論文 / 著書情報
Article / Book Information

題目(和文)	GPU を用いたコンピュータビジョンの高速化と最適化の研究
Title(English)	GPU-Based Acceleration and Optimization Research on Computer Vision
著者(和文)	WANGCHENYU
Author(English)	Chenyu Wang
出典(和文)	学位:博士(理学), 学位授与機関:東京工業大学, 報告番号:甲第12513号, 授与年月日:2023年9月22日, 学位の種別:課程博士, 審査員:遠藤 敏夫,坂本 龍一,髙邉 賢史,脇田 建,佐藤 育郎
Citation(English)	Degree:Doctor (Science), Conferring organization: Tokyo Institute of Technology, Report number:甲第12513号, Conferred date:2023/9/22, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

系・コース: Department of, Graduate major in	数理・計算科学 数理・計算科学	系 コース	申請学位(専攻分野): Academic Degree Requested	博士 Doctor of	(理学)
学生氏名: Student's Name	WANG CHENYU		指導教員(主): Academic Supervisor(main)		遠藤敏夫
			指導教員(副): Academic Supervisor(sub)		

要旨(英文 800 語程度)
Thesis Summary （approx.800 English Words ）

Research Background and Problem Statement:

Computer vision, a vibrant subfield of artificial intelligence, is transforming the way machines perceive and interpret the visual world. The complexity and computational requirements of vision-related tasks necessitate robust hardware resources and efficient software implementations. Despite significant advances, the intricate computations involved often exceed the processing capabilities of standard Central Processing Units (CPUs). As a solution, Graphics Processing Units (GPUs), with their parallel processing capabilities, have emerged as a promising platform for speeding up computer vision tasks. Nevertheless, leveraging GPUs for computationally intensive tasks, such as object detection and image classification, requires optimized models and algorithms specifically tailored for these parallel environments, a task that presents its own set of challenges.

Research Objective:

The primary objective of this dissertation is to address these challenges by focusing on optimizing computer vision model with GPU, to achieve more effective performance trade-off of the models, particularly optimizing object detection and transformer-based models. The research aims to enhance the detection speed of the Single Shot MultiBox Detector (SSD), a commonly used object detection model, and optimize the Swin Transformer, which is a transformer-based model known for its high performance but also for its computational intensity, both architectural and computing optimization. The scope of the study extends to proposing and evaluating strategies for enhancing computational efficiency, ultimately contributing to the broader aim of advancing the field of computer vision.

Research Content and Results:

In the first research, the study reveals a 22.53% performance enhancement of the Single Shot MultiBox Detector (SSD) model for object detection tasks by efficiently adapting its computations for GPUs. Despite the speed-up in the feature extraction layer being less pronounced, it plays a crucial role in accelerating the overall SSD512 framework, underlining the feasibility of using GPUs to boost performance in object detection frameworks. Our approach concentrates on optimizing the post-processing stage by enhancing the Non-Maximum Suppression (NMS) algorithm, a key component in object detection. We have re-engineered this algorithm to effectively suit object detection tasks, achieving a considerable reduction in processing times. The main innovation here is our use of shared memory to overcome the scalability limitations imposed by the traditional iterative clustering procedure, which dramatically speeds up object detection by minimizing data dependencies that tend

to serialize computations.

The second research introduces the Pyramid Swin Transformer to optimize the Swin Transformer, an architecture known for its high performance but also its computational intensity. Despite the benefits, a significant problem with this architecture is the lack of sufficient information exchange between windows in large-scale dimensions. To solve this, we adjusted the original architecture to foster greater connectivity. Our model splits the attention process into smaller blocks and implements a hierarchical progression of window sizes from larger to smaller, thereby balancing computational complexity and performance, yielding significant improvements in various tasks including image classification (85.4% top-1 accuracy on ImageNet), object detection (51.6 box AP with Mask R-CNN and 54.3 box AP with Cascade Mask R-CNN on COCO), semantic segmentation (49.0 mIoU on ADE20K), and video recognition (83.4 % top-1 accuracy on Kinetics-400). Then we focus on further GPU acceleration techniques for the Swin Transformer architecture, developing innovative parallel processing strategies for window-based multi-head self-attention (W-MSA) mechanisms. We adopted two parallelizable strategies to harness the full computational power of modern GPUs, which is particularly effective in speeding up the window-based multi-head self-attention mechanism, called 2D and 3D Parallelization Strategies. Our approaches yield substantial speed-ups and stability across increasing computational loads, effectively mitigating bottlenecks in standard PyTorch implementations. Our method breaks down the process into separate parallelizable subtasks, which contribute to the efficient computation of window-based multi-head self-attention. Extensive experiments are carried out to measure the execution times of these implementations under different conditions. In all cases, our 2D Parallel emerges as the top performer, showcasing an impressive speed-up of around 8.57x faster execution for a 64*64 input size, 8 head number with an 8*8 window size.

Conclusion:

This research underscores the advantageous trade-offs GPUs can accomplish in optimizing performance for computer vision models. By leveraging GPUs, our study was able to accelerate the Single Shot MultiBox Detector (SSD) target detection model, striking a favorable balance between precision and speed. Our investigation into the Swin Transformer method involved an initial optimization of the framework to enhance detection performance. This was subsequently followed by the acceleration of the Window-based Multi-head Self-Attention (W-MSA) using GPU, anticipating it would yield a more effective trade-off between speed and accuracy. The findings of this study present a promising outlook for the future of GPU-accelerated computer vision tasks, advocating for further exploration and application of these innovative techniques within this challenging field.