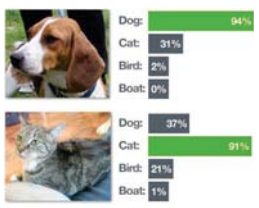
Accelerating Deep Convolutional Neural Networks on Low Power Embedded Devices

José Cano[1,2], Jack Turner[2], Valentin Radu[2], and Michael O'Boyle[2]

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

Deep Neural Networks

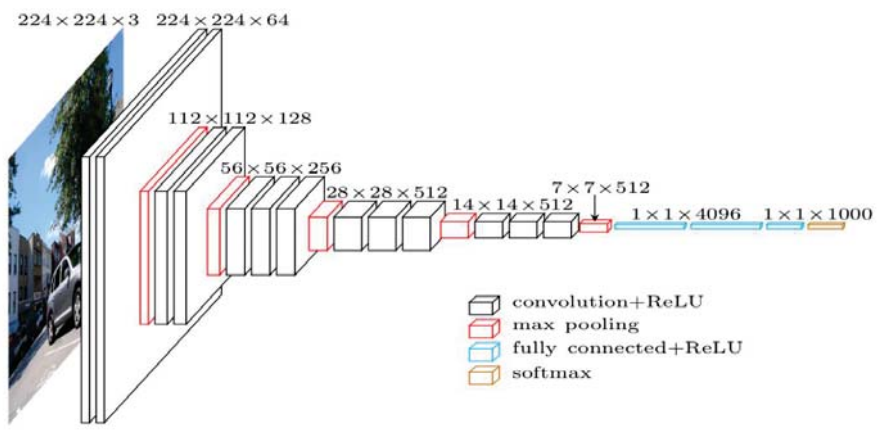
- Complex architecture
  - Transformations
  - Learnable parameters
- Phases
  - **Training**: dataset
  - **Inference**: prediction
- Widely adopted
  - Development of GPUs
  - Evolution of smartphones
- Types
  - **Feed forward**: numerical and linguistic data analysis
  - **Recurrent**: machine translation,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



AlphaGo

Google DeepMind

VGG-16 Convolutional Neural Network



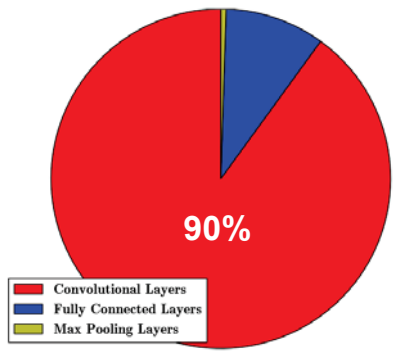
- Developed for ILSVRC 2014, top-1 accuracy of 70.5%
- 13 convolutional layers, 3 fully connected layers
- 3x3 kernels, 2x2 MAX pooling

Accelerating VGG-16

- **Objective: reduce inference time**
  - Pre-trained model (ImageNet dataset)
  - We focus on the convolutional layers
- **Initial code: serial version in C**
- **Contribution: parallel versions**

OpenMP

OpenCL



- **Optimisations**
  - Threads, work-groups, vectorisation (SIMD), CLBlast Library

Hardware platforms

Odroid-XU4

- big.LITTLE CPU
  - 4 Cortex A15 @ 2.0 GHz
  - 4 Cortex A7 @ 1.4 GHz

arm

- Mali T628 MP6 GPU: 6 cores @ 600 MHz
- 2GB shared LPDDR3 RAM @ 750 MHz

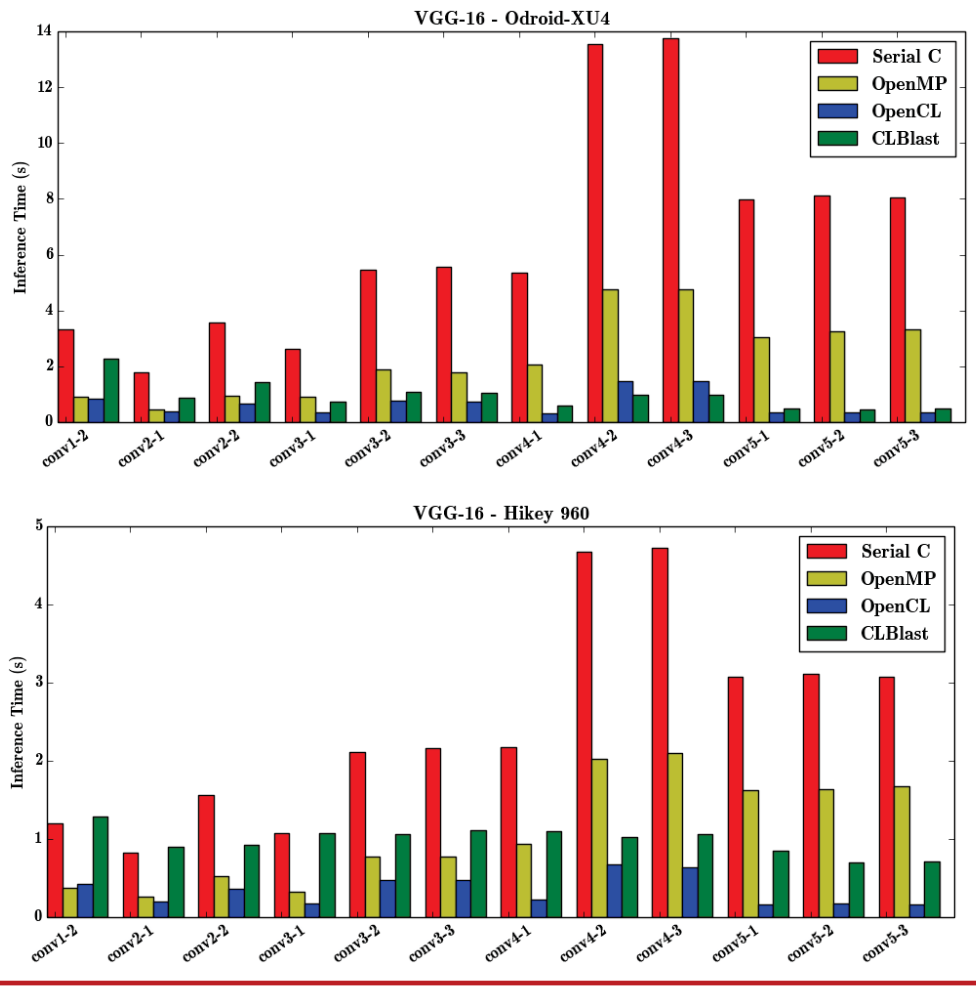
Hikey 960

- big.LITTLE CPU
  - 4 Cortex A73 @ 2.3 GHz
  - 4 Cortex A53 @ 1.8 GHz

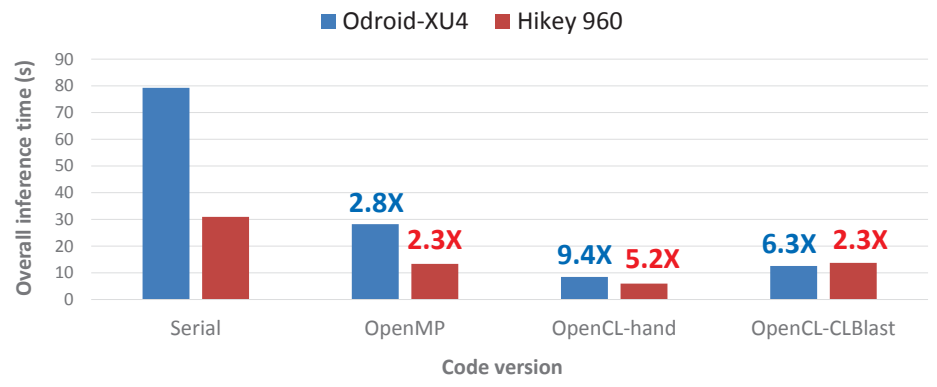
arm

- Mali G71 MP8 GPU: 8 cores @ 900 MHz
- 3GB shared LPDDR4 SDRAM @ 1866MHz

Results by layer



Overall results



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

University of Glasgow | School of Computing Science

THE UNIVERSITY of EDINBURGH informatics

ARM Research Summit - Cambridge, UK
September 17-19, 2018

BONSEYES ARTIFICIAL INTELLIGENCE MARKETPLACE