論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Bi-directional Exploration of High-performance Computing and Machine Learning: Mutual Benefits
著者(和文)	郭 箭
Author(English)	Jian Guo
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11033号, 授与年月日:2019年3月26日, 学位の種別:課程博士, 審査員:松岡 聡,遠藤 敏夫,額田 彰,脇田 建,横田 理央
Citation(English)	Degree:Doctor (Science), Conferring organization: Tokyo Institute of Technology, Report number:甲第11033号, Conferred date:2019/3/26, 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

専攻： Department of	数理・計算科学　専攻	申請学位（専攻分野）： Academic Degree Requested	博士　（　理学　） Doctor of
学生氏名： Student's Name	GUO Jian	指導教員（主）： Academic Supervisor(main)	特任教授　松岡 聡
		指導教員（副）： Academic Supervisor(sub)	

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

High-Performance Computing (HPC) puts together multiple computer technologies like computer architecture, algorithms, programming and operating system to meet increasing requirements in computing power and speed. HPC is used to effectively and quickly handle complex computational problems. Machine Learning is a sub-discipline of artificial intelligence (AI), that is closely related to mathematical optimization as well as computational statistics. HPC and Machine Learning are two independent fields of computer science respectively, which have been developed in their own fields for decades. They have achieved great results and have made great contributions to the academia and industry. In recent years, we can see that HPC and Machine Learning benefit each other rather than profiting unilaterally. There is a mutually beneficial relationship between the HPC and Machine Learning on some topics. In this thesis, we extend and strengthen the mutual benefit between HPC and Machine Learning to broader fields by bi-directional research exploration: speeding up Machine Learning with HPC's help, and benefiting the productivity and efficiency of HPC by use of Machine Learning.

For improving the overall efficiency of Machine Learning, we focus on accelerating the feature extraction and the training part in Machine Learning process. For acceleration of feature extraction, we speed up the feature extraction of bioacoustics sound data by employing a Just-in-Time (JIT) compiler technique, Numba, which implements automatic parallelization and performs other optimizations, as well as a GPU-accelerated library going by the principle of modifying Python-based baseline as little as possible. With small modifications in the baseline code of feature extraction, our optimized feature extraction yields maximum 86x speedup over the baseline without losing programming efficiency. In order to reduce the training time-cost of Machine Learning modeling, we speed up the training of Stacked Sparse Autoencoder (SSAE) with a soft-max layer for bird sounds classification by use of HPC hardware and software. Our experiment shows that using GPUs (K20, P100) results in up to 10x speed-up compared to CPU when training with different feature dimensions of bird sounds.

For helping HPC by improving the overall productivity and efficiency of HPC systems, we propose a Machine Learning based data-driven approach to predict runtime-underestimated jobs in HPC systems by learning complex hidden patterns between different HPC applications and different features of computing resource usage, as well as user behavior from job logs. The experiment results show that our prediction models are able to predict runtime-underestimated jobs with different accuracy at different checkpoints times in HPC systems, which would benefit HPC users by reducing time and monetary loss. It also helps HPC systems free up computing resources to a certain extent and allows users and administrators take the appropriate action (to terminate those on-going jobs that are predicted to be runtime-underestimated). All in all, our Machine Learning-based prediction model would be able to improve the overall productivity and efficiency of the HPC system. In the preliminary simulation, we evaluate benefits of the prediction model with a metric named Saved-Lost Rate (SLR), most of SLRs are around 0.05-0.12, 0.337 is the best one. If we set the checkpoints for different applications to their best points, we can estimate that our prediction model would save 24962 hours totally based on the result of preliminary simulation from the existing database.

All in all, in this thesis, we perform the bi-directional research exploration between HPC and Machine Learning: speeding up Machine Learning with HPC's help; benefiting the productivity and efficiency of HPC by use of Machine Learning. Through our three use cases, we can clearly conclude that HPC hardware and software approaches are able to supply powerful computing resource for acceleration and scaling of Machine Learning. Meanwhile, Machine Learning can benefit HPC in improving the overall productivity,

efficiency and etc. This is the role of interaction between HPC and Machine Learning: mutual benefit and promotion.

備考 ： 論文要旨は、和文 2000 字と英文 300 語を 1 部ずつ提出するか、もしくは英文 800 語を 1 部提出してください。

Note : Thesis Summary should be submitted in either a copy of 2000 Japanese Characters and 300 Words (English) or 1copy of 800 Words (English).

注意：論文要旨は、東工大リサーチリポジトリ(T2R2)にてインターネット公表されますので、公表可能な範囲の内容で作成してください。

Attention: Thesis Summary will be published on Tokyo Tech Research Repository Website (T2R2).