Qeios

Peer Review

Review of: "Efficiency Meets Fidelity: A Novel Quantization Framework for Stable Diffusion"

Shidi Tang[1]

1. School of Integrated Circuits, Southeast University, China

Summary

This paper introduces a novel quantization framework for Stable Diffusion (SD) models, aiming to balance efficiency and fidelity. The proposed Serial-to-Parallel (S2P) pipeline combines the strengths of serial (data-free) and parallel (training-like) approaches, addressing inconsistencies in prior post-training quantization (PTQ) methods. Key innovations include multi-timestep activation quantization, time-information precalculation, and a Hessian-based mixed-precision strategy. Experiments across SD v1–4, v2–1, and XL demonstrate significant improvements in distributional (FID-to-FP) and visual (SSIM, LPIPS) metrics over state-of-the-art methods.

Strength

1. Paper Writing: This manuscript is well written and organized.
2. Problem Relevance: The focus on output consistency between quantized and floating-point models addresses a critical gap in deploying diffusion models for real-world applications (e.g., edge devices, content creation).
3. Methodological Innovation:
    a. The S2P pipeline effectively balances training stability and data-free calibration.
    b. Multi-timestep activation quantization accounts for temporal variations in activation distributions.
    c. Time-information precalculation reduces memory overhead without sacrificing fidelity.

Major issues

1. Motivation

 doi.org/10.32388/KCJYXZ

This manuscript discusses the necessity of “the consistency of results produced by quantized and floating-point models” in the third paragraph of Section 1. I wonder what this “consistency” refers to? Does it refer to the metrics (like FID) gap between the quantized model and the fp model? If so, what is the difference between “high-quality image generation” and "consistency” as stated in the 1st sentence of Section 1? If I am correct, the methods proposed utilize “consistency” to improve the image generation quality. If that is so, the 1st sentence of Section 1 should be elaborated for readers to better understand the motivation of this manuscript. The authors should not treat “consistency” and “high-quality image generation” as two opposite concepts.

If the “consistency” differs from “high-quality image generation”, the authors should explain more in detail to better understand the motivation.

1. Method

The authors proposed the Mixed-Precision Quantization Strategy to quantize different layers with different precisions according to the sensitivity defined with the deviation of the loss function. I wonder what the differences are between the typical strategy and the proposed one in terms of the modeling of sensitivity (i.e., Equation 9). This could potentially help readers to better understand the novelty of this manuscript.

1. Experiments

The proposed method belongs to the QAT method, while several PTQ baselines (i.e., Q-diffusion, PTQ4DM, PCR) are selected for comparison. This comparison is unfair because typically the QAT method usually performs better than the PTQ method. Although the computational cost of the proposed method is less than that of the PTQ baselines, the effectiveness of their method should be better elaborated by directly comparing it with other QAT baselines (such as Q-DM [NeurIPS’23] and MPQ-DM [AAAI’25])

Minor issues

1. Figure 1 lacks the labels (e.g., (a), (b), (c)) for each sub-plot.
2. Figure 4 lacks the labels for the x-axis and y-axis.
3. What do numbers like “8.35”, “8.34”, “4.02” etc. mean in Table 4? More descriptions should be made in the manuscript.

4. Why are all the metrics in Table 7 for the FP32 method zeros? These metrics should be some positive values.
5. Why are all insensitive layers for the SD XL model related to q and k computation? Any insight into this?

Declarations

Potential competing interests: No potential competing interests to declare.