ART AND MACHINE LEARNING
CMU 2022 SPRING
PROJECT Final

What is your mood today?

Xiaofeng Gan (Social & Political History Major Undergrad),
Amanda Jin (Information Systems Undergrad with minimum knowledge in ML),
Gavin Deiss (Stat & ML Undergrad),
Yanxi Zhu (ECE graduate with some ML background)

DESCRIPTION

How to relate emotions to Arts and Machine Learning? It might seem to be novel and unrealistic. However, in our final project, we are going to examine different text, image and music generating models, link these model results with emotions, and more imperatively, interpret the results with artistic mindsets. For text generation, we first tried Markovify but failed to generate meaningful outputs and later we successfully gained expected results from the GPT-2 model for text generation. In terms of image generation, we use the CycleGan and CGan models to produce compelling outputs. Last but not least, we used Python's Magenta library to generate music for this project using melody attention RNN. Overall, the output of "delighted/happy" is the best we got so far given various text and image generating models. In addition, sad songs would be our best music output.

Concept

Emotions are at the center of metaphysics of humanity. They are prevalent in every inch of space, whether in imagery, sensation, or even the temperature of the air. Artists and writers imbue emotions within their works, creating vivid contexts and empathic understandings. However, emotions at the very core, are human beings. Everyone approaches and feels emotions differently. For example, surprise can be a positive emotion for ones and negative for the others. How might a machine approach emotions? This is the very question that our team asks and seeks to find out. Our project aims to explore the combination of art and the humanities. We want to see what a well-trained machine model can generate given a keyword corresponding to emotions, in terms of text, image and music. The emotions we are drawing from are the six universal emotions proposed by renowned psychologist Dr. Paul Ekman–happiness, sadness, fear, disgust, anger, and surprise. Beyond that, we also attempted to incorporate a series of synonyms and slangs into our test corpus.

Technique

Text

Markovify [2] is the first text-generating model we tried. Nevertheless, we attempted and failed to generate a text corpus based on only one descriptive word since Markovify works best with large, well-punctuated texts. If we have accidentally read the input text as one long sentence, markovify will be unable to generate new sentences from it due to a lack of beginning and ending delimiters. This issue can occur if we have read a newline delimited file using the markovify.Text command instead of markovify.NewlineText. To solve this problem, we have to pre-determine a text chunk with detailed information. Therefore, though Markovify is a more accurate and delicate text-generator, we still need to look for another model which guarantees texts based on one word or several key terms.

Following the failed trial, we switched our direction into the GPT-2 model given the paper titled "Language Models are Unsupervised Multitask Learners" [3]. In this text-generating model, we successfully get the expected outputs.

Image

Our major image-generation method for generating the final photos is CycleGan. Using CycleGAN, we experimented with several dataset combinations. We have played around with lambda and image size and also attempted least squares loss instead of sigmoid loss, which resulted in faster convergence of the models. More crucially, our training dataset contains over a thousand photos of various emotions, allowing our CycleGan-based model to discriminate some photos from background color and produce compellent results. Despite the fact that prospective and conditional failures are unforeseeable, our test results demonstrate the stability of our image model. In the end, we obtained six successfully blended outputs from six universal emotions —— happiness, sadness, fear, disgust, anger, and surprise. Unlike we did for project 2, in our final project, we selected our base and style images carefully. Taking similar color, shape and size into account, the model did not produce any blurry or unrecognizable images. Instead, our 6 trials of image generation given the CycleGan method all worked perfectly.

In addition to using CycleGan, we also tried a CGAN to explore what kind of images a machine is capable of generating. This model is trained by a labeled dataset called Wiki-Art Emotions which annotate paintings from places like museums and galleries with emotions. Each image is labelled by up to twenty emotions. For simplicity, the model we used compressed all the emotions down to three: positive, negative and neutral. For example, Happy will be categorized as positive; fear and disgust will be categorized as negative; and surprise will be categorized as neutral since it could be either way. Using the Emotion-to-Image generator, we can choose to generate either landscape, flower-painting, abstract or portraits along with the type of emotion we want the result to be.

Music

We used python's Magenta library to attempt to generate music for this project using melody attention RNN's, *attempt* being the key word here. Due to the high-dimensionality of mp3 files, and the difficulty in finding enough lower-dimensional midi files for each emotion, we ultimately scrapped music from our final app.

We had pretty good results using Magenta in Project 3, but it was because we were able to find a dataset of nearly 1000 Jazz midi's to work with. Here, even finding 30 midi files each for the emotions of happiness, sadness, and anger was challenging. It's not a widely-used file format on the internet. Further, finding files of ***any*** format that captured the other emotions we wanted to include in our app was borderline impossible. What exactly is "disgusting" music?

Regardless, we did train melody RNN's on our datasets for happiness, sadness, and anger. For anger and especially sadness, our results were actually quite good given our limitations. Still, there was no way we would be able to achieve decent results with the emotions of fear, disgust, and surprise, and the tradeoff of dropping these emotions from our final app wasn't worth it when we could generate great art & text for these emotions.

Process

Text Process

First, we try a single descriptive word for text generation based on the GPT-2 model.

Input: “Pessimistic” (with one descriptive word)

Output: “The world is a different place. There is a time for optimism about today. There is a time for cynicism about tomorrow. There is a time for the need to move ahead towards a more healthy world but also focus on the long-term goals that are important to a prosperous future. When I look at the future tomorrow morning in London I see a bright future. There are many wonderful opportunities in London for working hard, learning and building a better life. There is even a chance to join the ETS group. It's worth remembering that, despite this, it is not only London, but all over the UK, that will make these opportunities more available and more rewarding. My hope is that these opportunities will help, not disrupt, our current economic system. They are going to make it easier for people who do not currently see the kind of life that we want to have, and will only become more available and better.”

Moreover, though a player may only choose one adjective to represent his/her mindset, our model may lock several relevant keywords down and provide more precise and narrowed content. For instance:

Input: “Pessimistic, gloomy, you, children” (with multiple descriptive words)

Output: “I walk in a **gloomy** day and no one comes. "Mmm. Oh no!" **you** remark, as **you** lean over and make your approach. There is no sign of any good news to be found at your door! Your own children were in school, too. "Oh, boy! I've never heard of it," **you** say, and the children nod in agreement. "Oh, yeah. It's probably just that there aren't any good people in town." "Well, but it seems like a little place." A few of the parents smile and give a thumbs-up, and a girl smiles back, so that **you** are sure your boy is a little girl. **You** smile politely, and the little girl in particular looks at **you**, and **you** immediately tell her she is right. She looks at **you** the other way, and **you** are so very very very very wrong. When **you** get home, **you** come down to the door. "Shoot." **You** say to the old lady. She stands and looks at **you**, and she looks at your mother, and **you** realize she is crying. "Get her out of here," **you** shout towards the side of the stairs. **You** tell the girl she cannot get **you** out of here. She looks at **you**, but she smiles back and sees **you**. She stares at everything. And she is now a little girl.”

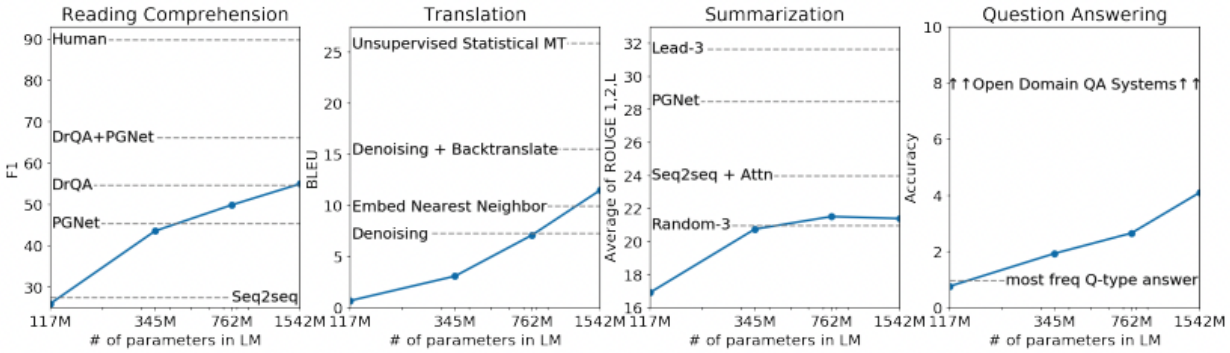
Here is another example of successful text generation with multiple descriptive words. We use the keywords, “delighted, university, basketball, scholarship”

Input: “delighted, university, basketball, scholarship”

Output: “I feel so delighted today. To make my name known to the whole community is really cool," he said. He added, "We can't wait to bring back our great city and be part of other people's lives!" The announcement marks the 18th year that the University of Utah is a member of the National Conference of Women Entrepreneurs. The university, meanwhile, has already opened its first restaurant in Reno in December before opening an outdoor tasting room next year. A team from the University of Utah has also been selected by NU to host a women's and men's basketball game at the University of Utah. "It provides opportunities for the students and visitors of Utah to learn from our coaches, players and athletic directors," said BYU President and President Rona Salih in a statement announcing the move. About $500,000 has been raised for scholarships for female athletes and $1 million for scholarships for women's basketball players. More than 439 students from over 35 different universities are scheduled at Brigham Young University-NU in Provo, Utah, for the Women's NBA Finals in 2017, which will be aired nationally on ESPN at 5 p.m. ET on Feb. 15.”

Now, the story is more comprehensive and logically clear. Below are some mathematical formulas and illustrations we have been using for our final project.

$$p(x) = \prod_{i=1}^{n} p(s_n | s_1, ..., s_{n-1}) \qquad (1)$$

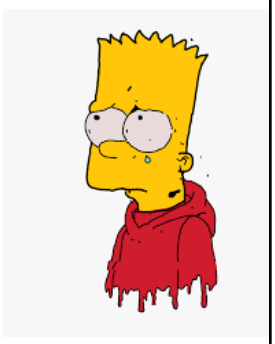
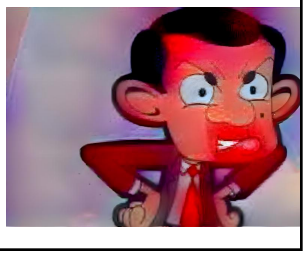


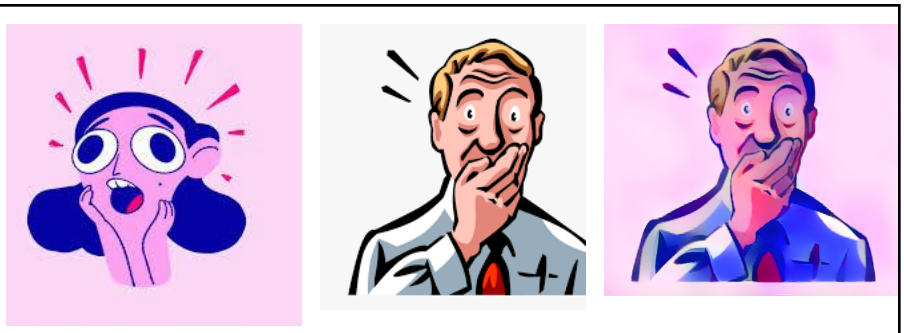
(2)

Image Process

Results from CycleGan:

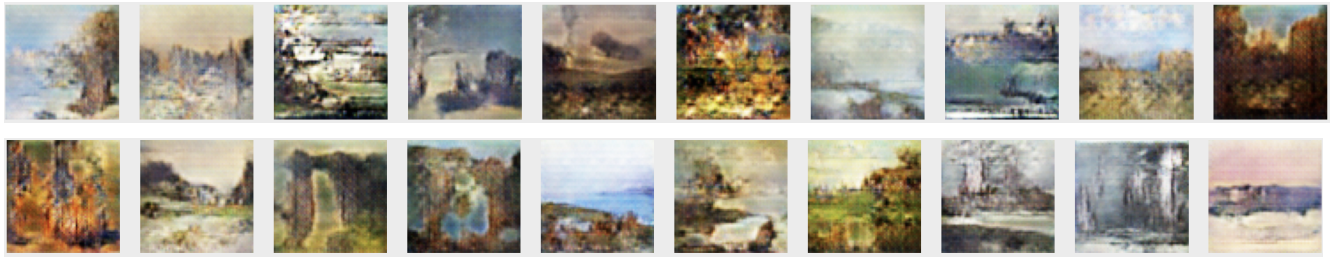
Number	Method	Image1 (style)	Image2 (base)	Output

1 (Happy)	**CycleGan**			
2 (Sad)	**CycleGan**			
3 (Fear)	**CycleGan**			
4 (Disgusted)	**CycleGan**			
5 (Angry)	**CycleGan**			

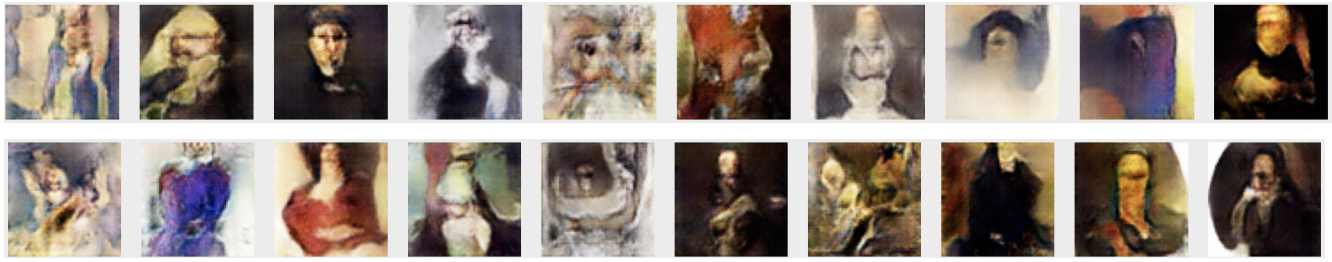
6 (Surprised)	**CycleGan**	

Results from CGAN[7]

Happy(positive), Landscape

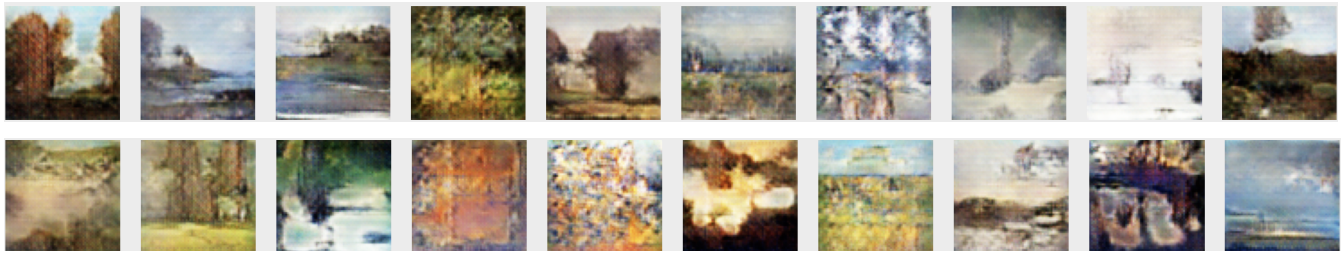


Sad(negative), Portrait



Fear(negative), Abstract

Disgust(negtaive), Landscape



Angry(negative), flower-painting

Surprise(neutral), Abstract

The result of this model wasn't on par with what we expected it to be. Although some of the results look good, one problem with it was that it didn't generate any emotion-based features in the paintings. In other words, positive landscape paintings look similar to negative landscape paintings generated by the model. There aren't many differences between the shape and the colors in paintings. However, this could also be due to the fact that the CNN used to label the data didn't have a good interpretation between the emotions within each painting.

Music Process

Our results with music were less successful due to difficulties in building datasets for each emotion. To generate music from mp3's would require literally millions of songs for each emotion due to the high-dimensionality of the data format. We thus looked to lower-dimensional midi files, but it's significantly more difficult to find midi files on the internet in general, much less for the six emotions we wanted to include in our app. There's plenty of happy and sad music out there, but disgusting music? Not so much.

So we had to make a decision: include just the 3 emotions we could find suitable music for in our app (happiness, anger, and sadness), even though said music wasn't that great; or, simply cut the music portion and focus on making the already functional portions better. We ultimately decided on the latter, but not before training a few models to see if there was any promise.

We compiled our midi datasets from bitmidi. We downloaded about 30 songs each for the emotions of happiness, sadness, and anger to see what results we would get. Our results for happiness weren't

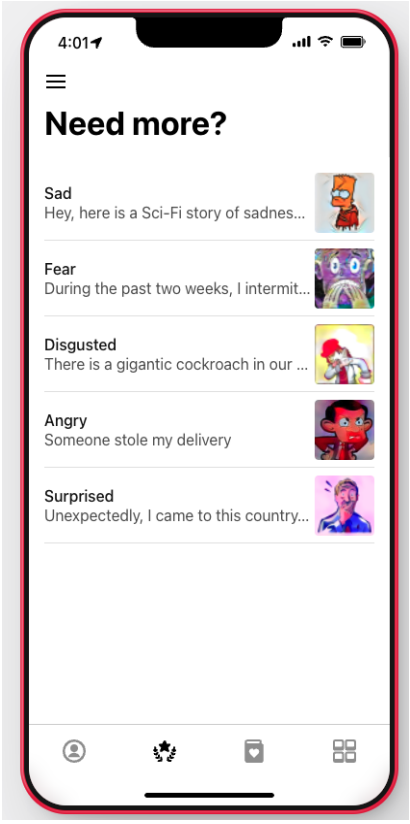
amazing, but for sadness and anger we actually got decent results considering our limitations. Overall, sad songs would be our best output.

- **Happiness**: Our dataset is comprised of various famous pop anthems, and songs with happy in their title. Generated results often lacked much of any structure and didn't necessarily sound happy, or sad, or really anything. They were very emotionally neutral
- **Anger:** Our dataset consists largely of video game boss battle music, since that was what was largely available in the way of midi files. We worried this might lead to homogeneity in our output, and while this is certainly a bit of a problem, there's actually a lot more variety than expected. A few popular, angry-sounding songs from bands like Metallica and Rage Against the Machine were included, perhaps contributing to the pleasantly-surprising amount of variety. And the output certainly does sound angry.
- **Sadness:** Arguably our most successful output comes from our dataset on sad songs. It is an unpleasant emotion, yet one responsible for a lot of amazing music, so perhaps this shouldn't come as a surprise.

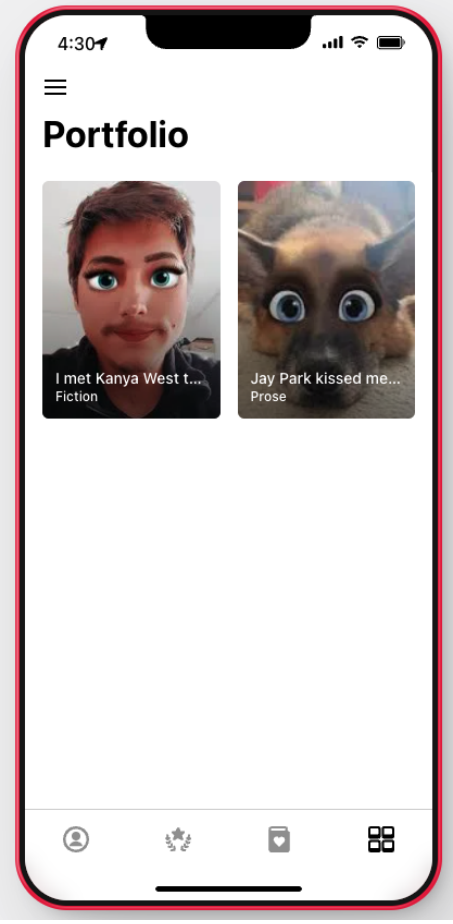
For each dataset, we created note sequences from the midi files; created sequence samples with an eval ratio of .10; and trained an attention RNN for 1000 epochs. We had a batch size of 64, and 2 layers each of size 64 for our RNN's. For each of the 3 models, we then generated 10 songs using just the note C as a primer 'melody'. All of this was done using python's magenta library. Some notable output can be found here (google docs doesn't allow you to insert audio clips).

UI/UX Design (Mobile Application)

Given the detailed instructions from Prof. Kang, we have all agreed that we might be able to create a mobile application demonstrating every aspect of our final project and holistically portraying our ideas in the end. We now have four different user interfaces displayed as below.



The first two pages contain several fully randomized outputs of different descriptive words of emotion. As demonstrated above, the main page is the best result being generated by GPT-2 and CycleGan models that day. In addition, we provide at least five more various options per day in order to meet users' different demands. For instance, we have listed "sad", "fear", "disgusted", "angry", and "surprised" in our second page.



The last two pages are about “Likes & Collected” and “Portfolio”. Obviously, the output that you found particularly intriguing and appealing will be directed to this page, as long as you clicked “like” or “collect”. Even though you might be unable to find something you like for one particular day, do not worry! You absolutely have the choice to create your own text and image generations given your keywords and photos.

Reflection

We have chosen “delighted/happy” as our final best output. It is undoubtedly to witness the fact that the story generated by “delighted, university, basketball, scholarship” is the most comprehensive and logically clear. At the same time, our base and style image selections for “delighted/happy” perfectly match and interplay with each other in terms of color, shape and size. However, for the RNN model of music generation, our best result goes to “sad”, since “happy” output sounds neutral and not significantly appealing. All things considered, we have manually opted “delighted/happy” as our main page example, since it is the general best result.

During the process of text and image generations, we by no means only use GPT-2 and CycleGan for text and image generations. For text, we also considered the markovify model to create different text corpus.

Nonetheless, the model is based on existing paragraphs instead of one or several keywords. Another possible path of future development of text generation would be generating one or two paragraphs first given GPT-2 and then applying the markovify model to produce a massive text chunk. As for image generation, there are a few more areas we could explore. For example, if there's a better CNN for labeling the paintings for our dataset, Conditional GAN could potentially provide us with better results. At the same time, we should also explore other image generation models and see what they're capable of generating. However, since emotions are often very subjective, the results of our model are heavily dependent on the dataset we curated. If we want consistency throughout the six emotions, we might need to have one person collect all the data.

As for music, there are a few directions we can go if we want to incorporate it into our app in the future. Certainly, building a dataset of millions of mp3's corresponding to each emotion is not practical, especially if we plan on an even larger variety of emotions. However, we could hand-pick songs manually and add a shuffle feature. This obviously wouldn't be an ML-heavy process, but would more so just be a quality of life tweak to the app. If we really want to focus on the ML, the best path forward might be to look at music more loosely and instead try to make soundscapes. This poses a similar problem as to whether anyone would even want to listen to a disgusting soundscape, but at least there's data for it out there opposed to disgusting music.

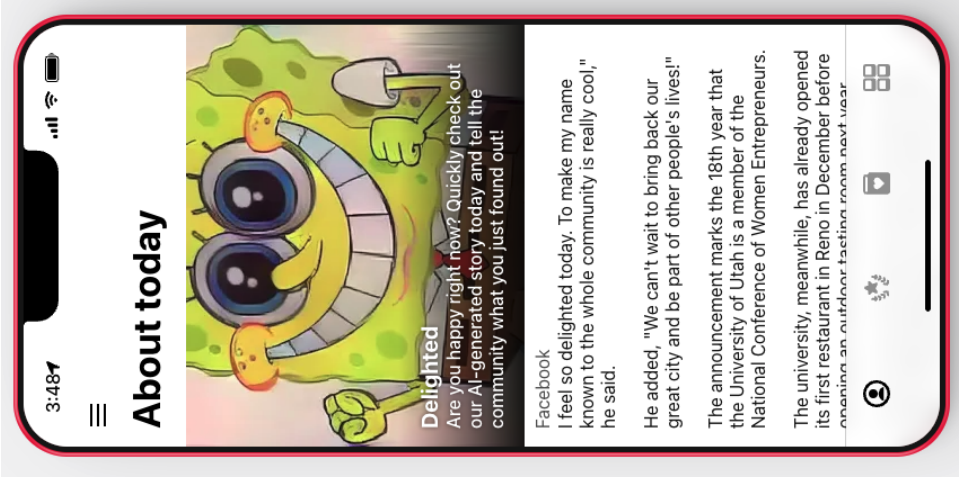
RESULT

Input: "delighted, university, basketball, scholarship"

Output: "I feel so delighted today. To make my name known to the whole community is really cool," he said. He added, "We can't wait to bring back our great city and be part of other people's lives!" The announcement marks the 18th year that the University of Utah is a member of the National Conference of Women Entrepreneurs. The university, meanwhile, has already opened its first restaurant in Reno in December before opening an outdoor tasting room next year. A team from the University of Utah has also been selected by NU to host a women's and men's basketball game at the University of Utah. "It provides opportunities for the students and visitors of Utah to learn from our coaches, players and athletic directors," said BYU President and President Rona Salih in a statement announcing the move. About $500,000 has been raised for scholarships for female athletes and $1 million for scholarships for women's basketball players. More than 439 students from over 35 different universities are scheduled at Brigham Young University-NU in Provo, Utah, for the Women's NBA Finals in 2017, which will be aired nationally on ESPN at 5 p.m. ET on Feb. 15."

Number	Method	Image1 (style)	Image2 (base)	Output
1 (Happy)	CycleGan			

https://docs.google.com/presentation/d/108FYUEDKhYwvqIaN_54AnCFSKxQj52445wufdWBTp7E/edit?usp=sharing (sad songs only)



https://go.glideapps.com/app/OdUvVlG7Rg4smbnyhr9k/layout (Link for our mobile app user interfaces)

CODE

https://github.com/xiaofenggan/10335-Final-Project/tree/main
https://github.com/gavindeiss/proj4_music_output

REFERENCE

https://github.com/openai/gpt-2 [1]
https://github.com/jsvine/markovify [2]
https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf [3]
https://github.com/MarvinMartin24/CGAN-Emotion-Art/ [4]