Cross-Stack Efficiency in Neural Networks

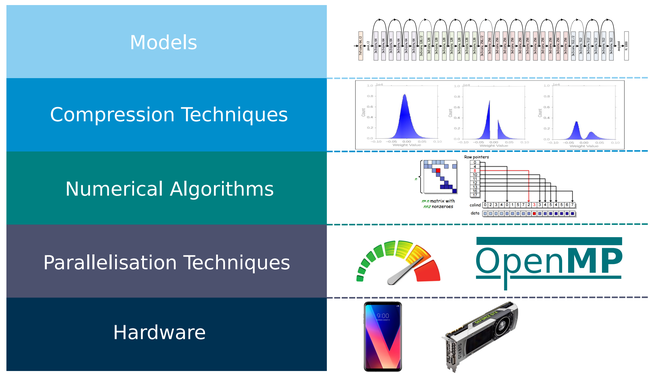
THE UNIVERSITY of EDINBURGH
informatics

Jack Turner and Michael O'Boyle
School of Informatics, University of Edinburgh
{jack.turner, mob}@inf.ed.ac.uk

EPSRC Centre for Doctoral Training in
Pervasive Parallelism

Characterising Across-Stack Optimisations For Deep Neural Networks

J. Turner, J. Cano, V. Radu, E.J. Crowley, M. O'Boyle and A. Storkey



Abstract

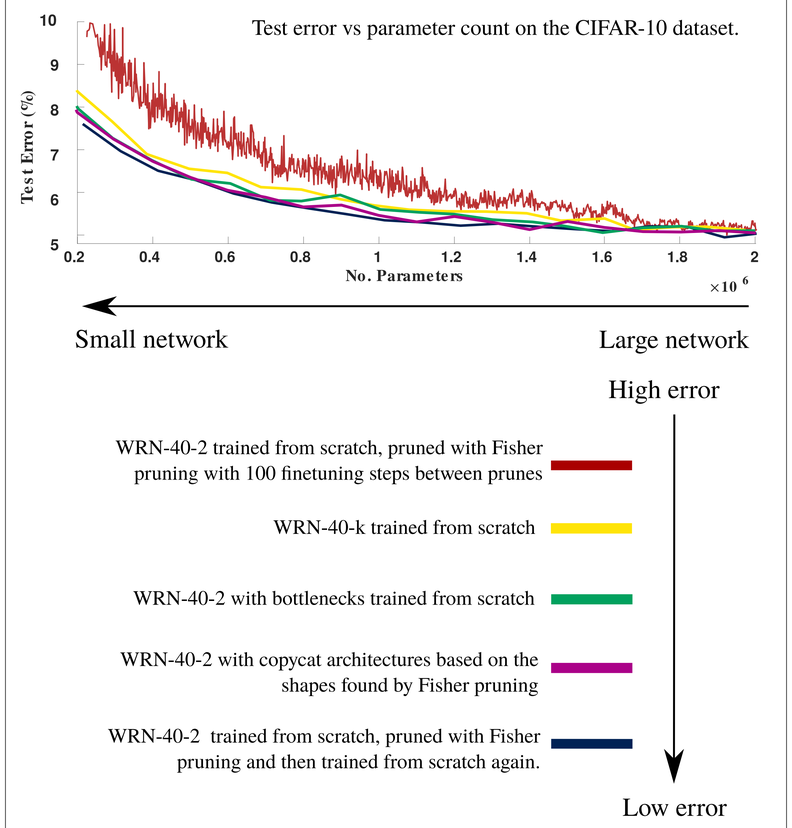
- Both the machine learning and computer architecture communities have developed acceleration methods for deep learning, but their interactions are not always clear.
- We unify the two viewpoints in an *Inference Stack* and compare weight pruning, channel pruning, and quantisation with a range of programming approaches (OpenMP, OpenCL) and hardware architectures (CPU, GPU).
- We provide comprehensive Pareto curves to instruct trade-offs under constraints of accuracy, execution time, and memory availability.

Results

- Sparse neural networks are very difficult to accelerate. Sparsity is often very irregular, with some layers left almost completely dense.
- Channel pruning is best from both an accuracy and performance perspective
- Large, old networks (VGG-16) often perform better after compression than newer networks that have been hand-crafted for efficient inference (MobileNet). The trend towards skinny, deep networks has reduced opportunities for data reuse.

Pruning neural networks: is it time to nip it in the bud?

E.J. Crowley, J.Turner, A. Storkey and M. O'Boyle



Pruning-and-tuning is not as effective as choosing smaller architectures. Better yet is to randomly re-initialise the pruned architectures and train from scratch than to fine-tune the model. We show that pruning patterns can be emulated by "copycat" architectures, and that these copycat architectures are robust to different tasks.

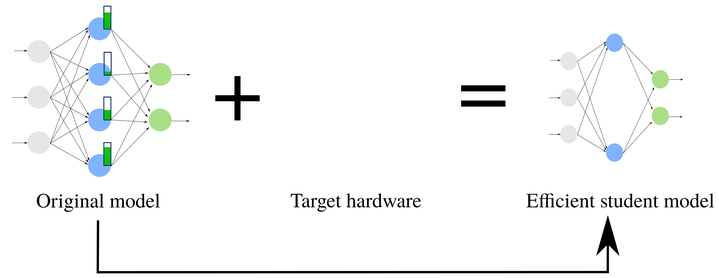
Finally, we show that our reduced architectures have faster inference speeds than pruned models.

Attention Transfer with Performance Aware Student Networks

J.Turner, E.J. Crowley, A. Storkey and M. O'Boyle

Problem: we have a large network, but it is too large/slow for our target hardware

Solution: combine channel saliency metrics ( ) with empirical observations of latency to generate efficient student networks



Student model is trained with the attention maps from our original teacher network. This gives us the flexibility to swap out operations and enhances the accuracy of the student network.

Results