DeepCompress-ViT: Rethinking Model Compression to Enhance Efficiency of Vision Transformers at the Edge (CVPR 2025)

Sabbir Ahmed
PhD Candidate

BINGHAMTON UNIVERSITY
STATE UNIVERSITY OF NEW YORK

NC STATE UNIVERSITY

NJIT
New Jersey Institute of Technology

Georgia Tech

Adobe

Problem Statement

- Deployment of ViTs on edge devices presents significant challenges due to their substantial memory requirements (ViT-S requires 84 MB)
- This requirement exceeds on chip memory limit (1-8MB) in edge devices
- Model parameters must be stored in off-chip memory, requiring a sequential processing approach during inference
- This involves iteratively loading each layer’s weights into on-chip memory and performing computations on that layer before proceeding to the next
- Introduces significant latency and increased energy consumption due to frequent off-chip memory access

Possible Solution (Model Compression)

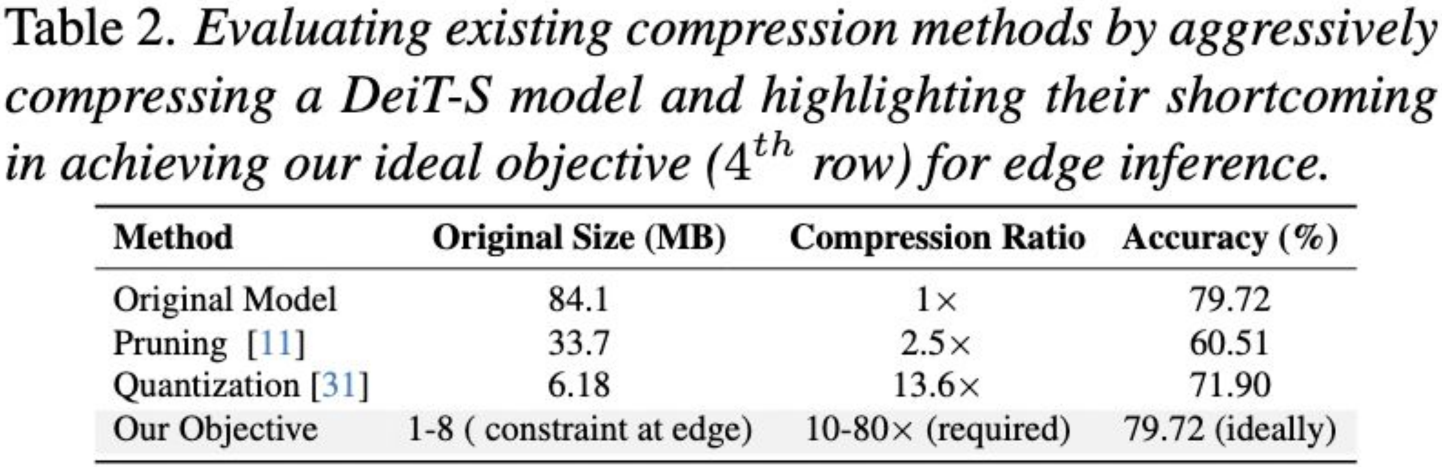
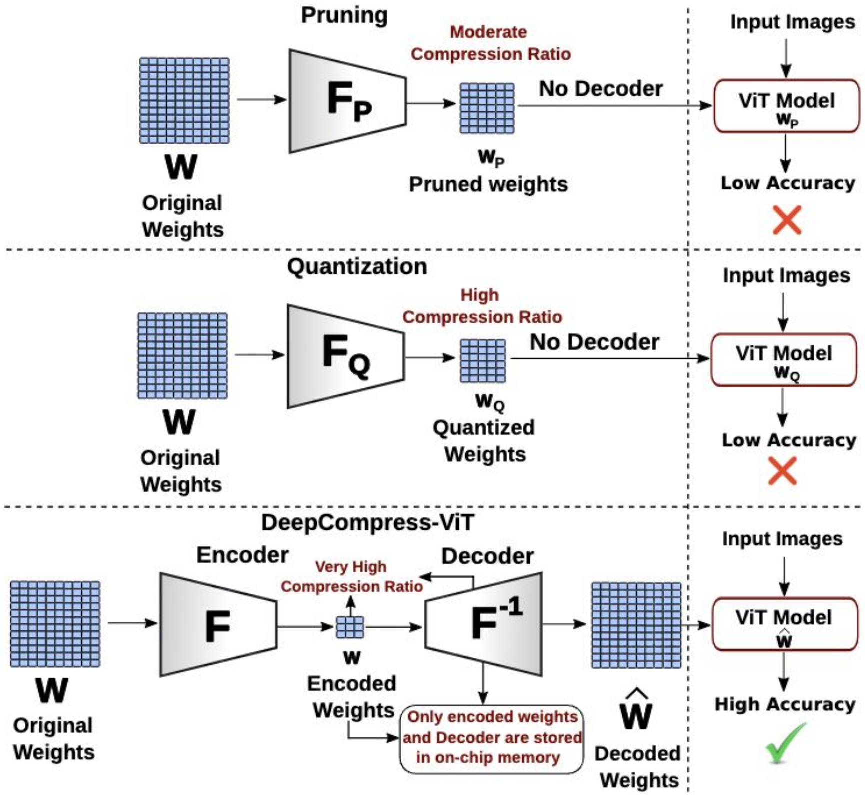


Table 2. *Evaluating existing compression methods by aggressively compressing a DeiT-S model and highlighting their shortcoming in achieving our ideal objective (4^{th} row) for edge inference.*

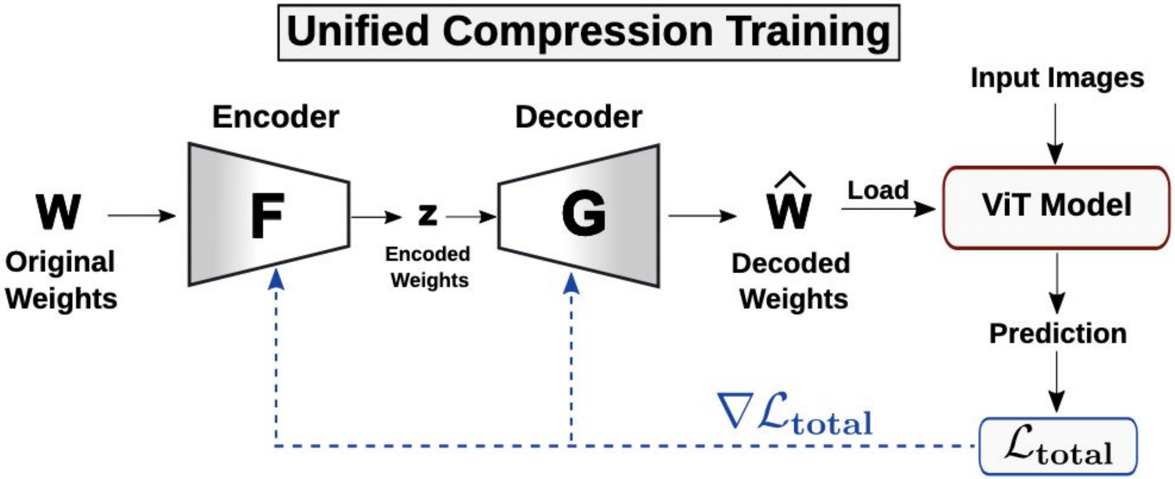
Method	**Original Size (MB)**	**Compression Ratio**	**Accuracy (%)**
Original Model	84.1	1×	79.72
Pruning [11]	33.7	2.5×	60.51
Quantization [31]	6.18	13.6×	71.90
Our Objective	1-8 (constraint at edge)	10-80× (required)	79.72 (ideally)

Bottleneck in existing Model Compression



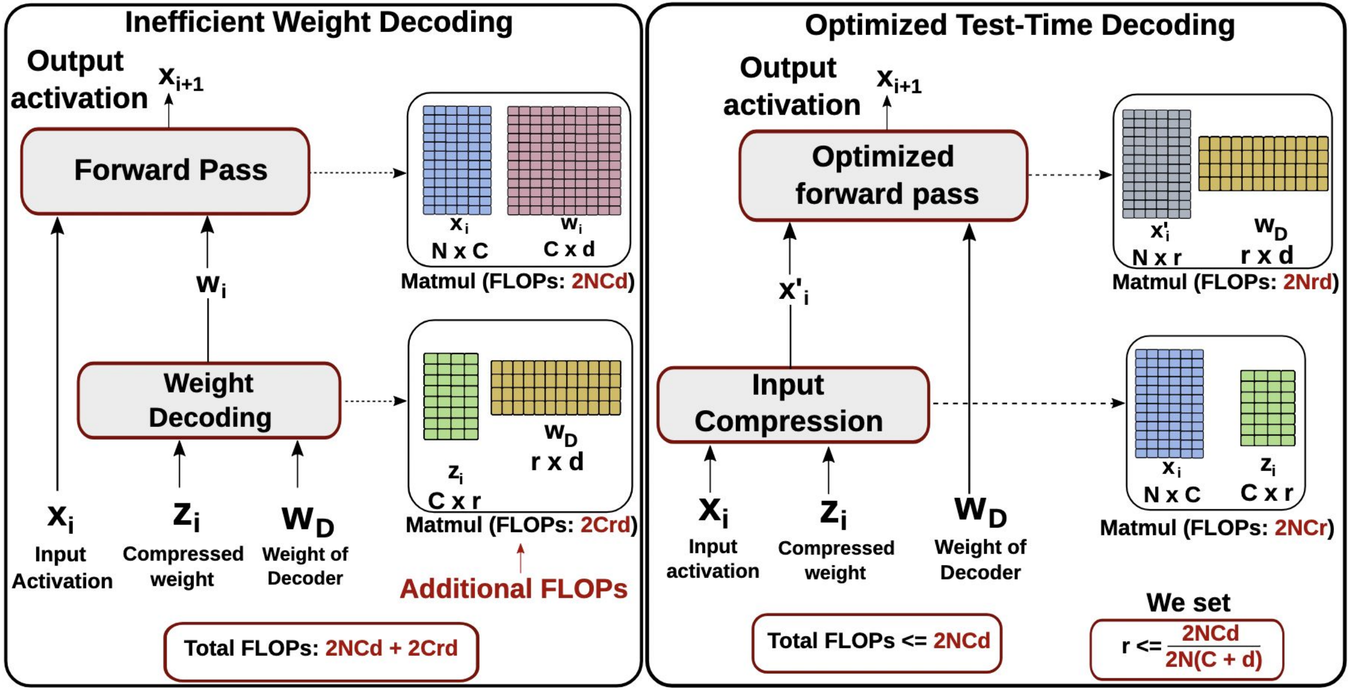
DeepCompress-ViT

To achieve high compression ratio we design UCT

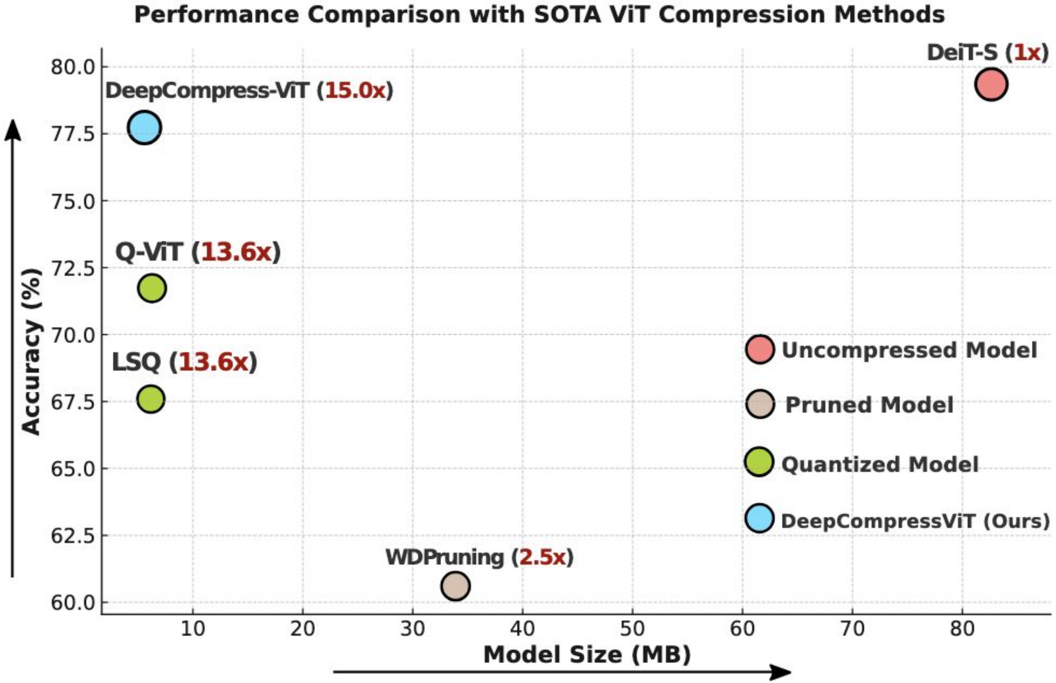


DeepCompress-ViT

To prevent FLOP count increase we design Optimized Test-Time Decoding



Results



Performance Comparison with SOTA ViT Compression Methods
DeiT-S (1x)
DeepCompress-ViT (15.0x)
Q-ViT (13.6x)
LSQ (13.6x)
WDPruning (2.5x)
Uncompressed Model
Pruned Model
Quantized Model
DeepCompressViT (Ours)
Accuracy (%)
80.0
77.5
75.0
72.5
70.0
67.5
65.0
62.5
60.0
Model Size (MB)
10
20
30
40
50
60
70
80

Summary

- Deployment of ViTs on edge devices presents significant challenges requiring frequent off-chip memory for loading weights
- Existing compression techniques are ineffective to compress ViTs to store the entire model weights on on-chip memory particularly because of large accuracy drop at high compression ratio
- To address this issue, in this work, we propose an orthogonal compression technique that involves encoding and efficient weight decoding at runtime to achieve high compression ratio with minimal performance degradation

NC STATE
UNIVERSITY

NJIT
New Jersey Institute
of Technology

GT
Georgia
Tech

Adobe

Thank You!