J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr

MCycleGAN: 잡음이 포함된 음성에서 아이 음성 추출을 위한 CycleGAN 기반의 딥러닝 모형

MCycleGAN: CycleGAN-based deep learning model for child speech extraction from noisy speech

손 수 락[1] 심 규 정[1] 정 이 나[1*]
Su-rak Son Kyu-jeong Sim Yi-na Jeong

요 약

아기를 돌보는 로봇에게 가장 중요한 기술은 아기의 상태를 파악하는 것이다. 아기는 주로 울음소리의 패턴으로 자신의 상태를 표현하기 때문에, 음성을 통해 아기의 상태를 분류하는 연구가 활발히 이루어지고 있다. 대부분 아기의 상태를 분류하는 연구들은 잡음이 포함되지 않은 깔끔한 아기 음성으로 아기의 상태를 파악했다. 하지만 실제 환경에서 수집된 아기의 음성 데이터는 내부에 잡음이 포함되어 있을 가능성이 크다. 따라서, 음성 데이터 속의 잡음을 처리해야 한다. 본 논문은 잡음 처리를 위해, Cycle GAN 기반 딥러닝 모델인 MCycle GAN(Multiple Cycle Generative Adversarial Net)을 제안한다. MCycle GAN은 더욱 정밀한 잡음 처리를 위해, 기존 Cycle GAN에서 Cycle을 다중으로 배치한 모델이다. 다수의 생성자와 소수의 판별자가 적대 관계로 학습함으로 판별자의 판별 성능을 향상하고, 생성자는 판별자를 속이기 위해 더 정밀한 위조 데이터를 생성해야 한다. 실험 결과, MCycle GAN 모델은 Cycle에 비해 더 많은 학습 시간이 소요되지만, 더 강화된 판별자의 판별 성능과 생성자의 위조 데이터 생성 성능을 보였다. 그러나 너무 많은 Cycle을 가질 경우, 늘어난 학습 시간에 비해 적은 성능 향상을 볼 수 있었다.

☞ 주제어 : 사이클적대생성망, 다중 사이클, 아기 상태 분석, 잡음 처리, 스펙트로그램

ABSTRACT

The most important technology for a robot to take care of a baby is to understand the baby's condition. Since babies mainly express their status through the pattern of crying sounds, research to classify the baby's status through voice is being actively conducted. Most of the studies that classify the baby's condition identified the baby's condition with a clean baby voice without noise. However, the baby's voice data collected in the real environment is likely to contain noise inside. Therefore, it is necessary to process the noise in the voice data. This paper proposes MCycle GAN (Multiple Cycle Generative Adversarial Net), which is a cycle GAN-based deep learning model for noise processing. MCycle GAN is a model in which multiple cycles are arranged in the existing Cycle GAN for more precise noise processing. The discrimination performance of the discriminator is improved by learning the adversarial relationship between a large number of generators and a small number of discriminators, and the generator needs to generate more precise forged data to deceive the discriminator. As a result of the experiment, the MCycle GAN model takes more training time than Cycle, but it showed stronger discriminant discrimination performance and generator forged data generation performance. However, when there are too many cycles, a small performance improvement can be seen compared to the increased learning time.

☞ keyword : CycleGAN, multiple Cycle, Baby condition analysis, noise processing, spectrogram

1. 서 론

로봇공학 분야에서 아기를 돌보는 로봇을 구현하고자 할 때, 가장 핵심적인 기술은 아기의 상태를 파악하는 것이다. 아기는 자신의 상태를 표정, 음성, 심박수 등으로 표현한다. 현재 기술은 표정과 심박수의 경우 카메라나 심박수 센서를 통해 아기의 관련 데이터를 인식한다. 이는 아기의 얼굴이 카메라를 계속 향해 있어야 하거나, 아기가 심박수 센서를 불편해하는 문제가 있다. 그러나 오디오 센서는 이들에 비해 비용이 저렴하고, 제한이 적다는 이점이 있다. 따라서 울음소리의 패턴을 분석하는 것은 아기의 상태를 파악하기 위해 매우 중요한 요소이다. 대부분 아기 상태를 분류하는 연구는 잡음이 포함되지

1 Software Department, Catholic Kwandong University, Gangneung-si, Gangwon-do, 25602, Korea.
* Corresponding author (lupinus07@nate.com)
[Received 18 October 2022, Reviewed 18 November 2022(R2 1 February 2023, R3 31 March 2023), Accepted 8 April 2023]

않은 깔끔한 아기 음성으로 인공지능의 학습을 진행한다. 하지만 실제 환경에서 데이터를 인식할 때, 오디오 센서가 감지한 음성 데이터는 아기 소리에 잡음이 포함되어 있을 가능성이 농후하다. 따라서 오디오 센서에서 받은 음성 데이터에서 아기 소리를 제외한 잡음의 처리가 필요하다.

본 논문은 잡음이 포함된 음성에서 아기 음성 추출을 위해 MCycle GAN(Multiple-Cycle Generative adversarial network, 다중 사이클 적대생성망)을 이용한 방법을 제안한다. MCycle GAN은 기존 Cycle GAN에서 Cycle을 다중으로 두어 학습하여, 판별자의 판별 능력을 강화하기 위해 고안했다. 학습 이후, Cycle 중 가장 성능이 좋은 Cycle의 생성자를 최종적으로 잡음 처리 모델로 사용한다. MCycle GAN 구조의 학습 모델을 구현 및 학습 진행하고, 학습이 완료된 MCycle GAN에서 잡음이 포함된 아이 음성(Noisy Baby Sound)을 입력받아, 가짜 깨끗한 아이 음성(Fake Clean Baby Sound)을 생성할 생성자(Generator)를 추출한다. 이때, 추출하는 생성자는 다수의 Cycle 중 가장 성능이 우수한 생성자이다. 추출한 생성자의 성능 평가를 위해 다른 학습 모델의 생성자와의 성능을 비교한다. 성능 지표로는 오차, 학습 시간, 판별자의 손실 함수를 사용한다.

2. 관련 연구

잡음 제거를 위해 아기 소리의 주파수 대역을 제외한 잡음의 주파수 대역을 제거하는 최소 오차 평균(Wiener) 필터링 방법이 있다. 하지만 최소 오차 평균 필터링 방법으로 실생활에서 발생하는 다양한 잡음을 모두 제거하기란 기대하기 어렵다. 따라서 최근 딥러닝 모델을 통해 잡음 제거를 시도하는 연구가 활발하다. 오토인코더(Auto Encoder)에 잡음을 학습시켜 잡음이 포함된 음성에서 잡음의 특징을 추출하여 잡음을 복원하고, 복원된 잡음으로 원본 음성 데이터에서 잡음을 제거한다 [1].

CNN(Convolution Neural Net)을 활용하여 인간의 음성 데이터에서 발화자의 음성 구간과 비음성 구간을 판별하여 불필요한 음성 신호의 제거가 가능하다 [2]. 잡음 제거 분야에서 딥러닝 모델들이 좋은 성능을 보이고 있으나, 학습되지 않은 잡음에 취약하다는 점을 주목하였고, 잡음 제거를 학습한 GAN 기반 생성모델을 사용하여 학습되지 않은 잡음 제거에서 향상된 성능을 보인다 [3][4]. SEGAN (Speech Enhancement Generative Adversarial Network)은 처음으로 GAN 기반 딥러닝 모델이 잡음 제거 분야에서도 효과적임을 보였다 [5].

영상 대 영상 변환 (Image-to-Image Translation) 분야에서 고전적인 방식은 훈련 데이터가 반드시 짝이 있어야 한다. 짝이 있는 훈련 데이터를 모으는 것은 매우 비용이 큰 작업이다. Cycle GAN은 일반적인 GAN에서 역방향 생성자를 추가하여, 변환한 이미지를 다시 원래 스타일로 복구하기 쉽게 학습을 진행한다. 이는 짝이 없는 훈련 데이터를 사용할 수 있다 [6]. 기존 Cycle GAN은 이미지 데이터 변환을 위해 설계된 모델이지만, 음성 데이터를 스펙트로그램으로 변환하여 음성을 이미지처럼 사용하여 Cycle GAN에서 학습할 수 있다 [7]. GAN 기반 음성 주파수 합성을 통해 음성에 포함된 감정을 추출할 수 있다 [8]. 본 논문에서 librosa 라이브러리를 사용하여, 스펙트로그램으로 변환하였다 [9].

일반적인 GAN에서는 주로 판별자의 손실 함수로 Cross Entropy loss를 사용한다. 이것을 사용할 때, 실제 데이터로부터 먼 가짜 데이터를 사용하여 가중치를 갱신할 경우, Vanishing Gradient 현상이 일어난다. 손실 함수로 Least Square loss를 사용하면, 이러한 현상을 방지하고, 더 안정적인 학습을 진행할 수 있다 [10]. 딥러닝 모델의 레이어를 너무 깊게 구성하면, 학습이 어렵고 성능이 오히려 떨어진다. Resnet(Residual neural network)은 Skip Connection을 이용하여 매우 깊은 층을 구성할 수 있다 [11][12]. 일반적인 GAN은 학습이 불안정하다. DCGAN은 Convolution, batch normalization, LeakyReLU를 사용하여, GAN의 학습을 안정적으로 만든 모델이다. 판별자의 활성화 함수로 ReLU 대신 LeakyReLU를 사용할 경우, 생성자에게 더 강한 기울기를 전달할 수 있다 [13].

아기의 감정(불편함, 배고픔, 졸음) 분류를 위해 32차원 FFT(Fast Fourier Transform, 고속 푸리에 변환)한 음성을 훈련 데이터로 사용하여, PCA(principal component analysis, 주성분 분석)가 가능하다 [14]. 생후 6개월 미만의 아기는 자신의 상태에 대한 의사소통 수단으로 표정, 음성, 심박수, 체온 등을 사용하는데, 대부분 울음소리를 사용한다. 따라서 음성을 통해 아기의 상태를 분석하는 것은 중요한 기술 중 하나이다 [15]. 딥러닝 모델로 아기 울음소리 감지 시, 음성을 시각적으로 처리하여 특수한 비대칭 커널을 사용한 CNN으로 아기 울음소리를 감지했을 때, 일반적인 CNN 아키텍처에 비해 좋은 결과를 얻을 수 있다 [16].

3. 본 론

3.1 MCycle GAN

MCycle GAN은 기존 Cycle GAN의 Cycle을 다중으로 두어 학습하여 판별자의 판별 능력을 강화하고, 학습시킨 Cycle 중 가장 성능이 좋은 Cycle에서 생성자를 최종적으로 선택하고자 설계하였다. 잡음이 포함된 아기 음성의 잡음 제거 시, 원본 아기 음성의 형태를 유지하기 위해서 원본 데이터의 형태를 보존하는 특성을 가진 Cycle GAN 기반 모델을 사용한다. 또한, Cycle GAN은 도메인끼리 서로 짝이 없는 훈련 데이터도 사용할 수 있기 때문에, 깨끗한 아이 음성, 잡음이 포함된 아이 음성 두 도메인의 데이터를 각각 수집하여 인공지능 학습을 진행할 수 있다.

본 논문에서 도메인 깨끗한 아이 음성을 X, 잡음이 포함된 아이 음성을 Y로 정의한다. X에 속한 데이터를 x, Y에 속한 데이터를 y라 할 때, x를 y로 변환하는 생성자(Generator)를 G_{xy}, 그 역을 G_{yx}라 한다. D_{domain}는 해당 $domain$에 대하여, 입력받은 데이터가 진짜 데이터인지, 생성자에 의해 생성된 가짜 데이터인지 진위를 판별하는 판별자(Discriminator)이다. MCycle GAN의 역전파를 위해 2가지 손실 함수(Loss Function)인 Adversarial Loss와 Cycle Consistency Loss를 통합한다. Adversarial Loss는 생성자가 판별자를 더 잘 속이도록, 판별자는 판별을 더 잘하도록 하는 GAN의 손실 함수 중 하나이다. 본 논문에서 Adversarial Loss로 CEE(Cross Entropy Error) 기반의 손실 함수를 사용할 경우, Gradient Vanishing 문제가 발생할 여지가 크기 때문에, SSE(Sum of Squares of Error) 기반의 손실 함수인 LSGAN을 사용한다 [10].

Adversarial Loss :

$$L_{gan}(G_{xy}, D_Y, X, Y) = E_{y \sim p_{data(Y)}}[(D_Y(y)-1)^2] + E_{x \sim pdata(X)}[D_Y(G_{xy}(x))^2]$$

Cycle Consistency Loss는 생성자의 목표로 , 가 되도록 하는 손실 함수이다. 이는 변환한 데이터가 원본 데이터로 복구되기 쉽게끔 학습되도록 한다 [5].

Cycle Consistency Loss :

$$L_{cyc}(G_{xy}, G_{yx}) = E_{x \sim p_{data(X)}}[\| G_{yx}(G_{xy}(x)) - x \|] + E_{y \sim p_{data(Y)}}[\| G_{xy}(G_{yx}(y) - y \|]$$

MCycle GAN은 위 두 가지 손실 함수를 통합한 MCycle Loss를 사용한다. MCycle GAN은 Cycle이 여러 개 존재하기 때문에 L_{gan}과 L_{cyc}이 다수 존재한다. 이때, 어떤 Cycle에 대한 생성자 G_{xy}와 G_{yx}를 각각 G_{xy_i}와 G_{yx_i}, 손실 함수 L_{gan}과 L_{cyc}을 각각 L_{gan_i}와 L_{cyc_i}로 정의한다. 이때, MCycle Loss의 수식은 다음과 같다.

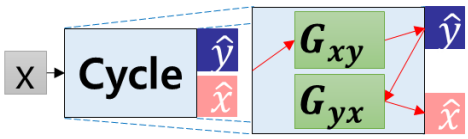
MCycle Loss :

$$L_{mcyc}(L_{gan_1}(G_{xy_1}, D_Y, X, Y), L_{gan_1}(G_{yx_1}, D_X, Y, X), L_{cyc_1}(G_{yx_1}, G_{xy_1}), \ldots, L_{gan_N}(G_{xy_N}, D_Y, X, Y), L_{gan_N}(G_{yx_N}, D_X, Y, X), L_{cyc_N}(G_{yx_N}, G_{xy_N})) = \sum_{i=1}^{N} L_{gan_i}(G_{xy_i}, D_Y, X, Y) + L_{gan_i}(G_{yx_i}, D_X, Y, X) + L_{cyc_i}(G_{yx_i}, G_{xy_i})$$

L_{mcyc}는 MCycle GAN의 모든 Cycle에 대한 L_{gan}과 L_{cyc}을 다중 Cycle에 맞게 조합한 손실 함수이다. 이는 다중의 Cycle을 동시에 학습하며, 판별자의 학습을 강화하는 형태를 가진다. 또한, 여러 개의 생성모델 학습에 하나의 판별자를 재활용하는 효과가 있다. 여러 개의 Cycle을 학습하였으므로, 생성자의 출력 결과를 확인하여 가장 품질이 좋은 결과를 출력하는 생성자를 선택하여 사용한다.

Algorithm1. MCycle GAN Loss Function

```
//Adversarial Loss
L_gan(gen_x2y, disc_y, real_x, real_y){
   fake_y = gen_x2y(real_x);
   loss_ad = (disc_y(real_y) - 1) *
   (disc_y(real_y) - 1) + disc_y(fake_y) *
   disc_y(fake_y);
   return loss_ad;
}
//Reverse Loss
L_rgan(gen_y2x, disc_x, real_y, real_x){
   fake_x = gen_y2x(real_y);
   loss_re = (disc_x(real_x) - 1) *
   (disc_x(real_x) - 1) + disc_x(fake_x) *
   disc_x(fake_x);
   return loss_re;
}
//Cycle Consistency Loss
L_cyc(gen_x2y, gen_y2x, real_x, real_y){
   fake_y = gen_x2y(real_x);
   fake_x = gen_y2x(real_y);
   loss_cy = abs(gen_y2x(fake_y) - real_x) +
   abs(gen_x2y(fake_x) - real_y);
   return loss_cy;
//MCycle Loss
L_mcyc(list_gen_x2y[], list_gen_y2x[], real_x, real_y){
   for(int i = 0; i < n; i++){
      L_gan(list_gen_x2y[i], disc_y, real_x,
      real_y) + L_rgan(list_gen_y2x[i], disc_x,
      real_y, real_x) + L_cyc(list_gen_x2y[i],
      list_gen_y2x[i], real_x, real_y);
   }
}
```



(그림 1) Cycle 구조
(Figure 1) The structure of cycle

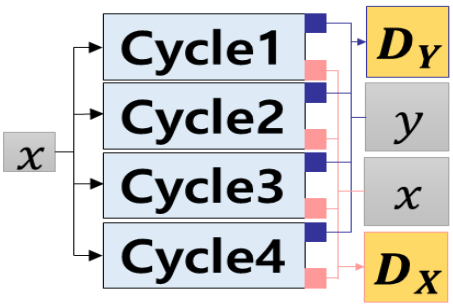
3.2 MCycle GAN 구조

Cycle은 깨끗한 아이 음성 도메인인 X의 데이터 x를 입력받아, 2가지의 결과물 데이터를 출력한다.

output 1 : $G_{xy}(x) = \hat{y}$

output 2 : $G_{yx}(\hat{y}) = \hat{x}$

$\hat{x}$와 $\hat{y}$는 각각 생성자에 의해 생성된 Fake Domain X data, Fake Domain Y data를 의미한다. output 1은 Domain X인 데이터 중에서 임의로 선택된 데이터인 x에서, 생성자 G_{xy}가 생성한 Fake Domain Y data인 $\hat{y}$이다. output 2는 output 1 데이터인 $\hat{y}$에서, 생성자 G_{yx}가 생성한 Fake Domain X data인 $\hat{x}$이다.



(그림 2) MCycle GAN (4 Cycle) 구조
(Figure 2) The structure of MCycle GAN (4 cycle)

MCycle GAN은 그림과 같이 Cycle을 여러 개 사용하는 구조를 가진다. 임의의 Domain X data인 x를 여러 개의 Cycle이 병렬적으로 입력받는다. Cycle들의 각각의 output 중에서 Fake Domain Y data인 $\hat{y}$는 판별자인 D_Y의 input으로 사용된다. 또한, output 중에서 Fake Domain X data인 $\hat{x}$는 판별자인 D_X의 input으로 사용된다. D_Y는 앞서 설명한 Fake data인 $\hat{y}$와 Real data인 y 중 하나를 입력받아 그 데이터가 Real data일 확률을 출력한다. 마찬가지로 D_X도 $\hat{x}$와 x를 입력받아, 그 입력 데이터가 Real data일 확률을 출력한다.

학습 시, MCycle Loss에 따라 각 Cycle의 생성자들과 판별자 D_X, D_Y의 가중치가 갱신된다. D_X, D_Y의 판별 정확도가 50%가 되었을 때, 학습을 완료한다. 최종적으로 Cycle들 중에서 가장 판별자를 잘 속이는 Cycle이 가장 성능이 좋으므로, 그것을 선택하여 사용한다. 본 논문의 경우, Cycle의 구성요소 중 하나인 G_{xy}는 깨끗한 아기 소리 데이터를 입력받아 잡음이 포함된 아기 소리를 생성하는 생성자이고, G_{yx}는 G_{xy}의 역이다. 따라서 잡음이 포함된 아기 소리 데이터에서 잡음을 제거된듯한 아기 소리 데이터를 생성하기 위해, G_{yx}를 최종적인 잡음 제거 모델로 사용한다.

MCycle GAN의 생성자는 ResNet 구조의 CNN 모델을 사용하여, 인풋 데이터의 디테일을 유지하며, 데이터의 병목 현상을 방지한다 [11]. 생성자는 Adversarial Loss가 작아지는 방향으로 학습한다. Adversarial Loss가 작을수록 판별자를 잘 속임을 의미한다.

판별자는 활성화 함수로 Leaky ReLU를 사용한 CNN 기반의 이미지 분류 모델을 사용한다. Leaky ReLU로 학습한 판별자는 생성자에게 ReLU에 비해 더 강한 기울기 전달이 가능하다 [13].

3.3 실험

3.3.1 실험 설계

(표 1) 성능 평가를 위한 학습 모델 설정
(Table 1) Learning model Setup for performance evaluation

	Cycle (개)	Loss Function	batch size	epoch
Cycle GAN	1	L_{gan}, L_{cyc}	1	40
MCycle GAN(2)	2	L_{mcyc}	1	40
MCycle GAN(4)	4	L_{mcyc}	1	40

MCycle GAN의 성능 평가를 위해 3개의 모델을 준비한다. Model 1은 일반적인 Cycle GAN 모델이다. Model 2, Model 3는 각각 Cycle 2개, 4개를 다중으로 사용하여 학습한 MCycle GAN 모델이다. 사이클 개수에 따른 성능 증감량과 학습 시간 변화량을 파악하기 위해 3가지 모델을 준비하였다. Model 1은 $D_X(G_{yx}(y))$와 역방향인 $D_Y(G_{xy}(G_{yx}(y)))$ 두 과정에서, L_{gan}과 L_{cyc}로 G_{yx},

G_{xy}, D_X, D_Y의 학습을 진행한다. Model 2, 3는 본 논문에서 소개하는 MCycle GAN의 학습 방법을 따른다. 3개의 모델 모두 훈련 단계에서 batch size 1, 40 epoch만큼 학습한다.

학습을 위한 데이터 세트를 위하여, donateacry-corpus에서 깨끗한 아이 음성(.wav) 500개 [17], ESC-50 : Dataset for Environmental Sound Classification에서 환경 소음(.wav) 2,000개를 참조하였다 [18]. donateacry-corpus는 Donate-a-cry 캠페인을 통해 구축된 아기 울음소리이다. 아기의 음성이 5종류의 감정(복통, 트림, 불편함, 배고픔, 피곤함)으로 분류되어있다. 깨끗한 아이 음성와 환경 소음 데이터를 각각 무작위로 추출 및 병합하여, 잡음이 포함된 아이 음성(.wav) 2,000개를 생성하였다. 깨끗한 아기 음성과 잡음의 신호 대비 잡음 비(Signal-to-Noise ratio)가 -5db이 되도록 음성 병합을 진행하였다 [19]. 생성자와 판별자는 이미지 처리 딥러닝 모델인 CNN 기반의 모델이므로, 음성 데이터를 스펙트로그램(.png, 196 * 128)으로 변환하여 이미지처럼 사용하였다 [9].

트레이닝 데이터 세트 : domain Y data는 1,500개의 잡음이 포함된 아이 음성 스펙트로그램, domain X data는 450개의 깨끗한 아이 음성 스펙트로그램을 사용한다.

테스트 데이터 세트 : domain Y data는 500개의 잡음이 포함된 아이 음성 스펙트로그램, domain X data는 50개의 깨끗한 아이 음성 스펙트로그램을 사용한다.

MCycle GAN은 훈련 시, 훈련 데이터 세트인 domain Y data, domain X data에서 각각 400개씩 무작위 추출하였다.

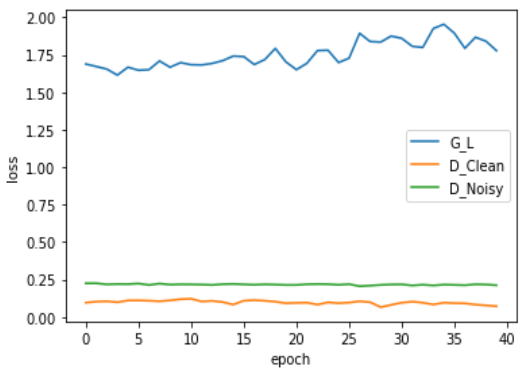
사용한 언어는 python3이고, 주요 라이브러리로 tensorflow 2.9.2, keras를 사용하였다. colab pro plus(OS : Linux-5.10.133+-x86_64-with-Ubuntu-18.04-bionic, GPU : Tesla P100-PCIE, CPU : Intel(R) Xeon(R) CPU @ 2.20GHz RAM : 27.4GB) 환경에서 인공지능 모델을 구현, 학습 및 예측을 진행하였다. 성능 평가는 원본 데이터와의 오차, 학습 시간, 그리고 판별자의 손실 함수, 3가지 지표로 평가한다. 원본 데이터와의 오차는 모델에서 추출한 생성자가 원본 데이터에 잡음을 혼합한 데이터 10개를 입력받아, 예측한 데이터가 원본 데이터와의 오차 제곱의 합이다. 낮을수록 원본 데이터와 유사한 데이터를 출력하는 모델이다. 학습 시간은 학습 모델이 40 epoch 학습 시, 걸리는 시간으로 낮을수록 학습이 빠른 모델이다. 판별자 손실 함수는 판별자가 생성자의 위조를 검출할수록 낮은 값을 보인다. 이 값이 클수록 생성자가 판별자를 잘 속일 가능성이 크거나, 판별자가 강화된 모델임을 의미한다. 각 모델에서 추출한 생성자가 50개의 테스트 데이터를 예측하고, 판별자의 손실 함수의 값이다.

(표 2) 각 학습 모델의 성능 지표
(Table 2) Performance metrics for each model

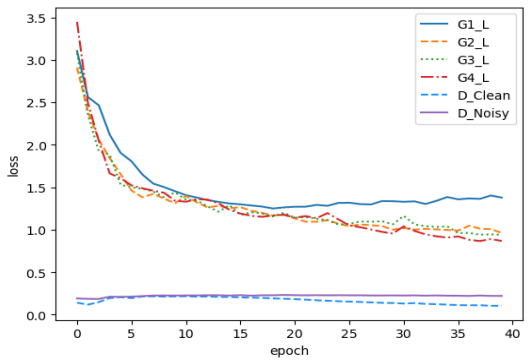
	Cycle GAN	MCycle GAN(2)	MCycle GAN(4)
오차	17,150	14,781	15,312
학습 시간(s)	2,037	3,838	7,590
판별자의 손실 함수	11.2139	0.2813	0.2409

3.3.2 실험 결과

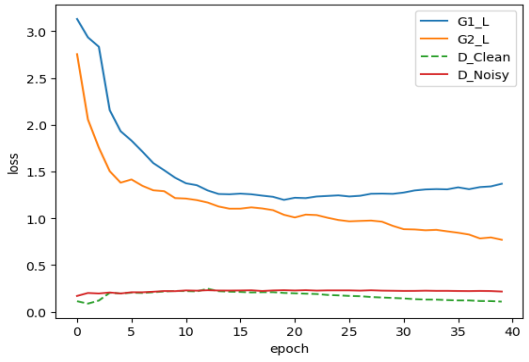
표 1은 각 모델의 잡음 제거 성능 평가로, MCycle GAN(2)가 14,781로 가장 낮은 오차를 가진다. 이는 MCycle GAN(2)이 다른 모델에 비해 원본 데이터와 유사한 형태를 가지면서 잡음을 처리했음을 알 수 있다. 학습 시간의 경우, Cycle GAN이 2,037(s)로 가장 짧았고, MCycle GAN(4)이 7,590(s)로 가장 길었다. MCycle GAN은 Cycle을 다중으로 사용하기 때문에 다른 모델들에 비해 학습 시간이 상대적으로 길다. 판별자의 손실 함수의 경우, Cycle GAN이 11.2139로 다른 모델들에 비해 상대적으로 높은 수치를 가지며, MCycle GAN은 Cycle 개수에 따라 2개는 0.2813, 4개는 0.2409로 근소한 차이를 가졌다. Cycle GAN은 생성자가 판별자를 충분히 속이지 못하였다. MCycle GAN의 Cycle 2개와 Cycle 4개를 비교했을 때, Cycle 4개는 Cycle 2개에 비해 학습 속도가 약 2배 더 소요되며, 오차가 오히려 늘어났다. Cycle 개수를 늘리다 보면 성능이 떨어지는 지점이 있을 것으로 예상한다.



(그림 3) Cycle GAN 학습 로그
(Figure 3) Training log of Cycle GAN



(그림 5) MCycle GAN (4 Cycle) 학습 로그
(Figure 5) The Training log of MCycle GAN (4 Cycle)



(그림 4) MCycle GAN (2 Cycle) 학습 로그
(Figure 4) The Training log of MCycle GAN (2 Cycle)

위 그림 3, 4, 5는 모델별 40 epoch 동안의 학습 중 생성자의 손실 함수, 판별자의 손실 함수이다. 같은 epoch 동안 MCycle GAN 모델은 생성자의 손실 함수가 감소하는 방향으로 가는 것을 볼 수 있다. epoch 0에서는 Cycle GAN에 비해 저조한 성능을 보이지만, epoch 5부터 Cycle GAN보다 높은 성능을 보인다. 하지만 Cycle GAN은 오히려 생성자의 손실 함수가 오히려 증가하는 방향으로 학습되는 것이 관측되었다. 따라서 Cycle GAN의 경우, 더 많은 epoch가 필요하다고 예측된다. 판별자의 손실 함수의 경우, 40번째 epoch에서 D_Clean은 각각 MCycle GAN(4) = 0.1051, MCycle GAN(2) = 0.1100, Cycle GAN = 0.0714로 MCycle GAN의 판별자가 더 많이 틀린다. Cycle GAN은 D_Clean은 작지만, G_L이 크기 때문에, 생성자가 위조 데이터를 능숙하게 만들지 못하는 상태로 볼 수 있다.

4. 결 론

본 논문은 잡음이 포함된 아기 음성 데이터에서 아기 음성만 추출하기 위한 인공지능 모델 MCycle GAN을 설계한다. 음성을 이미지처럼 사용하기 위해 스펙트로그램화하고, 이미지 학습을 진행한다. MCycle GAN 모델의 구조를 설계하고 구현하여, 여러 Cycle을 동시에 학습한다. 최종적으로 Cycle의 잡음이 포함된 아이 음성를 입력받아, 가짜 깨끗한 아이 음성을 생성하는 생성자 중 가장 성능이 좋은 생성자를 추출한다. 추출한 생성자는 결과적으로 잡음이 포함된 아기 음성 데이터를 깨끗한 아기 음성 데이터로 변환하는 효과를 가진다.

실험은 각각 Cycle을 1, 2, 4개 가진 모델의 성능을 평가하여 비교하는 방식으로 구성하였다. 성능 평가 지표로 원본 데이터와의 오차와 학습에 걸리는 시간, 그리고 판별자의 성능을 파악하기 위한 판별자의 손실 함수를 사용하였다.

실험 결과, 잡음 처리 시, MCycle GAN이 GAN, Cycle GAN에 비해 낮은 오차와 낮은 판별자의 손실 함수를 가진다. 이는 MCycle GAN이 원본 아기 소리의 형태를 유지하며 잡음을 처리하며, 잡음 처리한 데이터가 판별자를 잘 속일 수 있을 만큼, 깨끗한 아이 음성 도메인의 데이터와 크게 다르지 않음을 의미한다. 하지만 학습 시간이 오래 걸리는 단점이 있다. 이는 학습 시, 다중 Cycle을 사용함으로써 생기는 문제라고 생각된다.

MCycle GAN은 아기 음성을 통한 아기 상태 분석 분야의 발전에 기여하고, Cycle GAN 기반 모델의 Cycle을 다중으로 설계하였을 때, Cycle의 개수에 따른 성능의 변화를 파악한다. MCycle GAN은 다른 훈련 데이터 세트로 학습을 진행함으로써, 아기 음성뿐만 아니라, 다른 특수한 타깃 음성에 대한 잡음 처리 또한 수행할 수 있다. 추후, MCycle GAN에서 다수의 Cycle들이 병렬적으로 분산처리하여 학습 시간을 단축하는 연구가 필요하다. 또한, MCycle GAN 모델을 더 많은 epoch와 데이터를 사용하여 학습할 계획이다.

참고문헌(Reference)

[1] Park, Ji Hong, "A Noise Filtering Scheme with Machine Learning for Audio Content Recognition", 2019.02.
https://repository.hanyang.ac.kr/handle/20.500.11754/100017

[2] Hoo-Young Lee, "A Study on a Non-Voice Section Detection Model among Speech Signals using CNN Algorithm", Journal of Convergence for information Technology vol. 11. pp. 33-39, 2021. https://doi.org/10.22156/CS4SMB.2021.11.06.033

[3] Kyung-Hyun Lim, "GAN with Dual Discriminator for Non-stationary Noise Cancellation", 2019. http://sclab.yonsei.ac.kr/publications/Papers/KC/2019_KSC_KHL.pdf

[4] Wonsup Shin, Jin-Young Kim, Sung-Bae Cho, "GAN-based noise elimination model for high-quality speech database", 한국소프트웨어종합학술대회 논문집, pp. 557-559, 2019. http://sclab.yonsei.ac.kr/publications/Papers/KC/2019_KSC_WSS.pdf

[5] Santiago Pascual, Joan Serra, Antonio Bonafonte, "Time-domain Speech Enhancement Using Generative Adversarial Networks", Volume 114, Pages 10-21, 2019. https://doi.org/10.1016/j.specom.2019.09.001

[6] Zhu, Jun-Yan, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." Proceedings of the IEEE international conference on computer vision. 2017. https://doi.org/10.48550/arXiv.1703.10593

[7] Lai, Wen-Hsing, Wang, Siou-Lin, Xu, Zhi-Yao. "CycleGAN-Based Singing/Humming to Instrument Conversion Technique," Electronics, 11(11), 1724, 2022. https://doi.org/10.3390/electronics11111724

[8] H. Kwon, M. Kim, J. Baek and K. Chung, "Voice Frequency Synthesis using VAW-GAN based Amplitude Scaling for Emotion Transformation," KSII Transactions on Internet and Information Systems, vol. 16, no. 2, pp. 713-725, 2022. https://doi.org/10.3837/tiis.2022.02.018

[9] McFee, Brian, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. "librosa: Audio and music signal analysis in python." In Proceedings of the 14th python in science conference, pp. 18-25. 2015. https://www.researchgate.net/publication/328777063_librosa_Audio_and_Music_Signal_Analysis_in_Python

[10] Mao, Xudong, et al. "Least squares generative adversarial networks." Proceedings of the IEEE international conference on computer vision. 2017. https://doi.org/10.48550/arXiv.1611.04076

[11] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. https://doi.org/10.48550/arXiv.1512.03385

[12] J. Zhu, L. Sun, Y. Wang, S. Subramani, D. Peng and S. C. Nicolas, "A ResNet based multiscale feature extraction for classifying multi-variate medical time series," KSII Transactions on Internet and Information Systems, vol. 16, no. 5, pp. 1431-1445, 2022. https://doi.org/10.3837/tiis.2022.05.002

[13] Radford, Alec, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks." arXiv preprint arXiv:1511.06434. 2015. https://doi.org/10.48550/arXiv.1511.06434

[14] Yamamoto S, Yoshitomi Y, Tabuse M, Kushida K, Asada T. Recognition of a Baby's Emotional Cry towards Robotics Baby Caregiver. International Journal of Advanced Robotic Systems. 2013. https://doi.org/10.5772/55406

[15] Il-Kyu Hwang, Ho-Bum Song. "AI-based Infant State Recognition Using Crying Sound". The Journal of Korean Institute of Information Technology, 17(7), 13-21. 2019. http://dx.doi.org/10.14801/jkiit.2019.17.7.13

[16] Cohen, Rami, et al. "Baby Cry Detection: Deep Learning and Classical Approaches." PsyArXiv, 17 Dec. 2019. https://doi.org/10.1007/978-3-030-31764-5

[17] donateacry-corpus, Clean Baby Data Set, https://github.com/gveres/donateacry-corpus

[18] ESC-50: Dataset for Environmental Sound Classification, Noisy Sound Data Set, https://github.com/karolpiczak/ESC-50

[19] https://github.com/Sato-Kunihiko/audio-SNR

◐ 저 자 소 개 ◑

손 수 락(Su-rak Son)
2018년 가톨릭관동대학교 컴퓨터공학과(공학사)
2020년 가톨릭관동대학교 대학원 컴퓨터공학과(공학석사)
2022년 가톨릭관동대학교 대학원 컴퓨터공학과(공학박사)
2022년~현재 가톨릭관동대학교 강사
관심분야 : 빅데이터, 인공지능, 프로그래밍 언어, etc.
E-mail : sonsur@naver.com

심 규 정(Kyu-jeong Sim)
2020년 가톨릭관동대학교 소프트웨어학과(학사)
2021년~현재 가톨릭관동대학교 대학원 소프트웨어학과 수료중
관심분야 : 인공지능, 적대적생성망, 잡음처리.
E-mail : rbwjd1998@gmail.com

정 이 나(Yi-na Jeong)
2011년 가톨릭관동대학교 컴퓨터공학과(공학사)
2012년 가톨릭관동대학교 대학원 컴퓨터공학과(공학석사)
2018년 가톨릭관동대학교 대학원 컴퓨터공학과(공학박사)
2018년~현재 가톨릭관동대학교 소프트웨어학과 교수
관심분야 : Sensor Network, IT security, Network security
E-mail : lupinus07@nate.com