EfficientDM: Efficient Quantization-Aware Fine-Tuning of Low-Bit Diffusion Models

Yefei He[1], Jing Liu[2], Weijia Wu[1], Hong Zhou[1], Bohan Zhuang[2]

[1]Zhejiang University

[2]ZIP Lab, Monash University

The Twelfth International Conference on Learning Representations

Introduction

Background

- **The application of diffusion models is constrained by the massive parameters and computational complexity.** Even with advanced samplers, diffusion models are not yet ready for real-time applications. For example, even when executed on a high-performance platform such as the RTX 3090, Stable Diffusion with the DPM-Solver sampler still takes over a second to generate a 512×512 image.
- **Model quantization employs lower numerical bitwidth to represent weights and activations, alleviating both memory and computational burdens.** For instance, an 8-bit model's inference speed can be 2.03× faster than that of a full-precision (FP) model, and the acceleration ratio reaches 3.34× for a 4-bit model.

Introduction

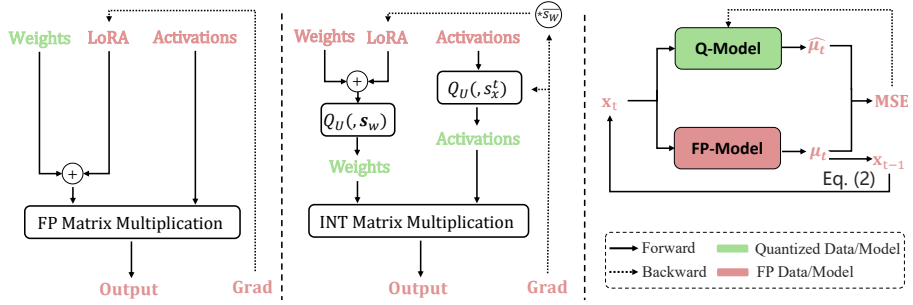
However, the challenges associated with low-bit quantization for diffusion models have not received adequate attention.

Challenges

- Post-training quantization (PTQ) is time- and data-efficient but can introduce substantial quantization errors at low bit-width.
- Quantization-aware training (QAT) can recover performance losses at lower bit-width by fine-tuning but requires original training datasets and significantly more computing resources, as evidenced by a $2.6\times$ increase in GPU memory consumption (31.4GB vs. 11.7GB) and a $18.9\times$ longer execution time (54.5 GPU hours vs. 2.88 GPU hours) when fine-tuning LDM-4 on ImageNet 256×256.

Method

In this paper, we introduce a data-free and parameter-efficient fine-tuning framework for low-bit diffusion models, denoted as EfficientDM, which demonstrates the capability to achieve **QAT-level performance** while upholding **PTQ-level efficiency** in terms of data and time.



(a) QLoRA layer (b) QALoRA layer (c) EfficientDM fine-tuning framework

Figure: An overview of the proposed EfficientDM fine-tuning framework.

Quantization-aware low-rank adapter

QLoRA fixes the original quantized weights $\hat{\mathbf{W}}_0$ and introduces updates as follows:

$$\mathbf{Y} = \mathbf{X}\hat{\mathbf{W}}_0 + \mathbf{XBA}, \tag{1}$$

However, it incurs limitations when activations are also quantized, denoted by $\hat{\mathbf{X}}$. In this case, the inner product between $\hat{\mathbf{W}}_0$ and $\hat{\mathbf{X}}$ can be efficiently implemented with bit-wise operations, whereas the operations involving $\mathbf{BA}$ and $\hat{\mathbf{X}}$ are computationally expensive during inference as $\mathbf{BA}$ is full-precision and has the same size as $\mathbf{W}_0$.
To address this, we propose Quantization-aware Low-rank Adapter (QALoRA), where the LoRA weights are first merged with FP model weights and then jointly quantized to the target bit-width:

$$\mathbf{Y} = \mathcal{Q}_U(\mathbf{X}, s_x)\mathcal{Q}_U(\mathbf{W}_0 + \mathbf{BA}, \mathbf{s}_w) = \hat{\mathbf{X}}\hat{\mathbf{W}}, \tag{2}$$

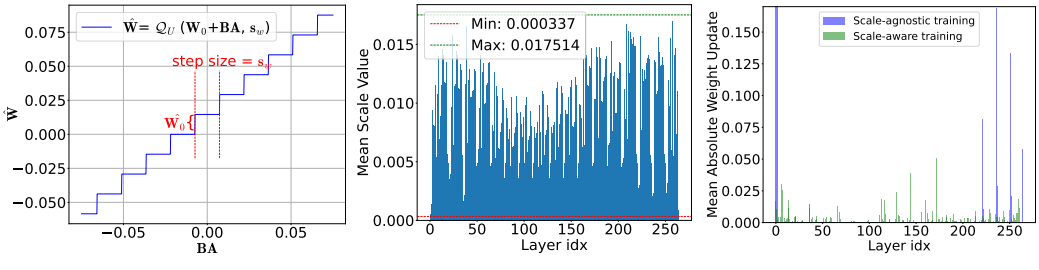
Data-free fine-tuning for diffusion models

To alleviate the dependency on the original dataset, we propose a data-free fine-tuning approach that distills the denoising capabilities of a full-precision model into its quantized counterpart. We input the same noise $\mathbf{x}_t$ to both FP and quantized denoising models at denoising step t and minimize the mean squared error (MSE) between their denoising results:

$$\mathcal{L}_t = \left\| \boldsymbol{\mu}_\theta(\mathbf{x}_t, t) - \hat{\boldsymbol{\mu}}_\theta(\mathbf{x}_t, t) \right\|^2, \tag{3}$$

The input data $\mathbf{x}_t$ is obtained by denoising random Gaussian noise $\mathbf{x}_T \sim \mathcal{N}(0, 1)$ with FP model iteratively for $T - t$ steps, as illustrated in the overview figure.

Scale-aware LoRA optimization



(a) (b) (c)

Figure: The motivation and effect of scale-aware LoRA optimization.

To facilitate the optimization of LoRA weights, we consider the ratio of $R = \frac{\nabla_{\mathbf{BA}}\mathcal{L}}{\mathbf{s}_w}$ should be roughly consistent in each layer.

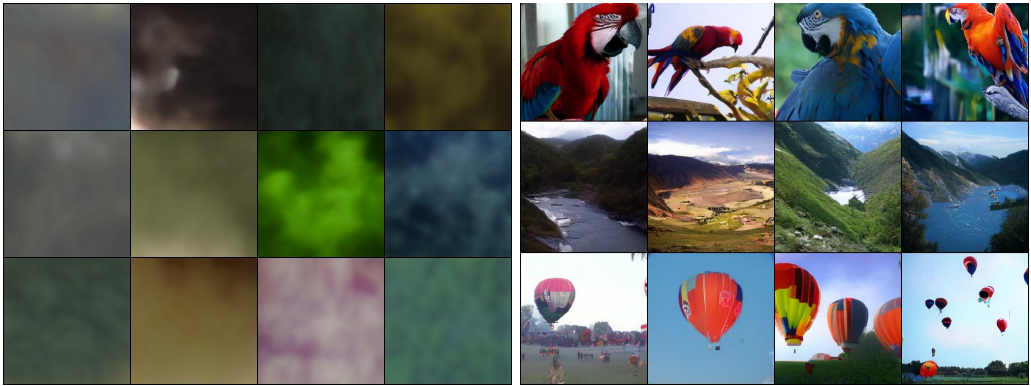
Temporal Activation LSQ (TALSQ)

Inspired by LSQ, we allocate temporal-aware quantization scales for activations and optimize them individually for each step, which we refer to as Temporal Activation LSQ (TALSQ):

$$S_x = \left\{ s_x^0, s_x^1, \ldots, s_x^{T-1} \right\}, \tag{4}$$

where T is the number of denoising steps for the fine-tuning. Recent advancements in efficient samplers have significantly reduced the number of sampling steps. Therefore, TALSQ introduces only a few trainable parameters for a single layer, which is negligible even compared to LoRA weights (which generally have thousands of parameters per layer).

Visualization Results on ImageNet



(a) Q-Diffusion (b) Ours

Figure: Samples generated by W4A4 LDM model on ImageNet 256×256.

Visualization Results on LSUN

(a) Q-Diffusion (b) Ours

Figure: Randomly generated samples by W4A4 LDM model on LSUN-Bedrooms.