Machine Learning for the Arts
UCSD SPRING 2019
FINAL PROJECT

"A Picture of Him" - Video Style Transfer and Chatbot Audio Generation

Jishi Lyu (jlyu@ucsd.edu)
Zijian Ding (zding@ucsd.edu)
Yidi Zhu (yizhu@ucsd.edu)
Changhao Shi (cshi@ucsd.edu)

1. Concept

“I don't even have a picture of him. He exists now only in my memory." The moving lines of Rose in *Titanic* have touched thousands of people deeply. In this project, we hope to eliminate the regret of Rose by applying style transfer to generate the scene of Jack and Rose “I’m flying”. To make the scene more interesting, we also use chatbot audio generation to produce new lines of the scene.

It can also be seen as a revisit of both Project 3 - Generative Audio and Project 4 - Generative Visual. The style transfer and generative audio will be conducted by GAN and DeepVoice3 model respectively.

2. Technique & Process

2.1 Style transfer

Basic idea
The main idea behind this is to leverage available image style transfer model to transfer the style of a frame sequence taken from input videos and convert this transferred frame sequence back to the original video format.

Method
First, we convert the input video into a sequence of image frame. we take every frame in the video and arrange them in order. The original format of the video is mp4.

Second, we transfer these consecutive frames into a target style using a pretrained model. The selected model is the arbitrary-style-transfer model which is more flexible and manageable.

Then, we convert these consecutive image frames in new style back to the video format, which is chosen to be avi due to its small memory consumption. As what we have expected, this new video is successfully transferred to the target style. we also combine the multiple styles by arrange those transferred frames in an ordered way (f1_sty1, f1_sty2, f1_sty3, f2_sty1, f2_sty2, f2_sty3...).

Model/Data
As mentioned, we use the arbitrary style transfer model. We have tried several differetn style images to choose the best effect. After comparison, we select the famous painting ***Udnie*** from the French artist Francis Picabia as shown in Figure 1. We consider that the bold color and the concise space abstraction is very suitable for style transfer.

Figure 1: style image Udnie

2.2 Generative Audio

Basic idea

Since we want to combine the idea of style-transferred movie with generated video, our idea is to apply DeepVoiceV3 of multi speakers to generate a dialogue between two chatbots. First, we input actor's lines and make the “multi speaker” to imitate the original lines to restore the scene. Then we want to try something interesting, so we use the chatterbot we trained before in project 3 to teach the machine some chatting pattern and make them to chat with themselves.

So the main idea is that we try some combination of original conversation and generated ones. The feed seed could be the question in the original text or something new.

Method

From previous work of chatterbot, we can build our own robot to generate the text of the speech. We build two robots: Jack and Rose. They are all trained on English dataset which contains following fields:

ai,botprofile,computers,conversations,emotion,food,gossip,greetings,health,history,humor,literature,money,movies,politics,psychology,science,sports,trivia.

Then we pick two speakers who speak clearly (number 4 and number 7) and then turn the conversation into audio file. Our method is that every time we turn one sentence into audio. Then we insert about a one second pause. After that we could follow next sentence. This method makes we understand who (which robot) is speaking and make the conversation clear.

Model/Data

In ChatterBot, we use the "chatterbot.corpus.english" set to train the robot. In DeepVoice, we download the pretrianed multi speaker synsethis model "20171222_deepvoice3_vctk108_checkpoint_step000300000.pth".

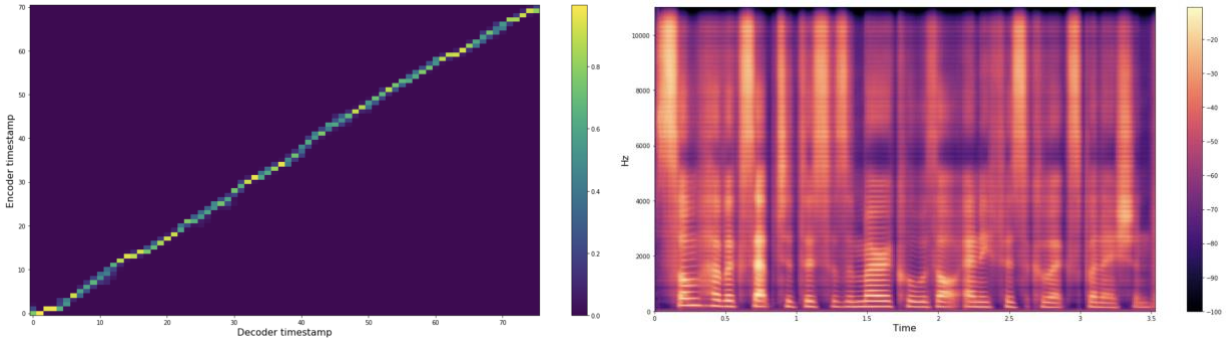

Figure 2: the attention plot of the multi speaker model

The line in the graph shows that the text position n progresses nearly linearly to the time t, i.e., n ~ at, where a ~ N/T. It means the model could just pronounce the sentence in a almost uniform speed. Furthermore, the fact is that we could see various frequency in the whole timestamp which may be a good synthesis of the sound.

Results

By feeding the response of each robot to the other, we get the response from the other. Finally, it forms like a dialogue. Then we turn it to .wav file. We could hear the speaking most time clearly. Also, we feed original actors' dialog and some horror stories to generate the audio file of them. The generated script is attached below:

Rose: Hello Jack. I changed my mind.

Rose: They said you might be out here.

Jack: Shhh. Gimme your hand.

Jack: Now close your eyes, go on.

Jack: Now step up. Now hold on to the railing.

Jack: Keep your eyes closed, don't peek.

Rose: 23 skiddoo! (generated)

Jack: Step up on the railing.

Jack: Hold on, hold on. Keep your eyes closed. Do you trust me?

Rose: Sort of. (generated)

Jack: All right. Open your eyes.

Rose: I'm flying, Jack!

Jack: Come, Josephine, in my flying machine, going up, she goes up, up she goes. (generated)

3. Final Result

The final artwork is uploaded to YouTube:
https://www.youtube.com/watch?v=bkWVaSbXI4w

Since the Udnie provides the best effects according to our trial, the overall effect is satisfactory. However, there still exists small flaw – the face becomes blurry sometimes.

Since we apply DeepVoice to generate the audio, interestingly, we find some words in the script generated by ChatterBot cannot be pronounced by DeepVoice, like “23 skiddoo!” – DeepVoice just generates other phrases instead.

Here is an example of the text generated by the chatter bot in the video. The screenshot for original question is:

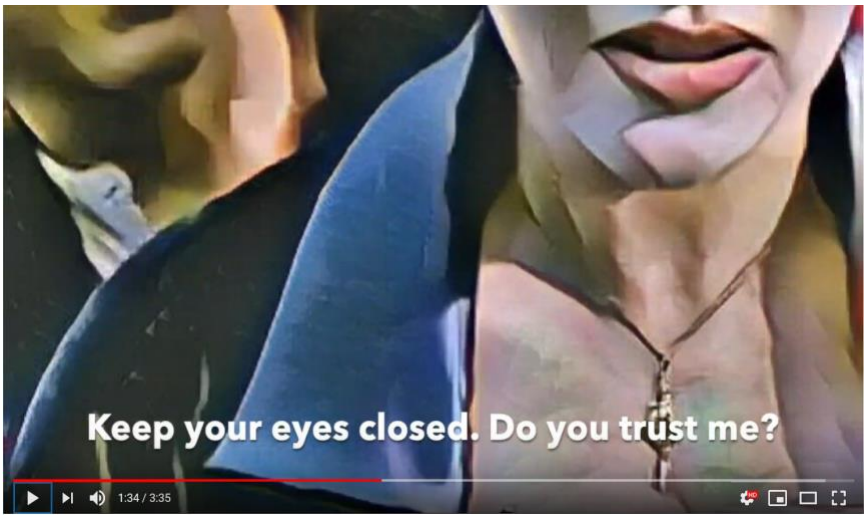

Figure 3: the screenshot for original question

The original answer is “I trust you”; while the generated one is “sort of” as shown in the screenshot:

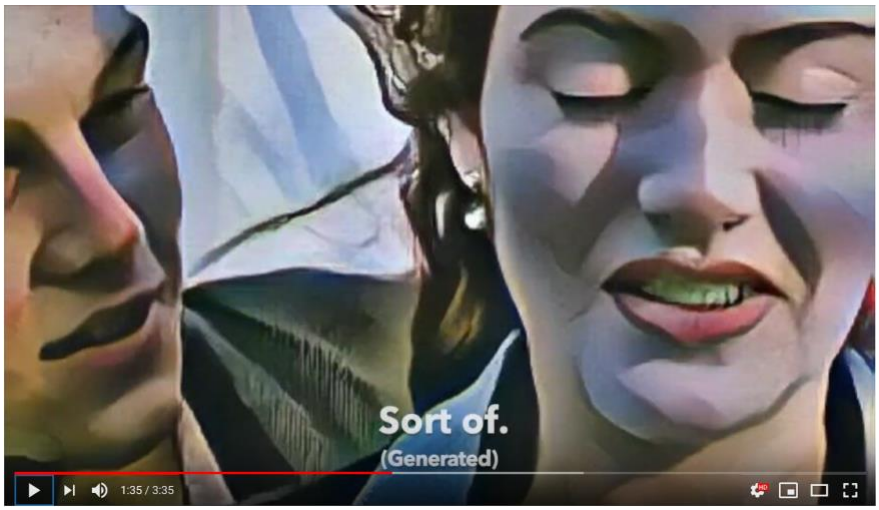

Figure 4: the screenshot for generated answer

More selected screenshots are attached below:

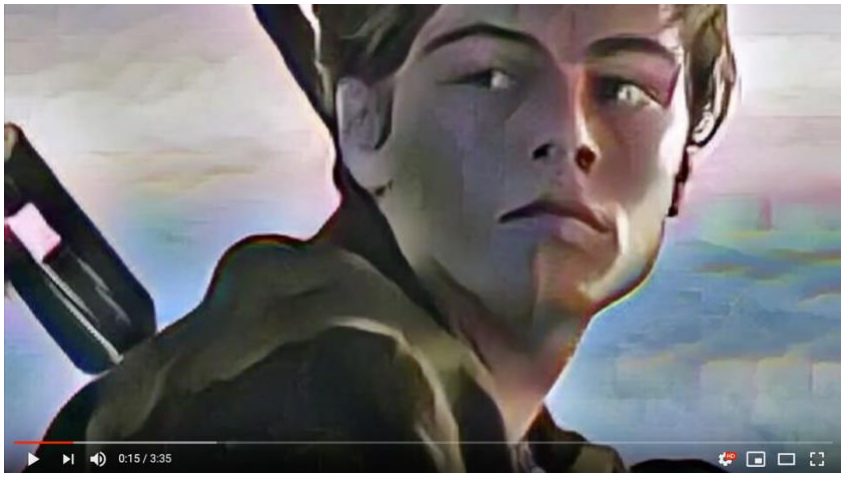

Figure 5: opening scene of Jack

Figure 6: scene with generated audio - 1

Figure 7: scene with generated audio - 2

4. REFERENCE

Reference Link
ChatterBot: https://github.com/gunthercox/ChatterBot
DeepVoicev3: https://github.com/r9y9/deepvoice3_pytorch
ArbitraryStyleTransfer: https://github.com/elleryqueenhomels/arbitrary_style_transfer