Doctoral School of Information and Biomedical Technologies

Polish Academy of Sciences (TIB PAN)

Subject:

Methodology for optimizing the learning process and energy consumption of deep neural networks

Supervisor:

dr hab. inż. Janusz Będkowski, januszbedkowski@gmail.com

Description:

Progress in the field of Artificial Intelligence (AI) is accelerating. The progressive increase in energy demand in the context of machine learning and numerous implementations of AI is significant. It is reasonable to develop a methodology for optimizing the learning process and energy consumption of AI systems. The direct benefit will be a reduction in the carbon footprint (positive impact on the environment) as well as a reduction in production and implementation costs (positive business benefit). Both aspects, environmental and business, have a direct impact on the implementation of this solution in the industry. The aim of the implementation doctorate is to develop a methodology for reducing energy consumption in selected deep neural networks that are widely used in the industry. This will allow to reduce energy consumption by up to 10%. The potential impact of the project will be global.

Key tasks of the project include:

The premise for starting work on the implementation doctorate is the significant experience of the future doctoral student with Lightweight Neural Networks (LWN) and for reducing energy consumption in AI systems, the work already has a solid foundation, which will then be developed as part of the doctorate. Research on these tools will also enable the development of a methodology for designing AI systems that minimise energy consumption. As part of the work, experiments will be carried out to verify the current state of knowledge in the field of AI.

On this basis, research problems will be clarified. Reducing energy consumption in systems based on AI models, so that the solution is generic and hardware-independent, comes down to:

1) Modification of the components of the model architecture to energy efficiency.

2) Modification of the training process and model inference.

3) Increasing the efficiency of algorithms used for training and inference.

Each of these paths will be properly verified experimentally and documented, and the result will combine elements of many approaches to the problem. Much of the

current work on this problem focuses on related issues related to minimizing energy consumption, so many of the experiments will concern their measurement and modification in order to maximize the impact on the work saving of the neural network model. Due to the breadth of the issue and the periodic verification of needs, regardless of technological progress, the product of the implementation doctorate will improve the situation of AI in the context of energy. The specificity of the problem allows the following milestones to be constructed:

KM1: Development of a research labolatory/tools.

KM2: Define laboratory tests.

KM3: Development of a methodology for optimizing energy consumption by AI.

KM4: Implementation of the solution through integration with business processes in chosen industrial partner.

Candidate Requirements:

1. A Master's degree (M.Sc.) in Computer Science or a relevant field.

2. Proficiency in Python programming.

3. Preferentially experience in at least one of the following areas:

- Implementation and training of neural networks.
- Profiling and refactoring code.
- A strong scientific curiosity and interest in AI.

Candidate should contact dr hab. inż. Janusz Będkowski (januszbedkowski@gmail.com) before formal submission of documents.

Bibliography:

[1] Thomas Bird, Friso H. Kingma, David Barber, Reducing the Computational Cost of Deep Generative Models with Binary Neural Networks, ICLR 2021

[2]Bartosz Wójcik, Marcin Przewięźlikowski, Filip Szatkowski, Maciej Wołczyk, Klaudia Bałazy, Bartłomiej Krzepkowski, Igor Podolak, Jacek Tabor, Marek Śmieja, Tomasz Trzciński, Zero time waste in pre-trained early exit neural networks, Neural Networks, Volume 168, 2023,

[3] Juhyoung Lee, Sangyeob Kim, Sangjin Kim, Wooyoung Jo, Hoi-Jun Yoo, GST: Group-sparse training for accelerating deep reinforcement learning (2021)

[4] Dor Livne, Kobi Cohen, PoPS: Policy pruning and shrinking for deep reinforcement learning (2020)

[5] Hongjie Zhang; Zhuocheng He; Jing Li, Accelerating the deep reinforcement learning with neural network compression international joint conference on neural networks (IJCNN) (2019)