[T2I]Generative Adversarial Text to Image Synthesis

2020. 4. 24. 12:00

#gan #Generative Adversarial Text to Image Synthesis #PAPER #T2i #TextToImage

다음 논문을 읽고 작성한 내용이며 잘못 해석한 내용이 있을 수도 있습니다.
https://arxiv.org/abs/1605.05396

Generative Adversarial Text to Image Synthesis

Abstract

Text로 실제 이미지와 같은 합성 이미지를 만드는 것은 흥미롭고 유용하지만 많은 연구가 되어 있지는 않다
최근 RNN과 GAN의 연구가 활발해지고 있고 text to image를 이에 적용시켜보도록 한다.

1. Introduction

single-sentence human written descriptions를 image pixel로 바꾸는 것에 관심이 있다.
ex) this small bird has a short, pointy orange beak and white belly -> image
natural language는 물체를 설명하는데 일반적이고 유연한 추론을 제공한다. 따라서 이상적으로 text description이 discriminative의 성능을 높힐 것이다.
최근에 text와 image 분야에서 다양한 연구(zero-shot)가 진행되고 있고, 이러한 연구를 바탕으로 char to pixel을 mapping 학습을 시도해본다.
이 문제를 해결하기 위해서는 두 가지 문제를 해결해야 한다.

1. 시각적으로 중요한 것을 잡아내는 text feature 표현을 학습해야 한다.
2. 이 feature를 사용해서 사람이 실제라고 착각할만한 이미지를 만들어내야 한다.

딥러닝은 이 두가지 subproblem을 각각을 잘 해내고 이를 한번에 푸는 것이 목표다.
하지만 딥러닝으로도 어려운 것이 있는데, text 설명을 조건으로 가지는 image의 분포가 매우 높은 multimodal이다

- 텍스트 설명에 맞는 이미지는 무수히 많기 때문에 쉽지 않다.
- image to text에서도 이러한 문제가 발생하지만 문장의 단어를 순차적으로
- 주어진 이미지와 생성할 단어들을 조합하면 well-defined prediction problem이 된다.
- image에서 text를 만드는 것은 한 단어를 기준으로 순차적으로 생성하면 되지만, text에서 image를 만드는 것은 한 번에 진행해야 할 일이기 때문에 어렵다.

새나 꽃 이미지를 text 설명으로 생성하는 것이 목표이며 다음과 같은 dataset을 사용할 것이다.

- Caltech-UCSD (Birds)
- Oxford-102 (Flowers)
- MS COCO (general)

2. Related work

1. Multimodal learning
   어떠한 modal을 사용할지?
2. 여러 modal에서 공유되는 representation 연구
   이미지가 여러개 생성된다고 해도 공통되는 특징이 어떠한 것인가

3. Deep Convolutional decoder network 관련 연구
   이미지를 만들어내는 과정이 decoder에 해당하는데 DCN을 기반으로 실제같은 이미지를 합성하는 연구
4. image to text
   이 논문의 반대

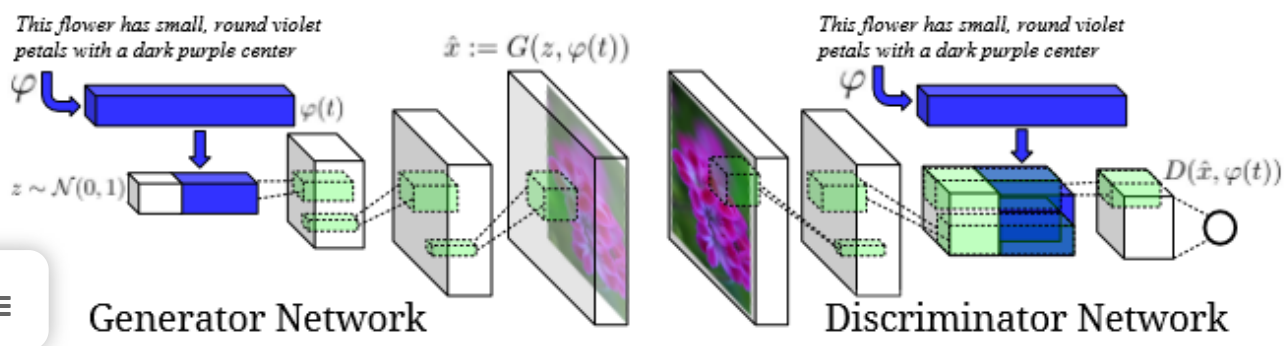
3.Background

생략

4.Method

text feature를 조건으로 가지는 DC-GAN을 학습한다. text feature는 hybrid character-level convolutional recurrent neural network로 encoding 했다.

4.1 Network architecture



G : R^Z x R^T -> R^D, D : R^D x R^T -> {0,1}

T는 text description embedding의 차원

D는 image의 차원

Z는 G의 noise input의 차원

G에서 noise는 정규분포에서, t는 text encoder를 통해 encoding한다. embedding은 fc를 사용해서 작은 차원으로 압축시킨다.(leaky ReLU 사용) 그리고 이를 noise vector z와 합친다. 이 과정을 거친 뒤에 normal deconvolutional network에 진행시켜서 합성된 이미지 x를 얻는다. 이미지 생성은 query text와 noise sample을 바탕으로 진행한다.

D에서, stride가 2인 spatial batch normalization를 수행하고, description embedding의 차원과 같아질 때까지 반복한다. 4x4의 차원이 될 때 텍스트 벡터를 복사해서 concat을 다시 수행하고 final score를 구해 분류를 한다.

4.2 Matching-aware discriminator(GAN-CLS)

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

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

conditional GAN을 훈련하기 위해 대부분 text와 image 쌍을 joint obeservation 하고, D가 진짜인지 가짜인지 판별하는 방법이다. 하지만 이렇게 naive한 discriminator는 진짜 image가 text embedding context와 일치하는지 알 수 없다.

discriminator가 conditioning information을 무시하고, 쉽게 거절해버리는데 G가 이상하게 생성하기 때문이다. G가 괜찮게 이미지를 생성하기 시작하면 이것이 information을 담고 있는지 판별해야 한다.

일반 naive GAN은 D는 두 개의 input을 가진다. real image with matching text, synthetic image with arbitrary text 그러므로 당연히 두 가지 에러를 가지게 될 것이다. unrealistic image, realistic image but mismatch conditioning information.

따라서 앞서 말한 내용들을 반영하기 위해서 GAN 알고리즘을 수정한다. 세번째 input type을 넣는데, D가 fake라고 반별하는 reacl image with mismatched text이다.

4.3 Learning with manifold interpolation (GAN-INT)

딥러닝은 data manifold 근처에 있는 embedding pari 사이에서 interpolations을 수행해 표현을 학습한다고 밝혀졌다. 이를 이용해서 training set의 embedding으로 부터 interpolation을 수행해 추가적인 text embedding을 얻을 수 있다. 그리고 이 것은 사람이 작성한 text가 아니기 때문에 labeling cost가 발생하지 않는다고 한다. 비슷한 이미지에 대해서 많은 텍스트 임베딩을 만들고 그들이 공유하는 것을 찾아낼 수 있기 때문에 그럴듯한 이미지를 만들 수 있다. 이 것은 다음 목적함수를 보면 알 수 있다.

$$\mathbb{E}_{t_1,t_2\sim p_{data}}[\log(1-D(G(z,\beta t_1+(1-\beta)t_2)))] \quad (5)$$

여기서 z는 noise 분포에서 온 것이고, B는 두 text embeddings를 interpolation한 값이며 보통 0.5로 고정한다.

interpolate 된 embedding은 합성이기 때문에 D는 이 와 연관된 image가 없기 때문에 real을 가지고 있지 않다. 그러나 image와 text가 match하는지 예측해야 하기 때문에 fake라고 예측할 것이다. 그러므로 D가 이

것을 잘 수행하면 training points 사이의 data manifold 상에 있는 gap을 메꿀 수 있다.

- training points는 유한개이기 때문에 빈 공간이 생길 수 밖에 없는데 무한히 생성한다면 이 공간들을 메꿀 수있을 것이다.

4.4 Inverting the generator for style transfer

text encoding은 image content를 잡아내고, real같은 이미지를 만들기 위해서 noise는 뒷 배경이나 pose 같은 style factor를 포착한다. GAN을 훈련시킬 때, query image의 style을 특정한 text 설명의 내용으로 전환시키고 싶다. 이를 하기 위해서 G를 x^에서 z로 거꾸로 흐르게 훈련을 시킨다. 따라서 다음과 같은 squared loss를 가지는 style encode를 가진다.

$$\mathcal{L}_{style} = \mathbb{E}_{t,z\sim\mathcal{N}(0,1)}||z - S(G(z,\varphi(t)))||_2^2 \qquad (6)$$

5. Experiments

Dataset

- CUB dataset : 11,788개의 새 이미지
- Oxford -102 dataset - 8,189개의 꽃 이미지

Encoder

- Text encoder - deep convolutional recurrent text encoder(1024 dimensional)
- Image encoder - 1024 dimensional GoogLeNet

Hyper Parameter

- Image size 64 x 64 x 3
- learning rate 0.0002
- momentum 0.5
- Minibatch size : 64
- 600 epochs

5.1 Qualitative results

GAN을 baseline으로 GAN-CLS, GAN-INT, GAN-INT-CLS를 비교했다.
CUB 데이터로 봤을 때, GAN과 GAN-CLS는 color information 정보는 맞지만 real 같아 보이지는 않다. 하지만 GAN-INT와 GAN-INT-CLS는 이미지가 text를 잘 반영하며 그럴듯한 이미지를 보여주고 있다.

Oxford-102 데이터로 봤을 때에는 4개의 방법 모두 설명을 잘 반영하며 그럴듯한 이미지를 생성하고 있다.
따라서 interpolation을 적용한 방법이 더 좋은 성능을 내는 것이라고 생각된다.

5.2 Disentangling style and content

style과 content를 구분한다. content는 새 자체의 시각적 특징을 의미하며 모양, 크기 각 몸의 색깔을 나타낸다. style은 image의 다른 요소들 뒷 배경이라 새 자체의 pose를 의미한다.
text embedding은 주로 content information을 담당하고, style에는 관여하지 않는다. 그러므로 실제같은 이미지를 생성하기 위해서 GAN은 noise sample z를 잘 학습해야 한다. style encoder에 미지를 줌으로써 style vector를 예측한다. GAN이 image content로 부터 z를 사용해 style을 풀어내고 싶다면 같은 style을 가지는 이미지들 사이의 유사성이 더 높을 것이다.
z를 위해서 이전에 4.4와 같이 구축하고, K-means를 이용해 100개의 cluster를 그룹지었다.
평가를 위해서 4가지 모델에 대해 style encoder를 적용해서 비교해 보았다. cosine 유사도를 사용해 점수를 매겼으며 AU-ROC로 기록했다.?
예상했던 것처럼 caption만을 사용하는 것은 style prediction에 도움을 주지 못했다. 게다가 interpolation을 적용한 것이 더 성능이 좋았다.

5.3 Pose and background style transfer

훈련된 style encoder를 가지고 있는 GAN-INT-CLS는 text 설명이 있는 보지 않은 query image로 부터 style transfer를 수행할 수 있다. style encoder를 이용해 분리해낸 스타일을 다른 이미지를 생성할 때 transfer 할 수 있다.
맨 위의 이미지에서 얻어낸 스타일을 통해서 새로운 이미지를 만들때 이를 사용해서 그럴듯한 이미지를 만드는 것이다. 텍스트만으로 새에 대한 정보를 알 수 없기 때문에 이러한 방법을 사용한다.

5.4 Sentence interpolation

intervening points를 위한 no ground-truth text가 없음에도 불구하고 그럴듯한 이미지를 만들어낼 수 있다.
noise 분포가 같다고 하면 우리가 사용하는 text embedding만 바뀌게 된다. interpolation이 새가 파란색에서 빨간색으로 변하는 것과 같은 color information을 반영할 수 있다. 즉 텍스트의 일부가 변경되어도 noise가 고정되기 때문에 새의 색만 바뀌게 되는 것이다.

반대로 text embedding 값을 interpolate하는 것처럼 두 noise 벡터를 interpolate한 결과로 text 부분이 고정되고 noise와 style이 smooth하게 연결되는 것을 확인할 수 있다.

5.5 Beyond birds and flowers

이제 일반성을 보기 위해서 bird, flower 말고 일반 데이터인 COCO 데이터를 통해서 진행한다. 나름 괜찮은 결과를 얻을 수 있었다. 하지만 중심이 되는 객체가 여러개일 경우에는 잘 반영을 하지 못하고 있으며 추후 연구가 필요하다고 합니다.

6. Conclusions

시각 설명에 기반한 이미지를 생성하는 효과적이고 간단한 모델을 개발했다.
manifold interpolation가 성능을 향상시켰으며, content와 style을 분해하여 복사해내는 모델임을 증명했다.
MS COCO 데이터에도 일반적으로 적용이 가능하다.
고해상도 이미지를 만들고, 많은 텍스트를 반영하기 위한 연구가 필요하다.

[T2I]Generative Adversarial Text to Image Synthesis (0) 2020.04.24

[NLP]Deep Reinforcement Learning for Dialogue Generation (0)

NAME

PASSWORD

HOMEPAGE

http://

SECRE WRITE

PREV 1 ··· 160 161 162 163 164 165 166 167 168 ··· 500 NEXT

+ Recent posts

[BOJ]20057. 마법사 상어와…

[BOJ] 15685. 드래곤 커브

[BOJ]20056. 마법사 상어와…