AUDIO DIFFUSION RESEARCH

楊佳誠、邱以中、蔡桔析

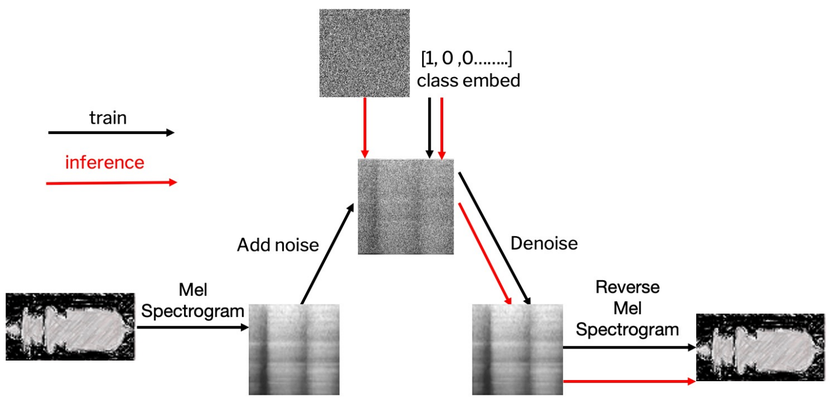
Introduction

- After reading the paper "Palette: A Simple, General Framework for Image-to-Image Translation," we found it interesting to investigate whether the ablation results are consistent with those in the audio domain.
- We will compare the L1 and L2 loss and also evaluate the significance of self-attention and normalization in the audio diffusion architecture.

Method

Model

- The audio diffusion backbone utilizes a U-Net architecture.
- The class_embedding of the UNet2DModel is utilized to incorporate both the time embedding and class_embedding, treating them as part of the conditional input of the model.



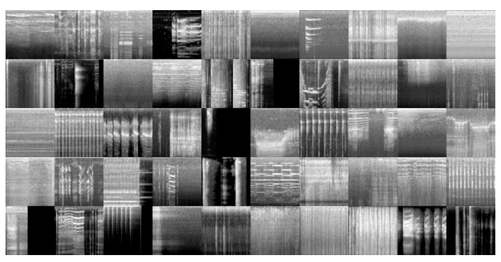
Dataset

- ESC-50 consists of 5-second-long recordings organized into 50 semantical classes, with 40 examples per class.
- This dataset consists of the following five main categories: Animals, Natural, Human non-speech sounds, Interior sounds, and Exterior noises.

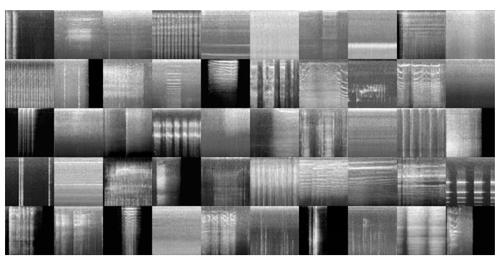
Data preprocessing

- The .wav data is preprocessed into Mel Spectrograms.
- The Mel spectrograms will be normalized according to the experiment setting.

Result



- ESC50

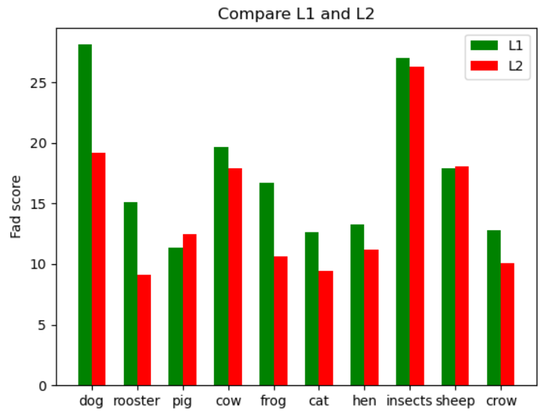


- Generate

Experiment

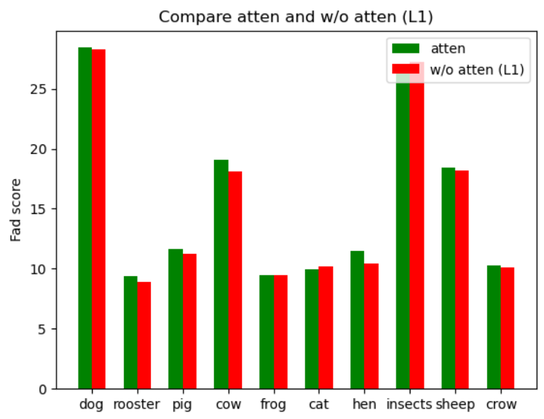
- We utilize our model to generate 50 classes of audio, producing 40 audio samples for each class as evaluation data.
- The FAD score is a metric employed to measure the similarity between evaluation data and original data. A lower FAD score indicates a closer match between the distributions of the generated and real audio.
- The CA score, utilizing pretrained Contrastive Language-Audio Pretraining (CLAP), is used to assess whether our model can successfully generate the correct voice.

L1 & L2 comparison



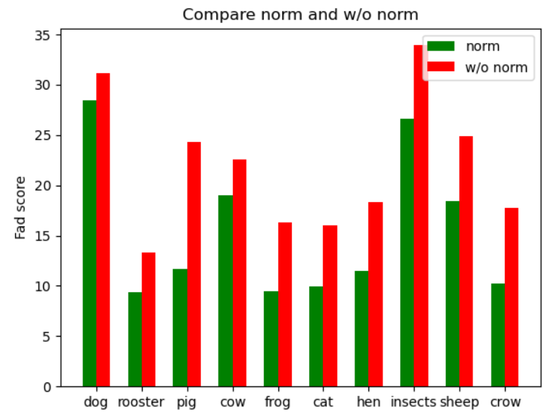
	Fad score	CA Top 1	CA Top 5	CA Top10	CA mAP10
L1	15.47	0.156	0.411	0.562	0.266
L2	**14.44**	**0.220**	**0.504**	**0.674**	**0.341**

Comparative Analysis of with and without Attention



	Fad scrore	CA Top 1	CA Top 5	CA Top10	CA mAP10
L1	15.47	0.156	0.411	0.562	0.266
L1 w/o attention	15.21	0.202	0.489	0.661	0.324

Comparative Analysis of with and without Normalization



	Fad scrore	CA Top 1	CA Top 5	CA Top10	CA mAP10
L1	**15.47**	**0.156**	**0.411**	**0.562**	**0.266**
L1 w/o normalize	21.85	0.076	0.262	0.425	0.160

Conclusion

- We conducted a series of experiments, including comparisons between L1 and L2 loss, as well as examining the effects of attention and normalization on the results.
- While we transformed audio into spectrograms, a visual representation, we observed variations in the outcomes compared to those obtained in image diffusion.
- In the future, we hope to conduct various evaluations on the results generated by the model, including generating diversity, among other factors. This will provide us with more reference points for structural optimization.

Reference
1. https://github.com/teticio/audio-diffusion?fbclid=IwAR2bj6KkH1OB-oX7HouirBzdlIGDbNjX7z6L3z_skTYqxm8ymGGe4tula5U
2. https://github.com/LAION-AI/CLAP?fbclid=IwAR3aweGGr-4o6JnjvlHT3xuYmB85eGPW6frVXAir7KkiVtmVHimkB-p7SuQ
3. https://github.com/gudgud96/frechet-audio-distance?fbclid=IwAR3Vvx41X9JS2eEkV5J7OlrtkInZ5HvLMHkj5LlQ9LcO9atnPmPAChyn9j0
4. https://github.com/huggingface/diffusers?fbclid=IwAR0bA1laZ8nGMkgiFGEvncPUHOgQXDOFOTRtL9F5yZSYmYR8YZtq6NAiz0I
5. https://arxiv.org/abs/2111.05826