(19)日本国特許庁(JP)　(12)公表特許公報(A)　(11)公表番号
特表2025-505291
(P2025-505291A)
(43)公表日　令和7年2月21日(2025.2.21)

(51)国際特許分類	FI		テーマコード（参考）
G06F 17/10 (2006.01)	G06F	17/10 A	5B056

審査請求　未請求　予備審査請求　未請求　（全29頁）

(21)出願番号　特願2024-548397(P2024-548397)
(86)(22)出願日　令和5年2月13日(2023.2.13)
(85)翻訳文提出日　令和6年10月15日(2024.10.15)
(86)国際出願番号　PCT/CN2023/075661
(87)国際公開番号　WO2023/155748
(87)国際公開日　令和5年8月24日(2023.8.24)
(31)優先権主張番号　17/673,490
(32)優先日　令和4年2月16日(2022.2.16)
(33)優先権主張国・地域又は機関
米国(US)
(81)指定国・地域　AP(BW,CV,GH,GM,KE,LR,LS,MW,MZ,NA,RW,SD,SL,ST,SZ,TZ,UG,ZM,ZW),EA(AM,AZ,BY,KG,KZ,RU,TJ,TM),EP(AL,AT,BE,BG,CH,CY,CZ,DE,DK,EE,ES,FI,FR,GB,GR,HR,HU,IE,IS,IT,LT,LU,LV
最終頁に続く

(71)出願人　523062486
モフェット インターナショナル カンパニー，リミティド
中華人民共和国，香港 ９９９０７７，クーロン，サン ポー コン，ルーク ホップ ストリート ２９，ワン フェイ インダストリアル ビルディング，１１／フロア，ルーム ８
(74)代理人　100099759
弁理士　青木 篤
(74)代理人　100123582
弁理士　三橋 真二
(74)代理人　100112357
弁理士　廣瀬 繁樹
(74)代理人　100114018
最終頁に続く

(54)【発明の名称】　スパースニューラルネットワークのための適応テンソル畳み込みカーネル

(57)【要約】
適応テンソル演算カーネルを使用したニューラルネットワーク計算の効率化のための、コンピュータ記憶媒体にエンコードされたコンピュータプログラムを含む方法、システム、および装置が提供される。第一に、適応テンソル演算カーネルは、並列処理用の処理要素（ＰＥ）アレイに重みと入力値を分配するために、入力テンソル／重みテンソルの異なる形状に応じて形状を変形することができる。テンソル演算カーネルの形状によっては、畳み込み演算を実行するために、クラスタ間またはクラスタ内の追加の加算器が必要になる場合がある。第二に、適応テンソル演算カーネルは、全ての種類の畳み込み計算に対応するために、１×１テンソル演算モードと３×３テンソル演算モードという２つの異なるテンソル演算モードをサポートすることができる。第三に、基盤となるＰＥアレイは、スパースニューラルネットワークの異なる圧縮率と疎粒度に対応するために、各ＰＥ内部バッファ（レジスタファイルなど）を異なるように構成することができる。

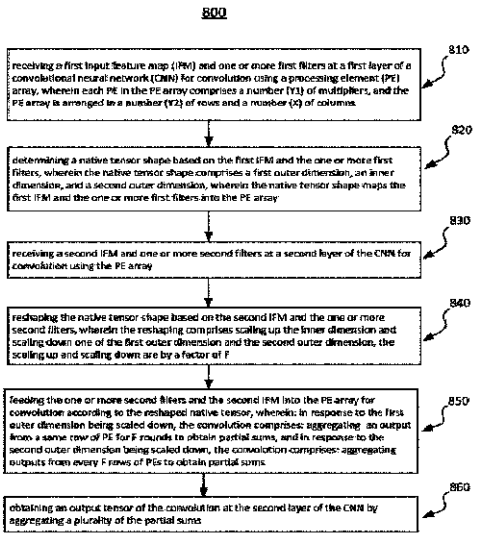

FIG. 8

【特許請求の範囲】
【請求項１】
コンピュータで実行される方法であって、
処理要素（ＰＥ）アレイを使用して畳み込みを行うための畳み込みニューラルネットワーク（ＣＮＮ）の第１層において、第１入力特徴マップ（ＩＦＭ）および１つまたは複数の第１フィルタを受信することであって、前記ＰＥアレイ内の各ＰＥは、（Ｙ１）個の乗算器を備え、前記ＰＥアレイは、（Ｙ２）個の行および（Ｘ）個の列に配置される、受信することと、
前記第１ＩＦＭおよび前記１つまたは複数の第１フィルタに基づいて、ネイティブテンソル形状を決定することであって、前記ネイティブテンソル形状は、第１外部次元、内部次元、および第２外部次元を含み、前記ネイティブテンソル形状は、前記第１ＩＦＭおよび前記１つまたは複数の第１フィルタを前記ＰＥアレイにマッピングする、決定することと、
前記ＰＥアレイを使用して畳み込みを行う前記ＣＮＮの第２層において、第２ＩＦＭおよび１つまたは複数の第２フィルタを受信することと、
前記第２ＩＦＭおよび前記１つまたは複数の第２フィルタに基づいて、前記ネイティブテンソル形状を変形することであって、前記変形は、前記内部次元の拡大と前記第１外部次元および前記第２外部次元のうちの一方の縮小を含み、前記拡大および縮小は係数Ｆによって実行される、変形することと、
変形された前記ネイティブテンソルに従って、前記１つまたは複数の第２フィルタおよび前記第２ＩＦＭを、畳み込みのために前記ＰＥアレイに供給することであって、
前記第１外部次元が縮小された場合、ＰＥの同一の行からの出力をＦラウンドについて合計して部分和を求めることを含み、
前記第２外部次元が縮小された場合、前記畳み込みは、ＰＥの全てのＦ倍した行からの出力を合計して部分和を求めることを含む、
供給することと、
複数の前記部分和を合計することにより、前記ＣＮＮの前記第２層において、前記畳み込みの出力テンソルを取得することであって、Ｙ１、Ｙ２、Ｘ、およびＦは全て１より大きい整数である、取得することと、
を含む、方法。
【請求項２】
前記ＣＮＮの前記第２層は、前記ＣＮＮの前記第１層の後にあり、前記第２ＩＦＭは、前記第１ＩＦＭよりも多くの入力チャンネルと、前記第１ＩＦＭよりも低い解像度と、を有する、請求項１に記載の方法。
【請求項３】
前記１つまたは複数の第２フィルタの各々が、２次元（２Ｄ）カーネルの複数のチャンネルを備え、各２Ｄカーネルは、１×１または３×３の次元を有する、請求項１に記載の方法。
【請求項４】
前記変形されたネイティブテンソルに従って、前記１つまたは複数の第２フィルタを前記ＰＥアレイに供給することは、
前記変形されたネイティブテンソルの前記第１外部次元および前記内部次元に従って、前記１つまたは複数の第２フィルタを行列に変換することであって、前記１つまたは複数の第２フィルタにおける各２Ｄカーネルが前記１×１の次元を有する場合、前記行列の各行は、前記１つまたは複数の第２フィルタの異なる入力チャンネルからの重みを含む、変換することと、
複数の入力チャンネルが一度に同時に処理されるように、前記行列の各行の重みをＰＥの異なる列に分配することと、
を含む、請求項３に記載の方法。
【請求項５】

前記変形されたネイティブテンソルに従って、前記ＰＥアレイに前記１つまたは複数の第２フィルタを供給することは、
前記変形されたネイティブテンソルの前記第１外部次元および前記内部次元に従って、前記１つまたは複数の第２フィルタを行列に変換することであって、前記１つまたは複数の第２フィルタにおける各２Ｄカーネルが前記３×３の次元を有し、９個の重みを含む場合、前記９個の重みは前記行列の同じ行に配置される、変換することと、
前記同じチャンネルからの前記重みが一度に同時に処理されるように、前記行列の前記同じ行からＰＥの異なる列に前記９個の重みを分配することと、
を含む、請求項３に記載の方法。
【請求項６】
前記変形されたネイティブテンソルに従って、前記ＩＦＭを前記ＰＥアレイへ供給することは、
前記変形されたネイティブテンソルの前記内部次元および前記第２外部次元に従って、前記ＩＦＭを行列に変換することと、
前記行列の列に対応する前記ＩＦＭの入力値を前記ＰＥの行のバッファに供給することと、
を含む、請求項５に記載の方法。
【請求項７】
前記１つまたは複数のフィルタのチャンネルを複数のチャンネルグループに分割することであって、各チャンネルグループは１より大きい整数である定数のチャンネルを含む、分割することと、
前記各チャンネルグループ内の重みの固定割合が非ゼロとなるように、前記複数のチャンネルグループのそれぞれをプルーニングすることと、
をさらに含む、請求項１に記載の方法。
【請求項８】
前記ＰＥアレイ内の各ＰＥに関連付けられたバッファの深さを決定することをさらに含み、
前記バッファの前記深さが前記定数より大きい場合、各ＰＥのプライベートメモリとして前記バッファを構成し、
前記バッファの前記深さが前記定数より小さい場合、前記ＰＥの前記バッファおよび隣接するＰＥの１つまたは複数のバッファを共有メモリとして結合する、
請求項７に記載の方法。
【請求項９】
各ＰＥの前記プライベートメモリは、前記ＰＥ内の前記（Ｙ１）個の乗算器によって取得可能な入力値を記憶する、請求項８に記載の方法。
【請求項１０】
前記共有メモリが、前記ＰＥおよび前記１つまたは複数の近隣ＰＥ内の前記（Ｙ１）個の乗算器によって取得可能な入力値を記憶する、請求項８に記載の方法。
【請求項１１】
ＰＥの各行は、各ＰＥ内の前記（Ｙ１）個の乗算器にそれぞれ対応する（Ｙ１）個の加算器ツリーと結合されており、各ＰＥ内の各乗算器は、合計のために、対応する加算器ツリーに積算出力を送信する、請求項１に記載の方法。
【請求項１２】
前記１つまたは複数の第２フィルタの各々は複数の非ゼロ重みを含み、前記１つまたは複数の第２フィルタを前記ＰＥアレイに畳み込みのために供給することは、
各非ゼロ重みを、前記非ゼロ重みおよび対応するインデックスを含むインデックス－値のペアとして、対応するＰＥの乗算器に供給することを含み、前記畳み込みは、
前記インデックスに従って、前記対応するＰＥのバッファから入力値を取得することと、
前記取得した値および前記非ゼロ重みを前記乗算器に送り、出力を取得することと、

前記対応するＰＥと同じ行にある他のＰＥの他の乗算器によって生成された出力と合計するために、対応する加算器ツリーに前記出力を送信することと、
を含む、請求項１に記載の方法。
【請求項１３】
各ＰＥ内の前記（Ｙ１）個の乗算器は、データを並列処理し、前記ＰＥアレイ内のＰＥは、データを並列処理する、請求項１に記載の方法。
【請求項１４】
１つまたは複数のプロセッサと、１つまたは複数のプロセッサに結合され、かつ、前記１つまたは複数のプロセッサによって実行可能な命令を備え、システムに演算を実行させるように構成された１つまたは複数の非一時的コンピュータ読み取り可能メモリと、を備えるシステムであって、
処理要素（ＰＥ）アレイを使用して畳み込みを行うための畳み込みニューラルネットワーク（ＣＮＮ）の第１層において、第１入力特徴マップ（ＩＦＭ）および１つまたは複数の第１フィルタを受信することであって、前記ＰＥアレイ内の各ＰＥは、（Ｙ１）個の乗算器を備え、前記ＰＥアレイは（Ｙ２）個の行および（Ｘ）個の列に配置される、受信することと、
前記第１ＩＦＭおよび前記１つまたは複数の第１フィルタに基づいて、ネイティブテンソル形状を決定することであって、前記ネイティブテンソル形状は、第１外部次元、内部次元、および第２外部次元を含み、前記ネイティブテンソル形状は、前記第１ＩＦＭおよび前記１つまたは複数の第１フィルタを前記ＰＥアレイにマッピングする、決定することと、
前記ＰＥアレイを使用して畳み込みを行う前記ＣＮＮの第２層において、第２ＩＦＭおよび１つまたは複数の第２フィルタを受信することと、
前記第２ＩＦＭおよび前記１つまたは複数の第２フィルタに基づいて、前記ネイティブテンソル形状を変形することであって、前記変形は、前記内部次元の拡大と前記第１外部次元および第２外部次元のうちの一方の縮小を含み、前記拡大および縮小は係数Ｆによって実行される、変形することと、
変形された前記ネイティブテンソルに従って、前記１つまたは複数の第２フィルタおよび前記第２ＩＦＭを、畳み込みのために前記ＰＥアレイに供給することであって、
前記第１外部次元が縮小された場合、前記畳み込みは、ＰＥの同一の行からの出力をＦラウンドについて合計して部分和を求めることを含み、
前記第２外部次元が縮小された場合、前記畳み込みは、ＰＥの全てのＦ倍した行からの出力を合計して部分和を求めることを含む、
供給することと、
複数の前記部分和を合計することにより、前記ＣＮＮの前記第２層において、前記畳み込みの出力テンソルを取得することであって、Ｙ１、Ｙ２、Ｘ、およびＦは全て１より大きい整数である、取得することと、
を含む、システム。
【請求項１５】
前記ＣＮＮの前記第２層は、前記ＣＮＮの前記第１層の後にあり、前記第２ＩＦＭは、前記第１ＩＦＭよりも多くの入力チャンネルと、前記第１ＩＦＭよりも低い解像度と、を有する、請求項１４に記載のシステム。
【請求項１６】
前記演算は、さらに、
前記１つまたは複数のフィルタのチャンネルを複数のチャンネルグループに分割することであって、各チャンネルグループは１より大きい整数である定数のチャンネルを含む、分割することと、
前記複数のチャンネルグループのそれぞれにおいて、１つのチャンネルのみが非ゼロの入力値を含み、前記各チャンネルグループの他のチャンネルは全てゼロを含むように、前記１つまたは複数のフィルタの各チャンネルをプルーニングすることと、

を含む、請求項１４に記載のシステム。
【請求項１７】
前記演算は、
前記ＰＥアレイ内の各ＰＥに関連付けられたバッファの深さを決定することさらに含み、
前記バッファの前記深さが前記定数より大きい場合、各ＰＥのプライベートメモリとして前記バッファを構成し、
前記バッファの前記深さが前記定数より小さい場合、前記ＰＥの前記バッファおよび隣接するＰＥの１つまたは複数のバッファを共有メモリとして結合する、
請求項１６に記載のシステム。
【請求項１８】
前記１つまたは複数の第２フィルタの各々は、２次元（２Ｄ）カーネルの複数のチャンネルを含み、各２Ｄカーネルは、１×１または３×３の次元を有する、請求項１４に記載のシステム。
【請求項１９】
前記変形されたネイティブテンソルに従って、前記１つまたは複数の第２フィルタを前記ＰＥアレイに供給することは、
前記変形されたネイティブテンソルの前記第１外部次元および前記内部次元に従って、前記１つまたは複数の第２フィルタを行列に変換することであって、前記１つまたは複数の第２フィルタにおける各２Ｄカーネルが前記１×１の次元を有する場合、前記行列の各行は、前記１つまたは複数の第２フィルタの異なる入力チャンネルからの重みを含む、変換することと、
複数の入力チャンネルが一度に同時に処理されるように、前記第１行列の各行の重みをＰＥの異なる列に分配することと、
を含む、請求項１８に記載のシステム。
【請求項２０】
１つまたは複数のプロセッサによって実行可能な命令を有するように構成された非一時的コンピュータ読み取り可能記憶媒体であって、前記１つまたは複数のプロセッサに、
処理素子（ＰＥ）アレイを使用した畳み込みのための畳み込みニューラルネットワーク（ＣＮＮ）の第１層において、第１入力特徴マップ（ＩＦＭ）および１つまたは複数の第１フィルタを受信することであって、前記ＰＥアレイ内の各ＰＥは、（Ｙ１）個の乗算器を備え、前記ＰＥアレイは（Ｙ２）個の行および（Ｘ）個の列に配置される、受信することと、
前記第１ＩＦＭおよび前記１つまたは複数の第１フィルタに基づいて、ネイティブテンソル形状を決定することであって、前記ネイティブテンソル形状は、第１外部次元、内部次元、および第２外部次元を含み、前記ネイティブテンソル形状は、前記第１ＩＦＭおよび前記１つまたは複数の第１フィルタを前記ＰＥアレイにマッピングする、決定することと、
前記ＰＥアレイを使用して畳み込みを行う前記ＣＮＮの第２層において、第２ＩＦＭおよび１つまたは複数の第２フィルタを受信することと、
前記第２ＩＦＭおよび前記１つまたは複数の第２フィルタに基づいて、前記ネイティブテンソル形状を変形することであって、前記変形は前記内部次元の拡大と前記第１外部次元および第２外部次元のうちの一方の縮小を含み、前記拡大および縮小は係数Ｆによって実行される、変形することと、
変形された前記ネイティブテンソルに従って、畳み込みを行うために、前記１つまたは複数の第２フィルタおよび前記第２ＩＦＭを前記ＰＥアレイに供給することであって、
前記第１外部次元が縮小された場合、前記畳み込みは、ＰＥの同一の行からの出力をＦラウンドについて合計して部分和を求めることを含み、
前記第２外部次元が縮小された場合、前記畳み込みは、ＰＥの全てのＦ倍した行からの出力を合計して部分和を求めることを含む、

供給することと、
複数の前記部分和を合計することにより、前記ＣＮＮの前記第２層において前記畳み込みの出力テンソルを取得することであって、Ｙ１、Ｙ２、Ｘ、およびＦは全て１より大きい整数である、取得することと、
を含む演算を実行させる、非一時的コンピュータ読み取り可能記憶媒体。
【発明の詳細な説明】
【技術分野】
【０００１】
本開示は、概してニューラルネットワークの計算効率の向上に関し、特に、スパースニューラルネットワークの様々な入力テンソルの形状や演算に適応するために、ネイティブテンソル（ｎａｔｉｖｅ　ｔｅｎｓｏｒ）の次元および演算モードを動的に調整することに関する。
【背景技術】
【０００２】
深層学習のＰＥ（処理要素（ｐｒｏｃｅｓｓｉｎｇ　ｅｌｅｍｅｎｔ））アレイは、ほとんど全てがネイティブテンソルの次元および演算モードで定められており、通常はコンパイラに依存して、様々なネストループマッピングアプローチを使用して、様々なテンソル形状（例えば、入力特徴マップやフィルタ）や演算に対応する。ＰＥアレイのネイティブテンソルの次元や演算モードと互換性のないテンソル形状や演算に対してＰＥアレイを使用して計算を行うのは明らかに非効率的である。柔軟性に欠けるＰＥアレイのネイティブテンソルの形状や演算モードでは、多数のゼロを持つテンソルを効率的に表現したり処理したりできないため、スパース特性を持つスパースニューラルネットワークではこの非互換性がさらに悪化する。
【発明の概要】
【０００３】
本明細書の種々の実施形態には、ニューラルネットワークの計算において適応テンソル計算カーネルを使用するためのシステム、方法、および非一時的なコンピュータ読み取り可能媒体が含まれ得る。
【０００４】
一態様によれば、ニューラルネットワークの計算における適応テンソル計算カーネルを使用するための方法は、処理要素（ＰＥ）アレイを使用した畳み込みのために、畳み込みニューラルネットワーク（ＣＮＮ）の第１層において、第１入力特徴マップ（ＩＦＭ）および１つまたは複数の第１フィルタを受信することであって、ＰＥアレイ内の各ＰＥは、（Ｙ１）個の乗算器を備え、ＰＥアレイは、（Ｙ２）個の行および（Ｘ）個の列に配置される、受信することと、第１ＩＦＭおよび１つまたは複数の第１フィルタに基づいてネイティブテンソル形状を決定することであって、ネイティブテンソル形状は、第１外部次元（ｆｉｒｓｔ　ｏｕｔｅｒ　ｄｉｍｅｎｓｉｏｎ）、内部次元（ｉｎｎｅｒ　ｄｉｍｅｎｓｉｏｎ）、および第２外部次元（ｓｅｃｏｎｄ　ｏｕｔｅｒ　ｄｉｍｅｎｓｉｏｎ）を含み、ネイティブテンソル形状は、第１ＩＦＭおよび１つまたは複数の第１フィルタをＰＥアレイにマッピングする、決定することと、ＰＥアレイを使用して畳み込みを行うＣＮＮの第２層において、第２ＩＦＭおよび１つまたは複数の第２フィルタを受信することと、第２ＩＦＭおよび１つまたは複数の第２フィルタに基づいてネイティブテンソル形状を変形すること（ｒｅｓｈａｐｉｎｇ）であって、その変形は、内部次元の拡大と、第１外部次元および第２外部次元のうちの一方の縮小とを含み、その拡大および縮小は係数Ｆによって実行される、変形することと、変形されたネイティブテンソルに従って畳み込みを行うために、１つまたは複数の第２フィルタおよび第２ＩＦＭをＰＥアレイに供給することであって、第１外部次元が縮小された場合、畳み込みは、ＰＥの同一の行からの出力をＦラウンドについて合計して部分和を求め、かつ、第２外部次元が縮小された場合、畳み込みは、ＰＥの全てのＦ倍した行からの出力を合計して部分和を求める、供給することと、複数の部分和を合計することにより、ＣＮＮの２層目における畳み込みの出力テンソル

を取得することであって、Ｙ１、Ｙ２、Ｘ、およびＦは全て１より大きい整数である、取得することと、を含んでもよい。

【０００５】

いくつかの実施形態において、ＣＮＮの第２層はＣＮＮの第１層の後にあり、第２ＩＦＭは、第１ＩＦＭよりも多くの入力チャンネルと、第１ＩＦＭよりも低い解像度と、を有する。

【０００６】

いくつかの実施態様において、１つまたは複数の第２フィルタの各々は、２次元（２Ｄ）カーネルの複数のチャンネルを備え、各２Ｄカーネルは、１×１（ｏｎｅ　ｂｙ　ｏｎｅ）または３×３（ｔｈｒｅｅ　ｂｙ　ｔｈｒｅｅ）の次元を有する。

【０００７】

いくつかの実施態様において、変形されたネイティブテンソルに従ったＰＥアレイへの１つまたは複数の第２フィルタの供給は、変形されたネイティブテンソルの第１外部次元および内部次元に従って、１つまたは複数の第２フィルタを行列に変換することであって、１つまたは複数の第２フィルタの各２Ｄカーネルが１×１（ｏｎｅ　ｂｙ　ｏｎｅ）の次元を有する場合、行列の各行は、１つまたは複数の第２フィルタの異なる入力チャンネルからの重みを含む、変換することと、複数の入力チャンネルが一度に同時に処理されるように、行列の各行の重みをＰＥの異なる列に分配することと、を含む。

【０００８】

いくつかの実施形態において、変形されたネイティブテンソルに従ってＰＥアレイに１つまたは複数の第２フィルタを供給することは、変形されたネイティブテンソルの第１外部次元および内部次元に従って、１つまたは複数の第２フィルタを行列に変換することであって、１つまたは複数の第２フィルタの各２Ｄカーネルが３×３の次元を有し、９個の重みを含む場合、９個の重みが行列の同じ行に配置される、変換することと、同じチャンネルからの重みが一度に同時に処理されるように、行列の同じ行からＰＥの異なる列に９個の重みを分配することと、を含む。

【０００９】

いくつかの実施例において、変形されたネイティブテンソルに従ったＰＥアレイへのＩＦＭの供給は、変形されたネイティブテンソルの内部次元および第２外部次元に従って、ＩＦＭを行列に変換することと、行列の列に対応するＩＦＭの入力値をＰＥの行のバッファに供給することと、を含む。

【００１０】

いくつかの実施例において、本方法はさらに、１つまたは複数のフィルタのチャンネルを複数のチャンネルグループに分割することであって、各チャンネルグループは１より大きい整数である定数のチャンネルを含む、分割することと、複数のチャンネルグループの各々におけるいくつかのチャンネルのみが非ゼロの入力値を含み、各チャンネルグループにおける他のチャンネルは全てゼロを含むように、１つまたは複数のフィルタの各々をプルーニングすることと、を含む。プルーニング後、複数のチャンネルグループのそれぞれは、非ゼロの重みと同一の割合の重みを含む。

【００１１】

いくつかの実施例において、本方法は、ＰＥアレイ内の各ＰＥに関連付けられたバッファの深さを決定することをさらに含んでもよく、バッファの深さが定数よりも大きい場合、各ＰＥのプライベートメモリとしてバッファを構成し、バッファの深さが定数よりも小さい場合、ＰＥのバッファと隣接するＰＥの１つまたは複数のバッファを共有メモリとして結合する。

【００１２】

いくつかの実施例において、各ＰＥのプライベートメモリは、ＰＥ内の（Ｙ１）個の乗算器によって取得可能な入力値を記憶する。

【００１３】

いくつかの実施例において、共有メモリは、ＰＥおよび１つまたは複数の隣接ＰＥ内の

（Ｙ１）個の乗算器によって取得可能な入力値を記憶する。
【００１４】
いくつかの実施例において、ＰＥの各行は、各ＰＥ内の（Ｙ１）個の乗算器にそれぞれ対応する（Ｙ１）個の加算器ツリーと結合されており、各ＰＥ内の各乗算器は、合計するために、対応する加算器ツリーに積算出力を送信する。
【００１５】
いくつかの実施例において、１つまたは複数の第２フィルタの各々は複数の非ゼロ重みを含み、畳み込みのために１つまたは複数の第２フィルタをＰＥアレイに供給することは、各非ゼロ重みを、非ゼロ重みおよび対応するインデックスを含むインデックス－値のペアとして、対応するＰＥの乗算器に供給することを含み、畳み込みは、インデックスに従って、対応するＰＥのバッファから入力値を取得することと、取得した値と非ゼロ重みを乗算器に送信して出力を取得することと、対応するＰＥと同じ行にある他のＰＥの他の乗算器によって生成された出力と合計するために、対応する加算器ツリーに出力を送信することと、を含む。
【００１６】
いくつかの実施例において、各ＰＥ内の（Ｙ１）個の乗算器は、データを並列処理し、ＰＥアレイ内のＰＥはデータを並列処理する。
【００１７】
さらに他の一態様によれば、システムは、１つまたは複数のプロセッサと、１つまたは複数のプロセッサに結合され、１つまたは複数のプロセッサによって実行可能な命令で構成され、システムに本明細書に記載のいずれかの方法を実行させるように構成された、１つまたは複数の非一時的コンピュータ可読メモリと、を備えてよい。
【００１８】
さらに他の一態様によれば、非一時的なコンピュータ読み取り可能な記憶媒体は、１つまたは複数のプロセッサが本明細書に記載された方法のいずれかを実行するように、１つまたは複数のプロセッサによって実行可能な命令で構成されてよい。
【００１９】
本明細書に開示されているシステム、方法、および非一時的コンピュータ読み取り可能媒体のこれらの機能およびその他の機能、ならびに構造の関連要素の動作方法および機能、ならびに一部の組み合わせおよび製造者の経済性は、以下の説明および添付の図面を参照した添付の特許請求の範囲を考慮することでより明らかにされ、これらは全て本明細書の一部を構成し、同様の参照番号は、様々な図における対応する部分を指す。ただし、図面は説明および例示のみを目的としており、本発明を限定する目的で作成されたものではないことを十分に理解されたい。
【図面の簡単な説明】
【００２０】
【図１】図１は、様々な実施形態によるＰＥアレイにおけるニューラルネットワーク計算処理のための例示的なシステム図を示す。
【図２】図２は、様々な実施形態によるＰＥアレイの例示的なアーキテクチャ図を示す。
【図３】図３は、様々な実施形態によるネイティブテンソル形状を使用するＰＥアレイにおける例示的なニューラルネットワーク計算を示す。
【図４Ａ】図４Ａは、様々な実施形態に従った適応テンソル形状を使用するＰＥアレイにおける例示的なニューラルネットワーク計算を示す。
【図４Ｂ】図４Ｂは、様々な実施形態に従った適応テンソル形状を使用するニューラルネットワーク計算のためのクラスタ間加算器を備える例示的なＰＥアレイを示す。
【図５Ａ】図５Ａは、様々な実施形態に従った適応テンソル形状を使用するＰＥアレイにおける他の例示的なニューラルネットワーク計算を示す。
【図５Ｂ】図５Ｂは、様々な実施形態に従った適応テンソル形状を使用するニューラルネットワーク計算のためのクラスタ内加算器を備えた他の例示的なＰＥアレイを示す。
【図６Ａ】図６Ａは、様々な実施形態に従った１×１テンソル演算モードを備えた例示的

なニューラルネットワーク計算を示す。
【図６Ｂ】図６Ｂは、様々な実施形態に従った３×３テンソル演算モードによる例示的なニューラルネットワーク計算を示す。
【図７】図７は、様々な実施形態に従った適応テンソル形状およびＰＥアレイにおける３×３テンソル演算モードを使用するニューラルネットワーク計算の例示的な方法を説明する図である。
【図８】図８は、様々な実施形態に従った適応テンソル形状を使用するニューラルネットワーク計算の例示的な方法を説明する図である。
【図９】図９は、本明細書で説明される実施形態のいずれかが実装され得る例示的なコンピュータシステムを示す図である。
【発明を実施するための形態】
【００２１】
本明細書で説明する実施形態は、適応テンソルの形状および演算モードを使用するＰＥアレイにおけるニューラルネットワーク計算のための方法、システム、装置を提供する。以下の説明では、適応テンソル計算カーネルは、異なる形状の入力特徴マップ（ＩＭＦ）および重みテンソル（例えば、フィルタ）を扱うために、複数のネイティブテンソルの次元および演算モードを持つものとして説明される。入力および出力テンソルの形状および演算モードに従って、適応テンソル計算カーネル（適応ネイティブテンソルとも称する）の次元および演算モードは、ＰＥアレイの基盤となるハードウェア資源を並列処理に完全に利用するために、動的に調整することができる。
【００２２】
これらの適応テンソル計算カーネルは、３つの技術的解決策を提供することにより、ニューラルネットワーク計算における技術的課題（背景のセクションで説明）に対処する。まず、適応テンソル計算カーネルは、入力テンソル／重みテンソルの様々な形状に適応して形状を調整することができる。入力テンソル／重みテンソルの様々な形状は、様々なニューラルネットワーク間に渡って存在するだけでなく、同じニューラルネットワークパイプライン内でも存在する場合がある。例えば、ニューラルネットワークの最初の数層を処理する際には、通常、テンソルは高解像度（高さおよび幅が大きい）ではあるが、入力および出力チャンネルは少なくなるように構成され、ニューラルネットワークの最後の数層を処理する際には、テンソルは低解像度（高さおよび幅が小さい）ではあるが、入力および出力チャンネルは多くなるように構成される。これは、ニューラルネットワークの最初の数層では、入力特徴マップからの特徴抽出に重点が置かれるのに対し、ニューラルネットワークの最後の数層では、抽出された特徴間の根本的な相関関係の学習に重点が置かれるためであると考えられる。
【００２３】
第二に、適応テンソル計算カーネルは、１×１テンソル演算モードと３×３テンソル演算モードという２つの異なるテンソル演算モードをサポートすることができる。行列の乗算が関わるこれらのニューラルネットワーク層では、１×１カーネルの畳み込み（例えば、重みテンソル内の各カーネルが１×１の形状を持つ）は１×１テンソル演算モードにマッピングされ、その他の畳み込み（３×３、５×５、７×７など）は３×３テンソル演算モードにマッピングすることができる。これらの異なるテンソル演算モードは、実行時に重みテンソルの形状に基づいて動的に決定することができる。
【００２４】
第三に、基盤となるＰＥアレイは、スパースニューラルネットワークの異なる圧縮率と疎粒度（ｓｐａｒｓｉｔｙ　ｇｒａｎｕｌａｒｉｔｉｅｓ）をサポートするために、各ＰＥ内部バッファ（例えば、レジスタファイル）を異なるように構成することができる。スパースニューラルネットワークがより細かい粒度でプルーニングされる場合（例えば、少数の入力チャンネルから１つまたは複数の非ゼロ入力チャンネルを選択し、その少数の数が閾値よりも小さい場合）、各ＰＥ内のレジスタファイルはプライベートメモリとして構成することができる（例えば、対応するＰＥのみが使用する）。スパースニューラルネッ

トワークが粗い粒度でプルーニングされる場合（例えば、多数の入力チャンネルから１つまたは複数の非ゼロ入力チャンネルを選択し、その多数の数が閾値よりも大きい場合）、隣接するＰＥ内の複数のレジスタファイルを、隣接するＰＥで共有されるマルチポートメモリとして構成することができる。

【００２５】

以下の説明において、図面を参照しながら、本発明の特定の非限定的な実施態様について説明する。本明細書で開示する任意の実施態様の特定の特徴および態様は、本明細書で開示する任意の他の実施態様の特定の特徴および態様と併用および／または組み合わせることができる。また、このような実施態様は例示的なものであり、あくまで、本願発明の範囲内の少数の実施態様を例示するものであることを理解されたい。当業者にとって明らかな、本発明に係る技術分野における様々な変更および修正は、添付の特許請求の範囲でさらに規定されているように、本発明の精神、範囲、および趣旨の範囲内であるとみなされる。

【００２６】

図１は、様々な実施形態に従ったＰＥアレイにおけるニューラルネットワーク計算処理のための例示的なシステム図を示す。図１の図は、ＰＥアレイを用いてパイプラインで実行される典型的なニューラルネットワーク計算のワークフローを示す。本開示で説明される実施形態は、図１におけるニューラルネットワーク計算の一部として、または他の適切な環境で実装されてもよい。

【００２７】

ニューラルネットワーク（例えば、ＣＮＮ）内の所定の（例えば、畳み込み）層において、１つまたは複数の入力特徴マップ（ＩＦＭ）１２０は、入力ソース（例えば、例えば入力画像など）または前の層（例えば、前の層から出力されたテンソルなど）から取得され、１つまたは複数の重みテンソル１１０が、様々な特徴を抽出するためにＩＦＭを畳み込むために使用されてよい。畳み込み処理は、ＰＥアレイ１６０と呼ばれる処理要素（ＰＥ）のアレイ内で並列に実行されてもよい。各ＰＥは、処理能力および記憶容量（例えば、バッファまたはキャッシュ）を有するプロセッサを指してもよい。ＰＥは、相互接続配線を有するＰＥアレイ内で特定の方法で配置されてもよく、実行時に動的に再配置されなくてもよい。ＰＥアレイは再利用され、ニューラルネットワークの異なる層において、または異なるニューラルネットワークおよび異なるユースケースに渡って演算に関与してよい。ＰＥアレイにおける固定された内部ＰＥ配置と、潜在的に無限の種類の（ＩＦＭおよび／または重みテンソルにおける）テンソル形状との間の非互換性は、通常、非効率的な資源利用と最適ではない並列処理につながる。

【００２８】

図１を参照すると、いくつかの実施形態において、ＩＦＭ１２０はＩＦＭキャッシュ１４０に記憶され、重みテンソル１１０はＰＥアレイ１６０の演算のために重みキャッシュ１３０に記憶されてもよい。ＰＥアレイ１６０は、ＰＥの行列（例えば、Ｘ×Ｙ）を含んでもよく、各ＰＥは、並列処理用の複数の乗算器を含んでもよい。いくつかの実施形態において、ＩＦＭキャッシュ１４０からの各ＩＦＭは、ＰＥアレイ１６０における計算を容易にするために、行列変換層１５０を通過してもよい。行列変換は、ｉｍ２ｃｏｌツールを使用して、オリジナルのＨＷＣフォーマット（Ｈは高さ、Ｗは重さ、Ｃはチャンネルを表す）からＲＳＣフォーマット（Ｒは行、Ｓは列、Ｃはチャンネルを表す）にＩＦＭを変形するＴｏｅｐｌｉｔｚ行列変換を含んでもよく、ここでＲＳＣフォーマットは重さテンソルの形状に基づいて決定される。ここで、変換とは、ＩＦＭ内の入力値を複製して配置し、変換されたＩＦＭを形成することであり、これにより、変換されたＩＦＭと重みテンソルとの間の行列乗算が、ＰＥアレイ１６０内のＰＥ間の依存性を最小限に抑えつつ、並列でＰＥアレイ１６０内で実行される。いくつかの実施形態において、ＰＥアレイ１６０における並列畳み込みの各ラウンドは、複数の部分和を生成する場合があり、その部分和は、累積バッファ１７０で合計されて、１つまたは複数の出力値を生成する場合がある。出力値は最終的に、現在の層によって生成される出力テンソルの一部となる場合がある。

【００２９】
図２は、様々な実施形態によるＰＥアレイの例示的なアーキテクチャ図を示す。図２のＰＥアレイにおけるＰＥの配列は例示的なものであり、ユースケースに応じて他の方法で実装することもできる。
【００３０】
図２の左部分に示されているように、ＰＥアレイ２００は、ＰＥの行列を含んでもよい。図２の右部分に示されているように、各ＰＥ２４０は、複数の乗算器（ＭＵＬゲート）を含んでもよい。各ＰＥ２４０内の乗算器は並列に演算してもよく、ＰＥアレイ２２０内のＰＥは並列に演算してもよい。参照を容易にするため、以下の説明では、ＰＥアレイ２００内のＰＥの列２２０の数をＸ、ＰＥアレイ２００内のＰＥの行２１０の数をＹ２、各ＰＥ２４０内の乗算器の数をＹ１と表記する。ＰＥ２１０の各行はＰＥクラスタと呼ばれ、各ＰＥクラスタは、ＰＥクラスタ内の乗算器によって生成された部分和を合計するために、Ｙ１個の加算器ツリー（Ｙ１加算器ツリー）２３０に結合されていてもよい。即ち、ＰＥクラスタ内の各ＰＥ２４０における第１乗算器は、合計用の第１加算器ツリー２３０に結合され、ＰＥクラスタ内の各ＰＥ２４０における第２乗算器は、合計用の第２加算器ツリー２３０に結合され、以下同様である。全てのＰＥクラスタにわたる加算器ツリー２３０（合計Ｙ１×Ｙ２加算器ツリー）を合計した結果は、合計のために加算器２５０に供給されてもよい。加算器２５０は、ネットワークオンチップ（ＮＯＣ）サブシステムの一部である数値の加算を行うデジタル回路を意味する場合がある。
【００３１】
いくつかの実施形態において、ＰＥアレイ２００が重みをＰＥにブロードキャストしてもよい。スパースニューラルネットワークの場合、重みの大部分はゼロであり、従ってＰＥにブロードキャストされる重みは全て非ゼロの重みである。非ゼロの重みは重みテンソル内の任意の位置から生じる可能性があるため、ブロードキャストされる各重みは重み値だけでなく、重み値の位置情報を示すインデックス、即ち（インデックス、重み値）のようなインデックス－値のペアも含む可能性がある。インデックスに基づいて、各ＰＥ２４０はＩＦＭから対応する入力値を取得し、重み値との乗算を行うことができる。乗算結果は対応する加算器ツリーに供給することができる。図１に示すように、第１乗算器ＭＵＬ１は、（インデックス１、値１）の形式で重みを受信し、インデックス１に基づいて（ＩＦＭを記憶する）ＩＢＵＦ２６０から入力値ＩＦＭ１を取得し、入力値ＩＦＭ１と値１に基づいて乗算を行い、合計のために結果を加算器ツリー１（例えば、ＰＥが位置するＰＥクラスタのＹ１加算器ツリー２３０の第１加算器ツリー）に送信する。
【００３２】
図３は、様々な実施態様に従ったネイティブテンソル形状を使用するＰＥアレイにおける例示的なニューラルネットワーク計算を示す。図３の例示的な計算では、変換された重みテンソルＡ（３１０）と変換されたＩＦＭテンソルＢ（３２０）との間の行列乗算が関与し、これにより出力特徴マップ（ＯＦＭ）テンソルＣ（３３０）が生成される。行列乗算では、ＸおよびＹの次元を有するＰＥアレイ３４０に対応するネイティブテンソル形状が使用される。
【００３３】
いくつかの実施形態において、変換された重みテンソルＡ（３１０）は、ＲＳＣフォーマット（３次元）の全ての重みテンソルをｍ´＊ｋ´と表記される２次元行列に統合することによって得られ（例えば、異なるチャンネルからの重みが同じチャンネルに再配置される）、ここで、ｍ´は重みテンソルの数（通常はＫと表記）によって決定される出力チャンネルの数であり、ｋ´は各重みテンソルのＲ、Ｓ、Ｃ次元の積である（ＲとＳは重みテンソル内の各カーネルの次元を指し、Ｃは入力チャンネルの数を指す）。
【００３４】
いくつかの実施形態において、変換されたＩＦＭテンソルＢ（３２０）は、ＲＳＣ形式に基づくＨＷＣ形式（３次元）の全てのＩＦＭを、ｋ´＊ｎ´と表記される２次元行列に統合することによって取得することができ、ここで、ｋ´は依然として各重みテンソルの

Ｒ、Ｓ、およびＣ次元の積であり、ｎ´はＩＦＭのＨおよびＷ次元の積である（Ｈは高さ、Ｗは幅を指す）。行列ｍ´＊ｋ´（重みテンソルＡ（３１０））と行列ｋ´＊ｎ´（ＩＦＭＢ（３２０））の行列積は、ｍ´＊ｎ´の行列としてＯＦＭテンソルＣ（３３０）を生成することができる。

【００３５】

上記の変換により、変換された重みテンソルＡ（３１０）および変換されたＩＦＭテンソルＢ（３２０）は、並列処理のためにＰＥアレイ３４０内のＰＥにマッピングされ得る。ＰＥアレイ３４０がＹ２個の行のＰＥ、Ｘ個の列のＰＥを含み、各ＰＥがＹ１個の乗算器を含むと仮定すると、テンソルＡおよびＢはＰＥアレイ３４０に以下のようにマッピングされ得る。即ち、テンソルＡ（３１０）およびテンソルＢ（３２０）の内部次元ｋ´はＰＥアレイ３４０のＸ（行）次元にマッピングされ、即ち、Ｘ＝ｋ´＝Ｒ＊Ｓ＊Ｃとなり、テンソルＡおよびテンソルＢの外部次元ｍ´＊＝ｎ´の乗算はＰＥアレイ３４０のＹ（列）次元にマッピングされる。ＰＥの各列にはＹ１＊Ｙ２乗算器が含まれるため、上記のマッピングにより、Ｙ＝ｍ’＊ｎ’＝Ｋ＊Ｈ＊Ｗの乗算はＹ１＊Ｙ２乗算器によって並列に処理されることになる。例えば、各ＰＥ内では、１つの乗算器が同じ出力チャンネルに対応する重み（例えば、全ての重みテンソルに渡る同じ位置からの重み）を処理し、即ち、Ｙ１＝Ｋ＝ｍ’であり、ＰＥの各列はＨ＊Ｗの重みを並列に処理し、即ち、Ｙ２＝Ｈ＊Ｗ＝ｎ’である。

【００３６】

上記の説明では、ワークロード（例えば、乗算のための重みと対応する入力値のペア）をＰＥアレイ３４０内のＰＥにマッピングするために、ネイティブテンソル形状ｍ’＊ｋ’＊ｎ’が確定され、ここで、Ｘ＝ｋ’、Ｙ１＊Ｙ２＝ｍ’＊ｎ’である。即ち、ＰＥアレイ内のＰＥのレイアウトに基づいて、ネイティブテンソル形状が決定される。ＰＥアレイのレイアウトが確定すると、ネイティブテンソルの形状も確定する。全ての入力テンソル（ＩＦＭやフィルタ／重みテンソルなど）は、確定したネイティブテンソルの形状に従って変換されなければならない。しかしながら、実際のアプリケーションにおける入力テンソルは形状が様々であり、ＰＥアレイ３４０におけるＰＥのレイアウトではなく、入力テンソルの形状に基づいて変換を行うと、最適な並列処理が可能になる。多くの場合、ＰＥレイアウトに基づいて決定された確定したネイティブテンソル形状を使用する変換は、ワークロードをＰＥにマッピングするとしても、特定のＰＥ間でいくつかの連続した依存関係（例えば、あるＰＥが他のＰＥの出力待ちになる）を引き起こす可能性がある。以下の説明では、ＩＦＭおよびフィルタの次元に基づいて決定され、同時にワークロードをＰＥにマッピングして並列性を最大化する適応ネイティブテンソル形状を使用する変換について説明する。

【００３７】

図４Ａは、様々な実施形態に従った適応テンソル形状を用いたＰＥアレイにおける例示的なニューラルネットワーク計算を示す。上述の通り（図３）、確定したネイティブテンソル形状ｍ’＊ｋ’＊ｎ’（即ち、行列ＡおよびＢをカバーする）を使用して入力テンソル（ＩＦＭ）および重みテンソルを行列Ａ（４１０）および行列Ｂ（４２０）に変換できる場合、変換されたテンソルは対応するＰＥアレイに分配することができる。しかしながら、実際のアプリケーションでは、乗算用のＩＦＭや重みテンソル（例えば、ＣＮＮの異なる層における、異なるレベルのスパース化を経たテンソル）は、ＰＥアレイに完全にマッピングできない様々な形状を有する可能性がある。確定したネイティブテンソル形状を使用してテンソルを強制的に変換すると、いくつかのＰＥがアイドル状態になるか、計算中にシーケンシャルな依存関係が発生する可能性がある。例えば、同じ畳み込みニューラルネットワーク（ＣＮＮ）内でも、最初の数層のＣＮＮのテンソルは高解像度（例えば、Ｈ＊Ｗ＝６４）で入力チャンネル数が少ない（Ｃ＝１６）場合があり、最後の数層のＣＮＮのテンソルは低解像度（例えば、Ｈ＊Ｗ＝１６）で入力チャンネル数が多い（Ｃ＝６４）場合がある。ここで、「少ない」と「多い」は閾値に基づいて決定される。即ち、同じＣＮＮ内の畳み込み処理でも、異なる形状のテンソルが現れる可能性がある。

【００３８】
いくつかの実施形態において、入力テンソルおよび重みテンソルの変化する形状に対応するために、ネイティブテンソルの形状が動的に変形される場合がある。例えば、入力テンソルが、入力チャンネル数が少ない（例えば、最初の数層のＣＮＮ層において）高解像度（画素数が多い）から、入力チャンネル数が多い（例えば、最後の数層のＣＮＮ層において）低解像度に変化する場合、ネイティブテンソル形状は、それに応じて変形することができる。いくつかの実施形態において、ネイティブテンソル形状は、第１外部次元、内部次元、および第２外部次元として示される３つの次元を有する。最初の２つの次元（第１外部次元および内部次元）は、重みテンソルを行列に変換するために使用することができ、最後の２つの次元（内部次元および第２外部次元）は、ＩＦＭを行列に変換するために使用することができる。変換された行列は、重みおよび入力値がＰＥアレイにどのようにマッピングされるか（例えば、最適な並列性を達成するために重みおよび入力値をどのように分配するか）についての指針を提供することができる。

【００３９】
いくつかの実施形態において、以前のテンソルがマッピングおよび変換にネイティブテンソル形状ｍ’＊ｋ‘＊ｎ’を使用し、入力テンソルが以前のテンソルと比較してより低い解像度でより多くの入力チャンネルを有する場合、ネイティブテンソル形状の３つの次元は、ｍ’＊（Ｆ＊ｋ’）＊（ｎ’／Ｆ）に変形することができ、ここで、Ｆは１より大きい整数であり、スケーリング係数を表し、最初の２つの次元（即ち、第１外部次元ｍ’および内部次元Ｆ＊ｋ’）は重みテンソル行列４２０を表し、次の２つの次元（即ち、内部次元Ｆ＊ｋ’および第２外部次元ｎ’／Ｆ）は畳み込みのためのＩＦＭテンソル行列４２２を表す。即ち、ネイティブテンソル形状は、その内部次元を係数Ｆで拡大し、第２外部次元（ＩＦＭテンソル行列に対応する）を係数Ｆで縮小することができる。この変形方法は、以下の説明ではｋ’＆ｎ’変形と呼ばれることがある。いくつかの実施形態では、Ｆは２、４、８などのいずれかである。

【００４０】
いくつかの実施形態において、変形されたテンソル形状の内部次元Ｆ＊ｋ’は、重みテンソル行列４２０およびＩＦＭテンソル行列４２２で共有され（例えば、同じ内部次元を有する）、ＰＥアレイにおけるＰＥの列数に対応し、第１外部次元（例えば、重みテンソル行列４２０の外部次元ｍ’）は、各ＰＥ内の乗算器の数に対応し、第２外部次元（例えば、ＩＦＭテンソル行列４２２の外部次元ｎ’／Ｆ）は、ＰＥアレイ内のＰＥの行の数に対応する。ここで、「対応する」とは、変換されたテンソル行列の重みおよび入力値がＰＥアレイにどのように分配されるかを指示するマッピング関係を指す。例えば、重みテンソル行列４２０の各外部次元（例えば、各列）における重みは、並列処理のために単一ＰＥ内の乗算器に分配されてもよく、ＩＦＭテンソル行列４２２の各外部次元（例えば、各列）における入力値は、ＰＥアレイ内のＰＥの行に渡って分配されてもよい。

【００４１】
図４Ａに示されているように、このように変形されたネイティブテンソル形状により、重みテンソル行列４１０は、その内部次元をＦ倍に拡大し、その外部次元ｍ’を同一に保つように変形され、それによって新たな重みテンソル行列４２０が形成される。即ち、重みテンソル行列の内部次元は、ｋ’＝Ｒ＊Ｓ＊Ｃ（例えば、４１０における行列Ａ）からＦ＊ｋ’＝Ｒ＊Ｓ＊（Ｆ＊Ｃ）（例えば、４２０における行列Ａ）に変化し、その結果、新しい行列４２０はより多くの入力チャンネル（ＣからＦ＊Ｃ）をサポートできる。同様に、ＩＦＭ行列４１２は、係数Ｆによってその外部次元を縮小し、重みテンソル行列４２０の縮小された内部次元と同じ方法でその内部次元を拡大し、それによって新しいＩＦＭテンソル行列４２２を形成する。即ち、ＩＦＭテンソル行列の内部次元は、ｋ’＝Ｒ＊Ｓ＊Ｃ（例えば、４１２における行列Ｂ）からＦ＊ｋ’＝Ｒ＊Ｓ＊（Ｆ＊Ｃ）（例えば、４２２における行列Ｂ）に変化し、ＩＦＭテンソル行列の外部次元はｎ’（例えば、４１２における行列Ｂ）からｎ’／Ｆ（例えば、４２２における行列Ｂ）に変化し、従って、新たな行列４２２はより少ない画素をサポートする可能性がある。従って、新たな行列４２

０および４２２は、より少ない入力チャンネルで低解像度を有する最後のいくつかのＣＮＮ層からのテンソルを表すのに、より適している。いくつかの実施形態において、「最初のいくつかのＣＮＮ層」および「最後のいくつかのＣＮＮ層」は、それぞれ、ＣＮＮ構造の最初から数えて第１の数のＣＮＮ層、およびＣＮＮ構造の最後から数えて第２の数のＣＮＮ層を指す場合がある。

【００４２】

例として、最初の新たなＣＮＮ層のテンソルは、高解像度Ｈ＊Ｗ＝６４で入力チャンネル数が少ないＣ＝１６であってよい。ここで、「少ない数」とは閾値よりも小さい数を指し、閾値は基盤となるＰＥアレイに従って構成されたコンパイラによって決定することができる。最初の数層のＣＮＮ層からのこれらのテンソルのネイティブテンソル形状は、ｍ’＝Ｋ＝１６、ｋ’＝１＊１＊１６、ｎ’＝６４の形状を有してよく、畳み込みが１＊１カーネルに基づいていると仮定する。畳み込みが最後の数層のＣＮＮ層に進むと、テンソルは低解像度のＨ＊Ｗ＝１６になってよく、入力チャンネルＣ＝６４（例えば、閾値よりも大きい）がより多くなると、ネイティブテンソルの形状はｍ’＝Ｋ＝１６，ｋ’＝１＊１＊６４，ｎ’＝１６というように変形されてよい。

【００４３】

上記のｋ’＆ｎ’変形を使用してテンソルを変換した後、変換されたテンソル行列４２０および４２２をＰＥアレイに分散させて並列処理を行うことができる。図４Ｂは、様々な実施形態に従ってｋ’＆ｎ’変形に基づく適応テンソル形状を使用するニューラルネットワーク計算のためのクラスタ間加算器を備えた例示的なＰＥアレイを示す。図４Ｂに示されるＰＥアレイを使用する並列処理スキームは、図４Ａに記述されたネイティブテンソルのｋ’＆ｎ’変形に対応してよい。一貫性を保つため、ＰＥアレイは、Ｙ２個の行とＸ個の列のＰＥを有し、各ＰＥはＹ１個の乗算器を有すると仮定する。

【００４４】

ｋ’＆ｎ’を変形することによって、重みテンソル行列とＩＦＭテンソル行列の内部次元はＦ倍に拡大され、ＩＦＭテンソル行列の外部次元はＦ分の１に縮小される。ＰＥアレイへの重みと入力値の分布は、重みテンソル行列の同じ行（即ち、拡大された内部次元／行に沿って）からの重みと、ＩＦＭテンソル行列の同じ列（即ち、同じく拡大された内部次元／列に沿って）からの入力値が、行ごとにＰＥに割り当てられるようにしてもよい。これは、これらの重みと入力値のペアが、ＰＥのＦ個の行に渡って分散される可能性があることを意味する。従って、ＰＥアレイは、クラスタ間（即ち、ＰＥクラスタ間または行間）の加算器４００を有し、畳み込み処理の部分和を求めるために、ＰＥの各行によって生成された出力を加算する。各クラスタ間加算器４００は、ＰＥのＦ個の行のＹ１加算器ツリーからの出力をＹ１部分和として合計することができる。これらの部分和は、畳み込みの結果として出力テンソルを構築するために合計される。このプロセスでは、部分和の合計はＹ１＊（Ｙ２／Ｆ）となり、これは、出力チャンネル数（例えば、畳み込みプロセスの出力テンソルのチャンネル数）がＹ１であり、出力画素数がＹ２／Ｆ＝Ｈ＊Ｗ／Ｆであることを意味する。

【００４５】

図５Ａは、様々な実施形態に従った適応テンソル形状を使用するＰＥアレイにおける他の例示的なニューラルネットワーク計算を示す。上述のｋ’＆ｎ’の変形と比較すると、ネイティブテンソル形状もまた、重みテンソルのスパース度に基づいて動的に変形することができる。多くの実用的な用途では、畳み込み処理における重みテンソルをプルーニングまたはスパース化して、計算効率を向上させ、ニューラルネットワークのフットプリントを削減することができる。慎重にプルーニングされた重みテンソルは、ゼロ値の重みを導入することで特徴抽出の精度を犠牲にすることなく畳み込み速度を向上させ、その結果、総計算数を削減することができる（例えば、ゼロ値の重みはスキップされる）。いくつかの実施形態では、重みテンソルのプルーニングは、重みテンソルのチャンネル（フィルタとも呼ばれる）を複数のチャンネルグループに分割し、全てのチャンネルグループが同じ数のチャンネルを持つようにすることと、各チャンネルグループの少数のチャンネルのみ

を非ゼロ入力チャンネル（例えば、非ゼロ重み）として保持し、そのチャンネルグループ内の他の全てのチャンネルをゼロ（例えば、全ての重みがゼロ）にすることと、を含んでもよい。プルーニング処理の後に、各チャンネルグループは、同じ割合の非ゼロ重みを含む。いくつかの実施形態において、プルーニングのためのチャンネルグループのサイズ（例えば、各チャンネルグループ内のチャンネルの数）は、重みテンソル（フィルタ）の数、即ち、出力チャンネルの数に基づいて決定されてもよい。一般的に、重みテンソルのプルーニングは、２つのレベルに分類されてもよく、即ち、出力チャンネル数（例えば、重みテンソル数）が第１閾値よりも大きく、非ゼロ入力チャンネル数が第２閾値よりも小さい、重みのスパース性が高い場合と、出力チャンネル数（例えば、重みテンソル数）が第１閾値よりも小さく、非ゼロ入力チャンネル数が第２閾値よりも大きい、重みのスパース性が低い場合である。例えば、重みのスパース性が高い場合（１６倍）のネイティブテンソル形状は、ｍ’＝Ｋ＝１６（例えば、１６個の重みテンソルまたはフィルタ）、ｋ’＝３＊３＊４（例えば、各カーネルは３＊３、１つのフィルタ内の非ゼロチャンネル数は４）、ｎ’＝６４となり、一方、重みのスパース性が低い場合（４倍）のネイティブテンソル形状は、ｍ’＝Ｋ＝４（例えば、４つの重みテンソルまたはフィルタ）、ｋ’＝３＊３＊１６（例えば、各カーネルは３＊３、１つのフィルタ内の非ゼロチャンネル数は１６）、ｎ’＝６４となる。

【００４６】

いくつかの実施形態において、重みのスパース性が高から低に変化する場合、ネイティブテンソルの形状（ｆｉｒｓｔ＿ｏｕｔｅｒ＿ｄｉｍｅｎｓｉｏｎ＊ｉｎｎｅｒ＿ｄｉｍｅｎｓｉｏｎ＊ｓｅｃｏｎｄ＿ｏｕｔｅｒ＿ｄｉｍｅｎｓｉｏｎと表記される）は、内部次元（重みテンソル行列とＩＦＭテンソル行列とで共有される）を係数Ｆで拡大し、第１外部次元（重みテンソル行列に対応する）を係数Ｆで縮小することによって変形されてよい。図５Ａに示されているように、元のネイティブテンソル形状ｍ’＊ｋ’＊ｎ’は、（ｍ’／Ｆ）＊（Ｆ＊ｋ’）＊ｎ’となり、重みテンソル行列５１０はｍ’＊ｋ’から、次元（ｍ’／Ｆ）＊（Ｆ＊ｋ’）の変形されたテンソル行列５２０に変化し、ＩＦＭテンソル行列５１２はｋ’＊ｎ’から（Ｆ＊ｋ’）＊ｎ’の次元を持つ変形されたＩＦＭ行列５２２に変化する。この変形方法は、以下の説明では、ｋ’＆ｍ’変形と呼ばれる場合がある。係数Ｆで拡大された内部次元は、より多くの入力チャンネル（ＣからＦ＊Ｃ）をサポートすることを示し、重みテンソル行列の縮小された外部次元は、より少ない出力チャンネル（ＫからＫ／Ｆ）をサポートすることを示す。

【００４７】

上記のｋ’＆ｍ’変形を使用してテンソルを変換した後、変換されたテンソル行列５２０および５２２をＰＥアレイに分散して並列処理を行うことができる。図５Ｂは、様々な実施形態に従ったｋ’＆ｍ’変形に基づく適応テンソル形状を使用するニューラルネットワーク計算のためのクラスタ間加算器を備えた他の例示的なＰＥアレイを示す。図５Ｂに示されたＰＥアレイを使用する並列処理スキームは、図５Ａに説明されたネイティブテンソルのｋ’＆ｍ’変形に対応し得る。

【００４８】

一貫性を保つため、ＰＥアレイは、Ｙ２個の行とＸ個の列のＰＥを有し、各ＰＥはＹ１個の乗算器を有すると仮定する。さらに、重み行列５２０とＩＦＭ行列５２２は、ＰＥアレイ内のＰＥの（Ｘ）個の列に対応する同一の内部次元を有し、重み行列５２０の外部次元はＰＥアレイにおける各ＰＥ内の（Ｙ１）個の乗算器に対応し、ＩＦＭ行列５２２の外部次元はＰＥアレイ内のＰＥの（Ｙ２）個の行に対応する。

【００４９】

変形されたネイティブテンソルは、（重み行列５２０の外部次元に対応する）第１外部次元をｍ’／Ｆとして有するため、重みテンソル行列の各列の重みは、各ＰＥ内のＹ１／Ｆ乗算器に供給することができる。ＰＥアレイから部分和を求めるために、Ｆラウンド分のＹ１加算器ツリーからの出力を記憶し合計するクラスタ内加算器５００が実装されてもよい。ここで、「ラウンド」とは、ＰＥ内の乗算器を使用して乗算を実行するサイクルを

指す。各ラウンドの間、Ｙ１／Ｆ乗算器の出力は、一時的に１つのクラスタ内加算器５００に記憶されてもよい。Ｆラウンド後、クラスタ内加算器５００には、Ｙ１加算器ツリーから収集されたＦ＊Ｙ１／Ｆ＝Ｙ１個の部分和が格納されていてもよい。これらの部分和は、畳み込みの結果として出力テンソルを構築するために合計されてもよい。このプロセスにおける部分和の合計は、＊Ｙ１／Ｆ）＊Ｙ２ととなり、これは出力チャンネル数（例えば、畳み込みプロセスの出力テンソルのチャンネル数）がＹ１／Ｆであり、出力画素数がＹ２＝Ｈ＊Ｗであることを意味する。

【００５０】

畳み込みニューラルネットワークの分野では、重みテンソルは、複数の２Ｄカーネルを含む３Ｄフィルタと呼ばれる場合がある。各３Ｄフィルタ内の２Ｄカーネルの数は、フィルタ内のチャンネル数と呼ばれる場合があり、各２Ｄカーネルは、１×１または３×３の行列であってよい。図６Ａは、様々な実施形態に従った１×１（ｏｎｅ　ｂｙ　ｏｎｅ）テンソル演算モード（即ち、１×１（ｏｎｅ　ｂｙ　ｏｎｅ）カーネルを使用）による例示的なニューラルネットワーク計算を示し、図６Ｂは、様々な実施形態に従った３×３テンソル演算モード（即ち、３×３カーネルを使用）による例示的なニューラルネットワーク計算を示す。

【００５１】

いくつかの実施形態において、一般的な行列乗算（ＧＥＭＭ）および１×１の畳み込み演算は、１×１の演算モードにマッピングされてよい（例えば、１×１カーネルを使用する）。図６Ａに示されているように、異なる入力チャンネル（または、スパース化された入力テンソル用のチャンネルグループ）からの２Ｄカーネル（即ち、重み）は、複数の入力チャンネルが一度に同時に処理されるように、ＰＥの異なる列に配置されてもよく、同じ入力チャンネルからの２Ｄカーネルは、乗算器が一度に同時に複数の出力チャンネルを処理できるように、１つのＰＥ内の複数の乗算器に分配されてもよい。例えば、チャンネル１（Ｃ＝１）からの重み数Ｙ１およびフィルタからの１～Ｙ１カーネル（即ち、複数のフィルタの同じ入力チャンネルからの重み）を第１ＰＥに供給し、チャンネル２（Ｃ＝２）からの重み数Ｙ１およびフィルタからの１～Ｙ１カーネルを第２ＰＥに供給することができる。このようにして、異なる入力チャンネルからの２ＤカーネルがＰＥの列に分散される。

【００５２】

スパース化された入力テンソルを含むいくつかの実施形態において、各重みはインデックス－値のペアとして表すことができる。インデックス－値のペアの値は、非ゼロの重みの値であり、インデックス－値のペアのインデックスは、非ゼロの重みのインデックスであり、これは、１つの乗算器で乗算を行うために、対応する入力値を識別するために使用することができる。いくつかの実施形態において、チャンネル数が各ＰＥクラスタ（各行）内のＰＥの数よりも少ない場合、残りのＰＥを他のベクトル演算に使用することができる。

【００５３】

いくつかの実施形態において、上述の１×１畳み込み演算以外の畳み込み演算は、１つまたは複数の３×３畳み込みに分解され、（例えば、３×３カーネルを使用して）３×３ネイティブ演算モードにマッピングされてもよい。図６Ｂに示されているように、各２Ｄの３×３カーネルは、同じ入力チャンネルからの９個の重み（（０，０）、（０，１）、（０，２）、（１，０）、（１，１）、（１，２）、（２，０）、（２，１）、（２，２））を有し、それらは同時に処理するためにＰＥの同じ行（異なる列）に分配され得る。異なる入力チャンネルからの９個の重みは、ＰＥの異なる行に分散させることができる。

【００５４】

図７は、様々な実施形態によるＰＥのアレイにおける内部バッファの例示的なアーキテクチャ図を示す。いくつかの実施形態において、ＰＥアレイ内の各ＰＥは、入力値を記憶するための入力バッファ（ＩＢＵＦ）７２２と結合される。これらの入力値は、対応する入力値を見つけるために所定の重みインデックスに基づいてＰＥによって取得することが

できる。取得された入力値は、ＰＥにおける乗算器内で重み値と乗算することができる。実際の実装では、ＩＢＵＦの深さは通常、制限されており、１つのＩＢＵＦ７２２は定数の入力チャンネルからの入力値のみを記憶できる。非ゼロの入力チャンネルの数も限られているため、この設計はスパース化された入力テンソルに対しては、有効に機能する。しかしながら、非ゼロの入力チャンネルの数がＩＢＵＦ７２２の深さを超えることもよく起こる。このような場合、ＩＢＵＦ７２２はキャッシュ置換を実行して外部メモリから必要な入力値を読み込む必要があるが、これはコストが高く非効率的である。

【００５５】

いくつかの実施形態において、重みテンソル（フィルタ）のスパース化の度合いに応じて、各ＰＥのＩＢＵＦをプライベートメモリまたは共有メモリとして構成することができる。例えば、１つまたは複数の重みテンソルをスパース化することは、１つまたは複数の重みテンソルの入力チャンネルを複数のチャンネルグループに分割することであって、各チャンネルグループは、１より大きい整数である定数のチャンネルを含む、分割することと、１つまたは複数の重みテンソルをそれぞれプルーニングして、複数のチャンネルグループの各々において、複数のチャンネルグループの各々における少数のチャンネルのみが非ゼロの入力値を含み、各チャンネルグループにおける他のチャンネルは全てゼロを含むようにすることと、を含み得る。即ち、プルーニングプロセス後、各チャンネルグループは、同じ割合の非ゼロ重みを含む。スパース化の粒度は、細粒度７１０と粗粒度７５０に分類することができる。細粒度７１０のスパース性は、非ゼロ入力チャンネルが、定数よりも小さい数のチャンネルから選択される場合に発生し、粗粒度７５０のスパース性は、非ゼロ入力チャンネルが、定数よりも大きい数のチャンネルから選択される場合に発生する。例えば、重みのスパース性が１５／１６（１６チャンネルのうちの１つが非ゼロ入力チャンネル）である場合、１６個の入力チャンネルごとに１つの非ゼロ入力チャンネルを選択すること（例えば、１つのチャンネルグループには１６個の入力チャンネルが含まれる）は、細粒度７１０のスパース性として決定されてよく、一方、６４個の入力チャンネルごとに４つの非ゼロ入力チャンネルを選択する（例えば、１つのチャンネルグループには６４個の入力チャンネルが含まれる）ことは、粗粒度７５０のスパース性として決定されてよい。

【００５６】

いくつかの実施形態において、ＩＢＵＦ７２２は、細粒度を有するスパース重みテンソル用のプライベートメモリとして、または、粗粒度を有するスパーステンソル用の共有メモリとして構成することができる。即ち、ＩＢＵＦ７２２の深さは、細粒度および粗粒度のスパース性を分類するために使用される定数と比較することができる。ＩＢＵＦ７２２の深さが定数よりも大きい場合、ＩＢＵＦ７２２は必要な入力値を記憶するのに十分であることを意味する。このように、専用プライベートメモリにより、データ取得性能が最適化される。ＩＢＵＦ７２２の深さが定数よりも小さい場合、複数の隣接するＰＥがそれらのＩＢＵＦを共有し、共有ＩＢＵＦと表記され、これにより、それらによって取得可能な入力値を記憶する。このように、重複する入力値が削減され、全体的な記憶効率が向上する。

【００５７】

図８は、様々な実施形態に従った適応テンソル形状を使用するニューラルネットワーク計算のための例示的な方法８００を示す。方法８００は、図１～７に記載されたデバイス、装置、またはシステムによって実行されてもよい。以下に提示される方法８００の動作は、例示的なものであることを意図している。実行内容によっては、方法８００は、様々な順序で、または並行して実行される追加の、より少ない、または代替のステップを含んでもよい。

【００５８】

ブロック８１０は、処理要素（ＰＥ）アレイを使用した畳み込みのために、畳み込みニューラルネットワーク（ＣＮＮ）の第１層において、第１入力特徴マップ（ＩＦＭ）および１つまたは複数の第１フィルタを受信することを含み、ＰＥアレイ内の各ＰＥは、（Ｙ

１）個の乗算器を備え、ＰＥアレイは、（Ｙ２）個の行および（Ｘ）個の列に配置される。いくつかの実施形態では、ＰＥの各行は、各ＰＥ内の（Ｙ１）個の乗算器にそれぞれ対応する（Ｙ１）個の加算器ツリーと結合されており、各ＰＥ内の各乗算器は、合計のために、対応する加算器ツリーに積算出力を送信する。各ＰＥ内の（Ｙ１）個の乗算器は並列に処理データを処理し、ＰＥアレイ内のＰＥは並列に処理データを処理する。

【００５９】

ブロック８２０は、第１ＩＦＭおよび１つまたは複数の第１フィルタに基づいてネイティブテンソル形状を決定することを含み、ネイティブテンソルの形状は、第１外部次元、内部次元、および第２外部次元を含み、ネイティブテンソル形状は、第１ＩＦＭおよび１つまたは複数の第１フィルタをＰＥアレイにマッピングする。

【００６０】

ブロック８３０は、ＰＥアレイを使用して畳み込みを行うＣＮＮの第２層において、第２ＩＭＦおよび１つまたは複数の第２フィルタを受信することを含む。いくつかの実施例において、ＣＮＮの第２層はＣＮＮの第１層の後にあり、第２ＩＦＭは第１ＩＦＭよりも多くの入力チャンネルと、第１ＩＦＭよりも低い解像度と、を有する。いくつかの実施例において、１つまたは複数の第２フィルタの各々は、複数の２次元（２Ｄ）カーネルのチャンネルを備え、各２Ｄカーネルは１×１（ｏｎｅ　ｂｙ　ｏｎｅ）または３×３（ｔｈｒｅｅ　ｂｙ　ｔｈｒｅｅ）の次元を有する。

【００６１】

ブロック８４０は、第２ＩＦＭおよび１つまたは複数の第２フィルタに基づいて、ネイティブテンソルの形状を変形することを含み、その変形は、内部次元を拡大し、第１外部次元および第２外部次元のうちの一方を縮小することを含み、拡大および縮小は、係数Ｆによって実行される。

【００６２】

ブロック８５０は、変形されたネイティブテンソルに従って畳み込みを行うために、１つまたは複数の第２フィルタおよび第２ＩＦＭをＰＥアレイに供給することを含み、第１外部次元が縮小された場合、畳み込みはＰＥの同一の行からの出力をＦラウンドについて合計して部分和を求めることを含み、第２外部次元が縮小された場合、畳み込みはＰＥの全てのＦ倍した行からの出力を合計して部分和を求めることを含む。いくつかの実施形態において、変形されたネイティブテンソルに従ってＰＥアレイに１つまたは複数の第２フィルタを供給することは、変形されたネイティブテンソルの第１外部次元および内部次元に従って１つまたは複数の第２フィルタを行列に変換することであって、１つまたは複数の第２フィルタの各２Ｄカーネルが１×１（ｏｎｅ　ｂｙ　ｏｎｅ）の次元をそれぞれ有していることに応じて、行列の各行は、１つまたは複数の第２フィルタの異なる入力チャンネルからの重みを含む、変換することと、複数の入力チャンネルが一度に同時に処理されるように、行列の各行の重みをＰＥの異なる列に分配することと、を含む。いくつかの実施形態において、変形されたネイティブテンソルに従ってＰＥアレイに１つまたは複数の第２フィルタを供給することは、変形されたネイティブテンソルの第１外部次元および内部次元に従って、１つまたは複数の第２フィルタを行列に変換することであって、１つまたは複数の第２フィルタの各２Ｄカーネルが３×３の次元を有し、９個の重みを含む場合、９個の重みは行列の同じ行に配置される、変換することと、同じチャンネルからの重みが一度に同時に処理されるように、行列の同じ行からＰＥの異なる列に９個の重みを分配することと、を含む。いくつかの実施形態において、変形されたネイティブテンソルに従ったＰＥアレイへのＩＦＭの供給は、変形されたネイティブテンソルの内部次元および第２外部次元に従ってＩＦＭをマトリックスに変換することと、行列の列に対応するＩＦＭの入力値をＰＥの行のバッファに供給することと、を含む。

【００６３】

ブロック８６０は、複数の部分和を合計することにより、ＣＮＮの第２層における畳み込みの出力テンソルを取得することを含む。

【００６４】

上記の説明において、Ｙ１、Ｙ２、Ｘ、およびＦは、全て１より大きい整数である。

【００６５】

いくつかの実施形態において、方法８００は、さらに、１つまたは複数のフィルタのチャンネルを複数のチャンネルグループに分割することであって、各チャンネルグループは１より大きい整数である定数のチャンネルを含む、分割することと、複数のチャンネルグループの各グループにおける少数のチャンネルのうちの１つのみが非ゼロの入力値を含み、各チャンネルグループの他のチャンネルは全てゼロを含むように、１つまたは複数のフィルタのそれぞれをプルーニングすることと、を含む。いくつかの実施形態において、方法８００は、さらに、ＰＥアレイ内の各ＰＥに関連するバッファの深さを決定することと、バッファの深さが定数より大きい場合、各ＰＥのプライベートメモリとしてバッファを構成することと、バッファの深さが定数より小さい場合、ＰＥのバッファと隣接するＰＥの１つまたは複数のバッファを共有メモリとして結合することと、を含んでもよい。いくつかの実施形態において、各ＰＥのプライベートメモリは、ＰＥ内の（Ｙ１）個の乗算器によって取得可能な入力値を記憶し、共有メモリは、ＰＥ内の（Ｙ１）個の乗算器および１つまたは複数の隣接するＰＥによって取得可能な入力値を記憶する。

【００６６】

いくつかの実施形態において、１つまたは複数の第２フィルタの各々は複数の非ゼロ重みを含み、１つまたは複数の第２フィルタを畳み込みのためにＰＥアレイに供給することは、各非ゼロ重みを、対応するＰＥの乗算器に、非ゼロ重みと対応するインデックスを含むインデックス－値のペアとして供給することを含み、畳み込みは、インデックスに従って、対応するＰＥのバッファから入力値を取得することと、取得した値および非ゼロ重みを乗算器に送信して出力を取得することと、対応するＰＥと同じ行にある他のＰＥの他の乗算器によって生成された出力と合計するために、対応する加算器ツリーにその出力を送信することと、を含む。

【００６７】

図９は、本明細書で説明されているいずれかの実施形態を実行することができる例示的な演算装置を示す。この演算装置は、図１から図８に示されているシステムおよび方法の１つまたは複数のコンポーネントを実装するために使用することができる。演算装置９００は、情報を伝達するためのバス９０２または他の通信メカニズムと、情報を処理するためのバス９０２に結合された１つまたは複数のハードウェアプロセッサ７０４と、を備えてよい。ハードウェアプロセッサ（複数可）７０４は、例えば、１つまたは複数の汎用マイクロプロセッサとすることができる。

【００６８】

演算装置９００は、情報およびプロセッサ（複数可）９０４によって実行される命令を記憶するためにバス９０２に結合された、ランダムアクセスメモリ（ＲＡＭ）などのメインメモリー９０７、キャッシュおよび／またはその他の記憶装置も備えてよい。メインメモリー９０７は、プロセッサ（複数可）９０４によって実行される命令の実行中に、一時変数またはその他の中間情報を記憶するためにも使用することができる。このような命令は、プロセッサ（複数可）９０４がアクセス可能な記憶媒体に記憶されると、演算装置９００を、命令で規定された演算を実行するようにカスタマイズされた専用機にすることができる。メインメモリー９０７は、不揮発性媒体および／または揮発性媒体を含んでもよい。不揮発性媒体は、例えば、光ディスクまたは磁気ディスクを含んでもよい。揮発性媒体は、ダイナミックメモリを含んでもよい。一般的な媒体の形態には、例えば、フロッピーディスク、フレキシブルディスク、ハードディスク、半導体ドライブ、磁気テープ、その他の磁気データ記憶媒体、ＣＤ－ＲＯＭ、その他の光データ記憶媒体、穿孔パターンを設けた物理的媒体、ＲＡＭ、ＤＲＡＭ、ＰＲＯＭ、ＥＰＲＯＭ、フラッシュ（登録商標）ＥＰＲＯＭ、ＮＶＲＡＭ、その他のメモリチップまたはカートリッジ、またはそれらのネットワーク化されたバージョンが含まれる。

【００６９】

演算装置９００は、カスタマイズされた配線ロジック、１つまたは複数のＡＳＩＣまた

はＦＰＧＡ、ファームウェアおよび／または演算装置と組み合わせることで演算装置９００を特殊用途の機械にすることができるプログラムロジックを、本明細書で説明されている技術を実装するために使用してもよい。一実施形態によると、本明細書に記載の技術は、メインメモリー９０７に含まれる１つまたは複数の命令の１つまたは複数のシーケンスを実行するプロセッサ９０４に応答して、演算装置９００によって実行される。このような命令は、記憶装置９０９などの他の記憶媒体からメインメモリー９０７に読み込むことができる。メインメモリー９０７に含まれる一連の命令の実行により、プロセッサ９０４が本明細書に記載された処理ステップを実行することができる。例えば、本明細書に開示されたプロセス／方法は、メインメモリー９０７に記憶されたコンピュータプログラム命令により実行することができる。これらの命令がプロセッサ（複数可）９０４によって実行される場合、対応する図に示され、上述されたステップを実行することができる。代替の実施形態では、配線回路がソフトウェア命令の代わりに、またはソフトウェア命令と組み合わせて使用されてもよい。

【００７０】

演算装置９００は、バス９０２に結合された通信インターフェース９１０も備える。通信インターフェース９１０は、１つまたは複数のネットワークに接続された１つまたは複数のネットワークリンクに結合された双方向データ通信を提供することができる。他の例として、通信インターフェース９１０は、互換性のあるＬＡＮ（または、ＷＡＮと通信するＷＡＮコンポーネント）にデータ通信接続を提供するローカルエリアネットワーク（ＬＡＮ）カードであってもよい。無線リンクを実装することもできる。

【００７１】

特定の動作の実行は、単一のマシン内だけでなく、多数のマシンに渡って展開されるプロセッサ間で分散されてもよい。いくつかの例示的な実施形態において、プロセッサまたはプロセッサ実装エンジンは単一の地理的位置（例えば、家庭環境、オフィス環境、またはサーバファーム内）に配置されてもよい。他の例示的な実施形態において、プロセッサまたはプロセッサ実装エンジンは複数の地理的位置に分散配置されてもよい。

【００７２】

前述の各セクションで説明されているプロセス、方法、アルゴリズムは、１つまたは複数のコンピュータシステム、またはコンピュータハードウェアを備えるコンピュータプロセッサによって実行されるコードモジュールに実装され、完全にまたは部分的に自動化されてよい。プロセスおよびアルゴリズムは、部分的にまたは完全に、アプリケーション固有の回路に実装されてよい。

【００７３】

本明細書で開示された機能がソフトウェア機能単位で実装され、独立した製品として販売または使用される場合、それらはプロセッサで実行可能な不揮発性のコンピュータ読み取り可能な記憶媒体に記憶することができる。本明細書で開示された特定の技術的解決策（全体または一部）または現在の技術に貢献する一態様は、ソフトウェア製品の形で具現化することができる。ソフトウェア製品は、記憶媒体に記憶され、演算装置（パーソナルコンピュータ、サーバー、ネットワーク装置など）に本発明の実施形態の方法の全てまたはいくつかのステップを実行させる複数の命令を含んでよい。記憶媒体は、フラッシュ（登録商標）ドライブ、ポータブルハードドライブ、ＲＯＭ、ＲＡＭ、磁気ディスク、光ディスク、プログラムコードを記憶するために動作可能な他の媒体、またはそれらの組み合わせを含んでよい。

【００７４】

特定の実施形態はさらに、プロセッサと、プロセッサが実行可能な命令を記憶する非一時的なコンピュータ可読記憶媒体と、を備え、システムが、上述の実施形態の任意の方法におけるステップに対応する動作を実行するように構成される。特定の実施形態は、さらに、１つまたは複数のプロセッサが実行可能な命令を含む非一時的なコンピュータ可読記憶媒体を備え、１つまたは複数のプロセッサが、上述の実施形態の任意の方法におけるステップに対応する動作を実行するように構成される。

【００７５】
本明細書で開示されている実施形態は、クライアントとやりとりするクラウドプラットフォーム、サーバー、またはサーバーグループ（以下、総称して「サービスシステム」という）を通じて実装することができる。クライアントは、端末機器であってもよいし、プラットフォームでユーザが登録したクライアントであってもよく、端末機器は、モバイル端末、パーソナルコンピュータ（ＰＣ）、およびプラットフォームアプリケーションプログラムがインストール可能な如何なる機器であってもよい。
【００７６】
上記で説明した様々な機能およびプロセスは、互いに独立して使用される場合もあれば、様々な方法で組み合わせられる場合もある。全ての可能な組み合わせおよび部分的な組み合わせは、本開示の範囲に含まれることが意図されている。さらに、いくつかの実装では、特定の方法またはプロセスブロックが省略される場合がある。また、本明細書で説明した方法およびプロセスは、特定の順序に限定されるものではなく、それらに関連するブロックまたは状態は、適切な他の順序で実行することができる。例えば、説明されたブロックまたは状態は、具体的に開示された順序以外の順序で実行されてもよく、または、複数のブロックまたは状態が単一のブロックまたは状態に結合されてもよい。例示的なブロックまたは状態は、直列、並列、またはいくつかの他の方法で実行されてもよい。ブロックまたは状態は、開示された例示的な実施形態に追加されるか、またはそこから削除されてもよい。本明細書で説明される例示的なシステムおよびコンポーネントは、説明されたものとは異なる方法で構成されてもよい。例えば、開示された例示的な実施形態と比較して、要素が追加、削除、または再配置される場合がある。
【００７７】
本明細書で説明した例示的な方法の様々な動作は、少なくとも部分的に、アルゴリズムによって実行することができる。アルゴリズムは、メモリ（例えば、上述の非一時的なコンピュータ読み取り可能な記憶媒体）に記憶されたプログラムコードまたは命令を含んでよい。このようなアルゴリズムは、機械学習アルゴリズムを含んでよい。いくつかの実施形態において、機械学習アルゴリズムは、機能を実行するようにコンピュータに明示的にプログラムするのではなく、機能を実行する予測モデルを作成するためにトレーニングサンプルから学習することができる。
【００７８】
本明細書で説明されている例示的な方法の様々な動作は、関連する動作を実行するように一時的に（例えば、ソフトウェアによって）または恒久的に構成された１つまたは複数のプロセッサによって、少なくとも部分的に実行することができる。一時的にまたは恒久的に構成されたか否かに関わらず、そのようなプロセッサは、本明細書で説明されている１つまたは複数の動作または機能を実行するように動作可能に構成されたプロセッサ実装エンジンを構成することができる。
【００７９】
同様に、本明細書に記載される方法は、少なくとも一部がプロセッサによって実行されてもよく、特定のプロセッサまたはプロセッサ群はハードウェアの一例である。例えば、方法の少なくともいくつかの動作は、１つまたは複数のプロセッサまたはプロセッサ実装エンジンによって実行されてもよい。さらに、１つまたは複数のプロセッサは、「クラウドコンピューティング」環境内または「サービスとしてのソフトウェア」（ＳａａＳ）として、関連動作の実行をサポートするように動作してもよい。例えば、本実施例におけるいくつかの動作は、プロセッサを含む機械の一例であるコンピュータのグループによって実行され、これらの動作は、ネットワーク（例えば、インターネット）および１つまたは複数の適切なインターフェース（例えば、アプリケーションプログラムインターフェース（ＡＰＩ））を介してアクセス可能である。
【００８０】
特定の動作の実行は、単一のマシン内だけでなく、複数のマシンに渡って展開されるプロセッサ間で分散されてもよい。いくつかの例示的な実施形態において、プロセッサまた

はプロセッサ実装エンジンは単一の地理的位置（例えば、家庭環境、オフィス環境、またはサーバファーム内）に配置されてもよい。他の例示的な実施形態では、プロセッサまたはプロセッサ実装エンジンは複数の地理的位置に分散配置されてもよい。

【００８１】

本明細書全体を通して、複数の例が、単一の例として説明されているコンポーネント、操作、または構造を実装する場合がある。１つまたは複数の方法の個々の動作は、個別の動作として図示および説明されているが、１つまたは複数の個々の動作は同時に実行されてもよく図示されている順序で動作を実行する必要はない。構成例における個別のコンポーネントとして示されている構造および機能は、統合された構造またはコンポーネントとして実装されてもよい。同様に、単一のコンポーネントとして提示された構造および機能は、個別のコンポーネントとして実装されてもよい。これらおよびその他のバリエーション、修正、追加、改善は、本願発明の範囲に含まれる。

【００８２】

本書で使用される場合、「または」は、他に明確な指示がない限り、または文脈によって他に明確な指示がない限り、包括的であり排他的ではない。従って、本書では、「Ａ、Ｂ、またはＣ」は、「Ａ、Ｂ、ＡおよびＢ、ＡおよびＣ、ＢおよびＣ、または、Ａ、Ｂ、およびＣ」を意味するが、他に明確な指示がない限り、または文脈によって他に明確な指示がない限り、この限りではない。さらに、「および」は、他に明確な指示がない限り、または文脈によって他に明確な指示がない限り、包括的であり個別的でもある。従って、本明細書では、「ＡおよびＢ」は「ＡおよびＢの両方またはどちらか一方」を意味し、別段の明示的な指示または文脈による別段の指示がない限り、この限りではない。さらに、本明細書で単一の例として説明されている資源、動作、または構造については、複数の例が提供される場合がある。さらに、各種の資源、動作、エンジン、およびデータ記憶装置の境界は、ある程度任意であり、特定の動作は、特定の例示的な構成の文脈で説明されている。その他の機能の割り当ても想定されており、本開示の様々な実施形態の範囲に含めることができる。一般的に、例示的な構成における個別の資源として提示された構造および機能は、統合された構造または資源として実装することができる。同様に、単一の資源として提示された構造および機能は、個別の資源として実装することができる。これらのバリエーション、修正、追加、および改善は、添付の特許請求の範囲に示される本開示の実施形態の範囲に含まれる。従って、本明細書および図面は、制限的な意味ではなく、むしろ、例示的な意味で解釈されるべきである。

【００８３】

「含む」または「備える」という用語は、後に述べられる特徴の存在を示すために使用されるが、他の特徴の追加を排除するものではない。「～できる」、「～できた」、「～かもしれない」、「～してもよい」などの条件文は、特に明記されていない限り、または文脈上、そう解釈されない限り、一般的に、特定の実施形態には特定の機能、要素および／またはステップが含まれるが、他の実施形態には含まれないことを意味する。従って、このような条件付きの表現は、機能、要素および／またはステップが１つまたは複数の実施形態に何らかの形で必要であることを意味するものではなく、また、１つまたは複数の実施形態が、これらの機能、要素および／またはステップが特定の実施形態に含まれるか、または実行されるべきであるか否かを、ユーザ入力またはプロンプトの有無にかかわらず決定するためのロジックを必ずしも含むことを意味するものではない。

【００８４】

特定の例示的な実施形態を参照しながら発明の概要を説明したが、本開示の実施形態のより広い範囲から逸脱することなく、これらの実施形態に対して様々な修正や変更を加えることができる。本発明のこのような実施態様は、便宜上、単に「発明」という用語によって、個別にまたは集合的に本明細書で言及される場合があるが、実際には複数の開示または概念が開示されている場合であっても、本出願の範囲を任意の単一の開示または概念に自主的に限定することを意図するものではない。

【００８５】

本明細書で説明されている実施形態は、当業者が開示された発明を実施できるように、十分詳細に説明されている。他の実施形態も利用でき、そこから派生させることも可能であり、構造的および論理的な置換や変更を、本開示の範囲から逸脱することなく行うことができる。従って、詳細な説明は限定的な意味で解釈されるべきではなく、様々な実施形態の範囲は、添付の請求の範囲のみによって定義され、そのような請求の範囲に含まれる全ての均等物とともに定義される。

【図面】

【図１】

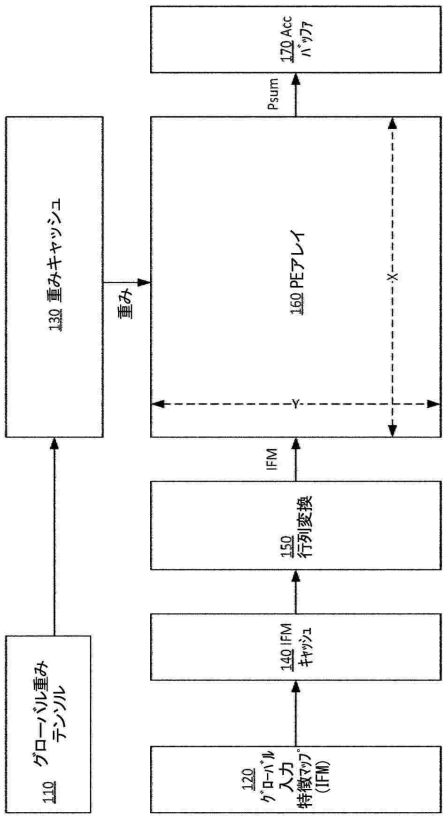

FIG. 1

【図２】

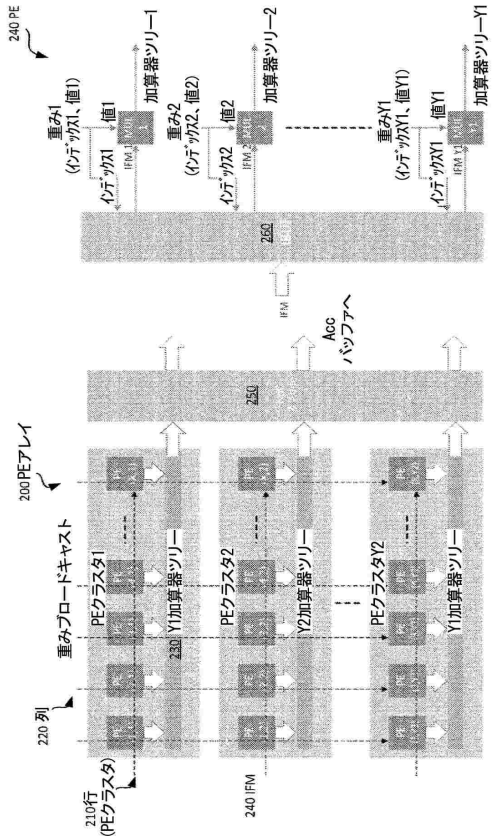

FIG. 2

【図３】

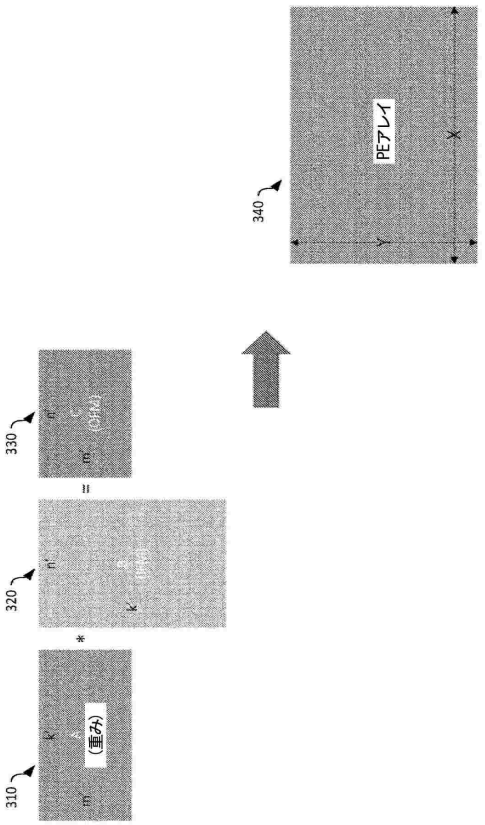

FIG. 3

【図４Ａ】

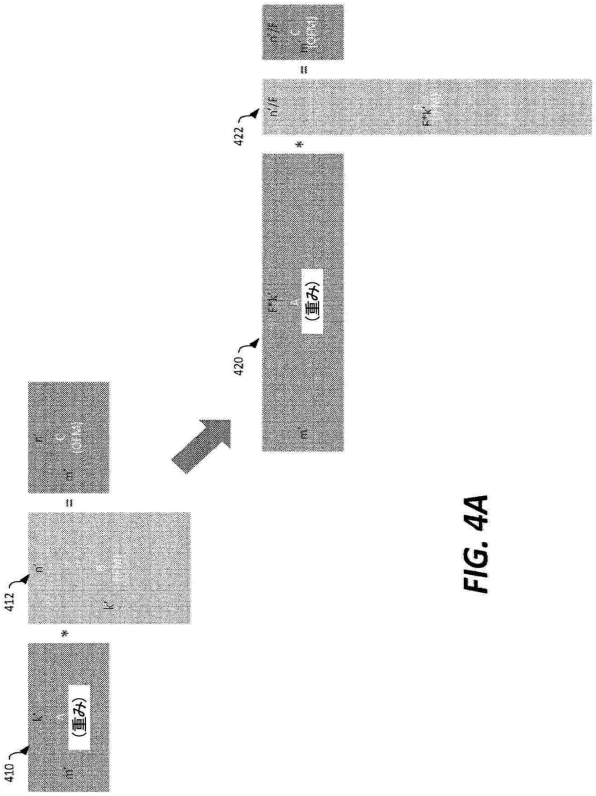

FIG. 4A

【図４Ｂ】

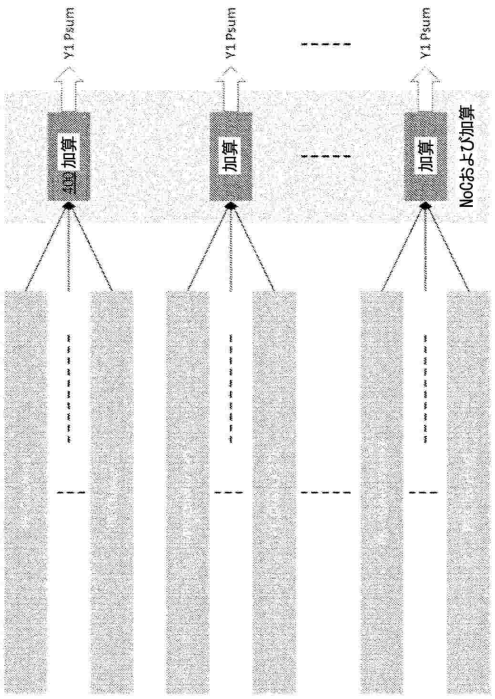

FIG. 4B

【図５Ａ】

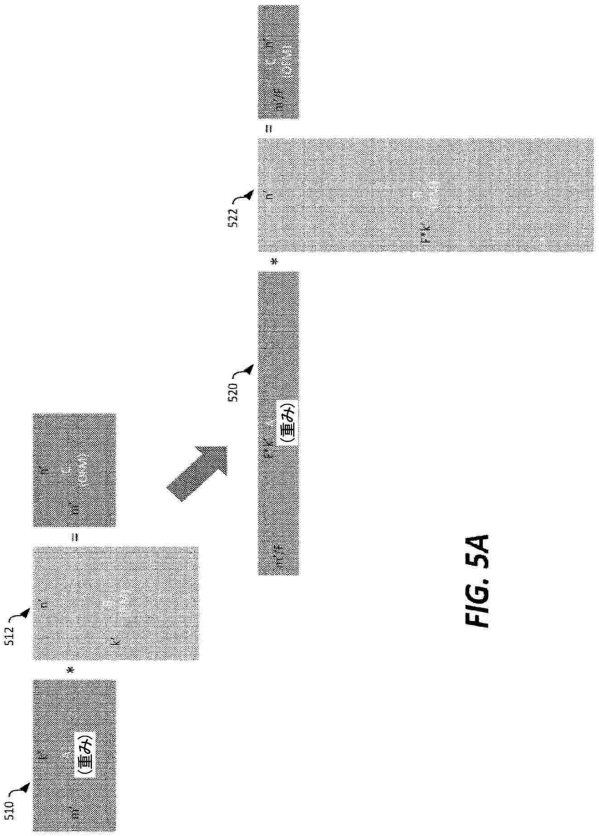

FIG. 5A

【図５Ｂ】

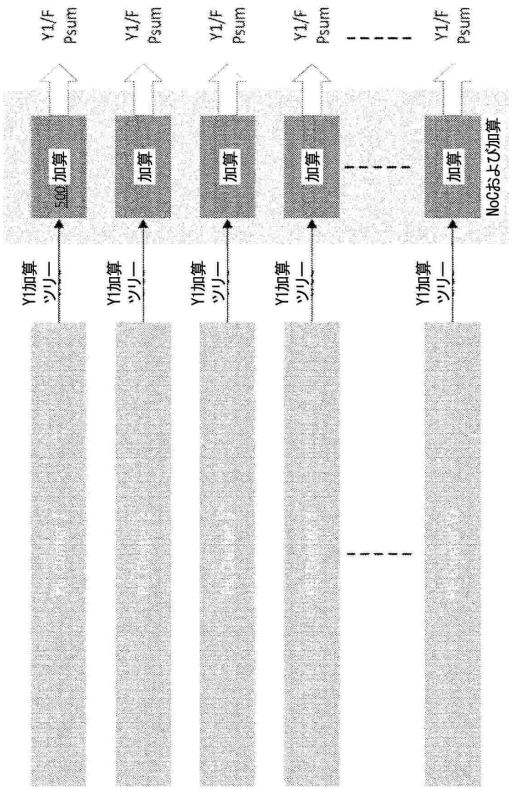

FIG. 5B

【図６Ａ】

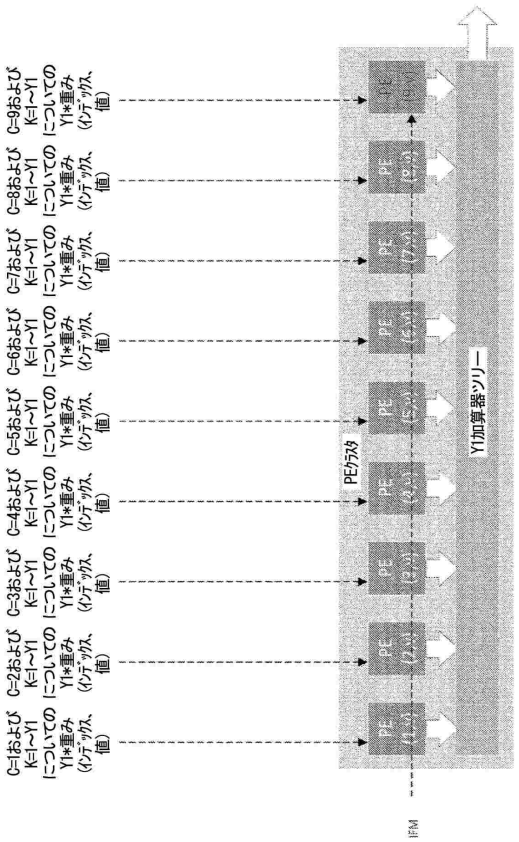

FIG. 6A

【図６Ｂ】

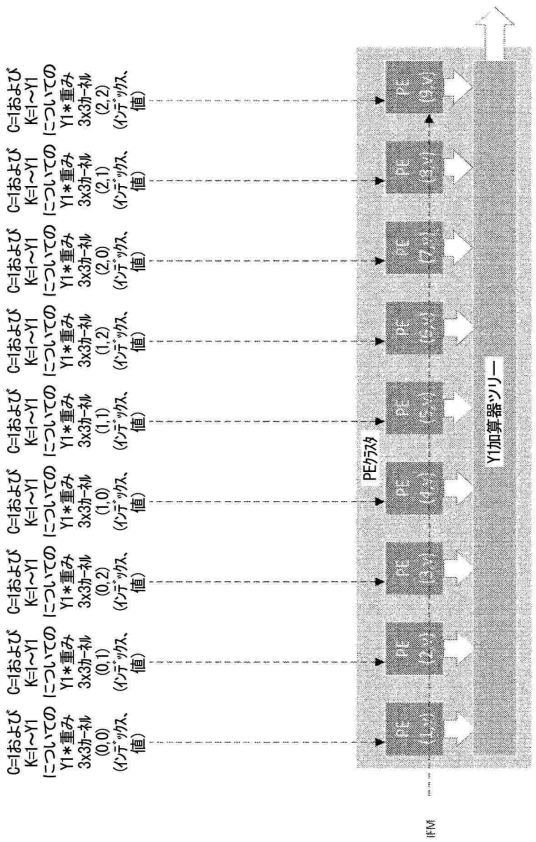

FIG. 6B

【図７】

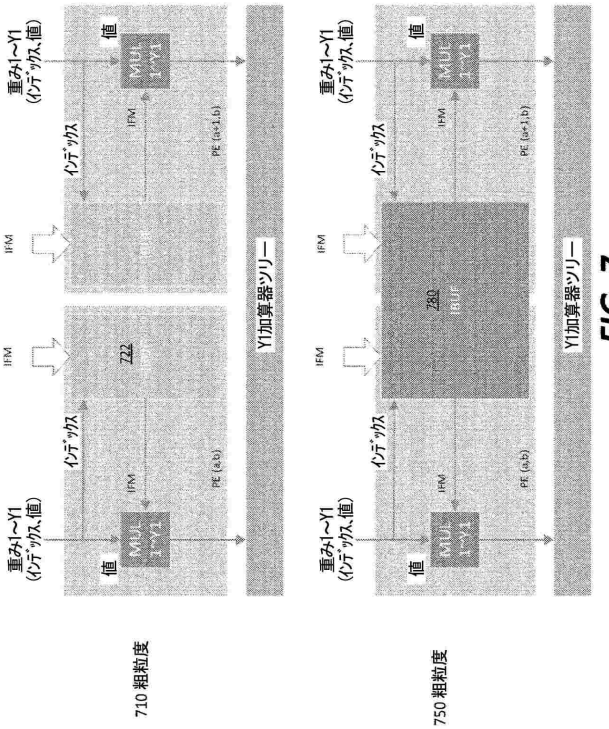

FIG. 7

【図８】

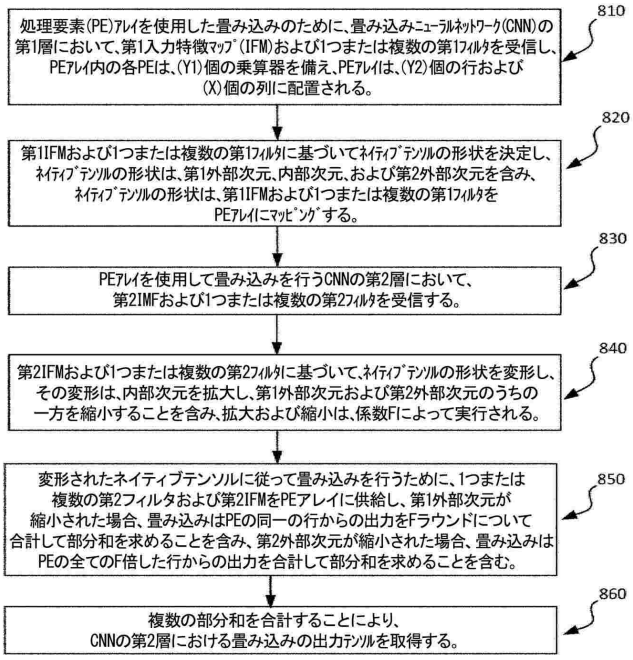

FIG. 8

【図９】

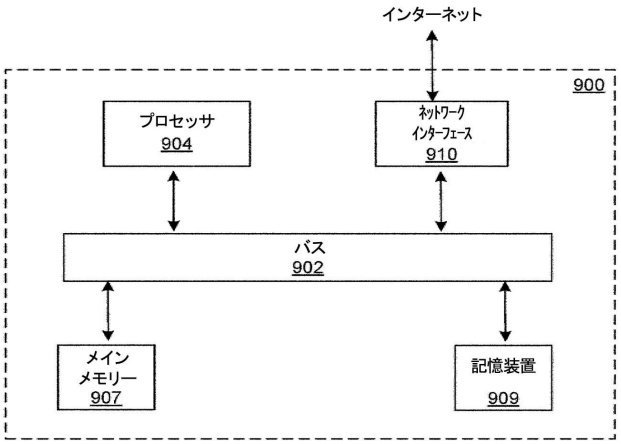

FIG. 9

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2023/075661

A. CLASSIFICATION OF SUBJECT MATTER

G06N3/04(2023.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC: G06N,G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

VEN, CNTXT, WOTXT, EPTXT, USTXT, CNKI, IEEE: processor element array, convolution, tensor, filter, neural network, kernel, shape, reshape, dimension, adaptive

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2020134417 A1 (INTEL CORPORATION) 30 April 2020 (2020-04-30) description, paragraphs [0021]-[0071]	1-20
A	CN 108875958 A (GUANGZHOU YIGOU INTELLIGENT TECHNOLOGY) 23 November 2018 (2018-11-23) the whole document	1-20
A	CN 113449857 A (HUAWEI TECHNOLOGIES CO., LTD.) 28 September 2021 (2021-09-28) the whole document	1-20
A	US 2020293858 A1 (SAMSUNG ELECTRONICS CO., LTD.) 17 September 2020 (2020-09-17) the whole document	1-20

☐ 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
"D" document cited by the applicant in the international application
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
27 April 2023	**06 May 2023**
Name and mailing address of the ISA/CN **CHINA NATIONAL INTELLECTUAL PROPERTY ADMINISTRATION** **6, Xitucheng Rd., Jimen Bridge, Haidian District, Beijing 100088, China**	Authorized officer **JIE,Xin** Telephone No. **(+86) 010-53961366**

Form PCT/ISA/210 (second sheet) (July 2022)

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2023/075661

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
US	2020134417	A1	30 April 2020	DE	102020131050	A1	24 June 2021
				CN	113033765	A	25 June 2021
				IN	202044041613	A	25 June 2021
				KR	20210082058	A	02 July 2021
				TW	202127324	A	16 July 2021
CN	108875958	A	23 November 2018	US	10223334	B1	05 March 2019
				US	10216704	B1	26 February 2019
				US	10169298	B1	01 January 2019
				US	10073816	B1	11 September 2018
CN	113449857	A	28 September 2021	WO	2021190127	A1	30 September 2021
				CN	115456159	A	09 December 2021
				CN	115456160	A	09 December 2021
				CN	115456161	A	09 December 2021
				EP	4123515	A1	25 January 2023
				US	2023023101	A1	26 January 2023
US	2020293858	A1	17 September 2020	IN	201941009806	A	18 September 2020
				KR	20200110165	A	23 September 2020

Form PCT/ISA/210 (patent family annex) (July 2022)

フロントページの続き

,MC,ME,MK,MT,NL,NO,PL,PT,RO,RS,SE,SI,SK,SM,TR),OA(BF,BJ,CF,CG,CI,CM,GA,GN,GQ,GW,KM,ML,MR,NE,SN,TD,TG),AE,AG,AL,AM,AO,AT,AU,AZ,BA,BB,BG,BH,BN,BR,BW,BY,BZ,CA,CH,CL,CN,CO,CR,CU,CV,CZ,DE,DJ,DK,DM,DO,DZ,EC,EE,EG,ES,FI,GB,GD,GE,GH,GM,GT,HN,HR,HU,ID,IL,IN,IQ,IR,IS,IT,JM,JO,JP,KE,KG,KH,KN,KP,KR,KW,KZ,LA,LC,LK,LR,LS,LU,LY,MA,MD,MG,MK,MN,MW,MX,MY,MZ,NA,NG,NI,NO,NZ,OM,PA,PE,PG,PH,PL,PT,QA,RO,RS,RU,RW,SA,SC,SD,SE,SG,SK,SL,ST,SV,SY,TH,TJ,TM,TN,TR,TT,TZ,UA,UG,US,UZ,VC,VN,WS,ZA,ZM,ZW

弁理士　南山 知広
(74)代理人　100153729
弁理士　森本 有一
(74)代理人　100151459
弁理士　中村 健一
(72)発明者　シアオチエン チャン
アメリカ合衆国，カリフォルニア ９４０２２，ロス アルトス，シェアウッド アベニュ ９４９，スイート ２００
(72)発明者　エンシュイ イェン
アメリカ合衆国，カリフォルニア ９４０２２，ロス アルトス，シェアウッド アベニュ ９４９，スイート ２００
(72)発明者　チーピン シアオ
アメリカ合衆国，カリフォルニア ９４０２２，ロス アルトス，シェアウッド アベニュ ９４９，スイート ２００

Ｆターム（参考）　5B056　BB26 BB38 FF01 FF02 FF05 FF07