Applying Cartoony Style Transfer to Realistic Videos

Chloe Toda
2025 Computer Systems Lab
Dr. Yilmaz, Period 3

Introduction

Cartoons are a popular form of media. However, in some cartoon shows (i.e. anime), animators must draw out each individual animation frame, and there are often over twenty frames per second. The animation process can be repetitive and physically taxing. I wanted to address this problem by converting realistic videos into cartoony ones using neural networks to mimic the animation process.

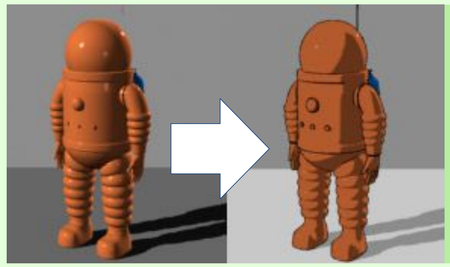

Figure 1: Generated images should have defined lines and cel shading.

Neural Style Transfer (NST)

NST is a relatively new process that involves applying the "style" of one image onto the "content" of another to generate a completely new image. For example, applying the style of Vincent van Gogh's *Starry Night* onto an image of a lion (Figure 2). NST is what's used to apply a cartoony "style" to videos of people.

Figure 2

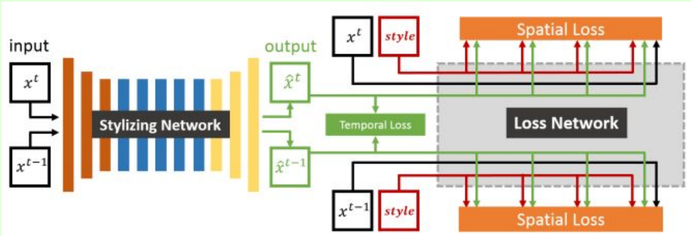

Figure 3: Video training flowchart

Method

I. Training Data

The content dataset is composed of images from the CelebA dataset, a dataset of over 200k photos of celebrities. The style data set is composed of frames of Disney characters from old Disney movies, which were found on an animation screencap website. Videos of myself or ones found on YouTube/stock video sites are used for the video dataset. Each video is 10-15 seconds long.

II. Training with Individual Images

The training neural networks were modeled after two previous research papers: CycleGAN and CartoonGAN. CycleGAN consists of two pairs of convolutional neural networks (CNN). Each pair has a generator CNN, which creates a fake image in the desired style, and a discriminator CNN, which determines whether the fake image is real or not. The goals of the generator and discriminator are to outperform each other, forcing each CNN to improve as training continues. The second project is CartoonGAN, which uses unique loss calculations to push its CNNs towards producing images with cartoon-like qualities (See *Figure 1*).

Figure 4: Training the network on both regular style images and those with their lineart blurred pushes it towards generating images with defined edges.

III. Training with Videos

After adequate training on the CelebA images, the network is trained on videos. With video style transfer, the network must take into account both frame n and the frame n-1 (See *Figure 3*). Calculating the loss between these frames is known as *temporal loss*, which pushes the network to generate adjacent frames with fewer differences, reducing the flickering effects that would be present if each frame was treated individually.

Results

Below are images generated by the GAN during the 35th epoch. While the style is not completely apparent, there are noticeable cartoony features like larger eyes and less detail in the hair and face. In terms of video style transfer, there was a noticeable decline in frame variation after each epoch.

Real input *Cartoon input* *Generated*

Figure 5: Training results

Conclusion

The goal of this project was to use a neural network to apply a cartoony style to realistic videos, while still preserving the original content of the video. Training a GAN takes a large amount of time and GPU. As such, with more time to train and stronger GPUs, the Disney style might be more pronounced the videos. Given the large amount of resources required to produce such outputs, perhaps it's better to leave animation to the actual animators, despite how interesting and novel this approach is. Some extensions of this project that I would be interested in pursuing are working with longer video frames, finding a way to reduce the time taken to train on them, and working with different cartoon styles.