様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ５ 月 ２０ 日現在

機関番号：１５４０１
研究種目：基盤研究(C)（一般）
研究期間：2020～2022
課題番号：２０Ｋ１１７３５

研究課題名（和文）回路シミュレーションによるGPU向け超並列機械学習計算環境の構築

研究課題名（英文）Development of a Massively Parallel Machine Learning Environment for GPUs using Circuit Simulation

研究代表者
伊藤　靖朗（Ito, Yasuaki）

広島大学・先進理工系科学研究科（工）・教授

研究者番号：４０３９７９６４

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：回路を用いた機械学習計算ではビットレベルの演算を行う手法が提案され,認識精度の低下を抑えることにより高速でかつ高精度な計算を実現している．本研究では,超並列計算のアイデアを用いた機械学習計算の高速化を目指し,回路シミュレーションとビット並列化を組み合わせた手法を提案した.具体的な成果として畳み込みニューラルネットワークを対象に,逐次計算に比べて最大300倍の高速化を実現した．さらに，ネットワークモデルの圧縮も行い，さらなる実行時間の削減を目指した．この手法は既存のネットワーク圧縮手法と比べて精度低下を抑えつつ高い圧縮率を達成しており，実行時間の大幅な削減が可能であることを示した．

研究成果の学術的意義や社会的意義
本研究は機械学習計算の高速化において、新たなアプローチとして回路シミュレーションとビット並列化を組み合わせた手法を提案した．従来のソフトウェアアプローチによる高速化手法とは異なる視点から，計算の高スループット化を実現した．さらに，ネットワークモデルの圧縮により，実行時間の削減という観点からも新たな機械学習計算の高速化手法の提案を行った．機械学習は現代社会において重要な役割を果たしている一方，その高い性能を実現するためには大量の計算リソースが要求される．本研究の成果により，既存のGPUをより効率的な利用が可能となり様々な分野での研究や実用化が進むことが期待される．

研究成果の概要（英文）：In machine learning computation using circuits, methods that perform bit-level operations have been proposed to achieve high-speed and high-accuracy computation by minimizing the degradation of recognition accuracy. In this study, we proposed a method that combines circuit simulation and bit parallelization, aiming to accelerate machine learning computation using the idea of massively parallel computation. We achieved a speedup of up to 300 times faster than sequential computation for convolutional neural networks. In addition, we also performed network model compression to further reduce execution time. Compared to existing network compression methods, this method achieves a high compression ratio with minimal loss of accuracy, and we have shown that it is possible to significantly reduce the execution time.

研究分野：高性能計算

キーワード：並列計算　GPU　機械学習　回路シミュレーション

様　式　Ｃ－１９，Ｆ－１９－１，Ｚ－１９（共通）

１．研究開始当初の背景

現在の GPU(Graphics Processing Unit)は，グラフィックス以外の汎用計算に利用する GPGPU の研究が盛んに行われている．現在の GPU は多数の演算コアと単一の命令スケジューラから成るマルチコアプロセッサが共有メモリで複数接続する構造になっており，現在の汎用 CPU や GPU では，基本的に 32 ビットや 64 ビット単位で演算などの処理を実行する (簡単のため，以降 64 ビットで説明する)．これは加算や乗算などの演算命令の基本となっており，64 ビット単位として扱う必要がある．一方，論理演算命令はビット毎の演算，つまり，64 回のビット演算を 1 命令で計算する．つまり，論理演算命令を 1 回の論理演算で 64 並列の論理演算が可能な並列演算命令と考えることができる．一方で，現在，深層学習を用いた機械学習計算のハードウェア実装に関する研究が盛んに行われている．これらの研究では，ハードウェアでの効率的に計算を行うために内部の計算を 2 値や 3 値といった少数のビットサイズで行う手法が提案されている．これは通常ネットワークの計算で用いられている浮動小数点数演算から二値演算に計算精度を下げることで，回路の簡単化とメモリの削減を図るものである．計算精度を下げるためネットワーク自体の性能である認識精度が低下するので，これらの研究では計算精度を低下しても，認識精度が低下しないように様々な工夫が提案されている．既存のハードウェアを用いた機械学習計算では，ソフトウェアアプローチの研究成果に対して，精度をできるだけ落とさないように専用ハードウェアとして回路化する研究が主流であった．

２．研究の目的

本研究では，上記背景をもとに，ソフトウェアアプローチによる機械学習計算の高スループット化を目的とする．深層学習の機械学習計算では学習時に複数の入力に対して学習することで，高精度な学習を実現するバッチ学習手法が広く用いられている．また，学習後のネットワークの利用を考えると，大量の入力に対して処理することを要求されることが多い．このような機械学習計算の現状と，大量の入力に対する計算であるバルク実行をターゲットとしたビット並列化手法はマッチしている．そこで本研究では，ハードウェア実装の機械学習計算をビットレベル並列化によるバルク実行に応用し論理演算命令を用いてシミュレートすることで，ハイスループットな機械学習計算を実現することを目的とする．この手法は元々多数のコアを搭載し，高メモリ帯域である GPU をターゲットとしており，GPU での機械学習計算の高速化を当初の目標とする．さらにこの手法を既存のソフトウェア実行環境に適用することで，ハードウェアの追加や変更なしで機械学習計算の高速化の実現を目指す．

３．研究の方法

ディープニューラルネットワークによる機械学習計算の高速化の研究はソフトウェア，ハードウェア，ともに盛んに行われている．特に現在の機械学習計算の発展の要因となったGPUでは，現在，機械学習計算をターゲットとし，浮動小数点演算の精度を倍精度(64 ビット)から単精度(32 ビット)，さらに半精度(16 ビット)とビット精度を下げることでスループットの向上が行われてきた．本研究課題である回路シミュレーションの超並列計算のアイデアを用いた機械学習計算は，それを 1 ビット単位で並列に計算を行う．さらに提案手法では，既存のソフトウェアアプローチと異なり，シミュレートする回路サイズが実行時間に直接的に影響することに注目し，対象となるネットワークモデルの圧縮を行った．通常，学習済みのネットワークモ

デルに対して行われるネットワークプルーニングだが，精度を維持したまま多くの重みを削除することは困難であるという課題がある．そのため，本研究では実行時間の削減の観点からネットワーク圧縮を行う．

４．研究成果

本研究の提案手法である回路シミュレーションによる超並列計算を具体的な機械学習ネットワークモデルである畳み込みニューラルネットワークに適用した．具体的には畳み込み層の計算を回路として論理演算に変換し，各論理演算を同時に GPU のマルチコアプロセッサに割当て，各マルチコアプロセッサ内でもビット並列演算を行うことで計算の並列化を実現した．その結果，逐次計算に対して最大 300 倍の高速化を達成して，提案手法の有用性を示した．さらに提案手法では，既存のソフトウェアアプローチと異なり，シミュレートする回路サイズが実行時間に直接的に影響することに注目し，対象となるネットワークモデルの圧縮を行った．通常，学習済みのネットワークモデルに対して行われるネットワークプルーニングだが，精度を維持したまま多くの重みを削除することは困難であるという課題がある．そのため，本研究では「未学習モデルから開始した方が，プルーニングの効果が高くなる」という盆栽仮説を立て，シンプルで効率的なチャネルプルーニングアルゴリズムを提案している．このアルゴリズムは従来学習後に精度に影響のない重みを削減する手法が一般的であるが，本手法では学習をする過程でネットワークを圧縮する点が既存手法と大きく異なる．本手法は任意の構造のネットワークに適用可能で，実験では CIFAR-10 と CIFAR-100 をターゲットに VGG16 に対して未学習モデルと学習済みモデルのプルーニングを行った結果，未学習モデルは既学習モデルよりも多くのチャネルを削減できることがわかった．具体的には，CIFAR-10 に対してネットワークの重みを 97.9%削減，CIFAR-100 に対してネットワークの重みを 84.0%削減することに成功した．表 1 に既存のネットワーク圧縮手法との比較を示す．表からもわかる通り，提案手法は精度低下を抑えたまま大幅な圧縮率を達成していることがわかる．ネットワークの重みの削減分が直接実行時間の短縮に相当することより，このネットワーク圧縮手法と組み合わせることにより，本提案手法である回路シミュレーションの実行時間を大幅な削減可能であることを示した．

表1: Performance comparison of different channel pruning methods to prune the VGG16 models

	CIFAR-10		CIFAR-100	
	Test accuracy	Remaining weights	Test accuracy	Remaining weights
Ref. [1]	92.7%	16.0%	72.0%	35.2%
Ref. [2]	93.1%	11.5%	71.6%	24.0%
Ref. [3]	93.8%	11.5%	73.5%	24.9%
Ref. [4]	92.6%	26.6%	73.3%	62.1%
This work	92.6%	2.1%	70.3%	16.0%

[1] Hu, Y., Sun, S., Li, J., Wang, X., Gu, Q.: A novel channel pruning method for deep neural network compression. ArXiv abs/1805.11394 (2018)

[2] Li, H., Kadav, A., Durdanovic, I., Samet, H., Graf, H.P.: Pruning filters for efficient ConvNets. In: 5th International Conference on Learning Representations. OpenReview.net (2017), https://openreview.net/forum?id=rJqFGTslg

[3] Liu, Z., Li, J., Shen, Z., Huang, G., Yan, S., Zhang, C.: Learning efficient convolutional networks through network slimming. In: Proc. of the IEEE international conference on computer vision. pp. 2736–2744 (2017)

[4] Zhao, C., Ni, B., Zhang, J., Zhao, Q., Zhang, W., Tian, Q.: Variational convolutional neural network pruning. In: Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2780–2789 (2019)

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 Matsumura Naoki、Ito Yasuaki、Nakano Koji、Kasagi Akihiko、Tabaru Tsuguchika	４．巻 －
２．論文標題 A novel structured sparse fully connected layer in convolutional neural networks	５．発行年 2021年
３．雑誌名 Concurrency and Computation: Practice and Experience	６．最初と最後の頁 －
掲載論文のDOI（デジタルオブジェクト識別子） 10.1002/cpe.6213	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計0件

〔図書〕　計0件

〔産業財産権〕

〔その他〕

－

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	中野　浩嗣 (Nakano Koji) (30281075)	広島大学・先進理工系科学研究科（工）・教授 (15401)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関