Multi-instance Text-to-Photo Image Generation Using Stacked Generative Adversarial Networks

Alex Fu | Yiju Hou
CS 231N Class Project, Spring 2017

INTRODUCTION

Image Generation

- Unsolved challenging problem in Computer Vision
- Potential applications: photo editing, video generation and digital design

Current State-of-Art

- Generate high-resolution images of a single instance of birds or flowers using StackGan
- Generate Images from captions with attention by extending DRAW

PROBLEM & DATASET

Task

Generate multi-instance images from multiple categories by interpreting the given text description.

Approach

- Modularized deep neural network based on TensorFlow and PyTorch
- Experiment with various convolutional neural network architectures, text encoders, decoders, attention mechanism, etc.

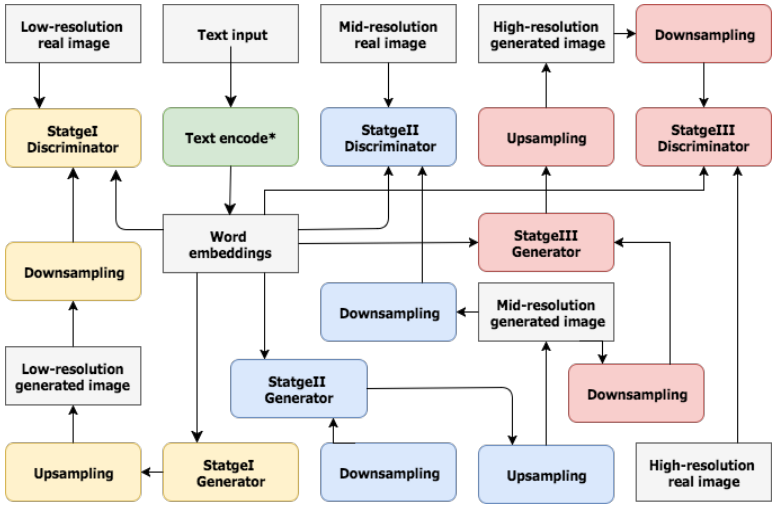
COCO Dataset

- Largest publicly available recognition, segmentation, captioning dataset.
- 80 categories, 300,000+ images.
- Multiple objects per image.
- More than 300,000 images.

Evaluation

Inception Score

MODEL



EXPERIENMENTS & TRAINING

Text

Text Encoder	Decoder
Char-CNN-RNN	Char-CNN-RNN
Skipthoughts	Skipthoughts
Match-LSTM	Ans-Ptr
Coattention	LSTM
Coattention	Ans-Ptr

Optimizations

- Dropout
- Trainable end-of-sentence sentinel
- Flexible RNN sequence length

Experiments

- Residual blocks
- Conditional augmentation
- Additional GAN stage to increase the resolution of generated images
- Feed additional COCO classification data to create word imbedding
- LR exponential annealing and manual tuning

RESULTS & CONCLUSIONS

Training

- Decreasing generator loss verifies model potential
- Linear loss drop indicates insufficient LR
- Gap between Train/Val indicates overfitting
- For dropout to improve the performance on Val, we need to reconsider its location and rate

Current Results

- The output images of stageI do not have clear composition and distinct object shapes as the results in training on CUB and Oxford-102 datasets
  - Hypnosis 1: Suffer from insufficient LR
  - Hypnosis 2: The problem will be solved by adding one extra GAN layer
  - Hypnosis 3: The current architecture is not expressive enough to capture the complexity of COCO dataset

FUTURE DIRECTIONS

Architecture

- Experiment with more expressive encoders, decoders and attention mechanism
- Increase the complexity of the model to better capture the interaction of objects
- Use attention mechanism to determine the pixel size of generated objects, so that the model can generate high-quality objects in the foreground

Training

- Use larger embeddings and better methods to preprocess text data
- Identify the right places to apply dropout
- Hyper-parameter fine tuning:
  - Dropout rate
  - Learning rate