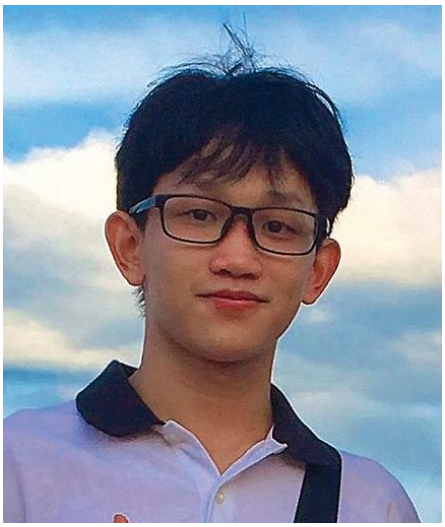
THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo: https://youtu.be/7qQL_j7FIKM
- Link slides (dạng .pdf đặt trên Github của nhóm): https://github.com/HungStark/CS519.O11/blob/main/Slide.pdf

• Họ và Tên: Phạm Văn Hùng
• MSSV: 21522124
 
	• Lớp: CS519.O11
	• Tự đánh giá (điểm tổng kết môn): 7/10
	• Số buổi vắng: 3
	• Số câu hỏi QT cá nhân: ??
	• Số câu hỏi QT của cả nhóm: ??
	• Link Github: https://github.com/HungStark/CS519.O11/
	• Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:
	  ○ Lên ý tưởng cho đồ án
	  ○ Tìm hiểu nội dung kiến thức
	  ○ Làm Slide báo cáo đồ án
	  ○ Quay và chỉnh sửa Video

• Họ và Tên: Nguyễn Tiến Đạt
• MSSV: 21520707
 
	• Lớp: CS519.O11
	• Tự đánh giá (điểm tổng kết môn): 8/10
	• Số buổi vắng: 0
	• Số câu hỏi QT cá nhân: ??
	• Số câu hỏi QT của cả nhóm: ??
	• Link Github: https://github.com/HungStark/CS519.O11/
	• Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:
	  ○ Lên ý tưởng cho đồ án
	  ○ Tìm hiểu nội dung kiến thức
	  ○ Viết nội dung báo cáo đồ án
	  ○ Làm Poster
	  ○ Quay Video

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

CƠ CHẾ CHÚ Ý DẢI KHÔNG CẦN TINH CHỈNH CHO BÀI TOÁN TỔNG HỢP VÀ CHỈNH SỬA HÌNH ẢNH NHẤT QUÁN

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

TUNING-FREE STRIP ATTENTION CONTROL FOR CONSISTENT IMAGE SYNTHESIS AND EDITING

TÓM TẮT

Các mô hình tạo hình ảnh từ văn bản hiện nay đã rất thành công với việc tạo ra các hình ảnh, tuy nhiên, vẫn luôn tồn tại những vấn đề với việc tổng hợp và chỉnh sửa hình ảnh nhất quán khiến việc ứng dụng của bài toán chỉ dừng lại ở các hình ảnh tĩnh. Tuy nhiên, nếu có thể tạo ra không chỉ các hình ảnh đơn lẻ mà là một chuỗi những hình ảnh nhất quán thì ứng dụng của bài toán này sẽ mở rộng hơn rất nhiều. Nhận thấy tiềm năng ấy, nhiều mô hình mới đã được tạo ra và nổi bật nhất là: Cơ chế tự chú ý không cần tinh chỉnh (Tuning-Free Mutual Self-Attention Control). Tuy nhiên, mô hình này vẫn chưa hoàn hảo, đặc biệt là độ phức tạp vẫn còn cao khi sử dụng cơ chế tự chú ý (Self-Attention). Vì vậy, chúng tôi đề xuất một giải pháp mới là sử dụng cơ chế chú ý dải (Strip Attention) thay cho tự chú ý từ đó tối ưu hóa thời gian chạy của mô hình. Cùng với sự cải tiến về thời gian chạy này, chúng tôi mô hình vẫn đảm bảo được tính chính xác, từ đó làm nền móng cho việc sử dụng AI để tạo ra đoạn video.

GIỚI THIỆU

Bài toán sinh ảnh từ một đoạn văn bản là một bài toán quan trọng trong thị giác máy tính, có nhiều ứng dụng trên các khía cạnh nghệ thuật, đời sống và còn có thể dùng để sinh bộ dữ liệu nhằm huấn luyện cho các mô hình thị giác máy tính. Tổng hợp và chỉnh sửa hình ảnh nhất quán là một bài toán con của bài toán sinh ảnh từ một đoạn văn bản. Mục tiêu của bài toán tổng hợp và chỉnh sửa hình ảnh nhất quán là nhằm

sinh ra nhiều ảnh của cùng một đối tượng/nhân vật/… nhưng với những tư thế, trạng thái khác nhau, vì vậy bài toán con này là cần thiết để có thể tạo truyện tranh và tạo video ngắn với các mô hình sinh ảnh từ văn bản [1].

Gần đây, các mô hình sinh ảnh từ văn bản đã đạt được những bước tiến lớn [2, 3]. Tuy nhiên, vẫn còn khoảng cách rất xa giữa nhu cầu và phương pháp để sinh ra và chỉnh sửa hình ảnh nhất quán [1]. Các mô hình chỉnh sửa hình ảnh từ văn ảnh, ví dụ như mô hình Prompt-to-Prompt với cơ chế chú ý chéo [4] cũng đã đạt được những thành tựu lớn, tuy nhiên nó thiếu khả năng thay đổi tư thế, trạng thái,… của đối tượng mà không làm thay đổi ngoại trang (textures) của đối tượng đó. Để giải quyết vấn đề này, người ta đã đề xuất mô hình MasaCtrl [1] cho phép thay đổi tư thế, trạng thái đối tượng với cơ chế tự chú ý chung không cần tinh chỉnh (Tuning-Free Mutual Self-Attention Control), tuy nhiên độ phức tạp bậc hai của cơ chế tự chú ý đã khiến cho mô hình này gặp khó khăn đối với ảnh có kích thước lớn [5].

Vì vậy, câu hỏi được đặt ra là: làm thế nào để có thể giảm thiểu độ phức tạp của cơ chế tự chú ý nhưng vẫn đảm bảo được mô hình có thể thực hiện được việc tổng hợp và chỉnh sửa hình ảnh nhất quán. Để trả lời câu hỏi trên, chúng tôi quan tâm đến một cơ chế chú ý mới, gọi là cơ chế chú ý dải (Strip Attention) [5]. Mặc dù đây là cơ chế chú ý mới nhưng nó có tính phổ quát cao và đã được áp dụng cho nhiều bài toán trong thị giác máy tính như: bài toán khôi phục hình ảnh (image restoration), theo dõi chuyển động vật thể (object tracking),...

Input: hình ảnh thực tế hoặc hình ảnh được sinh ra từ mô hình sinh ảnh, một đoạn văn bản thể hiện sự thay đổi trạng thái của ảnh mới so với ảnh gốc.

Output: một ảnh mới có xuất hiện sự thay đổi trạng thái từ ảnh gốc.

Input real image "... jumping ..." "A sitting boy" → "... standing ..."

MỤC TIÊU

● Xây dựng được mô hình mới sử dụng Tuning-Free Strip Attention Control thực hiện được việc tổng hợp và chỉnh sửa hình ảnh nhất quán.

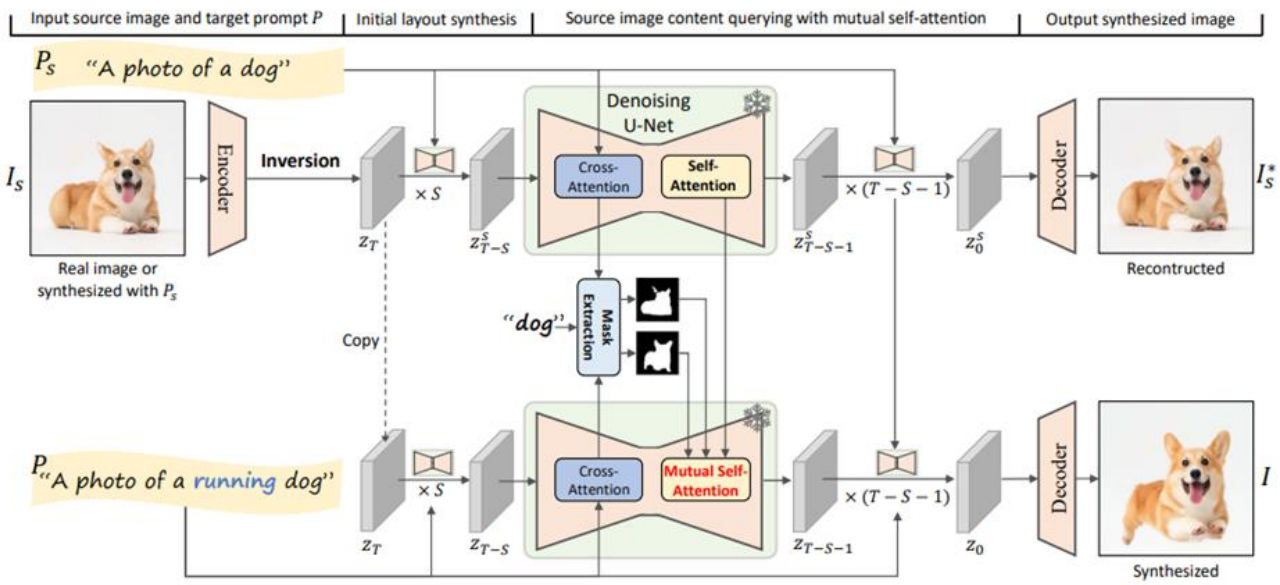
● Mô hình cải thiện về tốc độ nhưng vẫn đảm bảo về tính chính xác ở mức chấp nhận được.

● Đánh giá được độ tốt của mô hình mới so với mô hình cũ dựa trên tốc độ, độ chân thực, sự ổn định và khả năng tinh chỉnh từ đó kết luận về tính khả thi của giải pháp này.

NỘI DUNG VÀ PHƯƠNG PHÁP

1. Nội dung:

Sở dĩ cơ chế chú ý dải có thể giải quyết được triệt để vấn đề về độ phức tạp của cơ chế chú ý là vì cơ chế này đã thay đổi một cách bản chất cách tính mức chú ý trên mỗi điểm ảnh: thay vì tính các ma trận Query, Key, Value như cơ chế tự chú ý trước đây thì chỉ cần đưa vào một dải ngang (hoặc dải dọc) k điểm ảnh, rồi đưa qua một lớp Global Avarage Pooling, một lớp tích chập (convolution) kích thước 1x1 và một lớp sigmoid [5]. Hơn thế nữa, mô hình MasaCtrl là mô hình tốt nhất gần đây cho bài toán tổng hợp và chỉnh sửa hình ảnh nhất quán, nhưng có điểm yếu về độ phức tạp của cơ chế tự chú ý. Vì vậy, giải pháp chúng tôi đưa ra là thay thế cơ chế tự chú ý trong mô hình MasaCtrl bằng cơ chế chú ý dải nhằm giảm thiểu độ phức tạp của cơ chế chú ý trong mô hình.



Tuy nhiên, khi thay đổi cơ chế chú ý thành cơ chế chú ý dải thì xuất hiện thêm một vấn đề, đó là các thuật toán tuning-free và cơ chế tự chú ý chung (mutual self-attention) sẽ bị ảnh hưởng bởi vì những thuật toán trên được xây dựng trên nền của cơ chế tự chú ý [1]. Do đó, chúng tôi cần phải nghiên cứu điều chỉnh lại các thuật toán có liên quan nhằm đảm bảo những thuật toán trên có thể thực hiện tốt trên cơ chế chú ý dải.

2. Phương pháp:

● Tìm hiểu về phương pháp huấn luyện của mô hình gốc MasaCtrl và sau đó cài đặt lại mô hình.

● Tìm hiểu về cơ chế chú ý dải, và sau đó cài đặt cơ chế chú ý dải (Strip Attention) trên mô hình MasaCtrl.

● Fine-tuning các siêu tham số của cơ chế chú ý dải để từ đó chọn ra các siêu tham số phù hợp nhất, đảm bảo cho mô hình vẫn thực hiện tốt bài toán đồng thời đảm bảo độ phức tạp của cơ chế chú ý được giảm thiểu một cách đáng kể.

● Tìm hiểu thuật toán tuning-free và cơ chế tự chú ý chung trên mô hình gốc, sau đó nghiên cứu một thuật toán tuning-free mới và cơ chế chú ý chung mới đối với cơ chế chú ý dải.

● Chuẩn bị và thu thập các tập dữ liệu để huấn luyện và đánh giá mô hình (Có thể chụp ảnh hoặc sử dụng hệ thống sinh ảnh từ văn bản)

● Tiến hành cài đặt các thuật toán và cơ chế đã được nghiên cứu, sau đó đánh giá và so sánh với mô hình trước đó (mô hình MasaCtrl gốc) về thời gian chạy, độ chân thực, sự ổn định, khả năng tinh chỉnh và phân tích các trường hợp lỗi từ đó đề xuất hướng phát triển cho tương lai.

KẾT QUẢ MONG ĐỢI

● Mô hình xây dựng được giúp khắc phục nhược điểm của mô hình trước đó về độ phức tạp, đồng thời chạy nhanh hơn ít nhất 10 lần so với mô hình tốt nhất hiện tại là MasaCtrl, nhưng vẫn đảm bảo được độ chính xác và độ nhất quán của mô hình không bị suy giảm đáng kể (dưới 5%).

● Báo cáo về kĩ thuật, các bảng so sánh, đánh giá kết quả thực nghiệm giữa mô hình xây dựng được và các mô hình khác trên cùng bài toán.

● Bảng thống kê các trường hợp lỗi:

- Trạng thái của đối tượng không chính xác.
- Ngoại trang của đối tượng bị thay đổi so với ảnh gốc
- Hình ảnh của đối tượng được sinh ra không chân thực (ví dụ người có 3 tay, bàn tay có 6 ngón,...)

TÀI LIỆU THAM KHẢO

[1] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, Yinqiang Zheng: MasaCtrl: Tuning-Free Mutual Self-Attention Control for Consistent Image Synthesis and Editing. ICCV 2023: 22503-22513

[2] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, Dimitris N. Metaxas: StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks. CoRR abs/1612.03242 (2016)

[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer: High-Resolution Image Synthesis with Latent Diffusion Models. CoRR abs/2112.10752 (2021)

[4] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, Daniel Cohen-Or: Prompt-to-Prompt Image Editing with Cross-Attention Control. ICLR 2023

[5] Yuning Cui, Yi Tao, Luoxi Jing, Alois Knoll: Strip Attention for Image Restoration. IJCAI 2023: 645-653

[6] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, Xiaohu Qie: T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models. CoRR abs/2302.08453 (2023)