OUKA
The University of Osaka
Institutional Knowledge Archive

Title	Energy-efficient DNN Training with Approximate Computing and Voltage Scaling
Author(s)	鄭, 泰禹
Citation	大阪大学, 2021, 博士論文
Version Type	VoR
URL	https://doi.org/10.18910/82288
rights	
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

https://ir.library.osaka-u.ac.jp/

The University of Osaka

論 文 内 容 の 要 旨

氏 名 （ 鄭泰禹 ）	
論文題名	Energy-efficient DNN Training with Approximate Computing and Voltage Scaling （近似コンピューティングと電源電圧スケーリングによる高エネルギー効率DNN学習）

論文内容の要旨

This dissertation aims at providing a design methodology to achieve energy minimization for DNN training with useful implementation techniques. First, to reduce the primary MAC computation in training with FP, this work proposes to use logarithm-approximate multiplier (LAM) for training. By approximating float-point multiplication to cheap fixed-point addition, the NN training engine is expected to be implemented with smaller delay, fewer gates, and lower power consumption. With a dedicated hardware for NN training engine and a 2-D classification dataset, this dissertation demonstrates that LAM-based training can achieve 10% speed-up, 2.3X power and area improvement, respectively, compared with traditional training with FP32 exact computation. In addition, LAM also reveals its high compatibility with conventional BWS. If LAM is applied to NN where BWS is already implemented, LAM and BWS can enjoy synergy effect and up to 4.9X power reduction is attained with sacrificing up to 1% accuracy loss. Also, the processor level design can exploit the advantage of LAM. An experimental GPGPU embedded with LAM executing a NN training workload presents 28% power reduction improvement and the improvement reaches 41% with LAM + BWS. Finally, LAM is qualified for deeper-layer NN training. Up to four hidden layers, LAM- based training yields the same level of accuracy as training with accurate multipliers, even with aggressive BWS.

Finally, this dissertation provides an overall scheme for minimizing NN training energy as the main contribution of this work and gave it a name as adaptive bit-width and voltage scaling (ABVS) which can leverage the FP units with configurable bit-width. The key idea is that the NN training starts with smaller fraction bit-width (FB), and then FB is gradually increased depending on present training quality during the training phase. Since smaller FB possesses shorter latency, this training scheme can concurrently adapt the bit-width and voltage scaling and hence intensify energy reduction. Experimental results across a various scale of training datasets and DNN applications reveal that the proposed ABVS flow in training can achieve the comparable training quality as FP32, but up to 62% energy reduction, and at most 37% reduction even compared with the training under the least sufficient FB, with at most 0.5% quality loss. This work also gave an explanation of why ABVS works in modern DNN training. Furthermore, the investigation of different rounding schemes is conducted. Results show that conventional round-to-the-nearest (RTNE) provides better trade-off between training quality and energy efficiency than round-to-zero (RTO).

Overall, this dissertation seeks for achieving energy minimization with two main countermeasures, approximate computing and voltage scaling. LAM and BWS techniques are exploited in this work as the former, while MWVS belonging to the latter combines MCMM and ASA to achieve further voltage reduction. The ABVS flow can be recognized as an integrated solution for all the aforementioned strategies. In addition, all the methods apply to GPUs and the dedicated ASICs for achieving efficient DNN training.

Next, this dissertation addresses the disadvantage of non-zero timing error risk involved in conventional ASA methods and provide a design applying to mode-wise voltage-scaling (MWVS) design with enhancing ASA benefit on VS efficiency yet guaranteeing zero timing errors. First, the MWVS design is formulated as an optimization problem that minimizes the overall power consumption considering each mode duration, as well as the achievable voltage reduction, and accompanied circuit overhead. The proposed method explores the solution space with the downhill simplex algorithm (DSA). To attain a solution, i.e., a design, the feature of multi-corner multi-mode (MCMM) in a commercial tool is exploited to perform mode-wise ASA, where the ASA is realized by assigning with sets of false paths specialized for individual modes to obtain the maximum voltage scaling for each mode. Experimental results based on a popular RISC-V hardware design show that the proposed MWVS design can save 20% more energy compared with the conventional VS approach and acquire 15¥% more gain compared with single-mode ASA. In addition, the cycle-by-cycle fine-grained false paths identification is also proposed and it successfully reduced 42% leakage power.

様式７

論文審査の結果の要旨及び担当者

氏　名　（　鄭 泰禹　）		
	（職）	氏　　名
論文審査担当者		
主 査	教授	橋本 昌宜
副 査	教授	三浦 典之
副 査	准教授	粟野 皓光（京都大学大学院情報学研究科）
副 査	准教授	谷口 一徹

論文審査の結果の要旨

本論文は、ハイエンドGPUや専用学習エンジン設計で利用可能なエネルギー効率の高いDeep Neural Network (DNN)の学習向け設計手法を提案している。提案手法は、2つの低電力化技術(1)近似コンピューティングと(2)電源電圧スケーリングの組み合わせで構成される。主な成果は以下の通りである。

1. 近似コンピューティング
Neural Network (NN)の学習は膨大な浮動小数点演算を必要とする。本論文では、推論に適用されていた対数近似乗算器を学習に用いることで、積和演算を軽量化することを提案している。浮動小数点乗算を固定小数点加算に近似することで、NN学習に用いる浮動小数点演算器が高効率実装できる。対数近似乗算器を組み込んだGPUをFPGA実装してNN学習プログラムを実行し、28%の消費電力削減を確認している。

2. 電源電圧スケーリング
電源電圧スケーリングは消費電力を低減する有力な手法であるが、ゲートの遅延が増加し、タイミングエラーが発生する可能性がある。本論文では、モード毎の電源電圧スケーリングを適用して、タイミングエラーをゼロにしつつ、電力効率を向上させる設計手法を提案している。各モードの実行時間、達成可能な電源電圧降下、および回路オーバーヘッドを考慮して、消費電力を最小化する回路設計最適化問題を定式化し、その解法を提示している。RISC-Vプロセッサを用いた実験では、従来の電源電圧スケーリングと比較して20%以上、単一モード電源電圧スケーリングと比較して15%以上の電力削減を得ている。

3. 近似コンピューティングと電源電圧スケーリングの統合
DNNモデルの学習には32ビット浮動小数点演算(FP32)が用いられてきた。近年、高い計算効率とメモリ効率が実現できる16ビット浮動小数点演算に注目が集まっているが、実世界の新しいデータセットに適用することは難しい。本論文では、適応的な演算ビット幅と電圧のスケーリングを利用したエネルギー効率の高いDNN学習を提案した。ビット幅が設定可能な浮動小数点演算器を用いて、低電圧動作の小さな仮数ビット幅（FB）から学習を始め、学習品質に応じてFBと電圧を徐々に増加させる。実験の結果、FP32を用いた場合と同等の学習品質を達成しつつ最大で62%のエネルギー削減が可能であること、最小のFBを用いた学習と比較しても最大で37%のエネルギー削減が可能であることを示している。

本論文は、近似コンピューティングと電源電圧スケーリングを適用することで高エネルギー効率なDNN学習を実現するVLSI設計技術を提示している。本論文が提案する技術は、ハイエンドGPUから専用学習エンジン設計まで適用可能で、広い応用範囲を持つ。従って，博士（情報科学）の学位論文として価値のあるものと認める。