(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2019-3414
(P2019-3414A)
(43)公開日　平成31年1月10日(2019.1.10)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 15/80	(2006.01)	G06F	15/80		5B045
G06F 17/10	(2006.01)	G06F	17/10	A	5B056
G06N 3/063	(2006.01)	G06N	3/063		
G06F 15/173	(2006.01)	G06F	15/173	683D	

審査請求 未請求　請求項の数 10　ＯＬ　（全 19 頁）

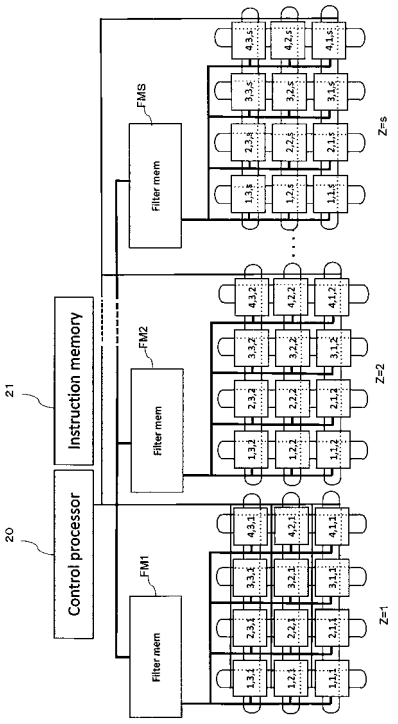
(21)出願番号　特願2017-117686 (P2017-117686)
(22)出願日　平成29年6月15日 (2017.6.15)

(71)出願人　506301140
公立大学法人会津大学
福島県会津若松市一箕町大字鶴賀字上居合９０番地
(74)代理人　100094525
弁理士　土井　健二
(74)代理人　100094514
弁理士　林　恒徳
(72)発明者　富岡　洋一
福島県会津若松市一箕町大字鶴賀字上居合９０番地　公立大学法人会津大学内
(72)発明者　スタニスラフ　セドゥーキン
福島県会津若松市一箕町大字鶴賀字上居合９０番地　公立大学法人会津大学内

Ｆターム(参考)　5B045 GG13
5B056 AA05　BB26　FF01　FF02　FF05

(54)【発明の名称】データ処理装置、及びこれにおけるデータ処理方法

(57)【要約】
【課題】演算素子プロセッシングエレメントを三次元的に多数配置し、並列性を保ったまま省力で高速の計算を行えるデータ処理方法及びデータ処理装置を提供する。
【構成】乗算及び加算機能を有する複数のプロセッシングエレメントを３次元方向に有するデータ処理装置であって、それぞれ複数の前記プロセッシングエレメントが２次元方向に配置され、Ｚ軸方向に積層された複数の２次元面を有し、前記複数の２次元面のそれぞれに対応して特徴重みが配置されるフィルタメモリを有し、入力データがＺ軸方向の上位面の２次元面から配置され、一の面に配置されたプロセッシングエレメントで前記乗算機能により前記入力データと前記特徴重みの積を順次演算して２次元畳み込みデータを演算し、更に下面から転送されるデータと自身のデータを加算する演算を行い、当該演算結果を隣接する上面のプロセッシングエレメントに転送することを特徴とする。
【選択図】図４



【特許請求の範囲】
【請求項１】
乗算及び加算機能を有する複数のプロセッシングエレメントを３次元方向に有するデータ処理装置であって、
それぞれ複数の前記プロセッシングエレメントが２次元方向に配置され、Ｚ軸方向に積層された複数の２次元面を有し、
前記複数の２次元面のそれぞれに対応して特徴重みが配置されるフィルタメモリを有し、
入力データがＺ軸方向の上位面の２次元面から配置され、
一の面に配置されたプロセッシングエレメントで前記乗算機能により前記入力データと前記特徴重みの積を順次演算して２次元畳み込みデータを演算し、更に下面から転送されるデータと自身のデータを加算する演算を行い、当該演算結果を隣接する上面のプロセッシングエレメントに転送する、
ことを特徴とするデータ処理装置。
【請求項２】
請求項１において、
前記２次元方向に配置されたプロセッシングエレメントはトーラスネットワークに接続され、Ｚ軸方向には、上下面に隣接するプロセッシングエレメントが、ネットワークで双方向に接続される、
ことを特徴とするデータ処理装置。
【請求項３】
請求項１において、
前記２次元畳み込みデータは、隣接するプロセッシングエレメントからの転送データと自身のデータを加算演算し、更にシフト方向に隣接するプロセッシングエレメントに前記加算演算結果を転送する、
ことを特徴とするデータ処理装置。
【請求項４】
請求項１において、
前記Ｚ軸方向の最上位面にあるプロセッシングエレメントは、下面の複数のプロセッシングエレメントから転送される２次元畳み込みデータと自身のデータを加算して２．５次元畳み込みデータを演算する、
ことを特徴とするデータ処理装置。
【請求項５】
請求項４において、
前記２．５次元畳み込みデータは、順次下面のプロセッシングエレメントにシフトされる、
ことを特徴とするデータ処理装置。
【請求項６】
請求項１乃至４の何れか１項において、
前記特徴重みは、前記入力データの配置された２次元面の数で分割され、前記入力データの配置された面毎に対応するフィルタメモリに配置し、前記畳み込みの演算の際、前記フィルタメモリに配置された特徴重みを、対応する面の全てのプロセッシングエレメントにブロードキャストする、
ことを特徴とするデータ処理装置。
【請求項７】
乗算及び加算機能を有する複数のプロセッシングエレメントを３次元方向に有するデータ処理装置におけるデータ処理方法であって、
前記データ処理装置は、それぞれ複数の前記プロセッシングエレメントが２次元方向に配置され、Ｚ軸方向に積層された複数の２次元面を有し、前記複数の２次元面のそれぞれに対応して特徴重みが配置されるフィルタメモリを有し、

入力データを前記Ｚ軸方向の上位面の２次元面から配置する工程と、
一の面に配置されたプロセッシングエレメントで前記乗算機能により前記入力データと前記特徴重みの積を順次演算して２次元畳み込みデータを演算する工程と、
更に下面から転送されるデータと自身のデータを加算する演算を行い、当該演算結果を隣接する上面のプロセッシングエレメントに転送する工程を、
有することを特徴とするデータ処理方法。
【請求項８】
請求項７において、
さらに、前記２次元畳み込みデータは、隣接するプロセッシングエレメントからの転送データと自身のデータを加算演算し、更にシフト方向に隣接するプロセッシングエレメントに前記加算演算結果を転送する工程を、
有することを特徴とするデータ処理方法。
【請求項９】
請求項７において、
前記Ｚ軸方向の最上位面にあるプロセッシングエレメントは、下面の複数のプロセッシングエレメントから転送される２次元畳み込みデータと自身のデータを加算して２．５次元畳み込みデータを演算する工程を、
有することを特徴とするデータ処理方法。
【請求項１０】
請求項９において、
前記２．５次元畳み込みデータを、順次下面のプロセッシングエレメントにシフトする工程を、
有することを特徴とするデータ処理方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、データ処理装置、及びこれにおけるデータ処理方法に関し、特に、畳み込みニューラルネットワークにおける畳み込み演算に適したデータ処理装置、及びこれにおけるデータ処理方法に関する。
【背景技術】
【０００２】
ニューラルネットワークに畳み込み(圧縮処理：Convolution))を追加した畳み込みニューラルネットワーク（Convolutional Neural Network:以下適宜ＣＮＮと表記）が、特に画像認識に有効な機械学習として広く認識されている。
【０００３】
図１は、ＣＮＮのシステム構成の概略を示す図である。入力データに対して、複数の層(レイヤーＬ１－Ｌ５))構造で処理を行う。
【０００４】
図１では、レイヤーＬ１、Ｌ２のそれぞれは、畳み込み層(Convolutional Layer）、プーリング層（Pooling Layer）を含みこれを繰り返す。
【０００５】
畳み込み層は、入力データに対して フィルタ(kernel)特徴を乗算する（特徴量を畳み込む）層である。入力データが画像データである場合、入力データ（画像）に対して、それぞれ異なるフィルタ特徴を乗算してフィルタの数に対応する画像を得ている。複数のフィルタを使うことにより入力画像のさまざまな特徴が捉えられ、特徴量の畳み込みによって画像内のパターンが検出出来る。
【０００６】
プーリング層は、畳み込み層の直後に置かれ、レイヤーを縮小して扱い易くし、抽出された特徴の位置感度を低下させる。
【０００７】

ＣＮＮは、次いで、レイヤーＬ３－Ｌ５により全結合した多層パーセプトロンを配置して入力データ（画像）を認識する。

【０００８】

ここで、序盤のレイヤーで行う畳み込み演算には膨大な計算回数が必要である。このためかかる部分の省電力化が非常に重要な課題となっている。

【０００９】

しかし、複数の演算素子（ＰＥ：プロセッシングエレメント）を、アレイ状に配置するアレイ型の並列演算処理素子とすると、周辺機能ブロックとの類似度計算をするために多くの配線資源や転送時間が必要となる。

【００１０】

かかる点に鑑みて、本発明者等は、先にＰＥ間の通信でのデータ衝突を回避し、かつＰＥを特定の方向に偏ることなく増加させることが可能な拡張性の高いデータ処理装置を提案している(特許文献１)。

【００１１】

かかる先の発明技術では、ｎ次元のネットワークを構成するｎ次元の方向に配置された全てのＰＥが、転送クロックに同期してデータを入出力する。そして、データを入出力する方向であるシフト方向に隣接する第１の隣接ＰＥから第１のデータを受け取るとともに、反対側に隣接する第２の隣接ＰＥに第２のデータを出力し、隣接するＰＥ間のデータ転送レートがシフト方向によらず等しいという特徴を有する演算装置である。

【先行技術文献】

【特許文献】

【００１２】

【特許文献１】特許第５９３９５７２号公報

【発明の概要】

【発明が解決しようとする課題】

【００１３】

これまでのＣＮＮ演算のための技術は、上記特許文献１に提案の発明に従う場合であっても、並列演算において個々のＰＥの処理量が大きくなるもの、即ち、技術的に最速であるがエネルギーに乏しいＰＥあるいはコアの数が、メモリに蓄積されるデータの数よりはるかに小さい。

【００１４】

換言すれば、各演算ステップにおけるアクティブなプロセッサ、メモリ動作の数が基本的にＣＮＮアルゴリズムにおける可能性より小さいものであった。

【００１５】

結果として、ＣＮＮの解決のための時間が、最小値よりはるかに大きく、解決すべきエネルギーが高くなる。

【００１６】

かかる点に鑑みて、本発明の目的は、演算素子ＰＥを三次元的に多数配置し、並列性を保ったまま省電力で高速の計算を行えるデータ処理装置、及びこれにおけるデータ処理方法を提供することにある。

【課題を解決するための手段】

【００１７】

上記目的を達成する本発明に従う第１の側面は、乗算及び加算機能を有する複数のプロセッシングエレメントを３次元方向に有するデータ処理装置であって、それぞれ複数の前記プロセッシングエレメントが２次元方向に配置され、Ｚ軸方向に積層された複数の２次元面を有し、前記複数の２次元面のそれぞれに対応して特徴重みが配置されるフィルタメモリを有し、入力データがＺ軸方向の上位面の２次元面から配置され、一の面に配置されたプロセッシングエレメントで前記乗算機能により前記入力データと前記特徴重みの積を順次演算して２次元畳み込みデータを演算し、更に下面から転送されるデータと自身のデータを加算する演算を行い、当該演算結果を隣接する上面のプロセッシングエレメントに

転送することを特徴とする。
【００１８】
上記目的を達成する本発明に従う第１の側面において、第１の態様として、前記２次元方向に配置されたプロセッシングエレメントはトーラスネットワークに接続され、Ｚ軸方向には、上下面に隣接するＰＥが、ネットワークで双方向に接続されることを特徴とする。
【００１９】
上記目的を達成する本発明に従う第１の側面において、第２の態様として、前記２次元畳み込みデータは、隣接するプロセッシングエレメントからの転送データと自身のデータを加算演算し、更にシフト方向に隣接するプロセッシングエレメントに前記加算演算結果を転送することを特徴とする。
【００２０】
上記目的を達成する本発明に従う第１の側面において、第３の態様として、前記Ｚ軸方向の最上位面にあるプロセッシングエレメントは、下面の複数のプロセッシングエレメントから転送される２次元畳み込みデータと自身のデータを加算して２．５次元畳み込みデータを演算することを特徴とする。
【００２１】
上記目的を達成する本発明に従う第１の側面における第３の態様において、第４の態様として、前記２．５次元畳み込みデータは、順次下面のプロセッシングエレメントのシフトされることを特徴とする。
【００２２】
上記目的を達成する本発明に従う第１の側面における上記何れかの態様において、前記特徴重みは、前記入力データの配置された２次元面の数で分割され、前記入力データの配置された面毎に対応するフィルタメモリに配置し、前記畳み込み演算の際、前記フィルタメモリに配置された特徴重みを、対応する面の全てのプロセッシングエレメントにブロードキャストすることを特徴とする。
【００２３】
上記目的を達成する本発明に従う第２の側面は、乗算及び加算機能を有する複数のプロセッシングエレメントを３次元方向に有するデータ処理装置におけるデータ処理方法であって、前記データ処理装置は、それぞれ複数の前記プロセッシングエレメントが２次元方向に配置され、Ｚ軸方向に積層された複数の２次元面を有し、前記複数の２次元面のそれぞれに対応して特徴重みが配置されるフィルタメモリを有し、入力データを前記Ｚ軸方向の上位面の２次元面から配置する工程と、一の面に配置されたプロセッシングエレメントで前記乗算機能により前記入力データと前記特徴重みの積を順次演算して２次元畳み込みデータを演算する工程と、更に下面から転送されるデータと自身のデータを加算する演算を行い、当該演算結果を隣接する上面のプロセッシングエレメントに転送する工程を有することを特徴とする。
【００２４】
上記目的を達成する本発明に従う第２の側面において、第１の態様として、前記２次元畳み込みデータは、隣接するプロセッシングエレメントからの転送データと自身のデータを加算演算し、更にシフト方向に隣接するプロセッシングエレメントに前記加算演算結果を転送する工程を有することを特徴とする。
【００２５】
上記目的を達成する本発明に従う第２の側面において、第２の態様として、前記Ｚ軸方向の最上位面にあるプロセッシングエレメントは、下面の複数のプロセッシングエレメントから転送される２次元畳み込みデータと自身のデータを加算して２．５次元畳み込みデータを演算する工程を有することを特徴とする。
【００２６】
上記目的を達成する本発明に従う第２の側面において、第３の態様として、前記２．５次元畳み込みデータを、順次下面のプロセッシングエレメントにシフトする工程を有する

ことを特徴とする。
【発明の効果】
【００２７】
上記本発明に従う特徴構成により、処理されるデータと同数のプロセッサで並列演算を行うことで、最小実行ステップ数でＣＮＮの各層の計算を実行できる。このため、リアルタイム処理に求められる実行時間制約を達成できる最小の動作クロック周波数で実行可能であり、リアルタイムかつ低消費電力の計算を行える。
【図面の簡単な説明】
【００２８】
【図１】ＣＮＮのシステム構成の概略を示す図である。
【図２】TAP(Tensor Array Processor)における３次元アレイ状のPME(Processor in Memory)を示す図である。
【図３】各PMEの機能構成例ブロック図である。
【図４】TAPの３次元アレイをそれぞれのXY面に展開して示す図である。
【図５】図４における一つの面に属するPMEを拡大して示す図である。
【図６】畳み込み演算による演算結果の変化を示す図である。
【図７Ａ】本発明のデータ処理方法におけるある層の初期状態を表す図である。
【図７Ｂ】１番目のフィルタについて２次元畳み込みを行っている状態を表す図である。
【図７Ｃ】下の面から１番目の２次元畳み込み演算データが転送され、自身の畳み込み演算結果と加算することで２．５次元畳み込み演算の結果を求める処理を示す図である。
【図７Ｄ】１番目の２．５次元畳み込みの結果（白丸）を下の面にデータシフトした状態を表す図である。
【図７Ｅ】２番目のフィルタについて２次元畳み込みを行っている状態を表す図である。
【図７Ｆ】下の面から転送される２番目のフィルタに対する２次元畳み込み演算データと自身の畳み込み演算結果との加算により２．５次元畳み込み演算の結果を求める処理を示す図である。
【図７Ｇ】１番目と２番目の２．５次元畳み込みの結果（白丸）を下の面にデータシフトした状態を表す図である。
【図７Ｈ】３番目のフィルタについて２次元畳み込みを行っている状態を表す図である。
【図７Ｉ】下の面から転送される３番目のフィルタに対する２次元畳み込み演算データと自身の畳み込み演算結果と加算により２．５次元畳み込み演算の結果を求める処理を示す図である。
【図７Ｊ】１番目から３番目の２．５次元畳み込みの結果（白丸）を下の面にデータシフトした状態を表す図である。
【図７Ｋ】４番目のフィルタについて２次元畳み込みを行っている状態を表す図である。
【図７Ｌ】下の面から転送される４番目のフィルタに対する２次元畳み込み演算データと自身の畳み込み演算結果と加算により２．５次元畳み込み演算の結果を求める処理を示す図である。
【図７Ｍ】４枚のフィルタを用いた畳み込み層の計算結果を表す図である
【図８】畳み込み演算の例を具体的数値で説明する図である。
【図９】特徴重みの縦横を異なるものとした時のデータ上のシフト方向を考察する図である。
【図１０】本発明に従う２．５次元畳み込み演算の様子を示すタイムチャート図である。
【発明を実施するための形態】
【００２９】
以下に、本発明の実施例を添付の図面に従い説明する。これらの実施例は本発明の理解を容易とするためのものであり、本発明の適用は、これら実施例に限定されるものではない。また、本発明の保護の範囲は、特許請求の範囲と同一又は類似の範囲にも及ぶ。
【００３０】
本発明に従うデータ処理装置は、３次元アレイ状に配置されたそれぞれメモリ機能を有

する演算素子であるPME(Processor in Memory)とネットワークから構成されるシステムであり、以降TAP(Tensor Array Processor)と称する。

【００３１】

図２は、かかるTAPにおける３次元アレイ状のPMEを示す図であり、それぞれ計算モジュールを有する複数のPMEが３次元（X,Y,Z）方向に積層配列されている。

【００３２】

かかる構造は、半導体技術により、３次元プロセッサとして作成可能である。すなわち、複数のPMEがX,Y方向に配列された２次元半導体面をZ方向に積み重ねて１チップで３次元構造とすることが可能である。

【００３３】

図２において、複数のPMEが、Ws×Hs×Csの３次元アレイ状（X軸方向にWs個,Y軸方向にHs個,Z軸方向にCz個）に配列されている。各XY面では、各PMEがトーラスネットワーク（Ｌｘ、Ly）に接続され、各PMEの有するデータをX軸正方向、X軸負方向、Y軸正方向、Y軸負方向の4方向にデータシフトする機能を有する。

【００３４】

また、Z軸方向では、上下面に隣接するPMEが、ネットワークLzで双方向に接続され、Z軸正方向（上面方向）とZ軸負方向（下面方向）のデータ転送を同時に行うことが可能に構成されている。かかるデータシフトの方向制御及び、そのための共通シフトクロックは、後にデータ処理装置を展開図で示す制御プロセッサにより供給される。

【００３５】

図３は、各PMEの機能構成例ブロック図であり、畳み込み層の計算に必要な乗算及び加算機能ブロック１０とプーリング層の計算に必要なMAX（最大値）演算機能ブロック１１を有している。さらに、必要に応じて、追加の機能ブロックを添えることは可能である。

【００３６】

図４は、上記TAPの３次元アレイをそれぞれのXY面に展開して示す図である。かかるTAPは、システムとして共通の制御プロセッサ２０と、指示メモリ２１を有している。

【００３７】

３次元アレイのそれぞれのXY面（Z=1, Z=2,・・・Z=S）に対応してフィルタメモリFM1-FMSを有し、フィルタメモリFM1-FMSのそれぞれは、対応する同一面に存在する全てのPMEと接続されている。計算の際に、フィルタメモリFM1-FMSからフィルタの特徴重みを対応する面の全てのPMEにブロードキャストすることが可能である。

【００３８】

ここで、本発明に従うデータ処理方法をＣＮＮの畳み込み演算処理に用いる場合を想定する。

【００３９】

指示メモリ２１には、事前の学習により得られた各層のフィルタサイズ及びフィルタ数が畳み込みニューラルネットワーク構造として入力される。これに基づき、制御プロセッサ２０により、各面のフィルタメモリFM1-FMSに対応する重み、及び共通のクロック信号の供給等が行われる。

【００４０】

PMEは、畳み込み層において、自身の有するデータと対応するフィルタメモリからブロードキャストされる重みとを乗算し、更に隣接するPMEから転送されるデータとを加算して、その結果を反対側に隣接するPMEに転送する。プーリング層においては、自身のデータと隣接するPMEから転送されるデータの最大値を求め、隣接するPMEに順次転送する。

【００４１】

先に、説明した様に指示メモリ２１からの予め学習によって得られた指示データに基づき、上記のPMEによる演算と転送の方向及び共通シフトのためのタイミングクロックが、制御プロセッサ２０から全てのPMEに送られる。

【００４２】

図５は、かかる図４における一つの面に属するPMEを拡大して示す図である。この例で

は、Ws=４， Hs=３の場合で、Z=kの面を示している。ファイルメモリFMkから共通に、Z=kの面にある全てのPMEにフィルタ（特徴重み）が供給される。

【００４３】

かかる構成のTAPに、列数W_o×行数H_o×チャンネル数C_oの３次元テンソルデータがＣＮＮの第１層の入力となる。

【００４４】

この入力データに繰り返しＣＮＮの畳み込み層の計算を適用することで、各層のデータサイズが変化する。

【００４５】

ここで、k層目のデータサイズを$W_k \times H_k \times C_k$とする。本発明のシステムでは、

【００４６】

【数１】

$$Ws \geq \max_k W_k, \quad Hs \geq \max_k H_k, \quad Cs \geq \max_k C_k$$

【００４７】

であると想定する。

【００４８】

入力のチャネル数がこの値より小さい場合は、TAPの上位のXY面から順に入力データを配置する。入力データが配置されたPMEは活性面（active）、そうでなければ非活性面（inactive）となる。ただし、PMEが非活性面であってもデータの転送は行われる。

【００４９】

例えば、入力データとして一枚の画像データを考えた時、次のように想定することが出来る。一枚の画像データを同じ大きさの領域ごとに区切り複数の領域データ（チャネル）として切り出し、各領域データをTAPの最上位の面から順に該当の面にあるPMEに配置していく。

【００５０】

この時、フィルタ（特徴重み）は、次のように処理される。一つの特徴重みを前記画像データの配置される面の数に対応して分割し、それぞれの分割特徴重みを対応する面のファイルメモリFMに格納する。そして、計算時に対応するフィルタメモリFMに配置されている特徴重みが当該面に属する全てのPMEにブロードキャストされる。

【００５１】

それぞれのPMEは、ブロードキャストされた特徴重みと自身のデータとの積を演算する。さらに、PMEは、一方向の隣接するメモリ要素から転送されるデータを前記の積の演算結果に加え、反対方向に隣接するPMEに転送する。かかる処理を繰り返し、２次元畳み込みを行う。なお、かかる場合の転送制御は、先に述べた特許文献１の発明に従い実行される。

【００５２】

さらに、本発明では、特徴として、入力データに対し２次元畳み込みを行ったデータが配置されたTAPの最下面から最上面まで、それぞれ２次元畳み込みデータを上方向に転送する。この時、各面のPMEは自身のデータと一つ下の面からの畳み込み演算結果を足し合わせ、その結果を一つ上の面のPMEに転送する。

【００５３】

最終的に最上位面の２次元アレイプロセッサで、全ての２次元アレイの２次元畳み込み結果を足し合わせた２．５次元畳み込み演算結果を得ることが出来る。

【００５４】

さらに、後に詳述するように、最上位面で得られた２．５次元畳み込み演算結果は、順次下面にシフトされる。

【００５５】

かかる畳み込み演算による演算結果の変化を図６に示す。図６（１）に示すように、N

×M×Cin個のPMEに配置された入力データが、畳み込み演算の結果N×M×Cout個のPMEに畳み込み演算結果が得られる。このときCoutの大きさは、特徴重み(kernel)の数に依存する。

【００５６】

この結果がＣＮＮの一つの層の畳み込み演算処理結果のデータであり、次いで、図６（２）に示すようにプーリング層の演算処理を行ってレイヤーを縮小して扱いやすくする。同時に、この演算結果は次の層の入力になる。

【００５７】

ここで、本発明のデータ処理装置において実行されるデータ処理に従う畳み込み演算処理の特徴を理解容易のために、更に図７Ａ～７Ｍにおいて各面における変化を模式的に示す。

【００５８】

図７Ａ～７Ｍにおいて、活性面は入力データが配置された面である。図７Ａは、初期状態を表す。実戦の直方体が活性面のPMEを表し、破線の直方体が付加性面のPMEを表す。以下、図７Ｂ～７Ｍにおいて同様である。さらに、灰色の丸で占められる表示は入力データDinを表している。図７Ｂは１番目のフィルタに対して各面において２次元畳み込みを行っている状態を示す。図の矢印は、各面にあるPMEに対するデータの転送方向を表し、黒丸は計算結果を示す。以下、図７Ｃ～７Ｍにおいて同様である。

【００５９】

図７Ｃは、下の面からその２次元畳み込み演算データが転送され、自身の２次元畳み込み演算結果と加算することで２．５次元畳み込み演算の結果を求める処理を示している。

【００６０】

図７Ｄは、２番目のフィルタに対する畳み込み計算を行う準備として、この２．５次元畳み込み演算結果を下の面にシフトする状態を示している。このシフトは、不活性面を含めて行われる。

【００６１】

図７Ｅは２番目のフィルタに対して各面において２次元畳み込みを行っている状態を示す。図７Ｆは下の面から２番目のフィルタに対する２次元畳み込み演算データが転送され、自身の畳み込み演算結果と加算することで、２．５次元畳み込み演算の結果を求める処理を示している。

【００６２】

同様に図７Ｇ～図７Ｌは３番目、４番目のフィルタに対する畳み込み計算の様子を示している。すなわち、図７Ｇは、１番目と２番目の２．５次元畳み込みの結果(白丸)を矢印のように上の面から下の面にデータシフトした状態を示している。

【００６３】

図７Ｈは、３番目のフィルタについて２次元畳み込みを行っている状態を示している。図７Ｉは、下から３番目のフィルタに対する２次元畳み込み演算データが転送され、自身の畳み込み演算結果と加算することで２．５次元畳み込み演算の結果を求める処理を示している。

【００６４】

図７Ｊは、１番目から３番目の２．５次元畳み込み結果（白丸）を上の面から下の面にデータシフトした状態を示している。この際、不活性面にもデータがシフトされている。

【００６５】

図７Ｋは、４番目のフィルタに対する２次元畳み込み演算を行っている状態を示している。図７Ｌは、図７Ｉの処理と同様であるが、下から４番目のフィルタに対する２次元畳み込み演算データが転送され、自身の畳み込み演算結果と加算することで２．５次元畳み込み演算お結果を求める処理を示している。

【００６６】

図７Ｍは、最終的に４枚のフィルタを用いた畳み込み層の計算結果を表し、これが次層の入力となる。

【００６７】
ここで、畳み込み演算を式で表すと下記(１)式のようになる。b_0はバイアス定数項である。バイアスb_0は、畳み込み演算の結果を一定値増加、減少するために使用される。このバイアスb_0とフィルタの重みωはともに、ＣＮＮの学習時に自動的に決定される。
【００６８】
【数２】

$$y_o^{(l)}(i,j) = b_o + \sum_{c=1}^{C^l} \sum_{\Delta i=-\lfloor\frac{\omega}{2}\rfloor}^{\lfloor\frac{\omega}{2}\rfloor} \sum_{\Delta j=-\lfloor\frac{\omega}{2}\rfloor}^{\lfloor\frac{\omega}{2}\rfloor} w_{o,c}^{(l)}(\Delta i,\Delta j)\cdot x_c^{(l)}(is-\Delta i,\ js-\Delta j), \quad \cdots(1)$$

【００６９】
ただし、sは自然数であり、畳み込み計算を行うときのストライドを表す。さらに、簡単化のため、本発明の説明ではストライドが１のときのみを説明しているが、ストライドが２以上であっても本発明の適用可能は、否定されない。
【００７０】
【数３】

$$w_{o,c}^{(l)}(\Delta i,\Delta j)$$

【００７１】
は、第ｌ層のo番目のフィルタの重み、
【００７２】
【数４】

$$x_c^{(l)}(is-\Delta i,\ js-\Delta j)$$

【００７３】
は、第ｌ層の入力データである。
【００７４】
それぞれの面にあるPMEは、（１）式の後半部分
【００７５】
【数５】

$$\sum_{\Delta i=-\lfloor\frac{\omega}{2}\rfloor}^{\lfloor\frac{\omega}{2}\rfloor} \sum_{\Delta j=-\lfloor\frac{\omega}{2}\rfloor}^{\lfloor\frac{\omega}{2}\rfloor} w_{o,c}^{(l)}(\Delta i,\Delta j)\cdot x_c^{(l)}(is-\Delta i,\ js-\Delta j),$$

【００７６】
の計算を行う。このとき、$C=C^l$の２次元畳み込み演算をTAPの一番上の面のPMEが計算しており、同様に$C=C^l-1$の２次元畳み込演算をその一つ下の面のPMEが計算している。各面で計算した上記の後半部分の計算結果を足し合わせることで（１）式全体の計算をしている。
【００７７】
図８は、上記の畳み込み演算の例を具体的数値で説明する図であり、２つの上下面の場合を例にしている。
【００７８】
一の面（Ch1）で入力（Input）ｘ６０と重み（kernel）ｗ６１を矢印方向に移動しながら乗算し、同時に下の面（Ch2）で入力（Input）ｘ６２と重み（kernel）ｗ６３を矢印方向に移動しながら乗算する。これにより、それぞれ２次元畳み込み演算結果６４が得られる。

【００７９】
一の面（Ch1）の初期時点での入力６０と重み６１との乗算結果は、次のようであり、
（－３＊１）＋（－２＊２）＋（１＊２）＋（３＊２）＝１
次いで、1桁分矢印方向にシフトした時の入力６０と重み６１との乗算結果は、次のようである。
【００８０】
（１＊１）＋（－３＊２）＋（－２＊３）＋（２＊２）＋（１＊２）＋（３＊１）＝－２
これらは、２次元畳み込み演算結果６４に示される通りである。
【００８１】
一方、下の面（Ch2）の初期時点での入力６２と重み６３との乗算結果は、次のようであり。
【００８２】
（－２＊２）＋（３＊３）＋（－３＊１）＋（１＊３）＝５
次いで、1桁分矢印方向にシフトした時の入力６０と重み６１との乗算結果は、次のようである。
【００８３】
（２＊２）＋（－２＊３）＋（３＊３）＋（２＊１）＋（３＊３）＋（１＊２）＝１である。
【００８４】
これらは、２次元畳み込み演算結果６５に示される通りである。
【００８５】
ついで、前記一の面（Ch1）では、自身の２次元畳み込み演算結果６４を得て、更に下の面（Ch2）から転送される２次元畳み込み演算結果６５が転送される。したがって、それら２次元畳み込み演算結果６４及び６５とバイアスb_0＝１とを加算して２．５次元畳み込み演算結果６６に示すように求める。
【００８６】
上記の様に、入力データ上で重みを順次所定桁数分ずつシフトして乗算及び加算を繰り返すことにより２次元畳み込み演算結果が得られる。
【００８７】
この際、指示メモリ２１に格納されている指示に基づき、PMEからのデータ転送の方向がデータを一筆書きに転送し、無駄な転送をなくし、同じPMEに複数のデータ転送が行われないように制御され、データの衝突を回避することが出来る。
【００８８】
ここで、上記図８に示す例では特徴重みを縦横３×３、即ち縦横の長さが同じｗ×ｗとしているが、縦横の長さが異なる様に一般化することが出来、これをｗ_1×ｗ_2として表す。
【００８９】
図９は、特徴重みの縦横を異なるものとした時のデータ上のシフト方向を考察する図である。図９（１）は、縦横長さが同じ奇数で、中心に向かう様に一筆書きでシフトすることが出来る。図９（２）は、縦の長さ、横の長さのいずれか一方が偶数であり、図９（１）と同様に、全ての点をちょうど１回ずつ通ハミルトンパスが存在する。これに対し、図９（３）は、縦の長さも横の長さも奇数の場合で有り、１個のPMEは２回通過することになるので、無駄な転送が発生する。
【００９０】
ここで、上記説明したように２．５次元の畳み込み演算の結果が得られるが、このデータはTAPの一番上のPMEが保有している。本発明に従うアルゴリズムでは、TAPの最下面をZ=1，最上面を$Z=C_S$とすると、l層目で0番目のフィルタを用いたときの２．５次元畳み込み演算の結果
【００９１】

【数６】

$$y_0^{(l)}(i,j)$$

を

【数７】

$$Z = C_s - C^{l+1} + 0$$

【００９２】

の面状のPMEに配置し、次の層の畳み込みの計算の準備に整える。このため、最上面で計算結果が得られる度に、各PMEが保有している２．５次元畳み込み演算結果を下面方向に１回シフトする(図７Ｂ参照)。

【００９３】

図１０は、更に本発明に従う２．５次元畳み込み演算の様子を示すタイムチャート図である。このタイムチャートでは、PME(i,j,1), PME(i,j,2),…PME(i,j,Cs)の動作を表している。また、この図ではCin個の面がactiveである。

【００９４】

タイムチャートにおいて、各面の墨塗り部分Ａで２次元畳み込み演算を行っている。この計算結果が終了した次のステップでその計算結果とbiasBを足し合わせて上の面のPMEにデータを転送する(上方向矢印)。

【００９５】

次の面のPMEは、下の面のPMEから転送されたデータと自身の２次元畳み込み演算結果を足し合わせて,その結果Ｃを更に一つ上の面のPMEに転送することを繰り返す。最終的に一番上の面のPMEでの演算結果が、２．５次元畳み込みの演算結果Ｄとなる。

【００９６】

次いで、この、２．５次元畳み込みの演算結果Ｄが、一つの重みについて２．５次元畳み込み演算が終わる都度、下向矢印の方向に下の面にシフトされる。この際、上の面からシフトされるデータはPMEでは、それを保存するだけで、その他の処理は行われない。２．５次元畳み込演算を求めるために、一度だけシフトを行う。

【００９７】

上記の動作をCout回繰り返し、一番上の面からCoutまでに２．５次元畳み込み演算結果が保持される。

【００９８】

ここで、ＣＮＮの各層において、上記したように２．５次元畳み込み処理が行われた後、プーリング（pooling）演算を行なって、次の層の入力データとされる。

【００９９】

プーリング演算は、２次元畳み込み演算の時と同じ方法で周辺のPMEが持つデータを受け取り次の式（２）で最大値を計算する。

【０１００】

【数８】

$$x_0^{(l+1)}(i,j) = \max_{\substack{-\left\lfloor\frac{\omega'}{2}\right\rfloor \le \Delta i \le \left\lfloor\frac{\omega'}{2}\right\rfloor, \\ -\left\lfloor\frac{\omega'}{2}\right\rfloor \le \Delta i \le \left\lfloor\frac{\omega'}{2}\right\rfloor}} \left\{ act\left(y_0^{(l)}(is' + \Delta i, js' + \Delta j)\right)\right\} \quad \cdots (2)$$

【０１０１】

ただし、s'は２以上の自然数であり、畳み込み計算を行うときのストライドを表す。さらに、actは活性化関数であり、例えば

【０１０２】

【数９】

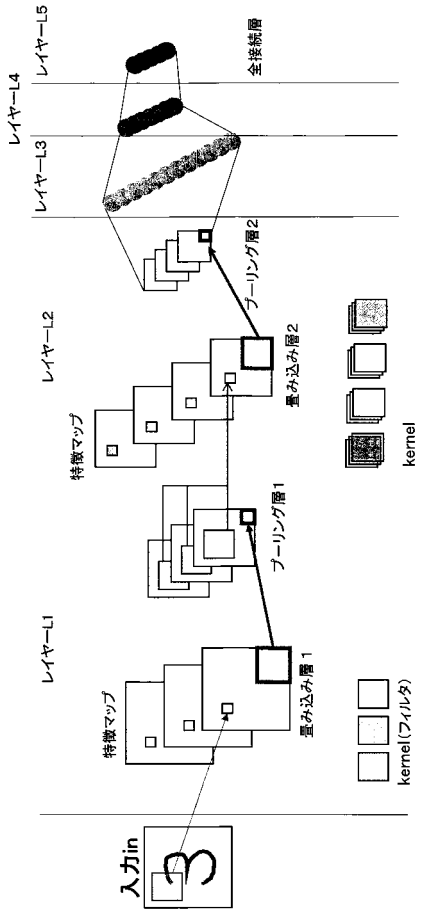
$$act(\cdot) = \max(0,\cdot)$$

【０１０３】
が用いられる。
【０１０４】
また、プーリングでは、２次元畳み込みと同様に上面側にデータ転送を行うが、各PMEは、自身の有するデータxと隣接した下面のPMEから受け取ったy_{in}を用いて
【０１０５】
【数１０】

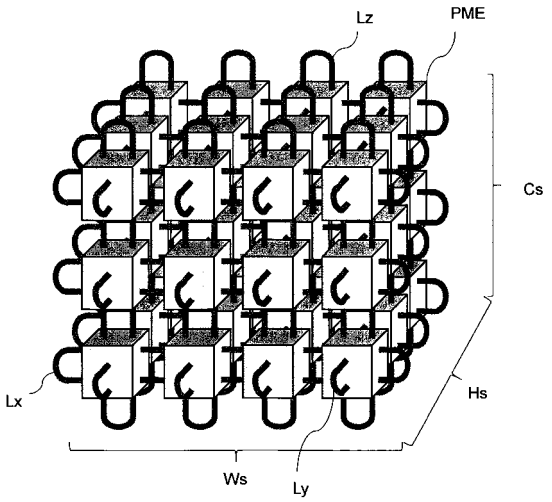
$$y_{out} = \max(y_{in}, x) \quad \cdots \quad (3)$$

【０１０６】
を計算する。
【０１０７】
この計算を行いながら先に、図９で説明した様にデータを転送することにより周辺のPMEの持つデータの最大値を求める。
【０１０８】
以上説明したように、本発明に従うデータ処理装置は、ＣＮＮにおけるデータ処理装置として使用される場合は、ＣＮＮの構造（学習によって得られた各層のフィルタサイズ、フィルタ数）が入力として与えられる。各フィルタの重みがTAPの各面のフィルタメモリ上に与えられる。さらに、ＣＮＮの主入力データがTAPのPMEに配置される。
【０１０９】
各面のPMEは、自身のデータと重みを乗算して２次元畳み込みデータを、周辺から転送されるデータとを加算して上位の面上のPMEに送る。したがって、最上位の面にあるPMEで、全ての下面の２次元畳み込み演算結果加算することにより並列性を保ったまま省力で高速の２．５次元畳み込み演算結果を得ることが出来る。
【符号の説明】
【０１１０】
PME　メモリ要素
１０　乗算及び加算機能ブロック
１１　MAX（最大値）演算機能ブロック
２０　制御プロセッサ
２１　指示メモリ
FM1-FMS　フィルタメモリ

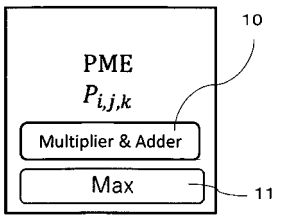
【図１】


レイヤーL5
レイヤーL4
レイヤーL3
レイヤーL2
レイヤーL1
全接続層
プーリング層2
畳み込み層2
特徴マップ
kernel
プーリング層1
畳み込み層1
特徴マップ
kernel(フィルタ)
入力in

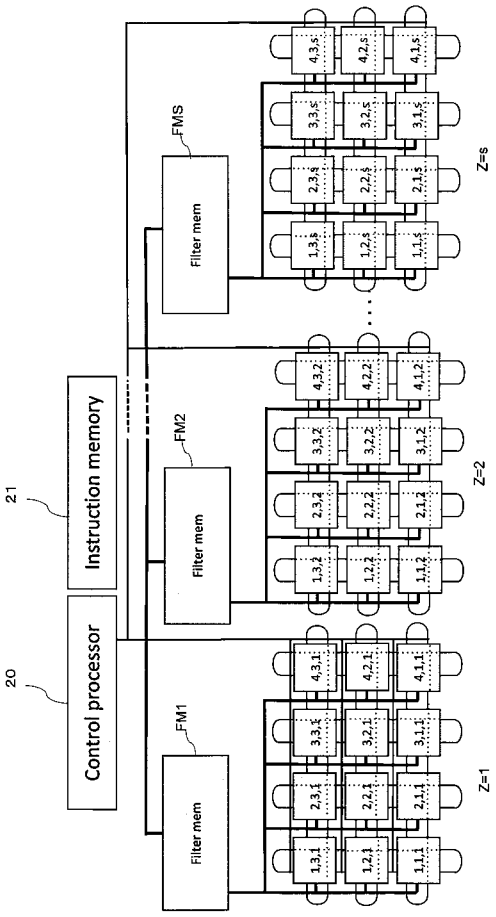
【図２】


Lz
PME
Cs
Lx
Hs
Ws
Ly

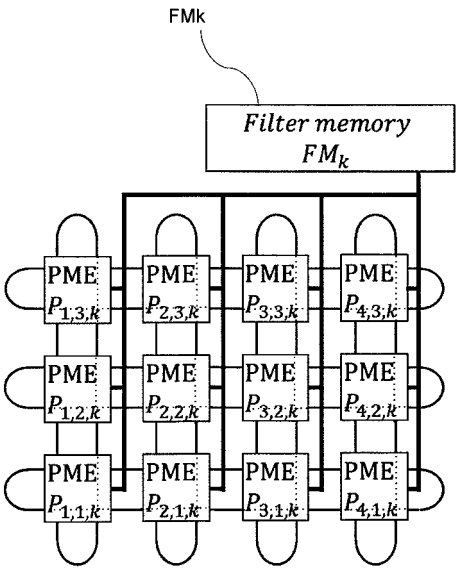
【図３】


PME
$P_{i,j,k}$
Multiplier & Adder
Max
10
11

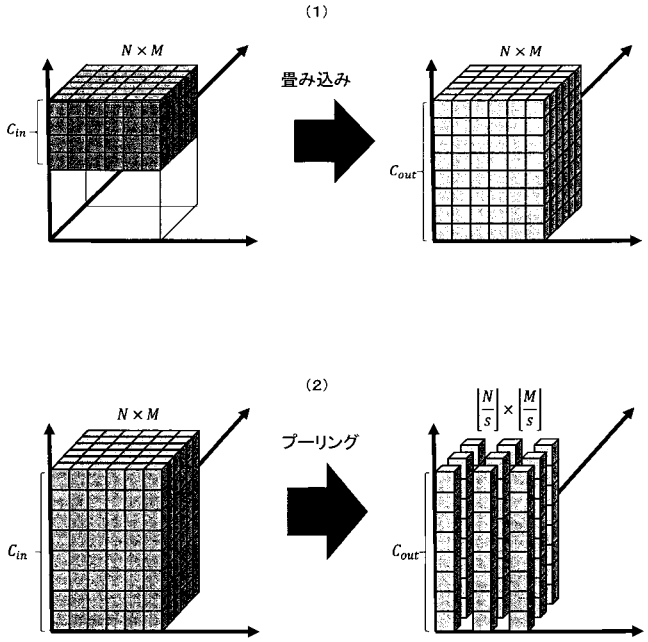
【図４】


Instruction memory
Control processor
21
20
FMS
FM2
FM1
Filter mem
Filter mem
Filter mem
Z=s
Z=2
Z=1

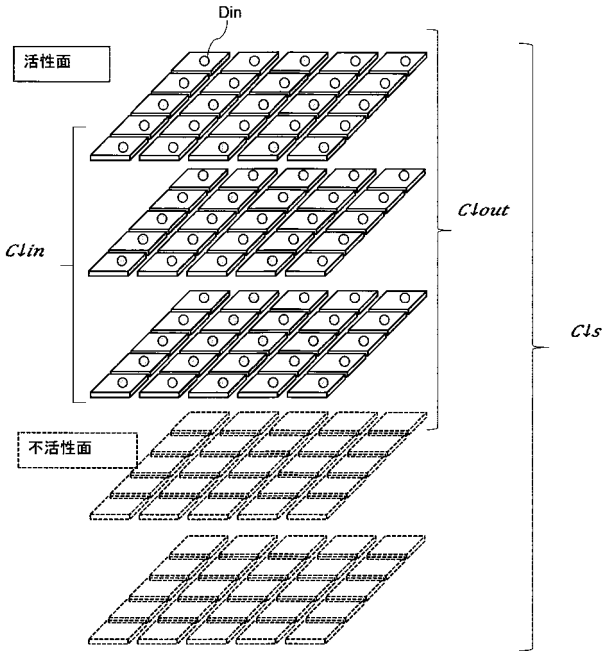
【図５】


FMk
Filter memory
FM_k
PME $P_{1,3,k}$
PME $P_{2,3,k}$
PME $P_{3,3,k}$
PME $P_{4,3,k}$
PME $P_{1,2,k}$
PME $P_{2,2,k}$
PME $P_{3,2,k}$
PME $P_{4,2,k}$
PME $P_{1,1,k}$
PME $P_{2,1,k}$
PME $P_{3,1,k}$
PME $P_{4,1,k}$

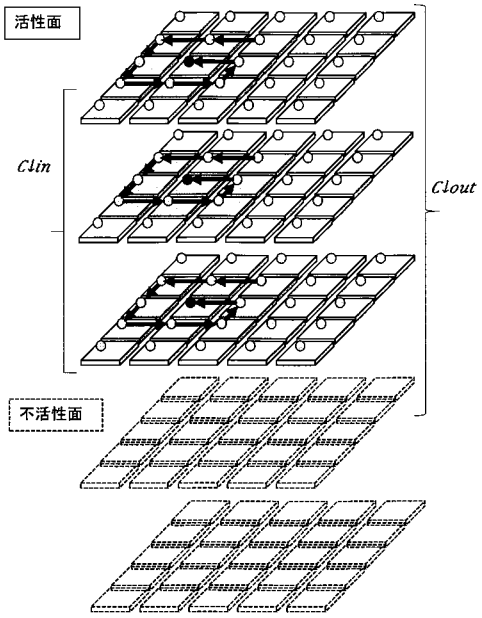
【図６】



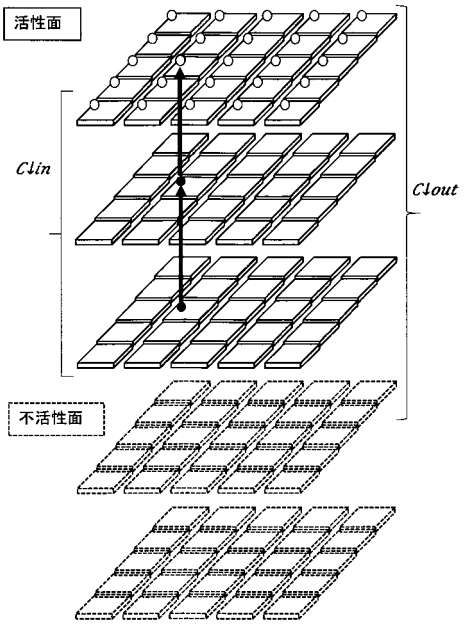
【図７Ａ】



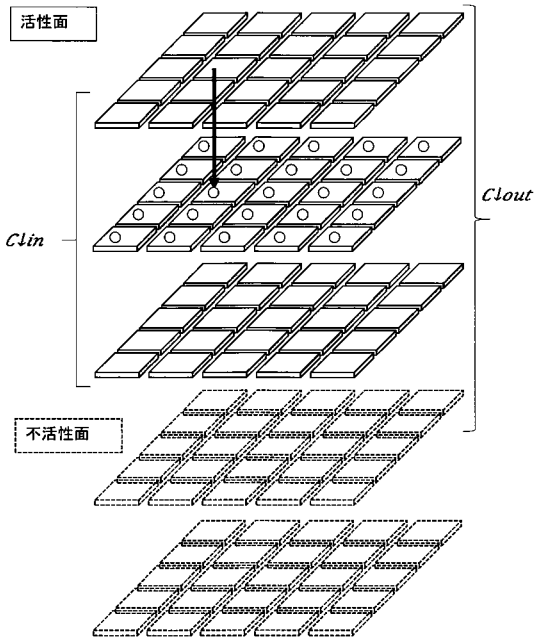
【図７Ｂ】



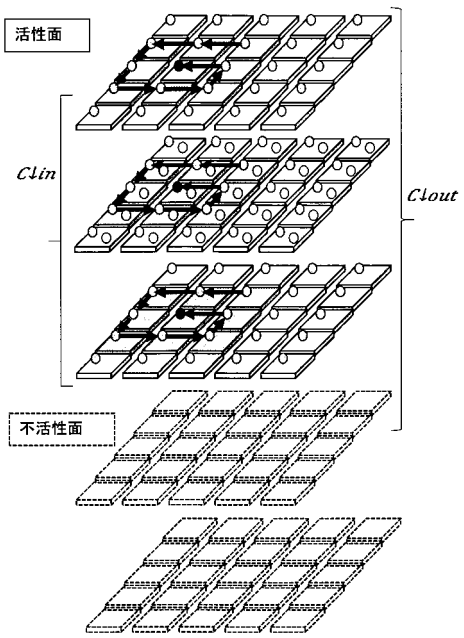
【図７Ｃ】



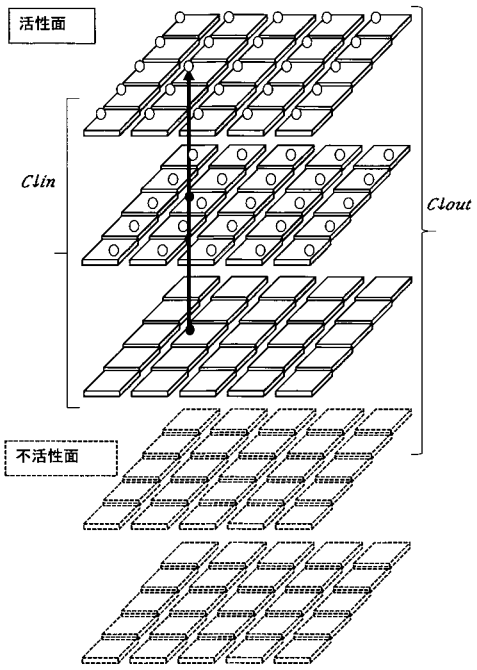
【図７Ｄ】



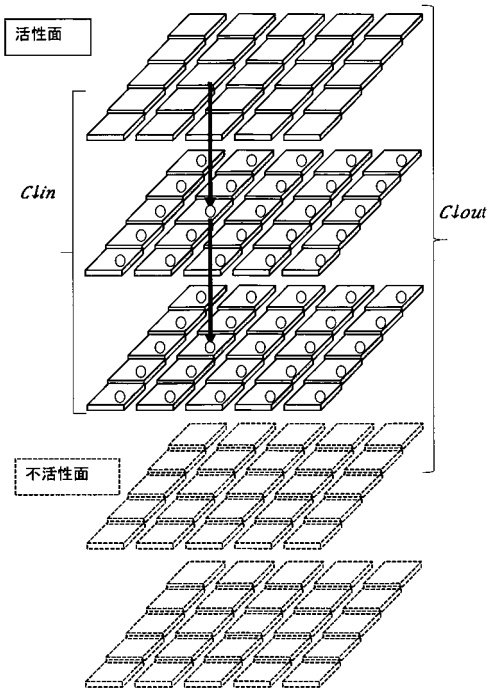
【図７Ｅ】



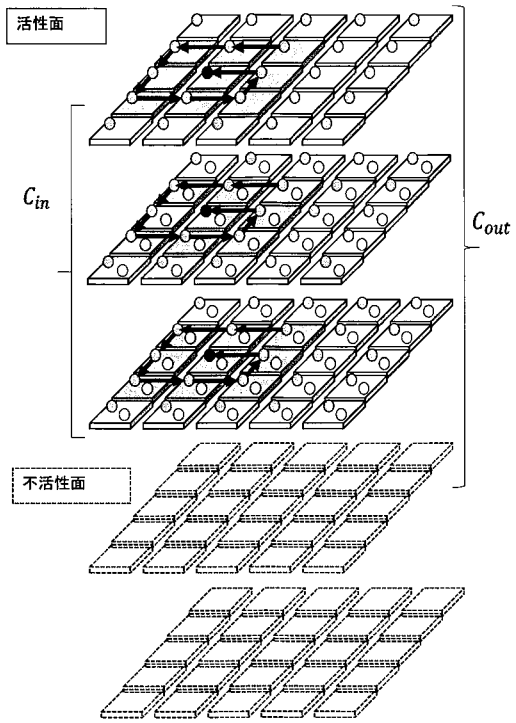
【図７Ｆ】



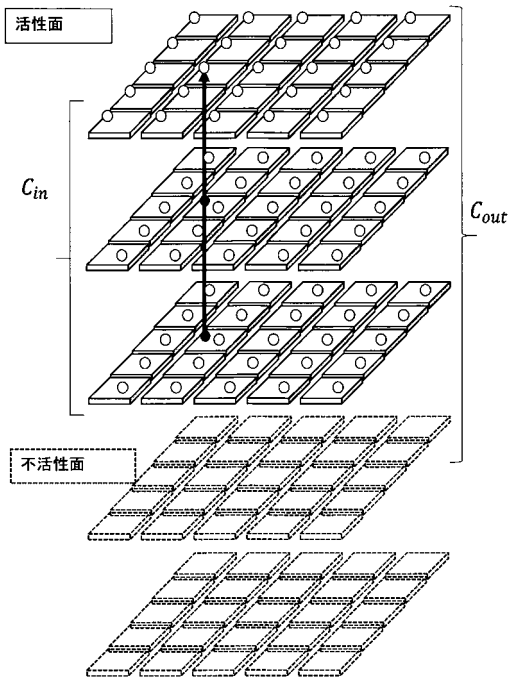
【図７Ｇ】



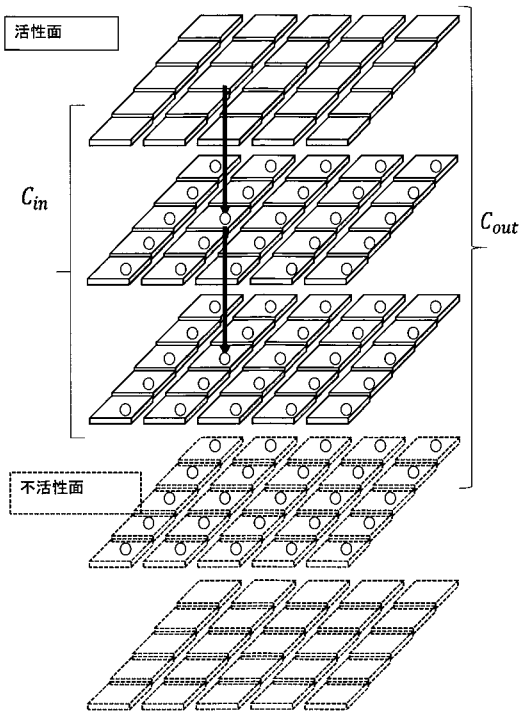
【図７Ｈ】



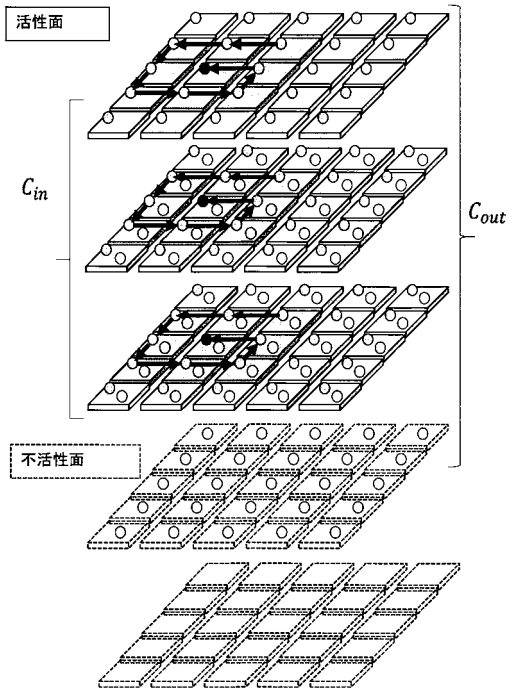
【図７Ｉ】



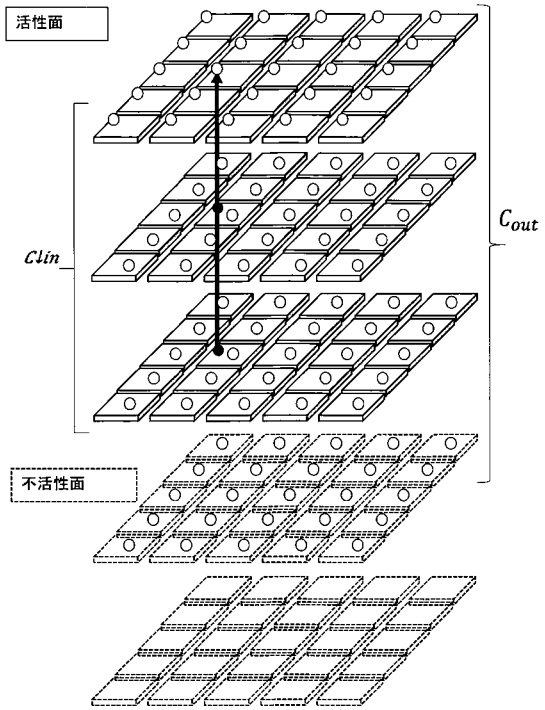
【図７Ｊ】



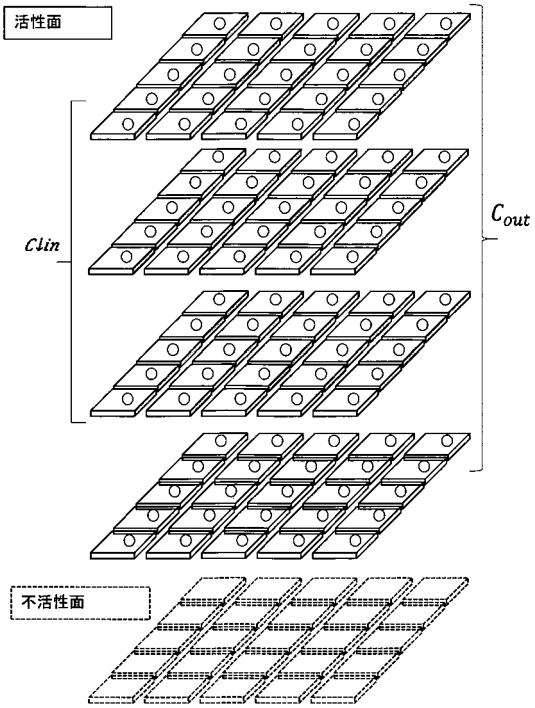
【図７Ｋ】



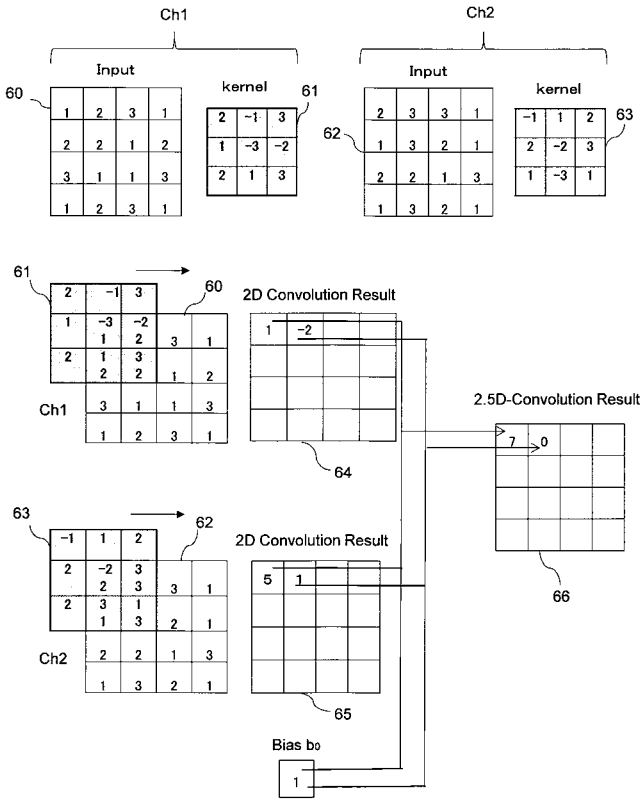
【図７Ｌ】


活性面
C_{in}
C_{out}
不活性面

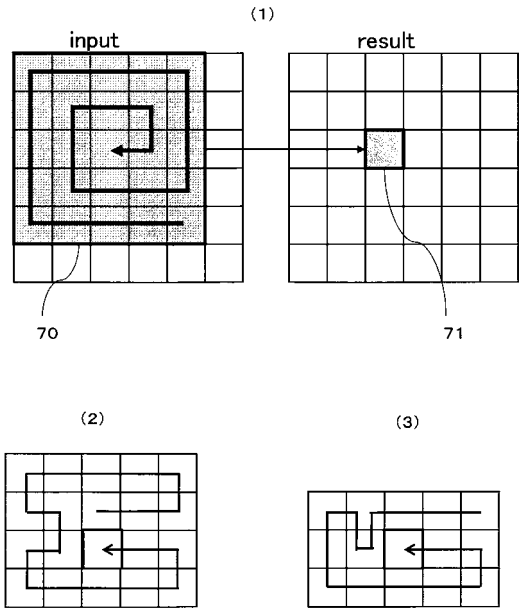
【図７Ｍ】


活性面
C_{in}
C_{out}
不活性面

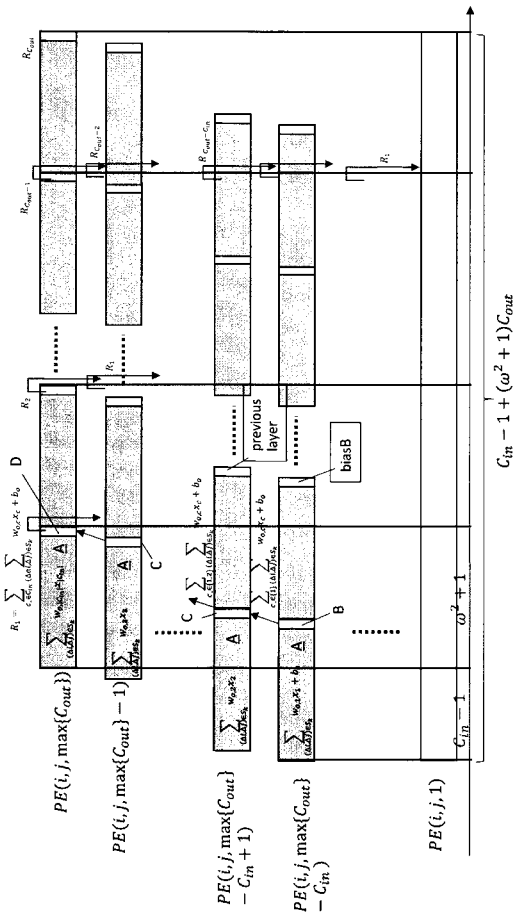
【図８】


Ch1
Input
kernel
60
61
Ch2
Input
kernel
62
63
2D Convolution Result
64
2.5D-Convolution Result
66
2D Convolution Result
65
Bias b0

【図９】


(1)
input
result
70
71
(2)
(3)

【図１０】


previous layer
biasB
A
B
C
D
$PE(i,j,\max\{C_{out}\})$
$PE(i,j,\max\{C_{out}\}-1)$
$PE(i,j,\max\{C_{out}\}-C_{in}+1)$
$PE(i,j,\max\{C_{out}\}-C_{in})$
$PE(i,j,1)$
$C_{in}-1$
ω^2+1
$C_{in}-1+(\omega^2+1)C_{out}$