(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2021-111016
(P2021-111016A)
(43)公開日　令和3年8月2日(2021.8.2)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 7/499	(2006.01)	G06F	7/499	101	5B013
G06F 9/38	(2006.01)	G06F	9/38	370A	5B056
G06F 17/10	(2006.01)	G06F	17/10	S	
G06N 3/063	(2006.01)	G06N	3/063		

審査請求 未請求　請求項の数 8　ＯＬ　（全 19 頁）

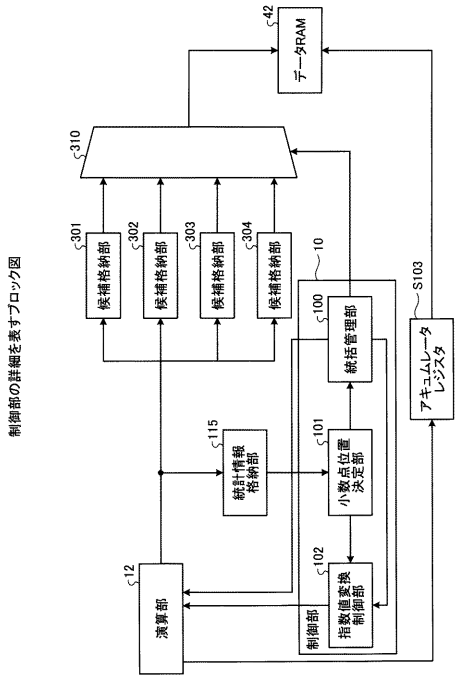
(21)出願番号　特願2020-886 (P2020-886)
(22)出願日　令和2年1月7日 (2020.1.7)

(71)出願人　000005223
富士通株式会社
神奈川県川崎市中原区上小田中４丁目１番１号
(74)代理人　110002147
特許業務法人酒井国際特許事務所
(72)発明者　依田　勝洋
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内

Ｆターム(参考)　5B013 DD01
5B056 BB72　BB73

(54)【発明の名称】演算処理装置、演算処理装置の制御方法及び演算処理プログラム

(57)【要約】
【課題】Ｄｅｆｉｎｅ　ｂｙ　Ｒｕｎにより深層学習を行う際の固定小数点を用いた学習における学習精度を向上させつつ学習時間を短縮する演算処理装置、演算処理装置の制御方法及び演算処理プログラムを提供する。
【解決手段】演算部１２は、演算を実行する。小数点位置決定部１０１は、演算の演算結果の統計情報を基に特定小数点位置を決定する。統括管理部１００は、演算部１２による候補小数点位置を用いた第１演算の結果である第１演算結果を取得し、第１演算結果の統計情報を基に小数点位置決定部１０１により決定された特定小数点位置を取得し、候補小数点位置及び特定小数点位置を基に、第１演算結果又は演算部１２による特定小数点位置を用いた第２演算の結果である第２演算結果を最終演算結果として取得する。
【選択図】図５



【特許請求の範囲】
【請求項１】
演算を実行する演算部と、
前記演算の演算結果の統計情報を基に特定小数点位置を決定する小数点位置決定部と、
前記演算部による候補小数点位置を用いた第１演算の結果である第１演算結果を取得し、前記第１演算結果の統計情報を基に前記小数点位置決定部により決定された前記特定小数点位置を取得し、前記候補小数点位置及び前記特定小数点位置を基に、前記第１演算結果又は前記演算部による前記特定小数点位置を用いた第２演算の結果である第２演算結果を最終演算結果として取得する管理部と
を備えたことを特徴とする演算処理装置。
【請求項２】
前記管理部は、前記特定小数点位置に一致する前記候補小数点位置が存在する場合、前記特定小数点位置に一致する前記候補小数点位置を用いた前記第１演算の結果である前記第１演算結果を前記最終演算結果とすることを特徴とする請求項１に記載の演算処理装置。
【請求項３】
前記管理部は、前記特定小数点位置に一致する前記候補小数点位置が存在しない場合、前記特定小数点位置を用いた前記第２演算の結果である前記第２演算結果を取得し、前記第２演算結果を前記最終演算結果とすることを特徴とする請求項１又は２に記載の演算処理装置。
【請求項４】
前記管理部は、予め決められた前記候補小数点位置を保持することを特徴とする請求項１～３のいずれか一つに記載の演算処理装置。
【請求項５】
前記管理部は、それぞれに対応する所定演算が決められた連続する複数の処理層毎に、対応する前記所定演算を前記第１演算及び前記第２演算として前記演算部に実行させて前記最終演算結果の取得を繰り返し、且つ、特定処理層における前記最終演算結果の取得に用いた前記候補小数点位置又は前記特定小数点位置を基に前記特定処理層の次の前記処理層における前記候補小数点位置を生成することを特徴とする請求項１～３のいずれか一つに記載の演算処理装置。
【請求項６】
前記管理部は、前記第１演算及び前記第２演算として前記演算部に所定積和演算を実行させる場合、前記所定積和演算を基に前記候補小数点位置を生成することを特徴とする請求項１～５のいずれか一つに記載の演算処理装置。
【請求項７】
演算回路を有する演算処理装置の制御方法であって、
前記演算回路による候補小数点位置を用いた第１演算の結果である第１演算結果を取得し、
前記第１演算結果の統計情報を基に特定小数点位置を決定し、
前記候補小数点位置及び前記特定小数点位置を基に、前記第１演算結果又は前記演算回路による前記特定小数点位置を用いた第２演算の結果である第２演算結果を最終演算結果として取得する
ことを特徴とする演算処理装置の制御方法。
【請求項８】
演算回路による候補小数点位置を用いた第１演算の結果である第１演算結果を算出させ、
前記第１演算結果の統計情報を基に特定小数点位置を決定し、
前記候補小数点位置及び前記特定小数点位置を基に、前記第１演算結果又は前記演算回路による前記特定小数点位置を用いた第２演算の結果である第２演算結果を最終演算結果として取得する

処理をコンピュータに実行させることを特徴とする演算処理プログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、演算処理装置、演算処理装置の制御方法及び演算処理プログラムに関する。
【背景技術】
【０００２】
今日、深層学習（ディープラーニング）へのニーズが高まっている。深層学習においては、乗算、積和演算、ベクトル乗算を含む様々な演算が実行される。ところで、深層学習では、個々の演算精度への要求は、他のコンピュータ処理ほど厳密ではない。例えば、従来の信号処理等では、プログラマは極力桁あふれを発生させないようにコンピュータプログラムを開発する。一方、深層学習では、大きな値がある程度飽和することは許容される。これは、深層学習では、複数の入力データを畳み込み演算するときの係数（重み）の調整が主な処理となり、入力データのうち極端なデータは重視されないことが多いためである。また、大量のデータを繰り返し用いて係数を調整するため、一度飽和された値も、学習の進行に合わせて桁調整を行なうことで、飽和されずに係数調整に反映できるようになるためである。
【０００３】
そこで、このような深層学習の特性を考慮し、深層学習用の演算処理装置のチップ面積の削減及び電力性能の向上等を図るため、浮動小数点数を用いずに固定小数点数による演算を用いることが考えられる。これは、浮動小数点数演算よりも固定小数点数を用いた演算の方が回路構成を簡素にできるためである。
【０００４】
また、近年、ディープラーニング用の専用アクセラレータの開発が盛んになっている。そこで、専用アクセラレータにおける演算の面積効率を上げるためにも、固定小数点数による演算を用いることが好ましい。例えば、演算ビット数を例えば３２ビットの浮動小数点数を８ビットの固定小数点数に減らして面積あたりの演算性能を向上させたハードウェアが開発されている。３２ビットの浮動小数点数を８ビットの固定小数点数に減らすことで、単純に面積当たり４倍の性能が得られる。このように、十分な精度を有する実数を少ないビット数で表現する処理は量子化と呼ばれる。
【０００５】
ただし、固定小数点数は、ダイナミックレンジが狭いため浮動小数点数より演算精度が劣化する場合がある。そのため、深層学習においても、小さな値を表現する精度、すなわち有効桁数について配慮が求められる。そこで、演算結果のビット位置の統計情報を用いて固定小数点の有効桁数を決定し小数点位置を最適化する従来技術が存在する。
【０００６】
従来技術では、前のイテレーション（iteration）の統計情報を用いて次のイテレーションの小数点位置が決定され、決定された小数点位置を用いて次のイテレーションの演算が実行される。イテレーションは、ミニバッチとも呼ばれる。
【０００７】
また、統計情報を用いて固定小数点における小数点位置を決定する技術として、最下位ビット位置から最上位ビット位置までの範囲を示す情報及び符号ビット位置から最下位ビットのビット位置まで範囲を示す情報を用いて小数点位置を決定する従来技術がある。また、固定小数点演算を行う技術として、指定された小数点位置を表すデータに基づいて出力された演算結果に対して丸め処理及び飽和処理を実行しつつ固定小数点演算を行う従来技術がある。
【先行技術文献】
【特許文献】
【０００８】
【特許文献１】特開２０１８－１２４６８１号公報

【特許文献２】特開２０１９－７４９５１号公報
【特許文献３】特開２００９－２７１５９８号公報
【発明の概要】
【発明が解決しようとする課題】
【０００９】
しかしながら、最近の深層学習のフレームワーク、特にｐｙＴｏｒｃｈやｃｈａｉｎｅｒで、Ｄｅｆｉｎｅ　ｂｙ　Ｒｕｎと呼ばれる処理方式が導入される機会が増加した。以下では、Ｄｅｆｉｎｅ　ｂｙ　ＲｕｎをＤｂＲと省略して表記する。ＤｂＲでは、ニューラルネットの構造となる計算グラフの構築が、深層学習の処理を実行しながら行われる。そして、ＤｂＲでは、最短で学習のイテレーション毎に計算グラフが変わる。そのため、過去に推定した小数点位置を記憶することが困難である。また、計算グラフが変わるということは、ある層を通過する際の計算グラフが複数あり、特定のイテレーションにおいて、その層でどちらの計算グラフが用いられるか特定困難ということである。なお、ＤｂＲではない従来型の深層学習における演算処理はＤｅｆｉｎｅ　ａｎｄ　Ｒｕｎと呼ばれ、学習開始時に計算グラフが確定する。
【００１０】
ＤｂＲにより深層学習を行う場合、仮に前回のイテレーションにおける統計情報を用いるとしても、前回自体が存在しないあるいは前回の統計情報が実際にはいくつものイテレーションも前の情報となる場合がある。このため、ＤｂＲにより深層学習を行う場合、過去の統計情報を用いると学習が破たんするおそれがあり、過去の統計情報を用いて小数点位置を決定することは困難である。
【００１１】
そこで、現在のレイヤの演算を実行し、その演算結果の統計情報から小数点位置を決定し、求めた小数点位置を用いて再度演算を実行する方法が考えられる。しかし、この方法では、確実に２度同じ演算を行うことになり学習時間が長くなってしまうという問題がある。
【００１２】
また、最下位ビット位置から最上位ビット位置までの範囲を示す情報及び符号ビット位置から最下位ビットのビット位置まで範囲を示す情報を用いて小数点位置を決定する技術でも、過去の統計情報を用いることからＤｂＲを用いた深層学習への適用は困難である。また、指定された小数点位置を表すデータに基づいて出力された演算結果に対して丸め処理及び飽和処理を行う従来技術では、小数点位置の決め方が考慮されておらず、ＤｂＲにより深層学習を行うことは困難である。
【００１３】
開示の技術は、上記に鑑みてなされたものであって、Ｄｅｆｉｎｅ　ｂｙ　Ｒｕｎにより深層学習を行う際の固定小数点を用いた学習における学習精度を向上させつつ学習時間を短縮する演算処理装置、演算処理装置の制御方法及び演算処理プログラム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】
【００１４】
本願の開示する演算処理装置、演算処理装置の制御方法及び演算処理プログラムの一つの態様において、演算部は、演算を実行する。小数点位置決定部は、前記演算の演算結果の統計情報を基に特定小数点位置を決定する。管理部は、前記演算部による候補小数点位置を用いた第１演算の結果である第１演算結果を取得し、前記第１演算結果の統計情報を基に前記小数点位置決定部により決定された前記特定小数点位置を取得し、前記候補小数点位置及び前記特定小数点位置を基に、前記第１演算結果又は前記演算部による前記特定小数点位置を用いた第２演算の結果である第２演算結果を最終演算結果として取得する。
【発明の効果】
【００１５】
１つの側面では、本発明は、Ｄｅｆｉｎｅ　ｂｙ　Ｒｕｎにより深層学習を行う際の固

定小数点を用いた学習における学習精度を向上させつつ学習時間を短縮することができる。

【図面の簡単な説明】

【００１６】

【図１】図１は、サーバの概略を表す構成図である。

【図２】図２は、ニューラルネットワークにおける深層学習の一例の図である。

【図３】図３は、ＤｂＲを説明するための図である。

【図４】図４は、演算用回路のブロック図である。

【図５】図５は、制御部の詳細を表すブロック図である。

【図６】図６は、実施例１に係る演算用回路による深層学習の処理のフローチャートである。

【図７】図７は、実施例２に係る演算用回路による深層学習の処理のフローチャートである。

【図８Ａ】図８Ａは、実施例３に係る演算用回路による深層学習の処理のフローチャートである。

【図８Ｂ】図８Ｂは、実施例３に係る演算用回路による深層学習の処理のフローチャートである。

【発明を実施するための形態】

【００１７】

以下に、本願の開示する演算処理装置、演算処理装置の制御方法及び演算処理プログラムの実施例を図面に基づいて詳細に説明する。なお、以下の実施例により本願の開示する演算処理装置、演算処理装置の制御方法及び演算処理プログラムが限定されるものではない。

【実施例１】

【００１８】

図１は、サーバの概略を表す構成図である。サーバ１は、深層学習を実行する。サーバ１は、ＣＰＵ（Central Processing Unit）２、メモリ３、及び演算用回路４を有する。ＣＰＵ２、メモリ３及び演算用回路４は、相互にＰＣＩｅ（Peripheral Component Interconnect Express）バス５で接続される。

【００１９】

ＣＰＵ２は、メモリ３に格納されたプログラムを実行しサーバ１としての各種機能を実現する。例えば、ＣＰＵ２は、ＰＣＩｅバス５経由で制御信号を送信して、演算用回路４が有する制御コアを起動する。また、ＣＰＵ２は、演算で用いるデータ及び実行する演算命令を演算用回路４へ出力して、演算用回路４に演算を行わせる。

【００２０】

演算用回路４は、深層学習におけるレイヤ毎の演算を実行する回路である。ここで、図２を参照して、ニューラルネットワークにおける深層学習の一例について説明する。図２は、ニューラルネットワークにおける深層学習の一例の図である。ニューラルネットワークは、例えば、画像を認識して識別するためのフォワード方向の処理と、フォワード方向の処理で使用するパラメータを決定するバックワード方向の処理を実行する。図２における上部の矢印の紙面に向かって右に進む方向がフォワード方向であり、紙面に向かって左に進む方向がバックワード方向である。

【００２１】

図２のニューラルネットワークは、入力画像に対して、畳み込み層（Convolution Layer）の処理及びプーリング層（Pooling Layer）の処理を実行し、画像の特徴を抽出して画像を識別する。図２の紙面中央に記載した処理はフォワード方向の処理を表す。

【００２２】

図２では、フォワード方向の処理における特徴抽出部において、入力画像に対して畳み込み層の処理及びプーリング層の処理が実行され特徴マップが生成される。その後、識別部において、特徴マップに対して全結合が行われ最終層から識別結果が出力される。畳み

込み層の処理は、畳み込み演算とも呼ばれる。また、プーリング層の処理は、プーリング演算とも呼ばれる。その後、識別結果は、正解データと比較され、比較結果である差分値が得られる。次に、バックワード方向の処理として、差分値からフォワード方向の畳み込み層及び全結合層における各層でのエラー及び各層での次の重みを計算する学習処理が行われる。

【００２３】

深層学習は、ミニバッチと呼ばれる処理の単位に区切られて実行される。ミニバッチとは、学習の対象となる入力データの集合を所定個の組に分割した複数個のデータの組み合わせである。図２では、Ｎ個の画像で１つのミニバッチである。そして、ミニバッチ毎のフォワード方向の処理及びバックワード方向の処理の一連の処理をまとめた単位を１イテレーションという。

【００２４】

さらに、本実施例では、ＤｂＲにより深層学習を実行する。図３は、ＤｂＲを説明するための図である。Ｄｅｆｉｎｅ　ａｎｄ　Ｒｕｎにより深層学習を実行する場合、計算グラフが固定されるため、レイヤ５１、５３、５４、５５と経由するか、レイヤ５１、５２、５５と経由するかは演算を実行する前に決定される。これに対して、ＲｂＤにより演算を行う場合、演算を実行するまで計算グラフが決定されずに、レイヤ５１からレイヤ５２に進むかレイヤ５３に進むかは確率的に決まり、演算時に計算グラフが動的に変化する。そのため、演算用回路４は、演算に用いる小数点位置を前もって決定することが困難である。そこで、以下の手順で演算用回路４は演算を実行する。

【００２５】

図２に戻って説明を続ける。演算用回路４は、深層学習中の所定数のミニバッチ毎に、各層の演算を行うとともに各層の各変数の統計情報を取得して蓄積し、深層学習に用いる変数の固定小数点位置を自動調整する。次に、演算用回路４の詳細について説明する。

【００２６】

図４は、演算用回路のブロック図である。図４に示すように、演算用回路４は、プロセッサ４０及び命令ＲＡＭ（Random　Access　Memory）４１及びデータＲＡＭ４２を有する。

【００２７】

プロセッサ４０は、制御部１０、レジスタファイル１１、演算部１２、統計情報集約部１３、メモリインタフェース１４及びメモリインタフェース１５を有する。メモリインタフェース１４は、プロセッサ４０における命令ＲＡＭ４１に接続するインタフェースである。また、メモリインタフェース１５は、プロセッサ４０におけるデータＲＡＭ４２に接続するインタフェースである。以下の説明では、プロセッサ４０の各部が命令ＲＡＭ４１又はデータＲＡＭ４２にアクセスする場合、メモリインタフェース１４及び１５の仲介を省略して説明する。

【００２８】

命令ＲＡＭ４１は、ＣＰＵ２から送信された命令を格納する記憶装置である。命令ＲＡＭ４１に格納された命令は、制御部１０によりフェッチされて実行される。データＲＡＭ４２は、命令で指定された演算を実行する際に使用するデータを格納する記憶装置である。データＲＡＭ４２に格納されたデータは、演算部１２で実行される演算に使用される。

【００２９】

レジスタファイル１１は、スカラレジスタファイル１１１、ベクタレジスタファイル１１２、アキュムレータレジスタ１１３、ベクタアキュムレータレジスタ１１４、統計情報格納部１１５及び候補格納部３００を有する。

【００３０】

スカラレジスタファイル１１１及びベクタレジスタファイル１１２は、入力されたデータや学習処理の実行途中のデータなど演算に用いるデータを格納する。アキュムレータレジスタ１１３及びベクタアキュムレータレジスタ１１４は、演算部１２が演算を実行する際に、累積などの演算を行う場合にデータを一時的に格納する。

【００３１】
統計情報格納部１１５は、統計情報集約部１３が集約した統計情報を取得して格納する。統計情報は、演算結果の小数点位置に関する情報である。統計情報は、例えば、非符号となる最上位ビット位置の分布、非ゼロの最下位ビット位置の分布、非符号となる最上位ビット位置の最大値又は非ゼロの最下位ビット位置の最小値などを含む複数の情報のうちのいずれか又はその組み合わせである。
【００３２】
候補格納部３００は、操作者が指定したＮ個の候補小数点位置をそれぞれ用いて演算部１２が実行した先行演算のＮ個の演算結果を格納する。候補小数点位置及び先行演算については後で詳細に説明する。
【００３３】
次に、演算部１２について説明する。演算部１２は、スカラユニット１２１及びベクタユニット１２２を有する。
【００３４】
スカラユニット１２１は、制御部１０、レジスタファイル１１及びメモリインタフェース１５に接続される。スカラユニット１２１は、演算器２１１、統計情報取得部２１２及びデータ変換部２１３を有する。本実施例では、スカラユニット１２１は、Ｎ個の候補小数点位置を用いた演算であり、且つ、統計情報を取得するための先行演算を実行する。また、統計情報から求められた小数点位置と一致する候補小数点位置が存在しなければ、スカラユニット１２１は、先行演算の統計情報から決定された小数点位置で演算を実行して演算結果を取得する本演算という２つの演算を実行する。
【００３５】
演算器２１１は、データＲＡＭ４２、スカラレジスタファイル１１１及びアキュムレータレジスタ１１３が保持するデータのうちの１つもしくはいくつかを用いて積和演算などの演算を実行する。この演算器２１１が演算に用いるデータが、「入力データ」の一例にあたる。演算器２１１は、先行演算及び本演算のいずれの演算でも同様の演算を実行する。演算器２１１は、演算結果を表すのに十分なビット幅を用いて演算を実行する。演算器２１１は、データＲＡＭ４２、統計情報取得部２１２及びデータ変換部２１３に演算結果を出力する。
【００３６】
統計情報取得部２１２は、演算結果のデータの入力を演算器２１１から受ける。そして、統計情報取得部２１２は、演算結果のデータから統計情報を取得する。その後、統計情報取得部２１２は、取得した統計情報を統計情報集約部１３へ出力する。ただし、統計情報取得部２１２は、本演算の場合統計情報の取得及び取得した統計情報の出力を行わなくてもよい。
【００３７】
データ変換部２１３は、演算器２１１による演算結果を取得する。次に、データ変換部２１３は、先行演算の場合、Ｎ個の候補小数点位置の入力を制御部１０から受ける。そして、候補少数点位置毎に、データ変換部２１３は、固定小数点数データを取得した各候補少数点位置で指定されたシフト量だけシフトさせる。また、データ変換部２１３は、シフトとともに、上位ビットの飽和処理および下位ビットの丸めを実行する。これにより、データ変換部２１３は、固定小数点数のデータの小数点位置を更新したＮ個の演算結果を算出する。その後、データ変換部２１３は、候補小数点位置を用いたＮ個の演算結果を候補格納部３００に格納する。
【００３８】
また、データ変換部２１３は、本演算の場合、先行演算により取得された統計情報から決定された小数点位置の入力を制御部１０から受ける。そして、データ変換部２１３は、固定小数点数データを取得した少数点位置で指定されたシフト量だけシフトさせる。また、データ変換部２１３は、シフトとともに、上位ビットの飽和処理および下位ビットの丸めを実行する。これにより、データ変換部２１３は、固定小数点数のデータの小数点位置

を更新する。データ変換部２１３は、小数点位置を更新した演算結果をスカラレジスタファイル１１１及びデータＲＡＭ４２に格納する。

【００３９】

ベクタユニット１２２は、制御部１０、レジスタファイル１１及びメモリインタフェース１５に接続される。ベクタユニット１２２は、演算器２２１、統計情報取得部２２２及びデータ変換部２２３の組を複数有する。本実施例では、ベクタユニット１２２は、Ｎ個の候補小数点位置を用いた演算であり、且つ、統計情報を取得するための先行演算を実行する。また、統計情報から求められた小数点位置と一致する候補小数点位置が存在しなければ、ベクタユニット１２２は、先行演算の統計情報から決定された小数点位置で演算を実行して演算結果を取得する本演算という２つの演算を実行する。

【００４０】

演算器２２１は、データＲＡＭ４２、ベクタレジスタファイル１１２又はベクタアキュムレータレジスタ１１４のうちの１つもしくはいくつかが保持するデータを用いて積和演算などの演算を実行する。演算器２２１は、演算結果を表すのに十分なビット幅を用いて演算を実行する。演算器２２１は、先行演算及び本演算のいずれの演算でも同様の演算を実行する。演算器２１１は、データＲＡＭ４２、統計情報取得部２２２及びデータ変換部２２３に演算結果を出力する。

【００４１】

統計情報取得部２２２は、演算結果のデータの入力を演算器２２１から受ける。この時、統計情報取得部２２２は、精度を維持できる十分なビット幅で表された演算結果のデータを取得する。

【００４２】

そして、統計情報取得部２２２は、演算結果のデータから統計情報を取得する。例えば、非符号となる最上位ビット位置を取得する場合、統計情報取得部２２２は、非符号最上位ビット検出器を用いて、非符号となる最上位ビット位置の値を１とし、他のビット位置の値を０とする出力データを生成する。その後、統計情報取得部２２２は、取得した統計情報を統計情報集約部１３へ出力する。ただし、統計情報取得部２２２は、本演算の場合統計情報の取得及び取得した統計情報の出力を行わなくてもよい。

【００４３】

データ変換部２２３は、演算器２２１による演算結果を取得する。次に、データ変換部２２３は、先行演算の場合、Ｎ個の候補小数点位置の入力を制御部１０から受ける。そして、候補少数点位置毎に、データ変換部２２３は、固定小数点数データを取得した各候補少数点位置で指定されたシフト量だけシフトさせる。また、データ変換部２２３は、シフトとともに、上位ビットの飽和処理および下位ビットの丸めを実行する。これにより、データ変換部２２３は、固定小数点数のデータの小数点位置を更新したＮ個の演算結果を算出する。その後、データ変換部２２３は、候補小数点位置を用いたＮ個の演算結果を候補格納部３００に格納する。

【００４４】

データ変換部２２３は、演算器２２１による演算結果を取得する。次に、データ変換部２２３は、本演算の場合、先行演算により取得された統計情報から決定された小数点位置の入力を制御部１０から受ける。そして、データ変換部２２３は、固定小数点数のデータを取得した少数点位置で指定されたシフト量だけシフトさせる。また、データ変換部２２３は、シフトとともに、上位ビットの飽和処理および下位ビットの丸めを実行する。これにより、データ変換部２２３は、固定小数点数のデータの小数点位置を更新する。データ変換部２２３は、小数点位置を更新した演算結果をアキュムレータ１０３に格納し、その後にベクタレジスタファイル１１２及びデータＲＡＭ４２に格納する。

【００４５】

統計情報集約部１３は、演算器２２１による演算結果のデータから取得された統計情報の入力を統計情報取得部２１２から受ける。また、統計情報集約部１３は、各演算器２２１による演算結果のデータから取得されたそれぞれの統計情報の入力を各統計情報取得部

２２２から受ける。統計情報集約部１３は、統計情報取得部２１２から取得した統計情報及び各統計情報取得部２２２から取得した各統計情報を集約して統計情報格納部１１５へ出力する。

【００４６】

次に、制御部１０について説明する。図５は、制御部の詳細を表すブロック図である。ここでは、候補小数点位置が４個の場合を例に説明する。ただし、Ｎは、２以上の整数であればよい。また、Ｎが多いほど後述する再演算の実行を回避することができるが、先行する演算の数が増えまた記憶領域も増大する。制御部１０は、図５に示すように、統括管理部１００、小数点位置決定部１０１及び指数値変換制御部１０２を有する。

【００４７】

ここで、例えば、候補格納部３００は、候補小数点位置の個数に対応する数の候補格納部３０１～３０４を有する。そして、候補格納部３０１～３０４は、演算部１２によりそれぞれの候補小数点位置を用いて実行された先行演算の各演算結果をそれぞれが格納する。

【００４８】

統括管理部１００は、演算部１２による先行演算及び本演算の実行を管理する。統括管理部１００は、深層学習において演算部１２に演算を実行させるレイヤの情報を保持する。統括管理部１００は、演算部１２に演算を実行させるレイヤが次のレイヤに移ると、先行演算の実行を決定する。

【００４９】

次に、統括管理部１００は、操作者が指定した４個の候補小数点位置を取得する。そして、統括管理部１００は、取得した４個の候補小数点位置を指数値変換制御部１０２に通知し、演算部１２に先行演算の実行を指示する。この演算部１２による先行演算が、「第１演算」の一例にあたり、先行演算の演算結果が「第１演算結果」の一例にあたる。

【００５０】

その後、統括管理部１００は、演算部１２による先行演算の実行が完了すると、新たに算出された小数点位置を小数点位置決定部１０１から取得する。そして、統括管理部１００は、小数点位置決定部１０１から取得した小数点位置に一致する候補小数点位置が存在するか否かを判定する。

【００５１】

小数点位置決定部１０１から取得した小数点位置に一致する候補小数点位置が存在する場合、統括管理部１００は、その候補小数点位置を用いて小数点位置が更新された固定小数点数を演算結果として決定する。例えば、図５に示すように、統括管理部１００は、候補格納部３０１～３０４が有する各候補小数点位置を用いて小数点位置が更新された演算結果の中からセレクタ３１０に選択させる。その後、統括管理部１００は、決定した演算結果をそのレイヤにおける最終演算結果としてデータＲＡＭ４２に格納する。

【００５２】

これに対して、小数点位置決定部１０１から取得した小数点位置に一致する候補小数点位置が存在しない場合、統括管理部１００は、本演算の実行を決定する。そして、統括管理部１００は、新たに決定された小数点位置の出力を指数値変換制御部１０２に指示し、演算部１２に本演算の実行を指示する。本演算の演算結果はそのレイヤにおける最終演算結果としてアキュムレータレジスタ１１３を経由してデータＲＡＭ４２に格納される。この演算部１２による本演算が、「第２演算」の一例にあたり、本演算の演算結果が「第２演算結果」の一例にあたる。統括管理部１００は、以上のような先行演算及び本演算を演算部１２に実行させる制御をレイヤ毎に繰り返す。

【００５３】

また、統括管理部１００は、深層学習におけるイテレーションの管理も行う。例えば、所定回数のイテレーションの実行が指示された場合、統括管理部１００は、実行されたイテレーションの回数をカウントし所定回数に達すると学習の終了を決定する。その後、統括管理部１００は、例えば、ＣＰＵ２に学習終了を通知して学習を終了する。この統括管

理部１００が、「管理部」の一例にあたる。

【００５４】

小数点位置決定部１０１は、演算部１２による先行演算が終了すると、統計情報を統計情報格納部１１５から取得する。そして、小数点位置決定部１０１は、取得した統計情報を用いて最適な小数点位置を決定する。その後、小数点位置決定部１０１は、決定した小数点位置を指数値変換制御部１０２へ出力する。小数点位置決定部１０１は、先行計算後の小数点位置の決定処理をレイヤ毎に繰り返す。

【００５５】

指数値変換制御部１０２は、各候補小数点位置を統括管理部１００から取得する。さらに、指数値変換制御部１０２は、Ｎ個の候補小数点位置の出力の指示を統括管理部１００から受ける。そして、指数値変換制御部１０２は、Ｎ個の候補小数点位置を演算部１２に出力する。

【００５６】

その後、演算部１２による先行演算が完了すると、指数値変換制御部１０２は、先行演算の演算結果を用いて新たに決定された小数点位置の入力を小数点位置決定部１０１から受ける。そして、小数点位置決定部１０１が算出した小数点位置に一致する候補小数点位置が存在しない場合、指数値変換制御部１０２は、新たに決定された小数点位置の出力の指示の入力を統括管理部１００から受ける。その後、指数値変換制御部１０２は、新たに決定された小数点位置の情報を演算部１２に出力する。

【００５７】

次に、図６を参照して、本実施例に係る演算用回路４による深層学習の処理の流れを説明する。図６は、実施例１に係る演算用回路による深層学習の処理のフローチャートである。

【００５８】

制御部１０の統括管理部１００は、操作者が指定したＮ個の候補小数点位置を取得する（ステップＳ１０１）。そして、統括管理部１００は、Ｎ個の候補小数点位置を指数値変換制御部１０２へ出力する。

【００５９】

制御部１０の指数値変換制御部１０２は、Ｎ個の候補小数点位置を演算部１２へ出力する。そして、各演算器２１１及び２２１は、Ｎ個の候補小数点位置のそれぞれを用いて、位置入力データを用いて先行演算における演算を実行する（ステップＳ１０２）。各統計情報取得部２１２及び２２２は、対応する各演算器２１１及び２２１による演算結果から統計情報を求める。統計情報集約部１３は、各統計情報取得部２１２及び２２２から統計情報を集約して統計情報格納部１１５に格納する。

【００６０】

データ変換部２１３及び２２３は、演算器２１１及び２２１による演算結果を取得する。そして、データ変換部２１３及び２２３は、Ｎ個の候補小数点位置のそれぞれを用いて小数点位置を更新する（ステップＳ１０３）。データ変換部２１３及び２２３は、Ｎ個の先行演算の演算結果を候補格納部３００に格納する。

【００６１】

制御部１０の小数点位置決定部１０１は、統計情報格納部１１５に格納された統計情報を初期化する。そして、小数点位置決定部１０１は、先行演算における演算結果の統計情報を用いて新たな小数点位置を決定する（ステップＳ１０４）。

【００６２】

制御部１０の統括管理部１００は、Ｎ個の候補小数点位置の中に小数点位置決定部１０１により決定された新たな小数点位置に一致する候補小数点位置が存在するか否かを判定する（ステップＳ１０５）。

【００６３】

新たな小数点位置に一致する候補小数点位置が存在しない場合（ステップＳ１０５：否定）、統括管理部１００は、新たな小数点位置の出力を指数値変換制御部１０２に指示す

るとともに、本演算の実行を演算部１２に指示する。演算部１２の各演算器２１１及び２２１は、入力データを用いて本演算における演算を実行する（ステップＳ１０６）。

【００６４】

演算部１２のデータ変換部２１３及び２２３は、指数値変換制御部１０２から入力された小数点位置で、演算器２１１及び２２１による演算結果の小数点位置を更新する（ステップＳ１０７）。このように、演算部１２は、本演算を実行する。演算部１２は、本演算の演算結果をそのレイヤの演算結果とする。

【００６５】

これに対して、新たな小数点位置に一致する候補小数点位置が存在する場合（ステップＳ１０５：肯定）、統括管理部１００は、新たな小数点位置に一致する候補小数点位置を用いて算出された固定小数点数を候補格納部３００から取得する。そして、統括管理部１００は、取得した固定小数点数をそのレイヤの演算結果とする（ステップＳ１０８）。

【００６６】

その後、制御部１０の統括管理部１００は、実行中の全てのレイヤが終了したか否かを判定する（ステップＳ１０９）。全てのレイヤが終了していない場合（ステップＳ１０９：否定）、統括管理部１００は、次のレイヤの演算を開始させる（ステップＳ１１０）。その後、深層学習の処理はステップＳ１０１へ戻る。

【００６７】

これに対して、全てのレイヤが終了した場合（ステップＳ１０９：肯定）、制御部１０の統括管理部１００は、学習が終了したか否かを判定する（ステップＳ１１１）。

【００６８】

学習が終了していない場合（ステップＳ１１１：否定）、統括管理部１００は、次のイテレーションを開始する（ステップＳ１１２）。その後、深層学習の処理はステップＳ１０１へ戻る。

【００６９】

これに対して、学習が終了した場合（ステップＳ１１１：肯定）、統括管理部１００は、学習完了をＣＰＵ２に通知して学習を終了する。

【００７０】

以上に説明したように、本実施例に係る演算用回路は、予め決められたＮ個の候補小数点位置のそれぞれを用いて先行演算を行い、Ｎ個の先行演算の演算結果を取得する。また、演算用回路は、先行演算における演算の結果から得た統計情報を用いて、入力データを用いた演算に対する適切な小数点位置を新たに決定する。そして、新たな小数点位置と一致する候補小数点位置が存在する場合、演算用回路は、その候補小数点位置を用いて算出した先行演算の演算結果をそのレイヤの演算結果とする。これに対して、新たな小数点位置と一致する候補小数点位置が存在しない場合、演算用回路は、新たな小数点位置を用いて本演算を実行し、その本演算の演算結果をそのレイヤの演算結果とする。このように、演算用回路は、Ｎ個の候補小数点位置を用いて先行演算を投機的に実行し、その投機的演算の演算結果が適切な小数点位置と一致する場合にはその投機的演算の演算結果をそのレイヤの演算結果とすることができる。

【００７１】

これにより、ニューラルネットの構造となる計算グラフの構築が、深層学習の処理を実行しながら行われるＤｅｆｉｎｅ　ｂｙ　Ｒｕｎにより深層学習を行う際に、適切な小数点位置を決定できる。そして、固定小数点を用いた学習における学習精度を向上させることが可能となる。それに加えて、投機的演算が適切な小数点位置を用いた場合の演算と同じ結果となる場合、その投機的演算の演算結果を用いることができるため、本演算の実行を回避することができる。すなわち、投機的演算を行わずに、先行演算の統計情報を用いて適切な小数点位置を求めて本演算を実行することで２回の演算を行う場合に比べて、オーバヘッドとなる演算時間を低減することができる。

【００７２】

ここで、投機的演算のため演算時間の低減の割合は確率的であるが、例えば、投機的演

算の演算結果がレイヤの演算結果として使用できる確率が２０％である場合を考える。１回の演算時間を１基準時間とすると、この場合、本実施例に係る演算用回路の演算時間は、１．２基準時間×０．８＋１回×（１．２＋１）基準時間×０．２と表される。これに対して、２回の演算を行う場合は、２回×１基準時間と表される。すなわち、本実施例に係る演算用回路は、２回演算を行う場合に比べてオーバヘッドを６０％削減することができる。これにより、Ｄｅｆｉｎｅ　ｂｙ　Ｒｕｎにより深層学習を行う際に、固定小数点を用いた学習における学習精度を向上させつつ、演算のオーバヘッドを軽減でき、学習時間を短縮することが可能となる。

【実施例２】

【００７３】

次に、実施例２について説明する。本実施例に係る演算用回路４は、１つ前のレイヤで使用した小数点位置から候補小数点位置を生成することが実施例１と異なる。本実施例に係る演算用回路４も図４及び５で表される。以下の説明では、実施例１と同様の各部の機能については説明を省略する。ここでは、候補小数点位置を４つ用いる場合で説明する。

【００７４】

統括管理部１００は、１つ前のレイヤである前レイヤで用いた小数点位置を取得する。例えば、前レイヤで用いた小数点位置はデータＲＡＭ４２に格納される。そして、統括管理部１００は、前レイヤで用いた小数点位置が指定するビット値に＋１、＋０、－１、－２をしたビット値を候補小数点位置として取得する。その後、統括管理部１００は、候補小数点位置を用いた先行演算を演算部１２に実行させる。この場合、投機的演算の演算結果を使用できる確率を向上させるために、前レイヤで用いた小数点位置を挟むように候補小数点位置を決定することが好ましい。

【００７５】

その後、先行演算における演算の統計情報から小数点位置決定部１０１により決定される新しい小数点位置に一致する候補小数点位置がある場合、統括管理部１００は、その候補小数点位置を用いた先行演算の演算結果をそのレイヤの演算結果とする。また、新しい小数点位置に一致する候補小数点位置がない場合、統括管理部１００は、新しい小数点位置を用いた本演算を演算部１２に実行させ、その演算結果をそのレイヤの演算結果とする。

【００７６】

次に、図７を参照して、本実施例に係る演算用回路４による深層学習の処理の流れを説明する。図７は、実施例２に係る演算用回路による深層学習の処理のフローチャートである。

【００７７】

制御部１０の統括管理部１００は、前レイヤの小数点位置を取得する（ステップＳ２０１）。

【００７８】

制御部１０の指数値変換制御部１０２は、前レイヤの小数点位置が表すビット位置を予め決められたビット数ずらすなどしてＮ個の候補小数点位置を生成する（ステップＳ２０２）。

【００７９】

次に、指数値変換制御部１０２は、Ｎ個の候補小数点位置を演算部１２へ出力する。そして、演算部１２の各演算器２１１及び２２１は、Ｎ個の候補小数点位置のそれぞれを用いて、位置入力データを用いて先行演算における演算を実行する（ステップＳ２０３）。各統計情報取得部２１２及び２２２は、対応する各演算器２１１及び２２１による演算結果から統計情報を求める。統計情報集約部１３は、各統計情報取得部２１２及び２２２から統計情報を集約して統計情報格納部１１５に格納する。

【００８０】

データ変換部２１３及び２２３は、演算器２１１及び２２１による演算結果を取得する。そして、データ変換部２１３及び２２３は、Ｎ個の候補小数点位置のそれぞれを用いて

小数点位置を更新する（ステップＳ２０４）。データ変換部２１３及び２２３は、Ｎ個の先行演算の演算結果を候補格納部３００に格納する。

【００８１】

制御部１０の小数点位置決定部１０１は、統計情報格納部１１５に格納された統計情報を初期化する。そして、小数点位置決定部１０１は、先行演算における演算結果の統計情報を用いて新たな小数点位置を決定する（ステップＳ２０５）。

【００８２】

制御部１０の統括管理部１００は、Ｎ個の候補小数点位置の中に小数点位置決定部１０１により決定された新たな小数点位置に一致する候補小数点位置が存在するか否かを判定する（ステップＳ２０６）。

【００８３】

新たな小数点位置に一致する候補小数点位置が存在しない場合（ステップＳ２０６：否定）、統括管理部１００は、新たな小数点位置の出力を指数値変換制御部１０２に指示するとともに、本演算の実行を演算部１２に指示する。演算部１２の各演算器２１１及び２２１は、入力データを用いて本演算における演算を実行する（ステップＳ２０７）。

【００８４】

演算部１２のデータ変換部２１３及び２２３は、指数値変換制御部１０２から入力された小数点位置で、演算器２１１及び２２１による演算結果の小数点位置を更新する（ステップＳ２０８）。このように、演算部１２は、本演算を実行する。統括管理部１００は、本演算の演算結果をそのレイヤの演算結果とする。

【００８５】

これに対して、新たな小数点位置に一致する候補小数点位置が存在する場合（ステップＳ２０６：肯定）、統括管理部１００は、新たな小数点位置に一致する候補小数点位置を用いて算出された固定小数点数を候補格納部３００から取得する。そして、統括管理部１００は、取得した固定小数点数をそのレイヤの演算結果とする（ステップＳ２０９）。

【００８６】

その後、制御部１０の統括管理部１００は、実行中の全てのレイヤが終了したか否かを判定する（ステップＳ２１０）。全てのレイヤが終了していない場合（ステップＳ２１０：否定）、統括管理部１００は、次のレイヤの演算を開始させる（ステップＳ２１１）。その後、深層学習の処理はステップＳ２０１へ戻る。

【００８７】

これに対して、全てのレイヤが終了した場合（ステップＳ２１０：肯定）、制御部１０の統括管理部１００は、学習が終了したか否かを判定する（ステップＳ２１２）。

【００８８】

学習が終了していない場合（ステップＳ２１２：否定）、統括管理部１００は、次のイテレーションを開始する（ステップＳ２１３）。その後、深層学習の処理はステップＳ２０１へ戻る。

【００８９】

これに対して、学習が終了した場合（ステップＳ２１２：肯定）、統括管理部１００は、学習完了をＣＰＵ２に通知して学習を終了する。

【００９０】

以上に説明したように、本実施例に係る演算用回路は、１つ前のレイヤの小数点位置から候補小数点位置を生成する。深層学習におけるＣＮＮ（Convolutional Neural Network）系のネットワークのように同じようなレイヤ構造が続く場合、小数点位置は安定する傾向にある。そこで、１つ前のレイヤの小数点位置を用いることで、投機的演算の演算結果を使用できる確率を向上させることができる。したがって、オーバヘッドとなる演算時間をより低減することができる。これにより、Ｄｅｆｉｎｅ ｂｙ Ｒｕｎにより深層学習を行う際に、固定小数点を用いた学習における学習精度を向上させつつ、演算のオーバヘッドを軽減でき、学習時間を短縮することが可能となる。

【実施例３】

【００９１】
次に、実施例３について説明する。本実施例に係る演算用回路４は、対象とするレイヤで実行される積和演算から小数点位置を限定して候補小数点位置を生成することが実施例１と異なる。すなわち、本実施例で実行される処理は、積和演算を実行するレイヤを対象とする。本実施例に係る演算用回路４も図４及び５で表される。以下の説明では、実施例１と同様の各部の機能については説明を省略する。ここでは、積和演算を実行するレイヤでの処理において候補小数点位置を４つ用いる場合で説明する。
【００９２】
統括管理部１００は、対象とするレイヤで実行させる積和演算をＫ分割する。例えば、統括管理部１００は、次の数式（１）で示されるように積和演算を４分割する。この、対象とするレイヤで実行させる積和演算が、「所定積和演算」の一例にあたる。
【００９３】
【数１】

$$\sum_N XW = \sum_0^{N1} XW + \sum_{N1}^{N2} XW + \sum_{N2}^{N3} XW + \sum_{N3}^{N4} XW \quad \cdots(1)$$

【００９４】
ここで、数式（１）の左辺は対象とするレイヤで実行される積和演算を表す。そして、数式（１）の右辺が左辺の積和演算を４分割した積和演算である。以下では、分割後の各積和演算を「分割積和演算」と言う。
【００９５】
次に、統括管理部１００は、Ｋ分割後の分割積和演算毎の演算結果を演算部１２に算出させる。次に、統括管理部１００は、各分割積和演算の演算結果の中から最大値となる演算結果を選択する。
【００９６】
ここで、対象とするレイヤの積和演算の演算結果は、各分割積和演算の演算結果の最大値のＫ倍以下となる。そこで、統括管理部１００は、分割積和演算の演算結果のうちの最大値となる演算結果をＫ倍した値を算出する。そして、統括管理部１００は、分割積和演算の演算結果のうちの最大値となる演算結果をＫ倍した値の小数点位置を上限小数点位置として求める。例えば、Ｋ分割後の各積和演算の演算結果の最大値をＶＭＡＸとした場合、統括管理部１００は、Ｂ＝ｌｏｇ_２（ＶＭＡＸ×Ｋ）として上限小数点位置を求める。ここで、Ｂは、上限小数点位置を表す。
【００９７】
対象とするレイヤの積和演算の演算結果の小数点位置は、上限小数点位置よりもビット位置が上になると考えられるので、統括管理部１００は、上限小数点位置が示すビット位置から上のビット位置を候補小数点位置として取得する。例えば、上述したＢが上限小数点位置の場合、統括管理部１００は、Ｂ、Ｂ－１、Ｂ－２及びＢ－３を４つの候補小数点位置として取得する。
【００９８】
次に、統括管理部１００は、Ｋ分割後の各積和演算の演算結果の総和を演算部１２に算出させる。そして、統括管理部１００は、算出された総和に対して候補小数点位置を用いが小数点位置の変更を行わせることで先行演算を演算部１２に実行させる。
【００９９】
その後、先行演算における演算の統計情報から小数点位置決定部１０１により決定される新しい小数点位置に一致する候補小数点位置がある場合、統括管理部１００は、その候補小数点位置を用いた先行演算の演算結果をそのレイヤの演算結果とする。また、新しい小数点位置に一致する候補小数点位置がない場合、統括管理部１００は、新しい小数点位置を用いた本演算を演算部１２に実行させ、その演算結果をそのレイヤの演算結果とする。
【０１００】

次に、図８Ａ及び図８Ｂを参照して、本実施例に係る演算用回路４による深層学習の処理の流れを説明する。図８Ａ及び図８Ｂは、実施例３に係る演算用回路による深層学習の処理のフローチャートである。

【０１０１】

制御部１０の統括管理部１００は、対象のレイヤで積和演算が実行されるか否かを判定する（ステップＳ３０１）。

【０１０２】

対象のレイヤで積和演算が実行される場合（ステップＳ３０１：肯定）、統括管理部１００は、積和演算をＫ分割する（ステップＳ３０２）。

【０１０３】

そして、統括管理部１００は、各分割積和演算の実行を演算部１２に指示する。演算部１２は、分割積和演算毎に演算を実行する（ステップＳ３０３）。

【０１０４】

統括管理部１００は、各分割積和演算の演算結果の最大値を取得する（ステップＳ３０４）。

【０１０５】

指数値変換制御部１０２は、分割積和演算の演算結果の最大値をＫ倍した値の小数点位置を求めて上限小数点位置とする（ステップＳ３０５）。

【０１０６】

指数値変換制御部１０２は、上限小数点位置が表すビット位置から上の位置のＮ個の候補小数点位置を生成する（ステップＳ３０６）。

【０１０７】

次に、指数値変換制御部１０２は、分割積和演算の総和の算出を演算部１２に指示する。演算部１２の各各演算器２１１及び２２１は、分割積和演算の総和を計算する（ステップＳ３０７）。各統計情報取得部２１２及び２２２は、対応する各演算器２１１及び２２１による演算結果から統計情報を求める。統計情報集約部１３は、各統計情報取得部２１２及び２２２から統計情報を集約して統計情報格納部１１５に格納する。

【０１０８】

データ変換部２１３及び２２３は、演算器２１１及び２２１による演算結果を取得する。そして、データ変換部２１３及び２２３は、Ｎ個の候補小数点位置のそれぞれを用いて小数点位置を更新する。データ変換部２１３及び２２３は、算出したＮ個の先行演算の演算結果を候補格納部３００に格納する。

【０１０９】

制御部１０の小数点位置決定部１０１は、統計情報格納部１１５に格納された統計情報を初期化する。そして、小数点位置決定部１０１は、先行演算における演算結果の統計情報を用いて新たな小数点位置を決定する（ステップＳ３０８）。

【０１１０】

制御部１０の統括管理部１００は、Ｎ個の候補小数点位置の中に小数点位置決定部１０１により決定された新たな小数点位置に一致する候補小数点位置が存在するか否かを判定する（ステップＳ３０９）。

【０１１１】

新たな小数点位置に一致する候補小数点位置が存在しない場合（ステップＳ３０９：否定）、統括管理部１００は、新たな小数点位置の出力を指数値変換制御部１０２に指示するとともに、本演算の実行を演算部１２に指示する。演算部１２の各演算器２１１及び２２１は、入力データを用いて本演算における演算を実行する（ステップＳ３１０）。

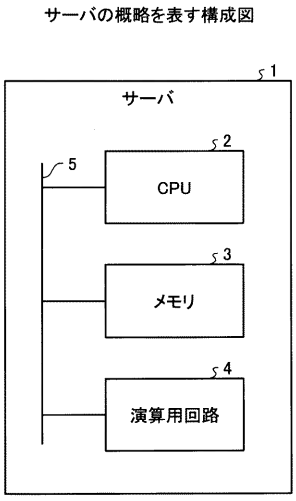
【０１１２】

演算部１２のデータ変換部２１３及び２２３は、指数値変換制御部１０２から入力された小数点位置で、演算器２１１及び２２１による演算結果の小数点位置を更新する（ステップＳ３１１）。このように、演算部１２は、本演算を実行する。統括管理部１００は、本演算の演算結果をそのレイヤの演算結果とする。

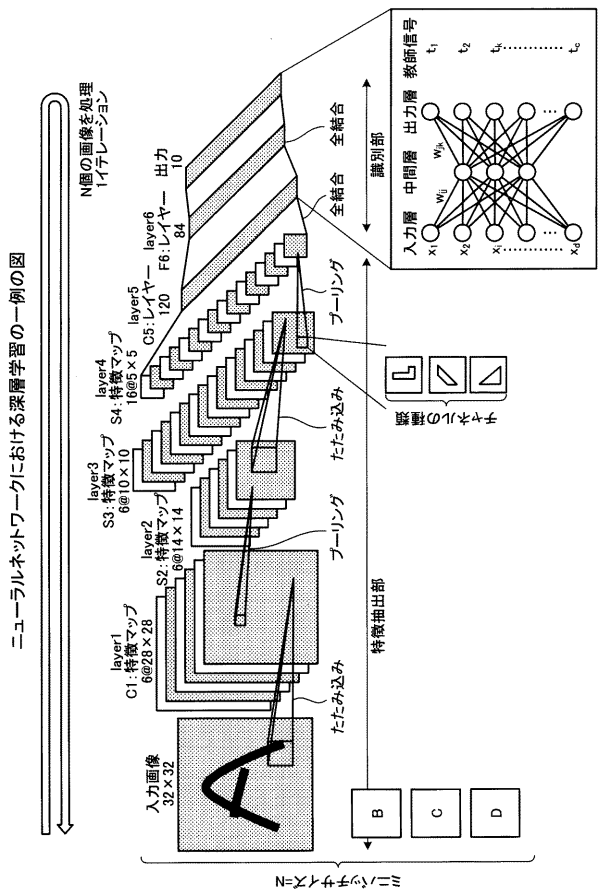
【０１１３】
一方、新たな小数点位置に一致する候補小数点位置が存在する場合（ステップＳ３０９：肯定）、統括管理部１００は、新たな小数点位置に一致する候補小数点位置を用いて算出された固定小数点数をそのレイヤの演算結果として選択する（ステップＳ３１２）。
【０１１４】
これに対して、対象のレイヤで積和演算が実行されない場合（ステップＳ３０１：否定）、統括管理部１００は、他の手順でそのレイヤの演算を実行する（ステップＳ３１３）。ここで、他の手順として、例えば、実施例１又は２で説明した手順を用いることができる。
【０１１５】
その後、制御部１０の統括管理部１００は、実行中の全てのレイヤが終了したか否かを判定する（ステップＳ３１４）。全てのレイヤが終了していない場合（ステップＳ３１４：否定）、統括管理部１００は、次のレイヤの演算を開始させる（ステップＳ３１５）。その後、深層学習の処理はステップＳ３０１へ戻る。
【０１１６】
これに対して、全てのレイヤが終了した場合（ステップＳ３１４：肯定）、制御部１０の統括管理部１００は、学習が終了したか否かを判定する（ステップＳ３１６）。
【０１１７】
学習が終了していない場合（ステップＳ３１６：否定）、統括管理部１００は、次のイテレーションを開始する（ステップＳ３１７）。その後、深層学習の処理はステップＳ３０１へ戻る。
【０１１８】
これに対して、学習が終了した場合（ステップＳ３１６：肯定）、統括管理部１００は、学習完了をＣＰＵ２に通知して学習を終了する。
【０１１９】
以上に説明したように、本実施例に係る演算用回路は、対象のレイヤで実行される積和演算を分割し、分割積和演算の演算結果の最大値から上限小数点位置を求め、求めた上限小数点位置から候補小数点位置を生成する。これにより、制限された範囲の小数点位置の中から候補小数点位置を選定することができ、投機的演算の演算結果を使用できる確率を向上させることができる。したがって、オーバヘッドとなる演算時間をより低減することができる。これにより、Ｄｅｆｉｎｅ　ｂｙ　Ｒｕｎにより深層学習を行う際に、固定小数点を用いた学習における学習精度を向上させつつ、演算のオーバヘッドを軽減でき、学習時間を短縮することが可能となる。
【符号の説明】
【０１２０】
１　サーバ
２　ＣＰＵ
３　メモリ
４　演算用回路
５　ＰＣＩｅバス
１０　制御部
１１　レジスタファイル
１２　演算部
１３　統計情報集約部
１４　メモリインタフェース
１５　メモリインタフェース
４０　プロセッサ
４１　命令ＲＡＭ
４２　データＲＡＭ
１００　統括管理部

１０１　小数点位置決定部
１０２　指数値変換制御部
１１１　スカラレジスタファイル
１１２　ベクタレジスタファイル
１１３　アキュムレータレジスタ
１１４　ベクタアキュムレータレジスタ
１１５　統計情報格納部
１２１　スカラユニット
１２２　ベクタユニット
２１１，２２１　演算器
２１２，２２２　統計情報取得部
２１３，２２３　データ変換部

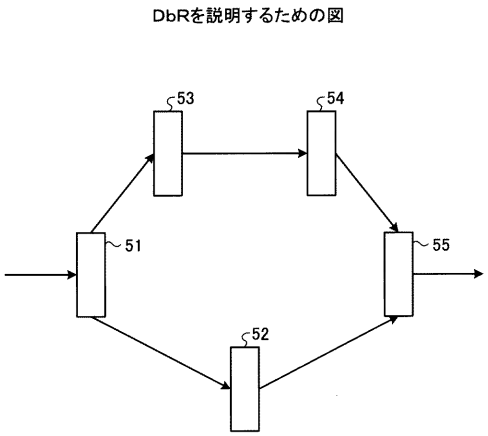
【図１】



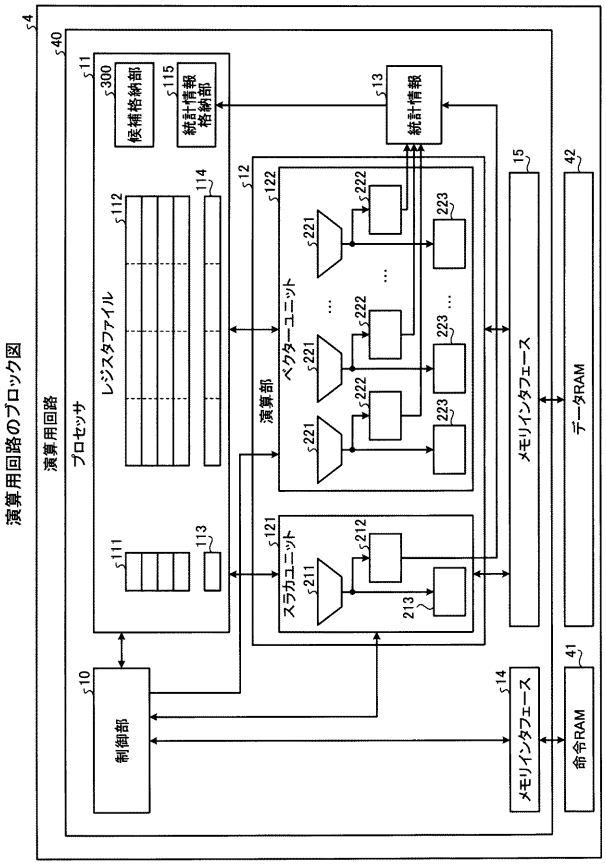
【図２】



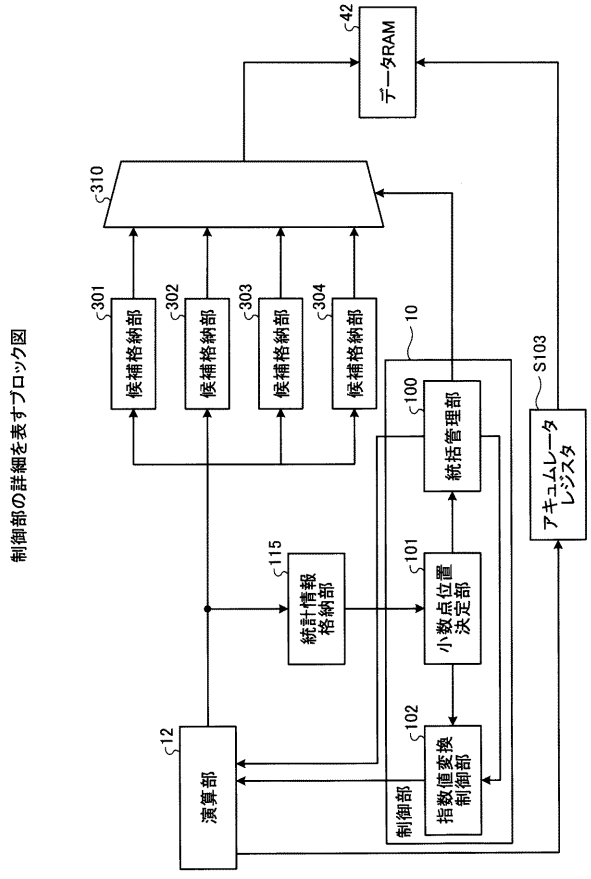
【図３】


DbRを説明するための図
53
54
51
55
52

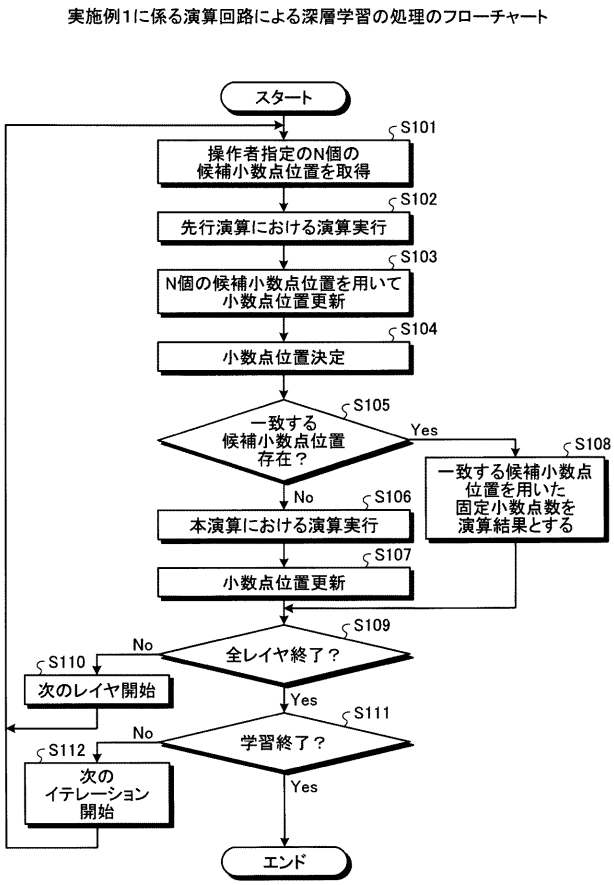
【図４】


演算用回路のブロック図
4
40
演算用回路
プロセッサ
11
300
候補格納部
115
統計情報格納部
13
統計情報
112
レジスタファイル
114
12
122
演算部
ベクターユニット
221
222
223
15
メモリインタフェース
42
データRAM
111
113
121
212
スラカユニット
211
213
10
制御部
14
メモリインタフェース
41
命令RAM

【図５】

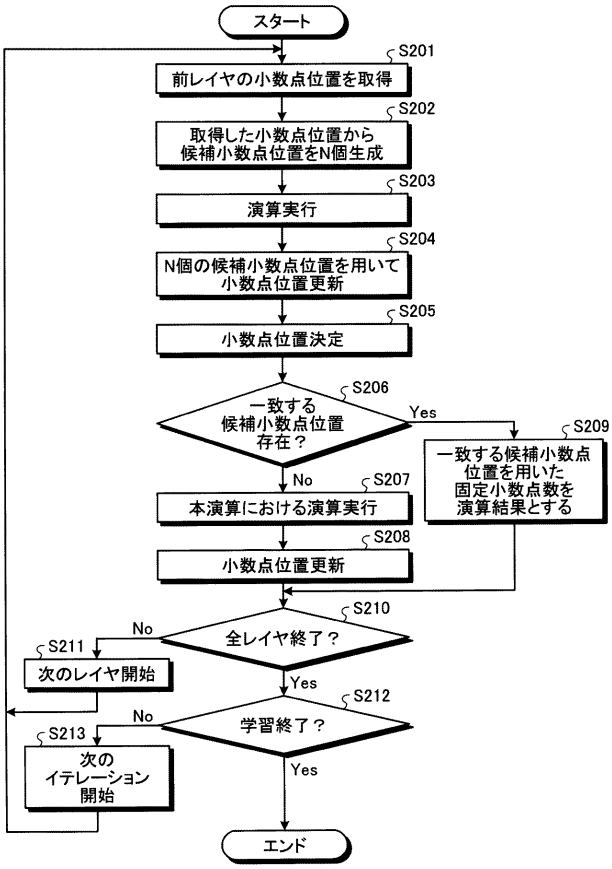

制御部の詳細を表すブロック図
42
データRAM
310
301
候補格納部
302
候補格納部
303
候補格納部
304
候補格納部
10
100
統括管理部
S103
アキュムレータレジスタ
115
統計情報格納部
101
小数点位置決定部
12
演算部
102
指数値変換制御部
制御部

【図６】


実施例1に係る演算回路による深層学習の処理のフローチャート
スタート
S101
操作者指定のN個の候補小数点位置を取得
S102
先行演算における演算実行
S103
N個の候補小数点位置を用いて小数点位置更新
S104
小数点位置決定
S105
一致する候補小数点位置存在？
Yes
S108
一致する候補小数点位置を用いた固定小数点数を演算結果とする
No
S106
本演算における演算実行
S107
小数点位置更新
S109
全レイヤ終了？
No
S110
次のレイヤ開始
Yes
S111
学習終了？
No
S112
次のイテレーション開始
Yes
エンド

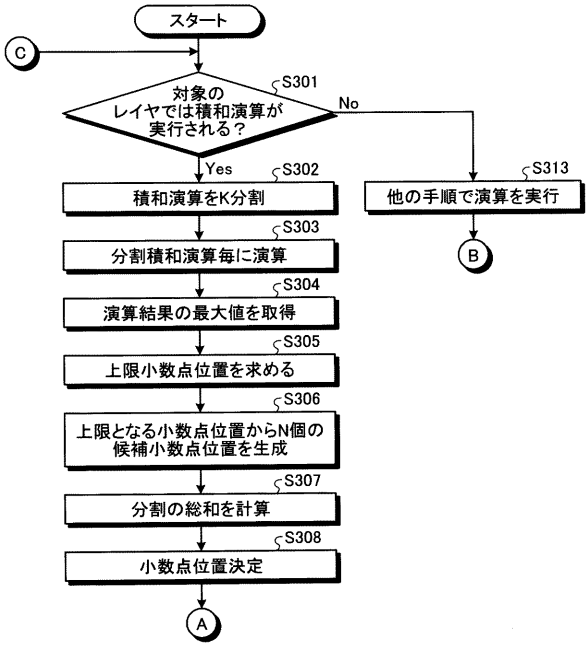
【図７】

実施例2に係る演算用回路による深層学習の処理のフローチャート



【図８Ａ】

実施例3に係る演算用回路による深層学習の処理のフローチャート



【図８Ｂ】

実施例3に係る演算用回路による深層学習の処理のフローチャート