ĐẢM BẢO NGỮ NGHĨA ẢNH TẠO BỞI MÔ HÌNH DIFFUSION BẰNG PHƯƠNG PHÁP GIÁM SÁT

Trần Siêu

Trường Đại học Công nghệ Thông tin, ĐHQG-HCM

Tóm tắt

Chúng tôi giới thiệu phương pháp giúp tăng độ chính xác về mặt ngữ nghĩa của hình ảnh sinh ra bằng mô hình Diffusions cho bài toán Text-to-Image. Cụ thể:

- Đề xuất giám sát ngữ nghĩa của kết quả đầu ra bằng Descriminator trong quá trình huấn luyện.
- Đề xuất phương pháp chuyển đổi text prompt thành segmentation map nhằm cung cấp thông tin trực quan hơn cho mô hình.

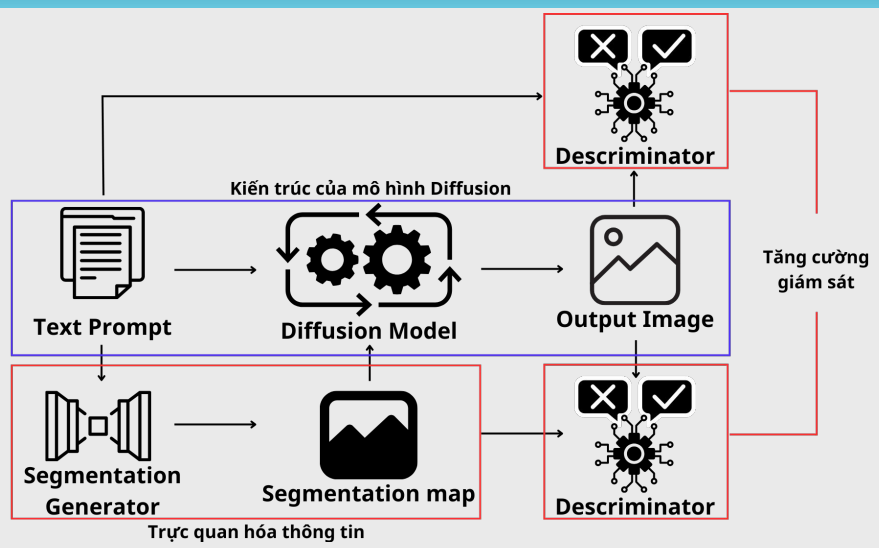
Động lực

Các mô hình Diffusion đang thể hiện những hiệu suất đáng kể trong bài toán Text-to-Image khi cho ra ảnh có độ phân giải, độ hài hòa và tính chân thực cao. Tuy nhiên, những mô hình này không đảm bảo được tính đúng đắn của hình ảnh đầu ra so với text prompt. Bởi vì:

- Ngữ nghĩa của ảnh đầu ra không được giám sát trong quá trình huấn luyện.
- Thông tin từ text prompt thiếu trực quan, khó biểu đạt được mối quan hệ giữa các đối tượng trong không gian.

Tổng quan

Hình 1. Minh họa sự khác biệt ngữ nghĩa giữa input và output.



Hình 2. Hình ảnh minh họa tổng quan phương pháp.

Mô tả

1. Nội dung nghiên cứu

- Giám sát về mặt ngữ nghĩa trong quá trình huấn luyện của mô hình Diffusion.
- Sinh ra segmentation map từ prompt để cung cấp thông tin không gian trực quan hơn cho mô hình.
- Kết hợp thông tin từ segmentation map sinh ra vào latent space của mô hình dưới dạng điều kiện và thực hiện giám sát trong huấn luyện để đảm bảo tính hài hòa trong bố cục đầu ra.

2. Phương pháp nghiên cứu

- Tìm hiểu về quá trình huấn luyện của mô hình Diffusion-based như Latent Diffusion. Tìm hiểu về phương pháp kết hợp Descriminator trong quá trình huấn luyện của mô hình Diffusion như ALDM. Tìm hiểu phương pháp so sánh khoảng cách ngữ nghĩa của hình ảnh và văn bản dựa trên các mô hình vision-language như CLIP.
- Tìm hiểu các phương pháp trực quan hóa thông tin không gian từ văn bản. Tìm hiểu các phương pháp tạo segmentation map đạt hiệu quả cao như dựa trên Unet, Diffusion. Đề xuất phương pháp tạo segmentation map bằng prompt dựa trên những phương pháp tìm hiểu được.
- Tìm hiểu các phương pháp trực quan hóa thông tin không gian từ văn bản. Tìm hiểu các phương pháp tạo segmentation map đạt hiệu quả cao như dựa trên Unet, Diffusion. Đề xuất phương pháp tạo segmentation map bằng prompt dựa trên những phương pháp tìm hiểu được.
- Tìm hiểu các phương pháp, độ đo để đánh giá sự khác biệt giữa nội dung hình ảnh và văn bản như TIFA, CLIP Score, đánh giá sự hài hòa, chính xác về bố cục như FID, mIoU.

3. Kết quả mong đợi

- Phương pháp huấn luyện có giám sát các mô hình Diffusion-based như Latent Diffusion, giúp nâng cao độ chính xác giữa nội dung prompt và hình ảnh được đánh giá bởi các độ đo như TIFA, CLIP Score, đồng thời bảo toàn được khả năng tạo ảnh có tính hài hòa trong bố cục của các mô hình pretrained được đánh giá bởi các độ đo như FID, mIoU.
- Mô hình cho phép tạo ra segmentation map từ prompt có độ chính xác cao, đánh giá bởi độ đo như mIoU. Giúp bổ sung thông tin vào latent space của mô hình Diffusion, từ đó nâng cao sự chính xác trong quá trình sinh ảnh.
- Báo cáo về kỹ thuật, cách cài đặt và kết quả thực nghiệm của các phương pháp.