Exploiting Pre-trained Feature Networks for Generative Adversarial Networks in Audio-Domain Loop Generation

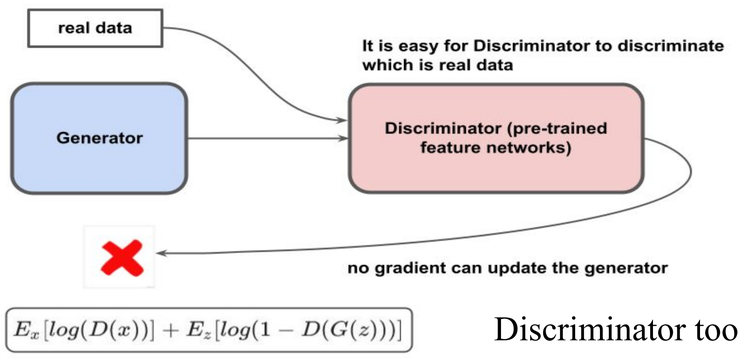
Yen-Tung Yeh[1,2], Bo-Yu Chen[1,2], Yi-Hsuan Yang[1,3]

[1] Academia Sinica, Taiwan, [2] National Taiwan University,[3] Taiwan AI Labs

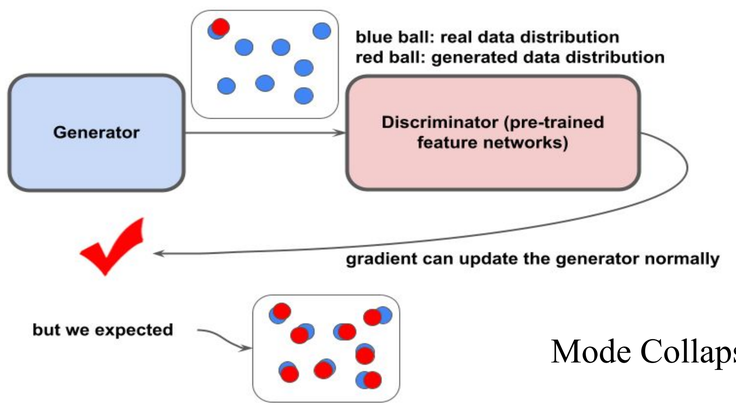
1. Introduction

GAN is a powerful generative model. It consists of a generator and a discriminator.
Many SOTA audio generation models use GAN; e.g., loop generation (ISMIR'21)

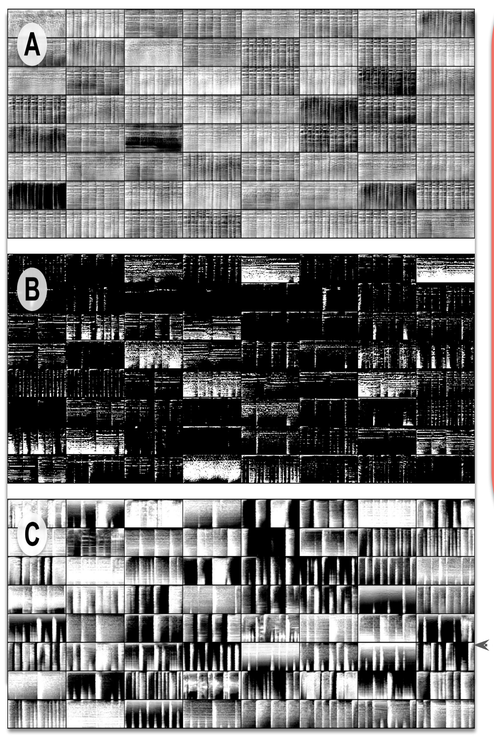
Failure Case



Discriminator too strong



Mode Collapse



2. Motivation

GAN can generate impressive results. However, it is unstable and hard to train. The first two are different failure cases of GAN, and the third is successful case.

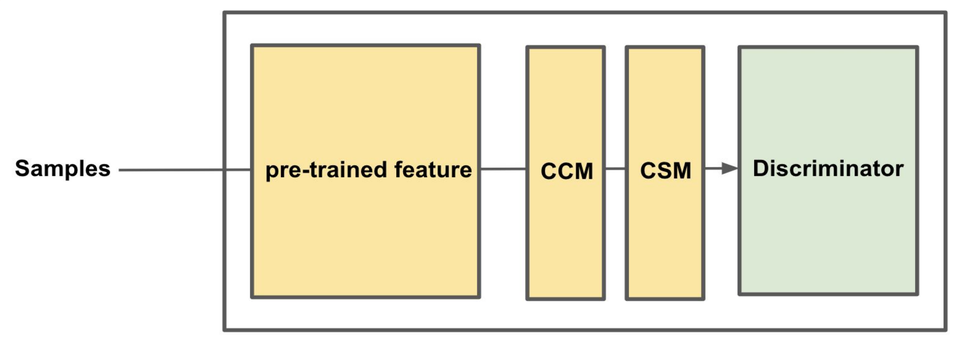
Failure cases:

A. Mode collapse: the generator only generate a certain type of data. (Q1)
B. Discriminator is too strong. It leads to vanishing gradients. (Q2)

Moreover, GAN is time-consuming to converge. (Q3)

How can we avoid Q1 and Q2 and improve Q3?

3. Keys



1. Use pre-trained feature to speed up training → Q3
2. Use CCM and CSM to prevent mode collapse → Q1
3. Use a "**proper**" feature network to prevent the discriminator from dominanting the training process → Q2

===> **1.** and **2.** are ideas from **Projected GAN,** but their effectiveness has not been evaluated for audio generation.

===> as for **3**: before our work, we didn't know which feature network is "**proper**", so we study it in this paper.

4. Setup

General

- VGGish
- Pre-trained on Youtube-100M
- Music + sound

Domain-specific

- ShortCunk CNN
- Pre-trained on MTAT (auto-tagging)
- Pre-trained on Loopermen (genre classifier)
- Music

Fusion

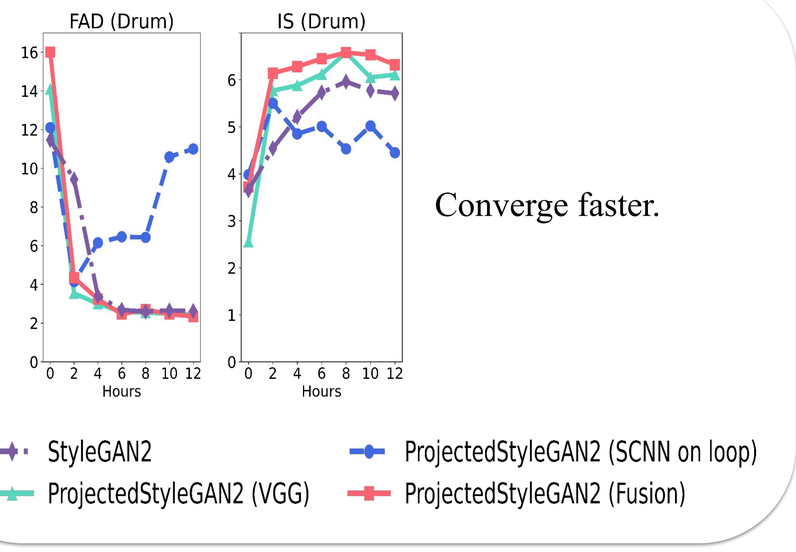
- combination of general and domain-specific features

5. Results

Models	Drum loops IS ↑	FAD ↓	D ↑	C ↑
A Real data	16.30	0.01	1.00	1.00
B StyleGAN2	5.58	2.50	1.01	**0.89**
D Projected StyleGAN2 (VGG)	5.87	3.03	1.06	0.70
E Projected StyleGAN2 ($SCNN_{MTAT}$)	4.75	8.18	0.00	0.00
F Projected StyleGAN2 ($SCNN_{Loop}$)	4.45	7.21	0.00	0.00
G Projected StyleGAN2 (VGG+$SCNN_{MTAT}$)	6.22	2.45	**1.11**	0.74
H Projected StyleGAN2 (VGG+$SCNN_{Loop}$)	**6.31**	**2.34**	1.08	0.73

We show the evaluation of drum loops. We also generate synth loops, which are similar to drum loops. Please see the paper and the demo page if you are interested in the result of synth loops.
D & C: Density and Coverage.
D: quality
C: diversity

- Domain-specific features lead to high FAD and nearly zero density and coverage.
- General features work better and achieve competable results.
- Fusion achieves lower FAD, and higher IS against general features.



Converge faster.

demo page

github page

6. Conclusion and Future work

- ➢ We find out general features are more helpful to loop generation.
- ➢ Loop generation is still far from human-made loops. The improvement of loop generation is largely unexplored.
- ➢ Vocoder limitation
- ➢ Unsupervised, Self-supervised pre-trained features
- ➢ If you are interested in our work, please check our paper and demo page.

ISMIR 2022 BENGALURU

National Taiwan University

ACADEMIA SINICA 1928

Music AI Lab