View Reviews

Paper ID
84

Paper Title
CLIPSynth: Learning Text-to-audio Synthesis from Videos using CLIP and Diffusion Models

Track Name
MMSP2023

Reviewer #2

Questions

1. How confident are you in your evaluation of this paper?
confident

2. Importance/Relevance to you
Of sufficient interest

3. Novelty/Originality
Has been done before

4. Technical correctness
Probably correct

5. Experimental Validation and Reproducibility
Limited but convincing

6. Clarity of presentation
Clear enough

7. Reference to prior work
Reference missing

8. Overall evaluation of the Paper
Strong reject

9. Justification (required if score of 1 or 2 has been selected for questions 3-7)
The paper presents an architecture for generating realistic sounds. It is not clear which is the novelty with respect with what appear to be previous works from the same authors.

10. Additional comments to author
The authors should cite their preceding works (e.g. is there a related work by the same authors at CVPR2023?) and clearly highlight the novelty of the current contribution.

Reviewer #3

Questions

1. How confident are you in your evaluation of this paper?
Less confident

2. Importance/Relevance to you
Of sufficient interest

3. Novelty/Originality

3. Novelty/Originality
Moderately original

4. Technical correctness
Probably correct

5. Experimental Validation and Reproducibility
Sufficient validation / Theoretical paper

6. Clarity of presentation
Clear enough

7. Reference to prior work
Reference adequate

8. Overall evaluation of the Paper
Weak accept

9. Justification (required if score of 1 or 2 has been selected for questions 3-7)
The paper presents a novel method for text-to-audio synthesis without using audio-text pairs and uses images to bridge this gap. The architecture proposed by the authors involves a pretrained CLIP model to generate embeddings, a diffusion model to map noise to spectrograms and a spectrogram inversion model. The empirical performance demonstrated is encouraging and shows the utility of the method.

The paper however could use some improvement in explaining the following aspects:

1. One crucial aspect of the success of the method involves the validity of the assumption that context embedding generated by an image query is similar to that of the text query (as pointed out in section 3.A). The authors should explain / try to motivate as to why this should hold true for such different modalities of inputs, since this is not entirely obvious. This could also be interesting since this can possibly hint as to why the model cannot sufficiently perform well in case of complex queries (such as combination of two instruments as highlighted in section 5.D)

2. It is not clear from the manuscript as to what the authors wish to achieve through the use of figure 4, where the frequency scale for different spectrograms are missing. I would suggest, instead of several spectrograms, choose perhaps 3-4 and mark the regions of interest for the reader to even visually evaluate the spectrogram.

Reviewer #4

Questions

1. How confident are you in your evaluation of this paper?
confident

2. Importance/Relevance to you
Of broad interest

3. Novelty/Originality
Very original

4. Technical correctness
Probably correct

5. Experimental Validation and Reproducibility
Limited but convincing

6. Clarity of presentation
Clear enough

7. Reference to prior work

Excellent reference

8. Overall evaluation of the Paper
Strong accept

9. Justification (required if score of 1 or 2 has been selected for questions 3-7)
The paper presents a great idea of training on self-supervised image-audio pairs from videos and using text embeddings from CLIP for inference. Evaluation of such generative models is tricky, I appreciate the effort of also presenting the retrieval baseline.
Some of the presented examples on the webpage however question if the model actually generalizes to sound generation. It rather seems it learned the classes/training examples. This is obvious for "sharpen knife" which is rather some humans chattering about how to sharpen knifes or the goat bleating example, which is some sound non-related to a goat, but similar as in the ground truth example.

Reviewer #5

Questions

1. How confident are you in your evaluation of this paper?
confident

2. Importance/Relevance to you
Of sufficient interest

3. Novelty/Originality
Has been done before

4. Technical correctness
Definitely correct

5. Experimental Validation and Reproducibility
Limited but convincing

6. Clarity of presentation
Very clear

7. Reference to prior work
Does not cite relevant reference (implies reject)

8. Overall evaluation of the Paper
Strong reject

9. Justification (required if score of 1 or 2 has been selected for questions 3-7)
This paper proposes a method that learns a model to synthesize sound from images during training and synthesizes sound from text during inference using CLIP, a multimodal model of image and text. Although the method is quite straight forward, it succeeds in synthesizing high-quality sound compared to SpecVQGAN and Im2Wav. I found the paper itself easy to read and clearly written.

First, the content of this paper was previously presented at the CVPR 2023 Sight and Sound Workshop. Since there is no difference, there is no novelty in this paper, and at least the workshop paper should be properly cited in this paper. Both papers contain the typo "0.712n" in Table II, for example, and there is no doubt that they were created by copying and pasting.

Also, I know that arXiv papers should not be considered for peer review of international conference papers, but a paper with almost identical content by the same first author was submitted to arXiv in mid-June (to days after the MMSP deadline) as the result of an internship at Dolvy [R1]. Both claim to be the first method for synthesizing sound from text using only unlabeled video without text/sound pairs, but I could not know which one to believe, and I

believe the first author is accountable for this. At the very least, the first author knew of both methods prior to submission of this paper, so the arXiv paper should be properly cited, their claims narrowed if necessary, and evaluated as a comparison in the experimental section. Unless this point is resolved, I as a reviewer cannot recommend acceptance of this paper.

[R1] Dong+, "CLIPSONIC: TEXT-TO-AUDIO SYNTHESIS WITH UNLABELED VIDEOS AND PRETRAINED LANGUAGE-VISION MODELS". Available at: https://arxiv.org/pdf/2306.09635.pdf