R1-15 : Convolutional Neural Network Layer Reordering for Acceleration

Vijay Daultani, Subhajit Chaudhury, Kazuhisa Ishizaka
(NEC Central Research Labs, Japan)

NEC

Overview

Issue: **Acceleration of CNN without accuracy drop**
Proposed Solution: **Reordering layers in CNN**

Key Contributions

- Proposed simple ***layer re-ordering technique*** for computational speed-up with same output.
- Obtained ***algorithmic speed up*** that benefits ***all types of hardware*** (CPU or GPU)

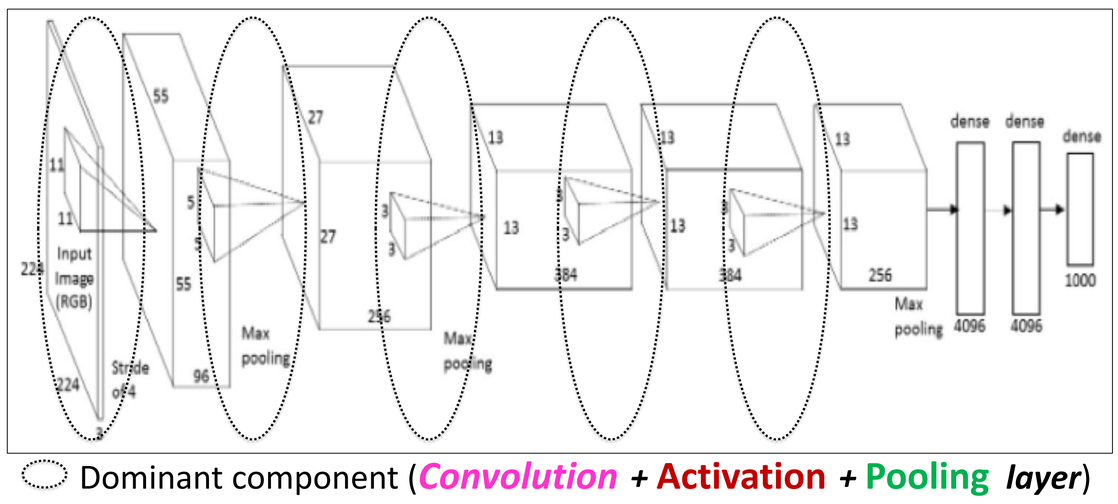
Results:

75% reduction of computational time ***for CPU***, ***20% to 67% for GPU***, in activation layer for 2x2 max pooling kernel in VGG16.

Motivation

Convolution-Activation-Pooling (CAP) pattern is very common in CNN architectures. Reducing time of computation for this pattern is key to achieve overall acceleration.

CNN (ex: AlexNet)



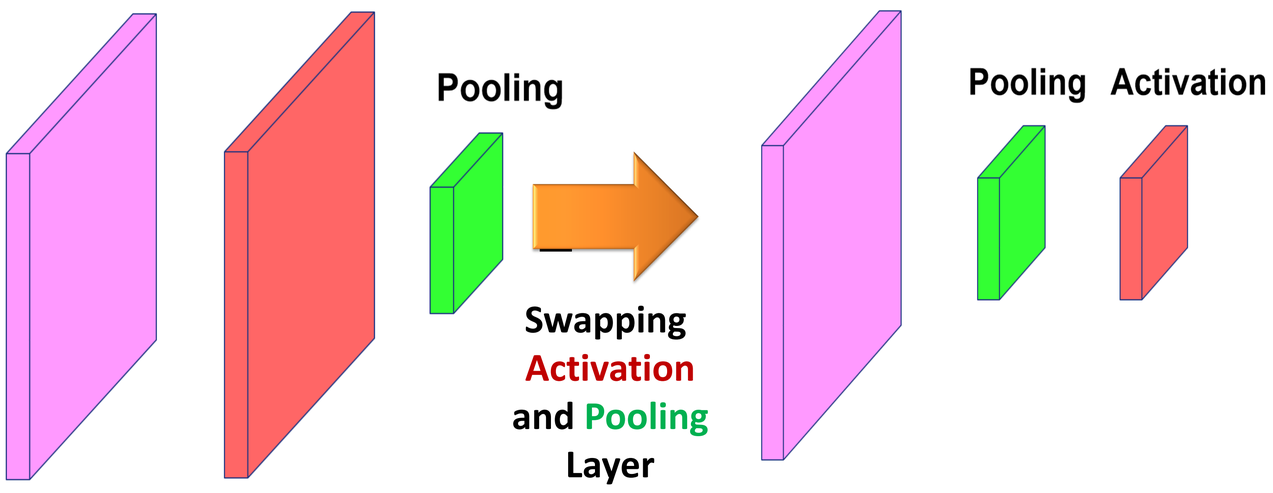
Dominant component (*Convolution* + Activation + Pooling *layer*)

Five instance of CAP pattern in Alex-net architecture

Proposed Layer re-ordering (CAP ➔ CPA)

Existing Configuration

Convolution Activation Pooling



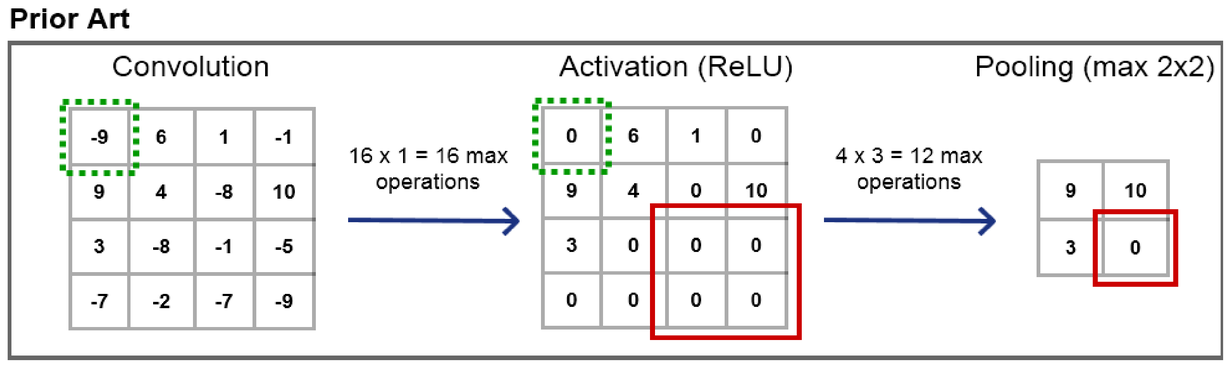
Proposed Configuration

Convolution Pooling Activation

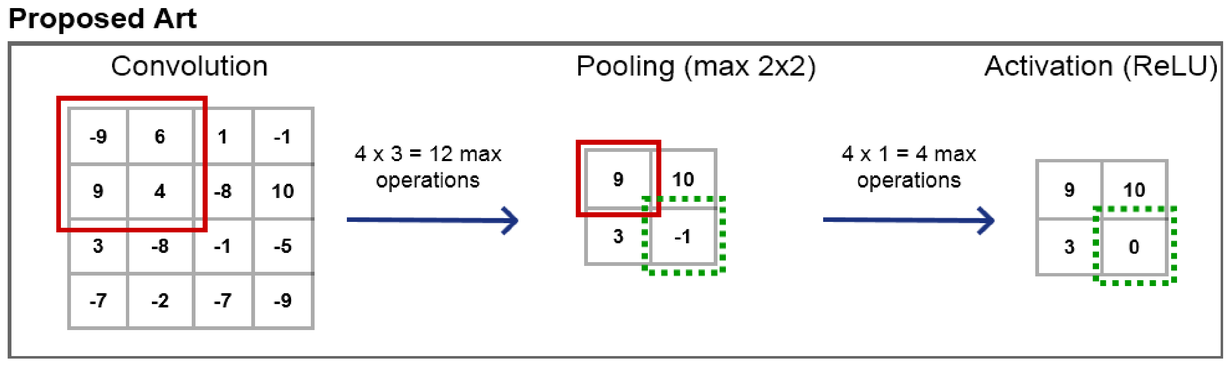
* K is size of pooling kernel. Typically 2.

Same result and less computation amount (1/K^2)

Prior Art



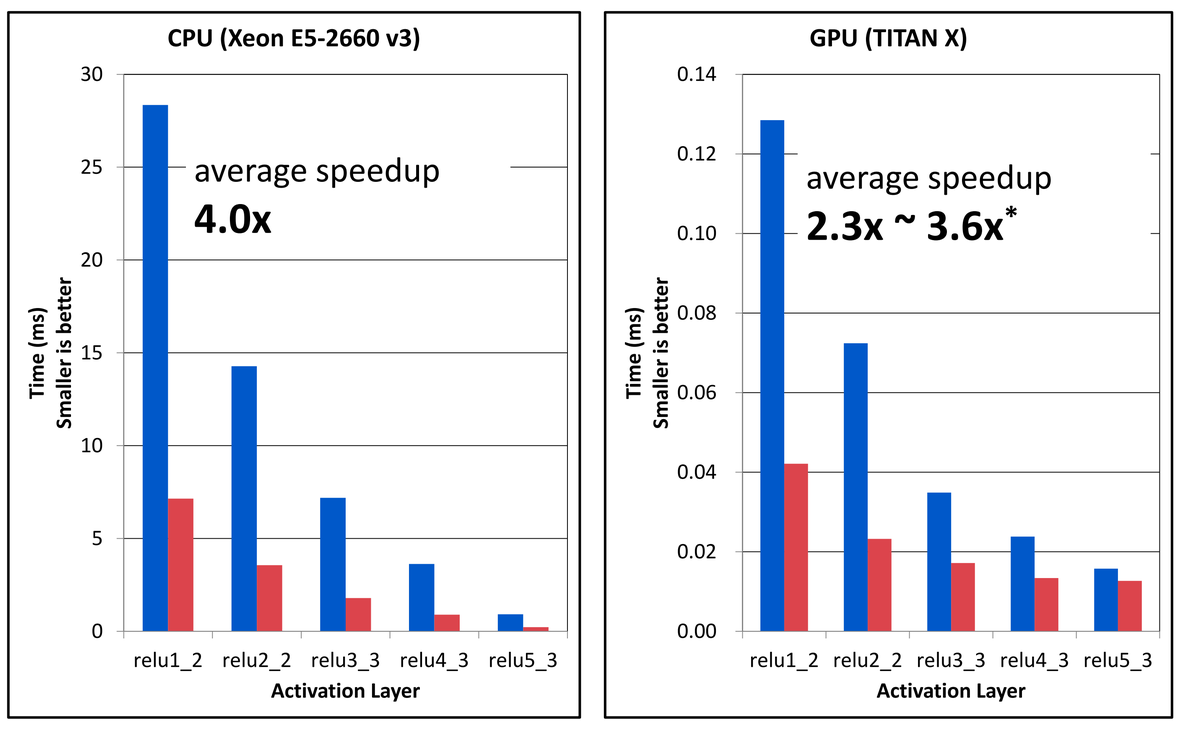
Proposed Art



Prior Art : 28 binary max vs. Proposed : 16 binary max

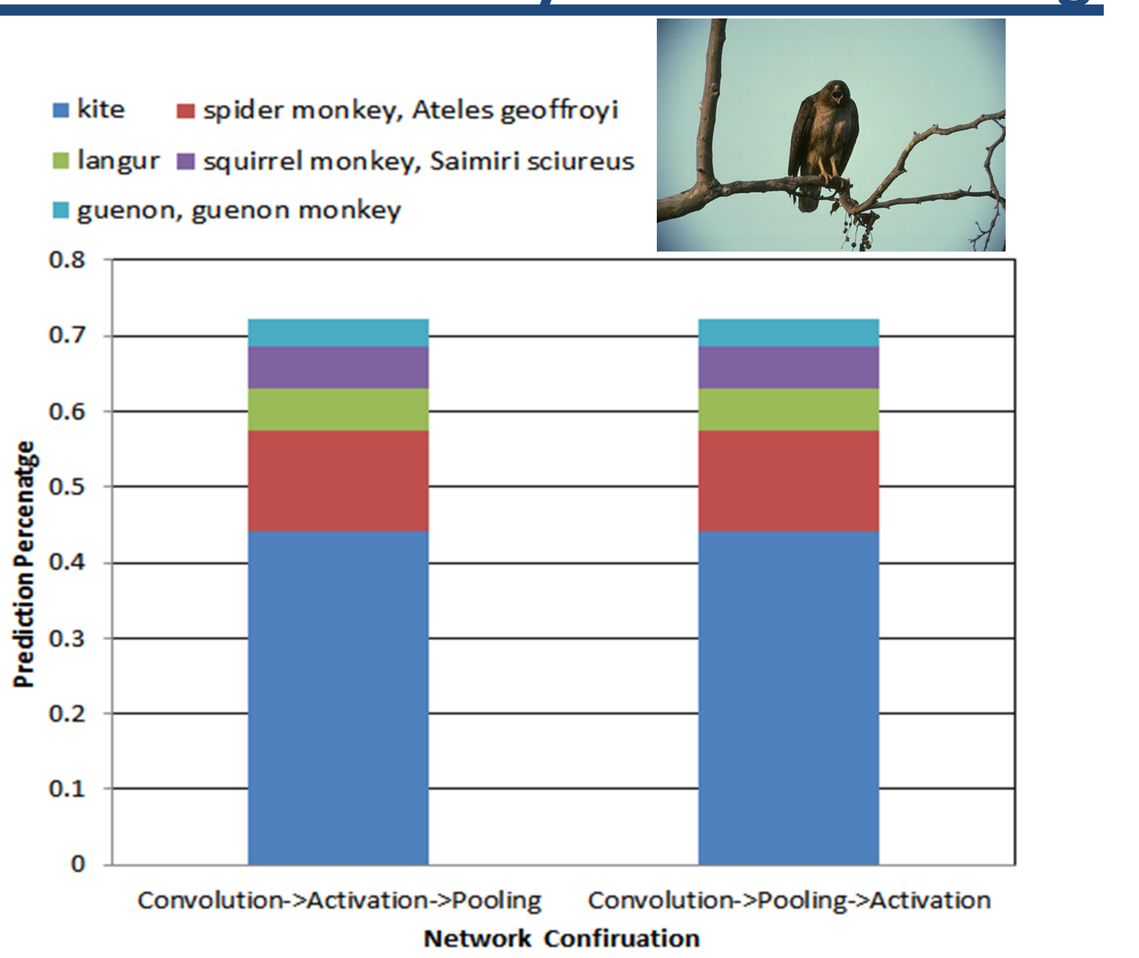
Computation time evaluation

Evaluation of proposed technique on CPU and GPU, for VGG16



Speed up by our proposed method on CPU and GPU for VGG16 with ReLU activation and 2 x 2 pooling. CPU gives speed-up close to theoretical limits.

Prediction accuracy after re-ordering



The top-five prediction scores remain unchanged due to proposed re-ordering, giving speed up with same output predication accuracy