Towards Energy-Aware Machine Learning in Geo-Distributed IoT Settings

Demetris Trihinas[1][0000−0002−9540−7342] and
Lauritz Thamsen[2][0000−0003−3755−1503]

[1] Department of Computer Science, University of Nicosia
trihinas.d@unic.ac.cy
[2] School of Computing Science, University of Glasgow
lauritz.thamsen@glasgow.ac.uk

Abstract. As the Internet of Things (IoT) increasingly empowers the network extremes with in-place intelligence through Machine Learning (ML), energy consumption and carbon emissions become crucial factors. ML is often computationally intensive, with state-of-the-art model architectures consuming significant energy per training round and imposing a large carbon footprint. This work, therefore, argues for the need to introduce novel mechanisms into the ML pipelines of IoT services, so that energy awareness is integrated in the decision-making process for when and where to initiate ML model training.

Keywords: Machine Learning · Internet of Things · Distributed Systems · Energy Profiling · Carbon Footprint · System Orchestration

1 Introduction

With recent advancements in IoT hardware, we are seeing the use of ML on IoT devices for highly responsive and intelligent services. However, ML is compute-hungry. In fact, the computational power required for training new state-of-the-art model architectures has been doubling every 4 months [2]. This computational effort results in higher and higher energy consumption and, in turn, increasing carbon emissions, contributing to global warming. Already, ICT organizations report that approximately 15% of their energy consumption can be attributed to AI/ML and this ratio is expected to rise considerably [3]. With Gartner [1] indicating that 75% of enterprise data will be created and processed outside of data centers, and the climate crisis demanding a rapid reduction in carbon emissions, a key emerging challenge is to adequately support the migration to sustainable AI-driven cloud edge IoT solutions [5].

This work discusses the challenges of deploying AI-driven IoT services in geo-distributed settings with a focus on energy consumption and carbon footprint. During the session we will broaden the discussion towards the need for extending ML orchestration frameworks so that their decision-making mechanisms cover energy-awareness by recommending *when* and *where* ML models should be trained and elaborate why these two inter-related challenges are not easy to overcome.

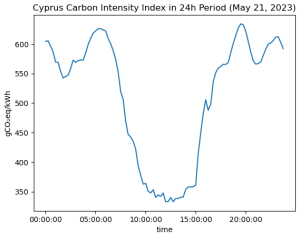

Fig. 1. Cyprus 24h carbon intensity

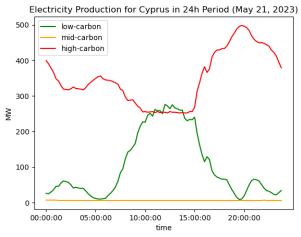

Fig. 2. Cyprus 24h power production

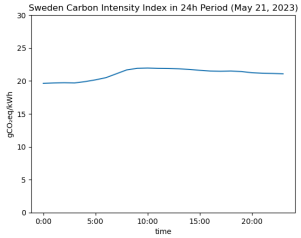

Fig. 3. Sweden 24h carbon intensity

2 Reference Use Case

To drive the discussion, let us consider a realistic ML-driven IoT application. This application features several road-side IoT units, using cameras and object detection for traffic monitoring. Several Mobile Edge Computing nodes (MECs) are scattered across the city and employed for local coordination as well as recurrent model training at a neighborhood level. For the evaluation we consider a MEC to be powered by a DELL PowerEdge R610 server and equipped with a Nvidia T4 GPU. The ML pipeline employs the TensorFlow benchmark suite[3] to output a CNN model for object detection, trained with the ImageNet dataset[4] (144GB, 1.3M images) for a duration of approximately 5 hours, when it reaches a satisfactory MLPerf accuracy.

3 When to Train a ML Model?

Deciding when to initiate repeated ML model training can highly impact the carbon footprint of an ML-based application. In particular, an application's operational carbon footprint depends on the energy mix powering the compute resources used. An illustrative example is given in Fig. 1, where for a given day in the country of Cyprus, the carbon intensity of the energy grid shows significant volatility. This is attributed to the mix of energy sources powering the grid (Fig. 2), where the low-carbon energy sources solar and wind generate to the greatest extent during the day, while high-carbon sources (i.e., oil) dominate production during the evening hours.

Taking this into account, power utilization data is extracted from the CNN model training runs over the use case testbed. Fig. 4 (red palette) showcases the estimated carbon footprint for model training with the training process initiated at different times in Cyprus. Specifically, it shows that initiating model training at mid-day versus 6pm reduces the carbon footprint by 1.93kg, while the carbon footprint is reduced even by 2.61kg in comparison to 9pm.

[3] https: //github.com/tensorflow/benchmarks
[4] https://www.image-net.org/

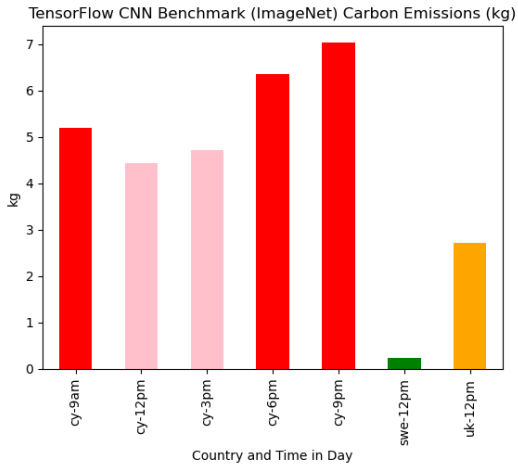

Fig. 4. Carbon footprint in kg for different periods of the day and countries

4 Where to Train a Model?

Arguably, a country or region that uses high-carbon energy sources, such Cyprus, may not be ideal for ML model training. In contrast, let us consider the country of Sweden, with Fig. 3 showing its carbon intensity measured over the same day. We can observe that the intensity is relatively stable across the day. This is ideal. First, when the model is trained does not make a huge difference. Second, the intensity is significantly lower, with Sweden usually being the EU state reporting the lowest carbon intensity. Considering now also the UK with a moderate carbon intensity that would rank it in the middle of the EU, let us go back to Fig. 4 and compare this with the training rounds initiated at different times in Cyprus. We can see that migrating an ML application to a different country can yield a significantly different environmental footprint, with model training in Sweden and the UK promising a 93% and 38% reduction in carbon emissions, respectively, in contrast to the Cyprus-based training, even during mid-day.

5 Energy-Aware Support for ML Workflow Orchestrators

Figure 5 depicts a high-level overview of the PowerML tool for aiding the decision-making of ML orchestration frameworks as to when and where to train ML models. To design such a tool the following steps are required. First, resource utilization must be mapped into energy consumption with different power models embraced for processors, memory, graphic and AI accelerators, as well as network links. In large-scale heterogeneous deployments this can easily become a configuration nightmare. To aid with this, we are building an open repository for power models that can be shared among users. Second, energy consumption must be used for estimating carbon emissions, which are dependent on the

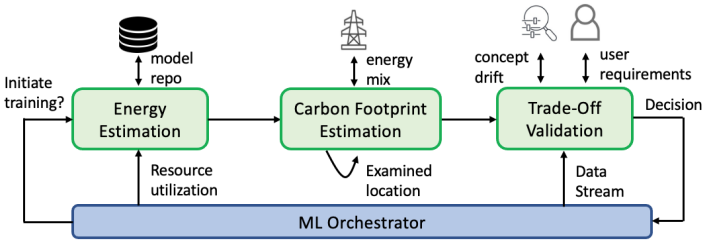

Fig. 5. The PowerML tool for aiding energy-aware orchestration of ML training

energy mix currently powering the power grid. Several grids provide live and historic data but this is either through websites or APIs, without a common data model. PowerML overcomes this challenge by providing an abstraction layer for accessing energy mix data from energy grids.

Moreover, there are many trade-offs to consider for the decision-making, commonly requiring human input as to which strategies should be explored. This is an inhibitor to a fully automated processes. One such trade-off is between accuracy and energy saving when postponing model training. That is, waiting for a low-carbon energy time window may come with a huge accuracy hit if the data distribution changes (concept drift) in the meantime [4]. Other trade-offs come with moving the workload to a different location. Moving large volumes of training data introduces delays and has a carbon cost of its own to consider. Moreover, moving data across regions is not a simple process with potential legal and privacy requirements, contradicting key arguments for in-place processing and edge intelligence.

Acknowledgements. This work is partially supported by the University of Nicosia Seed Grant Scheme for the FlockAI project.

References

1. Gartner: What Edge Computing Means for Infrastructure and Operations Leaders (2018), https://www.gartner.com/smarterwithgartner/what-edge-computing-means-for-infrastructure-and-operations-leaders
2. OpenAI: AI and Compute (2021), https://openai.com/blog/ai-and-compute/
3. Patterson, D., Gonzalez, J., Hölzle, U., Le, Q.H., Liang, C., Munguia, L.M., Rothchild, D., So, D., Texier, M., Dean, J.: The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink (4 2022)
4. Trihinas, D., Pallis, G., Dikaiakos, M.D.: Monitoring Elastically Adaptive Multi-Cloud Services. IEEE Transactions on Cloud Computing **6**(3) (2018)
5. Trihinas, D., Thamsen, L., Beilharz, J., Symeonides, M.: Towards Energy Consumption and Carbon Footprint Testing for AI-driven IoT Services. In: 2022 IEEE International Conference on Cloud Engineering (IC2E) (2022)