Inferring Semantic Layout for Hierarchical Text-to-Image Synthesis

Hongzhi Liu

December 28, 2018

CONTENTS

Introduction

1. Motivation

Generating images from text description has been an active research topic in computer vision.

Based on conditional GAN framework, recent approaches improve the prediction quality by generating images with text information.

Existing approaches **have not been successful** in generating reasonable images for complex text descriptions, because of the **complexity** of learning a direct text-to-pixel **mapping** from general images.

H. Zhang, et al. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017.

Introduction

1. Motivation

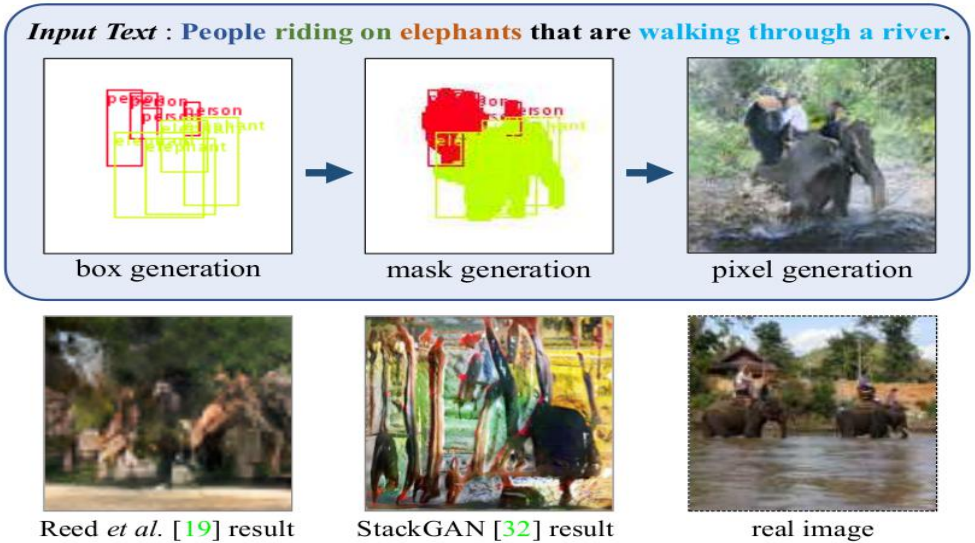

Figure 1. Overall framework of the proposed algorithm. Given a text description, the algorithm sequentially constructs a semantic structure of a scene and generates an image conditioned on the inferred layout and text. Best viewed in color.

S. Reed, et al. Generative adversarial text to image synthesis. In ICML, 2016.

2. Abstract

Instead of learning a **direct mapping** from text to image, the proposed algorithm decomposed the generation process into multiple steps, in which it **first** constructed a semantic layout from the text by the layout generator and converted the layout to an image by the image generator.

The team introduced a novel hierarchical approach for text-to-image synthesis by inferring semantic layout.

3. Related Work

Reed et al. proposed to learn both generator and discriminator conditioned on text embedding. Zhang et al. improved the image quality by increasing image resolution with a two-stage GAN.

Isola et al. proposed a pixel-to-pixel translation network that converts dense pixel-wise labels to an image, and Chen et al. proposed a cascaded refinement network that generates high-resolution output from dense semantic labels.

The most relevant work to the method of Reed et al., which predicted local key-points of bird or human for text-to-image synthesis.

P. Isola, et al. Image-to-image translation with conditional adversarial networks.. In *CVPR*, 2017.

Method

1. Overview

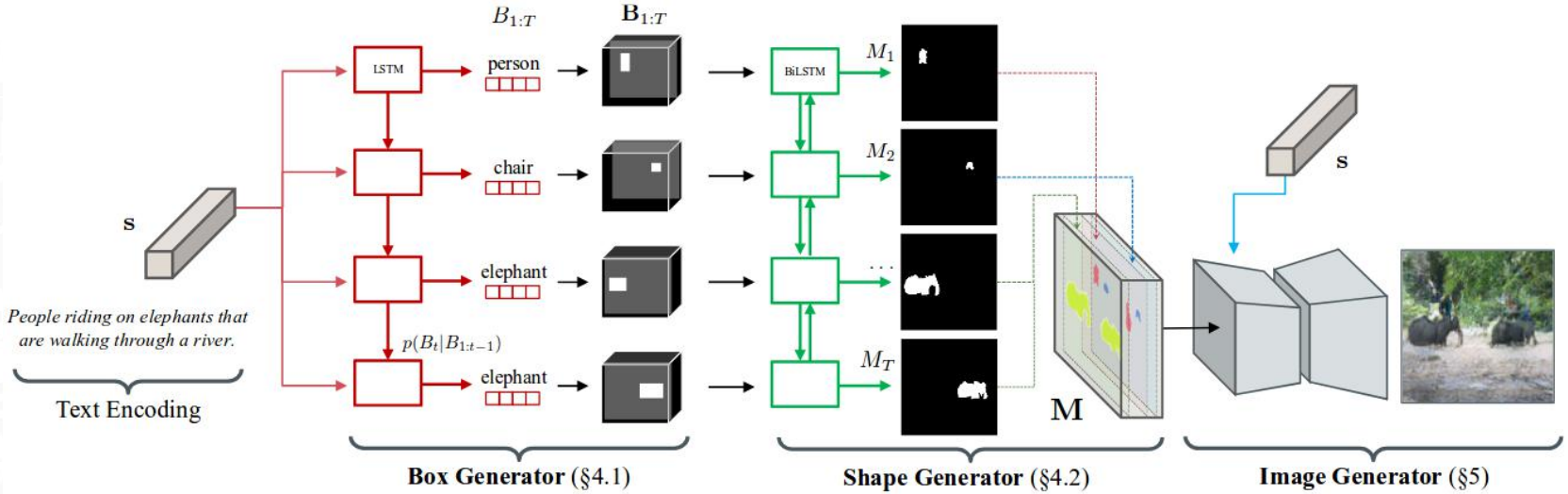

Figure 2. Overall pipeline of the proposed algorithm. Given a text embedding, the algorithm first generates a coarse layout of the image by placing a set of object bounding boxes using the box generator, and further refines the object shape inside each box using the shape generator. Combining outputs from the box and the shape generator leads to a semantic label map defining semantic structure of the scene. Conditioned on the inferred semantic layout and the text, a pixel-wise image is finally generated by the image generator.

2. Bounding Box Generation

Box generator takes a text embedding s as input, and generates a coarse layout by composing object instances in an image. The output of the box generator is a set of bounding boxes $B_{1:T} = \{B_1, ..., B_T\}$, where each bounding box B_t defines the location, size and category label of the t-th object.

They train the box generator by minimizing the negative log-likelihood of ground-truth bounding boxes:

$$\mathcal{L}_{\text{box}} = -\lambda_l \frac{1}{T} \sum_{t=1}^{T} l_t^* \log p(l_t) - \lambda_b \frac{1}{T} \sum_{t=1}^{T} \log p(\mathbf{b}_t^*)$$

where T is the number of objects. b_t and l_t are ground-truth bounding box coordinates and label of the t-th object. The hyper-parameters are set to $\lambda_l = 4$, $\lambda_b = 1$ in the experiments.

3. Shape Generation

Shape generator takes a set of bounding boxes generated from box generator, and predicts shapes of the object inside the boxes. The output of the shape generator is a set of binary masks $M_{1:T} = \{M_1,...,M_T\}$, where each mask M_t defines the foreground shape of the t-th object.

The overall training objective for the shape generator becomes:

$$\mathcal{L}_{\text{shape}} = \lambda_i \mathcal{L}_{\text{inst}} + \lambda_g \mathcal{L}_{\text{global}} + \lambda_r \mathcal{L}_{\text{rec}}$$

where λ_i , λ_g and λ_r are hyper-parameters that balance different losses, which are set to 1, 1 and 10 in the experiment.

3. Shape Generation

They train an instance-wise discriminator D_{inst} by optimizing the following instance-wise adversarial loss L_{inst}

$$\mathcal{L}_{\text{inst}}^{(t)} = \mathbb{E}_{(\mathbf{B}_t, M_t)}\left[\log D_{\text{inst}}(\mathbf{B}_t, M_t)\right] + \mathbb{E}_{\mathbf{B}_t, \mathbf{z}_t}\left[\log\left(1 - D_{\text{inst}}(\mathbf{B}_t, G_{\text{mask}}^{(t)}(\mathbf{B}_{1:T}, \mathbf{z}_{1:T}))\right)\right]$$

The global loss encourages all the instance-wise masks form a globally coherent context.

$$\mathcal{L}_{\text{global}} = \mathbb{E}_{(\mathbf{B}_{1:T}, M_{1:T})}\left[\log D_{\text{global}}(\mathbf{B}_{\text{global}}, M_{\text{global}})\right] + \mathbb{E}_{\mathbf{B}_{1:T}, \mathbf{z}_{1:T}}\left[\log\left(1 - D_{\text{global}}(\mathbf{B}_{\text{global}}, G_{\text{global}}(\mathbf{B}_{1:T}, \mathbf{z}_{1:T}))\right)\right]$$

The perceptual loss, which measures the distance of real and fake images in the feature space of a pre-trained CNN.

$$\mathcal{L}_{\text{rec}} = \sum_{l} \left\| \Phi_l(G_{\text{global}}) - \Phi_l(M_{\text{global}}) \right\|$$

4. Image generator

Image generator takes the semantic label map M obtained by aggregating instance-wise masks, and the text embedding as inputs, and generates an image by translating a semantic layout to pixels matching the text description.

They define the objective function by $L_{img} = \lambda_a L_{adv} + \lambda_r L_{rec}$, where X is a ground-truth image associated with semantic layout M. And they set the hyper-parameters $\lambda_a = 1$, $\lambda_r = 10$ in experiment.

$$\mathcal{L}_{\text{adv}} = \mathbb{E}_{(\mathbf{M},\mathbf{s},X)}\Big[\log D_{\text{img}}(\mathbf{M},\mathbf{s},X)\Big] + \mathbb{E}_{(\mathbf{M},\mathbf{s}),\mathbf{z}}\Big[\log\Big(1 - D_{\text{img}}\big(\mathbf{M},\mathbf{s},G_{\text{img}}(\mathbf{M},\mathbf{s},\mathbf{z})\big)\Big)\Big],$$

$$\mathcal{L}_{\text{rec}} = \sum_{l} \|\Phi_l(G_{\text{img}}(\mathbf{M},\mathbf{s},\mathbf{z})) - \Phi_l(X)\|,$$

4. Image generator

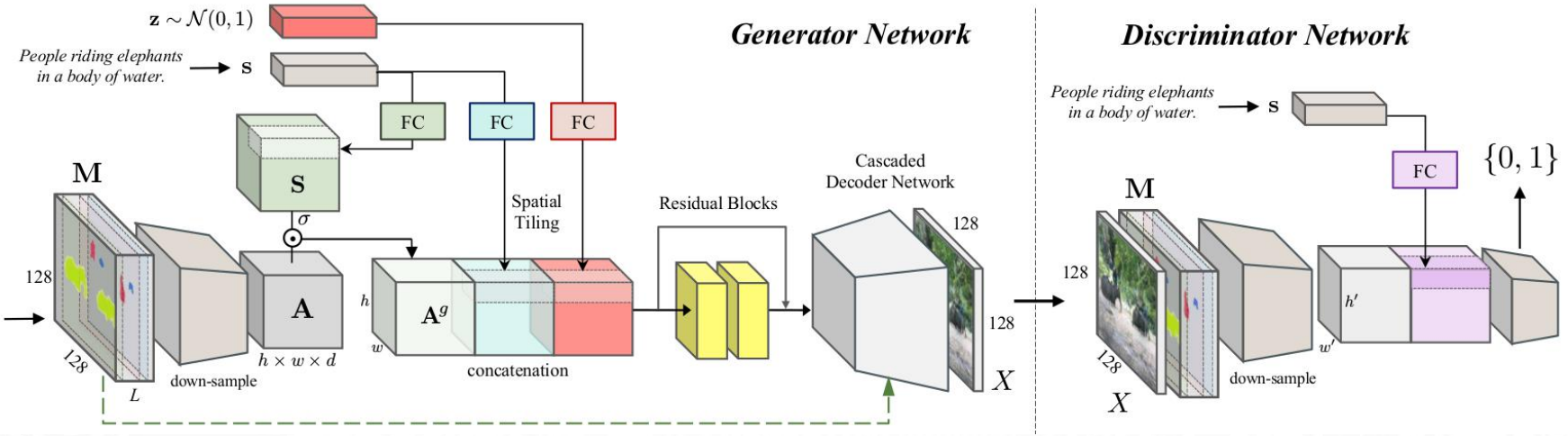

Figure 3. Architecture of the image generator. Conditioned on the text description and the semantic layout generated by the layout generator, it generates an image that matches both inputs.

Experimental Results

1. Experimental Setup

Dataset: MS-COCO 2014 dataset.

It contains 164,000 training images over 80 semantic classes, where each image is associated with instance-wise annotations and 5 text descriptions. They use the official train and validation splits from MS-COCO 2014 for training and evaluating the model.

Evaluation metrics: Inception score, caption generation, and human evaluation.

T.-Y. Lin, et al. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014.

Experimental Results

2. Inception Scores

Method	Box	Mask	Caption generation BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr	Inception [24]
Reed *et al.* [21]	-	-	0.470	0.253	0.136	0.077	0.122	0.160	7.88 ± 0.07
StackGAN [34]	-	-	0.492	0.272	0.152	0.089	0.128	0.195	8.45 ± 0.03
Ours	Pred.	Pred.	**0.541**	**0.332**	**0.199**	**0.122**	**0.154**	**0.367**	**11.46 ± 0.09**
Ours (control experiment)	GT	Pred.	0.556	0.353	0.219	0.139	0.162	0.400	11.94 ± 0.09
	GT	GT	0.573	0.373	0.239	0.156	0.169	0.440	12.40 ± 0.08
Real images (upper bound)	-	-	0.678	0.496	0.349	0.243	0.228	0.802	-

Table 1. Quantitative evaluation results. Two evaluation metrics based on caption generation and the Inception score are presented. The second and third columns indicate types of bounding box or mask layout used in image generation, where “GT” indicates ground-truth and “Pred.” indicates predicted one by the model. The last row presents the caption generation performance on real images, which corresponds to upper-bound of caption generation metric. Higher is better in all columns.

T.-Y. Lin, et al. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014.

3. Human Evaluation

Method	ratio of ranking 1st	*vs.* Ours
StackGAN [34]	18.4 %	29.5 %
Reed *et al.* [21]	23.3 %	32.3 %
Ours	**58.3** %	-

Table 2. Human evaluation results.

Experimental Results

4. Caption Generation

Figure 4. Qualitative examples of generated images conditioned on text descriptions on the MS-COCO validation set, using our method and baselines (StackGAN and Reed et al.). The input text and ground-truth image are shown in the first row. For each method, we provide a reconstructed caption conditioned on the generated image.

Q & A