



RUTGERS  
THE STATE UNIVERSITY  
OF NEW JERSEY



WASHINGTON STATE  
UNIVERSITY

# HALOC: Hardware-Aware Automatic Low-Rank Compression for Compact Neural Networks

---

Jinqi Xiao<sup>1</sup>, Chengming Zhang<sup>2</sup>, Yu Gong<sup>1</sup>, Miao Yin<sup>1</sup>, Yang Sui<sup>1</sup>,  
Lizhi Xiang<sup>3</sup>, Dingwen Tao<sup>2,3</sup>, Bo Yuan<sup>1</sup>  
<sup>1</sup> Rutgers University, <sup>2</sup> Indiana University,  
<sup>3</sup> Washington State University





# CONTENT

---

**01**

**Motivation**

**02**

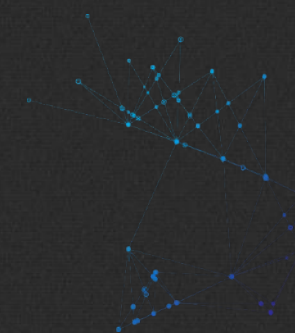
**Method**

**03**

**Ablation Study**

**04**

**Comparison**



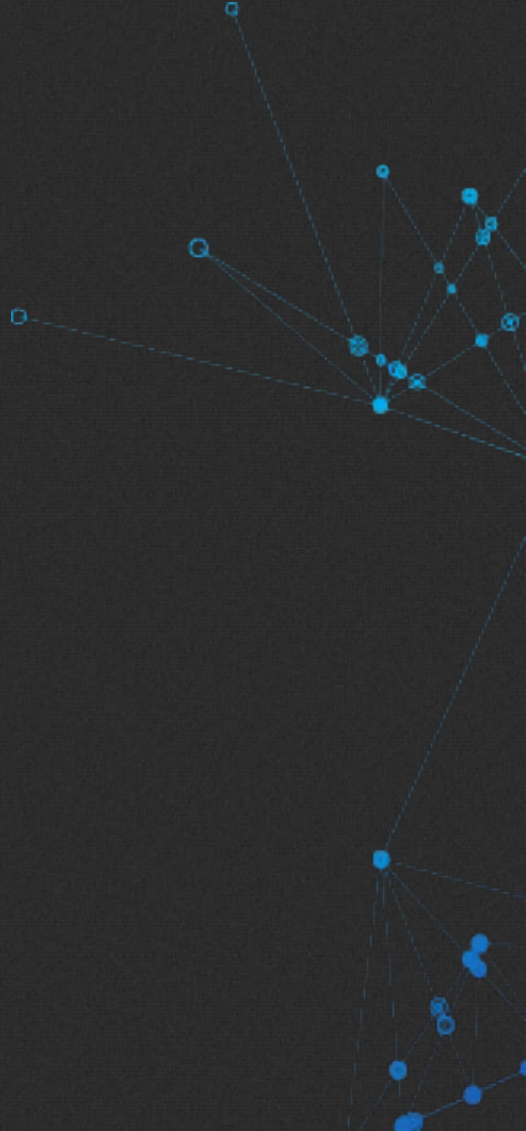


# Motivation



---

Low-rank compression is an important model compression strategy for obtaining compact neural network models.

- ❖ The rank values directly determine the model complexity and model accuracy; proper selection of layer-wise rank is very critical and desired.
  - ❖ All existing works are not designed in a hardware-aware way, limiting the practical performance of the compressed models on real-world hardware platforms.
- 

# Method

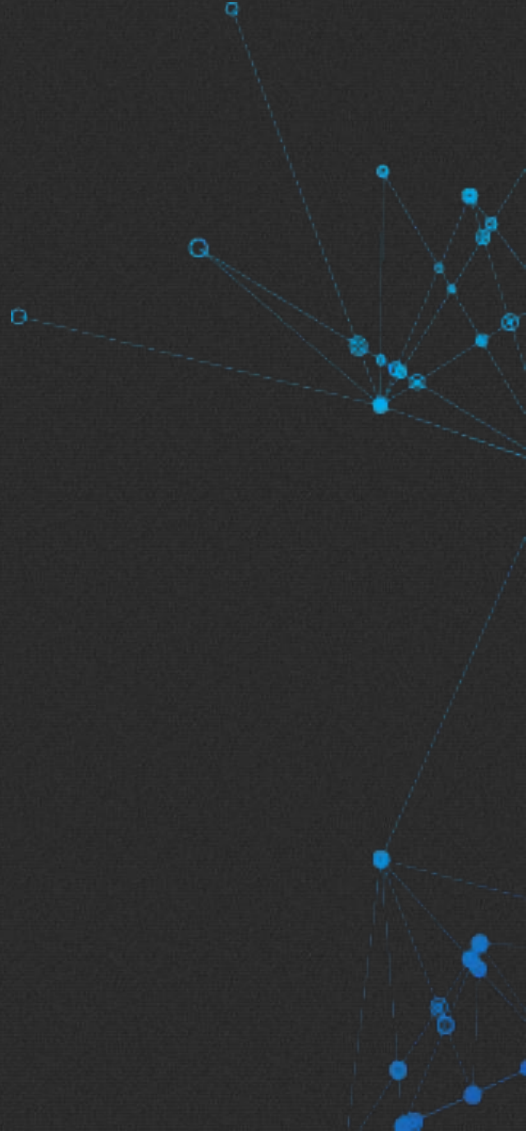


---

## ❖ Problem Formulation

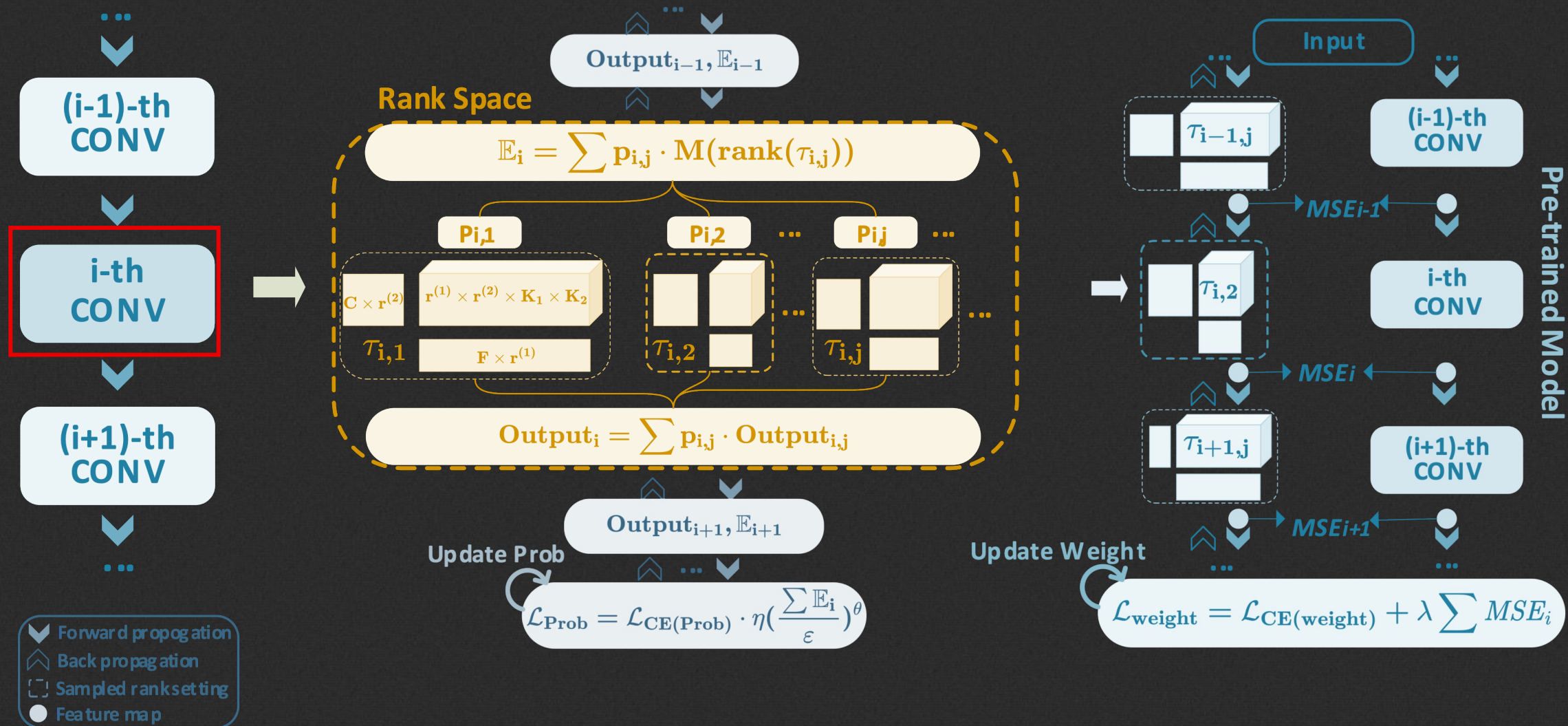
$$\min_{\{W_i\}_{i=1}^n} \mathcal{L}(\{W_i\}) \quad \text{s.t.} \quad \sum_{i=1}^n \hbar(\text{rank}(W_i)) \leq \varepsilon$$

## ❖ Design Challenges

- Insufficient exploration for the rank space
  - The search process of the state-of-the-art rank determination works cannot be extended to consider the hardware performance constraint
- 

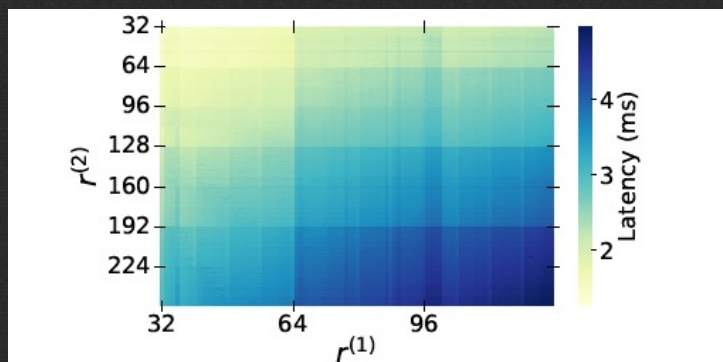


# Method

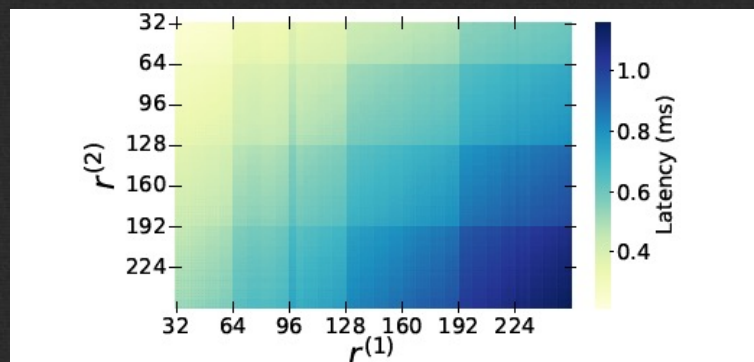


# Method

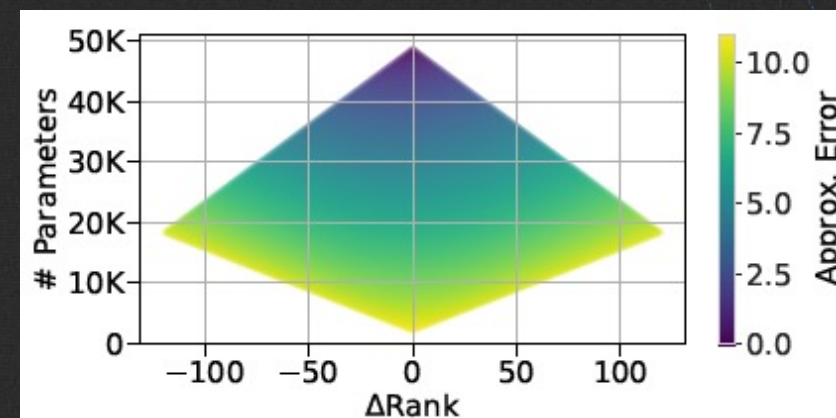
- ❖ How should we set the proper search scope to realize sufficient exploration in the rank space with affordable search cost?



(a) Layer3.0.conv on Nvidia RTX 2080. Batch size is 128.



(b) Layer3.0.conv2 on Nvidia Tesla V100. Batch size is 128.



- **Design Principle-1:** To make good balance between search cost and rank granularity, the rank candidates in HALOC is set as the multiples of a constant (typically 32).
- **Design Principle-2:** For a Tucker-2-format layer, equal rank setting ( $r_1 = r_2$ ) can be adopted to simplify the rank search process with good approximation performance.}



# Method



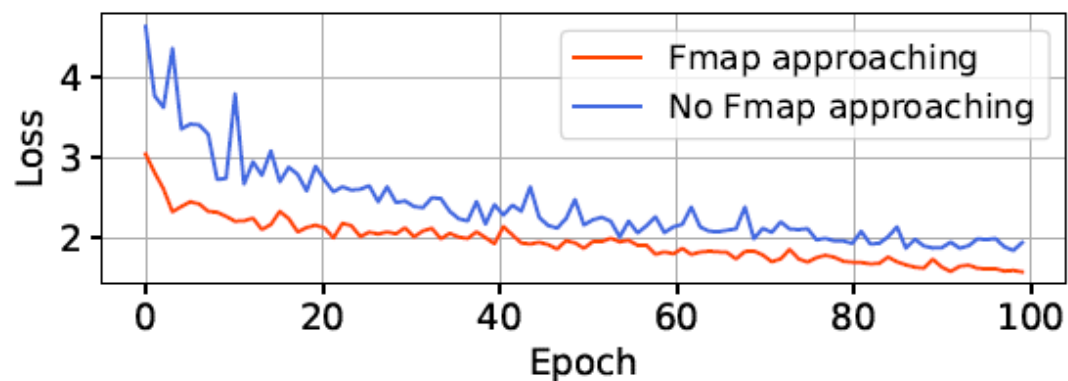
- ❖ What is the proper scheme to mitigate the interference between different selected rank settings?

$$\mathcal{L}_{approach} = \sum_{i=1}^n MSE(Fmap_{decomp,i}, Fmap_{org,i})$$

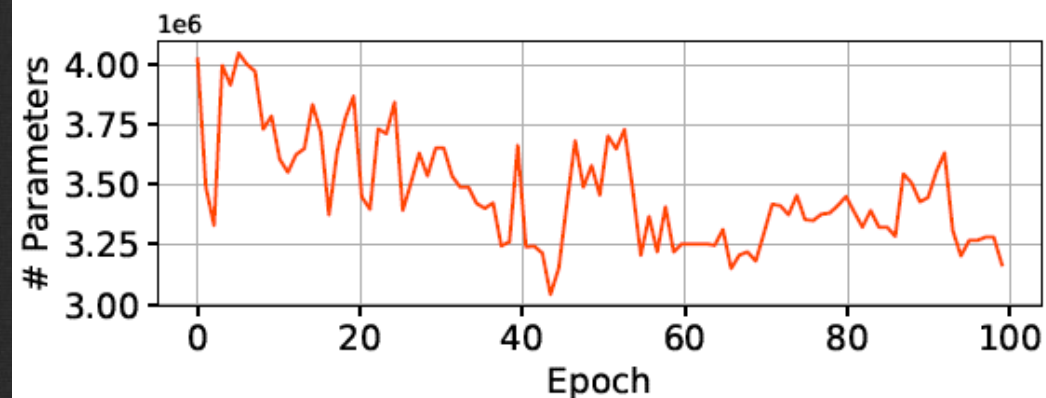
$$\mathcal{L}_{weight} = \mathcal{L}_{CE(weight)} + \lambda \mathcal{L}_{approach}$$


# Ablation Study

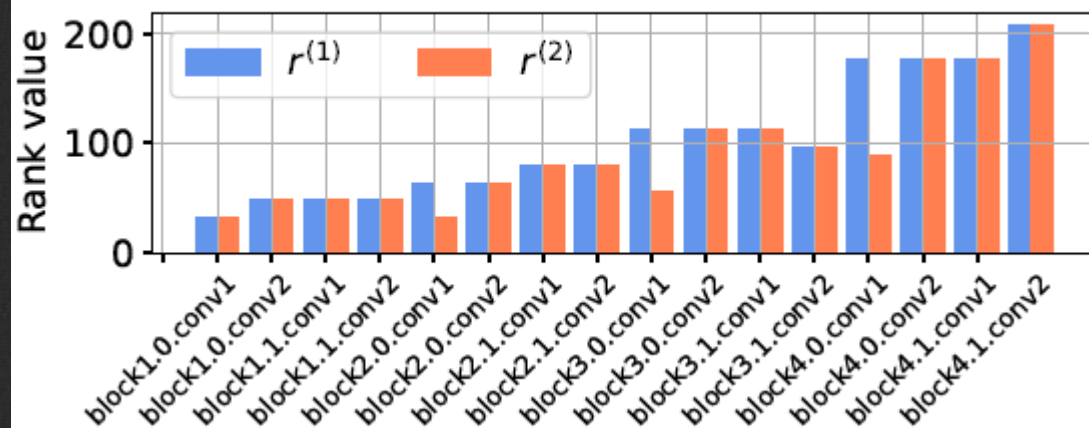
(a) The curve of training loss ( $\mathcal{L}_{weight}$ )



(b) The change of the model size in the rank search process



(c) The final rank distribution after automatic rank search





# Comparison

Method	Comp. Type	Auto. Rank	Top-1 (%)	Top-5 (%)	FLOPs (↓%)	Params. (↓%)
ResNet-18	Baseline	-	69.75	89.08	-	-
<b>HALOC</b>	Low-rank	✓	<b>70.65</b>	<b>89.42</b>	<b>66.16</b>	63.64
<b>HALOC</b>	Low-rank	✓	70.14	89.38	63.81	<b>71.31</b>
ALDS	Low-rank	✓	69.22	89.03	43.51	66.70
TETD	Low-rank	✗	-	89.08	59.51	-
Stable EPC	Low-rank	✓	-	89.08	59.51	-
MUSCO	Low-rank	✗	69.29	88.78	58.67	-
CHEX	Pruning	-	69.60	-	43.38	-
EE	Pruning	-	68.27	88.44	46.60	-
SCOP	Pruning	-	69.18	88.89	38.80	39.30
MobileNetV2	Baseline	-	71.85	90.33	-	-
<b>HALOC</b>	Low-rank	✓	<b>70.98</b>	<b>89.77</b>	24.84	40.03
<b>HALOC</b>	Low-rank	✓	66.37	87.02	<b>45.65</b>	<b>62.59</b>
ALDS	Low-rank	✓	70.32	89.60	11.01	32.97
HOSA	Pruning	-	64.43	-	43.65	27.13
DCP	Pruning	-	64.22	-	44.75	25.93
FT	Pruning	-	70.12	89.48	20.23	21.31

ResNet-20 and VGG-16 on CIFAR-10

Method	Comp. Type	Auto. Rank	Top-1 (%)	Top-5 (%)	FLOPs (↓%)	Params. (↓%)
ResNet-18	Baseline	-	69.75	89.08	-	-
<b>HALOC</b>	Low-rank	✓	<b>70.65</b>	<b>89.42</b>	<b>66.16</b>	63.64
<b>HALOC</b>	Low-rank	✓	70.14	89.38	63.81	<b>71.31</b>
ALDS	Low-rank	✓	69.22	89.03	43.51	66.70
TETD	Low-rank	✗	-	89.08	59.51	-
Stable EPC	Low-rank	✓	-	89.08	59.51	-
MUSCO	Low-rank	✗	69.29	88.78	58.67	-
CHEX	Pruning	-	69.60	-	43.38	-
EE	Pruning	-	68.27	88.44	46.60	-
SCOP	Pruning	-	69.18	88.89	38.80	39.30
MobileNetV2	Baseline	-	71.85	90.33	-	-
<b>HALOC</b>	Low-rank	✓	<b>70.98</b>	<b>89.77</b>	24.84	40.03
<b>HALOC</b>	Low-rank	✓	66.37	87.02	<b>45.65</b>	<b>62.59</b>
ALDS	Low-rank	✓	70.32	89.60	11.01	32.97
HOSA	Pruning	-	64.43	-	43.65	27.13
DCP	Pruning	-	64.22	-	44.75	25.93
FT	Pruning	-	70.12	89.48	20.23	21.31

ResNet-18 and MobileNetV2 on ImageNet

# Comparison

Hardware	Method	ResNet-18				MobileNetV2			
		Top-1 (%)	Top-5 (%)	FLOPs (M)	Throughput (images/s)	Top-1 (%)	Top-5 (%)	FLOPs (M)	Throughput (images/s)
NVIDIA Tesla V100	Original	69.75	89.08	1819.07	4362.1	71.85	90.33	314.19	3877.3
	<b>HALOC</b>	69.75	88.93	<b>553.13</b>	<b>6360.5</b>	70.86	89.77	<b>245.52</b>	<b>3993.6</b>
NVIDIA Jetson TX2	Original	69.75	89.08	1819.07	86.3	71.85	90.33	314.19	112.1
	<b>HALOC</b>	70.14	89.38	<b>658.26</b>	<b>151.0</b>	70.80	89.55	<b>240.99</b>	<b>117.0</b>
ASIC Eyeriss	Original	69.75	89.08	1819.07	121.4	71.85	90.33	314.19	496.3
	<b>HALOC</b>	70.65	89.42	<b>615.62</b>	<b>247.0</b>	70.83	89.65	<b>229.13</b>	<b>590.2</b>

Table 3: Measured Speedup for compressed ResNet-18 and MobileNetV2 on different computing platforms. Hardware-aware automatic rank selection is adopted in the low-rank compression process.





RUTGERS  
THE STATE UNIVERSITY  
OF NEW JERSEY



WASHINGTON STATE  
UNIVERSITY

A complex network diagram consisting of numerous blue circular nodes of varying sizes connected by thin blue lines. The nodes are distributed across the left and center of the slide, with a higher density in the upper-left and lower-left areas, and a few nodes extending towards the right.

Thank you for listening

---