様式3

論 文 内 容 の 要 旨

氏 名 （ 李彦辰 ）	
論文題名	Accelerating Computing-Intensive Applications by Pruning with GPU-Based Parallelization (計算集中型応用の枝刈りおよびGPUに基づく並列化による高速化)

論文内容の要旨

Computationally intensive applications present significant challenges for real-world deployment, necessitating effective acceleration strategies. Such acceleration can be achieved through two distinct strategies: algorithmic optimization and hardware-level enhancement. Pruning represents a fundamental algorithmic approach that reduces computational complexity by systematically eliminating redundant calculations, while graphics processing units (GPUs) are powerful hardware accelerators capable of parallelizing regular and repetitive operations. However, each approach presents distinct challenges: pruning techniques may compromise output quality, while GPU implementations require careful optimization to mitigate computational overhead.

This dissertation proposes two complementary approaches for integrating pruning with GPU-based parallelization to address these challenges. The first approach enhances pruning and GPU parallelization separately as independent components. The second approach cooperates pruning with GPU parallelization to minimize the overhead in parallel computations. Our analysis shows that the first approach is particularly effective for out-of-kernel pruning, where pruning maintains consistent data patterns for parallelization by eliminating entire kernel function calls. Conversely, the second approach proves more suitable for in-kernel pruning, where pruning modifies the underlying data patterns for parallelization in each kernel function call. These systematic approaches provide clear guidelines for selecting appropriate acceleration strategies based on specific pruning strategies.

We validate the first approach by applying it to accelerate best equivocation code (BEC) generation, a computing-intensive problem adaptive to out-of-kernel pruning. Traditional BEC generation uses a sequential algorithm with two main components: pruning to reduce the search space, and evaluation to select optimal BECs from candidates. For the algorithmic enhancements, we propose two methods: a dynamic programming (DP) method and a greedy method. The DP method achieves lower time complexity than previous methods by efficiently reusing intermediate evaluation data. The greedy method further improves upon the DP method by enhancing the pruning component, reducing the search space while maintaining comparable equivocation rates. We then parallelize the evaluation component, which is the primary bottleneck, on GPUs. Experimental results demonstrate that the proposed DP and greedy methods reduce the sequential generation time to a quarter. Moreover, the BEC generation enhanced by the GPU achieved 17-fold acceleration compared to the its CPU version.

We validate the second approach by applying it to accelerate DNN inference with fine-grained pruning. Fine-grained pruning represents an in-kernel approach that facilitates efficient GPU memory access by removing less important elements from weight matrices in structured patterns. However, this method faces two typical challenges of in-kernel pruning: increased pruning loss from structured pattern constraints and computational load imbalance from unevenly pruned workloads. To address these challenges, we propose two methods: TileTrans and adaptive tile pruning (ATP). TileTrans reduces pruning loss through strategic weight matrix transformation, while ATP ensures balanced computational workloads and efficient sparse matrix multiplication. The experimental results indicate that TileTrans can improve the accuracy of the pruned BERT-Base model by up to 5.7¥% on the question-answering natural language inference (QNLI) task. Meanwhile, results also show that ATP-accelerated models achieve superior speedup compared to previous methods while preserving inference accuracy.

This dissertation presents two complementary approaches for accelerating computing-intensive applications by pruning with GPU-based parallelization. Through rigorous empirical validation, we demonstrate the effectiveness of these approaches in two typical applications: BEC generation and DNN inference. The first approach accelerates BEC generation by separately enhancing the pruning and GPU-based parallelization processes, while the second approach accelerates DNN inference by cooperating pruning with GPU-based parallelization to eliminate the load imbalance. These results establish a theoretical and practical foundation for accelerating computing-intensive applications through the synergistic combination of algorithmic and hardware-level acceleration techniques.

様式 7

論文審査の結果の要旨及び担当者

氏　名	（　　李　彦辰　　）

	（職）	氏　名
論文審査担当者		
主 査	教 授	伊野　文彦
副 査	教 授	増澤　利光
副 査	教 授	中島　悠太

論文審査の結果の要旨

本学位論文は，Graphics Processing Unit（GPU）を装備する計算機において，枝刈りによる効率のよい計算およびGPU上の高速な並列計算を両立するための高速化技術をまとめたものである．GPUシステム向けの高速なプログラム設計を明らかにすることを目的として，以下の3つの研究成果を得ている．

１．最適な曖昧度符号を生成するための並列ヒューリスティック手法
学位論文の２章では，最適な曖昧度符号を高速に生成するために、動的計画法に基づく符号生成手法を提案し、枝刈り手法を併用することにより時間計算量を削減する貪欲法を提案している。評価実験の結果、提案手法の実行時間が時間計算量とともに削減されていることを示し、マルチコアCPUおよびGPUにおいて演算コア数の増大とともに性能を向上できる線形な速度向上を実現している。また、貪欲法は従来手法よりも高い曖昧度を持つ符号を得られることを示している。

２．深層ニューラルネットワークを効果的に枝刈りするための再パラメータ化手法
３章では、深層ニューラルネットワーク（DNN）における細粒度枝刈り損失を低減するための再パラメータ化手法を提案している。提案手法は、DNNの重み行列を変換することにより枝刈り損失を低減する。本章では、提案手法が枝刈り損失を低減できることを理論的に証明している。ResNet-34やBERT-BaseなどのDNNモデルを用いて提案手法を評価した結果、提案手法が枝刈り損失を低減することによりDNNの精度を向上できることを確認している。具体的には、ResNet-34およびBERT-Baseの精度がそれぞれ最大6.2%および5.4%向上し、提案手法の有用性を示している。

３．GPUにおける効率的な推論のための適応的なタイル枝刈り
４章では，細粒度枝刈りを施したDNNにおける推論を効率化するための手法を提案する。提案手法は、従来手法よりも低い枝刈り損失のもとで疎行列積の計算負荷を均等化する。計算負荷を均等に分散させた疎行列を構築し、GPU上で疎行列積を効率的に計算する。さらに、計算負荷が不均衡になる得る枝刈り済みDNNの推論時間を高速化する。DNNモデルとしてResNet-34およびBERT-Smallを用いた評価実験の結果、提案手法は従来手法と比較して枝刈りによる損失を抑えつつ、計算負荷を分散でき、両モデルを用いた推論をおよそ40%高速化できることを示している。

以上のように、本学位論文で得られた研究成果は、GPUシステムにおいて効率のよい枝刈りおよび並列計算を両立する際に役立つものであり、ひいては安全な符号の開発や高性能かつ高精度な人工知能技術の発展に寄与する。よって、本論文は博士（情報科学）の学位論文として価値のあるものと認める。