Energy efficiency measurement and optimization of ML models deployment in cloud providers

Project Proposal and Work Plan

WRITTEN BY

Name: Alec Lagarde Teixidó
Date: 22/02/2023

REVIEWED AND APPROVED BY

Name: Silverio Martínez-Fernández
Date: 07/02/2023

Project overview and goals

The project is carried out at the Universitat Politècnica de Catalunya (UPC) between January and June of 2023. The director is Silverio Martínez-Fernández and the codirector is Matias Martinez.

The research of this TFG consists of understanding how existing ML inference cloud providers optimize calculations for energy reduction. We will study the following aspects of ML models deployment: (i) energy consumption measurement after applying model optimization (e.g., quantization, pruning); (ii) impact regarding the optimization framework (Pytorch and Tensorflow); (iii) context-aware evaluation of the energy efficiency for diverse cloud providers (e.g., AWS, Azure).

We explore around 9 ML models for diverse domains (balanced among computer vision (3), NLP (3), and code (3)).

- RQ1 - What is the impact of model optimization techniques (such as quantization and pruning) in energy consumption and accuracy?
    - RQ1.1 - What is the energy consumption of applying the optimization strategy?
    - RQ1.2 - To what extent does the optimization framework (Pytorch and Tensorflow) affect the energy consumption.
    - RQ1.3 - To what extent does the optimization strategy affect the energy consumption of the ML models' inference?
    - RQ1.4 - To what extent does the optimization strategy affect the accuracy of the ML models' inference?
    - RQ1.5 - Can we optimize the tradeoff between energy consumption and accuracy?
    - RQ1.6 - To what extent does the cloud provider affect the energy consumption of the ML models' inference?
    - RQ1.7 - To what extent does the cloud provider affect the accuracy of the ML models' inference?

Project background

The project is performed in the framework of the Towards green AI-based software systems: an architecture-centric approach (GAISSA) project and takes some aspects of Daniel Escribano's TFG and amplifies it, adding more models, introducing optimization techniques, and analyzing the accuracy.

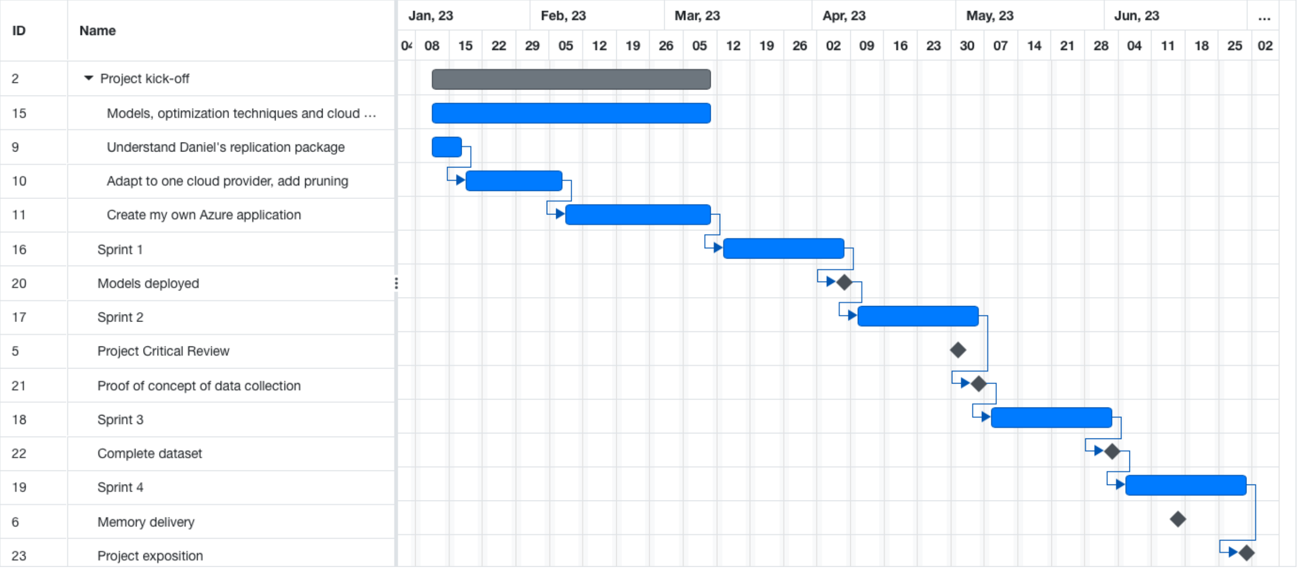
The main project initial ideas were provided by the supervisor, who posted the offer in the Racó and the current ideas have evolved from the initial ones that were posted in the first place.

Work Plan

Tasks and Milestones. Gantt Diagram

We divide this project in the project kick-off and 4 sprints:

1. **Project kick-off**
   a. **Models, optimization techniques and cloud providers selection**
   In this first part, a proposal with 9 models, 3 optimization techniques and 1-2 cloud providers has already been made. There is still not a definitive selection but that will probably include 3 natural language processing (NLP) models, 3 computer vision models, 3 code models and the optimization techniques and cloud providers are subject to change too.
   b. **Demo**
   The demo part is being performed concurrently to the first part. This consists of taking Daniel Escribano's replication package and adapting it to a single cloud provider and adding pruning as an optimization technique while keeping the T5 model and computing its accuracy. The tricky part about this is the Azure application, as it must be remade from scratch (as will be for the rest of cloud providers).
2. **Sprint 1: Deploy of baseline models and energy measurement of optimization**
   Deploy the baseline of all models in all cloud providers, getting one proof of concept of each optimization.
3. **Sprint 2: Analysis of optimization strategies**
   Design and execute data collection.
4. **Sprint 3: Comparison of deployed models in different cloud providers**
   Deploy all optimized models and execute data collection.
5. **Sprint 4: Analyzing collected data**
   Deep analysis of the dataset to answer RQs.



ID	Name
2	▾ Project kick-off
15	Models, optimization techniques and cloud ...
9	Understand Daniel's replication package
10	Adapt to one cloud provider, add pruning
11	Create my own Azure application
16	Sprint 1
20	Models deployed
17	Sprint 2
5	Project Critical Review
21	Proof of concept of data collection
18	Sprint 3
22	Complete dataset
19	Sprint 4
6	Memory delivery
23	Project exposition

Meeting and communication plan

A weekly meeting has been established every Tuesday morning. There, the progress made during the week will be discussed. If there is anything that needs to be discussed before, e-mail is used.

Generic skills

The following generic skills will be promoted and assessed during the development of the project.

#	Generic Skill	Assessed
GS1	Innovation and entrepreneurship	
GS2	Societal and environmental context	X
GS3	Oral and written communication	X
GS4	Teamwork	
GS5	Survey of information resources	
GS6	Autonomous learning	X
GS7	Communication in a foreign language	X
GS8	Gender perspective	