



DESIGN, AUTOMATION & TEST IN EUROPE

25 - 27 March 2024 · Valencia, Spain

The European Event for Electronic
System Design & Test

FMTT : Fused Multi-head Transformer with Tensor-compression for 3D Point Clouds Detection on Edge Devices

Zikun Wei¹, Tingting Wang¹, Chenchen Ding¹, Bohan Wang¹, Ziyi Guan^{2^}, Hantao Huang¹, and Hao Yu¹

¹*School of Microelectronics, Southern University of Science and Technology, Shenzhen, China*

²*Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong*

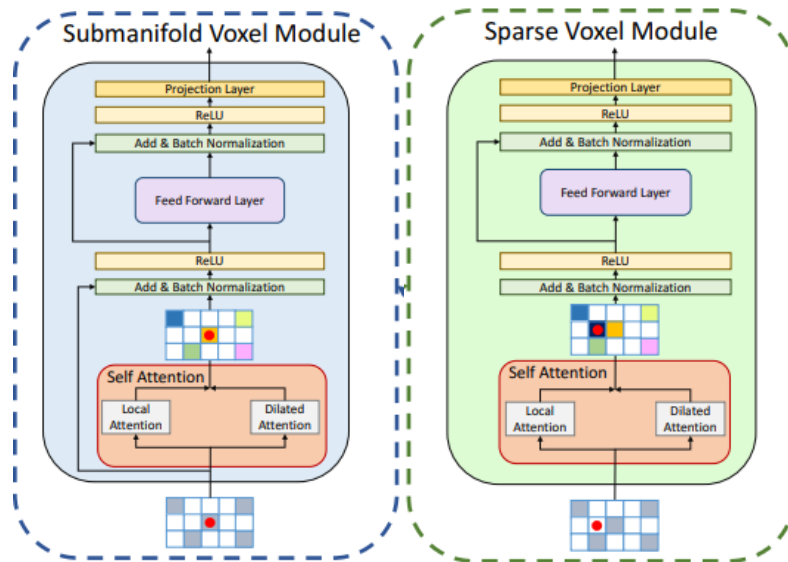
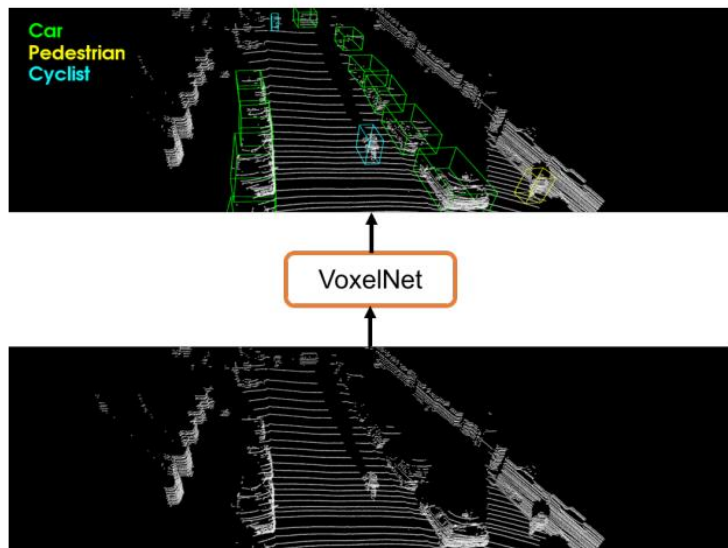
^Presenter



Outline

- **Background**
- **Fused Multi-head Tensor-compression for Attention**
- **Tensorized 3D Point Clouds Network**
- **Experiment Results**
- **Conclusion**

Background: 3D Point Clouds Detection

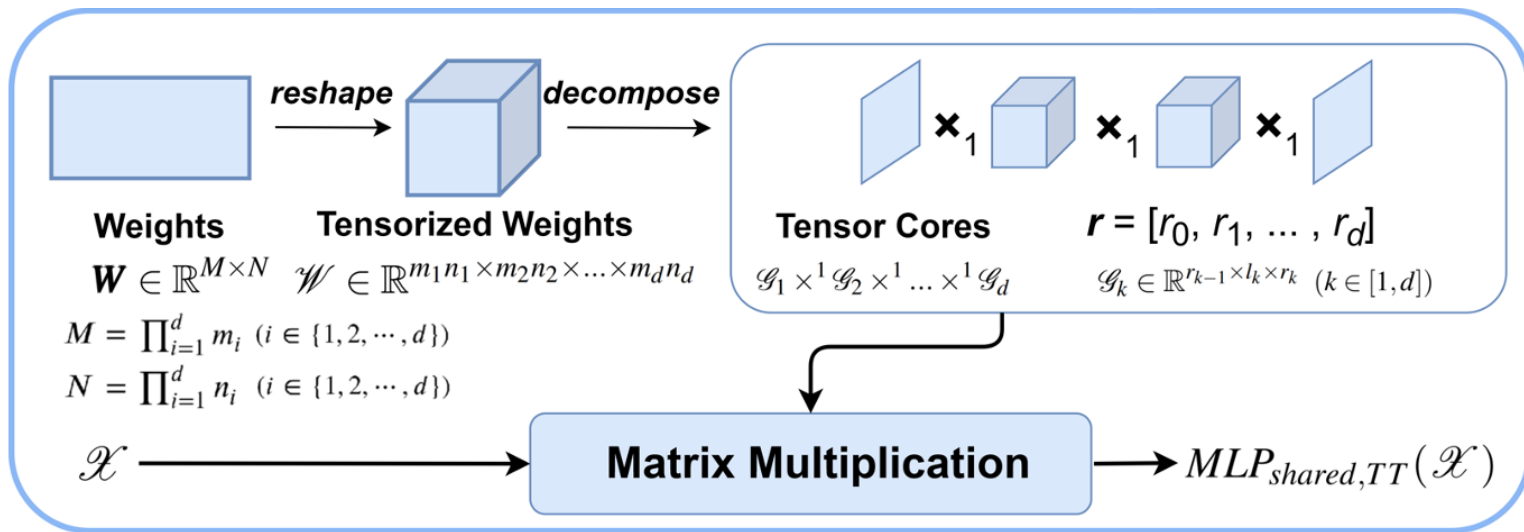


3D Point Clouds Detection Models [1][2]

[1] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in CVPR, 2018, pp. 4490–4499.

[2] J. Mao, Y. Xue, M. Niu, H. Bai, J. Feng, X. Liang, H. Xu, and C. Xu, "Voxel transformer for 3d object detection," in ICCV, 2021, pp. 3164–3173.

Background: Network Compression

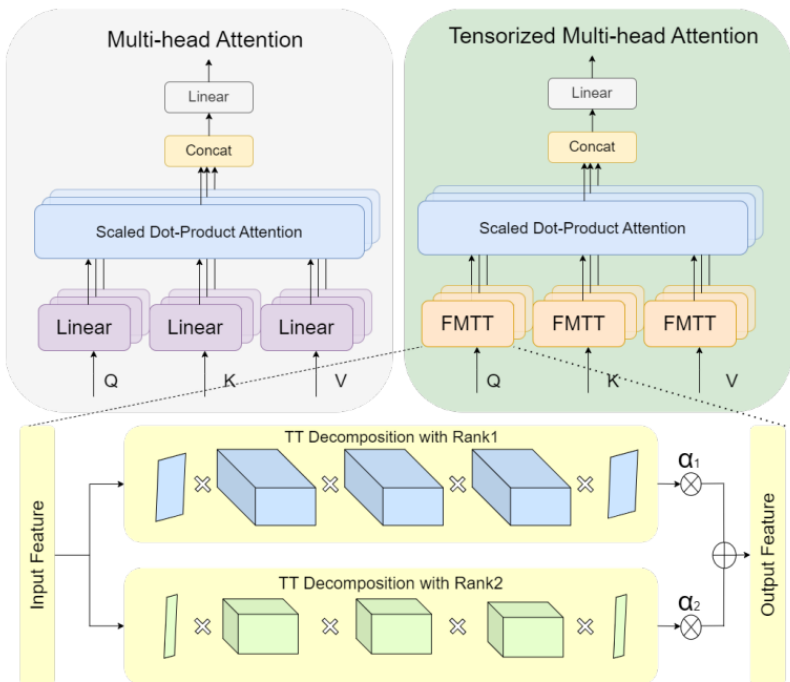


Tensor-train compression[3]

[3] C. Ding, H. Ren, Z. Guo, M. Bi, C. Man, T. Wang, S. Li, S. Luo, R. Zhang, and H. Yu, "Tt-lcd: Tensorized-transformer based loop closure detection for robotic visual slam on edge," in ICARM. IEEE, 2023, pp. 166–172.

Fused Multi-head Tensor-compression for Attention

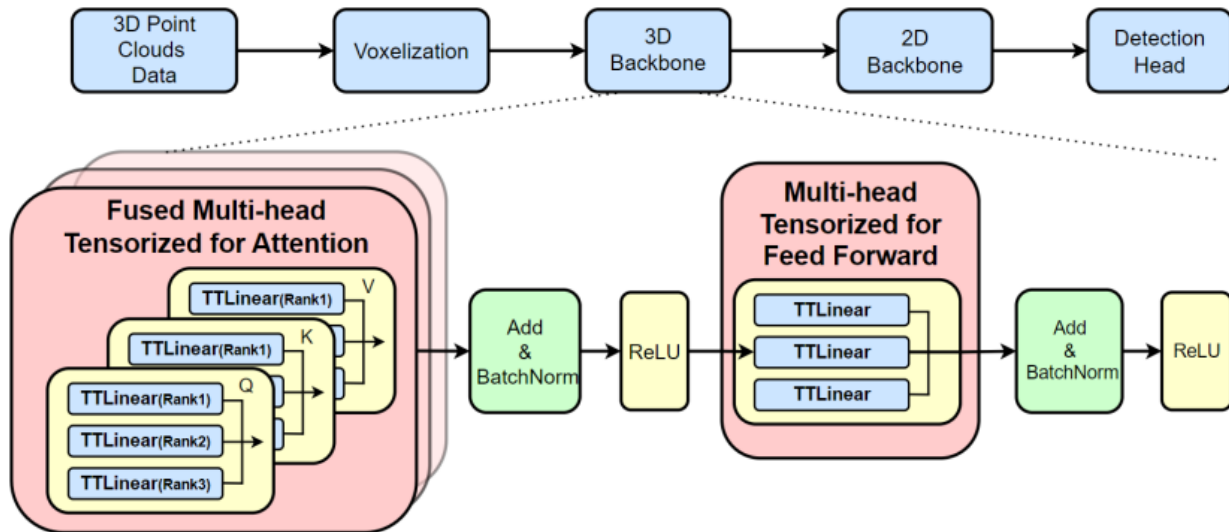
Comparison between Our Compressed Attention and Conventional Attention



Algorithm 1: Fused multi-head tensor-compression methods

- input :** the input feature x , the rank selection space
output: the output feature y , the selected ranks $r_1 \dots r_d$
- 1 Five epochs pre-train for different ranks in rank selection space to get the average loss, number of parameters and floating point operations.
 - 2 **for** $i \leftarrow 1$ to N **do**
 - 3 $C_{comp}^{r_i} = \theta \cdot \#PARAMS^{r_i} + \#FLOPS^{r_i}$
 - 4 $C_{rank}^{r_i} = C_{loss}^{r_i} + \gamma \cdot C_{comp}^{r_i}$
 - 5 $r_1 \dots r_d = \text{sorted}(C_{rank})[:d]$, sort the C_{rank} and select the minimum d items' ranks.
 - 6 **for** $i \leftarrow 1$ to d **do**
 - 7 $\beta_i = \text{Softmax}(\alpha) = \frac{\exp^{\alpha_i}}{\sum_1^d \exp^{\alpha_i}}$, normalize the coefficients of each heads.
 - 8 $y+ = TLinear_{r_i}(x) \cdot \beta_i$, apply the fused multi-head tensor-train compression and get the output feature
-

Tensorized 3D Point Clouds Network



- The Overall Model Architecture with Fused Multi-head Tensorized Blocks for Attention and Feed Forward

Experiment Results

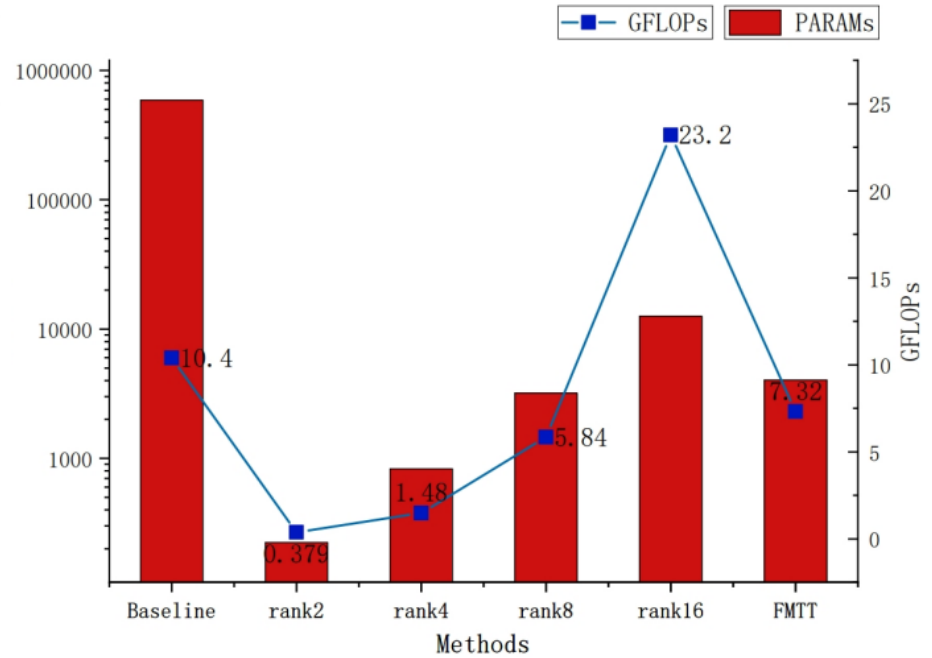
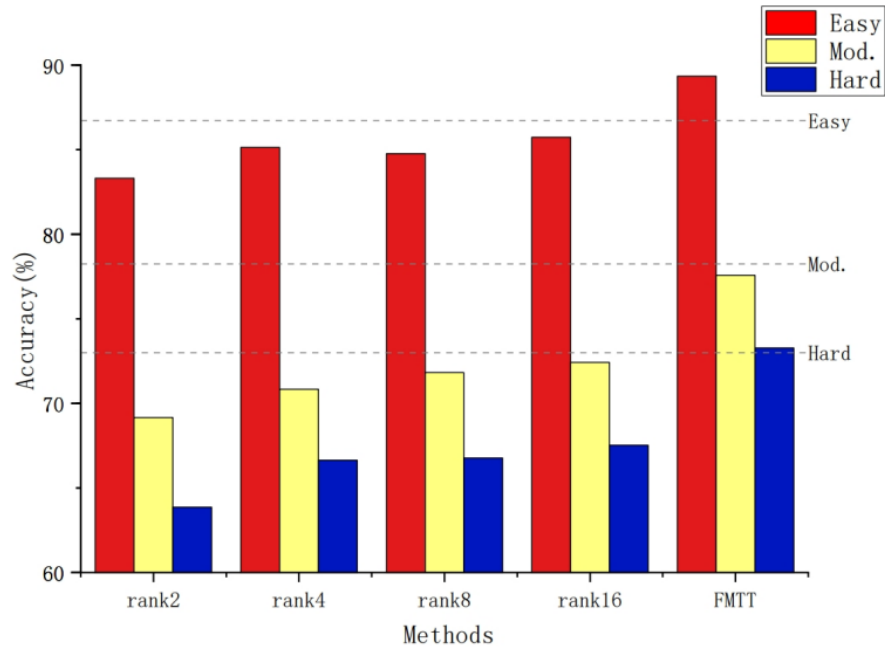
TABLE I: The average accuracy and model size comparison on the KITTI dataset car category

Method	Mode	Acc(%)			Size
		Easy	Mod.	Hard	
Part-A2 Net [16]	TSD	87.81	78.49	73.51	226MB
PV-RCNN [17]	TSD	90.25	81.43	76.82	50.1MB
PointRCNN [18]	TSD	86.96	75.64	70.70	14.9MB
VoxelNet [1]	SSD	77.47	65.11	57.73	78.26MB
Patches [19]	SSD	88.67	77.20	71.82	-
STD [20]	SSD	87.95	79.71	75.09	-
PointPillars [8]	SSD	82.58	74.31	68.99	18.4MB
HVNet [3]	SSD	87.21	77.58	71.79	77.29MB
3DSSD [21]	SSD	88.36	79.57	74.55	30.0MB
SA-SSD [22]	SSD	88.75	79.79	74.16	40.7MB
VoTr-SSD [4]	SSD	86.73	78.25	72.99	55.09MB
TT-VoTr-SSD(Ours)	SSD	89.35	77.58	73.27	9.12MB

TABLE II: Comparison between our fused multi-head tensorized method and other compressed method

Method	GFLOPs	Acc(%)			Size
		Easy	Mod.	Hard	
Baseline	10.4	86.73	78.25	72.99	55.09MB
Tensorized	5.84	84.77	71.82	66.77	8.96MB
QAT [5]	10.5	85.62	72.50	69.37	14.91MB
Sparsity [6]	5.26	85.46	70.81	67.44	28.31MB
FMTT	7.32	89.35	77.58	73.27	9.12MB

Experiment Results



Accuracy Comparison of Our Compressed Models with Different Ranks

Complexity Comparison of Our Compressed Operations with Different Ranks

Conclusion

- **An end-to-end 3D point clouds voxel transformer based model is fully compressed by the tensor-compression. In comparison with uncompressed model, it achieves $6.04\times$ times compression rate and 2.62% accuracy improvements.**
- **A novel fused multi-head tensor compression for both attention and convolution is proposed to compress the model.**
- **A tensor-train rank selection strategy is proposed with consideration of model size, computation load and accuracy during training.**

Thank you!
Q&A