CV-003, Generative Adversarial Text to Image Synthesis (2016-JMLR)

0. Abstract

- 이 당시에만 해도 text to image 생성 연구가 거의 없었다고 한다.
- 이미지에서 일반적인 GAN은 연구가 잘 되 엉 ㅣ ㅆ다.
- GAN을 이용해서 text to image을 해보겠다는 것이고, 즉 visual concepts을 character에서 pixels로 변환시키겠다.

1. Introduction

- single-sentence (사람이 쓴 설명)을 가지고 이미지 픽셀로 변환시키겠다.
    - 예) "this small bird has a short, pointy orange beak and white belly"
- NLP는 어떠한 물체를 설명하기에 일반적이고 유연하다. (단순히 레이블링을 속성으로 쓰기에는 도메인에 대한 정보가 필요하다는 듯)
    - 따라서 이상적으로는 text descriptions이 discriminator의 성능을 높여줄 것이다.
- Caltech-UCSD에서 zero-shot visual recognition의 방법을 응용해ㅐ ㅆ다는 것 같음.
- 문제를 해결하기엔 두 가지 문제가 있다.
    1. learn a text feature representation that captures the important visual details
    2. use these features to synthesize a compelling image that a human might mistake for real.
    - 딥러닝은 이 두가지 subproblem을 각각은 잘 해내고 결국 한꺼번에 푸는 것이 목표임.
- 어려운 점은 text descriptions을 조건으로 이미지의 분포를 생성하기에는 너무 많은 multimodal이다.
    - 즉, 텍스트 설명에 맞는 이미지는 무수히 많은게 존재하기 때문에 쉽지 않다.
    - Image to text도 똑같은 문제가 있으나, image to text을 학습할 때는 문장의 단어를 순차적으로 생성하게 된다.
    - 따라서 처음에 주어진 이미지와 생성할 단어 이전의 단어들을 조합하면 well-defined prediction problem이 되는 것이다.
    - 즉 text to image는 이미지를 한 번에 생성해야 하기 때문에 더 어렵다는 듯
- 데이터로는 다음과 같이 있다.
    - Caltech-UCSE Birds
    - Oxford-102 Flowers
    - MS COCO dataset

2. Related Word

- 생략

3. Background

- 생략

4. Method

- Text features을 조건으로 한 DC-GAN을 이용함.
    - text features은 hybrid character-level convolutional-RNN을 이용해서 뽑음.

4.1 Network architecture

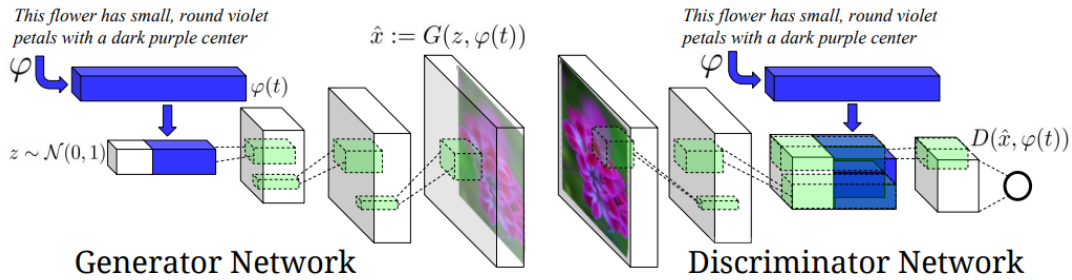


Figure 2. Our text-conditional convolutional GAN architecture. Text encoding $\varphi(t)$ is used by both generator and discriminator. It is projected to a lower-dimensions and depth concatenated with image feature maps for further stages of convolutional processing.

- $G : \mathbb{R}^Z \times \mathbb{R}^T \rightarrow \mathbb{R}^D$
- $D : \mathbb{R}^D \times \mathbb{R}^T \rightarrow \{0, 1\}$
- 여기서 T는 text description embedding 차원
- D는 image의 차원
- Z는 noise input의 차원
- $z \in \mathbb{R}^Z \sim N(0, 1)$
- text query t는 text encoder φ에 의해 encode 됨.
- φ는 128 fully-connected layer + leaky-ReLU임.
- 생성 문장 $\hat{x}$는 $\hat{x} \leftarrow G(z, \varphi(t))$로 생성이 됨.
- Generator은 feed-forward로 deconvolution 과정을 거침
- Discriminator D는 stride 2, spatial batch normalization with ReLU을 거침.

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

- 일반적인 GAN에서 text을 condition으로 이미지를 생성해서 discriminator가 real or fake인지 구하게 하
- 이것이 real training image와 text embedding context가 정말 매칭이 되는지 명시적인 개념이 없다고 한다.
- 학습을 할 때, non-conditional case과 다르게 discriminator은 conditioning information을 쉽게 무시한다고 한다.
  - 왜냐하면 애초에 generator가 생성하는 이미지가 그럴듯한 이미지가 아니기 때문에 condition은 따질만한 거리가 아니라고 생각한다.
  - 따라서 G가 일단 그럴듯한 이미지를 만든 후에 이것이 conditioning information과 align이 되는지를 판별해야 한다.
- Discriminator은 두 종류의 입력을 받는다.
  1. real image + matching text
  2. synthetic image + arbitrary text
- 따라서 Discriminator은 두 가지 에러를 잡아야한다.
  - unrealistic image가 생성되는 것
  - realistic image가 생성됐지만 conditioning information과 mismatch가 된 이미지
- 따라서 GAN 학습 알고리즘을 수정해야 한다.
  - 이 논문에서 3번째 type의 입력을 넣었다고 하는데
  - 3번째 유형의 입력은 real image + wrong text이고 이것에 대한 discriminator은 fake라고 판단을 해야한다.
  - 1,2번째는 기본 유형의 real image,right text → real로 판단, fake image,right text →

    - 1,2번째는 기본 유형인 real image+right text → real로 판단, fake image+right text → fake로 판단
- 위의 알고리즘 1 도식도를 보면 이에 대해 학습 흐름도가 있다.

4.3 Learning with manifold interpolation (GAN-INT)

- 딥러닝 네트워크는 embedding 쌍 사이의 interpolation 근처 data manifold 사이에서 representation을 학습함이 밝혀졌다고 한다.
- 따라서 additional text embedding의 많은 양을 생성하기 위해, 학습 데이터의 captions의 embedding 사이를 interpolation하면서 학습했다고 한다.
- 이러한 방법은 사람이 직접 쓴 text는 아니여서 (embedding representation에서의 값일 뿐) 추가적인 cost가 없다.
- 최소화할 Objective 함수는 다음과 같다.
    - $$\mathbb{E}_{t_1,t_2\sim p_{data}}[\log(1-D(G(z,\beta t_1+(1-\beta)t_2)))] \quad (5)$$
    - z는 noise distribution이고 β는 text embedding t1과 t2를 interpolation하는 값으로 β=0.5으로 설정하면 잘 작동한다고 한다.
    - 즉 여기서 잘못된 text을 interpolation을 통해서 생성해서 이용한다는 것인 듯
- t1과 t2 사이를 interpolation해서 생성된 text embedding을 t3라고 하면, t3하고 매칭되는 real image는 없다.
    - 하지만, D가 image와 text가 매칭되는 것을 배우기 때문에 t3가 들어가면 정답은 fake가 되어야 한다.
    - D가 잘 작동한다면, G는 training points 사이의 data manifold의 차이를 매꿀수 있게 된다.
        - 이게 핵심 효과인 것 같음
        - 원래라면, training data는 당연히 유한 개이므로 text embedding 또한 유한 개의 point라고 볼 수있다.
        - 따라서 discrete하여 text embedding 사이의 gap이 존재할 텐데, 이것으로 이 부분을 메꿀 수 있다는 것..!!
    - t1과 t2는 다른 이미지 카테고리에서 올 수도 있다.

4.4 Inverting the generator for style transfer

- text encoding $\varphi(t)$이 image content을 (예, 꽃의 모양, 색깔) 담아낸다면, realisitc한 이미지를 생성하기 위해서는 noise sample z는 style factors을 (바탕 색, 포즈 등) 담아내야 한다.
- 이렇기 위해서 $\hat{x}\leftarrow G(z,\varphi(t))$가 다시 z로 돌아가게끔 다음과 같이 학습을 한다.
    - $$\mathcal{L}_{style}=\mathbb{E}_{t,z\sim\mathcal{N}(0,1)}||z-S(G(z,\varphi(t)))||_2^2 \quad (6)$$
    - 학습 할 때, S(style encoder)와 G(generator)을 업데이트 함.
- 따라서 다음과 같이 되는 것
    - $$s\leftarrow S(x),\hat{x}\leftarrow G(s,\varphi(t))$$

5. Experiments

- 실험 부분은 결과 그림, 표말고는 생략을...

5.1 Qualitative results

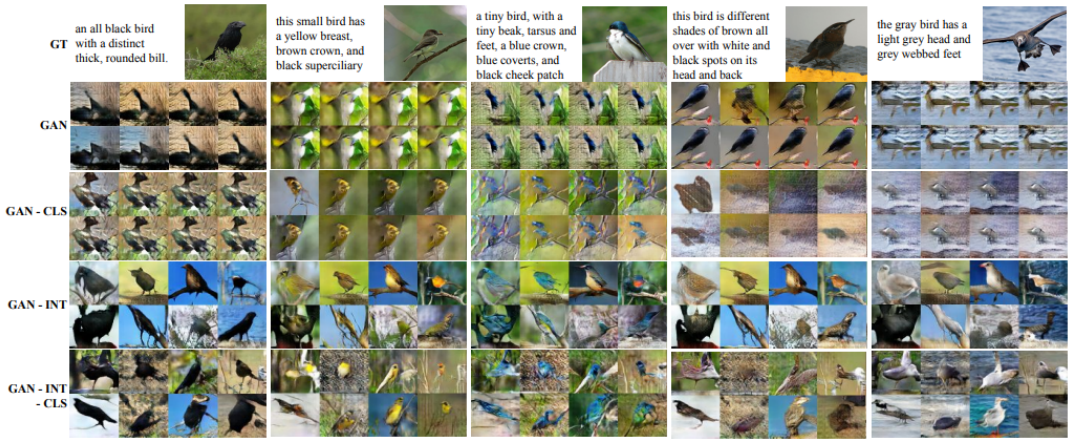


Figure 3. Zero-shot (i.e. conditioned on text from unseen test set categories) generated bird images using GAN, GAN-CLS, GAN-INT and GAN-INT-CLS. We found that interpolation regularizer was needed to reliably achieve visually-plausible results.

5.2 Disentangling style and content

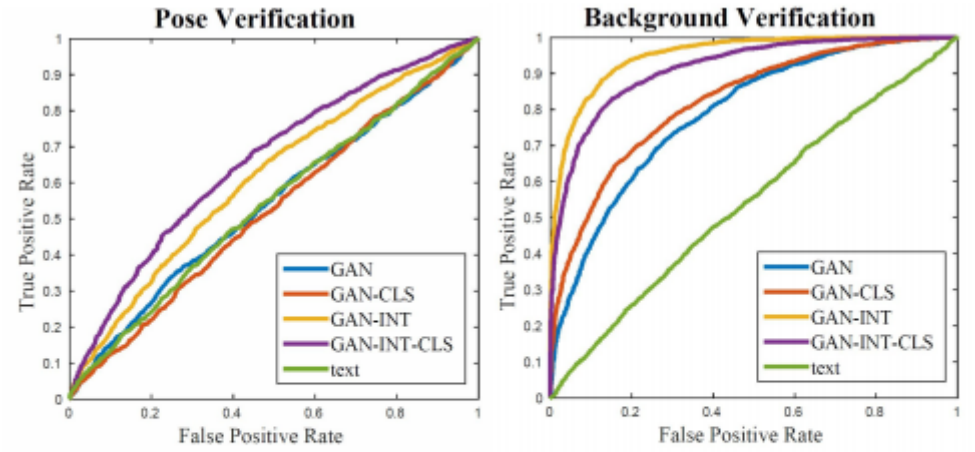
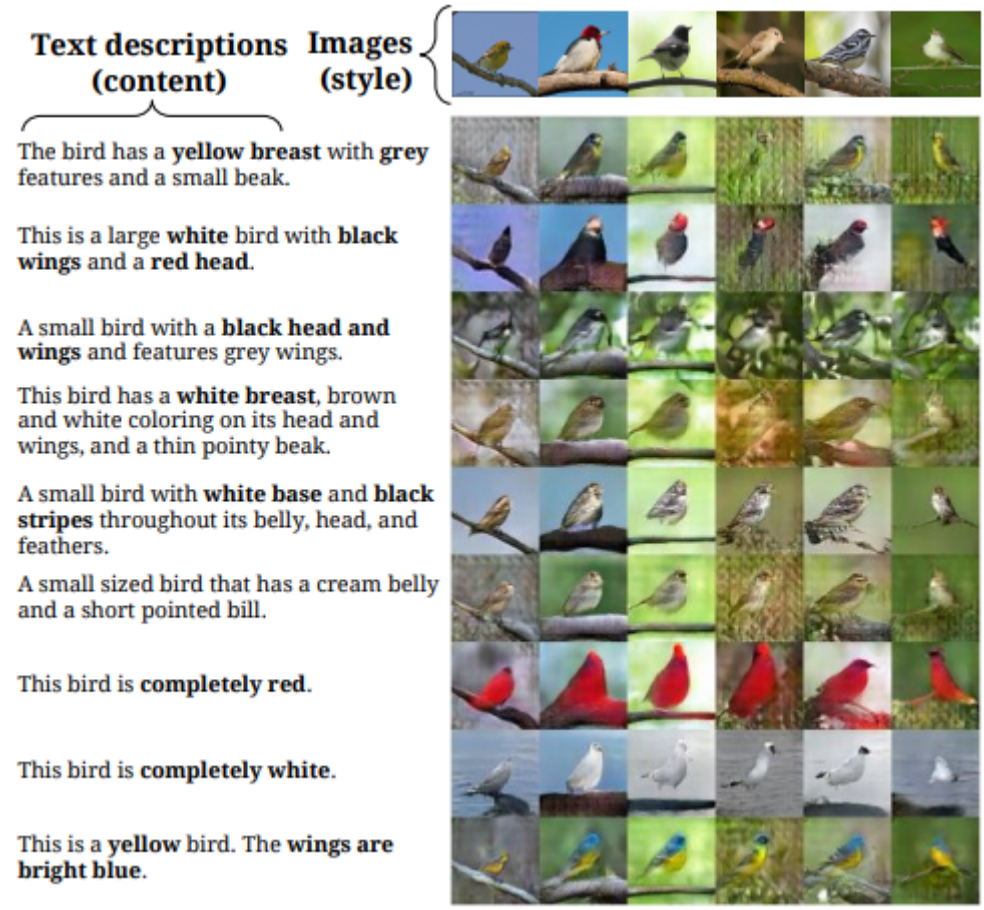


Figure 5. ROC curves using cosine distance between predicted style vector on same vs. different style image pairs. Left: image pairs reflect same or different pose. Right: image pairs reflect same or different average background color.

5.3 Pose and background style transfer



-

5.4 Sentence interpolation

Figure 8. Left: Generated bird images by interpolating between two sentences (within a row the noise is fixed). Right: Interpolating between two randomly-sampled noise vectors.

-

5.5 Beyond birds and flowers

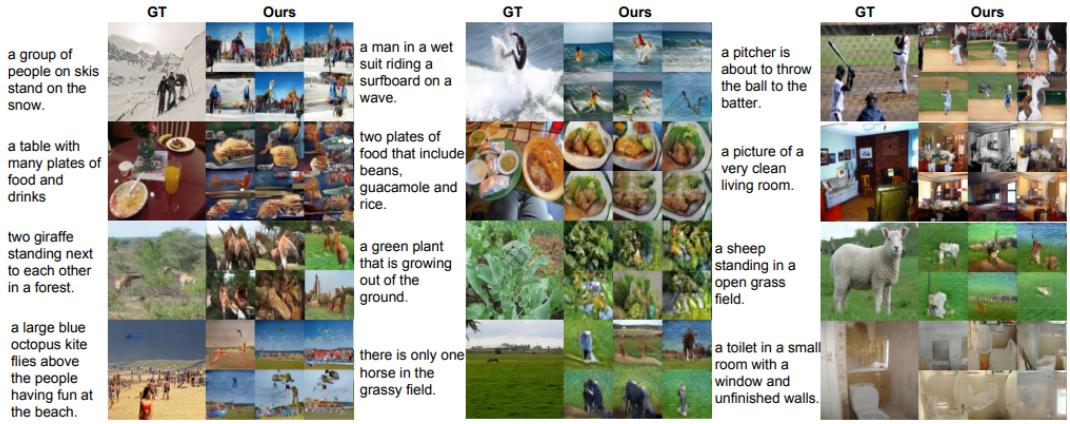


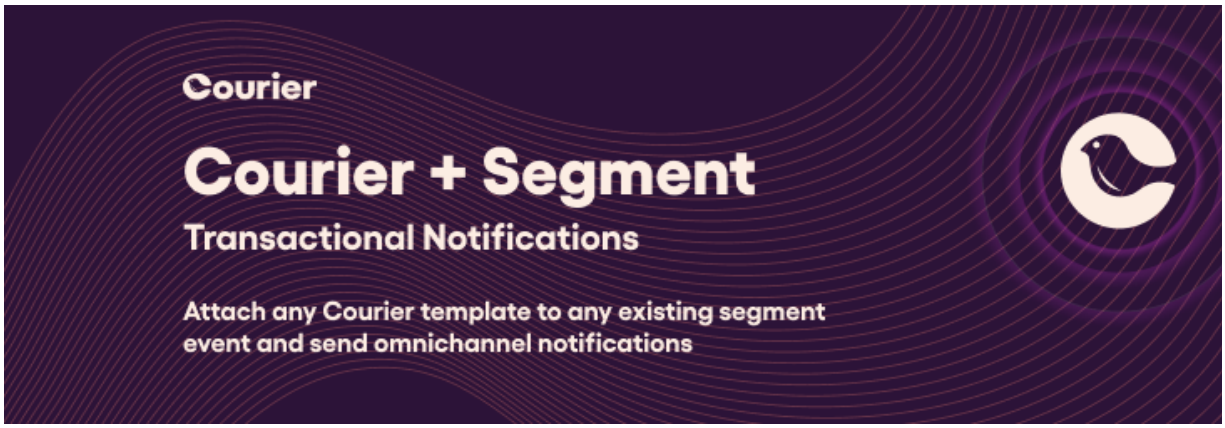
Figure 7. Generating images of general concepts using our GAN-CLS on the MS-COCO validation set. Unlike the case of CUB and Oxford-102, the network must (try to) handle multiple objects and diverse backgrounds.

6. Conclusions

- In this work we developed a simple and effective model for generating images based on detailed visual descriptions.
- We demonstrated that the model can synthesize many plausible visual interpretations of a given text caption.
- Our manifold interpolation regularizer substantially improved the text to image synthesis on CUB.
- We showed disentangling of style and content, and bird pose and background transfer from query images onto text descriptions.
- Finally we demonstrated the generalizability of our approach to generating images with multiple objects and variable backgrounds with our results on MS-COCO dataset.
- In future work, we aim to further scale up the model to higher resolution images and add more types of text.

Reference

- https://arxiv.org/pdf/1605.05396.pdf



mexade92@gmail.com

RUNGJOO

프로필로 이동

Introduction

AI papers

Papers Category

Materials

Challenge & Dataset

Youtube (주렁코드) & 코딩

신고하기