Deploying Vision Language Models (VLMs) on edge devices is crucial for real-time applications, yet their size and complexity pose significant challenges. To address these limitations in resource-constrained environments, efficient model compression techniques are necessary. Qin et al. (2024) investigates model compression strategies for edge Large Language Models (LLMs), finding that quantization, pruning, and distillation show distinct performance characteristics for edge devices. Qin et al. (2024) reveals knowledge distillation yields the most stable performance, while quantization offers the potential for peak performance despite a tradeoff in stability. Pruning is comparatively less suitable for edge deployment than both quantization and distillation.

Zhu et al. (2024) provide a comprehensive survey on model compression techniques for large language models (LLMs) including quantization and distillation methods. Quantization reduces the numerical precision of model parameters and can be applied either post training (PTQ) or during training (QAT). The PTQ approach is computationally inexpensive and convenient because it does not require retraining Zhu et al. (2024). Frantar et al., (2023) introduced GPTQ, an efficient PTQ technique that uses second-order information to quantize weights of language models. GPTQ employs a loss function that approximates the Hessian matrix of the model weights to calculate the sensitivity of different weight parameters. PD-Quant (Liu et al., 2023) leverages a prediction difference loss to optimize activation scaling factors by aligning the output predictions of the quantized model with those of its full-precision counterpart.

MobileQuant (Wu et al., 2023) employs weight equivalent transformations and joint optimization of per-tensor and per-channel quantization ranges, specifically targeting latency and energy reduction for mobile deployments. Shen et al. (2024) introduced EdgeQAT to combat LLM quantization degradation on edge devices. The framework utilizes Entropy and Distribution Guided Quantization for attention mechanisms and Token Importance-Aware Adaptive Quantization for activations. EdgeQAT achieves accuracy comparable to FP16 models while demonstrating up to 2.37x inference speedup on edge platforms.

QAT approach combines quantization process with model training to retain accuracy but with increased computational cost. It simulates quantization effects during training by employing techniques such as the Straight-Through Estimator (STE) to approximate gradients through non-differentiable quantization operations (Bengio et al., 2013). The training objective in QAT remains fundamentally aligned with the primary task loss function which can be adapted to account for for the representational constraints introduced by quantization. EfQAT Ashkboos et al. 2024) optimizes quantization by fine-tuning a subset of critical network and quantization parameters while freezing others, reducing the computational cost of the backward pass. It employs a standard cross-entropy loss while adjusting quantization scales and weight updates

selectively. QLoRA offers a memory-efficient approach to fine-tuning large language models by quantizing the pre-trained weights to 4-bit precision (Dettmers et al., 2023). Parameter-efficient adaptation is achieved through the introduction of Low-Rank Adapters, which are trained while the base model remains frozen, enabling customization with minimal trainable parameters (Dettmers et al., 2023).

Distillation transfers knowledge from a larger, teacher model to a smaller, student model, enabling the latter to effectively approximate the performance of the former. (Zhu et al. 2024) The effectiveness of distillation depends on a careful design of the knowledge transfer process to balance model size and performance. The Kullback–Leibler (KL) divergence loss is used to minimize the difference between the softened probability distributions of the teacher and the student, as initially proposed by Hinton et al. (2015). Supervised fine-tuning (SFT) employs the cross-entropy loss to directly align the student's output with ground-truth labels while divergence minimization techniques, such as reverse KL divergence and Jensen–Shannon (JS) divergence, are applied to further reconcile the probability distributions of both models. (Xu et al. 2024) A few similarity-based methods, including L2-norm loss and cosine similarity loss, are utilized to enforce the alignment of intermediate feature representations between the teacher and the student. (Xu et al. 2024)

We aim to propose a novel approach that combines quantization and distillation at the loss function level to achieve efficient model compression. By integrating quantization-aware losses with knowledge distillation losses, our method targets robust low-precision performance while preserving key model characteristics. Additionally, incorporating task-specific loss components allows the approach to adapt effectively to the unique requirements of different applications. This hybrid strategy is designed to deliver high accuracy and real-time performance on resource-constrained edge devices.

Github Repos:

EdgeQat: https://github.com/shawnricecake/edge-qat

GPTQ: https://github.com/IST-DASLab/gptq

Knowledge Distillation: https://github.com/haitongli/knowledge-distillation-pytorch

QLora: https://github.com/artidoro/qlora

Datasets:

Visual Question Answering (VQA): VQA v2, Visual Genome, e-VQA (edge)

Scene: Cityscapes, KITTI, BDD100K, ScanNet, ADE20K

Object Detection: COCO Detection, Pascal VOC, Cityscapes (driving)s

References:

Qin, R., Liu, D., Xu, C., Yan, Z., Tan, Z., Jia, Z., Nassereldine, A., Li, J., Jiang, M., Abbasi, A., Xiong, J., & Shi, Y. (2024). Empirical Guidelines for Deploying LLMs onto Resource-constrained Edge Devices. *ArXiv*. https://arxiv.org/abs/2406.03777

Xunyu Zhu, Jian Li, Yong Liu, Can Ma, Weiping Wang; A Survey on Model Compression for Large Language Models. *Transactions of the Association for Computational Linguistics* 2024; 12 1556–1577. doi: https://doi.org/10.1162/tacl_a_00704

Frantar, E., Ashkboos, S., Hoefler, T., & Alistarh, D. (2022). GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers. *ArXiv*. https://arxiv.org/abs/2210.17323

Ashkboos, S., Verhoef, B., Hoefler, T., Eleftheriou, E., & Dazzi, M. (2024). EfQAT: An Efficient Framework for Quantization-Aware Training. *ArXiv*. https://arxiv.org/abs/2411.11038

Shen, X., Kong, Z., Yang, C., Han, Z., Lu, L., Dong, P., Lyu, C., Li, C., Guo, X., Shu, Z., Niu, W., Leeser, M., Zhao, P., & Wang, Y. (2024). EdgeQAT: Entropy and Distribution Guided Quantization-Aware Training for the Acceleration of Lightweight LLMs on the Edge. *ArXiv*. https://arxiv.org/abs/2402.10787

Xu, X., Li, M., Tao, C., Shen, T., Cheng, R., Li, J., Xu, C., Tao, D., & Zhou, T. (2024). A Survey on Knowledge Distillation of Large Language Models. *ArXiv*. https://arxiv.org/abs/2402.13116

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient Finetuning of Quantized LLMs. *ArXiv*. https://arxiv.org/abs/2305.14314

Tan, F., Lee, R., Dudziak, Ł., Hu, S. X., Bhattacharya, S., Hospedales, T., Tzimiropoulos, G., & Martinez, B. (2024). MobileQuant: Mobile-friendly Quantization for On-device Language Models. *ArXiv*. https://arxiv.org/abs/2408.13933

Liu, J., Niu, L., Yuan, Z., Yang, D., Wang, X., & Liu, W. (2022). PD-Quant: Post-Training Quantization based on Prediction Difference Metric. *ArXiv*. https://arxiv.org/abs/2212.07048