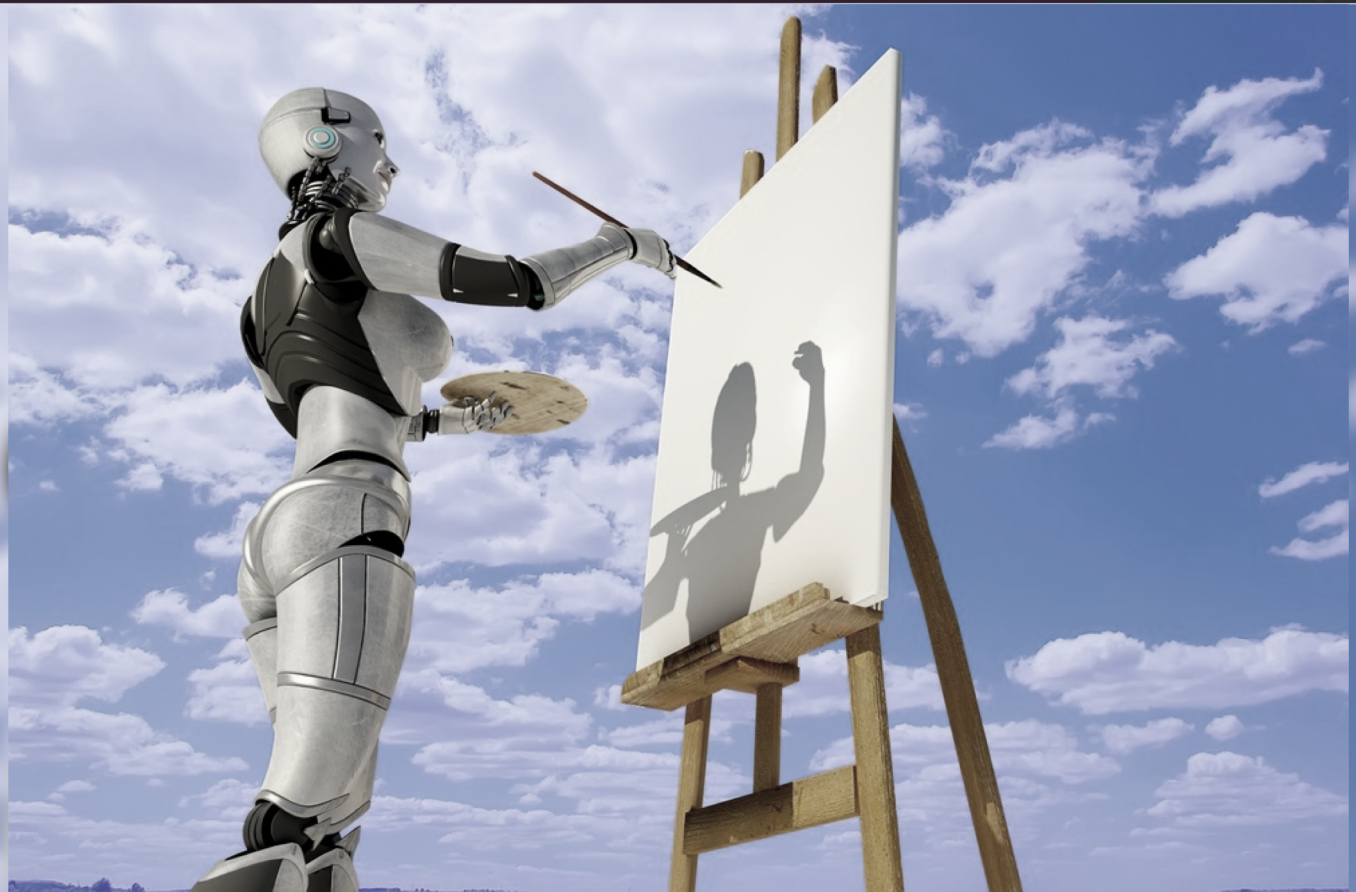
Music Generation

Hunter Hsieh, Hsien Fang Liu, Tsung-Min Huang

CMPE 256 Large-Scale Analytics

Introduction



Introduction

Dataset

- Piano dataset from “*Neural Translation of Musical Style*”

Method

- Feature Engineering
- Generative Adversarial Network (GAN)

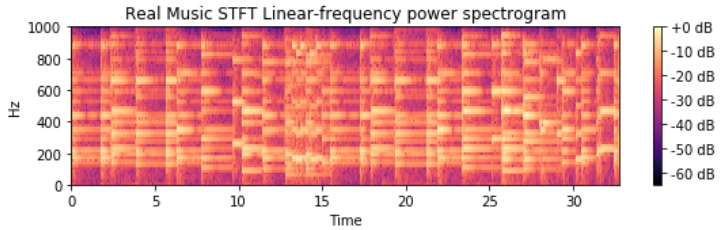
Feature Engineering

Mel Frequency Spectrum (MEL)

Constant Q Transform (CQT)

Difficult to inverse back to audio

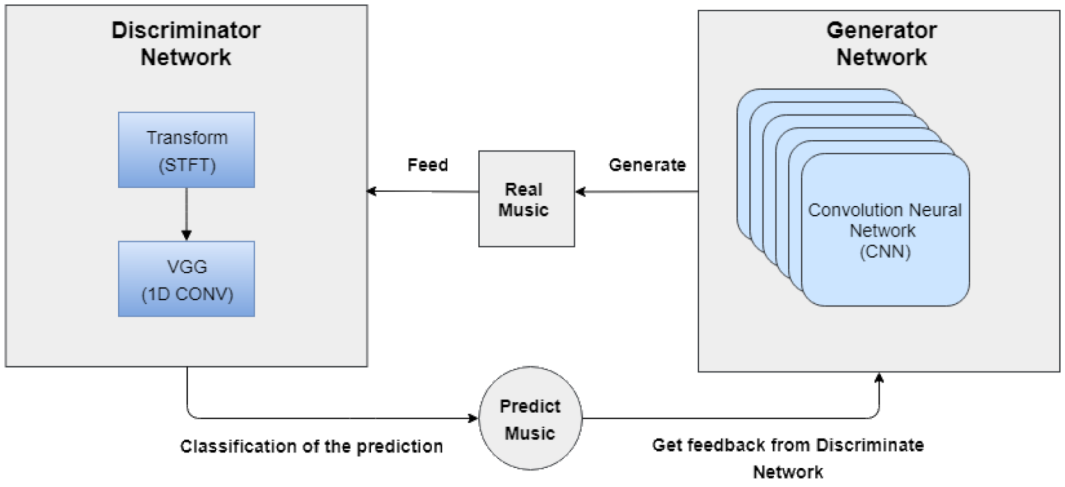
Short Time Fourier transform (STFT)



GAN Architecture

Discriminator Network:
Validated the generated music

Generator Network:
Generate new music from Gaussian noise



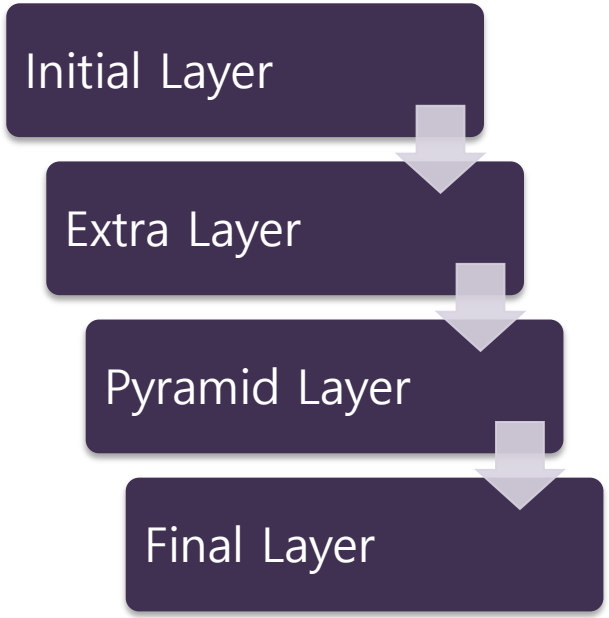
Discriminator Network Design

Initial layer:
Down Sample 2D Spectrum image.

Extra layer:
1D convolution along time axis.

Pyramid Layer:
Feature extraction along time and frequency.

Final Layer:
Leverage the input is similar to real sample.



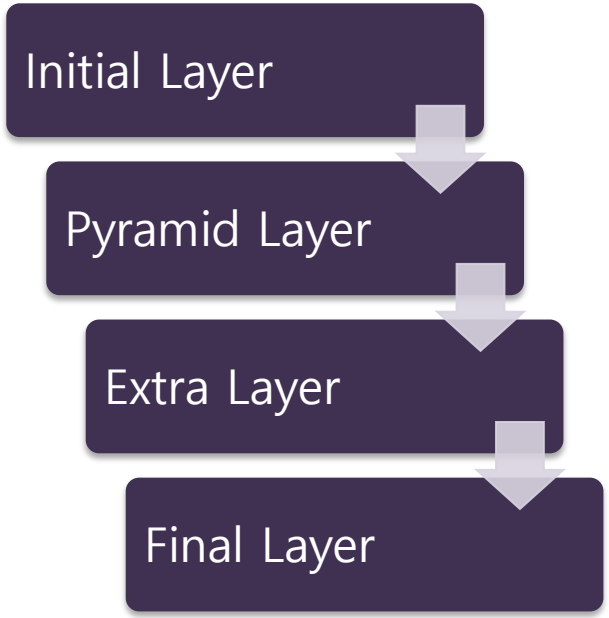
Generator Network Design

Initial layer:
Convert Gaussian noise to feature representation.

Pyramid layer:
Transpose convolution operation to unfold noise as feature representation along time and frequency axis.

Extra Layer:
Transpose convolution along the time axis.

Final Layer:
Generate fake samples.



Experiment

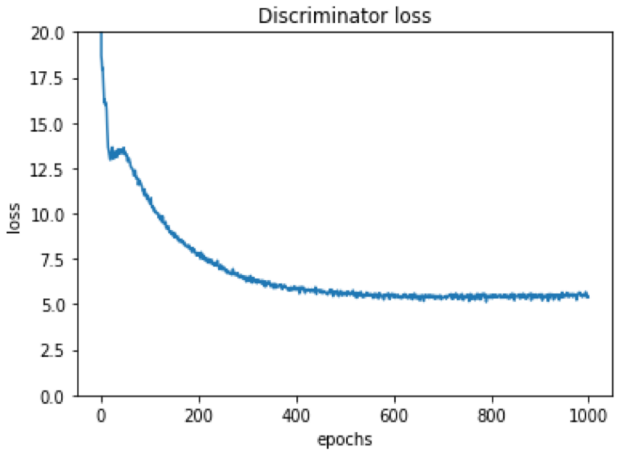
Preprocessing:

- MIDI to mp3 (44.1KHz)
- Down Sampling (44.1KHz -> 2KHz)
- Random slice 32 seconds

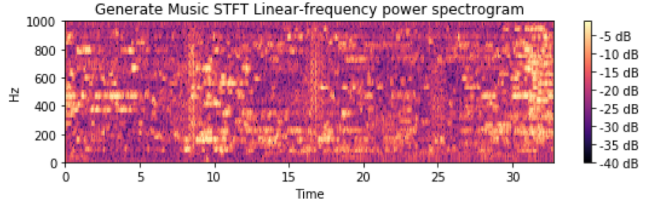
STFT

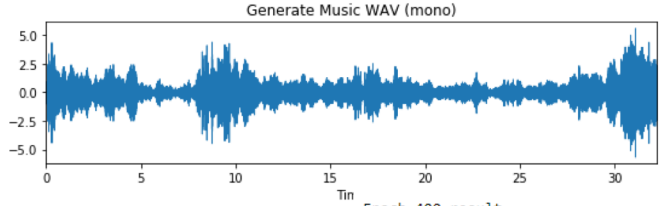
- FFT windows: 128
- Half overlap width: 64

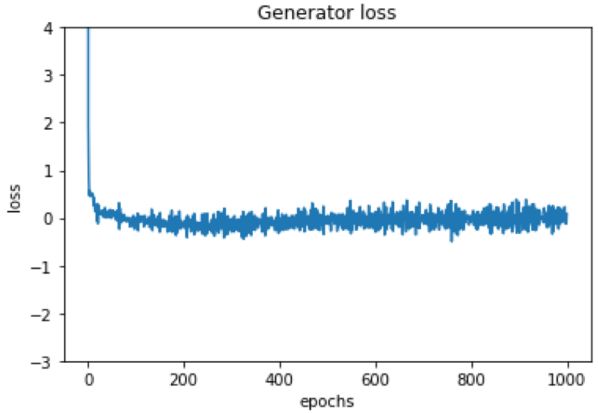
Experiment



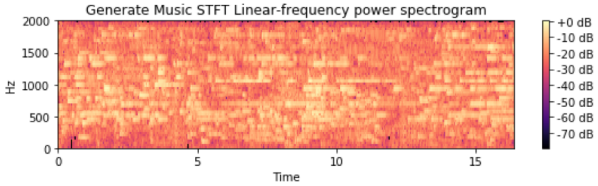
Epoch 100 result

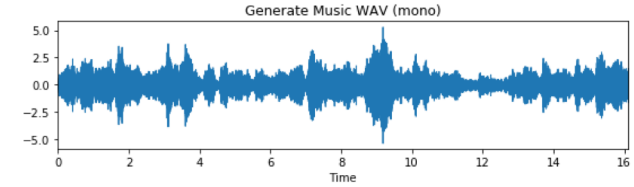






Epoch 400 result





Demonstration

Generated music with Epoch 100:

Generated music with Epoch 400:

Generated music with Epoch 1000:

Demonstration

Generated music with Epoch 1000:

MIDI conversion

Conclusion

- We demonstrate that music can be generated from spectrum image.
- Audio quality increases along with the number of epochs
- Network architecture optimization
  - Deeper convolution layer
- Different feature representation:
  - Wavelet transform