Towards Memory-Efficient Neural Networks via Multi-Level *in situ* Generation

Jiaqi Gu, Hanqing Zhu, Chenghao Feng, Mingjie Liu, Zixuan Jiang, Ray T. Chen, David Z. Pan

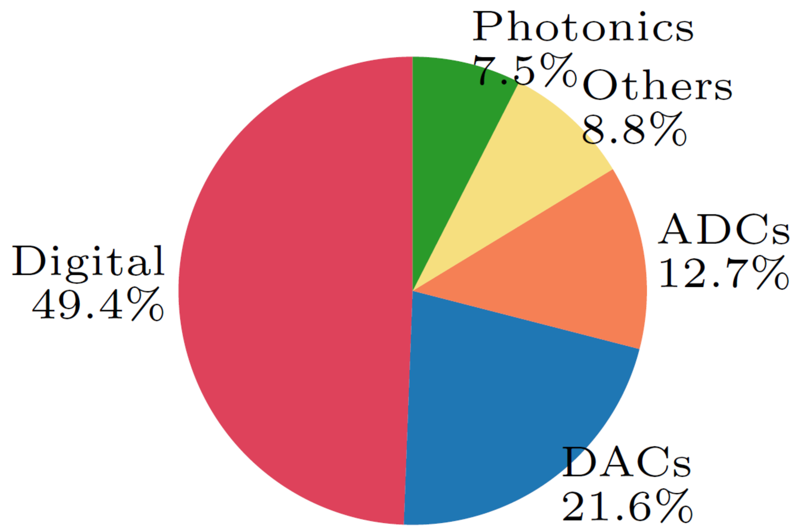
ECE Department, The University of Texas at Austin

TEXAS The University of Texas at Austin

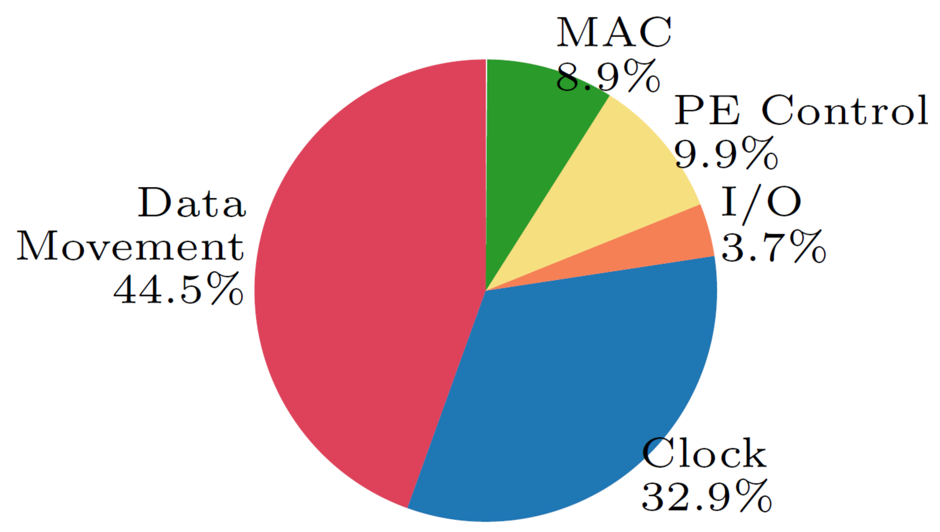
2021 ICCV OCTOBER 11-17 VIRTUAL

Introduction:

- Data movement is more costly than computations

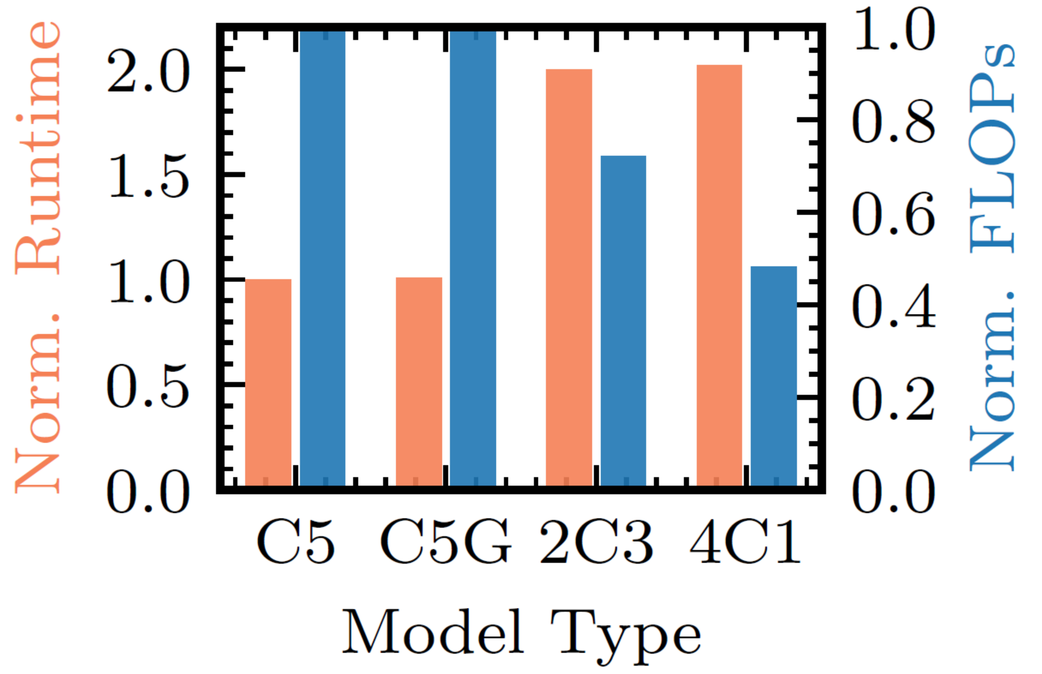


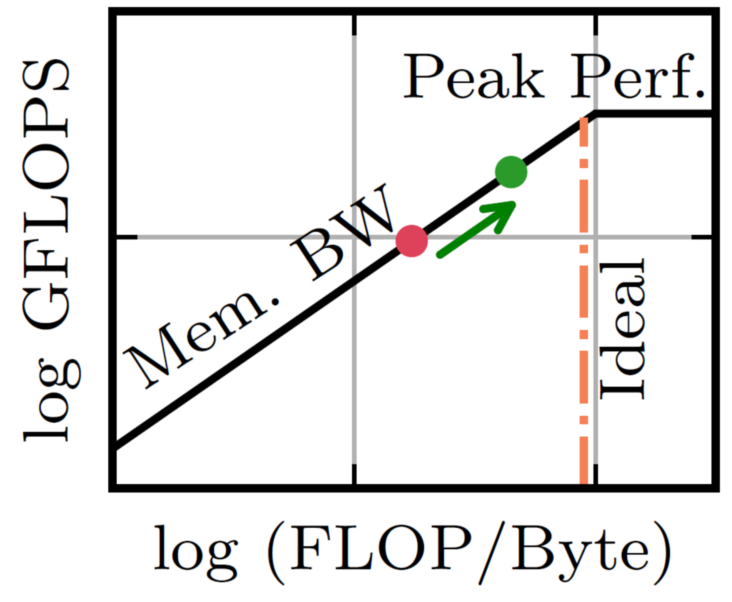
Power Breakdown for Eyeriss
Electrical digital



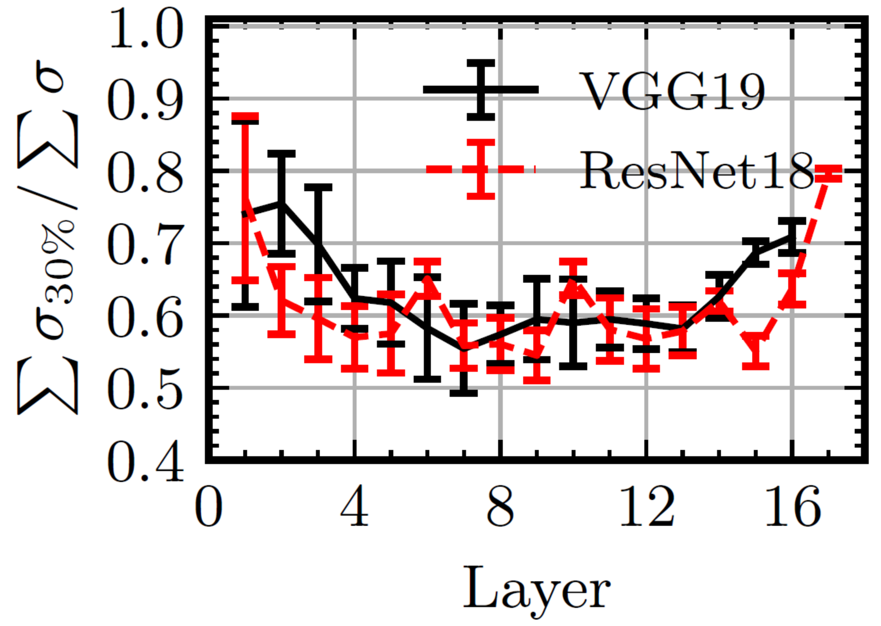
Power Breakdown for Mars
Photonics analog

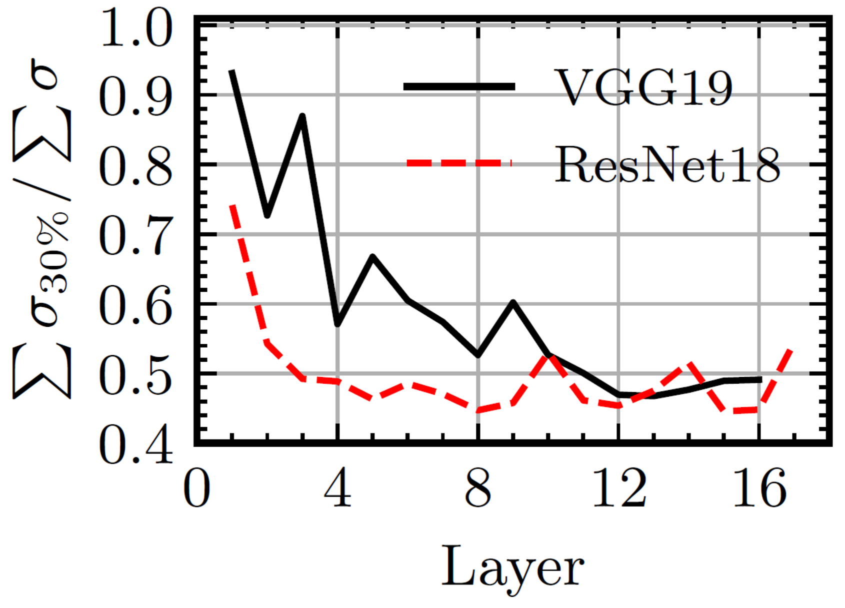
- Memory bottlenecked modern AI accelerators efficiency
- Naive layer decomposition does not save memory cost





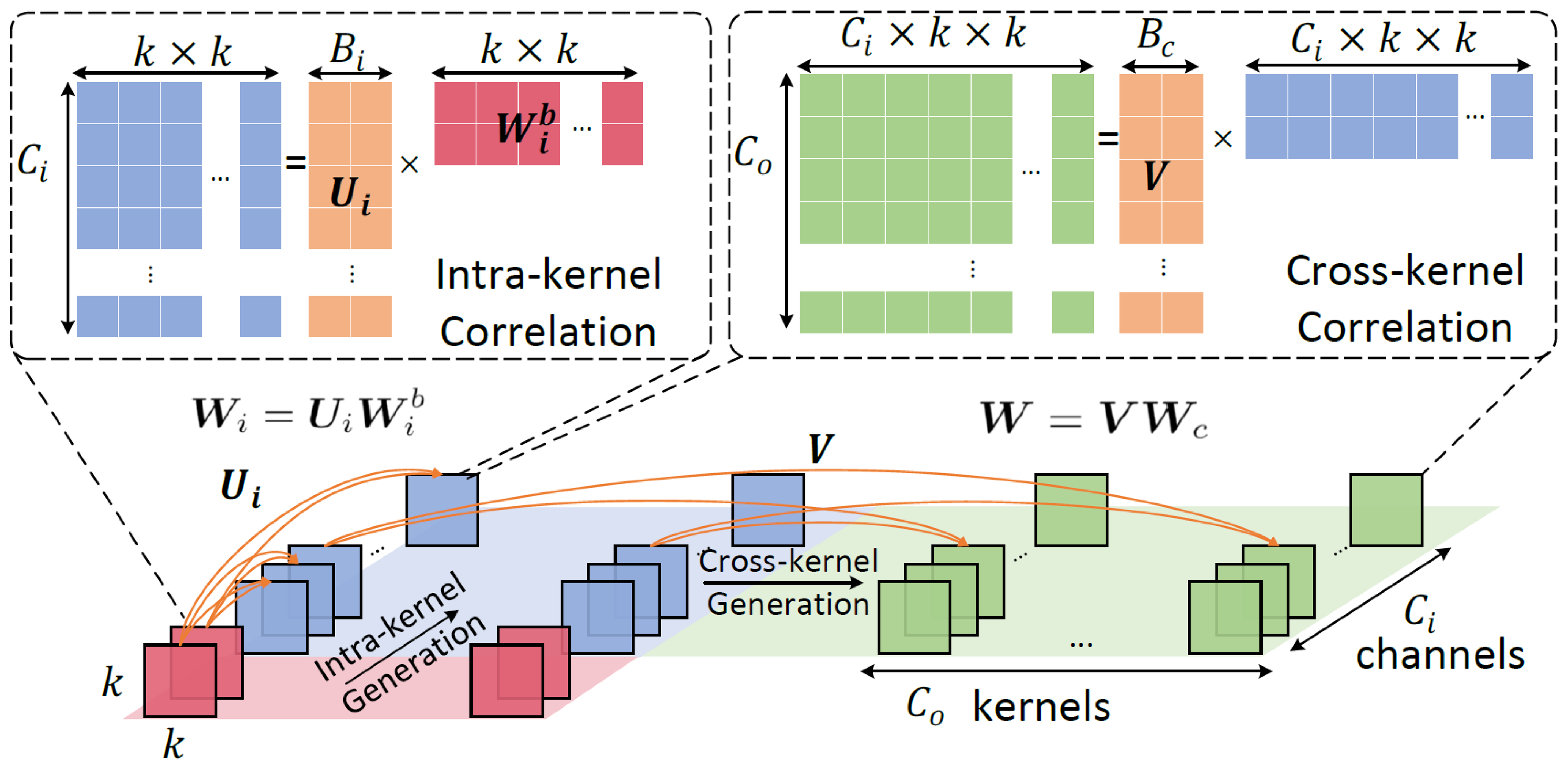
- Intrinsic intra-kernel and cross-kernel correlation in CNNs
- Motivate us to generate kernel on-the-fly with small basis





Proposed Multi-Level *in-situ* Generation:

- Intra-kernel generation $\boldsymbol{W}_i = \boldsymbol{U}_i \boldsymbol{W}_i^b, \quad \forall i \in [C_o]$
  - Span all input channels from a small basis $\boldsymbol{W}^b$
- Cross-kernel generation $\boldsymbol{W} = \boldsymbol{V}\boldsymbol{W}_c = \boldsymbol{V}\{\boldsymbol{U}_i \boldsymbol{W}_i^b\}_{i\in[B_c]}$
  - Span all kernels from a kernel basis
- Augmented mixed-precision generation
  - Assign different bitwidth to basis and coefficient



Proposed Training Flow:

- Project teacher onto decomposed low-bit students

$$\min \|\widehat{\mathcal{M}}(\widehat{\boldsymbol{W}}) - \mathcal{M}(\boldsymbol{W})\|_2^2 \approx \|\widehat{\boldsymbol{W}} - \boldsymbol{V}\{\boldsymbol{U}_i \boldsymbol{W}_i^b\}_{i\in[B_c]}\|_2^2$$

- Distill knowledge from teacher to students

$$\min \mathcal{L}_{KD} = \beta T^2 \mathcal{D}_{KL}(q_T, p_T) + (1-\beta) H(q, p_{T=1})$$

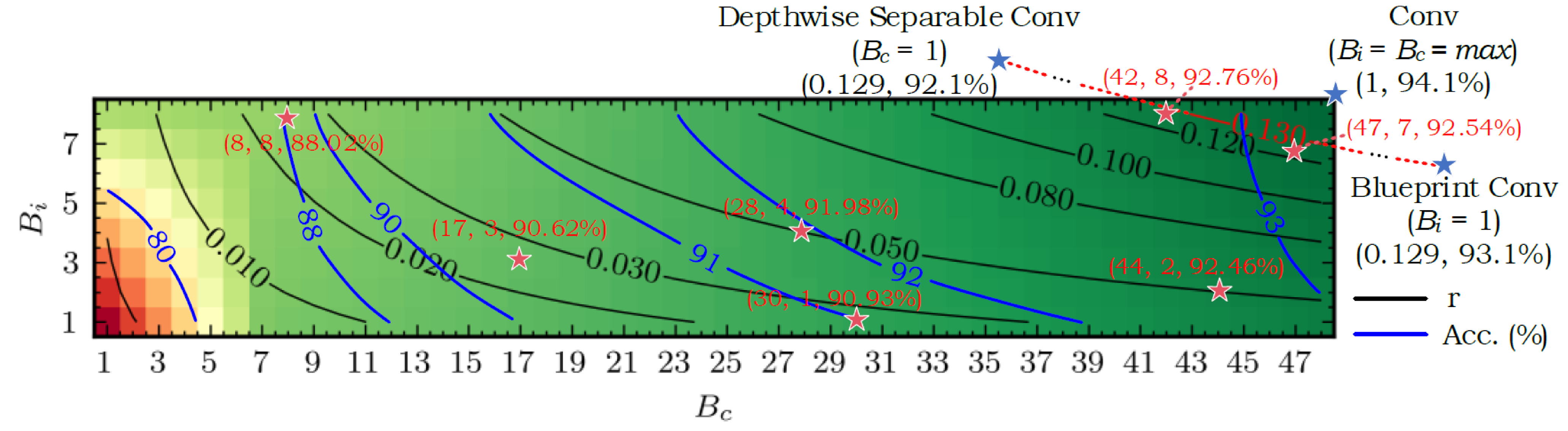
$$\text{s.t. } p_T = \frac{\exp(\frac{\mathcal{M}(\boldsymbol{W})}{T})}{\sum \exp(\frac{\mathcal{M}(\boldsymbol{W})}{T})}, q_T = \frac{\exp(\frac{\widehat{\mathcal{M}}(\widehat{\boldsymbol{W}})}{T})}{\sum \exp(\frac{\widehat{\mathcal{M}}(\widehat{\boldsymbol{W}})}{T})},$$

- Orthonormal regularization to encourage ranks

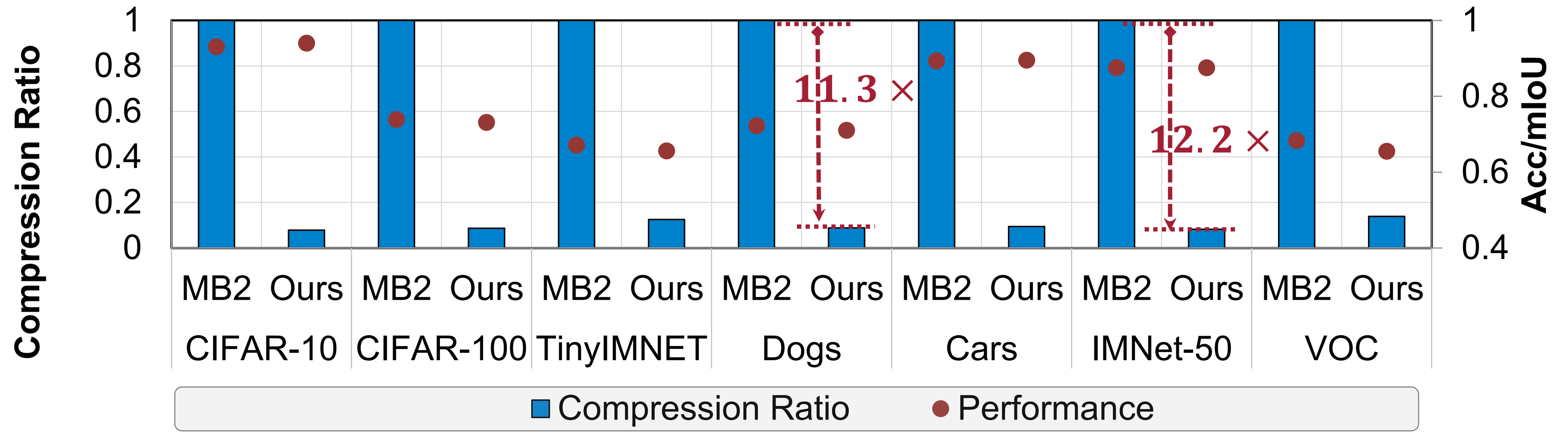
$$\sum_{i=1}^{B_c}\Big(\|\boldsymbol{W}_i^b(\boldsymbol{W}_i^b)^T - \boldsymbol{I}\|_2^2 + \|\tilde{\boldsymbol{U}}_i^T \tilde{\boldsymbol{U}} - \boldsymbol{I}\|_2^2\Big) + \|\tilde{\boldsymbol{V}}^T \tilde{\boldsymbol{V}} - \boldsymbol{I}\|_2^2$$

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× memory compression (3~5 bit quantization)
- ~30% less latency, ~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**× memory compression; **~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