Sustainability of neural network applications in training and inference - Some approaches and practices – Extended Abstract

Angela Ciocan Voinea [1,2] and Vincent Courboulay [3]

[1] *La Rochelle University – L3i, Avenue Michel Crépeau, La Rochelle, 17042, France*
[2] *Pôle Emploi – DSI Nantes, Rue Konrad Adenauer, Nantes, 44000, France*
[3] *La Rochelle University – L3i, Avenue Michel Crépeau, La Rochelle, 17042, France*

Keywords
sustainable AI, reducing energy consumption, power consumption, DNN, Green AI

A Deep Neural Networks (DNNs) have gained tremendous importance for companies in recent times. This has become possible with the emergence of big data and the great possibility of storage of these data, on the one hand, and the computing resources which allowed the development of deeper deep learning algorithms with a considerable number of parameters for the other hand [1],[2],[3],[4].
The development of more efficient hardware, such as GPUs and TPUs, has made it possible to train DNNs faster and more efficiently, which has made them more accessible to companies of all sizes [5],[6],[7],[8],[9]. The ability of DNNs to learn from vast amounts of data and extract useful insights has made them a powerful tool for companies across various industries, and as a result, DNNs have become a key tool for companies looking to gain a competitive edge in their respective industries [10],[11],[12],[13],[14].
As DNNs continue to advance in their ability to perform complex tasks, concerns have arisen about their sustainability [15],[16],[17],[18].
Overall, considering the environmental impact of DNNs development requires an exhaustive approach that considers the entire lifecycle of a model, from data collection to model deployment [19], [20],[21]. Adopting sustainable practices and technologies can help reduce the environmental impact of DNN development while still achieving high performance on various tasks.
Two key phases in the lifecycle of DNNs applications that can significantly impact its development and deployment are the training and inference. Both require significant computational resources, resulting in high energy consumption and carbon emissions [22],[23],[24],[25].
Considering the environmental impact of both training and inference, and adopting sustainable practices and technologies, can help reduce the environmental footprint of DNN models throughout their entire lifecycle. This holistic approach to DNN development can contribute to a more sustainable future for artificial intelligence and technology in general.
In this paper, we discuss techniques such as Model Compression, Knowledge Distillation, Transfer Learning and Data Management practices that can contribute to a more sustainable approach to DNN development and exploring strategies for reducing energy consumption without compromising performance [26],[27],[28],[29],[30].
This paper aims to highlight the importance of sustainability in DNNs development and to inspire further research into more sustainable approaches to training and inference.

ICT4S Workshop: Towards Sustainable and Energy-Aware Video Communications (SEViC), Month 06, 2023, Rennes, France
EMAIL: angela.ciocan@univ-lr.fr (A. 1,2); vincent.courboulay@univ-lr.fr (A. 3)
ORCID: 0000-0002-9651-0368 (A. 1,2); 0000-0002-3999-8408 (A. 3)

References

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

[2] Gadepally et al., 2019. AI Enabling Technologies: A Survey, 2019, https://arxiv.org/abs/1905.03592

[3] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." Nature, 2015.

[4] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. "Deep Learning." MIT Press, 2016.

[5] Tim Dettmers, Greg Crosswhite, and Greg Farquhar. "Theoretical and Empirical Speedup of Deep Learning on GPUs versus CPUs." International Conference on Learning Representations (ICLR), 2016.

[6] Norman P. Jouppi, Cliff Young, Nishant Patil, David Patterson, et al. "In-Datacenter Performance Analysis of a Tensor Processing Unit." International Symposium on Computer Architecture (ISCA), 2017.

[7] Volodymyr Mnih, Nicolas Heess, Alex Graves, Timothy Lillicrap, et al. "Asynchronous Methods for Deep Reinforcement Learning." International Conference on Machine Learning (ICML), 2016.

[8] Tim Dettmers. "The impact of convolutional neural networks on the field of computer vision." arXiv preprint arXiv:1704.06904, 2017.

[9] Sam S. Stone, Robert D. Kirchhoff, and Marc Snir. "Evaluating the Potential of Intel's Next Unit of Computing (NUC) for Deep Learning Research." International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA), 2019.

[10] H. Wu, Y. Chen, J. Wu, et al. "A Survey on Continual Learning in Deep Neural Networks." IEEE Transactions on Neural Networks and Learning Systems, 2022.

[11] H. Li, Z. Li, J. Xie, et al. "Progressive Learning for Deep Neural Networks: A Review." IEEE Transactions on Neural Networks and Learning Systems, 2022.

[12] R. Wu, Y. Wu, C. Dong, et al. "Deep Learning in Internet of Things: A Survey." IEEE Communications Surveys & Tutorials, 2022.

[13] Y. Wang, J. Cheng, X. Xu, et al. "A Review of Deep Learning Based Speech Synthesis." IEEE Transactions on Cognitive and Developmental Systems, 2021.

[14] Shervin Minaee, Amirali Abdolrashidi, and Yao Wang. "Deep Learning in Medical Imaging: A Review." arXiv preprint arXiv:1902.04208, 2019.

[15] Muhammad Abdullah Jamal, John B. Hill, and Tariq M. King. "Towards Sustainable Deep Learning: Perspective and Survey." arXiv preprint arXiv:2105.01653, 2021.

[16] Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, et al. "Hardware Challenges of Deep Learning." arXiv preprint arXiv:1903.05734, 2019.

[17] Payal Dhar. The carbon impact of artificial intelligence, Nature Machine Intelligence 2, 423-425, 2020.

[18] Wu et al. Sustainable AI: Environmental Implications, Challenges and Opportunities, 2022, https://arxiv.org/abs/2111.00364.

[19] Kavya Srinet and Gaurav Sharma. "Greening Data Science: Methodologies and Metrics for Sustainable Machine Learning." In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*), pp. 581-591, 2020.

[20] Tegan Maharaj, Nicholas Lane, Dong Yin, et al. "Towards Energy Efficient Deep Learning: A Survey." arXiv preprint arXiv:2104.02009, 2021.

[21] S. Han, H. Mao, and W. J. Dally. "Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding." arXiv preprint arXiv:1510.00149, 2015.

[22] M. Shoeybi et al. "Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism." In Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS), 2020.

[23] S. J. Hwang et al. "Stochastic Gradient Descent with Hyperbolic-Tangent Learning Rates." In Proceedings of the 37th International Conference on Machine Learning (ICML), 2020.

[24] S. Ma et al. "Efficient Stochastic Gradient Descent with Horizontally Updated Adaptive Learning Rates." In Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS), 2021.

[25] A. Vaswani et al. "Attention Is All You Need." In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS), pp. 5998-6008, 2017.

[26] Y. Chen et al. "Learning Efficient Object Detection Models with Knowledge Distillation." In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[27] D. N. Domingo et al. "Large Scale Pre-training for Computer Vision: A Study on When and Why." In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[28] K. Zhang et al. "Efficient Large-Scale Federated Learning with Differential Privacy and Weight Sparsification." In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[29] Y. Liu et al. "Meta-Pruning: Meta Learning for Automatic Neural Network Channel Pruning." In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[30] E. Chen et al. "Dropout Distillation for Resource-Efficient Inference in Large Neural Networks." In Proceedings of the 2021 International Conference on Learning Representations (ICLR), 2021.