IMAGAN: Learning Images from Captions

Samir Sen, Trevor Tsue, Karan Singhal | {samirsen, ttsue, ksinghal}@stanford.edu

Motivation

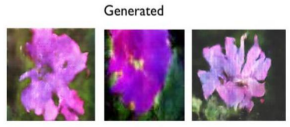
- Our task is to generate new images of flowers from text captions of flowers
- We were inspired by the idea that a machine could learn to *imagine*
- We hope this work opens avenues toward a more complete understanding of textual information

Data

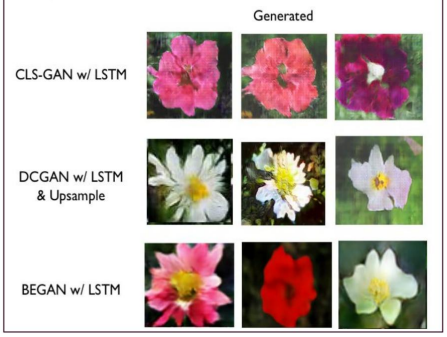
- **Oxford Flowers Dataset**
  - 8189 images of flowers
  - 5 human-labeled captions per flower
  - 17 overall categories of flowers represented
- **GloVe Word Embeddings**
  - 300 dim representation of individual words
  - trained on 6B words from Gigaword5 + Wikipedia2014
- **Skipthoughts**
  - 4800-dim learned representation of sentences
  - Trained on ~900 million words from BookCorpus

Results

Baseline

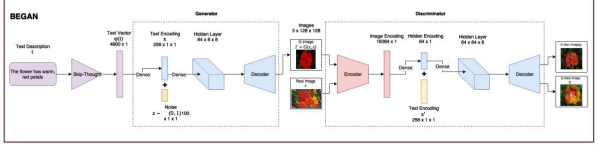


Experiments

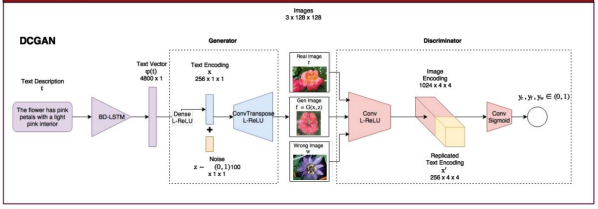


Generative Adversarial Network (GAN) Architecture

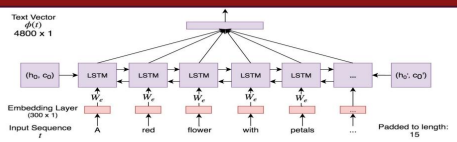
BEGAN



DCGAN



LSTM



Loss and Backpropagation

	L(Discriminator)	L(Generator)	L(LSTM)
CLS-GAN	$\mathcal{L}_D \leftarrow \log(y_r) + \log(1 - y_w) + \log(1 - y_f)$	$\mathcal{L}_G \leftarrow \log(y_f)$	$\mathcal{L}_G \leftarrow \log(y_f)$
BeGAN	$\mathcal{L}_D \leftarrow \mathcal{L}(y_r, r) - k\mathcal{L}(y_f, f)$	$\mathcal{L}_G \leftarrow \mathcal{L}(y_f, f)$	$\mathcal{L}_G \leftarrow \mathcal{L}(y_f, f)$

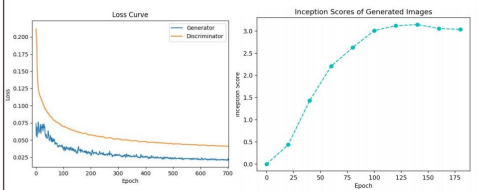
Future Work

- Explore different model architectures like StackGAN, AttnGAN
- Spend more time tuning architecture and hyperparameters to produce better images
- Video reconstruction from text

Analysis

Model Performances

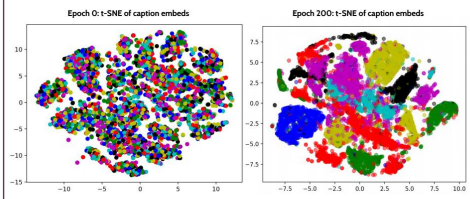
- We perform a variety of architectural enhancements from our baseline, including *Boundary Equilibrium* (BEGAN), *Deep Convolution* (DCGAN)



BeGAN achieves moderately high score (~3.0). We hypothesize bottleneck is likely ability to generate finer edges

Learned Embedding Representations

- We performed t-SNE dimensionality reduction on learned caption features over the categories of flowers generated



Training Techniques

LSTM

- Initialize hidden states to small random values
- Batch size = 64
- Bidirectional, 1-layer
- Xavier weights initialization

GAN

- Dataset augmentation with random crops + flips
- Gaussian noise for generator
- Normal initialization
- Better resolution/eliminated checkerboard with upsampling
- One-sided label smoothing
- Leaky ReLU instead of ReLU
- SGD instead of Adam for D
- Tanh for output
- Batchnorm
- Learning rate decay
- Average pool instead of max pool
- ELU instead of ReLU