

HN, HR, HU, ID, IL, IN, IR, IS, JO, 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:

- 一 国際調査報告 (条約第21条(3))

(57) 要約: データ転送のボトルネックを排除して推論演算の処理時間を削減することができる推論処理技術を提供することを目的とする。推論処理装置1は、入力データXを記憶する入力データ格納部120と、ニューラルネットワークの重みデータWを記憶する学習済みNN格納部121と、入力データXに関する情報に基づいて、バッチサイズを設定するバッチ処理制御部10と、設定されたバッチサイズに応じた入力データXを入力データ格納部120から読み出すメモリ制御部11と、バッチサイズに応じた入力データXおよび重みデータWを入力として、ニューラルネットワークの演算をバッチ処理して、入力データXの特徴を推論する推論演算部13とを備える。

明 細 書

発明の名称：推論処理装置および推論処理方法

技術分野

[0001] 本発明は、推論処理装置および推論処理方法に関し、特に、ニューラルネットワークを用いて推論を行う技術に関する。

背景技術

[0002] 近年、モバイル端末やInternet of Things (IoT) デバイス等のエッジデバイスの増加に伴い、生成されるデータが爆発的に増加している。この膨大なデータから有意義な情報を抽出するには、深層ニューラルネットワーク (Deep Neural Networks: DNN) とよばれる最先端の機械学習技術が優位である。近年のDNNに関する研究の進歩により、データの解析精度は大幅に改善されており、DNNを利用した技術のさらなる発展が期待されている。

[0003] DNNの処理は学習および推論の2つのフェーズがある。一般に、学習では、大量のデータを必要とするため、クラウドで処理される場合がある。一方、推論では学習済みのDNNモデルを使用し、未知の入力データに対して出力を推定する。

[0004] より詳細に説明すると、DNNにおける推論処理では、学習済みのニューラルネットワークモデルに時系列データまたは画像データなどの入力データを与えて、入力データの特徴を推論する。例えば、非特許文献1に開示されている具体的な例によると、加速度センサとジャイロセンサとを搭載したセンサ端末を用いて、ゴミ収集車の回転や停止といったイベントを検出することで、ゴミの量を推定している。このように、未知の時系列データを入力として、各時刻におけるイベントを推定するには、予め各時刻におけるイベントが既知である時系列データを用いて学習したニューラルネットワークモデルを用いる。

[0005] 非特許文献1では、センサ端末から取得される時系列データを入力データ

として用いており、リアルタイムにイベントを抽出する必要がある。そのため、推論処理をより高速化することが必要となる。そこで、従来から、推論処理を実現するFPGAをセンサ端末に搭載し、そのようなFPGAで推論演算を行い、処理の高速化を図っている（非特許文献2参照）。

先行技術文献

非特許文献

[0006] 非特許文献1: Kishino, et. al, “Detecting garbage collection duration using motion sensors mounted on a garbage truck toward smart waste management”, SPWID17

非特許文献2: Kishino, et. al, “Datafying city: detecting and accumulating spatio-temporal events by vehicle-mounted sensors”, BIGDATA 2017

発明の概要

発明が解決しようとする課題

[0007] しかし、従来の技術では、推論処理を行う際に、メモリから入力データおよびニューラルネットワークモデルの重みを推論処理の対象となるデータセットごとに読み出し、推論演算を行う回路に転送する必要があった。そのため、扱うデータ量が多くなるとデータ転送がボトルネックになり、推論演算の処理時間を削減することが困難であった。

[0008] 本発明は、上述した課題を解決するためになされたものであり、データ転送のボトルネックを排除して推論演算の処理時間を削減することができる推論処理技術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

[0009] 上述した課題を解決するために、本発明に係る推論処理装置は、入力デー

タを記憶する第1記憶部と、ニューラルネットワークの重みを記憶する第2記憶部と、前記入力データに関する情報に基づいて、バッチサイズを設定するバッチ処理制御部と、設定された前記バッチサイズに応じた前記入力データを前記第1記憶部から読み出すメモリ制御部と、前記バッチサイズに応じた前記入力データおよび前記重みを入力として、前記ニューラルネットワークの演算をバッチ処理して、前記入力データの特徴を推論する推論演算部とを備える。

[0010] また、本発明に係る推論処理装置において、前記バッチ処理制御部は、推論演算に用いられるハードウェアリソースに関する情報に基づいて、前記バッチサイズを設定してもよい。

[0011] また、本発明に係る推論処理装置において、前記推論演算部は、前記入力データと前記重みとの行列演算を行う行列演算部と、前記行列演算部による行列演算結果に対して活性化関数を適用する活性化関数演算部とを備え、前記行列演算部は、前記入力データと前記重みとを乗算する乗算器と、前記乗算器による乗算結果を加算する加算器とを有していてもよい。

[0012] また、本発明に係る推論処理装置において、前記行列演算部は、複数具備され、行列演算を並列に行ってもよい。

[0013] また、本発明に係る推論処理装置において、前記行列演算部は、前記乗算器および前記加算器をそれぞれ複数具備し、乗算および加算を並列に行ってもよい。

[0014] また、本発明に係る推論処理装置において、前記推論演算部に入力される前記入力データおよび前記重みのデータタイプを変換するデータ変換部をさらに備えていてもよい。

[0015] また、本発明に係る推論処理装置において、前記推論演算部は、複数具備され、推論演算を並列に行ってもよい。

[0016] 上述した課題を解決するために、本発明に係る推論処理方法は、第1記憶部に記憶された入力データに関する情報に基づいて、バッチサイズを設定する第1ステップと、設定された前記バッチサイズに応じた前記入力データを

前記第 1 記憶部から読み出す第 2 ステップと、前記バッチサイズに応じた前記入力データおよび第 2 記憶部に記憶されたニューラルネットワークの重みを入力として、前記ニューラルネットワークの演算をバッチ処理して、前記入力データの特徴を推論する第 3 ステップとを備える。

発明の効果

[0017] 本発明によれば、入力データに関する情報に基づいて設定されたバッチサイズに応じた入力データ、および重みを入力として、学習済みニューラルネットワークの演算をバッチ処理するので、扱うデータ量が多くなっても、データ転送のボトルネックを排除して、推論演算の処理時間を削減することができる。

図面の簡単な説明

[0018] [図1]図 1 は、本発明の第 1 の実施の形態に係る推論処理装置の構成を示すブロック図である。

[図2]図 2 は、第 1 の実施の形態に係る記憶部の構成を示すブロック図である。

[図3]図 3 は、第 1 の実施の形態に係る推論演算部の構成を示すブロック図である。

[図4]図 4 は、第 1 の実施の形態に係る行列演算部の構成を示すブロック図である。

[図5]図 5 は、第 1 の実施の形態に係る推論処理装置のハードウェア構成を示すブロック図である。

[図6]図 6 は、第 1 の実施の形態に係る推論処理プログラムのサンプルコードの一例を説明する図である。

[図7A]図 7 A は、第 1 の実施の形態に係るニューラルネットワークを用いた推論処理を説明するための図である。

[図7B]図 7 B は、第 1 の実施の形態に係るニューラルネットワークを用いた推論処理を説明するための図である。

[図8]図 8 は、第 1 の実施の形態に係る推論処理装置の動作を説明するための

フローチャートである。

[図9]図 9 は、第 1 の実施の形態に係るバッチサイズの設定処理を説明するためのフローチャートである。

[図10]図 10 は、従来例の推論処理装置におけるデータの転送を説明するための図である。

[図11]図 11 は、第 1 の実施の形態に係る推論処理装置におけるデータの転送を説明するための図である。

[図12]図 12 は、第 1 の実施の形態の効果を説明するための図である

[図13]図 13 は、第 2 の実施の形態に係る推論処理装置の構成を示すブロック図である。

[図14]図 14 は、第 2 の実施の形態に係る推論処理装置の動作を説明するためのフローチャートである。

[図15]図 15 は、第 2 の実施の形態の効果を説明するための図である。

[図16]図 16 は、第 3 の実施の形態に係る推論処理装置の構成を示すブロック図である。

[図17]図 17 は、第 4 の実施の形態に係る推論演算部の構成を示すブロック図である。

[図18]図 18 は、第 5 の実施の形態に係る行列演算部の構成を示すブロック図である。

[図19]図 19 は、第 6 の実施の形態に係る推論処理装置の構成を示すブロック図である。

[図20]図 20 は、従来例に係る推論処理装置の構成を示すブロック図である。

発明を実施するための形態

[0019] 以下、本発明の好適な実施の形態について、図 1 から図 20 を参照して詳細に説明する。

[0020] [第 1 の実施の形態]

図 1 は、本発明の第 1 の実施の形態に係る推論処理装置 1 の構成を示すブ

ロック図である。本実施の形態に係る推論処理装置 1 は、図 1 に示すように、外部のセンサ 2 などから取得された音声データや言語データなどの時系列データ、または画像データを推論対象の入力データ X として用いる。推論処理装置 1 は、学習済みのニューラルネットワークモデルを用いてニューラルネットワークの演算をバッチ処理し、入力データ X の特徴を推論する。

[0021] より詳細には、推論処理装置 1 は、各時刻におけるイベントが既知である時系列データなどの入力データ X を用いて予め学習したニューラルネットワークモデルを用いる。推論処理装置 1 は、設定されたバッチサイズに応じた未知の時系列データなどの入力データ X および学習済みのニューラルネットワークの重みデータ W を入力として、各時刻におけるイベントの推定をバッチ処理により行う。なお、入力データ X および重みデータ W は、行列データである。

[0022] 例えば、推論処理装置 1 は、加速度センサとジャイロセンサとを搭載したセンサ 2 から取得された入力データ X をバッチ処理して、ごみ収集車の回転や停止といったイベントを検出することで、ゴミの量を推定することができる（非特許文献 1 参照）。

[0023] [推論処理装置の構成]

推論処理装置 1 は、図 1 に示すように、バッチ処理制御部 10、メモリ制御部 11、記憶部 12、および推論演算部 13 を備える。

[0024] バッチ処理制御部 10 は、入力データ X に関する情報に基づいて、推論演算部 13 で入力データ X をバッチ処理するためのバッチサイズを設定する。バッチ処理制御部 10 は、メモリ制御部 11 に対して、設定されたバッチサイズに応じた入力データ X を記憶部 12 から読み出す指示を送る。

[0025] 例えば、バッチ処理制御部 10 は、後述する推論演算に用いられるハードウェアリソースに関する情報に基づいて、1 回のバッチ処理で扱う入力データ X の数、すなわちバッチサイズを設定することができる。

[0026] あるいは、バッチ処理制御部 10 は、記憶部 12 に記憶されているニューラルネットワークモデルの重みデータ W の行列サイズまたは入力データ X の

行列サイズに基づいて、バッチサイズを設定することができる。

[0027] 上記例の他にも、バッチ処理制御部 10 は、例えば、データの送受信時間およびデータの演算時間の最適化を行い、送受信時間と演算時間とのバランスにより最適なバッチサイズを設定することもできる。また、バッチ処理制御部 10 は、推論処理装置 1 全体の処理時間および推論精度に基づいて、バッチサイズを設定してもよい。

[0028] メモリ制御部 11 は、バッチ処理制御部 10 で設定されたバッチサイズに応じた入力データ X を記憶部 12 から読み出す。また、メモリ制御部 11 は、記憶部 12 からニューラルネットワークの重みデータ W を読み出す。メモリ制御部 11 は、読み出した入力データ X および重みデータ W を推論演算部 13 に転送する。

[0029] 記憶部 12 は、図 2 に示すように、入力データ格納部（第 1 記憶部）120 および学習済みニューラルネットワーク（NN）格納部（第 2 記憶部）121 を備える。

入力データ格納部 120 には、外部のセンサ 2 から取得された時系列データなどの入力データ X が格納される。

[0030] 学習済み NN 格納部 121 には、予め学習され構築された学習済みニューラルネットワーク、すなわち、ニューラルネットワークの学習済みパラメータである重みデータ W が格納される。例えば、予め外部のサーバなどで学習が行われて決定された重みデータ W がロードされ、学習済み NN 格納部 121 に格納されている。

[0031] なお、推論処理装置 1 において採用されるニューラルネットワークモデルとして、例えば、畳み込みニューラルネットワーク（CNN）、長期短期記憶（LSTM）、ゲート付き再帰型ユニット（GRU）、Residual Network（ResNet）CNN、その他の公知の少なくとも 1 層の中間層を有するニューラルネットモデル、あるいはこれらを組み合わせたニューラルネットワークを用いることができる。

[0032] また、行列である入力データ X および重みデータ W のサイズは推論処理装

置 1 で用いられるニューラルネットワークモデルにより定まる。入力データ X および重みデータ W は、例えば、32 ビットの浮動小数点型で表される。

[0033] 推論演算部 13 は、設定されたバッチサイズに応じた入力データ X および重みデータ W を入力としてニューラルネットワークの演算をバッチ処理し、入力データ X の特徴を推論する。より詳細には、メモリ制御部 11 によって読み出され、転送された入力データ X および重みデータ W が推論演算部 13 に入力されて、推論演算が行われる。

[0034] 推論演算部 13 は、図 3 に示すように、行列演算部 130、および活性化関数演算部 131 を備える。行列演算部 130 は、図 4 に示すように、乗算器 132 と加算器 133 とを有する。

[0035] 行列演算部 130 は、入力データ X と重みデータ W との行列演算を行う。より詳細には、図 4 に示すように、乗算器 132 は、入力データ X と重みデータ W との乗算を行う。乗算結果は、加算器 133 によって加算され、加算結果が出力される。この加算結果は、行列演算部 130 による行列演算結果 A として出力される。

[0036] 行列演算結果 A は、活性化関数演算部 131 に入力され、予め設定された活性化関数が適用されて、推論演算の結果である推論結果 Y が決定される。より具体的には、活性化関数演算部 131 は、活性化関数の適用により、行列演算結果 A がどのように活性化するかを決定し、行列演算結果 A を変換して推論結果 Y を出力する。活性化関数は、例えば、ステップ関数、シグモイド関数、 $\tanh$ 関数、ReLU 関数、softmax 関数などから選択することができる。

[0037] [推論処理装置のハードウェア構成]

次に、上述した構成を有する推論処理装置 1 のハードウェア構成の一例について図 5 を参照して説明する。

[0038] 図 5 に示すように、推論処理装置 1 は、例えば、バス 101 を介して接続されるプロセッサ 102、主記憶装置 103、通信インターフェース 104、補助記憶装置 105、入出力装置 106 を備えるコンピュータと、これら

のハードウェア資源を制御するプログラムによって実現することができる。
推論処理装置 1 は、例えば、表示装置 107 がバス 101 を介して接続され、表示画面に推論結果などを表示してもよい。また、センサ 2 がバス 101 を介して接続され、推論処理装置 1 において推論の対象となる音声データなどの時系列データからなる入力データ X を測定してもよい。

[0039] 主記憶装置 103 は、例えば、SRAM、DRAM、およびROMなどの半導体メモリによって実現される。主記憶装置 103 は、図 1 で説明した記憶部 12 を実現する。

[0040] 主記憶装置 103 には、プロセッサ 102 が各種制御や演算を行うためのプログラムが予め格納されている。プロセッサ 102 と主記憶装置 103 とによって、図 1 から図 4 に示したバッチ処理制御部 10、メモリ制御部 11、推論演算部 13 を含む推論処理装置 1 の各機能が実現される。

[0041] 通信インターフェース 104 は、通信ネットワーク NW を介して各種外部電子機器との通信を行うためのインターフェース回路である。推論処理装置 1 は、通信インターフェース 104 を介して外部から学習済みニューラルネットワークの重みデータ W を受信したり、外部に推論結果 Y を送出してもよい。

[0042] 通信インターフェース 104 としては、例えば、LTE、3G、無線 LAN、Bluetooth（登録商標）などの無線データ通信規格に対応したインターフェースおよびアンテナが用いられる。通信ネットワーク NW は、例えば、WAN（Wide Area Network）や LAN（Local Area Network）、インターネット、専用回線、無線基地局、プロバイダなどを含む。

[0043] 補助記憶装置 105 は、読み書き可能な記憶媒体と、その記憶媒体に対してプログラムやデータなどの各種情報を読み書きするための駆動装置とで構成されている。補助記憶装置 105 には、記憶媒体としてハードディスクやフラッシュメモリなどの半導体メモリを使用することができる。

[0044] 補助記憶装置 105 は、推論処理装置 1 がバッチ処理により推論を行うた

めのプログラムを格納するプログラム格納領域を有する。さらには、補助記憶装置105は、例えば、上述したデータやプログラムやなどをバックアップするためのバックアップ領域などを有していてもよい。補助記憶装置105は、例えば、図6に示す推論処理プログラムを記憶することができる。

[0045] 入出力装置106は、表示装置107など外部機器からの信号を入力したり、外部機器へ信号を出力したりするI/O端子により構成される。

[0046] なお、推論処理装置1は、1つのコンピュータによって実現される場合だけでなく、互いに通信ネットワークNWで接続された複数のコンピュータによって分散されていてもよい。また、プロセッサ102は、FPGA (Field-Programmable Gate Array)、LSI (Large Scale Integration)、ASIC (Application Specific Integrated Circuit) 等のハードウェアによって実現されていてもよい。

[0047] 特に、推論演算部13をFPGAなどの書き換え可能なゲートアレイを用いて構成することで、入力データXの構成や使用されるニューラルネットワークモデルに応じて柔軟に回路構成を書き換えることができる。この場合、様々なアプリケーションに対応することが可能な推論処理装置1を実現できる。

[0048] [推論処理方法の概要]

次に、本実施の形態に係る推論処理装置1による入力データXの推論処理の概要について、図7Aおよび図7Bに示す具体例を用いて説明する。

[0049] 図7Aに示すように、入力層、中間層、および出力層の3層からなるニューラルネットワークを例に挙げて説明する。活性化関数としては、図7Bに示すsoftmax関数を用いる。推論の対象となる入力データXの特徴は、M個 (Mは正の整数) の成分、推論結果Yの特徴は、N個 (Nは正の整数) の成分で表される。また、ニューラルネットワークの重みデータWのデータサイズはM×Nで表される。

[0050] 図7Aおよび図7Bの具体例に示すように、M=N=2とする。また、説

明の簡単のため、バッチ処理制御部 10 によって設定された 1 回のバッチ処理で扱われるバッチサイズは、 $Batch = 1$ とする。このとき、バッチサイズ $Batch = 1$ に対応する入力データ X は、 $X[x_1, x_2]$ である。また、重みデータ W は、4 つの成分を有する 2 行 2 列の行列で表される。

[0051] 図 7 B に示すように、まず、バッチサイズ $Batch = 1$ に応じた入力データ X と、重みデータ W との行列の積和演算が行われ、行列演算結果 A が得られる。行列演算結果 A のデータサイズは、 $Batch \times N$ 、すなわち 1×2 である。その後、行列演算結果 A には、活性化関数として $softmax$ 関数が適用されて、推論結果 Y が求められる。

[0052] 設定されたバッチサイズ $Batch$ に応じたデータ数の入力データ X に対しては、そのバッチサイズ $Batch$ に応じたデータ数の推論結果 Y が出力される。したがって、図 7 A および図 7 B の例では、 $Batch = 1$ に応じた 1 セットの入力データ $X[x_1, x_2]$ に対して、1 セットの推論結果 $Y[y_1, y_2]$ が出力されている。なお、バッチサイズ $Batch$ は、1 以上かつ入力データ X のデータ数以下の範囲の値である。

[0053] 活性化関数の演算処理においては、行列演算結果 A の各成分の値 a_k ($k = 1, \dots, n$) に対して $softmax$ 関数が適用されて、推論結果 Y の各成分 y_k ($k = 1, \dots, n$) の値が求められる。図 7 A および図 7 B に示す具体例では、行列演算結果 $A[a_1, a_2]$ の各成分に $softmax$ 関数が適用されて ($softmax(A[a_1, a_2])$)、推論結果 $Y[y_1, y_2]$ が出力される。

[0054] なお、設定されたバッチサイズに応じて、入力データ X のバッチ処理により推論演算を繰り返し行って推論結果 Y を出力する処理は、図 6 のサンプルコードの破線の枠 60 に示されている。この推論演算において、入力データ X の行の成分と重みデータ W の列の成分との積およびその和が計算される（図 6 のサンプルコードにおける破線の枠 61、62 を参照）。

[0055] [推論処理装置の動作]

次に、本実施の形態に係る推論処理装置 1 の動作について、図 8 および図

9のフローチャートを参照してより詳細に説明する。以下の説明では、記憶部12には、予め学習が行われて構築されたニューラルネットワークの重みデータWが格納されているものとする。また、外部のセンサ2によって測定された時系列データや画像データなどの入力データXは、記憶部12に保持されるものとする。

[0056] まず、図8に示すように、バッチ処理制御部10は、1回のバッチ処理で扱う入力データXのバッチサイズを設定する（ステップS1）。

[0057] より詳細には、バッチ処理制御部10は、図9に示すように、記憶部12に記憶されている重みデータWのデータサイズ、および入力データXのデータ数の情報を取得する（ステップS100）。次に、バッチ処理制御部10は、推論処理装置1全体におけるハードウェアリソースの情報を記憶部12から取得する（ステップS101）。なお、推論処理装置1全体のハードウェアリソースに関する情報は、予め記憶部12に記憶されている。

[0058] ここで、ハードウェアリソースとは、入力データXや重みデータWを記憶するために必要なメモリ容量や、加算、乗算などの演算処理を行うための回路を構成するのに必要なスタンダードセルの組合せ回路などを意味する。例えば、FPGAの場合であれば、フリップフロップ（FF）、ルックアップテーブル（LUT）、デジタルシグナルプロセッサ（DSP）などの組合せ回路がハードウェアリソースの例として挙げられる。

[0059] ステップS101では、推論処理装置1全体におけるメモリ容量、および推論処理装置1全体のデバイスとしての規模、すなわち推論処理装置1全体が演算回路として備えるハードウェアリソース、例えば、FPGAであればFF、LUT、DSPなどの数が記憶部12から取得される。

[0060] 次に、バッチ処理制御部10は、入力データXの全データ数を1回のバッチ処理で扱うバッチサイズの初期値として設定する（ステップS102）。すなわち、ステップS102では、バッチサイズの最大値である、入力データXの全データ数がバッチサイズの初期値として設定される。

[0061] その後、ステップS100で取得した重みデータWのデータサイズおよび

入力データXのデータ数、ステップS101で取得した推論処理装置1全体のハードウェアリソースの情報、およびステップS102で設定されたバッチサイズに基づいて、推論演算部13を実現する回路構成に必要なハードウェアリソースを求める（ステップS103）。例えば、バッチ処理制御部10は、推論演算部13のロジック回路を構築して、使用されるハードウェアリソースを取得することができる。

[0062] 次に、推論演算部13が推論演算を行う際に使用するハードウェアリソースの数が、推論処理装置1全体が備えるハードウェアリソースの数を超える場合には（ステップS104：YES）、バッチ処理制御部10は、ステップS102で初期設定されたバッチサイズを小さくする（ステップS105）。例えば、バッチ処理制御部10は、初期設定されたバッチサイズから1を減算する。

[0063] その後、より小さいバッチサイズに基づいて求められる推論演算部13のハードウェアリソースの数が、推論処理装置1全体のハードウェアリソースの数以下となる場合には（ステップS106：NO）、そのバッチサイズが設定値として用いられ、処理は図8に戻る。より詳細には、バッチ処理制御部10は、メモリ制御部11に対して、設定されたバッチサイズに応じた入力データXを読み出すことを指示する。

[0064] なお、ステップS106で推論演算部13が推論演算に使用するハードウェアリソースの数が、推論処理装置1全体の備えるハードウェアリソースの数を超える場合には（ステップS106：YES）、バッチ処理制御部10は、再びバッチサイズを減らす処理を行う（ステップS105）。

[0065] その後、メモリ制御部11は、設定されたバッチサイズに応じた入力データX、および重みデータWを記憶部12から読み出す（ステップS2）。より詳細には、メモリ制御部11は、入力データXおよび重みデータWを記憶部12から読み出して、推論演算部13へ転送する。

[0066] 次に、推論演算部13は、入力データXおよび重みデータWに基づいてニューラルネットワークの演算をバッチ処理し、推論結果Yを求める（ステッ

プS 3)。より詳細には、行列演算部 1 3 0 において、入力データ X と重みデータ W との積和演算が行われる。具体的には、乗算器 1 3 2 が入力データ X と重みデータ W との乗算を行う。乗算結果は加算器 1 3 3 によって加算されて行列演算結果 A が求められる。行列演算結果 A に対しては、活性化関数演算部 1 3 1 によって活性化関数が適用され、推論結果 Y が出力される（ステップ S 4）。

[0067] 上記処理によって推論処理装置 1 は、画像データや音声などの時系列データを入力データ X として、学習済みのニューラルネットワークを用いた入力データ X の特徴の推論を行うことができる。

[0068] ここで、本実施の形態に係るバッチ処理制御部 1 0 の効果について、図 1 0 および図 1 1、ならびに図 2 0 を用いて説明する。まず、比較のため、従来例の推論処理装置（図 2 0）として、本実施の形態に係るバッチ処理制御部 1 0 を備えていない推論処理装置について説明する。図 1 0 に示すように、従来例に係る推論処理装置では、入力データ X を n 個（ n は正の整数）処理すると、重みデータ W を n 回にわたって推論演算部に転送する必要がある。

[0069] これに対して、本実施の形態に係るバッチ処理制御部 1 0 を備える推論処理装置 1 では、図 1 1 に示すように、バッチ処理制御部 1 0 が、1 回の推論演算で処理されるバッチサイズ $B a t c h$ を設定し、設定されたバッチサイズに応じた入力データ X をまとめて処理する。そのため、入力データ X が、例えば、 n 個ある場合であっても、重みデータ W を $n / B a t c h$ 回だけ推論演算部 1 3 に転送すればよい。 $B a t c h = n$ の場合は、重みデータ W の推論演算部 1 3 への転送は 1 回だけで済む。したがって、推論処理装置 1 におけるバス帯域の負荷を低減することができる。

[0070] また、本実施の形態に係る推論処理装置 1 では、バッチ処理を行うことで比較的大きな行列計算を行うことができ、分割された、より小さい行列計算を実行するよりも計算速度は速く、推論演算をより高速化することができる。

[0071] 図12は、重みデータWのデータサイズが 30×30 の場合のバッチ処理による本実施の形態の効果を示している。図12において、破線はバッチ処理を行わない場合、実線は本実施の形態に係るバッチ処理を行った場合それぞれにおけるバッチサイズと、推論演算の正規化された処理時間との関係を示している。図12からわかるように、本実施の形態に係るバッチ処理を行った場合には、バッチ処理を行わない場合と比較して処理時間が短縮されている。

[0072] 以上説明したように、第1の実施の形態に係る推論処理装置1によれば、推論処理装置1全体のハードウェアリソースに対する推論演算部13が使用するハードウェアリソースに基づいて、1回のバッチ処理で扱う入力データXのバッチサイズを設定する。そのため、扱われるデータ量が多くなってもデータ転送におけるボトルネックを排除して、推論演算に要する処理時間を削減することができる。

[0073] [第2の実施の形態]

次に、本発明の第2の実施の形態について説明する。なお、以下の説明では、上述した第1の実施の形態と同じ構成については同一の符号を付し、その説明を省略する。

[0074] 第1の実施の形態では、推論演算部13は、例えば、32ビットの浮動小数点型の入力データXおよび重みデータWの推論演算を実行する場合について説明した。これに対して第2の実施の形態では、推論演算部13に入力されるデータのビット表現をより低ビット精度のデータに変換してから推論演算を実行する。以下、第1の実施の形態と異なる構成を中心に説明する。

[0075] [推論処理装置の構成]

図13は、本実施の形態に係る推論処理装置1Aの構成を示すブロック図である。

推論処理装置1Aは、バッチ処理制御部10、メモリ制御部11、記憶部12、推論演算部13、およびデータタイプ変換部（データ変換部）14を備える。

[0076] データタイプ変換部 14 は、推論演算部 13 に入力される入力データ X および重みデータ W のデータタイプを変換する。より詳細には、データタイプ変換部 14 は、メモリ制御部 11 によって記憶部 12 から読み出されて推論演算部 13 に転送される入力データ X および重みデータ W のデータタイプを、32 ビットの浮動小数点型から、予め設定されたデータタイプ、例えば、8 ビットや 16 ビットなど、より桁数を減らして精度を下げたデータ表現に変換する。データタイプ変換部 14 は、例えば、切り上げ、切り捨て、四捨五入などの丸め処理を行って小数点を含む入力データ X および重みデータ W を整数型へ変換することができる。

[0077] なお、データタイプ変換部 14 は、メモリ制御部 11 が記憶部 12 にアクセスして読み出した入力データ X および重みデータ W について、転送される前にデータタイプを変換することができる。また、データタイプ変換部 14 は、元のデータタイプよりも桁数の低い低ビット精度とすることができれば、入力データ X と重みデータ W とをそれぞれ異なるビット表現のデータタイプに変換してもよい。

[0078] メモリ制御部 11 は、データタイプ変換部 14 によってデータタイプが変換され、よりビット精度が低くなった入力データ X' および重みデータ W' を推論演算部 13 に転送する。より詳細には、メモリ制御部 11 は、バッチ処理制御部 10 によって設定されたバッチサイズに応じた入力データ X、および予め記憶部 12 に記憶されている重みデータ W を記憶部 12 から読み出す。その後、読み出された入力データ X および重みデータ W は、データタイプ変換部 14 によってデータタイプが変換され、変換された入力データ X' および重みデータ W' が推論演算部 13 に転送される。

[0079] [推論処理装置の動作]

次に、上述した構成を有する推論処理装置 1A の動作について、図 14 のフローチャートを参照して説明する。以下の説明では、記憶部 12 には、予め学習が行われて構築されたニューラルネットワークの重みデータ W が格納されているものとする。また、重みデータ W と、センサ 2 から取得されて記

憶部 12 に記憶されている入力データ X とは、ともに 32 ビットの浮動小数点型のデータであるものとする。

[0080] まず、図 14 に示すように、バッチ処理制御部 10 は、1 回のバッチ処理で扱う入力データ X のバッチサイズを設定する（ステップ S10）。なお、バッチサイズの設定処理は、第 1 の実施の形態と同様である（図 9）。

[0081] その後、メモリ制御部 11 は、バッチ処理制御部 10 によって設定されたバッチサイズに応じた入力データ X、および重みデータ W を記憶部 12 から読み出す（ステップ S11）。次に、データタイプ変換部 14 は、メモリ制御部 11 が読み出した入力データ X および重みデータ W のデータタイプを変換する（ステップ S12）。

[0082] より具体的には、データタイプ変換部 14 は、32 ビットの浮動小数点型の入力データ X および重みデータ W を、より低ビット精度のデータ、例えば、8 ビットの入力データ X' および重みデータ W' に変換する。データタイプが変換された入力データ X' および重みデータ W' は、メモリ制御部 11 によって推論演算部 13 に転送される。

[0083] その後、推論演算部 13 は、低ビット精度のデータに変換された入力データ X' および重みデータ W' に基づいて、ニューラルネットワークの演算をバッチ処理し、推論結果 Y を求める（ステップ S13）。より詳細には、行列演算部 130 において、入力データ X' と重みデータ W' との積和演算が行われる。具体的には、乗算器 132 が入力データ X' と重みデータ W' との乗算を行う。乗算結果は加算器 133 によって加算され、行列演算結果 A が求められる。行列演算結果 A に対しては、活性化関数演算部 131 により活性化関数が適用され、推論結果 Y が出力される（ステップ S14）。

[0084] 上記処理によって推論処理装置 1A は、画像データや音声などの時系列データを入力データ X として、学習済みのニューラルネットワークを用いた入力データ X の特徴の推論を行うことができる。

[0085] 次に、本実施の形態に係る推論処理装置 1A におけるデータの転送時間について、図 15 を用いて説明する。図 15 の上段に示すように、バス幅が 3

2ビットの場合、32ビットの入力データXの送信において1つの32ビットデータしか転送することができない。一方、図15の下段に示すように、32ビットの入力データXが8ビットの入力データX'に変換された場合には、4つの8ビットデータを転送することができる。

[0086] このように、メモリ制御部11が記憶部12から入力データXおよび重みデータWを読み出して転送する際に、低ビット精度のデータに変換したデータを転送するので、転送時間を削減することができる。

[0087] 以上説明したように、第2の実施の形態に係る推論処理装置1Aによれば、推論演算部13に入力される入力データXおよび重みデータWをより低ビット精度のデータに変換するので、キャッシュの使用率を向上させ、データバス帯域のボトルネックを減らすことができる。

[0088] また、推論処理装置1Aでは、低ビット精度の入力データX'および重みデータW'を用いてニューラルネットワークの演算を行うので、演算に必要な乗算器132および加算器133の数を削減することができる。その結果として、推論処理装置1Aは、より少ないハードウェアリソースによって実現でき、装置全体としての回路規模を小さくすることができる。

[0089] また、推論処理装置1Aでは、使用するハードウェアリソースを削減することができるため、消費電力および発熱を低減できる。

[0090] また、推論処理装置1Aでは、より低ビット精度の入力データX'および重みデータW'を用いてニューラルネットワークの演算を行うので、より高いクロック周波数で処理を行うことで、処理を高速化することができる。

[0091] また、推論処理装置1Aでは、32ビットよりも低ビット精度の入力データX'および重みデータW'を用いてニューラルネットワークの演算を行うので、32ビットで演算を行う場合と比較して、より多くの並列化やバッチ処理が可能となり、処理を高速化できる。

[0092] [第3の実施の形態]

次に、本発明の第3の実施の形態について説明する。なお、以下の説明では、上述した第1および第2の実施の形態と同じ構成については同一の符号

を付し、その説明を省略する。

[0093] 第1および第2の実施の形態では、1つの推論演算部13によってニューラルネットワークの演算処理が行われる場合について説明した。これに対して、第3の実施の形態では、複数の推論演算部13a、13bを用いて、図6のサンプルコードの破線の枠60に示す推論演算を並列処理する。以下、第1および第2の実施の形態と異なる構成を中心に説明する。

[0094] 図16に示すように、推論処理装置1Bは、バッチ処理制御部10、メモリ制御部11、記憶部12、および複数の推論演算部13a、13bを備える。

[0095] 本実施の形態では、推論演算部13a、13bが例えばK個（Kは2以上かつBatch（バッチサイズ）以下の整数、但しBatchは2以上の場合。）設けられている。推論演算部13a、13bは、それぞれが備える行列演算部130において、メモリ制御部11によって転送された入力データXおよび重みデータWの行列演算を行い、行列演算結果Aをそれぞれ出力する。

[0096] さらに、複数の推論演算部13a、13bのそれぞれが備える活性化関数演算部131において、行列演算結果Aに対して活性化関数が適用されて、出力である推論結果Yが求められる。

[0097] より具体的には、設定されたバッチサイズに応じた入力データXがBatch個の場合、入力データXはBatch行N列となる。図6のサンプルコードにおける破線の枠60に示すように、設定されたバッチサイズに応じた入力データXのデータ数分の推論結果Yを求めるためにBatch回繰り返すことが必要な演算が、本実施の形態ではK並列で行われる。

[0098] 以上説明したように、第3の実施の形態に係る推論処理装置1Bによれば、K個の推論演算部13a、13bが設けられ、Batch回繰り返すことが必要なニューラルネットワークの演算をK並列に行うので、繰り返し演算の回数が削減され、推論演算の処理を高速化できる。

[0099] [第4の実施の形態]

次に、本発明の第4の実施の形態について説明する。なお、以下の説明では、上述した第1から第3の実施の形態と同じ構成については同一の符号を付し、その説明を省略する。

[0100] 第1から第3の実施の形態では、推論演算部13が行列演算部130を1つだけ備え、行列の積和演算を行う場合について説明した。これに対して第4の実施の形態では、推論演算部13Cが複数の行列演算部130a、130bを備え、図6に示したサンプルコードの破線の枠61に示す行列の積和演算を並列に実行する。以下、第1から第3の実施の形態と異なる構成を中心に説明する。

[0101] 図17に示すように、本実施の形態に係る推論演算部13Cは、複数の行列演算部130a、130b、および1つの活性化関数演算部131を備える。本実施の形態に係る推論処理装置1が備えるその他の構成は、図1に示す推論処理装置1と同様である。

[0102] 推論演算部13Cは、K個（Kは2以上かつN以下の整数）の行列演算部130a、130bを備える。K個の行列演算部130a、130bは、入力データXと重みデータWとの行列演算をK並列で実行し、行列演算結果Aを出力する。前述したように、入力データXの成分がM個で、重みデータWのデータサイズが $M \times N$ である場合に、これらの行列の積和演算をN回繰り返すことで、バッチサイズ（Batch） $\times N$ のデータサイズを有する行列演算結果Aの1行分の計算を完了する。

[0103] 例えば、図7Aおよび図7Bで説明したように、 $M = N = 2$ 、およびBatch = 1であり、2個（ $K = 2$ ）の行列演算部130a、130bの場合を考える。行列演算部130a、130bのそれぞれには、M個の入力データXが入力される。例えば、行列演算部130aには、重みデータWの1列目の成分 W_{11} 、 W_{21} が入力され、行列演算部130bには、重みデータWの2列目の成分 W_{12} 、 W_{22} が入力される。メモリ制御部11は、行列演算部130a、130bの個数に応じて重みデータWの振り分けを制御することができる。

[0104] 行列演算部 130a は積和演算を行い、行列演算結果 A の成分 a1 を出力する。一方、行列演算部 130b も同様に積和演算を行い、行列演算結果 A の成分 a2 を出力する。行列演算部 130a、130b の演算結果は、活性化関数演算部 131 に入力されて、活性化関数にかけられ、推論結果 Y が決定される。

[0105] 以上説明したように、第 4 の実施の形態によれば、K 個の行列演算部 130a、130b が行列演算を K 並列に行うので、行列演算結果 A の 1 行分の行列演算における繰り返しの計算回数を削減することができる。特に、上述した具体例のように $K = N$ の場合、計算の繰り返しが不要となり、行列演算の処理時間を削減することができる。結果として、推論処理装置 1 の推論処理を高速化できる。

[0106] なお、第 4 の実施の形態に係る複数の行列演算部 130a、130b は、第 3 の実施の形態と組み合わせてもよい。第 3 の実施の形態で説明した複数の推論演算部 13a、13b のそれぞれが、複数の行列演算部 130a、130b を備えることにより、推論演算をより高速化することができる。

[0107] [第 5 の実施の形態]

次に、本発明の第 5 の実施の形態について説明する。なお、以下の説明では、上述した第 1 から第 4 の実施の形態と同じ構成については同一の符号を付し、その説明を省略する。

[0108] 第 1 から第 4 の実施の形態では、行列演算部 130 が 1 つの乗算器 132 と加算器 133 とを備える場合について説明した。これに対して、第 5 の実施の形態では、行列演算部 130D が複数の乗算器 132a、132b と加算器 133a、133b とを備え、図 6 のサンプルコードの破線の枠 62 で示す行列演算の内部処理を並列に行う。

[0109] 図 18 に示すように、行列演算部 130D は、K 個（K は 2 以上かつ M 以下の整数）の乗算器 132a、132b と加算器 133a、133b とを備える。本実施の形態に係る推論処理装置 1 の他の構成は、第 1 の実施の形態と同様である（図 1）。なお、以下においては、説明の簡単のため、 $M = 3$

の場合を例に挙げて説明する。

- [0110] 行列演算部 130D は、入力データ X と重みデータ W との積和演算を行って、行列演算結果 A の 1 行分の成分を計算する。行列演算部 130D は、 K 個の乗算器 132a、132b および加算器 133a、133b において、積和演算を K 並列で行う。行列演算では、 M 個の成分を有する入力データ X とデータサイズが $M \times N$ の重みデータ W との積和演算が行われる。
- [0111] 例えば、 $M=3$ の場合に、2 個 ($K=2$) の乗算器 132a、132b および加算器 133a、133b が設けられている場合を考える。なお、入力データ X は $[x_1, x_2, x_3]$ で表される。また、重みデータ W は、例えば、 3×2 ($M \times N$) のデータサイズを有する場合を考える。重みデータ W の第 1 列目は、 W_{11}, W_{21}, W_{31} で表される。また、行列演算結果 A は 2 つの成分を有し、 $A[a_1, a_2]$ で表される。
- [0112] この場合、例えば、乗算器 132a に入力データ X の成分 x_1 および重みデータ W の成分 W_{11} が入力される。一方、乗算器 132b には、入力データ X の成分 x_2 および重みデータ W の成分 W_{21} 、ならびに入力データ X の成分 x_3 および重みデータの成分 W_{31} が入力される。
- [0113] 乗算器 132a、132b のそれぞれは、乗算結果を出力する。上記具体例においては、乗算器 132a は、乗算結果 $x_1 W_{11}$ を出力し、乗算器 132b は、乗算結果 $x_2 W_{21}$ と乗算結果 $x_3 W_{31}$ を出力する。加算器 133b は、乗算器 132b の乗算結果 $x_2 W_{21}$ と乗算結果 $x_3 W_{31}$ とを加算する。加算器 133a は、乗算器 132a の乗算結果 $x_1 W_{11}$ と加算器 133b の加算結果 $x_2 W_{21} + x_3 W_{31}$ とを加算して、行列演算結果 A の成分 a_1 を出力する。
- [0114] 以上説明したように、第 5 の実施の形態によれば、行列演算部 130D において、 K 個の乗算器 132a、132b が入力データ X と重みデータ W との行列積を K 並列で実行するので、行列演算結果 A の各成分を計算する際の繰り返しの計算回数を削減することができる。特に、 $K=M$ の場合、1 回の計算で行列演算結果 A の 1 成分を出力することができる。結果として、行列

演算の処理時間を削減し、推論処理装置 1 における処理を高速化できる。

[0115] なお、第 5 の実施の形態は、第 3 および第 4 の実施の形態と組み合わせてもよい。例えば、第 3 の実施の形態の複数の推論演算部 1 3 a、1 3 b それぞれの行列演算部 1 3 0 が、本実施の形態に係る複数の乗算器 1 3 2 a、1 3 2 b を備えることで、第 3 の実施の形態に係る構成のみを採用した場合と比較して、推論演算をより高速化することができる。

[0116] また、第 4 の実施の形態に係る複数の行列演算部 1 3 0 a、1 3 0 b のそれぞれが本実施の形態に係る複数の乗算器 1 3 2 a、1 3 2 b を備えることで、第 4 の実施の形態に係る構成のみを採用した場合と比較して行列演算をより高速化することができる。

[0117] 第 3 から第 5 の実施の形態の構成をそれぞれ単独で採用した場合、例えば、バッチサイズ $B a t c h$ 、推論結果 Y の成分の数 N 、および入力データ X の成分の数 M の関係が、 $B a t c h > B > M$ であれば、第 3 の実施の形態に係る推論処理装置 1 B において最も処理が高速化できる。次いで、第 4 の実施の形態、第 5 の実施の形態の順に処理を高速化することができる。

[0118] なお、本実施の形態において $M = 2$ の場合には、加算器 1 3 3 を 1 つ設ければよい。その場合においても、乗算処理を並列に実行するので、行列演算を高速化することができる。本実施の形態は特に M が 4 以上の場合により有効である。

[0119] [第 6 の実施の形態]

次に、本発明の第 6 の実施の形態について説明する。なお、以下の説明では、上述した第 1 から第 5 の実施の形態と同じ構成については同一の符号を付し、その説明を省略する。

[0120] 第 1 から第 5 の実施の形態では、重みデータ W は予め記憶部 1 2 に記憶されている場合について説明した。これに対して、第 6 の実施の形態に係る推論処理装置 1 E は、通信ネットワーク NW を介して重みデータ W を受信する無線通信部 1 5 を備える。

[0121] 図 1 9 に示すように、第 6 の実施の形態に係る推論処理装置 1 E は、バッ

チ処理制御部 10、メモリ制御部 11、記憶部 12、推論演算部 13、および無線通信部 15 を備える。

[0122] 無線通信部 15 は、通信ネットワーク NW を介して外部のクラウドサーバなどから、推論処理装置 1 E で用いられるニューラルネットワークモデルの重みデータ W を受信し、記憶部 12 に格納する。例えば、推論処理装置 1 E で用いるニューラルネットワークモデルの重みデータ W を再度学習して更新した場合など、無線通信部 15 は、更新された重みデータ W を無線通信によりダウンロードし、記憶部 12 に記憶されている古い重みデータ W を書き換える。

[0123] また、推論処理装置 1 E において、別のニューラルネットワークモデルを用いて推論処理を行うような場合に、無線通信部 15 は、外部のクラウドサーバなどから受信した新しい学習済みのニューラルネットワークの重みデータ W を受信し、記憶部 12 に記憶させる。

[0124] このように、第 6 の実施の形態に係る推論処理装置 1 E によれば、ニューラルネットワークモデルの重みデータ W が書き換え可能であり、推論処理装置 1 E において最適な重みデータ W を用いることができるので、入力データ X の変動などにより推論精度が低下することを防止できる。

[0125] 以上、本発明の推論処理装置および推論処理方法における実施の形態について説明したが、本発明は説明した実施の形態に限定されるものではなく、請求項に記載した発明の範囲において当業者が想定し得る各種の変形を行うことが可能である。

[0126] 例えば、本発明の推論処理装置における推論演算部を除く各機能部は、コンピュータとプログラムによっても実現でき、プログラムを記録媒体に記録することも、ネットワークを通して提供することも可能である。

符号の説明

[0127] 1…推論処理装置、2…センサ、10…バッチ処理制御部、11…メモリ制御部、12…記憶部、13…推論演算部、120…入力データ格納部、121…学習済み NN 格納部、15…無線通信部、130…行列演算部、13

1…活性化関数演算部、132…乗算器、133…加算器、101…バス、
102…プロセッサ、103…主記憶装置、104…通信インターフェース
、105…補助記憶装置、106…入出力装置、107…表示装置。

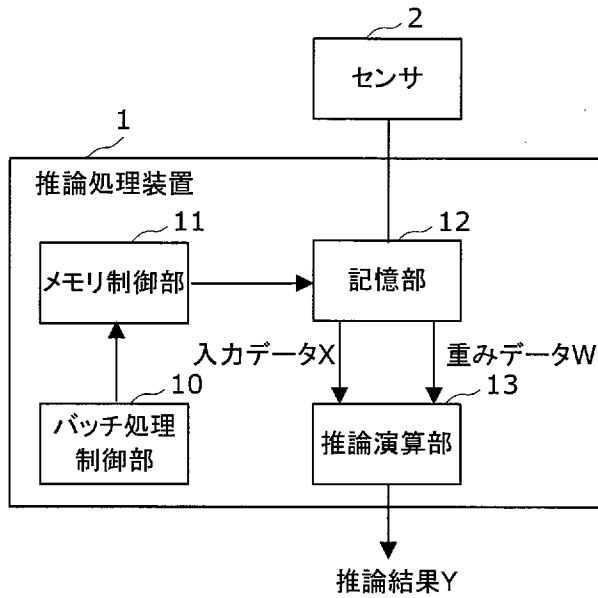
請求の範囲

- [請求項1] 入力データを記憶する第1記憶部と、
 ニューラルネットワークの重みを記憶する第2記憶部と、
 前記入力データに関する情報に基づいて、バッチサイズを設定する
バッチ処理制御部と、
 設定された前記バッチサイズに応じた前記入力データを前記第1記憶部から読み出すメモリ制御部と、
 前記バッチサイズに応じた前記入力データおよび前記重みを入力として、前記ニューラルネットワークの演算をバッチ処理して、前記入力データの特徴を推論する推論演算部と
 を備える推論処理装置。
- [請求項2] 請求項1に記載の推論処理装置において、
 前記バッチ処理制御部は、推論演算に用いられるハードウェアリソースに関する情報に基づいて、前記バッチサイズを設定することを特徴とする推論処理装置。
- [請求項3] 請求項1または請求項2に記載の推論処理装置において、
 前記推論演算部は、
 前記入力データと前記重みとの行列演算を行う行列演算部と、
 前記行列演算部による行列演算結果に対して活性化関数を適用する
活性化関数演算部と
 を備え、
 前記行列演算部は、
 前記入力データと前記重みとを乗算する乗算器と、
 前記乗算器による乗算結果を加算する加算器と
 を有する
 ことを特徴とする推論処理装置。
- [請求項4] 請求項3に記載の推論処理装置において、
 前記行列演算部は、複数具備され、行列演算を並列に行うことを特

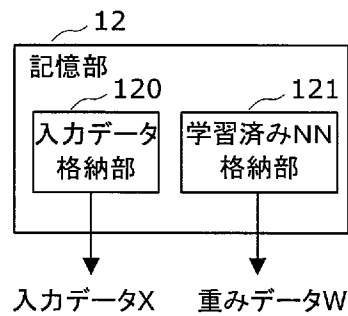
徴とする推論処理装置。

- [請求項5] 請求項3または請求項4に記載の推論処理装置において、
前記行列演算部は、前記乗算器および前記加算器をそれぞれ複数具備し、乗算および加算を並列に行うことを特徴とする推論処理装置。
- [請求項6] 請求項1から5のいずれか1項に記載の推論処理装置において、
前記推論演算部に入力される前記入力データおよび前記重みのデータタイプを変換するデータ変換部をさらに備えることを特徴とする推論処理装置。
- [請求項7] 請求項1から6のいずれか1項に記載の推論処理装置において、
前記推論演算部は、複数具備され、推論演算を並列に行うことを特徴とする推論処理装置。
- [請求項8] 第1記憶部に記憶された入力データに関する情報に基づいて、バッチサイズを設定する第1ステップと、
設定された前記バッチサイズに応じた前記入力データを前記第1記憶部から読み出す第2ステップと、
前記バッチサイズに応じた前記入力データおよび第2記憶部に記憶されたニューラルネットワークの重みを入力として、前記ニューラルネットワークの演算をバッチ処理して、前記入力データの特徴を推論する第3ステップと
を備える推論処理方法。

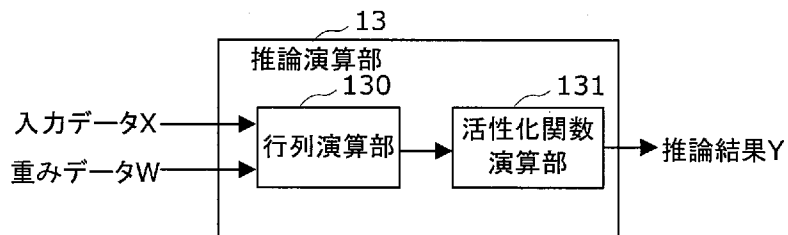
[図1]



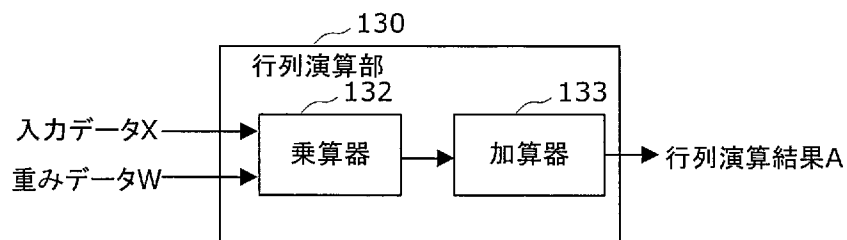
[図2]



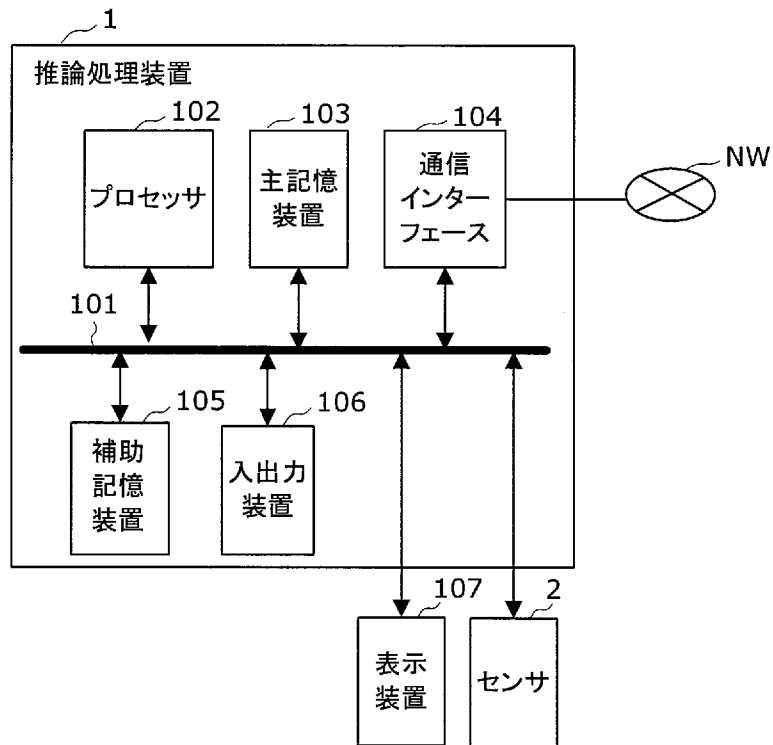
[図3]



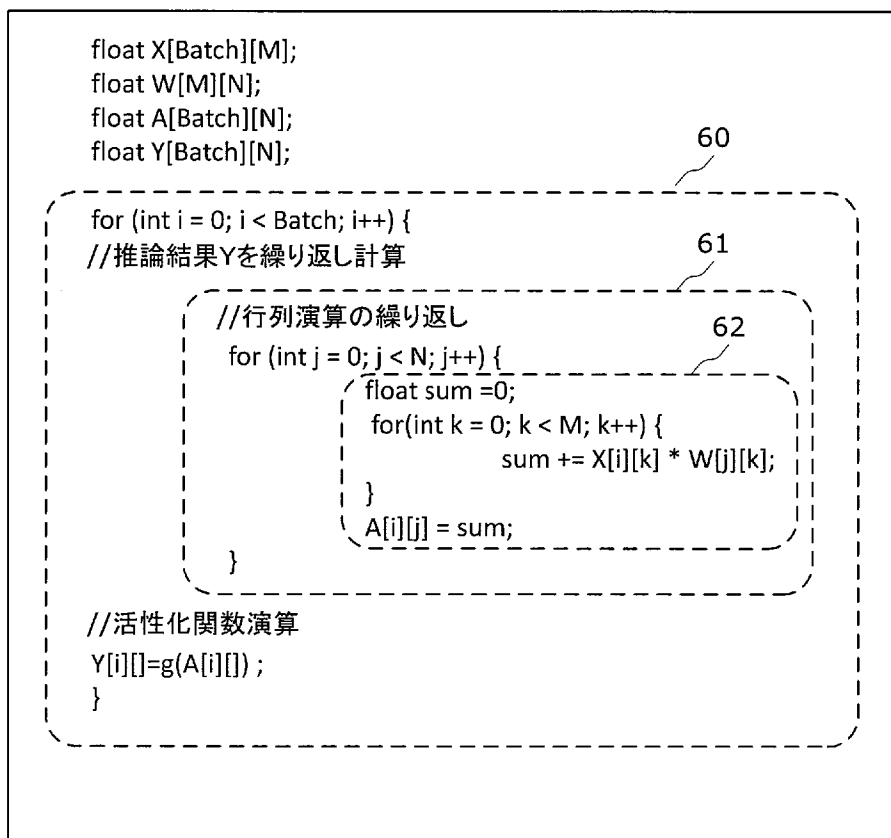
[図4]



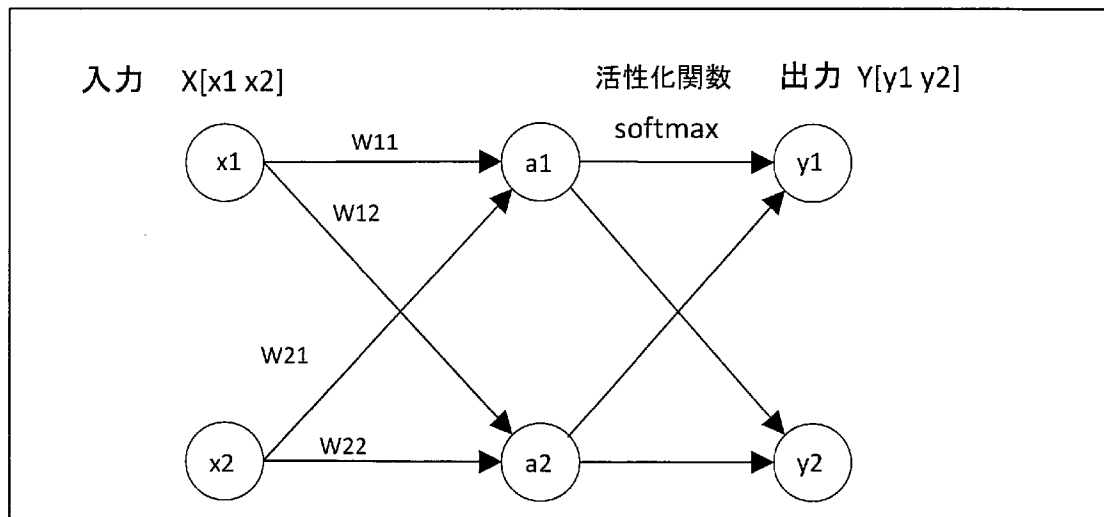
[図5]



[図6]



[図7A]



[図7B]

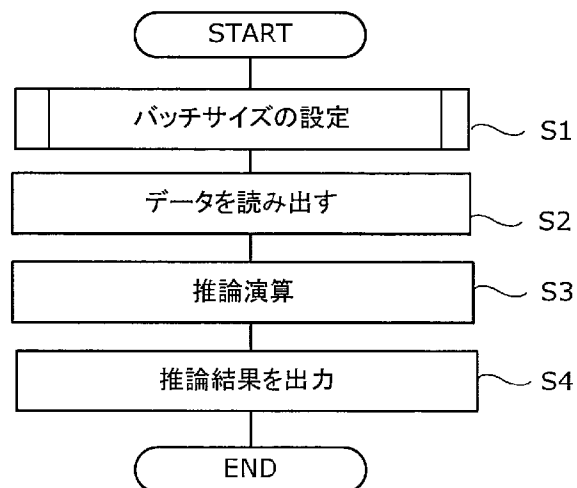
Batch =1 の場合

$$X[x1 \ x2] \times W \begin{bmatrix} w11 & w12 \\ w21 & w22 \end{bmatrix} = A[a1 \ a2]$$

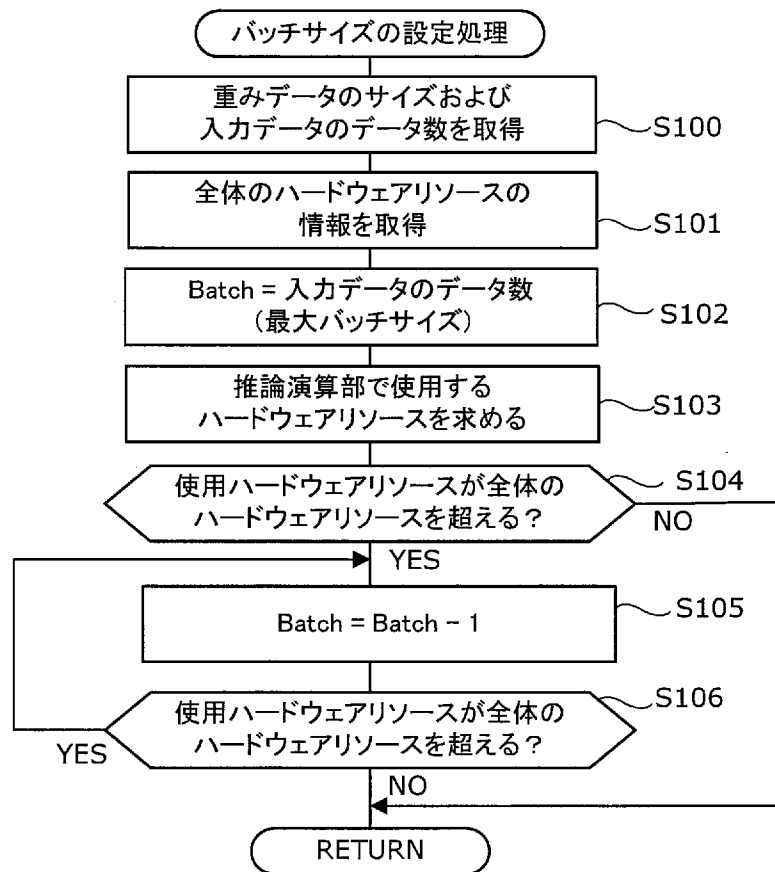
$$Y[y1 \ y2] = \text{softmax}(A[a1 \ a2])$$

softmax関数: $y_k = \frac{\exp(a_k)}{\sum_{i=1}^n \exp(a_i)}$ for $k = 1, \dots, n$

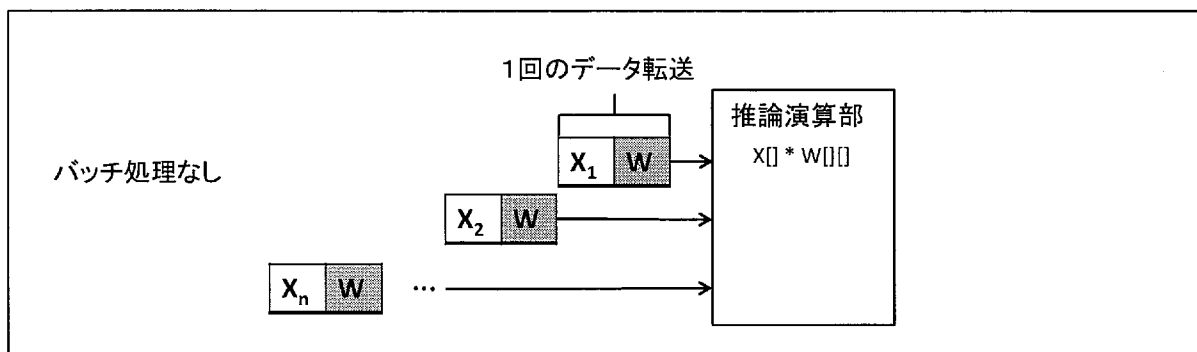
[図8]



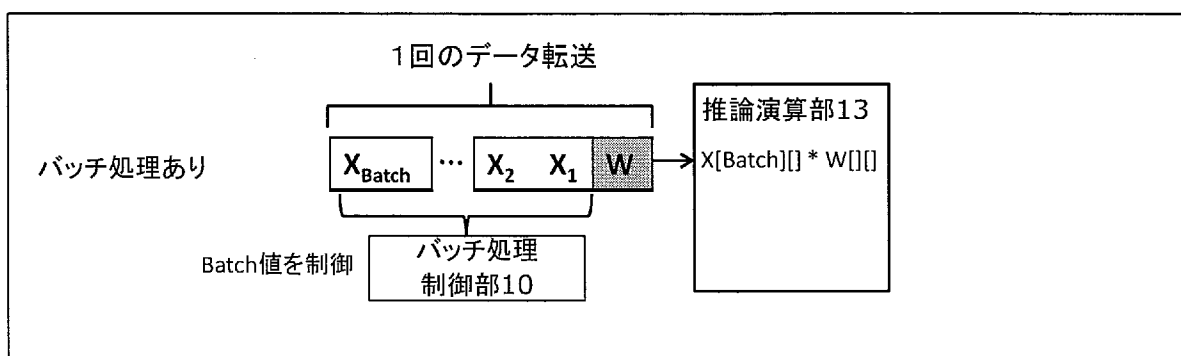
[図9]



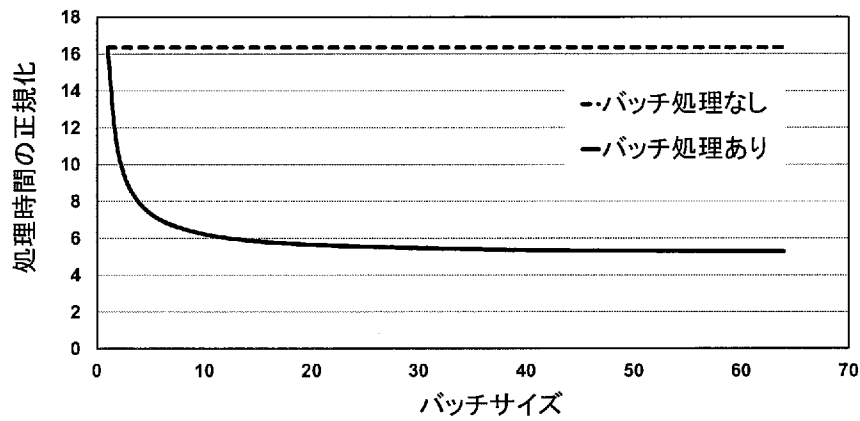
[図10]



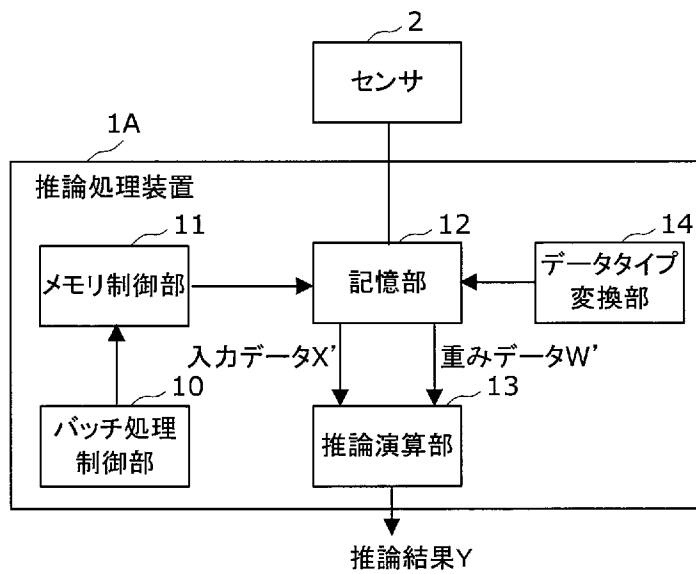
[図11]



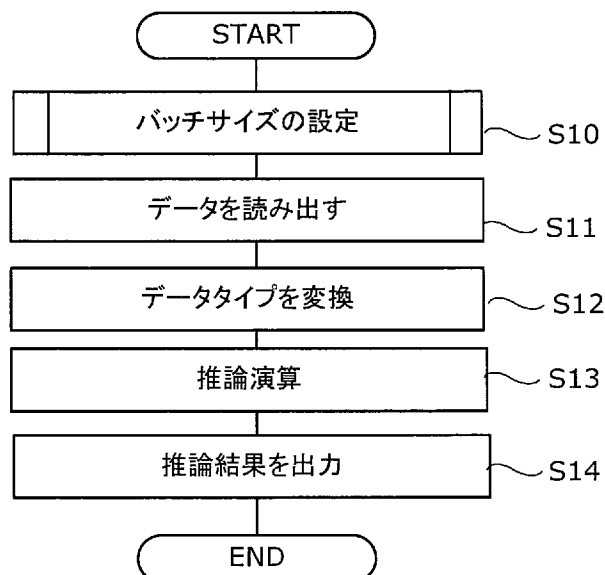
[図12]



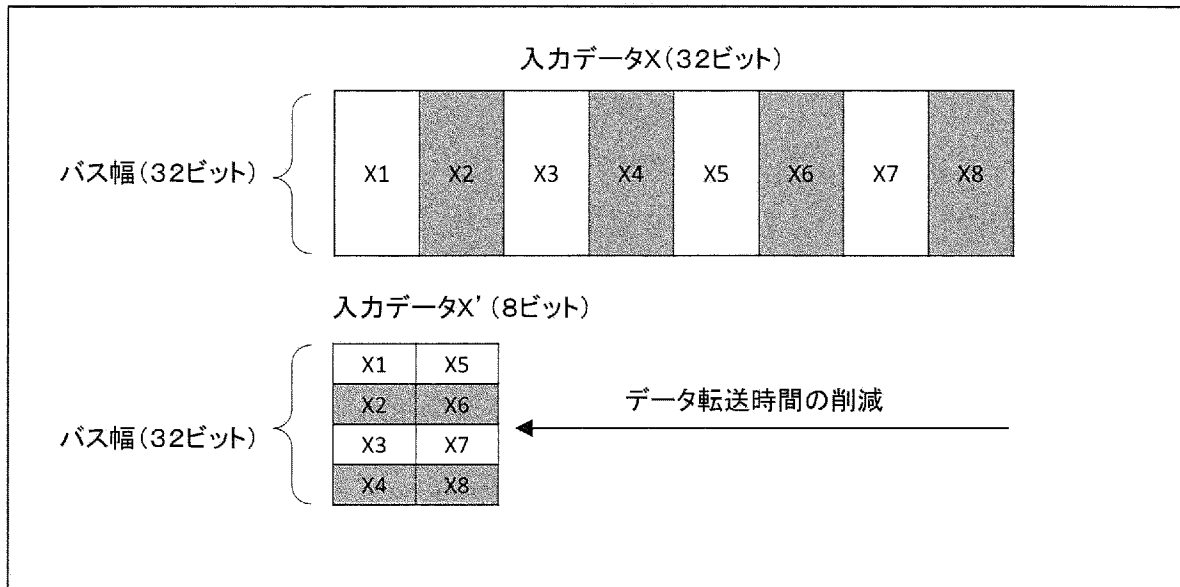
[図13]



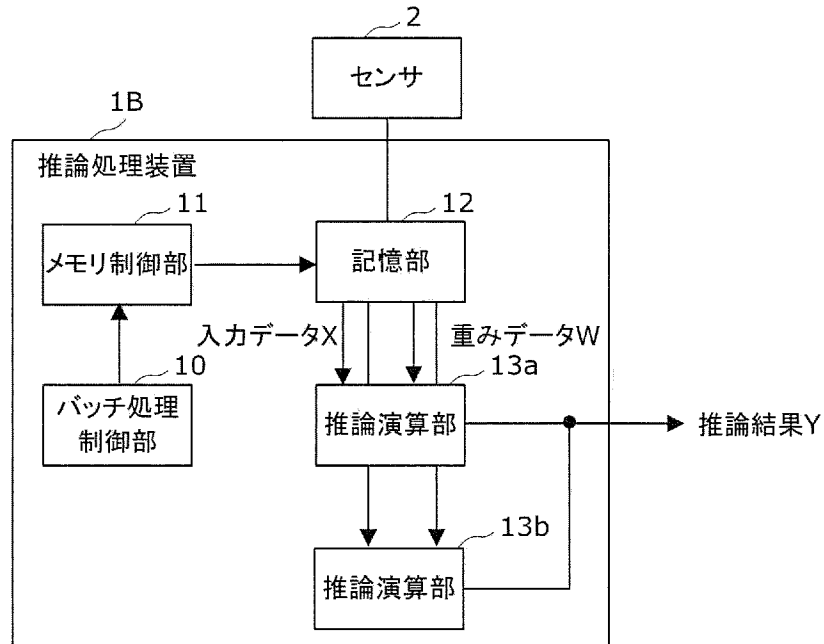
[図14]



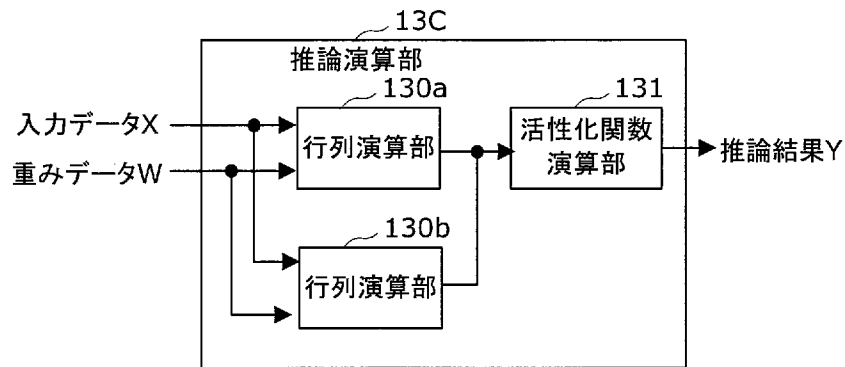
[図15]



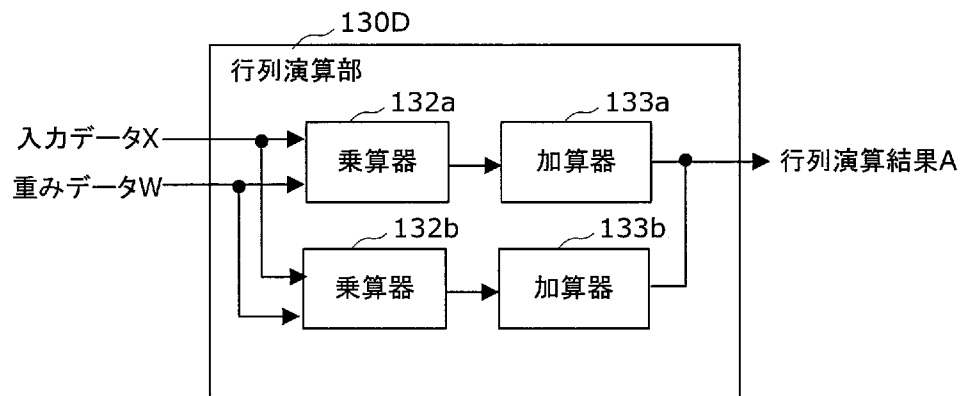
[図16]



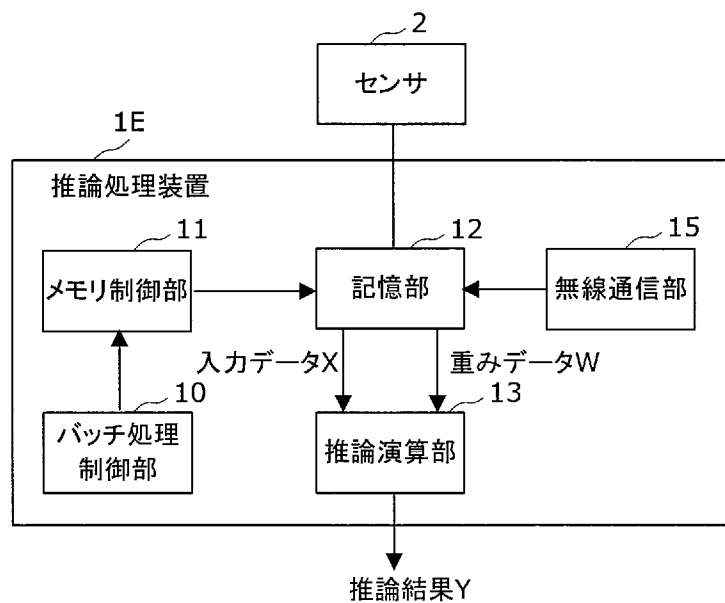
[図17]



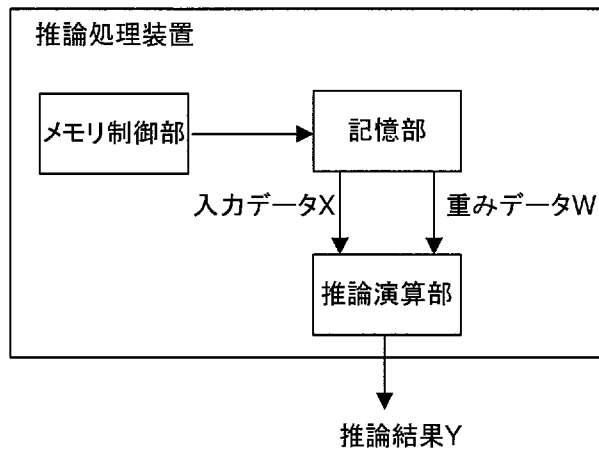
[図18]



[図19]



[図20]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2019/050832

A. CLASSIFICATION OF SUBJECT MATTER

G06N 3/08 (2006.01) i

FI: G06N3/08

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)

G06N 3/00-99/00

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 1971-2020

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.
Y	JP 2018-173814 A (FUJITSU LTD.) 08.11.2018 (2018-11-08) paragraphs [0090]-[0097], fig. 13-15	1-8
Y	渡邊雄也ほか, ディープラーニングによるヘイトスピーチ動画の判定, DEIM Forum 2017 論文集 [online], 31 March 2017, [retrieval date 21 February 2020], Internet: <URL:http://db-event.jpn.org/deim2017/papers/166.pdf>, in particular, "4. Experiment settings", non-official translation (WATANABE, Yuya et al., "The judgment of hate speech videos by deep learning", Proceedings of DEIM Forum 2017 [online])	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

"I" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search
21 February 2020 (21.02.2020)Date of mailing of the international search report
03 March 2020 (03.03.2020)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/JP2019/050832

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	岩室伸哉ほか, 異方性を考慮した画像の符号化に基づく時系列データ分類, 電子情報通信学会技術研究報告, 08 September 2017, vol. 117, no. 210, pp. 137-142, ISSN 0913-5685, in particular, p. 139, (IWAMURA, Shinya et al., "Time Series Classification based on Image Encoding Considering Anisotropy", IEICE technical report)	1-8
Y	三品気吹ほか, 高精度かつ高効率な秘密ロジスティック回帰の設計と実装, コンピュータセキュリティシンポジウム 2018 論文集, 15 October 2018, pp. 1229-1236, in particular, p. 1235, (MISHINA, Ibuki et al., "Designs and Implementations of Efficient and Accurate Secret Logistic Regression", Proceedings of Computer Security Symposium 2018)	1-8
Y	西川由理ほか, 一般物体検出YOLOの分散深層学習による性能評価, 情報処理学会研究報告[online], 20 September 2018, [retrieval date 21 February 2020], vol. 2018-HPC-166, no. 12, ISSN 2188-8841, Internet: <URL:https://ipsj.ixsq.nii.ac.jp/ej/?action=repository_uri&item_id=191420&file_id=1&file_no=1>, in particular, p. 3, (NISHIKAWA, Yuri et al., "Performance Evaluation of Object Detection Algorithm YOLO using Distributed Deep Learning", IPSJ SIG Technical Reports [online])	2-7
Y	竹中慎治ほか, スーパーコンピュータを活用したディープラーニング分散学習の高効率開発基盤, パナソニック技報[online], 15 May 2018, [retrieval date 21 February 2020], vol. 64, no. 1, pp. 39-44, Internet: <URL:https://www.panasonic.com/jp/corporate/technology-design/ptj/pdf/v6401/p0109.pdf>, in particular, pp. 41-42, (TAKENAKA, Shinji et al., "Highly Efficient Infrastructure for R&D by Utilizing Distributed Deep Learning on a Supercomputer", Panasonic technical journal [online])	2-7
Y	WO 2017/073000 A1 (PREFERRED NETWORKS, INC.) 04.05.2017 (2017-05-04) paragraphs [0197]-[0200]	6
Y	WO 2017/185412 A1 (CAMBRICON TECHNOLOGIES CO., LTD.) 02.11.2017 (2017-11-02) abstract	6

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/JP2019/050832

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
JP 2018-173814 A	08 Nov. 2018	US 2018/0285698 A1 paragraphs [0139]- [0153], fig. 13-15	
WO 2017/073000 A1	04 May 2017	US 2018/0349772 A1 paragraphs [0296]- [0313]	
WO 2017/185412 A1	02 Nov. 2017	US 2019/0122094 A1 abstract EP 3451164 A1 CN 107330515 A KR 10-2019-0003613 A CN 109934331 A	

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														
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														
最小限資料以外の資料で調査を行った分野に含まれるもの <table border="0"> <tr> <td>日本国実用新案公報</td> <td>1922 - 1996年</td> </tr> <tr> <td>日本国公開実用新案公報</td> <td>1971 - 2020年</td> </tr> <tr> <td>日本国実用新案登録公報</td> <td>1996 - 2020年</td> </tr> <tr> <td>日本国登録実用新案公報</td> <td>1994 - 2020年</td> </tr> </table>			日本国実用新案公報	1922 - 1996年	日本国公開実用新案公報	1971 - 2020年	日本国実用新案登録公報	1996 - 2020年	日本国登録実用新案公報	1994 - 2020年				
日本国実用新案公報	1922 - 1996年													
日本国公開実用新案公報	1971 - 2020年													
日本国実用新案登録公報	1996 - 2020年													
日本国登録実用新案公報	1994 - 2020年													
国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）														
C. 関連すると認められる文献														
引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求項の番号												
Y	JP 2018-173814 A（富士通株式会社）08.11.2018（2018 - 11 - 08） 段落[0090]-[0097]，図13-15	1-8												
Y	渡邊雄也ほか，ディープラーニングによるヘイトスピーチ動画の判定，DEIM Forum 2017 論文集 [online]，2017.03.31 [検索日 2020.02.21]，インターネット：<URL: http://db-event.jp/2017/papers/166.pdf > 特に「4. 実験設定」	1-8												
Y	岩室伸哉ほか，異方性を考慮した画像の符号化に基づく時系列データ分類，電子情報通信学会技術研究報告，2017.09.08，第117巻，第210号，pp.137-142，ISSN 0913-5685 特にp.139	1-8												
Y	三品気吹ほか，高精度かつ高効率な秘密ロジスティック回帰の設計と実装，コンピュータセキュリティシンポジウム2018論文集，2018.10.15，pp.1229-1236 特にp.1235	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。														
<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 21.02.2020		国際調査報告の発送日 03.03.2020												
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												

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
Y	西川由理ほか，一般物体検出YOL0の分散深層学習による性能評価，情報処理学会 研究報告 [online]，2018.09.20 [検索日 2020.02.21]，第2018-IPC-166巻，第12 号，ISSN 2188-8841，インターネット：<URL: https://ipsj.ixsq.nii.ac.jp/ej/? action=repository_uri&item_id=191420&file_id=1&file_no=1 > 特にp.3	2-7
Y	竹中慎治ほか，スーパーコンピュータを活用したディープラーニング分散学習の高 効率開発基盤，パナソニック技報 [online]，2018.05.15 [検索日 2020.02.21]，第 64巻，第1号，pp.39-44，インターネット：<URL: https://www.panasonic.com/jp/ corporate/technology-design/ptj/pdf/v6401/p0109.pdf > 特にpp.41-42	2-7
Y	WO 2017/073000 A1 (株式会社Preferred Networks) 04.05.2017 (2017 - 05 - 04) 段落[0197]-[0200]	6
Y	WO 2017/185412 A1 (CAMBRICON TECHNOLOGIES CO., LTD.) 02.11.2017 (2017 - 11 - 02) Abstract	6

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/JP2019/050832

引用文献			公表日	パテントファミリー文献	公表日
JP	2018-173814	A	08.11.2018	US 2018/0285698 A1 段落[0139]-[0153], 図 13-15	
WO	2017/073000	A1	04.05.2017	US 2018/0349772 A1 段落[0296]-[0313]	
WO	2017/185412	A1	02.11.2017	US 2019/0122094 A1 Abstract EP 3451164 A1 CN 107330515 A KR 10-2019-0003613 A CN 109934331 A	