N. Lane et al. DeepX: A Software Accelerator for Low Power Deep Learning Inference on Mobile Devices

Alex Gubbay

UNIVERSITY OF CAMBRIDGE

The Problem

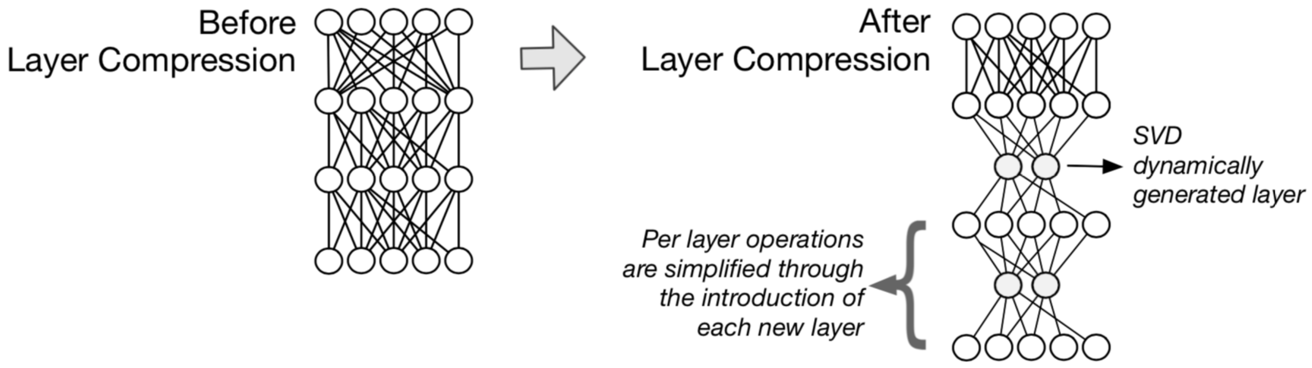
- Deep Learning Models are too resource intensive
- They often provide the best known solutions to problems
- Production mobile software using worse alternatives
- Supported in the cloud for high value use cases
- Handcrafted support

Solution: DeepX

- Software accelerator designed to reduce resource overhead
- Leverages Heterogeneity of SoC hardware
- Designed to be run as a black-box
- Two key Algorithms:
  - Runtime Layer Compression (RLC)
  - Deep Architecture Decomposition (DAD)

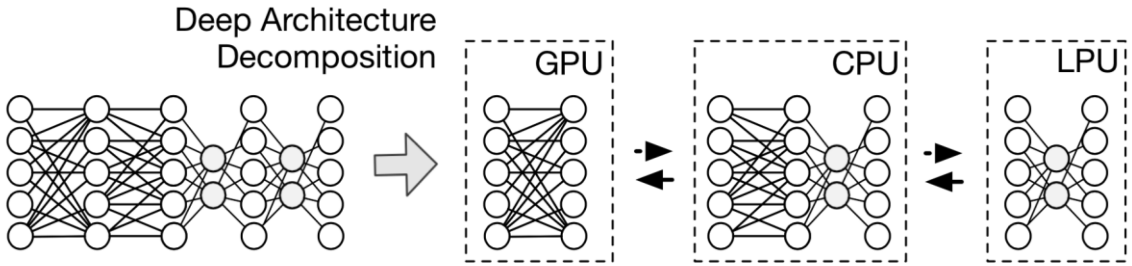
Runtime Layer Compression

- Provides runtime control of memory + compute
- Dimensionality reduction of individual layers
- Estimator - accuracy at a given level of reduction
- Error protection:
  - Conservative redundancy sought out
- Input: (L and L + 1), Error Limit



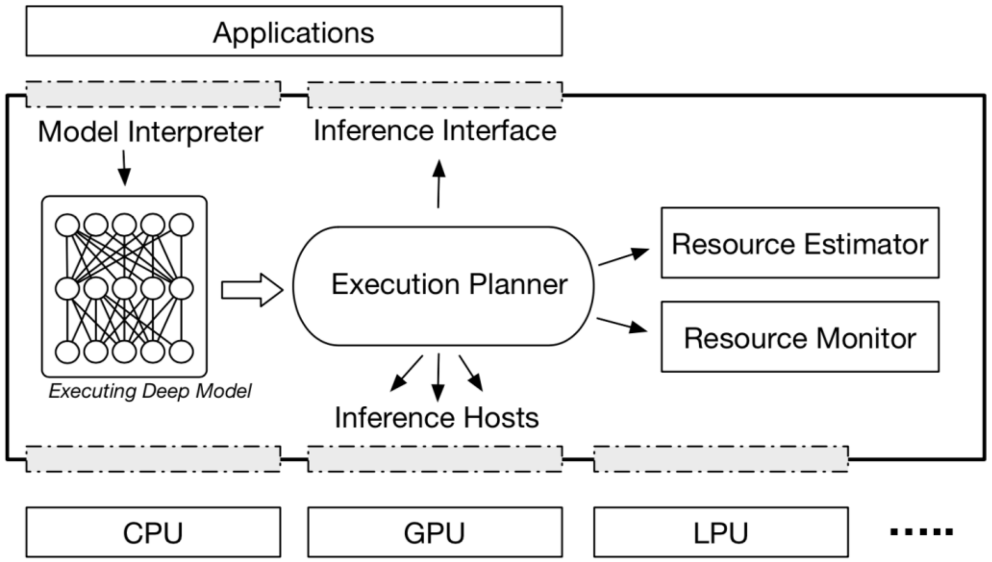
Deep Architecture Decomposition

- Input: deep model, and performance goals
- Creates unit blocks, in decomposition plan
- Considers dependencies:
  - Seriality
  - Hardware resources
  - Levels of compression
- Allocates unit blocks
- Recomposes and outputs model result



Testing

- Proof of Concept
  - Model interpreter
  - Inference APIs
  - OS Interface
  - Execution planner
  - Inference host
- Run on two SoCs:
  - Snapdragon 800 - CPU, DSP
  - Nivida Tegra K1 – CPU, GPU, LPC

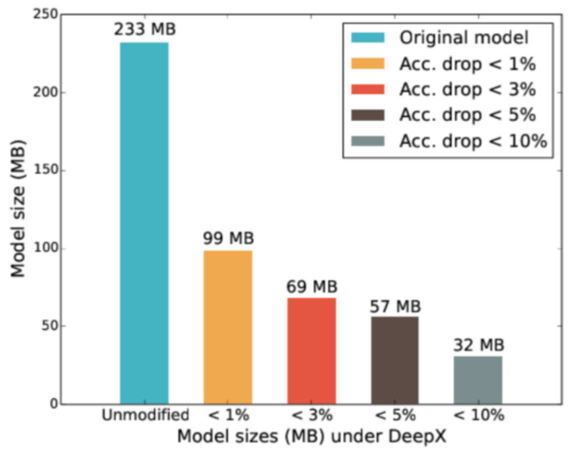


	Type	Size	Architecture
AlexNet	CNN	60.9M	$c{:}5^{\imath}$; $p{:}3^{\ddagger}$; $h{:}2^{\star}$; $n{:}\{\text{all } 4096\}^{\dagger}$
SVHN	CNN	313K	$c{:}2^{\imath}$; $p{:}2^{\ddagger}$; $h{:}2^{\star}$; $n{:}\{1600{,}128\}^{\dagger}$
SpeakerID	DNN	1.8M	$h{:}2^{\star}$; $n{:}\{\text{all } 1000\}^{\dagger}$
AudioScene	DNN	1.7M	$h{:}2^{\star}$; $n{:}\{\text{all } 1000\}^{\dagger}$

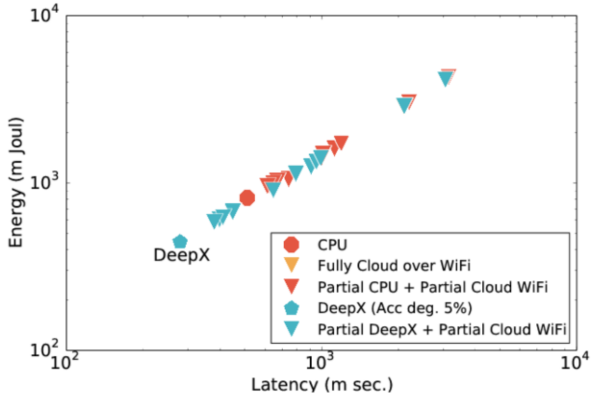
$^{\imath}$*convolution layers*; ‡*pooling layers*; **hidden layers*; †*hidden nodes*

Results

	CPU (only) (mJ)	DSP (only) (mJ)	Cloud (only) (mJ)
AlexNet	933.5 (2.1×)	–	4978.4 (11.2×)
SVHN	230.9 (2.6×)	142.1 (1.6×)	1101.1 (12.4×)
SpeakerID	113.4 (8.1×)	103.6 (7.4×)	124.2 (8.9×)
AudioScene	110.3 (8.0×)	99.3(7.2×)	122.7 (8.9×)



	CPU (only) (mJ)	LPU (only) (mJ)	GPU (only) (mJ)	Cloud (only) (mJ)
AlexNet	1681.3 (13.2×)	–	234.1 (1.8×)	2820 (22.1×)
SVHN	479.6 (4.3×)	–	167.3 (1.5×)	1382.9 (12.4×)
SpeakerID	7.1 (7.8×)	109.1 (120.4×)	1.3 (1.4×)	26.9 (29.7×)
AudioScene	6.7 (7.6×)	106.1 (120.3×)	1.2 (1.4×)	26.1 (29.4×)



	Relative Accuracy Loss (%)	Memory Reduction (%)
AlexNet	4.9 (77.5 to 72.6)	75.5 (233 MB to 57 MB)
SVHN	0.2 (83.9 to 83.7)	58.8 (16 MB to 7 MB)
SpeakerID	3.2 (93.7 to 90.5)	92.8 (28 MB to 2 MB)
AudioScene	4.3 (79.2 to 74.9)	77.8 (27 MB to 6 MB)

Conclusions

- It is possible to run full size Deep Learning models on mobile hardware
- Thorough experimentation
- Paper is candid about its limitations:
  - Changes in resource availability
  - Resource estimation
  - Architecture optimisation
  - Deep learning hardware