(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7689748号
(P7689748)

(45)発行日　令和7年6月9日(2025.6.9)

(24)登録日　令和7年5月30日(2025.5.30)

(51)国際特許分類　　ＦＩ
Ｇ０６Ｎ　3/047(2023.01)　　Ｇ０６Ｎ　3/047
Ｇ０６Ｎ　3/063(2023.01)　　Ｇ０６Ｎ　3/063

請求項の数　16　（全15頁）

(21)出願番号　特願2022-549079(P2022-549079)
(86)(22)出願日　令和3年2月11日(2021.2.11)
(65)公表番号　特表2023-513596(P2023-513596A)
(43)公表日　令和5年3月31日(2023.3.31)
(86)国際出願番号　PCT/ES2021/070096
(87)国際公開番号　WO2021/160914
(87)国際公開日　令和3年8月19日(2021.8.19)
審査請求日　令和6年1月30日(2024.1.30)
(31)優先権主張番号　P202030120
(32)優先日　令和2年2月13日(2020.2.13)
(33)優先権主張国・地域又は機関
スペイン(ES)

(73)特許権者　506119291
ウニベルシタット デ レス イリュス バレアルス
スペイン国 ０７１２２ パルマ デ マリョルカ、 クトラ．デ バルデモッサ カエメ．７、５
(74)代理人　100120031
弁理士　宮嶋 学
(74)代理人　100107582
弁理士　関根 毅
(74)代理人　100118843
弁理士　赤岡 明
(74)代理人　100202429
弁理士　石原 信人
(72)発明者　ジョゼップ、リュイス、ロセヨ、サンス
最終頁に続く

(54)【発明の名称】　確率信号生成要素、確率ニューロン及びそのニューラルネットワーク

(57)【特許請求の範囲】

【請求項１】

バイナリ信号（A）を受信する、第１入力と、ランダム信号（R）を受信する、第２入力と、を順に有し、前記ランダム信号（R）を用いて前記バイナリ信号（A）を第１確率信号（A*）に変換する、第１バイナリ－確率変換器（BSC）、

を備え、

前記第１確率信号（A*）を受信する、第１入力と、第２確率信号である参照確率信号（C*）を受信する第２入力であって前記参照確率信号（C*）は定数値信号（C）から前記ランダム信号（R）を用いて生成される、第２入力と、を備える、処理ユニット（１１）であって、前記第１確率信号（A*）及び前記参照確率信号（C*）を少なくとも１つの算術関数に従って処理し、前記処理を表す確率出力信号（S*）を生成する、処理ユニット（１１）、

を備え、

前記処理ユニット（１１）は、OR型論理ゲートであり、適用される前記算術関数は、正規化線形ユニット（ReLU）型の活性化関数から構成される、

確率信号生成要素（１）。

【請求項２】

バイナリ信号（A）を受信する、第１入力と、ランダム信号（R）を受信する、第２入力と、を順に有し、前記ランダム信号（R）を用いて前記バイナリ信号（A）を第１確率信号（A*）に変換する、第１バイナリ－確率変換器（BSC）、

を備え、
前記第１確率信号（A*）を受信する、第１入力と、第２確率信号である参照確率信号（C*）を受信する第２入力であって前記参照確率信号（C*）は定数値信号（Ｃ）から前記ランダム信号（Ｒ）を用いて生成される、第２入力と、を備える、処理ユニット（１１）であって、前記第１確率信号（A*）及び前記参照確率信号（C*）を少なくとも１つの算術関数に従って処理し、前記処理を表す確率出力信号（S*）を生成する、処理ユニット（１１）、
を備え、
前記処理ユニット（１１）は、AND型論理ゲートであり、適用される前記算術関数は、最小値型の活性化関数（A*、C*）から構成される、
確率信号生成要素（１）。
【請求項３】
請求項１又は請求項２に記載された確率信号生成要素（１）を備える、
計算ニューラルネットワーク用の確率ニューロン（１０）。
【請求項４】
複数の確率の入力信号（Y1* - Yn*）を受信する近似的並列カウンタ（APC）を備え、前記複数の入力信号（Y1* - Yn*）の和をとり、その出力において２の補数に符号化されたバイナリの出力信号に変換し、
前記出力信号は、前記第１バイナリ－確率変換器（BSC）に入力された前記バイナリ信号（A）である、
請求項３に記載の計算ニューラルネットワーク用の確率ニューロン（１０）。
【請求項５】
複数の処理サブユニットを備え、それぞれの処理サブユニットが外部確率信号（X1* - Xn*）及び確率重み信号（w1* - wn*）を受信し、算術関数を適用することでそれらを処理し、前記近似的並列カウンタ（APC）の前記入力信号（Y1* - Yn*）を構成する出力信号を生成する、
請求項４に記載の計算ニューラルネットワーク用の確率ニューロン（１０）。
【請求項６】
前記処理サブユニットは、XNOR論理ゲートから構成され、それぞれのXNOR論理ゲートは、前記外部確率信号（X1* - Xn*）と対応する前記確率重み信号（w1* - wn*）とをバイポーラ的に乗算する、
請求項５に記載の計算ニューラルネットワーク用の確率ニューロン（１０）。
【請求項７】
前記処理サブユニットは、AND論理ゲートから構成され、それぞれのAND論理ゲートは、前記外部確率信号（X1* - Xn*）と対応する前記確率重み信号（w1* - wn*）とをユニポーラ的に乗算する、
請求項５に記載の計算ニューラルネットワーク用の確率ニューロン（１０）。
【請求項８】
請求項３から請求項７のいずれかに記載の複数の確率ニューロン（１０）を備え、
前記複数の確率ニューロン（１０）の一部は、他と演算的に相互接続する、
計算ニューラルネットワーク。
【請求項９】
第２バイナリ－確率変換器（BSC2）を備え、前記第２バイナリ－確率変換器（BSC2）は、前記参照確率信号（C*）を前記定数値信号（C）及び前記ランダム信号（R）から生成し、前記複数の確率ニューロン（１０）の異なる前記処理ユニット（１１）に一斉に送信する、
請求項８に記載の計算ニューラルネットワーク。
【請求項１０】
乱数生成器（２）を備え、前記乱数生成器（２）は前記ランダム信号（R）を生成し、前記複数の確率ニューロン（１０）の異なる前記第１、第２バイナリ－確率変換器（BSC

、BSC2）に一斉に前記ランダム信号（Ｒ）を送信する、
請求項９に記載の計算ニューラルネットワーク。
【請求項１１】
前記乱数生成器（２）は線形フィードバックシフトレジスタ型である、
請求項１０に記載の計算ニューラルネットワーク。
【請求項１２】
ORグループゲート（３）を備え、前記ORグループゲート（３）は確率出力信号（S_0* - S_3*）を確率ニューロン（n_0 - n_3）のグループから受信し、その出力（S_{max}*）における最大値を取得する、
請求項８から請求項１１のいずれかに記載の計算ニューラルネットワーク。
【請求項１３】
バイナリ－確率変換器のアレイ（BSCアレイ）を備え、前記バイナリ－確率変換器のアレイは、それぞれの第１入力で受信される初期信号（x）を変換し、それぞれの第２入力で受信される前記ランダム信号（R）を用いてそれらの出力としてそれぞれ初期確率信号（x*）に変換する、
請求項８から請求項１２のいずれかに記載の計算ニューラルネットワーク。
【請求項１４】
第２乱数生成器（２’）と、バイナリ－確率重み変換器のアレイ（BSCアレイ’）と、を備え、前記バイナリ－確率重み変換器のアレイは第１入力で受信される重み信号（w）を変換し、前記第２乱数生成器（２’）から受信される第２乱数信号を用いてその出力として前記確率重み信号（w*）に変換する、
請求項５を引用する請求項８から請求項１３のいずれかに記載の計算ニューラルネットワーク。
【請求項１５】
請求項１に記載の確率ニューロン（１０）を複数備え、
前記処理ユニット（１１）がOR型論理ゲートである確率信号生成要素（１）を備える確率ニューロン（１０）を有する、マックスプーリング型層を備える、
請求項８から請求項１４のいずれかに記載の計算ニューラルネットワーク。
【請求項１６】
請求項２に記載の確率ニューロン（１０）を複数備え、
前記処理ユニット（１１）がAND型論理ゲートである確率信号生成要素（１）を備える確率ニューロン（１０）を有する、ミニマムプーリング型層を備える、
請求項８から請求項１４のいずれかに記載の計算ニューラルネットワーク。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、電子技術及び計算知能の分野に含まれるであろう。本発明の第１態様は、確率信号生成要素から構成され、確率信号生成要素は、デジタルハードウェア及び確率的計算技術に基づき、例えば、確率ニューロンの畳み込みネットワークに用いるようなコンピュータシステムに適している。本発明の他の観点は、確率ニューロンから構成され、確率ニューロンは、確率ニューロンから形成されるニューラルネットワーク（例えば、CNNで知られる畳み込み型）とと、確率信号生成要素と、を備え、それにより高い並列性が得られることが可能であり、この結果、パターン認識を非常に高速に実行することが可能となる。
【背景技術】
【０００２】
近年、広範なニューラルネットワーク（ディープニューラルネットワーク、Deep Neural Network：DNN）の使用が大量のデータから有用な情報を抽出する優れた能力から、大きな妥当性が獲得されている。そのハードウェアの実装は、シーケンシャルなVon-Neuman型計算アーキテクチャ（ソフトウェアソリューション）を用いるマイクロプロセッサ

の使用に基づく手法と比較して、並列に計算を実行するとき、実行速度の増加を可能にする。CNNは、フィードフォーワード（フィードバックがない）接続のDNNの型であり、次の文献にみられる通り、画像の形状認識に特に有用である：
- LeCun, Y., Bottou, L., Bengio, Y., Haffner, P. Gradient-based learning applied to document recognition (1998) Proceedings of the IEEE, 86 (11), pp. 2278-2323.
- Lawrence, S., Giles, C.L., Tsoi, A.C., Back, A.D. Face recognition: A convolutional neural-network approach (1997) IEEE Transactions on Neural Networks, 8 (1), pp. 98-113.

【０００３】

具体的に、CNNはニューロンが生物学的脳の一次視覚野のニューロンと非常に似た方法で受容野に接続する人工ネットワークの型であり、適用は２次元行列で実行される。したがって、他の適用の中でも画像の分類や分割のような人工視覚タスクにとても効果的である。適用は、１以上の次元の畳み込みフィルタとして振る舞う人工ニューロンの多層から構成され、通常、各層の後に関数は非線形因果マッピングを実行するように加算される。

【０００４】

ここで、処理画像（観測される対象により得られる生データ）から、一連の計算は層に配置された異なるニューロンを用いて実行される。各ニューロンは入力（処理画像から、又は他のニューロンからの入力でもよい）から比較的単純な処理を実行し、与えられる出力は入力の重みづけされた和の非線形関数（活性化関数と呼ばれる）から構成され、次式で示される。

【数１】

$$y_i = f\left(\sum_{i=1}^{N} \omega_{ij} x_j - T_i\right)$$

【０００５】

上記式はx_jとして定義されたN個の入力の総和に関連して、i番目のニューロンの出力(y_i)の依存として示し、同様に、これらの入力の各々と各ニューロン専用の重み（w_{ij}パラメータ）と、各ニューロン固有のバイアス値(T_i)の選択を示す。活性化関数fはHeaviside関数、シグモイド関数、又は正規化線形ユニット（Rectified Linear Unit：ReLU）として知られる正規化関数のような非線形関数であり、上記式の入力x_jの線形結合と0でもよい参照値間の最大値を評価する。

【０００６】

CNN型ネットワークにおいて、ニューロンは処理層に配置され、第１層の各ニューロンは処理画像から直接情報を受信し、内側の層に、先行する層に配置されたニューロンの応答が与えられる。各層の機能性は通常、畳み込みとリダクションの２種類である。畳み込み層はカーネル（よくインパルス応答とも呼ばれる）を用いて、先行する画像の畳み込みをすることから構成される。数学的に、畳み込み層はカーネルにより定義された特定の部分空間への入力情報の投影の表現として考えられることができ、情報の削減は比較的小さい。層の第２機能性は抜本的な情報削減処理から構成され、処理画像の異なる領域が選択され、信号で最も高い値（層の型はマックスプーリングと呼ばれる）、又は信号の平均値（平均プーリング）でもよい最も有力な値が選択される。最終的な結果は先行する畳み込み層からの入力信号において削減された情報である。

【０００７】

ニューラル層の２つの型（畳み込み及びプーリング）の連結で、入力画像の情報の方向は変化し、同一画像の特定の特徴が選択されて情報が削減され、一方で抽象度が増大する。このようにして、入力の情報ビットは処理画像の強度値を表すことができ、出力において認識したいカテゴリを得るであろう。

【０００８】

したがって、CNNの主な応用は画像認識及び推論の画像処理であろう。どんな場合でも

、その横断性により、CNNの応用分野は巨大であり、現在はニューラル処理アーキテクチャの使用に基づく何千もの科学研究が存在し、前述したLawrenceらの研究における顔認識、又は"Low resolution optical character recognition for camera acquired documents"と題される特許文書EP 1598770で開示されたようなテキストの認識といった処理を含む。

【０００９】

CNN処理の欠点の１つは、初期画像から生成した画像の全体に及ぶ多数の演算（連続する畳み込み及びプーリング処理）による高い計算コストである。計算コストは画像処理時間、同様に当該処理に関係づけられる使用電力により定義され、大量のデータに適用したい場合（ハイスピードビデオ処理の場合のような）には非常に高くなりうる。従来のデータ処理システム（ソフトウェアソリューション）の場合、プロセッサの使用に基づく汎用アーキテクチャはCNNの実装に対する応用に最適でなく、その結果ある種の応用において非常に高い消費及び応答時間は双方とも関連付けられる。このため、ハードウェアシステムは開発されており、これらの処理を実装し、最適に処理を並列化する試みがされている。ハードウェアを用いたCNNの最適化の試みの例として、次の刊行物がある。

-Zhang, C., Li, P., Sun, G., Guan, Y., Xiao, B., Cong, J. "Optimizing FPGA-based accelerator design for deep convolutional neural networks" (2015) FPGA 2015 - 2015 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, pp. 161-170.

【００１０】

この研究では、Zhangや共著者はCNNの最適化方法論を示し、再構成可能論理機器（FPGA）で実装している。彼らの方法は古典的なデジタル論理の使用に基づいており、並列性は汎用のマイクロプロセッサの使用に基づいた実装に関連して増加するが、完全にはCNNの固有な並列性は活用できない。

【００１１】

並列性を高め、処理ハードウェアを、可能な限り一般のDNNのアーキテクチャに適合することを目的として、計画は開発されてきており、確率計算のような慣例的でないデジタル論理が用いられる。確率計算は信号間の演算が確率的なルールに従うデジタル論理である。これは、ディープビリーフネットワークのようなディープネットワークの型が適用される次の刊行物の研究の場合であり、通常、形状認識に用いられるネットワークの型であり、統計物理学に触発されている。

-Sanni, K., Garreau, G., Molin, J.L., Andreou, A.G. "FPGA implementation of a Deep Belief Network architecture for character recognition using stochastic computation", (2015) 2015 49th Annual Conference on Information Sciences and Systems, CISS 2015, art. no. 7086904.

【００１２】

この研究は、非線形シグモイド関数に線形手法を用いているが、ニューロンがCNN型ネットワークの形に配置される必要がある場合、マックスプーリングのような特定の層の実装方法が明確ではない。一方、ネットワークの実装において無相関信号の使用のみ仮定され、確率信号間の相関を活用する可能性を残しており、ハードウェアの削減を非常に助けている。

【００１３】

ハードウェアの最適化における慣例的でないの方法論の実装に関連するが、畳み込みニューラルネットワークに適用される、次の刊行物は利用可能である。

-Alawad, M., Lin, M. "Stochastic-Based Deep Convolutional Networks with Reconfigurable Logic Fabric" (2016) IEEE Transactions on Multi-Scale Computing Systems, 2 (4), art. no. 7547913, pp. 242-256.

この研究では、確率論の特定の特性が用いられており、この確率論の特定の特性は、２つの独立した確率変数の和の確率密度関数と、これらの変数個々の確率密度（双方の畳み込みに関連する）との間の関係といったものである。CNN処理の高速化の基礎は個々のニ

ューラル要素を用いる代わりにこの確率特性の実装にある。

【００１４】

確率特性の使用にも基づく他の慣例的でないの方法論は、次の刊行物に発見される。

-Ren, A., Li, Z., Ding, C., Qiu, Q., Wang, Y., Li, J., Qian, X., Yuan, B. "SC-DCNN: Highly-scalable deep convolutional neural network using stochastic computing" (2017) International Conference on Architectural Support for Programming Languages and Operating Systems - ASPLOS, Part F127193, pp. 405-418.

この研究では、確率論理は畳み込み処理及びマックスプーリング処理双方の実装に用いられる。一方、正接バイポーラ関数（前述の式である関数f）の実装にステートマシンを用い、構成をかなり複雑化しうる（図６に見られるように）。また、相関信号の使用も活用せず、活性化関数の実装を簡潔化もしない。

【００１５】

一方、確率ニューラルネットワークにおける提案は次の近年の刊行物にも発見される。

-Li, Z., Li, J., Ren, A., Cai, R., Ding, C., Qian, X., Draper, J., Yuan, B., Tang, J., Qiu, Q., Wang, Y. "HEIF: Highly Efficient Stochastic Computing-Based Inference Framework for Deep Neural Networks" (2019) IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 38 (8), art. no. 8403283, pp. 1543-1556.

この研究では、図４ｃに示すように、近似的並列カウンタ（Approximate Parallel Counter：APC）として知られるバイナリブロックが用いられる。これは、ニューラル入力の重み付け和を実行するのに用いられるが、複雑な回路はReLU型活性化関数又はマックスプーリングブロック（前記文献の図６における）の実装において実行される。実際に、得られた活性化関数は古典的ReLUでなく、「クリップされたReLU」、すなわち飽和されたReLUである。

【００１６】

確率ロジック（又は確率計算）は、従来のバイナリロジックである量がビットのトグル周波数により符号化されるロジックに代替され、確率型性質にそのようなトグルを強制する。確率計算（SC）は時間依存のビット列のトグル周波数を用いた信号を表す近似的コンピュータ方法論である。各SC信号はビットの列において、高い値（論理’１’）を見つける可能性を表すパルスにより構成される。例えば、数値0.75はビット列の論理’１’を見つける可能性が75％であるビット列により表され、４ビットでは（１、１、０、１）、８ビットフレームでは（０、１、１、０、１、１、１、１）となる。この符号化は、ユニポーラの表現として参照され、各値は０と１の間である。負の値を含むためには、異なる符号化が求められ、バイポーラ符号化の場合と同様に、０値は１値から減算されて、最終的にビットの総数で除算される。この符号化は’p’が数のユニポーラ表現である変数p* = 2p - 1の変化の実装に等しい。バイポーラ符号化は取り得る値として［－１、１］の範囲を与える。

【００１７】

確率ロジックを用いる主な利点の１つは、低コストであるハードウェア資源で複雑な関数を実装し、信号間の相関にも依存することである。例えば、XNORゲートは双方の信号が一時的に無相関である場合、バイポーラ符号における乗算を実装し(f=x・y)、相関のある場合、単位引く２つの信号間の差の絶対値(f=1-¦x-y¦)を実装する。

【００１８】

文献では、異なる確率ニューロンの構成が提案されているが、CNNをかなり単純化可能な信号の相関を活用するものはない。本発明では、ニューロンの出力信号間の相関が畳み込みニューラルネットワークの効率的な実装に用いられる。

【発明の概要】

【課題を解決するための手段】

【００１９】

本発明は主に、確率信号生成要素と、この要素を備える計算ニューロン、特に確率ニュ

ーロンと、同様に、複数のこのようなニューロンを備える計算ニューラルネットワークと、から構成される。提案されたニューロンは、ディープ人工ニューラルネットワーク、特に畳み込みニューラルネットワークの実装に向いている。

【００２０】

本発明の第１態様は、確率信号生成要素から構成され、確率信号生成要素は、第１バイナリー確率変換器（Binary to Stochastic Converter：BSC）を備え、通常、BSCは2進の２の補数比較器から構成される。このBSCは、バイナリ信号を受信する第１入力及びランダム信号（ランダムは、純粋なランダム又は疑似ランダムのどちらか一方の型の乱雑さ含むと理解される）を受信する第２入力を順に備え、その出力においてバイナリ信号をランダム信号から第１確率信号に変換するように構成される。

【００２１】

この確率信号生成要素は、処理ユニットを備えることを特徴とし、処理ユニットは、BSCから第１確率信号を受信する第１入力、及び、BSCが第２確率信号を生成するのに用いるものと同一のランダム信号を用いた、参照確率信号として用いられ、定数値信号（例えば、0値）から生成された第２確率信号を受信する第２入力を順に有する。この処理ユニットは、少なくとも１つの算術関数に従ってBSCからの第１確率信号及び第２確率信号（参照確率信号）を処理し、結果として処理を表す確率出力信号を生成するように構成される。

【００２２】

処理ユニットにより受信される参照確率信号は、BSCに用いられるものと同一のランダム信号を用いて生成されることを条件として、異なる確率信号生成要素からの信号間の相関が得られる。

【００２３】

結果として、処理ユニットがOR型又はAND型論理ゲートを用いて適用されるような単純な算術関数を適用することが可能であり、それにより、不飽和活性化関数を実装して他の取りうる確率信号生成要素と相関可能な確率出力信号を得ることが可能である。

【００２４】

このように、本発明の第１態様の確率信号生成要素の例示的な実施形態として、処理ユニットはOR型論理ゲートであり、従って同一の論理ゲートにより適用された算術関数は最大値関数（max(a,b)）から構成され、最大値関数は正規化線形ユニット（ReLU）を具体化する基礎である。他の例示的な実施形態として、処理ユニットはAND型論理ゲートであり、従って同一の論理ゲートにより適用された算術関数はmin型(a,b)の活性化関数から構成される。

【００２５】

例えば、この確率信号生成要素は、計算ニューラルネットワーク用の確率ニューロンの一部であってもよい本発明の他の態様、具体的には第２態様であり、確率ニューロンが第１態様に従う確率信号生成要素を備える。

【００２６】

本発明の確率ニューロンは、好適には、近似的並列カウンタ（APC）を備え、APCは複数の確率入力信号を受信するように配置され、複数の入力信号を加算し、それらをその出力として２進の２の補数表記で符号化された出力信号に変換するよう構成される。この出力信号は確率信号生成要素の上記BSCのバイナリ入力信号となる。

【００２７】

同様に、例示的な実施形態として、本発明の第２態様の確率ニューロンは複数の処理サブユニットを備え、各々が外部確率信号（他のニューロンからの信号といった）と、信号結合重み信号とを受信するように配置され、これらの信号を、算術関数を適用して処理し、出力信号を生成するように構成される。処理サブユニットにより受信される双方の信号は無相関であるため、例示的な実施形態では、これらのサブユニットはXNOR論理ゲートから構成され、適用される関数はバイポーラ符号における双方の確率信号の乗算である。他のありうる実施形態として、これらのサブユニットはAND論理ゲートから構成され、適用される関数はユニポーラ符号における双方の確率信号の乗算である。その後、これらの

処理サブユニットにより生成される出力信号は、確率APC入力信号に対応するように、作動的にAPCに接続される。

【００２８】

本発明の他の態様、具体的には第３態様は、上述した発明の第２態様に従って定義された複数の確率ニューロンで実装された計算ニューラルネットワークから構成され、一部は演算的に他と相互接続する。

【００２９】

好適に、本計算ニューラルネットワークは、第２バイナリ－確率変換器すなわちBSC変換器を備え、第２BSC変換器は、確率参照信号を定数値信号の入力とランダム信号の入力から生成し、一斉に異なるニューロンの前記処理ユニットに送信するように設計される。同様に、例示的な実施形態として、本発明の第３態様のニューラルネットワークは、乱数生成器を備え、乱数生成器は、ランダム信号を生成し、一斉に異なるバイナリ－確率変換器に送信するように構成され、その結果ニューロンのAPCにより生成された全ての確率信号が相関する。

【００３０】

殆どのCNNの実装において、用いられる一般の演算は、乗算、加算及び最大値関数又はReLUである。これらの演算は、相関が正確に用いられる場合、確率回路において簡単に実装できる。畳み込みニューラルネットワークの実装のために相関信号を用いることで得られる主な利点は：

a)各ニューロンの異なる乱数生成器の実装を要求しないことによる、結果の精度を落とすことのない使用ハードウェア資源の節約をすること（確率論による実装において、資源の最大割合はランダム／乱数の生成に用いられる）

b)論理ゲートのような単純な計算ユニットのみ含むマックスプーリング関数及び畳み込みの実装の簡略化（いろいろな取り組み及び設計が文献で紹介され、ニューラルネットワークにおいてこの関数が実装可能となったが、無相関信号を用いる場合、提案された設計は相当なハードウェアスペースを要求する）

c)ネットワークの層の数に拘わらず生成器の数は一定であるため、各々の乱数生成器の使用を必要とせずにニューラルネットワークにおけるさらなる深さの層の加算が可能なこと（ネットワークの層が増えると、乱数生成器の数が増える他の実装と異なる）

から構成される。

【００３１】

好適な計算ニューラルネットワークの実施形態では、乱数生成器は線形帰還シフトレジスタ（LFSR: Linear Feedback Shift Register）型であり、ハードウェア資源において低コストかつ単純な実装である。

【００３２】

例示的な実施形態として、本発明の第３態様の計算ニューラルネットワークは、ORグループゲートを備え、ORグループゲートは、確率ニューロンのグループから確率出力信号を受信し、その出力においてその最大値を与えるように構成され、又は、ORグループゲートをANDグループゲートに置き換えることによりミニマムプーリング型関数を実装するように構成される。

【００３３】

他の例示的な実施形態として、計算ニューラルネットワークはバイナリ－確率変換器のアレイを備え、その変換器は初期信号を変換し、各々の初期信号は第１入力で受信したものであり、この信号を出力として初期確率信号に各々の第２入力で受信するランダム信号を用いて変換するように構成される。

【００３４】

好適に、本発明のニューラルネットワークは、第２乱数生成器と、バイナリ－確率変換器のアレイと、を備える。このアレイの変換器は、第１入力で受信された結合重み信号を変換し、この信号を出力として確率重み信号に変換し、第２生成器からの前述のランダム信号をこのために用いるように構成される。このようにして、無相関信号は入力及び重み

の積を生成するために使用でき、相関信号はニューロンの出力伝達関数を生成するために使用できる。

【００３５】

第１実施形態において、ニューラルネットワークは少なくとも１つのマックスプーリング型層を備え、すなわち、確率信号生成要素を備えるニューロンで実装され、確率信号生成要素の処理ユニットはOR型論理ゲートである。第２実施形態において、ニューラルネットワークはミニマムプーリング型層を備え、すなわち、確率信号生成要素を備えるニューロンで実装され、確率信号生成要素の処理ユニットはAND型論理ゲートである。そして、さらに好適な実施形態では、ニューラルネットワークは畳み込みニューラルネットワークとして実装され、畳み込みニューラルネットワークは複数のマックスプーリング型層と、複数のミニマムプーリング型とを備える。

【００３６】

２D又は３D画像処理、又は一般のN次元構造に最適な畳み込みニューラルネットワークでは、ネットワーク全体は、ニューロンの異なる層に配置される。特定の特徴が畳み込みニューラル層で抽出された後、サブサンプリング演算は通常適用され、次の層におけるデータ処理の次元を削減する。説明したように、最も用いられる演算の１つはマックスプーリングブロックであり、そのブロックではサブサンプリングが各ニューロンの出力の最大値を抽出することで実行され、この各ニューロンは畳み込みニューロン層のウィンドウ内に備えられる。最先端の実装では、ブロックのセット（アキュムレータ、コンパレータ及びカウンタ）を用い、マックスプーリング関数を実装し、これには相当な資源の使用が強いられる。しかしながら、本発明では、ネットワークの全てのニューロンで（各ニューロンの全てのBSC変換器で同一のランダム信号を用いることにより）実行される相関を用い、したがって、単一のOR論理ゲートを用いて最大値分類関数を実装することが可能である。すなわち、マックスプーリング関数（ディープニューラルネットワークの実装において広く用いられる）を、ORゲートを通してネットワーク層のニューロンの出力信号に実装し、又はミニマムプーリング型関数を、ANDゲートを通して実装することが可能である。

【図面の簡単な説明】

【００３７】

【図１】本発明の確率信号生成要素の概略図。

【図２】OR論理ゲートを用いて、それによりReLU型活性化関数を実装する確率ニューロンの好適実施形態の概略図。

【図３】OR論理ゲートに対応して２つの確率ニューロンを備えるニューラルネットワークのマックスプーリング型層の好適実施形態の概略図。

【図４】出力がOR論理ゲートに向けられ、ニューロンの出力伝達関数を生成する４つのニューロンのマックスプーリング型の単層ニューラルネットワークにおいて全ての信号が最大値である場合の好適実施形態の概略図。

【図５】無相関信号を使用して入力及び重み間の積を生成し、相関信号を用いてニューロンからの出力伝達関数を生成する畳み込み多層ニューラルネットワークの好適実施形態の概略図。

【図６】他のニューラルネットワークの実装に関する本ニューラルネットワークの性能比較表。

【発明を実施するための形態】

【００３８】

クレームされた異なる発明の好適実施形態は図１～６を参照して下記に記述される。

【００３９】

本発明の第１態様の対象である確率信号生成要素の概略図は図１に示される。見ればわかるように、確率信号生成要素（１）は第１バイナリ－確率変換器（BSC）を備え、このBSCは、バイナリ信号（A）を受信する第１入力とランダム信号（R）を受信する第２入力とを順に備え、その出力において、バイナリ信号（A）は確率信号（A*）に変換される。続いて、確率信号生成要素（１）は、確率信号（A*）と参照確率信号（C*）を受信する処

理ユニット（１１）を備える。参照確率信号（C*）は定数値信号（C）から前記ランダム信号（R）を用いて生成される。この処理ユニット（１１）の出力は、処理された算術関数を表す確率出力信号（S*）である。

【００４０】

本発明の第２態様に関して、図２は確率ニューロン（１０）の好適実施形態の概略図を示し、確率ニューロン（１０）は、確率信号生成要素（１）の好適実施形態を含み、確率信号生成要素（１）は、OR論理ゲートを、出力（S*）においてReLU型活性化関数を実装する処理ユニットとして備える。確率ニューロン（１０）のこの実施形態は、複数のXNOR論理ゲートを備え、それらの各々が外部確率信号（X1* - Xn*）と確率重み信号（w1* - wn*）を受信するように配置される。これら２つの信号が相関しないことを考慮して、XNORゲートにより適用された関数は、バイポーラ符号における双方の確率信号の乗算から構成される。

【００４１】

続いて、確率ニューロン（１０）の本実施形態のXNORゲートの出力は近似的並列カウンタ（APC）の入力に演算的に接続され、近似的並列カウンタ（APC）は、入力信号（Y1* - Yn*）の和をとり、それらの和を２の補数に符号化されたバイナリ値にその出力において変換するように構成される。言い換えれば、近似的並列カウンタ（APC）はいくつの’１’である信号がXNOR論理ゲートの出力においてあるかを評価し、それらの出力においてその和のバイナリ信号を与える。続いて、この和は確率信号生成要素（１）のバイナリ－確率変換器（BSC）に送信される。前述のように、確率信号生成要素（１）のこの実施形態はOR論理ゲートを備えてReLU型活性化関数を実装し、バイナリ－確率変換器（BSC）で受信したものと同一のランダム信号（R）を用いて出力（S＊）をする。

【００４２】

図３は、ニューラルネットワークの２つのニューロンの接続の例示的な実施形態である本発明の第３態様の主部を模式的に示す。この実施形態は、２つのOR論理ゲートを有し、図２に示され、上記された２つの確率ニューロン（１０）を備える。同様に、第２のバイナリ－確率変換器（BSC2）を備え、BSC2は参照確率信号（C*）を生成し、双方の確率ニューロン（１０）のOR論理ゲートに送信するように構成される。見ればわかるように、参照確率信号（C*）は第２バイナリ－確率変換器（BSC2）において定数値信号（C）及びランダム信号（R）から生成される。

【００４３】

当業者であればわかるように、本発明の相関ニューラル信号を生成することの実装はReLU関数を生成することがより複雑な最先端の実装と対照であり、飽和している欠点も持ち、その結果典型的な機械学習処理で用いられる標準ReLUではない。

【００４４】

本発明さらなる効果は、図４からわかるように、ネットワークニューロンの全ての出力が相関するため、ORグループゲート（３）を用いてネットワーク層の確率ニューロン（n_0 - n_3）のグループから確率出力信号（S_0* - S_3*）へのマックスプーリング関数（ディープニューラルネットワークの実装に広く用いられる）の効率的な実装が可能であること、及び、出力（S_{max}*）としてそれらの最大値を与えることが可能であること、又は、ORグループゲート（３）をANDグループゲートに置き換えることでミニマムプーリング型の関数の実装が可能であることである。

【００４５】

図５は、より一般的な態様のニューラルネットワークを示すブロック図であり、２つの乱数生成器（２、２’）が用いられる。見ればわかるように、第１乱数生成器（２）は畳み込みネットワークのニューロン（n_0 - n_n, n'_0 - n'_n）の近似的並列カウンタの入力信号（ｘ）、参照信号（０）及び出力信号の変換に用いられる。第２乱数生成器（２’）はネットワークの重み（ｗ）の確率に変換のみに用いられる。

【００４６】

上記からわかるように、本発明のニューラルネットワークの各層の出力は次の層のニュ

ーロンの入力信号が不正確であることのリスクなしに一致し、これは、第１乱数生成器（２）により生成された場合、この出力が確率重み信号（w^*）と乗算可能であり、この確率重み信号は入力信号と相関がなく、第２乱数生成器（２’）からの乱数を用いたバイナリー確率変換器のアレイ（BSCアレイ’）により生成されるためである。

【００４７】

見ればわかるように、確率ニューロンの本実装は従来では考えられなかったニューラルネットワーク全体をネットワーク全体に対して２つの乱数生成器（２、２’）のみで実行させることをできるようにし、ネットワーク全体に実装されたデジタルゲートの観点でコストを簡略化することをできるようにした。

【００４８】

図６は、本実装と機械学習の分野でLeNet-5（刊行物Y. LeCun, L. Bottou, Y. Bengio and P. Haffner: Gradient-Based Learning Applied to Document Recognition, Proceedings of the IEEE, 86(11):2278-2324, November 1998で示される）として知られる他のディープニューラルネットワークの実装との性能の比較表を示す。提案モデルでの比較されるハードウェアの実装は次の通りである。

-FPGA16. S. I. Venieris and C. Bouganis, “fpgaconvnet: A framework for mapping convolutional neural networks on fpgas,” in 2016 IEEE 24th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM) , May 2016, pp. 40-47.

-FPGA17a. Z. Liu, Y. Dou, J. Jiang, J. Xu, S. Li, Y. Zhou, and Y. Xu, “Throughput-optimized fpga accelerator for deep convolutional neural networks,” TRETS, vol. 10, pp. 17:1-17:23, 2017.

-FPGA17b. Z. Li, L. Wang, S. Guo, Y. Deng, Q. Dou, H. Zhou, and W. Lu, “Laius: An 8-bit fixed-point cnn hardware inference engine,” in 2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC), Dec 2017, pp. 143-150.

-FPGA18.S.-S. Park, K.-B. Park, and K. Chung, “Implementation of a cnn accelerator on an embedded soc platform using sdsoc,” 02 2018, pp.161-165.

【００４９】

比較するために、LeNet-5ネットワークはFPGAで実装されている。このCNN設計は、手書きの数字認識の機械学習（MNIST）において高く標準化されたデータベースを処理することを目的とし、60,000枚の訓練画像及び10,000枚のテスト画像で構成される。LeNet-5型のCNNアーキテクチャは２つの畳み込み層と３つの全結合層から構成される。このCNNにおけるMNISTの成功率は約98.5%であり、提案ニューラルモデルで用いられるハードウェアの実装は97.6%である(信号をバイナリ化することに起因し、また、確率計算方法論であるから、ソフトウェアの場合より幾分か低い)。

【００５０】

実装は、Arria 10 10AX115H2F34E1SG FPGA、150MHzのクロック周波数で作動、バイナリ信号の精度が8ビット、を用いて実行される。

【００５１】

このように、提案した手法は性能とエネルギー効率の観点で他のアーキテクチャより性能が優れている。その結果、提案した確率CNNの実装はFPGA17aと比較して27倍優れた性能（秒当たり及びメガヘルツ当たりの推論で測定）と、FPGA17bと比較して6.3倍優れたエネルギー効率（ジュール当たりの推論で測定）を成し遂げることを示し、組み込みシステムのアプリケーションに有望である。

【００５２】

提案された新しいニューロンの設計により、システム全体の消費面積を削減し、提案されたアーキテクチャは完全なCNNを単独のFPGAで完全に並列的に調整し、これは、他の３つの実装でなされたものであるネットワークの一部の実装のみをし、循環を基礎に反復

する（逐次実装）と対照的である。

【００５３】

この集合的な並列は提案された設計のより低レイテンシの主な理由であり、それはネットワークの逐次的な反復が必要ないからである。

【００５４】

図６の表はまた、タスク当たりに必要とされるハードウェア資源を強調する。上記実装は、RAM及びデジタル信号処理（Digital Signal Processing：DSP）ブロックといった特定のハードウェアブロックを用いることにより、大きい面積を必要とする。本提案では、DSPブロックは、従来にはない計算技術（確率計算）が古典的なバイナリロジックの代わりに用いられるため使用されない。同時に、計算が再起的に実行されないためメモリブロックが必要とされず、消費電力の主要な原因を減らし、この原因はチップ外のメモリとのデータ転換又はメモリアクセスの演算である。

【００５５】

本提案の第１態様を特徴づけることは、相関した確率信号を用いた処理ユニット（１１）と２進－確率変換器（BSC）の組み合わせである。発明の他の態様は、それに続く、XNORゲートといった処理サブユニットのアレイと確率ニューロン（１０）を実装するバイナリ－確率変換器（BSC）に接続された近似的並列カウンタ（APC）との組み合わせである。そして、第３態様は全てのニューロンに共通する乱数生成器（２）を備えたネットワークに実装することである。乱数生成器（２）が負及び正の数双方のランダム信号（R）（生成された符号ビットが’０’又は’１’に等しい確率は50%）を生成することを考慮すると、レベル０の確率信号はランダムに0と1の間を変動するビットから構成され、各レベルは50%の確率である。この発明の要点は、確率参照値と近似的並列カウンタ（APC）からの信号とを正しく組み合わせることであり、この信号はバイナリ－確率変換器（BSC）により確率に変換されたものであり、組み合わせは同一のランダム信号（R）が用いられる。したがって、０レベル確率信号と、バイナリ－確率変換器（BSC）により与えられる確率信号（A*）と、は完全に相関し、それらの組み合わせ、例えばORゲートにおいて、最大値信号が与えられてもよい。

【図面】

【図１】

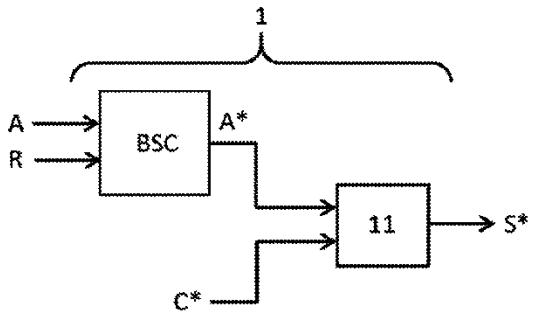


Fig. 1

【図２】

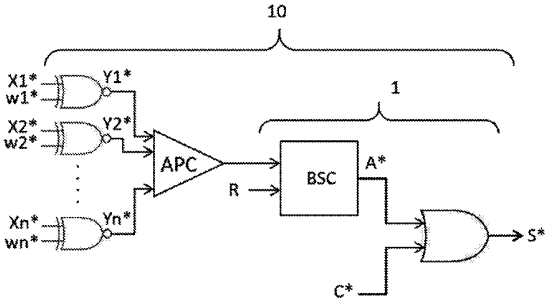


Fig. 2

【図３】

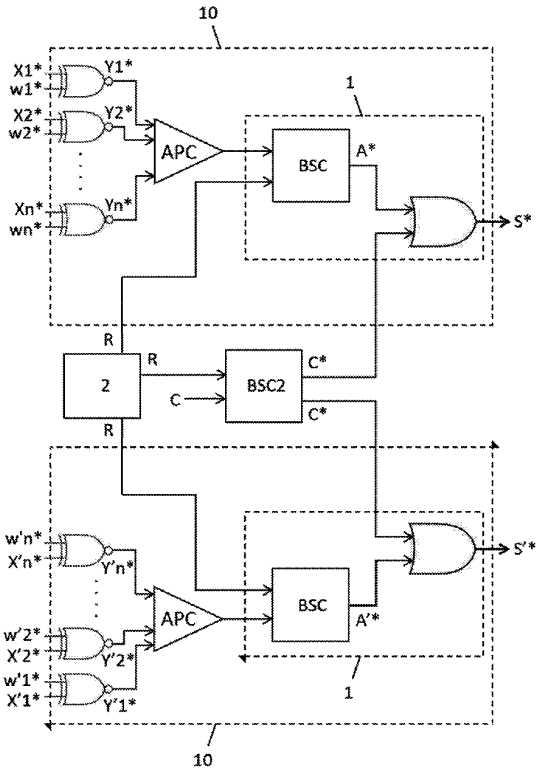


Fig. 3

【図４】

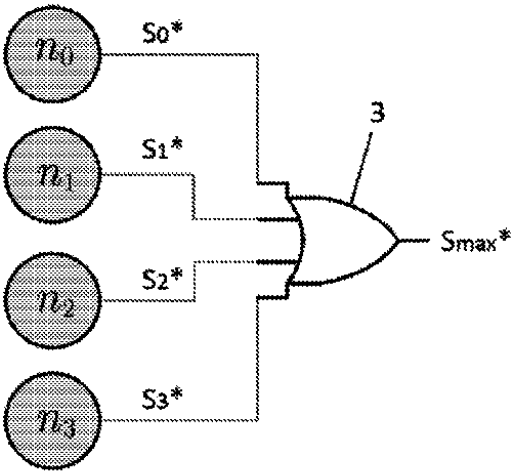
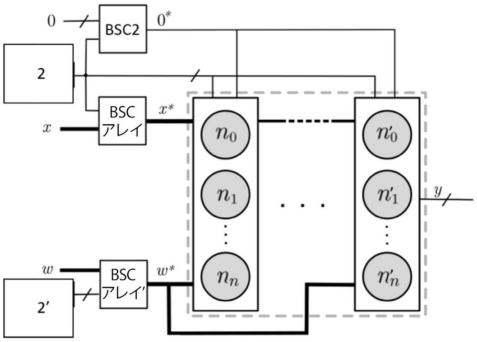


Fig. 4

【図５】



【図６】

LENET-5の実装である他研究との技術的特徴の比較

モデル	FPGA16	FPGA17a	FPGA17b	FPGA18	提案
用いたアーキテクチャ	シーケンシャル（2進）	シーケンシャル（2進）	シーケンシャル（2進）	シーケンシャル（2進）	パラレル（確率）
FPGAプラットフォーム	Zynq XC7Z020	Virtex7 VX690T	Virtex7 485t	Zynq zc706	Arria10 GX1150
周波数（MHz）	100	100	–	166	150
レイテンシ（us）	7916	94,2	960	1600	3,4
キロ干渉毎秒（KIPS）	0,13	10,6	1,04	0,625	294,1
処理速度（KIPS/MHz）	0,001	0,1	–	0,003	1,96
消費電力（W）	–	25,2	0,47	10,98	21
エネルギー効率（KI/J）	–	0,42	2,21	0,056	14
用いたロジック（LUT/Alm）	9682	233K	7204	39837	343,4K
用いたDSPの数	2,4	2907	574	59	0
用いたメモリブロック	6,16	477	343,3	97	0

フロントページの続き

スペイン国イスラス、バレアレス、パルマ、デ、マリョルカ、エディフィシ、ソン、リェド、カレテラ、デ、バルデモッサ、キロメトロ、７、５、カンプス、ウニベルシタリオ
(72)発明者　クリスティアム、カミロ、フランコ、フラッセル
スペイン国イスラス、バレアレス、パルマ、デ、マリョルカ、エディフィシ、ソン、リェド、カレテラ、デ、バルデモッサ、キロメトロ、７、５、カンプス、ウニベルシタリオ
審査官　今城 朋彬

(56)参考文献　Te-Hsuan Chen and John P. Hayes，Design of Division Circuits for Stochastic Computing，2016 IEEE Computer Society Annual Symposium on VLSI，IEEE [online]，2016年07月13日，pp.116-121，https://ieeexplore.ieee.org/abstract/document/7560183
Yi Xie, et al.，Fully-Parallel Area-Efficient Deep Neural Network Design using Stochastic Computing，IEEE Transactions on Circuits and Systems II: Express Briefs，IEEE [online]，2017年08月30日，pp.1382-1386，https://ieeexplore.ieee.org/document/8022910
T. LEE, Vincent et al.，Energy-Efficient Hybrid Stochastic-Binary Neural Networkds For Near-Sensor Computing，arXiv [online]，2017年，all 6 pages，Retrieved from The Internet: 〈URL: https://arxiv.org/pdf/1706.02344 〉

(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ０６Ｎ　３／０４７
Ｇ０６Ｎ　３／０６３