Deep Conditional Text to Image Generation

Course Name : Computer Vision
Course ID : CS7.505
Instructors : Prof. Anoop Namboodiri
Assigned TA : Avinash Prabhu

Team: Unnamed

Zeeshan Khan (2021701029)
K. N. Amruth Sagar(2021701013)
Seshadri Mazumder(2021801002)

Project Link : CV Project

Index

Introduction

Generation of images from the give text description is a multimodal problem, where there are two tasks involved.

- The first one is to ensure that all the characteristics of the object that we want to generate and the attributes of the object we want are properly encoded into some form of textual representation.
- The second task is learn a mapping function which uses this encoded vector with information about object to generate an image. Advances in deep learning techniques has provided ways to tackle this problem as we possess models which provide good feature representation and also are capable of generating images from noise. Still, this problem far from completely solved.

Problem Statement

Given a single-sentence human-written description $S = (w_1, w_2, w_3 \cdots w_n)$

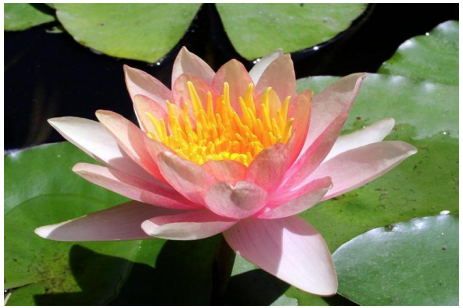
learn a text feature representation that captures all the visual features present in the sentence and use these features to synthesize a compelling image that a human might mistake for real.

Datasets

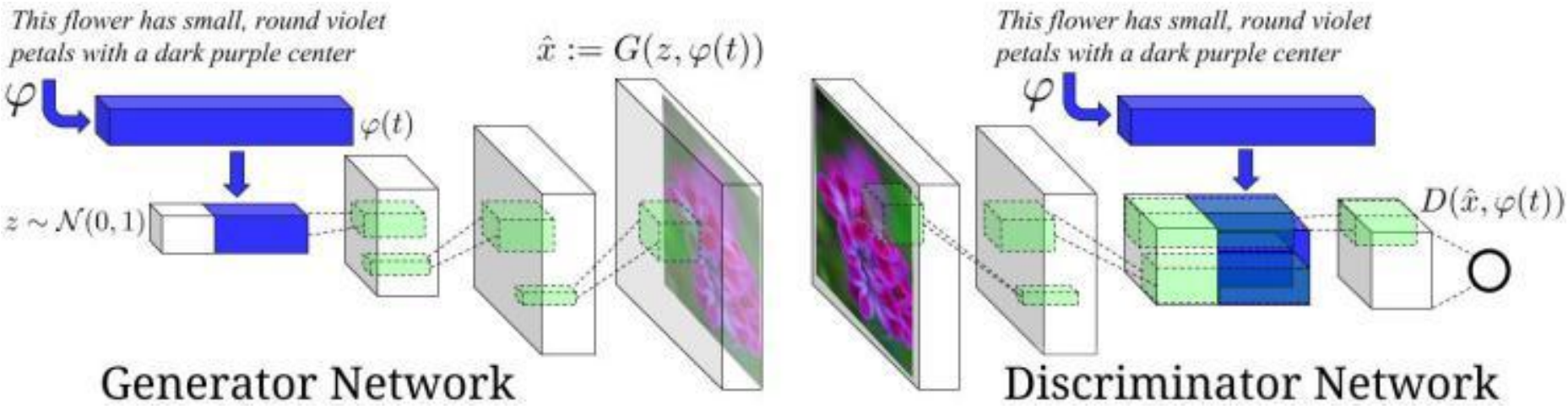
We have used the 102 Category Flower dataset for the task.

Stats:

- Images per category of flower : images per category
- No of images per category : 40 to 258
- No of description per image : 10



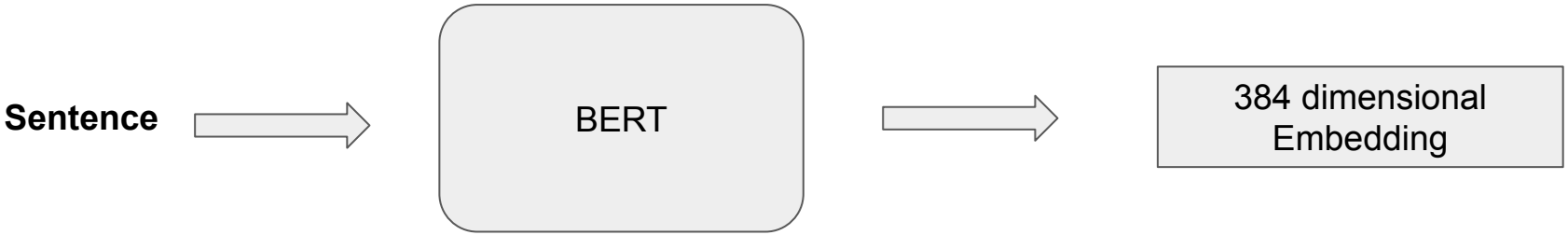
Pipeline



Text to image DC-GAN

Data Preprocessing

- All the images were resized to 64*64.
- For each image, 10 textual description were selected.
- Then, we get the embeddings of the textual descriptions using BERT, and they are saved separately.



Generator and discriminator Model Architecture

- Generator
  - generator

- Discriminator
  - discriminator

- Embedding+noise is fed into generator, which generates an image
- In discriminator, we again concatenate the embedding of the textual description in the last layers.
- This is to condition the output of the generator on the input sentence.

Loss

- **KL Divergence**
  - **Kullback-Leibler Divergence (KL divergence), or relative entropy, quantify the difference between probability distributions for a given random variable.**
- **LPIPS**
  - **Learned Perceptual Image Patch Similarity (LPIPS) metric. It uses deep network activations and evaluates the distance between image patches. Higher means more different. Lower means more similar.**
- **GAN discriminator (Real + Fake)**
  -

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D\left(\boldsymbol{x}^{(i)}\right) + \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right) \right]$$

Qualitative Evaluation : Sample 1

Predicted Image

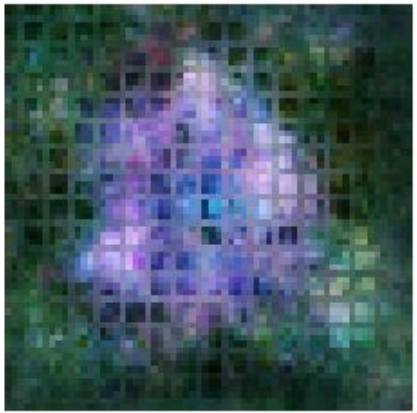
Target Image

Source Sentence :

- **this flower is orange and white in color, and has petals that are short and pointed.**
- **there are many overlapping pale orange petals, with very prominent green stigmas in the center and many branching sepals around the petals.**
- **a flower with white and orange petals and green anther filaments and pistils.**

Qualitative Evaluation : Sample 2

Predicted Image



Target Image

Source Sentence :

- **petals are light purple in color with longer stamens and a prominent green pistil**
- **this blue and purple flower has layered pointed petals and a dark green stigma.**
- **this flower is purple and blue in color, with petals that are oval shaped.**

Qualitative Evaluation : Sample 3

Predicted Image

Target Image

Source Sentence :

- **this flower has orange and white petals, and several yellow anthers.**
- **this flower has petals that are pink with green stamen**
- **the petals of this flower are pink with a long stigma**

References

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48 (ICML'16). JMLR.org, 1060–1069.