Neural Computing and Applications (2020) 32:5531–5533
https://doi.org/10.1007/s00521-020-04887-7

EDITORIAL

Editorial introduction: special issue on advances in parallel and distributed computing for neural computing

Jianguo Chen[1] · Ahmad Salah[2]

Published online: 2 April 2020

In recent years, the popularity of neural computing (NC), machine learning (ML), and artificial intelligence (AI) has grown substantially. A lot of research was carried out in both academia and industry, and it was applied in many fields. For example, deep learning achieved superhuman performance in image classification. NC/ML/AI technologies were used very successfully to play games such as Chess, Go, Atari, and Jeopardy. In addition, many companies used AI and ML technology in areas such as health care, natural resource management, and advertisement.

Most NC/ML/AI technologies and applications require heavy use of high-performance computers and accelerators for efficient processing. Consequently, parallel computing, distributed computing, cloud computing, and high-performance computing (HPC) are key components of these systems. In scientific research and practical applications, clusters of computers and accelerators (e.g., GPUs) are routinely used to train and run various neural network models. In addition, due to time-consuming iterative training processes and massive training datasets, NC/ML/AI technologies also become a "killer application" for parallel computing, distributed computing, and HPC. The above challenges have driven much of the research in distributed and parallel computing. For example, tailored computer architectures were devised and new parallel programming frameworks were developed to accelerate NC/ML/AI models. The objective of this special issue is to bring together the parallel and distributed computing and NC/ML/AI communities to present their applications and solutions to performance issues and also to present how NC/ML/AI can be used to solve performance problems.

The range of topics covered by this special issue is broad. The papers in the special issue represent a broad spectrum of parallel and distributed computing, machine learning models, and neural network models. Papers by Zheng Xiao, Zhao Tong, and Yikun Hu et al. focused on computing task scheduling in distributed and parallel computing environments. The paper by Yuedan Chen et al. focused on the partitioning and parallelization of the general sparse matrix-sparse matrix (SpGEMM) on HPC systems, which is used as the basic kernel in many NC/ML/AI algorithms. Papers by Shuang Yang and Hao Wang focused on parallelization of ML algorithms in distributed computing environments, including mobile social networks (MSNs) and of multi-view clustering (MvC) algorithms. Papers by Xiaofeng Zou, Titinunt Kitrungrotsakul, and Keyang Cheng focused on the parallelization and performance optimization of different neural network models in distributed and parallel computing environments, including cloud computing platforms, GPU-based parallel computing systems, and HPC systems. In addition, papers by Ao Liu, Minrong Lu, Jin Zhang, and Fan Wu focused on structural optimization of neural network models. Moreover, papers by Xiaofeng Zou, Minrong Lu, Xiaoyong Tang, Titinunt Kitrungrotsakul, and Fan Wu focused on performance optimization of neural network models and their applications in various fields, such as monetary policy prediction, bioinformatics, image classification, pedestrian re-identification, fingerprint pattern recognition, and human personality classification.

The paper by Zheng Xiao et al. focused on task scheduling and virtual machine (VM) allocation in distributed computing environments and described a workload-driven coordination mechanism between virtual machine allocation and task scheduling. The datasets acquired from machine learning and deep learning applications have Markov poverty and are modeled as Markov chains to extract workload characteristic operators of

✉ Jianguo Chen
jianguo.chen@utoronto.ca

Ahmad Salah
ahmad@zu.edu.eg

[1] Department of Computer Science, University of Toronto, Toronto, ON, Canada

[2] Faculty of Computes and Informatics, Zagazig University, Zagazig, Ash Sharqiyah, Egypt

persistence, recurrence, and entropy. After a nonlinear relationship is established between the workload characteristic operators and the number of VMs, the number of VM required for computing tasks in a specific application can be accurately predicted, which further affects the performance of subsequent parallel task scheduling.

The paper by Zhao Tong et al. introduced a reinforcement learning algorithm in task scheduling in cloud computing environments. Combining the Q-learning algorithm and the Heterogeneous Earliest Finish Time (HEFT) algorithm, a task scheduling algorithm called QL-HEFT is created. The QL-HEFT algorithm is divided into two phases: a Q-learning-based task sorting phase (for obtaining an optimal order) and a processor allocation phase (using the earliest finish time strategy). In addition, the EFT allocation strategy is used to allocate the best processor to tasks with the optimal order, which can effectively reduce the makespan of the computation tasks.

The paper by Yikun Hu et al. focused on the task scheduling and energy consumption of HPC systems and proposed a reformed task scheduling method with energy consumption constraints. The authors designed a pre-allocation mechanism based on the energy consumption level and provided proof to ensure energy consumption constraints. They also verified the effectiveness of the new algorithm through simulation experiments.

The paper by Yuedan Chen et al. focused on the general sparse matrix-sparse matrix (SpGEMM)—the basic kernels in many machine learning and neural computing, and designed a partitioning and parallelization method of CSR-based SpGEMM. The partitioning of SpGEMM is optimized based on the distribution of floating-point calculations. The proposed method can make SpGEMM and the system architecture Sunway TaihuLight supercomputer match well, so that it can further achieve the load balancing and performance improvement. The work will facilitate the parallel execution of various SpGEMM kernel-based applications in HPC systems.

The paper by Ao Liu et al. focused on the performance optimization of neural networks and introduced a water wave optimization (WWO)-based memetic algorithm to determine the optimal weights of neural networks. The authors performed WWO-based global search through individual improvement and population coevolution and then enhanced local refinement capabilities by using multiple local search components. They also used the Meta-Lamarckian learning strategy to select the appropriate local search component to concentrate their computational efforts on more promising solutions.

The paper by Xiaoyong Tang et al. focused on the parallel optimization of multi-fractal detrending fluctuation analysis (MF-DFA) and its application in agricultural image processing. The authors implemented the MF-DFA program for agricultural image analysis, including modules of image preprocessing, image segmentation, local area accumulation matrix calculation, and global q-order fluctuation. They further explored the segmentation scales of each module of MF-DFA and performed a parallel optimization scheme based on an OpenMP parallel computing system.

The paper by Shuang Yang et al. focused on the security threats of mobile social networks (MSNs) and proposed a distributed key management scheme. The proposed trust-based security routing scheme judges each node's behavior based on comprehensive trust and the strength of social relationships and then validates or revokes the certificate based on the evaluation of the social relationship strength. The authors implemented a feedback scheme to update the trust value between nodes and detect more malicious nodes to improve end-to-end communication security.

The paper by Hao Wang et al. focused on the parallelization of multi-view clustering (MvC) in distributed computing environments. A parallel MvC method based on the concept factorization with local manifold learning is proposed, which is represented by parallel multi-view concept clustering (PMCC). In a distributed computing environment, clustering calculations in each view are executed independently and in parallel. In addition, the authors performed local manifold learning to preserve locally intrinsic geometrical structure in the data. Finally, the weight of each view is shared in a cooperative normalized way.

The paper by Xiaofeng Zou et al. focused on the performance bottleneck of deep CNNs in object detection and image classification and proposed a parallel optimization framework: multi-task cascade deep CNN (MTCD-CNN). The object detection module locates and crops the areas that may contain objects for large-scale commodity recognition. The hierarchical image classification module uses hierarchical spectrum clustering to construct a tree-like image classification model. The authors verified the performance advantages of MTCD-CNN through experiments on large-scale commodity image detection.

The paper by Minrong Lu focused on the performance optimization and application of neural networks in the financial field and introduced a deep learning-based monetary policy prediction model. A large-scale training dataset for financial markets was gathered, and three neural network models with different network structures were established. The first model: back-propagation (BP) neural network model, was used as a forecasting model for monetary policy. The second model: weights BP (WBP) model, was established by considering important characteristics of financial index data. The third model: timing WBP (TWBP) model, was established on the basis of the timing characteristics of financial markets.

The paper by Titinunt Kitrungrotsakul et al. focused on the parallel training model of CNN and its application in bioinformatics. The authors discussed the limitations of the high false positive due to the complexity of mitotic and normal cells and the orientation of the mitosis when using a two-dimensional CNN model. They then proposed a 2.5-dimensional (2.5D) CNN model with a convolutional long short-term memory (LSTM) component to extract time series knowledge from large-scale microscopic images. The 2.5D CNN model is further combined with 3D anchors to collect spatial information for final mitotic detection. To improve the performance of the proposed model, a parallelism solution is implemented on a multi-GPU computing platform.

The paper by Keyang Cheng et al. focused on the performance improvement of CNN models. A parallel stochastic gradient descent (PSGD) method was proposed to train five-hierarchical parallel CNN models. A five-hierarchy parallel structure with attribute-based blocks was proposed to accelerate the training process. During the parallel training process, momentum correction and adaptive adjustment of the learning rate are applied, and the time interval for updating parameters is inspected during the optimization of parameter selection. In addition, the authors applied the PSGD model to pedestrian re-identification applications. With the help of existing unsupervised CNN, the pedestrian attributes are divided into five hierarchies from top to bottom. These five attribute modules are trained with the PSGD algorithm to improve the performance.

The paper by Jin Zhang et al. focused on the construction of biological neural system models and introduced the bionic model of olfactory neural system—KIII model. Based on neurophysiological experimental data, the authors created a KIII neuron model that accurately reflects the response of olfactory neurons to odor stimulation. The KIII model realistically simulates the structure of a real olfactory neural system. In addition, the process of odor molecules is gradually transformed by the core components of the olfactory system (including olfactory receptor, olfactory bulb, and olfactory cortex). Two groups of experiments were performed on the epileptic electroencephalograph (EEG) recognition tasks to evaluate the accuracy and performance of the proposed model.

The paper by Fan Wu et al. focused on the application of CNN models in fingerprint pattern recognition and human personality classification. The authors manually annotated a six-category fingerprint database and established a new CNN model to identify real fingerprint patterns. The proposed CNN model consists of four convolutional layers, three max-pooling layers, two norm layers, and three fully connected layers. Experiments were performed on a large-scale fingerprint database to evaluate the automatic learning and feature extraction abilities of the proposed model.

Acknowledgements We thank all authors who submitted to this special issue. We acknowledge the many reviewers who anonymously and promptly contributed their valuable insights and suggestions. We also appreciate the support from the editorial office.