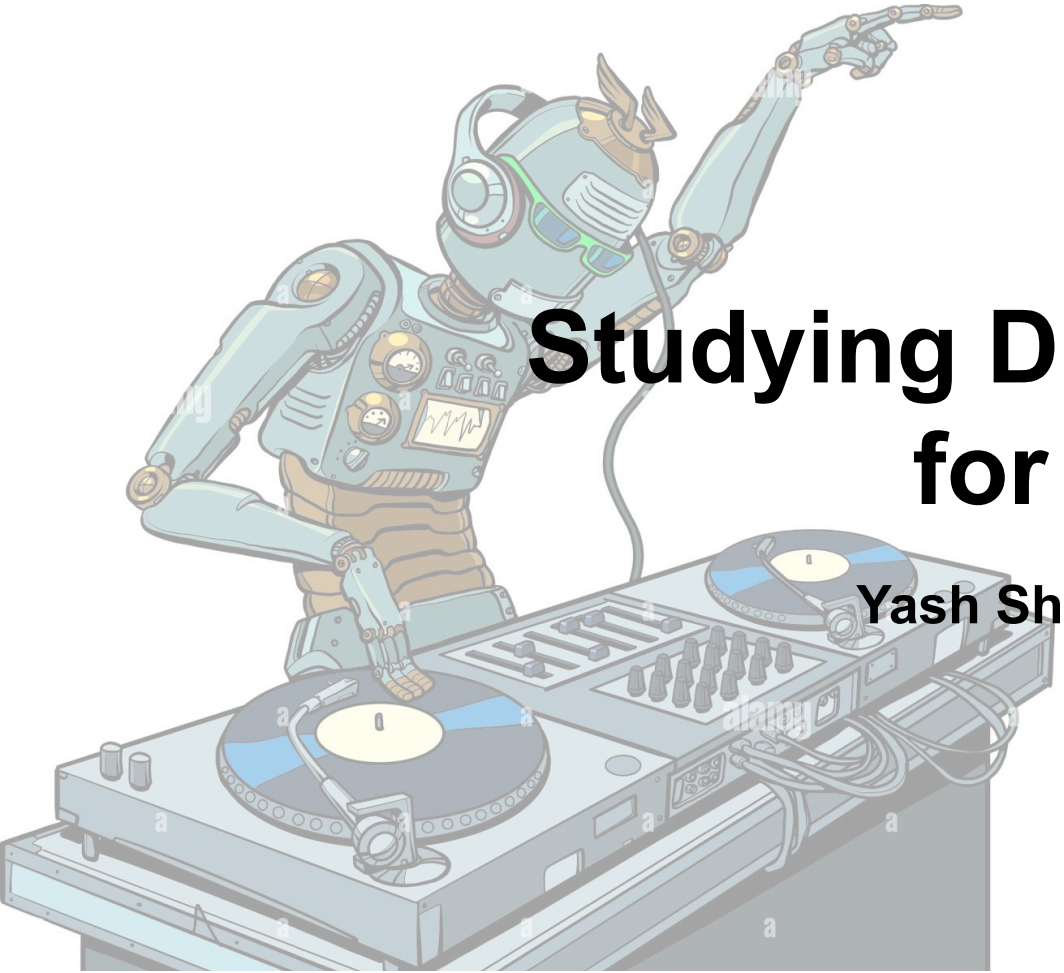


Studying Diffusion Model for Audio

Yash Sharma (ys4yh)

The Decade of **Generative** Model

Data

+

Model family, loss function, optimization algorithm, etc.

Learning

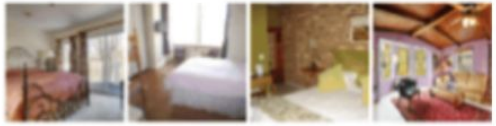
Prior Knowledge

Image x

A probability distribution p(x)

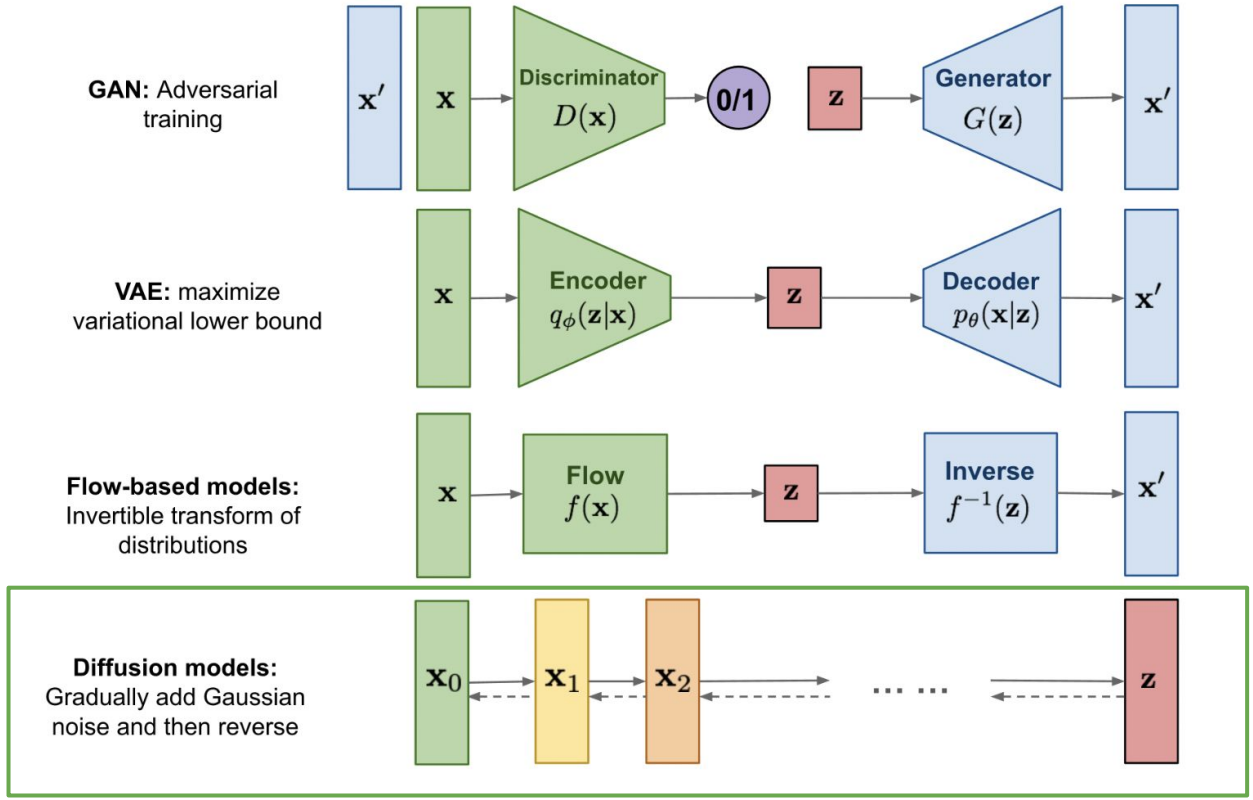
probability p(x)

Sampling from p(x) **generates** new images:



Source: http://home.ustc.edu.cn/~jclee/posts/2019/08/generative-models/

Different Types of **Generative** Models



Source: https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

Diffusion Models for **Audio**

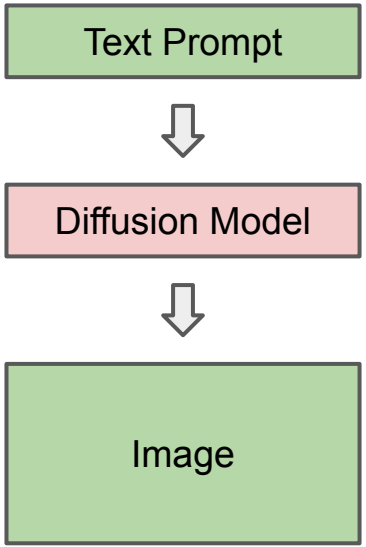
- Following the great success of the Text-to-Image Diffusion Model, many Audio Diffusion models have been proposed in recent months.

- As part of our project, we will be exploring two open-sourced audio diffusion projects:
  - **Riffusion**
  - **Hugging Face's Audio Diffusion**

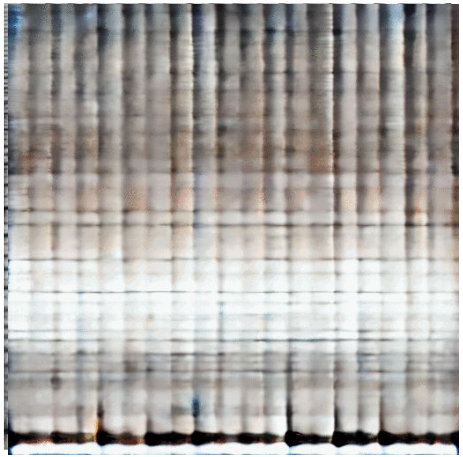
Riffusion

- Created by Seth Forsgren and Hayk Martiros as a hobby project.
- They **fine-tuned** the Stable Diffusion model to generate images of **spectrograms**!
- Did no modifications, just fine-tuned on images of **spectrograms paired with text**.

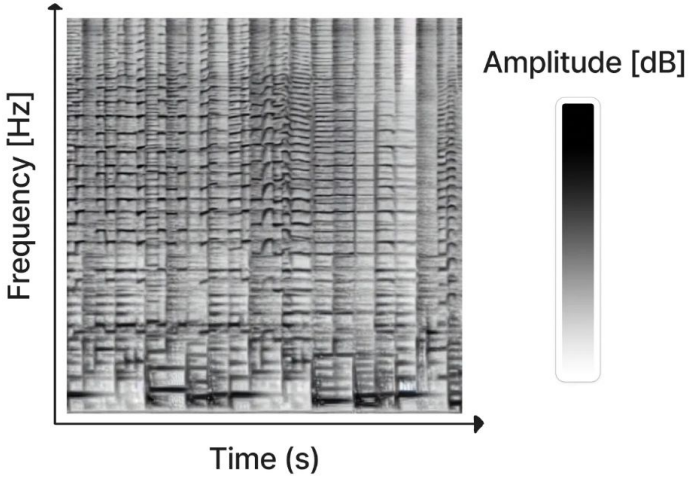
photograph of an astronaut riding a horse



funk bassline with a jazzy saxophone solo



How to convert **Spectrogram to Audio**?



- Visual representation of the frequency content of a sound clip.
- Can be computed from audio using our favorite STFT.

Journals & Magazines > IEEE Transactions on Acoustic...
> Volume: 32 Issue: 2

Signal estimation from modified short-time Fourier transform

Publisher: IEEE

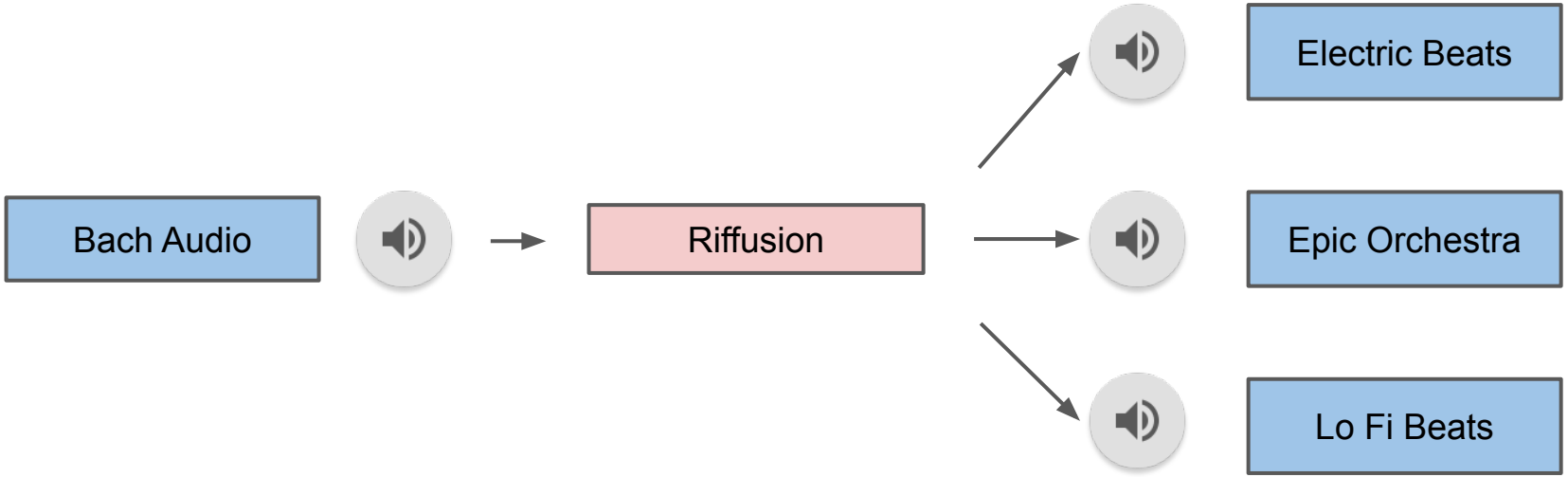
Cite This

D. Griffin ; Jae Lim

- Spectrogram images from model **only** contain the **amplitude** of the sine waves and **not** the phases, because the **phases** are chaotic and hard to learn.
- Use the **Griffin-Lim** algorithm to approximate the phase when reconstructing the audio clip.

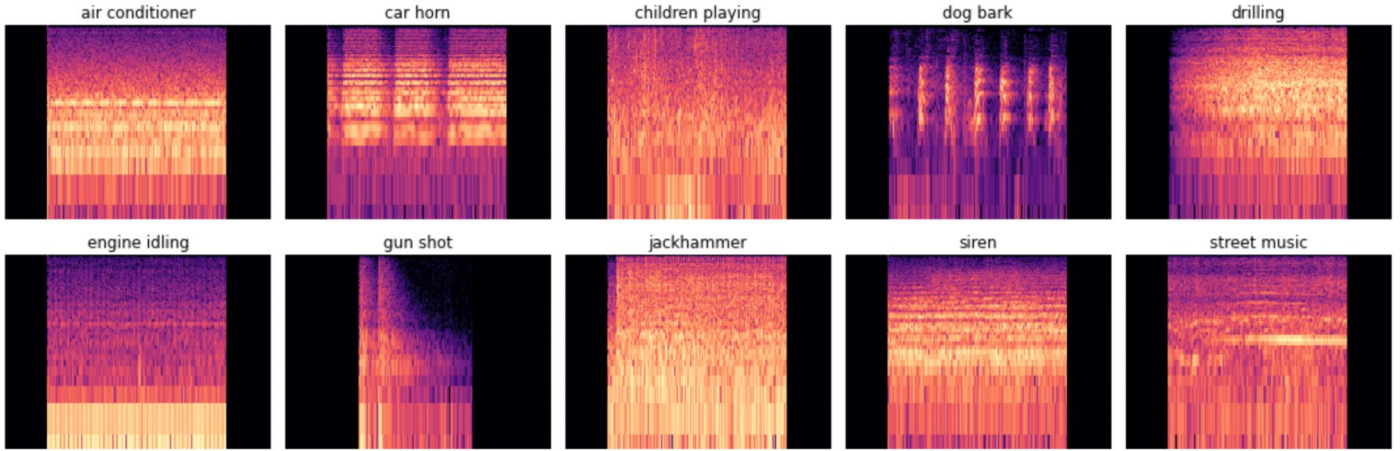
Style Transfer your own **Audio**

- We style transfer our BACH audio using different text prompts.



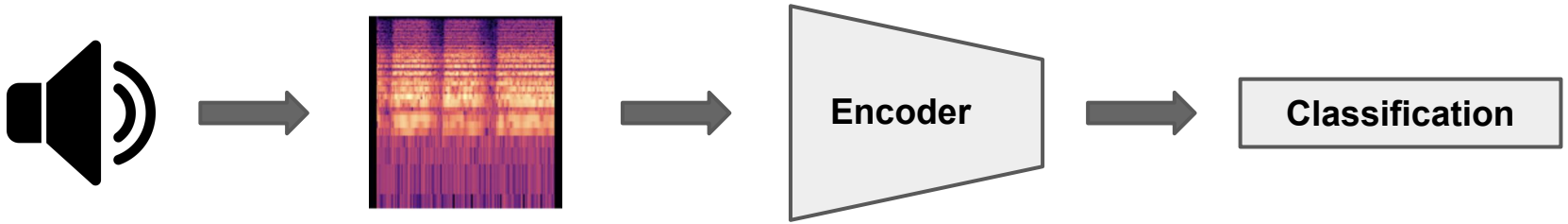
Audio **Classification**

- Using **UrbanSound8k Audio Dataset** for 10-class classification.
- Contains **8732** labeled sound excerpts (<=4s) of **urban sounds** from 10 classes: air_conditioner, car_horn, children_playing, dog_bark, drilling, enginge_idling, gun_shot, jackhammer, siren, and street_music.



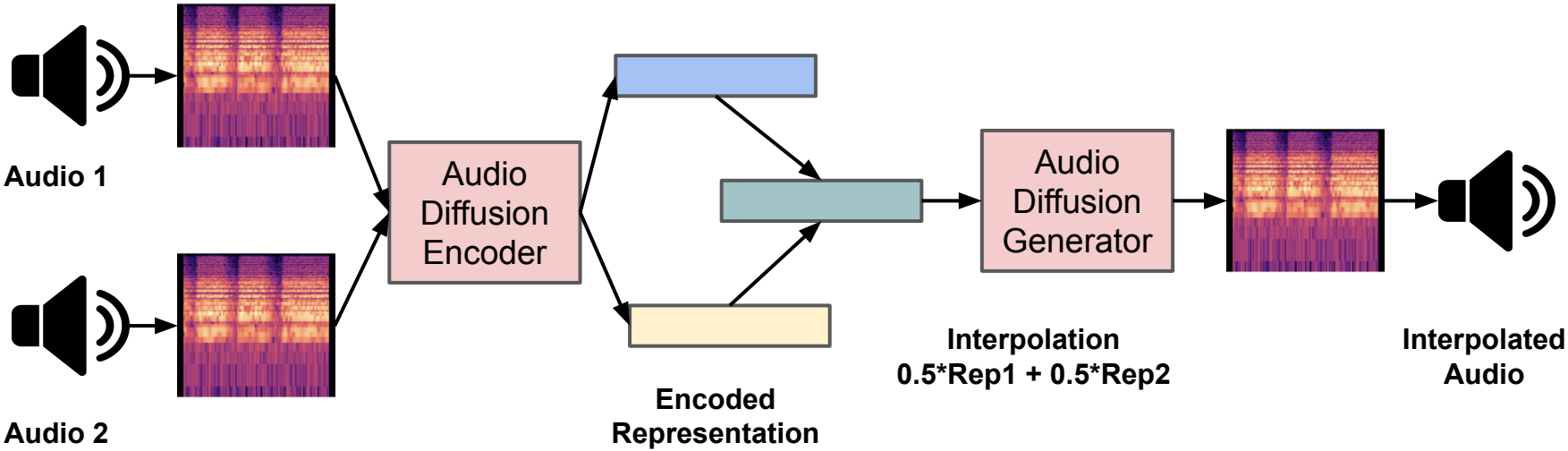
Deep Learning Experiment

- Converting **Audio** files **to Spectrogram** for CNN classification.
- Divided UrbanSound8k into 90:10 split for performance assessment.



Using **Generated** Audio for **Augmentation**

- One of the **downstream** applications of diffusion models can be **generating** audio files for **improving classification** performance.
- For scalability and implementation ease, used Hugging face audio diffusion model.
- Generated **100 new audios** using interpolation for each class.



Model Comparison

- Our model consisted of 3 blocks of CNN layers followed by two linear layers.
- **Base Model** Training - Trained model for 10-class classification for 50 epochs with a learning rate of 1e-3 and batch size of 128.
- **Augmented Model** Training - Trained model for 10-class classification with augmented audios included.
- **Observation:** We observe a **5% improvement** by the **inclusion** of **augmented audio**, demonstrating that diffusion models can be effectively used for audio signals to generate novel inputs for **improving** model training.

Model	Accuracy (%)
Base Model	57.73
Augmented Model	62.77

References

- **Riffusion**: https://www.riffusion.com/
- **Audio Diffusion**: https://github.com/teticio/audio-diffusion
- **UrbanSound8k Dataset**: https://urbansounddataset.weebly.com/urbansound8k.html
- **CNN Model Inspiration**: https://github.com/musikalkemist/pytorchforaudio