**Mô hình khuếch tán (Diffusion Models) đánh bại GAN (Generative Adversarial Networks - mạng sinh đối nghịch) về tổng hợp hình ảnh**

**Abstract**

Chúng tôi cho thấy rằng các mô hình khuếch tán có thể đạt chất lượng mẫu ảnh vượt trội so với các mô hình sinh hiện đại nhất hiện nay. Chúng tôi đạt được điều này trong tổng hợp ảnh không điều kiện bằng cách tìm ra một kiến trúc tốt hơn thông qua một loạt các thử nghiệm loại bỏ. Đối với tổng hợp ảnh có điều kiện, chúng tôi cải thiện chất lượng mẫu bằng cách sử dụng classifier guidance: một phương pháp đơn giản, hiệu quả về tính toán để đánh đổi đa dạng (diversity) với độ trung thực (fidelity) bằng cách sử dụng gradient từ một bộ phân loại.

* **Đa dạng (diversity)**: Sự khác nhau và phong phú giữa các mẫu ảnh được tạo ra. Một mô hình có tính đa dạng cao sẽ tạo ra nhiều mẫu ảnh khác nhau, không bị lặp lại hoặc quá giống nhau.
* **Độ trung thực (fidelity)**: Mức độ chính xác và chân thực của các mẫu ảnh so với dữ liệu gốc. Một mẫu ảnh có độ trung thực cao sẽ giống với ảnh thật và có chất lượng cao.
  1. Introduction

Trong vài năm qua, các mô hình sinh (generative models) đã đạt được khả năng tạo ra ngôn ngữ tự nhiên giống con người, hình ảnh tổng hợp chất lượng cao và giọng nói cùng âm nhạc phong phú. Các mô hình này có thể được sử dụng theo nhiều cách khác nhau, chẳng hạn như tạo hình ảnh từ các câu lệnh văn bản hoặc học các biểu diễn đặc trưng hữu ích. Mặc dù các mô hình này đã có khả năng tạo ra hình ảnh và âm thanh chân thực, nhưng vẫn còn rất nhiều tiềm năng để cải thiện vượt xa so với hiện tại, và các mô hình sinh tốt hơn có thể có tác động sâu rộng đến thiết kế đồ họa, trò chơi, sản xuất âm nhạc và vô số lĩnh vực khác.

Hiện tại, GANs đang giữ vị trí tiên tiến nhất trong hầu hết các nhiệm vụ tạo ảnh, được đo lường bằng các chỉ số chất lượng mẫu như FID, Inception Score và Precision. Tuy nhiên, một số chỉ số này không hoàn toàn phản ánh được tính đa dạng, và đã có nghiên cứu cho thấy rằng GANs nắm bắt ít tính đa dạng hơn so với các mô hình dựa trên xác suất tiên tiến nhất. Hơn nữa, GANs thường khó huấn luyện, dễ bị sụp đổ nếu không chọn cẩn thận các siêu tham số và các phương pháp điều chỉnh. Mặc dù GANs đang giữ vị trí tiên tiến nhất, nhưng những hạn chế của chúng khiến việc mở rộng quy mô và áp dụng vào các lĩnh vực mới trở nên khó khăn. Do đó, đã có nhiều nghiên cứu nhằm đạt được chất lượng mẫu tương tự GANs bằng các mô hình dựa trên xác suất. Mặc dù các mô hình này nắm bắt được nhiều tính đa dạng hơn và thường dễ mở rộng và huấn luyện hơn so với GANs, chúng vẫn chưa đạt được độ trung thực hình ảnh cao. Hơn nữa, ngoại trừ VAEs, việc lấy mẫu từ các mô hình này thường chậm hơn so với GANs về thời gian thực.

Các mô hình khuếch tán (diffusion models) là một lớp các mô hình dựa trên xác suất gần đây đã được chứng minh có thể tạo ra hình ảnh chất lượng cao, đồng thời cung cấp các tính chất mong muốn như bao phủ phân phối, mục tiêu huấn luyện không thay đổi và dễ mở rộng quy mô. Các mô hình này tạo mẫu bằng cách dần dần loại bỏ nhiễu khỏi tín hiệu, và mục tiêu huấn luyện của chúng có thể được biểu thị như một giới hạn dưới biến phân có trọng số lại. Lớp mô hình này đã đạt được vị trí tiên tiến nhất trên CIFAR-10, nhưng vẫn còn thua kém GANs trên các tập dữ liệu tạo hình khó như LSUN và ImageNet. Chúng tôi giả thuyết rằng khoảng cách này tồn tại ít nhất vì hai lý do: thứ nhất, kiến trúc mô hình được sử dụng trong tài liệu GAN gần đây đã được nghiên cứu và tinh chỉnh rất nhiều; thứ hai, GANs có thể đánh đổi giữa tính đa dạng và độ trung thực, tạo ra các mẫu chất lượng cao nhưng không bao phủ toàn bộ phân phối. Chúng tôi nhằm mang lại những lợi ích này cho các mô hình khuếch tán, trước tiên bằng cách cải thiện kiến trúc mô hình và sau đó bằng cách xây dựng một cơ chế để đánh đổi tính đa dạng với độ trung thực.

* 1. Background

Các mô hình khuếch tán lấy mẫu từ một phân phối bằng cách đảo ngược một quá trình nhiễu dần dần. Cụ thể, việc lấy mẫu bắt đầu với nhiễu xT và tạo ra các mẫu ít nhiễu dần xT-1, xT-2, ... cho đến khi đạt được mẫu cuối cùng x0. Đặc biệt, một mô hình khuếch tán học cách tạo ra xt-1 ít nhiễu hơn một chút từ xt. Ho và cộng sự tham số hóa mô hình này bằng cách sử dụng một hàm epsilon(theta) (xt , t) dự đoán thành phần nhiễu của một mẫu nhiễu xt. Để huấn luyện hàm này, mỗi mẫu trong một minibatch được tạo ra bằng cách chọn ngẫu nhiên một mẫu dữ liệu x0, một bước thời gian t và nhiễu epsilon, cùng nhau tạo ra một mẫu bị nhiễu xt. Mục tiêu huấn luyện sau đó là || epsilon(theta) (xt , t) - epsilon||^2, tức là một hàm lỗi bình phương trung bình đơn giản giữa nhiễu thực và nhiễu được dự đoán.

**Các chỉ số chất lượng mẫu:**

Chúng tôi sử dụng FID làm chỉ số mặc định để so sánh chất lượng mẫu tổng thể vì nó phản ánh cả độ trung thực và tính đa dạng và đã trở thành chỉ số tiêu chuẩn cho các mô hình sinh tiên tiến. Chúng tôi sử dụng Precision và Recall như là các đại diện cho việc đo lường riêng biệt độ trung thực và tính đa dạng. Chúng tôi bao gồm sFID như một chỉ số phản ánh tốt hơn các mối quan hệ không gian so với FID, và cũng bao gồm Inception Score (IS) như một đại diện khác cho độ trung thực. Khi so sánh với các phương pháp khác, chúng tôi tính lại các chỉ số này bằng cách sử dụng các mẫu công khai hoặc các mô hình bất cứ khi nào có thể. Điều này vì hai lý do: thứ nhất, một số bài báo so sánh với các tập con tùy ý của tập huấn luyện mà không có sẵn; và thứ hai, các khác biệt nhỏ trong triển khai có thể ảnh hưởng đến các giá trị FID. Để so sánh nhất quán, chúng tôi sử dụng toàn bộ tập huấn luyện làm lô tham chiếu và đánh giá các chỉ số cho tất cả các mô hình bằng cùng một mã nguồn.