Generating Cartoon Style Facial Expressions with StackGAN

Xiaoyi Li
Stanford University
xiaoyili@Stanford.edu

Xiaowen Yu
Stanford University
wyu1207@Stanford.edu

Introduction

In this project, we propose an end-to-end stacked jointly learning architecture stackGAN to transfer the facial expressions of real-world photos and convert the style to cartoon based on StarGAN and CartoonGAN.

Data and Features

- RAF-DB

The database used to generate facial expression is the Real-world Affected Faces Database (RAF-DB) which contains 12271 training samples and 3080 testing samples from real-world images. The database has 7 dimensional expression categories (from left to right: neutral, anger, fear, sadness, disgust, happiness, surprise).

- IIT-CFW

IIT-CFW dataset used for Cartoon style transform contains 8,928 annotated cartoon faces of celebrities with varying professions which are harvested from Google search.

RealFace

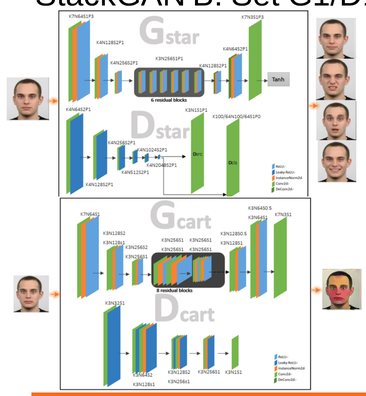
Cartoon Style Faces

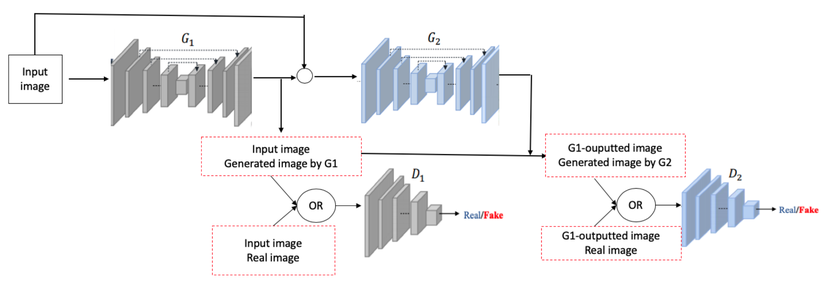
Models

We use StarGAN to generate facial expressions with the architecture Gstar and Dstar. We use the CartoonGAN for carton style transfer with the architecture Gcart and Dcart. The two GANs are stacked together and we train the stacked model from end-to-end with the structure shown on the right. A combined loss function where second GAN is trained conditionally on first GAN is used. We have two stacked GANs:

StackGAN A: Set G1/D1 to be StarGAN and G2/D2 to be CartoonGAN

StackGAN B: Set G1/D1 to be CartoonGAN and G2/D2 to be StarGAN



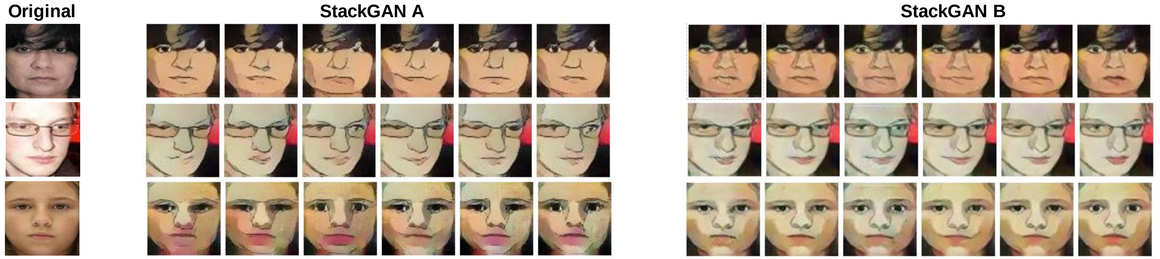


Loss Function: $L = \lambda_1 L_{GAN_1}(G_1, D_1) + \lambda_2 L_{GAN_2}(G_2, D_2|G_1)$

Results

We compare the stackGAN models in terms of output quality.

- Sample outputs (from left to right: anger, disgust, fear, happiness, sadness, surprise)



- Survey Result (15 photos in each survey, 40 responses)

Quality Scale (1-5)	StackGAN A	StackGAN B
Expression Quality	3.15	3.32
Cartoon Quality	3.06	3.25

Accuracy	anger	happiness	sadness	Total
StackGAN A	72%	76%	89%	79%
StackGAN B	86%	84%	86%	85%

Discussion

We compare the outputs from StackGAN A and StackGAN B and notice that:

- Almost no time difference in terms of training each epoch for both StackGANs.
- No significant difference on the convergence rate for both StackGANs.
- Switching the training sequence has some impact on the final output photos. Our survey assessing both quality and accuracy of the outputs shows that people tend to prefer StackGAN B a little more.

Future

- Improve expression training data quality
- Improve the architecture to make it more GPU-memory efficient
- Redesign some layers to improve the output data quality

Reference

Li, S etl. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Choi Y.etl. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. arXiv preprint arXiv:1711.09020v3

Chen Y, etl. CartoonGAN: Generative Adversarial Networks for Photo Cartoonization. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR).