DEPLOYING GENERATIVE AI ON MOBILE: COMPRESSING STABLE DIFFUSION WITH VRAM OPTIMIZATION AND DOMAIN-SPECIFIC PROMPTING

Do Trung Hieu

University of Information Technology
HCMC, Vietnam

What ?

We propose a Singular Value Decomposition Quantization technique to reduce latency and saved more memory:

- Designed a 4-bit quantization technique and Nukachu inference engine quantization speed up.
- Use low-rank branch to cut off redundant memory access, off LoRAs without re-quantization.
- Built a demo app with Diffusion model as SDXL and 12B FLUX.1 - inference time/ VRAM reduction

Why ?

- Generate high-quality image by Diffusion which is useful. However it scaled, increasing memory and higher latency.
- Significant rising the cost overhead due to extra data movement of activations.
- Facing a challenge deployment when Diffusion models applied for mobile or edge computing device.

Overview

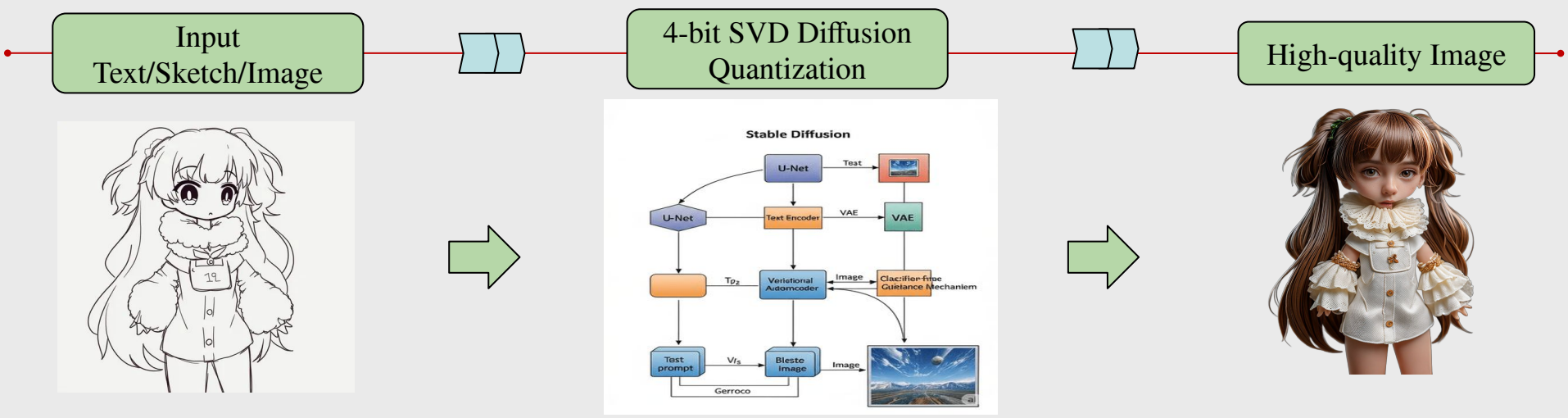


Figure 1. Diagram of the Stable Diffusion model after quantization to generate images

Description

1. CONTENT

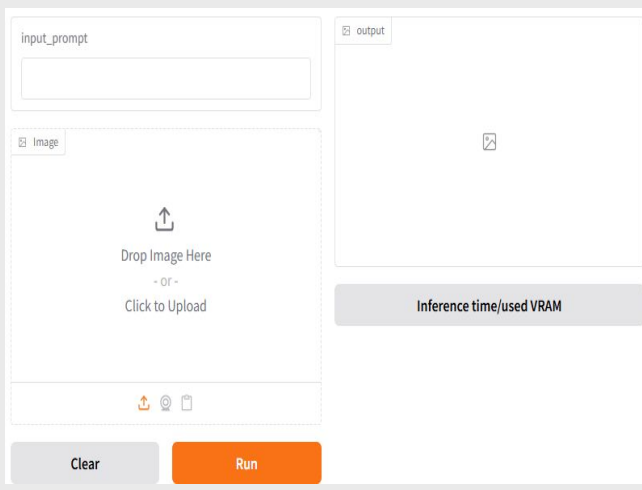
- Research the Diffusion Transformer (DiT) and UNet backbone.
- Study on Stable Diffusion and Low-Rank Adapter.
- Research the way to change floating point precision type from 8-bits to 4-bits.
- Preprocessing randomly prompt on COCO Captions 2024 and small Densely Captioned Images (sDCI) Dataset, including the labeled images.
- Training, evaluation and inference on Stable Diffusion architecture specially FLUX.1.
- Development of a demo application with text/image/sketch input and high-quality image, inference time and used VRAM.

2. METHOD RESEARCH

- Study and explore the how to **Low-rank decompose** can enhance computational and memory efficiency and shifted to new domain without more resources ahead**.**
- **Based-model** use the Diffusion model to know how it generates high-quality samples through denoising process shifted from convolutional-based UNet backbone.
- **Quantization** an effective approach for LLMs to reduce model size. We suppose to implement low-bit inference engine to reduce 4-bits support integer or floating-point data types
- **Develop a demo application** that enables users to use the prompt text/ given image or a sketch to generate high-quality with minimize inference time and VRAM.

3. EXPECTED RESULTS

- Experiment on Fréchet Inception Distance(FID), LPIPS - lower is better, Image Reward, Peak Signal Noise Ratio (PSNR) - higher is better based one FLUX.1 and SDXL version to compare together to get inference time and used Virtual RAM(VRAM).
- Develop an application with gradio to calculate the inference time and VRAM as below image.



NII

Đỗ Trung Hiếu – Trường Đại học Công nghệ Thông tin. ĐHQG TP.HCM
TEL : 0384896154 Email : hieudt.19@grad.uit.edu.vn