(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2021-5358
(P2021-5358A)
(43)公開日　令和3年1月14日(2021.1.14)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 17/16	(2006.01)	G06F	17/16	K	5B056
G06N 3/04	(2006.01)	G06N	3/04		
G06N 3/02	(2006.01)	G06N	3/02		
G06N 3/10	(2006.01)	G06N	3/10		
G06N 3/06	(2006.01)	G06N	3/06		

審査請求 有　請求項の数 20　ＯＬ　外国語出願　（全 23 頁）

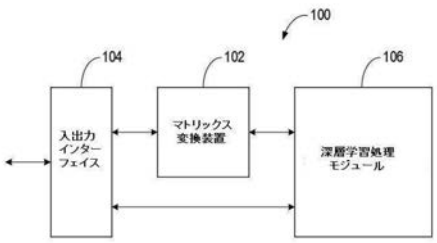
(21)出願番号　特願2020-2746(P2020-2746)
(22)出願日　令和2年1月10日(2020.1.10)
(31)優先権主張番号　201910559362.7
(32)優先日　令和1年6月26日(2019.6.26)
(33)優先権主張国・地域又は機関
中国(CN)

(71)出願人　514322098
ベイジン　バイドゥ　ネットコム　サイエンス　アンド　テクノロジー　カンパニー　リミテッド
中華人民共和国　ペキン　１０００８５，ハイディアン　ディストリクト，　シャンディ　テンス　ストリート，　１０番，バイドゥ　キャンパス　２階
(74)代理人　110001508
特許業務法人　津国

最終頁に続く

(54)【発明の名称】マトリックスを変換するための装置及び方法、データ処理システム

(57)【要約】　（修正有）
【課題】マトリックスを変換するための装置、方法及びデータ処理システムを提供する。
【解決手段】データ処理システム１００において、マトリックス装置は、マトリックスデータを受信し、マトリックスデータに対して第一の循環シフトを行うことによって、第一のデータを生成するように配置されている第一のシフトユニットと、第一のデータの中の各行のデータを、当該行のデータの中の各データと異なる配列の配列順で書き込むように配置されているキャッシュユニットと、キャッシュユニットから第二のデータを読み取り、第二のデータに対して第二の循環シフトを行うことによって、変換後のマトリックスデータを生成するように配置されている第二のシフトユニットとを含む。
【選択図】図１



【特許請求の範囲】
【請求項１】
マトリックスを変換するための装置であって、
マトリックスデータを受信し、前記マトリックスデータに対して第一の循環シフトを行うことによって、第一のデータを生成するように配置されている第一のシフトユニットと、
前記第一のデータの中の各行のデータを、前記行のデータの中の各データと異なる配列の配列順で書き込むことによって、前記第一のデータを第二のデータとして記憶するように配置されているキャッシュユニットと、
前記キャッシュユニットから前記第二のデータを読み取り、前記第二のデータに対して第二の循環シフトを行うことによって、変換後のマトリックスデータを生成するように配置されている第二のシフトユニットと、を含むことを特徴とする装置。
【請求項２】
前記キャッシュユニットは、それぞれが複数の記憶アドレスを有する複数の記憶ユニットグループを含み、前記第一のデータの中の各行のデータを、異なる記憶ユニットグループの異なる記憶アドレスにそれぞれ書き込むことによって、前記第一のデータを前記第二のデータとして記憶するように更に配置されていることを特徴とする請求項１に記載の装置。
【請求項３】
前記第二のシフトユニットは、
前記第二のデータの中の、異なる記憶ユニットグループの同じ記憶アドレスに記憶されているデータをそれぞれ読み取ることによって、前記第二のデータの中の相応する行のデータとし、
前記第二のデータの中の各行のデータに対して前記第二の循環シフトを行うことによって、前記変換後のマトリックスデータの中の相応する行のデータを生成するように更に配置されていることを特徴とする請求項２に記載の装置。
【請求項４】
前記第一のシフトユニットは、前記マトリックスデータの中の第i行のデータを（ｉ－１）桁右循環シフトすることによって、前記第一のデータを生成し、iは、自然数であるように更に配置されており、
前記第二のシフトユニットは、前記第二のデータの中の第i行のデータを（ｉ－１）桁左循環シフトすることによって、前記変換後のマトリックスデータを生成するように配置されていることを特徴とする請求項３に記載の装置。
【請求項５】
前記マトリックスデータで表されるマトリックスは、ｎ行及びｍ列を含み、ｎ及びｍは、それぞれ自然数であり、
前記キャッシュユニットは、
前記第一のデータの中の第１行のデータの中のｍ個の列のデータの中の第jデータを、前記複数の記憶ユニットグループの中の第j記憶ユニットグループの第j記憶アドレスにそれぞれ書き込み、jは、１以上且つｍ以下である自然数であり、
前記第一のデータの中の第i行のデータの中のｍ個の列のデータの中の第jデータを、前記複数の記憶ユニットグループの中の第j記憶ユニットグループの第一の記憶アドレス及び第二の記憶アドレスにそれぞれ書き込み、iは、2以上且つn以下である自然数であり、
jが１以上且つｉ－１以下であるとき、前記第一の記憶アドレスは、第m+j-i+1記憶アドレスであり、
jがi以上且つｍ以下であるとき、前記第二の記憶アドレスは、第j-i+1記憶アドレスであるように更に配置されていることを特徴とする請求項４に記載の装置。
【請求項６】
前記第一のシフトユニットは、
前記マトリックスデータで表されるマトリックスを、ｐ行を含む第一のマトリックス及

びｑ行を含む第二のマトリックス、又は、ｐ列を含む第一のマトリックス及びｑ列を含む第二のマトリックスに分割し、ｐ及びｑは、それぞれ自然数であり、
前記第一のマトリックスのマトリックスデータに対して前記第一の循環シフトを行うことによって、前記第一のマトリックスの前記第一のデータを生成し、
前記第二のマトリックスのマトリックスデータに対して前記第一の循環シフトを行うことによって、前記第二のマトリックスの前記第一のデータを生成するように更に配置されていることを特徴とする請求項１に記載の装置。
【請求項７】
前記キャッシュユニットは、
前記第一のマトリックスの前記第一のデータを、第一の記憶アドレスをスタートアドレスとして前記キャッシュユニットに書き込むことによって、前記第一のマトリックスの前記第一のデータを前記第一のマトリックスの前記第二のデータとして記憶し、
前記第二のマトリックスの前記第一のデータを、第ｋ＋１記憶アドレスをスタートアドレスとして前記キャッシュユニットに書き込むことによって、前記第二のマトリックスの前記第一のデータを前記第二のマトリックスの前記第二のデータとして記憶し、ｋは、ｐ以上の自然数であるように更に配置されていることを特徴とする請求項６に記載の装置。
【請求項８】
前記第二のシフトユニットは、
前記キャッシュユニットから前記第一のマトリックスの前記第二のデータを読み取り、前記第一のマトリックスの前記第二のデータに対して前記第二の循環シフトを行うことによって、変換後の第一のマトリックスデータを生成し、
前記キャッシュユニットから前記第二のマトリックスの前記第二のデータを読み取り、前記第二のマトリックスの前記第二のデータに対して前記第二の循環シフトを行うことによって、変換後の第二のマトリックスデータを生成し、
前記変換後の第一のマトリックスデータ及び前記変換後の第二のマトリックスデータを結合することによって、前記変換後のマトリックスデータを生成するように更に配置されていることを特徴とする請求項７に記載の装置。
【請求項９】
前記キャッシュユニットは、複数の記憶ユニットグループを含み、前記複数の記憶ユニットグループの中の各ｓ個の記憶ユニットグループが１グループの記憶ユニットグループに分割され、各グループの記憶ユニットグループは、複数の記憶アドレスを含み、前記複数の記憶アドレスの中の各ｔ個の記憶アドレスが１グループの記憶アドレスに分割され、
前記キャッシュユニットは、前記第一のデータの中の各ｔ行のデータの中の複数のグループのデータを、異なるグループの記憶ユニットグループの異なるグループの記憶アドレスにそれぞれ書き込むことによって、前記第一のデータを前記第二のデータとして記憶し、前記複数のグループのデータの中の各グループのデータは、ｓ×ｔ個のデータを含み、ｓ及びｔは、自然数であるように配置されていることを特徴とする請求項１に記載の装置。
【請求項１０】
前記第二のシフトユニットは、
前記第二のデータの中の、異なるグループの記憶ユニットグループの同じグループの記憶アドレスに記憶されている各グループのデータをそれぞれ読み取り、前記第二のデータの中の相応する行のデータとし、
前記第二のデータの中の各行のデータに対して前記第二の循環シフトを行うことによって、前記変換後のマトリックスデータの中の相応する行のデータを生成するように更に配置されていることを特徴とする請求項９に記載の装置。
【請求項１１】
前記第一のシフトユニットは、前記マトリックスデータの中の各列のデータの中の各ｓ個のデータを１グループのデータに分割し、前記マトリックスデータの中の第i行のデータの中の各グループのデータを(i-1)×ｓ桁右循環シフトすることによって、前記第一の

データを生成し、i及びｓは、自然数であるように更に配置されており、
前記第二のシフトユニットは、前記第二のデータの中の第i行のデータの中の各グループのデータを（i-1）×ｓ桁左循環シフトするように更に配置されていることを特徴とする請求項１０に記載の装置。
【請求項１２】
前記変換後のマトリックスデータで表されるマトリックスは、前記マトリックスデータで表されるマトリックスの転置マトリックスであることを特徴とする請求項１～１１の何れか一項に記載の装置。
【請求項１３】
請求項１～１２の何れか一項に記載の装置と、
前記装置に電気的に結合され、前記マトリックスデータを前記装置に送るように配置されている入出力インターフェイスと、
前記装置に電気的に結合され、深層学習モデルに基づいて前記変換後のマトリックスデータに対して処理を行う深層学習処理モジュールとを備えることを特徴とするデータ処理システム。
【請求項１４】
前記深層学習処理モジュールは、前記処理の結果を他のマトリックスデータとして前記装置に送るように更に配置されており、
前記装置は、前記他のマトリックスデータに基づいて変換後の他のマトリックスデータを生成し、前記変換後の他のマトリックスデータを前記入出力インターフェイスに送るように更に配置されていることを特徴とする請求項１３に記載のデータ処理システム。
【請求項１５】
前記深層学習処理モジュールは、前記入出力インターフェイスに電気的に結合され、前記処理の結果を前記入出力インターフェイスに送るように更に配置されていることを特徴とする請求項１３に記載のデータ処理システム。
【請求項１６】
マトリックスを変換するための方法であって、
マトリックスデータを受信し、前記マトリックスデータに対して第一の循環シフトを行うことによって、第一のデータを生成することと、
前記第一のデータの中の各行のデータを、当該行のデータの中の各データと異なる配列の配列順でキャッシュユニットに書き込むことによって、前記キャッシュユニットにおいて前記第一のデータを第二のデータとして記憶することと、
前記キャッシュユニットから前記第二のデータを読み取り、前記第二のデータに対して第二の循環シフトを行うことによって、変換後のマトリックスデータを生成することと、
を含むことを特徴とする方法。
【請求項１７】
前記キャッシュユニットは、それぞれが複数の記憶アドレスを有する複数の記憶ユニットグループを含み、
前記第一のデータを前記第二のデータとして記憶することは、前記第一のデータの中の各行のデータを、異なる記憶ユニットグループの異なる記憶アドレスにそれぞれ書き込むことを含むことを特徴とする請求項１６に記載の方法。
【請求項１８】
前記変換後のマトリックスデータを生成することは、
前記第二のデータの中の、異なる記憶ユニットグループの同じ記憶アドレスに記憶されているデータをそれぞれ読み取ることによって、前記第二のデータの中の相応する行のデータとすることと、
前記第二のデータの中の各行のデータに対して前記第二の循環シフトを行うことによって、前記変換後のマトリックスデータの中の相応する行のデータを生成することとを含むことを特徴とする請求項１７に記載の方法。
【請求項１９】

前記マトリックスデータの中の第i列のデータを（ｉ－１）桁右循環シフトすることによって、前記第一のデータを生成し、iは、自然数であり、
前記第二のデータの中の第i列のデータを（ｉ－１）桁左循環シフトすることによって、前記変換後のマトリックスデータを生成することを特徴とする請求項１８に記載の方法。
【請求項２０】
前記マトリックスデータで表されるマトリックスは、ｎ行及びｍ列を含み、ｎ及びｍは、それぞれ自然数であり、
前記第一のデータを前記第二のデータとして記憶することは、
前記第一のデータの中の第１行のデータの中のｍ個の列のデータの中の第jデータを、前記複数の記憶ユニットグループの中の第j記憶ユニットグループの第j記憶アドレスにそれぞれ書き込み、jは、１以上且つｍ以下である自然数であることと、
前記第一のデータの中の第i行のデータの中のｍ個の列のデータの中の第jデータを、前記複数の記憶ユニットグループの中の第j記憶ユニットグループの第一の記憶アドレス及び第二の記憶アドレスにそれぞれ書き込み、iは、2以上且つn以下である自然数であり、
jが１以上且つｉ－１以下であるとき、前記第一の記憶アドレスは、第m+j-i+1記憶アドレスであり、jがi以上且つｍ以下であるとき、前記第二の記憶アドレスは、第j-i+1記憶アドレスであることとを含むことを特徴とする請求項１９に記載の方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本願の実施形態は、主にマトリックスデータ処理の分野に関し、具体的には、マトリックスを変換するための装置、方法、及び当該装置を含むデータ処理システムに関する。
【背景技術】
【０００２】
深層学習アクセラレータにおいては、入力データの最も核心的なものとしての特徴、モデルパラメータ、中間結果等のデータは、何れもマトリックス方式で表され、例えば、一次元マトリックス、二次元マトリックス及び多次元マトリックスである。中央処理装置（ＣＰＵ）は、通常、クロックサイクルごとに処理する単位が１つの数であるが、深層学習アクセラレータの処理能力は、伝統的なＣＰＵの数十倍乃至数百倍である。故に、深層学習アクセラレータは、クロックサイクルごとに処理する単位が１つの数ではなく、１つの多次元マトリックスであり、例えば、入力マトリックスの１つのマトリックスブロック又はサブマトリックスである。
【０００３】
通常、深層学習モデルは、１つの複雑な計算プロセスであり、より良い性能を得るために、マトリックスのようなデータに対して幾つかのフォーマットの変化及び整理を行う必要がある。マトリックスの変換においては、マトリックス転置は、最も頻繁に使われている操作であり、このような操作では、マトリックス内部のデータを変えないが、マトリックスの次元及びデータの順番を調整することがある。
【０００４】
深層学習アクセラレータのハードウェアの設計においては、性能が高く、コストが低いマトリックスの変換案が期待されている。
【発明の概要】
【０００５】
本願の実施形態は、マトリックスを変換するための装置、方法、及び当該装置を含むデータ処理システムを提供する。これらの装置、方法、及びデータ処理システムは、深層学習アクセラレータにおいて、多次元のマトリックスの変換案を提供することができ、ストリーム処理、簡単な回路、低い消費電力等の特徴を有する。
【０００６】
本願の第一の態様は、マトリックスを変換するための装置を提供する。当該装置は、マトリックスデータを受信し、前記マトリックスデータに対して第一の循環シフトを行うこ

とによって、第一のデータを生成するように配置されている第一のシフトユニットと、前記第一のデータの中の各行のデータを、前記行のデータの中の各データと異なる配列の配列順で書き込むことによって、前記第一のデータを第二のデータとして記憶するように配置されているキャッシュユニットと、前記キャッシュユニットから前記第二のデータを読み取り、前記第二のデータに対して第二の循環シフトを行うことによって、変換後のマトリックスデータを生成するように配置されている第二のシフトユニットとを含む。

【０００７】

本願の第二の態様は、データ処理システムを提供する。当該データ処理システムは、上述した装置と、前記装置に電気的に結合され、前記マトリックスデータを前記装置に送るように配置されている入出力インターフェイスと、前記装置に電気的に結合され、深層学習モデルに基づいて前記変換後のマトリックスデータに対して処理を行うように配置されている深層学習処理モジュールとを備える。

【０００８】

本願の第三の態様は、マトリックスを変換するために用いられる方法を提供する。当該方法は、マトリックスデータを受信し、前記マトリックスデータに対して第一の循環シフトを行うことによって、第一のデータを生成することと、前記第一のデータの中の各行のデータを、当該行のデータの中の各データと異なる配列の配列順でキャッシュユニットに書き込むことによって、前記キャッシュユニットにおいて前記第一のデータを第二のデータとして記憶することと、前記キャッシュユニットから前記第二のデータを読み取り、前記第二のデータに対して第二の循環シフトを行うことによって、変換後のマトリックスデータを生成することとを含む。

【０００９】

発明の内容の部分に記載の内容は、本願の実施形態の肝心又は重要な特徴を制限せず、本願の範囲も制限しない。本願の他の特徴は、以下の説明により理解しやすくなる。

【図面の簡単な説明】

【００１０】

図面とともに以下の詳しい説明を参照することによって、本願の各実施形態の上述した及び他の特徴、メリットと態様は、より明らかになる。図面においては、同じ又は類似する図面符号は、同じ又は類似する部材を示す。

【図１】本願の実施形態に係るデータ処理システムを示すブロック図である。

【図２Ａ】マトリックスの変換を示す模式図である。

【図２Ｂ】マトリックスの変換を示す模式図である。

【図２Ｃ】マトリックスの変換を示す模式図である。

【図３】本願の実施形態に係るマトリックス変換装置を示すブロック図である。

【図４】本願の実施形態に係るキャッシュユニットを示す模式図である。

【図５】本願の実施形態に係るマトリックスの変換のプロセスを示す模式図である。

【図６】本願の実施形態に係るマトリックスの変換のプロセスを示す模式図である。

【図７】本願の実施形態に係るマトリックスの変換のプロセスを示す模式図である。

【図８】本願の実施形態に係るマトリックスの変換のプロセスを示す模式図である。

【図９】本願の他の一実施形態に係るマトリックスの変換のプロセスを示す模式図である。

【図１０】本願の他の一実施形態に係るマトリックスの変換のプロセスを示す模式図である。

【図１１】本願の実施形態に係るマトリックスを変換するために用いられる方法を示すフローチャート図である。

【発明を実施するための形態】

【００１１】

以下、図面を参照しながら、本願の実施形態をより詳しく説明する。図面は、本願の幾つかの実施形態を示しているが、本願は、様々な態様で実現することができ、ここで説明した実施形態に限定されないことが理解される。逆に、本願をより詳しく及び完全に理解

するために、これらの実施形態を提供する。本願の図面及び実施形態は、例示するためのものであり、本願の保護範囲を制限するためのものではない。

【００１２】

本願の実施形態の記載においては、用語「含む」及びその類似する用語は、開放的な含みであり、即ち、「含むが、限定されない」と理解すべきである。用語「基づく」は、「少なくとも部分的に基づく」と理解すべきである。用語「一つの実施形態」又は「当該実施形態」は、「少なくとも１つの実施形態」と理解すべきである。用語「第一」、「第二」等は、異なる又は同じ対象を示すことができる。後述は、他の明確な及び隠れた定義を含む可能性がある。

【００１３】

上述したように、深層学習アクセラレータの入力は、通常、マトリックスデータであり、深層学習モデルの計算プロセスにおいては、マトリックスデータに対して変換を行う必要がある。一つの伝統的な方案においては、汎用プロセッサ（ＣＰＵ又はＡＲＭ）により表示されるソフトウェア前処理又はコプロセッシングによって、マトリックスの次元拡張、転置等の操作を実現する。しかしながら、このような方案のマトリックスの変換性能が比較的悪く、ユーザーのコードが増えたので、不便である。また、もう一つの伝統的な方案においては、レジスタ行列によりマトリックスの転置を実現する。これは、一つのＮ×Ｎのマトリックスを縦方向でレジスタ行列に移し入れ、そして横方向でレジスタ行列から移し出すことによって、マトリックスの転置を実現するものである。しかし、このような方案は、大量のレジスタを消耗する必要があり、マトリックスの次元拡張をサポートすることができず、三次元以上のマトリックスの転置をサポートすることができず、フレキシビリティが比較的悪い。

【００１４】

深層学習アクセラレータのハードウェア設計は、ＦＰＧＡ回路設計又はＡＳＩＣチップ設計等の形式を含む。性能が高く、コストが低いマトリックス変換回路を如何に実現し、二次元のマトリックスの転置、三次元のマトリックスの転置、マトリックスの次元拡張及び転置等の様々なマトリックスの変換を如何に柔軟にサポートするかは、一つの難点である。

【００１５】

本願の実施形態は、マトリックスを変換するために用いられる装置を提供する。第一のシフトユニットにより、マトリックスデータを受信し、前記マトリックスデータに対して第一の循環シフトを行うことによって、第一のデータを生成する。キャッシュユニットにより、前記第一のデータの中の各行のデータを、前記行のデータの中の各データと異なる配列の配列順で前記キャッシュユニットに書き込むことによって、前記第一のデータを第二のデータとして記憶する。第二のシフトユニットにより、前記キャッシュユニットから前記第二のデータを読み取り、前記第二のデータに対して第二の循環シフトを行うことによって、変換後のマトリックスデータを生成する。このような方式では、本願の実施形態により深層学習アクセラレータにおいて多次元のマトリックスの変換方案を実現し、二次元のマトリックスの転置、三次元のマトリックスの転置、マトリックスの次元拡張及び転置等の様々なマトリックスの変換をサポートし、ストリーム処理、簡単な回路、低い消費電力等の特徴を有する。

【００１６】

以下、図面を参照しながら、様々な実施形態を用いて本願を詳しく説明する。

【００１７】

図１は、本願の一つの実施形態によるデータ処理システムを示すブロック図である。図１に示すデータ処理システム１００は、深層学習アクセラレータで実現することができる。図１に示すように、データ処理システム１００は、マトリックス変換装置１０２、入出力インターフェイス１０４及び深層学習処理モジュール１０６を含んでも良い。

【００１８】

入出力インターフェイス１０４は、データ処理システム１００の外部からマトリックス

データを受信し、データ処理システム１００の外部へデータ処理システム１００により処理された結果を送信するように配置されている。幾つかの実施形態においては、入出力インターフェイス１０４は、外部のメモリからマトリックスデータを読み取り、処理結果を示すデータを外部のメモリに書き込むように配置されている。幾つかの実施形態においては、入出力インターフェイス１０４は、マトリックスブロックデータ又はサブマトリックスデータを受信するように配置されている。また、入出力インターフェイス１０４は、受信されたマトリックスデータをマトリックス変換装置１０２へ送信するように更に配置されている。

【００１９】

マトリックス変換装置１０２は、入出力インターフェイス１０４に電気的に結合される。マトリックス変換装置１０２は、マトリックスデータを受信し、マトリックスデータに対して変換を行うことによって、変換後のマトリックスデータを生成するように配置されている。幾つかの実施形態においては、マトリックス変換装置１０２は、マトリックスデータに対してフォーマットの変換を行うように配置されている。幾つかの実施形態においては、マトリックス変換装置１０２は、二次元のマトリックスの転置、三次元のマトリックスの転置、マトリックスの次元拡張及び転置等のような様々なマトリックスの変換を実行するように配置されている。幾つかの実施形態においては、マトリックス変換装置１０２は、計算処理に用いられるように、変換後のマトリックスデータを深層学習処理モジュール１０６に送信するように更に配置されている。マトリックス変換装置１０２の詳しい配置については後に説明する。

【００２０】

深層学習処理モジュール１０６は、マトリックス変換装置１０２に電気的に結合される。深層学習処理モジュール１０６は、深層学習モデルに基づいて変換後のマトリックスデータに対して処理を行うように配置されている。幾つかの実施形態においては、深層学習処理モジュール１０６は、様々なマトリックス計算、ベクトル計算、非線形計算等の処理を実行するように配置されている。幾つかの実施形態においては、深層学習処理モジュール１０６は、当該分野の既知の深層学習モデルに基づく知的財産権（ＩＰ）の核とするように配置されている。

【００２１】

幾つかの実施形態においては、深層学習処理モジュール１０６は、処理結果を他のマトリックスデータとしてマトリックス変換装置１０２に送信するように更に配置されている。幾つかの実施形態においては、マトリックス変換装置１０２は、他のマトリックスデータに基づいて変換後の他のマトリックスデータを生成し、変換後の他のマトリックスデータを入出力インターフェイス１０４に送信することによって、変換後の他のマトリックスデータを外部の装置に出力するように更に配置されている。

【００２２】

幾つかの実施形態においては、深層学習処理モジュール１０６は、入出力インターフェイス１０４に直接に電気的に結合され、処理結果を入出力インターフェイス１０４に直接に送信することによって、処理結果を外部の装置に出力するように更に配置されている。幾つかの実施形態においては、深層学習処理モジュール１０６は、入出力インターフェイス１０４からマトリックスデータを直接に受信し、マトリックスデータに対して処理を行うように更に配置されている。

【００２３】

幾つかの実施形態においては、データ処理システム１００は、流れの方式で操作を行い、マトリックス変換装置１０２により目下のマトリックスデータに対して変換を行うプロセスにおいて、入出力インターフェイス１０４が次の１つのマトリックスデータを受信することができ、深層学習処理モジュール１０６が前の１つの変換後のマトリックスデータに対して処理を行うことができる。

【００２４】

図２Ａ～図２Ｃは、マトリックス変換装置により実行されるマトリックス変換を示す模

式図である。通常、二次元のマトリックスは、[n,m]で表され、nは、一番目の次元のサイズを表し、mは、二番目の次元のサイズを表す。また、三次元のマトリックスは、[n,m,k]で表され、順に類推する。

【００２５】

図２Ａは、（a）マトリックス[4,2]及び（b）マトリックス[2,4]を示しており、なお、左側のマトリックス[4,2]は、右側のマトリックス[2,4]に転置される。マトリックスデータの中の（0,0）データは、第一の行の第一の列のデータを表し、（2,1）データは、第三の行の第二の列のデータを表し、このように類推する。図２Ａは、マトリックス[n,m]の二次元のマトリックスの転置を示している。

【００２６】

図２Ｂは、（a）マトリックス[4,3,2]及び（b）マトリックス[3,4,2]を示しており、なお、左側のマトリックス[4,3,2]は、右側のマトリックス[3,4,2]に転置される。図２Ｂは、マトリックス[n,m,k]の三次元のマトリックスの転置を示している。

【００２７】

図２Ｃは、（a）[4,4]マトリックス及び（b）[2,4,2]マトリックスを示しており、なお、左側の[4,4]マトリックスは、まず、次元拡張されて[4,2,2]になり、その後、次元拡張されたマトリックスは、右側の[2,4,2]マトリックスに転置される。図２Ｃは、二次元のマトリックスが三次元のマトリックスに拡張され、三次元のマトリックスが更に転置される変換を示している。

【００２８】

図３は、本願の実施形態によるマトリックス変換装置を示すブロック図である。図３に示すように、マトリックス変換装置１０２は、第一のシフトユニット２０２、キャッシュユニット２０４及び第二のシフトユニット２０６を含む。

【００２９】

第一のシフトユニット２０２は、マトリックスデータを受信し、マトリックスデータに対して第一の循環シフトを行うことによって、第一のデータを生成するように配置されている。幾つかの実施形態においては、第一のシフトユニット２０２は、マトリックスデータの中の各行のデータに対して第一の循環シフトをそれぞれ行うように配置されている。幾つかの実施形態においては、第一のシフトユニット２０２は、マトリックスデータの中の第i行のデータを（ｉ－１）桁右循環シフトすることによって、第一のデータを生成するように配置されている。

【００３０】

キャッシュユニット２０４は、第一のシフトユニット２０２に電気的に結合され、第一のデータをキャッシュユニット２０４に書き込むことによって、第一のデータを第二のデータとして記憶するように配置されている。キャッシュユニット２０４は、第一のデータの中の各行のデータを、前記行のデータの中の各データと異なる配列の配列順でキャッシュユニット２０４に書き込む。幾つかの実施形態においては、キャッシュユニット２０４は、記憶コントローラ及び記憶ユニットグループを含み、記憶コントローラは、第一のデータを記憶ユニットグループに書き込むように制御する。幾つかの実施形態においては、キャッシュユニット２０４は、それぞれが複数の記憶アドレスを有する複数の記憶ユニットグループを含み、キャッシュユニット２０４は、第一のデータの中の各行のデータを、異なる記憶ユニットグループの異なる記憶アドレスにそれぞれ書き込むことによって、第一のデータを第二のデータとして記憶するように更に配置されている。第一のデータは、自身の元々の行及び列の配列順でキャッシュユニット２０４に書き込むのではないので、キャッシュユニット２０４に記憶されている第二のデータは、第一のデータの配列順又はフォーマットの変更後のデータであることを理解すべきである。しかしながら、第一のデータに比べ、第二のデータの中の各データの内容は、変化がない。後に、第一のデータをキャッシュユニットに書き込むことを詳しく説明する。

【００３１】

第二のシフトユニット２０６は、キャッシュユニット２０４に電気的に結合され、キャ

ッシュユニット２０４から第二のデータを読み取り、第二のデータに対して第二の循環シフトを行うことによって、変換後のマトリックスデータを生成するように配置されている。幾つかの実施形態においては、第二のシフトユニット２０６は、第二のデータの中の、異なる記憶ユニットグループの同じ記憶アドレスに記憶されているデータをそれぞれ読み取ることによって、第二のデータの中の相応する行のデータとするように配置されている。幾つかの実施形態においては、第二のシフトユニット２０６は、第二のデータの中の各行のデータに対して第二の循環シフトを行うことによって、変換後のマトリックスデータの中の相応する行のデータを生成するように配置されている。幾つかの実施形態においては、第二のシフトユニット２０６は、第二のデータの中の第i行のデータを（ｉ－１）桁左循環シフトすることによって、変換後のマトリックスデータを生成するように配置されている。また、第二のシフトユニット２０６は、変換後のマトリックスデータを図１の深層学習処理モジュール１０６に送信するように更に配置されている。

【００３２】

図４は、本願の実施形態によるキャッシュユニットを示す模式図である。図４に示すように、キャッシュユニット２０４は、複数の記憶ユニットグループ３０２を含む。

【００３３】

図４に示すように、幾つかの実施形態においては、複数の記憶ユニットグループ３０２は、ｘ軸の方向に沿って順に配列し、ｘ軸は、第i記憶ユニットグループ３０２を表す。また、それぞれの記憶ユニットグループ３０２は、複数の記憶アドレスを含む。説明の便宜上、幾つかの実施形態においては、それぞれの記憶ユニットグループ３０２の複数の記憶アドレスは、ｘ軸に直交するｙ軸の方向に沿って配列するように示されており、ｙ軸は、それぞれの記憶ユニットグループ３０２の第j記憶アドレスを示す。幾つかの実施形態においては、記憶アドレスは、深さと呼ばれても良い。幾つかの実施形態においては、記憶アドレス[i,j]は、第i記憶ユニットグループ３０２の第j記憶アドレスを示す。ここでは、i及びjは、それぞれ自然数である。

【００３４】

幾つかの実施形態においては、記憶ユニットグループ３０２は、スタティックランダムアクセスメモリ（ＳＲＡＭ）グループである。後述においては、記憶ユニットグループ３０２をＳＲＡＭグループとする例示を説明する。記憶ユニットグループ３０２は、ＳＲＡＭグループと限らず、他のタイプの記憶ユニットの集合を採用することもできる。

【００３５】

図５は、本願の実施形態によるマトリックス変換のプロセスを示す模式図である。図５に示される左側のデータは、マトリックスデータの中の第一の行のデータ（0,0）、（0,1）、（0,2）、（0,3）、（0,4）、（0,5）、（0,6）、（0,7）である。幾つかの実施形態においては、第一のシフトユニット２０２は、マトリックスデータの中から当該第一の行のデータを読み取り、当該第一の行のデータに対して（1-1=0）桁右循環シフトし、即ち、実質的に当該第一の行のデータに対して右循環シフトしない。キャッシュユニット２０４は、第一の対角線の記憶アドレスで当該第一の行のデータをキャッシュユニット２０４に書き込むように、当該第一の行のデータの書き込みを制御する。幾つかの実施形態においては、図５に示すように、キャッシュユニット２０４は、当該第一の行のデータを複数の記憶ユニットグループの記憶アドレス[1,1]、[2,2]、[3,3]、[4,4]、[5,5]、[6,6]、[7,7]、[8,8]にそれぞれ書き込む。このように、マトリックスデータの中の第一の行のデータは、自身の元々のフォーマット又は配列順ではなく、第一の対角線の記憶アドレスでキャッシュユニット２０４に記憶されている。

【００３６】

図６は、本願の実施形態によるマトリックスの変換のプロセスを示す模式図である。図６に示される左側のデータは、マトリックスデータの中の第二の行のデータである。幾つかの実施形態においては、第一のシフトユニット２０２は、マトリックスデータの中から当該第二の行のデータを読み取り、当該第二の行のデータに対して（2-1=1）桁右循環シフトすることによって、１桁右循環シフトした後の第二の行のデータ（1,7）、（1,0）、

（1,1）、（1,2）、（1,3）、（1,4）、（1,5）、（1,6）を獲得する。キャッシュユニット２０４は、右循環シフトした後の第二の行のデータの書き込みを制御することによって、第二の対角線の記憶アドレスで、右循環シフトした後の第二の行のデータをキャッシュユニット２０４に書き込む。幾つかの実施形態においては、図６に示すように、キャッシュユニット２０４は、右循環シフトした後の第二の行のデータを、複数の記憶ユニットグループの記憶アドレス[1,8]、[2,1]、[3,2]、[4,3]、[5,4]、[6,5]、[7,6]、[8,7]にそれぞれ書き込む。このように、マトリックスデータの中の右循環シフトした後の第二の行のデータは、自身の元々のフォーマット又は配列順ではなく、第二の対角線の記憶アドレスでキャッシュユニット２０４に記憶されている。

【００３７】

図７は、本願の実施形態によるマトリックスの変換のプロセスを示す模式図である。図７に示される左側のデータは、マトリックスデータの中の第三の行のデータである。幾つかの実施形態においては、第一のシフトユニット２０２は、マトリックスデータの中から当該第三の行のデータを読み取り、当該第三の行のデータに対して（3-1=2）桁右循環シフトすることによって、2桁右循環シフトした後の第三の行のデータ（2,6）、（2,7）、（2,0）、（2,1）、（2,2）、（2,3）、（2,4）、（2,5）を獲得する。キャッシュユニット２０４は、右循環シフトした後の第三の行のデータの書き込みを制御することによって、第三の対角線の記憶アドレスで、右循環シフトした後の第三の行のデータをキャッシュユニット２０４に書き込む。幾つかの実施形態においては、図７に示すように、キャッシュユニット２０４は、右循環シフトした後の第三の行のデータを、複数の記憶ユニットグループの記憶アドレス[1,7]、[2,8]、[3,1]、[4,2]、[5,3]、[6,4]、[7,5]、[8,6]にそれぞれ書き込む。このように、マトリックスデータの中の右循環シフトした後の第三の行のデータは、自身の元々のフォーマット又は配列順ではなく、第三の対角線の記憶アドレスでキャッシュユニット２０４に記憶されている。

【００３８】

幾つかの実施形態においては、これによって類推し、第一のシフトユニット２０２は、マトリックスデータの中から第i行のデータを読み取り、第i行のデータに対して（i-1）桁右循環シフトする。キャッシュユニット２０４は、第iの対角線の記憶アドレスで右循環シフトした後の第i行のデータをキャッシュユニット２０４に書き込むように、右循環シフトした後の第i行のデータの書き込みを制御する。ここでは、仮に、マトリックスデータで表されるマトリックスは、ｎ行及びｍ列を含むとする。なお、ｎ及びｍは、それぞれ自然数である。幾つかの実施形態においては、キャッシュユニット２０４は、右循環シフトした後の第i行のデータの中のｍ個の列のデータの中の第jデータを、第j記憶ユニットグループの第（m+j-i+1）記憶アドレス（jは、１以上且つi-1以下である）及び第（j-i+1）記憶アドレス（jは、i以上且つｍ以下である）にそれぞれ書き込み、なお、iは、２以上且つｎ以下である。また、iが１に等しい場合、上述したように、第一の行のデータの中のｍ個の列のデータの中の第jデータを第j記憶ユニットグループの第j記憶アドレスにそれぞれ書き込む。このように、キャッシュユニット２０４は、右循環シフトした後の各行のデータの中の各データを、相応する対角線の記憶アドレスで異なる記憶ユニットグループの異なる記憶アドレスにそれぞれ書き込むように制御する。これにより、第一の循環シフトの後のマトリックスデータを、自身の元々のフォーマット又は配列順と異なるフォーマット又は配列順でキャッシュユニットに記憶するので、読み取ることで変換後のマトリックスを生成することができる。

【００３９】

図８は、本願の実施形態によるマトリックスの変換のプロセスを示す模式図である。図８の左側の部分は、キャッシュユニット２０４に記憶されている第二のデータの一部を示している。幾つかの実施形態においては、第二のシフトユニット２０６は、キャッシュユニット２０４から異なる記憶ユニットグループの第一の記憶アドレスに記憶されているデータ（0,0）、（1,0）、（2,0）、（3,0）、（4,0）、（5,0）、（6,0）、（7,0）を順に読み取って第二のデータの中の第一の行のデータとする。そして、第二のシフトユニット

２０６は、第二のデータの中の第一の行のデータに対して（1-1=0）桁左循環シフトし、即ち、第二のデータの中の第一の行のデータに対して実質的に左循環シフトしない。第二のシフトユニット２０６は、第二のデータの中の第一の行のデータを変換後のマトリックスデータの中の第一の行のデータとして出力する。

【００４０】

幾つかの実施形態においては、第二のシフトユニット２０６は、キャッシュユニット２０４から異なる記憶ユニットグループの第二の記憶アドレスに記憶されているデータ（7,1）、（0,1）、（1,1）、（2,1）、（3,1）、（4,1）、（5,1）、（6,1）を順に読み取り、第二のデータの中の第二の行のデータとする。第二のシフトユニット２０６は、第二のデータの中の第二の行のデータに対して（2-1=1）桁左循環シフトし、左循環シフトした後の第二の行のデータ（0,1）、（1,1）、（2,1）、（3,1）、（4,1）、（5,1）、（6,1）、（7,1）を生成する。第二のシフトユニット２０６は、第二のデータの中の、左循環シフトした後の第二の行のデータを変換後のマトリックスデータの中の第二の行のデータとして出力する。

【００４１】

幾つかの実施形態においては、第二のシフトユニット２０６は、キャッシュユニット２０４から異なる記憶ユニットグループの第三の記憶アドレスに記憶されているデータ（6,2）、（7,2）、（0,2）、（1,2）、（2,2）、（3,2）、（4,2）、（5,2）を順に読み取り、第二のデータの中の第三の行のデータとする。第二のシフトユニット２０６は、第二のデータの中の第三の行のデータに対して（3-1=2）桁左循環シフトし、左循環シフトした後の第三の行のデータ（0,2）、（1,2）、（2,2）、（3,2）、（4,2）、（5,2）、（6,2）、（7,2）を生成する。第二のシフトユニット２０６は、第二のデータの中の、左循環シフトした後の第三の行のデータを変換後のマトリックスデータの中の第三の行のデータとして出力する。

【００４２】

幾つかの実施形態においては、これによって類推し、第二のシフトユニット２０６は、キャッシュユニット２０４から第二のデータの中の、異なるユニットグループの同じ記憶アドレスに記憶されているデータを順に読み取り、第二のデータの中の相応する行のデータとする。そして、第二のシフトユニット２０６は、第二のデータの中の第i行のデータに対して（i-1）桁左循環シフトすることによって、変換後のマトリックスデータの中の相応する行のデータを生成する。このように、マトリックス変換装置１０２は、転置後のマトリックスデータを出力する。

【００４３】

幾つかの実施形態においては、マトリックス[n, m]のｎがｍ以下である場合、上述した方式でマトリックスデータに対して第一の循環シフトを行うことによって第一のデータを生成し、第一のデータが書き込まれることを制御することによって第二のデータとして記憶し、第二のデータの中の各行のデータを読み取り、各行のデータに対して第二の循環シフトを行うことによって、変換後のマトリックスを生成する。例えば、ｎが３に等しく且つｍが８に等しい場合、マトリックスデータの中の各行のデータは、図７に示すように、右循環シフトされ、書き込まれる。幾つかの実施形態においては、第二のシフトユニット２０６は、異なる記憶ユニットグループの、第一の記憶アドレスに記憶されているデータ（0,0）、（1,0）、（2,0）、（dummy）、（dummy）、（dummy）、（dummy）、（dummy）、第二の記憶アドレスに記憶されているデータ（dummy）、（0,1）、（1,1）、（2,1）、（dummy）、（dummy）、（dummy）、（dummy）、・・・・・・第八の記憶アドレスに記憶されているデータ（1,7）、（2,7）、（dummy）、（dummy）、（dummy）、（dummy）、（dummy）、（0,7）をそれぞれ読み取り、第二のデータの中の相応する行のデータとし、なお、dummyは、ダミーデータであり、マトリックス変換装置１０２から出力される時に省略される。第二のシフトユニット２０６は、第二のデータの中の第i行のデータをそれぞれ（i-1）桁左循環シフトすることによって、第一の行のデータ（0,0）、（1,0）、（2,0）、第二の行のデータ（0,1）、（1,1）、（2,1）……及び第八の行のデータ（0,7）、（

1,7）、（2,7）を含む変換後のマトリックスデータを出力する。このように、マトリックス変換装置１０２は、マトリックス[3,8]の転置マトリックス[8,3]を出力する。

【００４４】

幾つかの実施形態においては、ｎがｍより大きい場合、第一のシフトユニット２０２は、マトリックス[n,m]をマトリックス[m,m]とマトリックス[n-m, m]に分割し、上述した方式でマトリックス[m,m]に対して変換を行い、変換後の第一のマトリックスを出力する。マトリックス[n-m, m]に対し、第一の循環シフトの後のマトリックスデータを、複数の記憶ユニットグループの第（m+1）記憶アドレスからキャッシュユニット２０４に書き込み、上述した方式で、マトリックス[n-m, m]に対して変換を行い、なお、第二のシフトユニットは、第（m+1）記憶アドレスからデータを読み取り、読み取られたデータに対して第二の循環シフトを行った後に変換後の第二のマトリックスを出力する。そして、第一のマトリックス及び第二のマトリックスを組み合わせて転置後のマトリックスを生成する。

【００４５】

幾つかの実施形態においては、ｍがキャッシュユニット２０４の総データの幅ｋより大きい場合、第一のシフトユニット２０２は、マトリックス[n, m]をマトリックス[n,k]及びマトリックス[n,m-k]に分割し、上述した方式で、マトリックス[n,k]に対して変換を行い、変換後の第一のマトリックスを出力する。マトリックス[n,m-k]に対し、第一の循環シフトの後のマトリックスデータを、記憶ユニットグループの第（k+1）記憶アドレスからキャッシュユニット２０４に書き込み、上述した方式で、マトリックス[n,m-k]に対して変換を行い、なお、第二のシフトユニットは、第（k+1）記憶アドレスから読み取り、読み取られたデータに対して第二の循環シフトを行った後に変換後の第二のマトリックスを出力する。そして、第一のマトリックス及び第二のマトリックスを組み合わせて転置後のマトリックスを生成する。

【００４６】

幾つかの実施形態においては、上述した、対角線の方式で書き込んで、行方向でデータを読み取る方式と逆に、まず、第一のデータの中の各行のデータを異なる記憶ユニットグループの同じアドレスに記憶し、その後、対応する対角線の方式でキャッシュユニットに記憶されているデータを読み取り、変換後のマトリックスを生成しても良い。

【００４７】

幾つかの実施形態においては、第一のデータの中の各行のデータの、キャッシュユニットにおける書き込みアドレスを好ましく設定し、及び/又は読み取り方式を変えることによって、次元の変換、データ位置の変換、データの入れ替え等を含む望ましいマトリックスの変換を行うことができる。

【００４８】

図９及び図１０は、本願のもう一つの実施形態によるマトリックス変換のプロセスを示す模式図である。図９及び図１０は、マトリックス[n, m]が[n, p,q]に拡張され、その後、マトリックス[n, p,q]に対して変換を行うプロセスの例示を示している。

【００４９】

幾つかの実施形態においては、複数の記憶ユニットグループ３０２の中の各ｓ個の記憶ユニットグループは、１グループの記憶ユニットグループに分けられ、各グループの記憶ユニットグループは、複数の記憶アドレスを含む。幾つかの実施形態においては、複数の記憶アドレスの中の各ｔ個の記憶アドレスが１グループの記憶アドレスに分けられる。幾つかの実施形態においては、キャッシュユニット２０４は、第一のデータの中の各ｔ行のデータの中の複数のグループのデータを異なるグループの記憶ユニットグループの異なるグループの記憶アドレスにそれぞれ書き込み、第一のデータを第二のデータとして記憶するように更に配置されており、なお、複数のグループのデータの中の各グループのデータは、ｓ×ｔ個のデータを含み、なお、ｓ及びｔは、自然数である。

【００５０】

幾つかの実施形態においては、第一のシフトユニット２０２は、マトリックスデータの中の各行のデータの中の各ｓ個のデータを１グループのデータに分け、マトリックスデー

タの中の第i行のデータの中の各グループのデータを（i-1）×s桁右循環シフトし、第一のデータを生成するように更に配置されている。

【００５１】

幾つかの実施形態においては、第二のシフトユニット２０６は、第二のデータの中の、異なるグループの記憶ユニットグループの同じグループの記憶アドレスに記憶されている各グループのデータをそれぞれ順に読み取り、第二のデータの中の相応する行のデータとするように更に配置されている。幾つかの実施形態においては、第二のシフトユニット２０６は、第二のデータの中の各行のデータに対して第二の循環シフトを行うことによって、変換後のマトリックスデータの中の相応する行のデータを生成する。幾つかの実施形態においては、第二のシフトユニット２０６は、第二のデータの中の第i行のデータの中の各グループのデータを（i-1）×ｓ桁左循環シフトするよう、即ち、（i-1）グループであるように更に配置されている。

【００５２】

図９及び図１０に示される例示においては、入力マトリックスは、例えば、二次元のマトリックス[4,8]であり、マトリックス変換装置１０２は、まず、二次元のマトリックス[4,8]を三次元のマトリックス[4,4,2]に拡張する。幾つかの実施形態においては、各行のデータの中のそれぞれの二つのデータが１グループのデータに分けられ、それぞれの二つの記憶ユニットグループが１グループの記憶ユニットグループに分けられ、異なるグループの記憶ユニットグループの同じ記憶アドレスが１グループの記憶アドレスに分けられる。

【００５３】

図９に示すように、第一のシフトユニット２０２は、第一の行のデータの中の４グループのデータを（1-1=0）×2桁右循環シフトし、即ち、実質的に右循環シフトを行わない。図９に示すように、キャッシュユニット２０４は、第一の行のデータの中の４グループのデータ{（0,0）、（0,1）}、{（0,2）、（0,3）}、{（0,4）、（0,5）}、{（0,6）、（0,7）}を、第一及び第二の記憶ユニットグループの第一の記憶アドレス、第三及び第四の記憶ユニットグループの第二の記憶アドレス、第五及び第六の記憶ユニットグループの第三の記憶アドレス、第七及び第八の記憶ユニットグループの第四の記憶アドレスにそれぞれ書き込む。

【００５４】

図１０に示すように、第一のシフトユニット２０２は、第二の行のデータの中の４グループのデータ{（1,0）、（1,1）}、{（1,2）、（1,3）}、{（1,4）、（1,5）}、{（1,6）、（1,7）}を（2-1=1）×2桁右循環シフトし、即ち、２桁又は１つのグループ右循環シフトし、右循環シフトした第二の行のデータ{（1,6）、（1,7）}、{（1,0）、（1,1）}、{（1,2）、（1,3）}、{（1,4）、（1,5）}を生成する。図１０に示すように、キャッシュユニット２０４は、当該右循環シフトした第二の行のデータを第一及び第二の記憶ユニットグループの第四の記憶アドレス、第三及び第四の記憶ユニットグループの第一の記憶アドレス、第五及び第六記憶ユニットグループの第二の記憶アドレス、第七及び第八の記憶ユニットグループの第四の記憶アドレスにそれぞれ記憶する。

【００５５】

これによって類推し、図５～図８を参照しながら説明したマトリックス変換方式と類似的にマトリックス[4,4,2]に対して第一の循環シフト、書き込み、読み取り及び第二の循環シフトを行うことによって、マトリックス[4,8]を拡張して変換し、変換後のマトリックス[4,4,2]を生成する。

【００５６】

本願の実施形態により、深層学習アクセラレータにおいては、深層学習処理モジュールに対してマトリックスを変換するために用いられる装置を実現し、当該装置は、チップの外から入力された、マトリックスブロック又はサブマトリックス等のようなマトリックスデータに対して変換を行い、変換後のマトリックスデータを深層学習処理モジュールに送信し、計算処理に用いられることができる。本願の実施形態の装置は、多次元のマトリッ

クスの変換方案を実現することができ、二次元のマトリックス転置、三次元のマトリックス転置、マトリックスの次元拡張及び転置等の様々なマトリックスの変換をサポートし、ストリーム処理、簡単な回路、低い消費電力等の特徴を有する。

【００５７】

図１１は、本願の実施形態によるマトリックスを変換するために用いられる方法を示すフローチャート図である。図１１に示すように、マトリックスを変換するために用いられる方法４００は、ブロック４０２～ブロック４０６を含む。

【００５８】

ブロック４０２においては、マトリックスデータを受信し、マトリックスデータに対して第一の循環シフトを行うことによって、第一のデータを生成する。幾つかの実施形態においては、マトリックスデータの中の第i行のデータを（i-1）桁右循環シフトすることによって、第一のデータを生成し、iは、自然数である。

【００５９】

ブロック４０４においては、第一のデータの中の各行のデータを、当該行のデータの中の各データの配列順と異なる配列順でキャッシュユニットに書き込むことによって、キャッシュユニットにおいて第一のデータを第二のデータとして記憶する。幾つかの実施形態においては、キャッシュユニットは、それぞれが複数の記憶アドレスを含む複数の記憶ユニットグループを含む。幾つかの実施形態においては、第一のデータの中の各行のデータを異なる記憶ユニットグループの異なる記憶アドレスにそれぞれ書き込む。

【００６０】

幾つかの実施形態においては、マトリックスデータで表されるマトリックスは、ｎ行及びｍ列を含み、なお、ｎ及びｍは、それぞれ自然数である。幾つかの実施形態においては、第一のデータの中の第１行のデータの中のｍ個の列のデータの中の第jデータを複数の記憶ユニットグループの中の第j記憶ユニットグループの第j記憶アドレスにそれぞれ書き込み、なお、jは、１以上且つｍ以下の自然数である。また、第一のデータの中の第i行のデータの中のｍ個の列のデータの中の第jデータを、複数の記憶ユニットグループの中の第j記憶ユニットグループの第（m+j-i+1）記憶アドレス（jは、１以上且つi-1以下である）及び第（j-i+1）記憶アドレス（jは、i以上且つｍ以下である）にそれぞれ書き込み、なお、iは、２以上且つｎ以下の自然数である。

【００６１】

ブロック４０６においては、キャッシュユニットから第二のデータを読み取り、第二のデータに対して第二の循環シフトを行うことによって、変換後のマトリックスデータを生成する。幾つかの実施形態においては、第二のデータの中の、異なる記憶ユニットグループの同じ記憶アドレスに記憶されているデータをそれぞれ読み取り、第二のデータの中の相応する行のデータとする。幾つかの実施形態においては、第二のデータの中の各行のデータに対して第二の循環シフトを行うことによって、変換後のマトリックスデータの中の相応する行のデータを生成する。幾つかの実施形態においては、第二のデータの中の第i行のデータを（ｉ－１）桁左循環シフトすることによって、変換後のマトリックスデータを生成する。

【００６２】

本願の実施形態を実施するための方法は、システムオンチップ（ＳｏＣ）アーキテクチャに適用するために、１つ又は複数のプログラミング言語の任意の組み合わせを用いてプログラミングすることができると理解すべきである。また、特定の順序でそれぞれの操作を説明したが、理想な結果が得られるために、このような操作は、示された特定の順序又は順に実行することが求められ、又は、全ての図示の操作は、実行する必要があると理解すべきである。一定の環境においては、マルチタスク及びパラレル処理が有利である可能性がある。

【００６３】

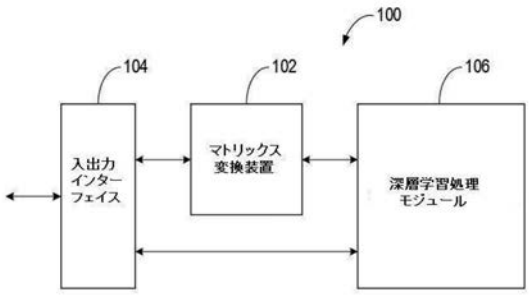
上述した記載は、若干の具体的な実現の細部を含むが、これらは、本願の範囲に対する制限であると解釈されるべきではない。１つの実施形態の上下の文脈に記載の幾つかの特

徴は、１つの実現において組合わせて実現されることもできる。逆に、１つの実現の上下の文脈に記載の様々な特徴は、複数の実現において単独又は任意の適切なサブ組み合わせの方式で実現されることもできる。

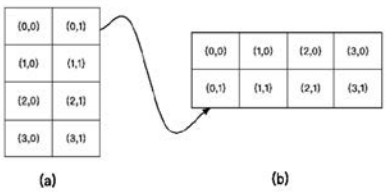
【００６４】

構造の特徴及び/又は方法のロジック動作に特定した言語を用いて本願の主題を記載したが、特許請求の範囲に限定される主題は、上述した記載の特定の特徴又は動作に限らないと理解すべきである。逆に、上述した記載の特定の特徴及び動作は、特許請求の範囲を実現する例示の形式に過ぎない。

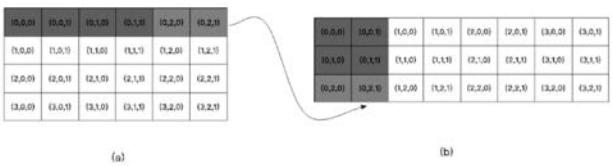
【図１】



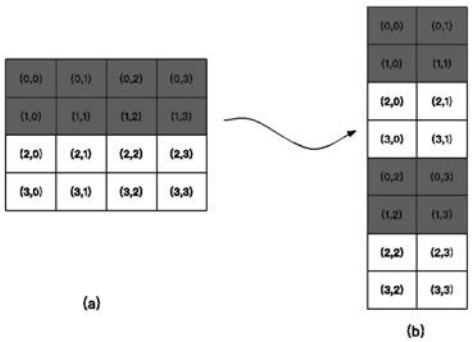
【図２Ａ】



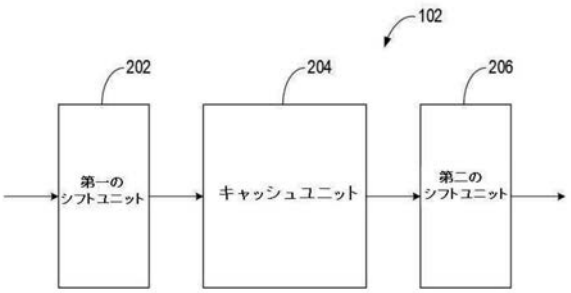
【図２Ｂ】



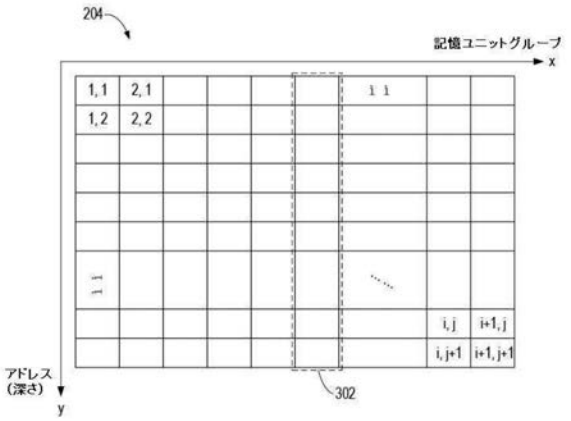
【図２Ｃ】



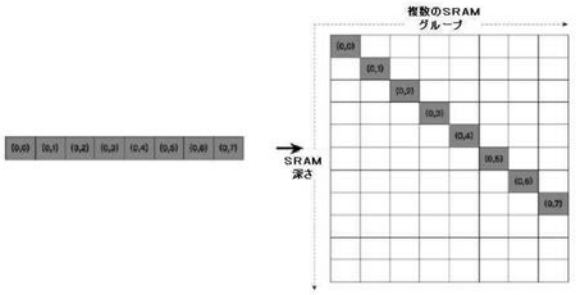
【図３】



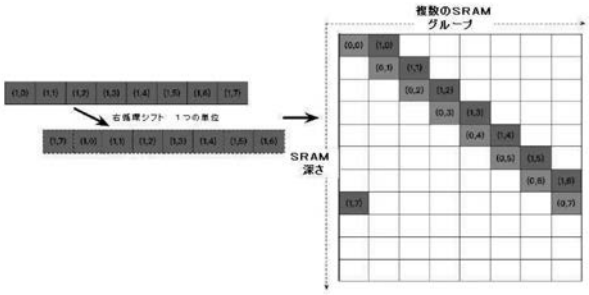
【図４】



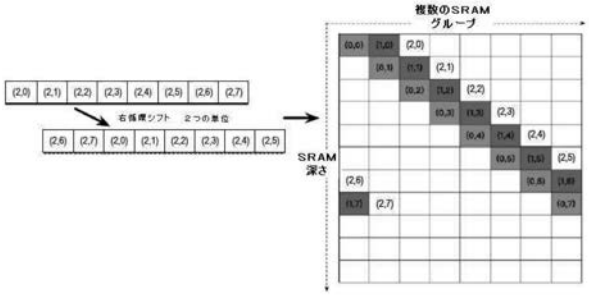
【図５】



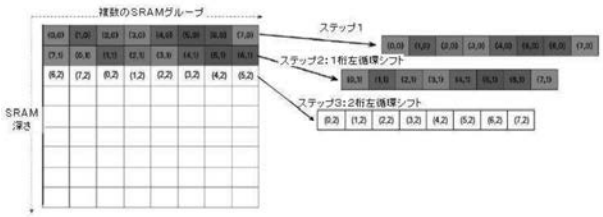
【図６】



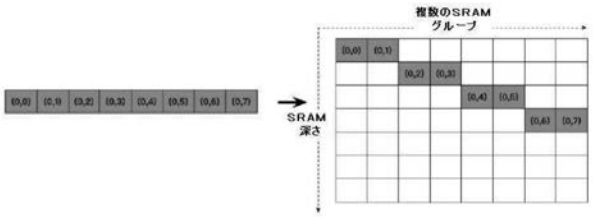
【図７】



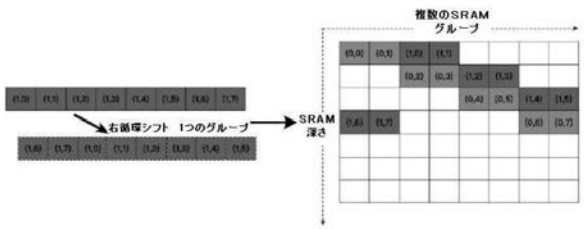
【図８】



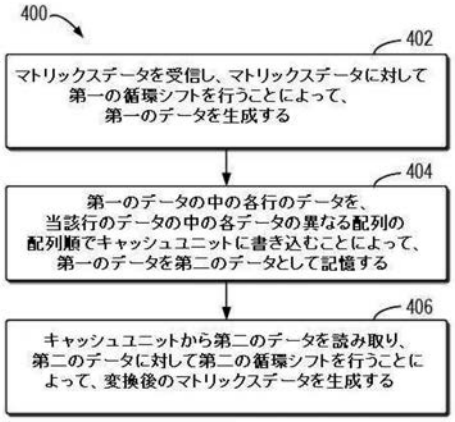
【図９】



【図１０】



【図１１】



【手続補正書】
【提出日】令和2年8月27日(2020.8.27)
【手続補正１】
【補正対象書類名】特許請求の範囲
【補正対象項目名】全文
【補正方法】変更
【補正の内容】
【特許請求の範囲】
【請求項１】
マトリックスを変換するための装置であって、
マトリックスデータを受信し、前記マトリックスデータに対して第一の循環シフトを行うことによって、第一のデータを生成するように配置されている第一のシフトユニットと、
前記第一のデータの中の各行のデータを、前記行のデータの中の各データと異なる配列の配列順で書き込むことによって、前記第一のデータを第二のデータとして記憶するように配置されているキャッシュユニットと、
前記キャッシュユニットから前記第二のデータを読み取り、前記第二のデータに対して第二の循環シフトを行うことによって、変換後のマトリックスデータを生成するように配置されている第二のシフトユニットと、を含むことを特徴とする装置。
【請求項２】
前記キャッシュユニットは、それぞれが複数の記憶アドレスを有する複数の記憶ユニットグループを含み、前記第一のデータの中の各行のデータを、異なる記憶ユニットグループの異なる記憶アドレスにそれぞれ書き込むことによって、前記第一のデータを前記第二のデータとして記憶するように更に配置されていることを特徴とする請求項１に記載の装置。
【請求項３】
前記第二のシフトユニットは、
前記第二のデータの中の、異なる記憶ユニットグループの同じ記憶アドレスに記憶されているデータをそれぞれ読み取ることによって、前記第二のデータの中の相応する行のデータとし、
前記第二のデータの中の各行のデータに対して前記第二の循環シフトを行うことによって、前記変換後のマトリックスデータの中の相応する行のデータを生成するように更に配置されていることを特徴とする請求項２に記載の装置。
【請求項４】
前記第一のシフトユニットは、前記マトリックスデータの中の第i行のデータを（i－１）桁右循環シフトすることによって、前記第一のデータを生成し、iは、自然数であるように更に配置されており、
前記第二のシフトユニットは、前記第二のデータの中の第i行のデータを（i－１）桁左循環シフトすることによって、前記変換後のマトリックスデータを生成するように配置されていることを特徴とする請求項３に記載の装置。
【請求項５】
前記マトリックスデータで表されるマトリックスは、ｎ行及びｍ列を含み、ｎ及びｍは、それぞれ自然数であり、
前記キャッシュユニットは、
前記第一のデータの中の第１行のデータの中のｍ個の列のデータの中の第jデータを、前記複数の記憶ユニットグループの中の第j記憶ユニットグループの第j記憶アドレスにそれぞれ書き込み、jは、１以上且つｍ以下である自然数であり、
前記第一のデータの中の第i行のデータの中のｍ個の列のデータの中の第jデータを、前記複数の記憶ユニットグループの中の第j記憶ユニットグループの第一の記憶アドレス及び第二の記憶アドレスにそれぞれ書き込み、iは、2以上且つn以下である自然数であり、

jが１以上且つi－１以下であるとき、前記第一の記憶アドレスは、第m+j-i+1記憶アドレスであり、
jがi以上且つm以下であるとき、前記第二の記憶アドレスは、第j-i+1記憶アドレスであるように更に配置されていることを特徴とする請求項４に記載の装置。
【請求項６】
前記第一のシフトユニットは、
前記マトリックスデータで表されるマトリックスを、ｐ行を含む第一のマトリックス及びｑ行を含む第二のマトリックス、又は、ｐ列を含む第一のマトリックス及びｑ列を含む第二のマトリックスに分割し、ｐ及びｑは、それぞれ自然数であり、
前記第一のマトリックスのマトリックスデータに対して前記第一の循環シフトを行うことによって、前記第一のマトリックスの前記第一のデータを生成し、
前記第二のマトリックスのマトリックスデータに対して前記第一の循環シフトを行うことによって、前記第二のマトリックスの前記第一のデータを生成するように更に配置されていることを特徴とする請求項１に記載の装置。
【請求項７】
前記キャッシュユニットは、
前記第一のマトリックスの前記第一のデータを、第一の記憶アドレスをスタートアドレスとして前記キャッシュユニットに書き込むことによって、前記第一のマトリックスの前記第一のデータを前記第一のマトリックスの前記第二のデータとして記憶し、
前記第二のマトリックスの前記第一のデータを、第ｋ＋１記憶アドレスをスタートアドレスとして前記キャッシュユニットに書き込むことによって、前記第二のマトリックスの前記第一のデータを前記第二のマトリックスの前記第二のデータとして記憶し、ｋは、ｐ以上の自然数であるように更に配置されていることを特徴とする請求項６に記載の装置。
【請求項８】
前記第二のシフトユニットは、
前記キャッシュユニットから前記第一のマトリックスの前記第二のデータを読み取り、前記第一のマトリックスの前記第二のデータに対して前記第二の循環シフトを行うことによって、変換後の第一のマトリックスデータを生成し、
前記キャッシュユニットから前記第二のマトリックスの前記第二のデータを読み取り、前記第二のマトリックスの前記第二のデータに対して前記第二の循環シフトを行うことによって、変換後の第二のマトリックスデータを生成し、
前記変換後の第一のマトリックスデータ及び前記変換後の第二のマトリックスデータを結合することによって、前記変換後のマトリックスデータを生成するように更に配置されていることを特徴とする請求項７に記載の装置。
【請求項９】
前記キャッシュユニットは、複数の記憶ユニットグループを含み、前記複数の記憶ユニットグループの中の各ｓ個の記憶ユニットグループが１グループの記憶ユニットグループに分割され、各グループの記憶ユニットグループは、複数の記憶アドレスを含み、前記複数の記憶アドレスの中の各ｔ個の記憶アドレスが１グループの記憶アドレスに分割され、
前記キャッシュユニットは、前記第一のデータの中の各ｔ行のデータの中の複数のグループのデータを、異なるグループの記憶ユニットグループの異なるグループの記憶アドレスにそれぞれ書き込むことによって、前記第一のデータを前記第二のデータとして記憶し、前記複数のグループのデータの中の各グループのデータは、ｓ×ｔ個のデータを含み、ｓ及びｔは、自然数であるように配置されていることを特徴とする請求項１に記載の装置。
【請求項１０】
前記第二のシフトユニットは、
前記第二のデータの中の、異なるグループの記憶ユニットグループの同じグループの記憶アドレスに記憶されている各グループのデータをそれぞれ読み取り、前記第二のデータの中の相応する行のデータとし、

前記第二のデータの中の各行のデータに対して前記第二の循環シフトを行うことによって、前記変換後のマトリックスデータの中の相応する行のデータを生成するように更に配置されていることを特徴とする請求項９に記載の装置。
【請求項１１】
前記第一のシフトユニットは、前記マトリックスデータの中の各列のデータの中の各ｓ個のデータを１グループのデータに分割し、前記マトリックスデータの中の第i行のデータの中の各グループのデータを(i-1)×ｓ桁右循環シフトすることによって、前記第一のデータを生成し、i及びｓは、自然数であるように更に配置されており、
前記第二のシフトユニットは、前記第二のデータの中の第i行のデータの中の各グループのデータを（i-1）×ｓ桁左循環シフトするように更に配置されていることを特徴とする請求項１０に記載の装置。
【請求項１２】
前記変換後のマトリックスデータで表されるマトリックスは、前記マトリックスデータで表されるマトリックスの転置マトリックスであることを特徴とする請求項１～１１の何れか一項に記載の装置。
【請求項１３】
請求項１～１２の何れか一項に記載の装置と、
前記装置に電気的に結合され、前記マトリックスデータを前記装置に送るように配置されている入出力インターフェイスと、
前記装置に電気的に結合され、深層学習モデルに基づいて前記変換後のマトリックスデータに対して処理を行う深層学習処理モジュールとを備えることを特徴とするデータ処理システム。
【請求項１４】
前記深層学習処理モジュールは、前記処理の結果を他のマトリックスデータとして前記装置に送るように更に配置されており、
前記装置は、前記他のマトリックスデータに基づいて変換後の他のマトリックスデータを生成し、前記変換後の他のマトリックスデータを前記入出力インターフェイスに送るように更に配置されていることを特徴とする請求項１３に記載のデータ処理システム。
【請求項１５】
前記深層学習処理モジュールは、前記入出力インターフェイスに電気的に結合され、前記処理の結果を前記入出力インターフェイスに送るように更に配置されていることを特徴とする請求項１３に記載のデータ処理システム。
【請求項１６】
マトリックスを変換するための方法であって、
マトリックスデータを受信し、前記マトリックスデータに対して第一の循環シフトを行うことによって、第一のデータを生成することと、
前記第一のデータの中の各行のデータを、当該行のデータの中の各データと異なる配列の配列順でキャッシュユニットに書き込むことによって、前記キャッシュユニットにおいて前記第一のデータを第二のデータとして記憶することと、
前記キャッシュユニットから前記第二のデータを読み取り、前記第二のデータに対して第二の循環シフトを行うことによって、変換後のマトリックスデータを生成することと、
を含むことを特徴とする方法。
【請求項１７】
前記キャッシュユニットは、それぞれが複数の記憶アドレスを有する複数の記憶ユニットグループを含み、
前記第一のデータを前記第二のデータとして記憶することは、前記第一のデータの中の各行のデータを、異なる記憶ユニットグループの異なる記憶アドレスにそれぞれ書き込むことを含むことを特徴とする請求項１６に記載の方法。
【請求項１８】
前記変換後のマトリックスデータを生成することは、

前記第二のデータの中の、異なる記憶ユニットグループの同じ記憶アドレスに記憶されているデータをそれぞれ読み取ることによって、前記第二のデータの中の相応する行のデータとすることと、

前記第二のデータの中の各行のデータに対して前記第二の循環シフトを行うことによって、前記変換後のマトリックスデータの中の相応する行のデータを生成することとを含むことを特徴とする請求項１７に記載の方法。

【請求項１９】

前記マトリックスデータの中の第i列のデータを（ｉ－１）桁右循環シフトすることによって、前記第一のデータを生成し、iは、自然数であり、

前記第二のデータの中の第i列のデータを（ｉ－１）桁左循環シフトすることによって、前記変換後のマトリックスデータを生成することを特徴とする請求項１８に記載の方法。

【請求項２０】

前記マトリックスデータで表されるマトリックスは、ｎ行及びｍ列を含み、ｎ及びｍは、それぞれ自然数であり、

前記第一のデータを前記第二のデータとして記憶することは、

前記第一のデータの中の第１行のデータの中のｍ個の列のデータの中の第jデータを、前記複数の記憶ユニットグループの中の第j記憶ユニットグループの第j記憶アドレスにそれぞれ書き込み、jは、１以上且つｍ以下である自然数であることと、

前記第一のデータの中の第i行のデータの中のｍ個の列のデータの中の第jデータを、前記複数の記憶ユニットグループの中の第j記憶ユニットグループの第一の記憶アドレス及び第二の記憶アドレスにそれぞれ書き込み、iは、2以上且つn以下である自然数であり、

jが１以上且つｉ－１以下であるとき、前記第一の記憶アドレスは、第m+j-i+1記憶アドレスであり、jがi以上且つｍ以下であるとき、前記第二の記憶アドレスは、第j-i+1記憶アドレスであることとを含むことを特徴とする請求項１９に記載の方法。

【請求項２１】

コンピュータープログラムであって、

プロセッサにより実行されると、請求項１６～２０の何れか一項に記載の方法を実行させるコンピュータープログラム。

フロントページの続き

(72)発明者　シャオチャン・ゴン
中華人民共和国、ベイジン　１０００８５、ハイディアン・ディストリクト、シャンディ・１０ティーエイチ・ストリート、ナンバー１０、バイドゥ・キャンパス　２／エフ

Ｆターム(参考) 5B056 BB42

【外国語明細書】
2021005358000001.pdf