Reproducing the sparse Huffman Address Map compression for deep neural networks

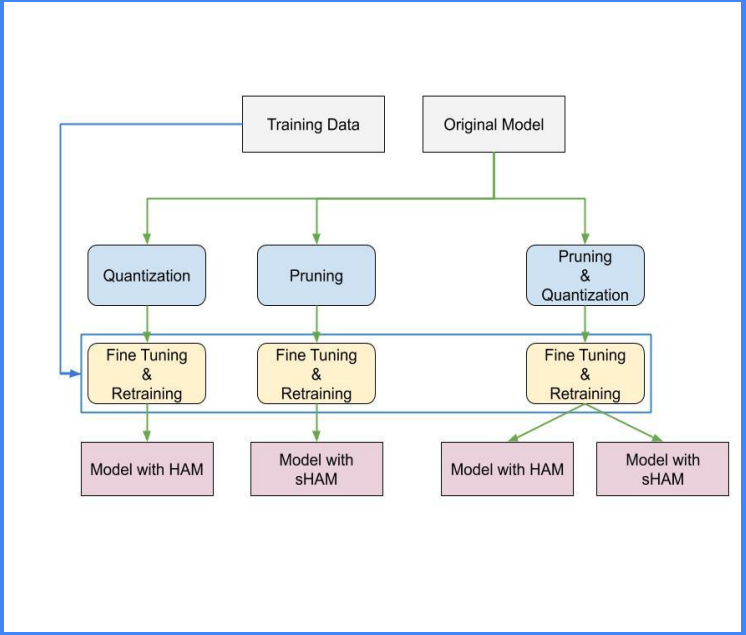
Giosuè Marinò Gregorio Ghidoli Marco Frasca Dario Malchiodi


UNIVERSITÀ
DEGLI STUDI
DI MILANO

Dipartimento di Informatica
Università degli Studi di Milano

RRPR 2021, Milan, Jan 11th 2021

Goal



- Extract relevant information in a deep neural network
  - Pruning
  - Weight sharing
  - Probabilistic quantization NEW
- Organize it exploiting succinct storage formats
  - Huffman Address Map NEW
  - Sparse HAM

G. C. Marinò, G. Ghidoli, M. Frasca and D. Malchiodi, Compression strategies and space-conscious representations for deep neural networks. In: Proceedings of ICPR 2020.

Results

NN + dataset	Compression	Δ Performance
VGG19 + MNIST	0.018	0.0011 (Acc)
VGG19 + CIFAR10	0.006	0.0019 (Acc)
DeepDTA + DAVIS	0.060	0.0552 (-MSE)
DeepDTA + KIBA	0.127	0.0017 (-MSE)

Implementation

https://github.com/giosumarin/ICPR2020_sHAM

Replicability

Python + shell scripts

- All experiments
- Specific compression
- One-shot experiment

Jupyter notebook

- Tables and graphics

Requirements

- GPU (recommended)
- ~10GB RAM

Giosuè Marinò
giosue.marino@studenti.unimi.it

Gregorio Ghidoli
gregorio.ghidoli@studenti.unimi.it

Marco Frasca
marco.frasca@unimi.it
http://frasca.di.unimi.it

Dario Malchiodi
dario.malchiodi@unimi.it
https://malchiodi.di.unimi.it
@dariomalchiodi

Thanks!