(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6987860号
(P6987860)
(45)発行日　令和4年1月5日(2022.1.5)　(24)登録日　令和3年12月3日(2021.12.3)

(51)Int.Cl.　　F I
G06N 3/08 (2006.01)　　G06N 3/08
G06N 3/063 (2006.01)　　G06N 3/063

請求項の数 13　(全 28 頁)

(21)出願番号　特願2019-524156(P2019-524156)
(86)(22)出願日　平成29年8月23日(2017.8.23)
(65)公表番号　特表2019-537139(P2019-537139A)
(43)公表日　令和1年12月19日(2019.12.19)
(86)国際出願番号　PCT/US2017/048123
(87)国際公開番号　WO2018/089079
(87)国際公開日　平成30年5月17日(2018.5.17)
審査請求日　令和2年5月7日(2020.5.7)
(31)優先権主張番号　15/348,199
(32)優先日　平成28年11月10日(2016.11.10)
(33)優先権主張国・地域又は機関
米国(US)
(31)優先権主張番号　15/467,382
(32)優先日　平成29年3月23日(2017.3.23)
(33)優先権主張国・地域又は機関
米国(US)

(73)特許権者　502208397
グーグル　エルエルシー
Ｇｏｏｇｌｅ　ＬＬＣ
アメリカ合衆国　カリフォルニア州　９４０４３　マウンテン　ビュー　アンフィシアター　パークウェイ　１６００
１６００　Ａｍｐｈｉｔｈｅａｔｒｅ　Ｐａｒｋｗａｙ　９４０４３　Ｍｏｕｎｔａｉｎ　Ｖｉｅｗ，　ＣＡ　Ｕ．Ｓ．Ａ．
(74)代理人　110001195
特許業務法人深見特許事務所
(72)発明者　ヤング，レジナルド・クリフォード
アメリカ合衆国、９４０４３　カリフォルニア州、マウンテン・ビュー、アンフィシアター・パークウェイ、１６００
最終頁に続く

(54)【発明の名称】ハードウェアにおけるカーネルストライドの実行

(57)【特許請求の範囲】
【請求項１】
コンピュータにより実現される方法であって、
専用ハードウェア回路上に畳み込みニューラルネットワークを実装する要求を受け取り、前記畳み込みニューラルネットワークを用いて、前記専用ハードウェア回路に命令を実行させることによって、ニューラルネットワーク入力を受け取って処理することを備え、前記畳み込みニューラルネットワークは、１より大きいストライドを有する第１の畳み込みニューラルネットワーク層を含み、前記専用ハードウェア回路は、ニューラルネットワーク計算を実行するための集積回路であり、ベクトル-行列乗算を実行するようにされた行列計算ユニットと、前記行列計算ユニットの出力に対してプーリングを実行するようにされたプーリング回路を含むベクトル計算ユニットとを含み、前記方法はさらに、
前記専用ハードウェア回路によって実行されると、前記専用ハードウェア回路に、前記畳み込みニューラルネットワークによる入力テンソルの処理中に、動作を実行することによって前記第１の畳み込みニューラルネットワーク層の出力と等価な層出力テンソルを生成させる命令を生成することを備え、前記動作は、
前記行列計算ユニットが、前記第１の畳み込みニューラルネットワーク層への前記入力テンソルを、１に等しいストライドを有するがそれ以外は前記第１の畳み込みニューラルネットワーク層と同等である第２の畳み込みニューラルネットワーク層を用いて処理することにより、第１のテンソルを生成することと、
前記ベクトル計算ユニットが、前記第２の畳み込みニューラルネットワーク層が前記第

１の畳み込みニューラルネットワーク層のストライドを有する場合には生成されなかったであろう前記第１のテンソルの要素を零出力して、第２のテンソルを生成することと、
前記ベクトル計算ユニットの前記プーリング回路が、前記第２のテンソルに対して最大プーリングを実行して前記層出力テンソルを生成することとを含み、
前記ベクトル計算ユニットが前記第１のテンソルの要素を零出力することは、
前記ベクトル計算ユニットがマスキングテンソルと前記第１のテンソルとの要素ごとの乗算を実行して前記第２のテンソルを生成することを含み、前記マスキングテンソルは、（ｉ）前記第２の畳み込みニューラルネットワーク層が前記第１の畳み込みニューラルネットワーク層のストライドを有する場合には生成されなかったであろう前記第１のテンソルの要素に対応する前記マスキングテンソルの各要素位置において０を含み、（ｉｉ）前記マスキングテンソルの各他の要素位置において１を含む、コンピュータにより実現される方法。

【請求項２】
前記ベクトル計算ユニットが前記第１のテンソルの要素を零出力することは、
前記第１のテンソルの要素のサブセットに０を乗算することと、
前記サブセットに含まれていない前記第１のテンソルの要素に１を乗算することとを含み、
前記サブセットは、前記第２の畳み込みニューラルネットワーク層が前記第１の畳み込みニューラルネットワーク層のストライドを有する場合には生成されなかったであろう前記第１のテンソルの要素に対応する前記マスキングテンソルの各要素からなる、請求項１に記載の方法。

【請求項３】
前記マスキングテンソルは、前記専用ハードウェア回路によってアクセス可能なメモリに格納される、請求項１または２に記載の方法。

【請求項４】
コンピュータにより実現される方法であって、
専用ハードウェア回路上に畳み込みニューラルネットワークを実装する要求を受け取り、前記畳み込みニューラルネットワークを用いて、前記専用ハードウェア回路に命令を実行させることによって、ニューラルネットワーク入力を受け取って処理することを備え、前記畳み込みニューラルネットワークは、１より大きいストライドを有する第１の畳み込みニューラルネットワーク層を含み、前記専用ハードウェア回路は、ニューラルネットワーク計算を実行するための集積回路であり、ベクトル－行列乗算を実行するようにされた行列計算ユニットと、前記行列計算ユニットの出力に対してプーリングを実行するようにされたプーリング回路を含むベクトル計算ユニットとを含み、前記方法はさらに、
前記専用ハードウェア回路によって実行されると、前記専用ハードウェア回路に、前記畳み込みニューラルネットワークによる入力テンソルの処理中に、動作を実行することによって前記第１の畳み込みニューラルネットワーク層の出力と等価な層出力テンソルを生成させる命令を生成することを備え、前記動作は、
前記行列計算ユニットが、前記第１の畳み込みニューラルネットワーク層への前記入力テンソルを、１に等しいストライドを有するがそれ以外は前記第１の畳み込みニューラルネットワーク層と同等である第２の畳み込みニューラルネットワーク層を用いて処理することにより、第１のテンソルを生成することと、
前記ベクトル計算ユニットが、前記第２の畳み込みニューラルネットワーク層が前記第１の畳み込みニューラルネットワーク層のストライドを有する場合には生成されなかったであろう前記第１のテンソルの要素を零出力して、第２のテンソルを生成することと、
前記ベクトル計算ユニットの前記プーリング回路が、前記第２のテンソルに対して最大プーリングを実行して前記層出力テンソルを生成することとを含み、
前記ベクトル計算ユニットが前記第１のテンソルの要素を零出力することは、
前記ベクトル計算ユニットが、第１のマスキングテンソルと前記第１のテンソルとの要素ごとの乗算を実行して、修正された第１のテンソルを生成することを含み、前記第１の

マスキングテンソルは、（ｉ）前記第２の畳み込みニューラルネットワーク層が前記第１の畳み込みニューラルネットワーク層のストライドを有する場合には生成されなかったであろう前記第１のテンソルの要素に対応する前記第１のマスキングテンソルの各要素位置において０を含み、（ｉｉ）前記第２の畳み込みニューラルネットワーク層が前記第１の畳み込みニューラルネットワーク層のストライドを有する場合には生成されたであろう前記第１のテンソルの要素に対応する前記第１のマスキングテンソルの各要素位置においてそれぞれの非０値を含み、前記ベクトル計算ユニットが前記第１のテンソルの要素を零出力することはさらに、

前記ベクトル計算ユニットが第２のマスキングテンソルと前記修正された第１のテンソルとの要素ごとの乗算を実行することを含み、前記第２のマスキングテンソルは、前記第２の畳み込みニューラルネットワーク層が前記第１の畳み込みニューラルネットワーク層のストライドを有する場合に生成されるであろう前記第１のテンソルの要素に対応する各要素位置において、前記第１のマスキングテンソルの前記それぞれの非０値の逆数を含む、コンピュータにより実現される方法。

【請求項５】

前記第１のマスキングテンソルおよび前記第２のマスキングテンソルは、前記専用ハードウェア回路によってアクセス可能なメモリに格納される、請求項４に記載の方法。

【請求項６】

１より大きい複数のストライドにそれぞれ対応するように、複数のマスキングテンソルが前記メモリに格納され、

前記方法はさらに、前記複数のマスキングテンソルの中から、前記第１の畳み込みニューラルネットワーク層のストライドに対応するマスキングテンソルを選択することを備える、請求項５に記載の方法。

【請求項７】

前記ベクトル計算ユニットの前記プーリング回路が最大プーリングを実行することは、前記第１の畳み込みニューラルネットワーク層のストライドによって定義される前記第２のテンソルの１つまたは複数のウィンドウの各々について、前記ウィンドウ内の要素の最大値要素を取得することを含む、請求項１～６のいずれか１項に記載の方法。

【請求項８】

前記第２のテンソルの前記１つまたは複数のウィンドウの各々は、前記第１の畳み込みニューラルネットワーク層のストライドに対応する次元を有する矩形ウィンドウであり、前記第２のテンソルの異なる要素を含む、請求項７に記載の方法。

【請求項９】

前記ベクトル計算ユニットの前記プーリング回路が最大プーリングを実行することは、前記第２のテンソルの要素の１つまたは複数のサブセットの各々について、前記サブセットの最大値要素を取得することを含む、請求項１～８のいずれか１項に記載の方法。

【請求項１０】

前記第１の畳み込みニューラルネットワーク層は、前記畳み込みニューラルネットワーク内の複数の畳み込みニューラルネットワーク層のうちの第１のニューラルネットワーク層であり、前記入力テンソルは、デジタル画像の、前記デジタル画像の画素に対応する要素を含む表現である、請求項１～９のいずれか１項に記載の方法。

【請求項１１】

前記入力テンソルは前記専用ハードウェア回路のユニファイドバッファに格納され、前記第２の畳み込みニューラルネットワーク層の重みは前記専用ハードウェア回路のダイナミックメモリに格納され、前記第２の畳み込みニューラルネットワーク層を用いて前記第１の畳み込みニューラルネットワーク層への前記入力テンソルを処理することは、

前記入力テンソルを前記ユニファイドバッファから前記行列計算ユニットに送ることと、

前記ダイナミックメモリから前記行列計算ユニットに前記第２の畳み込みニューラルネットワーク層の前記重みを送ることと、

前記行列計算ユニットによって、前記第２の畳み込みニューラルネットワーク層の前記重みを用いて前記入力テンソルを処理して、前記第１のテンソルを生成することとを含む、請求項１～１０のいずれか１項に記載の方法。

【請求項１２】

システムであって、

請求項１に記載の専用ハードウェア回路と、

命令を格納する１つまたは複数の記憶装置とを備え、前記命令は、前記専用ハードウェア回路によって実行されると、前記専用ハードウェア回路に請求項１～１１のいずれか１項に記載の方法を実行させるよう動作可能である、システム。

【請求項１３】

１つまたは複数のコンピュータによって実行されると、前記１つまたは複数のコンピュータに請求項１～１１のいずれか１項に記載の方法を実行させる命令を含むコンピュータプログラム。

【発明の詳細な説明】

【背景技術】

【０００１】

背景

本明細書は、ハードウェアにおけるニューラルネットワーク推論の計算に関する。

【０００２】

ニューラルネットワークは、受け取られた入力に対する出力、たとえば分類を生成するために１つまたは複数の層を用いる機械学習モデルである。いくつかのニューラルネットワークは、出力層に加えて１つまたは複数の隠れ層を含む。各隠れ層の出力は、ネットワーク内の別の層、たとえばネットワークの次の隠れ層または出力層への入力として用いられる。ネットワークの各層は、それぞれのパラメータセットの現在の値に従って、受け取られた入力から出力を生成する。

【発明の概要】

【課題を解決するための手段】

【０００３】

概要

一般に、この明細書はニューラルネットワーク推論を計算する専用ハードウェア回路について記載する。

【０００４】

一般に、この明細書に記載される主題の１つの革新的な局面は、ハードウェア回路上でニューラルネットワークを処理するよう要求を受け取ることを備え、ニューラルネットワークは、１より大きいストライドを有する第１の畳み込みニューラルネットワーク層を含み、さらに、これに応答して、ハードウェア回路によって実行されると、ハードウェア回路に、ニューラルネットワークによる入力テンソルの処理中に、以下の動作を実行することによって第１の畳み込みニューラルネットワーク層の出力と等価な層出力テンソルを生成させる命令を生成することを備える方法およびシステムにおいて実施され得る。この動作は、第１の畳み込みニューラルネットワーク層への入力テンソルを、１に等しいストライドを有するがそれ以外は第１の畳み込みニューラルネットワーク層と同等である第２の畳み込みニューラルネットワーク層を用いて処理することにより、第１のテンソルを生成することと、第２の畳み込みニューラルネットワーク層が第１の畳み込みニューラルネットワーク層のストライドを有する場合には生成されなかったであろう第１のテンソルの要素を零出力して、第２のテンソルを生成することと、第２のテンソルに対して最大プーリングを実行して層出力テンソルを生成することとを含む。

【０００５】

実装形態は、以下の特徴の１つ以上を含み得る。いくつかの実装形態では、第１のテンソルの要素を零出力することは、第１のテンソルの要素のサブセットに０を乗算することと、サブセットに含まれていない第１のテンソルの要素に１を乗算することとを含む。第

１のテンソルの要素を零出力することは、マスキングテンソルと第１のテンソルとの要素ごとの乗算を実行して第２のテンソルを生成することを含み、マスキングテンソルは、（ｉ）第２の畳み込みニューラルネットワーク層が第１の畳み込みニューラルネットワーク層のストライドを有する場合には生成されなかったであろう第１のテンソルの要素に対応するマスキングテンソルの各要素位置において０を含み、（ｉｉ）マスキングテンソルの各他の要素位置において１を含む。いくつかの実装形態では、マスキングテンソルは、ハードウェア回路によってアクセス可能なメモリに格納され、マスキングテンソルと第１のテンソルとの要素ごとの乗算は、ハードウェア回路に含まれる、ハードウェアで実装されるベクトル計算ユニットによって実行される。

【０００６】

実装形態は、さらに、以下の特徴の１つ以上を含み得る。いくつかの実装形態では、第１のテンソルの要素を零出力することは、第１のマスキングテンソルと第１のテンソルとの要素ごとの乗算を実行して、修正された第１のテンソルを生成することを含み、第１のマスキングテンソルは、（ｉ）第２の畳み込みニューラルネットワーク層が第１の畳み込みニューラルネットワーク層のストライドを有する場合には生成されなかったであろう第１のテンソルの要素に対応するマスキングテンソルの各要素位置において０を含み、（ｉｉ）第２の畳み込みニューラルネットワーク層が第１の畳み込みニューラルネットワーク層のストライドを有する場合には生成されたであろう第１のテンソルの要素に対応するマスキングテンソルの各要素位置においてそれぞれの非０値を含み、第１のテンソルの要素を零出力することはさらに、第２のマスキングテンソルと修正された第１のテンソルとの要素ごとの乗算を実行することを含み、第２のマスキングテンソルは、第２の畳み込みニューラルネットワーク層が第１の畳み込みニューラルネットワーク層のストライドを有する場合に生成されるであろう第１のテンソルの要素に対応する各要素位置において、第１のマスキングテンソルのそれぞれの非０値の逆数を含む。

【０００７】

実装形態は、さらに、以下の特徴の１つ以上を含み得る。いくつかの実装形態では、最大プーリングを実行することは、第１の畳み込みニューラルネットワーク層のストライドによって定義される第２のテンソルの１つまたは複数のウィンドウの各々について、ウィンドウ内の要素の最大値要素を取得することを含む。第２のテンソルの１つまたは複数のウィンドウの各々は、畳み込みニューラルネットワーク層のストライドに対応する次元を有する矩形ウィンドウであり、第２のテンソルの異なる要素を含む。いくつかの実装形態では、最大プーリングを実行することは、第２のテンソルの要素の１つまたは複数のサブセットの各々について、サブセットの最大値要素を取得することを含む。第２のテンソル上で実行される最大プーリングは、ハードウェア回路のプーリング回路によって実行される。畳み込みニューラルネットワーク層は、ニューラルネットワーク内の第１のニューラルネットワーク層であり、入力テンソルは、デジタル画像の、当該デジタル画像の画素に対応する要素を含む表現である。

【０００８】

実装形態は、さらに、以下の特徴の１つ以上を含み得る。いくつかの実装形態では、入力テンソルはハードウェア回路のユニファイドバッファに格納され、第２の畳み込みニューラルネットワーク層の重みはハードウェア回路のダイナミックメモリに格納され、第２の畳み込みニューラルネットワーク層を用いて第１の畳み込みニューラルネットワーク層への入力テンソルを処理することは、入力テンソルをユニファイドバッファからハードウェアで実装されるハードウェア回路の行列計算ユニットに送ることと、ダイナミックメモリからハードウェア回路の行列計算ユニットに第２の畳み込みニューラルネットワーク層の重みを送ることと、ハードウェア回路の行列計算ユニットによって、第２の畳み込みニューラルネットワーク層の重みを用いて入力テンソルを処理して、第１のテンソルを生成することとを含む。

【０００９】

この明細書において記載される主題の特定の実施形態は、以下の利点の１つ以上を実現

するように実現することができる。ハードウェア回路が１より大きいストライドを有する畳み込みニューラルネットワークを用いて入力テンソルを直接処理することができない場合でも、１より大きいストライドを有する畳み込みニューラルネットワーク層に対応する出力テンソルを、専用ハードウェア回路によってハードウェアで生成できる。専用ハードウェア回路を用いて適切な出力を生成することにより、１より大きいストライドを有するニューラルネットワーク層の処理は、たとえ専用ハードウェア回路がそのような処理を直接サポートしていなくても、データをホストコンピュータに返送することなく、すなわち計算の少なくとも一部をオフチップで実行することなく、実行できる。これにより、専用ハードウェア回路のハードウェアアーキテクチャを変更することなく、１より大きいストライドを有する畳み込み層を含むニューラルネットワークの推論を効率的に判断することが可能になる。すなわち、処理の一部をオフチップで、ソフトウェアで、またはその両方で実行することから生じる処理遅延が回避される。

【００１０】

本明細書に記載される主題は、たとえば、ニューラルネットワーク推論を計算するときにカーネルストライドを実行するための、開示された技法およびハードウェアを用いる画像認識または分類方法およびシステムにも関する。

【００１１】

この明細書において記載される主題の１つ以上の実施形態の詳細は、添付の図面および以下の詳細な説明において述べられる。主題の他の特徴、局面および利点は、詳細な説明、図面および特許請求の範囲から明らかになる。

【図面の簡単な説明】

【００１２】

【図１】例示的なニューラルネットワーク処理システムを示す。

【図２】ニューラルネットワークの所与の層について計算を実行するための例示的な方法の流れ図である。

【図３】例示的なニューラルネットワーク処理システムを示す。

【図４】行列計算ユニットを含む例示的なアーキテクチャを示す。

【図５】シストリックアレイ内のセルの例示的アーキテクチャを示す。

【図６】ベクトル計算ユニットの例示的アーキテクチャを示す。

【図７】プーリング回路のための例示的なアーキテクチャを示す。

【図８】１より大きいストライドでニューラルネットワークの所与の層に対して計算を実行するようにニューラルネットワーク処理システムに命令するための例示的な方法の流れ図である。

【図９】１より大きいストライドを有するニューラルネットワークの所与の層に対して計算を実行するための例示的な方法の流れ図である。

【図１０】１より大きいストライドでのニューラルネットワークの所与の層に対する計算の例である。

【発明を実施するための形態】

【００１３】

さまざまな図面中の同様の参照番号および指定は、同様の要素を示す。

詳細な説明

複数の層を有するニューラルネットワークを用いて推論を計算することができる。たとえば、入力が与えられると、ニューラルネットワークはその入力に対する推論を計算することができる。ニューラルネットワークは、ニューラルネットワークの各層を通って入力を処理することによって、この推論を計算する。各層は入力を受け取り、その層に対する重みのセットに従って入力を処理して出力を生成する。

【００１４】

したがって、受け取った入力から推論を計算するために、ニューラルネットワークは入力を受け取り、それを各ニューラルネットワーク層の各々を通して処理して推論を生成し、１つのニューラルネットワーク層からの出力は次のニューラルネットワーク層への入力

として与えられる。ニューラルネットワーク層へのデータ入力、たとえば、ニューラルネットワークへの入力、またはシーケンス内におけるその層の下の層の、あるニューラルネットワーク層への出力は、その層への活性化入力と呼ぶことができる。

【００１５】

いくつかの実装形態では、ニューラルネットワークの層はシーケンスで配置される。ある他の実装形態では、層は有向グラフとして配される。つまり、任意の特定の層が複数の入力、複数の出力、またはそれらの両方を受け取ることができる。ニューラルネットワークの層は、ある層の出力を前の層への入力として送り返すことができるように構成することもできる。

【００１６】

いくつかのニューラルネットワークは、１つまたは複数のニューラルネットワーク層からの出力をプーリングして、後続のニューラルネットワーク層への入力として用いられるプーリングされた値を生成する。いくつかの実装形態では、ニューラルネットワークは、出力のグループの最大値、最小値、または平均値を判断し、最大値、最小値、または平均値をグループのプーリングされた出力として用いることによって、出力のグループをプーリングする。出力をプーリングすることは、空間的不変性をある程度維持することができるので、さまざまな構成で配置される出力を、同じ推論を有するように処理することができる。出力をプーリングすることはまた、プーリングする前の出力の所望の特性を維持しながら後続のニューラルネットワーク層で受け取られる入力の次元を低減することができ、それはニューラルネットワークによって生成される推論の品質を著しく損なうことなく効率を改善できる。

【００１７】

いくつかのニューラルネットワークは、１より大きいストライドを有する１つまたは複数の畳み込みニューラルネットワーク層を含む。概念的には、１のストライドの場合、畳み込みニューラルネットワーク層は、重みのセットを活性化入力に順次適用することができる。すなわち、活性化入力アレイの場合、重みを活性化入力のサブセットに適用し、畳み込み計算が完了するまで、活性化入力の各他のサブセットに、１つの位置、たとえば行または列だけ、移動させることができる。１より大きい整数であるストライドを有する畳み込みニューラルネットワーク層の場合、重みを活性化入力のサブセットに適用し、畳み込み計算が完了するまで、活性化入力の各他のサブセットに、ストライドに等しい数の位置だけ、たとえば、ストライドによって示される行または列の数だけ、移動させることができる。

【００１８】

本明細書は、ニューラルネットワーク層を処理し、任意選択で１つまたは複数のニューラルネットワーク層の出力に対してプーリングを実行する専用ハードウェア回路を記載する。専用ハードウェア回路は、１のストライドを有するニューラルネットワーク層を処理することができる回路を含む。専用ハードウェア回路は、１より大きいストライドを有するニューラルネットワーク層の処理を直接サポートしないが、専用ハードウェア回路は、１より大きいストライドを有するニューラルネットワーク層の出力と等価の出力を生成するように制御され得る。したがって、開示された技術の１つの技術的効果および利点は、１のストライドを有するニューラルネットワーク層を処理することができる回路を、より柔軟な態様で、１より大きいストライドを有するニューラルネットワーク層についてニューラルネットワーク推論を計算するために用いることができることである。

【００１９】

図１は、例示的なニューラルネットワーク処理システム１００を示す。ニューラルネットワーク処理システム１００は、以下に記載されるシステム、コンポーネント、および技術が実装され得る１つまたは複数の位置に１つまたは複数コンピュータとして実装されるシステムの例である。

【００２０】

ニューラルネットワーク処理システム１００は、専用ハードウェア回路１１０を用いて

ニューラルネットワーク計算を実行するシステムである。ハードウェア回路１１０は、ニューラルネットワーク計算を実行するための集積回路であり、ハードウェアでベクトル - 行列乗算を実行する行列計算ユニット１２０を含む。ハードウェア回路１１０はまた、行列計算ユニット１２０の出力に対してプーリングを実行するためのプーリング回路を含むベクトル計算ユニット１４０を含む。例示的な専用ハードウェア回路１２０が、図３を参照して以下により詳細に記載される。

【００２１】

特に、ニューラルネットワーク処理システム１００は、専用ハードウェア回路１１０上にニューラルネットワークを実装する要求を受信し、専用ハードウェア回路１１０上にニューラルネットワークを実装し、所与のニューラルネットワークが実装されると、ニューラルネットワーク推論を生成するために専用集積回路１１０を用いてニューラルネットワークへの入力を処理する。

【００２２】

すなわち、ニューラルネットワーク処理システム１００は、入力を処理するために用いられるべきニューラルネットワークのためのニューラルネットワークアーキテクチャを指定する要求を受け取ることができる。ニューラルネットワークアーキテクチャは、ニューラルネットワーク内の層の数および構成、ならびにパラメータを有する各層のパラメータの値を定義する。

【００２３】

専用集積回路１１０上にニューラルネットワークを実装するために、ニューラルネットワーク処理システム１００は、１つまたは複数の物理的位置にある１つまたは複数のコンピュータ上の１つまたは複数のコンピュータプログラムとして実装されるニューラルネットワーク実装エンジン１５０を含む。

【００２４】

ニューラルネットワーク実装エンジン１５０は命令を生成し、命令は、専用ハードウェア回路１１０によって実行されると、ハードウェア回路１１０に、ニューラルネットワークによって指定される動作を実行させて、受け取られたニューラルネットワーク入力からニューラルネットワーク出力を生成させる。

【００２５】

命令がニューラルネットワーク実装エンジン１５０によって生成され、ハードウェア回路１１０に与えられると、ニューラルネットワーク処理システム１００は、ニューラルネットワーク入力を受け取り、ニューラルネットワークを用いて、ハードウェア回路１１０に、生成された命令を実行させることによって、ニューラルネットワーク入力を処理することができる。

【００２６】

しかしながら、いくつかのニューラルネットワークは、１つまたは複数の互換性のないニューラルネットワーク層を含む。本明細書で用いられる互換性のないニューラルネットワーク層という用語は、専用ハードウェア回路１１０によってハードウェアで直接実行することができない操作を指定するニューラルネットワーク層を指す。ハードウェア回路１１０上にこれらのニューラルネットワークを実装するために、ニューラルネットワーク実装エンジン１５０は、ハードウェア回路１１０によって実行されると、ハードウェア回路１１０に、ハードウェアにおいて以下の操作を実行することによって互換性のないニューラルネットワーク層についての出力を生成させる命令を生成する。それらの操作は、ニューラルネットワーク層によって指定されるものとは異なる操作であるが、互換性のないニューラルネットワーク層の仕様を満たす層出力、たとえば層出力テンソル、つまり、層によって指定される操作を直接実行することによって生成されたであろう出力と同じ層出力が生成される結果となる。

【００２７】

特に、いくつかのニューラルネットワークは、１より大きいストライドを有する畳み込みニューラルネットワーク層を含む。そのようなニューラルネットワーク層は、入力テン

ソルを用いて非順次的に処理される１つまたは複数のカーネルを特徴とする。たとえば、１のストライドでカーネルストライドを実行するとき、カーネルは入力テンソルの要素に順次適用される。しかしながら、２のストライドでカーネルストライドを実行するとき、ニューラルネットワーク層のカーネルは、カーネルの特定の要素が入力テンソルの１つおきの要素に適用されて出力テンソルを生成するように、シフトされる。出力テンソルは、ニューラルネットワークの別の層により入力として用いることができる。

【００２８】

ハードウェア回路１１０上で行列演算を実行する主ハードウェアユニットは行列計算ユニット１２０であるので、集積回路は１より大きいストライドを有するニューラルネットワーク層を直接計算することはできない。１より大きいストライドを有する層を含むニューラルネットワークを実現するために、ニューラルネットワーク実現エンジン１５０は命令を生成し、命令は、ニューラルネットワークによるニューラルネットワーク入力の処理中に専用ハードウェア回路１１０によって実行されると、ハードウェア回路１１０にハードウェアで他の操作を実行させて、行列乗算ユニット１２０およびプーリング回路を特徴とするベクトル計算ユニット１４０を用いて、１より大きいストライドを有するニューラルネットワーク層の仕様を満たす出力テンソルを生成する。これらの命令および他の操作は、図７～図１０を参照して以下により詳細に説明される。

【００２９】

図２は、専用ハードウェア回路を用いてニューラルネットワークの所与の層について計算を実行するための例示的なプロセス２００の流れ図である。便宜上、方法２００は、方法２００を実行する１つまたは複数の回路を有するシステムに関して説明される。方法２００は、受け取られた入力から推論を計算するために、ニューラルネットワークの各層に対して実行することができる。

【００３０】

システムは、所与の層について重み入力のセット（ステップ２０２）および活性化入力のセット（ステップ２０４）を受け取る。重み入力のセットおよび活性化入力のセットは、専用ハードウェア回路のダイナミックメモリおよびユニファイドバッファからそれぞれ受け取ることができる。いくつかの実装形態では、重み入力のセットと活性化入力のセットとの両方をユニファイドバッファから受け取ることができる。

【００３１】

システムは、専用ハードウェア回路の行列乗算ユニットを用いて、重み入力および活性化入力から累積値を生成する（ステップ２０６）。いくつかの実装形態では、累積値は、重み入力のセットと活性化入力のセットとの内積である。すなわち、層内のすべての重みのサブセットである１セットの重みについて、システムは各重み入力を各活性化入力と乗算し、それらの積を合計して累積値を形成することができる。システムは、次いで、他のセットの重みと他のセットの活性化入力との内積を計算することができる。いくつかの実装形態では、専用ハードウェア回路は、特定のニューラルネットワーク層のストライド、すなわちニューラルネットワーク層が１のストライドまたは１より大きいストライドを有するかどうかにかかわらず、同様にそのような操作を実行し得る。行列乗算ユニットからの出力のその後の処理は、ニューラルネットワーク層が１より大きい指定されたストライドで処理された場合に生成されるであろう出力と等価な出力を生成するよう実行することができる。

【００３２】

システムは、専用ハードウェア回路のベクトル計算ユニットを用いて累積値から層出力を生成することができる（ステップ２０８）。いくつかの実装形態では、ベクトル計算ユニットは、累積値に活性化関数を適用する。これについては、図５を参照して以下でさらに説明する。層の出力は、ニューラルネットワーク内の次の層への入力として用いるためにユニファイドバッファに格納することができ、または推論を決めるために用いることができる。いくつかの実装形態では、ニューラルネットワーク層は、１より大きいストライドを指定し得、システムは、１より大きいストライドを有するニューラルネットワーク層

の出力と等価である層出力を得るために、累積値に対して追加の処理を行い得る。受け取られた入力がニューラルネットワークの各層を介して処理されて、受け取られた入力に対する推論を生成すると、システムはニューラルネットワークの処理を終了する。

【００３３】

図３は、ニューラルネットワーク計算を実行するための例示的な専用ハードウェア回路３００を示す。システム３００はホストインターフェース３０２を含む。ホストインターフェース３０２は、ニューラルネットワーク計算のためのパラメータを含む命令を受け取ることができる。パラメータは、以下のうちの１つまたは複数を含むことができる。処理すべき層の数、モデルの各層に対する対応する重み入力のセット、活性化入力の初期セット、すなわち推論が計算されるニューラルネットワークへの入力、各層の対応する入力および出力サイズ、ニューラルネットワーク計算のためのストライド値、および処理されるべき層のタイプ、たとえば畳み込み層または全結合層。

【００３４】

ホストインターフェース３０２は、命令をシーケンサ３０６に送ることができ、シーケンサ３０６は、命令を、ニューラルネットワーク計算を実行するように回路を制御する低レベル制御信号に変換する。いくつかの実装形態では、制御信号は、回路内のデータフロー、たとえば重み入力のセットおよび活性化入力のセットが回路をどのように流れるか、を調整する。シーケンサ３０６は、制御信号をユニファイドバッファ３０８、行列計算ユニット３１２、およびベクトル計算ユニット３１４に送ることができる。いくつかの実装形態では、シーケンサ３０６はまた、ダイレクトメモリアクセスエンジン３０４およびダイナミックメモリ３１０に制御信号を送る。いくつかの実装形態では、シーケンサ３０６は制御信号を生成するプロセッサである。シーケンサ３０６は、適切なときに制御信号を回路３００の各構成要素に送るために、制御信号のタイミングを用いることができる。いくつかの他の実装形態では、ホストインターフェース３０２は外部プロセッサから制御信号を渡す。

【００３５】

ホストインターフェース３０２は、重み入力のセットおよび活性化入力の初期セットをダイレクトメモリアクセスエンジン３０４に送ることができる。ダイレクトメモリアクセスエンジン３０４は、ユニファイドバッファ３０８に活性化入力のセットを格納することができる。いくつかの実装形態では、ダイレクトメモリアクセスは、メモリユニットであり得るダイナミックメモリ３１０に重みのセットを格納する。いくつかの実装形態では、ダイナミックメモリ３１０は回路の外に配置されている。

【００３６】

ユニファイドバッファ３０８はメモリバッファである。それは、ダイレクトメモリアクセスエンジン３０４からの活性化入力のセットおよびベクトル計算ユニット３１４の出力を格納するために用いることができる。ベクトル計算ユニット３１４は、図６を参照して以下により詳細に説明される。ダイレクトメモリアクセスエンジン３０４は、ユニファイドバッファ３０８からベクトル計算ユニット３１４の出力を読み出すこともできる。

【００３７】

ダイナミックメモリ３１０およびユニファイドバッファ３０８は、重み入力のセットおよび活性化入力のセットをそれぞれ行列計算ユニット３１２に送ることができる。いくつかの実装形態では、行列計算ユニット３１２は二次元シストリックアレイである。行列計算ユニット３１２は、数学的演算、たとえば乗算および加算を実行することができる一次元シストリックアレイまたは他の回路とすることもできる。いくつかの実装形態では、行列計算ユニット３１２は汎用の行列プロセッサである。

【００３８】

行列計算ユニット３１２は、重み入力および活性化入力を処理し、出力のベクトルをベクトル計算ユニット３１４に与えることができる。いくつかの実装形態では、行列計算ユニット３１２は、出力ベクトルをユニファイドバッファ３０８に送り、ユニファイドバッファ３０８は、出力ベクトルをベクトル計算ユニット３１４に送る。ベクトル計算ユニッ

ト３１４は、出力のベクトルを処理し、処理された出力のベクトルをユニファイドバッファ３０８に格納することができる。１より大きいストライドを有するニューラルネットワーク層の場合、ベクトル計算ユニット３１４は、出力のベクトルを処理して、１より大きいストライドを有するニューラルネットワーク層の出力と等価である層出力テンソルを生成し、層出力テンソルをユニファイドバッファ３０８に格納することができる。処理された出力のベクトルは、たとえばニューラルネットワーク内の後続の層で用いるために、行列計算ユニット３１２への活性化入力として用いることができる。行列計算ユニット３１２およびベクトル計算ユニット３１４は、図４および図６をそれぞれ参照して以下により詳細に説明される。

【００３９】

図４は、行列計算ユニットを含む例示的アーキテクチャ４００を示す。行列計算ユニットは、二次元シストリックアレイ４０６である。アレイ４０６は複数のセル４０４を含む。いくつかの実装形態では、シストリックアレイ４０６の第１の次元４２０はセルの列に対応し、シストリックアレイ４０６の第２の次元４２２はセルの行に対応する。シストリックアレイは、列よりも多い行、行よりも多い列、または同数の列と行とを有することができる。

【００４０】

図示の例では、値ローダ４０２は活性化入力をアレイ４０６の行に送り、重みフェッチャインターフェース（weight fetcher interface）４０８は重み入力をアレイ４０６の列に送る。しかしながら、いくつかの他の実装形態では、活性化入力は列に転送され、重み入力はアレイ４０６の行に転送される。

【００４１】

値ローダ４０２は、ユニファイドバッファ、たとえば、図３のユニファイドバッファ３０８から、活性化入力を受け取ることができる。各値ローダは、対応する活性化入力をアレイ４０６の最も左側の異なるセルに送ることができる。たとえば、値ローダ４１２はセル４１４に活性化入力を送ることができる。

【００４２】

重みフェッチャインターフェース４０８は、メモリユニット、たとえば図３のダイナミックメモリ３１０から重み入力を受け取ることができる。重みフェッチャインターフェース４０８は、対応する重み入力をアレイ４０６の最も上の異なるセルに送ることができる。たとえば、重みフェッチャインターフェース４０８は、重み入力をセル４１４および４１６に送ることができる。重みフェッチャインターフェース４０８はさらに、メモリユニット、たとえばダイナミックメモリ３１０から複数の重みを受け取り、複数の重みをアレイ４０６の最も上の別個のセルに並列に送ることができる。たとえば、重みフェッチャインターフェース４０８は、異なる重みをセル４１４および４１６に同時に送ることができる。

【００４３】

いくつかの実装形態では、ホストインターフェース、たとえば、図３のホストインターフェース３０２は、活性化入力をアレイ４０６全体にわたって１つの次元に沿って、たとえば右にシフトしながら、重み入力をアレイ４０６全体にわたって別の次元に沿って、たとえば下にシフトする。たとえば、１クロックサイクルにわたって、セル４１４における活性化入力は、セル４１４の右にあるセル４１６の活性化レジスタにシフトすることができる。同様に、セル４１６における重み入力は、セル４１４の下にあるセル４１８における重みレジスタにシフトすることができる。

【００４４】

各クロックサイクルで、各セルは、所与の重み入力、所与の活性化入力、および隣接セルからの累積出力を処理して、累積出力を生成することができる。累積出力は、与えられた重み入力と同じ次元に沿って隣接セルに渡すこともできる。各セルは、隣接セルからの累積出力を処理することなく、所与の重み入力および所与の活性化入力を処理して出力を生成することもできる。出力は、累積されることなく、与えられた重み入力および出力と

同じ次元に沿って隣接セルに渡されることができる。個々のセルは、図５を参照して以下にさらに説明される。

【００４５】

いくつかの実装形態では、恒等行列、すなわち、主対角線上に１および他の場所に０を有する行列をアレイ４０６に渡すことができ、それによって値ローダ４０２で与えられる入力を修正なしでアキュムレータ４１０に渡すことができる。これは、２つの入力の要素ごとの乗算を実行するために用いられ得、ここで、アキュムレータでの第１の出力は、output = MatMul(input1, identity)として表すことができ、MatMulは、行列計算ユニットが行列乗算を実行するための命令であり、要素ごとの乗算結果に対応する第２の出力は、output *= MatMul(input2, identity)として表される。*=演算、すなわち演算output = output * MatMul(input2, identity)を実行するために、アーキテクチャ４００は、 +=または*=計算を実行するための構成要素を含み得る。+=または*=演算を実行するための構成要素は、アキュムレータ４１０の前、すなわちセル４０４の最後の行の後に配置することができる。いくつかの実装形態では、図３のベクトル計算ユニット３１４が、+=または*=演算を実行するための構成要素を含んでもよく、すなわち、その場合、ベクトル計算ユニット３１４が、要素ごとの乗算を実行するために、output = output * MatMul(input2, identity)演算を実行する。

【００４６】

累積出力は、重み入力と同じ列に沿って、たとえばアレイ４０６内の列の一番下に向かって渡すことができる。いくつかの実装形態では、各列の一番下において、アレイ４０６は、行よりも多い活性化入力を有する層の計算を実行するときに各列から出力される各累積出力を格納および累積する、アキュムレータユニット４１０を含み得る。いくつかの実装形態では、各アキュムレータユニットは複数の並列累積値を格納する。アキュムレータユニット４１０は、各累積出力を累積して最終累積値を生成することができる。最終累積値はベクトル計算ユニット、たとえば図６のベクトル計算ユニットに転送することができる。いくつかの他の実装形態では、アキュムレータユニット４１０は、行よりも少ない活性化入力を有する層を処理するときに累積を実行せずに累積値をベクトル計算ユニットに渡す。

【００４７】

図５は、シストリックアレイ内のセル、たとえば図４のシストリックアレイ４０６のセル４１４、４１６、または４１８のうちの１つ、の例示的アーキテクチャ５００を示す。

【００４８】

セルは、活性化入力を格納する活性化レジスタ５０６を含み得る。活性化レジスタは、シストリックアレイ内のセルの位置に応じて、左側の隣接セル、すなわち所与のセルの左側に位置する隣接セルから、またはユニファイドバッファから、活性化入力を受け取ることができる。セルは、重み入力を格納する重みレジスタ５０２を含み得る。重み入力は、シストリックアレイ内のセルの位置に応じて、上の隣接セルまたは重みフェッチャインターフェースから転送され得る。セルは総和レジスタ５０４を含むこともできる。総和レジスタ５０４は、上の隣接セルからの累積値を格納することができる。乗算回路５０８を用いて、重みレジスタ５０２からの重み入力を活性化レジスタ５０６からの活性化入力と乗算することができる。乗算回路５０８は積を合計回路５１０に出力することができる。

【００４９】

合計回路５１０は、積と総和レジスタ５０４からの累積値とを合計して新たな累積値を生成することができる。次いで、合計回路５１０は、新たな累積値を、下の隣接セルに位置する別の総和レジスタに送ることができる。新たな累積値は、下の隣接セルにおける合計のためのオペランドとして用いることができる。合計回路５１０はまた、総和レジスタ５０４からの値を受け入れ、総和レジスタ５０４からの値を、乗算回路５０８からの積と合計することなく、下の隣接セルに送ることもできる。

【００５０】

セルは、重み入力および活性化入力を、処理のために、隣接セルにシフトすることもで

きる。たとえば、重み経路レジスタ５１２は、重み入力を下の隣接セル内の別の重みレジスタに送ることができる。活性化レジスタ５０６は、活性化入力を右の隣接セル内の別の活性化レジスタに送ることができる。したがって、重み入力と活性化入力との両方を、後続のクロックサイクルでアレイ内の他のセルによって再利用することができる。

【００５１】

いくつかの実装形態では、セルは制御レジスタも含む。制御レジスタは、セルが重み入力または活性化入力のいずれかを隣接セルにシフトすべきかを決定する制御信号を記憶することができる。いくつかの実装形態では、重み入力または活性化入力をシフトすることは、１つまたは複数のクロックサイクルを要する。制御信号は、活性化入力または重み入力が乗算回路５０８に転送されるかどうかも決定し得るか、または乗算回路５０８が活性化入力および重み入力で演算を行うかどうかも決定し得る。制御信号は、たとえば配線を用いて、１つまたは複数の隣接セルに渡すこともできる。

【００５２】

いくつかの実装形態では、重みは重み経路レジスタ５１２に事前にシフトされる。重み経路レジスタ５１２は、重み入力をたとえば上の隣接セルから受け取り、重み入力を制御信号に基づいて重みレジスタ５０２に転送することができる。重みレジスタ５０２は、活性化入力が複数のクロックサイクルにわたってたとえば活性化レジスタ５０６を介してセルに転送されるとき、重み入力がセル内に留まり、隣接セルに転送されないように、重み入力を静的に格納することができる。したがって、重み入力は、たとえば乗算回路５０８を用いて、複数の活性化入力に適用することができ、それぞれの累積値は隣接セルに転送することができる。

【００５３】

図６は、ベクトル計算ユニット６０２の例示的アーキテクチャ６００を示す。ベクトル計算ユニット６０２は、行列計算ユニット、たとえば図３を参照して説明した行列計算ユニット３１２または図４の行列計算ユニットのアキュムレータ４１０から、累積値のベクトルを受け取ることができる。

【００５４】

ベクトル計算ユニット６０２は、活性化ユニット６０４で累積値のベクトルを処理することができる。いくつかの実装形態では、活性化ユニットは、活性化値を生成するために各累積値に非線形関数を適用する回路を含む。たとえば、非線形関数はtanh(x)とすることができ、ここで、xは累積値である。

【００５５】

任意選択で、ベクトル計算ユニット６０２は、プーリング回路６０８を用いて、値、たとえば活性化値をプーリングすることができる。プーリング回路６０８は、プーリングされた値を生成するために値の１つまたは複数に集約関数を適用することができる。いくつかの実装形態では、集約関数は、値、または値のサブセットの、最大値、最小値、もしくは平均値を返す関数である。

【００５６】

制御信号６１０は、たとえば、図３のシーケンサ３０６によって転送することができ、ベクトル計算ユニット６０２がどのように累積値のベクトルを処理するかを調整することができる。すなわち、制御信号６１０は、活性化値がプーリングされるかどうかを調整することができ、その場合、活性化値はたとえばユニファイドバッファ３０８に格納され、またはそうでなければ、制御信号６１０は、活性化値の取り扱いを調整することができる。制御信号６１０は、活性化関数またはプーリング関数、および活性化値またはプーリング値、たとえばストライド値を処理するための他のパラメータを指定することもできる。

【００５７】

ベクトル計算ユニット６０２は、値、たとえば活性化値またはプーリングされた値を、ユニファイドバッファ、たとえば図３のユニファイドバッファ３０８に送ることができる。いくつかの実装形態では、プーリング回路６０８は、活性化値またはプーリングされた値を受け取り、活性化値またはプーリングされた値をユニファイドバッファに格納する。

【００５８】
図７は、プーリング回路のための例示的アーキテクチャ７００を示す。プーリング回路は、プーリングされた値を生成するために、１つまたは複数の活性化された値に集約関数を適用することができる。例示として、アーキテクチャ７００は、４×４セットの活性化された値のプーリングを実行することができる。図７に示されるプーリングは正方形の領域、すなわち４×４を有するが、長方形の領域も可能である。たとえば、領域がｎ×ｍのウィンドウを有する場合、アーキテクチャ７００はｎ＊ｍ個のレジスタ、すなわちｎ個の列およびｍ個の行を有することができる。
【００５９】
プーリング回路アーキテクチャ７００は、値のベクトルから、たとえば図６の活性化回路６０４から、要素のシーケンスを受け取ることができる。たとえば、シーケンスは画像の８×８部分の画素を表すことができ、プーリング回路アーキテクチャ７００は８×８部分の４×４サブセットからの値をプーリングすることができる。いくつかの実装形態では、プーリングされた値は、いったんプーリング回路アーキテクチャ７００によって計算されるとシーケンスに追加される。いくつかの実装形態では、ニューラルネットワークプロセッサは、複数の並列プーリング回路を含む。各クロックサイクルにわたって、各プーリング回路は活性化回路６０４から、値のベクトルからのそれぞれの要素を受け取ることができる。各プーリング回路は、活性化回路６０４から受け取った要素を、ラスタ順に到着する二次元画像として解釈することができる。
【００６０】
プーリング回路は、一連のレジスタおよびメモリユニットを含み得る。各レジスタは、レジスタ内部に格納されている値にわたって集約関数を適用する集約回路７０６に出力を送ることができる。集約関数は、値のセットから最小値、最大値、または平均値を返すことができる。
【００６１】
第１の値は、レジスタ７０２に送られてその内部に格納され得る。後続のクロックサイクルで、第１の値は後続のレジスタ７０８にシフトしてメモリ７０４に格納されることができ、第２の値はレジスタ７０２に送られてレジスタ７０２内に格納されることができる。
【００６２】
４クロックサイクル後、４つの値が最初の４つのレジスタ７０２、７０８～７１２の内部に格納される。いくつかの実装形態では、メモリユニット７０４は先入れ先出し（ＦＩＦＯ）の下で動作する。各メモリユニットは最大８つの値を格納できる。メモリユニット７０４が画素の完全な行を含んだ後、メモリユニット７０４はレジスタ７１４に値を送ることができる。
【００６３】
任意の所与の時点で、集約回路７０６は各レジスタからの値にアクセスすることができる。レジスタ内の値は、画像の４×４部分の値を表す。
【００６４】
プーリング回路は、集約回路７０６を用いることによって、アクセスされた値から、プーリングされた値、たとえば最大値、最小値、または平均値を生成することができる。プーリングされた値は、ユニファイドバッファ、たとえば図３のユニファイドバッファ３０８に送ることができる。
【００６５】
第１のプーリングされた値を生成した後、プーリング回路は、新たな値がレジスタに格納され集約回路７０６によってプーリングされることができるように、各レジスタを通して値をシフトすることによって、プーリングされた値を生成し続けることができる。たとえば、アーキテクチャ７００では、プーリング回路は値をさらに４クロックサイクルにわたってシフトすることができ、それによってメモリユニット内の値をレジスタにシフトする。いくつかの実施態様では、プーリング回路は、新たな値が最後の最上位レジスタ、た

とえばレジスタ７１６に格納されるまで、新たな値をシフトする。
【００６６】
次いで、集約回路７０６は、レジスタに格納されている新たな値をプーリングすることができる。新たな値をプーリングした結果は、ユニファイドバッファに格納できる。
【００６７】
図８は、１より大きいストライドでニューラルネットワークの所与の畳み込み層について計算を実行するための例示的なプロセス８００のフローチャートである。一般に、プロセス７００は、専用ハードウェア回路を含む１つまたは複数のコンピュータのシステムによって実行される。いくつかの実装形態では、例示的なプロセス８００は、図１のシステムによって実行され得る。
【００６８】
システムは、専用ハードウェア回路上にニューラルネットワークを実装するための要求を受信する（ステップ８０２）。特に、ニューラルネットワークは、１より大きいストライドを有する畳み込みニューラルネットワーク層を含む。要求は、さらに、ニューラルネットワークを用いて処理する入力、ニューラルネットワークによって生成された出力テンソルを格納するための位置、または他のパラメータなど、ニューラルネットワークを実装するための他のパラメータを指定し得る。
【００６９】
システムは、要求に基づいて、１より大きいストライドを有するニューラルネットワーク層を処理する際に用いられるべきマスキングテンソルを生成する（ステップ８０４）。たとえば、ニューラルネットワークを実施する要求およびニューラルネットワークへの入力を特定する情報を受け取ることに基づいて、システムは、１より大きいストライドを有するニューラルネットワーク層を処理するためのマスキングテンソルを生成する。
【００７０】
マスキングテンソルのサイズは、特定される入力の次元、または１より大きいストライドを有するニューラルネットワーク層への入力テンソルの予想サイズに基づいて、判断することができる。マスキングテンソルに含まれる値は、１より大きいストライドを有するニューラルネットワーク層の指定されたストライドに基づいて判断されてもよい。たとえば、ニューラルネットワーク層の指定されたストライドが４である場合、マスキングテンソルの４つおきの要素は１に設定され得る一方、マスキングテンソルの他のすべてのエントリは０に設定され得る。いくつかの実装形態では、ニューラルネットワークは、１より大きいストライドを有する複数の層を含み得、システムは、１より大きいストライドを有する層の各々について対応するマスキングテンソルを生成し得る。さらに、いくつかの実装形態では、システムは、たとえばメモリに、マスキング行列またはマスキング行列成分のライブラリを格納し、そのライブラリを用いることに基づいてマスキング行列を選択または生成することができる。
【００７１】
システムは、専用ハードウェア回路１１０によって実行されると、専用ハードウェア回路１１０に、ニューラルネットワークによる入力テンソルの処理中に、マスキングテンソルを用いて１より大きいストライドを有する畳み込みニューラルネットワーク層の出力と等価な層出力テンソルを生成させる命令を生成する（ステップ８０６）。たとえば、要求に応答して、ニューラルネットワーク実装エンジン１５０は、専用ハードウェア回路１１０に指示またはそれを制御して、専用ハードウェア回路１１０が１より大きいストライドを有する畳み込みニューラルネットワーク層を用いて入力テンソルを処理した場合に等価である出力テンソル、すなわち出力ベクトルを生成するよう、命令を生成することができる。
【００７２】
システムは命令およびマスキングテンソルを専用ハードウェア回路１１０に送る（ステップ８０８）。たとえば、ニューラルネットワーク実装エンジン１５０は、命令を専用ハードウェア回路１１０に与えることができ、専用ハードウェア回路１１０は、たとえば図

３のホストインターフェース３０２において命令を受け取ることができる。ニューラルネットワーク実装エンジン１５０は、ホストインターフェース３０２によっても受け取られ得る、ニューラルネットワーク計算のための他の命令および／またはパラメータも与え得る。

【００７３】

図９は、１より大きいストライドを有するニューラルネットワーク計算層を計算するための例示的なプロセス９００のフローチャートである。たとえば、プロセス９００は、ニューラルネットワーク実装エンジン１５０から受け取られる命令に基づいて、図１の専用ハードウェア回路１１０によって実行することができる。

【００７４】

たとえば、１より大きいストライドを有するニューラルネットワーク層を実装するための命令を受け取ると、ホストインターフェース３０２は命令を図３のシーケンサ３０６に送ることができ、シーケンサ３０６は、ニューラルネットワーク計算を実行するよう、命令を、図３の専用ハードウェア回路３００を制御する低レベル制御信号に変換することができる。

【００７５】

受け取られた命令に基づいて、専用ハードウェア回路３００は、畳み込みニューラルネットワーク層への入力テンソルを、１のストライドを有する第２の畳み込みニューラルネットワーク層を用いて処理する（ステップ９０２）。たとえば、受け取られた命令から生成される制御信号は、畳み込まれたテンソルを生成するべく、専用ハードウェア回路３００を制御して、１に等しいがそれ以外は畳み込みニューラルネットワーク層に等しいストライドを有する第２の畳み込みニューラルネットワーク層を用いて、入力テンソル、たとえばユニファイドバッファ３０８に格納されるニューラルネットワークの先行する層の出力または指定されたもしくは専用ハードウェア回路３００に与えられるニューラルネットワークへの入力を処理する。

【００７６】

第２の畳み込みニューラルネットワーク層を用いて入力テンソルを処理するために、制御信号は、入力テンソル、すなわちニューラルネットワークへの入力または先行するニューラルネットワークの出力に対応し得る活性化入力を図３の行列計算ユニット３１２に供給するように、ユニファイドバッファ３０８を制御し得る。制御信号はまた、図３のダイレクトメモリアクセスエンジン３０４および／またはダイナミックメモリ３１０に命令して、１のストライド、すなわち単位ストライドを有するが、それ以外は１より大きいストライドを有するニューラルネットワーク層と同等である第２のニューラルネットワーク層に対応する重みを行列計算ユニット３１２に与えてもよい。

【００７７】

シーケンサ３０６はさらに、重みを用いて、たとえば図３に関して説明したプロセスを用いて、入力テンソルを処理するように行列計算ユニット３１２を制御する命令を生成することができる。いくつかの実装形態では、行列計算ユニット３１２は、２０１５年９月３日に提出された米国特許出願第１４／８４４，７３８号に記載されている技法を用いて畳み込みを実行し、その全体をここに引用により援用する。

【００７８】

行列計算ユニット３１２は、制御信号に基づいて計算を行い、畳み込まれたテンソルをベクトル計算ユニット３１４に出力する。たとえば、行列計算ユニット３１２は、行列計算ユニット３１２が生成した出力のベクトルをベクトル計算ユニット３１４に送る。出力のベクトルは、１のストライドを有するがそれ以外の点では１より大きいストライドを有するニューラルネットワーク層と同等であるニューラルネットワーク層に対応する重みを用いて入力テンソルを処理することに基づいて判定され得る。ベクトル計算ユニット３１４は、畳み込まれたテンソルをユニファイドバッファ３０８に格納することができる。

【００７９】

畳み込まれたテンソルを生成するために畳み込みニューラルネットワーク層を介して１

のストライドで活性化入力を処理した後、専用ハードウェア回路３００は、第２の畳み込みニューラルネットワーク層が１より大きいストライドを有する畳み込みネットワーク層のストライドを有すると仮定した場合に生成されなかったであろう要素を零出力する（ステップ９０４）。要素を零出力するとは、通常、その要素の現在の値を０に置き換えることを指す。値を取り消す、すなわち０にすることは、畳み込まれたテンソルとマスキングテンソル、すなわちニューラルネットワーク処理エンジン１５０によって生成され専用ニューラルネットワークに送られるマスキングテンソルとの要素ごとの乗算を実行することによって達成することができる。

【００８０】

入力テンソルが指定されたストライドで畳み込みニューラルネットワーク層によって処理された場合に生成されなかったであろう畳み込まれたテンソルの値を取り消すために、シーケンサ３０６は、制御信号を送って、行列乗算ユニット３１２に、畳み込まれたテンソルとマスキングテンソルとの要素ごとの乗算を実行させることができる。畳み込まれたテンソルは、シーケンサ３０６からの他の制御信号に基づいてユニファイドバッファ３０８から行列乗算ユニット３１２に送られてもよく、マスキングテンソルは、シーケンサ３０６からダイレクトメモリアクセスエンジン３０４またはダイナミックメモリ３１０への制御信号に基づいて、すなわち、マスキングテンソルが専用ハードウェア回路３００によって受け取られ、ダイナミックメモリ３１０に格納された後、行列計算ユニット３１２に送られてもよい。

【００８１】

一般に、図８に関して説明したように、マスキングテンソルは、１より大きいストライドを有する畳み込みニューラルネットワーク層を用いて入力テンソルを処理することによって生成されるであろう要素に対応する要素位置に単位値要素、すなわち１の値を含み、他のすべての位置、すなわち１より大きいストライドを有する畳み込みニューラルネットワーク層を用いて活性化値を処理することによって生成されないであろう要素に対応する位置に０値要素を含むベクトルである。

【００８２】

マスキングテンソルは、たとえば、ダイナミックメモリ３１０に格納されてもよく、シーケンサ３０６は、マスキングテンソルをダイナミックメモリ３１０から行列計算ユニット３１２に送るよう制御信号を送信してもよい。たとえば、専用ハードウェア回路３００に与えられる命令は、マスキングテンソルを識別してもよく、たとえばマスキングテンソルのダイナミックメモリ３１０内の位置を与えてもよく、またはその場合にダイナミックメモリ３１０に格納されるマスキングテンソルを定義するデータを含んでもよく、シーケンサ３０６は、ダイナミックメモリ３１０内のその位置に格納されるマスキングテンソルを行列計算ユニット３１２に送らせる制御信号を送信してもよい。さらに、シーケンサ３０６は、ユニファイドバッファ３０８に格納される畳み込まれたテンソルを行列計算ユニット３１２に与えさせるための制御信号を与えることができる。そして、行列計算ユニット３１２は、畳み込まれたテンソルとマスキングテンソルとの要素ごとの乗算を行い、修正された畳み込まれたテンソルを生成する。修正された畳み込まれたテンソルは、行列計算ユニット３１２からベクトル計算ユニット３１４によって受け取られることができる。ベクトル計算ユニット３１４は、任意選択で、修正された畳み込まれたテンソルをユニファイドバッファ３０８に格納することができる。

【００８３】

マスキングテンソルでの要素ごとの乗算のため、修正された畳み込まれたテンソルは、入力テンソルが１より大きい指定されたストライドを有するニューラルネットワーク層を用いて処理された場合に出力されるであろう値を含む。修正された畳み込まれたテンソルは、入力テンソルが指定されたストライドを有する畳み込みニューラルネットワークで処理された場合に出力されなかったであろう、１のストライドでの畳み込みニューラルネットワーク層を用いた入力テンソルの計算において出力される値に対応する位置に０を含む。他の実装形態では、畳み込まれたテンソルの要素を０にする他の方法を利用することが

できる。たとえば、畳み込まれた行列は、修正された形でユニファイドバッファ３０８または他のメモリにおいて書き直され得、指定されたストライドを有する畳み込みニューラルネットワークを用いての入力テンソルの計算において出力される値に対応する要素は変更されず、他の要素は０として書き込まれる。

【００８４】

ベクトル計算ユニット３１４は、修正された畳み込まれたテンソルを受け取り、修正された畳み込まれたテンソルに対して最大プーリングを実行して、１より大きいストライドを有する畳み込みニューラルネットワーク層の層出力テンソルを生成する（ステップ９０６）。たとえば、ベクトル計算ユニット３１４は、行列計算ユニット３１２から修正された畳み込まれたテンソルを受け取り、プーリング回路６０８を用いて、修正された畳み込まれたテンソルに対して最大プーリングを実行することができる。最大プーリングは、データのセットを受け取り、そのデータの１つまたは複数のサブセットの各々について、サブセット内の要素の最大値を出力する操作である。修正された畳み込まれたテンソルに対して最大プーリングを実行すると、修正された畳み込まれたテンソルの要素の複数のサブセットの各々について、サブセットの最大値を含むテンソルが得られる結果となる。ベクトル計算ユニット３１４は、畳み込みニューラルネットワーク層の指定されたストライドに基づいて決定される修正された畳み込まれたテンソルのウィンドウについて最大プーリングを実行することができる。たとえば、ストライドが２の場合、プーリング回路６０８は、２×２ウィンドウを用いて最大プーリングを実行し、各２×２ウィンドウからの最大値要素を含む層出力テンソルを生成する。４のストライドを有するニューラルネットワーク層の場合、プーリング回路６０８は、４×４ウィンドウを用いて最大プーリングを実行し、各４×４ウィンドウからの最大値要素を含む層出力テンソルを生成する。最大プーリング操作の結果は、ベクトル計算ユニット３１４によってユニファイドバッファ３０８に格納され、その結果とは、専用ハードウェア回路３００が１より大きいストライドを有するニューラルネットワーク層を用いて入力テンソルを処理した場合に生成されるであろう出力と等価の出力テンソルである。ニューラルネットワークの後続の層の処理は、最終的にニューラルネットワークの推論を取得するよう、層出力テンソルを用いて実行されてもよい。

【００８５】

図１０は、１より大きいストライドでのニューラルネットワークの所与の層に対する計算の例を示す。図１０の例は、図７のプロセスおよび図２の専用ハードウェア回路３００を用いて実行することができる。例として、図１０の例は、４のストライド有する畳み込みニューラルネットワーク層を活性化値の８×８アレイに適用する。畳み込みニューラルネットワーク層は、活性化値の８×８アレイに適用される重みの４×４カーネルを有してもよい。活性化値は、ニューラルネットワークに入力される画像の８×８部分、すなわち画像の８×８部分に対応する値のシーケンスを表すことができる。代替的に、活性化値の８×８アレイは、別の入力テンソル、たとえばニューラルネットワークの先行する層の出力に対応する入力テンソルの８×８部分を表すことができる。

【００８６】

図１０の部分（ａ）において、８×８入力テンソルは、１のストライドを有するが他の点では１より大きいストライドを有する畳み込みニューラルネットワーク層と同等である畳み込みニューラルネットワーク層を用いて処理される。したがって、部分（ａ）に示される重みの４×４のカーネルは、最初に、入力テンソルの最初の４行および最初の４列に対応する入力テンソルの要素に適用され得る（値は示されてはいない）。処理の結果は、結果として得られる畳み込まれたテンソルの第１の要素、すなわち、図１０の部分（ａ）に示される、結果として得られる畳み込まれたテンソルの要素「ａ」であり得る。

【００８７】

入力テンソルの処理は、指定された４のストライドではなく、１のストライドで畳み込みニューラルネットワーク層を用いて実行されるので、部分（ａ）に示された重みの４×４セットは、活性化値アレイの最初の４行および入力テンソルの第２列から第５列に対応

する入力テンソルの要素に適用されてもよい（値は示されず）。処理結果は、畳み込まれたテンソルの第２の要素、すなわち、図１０の部分（ａ）に示す畳み込み結果の要素「ｂ」である。重みの４×４セットを活性化値アレイに１のストライドを用いて適用することによって、すなわち重みの４×４セットを活性化値アレイに増分的に列方向および行方向の両方に適用することによって、プロセスを繰り返してもよい。この処理の結果、図１０の部分（ａ）に示す８×８の畳み込まれたテンソルが得られる。

【００８８】

次に、図９の部分（ｂ）に示すように、畳み込まれたテンソルとマスキングテンソルとの間で要素ごとの乗算が行われ、修正された畳み込まれたテンソルが得られる。マスキングテンソルのサイズは、入力テンソルのサイズまたは畳み込まれたテンソルのサイズに基づいて判断され、それは、１のストライドを有する畳み込みニューラルネットワーク層を用いる図１０の部分（ａ）での処理のため、一般的に等しい。マスキングテンソルは、入力テンソルが指定されたストライドを有する畳み込みニューラルネットワーク層を用いて処理された場合に生成されるであろう値に対応する位置に、単位値、すなわち１を含む。その場合、一般に、マスキングテンソルにおける単位値エントリの位置は、畳み込みニューラルネットワーク層の指定されたストライドに依存する。図１０の例では、畳み込みニューラルネットワーク層は４のストライドを有するので、マスキングテンソルは列方向および行方向の両方において４つおきの位置に単位値を含むであろう。マスキングテンソルの他のエントリには０値が割り当てられ、畳み込まれたテンソルとマスキングテンソルとの要素ごとの乗算は、畳み込みニューラルネットワークが指定されたストライドを有する状態で入力テンソルが処理された場合に生成されないであろうすべての値を０にする結果となることになる。

【００８９】

修正された畳み込まれたテンソルを生成するために、畳み込まれたテンソルとマスキングテンソルとの要素ごとの乗算が実行される。図１０に示すように、要素ごとの乗算の後、畳み込まれたテンソルの４つおきの要素は維持され、畳み込まれたテンソルの要素の残りはマスキング行列の対応する０値要素との乗算により０になる。したがって、８×８の畳み込まれたテンソルの要素のうち、４つの要素だけが非０のままである。

【００９０】

いくつかの実装形態では、最初に、畳み込まれたテンソルの要素に非単位係数を乗算し、続いて、それらの要素に第２の非単位係数を乗算することによって、同様の結果を得ることができる。たとえば、マスキングテンソルは、入力テンソルが指定されたストライドを有する畳み込みニューラルネットワーク層を用いて処理された場合に生成されるであろう値に対応する位置に、２（または他の値）を含んでもよい。したがって、上記の例に従って、畳み込まれたテンソルとマスキングテンソルとの要素ごとの乗算は、畳み込まれたテンソルの４つおきの要素が２倍になり、要素の残りが０である、修正された畳み込まれたテンソルを生成する。その後、修正された畳み込まれたテンソルの半分（または他の値の逆数）によるスカラー乗算が実行されてもよい。代替的に、修正された畳み込まれたテンソルと第２のマスキングテンソルとの要素ごとの乗算が実行されてもよく、第２のマスキングテンソルは、入力テンソルが指定されたストライドを有する畳み込みニューラルネットワーク層を用いて処理された場合に生成されるであろう値に対応する位置に２分の１の値を含む。

【００９１】

続いて、図１０の部分（ｃ）において修正された畳み込み結果アレイに対して最大プーリングが行われる。最大プーリングの結果は、入力テンソルがストライドが４の畳み込みニューラルネットワーク層によって処理された場合に得られるであろう結果と同等である。図６のプロセスを用いて、修正された畳み込まれたテンソルに対して最大プーリングが実行され、修正された畳み込まれたテンソルの各４×４ウィンドウの最大値が識別される。次いで、最大プーリングの結果は、４のストライドを有する畳み込みニューラルネットワーク層の出力テンソルとして格納される。入力テンソルは８×８アレイであったため、

４のストライドを有するニューラルネットワーク層による処理は２×２出力アレイをもたらす。２×２出力アレイは、図２のユニファイドバッファ３０８に、たとえばラスタ順で格納することができる。２×２出力アレイの値は、ニューラルネットワークの次の層への入力として与えられてもよい。

【００９２】

本明細書において記載される主題および機能的動作の実施形態は、本明細書に開示される構造およびそれらの構造的等価物を含む、デジタル電子回路系において、有形で実施されるコンピュータソフトウェアもしくはファームウェアにおいて、コンピュータハードウェアにおいて、またはそれらの１つ以上の組合せにおいて実現され得る。本明細書に記載される主題の実施形態は、１つ以上のコンピュータプログラムとして、すなわち、データ処理装置による実行のために、または、データ処理装置の動作を制御するために有形の非一時的なプログラム担体上でエンコードされたコンピュータプログラム命令の１つ以上のモジュールとして実現され得る。代替的に、または加えて、プログラム命令は、データ処理装置による実行に対して好適な受信側装置への送信のために情報をエンコードするよう生成される、たとえばマシンにより生成された電気信号、光信号、または電磁気信号などの、人為的に生成された伝搬される信号上でエンコードすることができる。コンピュータ記憶媒体は、コンピュータ可読記憶装置、コンピュータ可読記憶基板、ランダムもしくはシリアルアクセスメモリデバイス、または、それらの１つ以上の組合せであり得る。

【００９３】

「データ処理装置」という用語は、例としてプログラマブルプロセッサ、コンピュータ、または複数のプロセッサもしくはコンピュータを含む、データを処理するためのすべての種類の装置、デバイスおよびマシンを包含する。当該装置は、たとえばＦＰＧＡ（フィールドプログラマブルゲートアレイ）またはＡＳＩＣ（特定用途向け集積回路）といった特定目的論理回路を含み得る。当該装置は、ハードウェアに加えて、たとえばプロセッサファームウェア、プロトコルスタック、データベース管理システム、オペレーティングシステム、または、それらの１つ以上の組合せを構成するコードといった、当該コンピュータプログラムについて実行環境を作成するコードをさらに含み得る。

【００９４】

（プログラム、ソフトウェア、ソフトウェアアプリケーション、モジュール、ソフトウェアモジュール、スクリプトまたはコードとも称され、または記載され得る）コンピュータプログラムは、コンパイル型もしくはインタープリタ型言語、または宣言型もしくは手続き型言語を含む任意の形態のプログラミング言語で記述され得、スタンドアロンプログラムとして、または、モジュール、コンポーネント、サブルーチン、もしくは、コンピューティング環境で使用するのに好適な他のユニットとして任意の形態で展開され得る。コンピュータプログラムは、ファイルシステムにおけるファイルに対応し得るが、対応する必要があるわけではない。プログラムは、当該プログラムに専用である単一のファイルにおいて、または、複数の連携ファイル（ｃｏｏｒｄｉｎａｔｅｄ　ｆｉｌｅｓ）（たとえばコードの１つ以上のモジュール、サブプログラムまたは部分を格納するファイル）において、他のプログラムまたはデータ（たとえばマークアップ言語ドキュメントに格納される１つ以上のスクリプト）を保持するファイルの一部に格納され得る。コンピュータプログラムは、１つの場所に位置するかもしくは複数の場所にわたって分散され通信ネットワークによって相互接続される１つのコンピュータまたは複数のコンピュータ上で実行されるように展開され得る。

【００９５】

本明細書に記載されるプロセスおよび論理フローは、入力データ上で動作し出力を生成することにより機能を実行するよう１つ以上のプログラマブルコンピュータが１つ以上のコンピュータプログラムを実行することによって実行され得る。本プロセスおよび論理フローの実行、ならびに本装置の実施は、さらに、たとえばＦＰＧＡ（フィールドプログラマブルゲートアレイ）またはＡＳＩＣ（特定用途向け集積回路）といった特殊目的論理回路系によってもなされ得る。

【００９６】
コンピュータプログラムの実行に好適であるコンピュータは、例として、汎用マイクロプロセッサもしくは特殊目的マイクロプロセッサもしくはその両方または任意の種類の中央処理ユニットに基づき得る。一般に、中央処理ユニットは、リードオンリメモリもしくはランダムアクセスメモリまたはその両方から命令およびデータを受け取る。コンピュータの必須の要素は、命令を実行するための中央処理ユニットと、命令およびデータを格納するための１つ以上のメモリデバイスとである。一般に、コンピュータはさらに、たとえば磁気ディスク、光磁気ディスクまたは光ディスクといった、データを格納するための１つ以上の大容量記憶装置を含むか、当該１つ以上の大容量記憶装置からデータを受け取るかもしくは当該１つ以上の大容量記憶装置にデータを転送するよう動作可能に結合されるか、またはその両方を行う。しかしながら、コンピュータはそのような装置を有する必要はない。さらに、コンピュータはたとえば、携帯電話、携帯情報端末（ＰＤＡ）、モバイルオーディオまたはビデオプレーヤ、ゲームコンソール、全地球測位システム（ＧＰＳ）受信機、またはポータブル記憶装置（たとえばユニバーサルシリアルバス（ＵＳＢ）フラッシュドライブ）といった別のデバイスに埋め込まれ得る。

【００９７】
コンピュータプログラム命令およびデータを格納するのに好適であるコンピュータ可読媒体は、例として、たとえばＥＰＲＯＭ、ＥＥＰＲＯＭおよびフラッシュメモリデバイスといった半導体メモリデバイスを含むすべての形態の不揮発性メモリ、媒体およびメモリデバイス；たとえば内部ハードディスクまたはリムーバブルディスクといった磁気ディスク；光磁気ディスク；ならびにＣＤ－ＲＯＭおよびＤＶＤ－ＲＯＭディスクを含む。プロセッサおよびメモリは、特殊目的論理回路によって補足され得るか、または特殊目的論理回路に組み込まれ得る。

【００９８】
ユーザとの対話を求めて、本明細書に記載される主題の実施形態は、たとえばＣＲＴ（陰極線管）またはＬＣＤ（液晶ディスプレイ）モニタといったユーザに対して情報を表示するための表示デバイスと、たとえばマウス、トラックボールといったユーザがコンピュータに入力を送ることができるキーボードおよびポインティングデバイスとを有するコンピュータ上で実現され得る。他の種類のデバイスが、同様に、ユーザとの対話を求めて用いられ得；たとえば、ユーザに提供されるフィードバックは、たとえば視覚フィードバック、聴覚フィードバックまたは触覚フィードバックといった任意の形態の感覚フィードバックであり得；ユーザからの入力は、音響入力、音声入力、または触覚入力を含む任意の形態で受け取られ得る。加えて、コンピュータは、ユーザが使用するデバイスにドキュメントを送信しユーザが使用するデバイスからドキュメントを受信することによって、たとえば、ウェブブラウザから受信された要求に応答してユーザのクライアントデバイス上のウェブブラウザにウェブページを送信することによって、ユーザと対話し得る。

【００９９】
本明細書に記載される主題の実施形態は、たとえばデータサーバとしてバックエンドコンポーネントを含む計算システムにおいて実現され得るか、たとえばアプリケーションサーバといったミドルウェアコンポーネントを含む計算システムにおいて実現され得るか、たとえば本明細書に記載される主題の実現例とユーザが対話することが可能であるグラフィカルユーザーインターフェイスもしくはウェブブラウザを有するクライアントコンピュータといったフロントエンドコンポーネントを含む計算システムにおいて実現され得るか、または１つ以上のそのようなバックエンドコンポーネント、ミドルウェアコンポーネントもしくはフロントエンドコンポーネントの任意の組合せの計算システムにおいて実現され得る。システムのコンポーネントは、たとえば通信ネットワークといったデジタルデータ通信の任意の形態または媒体によって相互接続され得る。通信ネットワークの例は、ローカルエリアネットワーク（「ＬＡＮ」）およびワイドエリアネットワーク（「ＷＡＮ」）、たとえばインターネットを含む。

【０１００】

計算システムはクライアントおよびサーバを含むことができる。クライアントとサーバとは一般に互いから遠隔にあり、典型的には通信ネットワークを通じて対話する。クライアントとサーバとの関係は、それぞれのコンピュータ上で実行されるとともに互いに対してクライアント－サーバ関係を有するコンピュータプログラムによって発生する。

【０１０１】

本明細書は多くの特定の実現例の詳細を含んでいるが、これらは如何なる発明の範囲または請求され得るものの範囲に対する限定としても解釈されるべきではなく、特定の発明の特定の実施形態に特有の特徴であり得る記載として解釈されるべきである。別個の実施形態の文脈で本明細書において記載されるある特徴は、単一の実施形態において組合せでも実現され得る。反対に、単一の実施形態の文脈において記載されるさまざまな特徴は、複数の実施形態において別々に、または任意の好適な部分的組合わせでも実現され得る。さらに、特徴は、ある組合せにおいて作用すると上で記載され、最初はそのように請求されていさえする場合もあるが、請求される組合せからの１つ以上の特徴はいくつかの場合には当該組合せから削除され得、請求される組合せは、部分的組合わせまたは部分的組合わせの変形例に向けられ得る。

【０１０２】

同様に、動作が図においては特定の順に示されているが、そのような動作は、望ましい結果を達成するために、示された当該特定の順もしくは連続した順で実行される必要があると理解されるべきではなく、または、すべての示された動作が実行される必要があると理解されるべきではない。ある状況においては、マルチタスキングおよび並列処理が有利であり得る。さらに、上述の実施形態におけるさまざまなシステムモジュールおよびコンポーネントの分離は、すべての実施形態においてそのような分離を必要とすると理解されるべきではなく、記載されるプログラムコンポーネントおよびシステムは一般に単一のソフトウェア製品に統合され得るかまたは複数のソフトウェア製品にパッケージ化され得ることが理解されるべきである。

【０１０３】

その他の実現例は、以下の例にまとめられる。

例１：ハードウェア回路上でニューラルネットワークを処理するよう要求を受け取ることを備え、ニューラルネットワークは、１より大きいストライドを有する第１の畳み込みニューラルネットワーク層を含み、さらに、これに応答して、ハードウェア回路によって実行されると、ハードウェア回路に、ニューラルネットワークによる入力テンソルの処理中に、動作を実行することによって第１の畳み込みニューラルネットワーク層の出力と等価な層出力テンソルを生成させる命令を生成することを備え、動作は、第１の畳み込みニューラルネットワーク層への入力テンソルを、１に等しいストライドを有するがそれ以外は第１の畳み込みニューラルネットワーク層と同等である第２の畳み込みニューラルネットワーク層を用いて処理することにより、第１のテンソルを生成することと、第２の畳み込みニューラルネットワーク層が第１の畳み込みニューラルネットワーク層のストライドを有する場合には生成されなかったであろう第１のテンソルの要素を零出力して、第２のテンソルを生成することと、第２のテンソルに対して最大プーリングを実行して層出力テンソルを生成することとを含む、方法。

【０１０４】

例２：第１のテンソルの要素を零出力することは、第１のテンソルの要素のサブセットに０を乗算することと、サブセットに含まれていない第１のテンソルの要素に１を乗算することとを含む、例１の方法。

【０１０５】

例３：第１のテンソルの要素を零出力することは、マスキングテンソルと第１のテンソルとの要素ごとの乗算を実行して第２のテンソルを生成することを含み、マスキングテンソルは、（ｉ）第２の畳み込みニューラルネットワーク層が第１の畳み込みニューラルネットワーク層のストライドを有する場合には生成されなかったであろう第１のテンソルの要素に対応するマスキングテンソルの各要素位置において０を含み、（ｉｉ）マスキング

テンソルの各他の要素位置において１を含む、例１の方法。
【０１０６】
例４：マスキングテンソルは、ハードウェア回路によってアクセス可能なメモリに格納され、マスキングテンソルと第１のテンソルとの要素ごとの乗算は、ハードウェア回路に含まれる、ハードウェアにおいて実装されるベクトル計算ユニットによって実行される、例３の方法。
【０１０７】
例５：第１のテンソルの要素を零出力することは、第１のマスキングテンソルと第１のテンソルとの要素ごとの乗算を実行して、修正された第１のテンソルを生成することを含み、第１のマスキングテンソルは、（ｉ）第２の畳み込みニューラルネットワーク層が第１の畳み込みニューラルネットワーク層のストライドを有する場合には生成されなかったであろう第１のテンソルの要素に対応するマスキングテンソルの各要素位置において０を含み、（ｉｉ）第２の畳み込みニューラルネットワーク層が第１の畳み込みニューラルネットワーク層のストライドを有する場合には生成されたであろう第１のテンソルの要素に対応するマスキングテンソルの各要素位置においてそれぞれの非０値を含み、第１のテンソルの要素を零出力することはさらに、第２のマスキングテンソルと修正された第１のテンソルとの要素ごとの乗算を実行することを含み、第２のマスキングテンソルは、第２の畳み込みニューラルネットワーク層が第１の畳み込みニューラルネットワーク層のストライドを有する場合に生成されるであろう第１のテンソルの要素に対応する各要素位置において、第１のマスキングテンソルのそれぞれの非０値の逆数を含む、例１の方法。
【０１０８】
例６：最大プーリングを実行することは、第１の畳み込みニューラルネットワーク層のストライドによって定義される第２のテンソルの１つまたは複数のウィンドウの各々について、ウィンドウ内の要素の最大値要素を取得することを含む、例１～例５の１つの方法。
【０１０９】
例７：第２のテンソルの１つまたは複数のウィンドウの各々は、畳み込みニューラルネットワーク層のストライドに対応する次元を有する矩形ウィンドウであり、第２のテンソルの異なる要素を含む、例６の方法。
【０１１０】
例８：最大プーリングを実行することは、第２のテンソルの要素の１つまたは複数のサブセットの各々について、サブセットの最大値要素を取得することを含む、例１～例７の１つの方法。
【０１１１】
例９：第２のテンソル上で実行される最大プーリングは、ハードウェア回路のプーリング回路によって実行される、例１～例８の１つの方法。
【０１１２】
例１０：畳み込みニューラルネットワーク層は、ニューラルネットワーク内の第１のニューラルネットワーク層であり、入力テンソルは、デジタル画像の、デジタル画像の画素に対応する要素を含む表現である、例１～例９の１つの方法。
【０１１３】
例１１：入力テンソルはハードウェア回路のユニファイドバッファに格納され、第２の畳み込みニューラルネットワーク層の重みはハードウェア回路のダイナミックメモリに格納され、第２の畳み込みニューラルネットワーク層を用いて第１の畳み込みニューラルネットワーク層への入力テンソルを処理することは、入力テンソルをユニファイドバッファからハードウェアで実装されるハードウェア回路の行列計算ユニットに送ることと、ダイナミックメモリからハードウェア回路の行列計算ユニットに第２の畳み込みニューラルネットワーク層の重みを送ることと、ハードウェア回路の行列計算ユニットによって、第２の畳み込みニューラルネットワーク層の重みを用いて入力テンソルを処理して、第１のテンソルを生成することとを含む、例１～例１０の１つの方法。

【０１１４】
例１２：システムであって、ハードウェア回路と、命令を格納する１つまたは複数の記憶装置とを備え、命令は、ハードウェア回路によって実行されると、ハードウェア回路に動作を実行させるよう動作可能であり、動作は、１より大きいストライドを有する畳み込みニューラルネットワーク層への入力テンソルを、１に等しいストライドを有するがそれ以外は畳み込みニューラルネットワーク層と同等である第２の畳み込みニューラルネットワーク層を用いて処理することにより、第１のテンソルを生成することと、第２の畳み込みニューラルネットワーク層が畳み込みニューラルネットワーク層のストライドを有する場合には生成されなかったであろう第１のテンソルの要素を零出力して、第２のテンソルを生成することと、第２のテンソルに対して最大プーリングを実行して層出力テンソルを生成することとを含む、システム。
【０１１５】
例１３：第１のテンソルの要素を零出力することは、マスキングテンソルと第１のテンソルとの要素ごとの乗算を実行して第２のテンソルを生成することを含み、マスキングテンソルは、（ｉ）第２の畳み込みニューラルネットワーク層が第１の畳み込みニューラルネットワーク層のストライドを有する場合には生成されなかったであろう第１のテンソルの要素に対応するマスキングテンソルの各要素位置において０を含み、（ｉｉ）マスキングテンソルの各他の要素位置において１を含む、例１２のシステム。
【０１１６】
例１４：マスキングテンソルは、ハードウェア回路によってアクセス可能なメモリに格納され、マスキングテンソルと第１のテンソルとの要素ごとの乗算は、ハードウェア回路に含まれる、ハードウェアで実装されるベクトル計算ユニットによって実行される、例１３のシステム。
【０１１７】
例１５：最大プーリングを実行することは、第１の畳み込みニューラルネットワーク層のストライドによって定義される第２のテンソルの１つまたは複数のウィンドウの各々について、ウィンドウ内の要素の最大値要素を取得することを含む、例１２～例１４の１つのシステム。
【０１１８】
例１６：第２のテンソルの１つまたは複数のウィンドウの各々は、畳み込みニューラルネットワーク層のストライドに対応する次元を有する矩形ウィンドウであり、第２のテンソルの異なる要素を含む、例１５のシステム。
【０１１９】
例１７：第２のテンソル上で実行される最大プーリングは、ハードウェア回路のプーリング回路によって実行される、例１２～例１６の１つのシステム。
【０１２０】
例１８：畳み込みニューラルネットワーク層は、ニューラルネットワーク内の第１のニューラルネットワーク層であり、入力テンソルは、デジタル画像の、デジタル画像の画素に対応する要素を含む表現である、例１２～例１７の１つのシステム。
【０１２１】
例１９：入力テンソルはハードウェア回路のユニファイドバッファに格納され、第２の畳み込みニューラルネットワーク層の重みはハードウェア回路のダイナミックメモリに格納され、第２の畳み込みニューラルネットワーク層を用いて第１の畳み込みニューラルネットワーク層への入力テンソルを処理することは、入力テンソルをユニファイドバッファからハードウェアで実装されるハードウェア回路の行列計算ユニットに送ることと、ダイナミックメモリからハードウェア回路の行列計算ユニットに第２の畳み込みニューラルネットワーク層の重みを送ることと、ハードウェア回路の行列計算ユニットによって、第２の畳み込みニューラルネットワーク層の重みを用いて入力テンソルを処理して、第１のテンソルを生成することとを含む、例１２～例１８の１つのシステム。
【０１２２】

例２０：コンピュータプログラムでエンコードされたコンピュータ可読記憶装置であって、コンピュータプログラムは、１つまたは複数のコンピュータによって実行されると、１つまたは複数のコンピュータに動作を実行させる命令を含み、動作は、ハードウェア回路上でニューラルネットワークを処理するよう要求を受け取ることを含み、ニューラルネットワークは、１より大きいストライドを有する第１の畳み込みニューラルネットワーク層を含み、動作はさらに、これに応答して、ハードウェア回路によって実行されると、ハードウェア回路に、ニューラルネットワークによる入力テンソルの処理中に、動作を実行することによって第１の畳み込みニューラルネットワーク層の出力と等価な層出力テンソルを生成させる命令を生成することを含み、動作は、第１の畳み込みニューラルネットワーク層への入力テンソルを、１に等しいストライドを有するがそれ以外は第１の畳み込みニューラルネットワーク層と同等である第２の畳み込みニューラルネットワーク層を用いて処理することにより、第１のテンソルを生成することと、第２の畳み込みニューラルネットワーク層が第１の畳み込みニューラルネットワーク層のストライドを有する場合には生成されなかったであろう第１のテンソルの要素を零出力して、第２のテンソルを生成することと、第２のテンソルに対して最大プーリングを実行して層出力テンソルを生成することとを含む、コンピュータ可読記憶装置。

【０１２３】

主題の特定の実施形態が記載された。他の実施形態は以下の請求の範囲内にある。たとえば、請求項において記載されるアクションは、異なる順で実行され得、それでも望ましい結果を達成し得る。一例として、添付の図において示されるプロセスは、望ましい結果を達成するために、示された特定の順または連続する順であることを必ずしも必要としない。ある実現例においては、マルチタスキングおよび並列処理が有利であり得る。

【図１】

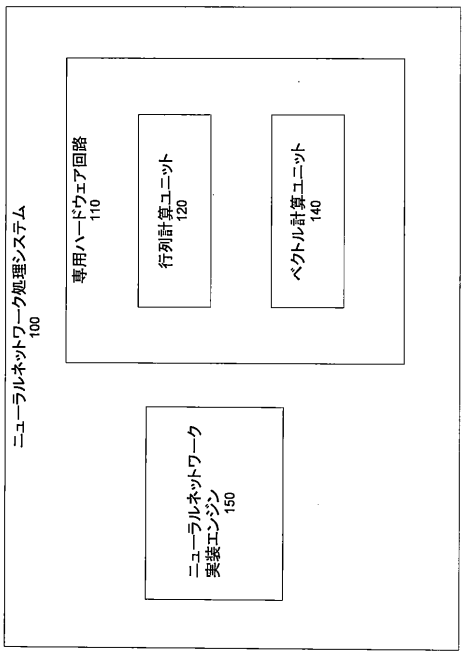


FIG. 1

【図２】

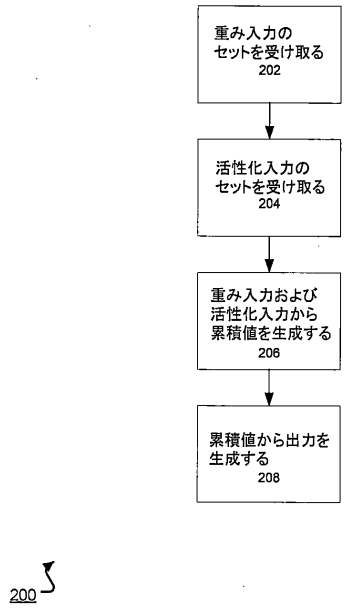


FIG. 2

【図３】

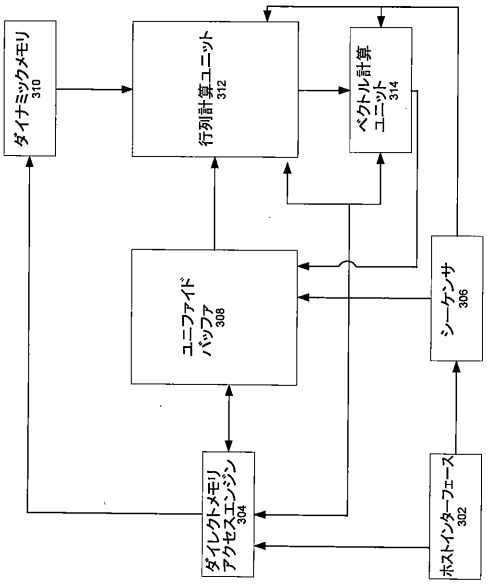


FIG. 3

【図４】

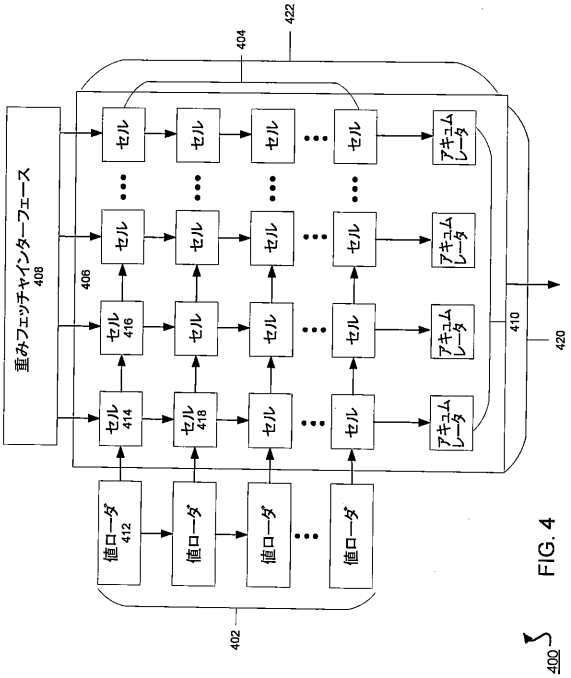


FIG. 4

【図５】

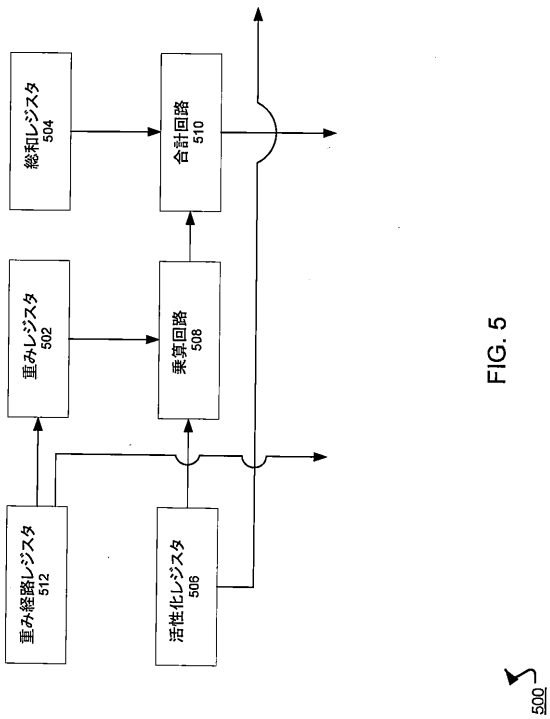


FIG. 5

【図６】

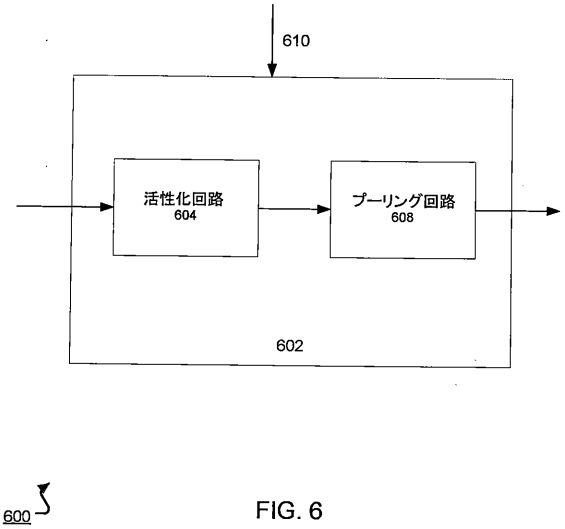


FIG. 6

【図７】

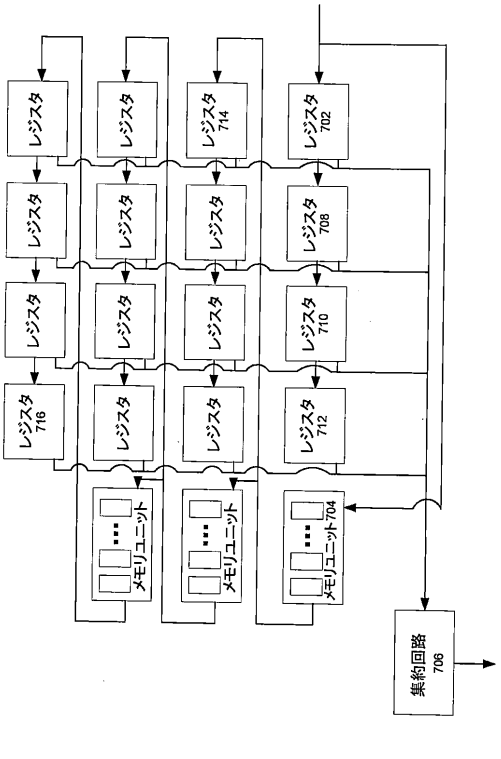


FIG. 7

【図８】

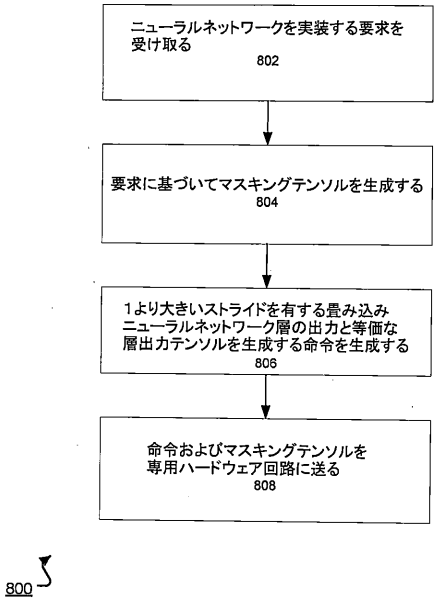


FIG. 8

【図９】

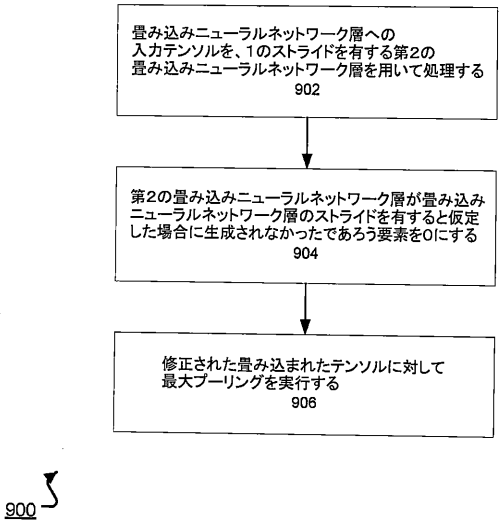
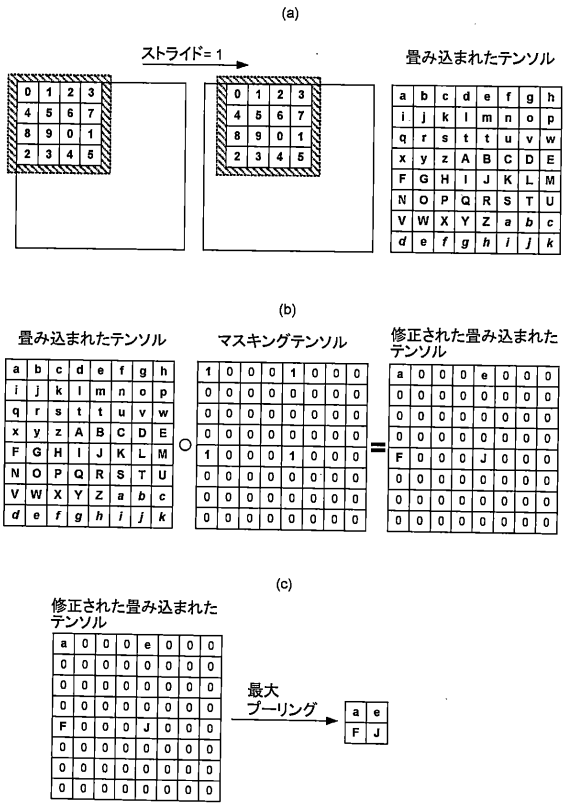


FIG. 9

【図１０】



フロントページの続き

(72)発明者　ガランド，ウィリアム・ジョン
アメリカ合衆国、９４０４３　カリフォルニア州、マウンテン・ビュー、アンフィシアター・パークウェイ、１６００

審査官　金田　孝之

(56)参考文献　特開２０１６－１５３９８４（ＪＰ，Ａ）
米国特許出願公開第２０１５／０２７８６４２（ＵＳ，Ａ１）
Chen Zhang et al., Optimizing FPGA-based Accelerator Design for Deep Convolutional Neural Networks, Proceedings of the 2015 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, 2015年02月22日, pp.161-170
Clement Farabet et al., CNP: An FPGA-based processor for Convolutional Networks, 2009 International Conference on Field Programmable Logic and Applications, 2009年08月31日, pp.32-37

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｎ　　３／００－９９／００