DEPLOYING GENERATIVE AI ON MOBILE: COMPRESSING STABLE DIFFUSION BY 4-BITS QUANTIZATION AND DOMAIN-SPECIFIC PROMPTING

TRIỂN KHAI GENERATIVE AI TRÊN MOBILE: NÉN STABLE DIFFUSION BẰNG VIỆC LƯỢNG TỬ HÓA 4-BITS KẾT HỢP VỚI KỸ THUẬT PROMPTING TRONG LĨNH VỰC CỤ THỂ.

GVHD: PGS. TS Lê Đình Duy
Học viên: Đỗ Trung Hiếu - 240101045

Tóm tắt

- Lớp: **CS2205.FEB2025**
- Link Github của nhóm: https://github.com/dotrunghieu0903/CS2205.FEB2025
- Link YouTube video: https://youtu.be/0AdwczYnjKo
- Họ và tên: **Đỗ Trung Hiếu - 240101045**

Giới thiệu

- Các nhà sản xuất smartphone hàng đầu như Apple và Samsung đang chạy đua để tích hợp các công nghệ AI vào thiết bị của họ, đặc biệt tập trung vào AI Camera và **Generative AI** là công cụ mạnh để tạo và sửa ảnh.
- Nó cung cấp cho người dùng khả năng tạo ra và chỉnh sửa những bức ảnh đẹp, chân thực mà không yêu cầu kỹ năng về các công cụ đồ họa phức tạp như Photoshop. Mô hình tạo sinh ảnh(Stable Diffusion) như **text-to-image**, **sketch-to-image** và **inpainting**, cho phép người dùng tạo hoặc chỉnh sửa hình ảnh thông qua các mô tả đơn giản.
- Vì tập trung vào chất lượng hình ảnh cao và đa dạng phong cách nên số lượng tham số của mô hình càng lớn dẫn tới tiêu tốn tài nguyên và khó triển khai trên mobile và edge device.

Mục tiêu

- **Tối ưu hóa khả năng thích nghi và đa dạng hóa mô hình**. Triển khai và hiệu chỉnh (fine-tune) các mô hình khuếch tán (stable diffusion model) với lượng tham số phù hợp trên mobile hay edge device.
- **Cực tiểu hóa tài nguyên tính toán và thời gian suy luận**: Giảm mức tiêu thụ VRAM, thời gian suy luận (inference time) trong quá trình tạo ảnh giữ nguyên chất lượng ảnh và video bằng cách giảm floating-point type precision từ 32-bits, 16-bits xuống 8-bits hoặc 4-bits của kiến trúc Diffusion Transformer(DiT) và UNet backbones:
- **Tối ưu hóa và hiệu chỉnh mô hình với câu mô tả - prompting trong domain cụ thể**: sử dụng Low-rank Adaptation(LoRA) trong Large Language Model(LLM) để giảm lượng tham số của model trong quá trình fine-tune LLM .

Nội dung và Phương pháp

- Nội dung
  - Nghiên cứu kiến trúc của Stable Diffusion như VAE(Variational Autoencoder), UNet Architecture, Text Encoding, Latent Space - Diffusion Process - Denoising step, cơ chế Cross-Attention, Skip Connection.
  - So sánh sự khác biệt UNet và VAE.
  - Thu thập, kế thừa và hiệu chỉnh bộ dataset COCO Captions 2024, MJHQ-30K và Densely Captioned Images với mẫu data đủ lớn tránh overfitting.
  - Chuẩn bị công thức tính VRAM và thời gian suy luận từ lúc bắt đầu tạo sinh ảnh tới lúc hoàn thành để đánh giá mức độ tiêu tốn tài nguyên và thời gian suy luận.

Nội dung và Phương pháp

- Phương pháp
  - Làm thực nghiệm với các Stable diffusion version XL hay Flux.1, thay đổi floating point precision dtype trong pytorch bằng bfloat16, int8, int4(W4A4) để đánh giá kết quả xem inference time, chất lượng ảnh.
  - Lập bảng dữ liệu các metrics để so sánh Stable Diffusion version ví dụ như chất lượng hình ảnh FID: độ đo tương đồng giữa phân phối của tập dữ liệu hình ảnh tạo ra với ảnh gốc - chỉ số thấp càng tốt. Image Reward: độ đo đánh giá chất lượng hình ảnh tạo ra so với prompt input từ user - chỉ số càng cao càng tốt. Về sự tương đồng sử dụng LPIPS (Learned Perceptual Image Patch Similarity) đánh giá sự tương đồng hai hình ảnh PSNR(Peak Signal-to-Noise Ratio) độ đo đánh giá chất lượng của hình ảnh được tái tạo hoặc tạo ra so với một hình ảnh gốc (ground truth).

Kết quả dự kiến

- Phương pháp này dựa trên Stable Diffusion version như XL version, Turbo, Flux.1 version để hiệu chỉnh tham số và giữ nguyên chất lượng ảnh tạo ra.
- So sánh hiệu quả giữa các mô hình qua các độ đo: CLIP, FID, Image Reward, LPIPS, PSNR.
- Dùng các dataset như COCO Captions 2024, MJHQ-30K và Densely Captioned Images (small-DCI) để đánh giá chất lượng hình ảnh sau khi được tạo ra model được optimize lớn nhất qua việc random các câu prompt trong 1 lĩnh vực.

Tài liệu tham khảo

[1]. Muyang Li, Yujun Lin, Zhekai Zhang, Tianle Cai, Xiuyu Li, Junxian Guo, Enze Xie, Chenlin Meng, Jun-Yan Zhu, Song Han: Svdquant: Absorbing Outliers By Low-Rank Components For 4-Bit Diffusion Models. ICLR 2025: 2411-05007.

[2]. Yefei He, Jing Liu, Weijia Wu, Hong Zhou, and Bohan Zhuang. Efficientdm: Efficient quantization-aware fine-tuning of low-bit diffusion models. In ICLR, 2024.

[3]. Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. Quarot: Outlier-free 4-bit inference in rotated llms. NeurIPS, 2024.

[4]. Junyu Chen, Han Cai, Junsong Chen, Enze Xie, Shang Yang, Haotian Tang, Muyang Li, Yao Lu, and Song Han. Deep compression autoencoder for efficient high-resolution diffusion models. ICLR, 2025.

[5]. Yujun Lin, Haotian Tang, Shang Yang, Zhekai Zhang, Guangxuan Xiao, Chuang Gan, and Song Han. Qserve: W4a8kv4 quantization and system co-design for efficient llm serving. In MLSys, 2025.

[6]. Yixiao Li, Yifan Yu, Chen Liang, Nikos Karampatziakis, Pengcheng He, Weizhu Chen, and Tuo Zhao. Loftq: Lora-fine-tuning-aware quantization for large language models. In ICLR, 2024c.

[7]. Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. ICLR, 2025.

[8]. Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. GaLore: Memory-efficient LLM training by gradient low-rank projection. In ICML, 2024a.