Generative Adversarial Text to Image Synthesis

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

Presented by: Jingyao Zhan

Contents

Introduction

- RNN -> discriminative text feature representations
- GAN -> compelling images of specific categories.
- Problem: multimodal distribution
  - Multi-modality is a very well suited property for GANs to learn.
- Main contribution:
  - A simple and effective GAN architecture and training strategy that enables compelling text to image synthesis
  - Performance is demonstrated by “zero-shot” text to image synthesis
    - Q: What is “zero-shot” learning?

Generate images
from text descriptions

Related Work

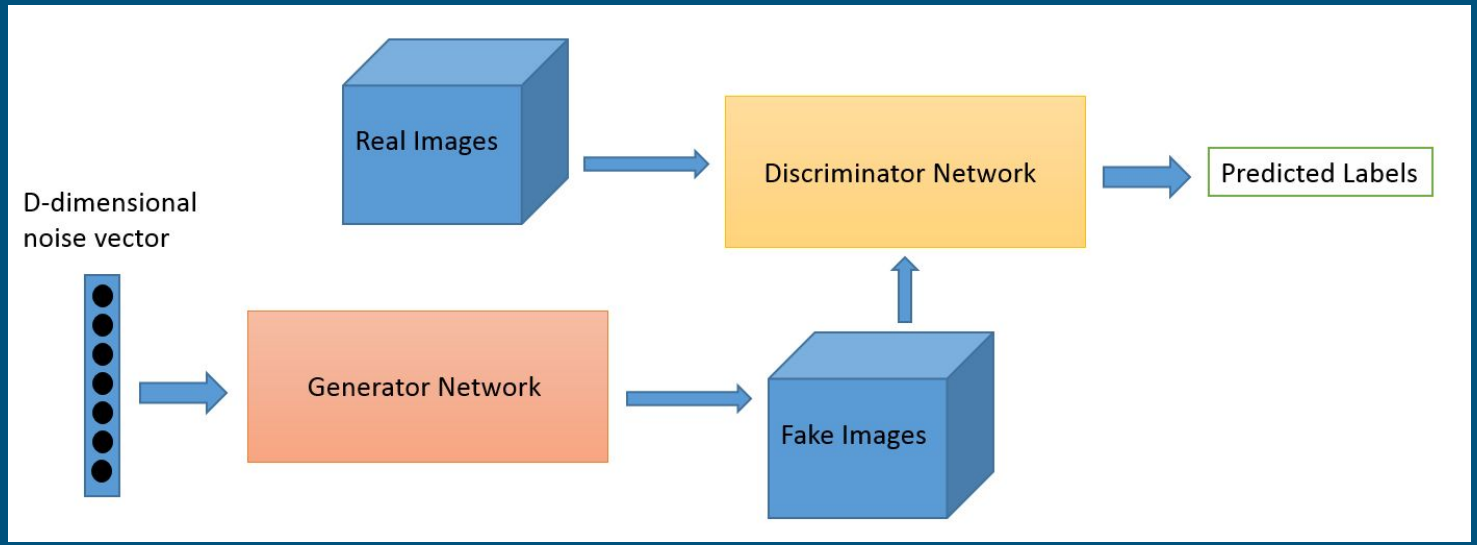
- Learning *shared representation* across modalities and predict data in one modality conditioned on another.
- Deep convolutional decoder networks for generating realistic images
- Generative adversarial networks
  - Denton et al (2015): Laplacian pyramid of adversarial generator and discriminators
    - Generated high-resolution images; condition on class labels
  - Radford et al (2016): standard convolutional decoder incorporating batch normalization
- How this work is different?
  - Conditions on *text description* rather than *class labels*
  - Manifold interpolation regularizer for GAN generator

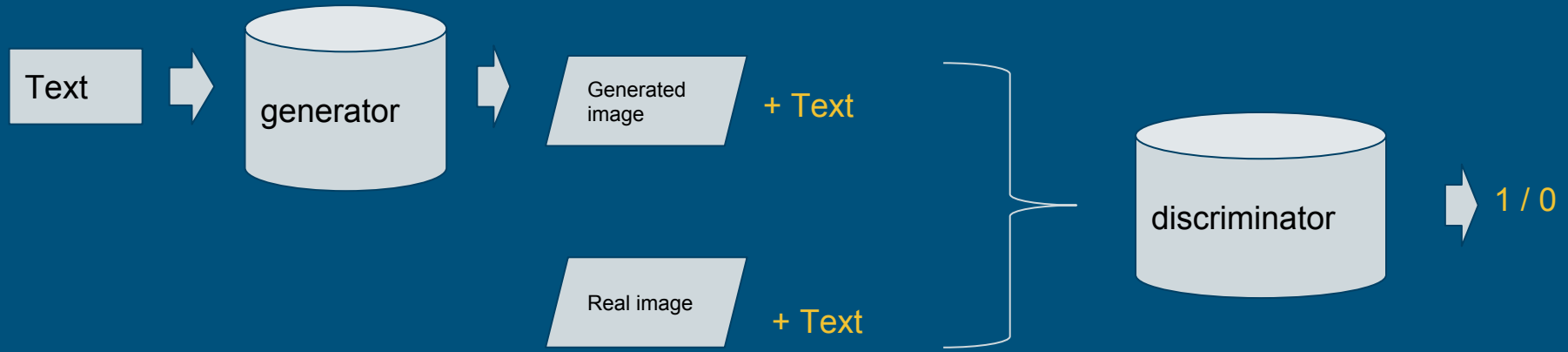
Related Work

- Using RNN decoders to generate text descriptions conditioned on images
  - Condition Long Short-term Memory on the top layer features of a deep convolutional network to generate captions
  - Recurrent visual attention mechanism
- Generating images from text:
  - Mansimov et al. (2016): variational recurrent autoencoder w/ attention
  - Generates unrealistic images

Background: GAN

- Generative Adversarial Networks
  - Generator G: generates new data instances
  - Discriminator D: evaluate them for authenticity




Text
generator
Generated image
+ Text
Real image
+ Text
discriminator
1 / 0

Background

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{x\sim p_z(z)}[\log(1 - D(G(z)))] \quad (1)$$

- A Minimax game which has a global optimum precisely when p_g = p_data (p_g converges to p_data under mild conditions)

$$\frac{1}{N}\sum_{n=1}^{N} \Delta(y_n, f_v(v_n)) + \Delta(y_n, f_t(t_n))$$

(vn, tn, yn): training data set.
vn are the images, tn are the corresponding text descriptions, and yn are the class labels.

Background

$$f_v(v) = \arg\max_{y \in \mathcal{Y}} \mathbb{E}_{t \sim \mathcal{T}(y)} [\phi(v)^T \varphi(t))]$$

Image classifier

$$f_t(t) = \arg\max_{y \in \mathcal{Y}} \mathbb{E}_{v \sim \mathcal{V}(y)} [\phi(v)^T \varphi(t))]$$

Text classifier

φ: the image encoder

Φ: the text encoder

T (y) is the set of text descriptions of class y and likewise V(y) for images.

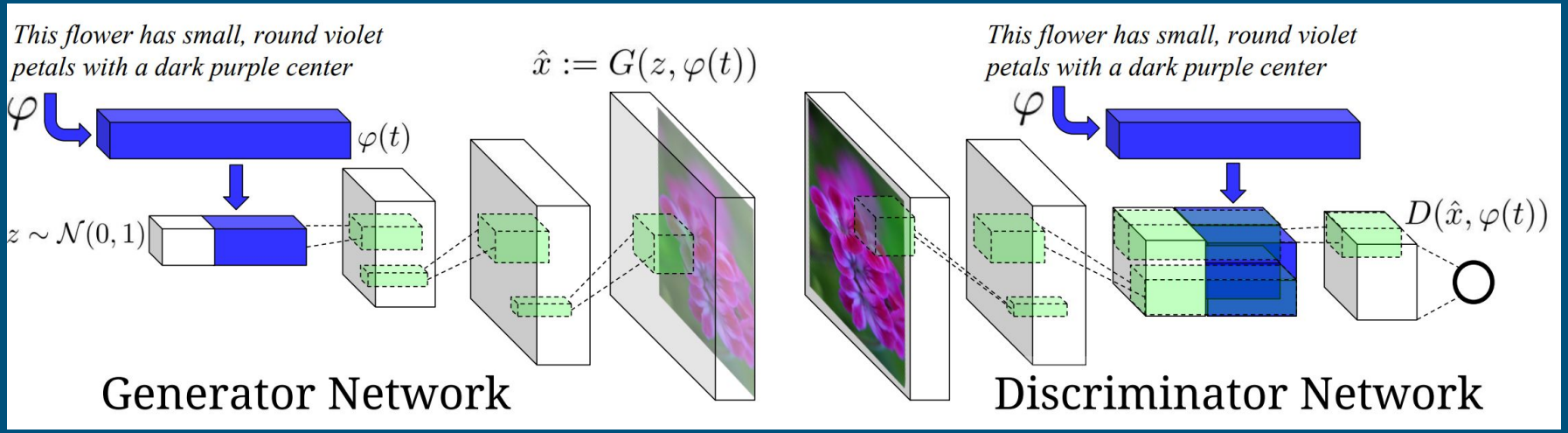
$$F(v,t) = \phi(v)^T \varphi(t)$$

Compatibility function with Φ(v) as image encoding and φ(t) as text embedding

Intuition: text encoding should have a higher compatibility score with images of the corresponding class compared to any other class and vice-versa

Method

- train a deep convolutional generative adversarial network (**DC-GAN**) conditioned on text features encoded by a **hybrid character-level** convolutional recurrent neural network.



Q: what is rectification? What does it do?

DC-GAN Structure Revisit

- “The generative model, G, takes a random 100 dimensional vector drawn from a uniform distribution between [−1, 1] and generates a 64×64×3 image. The discriminator model, D, is structured essentially in reverse order. The input layer is an image of dimension 64 × 64 × 3, followed by a series of convolution layers where the image dimension is half, and the number of channels is double the size of the previous layer, and the output layer is a two class softmax. ”

  Reference: Semantic Image Inpainting With Perceptual and Contextual Losses

- most straightforward way to train a conditional GAN is to view (text, image) pairs as joint observations and train the discriminator to judge pairs as real or fake
  - discriminator has no explicit notion of whether real training images match the text embedding context
- The paper proposes two possible improvements on DC-GAN model.

GAN-CLS

- Motivation: discriminator must implicitly separate two sources of error (Q: which two?)
- Main idea: Adding the third type of input consisting of real images with mismatched text, which discriminator must learn to score as fake.
  - Can learn to optimize image/text matching better
- Introducing new loss functions:

$$\mathcal{L}_D \leftarrow \log(s_r) + (\log(1 - s_w) + \log(1 - s_f))/2$$

$$\mathcal{L}_G \leftarrow \log(s_f)$$

GAN-CLS

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

1: **Input:** minibatch images x, matching text t, mismatching $\hat{t}$, 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**

Encoding text, image, and noise

Generate fake image

scores

Ln 11 and 13: taking a gradient step to update network parameters

GAN-INT

- Motivation: deep networks can learn representations in which interpolations between embedding pairs tend to be near the data manifold
- Main idea: Generating a large amount of additional text embeddings by simply **interpolating** between embeddings of training set captions.
- Adding an additional term to the generator objective to minimize

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

- Parameter β: interpolates between text embeddings t_1 and t_2.

GAN-INT

- D does not have “real” corresponding image and text pairs to train on, but D learns to predict if image and text pairs match or not.
- By satisfying D on interpolated text embeddings G can learn to **fill in gaps** on the data manifold between training points.

Inverting the generator for style transfer

- Image content: visual attributes of the objects such as shape, size and color of each body part of a bird.
- Style: all the other factors of variation in the image such as background color and the pose orientation of the bird.
- φ(t) captures image content, noise sample z should capture style factors such as background color and pose
- The authors train a CNN to invert G to regress from samples back to style z.

Squared loss to train style encoder:

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

Style transfer from a query image x on to text t:

$$s \leftarrow S(x), \hat{x} \leftarrow G(s,\varphi(t))$$

Experiments

- Data sets used:
  - CUB dataset of bird images - 11,788 images of birds of 200 different categories
  - Oxford-102 dataset of flower images- 8,189 images of flowers from 102 different categories
- Class-disjoint training and test sets:
  - CUB: 150 train+val, 50 test classes
  - Oxford-102 82 train+val, 20 test classes
  - 5 captions per image
  - Randomly pick an image view + one of the captions
- Pre-train a deep convolutional recurrent text encoder on structured joint embedding of text captions
  - Q: what is the reason for pre-training?
- Hybrid of character-level ConvNet with a recurrent neural network (char-CNN-RNN)

Qualitative Results: CUB

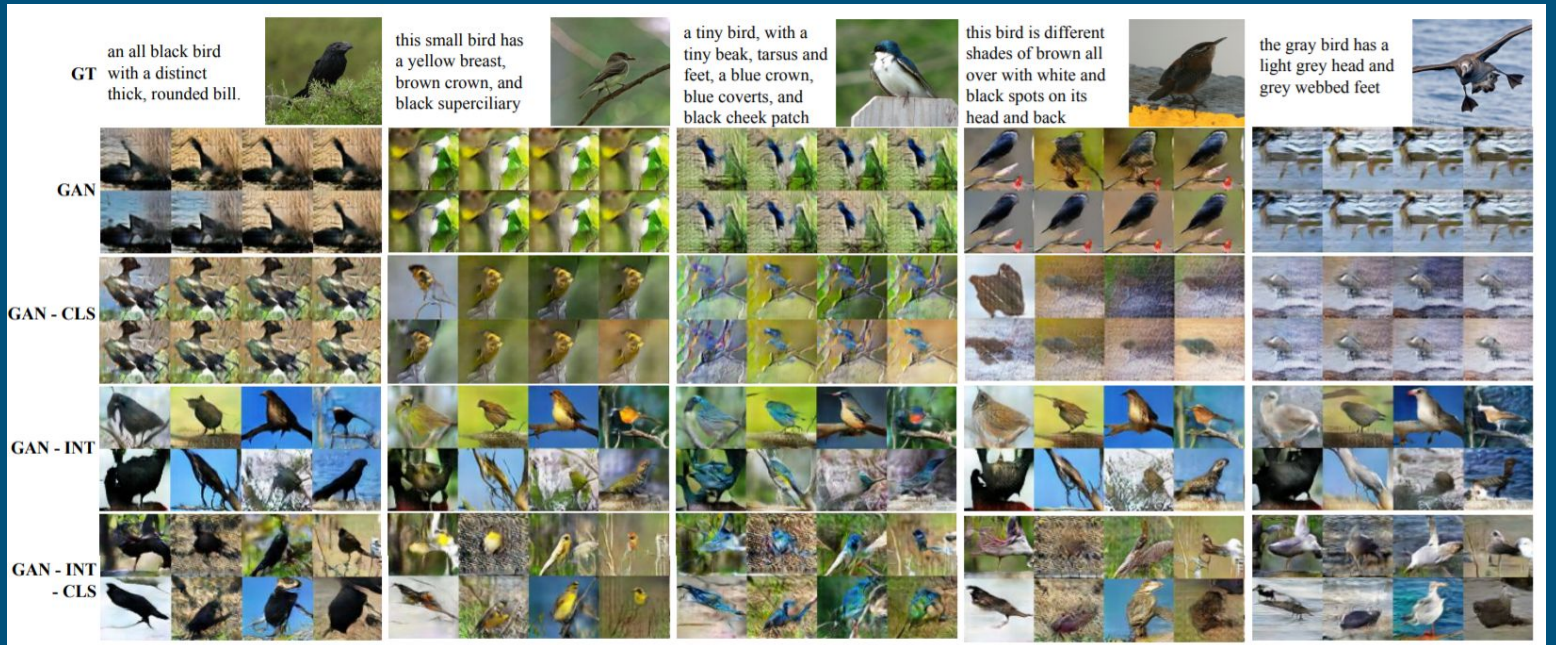


Figure 3. Zero-shot (i.e. conditioned on text from unseen test set categories) generated bird images using GAN, GAN-CLS, GAN-INT and GAN-INT-CLS. We found that interpolation regularizer was needed to reliably achieve visually-plausible results.

Qualitative Results: Oxford-102

Figure 4. Zero-shot generated flower images using GAN, GAN-CLS, GAN-INT and GAN-INT-CLS. All variants generated plausible images. Although some shapes of test categories were not seen during training (e.g. columns 3 and 4), the color information is preserved.

Disentangling style and content

- Since captions provide no information on styles, GAN must learn to use noise sample z to account for style variations.
- Two predictive tasks are set with z as the input:
  - Pose verification
  - Background color verification
- 1) construct similar and dissimilar pairs of images
- 2) compute the predicted style vectors using style encoder
- 3) quantify degree of disentanglement: similarity between images of the same style (e.g. similar pose) should be **higher than** that of different styles (e.g. different pose)

- Pairs of verification is constructed by grouping images into 100 clusters using K-means where images from the same cluster share the same style.
  - Average background color and using 6 keypoint coordinates

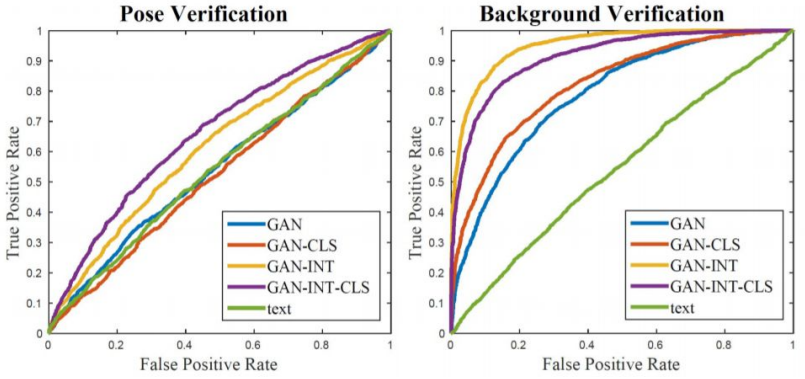


Figure 5. ROC curves using cosine distance between predicted style vector on same vs. different style image pairs. Left: image pairs reflect same or different pose. Right: image pairs reflect same or different average background color.

Compute actual predicted style variables and using cosine similarity to verify. Baseline is cosine similarity between text features from our text encoder.

Pose and background style transfer

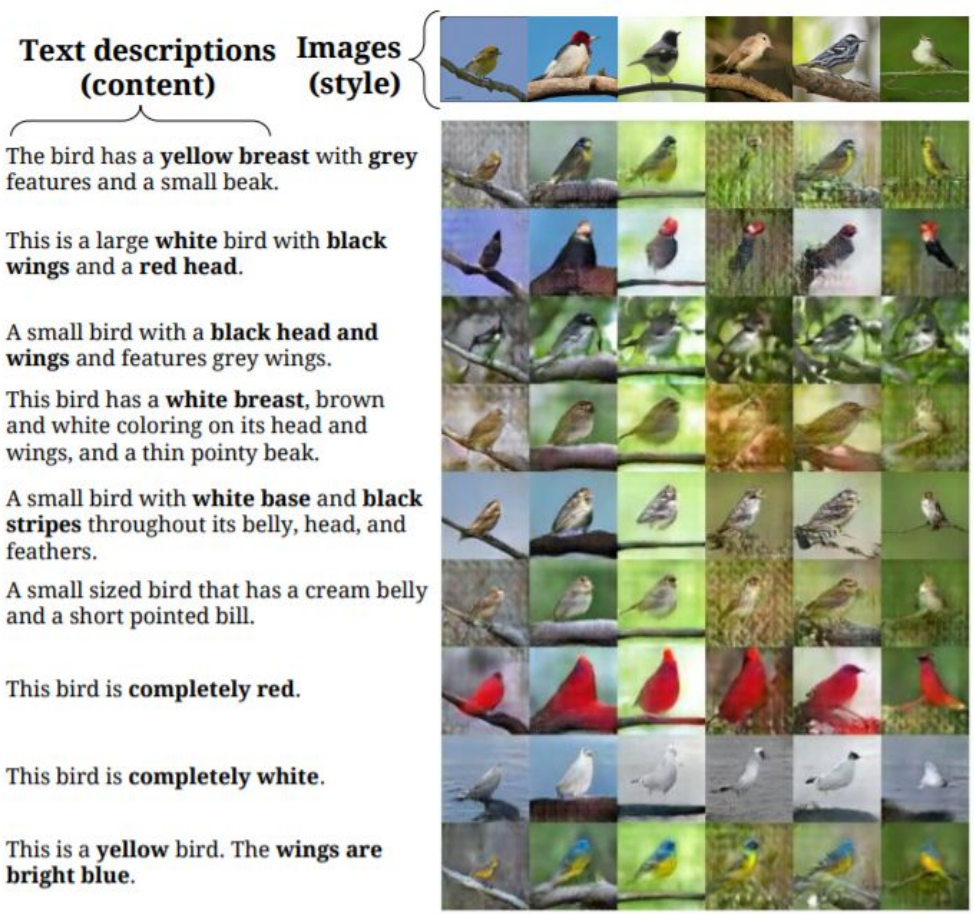


Figure 6. Transfering style from the top row (real) images to the content from the query text, with G acting as a deterministic decoder. The bottom three rows are captions made up by us.

Sentence interpolation

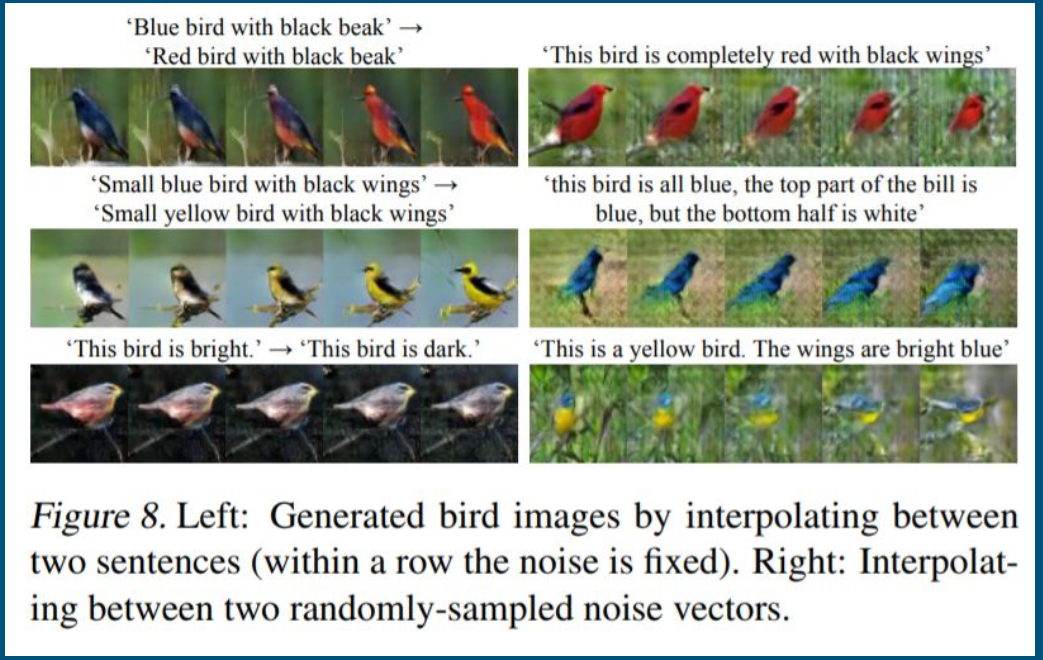


Figure 8. Left: Generated bird images by interpolating between two sentences (within a row the noise is fixed). Right: Interpolating between two randomly-sampled noise vectors.

Results on MS-COCO

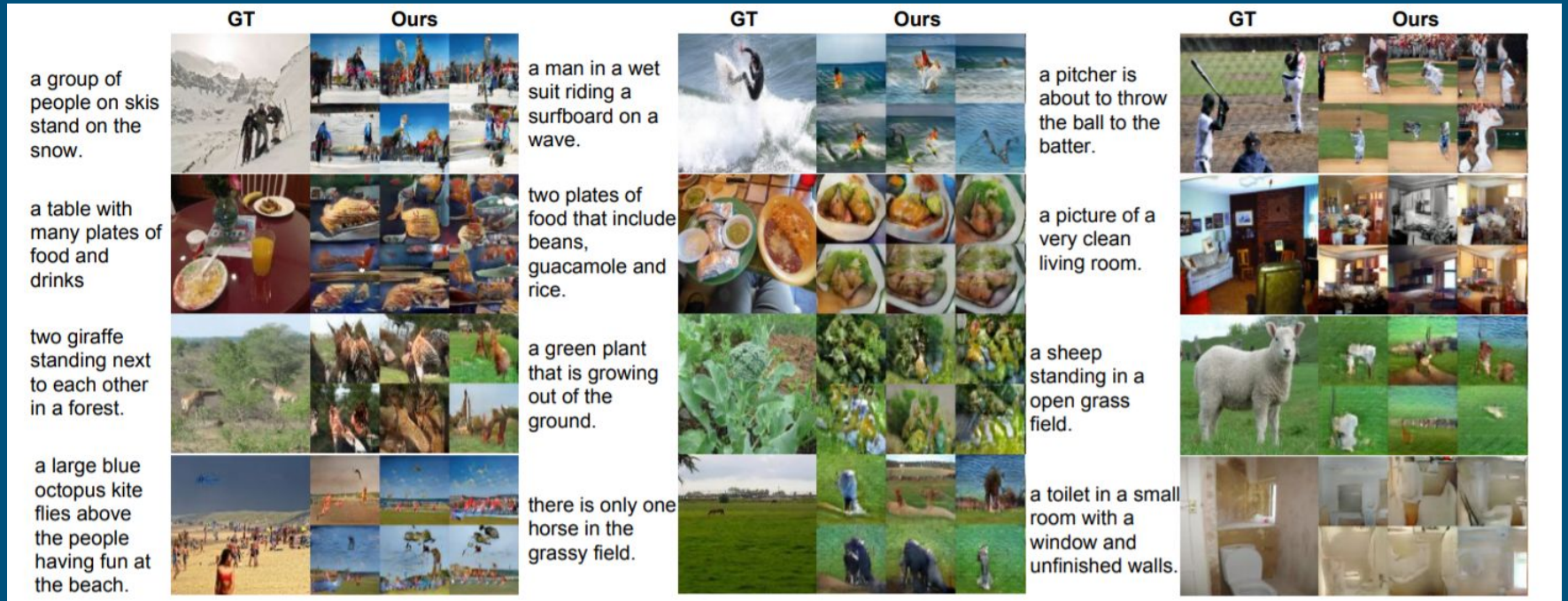


Figure 7. Generating images of general concepts using our GAN-CLS on the MS-COCO validation set. Unlike the case of CUB and Oxford-102, the network must (try to) handle multiple objects and diverse backgrounds.

Conclusion

- Pros:
  - Using GAN to address the problem of multimodal distribution.
  - Able to successfully generate images that are highly related to the text
  - Extended GAN by adding image-text matching discriminator and text manifold interpolation to achieve better results
  - Able to transfer style from real images to the content from the query text
  - Able to generate plausible images when doing sentence interpolation
- Cons:
  - Lack of evaluation metrics
  - Limited success in complex scenes involving human figures
  - The size of generated image is limited at 64 x 64 with no details and vivid object parts.

Extension: GAWWN

Abstract

Generative Adversarial Networks (GANs) have recently demonstrated the capability to synthesize compelling real-world images, such as room interiors, album covers, manga, faces, birds, and flowers. While existing models can synthesize images based on global constraints such as a class label or caption, they do not provide control over pose or object location. We propose a new model, the Generative Adversarial What-Where Network (GAWWN), that synthesizes images given instructions describing what content to draw in which location. We show high-quality 128×128 image synthesis on the Caltech-UCSD Birds dataset, conditioned on both informal text descriptions and also object location. Our system exposes control over both the bounding box around the bird and its constituent parts. By modeling the conditional distributions over part locations, our system also enables conditioning on arbitrary subsets of parts (e.g. only the beak and tail), yielding an efficient interface for picking part locations. We also show preliminary results on the more challenging domain of text- and location-controllable synthesis of images of human actions on the MPII Human Pose dataset.

References

https://arxiv.org/pdf/1610.02454.pdf
https://arxiv.org/pdf/1607.07539.pdf