(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7004741号
(P7004741)

(45)発行日　令和4年1月21日(2022.1.21)

(24)登録日　令和4年1月6日(2022.1.6)

(51)国際特許分類		FI		
G06N 3/06 (2006.01)		G06N	3/06	
G06F 12/04 (2006.01)		G06F	12/04	530
H03M 7/40 (2006.01)		H03M	7/40	

請求項の数　15　（全31頁）

(21)出願番号　特願2019-555659(P2019-555659)
(86)(22)出願日　平成30年4月16日(2018.4.16)
(65)公表番号　特表2020-517014(P2020-517014A)
(43)公表日　令和2年6月11日(2020.6.11)
(86)国際出願番号　PCT/US2018/027840
(87)国際公開番号　WO2018/194998
(87)国際公開日　平成30年10月25日(2018.10.25)
審査請求日　令和3年3月10日(2021.3.10)
(31)優先権主張番号　62/486,432
(32)優先日　平成29年4月17日(2017.4.17)
(33)優先権主張国・地域又は機関
米国(US)
(31)優先権主張番号　15/953,356
(32)優先日　平成30年4月13日(2018.4.13)
最終頁に続く

(73)特許権者　314015767
マイクロソフト テクノロジー ライセンシング，エルエルシー
アメリカ合衆国 ワシントン州 98052 レッドモンド ワン マイクロソフト ウェイ
(74)代理人　100107766
弁理士　伊東 忠重
(74)代理人　100070150
弁理士　伊東 忠彦
(74)代理人　100091214
弁理士　大貫 進介
(72)発明者　コーカリー，ジョセフ レオン
アメリカ合衆国 ワシントン州 98052 レッドモンド ワン マイクロソフト
最終頁に続く

(54)【発明の名称】　メモリ帯域幅利用を低減するために活性化データの圧縮及び復元を使用するニューラルネットワークプロセッサ

(57)【特許請求の範囲】

【請求項１】

ニューラルネットワークプロセッサであって、

１つ以上のニューロンと、圧縮ユニットとを有し、

前記圧縮ユニットは、

当該ニューラルネットワークプロセッサ内の前記ニューロンの少なくとも１つによって生成されるデータの非圧縮チャンクを受け取り、該データの非圧縮チャンクが一定数のバイトを含み、

圧縮された出力チャンクのマスク部分を生成し、該マスク部分が、前記データの非圧縮チャンク内の前記一定数のバイトに等しいビットの数を含み、前記マスク部分内の各ビットが、前記データの非圧縮チャンク内のバイトに対応し、前記マスク部分内の各ビットが、前記データの非圧縮チャンク内の対応するバイトがゼロである場合に論理０に設定され、前記データの非圧縮チャンク内の対応するバイトが非ゼロである場合に論理１に設定され、前記データの非圧縮チャンク内の非ゼロバイトの数を決定することと、前記データの非圧縮チャンク内の前記非ゼロバイトの数に基づき、前記データの非圧縮チャンクの切り捨てられた非ゼロバイトを格納するために利用可能な前記圧縮された出力チャンクのデータ部分内のビットの数を決定することと、該決定されたビットの数まで前記データの非圧縮チャンク内の前記非ゼロバイトを切り捨てることと、該切り捨てられた非ゼロバイトを前記圧縮された出力チャンクの前記データ部分に格納することとによって、前記圧縮された出力チャンクの前記データ部分を生成し、

前記マスク部分及び前記データ部分を含む前記圧縮された出力チャンクを出力する
よう構成される、
ニューラルネットワークプロセッサ。
【請求項２】
当該ニューラルネットワークプロセッサは、復元ユニットを更に有し、
前記復元ユニットは、
前記圧縮された出力チャンクを受け取り、
前記圧縮された出力チャンクの前記マスク部分に基づき、前記データの非圧縮チャンク内の非ゼロバイトの数を決定し、
前記圧縮された出力チャンクの前記マスク部分に基づき、前記データの非圧縮チャンク内の前記非ゼロバイトの位置を決定し、
前記圧縮された出力チャンクの前記データ部分に前記切り捨てられた非ゼロバイトを格納するために前記圧縮ユニットによって使用されたビットの数を決定し、
論理０である前記圧縮された出力チャンクの前記マスク部分内の各ビット位置について、ゼロバイトを、復元された出力チャンクの対応する位置に挿入し、
論理１である前記マスク部分内の各ビット位置について、前記圧縮された出力チャンクの前記データ部分内の対応する位置からの前記切り捨てられた非ゼロバイトを、前記復元された出力チャンクの対応する位置に、前記圧縮された出力チャンクの圧縮中に切り捨てられたビットの数に等しい数のゼロビットとともに挿入する
よう構成される、
請求項１に記載のニューラルネットワークプロセッサ。
【請求項３】
前記圧縮ユニットは更に、
前記データの非圧縮チャンクの切り捨てられた非ゼロバイトを格納するために利用可能な前記圧縮された出力チャンクの前記データ部分内の追加ビットの数を決定し、
前記データの非圧縮チャンク内の前記非ゼロバイトのうちの１つ以上の非ゼロバイトを切り捨てる前に、該１つ以上の非ゼロバイトに前記追加ビットを割り当てる
よう構成される、
請求項２に記載のニューラルネットワークプロセッサ。
【請求項４】
前記復元ユニットは更に、前記圧縮された出力チャンクの前記データ部分内に格納された前記１つ以上の非ゼロバイトに割り当てられている前記追加ビットの数を決定するよう構成される、
請求項３に記載のニューラルネットワークプロセッサ。
【請求項５】
前記復元ユニットは更に、前記復元された出力チャンクに格納された前記切り捨てられた非ゼロバイトのうちの１つ以上の非ゼロバイトにオフセットを加えるよう構成される、
請求項２に記載のニューラルネットワークプロセッサ。
【請求項６】
前記非ゼロバイトの１つ以上の最下位ビット（ＬＳＢ）が切り捨てられる、
請求項１に記載のニューラルネットワークプロセッサ。
【請求項７】
ニューラルネットワークプロセッサであって、
１つ以上のニューロンと、復元ユニットとを有し、
前記復元ユニットは、
マスク部分及びデータ部分を含むデータの圧縮されたチャンクを受け取り、
前記マスク部分内のビットに基づき、データの復元されたチャンク内の非ゼロバイトの数を決定し、
前記非ゼロバイトの数に少なくとも部分的に基づき、前記データの圧縮されたチャンクの前記データ部分に切り捨てられた非ゼロバイトを格納するために使用されたビットの数を

決定し、
論理０である前記データの圧縮されたチャンクの前記マスク部分内の各ビット位置について、ゼロバイトを、前記データの復元されたチャンクの対応する部分に挿入し、
論理１である前記データの圧縮されたチャンクの前記マスク部分内の各ビット位置について、前記データの圧縮されたチャンクの前記データ部分内の対応する位置からの前記切り捨てられた非ゼロバイトを、前記データの復元されたチャンクの対応する位置に、前記データの圧縮されたチャンクの圧縮中に切り捨てられたビットの数に等しい数のゼロビットとともに挿入する
よう構成される、
ニューラルネットワークプロセッサ。
【請求項８】
圧縮ユニットを更に有し、
前記圧縮ユニットは、
当該ニューラルネットワークプロセッサ内の前記ニューロンの少なくとも１つによって生成されるデータの非圧縮チャンクを受け取り、該データの非圧縮チャンクが一定数のバイトを含み、
前記データの圧縮されたチャンクの前記マスク部分を生成し、該マスク部分が、前記データの非圧縮チャンク内の前記一定数のバイトに等しいビットの数を含み、前記マスク部分内の各ビットが、前記データの非圧縮チャンク内のバイトに対応し、前記マスク部分内の各ビットが、前記データの非圧縮チャンク内の対応するバイトがゼロである場合に論理０を有し、前記データの非圧縮チャンク内の対応するバイトが非ゼロである場合に論理１を有し、
前記データの非圧縮チャンク内の非ゼロバイトの数を決定することと、前記データの非圧縮チャンク内の前記非ゼロバイトの数に基づき、前記データの非圧縮チャンクの切り捨てられた非ゼロバイトを格納するために利用可能な前記データの圧縮されたチャンクの前記データ部分内のビットの数を決定することと、該決定されたビットの数まで前記データの非圧縮チャンク内の前記非ゼロバイトを切り捨てることと、該切り捨てられた非ゼロバイトを前記データの圧縮されたチャンクの前記データ部分に格納することとによって、前記データの圧縮されたチャンクの前記データ部分を生成し、
前記マスク部分及び前記データ部分を含む前記データの圧縮されたチャンクを出力する
よう構成される、
請求項７に記載のニューラルネットワークプロセッサ。
【請求項９】
前記圧縮ユニットは更に、前記データの非圧縮チャンク内の非ゼロバイトの数が前記データの圧縮されたチャンクの前記データ部分内のバイトの数以下である場合には切り捨てなしで、前記データの非圧縮チャンク内の前記非ゼロバイトを前記データの圧縮されたチャンクの前記データ部分に格納するよう構成される、
請求項８に記載のニューラルネットワークプロセッサ。
【請求項１０】
前記圧縮ユニットは更に、
前記データの非圧縮チャンクの切り捨てられた非ゼロバイトを格納するために利用可能な前記データの圧縮されたチャンクの前記データ部分内の追加ビットの数を決定し、
前記データの非圧縮チャンク内の前記非ゼロバイトのうちの１つ以上の非ゼロバイトを切り捨てる前に、該１つ以上の非ゼロバイトに前記追加ビットを割り当てる
よう構成される、
請求項８に記載のニューラルネットワークプロセッサ。
【請求項１１】
前記復元ユニットは更に、前記データの圧縮されたチャンクの前記データ部分内に格納された前記１つ以上の非ゼロバイトに割り当てられている前記追加ビットの数を決定するよう構成される、

請求項１０に記載のニューラルネットワークプロセッサ。
【請求項１２】
ニューラルネットワークプロセッサの圧縮ユニットで、前記ニューラルネットワークプロセッサ内の少なくとも１つのニューロンによって生成されるデータの非圧縮チャンクを受け取り、該データの非圧縮チャンクが一定数のバイトを含む、ことと、
圧縮された出力チャンクのマスク部分を生成し、該マスク部分が、前記データの非圧縮チャンク内の前記一定数のバイトに等しいビットの数を含み、前記マスク部分内の各ビットが、前記データの非圧縮チャンク内のバイトに対応し、前記マスク部分内の各ビットが、前記データの非圧縮チャンク内の対応するバイトがゼロである場合に論理０を有し、前記データの非圧縮チャンク内の対応するバイトが非ゼロである場合に論理１を有する、ことと、
前記データの非圧縮チャンク内の非ゼロバイトの数を決定すること、前記データの非圧縮チャンク内の前記非ゼロバイトの数に基づき、前記データの非圧縮チャンクの切り捨てられた非ゼロバイトを格納するために利用可能な前記圧縮された出力チャンクのデータ部分内のビットの数を決定することと、該決定されたビットの数まで前記データの非圧縮チャンク内の前記非ゼロバイトを切り捨てることと、該切り捨てられた非ゼロバイトを前記圧縮された出力チャンクの前記データ部分に格納することとによって、前記圧縮された出力チャンクの前記データ部分を生成することと、
前記マスク部分及び前記データ部分を含む前記圧縮された出力チャンクを前記ニューラルネットワークプロセッサのメモリに記憶することと
を有する、コンピュータにより実施される方法。
【請求項１３】
前記データの非圧縮チャンクの切り捨てられた非ゼロバイトを格納するために利用可能な前記圧縮された出力チャンクの前記データ部分内の追加ビットの数を決定することと、
前記データの非圧縮チャンク内の前記非ゼロバイトのうちの１つ以上の非ゼロバイトを切り捨てる前に、該１つ以上の非ゼロバイトに前記追加ビットを割り当てることと
を更に有する、
請求項１２に記載のコンピュータにより実施される方法。
【請求項１４】
前記データの非圧縮チャンク内の非ゼロバイトの数が前記圧縮された出力チャンクの前記データ部分内のバイトの数以下である場合には切り捨てなしで、前記データの非圧縮チャンク内の前記非ゼロバイトを前記圧縮された出力チャンクの前記データ部分に格納することを更に有する、
請求項１２に記載のコンピュータにより実施される方法。
【請求項１５】
前記ニューラルネットワークプロセッサの復元ユニットで、前記圧縮された出力チャンクを受け取ることと、
前記圧縮された出力チャンクの前記マスク部分に基づき、前記データの非圧縮チャンク内の非ゼロバイトの数を決定することと、
前記圧縮された出力チャンクの前記マスク部分に基づき、前記データの非圧縮チャンク内の前記非ゼロバイトの位置を決定することと、
前記圧縮された出力チャンクの前記データ部分に前記切り捨てられた非ゼロバイトを格納するために前記圧縮ユニットによって使用されたビットの数を決定することと、
論理０である前記圧縮された出力チャンクの前記マスク部分内の各ビット位置について、ゼロバイトを、復元された出力チャンクの対応する位置に挿入することと、
論理１である前記マスク部分内の各ビット位置について、前記圧縮された出力チャンクの前記データ部分内の対応する位置からの前記切り捨てられた非ゼロバイトを、前記復元された出力チャンクの対応する位置に、前記圧縮された出力チャンクの圧縮中に切り捨てられたビットの数に等しい数のゼロビットとともに挿入することと
を更に有する、

請求項１２に記載のコンピュータにより実施される方法。
【発明の詳細な説明】
【背景技術】
【０００１】
ディープニューラルネットワーク（Deep Neural Network(s)）（“ＤＮＮ”）は、人間の脳などの生体神経系における情報処理及び通信パターンを大まかにモデル化している。ＤＮＮは、制限なしに、物体検出、意味的標識付け（semantic labeling）、及び特徴抽出などの複雑な分類問題を解くために利用可能である。結果として、ＤＮＮは、コンピュータビジョン、音声認識、及び機械翻訳などの多くの人工知能（Artificial Intelligence）（“ＡＩ”）アプリケーションの基礎を形成している。ＤＮＮは、それらの分野の多くで人間の精度以上であることができる。
【０００２】
ＤＮＮの高い性能は、入力空間の効果的表現を得るために大規模データセットに対する統計学習を使用した後に入力データから高次特徴（high-level features）を抽出するそれらの能力に由来する。しかし、ＤＮＮの優れた性能は、高い計算複雑性という代償の上に成り立っている。グラフィクスプロセッシングユニット（Graphics Processing Unit(s)）（“ＧＰＵ”）などの高性能の汎用プロセッサが、多くのＤＮＮアプリケーションによって必要とされる高度な計算性能を提供するために広く利用されている。
【０００３】
ＧＰＵのような汎用プロセッサは、ＤＮＮを実装するための高度な計算性能を提供することができるが、他方で、そのようなプロセッサは、低電力消費が重要であるコンピュータ装置でＤＮＮ動作を長期にわたって実行することにおける使用には、通常は適さない。例えば、ＧＰＵなどの汎用プロセッサは、スマートフォン又は代替／仮想現実（Alternate/Virtual Reality）（“ＡＲ／ＶＲ”）デバイスのような、バッテリ寿命を延ばすために電力消費の削減が求められるバッテリ駆動のポータブルデバイスで長期にわたるＤＮＮタスクを実行することにおける使用には不適切であり得る。
【０００４】
人間の動作の検出などの連続的なＤＮＮタスクを実行しながら電力消費を低減することは、例えば、パワー・オーバー・イーサネット（Power-over-Ethernet）（“ＰＯＥ”）などの非バッテリ駆動デバイスでも重要であり得る。この具体的な例において、ＰＯＥスイッチは、限られた量の電力しか供給することができず、防犯カメラなどのＰＯＥデバイスの電力消費量を減らすことは、電力供給量が少ないＰＯＥスイッチの使用を可能にする。
【０００５】
高性能のＤＮＮ処理を提供しながら同時に、汎用プロセッサと比較して電力消費量を削減することができる特定用途向け集積回路（Application-Specific Integrated Circuit(s)）（“ＡＳＩＣ”）が開発されてきた。しかし、この分野の進歩にかかわらず、特に、低電力消費が重要であるコンピュータ装置での使用のために、ＤＮＮ処理を実行するＡＳＩＣの性能の改善及び電力消費量の削減に対するニーズが依然として存在する。
【０００６】
本明細書でなされた開示が提示されるのは、これら及び他の技術的課題に関してである。
【発明の概要】
【０００７】
メモリバス帯域幅の利用量を低減するために活性化データを圧縮及び復元することができるＤＮＮモジュール又はプロセッサが、開示される。特に、ＤＮＮモジュールは、ニューロン出力とオンボード又はオフボードのメモリとの間のバス帯域幅の利用量を減らすために圧縮を利用することができる。ＤＮＮモジュールはまた、オンボード又はオフボードのメモリとニューロン入力との間のメモリバス帯域幅の利用量を減らすために復元を利用することもできる。帯域幅利用量の低減は、より高速な処理を可能にすることができ、その結果、電力量消費量も減らすことができる。本明細書で具体的に述べられない他の技術的利点も、開示されている対象の実施を通じて実現され得る。

【０００８】
先に簡単に述べられた技術的利点を実現するために、１つ以上のニューロン及び圧縮ユニットを含むＤＮＮプロセッサが、開示される。圧縮ユニットは、１つ以上のヒューロンによって生成されるデータの非圧縮チャンクを受け取ることができる。いくつかの実施形態において、データの非圧縮チャンクは、６４バイトなどの、一定数のバイトを含む。
【０００９】
データの非圧縮チャンクを圧縮するために、圧縮ユニットは、圧縮された出力チャンクのマスク部分及びデータ部分を生成することができる。圧縮された出力チャンクのマスク部分は、データの非圧縮チャンク内の一定数のバイトに等しいビットの数を含む。例えば、データの非圧縮チャンクが６４バイトのデータを含む場合には、マスク部分は６４ビット（すなわち、８バイト）を含むことになる。
【００１０】
いくつかの実施形態において、圧縮された出力チャンクのマスク部分内の各ビットは、データの非圧縮チャンク内のバイトに対応する。例えば、マスク部分のビット１は、データの非圧縮チャンク内の第１バイトに対応することができ、マスク部分のビット２は、データの非圧縮チャンク内の第２バイトに対応することができる、など。他の実施形態では、圧縮された出力チャンクのマスク部分内の２つ以上のビットが、データの非圧縮チャンク内のあるバイトに対応する。かような実施形態で、圧縮された出力チャンクのマスク部分内のビットは、非圧縮チャンク内の対応するバイトだけでなくそのおおよその大きさも示すことができる。
【００１１】
マスク部分の個々のビットが非圧縮チャンク内のバイトに対応する場合に、圧縮ユニットは、圧縮された出力チャンクのマスク部分内の各ビットを、データの非圧縮チャンク内の対応するバイトが全てゼロを含むときに（すなわち、“ゼロバイト”）論理偽（false）（本明細書中「論理０」とも呼ばれ得る。）に設定する。圧縮ユニットはまた、データの非圧縮チャンク内の対応するバイトが少なくとも１つの非ゼロビットを含むときに（すなわち、“非ゼロバイト”）、圧縮された出力チャンクのマスク部分内の各ビットを論理真（true）（本明細書中「論理１」とも呼ばれ得る。）に設定する。このようにして、圧縮された出力チャンクのマスク部分は、データの非圧縮チャンクのゼロバイト及び非ゼロバイトの存在及び位置を符号化する。
【００１２】
圧縮ユニットは、データの非圧縮チャンク内の非ゼロバイトの数を決定することによって、圧縮された出力チャンクのデータ部分を生成する。圧縮ユニットは次いで、データの非圧縮チャンク内の非ゼロバイトの数と、圧縮された出力チャンクのデータ部分内で利用可能なバイトの数とに基づき、データの非圧縮チャンクの各非ゼロバイトを格納するために利用可能である圧縮された出力チャンクのデータ部分内のビットの数を決定する。例えば、データの圧縮されたチャンクのデータ部分が２４バイト幅（すなわち、１９２ビット）であり、データの非圧縮チャンク内に４７の非ゼロバイトがある場合に、データの非圧縮チャンクから各非ゼロバイトを格納するためには、４ビットがデータ部分内で利用可能である。
【００１３】
いくつかの実施形態において、圧縮ユニットはまた、データの非圧縮チャンクの非ゼロバイトを格納するために利用可能である圧縮された出力チャンクのデータ部分内で、もしあれば、追加ビットの数を決定することもできる。上記の例において、例えば、４つの追加ビットが、非ゼロバイトを格納するために利用可能である（すなわち、１９２　ｍｏｄ　４７＝４ビット）。圧縮ユニットは、それらの追加ビットを、データの非圧縮チャンク内の１つ以上の非ゼロバイトを切り捨てる前に、その１つ以上の非ゼロバイトに割り当てることができる。例えば、圧縮ユニットは、それらの追加ビットを、圧縮された出力チャンクのデータ部分内の最初の数バイトに割り当ててもよい。
【００１４】

圧縮ユニットは、次いで、データの非圧縮チャンク内の非ゼロバイトを、各非ゼロバイトを格納するためにデータ部分内で利用可能な決定されたビット数（すなわち、上記の例では、４）まで切り捨てる。一実施形態において、圧縮ユニットは、データ部分内の利用可能なビット数に収まるように非ゼロバイトの最下位ビット（Least Significant Bit(s)）（“ＬＢＳ”）を切り捨てる。他の実施形態では、圧縮ユニットは、非ゼロバイトの最上位ビット（Most Significant Bit(s)）（“ＭＳＢ”）を切り捨てる。圧縮ユニットは、次いで、切り捨てられた非ゼロバイトを、圧縮された出力チャンクのデータ部分に格納する。マスク部分及びデータ部分を含む圧縮された出力チャンクは、次いで、例えば、ＤＮＮプロセッサ内のオンボード・メモリ、又はＤＮＮプロセッサのアプリケーションホストのオフボード・メモリへ、出力され得る。

【００１５】

ＤＮＮモジュールはまた、上述されたように圧縮されたデータのチャンクを復元することができる復元ユニットを含むこともできる。例えば、復元ユニットは、ＤＮＮプロセッサ内のメモリ又はアプリケーションホストのメモリからデータの圧縮されたチャンクを受け取ることができる。復元ユニットは、次いで、圧縮された出力チャンクのマスク部分内の論理真ビットの数に基づき、データの非圧縮チャンク内の非ゼロバイトの数を決定することができる。復元ユニットはまた、圧縮された出力チャンクのマスク部分内の論理真ビットの位置に基づき、データの非圧縮チャンク内の非ゼロバイトの位置を決定することもできる。復元ユニットは、同様にして、データの非圧縮チャンク内のゼロバイトの位置を決定することができる。

【００１６】

復元ユニットはまた、切り捨てられた非ゼロバイトを圧縮された出力チャンクのデータ部分に格納するために圧縮ユニットによって使用されたビットの数を決定することもできる。復元ユニットは、データの非圧縮チャンク内の非ゼロバイトの数と、圧縮された出力チャンクのデータ部分内で利用可能なバイトの数とに基づき、切り捨てられた各非ゼロバイトを格納するために使用されたビットの数を決定することができる。

【００１７】

上記の例では、例えば、データの圧縮されたチャンクのデータ部分が２４バイト幅（すなわち、１９２ビット）であり、データの非圧縮チャンク内に４７の非ゼロバイトがある場合に、圧縮ユニットは、データの非圧縮チャンクの切り捨てられた各非ゼロバイトを格納するためにデータ部分において４ビットを利用した。復元ユニットはまた、もしあれば、圧縮ユニットが圧縮された出力チャンクのデータ部分に格納された切り捨てられた非ゼロバイトのうちの１つ以上に割り当てた追加ビットの数を決定することもできる。

【００１８】

論理０である圧縮された出力チャンクのマスク部分内の各ビット位置について、復元ユニットは、ゼロバイトを、復元された出力チャンクの対応する部分に挿入する。論理１であるマスク部分内の各ビット位置について、復元ユニットは、圧縮された出力チャンクのデータ部分内の対応する位置からの切り捨てられた非ゼロバイトを、圧縮された出力チャンクの圧縮中に切り捨てられたビットの数に等しい数のゼロビットとともに、復元された出力チャンクの対応する位置に挿入する。ゼロビットは、どのビットが圧縮中に切り捨てられたかに応じて、切り捨てられた非ゼロバイトのＬＳＢ又はＭＳＢに挿入され得る。

【００１９】

いくつかの実施形態において、復元ユニットはまた、復元された出力チャンクに格納された切り捨てられた非ゼロバイトのうちの１つ以上にオフセット（例えば、００００００１）を付加する。例えば、オフセットは、圧縮後にゼロバイトになるデータの非圧縮チャンクの非ゼロバイトに加えられ得る。このようにして、非ゼロバイトは、圧縮及び復元される場合にゼロバイトにならない。他の実施形態では、オフセットは、復元された出力チャンク内の全てのバイトに付加され得る。

【００２０】

先に簡単に説明されたように、本明細書で開示される技術の実装は、ＤＮＮモジュールに

おけるメモリバス帯域幅利用量を低減し、ＤＮＮモジュールが処理動作をより速く完了することを可能にし、電力消費量を削減することができる。本明細書で具体的に特定されていない技術的利点も、開示されている技術の実施を通じて実現され得る。

【００２１】

当然ながら、上記の対象は、コンピュータにより制御される装置、コンピュータにより実施される方法、コンピュータ装置として、又はコンピュータ可読媒体などの製品として実施され得る。それら及び様々な他の特徴は、続く詳細な説明を読み、添付の図面を参照することで、明らかになるだろう。

【００２２】

この概要は、開示される技術のいくつかの態様について概要を簡単に紹介するために設けられており、それらの態様は、詳細な説明において更に後述される。この概要は、請求されている対象の重要な特徴又は必須の特徴を特定する意図はなく、この概要が、請求されている対象の適用範囲を制限するために使用されることも意図しない。更には、請求されている対象は、本開示のいずれかの部分で述べられているありとあらゆる欠点を解消する実施に限られない。

【図面の簡単な説明】

【００２３】

【図１】一実施形態に従って、本明細書で開示される技術の態様を実装するＤＮＮモジュールの構成及び動作の態様を示すコンピューティングアーキテクチャ図である。

【図２Ａ】一実施形態に従って、活性化データを圧縮するＤＮＮモジュールの構成及び動作の態様を示すコンピューティングシステムアーキテクチャ図である。

【図２Ｂ】一実施形態に従って、活性化データを圧縮するＤＮＮモジュールの構成及び動作の態様を示すコンピューティングシステムアーキテクチャ図である。

【図３】一実施形態に従って、圧縮されていない活性化データのチャンクの例を参照して、活性化データを圧縮するためのＤＮＮモジュールの動作の態様を説明するデータ構造図である。

【図４】本明細書で開示される一実施形態に従って、活性化データを圧縮するための開示されているＤＮＮモジュールの動作の態様を説明するルーチンを示すフロー図である。

【図５Ａ】一実施形態に従って、活性化データを復元するためのＤＮＮモジュールの構成及び動作の態様を示すコンピューティングシステムアーキテクチャ図である。

【図５Ｂ】一実施形態に従って、活性化データを復元するためのＤＮＮモジュールの構成及び動作の態様を示すコンピューティングシステムアーキテクチャ図である。

【図６】一実施形態に従って、圧縮された活性化データのチャンクの例を参照して、活性化データを復元するためのＤＮＮモジュールの動作の態様を説明するデータ構造図である。

【図７】本明細書で開示される一実施形態に従って、活性化データを復元するための開示されているＤＮＮモジュールの動作の態様を説明するルーチンを示すフロー図である。

【図８】一実施形態に従って、本明細書で提示されるＤＮＮモジュールのためのアプリケーションホストとして動作することができるコンピュータ装置のためのコンピュータハードウェア及びソフトウェアアーキテクチャを例示するコンピュータアーキテクチャ図である。

【図９】本明細書で提示される様々な実施形態に従って、開示される技術の態様が実装され得る分散コンピューティング環境を説明するネットワーク図である。

【発明を実施するための形態】

【００２４】

続く詳細な説明は、メモリバス帯域幅の利用量を低減するために活性化データを圧縮及び復元することができるＤＮＮモジュールを対象とする。先に簡単に説明されたように、開示される技術の実装は、ＤＮＮモジュールにおけるメモリバス帯域幅利用量を低減し、ＤＮＮモジュールが処理動作をより速く完了することを可能にし、電力消費量を削減することができる。本明細書で具体的に述べられない他の技術的利点も、開示されている対象の実施を通じて実現され得る。

【００２５】
本明細書中に記載される対象は、ハードウェアＤＮＮモジュールの一般的状況において提示されるが、当業者は、他のタイプのコンピューティングシステム及びモジュールと組み合わせて他の実施が実行され得ると認めるだろう。当業者はまた、本明細書中に記載される対象が、手持ち式デバイス、マルチプロセッサシステム、マイクロプロセッサに基づく又はプログラム可能な家電製品、デバイス（例えば、装用式コンピュータ装置、自動車、ホームオートメーション、など）に埋め込まれたコンピューティング又はプロセッシングシステム、ミニコンピュータ、メインフレームコンピュータ、などを含む他のコンピュータシステム構成により実施され得るであると認めるだろう。
【００２６】
以下で更に詳細に記載されるように、そのニューロンの出力を圧縮するよう構成されるＤＮＮモジュールが、開示される。圧縮された出力は、ＤＮＮモジュールのメモリに、又はＤＮＮモジュールのためのアプリケーションホストによって提供されるメモリなどの、ＤＮＮモジュールの外にあるメモリに、記憶され得る。ＤＮＮモジュールは、前に圧縮されたデータを後に復元し、そして、復元されたデータをニューロンへ供給することができる。
【００２７】
一実施形態によれば、ＤＮＮプロセッサ内の圧縮ユニットは、圧縮されていない活性化データの固定長のチャンク（例えば、６４バイト）を一定の圧縮比（例えば、２：１）で圧縮する。圧縮ユニットによって生成される圧縮された活性化データは、固定長（例えば、３２バイト）を有するデータのチャンクを含むことができ、固定長のマスク部分（例えば、８バイト）と、固定長のデータ部分（例えば、２４バイト）とを含む。
【００２８】
一実施形態において、圧縮された出力チャンクのマスク部分のビットは、圧縮されていない入力チャンク内のバイトに対応する。例えば、マスク部分の第１ビットは、圧縮されていない入力チャンク内の第１バイトに対応することができ、マスク部分の第２ビットは、圧縮されていない入力チャンク内の第２バイトに対応することができる、など。圧縮された活性化データのマスク部分内のビットは、圧縮されていない入力チャンク内の対応するバイトがゼロである場合には論理０にセットされ得、圧縮されていない入力チャンク内の対応するバイトが非ゼロである場合には論理１にセットされ得る。
【００２９】
先に簡単に説明されたように、いくつかの実施形態において、圧縮された出力チャンクのマスク部分内の２つ以上のビットが、データの非圧縮チャンク内のあるバイトに対応する。かような実施形態で、圧縮された出力チャンクのマスク部分内のビットは、非圧縮チャンク内の対応するバイトだけでなくそのおおよその大きさも示すことができる。
【００３０】
圧縮された出力チャンクのデータ部分は、圧縮されたデータ部分内の利用可能なビットの数を用いて非圧縮入力チャンクの非ゼロバイトを表すよう、切り捨てられた非圧縮入力チャンクの非ゼロバイトを含む。いくつかの実施形態において、各非ゼロバイトごとの、圧縮された出力チャンクのデータ部分内の利用可能なビットの数は、データ部分内の利用可能なビットの総数（例えば、１９２ビット）を非圧縮入力チャンク内の非ゼロバイトの数で割ることによって決定される。この計算の結果は、非圧縮入力チャンク内の非ゼロデータの各バイトを表すために利用可能である圧縮された出力チャンクのデータ部分内のビットの数を示す。如何なる残りのビットも、圧縮された出力チャンクのデータ部分において、非ゼロ値のうちのいくつかを表すための追加ビットを提供するために使用され得る。
【００３１】
非圧縮入力チャンク内の各非ゼロバイトを表すための圧縮された出力チャンクのデータ部分内で利用可能なビットの数が決定されると、非圧縮入力チャンク内の非ゼロ値のＬＳＢが、利用可能なビットの数に収まるよう切り捨てられる。他の実施形態では、非ゼロ値のＭＳＢが切り捨てられてもよい。次いで、切り捨てられた非ゼロ値が、圧縮された出力チャンクのデータ部分に格納され得る。このプロセスは、圧縮されていない入力活性化デー

タの各チャンクごとに繰り返され得る。次いで、圧縮された出力チャンクが、後の復元及びニューロンによる使用のためにオン又はオフモジュール・メモリに記憶され得る。

【００３２】

開示されるＤＮＮモジュールはまた、上述されたように圧縮ユニットによって圧縮された活性化値を復元する復元ユニットを含むこともできる。復元ユニットは、マスク部分及びデータ部分を含む圧縮された活性化データのチャンクを受け取る。復元ユニットは、復元された出力チャンクに存在することになる非ゼロバイトの数と、復元された出力チャンク内のそれらの位置とを特定するために、マスク部分のビットを利用することができる。マスクはまた、復元された出力チャンク内のゼロバイトの位置も示す。

【００３３】

いくつかの実施形態において、復元ユニットは、圧縮されたチャンクのデータ部分内の利用可能なビットの総数（例えば、１９２ビット）を、マスクによって指定されている非圧縮入力チャンク内の非ゼロバイトの数で割ることによって、各非ゼロバイトを表すために圧縮ユニットによって使用されたビットの数を決定する。復元ユニットはまた、圧縮されたチャンクのデータ部分内の非ゼロ値のうちのいくつか（例えば、最初のＮ個の値）を表す追加ビットを供給するために圧縮ユニットがいずれかの残りのビットを使用したと仮定することもできる。

【００３４】

論理０であるマスク内の各ビット位置について、復元ユニットは、ゼロバイトを、復元された出力チャンク内にその対応する位置で挿入することができる。論理１であるマスク内の各ビット位置について、復元ユニットは、圧縮された入力チャンクのデータ部分内の対応する位置からの切り捨てられた非ゼロバイトを、復元された出力チャンク内の対応する位置に挿入する。復元ユニットはまた、非ゼロ値の、必要に応じてＬＳＢ又はＭＳＢにおいて、圧縮中に切り捨てられたそれらのビットを置き換えるようゼロを挿入する。

【００３５】

いくつかの実施形態において、復元ユニットは、非ゼロの非圧縮値が復元されるときにゼロバイトにならないことを確かにするよう、切り捨てられた非ゼロ値にオフセット値を付加する。次いで、復元された出力チャンクが、ニューロンによる使用のためにオン又はオフモジュール・メモリに記憶され得る。ＤＮＮモジュール、圧縮ユニット、及び復元ユニットの動作に関する更なる詳細は、以下で与えられる。

【００３６】

続く詳細な説明では、本願の部分を形成し、実例として具体的な構成又は例が示されている添付の図面が、参照される。これより図面を参照すると、図面では、いくつかの図を通して同じ番号が同じ要素を表しており、メモリバス帯域幅の利用量を低減するために活性化データを圧縮及び復元することができるＤＮＮモジュールの態様が記載される。

【００３７】

図１は、一実施形態に従って、本明細書で開示される技術の態様を実装するＤＮＮモジュール１０５の構成及び動作の態様を示すコンピューティングアーキテクチャ図である。本明細書で開示されるＤＮＮモジュール１０５は、例えば、制限なしに、物体検出、意味的標識付け、及び特徴抽出などの分類問題（及び関連する問題）を解くよういくつかの実施形態で構成される。

【００３８】

この機能を提供するために、ＤＮＮモジュール１０５は、リコールオンリー（recall-only）ニューラルネットワークを実装し、多種多様なネットワーク構造をプログラムでサポートすることができる。ＤＮＮモジュール１０５によって実装されるネットワークの教育は、サーバーファーム、データセンター、又は他の適切なコンピューティング環境においてオンラインで実行され得る。ＤＮＮを教育する結果は、“重み”又は“カーネル”として知られ得るパラメータの組である。それらのパラメータは、入力に適用可能な変換関数を表し、その結果が分類又は意味的に標識付けされた出力である。

【００３９】

本明細書で開示されるＤＮＮモジュール１０５は、スーパースカラープロセッサと見なされ得る。ＤＮＮモジュール１０５は、ニューロン１０５Ｆと呼ばれる複数の実行ユニットへ１つ以上の命令をディパッチすることができる。実行ユニットは、“同時ディスパッチ同時完了”（simultaneous dispatch simultaneous complete）であることができ、各実行ユニットが他の実行ユニットの夫々と同期する。ＤＮＮモジュール１０５は、単一命令ストリーム多重データストリーム（single instruction stream, multiple data stream）（“ＳＩＭＤ”）アーキテクチャとして分類され得る。

【００４０】

ＤＮＮモジュール１０５は、多数のニューロン１０５Ｆ（例えば、２の累乗）を含む。ニューロン１０５Ｆは、脳内の生体ニューロンをモデル化するために使用される人工ニューラルネットワーク内の基本単位である。ニューロン１０５Ｆのモデルは、活性化関数を作用させて入力ベクトルと重みベクトルとの内積にバイアスを加えたものを含むことができる。本明細書中に記載されるＤＮＮモジュール１０５内のニューロン１０５Ｆによって実行される処理は、人工ニューロンに密接にマッピングされる。

【００４１】

ＤＮＮモジュール１０５内の各ニューロン１０５Ｆは、加重加算、最大プーリング（max pooling）、バイパス、及び場合によっては、他のタイプの演算を実行可能である。ニューロン１０５Ｆは、クロック周期ごとに入力及び重みデータを処理する。各ニューロン１０５Ｆは、ＤＮＮモジュール１０５内のカーネルデータのフローを最小限にするよう、カーネル内の進捗状況に関して、他の全てのニューロン１０５Ｆと同期する。

【００４２】

各ニューロン１０５Ｆは、乗算器、加算器、比較器、及び多数の積算器（図１に図示せず。）を含むことができる。複数の積算器を有することによって、ニューロン１０５Ｆは、一度に複数の異なるアクティブなカーネルのコンテキストを保持することができる。各積算器は、ＢａＳＲＡＭ１５０（後述される。）の読み出しからロード可能である。積算器は、それら自体を、他のニューロン１０５Ｆからの他の積算器の内容と加算することができる。

【００４３】

ＤＮＮモジュール１０５は、画像データなどの平面（planar）データを入力として受け入れる。ＤＮＮモジュール１０５への入力は、しかしながら、画像データに限られない。むしろ、ＤＮＮモジュール１０５は、一様平面の形でＤＮＮモジュール１０５へ与えられる如何なる入力データにも作用することができる。１つの特定の実施形態では、ＤＮＮモジュール１０５は、入力として、多断面（multi-planar）１バイト又は２バイトデータフレームを受け入れることができる。

【００４４】

各入力フレームは、Ｎ×Ｋ×Ｈ×Ｗ組のカーネルにより畳み込まれ得る。ここで、Ｎは、カーネルの数であり、Ｋは、カーネルごとのチャネルの数であり、Ｈは高さであり、Ｗは幅である。畳み込みは、重なり合ったインターバルで入力データに対して実行され、インターバルは、Ｘ及びＹ方向におけるストライドによって定義される。これらの関数は、ニューロン１０５Ｆによって実行され、ＤＮＮモジュール１０５及びソフトウェア管理下の制御レジスタによって管理される。

【００４５】

ＤＮＮモジュール１０５は、３つの主なデータタイプ、すなわち、重み、入力データ／特徴マップ、及び活性化データをサポートする。入力データ／特徴マップ及び活性化データは、ほとんどの場合に、レイヤの出力に言及するときに語「活性化データ」が使用されるとの区別により、同じデータについての２つの名称である。レイヤの入力を言及するときには、語「入力データ／特徴マップ」が使用される。

【００４６】

ＤＮＮモジュール１０５内のニューロン１０５Ｆは、それらの入力の加重和を計算し、加重和を“活性化関数”又は“伝達関数”に通す。伝達関数は、一般にシグモイド関数を有する

が、区分線形関数（piecewise linear function）、ステップ関数、又は他のタイプの関数でも構成され得る。活性化関数は、ニューロン１０５Ｆが、分類境界が非線形である入力及び所望の出力のより大きい組を教育することを可能にする。

【００４７】

ＤＮＮモジュール１０５は、ニューラルネットワークのレイヤに対応するレイヤ記述子のリストに作用する。レイヤ記述子のリストは、ＤＮＮモジュール１０５によって命令として教育され得る。それらの記述子は、メモリからＤＮＮモジュール１０５にプリフェッチされ、順に実行され得る。記述子リストは、ＤＮＮモジュール１０５への命令の組として働く。ソフトウェアツール及び／又はコンパイラが、ＤＮＮモジュール１０５で実行される記述子リストを生成するよう、ＤＮＮモジュール１０５の外部にあるデバイスで実行され得る。

【００４８】

一般に、記述子には２つの主な種類があり、メモリ間移動（memory-to-memory move）（“Ｍ２Ｍ”）記述子及びオペレーション記述子である。Ｍ２Ｍ記述子は、オペレーション記述子による消費のためにメインメモリからローカルバッファ（すなわち、後述されるラインバッファ１２５）へ又はローカルバッファからメインメモリへデータを移動させるために使用され得る。Ｍ２Ｍ記述子のターゲットパイプラインは、内部ＤＭＡエンジン１０５Ｂ又はコンフィグレーションレジスタ１０５Ｇであることができ、一方、オペレーション記述子のターゲットパイプラインは、ニューロン１０５Ｆであることができる。

【００４９】

オペレーション記述子は、ローカルの静的ランダムアクセスメモリ（“ＳＲＡＭ”）に位置するデータ構造に対してニューロン１０５Ｆが実行すべき具体的な動作を指定する。オペレーション記述子は、順に処理され、多種多様なレイヤ動作が可能であり、そのうちの少なくともいくつかが本明細書中に記載されている。

【００５０】

図１に表されているように、ＤＮＮモジュール１０５は、一意のＬ１及びＬ２バッファ構造を有するメモリサブシステムを具備する。図１に示されるＬ１及びＬ２バッファは、特にニューラルネットワーク処理のために設計される。例として、Ｌ２バッファ１５０は、選択された周波数で動作する高速プライベートインターフェイスとともに、選択された記憶容量を保持することができる。Ｌ１バッファ１２５は、カーネルと活性化データとの間で分けられ得る選択された記憶容量を保持することができる。Ｌ１バッファ１２５は、本明細書で“ラインバッファ１２５”と呼ばれてもよく、Ｌ２バッファ１５０は、本明細書でＢａＳＲＡＭ１５０と呼ばれてもよい。

【００５１】

計算データ（すなわち、入力データ、重み及び活性化データ）は、いくつかの実施形態において、行優先（row-major）でＢａＳＲＡＭ１５０に格納される。計算データは、２つのラインバッファとして体系化され得る。一方のラインバッファは入力データを含み、本明細書で“入力バッファ”と呼ばれてよく、他方のラインバッファは、本明細書で“重みバッファ”と呼ばれてよく、カーネル重みを含む。ラインバッファは、ロード／ストアユニット１０５ＣによってＢａＳＲＡＭ１５０から満たされる。データは、各ラインバッファにおいて、その所定の容量に達するまで蓄積される。ラインバッファデータは、次いで、いくつかの実施形態において、シャドウバッファにコピーされ、ニューロン１０５Ｆに渡される。

【００５２】

ＤＮＮモジュール１０５は、プリフェッチユニット１０５Ａ、セーブ／リストアユニット１０５Ｅ、レイヤコントローラ１０５Ｄ、及びレジスタインターフェイス１０５Ｇを含むがこれらに限られない多数の他のコンポーネントを有することもできる。ＤＮＮモジュール１０５は、いくつかの実施形態において、追加又は代替のコンポーネントを含むことができる。

【００５３】

ＤＮＮモジュール１０５は、いくつかの構成において、他の外部コンピューティングコンポーネントと協働して動作する。例えば、ＤＮＮモジュール１０５は、いくつかの実施形態において、ホストアプリケーションプロセッサ・システムオンチップ（system on chip）（“ホストＳＯＣ”）１３０へ接続される。ＤＮＮモジュール１０５は、例えば、ＰＣＩｅインターフェイスを通じてホストＳｏＣ１３０へ接続され得る。ＰＣＩｅエンドポイント１３５などの適切なＰＣＩｅコンポーネントが、それらの接続を可能にするために利用され得る。

【００５４】

ホストＳｏＣ１３０は、ＤＮＮモジュール１０５のためのアプリケーションプロセッサとなる。メインオペレーティングシステム、アプリケーション、及び補助センサ処理が、ホストＳｏＣ１３０によって実行される。ホストＳｏＣ１３０はまた、画像データなどの入力データをＤＮＮモジュール１０５へ供給する外部カメラなどの入力データ源１０２へも接続され得る。

【００５５】

ＤＤＲ　ＤＲＡＭ１５５も、ホストＳｏＣ１３０へ接続され得、メインシステムメモリとして使用され得る。このメモリは、メモリコントローラ１４５を用いて高帯域幅ファブリック１２０（例えば、ＰＣＩｅバス）を越えてホストＳｏＣ１３０からアクセス可能である。高帯域幅ファブリック１２０は、双方向の直接メモリアクセス（“ＤＭＡ”）スモールメッセージングトランザクション及びより大きいＤＭＡトランザクションを提供する。ブリッジ１１５及び低帯域幅ファブリック１１０は、サブモジュール構成及び他の機能のためにＤＮＮモジュール１０５をホストＳｏＣ１３０へ接続することができる。

【００５６】

ＤＮＮモジュール１０５は、メインメモリ１５５へ及びそれからデータを移動させるよう構成されるＤＭＡエンジン１０５Ｂを含むことができる。ＤＭＡエンジン１０５Ｂは、いくつかの実施形態において、２つのチャネルを具備する。一方のチャネルは、フェッチ動作記述子に専用であり、一方、他方のチャネルは、Ｍ２Ｍ動作に専用である。本文脈中の記述子は、メモリの中身を移動させるために使用されるＤＭＡ記述子であり、上記のオペレーション記述子と混同されるべきではない。

【００５７】

ローカルＢａＳＲＡＭ１５０をオフロードするよう、且つ、入力及び重みデータのための更なる空間を提供するよう、活性化出力は、任意に、ＤＤＭメモリ１５５に直接にストリーミングされ得る。データをＤＤＲメモリ１５５にストリーミングするとき、ＤＮＮモジュール１０５は、高帯域幅ファブリック１２０上でのバーストトランザクションのために十分なデータを蓄積することになり、且つ、ニューロン１０５Ｆに対するバックプレッシャを最小限とするよう十分なトランザクションをバッファすることになる。ＤＮＮモジュール１０５の動作に関する更なる詳細は、以下で与えられる。

【００５８】

図２Ａ及び図２Ｂは、一実施形態に従って、活性化データを圧縮するＤＮＮモジュール１０５の構成及び動作の態様を示すコンピューティングシステムアーキテクチャ図である。図２Ａに示され且つ先に簡単に記載されたように、ＤＮＮモジュール１０５は、１つ以上のニューロン１０５Ｆと、圧縮ユニット２００とを含む。圧縮ユニット２００は、いくつかの実施形態において、ロード／ストアユニット１０５Ｃによって実装されるが、他の実施形態では他の方法で実装されてもよい。

【００５９】

圧縮ユニット２００は、ニューロン１０５Ｆの１つ以上によって生成される活性化データの非圧縮チャンク２０２を受け取ることができる。データの非圧縮チャンク２０２は、いくつかの実施形態において、６４バイトなどの一定数のバイトを含む。

【００６０】

圧縮ユニット２００は、活性化データの圧縮されたチャンク２０４を生成するようデータの非圧縮チャンク２０２を圧縮することができる。活性化データの圧縮されたチャンク２

０４は、次いで、メモリ２０６に格納され得る。例えば、活性化データの圧縮されたチャンク２０４は、アプリケーションホストによって提供されるＬＰＤＤＲ４メモリ１５５に格納されても、あるいは、ＤＮＮモジュール１０５によって提供されるＢａＳＲＡＭ１５０に格納されてもよい。以下で更に詳細に開示されるように、本明細書で開示される技術は、ＬＰＤＤＲ４メモリ１５５又はＢａＳＲＡＭ１５０から圧縮又は復元された活性化データを記憶し又は読み出すときにメモリバス利用の利用量を低減するために圧縮及び復元を利用することができる。これらの技術に関する更なる詳細は、図２Ａ～９に関して以下で開示される。

【００６１】

図２に表されているように、圧縮ユニット２００は、データの圧縮された出力チャンク２０４のマスク部分２０８及びデータ部分２１を生成することができる。圧縮された出力チャンク２０４のマスク部分２０８は、データの非圧縮チャンク２０２内の一定数のバイトに等しいビットの数を含む。例えば、データの非圧縮チャンク２０２が６４バイトのデータを含む場合には、圧縮された出力チャンク２０４のマスク部分２０８は６４ビット（すなわち、８バイト）を含むことになる。

【００６２】

圧縮された出力チャンク２０４のマスク部分２０８内の各ビットは、いくつかの実施形態において、データの非圧縮チャンク２０２内のバイトに対応する。例えば、マスク部分２０８のビット１は、データの非圧縮チャンク２０２内の第１バイトに対応することができ、マスク部分２０８のビット２は、データの非圧縮チャンク２０２内の第２バイトに対応することができる、など。

【００６３】

圧縮ユニット２００は、圧縮された出力チャンク２０４のマスク部分２０８内の各ビットを、データの非圧縮チャンク２０２内の対応するバイトがゼロバイトである場合に論理０にセットする。圧縮ユニット２００はまた、圧縮された出力チャンク２０４のマスク部分２０８内の各ビットを、データの非圧縮チャンク２０２内の対応するバイトが非ゼロバイトである場合に論理１にセットする。このようにして、圧縮された出力チャンク２０４のマスク部分２０８は、データの非圧縮チャンク２０２内のゼロバイト及び非ゼロバイトの存在及び位置を符号化する。

【００６４】

圧縮ユニット２００は、データの非圧縮チャンク２０２内の非ゼロバイトの数を決定することによって、圧縮された出力チャンク２０４のデータ部分２１０を生成する。次いで、圧縮ユニット２００は、データの非圧縮チャンク２０２内の非ゼロバイトの数と、圧縮された出力チャンク２０４のデータ部分２１０で利用可能なバイトの数とに基づき、データの非圧縮チャンク２０２の各非ゼロバイトを格納するために利用可能である圧縮された出力チャンク２０４のデータ部分２１０内のビットの数を決定する。例えば、データの圧縮されたチャンク２０４のデータ部分２１０が２４バイト幅（すなわち、１９２ビット）であり、データの非圧縮チャンク２０２内に４７の非ゼロバイトがある場合に、データの非圧縮チャンク２０２から各非ゼロバイトを格納するためには、４ビットがデータ部分２１０内で利用可能である。

【００６５】

いくつかの実施形態において、圧縮ユニット２００はまた、データの非圧縮チャンク２０２の非ゼロバイトを格納するために利用可能である圧縮された出力チャンク２０４のデータ部分２１０内で、もしあれば、追加ビットの数を決定することもできる。上記の例において、例えば、４つの追加ビットが、非ゼロバイトを格納するために利用可能である（すなわち、１９２　ｍｏｄ　４７＝４ビット）。圧縮ユニット２００は、それらの追加ビットを、データの非圧縮チャンク２０２内の１つ以上の非ゼロバイトを切り捨てる前に、その１つ以上の非ゼロバイトに割り当てることができる。例えば、圧縮ユニット２００は、それらの追加ビットを、圧縮された出力チャンク２０４のデータ部分２１０内の最初のＮバイトに割り当ててもよい。

【００６６】
圧縮ユニット２００は、次いで、データの非圧縮チャンク２０２内の非ゼロバイトを、各非ゼロバイトを格納するためにデータ部分２１０内で利用可能な決定されたビット数（すなわち、上記の例では、４）まで切り捨てる。一実施形態において、圧縮ユニット２００は、データ部分２１０内の利用可能なビット数に収まるように非ゼロバイトのＬＢＳを切り捨てる。他の実施形態では、圧縮ユニット２００は、非ゼロバイトのＭＳＢを切り捨てる。圧縮ユニット２００は、次いで、切り捨てられた非ゼロバイトを、圧縮された出力チャンク２０４のデータ部分２１０に格納する。マスク部分２０８及びデータ部分２１０を含む圧縮された出力チャンク２０４は、次いで、例えば、ＤＮＮモジュール１０５内のオンボード・メモリ、又はＤＮＮモジュール１０５のアプリケーションホストのオフボード・メモリへ、出力され得る。
【００６７】
先に簡単に述べられたように、圧縮された出力チャンク２０４のマスク部分２０８内の２つ以上のビットは、いくつかの実施形態において、データの非圧縮チャンク２０２内のあるバイトに対応する。かような実施形態で、圧縮された出力チャンク２０４のマスク部分２０８内のビットは、非圧縮チャンク２０２内の対応するバイトだけでなくそのおおよその大きさも示すことができる。例えば、制限なしに、マスク部分２０８は、データの非圧縮チャンク２０２内のバイトごとに２ビットを含んでよい。この例では、００は、データの非圧縮チャンク２０２内の対応する非ゼロバイトのＭＳＢが０であることを示すことができ、０１は、ＭＳＢが＜６４であることを示すことができ、１０は、ＭＳＢが＜１２８であることを示すことができ、１１は、ＭＳＢ＞１２８であることを示すことができる。これらの値は、データの非圧縮チャンク２０２内のバイトのどのＭＳＢが切り捨てられ得るかを特定するために利用され得る。例えば、特定のバイトのＭＳＢが＜６４である場合に、上から２つのＭＳＢがデータの損失なしで切り捨てられ得る。
【００６８】
図３は、一実施形態に従って、圧縮されていない活性化データのチャンク２０２の例を参照して、圧縮されていない活性化データのチャンク２０２を圧縮するためのＤＮＮモジュール１０５の動作の態様を説明するデータ構造図である。図３に示される例では、活性化データの非圧縮チャンク２０２は、６４バイト長である。圧縮されていない活性化データのチャンク２０２のバイト０、１及び６３はゼロバイトである。圧縮されていない活性化データのチャンク２０２のバイト２、３及び６２は非ゼロバイトであり、夫々値１１２、１２１及び２を格納する。例となる圧縮されていない活性化データのチャンク２０２のバイト４乃至６１は、ゼロ又は非ゼロバイトを格納することができる。
【００６９】
上述されたように、圧縮ユニット２００は、活性化データの非圧縮チャンク２０２内のゼロバイト及び非ゼロバイトの存在及び位置を符号化するマスク部分２０８を生成することができる。この例では、例えば、マスク部分２０８のビット０、１及び６３が、活性化データの非圧縮チャンク２０２内の対応する位置にあるゼロバイトの存在を示すよう、論理０にセットされている。同様に、マスク部分２０８のビット２、３及び６２は、活性化データの非圧縮チャンク２０２のバイト２、３及び６２が非ゼロバイトを格納していることを示すよう、論理１にセットされている。
【００７０】
上述されたように、圧縮ユニット２００は、データの非圧縮チャンク２０２内の非ゼロバイトの数を決定することによって、圧縮された出力チャンク２０４のデータ部分２１０を生成する。図３に示される例では、例えば、データの非圧縮チャンク２０２は、４７個の非ゼロバイト（図３では、その全てが示されているわけではない。）を含む。圧縮ユニット２００は、次いで、データの非圧縮チャンク２０２内の非ゼロバイトの数と、圧縮された出力チャンク２０４のデータ部分２１０で利用可能なバイトの数とに基づき、データの非圧縮チャンク２０２の各非ゼロバイトを格納するために利用可能である圧縮された出力チャンク２０４のデータ部分２１０内のビットの数を決定する。

【００７１】
図３に示される例では、例えば、データの圧縮されたチャンク２０４のデータ部分２１０は２４バイト幅（すなわち、１９２ビット）であり、データの非圧縮チャンク２０２内には４７の非ゼロバイトがある。結果として、データの非圧縮チャンク２０２から各非ゼロバイトを格納するためには、４ビットがデータ部分２１０内で利用可能である（すなわち、１９２／４７＝４余り４）。
【００７２】
やはり上述されたように、圧縮ユニット２００はまた、データの非圧縮チャンク２０２の非ゼロバイトを格納するために利用可能である圧縮された出力チャンク２０４のデータ部分２１０内で、もしあれば、追加ビットの数を決定することもできる。図３に示される例では、例えば、４つの追加ビットが、非ゼロバイトを格納するために利用可能である（すなわち、１９２　ｍｏｄ　４７＝４ビット）。圧縮ユニット２００は、それらの追加ビットを、データの非圧縮チャンク２０２内の１つ以上の非ゼロバイトを切り捨てる前に、その１つ以上の非ゼロバイトに割り当てることができる。図３に示される例では、追加ビットのうちの１つが、圧縮されていない活性化データのチャンク２０２の最初の４つの非ゼロバイトの夫々に割り当てられている。結果として、活性化データの非圧縮チャンク２０２の最初の４バイトは、４ではなく５ビットに切り捨てられる。
【００７３】
圧縮ユニット２００は、次いで、データの非圧縮チャンク２０２内の非ゼロバイトを、各非ゼロバイトを格納するためにデータ部分２１０内で利用可能な決定されたビット数（すなわち、上記の例では、４ビット。ただし、最初の４つの非ゼロバイトのためには５ビット。）まで切り捨てる。図３に示される例では、一実施形態において、圧縮ユニット２００は、利用可能なビット数に収まるように非ゼロバイトのＬＢＳを切り捨てる。他の実施形態では、圧縮ユニット２００は、非ゼロバイトのＭＳＢを切り捨てる。
【００７４】
図３に示されるように、活性化データの非圧縮チャンク２０２の第２のバイトは値１１３（０１１１０００１）を格納する。活性化データの非圧縮チャンク２０２内の最初の４つの非ゼロ値には５ビットが割り当てられているので、この値の３つのＬＳＢが切り捨てられ、その結果、値０１１１０が、活性化データの圧縮されたチャンク２１０内の最初の位置に格納される。活性化データの非圧縮チャンク２０２の第３のバイトは値１２１（０１１１１００１）を格納する。活性化データの非圧縮チャンク２０２内の最初の４つの非ゼロ値には５ビットが割り当てられているので、この値の３つのＬＳＢが切り捨てられ、その結果、値０１１１１が、活性化データの圧縮されたチャンク２１０内の第２の位置に格納される。
【００７５】
図３に示される例では、活性化データの非圧縮チャンク２０２の６２番目のバイトは値２（００００００１０）を格納する。活性化データの非圧縮チャンク２０２内の５番目から６３番目の非ゼロ値には４ビットが割り当てられているので、この値の４つのＬＳＢが切り捨てられ、その結果、値０００００が、活性化データの圧縮されたチャンク２１０内の６２番目の位置に格納される。活性化データの圧縮されたチャンク２１０内の他の非ゼロバイトは、同様にして切り捨てられ、活性化データの圧縮されたチャンク２０４のデータ部分２１０に格納され得る。
【００７６】
圧縮されていない活性化データのチャンク２０２の全ての非ゼロバイトがデータ部分２１０に格納されると、圧縮ユニット２００は、例えば、ＤＮＮモジュール１０５内のオンボード・メモリ又はＤＮＮモジュール１０５のアプリケーションホストのオフボード・メモリに、マスク部分２０８及びデータ部分２１０を含む圧縮された出力チャンク２０４を格納する。圧縮プロセスに関する更なる詳細は、図４に関して以下で与えられる。
【００７７】
図４は、本明細書で開示される一実施形態に従って、圧縮されていない活性化データのチ

ャンク２０２を圧縮するためのＤＮＮモジュール１０５の動作の態様を説明するルーチン４００を示すフロー図である。当然ながら、図４及び他の図に関して本明細書中に記載される論理動作は、（１）コンピュータにより実施される動作又はコンピュータ装置で実行されるプログラムモジュールのシーケンスとして、及び／又は（２）コンピュータ装置内の相互接続された機械論理回路又は回路モジュールとして、実装可能である。

【００７８】

本明細書で開示される技術の特定の実施は、コンピュータ装置の性能及び他の要件に依存して選択できる問題である。従って、本明細書中に記載される論理動作は、状態、操作、構造的デバイス、動作、又はモジュールと様々に呼ばれる。これらの状態、操作、構造的デバイス、動作、及びモジュールは、ハードウェア、ソフトウェア、ファームウェア、特別目的のデジタルロジック、及びそれらの任意の組み合わせで実装され得る。当然ながら、図示及び本明細書中に記載されているよりも多くの又は少ない動作が実行されてよい。これらの動作は、本明細書中に記載されているのとは異なる順序で実行されてもよい。

【００７９】

ルーチン４００は動作４０２から開始する。４０２で、圧縮ユニット２００は、活性化データの圧縮されたチャンク２１０内の非ゼロバイトの数を決定する。ルーチン４００は次いで、動作４０４へ進む。４０４で、圧縮ユニット２００は、活性化データの圧縮されたチャンク２１０内の非ゼロバイトの数が、活性化データの圧縮されたチャンク２０４のデータ部分２１０で利用可能なバイトの数以下であるかどうかを判定する。活性化データの非圧縮チャンク２０２の非ゼロバイトの数が、活性化データの圧縮されたチャンク２０４のデータ部分２１０で利用可能なバイトの数以下である場合には、非ゼロバイトは圧縮される必要がない。従って、この場合に、ルーチン４００は動作４０８へ進む。４０８で、非ゼロバイトは、切り捨てなしでデータ部分２１０に格納される。

【００８０】

活性化データの非圧縮チャンク２０２内の非ゼロバイトの数が、活性化データの圧縮されたチャンク２０４のデータ部分２１０で利用可能なバイトの数よりも多い場合には、ルーチン４００は動作４０６から動作４１２へ進む。動作４１２で、圧縮ユニット２００は、上述されたようにして、活性化データの非圧縮チャンク２０２の切り捨てられた非ゼロバイトを格納するために利用可能な出力データの圧縮されたチャンク２０４のデータ部分２１０のビットの数を決定する。次いで、ルーチン４００は動作４１２から動作４１４へ進む。

【００８１】

動作４１４で、圧縮ユニット２００は、データの非圧縮チャンク２０２の非ゼロバイトを格納するために利用可能である圧縮された出力チャンク２０４のデータ部分２１０内で、もしあれば、追加ビットの数を決定する。上述されたように、圧縮ユニット２００は、それらの追加ビットを、データの非圧縮チャンク２０２内の非ゼロバイトの１つ以上を切り捨てる前に、その１つ以上の非ゼロバイトに割り当てることができる。これは動作４１６で行われる。

【００８２】

動作４１６から、ルーチン４００は動作４１８へ進む。４１８で、圧縮ユニット２００は、活性化データの圧縮されたチャンク２０４のマスク部分２０８内のビットを、活性化データの非圧縮チャンク２０２内の対応するバイトが非ゼロである場合に、論理１にセットする。圧縮ユニット２００はまた、活性化データの圧縮されたチャンク２０４のマスク部分２０８内のビットを、活性化データの非圧縮チャンク２０２内の対応するバイトがゼロである場合に、論理０にセットする。

【００８３】

動作４１８から、ルーチン４００は次いで動作４２０へ進む。４２０で、圧縮ユニット２００は、データの非圧縮チャンク２０２内の非ゼロバイトのＬＳＢ又はＭＳＢを、各非ゼロバイトごとにデータ部分２１０内で利用可能な決定されたビット数まで切り捨てる。切り捨てられた非ゼロバイトは、次いで、活性化データの圧縮されたチャンク２０４のデー

タ部分２１０に格納される。圧縮ユニット２００は次いで、マスク部分２０８及びデータ部分２１０を含む圧縮された出力チャンク２０４を、ＤＮＮモジュール１０５のオンボード・メモリ又はＤＮＮモジュール１０５のアプリケーションホストのオフボード・メモリに格納する。動作４０８及び４２０から、メモリ４００は動作４１０へ進み、終了する。

【００８４】

図５Ａ及び５Ｂは、一実施形態に従って、圧縮された活性化データを復元するためのＤＮＮモジュール１０５の構成及び動作の態様を示すコンピューティングシステムアーキテクチャ図である。先に簡単に説明されたように、且つ、図５Ａ及び５Ｂに示されるように、ＤＮＮモジュール１０５はまた、復元ユニット５００も含むことでき、復元ユニット５００は、上述されたようにして圧縮された活性化データのチャンク２０４を復元することができる。

【００８５】

例えば、復元ユニット５００は、ＤＮＮプロセッサ内のメモリ又はアプリケーションホストのメモリなどのストレージ２０６から活性化データの圧縮されたチャンク２０４を受け取ることができる。復元ユニット５００は、次いで、圧縮されたチャンク２０４のマスク部分２０８内の論理真ビットの数に基づき、データの圧縮されたチャンク２０４のデータ部分２１０内の非ゼロバイトの数を決定することができる。復元ユニット５００はまた、圧縮された出力チャンク２０４のマスク部分２０８内の論理真ビットの位置に基づき、データの復元されたチャンク５０２内の非ゼロバイトの位置を決定することもできる。復元ユニット５００は、同様にして、データの復元されたチャンク５０２内のゼロバイトの位置を決定することができる。

【００８６】

復元ユニット５００はまた、切り捨てられた非ゼロバイトの夫々を圧縮された出力チャンク２０４のデータ部分２１０に格納するために圧縮ユニット２００によって使用されたビットの数を決定することもできる。復元ユニット５００は、データの圧縮されたチャンク２０４内の非ゼロバイトの数（マスク部分２０８によって示される。）と、復元された出力チャンク５０２の目標サイズとに基づき、切り捨てられた各非ゼロバイトを格納するために使用されたビットの数を決定することができる。

【００８７】

上記の例では、例えば、データの圧縮されたチャンク２０４のデータ部分が２４バイト幅（すなわち、１９２ビット）であり、データの非圧縮チャンク２０２内に４７の非ゼロバイトがある場合に、これは、圧縮ユニット２００が、データの非圧縮チャンク２０２の切り捨てられた各非ゼロバイトを格納するためにデータ部分２１０において４ビットを利用したことを意味する。復元ユニット５００はまた、もしあれば、圧縮ユニット２００が圧縮された出力チャンク２０４のデータ部分２１０に格納された切り捨てられた非ゼロバイトのうちの１つ以上に割り当てた追加ビットの数を決定することもできる。

【００８８】

論理０である圧縮された出力チャンク２０４のマスク部分２０８内の各ビット位置について、復元ユニット５００は、ゼロバイトを、復元された出力チャンク５０２の対応する部分に挿入する。論理１であるマスク部分２０８内の各ビット位置について、復元ユニット５００は、圧縮された出力チャンク２０４のデータ部分２１０内の対応する位置からの切り捨てられた非ゼロバイトを、圧縮された出力チャンク２０４の圧縮中に切り捨てられたビットの数に等しい数のゼロビットとともに、復元された出力チャンク５０２の対応する位置に挿入する。ゼロビットは、どのビットが圧縮中に切り捨てられたかに応じて、切り捨てられた非ゼロバイトのＬＳＢ又はＭＳＢに挿入され得る。

【００８９】

上述されたように、いくつかの実施形態において、復元ユニット５００はまた、復元された出力チャンク５０２に格納された切り捨てられた非ゼロバイトのうちの１つ以上にオフセット（例えば、００００００００１）を付加する。例えば、オフセットは、圧縮後にゼロバイトになるデータの非圧縮チャンク２－２の非ゼロバイトに加えられ得る。このように

して、非ゼロバイトは、復元時にゼロバイトにならない。

【００９０】

図６は、一実施形態に従って、圧縮された活性化データのチャンクの例を参照して、活性化データを復元するためのＤＮＮモジュール１０５の動作の態様を説明するデータ構造図である。図６に示される例は、図３に関して上述された例で生成された圧縮された活性化データ２０４の復元について説明する。図６に示されるように、マスク部分２０８は、ビット０、１及び６３に０を格納し、ビット２、３及び６２に１を格納する。データ部分２１０は、図６に示されるように値０１１１０、０１１１１及び００００を格納する。

【００９１】

復元ユニット５００が上記の処理動作を実行する場合に、マスク部分２０８の最初のビット位置にある論理０は、復元ユニット５００に、活性化データの復元されたチャンク５０２の最初のバイトとしてゼロバイトを格納させることになる。同様に、マスク部分２０８の第２のビット位置にある論理０は、復元ユニット５００に、データの復元されたチャンク５０２の第２のバイトとしてゼロバイトを格納させることになる。

【００９２】

マスク部分２０８の第３のビット位置にある論理１は、復元ユニット５００に、データ部分２１０の最初の５ビット（すなわち、０１１１０）を読み出させ且つ３つのＬＳＢを挿入させて、その結果、値０１１１００００（１１２）が活性化データの復元されたチャンク５０２の第３のバイトとして格納されることになる。同様に、マスク部分２０８の第４のビット位置にある論理１は、復元ユニット５００に、データ部分２１０の第２の５ビット（すなわち、０１１１１）を読み出させ且つ３つのＬＳＢを挿入させて、その結果、値０１１１１０００（１２０）が活性化データの復元されたチャンク５０２の第４のバイトとして格納されることになる。

【００９３】

マスク部分２０８の６２番目のビット位置にある論理１は、復元ユニット５００に、データ部分２１０の最後の４つのビット（すなわち、００００）を読み出させ且つ４つのゼロＬＳＢを挿入させ、その結果、０の値が活性化データの復元されたチャンク５０４の６２番目のバイト部分に格納されることになる。マスク部分２０８の最後のビット位置にある論理０は、復元ユニット５００に、データの復元されたチャンク５０２の最後のバイトとしてゼロバイトを格納させることになる。

【００９４】

上述されたように、復元ユニット５００は、活性化データの復元されたチャンク５０２内の特定のバイトにオフセット値を加えることができる。例えば、復元ユニット５００は、活性化データの非圧縮チャンク２０２内で非ゼロであったが、活性化データの圧縮されたチャンク２０４内でゼロバイトに圧縮されたバイトに、０００００００１などのオフセット値を付加することができる。

【００９５】

図６に示される例では、データ部分２１０内の最後のバイトは、活性化データの非圧縮チャンク２０２では非ゼロ（すなわち、２）であったが、活性化データの圧縮されたチャンク２０４ではゼロになった。従って、復元ユニット５００は、このバイトに０００００００１などのオフセット値を付加し、それによって、活性化データの非圧縮チャンク２０２内の非ゼロバイトがゼロバイトに圧縮されないことを確かにすることができる。

【００９６】

図７は、本明細書で開示される一実施形態に従って、活性化データを復元するためのＤＮＮモジュール１０５の動作の態様を説明するルーチン７００を示すフロー図である。ルーチン７００は動作７０２から開始し、７０２で、復元ユニット５００は、活性化データの復元されたチャンク５０２内の非ゼロバイトの数及びそれらの位置を決定するために、活性化データの圧縮されたチャンク２０４のマスク部分２０８を利用する。

【００９７】

ルーチン７００は、動作７０２から動作７０４へ進み、動作７０４で、復元ユニット５０

０は、活性化データの圧縮されたチャンク内の非ゼロバイトの数が活性化データの復元されたチャンク５０２のバイトの数以下であるかどうかを判定する。上述されたように、活性化データの圧縮されたチャンク２０４の非ゼロバイトは、その非ゼロバイトの数が活性化データの復元されたチャンク５０２のバイトの数以下である場合には、復元される必要がない。従って、この場合に、ルーチン７００は動作７０８へ進み、７０８で、活性化データの圧縮されたチャンク２０４内の非ゼロバイトは、変更なしで、活性化データの復元されたチャンク５０２に格納される。

【００９８】

活性化データの圧縮されたチャンク内の非ゼロバイトの数が活性化データの復元されたチャンク５０２内のバイトの数よりも多い場合には、ルーチン７００は、動作７０６から動作７１２へ進む。動作７１２で、復元ユニット５００は、活性化データの非圧縮チャンク２０２の切り捨てられた各非ゼロバイトを格納するために圧縮ユニット２００が使用した出力データの圧縮されたチャンク２０４のデータ部分２１０のビットの数を決定する。ルーチン７００は次いで、上述されたように動作７１２から動作７１４へ進む。

【００９９】

動作７１４で、復元ユニット５００は、もしあれば、データの非圧縮チャンク２０２の非ゼロバイトを格納するために使用された追加ビットの数を決定する。復元ユニット５００は、上述されたように、それらの追加ビットを、データの復元されたチャンク５０２内の非ゼロバイトのうちの１つ以上に割り当てることができる。

【０１００】

動作７１６から、ルーチン７００は動作７１８へ進み、動作７１８で、復元ユニット５００は、論理０である圧縮された出力チャンク２０４のマスク部分２０８内の各ビット位置について、ゼロバイトを、復元された出力チャンク５０２の対応する位置に挿入する。論理１である圧縮された出力チャンク２０４のマスク部分２０８内の各ビット位置について、復元ユニット５００は、圧縮された出力チャンク２０４の対応する位置からの切り捨てられた非ゼロバイトを、圧縮された出力チャンク２０４の圧縮中に切り捨てられたビットの数に等しい数のゼロビットとともに、復元された出力チャンク５０２の対応する位置に挿入する。ゼロビットは、どのビットが圧縮中に切り捨てられたかに応じて、切り捨てられた非ゼロバイトのＬＳＢ又はＭＳＢに挿入され得る。これは、動作７２０で行われる。

【０１０１】

いくつかの実施形態において、復元ユニット５００はまた、復元された出力チャンク５０２に格納された切り捨てられた非ゼロバイトのうちの１つ以上にオフセット値を付加することができる。例えば、オフセットは、圧縮後にゼロバイトになるデータの非圧縮チャンク２０２の非ゼロバイトに加えられ得る。このようにして、非ゼロバイトは、圧縮及び復元される場合にゼロバイトにならない。他の実施形態では、オフセットは、活性化データの復元されたチャンク５０２内の全てのバイトに付加され得る。

【０１０２】

復元ユニット５００は次いで、復元された出力チャンク５０２を、ニューロン１０５Ｆによる使用のために、ＤＮＮモジュール１０５のオンボード・メモリ又はＤＮＮモジュール１０５のアプリケーションホストのオフボード・メモリに格納する。動作７０８及び７２０から、ルーチン７００は動作７１０へ進み、終了する。

【０１０３】

図８は、本明細書で提示されるＤＮＮモジュール１０５のためのアプリケーションホストとして動作することができるコンピュータ装置のためのコンピュータハードウェア及びソフトウェアアーキテクチャを例示するコンピュータアーキテクチャ図である。特に、図８に表されているアーキテクチャは、サーバーコンピュータ、携帯電話機、電子書籍リーダー、スマートフォン、デスクトップコンピュータ、ＡＲ／ＶＲデバイス、タブレットコンピュータ、ラップトップコンピュータ、又はＤＮＮモジュール１０５との使用に適した他のタイプのコンピュータ装置を実装するために利用され得る。

【０１０４】

図８に表されているコンピュータ８００は、中央演算処理装置８０２（“ＣＰＵ”）と、ランダムアクセスメモリ（“ＲＡＭ”）及びリードオンリーメモリ（“ＲＯＭ”）を含むシステムメモリ８０４と、メモリ８０４をＣＰＵ８０２へ結合するシステムバス８１０とを含む。起動中などのように、コンピュータ８００内の要素間で情報を伝送するのを助ける基本ルーチンを含む基本入出力システム（“ＢＩＯＳ”又は“ファームウェア”）が、ＲＯＭ８０８に記憶され得る。コンピュータ８００は、オペレーティングシステム８２２、アプリケーションプログラム、及び他のタイプのプログラムを記憶する大容量記憶装置８１２を更に含む。大容量記憶装置８１２はまた、他のタイプのプログラム及びデータも記憶するよう構成され得る。

【０１０５】

大容量記憶装置８１２は、バス８１０へ接続されている大容量記憶コントローラ（図示せず。）を通じてＣＰＵ８０２へ接続されている。大容量記憶装置８１２及びその関連するコンピュータ可読媒体は、コンピュータ８００のための不揮発性記憶を提供する。本明細書で含まれているコンピュータ可読媒体についての記載は、ハードディスク、ＣＤ－ＲＯＭドライブ、又はＵＳＢストレージキーなどの大容量記憶装置を指すが、当業者には当然ながら、コンピュータ可読媒体は、コンピュータ８００によってアクセス可能な如何なる利用可能なコンピュータ記憶媒体又は通信媒体でもあることができる。

【０１０６】

通信媒体は、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、又は搬送波若しくは他の伝送メカニズムなどの変調データ信号内の他のデータを含み、如何なる配信媒体も含む。語「変調データ信号」は、信号内の情報を符号化するように変更又は設定されたその特徴の１つ以上が変更又は設定されている信号を意味する。例として、制限なしに、通信媒体は、有線ネットワーク又は直接配線接続などの有線媒体と、音響、無線周波数、赤外線及び他の無線媒体等の無線媒体とを含む。上記のうちのいずれかの組み合わせも、コンピュータ可読媒体の範囲内に含まれるべきである。

【０１０７】

例として、制限なしに、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、又は他のデータなどの情報の記憶のための如何なる方法及び技術でも実装される揮発性及び不揮発性のリムーバブル及び非リムーバブル媒体を含むことができる。例えば、コンピュータ記憶媒体は、ＲＡＭ、ＲＯＭ、ＥＰＲＯＭ、ＥＥＰＲＯＭ、フラッシュメモリ若しくは他のソリッドステートメモリ技術、ＣＤ－ＲＯＭ、デジタルバーサタイルディスク（“ＤＶＤ”）、ＨＤ－ＤＶＤ、ＢＬＵＥ－ＲＡＹ、若しくは他の光学ストレージ、磁気カセット、磁気テープ、磁気ディスクストレージ若しくは他の磁気記憶デバイス、又は所望の情報を記憶するために使用可能であって、コンピュータ８００によってアクセス可能であるあらゆる他の媒体を含むが、これらに限られない。特許請求の範囲のために、語句「コンピュータ記憶媒体」及びその変形は、波若しくは信号自体又は通信媒体を含まない。

【０１０８】

様々な構成によれば、コンピュータ８００は、ネットワーク８２０などのネットワークを通じた遠隔のコンピュータへの論理接続を使用するネットワーク化された環境で動作することができる。コンピュータ８００は、バス８１０へ接続されたネットワークインターフェイスユニット８１６を通じてネットワーク８２０へ接続することができる。当然ながら、ネットワークインターフェイスユニット８１６は、他のタイプのネットワーク及び遠隔コンピュータシステムへ接続するためにも利用可能である。コンピュータ８００はまた、キーボード、マウス、タッチ入力、電子スタイラス（図８に図示せず。）、又はビデオカメラなどの物理センサを含む多数の他のデバイスから入力を受け取って処理する入出力（“Ｉ／Ｏ”）コントローラ８１８も含むことができる。同様に、入出力コントローラ８１８は、表示スクリーン又は他のタイプの出力デバイス（図８に図示せず。）へ出力を供給することができる。

【０１０９】

当然ながら、本明細書で記載されるソフトウェアコンポーネントは、ＣＰＵ８０２にロードされ実行される場合に、ＣＰＵ８０２及びコンピュータ８００の全体を、汎用のコンピュータ装置から、本明細書で提示される機能を容易にするようカスタマイズされた特別目的のコンピュータ装置へ変えることができる。ＣＰＵ８０２は、個々に又は集合的に任意数の状態を前提とすることができる任意数のトランジスタ又は他のディスクリート回路素子から構成され得る。より具体的には、ＣＰＵ８０２は、本明細書で開示されるソフトウェアモジュール内に含まれる実行可能命令に応答して有限状態機械として動作することができる。これらのコンピュータ実行可能命令は、如何にしてＣＰＵ８０２が状態間を遷移するかを特定し、それによってＣＰＵ８０２を構成するトランジスタ又は他のディスクリートハードウェア素子を変化させることによって、ＣＰＵ８０２を変化させることができる。

【０１１０】

本明細書で提示されるソフトウェアモジュールを符号化することは、本明細書で提示されるコンピュータ可読媒体の物理構造も変えることができる。物理構造の具体的な変形は、本明細書の種々の実施において、様々な因子に依存する。かような因子の例には、コンピュータ可読媒体を実装するために使用される技術、コンピュータ可読媒体が一次又は二次記憶と見なされるかどうか、などがあるが、これらに限られない。例えば、コンピュータ可読媒体が半導体に基づくメモリとして実装される場合には、本明細書で開示されるソフトウェアは、半導体メモリの物理状態を変えることによってコンピュータ可読媒体上で符号化され得る。例えば、ソフトウェアは、半導体メモリを構成するトランジスタ、キャパシタ、又は他のディスクリート回路素子の状態を変えることができる。ソフトウェアはまた、かような構成要素の物理状態を、それにデータを格納するために変えることもできる。

【０１１１】

他の例として、本明細書で開示されるコンピュータ可読媒体は、磁気又は光学技術を用いて実装可能である。かような実施において、本明細書で提示されるソフトウェアは、ソフトウェアが磁気又は光学媒体で符号化されるときに、磁気又は光学媒体の物理状態を変えることができる。これらの変形は、所与の磁気媒体内の特定の位置の磁気特性を変更することを含むことができる。これらの変形は、所与の光学媒体内の特定の位置の物理的な特徴又は特性を変更してそれらの位置の光学特性を変えることを含むこともできる。物理媒体の他の変形は、本明細書の適用範囲及び精神から外れることなしに可能であり、上記の例は、本議論を助けるためにのみ与えられている。

【０１１２】

上記を鑑みて、当然ながら、本明細書で提示されるソフトウェアコンポーネントを記憶及び実行するために、多数のタイプの物理変形がコンピュータ８００において行われる。また、当然ながら、コンピュータ８００に関して図８に示されるアーキテクチャ、又は同様のアーキテクチャは、手持ち式コンピュータ、ビデオゲーム機、埋込型コンピュータシステム、スマートフォン、タブレット、及びＡＲ／ＶＲ装置などのモバイル機器、並びに当業者に知られている他のタイプのコンピュータ装置を含む他のタイプのコンピュータ装置を実装するために利用可能である。また、コンピュータ８００は、図８に示されている全てのコンポーネントを含まなくてもよく、図８に明示的に示されていない他のコンポーネントを含むことができ、あるいは、図８に示されているのとは全く異なるアーキテクチャを利用することができる。

【０１１３】

図９は、本明細書で提示される様々な実施形態に従って、開示される技術の態様が実装され得る分散ネットワークコンピューティング環境９００帯域幅を説明するネットワーク図である。図９に示されるように、１つ以上のサーバーコンピュータ９００Ａは、通信ネットワーク８２０（固定配線若しくは無線ＬＡＮ、ＷＡＮ、イントラネット、エクストラネット、ピア・ツー・ピアネットワーク、仮想プライベートネットワーク、インターネット、Ｂｌｕｅｔｏｏｔｈ（登録商標）通信網、固有低電圧通信網、又は他の通信ネットワークのいずれか１つ又は組み合わせであってよい。）を介して、例えば、制限なしに、タブ

レットコンピュータ９００Ｂ、ゲーム機９００Ｃ、スマートウォッチ９００Ｄ、スマートフォンなどの電話機９００Ｅ、パーソナルコンピュータ９００Ｆ、及びＡＲ／ＶＲデバイス９００Ｇなどの多数のクライアントコンピュータ装置と相互接続され得る。

【０１１４】

例えば、通信ネットワーク８２０がインターネットであるネットワーク環境で、サーバーコンピュータ９００Ａは、ハイパーテキスト転送プロトコル（Hypertext Transfer Protocol）（“ＨＴＴＰ”）、ファイル転送プロトコル（File Transfer Protocol）（“ＦＴＰ”）、又はシンプル・オブジェクト・アクセス・プロトコル（Simple Object Access Protocol）（“ＳＯＡＰ”）などの多数の既知のプロトコルのいずれかを介して、クライアントコンピュータ装置９００Ｂ～９００Ｇへのデータ及びそれらからのデータを処理及び通信するよう動作可能な専用サーバーコンピュータであることができる。更に、ネットワーク化されたコンピューティング環境９００は、セキュアド・ソケット・レイヤ（Secured Socket Layer）（“ＳＳＬ”）又はプリティ・グッド・プライバシー（Pretty Good Privacy）（“ＰＧＰ”）などの様々なデータセキュリティプロトコルを利用することができる。クライアントコンピュータ装置９００Ｂ～９００Ｇの夫々は、サーバーコンピュータ９００Ａへのアクセスを得るようにウェブブラウザ（図９に図示せず。）などの１つ以上のコンピューティングアプリケーション若しくは端末セッション、又は他のグラフィカルユーザインターフェイス（図９に図示せず。）、又はモバイルデスクトップ環境（図９に図示せず。）をサポートするよう動作可能なオペレーティングシステムを装備され得る。

【０１１５】

サーバーコンピュータ９００Ａは、他のコンピューティング環境（図９に図示せず。）へ通信上結合され、参加ユーザのインタラクション／リソースネットワークに関するデータを受信することができる。実例となる動作では、ユーザ（図９に図示せず。）は、所望のデータを得るよう及び／又は他のコンピューティングアプリケーションを実行するようクライアントコンピュータ装置９００Ｂ～９００Ｇで実行されるコンピューティングアプリケーションと相互作用してもよい。

【０１１６】

データ及び／又はコンピューティングアプリケーションは、サーバ９００Ａ又は複数のサーバ９００Ａに記憶され、実例となる通信ネットワーク８２０上でクライアントコンピュータ装置９００Ｂ～９００Ｇを通じて協調するユーザへ送られ得る。参加ユーザ（図９に図示せず。）は、完全に又は部分的にサーバーコンピュータ９００Ａに格納された特定のデータ及びアプリケーションへのアクセスを要求してもよい。これらのデータは、処理及び記憶のためにクライアントコンピュータ装置９００Ｂ～９００Ｇとサーバーコンピュータ９００Ａとの間で通信されてもよい。

【０１１７】

サーバーコンピュータ９００Ａは、データ及びアプリケーションの生成、認証、暗号化、及び通信のためのコンピューティングアプリケーション、プロセス及びアプレットをホストすることができ、他のサーバーコンピューティング環境（図９に図示せず。）、第三者サービスプロバイダ（図９に図示せず。）、ネットワーク・アタッチト・ストレージ（Network Attached Storage）（“ＮＡＳ”）及びストレージ・エリア・ネットワーク（Storage Area Network(s)）（“ＳＡＮ”）と協調してアプリケーション／データトランザクションを実現し得る。

【０１１８】

当然ながら、図８に示されているコンピューティングアーキテクチャ及び図９に示されている専用ネットワークコンピューティング環境は、議論を簡単にするために簡略化されている。また、当然ながら、コンピューティングアーキテクチャ及び専用コンピューティングネットワークは、本明細書で具体的に記載されていない多くの更なるコンピューティングコンポーネント、デバイス、ソフトウェアプログラム、ネットワーキングデバイス、及び他のコンポーネントを含み、利用することができる。

【０１１９】
本明細書で提示される開示は、以下の付記に記載されている対象も包含する。
【０１２０】
付記１．ニューラルネットワークプロセッサであって、１つ以上のニューロンと、圧縮ユニットとを有し、前記圧縮ユニットは、
当該ニューラルネットワークプロセッサ内の前記ニューロンの少なくとも１つによって生成されるデータの非圧縮チャンクを受け取り、該データの非圧縮チャンクが一定数のバイトを含み；
圧縮された出力チャンクのマスク部分を生成し、該マスク部分が、前記データの非圧縮チャンク内の前記一定数のバイトに等しいビットの数を含み、前記マスク部分内の各ビットが、前記データの非圧縮チャンク内のバイトに対応し、前記マスク部分内の各ビットが、前記データの非圧縮チャンク内の対応するバイトがゼロである場合に論理０に設定され、前記データの非圧縮チャンク内の対応するバイトが非ゼロである場合に論理１に設定され；
前記データの非圧縮チャンク内の非ゼロバイトの数を決定することと、前記データの非圧縮チャンク内の前記非ゼロバイトの数に基づき、前記データの非圧縮チャンクの切り捨てられた非ゼロバイトを格納するために利用可能な前記圧縮された出力チャンクのデータ部分内のビットの数を決定することと、該決定されたビットの数まで前記データの非圧縮チャンク内の前記非ゼロバイトを切り捨てることと、該切り捨てられた非ゼロバイトを前記圧縮された出力チャンクの前記データ部分に格納することとによって、前記圧縮された出力チャンクの前記データ部分を生成し；
前記マスク部分及び前記データ部分を含む前記圧縮された出力チャンクを出力する
よう構成される、ニューラルネットワークプロセッサ。
【０１２１】
付記２．当該ニューラルネットワークプロセッサは、復元ユニットを更に有し、該復元ユニットは、
前記圧縮された出力チャンクを受け取り；
前記圧縮された出力チャンクの前記マスク部分に基づき、前記データの非圧縮チャンク内の非ゼロバイトの数を決定し；
前記圧縮された出力チャンクの前記マスク部分に基づき、前記データの非圧縮チャンク内の前記非ゼロバイトの位置を決定し；
前記圧縮された出力チャンクの前記データ部分に前記切り捨てられた非ゼロバイトを格納するために前記圧縮ユニットによって使用されたビットの数を決定し；
論理０である前記圧縮された出力チャンクの前記マスク部分内の各ビット位置について、ゼロバイトを、復元された出力チャンクの対応する位置に挿入し；
論理１である前記マスク部分内の各ビット位置について、前記圧縮された出力チャンクの前記データ部分内の対応する位置からの前記切り捨てられた非ゼロバイトを、前記復元された出力チャンクの対応する位置に、前記圧縮された出力チャンクの圧縮中に切り捨てられたビットの数に等しい数のゼロビットとともに挿入する
よう構成される、付記１に記載のニューラルネットワークプロセッサ。
【０１２２】
付記３．前記圧縮ユニットは更に、
前記データの非圧縮チャンクの切り捨てられた非ゼロバイトを格納するために利用可能な前記圧縮された出力チャンクの前記データ部分内の追加ビットの数を決定し；
前記データの非圧縮チャンク内の前記非ゼロバイトのうちの１つ以上の非ゼロバイトを切り捨てる前に、該１つ以上の非ゼロバイトに前記追加ビットを割り当てる
よう構成される、付記１及び２のいずれかに記載のニューラルネットワークプロセッサ。
【０１２３】
付記４．前記復元ユニットは更に、前記圧縮された出力チャンクの前記データ部分内に格納された前記１つ以上の非ゼロバイトに割り当てられている前記追加ビットの数を決定するよう構成される、付記１乃至３のうちいずれかに記載のニューラルネットワークプロセ

ッサ。
【０１２４】
付記５．前記復元ユニットは更に、前記復元された出力チャンクに格納された前記切り捨てられた非ゼロバイトのうちの１つ以上の非ゼロバイトにオフセットを加えるよう構成される、付記１乃至４のうちいずれかに記載のニューラルネットワークプロセッサ。
【０１２５】
付記６．前記非ゼロバイトの１つ以上の最下位ビット（ＬＳＢ）が切り捨てられる、付記１乃至５のうちいずれかに記載のニューラルネットワークプロセッサ。
【０１２６】
付記７．前記非ゼロバイトの１つ以上の最上位ビット（ＭＳＢ）が切り捨てられる、付記１乃至６のうちいずれかに記載のニューラルネットワークプロセッサ。
【０１２７】
付記８．ニューラルネットワークプロセッサであって、１つ以上のニューロンと、復元ユニットとを有し、前記復元ユニットは、
マスク部分及びデータ部分を含むデータの圧縮されたチャンクを受け取り；
前記マスク部分内のビットに基づき、データの復元されたチャンク内の非ゼロバイトの数を決定し；
前記非ゼロバイトの数に少なくとも部分的に基づき、前記データの圧縮されたチャンクの前記データ部分に切り捨てられた非ゼロバイトを格納するために使用されたビットの数を決定し；
論理０である前記データの圧縮されたチャンクの前記マスク部分内の各ビット位置について、ゼロバイトを、前記データの復元されたチャンクの対応する部分に挿入し；
論理１である前記データの圧縮されたチャンクの前記マスク部分内の各ビット位置について、前記データの圧縮されたチャンクの前記データ部分内の対応する位置からの前記切り捨てられた非ゼロバイトを、前記データの復元されたチャンクの対応する位置に、前記データの圧縮されたチャンクの圧縮中に切り捨てられたビットの数に等しい数のゼロビットとともに挿入する
よう構成される、ニューラルネットワークプロセッサ。
【０１２８】
付記９．圧縮ユニットを更に有し、該圧縮ユニットは、
当該ニューラルネットワークプロセッサ内の前記ニューロンの少なくとも１つによって生成されるデータの非圧縮チャンクを受け取り、該データの非圧縮チャンクが一定数のバイトを含み；
前記データの圧縮されたチャンクの前記マスク部分を生成し、該マスク部分が、前記データの非圧縮チャンク内の前記一定数のバイトに等しいビットの数を含み、前記マスク部分内の各ビットが、前記データの非圧縮チャンク内のバイトに対応し、前記マスク部分内の各ビットが、前記データの非圧縮チャンク内の対応するバイトがゼロである場合に論理０を有し、前記データの非圧縮チャンク内の対応するバイトが非ゼロである場合に論理１を有し；
前記データの非圧縮チャンク内の非ゼロバイトの数を決定することと、前記データの非圧縮チャンク内の前記非ゼロバイトの数に基づき、前記データの非圧縮チャンクの切り捨てられた非ゼロバイトを格納するために利用可能な前記データの圧縮されたチャンクの前記データ部分内のビットの数を決定することと、該決定されたビットの数まで前記データの非圧縮チャンク内の前記非ゼロバイトを切り捨てることと、該切り捨てられた非ゼロバイトを前記データの圧縮されたチャンクの前記データ部分に格納することとによって、前記データの圧縮されたチャンクの前記データ部分を生成し；
前記マスク部分及び前記データ部分を含む前記データの圧縮されたチャンクを出力する
よう構成される、付記８に記載のニューラルネットワークプロセッサ。
【０１２９】
付記１０．前記圧縮ユニットは更に、前記データの非圧縮チャンク内の非ゼロバイトの数

が前記データの圧縮されたチャンクの前記データ部分内のバイトの数以下である場合には切り捨てなしで、前記データの非圧縮チャンク内の前記非ゼロバイトを前記データの圧縮されたチャンクの前記データ部分に格納するよう構成される、請求項８及び９のいずれかに記載のニューラルネットワークプロセッサ。

【０１３０】

付記１１．前記圧縮ユニットは更に、

前記データの非圧縮チャンクの切り捨てられた非ゼロバイトを格納するために利用可能な前記データの圧縮されたチャンクの前記データ部分内の追加ビットの数を決定し；

前記データの非圧縮チャンク内の前記非ゼロバイトのうちの１つ以上の非ゼロバイトを切り捨てる前に、該１つ以上の非ゼロバイトに前記追加ビットを割り当てる

よう構成される、付記８乃至１０のうちいずれかに記載のニューラルネットワークプロセッサ。

【０１３１】

付記１２．前記復元ユニットは更に、前記データの圧縮されたチャンクの前記データ部分内に格納された前記１つ以上の非ゼロバイトに割り当てられている前記追加ビットの数を決定するよう構成される、付記８乃至１１のうちいずれかに記載のニューラルネットワークプロセッサ。

【０１３２】

付記１３．前記非ゼロバイトの１つ以上の最下位ビット（ＬＳＢ）が、前記データの圧縮されたチャンクの圧縮中に切り捨てられる、付記８乃至１２のうちいずれかに記載のニューラルネットワークプロセッサ。

【０１３３】

付記１４．前記非ゼロバイトの１つ以上の最上位ビット（ＭＳＢ）が、前記データの圧縮されたチャンクの圧縮中に切り捨てられる、付記８乃至１３のうちいずれかに記載のニューラルネットワークプロセッサ。

【０１３４】

付記１５．コンピュータにより実施される方法であって、

ニューラルネットワークプロセッサの圧縮ユニットで、前記ニューラルネットワークプロセッサ内の少なくとも１つのニューロンによって生成されるデータの非圧縮チャンクを受け取り、該データの非圧縮チャンクが一定数のバイトを含む、ことと、

圧縮された出力チャンクのマスク部分を生成し、該マスク部分が、前記データの非圧縮チャンク内の前記一定数のバイトに等しいビットの数を含み、前記マスク部分内の各ビットが、前記データの非圧縮チャンク内のバイトに対応し、前記マスク部分内の各ビットが、前記データの非圧縮チャンク内の対応するバイトがゼロである場合に論理０を有し、前記データの非圧縮チャンク内の対応するバイトが非ゼロである場合に論理１を有する、ことと、

前記データの非圧縮チャンク内の非ゼロバイトの数を決定すること、前記データの非圧縮チャンク内の前記非ゼロバイトの数に基づき、前記データの非圧縮チャンクの切り捨てられた非ゼロバイトを格納するために利用可能な前記圧縮された出力チャンクのデータ部分内のビットの数を決定することと、該決定されたビットの数まで前記データの非圧縮チャンク内の前記非ゼロバイトを切り捨てることと、該切り捨てられた非ゼロバイトを前記圧縮された出力チャンクの前記データ部分に格納することとによって、前記圧縮された出力チャンクの前記データ部分を生成することと、

前記マスク部分及び前記データ部分を含む前記圧縮された出力チャンクを前記ニューラルネットワークプロセッサのメモリに記憶することと

を有する、方法。

【０１３５】

付記１６．前記データの非圧縮チャンクの切り捨てられた非ゼロバイトを格納するために利用可能な前記圧縮された出力チャンクの前記データ部分内の追加ビットの数を決定することと、前記データの非圧縮チャンク内の前記非ゼロバイトのうちの１つ以上の非ゼロバ

イトを切り捨てる前に、該１つ以上の非ゼロバイトに前記追加ビットを割り当てることとを更に有する、付記１５に記載のコンピュータにより実施される方法。

【０１３６】

付記１７．前記データの非圧縮チャンク内の非ゼロバイトの数が前記圧縮された出力チャンクの前記データ部分内のバイトの数以下である場合には切り捨てなしで、前記データの非圧縮チャンク内の前記非ゼロバイトを前記圧縮された出力チャンクの前記データ部分に格納することを更に有する、付記１５及び１６のうちいずれかに記載のコンピュータにより実施される方法。

【０１３７】

付記１８．前記ニューラルネットワークプロセッサの復元ユニットで、前記圧縮された出力チャンクを受け取ることと、

前記圧縮された出力チャンクの前記マスク部分に基づき、前記データの非圧縮チャンク内の非ゼロバイトの数を決定することと、

前記圧縮された出力チャンクの前記マスク部分に基づき、前記データの非圧縮チャンク内の前記非ゼロバイトの位置を決定することと、

前記圧縮された出力チャンクの前記データ部分に前記切り捨てられた非ゼロバイトを格納するために前記圧縮ユニットによって使用されたビットの数を決定することと、

論理０である前記圧縮された出力チャンクの前記マスク部分内の各ビット位置について、ゼロバイトを、復元された出力チャンクの対応する位置に挿入することと、

論理１である前記マスク部分内の各ビット位置について、前記圧縮された出力チャンクの前記データ部分内の対応する位置からの前記切り捨てられた非ゼロバイトを、前記復元された出力チャンクの対応する位置に、前記圧縮された出力チャンクの圧縮中に切り捨てられたビットの数に等しい数のゼロビットとともに挿入することと

を更に有する、付記１５乃至１７のうちいずれかに記載のコンピュータにより実施される方法。

【０１３８】

付記１９．前記復元された出力チャンクに格納される前記切り捨てられた非ゼロバイトのうちの１つ以上にオフセットを付加することを更に有する、付記１５乃至１８のうちいずれかに記載のコンピュータにより実施される方法。

【０１３９】

付記２０．前記オフセットは、前記復元された出力チャンクに格納される前記切り捨てられた非ゼロバイトの１つ以上の最下位ビット（ＬＳＢ）に加えられる、付記１５乃至１９のうちいずれかに記載のコンピュータにより実施される方法。

【０１４０】

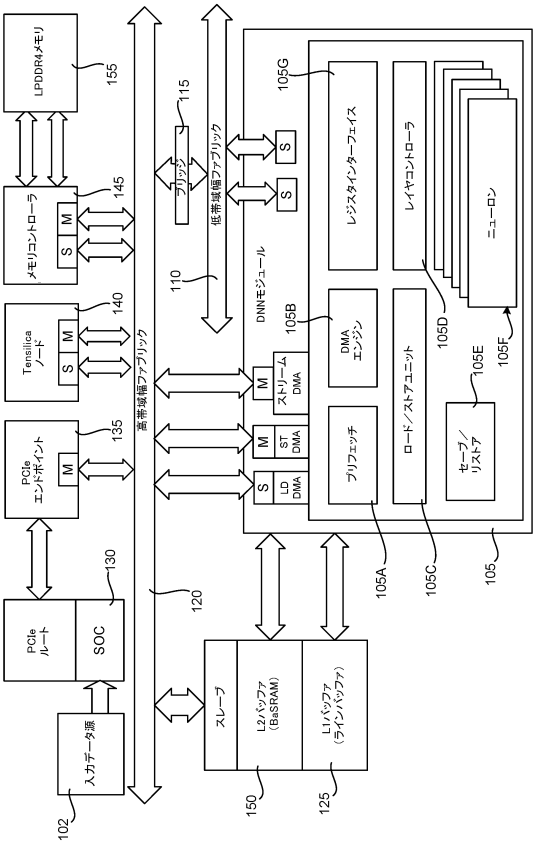
上記に基づき、メモリバス帯域幅の利用量を低減するために活性化データを圧縮及び復元することができるＤＮＮモジュールが本明細書で開示されてきた、ことが認識されるべきである。本明細書で提示される対象は、コンピュータ構造特徴、方法論的及び変形可能な動作、具体的な計算器、並びにコンピュータ可読媒体に特有の言語で記載されてきたが、添付の特許請求の範囲に示されている対象は、必ずしも、本明細書で記載される具体的な特徴、動作、又は媒体に限られないことが理解されるべきである。むしろ、具体的な特徴、動作及び媒体は、請求される対象を実施する形態の例として開示されている。

【０１４１】

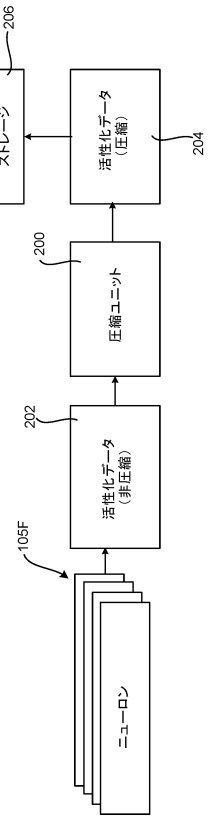
上記の対象は、実例としてのみ与えられており、限定として解釈されるべきではない。様々な修正及び変更が、図示及び記載されている例となる構成及び適用に従うことなく、且つ、続く特許請求の範囲で示されている本開示の適用範囲から外れることなしに、本明細書で記載されている対象に対して行われ得る。

【図面】

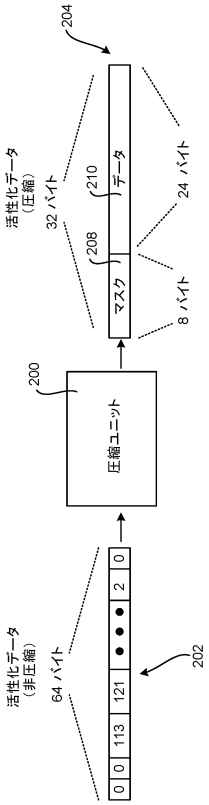
【図１】



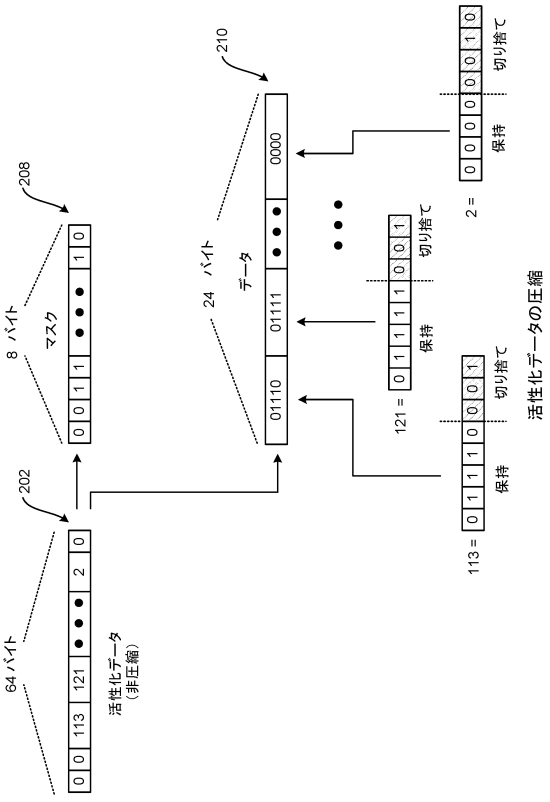
【図２Ａ】



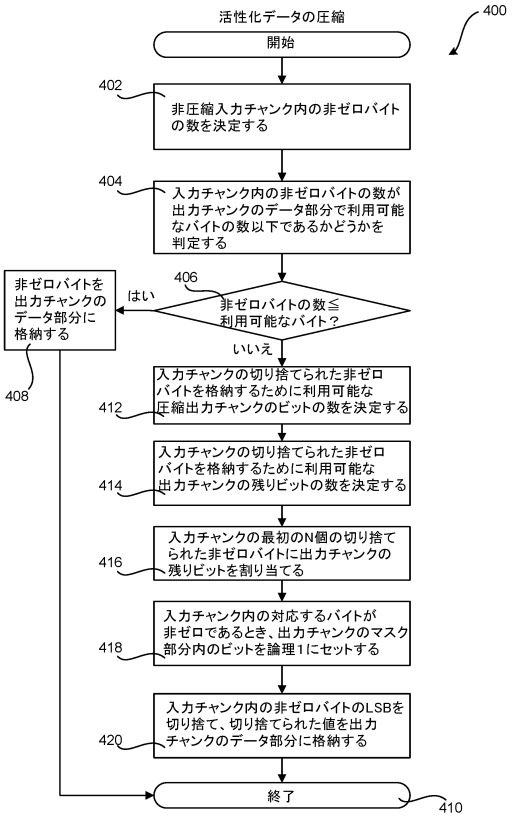
【図２Ｂ】



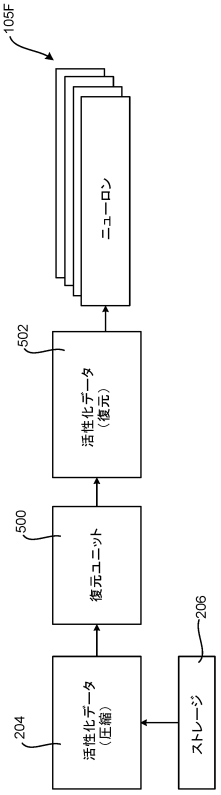
【図３】



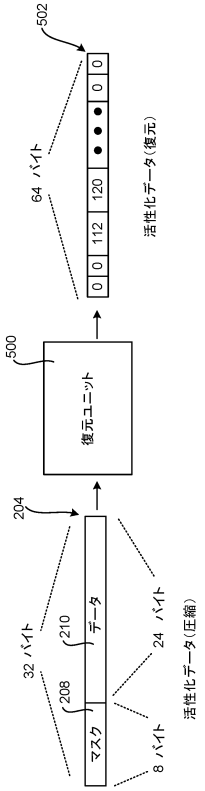
【図４】


活性化データの圧縮
400
開始
402
非圧縮入力チャンク内の非ゼロバイトの数を決定する
404
入力チャンク内の非ゼロバイトの数が出力チャンクのデータ部分で利用可能なバイトの数以下であるかどうかを判定する
406
非ゼロバイトの数≦利用可能なバイト？
はい
いいえ
非ゼロバイトを出力チャンクのデータ部分に格納する
408
入力チャンクの切り捨てられた非ゼロバイトを格納するために利用可能な圧縮出力チャンクのビットの数を決定する
412
入力チャンクの切り捨てられた非ゼロバイトを格納するために利用可能な出力チャンクの残りビットの数を決定する
414
入力チャンクの最初のN個の切り捨てられた非ゼロバイトに出力チャンクの残りビットを割り当てる
416
入力チャンク内の対応するバイトが非ゼロであるとき、出力チャンクのマスク部分内のビットを論理1にセットする
418
入力チャンク内の非ゼロバイトのLSBを切り捨て、切り捨てられた値を出力チャンクのデータ部分に格納する
420
終了
410

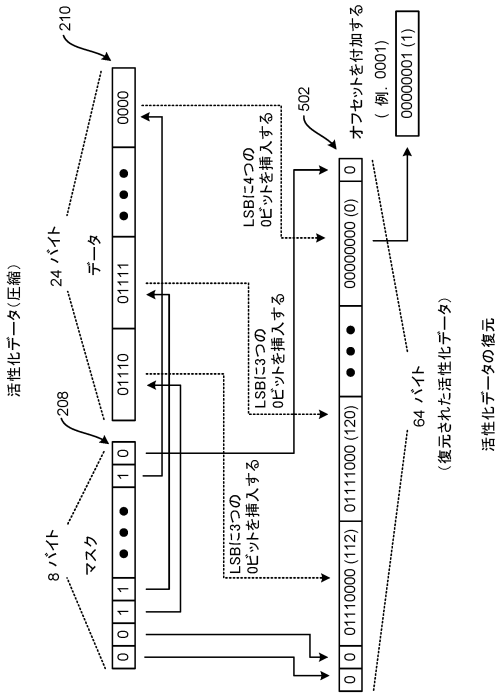
【図５Ａ】


105F
ニューロン
502
活性化データ（復元）
500
復元ユニット
206
ストレージ
204
活性化データ（圧縮）

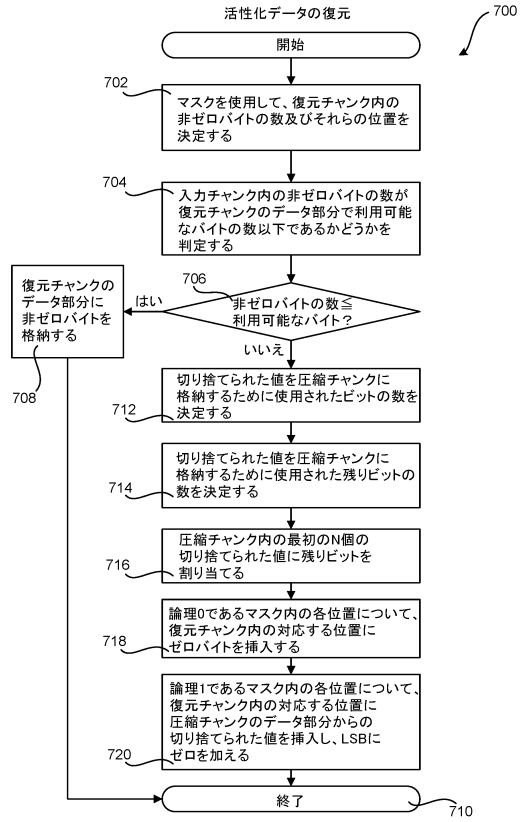
【図５Ｂ】


502
64 バイト
0 0 112 120 ● ● ● 0 0
活性化データ（復元）
500
復元ユニット
204
32 バイト
210
データ
208
マスク
24 バイト
8 バイト
活性化データ（圧縮）

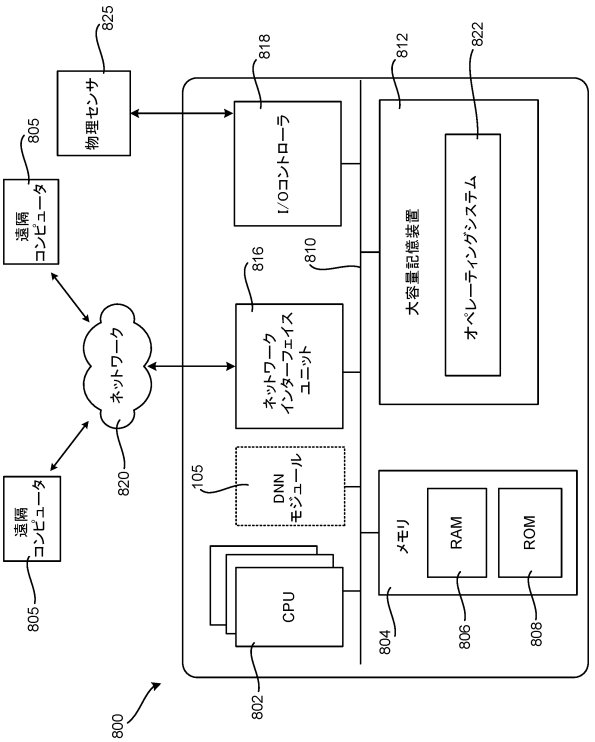
【図６】


210
活性化データ（圧縮）
24 バイト
データ
01110 01111 ● ● ● 0000
208
8 バイト
マスク
0 0 1 1 ● ● ● 1 0
LSBに3つの0ビットを挿入する
LSBに3つの0ビットを挿入する
LSBに4つの0ビットを挿入する
502
オフセットを付加する
（例. 0001）
00000001 (1)
0 0 01110000 (112) 01111000 (120) ● ● ● 00000000 (0) 0
64 バイト
（復元された活性化データ）
活性化データの復元

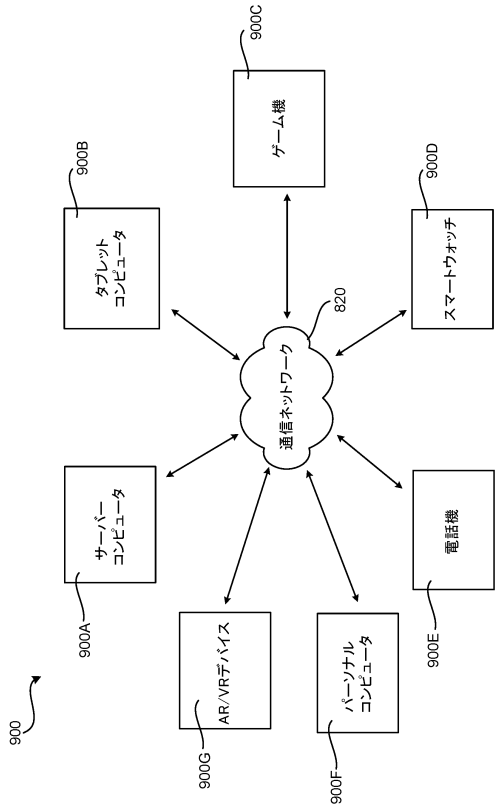
【図７】


活性化データの復元
700
開始
702
マスクを使用して、復元チャンク内の非ゼロバイトの数及びそれらの位置を決定する
704
入力チャンク内の非ゼロバイトの数が復元チャンクのデータ部分で利用可能なバイトの数以下であるかどうかを判定する
706
非ゼロバイトの数≦利用可能なバイト？
はい
いいえ
復元チャンクのデータ部分に非ゼロバイトを格納する
708
712
切り捨てられた値を圧縮チャンクに格納するために使用されたビットの数を決定する
714
切り捨てられた値を圧縮チャンクに格納するために使用された残りビットの数を決定する
716
圧縮チャンク内の最初のN個の切り捨てられた値に残りビットを割り当てる
718
論理0であるマスク内の各位置について、復元チャンク内の対応する位置にゼロバイトを挿入する
720
論理1であるマスク内の各位置について、復元チャンク内の対応する位置に圧縮チャンクのデータ部分からの切り捨てられた値を挿入し、LSBにゼロを加える
終了
710

【図８】


825
物理センサ
818
I/Oコントローラ
812
822
大容量記憶装置
オペレーティングシステム
805
遠隔コンピュータ
816
ネットワークインターフェイスユニット
810
ネットワーク
820
105
DNNモジュール
メモリ
RAM
ROM
CPU
804
806
808
802
800

【図９】


900C
ゲーム機
900B
タブレットコンピュータ
900D
スマートウォッチ
820
通信ネットワーク
サーバーコンピュータ
900A
電話機
900E
AR/VRデバイス
900G
パーソナルコンピュータ
900F
900

フロントページの続き

(33)優先権主張国・地域又は機関
米国(US)
ウェイ マイクロソフト テクノロジー ライセンシング，エルエルシー（番地無し）
(72)発明者 ルンデル，ベンジャミン エリオット
アメリカ合衆国 ワシントン州 ９８０５２ レッドモンド ワン マイクロソフト ウェイ マイクロソフト テクノロジー ライセンシング，エルエルシー（番地無し）
(72)発明者 ウォール，ラリー マーヴィン
アメリカ合衆国 ワシントン州 ９８０５２ レッドモンド ワン マイクロソフト ウェイ マイクロソフト テクノロジー ライセンシング，エルエルシー（番地無し）
(72)発明者 マクブライド，チャド ボーリング
アメリカ合衆国 ワシントン州 ９８０５２ レッドモンド ワン マイクロソフト ウェイ マイクロソフト テクノロジー ライセンシング，エルエルシー（番地無し）
(72)発明者 アンバルデカル，アモール アショク
アメリカ合衆国 ワシントン州 ９８０５２ レッドモンド ワン マイクロソフト ウェイ マイクロソフト テクノロジー ライセンシング，エルエルシー（番地無し）
(72)発明者 ピーター，ジョージ
アメリカ合衆国 ワシントン州 ９８０５２ レッドモンド ワン マイクロソフト ウェイ マイクロソフト テクノロジー ライセンシング，エルエルシー（番地無し）
(72)発明者 セドラ，ケント ディー
アメリカ合衆国 ワシントン州 ９８０５２ レッドモンド ワン マイクロソフト ウェイ マイクロソフト テクノロジー ライセンシング，エルエルシー（番地無し）
(72)発明者 ボブロフ，ボリス
アメリカ合衆国 ワシントン州 ９８０５２ レッドモンド ワン マイクロソフト ウェイ マイクロソフト テクノロジー ライセンシング，エルエルシー（番地無し）
審査官 北川 純次

(56)参考文献 米国特許出願公開第２０１０／３１２７３５（ＵＳ，Ａ１）
米国特許第５３６９７７３（ＵＳ，Ａ）
特開２０００－５９２２７（ＪＰ，Ａ）

(58)調査した分野 (Int.Cl.，ＤＢ名)
Ｇ０６Ｎ ３／０６
Ｇ０６Ｆ １２／０４
Ｈ０３Ｍ ７／４０