

(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10月22日(22.10.2020)



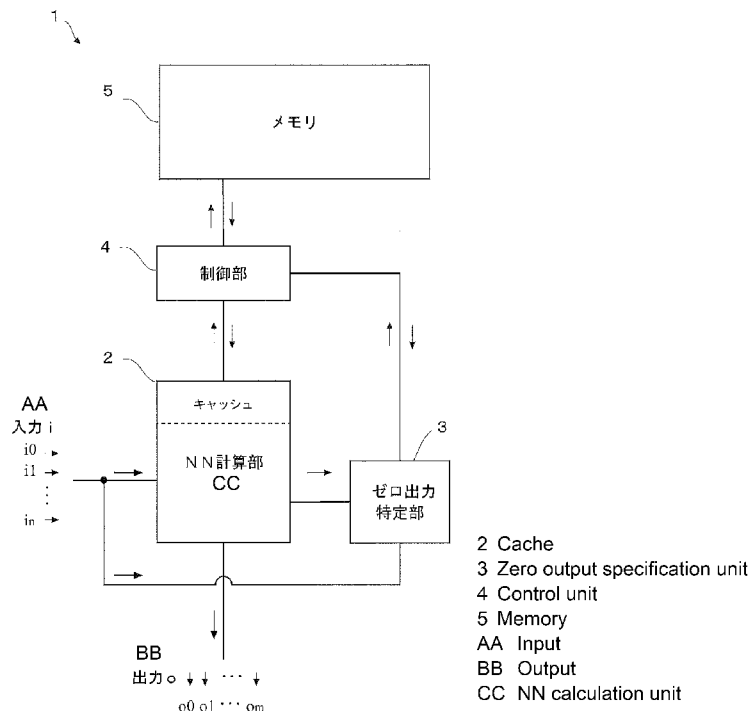
(10) 国際公開番号

WO 2020/213670 A1

- (51) 国際特許分類:
G06N 3/02 (2006.01) *G06N 3/08* (2006.01) 北海道札幌市北区北8条西5丁目 Hokkaido (JP).
- (21) 国際出願番号: PCT/JP2020/016684 (72) 発明者: 高前田 伸也 (TAKAMAEDA Shinya); 〒0600808 北海道札幌市北区北8条西5丁目 国立大学法人北海道大学内 Hokkaido (JP).
- (22) 国際出願日: 2020年4月16日(16.04.2020)
- (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 2019-080194 2019年4月19日(19.04.2019) JP (74) 代理人: 特許業務法人 インテクト国際特許事務所, 外 (INTECT INTERNATIONAL PATENT OFFICE et al.); 〒1020083 東京都千代田区麹町四丁目7番2号 Daiwa 麹町4丁目ビル4階 Tokyo (JP).
- (71) 出願人: 国立大学法人北海道大学 (NATIONAL UNIVERSITY CORPORATION HOKKAIDO UNIVERSITY) [JP/JP]; 〒0600808

(54) Title: NEURAL CALCULATION DEVICE AND NEURAL CALCULATION METHOD

(54) 発明の名称: ニューラル計算装置、および、ニューラル計算方法



(57) Abstract: The present invention provides a neural calculation device 1, etc., with which it is possible to dynamically omit calculations relating to a neuron. The present invention comprises: a neural net calculation unit 2 composed of a plurality of neuron calculators 21a that add up the results obtained by multiplying input data *i* and a weighting coefficient *w* and applying an activation function to the addition result to output output data *o*; a zero output specification unit 3 for performing a calculation to specify a neuron calculator 21a for which the output is zero depending on the input data *i*; and

(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, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, 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, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, 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))

a control unit 4 for controlling the calculation of the neural net calculation unit 2 on the basis of the output of the zero output specification unit 3 so as to omit the calculation of the neuron calculator 21a for which the output is zero.

(57) 要約: ニューロンに関する計算を動的に省略できるニューラル計算装置1等を提供する。入力データ*i*と重み付け係数*w*との乗算結果を加算し、当該加算結果に対して、活性化関数を適用して出力データ*o*を出力する複数のニューロン計算器21aから構成されるニューラルネット計算部2と、入力データ*i*に応じて出力がゼロになるニューロン計算器21aを特定する計算を行うゼロ出力特定部3と、ゼロ出力特定部3の出力に基づき、出力がゼロになるニューロン計算器21aの計算を省略するようにニューラルネット計算部2の計算を制御する制御部4と、を備える。

明 細 書

発明の名称：ニューラル計算装置、および、ニューラル計算方法
技術分野

[0001] 本発明は、ニューラル計算装置、および、ニューラル計算方法の技術分野に属する。

背景技術

[0002] 近年、ＩｏＴ（Internet of Things）・自動運転等の用途で、深層ニューラルネットワーク（ＤＮＮ）を高速・低消費電力に処理するコンピュータ・ハードウェアが求められている。ニューラルネットワークの計算の大部分は、入力データに対する重み付け係数と活性値の乗算と、乗算結果の積算である。そのため、最終的な認識結果に影響を大きく与えない重みを省略し、乗算と加算の回数を削減する手法として枝刈り(Pruning)と呼ばれる技術（例えば、非特許文献１）や、同様の技術に認識結果に影響を大きく与えないニューロンを取り除くニューロン刈り(Neuron Pruning)と呼ばれる技術（例えば、非特許文献２）が存在する。

先行技術文献

非特許文献

[0003] 非特許文献１：「Learning both Weights and Connections for Efficient Neural Networks」論文、Song Han他、arXiv:1506.02626v3 [cs.NE]、2015年10月30日

非特許文献２：「A Threshold Neuron Pruning for a Binarized Deep Neural Network on an FPGA」論文、Tomoya Fujii他、IEICE Transactions on Information and Systems, 2018 年 E101.D 巻 2 号 p. 376-386 (URL : https://www.jstage.jst.go.jp/article/transinf/E101.D/2/E101.D_2017RCP0013/_article/-char/ja/)

発明の概要

発明が解決しようとする課題

[0004] しかしながら、従来技術において、ニューラルネットワークの学習時に静的に重み付け係数やニューロンを削るため、入力パターンに応じて動的に発現する、無効な重みやニューロンに関する計算を省略することはできなかった。

[0005] そこで本発明は、上記の各問題点及び要請等に鑑みて為されたもので、その課題の一例は、ニューロンに関する計算を動的に省略できるニューラル計算装置を提供することにある。

課題を解決するための手段

[0006] 上記の課題を解決するために、請求項 1 に記載の発明は、入力データと重み付け係数との乗算結果を加算し、当該加算結果に対して、活性化関数を適用して出力データを出力する複数のニューロン計算器から構成されるニューラルネット計算手段と、前記入力データに応じて、出力がゼロになる前記ニューロン計算器を特定するための計算を行うゼロ出力特定手段と、前記ゼロ出力特定手段の計算結果に基づき、前記出力がゼロになるニューロン計算器の計算を省略するように前記ニューラルネット計算手段の計算を制御する制御手段と、を備えたことを特徴とする。

[0007] 請求項 2 に記載の発明は、請求項 1 に記載のニューラル計算装置において、前記ゼロ出力特定手段が、前記ニューラルネット計算手段の計算を模した 2 値化ニューラルネット計算器であり、前記制御手段が、前記入力データから変換して得られる 1 ビットのデータを入力した前記ゼロ出力特定手段の出力に基づき、前記ニューラルネット計算手段の計算を制御することを特徴とする。

[0008] 請求項 3 に記載の発明は、請求項 2 に記載のニューラル計算装置において、前記ゼロ出力特定手段が、前記ニューラルネット計算手段において前記活性化関数を適用する前までの計算を模した 2 値化ニューラル計算器であることを特徴とする。

[0009] 請求項 4 に記載の発明は、請求項 3 に記載のニューラル計算装置において、前記ゼロ出力特定手段の出力と前記ニューラルネット計算手段における前

記加算結果との差異に基づき、前記 2 値化ニューラル計算器を学習させる学習手段を更に備えたことを特徴とする。

[0010] 請求項 5 に記載の発明は、請求項 1 から請求項 4 のいずれか 1 項に記載のニューラル計算装置において、前記ニューラルネット計算手段が、複数のニューラルレイヤーから構成され、各前記ニューラルレイヤーに対応する複数の前記ゼロ出力特定手段があり、前記ニューラルレイヤーに対応するゼロ出力特定手段が、前の層のニューラルレイヤーの出力データに応じて、当該ニューラルレイヤーに属する前記ニューロン計算器の出力がゼロになるニューロン計算器を特定する計算を行うことを特徴とする。

[0011] 請求項 6 に記載の発明は、請求項 5 に記載のニューラル計算装置において、前記制御手段が、前記ニューラルレイヤーに対応するゼロ出力特定手段により出力がゼロで無いと特定された前記ニューロン計算器から計算を先行して開始するように、前記ニューラルネット計算手段の計算を制御することを特徴とする。

[0012] 請求項 7 に記載の発明は、請求項 5 または請求項 6 に記載のニューラル計算装置において、前記制御手段が、前記ニューラルレイヤーにおける計算と次の層の前記ニューラルレイヤーにおける計算との間に時間間隔を設けて消費電力を減少させるように、前記ニューラルネット計算手段の計算を制御することを特徴とする。

[0013] 請求項 8 に記載の発明は、請求項 1 から請求項 7 のいずれか 1 項に記載のニューラル計算装置において、前記活性化関数が正規化線形関数であることを特徴とする。

[0014] 請求項 9 に記載の発明は、請求項 8 に記載のニューラル計算装置において、前記ゼロ出力特定手段が、前記入力データのパターンから、前記加算結果の符号がゼロ以下になる前記ニューロン計算器を特定する計算を行うことを特徴とする。

[0015] 請求項 10 に記載の発明は、入力データと重み付け係数との乗算結果を加算し、当該加算結果に対して、活性化関数を適用して出力データを出力する

複数のニューロン計算器から構成されるニューラルネット計算装置において、ゼロ出力特定手段が、前記入力データに応じて、出力がゼロになる前記ニューロン計算器を特定するための計算を行う出力特定ステップと、制御手段が、前記ゼロ出力特定手段の計算結果に基づき、前記出力がゼロになるニューロン計算器の計算を省略するように前記ニューラルネット計算装置の計算を制御する制御ステップと、を含むことを特徴とする。

発明の効果

[0016] 本発明によれば、入力データと重み付け係数との乗算結果を加算し、当該加算結果に対して、活性化関数を適用して出力データを出力する複数のニューロン計算器から構成されるニューラルネット計算において、入力データに応じて出力がゼロになるニューロン計算器を特定するための計算を行い、この計算結果に基づき、出力がゼロになるニューロン計算器の計算を省略するようにニューラルネット計算を制御することにより、入力データのパターンに応じて、出力がゼロになるニューロン計算器を特定できるため、ニューロンに関する計算を動的に省略できる。

図面の簡単な説明

[0017] [図1]実施形態に係るニューラル計算装置の一例を示すブロック図である。

[図2]図1のニューラルネット計算部およびゼロ出力特定部の概要構成例を示すブロック図である。

[図3]図2のニューロン計算器の一例を示すブロック図である。

[図4]ニューロン計算器の計算を省略の一例を示す模式図である。

[図5]2値化ニューラルネットワーク・システムの概要構成例を示すブロック図である。

[図6]図5のニューラル電子回路の一例を示すブロック図である。

[図7]ニューラル計算装置の動作例を示すフローチャートである。

[図8]ニューラルネット計算部およびゼロ出力特定部の一例を示すブロック図である。

[図9]ニューラルネット計算部およびゼロ出力特定部の動作のタイミングの一

例を示す模式図である。

[図10]ニューラルネット計算部およびゼロ出力特定部の動作のタイミングの一例を示す模式図である。

[図11]ニューラル計算装置による活性予測精度の結果の一例を示す模式図である。

[図12]ニューラル計算装置による計算削減率の結果の一例を示す模式図である。

[図13]ゼロ出力特定部の変形例を示すブロック図である。

[図14]ゼロ出力特定部がサポートベクターマシンの識別関数等を説明する模式図である。

[図15]ニューラルネット計算部およびゼロ出力特定部の変形例を示すブロック図である。

発明を実施するための形態

[0018] 以下、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、以下に説明する実施の形態は、ニューラル計算装置に対して本発明を適用した場合の実施形態である。

[0019] (1. ニューラル計算装置の構成および機能概要)

まず、本発明の一実施形態に係るニューラル計算装置の構成および概要機能について、図1を用いて説明する。

[0020] 図1は、実施形態に係るニューラル計算装置の一例を示すブロック図である。図2は、ニューラルネット計算部およびゼロ出力特定部の概要構成例を示すブロック図である。図3は、ニューロン計算器の一例を示すブロック図である。図4は、ニューロン計算器の計算を省略の一例を示す模式図である。

[0021] 図1に示すように、ニューラル計算装置1は、入力データ i と重み付け係数との乗算結果を加算し、加算結果に対して、活性化関数を適用して出力データ o を出力する複数のニューロン計算器から構成されるニューラルネット計算部(NN計算部)2と、ニューラルネット計算部2のニューロン計算器

の出力がゼロになるニューロン計算器を特定する計算を行うゼロ出力特定部 3 と、ゼロ出力特定部 3 に基づき、出力がゼロになるニューロン計算器の計算を省略するようにニューラルネット計算部 2 の計算を制御する制御部 4 と、重み付け係数 w および計算の中間データ等を記憶するメモリ 5 と、を備えている。

[0022] ニューラルネット計算部 2 は、入力データと重み付け係数との乗算結果を加算し、当該加算結果に対して、活性化関数を適用して出力データを出力する複数のニューロン計算器から構成されるニューラルネット計算手段の一例である。ゼロ出力特定部 3 は、前記入力データに応じて、出力がゼロになる前記ニューロン計算器を特定するための計算を行うゼロ出力特定手段の一例である。制御部 4 は、前記ゼロ出力特定手段の計算結果に基づき、前記出力がゼロになるニューロン計算器の計算を省略するように前記ニューラルネット計算手段の計算を制御する制御手段の一例である。

[0023] ニューラル計算装置 1 は、入力データ i_0 、入力データ i_1 、入力データ i_2 、…、入力データ i_n (n は自然数。以下同様。) に対して、出力データ o_0 、出力データ o_1 、出力データ o_2 、…、入出力データ o_m (m は自然数。以下同様。) を出力する。

[0024] 図 2 に示すように、ニューラルネット計算部 2 は、複数のニューロン計算器 20a を有する多層のニューラルレイヤー 20 から構成されている。

[0025] 図 3 に示すように、ニューロン計算器 20a は、入力データ i_0 、入力データ i_1 、入力データ i_2 、…、入力データ i_n と、入力データ i のそれぞれに対する重み付け係数 w_0 、重み付け係数 w_1 、重み付け係数 w_2 、…、重み付け係数 w_n の乗算結果を加算する乗算加算器 21a と、加算結果に対して活性化関数を適用する活性化関数器 25a と、を有する。活性化関数器 25a の活性化関数が正規化線形関数である場合、乗算加算器 21a の出力がゼロ以下の場合、ゼロを出力し、乗算加算器 21a の出力が正の値 x の場合、値 x を出力する。

[0026] なお、図 2 において、ニューラルネットの各層が出力側のニューロンと入

力側のニューロンとが、隣り合ったニューラルレイヤー20の間で分離して模式的に描かれている。1番目のニューラルレイヤー20内では、入力層L0のニューロンに対して、第1層L1のニューロン計算器20a群が構成されている。2番目のニューラルレイヤー20内では、第1層L1のニューロンに対して、第2層L2のニューロン計算器20a群が構成されている。n番目のニューラルレイヤー20内では、第n-1層L n-1のニューロンに対して、第n層Lnのニューロン計算器20a群が構成されている。

[0027] 図2に示すように、ゼロ出力特定部3は、複数の活性予測器30から構成されている。各活性予測器30は、各ニューラルレイヤー20に対応して設置される。

[0028] 活性予測器30には、対応するニューラルレイヤー20と同じ入力データが入力される。第2層以降のニューラルレイヤー20に対応する活性予測器30には、1つ前の層のニューラルレイヤー20の出力が、入力される。

[0029] 活性予測器30は、対応するニューラルレイヤー20と同じ入力データのパターンに基づき、ニューラルレイヤー20の各ニューロン計算器20aにおける計算のうち活性化関数を適用する前までの計算を模した計算を行う。活性予測器30における計算結果は、制御部4に出力される。

[0030] この計算結果でニューロン計算器20aの出力がゼロ以下になると予測されるニューロン計算器20aに対して、制御部4は、出力がゼロになるニューロン計算器20aに関わる乗算や加算の計算を省略するようにニューラルレイヤー20の計算を制御する。

[0031] 例えば、図4に示すように、ニューラルネット計算部2は、計算を省略しないニューロン計算器20aのネットワークで連結されたニューロン計算器20aのみの計算を行う。

[0032] 活性予測器30は、ニューラルレイヤー20より速く計算できる計算器ならばよい。例えば、活性予測器30の一例として、2値化ニューラル計算器、整数化ニューラル計算器、サポートベクターマシン（SVM）、ランダムフォレスト等が挙げられる。活性予測器30が2値化ニューラル計算器の場合、

活性予測器 30 は、入力側で、例えば、入力データの最上ビットを取り出し、2 値化ニューラルネットワークの計算がされる。第 1 層目では、最上ビットは、入力データの符号（例えば、0 がマイナス符号を示し、1 がプラス符号）を表し、第 2 層目以降では、最上ビットは、データの値のマグニチュード（例えば、0 ならば 0 を示し、1 ならばそのビット数で示せる最大値を示す）を表している。なお、2 値化ニューラルネットワークの計算に使用される入力は、入力データの最上ビットの他に、入力データから変換して得られる 1 ビットのデータならばよい。

[0033] ニューラルネット計算部 2 およびゼロ出力特定部 3 の計算は、専用の電子回路のハードウェアで計算されてもよいし、ノイマン型等の CPU による計算でもよいし、GPU 等の AI チップによる計算でもよいし、CPU と GPU とを組み合わせた計算でもよし、ニューロモーフィックチップによる計算でもよい。

[0034] ゼロ出力特定部 3 における 2 値化ニューラルネットワークの計算は、例えば、1 ビットの重み付け係数と 1 ビットの入力データとの排他的論理和（XNOR）によって 2 値化乗算を行うニューラル電子回路で行われてもよい。2 値化ニューラルネットワークの計算は、ニューロン間の 1 ビットの接続有無情報に基づき、2 値化乗算の乗算結果に加算をする。2 値化ニューラルネットワークの計算は、加算結果に活性化関数を適用して 1 ビットの入力データを出力する。

[0035] （2. ゼロ出力特定部 3 における 2 値化ニューラルネットワーク・システムの構成および機能）

次に、ゼロ出力特定部 3 が 2 値化ニューラルネットワークの場合における 2 値化ニューラルネットワーク・システムの構成および概要機能について、図を用いて説明する。

[0036] 図 5 は、2 値化ニューラルネットワーク・システムの概要構成例を示すブロック図である。図 6 は、図 5 のニューラル電子回路の一例を示すブロック図である。

[0037] 図5に示すように、2値化ニューラルネットワーク・システムNNSは、様々なタイプのニューラルネットワークを電子回路で実現可能な複数のコア電子回路Coreと、コア電子回路Core同士を接続するシステムバスbusと、を備えている。

[0038] コア電子回路Coreは、様々なタイプのニューラルネットワークを電子回路で実現可能な2値化ニューラル電子回路NNと、2値化ニューラル電子回路NNの重み付け係数等を設定するメモリアクセス制御部MCntと、ニューラル電子回路NNおよびメモリアクセス制御部MCntを制御する制御部Cntと、を有する。ここで、様々なタイプのニューラルネットワークの一例として、ニューロン層間のニューロン同士がフル結合したフル結合のタイプのニューラルネットワーク、コンボリユーション演算をするニューラルネットワーク、ニューロン層における層内拡張をしたニューラルネットワーク、層数を拡張するニューラルネットワーク等が挙げられる。

[0039] 2値化ニューラル電子回路NNは、入力データI1、・・・、Ip（pは自然数。以下同様）を並列で順次供給する入力メモリアレイ部MAiと、重み付け係数のデータを並列で順次供給するメモリセル部MCと、供給された入力データI1、・・・、Ipと重み付け係数とを乗算する乗算機能を実現して乗算結果を出力する複数のプロセスエレメント部Peと、並列の各入力データのプロセスエレメント部Peからの各乗算結果を加算して加算結果に活性化関数を適用する加算活性化部Actと、各加算活性化部Actからの1ビットの出力データO1、・・・、Oq（qは自然数。以下同様）をそれぞれ順次記憶する出力メモリアレイ部MAoと、各加算活性化部Actにバイアス用の値を順次提供するバイアス用メモリアレイ部MAbとを有する。

[0040] メモリアクセス制御部MCntは、例えば、Direct Memory Access Controllerである。メモリアクセス制御部MCntは、制御部Cntの制御に従い、各プロセスエレメント部Peに逐次供給する入力データを、入力メモリアレイ部MAiに設定する。また、メモリアクセス制御部MCntは、制御部Cntの制御に従い、重み付け係数およびニューロン間の接続の有無を示す

所定値を、各メモリセル部MCに予め設定する。また、メモリアクセス制御部MC_{nt}は、制御部C_{nt}の制御に従い、加算活性化部A_{ct}から出力された出力データを、出力メモリアレイ部MA_oから取り出す。

[0041] 制御部C_{nt}は、CPU (Central Processing Unit) 等を有する。制御部C_{nt}は、2値化ニューラル電子回路NNの各素子の同期等のタイミングを計ったり、計算やデータの転送の同期を取ったりする。また、制御部C_{nt}は、2値化ニューラル電子回路NN内の後述のセクタ素子の切り替え制御を行う。

[0042] 制御部C_{nt}は、メモリアクセス制御部MC_{nt}を制御して、他のコア電子回路C_{ore}から出力されたデータを入力メモリアレイ部MA_i用に整えて、入力データとして入力メモリアレイ部MA_iに供給するように制御する。制御部C_{nt}は、メモリアクセス制御部MC_{nt}が、出力メモリアレイ部MA_oから取得した出力データを、他のコア電子回路C_{ore}に転送するように制御する。

[0043] なお、上位コントローラ（例えば、制御部4）が、ニューラルネットワーク・システムNNSや、各コア電子回路C_{ore}の制御部C_{nt}を制御してもよい。また、上位コントローラが、制御部C_{nt}の代わりに、2値化ニューラル電子回路NNおよびメモリアクセス制御部MC_{nt}を制御してもよい。上位コントローラは、外付けのコンピュータでもよい。

[0044] バイアス用メモリアレイ部MA_bは、各加算活性化部A_{ct}に提供するバイアス用のデータを予め記憶している。

[0045] 図6に示すように、2値化ニューラル電子回路NNは、例えば、入力p個×出力q個の2層のニューラルネットワークを実現する。

[0046] メモリセル部MCは、重み付け係数を記憶するメモリセル10を有する。メモリセル10は、構築するニューラルネットワークによる実現する脳機能に基づいて予め設定された「1」または「0」の1ビットの重み付け係数を記憶している。

[0047] なお、メモリセル部MCは、上記脳機能に基づいて予め設定されたニュー

ロン間の接続有無情報を記憶している別の接続有無情報用のメモリセル（図示せず）も有していてもよい。ここで、接続無情報は、例えば、NC（Not Connected（接続なし））を意味する1ビット所定値であり、所定値として「1」または「0」等が割り当てられる。

[0048] メモリセル10が並んで、メモリセルの列が形成される。同時に各プロセスエレメント部Peに出力されるメモリセル10をまとめてメモリセルブロックCBが形成される。メモリセルブロックCBのメモリセル10は、並列で入力される各入力データに対応している。

[0049] メモリセル部MCは、入力メモリアレイ部MAiから並列で入力される入力データI1、・・・、Ipの入力並列数p個以上のメモリセルブロックCBを有することが好ましい。メモリセルブロックCBにおいて、メモリセル10の数は、入力メモリアレイ部MAiから1ビット順次入力されるシリアルの入力データのサイクル数以上が好ましい。

[0050] メモリセル部MCは、メモリセルブロックCB毎に、1ビットの重み付け係数を、1ビット順次入力されるシリアルの入力データに対応してプロセスエレメント部Peに、順次出力する。各プロセスエレメント部Peに、メモリセルブロックCBからの重み付け係数と、入力メモリアレイ部MAiからの入力データとが入力される。

[0051] メモリセルブロックCBは、1ビットの重み付け係数と1ビットの接続有無情報とを、交互に、プロセスエレメント部Peに順次、出力してもよい。メモリセル10は、プロセスエレメント部Peに対して独立した結線を有し、別々に順次、プロセスエレメント部Peに出力してもよい。

[0052] メモリセル部MCは、図5および図6に示すように、出力メモリアレイ部MAoに並列に出力される出力データの出力並列数q個、並列に出力される出力データO1、・・・、Oqに対応して、2値化ニューラル電子回路NNに配置される。

[0053] 図6に示すように、並列で入力される並列の各入力データに配置された入力並列数p個のプロセスエレメント部Peは、2値化ニューラル電子回路N

Nにおいて、プロセスエレメント・コラム（例えば、プロセスエレメント・コラムPC1）を形成する。出力並列数q個のプロセスエレメント・コラムPC1からPCqは、並列に出力される出力データに対応して、2値化ニューラル電子回路NNにおいて、q列に配置される。プロセスエレメント部Peは、図3に示すように、2値化ニューラル電子回路NNにおいて、p行×q列に2次元の演算器アレイとして設定される。

[0054] 行列(1, 1)、(1, 2)、・・・、(1, q)のプロセスエレメント部Peには、入力データI1が共通で入力される結線になっている。行列(2, 1)、(2, 2)、・・・、(2, q)のプロセスエレメント部Peには、入力データI2が共通で入力される結線になっている。行列(p, 1)、(p, 2)、・・・、(p, q)のプロセスエレメント部Peには、入力データIpが共通で入力される結線になっている。

[0055] プロセスエレメント部Peは、対応するメモリセル10から出力される1ビットの重み付け係数と、1ビットの入力データとの排他的論理和(XNOR)を乗算結果として算出して出力する。

[0056] なお、接続有無情報用のメモリセルからの接続無情報（例えば、「NC」を意味する所定値）が出力された場合、加算活性化部Actにおいて、乗算結果が加算されない。例えば、乗算結果と接続有無情報とが交互にペアで出力されてもよい。また、接続有無情報に関して、プロセスエレメント部Peから加算活性化部Actへ、乗算結果とは独立の結線を有し、乗算結果と接続有無情報とは別々に出力されてもよい。

[0057] なお、プロセスエレメント部Peで、乗算結果の部分加を計算するとき、接続有無情報用のメモリセルからの接続無情報（例えば、「NC」を意味する所定値）が出力された場合、乗算結果の部分加に加算されない。

[0058] プロセスエレメント・コラムPC1、・・・、PCqは、各プロセスエレメント部Peからの乗算結果または一部の乗算結果を加算した部分加結果等を加算活性化部Actに出力する。

[0059] 図5および図6に示すように、加算活性化部Actは、並列で出力される

各出力データ O_1 、 $\dots$ 、 O_q に応じて配置されている。

[0060] 加算活性化部 Act は、プロセスエレメント・コラムから逐次出力される乗算結果を、接続有無情報に基づき加算して、加算結果に活性化関数を適用して1ビットの出力データを出力メモリアレイ部 MAo に出力する。プロセスエレメント部 Pe が乗算結果の部分和を出力する場合、加算活性化部 Act は、プロセスエレメント・コラムから逐次出力される乗算結果を加算して、加算結果に活性化関数を適用して1ビットの出力データを出力メモリアレイ部 MAo に出力する。

[0061] 加算活性化部 Act は、プロセスエレメント・コラムにおいて、入力データの1サイクル単位で、乗算結果として「1」を算出した回数から、乗算結果として「0」を算出した回数を減じた値が予め定められた閾値以上の場合に「1」を出力データとして出力し、減じた値が閾値未満の場合に「0」を出力データとして出力する。

[0062] 図6に示すように、ビットシリアル入力の並列化が行われ、プロセスエレメント部 Pe の行が入力データに対して共有し、プロセスエレメント部 Pe の列である各プロセスエレメント・コラムが独立して出力データを出力する。

[0063] なお、2値化ニューラル電子回路 NN が、2ビットのデータの他に、複数ビットのデータも扱えるように、対数化する回路を有してもよい。例えば、2値化ニューラル電子回路 NN が、入力データと重み付け係数との乗算結果を、対数化入力データと対数化重み付け係数とを加算した対数加算して逆変換で線形化する回路を有してもよい。また、ゼロ出力特定部3は、2値化ニューラル電子回路 NN の他に、最終的に0/1を返す機械学習方式であればよい。

[0064] (3. ニューラル計算装置の動作例)

次に、ニューラル計算装置1の動作例について図7から図12を用いて説明する。なお、ニューラル計算装置1が畳み込みニューラルネット、ゼロ出力特定部3が2値化ニューラルネットワークである場合を例として、動作

を説明する。

[0065] 図7は、ニューラル計算装置の動作例を示すフローチャートである。図8は、ニューラルネット計算部およびゼロ出力特定部の一例を示すブロック図である。図9および図10は、ニューラルネット計算部およびゼロ出力特定部の動作のタイミングの一例を示す模式図である。図11は、ニューラル計算装置による活性予測精度の結果の一例を示す模式図である。図12は、ニューラル計算装置による計算削減率の結果の一例を示す模式図である。

[0066] 図7に示すように、ニューラル計算装置1は、学習済みニューラルネットモデルを取得する（ステップS1）。具体的には、ニューラル計算装置1の制御部4は、畳み込みニューラルネットを構築に必要な重み付け係数の値をメモリ5から取得する。制御部4は、ニューラルネット計算部2に重み付け係数の値を出力して、学習済みニューラルネットモデルを設定する。また、制御部4は、活性化関数を正規化線形関数に設定する。

[0067] 例えば、図8に示すように、ニューラルネット計算部2において、ニューラルレイヤー20が、畳み込み層である場合、畳み込み演算を行う乗算加算部21と、正規化線形関数部25とが設定される。なお、出力層のニューラルレイヤー20は、フル結合層でもよい。隠れ層のニューラルレイヤー20に、フル結合層が含まれてもよい。

[0068] なお、学習済みニューラルネットモデルは、ニューラルネット計算部2に対して、所定の学習用のデータセットによって予め学習が行われて、構築される。例えば、ニューラルネット計算部2に学習用のデータセットを適用して、誤差伝搬法により学習する。

[0069] 次に、ニューラル計算装置1は、ゼロ出力特定部3を学習させる（ステップS2）。例えば、図8に示すように、活性予測器30が2値化ニューラル計算器31の場合、制御部4は、ゼロ出力特定部3の各2値化ニューラル計算器31に、ニューラルネット計算部2に対して適用した同じ学習用のデータセットを適用して、ゼロ出力特定部3を学習させる。さらに具体的には、制御部4は、各ニューラルレイヤー20における乗算加算部21の出力であ

るOFM (Output Feature Map) と、ニューラルレイヤー20に対応する2値化ニューラル計算器31の出力であるOFMとを取得する。

[0070] 制御部4は、乗算加算部21のOFMを教師信号として、2値化ニューラル計算器31のOFMとの誤差を求め、この誤差に基づき誤差伝搬法により各2値化ニューラル計算器31を学習させる。制御部4は、誤差として損失関数の値が所定以下になるまで、ニュートン法、最急降下法等の最適化手法により2値化ニューラル計算器31の重み付けを修正して、学習を繰り返す。なお、誤差は、例えば2値化ニューラル計算器31の出力と乗算加算部21における加算結果との差分の2乗である。学習用のデータセットを適用して、2値化ニューラル計算器31に対する重み付け係数が修正されて、乗算加算部21を模した2値化ニューラル計算器31が形成される。

[0071] ここで、2値化ニューラル計算器31は、例えば、2値化ニューラル電子回路NNにより実現される。また、ゼロ出力特定部3が、複数のコア電子回路Coreにより実現されてもよい。

[0072] 2値化ニューラル計算器31用の重み付け係数は、メモリ5に記憶されていて、学習が進む毎に修正されている。制御部4は、メモリ5上の重み付け係数を逐次書き換えられて学習して行く。修正後の重み付け係数は、メモリ5からゼロ出力特定部3のメモリアクセス制御部MCntに転送され、メモリセル部MCに設定される。

[0073] なお、畳み込み演算を行う場合、2値化ニューラル電子回路NNにおいて、入力メモリアレイ部MAiにおける入力データと、メモリセル部MCからプロセスエレメント部Peに順次出力する重み付け係数とが、畳み込み演算になるように制御されて、畳み込み演算が実現される。なお、RGB画像の場合、 $p=3$ の入力データになる。入力画像が $k \times k$ ピクセルの場合、入力データ i_1 、 i_2 、 $\dots$ 、 i_k 、 $\dots$ 、 i_{k^2} が逐次入力される。 q の値は、次の層のニューロン数になる。

[0074] 学習した活性予測器30または2値化ニューラル計算器31は、前記ニューラルネット計算手段の計算を模した2値化ニューラルネット計算器の一例

である。学習した活性予測器 30 または 2 値化ニューラル計算器 31 は、前記ニューラルネット計算手段において前記活性化関数を適用する前までの計算（乗算加算部 21 の計算）を模した 2 値化ニューラル計算器の一例である。制御部 4 は、前記ゼロ出力特定手段の出力と前記ニューラルネット計算手段における前記加算結果との差異に基づき、前記 2 値化ニューラル計算器を学習させる学習手段の一例として機能する。

[0075] 次に、ニューラル計算装置 1 は、ゼロ出力特定部 3 を用いたニューラルネット計算を実行する（ステップ S3）。制御部 4 は、ニューラルネット計算部 2 用の学習済み重み付け係数を、メモリ 5 から読み出し、ニューラルネット計算部 2 にロードする。制御部 4 は、ゼロ出力特定部 3 用の学習済み重み付け係数を、メモリ 5 から読み出し、ゼロ出力特定部 3 にロードする。2 値化ニューラル電子回路 NN の場合は、各 2 値化ニューラル計算器 31 に対応したメモリセル部 MC に、ゼロ出力特定部 3 用の学習済み重み付け係数がセットされる。

[0076] 次に、図 9 に示すように、1 番目のニューラルレイヤー 20 に対応する 2 値化ニューラル計算器 31 の計算が開始される。1 番目の 2 値化ニューラル計算器 31 に、入力データのパターンとして、入力データが逐次入力され、入力データの最上ビット（例えば、0 がマイナス符号を示し、1 がプラス符号）により、2 値化ニューラル計算器 31 おける計算が行われる。2 値化ニューラル計算器 31 の計算結果として、加算活性化部 Act の出力データ O1、O2・・・が、1 番目の 2 値化ニューラル計算器 31 から制御部 4 に送信される。

[0077] 出力データ O1 が「0」ならば、出力データ O1 に対応するニューロン計算器 20a の乗算加算器 21a の計算を省略するように、制御部 4 が、1 番目の乗算加算部 21 を制御する。出力データ O1 が「1」ならば、出力データ O1 に対応するニューロン計算器 20a の乗算加算器 21a の計算を、同じ入力データで計算を開始するように、制御部 4 が、1 番目の乗算加算部 21 を制御する。乗算加算部 21 の計算結果は、正規化線形関数部 25 が適用

されて出力される。

[0078] このように、活性予測器 30 または 2 値化ニューラル計算器 31 は、前記入力データに応じて、出力がゼロになる前記ニューロン計算器を特定するための計算を行うゼロ出力特定手段の一例として機能する。

[0079] 次に、出力データ 02 が「0」ならば、出力データ 02 に対応するニューロン計算器 20 a の乗算加算器 21 a の計算を省略するように、制御部 4 が、1 番目の乗算加算部 21 を制御する。出力データ 02 が「1」ならば、出力データ 02 に対応するニューロン計算器 20 a の乗算加算器 21 a の計算を、同じ入力データで計算を開始するように、制御部 4 が、1 番目の乗算加算部 21 を制御する。各出力データに対して、制御部 4 が、1 番目の乗算加算部 21 を制御する。例えば、図 4 に示すように、第 1 層 L1 において、出力がゼロとならないと予想される乗算加算器 21 a の計算が行われる。

[0080] 1 番目の乗算加算部 21 および正規化線形関数部 25 からの出力が、入力データのパターンとして、2 番目の 2 値化ニューラル計算器 31 に逐次入力され、入力データの最上ビット（例えば、0 ならば 0 を示し、1 ならばそのビット数で示せる最大値を示す）により、2 値化ニューラル計算器 31 おける計算が行われる。

[0081] なお、1 番目の乗算加算部 21 および正規化線形関数部 25 からの出力が、中間データとして、ニューラルネット計算部 2 のキャッシュまたはメモリ 5 に記憶されてもよい。

[0082] 2 番目の 2 値化ニューラル計算器 31 の出力データが逐次出力され、各出力の値（「0」または「1」）に応じて、対応するニューロン計算器 20 a の乗算加算器 21 a の計算を省略、または、乗算加算器 21 a の計算を行うように、制御部 4 が、2 番目の乗算加算部 21 を制御する。例えば、図 4 に示すように、第 2 層 L1 において、出力がゼロとならないと予想される乗算加算器 21 a の計算が行われる。

[0083] 2 番目の乗算加算部 21 および正規化線形関数部 25 からの出力が、入力データとして、3 番目の 2 値化ニューラル計算器 31 に逐次入力され、入力

データの最上ビット（例えば、0ならば0を示し、1ならばそのビット数で示せる最大値を示す）により、2値化ニューラル計算器31における計算が行われる。

[0084] 最後のニューラルレイヤー20まで計算されて、ニューラルネット計算部2は出力データを出力する。

[0085] 図9に示すように、出力がゼロとなると予測されるニューロン計算器20aの乗算加算器21aの計算を省くことにより、従来のニューラルネット計算部より速く計算できるようになる。従来のニューラルネット計算部があるレイヤーで計算時間 t_0 であることに対して、ニューラルネット計算部2の乗算加算部21は計算時間 t_2 ($< t_0$) で計算し、ゼロ出力特定部3の2値化ニューラル計算器31は計算時間 t_1 ($< t_0$) で計算し、かつ、2値化ニューラル計算器31と乗算加算部21とが少し時間をずらして、計算が可能となったニューロン計算器20aから計算始めるといふ並列的に処理することで、計算時間 t_3 ($< t_0$) で計算可能となる。

[0086] 制御部4は、前記ニューラルレイヤーに対応するゼロ出力特定手段により出力がゼロで無いと特定された前記ニューロン計算器から計算を先行して開始するように、前記ニューラルネット計算手段の計算を制御する制御手段の一例として機能する。

[0087] なお、図10に示すように、ニューラルネット計算部2における乗算加算部21と次の層の乗算加算部21との計算の時間間隔を空けてもよい。ゼロ出力特定部3の2値化ニューラル計算器31の計算が終わってから、次の層の乗算加算部21の計算を開始してもよい。また、同じ層の乗算加算部21と2値化ニューラル計算器31との計算で重なる時間があってもよい。

[0088] 制御部4は、前記ニューラルレイヤーにおける計算と次の層の前記ニューラルレイヤーにおける計算との間に時間間隔を設けて消費電力を減少させるように、前記ニューラルネット計算手段の計算を制御する制御手段の一例として機能する。

[0089] このように、制御部4は、前記入力データから変換して得られる1ビット

のデータを入力した前記ゼロ出力特定手段の出力に基づき、前記ニューラルネット計算手段の計算を制御する制御手段の一例である。ニューラルレイヤー20に対応する2値化ニューラル計算器31が、前の層のニューラルレイヤーの出力データに応じて、当該ニューラルレイヤーに属する前記ニューロン計算器の出力がゼロになるニューロン計算器を特定する計算を行うゼロ出力特定手段の一例である。活性予測器30または2値化ニューラル計算器31は、前記入力データのパターンから、前記加算結果の符号がゼロ以下になる前記ニューロン計算器を特定する計算を行うゼロ出力特定手段の一例として機能する。

[0090] 次に、シミュレーションの結果を図11および図12を用いて説明する。畳み込み層が8層でフル結合層が2層で、入力層を含めて11層のモデルでシミュレーションを行った。

[0091] 図11に示すように、各2値化ニューラル計算器31の活性予測精度は、学習前（破線）で層が先に進むほど、低下する傾向であったが、学習後は、3層目（conv2_1）以降で、活性予測精度が向上した。なお、横軸が各層で、縦軸が各層の活性予測精度である。各層の活性予測精度は、ゼロ出力特定部3を使用せずニューラルネット計算部2で計算した場合を正解とした場合において、ゼロと予測されたニューロン計算器20aの数を活性予測精度の分母とし、予測が正解であったニューロン計算器20aの数を活性予測精度の分子とした値である。

[0092] 図12に示すように、計算削減率に関して、3層目（conv2_1）以降で、計算削減率が向上した。各層の計算削減率は、各層のニューロン計算器20aの数を計算削減率の分母として、2値化ニューラル計算器31によりゼロと予測された各層のニューロン計算器20aの数とした値である。

[0093] これらのシミュレーション結果から、第1層L1に対応する活性予測器30と、第2層L2に対応する活性予測器30と、を省いたゼロ出力特定部3でもよい。

[0094] 以上説明したように、本実施形態によれば、入力データiと重み付け係数

wとの乗算結果を加算し、当該加算結果に対して、活性化関数を適用して出力データoを出力する複数のニューロン計算器20aから構成されるニューラル計算装置1のニューラルネット計算において、ゼロ出力特定部3が入力データiに応じて出力がゼロになるニューロン計算器20aを特定するための計算を行い、制御部4が、この計算結果に基づき、出力がゼロになるニューロン計算器20aの計算を省略するようにニューラルネット計算部2のニューラルネット計算を制御することにより、入力データのパターンに応じて、出力がゼロになるニューロン計算器20aを特定できるため、ニューロンに関する計算を動的に省略できる。

[0095] ゼロ出力特定部3が、ニューラルネット計算部2の計算を模した2値化ニューラル計算器31であり、制御部4が、入力データの最上ビットを入力した出力特定部3の出力に基づき、ニューラルネット計算部2の計算を制御する場合、2値化して計算することにより、ニューラルネット計算手段より速く計算できるため、ニューラルネット計算部2を計算する前に、出力がゼロになるニューロン計算器20aを特定できる。

[0096] ゼロ出力特定部3が、ニューラルネット計算部2において活性化関数を適用する前までの計算を模した2値化ニューラル計算器31である場合、活性化関数を適用する計算を省いて、2値化ニューラル計算器31を適用するため、より速く、出力がゼロになるニューロン計算器20aを特定できる。

[0097] ゼロ出力特定部3の出力とニューラルネット計算部2における加算結果との差異に基づき、2値化ニューラル計算器31を学習させる場合、学習させることにより、ゼロ出力特定部3の予測精度が向上する。

[0098] ニューラルネット計算部2が、複数のニューラルレイヤー20から構成され、各ニューラルレイヤー20に対応する複数の活性予測器30があり、ニューラルレイヤー20に対応する活性予測器30が、前の層のニューラルレイヤー20の出力データに応じて、当該ニューラルレイヤー20に属するニューロン計算器20aの出力がゼロになるニューロン計算器20aを特定する計算を行う場合、層毎に対応した活性予測器30が出力がゼロになるニュー

ーロン計算器 20a をより精度よく特定できる。

[0099] 制御部 4 が、ニューラルレイヤー 20 に対応するゼロ出力特定部 3 により出力がゼロで無いと特定されたニューロン計算器 20a から計算を先行して開始するように、ニューラルネット計算部 2 の計算を制御する場合、ニューラル計算装置 1 の計算をより速く実現できる。

[0100] 制御部 4 が、ニューラルレイヤー 20 における計算と次の層のニューラルレイヤー 20 における計算との間に時間間隔を設けて消費電力を減少させるように、ニューラルネット計算部 2 の計算を制御する場合、最も消費電力が高いニューラルネット計算部 2 を省力化でき、ニューラル計算装置 1 全体のエネルギー消費量を節約することができる。

[0101] 活性化関数が正規化線形関数である場合、正規化線形関数の適用においてゼロ以下であれば、出力がゼロであるので、ニューラルネット計算部 2 の計算を省略しやすくなる。

[0102] ゼロ出力特定部 3 が、入力データのパターンから、加算結果の符号がゼロ以下になるニューロン計算器 20a を特定する計算を行う場合、出力がゼロになるニューロン計算器 20a をより特定しやすくなる。

[0103] (4. 変形例)

次に、ゼロ出力特定部の変形例について図を用いて説明する。

図 13 は、ゼロ出力特定部がサポートベクターマシンの場合のブロック図である。図 14 は、ゼロ出力特定部がサポートベクターマシンの識別関数等を説明する模式図である。図 15 は、ニューラルネット計算部およびゼロ出力特定部の変形例を示すブロック図である。

[0104] 図 13 に示すように、活性予測器 30 がサポートベクターマシン識別器 32 であってもよい。ニューラル計算装置 1 は、ニューラルネット計算部 2 と、ゼロ出力特定部 3A と、を有する。ゼロ出力特定部 3A は、2 値化ニューラル計算器 31 の代わりに複数のサポートベクターマシン識別器 32 を有する。サポートベクターマシン識別器 32 は、各層のニューラルレイヤー 20 に対応して設置される。

- [0105] ステップS 2において、活性予測器30がサポートベクターマシン識別器32の場合、制御部4は、ゼロ出力特定部3Aの各サポートベクターマシン識別器32に、ニューラルネット計算部2に対して適用した同じ学習用のデータセットを適用して、ゼロ出力特定部3Aを学習させる。さらに具体的には、制御部4は、ニューラルレイヤー20における乗算加算部21の出力であるOFMと、ニューラルレイヤー20に対応するサポートベクターマシン識別器32の出力であるOFMとを取得する。
- [0106] 制御部4は、図14に示すように乗算加算部21のOFMを正解ラベル t として、サポートベクターマシン識別器32のOFMとの差異を求め、この差異による損失関数に基づき、制約つきニュートン法や最急降下法等の最適化手法により、各サポートベクターマシン識別器32を学習させる。制御部4は、損失関数が所定以下になるまで、学習を繰り返す。学習用のデータセットを適用して、サポートベクターマシン識別器32に対する重み付け係数が修正されて、乗算加算部21を模したサポートベクターマシン識別器32が形成される。なお、図14に示すように、サポートベクターマシンの識別関数が、 $y = w^{(\text{sum})} i + w_0^{(\text{sum})}$ である。
- [0107] サポートベクターマシン識別器32の学習が完了したら、ステップS3のように、ニューラル計算装置1は、2値化ニューラル計算器31の代わりにサポートベクターマシン識別器32を用いたニューラルネット計算を実行する。なお、サポートベクターマシン識別器32は、最上ビット化しない入力データでも、2値化ニューラル計算器31のように最上ビット化した入力データでもよい。
- [0108] 次に、図15に示すように、畳み込み層でなく、フル結合層の場合にも、本願発明を適用できる。すなわち、ニューラルレイヤー20がフル結合層の場合、活性予測器30がフル結合の2値化ニューラル計算器33であってもよい。各層のニューラルレイヤー20フル結合の乗算加算部22と、ゼロ出力特定部3Bにおけるフル結合の2値化ニューラル計算器33とが対応する。

- [0109] なお、フル結合の場合、2値化ニューラル電子回路NNにおいて、入力メモリアレイ部MA_iにおける入力データと、メモリセル部MCからプロセスエレメント部P_eに順次出力する重み付け係数とが、フル結合の演算になるように制御されて、フル結合の演算が実現される。例えば、図6に示すように、 $p \times q$ のフル結合のニューラルネットワークの場合、入力並列数 p 個のプロセスエレメント部P_eが並んだプロセスエレメント・コラムと、出力並列数 q 個のプロセスエレメント・コラムPC₁、PC₂、 $\dots$ 、PC_qと、出力並列数 q 個のメモリセル部MCとを有する。制御部C_{nt}は、ニューラル電子回路NNのうち、プロセスエレメント・コラムPC₁、PC₂、 $\dots$ 、PC_qと、 q 個のメモリセル部MCとを、使用するように制御する。また、 $p \times q$ 以上の場合、コア電子回路C_{ore}同士を、直列または並列に連結して、実現してもよい。
- [0110] 畳み込み演算を行う2値化ニューラル計算器31と、フル結合の2値化ニューラル計算器33とは、構成は同じで、演算を行う上で、メモリセル部MCから供給する重み付け等の違いに過ぎないので、2値化ニューラル計算器31と、2値化ニューラル計算器33との動作は、基本的に同じである。
- [0111] 以上のように、活性予測器30がサポートベクターマシン識別器32でも、2値化ニューラル計算器33でも、2値化ニューラル計算器31と同様の効果が得られる。

符号の説明

- [0112] 1：ニューラル計算装置
2：ニューラルネット計算部（ニューラルネット計算手段）
3、3A、3B：ゼロ出力特定部（ゼロ出力特定手段）
4：制御部（制御手段）
20：ニューラルレイヤー（ニューラルネット計算手段）
20a：ニューロン計算器
21、22：乗算加算部
25：正規化線形関数部

30 : 活性予測器 (ゼロ出力特定手段)

31、33 : 2値化ニューラル計算器 (ゼロ出力特定手段)

32 : サポートベクターマシン識別器 (ゼロ出力特定手段)

i : 入力データ

w : 重み付け係数

o : 出力データ

請求の範囲

- [請求項1] 入力データと重み付け係数との乗算結果を加算し、当該加算結果に対して、活性化関数を適用して出力データを出力する複数のニューロン計算器から構成されるニューラルネット計算手段と、
- 前記入力データに応じて、出力がゼロになる前記ニューロン計算器を特定するための計算を行うゼロ出力特定手段と、
- 前記ゼロ出力特定手段の計算結果に基づき、前記出力がゼロになるニューロン計算器の計算を省略するように前記ニューラルネット計算手段の計算を制御する制御手段と、
- を備えたことを特徴とするニューラル計算装置。
- [請求項2] 請求項1に記載のニューラル計算装置において、
- 前記ゼロ出力特定手段が、前記ニューラルネット計算手段の計算を模した2値化ニューラルネット計算器であり、
- 前記制御手段が、前記入力データから変換して得られる1ビットのデータを入力した前記ゼロ出力特定手段の出力に基づき、前記ニューラルネット計算手段の計算を制御することを特徴とするニューラル計算装置。
- [請求項3] 請求項2に記載のニューラル計算装置において、
- 前記ゼロ出力特定手段が、前記ニューラルネット計算手段において前記活性化関数を適用する前までの計算を模した2値化ニューラル計算器であることを特徴とするニューラル計算装置。
- [請求項4] 請求項3に記載のニューラル計算装置において、
- 前記ゼロ出力特定手段の出力と前記ニューラルネット計算手段における前記加算結果との差異に基づき、前記2値化ニューラル計算器を学習させる学習手段を更に備えたことを特徴とするニューラル計算装置。
- [請求項5] 請求項1から請求項4のいずれか1項に記載のニューラル計算装置において、

前記ニューラルネット計算手段が、複数のニューラルレイヤーから構成され、

各前記ニューラルレイヤーに対応する複数の前記ゼロ出力特定手段があり、

前記ニューラルレイヤーに対応するゼロ出力特定手段が、前の層のニューラルレイヤーの出力データに応じて、当該ニューラルレイヤーに属する前記ニューロン計算器の出力がゼロになるニューロン計算器を特定する計算を行うことを特徴とするニューラル計算装置。

[請求項6] 請求項5に記載のニューラル計算装置において、

前記制御手段が、前記ニューラルレイヤーに対応するゼロ出力特定手段により出力がゼロで無いと特定された前記ニューロン計算器から計算を先行して開始するように、前記ニューラルネット計算手段の計算を制御することを特徴とするニューラル計算装置。

[請求項7] 請求項5または請求項6に記載のニューラル計算装置において、

前記制御手段が、前記ニューラルレイヤーにおける計算と次の層の前記ニューラルレイヤーにおける計算との間に時間間隔を設けて消費電力を減少させるように、前記ニューラルネット計算手段の計算を制御することを特徴とするニューラル計算装置。

[請求項8] 請求項1から請求項7のいずれか1項に記載のニューラル計算装置において、

前記活性化関数が正規化線形関数であることを特徴とするニューラル計算装置。

[請求項9] 請求項8に記載のニューラル計算装置において、

前記ゼロ出力特定手段が、前記入力データのパターンから、前記加算結果の符号がゼロ以下になる前記ニューロン計算器を特定する計算を行うことを特徴とするニューラル計算装置。

[請求項10] 入力データと重み付け係数との乗算結果を加算し、当該加算結果に対して、活性化関数を適用して出力データを出力する複数のニューロ

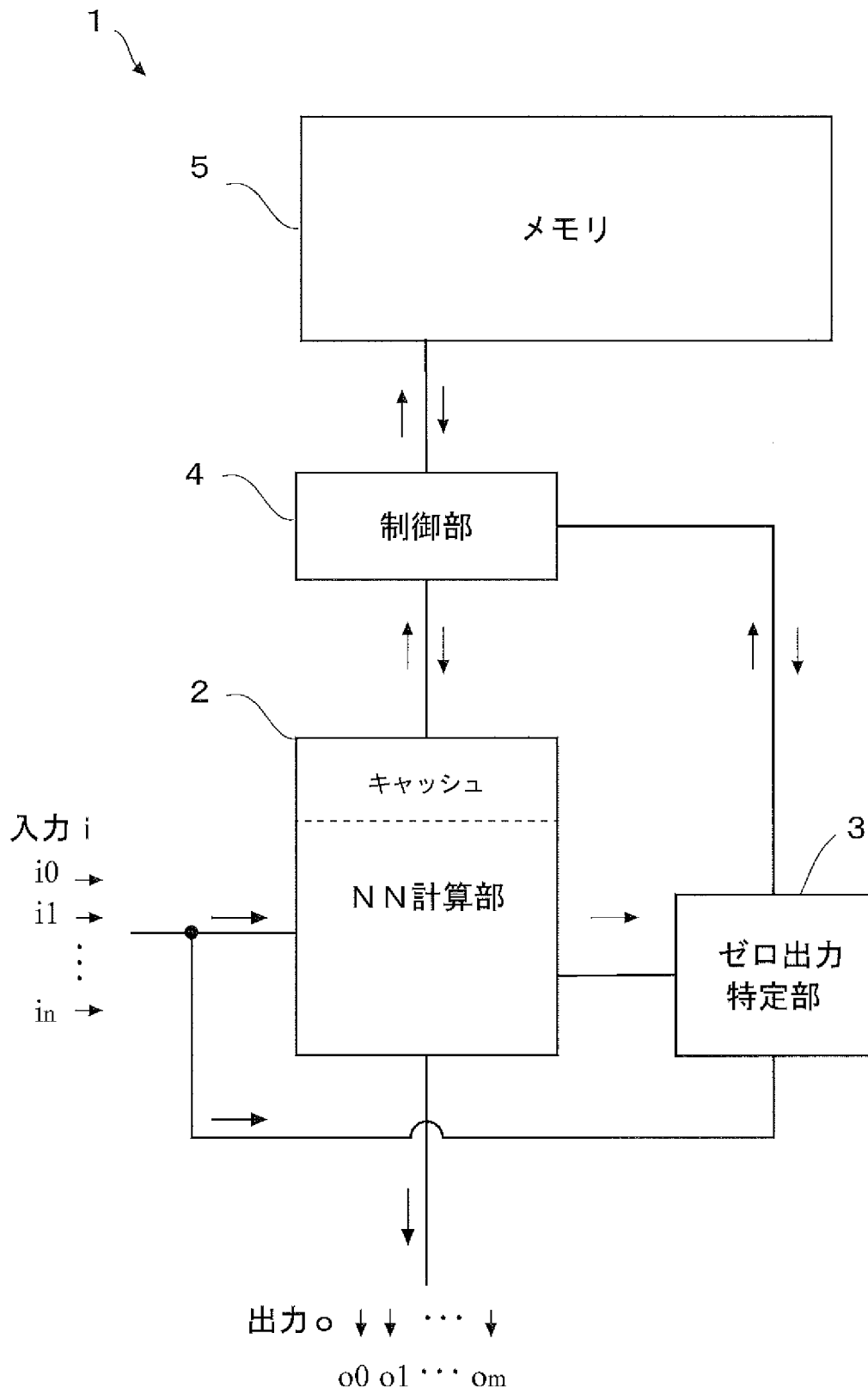
ン計算器から構成されるニューラルネット計算装置において、

ゼロ出力特定手段が、前記入力データに応じて、出力がゼロになる前記ニューロン計算器を特定するための計算を行う出力特定ステップと、

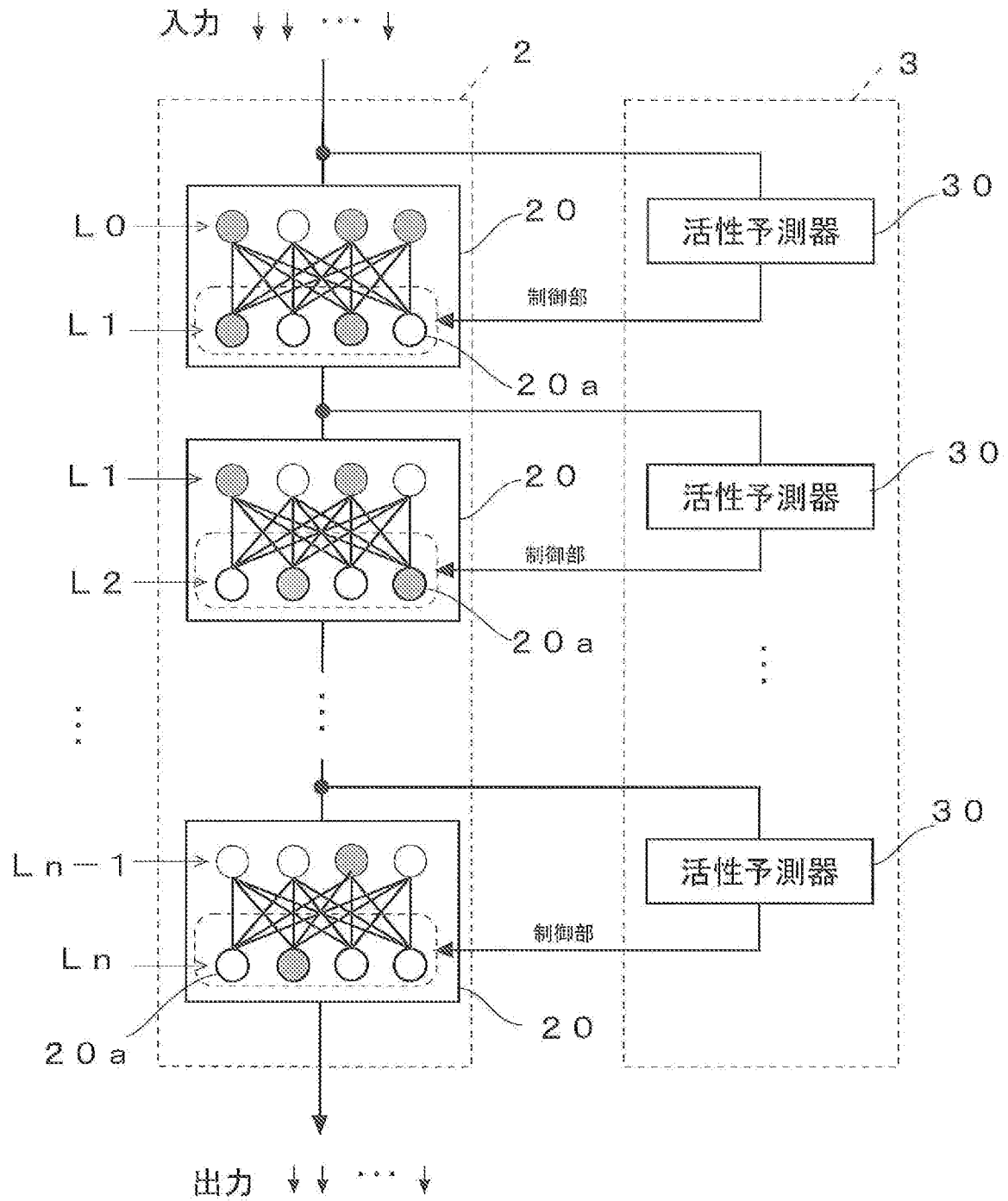
制御手段が、前記ゼロ出力特定手段の計算結果に基づき、前記出力がゼロになるニューロン計算器の計算を省略するように前記ニューラルネット計算装置の計算を制御する制御ステップと、

を含むことを特徴とするニューラル計算方法。

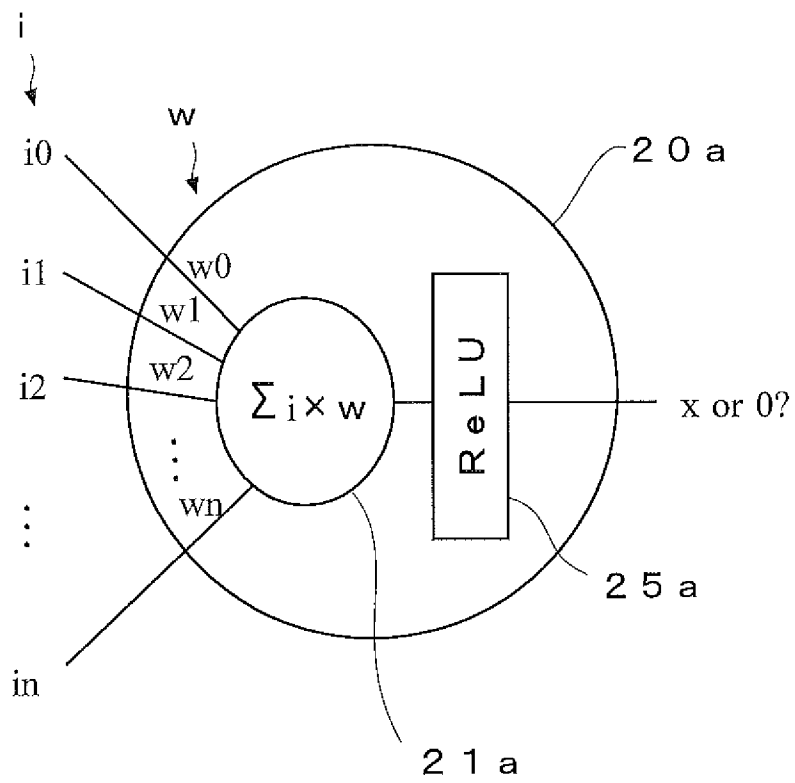
[図1]



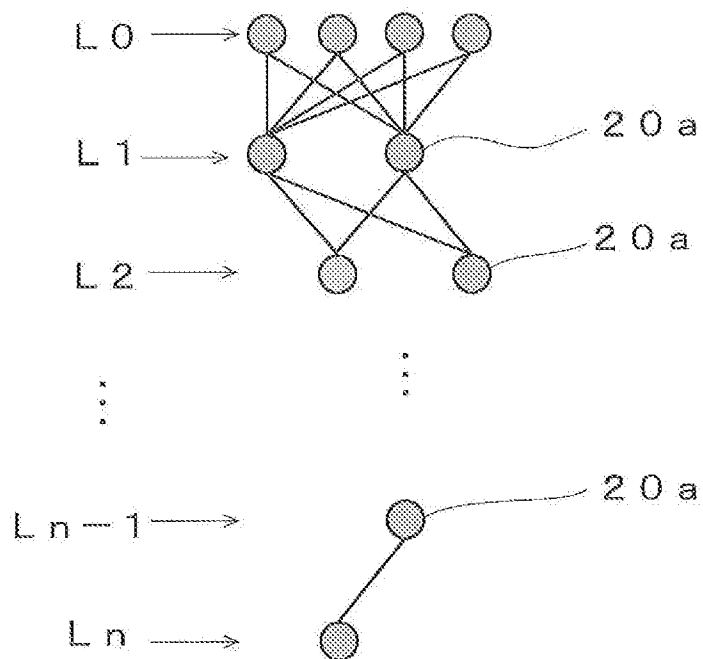
[図2]



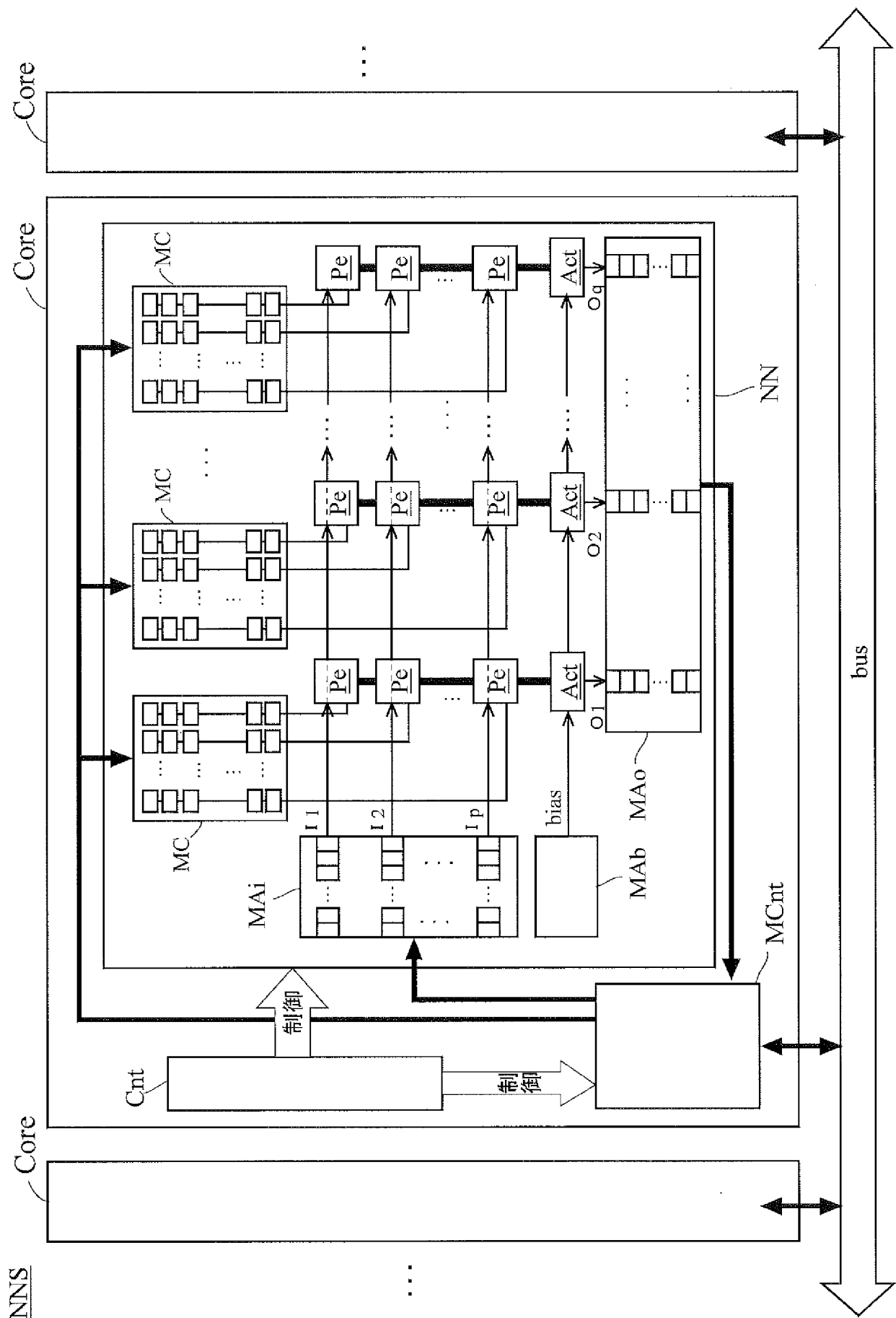
[図3]



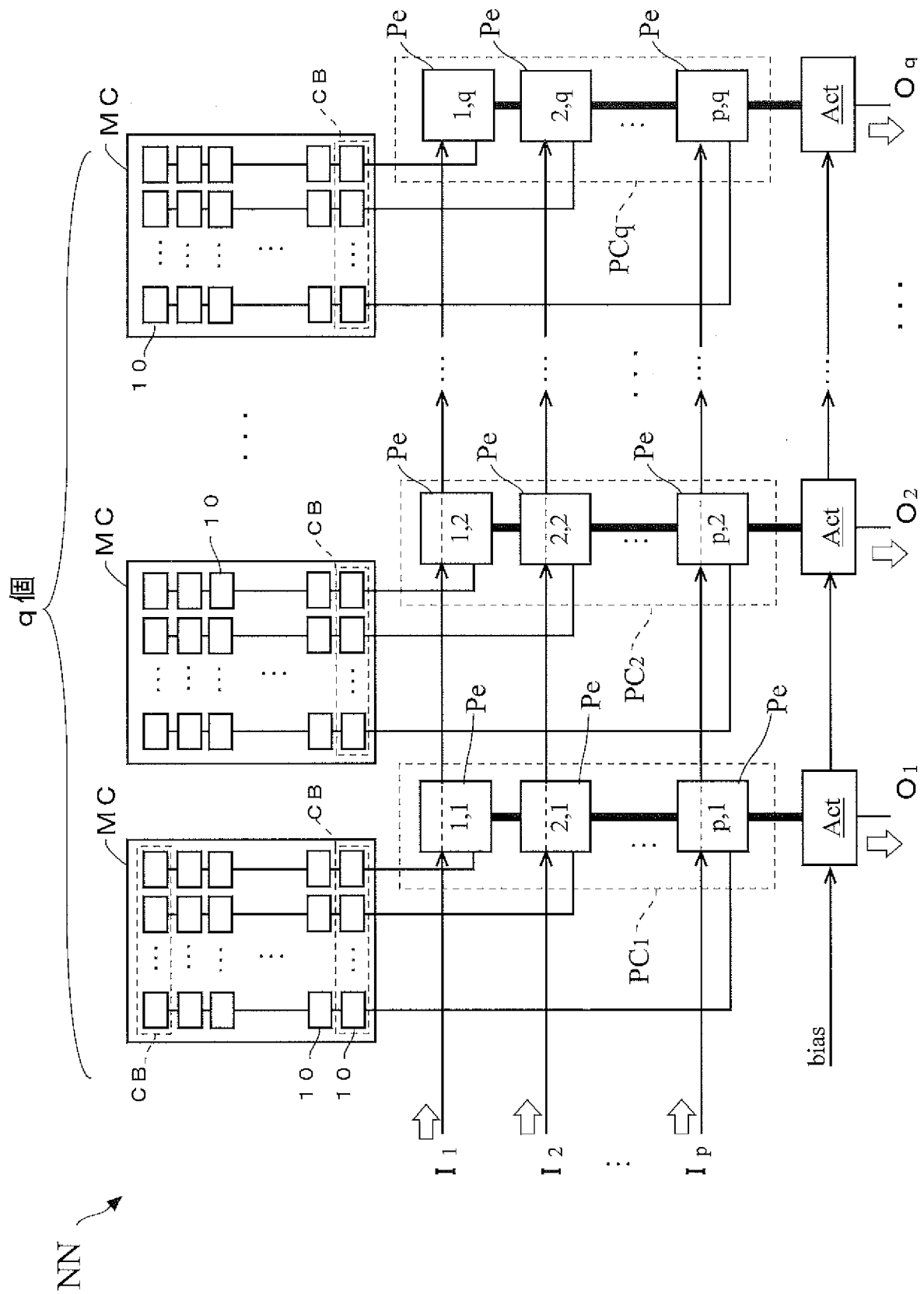
[図4]



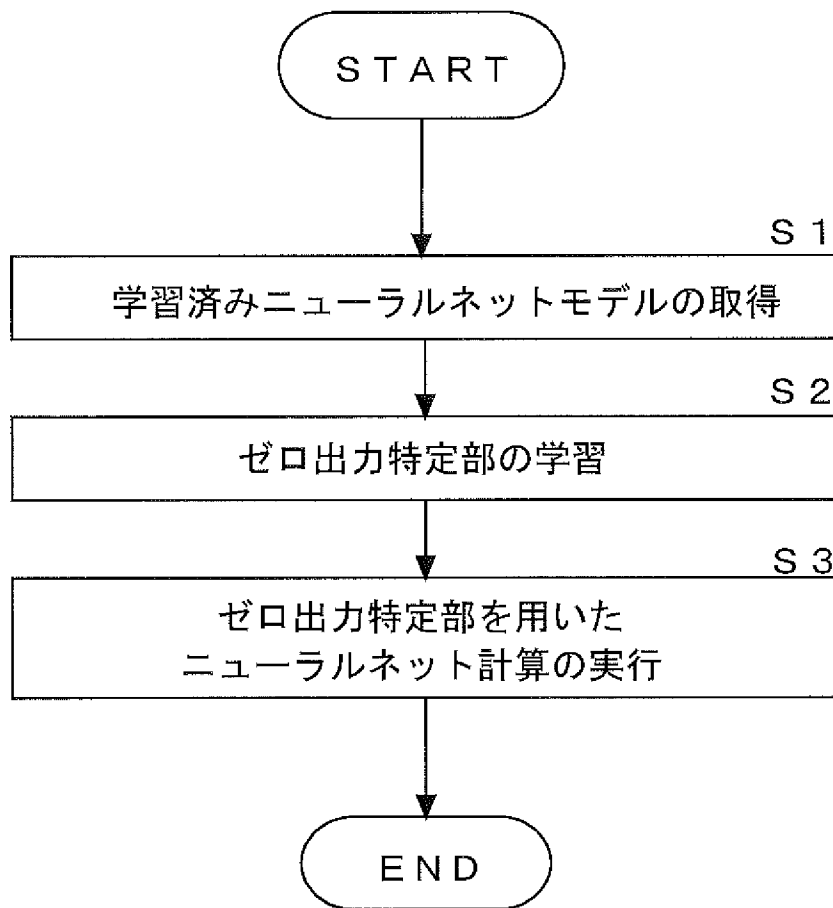
[図5]



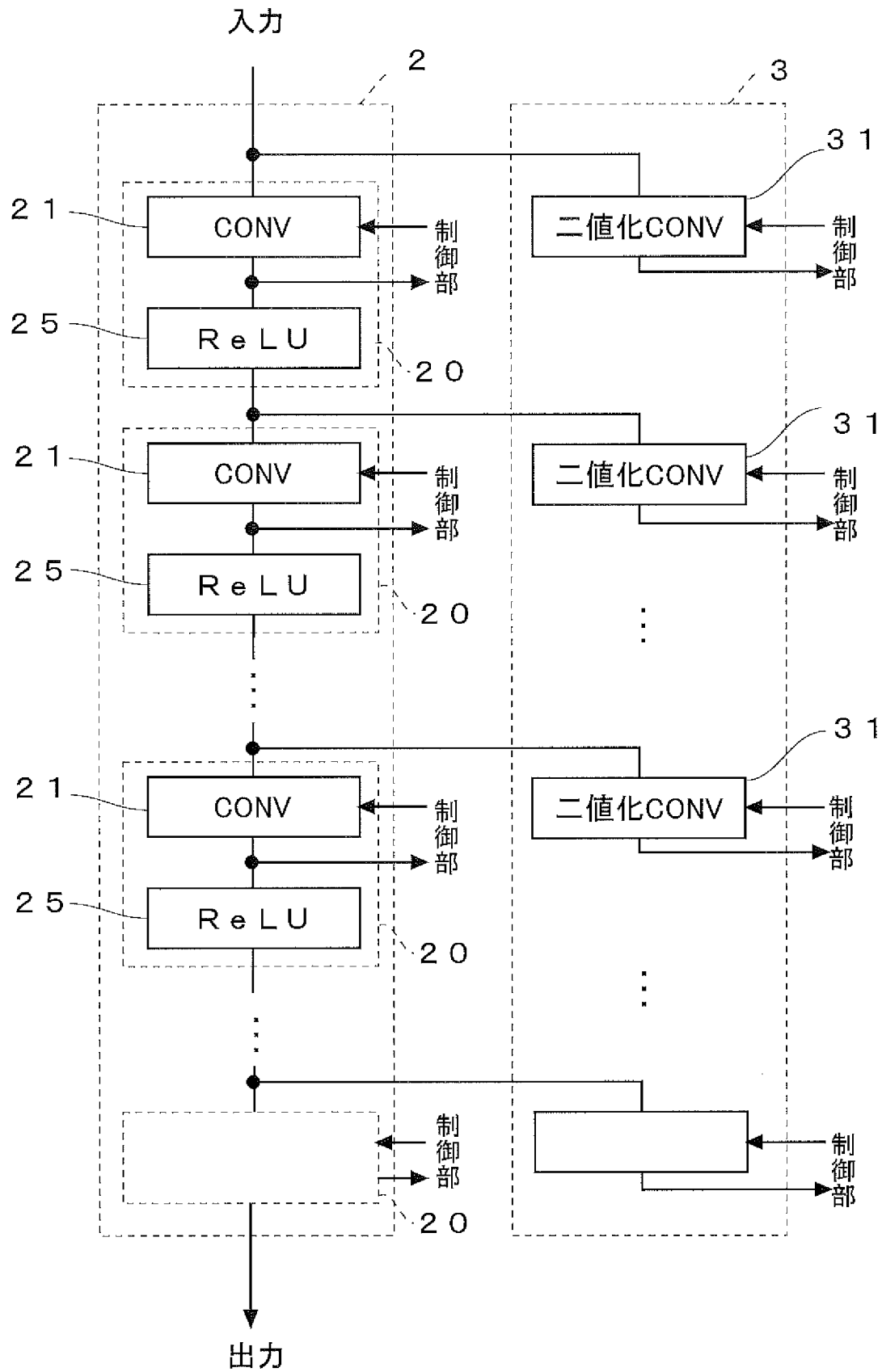
[図6]



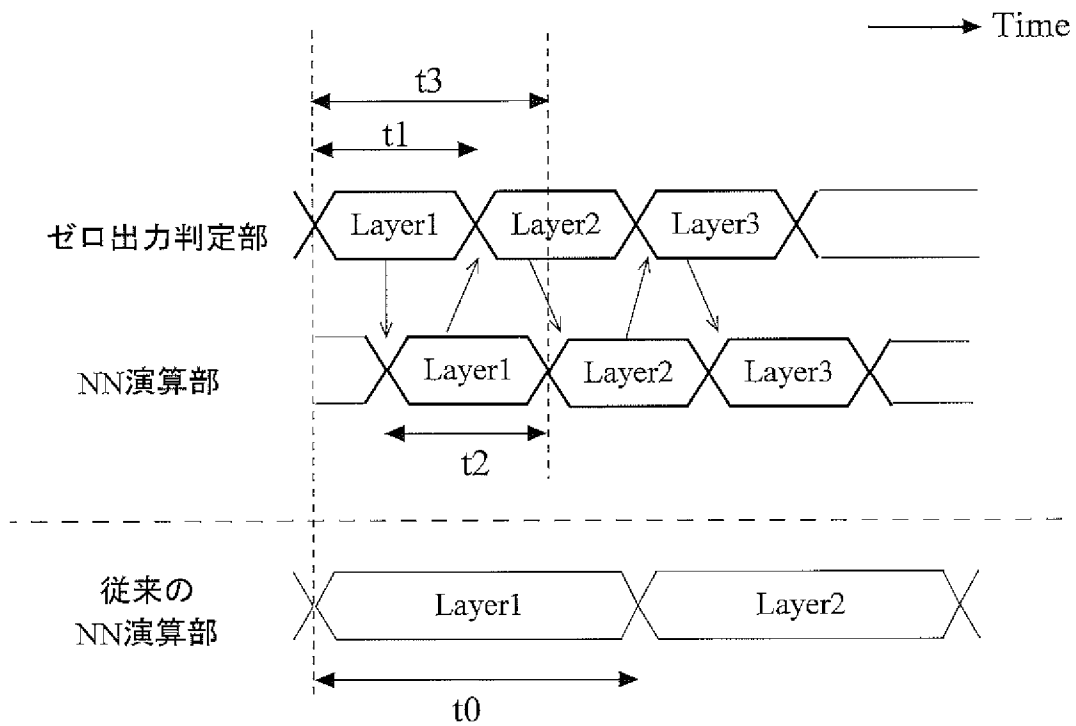
[図7]



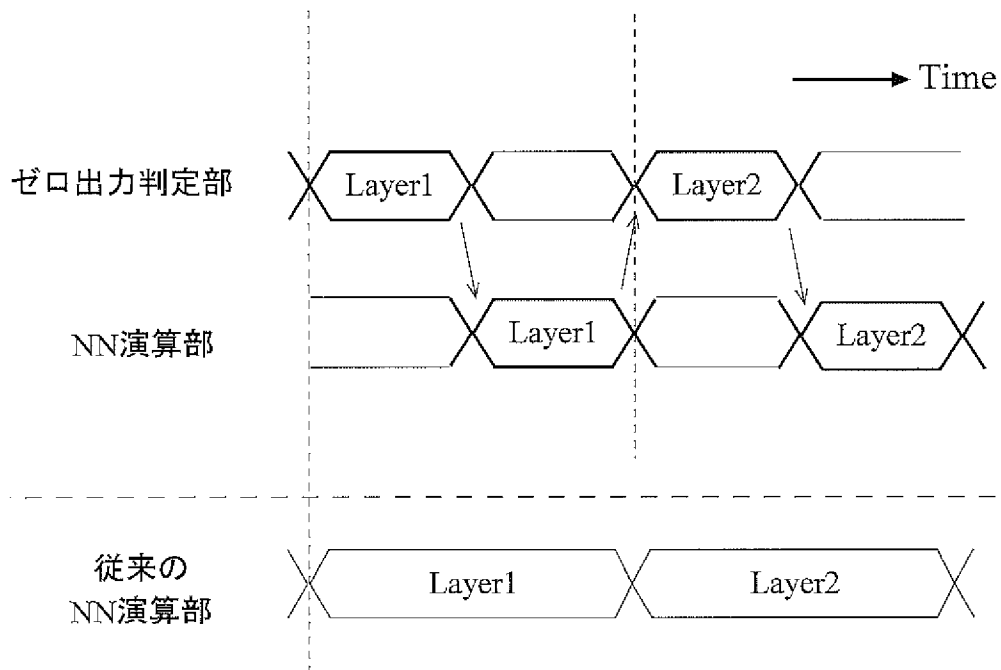
[図8]



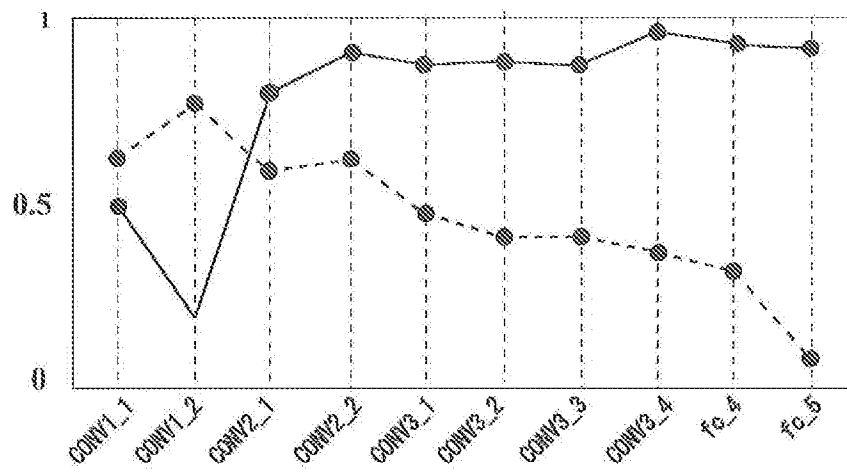
[図9]



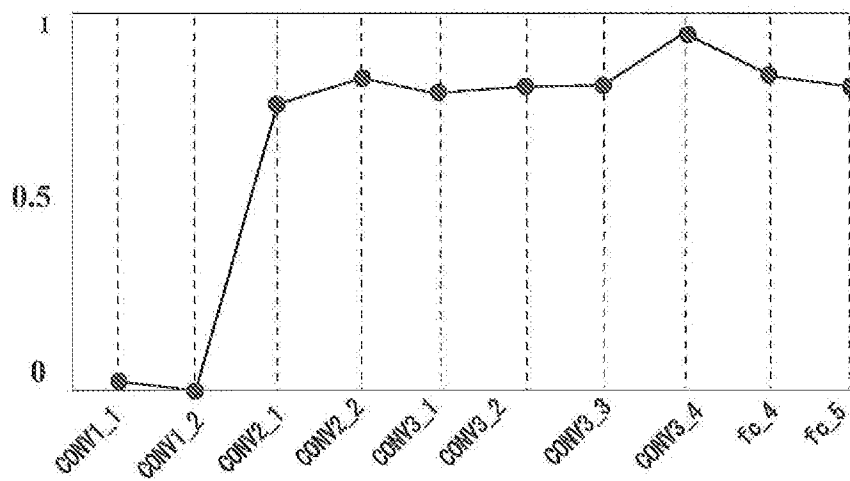
[図10]



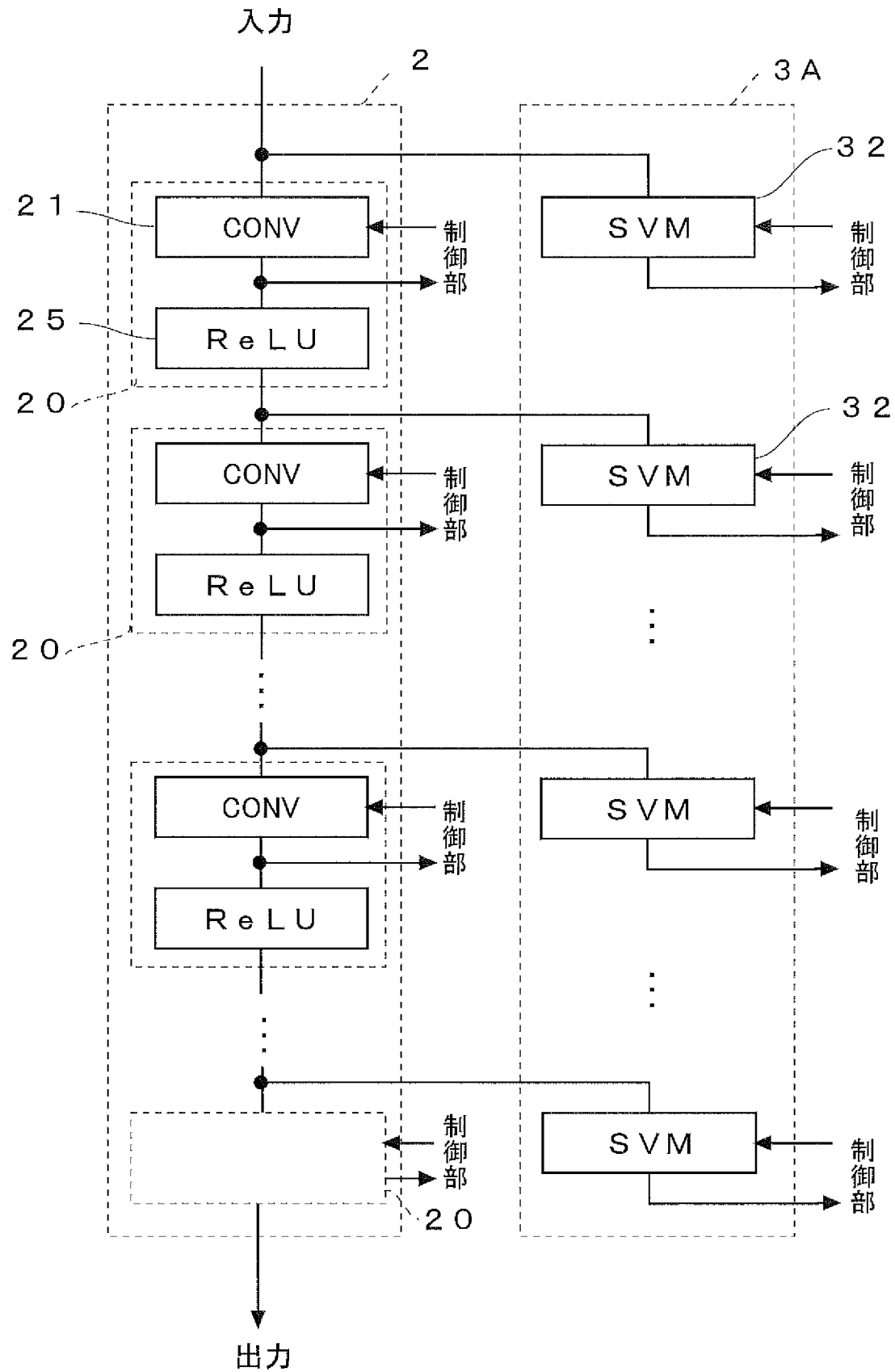
[図11]



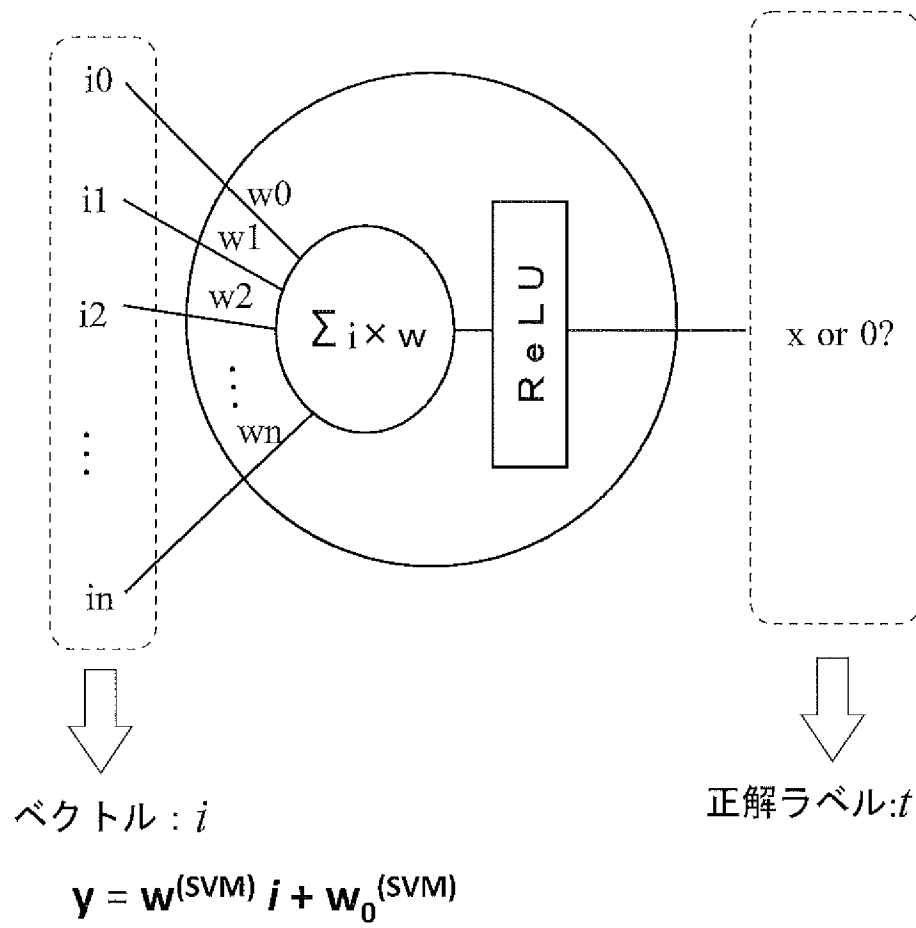
[図12]



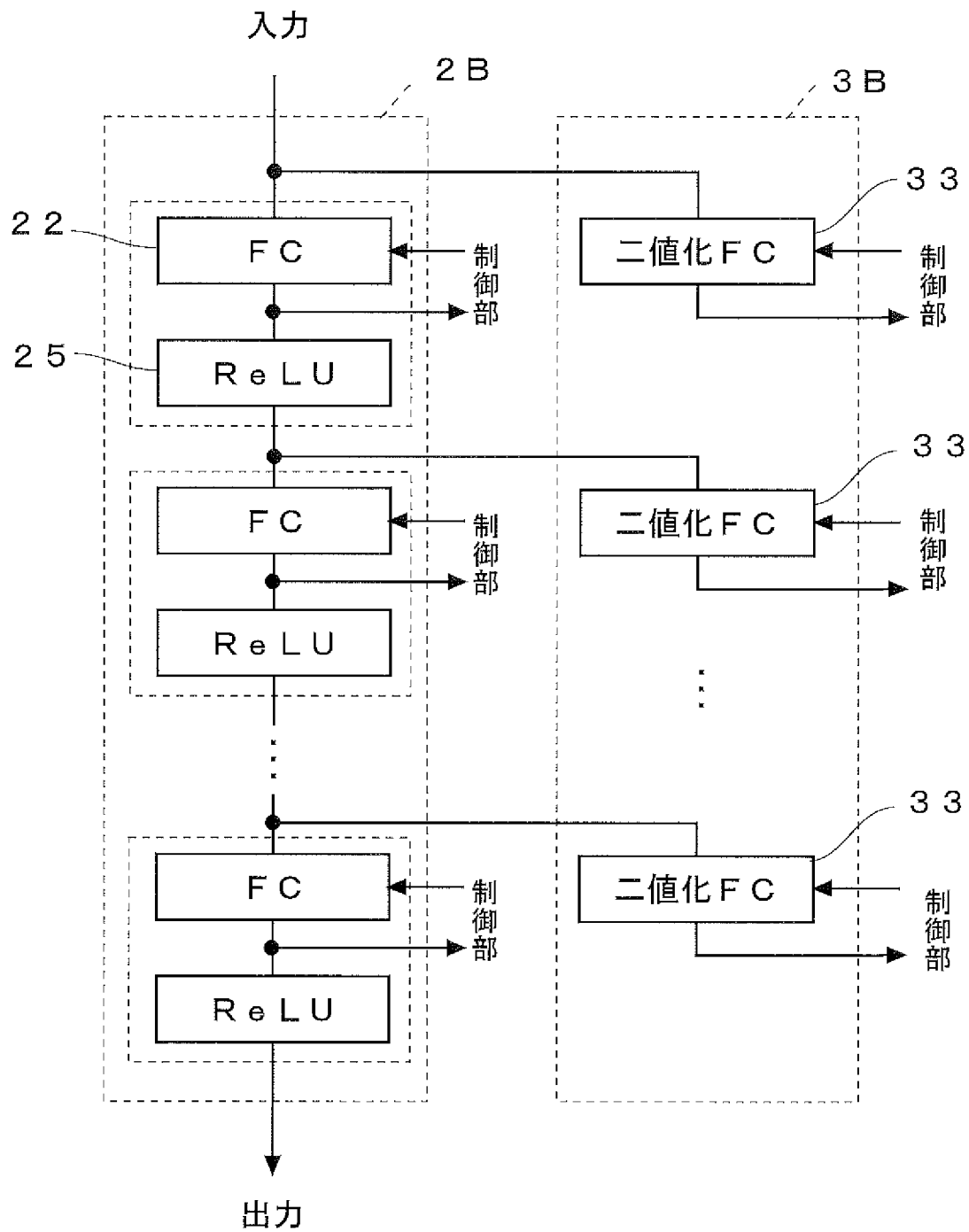
[図13]



[図14]



[図15]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2020/016684

A. CLASSIFICATION OF SUBJECT MATTER

G06N 3/02 (2006.01) i; G06N 3/08 (2006.01) i
FI: G06N3/08 120; G06N3/02

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)

G06N3/02; G06N3/08

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	1971-2020
Registered utility model specifications of Japan	1996-2020
Published registered utility model applications of Japan	1994-2020

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2018/0174025 A1 (SK HYNIX INC.) 21.06.2018 (2018-06-21) paragraphs [0044]-[0045], [0092], [0108], fig. 11-12	1, 5, 8-10
X	JP 7-302249 A (HITACHI, LTD.) 14.11.1995 (1995-11-14) paragraphs [0021], [0033]-[0038], [0046]-[0051]	1, 5, 8-10
A	JP 2018-129033 A (AXIS AB) 16.08.2018 (2018-08-16) paragraph [0003]	1-10
A	JP 2017-182319 A (MEGACHIPS CORPORATION) 05.10.2017 (2017-10-05) paragraphs [0050], [0053], [0057], [0066], [0075]-[0076]	2-4



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

"I" 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

16 June 2020 (16.06.2020)

Date of mailing of the international search report

23 June 2020 (23.06.2020)

Name and mailing address of the ISA/

Japan Patent Office
3-4-3, Kasumigaseki, Chiyoda-ku,
Tokyo 100-8915, Japan

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/JP2020/016684

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
US 2018/0174025 A1	21 Jun. 2018	KR 10-2018-0070194 A paragraphs [0028]- [0029], [0079], [0094], fig. 11-12	
JP 7-302249 A	14 Nov. 1995	EP 0583217 A2 page 3, line 58 to page 4, line 5, page 6, line 1 to page 6, line 32, page 7, line 2 to page 7, line 19	
JP 2018-129033 A	16 Aug. 2018	US 2018/0181867 A1 paragraph [0003]	
JP 2017-182319 A	05 Oct. 2017	(Family: none)	

A. 発明の属する分野の分類（国際特許分類（IPC）） G06N 3/02(2006.01)i; G06N 3/08(2006.01)i FI: G06N3/08 120; G06N3/02														
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）） G06N3/02; G06N3/08														
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														
<table border="0"> <tr> <td>日本国実用新案公報</td> <td>1922 - 1996年</td> </tr> <tr> <td>日本国公開実用新案公報</td> <td>1971 - 2020年</td> </tr> <tr> <td>日本国実用新案登録公報</td> <td>1996 - 2020年</td> </tr> <tr> <td>日本国登録実用新案公報</td> <td>1994 - 2020年</td> </tr> </table>			日本国実用新案公報	1922 - 1996年	日本国公開実用新案公報	1971 - 2020年	日本国実用新案登録公報	1996 - 2020年	日本国登録実用新案公報	1994 - 2020年				
日本国実用新案公報	1922 - 1996年													
日本国公開実用新案公報	1971 - 2020年													
日本国実用新案登録公報	1996 - 2020年													
日本国登録実用新案公報	1994 - 2020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）														
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												
X	US 2018/0174025 A1 (SK HYNIX INC.) 21.06.2018 (2018 - 06 - 21) [0044]-[0045], [0092], [0108], Figs.11-12	1, 5, 8-10												
X	JP 7-302249 A (株式会社日立製作所) 14.11.1995 (1995 - 11 - 14) [0021], [0033]-[0038], [0046]-[0051]	1, 5, 8-10												
A	JP 2018-129033 A (アクシス アーバー) 16.08.2018 (2018 - 08 - 16) [0003]	1-10												
A	JP 2017-182319 A (株式会社メガチップス) 05.10.2017 (2017 - 10 - 05) [0050], [0053], [0057], [0066], [0075]-[0076]	2-4												
☐ 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。														
<table border="0"> <tr> <td>* 引用文献のカテゴリー</td> <td>"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抵触するものではなく、発明の原理又は理論の理解のために引用するもの</td> </tr> <tr> <td>"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</td> <td>"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</td> </tr> <tr> <td>"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</td> <td>"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</td> </tr> <tr> <td>"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）</td> <td>"&" 同一パテントファミリー文献</td> </tr> <tr> <td>"O" 口頭による開示、使用、展示等に言及する文献</td> <td></td> </tr> <tr> <td>"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</td> <td></td> </tr> </table>			* 引用文献のカテゴリー	"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抵触するものではなく、発明の原理又は理論の理解のために引用するもの	"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	"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	"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	"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	"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）	"&" 同一パテントファミリー文献	"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	
* 引用文献のカテゴリー	"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抵触するものではなく、発明の原理又は理論の理解のために引用するもの													
"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	"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													
"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	"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													
"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）	"&" 同一パテントファミリー文献													
"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														
国際調査を完了した日 16.06.2020	国際調査報告の発送日 23.06.2020													
名称及びあて先 日本国特許庁(ISA/JP) 〒100-8915 日本国 東京都千代田区霞が関三丁目4番3号	権限のある職員（特許庁審査官） 渡部 博樹 5B 3868 電話番号 03-3581-1101 内線 3545													

国際調査報告
 パテントファミリーに関する情報

国際出願番号

PCT/JP2020/016684

引用文献			公表日	パテントファミリー文献	公表日
US	2018/0174025	A1	21.06.2018	KR 10-2018-0070194 A [0028]-[0029], [0079], [0094], Figs. 11-12	
JP	7-302249	A	14.11.1995	EP 0583217 A2 P.3 Line 58 - P.4 Line 5, P.6 Line 1 - P.6 Line 32, P.7 Line 2 - P.7 Line 19	
JP	2018-129033	A	16.08.2018	US 2018/0181867 A1 [0003]	
JP	2017-182319	A	05.10.2017	(ファミリーなし)	