GEN AI DEPLOYMENT: COMPRESS STABLE DIFFUSION BY REDUCING VRAM AND COMBINING DOMAIN-SPECIFIC LANGUAGE OPTIMIZATION ON A MOBILE PHONE.

TRIỂN KHAI GENERATIVE AI TRÊN MOBILE: NÉN STABLE DIFFUSION BẰNG TỐI ƯU VRAM KẾT HỢP VỚI KỸ THUẬT PROMPTING TRONG LĨNH VỰC CỤ THỂ.

Tóm tắt(max 400 words)
Hiện nay các tập đoàn điện thoại hàng đầu của thế giới như Apple, SamSung, etc họ đua nhau về việc nâng cấp camera và công nghệ đi cùng nó như AI Camera, AI Assistant. Một trong những mục tiêu của họ là mang lại những bức ảnh đẹp chân thực cho người dùng mà không cần biết những công cụ khó và phức tạp mà dân chuyên nghiệp đồ họa mới sử dụng như Photoshop. Chính vì vậy mà Generative Artificial Intelligence - AI tạo sinh ảnh trên mobile hoặc edge device cụ thể là text to image, sketch to image, depth to image, inpainting ra đời.
Người dùng cuối chỉ cần mô tả yêu cầu của mình để tạo ảnh hay chỉnh sửa ảnh để tạo ra bức ảnh đẹp, có thể là bức ảnh nghệ thuật với nhiều phong cách khác nhau từ một câu prompt - câu mô tả đơn giản từ người dùng và chỉnh nó dễ dàng theo ý người dùng. Hơn nữa những mô hình tạo sinh này lại có thể triển khai ở những thiết bị có tài nguyên thấp và bảo mật thông tin người dùng và theo phong cách cá nhân(personalization).
Có thể tạo ra các image với input đầu vào là câu prompt trong lĩnh vực cụ thể như marketing, design và các domain khác sử dụng thêm kỹ thuật low-rank adapter(LoRa) để điều chỉnh tham số ứng với domain cụ thể, không những có thể fine-tuning câu prompt với domain cụ thể mà còn tiết kiệm tài nguyên tiêu tốn ít bộ nhớ, etc. Mặt khác các mô hình diffusion điển hình là stable diffusion được chứng minh có hiệu quả cao trong việc tạo ra ảnh chất lượng cao tuy nhiên thì những models này lại yêu cầu nhiều bộ nhớ và có độ trễ cao hơn dẫn tới nhiều thách thức trong triển khai. Để vượt qua những điều này phương pháp Compression 4bits ra đời. Mục đích phương pháp này là để chia nhỏ hay lượng tử hóa trọng số và hàm kích hoạt của nó thành bits nhỏ hơn và mô tả câu prompt ngắn gọn súc tích trong một lĩnh vực cụ thể. Không chỉ giữ nguyên chất lượng ảnh khi triển khai xuống thiết bị tài nguyên giới hạn mà còn tốc độ tạo ảnh còn được tối ưu và VRAM tiêu tốn cho quá trình tạo sinh ảnh cũng giảm.

Giới thiệu(max 1 page A4)
Sự phát triển vũ bão của Generative AI đặc biệt là các stable diffusion (mô hình khuếch tán) trong việc tạo ra hình ảnh, videos từ text (văn bản) đem lại sự phát triển không ngừng mở ra đa dạng định hướng phát triển trong nhiều lĩnh vực. Từ tiếp thị giới thiệu sản phẩm, thiết kế đồ họa, hội họa, điện ảnh cho đến các lĩnh vực sáng tạo nội dung khả năng tạo ra hình ảnh, video hay nội dung quảng cáo hay chiến lược marketing với chi phí tối ưu khiến cho khả năng sáng tạo không giới hạn và có sự hỗ trợ đắc lực từ Generative AI có thể hiện thực hầu như tất cả các ý tưởng, v,v. Sự bùng nổ này thể hiện rõ qua việc 2 năm trở lại đây từ 2023 các Diffusion models liên tục ra đời và được nâng cấp các phiên bản vì vậy mà các nghiên cứu của các nhà khoa học có hơi nghiêng theo hướng tập trung vào chất lượng hình ảnh tạo ra và độ đa dạng phong cách của nội dung tạo sinh ảnh video, điều này vô hình trung dẫn tới việc gia tăng đáng kể số lượng tham số(parameters) của mô hình.
Quá trình tạo và khử nhiễu hình ảnh, số bước suy luận(inference steps), cùng với thời gian hay độ trễ để tạo ra một hình ảnh hoặc video chất lượng cao đang đòi hỏi lượng tài nguyên tính toán đặc biệt như VRAM ngày càng lớn. Điều này dẫn tới khó khăn khi triển khai các mô hình này xuống thiết bị di động hay edge device vốn có tài nguyên hạn chế. Không những vậy chi phí để có thể hiệu chỉnh (fine tune) một mô hình với lượng tham số khổng lồ cũng là một thách thức với các nhà nghiên cứu. Hơn thế nữa việc fine tune lại mô hình với domain mới cụ thể đòi hỏi sự đầu tư về thời gian và tài nguyên tăng thêm gánh nặng cho quá trình nghiên cứu, phát triển và đưa sản phẩm tiếp cận người dùng, ứng dụng rộng rãi.

Nhận thức được hạn chế và thách thức trên nghiên cứu này đặt ra mục tiêu kép là không những tiết kiệm tài nguyên cho các nhà nghiên cứu trong quá trình huấn luyện một mô hình với domain cụ thể mà quan trọng hơn là mô hình sau khi huấn luyện, triển khai có thể giảm tài nguyên tiêu thụ. Cụ thể, mục tiêu của nghiên cứu là phát triển một phương pháp hoặc kiến trúc mô hình có thể giảm đáng kể lượng VRAM cần thiết trong quá trình tạo sinh hình ảnh hoặc video. Đồng thời, nghiên cứu hướng tới việc tối ưu hóa thời gian suy luận (inference time) xuống khoảng 10-12 giây hoặc giảm thiểu số bước khử nhiễu (inference steps) xuống mức thấp (ví dụ: 10-20 bước), mà vẫn đảm bảo rằng hình ảnh và video được tạo ra đạt chất lượng cao. Điều cốt yếu là các sản phẩm tạo sinh này phải đúng với ý định của người dùng cuối, dù đó là từ một câu lệnh văn bản mô tả (prompt), một bản phác thảo (sketch), hay một hình ảnh đầu vào được cung cấp, tạo ra một tác phẩm có phong cách nghệ thuật hay nhiều phong cách khác nhau so với ảnh gốc của người dùng. Nghiên cứu này kỳ vọng sẽ góp phần giải quyết những nút thắt hiện tại, mở đường cho việc ứng dụng rộng rãi Generative AI trên đa dạng các nền tảng, đưa Generative AI đến gần hơn với người dùng cuối và thúc đẩy sự phát triển bền vững của lĩnh vực AI tạo sinh.

Input: Có thể câu prompt mô tả, hoặc một phác họa hoặc hình ảnh gốc của người dùng.
Output: Hình ảnh tạo sinh từ mô tả, phác họa, ảnh gốc của người dùng.

Mục tiêu(viết trong 3 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 với mobile hay edge device. Đa dạng các models với phong cách khác nhau chính là việc sử dụng các pre-trained models hoặc kỹ thuật quantization models và kiến trúc Diffusion Transformer(DiT) và UNet backbones.

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 tức là thay vì dùng dtype truyền thống nhu 32-bit(FP32), 16-bit(FP16 hoặc BF16) hay 8-bit(INT8), 4-bit(INT4) thì phương pháp này dùng 4 bits hoặc thấp hơn.

Tối ưu hóa hiệu chỉnh mô hình với câu mô tả - prompt input từ người dùng cuối: sử dụng kỹ thuật Low-rank Adaptation(LoRA) trong large language model kỹ thuật này kỳ vọng điều chỉnh các mô hình có tham số lớn hiệu quả để giảm tài nguyên tính toán cần thiết.

Nội dung và phương pháp

Để hiện thực mục tiêu trên nội dung cần chuẩn bị:

- Nghiên cứu kiến trúc của Stable Diffusion models như VAE(Variational Autoencoder), UNet Architecture, Text Encoding, Latent Space - Diffusion Process - Denoising step, cơ chế Cross-Attention, Skip Connection, Reverse Process.
- So sánh sự khác biệt để biết khi nào dùng UNet hay VAE.
- Thu thập, kế thừa và hiệu chỉnh bộ dataset MJHQ-30K và Densely Captioned Images với lượng data đủ lớn để tránh overfitting.
- Làm thực nghiệm với các model Stable diffusion version XL hay Flux, thay đổi floating point precision dtype trong pytorch bằng bfloat16, int8, int4(W4A4) để đánh giá kết quả xem inference time, chất lượng ảnh.
- Chuẩn bị công thức cách tính VRAM và thời gian suy luận từ lúc bắt đầu tạo sinh ảnh tới lúc hoàn thành để đánh giá được mức độ sử dụng tài nguyên và thời gian tiêu tốn.
- Lập bảng dữ liệu các metrics để so sánh các Stable Diffusion để chứng minh hypothesis, câu hỏi nghiên cứu có đúng không 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 này càng thấp càng tốt. Image Reward: độ đo này đánh giá chất lượng hình ảnh tạo ra so với prompt input từ user - chỉ số càng cao càng tốt. Về sự tương đồng(Similarity) sử dụng LPIPS (Learned Perceptual Image Patch Similarity) đánh giá sự tương đồng hai hình ảnh và PSNR(Peak Signal-to-Noise Ratio) độ đo này đánh giá chất lượng của hình ảnh được tái tạo hoặc tạo ra so với một hình ảnh gốc (ground truth) - chỉ số này càng cao càng tốt.

Kết quả mong đợi

- Phương pháp này được đánh giá trên các stable diffusion backbone version như stable diffusion version 1.5, version turbo, version Flux.
- Phương pháp này được đánh giá bằng các độ đo (metrics) như CLIP, FID, ImageReward, LPIPS, PSNR
- Dùng các dataset từ MJHQ-30K và Densely Captioned Images (small-DCI) để đánh giá chất lượng hình ảnh sau khi được tạo ra sau khi optimize model bởi random tạo ra các câu prompt trong 1 lĩnh vực.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[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.