|  |  |  |
| --- | --- | --- |
| |  | | --- | |  | |  | |
|

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

Project Proposal and Work Plan

**WRITTEN BY REVIEWED AND APPROVED BY**

Name: Alec Lagarde Teixidó Name: Silverio Martínez-Fernández

Date: xx/02/2023 Date: xx/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) context-aware evaluation of the energy efficiency for diverse cloud providers (e.g., AWS, Azure).

We explore 10 ML models for diverse domains.

* RQ1 - What is the impact of model optimization techniques (such as quantization and pruning) in energy consumption and accuracy?
  + RQ1.1 - To what extent does the optimization strategy affect the energy consumption of the ML models’ inference?
  + RQ1.2 - To what extent does the optimization strategy affect the accuracy of the ML models’ inference?
  + RQ1.3 - Can we optimise the tradeoff between energy consumption and accuracy?
* RQ2 - What is the impact of deploying ML models in different cloud providers (such as Azure, AWS) in energy consumption and accuracy in production (i.e., inference)?
  + RQ2.1 - To what extent does the cloud provider affect the energy consumption of the ML models’ inference?
  + RQ2.2 - To what extent does the cloud provider affect the accuracy of the ML models’ inference?
  + RQ2.3 - Can we optimise the tradeoff between energy consumption and accuracy?

##### Project background

In order to allow determining the originality and scope of the contributions made by the project author, a short statement describing the origin of the main ideas should be made:

* This project takes some aspects of Daniel Escribano’s TFG and amplifies it, adding more models, introducing optimization techniques (RQ1), and analyzing the accuracy.
* State if the project is independent or it is performed in the framework of the department or company research or development project
* The main project initial ideas were provided by the supervisor, who posted the offer in the Racó, but the current ideas have evolved from the initial ones that were posted in the first place.

##### Work Plan

## Tasks and Milestones. Gantt Diagram

Describe the main envisaged tasks and milestones of the project providing their estimated periods of time and deadlines, respectively. If informative, include a Gantt diagram showing the dependencies among tasks.

We divide this project in 4 parts:

1. **Models, optimization techniques and cloud providers selection**

In this first part, a proposal with 10 models, 3 optimization techniques and 3 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.

1. **Demo**

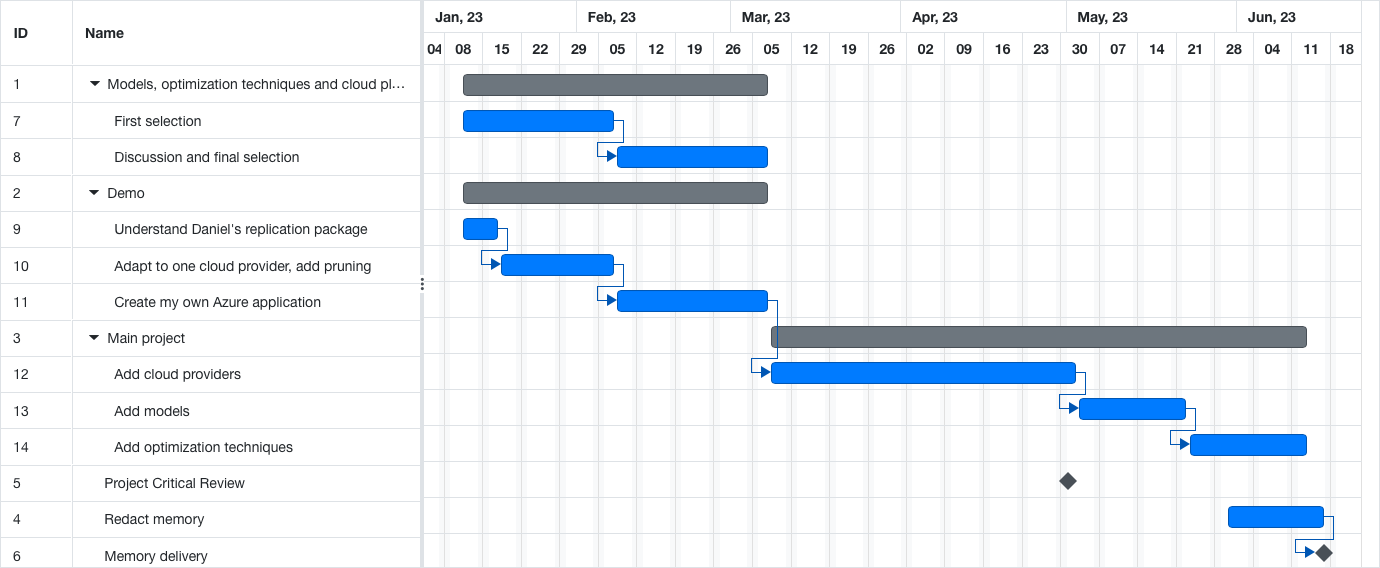
The demo part is being performed concurrently to the first part. This consists of taking Daniel Escribano’s replication package and adapting it to a single cloud provider and adding pruning as an optimization technique while keeping the T5 model and computing its accuracy. The tricky part about this is the Azure application, as it must be remade from scratch (as will be for the rest of cloud providers).

1. **Main project**

After having a clearer idea as to how it all must work, the demo will be adapted adding the rest of cloud providers, models and optimization techniques.

1. **Redact memory**

Once most of the work is done, the writing of the memory can start, which has to be delivered on the 16th of June.



## 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. Note that an initial set of generic skills has been defined by your advisor when creating the project proposal. Check whether this initial proposal is still the most suitable one for the project and make, if necessary, the required changes both in this document and in the academic platform (Racó).

Be aware that if you have some of the third level generic skills not scored yet with A or B, you can work them in your TFG in order to obtain your Bachelor degree with the set of generic skills completely acquired.

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