EMPIRICAL RESEARCH ON QUANTIZATION FOR 3D MULTI-MODAL VIT MODELS

Zicong Hu[1,†], *Jian Cao*[1,⋆,†], *Weichen Xu*[1], *Ruilong Ren*[1], *Tianhao Fu*[1], *Xinxin Xu*[1], *Xing Zhang*[1,2,⋆]

[1]School of Software and Microelectronics, Peking University, China
[2]Shenzhen Graduate School, Peking University, China

2024 IEEE International Conference on Image Processing (ICIP) | 979-8-3503-4939-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICIP51287.2024.10647855

ABSTRACT

Model quantization finds success in simplifying model inference in practical applications. However, it predominantly focuses on CNNs and 2D ViT models, with limited attention given to quantizing 3D models. We pensively explore 3D model quantization challenges and discover similar numerical distributions of Softmax and LayerNorm between 3D and 2D models. Consequently, we apply the quantization algorithms FQ-ViT and I-ViT designed for 2D ViT models to 3D model quantization to address performance issues caused by uneven numerical distributions in Softmax and LayerNorm. Our research includes extensive experiments using transformer architectures and establishes benchmarks, demonstrating successful quantization of 3D multimodal model UNITR. Notably, our approach experiences a slight decrease compared to FP32 while outperforming other state-of-the-art models. For example, in the 3D object detection task on the nuScenes dataset, the 8-bit UNITR (FQ-ViT) achieves impressive NDS and mAP scores of 73.0% and 70.0%, surpassing the full precision BEVFusion model.

***Index Terms*—** ViT, Model Quantization, 3D Object Detection, BEV map segmentation

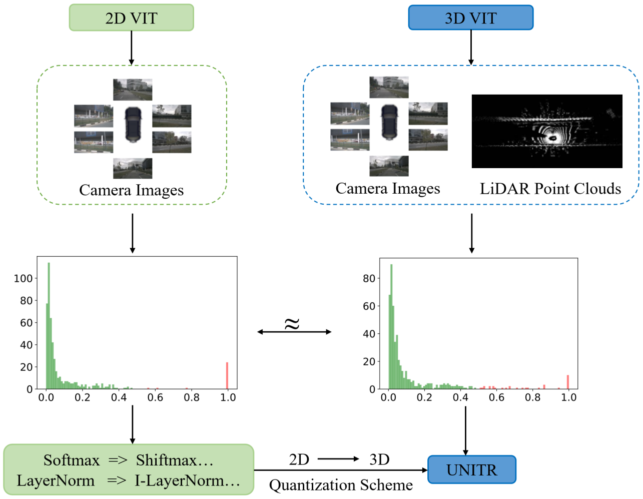

Fig. 1: illustrates the workflow of the entire paper. Initially, a comparison is made between the data constructions in 2D and 3D, followed by an investigation into their data distributions. Finally, the quantization strategy employs in 2D is applied to the 3D vision.

1. INTRODUCTION

Quantization enhances computational efficiency by reducing the bit widths of model weights and activations, proving beneficial for hardware implementation[1]. However, current quantization approaches, developed mainly for CNNs and 2D vision transformers, need to accommodate the intricacies of 3D vision, where data representation and processing differ markedly, especially in handling sparse point clouds[2]. The significant gap is primarily attributed to the distinct numerical distributions between 2D and 3D data, especially in applications that involve the fusion of images and point clouds in 3D vision[3]. Consequently, there is a pressing need for quantization methods specifically designed for 3D vision transformers to address the unique challenges of three-dimensional data environments.

Recent research has highlighted the detrimental impact of quantizing LayerNorm and Softmax in 2D vision transformers on model accuracy. This is primarily due to significant variability in LayerNorm input channels, some deviating up to 50 times the median. Additionally, attention map values in transformers show a skewed distribution, mainly concentrated between 0 and 0.01, with outliers[4]. These issues reduce the effectiveness of standard quantization methods, leading to a notable decline in model performance. Therefore, developing and exploring quantization techniques that address these challenges is crucial to maintaining model integrity.

As shown in Figure 1, a notable discovery in applying uniform quantization to the 3D multimodal algorithm UNITR[5] is the similarity in numerical distributions between the LayerNorm and Softmax layers of 3D and 2D visual data despite their differing data compositions. This similarity justifies exploring the transfer of 2D vision quantization techniques, known for managing channel variances in LayerNorm and skewed distributions in Softmax, to 3D vision. Consequently, we introduce two specialized methods, FQ-ViT and I-ViT[6], designed for the quantization of these layers in 3D

[⋆]means equal contribution, and [†] means corresponding author contexts. These approaches effectively address the unique quantization challenges in 3D vision, enhancing the UNITR model's performance after quantization.

Integrating FQ-ViT and I-ViT into UNITR model significantly enhance performance at 8-bit precision, achieving 73.0% in NDS and 70.0% in mAP. This result significantly exceeds that of full-precision BEVFusion, surpassing it by over 1.6% in NDS and 1.5% in mAP. Comparative and ablation studies confirm the superiority and reliability of FQ-ViT and I-ViT in 3D quantization, highlighting their crucial role in improving model efficiency in resource-limited settings.

The main contributions are summarized as follows:

1.We primarily analyze the numerical distribution of LayerNorm and Softmax in 3D vision. The findings indicate that the performance reduction in quantized models across 2D and 3D dimensions is primarily due to marked inter-channel variations in LayerNorm and substantial non-uniformity in softmax attention maps.

2.The Power-of-Two Factor (PTF) and Log Int Softmax (LIS) from the 2D quantization strategy FQ-ViT and Shiftmax and I-LayerNorm from I-ViT are introduced into 3D vision. FQ-ViT and I-ViT address the issue of uneven numerical distribution in Softmax and LayerNorm, enabling more accurate quantization.

3.We conduct extensive model quantization experiments on the UNITR using FQ-ViT and I-ViT, comparing them with other 3D vision models. The results show that FQ-ViT and I-ViT achieve competitive performance in quantizing the UNITR to 8 bits compared to other 3D vision models.

2. RELATED WORK

2.1. Multi-Sensor 3D Perception

In autonomous driving, the combined use of lidar and cameras is critically explored for improved reliability, supported by extensive research[7, 8]. 3D perception methods using these sensors are mainly divided into point-based[9], proposal-based[10], and Bird's Eye View (BEV)-based[11] approaches. Point and proposal-based techniques enhance lidar data with image features, whereas BEV-based methods integrate camera and lidar data in the BEV space, applying 2D convolution to achieve adequate 3D perception.

2.2. Network Quantization

Model quantization, converting floating-point to lower-bit parameters, is critical for efficiency on constrained hardware and is divided into Post-Training Quantization (PTQ) and Quantization-Aware Training (QAT). While QAT improves performance through full dataset retraining, it is resource-heavy. PTQ is faster and less resource-demanding but may reduce performance.Current quantization algorithms like DFQ[12], AdaRound[13], and BRECQ[14], optimized for CNNs, underperform with ViTs. The distinct architecture of ViTs necessitates bespoke quantization methods, underscoring the need for new algorithmic approaches in this domain.

To improve ViT efficiency, researchers have developed various quantization strategies. I-BERT[15] and FQ-ViT introduce innovative methods like Powers-of-Two Scale and Logarithmic Integer Softmax, targeting key components like LayerNorm and Softmax, thus advancing ViT quantization. PSAQ-ViT[16] presents a data-free approach leveraging patch similarity for enhanced quantization efficiency. In contrast, RepQ-ViT[17] separates quantization from inference, addressing distribution imbalances in LayerNorm and Softmax activations. These methods signify the ongoing advancements and refinement in ViT quantization.

However, existing quantization methods primarily address 2D ViT, leaving 3D ViT models less explored. Thus, this paper investigates quantization challenges specific to 3D perception models.

3. METHOD

3.1. 2D ViTs' Standard Structure

The vision transformers introduce a novel shift in image processing by replacing traditional CNN convolutional layers with self-attention mechanisms. ViT breaks down an image into patches processed by a transformer encoder, focusing on key features and diminishing minor ones, as shown in Figure 2. To enhance performance on larger images, a hybrid model merges convolutional layers for spatial reduction with self-attention layers to keep long-range patch dependencies. This approach aims to blend the best of both worlds, offering a robust solution for complex image-processing tasks.

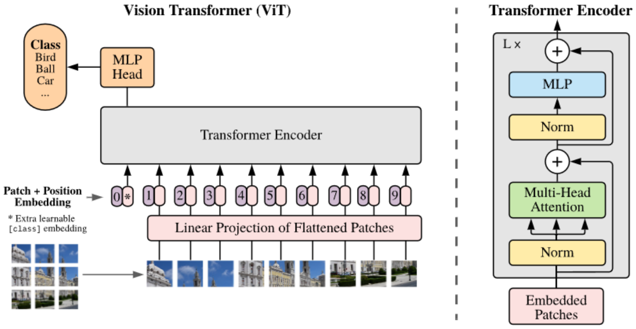

Fig. 2: The architecture of ViT, the left panel shows the image division and position embedding process and the right panel presents a standard encoder architecture that contains the multihead attention layer.

3.2. 3D Multi-Modal ViT Model

In autonomous driving, integrating data from diverse sensors is critical. Traditional approaches often employ separate encoders for each sensor, leading to complex fusion algorithms that hinder system speed and complicate training. UNITR addresses these issues by introducing two modal-agnostic transformer blocks that leverage the complementary features of 2D and 3D data. This method simplifies fusion, enhancing system efficiency and training manageability.

In image perspective analysis, we classify multi-modal tokens by their camera perspective positions into specific sets, leveraging DSVT[18] for 2D cross-modal interactions within these mixed-modality groups.

UNITR addresses the computational intensity of traditional 2D to 3D conversion methods by introducing a non-trainable, precomputable technique for efficiently mapping image patches to 3D, enhancing 3D segmentation efficiency and overcoming conventional method limitations.

3.3. Distributions of Attention Score and LayerNorm Activations in 2D and 3D ViT Model

Analyzing 2D ViT and UNITR for 3D object detection reveals apparent numerical differences. Reluctance to quantize Softmax to avoid accuracy[19] loss leads to CPU-GPU data transfers for de/re-quantization, maintaining hardware's dependence on floating-point operations. This increases resource usage and slows inference, complicating performance optimization.

In 2D ViT, the Softmax operation transforms the attention scores of the MSA module into probabilities, constraining the values within the (0, 1) interval:

$$\mathrm{Softmax}(\mathbf{x})_i := \frac{\exp x_i}{\sum_{j=1}^{k} \exp x_j}, \text{ where } \mathrm{x} = [x_1, \ldots, x_k] \quad (1)$$

Utilizing the exponential function in the Softmax activation produces a skewed attention map distribution, primarily focused on lower values, with more than 99% of activations falling below 0.4, as illustrated in Figure 3. The remaining 1% of higher activations are crucial, representing key patch correlations utilized by the MSA module. Previous work highlights that higher image resolutions and more minor patches benefit model performance. However, these adjustments significantly increase attention maps' computational load and storage, impacting inference efficiency. Therefore, it is crucial to efficiently preserve these essential components during quantization to maintain model performance.

LayerNorm is commonly employed in transformers and involves several nonlinear operations. This operation normalizes input activations along the channel dimension, computes statistical metrics μX and σX at each forward step, and normalizes the input X. Subsequently, affine parameters γ and β rescale the normalized input to another learned distribution. We describe the normalization process as follows:

$$\mathrm{LayerNorm}(X_{n,:}) = \frac{X_{n,:} - \mathrm{E}[X_{n,:}]}{\sqrt{\mathrm{Var}[X_{n,:}] + \epsilon}} \odot \gamma + \beta \quad (2)$$

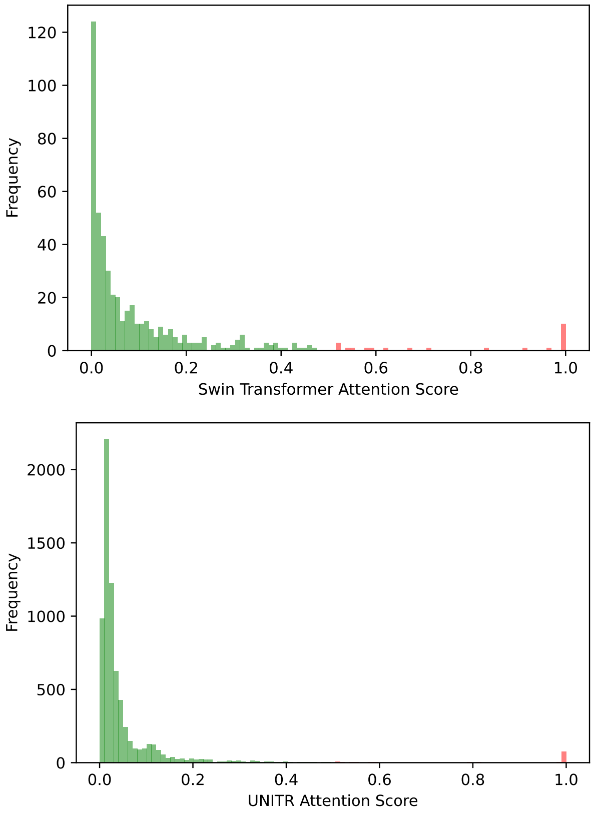

Fig. 3: The histogram of the Softmax activation after the first MSA module in Swin-Transformer and UNITR is presented. It is evident that the distribution is highly unbalanced, with the majority concentrated in small values (green) and a minority scattered in large values (red).

Figure 4 displays boxplots of post-LayerNorm activation distributions in Swin-Transformer and UNITR. Both models show notable inter-channel variations, with some channels having significant min-max differences. Traditional quantization struggles with these fluctuations, risking significant errors. Uniformly applying quantization scales across channels in such cases results in unacceptable inaccuracies. Alternatives like group quantization[20] or channel quantization[21], assigning unique parameters to different groups or channels, might be more effective.

3.4. Quantization Scheme For 3D Multi-Modal ViT Model

3.4.1. Softmax

In 2D ViT research, uniform quantization of attention maps caused significant performance declines in various models. FQ-ViT introduces log2 quantization with i-exp, ensuring consistency between full-precision and quantized maps, en-

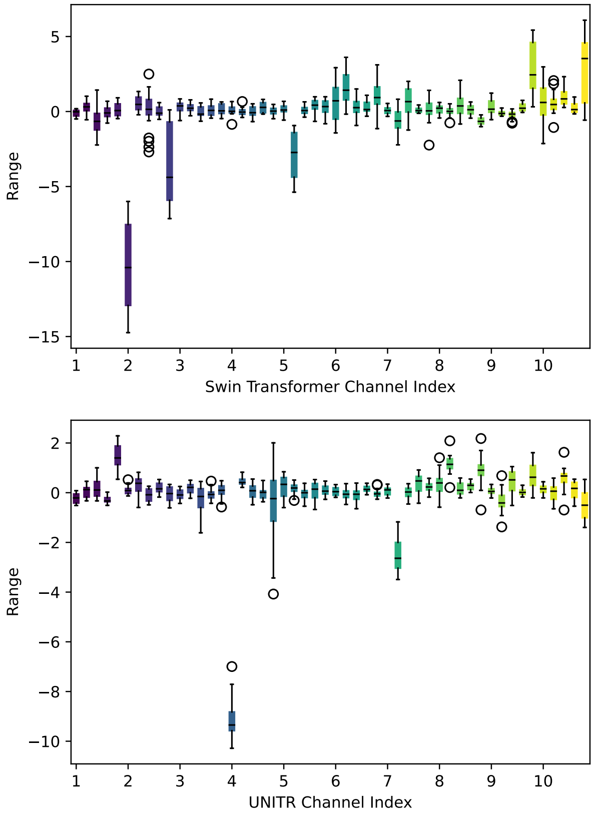

Fig. 4: Boxplots of channel activations after the first module of LayerNorm in Swin-Transformer and UNITR are displayed. Clearly, there are significant inter-channel variations.

hancing quantization efficiency.

$$\mathrm{LIS}(s \cdot \mathrm{X_Q}) = \mathrm{N} - \mathrm{clip}\left(\log_2 \left\lfloor \frac{\sum \mathrm{i-exp(X_Q)}}{\mathrm{i-exp(X_Q)}} \right\rceil, 0, 2^b - 1\right) \\ = \mathrm{N} - \mathrm{Attn_Q} \tag{3}$$

The equation $N = 2^b - 1$ relates to the quantized input X_Q, where s represents the quantized input and scale.

The authors achieve a fully fixed-point inference for Softmax based on this process. However, due to non-linearity, Softmax cannot follow a binary arithmetic pipeline, and the exponential arithmetic within it is typically not supported by pure integer logic units. To address this issue, I-ViT proposes the Shiftmax approximation method. Expressed by the equation 4, Shiftmax approximates the non-linear function, allowing for the precise and efficient integer computation of Softmax using straightforward hardware logic.

$$2^{S_\Delta \cdot (-r)} \approx [S_\Delta \cdot (-r)]/2 + 1 \\ = S_\Delta \cdot [((-r) \gg 1) + I_0] \tag{4}$$

In equation 4, $I_0 = \lfloor 1/S_\Delta \rceil$. The above completes the approximation, namely $S_\Delta \cdot I_{exp} \approx e^{S_\Delta \cdot I_\Delta}$, where $S\Delta$ can be simplified through fractional simplification.

3.4.2. *LayerNorm*

Figure 4 shows notable channel variations in the LayerNorm layer input. FQ-ViT introduces Power-of-Two Factor (PTF), a simple yet effective LayerNorm quantization method to tackle this. PTF assigns unique factors to each channel, bypassing the need for varied quantization parameters.

With $N = 2^b - 1$, X_Q and s representing the quantized input and scale, the authors accomplished fully fixed-point inference for Softmax based on the described procedure.

$$\mathrm{Y_Q} = \lfloor \mathrm{A\widehat{X}_Q + B} \rceil + zp_{out} \\ = \left\lfloor \frac{\mathrm{sign(A) \cdot N_2 \widehat{X}_Q} + \lfloor \mathrm{B} \cdot 2^{\mathrm{N_1}} \rceil}{2^{\mathrm{N_1}}} \right\rceil + zp_{out} \tag{5}$$

I-ViT improves the lightweight integer iteration method through a shift-based approach, aiming to achieve maximum convergence through iterative exploration. We modify the stopping criterion to the number of iterations for convenient hardware implementation.

$$I_{i+1} = (I_i + \lfloor \mathrm{Var}(I_x)/I_i \rfloor)/2 \\ = (I_i + \lfloor \mathrm{Var}(x)/I_i \rfloor) \gg 1 \tag{6}$$

In equation 6, I_i represents the result of the $i - th$ iteration, and I_0 is initialized. The entire iteration stops when $I_{i+1} \geq I_i$.

4. EXPERIMENTS

4.1. Implementation Details

The UNITR architecture employs the DVFE layer to tokenize image and LiDAR data for voxelizing point clouds. Detection tasks utilize a grid of 0.3m×0.3m×8.0m, while segmentation tasks utilize 0.4m×0.4m×8.0m. Patch tokenizer downscales images to a 32×88 resolution. UNITR blocks utilize weight-sharing to enhance modality representation learning, optimizing multimodal data integration.

Our research employs the comprehensive nuScenes dataset, known for its extensive annotations ideal for 3D object detection tasks[22]. This dataset comprises 40,157 instances, each with six monocular camera images for a 360-degree view and 32-beam LiDAR data. We emphasize metrics like the nuScenes detection score (NDS) and mean average precision (mAP) for 3D object detection. We randomly select a calibration subset of 20 training images and assess the model's performance using the validation set.

Our work adopts symmetric channel-wise quantization for weights and asymmetric per-layer quantization for activation functions, ensuring model optimization. The MinMax algorithm serves as the standard for weight quantization, facilitating consistent comparative analyses across different model architectures.We conduct quantitative experiments on Weights, Activations, Attention scores, and Layernorm activations. In the experimental tables, we abbreviate these components as W, A, Attn, and LN to streamline presentation and enhance clarity.

4.2. Quantization Results on NuScenes Dataset

As depicted in Table 1, we subject UniTR to 8-bit FQ-ViT quantization. While there is a slight performance decrease compared to FP32, it exhibits outstanding performance in LiDAR and camera fusion methods. On the validation dataset, employing FQ-ViT for 8-bit quantization of UNITR yields NDS and mAP scores of 73.0% and 70.0%, respectively, surpassing FP32 BEVFusion and achieving quantization with lower bit precision. Even when compressing UniTR to 4 bits, its accuracy remains 71.5%. The performance of the model quantization algorithm using I-ViT on UniTR mirrors these results. Furthermore, as shown in Table 2, we conduct quantization experiments on UNITR for the BEV map segmentation task. Despite a reduction in performance compared to FP32, our quantized results outperform other listed 3D models.

Methods	W/A/Attn/LN	NDS	mAP
FusionPainting[23]	FP32	0.707	0.665
TransFusion[24]	FP32	0.713	0.665
AutoAlignV2[25]	FP32	0.712	0.671
UVTR[26]	FP32	0.702	0.654
DeepInteraction[27]	FP32	0.726	0.699
BEVFusion[11]	FP32	0.714	0.685
	FP16 INT8	0.708	0.677
UNITR	FP32	0.731	0.701
UNITR(FQ-ViT)	8/8/8/8	0.730	0.700
	4/8/8/8	0.715	0.679
UNITR(I-ViT)	8/8/8/8	0.729	0.693
	4/8/8/8	0.708	0.655

Table 1: Performance of FQ-ViT and I-ViT methods for 3D object detection tasks on nuScenes (val) dataset.

4.3. Comparison with State-of-the-art Methods

We explore popular post-training quantization methods, including MinMax, EMA[32], Percentile[33], and OMSE[34]. We opt for FQ-ViT and I-ViT, two methods capable of fully quantizing the transformer structure in the 2D domain. Our numerical analysis predicts these methods would also be effective in the 3D domain. To validate the effectiveness of our proposed methods, we conduct experiments on different quantization strategies using the nuScenes dataset and report the overall NDS and mAP results in Table 3. Most current methods have yet to achieve complete quantization of the vision transformer, whereas our FQ-ViT and I-ViT methods have successfully fully quantized the transformer results. Furthermore, our experimental results indicate that the quantization methods used in the traditional 2D domain are equally effective in 3D object detection. For example, our FQ-ViT achieves 73.0% NDS and 70.0% mAP on UNITR with 8-bit quantization of all modules, while I-ViT achieves 72.9% NDS and 69.3% mAP.

Methods	W/A/Attn/LN	NDS	mAP
Full Precision	FP32	0.731	0.701
MinMax	8/8/8/8	0.676	0.627
EMA[32]	8/8/8/8	0.689	0.646
Percentile[33]	8/8/8/8	0.255	0.147
OMSE[34]	8/8/8/8	0.260	0.155
FQ-ViT	8/8/8/8	0.730	0.700
I-ViT	8/8/8/8	0.729	0.693

Table 3: Comparison of the 3D object detection results with state-of-the-art quantization methods on the nuScenes dataset.

4.4. Ablation Studies

Table 4 presents the effects of PTF and LIS in FQ-ViT, and Shiftmax and ShiftGELU in I-ViT—on performance. These are assessed on the nuScenes validation set across full-precision and quantized UNITR models, using 8-bit weights and activations via the MinMax method as the benchmark.

To delve deeper into the effectiveness of our methods, we conducted performance tests and quantization effect analyses on the PTF and LIS of FQ-ViT, as well as the Shiftmax and I-LayerNorm of I-ViT, as part of our ablation studies. The experimental results indicate that the FQ-ViT and I-ViT quantization algorithms proposed for the 2D domain maintained robust performance in 3D object detection and surpassed traditional quantization methods.

Methods	PTF	LIS	NDS	mAP
Full Precision	-	-	0.731	0.701
FQ-ViT	×	×	0.676	0.627
	×	✓	0.700	0.655
	✓	×	0.697	0.650
	✓	✓	0.730	0.700
Methods	Shiftmax	I-LayerNorm	NDS	mAP
I-ViT	×	×	0.676	0.627
	×	✓	0.702	0.658
	✓	×	0.695	0.648
	✓	✓	0.729	0.693

Table 4: Ablation studies of 3D object detection results for FQ-ViT and I-ViT on the nuScenes dataset.

Methods	W/A/Attn/LN	Drivable	Ped. Cross.	Walkway	Stop Line	Carpark	Divider	Mean IoU
PointPillars[28]	FP32	72.0	43.1	53.1	29.7	27.7	37.5	43.8
CenterPoint[29]	FP32	75.6	48.4	57.5	36.5	31.7	41.9	48.6
PointPainting[30]	FP32	75.9	48.5	57.1	36.9	34.5	41.9	49.1
MVP[31]	FP32	76.1	48.7	57.0	36.9	33.0	42.2	49.0
BEVFusion[11]	FP32	85.5	60.5	67.6	52.0	57.0	53.7	62.7
UNITR	FP32	90.4	73.1	78.0	67.2	67.7	63.4	73.4
UNITR(FQ-ViT)	8/8/8/8	89.7	71.8	76.3	65.3	66.6	61.8	71.9
UNITR(FQ-ViT)	4/8/8/8	84.9	61.0	64.8	53.9	54.1	46.7	60.9
UNITR(I-ViT)	8/8/8/8	90.0	72.0	76.8	65.7	66.8	62.5	72.3
UNITR(I-ViT)	4/8/8/8	90.0	60.9	53.8	54.6	66.6	46.8	60.9

Table 2: The quantized UniTR outperforms state-of-the-art multi-sensor fusion methods in BEV map segmentation on the nuScenes validation set, demonstrating the effectiveness of our quantization strategy for semantic 3D perception tasks.

5. CONCLUSIONS

In this paper, we initially investigate the numerical distribution of 2D and 3D vision on Layernorm and Softmax, where we observe an issue of uneven distribution in both. Consequently, we introduce the methods FQ-ViT and I-ViT, which demonstrate effective performance in 2D, into the 3D multimodal algorithm UNITR. Compared to traditional uniform quantization methods, which significantly degrade model performance when quantizing 3D vision, our introduced FQ-ViT and I-ViT maintain commendable results in 3D vision. The quantized 3D vision transformer achieves performance comparable to the full-precision model, and even at lower bit rates, such as 4-bit, the performance of mmodels remains robust. In summary, we provide a higher baseline for future work and consider implementing full integer quantization for 3D multimodal vision.

6. REFERENCES

[1] Gholami A, Kim S, Dong Z, et al. A survey of quantization methods for efficient neural network inference[M]//Low-Power Computer Vision. Chapman and Hall/CRC, 2022: 291-326.

[2] Liu Z, Wang Y, Han K, et al. Post-training quantization for vision transformer[J]. Advances in Neural Information Processing Systems, 2021, 34: 28092-28103.

[3] Wang D, Devin C, Cai Q Z, et al. Monocular plan view networks for autonomous driving[C]//2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2019: 2876-2883.

[4] Lin Y, Zhang T, Sun P, et al. Fq-ViT: Post-training quantization for fully quantized vision transformer[J]. arXiv preprint arXiv:2111.13824, 2021.

[5] Wang H, Tang H, Shi S, et al. UniTR: A Unified and Efficient Multi-Modal Transformer for Bird's-Eye-View Representation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 6792-6802.

[6] Li Z, Gu Q. I-ViT: integer-only quantization for efficient vision transformer inference[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 17065-17075.

[7] Shi S, Wang X, Li H. Pointrcnn: 3d object proposal generation and detection from point cloud[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 770-779.

[8] Yan Y, Mao Y, Li B. Second: Sparsely embedded convolutional detection[J]. Sensors, 2018, 18(10): 3337.

[9] Wang Y, Guizilini V C, Zhang T, et al. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries[C]//Conference on Robot Learning. PMLR, 2022: 180-191.

[10] Wu H, Wen C, Shi S, et al. Virtual Sparse Convolution for Multimodal 3D Object Detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 21653-21662.

[11] Liang T, Xie H, Yu K, et al. Bevfusion: A simple and robust lidar-camera fusion framework[J]. Advances in Neural Information Processing Systems, 2022, 35: 10421-10434.

[12] Nagel M, Baalen M, Blankevoort T, et al. Data-free quantization through weight equalization and bias correction[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 1325-1334.

[13] Nagel M, Amjad R A, Van Baalen M, et al. Up or down? adaptive rounding for post-training quantization[C]//International Conference on Machine Learning. PMLR, 2020: 7197-7206.

[14] Li Y, Gong R, Tan X, et al. Brecq: Pushing the limit of post-training quantization by block reconstruction[J]. arXiv preprint arXiv:2102.05426, 2021.

[15] Kim S, Gholami A, Yao Z, et al. I-bert: Integer-only bert quantization[C]//International conference on machine learning. PMLR, 2021: 5506-5518.

[16] Li Z, Ma L, Chen M, et al. Patch similarity aware data-free quantization for vision transformers[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 154-170.

[17] Li Z, Xiao J, Yang L, et al. Repq-ViT: Scale reparameterization for post-training quantization of vision transformers[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 17227-17236.

[18] Wang H, Shi C, Shi S, et al. Dsvt: Dynamic sparse voxel transformer with rotated sets[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 13520-13529.

[19] Liu Z, Wang Y, Han K, et al. Post-training quantization for vision transformer[J]. Advances in Neural Information Processing Systems, 2021, 34: 28092-28103.

[20] Shen S, Dong Z, Ye J, et al. Q-bert: Hessian based ultra low precision quantization of bert[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(05): 8815-8821.

[21] Li R, Wang Y, Liang F, et al. Fully quantized network for object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 2810-2819.

[22] Caesar H, Bankiti V, Lang A H, et al. nuscenes: A multimodal dataset for autonomous driving[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 11621-11631.

[23] Xu S, Zhou D, Fang J, et al. Fusionpainting: Multimodal fusion with adaptive attention for 3d object detection[C]//2021 IEEE International Intelligent Transportation Systems Conference (ITSC). IEEE, 2021: 3047-3054.

[24] Bai X, Hu Z, Zhu X, et al. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 1090-1099.

[25] Chen Z, Li Z, Zhang S, et al. Autoalignv2: Deformable feature aggregation for dynamic multi-modal 3d object detection[J]. arXiv preprint arXiv:2207.10316, 2022.

[26] Li Y, Chen Y, Qi X, et al. Unifying voxel-based representation with transformer for 3d object detection[J]. Advances in Neural Information Processing Systems, 2022, 35: 18442-18455.

[27] Yang Z, Chen J, Miao Z, et al. Deepinteraction: 3d object detection via modality interaction[J]. Advances in Neural Information Processing Systems, 2022, 35: 1992-2005.

[28] Lang A H, Vora S, Caesar H, et al. Pointpillars: Fast encoders for object detection from point clouds[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 12697-12705.

[29] Yin T, Zhou X, Krahenbuhl P. Center-based 3d object detection and tracking[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 11784-11793.

[30] Vora S, Lang A H, Helou B, et al. Pointpainting: Sequential fusion for 3d object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 4604-4612.

[31] Yin T, Zhou X, Krähenbühl P. Multimodal virtual point 3d detection[J]. Advances in Neural Information Processing Systems, 2021, 34: 16494-16507.

[32] Jacob B, Kligys S, Chen B, et al. Quantization and training of neural networks for efficient integer-arithmetic-only inference[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 2704-2713.

[33] Lian X, Liu Z, Song Z, et al. High-performance FPGA-based CNN accelerator with block-floating-point arithmetic[J]. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 2019, 27(8): 1874-1885.

[34] Choukroun Y, Kravchik E, Yang F, et al. Low-bit quantization of neural networks for efficient inference[C]//2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). IEEE, 2019: 3009-3018.