

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号  
特許第7654013号  
(P7654013)

(45)発行日 令和7年3月31日(2025.3.31)

(24)登録日 令和7年3月21日(2025.3.21)

(51)国際特許分類

FI

G 0 6 N 3/065(2023.01) G 0 6 N 3/065

G 0 6 N 3/0464(2023.01) G 0 6 N 3/0464

請求項の数 18 (全27頁)

|                   |                             |          |                      |
|-------------------|-----------------------------|----------|----------------------|
| (21)出願番号          | 特願2022-568491(P2022-568491) | (73)特許権者 | 390009531            |
| (86)(22)出願日       | 令和3年5月13日(2021.5.13)        |          | インターナショナル・ビジネス・マシー   |
| (65)公表番号          | 特表2023-526915(P2023-526915  |          | ンズ・コーポレーション          |
|                   | A)                          |          | INTERNATIONAL BUSI   |
| (43)公表日           | 令和5年6月26日(2023.6.26)        |          | NESS MACHINES CORPO  |
| (86)国際出願番号        | PCT/IB2021/054105           |          | RATION               |
| (87)国際公開番号        | WO2021/240286               |          | アメリカ合衆国10504 ニューヨー   |
| (87)国際公開日         | 令和3年12月2日(2021.12.2)        |          | ク州 アーモンク ニュー オーチャード  |
| 審査請求日             | 令和5年10月26日(2023.10.26)      |          | ロード                  |
| (31)優先権主張番号       | 16/884,128                  |          | New Orchard Road, A  |
| (32)優先日           | 令和2年5月27日(2020.5.27)        |          | rmonk, New York 105  |
| (33)優先権主張国・地域又は機関 | 米国(US)                      | (74)代理人  | 04, United States of |
|                   |                             |          | America              |
|                   |                             |          | 100112690            |
|                   |                             |          | 弁理士 太佐 種一            |
|                   |                             |          | 最終頁に続く               |

(54)【発明の名称】 アナログ人工知能ネットワーク推論のための行単位畳み込みニューラル・ネットワーク・マッピングのための効率的なタイル・マッピング

(57)【特許請求の範囲】

【請求項1】

クロスポイント・アレイを使用して畳み込みニューラル・ネットワーク（CNN）を実装するためのコンピュータ実装方法であって、

前記クロスポイント・アレイの1つまたは複数のクロスポイント・デバイス内に前記CNN内の畳み込み層の1つまたは複数の畳み込みカーネルを記憶することによって、前記クロスポイント・アレイを構成することであって、前記クロスポイント・アレイが、前記畳み込み層に対応する、前記構成することと、

所定の回数にわたって操作のセットを反復することによって、前記クロスポイント・アレイを介して前記CNNについての計算を実施することであって、操作の前記セットが、

前記畳み込み層の入力データのベクトルの副部分に対応する電圧パルスを前記クロスポイント・アレイに送ることと、

前記クロスポイント・アレイ内の前記1つまたは複数のクロスポイント・デバイスで乗算演算を実施することを表す電流を出力することであって、前記電流が、前記クロスポイント・デバイスによって記憶された重み値と、前記入力データからの前記電圧パルスとに基づく、前記出力することと、

積分器のセットによって、前記クロスポイント・デバイスからの出力電流に基づく電荷を蓄積することと

を含む、前記実施することと、

前記所定の回数にわたって反復した後に、積分器の前記セットによって、蓄積した電荷

を出力することであって、前記蓄積した電荷が、入力データの前記ベクトルと前記1つまたは複数の畳み込みカーネルとの乗算加算結果を表す、前記出力することを含む方法。

【請求項2】

積分器の前記セット内の前記蓄積した電荷を出力することが、前記蓄積した電荷をプールすることを含む、請求項1に記載の方法。

【請求項3】

入力データの各ベクトルの前記副部分が積分器の前記セットに関連付けられる、請求項1に記載の方法。

【請求項4】

前記クロスポイント・アレイが複数のクロスポイント・アレイであり、入力データの前記ベクトルの第1の副部分が第1のクロスポイント・アレイに送られ、入力データの前記ベクトルの第2の副部分が第2のクロスポイント・アレイに送られる、請求項1に記載の方法。

【請求項5】

積分器の前記セットによって前記電荷を蓄積することが、前記第1のクロスポイント・アレイの積分器の前記セットにより、前記第2のクロスポイント・アレイの積分器の前記セットによって蓄積される前記電荷を蓄積することを含む、請求項4に記載の方法。

【請求項6】

前記クロスポイント・デバイスが、前記CNNの所与の層の畳み込みカーネルの1つまたは複数の列を実装するように構成され、入力データの前記ベクトルが、前記入力データから一度に1行ずつ提示される前記CNNの前記所与の層に対するニューロン励起を表す、請求項1に記載の方法。

【請求項7】

積分器の前記セットのうちのある積分器によって蓄積された前記電荷が、前記CNNの前記所与の層に従う出力励起を表し、前記出力励起が、前記畳み込みカーネルのすべての行が統合された後にのみ、変換され、送られる、請求項6に記載の方法。

【請求項8】

前記クロスポイント・デバイスが、前記CNNの所与の層の畳み込みカーネルの1つまたは複数の行を実装するように構成され、前記入力データが、一度に1列ずつ提示される前記CNNの前記層に対するニューロン励起を表す、請求項1に記載の方法。

【請求項9】

積分器の前記セットのうちのある積分器によって蓄積された前記電荷が、前記CNNの前記所与の層に従う出力励起を表し、前記出力励起が、前記畳み込みカーネルのすべての列が統合された後にのみ、変換され、送られる、請求項8に記載の方法。

【請求項10】

トレーニング済み畳み込みニューラル・ネットワーク(CNN)の計算を実施するための電子回路であって、

クロスポイント・アレイと、

積分器のセットを備える出力回路と、

前記クロスポイント・アレイの1つまたは複数のクロスポイント・デバイス内に前記CNN内の畳み込み層の1つまたは複数の畳み込みカーネルを記憶することによって、前記畳み込み層に対応する前記クロスポイント・アレイを構成する回路と、

所定の回数にわたって操作のセットを反復する回路であって、操作の前記セットが、

前記畳み込み層の入力データのベクトルの副部分に対応する電圧パルスを前記クロスポイント・アレイに送ることと、

前記クロスポイント・アレイ内の前記1つまたは複数のクロスポイント・デバイスで乗算演算を実施することを表す電流を出力することであって、前記電流が、前記クロスポイント・デバイスによって記憶された重み値と、前記入力データからの前記電圧パルスとに基づく、前記出力することと、

10

20

30

40

50

積分器の前記セットによって、前記クロスポイント・デバイスからの出力電流に基づく電荷を蓄積することと、

前記所定の回数にわたって反復した後に、積分器の前記セットによって、蓄積した電荷を出力することであって、前記蓄積した電荷が、入力データの前記ベクトルと前記1つまたは複数の畳み込みカーネルとの乗算加算結果を表す、前記出力することと

を含む、前記回路と  
を備える回路。

【請求項11】

積分器の前記セット内の前記蓄積した電荷を出力することが、前記蓄積した電荷をプールすることを含む、請求項10に記載の回路。

10

【請求項12】

入力データの各ベクトルの前記副部分が積分器の前記セットに関連付けられる、請求項10に記載の回路。

【請求項13】

前記クロスポイント・アレイが複数のクロスポイント・アレイであり、入力データの前記ベクトルの第1の副部分が第1のクロスポイント・アレイに送られ、入力データの前記ベクトルの第2の副部分が第2のクロスポイント・アレイに送られる、請求項10に記載の回路。

【請求項14】

積分器の前記セットによって前記電荷を蓄積することが、前記第1のクロスポイント・アレイの積分器の前記セットにより、前記第2のクロスポイント・アレイの積分器の前記セットによって蓄積される前記電荷を蓄積することを含む、請求項13に記載の回路。

20

【請求項15】

前記クロスポイント・デバイスが、前記CNNの所与の層の畳み込みカーネルの1つまたは複数の列を実装するように構成され、入力データの前記ベクトルが、前記入力データから一度に1行ずつ提示される前記CNNの前記所与の層に対するニューロン励起を表す、請求項10に記載の回路。

【請求項16】

積分器の前記セットのうちのある積分器によって蓄積された前記電荷が、前記CNNの前記所与の層に従う出力励起を表し、前記出力励起が、前記畳み込みカーネルのすべての行が統合された後にのみ、変換され、送られる、請求項15に記載の回路。

30

【請求項17】

前記クロスポイント・デバイスが、前記CNNの所与の層の畳み込みカーネルの1つまたは複数の行を実装するように構成され、前記入力データが、一度に1列ずつ提示される前記CNNの前記層に対するニューロン励起を表す、請求項10に記載の回路。

【請求項18】

積分器の前記セットのうちのある積分器によって蓄積された前記電荷が、前記CNNの前記所与の層に従う出力励起を表し、前記出力励起が、前記畳み込みカーネルのすべての列が統合された後にのみ、変換され、送られる、請求項17に記載の回路。

【発明の詳細な説明】

40

【技術分野】

【0001】

本発明は一般にはコンピューティング技術に関し、より詳細には人工ニューラル・ネットワーク（ANN）に関する。より詳細には、本発明の実施形態は、前向き推論段階中にトレーニング済み畳み込みニューラル・ネットワーク（CNN）から出力を提供するためのアナログ・メモリ・ベースのハードウェアなどのクロスポイント・アレイ内のクロスポイント・デバイスにCNNをマッピングすることに関する。

【背景技術】

【0002】

コンピュータによる文字認識やイメージ認識などの技術的問題は、機械学習技術によっ

50

てうまく処理されることが知られている。「機械学習」は、データから学習する電子システムの主要な機能を大まかに説明するために用いられる。機械学習および認知科学では、ニューラル・ネットワークは、動物の生物学的神経回路網、特に脳から着想を得た1群の統計的学習モデルである。ニューラル・ネットワークは、多数の入力に依存し、一般には未知であるシステムおよび機能を推定し、または近似するために使用され得る。ニューラル・ネットワークは、相互接続された「ニューロン」の概念に基づくアルゴリズムのクラスを使用する。典型的なニューラル・ネットワークでは、ニューロンは、入力に対して作用する所与の活性化関数を有する。適切な接続重みを決定すること（「トレーニング」とも呼ばれるプロセス）により、ニューラル・ネットワークは、イメージや文字などの所望のパターンの効率的な認識を達成する。しばしば、グループ間の接続をより明白にし、計算プロセスを編成するために、こうしたニューロンが「層」内にグループ化される。こうした適切な接続重みと共に、トレーニング中にネットワークによって決して確認されなかった当該の他のパターンも正しく認識され得、プロセスは「前向き推論」と呼ばれる。

【発明の概要】

【0003】

本発明の1つまたは複数の実施形態によれば、クロスポイント・アレイを使用して畳み込みニューラル・ネットワーク（CNN）を実装するためのコンピュータ実装方法が説明される。方法は、CNN内の畳み込み層を実装するクロスポイント・アレイを構成することを含む。構成することは、クロスポイント・アレイの1つまたは複数のクロスポイント・デバイス内に畳み込み層の1つまたは複数の畳み込みカーネルを記憶することによって実施される。方法は、操作のセットを所定の回数にわたって反復することによって、クロスポイント・アレイを介してCNNについての計算を実施することをさらに含む。操作のセットは、畳み込み層の入力データのベクトルの副部分に対応する電圧パルスをクリックポイント・アレイに送ることを含む。操作のセットは、クロスポイント・アレイ内の1つまたは複数のクロスポイント・デバイスで乗算演算を実施することを表す電流を出力することをさらに含み、電流は、クロスポイント・デバイスによって記憶された重み値と、入力データからの電圧パルスとに基づく。操作のセットはまた、積分器のセットによって、クロスポイント・デバイスからの出力電流に基づく電荷を蓄積することをも含む。方法は、所定の回数にわたって反復した後、積分器のセットによって、蓄積した電荷を出力することをさらに含み、蓄積した電荷は、入力データのベクトルの乗算加算結果(multiply-add result)と、1つまたは複数の畳み込みカーネルとを表す。

【0004】

本発明の1つまたは複数の実施形態では、積分器のセット内の蓄積した電荷を出力することは、蓄積した電荷をプールすることを含む。本発明の1つまたは複数の実施形態では、入力データの各ベクトルの副部分が積分器のセットに関連付けられる。

【0005】

本発明の1つまたは複数の実施形態では、クロスポイント・アレイはいくつかのクロスポイント・アレイを含み、入力データのベクトルの第1の副部分が第1のクロスポイント・アレイに送られ、入力データのベクトルの第2の副部分が第2のクロスポイント・アレイに送られる。本発明の1つまたは複数の実施形態では、積分器のセットによって電荷を蓄積することは、第1のクロスポイント・アレイの積分器のセットにより、第2のクロスポイント・アレイの積分器のセットによって蓄積される電荷を蓄積することを含む。

【0006】

本発明の1つまたは複数の実施形態では、クロスポイント・デバイスは、CNNの所与の層の畳み込みカーネルの1つまたは複数の列を実装するように構成され、入力データのベクトルは、入力データから一度に1行ずつ提示されるCNNの所与の層に対するニューロン励起を表す。積分器のセットのうちのある積分器によって蓄積された電荷は、CNNの所与の層に従う出力励起を表し、出力励起は、前記畳み込みカーネルのすべての行が統合された後にのみ、変換され、送られる。

【0007】

本発明の1つまたは複数の実施形態では、クロスポイント・デバイスは、CNNの所与の層の畳み込みカーネルの1つまたは複数の行を実装するように構成され、入力データは、一度に1列ずつ提示されるCNNの前記層に対するニューロン励起を表す。積分器のセットのうちのある積分器によって蓄積された電荷は、CNNの所与の層に従う出力励起を表し、出力励起は、前記畳み込みカーネルのすべての列が統合された後にのみ、変換され、送られる。

【0008】

本発明の1つまたは複数の実施形態によれば、トレーニング済み畳み込みニューラル・ネットワーク(CNN)の計算を実施するための電子回路が説明される。電子回路はクロスポイント・アレイと出力回路とを含み、出力回路は1つまたは複数の積分器を含む。方法はクロスポイント・アレイを設けることと、出力回路を設けることとをさらに含む。方法は、クロスポイント・アレイの1つまたは複数のクロスポイント・デバイス内にCNN内の畳み込み層の1つまたは複数の畳み込みカーネルを記憶することによって、畳み込み層に対応するクロスポイント・アレイを構成することをさらに含む。方法は、所定の回数にわたって操作のセットを反復することをさらに含む。CNNのトレーニングが前述の方法を使用して実施される。

【0009】

本発明の1つまたは複数の実施形態によれば、抵抗性メモリ素子のアレイを含む電子回路が説明される。アレイは、(i) アナログ入力値のベクトルを符号化するアレイに対する電圧入力のベクトルと、(ii) アレイ内のアナログ抵抗性重みの行列との間のアナログ・ベクトル-行列積に等しい電流出力のベクトルを提供する。電子回路は、抵抗性メモリ素子の専用サブセットからの電流を集約する蓄積ワイヤおよび回路をさらに含む。電子回路は統合コンデンサをさらに含み、統合コンデンサのそれぞれは、単一の統合ステップの間に複数の蓄積ワイヤのうちの1つからの電流を集約するように電気的に切換え可能である。電子回路は、いくつかの統合ステップにわたって蓄積した統合コンデンサのサブセットからの統合電荷を、アナログ持続時間または2進数字を使用するデジタル表現のどちらかとして適切に変換し、送ることを可能にするデータ出力回路をさらに含む。抵抗性メモリ素子は、畳み込みニューラル・ネットワークの所与の層のシナプス重みカーネルのベクトルを実装するように構成される。

【0010】

本発明の1つまたは複数の実施形態によれば、方法は、前記重みカーネルの複数の部分ベクトルにわたる乗算累積演算(multiply-accumulate operation)を実装する、いくつかの統合ステップにわたる蓄積を実施するために電子回路を使用する。蓄積は、所定の回数にわたって操作のセットを反復することによってクロスポイント・アレイの抵抗性メモリ素子による計算を実施することを含む。操作のセットは、アナログ入力値の各ベクトルを複数の部分ベクトルに区分化することを含む。操作のセットはまた、複数の部分ベクトルのそれぞれに対応する部分出力励起をアナログ・メモリ内に蓄積することをも含む。操作のセットはまた、統合電荷を蓄積する統合コンデンサに部分出力励起をルーティングすることによって部分出力励起を組み合わせることをも含む。さらに、蓄積は、出力励起を表す統合コンデンサ上の統合電荷を送ることをさらに含む。

【0011】

本発明の1つまたは複数の実施形態では、統合コンデンサ上の統合電荷が、統合電荷を送る前に局所的にプールされる。本発明の1つまたは複数の実施形態では、抵抗性メモリ素子は不揮発性メモリ・デバイスである。本発明の1つまたは複数の実施形態では、抵抗性メモリ素子のサブセットは、アレイの1つまたは複数の列に対応する。本発明の1つまたは複数の実施形態では、抵抗性メモリ素子のサブセットは、アレイの1つまたは複数の行に対応する。

【0012】

本発明の1つまたは複数の実施形態では、クロスポイント・デバイスは、畳み込みニューラル・ネットワークの所与の層の畳み込みカーネルの1つまたは複数の行を実装するよ

10

20

30

40

50

うに構成され、入力データは、一度に1列ずつ提示される畳み込みニューラル・ネットワークの前記層に対するニューロン励起を表す。

【0013】

本発明の1つまたは複数の実施形態では、クロスポイント・デバイスは、畳み込みニューラル・ネットワークの所与の層の畳み込みカーネルの1つまたは複数の列を実装するように構成され、入力データのベクトルは、入力データから一度に1行ずつ提示される畳み込みニューラル・ネットワークの所与の層に対するニューロン励起を表す。

【0014】

技術的解決策は、以下の説明で述べられ、または図面に示される構成要素の構成および配置の詳細に適用が限定されないことを理解されたい。技術的解決策は、記載のものに加えて実施形態が可能であり、様々な方式で実践および実施することが可能である。さらに、本明細書ならびに要約で利用される表現および用語は説明のためのものであり、限定と見なされるべきではないことを理解されたい。したがって、本開示の基礎となる概念が、現在説明している技術的解決策のいくつかの目的を達成するための他の構造、方法、およびシステムの設計のための基礎として容易に利用され得ることを当業者は理解されよう。

【0015】

以下の説明および説明を参照して、本文書全体を通して説明される例をより良く理解されよう。図の構成要素は必ずしも原寸に比例しない。さらに、図では、様々な図全体にわたって、同様の参照番号は対応する部分を示す。

【図面の簡単な説明】

【0016】

【図1】 数学的ニューロンの入力および出力の接続の簡略化した図である。

【図2】 図1に示される数学的ニューロンの簡略化したモデルを示す図である。

【図3】 図2に示される数学的ニューロン・モデルを組み込むANNの簡略化したモデルを示す図である。

【図4】 サンプル入力マップを解釈している代表的なCNNの簡略化したブロック図である。

【図5】 入力マップおよび畳み込みカーネルを含むトレーニング・データを使用してトレーニングされているCNN内の例示的畳み込み層を示す図である。

【図6】 本発明の1つまたは複数の実施形態による、クロスポイント・アレイを使用して行列-行列乗算を実施するためのシステムを示す図である。

【図7】 本説明による、前向き行列乗算(forward matrix multiplication)、後ろ向き行列乗算(backward matrix multiplication)、および重み更新を実施する2次元(2D)クロスバー・システムを示す図である。

【図8】 本発明の1つまたは複数の実施形態によるクロスポイント・アレイの拡大図である。

【図9】 クロスバー・システム内の典型的な出力回路を示す図である。

【図10】 クロスポイント・アレイを使用して前向き推論演算を実施するための既存の演算を示す図である。

【図11】 本発明の1つまたは複数の実施形態による、部分蓄積を使用して前向き推論演算を実施することを示し、部分蓄積が時間区分化に基づく図である。

【図12】 本発明の1つまたは複数の実施形態による、複数のクロスポイント・アレイにわたって部分蓄積を使用して前向き推論演算を実施することを示す図である。

【図13】 本発明の1つまたは複数の実施形態による、部分蓄積を使用して前向き推論演算を実施することを示し、部分蓄積が空間区分化に基づく図である。

【発明を実施するための形態】

【0017】

本明細書で説明される技術的解決策は、既存の技術よりも効率的な方式で畳み込みニューラル・ネットワークを使用するディープ・ラーニング技術の実装を促進する。ディープ・ラーニング技術は、イメージ認識や音声認識などのマシン・ベースのパターン認識問題

10

20

30

40

50

で広く使用される。ディープ・ラーニングは本質的に、(ビッグ・データの使用と共に高まる)大規模なトレーニング・データ・セットおよび(ムーアの法則に従って増大すると予想される)計算能力の可用性を活用する。

#### 【0018】

本発明の実施形態は、クロスポイント・アレイを使用する人工ニューラル・ネットワーク(ANN)などの、アナログ人工知能システムを実装するときの、畳み込みニューラル・ネットワーク(CNN)の、アナログ・アレイへの効率的な作業負荷マッピングを促進する。既存の技術は、CNNの各層を介する活性化が効率的に使用され、能率化され、記憶要件が限定されるように、CNN推論作業負荷についての重みの「行単位」マッピングを記述する。しかしながら、そのような「行単位」マッピング技術に伴うアナログ・アレイ面積使用率が低く、そのような技術のスケラビリティに影響を及ぼす点で、既存の技術に伴う技術的課題が存在する。たとえば、大規模なCNN(ResNet-50など)をマッピングすることは、CNNを実装するために多数のアナログ・アレイを必要とし、そのことにより、実装が非効率となり、扱いにくくなり、費用が法外に高いもの(cross prohibitive)となり得る。

#### 【0019】

本発明の実施形態は、行単位マッピング技術のためのCNN層のコンパクトなマッピングを促進する柔軟なアレイ間ルーティング方式を提供することにより、ANN、特にCNNの実装中のそのような技術的課題に対処する。本発明の1つまたは複数の実施形態は、活性化が能率化されず、または再利用されない場合に、汎用マッピング技術を使用する既存の行単位マッピング技術に対して、必要とされるアナログ・アレイ(タイル)数を評価する。したがって、本発明の実施形態は、行単位マッピングについての能率化された活性化の利点を保持しながら、広範なCNNについて同程度であるアレイ使用率を促進する。

#### 【0020】

1つまたは複数の実施形態が、生物学的神経回路網の文脈で、脳構造および機能をモデル化することを特に強調して説明されるが、本明細書に記載の教示の実装は特定の環境をモデル化することに限定されないことをあらかじめ理解されたい。むしろ、本発明の実施形態は、環境に対する様々な入力をベクトルに変えることができる限り、たとえば、気象パターン、インターネットから収集された任意のデータなどを含む任意のタイプの環境をモデル化することができる。

#### 【0021】

ANNはしばしば、シミュレートされた「ニューロン」として働き、電子信号の形態で互いの間で「メッセージ」を交換する、相互接続されたプロセッサ素子のいわゆる「ニューロモーフィック」システムとして実施される。生物学的ニューロン間でメッセージを搬送するシナプス神経伝達物質接続のいわゆる「可塑性」と同様に、シミュレートされたニューロン間で電子メッセージを搬送するANN内の接続は、所与の接続の強さまたは弱さに対応する数値重みを与えられ。重みは、経験に基づいて調節および調整され得、ANNが入力に適応し、学習が可能となる。たとえば、手書き認識についてのANNは、入力イメージのピクセルによって活性化され得る入力ニューロンのセットによって定義される。ネットワークの設計者によって決定される関数によって重み付けされ、変換された後、こうした入力ニューロンの活性化が、しばしば「隠れ」ニューロンと呼ばれる他の下流側ニューロンに渡される。出力ニューロンが活性化されるまでこのプロセスが反復される。活性化された出力ニューロンは、どの文字が読み取られたかを判定する。

#### 【0022】

クロスバー・アレイは、クロスポイント・アレイ、クロスワイヤ・アレイ、または抵抗性処理装置(RPU)アレイとも呼ばれ、ANNアーキテクチャ、ニューロモーフィック・マイクロチップ、および超高密度不揮発性メモリを含む様々な電子回路およびデバイスを形成するために使用される、高密度で低コストの回路アーキテクチャである。基本的クロスポイント・アレイ構成は、導電性行ワイヤのセットと、導電性行ワイヤのセットと交差するように形成された導電性列ワイヤのセットとを含む。ワイヤの2つのセットの間の

交差は、いわゆるクロスポイント・デバイスによって分離され、クロスポイント・デバイスは薄膜材料から形成され得る。

#### 【0023】

クロスポイント・デバイスは実際には、ニューロン間のANNの重み付き接続として機能する。ナノスケール2端子デバイスデバイス、たとえば「理想的な」伝導状態切換え特性を有するメモristaが、高いエネルギー効率を有するシナプス可塑性をエミュレートするためにクロスポイント・デバイスとしてしばしば使用される。理想的なメモrista材料の伝導状態（たとえば、抵抗）が、行ワイヤおよび列ワイヤの個々のワイヤ間に印加される電圧を制御することによって変更され得る。デジタル・データが、高伝導状態または低伝導状態を達成するための交点でのメモrista材料の伝導状態の変更によって記憶され得る。メモrista材料はまた、材料の伝導状態を選択的に設定することによって2つ以上の別個の伝導状態を維持するようにプログラムされ得る。メモrista材料の伝導状態は、材料の両端間に電圧を印加し、ターゲット・クロスポイント・デバイスを通して電流を測定することによって読み取られ得る。

10

#### 【0024】

電力消費を制限するために、ANNチップ・アーキテクチャのクロスポイント・デバイスはしばしば、オフライン学習技術を利用するように設計され、初期トレーニング段階が解決された後は、目的関数の近似が変化しない。オフライン学習は、電力をほとんど消費しないようにクロスバー型ANNアーキテクチャのクロスポイント・デバイスを簡略化することを可能にする。

20

#### 【0025】

低電力消費、高計算スループット、および低待ち時間で、以前にトレーニングされたANNネットワークの前向き推論を実装し得る単純なクロスポイント・デバイスを提供することにより、全体のANN性能が改善され、より広い範囲のANN応用が可能となる。

#### 【0026】

本発明は電子システムを対象とするが、参照および説明を容易にするために、記載の電子システムの様々な態様が、たとえばニューロン、可塑性、シナプスなどの神経学的用語を用いて説明される。電子システムの本明細書でのどんな議論または図示についても、神経学的用語または神経学的略式表記の使用は、参照を容易にするためのものであり、記載の神経学的機能または神経学的構成要素のニューロモर्फック、ANN同等物を包含することを意味することを理解されよう。

30

#### 【0027】

ANNは、ニューロモर्फック・システムまたはシナプトロニック・システムとも呼ばれ、たとえば生物学的神経系、人間の脳、およびイメージ認識、音声認識などの脳のような機能を含む他の機能またはシステムを推定または近似し得る計算システムである。ANNは、神経生理学、認知科学／心理学、物理学（統計力学）、制御理論、計算機科学、人工知能、統計／数学、パターン認識、コンピュータ・ビジョン、並列処理、およびハードウェア（たとえば、デジタル／アナログ／VLSI／光学）を含む様々な分野から知識を組み込む。

#### 【0028】

0および1を操作する従来のデジタル・モデルを利用する代わりに、ANNは、推定または近似されているコア・システム機能の実質的に機能的同等物である処理要素間の接続を生み出す。たとえば、電子ニューロモर्फック・マシンの中心構成要素であるコンピュータ・チップが、哺乳動物の脳と類似の形態、機能、およびアーキテクチャを提供しようと試みる。コンピュータ・チップは従来のコンピュータ・チップと同一の基本的トランジスタ構成要素を使用するが、そのトランジスタは、ニューロンおよびそのシナプス接続の挙動を模倣するように構成される。コンピュータ・チップは、百万を超えるシミュレートされた「ニューロン」のネットワークを使用して情報を処理し、ニューロンは、生物学的ニューロン間のシナプス通信と同様に、電気的スパイクを使用して互いに通信する。そのようなコンピュータ・チップのアーキテクチャは、メモリ（すなわち、シミュレートさ

40

50

れた「シナプス」)を読み取り、単純な演算を実施するプロセッサ(すなわち、シミュレートされた「ニューロン」)の構成を含む。通常は相異なるコア内に配置される、こうしたプロセッサ間の通信(経路)は、オンチップ・ネットワーク・ルータによって実施される。

#### 【0029】

背景として、図1、2、および3を参照しながら、典型的なANNがどのように動作するかについての一般的説明が次に与えられる。本明細書で先に言及したように、典型的なANNは、ニューロンと呼ばれる約1000億の相互接続された細胞を含む人間の脳から着想を得た数学的モデルである。図1は、経路104、106、108、110を有する数学的ニューロン102の簡略化した図であり、経路104、106、108、110は、図示されるように構成され、配置された、上流側入力112、114、下流側出力116、および下流側の「他の」ニューロン118に数学的ニューロン102を接続する。各数学的ニューロン102は、経路104、106、108、110を通じて電氣的インパルスを送り、受け取る。こうした電氣的インパルスの性質、およびどのように生物学的ニューロン(図示せず)で電氣的インパルスが処理されるかが、主に全体の脳機能を担う。この機能を模倣することが、ネットワークに編成された数学的ニューロン102から構築された数学的ANNの意図である。生物学的ニューロン間の経路接続は強いことがあり、または弱いことがあるように、数学的ニューロン間の経路も同様であり得る。所与のニューロンが入力インパルスを受け取るとき、ニューロンは、ニューロンの機能に従って入力を処理し、機能の結果を下流側出力または下流側の「他の」ニューロンあるいはその両方に送る。

#### 【0030】

図2で、数学的ニューロン102が、図2に示される式によって示される数学関数 $f(x)$ を有するノード202としてモデル化される。ノード202は、入力212、214から電気信号を取り、各入力212、214にそれぞれの接続経路204、206の強さを乗算し、入力の和を取り、関数 $f(x)$ を通じて和を渡し、結果216を生成し、結果216は、最終的出力、または別のノードに対する入力、またはその両方であり得る。この説明では、アスタリスク(\*)が乗算を表すために用いられ、乗算は行列乗算であり得る。たとえば、入力データと1つまたは複数の畳み込みカーネルとの間の畳み込み演算を実施して出力マップを生成するために、行列乗算が使用され得る。弱い入力信号には非常に小さい接続強度数が乗算され、したがって関数に対する弱い入力信号の影響は非常に低い。同様に、強い入力信号には、より高い接続強度数が乗算され、したがって関数に対する強い入力信号の影響はより大きい。関数 $f(x)$ は設計の選択であり、様々な関数が使用され得る。 $f(x)$ についての典型的な設計の選択は双曲線正接関数であり、双曲線正接関数は、前の和の関数を取り、-1から+1の間の数を出力する。 $f(x)$ の代替の設計の選択は、正の入力について出力が入力に合致し、そうでない場合は出力がゼロである関数である、正規化線形ユニット(ReLU)である。

#### 【0031】

図3は、重み付き有向グラフとして編成された、簡略化したANNモデル300を示し、人工ニューロンはノード(たとえば、302、308、316)であり、重み付き有向端(たとえば、m1からm20)がノードを接続する。ANNモデル300は、ノード302、304、306が入力層ノードであり、ノード308、310、312、314が隠れ層ノードであり、ノード316、318が出力層ノードであるように編成される。各ノードは、接続経路によって、隣接する層内のあらゆるノードに接続され、図3では、接続経路は、接続強度m1からm20を有する有向矢印として示される。ただ1つの入力層、1つの隠れ層、1つの出力層が示されているが、実際には、複数の入力層、隠れ層、および出力層が設けられ得る。

#### 【0032】

人間の脳の機能を模倣するこの試みにおいて、ANN300の各入力層ノード302、304、306が、接続強度調節およびノード合計を行わずに、ソース(図示せず)から

10

20

30

40

50

直接的に入力  $x_1$ 、 $x_2$ 、 $x_3$  を受け取る。したがって、図 3 の下端に列挙される式によって示されるように、 $y_1 = f(x_1)$ 、 $y_2 = f(x_2)$ 、および  $y_3 = f(x_3)$  である。各隠れ層ノード 308、310、312、314 は、関連する接続経路に関連付けられる接続強度に従って、すべての入力層ノード 302、304、306 からその入力を受け取る。したがって、隠れ層ノード 308 では、 $y_4 = f(m_1 * y_1 + m_5 * y_2 + m_9 * y_3)$  であり、 $*$  は乗算を表す。1 つまたは複数の例では、乗算は、畳み込み演算を実施するために使用される行列乗算であり得る。図 3 の下端に示される関数  $y_5$  から  $y_9$  を定義する式によって示されるように、隠れ層ノード 310、312、314、出力層ノード 316、318 について同様の接続強度乗算およびノード合計が実施される。

#### 【0033】

ANN モデル 300 は、データ・レコードを一度に 1 つずつ処理し、レコードの当初は任意の分類をレコードの既知の実際の分類と比較することによって「学習する」。「逆伝播」（すなわち、「誤差の逆伝播」）と呼ばれるトレーニング方法を使用して、第 1 のレコードの初期分類からの誤差がネットワーク内にフィードバックされ、2 回目にネットワークの重み付き接続を修正するために使用され、このフィードバック・プロセスが何回かの反復にわたって続行される。ANN のトレーニング段階では、各レコードについての正しい分類が知られており、したがって出力ノードに「正しい」値が割り当てられ得、たとえば正しいクラスに対応するノードについてノード値「1」（または 0.9）、他のノードについてノード値「0」（または 0.1）が割り当てられる。したがって、出力ノードについてのネットワークの計算値をこうした「正しい」値と比較し、各ノードについての誤差項（すなわち、「デルタ」規則）を計算することが可能である。次いで、こうした誤差項が、次の反復で出力値が「正しい」値に近づくように、隠れ層内の重みを調節するために使用される。

#### 【0034】

多くのタイプのニューラル・ネットワークがあるが、2 つの最も広いカテゴリは、フィードフォワードおよびフィードバック／再帰ネットワークである。ANN モデル 300 は、入力、出力、および隠れ層を有する非再帰フィードフォワード・ネットワークである。信号は一方向にのみ移動し得る。入力データが、計算を実施する処理要素の層に対して渡される。各処理要素は、その入力の重み付き和に基づいてその計算を行う。次いで、新しい計算値が、次の層に供給する新しい入力値となる。このプロセスが、すべての層を経て、出力を決定するまで続行される。しきい値伝達関数が、出力層内のニューロンの出力を定量化するために使用されることがある。

#### 【0035】

フィードバック／再帰ネットワークはフィードバック経路を含み、そのことは、ループを使用して信号が両方向に移動し得ることを意味する。ノード間のすべての可能な接続が許可される。このタイプのネットワークではループが存在するので、一定の操作の下で、このタイプのネットワークは、平衡の状態に達するまで継続的に変化する非線形動的システムとなり得る。フィードバック・ネットワークはしばしば、連想記憶および最適化問題で使用され、ネットワークは相互接続された要素の最良の配置を探す。

#### 【0036】

フィードフォワードおよび再帰 ANN アーキテクチャでの機械学習の速度および効率は、ANN クロスポイント・アレイのクロスポイント・デバイスが典型的な機械学習アルゴリズムのコア動作をどのように効果的に実施するかに依存する。機械学習の厳密な定義を定式化することは難しいが、ANN の文脈での学習プロセスは、ネットワークが特定のタスクを効率的に実施し得るようにクロスポイント・デバイス接続重みを更新する問題と見なすことができる。クロスポイント・デバイスは通常、利用可能なトレーニング・パターンから必要な接続重みを学習する。ネットワーク内の重みを反復的に更新することによって性能が経時的に改善される。人間のエキスパートによって指定される規則のセットに従う代わりに、ANN は、代表例の所与の集合から（入力ー出力関係のような）基礎となる規則を「学習」する。したがって、学習アルゴリズムは一般に、関連する重みを更新また

10

20

30

40

50

は調節あるいはその両方を行うために学習規則が使用される手順と定義され得る。

【0037】

3つの主な学習アルゴリズム・パラダイムは、教師あり、教師なし、およびハイブリッドである。教師あり学習、すなわち「教師」を伴う学習では、ネットワークに、あらゆる入力パターンについての正しい回答（出力）が与えられる。既知の正しい回答に可能な限り近い回答をネットワークが生成するように重みが求められる。強化学習は、正しい回答自体ではなく、ネットワーク出力の正しさに関する論評だけがネットワークに与えられる、教師あり学習の変種である。一方、教師なし学習、すなわち教師を伴わない学習は、トレーニング・データ・セット内の各入力パターンに関連付けられる正しい回答を必要としない。データ内の基礎となる構造、またはデータ内のパターン間の相関を探索し、こうした相関からパターンをカテゴリに編成する。ハイブリッド学習は、教師あり学習と教師なし学習を組み合わせる。重みの各部分は通常、教師あり学習を通じて決定され、他の部分は教師なし学習を通じて取得される。ANNおよび学習規則の追加の詳細が、Artificial Neural Networks: A Tutorial, by Anil K. Jain, Jianchang Mao, and K.M. Mohiuddin, IEEE, March 1996で説明されており、その説明全体が、参照により本明細書に組み込まれる。

【0038】

ANNのトレーニングの応用を越えて、既にトレーニングされたネットワークの前向き推論は、ANN上で構築されるクラウド・ベースのサービスの実装から、スマートフォン、モノのインターネット（IoT）、極めて低電力の動作を必要とする他の電池が制限された適用例にまで及ぶ適用例を含む。一般には、トレーニングは（多くのトレーニング例から学習するために）高スループットを必要とする適用例であり、前向き推論は、（所与の新しいテスト例を可能な限り迅速に分類し、認識し、あるいは処理することができるように）高速な待ち時間を必要とする適用例である。

【0039】

CNNでは、カーネルは、視野内などの重複する領域を畳み込み、したがって特徴検出で空間的局所性の重要性を強調する。CNNの畳み込み層を計算することは通常、ニューラル・ネットワーク・トレーニングおよび推論での計算時間の90%超を包含する。最小限の関係のないデータの移動または計算で、CNNをアナログ・アレイにマッピングし、畳み込み層の数学的演算を実施している間に使用される電力の効率的な使用を保証することは技術的課題である。技術的課題には、推論のためにCNNをマッピングすること、ならびにResNet-50などの大規模なCNNであっても実装され得るようにそのようなマッピングのスケラビリティを維持することが含まれる。行単位マッピングを使用する既存の解決策はアレイ間ルーティング回路に関する一定の制限を仮定するが、本発明の1つまたは複数の実施形態は、行単位技術についてCNN層のクロスポイント・アレイへのコンパクトなマッピングを可能にするデータの柔軟なアレイ間ルーティングを促進する。

【0040】

本発明の実施形態によって実装される技術的解決策は、行単位マッピングについての効率化された活性化の利点を保持しながら、広範なCNNネットワークについて非常に同程度であるアレイ使用率を実現することによってそのような技術的問題に対処する。

【0041】

図4はCNNの簡略化したブロック図を示す。図示される例では、CNNは、サンプル入力マップ400を解釈するために使用されており、この特定の例では、手書き文字「w」を入力マップとして使用する。しかしながら、他のタイプの入力マップが可能であること、さらには本明細書で説明される技術的解決策が他のタイプの特徴検出などの他の演算を実施するCNNに適用可能であることを理解されたい。図示される例では、入力マップ100が、入力層410または「layer-1」についての値のセットを生成するために使用される。たとえば、layer-1は、サンプル入力マップ400のピクセルの、layer-1内の特定のニューロンへの直接マッピングによって生成され得、したがってニューロンは、ピクセルが特定の属性を示すかどうかに応じて1または0を示す。値を

ニューロンに割り当てる別の例示的方法が、畳み込みニューラル・ネットワークを参照しながら以下で論じられる。ニューラル・ネットワークの変動、および解決するように作成される問題に応じて、ネットワークの各層は、異なる数のニューロンを有し得、これらは、入力データの特定の品質に関係することがあり、または関係しないことがある。

#### 【0042】

図4を参照すると、前述のように（図3参照）、layer-1 410内のニューロンが、次の層であるlayer-2 420内のニューロンに接続される。図4のニューロンは、図1を参照して説明したものと同様である。したがって、layer-2 420内のニューロンは、layer-1 410内のニューロンのそれぞれから入力値を受け取る。次いで入力値が合計され、この和がバイアスと比較される。値が特定のニューロンについてのバイアスを超過する場合、そのニューロンは値を保持し、値は、ニューロンの次の層内のニューロンに対する入力として使用され得る。この計算が、少なくとも1つのFC層450を含むCNNの様々な層430~450を通じて、図4で「出力」と呼ばれる最終層460に達するまで続行される。いくつかのCNNネットワークでは、前の層からの「残留」結果が後の層の結果と組み合わせられ得、その間の層がスキップされる。文字認識のために使用されるCNNの一例では、層内の各値が特定の文字に割り当てられる。分類作業のために設計されるとき、ネットワークは、1つのニューロンでただ1つの大きい正の値を有する出力層で終了するように構成され、次いでその出力層は、どの文字をネットワークが最も可能性の高い手書き入力文字であると計算したかを実証する。他のシナリオでは、ネットワークは、出力ニューロン値が確率（尤度）、信頼度、または他の注目のメトリックを推定するために使用され得るように設計されていることがある。

#### 【0043】

CNN内の各層についてのデータ値は、通常は行列（または、いくつかの例ではテンソル）を使用して表現され、計算は行列計算として実施される。図4に示されるように、行列の添字（またはサイズあるいはその両方）は層ごとに、およびネットワークごとに様々である。異なる実装は、様々に行列を適応させ、または様々に行列をコンピュータ・メモリにマッピングする。図4を参照すると、図示される例示的CNNでは、ニューラル・ネットワークの各層についての行列の次元によって示されるように、各レベルはニューロン値のテンソルである。CNNの入力では、一例は、それぞれ2次元イメージである複数の入力「平面」であり得る。たとえば、フル・カラー・イメージから生じる赤色平面、緑色平面、および青色平面があり得る。CNN内により深く入ると、層は多くの「平面」の形態の中間データを取り、次の層に対して多数の出力平面を生成し得る。層の入力テンソル内の値に、フィルタと呼ばれる変換テンソル内にある接続強度が乗算される。この行列乗算は、接続強度に従って前の層内の各値をスケーリングし、次いでこうした寄与の集約合計が合計される。この基本演算は乗算累積演算と呼ばれる。次いでバイアス行列が、得られる積行列に加えられ、次のレベルの各ニューロンのしきい値が調節される。さらに、得られた各値に活性化関数が適用され、得られた値が、次の層に適用される出力テンソル内に配置される。一例では、活性化関数は正規化線形ユニット、S字形、またはtanh（）であり得る。したがって、図4が示すように、各層の間の接続、したがってネットワーク全体が、一連の行列として表現され得る。CNNをトレーニングすることには、こうした行列についての適切な値を見つけることが含まれる。

#### 【0044】

全結合ニューラル・ネットワークは、適切にトレーニングされるとき、手書きや家庭のペットの写真などの入力パターンを認識することができるが、シフト不変性を示さない。ネットワークが猫のひげを認識するために、イメージ内の多数の異なる2D位置に位置するひげのある猫のイメージを供給しなければならない。それぞれの異なるイメージ位置は、そのような全結合ネットワーク内の異なる重みと相互作用するニューロン値となる。一方、CNNでは、接続強度は畳み込みカーネルである。畳み込み演算はシフト不変性を導入する。したがって、ひげのある猫の複数のイメージが提示されるとき、ひげの縮尺、カラー、および回転がイメージごとに不変である限り、イメージ内の2D位置はもはや問題

ではない。したがって、トレーニングの間に、2Dイメージ内の特徴位置とは無関係に、同様の特徴のすべての例がこの特徴を学習する助けとなるように共に働く。トレーニングの後、フィルタの単一のセットまたはずっと小さいセットが、そのようなイメージ特徴を認識するために十分であり、次いで多くのフィルタのバンク（CNN層が何であるか）が、イメージを区別するのに有用である多くの異なる特徴（猫と犬、さらには異なる猫の品種を表す微細な点）を認識することを可能にする。

【0045】

図5は、入力マップ510および畳み込みカーネル520を含むトレーニング・データを使用してトレーニングされているCNN内の例示的畳み込み層500を示す。簡単のために、図5はバイアス行列525を示さない。入力マップ510（入力平面とも呼ばれる）は、複数の入力パターン、たとえばD個の入力マップを含み得る。各入力マップは、サイズ $N \times M$ の行列などの行列である。したがって、このケースでは入力ニューロンの総数は $N \times M \times D$ である。入力マップは、図示されるように、サイズ $k \times k$ のF個の畳み込みカーネル520と共に畳み込まれ、対応する出力であるマップ530が生成される。各出力マップは次元 $N' \times M'$ を有し得る。入力マップがサイズ $n$ の正方行列であるケースでは、出力マップはサイズ $n - k + 1 \times n - k + 1$ である。各畳み込みはD個の入力マップを伴う3D畳み込みである。CNNは、複数のそのような層を含み得、前の層からの出力マップ530が、後続の層のための入力マップ510として使用される。逆伝播アルゴリズムがフィルタの $k \times k \times D \times F$ 重み値を学習するために使用され得る。

【0046】

たとえば、入力マップ510が各フィルタ・バンクと共に畳み込まれ、対応する出力マップが生成される。たとえば、CNNが手書きを識別するためにトレーニングされているケースでは、入力マップ510が、垂直線を表す畳み込みカーネルを含むフィルタ・バンクと組み合わせられる。得られる出力マップは、入力マップ510内に存在する垂直線を識別する。さらに、別のフィルタ・バンクは、右上に進むような斜線を表す畳み込みカーネルを含み得る。入力マップ510と第2のフィルタ・バンクの畳み込みから得られる出力マップが、斜線を含むトレーニング・データのサンプルを識別する。2つの出力マップは文字についての異なる情報を示すと共に、ピクセル隣接を保持する。この結果、より効率的な文字認識となり得る。

【0047】

図6は、本発明の1つまたは複数の実施形態による、とりわけ行列－行列乗算を実施するためにコントローラ610を使用してクロスポイント・アレイ700が制御されるシステム600を示す。たとえば、コントローラ610は、クロスポイント・アレイ700によって乗算される入力データ510を送る。1つまたは複数の例では、コントローラ610はクロスポイント・アレイ700内の畳み込みカーネル520などからの重み値を記憶し、入力ベクトルを送る。1つまたは複数の例では、コントローラ610とクロスポイント・アレイ700がワイヤードまたはワイヤレスまたはそれらの組合せで結合される。コントローラ610はさらに、CNN内の1つまたは複数の層についての演算を開始するために、命令／コマンドをクロスポイント・アレイ700に送る。コントローラ610はさらに、計算が実施されたという通知を受け取った後、クロスポイント・アレイ700から出力データ530を読み取り得る。コントローラ610は、処理装置、またはサーバ、デスクトップ・コンピュータ、タブレット・コンピュータ、電話機などのコンピューティング・システムであり得る。コントローラ610は、コンピュータ実行可能命令を記憶したメモリ・デバイスを含み得、命令は、コントローラによって実行されるとき、行列－行列計算を引き起こす。

【0048】

次にこの説明の概要を参照すると、1つまたは複数の実施形態は、クロスバー・ワイヤの各交点にクロスポイント・デバイスを有するクロスポイント・アレイを対象とし、クロスポイント・アレイはCNNを実装するために使用される。クロスポイント・デバイスの一例は、ローカル・データ・ストレージ機能およびローカル・データ処理機能を提供する

10

20

30

40

50

、本明細書で抵抗性処理装置（RPU）と呼ばれる２端子プログラマブル抵抗性クロスポイント構成要素である。データ処理を実施するとき、各クロスポイント・デバイスによって表される重み付き寄与が、データの記憶された位置で実施される大規模並列乗算累積演算に寄与する。これにより、プロセッサおよび別々の記憶素子の中および外に関連データを移動する必要が解消される。したがって、記載のクロスポイント・デバイスを有する機械学習CNNアーキテクチャを実装することは、CNNをトレーニングし、その後でトレーニング済みCNNモデルを使用して推論を実施することを促進するオンライン機械学習機能の実装を可能にする。記載のクロスポイント・デバイスおよび得られるCNNアーキテクチャは、全体のCNN性能を改善し、より広い範囲の実際のCNN適用を可能にする。

【0049】

10

記載のクロスポイント・デバイスは、２端子抵抗性クロスポイント・デバイスとして実装され得る。たとえば、記載のクロスポイント・デバイスは、抵抗性ランダム・アクセス・メモリ（RRAM）、相変化メモリ（PCM）、プログラマブル・メタライゼーション・セル（PMC）メモリ、非線形メモリスタ・システム、または経時的に十分に安定な、広範なアナログ可同調不揮発性抵抗性メモリ状態を提供する任意の他のデバイスと共に実装され得る。

【0050】

図7は、この説明による、前向き推論を実施する２次元（2D）クロスバー・システム700を示す。クロスバー・システム700は、逆伝播アルゴリズムに従って、単純な行列乗算、後ろ向き行列乗算、さらにはin-situ重み更新を実装するために使用され得る。クロスバー・システム700は、とりわけクロスポイント・アレイ705、入力回路710、および出力回路720を含む。入力回路710および出力回路720は、一緒に周辺回路と呼ばれることがある。クロスバー・システム700は、１つまたは複数の例ではコンピュータ・チップであり得る。

20

【0051】

図8は、１つまたは複数の実施形態によるクロスポイント・アレイ705の拡大図を示す。クロスポイント・アレイ705は、導電性行ワイヤ802、804、806のセットと、導電性行ワイヤ802、804、806のセットと交差する導電性列ワイヤ808、810、812、814のセットとから形成される。行ワイヤのセットと列ワイヤのセットの間の交点は、クロスポイント・デバイスによって分離され、図8では、クロスポイント・デバイスは、 $\sigma_{11}$ 、 $\sigma_{21}$ 、 $\sigma_{31}$ 、 $\sigma_{41}$ 、 $\sigma_{12}$ 、 $\sigma_{22}$ 、 $\sigma_{32}$ 、 $\sigma_{42}$ 、 $\sigma_{13}$ 、 $\sigma_{23}$ 、 $\sigma_{33}$ 、 $\sigma_{43}$ としてそれぞれ示される、それ自体の調節可能／更新可能抵抗性重みをそれぞれ有する抵抗素子として示される。図を簡単にするために、図8ではただ１つのクロスポイント・デバイス820に参照番号が付けられている。前向き行列乗算では、クロスポイント・デバイスの両端間に電圧を印加し、クロスポイント・デバイスを通過する電流を測定することによって、クロスポイント・デバイスの伝導状態（すなわち、記憶された重み）が読み取られ得る。

30

【0052】

入力電圧 $V_1$ 、 $V_2$ 、 $V_3$ が、それぞれ行ワイヤ802、804、806に印加される。各列ワイヤ808、810、812、814は、コンデンサなどの積分器を使用して、特定の列ワイヤに沿った各クロスポイント・デバイスによって生成された電流 $I_1$ 、 $I_2$ 、 $I_3$ 、 $I_4$ を合計する。たとえば、図8に示されるように、列ワイヤ814によって生成される電流 $I_4$ は、式 $I_4 = V_1 \sigma_{41} + V_2 \sigma_{42} + V_3 \sigma_{43}$ によって与えられる。したがって、アレイ705は、クロスポイント・デバイス内に記憶された値に、電圧 $V_1$ 、 $V_2$ 、 $V_3$ によって定義される行ワイヤ入力を乗算することによって前向き行列乗算を計算する。

40

【0053】

図7を参照すると、１つまたは複数の例では、入力回路710は、少なくとも支持回路712、共有回路714、および行回路716を含む。行回路は、各行ワイヤ802、804、および806に関連付けられるハードウェア構成要素を含む。入力回路710は、

50

クロスポイント・アレイ 705 に入力電圧を与えることを促進する。

【0054】

図 9 は典型的な出力回路 720 を示す。出力回路は、列ワイヤ 808、810、812、および 814 に対応する積分器 908、910、912、914 を含む。1 つまたは複数の例では、積分器 908、910、912、および 914 はコンデンサである。各列ワイヤに沿った出力電流が積分器内に蓄積され、CNN の次の層に渡される。前述のように、そのような積分器の構成により、FC 層の計算が非常に効率的になる。しかしながら、畳み込み演算では、そのような積分器の構成を使用することは、データ移送、記憶、編成、および後続のデータ移送の点で著しい追加のオーバーヘッドを招く。そのような動作は、時間、電力、追加の回路面積などの追加のリソースを必要とし、したがってシステム全体が非効率となる。

10

【0055】

図 10 は、クロスポイント・アレイを使用して前向き推論演算を実施するための既存の演算を示す。図 10 に示されるように、すべての入力平面 510 の 1 つのイメージ行 (512、514、および 516) が、クロスバー・システム 700 のクロスポイント・アレイ 705 のアレイ行 (802、804、806) に対する入力の列と同時に提示される。各クロスポイントのクロスポイント・デバイス 820 は、フィルタ 525 からの重み要素を含み、それぞれは、オームの法則 (電圧とコンダクタンスの積が電流に等しい) によってアレイ行励起  $x_i$  と記憶された重み  $w_{ij}$  との間の乗算となる。すべてのそのような読取り電流寄与の統合が、各アレイ列に沿って合計され、アレイ列 (808、810、812、および 814) の対応する積分器 (908、910、912、および 914) に記憶される。計算は、列 # 1 (808) 上の電流  $I_1$  がコンデンサ  $C_1$  (908) 上に記憶され、 $I_2$  がコンデンサ  $C_2$  上に記憶され、 $I_3$  が  $C_3$  上に記憶され、以下同様であると表現され得る。そのようなクロスポイント・アレイ 705 を使用する既存の技術的解決策では、コンデンサ (908、910、912、および 914) 上の統合電荷は乗算累積の出力として扱われ、次のアレイ 705 に送るためにデジタル数またはパルス持続時間のどちらかに変換される。

20

【0056】

このようにして、各時間ステップ (すなわち、アレイ 705 によって実施される各計算) で、すべての入力平面 510 にわたる値が統合され、すべての出力平面 530 についての出力が生成される。

30

【0057】

さらに、畳み込み層  $i$  からのあらゆる出力を、プーリングの部分として他の畳み込み層からの出力と組み合わせなければならない。出力をプールすべき他の畳み込み層は、フィルタ・カーネル 520 内の要素の数に依存する。代替または追加として、層  $i$  からのあらゆる出力を、畳み込み層  $i+1$  についての入力平面 510 内の様々なスポットに配置しなければならない。プーリングのためのそのような出力値の編成はまた、読取り-書き込みアクセス、電力などの追加のコンピューティング・リソースを必要とし得る。

【0058】

したがって、既存のシステムでは、時間ステップ 1 で、システム 700 が結果をコンデンサ 908、910、912、および 914 内に統合するが、直ちに結果を次の層に送らない。それは、システム 700 がいくつかの異なる列から統合コンデンサ 908、910、912、および 914 に対して電流を読み取ることをステアリングしなければならないからである。システム 700 は、後続の時間ステップで他の列からの結果のそのようなステアリングを実施する。同様に、システム 700 は、それぞれの  $k$  番目の出力行を計算するのに  $k$  個の時間ステップを必要とする。したがって、行単位マッピングを使用する既存の技術では、各出力行は、生成するのに  $k$  個の時間ステップを必要とする。

40

【0059】

図 10 は、既存の技術による、前向き推論中にアレイ 705 によって実施される演算を示す。図 10 では、時間ステップ 1、2、および 3 が示されている。各時間ステップでは

50

、入力がクロスポイント・アレイ 705 内の行にマッピングされる。各時間ステップで、積分器 (908、910、912、および 914) のそれぞれは、 $k \times p$  個の乗算累積項からの寄与を受け、ただし  $p$  は入力平面 510 の数である。 $k$  個のそのような時間ステップの後、積分器上の全電荷は、すべての  $k \times k \times p$  個の項を含み、次の畳み込み層に出力する準備ができています。最初の  $k$  個または最後の  $k$  個の時間ステップの間を除いて、各統合ステップの後、出力回路 720 からの  $k$  番目ごとの積分器がこのステータスに達し、したがって、畳み込み層出力の 1 つのイメージ行 (512-A、514-A、516-A) のすべての出力ピクセルを生成する準備ができています。すべての他の  $j$  番目の積分器は、 $j$  の値に応じて、それぞれの統合段階で異なる段階を有する。

【0060】

たとえば、図 10 に示されるように、順伝播の時間ステップ 1 で、各入力平面 512-A、514-A、516-A の第 1 の行が畳み込み層に入力される。図示されるように、クロスポイント・アレイ 705 のクロスポイント・デバイス 820 がフィルタ 520 と共にロードされる。具体的には、フィルタ・カーネル 522-A および 522-B がクロスポイント・デバイス 820 内にロードされ、第 1 の入力平面 516-A の第 1 の行と共に畳み込みが実施される。同様に、フィルタ・カーネル 520 の第 2 のバンクからのフィルタ・カーネル 524-A および 524-B が、第 2 の入力平面 514-A の第 1 の行と共に畳み込まれ、以下同様である。得られるそれぞれの畳み込みの結果が、出力コントローラ 1110 によって出力回路 720 から積分器 (908、910、912、914) のうちの 1 つまたは複数に転送される。

【0061】

出力コントローラ 1110 は、出力回路 720、または出力回路 720 に結合される外部コントローラの部分であり得る。出力コントローラ 1110 は、アレイ 705 内の各列からの乗算累積演算の出力を出力回路 720 内の特定の積分器にステアリングする。1 つまたは複数の例では、出力コントローラ 1110 は、各時間ステップでの各列についての積分器の選択を与えるモード信号を受け取る。代替として、出力コントローラ 1110 には、すべての畳み込み層が実行されるまで、各列についての積分器の選択を示すモード信号が提供される。1 つまたは複数の例では、モード信号は、各列についての選択された積分器を示すビット・パターンであり得る。

【0062】

図 10 の例では、時間ステップ 1 で、列 808 および 814 からの出力が、それぞれ積分器 908 および 912 内に記憶される。時間ステップ #2 で、入力平面 510 からの第 2 の行 512-B、514-B、および 516-B がクロスポイント・アレイ 705 への入力として使用される。クロスポイント・デバイス 820 は、時間ステップ #1 と同様に、カーネル・フィルタ 520 と共に依然としてロードされる (図 10)。時間ステップ 2 では、出力コントローラ 1110 は、列 810 および 816 (時間ステップ 1 とは異なる列) の出力のために同一の積分器 908 および 912 を選択する。したがって、このケースでは、積分器 908 および 912 (およびその他) が、異なる時間ステップで異なる列から出力を受け取る。

【0063】

時間ステップ 3 では、最初の 2 つの時間ステップと同様に、入力平面 510 からの第 3 の行 512-C、514-C、および 516-C が、クロスポイント・アレイ 705 への入力として使用される。時間ステップ 3 では、出力コントローラ 1110 が、列 812 および 818 (時間ステップ 1、2 とは異なる列) の出力のために同一の積分器 908 および 912 を選択する。したがって、このケースでは、積分器 908 および 912 (およびその他) が、異なる時間ステップで異なる列から出力を受け取る。このようにして、一般には、 $k$  個の時間ステップの後に、出力平面 530 内の行全体が計算される。

【0064】

出力平面 530 内の最初の出力行からの最初の 2 つのエントリ (A および B) の計算のみが上記で説明されたが、同様にして、出力平面 530 の他の部分が、クロスポイント・

10

20

30

40

50

アレイ 705 の他の部分によって並列に計算されることに留意されたい。さらに、図 10 に示されるように、クロスポイント・アレイ 705 は、他の積分器（910、914、916、および 918）を使用して、各時間ステップで他の出力行（C および D）について蓄積中であり得る。

【0065】

したがって、出力コントローラ 1110 がクロスポイント・アレイ 705 の出力をステアリングした結果として、すべての入力、すべての入力平面にわたって完全かつ連続するイメージ行の形である。さらに、何らかの出力が利用可能となる前の最初の  $k$  個の時間ステップの後、すなわち  $k+1$  番目の時間ステップから、すべての出力平面にわたる完全かつ連続するイメージ行が各時間ステップで生成される。したがって、そのような演算によって生成された出力マップ 530 が、ニューロン励起のどんな中間記憶も用いずに後続の畳み込み層にパイプライン処理され得る。合計、平均、最大などのプーリング演算はデータが到着するときにデータに対して増分式に実施され得るので、任意のプーリング演算は、出力イメージ行にとって十分な一時記憶のみを必要とする。こうした中間結果が記憶され、行単位プーリング演算が完了するまで、ニューロン励起の各セットが到着するときに更新され、行単位プーリング演算が完了した時点で、中間結果のバッファは実質的にプーリング層の出力である。

【0066】

前述のように、既存の技術に伴う技術的課題は、イメージなどの入力データセットの数、または実装されるネットワークのタイプと共に、CNN を実装するために必要とされるクロスポイント・アレイの数が増加し得ることである。本発明の実施形態は、CNN 深度を介する重みコピーの数を低減して、行単位マッピングを促進する。したがって、本発明の実施形態は、CNN 重み再利用因子の変化を調節するようにロード・バランシングを促進する。さらに、本発明の実施形態は、入力回路 710 から、クロスポイント・アレイ 705 まで、およびクロスポイント・アレイ 705 を介して、出力回路 720 までのデータの柔軟なルーティングを使用して、よりコンパクトな重みマッピングで、行単位マッピングを促進する。

【0067】

本発明の 1 つまたは複数の実施形態では、本明細書で説明される技術的解決策が、部分行入力を伴う行単位畳み込みを促進することによって既存の技術的解決策に伴うそのような技術的課題に対処し、入力データが時間的に区分化される。本発明の他の実施形態では、行単位畳み込みが部分行入力と共に促進され、入力データが空間的に区分化される（クロスポイント・アレイ）。

【0068】

図 11 は、部分行入力を伴う行単位畳み込みマッピングを示し、入力データが、本発明の 1 つまたは複数の実施形態に従って、時間的に区分化される。この場合、異なる入力行セグメントからの部分和が、コンデンサの別々のセット上に記憶される。図示される例では、第 1 の行からの入力データの第 1 のサブセット 1210 が、コンデンサ（または積分器）の第 1 のセット 1230 にマッピングされ、第 1 の行からの入力データの第 2 のサブセット 1220 が、コンデンサの第 2 のセット 1240 にマッピングされる。そのようなマッピングでは、パーティションが公式  $L = D * (\text{入力イメージ幅} / N + K - \text{ストライド})$  を使用して求められる。ただし  $N$  は、前向き推論のための計算のためにクロスポイント・アレイ 705 を再利用することを促進するために使用されるコンデンサのコピーの数である。 $N$  は、イメージ幅に基づいてあらかじめ決定され得る。たとえば、 $N$  は、重みコピーの数を低減して再利用因子を同一に保つために、イメージ・サイズが縮小するにつれて増加し得る。畳み込みニューラル・ネットワーク（CNN）では、重みカーネルが入力イメージにわたって畳み込まれ、すなわち、同一の重みが、出力を生成するために入力イメージの様々な部分で複数回再利用される。重みが再利用される回数は再利用因子と呼ばれる。

【0069】

さらに、ストライドは、第1の行のサブセット内で重複がどれほど存在するかを定義する所定のパラメータである。重複＝(k-ストライド)、ただしkはカーネル次元である。図示される例では、コンデンサの2つのセット1230および1240が、クロスポイント・アレイ705内に記憶される重みを再利用し得る。再利用が機能するために、クロスポイント・デバイス820によって行単位畳み込みが計算されるように入力データがマッピングされる。そのようなマッピングでは、計算されるLは、クロスポイント・アレイ705に入力される入力データ要素の数であり、各順次データ要素は順次入力平面からのものである。たとえば、図示される例示的シナリオでは、 $D=3$ つの入力平面および $L=15$ 、 $k=3$ 、ストライド=1、および $N=2$ で、 $L1=D1(1,1)$ 、 $L2=D2(1,1)$ 、 $L3=D3(1,1)$ である。ただし、 $D1(1,1)$ という表記は、 $D1$ の第1の行および第1の列からの要素を指す。同様に、 $L4=D1(1,2)$ 、 $L5=D2(1,2)$ 、および $L6=D3(1,2)$ である。クロスポイント・アレイ705は、 $N$ =カーネル520からの重みの2つのコピーと共に構成され、コピーは $D$ \*ストライド行(または列)だけ互いにオフセットする。

【0070】

本明細書で説明されるようにデータ要素が入力された後は、クロスポイント・デバイス820は、記憶された重みとデータ要素の積の求められた部分和に対するメモリ内計算を実施する。計算は、アナログ式にメモリ内で実施される。得られる部分和が、セット1230、1240内のコンデンサ内に記憶される。

【0071】

コンデンサのセット2130、2140のそれぞれの中のコンデンサの数は、重みのコピーの数が低下する場合に増加する。本発明の1つまたは複数の実施形態では、出力回路720内のコンデンサ面積の効率を改善するために、部分和が宛先コンデンサ(次の層のクロスポイント・アレイ705の入力側)に送られる。重複の結果、冗長な計算となるとしても、再利用によって達成されるクロスポイント・アレイ705の効率の改善の結果、CNNを実装するのに必要とされるクロスポイント・アレイ705の数が削減される。

【0072】

図12は、全行入力または部分行入力を伴う行単位畳み込みマッピングを示し、入力データが、本発明の1つまたは複数の実施形態に従って、時間的に区分化される。図示されるマッピング方式は、複数のクロスポイント・アレイ705にわたる複数の再利用される重みのコピーを使用することにより、さらにコンパクトなマッピングを促進する。本発明の1つまたは複数の実施形態では、出力イメージ・チャンネルの数がFであり、入力イメージ・チャンネルの数がDである。この場合、重みの各グループは、(入力イメージ幅\*D)次元に及ぶように( $D$ \*ストライド)のオフセットを有するストライドを有する。そのような(出力イメージ幅)個の重みのコピーのあらゆるセットの後、次のセットがどんなオフセット( $D$ \*ストライド)もなしに構成される。オフセットは各グループ内で使用され、そのグループ内の重みのコピーが分離される。そのような重みのコピーのグループは、別々のクロスポイント・アレイ705Aおよび705Bに及び得る。たとえば、図12に示される例では、グループ1280は、第1のクロスポイント・アレイ705A内に記憶される2つの重みのコピー1282および1284と、第2のクロスポイント・アレイ705B上に記憶される重みの第3のコピー1286とを有する。

【0073】

本明細書で説明される図および例の寸法は、本発明の1つまたは複数の実施形態で様々であり得ることを理解されたい。さらに、クロスポイント・アレイ705の数も、本発明の1つまたは複数の実施形態では、本明細書で説明された例から変化し得る。

【0074】

図13は、部分行入力を有する別の行単位畳み込みマッピングを示し、入力データは、本発明の1つまたは複数の実施形態に従って空間的に区分化される。この場合、コンデンサの単一のセット1320が、クロスポイント・アレイ705内に記憶されたカーネル重みに基づいて、得られる部分和を計算するために使用される。入力データ要素は、単一の

行のサブセットが所与のCNN層を実装している別々のクロスポイント・アレイ705に送られるように分割される。部分和を表す、コンデンサ1320上で蓄積する電荷が、CNNの次の層を実装しているシステム700の入力回路710に送られる。

【0075】

入力回路710は部分和を組み合わせ、次の層についての別々のクロスポイント・アレイ705内に記憶された重みについての入力データを編成することを含む。たとえば、入力回路710は、入力データ要素1320に対応する出力と同一の、次の層内のカーネル重みに、入力データ要素1310に対応する出力を向かわせる。

【0076】

(イメージ・サイズ) \* (#入力チャネル) がカーネル重みのサイズ (カーネル・サイズ \* #入力チャネル) と比べて大きいとき、前述のように入力回路710によって提供される柔軟なルーティングが、異なる出力チャネルについての重みを既存の解決策と比べて、よりコンパクトにマッピングすることを促進することが示され得る。次の層に対する入力の間の、ある層からの出力の記憶および再順序付けのためのコストは、存在している行単位マッピング技術よりも低い。したがって、本発明の1つまたは複数の実施形態は、既存の行単位マッピング技術のスケラビリティを改善するように柔軟な信号ルーティング方式を構成することを促進する。本発明の1つまたは複数の実施形態では、各ネットワークの特定の動作詳細に従ってCNNが微調整され得る。たとえば、CNNカーネル・サイズまたはCNNカーネルの数が、アナログ・クロスポイント・アレイへのマッピングをさらに最適化するように調節され得る。

【0077】

本明細書の図に示される行列の次元は単なる例であり、1つまたは複数の例では、異なる次元が使用され得ることに留意されたい。さらに、前向き推論演算の間にCNNが既にトレーニングされること、およびCNNをトレーニングするために使用される技術の如何に関わらず、本発明の実施形態が適用可能であることに留意されたい。

【0078】

このようにして、本発明の実施形態は、トレーニング済みCNNの前向き推論演算のための行単位マッピングを促進し、マッピングが、クロスポイント・アレイおよび支持回路が再利用するようにコンパクトな方式で実施され、任意のスケールのCNNの実装が促進され得る。

【0079】

本発明の1つまたは複数の実施形態では、記載の技術的解決策が、抵抗性メモリ素子のクロスポイント・アレイを含む電子回路によって実装される。アレイは、(i) アナログ入力値のベクトルを符号化するアレイに対する電圧入力ベクトルと、(ii) アレイ内のアナログ抵抗性重みの行列との間のアナログ・ベクトル-行列積に等しい電流出力のベクトルを提供する。電子回路700は、抵抗性メモリ素子の専用サブセットからの電流を集約する蓄積ワイヤおよび回路を共に含む、支持回路712、722と、入力回路710と、出力回路720とをさらに含む。支持回路722は統合コンデンサを含み、統合コンデンサのそれぞれが、単一の統合ステップの間に蓄積ワイヤのうちの1つから電流を集約するように電氣的に切換え可能である。出力回路720は、所定の数の統合ステップにわたって蓄積した統合コンデンサのサブセットからの統合電荷を、アナログ持続時間または2進数字を使用するデジタル表現のどちらかとして適切に変換し、送る。抵抗性メモリ素子は、畳み込みニューラル・ネットワークの所与の層のシナプス重みカーネルの列(行)を実装するように構成される。

【0080】

畳み込みニューラル・ネットワークの前記層に対する入力ニューロン励起が統合の反復ごとに1行(列)提示されるとき、所定の数の統合ステップにわたる蓄積は、前記重みカーネルの複数の部分行(列)にわたる乗算累積演算を実装する。本発明の1つまたは複数の実施形態では、第1の層の入力ニューロン励起が常に一度に1つの全行(列)提示され、入力ニューロン励起の後続の層が複数の部分行(列)に区分化され、ローカル・アナロ

10

20

30

40

50

グ・メモリ（たとえば、コンデンサ）内に部分的に記憶され、複数の統合サイクルにわたってクロスポイント・アレイ内で処理され得る。

【0081】

全出力励起または部分出力励起あるいはその両方を表す統合電荷が、前記重みカーネルのすべての行（列）が完全に統合された後にのみ、適切に変換され、送られる。複数のクロスバー・アレイからの部分和が、統合コンデンサのうちの1つで組み合わせられるように柔軟にルーティングされ、その後に全出力励起に変換され、次いですべての部分和が完全に統合された後に送られる。統合コンデンサ上の統合電荷は出力励起を表し、出力励起は適切に変換される。さらに、適切にプールされた結果（たとえば、前記出力励起の最大値、和、または平均）が局所的に計算され、次いですべての関連する重みカーネルが完全に統合された後にのみ送られる。

10

【0082】

この技術的解決策は、任意の可能な統合の技術的詳細レベルでのシステム、方法、またはコンピュータ・プログラム製品、あるいはその組合せであり得る。コンピュータ・プログラム製品は、この技術的解決策の態様をプロセッサに実施させるコンピュータ可読プログラム命令をその上に有するコンピュータ可読記憶媒体を含み得る。

【0083】

コンピュータ可読記憶媒体は、命令実行デバイスによる使用のために命令を保持および記憶し得る有形デバイスであり得る。コンピュータ可読記憶媒体は、たとえば、限定はしないが、電子記憶デバイス、磁気記憶デバイス、光記憶デバイス、電磁記憶デバイス、半導体記憶デバイス、または上記の任意の適切な組合せであり得る。コンピュータ可読記憶媒体のより具体的な例の非網羅的なリストには、ポータブル・コンピュータ・ディスクレット、ハード・ディスク、ランダム・アクセス・メモリ（RAM）、読取り専用メモリ（ROM）、消去可能プログラマブル読取り専用メモリ（EPROMまたはフラッシュ・メモリ）、静的ランダム・アクセス・メモリ（SRAM）、ポータブル・コンパクト・ディスク読取り専用メモリ（CD-ROM）、デジタル・バーサタイル・ディスク（DVD）、メモリ・スティック、フロッピー・ディスク、命令が記録されたパンチ・カードや溝の中の隆起構造などの機械的に符号化されたデバイス、および上記の任意の適切な組合せが含まれる。本明細書では、コンピュータ可読記憶媒体は、電波または他の自由伝播電磁波、導波路または他の伝送媒体を通じて伝播する電磁波（たとえば、光ファイバ・ケーブルを通過する光パルス）、ワイヤを通じて伝送される電気信号など、本質的に一時的信号であると解釈されるべきではない。

20

30

【0084】

本明細書で説明されるコンピュータ可読プログラム命令は、コンピュータ可読記憶媒体からそれぞれのコンピューティング／処理デバイスに、あるいはネットワーク、たとえばインターネット、ローカル・エリア・ネットワーク、広域ネットワーク、もしくはワイヤレス・ネットワーク、またはその組合せを介して外部コンピュータまたは外部記憶デバイスにダウンロードされ得る。ネットワークは、銅伝送ケーブル、光伝送ファイバ、ワイヤレス伝送、ルータ、ファイアウォール、スイッチ、ゲートウェイ・コンピュータ、またはエッジ・サーバ、あるいはその組合せを含み得る。各コンピューティング／処理デバイス内のネットワーク・アダプタ・カードまたはネットワーク・インターフェースが、ネットワークからコンピュータ可読プログラム命令を受信し、それぞれのコンピューティング／処理デバイス内のコンピュータ可読記憶媒体内に記憶するためにコンピュータ可読プログラム命令を転送する。

40

【0085】

この技術的解決策の動作を実施するためのコンピュータ可読プログラム命令は、アセンブラ命令、命令セット・アーキテクチャ（ISA）命令、機械語命令、マシン依存命令、マイクロコード、ファームウェア命令、状態設定データ、集積回路のための構成データ、あるいはSmalltalk、C++などのオブジェクト指向プログラミング言語と、「C」プログラミング言語や類似のプログラミング言語などの従来の手続型プログラミング

50

言語とを含む1つまたは複数のプログラミング言語の何らかの組合せで書かれたソース・コードまたはオブジェクト・コードであり得る。コンピュータ可読プログラム命令は、完全にユーザのコンピュータ上で、スタンド・アロン・ソフトウェア・パッケージとして部分的にユーザのコンピュータ上で、部分的にユーザのコンピュータ、および部分的にリモート・コンピュータ上で、または完全にリモート・コンピュータもしくはサーバ上で実行され得る。後者のシナリオでは、リモート・コンピュータは、ローカル・エリア・ネットワーク（LAN）または広域ネットワーク（WAN）を含む任意のタイプのネットワークを通じてユーザのコンピュータに接続され得、または接続が外部コンピュータに対して（たとえば、インターネット・サービス・プロバイダを使用してインターネットを通じて）行われ得る。いくつかの実施形態では、たとえばプログラマブル論理回路、フィールド・プログラマブル・ゲート・アレイ（FPGA）、またはプログラマブル論理アレイ（PLA）を含む電子回路が、この技術的解決策の態様を実施するために、コンピュータ可読プログラム命令の状態情報を利用して電子回路を個別化することによってコンピュータ可読プログラム命令を実行し得る。

#### 【0086】

この技術的解決策の態様が、技術的解決策の実施形態による方法、装置（システム）、およびコンピュータ・プログラム製品のフローチャート図またはブロック図あるいはその両方を参照して本明細書で説明される。フローチャート図またはブロック図あるいはその両方の各ブロック、フローチャート図またはブロック図あるいはその両方の中のブロックの組合せが、コンピュータ可読プログラム命令によって実装され得ることを理解されよう。

#### 【0087】

こうしたコンピュータ可読プログラム命令は、コンピュータまたは他のプログラム可能データ処理装置のプロセッサを介して実行される命令が、フローチャートまたはブロック図あるいはその両方の1つまたは複数のブロックで指定される機能／動作を実装するための手段を生み出すように、汎用コンピュータ、専用コンピュータ、または他のプログラム可能データ処理装置のプロセッサに与えられ、マシンが作り出され得る。こうしたコンピュータ可読プログラム命令はまた、命令を記憶するコンピュータ可読記憶媒体がフローチャートまたはブロック図あるいはその両方の1つまたは複数のブロックで指定される機能／動作の態様を実装する命令を含む製造品を含むように、コンピュータ、プログラム可能データ処理装置、または他のデバイス、あるいはその組合せに特定の方式で機能するように指示し得るコンピュータ可読記憶媒体内に記憶され得る。

#### 【0088】

コンピュータ可読プログラム命令はまた、コンピュータ、他のプログラム可能装置、または他のデバイス上で実行される命令が、フローチャートまたはブロック図あるいはその両方の1つまたは複数のブロックで指定される機能／動作を実装するように、コンピュータ、他のプログラム可能データ処理装置、または他のデバイス上にロードされ、コンピュータ、他のプログラム可能装置、または他のデバイス上で一連の動作ステップを実施させて、コンピュータ実装プロセスが生成され得る。

#### 【0089】

図中のフローチャートおよびブロック図は、この技術的解決策の様々な実施形態によるシステム、方法、およびコンピュータ・プログラム製品の可能な実装のアーキテクチャ、機能、および動作を示す。この点で、フローチャートまたはブロック図の各ブロックは、指定の論理的機能を実装するための1つまたは複数の実行可能命令を含む命令のモジュール、セグメント、または部分を表し得る。いくつかの代替実装では、ブロック内に記載の機能は、図に記載されている以外の順序で行われ得る。たとえば、連続して示される2つのブロックは、実際にはほぼ同時に実行され得、またはブロックは、関係する機能に応じて、時には逆の順序で実行され得る。ブロック図またはフローチャート図あるいはその両方の各ブロック、およびブロック図またはフローチャート図あるいはその両方のブロックの組合せが、指定の機能または動作を実施し、あるいは専用ハードウェアおよびコンピュータ命令の組合せを実施する専用ハードウェア・ベースのシステムによって実装され得る

ことにも留意されよう。

【0090】

第1の動作から直接的または間接的に第2の動作が生じるかどうかとは無関係に、第2の動作が第1の動作「に」応答すると言われることがある。第2の動作は、第1の動作よりも実質的に後の時間に行われ得、依然として第1の動作に応答し得る。同様に、介在する動作が第1の動作と第2の動作との間で行われる場合であっても、かつ介在する動作のうちの1つまたは複数が直接的に第2の動作を実施させる場合であっても、第2の動作は第1の動作に応答すると言われることがある。たとえば、第1の動作がフラグをセットし、フラグがセットされるときにはいつでも後に第3の動作が第2の動作を開始する場合、第2の動作は第1の動作に応答し得る。

10

【0091】

語句の使用を明白にし、本明細書によって公衆に通知を与えるために、「<A>、<B>、... および<N>のうちの少なくとも1つ」、または「<A>、<B>、... <N>のうちの少なくとも1つ、またはそれらの組合せ」、または「<A>、<B>、... および/または<N>」という語句は、別段に明白に表明されていない限り、本明細書の前または後で任意の他の示唆される定義に優先して、A、B、... およびNを含むグループから選択された1つまたは複数の要素を意味するように最も広い意味で解釈されるべきである。言い換えれば、この語句は、任意の1つの要素のみ、または列挙されていない追加の要素も組み合わせて含み得る他の要素のうちの1つまたは複数と組み合わせた1つの要素を含む要素A、B、... およびNのうちの1つまたは複数の任意の組合せを意味する。

20

【0092】

命令を実行する本明細書で例示される任意のモジュール、ユニット、構成要素、サーバ、コンピュータ、端末、またはデバイスが、記憶媒体、コンピュータ記憶媒体、たとえば磁気ディスク、光ディスク、テープなどのデータ記憶デバイス（取外し可能または取外し不能あるいはその両方）などのコンピュータ可読媒体を含み、またはコンピュータ可読媒体にアクセスでき得ることも理解されよう。コンピュータ記憶媒体は、コンピュータ可読命令、データ構造、プログラム・モジュール、他のデータなどの、情報の記憶のための任意の方法または技術で実装された揮発性および不揮発性の取外し可能および取外し不能媒体を含み得る。そのようなコンピュータ記憶媒体は、デバイスの部分であり得、またはデバイスにアクセス可能もしくは接続可能であり得る。本明細書で説明された任意のアプリケーションまたはモジュールは、そのようなコンピュータ可読媒体によって記憶され、あるいは保持され得るコンピュータ可読／実行可能命令を使用して実装され得る。

30

【0093】

本明細書の技術的特徴の様々な実施形態の説明が例示のために提示されたが、網羅的なものでなく、開示される実施形態に限定されないものとする。記載の実施形態の範囲から逸脱することなく、多くの修正形態および変形形態が当業者には明らかとなるであろう。本明細書で使用される用語は、実施形態の原理、市場で見出される技術に勝る実際の応用または技術的改善を最良に説明するように、あるいは当業者が本明細書で開示される実施形態を理解することを可能にするように選ばれた。

40

【0094】

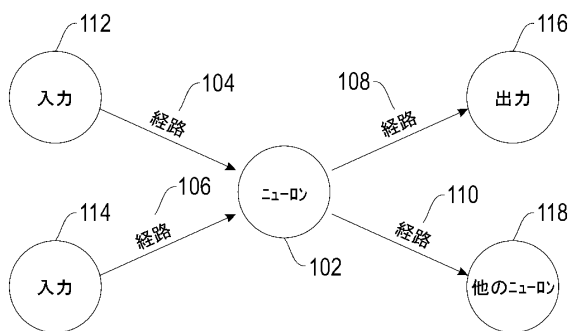
本明細書で説明される本発明の好ましい実施形態では、抵抗性メモリ素子のアレイであって、(i) アナログ入力値のベクトルを符号化するアレイに対する電圧入力のベクトルと、(ii) アレイ内のアナログ抵抗性重みの行列との間のアナログ・ベクトル-行列積に等しい電流出力のベクトルを提供する、アレイと、統合コンデンサであって、統合コンデンサのそれぞれが、単一の統合ステップの間に複数の蓄積ワイヤのうちの1つからの電流を集約するように電気的に切換え可能である、統合コンデンサと、統合電荷を蓄積する統合コンデンサに部分出力励起をルーティングすることによって抵抗性メモリ素子の専用サブセットからの電流を集約する蓄積ワイヤおよび回路と、複数の統合ステップにわたって蓄積された統合コンデンサのサブセットからの統合電荷を、アナログ持続時間または2

50

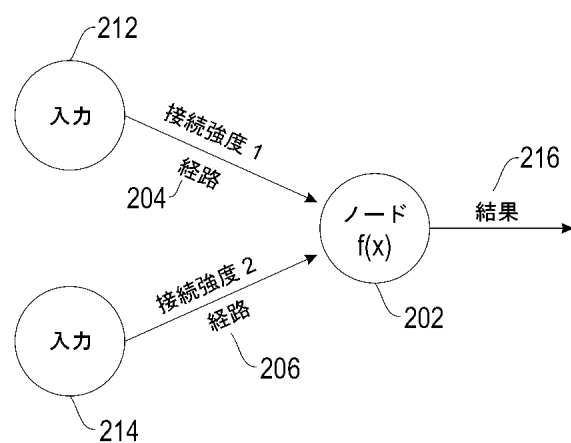
進数字を使用するデジタル表現のどちらかとして適切に変換し、送ることを可能にするためのデータ出力回路とを備える電子回路が提供され、抵抗性メモリ素子は、畳み込みニューラル・ネットワークの所与の層のシナプス重みカーネルのベクトルを実装するように構成される。好ましくは抵抗性メモリ素子は不揮発性メモリ・デバイスである。抵抗性メモリ素子のサブセットは、アレイの1つまたは複数の列に対応し得る。抵抗性メモリ素子のサブセットは、アレイの1つまたは複数の行に対応し得る。本明細書で説明される本発明の一実施形態では、このパラグラフの上記で説明したような回路を使用してトレーニング済み畳み込みニューラル・ネットワーク（CNN）の計算を実施するための方法であって、アナログ入力値の各ベクトルを複数の部分ベクトルに区分化することと、複数の部分ベクトルのそれぞれに対応する部分出力励起をアナログ・メモリ内に蓄積することと、統合電荷を蓄積する統合コンデンサに部分出力励起をルーティングすることによって部分出力励起を組み合わせることと、複数の出力励起を表す複数の統合コンデンサ上の統合電荷を送ることを含む操作のセットを所定の回数にわたって反復することによってクロスポイント・アレイの抵抗性メモリ素子による計算を実施することを含む方法が提供される。好ましくは、複数の統合コンデンサ上の統合電荷は、統合電荷を送る前に局所的にプールされた結果である。クロスポイント・デバイスは、畳み込みニューラル・ネットワークの所与の層の畳み込みカーネルの1つまたは複数の行を実装するように構成され得、入力データは、一度に1列ずつ提示される畳み込みニューラル・ネットワークの前記層に対するニューロン励起を表す。クロスポイント・デバイスは、畳み込みニューラル・ネットワークの所与の層の畳み込みカーネルの1つまたは複数の列を実装するように構成され得、入力データのベクトルは、入力データから一度に1行ずつ提示される畳み込みニューラル・ネットワークの所与の層に対するニューロン励起を表す。

【図面】

【図 1】

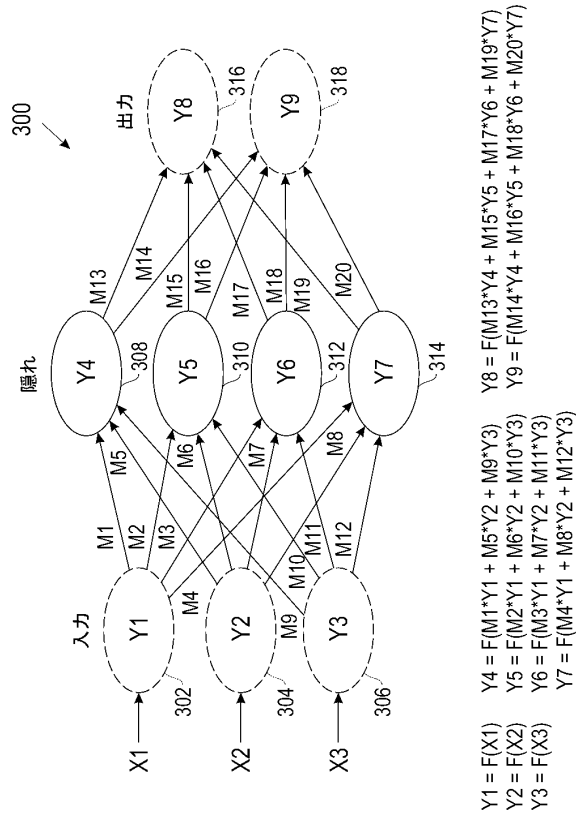


【図 2】

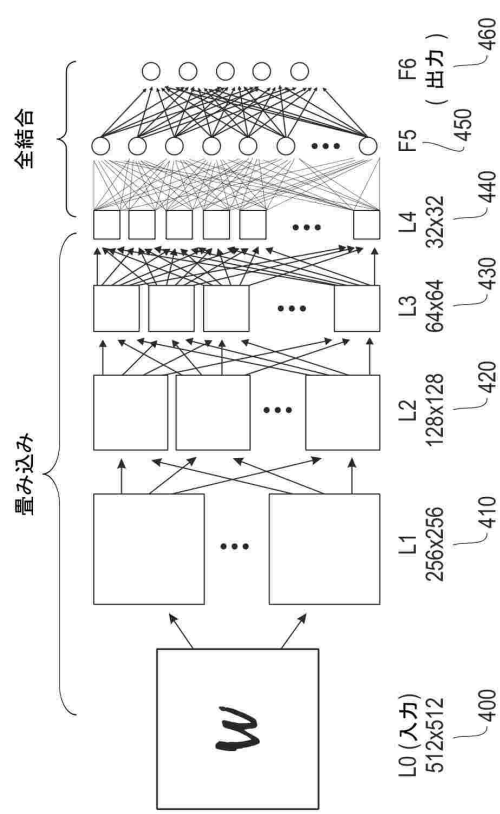


$$f(x)=f(\text{入力 } 1 \times \text{接続強度 } 1 + \text{入力 } 2 \times \text{接続強度 } 2)$$

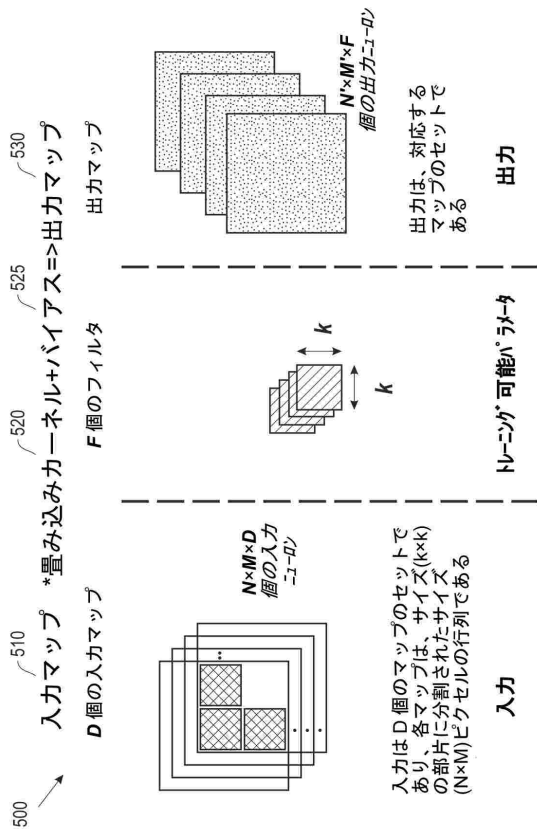
【図 3】



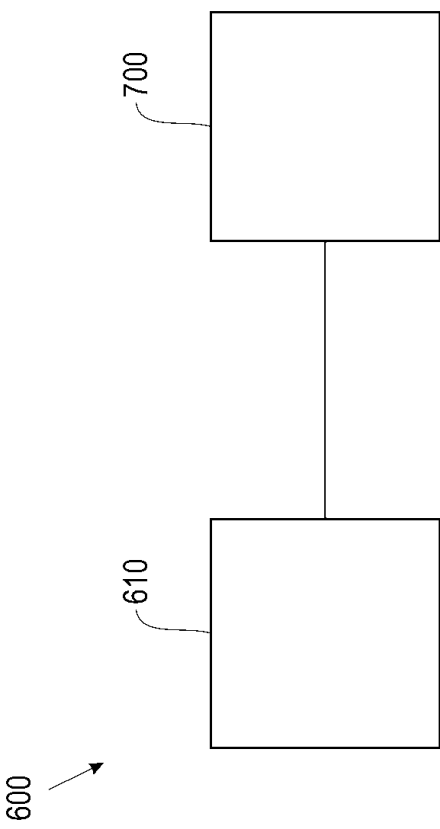
【図 4】



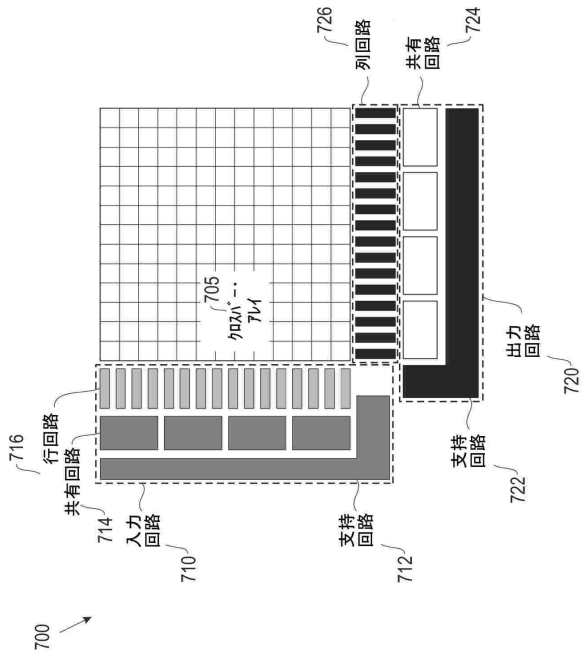
【図 5】



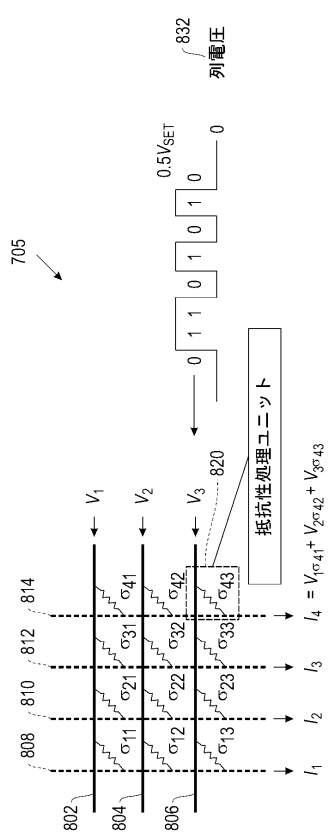
【図 6】



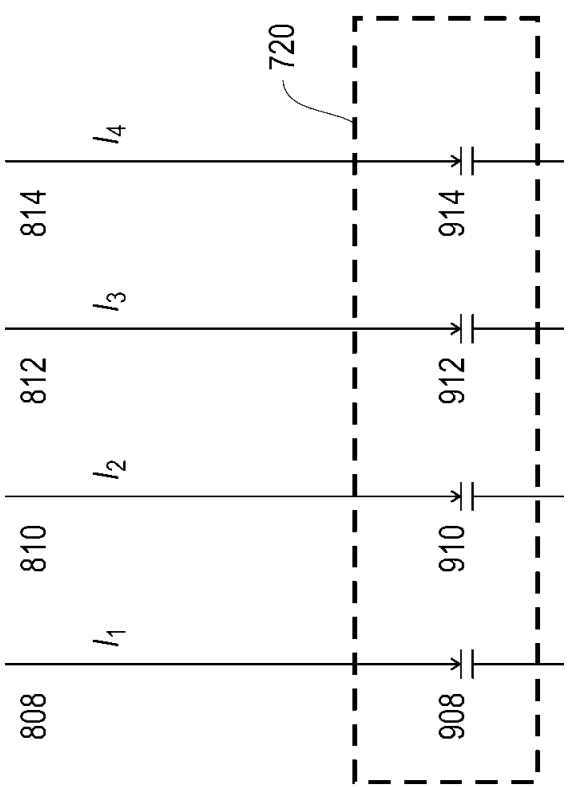
【図 7】



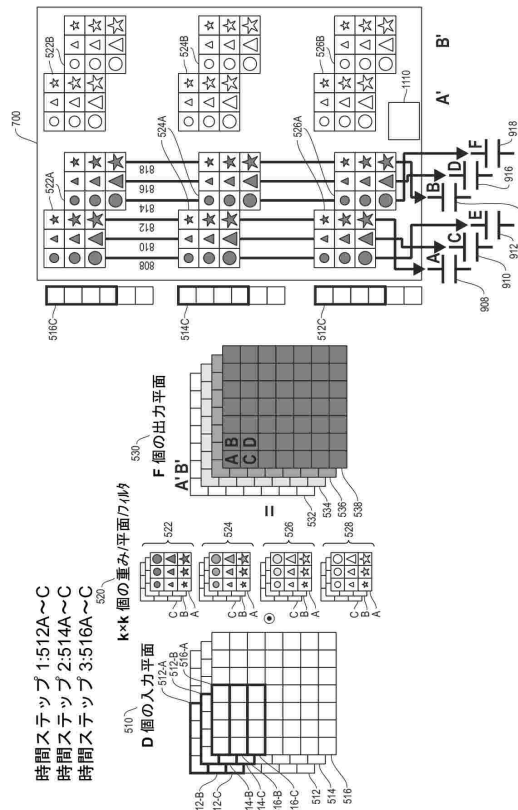
【図 8】



【図 9】

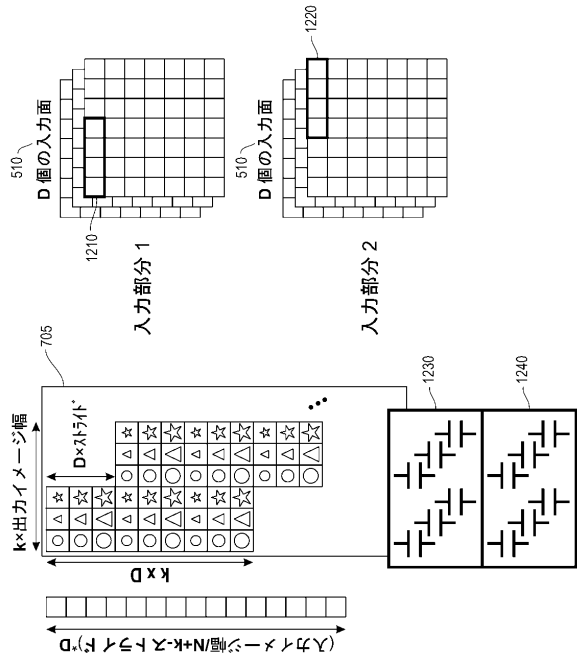


【図 10】

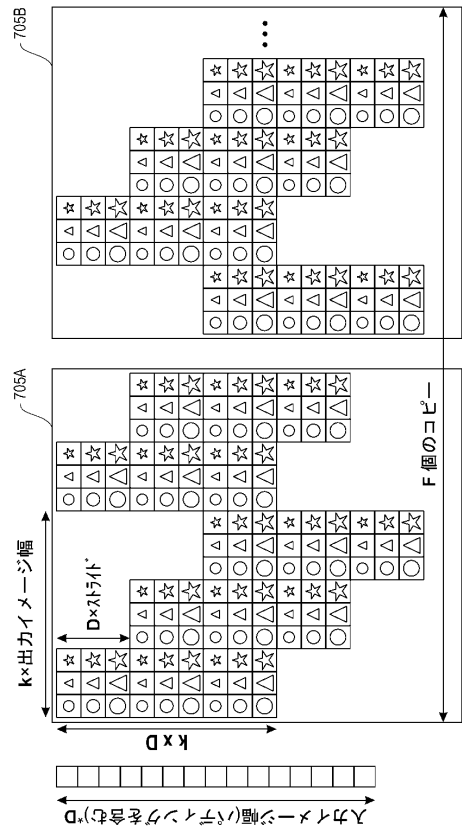


(従来技術)

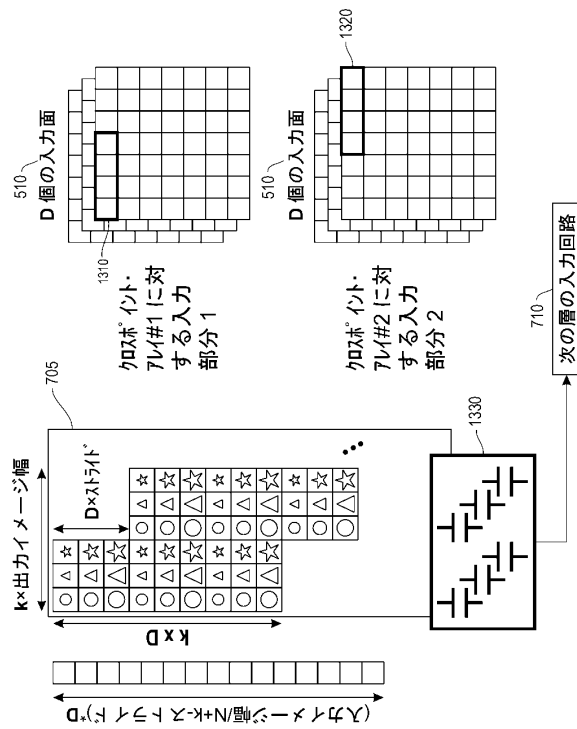
【図 1 1】



【図 1 2】



【図 1 3】



10

20

30

40

50

## フロントページの続き

- (74)代理人 100120710  
弁理士 片岡 忠彦
- (72)発明者 ツアイ、シンユ  
アメリカ合衆国 9 5 1 2 0 - 6 0 9 9 カリフォルニア州サンノゼ アルマデン・リサーチ・センタ  
ー ハリー・ロード 6 5 0
- (72)発明者 バール、ジェフリー  
アメリカ合衆国 9 5 1 2 0 - 6 0 9 9 カリフォルニア州サンノゼ アルマデン・リサーチ・センタ  
ー ハリー・ロード 6 5 0
- (72)発明者 ナラヤナン、プリティッシュ  
アメリカ合衆国 9 5 1 2 0 - 6 0 9 9 カリフォルニア州サンノゼ アルマデン・リサーチ・センタ  
ー ハリー・ロード 6 5 0
- 審査官 多賀 実
- (56)参考文献 米国特許出願公開第 2 0 2 0 / 0 1 1 7 9 8 6 (US, A 1)
- (58)調査した分野 (Int.Cl., DB名)  
G 0 6 N 3 / 0 0 - 9 9 / 0 0  
G 0 6 G 7 / 6 0