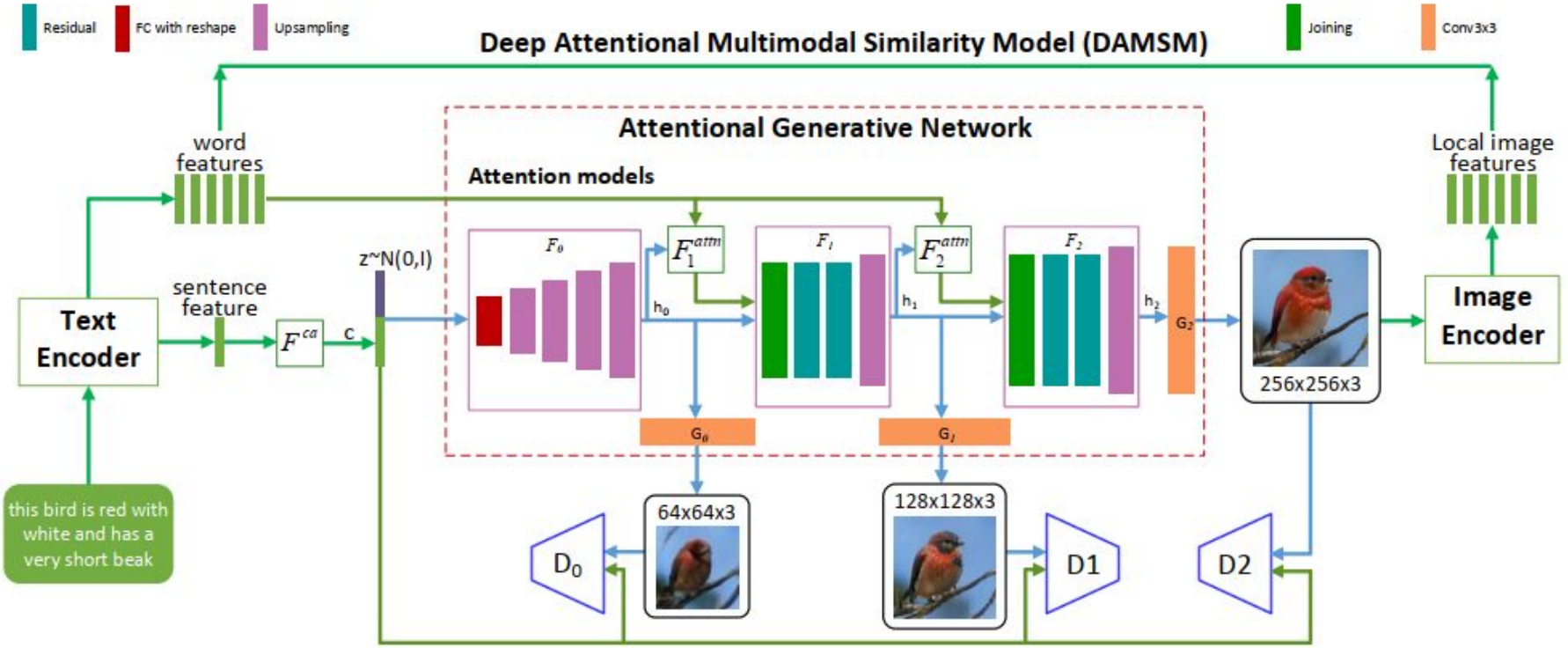
Fine grained Text to Image Generation with Attentional Generative Networks

Prakhar Shankar - 234161007
Pragati Mangalsing Patil - 234161019
Gayathri Gutla-234161016

Introduction

Generating high-quality images from text descriptions is essential for various applications such as art generation and Image captioning. While recent approaches based on Generative Adversarial Networks (GANs) have shown promising results, they often lack fine-grained information at the word level, hindering the generation of detailed images, especially for complex scenes. This limitation poses a significant challenge, particularly in datasets like CUB.

Architecture



Datasets

Our model is evaluated and trained on CUB dataset - The Caltech-UCSD Birds-200-2011 (CUB-200-2011) dataset.

It contains 11,788 images of 200 subcategories belonging to birds.

Training set - 8855 images (75%)

Test set - 2933 images (10%)

No. of captions per image used - 10

Attentional Generative Network

Attentional generative network has m generators (G0, G1, …. , Gm-1), which take the hidden states (h0, h1, … ,hm-1) as input and generate images of small-to-large scales.

The attention model F has two inputs: the word features and the image features from the previous hidden layer h.

The **word features** are first converted into the common semantic space of the image features by adding a new perceptron layer. Then, a word-context vector is computed for each sub-region of the image based on its hidden features h (query).

Finally, **image features** and the corresponding word-context features are combined to generate images at the next stage.

Deep Attentional Multimodal Similarity Model(DAMSM)-:

Text Encoder: **bi-directional Long Short Term Memory(LSTM)** - This part is responsible for encoding text descriptions into a feature vector. The output is word embeddings and sentence embeddings. We have three layers in it namely given as Embedding layer, Dropout layer, and LSTM.

Attention Models: This part includes two attention models that focus on specific regions of the image based on the encoded text description.

Image encoder: a **Convolutional Neural Network (CNN)** that maps images to semantic vectors. The intermediate layers of the CNN learn local features of different sub-regions of the image, while the later layers learn global features of the image. More specifically, image encoder is built upon the Inception-v3 model pretrained.

Deep Attentional Multimodal Similarity Model (DAMSM): This part uses the attention weights and features from other parts to compute a multimodal similarity score.

- The text encoder takes a text description as input and encodes it into a feature vector.
- The two attention models then focus on specific regions of the image based on the encoded text description.
- The local image features are then extracted from the specific regions of the image that the attention models focused on.

Results

Input by the user:

```
|                          Welcome to our platform for generating images from text, designed with simplicity in mind!                          |
```

Generate!!

```
********** "Prepare yourselves. Behold our exquisite platform, tailored for the transformation of textual essence into resplendent imagery." **********
```

Console Output | Final Image 256x256 | Intermediate Image 128x128 | Low resolution image 64x64

```
SEEEEEEEEEEEEEEEE ../output/birds_attn2_2024_04_26_14_52_07
/content/drive/MyDrive/IPML/AttnGAN-master/data/birds/CUB_200_2011/bounding_boxes.txt
Total filenames:  11788 001.Black_footed_Albatross/Black_Footed_Albatross_0046_18.jpg
--------- /content/drive/MyDrive/IPML/AttnGAN-master/data/birds/test
Load filenames from: /content/drive/MyDrive/IPML/AttnGAN-master/data/birds/train/filenames.pickle (8855)
Load filenames from: /content/drive/MyDrive/IPML/AttnGAN-master/data/birds/test/filenames.pickle (2933)
Load from:  /content/drive/MyDrive/IPML/AttnGAN-master/data/birds/captions.pickle
NUMBER OF WORDSSSS 5450
path..... ../output/birds_attn2_2024_04_26_14_52_07
Enter your sentences (separated by newline):
a white color bird with a red long beak
SENTENCES..... ['a white color bird with a red long beak']
TOKENS..... ['a', 'white', 'color', 'bird', 'with', 'a', 'red', 'long', 'beak']
/usr/local/lib/python3.10/dist-packages/torch/nn/modules/rnn.py:83: UserWarning: dropout option adds dropout after all but last recurrent layer, so non-zero dropout expects
num_layers greater than 1, but got dropout=0.5 and num_layers=1
  warnings.warn("dropout option adds dropout after all but last "
Load text encoder from: /content/drive/MyDrive/IPML/AttnGAN-master/DAMSMencoders/bird/text_encoder200.pth
Load G from:  /content/drive/MyDrive/IPML/AttnGAN-master/models/bird_AttnGAN2.pth
/content/drive/MyDrive/IPML/AttnGAN-master/code/trainer.py:474: UserWarning: volatile was removed and now has no effect. Use `with torch.no_grad():` instead.
  captions = Variable(torch.from_numpy(captions), volatile=True)
/content/drive/MyDrive/IPML/AttnGAN-master/code/trainer.py:475: UserWarning: volatile was removed and now has no effect. Use `with torch.no_grad():` instead.
  cap_lens = Variable(torch.from_numpy(cap_lens), volatile=True)
/content/drive/MyDrive/IPML/AttnGAN-master/code/trainer.py:480: UserWarning: volatile was removed and now has no effect. Use `with torch.no_grad():` instead.
  noise = Variable(torch.FloatTensor(batch_size, nz), volatile=True)
```

Output 1 - 64x64 low resolution image

Welcome to our platform for generating images from text, designed with simplicity in mind!

Generate!!

********** "Prepare yourselves. Behold our exquisite platform, tailored for the transformation of textual essence into resplendent imagery." **********

Console Output | Final Image 256x256 | Intermediate Image 128x128 | **Low resolution image 64x64**

Output 2 - 128x128 intermediate resolution image

Welcome to our platform for generating images from text, designed with simplicity in mind!

Generate!!

********** "Prepare yourselves. Behold our exquisite platform, tailored for the transformation of textual essence into resplendent imagery." **********

Console Output | Final Image 256x256 | **Intermediate Image 128x128** | Low resolution image 64x64

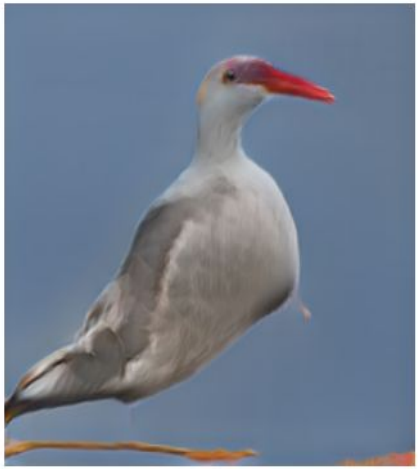
Output 3 - 256x256 high resolution image

```
------------------------------------------------------------------------------------------------------------------------------------------------------
|                                  Welcome to our platform for generating images from text, designed with simplicity in mind!                        |
------------------------------------------------------------------------------------------------------------------------------------------------------
```

Generate!!

********** "Prepare yourselves. Behold our exquisite platform, tailored for the transformation of textual essence into resplendent imagery." **********

Console Output | **Final Image 256x256** | Intermediate Image 128x128 | Low resolution image 64x64



this bird has wings that are black and has a white belly

this bird has wings that are red and has a yellow belly

this bird has wings that are blue and has a red belly

Inception Score of the model: 4.36

CHALLENGES

- **Fine-Grained Details:**
  - Capturing fine-grained details accurately from text descriptions, especially in complex scenes, remains a challenge.
- **Training Stability:**
  - Ensuring stable training of the AttGAN model, particularly with multi-stage refinement and attention mechanisms, can be challenging.
- **Evaluation Metrics:**
  - Developing appropriate evaluation metrics beyond visual inspection for assessing the quality of generated images. But we tried getting an inception score for our model.
- Getting google colab for the 3rd Time.

Conclusions

We tried building an attentional generative network for the AttnGAN to generate high quality image through a multi-stage process.

Extensive experimental results clearly show that the AttnGAN is actually helpful and succeeds critically for text-to-image generation for complex scenes.

References

S. E. Reed, A. van den Oord, N. Kalchbrenner, S. G. Colmenarejo,

Z. Wang, Y. Chen, D. Belov, and N. de Freitas. Parallel multiscale

autoregressive density estimation. In ICML, 2017.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma,

Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and

L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge.

IJCV, 115(3):211–252, 2015.

T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford,

and X. Chen. Improved techniques for training gans. In NIPS, 2016.

Thank You