THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút): https://youtu.be/HwZd5E57zPA
  Link slides (dạng .pdf đặt trên Github của nhóm):
  https://github.com/truongvanchinh/CS519.O11/blob/main/slide.pdf

• Họ và Tên: Trương Văn Chinh
• MSSV: 20521137
 
	• Lớp: CS519.O11
	• Tự đánh giá (điểm tổng kết môn): 8.5/10
	• Số buổi vắng: 3
	• Số câu hỏi QT cá nhân: 11
	• Link Github: https://github.com/truongvanchinh/CS519.O11
	• Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:
	  ○ Lên ý tưởng đề tài.
	  ○ Viết báo cáo, slide, poster.
	  ○ Làm video YouTube.

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

CONTROLNET: ĐIỀU KHIỂN BỐ CỤC KHÔNG GIAN TRONG MÔ HÌNH KHUẾCH TÁN VĂN BẢN-SANG-ẢNH

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

CONTROLNET: SPATIAL LAYOUT CONTROL IN TEXT-TO-IMAGE DIFFUSION MODELS

TÓM TẮT *(Tối đa 400 từ)*

Nghiên cứu này đề xuất ControlNet, một kiến trúc mạng nơ-ron cho phép bổ sung điều khiển bố cục không gian vào các mô hình khuếch tán văn bản-sang-ảnh đã được đào tạo trước trên quy mô lớn. ControlNet tận dụng các lớp mã hóa sâu và mạnh mẽ của mô hình khuếch tán gốc, đồng thời sử dụng các "zero convolution" để học các điều khiển bổ sung theo không gian một cách hiệu quả. Nghiên cứu đánh giá ControlNet với các điều kiện đầu vào đa dạng như đường viền, độ sâu, phân vùng, tư thế người, v.v., trên mô hình Stable Diffusion. Nghiên cứu này mong đợi ControlNet sẽ là một kỹ thuật hiệu quả để cải thiện khả năng kiểm soát bố cục của các mô hình khuếch tán văn bản-sang-ảnh, hứa hẹn mở rộng việc sử dụng chúng trong các ứng dụng thực tế.

GIỚI THIỆU *(Tối đa 1 trang A4)*

Con người thường có những ý tưởng hình ảnh độc đáo trong tâm trí mà muốn nắm bắt. Sự ra đời của các mô hình khuếch tán văn bản-sang-ảnh (text-to-image diffusion models) mở ra khả năng tạo ra những hình ảnh đẹp mắt chỉ bằng cách nhập mô tả text. Tuy nhiên, các mô hình này có giới hạn trong việc kiểm soát bố cục không gian của hình ảnh. Cụ thể, việc diễn tả chính xác các bố cục, tư thế, vị trí, hình dạng phức tạp chỉ bằng text đôi khi khó khăn.

Việc tạo ra một hình ảnh khớp chính xác với hình ảnh trong tâm trí thường đòi hỏi

nhiều thử nghiệm, sửa text, xem kết quả rồi lặp lại. Điều này có thể tốn thời gian và công sức. Vậy có cách nào tốt hơn cho phép người dùng cung cấp thêm hình ảnh trực tiếp thể hiện bố cục mong muốn không?

Trong bài nghiên cứu này, chúng tôi đề xuất một phương pháp mới để kiểm soát bố cục không gian của các mô hình khuếch tán văn bản-sang-ảnh bằng cách thêm các điều kiện đầu vào cục bộ theo không gian. Phương pháp này sử dụng một kiến trúc mạng nơ-ron gọi là ControlNet để kết hợp các điều kiện đầu vào với mô hình khuếch tán. ControlNet được thiết kế để học cách kết hợp các điều kiện đầu vào một cách hiệu quả, đồng thời vẫn duy trì chất lượng và khả năng của mô hình khuếch tán gốc.

Input:

- Đoạn văn bản mô tả về hình ảnh, kèm nhiều thông tin như vị trí, màu sắc.
- Điều kiện bổ sung: 1 hình ảnh thể hiện thông tin điều kiện bổ sung như: khung xương, nét vẽ, đường viền.

Output:

- Hình ảnh được tạo ra từ văn bản mô tả và điều kiện bổ sung.

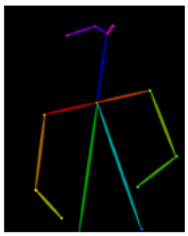


Figure 1: Human pose

Figure 2: Output image

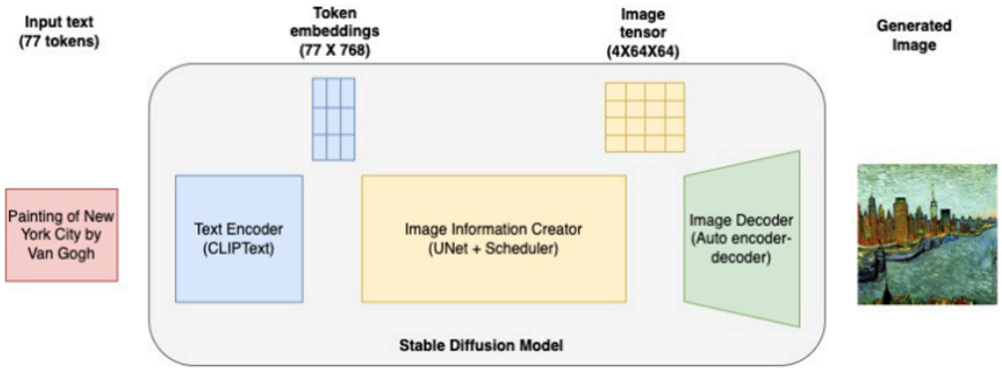
MỤC TIÊU

- Tìm hiểu tổng quan về Latent Diffusion Model trong bài toán sinh ảnh từ văn bản.
- Cải tiến Latent Diffusion Model.

NỘI DUNG VÀ PHƯƠNG PHÁP

Nội dung 1: Tìm hiểu tổng quan, cách thức hoạt động và các ưu điểm của Latent Diffusion Models so với các phương pháp trước đó (GANs[4], Diffusion Models[2]).

- Tìm hiểu các thông tin tổng quan của LDMs[3], GANs[4], Diffusion Models[2] sau đó tiến hành so sánh độ hiệu quả đối với ảnh độ phân giải cao, khả năng tiết kiệm tài nguyên, độ ổn định.
- Cách mà LDMs sử dụng không gian tiềm ẩn (Departure to Latent Space). Không gian tiềm ẩn là một không gian chiều thấp hơn không gian dữ liệu gốc, nhưng vẫn giữ lại thông tin quan trọng về dữ liệu gốc.
- LDMs chia quá trình huấn luyện thành 2 giai đoạn: Huấn luyện bộ mã hóa tự động (autoencoder) thành không gian tiềm ẩn hiệu quả về tính toán. Huấn luyện mô hình khuếch tán (Diffusion Models [2]) trong không gian tiềm ẩn này.
- Xem xét một vài ưu điểm của mô hình về khả năng huấn luyện và tốc độ đánh giá, tính tiết kiệm tài nguyên, có cho phép điều kiện đầu vào đa dạng, có thể được sử dụng cho nhiều tác vụ tổng hợp ảnh khác nhau, ... tổng hợp ảnh độ phân giải cao (megapixel) thì chất lượng có giảm không?
- Tìm hiểu về cấu trúc và cách hoạt động của LDMs[3]:
    + Tìm hiểu về CLIP[6] trong việc chuyển đổi text input thành text encoder.
    + Tìm hiểu về UNet hoạt động như thế nào trong LDMs.
    + Tìm hiểu về Auto encoder-decoder.

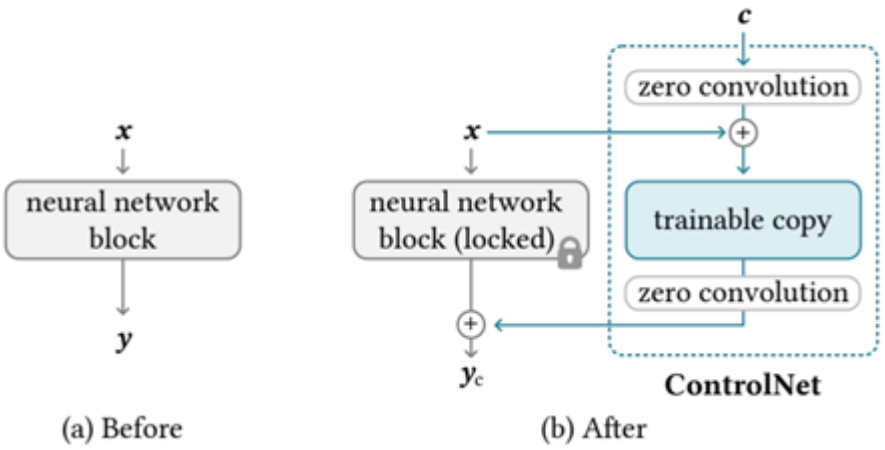


Hình 3: Cấu trúc mạng LDMs

Nội dung 2: Cải tiến Latent Diffusion Model bằng cách thêm điều kiện điều khiển bổ sung như bố cục, hình ảnh khung xương, vị trí các đối tượng với ControlNet[1].

- ControlNet[1] giúp cải thiện mô hình tạo ảnh bằng cách cung cấp thông tin hình ảnh cụ thể.
- ControlNet[1] chèn các điều kiện bổ sung vào các khối của một mạng nơ-ron bằng cách đóng băng lại khối gốc sau đó nhân bản nó ra và khối nhân bản sẽ sử dụng vector điều kiện đầu vào c làm input.
- ControlNet[1] sử dụng 2 lớp tích chập không (zero convolution layers) để kết nối bản sao và khối mạng gốc. Lớp tích chập không có kích thước 1×1 và có trọng số và bias được khởi tạo là 0. Mục đích giúp bảo vệ cốt lõi đã được huấn luyện trước đó trong giai đoạn huấn luyện ban đầu và tích hợp thông tin từ cả hai nguồn để tạo ra kết quả đầu ra tốt nhất.
- Khi đó nếu áp dụng khối mới này cho các mô hình lớn như Stable Diffusion[3] thì các tham số đã bị đóng băng vẫn giữ nguyên, trong khi bản sao có thể sử dụng lại mô hình huấn luyện trước giúp nó nhanh chóng học được các điều kiện mới và xử lý đa dạng đầu vào mà không cần phải đào tạo lại toàn bộ mạng từ

đầu.



Hình 4: Khối mạng nơ-ron trước (a) và sau khi áp dụng ControlNet (b)

- Sau khi áp dụng ControlNet[1] vào Stable Diffusion[3]:
  + Tìm hiểu cách chuyển đổi từng khối trong mô hình ban đầu thành từng khối mới cho mỗi tầng mã hóa của U-Net[5].
  + Tạo một bản sao có thể đào tạo của 12 khối mã hóa và 1 khối trung gian của Stable Diffusion[3].
  + Tạo ra các kết nối đi từ các khối encoder đến zero convolution, các kết nối này có tác dụng truyền thông tin từ các khối mã hóa đến các lớp zero convolution của ControlNet[1].
  + Kiểm tra độ hiệu quả tính toán sau khi áp dụng khối kết nối ControlNet[1].

KẾT QUẢ MONG ĐỢI

- Hiểu rõ thông tin cơ bản về cấu trúc và cách hoạt động của LDMs[3] với bài toán sinh ảnh từ văn bản.
- Ảnh được sinh ra đúng với yêu cầu hơn so với phương pháp chỉ dùng các đoạn text thông thường, tăng tính chính xác của kết quả.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1]. Lvmin Zhang, Maneesh Agrawala:

Adding Conditional Control to Text-to-Image Diffusion Models. CoRR abs/2302.05543 (2023).

[2]. Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, Surya Ganguli:

Deep Unsupervised Learning using Nonequilibrium Thermodynamics. CoRR abs/1503.03585 (2015).

[3]. Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer:

High-Resolution Image Synthesis with Latent Diffusion Models. CVPR 2022: 10674-10685.

[4]. Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, Yoshua Bengio:

Generative Adversarial Nets. NIPS 2014: 2672-2680.

[5]. Olaf Ronneberger, Philipp Fischer, Thomas Brox:

U-Net: Convolutional Networks for Biomedical Image Segmentation. CoRR abs/1505.04597 (2015).

[6]. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever:

Learning Transferable Visual Models From Natural Language Supervision. CoRR abs/2103.00020 (2021).