2022 ICASSP

Compression-aware Projection with Greedy Dimension Reduction for Activations

Speaker: Yu-Shan (Clover) Tai

Email: clover@access.ee.ntu.edu.tw

Advisor: Prof. An-Yeu Wu

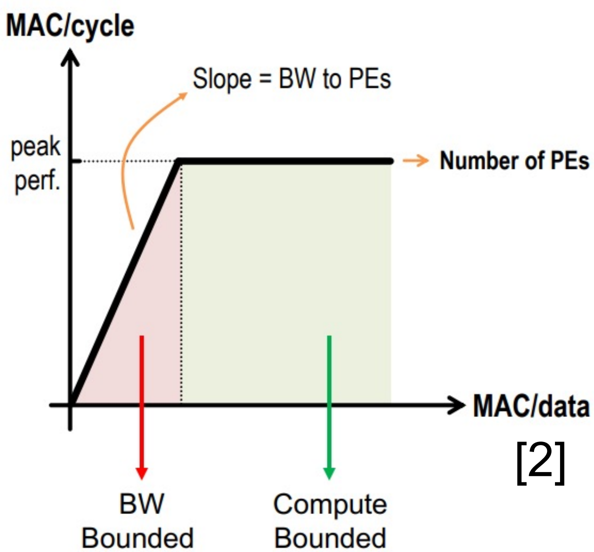
Date: 2022/5/

Outline

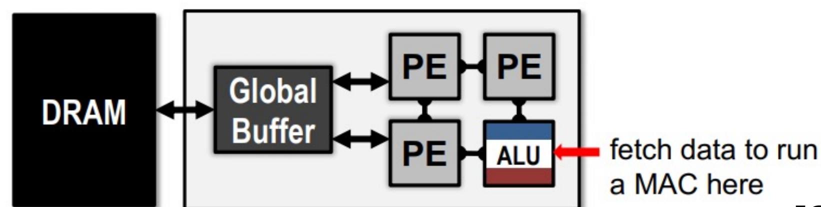
- Motivation
- Activation compression
- Proposed methods
  - Learnable projection
  - Select metric for greedy dimension reduction
- Simulation results and analysis
- Conclusion

Motivation: Limited Memory Bandwidth

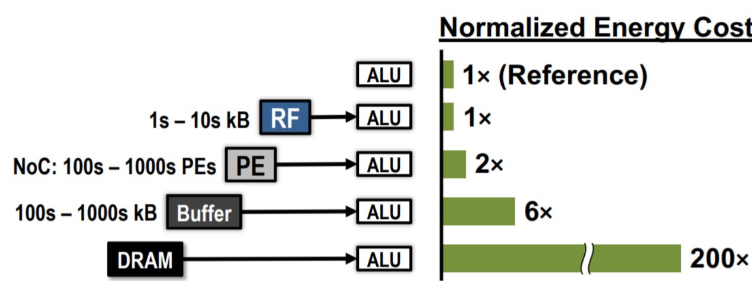
- **Large CNN model** is hard to deploy on **edge device**
- **Data movement** is more expensive than computation
  - Data movement from/to off-chip memory dominates energy footprint
  - Ex. in GoogLeNet, **68%** of energy consumption is due to data movement [1]



[2]



[2]



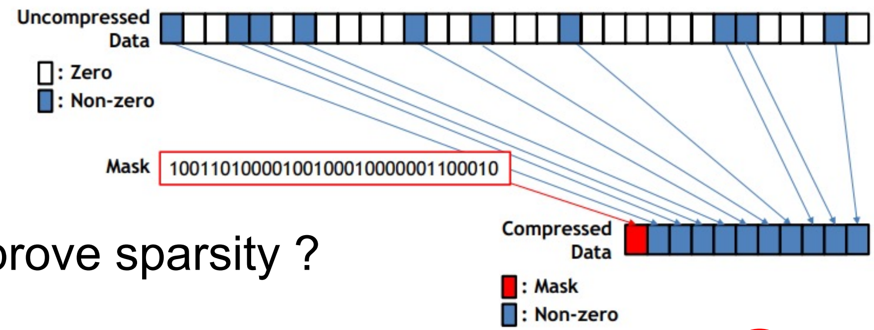
[1]

PE : Processing Element BW : Bandwidth MAC: multiply–accumulate operation

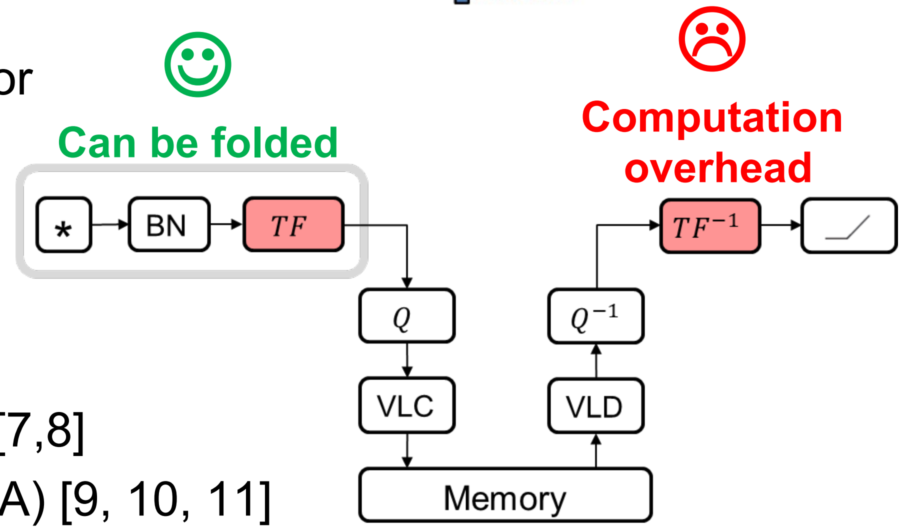
Require activation compression to reduce memory bandwidth!

Activation Compression

- Lossless encoder
  - RLE [3], ZVC [4], Huffman [5, 6]
  - **Too sensitive to sparsity**
    ➔ another method to further improve sparsity ?



- Transformation-based AC
  - Project data to a domain suitable for compression
    - **Decorrelate** data into important/ unimportant components
    - Enhance **sparsity** by discarding unimportant information
  - Discrete Cosine Transform (DCT) [7,8]
  - Principal Component Analysis (PCA) [9, 10, 11]

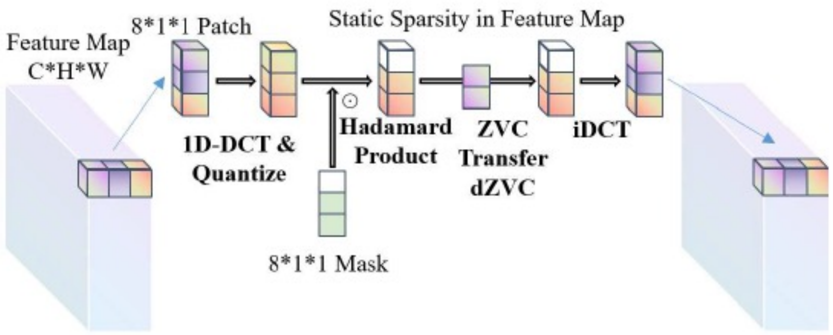


Transformation-based method projects data to domain with higher sparsity!

Transformation-based AC

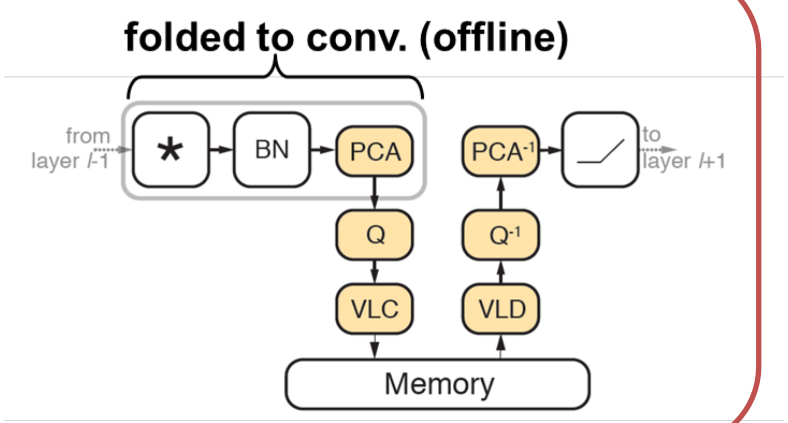
DCT [8]

- 1D-DCT on channel dimension
- Channel domain is **different** from natural figure
- Need to design a special mask for **channel sorting**



PCA [9]

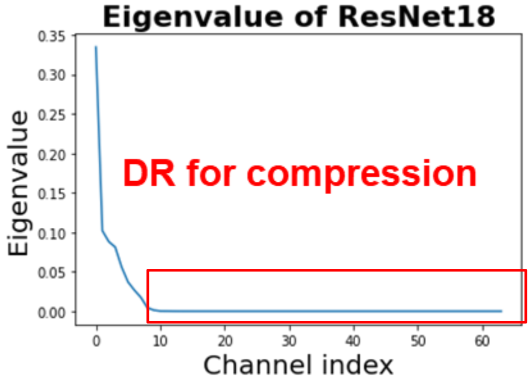
- PCA on channel dimension
- **Data dependent**
  ➔ enhance compressibility
- Eigenvalues helps to distinguish important/unimportant channels
  ➔ **dimension reduction**

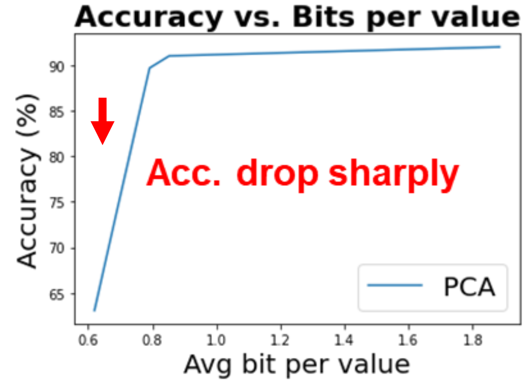


PCA is suitable for activation compression!

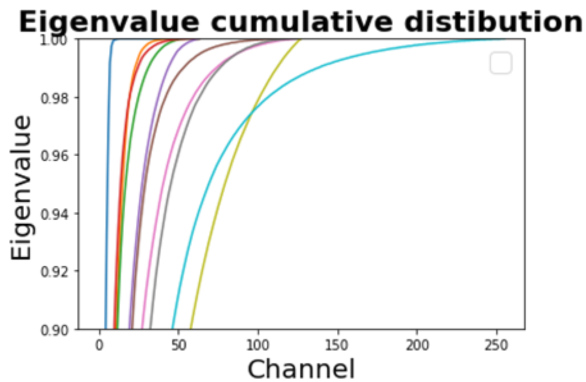
Challenges

- Further dimension reduction results in **severe performance degrade**
  - Enhance compression ratio but sacrifice accuracy disastrously

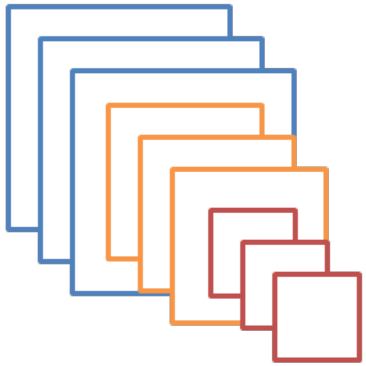




- DR to **same eigenvalue threshold** for every layer is non-ideal
  - The difference of distribution and size among each layer are ignored

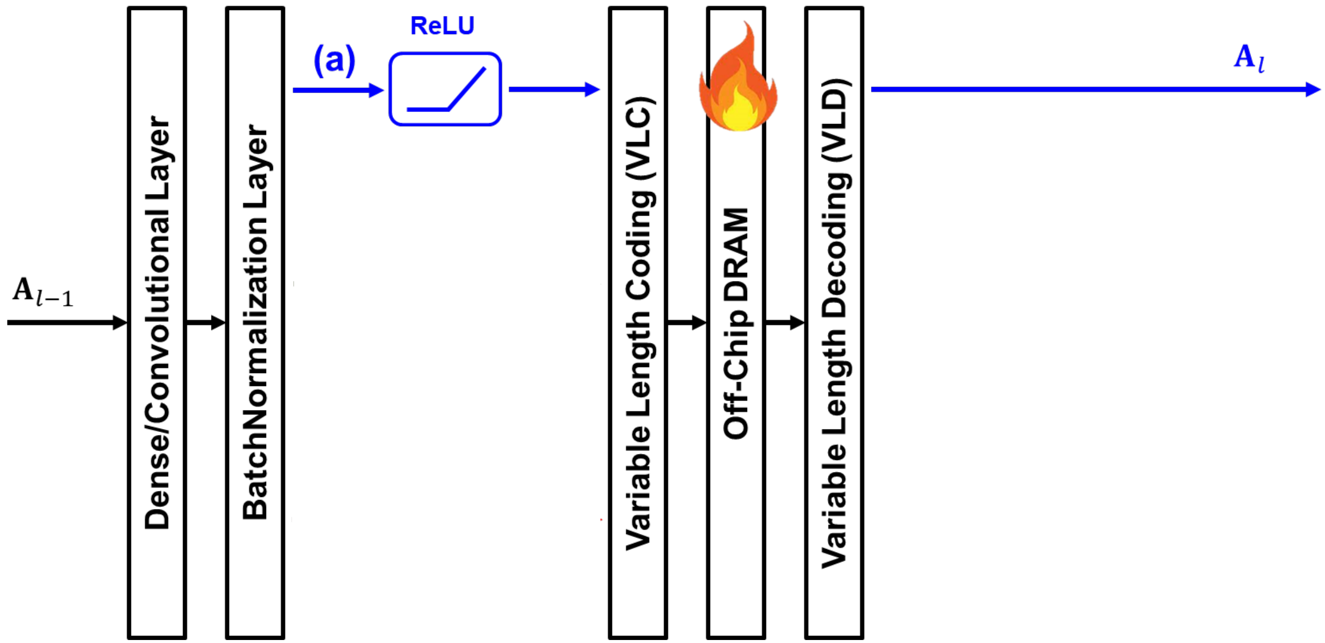


Different eigenvalue distribution



Different feature map size

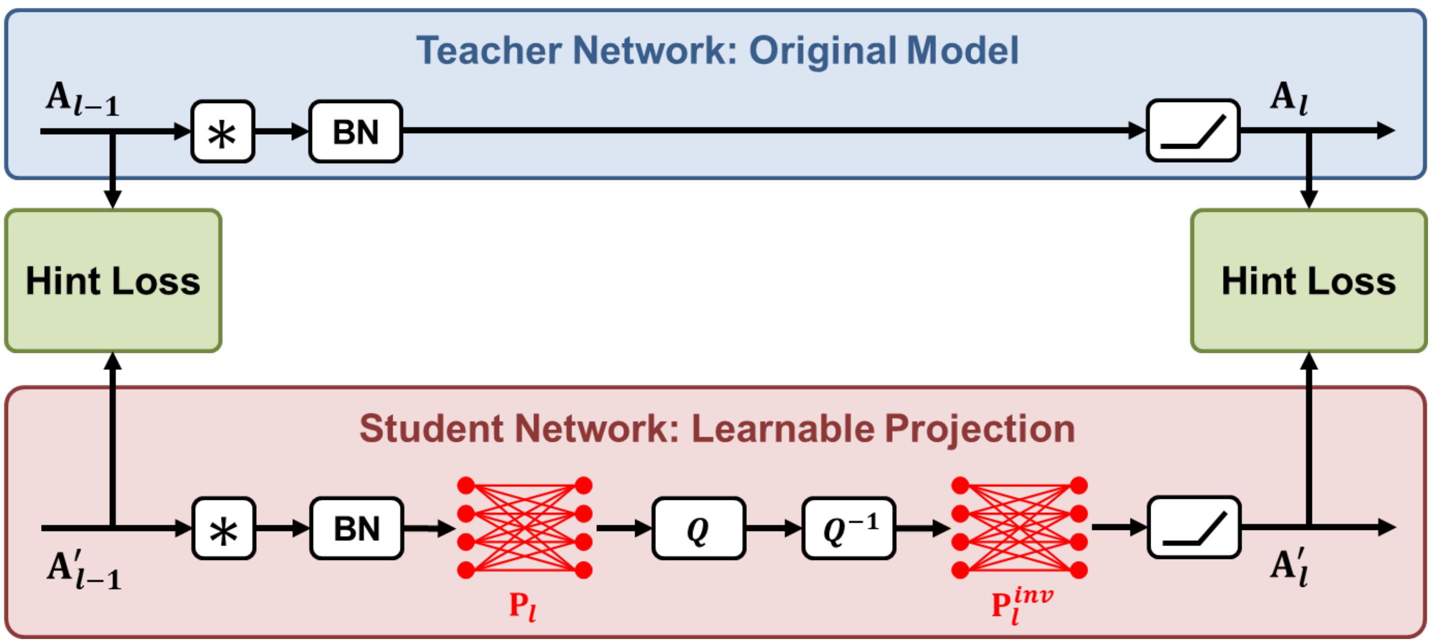
Proposed Compression-aware Projection with Greedy Dimension Reduction



- **Learnable Projection**
  - Replace PCA matrix with a trainable projection to compensate loss of DR
- **Selection Metric for Greedy DR**
  - Reduce dimension with consideration of accuracy and compression trade-off

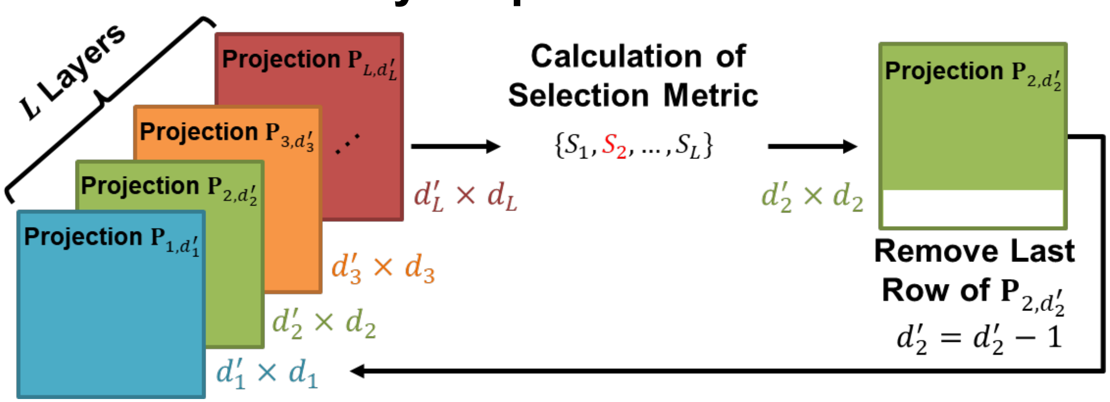
Learnable Projection

- Use **knowledge distillation** and **hint learning** to train PCA transformation matrix **without directly accessing labeled data**
  - Original model teach learnable projection model how to reconstruct well
  - **Mean square error** for hidden layer activation (hint loss)
  - **The Kullback-Leibler divergence** for output (knowledge distillation)



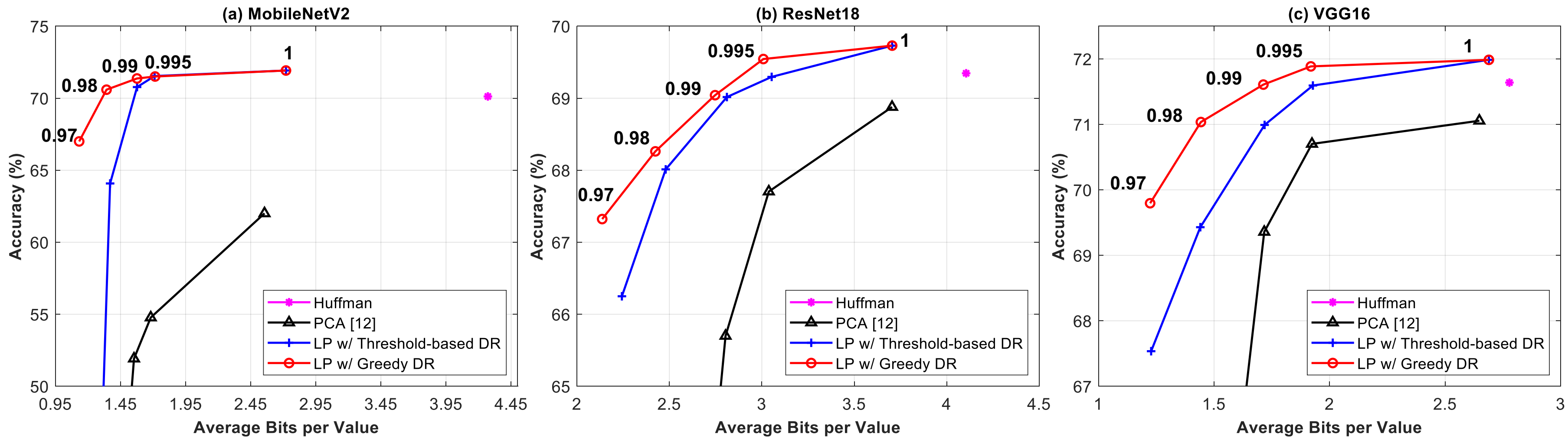
Selection Metric for Greedy Dimension Reduction

- Define a criterion to prioritize which layer for **dimension reduction**
- Selection metric: $S = \frac{\Delta accuracy}{\Delta activation\ bits}$
  - Use eigenvalues to approximate accuracy [11]: $\Delta accuracy = \sigma_{l,d'_l} / \sum_{c=1}^{d'_l} \sigma_{l,c}$
  - $\Delta activation\ bits = B\left(\mathbf{P}_{l,d'_l} \times \mathbf{A}'_l\right) - B\left(\mathbf{P}_{l,d'_l-1} \times \mathbf{A}'_l\right),$
    $B(\mathbf{A}) = \#bits\left(VLC\left(Q(\mathbf{A})\right)\right)$
- Lower S implies **low accuracy drop** and **high activation bits reduction**
  - Jointly consider **accuracy drop** and **bits reduction** to achieve better tradeoff



Simulation Results (1/3)

Comparison between Different AC Methods

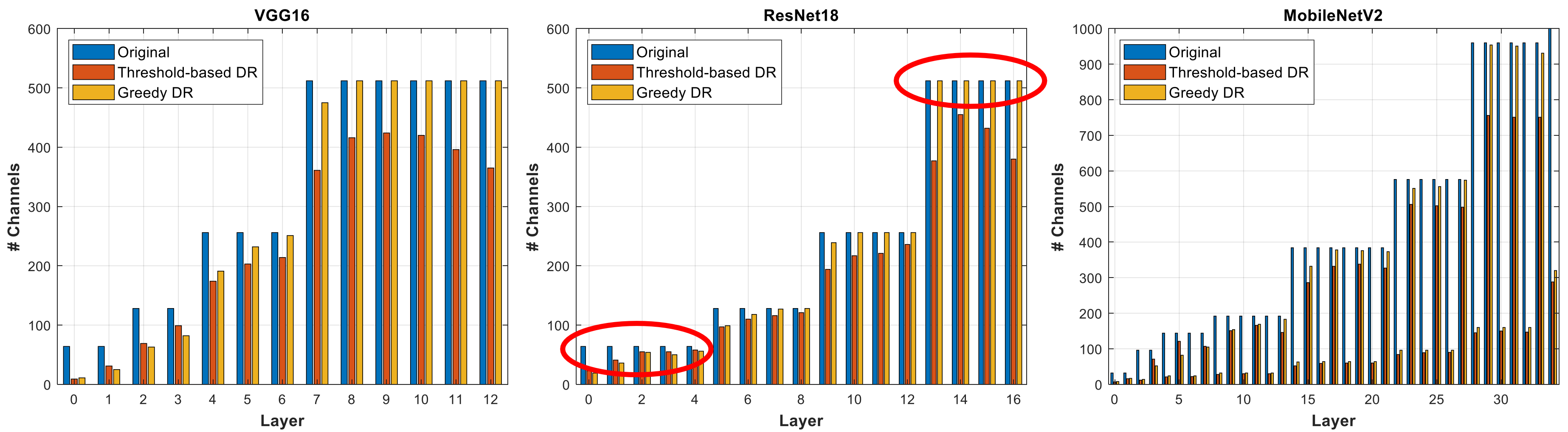


Simulation Settings	
Dataset	ImageNet [12]
Model	MobileNetV2/ ResNet18/VGG16
Weight Bit Width	8
Activation Bit Width	8
Eigenvalue Threshold	[0.97,0.98,0.99 0.995,1]
Learning rate	0.001
# epoch	3

- **MobileNetV2** reaches **0.6%** accuracy drop with average **1.34 bits(5.97x)** per value
- **ResNet18** reaches **0.4%** accuracy drop with average **2.75 bits(2.91x)** per value
- **VGG16** reaches **0.6%** accuracy drop with average **1.44 bits(5.56x)** per value

Simulation Results (2/3)

Analysis of Dimension Reduction Distribution



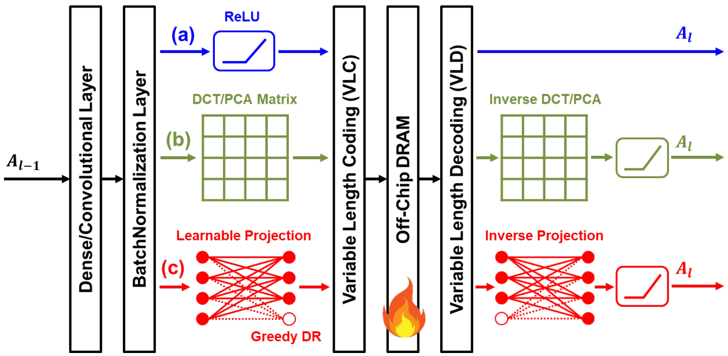
- **Greedy DR** tends to maintain higher #channels for deep layers than threshold-based DR
- Compressing shallow layers leads to high bits reduction but low accuracy drop
  - For ResNet18, size of 1st layer activation is 112 x 112 while that of last layer is 7 x 7

Simulation Results (3/3)

Computation Analysis under Different Threshold

- Forward transform can be folded into Conv. and BN
- The only induced computation comes from inverse projection
- Relative computation

$$= \frac{(C_{Original} + C_{Learnable}) + C_{Inverse}}{C_{Original}} \times 100\% = \frac{C_{Folded} + C_{Inverse}}{C_{Original}} \times 100\%$$

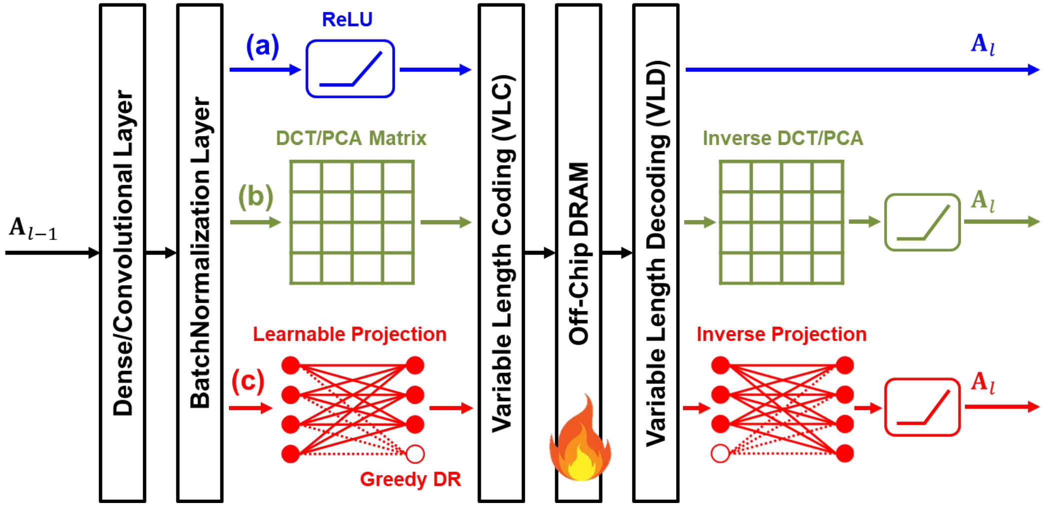


Model	0.97	0.98	0.99	0.995	1
MobileNetV2	59.7%	74.3%	89.5%	98.0%	126.8%
ResNet18	78.5%	86.4%	93.8%	99.3%	114.3%
VGG16	57.8%	70.0%	83.1%	92.2%	114.4%

- Our method needs less computation than PCA transformation method
  - Even **lower** computation than original model

Conclusion

- Our method reduces **2.91x~5.97x** memory access with **0.4~0.7%** negligible accuracy drop on MobileNetV2/ResNet18/VGG16
- **Learnable projection** can compensate compression loss without directly accessing labeled data
- Selection metric for **greedy DR**
  - Consider both bits reduction and accuracy drop simultaneously
  - Decide DR ratio for each layer automatically



Reference

[1] V. Sze, Y.-H. Chen, T.-J. Yang and J. S. Emer, “Efficient Processing of Deep Neural Networks: A Tutorial and Survey,” *Proceedings of the IEEE*, vol. 105, pp. 2295-2329, 2017.

[2] T.-J. Yang, Y.-H. Chen, J. Emer and V. Sze, “A method to estimate the energy consumption of deep neural networks,” in *Proc. Asilomar Conference on Signals, Systems, and Computers*, 2017.

[3] A. Parashar, et al., “SCNN: An accelerator for compressed-sparse convolutional neural networks,” in *Proc. ACM/IEEE Annual International Symposium on Computer Architecture (ISCA)*, 2017, pp. 27–40.

[4] M. Rhu, M. O'Connor, N. Chatterjee, J. Pool, Y. Kwon and S. W. Keckler, “Compressing DMA Engine: Leveraging Activation Sparsity for Training Deep Neural Networks,” in *IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 2018.

[5] M. Chandra, “Data Bandwidth Reduction in Deep Neural Network SoCs using History Buffer and Huffman Coding,” in *Proc. International Conference on Computing, Power and Communication Technologies (GUCON)*, 2018.

[6] S. Han, H. Mao and W. J. Dally, “Deep Compression: Compressing Deep Neural Network with Pruning, Trained Quantization and Huffman Coding,” in *International Conference on Learning Representations (ICLR)*, San Juan, Puerto Rico, May 2-4, 2016.

Reference

[7] R. D. Evans, L. Liu and T. M. Aamodt, "JPEG-ACT: Accelerating Deep Learning via Transform-based Lossy Compression," in *Proc. ACM/IEEE Annual International Symposium on Computer Architecture (ISCA)*, 2020.

[8] Y. Shi, M. Wang, S. Chen, J. Wei and Z. Wang, "Transform-Based Feature Map Compression for CNN Inference," in *Proc. IEEE International Symposium on Circuits and Systems (ISCAS)*, 2021.

[9] B. Chmiel, C. Baskin, R. Banner, E. Zheltonozhskii, Y. Yermolin, A. Karbachevsky, A. Bronstein and A. Mendelson, "Feature Map Transform Coding for Energy-Efficient CNN Inference," in *Proc. International Joint Conference on Neural Networks (IJCNN)*, pp. 1-9, 2020.

[10] F. Xiong, F. Tu, M. Shi, Y. Wang, L. Liu, S. Wei and S. Yin, "STC: Significance-aware Transform-based Codec Framework for External Memory Access Reduction," in *2020 57th ACM/IEEE Design Automation Conference (DAC)*, 2020.

[11] X. Zhang, J. Zou, X. Ming, K. He and J. Sun, "Efficient and accurate approximations of nonlinear convolutional networks," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1984-1992, 2015.

[12] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, pp. 211-252, 2015.

The end

Thank you for your listening