Source code

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

Tao Xu[1], Pengchuan Zhang[2], Qiuyuan Huang[2], Han Zhang[3], Zhe Gan[2], Xiaolei Huang[1], Xiaodong He[4]
[1]Lehigh University [2]Microsoft Research [3]Rutgers University [4]JD AI Research

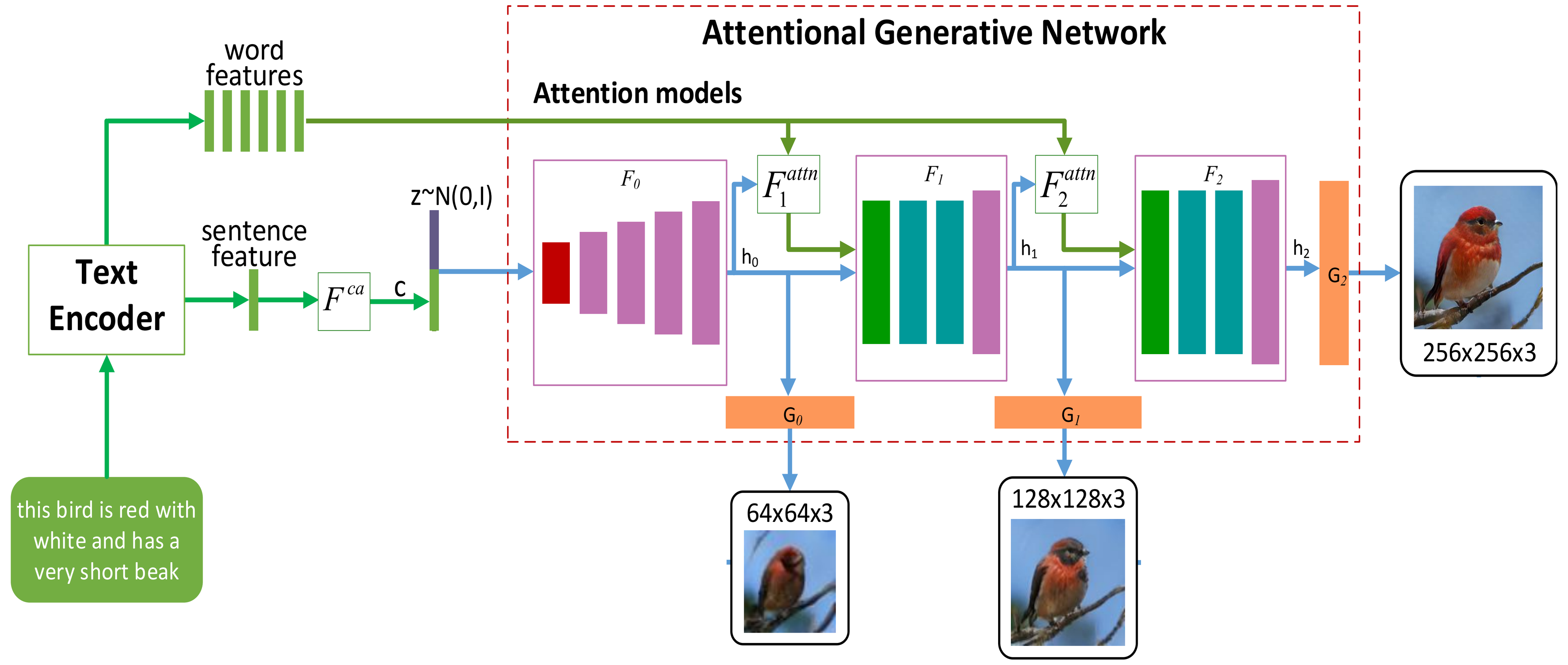
2018 CVPR SALT LAKE CITY • JUNE 18-22

Introduction

- ❖ Automatically generating images according to natural language descriptions is a fundamental problem in many applications, such as art generation and computer-aided design.
- ❖ Current text-to-image GAN models condition only on the global sentence vector which lacks important fine-grained information at the word level and prevents the generation of high quality images.

Our AttnGAN

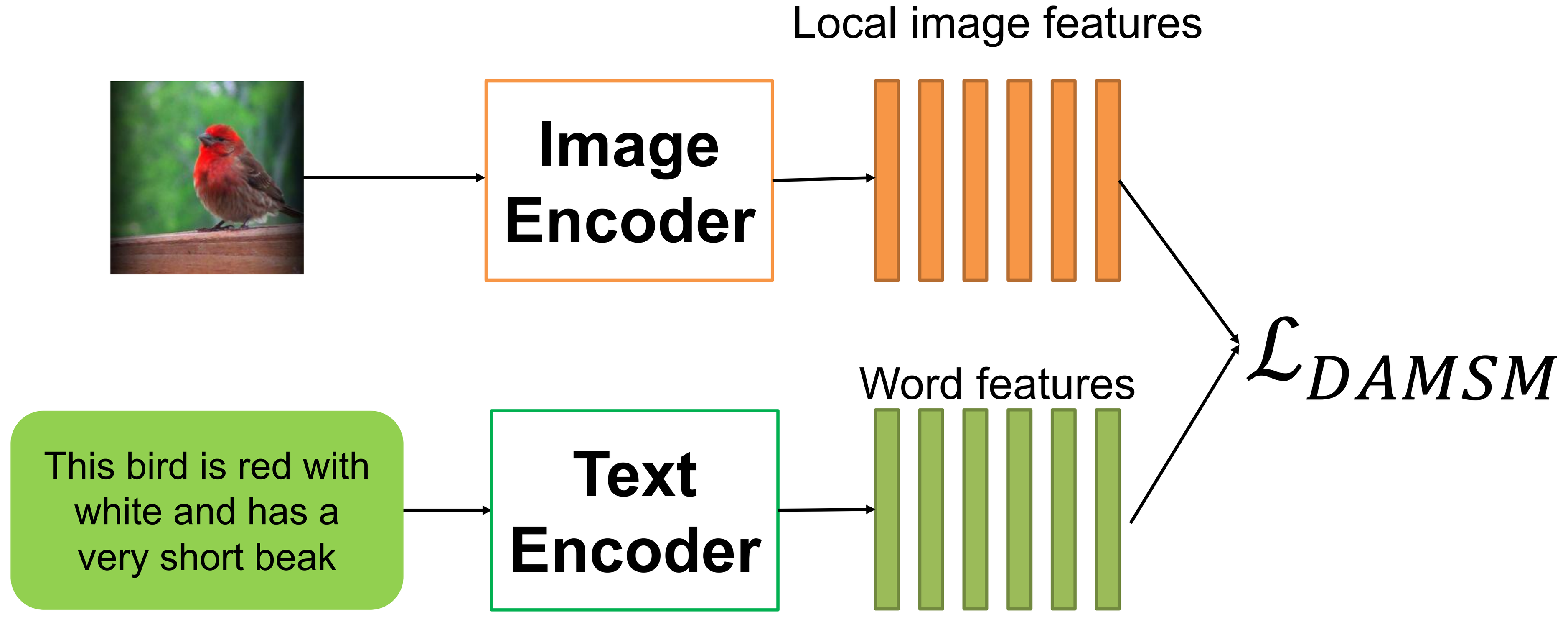
❖ A novel attentional generative network



- ➢ Progressively generate low-to-high resolution images with m generators
- ➢ **Attention model** F^{attn}
  - For each region feature of previous generated image, query its most relevant words.
  - Synthesizes fine-grained details at different sub-regions of the image by paying attentions to the relevant words in the natural language description.
- ➢ **The final objective function**

$$\mathcal{L} = \sum_{i=0}^{m-1} \mathcal{L}_{GAN}^{i} + \lambda \mathcal{L}_{DAMSM}$$

❖ A Deep Attentional Multimodal Similarity Model (DAMSM)



- ➢ **Text encoder** (LSTM) extracts word features $e_1, e_2, \dots, e_T$
- ➢ **Image encoder** (CNN) extracts image region features $v_1, v_2, \dots, v_N$
- ➢ **Attention mechanism:** for the i-th word, compute its region-context vector c_i,

$$c_i = \sum_{j=0}^{N-1} \alpha_j v_j, \qquad \text{where } \alpha_j = \frac{\exp(\gamma_1 \bar{s}_{i,j})}{\sum_{k=0}^{N-1} \exp(\gamma_1 \bar{s}_{i,k})}$$

- $\bar{s}_{i,j}$ is the dot product between features of the i-th word and the j-th image region

- ➢ **The similarity** between the image (Q) and the sentence (D)

$$R(Q,D) = \log\left(\sum_{i=0}^{T-1} \exp\big(\gamma_2 R(c_i, e_i)\big)\right)^{\frac{1}{\gamma_2}}$$

- $R(c_i, e_i)$ is the cosine similarity between c_i and e_i

- ➢ **The negative log posterior probability** that the images are matched with their ground truth text descriptions

$$\mathcal{L}_{DAMSM} = -\sum_{i=1}^{M} \log P(D_i|Q_i), \text{where } P(D_i|Q_i) = \frac{\exp\big(\gamma_3 R(Q_i, D_i)\big)}{\sum_{j=1}^{M} \exp\big(\gamma_3 R(Q_i, D_j)\big)}$$

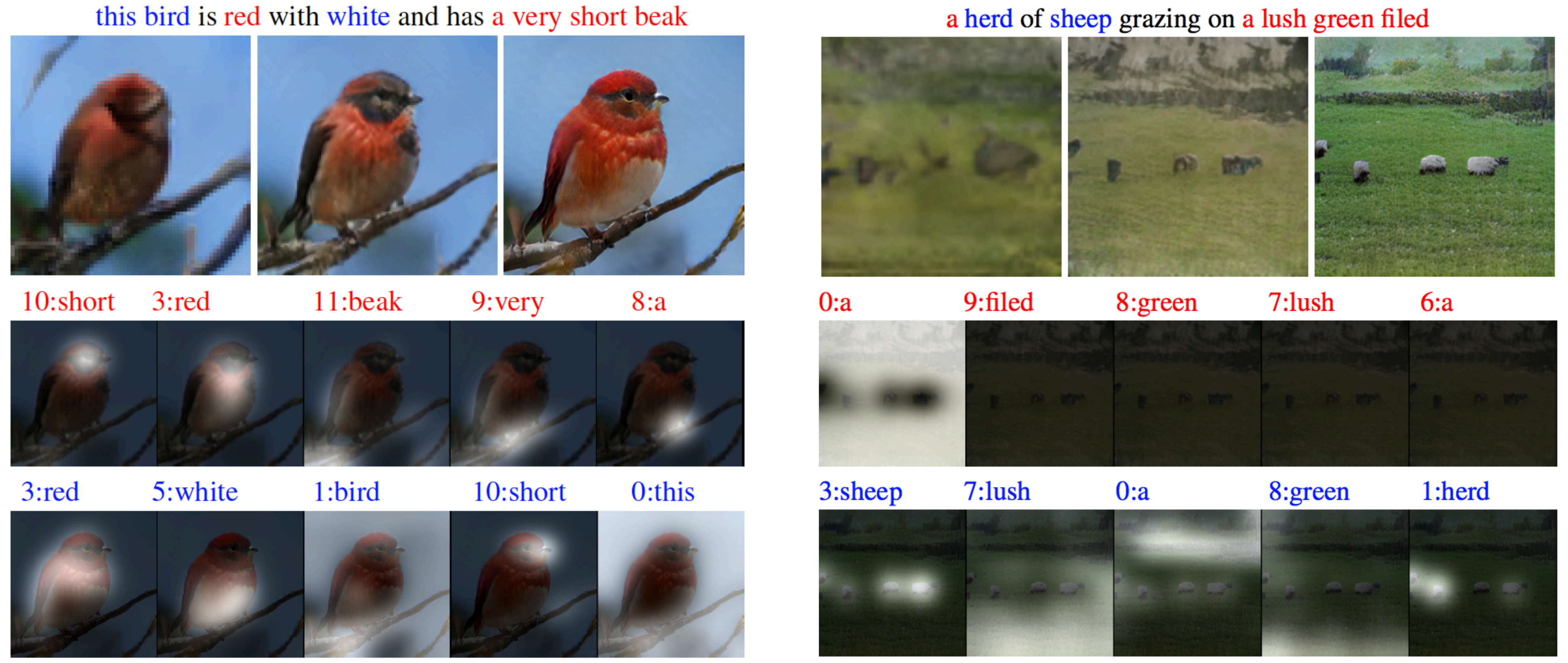
- M is the number of training pairs
- $\lambda, \gamma_1, \gamma_2$ and γ_3 are hyper-parameters
- The $\mathcal{L}_{DAMSM}$ provides a fine-grained image-text matching loss for training the generator

Results

- ❖ The DAMSM loss is important
- ❖ Stacking more attention models helps

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

- ❖ Attention maps on CUB (left) and COCO (right)



- ❖ Novel images on CUB (left) and COCO (right)

- ❖ **Compare** with state-of-the-art

Dataset	GAN-INT-CLS	GAWWN	StackGAN	StackGAN-v2	PPGN	Our AttnGAN
CUB	2.88 ± .04	3.62 ± .07	3.70 ± .04	3.82 ± .06	\	**4.36 ± .03**
COCO	7.88 ± .07	\	8.45 ± .03	\	9.58 ± .21	**25.89 ± .47**

- ❖ **Generalize** the proposed attention mechanisms to DCGAN framework
  - ➢ Vanilla DCGAN on CUB: 2.47 inception score 3.69% R-precision
  - ➢ Our AttnDCGAN on CUB: **4.12** inception score **38.45%** R-precision