CSCE883 Final Project

Proposal

Title: Generating Video from Audio input

Team Members: Sai Vuruma

Abstract: I wish to build a model/pipeline that can generate videos from input audio signals. This work will be the baseline in creating world-building videos from descriptive audio inputs like audiobooks. In this project, I will focus on the first step – generating short videos from audio inputs.

Introduction: Video generation from audio inputs is a cutting-edge research field that has applications in various domains from movies to education [1] [2]. The idea for this project came from a larger idea to generate world-building video from audiobooks. It is said that most of the information that we acquire is visual – through the eyes. There has been a lot of research on generating image and video outputs from text inputs but the area of generating video from audio inputs is still relatively new [3] [4].

Data: I will use the ACAV100M [5] dataset for this project. It is a large-scale dataset that contains 100 million videos with high audio-visual correspondence. Each video is a 10 second clip and in total the dataset adds up to 31 years' worth of videos. Given the size of the dataset, I will use a smaller sample of the 100M dataset to train my model.

- The input to the model will be the audio signal and the model will be trained to generate video outputs.
- The true videos from the dataset will act as the ground truth for judging the output generated by the model.

Model: I have not fully decided on the model yet. Given the nature of the problem, I would say right now that Generative Adversarial Networks (GANs) and Variational Auto Encoders (VAEs) are prime candidates. Meta's ImageBind [6] model will also feature to some extent given its strong performance in identifying commonalties between multi-modal inputs. A hybrid structure involving all three models seems like the best solution, however I am yet to finalize that.

Evaluation: The project will be judged on the following criteria, in order of precedence:

1. Relevance of generated video output to the input audio signal
2. Similarity between generated video output and expected video output
3. Quality of generated video output

I will look into relevant research and finalize the exact metrics to use for the above criteria. As mentioned in the Models section, ImageBind's similarity score is a great candidate at this moment. Signal-to-Noise ratio is another metric that will suit this use case. I will add more metrics according to the literature.

References:

[1] Xing, Yazhou, et al. "Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

[2] Kumar, Neeraj, et al. "Robust one shot audio to video generation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020.

[3] Chen, Lele, et al. "Deep cross-modal audio-visual generation." Proceedings of the on Thematic Workshops of ACM Multimedia 2017. 2017.

[4] Żelaszczyk, Maciej, and Jacek Mańdziuk. "Audio-to-image cross-modal generation." 2022 International Joint Conference on Neural Networks (IJCNN). IEEE, 2022.

[5] Lee, Sangho, et al. "Acav100m: Automatic curation of large-scale datasets for audio-visual video representation learning." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[6] Girdhar, Rohit, et al. "Imagebind: One embedding space to bind them all." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.