ディープラーニング ハードウェアの低消費電力化に寄与する学習時メモリーエラー解析手法

Memory Error Analysis Methodology Contributing to Reduction in Power Consumption of Deep Learning Hardware

丸亀 孝生 ■MARUKAME Takao
西 義史 ■NISHI Yoshifumi
三谷 祐一郎 ■MITANI Yuichiro

近年，様々な分野でディープラーニング（深層学習）技術の導入が進んでいるが，最適なパラメーターを長時間にわたって探索することから，ハードウェアの消費電力が増大するという課題がある。

ディープラーニング ハードウェアの低消費電力化に向けた取組みの一環として，東芝は，教師なし学習器である制限付きボルツマンマシン（RBM）をモチーフに，RBMアーキテクチャーをFPGA（Field Programmable Gate Array）に実装してDeep Belief Network（DBN）を構築し，メモリーエラーがRBMの性能に及ぼす影響を調べた。学習時のメモリー消費電力を低減するため，メモリーを低電圧で動作させた際のビットエラー率と学習性能に及ぼす影響を定量的に明らかにする手法を確立した。この手法を用いることで，ディープラーニング ハードウェアの低消費電力化を図ることができる。

The movement toward the introduction of deep learning technologies has recently accelerated in various fields. However, the high power consumption of hardware for deep learning caused by prolonged searching for optimal parameters presents a serious issue.

With this as a background, Toshiba, in cooperation with the Swiss Federal Institute of Technology in Lausanne and Hokkaido University, has carried out studies on the low-power operation of deep learning hardware using the motif of a restricted Boltzmann machine (RBM), one of the unsupervised learning algorithms, in a deep belief network (DBN). We have conducted simulation experiments using RBMs implemented on a field-programmable gate array (FPGA) and confirmed the robustness of this system against memory errors during and after learning. Furthermore, we have established a methodology to quantitatively analyze bit error rates versus learning performance of memory performing low-voltage operation. This methodology is expected to contribute to a reduction in the power consumption of deep learning hardware during learning sequences.

1 まえがき

近年，ディープラーニング技術は，従来型の機械学習アルゴリズムよりも優れた，実用的に高い認識性能や分類性能を持つことから大きな注目を集めている。十分に高い性能を得るには，最適なパラメーターを探索するための長時間にわたる学習処理が必要であり，新たなアプリケーションを探索するには，演算処理を加速するディープラーニング技術に特化した，演算効率の高いハードウェアが求められている。

そこで東芝は，ディープラーニング ハードウェアの低消費電力化の実現に向け，RBMをモチーフに，DBNの最終的な認識正答率を指標として，使用するメモリーを低電圧駆動した際のエラー耐性を詳しく解析した。

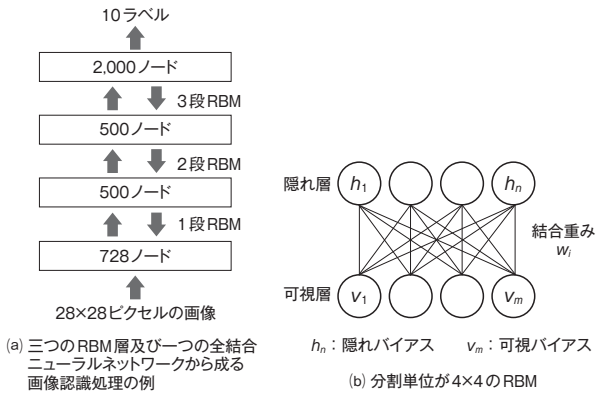
ここでは，開発した手法と定量的評価について述べる。

2 RBMの概要

ディープラーニング技術は，多層の人工ニューラルネットワーク構造から成り，ネットワークの重み情報を更新してより良い答えを導くには，ネットワークの結合重みを更新していく必要がある。このため，汎用GPU（Graphics Processing Unit）に代表されるアクセラレーターが用いられる。しかしGPUは，CPUに比べて消費電力が大きいことから，FPGAやASIC（用途特定IC）などを用いてディープラーニング技術に特化した，演算効率の高いハードウェアの開発が求められている。

ディープラーニング技術は，“教師あり学習”と“教師なし学習”の二つのタイプに大別される。一般に，扱うデータの種類や目的がはっきりしている場合は，教師あり学習で最適化していくと優れた性能が期待できる。一方，目的はまだはっきりしていないが，得られるデータが非常に膨大で下準備で粗いデータ分類をしておく場合は，教師なし学習が有用な場合がある。もっとも著名な教師なし学習を利用したディープラーニング技術の一つが，2006年にG. Hintonらにより示されたDBNである[1]。ブームの火付け役となったアルゴリズムであり，今日でも理論研究や新規の応用探索において用いられることが多い。

DBNは，複数のRBMという教師なし学習器と，全結合の教師あり学習器から成り，RBMを効率的なハードウェアで作ることにより，全体の学習に関わる演算性能を向上させることができる（**図1**(a)）。RBMは，全結合ニューラルネットワークの一種のため（図1(b)），入力数と出力数の増加に伴って指数関数的に計算量とメモリー使用量が増加する。したがって，高速かつ小面積な全結合ニューラルネットワークの回路を作ることは実用上の重要な課題である。また，計算が複雑になる



(a) 三つのRBM層及び一つの全結合ニューラルネットワークから成る画像認識処理の例

(b) 分割単位が4×4のRBM

図1. DBNとRBMの構造 — 三つのRBM層と一つの全結合ニューラルネットワークで画像認識処理を行う例と，分割単位が4×4のRBMである。

Examples of DBN and RBM structures

につれて計算中やメモリーの読み書きにおいて，データの信頼性を保証することも重要になる。LSIで低消費電力動作を目指すときは，回路に供給される電圧が小さくなるため，ノイズなどの影響による誤動作に対する耐障害性を設計の段階から作り込む必要があるが，ニューラルネットワークは内部構造に冗長性を持つことから，エラーに強いアーキテクチャーが期待できる。

RBMは，冗長性評価の良いモチーフになるため，RBMアーキテクチャーをFPGAに実装して評価を行った。

3　RBMハードウェアの構成

RBMアーキテクチャーにおける処理の流れを**図2**に示す。

RBMは，一層ごとの逐次処理が可能なので，多層にした場合でも計算量は指数関数的ではなく線形に増加する。一般的なニューラルネットワークと同様に，あるニューロン（ノード）には複数のノードからの入力信号が与えられ，それぞれで結合重みとの積が計算された後に全てが足し合わされる，いわゆる積和演算で信号処理が行われる。積和演算の結果は，シグモイド関数などのしきい値関数で評価され，RBMの場合はその値が確率値としてふるまうことがモデルで定義づけられているため，ランダムな0又は1の値と比較されてノードから出力される（図2の二項分布）。一つのRBMは，可視層と隠れ層の2層から構成され，可視層と隠れ層のノードは結合されて結合の強さを表す結合重み値と各ノードのバイアス値で表現される（図1(b)）。

一般に，RBM学習ルールにはCD（Contrastive Divergence）法が用いられる[1]。可視層に与えられる入力データは，隠れ層に向かって積和演算によってそれぞれのノードに対して信号処理され，その出力が今度は隠れ層にとっての入力値になり，可視層に与えられて逆方向に計算される。この過程を繰り返すなかで，多くのデータによってしだいに特徴的なデータが重みデータとして学習され，その結果がメモリーに蓄積されていく。

この学習アルゴリズムを実行するハードウェアアーキテクチャーを作成した。入力バッファーにデータが取り込まれ，CD法による演算が同じユニットで繰り返されると同時に，更新すべき結合重み値と可視バイアス値が計算され，ローカルメモリーに保存される。全体のシーケンスは，有限ステートマシン（FSM）で制御される。ハードウェアの構成は，設計時にRTL（Register Transfer Level）モデルとしてシミュレーションで検証している。

このハードウェアは，時分割で積和演算を処理できる[2]。ネットワークのノードをある単位で分割し，それぞれ4×4（図1(b)），8×8，及び16×16を単位とする。ネットワーク単位を大きくすると，**図3**で示すようにデータ処理に掛かる時間は短くなる。すなわち，並列数の増加に伴ってデータ処理速度は速くなる。

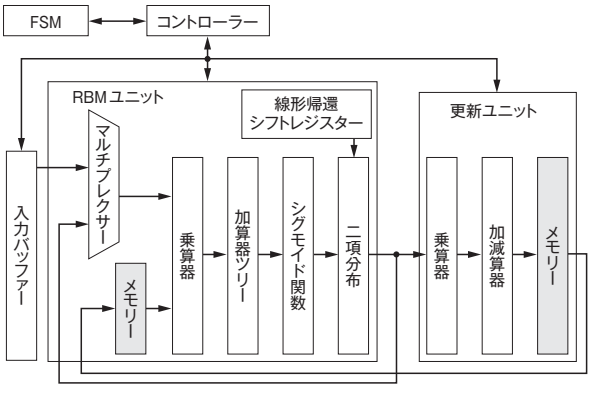


図2. RBMアーキテクチャーのフローチャート — CD法により演算が繰り返され，更新された結合重み値と可視バイアス値がメモリーに蓄積される。

Flowchart of operations in overall architecture of RBM hardware

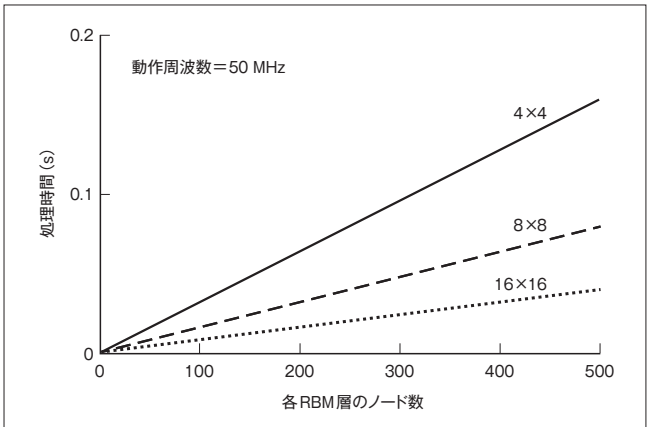


図3. 並列化構造ごとの処理速度のシミュレーション結果 — 分割単位が大きくなるほど，処理速度が速くなる。

Results of simulation of processing speed for parallelized architectures

4　メモリーに対するエラー注入解析

RBMハードウェアの性能をシミュレーションで確認した後，低消費電力化を目指して低消費電圧で動作するメモリーを想定し，メモリーエラーがRBMの性能に及ぼす影響について調べた。ここでは，RBMハードウェアと同様のふるまいをするシミュレーターを作成し，ソフトウェアプログラムの中でメモリーに対する疑似的なエラー注入解析を実施した。DBNは，3層のRBMと最終段の全結合層から構成され，RBMは1層ずつの学習を逐次的に進めて最後に全結合層の最適化を行うが，このとき，ある学習試行回数（単位：エポック）からは，前段3層のRBMに対しても学習した値を微調整するために，教師あり学習である“Fine-tuning”を行っている（今回は7エポックから実施）。

最初に，RBMの学習時だけにエラーを注入したケース，次にメモリーエラーを想定した2ケースの合計3ケースで解析を行った結果を**図4**に示す。

まず，メモリー内のセルごとにセル製造後の初期不良を意味する固定的なエラーがあるケースを想定した（図4(a)）。ビット

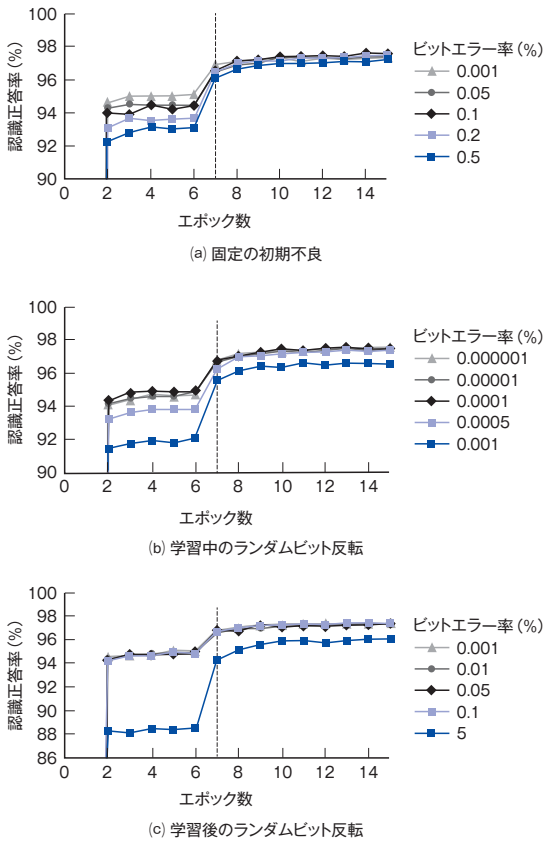


図4．ビットフリップ（反転）モデルに基づくエラー注入解析の結果 ― 7エポックからFine-tuningを開始したが，学習中にエラー注入したケースの認識性能は十分に回復していない。

Fault injection analysis results based on bit-flipping model

エラー率に依存して6エポックまでの結果に認識正答率の劣化が見られるが，7エポック以降は性能が回復し，97％に漸近する結果が得られた。

次に，学習中にメモリー内のランダムな箇所にエラーが発生するケースについて調べた（図4(b)）。初期不良の場合と同様に，ビットエラー率の増加とともに6エポックまで認識正答率が劣化する傾向が見られた。一部で例外的に性能が良い場合もあるが，仮定したビットエラー率が十分小さいことから誤差とも考えられる。重要なことは，初期不良の場合に比べて低いビットエラー率であるにもかかわらず，大きく性能が劣化している点である。特にビットエラー率0.001％の場合，6エポックまでの段階で，初期不良のケースの認識正答率が約95％であるのに対し，学習中にエラーを注入するケースでは約92％となっている。より大きな特徴としては，7エポック以降のFine-tuning開始後でも認識性能が十分に回復せず，学習がうまく働かないことを示唆している。

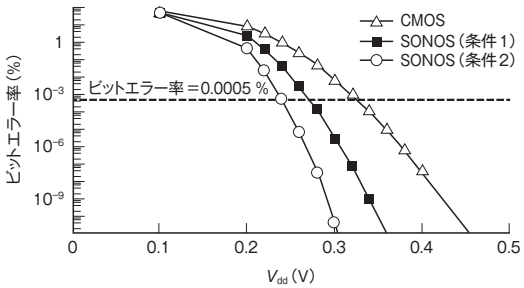
最後に，学習後にエラーを注入するケースについて評価した（図4(c)）。ビットエラー率0.001％に着目すると，認識性能にはほとんど劣化がなく，ビットエラー率が増加して0.1％になってもまだ劣化が始まっていない。5％と極端な場合には顕著に性能が劣化するが，現実的に製造されるメモリーではほとんど観測されないビットエラー率である。

これらの結果から，3ケースのうちでもっともエラーの影響を受けやすいのは学習中のエラー発生であり，もっともエラーの影響に強いのは学習後のエラー注入のケースであった。

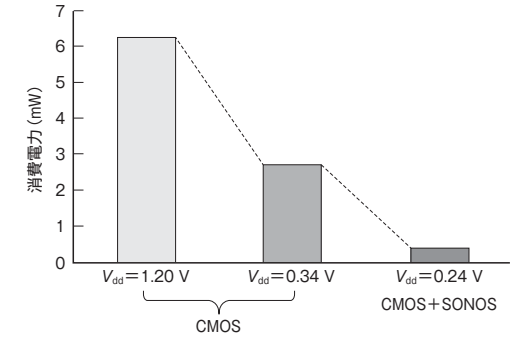
この他，ランダムな箇所のエラー発生だけでなく，エラーが発生するビットの位置が与える影響も調べた。データは，ハードウェアの中では固定小数点表現で処理されるため，例えば，先頭ビットのMSB（Most Significant Bit）は符号ビットで，それ以降は，LSB（Least Significant Bit）に向かって整数ビット，小数ビットと並ぶ構成となる。今回，データを強制的に元の値からある特定の値に変化させ，符号，整数値，及び小数値の変化，並びにそれらが与える影響度について調べた。その結果，値は大きく変化し，符号が影響する場合には大きな性能劣化を確認した。

5　SRAMのエラー耐性と低消費電力化の定量的評価

一般に，演算器の速度と同等のアクセス速度を実現できるSRAM（Static RAM）をメモリーとして用いた場合，SRAMの製造プロセスに依存したリーク電流に起因する静的な消費電力と，主に動作速度に依存する動的な消費電力の，二つの観点から低消費電力化を検討できる。これらへの影響を併せ持ったパラメーターはSRAMの電源電圧であり，どのようなメモリー構成でも，電源電圧を低下させることで全体の低消費電力化が可能になる。ただし，ここでは速度の観点での議



(a) ビットエラー率のV_{dd}依存性



(b) SRAMのV_{dd}と消費電力の関係

CMOS：相補型MOS

図5．SRAMのビットエラー率のV_{dd}依存性と低消費電力化 ― SRAMのV_{dd}を下げることで低消費電力化が可能である。

Results of analysis of bit error rate and power consumption dependence on power supply voltage (V_{dd}) of static random-access memories (SRAMs)

論は行わない。

SRAMへの供給電圧V_{dd}を低下させると，セルを構成する六つのMOS（金属酸化膜半導体）トランジスターの動作マージン（スタティックノイズマージン）が減少し，記憶保持に対するノイズの影響が大きくなる[3]。これは，V_{dd}に依存してSRAMにエラーが生じることを意味しており，セルごとに製造時における初期ばらつきの影響が異なるため，V_{dd}を低下させたとき，あるセルではエラーが頻発するようになる（**図5**(a)）。前述したビットエラー率と認識性能の関係から，全体性能が劣化しないSRAMのビットエラー率まではV_{dd}を下げることができるので，低消費電力化が可能になる。図5(a)に示すように，V_{dd}を0.4 V付近まで低下させても認識性能の劣化はない。また，SONOS（Silicon-Oxide-Nitride-Oxide-Silicon）トランジスターのように，しきい値を後から調整できる素子によってSRAMのセルばらつきを抑制してビットエラー率を低くすることも期待でき，更にV_{dd}を下げることができる[4]。これらにより，SRAMでは通常のV_{dd}に比べて1/4程度まで低消費電力化できる（図5(b)）。

6　あとがき

ここでは，ディープラーニング技術の代表的アルゴリズムの一つであるDBNに焦点を当て，教師なし学習器であるRBMに対するエラーの影響を，教師あり学習であるFine-tuningでの改善効果も含めて系統的に調べた。ディープラーニング技術は，ニューラルネットワークが持つ冗長性に由来してハードウェア化した際もエラー耐性が備わっていることを解析で明らかにした。また，定量的にエラーの影響を調べることで，どのような場合にエラーの影響が顕著に生じるかを，ハードウェア開発の前に調べるための手法を確立できた。この手法を用いたエラー耐性の解析結果を積極的に活用することで，メモリーの低電圧化，すなわちディープラーニング ハードウェアの低消費電力化が可能になる。

今後は，今回確立した手法をベースに，新規開発するメモリーも用いた更に低消費電力のディープラーニング ハードウェアを構築していく。

この研究の一部は，北海道大学 浅井哲也教授，及びスイス連邦工科大学ローザンヌ校　Alexandre Schmid博士と共同で実施した。

文　献

(1) Hinton, G. E. ; Salakhutdinov, R.R. Reducing the dimensionality of data with neural networks. Science. **313**, 5786, 2006, p.504－507.
(2) Ueyoshi, K. et al. "Scalable and Highly Parallel Architecture for Restricted Boltzmann Machines". 2015 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP'15). Kuala Lumpur, Malaysia, 2015-02, RISP. 2015, p.369－372.
(3) Marukame, T. et al. Error tolerance analysis of deep learning hardware using a restricted Boltzmann machine towards low-power memory implementation. IEEE Transactions on Circuits and Systems II: Express Briefs. **64**, 4, 2016, p.462－466.
(4) Suzuki, M. et al. "Improvement of gate disturb degradation in SONOS FETs for Vth mismatch compensation in CMOS analog circuits". Proceedings of 2013 International Conference on IC Design and Technology (ICICDT). Pavia, Italy, 2013-05, IEEE. 2013, p.195－198.

丸亀　孝生　MARUKAME Takao, D.Eng.
技術統括部 研究開発センター LSI基盤技術ラボラトリー研究主務，博士（工学）。脳型コンピューティングLSI技術の研究・開発に従事。応用物理学会，人工知能学会，IEEE会員。
Advanced LSI Technology Lab.

西　　義史　NISHI Yoshifumi, D.Sci.
技術統括部 研究開発センター 研究企画部参事，博士（理学）。研究企画業務に従事。応用物理学会会員。
Strategic Planning Dept.

三谷　祐一郎　MITANI Yuichiro, D.Eng.
技術統括部 研究開発センター LSI基盤技術ラボラトリー研究主幹，博士（工学）。ゲート絶縁膜の信頼性技術の研究・開発に従事。応用物理学会会員。
Advanced LSI Technology Lab.