

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B1)

(11) 特許番号

特許第6183980号
(P6183980)

(45) 発行日 平成29年8月23日(2017.8.23)

(24) 登録日 平成29年8月4日(2017.8.4)

(51) Int.Cl. F I
G O 6 N 3/063 (2006.01) G O 6 N 3/063

請求項の数 11 (全 22 頁)

<p>(21) 出願番号 特願2016-235383 (P2016-235383)</p> <p>(22) 出願日 平成28年12月2日 (2016.12.2)</p> <p>審査請求日 平成29年3月24日 (2017.3.24)</p> <p>早期審査対象出願</p>	<p>(73) 特許権者 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</p> <p>(74) 代理人 110001807 特許業務法人磯野国際特許商標事務所</p> <p>(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</p> <p>(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</p> <p>審査官 石川 亮</p>
--	---

最終頁に続く

(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) 【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 x_i および重み w_i 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】

$$\begin{aligned}
 Y &= \sum_{i=0}^n w_i X_i - \mu_B + \frac{\sqrt{\sigma_B^2 + \epsilon}}{\gamma} \beta \\
 &= \sum_{i=1}^n w_i X_i + \left(w_0 - \mu_B + \frac{\sqrt{\sigma_B^2 + \epsilon}}{\gamma} \beta \right) \\
 &= \sum_{i=1}^n w_i X_i + W' .
 \end{aligned}
 \tag{3}$$

10

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 x_i および前記重み w_i 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 x_i および前記重み w_i 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。

20

【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。

30

【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】

$$\begin{aligned}
 Y' &= \gamma \frac{Y - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} + \beta \\
 &= \frac{\gamma}{\sqrt{\sigma_B^2 + \epsilon}} \left(Y - \left(\mu_B - \frac{\sqrt{\sigma_B^2 + \epsilon}}{\gamma} \beta \right) \right) .
 \end{aligned}
 \tag{1}$$

40

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】

$$\begin{aligned}
 Y &= \sum_{i=0}^n w_i X_i - \mu_B + \frac{\sqrt{\sigma_B^2 + \epsilon}}{\gamma} \beta \\
 &= \sum_{i=1}^n w_i X_i + \left(w_0 - \mu_B + \frac{\sqrt{\sigma_B^2 + \epsilon}}{\gamma} \beta \right) \\
 &= \sum_{i=1}^n w_i X_i + W' .
 \end{aligned}
 \tag{3}$$

10

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、

20

前記中間層の中で、入力値 x_i および重み w_i 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】

$$\begin{aligned}
 Y &= \sum_{i=0}^n w_i X_i - \mu_B + \frac{\sqrt{\sigma_B^2 + \epsilon}}{\gamma} \beta \\
 &= \sum_{i=1}^n w_i X_i + \left(w_0 - \mu_B + \frac{\sqrt{\sigma_B^2 + \epsilon}}{\gamma} \beta \right) \\
 &= \sum_{i=1}^n w_i X_i + W' .
 \end{aligned}
 \tag{3}$$

30

ことを特徴とするニューラルネットワーク処理方法。

40

【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 x_i および重み w_i 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】

$$\begin{aligned}
 Y &= \sum_{i=0}^n w_i X_i - \mu_B + \frac{\sqrt{\sigma_B^2 + \epsilon}}{\gamma} \beta \\
 &= \sum_{i=1}^n w_i X_i + \left(w_0 - \mu_B + \frac{\sqrt{\sigma_B^2 + \epsilon}}{\gamma} \beta \right) \\
 &= \sum_{i=1}^n w_i X_i + W' \cdot
 \end{aligned}
 \tag{3}$$

10

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】

20

【0002】

古典的な順伝搬型ニューラルネットワーク（F F N N : Feedforward Neural Network）、R B F（Radial Basis Function）ネットワーク、正規化したR B Fネットワーク、自己組織化マップなどがある。R B F N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、A D A S（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（C N N : Convolutional Neural Network）（層間が全結合でないN N）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。C N Nは、ディープニューラルネットワーク（D N N : Deep Neural Network）に畳込み演算を付加したものである。

30

【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C N N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・消費電力が多くなる欠点があった。そこで、2値化した精度、すなわち+1と-1のみ用いてC N N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）。

40

【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N N

50

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回路装置であって、前記中間層の中で、入力値 x_i および重み w_i 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】

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

・・・(3)

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

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

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 $f_{act}(Y)$ を示す図である。 10

【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 $f_{sgn}(B)$ 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()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()による $-1 \sim +1$ の制限を示す図である。 20

【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 30

【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。 40

(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、 50

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 1 1 と隠れ層 1 2 のノード間が全て結合し、隠れ層 1 2 と出力層 1 3 のノード間が全て結合している。

【 0 0 1 3 】

入力層 1 1、隠れ層 1 2 および出力層 1 3 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 1 1 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ノ境界を決定できる。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 1 3 から伝播させていき、各層の重みを調整する。

10

【 0 0 1 4 】

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。

20

【 0 0 1 5 】

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 1 2 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 1 1 または出力層 1 3 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 1 3 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。

30

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 w_i は既に得られている。

【 0 0 1 6 】

<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 2 0 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。

40

ニューラルネットワーク回路 2 0 は、入力値（判別データ） $X_1 \sim X_n$ （多ビット）を入力する入力ノード、重み $W_1 \sim W_n$ （多ビット）およびバイアス W_0 （多ビット）を入力する入力部 2 1 と、入力値 $X_1 \sim X_n$ および重み $W_1 \sim W_n$ を受け取り、入力値 $X_1 \sim X_n$ に重み $W_1 \sim W_n$ をそれぞれ乗算する複数の乗算回路 2 2 と、各乗算値とバイアス W_0 との総和を取る総和回路 2 3 と、総和を取った信号 Y を活性化関数 $f_{act}(Y)$ で変換する活性化関数回路 2 4 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 2 0 は、入力値 $X_1 \sim X_n$ （多ビッ

50

ト)を受け取り、重み $W_1 \sim W_n$ を乗算した後に、バイアス W_0 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 $f_{act}(Y)$ 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 $f_{act}(Y)$ 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 \times 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 $f_{act}(Y)$ 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 $f_{act}(Y)$ 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・消費電力が多くなる欠点があった。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 $x_1 \sim x_n$ (2値)を入力する入力ノード、重み $w_1 \sim w_n$ (2値)およびバイアス w_0 (2値)を入力する入力部31と、入力値 $x_1 \sim x_n$ および重み $w_1 \sim w_n$ 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 w_0 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 $f_{sgn}(B)$ 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 $x_1 \sim x_n$ 、出力値 z 、および重み $w_1 \sim w_n$ 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 $f_{sgn}(B)$ 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 $f_{sgn}(B)$ 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 \times 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 $x_1 \sim x_n$ および重み $w_1 \sim w_n$ 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

10

20

30

40

50

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ぬ。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 $x_1 \sim x_n$ （2値）を入力する入力ノード $x_1 \sim x_n$ 、重み $w_1 \sim w_n$ （2値）およびバイアス w_0 （2値）を入力する入力部31と、入力値 $x_1 \sim x_n$ および重み $w_1 \sim w_n$ 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 w_0 （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 $f_{\text{sgn}}(B)$ 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（ ）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（ ）値による正規化後、シフト（ ）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（ ）値およびシフト（ ）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 $x_1 \sim x_n$ と重み $w_1 \sim w_n$ 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（ ）値およびシフト（ ）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（ ）による正規化を示す図、図8は、シフト（ ）による $-1 \sim +1$ 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（ ）値による正規化を行った後、シフト（ ）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（ ）値により、幅「2」（

10

20

30

40

50

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()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()値による正規化後の値を、シフト()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()値により、スケーリング()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。

10

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()値およびシフト()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。

20

【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。

30

【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 $f_{act}(Y)$ 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。

40

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】

$$\begin{aligned}
 Y' &= \gamma \frac{Y - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} + \beta \\
 &= \frac{\gamma}{\sqrt{\sigma_B^2 + \epsilon}} \left(Y - \left(\mu_B - \frac{\sqrt{\sigma_B^2 + \epsilon}}{\gamma} \beta \right) \right) \dots (1)
 \end{aligned}$$

ただし、

: スケーリング値

: シフト値

 μ_B : 平均値 σ_B^2 : 二乗和誤差

: パラメータ (調整用)

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

【0031】

【数 2】

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

【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】

$$\begin{aligned}
 Y &= \sum_{i=0}^n w_i x_i - \mu_B + \frac{\sqrt{\sigma_B^2 + \epsilon}}{\gamma} \beta \dots (3) \\
 &= \sum_{i=1}^n w_i x_i + \left(w_0 - \mu_B + \frac{\sqrt{\sigma_B^2 + \epsilon}}{\gamma} \beta \right) \\
 &= \sum_{i=1}^n w_i x_i + W' \cdot
 \end{aligned}$$

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】

10

20

30

40

50

[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 $x_1 \sim x_n$ (x_i)（2値）を入力する入力ノードおよび重み $w_1 \sim w_n$ (w_i)（2値）を入力する入力部101と、入力値 $x_1 \sim x_n$ および重み $w_1 \sim w_n$ 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 x_i （2値）および重み w_i （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 $W_1 \sim W_n$ およびバイアス W_0 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 W_0 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 $x_1 \sim x_n$ と重み $w_1 \sim w_n$ 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、符号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

10

20

30

40

50

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 $f \text{sgn}(Y)$ と表記されているが、図6に示すスケールリング()による正規化およびシフト()による $-1 \sim +1$ 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 $x_1 \sim x_n$ 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 $x_1 \sim x_n$ (2値)および重み $w_1 \sim w_n$ (2値)が入力される。

XNORゲート回路102では、入力値 $x_1 \sim x_n$ および重み $w_1 \sim w_n$ 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 $x_1 \sim x_n$ および重み $w_1 \sim w_n$ 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 w_0 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 W_0 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 w_0 (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、符号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(符号ビット)だけを入力する。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 x_i と重み w_i 、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

10

20

30

40

50

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 $f \operatorname{sgn}(Y)$ と同様な機能、すなわち活性化関数回路 $f \operatorname{sgn}(Y)$ 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

10

20

30

40

50

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。

10

【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。

20

【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。

30

<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】

40

<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。

50

<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 $x_1 \sim x_n$ (x_i) (2 値) を入力する入力ノードおよび重み $w_1 \sim w_n$ (w_i) (2 値) を入力する入力部 101 と、入力値 $x_1 \sim x_n$ および重み $w_1 \sim w_n$ 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。

10

【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。

20

【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。

30

【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 $x_1 \sim x_n$ (x_i) (2 値) を入力する入力ノード $x_1 \sim x_n$ および重み $w_1 \sim w_n$ (2 値) を入力する入力部 101 と、入力値 $x_1 \sim x_n$ および重み $w_1 \sim w_n$ 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 LUT 202 (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、LUT 202 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。

40

【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。

50

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（ x_1 、 w_1 ）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、ニューラルネットワークプログラム等であってもよい。

【符号の説明】

10

20

30

40

50

【 0 0 6 5 】

- 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
 - 1 1 入力層
 - 1 2 隠れ層 (中間層)
 - 1 3 出力層
- 1 0 0 , 2 0 0 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)
 - 1 0 1 入力部
 - 1 0 2 X N O R ゲート回路 (論理回路部 , 論理回路手段)
 - 1 0 3 総和回路 (総和回路部 , 総和回路手段)
 - 1 1 0 多ビットバイアス入力部
 - 1 2 0 活性化回路 (活性化回路部 , 活性化回路手段)
 - 2 0 2 L U T (論理回路部)
 - $x_1 \sim x_n (x_i)$ 入力値 (2 値)
 - $w_1 \sim w_n (w_i)$ 重み (2 値)
 - W' 多ビットバイアス

10

【要約】

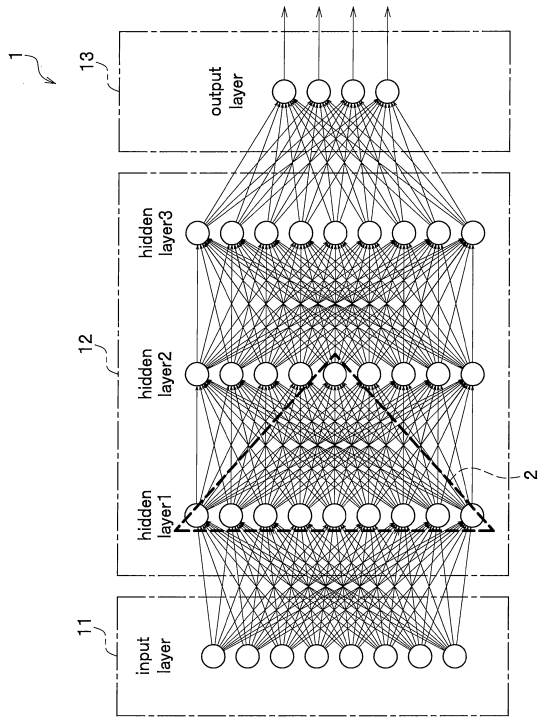
【課題】バッチ正規化回路が不要なニューラルネットワーク回路装置、ニューラルネットワーク、ニューラルネットワーク処理方法およびニューラルネットワークの実行プログラムを提供する。

20

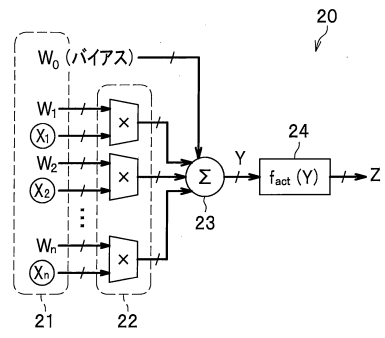
【解決手段】2 値化ニューラルネットワーク回路 1 0 0 は、入力値 $x_1 \sim x_n (x_i)$ (2 値) を入力する入力ノードおよび重み $w_1 \sim w_n (w_i)$ を入力する入力部 1 0 1 と、入力値 $x_1 \sim x_n$ および重み $w_1 \sim w_n$ を受け取り、X N O R 論理を取る X N O R ゲート回路 1 0 2 と、多ビットバイアス W' を入力する多ビットバイアス W' 入力部 1 1 0 と、各 X N O R 論理値と多ビットバイアス W' との総和を取る総和回路 1 0 3 と、総和を取った信号 Y に対して符号ビットのみを出力する活性化回路 1 2 0 と、を備える。

【選択図】図 9

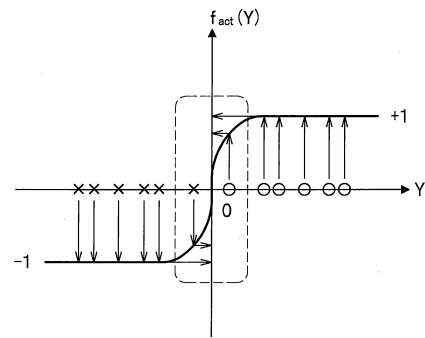
【図1】



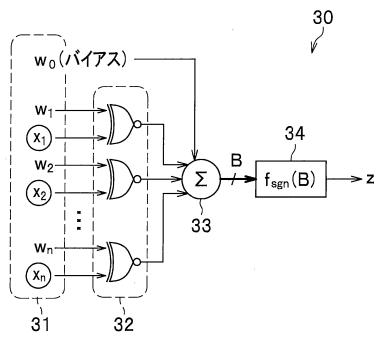
【図2】



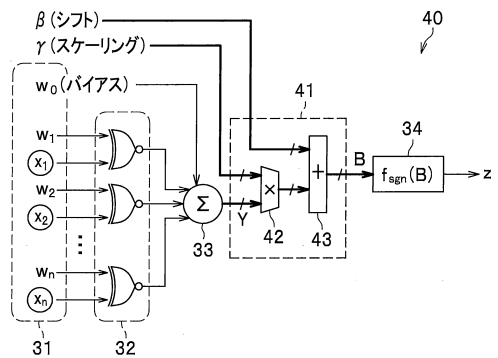
【図3】



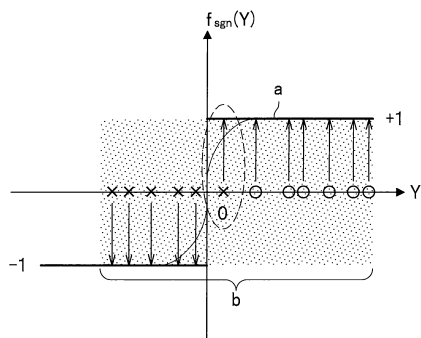
【図4】



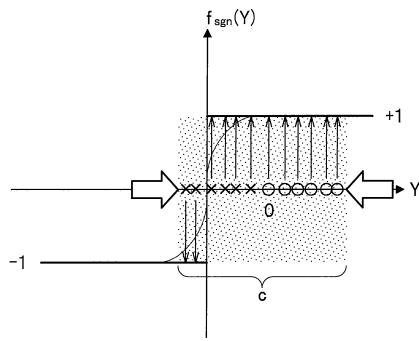
【図6】



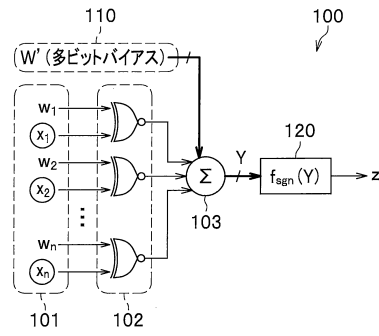
【図5】



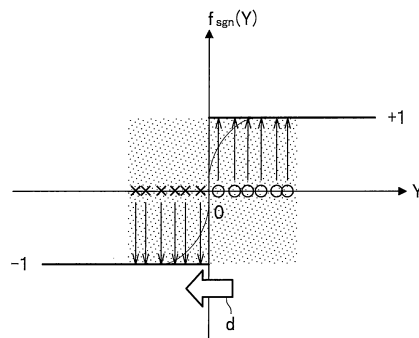
【図7】



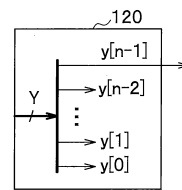
【図9】



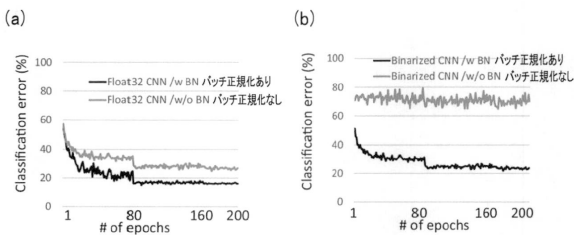
【図8】



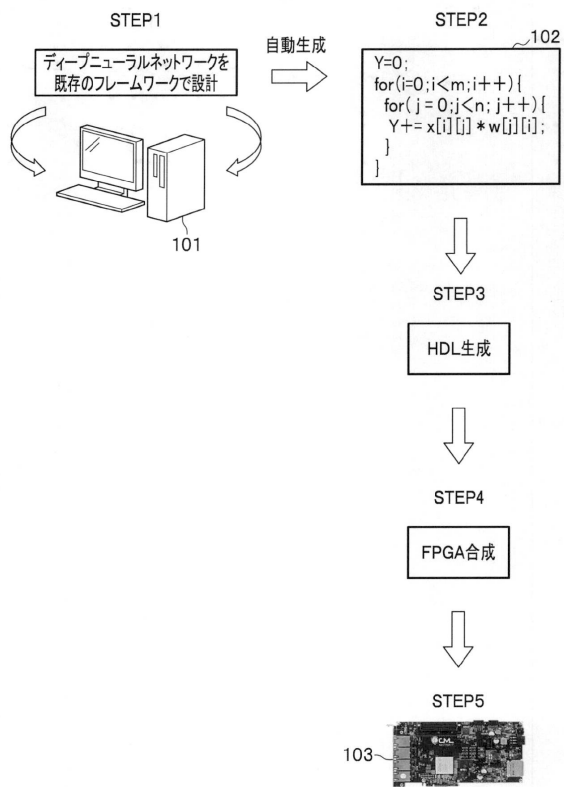
【図10】



【図11】



【図13】



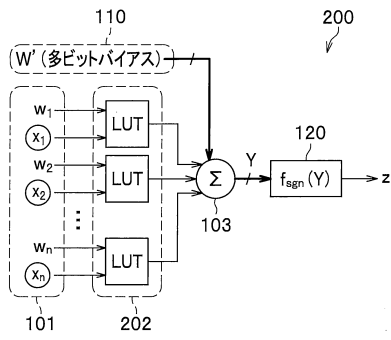
【図12】

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 1 4 】



【 図 1 5 】

LUT		
$x1$	$w1$	Y
-1	-1	1
-1	+1	-1
+1	-1	-1
+1	+1	1

フロントページの続き

(56)参考文献 特開2016-033806(JP,A)

特開2006-033579(JP,A)

中原啓貴, 米川晴義, 笹尾勤, 岩本久, 本村真人, メモリベースに基づく2値化深層畳込みニューラルネットワークの実現, 電子情報通信学会技術研究報告, 日本, 一般社団法人電子情報通信学会, 2016年 8月29日, Vol.116 No.210, pp.63-68, ISSN:0913-5685

(58)調査した分野(Int.Cl., DB名)

G06N 3/02

G06N 3/06 - 3/063