

2021

Text-to-Image Survey

汇报人： 汪志伟

时间： 2021.4.28

不同角度对text2image进行分类:

1. 是否多阶段: single-stage、multi-stage
2. 信息融合方式的不同: concatenating、cross-modal attention、Conditional Batch Normalization (CBN)
3. 采用的损失函数: pair-wise ranking loss、CE-based reconstruction loss、
4. 是否需要引入额外的网络: SD-GAN、DAMSM loss、Mirror-GAN



01

Concatenating

02

Cross-modal Attention

03

CBN

Conditional Batch Normalization

$$\text{BN}(x) = \gamma \cdot \frac{x - \mu(x)}{\sigma(x)} + \beta,$$

$$\text{BN}(x|c) = (\gamma + \gamma_c) \cdot \frac{x - \mu(x)}{\sigma(x)} + (\beta + \beta_c).$$

Evaluation Metric

IS:

$$IS(G) = \exp(\mathbb{E}_{\mathbf{x} \sim p_g} D_{KL}(p(y|\mathbf{x}) || p(y)))$$

原理:

基于一个假设：一个清晰、高质量的图像，它被分类正确的概率就非常大。利用Inception V3 network, 得到生成图片相应的评分。

缺点:

1. 没有考虑真实图像，无法反映生成图像和真实图像之间是否相近。
2. 使用Inception network，不同框架的内置网络可能有些许不同，则会导致IS值出现偏差。

FID:

$$d^2((\mu_r, \Sigma_r), (\mu_g, \Sigma_g)) = \|\mu_r - \mu_g\|^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{\frac{1}{2}})$$

原理:

首先分别选取真实图像和生成图像各N张，计算得到的特征向量有n*n维，之后分别计算这N个样本对应的均值向量 μ 和协方差矩阵 Σ 。

缺点:

1. FID与IS都是基于特征的方法，不能很好区分指标的好坏是由生成图像质量问题还是多样性问题造成。
2. 不能描述特征的空间关系。

Dataset:



Black_Footed_A
lbatross_0031_
100.jpg



Black_Footed_A
lbatross_0032_
796115.jpg



Black_Footed_A
lbatross_0033_
796086.jpg



Black_Footed_A
lbatross_0053_
796109.jpg



Black_Footed_A
lbatross_0056_
796078.jpg



Black_Footed_A
lbatross_0057_
796106.jpg



COCO_train201
4_00000000014
9.jpg



COCO_train201
4_00000000015
1.jpg



COCO_train201
4_00000000015
4.jpg



COCO_train201
4_00000000032
6.jpg



COCO_train201
4_00000000033
2.jpg



COCO_train201
4_00000000034
9.jpg

Dataset	CUB [28]		COCO [14]	
	train	test	train	test
#samples	8,855	2,933	80k	40k
caption/image	10	10	5	5

01

Part one

concatenating

Generative Adversarial Text to Image Synthesis

Scott Reed, Zeynep Akata, Xincheng Yan, Lajanugen Logeswaran
Bernt Schiele, Honglak Lee

REEDSCOT¹, AKATA², XCYAN¹, LLAJAN¹
SCHIELE², HONGLAK¹

¹ University of Michigan, Ann Arbor, MI, USA (UMICH.EDU)

² Max Planck Institute for Informatics, Saarbrücken, Germany (MPI-INF.MPG.DE)

ICML 2016

动机

- GAN网络在生成任务上取得巨大成功，可以生成高质量图像，如人脸、唱片封面等。
- 然而在跨模态领域的生成任务挑战巨大，作者首次提出该项任务----Text2Image。

方法

1. Matching-aware discriminator

Algorithm 1 GAN-CLS training algorithm with step size α , using minibatch SGD for simplicity.

```
1: Input: minibatch images  $x$ , matching text  $t$ , mis-  
   matching  $\hat{t}$ , number of training batch steps  $S$   
2: for  $n = 1$  to  $S$  do  
3:    $h \leftarrow \varphi(t)$  {Encode matching text description}  
4:    $\hat{h} \leftarrow \varphi(\hat{t})$  {Encode mis-matching text description}  
5:    $z \sim \mathcal{N}(0, 1)^Z$  {Draw sample of random noise}  
6:    $\hat{x} \leftarrow G(z, h)$  {Forward through generator}  
7:    $s_r \leftarrow D(x, h)$  {real image, right text}  
8:    $s_w \leftarrow D(x, \hat{h})$  {real image, wrong text}  
9:    $s_f \leftarrow D(\hat{x}, h)$  {fake image, right text}  
10:   $\mathcal{L}_D \leftarrow \log(s_r) + (\log(1 - s_w) + \log(1 - s_f))/2$   
11:   $D \leftarrow D - \alpha \partial \mathcal{L}_D / \partial D$  {Update discriminator}  
12:   $\mathcal{L}_G \leftarrow \log(s_f)$   
13:   $G \leftarrow G - \alpha \partial \mathcal{L}_G / \partial G$  {Update generator}  
14: end for
```

2. Learning with manifold interpolation

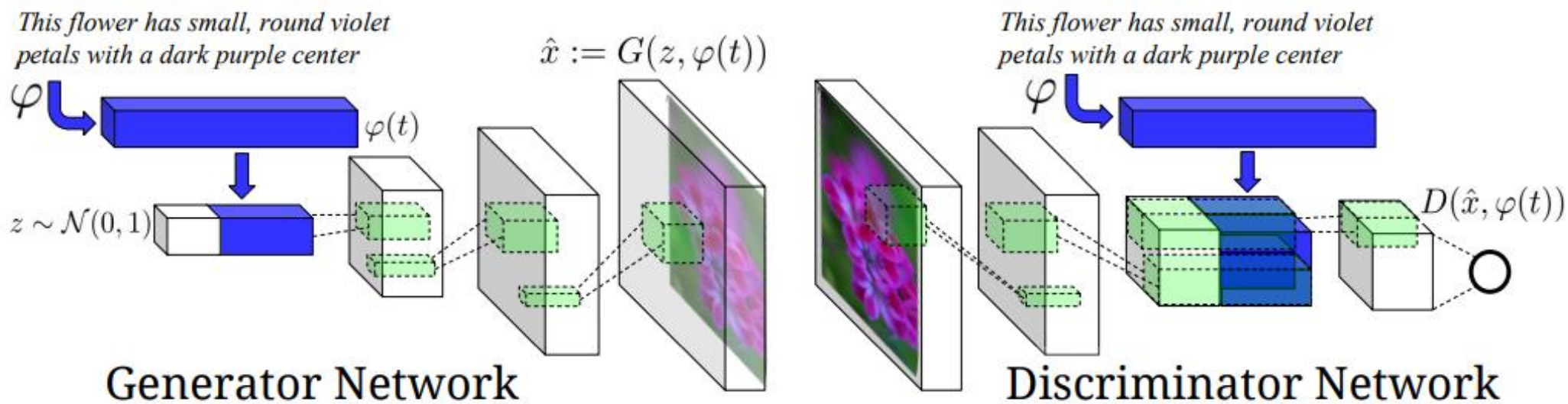
$$\mathbb{E}_{t_1, t_2 \sim p_{data}} [\log(1 - D(G(z, \beta t_1 + (1 - \beta)t_2)))]$$

3. Inverting the generator for style transfer

$$\mathcal{L}_{style} = \mathbb{E}_{t, z \sim \mathcal{N}(0, 1)} \|z - S(G(z, \varphi(t)))\|_2^2$$

模型

Generative Adversarial Text to Image Synthesis



总结

1. 首次提出跨模态生成任务，并且也获得了一定的成功。
2. 该论文提出了sentence interpolation技巧作为数据增强，来获取更好效果。
3. 该论文尝试将图像风格和内容分离，相当于一个正则项，使得生成图片具有更真实的背景、姿态特征。

StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks

Han Zhang¹, Tao Xu², Hongsheng Li³,
Shaoting Zhang⁴, Xiaogang Wang³, Xiaolei Huang², Dimitris Metaxas¹

¹Rutgers University ²Lehigh University ³The Chinese University of Hong Kong ⁴Baidu Research
{han.zhang, dnm}@cs.rutgers.edu, {tax313, xih206}@lehigh.edu
{hsli, xgwang}@ee.cuhk.edu.hk, zhangshaoting@baidu.com

ICCV2017

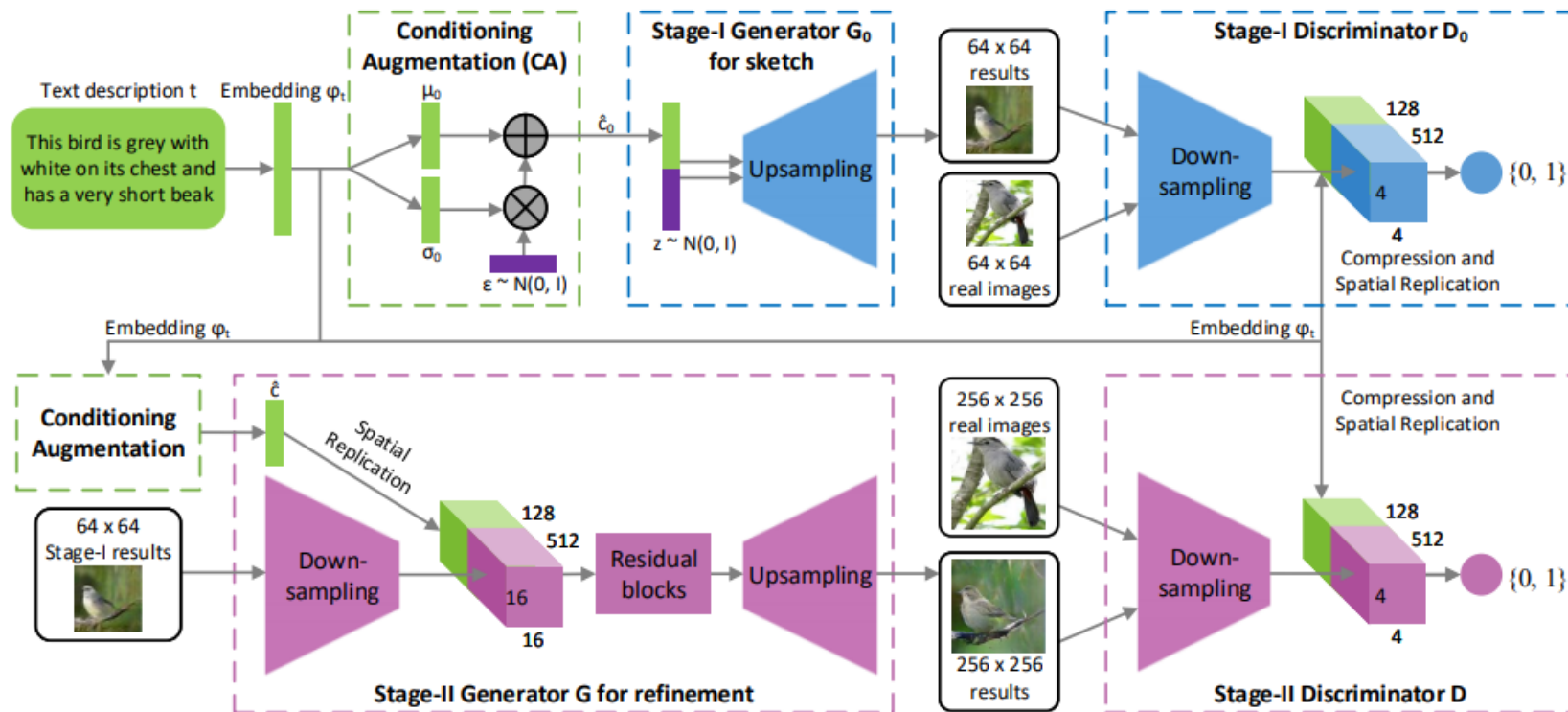
动机

- 之前的工作生成的图像，没能包含必须的细节部分，同时也没办法生成高质量、生动的物体。
- 根据文本信息，生成高分辨率的图像是一件很困难的工作。

方法

1. 阶段一根据给出的文本信息，粗略地生成原始物体的形状和颜色。
2. 阶段二将阶段一和文本信息作为输入，修改初始图像的缺陷，并为初始图像添加细节，从而生成更加真实，分辨率更高的图像。
3. 阶段三为了让生成多样化和训练稳定，使用了Conditioning Augmentation技巧。

模型



$$\mathcal{L}_D = \mathbb{E}_{(I, t) \sim p_{data}} [\log D(I, \varphi_t)] + \mathbb{E}_{s_0 \sim p_{G_0}, t \sim p_{data}} [\log(1 - D(G(s_0, \hat{c}), \varphi_t))],$$

$$\mathcal{L}_G = \mathbb{E}_{s_0 \sim p_{G_0}, t \sim p_{data}} [\log(1 - D(G(s_0, \hat{c}), \varphi_t))] + \lambda D_{KL}(\mathcal{N}(\mu(\varphi_t), \Sigma(\varphi_t)) \parallel \mathcal{N}(0, I)),$$

总结

将复杂的生成问题，分解成多个阶段，逐步生成更具细节、更高分辨率的图像。

提出CA技术，提高训练的稳定性，和生成图片的多样性。

这两项贡献，这之后的text to image任务中，被广泛采用。

StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, *Senior Member, IEEE*,
Xiaogang Wang, *Member, IEEE*, Xiaolei Huang, *Member, IEEE*, Dimitris N. Metaxas*, *Fellow, IEEE*

TPAMI

动机

- 1.提高生成图片的质量。
- 2.使得训练过程更加稳定。
- 3.在一定程度上解决mode collapse问题。

方法

1. JCU (Joint Conditional and Unconditional)

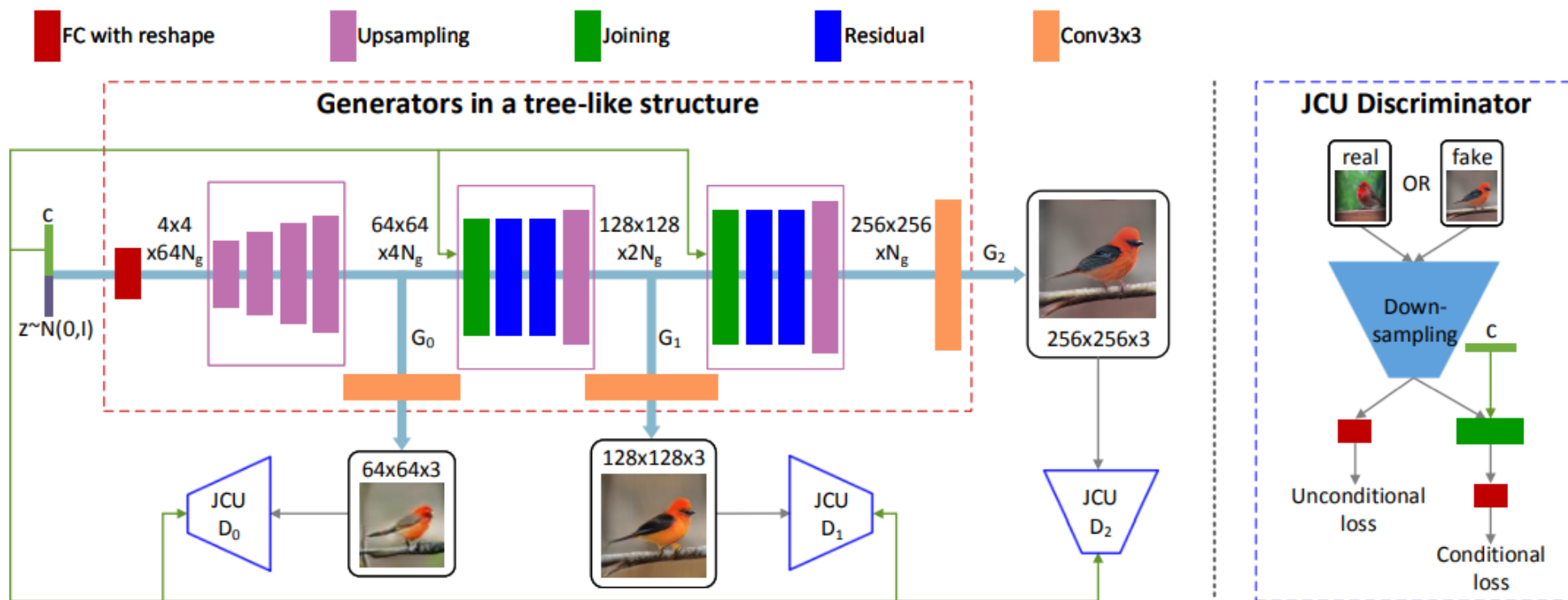
$$\mathcal{L}_{D_i} = \underbrace{-\mathbb{E}_{x_i \sim p_{data_i}} [\log D_i(x_i)] - \mathbb{E}_{s_i \sim p_{G_i}} [\log(1 - D_i(s_i))]}_{\text{unconditional loss}} + \underbrace{-\mathbb{E}_{x_i \sim p_{data_i}} [\log D_i(x_i, c)] - \mathbb{E}_{s_i \sim p_{G_i}} [\log(1 - D_i(s_i, c))]}_{\text{conditional loss}}$$

$$\mathcal{L}_{G_i} = \underbrace{-\mathbb{E}_{s_i \sim p_{G_i}} [\log D_i(s_i)]}_{\text{unconditional loss}} + \underbrace{-\mathbb{E}_{s_i \sim p_{G_i}} [\log D_i(s_i, c)]}_{\text{conditional loss}}$$

2. Color-consistency regularization

$$\mathcal{L}_{C_i} = \frac{1}{n} \sum_{j=1}^n \left(\lambda_1 \|\boldsymbol{\mu}_{s_i^j} - \boldsymbol{\mu}_{s_{i-1}^j}\|_2^2 + \lambda_2 \|\boldsymbol{\Sigma}_{s_i^j} - \boldsymbol{\Sigma}_{s_{i-1}^j}\|_F^2 \right)$$

模型



结论

相对于第一版本的Stack-GAN,该论文进行了进一步改进。

(1) 将多个生成器组织成树状结构，每个中间节点代表中间生成图片，叶子节点代表最终生成结果。

(2) 提出色彩正则项，在多阶段生成任务中，这一点倒是感觉合理，不同阶段图片应该保持风格、颜色的一致。

问题在于：

(1) 将多个生成器组织成树状结构，是否会让整个训练过程变得很慢。

(2) 为什么只提出色彩一致性，不研究其他角度的问题，如风格、姿态等。

02
Part two

Cross-modal attention

AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks

Tao Xu^{*1}, Pengchuan Zhang², Qiuyuan Huang²,
Han Zhang³, Zhe Gan⁴, Xiaolei Huang¹, Xiaodong He²

¹Lehigh University ²Microsoft Research ³Rutgers University ⁴Duke University
{tax313, xih206}@lehigh.edu, {penzhan, qihua, xiaohe}@microsoft.com
han.zhang@cs.rutgers.edu, zhe.gan@duke.edu

CVPR2018

动机

之前的工作都只编码了global sentence embedding，这个方法缺失了很多细粒度的词汇信息，从而生成不了高质量的结果。

方法

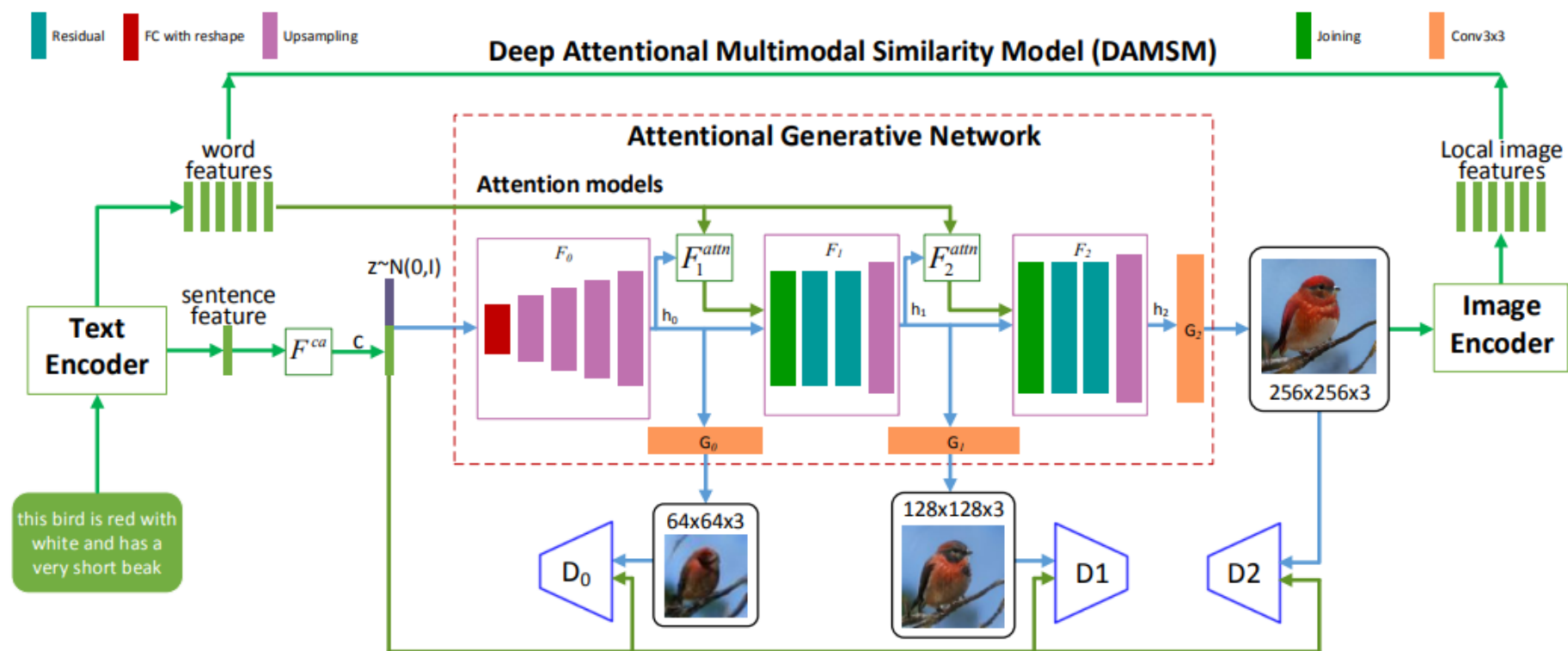
For fine-grained text-to-image generation:

1. attention-driven
2. Multi-stage refinement

Two novel components:

1. Attentional generative network
2. DAMSM

模型



实验

相关公式:

$$\begin{aligned} h_0 &= F_0(z, F^{ca}(\bar{e})); \\ h_i &= F_i(h_{i-1}, F_i^{attn}(e, h_{i-1})) \text{ for } i = 1, 2, \dots, m-1; \\ \hat{x}_i &= G_i(h_i). \end{aligned} \tag{1}$$

$$\mathcal{L} = \mathcal{L}_G + \lambda \mathcal{L}_{DAMSM}, \text{ where } \mathcal{L}_G = \sum_{i=0}^{m-1} \mathcal{L}_{G_i}. \tag{3}$$

$$\mathcal{L}_{G_i} = \underbrace{-\frac{1}{2} \mathbb{E}_{\hat{x}_i \sim p_{G_i}} [\log(D_i(\hat{x}_i))]}_{\text{unconditional loss}} \underbrace{-\frac{1}{2} \mathbb{E}_{\hat{x}_i \sim p_{G_i}} [\log(D_i(\hat{x}_i, \bar{e}))]}_{\text{conditional loss}}, \tag{4}$$

$$\begin{aligned} \mathcal{L}_{D_i} &= \underbrace{-\frac{1}{2} \mathbb{E}_{x_i \sim p_{data_i}} [\log D_i(x_i)] - \frac{1}{2} \mathbb{E}_{\hat{x}_i \sim p_{G_i}} [\log(1 - D_i(\hat{x}_i))]}_{\text{unconditional loss}} + \\ &\quad \underbrace{-\frac{1}{2} \mathbb{E}_{x_i \sim p_{data_i}} [\log D_i(x_i, \bar{e})] - \frac{1}{2} \mathbb{E}_{\hat{x}_i \sim p_{G_i}} [\log(1 - D_i(\hat{x}_i, \bar{e}))]}_{\text{conditional loss}}, \end{aligned} \tag{5}$$

DAMSM (Deep Attentional Multimodal Similarity Model)

采用LSTM作为文本编码器
采用CNN作为图像编码器

$$P(D_i|Q_i) = \frac{\exp(\gamma_3 R(Q_i, D_i))}{\sum_{j=1}^M \exp(\gamma_3 R(Q_i, D_j))}, \quad (11)$$

$$\mathcal{L}_1^w = - \sum_{i=1}^M \log P(D_i|Q_i), \quad (12)$$

DAMSM loss:

$$\mathcal{L}_2^w = - \sum_{i=1}^M \log P(Q_i|D_i), \quad (13)$$

$$\mathcal{L}_{DAMSM} = \mathcal{L}_1^w + \mathcal{L}_2^w + \mathcal{L}_1^s + \mathcal{L}_2^s. \quad (14)$$

总结

- 1.注意到句子嵌入含有的信息具有噪声，比较粗糙。引入了词嵌入这种细粒度的信息。
- 2.Attention机制的出现，将细粒度的信息与图像不同区域联系起来，使得生成的图片细节更多。
- 3.提出DAMSM模块，使得image-text能够得到更好的匹配，使得matching-loss更小。

DM-GAN: Dynamic Memory Generative Adversarial Networks for Text-to-Image Synthesis

Minfeng Zhu^{1,3*} Pingbo Pan³ Wei Chen¹ Yi Yang^{2,3†}

¹ State Key Lab of CAD&CG, Zhejiang University ² Baidu Research

³ Centre for Artificial Intelligence, University of Technology Sydney

{minfeng.zhu@, chenwei@cad}zju.edu.cn {pingbo.pan@student, Yi.Yang@}uts.edu.au

CVPR2019

动机

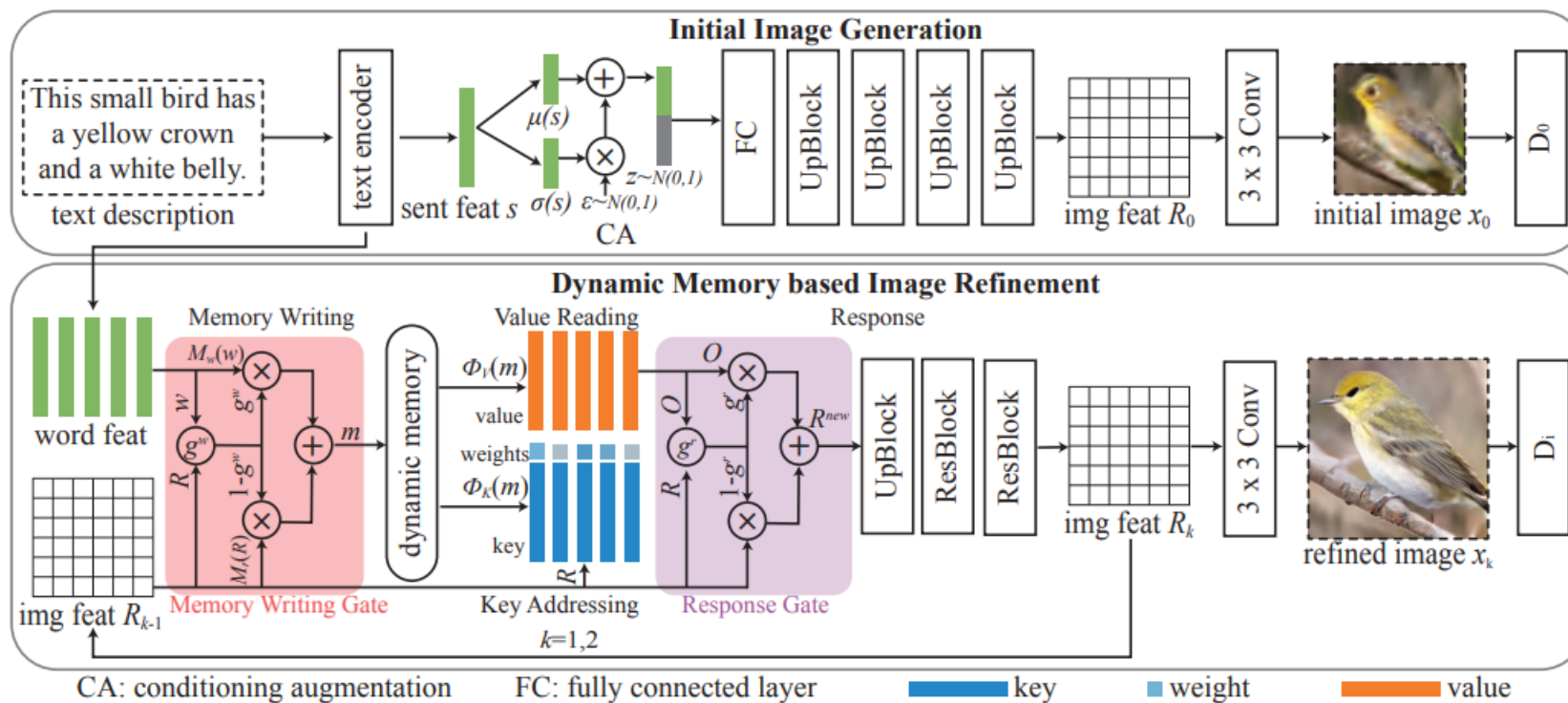
- 多阶段生成的模型，非常依赖最初始的生成图像质量。如果一开始质量不好，后面的阶段很难弥补。
- 每个词对图像生成都有着不同的贡献，而在整个精化过程中，文本嵌入都是不变的。

方法

1. Dynamic Memory GAN
2. Gate memory writing
3. Gate memory response

1. 解决第一个问题，即使初始图片也可以生成好的结果。
2. 能根据初始图片，选取相关的词语。
3. 将记忆单元里的信息和图像信息进行适应性地融合。

模型



$$r_i^{new} = [o_i, r_i],$$

$$g_i^r = \sigma(W[o_i, r_i] + b),$$

$$r_i^{new} = o_i * g_i^r + r_i * (1 - g_i^r),$$

总结

1. 引入了dynamic memory来精化初始的图像
2. 改进了memory writing gate来获取更加重要的文本信息
3. 改进了response gate来更好融合图片和记忆表示

MirrorGAN: Learning Text-to-image Generation by Redescription

Tingting Qiao^{1,3}, Jing Zhang^{2,3,*}, Duanqing Xu^{1,*}, and Dacheng Tao³

¹College of Computer Science and Technology, Zhejiang University, China

²The institute of System Science and Control Engineering, Hangzhou Dianzi University, China

³UBTECH Sydney AI Centre, School of Computer Science, FEIT, University of Sydney, Australia

qiaott@zju.edu.cn, jing.zhang@uts.edu.au, xdq@zju.edu.cn, dacheng.tao@sydney.edu.au

CVPR2019

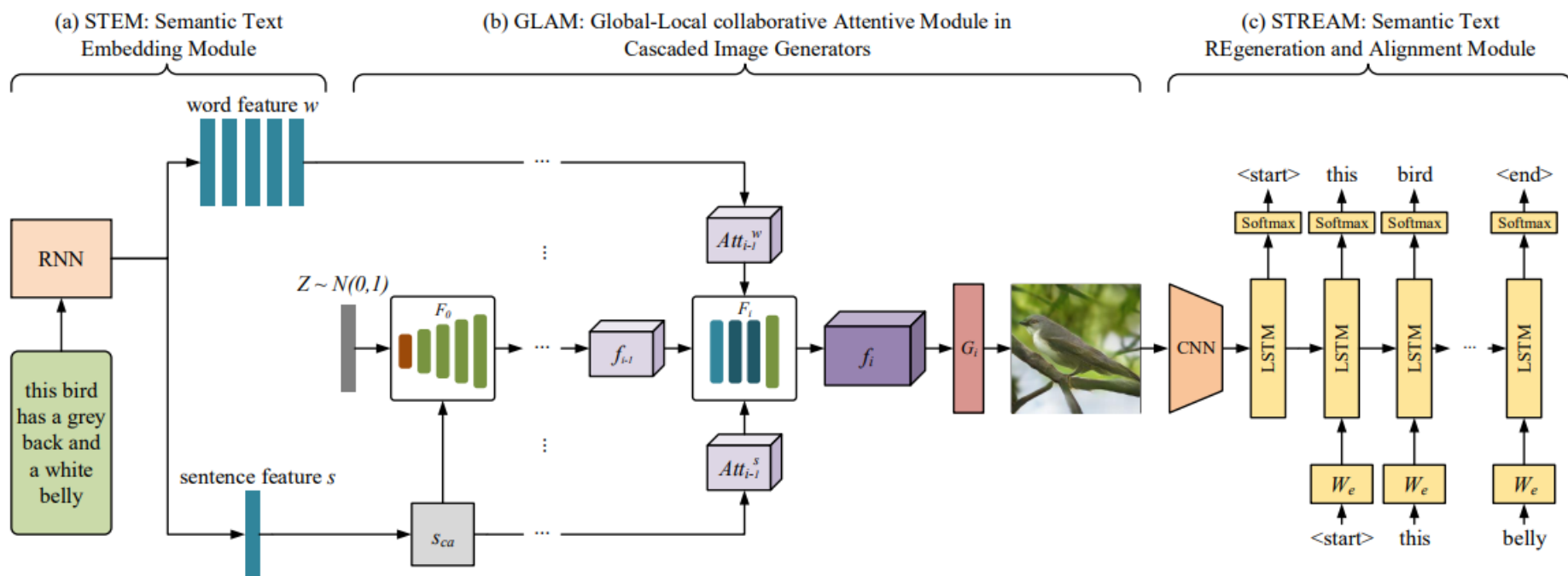
动机

提高文本描述和生成图片之间的语义一致性

Semantic-preserving text-to-image-to-text framework

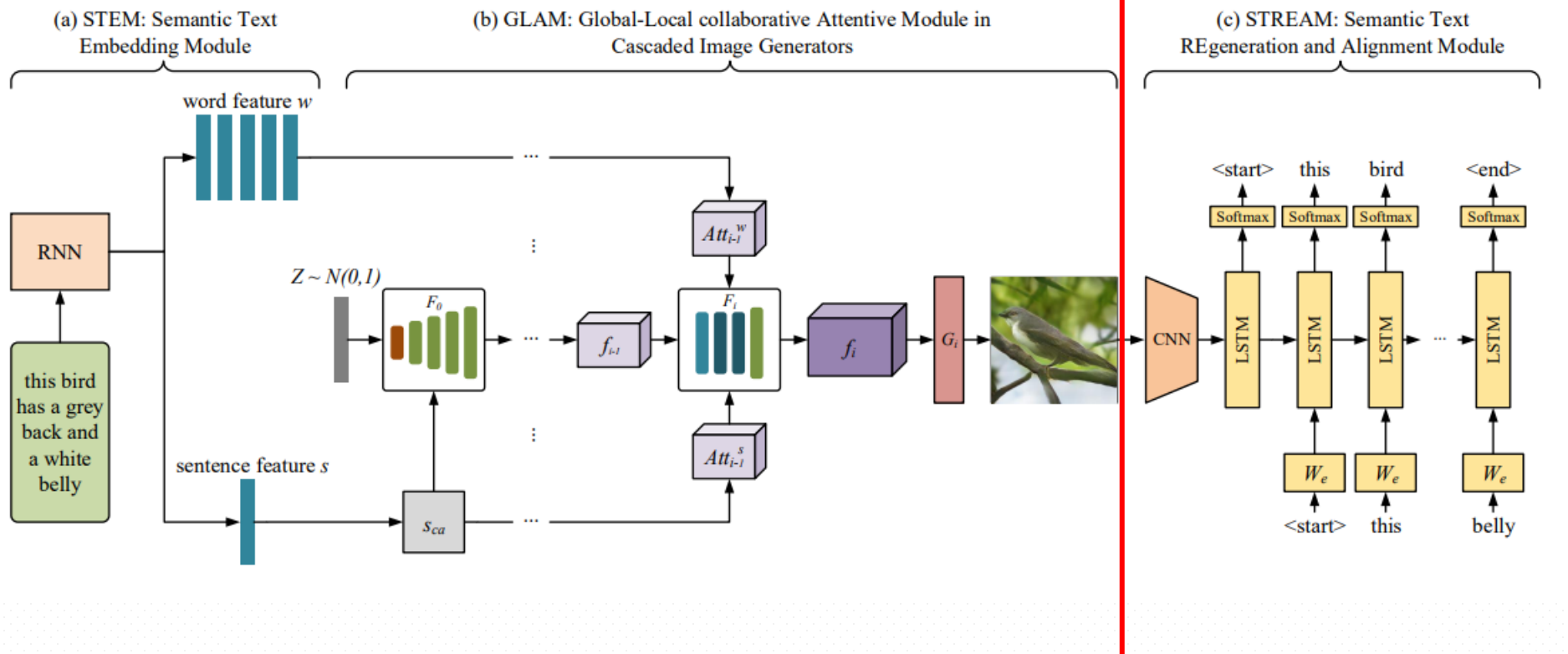
- (1) a semantic text embedding module (STEM)
- (2) a global-local collaborative attentive module (GLAM)
- (3) a semantic text regeneration and alignment module (STREAM)

模型



$$\begin{aligned}
 f_0 &= F_0(z, s_{ca}), \\
 f_i &= F_i(f_{i-1}, F_{att_i}(f_{i-1}, w, s_{ca})), i \in \{1, 2, \dots, m-1\}, \\
 I_i &= G_i(f_i), i \in \{0, 1, 2, \dots, m-1\},
 \end{aligned} \tag{3}$$

STREAM



$$x_{-1} = CNN(I_{m-1}),$$

$$x_t = W_e T_t, t \in \{0, \dots, L-1\},$$

$$p_{t+1} = RNN(x_t), t \in \{0, \dots, L-1\},$$

$$\mathcal{L}_{stream} = - \sum_{t=0}^{L-1} \log p_t(T_t).$$

结论

该论文提出的三个模块：

1. STEM和GLAM并没有什么很新的创新，都是这个任务中最近常用的一些方法。
2. STREAM模块是这篇论文最大的创新点。通过text-image-text这种方式，使得训练更加稳定，语义更加一致。

03
Part three

CBN

Semantics Disentangling for Text-to-Image Generation

Guojun Yin^{1,2}, Bin Liu¹, Lu Sheng^{2,4*}, Nenghai Yu¹, Xiaogang Wang², Jing Shao³

¹University of Science and Technology of China, Key Laboratory of Electromagnetic Space Information,
The Chinese Academy of Sciences, ²CUHK-SenseTime Joint Lab, The Chinese University of Hong Kong

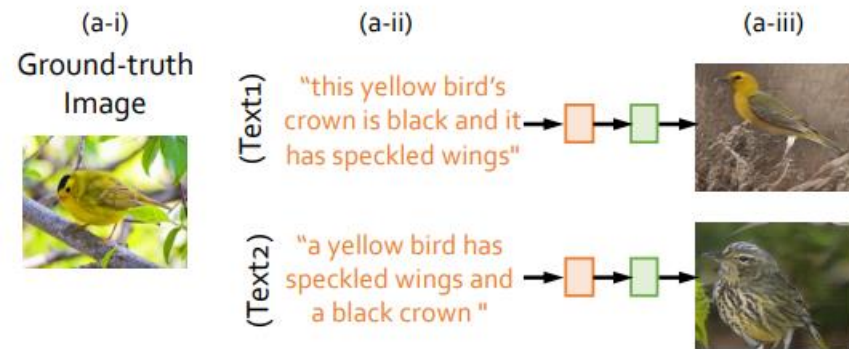
³SenseTime Research, ⁴College of Software, Beihang University

gjyin@mail.ustc.edu.cn, lsheng@buaa.edu.cn,

{flowice, ynh}@ustc.edu.cn, xgwang@ee.cuhk.edu.hk, shaojing@sensetime.com

CVPR2019

动机

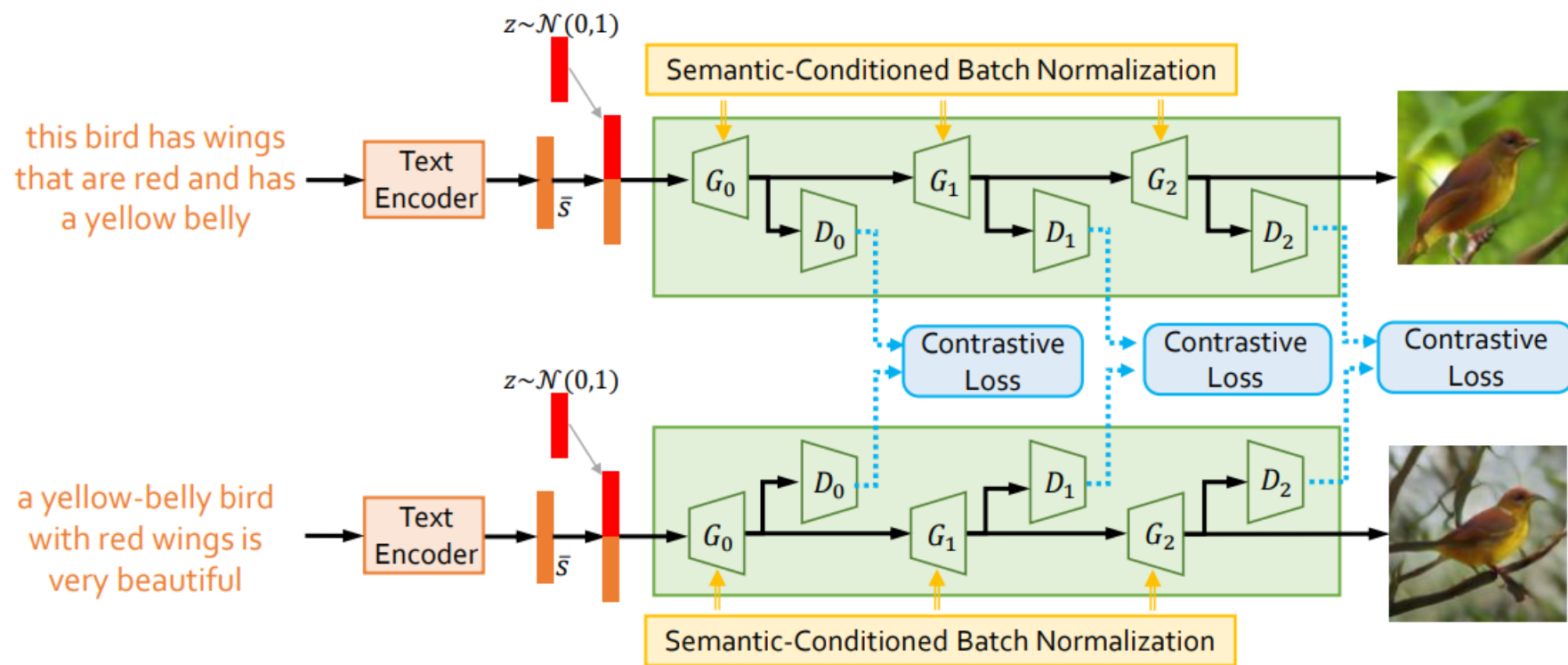


一个重要的现象：描述同一个物体的文本具有很大的主观性和多样性。这样的现象对于生成图文一致的任务具有很大挑战。

方法

1. Distill Semantic Commons from Text
2. Retain Semantic Diversities & Details from Text

模型



结论

- 1.具有不同表达的但具有相似语义的文本描述，在之前的模型中没有取得很好的效果。该模型针对这一现象，引入Siamese 网络，来解决这一问题。
- 2.因为Siamese 网络会损失语义的多样性，又引入SCBN模块来增强语义多样性，使得生成结果更加多样。

DF-GAN: Deep Fusion Generative Adversarial Networks for Text-to-Image Synthesis

Ming Tao, Hao Tang, Songsong Wu, Nicu Sebe, *Senior Member, IEEE*, Fei Wu,
and Xiao-Yuan Jing, *Member, IEEE*,

IEEE Transaction on multimedia

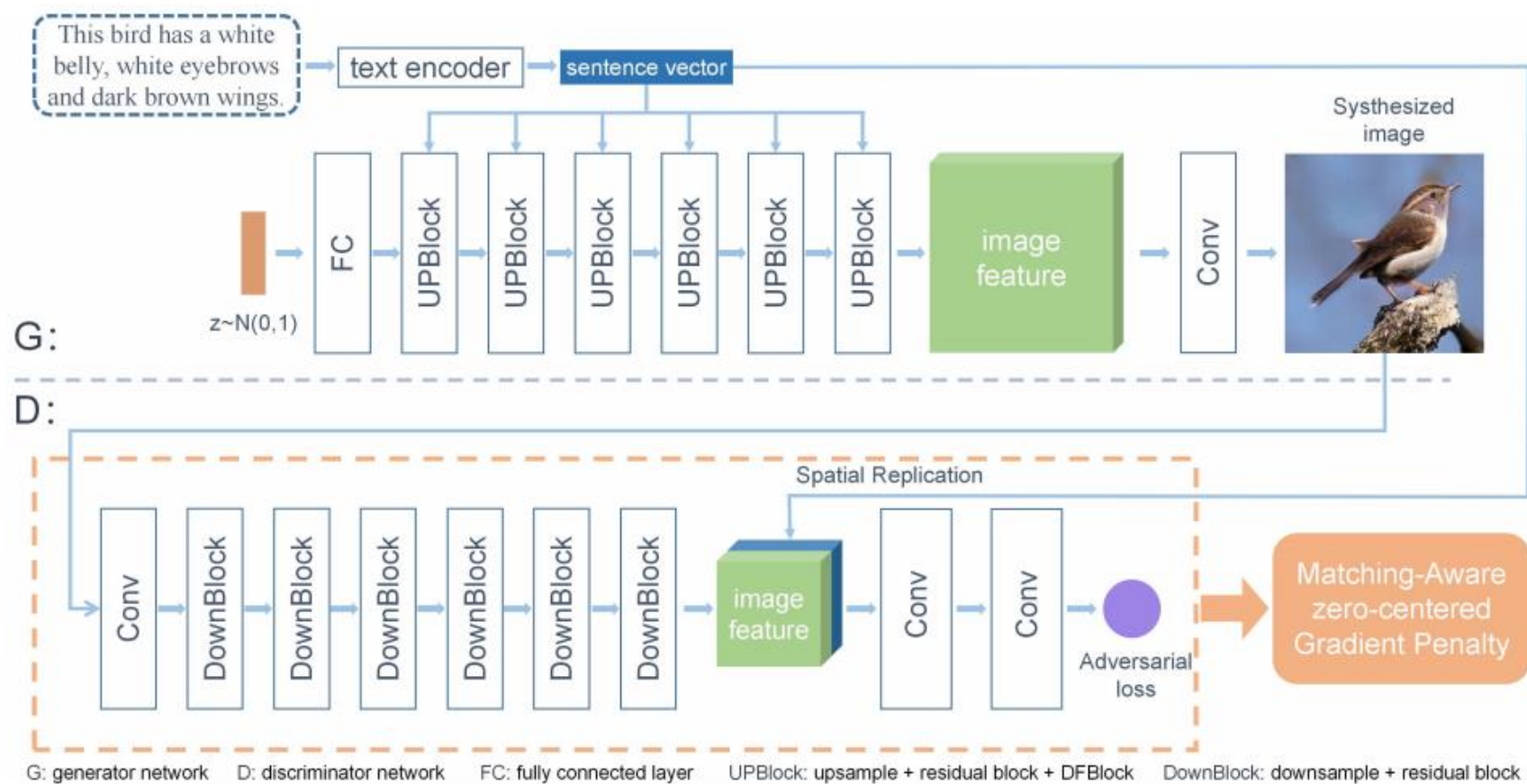
动机

- 多层次的生成器和判别器，让训练过程变得很慢、很低效
- 很多模型引入额外的网络来加强cross-modal之间的语义一致性
- Cross-modal attention仅仅在生成阶段应用几次，这样不能更深层地将文本和图像信息融合起来。

方法

1. 使用单个生成器和判别器，简化了模型的主干网络。
2. 使用了一种MA-GP的正则化方法，来获取更真实的图像生成效果、获得更加符合文本信息的图像。
3. 使用one-way discriminator来辅助该正则方法快速收敛。
4. 提出DFB结构，使得生成阶段能更有效地利用文本的语义特征。

模型



MA-GP

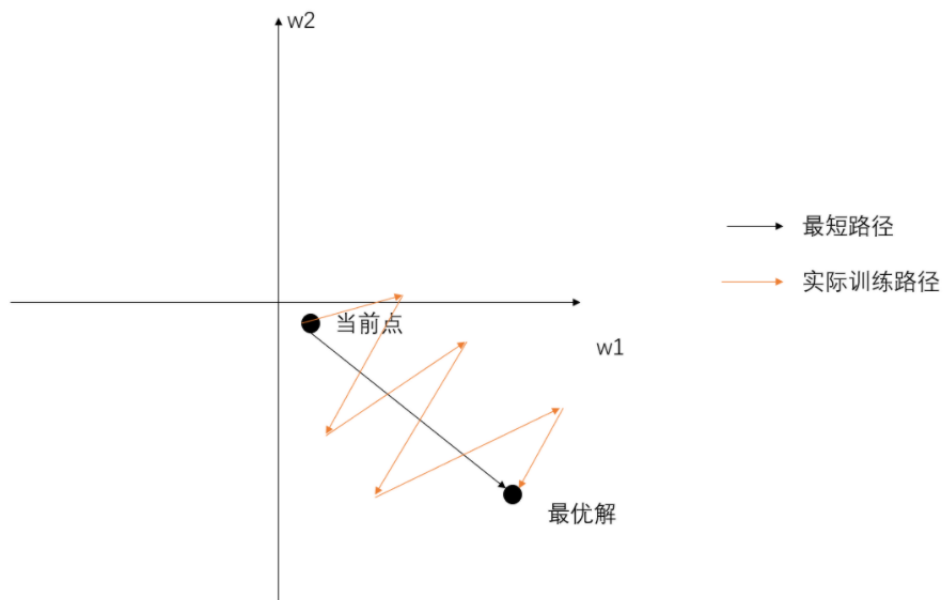
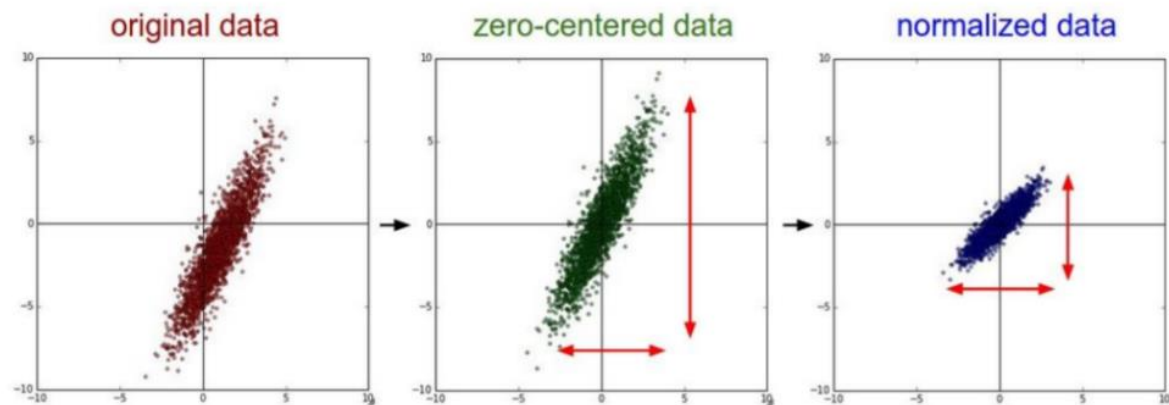


图3 Z型更新



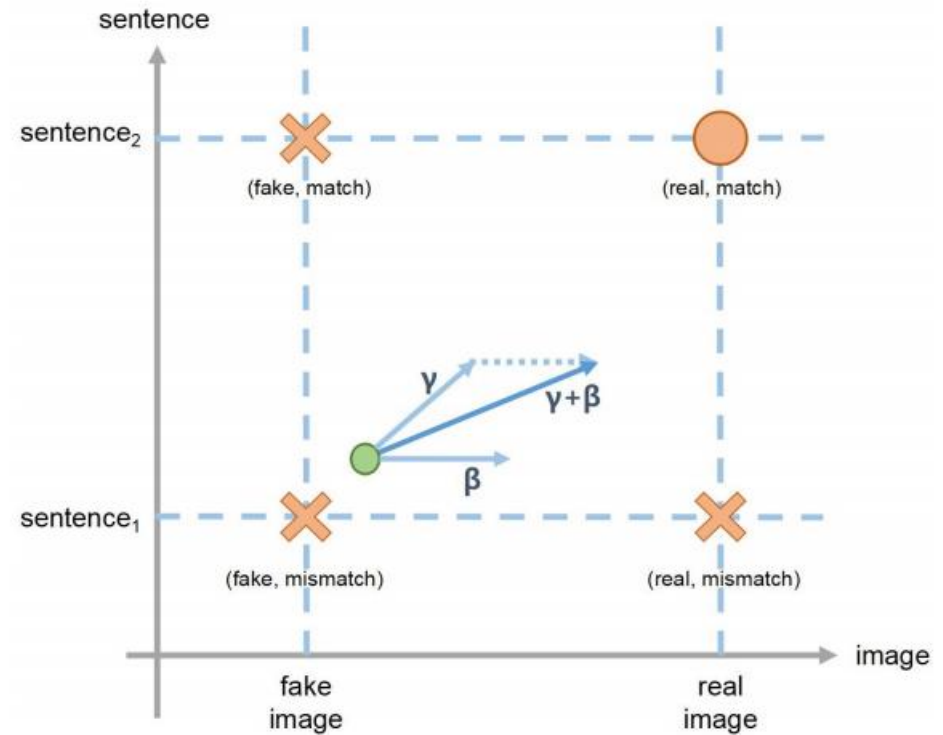
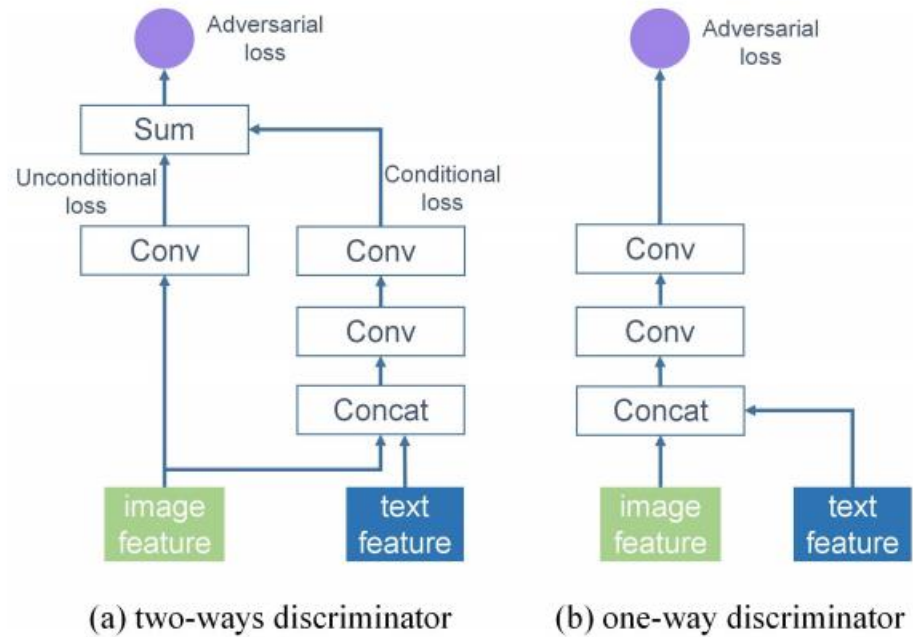
<https://blog.csdn.net/wtmash>

$$L_{0-GP} = k \mathbb{E}_{x \sim \mathbb{P}_r} [\|\nabla_x D(x)\|^p],$$

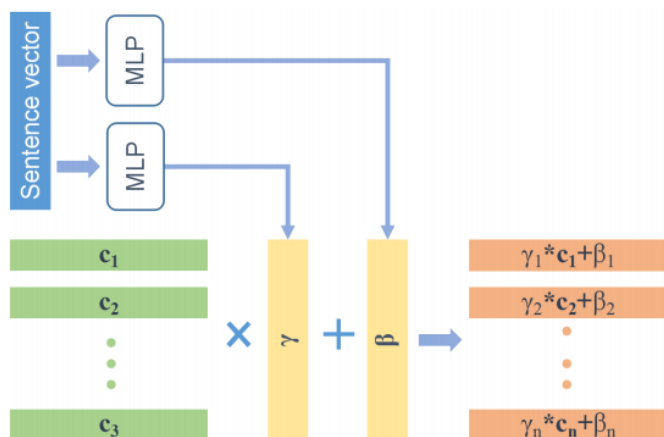
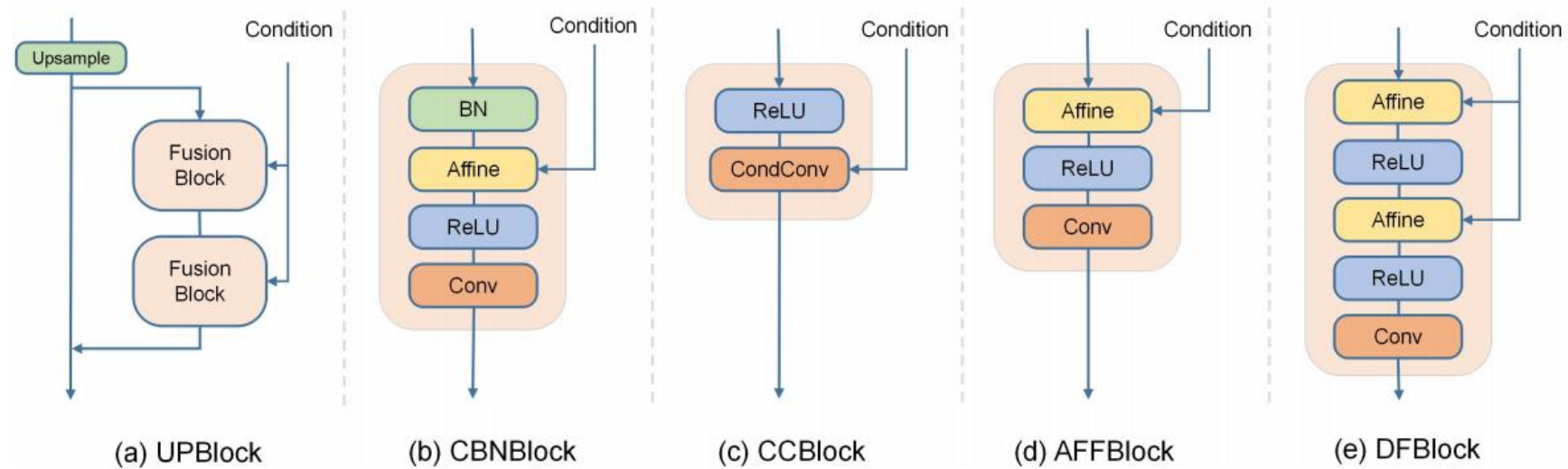
$$\begin{aligned} L_D = & -\mathbb{E}_{x \sim \mathbb{P}_r} [\min(0, -1 + D(x, e))] \\ & - (1/2) \mathbb{E}_{G(z) \sim \mathbb{P}_g} [\min(0, -1 - D(G(z), e))] \\ & - (1/2) \mathbb{E}_{x \sim \mathbb{P}_{mis}} [\min(0, -1 - D(x, e))] \\ & + k \mathbb{E}_{x \sim \mathbb{P}_r} [(\|\nabla_x D(x, e)\| +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

$$L_G = -\mathbb{E}_{G(z) \sim \mathbb{P}_g} D(G(z), e)$$

One-way discriminator



DFB



$$y = \frac{x - \mathbb{E}[x]}{\sqrt{\text{Var}[x] + \epsilon}} \cdot \gamma + \beta$$

结论

- 1.没有采用stacking结构，直接生成高质量的图文相关的图像。
- 2.提出了MA-GP策略，使得图文语义保持一致性，使生成器收敛得更快。
- 3.提出DFB模块，进一步融合图文信息，且提高效率。
- 4.改进判别器，使用one-way判别器，使得训练过程更加稳定，更加有效。

未来工作:

1. 基于few-shot工作的 text-to-image
2. 基于因果关系的text-to-image

1. 怎么更加有效融合文本和图片信息
2. 是否可以添加外部的模块来帮助模型学习
3. 从视觉领域引入新的技巧，来提高效果

2021

Thanks

Results :



DF-GAN

Results :

TABLE I
THE INCEPTION SCORE (IS) OF OUR PROPOSED DF-GAN COMPARED
WITH THE STATE-OF-THE ARTS ON THE TEST SET OF CUB.

Methods	Inception Score \uparrow
StackGAN++ [8]	4.04 ± 0.06
AttnGAN [10]	4.36 ± 0.03
MirrorGAN [13]	4.56 ± 0.05
SD-GAN [15]	4.67 ± 0.09
DM-GAN [14]	4.75 ± 0.07
DF-GAN (Ours)	4.86 ± 0.04

TABLE II
THE FID OF ATTNGAN, DM-GAN AND DF-GAN ON THE TEST SET OF
CUB AND COCO.

Methods	CUB-FID \downarrow	COCO-FID \downarrow
AttnGAN [10]	23.98	35.49
DM-GAN [14]	16.09	32.64
DF-GAN (Ours)	19.24	28.92

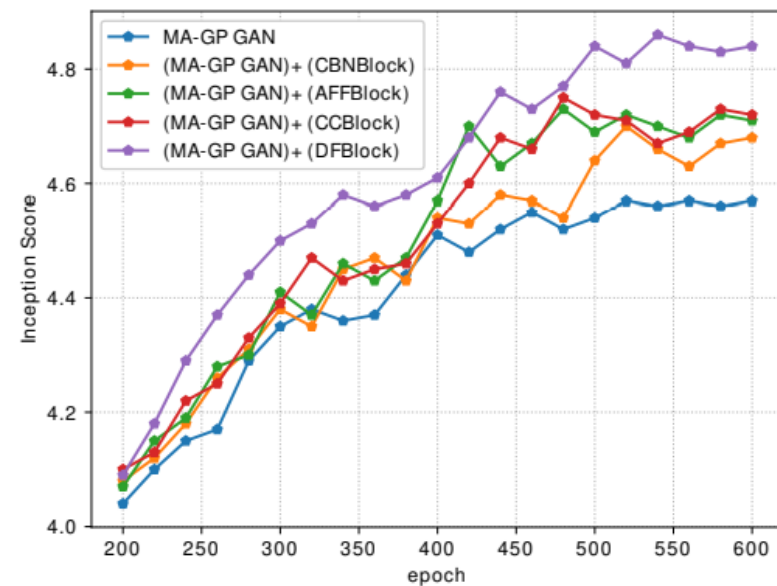
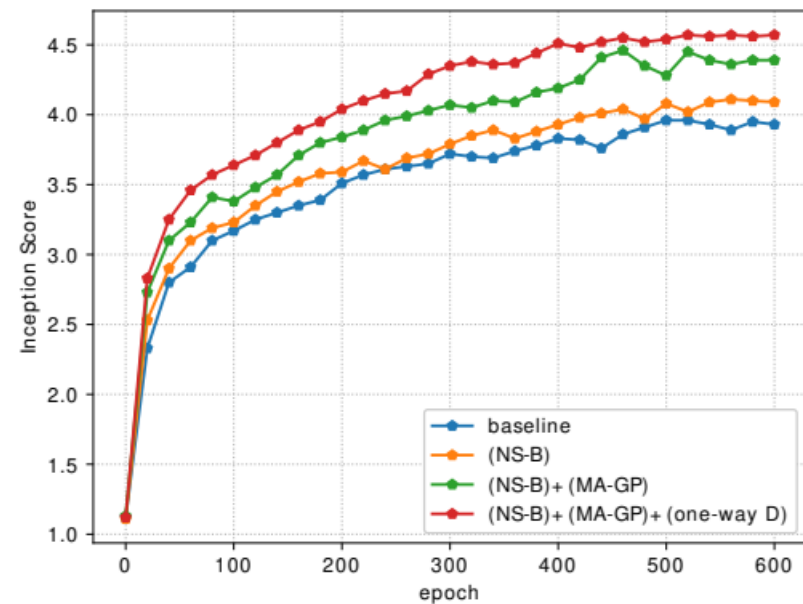
DF-GAN

试验

Methods	Inception Score \uparrow
StackGAN++ [8]	4.04 ± 0.06
AttnGAN [10]	4.36 ± 0.03
MirrorGAN [13]	4.56 ± 0.05
SD-GAN [15]	4.67 ± 0.09
DM-GAN [14]	4.75 ± 0.07
DF-GAN (Ours)	4.86 ± 0.04

Methods	CUB-FID \downarrow	COCO-FID \downarrow
AttnGAN [10]	23.98	35.49
DM-GAN [14]	16.09	32.64
DF-GAN (Ours)	19.24	28.92

DF-GAN



Results :

Method	inception score	R-precision(%)
AttnGAN1, no DAMSM	$3.98 \pm .04$	10.37 ± 5.88
AttnGAN1, $\lambda = 0.1$	$4.19 \pm .06$	16.55 ± 4.83
AttnGAN1, $\lambda = 1$	$4.35 \pm .05$	34.96 ± 4.02
AttnGAN1, $\lambda = 5$	$4.35 \pm .04$	58.65 ± 5.41
AttnGAN1, $\lambda = 10$	$4.29 \pm .05$	63.87 ± 4.85
AttnGAN2, $\lambda = 5$	$4.36 \pm .03$	67.82 ± 4.43
AttnGAN2, $\lambda = 50$ (COCO)	$25.89 \pm .47$	85.47 ± 3.69

Table 2. The best inception score and the corresponding R-precision rate of each AttnGAN model on CUB (top six rows) and COCO (the last row) test sets. More results in Figure 3.

Dataset	GAN-INT-CLS [20]	GAWWN [18]	StackGAN [31]	StackGAN-v2 [32]	PPGN [16]	Our AttnGAN
CUB	$2.88 \pm .04$	$3.62 \pm .07$	$3.70 \pm .04$	$3.82 \pm .06$	/	$4.36 \pm .03$
COCO	$7.88 \pm .07$	/	$8.45 \pm .03$	/	$9.58 \pm .21$	$25.89 \pm .47$

Table 3. Inception scores by state-of-the-art GAN models [20, 18, 31, 32, 16] and our AttnGAN on CUB and COCO test sets.

Attn-GAN