Few hints towards more sustainable AI

Marc Duranton
Université Paris-Saclay, CEA, List,
F-91120, Palaiseau, France
marc.duranton@cea.fr

***Abstract*— Artificial Intelligence (AI) is now everywhere and its domains of application grow every day. But its demand in data and in computing power is also growing at an exponential rate, faster than used to be the "Moore's law". The largest structures, like GPT-3, have impressive results but also trigger questions about the resources required for their learning phase, in the order of magnitude of hundreds of MWh. Once the learning done, the use of Deep Learning solutions (the "inference" phase) is far less energy demanding, but the systems are often duplicated in quantities (e.g. for consumer applications) and reused multiple times, so the cumulative energy consumption is also important. It is therefore of paramount importance to improve the efficiency of AI solutions in all their lifetime. This can only be achieved by combining efforts on several domains: on the algorithmic side, on the codesign application/algorithm/hardware, on the hardware architecture and on the (silicon) technology for example. The aim of this short tutorial is to raise awareness on the energy consumption of AI and to show different tracks to improve this problem, from distributed and federated learning, to optimization of Neural Networks and their data representation (e.g. using "Spikes" for information coding), to architectures specialized for AI loads, including systems where memory and computation are near, and systems using emerging memories or 3D stacking.**

Keywords—Artificial Intelligence, sustainability, Deep Learning, Low power, Optimization, Codesign.

I. Introduction

Artificial intelligence is omnipresent and is often marketed as an easy way to exploit big data and large computer infrastructure to solve business processes, with the promise of finding optimizations to open up even unknown market potential. AI, and more specifically Deep Learning (DL), was first a necessity for the major technology companies in the USA–GAFA– and in China –BATX: for example, to check if the millions of pictures uploaded everyday are "correct" (a typical Facebook deep learning use case) or for their virtual assistants. However, their increasing performance is done during the learning phase at the expense of more and more computing power (especially in the field of NLP). According to D. Amodei et al., "*since 2012, the amount of compute used in the largest AI training runs has been increasing exponentially with a 3.4-month doubling time"*[1]. This raised concerns about the energy required (and the access to such a computing power), as high as 656,347 kWh for a transformer network with 213M parameters with neural architecture search [2,3], and the latest GTP-3, from Open-AI, is using 175 billion parameters [4]... On the consumer side, if DL inference is used in IoT devices or terminals such as smartphones or smart speakers (more that 87M in USA in 2020 according to VoiceBot[5]), their multiplicity will also have an important impact in term of energy. A major challenge is therefore to reduce energy consumption of Artificial Intelligence, and more particularly of Deep Learning.

II. How to reduce Energy of Deep Learning?

This will be the main topic of this short tutorial, which will put the emphasis on an holistic approach, from better algorithms [6], efficient accelerators such as [7] embedded in systems using what technologies such as FDSOI can bring in term of energy reduction[8]. Energy used by DL can also be reduced by simplifying the topology of the Neural Network, and using less bits (quantization) for operations. This can be achieved by tools[9], helping the designer to take the best options. Like in all electronic systems, energy can be greatly minimized by reducing the cost of moving data, and computing in or near memory is very suitable for DL architectures. Furthermore, sparse coding of information, like using "spikes" instead of "bits", can lead to even further increase of efficiency of DL accelerators[10]. Sometimes, a physical phenomenon can be used to replace a digital computation, if its law follows a similar kind of equation than the computation. Since DL uses quite simple equations (sum of products in inference, minimizing an "energy" or a gradient descent in the learning phase), several materials express properties than can be used in "analog" realization of DL, either electronically[10] or optically.

References

[1] D. Amodei, D. Hernandez, "AI and Compute", OpenAI blog, https://openai.com/blog/ai-and-compute/, May 16, 2018.

[2] K. Hao, "Training a single AI model can emit as much carbon as five cars in their lifetimes", MIT technology Review, June 6, 2019.

[3] E. Strubell, A. Ganesh, A. McCallum, "Energy and Policy Considerations for Deep Learning in NLP", 57th Annual Meeting of the Association for Computational Linguistics (ACL). Florence, Italy, July 2019, arXiv:1906.02243v1.

[4] T. B. Brown, "Language Models are Few-Shot Learners", May 28, 2020, arXiv:2005.14165v4.

[5] B. Kinsella, "Nearly 90 Million U.S. Adults Have Smart Speakers, Adoption Now Exceeds One-Third of Consumers", April 28, 2020, from https://voicebot.ai/2020/04/28/nearly-90-million-u-s-adults-have-smart-speakers-adoption-now-exceeds-one-third-of-consumers/.

[6] D. Hernandez, T. Brown, "AI and efficiency", OpenAI blog, https://openai.com/blog/ai-and-efficiency/, May 5, 2020.

[7] A. Carbon *et al*., "PNeuro: A scalable energy-efficient programmable hardware accelerator for neural networks," 2018 Design, Automation & Test in Europe Conference & Exhibition (DATE), Dresden, 2018, pp. 1039-1044.

[8] I. Miro-Panades *et al*., "SamurAI: A 1.7MOPS-36GOPS Adaptive Versatile IoT Node with 15,000× Peak-to-Idle Power Reduction, 207ns Wake-Up Time and 1.3TOPS/W ML Efficiency," 2020 IEEE Symposium on VLSI Circuits, Honolulu, HI, USA, 2020, pp. 1-2.

[9] O. Bichler *et al.,* "N2D2, an open source CAD framework for Deep Neural Network simulation and full DNN-based applications building", available at https://github.com/CEA-LIST/N2D2.

[10] A. Valentian *et al*., "Fully Integrated Spiking Neural Network with Analog Neurons and RRAM Synapses," 2019 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2019, pp. 14.3.1-14.3.4.