

Tạo văn bản từ hình ảnh

Phạm Trung Sơn, Nguyễn Tuấn Sinh, Nguyễn Tiến Sỹ, Trần Hồ Phi Long

Đại học Công nghiệp TP. Hồ Chí Minh

Thị giác máy tính
Ngày 18 tháng 12 năm 2022



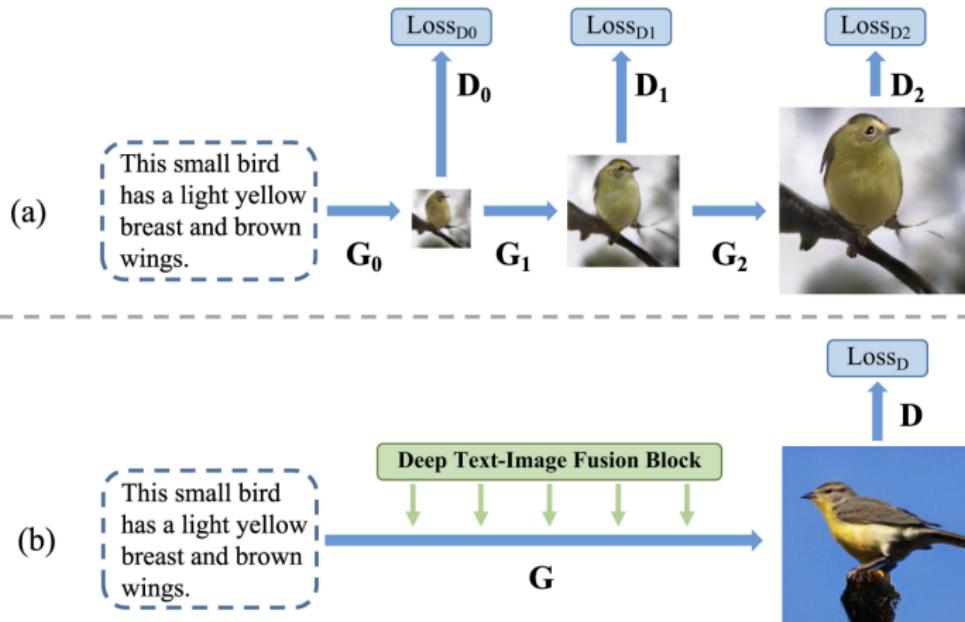
Nội dung

- ① Giới thiệu
- ② Các nghiên cứu liên quan
- ③ Phương pháp tiếp cận
 - Kiến trúc DF-GAN
 - Tổng quan mô hình
 - Trục chuyển văn bản thành hình ảnh một giai đoạn
 - Công cụ Discriminator nhận biết mục tiêu
 - Hình phạt chuyển màu nhận biết phù hợp (Matching-Aware Gradient Penalty)
 - Đầu ra một chiều (One-Way Output)
 - Kết hợp văn bản - hình ảnh hiệu quả (Efficient TextImage Fusion)
- ④ Thực nghiệm
 - Đánh giá định lượng
 - Đánh giá định tính
 - Hạn chế
- ⑤ Kết luận

Nội dung

- 1 Giới thiệu
- 2 Các nghiên cứu liên quan
- 3 Phương pháp tiếp cận
 - Kiến trúc DF-GAN
 - Tổng quan mô hình
 - Trục chuyển văn bản thành hình ảnh một giai đoạn
 - Công cụ Discriminator nhận biết mục tiêu
 - Hình phạt chuyển màu nhận biết phù hợp (Matching-Aware Gradient Penalty)
 - Đầu ra một chiều (One-Way Output)
 - Kết hợp văn bản - hình ảnh hiệu quả (Efficient TextImage Fusion)
- 4 Thực nghiệm
 - Đánh giá định lượng
 - Đánh giá định tính
 - Hạn chế
- 5 Kết luận

Giới thiệu



Hình: (a) Stack GAN (b) DF-GAN

Nội dung

- 1 Giới thiệu
- 2 Các nghiên cứu liên quan
- 3 Phương pháp tiếp cận
 - Kiến trúc DF-GAN
 - Tổng quan mô hình
 - Trục chuyển văn bản thành hình ảnh một giai đoạn
 - Công cụ Discriminator nhận biết mục tiêu
 - Hình phạt chuyển màu nhận biết phù hợp (Matching-Aware Gradient Penalty)
 - Đầu ra một chiều (One-Way Output)
 - Kết hợp văn bản - hình ảnh hiệu quả (Efficient TextImage Fusion)
- 4 Thực nghiệm
 - Đánh giá định lượng
 - Đánh giá định tính
 - Hạn chế
- 5 Kết luận

Các nghiên cứu liên quan

- cGAN: là mô hình đầu tiên giải quyết bài toán text-to-image.
- StackGAN: sinh ra các hình ảnh có độ phân giải cao.
- AttnGAN: sử dụng cơ chế cross-modal attention giúp mô hình sinh ra ảnh với nhiều chi tiết hơn.
- MirrorGAN: khôi phục lại văn bản từ hình ảnh được tạo cho tính nhất quán giữa ngữ nghĩa và hình ảnh.
- SD-GAN: sử dụng cấu trúc Siamese để chắt lọc những điểm chung về ngữ nghĩa từ các văn bản để tạo ra hình ảnh nhất quán.
- DM-GAN: sử dụng memory network để tinh chỉnh hình ảnh mờ khi hình ảnh ban đầu không được tạo tốt trong kiến trúc stack.

Nội dung

- 1 Giới thiệu
- 2 Các nghiên cứu liên quan
- 3 Phương pháp tiếp cận
 - Kiến trúc DF-GAN
 - Tổng quan mô hình
 - Trục chuyển văn bản thành hình ảnh một giai đoạn
 - Công cụ Discriminator nhận biết mục tiêu
 - Hình phạt chuyển màu nhận biết phù hợp (Matching-Aware Gradient Penalty)
 - Đầu ra một chiều (One-Way Output)
 - Kết hợp văn bản - hình ảnh hiệu quả (Efficient TextImage Fusion)
- 4 Thực nghiệm
 - Đánh giá định lượng
 - Đánh giá định tính
 - Hạn chế
- 5 Kết luận

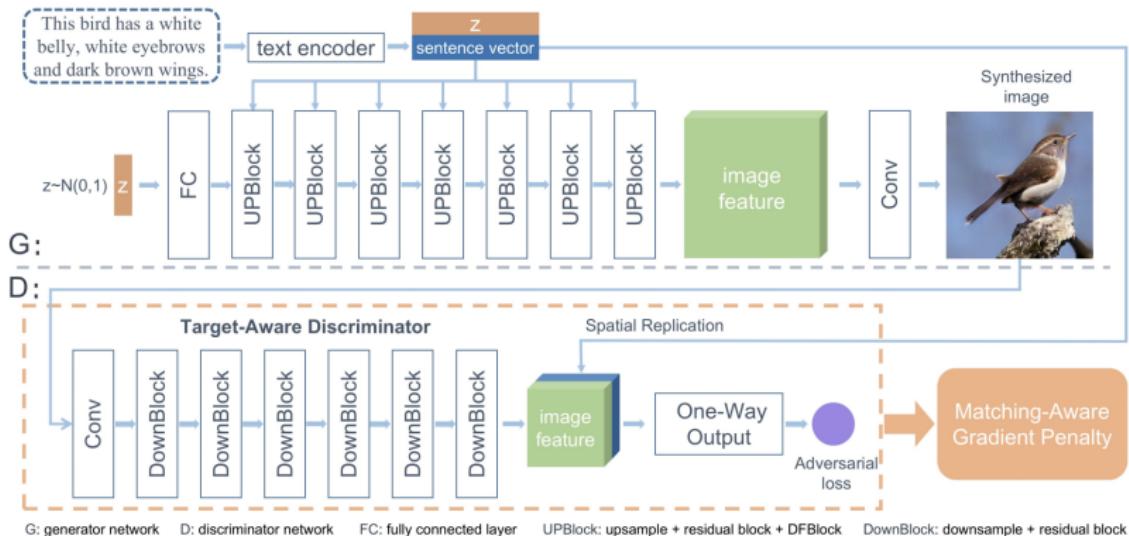
Kiến trúc DF-GAN

- Một đường trục chuyển văn bản thành hình ảnh một giai đoạn có thể tổng hợp trực tiếp các hình ảnh có độ phân giải cao mà không vướng víu về tính năng thực tế.
- Một Discriminator nhận biết mục tiêu mới bao gồm Matching-Aware Gradient Penalty (MA-GP), giúp tăng cường tính nhất quán ngữ nghĩa hình ảnh văn bản mà không cần giới thiệu thêm mạng.
- Một khối kết hợp hình ảnh văn bản sâu mới (DF-Block), kết hợp đầy đủ hơn các tính năng hình ảnh và văn bản.

Tổng quan mô hình

- Bao gồm 1 bộ Generator, một Bộ discriminator và bộ mã hóa văn bản. Trình tạo có hai đầu vào, một vectơ câu được mã hóa bằng bộ mã hóa văn bản và một vectơ nhiễu được lấy mẫu từ phân phối Gaussian.
- Bộ discriminator chuyển đổi hình ảnh thành các tính năng hình ảnh thông qua một loạt DownBlocks. Adversarial loss sẽ được sử dụng để đánh giá tính hiện thực trực quan và tính nhất quán về ngữ nghĩa của đầu vào. Bộ discriminator thúc đẩy trình tạo tổng hợp hình ảnh với chất lượng cao hơn và nhất quán ngữ nghĩa hình ảnh văn bản.
- Bộ mã hóa văn bản là Long Short-Term Memory (LSTM) trích xuất các vectơ ngữ nghĩa từ mô tả văn bản.

Tổng quan mô hình



Hình: Kiến trúc của DF-GAN

Bộ mã hóa văn bản (Text encoder)



Hình: Text Encoder

Trục chuyển văn bản thành hình ảnh một giai đoạn

- Mô hình GAN chuyển văn bản thành hình ảnh trước đây thường sử dụng kiến trúc xếp chồng. Tuy nhiên, kiến trúc xếp chồng gây vướng víu giữa các bộ tạo khác nhau và nó làm cho hình ảnh tinh chỉnh cuối cùng trông giống như sự kết hợp đơn giản giữa hình dạng mờ và một số chi tiết.
- Từ các nghiên cứu gần đây về tạo hình ảnh vô điều kiện, nhóm đề xuất một đường trục chuyển văn bản thành hình ảnh một giai đoạn có thể tổng hợp hình ảnh có độ phân giải cao trực tiếp bằng cách một cặp generator và discriminator.

Trục chuyển văn bản thành hình ảnh một giai đoạn (One-Stage Text-to-Image Backbone)

Sử dụng hinge loss để ổn định quá trình đào tạo đối thủ. Hàm loss:

$$\begin{aligned} L_D = & -\mathbb{E}_{x \sim \mathbb{P}_r} [\min(0, -1 + D(x, e))] \\ & - (1/2)\mathbb{E}_{G(z) \sim \mathbb{P}_g} [\min(0, -1 - D(G(z), e))] \\ & - (1/2)\mathbb{E}_{x \sim \mathbb{P}_{mis}} [\min(0, -1 - D(x, e))] \\ L_G = & -\mathbb{E}_{G(z) \sim \mathbb{P}_g} [D(G(z), e)] \end{aligned}$$

Trong đó:

z : vectơ ngẫu nhiên được lấy mẫu từ phân phối Gaussian.

e : vectơ câu.

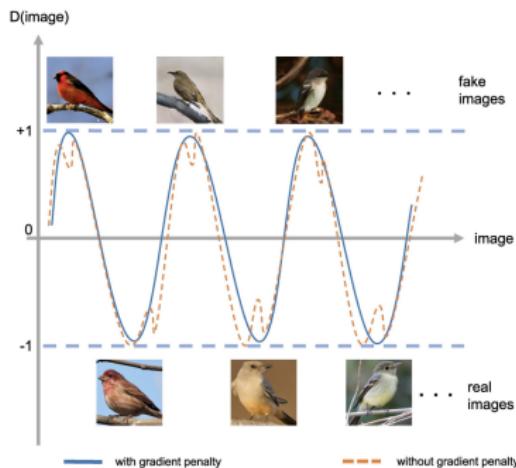
P_g, P_r, P_{mis} : lần lượt biểu thị phân phối dữ liệu tổng hợp, phân phối dữ liệu thực và phân phối dữ liệu không khớp.

Công cụ Discriminator nhận biết mục tiêu

- Hình phạt chuyển màu nhận biết phù hợp (Matching-Aware Gradient Penalty).
- Đầu ra một chiều (One-Way Output).

Hình phạt chuyển màu nhận biết phù hợp (Matching-Aware Gradient Penalty)

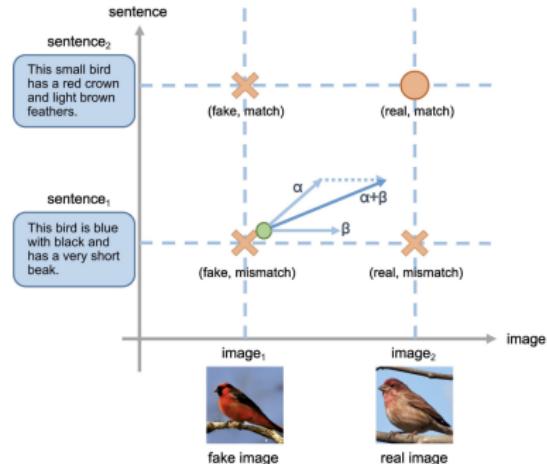
Hình phạt Gradient không tập trung vào Matching-Aware (MA-GP) nhằm để nâng cao hình ảnh văn bản nhất quán ngữ nghĩa.



Hình: So sánh cảnh quan mắt mát trước và sau khi áp dụng hình phạt độ dốc. Hình phạt gradient làm mịn bộ discriminator bề mặt mắt mát hữu ích cho sự hội tụ của bộ tạo.

Hình phạt chuyển màu nhận biết phù hợp (Matching-Aware Gradient Penalty)

Đối với tính nhất quán ngữ nghĩa văn bản-hình ảnh, nhóm có xu hướng áp dụng hình phạt gradient trên dữ liệu thực phù hợp với văn bản, mục tiêu của tổng hợp văn bản thành hình ảnh. Do đó, trong MA-GP, hình phạt gradient nên được áp dụng trên các hình ảnh thực có khớp chữ.



Hình: Một sơ đồ của MA-GP. Điểm dữ liệu (thực, khớp) nên được áp dụng MA-GP.

Hình phạt chuyển màu nhận biết phù hợp (Matching-Aware Gradient Penalty)

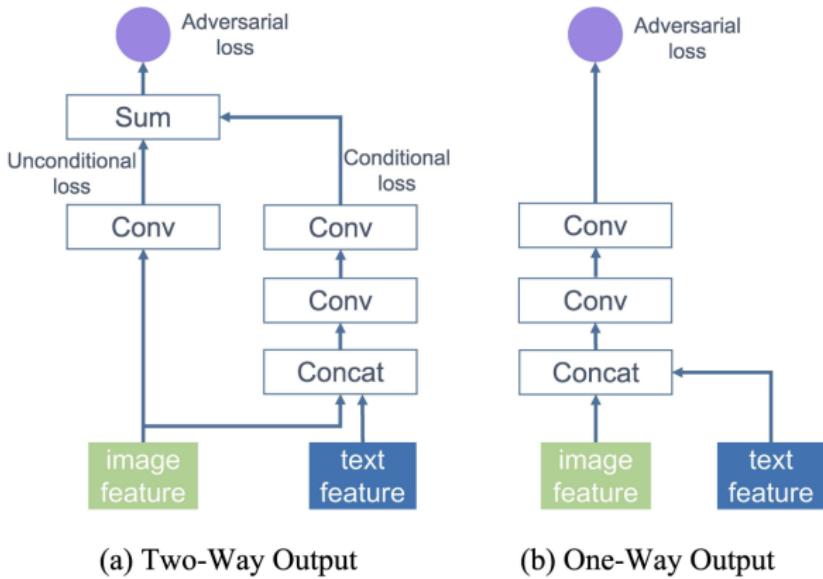
Toàn bộ công thức của mô hình của chúng tôi với MA-GP là như sau:

$$\begin{aligned}
 L_D = & -\mathbb{E}_{x \sim \mathbb{P}_r} [\min(0, -1 + D(x, e))] \\
 & - (1/2)\mathbb{E}_{G(z) \sim \mathbb{P}_g} [\min(0, -1 - D(G(z), e))] \\
 & - (1/2)\mathbb{E}_{x \sim \mathbb{P}_{mis}} [\min(0, -1 - D(x, e))] \\
 & + k\mathbb{E}_{x \sim \mathbb{P}_r} [(\|\nabla_x D(x, e)\| + \|\nabla_e D(x, e)\|)^p] \\
 L_G = & -\mathbb{E}_{G(z) \sim \mathbb{P}_g} [D(G(z), e)]
 \end{aligned}$$

Trong đó:

k và p : hai siêu tham số để cân bằng hiệu quả của hình phạt độ dốc.

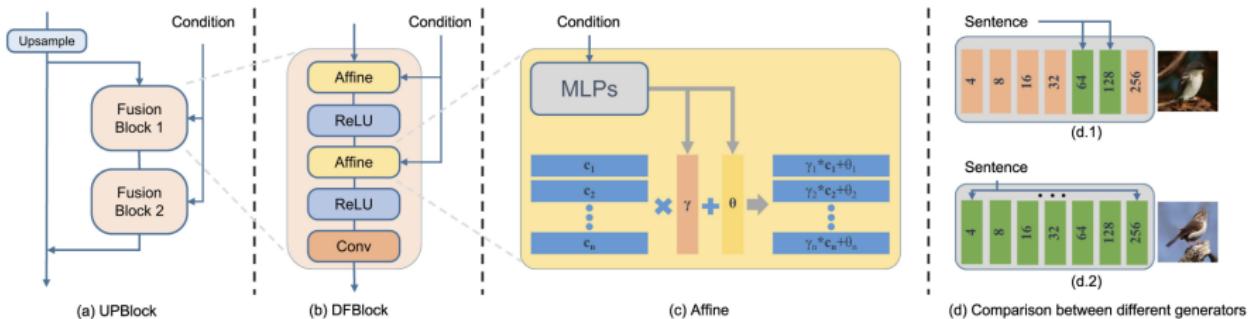
Đầu ra một chiều (One-Way Output)



Hình: So sánh giữa Đầu ra hai chiều và Đầu ra một chiều. (a) Đầu ra hai chiều dự đoán tổn thất có điều kiện và mất mát vô điều kiện và tổng hợp chúng như là đối thủ cuối cùng sự mất mát. (b) Đầu ra một chiều dự đoán toàn bộ tổn thất đối thủ trực tiếp.

Kết hợp văn bản - hình ảnh hiệu quả

Để sử dụng đầy đủ thông tin văn bản trong quá trình hợp nhất, nhóm đề xuất Khối hợp nhất hình ảnh văn bản sâu (DFBlock) ngăn xếp nhiều Biến đổi Affine và Các lớp ReLU trong Fusion Block.



Hình: (a) Một UPBlock điển hình trong mạng generator. UPBlock lấy mẫu các tính năng hình ảnh và hợp nhất các tính năng văn bản và hình ảnh bởi hai Khối Fusion. (b) DFBlock bao gồm hai lớp Affine, hai lớp kích hoạt ReLU và một lớp Convolution. (c) Các minh họa của Biến đổi Affine. (d) So sánh giữa (d.1) trình generator với sự chú ý đa phương thức [50, 60] và (d.2) của generator với DFBlock.

Kết hợp văn bản - hình ảnh hiệu quả

2 lợi ích mà DFBLOCK mang lại cho việc chuyển văn bản thành hình ảnh:

- Làm cho trình generator khai thác đầy đủ hơn thông tin văn bản khi kết hợp các tính năng văn bản và hình ảnh.
- Đào sâu quá trình hợp nhất sẽ mở rộng không gian biểu diễn của mô-đun hợp nhất, giúp tạo ngữ nghĩa hình ảnh nhất quán từ các mô tả văn bản khác nhau.

Nội dung

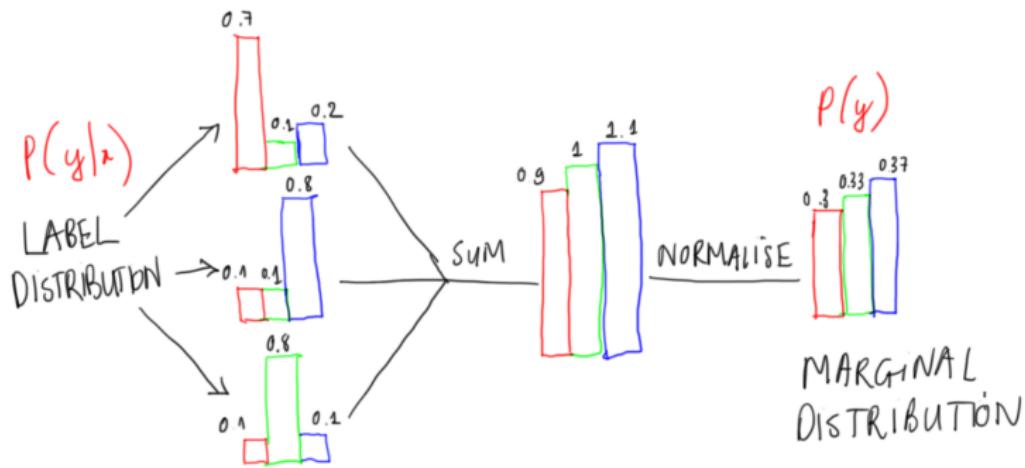
- 1 Giới thiệu
- 2 Các nghiên cứu liên quan
- 3 Phương pháp tiếp cận
 - Kiến trúc DF-GAN
 - Tổng quan mô hình
 - Trục chuyển văn bản thành hình ảnh một giai đoạn
 - Công cụ Discriminator nhận biết mục tiêu
 - Hình phạt chuyển màu nhận biết phù hợp (Matching-Aware Gradient Penalty)
 - Đầu ra một chiều (One-Way Output)
 - Kết hợp văn bản - hình ảnh hiệu quả (Efficient TextImage Fusion)
- 4 Thực nghiệm
 - Đánh giá định lượng
 - Đánh giá định tính
 - Hạn chế
- 5 Kết luận

Thực nghiệm

- Dữ liệu: CUB bird ,chứa 11.788 hình ảnh thuộc về 200 loài chim. Mỗi hình ảnh con chim có 10 câu mô tả.
- Training: tối ưu hóa mạng của mình bằng Adam, learning rate là 0,0001 cho generator và 0,0004 cho discriminator.
- Đánh giá: Inception Score (IS) và Frechet Inception Distance (FID).

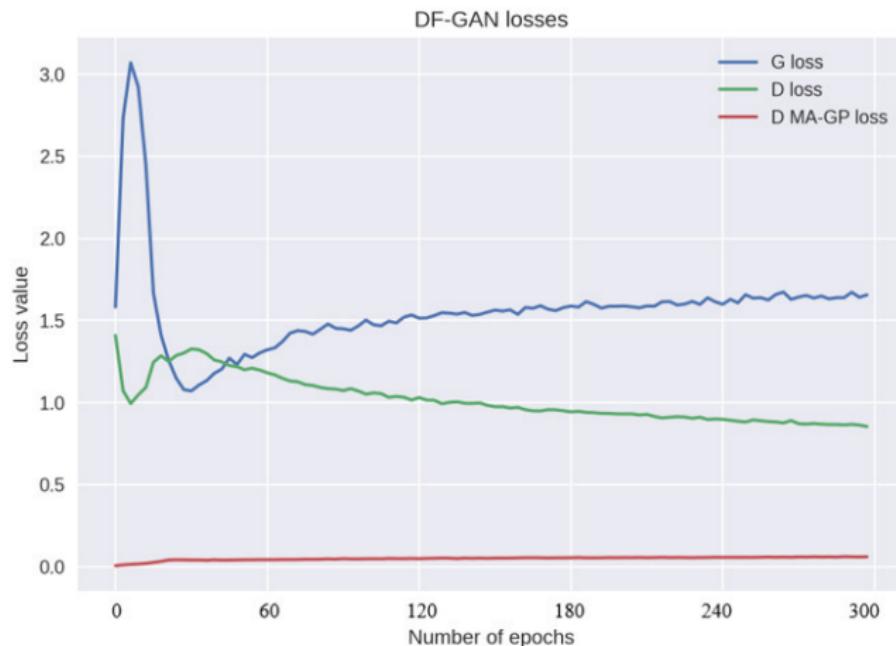
Thực nghiệm

$$D_{KL}(P\|Q) = \int P(x) \log \frac{P(x)}{Q(x)} dx$$



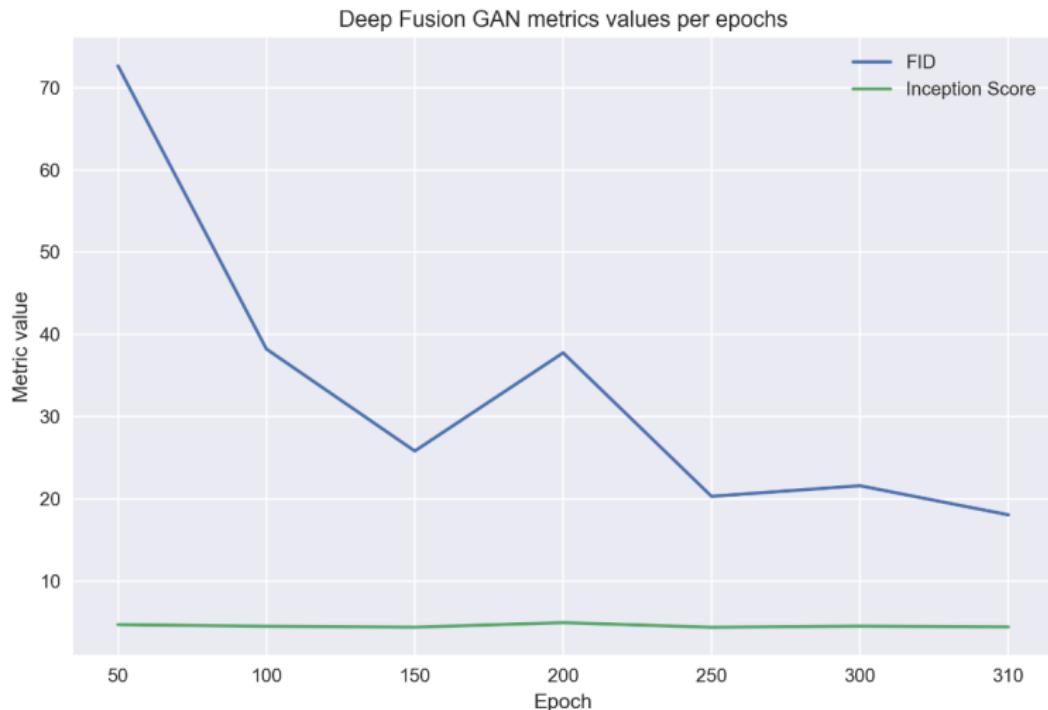
$$FID = \|\mu_x - \mu_g\|_2^2 + Tr(\Sigma_x + \Sigma_g - 2(\Sigma_x \Sigma_g)^{\frac{1}{2}})$$

Thực nghiệm



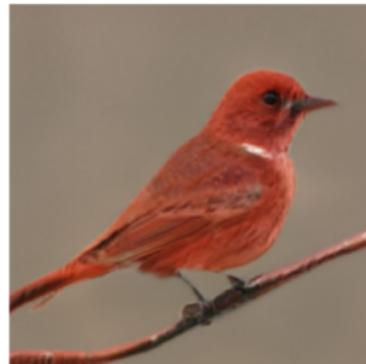
Hình: Loss trên 310 epoch

Thực nghiệm



Hình: Số liệu trên mỗi epoch

Thực nghiệm



(a)



(b)



(c)

Hình: (a) This bird has red pink feathers and dark brown wings. (b) This is a bird purple with brown wings. (c) it is a white bird with blue wings.

Dánh giá định lượng

Bảng: Kết quả IS, FID so với các phương pháp hiện đại trên bộ thử nghiệm của CUB

Model	IS ↑	FID ↓
StackGAN	3.70	-
StackGAN++	3.84	-
AttnGAN	4.36	23.98
MirrorGAN	4.56	18.34
SD-GAN	4.67	-
DM-GAN	4.75	16.09
DAE-GAN	4.42	15.19
TIME	4.91	14.30
DF-GAN (author)	5.10	14.81
DF-GAN (Ours)	4.31	18.27

Dánh giá định tính

A bird with a brown and black wings, red crown and throat and the bill is short and pointed.



This is a white and grey bird with black wings and a black stripe by its eyes.



This bird has a yellow throat, belly, abdomen and sides with lots of brown streaks on them.



This bird has a white belly and breast, with a blue crown and nape.



AttnGAN

DM-GAN

DF-GAN



Hình: Ví dụ về hình ảnh được tổng hợp bởi DF-GAN được đề xuất của nhóm dựa trên mô tả văn bản từ bộ dữ liệu thử nghiệm CUB

Hạn chế

- Chỉ thực hiện văn bản cấp độ câu, làm hạn chế khả năng của tổng hợp tính năng hình ảnh chi tiết.
- Giới thiệu các mô hình ngôn ngữ lớn được đào tạo trước để cung cấp kiến thức bổ sung có thể cải thiện hơn nữa hiệu suất.

Nội dung

- 1 Giới thiệu
- 2 Các nghiên cứu liên quan
- 3 Phương pháp tiếp cận
 - Kiến trúc DF-GAN
 - Tổng quan mô hình
 - Trục chuyển văn bản thành hình ảnh một giai đoạn
 - Công cụ Discriminator nhận biết mục tiêu
 - Hình phạt chuyển màu nhận biết phù hợp (Matching-Aware Gradient Penalty)
 - Đầu ra một chiều (One-Way Output)
 - Kết hợp văn bản - hình ảnh hiệu quả (Efficient TextImage Fusion)
- 4 Thực nghiệm
 - Đánh giá định lượng
 - Đánh giá định tính
 - Hạn chế
- 5 Kết luận

Kết luận

Trong bài báo này, nhóm đề xuất mô hình DF-Gan cho bài toán text-to-image thực hiện One stage để sinh ra hình ảnh có độ phân giải cao trực tiếp mà không có sự vướng víu giữa các mạng Generator khác nhau. Nhóm thực hiện Target-Aware Discriminator composed gồm Matching-Aware Gradient Penalty (MAGP) và One-Way Output. Giới thiệu Deep text-image Fusion Block (DFBlock) kết hợp đầy đủ các tính năng văn bản và hình ảnh một cách hiệu quả và sâu sắc hơn. Các kết quả thử nghiệm mở rộng chứng minh rằng DF-GAN được đề xuất của chúng tôi hoạt động tốt hơn đáng kể so với các mô hình tiên tiến hiện tại trên bộ dữ liệu CUB

Thank you!