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　 ６ 月　 １ 日現在

機関番号：１４５０１
研究種目：若手研究
研究期間：2020～2022
課題番号：２０Ｋ１９８２３

研究課題名（和文）Scaling up CNN computations for data-intensive scientific applications

研究課題名（英文）Scaling up CNN computations for data-intensive scientific applications

研究代表者
ＨＡＳＣＯＥＴ　ＴＲＩＳＴＡＮ（HASCOET, TRISTAN）

神戸大学・経営学研究科・助教

研究者番号：６０８４８４４８

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：この研究では、様々な科学的およびエンジニアリングの問題に対応するために新たなアルゴリズムを開発しました。
深層学習モデルの計算コストを改善する方法を考案しました。アルゴリズムの効果を示すために、神経科学、生物多様性監視、材料科学の分野を跨いだ幅広い科学的応用例でそれらをベンチマークしました。
計算効率の具体的な改善だけでなく、本研究では研究は、異分野間の協力と革新のための新たな道を開いています：イスラエルのスタートアップが視覚モデルのプロトタイプ開発に本研究の成果を使用し、パリ天文台とのパートナーシップで、水資源管理と持続可能性の理解を深めるための効率的な水文学的モデルを開発してきました。

研究成果の学術的意義や社会的意義
深層学習は、現行の技術では手が届かないとされていた技術的な課題への解決策を開いてきましたが大量の計算力と高額なインフラが必要となるため、技術の開発と応用は大規模な技術機関内に大きく集中しています。したがって、計算効率を改善し、その応用の利益を広範囲の人々に広げることが必要となっています。本研究では、限定的な計算力で展開できる技術を開発し、その適用性を複数の実用的な科学的な応用例を通じて示しました。

研究成果の概要（英文）：In this research, our objective has been to develop new computational tools that can be applied to tackle diverse scientific and engineering problems. We have focused our efforts on devising methods to improve the resource consumption of deep learning models.
To demonstrate the effectiveness of our algorithms, we have benchmarked them on a wide array of scientific applications across the fields of neuro-science, biodiversity monitoring and material science.
Beyond the tangible improvements in the computational efficiency, our work has also opened up new avenues for interdisciplinary collaboration and innovation: Our software has been used to help an Israeli startup prototype low-level vision models and, in partnership with the Paris Observatory, we have been developing efficient hydrological models to improve our understanding of water resource management and sustainability.

研究分野：機械学習・深層学習

キーワード：Deep Learning 4 science　Computational Efficiency　Computer Vision　ConvNets

１．研究開始当初の背景

The fourth paradigm of science [1] refers to the idea that the amount of scientific data collected is increasing in scales much larger than scientists can individually make sense of. Hence, the ability of our society to tackle future scientific challenges will depend on our ability to build information systems that enable scientists to make sense of this data. Ten years after the formulation of *"the fourth paradigm"* idea, progress in sensing and storage technologies have enabled various scientific communities to collect and store scientific data on a new scale. Furthermore, in the past few years computer vision (CV) and machine learning (ML) technologies have really begun to unlock key scientific breakthroughs: In late 2017 the Celeste project [2] has applied CV and ML to extract a complete catalogue of celestial objects in the visible universe from 178 terabytes of telescopic images. In 2018, the AlphaFold [3] project leveraged recent progress on deep reinforcement learning to provide unprecedented progress on the problem of protein folding, a cornerstone for biological research.

These recent successes are illustrative of a deeper underlying phenomenon: As the scale of scientific data collection increases, pattern recognition technologies will play an increasingly important role in future scientific discoveries. Foreseeable applications of CV and ML to scientific challenges include, among others, the mapping of the human brain's connectome [4], scaling up physical simulations for the development of energy systems [5], material discovery [6,7] and large-scale climate modeling [5]. At large, recent scientific breakthroughs have been increasingly powered by advances in information technologies: Advances in sensing and storage technologies have enabled various scientific communities to collect very large amounts of data, while progress in machine learning and pattern recognition techniques have enabled to extract key information from this large scale data. However, recent progress in machine learning have been enabled by an exponential growth in computation power. In particular, the advent of Graphical Processing Units (GPU) in the last ten years have unlocked the potential of Deep Learning models for Computer Vision. As the computational cost of these models continue scaling up, access to these models becomes increasingly more expensive to that the development and application of the best performing models become increasingly concentrated in large technological institutions.

In this context, providing computational tools enabling the development of such models on more modest infrastructure becomes increasingly important to the democratization and widespread benefit of these technological benefits to a wider audience.

２．研究の目的

The purpose of this project has been two-folds: First, we have aimed to contribute such toolset: our goal is to scale up the computational efficiency of Computer Vision models for data-intensive scientific applications. Second, our goal is to find practical and tractable technological challenges to address with this toolset.

Regarding our first goal, we identify two axis optimization to improve the resource consumption cost of large deep learning models: The computational efficiency (i.e.; reducing the computation time needed to run one iteration of the iterative training algorithm) and the algorithmic efficiency (i.e.; reducing the number of iterations needed by the optimization algorithm to train the model to convergence).

(1) Computational optimization

This project has focused on optimizing the memory consumption of Convolutional Neural Network training and inference memory consumption, as well as computational efficiency through single node performance tuning.

(2) Algorithmic optimization

The convergence rate of first-order iterative training algorithms like SGD is limited by the impact of second order phenomenons. Hence, understanding the curvature of the loss landscape along the optimization trajectory is key to optimize the algorithmic efficiency. My students and I have focused on developing tools allowing to visualize and investigate the structure of curvature along the training trajectories followed by SGD.

Regarding our second goal of practical applications, we have aimed to both optimize the performance of computation workloads of models for scientific tasks that we had already identify, as well as find new

innovative applications to such workloads. These applications are described in more details in the following section.

３．研究の方法

This project has been conducted in tight collaboration with Professor Takiguchi's laboratory, and a number of academic and industrial partners. Professor Takiguchi has entrusted me with the supervision of six graduate students from his laboratory. Together, we have defined a number of tasks differentiated into core technical contributions and applicative benchmarks.

Core technological tasks:

(1) Single Node performance tuning: Investigation of workload-specific task-specific single node performance tuning using the TVM framework.

(2) Training memory resource optimization: Development of algorithm allowing to reduce the memory consumption of training large CNN so as to alleviate the GPU memory bottleneck of training CNN on limited hardware

(3) Inference memory resource optimization: Development of algorithm allowing to reduce the memory consumption of CNN inference so as to allow processing of large gridded data on limited hardware.

Applicative tasks: The following tasks have been defined in order to benchmark the efficiency of the developed algorithms methods.

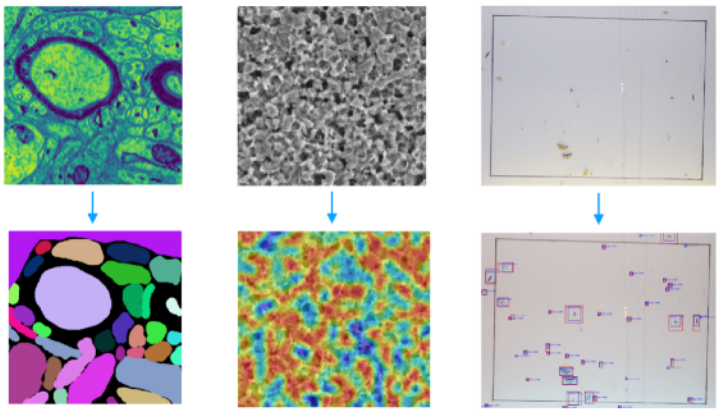

Figure 1: Illustration of the three original applications considered for this research. (Left): Neuron segmentation, illustrated in 2D for convenience. (Middle) Copper surface analysis using weakly supervised image segmentation. (Right): Insect population monitoring using object detection models. These three problems involve processing large datasets of very high resolution images.

- *Neuroscience*: Observing neural tissues at the nano metric scale is required to accurately map neural connections. To do so, terabytes of electron microscopy imaging of neural tissues at a resolution of a few nano-meters need to be collected from which individual neuron connections are extracted. Automating such extraction process is needed to map large neural structures.

- *Material Science*: Modeling the structure-property relationship of materials often require analyzing vast amount of observations of the microstructure of their surfaces. In this project, we aim to leverage our computational tool to analyse the bonding strength of copper surfaces used in the semi-conductor industry.

- *Ecosystem monitoring*: At the bottom of the food chain, insect populations play a crucial role in sustaining forest ecosystems. Our goal is to develop a system to monitor the population of insects based on a light trap imaging.

While the three applications described above had been identified at the beginning of this project, two additional applications have been found during the course of this research project:

- *Hydrology:* Vast amount of climatic data are analyzed to infer and predict the movement of water resources. We have developed ML models for both monthly evaporation estimation and river discharge modeling.

- *Low-level computer vision:* Recording videos in either very high speed or low luminosity settings require heavy deep learning denoising, which is computationally complex to process with hard computational and latency constraints for real-time usage. We investigate the application of our computational tools to this additional use case.

The above projects all share the same computational constraint to train and evaluate large deep learning models on large quantities of data within limited computational resources, which makes them suitable benchmarks to evaluate our developed algorithms.

４．研究成果

We have focused our efforts on devising methods to improve the efficiency and performance of deep learning models, thus reducing their computational burden. To demonstrate the effectiveness of our algorithms, we have benchmarked them on a wide array of scientific applications across the fields of neuro-science, biodiversity monitoring and material science.

Our work has not only led to tangible improvements in the computational efficiency of deep learning models but has also opened up new avenues for interdisciplinary collaboration and innovation. Our software has been used to help an Israeli startup prototype low-level vision models. Furthermore, in partnership with the Paris Observatory, we have been developing efficient hydrological models that can significantly improve our understanding of water resource management and sustainability.

In terms of academic publications, core methods derived in this research project have been presented in dedicated publications for training memory optimization [8] and inference memory optimization [9]. On the practical scientific application side, we have proposed a number of high performance models for applications across hydrology [10,11], material sciences [12,13], high speed imaging [14,15], and infrastructure monitoring [16].

References

[1] Tansley, Stewart, and Kristin M. Tolle. The fourth paradigm: data-intensive scientific discovery. Ed. Anthony JG Hey. Vol. 1. Redmond, WA: Microsoft research, 2009.

[2] Regier, Jeffrey, et al. "Cataloging the Visible Universe through Bayesian Inference at Petascale." 2018 IEEE International Parallel and Distributed Processing Symposium (IPDPS). IEEE, 2018.

[3] De novo structure prediction with deep-learning based scoring. DeepMind, Assessment of Techniques for Protein Structure Prediction (Abstracts) 1-4 December 2018.

[4] Seung, Sebastian. Connectome: How the brain's wiring makes us who we are. HMH, 2012.

[5] Rolnick, David, et al. "Tackling Climate Change with Machine Learning." arXiv preprint arXiv: 1906.05433 (2019).

[6] Jain, Anubhav, et al. "Commentary: The Materials Project: A materials genome approach to accelerating materials innovation." Apl Materials 1.1 (2013): 011002.

[7] Agrawal, Ankit, and Alok Choudhary. "Deep materials informatics: Applications of deep learning in materials science." MRS Communications (2019): 1-14.

[8] Hascoet, T., Febvre, Q., Zhuang, W., Ariki, Y., & Takiguchi, T. (2023). Reversible designs for extreme memory cost reduction of CNN training. EURASIP Journal on Image and Video Processing, 2023(1), 1-30. SpringerOpen.

[9] Zhuang, W., Hascoet, T., Chen, X., Takashima, R., Takiguchi, T., & Ariki, Y. (2021). Convolutional Neural Networks Inference Memory Optimization with Receptive Field-Based Input Tiling. APSIPA Transactions on Signal and Information Processing, 12(1),

[10] Hascoet, T., Yoshimi, K., Dossa, R., & Takiguchi, T. (2023). Optimizing River Discharge Forecasts with Machine Learning for Japanese Public Dams Operation. 国民経済雑誌= Journal of economics & business administration, 227(2), 45-64. 神戸大学経済経営学会.

[11] Hascoet T, Pellet V., Aires F., Learning global evapotranspiration corrections from a water cycle closure supervision, Journal of Hydrology (Under review).

[12] Zhuang, W., Hascoet, T., Takashima, R., & Takiguchi, T. (2022). Optical Flow Regularization of Implicit Neural Representations for Video Frame Interpolation. arXiv preprint arXiv:2206.10886.

[13] Zhuang, W., Hascoet, T., Takashima, R., & Takiguchi, T. (2022). Learn to See Faster: Pushing the Limits of High-Speed Camera with Deep Underexposed Image Denoising. arXiv preprint arXiv:2211.16034.

[14] Hascoet, T., Zhang, Y., Persch, A., Takashima, R., Takiguchi, T., & Ariki, Y. (2020). FasterRCNN monitoring of road damages: Competition and deployment. In 2020 IEEE International Conference on Big Data (Big Data) (pp. 5545-5552).

５．主な発表論文等

〔雑誌論文〕　計5件（うち査読付論文　2件／うち国際共著　2件／うちオープンアクセス　4件）

１．著者名 Weihao Zhuang, Tristan Hascoet, Xunquan Chen, Ryoichi Takashima , Tetsuya Takiguchi , Yasuo Ariki	４．巻 12
２．論文標題 Convolutional Neural Networks Inference Memory Optimization with Receptive Field-Based Input Tiling	５．発行年 2023年
３．雑誌名 APSIPA Transactions on Signal and Information Processing	６．最初と最後の頁 1～20
掲載論文のDOI（デジタルオブジェクト識別子） 10.1561/116.00000015	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Tristan Hascoet, Keisuke Yoshimi, Rousslan Dossa, Tetsuya Takiguchi	４．巻 227(2)
２．論文標題 Optimizing River Discharge Forecasts with Machine Learning for Japanese Public Dams Operation	５．発行年 2023年
３．雑誌名 国民経済雑誌	６．最初と最後の頁 45～64
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
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 －

１．著者名 Weihao Zhuang, Tristan Hascoet, Ryoichi Takashima, Tetsuya Takiguchi	４．巻 2206.10886v1
２．論文標題 Optical Flow Regularization of Implicit Neural Representations for Video Frame Interpolation	５．発行年 2022年
３．雑誌名 Arxiv	６．最初と最後の頁 1～10
掲載論文のDOI（デジタルオブジェクト識別子） 10.48550/arXiv.2206.10886	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Weihao Zhuang, Tristan Hascoet, Ryoichi Takashima, Tetsuya Takiguchi	４．巻 2211.16034v1
２．論文標題 Learn to See Faster: Pushing the Limits of High-Speed Camera with Deep Underexposed Image Denoising	５．発行年 2022年
３．雑誌名 Arxiv	６．最初と最後の頁 1～20
掲載論文のDOI（デジタルオブジェクト識別子） 10.48550/arXiv.2211.16034	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Tristan Hascoet , Quentin Febvre , Weihao Zhuang , Yasuo Ariki,Tetsuya Takiguchi	４．巻 -
２．論文標題 Reversible designs for extreme memory cost reduction of CNN training	５．発行年 2022年
３．雑誌名 EURASIP Journal on Image and Video Processing	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） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

〔学会発表〕　計4件（うち招待講演　0件／うち国際学会　4件）

１．発表者名 Masato Ikegawa, Tristan Hascoet, Victor Pellet, Xudong Zhou, Tetsuya Takiguchi, Dai Yamazaki
２．発表標題 Levee protected area detection for improved flood risk assessment in global hydrology models
３．学会等名 NeurIPS 2022 Workshop on Tackling Climate Change with Machine Learning（国際学会）
４．発表年 2022年

１．発表者名 Tristan Hascoet, Victor Pellet, Filipe Aires
２．発表標題 Learning evapotranspiration dataset corrections from water cycle closure supervision
３．学会等名 NeurIPS 2022 Workshop on Tackling Climate Change with Machine Learning（国際学会）
４．発表年 2022年

１．発表者名 Keisuke Yoshimi, Tristan Hascoet, Rousslan Dossa, Tetsuya Takiguchi, Satoru Oishi
２．発表標題 Optimizing Japanese dam reservoir inflow forecast for efficient operation
３．学会等名 NeurIPS 2022 Workshop on Tackling Climate Change with Machine Learning（国際学会）
４．発表年 2022年

１．発表者名 Tristan Hascoet; Yihao Zhang; Andreas Persch; Ryoichi Takashima; Tetsuya Takiguchi; Yasuo Ariki
２．発表標題 FasterRCNN Monitoring of Road Damages: Competition and Deployment
３．学会等名 IEEE Big Data Cup Challenge 20201（国際学会）
４．発表年 2020年

〔図書〕　計0件

〔出願〕　計0件

〔取得〕　計1件

産業財産権の名称	発明者	権利者
物性値予測方法、物性値予測システム及びプログラム	赤木 雅子，ハスコエ トリスタン，トウ セツコウ	メック株式会社，国立大学法人神戸大学
産業財産権の種類、番号	取得年	国内・外国の別
特許、特許第7078944号	2022年	国内

〔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
研究協力者	ペレット　ビクター (PELLET Victor)		
研究協力者	滝口　哲也 (TAKIGUCHI Tetsuya) (40397815)	神戸大学・都市安全研究センター・教授 (14501)	

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			
フランス	LERMA, Paris Observatory			