Summary of “Fine-tuning StyleGan2 for Cartoon Face Generation”

Abstract

-Problem: while generated images look realistic, maintaining the structure of the source image is difficult

-Paper proposes methods to "preserve the structure of the source images", and they show those methods with a fine-tuned StyleGAN2 pretrained model

Introduction

Paper states that "Because style-based architecture trains sequentially from low-resolution to high-resolution, simple fine-tuning techniques make it easier to transfer models of source domains to target domains"

(that is to say that transfer learning is a good method for image-to-image translation)

Methods proposed:

1. **FreezeSG**

    -freeze initial blocks of style vectors AND generator; this causes structure of image to be preserved

2. **Structure Loss**

    -new loss function

Related Work

Mentioned GANs, Image-to-Image Translation, and Transfer Learning

-under 'Transfer Learning' section, paper mentioned FreezeD model which freezes "the highest-resolution layer of the discriminator" during fine-tuning

-also mentioned "adaptive discriminator augmentation" (ADA) that "stabilizes training in limited data", and that "ADA performs better with FreezeD"

-FreezeG, which freezes the generator's low-level resolution (or earlier) layer to keep structure of image

-**Layer Swapping** method "combines the high-resolution layer of the FFHQ (real-life face) model with the low-resolution layer of the animation model to generate a photo-realistic face"

-taking the fine details of the real-life face and combining it with the structure/shape of the animation model, or vice versa

Method

The paper experiments on two methods:

1. FreezeSG
2. Structure Loss

Method: FreezeSG

-Paper builds on idea of FreezeG by freezing not just the initial blocks of generator, but also initial style vectors in stylegan2. This method is simple and they name it ‘FreezeSG’.

    -style vectors are injected into the generator during fine-tuning

Method: Structure Loss

-Use a new loss function that is calculated as below:

2. Calculate the mse loss between $G^s_{l=k}(w_s)$ and $G^t_{l=k}(w_s)$ of each resolution, and add it up to the n-th layer.

$$L_{structure} = \sum_{k=1}^{n} \mathbb{E}\left[G^s_{l=k}(w_s) - G^t_{l=k}(w_t)\right] \quad (2)$$

-They apply structure loss to the low-resolution layer because “the structure of an image is determined at low resolution”, so by doing this, they preserve structure better. They mention that “jaws and heads are well generated” as a result; the structure of those parts have been preserved well.

Experiments - Dataset

-Source domain dataset = FFHQ dataset (contains real-life human faces), 70000 images

-Target domain dataset = cartoon faces, 8000 images of ~15 kinds of webtoons

Both dataset sizes in 256 resolution (256x256 size)

Experiments - training details

FreezeSG method - “freeze initial blocks of the generator and style vector, and then train the model by fine-tuning the stylegan2 pre-trained model”. They kept the default objective function used in stylegan2.

Structure Loss method - applied structure loss in three low-resolution layers for both source and target generators

Experiments - Results & Conclusion

-Paper’s proposed model was more effective in making ‘source image and target image similar’ than both the baseline (FreezeD + ADA) and FreezeG

-layer swapping technique = helps maintain structure of source image

-further experiments showed that highest quality images “were generated when structural loss and layer swapping were used together”

-FreezeSG method produced less natural images compared to structure loss method

Experiments - Results & Conclusion

paper says that images on the right (structureLoss + LS used together) were highest quality

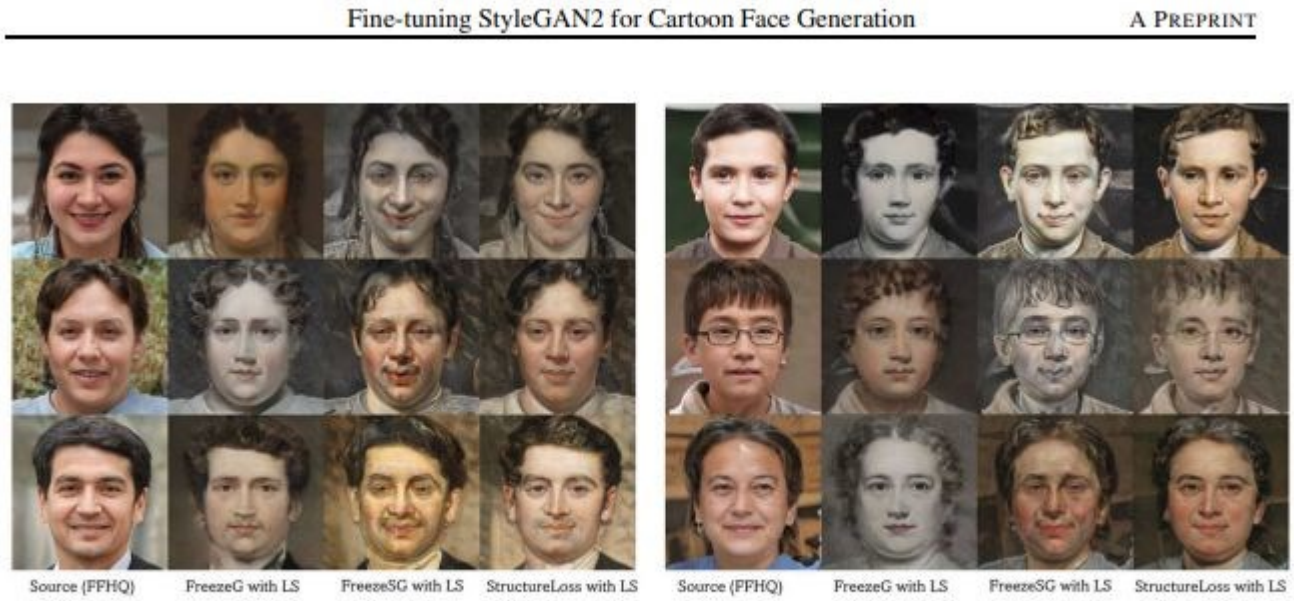

Figure 4: Comparison between freezeG and our models.

Future work

-”adjust layers to optimize for each dataset”

-”expect to be more stable and performance-enhancing”

Reference

Back, J. (2021). Fine-tuning stylegan2 for cartoon face generation. *arXiv preprint arXiv:2106.12445*.