Exploiting Approximation in DNN Hardware Accelerators

Giuseppe Ascia [1], Vincenzo Catania [1], Salvatore Monteleone [2], Maurizio Palesi [1], Davide Patti [1], Enrico Russo [1]

[1] University of Catania, Italy

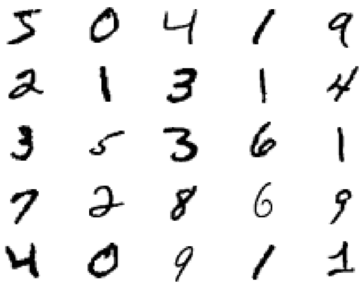
[2] Niccolò Cusano University, Italy

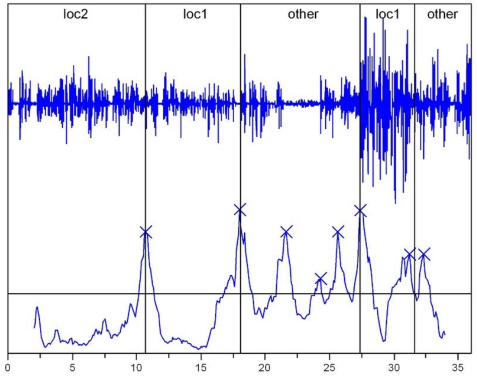
7th Italian Workshop on Embedded Systems (IWES 2022)

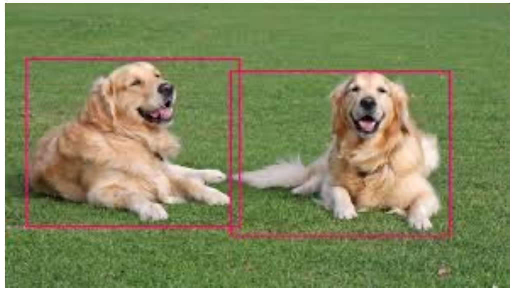
September 23rd, 2022 – Bari, Italy

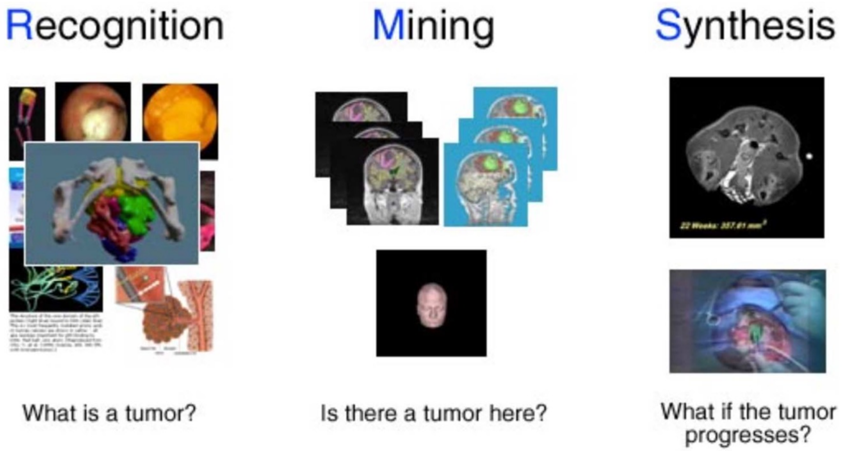
Scenario

- AI-based techniques, especially DNNs, are widespread
  - Image, video, audio, and text processing
  - RMS applications









Elements of Interest

- DNN models
- Hardware for DNNs
    - Domain Specific Architectures => Domain Specific Hardware Accelerators
    - Range from the edge to the cloud
    - Expose DNN “forgiving” nature
- Software for DNNs

… and

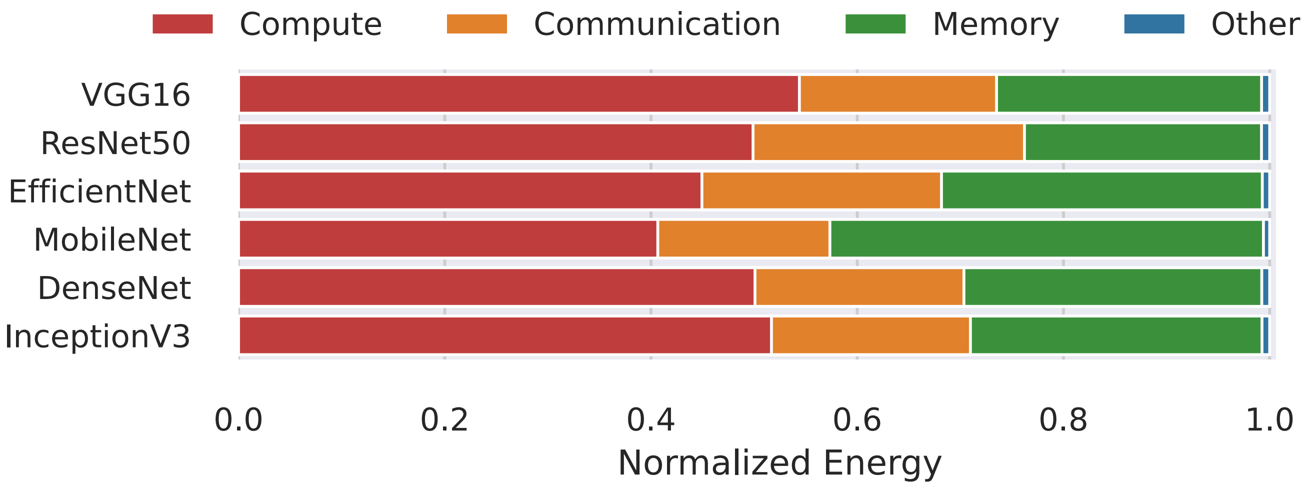
- Approximate computing paradigm

Aim

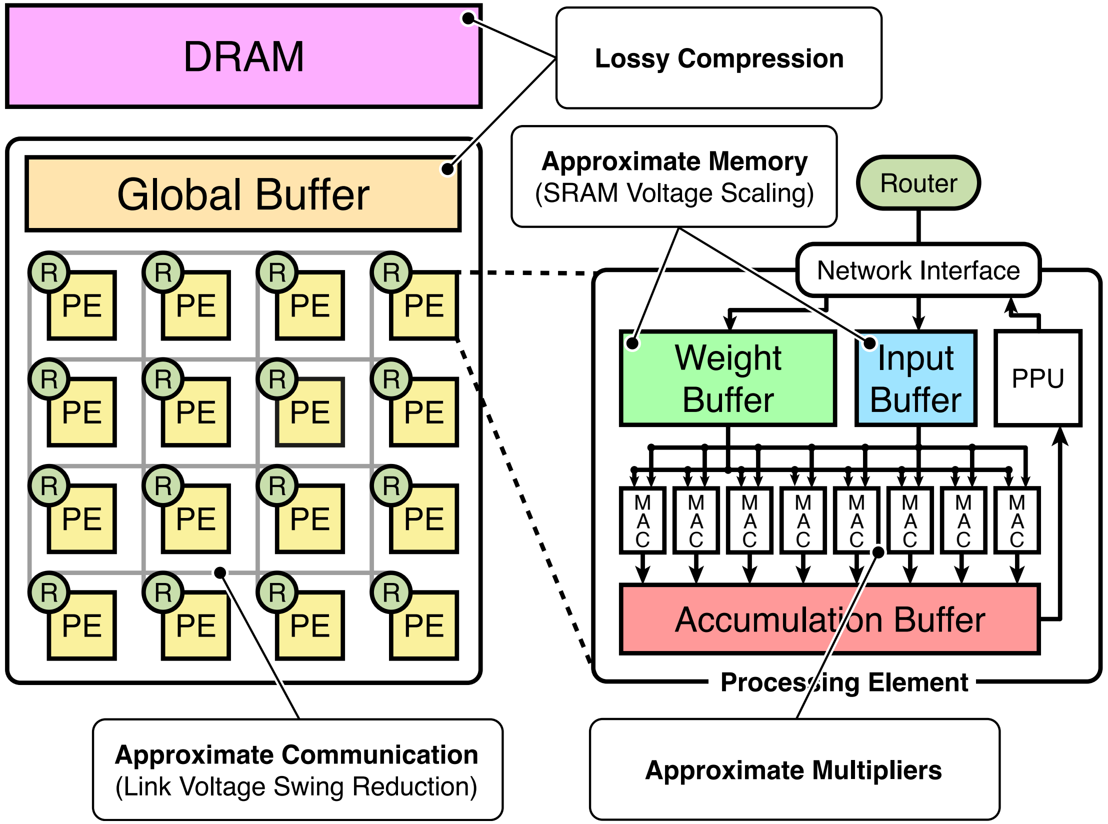
We are interested in the accuracy vs. energy trade-off.

Three main sub-systems which form the accelerator to consider:

1. Computing
2. Communication
3. Memory

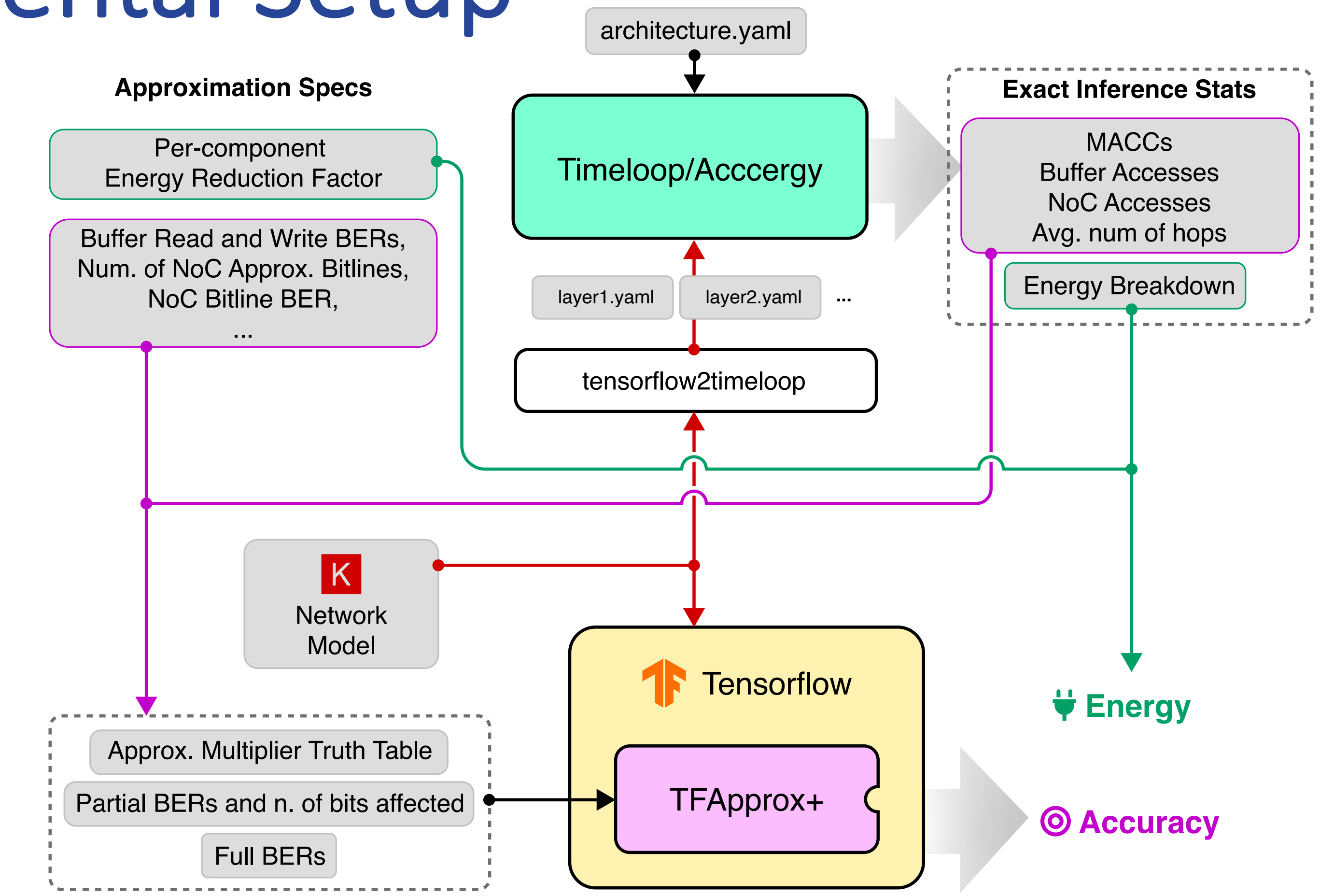


Reference Architecture + Approximation Knobs



Shao et al., “Simba: Scaling deep-learning inference with multi-chip-module-based architecture“, MICRO 2019

Experimental Setup



Parashar et al. “Timeloop: A systematic approach to dnn accelerator evaluation” ISPASS 2019
Vaverka et al. “TFApprox: Towards a fast emulation of DNN approximate hardware accelerators on gpu” DATE 2020
TFApprox extended version is available at https://github.com/Haimrich/tf-approximate

Experiments

In isolation:

Approximate Computing
Approximate Multipliers

Approximate Communication
Link Voltage Swing

Approximate Memory
SRAM Voltage Scaling

In conjunction with each other:

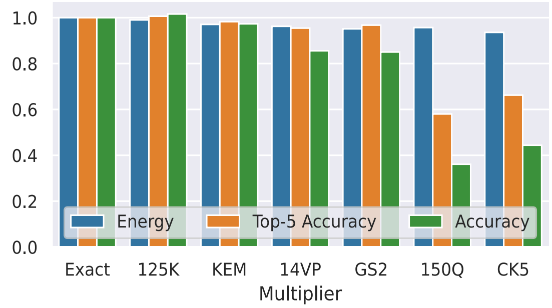
Combined Approximate Techniques Application
deriving Pareto front

Weight Compression
on a representative Pareto configuration

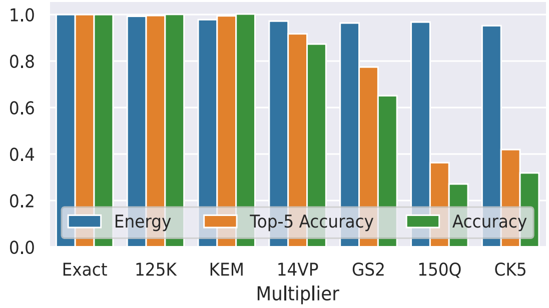
Results

Computing: Approximated multipliers from EvoApproxLib

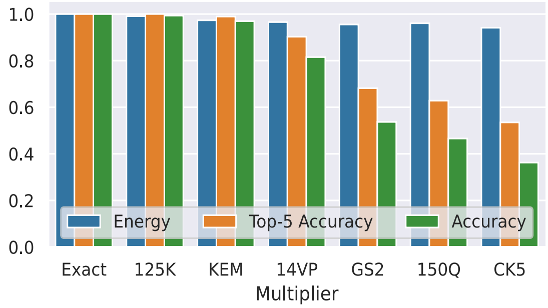
VGG16



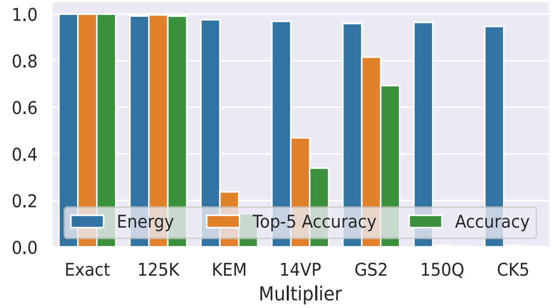
MobileNet



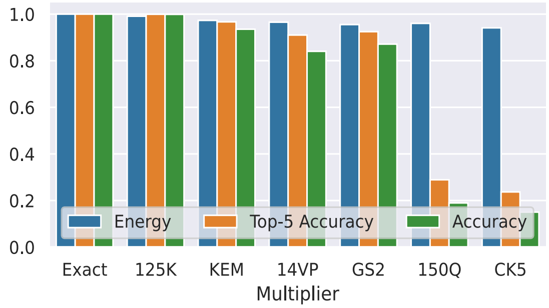
ResNet50



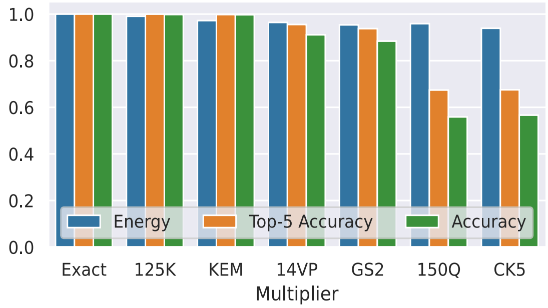
EfficientNetB0



DenseNet



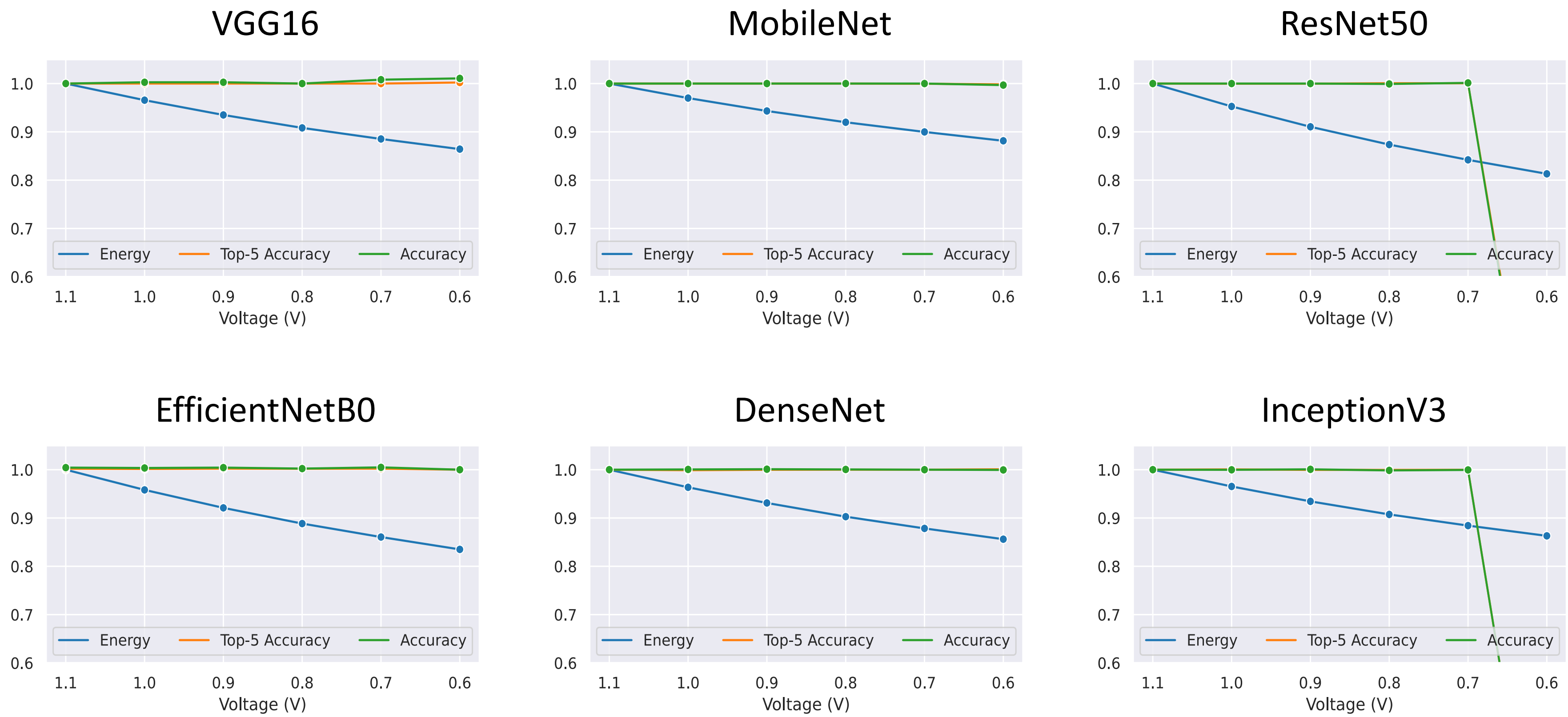
InceptionV3



Mrazek et al. “EvoApprox8b: Library of approximate adders and multipliers for circuit design and benchmarking of approximation methods” DATE 2017

Results

Communication: Network-on-Chip Link Voltage Swing

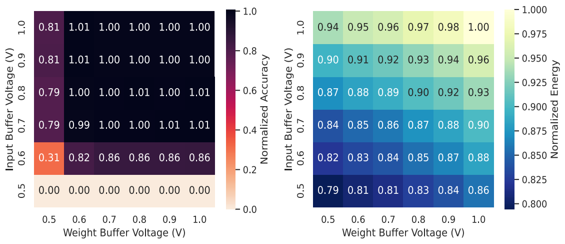


Ascia et al. “Exploiting data resilience in wireless network-on-chip architectures” in ACM Journal on Emerging Technologies in Computing Systems

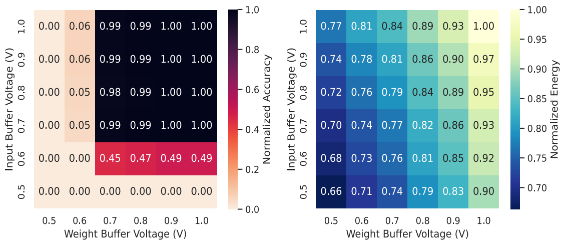
Results

Memory: SRAM Voltage Scaling (Input/Weight buffers)

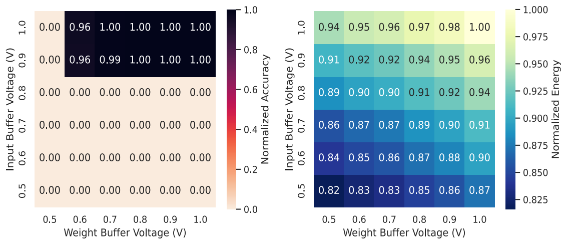
VGG16



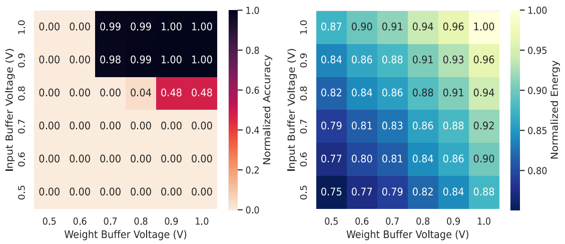
MobileNet



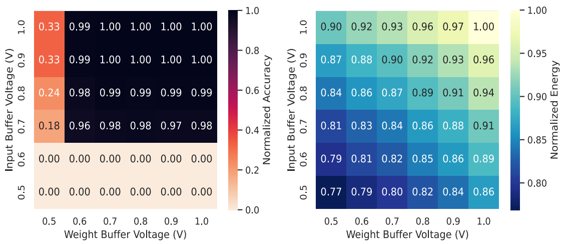
ResNet50



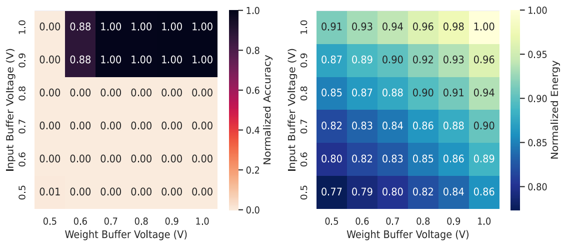
EfficientNetB0



DenseNet



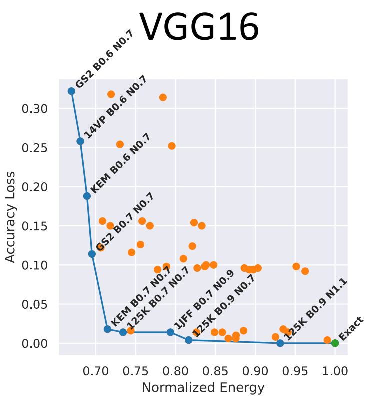
InceptionV3

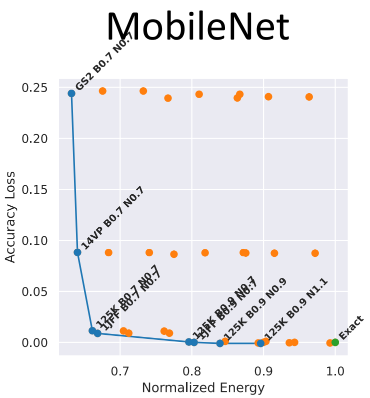


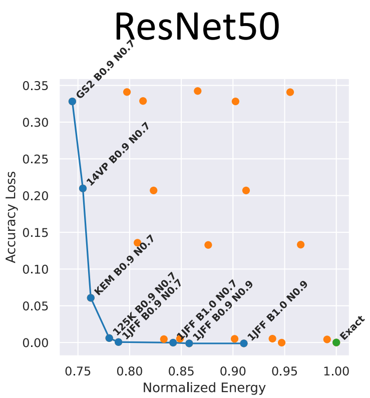
Ha et al. “Hierarchical approximate memory for deep neural network applications” ACSSC 2020

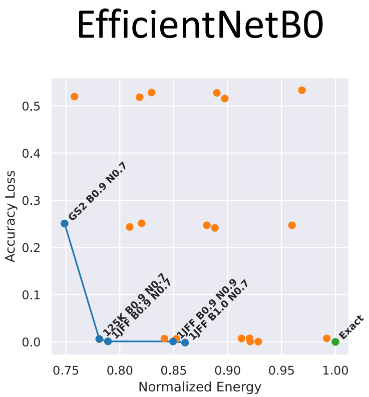
Results

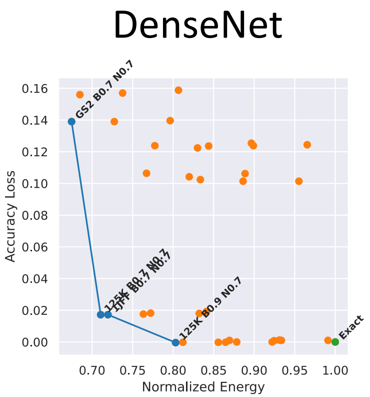
Combined Approximation: accuracy loss vs energy Pareto-sets

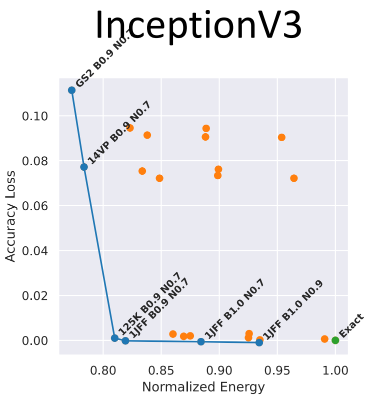

VGG16


MobileNet


ResNet50

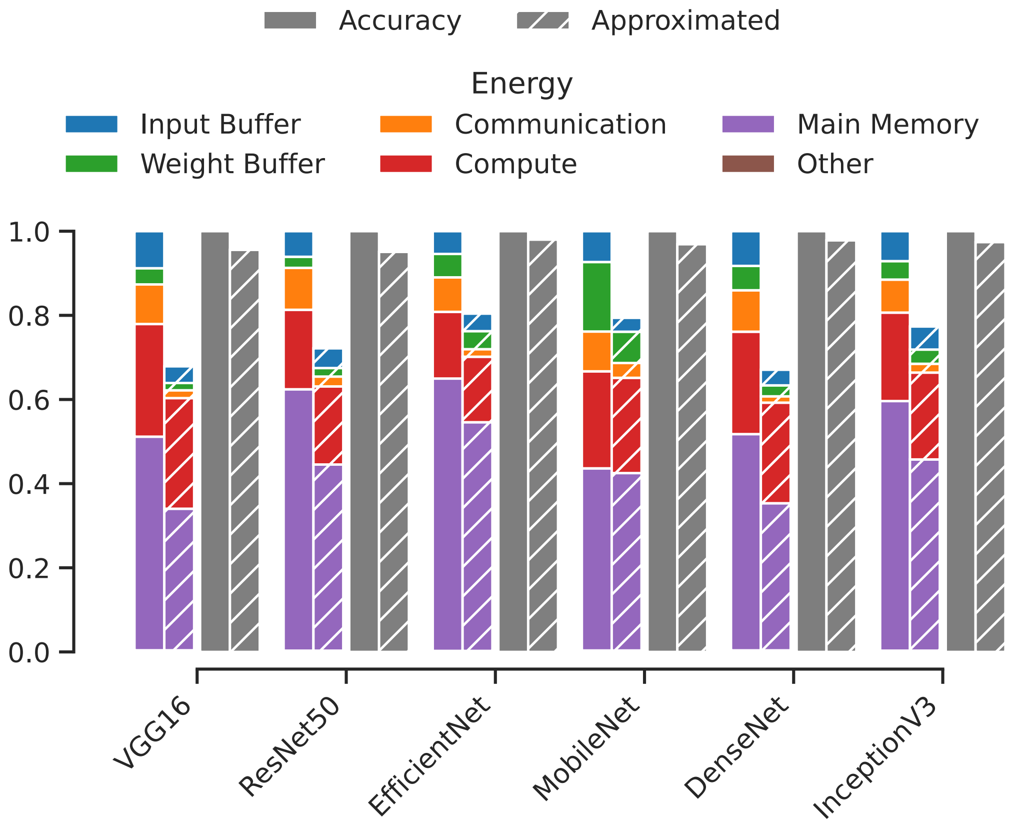

EfficientNetB0


DenseNet


InceptionV3

Results

Model Weight Compression applied on top of a representative Pareto configuration



Russo et al. "DNN model compression for IoT domain specific hardware accelerators" in IEEE Internet of Things Journal

Thank you!