Generating Stylized Images from Captions: Generative Styled Network

Stephanie Dong, Jeff Chen, Nick Guo

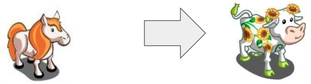
{sxdong11, jc1, nickguo}@stanford.edu - June 2018

Motivation

- Graphic artists need to create images from text captions for use within software programs
- Lots of repetitive and wasted work since many assets are discarded
- GANs have shown potential to generate images [1]
- Can we train a computer to produce original images in a particular style?

Input: **"A white cow with yellow flowers"**

Output



We wish to create a system that takes as input a text caption along with some style images and output an image that fit both the caption and the style of the input.

Datasets

- Generation: MS-COCO
- 118k training images, 5k validation images
- 80 fine-grained categories
  - Focus on 3 categories: elephant, laptop, train

- Style: iPhone Mafia Game Assets
- 3,000 images

Metrics & Baseline

- **Inception Score:**

$$\exp(\mathbb{E}_x \mathbf{KL}(p(y|\mathbf{x})||p(y)))$$

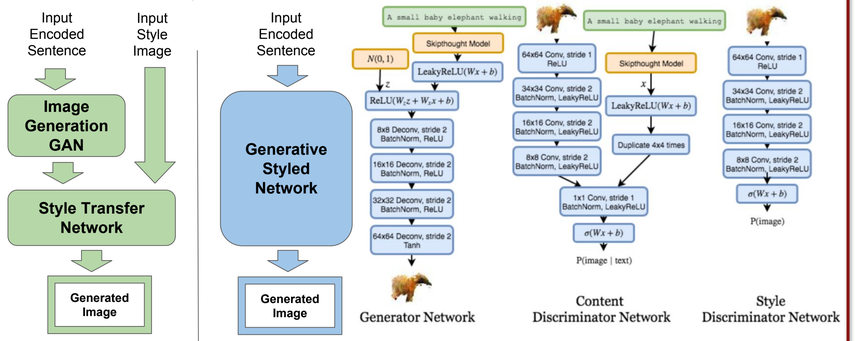
  - Measures both quality and diversity of generated images [3]
  - Metric tested and validated against human intuition for 20+ experiments

- **Baseline Evaluation (Inception Score: 2.49)**
  - MS-COCO images used as-is without any processing

Baseline image generation is incomprehensible. First we evaluated the baseline on a multi-category model (left), then we evaluated the baseline for single category models elephant, laptop, train. (3 images on the right)

Methods

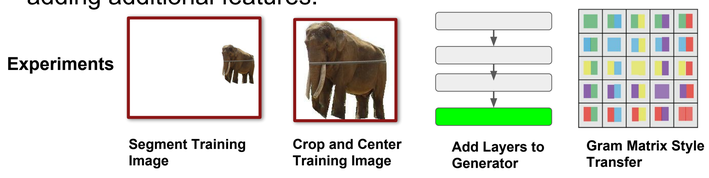
- 2 Approaches: first try 2-stage generate + style, then try integrated Generative Styled Network



Experiments

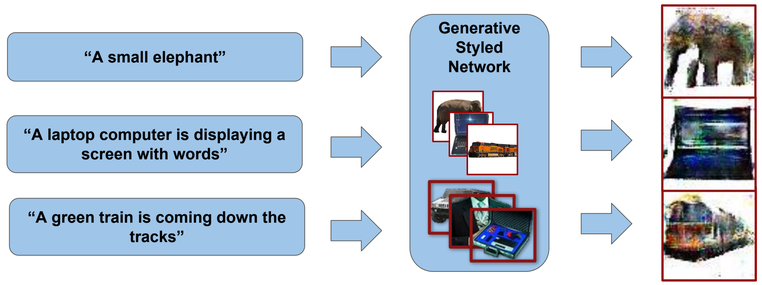
Approach One: 2 Separate Stages

- It's important to get a GAN working in the simplest fashion before adding additional features.



Generated images for caption "A small elephant walking across a dirt field"
Image generation performed much better than baseline, though Gram Matrix Style Transfer is not performant.

Approach Two: Single Stage Generative Styled Network



Generative Style Network is trained with both MS-COCO segmented and cropped images as well as style images from the iPhone Mafia Game. Generated images fit both the text input and style images

Final Results and Evaluation

- Quantitatively, single category model performance improved dramatically with Inception score increasing from 2.48 to 3.63

Single Category Improvements	Elephant Score	Laptop Score	Train Score
Integrated GAN and Style Transfer	3.33	4.87	3.94
5 Layer Network	**3.63**	-	-
VGG Architecture	3.27	-	-
6 Layer Network	3.50	-	-
Transfer Learning	2.84	-	-
Segmented and Cropped Images	3.55	4.4	5.12
Segmented Images	3.54	2.93	4.4
Baseline	2.48	3.75	2.2

- Qualitatively, images are much better: easily recognizable vs. incomprehensible for the baseline

"A yellow and black striped train next to sidewalk"

"A red train is docked at the station"

Information from captions is captured by the generated images. Left: images are without applying style. Middle: images are through our GSN. Right: results from multi-category model

- Multi-category performance also improved dramatically

Multi Category Improvements	Inception Score
3-Category Segmented and Cropped	7.15
3-Category Segmented and Cropped 5-Layer	6.03
3-Category Segmented and Cropped with Transfer Learning	5.67
Multi-Category Baseline	2.49
3-Category Integrated GAN and Style Transfer	1.86

Challenges and Error Analysis

- Poorly generated results (single category model)

"A picture of an elephant standing in some bush"

"A train with a striped door waiting on a train track."

"A laptop computer has a white screen."

- Single category elephant GSN: 26% generated correct images
  - Common failures: multiple elephants eg "two elephants," "elephant with a baby elephant"; pose eg "elephant walking away"; specific features "long tusks", "eating food"
- Multi category generation:
  - Model confusion: 'red' color associated with train, so a red elephant is in the shape of a train
- Future challenges:
  - Demonstrated that objects belonging to classes can be generated quite well. Moving forwards, can the model learn to generate relations as well? (e.g. "elephant next to a train")
  - Failure rate still too high, likely due to insufficient data. Try to get more data or augmentation
  - Multi category GSN not working as well as single category.

References

1. I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu,D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio.Generative Adversarial Networks.ArXiv e-prints, June 2014
2. S. E. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele,and H. Lee. Generative adversarial text to image synthesis.CoRR, abs/1605.05396, 2016.
3. T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. *CoRR*, abs/1606.03498, 2016.