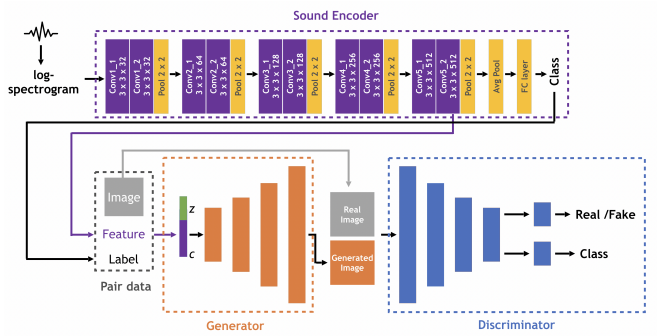
S2I-Bird: Sound-to-Image Generation of Bird Species using Generative Adversarial Networks

Joo Yong Shim, Joongheon Kim, and Jong-Kook Kim
School of Electrical Engineering, Korea University

Abstract

Generating images from sound is a challenging task. This paper proposes a novel deep learning model that generates bird images from their corresponding sound information. Our proposed model includes a sound encoder in order to extract suitable feature representations from audio recordings, and then it generates bird images that corresponds to its calls using conditional generative adversarial networks (GANs) with auxiliary classifiers. We demonstrate that our model produces better image generation results which outperforms other state- of-the-art methods in a similar context.

Model



Dataset

- Caltech- UCSD Birds-200-2011 (CUB) dataset
- Xeno-Canto collaborated database

Training

Loss function for realness prediction *Ls* :

$$L_S = \mathbb{E}\left[\log P(S = real|X_{real})\right] + \mathbb{E}[\log P(S = fake|X_{fake})]$$

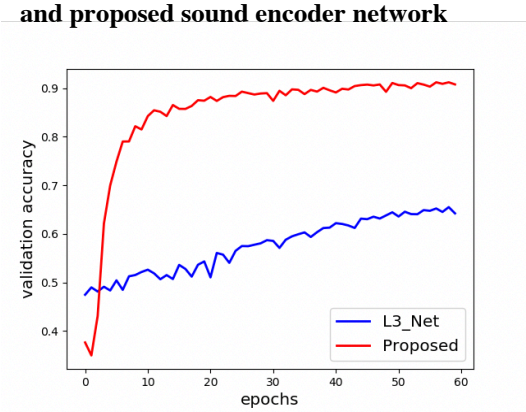
Loss function for class prediction L_L :

$$L_L = \mathbb{E}[\log P(L = label|X_{real})] + \mathbb{E}[\log P(L = label|X_{fake})]$$

GAN discriminator is trained to maximize $Ls + L_L$ as well as the GAN generator is trained to maximize $-Ls + L_L$.

Result

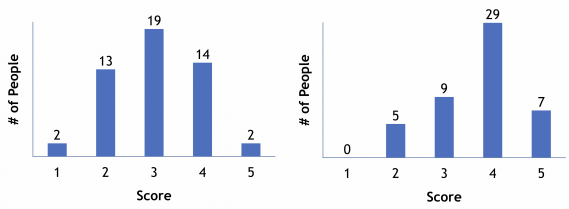
- **Training history on classification accuracy of L3 net and proposed sound encoder network**



- **Inception score of generated images**

Input condition	Inception Score
Upper Bound	4.10 ± 0.19
Label	2.81 ± 0.19
Mel-Spectrogram	1.76 ± 0.07
Our Sound Encoder	3.86 ± 0.31

- **Human evaluation on generated images**



Label Sound encoded features

- **Samples of generated images on sound encoded feature condition.**