Cross-Modal Quantization of BLIP-2 Using Activation-Aware Weight Quantization

Hui-Yun Deng*
Yuan Ze University
Taiwan, 320
e-mail : s1113724@mail.yzu.edu.tw

Chia-Yun Chiang*
Yuan Ze University
Taiwan, 320
e-mail : s1113701@mail.yzu.edu.tw

Yu-Hui Huang*
Yuan Ze University
Taiwan, 320
e-mail : yhhuang@saturn.yzu.edu.tw

This work extends Activation-aware Weight Quantization (AWQ) to the multimodal BLIP-2 framework, targeting language and vision modules. We apply AWQ to OPT-2.7B and perform quantization of the Value projection within the fused QKV structure in EVA-ViT-G. Our method enables efficient inference on memory-limited edge devices. Evaluation on COCO VQA v2 shows AWQ reduces memory and preserves language accuracy, while full quantization boosts efficiency but degrades accuracy and latency, highlighting trade-offs in cross-modal quantization.

I. Introduction

The increasing demand for real-time vision-language understanding on edge devices calls for efficient deployment of large pre-trained models. Recent advances in vision-language architectures, such as BLIP-2 [2] (Bootstrapping Language-Image Pre-training), combine large language models (LLMs) like OPT-2.7B (Open Pre-trained Transformer 2.7B, a large-scale causal language model developed by Meta AI) [3] with powerful vision encoders, including EVA-ViT-G (a high-capacity Vision Transformer model optimized for large-scale visual recognition) [4] in a two-stage framework. BLIP-2 [2] employs a pretrained vision transformer and a language model bridged by a learned query interface, enabling efficient multimodal learning. However, their large parameter sizes and high memory footprints pose significant challenges for low-power inference. Post-training quantization techniques like Activation-aware Weight Quantization (AWQ) [1], a post-training method that groups and scales weights based on activation sensitivity, offer a promising solution for compressing LLMs. However, applying AWQ [1] to vision transformers (ViTs) [6] remains limited and non-trivial due to architectural differences.

This work applies modular AWQ [1] quantization to BLIP-2 [2], encompassing both OPT-2.7B [3] and EVA-ViT-G [4]. While OPT integrates with AWQ seamlessly, EVA-ViT-G's [4] fused QKV structure requires structural adaptations. We introduce dequantization and weight-decoding mechanisms to extend AWQ [1] for vision modules and analyze the trade-offs.

The key contributions of this work are: 1. A unified AWQ-based quantization [1] pipeline for both language and vision modules in BLIP-2 [2]. 2. Structural adaptations to support ViT-specific [6] quantization. 3. Evaluation of quantized models on COCO VQA v2 [5], analyzing memory, accuracy, and latency.

II. Method

This study applies Activation-aware Weight Quantization (AWQ) [1] to BLIP-2 [2] to enhance efficiency on mid-range GPUs. Separate workflows are designed for OPT-2.7B [3] and EVA-ViT-G [4], accounting for architectural differences. Experiments are conducted on an NVIDIA RTX 3060 GPU with 12GB memory.

A. Adaptive Weight Quantization (AWQ) Technique

AWQ is a finetuning-free post-training quantization method that groups weights by output sensitivity, compressing them to INT4 with group-wise scaling. It performs layer-wise grouping and calibration using representative input data to preserve model behavior.

B. Quantization Strategy for the OPT Model

The OPT-2.7B [3] model is fully compatible with AWQ [1]. We quantize it to INT4 with group size 128 using 256 calibration prompts from COCO VQA v2 [5]. No structural changes are needed, and the quantized model maintains language generation capability within BLIP-2 [2].

C. Quantization Strategy for the EVA-ViT-G

EVA-ViT-G [4] features a fused QKV attention layer incompatible with AWQ [1]. We isolate the Value (V) projection and quantize it to INT4 while keeping Query (Q) and Key (K) in full precision. Other linear layers are quantized with group size 64. We implement a custom deployment routine to ensure AWQ [1] compatibility with the weight loading behavior of EVA-ViT-G [4].

III. Experimental Design and Evaluation

We evaluate our method on 1000 samples from the COCO VQA v2 [5] validation set, encompassing question types such as yes/no, number, and open-ended. Evaluation metrics include accuracy, inference latency, and GPU memory consumption. Notably, the quantized models retain strong performance on yes/no questions, achieving 70.74% with the full-precision model, 66.52% with OPT-2.7B quantized, and 60.72% with both language and vision modules quantized, demonstrating robustness in binary reasoning. The full-precision baseline achieves 51.46% overall accuracy, with 1187.09 seconds of inference time and a memory footprint of 8332 MiB. Quantizing only the language model yields a 39.3% memory reduction with a moderate accuracy drop to 49.43%. Full quantization of both modules further reduces memory usage to 4214 MiB, but results in increased latency (3167.58 seconds) and a larger decline in accuracy (26.59%).

IV. Quantization Strategy and Architectural Adaptations

AWQ [1] quantization of OPT-2.7B [3] yields significant memory savings with minimal accuracy loss. However, direct INT4

[*] All of the authors contributed equally.

TABLE I
VQA Accuracy, Inference Time, and Memory Usage for BLIP-2 Under Different Quantization Configurations

Models	**VQA acc.**				**Time**[a]	**GPU Mem.**
BLIP-2	Overall(%)	yes/no(%)	number(%)	other(%)		
ViT-g + OPT$_{2.7B}$	51.46	70.15	27.95	43.59	1187.09 s	8332 MiB
ViT-g + AWQ OPT$_{2.7B}$	49.43	66.52	28.75	42.00	2180.59 s	5054 MiB
AWQ ViT-g (V-Only) + AWQ OPT$_{2.7B}$	26.59	60.72	10.41	5.20	3160.13 s	4210 MiB
AWQ ViT-g (Fused-QKV) + AWQ OPT$_{2.7B}$	24.19	57.57	6.27	3.83	3327.14 s	3790 MiB

[a]Total time required for the model to perform inference on 1000 images from the VQAv2 Validation Set using a single GPU.

quantization of EVA's [4] fused QKV matrix impairs performance due to disrupted attention computations. In particular, direct quantization of the fused QKV projection layer in EVA-ViT-G [4] leads to severe performance degradation, with the "other" category dropping to only 3.83%, indicating that this layer requires separate handling during quantization. To mitigate this issue, we adopt a selective quantization strategy whereby only the Value (V) projection is quantized to INT4, while the Query (Q) and Key (K) projections remain in full precision. This approach preserves the integrity of the attention mechanism and ensures reliable inference, enabling efficient cross-modal quantization.

To elucidate the trade-offs involved, we analyze the accuracy degradation in relation to memory savings across different quantization configurations. Quantizing only the language module (OPT-2.7B [3]) results in a 39.3% reduction in GPU memory consumption with a modest 3.6 percentage point decrease in yes/no accuracy. This outcome demonstrates the effectiveness of AWQ [1] when applied to transformer-based language models. Conversely, extending AWQ [1] to both the language and vision modules achieves a greater memory reduction (−49.4%) but incurs a substantial overall accuracy decline, especially in number and open-ended question categories. Notably, the yes/no accuracy remains relatively stable, suggesting that simpler binary reasoning tasks are less sensitive to degradation in the vision encoder.

Additionally, we observe an increase in inference latency from 1187.09 seconds to 3327.14 seconds upon full quantization, using the same GPU. This latency overhead primarily stems from mixed-precision kernel dispatch and suboptimal memory alignment within the customized EVA-ViT-G [4] quantization kernel. These findings indicate that while full-model quantization is feasible for memory-constrained environments, further optimization is required to meet latency demands in real-time applications.

Finally, we evaluate model robustness by examining performance variance across question types. The results reveal that numerical questions are particularly susceptible to accuracy degradation due to reduced precision in the vision encoder, highlighting the potential necessity for selectively preserving precision in value-sensitive attention layers.

V. Conclusion

This work demonstrates the feasibility of applying Activation-aware Weight Quantization to both the language and vision modules in cross-modal architectures like BLIP-2 [2]. While AWQ [1] is directly applicable to LLMs, extending it to fused QKV structures in vision transformers requires structural adaptation. Our proposed selective quantization strategy preserves inference functionality and significantly reduces memory usage, making it suitable for deployment on memory-constrained hardware. Future work includes further optimizing accuracy through quantization-aware training and evaluating the design on actual AI accelerator platforms.

References

[1] J. Lin et al., "AWQ: activation-aware weight quantization for on-device LLM compression and acceleration," in *Proc. Mach. Learn. Syst.*, vol. 6, pp. 87–100, 2024.

[2] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. 40th Int. Conf. Mach. Learn. (ICML)*, vol. 202, pp. 19730–19742, 2023.

[3] S. Zhang et al., "OPT: open pre-trained transformer language models," unpublished, 2022.

[4] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, "Scaling vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 12104–12113, 2022.

[5] T.-Y. Lin et al., "Microsoft COCO: common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 8693, pp. 740–755, 2014.

[6] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.