PhIMT0003 Model Optimization Log

Version 1.0

Documentation Information

Contract Number	101069595
Project Website	www.safexplain.eu
Contratual Deadline	31.04.2024
Dissemination Level	PU
Nature	R
Author	Javier Fernández
Modified by	Javier Fernández
Reviewed by	Lorea Belategi, Irune Agirre
Approved by	Irune Agirre
Keywords	DL, Inference Management, Model Optimization

Table of Contents

1 Review / Modification History

Version	Date	Description Change
V1.0	01/12/2023	First version after complete internal review
V0.2	20/11/2023	Modifications and improvements based on internal review
V0.1	11/11/2023	First draft

> *Note: The paragraphs/name of the project/Rev./Ref./history table in* **blue** *must be replaced with the information for the specific project. The paragraphs written in* **red** *are instructions that can be used as a guide, so they must be deleted.*

2 Objective

The objective of this document is to define the information relative to the optimization of the DL model.

3 Scope

This template applies to the model optimization steps performed through the Artificial Intelligence - Functional Safety Management (AI-FSM).

4 Model Optimization

The following table collects the information relative to the model optimization:

Table 1. Model optimization information

Model optimization	<Model_optimization_ID>
Date	Date of design: Format YYYY/MM/DD (year, moth, day)
Responsible	The person who converts the model
Phase of the lifecycle	Inference Management
Input Model Specifications	
Verified Learning Model ID or Model Conversion ID	<Model_ID>_<Model_ID_version> or, if the model have just been converted: <Model_conversion_ID>
Calibration fundamentals operations (preprocessing operations before post-quantization)	
Calibration	Set the range to a maximum absolute value seen during calibration, to a percentile of the distribution of absolute values, use specific methods such as the KL divergence method to obtain an entropy value…
Transformation function	For instance: $f(x)=s \cdot x$
Scale factor	I.e., $s = (2^b - 1) / (\alpha - \beta)$
Post-training quantization specifications	
Framework and version	Specify the framework used to convert the model and its version: TensorFlow, pytorch, keras, etc.
Packages and version	Tensorflow (keras, tensorflow), onnx-tf (onnx), torch (pythorch)…
Quantization precision	Precision level for quantization: 8-bit (int8_t, uint8_t), int8, 16-bit (int16_t,uint16_t)
Quantization scheme	Symmetric/asymmetric
Quantization technique	Weight quantization, integer quantization…
Quantization granularity	Layerwise quantization, channelwise quantization, groupwise quantization… In case of being a particular quantization for each layer, group of layers… there would be specified configurations for each of the quantizations.
Additional configurations	Include here all the information that makes the quantization reproducible
Pruning specifications	
Framework and version	Specify the framework used to convert the model and its version: TensorFlow, pytorch, keras, etc.
Packages and version	Tensorflow (keras, tensorflow), onnx-tf (onnx), torch (pythorch)…
Pruning criteria	Weight magnitude, gradient magnitude, global or local threshold…
Pruning patterns	Element-wise, vector-wise, block-wise, group-wise…
Additional configurations	
Techniques to recover accuracy	
Partial quantization configurations	
Quantization-aware training configurations	
Learning quantization parameters configurations	

5 Acronyms and Abbreviations

Below is a list of acronyms and abbreviations employed in this document:

- AI - FSM – Artificial Intelligence - Functional Safety Management

6 Bibliography

Add here the reference to used bibliography / references (if any).