のニューロンの特徴量に応じて、縮約すべきニューロンと維持すべきニューロンを分類するようにしてもよい。

[0051] このように、縮約部３０は、特徴量に基づいてニューラルネットワーク毎の縮約率を決定することで、ニューロンの数を大幅に削減してから縮約パターンの最適化などの演算を行うことが可能となって、演算時間を短縮することが可能となる。

[0052] 次に、縮約部３０は、複数の縮約パターンから絞り込みを実行して、縮約

処理の演算時間を現実的な値とする。絞り込みの例としては、ＤＮＮ全体の縮約率と認識精度への感度の降順で上位から所定順位までの縮約パターンに絞り込む。あるいは、縮約率が所定値以上の縮約パターンに絞り込む等、絞り込みについては、公知または周知の技術を適用すればよい。

[0053] 縮約部３０は、絞り込みを行った複数の縮約パターンについて、縮約を実施して縮約後のＤＮＮ候補１１０として出力する。

[0054] 再学習部４０は、上述したように、縮約部３０で縮約中のＤＮＮに対してデータセット２００による再学習を実施する。これにより、汎化性能の高い(＝縮約に対して頑健な)ＤＮＮを構築することができる。

[0055] 再学習部４０は、縮約中のＤＮＮと当該ＤＮＮのパラメータ（重み係数）の最適解の候補を入力として受け付けて、受け付けたＤＮＮおよびパラメータを初期値として、再び学習を行うことでＤＮＮを再構築する。再構築された結果は、再学習済みのニューラルネットワーク及び再学習済みの重み係数として出力される。

[0056] 最適化エンジン部５０は、縮約部３０から出力された複数のＤＮＮ候補１１０について、データセット２００による推論を実施して推論誤差を算出し、推論誤差に基づいて最適な縮約率の組み合わせ（縮約パターン）を推定する。すなわち、最適化エンジン部５０は、ベイズ推論に基づく確率論的探索を実施して、ニューロン毎に適切な縮約率を確率的に決定する。そして、最適化エンジン部５０は、決定された縮約率の組み合わせ（縮約パターン）を縮約率補正部６０へ出力する。

[0057] 最適化エンジン部５０は、縮約部３０から出力された、複数のＤＮＮ候補１１０に対応する縮約パターンの中から推論誤差が最小となる縮約パターンを算出する。

[0058] また、最適化エンジン部５０は、縮約部３０から、複数のＤＮＮ候補１１０および再学習済みの重み係数を入力として受け付けて、ベイズ推論に基づく確率論的探索を用いて、縮約パターンを推定してもよい。

[0059] 縮約率補正部６０は、最適化エンジン部５０から受け付けた縮約率で縮約

後のＤＮＮ候補１１０の縮約率を補正して再構築する。

[0060] 精度判定部７０は、ＤＮＮ候補１１０へ正解付きデータを入力して推論を実施し、推論の結果と正解から縮約後のＤＮＮ候補１１０の推論誤差が所定の閾値ｔｈ未満であれば、縮約が完了した縮約済みＤＮＮ３００として出力する。

[0061] 一方、精度判定部７０は、推論誤差が所定の閾値ｔｈ以上の場合には、スケジューラ８０へ処理の繰り返しを通知する。スケジューラ８０は、精度判定部７０から処理の繰り返しの通知を受け付けると、縮約部３０に処理の繰り返しを実行させる。

[0062] 以上のように、ＤＮＮ縮約自動化装置１は、ネットワーク解析部２０がニューロンの発火状態に基づいて特徴量を算出し、縮約部３０が有望な縮約パターンに絞り込んでから縮約を実行してＤＮＮ候補１１０を複数出力し、再学習部４０が縮約対象のＤＮＮ候補１１０の再学習を実施し、最適化エンジン部５０が、推論の誤差に基づいて適正な縮約率を算出し、縮約率補正部６０が適正な縮約率でＤＮＮ候補１１０を再構築し、精度判定部７０が縮約後のＤＮＮ候補１１０の推論誤差の判定を実施し、複数の縮約パターン（ＤＮＮ候補１１０）の中から推論誤差が閾値ｔｈ未満のＤＮＮ３００を自動的に出力することができる。

[0063] ＤＮＮ縮約自動化装置１は、ＤＮＮ３００の適用先のアプリケーション（または装置）に応じたデータセット２００で発火状態に基づく特徴量を算出することで、縮約率と認識精度に優れた縮約パターンに絞り込むことが可能となって、縮約の際の演算量を削減して短時間で縮約処理を完了させることが可能となる。また、ＤＮＮ縮約自動化装置１は、ＤＮＮ１００の縮約処理に、人手を必要としないため、ＤＮＮ１００の縮約にかかる労力を大幅に削減することが可能となる。

[0064] また、本実施例１のＤＮＮ縮約自動化装置１は、適用先のアプリケーションに応じたデータセット２００でニューロンの発火状態を推定するので、縮約済みＤＮＮ３００の適用先の環境に応じたＤＮＮを生成することが可能と

なる。

[0065] 図４は、ＤＮＮ１００の縮約に要する設計期間と縮約率の関係を示すグラフである。図示のグラフは、横軸を縮約率とし、縦軸を縮約の設計期間とする。

[0066] 図中実線は、本実施例１のＤＮＮ縮約自動化装置１で大規模なＤＮＮ１００を縮約した場合の縮約率と設計期間（処理時間）の関係を示す。図中破線は、大規模なＤＮＮ１００を人手によって縮約を実施した例を示す。

[0067] 本実施例１のＤＮＮ縮約自動化装置１では、人手によって７．８日を要していた縮約率＝７０％の縮約を、約１／１０の≒１０時間で完了することができる。さらに、本実施例１のＤＮＮ縮約自動化装置１では、ネットワーク解析部２０による有望な縮約率の組み合わせ（縮約パターン）を絞り込むことにより、縮約の設計期間を大幅に短縮し、かつ、認識精度を向上させることが可能となる。

実施例 2

[0068] 図５は、本発明の実施例２を示し、ＤＮＮ縮約自動化装置を車両に搭載する例を示す車両制御システムのブロック図である。本実施例２では、前記実施例１に示したＤＮＮ縮約自動化装置１を、自動運転が可能な車両（エッジ）３とデータセンタ（クラウド）４にそれぞれ配置して、自動運転を行う車両３の走行環境に応じてＤＮＮ１００Ｂの縮約を最適化する例を示す。

[0069] データセンタ４は、ＤＮＮ縮約自動化装置１Ａと、大規模なＤＮＮ１００Ａに学習を行う学習装置５を含み、ＤＮＮ１００Ａの大幅なアップデートを実行する。データセンタ４は、無線ネットワーク（図示省略）を介して車両３と接続される。

[0070] 学習装置５は、車両３から走行環境や走行状態に関する情報を取得する。学習装置５は、車両３から取得した情報でＤＮＮ１００Ａの学習を実行する。学習装置５は、学習が完了したＤＮＮ１００Ａを縮約前のＤＮＮとしてＤＮＮ縮約自動化装置１Ａに入力する。

[0071] ＤＮＮ縮約自動化装置１Ａは、前記実施例１と同様に構成され、縮約済み

ＤＮＮを出力する。データセンタ４は、ＤＮＮ縮約自動化装置１Ａから出力されたＤＮＮを、所定のタイミングで車両３へ送信し、アップデートを依頼する。

[0072] 車両３は、カメラ２１０と、ＬｉＤＡＲ（Ｌｉｇｈｔ Ｄｅｔｅｃｔｉｏｎ Ａｎｄ Ｒａｎｇｉｎｇ）２２０と、レーダ２３０のセンサ類と、センサからのデータを組み合わせるフュージョン２４０と、カメラ２１０とフュージョン２４０からの情報に基づいて自動運転を実行する自動運転ＥＣＵ（Ｅｌｅｃｔｒｏｎｉｃ Ｃｏｎｔｒｏｌ Ｕｎｉｔ）２を有する。なお、カメラ２１０とフュージョン２４０で収集した情報は、無線ネットワークを介してデータセンタ４へ送信される。

[0073] 自動運転ＥＣＵ２は、運転シーン識別部１２０と、ＤＮＮ縮約自動化装置（エッジ）１Ａと、ＤＮＮ１００Ｂと、推論回路７００を含む。

[0074] 運転シーン識別部１２０は、カメラ２１０からの画像と、フュージョン２４０からのセンサデータに基づいて車両３の走行環境を検出し、走行環境が変化した場合には、ＤＮＮ１００Ｂの補正をＤＮＮ縮約自動化装置１Ｂに指令する。運転シーン識別部１２０が検出する走行環境としては、例えば、一般道や高速道路などの道路種別や、時間帯や、天候等が含まれる。

[0075] 運転シーン識別部１２０がＤＮＮ縮約自動化装置１Ｂへ指令するＤＮＮ１００Ｂの補正の内容としては、例えば、縮約の条件や、縮約の方法であり、これらの補正内容が走行環境に応じて予め設定される。

[0076] ＤＮＮ縮約自動化装置１Ｂは、指令された補正内容でＤＮＮ１００Ｂを縮約して、縮約後のＤＮＮを推論回路７００に出力する。推論回路７００は、縮約後のＤＮＮを用いて、センサデータやカメラ２１０の画像データから所定の認識処理を実行し、制御系（図示省略）へ出力する。なお、制御系とは、駆動力制御装置や操舵装置や制動装置やナビゲーション装置を含む。

[0077] データセンタ４では、車両３から収集したセンサデータで大規模なＤＮＮ１００Ａの学習処理を行い、ＤＮＮ縮約自動化装置１Ａは、学習済みのＤＮＮ１００Ａに集約処理を行ってアップデートを行う。アップデートの内容は

、例えば、認識対象の追加や、誤認識の低減などが含まれ、ＤＮＮ１００Ａの認識精度を向上させる。

[0078] 車両３では、運転シーン識別部１２０が、走行環境の変化を検出すると、ＤＮＮ縮約自動化装置１ＢがＤＮＮ１００Ｂの補正を実行することで、走行環境に適応した認識精度を確保することが可能となる。

[0079] また、車両３では、データセンタ４からアップデートが行われたＤＮＮを受信して、ＤＮＮ１００Ｂを更新することで、最新のＤＮＮで自動運転を実現することが可能となる。

実施例 3

[0080] 図６は、本発明の実施例３を示し、ＤＮＮ縮約自動化装置で行われる処理の一例を示す図である。本実施例３では、前記実施例１に示したＤＮＮ縮約自動化装置１の特徴量の算出方法と、縮約方法を複数有する例を示す。なお、その他の構成は、前記実施例１のＤＮＮ縮約自動化装置１と同様である。

[0081] ネットワーク解析部２０は、ＳｍｏｏｔｈＧｒａｄ２１と、発火状態抽出２２と、重み係数解析２３と、解析結果マージ２４を含む。

[0082] ＳｍｏｏｔｈＧｒａｄ２１は、ＤＮＮ１００が物体を認識する際に、ニューラルネットワークは注視する入力画像の領域を出力する。発火状態抽出２２は、データの認識の際にニューラルネットワークのニューロンがゼロか非ゼロかを出力する。重み係数解析２３は、ＤＮＮ１００のニューロンの結合の強さ（重み）を解析し、結合の弱い部分を縮約の対象とすることができる。

[0083] 解析結果マージ２４は、ＳｍｏｏｔｈＧｒａｄ２１と、発火状態抽出２２と、重み係数解析２３の結果を統合し、ニューラルネットワークの特徴量を算出する。

[0084] 縮約部３０は、プルーニング３１と、Ｌｏｗ　ｒａｎｋ（低ランク）近似３２と、Ｗｅｉｇｈｔ　Ｓｈａｒｉｎｇ（重み共有）３３と、低ビット化３４を含む。

[0085] プルーニング３１と、Ｌｏｗ　ｒａｎｋ近似３２では、不要または影響の

少ないニューロンを削減して、縮約を実行する。Ｗｅｉｇｈｔ　Ｓｈａｒｉｎｇ３３では、複数のニューロンの結合で重み係数を共有することでデータ量を削減する。低ビット化３４では、演算で使用するビット幅を制限することで演算負荷を削減する。ただし、ビット幅の制限は、推論誤差が許容される範囲内とする。

[0086] 縮約部３０は、上記４つの縮約方法のいずれか、あるいは複数の縮約方法の組み合わせで縮約を実行する。いずれの縮約方法を適用するかは、スケジューラ８０が指令してもよい。

[0087] また、再学習部４０の一例としては、ＢＣ（Ｂｅｔｗｅｅｎ－ｃｌａｓｓ）　ｌｅａｒｎｉｎｇ４１を適用することで、縮約した後にも認識精度を確保可能なＤＮＮを生成することができる。

[0088] ネットワーク解析部２０と、縮約部３０及び再学習部４０は、上記のような構成要素を利用することで、縮約率と認識精度に優れたＤＮＮを生成することができる。例えば、前記実施例２のように、エッジデバイス（自動運転ＥＣＵ２）のように、走行環境に応じてＤＮＮの補正を行う場合には、縮約部３０の縮約方法を上記のプルーニング３１～低ビット化３４から選択するようにしてもよい。

[0089] また、縮約部３０は、縮約手法が異なる複数の縮約実行部として、プルーニング３１と、Ｌｏｗ　ｒａｎｋ（低ランク）近似３２と、Ｗｅｉｇｈｔ　Ｓｈａｒｉｎｇ（重み共有）３３と、低ビット化３４を例示したが、これらに限定されるものではない。縮約済み胃ＤＮＮ３００の適用先に応じた縮約手法を適宜採用すればよい。

実施例 4

[0090] 図７は、本発明の実施例４を示し、ＤＮＮ縮約自動化装置１で行われる処理の一例を示す図である。本実施例４では、前記実施例３に示したＤＮＮ縮約自動化装置１の縮約部３０で、プルーニング３１と、Ｌｏｗ　ｒａｎｋ近似３２で、縮約情報を共有させる。

[0091] プルーニング３１で縮約するニューロンと、Ｌｏｗ　ｒａｎｋ近似３２で

縮約する行列を連携させることで、不要な演算を削減して処理の高速化を図ることができ、縮約部３０における演算量を削減して、ＤＮＮ縮約自動化装置１の縮約にかかる時間を短縮することができる。

[0092] ＜まとめ＞

以上のように、上記実施例１～４のＤＮＮ縮約自動化装置１は、以下のような構成とすることができる。

[0093] （１）．入力データ（データセット２００）と、重み係数を用いて演算を行うニューラルネットワーク（ＤＮＮ１００）を有する演算装置（ＤＮＮ縮約自動化装置１）であって、前記入力データ（２００）による前記ニューラルネットワーク（ＤＮＮ１００）のニューロンの発火状態を算出するネットワーク解析部（２０）と、前記ニューロンの発火状態に基づいて、前記ニューラルネットワーク（１００）の縮約率が設定された複数の縮約パターンから縮約パターンの候補を絞り込み、前記絞り込んだ縮約パターンの候補に基づいて前記ニューラルネットワーク（１００）の縮約を実行して縮約後のニューラルネットワーク（１１０）を生成する縮約部３０と、を有する。

[0094] ネットワーク解析部２０は、適用先の特徴に応じて発火するニューロンと、発火しないニューロンの分布が異なる点に着目し、ＤＮＮ３００の適用先のアプリケーションに応じたデータセット２００によるニューロンの発火状態を特徴量とする。そして、ネットワーク解析部２０は、ニューラルネットワーク（ＤＮＮ１００）の特徴量を認識精度に対する感度と関連させることで、縮約率と認識精度の最適解を探索することが可能となる。

[0095] 縮約部３０は、特徴量に基づいてニューラルネットワーク毎の縮約率を決定することで、ニューロンの数を大幅に削減してから縮約パターンの最適化などの演算を行うことが可能となって、縮約に要する演算時間を短縮することが可能となる。

[0096] （２）．上記（１）に記載の演算装置であって、前記縮約部（３０）で生成された前記縮約後のニューラルネットワーク（１１０）に推論を実行して推論誤差を算出し、前記複数の縮約パターンの中から前記推論誤差に基づい

て縮約パターンを抽出する最適化エンジン部を（５０）さらに有する。

[0097] 上記構成により、ＤＮＮ縮約自動化装置１は、最適化エンジン部５０では、縮約後のＤＮＮ候補１１０の推論誤差を縮約率（縮約パターン）にフィードバックすることで、認識精度の高い縮約済みＤＮＮ３００を生成することができる。

[0098] （３）．上記（２）に記載の演算装置であって、前記最適化エンジン部（５０）は、前記推論誤差が最小となる縮約パターンを抽出する。

[0099] 上記構成により、ＤＮＮ縮約自動化装置１は、推論誤差が最小となる縮約パターンによって、認識精度の高い縮約済みＤＮＮ３００を生成することができる。

[0100] （４）．上記（１）に記載の演算装置であって、前記縮約部（３０）で生成された前記縮約後のニューラルネットワーク（１１０）に対して、前記入力データ（２００）によって再度学習を実施する再学習部（４０）を、さらに有する。

[0101] 上記構成により、汎化性能の高い(縮約に対して頑健な)ＤＮＮを構築することができる

[0102] （５）．上記（２）に記載の演算装置であって、前記縮約部（３０）で生成された前記縮約後のニューラルネットワーク（１１０）に対して、前記入力データ（２００）によって再度学習を実施する再学習部（４０）を、さらに有し、前記ネットワーク解析部（２０）と、前記縮約部（３０）および前記最適化エンジン部（５０）ならびに前記再学習部（４０）の演算途中における中間データを一時的に格納するメモリ（１０）と、前記ネットワーク解析部（２０）と、前記縮約部（３０）と、前記再学習部（４０）と、前記最適化エンジン部（５０）および前記メモリ（１０）をスレーブとし、前記スレーブを制御するマスタとしてのスケジューラ（８０）と、前記マスタと前記スレーブを接続するインターコネクト（５）と、をさらに有する。

[0103] 上記構成により、ＤＮＮ縮約自動化装置１をハードウェアで構成することで、縮約処理の高速化を図ることが可能となる。

[0104] （６）．上記（１）に記載の演算装置であって、前記ネットワーク解析部（２０）は、前記ニューラルネットワーク（１００）と、前記縮約後のニューラルネットワーク（３００）の適用先に応じた入力データ（２００）を受け付け、前記ニューラルネットワーク（１００）の各ニューロンの発火状態を推定して数値化した特徴量を算出し、当該特徴量を前記適用先に特有の特徴を包含した解析結果として出力する。

[0105] 縮約済みＤＮＮ３００の適用先に応じたデータセット２００によるニューロンの発火状態に基づく特徴量を解析結果とすることで、適用先のアプリケーションに最適な縮約率の組み合わせを提供することが可能となる。

[0106] （７）．上記（６）に記載の演算装置であって、前記縮約部（３０）は、前記ネットワーク解析部（２０）の解析結果を受け付け、前記解析結果において数値化された特徴量に基づいてニューラルネットワーク（１００）の縮約を実行し、前記縮約後のニューラルネットワーク（１１０）および前記重み係数の最適解候補を複数出力する。

[0107] 上記構成によって、ＤＮＮ縮約自動化装置１は、特徴量を算出しておくことで、縮約率と認識精度に優れた縮約パターンに絞り込むことが可能となって、縮約の際の演算量を削減して短時間で縮約処理を完了させることが可能となる。また、ＤＮＮ縮約自動化装置１は、ＤＮＮ１００の縮約処理に、人手を必要としないため、ＤＮＮ１００の縮約にかかる労力を大幅に削減することが可能となる。

[0108] （８）．上記（１）に記載の演算装置であって、前記縮約部（３０）は、縮約手法が異なる複数の縮約実行部（プルーニング３１、低ランク近似３２、重み共有３３、低ビット化３４）を有し、前記ニューラルネットワーク（３００）の適用先に応じて前記縮約実行部（３１）～３４を切り替える。

[0109] 上記構成によって、縮約部３０は、縮約済みＤＮＮ３００の適用先に応じた縮約の手法を選択することが可能となって、処理時間の削減と、認識精度の向上を図ることができる。

[0110] （９）．上記（７）に記載の演算装置であって、前記縮約部（３０）が出

力する縮約後のニューラルネットワーク（１１０）に対して、前記入力データ（２００）によって再度学習を実施する再学習部（４０）を、さらに有し、前記再学習部（４０）は、前記ニューラルネットワーク（２００）および前記重み係数の最適解候補を入力として受け付け、前記ニューラルネットワーク（２００）および前記重み係数を初期値として、再び学習を行うことで、再学習済みのニューラルネットワーク（１１０）及び再学習済みの重み係数を出力する。

[0111] 上記構成によって、再学習部４０は、縮約した後にも認識精度を確保可能なＤＮＮ３００を生成することができる。

[0112] （１０）．上記（９）に記載の演算装置であって、前記縮約部（３０）で縮約を実行した縮約後のニューラルネットワーク（１１０）に推論を実行して推論誤差を算出し、前記複数の縮約パターンの中から前記推論誤差に基づいて縮約パターンを抽出する最適化エンジン部（５０）を、さらに有し、前記最適化エンジン部（５０）は、複数の前記ニューラルネットワーク（１１０）および前記再学習済みの重み係数を入力として受け付けて、予め設定された確率論的探索を用いて、前記縮約パターンを算出する。

[0113] 上記構成によって、最適化エンジン部５０は、推論誤差を低減可能な縮約パターンを推定することが可能となる。

[0114] なお、本発明は上記した実施例に限定されるものではなく、様々な変形例が含まれる。例えば、上記した実施例は本発明を分かりやすく説明するために詳細に記載したものであり、必ずしも説明した全ての構成を備えるものに限定されるものではない。また、ある実施例の構成の一部を他の実施例の構成に置き換えることが可能であり、また、ある実施例の構成に他の実施例の構成を加えることも可能である。また、各実施例の構成の一部について、他の構成の追加、削除、又は置換のいずれもが、単独で、又は組み合わせても適用可能である。

[0115] また、上記の各構成、機能、処理部、及び処理手段等は、それらの一部又は全部を、例えば集積回路で設計する等によりハードウェアで実現してもよ

い。また、上記の各構成、及び機能等は、プロセッサがそれぞれの機能を実現するプログラムを解釈し、実行することによりソフトウェアで実現してもよい。各機能を実現するプログラム、テーブル、ファイル等の情報は、メモリや、ハードディスク、ＳＳＤ（Ｓｏｌｉｄ　Ｓｔａｔｅ　Ｄｒｉｖｅ）等の記録装置、または、ＩＣカード、ＳＤカード、ＤＶＤ等の記録媒体に置くことができる。

[0116] また、制御線や情報線は説明上必要と考えられるものを示しており、製品上必ずしも全ての制御線や情報線を示しているとは限らない。実際には殆ど全ての構成が相互に接続されていると考えてもよい。

請求の範囲

[請求項1] 入力データと、重み係数を用いて演算を行うニューラルネットワークを有する演算装置であって、

前記入力データによる前記ニューラルネットワークのニューロンの発火状態を算出するネットワーク解析部と、

前記ニューロンの発火状態に基づいて、前記ニューラルネットワークの縮約率が設定された複数の縮約パターンから縮約パターンの候補を絞り込み、前記絞り込んだ縮約パターンの候補に基づいて前記ニューラルネットワークの縮約を実行して縮約後のニューラルネットワークを生成する縮約部と、

を有することを特徴とする演算装置。

[請求項2] 請求項１に記載の演算装置であって、

前記縮約部で生成された前記縮約後のニューラルネットワークに推論を実行して推論誤差を算出し、前記複数の縮約パターンの中から前記推論誤差に基づいて縮約パターンを抽出する最適化エンジン部をさらに有することを特徴とする演算装置。

[請求項3] 請求項２に記載の演算装置であって、

前記最適化エンジン部は、

前記推論誤差が最小となる縮約パターンを抽出することを特徴とする演算装置。

[請求項4] 請求項１に記載の演算装置であって、

前記縮約部で生成された前記縮約後のニューラルネットワークに対して、前記入力データによって再度学習を実施する再学習部を、さらに有することを特徴とする演算装置。

[請求項5] 請求項２に記載の演算装置であって、

前記縮約部で生成された前記縮約後のニューラルネットワークに対して、前記入力データによって再度学習を実施する再学習部を、さらに有し、

前記ネットワーク解析部と、前記縮約部および前記最適化エンジン部ならびに前記再学習部の演算途中における中間データを一時的に格納するメモリと、

前記ネットワーク解析部と、前記縮約部と、前記再学習部と、前記最適化エンジン部および前記メモリをスレーブとし、前記スレーブを制御するマスタとしてのスケジューラと、

前記マスタと前記スレーブを接続するインターコネクトと、をさらに有することを特徴とする演算装置。

[請求項6] 請求項１に記載の演算装置であって、

前記ネットワーク解析部は、

前記ニューラルネットワークと、前記縮約後のニューラルネットワークの適用先に応じた入力データとして受け付け、前記ニューラルネットワークの各ニューロンの発火状態を推定して数値化した特徴量を算出し、当該特徴量を前記適用先に特有の特徴を包含した解析結果として出力することを特徴とする演算装置。

[請求項7] 請求項６に記載の演算装置であって、

前記縮約部は、

前記ネットワーク解析部の解析結果を受け付け、前記解析結果において数値化された特徴量に基づいてニューラルネットワークの縮約を実行し、前記縮約後のニューラルネットワークおよび前記重み係数の最適解候補を複数出力することを特徴とする演算装置。

[請求項8] 請求項１に記載の演算装置であって、

前記縮約部は、

縮約手法が異なる複数の縮約実行部を有し、前記ニューラルネットワークの適用先に応じて前記縮約実行部を切り替えることを特徴とする演算装置。

[請求項9] 請求項７に記載の演算装置であって、

前記縮約部が出力する縮約後のニューラルネットワークに対して、

前記入力データによって再度学習を実施する再学習部を、さらに有し、

前記再学習部は、

前記ニューラルネットワークおよび前記重み係数の最適解候補を入力として受け付け、前記ニューラルネットワークおよび前記重み係数を初期値として、再び学習を行うことで、再学習済みのニューラルネットワーク及び再学習済みの重み係数を出力することを特徴とする演算装置。

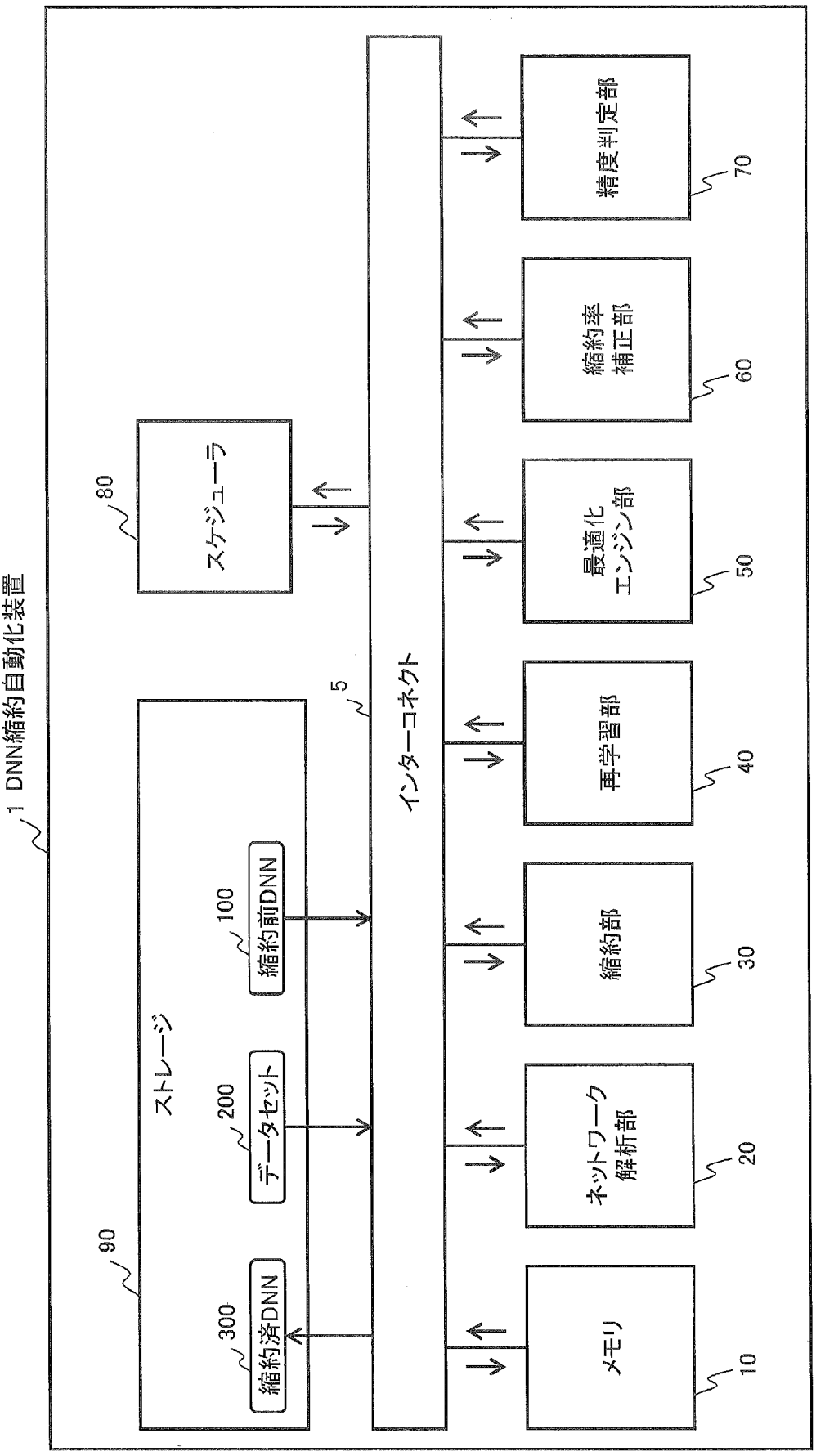
[請求項10] 請求項９に記載の演算装置であって、

前記縮約部で縮約を実行した縮約後のニューラルネットワークに推論を実行して推論誤差を算出し、前記複数の縮約パターンの中から前記推論誤差に基づいて縮約パターンを抽出する最適化エンジン部を、さらに有し、

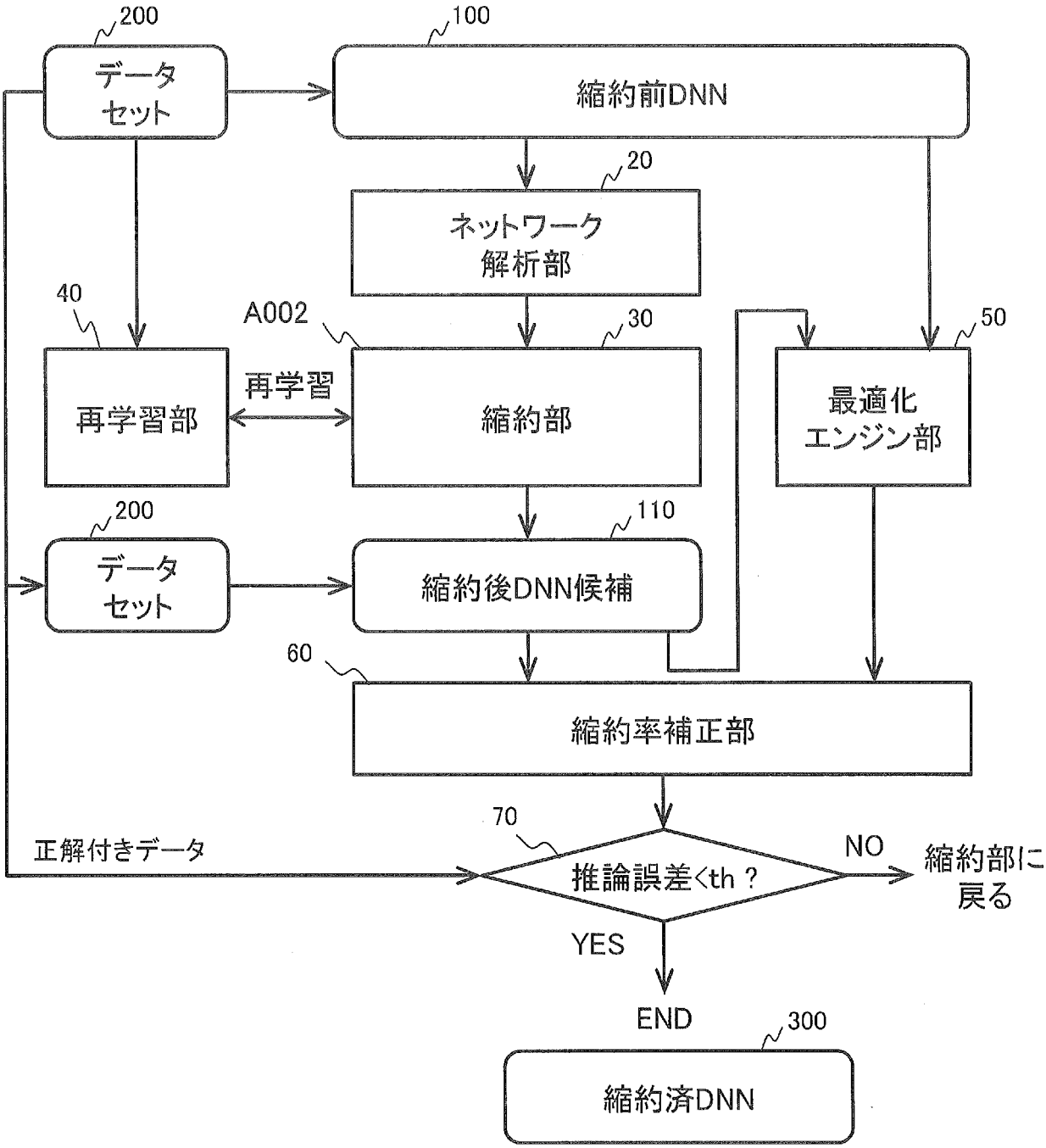
前記最適化エンジン部は、

複数の前記ニューラルネットワークおよび前記再学習済みの重み係数を入力として受け付けて、予め設定された確率論的探索を用いて、前記縮約パターンを算出することを特徴とする演算装置。

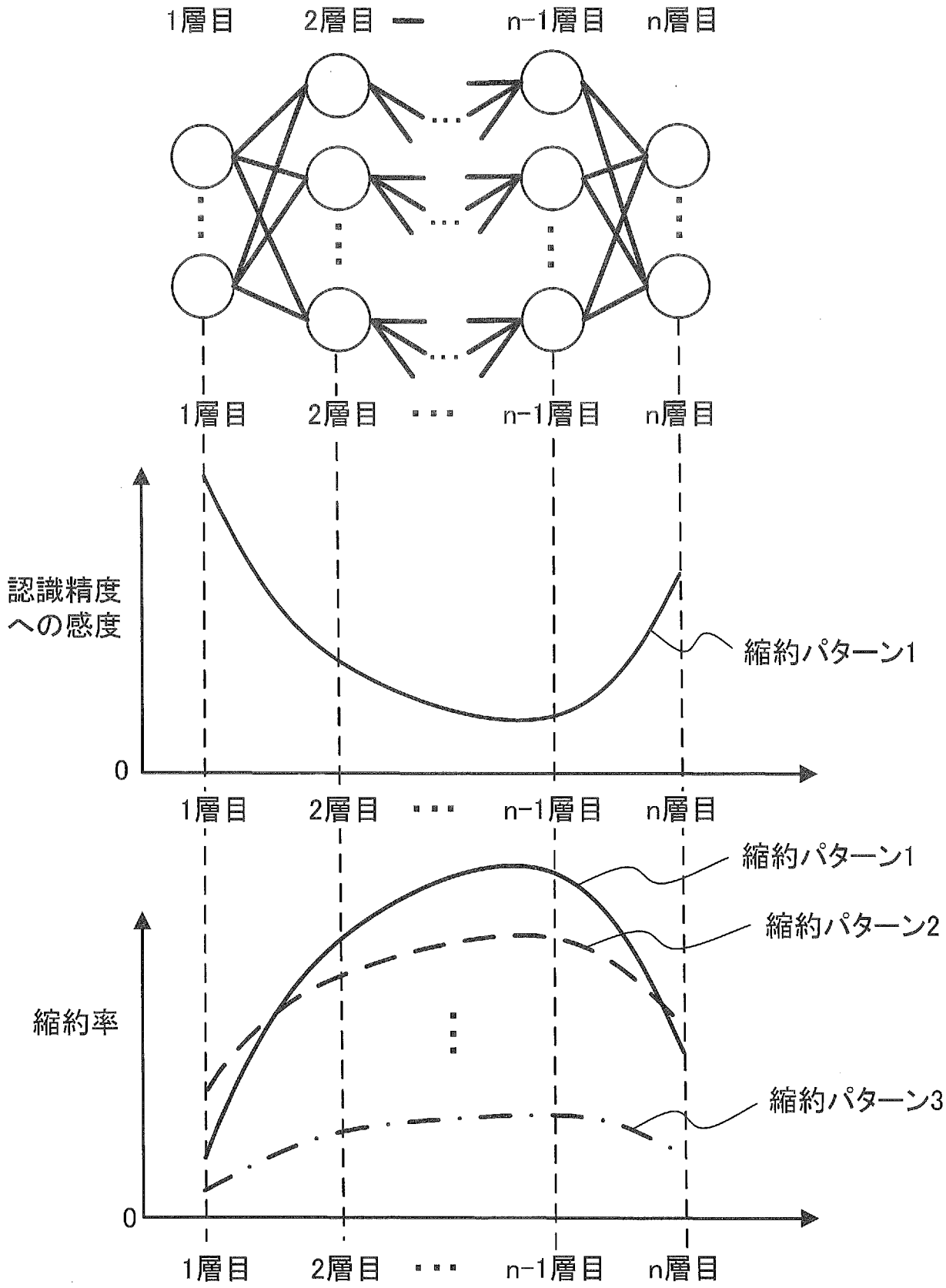
[図1]


1 DNN縮約自動化装置
90
ストレージ
300
縮約済DNN
200
データセット
100
縮約前DNN
5
インターコネクト
80
スケジューラ
10
メモリ
20
ネットワーク解析部
30
縮約部
40
再学習部
50
最適化エンジン部
60
縮約率補正部
70
精度判定部

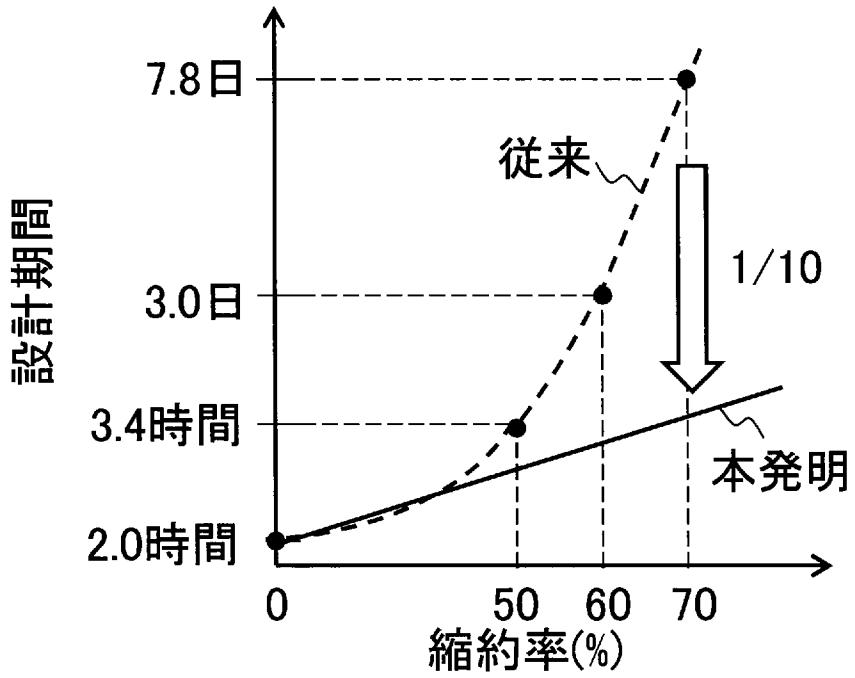
[図2]


200
データセット
100
縮約前DNN
20
ネットワーク解析部
40
A002
30
50
再学習部
再学習
縮約部
最適化エンジン部
200
データセット
110
縮約後DNN候補
60
縮約率補正部
70
正解付きデータ
推論誤差<th？
NO
縮約部に戻る
YES
END
300
縮約済DNN

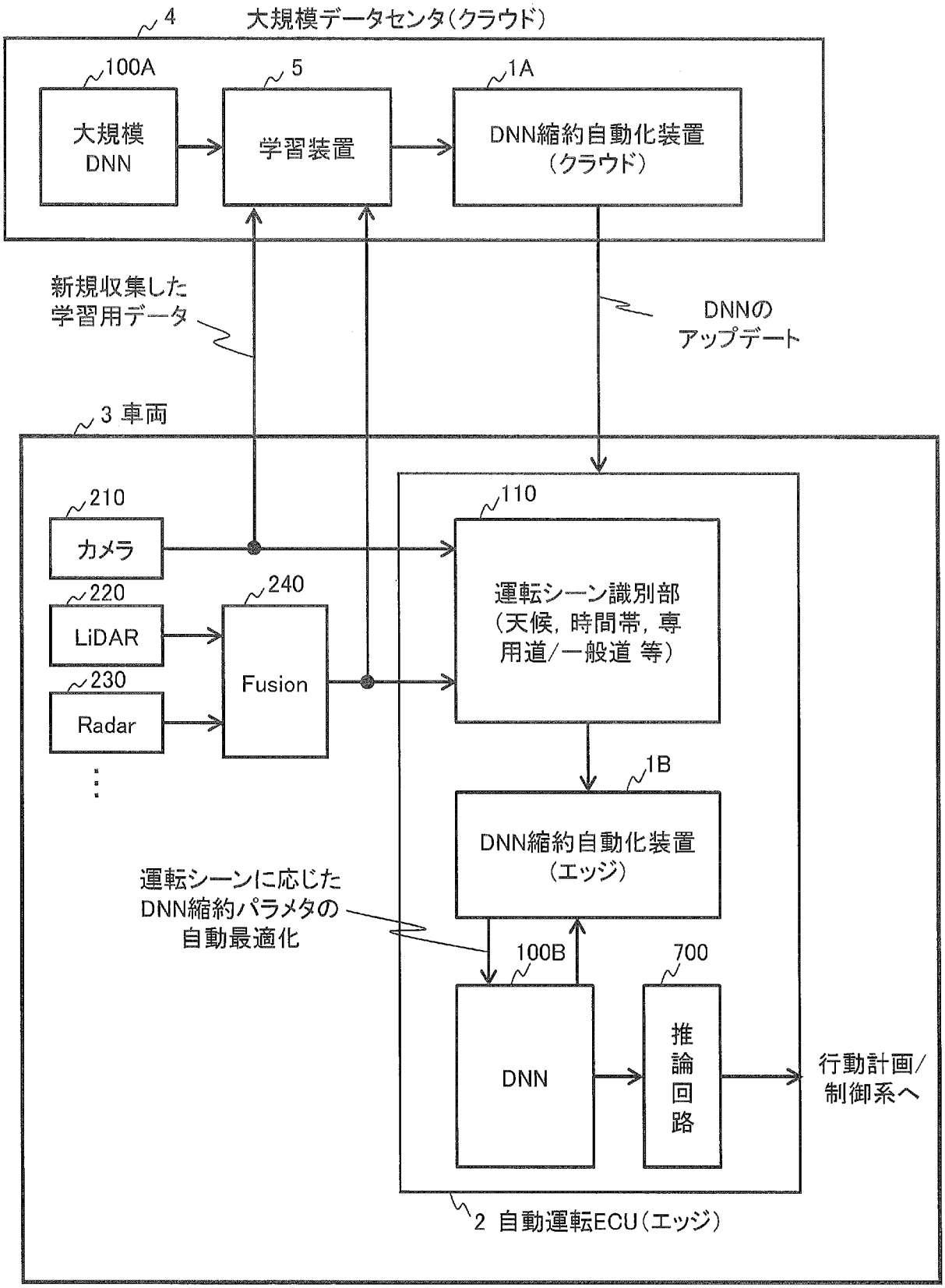
[図3]


1層目
2層目
n-1層目
n層目
1層目
2層目
n-1層目
n層目
認識精度
への感度
0
縮約パターン1
1層目
2層目
n-1層目
n層目
縮約パターン1
縮約パターン2
縮約率
縮約パターン3
0
1層目
2層目
n-1層目
n層目

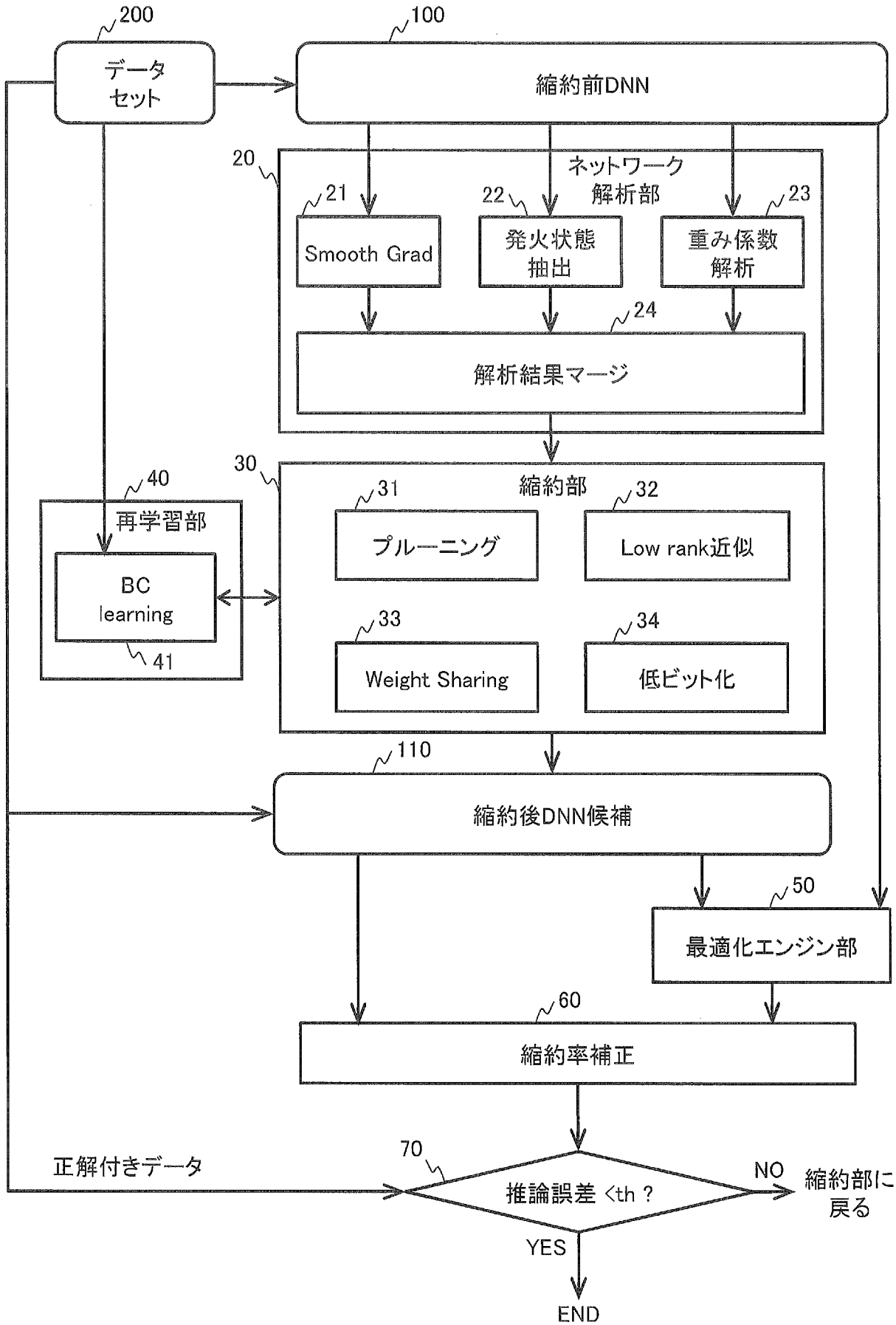
[図4]



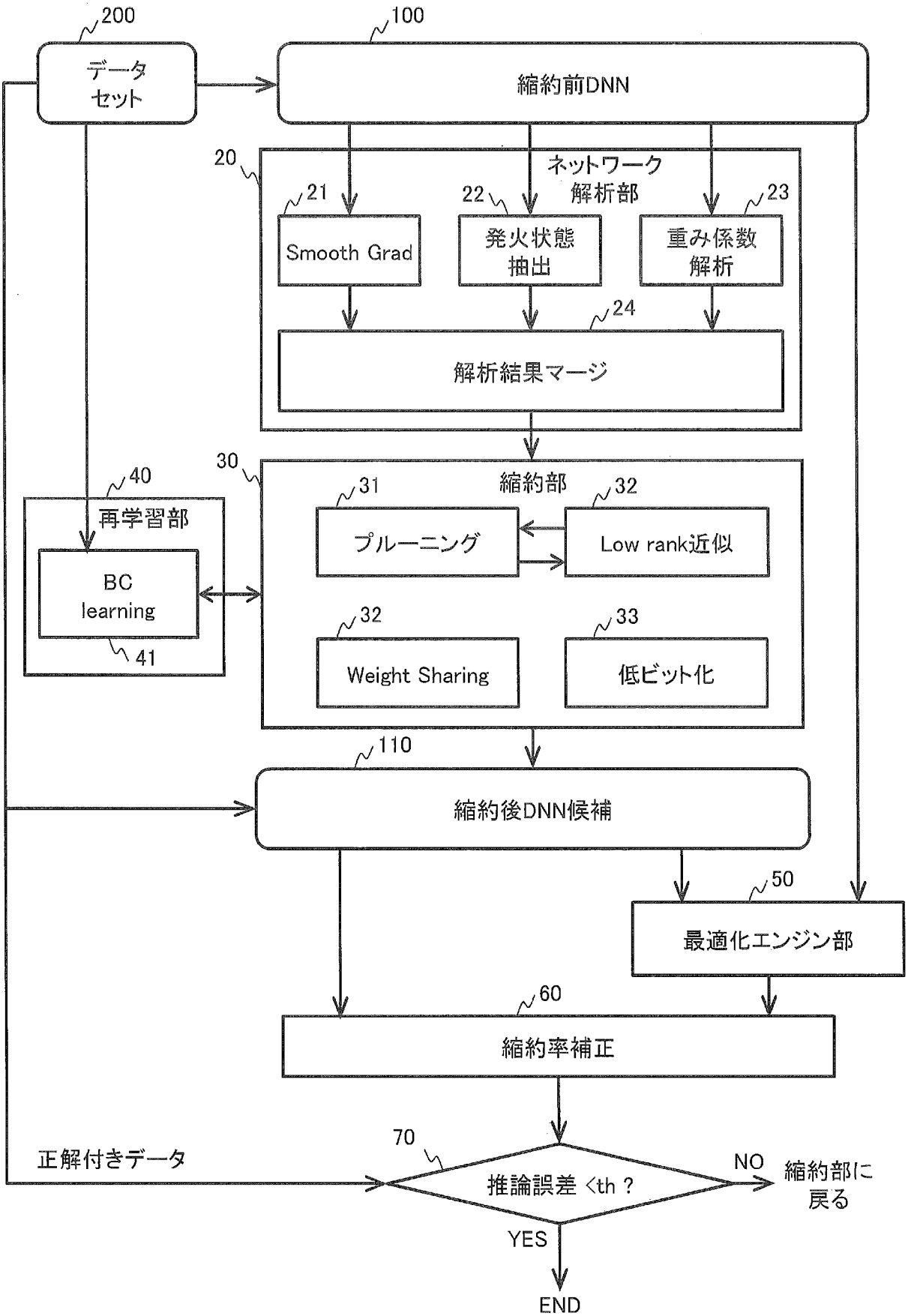
[図5]


4
大規模データセンタ(クラウド)
100A
大規模
DNN
5
学習装置
1A
DNN縮約自動化装置
(クラウド)
新規収集した
学習用データ
DNNの
アップデート
3 車両
210
カメラ
220
LiDAR
230
Radar
240
Fusion
110
運転シーン識別部
(天候, 時間帯, 専
用道/一般道 等)
1B
DNN縮約自動化装置
(エッジ)
運転シーンに応じた
DNN縮約パラメタの
自動最適化
100B
DNN
700
推
論
回
路
行動計画/
制御系へ
2 自動運転ECU(エッジ)

[図6]


200
データセット
100
縮約前DNN
20
ネットワーク解析部
21
Smooth Grad
22
発火状態抽出
23
重み係数解析
24
解析結果マージ
30
縮約部
31
プルーニング
32
Low rank近似
33
Weight Sharing
34
低ビット化
40
再学習部
BC learning
41
110
縮約後DNN候補
50
最適化エンジン部
60
縮約率補正
70
推論誤差 <th ?
NO
縮約部に戻る
YES
END
正解付きデータ

[図7]


200
データセット
100
縮約前DNN
20
ネットワーク解析部
21
Smooth Grad
22
発火状態抽出
23
重み係数解析
24
解析結果マージ
30
縮約部
31
プルーニング
32
Low rank近似
32
Weight Sharing
33
低ビット化
40
再学習部
BC learning
41
110
縮約後DNN候補
50
最適化エンジン部
60
縮約率補正
70
推論誤差 <th ?
NO
縮約部に戻る
YES
END
正解付きデータ

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2019/040272

A. CLASSIFICATION OF SUBJECT MATTER
Int. Cl. G06N3/08(2006.01)i, G06N3/04(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
Int. Cl. G06N3/08, G06N3/04

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Published examined utility model applications of Japan 1922-1996
Published unexamined utility model applications of Japan 1971-2019
Registered utility model specifications of Japan 1996-2019
Published registered utility model applications of Japan 1994-2019

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
JSTPlus/JMEDPlus/JST7580 (JDreamIII)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 2005-346472 A (CANON INC.) 15 December 2005, paragraphs [0055]-[0070], fig. 6, 7 & US 2007/0244842 A1, paragraphs [0084]-[0100], fig. 6, 7	1-10
A	村田大智外5名, "自己学習型超解像に適用するK-meansクラスタリング処理のハードウェアによる実現", FIT2015第14回情報科学技術フォーラム講演論文集, 24 August 2015, Part I, pp. 253-254 (MURATA, Daichi et al. A Hardware Architecture to Perform K-means Clustering for Learning-Based Super-Resolution Combining Self-Learning and Prior-Learning Dictionaries. Forum on Information Technology 2015.)	1-10
P, X	JP 2019-159693 A (KDDI CORP.) 19 September 2019, abstract, paragraphs [0015]-[0041], fig. 1-4,	1-2, 8

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
13.12.2019	24.12.2019
Name and mailing address of the ISA/ Japan Patent Office 3-4-3, Kasumigaseki, Chiyoda-ku, Tokyo 100-8915, Japan	Authorized officer Telephone No.

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ２０１９／０４０２７２

Ａ．　発明の属する分野の分類（国際特許分類（ＩＰＣ））

Int.Cl.　G06N3/08(2006.01)i, G06N3/04(2006.01)i

Ｂ．　調査を行った分野

調査を行った最小限資料（国際特許分類（ＩＰＣ））

Int.Cl.　G06N3/08, G06N3/04

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	１９２２－１９９６年
日本国公開実用新案公報	１９７１－２０１９年
日本国実用新案登録公報	１９９６－２０１９年
日本国登録実用新案公報	１９９４－２０１９年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

JSTPlus/JMEDPlus/JST7580（JDreamIII）

Ｃ．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
A	JP 2005-346472 A（キヤノン株式会社） 2005.12.15, 段落［００５５］－［００７０］及び図６－７ & US 2007/0244842 A1, 段落[0084]-[0100]及び図 6-7	1-10
A	村田　大智　外５名，“自己学習型超解像に適用するＫ－ｍｅａｎｓクラスタリング処理のハードウェアによる実現”，ＦＩＴ２０１５　第１４回情報科学技術フォーラム　講演論文集，2015.08.24，第１分冊，pp. 253-254	1-10

☑　Ｃ欄の続きにも文献が列挙されている。

☐　パテントファミリーに関する別紙を参照。

＊　引用文献のカテゴリー

「Ａ」特に関連のある文献ではなく、一般的技術水準を示すもの

「Ｅ」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの

「Ｌ」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）

「Ｏ」口頭による開示、使用、展示等に言及する文献

「Ｐ」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献

「Ｔ」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの

「Ｘ」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの

「Ｙ」特に関連のある文献であって、当該文献と他の１以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの

「＆」同一パテントファミリー文献

国際調査を完了した日	国際調査報告の発送日
１３．１２．２０１９	２４．１２．２０１９

国際調査機関の名称及びあて先
日本国特許庁（ＩＳＡ／ＪＰ）
郵便番号１００－８９１５
東京都千代田区霞が関三丁目４番３号

特許庁審査官（権限のある職員）　５Ｂ　２９４５
桜井　茂行
電話番号　０３－３５８１－１１０１　内線　３５４５

Ｃ（続き）．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
P, X	JP 2019-159693 A（ＫＤＤＩ株式会社） 2019.09.19, 要約, 段落［００１５］－［００４１］及び図１－４, 段落［００４６］－［００４７］及び図６ (ファミリーなし)	1-2, 8