***Architecture Improvements/ Cải tiến:***

Sử dụng kiến trúc UNet trong mô hình diffusion model để cải thiện chất lượng mẫu so với mô hình trước đó.

Mô hình UNet sử dụng một chuỗi các lớp residual và các lớp downsampling convolution, sau đó là một chuỗi các lớp residual với các lớp upsampling convolution, và các kết nối skip nối các lớp có cùng kích thước không gian. Ngoài ra, một lớp attention toàn cục được sử dụng tại độ phân giải 16x16 với một head duy nhất, và thêm một phép chiếu của timestep embedding vào mỗi khối residual.

Ảnh có chứa biểu đồ, hàng, Kế hoạch, Song song

Mô tả được tạo tự động

Hình minh họa UNet

Một số cải tiến khác:

* Tăng độ sâu so với độ rộng trong khi giữ kích thước mô hình tương đối không đổi.
* Tăng số lượng đầu chú ý. Sử dụng attention ở các độ phân giải khác nhau (32x32, 16x16 và 8x8) thay vì chỉ ở 16x16.
* Sử dụng khối residual của BigGAN để thực hiện việc tăng kích thước (upsampling) và giảm kích thước (downsampling) của các đặc trưng (activations) để cải thiện hiệu suất của mô hình.
* Thay đổi tỷ lệ (Rescaling) lớp residual với hệ số p=1/2.

Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ

Mô tả được tạo tự động

Hình minh họa BigGAN và ResBlock Upsampling/ ResBlock Downsampling

Mô hình với những thay đổi kiến trúc này được huấn luyện trên ImageNet (128x128) và so sánh hiệu suất của chúng bằng chỉ số Fréchet Inception Distance (FID) tại hai điểm huấn luyện khác nhau (700K and 1200K iterations) như bảng dưới. . Ngoại trừ việc thay đổi tỷ lệ lớp residual, tất cả các sửa đổi khác đều cải thiện hiệu suất của mô hình.

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, số

Mô tả được tạo tự động

Ngoài ra, 2 kiểu cấu hình attention được nghiên cứu sử dụng là constant attention heads và constant channels per head. Bảng dưới so sánh kết quả của 2 loại cấu hình và cho thấy việc tăng hay giảm số lượng đầu chú ý và số lượng kênh trên 1 đầu chú ý ảnh hưởng thế nào đến chỉ số FID.

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, số

Mô tả được tạo tự động

Ngoài ra, 1 lớp được gọi là adaptive group normalization (AdaGN) cũng được thử nghiệm và sử dụng, trong đó lớp timestep và class embedding được tích hợp vào từng khối ResBlock sau thao tác chuẩn hóa nhóm với công thức: AdaGN(h, y) = ys GroupNorm(h) + yb trong đó h là kích hoạt trung gian của khối sau lần tích chập đầu tiên, và y = [ys, yb] thu được từ phép chiếu tuyến tính của lớp timestep và class embedding.

Trong phần còn lại của bài báo, mô hình cuối cùng được chọn có cải thiện như sau:

* Chiều rộng thay đổi với 2 khối ResBlock cho mỗi độ phân giải.
* Lớp chú ý nhiều đầu (multiple heads) với 64 kênh trên 1 đầu.
* Attention tại độ phân giải 32, 16 và 8.
* Sử dụng BigGAN cho upsampling và downsampling, cũng như sư dụng AdaGN để chèn timestep và class embedding vào các khối.