DCGAN 의 잠재 벡터 보간을 활용한 두 음성 합성 방법

A method of mixing two audio signals using interpolation of latent vectors with DCGAN.

허찬영
정보통신공학과
명지대학교
용인시, 대한민국
hcn98@naver.com

정재희*
정보통신공학과
명지대학교
용인시, 대한민국
jhjung@mju.ac.kr

요 약

기계 학습 및 딥러닝 기술의 발전은 문학 분야를 비롯한 다양한 예술 분야에서 인공지능이 그림을 그리고 소설을 쓰거나 음악을 작곡, 작사하는 것과 같이 큰 영향력을 끼치고 있다. 이 중 인공지능이 음악을 작곡, 작사하는 음성을 생성하는 분야에서도 이미지 생성에 특화된 GANs(Generative Adversarial Nets) 모델을 사용하여 음성을 생성하는 연구를 적용할 수 있다. 하지만 음성 데이터 자체로 학습하여 음성을 생성하는 데에는 GANs 를 사용할 경우 적절한 음성 생성의 결과를 얻지 못한다. 따라서 음성을 이미지로 변환하여 GANs 을 학습한 후, 이미지를 생성하여 이를 다시 음성으로 생성하는 방법으로 음성 생성을 할 수 있다. 본 연구에서는 CNN(Convolution Neural Network) 기반의 GANs 모델인 DCGAN(Deep Convolutional Generative Adversarial Network) 모델을 활용하여, 두 개의 생성된 음성 이미지에서 추출된 잠재 벡터 z 들의 보간의 정도에 따라 생성된 이미지가 부드럽게 변하는 특징을 적용하여 음성 합성 방법을 제안한다. 두 개의 서로 다른 음성 포맷인 midi 파일과 wav 파일을 각각 이미지로 변환 후 모델을 학습시켰다. 두 포맷 모두 두개의 음성 이미지의 잠재 벡터의 보간 정도에 따라 생성된 이미지가 부드럽게 변환되었고, 각 보간 값의 정도에 따라 생성된 이미지들을 다시 음성으로 변환시켜 적절히 합성된 음성을 확인할 수 있었다.

키워드 : GANs(Generative Adversarial Nets), 음성 합성, 잠재 벡터의 보간, 음성 생성

1. 서론

인공지능 분야에서의 기계학습 및 딥러닝 기술의 발전은 문학 분야를 비롯한 다양한 예술 분야에서도 큰 영향력을 끼치고 있다. 최근에는 인공지능이 소설과 같은 글을 작성하거나 그림을 그리거나 음악을 작곡하는 등 다양한 예술 작품을 생성하는 데 성공하는 결과를 보이고 있다.

이미지 생성 분야에서는 GANs(Generative Adversarial Nets)[1]의 발전으로 인해 매우 높은 성능을 보이고 있다. 하지만 음성 생성 분야에서는 아직 인간의 청각에 대한 본질적인 문제로 인해 GANs 모델이 상대적으로 낮은 효율을 보이고 있다.

따라서 음성 생성 분야에서의 GANs 기술의 한계를 극복하기 위해 음성을 이미지화 하여 CNN(Convolutional Neural Network)기반의 GANs 모델인 DCGAN(Deep Convolutional Generative Adversarial Network)[2]모델을 학습시켜 음성을 생성하는, 즉 음성으로 음성을 생성하는 것이 아닌 음성을 이미지로 변환하여 이미지를 생성하고 이를 다시 음성으로 변환하여 음성을 생성하려는 연구가 활발히 진행되고 있다.

본 연구에서는 GANs 의 Base 모델인 DCGAN 모델에서 생성된 이미지의 특징 중 하나인 두 생성된 이미지에 대한 잠재 벡터(latent vector) z 에 보간을 적용 시, z 에 따라 생성된 이미지의 부드러운 변화가 일어나는 특징을 활용하고자 한다. 음성을 이미지로 변환하고 생성된 이미지의 잠재 벡터 보간에 따른 변화가 음성의 합성에 대해 유효한지 검증하고자 한다.

2. 관련연구

WaveGAN[3]은 2018 년 GANs 를 통해 음성을 생성하고자 하는 첫 번째 시도에 해당하는 연구이다. WaveGAN 은 기존 2 차원 이미지를 학습하는 DCGAN 구조에 전치 컨볼루션 (transposed convolution)을 적용해, 입력을 1 차원으로 변형함으로써 16,384 차원의 1 차원 벡터 이미지를 생성하여 음성을 생성하고자 하였다. 또한, 음성을 Spectrogram 으로 이미지로 변환하여 음성을 생성하는 SpecGAN 또한

* 교신저자

제안하였다.

2019 년 Google AI 에서는 음성을 이미지 형태의 고해상도 Spectrogram 으로 변환하여 GANs 를 학습하고 음성을 생성하는 GANsynth[4] 를 제안했다. WaveGAN 에서도 음성을 Spectrogram 화하여 이미지로 학습하는 SpecGAN 을 제안하였으나 SpecGAN 이 고해상도 Spectrogram 을 학습하지 못하는 DCGAN 기반으로 구성되어 성능이 떨어지는 문제가 존재했다. GANsynth 에서는 이를 개선하여 고해상도 이미지를 안정적으로 생성하는 PGGAN[5] 구조를 사용하였다.

이처럼 음성을 이미지로 변환하여 음성을 생성하려는 연구는 지금까지 활발하게 이어지고 있다.

3. 실험방법

본 연구에서는 DCGAN 의 두 잠재 벡터 z 의 보간된 z 에 따라 생성된 이미지의 부드러운 변화가 일어나는 특징을 활용하였다.

3.1. Data

본 연구에서는 midi 음성 포맷 데이터와 wav 음성 포맷 데이터를 활용하여 두 음성 데이터를 이미지로 변환 후 실험을 각각 진행하였다.

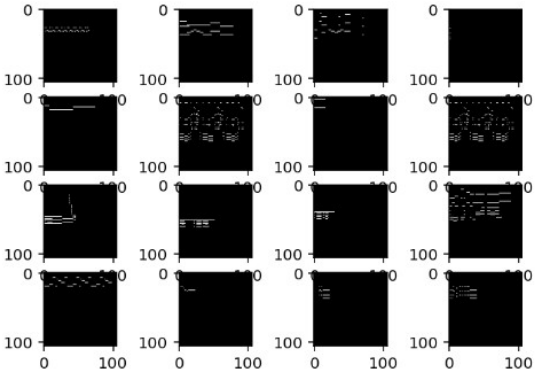

그림 1. midi 음성 포맷의 이미지 예시

그림 1 은 midi 음성 포맷을 검은색 바탕에 악보를 의미하는 흰색으로 표현되는 2 차원 이미지로 변환한 예시이다. midi 음성 데이터는 Lakh MIDI Dataset 을 활용하였으며, Midi2image 패키지를 통해 midi 파일들을 (106,106,1)형태의 이미지로 변환하였다. midi 음성 데이터의 실험을 위해 CNN 으로 2 차원 이미지로 생성하는 DCGAN 모델을 사용하였다.

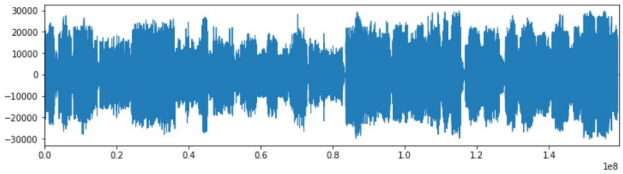

그림 2. wav 음성 포맷의 이미지 예시

그림 2 는 wav 음성 포맷을 시간에 따른 파형 값의 1 차원 이미지로 변환한 예시이다. wav 음성 데이터는 wav 포맷의 1 시간의 피아노 연주 음악을 librosa 패키지를 통해 (57,794,996, 1)형태의 tensor 로 변환한 후, 16,384 차원의 vector 들로 slice 하여 (3527, 16,384, 1)으로 구성하였다. wav 음성 데이터를 실험을 위해 CNN 으로 1 차원 이미지로 생성하는 WaveGAN 모델을 사용하였다.

3.2. 보간 방법

실험 단계는 첫째, 음성을 이미지로 변환하여 DCGAN 을 통해 생성하고, 둘째, 이미지를 음성파일로 변환하기 전, 잠재벡터 z 에 따라 보간된 DCGAN 의 이미지를 생성하게 된다. 마지막으로, 보간된 이미지를 다시 음성으로 변환하며 z 에 따른 다양한 생성물을 확인하였다.

알고리즘 1. 두 잠재 벡터 간의 보간

Algorithm. Interpolate latent vectors
$n\_steps : num\ of\ interpolations$ $z1 : latent\ vector\ 1$ $z2 : latent\ vector\ 2$
$1 : ratios = [\,]$ $2 : vecotrs = [\,]$ $3 : interpolation\_length = (1 - 0) / n\_steps$ $4 : for\ \ i\ in\ n\_steps$ $5 : \quad ratios.append(0 + (interpolation\_length \times i))$ $6 : for\ r\ in\ ratios$ $7 : \quad vector = (1.0 - r) \times z1 + r \times z2$ $8 : \quad vectors.append(vector)$

알고리즘 1 은 두 잠재 벡터에 대해 설정한 값만큼 나누어 보간하는 과정을 나타낸 표이다. 입력은 두 음성에서 이미지로 변환된 잠재 벡터 2 개와 보간의 정도이다. 몇 개의 구간으로 내분, 즉 보간할지

정하고 정한 구간 수만큼 두 잠재 벡터를 보간하게 된다.

4. 실험결과

본 실험은 3.1 의 Data 로 학습된 WaveGAN 과 DCGAN 모델에 대해 생성된 이미지를 다시 음성으로 변환하기 전 z1 과 z2 두 개의 잠재 벡터에 대해 10 개의 구간으로 보간하여 이미지의 변화를 확인하였다.

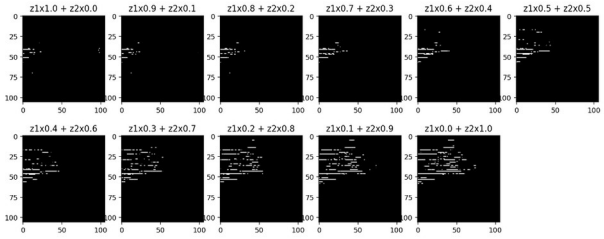

그림 3. midi 파일의 이미지 보간 예시

그림 3 은 1 행 1 열의 이미지가 z1, 2 행 5 열의 이미지가 z2 에 해당한다. 1 행에서 1 열에서 5 열로 이어 2 행에서 1 열에서 5 열로 갈수록 부드럽게 변하는 것을 확인할 수 있다.

이는 두 잠재 벡터의 이미지가 보간 정도가 반영되어 이미지가 합성됨을 의미한다.

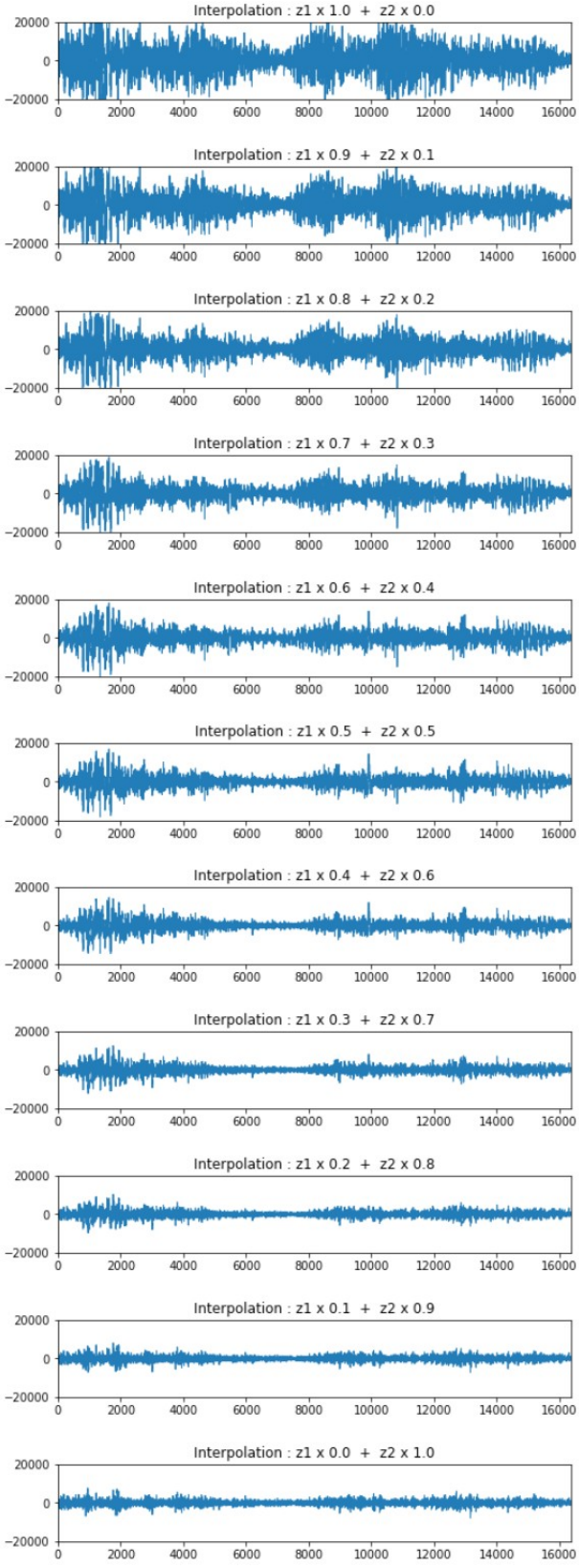

그림 4. wav 파일의 이미지 보간 예시

그림 4 는 첫 행이 z1, 마지막 행이 z2 로 생성한 이미지에 해당하고 첫 행과 마지막 행 사이의 이미지들은 z1 과 z2 의 보간을 통해 생성한 이미지이다. z1 에서 z2 로 갈수록 부드럽게 파형이

z2 에 가깝게 변화하는 것을 역시 확인할 수 있다.

또한, 그림 4 의 6 행은 z1 과 z2 의 잠재 벡터를 0.5 * z1 + 0.5 * z2 로 보간한 결과이고 z1 과 z2 로 생성된 이미지가 절반씩 합성되었고, 이는 보간 값에 따라 z1 과 z2 가 합성된 것이다.

이 두 차례 실험 결과를 통해 두 개의 잠재 벡터의 보간으로 생성된 이미지들이 보간 값에 따라 부드럽게 변하기에, 이미지를 음성으로 변환할 시 부드럽게 합성됨을 확인할 수 있다.

5. 결론

본 연구에서는 DCGAN 을 활용하여 음성을 이미지로 변환하여 서로 다른 두 이미지의 보간에 따른 이미지를 생성 후 음성을 확인하였다. DCGAN 의 잠재 벡터 z 의 보간을 통해 z 에 따라 생성된 이미지의 부드러운 변화가 일어나는 특징을 활용한 생성된 이미지를 다시 음성으로 변환하여 음성을 합성하는 방법을 제안한다. 두 개의 서로 다른 음성 포맷 midi, wav 데이터를 통해 음성을 이미지로 변환하여 학습 후 보간을 적용한 결과를 확인하였다. 이미지가 음성을 완전히 대표하여 표현할 수 있을 때 강하게 적용할 수 있다는 단점이 있을 수 있으나, DCGAN 의 특징만을 이용하여 생성한 음성을 합성할 수 있는 새로운 방법이다. 추후에 다른 방법을 통해 음성을 이미지로 완전히 변환한다면 더 효과적인 합성이 가능할 것으로 전망한다.

참 고 문 헌

[1] Goodfellow, Ian, et al, "Generative adversarial networks.", Communications of the ACM, 63.11 ,139-144, 2020

[2] Radford, Alec, Luke Metz, and Soumith Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks.", arXiv 1511.06434, 2015

[3] Donahue, Chris, Julian McAuley, and Miller Puckette, "Adversarial audio synthesis." , arXiv:1802.04208, 2018

[4] Engel, Jesse, et al, "Gansynth: Adversarial neural audio synthesis." , arXiv:1902.08710, 2019

[5] Karras, Tero, et al, "Progressive growing of gans for improved quality, stability, and variation." , arXiv:1710.10196, 2017