Speak Out and Get Your Own Bird

Yifan Li, Andong Ma, Risheng Wang, Changsheng Yan and Yi Xiao

SFU Simon Fraser University

INTRODUCTION

Topic:
Audio to Image Synthesis with Intermediate Text Representation

Procedure:

- Audio to Text Conversion (package: speech recognition)
  Develop a Graphical User Interface for interaction and connection
- Text to Image Generation through AttnGAN
  Propose the idea of adding another stage of attention model in order to generate images with higher resolution.
- Apply both pre-trained original AttnGAN model and our re-trained refined AttnGAN model on CUB dataset

Expected Outcome:
We aim at achieving the goal of audio to image synthesis with generating HIGHER quality images(512x512x3) by improving the original AttnGAN, from which the image generated is 256x256x3. Due to the constraints of our computers and laptops, training process may be long and the images shown below are not the best ones that we could get.

MODEL FORMULATION

Background

GANs consist of a generator G and a discriminator D that compete in a two-player minimax game. Concretely, D and G play the following game on V(D,G):

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

Attentional Generative Network

$$h_0 = F_0(z, F^{ca}(\bar{e}));$$
$$h_i = F_i(h_{i-1}, F_i^{attn}(e, h_{i-1})) \text{ for } i = 1, 2, ..., m-1;$$
$$\hat{x}_i = G_i(h_i).$$

G_i: *generator*
h_i: *hidden states*
e: *word feature*
z: *noise vector*

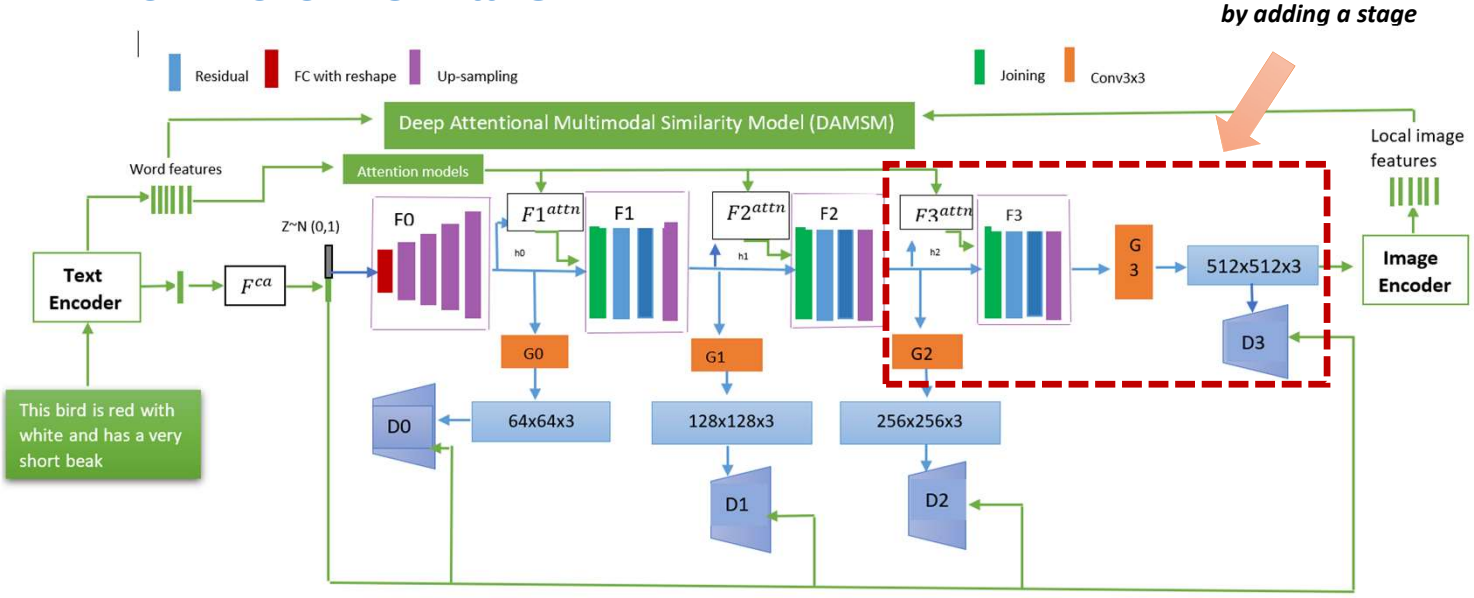
Deep Attentional Multimodal Similarity Model

The DAMSM learns two neural networks that map sub-regions of the image and words of the sentence to a common semantic space, thus measures the image-text similarity at the word level to compute a fine-grained loss for image generation.

- the first row gives 64×64 images by G_0. 128×128 images by G_1 and 256×256 images by G_2;
- the second and third row shows the top-5 most attended words by F^{attn} of the AttnGAN.

THE ARCHITECTURE OF AttnGAN

We refined AttnGAN by adding a stage



EXPERIMENTS (Now it's your turn to have a try)

Shh... It's our magic time
Image generated by original AttnGAN

Not the best, only better
Image generated by our refined AttnGAN