ART AND MACHINE LEARNING
CMU 2022 SPRING
PROJECT 4

Interactive Immersive Multimedia Generation

Zhouyao Xie
Nikhil Yadala
Yifan He
Guannan Tang

DESCRIPTION

Throughout the course of this semester, we have explored the translation between all sorts of modalities: natural language, images, audio, etc. In this final project, we would like to explore one more modality: human consciousness.

The goal of our project is to translate our stream of consciousness into a multimedia experience that consists of music, lyrics, and images. Instead of using brain computing interface (BCI) techniques to directly obtain consciousness as brain signals (which is definitely one of the future directions of this project), we use spoken language to represent consciousness. The user speaks anything that's on their mind right now, and our system turns that spoken language into a film that consists of music, lyrics, and images, thereby creating an immersive multimedia experience that represents and extends the user's streams of thoughts. To do so, we put together a whole suite of pretrained models and existing tools that perform speech-to-text, lyric generation, text-to-image, and text-to-music. Lastly, we combine the lyrics, image, and music generated into a complete film.

Ideally, this translation process would happen in real-time. As the user speaks out whatever thoughts that come to their mind at the very moment, the generated multimedia experience would get updated accordingly, making it an interactive experience. Unfortunately, due to constraints in the inference time of current generative models, we are not able to develop a real-time system. Currently, our whole pipeline takes from several minutes to about an hour to execute, depending on the length of the film. However, with the development of generative models with shorter inference times, we hope that a real-time, interactive system would become feasible in the future.

TECHNIQUE

Multimodal alignment: We use the CLIP model [2] to guide the generation of the images, music, lyrics that align with the text. During the inference, each of the generator is backpropogated multiple times until the CLIP score for the multimodal alignment is good enough.

Speech-to-Text: We use synchronous recognition Speech-to-Text API provided by Google Cloud. The API sends the audio data to perform speech-to-text recognition and return results after the entire audio has been processed. The maximum speech duration is 1 minute since the limited audio data of synchronous recognition requests are limited to audio data of 1 minute or less.

Lyric Generation: We use GPT-2 model pretrained on different artists' lyrics datasets available on HuggingFace [3] to generate lyrics from the output of Speech-to-Text module. Specifically, the examples we presented use models trained on lyrics of Muse, Bob Dylan, and Eminem.

Jukebox: We use Jukebox [4] to generate music with features conditioned on our chosen artist and input lyrics. The model uses a multi-scale VQ-VAE to compress the target audio domain and then generates the output music with an autoregressive Transformers.

FuseDream: FuseDream [1] is a model based on GAN and CLIP that can generate images given natural language prompts. The model was pretrained on the MS COCO dataset. We use this pretrained model to approach the lyrics-to-images transformation in our project. Each line of our generated lyrics will be transformed into an image that describes its semantic meaning. The synthesized images were collected eventually and formed into short films.

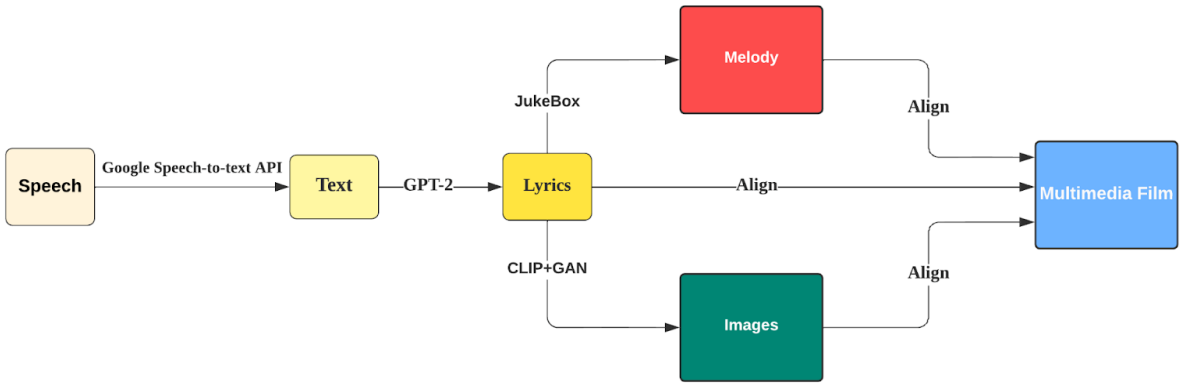

Fig: our system design diagram

RESULTS

Our results are available on Github:
https://github.com/ZhouyaoXie/interactive-immersive-multimedia-generation/tree/main/video

To perform smooth translation from one videoframe to the anothe, we gradually increased the weight of subsequent frame and gradualy decreased weight for previous frames. As the frame switches completely, the related lyric is added to the video.

REFLECTION

We encountered three main challenges while working on this project. The first challenge is that compared with generating instrumental music in symbolic format (e.g. MIDI, piano-roll representation, etc.), it is much harder to generate music with lyrics in waveform. Although Jukebox is one of the state-of-the-art music generation models for this task, the music it generates is pretty coarse and it is hard to distinguish the lyrics that are being sung. Moreover, it is also quite challenging to generate songs that are longer than one minute while maintaining long-term coherency. We observe that the quality of our generated music deteriorates sharply after around 50-60 seconds. Lastly, all the generative models we used in this project, whether it's music or image generation, suffer from high inference latencies. In fact, it takes almost a whole day for our pipeline to generate the lyric, music, and images for one film. In

the future, with the development of generative models that optimize for inference time, we hope that we will be able to develop a real-time, interactive generation system.

Moreover, Although we do not directly re-use our code or results from previous projects, we feel that they are really helpful in preparing us for this project. First of all, we are now much clearer about how to set reasonable goals for our project and how to best manage our timeline. While we were in trouble for setting overly ambitious deadlines in project 2, this time we were able to scope our project appropriately and achieve a pretty nice balance between complexity and feasibility. Moreover, after working together in project 2 and 3, our team members are much more familiar with each other's working style, which helps enhance our efficiency. While working on this project, we were able to communicate and collaborate much more efficiently compared with when we just teamed up. As a result, we feel that the previous team projects help us lay a great foundation for this final project. We had lots of fun developing this project, and we are looking forward to applying the skills we have learned in this course to future projects!

CODE

Our code can be accessed via this link:
https://github.com/ZhouyaoXie/interactive-immersive-multimedia-generation

REFERENCE

[1] Liu, Xingchao, Chengyue Gong, Lemeng Wu, Shujian Zhang, Hao Su, and Qiang Liu. "FuseDream: Training-Free Text-to-Image Generation with Improved CLIP+ GAN Space Optimization." arXiv preprint arXiv:2112.01573 (2021).
[2] Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry et al. "Learning transferable visual models from natural language supervision." In International Conference on Machine Learning, pp. 8748-8763. PMLR, 2021.
[3] HuggingfaceArtists models: https://huggingface.co/huggingartists
[4] Dhariwal, Prafulla, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. "Jukebox: A generative model for music." arXiv preprint arXiv:2005.00341 (2020).