(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
2021年5月6日(06.05.2021)

WIPO | PCT

(10) 国際公開番号
WO 2021/084717 A1

(51) 国際特許分類:
G06N 3/063 (2006.01) *G06F 17/16* (2006.01)

(21) 国際出願番号： PCT/JP2019/042927

(22) 国際出願日： 2019年10月31日(31.10.2019)

(25) 国際出願の言語： 日本語

(26) 国際公開の言語： 日本語

(71) 出願人: 日本電気株式会社 (**NEC CORPORATION**) [JP/JP]; 〒1088001 東京都港区芝五丁目７番１号 Tokyo (JP).

(72) 発明者: 竹中 崇 (**TAKENAKA Takashi**); 〒1088001 東京都港区芝五丁目７番１号 日本電気株式会社内 Tokyo (JP). 井上 浩明(**INOUE Hiroaki**); 〒1088001 東京都港区芝五丁目７番１号 日本電気株式会社内 Tokyo (JP).

(74) 代理人:岩壁 冬樹, 外(**IWAKABE Fuyuki** et al.); 〒1040031 東京都中央区京橋二丁目８番７号 読売八重洲ビル６階 サンライズ国際特許事務所 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, 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.

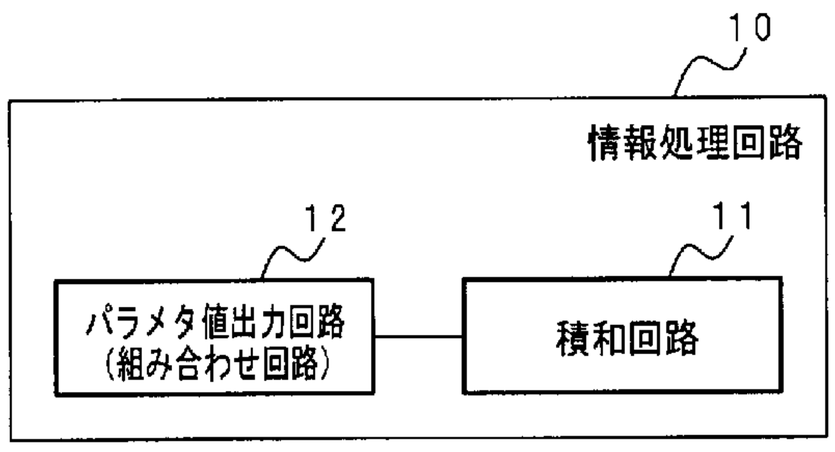
(84) 指定国(表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, 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：
— 国際調査報告（条約第21条(3)）

(54) Title: IINFORMATION PROCESSING CIRCUIT AND METHOD OF DESIGNING INFORMATION PROCESSING CIRCUIT

(54) 発明の名称： 情報処理回路および情報処理回路の設計方法

[図9]



10... INFORMATION PROCESSING CIRCUIT
11... PRODUCT SUM CIRCUIT
12... PARAMETER VALUE OUTPUT CIRCUIT (COMBINATIONAL CIRCUIT)

(57) Abstract: This information processing circuit 10 comprises: a product sum circuit 11 which executes a layered operation in deep learning, and performs a product sum operation by using input data and a parameter value; and a parameter value output circuit 12 which outputs the parameter value, wherein the parameter value output circuit 12 is configured from a combinational circuit.

(57) 要約：情報処理回路１０は、深層学習における層の演算を実行し、入力データとパラメタ値とを用いて積和演算を行う積和回路１１と、パラメタ値を出力するパラメタ値出力回路１２とを含み、パラメタ値出力回路１２は、組み合わせ回路で構成されている。

WO 2021/084717 A1

明　細　書

発明の名称 ： 情報処理回路および情報処理回路の設計方法

技術分野

[0001] 本発明は、深層学習の推論フェーズを実行する情報処理回路、およびそのような情報処理回路の設計方法に関する。

背景技術

[0002] 深層学習は、多層のニューラルネットワーク（以下、ネットワークという。）を使用するアルゴリズムである。深層学習では、各々のネットワーク（層）を最適化してモデル（学習モデル）を作成する学習フェーズと、学習モデルに基づいて推論が行われる推論フェーズとが実行される。なお、モデルは、推論モデルといわれることもある。また、以下、モデルを推論器と表現することがある。

[0003] 学習フェーズおよび推論フェーズにおいて、パラメタとしての重みを調整するための演算が実行されたり、入力データと重みとを対象とする演算が行われるが、それらの演算の計算量は多い。その結果、各々のフェーズの処理時間が長くなる。

[0004] 深層学習を高速化するために、ＣＰＵ（Central Processing Unit ）によって実現される推論器ではなく、ＧＰＵ（Graphics Processing Unit）によって実現される推論器がよく用いられる。さらに、深層学習専用のアクセラレータが実用化されている。

[0005] 図１１は、畳み込みニューラルネットワーク（ＣＮＮ：Convolutional Neural Network）の一例であるＶＧＧ（Visual Geometry Group ）－１６の構造を示す説明図である。ＶＧＧ－１６は、１３層の畳み込み層および３層の全結合層を含む。畳み込み層で、または畳み込み層とプーリング層とで抽出された特徴は、全結合層で分類される。

[0006] 図１１において、「Ｉ」は入力層を示す。「Ｃ」は畳み込み層を示す。図１１において、畳み込み層は３×３の畳み込みである。よって、たとえば、

図１１の最初の畳み込み演算には１画素あたり３（縦サイズ）×３（横サイズ）×３（入力チャネル）×６４（出力チャネル）個の積和演算を含む。また例えば図１１の５ブロック目の畳み込み層には、１画素あたり３（縦サイズ）×３（横サイズ）×５１２（入力チャネル）×５１２（出力チャネル）個の積和演算を含む。「Ｐ」はプーリング層を示す。図１１に示すＣＮＮでは、プーリング層は、Max Pooling 層である。「Ｆ」は全結合層を示す。「Ｏ」は出力層を示す。出力層では、softmax関数が使用される。なお、畳み込み層および全結合層は、正規化線形ユニット（Rectified Linear Unit ：ReLU）を含む。各層に付されている乗算式は、一枚の入力画像に対応するデータの縦サイズ×横サイズ×チャネル数を表す。また、層を表す直方体の体積は、層におけるアクティベーションの量に対応する。

先行技術文献

特許文献

[0007] 特許文献1：特開２０１９－１３９７４２号公報

非特許文献

[0008] 非特許文献1：P. N. Whatmough et al., ”FixyNN: Efficient Hardware for Mobile Computer Vision via Transfer Learning”, Feb, 27 2019

発明の概要

発明が解決しようとする課題

[0009] アクセラレータで推論器を実現する場合、主として２つの方法が考えられる。

[0010] ＣＮＮを例にすると、第１の方法では、ＣＮＮは、ＣＮＮを構成する複数の層の演算が共通の演算器で実行されるように構成される（例えば、特許文献１の段落００３３等参照。）。

[0011] 図１２は、複数の層の演算が共通の演算器で実行されるように構成されたＣＮＮの演算器を模式的に示す説明図である。推論器における演算を実行する部分は、演算器７００とメモリ（例えば、ＤＲＡＭ（Dynamic Random Acce

ss Memory）９００とで構成される。図１２に示す演算器７００には、多数の加算器と多数の乗算器とが形成される。図１２において、「＋」は加算器を示す。「＊」は乗算器を示す。なお、図１２には、３つの加算器と６個の乗算器とが例示されているが、ＣＮＮにおける全ての層の各々の演算が実行可能な数の加算器と乗算器とが形成されている。

[0012] 推論器の各層の演算が実行される場合、演算器７００は、演算実行対象の一層についてのパラメタをＤＲＡＭ９００から読み出す。そして、演算器７００は、一層における積和演算を、パラメタを係数として実行する。

[0013] 第２の方法では、ＣＮＮは、ＣＮＮを構成する全ての層の各々（特に、畳み込み層）の演算を、各層に対応する演算器で実行されるように構成される（例えば、非特許文献１参照）。なお、非特許文献１には、ＣＮＮが２つのステージに分割され、前段のステージにおいて、各々の層に対応する演算器が設けられることが記載されている。

[0014] 図１３は、各々の層に対応する演算器が設けられたＣＮＮを模式的に示す説明図である。図１３には、ＣＮＮにおける６つの層８０１，８０２，８０３，８０４，８０５，８０６が例示されている。層８０１，８０２，８０３，８０４，８０５，８０６のそれぞれに対応する演算器（回路）７０１，７０２，７０３，７０４，７０５，７０６が設けられている。

[0015] 演算器７０１～７０６は、対応する層８０１～８０６の演算を実行するので、パラメタが不変であれば、固定的に回路構成される。そして、非特許文献１には、パラメタを固定値にすることが記載されている。

[0016] 上記の第１の方法では、ＤＲＡＭ９００が備えられているので、パラメタが変更されても、演算器７０１～７０６の回路構成を変更することなく、ＣＮＮの機能が実行される。しかし、ＤＲＡＭ９００のデータ転送速度は、演算器７００の演算速度と比較すると低速である。すなわち、ＤＲＡＭ９００のメモリ帯域は狭い。したがって、演算回路７００とメモリの間のデータ転送がボトルネックになる。その結果、ＣＮＮの演算速度が制限される。

[0017] 上記の第２の方法では、各層のそれぞれに対応する演算器７０１～７０６

が設けられるので、ＣＮＮ全体としての回路規模が大きくなる。

[0018] 非特許文献１に記載された方法では、パラメタおよびネットワーク構成を固定することによって、ＣＮＮ全体としての加算器と乗算器の回路規模が小さくなる。ただし、非特許文献１に記載された方法では、各層に関して、完全に並列処理が可能であるように（fully-parallel）回路構成されるので、そのような回路構成によって、回路規模は大きくなる。なお、各層に関して各入力チャネル、各出力チャネルに対応する演算を並列処理するように回路構成されるので、そのような回路構成によって、回路規模は大きくなる。また、各層に関して、完全に並列処理が可能であるように回路構成されるので、一枚の画像に対応する入力データの処理時間は各層において同じ時間であることが望ましい。

[0019] ＣＮＮでは、先の層（出力層に近い層）であるほど、一枚の画像に対応する入力データの縦サイズや横サイズが小さくなる場合がある。例えばプーリング層によって一枚の画像に対応する入力データの縦サイズと横サイズが縮小される。各層が同じ時間で一枚の入力画像に対応するデータを処理するとした場合、先の層のチャネル数を極端に多くしない限り、先の層での計算量は小さくなる。換言すれば、本来、先の層であるほど、その層の演算を実行する回路規模は小さくてよい。しかし、非特許文献１に記載された方法では、演算器７００は、すべての入力チャネルと出力チャネルの演算を並列に実行可能に構成されるので、入力データの縦サイズと横サイズが少ない層については、一枚の画像に対応する入力データの処理が早く終わり、次の画像に対応する入力データが供給されるまで待ち時間が発生する。換言すれば演算器７００の利用率は低くなる。

[0020] また、非特許文献１に記載されたＣＮＮの構成は、ＣＮＮが２つのステージに分割され、前段のステージにおいて各々の層に対応する演算器が設けられるという構成である。そして、後段のステージは、ＤＲＡＭにパラメタが転送され、演算器としてプログラマブルなアクセラレータを用いるように構成される。すなわち、ＣＮＮは、ある程度のパラメタの変更やネットワーク

構成の変更に応えられるように構成され、ＣＮＮ全体として、すなわち、推論器全体として、パラメタおよびネットワーク構成を固定することは、非特許文献１に記載されていない。

[0021] 本発明は、推論器がハードウエアで実現される場合に、メモリ帯域の制約から解放され、かつ、推論器における各層の演算器の利用率が向上する情報処理回路および情報処理回路の設計方法を提供することを目的とする。

課題を解決するための手段

[0022] 本発明による情報処理回路は、深層学習における層の演算を実行し、入力データとパラメタ値とを用いて積和演算を行う積和回路と、パラメタ値を出力するパラメタ値出力回路とを含み、パラメタ値出力回路は、組み合わせ回路で構成されている。

[0023] 本発明による情報処理回路の設計方法は、深層学習における層の演算を実行する情報処理回路を生成する設計方法であって、学習済みの複数のパラメタ値とネットワーク構造を特定可能なデータとを入力し、入力データとパラメタ値とを用いて積和演算を行う回路であってネットワーク構造における層に特化した積和回路を作成し、複数のパラメタ値を出力する組み合わせ回路を作成する。

[0024] 本発明による情報処理回路の設計プログラムは、深層学習における層の演算を実行する情報処理回路を生成するためのプログラムであって、コンピュータに、学習済みの複数のパラメタ値とネットワーク構造を特定可能なデータとを入力する処理と、入力データとパラメタ値とを用いて積和演算を行う回路であってネットワーク構造における層に特化した積和回路を作成する処理と、複数のパラメタ値を出力する組み合わせ回路を作成する処理とを実行させる。

[0025] 本発明による情報処理回路設計装置は、深層学習における層の演算を実行する情報処理回路を生成する装置であって、学習済みの複数のパラメタ値とネットワーク構造を特定可能なデータとを入力する入力手段と、入力データとパラメタ値とを用いて積和演算を行う回路であってネットワーク構造にお

ける層に特化した積和回路を作成する演算器生成手段と、複数のパラメタ値を出力する組み合わせ回路を作成するパラメタ値出力回路作成手段とを含む。

発明の効果

[0026] 本発明によれば、メモリ帯域の制約から解放され、かつ、推論器における各層の演算器の利用率が向上する情報処理回路を得ることができる。

図面の簡単な説明

[0027] [図1]本実施形態の情報処理回路を模式的に示す説明図である。

[図2]基本回路の構成例を示す説明図である。

[図3]パラメタテーブルの回路構成例を説明するための説明図である。

[図4]情報処理回路設計装置の一例を示すブロック図である。

[図5]ＣＰＵを有するコンピュータの一例を示すブロック図である。

[図6]情報処理回路設計装置の動作を示すフローチャートである。

[図7]パラメタテーブルを最適化する処理の一例を示すフローチャートである。

[図8]パラメタ値の変更方法の一例を示す説明図である。

[図9]情報処理回路の主要部を示すブロック図である。

[図10]情報処理回路設計装置の主要部を示すブロック図である。

[図11]ＶＧＧ－１６の構造を示す説明図である。

[図12]複数の層の演算が共通の演算器で実行されるように構成されたＣＮＮの演算器を模式的に示す説明図である。

[図13]各々の層に対応する演算器が設けられたＣＮＮを模式的に示す説明図である。

発明を実施するための形態

[0028] 以下、本発明の実施形態を図面を参照して説明する。以下、情報処理回路として、ＣＮＮの推論器を例にする。また、ＣＮＮに入力されるデータとして、画像（画像データ）を例にする。

[0029] 図１３に例示された構成と同様に、情報処理回路は、ＣＮＮの各々の層に

対応する演算器が設けられたＣＮＮの推論器である。そして、情報処理回路は、パラメタが固定され、かつ、ネットワーク構成（深層学習アルゴリズムの種類、どのタイプの層を幾つどういった順で配置するのか、各層の入力データのサイズや出力データのサイズなど）が固定されたＣＮＮの推論器を実現する。すなわち、情報処理回路は、ＣＮＮの各層（例えば、畳み込み層および全結合層のそれぞれ）に特化した回路構成の回路である。特化するというのは、専ら当該層の演算を実行する専用回路であるということである。

[0030] なお、パラメタが固定されているということは、学習フェーズの処理が終了して、適切なパラメタが決定され、決定されたパラメタが使用されることを意味する。ただし、本実施形態では、学習フェーズで決定されたパラメタが変更されることがある。以下、パラメタが変更されることを、パラメタが最適化されると表現することがある。

[0031] また、本発明による情報処理回路を用いる推論器では、並列度は、データ入力速度や処理速度などを勘案して決定される。推論器におけるパラメタ（重み）と入力データとの乗算器は、組み合わせ論理回路（組み合わせ回路）で構成される。もしくは、パイプライン演算器で構成されてもよい。もしくは、順序回路で構成されてもよい。

[0032] 図１は、本実施形態の情報処理回路を模式的に示す説明図である。図１には、ＣＮＮを実現する情報処理回路１００における演算器２０１，２０２，２０３，２０４，２０５，２０６が例示されている。すなわち、図１には、ＣＮＮのうちの６層が例示されている。各演算器２０１，２０２，２０３，２０４，２０５，２０６は、層で使用されるパラメタ２１１，２１２，２１３，２１４，２１５，２１６と入力データとを対象として積和演算を実行する。演算器２０１～２０６は、複数の組み合わせ回路で実現される。パラメタ２１１～２１６も、組み合わせ回路で実現される。

[0033] なお、組み合わせ回路は、否定論理積回路（ＮＡＮＤ回路）、否定論理和回路（ＮＯＲ回路）、否定回路（反転回路：ＮＯＴ回路）、および、その組み合わせなどである。以下の説明において、１つの回路素子を組み合わせ回

路と表現することもあるが、複数の回路素子（ＮＡＮＤ回路、ＮＯＲ回路、ＮＯＴ回路など）を含む回路を組み合わせ回路と表現することもある。

[0034] 図１において、「＋」は加算器を示す。「＊」は乗算器を示す。なお、図１に例示された各層の演算器２０１～２０６のブロックに示されている加算器の数および乗算器の数は、表記のための単なる一例である。

[0035] 本実施形態では、演算器２０１～２０６のそれぞれにおいて並列演算が実行されるが、並列演算における１つの演算を実行する回路を基本回路とする。基本回路は、層の種類に応じてあらかじめ決定されている。

[0036] 図２は、基本回路の構成例を示す説明図である。６つの層のそれぞれの演算器（回路）２０１，２０２，２０３，２０４，２０５，２０６が例示されている。各層において、並列処理数の基本回路３００が設けられる。図２には、演算器２０３に含まれる基本回路３００が例示されているが、他の層の演算器２０１，２０２，２０４，２０５，２０６も同様の回路構成を有する。

[0037] 図２に示す例では、基本回路３００は、入力データとパラメタテーブル（重みテーブル）３０２からのパラメタ値を乗算し、乗算値を加算する積和回路３０１を含む。入力データは１つの値であってもよい。また、入力データは複数の値の組であってもよい。なお、図２には、パラメタ値を格納するパラメタテーブル３０２が示されているが、実際には、パラメタ値は記憶部（記憶回路）に記憶されているのではなく、パラメタテーブル３０２は、組み合わせ回路で実現される。本実施形態では、パラメタが固定されているので、パラメタテーブル３０２から、固定的な値であるパラメタ値が出力される。パラメタテーブル３０２は、１つの値を出力してもよい。また、パラメタテーブル３０２は、複数の値の組を出力してもよい。積和回路３０１は、１つの入力値と１つのパラメタ値の乗算を行ってもよい。また、積和演算器３０１は、入力値の組とパラメタ値の組との乗算を行ってもよい。入力値の組とパラメタ値の組との乗算結果の組の集約和の計算を行ってもよい。なお、一般に、１つの層に関して複数のパラメタ、もしくは、複数の組のパラメタ

が使用される、どのパラメタを出力するかは制御部４００が制御する。

[0038] 基本回路３００は、積和演算値を一時格納するレジスタ３０３を含んでもよい。積和回路３０１は、レジスタ３０３に一時格納された複数の乗算値を加算する加算器を含んでもよい。基本回路３００の入力には、別の基本回路３００の出力が接続されていてもよい。

[0039] 図３は、パラメタテーブル３０２の回路構成例を説明するための説明図である。図３（Ａ）には、真理値表３１１の一例が示されている。組み合わせ回路で、真理値表３１１を実現することができる。Ａ，Ｂ，Ｃのそれぞれは、組み合わせ回路の入力である。Ｚ１，Ｚ２は、組み合わせ回路の出力である。図３（Ａ）には、一例として、全加算器の真理値表３１１が示されているが、Ａ，Ｂ，Ｃをアドレスと見なし、Ｚ１，Ｚ２を出力データと見なすことができる。すなわち、Ｚ１，Ｚ２を、指定アドレスＡ，Ｂ，Ｃに対する出力データと見なすことができる。出力データをパラメタ値に対応づけると、何らかの入力（指定アドレス）に応じて、所望のパラメタ値を得ることができる。

[0040] 例えば、所望のパラメタ値が、ある特定の入力値（真理値表３１１ではＡ）によらず決定できるとすると、真理値表３１１における入力Ｂ、Ｃでパラメタ値を決定するように簡略化された真理値表３１２を用いるだけでよい。換言すれば、パラメタテーブル３０２を組み合わせ回路で実現する場合、パラメタを決定する入力の異種類が少ないほど、組み合わせ回路の回路規模が小さくなる。一般には、真理値表の簡単化にはクワイン・マクラスキー法などの公知技術が使われる。

[0041] 図２に示された演算器２０３は、制御部４００を含む。パラメタテーブル３０２におけるパラメタ値が、図２に示されたように指定アドレスに応じた出力データとして実現される場合には、制御部４００は、所望のタイミングで、出力データに対応する指定アドレスのデータをパラメタテーブル３０２に供給する。パラメタテーブル３０２は、指定アドレスに応じた出力データすなわちパラメタ値を積和回路３０１に出力する。なお、所望のタイミング

は、積和回路３０１が、パラメタテーブル３０２から出力されるべきパラメタ値を用いて乗算処理を実行する時点である。

[0042] 次に、図２に例示された演算器の設計方法を説明する。

[0043] 図４は、ＣＮＮの各層のパラメタテーブルの回路構成および演算器の回路構成を設計する情報処理回路設計装置の一例を示すブロック図である。図４に示す例では、情報処理回路設計装置５００は、パラメタテーブル最適化部５０１、パラメタテーブル生成部５０２、並列度決定部５０３、および演算器生成部５０４を含む。

[0044] 並列度決定部５０３は、ネットワーク構造（具体的には、ネットワーク構造を示すデータ。）を入力する。演算器生成部５０４は、層毎の演算器の回路構成を出力する。パラメタテーブル最適化部５０１は、学習フェーズで学習されたパラメタセット（各層における重み）と、並列度決定部５０３が決定した並列度を入力する。パラメタテーブル生成部５０２は、パラメタテーブルの回路構成を出力する。

[0045] 並列度決定部５０３は、層毎の並列度を決定する。パラメタテーブル最適化部５０１は、入力された層毎のパラメタと、並列度決定部５０３が決定した層毎の並列度とに基づいて、パラメタテーブルを最適化する。パラメタテーブルの個数は並列度で決まるが、パラメタテーブル最適化部５０１は、複数のパラメタテーブル３０２におけるそれぞれのパラメタを最適化する。ここで、最適化とは、パラメタテーブルに対応する組み合わせ回路の回路面積を小さくすることである。

[0046] 例えば、並列度決定対象の層（対象層）で実行される畳み込み演算が３×３×１２８×１２８（＝１４７,４５６の積和演算（パラメタ値とアクティベーション値とを対象とする積和演算）で構成されている場合を例にすると、並列度が「１２８」に決定されると、基本回路３００の数（並列度）は１２８である。各々の基本回路３００は、１１５２個の積和演算（１４７，４５６／１２８）に対する処理を実行する。その場合、基本回路３００において、１１５２のパラメタ値を有するパラメタテーブルが１２８個だけ備えられ

る。なお、上述したように、パラメタテーブル３０２は、記憶回路で実現されるのではなく、組み合わせ回路で実現される。

[0047] 後述するように、パラメタテーブル最適化部５０１は、あらかじめ定められた方法を用いて、パラメタテーブル３０２のパラメタ値を最適化する。パラメタテーブル生成部５０２は、最適化されたパラメタ値を有するパラメタテーブル３０２を実現するための回路構成を、パラメタテーブルの回路構成として出力する。

[0048] 演算器生成部５０４は、並列度決定部５０３が決定した層毎の並列度を入力する。演算器生成部５０４は、並列度が示す数の基本回路３００を並べた回路構成を、層毎に生成する。そして、演算器生成部５０４は、生成した層毎の回路構成を、演算器回路の構成として出力する。

[0049] 図４に示された情報処理回路設計装置５００における各構成要素は、１つのハードウエア、または１つのソフトウエアで構成可能である。また、各構成要素は、複数のハードウエア、または、複数のソフトウエアでも構成可能である。また、各構成要素の一部をハードウエアで構成し、他部をソフトウエアで構成することもできる。

[0050] 情報処理回路設計装置５００における各構成要素が、ＣＰＵ（Central Processing Unit ）等のプロセッサやメモリ等を有するコンピュータで実現される場合には、例えば、図５に示すＣＰＵを有するコンピュータで実現可能である。コンピュータは、ＣＰＵ１０００は、記憶装置１００１に格納されたプログラムに従って処理（情報処理回路設計処理）を実行することによって、図４に示された情報処理回路設計装置５００における各機能を実現する。すなわち、コンピュータは、図４に示された情報処理回路設計装置５００におけるパラメタテーブル最適化部５０１、パラメタテーブル生成部５０２、並列度決定部５０３、および演算器生成部５０４の機能を実現する。

[0051] 記憶装置１００１は、例えば、非一時的なコンピュータ可読媒体（non-transitory computer readable medium ）である。非一時的なコンピュータ可読媒体は、様々なタイプの実体のある記録媒体（tangible storage medium）の

いずれかである。非一時的なコンピュータ可読媒体の具体例として、磁気記録媒体（例えば、ハードディスクドライブ）、光磁気記録媒体（例えば、光磁気ディスク）、ＣＤ－ＲＯＭ（Compact Disc-Read Only Memory ）、ＣＤ－Ｒ（Compact Disc-Recordable ）、ＣＤ－Ｒ／Ｗ（Compact Disc-ReWritable ）、半導体メモリ（例えば、マスクＲＯＭ、ＰＲＯＭ（Programmable ROM ）、ＥＰＲＯＭ（Erasable PROM ）、フラッシュＲＯＭ）がある。

[0052] また、プログラムは、様々なタイプの一時的なコンピュータ可読媒体（transitory computer readable medium ）に格納されてもよい。一時的なコンピュータ可読媒体には、例えば、有線通信路または無線通信路を介して、すなわち、電気信号、光信号または電磁波を介して、プログラムが供給される。

[0053] メモリ１００２は、例えばＲＡＭ（Random Access Memory）で実現され、ＣＰＵ１０００が処理を実行するときに一時的にデータを格納する記憶手段である。メモリ１００２に、記憶装置１００１または一時的なコンピュータ可読媒体が保持するプログラムが転送され、ＣＰＵ１０００がメモリ１００２内のプログラムに基づいて処理を実行するような形態も想定しうる。

[0054] 次に、図６のフローチャートを参照して、情報処理回路設計装置の動作を説明する。

[0055] パラメタテーブル最適化部５０１は、学習フェーズで学習されたパラメタセット（複数のパラメタ値）を入力し、並列度決定部５０３は、あらかじめ決められているネットワーク構造を示すデータを入力する（ステップＳ１１）。

[0056] なお、本実施形態におけるネットワーク構造の概念の１つである深層学習アルゴリズムの種類として、例えば、ＡｌｅｘＮｅｔ、ＧｏｏｇＬｅＮｅｔ、ＲｅｓＮｅｔ（Residual Network）、ＳＥＮｅｔ（Squeeze-and-Excitation Networks ）、ＭｏｂｉｌｅＮｅｔ、ＶＧＧ－１６、ＶＧＧ－１９がある。また、ネットワーク構造の概念の１つである層数として、例えば、深層学習アルゴリズムの種類に応じた層数が考えられる。また、ネットワーク構造の概念として、フィルタサイズなども含められ得る。

[0057] 以下、ネットワーク構造を示すデータを入力することを、ネットワーク構造を入力すると表現する。

[0058] 並列度決定部５０３は、層毎の並列度を決定する（ステップＳ１２）。一例として、並列度決定部５０３は、（１）式で並列度Ｎを決定する。例えば、入力された深層学習アルゴリズムの種類で特定される層の数が１９である場合には、並列度決定部５０３は、１９の層のそれぞれの並列度を決定する。

[0059] $N = C_L / D_L$ ・・・（１）

[0060] （１）式において、C_Lは、並列度決定対象の層（対象層）において１画面の全画素を１つの積和演算器で処理するのに必要なクロック数を示す。D_Lは、対象層において１画面の処理に要するクロック数（許容されるクロック数）を示す。

[0061] 図１１に示されたＣＮＮを例にすると、１画面が縦サイズ２２４、横サイズ２２４（５０，１７６画素）の層（第１ブロックにおける層とする。）において１クロックで縦横１画素の処理し、１画面全体を５０，１７６クロックで実行されるとする。これに対して、１画面が縦サイズ１４、横サイズ１４の層（第５ブロックにおける層とする）では、同じ時間で１画面の処理を完了するためには２５６クロックで縦横１画素の処理が実行すれば、１画面分の処理を第１クロックと同じ５０、１７６クロックで完了できる。第１ブロックの畳み込み層の処理は、１画素あたり３（縦サイズ）×３（横サイズ）×３（入力チャネル）×６４（出力チャネル）（＝１７２８個）である。したがって、全画素を一つの積和演算器で処理するのに必要なクロック数は１７２８個×５０，１７６画素＝８６、７０４、１２８個である。１画面全体を５０，１７６クロックで完了するために、第１ブロックの層の並列度は、１７２８である。一方、第５ブロックの畳み込み層の処理は、１画素あたり３（縦サイズ）×３（横サイズ）×５１２（入力チャネル）×５１２（出力チャネル）（＝２、３５９、２９６個）である。したがって、全画素を一つの積和演算器で処理するのに必要なクロック数は２、３５９、２９６個×

１９６画素＝４６２、４２２、０１６個である。１画面全体を５０，１７６クロックで完了するために、第５ブロックの層の並列度は、９、２１６である。

[0062] 所望される演算速度（１画面の処理量／所要クロック数）に応じて、各層の並列度が決定されることによって、例えば、（１）式に基づいて各層の並列度が決定されることによって、各層の演算器（具体的には、演算器に含まれる複数の基本回路３００）を常に稼働する状態にすることができる。図１３に示された構成において、演算器７０１～７０６に対して何らの工夫も施されない場合には、演算器７０６の稼働率は、演算器７０１の稼働率よりも低い。非特許文献１に記載された構成を例にすると、各層はfully-parallelで構成されるので、出力層に近い層では、演算器の稼働率はより低い。しかし、本実施形態では、全ての層の演算器の稼働率を高く維持することができる。

[0063] パラメタテーブル最適化部５０１は、層毎に、決定された並列度に応じて、パラメタテーブル３０２を生成する（ステップＳ１３）。さらに、パラメタテーブル最適化部５０１は、生成したパラメタテーブル３０２を最適化する（ステップＳ１４）。

[0064] 図７は、パラメタテーブル３０２を最適化する処理（パラメタテーブル最適化処理）の一例を示すフローチャートである。

[0065] パラメタテーブル最適化処理において、パラメタテーブル最適化部５０１は、ＣＮＮ（推論器）の認識精度を測定する（Ｓ１４１）。ステップＳ１４１では、パラメタテーブル最適化部５０１は、決定された並列度に応じた数の基本回路３００とパラメタテーブルの回路構成とを用いた推論器を使用してシミュレーションを実行する。シミュレーションは、適当な入力データを用いた推論である。そして、シミュレーション結果を正解と比較すること等によって、認識精度を得る。

[0066] パラメタテーブル最適化部５０１は、認識精度が第１の基準値以上であるか否か確認する（ステップＳ１４２）。第１の基準値は、あらかじめ定めら

れたしきい値である。認識精度が第１の基準値以上である場合には、パラメタテーブル最適化部５０１は、パラメタテーブル３０２の回路面積を見積もる。そして、パラメタテーブル３０２の回路面積が第２の基準値以下であるか否か確認する（ステップＳ１４４）。第２の基準値は、あらかじめ定められたしきい値である。パラメタテーブル最適化部５０１は、例えば、パラメタテーブル３０２を構成する組み合わせ回路における論理回路の数に基づいて、パラメタテーブル３０２の回路面積を見積もることができる。

[0067] パラメタテーブル３０２の回路面積が第２の基準値以下である場合には、パラメタテーブル最適化部５０１は、パラメタテーブル最適化処理を終了する。

[0068] 認識精度が第１の基準値未満である場合、または、パラメタテーブル３０２の回路面積が第２の基準値を超える場合には、パラメタテーブル最適化部５０１は、パラメタ値を変更する（ステップＳ１４３）。そして、ステップＳ１４１に移行する。

[0069] ステップＳ１４３において、パラメタテーブル最適化部５０１は、認識精度が第１の基準値未満である場合には、認識精度が向上すると想定される方向にパラメタ値を変更する。認識精度が向上すると想定される方向が不明である場合には、パラメタテーブル最適化部５０１は、カットアンドトライ（cut and try ）でパラメタ値を変更してもよい。

[0070] ステップＳ１４３において、パラメタテーブル最適化部５０１は、パラメタテーブル３０２の回路面積が第２の基準値を超える場合には、パラメタテーブル３０２の回路面積が小さくなるようにパラメタ値を変更する。パラメタテーブル３０２の回路面積を小さくするためのパラメタ値の変更方法として、例えば、以下のような方法がある。

[0071] ・パラメタテーブル３０２において、絶対値が所定のしきい値よりも小さいパラメタ値を０に変更する。

・パラメタテーブル３０２において、所定のしきい値よりも大きいパラメタ値（正数）を、パラメタテーブル３０２における最大のパラメタ値で置き換

える。

・所定のしきい値よりも小さいパラメタ値（負数）を、パラメタテーブル３０２における最小のパラメタ値で置き換える。

・パラメタテーブル３０２における所定の領域毎に、代表的な値を設定し、領域内の全てのパラメタ値を代表的な値に置き換える。なお、代表的な値は、一例として、偶数の値、奇数の値、最頻値などである。

・パラメタ値を、パラメタテーブル３０２における近傍のパラメタ値に置き換える。

[0072] なお、パラメタテーブル最適化部５０１は、上記の複数の方法のうちの１つの方法を用いてもよいが、上記の複数の方法のうちの２つ以上の方法を併用してもよい。

[0073] 図８は、パラメタ値の変更方法の一例を示す説明図である。図８には、３×３のサイズのパラメタテーブルが例示されている。図８（Ａ）には、パラメタ値が変更される前のパラメタテーブル３０２ａが示されている。図８（Ｂ）には、パラメタ値が変更された後のパラメタテーブル３０２ｂが示されている。

[0074] 図８に示す例では、所定のしきい値である「３」よりも小さいパラメタ値が「０」に変更されている。

[0075] 上記の各方法に共通する目的は、パラメタテーブル３０２において、同じ値が頻出する、すなわち、同値のパラメタ値が増加するか、または、同じパターンが連続するようにすることである。なお、同じパターンが連続するという意味は、例えば、パラメタ値「１」「２」「３」（同じパターンの一例）のパターンが連続して出現するということである。

[0076] 上述したように、パラメタテーブル３０２が組み合わせ回路で実現される場合、パラメタ値の種類が少ないほど、組み合わせ回路の回路規模が小さくなる。また、同じパターンが連続する場合にも、組み合わせ回路の回路規模が小さくなることが期待される。

[0077] 本実施形態では、情報処理回路設計装置５００は、推論器の認識精度が所

望のレベル以上（具体的には、第１の基準値以上）であり、かつ、回路面積が所望のサイズ以下（具体的には、第２の基準値以下）になった場合に、パラメタテーブル最適化処理を終了する。

[0078] 図６に示すように、演算器生成部５０４は、層毎の演算器の回路構成を生成して出力する（ステップＳ１５，Ｓ１７）。すなわち、演算器生成部５０４は、並列度決定部５０３が決定した層毎の並列度に応じた演算器の回路構成を出力する。なお、本実施形態では、各層の基本回路３００があらかじめ決められているので、演算器生成部５０４は、並列度決定部５０３が決定した並列度に応じた数の基本回路３００（具体的には、層に特化した積和回路３０１）を生成する。

[0079] パラメタテーブル生成部５０２は、パラメタテーブル３０２の回路構成を生成して出力する（ステップＳ１６，Ｓ１７）。すなわち、パラメタテーブル生成部５０２は、パラメタテーブル最適化部５０１が最適化したパラメタ値を出力するための回路構成を生成して出力する。パラメタ値を出力するための回路構成は、例えば、図３（Ｂ）に例示されたような真理値表を実現する組み合わせ回路の構成である。

[0080] なお、図６のフローチャートでは、ステップＳ１４～Ｓ１６の処理が順次に実行されるが、ステップＳ１４，Ｓ１６の処理とステップＳ１５の処理とは、並行して実行可能である。

[0081] また、ステップＳ１４の処理を実行するパラメタテーブル最適化部５０１が設けられていない場合でも、並列度決定部５０３が適切な並列度を決定することによって、回路規模が小さくなるという効果を得ることができる。

[0082] 以上に説明したように、本実施形態の情報処理回路としての推論器において、パラメタテーブル３０２は組み合わせ回路で実現されているので、図１２に示されたパラメタ値をメモリから読み出すように構成された情報処理回路に比べて処理速度が向上する。また、推論器において各層の並列度がその層に所望される演算速度などに応じて定められているので、各層がfully-parallelで構成される場合に比べて、全ての層の演算器の稼働率を高く維持する

ことができる。また、本実施形態の推論器は、各層がfully-parallelで構成される場合に比べて、回路規模が小さくなる。その結果、推論器の消費電力が低減する。

[0083] また、情報処理回路設計装置５００がパラメタ値を最適化するように構成される場合には、推論器の回路規模をより小さくすることができる。

[0084] なお、本実施形態では、ＣＮＮの推論器を例にして情報処理回路が説明されたが、入力データとパラメタ値とを用いる演算を行う層を有する他のネットワークに本実施形態を適用することができる。また、本実施形態では、入力データとして画像データが用いられているが、画像データ以外を入力データとするネットワークでも、本実施形態を活用することができる。

[0085] データセンタの電力消費量は多いので、データセンタにおいて深層学習のアルゴリズムが実行される場合に、低消費電力で実行されることが望ましい。本実施形態の情報処理回路を用いる場合には消費電力が低減するので、本実施形態の情報処理回路は、データセンタにおいて有効に活用可能である。

[0086] また、エッジ側でも、低消費電力が求められる。本実施形態の情報処理回路は、エッジ側においても有効に活用可能である。

[0087] 図９は、情報処理回路の主要部を示すブロック図である。情報処理回路１０は、深層学習における層の演算を実行し、入力データとパラメタ値とを用いて積和演算を行う積和回路１１（実施形態では、積和回路３０１で実現される。）と、パラメタ値を出力するパラメタ値出力回路１２（実施形態では、パラメタテーブル３０２で実現される。）とを含み、パラメタ値出力回路１２は、組み合わせ回路で構成されている。

[0088] 図１０は、情報処理回路設計装置の主要部を示すブロック図である。情報処理回路設計装置２０は、深層学習における層の演算を実行する情報処理回路を生成する装置であって、学習済みの複数のパラメタ値とネットワーク構造を特定可能なデータとを入力する入力手段２１（実施形態では、パラメタテーブル最適化部５０１の一部および並列度決定部５０３の一部として実現される。）と、入力データとパラメタ値とを用いて積和演算を行う回路であ

ってネットワーク構造における層に特化した積和回路を作成する演算器生成手段２２（実施形態では、演算器生成部５０４で実現される。）と、複数のパラメタ値を出力する組み合わせ回路を作成するパラメタ値出力回路作成手段２３（実施形態では、パラメタテーブル生成部５０２で実現される。）とを備えている。

[0089] 上記の実施形態の一部または全部は、以下の付記のようにも記載され得るが、以下に限定されるわけではない。

[0090] （付記１）深層学習における層の演算を実行する情報処理回路であって、

入力データとパラメタ値とを用いて積和演算を行う積和回路と、

前記パラメタ値を出力するパラメタ値出力回路とを備え、

前記パラメタ値出力回路は、組み合わせ回路で構成されている

ことを特徴とする情報処理回路。

[0091] （付記２）並列処理数に応じた数の基本回路を備え、

複数の前記基本回路の各々は、前記積和回路と前記パラメタ値出力回路とを含む

付記１の情報処理回路。

[0092] （付記３）前記基本回路は、層に特化した回路構成を有し、

前記パラメタ値出力回路は、固定値である前記パラメタ値を出力する

付記２の情報処理回路。

[0093] （付記４）深層学習における層の演算を実行する情報処理回路を生成する情報処理回路の設計方法であって、

学習済みの複数のパラメタ値とネットワーク構造を特定可能なデータとを入力し、

入力データとパラメタ値とを用いて積和演算を行う回路であって前記ネットワーク構造における層に特化した積和回路を作成し、

前記複数のパラメタ値を出力する組み合わせ回路を作成する

ことを特徴とする情報処理回路の設計方法。

[0094] （付記５）深層学習が複数の層で実現される場合に、層毎の前記積和回路と

層毎の前記組み合わせ回路とを作成する

付記４の情報処理回路の設計方法。

[0095] （付記６）前記層に求められる演算速度に基づく並列度を決定し、

前記並列度に応じた数の積和回路を作成する

付記４または付記５の情報処理回路の設計方法。

[0096] （付記７）入力された前記複数のパラメタ値のうちの１つ以上を、同値のパラメタ値が増加するように変更する

付記４から付記６のうちのいずれかの情報処理回路の設計方法。

[0097] （付記８）入力された前記複数のパラメタ値のうちの１つ以上を、複数のパラメタ値によるパターンが連続して出現するように変更する

付記４から付記７のうちのいずれかの情報処理回路の設計方法。

[0098] （付記９）情報処理回路の精度を測定し、

前記組み合わせ回路の面積を見積り、

前記情報処理回路の精度が第１の基準値以上であり、かつ、前記組み合わせ回路の面積が第２の基準値以下であるという条件が満たされるまで、前記パラメタ値を繰り返し変更する

付記７または付記８の情報処理回路の設計方法。

[0099] （付記１０）深層学習における層の演算を実行する情報処理回路を生成するための情報処理回路の設計プログラムが格納されたコンピュータ読み取り可能な記録媒体であって、

前記情報処理回路の設計プログラムは、

学習済みの複数のパラメタ値とネットワーク構造を特定可能なデータとを入力する処理と、

入力データとパラメタ値とを用いて積和演算を行う回路であって前記ネットワーク構造における層に特化した積和回路を作成する処理と、

前記複数のパラメタ値を出力する組み合わせ回路を作成する処理と

をプロセッサに実行させることを特徴とする。

[0100] （付記１１）前記情報処理回路の設計プログラムは、

深層学習が複数の層で実現される場合に、層毎の前記積和回路と層毎の前記組み合わせ回路とを作成する処理をプロセッサに実行させる

付記１０の記録媒体。

[0101] （付記１２）前記情報処理回路の設計プログラムは、

前記層に求められる演算速度に基づく並列度を決定する処理と、

前記並列度に応じた数の積和回路を作成する処理と

をプロセッサに実行させる付記１０また付記１１の記録媒体。

[0102] （付記１３）前記情報処理回路の設計プログラムは、

入力された前記複数のパラメタ値のうちの１つ以上を、同値のパラメタ値が増加するように変更する処理をプロセッサに実行させる

付記１０から付記１２のうちのいずれかの記録媒体。

[0103] （付記１４）深層学習における層の演算を実行する情報処理回路を生成する情報処理回路設計装置であって、

学習済みの複数のパラメタ値とネットワーク構造を特定可能なデータとを入力する入力手段と、

入力データとパラメタ値とを用いて積和演算を行う回路であって前記ネットワーク構造における層に特化した積和回路を作成する演算器生成手段と、

前記複数のパラメタ値を出力する組み合わせ回路を作成するパラメタ値出力回路作成手段と

を備えたことを特徴とする情報処理回路設計装置。

[0104] （付記１５）深層学習が複数の層で実現される場合に、前記演算器生成手段は、層毎の前記積和回路を作成し、前記パラメタ値出力回路作成手段は、層毎の前記組み合わせ回路を作成する

付記１４の情報処理回路設計装置。

[0105] （付記１６）前記層に求められる演算速度に基づく並列度を決定する並列度決定手段を備え、

前記演算器生成手段は、前記並列度に応じた数の積和回路を作成する

付記１４または付記１５の情報処理回路設計装置。

[0106] （付記１７）入力された前記複数のパラメタ値のうちの１つ以上を、同値のパラメタ値が増加するように変更するパラメタ最適化手段を備えた
付記１４から付記１６のうちのいずれかの情報処理回路設計装置。

[0107] （付記１８）深層学習における層の演算を実行する情報処理回路を生成するためのプログラムであって、
コンピュータに、
学習済みの複数のパラメタ値とネットワーク構造を特定可能なデータとを入力する処理と、
入力データとパラメタ値とを用いて積和演算を行う回路であって前記ネットワーク構造における層に特化した積和回路を作成する処理と、
前記複数のパラメタ値を出力する組み合わせ回路を作成する処理と
を実行させるための情報処理回路の設計プログラム。

[0108] （付記１９）コンピュータに、
深層学習が複数の層で実現される場合に、層毎の前記積和回路と層毎の前記組み合わせ回路とを作成させる
付記１８の情報処理回路の設計プログラム。

[0109] （付記２０）コンピュータに、
前記層に求められる演算速度に基づく並列度を決定する処理と、
前記並列度に応じた数の積和回路を作成する処理と
を実行させる付記１８または付記１９の情報処理回路の設計プログラム。

[0110] （付記２１）コンピュータに、
入力された前記複数のパラメタ値のうちの１つ以上を、同値のパラメタ値が増加するように変更する処理を実行させる
付記１８から付記２０のうちのいずれかの情報処理回路の設計プログラム。

[0111] 以上、実施形態を参照して本願発明を説明したが、本願発明は上記の実施形態に限定されない。本願発明の構成や詳細には、本願発明のスコープ内で当業者が理解し得る様々な変更をすることができる。

符号の説明

[0112] １０　情報処理回路

１１　積和回路

１２　パラメタ値出力回路

２０　情報処理回路設計装置

２１　入力手段

２２　演算器生成手段

２３　パラメタ値出力回路作成手段

１００　情報処理回路

２０１，２０２，２０３，２０４，２０５，２０６　演算器

２１１，２１２，２１３，２１４，２１５，２１６　パラメタ

３００　基本回路

３０１　積和回路

３０２　パラメタテーブル

３０３　レジスタ

４００　制御部

５００　情報処理回路設計装置

５０１　パラメタテーブル最適化部

５０２　パラメタテーブル生成部

５０３　並列度決定部

５０４　演算器生成部

１０００　ＣＰＵ

１００１　記憶装置

１００２　メモリ

請求の範囲

[請求項1] 深層学習における層の演算を実行する情報処理回路であって、

入力データとパラメタ値とを用いて積和演算を行う積和回路と、

前記パラメタ値を出力するパラメタ値出力回路とを備え、

前記パラメタ値出力回路は、組み合わせ回路で構成されている

ことを特徴とする情報処理回路。

[請求項2] 並列処理数に応じた数の基本回路を備え、

複数の前記基本回路の各々は、前記積和回路と前記パラメタ値出力回路とを含む

請求項１記載の情報処理回路。

[請求項3] 前記基本回路は、層に特化した回路構成を有し、

前記パラメタ値出力回路は、固定値である前記パラメタ値を出力する

請求項２記載の情報処理回路。

[請求項4] 深層学習における層の演算を実行する情報処理回路を生成する情報処理回路の設計方法であって、

学習済みの複数のパラメタ値とネットワーク構造を特定可能なデータとを入力し、

入力データとパラメタ値とを用いて積和演算を行う回路であって前記ネットワーク構造における層に特化した積和回路を作成し、

前記複数のパラメタ値を出力する組み合わせ回路を作成する

ことを特徴とする情報処理回路の設計方法。

[請求項5] 深層学習が複数の層で実現される場合に、層毎の前記積和回路と層毎の前記組み合わせ回路とを作成する

請求項４記載の情報処理回路の設計方法。

[請求項6] 前記層に求められる演算速度に基づく並列度を決定し、

前記並列度に応じた数の積和回路を作成する

請求項４または請求項５記載の情報処理回路の設計方法。

[請求項7] 入力された前記複数のパラメタ値のうちの１つ以上を、同値のパラメタ値が増加するように変更する

請求項４から請求項６のうちのいずれか１項に記載の情報処理回路の設計方法。

[請求項8] 入力された前記複数のパラメタ値のうちの１つ以上を、複数のパラメタ値によるパターンが連続して出現するように変更する

請求項４から請求項７のうちのいずれか１項に記載の情報処理回路の設計方法。

[請求項9] 情報処理回路の精度を測定し、

前記組み合わせ回路の面積を見積り、

前記情報処理回路の精度が第１の基準値以上であり、かつ、前記組み合わせ回路の面積が第２の基準値以下であるという条件が満たされるまで、前記パラメタ値を繰り返し変更する

請求項７または請求項８記載の情報処理回路の設計方法。

[請求項10] 深層学習における層の演算を実行する情報処理回路を生成するための情報処理回路の設計プログラムが格納されたコンピュータ読み取り可能な記録媒体であって、

前記情報処理回路の設計プログラムは、

学習済みの複数のパラメタ値とネットワーク構造を特定可能なデータとを入力する処理と、

入力データとパラメタ値とを用いて積和演算を行う回路であって前記ネットワーク構造における層に特化した積和回路を作成する処理と、

前記複数のパラメタ値を出力する組み合わせ回路を作成する処理とをプロセッサに実行させることを特徴とする。

[請求項11] 前記情報処理回路の設計プログラムは、

深層学習が複数の層で実現される場合に、層毎の前記積和回路と層毎の前記組み合わせ回路とを作成する処理をプロセッサに実行させる

請求項１０記載の記録媒体。

[請求項12] 前記情報処理回路の設計プログラムは、

前記層に求められる演算速度に基づく並列度を決定する処理と、

前記並列度に応じた数の積和回路を作成する処理と

をプロセッサに実行させる請求項１０または請求項１１記載の記録媒体。

[請求項13] 前記情報処理回路の設計プログラムは、

入力された前記複数のパラメタ値のうちの１つ以上を、同値のパラメタ値が増加するように変更する処理をプロセッサに実行させる

請求項１０から請求項１２のうちのいずれか１項に記載の記録媒体。

[請求項14] 深層学習における層の演算を実行する情報処理回路を生成する情報処理回路設計装置であって、

学習済みの複数のパラメタ値とネットワーク構造を特定可能なデータとを入力する入力手段と、

入力データとパラメタ値とを用いて積和演算を行う回路であって前記ネットワーク構造における層に特化した積和回路を作成する演算器生成手段と、

前記複数のパラメタ値を出力する組み合わせ回路を作成するパラメタ値出力回路作成手段と

を備えたことを特徴とする情報処理回路設計装置。

[請求項15] 深層学習が複数の層で実現される場合に、前記演算器生成手段は、層毎の前記積和回路を作成し、前記パラメタ値出力回路作成手段は、層毎の前記組み合わせ回路を作成する

請求項１４記載の情報処理回路設計装置。

[請求項16] 前記層に求められる演算速度に基づく並列度を決定する並列度決定手段を備え、

前記演算器生成手段は、前記並列度に応じた数の積和回路を作成す

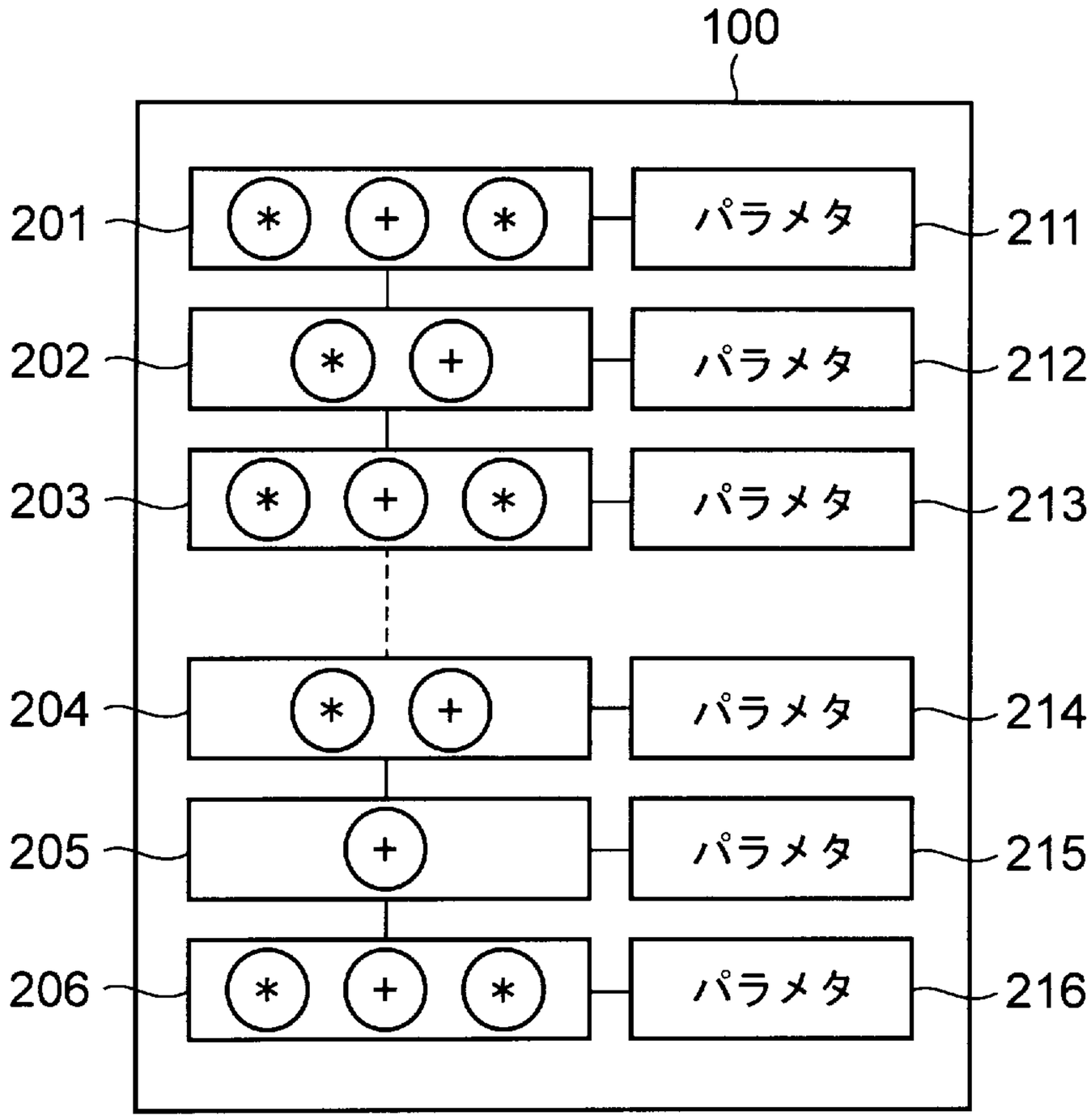
る

請求項１４または請求項１５記載の情報処理回路設計装置。

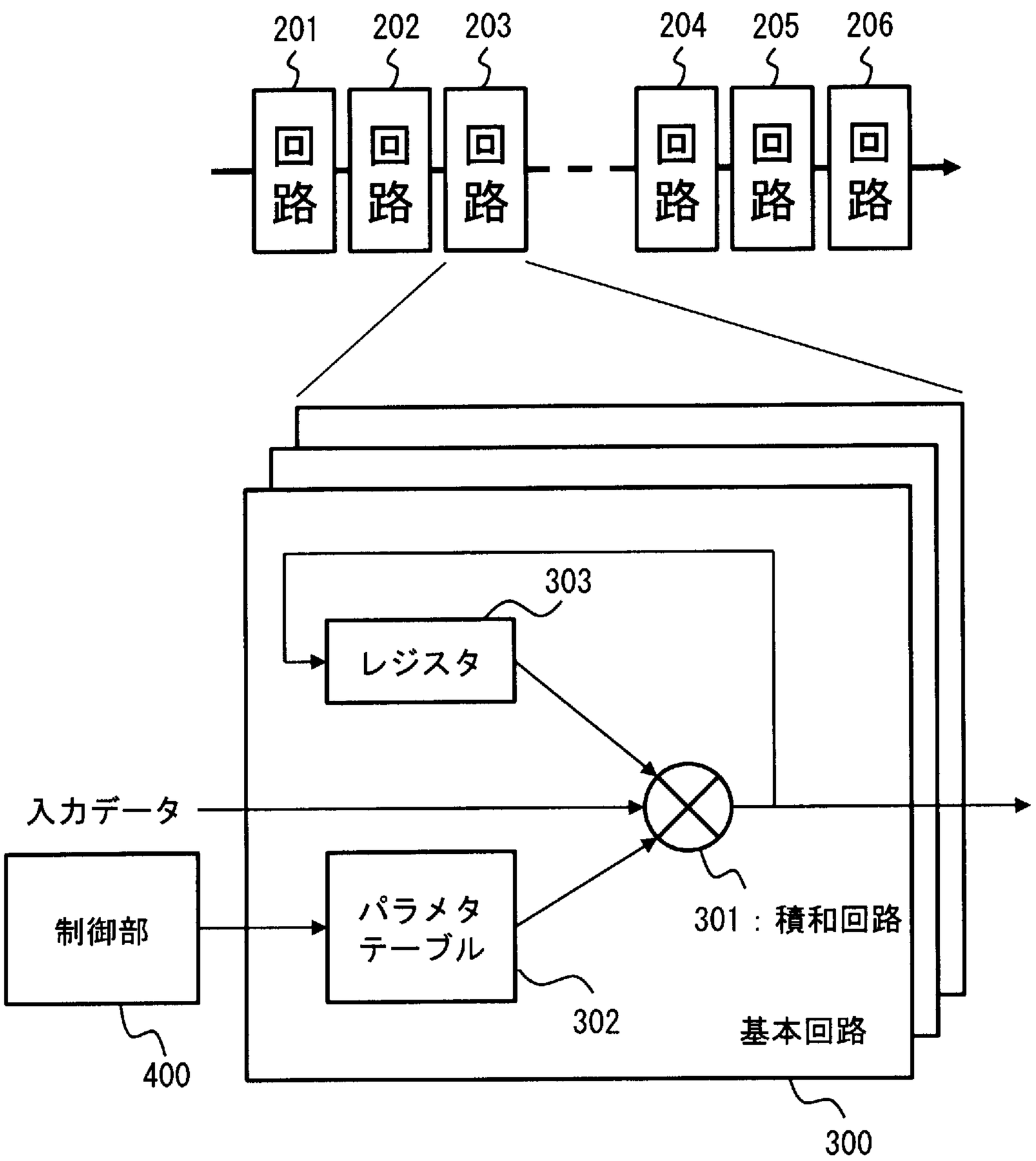
[請求項17] 入力された前記複数のパラメタ値のうちの１つ以上を、同値のパラメタ値が増加するように変更するパラメタ最適化手段を備えた

請求項１４から請求項１６のうちのいずれか１項に記載の情報処理回路設計装置。

[図1]



[図2]


201
202
203
204
205
206
回路
回路
回路
回路
回路
回路
303
レジスタ
入力データ
制御部
パラメタ
テーブル
301：積和回路
302
基本回路
400
300

[図3]

311

A	B	C	Z1	Z2
0	0	0	0	0
0	0	1	0	1
0	1	0	0	1
0	1	1	1	0
1	0	0	0	0
1	0	1	0	1
1	1	0	0	1
1	1	1	1	0

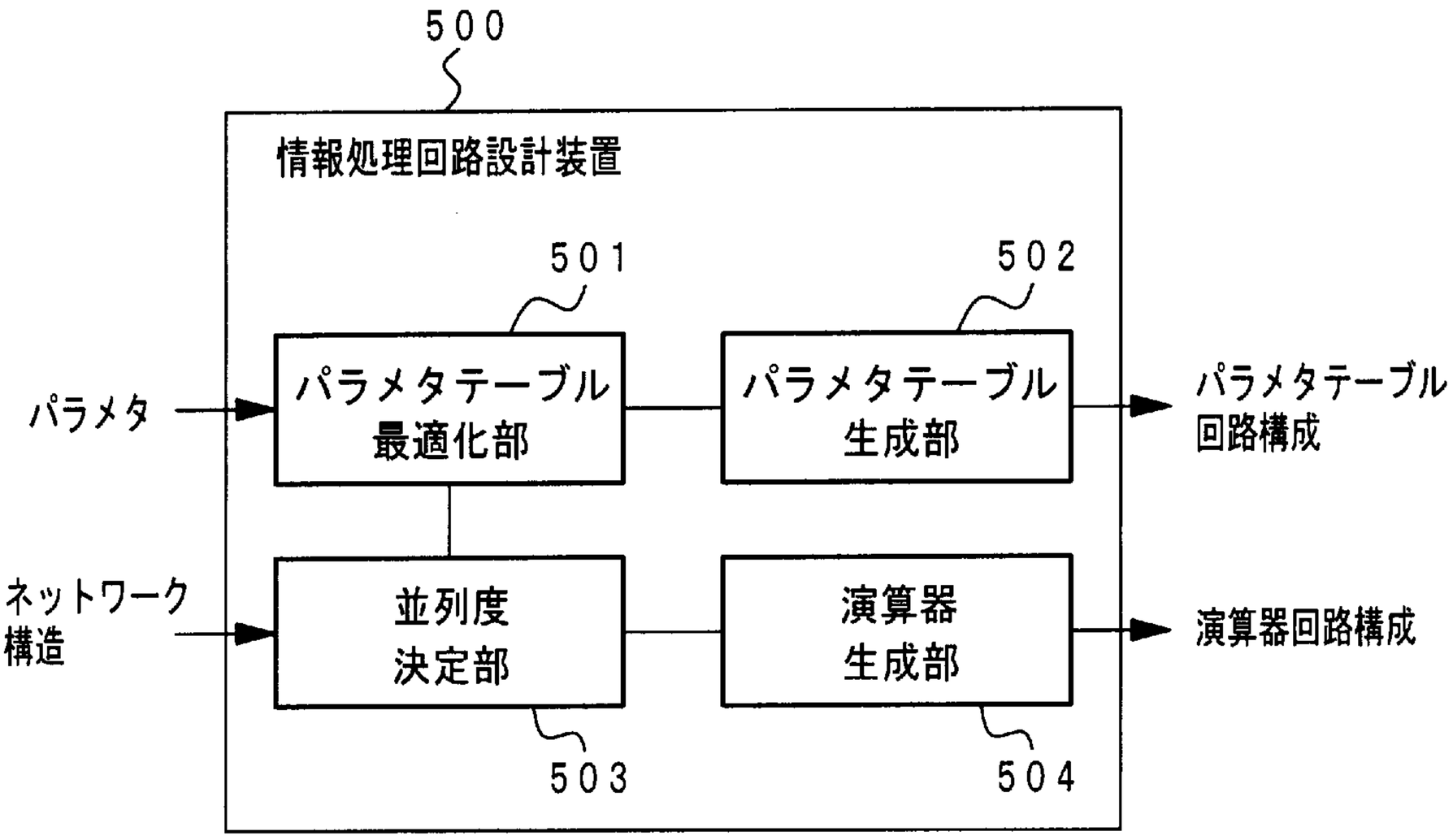
(A)

312

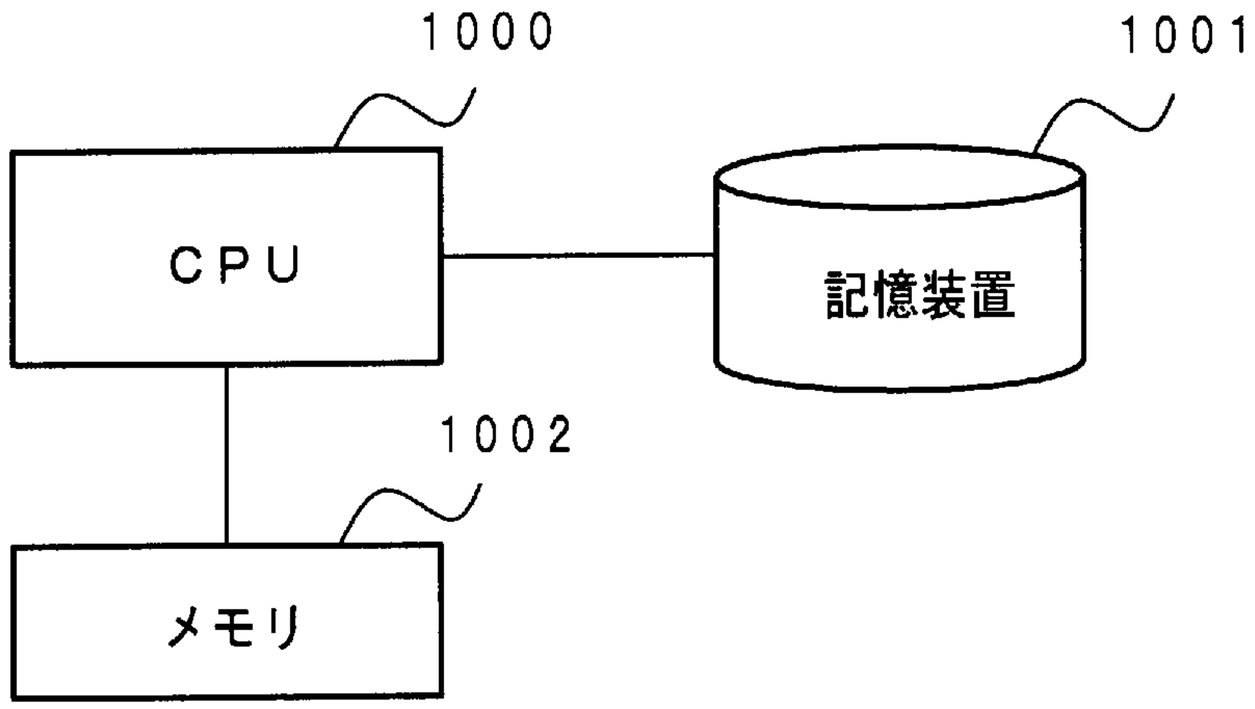
A	B	C	Z1	Z2
*	0	0	0	0
*	0	1	0	1
*	1	0	0	1
*	1	1	1	0

(B)

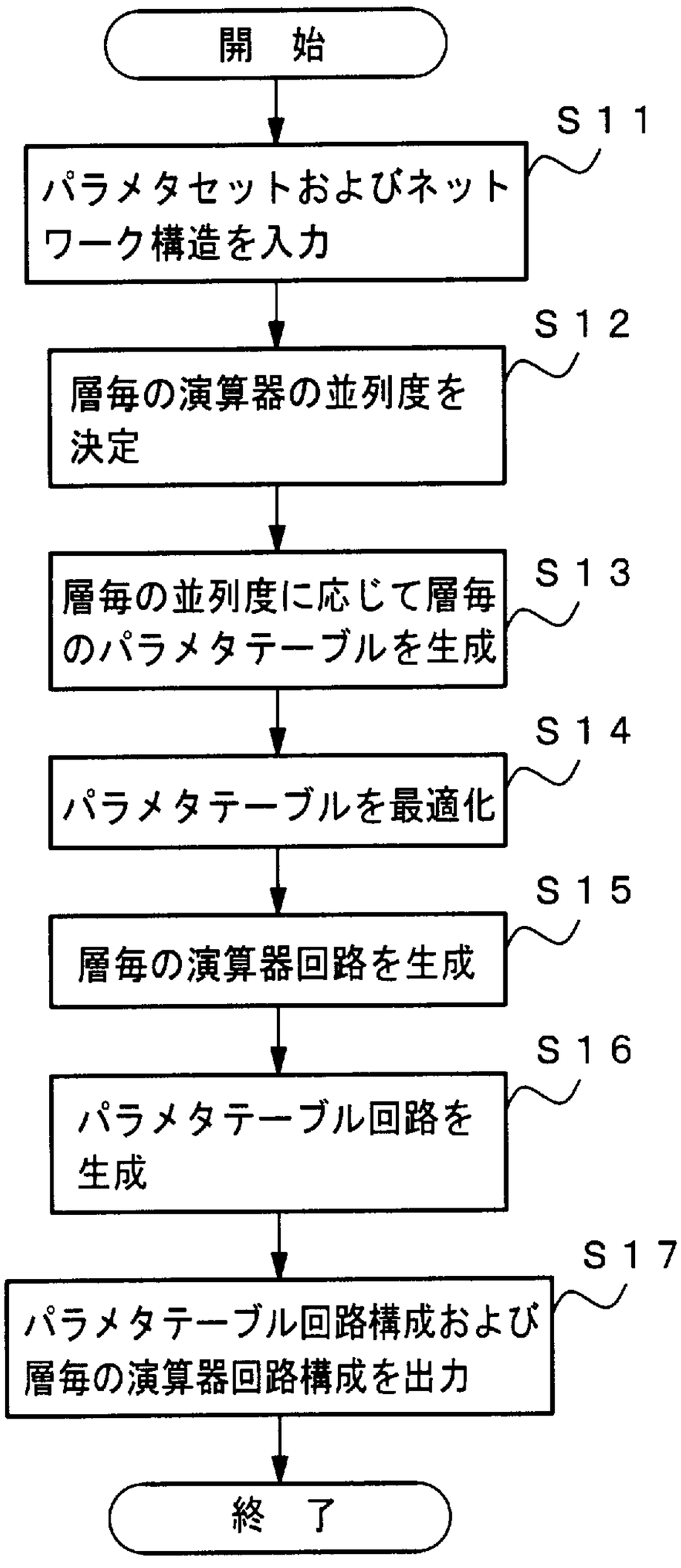
[図4]



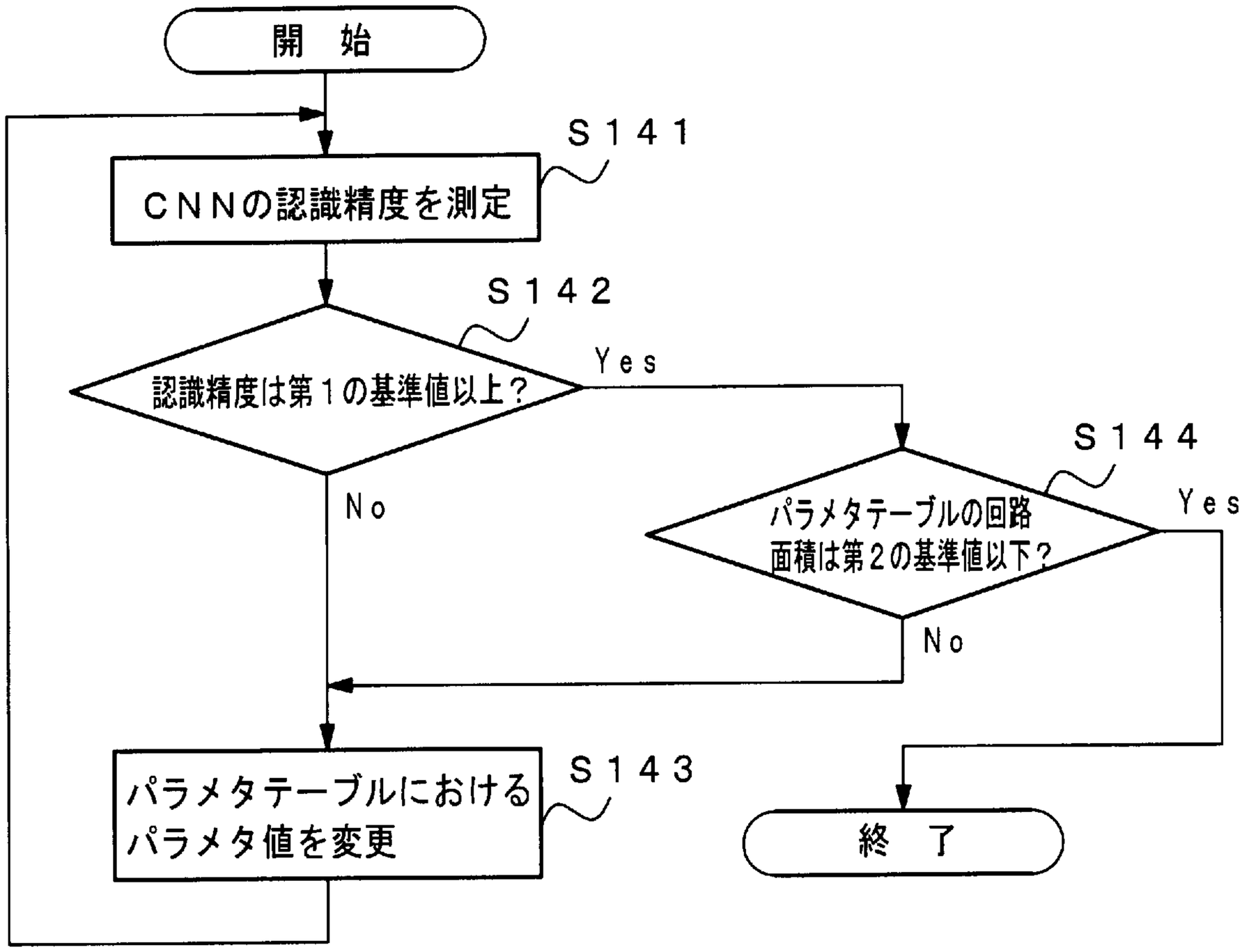
[図5]



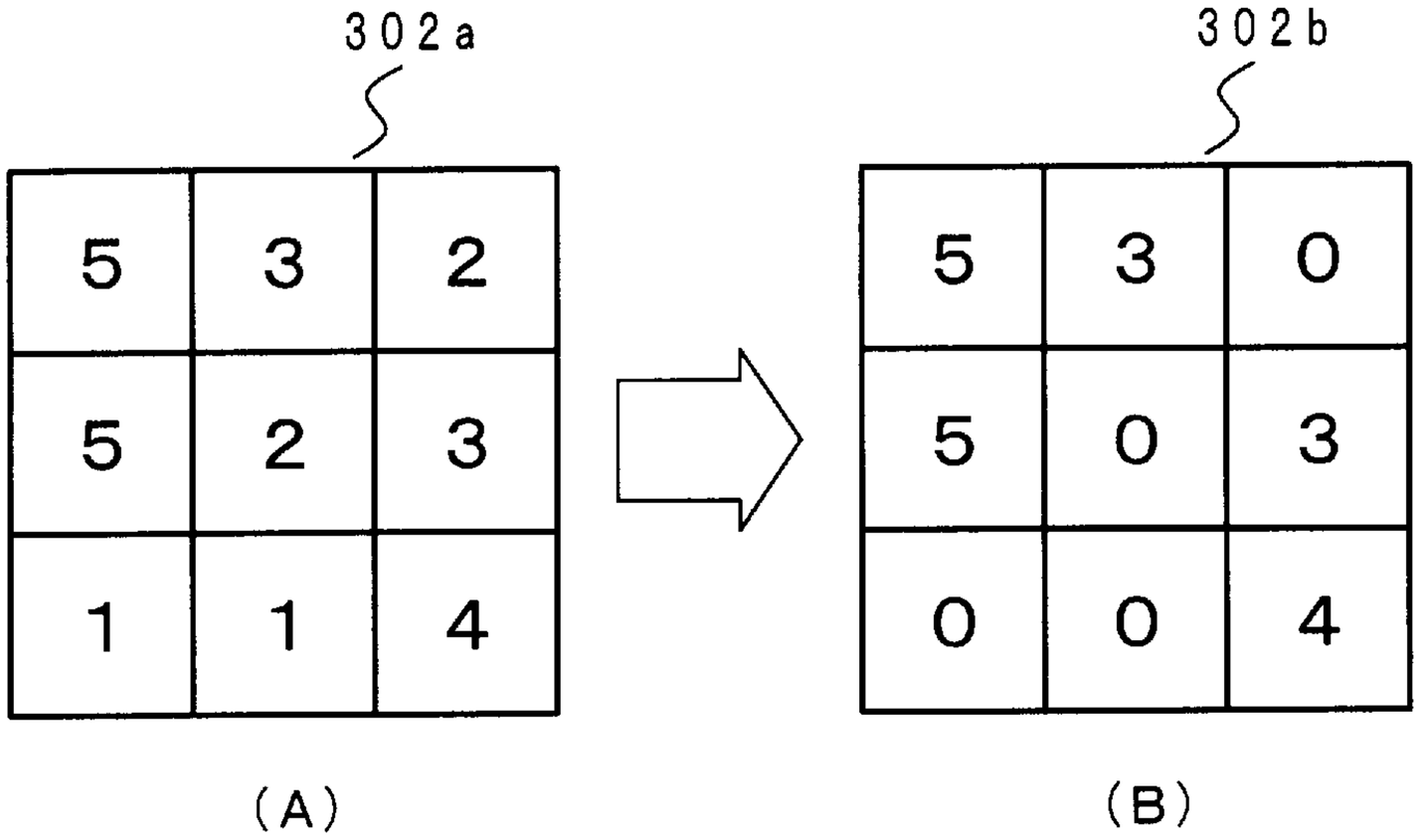
[図6]


開　始
S11
パラメタセットおよびネットワーク構造を入力
S12
層毎の演算器の並列度を決定
S13
層毎の並列度に応じて層毎のパラメタテーブルを生成
S14
パラメタテーブルを最適化
S15
層毎の演算器回路を生成
S16
パラメタテーブル回路を生成
S17
パラメタテーブル回路構成および層毎の演算器回路構成を出力
終　了

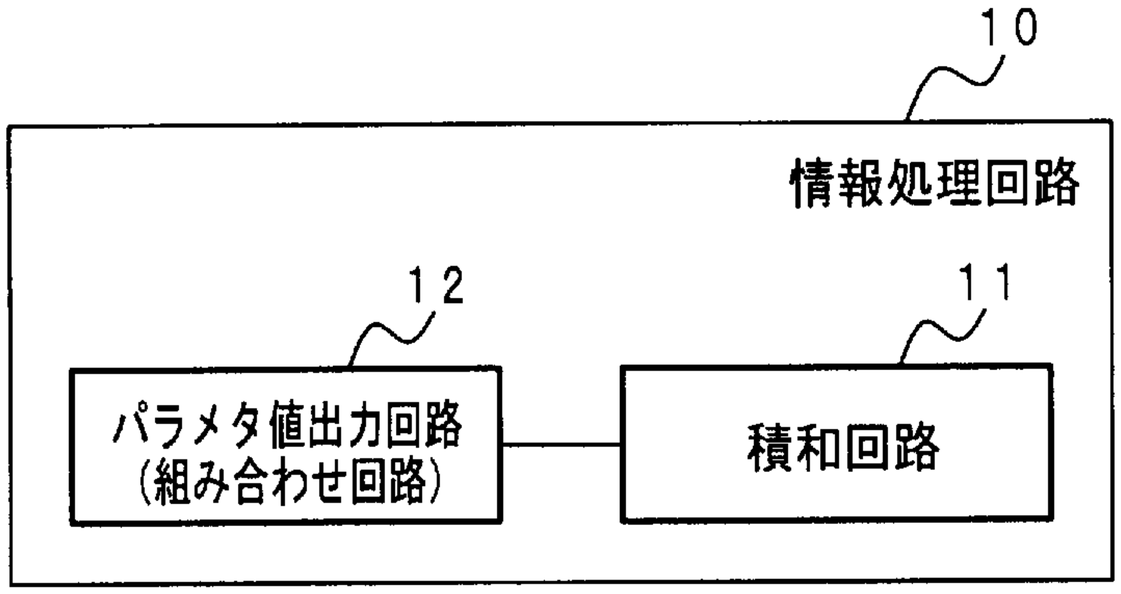
[図7]


開　始
S141
CNNの認識精度を測定
S142
認識精度は第1の基準値以上？
Yes
No
S144
パラメタテーブルの回路
面積は第2の基準値以下？
Yes
No
S143
パラメタテーブルにおける
パラメタ値を変更
終　了

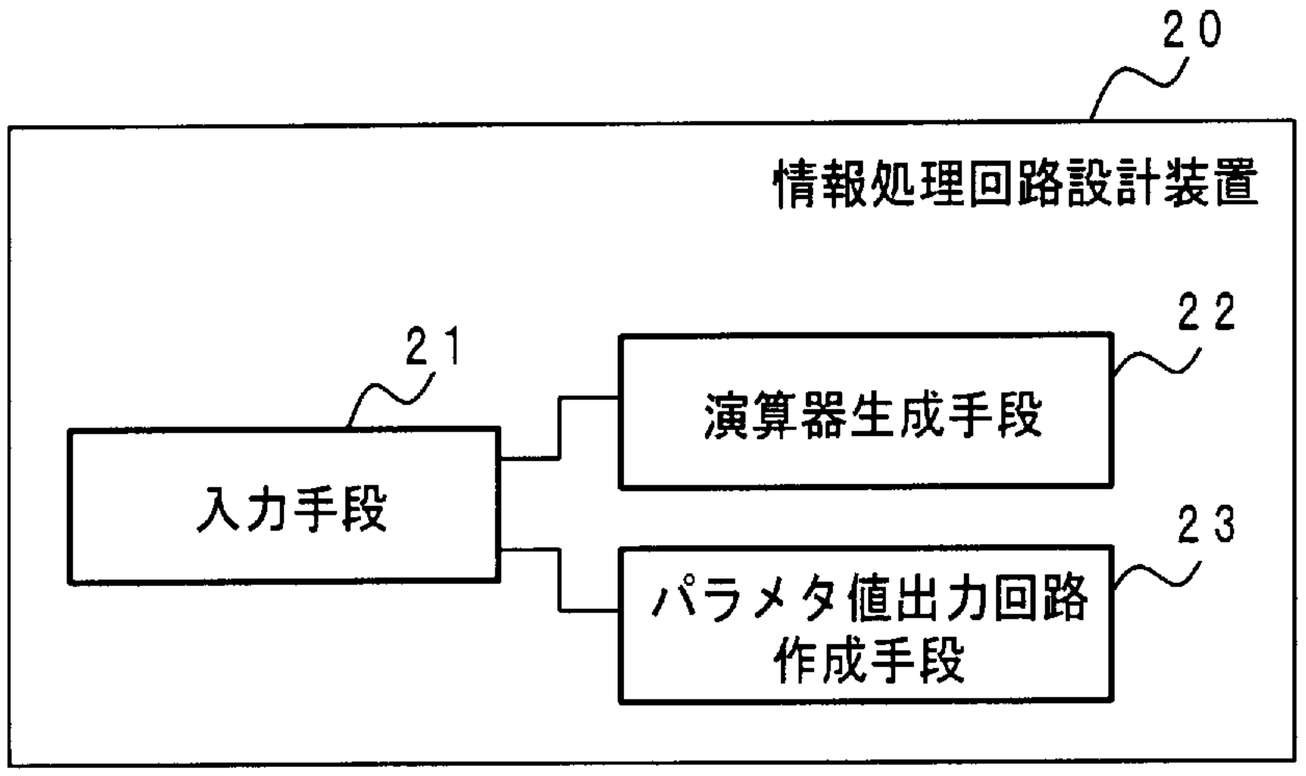
[図8]


302a
5 3 2
5 2 3
1 1 4
(A)
302b
5 3 0
5 0 3
0 0 4
(B)

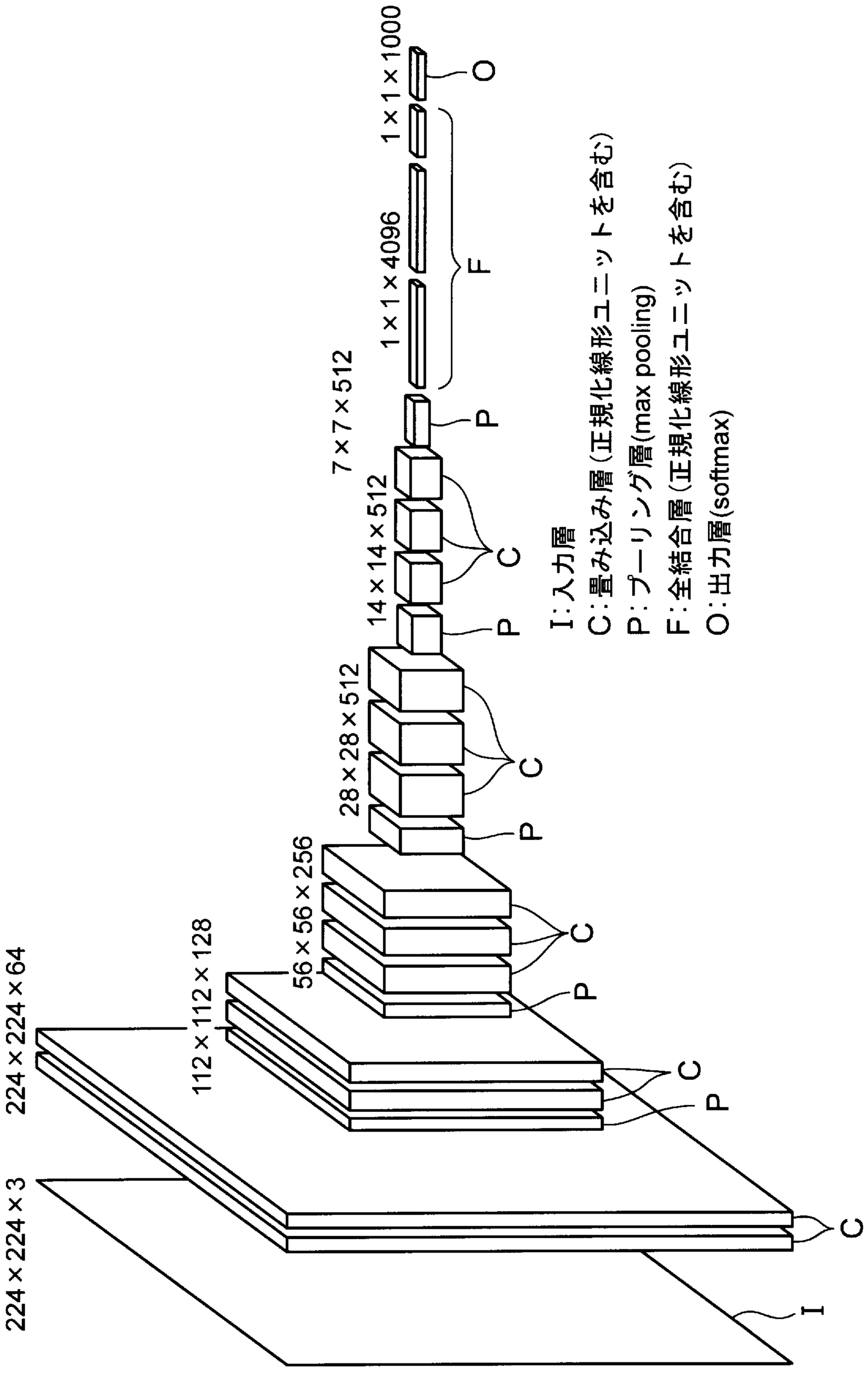
[図9]


10
情報処理回路
12
パラメタ値出力回路
(組み合わせ回路)
11
積和回路

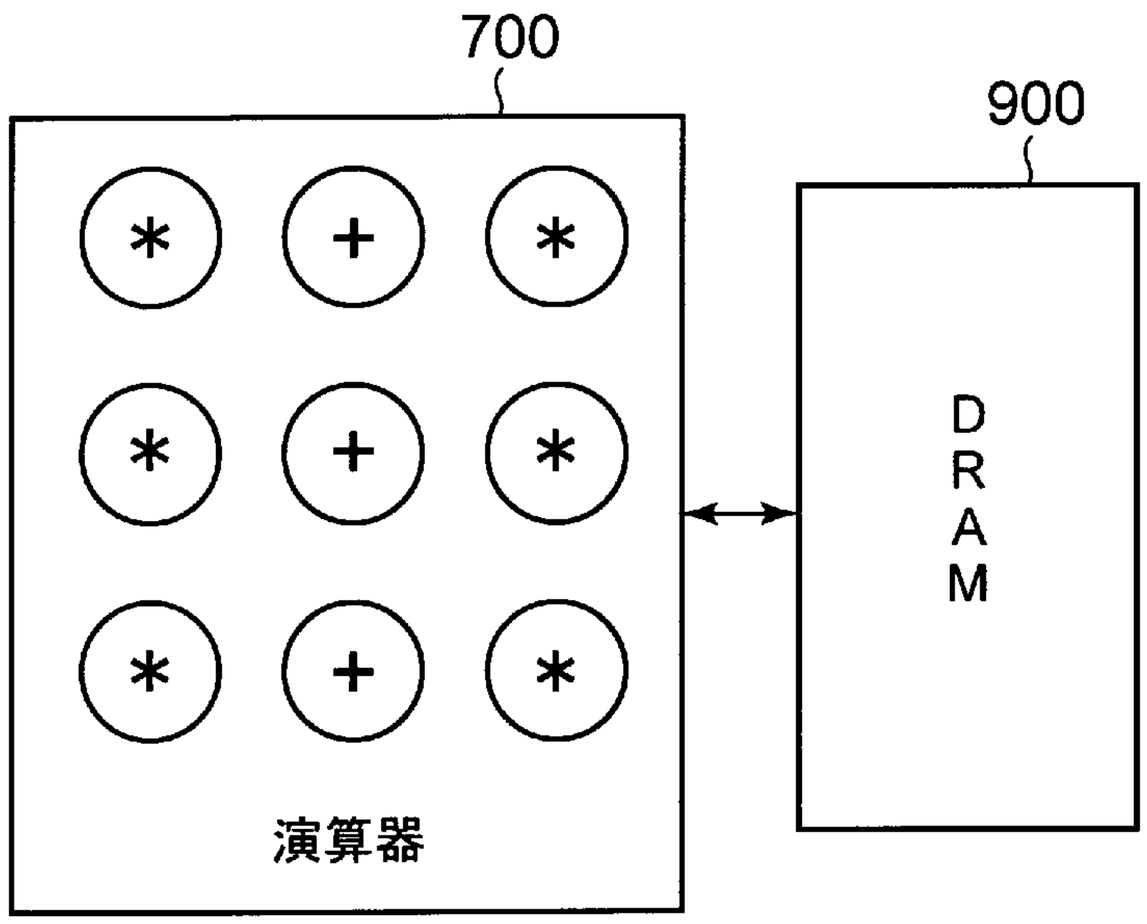
[図10]


20
情報処理回路設計装置
21
入力手段
22
演算器生成手段
23
パラメタ値出力回路
作成手段

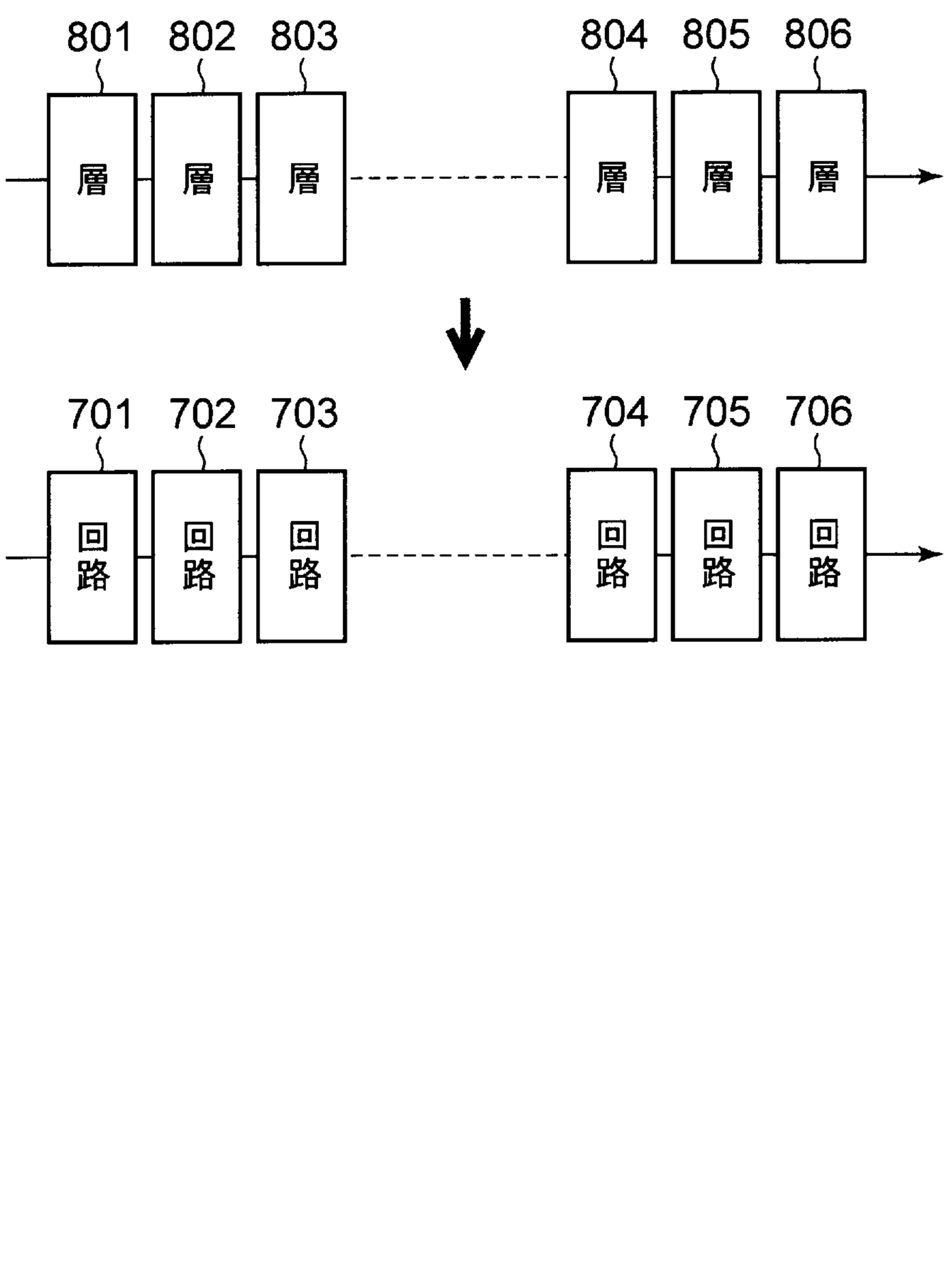
[図11]


224 × 224 × 3
224 × 224 × 64
112 × 112 × 128
56 × 56 × 256
28 × 28 × 512
14 × 14 × 512
7 × 7 × 512
1 × 1 × 4096
1 × 1 × 1000
I
C
P
F
O
I: 入力層
C: 畳み込み層 (正規化線形ユニット を含む)
P: プーリング層(max pooling)
F: 全結合層 (正規化線形ユニット を含む)
O: 出力層(softmax)

[図12]


700
900
DRAM
演算器

[図13]


801 802 803
804 805 806
層
701 702 703
704 705 706
回路

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2019/042927

A. CLASSIFICATION OF SUBJECT MATTER
Int. Cl. G06N3/063(2006.01)i, G06F17/16(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. G06N3/063, G06F17/16

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)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X A	JP 2018-132830 A (LEAPMIND INC.) 23 August 2018, paragraphs [0003], [0004], [0009]-[0012], [0030]-[0032] (Family: none)	1, 4-5, 10-11, 14-15 2-3, 6-9, 12-13, 16-17
A	JP 2004-86374 A (RICOH CO., LTD.) 18 March 2004, paragraphs [0033]-[0035], [0043] (Family: none)	1-17
A	JP 2018-124754 A (NIPPON TELEGRAPH AND TELEPHONE CORP.) 09 August 2018, paragraphs [0094]-[0098], fig. 12 (Family: none)	1-17

☒ 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
19.12.2019	07.01.2020
Name and mailing address of the ISA/ Japan Patent Office 3-4-3, Kasumigaseki, Chiyoda-ku, Tokyo 100-8915, Japan	Authorized officer Telephone No.

Form PCT/ISA/210 (second sheet) (January 2015)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2019/042927

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WHATMOUGH Pau l N., et al., "FIXYNN: Efficient Hardware for Mobile Computer Vision via Transfer Learning" arXiv [online], Cornell University Library, 27 February 2019, [retrieved on 19 December 2019], Internet: URL:https://arxiv.org/pdf/1902.11128.pdf	1-17

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ２０１９／０４２９２７

Ａ．　発明の属する分野の分類（国際特許分類（ＩＰＣ））

Int.Cl.　G06N3/063(2006.01)i, G06F17/16(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.　G06N3/063, G06F17/16

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）

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	JP 2018-132830 A（ＬｅａｐＭｉｎｄ株式会社）2018.08.23, 段落[0003]-[0004],[0009]-[0012],[0030]-[0032] （ファミリーなし）	1,4-5,10-11, 14-15
A		2-3,6-9, 12-13,16-17
A	JP 2004-86374 A（株式会社リコー）2004.03.18, 段落[0033]-[0035],[0043] （ファミリーなし）	1-17

☑　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 １９．１２．２０１９	国際調査報告の発送日 ０７．０１．２０２０
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　６２９６ 加藤　優一 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
A	JP 2018-124754 A（日本電信電話株式会社）2018.08.09, 段落[0094]-[0098],図 12 (ファミリーなし)	1-17
A	WHATMOUGH Paul N., et al., “FIXYNN: EFFICIENT HARDWARE FOR MOBILE COMPUTER VISION VIA TRANSFER LEARNING”, arXiv [online], Cornell University Library, 2019.02.27, ［検索日 2019.12.19], インターネット＜ＵＲＬ：https://arxiv.org/pdf/1902.11128.pdf＞	1-17