

(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7月23日(23.07.2020)



(10) 国際公開番号

WO 2020/149178 A1

(51) 国際特許分類:

G06N 3/08 (2006.01)

(21) 国際出願番号: PCT/JP2020/000231

(22) 国際出願日: 2020年1月8日(08.01.2020)

(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 2019-006660 2019年1月18日(18.01.2019) 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 2 5 2 0 番地 Ibaraki (JP).

(72) 発明者: 伊藤 浩朗(ITO Hiroaki); 〒3128503 茨城県ひたちなか市高場 2 5 2 0 番地 日立オ

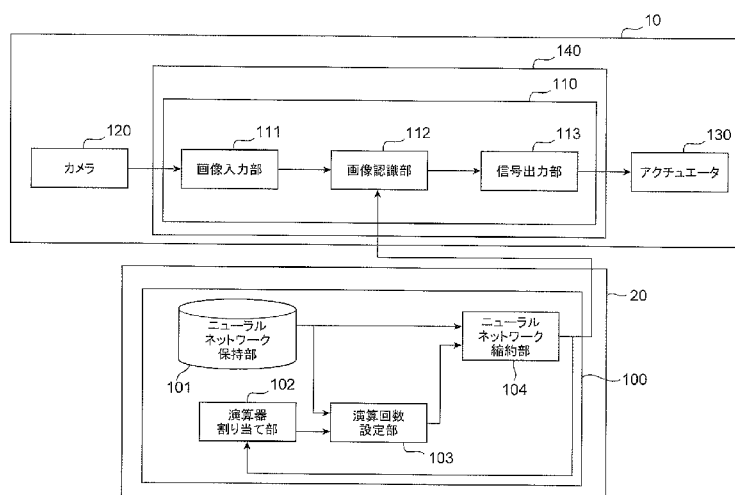
ートモティブシステムズ株式会社内 Ibaraki (JP). 小野 豪一(ONO Goichi); 〒1008280 東京都千代田区丸の内一丁目 6 番 6 号 株式会社日立製作所内 Tokyo (JP). 平井 理宇(HIRAI Riu); 〒1008280 東京都千代田区丸の内一丁目 6 番 6 号 株式会社日立製作所内 Tokyo (JP).

(74) 代理人: 戸田 裕二(TODA Yuji); 〒1008220 東京都千代田区丸の内一丁目 6 番 1 号 株式会社日立製作所内 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, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY,

(54) Title: NEURAL NETWORK COMPRESSION DEVICE

(54) 発明の名称: ニューラルネットワークの縮約装置



101 Neural network retaining unit
102 Computation unit allocation unit
103 Computation number setting unit
104 Neural network compression unit
111 Image input unit
112 Image recognition unit
113 Signal output unit
120 Camera
130 Actuator

(57) Abstract: The present invention makes it possible to shorten neural network processing time and reduce the number of neural network computations such that a plurality of computation units can be used effectively. Provided is a neural network compression device (100) which, by compressing a neural network, reduces the number of neural network computations performed by a computation device (140) comprising a plurality of computation units, said compression device (100) configured to comprise: a computation unit allocation unit (102) which sets an allocation quantity of computation units to a computation process of a neural network; a computation number setting unit (103) which, on the basis of the allocation quantity of the computation units, sets a number of computations of the neural network after compression thereof; and a neural network compression unit (104) which compresses the neural network such that the number of computations of the neural network by the computation device (140) is equal to the number of computations set by the computation number setting unit (103).

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))

(57) 要約：ニューラルネットワークの処理時間を短縮すると共に複数の演算器を有効利用できるようにニューラルネットワークの演算回数を低減する。複数の演算器を備えた演算装置（140）によるニューラルネットワークの演算回数を、ニューラルネットワークを縮約することによって低減するニューラルネットワークの縮約装置（100）であり、ニューラルネットワークの演算処理に対する演算器の割り当て数を設定する演算器割り当て部（102）と、演算器の割り当て数に基づいて縮約後のニューラルネットワークの演算回数を設定する演算回数設定部（103）と、演算装置（140）によるニューラルネットワークの演算回数が、演算回数設定部（103）によって設定された演算回数と等しくなるように、ニューラルネットワークを縮約するニューラルネットワーク縮約部（104）とを備える構成にした。

明 細 書

発明の名称：ニューラルネットワークの縮約装置

技術分野

[0001] 本発明はニューラルネットワークの縮約装置に関する。

背景技術

[0002] 近年、機械学習を用いた物体認識や行動予測を車両の自動運転に適用する技術が進展している。物体認識等に使用される機械学習手法として、ニューラルネットワークが知られている。ニューラルネットワークでは演算回数が多い畳み込み演算等が実施され、車両に搭載される演算装置がリアルタイムに処理を行う際の課題になっていた。上記の課題に対し、ニューラルネットワークの各ノードの重み係数が閾値以下となる枝（層間結合）を削減することによって、認識性能に対する影響を抑制しながらニューラルネットワークの演算負荷を軽減する縮約方法（圧縮方法）が開示されている（非特許文献1参照）。

先行技術文献

非特許文献

[0003] 非特許文献1：Song Han、Jeff Pool、John Tran、William J. Dally、“Learning both Weights and Connections for Efficient Neural Networks”、[online]、2015年10月30日、[平成30年12月24日検索]、インターネット<URL :<https://arxiv.org/pdf/1506.02626.pdf>>

発明の概要

発明が解決しようとする課題

[0004] 非特許文献1にはニューラルネットワークの一部の枝を削減することにより、ニューラルネットワークの演算回数を低減する方法が示されているが、演算回数が低減されても演算処理の処理時間が短縮されとは限られない。また、ニューラルネットワークの演算処理に対して、演算器に設けられた複数の演算器が有効に利用されない場合がある。このような問題は、車載用の

演算装置に限らず、ニューラルネットワークを利用した他の演算装置でも生じる可能性がある。

[0005] 本発明は前記課題を解決するもので、その目的とするところは、ニューラルネットワークの処理時間を短縮すると共に複数の演算器を有効利用できるようにニューラルネットワークの演算回数を低減することができるニューラルネットワークの縮約装置を提供することである。

課題を解決するための手段

[0006] 本発明の一態様のニューラルネットワークの縮約装置は、複数の演算器を備えた演算装置によるニューラルネットワークの演算回数を、ニューラルネットワークを縮約することによって低減するニューラルネットワークの縮約装置であって、ニューラルネットワークの演算処理に対する前記演算器の割り当て数を設定する演算器割り当て部と、前記演算器の割り当て数に基づいて縮約後のニューラルネットワークの演算回数を設定する演算回数設定部と、前記演算装置によるニューラルネットワークの演算回数が、前記演算回数設定部によって設定された演算回数と等しくなるように、ニューラルネットワークを縮約するニューラルネットワーク縮約部とを備えたことを特徴とする。

発明の効果

[0007] 本発明によれば、ニューラルネットワークの処理時間を短縮すると共に複数の演算器を有効利用できるようにニューラルネットワークの演算回数を低減することができる。本発明に関連する更なる特徴は、本明細書の記述、添付図面から明らかになるものである。また、上記した以外の、課題、構成及び効果は、以下の実施形態の説明により明らかにされる。

図面の簡単な説明

[0008] [図1]第1の実施形態に係る自動運転システムのブロック図。

[図2]第1の実施形態に係るニューラルネットワークの一例を示す図。

[図3]第1の実施形態に係る第1の中間層のノード処理の一例を示す図。

[図4]第1の実施形態に係る第1の中間層のノード処理の他の一例を示す図。

[図5]第1の実施形態に係る第2の中間層のノード処理の一例を示す図。

[図6]第1の実施形態に係る出力層のノード処理の一例を示す図。

[図7]第1の実施形態に係る縮約処理の一例を示す図。

[図8]第1の実施形態に係る演算装置の一部を示すブロック図。

[図9]ニューラルネットワークの演算処理のタイミングチャートの一例を示す図。

[図10]ニューラルネットワークの演算処理のタイミングチャートの一例を示す図。

[図11]ニューラルネットワークの演算処理のタイミングチャートの一例を示す図。

[図12]ニューラルネットワークの演算処理のタイミングチャートの一例を示す図。

[図13]第2の実施形態に係る自動運転システムのブロック図。

[図14]第3の実施形態に係る自動運転システムのブロック図。

[図15]第4の実施形態に係る自動運転システムのブロック図。

発明を実施するための形態

[0009] [第1の実施形態]

以下、添付の図面を参照して、ニューラルネットワークの縮約装置を適用した自動運転システムについて説明する。図1は、第1の実施形態に係る自動運転システムのブロック図である。なお、以下の説明では、ニューラルネットワークの縮約装置を自動運転システムに適用した一例について説明するが、ニューラルネットワークの縮約装置は、監視カメラシステム、医療用の画像診断システム、音声認識システム等の他のシステムに適用可能である。

[0010] 図1に示すように、自動運転システムは、ニューラルネットワークを用いて自動運転する車両10と、車両10に対してニューラルネットワークのパラメータを定期的に通知するサーバ20とが無線通信によって接続されている。車両10には、車両周辺を撮像するカメラ120と、車両10の自動運転を制御する車両制御部110と、アクセル、ブレーキ、操舵等の各種操作

に用いるアクチュエータ１３０とが設けられている。サーバ２０には、縮約技術によってニューラルネットワークの演算回数を低減するニューラルネットワークの縮約装置１００が実装されている。

[0011] 車両制御部１１０には、カメラ１２０から入力された画像を所定形式に変換する画像入力部１１１と、ニューラルネットワークを用いて画像を認識する画像認識部１１２と、認識結果に応じた制御信号をアクチュエータ１３０に出力する信号出力部１１３とが設けられている。画像認識部１１２には、ニューラルネットワークの縮約装置１００から縮約後のニューラルネットワークのパラメータが通知される。ニューラルネットワークのパラメータとは、画像認識部１１２において縮約後のニューラルネットワークを構築するためのパラメータであり、例えばニューラルネットワークで使用される重み係数等である。

[0012] 画像認識部１１２は、ニューラルネットワークの縮約装置１００から通知されたパラメータに基づいて縮約後のニューラルネットワークを構築して、ニューラルネットワークを用いて画像入力部１１１から入力された画像を認識する。信号出力部１１３は、画像の認識結果に基づいて自車両の走行方向及び走行速度等の自動運転に必要な車両制御情報を含む制御信号をアクチュエータ１３０に出力する。アクチュエータ１３０は、アクセル用の駆動アクチュエータ、ブレーキ用の制動アクチュエータ、ステアリング用の操舵アクチュエータ等を含んでいる。アクチュエータ１３０は、信号出力部１１３からの制御信号に基づいて車両１０を制御する。

[0013] 車両制御部１１０は、複数の演算器を備えた演算装置１４０によって実現されている。車両制御部１１０の各種処理に対して演算装置１４０の演算器が割り当てられている。演算装置１４０の複数の演算器によって車両制御部１１０の各種処理が実行されることで、ニューラルネットワークを用いた画像認識が実施されて、認識結果に基づいて車両１０の自動運転制御が実施される。なお、車両制御部１１０の各種処理のうち、画像認識処理は演算負荷が大きいため、車両制御部１１０の他の処理と比較して多数の演算器が画像

認識処理に対して割り当てられている。

[0014] ニューラルネットワークの縮約装置 100 には、ニューラルネットワーク保持部 101 と、演算器割り当て部 102 と、演算回数設定部 103 と、ニューラルネットワーク縮約部 104 とが設けられている。ニューラルネットワーク保持部 101 には、PC、サーバ等において学習された後の学習済みのニューラルネットワーク、すなわち推論処理に用いるニューラルネットワークが保持されている。演算器割り当て部 102 は、画像認識部 112 によってニューラルネットワークを用いて画像認識を行う際に、ニューラルネットワークの演算処理に対する演算装置 140 の演算器の割り当て数を設定する。この場合、演算器割り当て部 102 は、入力デバイス等を介してユーザから入力された演算器の割り当て数を縮約装置 100 に設定している。

[0015] 演算回数設定部 103 は、ニューラルネットワーク保持部 101 に保持されたニューラルネットワークの演算処理に必要な演算回数を算出し、演算器の割り当て数に基づいて縮約後のニューラルネットワークの演算回数を設定する。ニューラルネットワーク縮約部 104 は、演算装置 140 によるニューラルネットワークの演算回数が、演算回数設定部 103 によって設定された演算回数と等しくなるように、ニューラルネットワーク保持部 101 に保持されたニューラルネットワークを縮約する。また、ニューラルネットワーク縮約部 104 は、縮約後のニューラルネットワークのパラメータを車両 10 の画像認識部 112 に通知して、画像認識部 112 の画像認識処理に用いられるニューラルネットワークの演算回数を低減している。

[0016] なお、ニューラルネットワークの縮約装置 100 の各部は、プロセッサを用いてソフトウェアによって実現されてもよいし、集積回路等に形成された論理回路（ハードウェア）によって実現されてもよい。プロセッサを用いる場合には、プロセッサがメモリに格納されているプログラムを読み出して実行することで各種処理が実施される。プロセッサとしては、例えば、CPU (Central Processing Unit)、MPU (Micro Processing Unit)、GPU (Graphics Processing Unit) 等が使用される。また、メモリは、用

途に応じてROM(Read Only Memory)、RAM(Random Access Memory)等の一つ又は複数の記録媒体によって構成されている。

[0017] 以下、ニューラルネットワークの縮約装置100について具体例を用いて説明する。図2は、第1の実施形態に係るニューラルネットワークの一例を示す図である。図3は、第1の実施形態に係る第1の中間層のノード処理の一例を示す図である。図4は、第1の実施形態に係る第1の中間層のノード処理の他の一例を示す図である。図5は、第1の実施形態に係る第2の中間層のノード処理の一例を示す図である。図6は、第1の実施形態に係る出力層のノード処理の一例を示す図である。図7は、第1の実施形態に係る縮約処理の一例を示す図である。また、以下の説明では、説明の便宜上、図1の符号を適宜使用して説明する。

[0018] 図2に示すように、ニューラルネットワークは、入力層200と、第1の中間層201と、第2の中間層202と、出力層203とによって構成されている。ニューラルネットワークは、入力層200に5つの入力値 $X_0 - X_4$ が入力されると、第1、第2の中間層201、202を経由して出力層203から2つの出力値 Y_0 、 Y_1 を出力する。入力層200には入力要素として5つの入力ノード $N_{x0} - N_{x4}$ が設けられ、出力層203には出力要素として2つの出力ノード N_{y0} 、 N_{y1} が設けられている。第1の中間層201には演算要素としてノード $N_{o0} - N_{o3}$ が設けられ、第2の中間層202にはノード $N_{i0} - N_{i3}$ が設けられている。

[0019] 第1の中間層201のノード $N_{o0} - N_{o3}$ は、入力層200から入力された入力値 $X_0 - X_4$ に対して層間結合(枝)毎に設定された重み係数を乗算し、各乗算結果を加算した結果を第2の中間層202のノード $N_{i0} - N_{i3}$ に出力する。第2の中間層202のノード $N_{i0} - N_{i3}$ は、第1の中間層201のノード $N_{o0} - N_{o3}$ から入力された値に対して層間結合毎に設定された重み係数を乗算し、各乗算結果を加算した結果を出力層203の出力ノード N_{y0} 、 N_{y1} に出力する。出力層203の出力ノード N_{y0} 、 N_{y1} は、第2の中間層202のノード $N_{i0} - N_{i3}$ から入力された値に対して層間結合毎に設定された重み

係数を乗算し、各乗算結果を加算した結果を出力値 Y_0 、 Y_1 として出力する。

[0020] ここで、各ノードの演算処理について詳細に説明する。図3は、第1の中間層201のノード N_{00} の演算処理を示している。図3において、第1の中間層201のノード N_{00} では入力層200の入力ノード $N_{x0} - N_{x4}$ とノード N_{00} の層間の重み係数 $W_{000} - W_{040}$ が使用され、ノード N_{00} から出力される出力値 Z_{00} は次式(1)によって求められる。

$$Z_{00} = X_0 * W_{000} + X_1 * W_{010} + X_2 * W_{020} + X_3 * W_{030} + X_4 * W_{040} \dots (1)$$

[0021] 図4は、第1の中間層201のノード N_{01} の演算処理を示している。図4において、第1の中間層201のノード N_{01} では入力層200の入力ノード $N_{x0} - N_{x4}$ とノード N_{01} の層間の重み係数 $W_{001} - W_{041}$ が使用され、ノード N_{01} から出力される出力値 Z_{01} は次式(2)によって求められる。第1の中間層201の他のノード N_{02} 、 N_{03} についても同様にして出力値 Z_{02} 、 Z_{03} を出力する。

$$Z_{01} = X_0 * W_{001} + X_1 * W_{011} + X_2 * W_{021} + X_3 * W_{031} + X_4 * W_{041} \dots (2)$$

[0022] 図5は、第2の中間層202のノード N_{10} の演算処理を示している。図5において、第2の中間層202のノード N_{10} では第1の中間層201のノード $N_{00} - N_{03}$ とノード N_{10} の層間の重み係数 $W_{100} - W_{130}$ が使用され、ノード N_{10} から出力される出力値 Z_{11} は次式(3)によって求められる。第2の中間層202の他のノード $N_{11} - N_{13}$ についても同様にして出力値 $Z_{11} - Z_{13}$ を出力する。

$$Z_{11} = Z_{00} * W_{100} + Z_{01} * W_{110} + Z_{02} * W_{120} + Z_{03} * W_{130} \dots (3)$$

[0023] 図6は、出力層203の出力ノード N_{Y0} の演算処理を示している。図6において、出力層203の出力ノード N_{Y0} では第2の中間層202のノード $N_{10}-N_{13}$ と出力ノード N_{Y0} の層間の重み係数 $W_{200}-W_{230}$ が使用され、出力ノード N_{Y0} から出力される出力値 Y_0 は次式(4)によって求められる。出力層203の他の出力ノード N_{Y1} についても同様にして出力値 Y_1 を出力する。

$$Y_0 = Z_{10} * W_{200} + Z_{11} * W_{210} + Z_{12} * W_{220} + Z_{13} * W_{230} \cdots \quad (4)$$

[0024] 本実施形態のニューラルネットワークは、自動運転用の画像認識部112に用いられるものである。画像入力部111から入力層200の入力ノード $N_{X0}-N_{X4}$ に画像の入力値 X_0-X_4 が入力されて、出力層203の出力ノード N_{Y0} 、 N_{Y1} から画像認識結果である出力値 Y_0 、 Y_1 が出力される。例えば、画像が他車両である確率が Y_0 として出力され、画像が歩行者である確率が Y_1 として出力される。画像が他車両である確率が高い場合には車両制御部110によって自車両が他車両に追従するように制御され、画像が歩行者である確率が高い場合には自車両が歩行者を避けるように制御される。このように、ニューラルネットワークは、入力値から所望の出力値を得るためのネットワーク構成及び重み係数等から成り、ニューラルネットワーク保持部101にはネットワーク構成及び重み係数等の各種情報が記憶されている。

[0025] 次に、ニューラルネットワークの縮約装置100の具体的な動作について説明する。演算器割り当て部102は、ニューラルネットワークの演算処理、すなわち画像認識部112の画像認識処理に対する演算器の割り当て数を設定する。この場合、車両10に搭載された演算装置140の複数の演算器のうち、車両制御部110の各部の処理負荷を考慮して画像認識処理に対する演算器の割り当て数が設定される。本実施形態の自動運転システム等の組み込みシステムに用いられる演算装置140としてはFPGA(Field Programmable Gate Array)が挙げられる。FPGAは所望の処理を論理回路の形式によって設計及び実行できる演算装置140である。FPGAには積和演

算を高速に処理することが可能なDSP (Digital Signal Processor) と呼ばれる演算器が搭載されており、ニューラルネットワークの乗算及び加算を高速に処理するのに有用である。DSPはFPGA毎に搭載数が予め決められているため、演算器割り当て部102によってニューラルネットワークの演算処理に用いるDSP数が設定される。なお、ニューラルネットワークの演算回数は、一般にFPGAのDSP数よりもはるかに多く、ニューラルネットワーク全体を一度に処理することは困難である。このため、ニューラルネットワークの演算処理が複数回に分けて実施される。また、演算装置140は、FPGAに限定されるものではなく、ニューラルネットワークに対して使用可能な複数の積和演算器を備えたものであればよい。

[0026] 演算回数設定部103は、縮約前のニューラルネットワークの演算回数を算出する。一例として、図2に示したニューラルネットワークの演算処理に必要な演算回数について説明する。第1の中間層201の各ノードの演算処理には、上記の式(1)に示すように5回の乗算と4回の加算の計9回の演算が必要となる。第1の中間層201にはノード N_{00} – N_{03} まで4つのノードがあるため、第1の中間層201の演算処理に必要な演算回数は $9 \times 4 = 36$ 回となる。同様に、第2の中間層202の各ノードの演算処理には、上記の式(3)に示すように4回の乗算と3回の加算の計7回の演算が必要となる。第2の中間層202にはノード N_{10} – N_{13} まで4つのノードがあるため、第2の中間層202の演算処理に必要な演算回数は $7 \times 4 = 28$ 回となる。出力層203の各出力ノードの演算処理には、上記の式(4)に示すように4回の乗算と3回の加算の計7回の演算が必要となる。出力層203には出力ノード N_{y0} 、 N_{y1} の2つのノードがあるため、出力層203の演算処理に必要な演算回数は $7 \times 2 = 14$ 回となる。以上の各層の演算回数を合計すると、図2に示すニューラルネットワークの演算処理に必要な演算回数は次式(5)によって示される。

$$36\text{回} + 28\text{回} + 14\text{回} = 78\text{回} \cdots (5)$$

[0027] 実際には使用されるニューラルネットワークには演算回数が100万回を超える大規模なニューラルネットワークもあるが、本実施形態では説明の便宜上、ニューラルネットワークを簡略化して説明している。

[0028] 次に、演算回数設定部103は、縮約前のニューラルネットワークの演算回数よりも少なく、かつ演算器割り当て部102によって設定された演算器の割り当て数の整数倍になるように縮約後のニューラルネットワークの演算回数を設定する。上記の式(5)によって求めたように、縮約前のニューラルネットワークの演算回数は78回である。ここでは、一例として演算器の割り当て数を20個とすると、縮約後のニューラルネットワークの演算回数は次式(6)によって設定される。なお、式(6)の $\text{mod}(A, B)$ はAをBによって除算した時の剰余を算出する処理を示している。

$$78 - \text{mod}(78, 20) = 60 \cdots (6)$$

[0029] ニューラルネットワーク縮約部104は、縮約前のニューラルネットワークの演算回数78回と演算回数設定部103によって設定された演算回数60回との差分である18回分の演算を削減するようにニューラルネットワークを縮約する。縮約前のニューラルネットワークの演算回数については、ニューラルネットワーク縮約部104が演算回数設定部103から取得してもよいし、ニューラルネットワーク縮約部104が演算回数設定部103と同様に算出してもよい。ニューラルネットワークの縮約方法は、特に限定されないが、本実施形態ではPruning処理を例示して説明する。Pruning処理は、重み係数の絶対値が予め定められた閾値未満の場合には、出力値 Z_{oo} に対する影響が十分小さいとして、入力値に重み係数を乗算する乗算処理を削減するものである。ニューラルネットワークの重み係数の絶対値が小さい重み係数から順に削減することでニューラルネットワークが縮約される。

[0030] 例えば、図3に示すように、第1の中間層201のノード N_{oo} の出力値 Z_{oo} を得るためには、上記の式(1)に示すように5回の乗算と4回の加算の計

9回の演算が必要となる。このとき、図7に示すように、重み係数 W_{010} 、 W_{040} の絶対値が閾値未満の場合、入力値 X_1 と重み係数 W_{010} の乗算処理と、入力値 X_4 と重み係数 W_{040} の乗算処理が削減される。この結果、第1の中間層201のノード N_{00} の出力値 Z_{00} は次式(7)によって求められる。式(7)では、ノード N_{00} の演算回数が3回の乗算と2回の加算の計5回になっており、式(1)と比べて演算回数が4回削減されている。このように、Pruning処理は、出力に対する影響が小さなノード間の結合を削除することで演算回数を低減する。

$$Z_{00} = X_0 * W_{000} + X_2 * W_{020} + X_3 * W_{030} \cdots (7)$$

[0031] ニューラルネットワークの演算処理の処理時間について説明する。図8は、第1の実施形態に係る演算装置の一部を示すブロック図である。図9は、ニューラルネットワークの演算処理のタイミングチャートの一例を示す図である。図10は、ニューラルネットワークの演算処理のタイミングチャートの他の一例を示す図である。図11は、ニューラルネットワークの演算処理のタイミングチャートの他の一例を示す図である。

[0032] 図8に示すように、メモリ301には演算器302によって演算された中間データ303が格納される。また、演算器302によって入力値304とメモリ301に格納された中間データ305とを用いて演算処理が繰り返される。演算器302によって全ての演算処理が終了すると、演算器302から演算結果306が出力される。演算器302の割り当て数が20個、縮約後のニューラルネットワークに必要な演算回数が60回とすると、1個の演算器302につき3サイクルの演算処理が実施された後に演算結果306が出力される。図9に示すように、各演算器302が1サイクル(1回)の演算処理に要する処理時間をTとすると、ニューラルネットワーク全体として3サイクル分の処理時間Tで60回(#0-#59)の演算処理が終了する。なお、処理時間Tは1個の演算器302による1サイクル分の処理時間を示しているが、処理時間Tで20個の演算器302によって並列に演算処理

が実施されている。

[0033] 図10は、演算器302の割り当て数を考慮せずにニューラルネットワークの演算回数を低減した場合を示している。ここでは、一例として、演算器302の割り当て数を20個とし、ニューラルネットワークの演算回数を78回から62回に低減した場合について説明する。ニューラルネットワークの演算回数が60回の場合と比較して、演算回数が2回（#60、#61）増えただけであるが、ニューラルネットワーク全体として4サイクル分の処理時間Tが必要になって1サイクル分の処理時間Tが増える。

[0034] 同様に、図11は、演算器302の割り当て数を考慮せずにニューラルネットワークの演算回数を低減した場合を示している。ここでは、一例として、演算器302の割り当て数を20個とし、ニューラルネットワークの演算回数を78回から45回に低減した場合について説明する。1サイクル目の演算処理と2サイクル目の演算処理では20個の演算器302の全てが使用されているが、3サイクル目の演算処理では20個の演算器302のうち5個しか使用されていない。このため、ニューラルネットワークの演算回数が60回の場合と比較して、15回分の演算回数が低減されているにも関わらず、演算処理の処理時間が短縮されていない。特に、3サイクル目の演算処理で演算器302が有効利用されておらず、画像認識部112の認識性能が低下するおそれがある。

[0035] 以上のように、第1の実施形態の自動運転システムでは、カメラから入力した画像の画像認識に用いるニューラルネットワークの演算回数を、演算器割り当て部102によって設定された演算器302の割り当て数の整数倍になるように低減している。これにより、ニューラルネットワークの演算処理に割り当てられた複数の演算器302によって同時に演算処理が開始されて同時に演算処理が終了される。一部の演算器302だけで演算するサイクルが無くなり、演算処理の処理時間を短縮すると共に複数の演算器302の使用効率を最大限に高めることができる。

[0036] また、ニューラルネットワークの第1の中間層201から出力層203ま

での合計の演算回数を低減する構成に限定されず、ニューラルネットワークの層毎に演算回数が低減されてもよい。

[0037] 上記したように、図2に示すニューラルネットワークの第1の中間層201の演算回数は36回、第2の中間層202の演算回数は28回、出力層203の演算回数は14回である。この場合、演算器割り当て部102（図1参照）によってニューラルネットワークの層毎に演算器302の割り当て数が設定される。第1の中間層201に対する演算器302の割り当て数が8個、第2の中間層202に対する演算器302の割り当て数が8個、出力層203に対する演算器302の割り当て数が4個に設定される。

[0038] また、演算回数設定部103（図1参照）によって縮約後のニューラルネットワークの層毎に演算回数が設定される。第1の中間層201に対する演算回数は次式（8）によって設定され、第2の中間層202に対する演算回数は次式（9）によって設定され、出力層203に対する演算回数は次式（10）によって設定される。なお、式（8）－（10）には、各層に対して演算器302の割り当て数が個別に設定されているが、各層に同じ演算器302の割り当て数が設定されてもよい。

$$36 - \text{mod}(36, 8) = 32 \cdots (8)$$

$$28 - \text{mod}(28, 8) = 24 \cdots (9)$$

$$14 - \text{mod}(14, 4) = 12 \cdots (10)$$

[0039] ニューラルネットワークの層毎に演算回数が設定されると、ニューラルネットワーク縮約部104によってPruning等の縮約方法を用いてニューラルネットワークの層毎に演算回数が低減される。第1の中間層201における4回分の演算処理が削減され、第2の中間層202における4回分の演算処理が削減され、出力層203における2回分の演算処理が削減される。このように、演算回数設定部103によってニューラルネットワークの層毎に演算回数が設定されることで、比較的少ない演算器302の割り当て数（例えば、8個）によって演算処理できるため、演算装置140の実装規模

を小さくすることができる。

[0040] 演算回数設定部 103 は、ニューラルネットワークの演算処理に必要な演算回数を演算器 302 の割り当て数によって除算した剰余を、この演算回数から減算することで縮約後のニューラルネットワークの演算回数を設定する構成に限定されない。演算回数設定部 103 は、演算器 302 の割り当て数の整数倍を設定すればよい。例えば、演算回数設定部 103 は、次式 (11) を用いて演算回数を設定してもよい。

$$\text{低減後の演算回数} = 78 - \text{mod}(78, 20) - 20 \times n \cdots (11)$$

($n = 0, 1, 2 \cdots$ 但し、低減後の演算回数は正数)

[0041] $n = 0$ の場合には、式 (11) は式 (6) と同じになる。 $n = 1$ の場合には、次式 (12) によって演算回数が 40 回に設定される。演算回数が 40 回に設定されると、ニューラルネットワーク縮約部 104 によって、 $78 - 40 = 38$ 回分の演算を削減するように Pruning 処理が実施される。この場合、式 (6) から求められた演算回数よりも削減数が多い分、認識精度の低下を考慮する必要があるが、1 サイクル分の処理時間 T を短縮することができる。すなわち、 $n = 0$ の場合にはニューラルネットワークの演算処理に 3 サイクル分の処理時間 T が必要なのに対して、 $n = 1$ の場合にはニューラルネットワークの演算処理に 2 サイクル分の処理時間 T で処理を終えることができる。なお、 n の値はユーザによって任意に設定されてもよいし、予め設定されてもよい。

$$78 - \text{mod}(78, 20) - 20 = 40 \cdots (12)$$

[0042] 演算回数設定部 103 は、演算器 302 の割り当て数の整数倍になるように縮約後のニューラルネットワークの演算回数を設定する構成に限定されない。演算回数設定部 103 は、縮約後のニューラルネットワークの演算回数を、必ずしも演算器 302 の割り当て数の整数倍に設定しなくてもよい。演

算回数設定部 103 は、演算器 302 の割り当て数で、縮約前のニューラルネットワークの演算回数を除算した余りが演算器 302 の割り当て数の半分以上になるように、縮約後のニューラルネットワークの演算回数を設定してもよい。例えば、演算回数設定部 103 は、次式 (13) を用いて演算回数を設定してもよい。なお、 k は、演算器割り当て部 102 によって設定された演算器 302 の割り当て数の $1/2$ 以下である。

$$\begin{aligned} \text{低減後の演算回数} &= 78 - m \bmod (78, 20) - m \cdots (13) \\ (m &= 0, 1, 2, \cdots, k) \end{aligned}$$

[0043] 式 (13) では、低減後の演算回数は、 m の値によって変更される。例えば、 $m = 0$ の場合には演算回数が 60 回、 $m = 1$ の場合には演算回数が 59 回、 $m = k$ の場合には演算回数が $60 - k$ 回に設定される。図 12 に示すように、低減後の演算回数が 59 回の場合には、3 サイクル目の演算処理では 20 個の演算器 302 のうち 19 個の演算器 302 のみが使用され、1 個の演算器 302 は未使用になる。このため、演算器 302 の使用効率が低下するが、縮約後のニューラルネットワークの認識精度に応じて、ニューラルネットワークの縮約率を柔軟性に変更することができる。なお、 m の値はユーザによって任意に設定されてもよいし、予め設定されていてもよい。

[0044] [第 2 の実施形態]

以下に、第 2 の実施形態に係る自動運転システムについて説明する。第 2 の実施形態に係る自動運転システムは、ニューラルネットワークの認識精度を検証しながら、演算器の割り当て数の調整によってニューラルネットワークを縮約する点について第 1 の実施形態と相違している。したがって、第 1 の実施形態と同様な構成については説明を省略する。図 13 は、第 2 の実施形態に係る自動運転システムのブロック図である。図 13 では、第 1 の実施形態の自動運転システムと同一のブロックについては同一の符号を付している。

[0045] 図 13 に示すように、第 2 の実施形態に係るニューラルネットワークの縮

約装置 100 には、ニューラルネットワークの認識精度を検証する機能として、検証用データ保持部 106 と認識精度検証部 107 とが設けられている。検証用データ保持部 106 には、縮約後のニューラルネットワークの認識精度を検証するための検証用データとして検証値及び正解値が保存されている。認識精度検証部 107 は、縮約後のニューラルネットワークの認識精度を算出して、当該認識精度と予め定められた目標精度を比較することで認識精度を検証する。この場合、認識精度検証部 107 は、検証用データ保持部 106 から読み出した検証値を縮約後のニューラルネットワークに入力し、ニューラルネットワークから出力される出力値と検証用データ保持部 106 から読み出した正解値とを照らし合わせて認識精度を算出する。そして、認識精度検証部 107 は、認識精度と目標精度の比較することで、目標精度を基準として認識精度を検証する。

[0046] 認識精度検証部 107 に検証された認識精度が目標精度未満の場合には、認識精度検証部 107 から演算器割り当て部 102 に演算器の割り当て数の増加指令が通知される。演算器割り当て部 102 は、認識精度検証部 107 から演算器の割り当て数の増加指令を受信すると、演算器の割り当て数を 1 個増加して割り当て数を更新する。演算回数設定部 103 は、更新後の演算器の割り当て数に基づいて、縮約後のニューラルネットワークの演算回数を設定する。一方で、認識精度検証部 107 に検証された認識精度が目標精度以上の場合には、認識精度検証部 107 から演算器割り当て部 102 に演算器の割り当て数の減少指令が通知される。演算器割り当て部 102 は、認識精度検証部 107 から演算器の割り当て数の減少指令を受信すると、演算器の割り当て数を 1 個減少して割り当て数を更新する。演算回数設定部 103 は、更新後の演算器の割り当て数に基づいて、縮約後のニューラルネットワークの演算回数を設定する。

[0047] 一例として、第 1 の実施形態と同様に、図 2 に示すニューラルネットワークを用いて具体的に説明する。縮約前のニューラルネットワーク演算回数は 78 回である。演算器割り当て部 102 によって演算器の割り当て数が 20

個に設定されると、演算回数設定部 103 によって次式 (14) を用いて演算回数が 60 回に設定される。ニューラルネットワーク縮約部 104 によってニューラルネットワークの 18 回分の演算処理が削減されてニューラルネットワークが縮約される。

$$78 - \text{mod}(78, 20) = 60 \cdots (14)$$

[0048] 次に、認識精度検証部 107 によって縮約後のニューラルネットワークに対して認識精度が検証される。認識精度検証部 107 による認識精度が目標精度未満である場合には、演算器割り当て部 102 によって演算器の割り当て数が 1 個増加されて 21 個に更新される。演算器の割り当て数が更新されると、演算器割り当て部 102 によって次式 (15) を用いて演算回数が 63 回に再設定される。ニューラルネットワーク縮約部 104 によってニューラルネットワークの 15 回分の演算処理が削減されてニューラルネットワークが再び縮約される。ニューラルネットワークの演算回数が 60 回から 63 回に増加しているため認識精度の改善が期待できる。

$$78 - \text{mod}(78, 21) = 63 \cdots (15)$$

[0049] 次に、認識精度検証部 107 によって再縮約後のニューラルネットワークに対して認識精度が検証される。認識精度検証部 107 による認識精度が目標精度以上になった場合には、ニューラルネットワークのパラメータが画像認識部 112 に通知され、カメラ 120 から入力された画像に対する画像認識処理に用いられる。一方、認識精度検証部 107 による認識精度が目標精度未満になった場合には、再び演算器割り当て部 102 によって演算器の割り当て数が 1 個増加される。この処理を繰り返すことで、ニューラルネットワークの認識精度が目標精度に近づけられる。

[0050] なお、縮約後のニューラルネットワークの認識精度が不足している場合に、認識精度を改善するように認識精度検証部 107 から演算器割り当て部 102 に演算器の割り当て数の増加指令がフィードバックされる構成に限られ

ない。縮約後のニューラルネットワークの認識精度が必要以上に高い場合に、認識精度を目標精度まで下げるように認識精度検証部 107 から演算器割り当て部 102 に演算器の割り当て数の減少指令がフィードバックされてもよい。

[0051] 以上のように、第 2 の実施形態に係る自動運転システムでは、第 1 の実施形態と同様に、演算処理の処理時間を短縮すると共に複数の演算器の使用効率を高めることができる。

また、縮約後のニューラルネットワークの認識精度に応じて、ニューラルネットワークの演算処理に対する演算器の割り当て数を調整することができる。よって、ニューラルネットワークの認識精度が目標精度を維持できる程度に、ニューラルネットワークの演算回数を低減することができる。

[0052] [第 3 の実施形態]

以下に、第 3 の実施形態に係る自動運転システムについて説明する。第 3 の実施形態に係る自動運転システムは、ニューラルネットワークの認識精度を検証しながら、演算回数の調整によってニューラルネットワークを縮約する点について第 2 の実施形態と相違している。したがって、第 1 の実施形態と同様な構成については説明を省略する。図 14 は、第 3 の実施形態に係る自動運転システムのブロック図である。図 14 では、第 2 の実施形態の自動運転システムと同一のブロックについては同一の符号を付している。

[0053] 図 14 に示すように、第 3 の実施形態に係るニューラルネットワークの縮約装置 100 では、認識精度検証部 107 の検証結果に基づいて演算回数設定部 103 の演算回数を調整するように構成されている。すなわち、認識精度検証部 107 によって検証された認識精度が目標精度未満の場合には、認識精度検証部 107 から演算回数設定部 103 に演算回数の増加指令が通知される。演算回数設定部 103 は、認識精度検証部 107 から演算回数の増加指令を受信すると、縮約後のニューラルネットワークの演算回数を増加して演算回数を再設定する。一方で、認識精度検証部 107 によって検証された認識精度が目標精度以上の場合には、認識精度検証部 107 から演算回数

設定部 103 に演算回数の減少指令が通知される。演算回数設定部 103 は、認識精度検証部 107 から演算回数の減少指令を受信すると、縮約後のニューラルネットワークの演算回数を減少して演算回数を再設定する。

[0054] 一例として、第 1 の実施形態と同様に、図 2 のニューラルネットワークを用いて具体的に説明する。縮約前のニューラルネットワーク演算回数は 78 回である。演算器割り当て部 102 によって演算器の割り当て数が 20 個に設定されると、演算回数設定部 103 によって次式 (16) を用いて演算回数が設定される。一例として、 $n = 2$ が設定されていた場合には、演算回数設定部 103 によって演算回数が 20 回に設定される。ニューラルネットワーク縮約部 104 によってニューラルネットワークの 58 回分の演算処理が削減されてニューラルネットワークが縮約される。

$$\text{低減後の演算回数} = 78 - \text{mod}(78, 20) - 20 \times n \cdots (16)$$

($n = 0, 1, 2, \dots$ 。但し、低減後の演算回数は正数)

[0055] 次に、認識精度検証部 107 によって縮約後のニューラルネットワークに対して認識精度が検証される。認識精度検証部 107 による認識精度が目標精度未満である場合には、演算回数設定部 103 によって式 (16) の n の値が 1 つ減らされて $n = 1$ になり、演算回数が 40 回に再設定される。ニューラルネットワーク縮約部 104 によってニューラルネットワークの 38 回分の演算処理が削減されてニューラルネットワークが再び縮約される。ニューラルネットワークの演算回数が 20 回から 40 回に増加しているため認識精度の改善が期待できる。

[0056] 次に、認識精度検証部 107 によって再縮約後のニューラルネットワークに対して認識精度が検証される。認識精度検証部 107 による認識精度が目標精度以上になった場合には、ニューラルネットワークのパラメータが画像認識部 112 に通知され、カメラ 120 から入力された画像に対する画像認識処理に用いられる。一方、認識精度検証部 107 による認識精度が目標精

度未満になった場合には、再び演算回数設定部 103 によって式 (16) の n の値が 1 つ増加される。この処理を繰り返すことで、ニューラルネットワークの認識精度が目標精度に近づけられる。

[0057] なお、縮約後のニューラルネットワークの認識精度が不足している場合に、認識精度を改善するように認識精度検証部 107 から演算回数設定部 103 に演算回数の増加指令がフィードバックされる構成に限られない。縮約後のニューラルネットワークの認識精度が必要以上に高い場合に、認識精度を目標精度まで下げるように認識精度検証部 107 から演算回数設定部 103 に演算回数の減少指令がフィードバックされてもよい。

[0058] 以上のように、第 3 の実施形態に係る自動運転システムでは、第 2 の実施形態と同様に、演算処理の処理時間を短縮すると共に複数の演算器の使用効率を高めることができる。

また、縮約後のニューラルネットワークの認識精度に応じて、ニューラルネットワークの演算回数を調整することができる。よって、ニューラルネットワークの認識精度が目標精度を維持できる程度に、ニューラルネットワークの演算回数を低減することができる。

[0059] [第 4 の実施形態]

以下に、第 4 の実施形態に係る自動運転システムについて説明する。第 4 の実施形態に係る自動運転システムは、縮約後のニューラルネットワークの演算回数を、演算器の割り当て数の調整によって補正する点について第 1 の実施形態と相違している。したがって、第 1 の実施形態と同様な構成については説明を省略する。図 15 は、第 4 の実施形態に係る自動運転システムのブロック図である。図 15 では、第 1 の実施形態の自動運転システムと同一のブロックについては同一の符号を付している。

[0060] 第 4 の実施形態に係るニューラルネットワークの縮約装置 100 には、縮約後のニューラルネットワークの演算回数を補正する機能として演算回数補正部 108 が設けられている。演算回数補正部 108 は、演算器割り当て部 102 によって設定された演算器の割り当て数に基づいて縮約後のニューラ

ルネットワークの演算回数を補正する。この場合、演算回数補正部 108 は、縮約後のニューラルネットワークに対し、演算器割り当て部 102 によって設定された演算器の割り当て数の整数倍となるようにニューラルネットワークの演算回数を補正する。

[0061] 一例として、第 1 の実施形態と同様に、図 2 に示すニューラルネットワークを用いて具体的に説明する。縮約前のニューラルネットワーク演算回数は 78 回である。演算回数設定部 103 によって縮約後のニューラルネットワークの演算回数が 39 回に設定される。この縮約後のニューラルネットワークの演算回数は、縮約前のニューラルネットワークの演算回数の半分にした値である。ニューラルネットワーク縮約部 104 によって 39 回部の演算処理が削減されてニューラルネットワークが縮約される。また、演算器割り当て部 102 によって演算器の割り当て数が 20 個に設定されると、演算回数補正部 108 によって次式 (17) を用いて縮約後のニューラルネットワークの演算回数が補正される。

$$\text{補正後の演算回数} = (\text{div}(39, 20) + \text{INT}(\text{div}(\text{mod}(39, 20), 20) + 0.5)) * 20 \cdots (17)$$

[0062] 式 (17) の $\text{div}(A, B)$ は A を B によって除算する処理を示し、 $\text{INT}(X)$ は X の小数点以下を切り捨てする処理を示している。式 (17) の一例では、 $\text{div}(39, 20) = 1$ 、 $\text{INT}(\text{div}(\text{mod}(39, 20), 20) + 0.5) = 1$ であるため、式 (17) が $(1 + 1) * 20$ になって補正後の演算回数が 40 回に設定される。これにより、縮約後のニューラルネットワークの演算回数を、演算器の割り当て数の整数倍になるように補正することが可能となる。

[0063] 以上のように、本発明の第 4 の実施形態の自動運転システムでは、第 1 の実施形態と同様に、演算処理の処理時間を短縮すると共に複数の演算器の使用効率を高めることができる。また、最初にニューラルネットワークを縮約して、演算器の割り当て数を考慮しながら縮約後のニューラルネットワーク

の演算回数を補正することが可能である。

[0064] なお、上記した各実施形態では、画像認識処理に用いるニューラルネットワークの縮約装置について説明したが、この構成に限定されない。ニューラルネットワークの縮約装置は、音声認識処理等の他の認識処理の他、将来データの予測処理に用いるニューラルネットワークの縮約にも適用可能である。

[0065] また、上記した各実施形態では、ニューラルネットワークの縮約装置100がサーバ20に設けられる構成にしたが、ニューラルネットワークの縮約装置100が車両10に設けられていてもよい。

[0066] また、上記した各実施形態では、演算回数設定部103及びニューラルネットワーク縮約部104が、演算装置140による縮約前のニューラルネットワークの演算回数を算出する構成にしたが、この構成に限定されない。縮約前のニューラルネットワークの演算回数はニューラルネットワーク保持部101に予め保持されていてもよい。また、ニューラルネットワークの縮約装置100に縮約前のニューラルネットワークの演算回を算出する演算回数算出部が設けられていてもよい。

[0067] また、上記した各実施形態では、演算器割り当て部102がニューラルネットワークの演算処理に割り当てる演算器302の個数を設定する構成にしたが、この構成に限定されない。演算器割り当て部102は、演算器302の全体数のうちニューラルネットワークの演算処理に割り当てる割合を設定してもよい。

[0068] また、上記した第4の実施形態では、演算回数設定部103がニューラルネットワークの演算回数を設定する構成にしたが、この構成に限定されない。演算回数設定部103は、ニューラルネットワークの演算回数の代わりに、ニューラルネットワークの縮約率を設定してもよい。

[0069] また、上記した第4の実施形態では、演算回数補正部108が演算器の割り当て数の整数倍となるように縮約後のニューラルネットワークの演算回数を補正する構成にしたが、この構成に限定されない。演算回数補正部108

は、ニューラルネットワークの層毎に演算回数を補正してもよいし、演算器の割り当て数によって除算した余りが演算器の割り当て数の半分以上になるようにニューラルネットワークの演算回数を補正してもよい。

[0070] 以上の通り、本実施形態に記載のニューラルネットワークの縮約装置（１００）は、複数の演算器（３０２）を備えた演算装置（１４０）によるニューラルネットワークの演算回数を、ニューラルネットワークを縮約することによって低減するニューラルネットワークの縮約装置（１００）であって、ニューラルネットワークの演算処理に対する演算器（３０２）の割り当て数を設定する演算器割り当て部（１０２）と、演算器（３０２）の割り当て数に基づいて縮約後のニューラルネットワークの演算回数を設定する演算回数設定部（１０３）と、演算装置（１４０）によるニューラルネットワークの演算回数が、演算回数設定部（１０３）によって設定された演算回数と等しくなるようにニューラルネットワークを縮約するニューラルネットワーク縮約部（１０４）とを備えている。

[0071] この構成によれば、ニューラルネットワークの演算処理に対する演算器（３０２）の割り当て数を考慮してニューラルネットワークの演算処理に必要な演算回数が低減される。すなわち、ニューラルネットワークの演算処理に割り当てられた複数の演算器（３０２）によって同時に処理可能な１サイクル当たりの演算回数が考慮される。よって、少数の演算器（３０２）だけで演算されるサイクルを無くすようにニューラルネットワークの演算回数を低減して、演算処理の処理時間を短縮すると共に複数の演算器（３０２）の使用効率を高めることができる。

[0072] 本実施形態に記載のニューラルネットワークの縮約装置（１００）において、演算回数設定部（１０３）は、縮約前のニューラルネットワークの演算回数よりも小さく、かつ演算器割り当て部（１０２）によって設定された演算器（３０２）の割り当て数の整数倍に縮約後のニューラルネットワークの演算回数を設定する。この構成によれば、ニューラルネットワークの演算処理に割り当てられた複数の演算器（３０２）によって同時に演算処理が開始

されて同時に演算処理が終了される。よって、一部の演算器（３０２）だけで演算するサイクルが無くなり、演算処理の処理時間を短縮すると共に複数の演算器（３０２）の使用効率を最大限に高めることができる。

[0073] 本実施形態に記載のニューラルネットワークの縮約装置（１００）において、演算回数設定部（１０３）は、演算器割り当て部（１０２）によって設定された演算器（３０２）の割り当て数で、ニューラルネットワークの演算回数を除算した余りが当該演算器（３０２）の割り当て数の半分以上になるように、縮約後のニューラルネットワークの演算回数を設定する。この構成によれば、ニューラルネットワークの演算処理に割り当てられた複数の演算器（３０２）によって同時に演算処理が開始されて、複数の演算器（３０２）のうち半分以上の演算器（３０２）によって同時に演算処理が終了される。よって、少数の演算器（３０２）だけで演算するサイクルが無くなり、演算処理の処理時間を短縮すると共に複数の演算器（３０２）の使用効率を高めることができる。

[0074] 本実施形態に記載のニューラルネットワークの縮約装置（１００）において、ニューラルネットワークは複数の層を有しており、演算器割り当て部（１０２）はニューラルネットワークの層毎に演算器（３０２）の割り当て数を設定し、演算回数設定部（１０３）はニューラルネットワークの層毎に縮約後のニューラルネットワークの演算回数を設定する。この構成によれば、比較的少ない演算器によって演算処理することができるため、演算装置（１４０）の実装規模を小さくすることができる。

[0075] 本実施形態に記載のニューラルネットワークの縮約装置（１００）において、ニューラルネットワーク縮約部（１０４）はPruning処理によってニューラルネットワークを縮約する。この構成によれば、Pruning処理によってニューラルネットワークの精度の低下を抑えつつ、ニューラルネットワークの演算回数を低減することができる。

[0076] 本実施形態に記載のニューラルネットワークの縮約装置（１００）において、縮約後のニューラルネットワークの精度を算出して、当該精度と目標精

度を比較する精度検証部（認識精度検証部１０７）を備え、精度が目標精度以上の場合に演算器割り当て部（１０２）が演算器（３０２）の割り当て数を小さく設定し、精度が目標精度未満の場合に演算器割り当て部（１０２）が演算器（３０２）の割り当て数を大きく設定する。この構成によれば、縮約後のニューラルネットワークの精度に応じて、ニューラルネットワークの演算処理に対する演算器（３０２）の割り当て数を調整することができる。よって、ニューラルネットワークの精度が目標精度を維持できる程度に、ニューラルネットワークの演算回数を低減することができる。

[0077] 本実施形態に記載のニューラルネットワークの縮約装置（１００）において、縮約後のニューラルネットワークの精度を算出して、当該精度と目標精度を比較する精度検証部（認識精度検証部１０７）を備え、精度が目標精度以上の場合に演算回数設定部（１０３）が縮約後のニューラルネットワークの演算回数を小さく設定し、精度が目標精度未満の場合に演算回数設定部（１０３）が縮約後のニューラルネットワークの演算回数を大きく設定する。この構成によれば、縮約後のニューラルネットワークの精度に応じて、ニューラルネットワークの演算回数を調整することができる。よって、ニューラルネットワークの精度が目標精度を維持できる程度に、ニューラルネットワークの演算回数を低減することができる。

[0078] 本実施形態に記載のニューラルネットワークの縮約装置（１００）において、複数の演算器（３０２）を備えた演算装置（１４０）によるニューラルネットワークの演算回数を、ニューラルネットワークを縮約することによって低減するニューラルネットワークの縮約装置（１００）であって、縮約後のニューラルネットワークの演算回数を設定する演算回数設定部（１０３）と、演算装置（１４０）によるニューラルネットワークの演算回数が、演算回数設定部（１０３）によって設定された演算回数と等しくなるように、ニューラルネットワークを縮約するニューラルネットワーク縮約部（１０４）と、ニューラルネットワークの演算処理に対する演算器（３０２）の割り当て数を設定する演算器割り当て部（１０２）と、演算器（３０２）の割り当

て数に基づいて縮約後のニューラルネットワークの演算回数を補正する演算回数補正部（１０８）とを備えている。

[0079] この構成によれば、ニューラルネットワークの演算処理に対する演算器（３０２）の割り当て数を考慮して、縮約後のニューラルネットワークの演算処理に必要な演算回数が補正される。すなわち、ニューラルネットワークの演算処理に割り当てられた複数の演算器（３０２）によって同時に処理可能な１サイクル当たりの演算回数が考慮される。よって、少数の演算器だけで演算されるサイクルを無くすようにニューラルネットワークの演算回数を補正して、演算処理の処理時間を短縮すると共に複数の演算器（３０２）の使用効率を高めることができる。

[0080] 以上、本発明の実施形態について詳述したが、本発明は、前記の実施形態に限定されるものではなく、特許請求の範囲に記載された本発明の精神を逸脱しない範囲で、種々の設計変更を行うことができるものである。例えば、前記した実施の形態は本発明を分かりやすく説明するために詳細に説明したものであり、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、ある実施形態の構成の一部を他の実施形態の構成に置き換えることが可能であり、また、ある実施形態の構成に他の実施形態の構成を加えることも可能である。さらに、各実施形態の構成の一部について、他の構成の追加・削除・置換をすることが可能である。

符号の説明

[0081] １００ ニューラルネットワークの縮約装置、１０２ 演算器割り当て部、
１０３ 演算回数設定部、１０４ ニューラルネットワーク縮約部、１０７
認識精度検証部（精度検証部）、１０８ 演算回数補正部、１４０ 演算
装置、３０２ 演算器

請求の範囲

- [請求項1] 複数の演算器を備えた演算装置によるニューラルネットワークの演算回数を、ニューラルネットワークを縮約することによって低減するニューラルネットワークの縮約装置であって、
- ニューラルネットワークの演算処理に対する前記演算器の割り当て数を設定する演算器割り当て部と、
- 前記演算器の割り当て数に基づいて縮約後のニューラルネットワークの演算回数を設定する演算回数設定部と、
- 前記演算装置によるニューラルネットワークの演算回数が、前記演算回数設定部によって設定された演算回数と等しくなるように、ニューラルネットワークを縮約するニューラルネットワーク縮約部とを備えたことを特徴とするニューラルネットワークの縮約装置。
- [請求項2] 前記演算回数設定部は、縮約前のニューラルネットワークの演算回数よりも小さく、かつ前記演算器割り当て部によって設定された前記演算器の割り当て数の整数倍に縮約後のニューラルネットワークの演算回数を設定することを特徴とする請求項1に記載のニューラルネットワークの縮約装置。
- [請求項3] 前記演算回数設定部は、前記演算器割り当て部によって設定された前記演算器の割り当て数で、縮約前のニューラルネットワークの演算回数を除算した余りが当該演算器の割り当て数の半分以上になるように、縮約後のニューラルネットワークの演算回数を設定することを特徴とする請求項1に記載のニューラルネットワークの縮約装置。
- [請求項4] ニューラルネットワークは複数の層を有しており、
- 前記演算器割り当て部はニューラルネットワークの層毎に前記演算器の割り当て数を設定し、
- 前記演算回数設定部はニューラルネットワークの層毎に縮約後のニューラルネットワークの演算回数を設定することを特徴とする請求項1に記載のニューラルネットワークの縮約装置。

[請求項5] 前記ニューラルネットワーク縮約部はPruning処理によってニューラルネットワークを縮約することを特徴とする請求項1に記載のニューラルネットワークの縮約装置。

[請求項6] 縮約後のニューラルネットワークの精度を算出して、当該精度と目標精度を比較する精度検証部を備え、

前記精度が目標精度以上の場合に前記演算器割り当て部が前記演算器の割り当て数を小さく設定し、前記精度が目標精度未満の場合に前記演算器割り当て部が前記演算器の割り当て数を大きく設定することを特徴とする請求項1から請求項5のいずれか一項に記載のニューラルネットワークの縮約装置。

[請求項7] 縮約後のニューラルネットワークの精度を算出して、当該精度と目標精度を比較する精度検証部を備え、

前記精度が目標精度以上の場合に前記演算回数設定部が縮約後のニューラルネットワークの演算回数を小さく設定し、前記精度が目標精度未満の場合に前記演算回数設定部が縮約後のニューラルネットワークの演算回数を大きく設定することを特徴とする請求項1から請求項5のいずれか一項に記載のニューラルネットワークの縮約装置。

[請求項8] 複数の演算器を備えた演算装置によるニューラルネットワークの演算回数を、ニューラルネットワークを縮約することによって低減するニューラルネットワークの縮約装置であって、

縮約後のニューラルネットワークの演算回数を設定する演算回数設定部と、

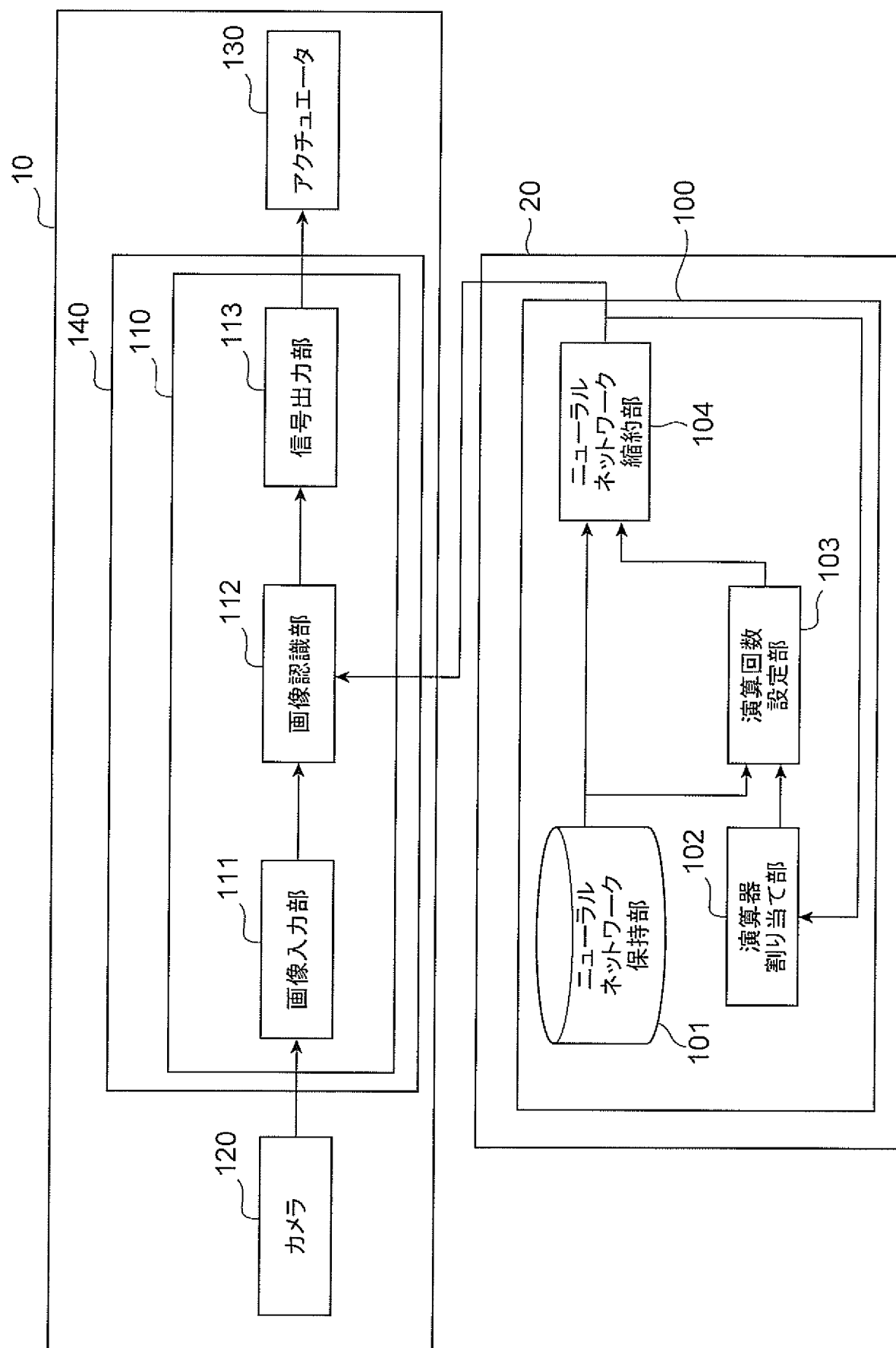
前記演算装置によるニューラルネットワークの演算回数が、前記演算回数設定部によって設定された演算回数と等しくなるように、ニューラルネットワークを縮約するニューラルネットワーク縮約部と、

ニューラルネットワークの演算処理に対する前記演算器の割り当て数を設定する演算器割り当て部と、

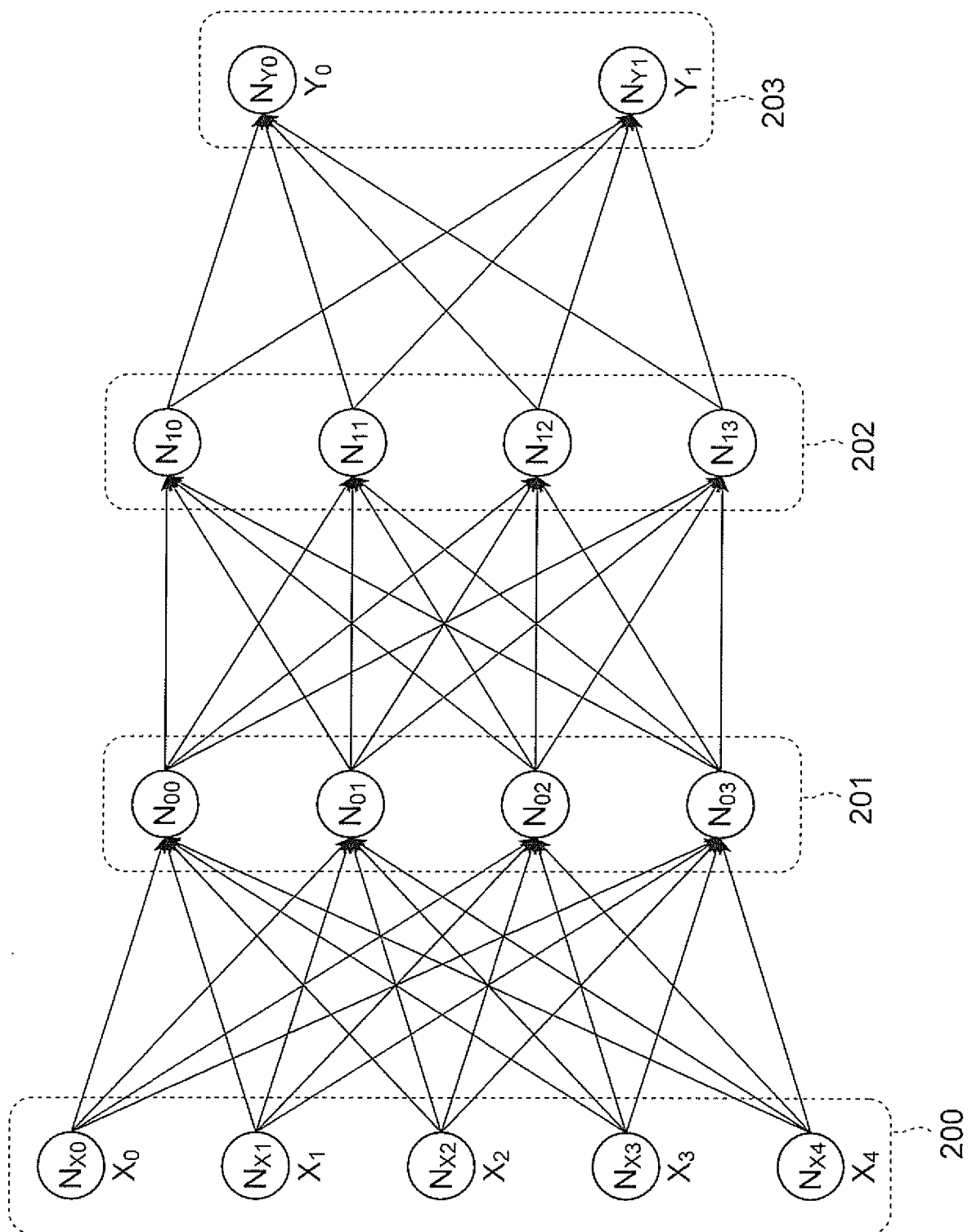
前記演算器の割り当て数に基づいて縮約後のニューラルネットワー

クの演算回数を補正する演算回数補正部とを備えたことを特徴とするニューラルネットワークの縮約装置。

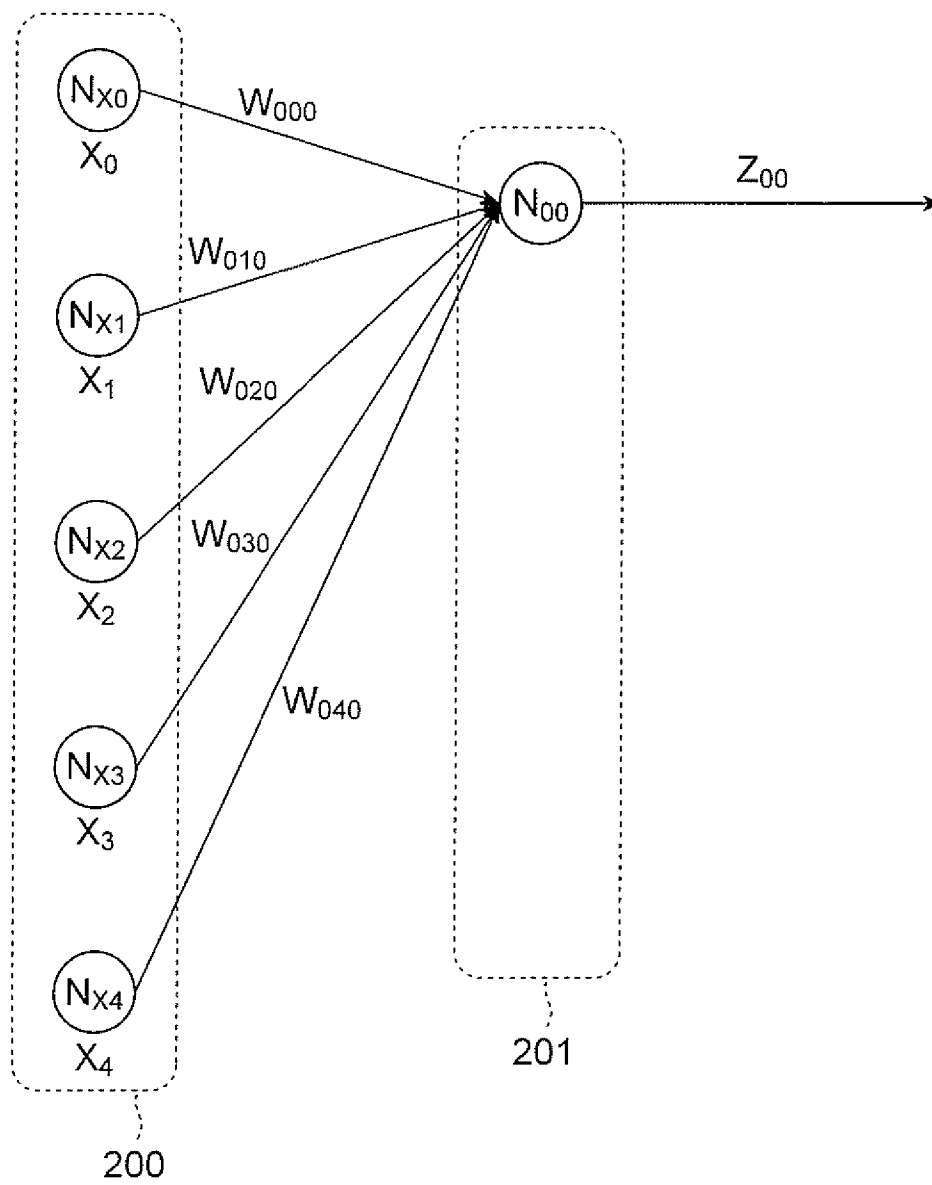
[図1]



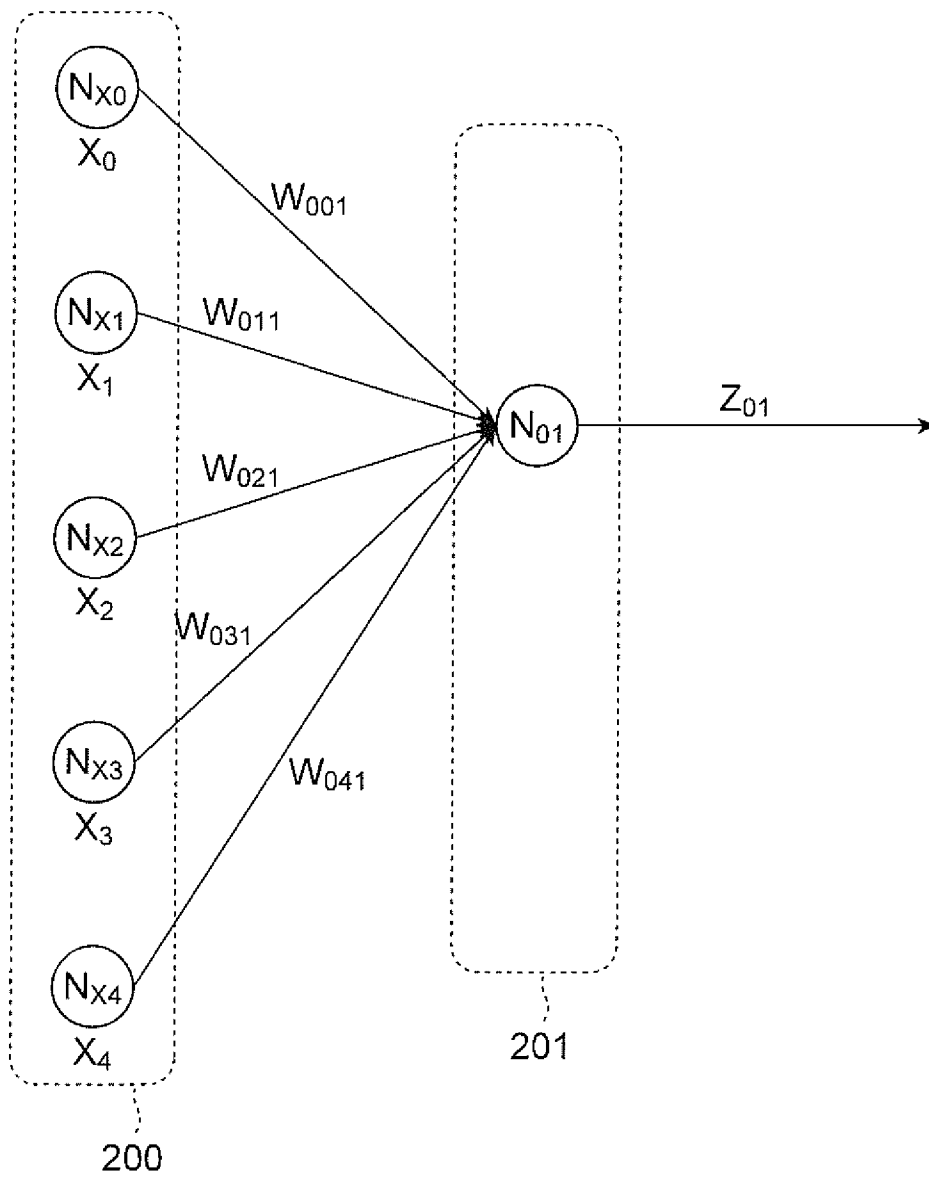
[図2]



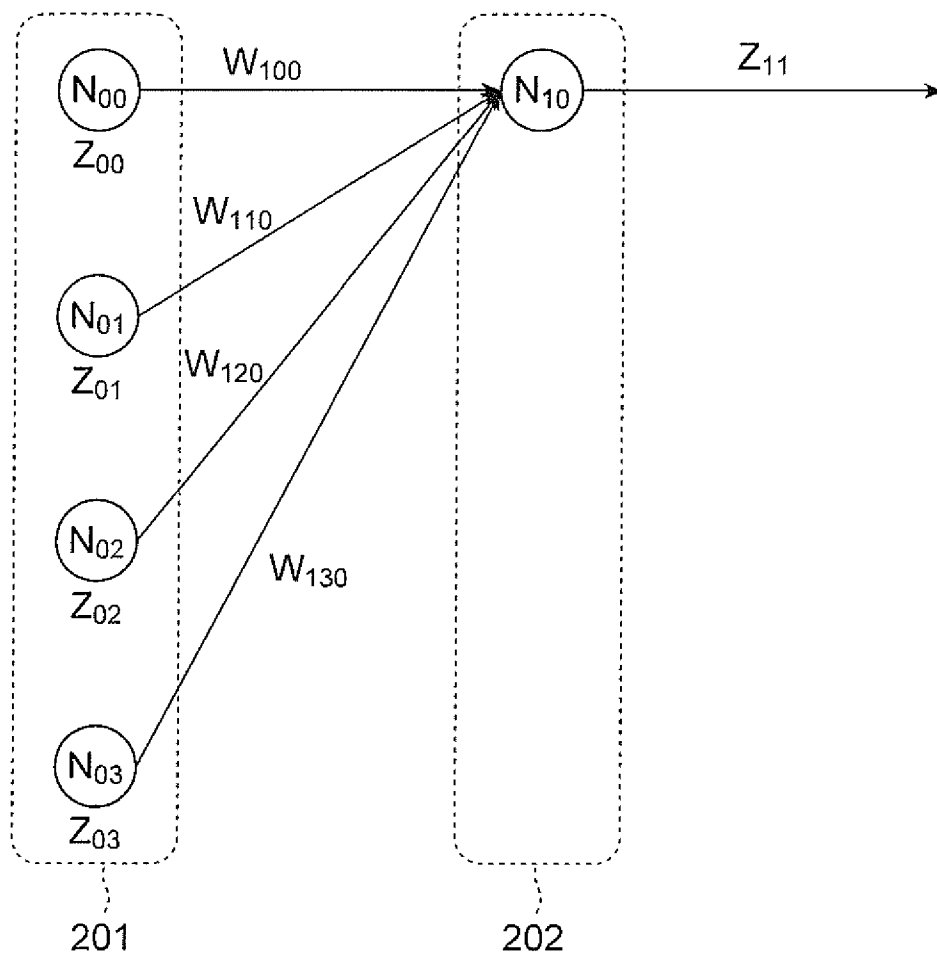
[図3]



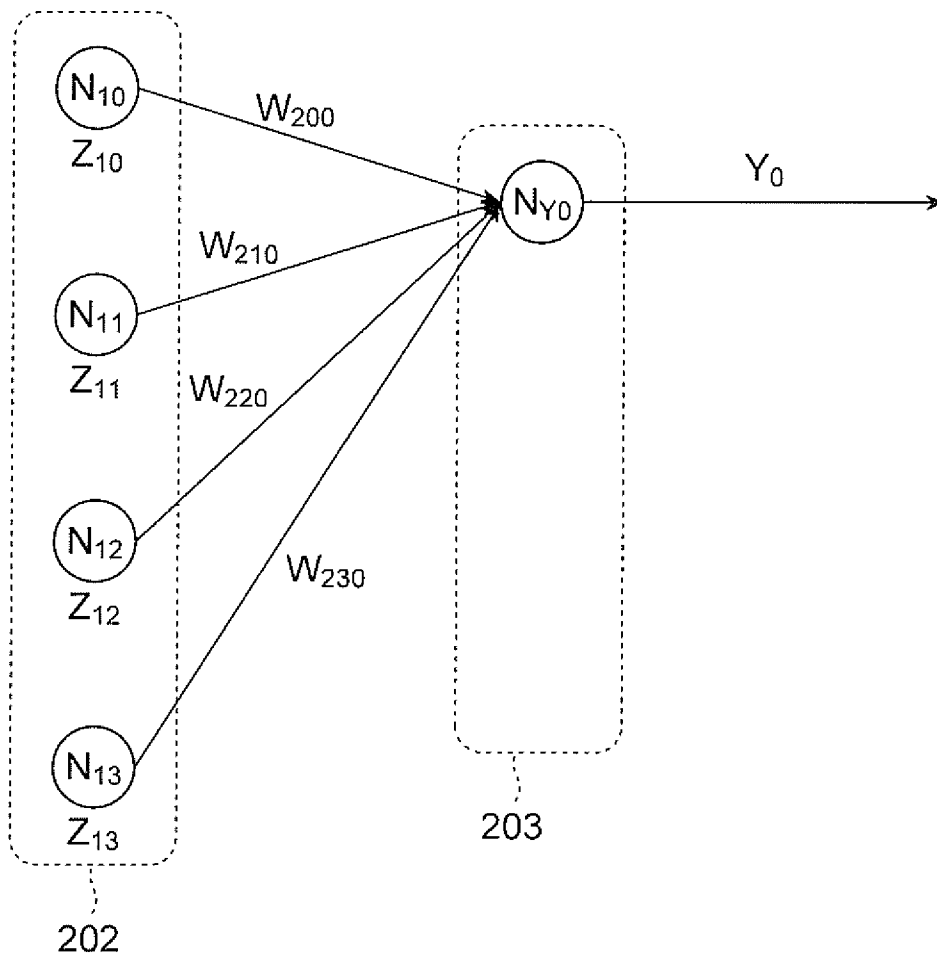
[図4]



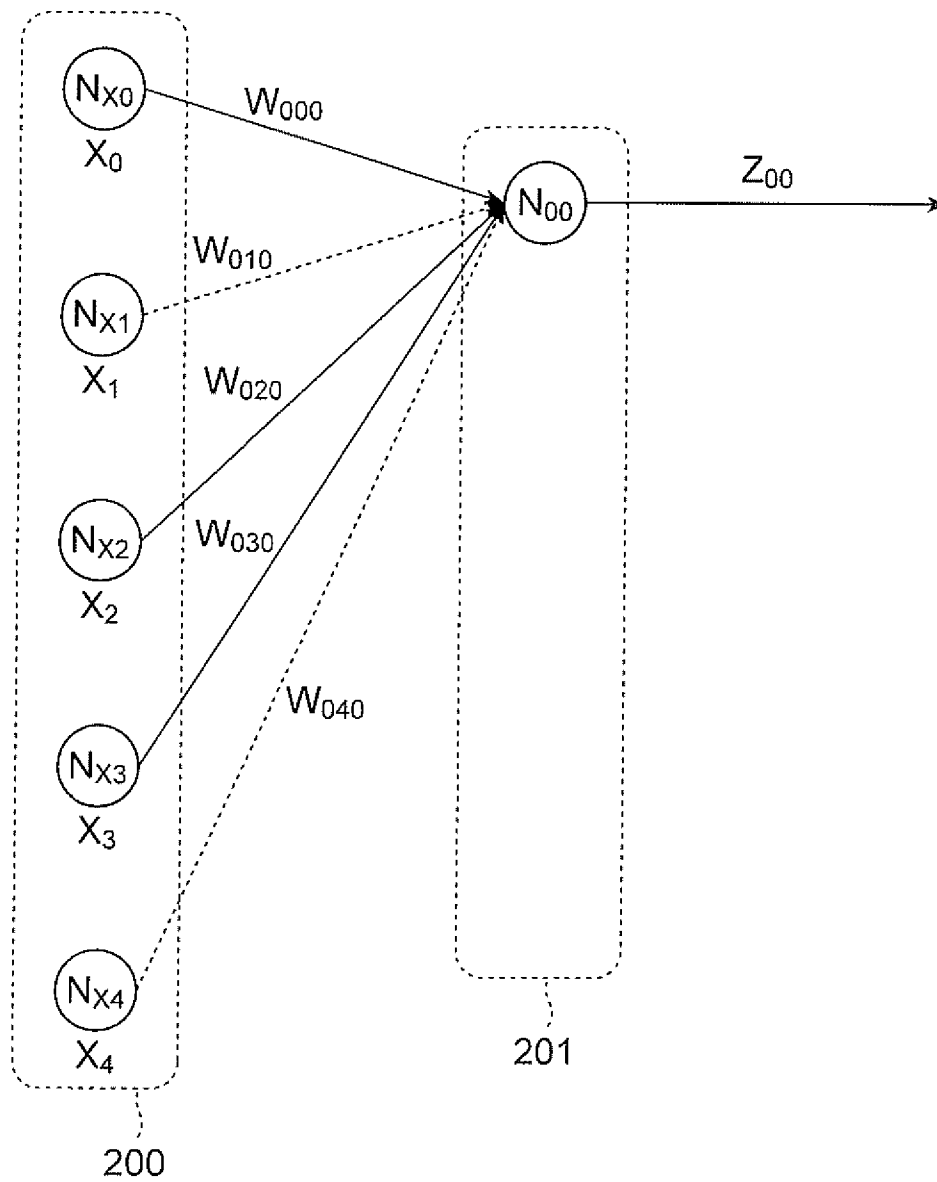
[図5]



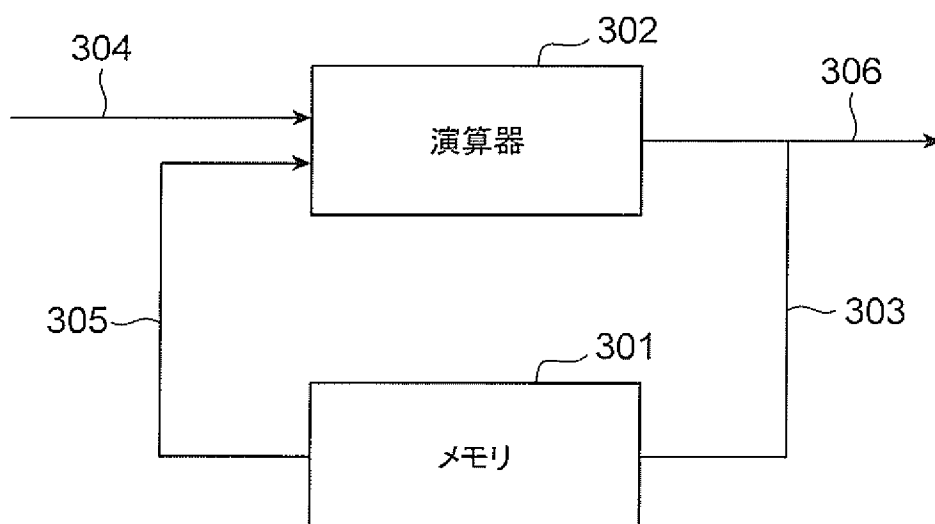
[図6]



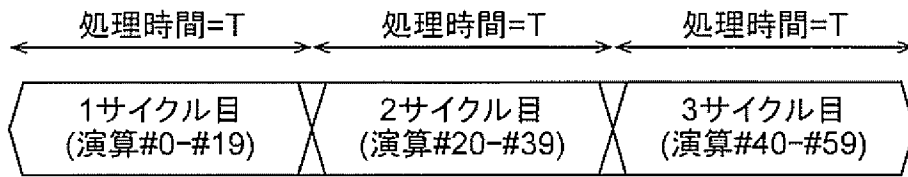
[図7]



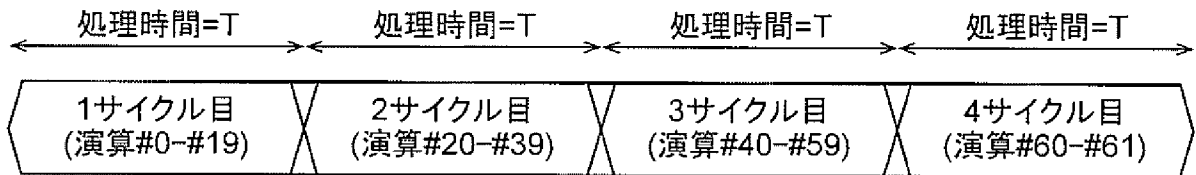
[図8]



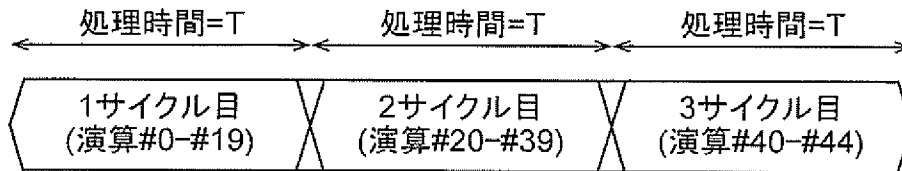
[図9]



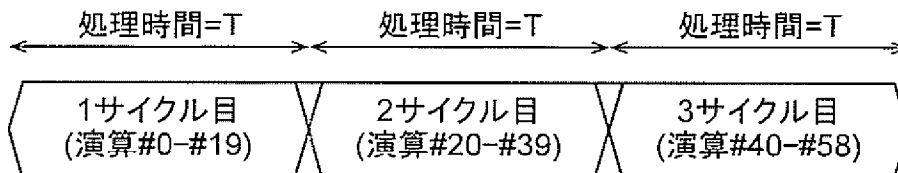
[図10]



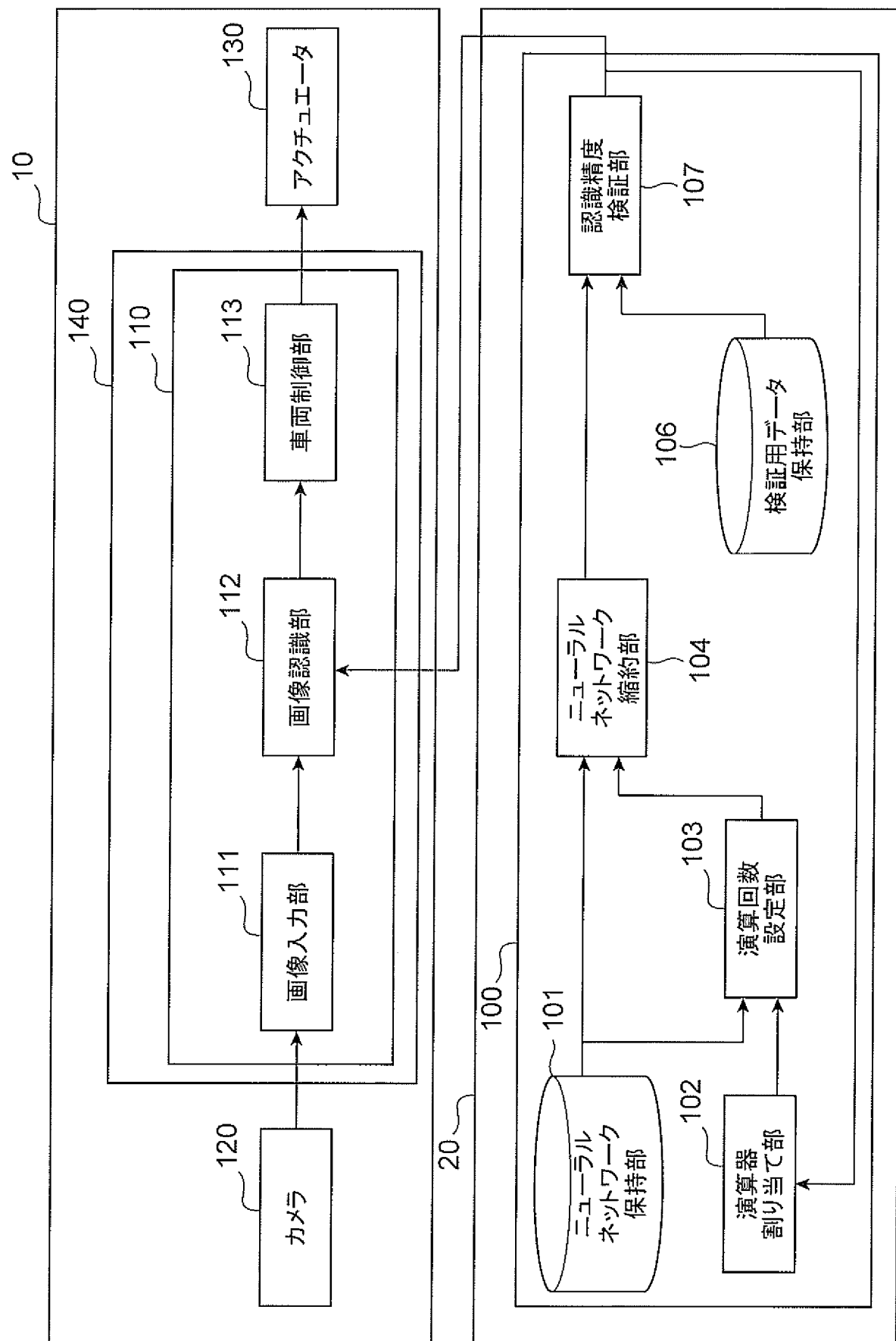
[図11]



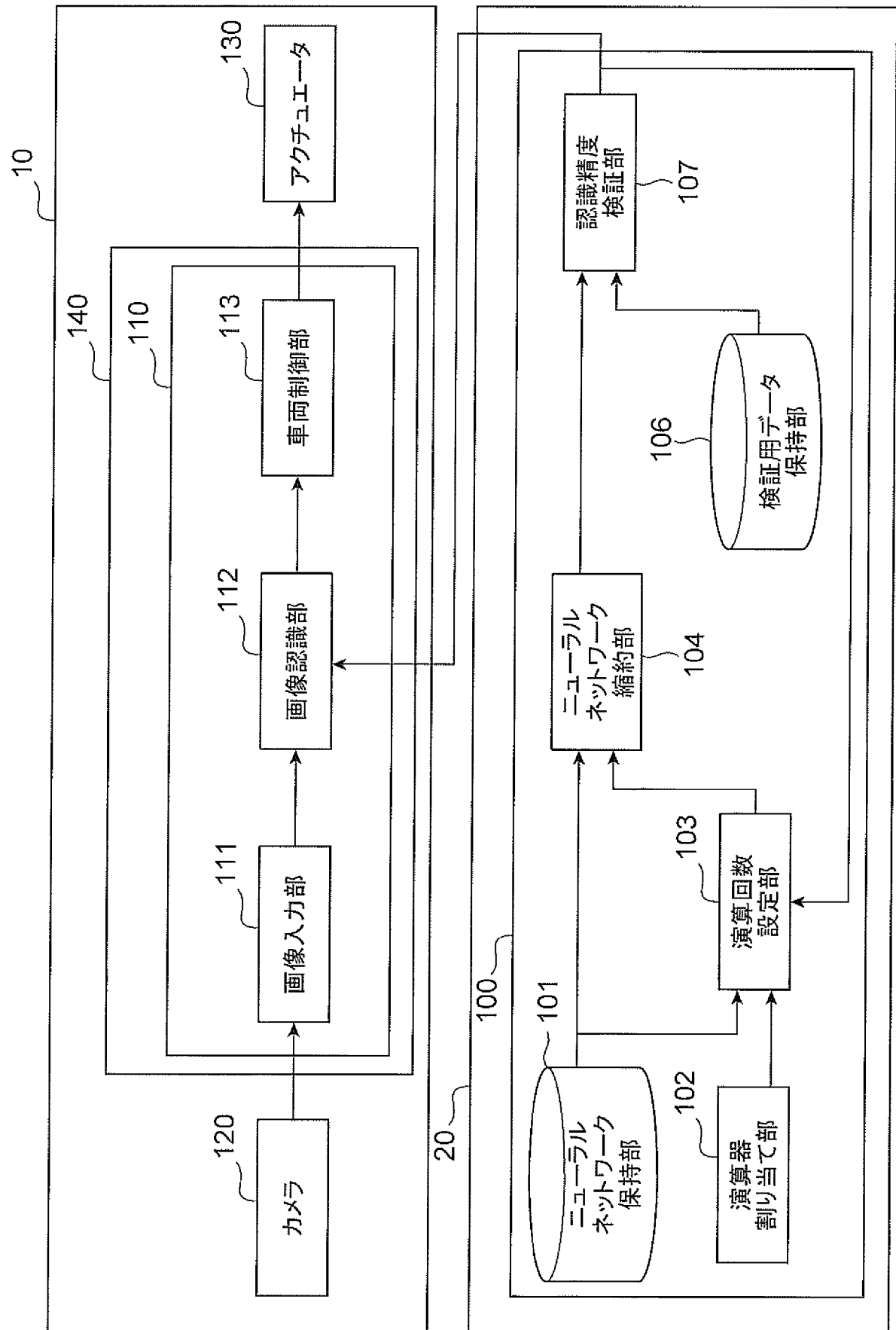
[図12]



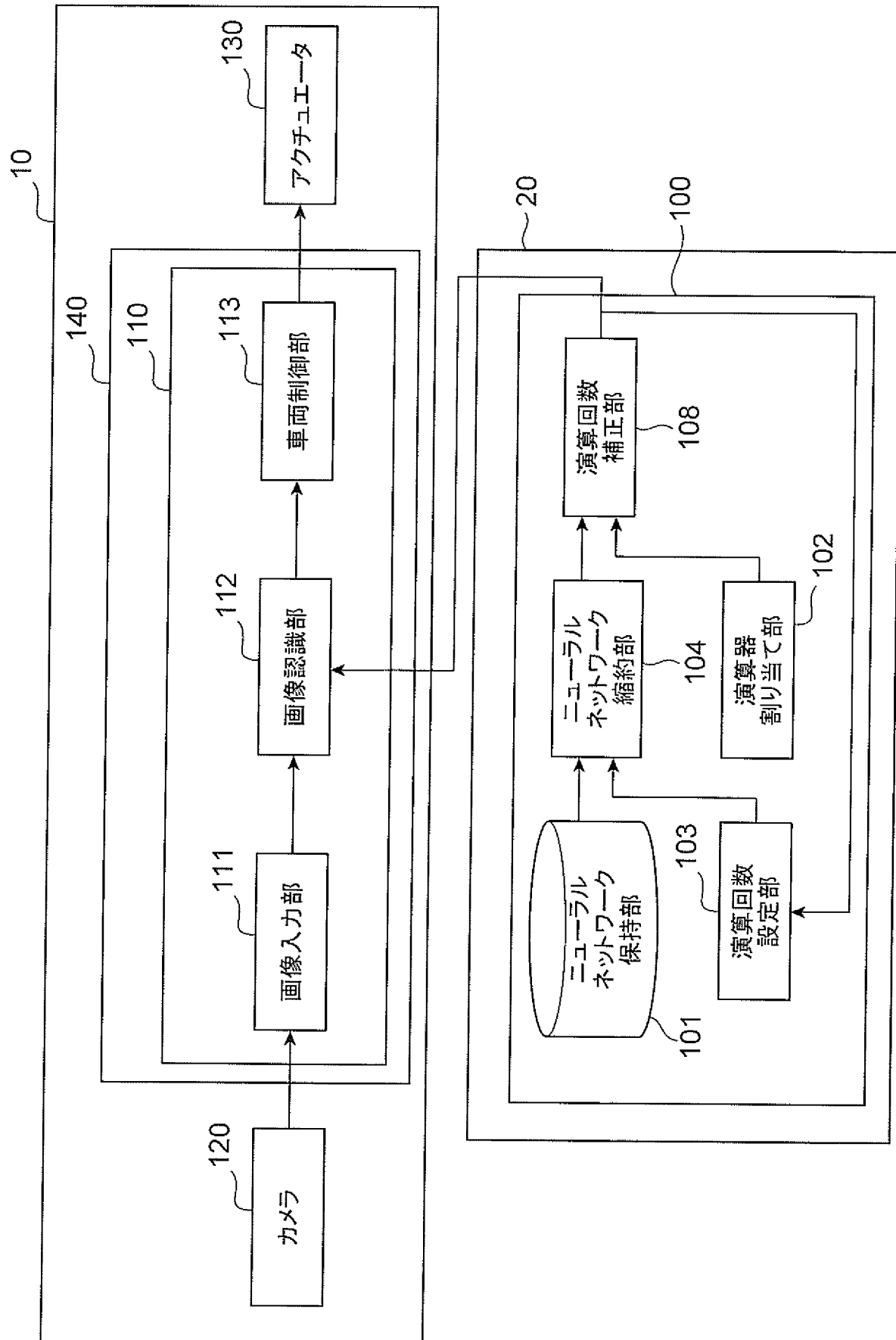
[図13]



[図14]



[図15]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2020/000231

A. CLASSIFICATION OF SUBJECT MATTER

Int.Cl. G06N3/08 (2006.01) i

FI: G06N3/08120

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. G06N3/00-99/00, G06F9/00-9/54

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.
A	児島隆生ほか, 自動運転の高度化を支える知能化技術, 日立評論, 20 October 2017, vol. 99, no. 5, pp. 52-56, ISSN 0367-5874, particularly, p. 55, 「4.1 ニューラルネットワークの縮約技術」, (KOJIMA, Takao, Hitachi Review), non-official translation (Intelligent technology to support the advancement of automated driving, "4.1 Contracting technology of the neural network")	1-8
A	JP 11-215380 A (RICOH CO., LTD.) 06.08.1999 (1999-08-06), paragraphs [0007]-[0009], fig. 1, 2	1-8



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

16.03.2020

Date of mailing of the international search report

24.03.2020

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

International application No.

PCT/JP2020/000231

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	吉田明正, 粗粒度タスク並列処理のための階層統合型実行制御手法, 情報処理学会論文誌, 15 December 2004, vol. 45, no. 12, pp. 2732-2740, ISSN 0387-5806, particularly, pp. 2733, 2734 「2.2 粗粒度タスク並列処理の実行方式」, p. 2738, 「6.1 Turb3d プログラムを用いた階層統合型実行制御と階層型実行制御の性能比較」, (YOSHIDA, Akimasa, Layer-unified execution control scheme for coarse grain task parallel processing, IPSJ Journal), non-official translation("2.2 Execution method of coarse grain task parallel processing", "6.1 Performance comparison between hierarchical integrated execution control and hierarchical execution control using Turb3d program")	1-8

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/JP2020/000231

JP 11-215380 A 06.08.1999 (Family: none)

A. 発明の属する分野の分類（国際特許分類（IPC）） G06N 3/08(2006.01)i FI: G06N3/08 120		
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）） G06N 3/00-99/00; G06F 9/00-9/54		
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 1971-2020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
A	児島隆生ほか、自動運転の高度化を支える知能化技術、日立評論、2017.10.20、第99巻、第5号、pp.52-56、ISSN 0367-5874 特にp.55「4.1 ニューラルネットワークの縮約技術」	1-8
A	JP 11-215380 A（株式会社リコー）06.08.1999（1999-08-06） 段落[0007]-[0009]、図1-2	1-8
A	吉田明正、粗粒度タスク並列処理のための階層統合型実行制御手法、情報処理学会論文誌、2004.12.15、第45巻、第12号、pp.2732-2740、ISSN 0387-5806 特にpp.2733-2734「2.2 粗粒度タスク並列処理の実行方式」、p.2738「6.1 Turb3dプログラムを用いた階層統合型実行制御と階層型実行制御の性能比較」	1-8
☐ 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 “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 16.03.2020	国際調査報告の発送日 24.03.2020	
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） 北元 健太 5B 3856 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/JP2020/000231

引用文献	公表日	パテントファミリー文献	公表日
JP 11-215380 A	06.08.1999	(ファミリーなし)	