Machine Learning for the Arts
UCSD SPRING 2019
FINAL PROJECT

Music Video Generation using Pix-to-Pix

Mikhail Kardash
Sean Liu

DESCRIPTION:

Concept

Does music make the video, or is the visual medium not influenced much by a musical one? In other words, can we treat music as a latent space for a music video? We were motivated to answer these questions because of our understanding of music as a collection of digital signals and video as a collection of digital images, which in itself also consist of digital signals. Therefore, we wanted to first establish whether the images of the video (the visual medium) was in any way correlated with the music at the corresponding timestamp (the aural medium). Next, we wanted we want to produce digital images completely based on the relationship between music and video signals. Finally, we wanted to use these digital images to produce brand new visual artwork; in other words, a music video if it were based on that relationship between the video and audio mediums.

Technique

In order to successfully accomplish our goals, we had to come up with an effective plan. First, we downloaded music videos from YouTube using Python modules such as PyTube and youtube-dl. We next processed our downloaded data by taking every 45th frame from the music video between 30 and 90 second timestamps and resizing the obtained image into a 256x256 shape. This was done using numpy and opencv libraries. Then, for each of the frames, we took 1.5 seconds of the corresponding audio to create a spectrogram image. This involved using numpy and scipy.spectrogram, setting the sample frequency to 44100 and setting binnings to produce the 256x256 output. The 256x256 outputs would then be paired together, placed adjacent, as a new training image for our network. This process of creating new training images would be repeated for however many music videos that we had downloaded.

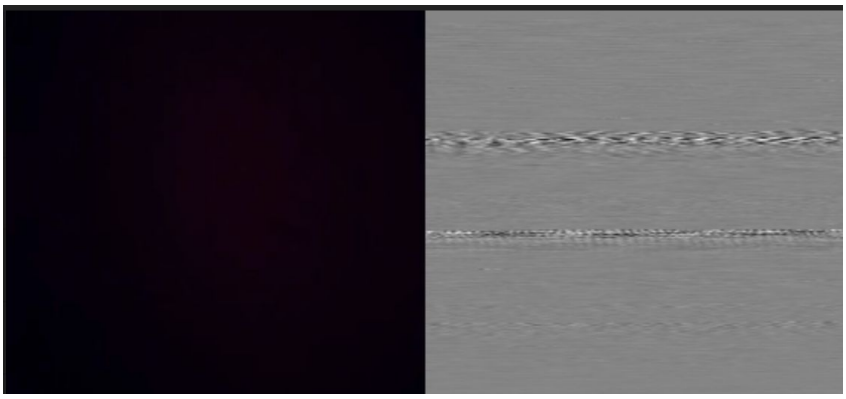
We then used pix2pix, a conditional GAN (Generative Adversarial Network) designed to do any general image to image translation. In other words, we trained our pix2pix network on the training images generated from pre-processing the data. Finally, we created a series of images based on another song, representative of the music video had it been shot based off of the music.

Process

In order to accomplish our goal, we had to ensure as much uniformity as possible. In our first iteration for project 4, we found that by downloading around ten music videos of various genres, it would be much more difficult to conclusively answer our question from our concept due to the excess noise that comes from different genres of music. In addition, our music videos didn't always contain music for some frames, as some videos had talking or cut scenes prior to the music. Therefore, for our final project's training data, we only used music videos by Cole

Bennett, a much sought after director for many rap music video productions. We chose him as our artist because we believe that his artistry makes his work a perfect candidate for analysis since we could effectively explore the relationship between rap music and music videos of a single director. By adding consistency to our initial training data, we eliminate the possibility of the excess noise explained earlier. Our ultimate goal is to train our network on a variety of his videos to see if different rap songs influence the aesthetics of the music videos, especially being fans of the rap genre.

Next, we needed to pre-process our data in a way that represented the image from every 45th frame and the corresponding audio of 1.5 seconds. We chose every 45th frame because we were downloading our music videos at 30 frames per second, which means that every 45th frame represents every 1.5 seconds of audio. Originally, for our project 4, we obtained every 1.5 seconds of audio along with every 45th frame of video for the entirety of each music video, which therefore includes, as explained earlier, audio at the end of the video (fade out) and audio at the beginning of the video (cut scenes, talking, fade in, etc.). Also, we did not use a spectrogram image, but actually obtained every 1.5 seconds of audio and reshaped the corresponding signal into 256 by 256 images. This inconsistency of results can be seen below:

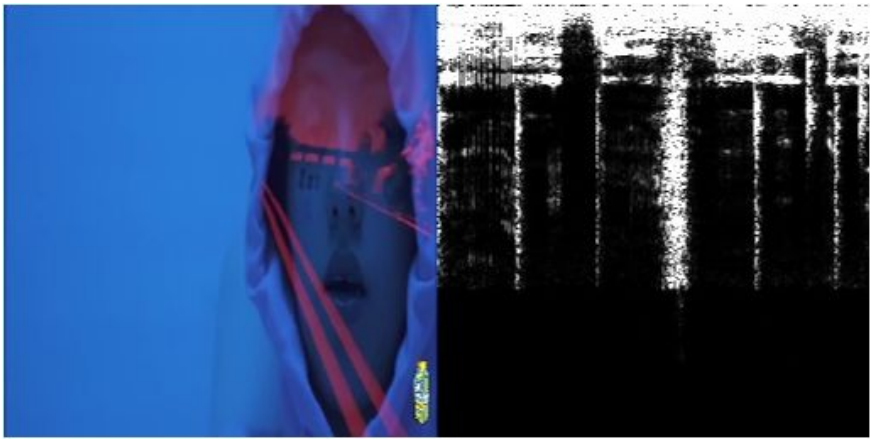


Training data: Image from music video on left, image created from audio on right

It's very clear that in the grand scheme of things, using images such as the one above would only create superfluous noise in our final results along with decreasing the effectiveness of answering our question posed in the concept. Outputs created would therefore look something like this:

Final Images: Image created from audio on left,
Image from music video in middle, Generated output image on right

Therefore, for our final project, instead of downloading 10 entire music videos of various genres as previously mentioned, we downloaded around 30-40 Cole Bennett directed music videos and obtained every 45th frame from the music videos between 30 and 90 second timestamps and resizing the obtained images into a 256x256 shape, in addition to creating spectrograms representative of the 1.5 seconds of audio in that same timeframe. This ensures that we will always get legitimate music to go along with legitimate video frames that go along with them, therefore ensuring the legitimacy of our training data. Examples can once again be seen below.

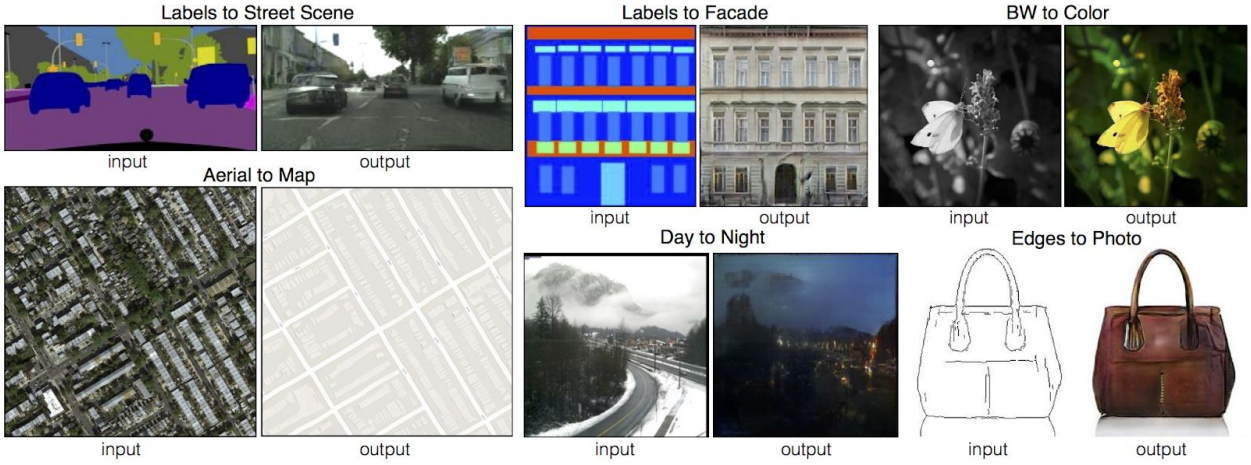


Revised Training data: Image from music video on left, image created from audio on right

Clearly, the revised training data not only has video frames which ensure have song audio, but the spectrogram images are also a more interesting way to represent the audio, especially compared to the mostly static and noisy audio image from our project 4 iteration.

Despite all the work and time spent done pre-processing, we would not be able to reach our end goal without a network. Therefore, we decided on pix2pix because of its solitary condition of inputs and outputs having the same dimensionality. In addition to that, pix2pix translates one

possible representation of a scene into another, given sufficient training data.[1] In other words, pix2pix essentially uses our paired input images from our pre-processing stage to establish relationships between the video image and the audio spectrogram image based on its image translation usage. This means that by implementing and establishing our audio spectrogram images as inputs and our video images as outputs for overall pix2pix training images, we therefore legitimize our relationship between the audio and video mediums. According to these training images, our audio spectrogram input should lead to the music video output. While we obviously the usage of pix2pix, whose training data already has relatively matching input and output images, will be much different from our audio input and video output scenes, we nonetheless believe this to be the best way to establish that relationship between the video frame vs the corresponding audio for each of the frames. Examples of pixtopix is shown below.



After training our data on the pix2pix network over epochs of 95, 180, and 200, taking place over multiple hours, we were finally able to make a ten second video of our generated images overlayed onto the original audio. We thought this was the most effective way of using the generated images- by making a revised music video based on the relationship between the signals of the image frames and the corresponding audio. In other words, our new music video is what would have been the result had the music video director had taken inspiration from the music itself to create his/her video.

Result

While we expected our generated outputs to have heavy bias with variation of shape or color, we were genuinely surprised with what our network actually spit out. Our outputs' overall shape and color are not only varying but transitioning smoothly in ways that make sense. Instead of being detrimental, the bias actually provided context to the overall output result. The transitioning shapes as a whole expressed ideas, especially when considering how the boundaries

were ambiguous and undefined. Finally, our video was pretty cool because there would be contrasting usage of light and shadow throughout, possibly reflective of its original musical medium. And the texture of the overall video appears spontaneous and decorative.

Generated Music Video in gif form

Individual frames of our music video

Reflection

Overall, we were quite satisfied with our final results because the results that we ended up getting were much better than we could have imagined. The network adequately produced color transitions that made sense and communicated aesthetic ideas through shapes. While we believed the bias of our network would be detrimental, especially when each generated image was taken in isolation, the overall video result was actually very much helped by the bias of our network. We clearly found the 200 epochs (more heavily trained) version to produce the best overall artistic results, which was why it was chosen over the results from less trained models. Although we have learned that we were able to answer our original question posed in the concept

with a more or less resounding “no”, our generated results can nonetheless lead us to conclude that music can have a legitimate influence in creating abstract visual media. For the future, one possible extension could be creating music videos where the audio is created from our generated images, therefore showing what a music video with video inspired by the music and music inspired by the video looks like.

Reference

[1] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros “Image-to-Image Translation with Conditional Adversarial Networks” 26 Nov 2018

CODE:

https://github.com/ucsd-ml-arts/ml-art-final-dem_boyz

RESULT:

https://docs.google.com/document/d/13X96gT69Ub33Fm_NqFLDhC8jfarBdFzUCIIZqg0IF-w/edit?usp=sharing

Output video:
Source of output video: https://www.youtube.com/watch?v=tc-v8MVw0S8&t=30s

Our output video can be seen in the github link, contained as “movie.mp4” (180 epoch) and “movie200.mp4” (200 epoch).