Exploring the Deep Learning Inference Stack for Resource Constrained Devices

Valentin Radu, Jack Turner, José Cano, Elliot J. Crowley,
Michael O'Boyle, and Amos Storkey

School of Informatics, University of Edinburgh

There is a growing demand in ubiquitous computing for performing even more complex detection tasks at the edge of the Internet, on resource constrained devices, due to privacy concerns related to transferring user sensitive data, and to operating in locations without network connection. At the same time, deep learning has emerged as the dominant solution for improved detection accuracy in several areas of interest, computer vision, speech, translations and context detection [Radu et al., 2018]. Although outstandingly accurate, deep neural networks are known for their high computation demand, so using these to perform detections running on resource constrained devices is an open challenge.

In the EU project, Bonseyes, we are developing new solutions to facilitate portability of deep neural networks to resource constrained mobile devices. We define the Deep Learning Inference Stack (DLIS) as the set of techniques that work together to produce deep neural network based inferences, spread over the following layers: (1) Neural Network Model; (2) Machine Learning Compression Technique; (3) Data Format; (4) Computation and Workload Parallelization; (5) Hardware. For each of these layers we select specific candidates and evaluate their impact on performance in combinations across layers. Although the biggest gains are generally expected to come from layers (1) and (5), by designing smaller more refined neural networks and using specialized inference hardware respectively, our investigation is focused primarily on layers (2)-(4), which we believe still hold potential for further optimizations. Promising candidates are selected at each layer, obtaining empirical observations to produce guidelines for improved solutions, with memory, inference time and detection accuracy being the main metrics in this investigation.

Specifically, at layer (1) we select three popular deep neural network models, designed for visual recognition tasks (VGG, ResNet, MobileNet). At layer (2) we explore a range of machine learning techniques to compress the previously mentioned networks: Weights Pruning, Channel Pruning and Quantization. These are applied with model retraining to limit accuracy degradation due to loss of information. Here we find that channel pruning clearly outperforms all the other compression techniques in accuracy and memory space reduction. Quantization and pruning produce sparse weight matrices, which we choose to represent in memory either as dense or sparse format at layer (3). The most notable observation here is that contrary to general belief, sparsity is not beneficial to reducing inference time in convolutional networks for the networks we have considered, due to the small filters (3x3); alternatively the Winograd transformation being more efficient, while a cross-layer solution like Sparse-Winograd [Liu et al., 2018] improving performance even further. At layer (4) we explore parallelization techniques on heterogeneous systems, using both the CPU and GPU with OpenCL. Here we experiment with scheduling techniques to balance the workload efficiently across compute resources. The hardware selection at level (5) is represented by the ARM big.LITTLE architecture, sharing the memory space with the Mali embedded GPU, a growing presence on many mobile devices.

It is important to observe the interoperability between these layers. Decisions made at each layer can influence the operations at the other layers, so we do a cross-layer exploration of these combinations. Crucially, decisions at each layer are directly influenced by the available hardware so any choice of technique should be hardware-aware by design.

References

[Liu et al., 2018] Liu, X., Pool, J., Han, S., and Dally, W. J. (2018). Efficient sparse-winograd convolutional neural networks. In *Proc. ICLR*.

[Radu et al., 2018] Radu, V., Tong, C., Bhattacharya, S., Lane, N., Mascolo, C., Marina, M., and Kawsar, F. (2018). Multimodal deep learning for activity and context recognition. *IMWUT*, 1(4).