Baselines Report

Team Rocket[1]

[1]School of Engineering and Applied Science, The University of Pennsylvania

Abstract

We aim to develop a model that can create video content in the form of video from audio input, such as music, speech, or environmental sounds. We leverage contrastive audio-visual pretraining and existing video generation models to translate audio patterns into corresponding visual representations. This technology has potential applications in multimedia content creation, accessibility tools for hearing-impaired individuals, and enhanced user experiences in virtual environments.

Introduction

Humans can imagine the scenes from the audio they hear, but some audio is hard to imagine like music. If there is a model that can convert complex sounds into visual scenes, people would better understand these sounds. For instance, individuals can enjoy music using more senses in concerts, which lowers the threshold for enjoying an art. We propose an audio-to-video generation model using CAVP from Diff-foley[1], which has a similar structure as un-CLIP[2] did in the text-to-image area.

Method

Datasets

We will use VGGSound [3], which contains 200,000 10-second clips from Youtube. There are 310+ classes for the clip from the dataset, for example, kid specking, gun shooting, vehicle horn, and sound for operating metro.

Evaluation Metrics

To measure the video quality, we use the following metrics: (i) Frechet Video Distance[4], which calculates the visual difference between the generated video and target features extracted from a pre-trained model; (ii) CLIP Similarity Metric[5], which is computed by averaging the CLIP similarity of each frame with the text input; (iii) Inception Score[6], which is computed with a pre-trained C3D model[7]; (iv) and Audio-Video Alignment proposed in Tempotoken[8], which measures alignment of changes in video and audio.

Table 1: *Result for the baseline*

Model	FVD(↓)	CLIPSIM(↑)	IS(↑)	AV-Align(↑)
	VGGSound			
ModelScope Text2Vid	**801**	**0.69**	**15.55**	0.27
ModelScope Random	1023	0.47	6.32	0.26
Tempotoken	923	0.47	6.32	**0.35**

Experiment Results

Baseline

As a baseline, we use the TempoTokens model [8]. Inputs are formatted as .wav files, and outputs are corresponding .mp4 files. Performance metrics, as reproduced from their paper, are reported in table 1.

References

[1] Simian Luo et al. *Diff-Foley: Synchronized Video-to-Audio Synthesis with Latent Diffusion Models*. 2023. arXiv: 2306.17203 [cs.SD].

[2] Aditya Ramesh et al. "Hierarchical text-conditional image generation with clip latents". In: *arXiv preprint arXiv:2204.06125* 1.2 (2022), p. 3.

[3] Honglie Chen et al. "VGGSound: A Large-scale Audio-Visual Dataset". In: *CoRR* abs/2004.14368 (2020). arXiv: 2004.14368. URL: https://arxiv.org/abs/2004.14368.

[4] Thomas Unterthiner et al. "Towards Accurate Generative Models of Video: A New Metric & Challenges". In: *CoRR* abs/1812.01717 (2018). arXiv: 1812.01717. URL: http://arxiv.org/abs/1812.01717.

[5] Chenfei Wu et al. "GODIVA: Generating Open-DomaIn Videos from nAtural Descriptions". In: *CoRR* abs/2104.14806 (2021). arXiv: 2104.14806. URL: https://arxiv.org/abs/2104.14806.

[6] Tim Salimans et al. "Improved Techniques for Training GANs". In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee et al. Vol. 29. Curran Associates, Inc., 2016. URL: https://proceedings.neurips.cc/paper_files/paper/2016/file/8a3363abe792db2d8761d6403605aeb7-Paper.pdf.

[7] Du Tran et al. "Learning spatiotemporal features with 3d convolutional networks". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 4489–4497.

[8] Guy Yariv et al. *Diverse and Aligned Audio-to-Video Generation via Text-to-Video Model Adaptation*. 2023. arXiv: 2309.16429 [cs.LG].