

(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
2022年9月29日(29.09.2022)



(10) 国際公開番号

WO 2022/201352 A1

(51) 国際特許分類:
G06N 20/00 (2019.01)

(21) 国際出願番号: PCT/JP2021/012193

(22) 国際出願日: 2021年3月24日(24.03.2021)

(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:三菱電機株式会社(MITSUBISHI ELECTRIC CORPORATION) [JP/JP]; 〒1008310 東京都千代田区丸の内二丁目7番3号 Tokyo (JP).

(72) 発明者: 出口 昌弘 (DEGUCHI, Masahiro); 〒1008310 東京都千代田区丸の内二丁目7番3号 三菱電機株式会社内 Tokyo (JP). 水

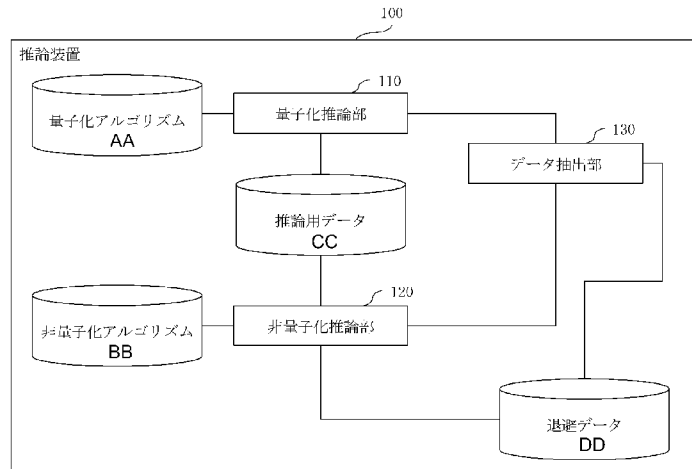
口 武尚(MIZUGUCHI, Takehisa); 〒1008310 東京都千代田区丸の内二丁目7番3号 三菱電機株式会社内 Tokyo (JP).

(74) 代理人: 弁理士法人クロスボーダー特許事務所 (CROSS-BORDER PATENT FIRM); 〒2470056 神奈川県鎌倉市大船二丁目17番10号3階 Kanagawa (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, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY,

(54) Title: INFERENCE DEVICE, INFERENCE METHOD, AND INFERENCE PROGRAM

(54) 発明の名称: 推論装置、推論方法、及び、推論プログラム



100 Inference device
110 Quantization inference unit
120 Non-quantization inference unit
130 Data extraction unit
AA Quantization algorithm
BB Non-quantization algorithm
CC Inference data
DD Evacuated data

(57) Abstract: An inference device (100) is provided with a quantization inference unit (110), a non-quantization inference unit (120), and a feature data extraction unit (130). The quantization inference unit (110) performs, by using inference data, one or more quantization calculations based on a machine learning method. The non-quantization inference unit (120) performs, by using the inference data, at least one of one or more non-quantization calculations corresponding to the one or more individual quantization calculations. When overflow occurs in at least one of the one or more quantization

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))

calculations, the feature data extraction unit (130) extracts quantization feature data corresponding to each quantization calculation in which overflow has occurred, and non-quantization feature data corresponding to a non-quantization calculation corresponding to each quantization calculation in which overflow has occurred.

(57) 要約：推論装置(100)は、量子化推論部(110)と非量子化推論部(120)と特徴データ抽出部(130)とを備える。量子化推論部(110)は、推論用データを用いて機械学習の手法に基づいた少なくとも1回の量子化演算を実行する。非量子化推論部(120)は、推論用データを用いて少なくとも1回の量子化演算それぞれに対応する少なくとも1回の非量子化演算の少なくともいずれかを実行する。特徴データ抽出部(130)は、少なくとも1回の量子化演算の少なくともいずれかにおいてオーバーフローが発生した場合に、オーバーフローが発生した量子化演算の各々に対応する量子化特徴データと、オーバーフローが発生した量子化演算の各々に対応する非量子化演算に対応する非量子化特徴データとを抽出する。

明 細 書

発明の名称：推論装置、推論方法、及び、推論プログラム

技術分野

[0001] 本開示は、推論装置、推論方法、及び、推論プログラムに関する。

背景技術

[0002] ディープラーニングの推論環境を組込機器等のリソース制約がある装置に搭載する場合、ディープラーニングに関する演算を軽量化する必要がある。軽量化する手法の一つとして、浮動小数点演算を固定小数点演算又は整数演算に置き換える技術がある。当該技術は、一般的に量子化と呼ばれる。特許文献1は、推論を高速化するための量子化を比較的高い精度で実施するために統計的手法を用いてデータ分布を推定する技術を開示している。

先行技術文献

特許文献

[0003] 特許文献1：特開2018-010618号公報

発明の概要

発明が解決しようとする課題

[0004] 特許文献1が開示する技術によれば、量子化において、学習用データセット等を用いた場合においてオーバーフローが発生しないよう、推論に関する演算のパラメータ等が調整される。しかしながら、学習用データセット等があらゆる推論用データを網羅することはできない。そのため、特許文献1が開示する技術によれば、どのような推論用データを用いた場合であってもオーバーフローが発生しないことを保証することはできず、推論用データによってはオーバーフローが発生し得るという課題がある。ここで、推論用データは機密情報であることもあるため、ある推論用データを用いて推論を実行した際にオーバーフローが発生した場合であっても、発生したオーバーフローを当該ある推論用データを用いて解析することができるとは限らない。

[0005] 本開示は、ある推論用データを用いて推論を実行した際にオーバーフロー

が発生した場合において、発生したオーバーフローを解析するためのデータであって当該ある推論用データとは異なるデータを取得することを目的とする。

課題を解決するための手段

[0006] 本開示に係る推論装置は、

推論用データを用いて機械学習の手法に基づいた少なくとも1回の量子化演算を実行する量子化推論部と、

前記推論用データを用いて前記少なくとも1回の量子化演算それぞれに対応する少なくとも1回の非量子化演算の少なくともいずれかを実行する非量子化推論部と

を備える推論装置であって、

前記少なくとも1回の量子化演算それぞれは、前記少なくとも1回の量子化演算それぞれの特徴を示す少なくとも1つの量子化特徴データそれぞれに応じた演算であり、

前記少なくとも1回の非量子化演算それぞれは、前記少なくとも1回の非量子化演算それぞれの特徴を示す少なくとも1つの非量子化特徴データそれぞれに応じた演算であり、

前記推論装置は、さらに、

前記少なくとも1回の量子化演算の少なくともいずれかにおいてオーバーフローが発生した場合に、オーバーフローが発生した量子化演算の各々に対応する量子化特徴データと、オーバーフローが発生した量子化演算の各々に対応する非量子化演算に対応する非量子化特徴データとを抽出する特徴データ抽出部

を備える。

発明の効果

[0007] 本開示によれば、機械学習の手法に基づいた推論においてオーバーフローが発生した場合に、特徴データ抽出部が、発生したオーバーフローに係のある量子化特徴データと非量子化特徴データとを抽出する。ここで、量子化

特徴データと非量子化特徴データとの各々は推論用データとは異なるデータである。そのため、本開示によれば、ある推論用データを用いて推論を実行した際にオーバーフローが発生した場合において、発生したオーバーフローを解析するためのデータであって当該ある推論用データとは異なるデータを取得することができる。

図面の簡単な説明

- [0008] [図1]実施の形態1に係る推論装置100の構成例を示す図。
[図2]実施の形態1に係るプロセスの優先度を説明する図。
[図3]実施の形態1に係る推論装置100のハードウェア構成例を示す図。
[図4]実施の形態1に係る推論装置100の動作を示すフローチャート。
[図5]実施の形態1に係る推論装置100の動作を説明する図。
[図6]実施の形態1の変形例に係る推論装置100のハードウェア構成例を示す図。
[図7]実施の形態1の変形例に係る推論装置100のハードウェア構成例を示す図。
[図8]実施の形態2に係る推論装置100の構成例を示す図。
[図9]実施の形態2に係る推論装置100の動作を示すフローチャート。

発明を実施するための形態

- [0009] 実施の形態の説明及び図面において、同じ要素及び対応する要素には同じ符号を付している。同じ符号が付された要素の説明は、適宜に省略又は簡略化する。図中の矢印はデータの流れ又は処理の流れを主に示している。また、「部」を、「回路」、「工程」、「手順」、「処理」又は「サーキットリ」に適宜読み替えてもよい。

- [0010] 実施の形態1.

以下、本実施の形態について、図面を参照しながら詳細に説明する。

- [0011] ***構成の説明***

図1は、本実施の形態に係る推論装置100の構成例を示している。推論装置100は、本図に示すように、量子化推論部110と、非量子化推論部

120と、データ抽出部130とを備える。

推論装置100は、典型的には組込みシステムの一部である。なお、推論装置100が推論プロセス管理部を備え、推論プロセス管理部が量子化推論部110と非量子化推論部120とを制御する構成であってもよい。

[0012] 量子化推論部110は、量子化推論プロセスを実行する、即ち、機械学習の手法に基づいた少なくとも1回の量子化演算を実行する。少なくとも1回の量子化演算それぞれは、少なくとも1つの量子化特徴データそれぞれに応じた演算である。少なくとも1つの量子化特徴データそれぞれは、少なくとも1回の量子化演算それぞれの特徴を示しており、また、少なくとも1回の量子化演算それぞれに対応するパラメータを含んでもよい。また、量子化推論部110は推論用データを用いる。推論用データは推論を実行する際に学習済モデルに入力されるデータである。

量子化推論プロセスは量子化済推論プロセスとも呼ばれる。量子化推論プロセスにおいて量子化アルゴリズムが実行される。量子化アルゴリズムは量子化推論アルゴリズムとも呼ばれる。また、量子化推論部110は、量子化推論プロセスにおいてオーバーフロー等が発生した場合に、発生したオーバーフロー等に関する情報を記録する。

[0013] 非量子化推論部120は、非量子化推論プロセスを実行する、即ち、少なくとも1回の量子化演算それぞれに対応する少なくとも1回の非量子化演算の少なくともいずれかを実行する。少なくとも1回の量子化演算と、少なくとも1回の非量子化演算とは、同一の機械学習の手法に基づいた演算である。機械学習の手法がニューラルネットワークである場合において、各量子化演算と各非量子化演算とはニューラルネットワークの各レイヤの演算である。少なくとも1回の非量子化演算それぞれは、少なくとも1つの非量子化特徴データそれぞれに応じた演算である。少なくとも1つの非量子化特徴データそれぞれは、少なくとも1回の非量子化演算それぞれの特徴を示しており、また、少なくとも1回の非量子化演算それぞれに対応するパラメータを含んでもよい。非量子化推論部120は推論用データと後述の退避データとを

用いる。

非量子化推論プロセスは未量子化推論プロセスとも呼ばれる。非量子化推論プロセスにおいて非量子化アルゴリズムが実行される。非量子化アルゴリズムは非量子化推論アルゴリズムとも呼ばれる。非量子化アルゴリズムは学習用のデータを用いて学習を実行した結果得られたアルゴリズムである。量子化アルゴリズムと、非量子化アルゴリズムとは基本的には同じである。量子化アルゴリズムは、非量子化アルゴリズムを、量子化に対応するよう適宜変更したものである。量子化は、典型的には浮動小数点演算を固定小数点演算又は整数演算に置き換えることである。また、量子化アルゴリズムと、非量子化アルゴリズムとの各々を、推論装置 100 が生成してもよく、他の装置が生成してもよい。また、非量子化推論部 120 は、量子化推論プロセスにおいてオーバーフローが発生した際に用いられた入力データ等に応じて演算を実行する。

[0014] データ抽出部 130 は、量子化推論プロセスにおいてオーバーフローが発生した場合に、量子化推論プロセスと非量子化推論プロセスとから退避データを抽出する。退避データは、量子化推論プロセスにおけるオーバーフローを解析する際に用いられるデータであり、具体例として、入力データと特徴データとから成る。入力データは、少なくとも 1 回の量子化演算の各々と少なくとも 1 回の非量子化演算の各々を実行する際に入力されるデータである。特徴データは、量子化演算において活用されるデータである。特徴データは、量子化特徴データと非量子化特徴データとの総称であり、少なくとも 1 回の量子化演算と少なくとも 1 回の非量子化演算との各々の演算ごとに存在し、具体例として演算における入力データの振れ幅を表す。入力データの振れ幅は、入力データとして想定されるデータが示す値の最小値から最大値までの範囲である。データ抽出部 130 は、少なくとも 1 回の量子化演算の少なくともいずれかにおいてオーバーフローが発生した場合に、オーバーフローが発生した量子化演算の各々に対応する量子化特徴データと、オーバーフローが発生した量子化演算の各々に対応する非量子化演算に対応する非量

量子化特徴データとを抽出する。

機械学習の手法がディープラーニングである場合において、特徴データ抽出部130は、量子化特徴データとしてオーバーフローが発生した量子化演算に対応するレイヤについてのパラメータを示すデータを抽出してもよく、非量子化特徴データとしてオーバーフローが発生した量子化演算に対応する非量子化演算に対応するレイヤについてのパラメータを示すデータを抽出してもよい。

[0015] 図2は、推論装置100が実行する各プロセスの優先度を示している。本図に示すように、量子化推論プロセスの優先度は、非量子化推論プロセスの優先度よりも高い。また、量子化推論プロセスの優先度よりも優先度が低く、かつ、非量子化推論プロセスの優先度よりも優先度が高いプロセスがあってもよい。なお、他プロセスという表記を挟む丸括弧は、他プロセスがあってもなくてもよいことを示している。

[0016] 図3は、本実施の形態に係る推論装置100のハードウェア構成例を示している。推論装置100は、コンピュータから成る。推論装置100は、複数のコンピュータから成ってもよい。

[0017] 推論装置100は、本図に示すように、プロセッサ11と、メモリ12と、補助記憶装置13と、入出力IF (Interface) 14と、通信装置15等のハードウェアを備えるコンピュータである。これらのハードウェアは、信号線19を介して適宜接続されている。

[0018] プロセッサ11は、演算処理を行うIC (Integrated Circuit) であり、かつ、コンピュータが備えるハードウェアを制御する。プロセッサ11は、具体例として、CPU (Central Processing Unit)、DSP (Digital Signal Processor)、又はGPU (Graphics Processing Unit) である。

推論装置100は、プロセッサ11を代替する複数のプロセッサを備えてもよい。複数のプロセッサは、プロセッサ11の役割を分担する。

[0019] メモリ 1 2 は、典型的には、揮発性の記憶装置である。メモリ 1 2 は、主記憶装置又はメインメモリとも呼ばれる。メモリ 1 2 は、具体例として、RAM (Random Access Memory) である。メモリ 1 2 に記憶されたデータは、必要に応じて補助記憶装置 1 3 に保存される。

[0020] 補助記憶装置 1 3 は、典型的には、不揮発性の記憶装置である。補助記憶装置 1 3 は、具体例として、ROM (Read Only Memory)、HDD (Hard Disk Drive)、又はフラッシュメモリである。補助記憶装置 1 3 に記憶されたデータは、必要に応じてメモリ 1 2 にロードされる。

メモリ 1 2 及び補助記憶装置 1 3 は一体的に構成されていてもよい。

[0021] 入出力 I F 1 4 は、入力装置及び出力装置が接続されるポートである。入出力 I F 1 4 は、具体例として、USB (Universal Serial Bus) 端子である。入力装置は、具体例として、カメラ、キーボード及びマウスである。出力装置は、具体例として、ディスプレイである。

[0022] 通信装置 1 5 は、レシーバ及びトランスミッタである。通信装置 1 5 は、具体例として、通信チップ又はNIC (Network Interface Card) である。

[0023] 推論装置 1 0 0 の各部は、他の装置等と通信する際に、通信装置 1 5 を適宜用いてもよい。推論装置 1 0 0 の各部は、入出力 I F 1 4 を介してデータを受け付けてもよく、また、通信装置 1 5 を介してデータを受け付けてもよい。

[0024] 補助記憶装置 1 3 は、推論プログラムを記憶している。推論プログラムは、推論装置 1 0 0 が備える各部の機能をコンピュータに実現させるプログラムである。推論プログラムは、メモリ 1 2 にロードされて、プロセッサ 1 1 によって実行される。推論装置 1 0 0 が備える各部の機能は、ソフトウェアにより実現される。

また、補助記憶装置 1 3 は、OS (Operating System) を記憶している。OS の少なくとも一部は、メモリ 1 2 にロードされて、プ

ロセッサ 11 によって実行される。つまり、プロセッサ 11 は、OS を実行しながら、推論プログラムを実行する。

[0025] 推論プログラムを実行する際に用いられるデータと、推論プログラムを実行することによって得られるデータと等は、記憶装置に適宜記憶される。推論装置 100 の各部は、適宜記憶装置を利用する。記憶装置は、具体例として、メモリ 12 と、補助記憶装置 13 と、プロセッサ 11 内のレジスタと、プロセッサ 11 内のキャッシュメモリとの少なくとも 1 つから成る。なお、データと情報とは、同等の意味を有することもある。記憶装置は、コンピュータと独立したものであってもよい。

メモリ 12 及び補助記憶装置 13 の機能は、他の記憶装置によって実現されてもよい。

[0026] 推論プログラムは、コンピュータが読み取り可能な不揮発性の記録媒体に記録されていてもよい。不揮発性の記録媒体は、具体例として、光ディスク又はフラッシュメモリである。推論プログラムは、プログラムプロダクトとして提供されてもよい。

[0027] ***動作の説明***

推論装置 100 の動作手順は、推論方法に相当する。また、推論装置 100 の動作を実現するプログラムは、推論プログラムに相当する。

[0028] 図 4 は、推論装置 100 の動作の一例を示すフローチャートである。また、図 5 は、推論装置 100 の動作の一例を模式的に説明する図である。図 5 において、平行四辺形は、出力データを除いて演算において用いられるデータを表現している。図 4 及び図 5 を参照して推論装置 100 の動作を説明する。なお、本実施の形態は、量子化したデータを用いることもできる機械学習の手法に対して適用することができるが、説明の便宜上、推論装置 100 の動作の説明において機械学習の手法はディープラーニングとする。

[0029] (ステップ S101)

量子化推論部 110 の量子化推論プロセスが起動される。

[0030] (ステップ S102)

推論を開始する指示がある場合、推論装置１００はステップＳ１０３に進む。それ以外の場合、推論装置１００は本ステップの処理を再度実行する。

[0031] (ステップＳ１０３)

量子化推論部１１０は、推論用データを用いて推論処理を開始する。推論用データは、推論を開始する指示と併せて推論装置１００が受け取ったデータであってもよい。

[0032] (ステップＳ１０４)

量子化推論部１１０は、対象レイヤについてのレイヤ演算を実行する。ここで、レイヤはディープラーニングにおけるレイヤであり、ステップＳ１０４からステップＳ１０６から成るループ処理においてレイヤ演算を初めて実行する場合に先頭のレイヤを対象レイヤとし、それ以外の場合に１回前に実行したレイヤ演算における対象レイヤの次のレイヤを対象レイヤとする。

図５における演算がレイヤ演算に相当する。また、量子化推論部１１０は、演算の直下に示されている太い矢印の左側に示されているデータを入力データとしてレイヤ演算を実行し、当該矢印の右側に示されているデータを出力する。

[0033] (ステップＳ１０５)

量子化推論部１１０は、ステップＳ１０４の処理においてオーバーフローが発生したか否かを確認する。オーバーフローが発生したか否かを確認する方法は、具体例として、ＣＰＵのフラグを参照する方法、演算に関する変数を演算の前後に確認する方法、又は、ＦＰＧＡ等の独自の回路により確認する方法である。

オーバーフローが発生した場合、推論装置１００はステップＳ１０９に進む。それ以外の場合、推論装置１００はステップＳ１０６に進む。

[0034] (ステップＳ１０６)

全てのレイヤ演算が終了した場合、推論装置１００はステップＳ１０７に進む。それ以外の場合、推論装置１００はステップＳ１０４に進む。

[0035] (ステップＳ１０７)

量子化推論部 110 は、推論を指示した者に推論の結果を適宜通知する。

[0036] (ステップ S 108)

少なくともいずれかのレイヤ演算においてオーバーフローが発生した場合、推論装置 100 はステップ S 111 に進む。それ以外の場合、推論装置 100 はステップ S 102 に戻る。

[0037] (ステップ S 109)

データ抽出部 130 は、対象レイヤに対応する退避データを退避する。具体例として、データ抽出部 130 は、対象レイヤについてのデータを抽出し、抽出したデータを退避データとして退避する。対象レイヤについてのデータは、具体例として、対象レイヤに対する入力データと、対象レイヤにおける変数データと、対象レイヤにおける特徴データとである。ただし、対象レイヤが先頭のレイヤである場合、データ抽出部 130 は、入力データが推論用データであるため、退避データとして入力データを抽出しない。なお、データ抽出部 130 は推論用データを退避してもよい。ここで、推論用データは先頭のレイヤに入力されるデータである。

図 5 は、対象レイヤに対応する退避データとして、オーバーフローが発生した対象レイヤについてのデータと、対象レイヤの 1 つ前のレイヤについてのデータとをデータ抽出部 130 が退避する様子を示している。なお、データ抽出部 130 が退避するデータは、対象レイヤについてのデータのみであってもよく、対象レイヤから対象レイヤの n (n は自然数) 個前のレイヤまでの各々のレイヤについてのデータであってもよい。

[0038] (ステップ S 110)

量子化推論部 110 は、飽和演算を実行し、その後、ステップ S 106 に進む。

[0039] (ステップ S 111)

非量子化推論部 120 の非量子化推論プロセスが起動される。

以下、推論装置 100 は、ステップ S 102 以降の処理と、ステップ S 121 以降の処理とを並列に実行する。

[0040] (ステップS 1 2 1)

非量子化推論部 1 2 0 は、推論用データを用いて非量子化推論プロセスを実行する。

なお、図 5 において、非量子化推論部 1 2 0 が非量子化推論プロセスを最後まで実行することによって出力データを出力しているが、非量子化推論部 1 2 0 は、量子化推論プロセスにおいてオーバーフローが発生したレイヤに対応するレイヤまでのレイヤ演算を実行すれば十分である。

[0041] (ステップS 1 2 2)

データ抽出部 1 3 0 は、退避データとして、オーバーフローが発生したレイヤに対応するレイヤについてのデータを抽出する。ここで、データ抽出部 1 3 0 は、量子化推論プロセスにおいて退避したデータに対応するレイヤと同一のレイヤについてのデータを退避する。具体例として、図 5 に示すように、データ抽出部 1 3 0 が量子化推論プロセスにおいてオーバーフローが発生したレイヤと当該レイヤの 1 つ前のレイヤとの各々についてのデータを退避した場合、データ抽出部 1 3 0 は、オーバーフローが発生したレイヤに対応するレイヤとオーバーフローが発生したレイヤの 1 つ前のレイヤとの各々に対応する非量子化推論プロセスのレイヤについてのデータを退避データとして退避する。

データ抽出部 1 3 0 が退避した量子化推論プロセスにおけるデータと非量子化推論プロセスにおけるデータとは推論装置 1 0 0 の外部に出力されてもよい。エンジニア等は、出力されたデータを入手し、入手したデータに基づいて量子化推論プロセスのパラメータを再設定してもよく、入手したデータに基づいてソースコード内の量子化推論プロセスに関するパラメータを変更してもよい。

[0042] (ステップS 1 2 3)

非量子化推論部 1 2 0 の非量子化推論プロセスは終了される。

[0043] ***実施の形態 1 の効果の説明***

以上のように、本実施の形態によれば、推論用データが機密情報である場

合であっても、推論用データを抽出せず、オーバーフローが発生した演算に関するデータのみを抽出する。ここで、オーバーフローが発生した演算に関するデータは、推論用データに存在する機密性が排除されたデータである。そのため、本実施の形態によれば、推論環境における精度を維持しつつ、機密性を保持したまま量子化演算において発生したオーバーフローを解析することができる。そのため、本実施の形態によれば、発生したオーバーフローに対処するための学習済モデルの改良に必要なデータを取得することができる。

また、本実施の形態によれば、学習用データ等のみならず、実際の推論用データに基づいて量子化を伴う機械学習におけるパラメータ等を調整することができる。

[0044] ***他の構成***

<変形例 1>

量子化推論部 110 は、オーバーフローが発生した場合に推論を途中で打ち切ってもよい。具体例として、量子化推論部 110 は、少なくとも 1 回の量子化演算のいずれかにおいてオーバーフローが発生した場合に、オーバーフローが発生した演算よりも後の演算を実行しない。

[0045] <変形例 2>

図 6 は、本変形例に係る推論装置 100 のハードウェア構成例を示している。本変形例に係る推論装置 100 は、本図に示すように、オフロードデバイス 16 を備える。

量子化推論プロセスは、オフロードデバイス 16 によって実行されてもよい。オフロードデバイス 16 は具体例として、GPU 又は FPGA である。本変形例において、非量子化推論プロセスはプロセッサ 11 で実行されていてもよく、このとき、プロセッサ 11 とオフロードデバイス 16 との間で適宜通信が実行される。

[0046] <変形例 3>

図 7 は、本変形例に係る推論装置 100 のハードウェア構成例を示してい

る。

推論装置 100 は、プロセッサ 11、プロセッサ 11 とメモリ 12、プロセッサ 11 と補助記憶装置 13、あるいはプロセッサ 11 とメモリ 12 と補助記憶装置 13 に代えて、処理回路 18 を備える。

処理回路 18 は、推論装置 100 が備える各部の少なくとも一部を実現するハードウェアである。

処理回路 18 は、専用のハードウェアであってもよく、また、メモリ 12 に格納されるプログラムを実行するプロセッサであってもよい。

[0047] 処理回路 18 が専用のハードウェアである場合、処理回路 18 は、具体例として、単回路、複合回路、プログラム化したプロセッサ、並列プログラム化したプロセッサ、ASIC (ASIC は Application Specific Integrated Circuit)、FPGA (Field Programmable Gate Array) 又はこれらの組み合わせである。

推論装置 100 は、処理回路 18 を代替する複数の処理回路を備えてもよい。複数の処理回路は、処理回路 18 の役割を分担する。

[0048] 推論装置 100 において、一部の機能が専用のハードウェアによって実現されて、残りの機能がソフトウェア又はファームウェアによって実現されてもよい。

[0049] 処理回路 18 は、具体例として、ハードウェア、ソフトウェア、ファームウェア、又はこれらの組み合わせにより実現される。

プロセッサ 11 とメモリ 12 と補助記憶装置 13 と処理回路 18 とを、総称して「プロセッシングサーキットリー」という。つまり、推論装置 100 の各機能構成要素の機能は、プロセッシングサーキットリーにより実現される。

他の実施の形態に係る推論装置 100 についても、本変形例と同様の構成であってもよい。

[0050] 実施の形態 2.

以下、主に前述した実施の形態と異なる点について、図面を参照しながら説明する。

[0051] ***構成の説明***

図8は、本実施の形態に係る推論装置100の構成例を示している。推論装置100は、本図に示すように再量子化部140を備える。

再量子化部140は、データ抽出部130が抽出したデータに基づいて量子化アルゴリズムを変更することにより再量子化アルゴリズムを生成する。ここで、再量子化アルゴリズムは、再量子化部140によって変更された量子化アルゴリズムである。また、再量子化部140は、再量子化を実行するタイミングを管理する。再量子化は、再量子化アルゴリズムを変更することであり、量子化アルゴリズムを再量子化アルゴリズムに置き換えることを含んでもよい。再量子化部140は、データ抽出部130によって抽出された量子化特徴データと非量子化特徴データとに基づいて、抽出された量子化特徴データに対応する量子化演算である対象演算においてオーバーフローが発生しないよう、対象演算に対応する量子化特徴データを変更する。

本実施の形態に係る量子化推論部110は、再量子化部140が生成した再量子化アルゴリズムを適宜利用し、変更された量子化特徴データに応じた量子化演算を実行する。また、量子化推論部110は、再量子化アルゴリズムと動作している量子化アルゴリズムとを入れ替えるタイミングを管理する。

[0052] ***動作の説明***

図9は、推論装置100の動作の一例を示すフローチャートである。本図を参照して、実施の形態1に係る推論装置100の動作と、本実施の形態に係る推論装置100の動作との差異を主に説明する。

[0053] (ステップS102)

推論装置100は、本ステップの処理を再度実行する代わりに、ステップS201に進む。

[0054] (ステップS108)

推論装置 100 はステップ S 102 に戻る代わりに、ステップ S 201 に進む。

[0055] (ステップ S 201)

再量子化部 140 によって再量子化アルゴリズムが準備されていない場合、量子化推論部 110 はステップ S 102 に進む。それ以外の場合、量子化推論部 110 はステップ S 202 に進む。

[0056] (ステップ S 202)

量子化推論部 110 は、量子化推論プロセスにおける量子化アルゴリズムを再量子化部 140 が準備した再量子化アルゴリズムに入れ替える。

[0057] (ステップ S 221)

再量子化部 140 は、退避データを用いて再量子化を実行することにより再量子化アルゴリズムを生成する。再量子化は、オーバーフローが発生したレイヤにおいてオーバーフローが発生しないように量子化アルゴリズムを調整することである。再量子化部 140 は、具体例として、オーバーフローが発生したレイヤにおける特徴データを変更する。

なお、複数のレイヤにおいてオーバーフローが発生した場合に、再量子化部 140 は、最も早く処理されるレイヤに対応する特徴データのみを調整してもよく、オーバーフローが発生した全てのレイヤの各々に対応する特徴データを一括で調整してもよい。

[0058] (ステップ S 222)

再量子化部 140 は、ステップ S 221 において生成した再量子化アルゴリズムを保存する。

[0059] ***実施の形態 2 の効果の説明***

以上のように、本実施の形態によれば、オーバーフローが発生しないように量子化アルゴリズムを変更する処理を自動的に実行することができる。

[0060] ***他の構成***

<変形例 4>

量子化推論プロセスにおける量子化アルゴリズムを再量子化部 140 が準

備した量子化アルゴリズムに量子化推論部 110 が入れ替えるタイミングは、推論装置 100 を備える組込みシステムの再起動時であってもよい。また、量子化推論部 110 は、次に推論を実行するタイミングを考慮して次に実行する推論に影響がないと判断した際に量子化アルゴリズムを入れ替えるよう動作予約する方式により量子化アルゴリズムを入れ替えてもよい。

[0061] ***他の実施の形態***

前述した各実施の形態の自由な組み合わせ、あるいは各実施の形態の任意の構成要素の変形、もしくは各実施の形態において任意の構成要素の省略が可能である。

また、実施の形態は、実施の形態 1 から 2 で示したものに限定されるものではなく、必要に応じて種々の変更が可能である。フローチャート等を用いて説明した手順は、適宜変更されてもよい。

符号の説明

[0062] 11 プロセッサ、12 メモリ、13 補助記憶装置、14 入出力 I/F、15 通信装置、16 オフロードデバイス、18 処理回路、19 信号線、100 推論装置、110 量子化推論部、120 非量子化推論部、130 データ抽出部、140 再量子化部。

請求の範囲

[請求項1] 推論用データを用いて機械学習の手法に基づいた少なくとも1回の量子化演算を実行する量子化推論部と、

前記推論用データを用いて前記少なくとも1回の量子化演算それぞれに対応する少なくとも1回の非量子化演算の少なくともいずれかを実行する非量子化推論部と
を備える推論装置であって、

前記少なくとも1回の量子化演算それぞれは、前記少なくとも1回の量子化演算それぞれの特徴を示す少なくとも1つの量子化特徴データそれぞれに応じた演算であり、

前記少なくとも1回の非量子化演算それぞれは、前記少なくとも1回の非量子化演算それぞれの特徴を示す少なくとも1つの非量子化特徴データそれぞれに応じた演算であり、

前記推論装置は、さらに、

前記少なくとも1回の量子化演算の少なくともいずれかにおいてオーバーフローが発生した場合に、オーバーフローが発生した量子化演算の各々に対応する量子化特徴データと、オーバーフローが発生した量子化演算の各々に対応する非量子化演算に対応する非量子化特徴データとを抽出する特徴データ抽出部
を備える推論装置。

[請求項2] 前記少なくとも1つの量子化特徴データそれぞれは、前記少なくとも1回の量子化演算それぞれに対応するパラメータを含み、

前記少なくとも1つの非量子化特徴データそれぞれは、前記少なくとも1回の非量子化演算それぞれに対応するパラメータを含む請求項1に記載の推論装置。

[請求項3] 前記機械学習の手法は、ディープラーニングであり、

前記特徴データ抽出部は、

前記量子化特徴データとして、オーバーフローが発生した量子化演

算に対応するレイヤについてのパラメータを示すデータを抽出し、

前記非量子化特徴データとして、オーバーフローが発生した量子化演算に対応する非量子化演算に対応するレイヤについてのパラメータを示すデータを抽出する請求項 1 又は 2 に記載の推論装置。

[請求項4]

前記推論装置は、さらに、

抽出された量子化特徴データと非量子化特徴データとに基づいて、抽出された量子化特徴データに対応する量子化演算である対象演算においてオーバーフローが発生しないよう、前記対象演算に対応する量子化特徴データを変更する再量子化部を備え、

前記量子化推論部は、変更された量子化特徴データに応じた量子化演算を実行する請求項 1 から 3 のいずれか 1 項に記載の推論装置。

[請求項5]

推論用データを用いて機械学習の手法に基づいた少なくとも 1 回の量子化演算を実行し、

前記推論用データを用いて前記少なくとも 1 回の量子化演算それぞれに対応する少なくとも 1 回の非量子化演算の少なくともいずれかを実行する推論方法であって、

前記少なくとも 1 回の量子化演算それぞれは、前記少なくとも 1 回の量子化演算それぞれの特徴を示す少なくとも 1 つの量子化特徴データそれぞれに応じた演算であり、

前記少なくとも 1 回の非量子化演算それぞれは、前記少なくとも 1 回の非量子化演算それぞれの特徴を示す少なくとも 1 つの非量子化特徴データそれぞれに応じた演算であり、

前記少なくとも 1 回の量子化演算の少なくともいずれかにおいてオーバーフローが発生した場合に、オーバーフローが発生した量子化演算の各々に対応する量子化特徴データと、オーバーフローが発生した量子化演算の各々に対応する非量子化演算に対応する非量子化特徴データとを抽出する推論方法。

[請求項6] 推論用データを用いて機械学習の手法に基づいた少なくとも1回の量子化演算を実行する量子化推論処理と、

前記推論用データを用いて前記少なくとも1回の量子化演算それぞれに対応する少なくとも1回の非量子化演算の少なくともいずれかを実行する非量子化推論処理と

をコンピュータである推論装置に実行させる推論プログラムであって、

前記少なくとも1回の量子化演算それぞれは、前記少なくとも1回の量子化演算それぞれの特徴を示す少なくとも1つの量子化特徴データそれぞれに応じた演算であり、

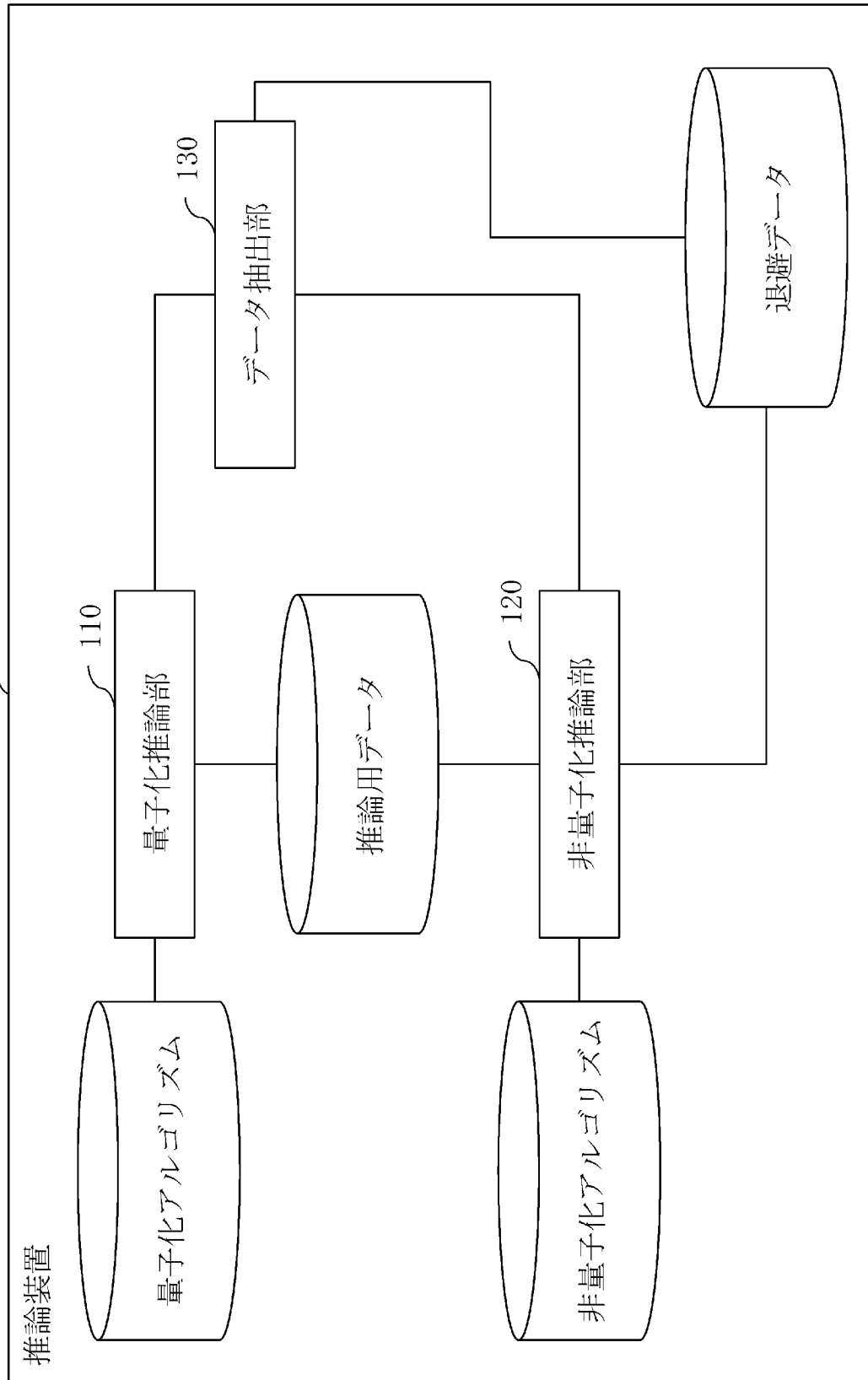
前記少なくとも1回の非量子化演算それぞれは、前記少なくとも1回の非量子化演算それぞれの特徴を示す少なくとも1つの非量子化特徴データそれぞれに応じた演算であり、

前記推論プログラムは、さらに、

前記少なくとも1回の量子化演算の少なくともいずれかにおいてオーバーフローが発生した場合に、オーバーフローが発生した量子化演算の各々に対応する量子化特徴データと、オーバーフローが発生した量子化演算の各々に対応する非量子化演算に対応する非量子化特徴データとを抽出する特徴データ抽出処理を前記推論装置に実行させる推論プログラム。

[図1]

図1



[図2]

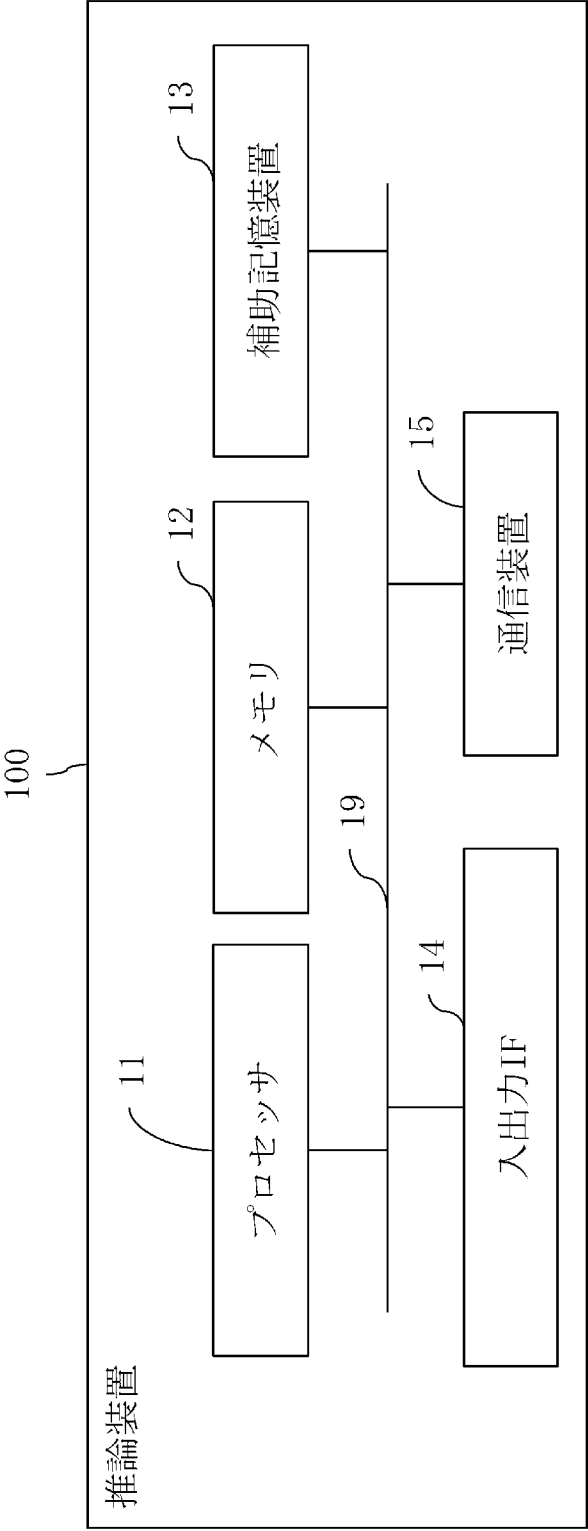
図2

優先度	プロセス名
最高	(他プロセス)
	量子化推論プロセス
	(他プロセス)
	非量子化推論プロセス
最低	(他プロセス)



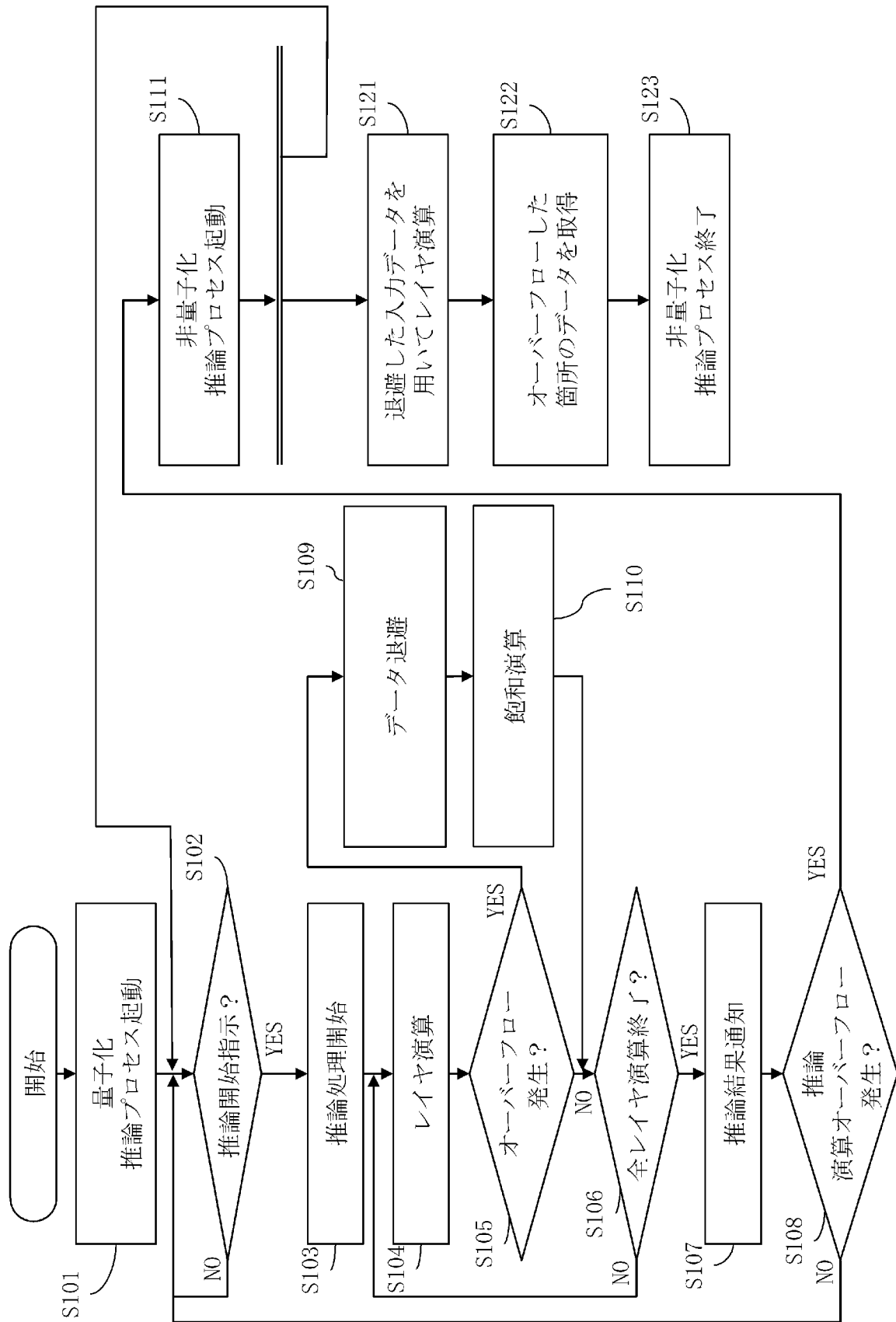
[図3]

図3



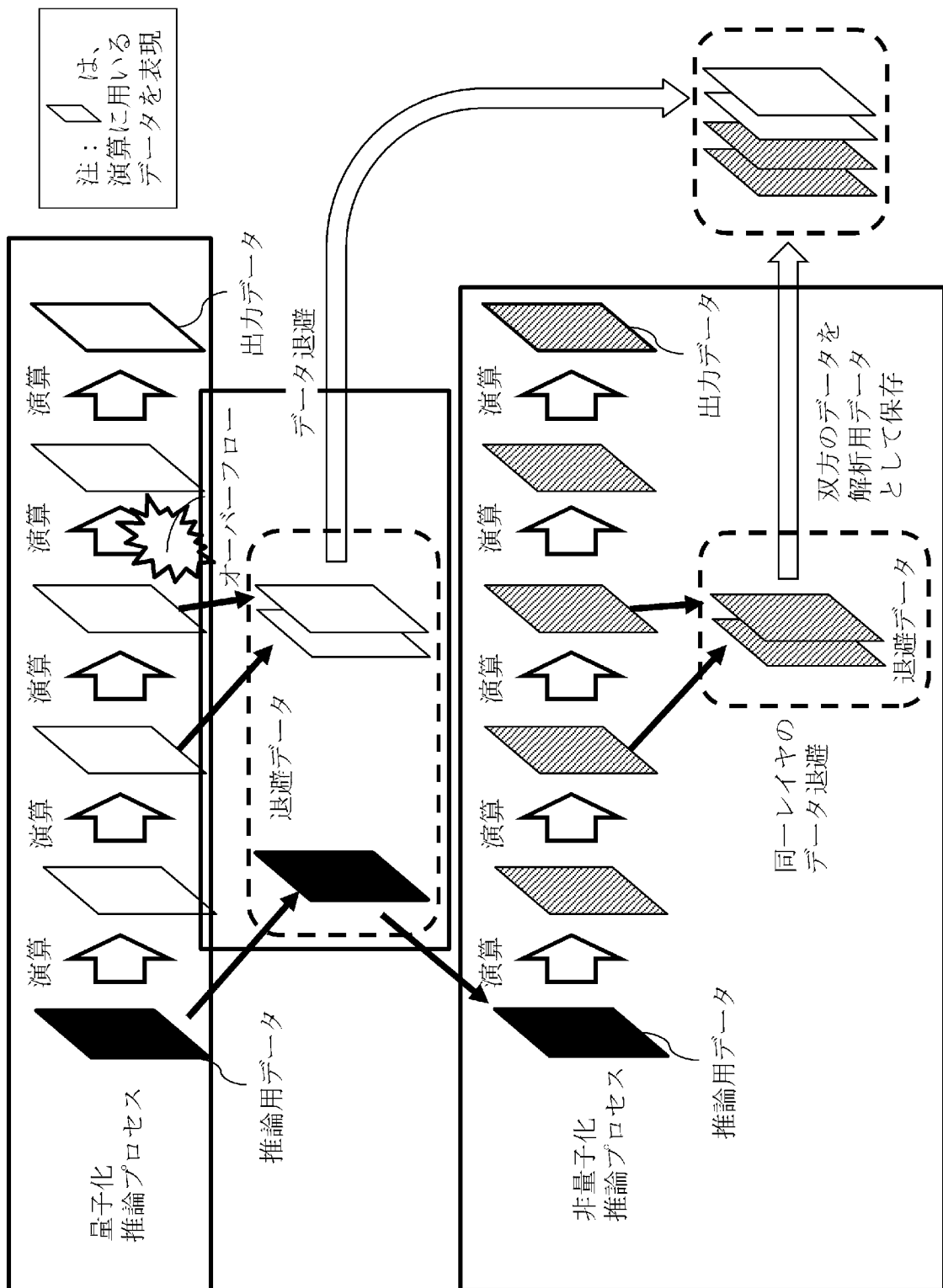
[図4]

図4



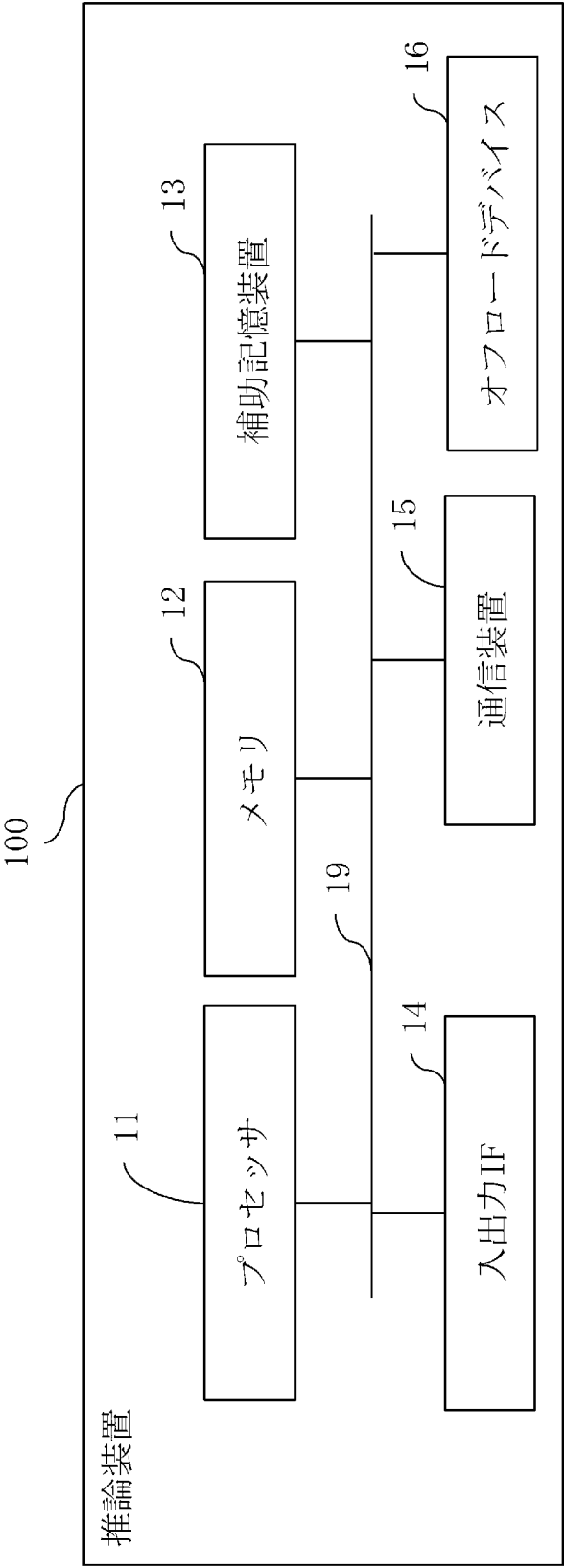
[図5]

図5



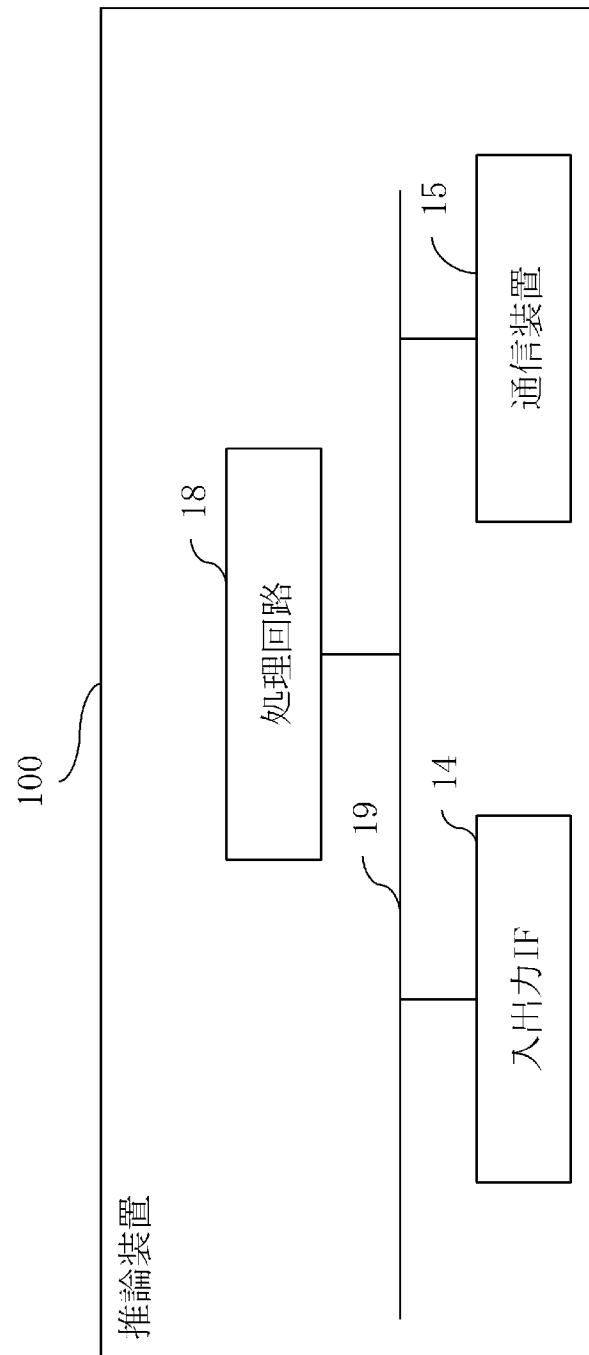
[図6]

図6



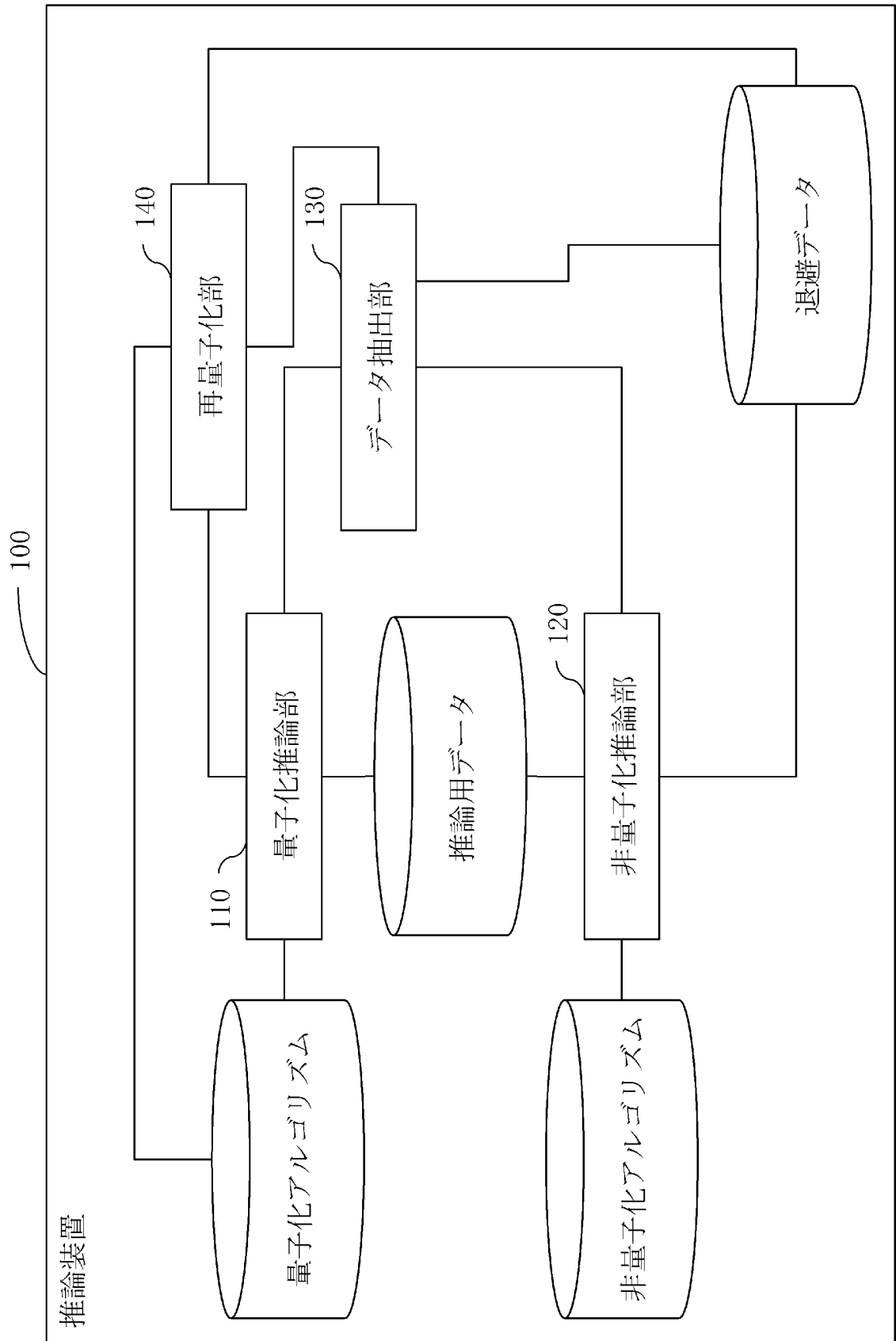
[図7]

図7



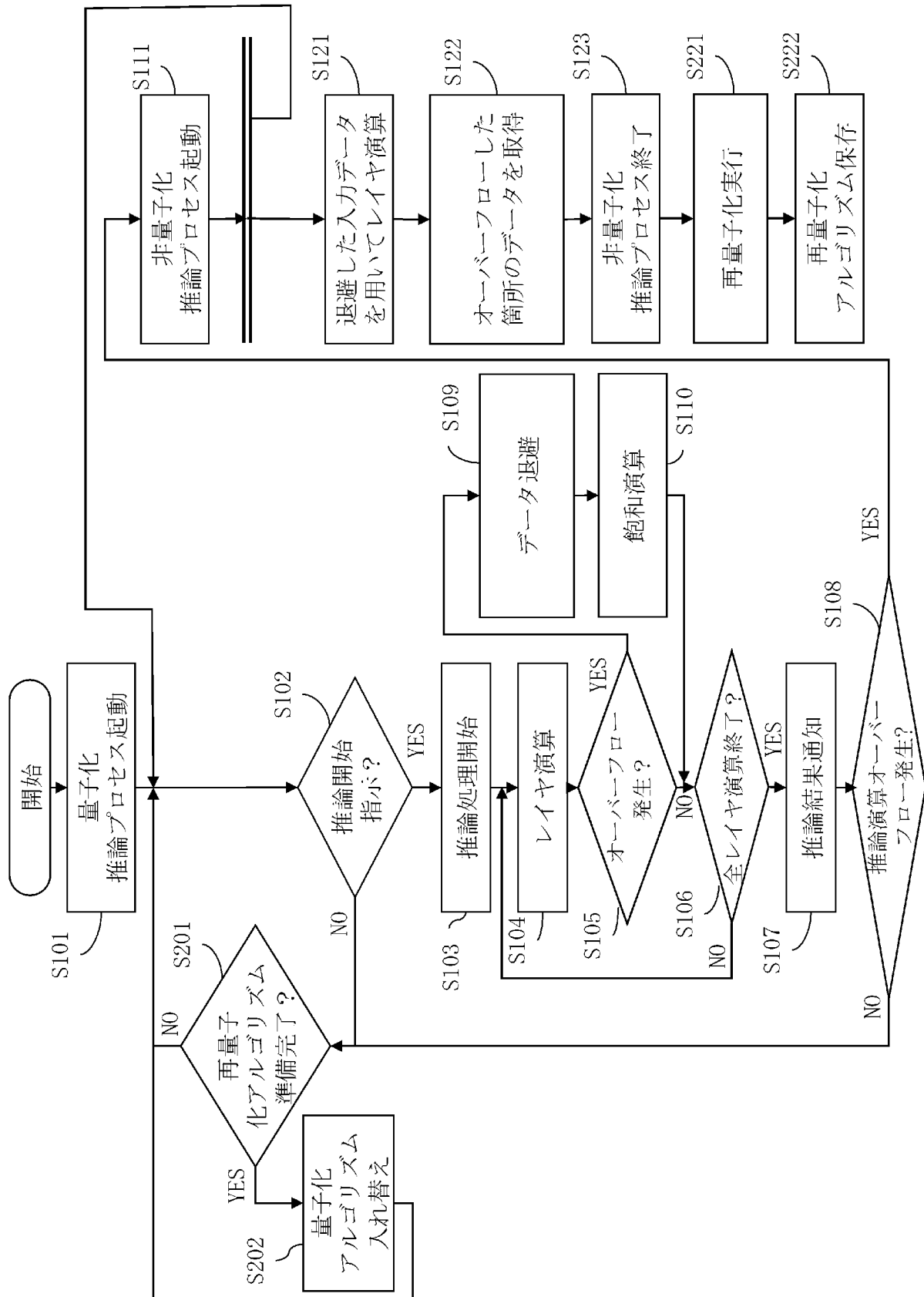
[図8]

図8



[図9]

図9



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2021/012193

A. CLASSIFICATION OF SUBJECT MATTER

G06N 20/00(2019.01)i

FI: G06N20/00

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)

G06N20/00

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 1971-2021

Registered utility model specifications of Japan 1996-2021

Published registered utility model applications of Japan 1994-2021

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	JP 2020-64635 A (SAMSUNG ELECTRONICS CO LTD) 23 April 2020 (2020-04-23)	1-6
A	US 2017/0316307 A1 (INTEL CORPORATION) 02 November 2017 (2017-11-02)	1-6
A	JP 2020-009048 A (HITACHI LTD) 16 January 2020 (2020-01-16)	1-6

☐ 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
17 June 2021 (17.06.2021)

Date of mailing of the international search report
29 June 2021 (29.06.2021)

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/JP2021/012193

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
JP 2020-64635 A	23 Apr. 2020	US 2020/0125947 A1 EP 3640858 A1 KR 10-2020-0043169 A CN 111062475 A (Family: none)	
US 2017/0316307 A1	02 Nov. 2017		
JP 2020-009048 A	16 Jan. 2020	US 2020/0012926 A1	

A. 発明の属する分野の分類（国際特許分類（IPC）） G06N 20/00(2019.01)i FI: G06N20/00														
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）） G06N20/00														
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 - 2021年</td> </tr> <tr> <td>日本国実用新案登録公報</td> <td>1996 - 2021年</td> </tr> <tr> <td>日本国登録実用新案公報</td> <td>1994 - 2021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 - 2021年	日本国実用新案登録公報	1996 - 2021年	日本国登録実用新案公報	1994 - 2021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 - 2021年													
日本国実用新案登録公報	1996 - 2021年													
日本国登録実用新案公報	1994 - 2021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	JP 2020-64635 A（三星電子株式会社）23.04.2020（2020 - 04 - 23）	1-6												
A	US 2017/0316307 A1（INTEL CORPORATION）02.11.2017（2017 - 11 - 02）	1-6												
A	JP 2020-009048 A（株式会社日立製作所）16.01.2020（2020 - 01 - 16）	1-6												
☐ 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 17.06.2021	国際調査報告の発送日 29.06.2021													
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） 渡辺 順哉 5B 1157 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/JP2021/012193

引用文献			公表日	パテントファミリー文献			公表日
JP	2020-64635	A	23.04.2020	US	2020/0125947	A1	
				EP	3640858	A1	
				KR	10-2020-0043169	A	
				CN	111062475	A	
US	2017/0316307	A1	02.11.2017	(ファミリーなし)			
JP	2020-009048	A	16.01.2020	US	2020/0012926	A1	