Phd title

Multi-criteria automated design and optimization of deep neural networks

Supervisor: Prof. El-ghazali TALBI

e-mail: el-ghazali.talbi@univ-lille.fr

Context

Over the last years, Deep Neural Networks (DNNs) [1] has enabled significant progress in many applications including image/speech recognition, language translation, computer vision, etc. Among the crucial contributing aspects in AutoDNN (Automated DNNs) are the design of novel neural architectures [2] and optimization of their associated hyper-parameters [3] and the advent of ultra-scale GPU-powered supercomputers [4]. Currently employed neural architectures have mostly been developed manually by human experts, which is a time-consuming and error-prone process. Consequently, there is growing interest in automated neural architecture search methods and hyper-parameter optimization. In addition, the rise of DNNs is continuing to be fueled by the improvements in accelerators. Important efforts have been directed towards improving the DNNs methods to deal with neural architecture engineering, hyper-parameter optimization and their hardware-accelerated implementation. However, despite these efforts and the impressive growth of high-performance computing technologies [5], the practical impact of these methods fail to meet expectation due mainly to the huge computational complexity when it comes to deal with big networks [6]. Indeed, dealing with many network layers and millions of hyper-parameters is a tedious complex task.

Automated machine learning is a big part of Deep Learning. The performance of deep learning models are highly dependent on the architecture and the choice of the hyperparameters. The automatic search for the architecture and optimal values of hyperparameters is of crucial importance in practice and tedious task for many supervised machine learning applications. Hyperparameters are the variables which will determine the network structure (ex. number of Hidden layers and units) and the variables which determine how the network is trained (ex. learning rate, network weight initialization, activation function, momentum, number of epochs, batch size).

The scientific challenges for this project can be described by the following difficulties:

- Size of the search space: DNNs are notoriously difficult to set because of the huge number of parameters to configure.
- Expensive objective function: the objective function of the problem is a whole classification, which is highly expensive to evaluate. Hence Bayesian optimization must be applied to handle this issue. It consists in using surrogate models (e.g. meta-models, reduced models) such as Gaussian processes (e.g. Kriging) to approximate the objective function.

- Multiple objectives: Most of the work on AutoDNN formulate the problem as a single-objective problem based on the accuracy. DNNs give high accuracy at the cost of high-computational complexity (e.g., billions of FLOPs). Recently, AutoDNN approaches have been applied to applications requiring light-weight models and fast run-time. It can be infeasible to run real-time applications on resource constrained platforms such as IoT, smartphones, robots, drones, autonomous vehicles, and embedded systems. Indeed, those platforms are often constrained by hardware resources in terms of power consumption, available memory, available FLOPs, and latency constraints. Optimizing those multiple objectives will enable efficient processing of DNNs to improve energy efficiency and throughput without sacrificing application accuracy or increasing hardware cost. This is a critical aspect to the wide deployment of DNNs in AI systems. Many device-related and device-agnostic objectives must be investigated for the optimization and/or the inference steps: energy consumption, inference speed, hardware cost, computational and memory cost.

The focus of this proposal is the design and implementation of multi-objective approaches for the automated design and optimization of DNNs. The planning of this Phd can be scheduled as follows:

- State of the art on the major topics related to the proposal (Neural architecture search and hyper-parameters optimization, multi-objective optimization, ...)
- Multi-objective models for automated design and optimization of deep neural networks using various criteria (e.g. accuracy, inference time, energy consumption). The inference time and energy consumption depends on the hardware used (e.g. GPU, FPGA, Asics). In addition to the configuration of DNNs, hardware configuration will be part of the multi-objective optimization problem.
- Design and implementation of multi-objective optimization approaches for DNN using large GPU-powered clusters. Validation through extensive experimentation on real applications (e.g. critical applications in automotive).

Bibliography

[1] Yann LeCun, Y Bengio, and Geffrey Hinton. Deep learning. Nature, 521: 436{44, 05 2015. doi: 10.1038/nature14539.

[2] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search: A survey. J. Mach. Learn. Res., 20: 55:1-55:21, 2018.

[3] Matthias Feurer and Frank Hutter. Hyperparameter optimization. In Hutter *et al.* (2019), pages 3-38. In press, available at http://automl.org/book.

[4] Sparsh Mittal and Shraiysh Vaishay. A survey of techniques for optimizing deep learning on GPUs, Journal of Systems Architecture, Vol. 99, 2019.

[5] Top500 international ranking: https://www.top500.org/

[6] E-G. Talbi, Automated design of deep neural networks: a survey and unified taxonomy, ACM Computing Surveys, 2021. https://doi.org/10.1145/3439730

[7] Barret Zoph and Quoc V. Le. Neural architecture search with reinforcement learning. In Intl. Conference on Learning Representations, 2017.

[8] Léo Souquet. Design of optimization algorithms for large scale continuous problems. Application on Deep Learning. Université Paris Est Créteil (UPEC), Déc. 2019.

[9] Amir Nakib, Leo Souquet, El-Ghazali Talbi: Parallel fractal decomposition based algorithm for big continuous optimization problems. J. Parallel Distrib. Comput. 133: 297-306 (2019).

[10] Ali Hebbal, Loïc Brevault, Mathieu Balesdent, El-Ghazali Taibi, Nouredine Melab. Efficient Global Optimization Using Deep Gaussian Processes. IEEE CEC 2018: 1-8.

[11] Chris Ying, Aaron Klein, Esteban Real, Eric Christiansen, Kevin Murphy, and Frank Hutter. Nas-bench-101: Towards reproducible neural architecture search. arXiv preprint arXiv:1902.09635, 2019.

Main activities

Principal activities:

- State of the art on the major topics related to the proposal (Neural architecture search and hyper-parameters optimization, multi-objective optimization, ...)
- Multi-objective models for automated design and optimization of deep neural networks.
- Design and implementation of multi-objective optimization approaches for AutoDNN using large GPU-powered clusters.
- Validation through extensive experimentation on real applications (e.g. critical applications in autonomous vehicles).
- Publications in high-ranked journals.

Other activities:

- Software packaging and publication on Gitlab
- Participation to scientific animation of the team (organisation of workhops/conferences)

Skills

Technical skills and required level:

- Being technically familiar with at least one of the topics: Optimization, Metaheuristics, Deep Learning, High-performance computing.
- Programming (Python, parallel programming libraries).

Language: English

Social skills:

- Spirit of collaboration and sharing