Aurora González Vidal
University of Murcia

Alexander Isenko
Technical University of Munich

K.R. Jayaram
IBM Thomas J. Watson Research Center

Contact: aurora.gonzalez2@um.es

On Serving Image Classification Models

This study forms part of the ThinkInAzul programme and was supported by MCIN with frunding from European Union NextGenerationEU (PRTR-C17.I1) and by Comunidad Autónoma de la Región de Murcia – Fundación Séneca

Funded by the European Union

Index

Index

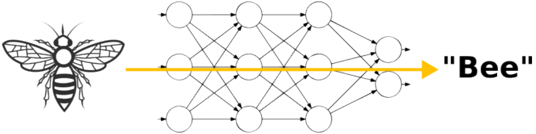
Model Inference - Motivation

- Up to 90 % of the infrastructure cost for developing and running a deep learning application is spent on inference.
- Needs: scalable, guarantee high system goodput, and maximize resource utilization.
- Intention: Set the foundations for model inference serving in serverless computing environments

Model Inference - Motivation

Objective: analyse the factors independently and together to build up a generalizable optimization model to assist in scheduling decisions

Use case: Image classification inference because its many applications such as e-comerce and retail (Amazon or Pinterest), social media such as instagram, autonomous vehicles, medical image analysis etc



Index

Background

Types of inference according to deadline guarantees.

- "Hard" Real-time Inference
- "Soft" Real-time
- Relaxed Inference
- Best-effort Inference

Equipment: TPU, GPU, CPU, etc.

Our study case: 1 GPU (NVIDIA A100 with 40 GB of VRAM), "Soft" Real-time and Relaxed Inference.

Index

CLOUD STARS

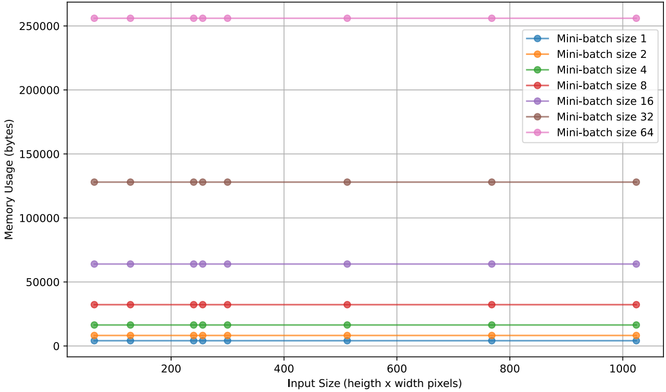
Methodology

- Selection of an image classification model: EfficienNet-B0
- Creation of dummy images with different input sizes
- Measuring inference times (repeated) over the different input sizes and mini-batch sizes looking for dependencies (for later on defining functions)
- Hardware monitoring [1] (164 features including network bandwidth, disk read/write bandwidth and counters, CPU parameters, memory utilization, GPU (pynvml and torch): temperature, memory fragmentation, etc.)
- Proposition of mathematical models for the optimization of the inference process

[1] `https://github.com/cirquit/py-hardware-monitor`

Index

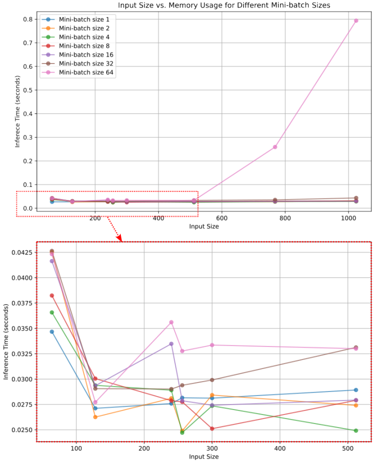
Results



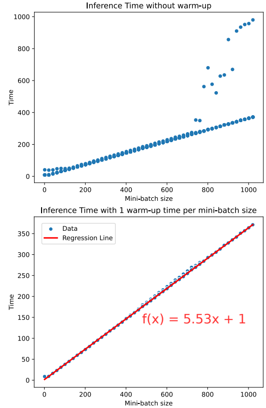
Memory usage using different image input sizes and mini-batch sizes

Results

Introduction
Background
Methodology
Results
Future Work

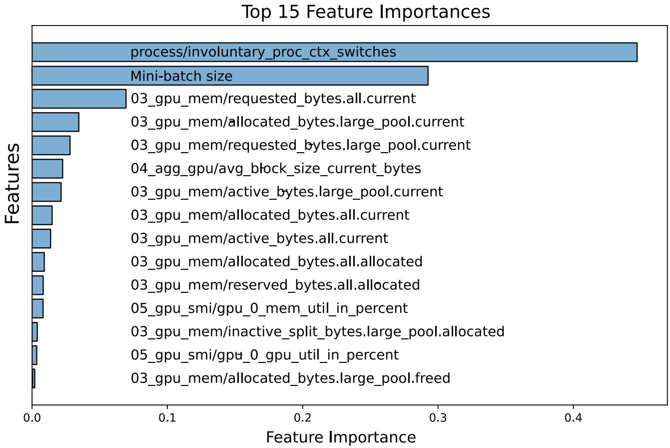


Memory usage using different image input sizes and mini-batch sizes



Inference time using different mini-batch sizes without considering warm-up (above) and considering warm-up (below) with fixed input size = 224

Results



15 most important features to determining first inference time / warm up

Optimization definitions

Decision variables:

- t_i: The number of times GPU$_i$ is used (an integer).
- mbs_i: The mini-batch size chosen for GPU$_i$ (an integer).
- N_G: The number of GPUs to be used (an integer)

The constants:

- T: The total available time. This should not be exceeded by any of the GPUs, given that they work in parallel (a decimal number).
- N: The number of images that need to be processed in total in the given time (an integer).
- $NGPU$: The maximum number of GPUs available (an integer)
- M_i: The maximum number of times GPU$_i$ can be used (a constant)
- $Size_i$: The images' input size for GPU$_i$

The functions:

- L_i: Latency per mbs$_i$ for GPU$_i$
- W_i: Warm-up time for GPU$_i$
- MB_i: The maximum mini-batch size for GPU$_i$ (a function of $Size_i$).

Optimization model - "Soft" real-time inference

$$
\begin{aligned}
\text{mín} \quad & N_G \\
\text{s.t.} \quad & \mathsf{Maximum}_i(W_i(\mathsf{mbs}_i) + t_i \cdot L_i(\mathsf{mbs}_i)) \leqslant T \\
& \sum_i (t_i + 1) \cdot \mathsf{mbs}_i \geqslant N \\
& 1 \leqslant \mathsf{mbs}_i \leqslant MB_i \quad \text{for all } i \\
& 0 \leqslant t_i \leqslant M_i \quad \text{for all } i \\
& 1 \leqslant N_G \leqslant NGPU
\end{aligned}
\tag{1}
$$

Optimization model - relaxed inference

$$
\begin{aligned}
\text{máx} \quad & NGPU \times \sum_i (t_i + 1) \cdot \mathsf{mbs}_i \\
\text{s.t.} \quad & \mathsf{Maximum}_i(W_i(\mathsf{mbs}_i) + t_i \cdot L_i(\mathsf{mbs}_i)) \leqslant T \\
& 1 \leqslant \mathsf{mbs}_i \leqslant MB_i \quad \forall i \\
& 0 \leqslant t_i \leqslant M_i \quad \forall i
\end{aligned} \tag{2}
$$

Index

Conclusions and Future Work

Conclusion: we have established a foundation for exploring the optimal way of serving AI models for image inference serving.

Future work:

- Optimal Mini-Batch Determination
- Resource Management and Load Times
- Concurrency and Cost-Energy Limits
- Versatility and Heterogeneous Serving
- Resolution of the optimization models
- Adaptation and Integration

Aurora González Vidal
University of Murcia
Alexander Isenko
Technical University of Munich
K.R. Jayaram
IBM Thomas J. Watson Research Center

Contact: aurora.gonzalez2@um.es

On Serving Image Classification Models

Funded by the European Union

This study forms part of the ThinkInAzul programme and was supported by MCIN with frunding from European Union NextGenerationEU (PRTR-C17.I1) and by Comunidad Autónoma de la Región de Murcia – Fundación Séneca