On energy-aware training of Deep Neural Networks in containerized environments

Dimitrios Tomaras, Vana Kalogeraki
{tomaras,vana}@aueb.gr
Athens University of Economics and Business
Athens, Greece

Nikolaos Panagiotou, Dimitrios Gunopulos
{npanagiotou,dg}@di.uoa.gr
National and Kapodistrian University of Athens
Athens, Greece

ABSTRACT

Containerized environments have emerged as the appropriate means to train a wide range of deep neural network (DNN) models for supporting data-intensive applications at scale. The containerized environment may benefit from the multiple and diverse types of accelerators available (such as CPUs, GPUs, TPUs and FPGAs) in order to speed up the training process. However, existing orchestrator platforms ignore the aspect of the cluster's energy consumption, that significanlty impacts the environmental footprint of the infrastructure. In this paper, we present SCORER, a reSource sCheduling framework fOr tRaining dEep neuRal networks in containers to address the aforementioned energy consumption challenge. Our experimental results illustrate the benefits of our approach.

ACM Reference Format:
Dimitrios Tomaras, Vana Kalogeraki and Nikolaos Panagiotou, Dimitrios Gunopulos. 2024. On energy-aware training of Deep Neural Networks in containerized environments. In . ACM, New York, NY, USA, 1 page.

1 INTRODUCTION

Deep Neural Networks have revolutionized various industries, including financial risk analysis, intelligent transportation, healthcare and smart cities. However, training DNNs can be a challenging problem due to the inherent model complexity, since, the more complex the deployed model is, the more expensive the training procedure will be. To this end, the training process of DNNs can benefit from the use of multiple and heterogeneous accelerators in cluster infrastructures (such as CPUs, GPUs, TPUs and FPGAs) in order to parallelize its execution, thus resulting in low values of training time required. However, the great diversity of DNN models may result in under-utilization of the cluster infrastructure, since not all jobs have the same computing requirements. If the training process opts for only a specific type of accelerator (i.e. train only with GPUs), the rest of the available resources remains unused, thus the cluster infrastructure is not utilized at its full scale.

For speeding up the DNN training process, approaches like in [1] have been proposed, exploiting multiple types of parallelism. However, these works are still limited since they focus on using only homogeneous cluster infrastructures and fail to address the energy consumption aspect, significantly impacting the clusters' environmental footprint.

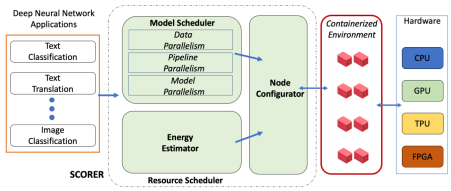

Figure 1: SCORER overview

2 RESEARCH PROBLEM AND APPROACH

In this work, we aim to address the aforementioned challenges by proposing SCORER, an energy-aware reSource sCheduling framework fOr tRaining dEep neuRal networks in containers. SCORER solves the energy-aware resource scheduling problem by balancing the trade-off between the timely execution of DNN training processes and minimizing the energy consumption of the cluster infrastructure. We select the appropriate number of instances of CPU and GPU-enabled containers, which can be used for speeding up the training process, but are also minimizing the energy consumption, based on the needs of each DNN model to be trained. SCORER incorporates two major components, the **Model Scheduler**, which is responsible for selecting the type of parallelism utilized to speedup the training process, based on specific attributes of the DNN model to be trained, and the **Energy Estimator**, which, based on profiling runs, examines the possible configurations that minimize both the training execution time and the energy consumed for each specific configuration. As soon as the **Energy Estimator** component finishes the generation of the appropriate configuration, this is forwarded to the **Node Configurator** component that deploys the appropriate number of CPU and GPU containers and initiates the training process of a DNN model. Our experimental evaluation has shown the significant benefits of our approach, minimizing the energy consumption of the DNN training process.

ACKNOWLEDGMENTS

This research has been financed by the European Union through the EU ICT-48 2020 project TAILOR (No. 952215), the H2020 AutoFair project (No. 101070568) & the Horizon Europe CoDiet project (No. 101084642).

REFERENCES

[1] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20*: IEEE, 1–16.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM