ART AND MACHINE LEARNING
CMU 2019 SPRING
PROJECT 4

IMAGICAL

Piyusha Sane
Simral Chaudhary
Manisha Chaurasia

DESCRIPTION

Concept:

Human perception is what results when threads of numerous sensations are stitched together. The more senses that are stimulated together, the more neurons that spark to form a deep network, fostering an enhanced impression on the state of mind. Most naturally, sketches and paintings tend to catch our eye and leave a visual impact. Now, if these images are mentally associated with some music which tinkles in our ear, it creates an experience for the mind. When visual and auditory artworks are presented together, the brain correlates the two and seeks to perceive the latent meaning, allowing for a more wholesome intuition.

We wish to bring out the innate artist in every individual by transforming their sketch into a colorful picture and further generating music from that image to allow the artist to connect to their own artwork. Fusing the impression of two senses, vision and audition, the artwork will create a deeper and meaningful impact on the artist's perception. As an added feature, we will allow the artist to select from a variety of paintings to adopt the style into their own sketch, providing a personalized touch to their artwork.

Various work has been performed in the image and music domain, but not much relevant work combines the visual and audio to create a two-part project. Image to image translation has been performed with the use of conditional adversarial networks by the Berkeley AI Research lab. The released software and technique, named Pix2Pix, has proven successful in mapping edges pixel by pixel to generate an photo from the original sketch [1]. Various means of style transfer have been researched in the past few years. Lately, an image reconstruction network with built-in feature and color transforms has been shown to produce universal style transfer so that the content in images can be stylized with arbitrary visual styles [2]. Interpolation of music has been successfully performed by Magenta's Music VAE model which learns latent spaces of two music tracks and merges them to form a smoothly transitioning result which captures the essence of both auditory artworks [3].

We improve upon these methods and integrate it into an interactive system where an artist can draw a sketch, transform it into an image with a personalized style, and then listen to the corresponding melody generated by our system. We hope that the visual and audio artwork duo provides each person with a creative platform to connect with.

Technique:

The technique used in our final project was composed of four different components. The specifics of these are described below.

<u>Pix2Pix</u>

The original sketch is passed first through the pix2pix model, which is a conditional generative adversarial network (GAN) is composed of generator which uses a "U-Net" based

network while the discriminator leverages a PatchGAN network composed of a convolutional neural architecture. The "U-Net" model builds upon the traditional encoder-decoder setup as it introduces skip connections between counterpart layers in the encoder and decoder. This architecture allows for low-level information to be channeled across the network so that the key information is not lost in the downsampling during encoding. The discriminator PatchGAN considers the image as a Markov random field, treating the distant pixels independent from one another. The discriminator then runs convolutionally on square patches of the image and eventually the average of the outputs is deemed to be the final output of the discriminator. This results in a model with less parameters, making it a faster approach and one that can be scaled to fit any size of image [1].

Style Transfer

In order to stylize the image generated through the Pix2Pix model, we leveraged a universal style transfer method with embedded feature transforms modeled off of Li et al.'s work [2]. In this network, the output of the Pix2Pix model is passed through a VGG-19 model, which performs the feature extraction in the encoder. Then, a systematic decoder is trained to output the original image from the extracted features. This trained encoder-decoder system will then be used ahead.

To actually perform the style transfer, the model uses a series of whitening and coloring transforms (WCT) on the style features at multiple feature layers as shown in Figure 1. These WCT transforms work by pushing the covariance matrix of vectorized VGG feature maps of the content image to that of the style image. The whitening transform is achieved by Equation 1, and then inverted into RGB space.

$$\hat{f}_c = E_c \, D_c^{-\frac{1}{2}} \, E_c^{\top} \, f_c \qquad (1)$$

This transform essentially preserves the content and overall structure in the image, but wipes out the style. Hence, the output of the whitening transform is now ready to be stylized with another desired style. The coloring transform is basically the inverse of the whitening transform as defined in Equation 2.

$$\hat{f_{cs}} = E_s \, D_s^{\frac{1}{2}} \, E_s^{\top} \, \hat{f}_c \qquad (2)$$

The output is a stylized picture of the original image. This technique differs from traditional style transfer methods as it does not train using pre-set styles, but rather offers generalizability by allowing style transfer of any style.

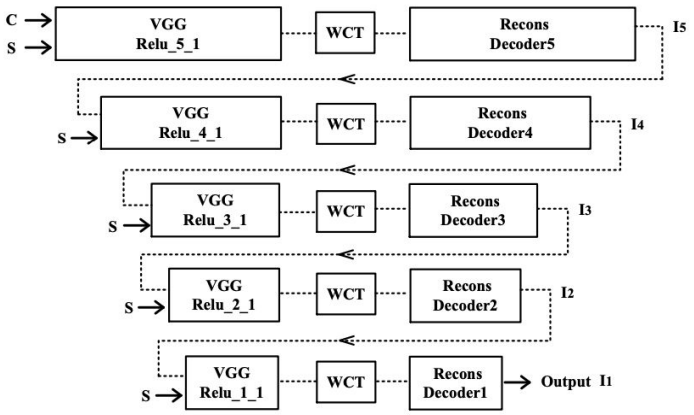

Figure 1. WCT transforms applied on the style features at the intermediate layers

Image To Midi

The stylized image is translated into a pixel-based musical sequence. The image consists of hundreds of pixels, which are defined by values in range 0 to 255 for the Red, Green, and Blue (RGB) scales. These pixel colors are transformed into musical notes in a series of steps. The average of the RGB components is divided by two and normalized. This normalized value is then linearly interpolated to the midi scale which takes value from 21 to 108, corresponding respectively to a low A in the 0th octave to a high C in the 8th octave. The resulting sequence of midi notes is packaged into a midi file corresponding to the stylized image.

Music Interpolation

After receiving the musical representation of the stylized image, the musical track is further interpolated with the music track corresponding to the specific style using a MusicVAE-based model. MusicVAE is a hierarchical recurrent variational autoencoder which learns the latent space of music. Fundamentally, it learns the innate characteristics of the high dimensional musical sequence and clusters examples based on the similarity amongst three traits: 1) *expression* (mapping any real music example to some point in latent space and reconstructing from it), 2) *realism* (any point in the latent space represents some real music sample), and 3) *smoothness* (nearby points in latent space represent similar music samples).

As shown in Figure 2, the model consists of an encoder, composed of a bidirectional recurrent neural network (RNN), which encodes the input into a latent representation. It is then followed by a decoder which uses another RNN to decode the latent representation back into a musical form. This network is trained while minimizing the variational loss. After interpolation of the two tracks, the resulting track can be seen to capture the flavors of both types of music. The latent representation in this case helps to further form a smooth result where the mix of the two tracks sounds more natural in its change in pitch and frequency, as compared to data space interpolation which sounds more choppy and noncoherent.

The interpolation between the stylized image's music and the style's associated image creates a meaningful result that helps the artist connect to their visual artwork even more.

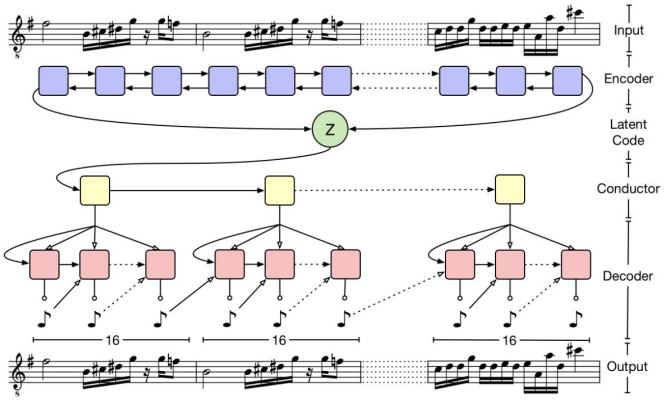

Figure 2. MusicVAE Network Architecture

Process:

When we began the project, we thought to use the Interactive Generative Adversarial Network (iGAN) for the sketch to image part of the project, which changed the produced image based on every stroke that the artist sketched as input. However, due to technical difficulties related to Python package versions, we could not leverage this model. We proceeded then to try a alternate model based on Pix2Pix, which showed promising results.

With the incorporation of the "U-Net" in the generator and PatchGAN in the discriminator (as described above in the technique), we saw colorful outputs that created a multitude of artistic perspectives in our minds. Figure 3 shows some examples of these images. As can be seen in Figure 3a, a simple zig-zag sketch produced a scene with two hills with a dusk background. It appears as though the sun has just set as the lighter blue color can be seen closer to the horizon. Meanwhile, Figure 3b, shows the output of a sketch of multiple squares. This artwork depicts the effect of heat or fire, and was perceived by us to represent a wildfire. Lastly, Figure 3c shows how a sketch of multiple overlapping criss-crosses turned into a colorful mountain range full of wildflowers.

(a) (b) (c)

Figure 3. Examples of the output of the Pix2Pix model

Based on these images, we thought we could further enhance the images with a completely different style inspiration. Hence, we progressed to incorporating style transfer on the generated image. In an attempt to allow a wide variety of styles to suit the personal interests of the artist, we wanted to integrate a style transfer method that was not restrictive of the specific styles. We hence proceeded to try a universal style transfer method as described in the technique. Figure 4 depicts a few of the images that resulted from the style transfer integration. Figure 4a shows the result of a sketch of a lady's face. The bold look of the woman intermingled with the vibrant colors depicts accomplishment and strength of the woman. Figure 4b shows a sketch of a flower stylized with the flower style. The image has successfully captured the colors of the petal with a background that imitates the sky and grass. Overall, the flower displays an earthy feel, sure to create a meaningful impact on artist. In the Figure 4c, the lollipop sketch has been stylized by a rainbow brush stroked style, which captures a sense of childhood and sweetness that is innate to the lollipop.

(a) (b) (c)

Figure 4. Output of the Style Transfer model

Based on our final results of the style transfer, we decided to proceed to the final component of our project, music generation. In this project, first we tried a naive way of image to midi conversion, to output a music track corresponding to the final image. Although this created an interesting set of musical notes, we felt that the final audio result could be enhanced if it blended a musical track representative of the style with the audio track generated from the image. For this, tracks were chosen to capture the innate theme of the default styles. In accomplishing this task, we wanted to fuse the latent meaning of the style audio and generated image audio. Consequently, we decided to use MusicVAE in interpolating the two musical sequences. The final audio result was thus generated.

Result:

(a) (b) (c)

Figure 5. Final results

We have chosen three image-music duos as our final result. **Figure 5a** is the styled image of a sketch of range of mountains in the far distance and a river closer in the view. The final picture turned out to be more detailed than the sketch. A boat, parked on the shore can be spotted closer to the bottom right region, with another bigger boat in the see above it, on which a man standing and waving his hand can be pictured. Small fish can also be pictured in the bottom region of the image. The sky seems to be much darker as if its approaching night, but the see has some colors in it possibly due to the artificial lights of the boats. The generated piece of interpolated music combines the music corresponding to the style of Van Gogh and sketch of mountains and river. This short piece of music resembles the interplay of calm of the sea and the contrasting occasional tides. It begins with a calmer note and turns into a variant pitch of notes signifying the two qualities of the sea.

In **Figure 5b**, a sketch of a zebra combined with the optical illusion-based style has actually colored in stripes on the zebra. Interestingly, the area above and below the horizon have been styled differently and can represent the sky in contrast with the ground. The second piece of

music evidently captures the confusion associated with the optical illusion style that was chosen for this. It starts with a confusing note and becomes repetitive set of two pitch notes which depict the black and white colors of the style. These notes can also be pictured as the stripes of the zebra, all of which are unique but still indistinguishable.

Finally, the last image, **Figure 5c**, shows an intricate sketch with a nature-themed style. The colors have beautifully captured a snowy mountain with evergreen trees. It has even captured the snow on the roof of the cabin. The corresponding piece of music is a beautiful depiction of the seamless harmony of the diverse elements of the nature, incorporating the different colors from the style and the different components like mountains, sunset sky, trees, snow and cabin in the woods, all combined to form a melodious harmony.

Reflection:

The final result of image and music was chosen as we felt it contributed to a variety of emotions that are expressed through both visual and audio art forms. The contrast of emotions of calmness vs magnificence is depicted by the sea in Van Gogh's style and justified by the generated music. The uncertainty of thoughts, illusion in the real world and nature itself is captured by the second image and its music. The harmony of all components of nature is beautifully conceptualised in the third image and its music. Contrast, uncertainty and harmony are important elements of the society and we have tried to represent these with our artworks.

We were excited with the results from our project as we used a combination of complex techniques to create an artwork stimulating more than just one of our senses. One point that we thought may further ameliorate the experience is adding a video of visualizations representing the musical rhythm and pitch. Additionally, we would like to enrich the music being generated through an alternate image to midi conversion process. The current method focuses on converting the RGB values directly to musical notes through linear interpolation. We would instead like to use a latent space model to characterize the generated image and then translate it into a musical sequence. This will likely result in a more meaningful music track, more representative of the image. Over the past few weeks, we learned that there are several avenues that can be followed to integrate image and music. In completing this project, we have realized how much senses can influence human perception. Incorporating multiple senses only enhances this perception and allows for a more magical experience when admiring artwork.

Reference:

[1] Isola et al. Image-to-image Translation with Conditional Adversarial Networks. 26 Nov 2018.

[2] Li et al. Universal Style Transfer via Feature Transforms. 17 Nov 2017.

[3] Roberts et al. A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music. 20 Jul 2018.
[4] https://github.com/mtobeiyf/sketch-to-art/
[5]https://colab.research.google.com/notebooks/magenta/music_vae/music_vae.ipynb#scrollTo=8J4vloU3Pgtz

CODE: **https://github.com/man11sha/Imagical**

RESULT: attached in email