Towards Memory-Efficient Neural Networks via Multi-Level \textit{in situ} Generation
Jiaqi Gu, Hanqing Zhu, Chenghao Feng, Mingjie Liu, Zixuan Jiang, Ray T. Chen, David Z. Pan
ECE Department, The University of Texas at Austin

Introduction:
- Data movement is more costly than computations
- Memory bottlenecked modern AI accelerators efficiency
- Naive layer decomposition does not save memory cost
- Intrinsic intra-kernel and cross-kernel correlation in CNNs
- Motivate us to generate kernel on-the-fly with small basis

Proposed Multi-Level \textit{in situ} Generation:
- Intra-kernel generation $W_i = U_i W_k^b$, $\forall i \in [C_o]$
  - Span all input channels from a small basis $W_k^b$
- Cross-kernel generation $W = V W_c = V \{ U_i W_k^b \}_{i \in [B_c]}
  - Span all kernels from a basic kernel
- Augmented mixed-precision generation
  - Assign different bitwidth to basis and coefficient

Proposed Training Flow:
- Project teacher onto decomposed low-bit students
- Distill knowledge from teacher to students
- Orthonormal regularization to encourage ranks

Experimental Results:
- Intra-kernel correlation is stronger than cross-kernel correlation
- Outperforms separable CONV and Blueprint CONV
- Comparable performance on various tasks with compact models
- $>10 \times$ memory compression (3~5 bit quantization)
- $\sim$30\% less latency, $\sim$86\% less energy on simulated ReRAM Accel.

Summary/Conclusion
- Multi-level in-situ generation for memory-efficient DNN design
- Mixed-precision for fine-grained design space exploration
- $10$-$20 \times$ memory compression; $\sim$97\% less weight loading time
- New design paradigm to break through the ultimate memory bottleneck for emerging DNN accelerators

Personal Website: https://jeremiemelo.github.io
Email: jqgu@utexas.edu