(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2018-92377
(P2018-92377A)
(43)公開日　平成30年6月14日(2018.6.14)

(51)Int.Cl.	F I	テーマコード（参考）
G06N 3/063 (2006.01)	G06N 3/063	

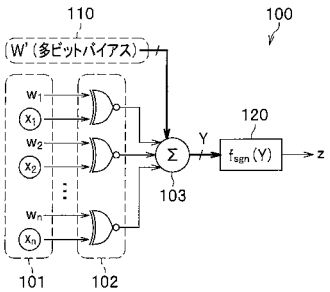
審査請求 有　請求項の数 11　ＯＬ　（全 23 頁）

(21)出願番号　特願2016-235383 (P2016-235383)
(22)出願日　平成28年12月2日 (2016.12.2)
(11)特許番号　特許第6183980号 (P6183980)
(45)特許公報発行日　平成29年8月23日 (2017.8.23)

(71)出願人　304021417
国立大学法人東京工業大学
東京都目黒区大岡山２丁目１２番１号
(74)代理人　110001807
特許業務法人磯野国際特許商標事務所
(72)発明者　中原　啓貴
東京都目黒区大岡山２－１２－１　国立大学法人東京工業大学内
(72)発明者　米川　晴義
東京都目黒区大岡山２－１２－１　国立大学法人東京工業大学内

(54)【発明の名称】ニューラルネットワーク回路装置、ニューラルネットワーク、ニューラルネットワーク処理方法およびニューラルネットワークの実行プログラム

(57)【要約】
【課題】バッチ正規化回路が不要なニューラルネットワーク回路装置、ニューラルネットワーク、ニューラルネットワーク処理方法およびニューラルネットワークの実行プログラムを提供する。
【解決手段】２値化ニューラルネットワーク回路１００は、入力値ｘ１～ｘｎ（ｘｉ）（２値）を入力する入力ノードおよび重みｗ１～ｗｎ（ｗｉ）を入力する入力部１０１と、入力値ｘ１～ｘｎおよび重みｗ１～ｗｎを受け取り、ＸＮＯＲ論理を取るＸＮＯＲゲート回路１０２と、多ビットバイアスＷ’を入力する多ビットバイアスＷ’入力部１１０と、各ＸＮＯＲ論理値と多ビットバイアスＷ’との総和を取る総和回路１０３と、総和を取った信号Ｙに対して符号ビットのみを出力する活性化回路１２０と、を備える。
【選択図】図９



【特許請求の範囲】
【請求項１】
入力層、１以上の中間層、および、出力層を少なくとも含むニューラルネットワークにおいて、前記中間層の中で、入力値に重みづけとバイアスを乗算するニューラルネットワーク回路装置であって、
入力値ｘｉおよび重みｗｉを受け取り、論理演算を行う論理回路部と、
多ビットバイアスW'を受け取り、前記論理回路部の出力と前記多ビットバイアスW'との総和を取る総和回路部と、
総和を取った多ビット信号Ｙに対して符号ビットのみを出力する活性化回路部と、を備える
ことを特徴とする記載のニューラルネットワーク回路装置。
【請求項２】
前記入力値ｘｉおよび前記重みｗｉを入力する入力部と、
前記多ビットバイアスW'を入力する多ビットバイアス入力部と、を備える
ことを特徴とする請求項１に記載のニューラルネットワーク回路装置。
【請求項３】
前記入力値ｘｉおよび前記重みｗｉは、２値信号である
ことを特徴とする請求項１または請求項２に記載のニューラルネットワーク回路装置。
【請求項４】
前記多ビットバイアスW'は、学習後の多ビットバイアス値である
ことを特徴とする請求項１または請求項２に記載のニューラルネットワーク回路装置。
【請求項５】
前記論理回路部は、否定排他的論理和または排他的論理和を含む
ことを特徴とする請求項１に記載のニューラルネットワーク回路装置。
【請求項６】
前記論理回路部は、ＬＵＴ（Look-Up Table）である
ことを特徴とする請求項１に記載のニューラルネットワーク回路装置。
【請求項７】
前記符号ビットは、総和を取った前記多ビット信号Ｙを活性化するかしないかで示す２値信号である
ことを特徴とする請求項１に記載のニューラルネットワーク回路装置。
【請求項８】
前記多ビット信号Ｙおよび多ビットバイアスW'は、下記式で示される
【数４】

$$Y = \sum_{i=1}^{n} w_i x_i + W' \cdot$$

ことを特徴とする請求項１に記載のニューラルネットワーク回路装置。
【請求項９】
請求項１乃至８のいずれか１項に記載のニューラルネットワーク回路装置を備えるニューラルネットワーク。
【請求項１０】
入力層、１以上の中間層、および、出力層を少なくとも含むニューラルネットワークにおいて、前記中間層の中で、入力値に重みづけとバイアスを乗算するニューラルネットワーク処理方法であって、
入力値ｘｉおよび重みｗｉを受け取り、論理演算を行うステップと、
多ビットバイアスW'を受け取り、前記論理回路部の出力と前記多ビットバイアスW'との総和を取るステップと、
総和を取った多ビット信号Ｙに対して符号ビットのみを出力するステップと、を有する

ことを特徴とするニューラルネットワーク処理方法。
【請求項１１】
入力層、１以上の中間層、および、出力層を少なくとも含むニューラルネットワークにおいて、前記中間層の中で、入力値に重みづけとバイアスを乗算するニューラルネットワーク回路装置としてのコンピュータを、
入力値ｘｉおよび重みｗｉを受け取り、論理演算を行う論理回路手段、
多ビットバイアスW'を受け取り、前記論理回路部の出力と前記多ビットバイアスW'との総和を取る総和回路手段、
総和を取った多ビット信号Ｙに対して符号ビットのみを出力する活性化回路手段、
として機能させるためのニューラルネットワークの実行プログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、ニューラルネットワーク回路装置、ニューラルネットワーク、ニューラルネットワーク処理方法およびニューラルネットワークの実行プログラムに関する。
【背景技術】
【０００２】
古典的な順伝搬型ニューラルネットワーク（ＦＦＮＮ：Feedforward Neural Network）、ＲＢＦ（Radial Basis Function）ネットワーク、正規化したＲＢＦネットワーク、自己組織化マップなどがある。ＲＢＦＮは、誤差逆伝搬法に用いる活性化関数に放射基底関数を用いる。しかし、中間層が多く取れず高精度認識判定が難しかったり、ＨＷ規模が大きく処理時間がかかる、などの問題があり手書き文字認識など応用分野が限定されていた。
近年、ＡＤＡＳ（advanced driver assistance system）用の画像認識や自動翻訳などで注目を集める新方式として畳み込みニューラルネットワーク（ＣＮＮ：Convolutional Neural Network）（層間が全結合でないＮＮ）や再帰型ニューラルネットワーク（双方向伝搬）が登場している。ＣＮＮは、ディープニューラルネットワーク（ＤＮＮ：Deep Neural Network）に畳込み演算を付加したものである。
【０００３】
特許文献１には、誤り訂正符号の検査行列に基づいて、階層型ニューラルネットワークにおける疎結合のノード間で学習された重みの値と入力信号とを用いて、問題を解く処理部を備える処理装置が記載されている。
【０００４】
既存のＣＮＮは、短精度（多ビット）による積和演算回路で構成されており、多数の乗算回路が必要である。このため、面積・消費電力が多大になる欠点があった。そこで、２値化した精度、すなわち＋１と－１のみ用いてＣＮＮを構成する回路が提案されている（例えば、非特許文献１～４参照）。
【先行技術文献】
【特許文献】
【０００５】
【特許文献１】特開２０１６－１７３８４３号公報
【非特許文献】
【０００６】
【非特許文献１】M. Courbariaux, I. Hubara, D. Soudry, R.E.Yaniv, Y. Bengio, “Binarized neural networks: Training deep neural networks with weights and activations constrained to +1 or -1," Computer Research Repository (CoRR)、「２値化ＮＮのアルゴリズム」、[online]、２０１６年３月、［平成２８年１０月５日検索］、<URL:http:// arxiv.org/pdf/1602.02830v3.pdf >
【非特許文献２】Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, Ali Farhadi, “XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks,

”Computer Vision and Pattern recognition、「２値化ＮＮのアルゴリズム」、[online]、２０１６年３月、［平成２８年１０月５日検索］、<URL: https://arxiv.org/pdf/1603.05279v4 >
【非特許文献３】Hiroki Nakahara, Haruyoshi Yonekawa, Tsutomu Sasao, Hisashi Iwamoto and Masato Motomura, ” A Memory-Based Realization of a Binarized Deep Convolutional Neural Network,” Proc. of the 2016 International Conference on Field-Programmable Technology (FPT), Xi'an, China, Dec 2016 (To Appear).
【非特許文献４】Eriko Nurvitadhi, David Sheffield, Jaewoong Sim, Asit Mishra, Ganesh Venkatesh, Debbie Marr, ”Accelerating Binarized Neural Networks: Comparison of FPGA, CPU, GPU, and ASIC,” Proc. of the 2016 International Conference on Field-Programmable Technology (FPT), Xi'an, China, Dec 2016 (To Appear).

【発明の概要】
【発明が解決しようとする課題】
【０００７】
非特許文献１～４の技術では、精度を２値に落とすことでＣＮＮの認識精度も落としてしまう。これを避けて２値化ＣＮＮの精度を維持するためには、バッチ正規化回路が必要であるが、バッチ正規化回路は、複雑な回路であり、面積・消費電力が増大するという課題があった。
【０００８】
本発明は、このような事情に鑑みてなされたものであり、バッチ正規化回路が不要なニューラルネットワーク回路装置、ニューラルネットワーク、ニューラルネットワーク処理方法およびニューラルネットワークの実行プログラムを提供することを課題とする。
【課題を解決するための手段】
【０００９】
前記した課題を解決するため、本発明に係るニューラルネットワーク回路装置は、入力層、１以上の中間層、および、出力層を少なくとも含むニューラルネットワークにおいて、前記中間層の中で、入力値に重みづけとバイアスを乗算するニューラルネットワーク回路装置であって、入力値ｘｉおよび重みｗｉを受け取り、論理演算を行う論理回路部と、多ビットバイアスW'を受け取り、前記論理回路部の出力と前記多ビットバイアスW'との総和を取る総和回路部と、総和を取った多ビット信号Ｙに対して符号ビットのみを出力する活性化回路部と、を備えることを特徴とする。
【発明の効果】
【００１０】
本発明によれば、バッチ正規化回路が不要なニューラルネットワーク回路装置、ニューラルネットワーク、ニューラルネットワーク処理方法およびニューラルネットワークの実行プログラムを提供することができる。
【図面の簡単な説明】
【００１１】
【図１】ディープニューラルネットワーク（ＤＮＮ）の構造の一例を説明する図である。
【図２】比較例のニューラルネットワークのニューラルネットワーク回路の構成の一例を示す図である。
【図３】図２に示すニューラルネットワーク回路における活性化関数ｆact(Y)を示す図である。
【図４】図２に示すニューラルネットワーク回路の乗算回路をＸＮＯＲゲート回路に置き換えた２値化ニューラルネットワーク回路の構成の一例を示す図である。
【図５】図４に示す２値化ニューラルネットワーク回路における活性化関数ｆsgn(B)を示す図である。
【図６】比較例のバッチ正規化回路を備える２値化ニューラルネットワーク回路の構成の一例を示す図である。
【図７】ニューラルネットワークの２値化ニューラルネットワーク回路のスケーリング（

γ）による正規化を示す図である。
【図８】ニューラルネットワークの２値化ニューラルネットワーク回路のシフト（β）による－１～＋１の制限を示す図である。
【図９】本発明の実施形態に係るディープニューラルネットワークの２値化ニューラルネットワーク回路の構成を示す図である。
【図１０】本発明の実施形態に係るディープニューラルネットワークの２値化ニューラルネットワーク回路の活性化回路を示す図である。
【図１１】本発明の実施形態に係るディープニューラルネットワークの多ビット構成のニューラルネットワーク回路と２値化ニューラルネットワーク回路の認識精度を説明する図である。
【図１２】本発明の実施形態に係るディープニューラルネットワークの２値化ニューラルネットワーク回路と既存の多ビット実装法との比較を行った結果を表にして示す図である。
【図１３】本発明の実施形態に係るディープニューラルネットワークの２値化ニューラルネットワーク回路の実装例を説明する図である。
【図１４】変形例のディープニューラルネットワークの２値化ニューラルネットワーク回路の構成を示す図である。
【図１５】変形例の２値化ニューラルネットワーク回路のＬＵＴの構成を示す図である。
【発明を実施するための形態】
【００１２】
以下、図面を参照して本発明を実施するための形態（以下、「本実施形態」という）におけるディープニューラルネットワークについて説明する。
（背景説明）
図１は、ディープニューラルネットワーク（ＤＮＮ）の構造の一例を説明する図である。
図１に示すように、ディープニューラルネットワーク（ＤＮＮ）１は、入力層（input layer）１１、任意の数の中間層である隠れ層（hidden layer）１２、出力層（output layer）１３から構成される。
入力層（input layer）１１は、複数個（ここでは８）の入力ノード（ニューロン）を有する。隠れ層１２は、複数（ここでは３層（hidden layer1, hidden layer2, hidden layer3））である。実際には、隠れ層１２の層数ｎは、例えば２０～１００に達する。出力層１３は、識別対象の数（ここでは４）の出力ノード（ニューロン）を有する。なお、層数およびノード数（ニューロン数）は、一例である。
ディープニューラルネットワーク１は、入力層１１と隠れ層１２のノード間が全て結合し、隠れ層１２と出力層１３のノード間が全て結合している。
【００１３】
入力層１１、隠れ層１２および出力層１３には、任意の数のノード（図１の○印参照）が存在する。このノードは、入力を受け取り、値を出力する関数である。入力層１１には、入力ノードとは別に独立した値を入れるバイアス（bias）ノードがある。構成は、複数のノードを持つ層を重ねることで構築される。伝播は、受け取った入力に対して重み（weight）をかけ、受け取った入力を次層に活性化関数（activation function）で変換して出力する。活性化関数は、sigmoid関数やtanh関数などの非線形関数、ReLU（Rectified Linear Unit function：正規化線形関数）がある。ノード数を増やすことで、扱う変数を増やし、多数の要素を加味して値／境界を決定できる。層数を増やすことで、直線境界の組み合わせ、複雑な境界を表現できる。学習は、誤差を計算し、それを基に各層の重みを調整する。学習は、誤差を最小化する最適化問題を解くことであり、最適化問題の解法は誤差逆伝播法（Backpropagation）を使うのが一般的である。誤差は、二乗和誤差を使うのが一般的である。汎化能力を高めるために、誤差に正則化項を加算する。誤差逆伝播法は、誤差を出力層１３から伝播させていき、各層の重みを調整する。
【００１４】

図１のディープニューラルネットワーク１の構成を２次元に展開することで画像処理に適したＣＮＮを構築できる。また、ディープニューラルネットワーク１にフィードバックを入れることで、双方向に信号が伝播するＲＮＮ（Recurrent Neural Network：再帰型ニューラルネットワーク）を構成することができる。

【００１５】

図１の太破線三角部に示すように、ディープニューラルネットワーク１は、多層のニューラルネットワークを実現する回路（以下、ニューラルネットワーク回路という）２から構成されている。

本技術は、ニューラルネットワーク回路２を対象とする。ニューラルネットワーク回路２の適用箇所および適用数は限定されない。例えば、隠れ層１２の層数ｎ：２０～３０の場合、これらの層のどの位置に適用してもよく、またどのノードを入出力ノードとするものでもよい。さらに、ディープニューラルネットワーク１に限らず、どのようなニューラルネットワークでもよい。ただし、入力層１１または出力層１３のノード出力には、２値化出力ではなく多ビット出力が求められるので、ニューラルネットワーク回路２は、対象外である。ただし、出力層１３のノードを構成する回路に、乗算回路が残ったとしても面積的には問題にはならない。

なお、入力データに対し学習済のものを評価していくことを前提としている。したがって、学習結果として重みｗｉは既に得られている。

【００１６】

<ニューラルネットワーク回路>

図２は、比較例のニューラルネットワーク回路の構成の一例を示す図である。

比較例のニューラルネットワーク回路２０は、図１のディープニューラルネットワーク１を構成するニューラルネットワーク回路２に適用できる。なお、以下の各図の表記において、値が多ビットである場合は太実線矢印、値が２値である場合は細太実線矢印で示す。

ニューラルネットワーク回路２０は、入力値（判別データ）Ｘ１～Ｘｎ（多ビット）を入力する入力ノード、重みＷ１～Ｗｎ（多ビット）およびバイアスＷ０（多ビット）を入力する入力部２１と、入力値Ｘ１～Ｘｎおよび重みＷ１～Ｗｎを受け取り、入力値Ｘ１～Ｘｎに重みＷ１～Ｗｎをそれぞれ乗算する複数の乗算回路２２と、各乗算値とバイアスＷ０との総和を取る総和回路２３と、総和を取った信号Ｙを活性化関数ｆact(Y)で変換する活性化関数回路２４と、を備えて構成される。

以上の構成において、ニューラルネットワーク回路２０は、入力値Ｘ１～Ｘｎ（多ビット）を受け取り、重みＷ１～Ｗｎを乗算した後に、バイアスＷ０を含めて総和を取った信号Ｙを活性化関数回路２４を通すことで人間のニューロンに模した処理を実現している。

【００１７】

図３は、前記図２に示すニューラルネットワーク回路２０における活性化関数ｆact(Y)を示す図である。図３は、横軸に総和を取った信号Ｙ、縦軸に活性化関数fact(Y)の値をとる。図３の符号○印は、±１の範囲の値をとる正側の活性化値（状態値）、図３の符号×印は、±１の範囲の値をとる負側の活性化値である。

ニューラルネットワーク回路２０（図２参照）は、多ビットで高い認識精度を実現している。このため、活性化関数回路２４（図２参照）において、非線形な活性化関数ｆact(Y)を用いることができる。すなわち、図４に示すように、非線形な活性化関数ｆact(Y)は、傾きが非ゼロとなる部分（図４の破線囲み部分参照）に±１の範囲の値をとる活性化値を設定できる。このため、ニューラルネットワーク回路２０は、多様な活性を実現でき、認識精度は実用的な値になっていた。しかし、ニューラルネットワーク回路２０は、大量の乗算回路２２が必要になる。加えて、ニューラルネットワーク回路２０は、入出力・重みが多ビットであることにより、大量のメモリが必要であり、読み書きの速度（メモリ容量・帯域）も問題である。

【００１８】

<単に２値化した２値化ニューラルネットワーク回路>

図２に示す比較例のニューラルネットワーク回路２０は、短精度（多ビット）による積和演算回路で構成されている。このため、多数の乗算回路２１が必要であり、面積・消費電力が多大になる欠点があった。また、入出力・重みが多ビットであることで大量のメモリが必要であり、読み書きの速度（メモリ容量・帯域）が問題となっていた。

そこで、２値化した精度、すなわち＋１と－１のみ用いてニューラルネットワーク回路２（図１参照）を構成する回路が提案された（非特許文献１～４）。具体的には、図２に示すニューラルネットワーク回路２０の乗算回路２１を、論理ゲート（例えばＸＮＯＲゲート回路）に置き換えることが考えられる。

【００１９】

図４は、図２に示すニューラルネットワーク回路２０の乗算回路２１をＸＮＯＲゲート回路に置き換えた２値化ニューラルネットワーク回路の構成の一例を示す図である。

比較例の２値化ニューラルネットワーク回路３０は、図１のニューラルネットワーク回路２に適用できる。

図４に示すように、比較例の２値化ニューラルネットワーク回路３０は、入力値ｘ１～ｘｎ（２値）を入力する入力ノード、重みｗ１～ｗｎ（２値）およびバイアスｗ０（２値）を入力する入力部３１と、入力値ｘ１～ｘｎおよび重みｗ１～ｗｎを受け取り、ＸＮＯＲ（Exclusive NOR：否定排他的論理和）論理を取る複数のＸＮＯＲゲート回路３２と、ＸＮＯＲゲート回路３２の各ＸＮＯＲ論理値とバイアスｗ０との総和を取る総和回路３３と、総和を取った信号Ｙのバッチ正規化した信号Ｂを活性化関数ｆsgn(B)で変換する活性化関数回路３４と、を備えて構成される。

２値化ニューラルネットワーク回路３０は、乗算回路２１（図２参照）がＸＮＯＲ論理を実現するＸＮＯＲゲート回路３２に置き換えられている。このため、乗算回路２１を構成する際に必要であった面積を削減することができる。また、入力値ｘ１～ｘｎ、出力値ｚ、および重みｗ１～ｗｎは、いずれも２値（－１と＋１）であるため、多値である場合と比較してメモリ量を大幅に削減でき、メモリ帯域を向上させることができる。

【００２０】

図５は、前記図４に示す２値化ニューラルネットワーク回路３０における活性化関数ｆsgn(B)を示す図である。図５は、横軸に総和を取った信号Ｙ、縦軸に活性化関数ｆsgn(B)の値をとる。図５の符号○印は、±１の範囲の値をとる正側の活性化値、図５の符号×印は、±１の範囲の値をとる負側の活性化値である。

２値化ニューラルネットワーク回路３０は、入力値ｘ１～ｘｎおよび重みｗ１～ｗｎを単に２値化している。このため、図５の符号ａに示すように、±１のみ扱う活性化関数しか扱えないため、誤差が頻繁に生じてしまう。また、傾きが非ゼロとなる区間（図５の破線囲み部分参照）が不均等となり学習が上手く行われない。すなわち、図６の符号ｂに示すように、不均等な幅により微分が定義できない。その結果、単に２値化した２値化ニューラルネットワーク回路４０は、認識精度が大幅に落ち込んでしまう。

そこで、非特許文献１～４には、既存の２値化ニューラルネットワークの精度を維持するためにバッチ正規化を行う技術が記載されている。

【００２１】

<バッチ正規化回路を備える２値化ニューラルネットワーク回路>

図６は、２値化した精度を是正して、ＣＮＮの認識精度を保つバッチ正規化回路を備える２値化ニューラルネットワーク回路４０の構成の一例を示す図である。図４と同一構成部分には同一符号を付している。

図６に示すように、比較例の２値化ニューラルネットワーク回路４０は、入力値ｘ１～ｘｎ（２値）を入力する入力ノードｘ１～ｘｎ、重みｗ１～ｗｎ（２値）およびバイアスｗ０（２値）を入力する入力部３１と、入力値ｘ１～ｘｎおよび重みｗ１～ｗｎを受け取り、ＸＮＯＲ（Exclusive NOR：否定排他的論理和）論理を取る複数のＸＮＯＲゲート回路３２と、ＸＮＯＲゲート回路３２の各ＸＮＯＲ論理値とバイアスｗ０（２値）との総和を取る総和回路３３と、２値化によるバラツキの偏りを正規化範囲を広げ中心をシフトさせる処理で是正するバッチ正規化回路４１と、総和を取った信号Ｙのバッチ正規化した信

号Ｂを活性化関数ｆsgn(B)で変換する活性化関数回路３４と、を備えて構成される。
【００２２】
バッチ正規化回路４１は、重み総和後、スケーリング（γ）値（多ビット）による正規化を行う乗算回路４２と、スケーリング（γ）値による正規化後、シフト（β）値（多ビット）によりシフトして２分類を行う加算器４３と、からなる。スケーリング（γ）値およびシフト（β）値の各パラメータは、事前に学習時に求めておく。
２値化ニューラルネットワーク回路４０は、バッチ正規化回路４１を備えることで、２値化した精度を是正して、ＣＮＮの認識精度を保つようにする。
なお、入力値ｘ１～ｘｎと重みｗ１～ｗｎとのＸＮＯＲ論理を取る論理回路であれば、ＸＮＯＲゲートに限らずどのような論理ゲートでもよい。
【００２３】
しかしながら、図６に示すように、バッチ正規化回路４０は、乗算回路４２と加算器４３が必要である。また、スケーリング（γ）値およびシフト（β）値をメモリに格納しておく必要がある。メモリは、外付けであり大面積となり、読み出しの速度遅延となる。
２値化ニューラルネットワーク回路４０は、図２に示すニューラルネットワーク回路２０のように、多数の乗算回路２１を必要とはしないものの、バッチ正規化回路４１において、面積が大きい乗算回路４２と加算器４３が必要となる。バッチ正規化回路４１では、パラメータ格納用のメモリも必要となり、面積とメモリ帯域の削減が求められている。
【００２４】
<バッチ正規化回路が必要となる理由>
比較例の２値化ニューラルネットワーク回路４０のバッチ正規化回路４１が必要となる理由について説明する。
図７および図８は、比較例の２値化ニューラルネットワーク回路４０のバッチ正規化による効果を説明する図である。図７は、スケーリング（γ）による正規化を示す図、図８は、シフト（β）による－１～＋１の制限を示す図である。
バッチ正規化とは、２値化によるバラツキの偏りを是正する回路であり、重み総和後、スケーリング(γ)値による正規化を行った後、シフト(β)値による適切な活性化による２分類を行う。これらのパラメータは事前に学習時に求めておく。具体的には、下記の通りである。
【００２５】
図７の白抜矢印および符号ｃに示すように、バッチ正規化回路４１の乗算回路４２（図６参照）は、重み総和後の信号（結果）Ｙを、スケーリング（γ）値により、幅「２」（図７の網掛け部参照）に正規化する。これにより、図５の幅（図５の網掛け部参照）と比較して分かるように、単に２値化した２値化ニューラルネットワーク回路３０では、不均等な幅により微分が定義できなかった不具合が、スケーリング（γ）値により幅「２」に正規化することで、不均等な幅が抑制される。
【００２６】
その上で、図８の白抜矢印および符号ｄに示すように、バッチ正規化回路４１の加算器４３（図６参照）は、スケーリング（γ）値による正規化後の値を、シフト(β)値により－１～＋１の範囲になるよう制限する。すなわち、図５の幅（図５の網掛け部参照）と比較して分かるように、図５の幅（図５の網掛け部参照）が、＋１側により多くシフトしている場合には、シフト(β)値により、スケーリング（γ）値による正規化後の値を－１～＋１に制限することで、この幅の中心を０とする。図５の例では、負側の活性化値（図５の破線囲み部の符号×印参照）が、本来あるべき負側に戻される。これにより、誤差の発生が減少し、認識精度を高めることができる。
このように、２値化ニューラルネットワーク回路４０には、バッチ正規化回路４１が必要である。
【００２７】
<バッチ正規化回路を備える２値化ニューラルネットワーク回路の課題>
上記バッチ正規化回路４１を導入することで、２値化ニューラルネットワーク回路４０

の認識精度は、図２に示したニューラルネットワーク回路２０とほぼ等しくなる。しかしながら、バッチ正規化回路４１は、乗算回路４２と加算器４３が必要であり、多ビットのスケーリング（γ）値およびシフト(β)値をメモリに格納しておく必要がある。このため、２値化ニューラルネットワーク回路４０は、依然として複雑な回路であり、面積・消費電力を低減したいという切実な要請がある。

【００２８】

２値化ニューラルネットワーク回路２０では、例えば８，９ビットを１ビットに落とすので、計算精度が落ちる。ＮＮに用いた場合、誤認識（認識失敗率）が８０％に増加し、使用に耐えない。そこで、バッチ正規化で対処することになる。しかし、バッチ正規化回路４１は、除算、若しくは浮動小数点の乗算と加算が必要であり、ハードウェア(ＨＷ)化して実装することが非常に困難である。また、外部メモリが必要となり、外部メモリとのアクセスのために遅延となる。

【００２９】

（本発明の原理説明）

本発明者らは、バッチ正規化の操作を導入したＮＮに対して、これと等価なＮＮを解析的に求めると、バッチ正規化が不要なＮＮを得ることができることを発見したことが着眼点である。例えば、従来では、図３のような非線形な活性化関数ｆact(Y)に対し、図３の符号○印に示す状態値が得られた場合、不均等な幅を正規化するためスケーリングを行っていた。多ビットを２値して乗算する場合の演算精度を確保するためである。しかしながら、本発明者らは、ニューラルネットワーク回路における２値化の本質は、活性化するかしないかの（２値）だけに帰着することに着目した。スケーリングは、不要となりシフトのみで対応できることになる。

すなわち、重み積和後に２値化ニューラルネットワーク回路４０のバッチ正規化回路４１（図６参照）に入力される信号をＹとすると、バッチ正規化回路４１から出力される信号（Ｙと等価となる信号）Ｙ'は、次式（１）で示される。

【００３０】

【数１】

$$Y' = \gamma \frac{Y - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} + \beta$$

$$= \frac{\gamma}{\sqrt{\sigma_B^2+\epsilon}} \left(Y - \left(\mu_B - \frac{\sqrt{\sigma_B^2+\epsilon}}{\gamma} \beta \right) \right) \cdot \quad \cdots (1)$$

ただし、

γ：スケーリング値

β：シフト値

μ_B：平均値

σ^2_B：二乗和誤差

ε：パラメータ（調整用）

したがって、２値化活性化関数の値ｆ'sgn(Y)は、下記式（２）の条件で決まる。

【００３１】

【数２】

$$f'_{sgn}(Y)=\begin{cases} 1 & \left(if\ Y<-\mu_B+\frac{\sqrt{\sigma_B^2+\epsilon}}{\gamma}\beta\right) \\ -1 & (otherwise)\end{cases} \quad \cdots(2)$$

【００３２】

よって、これらの解析的な操作から重み積和演算は、下記式（３）のように得られる。

【００３３】

【数３】

$$Y=\sum_{i=0}^{n} w_i x_i-\mu_B+\frac{\sqrt{\sigma_B^2+\epsilon}}{\gamma}\beta$$

$$=\sum_{i=1}^{n} w_i x_i+\left(w_0-\mu_B+\frac{\sqrt{\sigma_B^2+\epsilon}}{\gamma}\beta\right) \quad \cdots(3)$$

$$=\sum_{i=1}^{n} w_i x_i+W'\cdot$$

ただし、

W'：多ビットバイアス

【００３４】

バッチ正規化学習後に、これらの数学的な操作により、バッチ正規化と等価なＮＮの演算が得られる。

上記式（３）は、回路的にはバイアス値のみ多ビット構成にすればよいことを示している。回路は単純だが、単にバイアス値のみ多ビットにしただけでは認識精度は向上せず、これらの解析的な洞察がなければ成り立たない。

【００３５】

［実施形態の構成］

図９は、本発明の実施形態に係るニューラルネットワークの２値化ニューラルネットワーク回路の構成を示す図である。本実施形態の２値化ニューラルネットワーク回路は、ディープニューラルネットワークへの実装技術を提供する。

２値化ニューラルネットワーク回路１００は、図１のニューラルネットワーク回路２に適用できる。

図９に示すように、２値化ニューラルネットワーク回路１００（ニューラルネットワーク回路装置）は、入力値ｘ１～ｘｎ（ｘｉ）（２値）を入力する入力ノードおよび重みｗ１～ｗｎ（ｗｉ）（２値）を入力する入力部１０１と、入力値ｘ１～ｘｎおよび重みｗ１～ｗｎを受け取り、ＸＮＯＲ論理を取るＸＮＯＲゲート回路１０２（論理回路部）と、多ビットバイアスW'（式（３）参照）を入力する多ビットバイアスW'入力部１１０と、各ＸＮＯＲ論理値と多ビットバイアスW'との総和を取る総和回路１０３（総和回路部）と、総和を取った信号Ｙに対して符号ビットのみを出力する活性化回路１２０（活性化回路部）と、を備えて構成される。

【００３６】

上記入力値ｘｉ（２値）および重みｗｉ（２値）は、２値信号である。

上記多ビット信号Ｙおよび多ビットバイアスW'は、前記式（３）で示される。

【００３７】

２値化ニューラルネットワーク回路１００は、ディープニューラルネットワーク１の隠れ層１２（図１参照）に適用される。ここでは、ディープニューラルネットワーク１において、入力値に対し学習済のものを評価していくことを前提としている。したがって、学習結果として重みの多ビットバイアスW'は既に得られている。多ビットバイアスW'は、学習後の多ビットバイアス値である。ちなみに、図２のニューラルネットワーク回路２０では、多ビットの重みＷ１～Ｗｎおよびバイアスｗ０を用いるが、本実施形態の多ビットバイアスW'は、図２のニューラルネットワーク回路２０における多ビットのバイアスＷ０とは異なるものである。

なお、ＮＮでは、重みが、クライアントの認識物体毎に全て異なる。また学習により毎回異なることがある。画像処理では係数は、全て同じであり、この点でＮＮと画像処理では、ＨＷが大きく異なる。

【００３８】

ＸＮＯＲゲート回路１０２は、排他的論理和を含むどのような論理回路部でもよい。すなわち、入力値ｘ１～ｘｎと重みｗ１～ｗｎとの論理を取る論理回路であれば、ＸＮＯＲゲートに限らずどのようなゲート回路でもよい。例えば、ＸＯＲゲートにＮＯＴゲートを組み合わせる、ＡＮＤ，ＯＲゲートを組み合わせる、さらにはトランジスタスイッチを用いて作製するなど、論理的に等しいものであればどのようなものでもよい。

【００３９】

活性化回路１２０は、総和を取った信号Ｙに対して符号ビットのみを出力する活性化関数回路を模擬する回路となっている。符号ビットは、総和を取った多ビット信号Ｙを活性化するかしないかで示す２値信号である。

【００４０】

このように、２値化ニューラルネットワーク回路１００は、バイアス値のみ多ビット構成にし、バイアス値を含めた総和から、符合ビットのみを出力する活性化回路１２０を備える。すなわち、２値化ニューラルネットワーク回路１００は、図６に示す２値化ニューラルネットワーク回路４０のバッチ正規化回路４１および活性化関数回路３４を、符号ビットのみを出力する活性化回路１２０で置き換えた構成となっている。このため、２値化ニューラルネットワーク回路１００は、複雑なバッチ正規化回路４１が不要なニューラルネットワーク回路となっている。

【００４１】

図１０は、２値化ニューラルネットワーク回路の活性化回路を示す図である。

図１０に示すように、活性化回路１２０は、バイアス値を含めた総和の出力Ｙから、符合ビットのみを出力する回路である。図１０の回路では、符号ビットは、出力ｙ[0]，ｙ[1]，…，ｙ[n-1]のうち、最上位ビットｙ[n-1]であるとすると、符号ビットとして最上位ビットｙ[n-1]のみを出力する。活性化回路１２０は、最上位ビットｙ[n-1]のみが出力ｚとして出力される。図９では、活性化回路１２０は、活性化関数ｆsgn(Y)と表記されているが、図６に示すスケーリング（γ）による正規化およびシフト（β）による－１～＋１の制限は行っておらず、最上位ビットｙ[n-1]のみを出力する回路である。

【００４２】

以下、上述のように構成された２値化ニューラルネットワーク回路１００の動作について説明する。

２値化ニューラルネットワーク回路１００は、図１に示すディープニューラルネットワーク１のニューラルネットワーク回路２に用いられる。この場合、２値化ニューラルネットワーク回路１００の入力ノードｘ１～ｘｎは、図１に示すディープニューラルネットワーク１のhidden layer1の入力ノードである。入力部１０１には、隠れ層１２のhidden layer1の入力ノードの入力値ｘ１～ｘｎ（２値）および重みｗ１～ｗｎ（２値）が入力される。

ＸＮＯＲゲート回路１０２では、入力値ｘ１～ｘｎおよび重みｗ１～ｗｎを受け取り、

ＸＮＯＲ論理により２値（－１／＋１）の乗算を行う。
２値化ニューラルネットワーク回路１００は、多ビット構成の乗算回路２１（図２参照）がＸＮＯＲ論理を実現するＸＮＯＲゲート回路１０２に置き換えられている。このため、乗算回路２１を構成する際に必要であった面積を削減することができる。また、入力値ｘ１～ｘｎおよび重みｗ１～ｗｎは、いずれも２値（－１／＋１）であるため、多ビット（多値）である場合と比較してメモリ容量を大幅に削減でき、メモリ帯域を向上させることができる。
【００４３】
一方、前記式（３）に従った多ビットバイアスW'を入力する。多ビットバイアスW'は、２値化ニューラルネットワーク回路３０，４０（図４および図６参照）のような２値のバイアスｗ０ではない。また、多ビットではあっても２値化ニューラルネットワーク回路２０（図２参照）のようなバイアスＷ０とは異なる。多ビットバイアスW'は、前記式（３）に示すように、前記バイアスｗ０（２値）からバッチ正規化分を調整した、学習後のバイアス値である。
総和回路１０３には、バイアス値のみ多ビット構成にした多ビットバイアスW'が入力される。総和回路１０３は、ＸＮＯＲゲート回路１０２の各ＸＮＯＲ論理値と多ビットバイアスW'との総和を取り、総和の出力Ｙ（多ビット）を活性化回路１２０に出力する。
【００４４】
図１０に示すように、活性化回路１２０では、バイアス値を含めた総和の出力Ｙ（多ビット）から、符合ビットのみを出力する。図１０の回路では、符号ビットは、出力ｙ[0]，ｙ[1]，…，ｙ[n-1]のうち、最上位ビットｙ[n-1]である。活性化回路１２０は、バイアス値を含めた総和の出力Ｙから、最上位ビットｙ[n-1]のみを出力ｚとして出力する。換言すれば、活性化回路１２０は、ｙ[0]，ｙ[1]，…，ｙ[n-2]の数値を出力しない（ｙ[0]，ｙ[1]，…，ｙ[n-2]の数値は使用しない）。
例えば、活性化回路１２０の入力Ｙとして４～５ｂｉｔの信号が入力された場合、ＨＷでは、通常最上位ビットを符号ビットとするので、最上位ビット（符合ビット）だけを出力する。すなわち、活性化回路１２０からは、活性化するかしないかの二通り（２値、すなわち＋１か－１）が出力され、それが後段の中間層（隠れ層）のノードに伝達される。
【００４５】
２値化ニューラルネットワーク回路１００は、前記式（３）に示されるように、バッチ正規化の操作を導入したＮＮと等価なＮＮである。前記式（３）は、下記により実現される。すなわち、２値（１ビットのみ）にした入力値ｘｉと重みｗｉ、多ビットバイアスW'を入力とし、乗算の代わりとなるＸＮＯＲ論理を取った後、バイアス値を含めたそれらの総和を取り（前記式（３）の第１項）、活性化回路１２０がバイアス値を含めた総和の出力Ｙから、符合ビットのみを出力する（前記式（３）の第２項）。
このため、活性化回路１２０は、バイアス値を含めた総和の出力Ｙから、符号ビットのみを出力する回路ではあるが、機能的には、活性化関数回路ｆsgn(Y)と同様な機能、すなわち活性化関数回路ｆsgn(Y)を模擬した回路となっている。
【００４６】
本実施形態の効果を確認するため、VGG16（隠れ層が１６層） ベンチマークＮＮを実装した。VGG16は、良く使われているベンチマークで再現性があるものである。
図１１は、多ビット構成のニューラルネットワーク回路と２値化ニューラルネットワーク回路の認識精度を説明する図である。図１１（ａ）は多ビット（３２ビット浮動小数点）で構成したニューラルネットワーク回路２０（図２参照）の認識精度、図１１（ｂ）は２値化ニューラルネットワーク回路１００の認識精度を示す。図１１の横軸は、利用した学習データに対して更新を終えたサイクルであるエポック（epoch）数、縦軸は誤認識(誤差)（Classification error）である。図１１は、本実施形態をVGG16 ベンチマークＮＮで実装し確認したものである。また、図１１（ａ）は、ディープニューラルネットワーク用のフレームワークソフトウェアChainer（登録商標）ののfloat32 CNNを用いている。また、図１１（ｂ）は、ディープニューラルネットワーク用のフレームワークソフトウェア

Chainer（登録商標）のfloat32 CNNを用いている。また、バッチ正規化なし、バッチ正規化ありを示している。

【００４７】

図１１（ａ）に示すように、多ビット構成のニューラルネットワーク回路２０では、誤差（Classification error）が低く認識精度は高い。この多ビット構成のニューラルネットワーク回路２０の認識精度を比較対象として、２値化ニューラルネットワーク回路の認識精度を検討する。

図１１（ｂ）の「バッチ正規化なし」に示すように、単に２値化した２値化ニューラルネットワーク回路３０（図４参照）では、誤差率（Classification error）が大きく（約８０％）認識精度は悪い。また、学習を続けても誤差率の改善は見られない（学習が収束しない）。

これに対して、図１１（ｂ）の「バッチ正規化あり」で示される本実施形態の２値化ニューラルネットワーク回路１００は、多ビット構成のニューラルネットワーク回路２０と比較して約６％の誤差（VGG‐16を使用）に収まることが確認された。ただし、ニューロン数は同じ場合であるのでニューロン数を増やすとその差は縮まる。また、本実施形態の２値化ニューラルネットワーク回路１００は、多ビット構成のニューラルネットワーク回路２０と同様に、学習を続けるに従って収束していくことが確認された。

【００４８】

本実施形態では、２値化ニューラルネットワーク回路４０（図６参照）で必須であったバッチ正規化回路４１（図６参照）自体を不要であり、それらのパラメータも不要であることから面積が削減でき、メモリ量も削減できる。また、図１１（ａ）の「バッチ正規化あり」と図１１（ｂ）の「バッチ正規化あり」とを比較してわかるように、本実施形態の２値化ニューラルネットワーク回路１００は、認識精度について、多ビット構成のニューラルネットワーク回路２０（図２参照）と数％異なるだけであった。

【００４９】

図１２は、本実施形態の２値化ニューラルネットワーク回路１００をFPGA(Digilent 社 NetFPGA-1G-CML)上に実装し、既存の多ビット実装法との比較を行った結果を表にして示す図である。

図１２の表は、表下欄外に表記した[1]～[4]の学会発表者（論文発表年）のニューラルネットワークと本実施形態のニューラルネットワークをFPGA 上に実現した場合に、各項目を対比して示したものである。「Platform」(プラットフォーム)、「Clock（MHz）」(同期化のための内部クロック)、「Bandwidth（GB/s）」(データ転送のバンド幅／外部にメモリを付けた場合の転送速度)、「Quantization Strategy」(量子化ビット数)、「Power（W）」(消費電力)、「Performance（GOP/s）」(チップ面積に対する性能)、「Resource Efficiency（GOP/s/Slices）」(リソース効率)、および「Power Efficiency（GOP/s/W）」(性能パワー効率)の各項目を対比して示した。この表において、特に注目すべき事項は下記の通りである。

【００５０】

<消費電力>

本実施形態の２値化ニューラルネットワーク回路１００は、表の従来例と比較して、電力のバランスが取れていることが挙げられる。従来例では、「Power（W）」に示すように、消費電力が大きい。消費電力が大きいので、これを回避する制御方法が複雑である。「Power（W）」に示すように、本実施形態では、従来例と比較して消費電力を１／２～１／３に低減することができた。

【００５１】

<チップ面積>

本実施形態の２値化ニューラルネットワーク回路１００は、バッチ正規化回路がなくメモリが不要であること、乗算回路が２値論理ゲートであること、活性化関数が単純であること（活性化関数回路ではなく活性化関数回路を模擬する活性化回路１２０であること）、から、表の「Performance（GOP/s）」に示すように、チップ面積に対する性能は、従来

例と比較して約３０倍となる。すなわち、チップ面積が減る、外付けメモリが不要となる、メモリコントローラおよび活性化関数が単純になることなどの効果がある。チップ面積は価格に比例するので、価格も２桁程度安くなることが期待できる。

【００５２】

<性能等価>

本実施形態の２値化ニューラルネットワーク回路１００は、表の「Bandwidth（GB/s）」に示すように、従来例と比較してほぼ同等である。また、表の「Power（W）」に示すように、性能パワー効率は、面積を見ずにパワー効率だけを見たものでも約２倍となっている。さらに、表の「Power Efficiency（GOP/s/W）」に示すように、単位ワット数当たりの処理能力（基板全体のワット数）も約２倍となっている。

【００５３】

［実装例］

図１３は、本発明の実施形態に係る２値化ニューラルネットワーク回路の実装例を説明する図である。

<STEP1>

まず、与えられたデータセット（今回はImageNet、画像認識タスク用にデータ）を既存のディープニューラルネットワーク用のフレームワークソフトウェアであるChainer（登録商標）を用いてＣＰＵ（Central Processing Unit）１０１を有するコンピュータ上で学習を行った。このコンピュータは、ＡＲＭプロセッサなどのＣＰＵ１０１と、メモリと、ハードディスクなどの記憶手段（記憶部）と、ネットワークインタフェースを含むＩ／Ｏポートとを有する。このコンピュータは、ＣＰＵ１０１が、メモリ上に読み込んだプログラム（２値化したニューラルネットワークの実行プログラム）を実行することにより、後記する各処理部により構成される制御部（制御手段）を動作させる。

【００５４】

<STEP2>

次に、自動生成ツールを用いて、本実施形態の２値化ニューラルネットワーク回路１００と等価なC++コードを自動生成し、C++コード１０２を得た。

<STEP3>

次に、FPGA ベンダの高位合成ツール(Xilinx 社SDSoC)（登録商標）を用いて、ＦＰＧＡ（field-programmable gate array）合成用にＨＤＬ（hardware description language）を生成した。

<STEP4>

次に、従来のＦＰＧＡ合成ツールVivado（登録商標）を用いて、ＦＰＧＡ上に実現して画像認識タスクの検証を行った。

<STEP5>

検証後、基板１０３を完成させた。基板１０３には、２値化ニューラルネットワーク回路１００がハードウェア化されて実装されている。

【００５５】

以上説明したように、本実施形態に係る２値化ニューラルネットワーク回路１００（図９参照）は、入力値ｘ１～ｘｎ（ｘｉ）（２値）を入力する入力ノードおよび重みｗ１～ｗｎ（ｗｉ）（２値）を入力する入力部１０１と、入力値ｘ１～ｘｎおよび重みｗ１～ｗｎを受け取り、ＸＮＯＲ論理を取るＸＮＯＲゲート回路１０２と、多ビットバイアスW'（式（３）参照）を入力する多ビットバイアスW'入力部１１０と、各ＸＮＯＲ論理値と多ビットバイアスW'との総和を取る総和回路１０３と、総和を取った信号Ｙに対して符号ビットのみを出力する活性化回路１２０と、を備える。

【００５６】

この構成により、バッチ正規化回路自体が不要であり、それらのパラメータも不要であることから面積が削減でき、メモリ量も削減できる。また、本実施形態では、バッチ正規化回路がないにも拘わらず、バッチ正規化回路４１を備える２値化ニューラルネットワーク回路４０（図６参照）と性能的には等価な回路構成となっている。このように、バッチ

正規化回路の面積とパラメータを格納するメモリ面積・メモリ帯域を無くすことができ、かつ、性能的には等価な回路構成を実現することができる。例えば、図１２の表に示すように、本実施形態に係る２値化ニューラルネットワーク回路１００は、消費電力を半分に削減でき、面積を約３０分の１に削減できた。

【００５７】

本実施形態によれば、既存のバッチ正規化回路を備える２値化ニューラルネットワーク回路と比較して、面積を約３０分の１に削減しつつ、認識精度はほぼ等価なＣＮＮを構成できることが判明した。ディープラーニングを用いたＡＤＡＳ（Advanced Driver Assistance System：先進運転支援システム）カメラ画像認識用のエッジ組み込み装置ハードウェア方式として実用化が期待される。特にＡＤＡＳでは、車載する上で高信頼性と低発熱が要求される。本実施形態に係る２値化ニューラルネットワーク回路１００は、図１２の表に示すように、消費電力が格段に低減していることに加え、外付けメモリが不要であるので、メモリを冷却する冷却ファンや冷却フィンも不要である。ＡＤＡＳカメラに搭載して好適である。

【００５８】

［変形例］

図１４は、変形例のディープニューラルネットワークの２値化ニューラルネットワーク回路の構成を示す図である。図９と同一構成部分には同一符号を付している。

本変形例は、乗算回路としての論理ゲートに代えて、ＬＵＴ（Look-Up Table）を用いた例である。

２値化ニューラルネットワーク回路２００は、図１のニューラルネットワーク回路２に適用できる。

図１４に示すように、２値化ニューラルネットワーク回路２００（ニューラルネットワーク回路装置）は、入力値ｘ１～ｘｎ（ｘｉ）（２値）を入力する入力ノードｘ１～ｘｎおよび重みｗ１～ｗｎ（２値）を入力する入力部１０１と、入力値ｘ１～ｘｎおよび重みｗ１～ｗｎを受け取り、２値（－１／＋１）の乗算を行うためのテーブル値を格納し演算時に参照されるＬＵＴ２０２（論理回路部）と、多ビットバイアスW'（式（３）参照）を入力給する多ビットバイアスW'入力手段１１０と、ＬＵＴ２０２から参照した各テーブル値と多ビットバイアスW'との総和を取る総和回路１０３と、総和を取った信号Ｙに対して符号ビットのみを出力する活性化関数回路を模擬する活性化回路１２０と、を備えて構成される。

【００５９】

本変形例では、乗算回路としての論理ゲートに代えて、ＬＵＴ（Look-Up Table）を用いた例である。

ＬＵＴ２０２は、ＸＮＯＲ論理を行うＸＮＯＲゲート回路１０２（図９参照）に代えて、FPGA の基本構成要素であるルックアップテーブルを用いる。

【００６０】

図１５は、変形例の２値化ニューラルネットワーク回路２００のＬＵＴ２０２の構成を示す図である。

図１５に示すように、ＬＵＴ２０２は、２入力（ｘ１，ｗ１）に対する２値（－１／＋１）のＸＮＯＲ論理結果Ｙを格納する。

【００６１】

このように、変形例の２値化ニューラルネットワーク回路２００は、図９のＸＮＯＲゲート回路１０２を、ＬＵＴ２０２に置き換えた構成となっている。変形例では、実施形態と同様に、バッチ正規化回路の面積とパラメータを格納するメモリ面積・メモリ帯域を無くすことができ、かつ、性能的には等価な回路構成を実現することができる。

また、本変形例では、ＸＮＯＲ演算を行う論理ゲートとして、ＬＵＴ２０２を用いている。ＬＵＴ２０２は、ＦＰＧＡの基本構成要素であり、ＦＰＧＡ合成の際の親和性が高く、ＦＰＧＡによる実装が容易である。

【００６２】

本発明は上記の実施形態例に限定されるものではなく、特許請求の範囲に記載した本発明の要旨を逸脱しない限りにおいて、他の変形例、応用例を含む。

また、上記した実施形態例は本発明をわかりやすく説明するために詳細に説明したものであり、必ずしも説明した全ての構成を備えるものに限定されるものではない。また、ある実施形態例の構成の一部を他の実施形態例の構成に置き換えることが可能であり、また、ある実施形態例の構成に他の実施形態例の構成を加えることも可能である。また、実施形態例は、その他の様々な形態で実施されることが可能であり、発明の要旨を逸脱しない範囲で、種々の省略、置き換え、変更を行うことができる。これら実施形態やその変形例は、発明の範囲や要旨に含まれるとともに、特許請求の範囲に記載された発明とその均等の範囲に含まれる。

【００６３】

また、上記実施形態において説明した各処理のうち、自動的に行われるものとして説明した処理の全部または一部を手動的に行うこともでき、あるいは、手動的に行われるものとして説明した処理の全部または一部を公知の方法で自動的に行うこともできる。この他、上述文書中や図面中に示した処理手順、制御手順、具体的名称、各種のデータやパラメータを含む情報については、特記する場合を除いて任意に変更することができる。

また、図示した各装置の各構成要素は機能概念的なものであり、必ずしも物理的に図示の如く構成されていることを要しない。すなわち、各装置の分散・統合の具体的形態は図示のものに限られず、その全部または一部を、各種の負荷や使用状況などに応じて、任意の単位で機能的または物理的に分散・統合して構成することができる。

【００６４】

また、上記の各構成、機能、処理部、処理手段等は、それらの一部または全部を、例えば集積回路で設計する等によりハードウェアで実現してもよい。また、上記の各構成、機能等は、プロセッサがそれぞれの機能を実現するプログラムを解釈し、実行するためのソフトウェアで実現してもよい。各機能を実現するプログラム、テーブル、ファイル等の情報は、メモリや、ハードディスク、ＳＳＤ（Solid State Drive）等の記録装置、または、ＩＣ（Integrated Circuit）カード、ＳＤ（Secure Digital）カード、光ディスク等の記録媒体に保持することができる。

また、上記実施の形態では、装置は、ニューラルネットワーク回路装置という名称を用いたが、これは説明の便宜上であり、名称はディープニューラルネットワーク回路、ニューラルネットワーク装置、パーセプトロン等であってもよい。また、方法およびプログラムは、ニューラルネットワーク処理方法という名称を用いたが、ニューラルネットワーク演算方法、ニューラルネットプログラム等であってもよい。

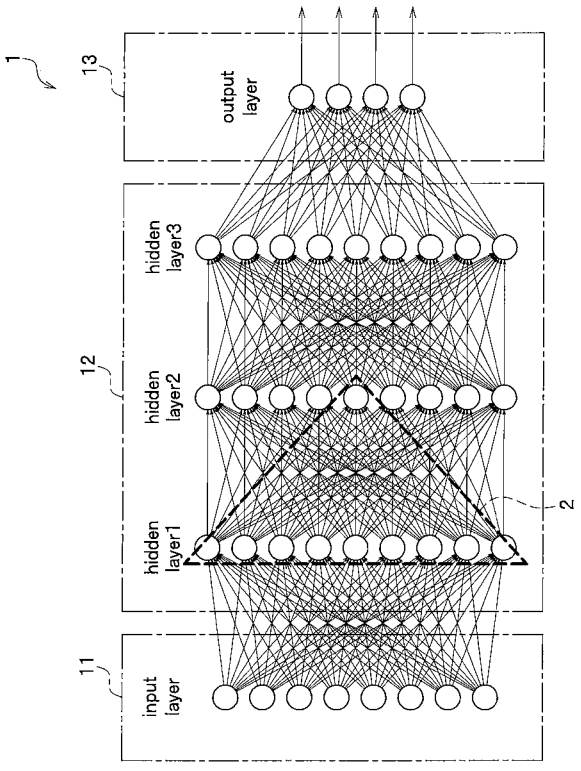
【符号の説明】

【００６５】

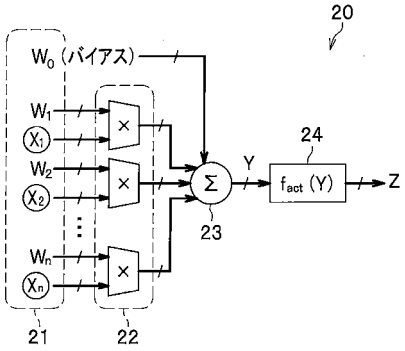
１　ディープニューラルネットワーク
２　ニューラルネットワーク回路
１１　入力層
１２　隠れ層（中間層）
１３　出力層
１００，２００　２値化ニューラルネットワーク回路（ニューラルネットワーク回路装置）
１０１　入力部
１０２　ＸＮＯＲゲート回路（論理回路部，論理回路手段）
１０３　総和回路（総和回路部，総和回路手段）
１１０　多ビットバイアス入力部
１２０　活性化回路（活性化回路部，活性化回路手段）
２０２　ＬＵＴ（論理回路部）
ｘ１～ｘｎ（ｘｉ）　入力値（２値）
ｗ１～ｗｎ（ｗｉ）　重み（２値）

W' 多ビットバイアス

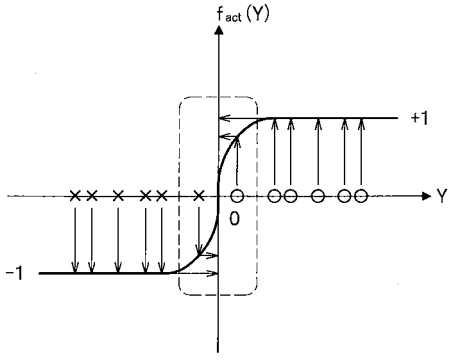
【図１】



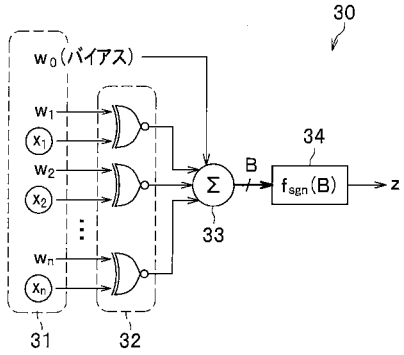
【図２】



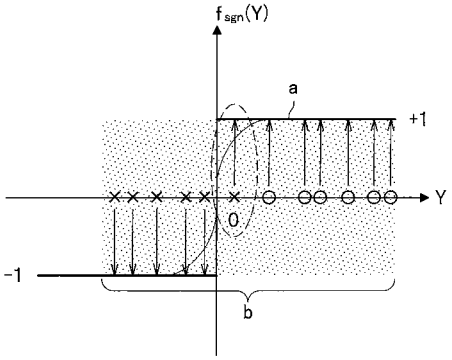
【図３】



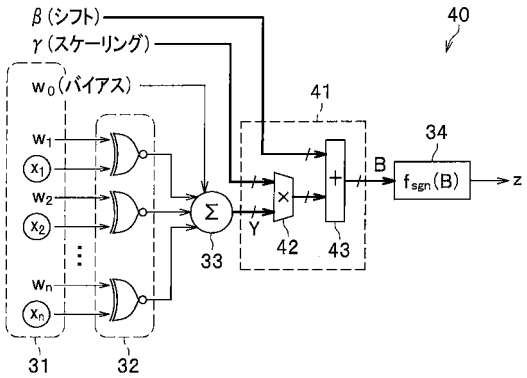
【図４】



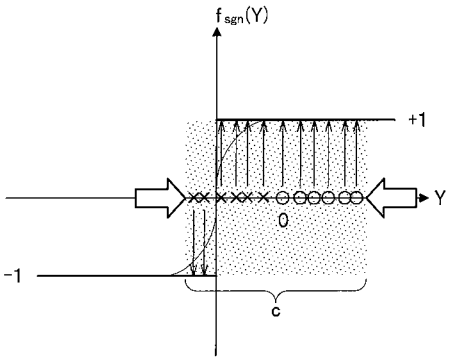
【図５】



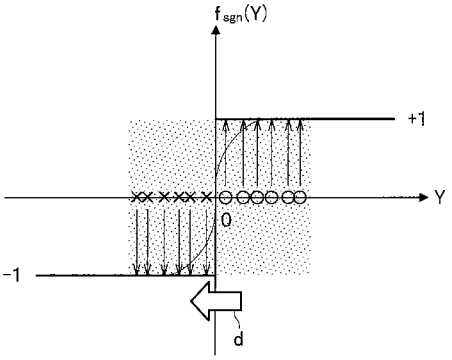
【図６】



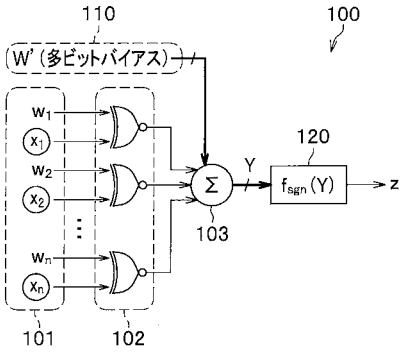
【図７】



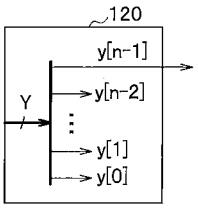
【図８】



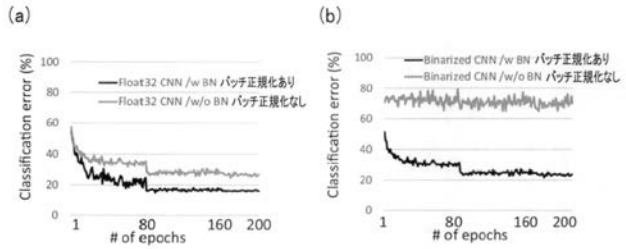
【図９】



【図１０】



【図１１】

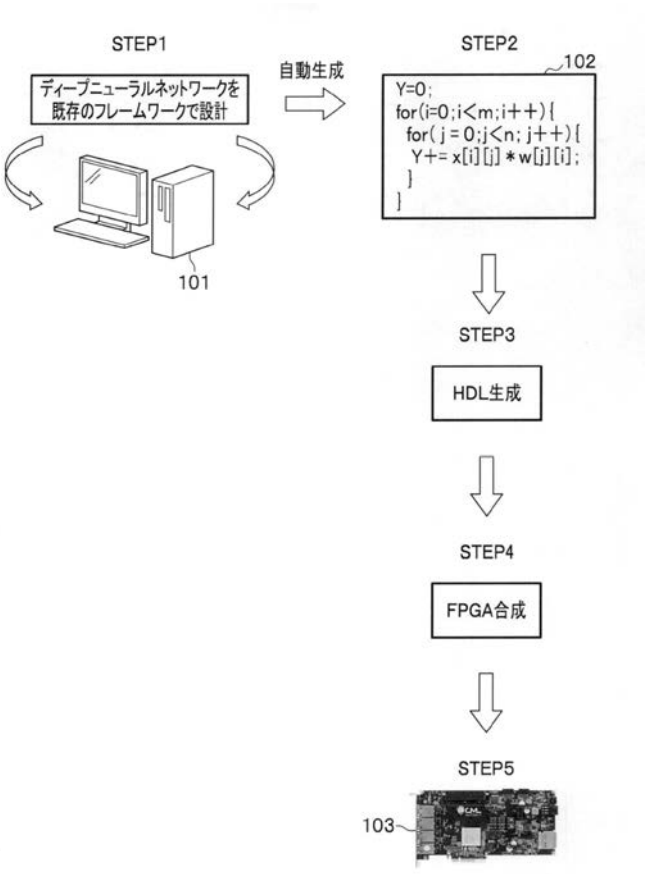


【図１２】

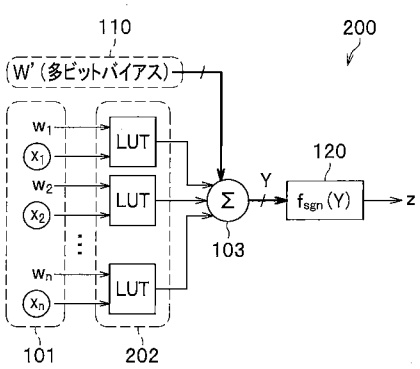
Year	2010 [1]	2014 [2]	2015 [3]	2016 [4]	提案手法
Platform	Virtex 5 SX240t	Zynq XC7Z045	Virtex7 VX485t	Zynq XC7Z045	Kintex7 325t
Clock(MHz)	120	150	100	150	450
Bandwidth(GB/s)	---	4.2	12.8	4.2	3.6
Quantization Strategy	48-bit fixed	16-bit fixed	32-bit float	16-bit fixed	binary (1bit)
Power(W)	14	8	18.61	9.63	**4.48**
Performance (GOP/s)	16	23.18	61.62	**187.80**	176.8
Resource Efficiency (GOP/s/Slices)	4.30×10^{-4}	---	8.12×10^{-4}	35.8×10^{-4}	**1389.9×10^{-4}**
Power Efficiency (GOP/s/W)	1.14	2.90	3.31	19.50	**39.46**

[1] S.Chakradhar et.al, "A dynamically configurable coprocessor for convolutional neural networks," in ISCA.
[2] V.Gokhale et.al, "A 240 g-ops/s mobile coprocessor for deep neural networks," in CVPRW.
[3] C. Zhang et.al, "Optimizing fpga-based accelerator design for deep convolutional neural networks," in FPGA.
[4] J. Qiu et. al, "Going deeper with embedded FPGA platform for convolutional neural network," in FPGA.

【図１３】



【図１４】



【図１５】

LUT 202

x1	w1	Y
−1	−1	1
−1	+1	−1
+1	−1	−1
+1	+1	1

【手続補正書】
【提出日】平成29年6月23日(2017.6.23)
【手続補正１】
【補正対象書類名】特許請求の範囲
【補正対象項目名】全文
【補正方法】変更
【補正の内容】
【特許請求の範囲】
【請求項１】
入力層、１以上の中間層、および、出力層を少なくとも含むニューラルネットワーク回路装置であって、
前記中間層の中で、入力値ｘｉおよび重みｗｉを受け取り、論理演算を行う論理回路部と、
多ビットバイアスＷ’を受け取り、前記論理回路部の出力と前記多ビットバイアスＷ’との総和を取る総和回路部と、
総和を取った多ビット信号Ｙに対して符号ビットのみを出力する活性化回路部と、を備え、
前記多ビット信号Ｙおよび前記多ビットバイアスＷ’は、下記式で示される
【数３】

$$Y = \sum_{i=0}^{n} w_i x_i - \mu_B + \frac{\sqrt{\sigma_B^2 + \epsilon}}{\gamma}\beta$$

$$= \sum_{i=1}^{n} w_i x_i + \left(w_0 - \mu_B + \frac{\sqrt{\sigma_B^2 + \epsilon}}{\gamma}\beta \right) \quad \cdots(3)$$

$$= \sum_{i=1}^{n} w_i x_i + W' \cdot$$

ことを特徴とする記載のニューラルネットワーク回路装置。
【請求項２】
前記入力値ｘｉおよび前記重みｗｉを入力する入力部と、
前記多ビットバイアスＷ’を入力する多ビットバイアス入力部と、を備える
ことを特徴とする請求項１に記載のニューラルネットワーク回路装置。
【請求項３】
前記入力値ｘｉおよび前記重みｗｉは、２値信号である
ことを特徴とする請求項１または請求項２に記載のニューラルネットワーク回路装置。
【請求項４】
前記多ビットバイアスＷ’は、学習後の多ビットバイアス値である
ことを特徴とする請求項１または請求項２に記載のニューラルネットワーク回路装置。
【請求項５】
前記論理回路部は、否定排他的論理和または排他的論理和を含む
ことを特徴とする請求項１に記載のニューラルネットワーク回路装置。
【請求項６】
前記論理回路部は、ＬＵＴ（Look-Up Table）である
ことを特徴とする請求項１に記載のニューラルネットワーク回路装置。
【請求項７】

前記符号ビットは、総和を取った前記多ビット信号Ｙを活性化するかしないかで示す２値信号である

ことを特徴とする請求項１に記載のニューラルネットワーク回路装置。

【請求項８】

前記多ビット信号Ｙを、正規化範囲を広げ中心をシフトさせるバッチ正規化処理を行い出力される信号Ｙ’が、式（１）で示される場合、

【数１】

$$Y' = \gamma \frac{Y - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} + \beta$$

$$= \frac{\gamma}{\sqrt{\sigma_B^2+\epsilon}} \left(Y - \left(\mu_B - \frac{\sqrt{\sigma_B^2+\epsilon}}{\gamma} \beta \right) \right) \cdot \quad \cdots (1)$$

前記多ビットバイアスＷ’は、

前記バッチ正規化処理による前記信号Ｙ’を含まない式（３）で示される前記多ビット信号Ｙで与えられる

【数３】

$$Y = \sum_{i=0}^{n} w_i x_i - \mu_B + \frac{\sqrt{\sigma_B^2 + \epsilon}}{\gamma} \beta \quad \cdots (3)$$

$$= \sum_{i=1}^{n} w_i x_i + \left(w_0 - \mu_B + \frac{\sqrt{\sigma_B^2 + \epsilon}}{\gamma} \beta \right)$$

$$= \sum_{i=1}^{n} w_i x_i + W' \cdot$$

ことを特徴とする請求項１に記載のニューラルネットワーク回路装置。

【請求項９】

請求項１乃至８のいずれか１項に記載のニューラルネットワーク回路装置を備えるニューラルネットワーク。

【請求項１０】

入力層、１以上の中間層、および、出力層を少なくとも含むニューラルネットワーク処理方法であって、

前記中間層の中で、入力値ｘｉおよび重みｗｉを受け取り、論理演算を行うステップと、

多ビットバイアスＷ’を受け取り、前記論理演算ステップの出力と前記多ビットバイアスＷ’との総和を取るステップと、

総和を取った多ビット信号Ｙに対して符号ビットのみを出力するステップと、を有し、

前記多ビット信号Ｙおよび前記多ビットバイアスＷ’は、下記式で示される

【数３】

$$Y = \sum_{i=0}^{n} w_i x_i - \mu_B + \frac{\sqrt{\sigma_B^2 + \epsilon}}{\gamma}\beta$$

$$= \sum_{i=1}^{n} w_i x_i + \left(w_0 - \mu_B + \frac{\sqrt{\sigma_B^2 + \epsilon}}{\gamma}\beta \right) \quad \cdots(3)$$

$$= \sum_{i=1}^{n} w_i x_i + W' \cdot$$

ことを特徴とするニューラルネットワーク処理方法。

【請求項１１】

入力層、１以上の中間層、および、出力層を少なくとも含むニューラルネットワーク回路装置としてのコンピュータを、

前記中間層の中で、入力値ｘｉおよび重みｗｉを受け取り、論理演算を行う論理回路手段、

多ビットバイアスＷ’を受け取り、前記論理手段の出力と前記多ビットバイアスＷ’との総和を取る総和回路手段、

総和を取った多ビット信号Ｙに対して符号ビットのみを出力する活性化回路手段、

ただし、前記多ビット信号Ｙおよび前記多ビットバイアスＷ’は、下記式で示される、

【数３】

$$Y = \sum_{i=0}^{n} w_i x_i - \mu_B + \frac{\sqrt{\sigma_B^2 + \epsilon}}{\gamma}\beta$$

$$= \sum_{i=1}^{n} w_i x_i + \left(w_0 - \mu_B + \frac{\sqrt{\sigma_B^2 + \epsilon}}{\gamma}\beta \right) \quad \cdots(3)$$

$$= \sum_{i=1}^{n} w_i x_i + W' \cdot$$

として機能させるためのニューラルネットワークの実行プログラム。

【手続補正２】

【補正対象書類名】明細書

【補正対象項目名】０００９

【補正方法】変更

【補正の内容】

【０００９】

前記した課題を解決するため、本発明に係るニューラルネットワーク回路装置は、入力層、１以上の中間層、および、出力層を少なくとも含むニューラルネットワーク回路装置であって、前記中間層の中で、入力値ｘｉおよび重みｗｉを受け取り、論理演算を行う論理回路部と、多ビットバイアスＷ’を受け取り、前記論理回路部の出力と前記多ビットバイアスＷ’との総和を取る総和回路部と、総和を取った多ビット信号Ｙに対して符号ビッ

トのみを出力する活性化回路部と、を備え、前記多ビット信号Ｙおよび前記多ビットバイアスＷ’は、下記式で示される

【数３】

$$Y = \sum_{i=0}^{n} w_i x_i - \mu_B + \frac{\sqrt{\sigma_B^2 + \epsilon}}{\gamma}\beta$$

$$= \sum_{i=1}^{n} w_i x_i + \left(w_0 - \mu_B + \frac{\sqrt{\sigma_B^2 + \epsilon}}{\gamma}\beta \right) \quad \cdots(3)$$

$$= \sum_{i=1}^{n} w_i x_i + W' \cdot$$

ことを特徴とする。