Accelerating Deep Convolutional Neural Networks on Low Power Embedded Devices

José Cano¹,², Jack Turner², Valentin Radu², and Michael O’Boyle²

¹School of Computing Science, University of Glasgow, UK - ²School of Informatics, University of Edinburgh, UK

Deep Neural Networks

- Complex architecture
  - Transformations
  - Learnable parameters
- Phases
  - Training: dataset
  - Inference: prediction
- Widely adopted
  - Development of GPUs
  - Evolution of smartphones
- Types
  - Feed forward: numerical and linguistic data analysis
  - Recurrent: machine translation, natural language processing
  - Convolutional: image classification, speech recognition

VGG-16 Convolutional Neural Network

- Developed for ILSVRC 2014, top-1 accuracy of 70.5%
- 13 convolutional layers, 3 fully connected layers
- 3x3 kernels, 2x2 MAX pooling

Hardware platforms

- Odroid-XU4
  - big.LITTLE CPU
    - 4 Cortex A15 @ 2.0 GHz
    - 4 Cortex A7 @ 1.4 GHz
  - Mali T628 MP6 GPU: 6 cores @ 600 MHz
  - 2GB shared LPDDR3 RAM @ 750 MHz
- Hikey 960
  - big.LITTLE CPU
    - 4 Cortex A73 @ 2.3 GHz
    - 4 Cortex A53 @ 1.8 GHz
  - Mali G71 MP8 GPU: 8 cores @ 900 MHz
  - 3GB shared LPDDR4 SDRAM @ 1866 MHz

Results by layer

- VGG-16 - Odroid-XU4
- VGG-16 - Hikey 960

Overall results

- Odroid-XU4
- Hikey 960

Conclusions

- Important to understand the architecture of the target platform
  - E.g. number/type of cores, memory type/size, number of SIMD lines
- Transformations of the input matrices are important
  - Flatten by row vs by depth
- Naive parameter selection can lead to poor results
  - E.g. work-group size
- Auto-tuning is not always the best solution
  - CLBlast provides less improvement than hand-tuned for OpenCL