Journal of the Korean Data & Information Science Society
2018, **29**(5), 1155–1167

http://dx.doi.org/10.7465/jkdi.2018.29.5.1155
한국데이터정보과학회지

대립생성망의 성능 비교에 관한 연구[†]

이영재[1] · 석경하[2]

[12]인제대학교 통계학과

접수 2018년 7월 5일, 수정 2018년 8월 10일, 게재확정 2018년 8월 20일

요 약

대립생성망 (generative adversarial networks, GAN)은 실제 자료와 유사한 자료를 만들어주는 생성형 딥러닝 (generative deep learning) 모형이다. 2014년에 발표된 이래로 많은 파생 모형들이 개발되어 다양한 분야에 활용되고 있다. 본 연구에서는 성능이 우수하다고 평가된 파생 GAN들을 요약 및 정리하고 성능을 비교하였다. 그리고 GAN의 입력 잠재공간 (input latent space)의 적절한 차원크기를 추정하고 생성자료의 품질을 평가하는 프레쳇 인셉션 거리 (Fréchet Inception distance, FID)와 인셉션 점수 (Inception score)의 적절성도 평가하였다. 실험결과 GAN-NS와 LSGAN이 안정적으로 우수한 성능을 보였으며 FID가 더 좋은 측도로 평가되었다. 그리고 잠재공간은 10차원에서도 전형적인 100차원과 차이가 없는 좋은 결과를 보였다.

주요용어: 대립생성망, 딥러닝, 인셉션 점수, 잠재공간, 프레쳇 인셉션 거리.

1. 서론

기계학습 (machine learning)은 지도학습 (supervised learning), 자율학습 (unsupervised learning) 그리고 강화학습 (reinforcement learning)으로 나누어진다. 자율학습은 지도학습과는 달리 목표값이 주어지지 않는 자료를 학습하여 차원축소, 군집화 그리고 특징추출 등의 작업을 주로 하는데 오토인코더 (autoencoder), GAN (generative adversarial networks), RBM (restricted Boltzmann machine)등이 있다.

딥러닝 분야에서는 지도학습이 많은 부분을 차지했지만 목표값 (target)을 구해야하는 이유로 활용에 한계가 있었다. 반면 자율학습은 Goodfellow(2014)의 GAN을 중심으로 넓은 활용도를 보이며 현재까지 이미지와 음성분야를 비롯한 다양한 분야에서 접목되고 있다.

GAN은 생성망 (generator)과 판별망 (discriminator)을 대립하며 학습하는 신경망이다. 생성망은 실제 자료와 같은 자료를 만들기 위해 노력하고, 판별망은 실제 자료와 생성자료를 구분하기 위해 노력하여 최종적으로 생성망이 실제 자료와 같은 자료를 만드는 것이 GAN의 목표다. 초기 GAN은 학습의 불안정과 모드붕괴 (mode-collapse) 등의 문제점을 수반하였다. 이를 개선하는 많은 파생 GAN 모형들이 개발되었으며, 인셉션 모형 (Inception model)을 이용한 인셉션 점수 (Inception score, IS)와 프레쳇 인셉션 거리 (Fréchet Inception distance, FID)등이 개발되어 GAN의 성능을 평가할 수 있게 되었다. 그렇지만 이들 측도에 대한 평가는 아직 이루어지지 않고 있다.

[†] 본 논문은 2017학년도 인제대학교 학술연구조성비 지원으로 수행된 연구결과임.
[1] (50834) 경남 김해시 인제로 197, 인제대학교 통계학과, 석사.
[2] 교신저자: (50834) 경남 김해시 인제로 197, 인제대학교 통계학과, 인제대학교 통계정보연구소, 교수.
E-Mail: statskh@inje.ac.kr

생성망은 균등분포 혹은 정규분포에서 무작위로 추출된 랜덤 벡터 ()를 실제 자료의 분포와 유사하게 만들어주는데 랜덤 벡터가 존재하는 공간을 잠재 공간 (latent space)이라 한다. 일반적으로 잠재공간의 차원크기가 충분히 커야 실제 자료의 특징을 잘 표현한다는 믿음으로 100차원의 잠재공간을 사용한다. 그러나 잠재공간의 크기는 매개변수의 수와 수렴속도 그리고 계산시간으로 이어지는 중요한 문제이기 때문에 잠재공간의 적절한 차원크기를 추정하는 것은 GAN의 최적화에 필요한 과제이다.

이에 본 연구에서는 성능이 우수하다고 알려진 파생 GAN 모형들을 요약 및 정리하고 성능비교를 한다. 그리고 잠재공간의 적절한 차원크기를 실험으로 추정하고 FID와 IS의 적절성을 평가하였다. 실험결과 대다수의 GAN들은 10차원까지는 성능이 향상되는 모습을 보였고 10, 50, 100,차원에서는 성능을 유지하여 10차원에서도 전형적인 100차원과 차이가 없는 것을 확인 할 수 있었다. 모형의 성능을 비교한 결과 GAN-NS와 LSGAN은 회색조이미지와 컬러이미지 모두에서 안정적으로 우수한 성능을 보였다. 그리고 IS는 일부 자료에서는 1차원에서도 성능이 좋은 것으로 잘 못 평가하는 등의 문제점을 안고 있어 FID 보다 부족한 모습을 보였다.

2절에서는 GAN에 대한 구조와 학습과정 그리고 문제점에 대해 설명하고 파생 GAN들의 특징을 설명한다. 3절에서는 연구에 사용한 자료들을 소개하고 평가 척도와 분석방법 그리고 결과를 기술하였고, 연구결과에 대한 결론과 제안은 4절에서 다루었다.

2. GAN

2.1. GAN

GAN은 실제 자료 (x)와 동일한 분포를 가지는 자료를 생성하는 생성망 (G)과 실제 자료와 생성자료 ($G(z) = \hat{x}$)를 구별하는 판별망 (D)이 서로 대립하며 최적화를 수행해 나가는 모형이다. 이는 판별망과 생성망이 최소최대 게임을 하는 것으로 표현 가능하며 다음과 같은 가치함수 (value function)로 정의 할 수 있다 (Goodfellow 등, 2014).

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_d}[\log D(x)] + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))], \tag{2.1}$$

여기에서 p_d와 p_z는 x와 z의 분포다.

Figure 2.1은 GAN의 구조를 설명한 그림으로 D는 실제 자료 x에 대해서는 $D(x) = 1$, 생성자료 $G(z)$에 대해서는 $D(G(z)) = 0$가 되도록 노력한다. 즉, D는 생성자료와 실제 자료를 잘 구분하려고 노력한다. 그리고 G는 실제 자료와 아주 유사한 자료를 생성하여 $D(G(z)) = 1$이 되도록 노력한다.

주어진 생성망에서 판별망이 최적의 값을 가질 때 (2.1)의 GAN의 가치함수를 아래와 같이 재표현할 수 있다.

$$\begin{aligned} C(G) &= \max_D V(G,D) \\ &= \mathbb{E}_{x\sim p_d}[\log D_G^*(x)] + \mathbb{E}_{x\sim p_g}[\log(1 - D_G^*(x))] \\ &= -\log(4) + KL(p_d||\frac{p_d + p_g}{2}) + KL(p_g||\frac{p_d + p_g}{2}) \\ &= -\log(4) + 2JSD(p_d||p_g), \end{aligned}$$

여기에서 p_g는 $\hat{x} = G(z)$의 분포이고 KL (Kullback-Leibler divergence)과 JSD (Jenson-Shannon divergence)는 두 분포 간의 거리를 나타내는 측도인데 JSD는 0이상의 값을 가지며 두 분포가 같으면 0이다. $C(G)$는 전역 최소값으로 $-\log(4)$를 가지는 것을 확인할 수 있다.

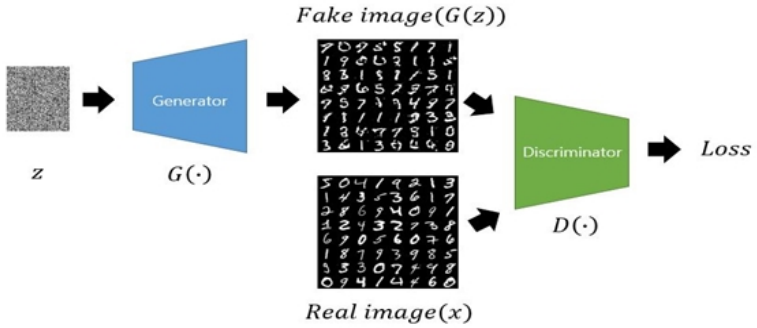

Figure 2.1 Structure of GAN

2.2. GAN의 문제점

GAN은 이론적으로는 수렴이 가능하지만 실제에서는 구조적인 불안정으로 인하여 다음과 같은 문제들이 발생할 수 있다. 그 첫 번째로는 판별망과 생성망은 최적 상태인 내쉬 균형점 (Nash equilibrium)에 도달하기 위해 각각의 가치함수를 구해 학습이 진행되지만 실제로는 수렴이 보장되지 않는다는 것이고 두 번째는 생성망이 다양한 실제 자료를 폭 넓게 생성하지 못하고 일부 자료만을 생성하게되는 모드붕괴 (mode collapse) 현상이다. 이는 생성자료가 판별망을 속일 수는 있겠지만 실제 자료의 분포와 다른 분포를 나타낸다. 세 번째는 처음부터 판별망의 성능이 매우 우수하다면 실제 자료와 생성자료를 완벽하게 구별하여 판별망이 가지는 손실이 0이 되기 때문에 생성망의 학습이 더 이상 이루어지지 않게 되고, 반면 판별망의 성능이 약하면 생성망에 정확한 피드백을 줄 수 없어 생성망이 제대로 학습할 수 없게 된다는 문제점을 안고 있다. 이러한 문제들을 해결하기 위해 다양한 파생 GAN들이 발표되었는데 그중에서 대표적인 것들을 살펴본다.

2.3. 파생 GAN

2.3.1. DCGAN

GAN의 불안정성과 모드붕괴 등의 문제점 극복하기 위해 이루어진 다양한 접근 중 가장 주목할만한 결과를 보인 것이 Radford 등 (2015)이 개발한 심층 합성곱 대립생성망 (deep convolutional GAN, DCGAN)이다. DCGAN은 이미지 지도학습에서 좋은 성능을 보이는 합성곱 신경망 (convolutional neural networks, CNN)을 이용하였는데 이 후 나온 다양한 파생 GAN들은 모두 DCGAN의 구조를 기반으로 하고 있다.

2.3.2. GAN-NS

학습 초기의 생성자료는 실제 자료와 많은 차이가 있기 때문에 판별망은 우수한 결과를 보인다. 그러므로 $\log(1 - D(G(z)))$가 0에 가까워 생성망이 학습되기에 충분히 큰 기울기를 제공하지 못하기 때문에 $\log(1-D(G(z)))$를 최소화하는 대신 $log(D(G(z)))$를 최대화 하도록 훈련하는 GAN-non saturating (GAN-NS)를 Goodfellow 등 (2014)이 제안하였다. 그 결과 학습 초기에도 생성망은 충분한 기울기를 제공받아 빠른 학습을 진행할 수 있다. 가치 함수는 아래와 같이 정의 된다.

$$L_D^{GAN-NS} = \mathbb{E}_{x\sim p_d}[\log(D(x))] + \mathbb{E}_{\hat{x}\sim p_g}[1-\log(D(\hat{x}))],$$
$$L_G^{GAN-NS} = \mathbb{E}_{\hat{x}\sim p_g}[\log(D(\hat{x}))],$$

여기에서 $\hat{x} = G(z)$, $\hat{x} \sim p_g$이다.

2.3.3. LSGAN

GAN에서 판별망의 손실함수로 사용된 시그모이드 교차 엔트로피 함수 (sigmoid cross entropy function)는 생성자료가 실제 자료로 분류되면 생성자료가 실제 자료와 아무리 멀리 있어도 이 자료들은 생성망에서 더 이상 학습이 진행되지 않는 문제를 야기한다. 이를 해결하기 위해 최소 제곱 손실함수 (least square loss function)를 사용하여 실제 자료로부터 멀리 있는 생성자료에 불이익을 주어 실제 자료에 더 다가 갈수 있도록 하여 학습의 안정성을 향상 시키는데 도움을 주도록 제안하는 least squares GAN (LSGAN)을 Mao 등 (2017)이 개발하였는데 그 가치함수는 아래와 같다.

$$L_D^{LSGAN} = \mathbb{E}_{x\sim p_d}[(D(x)-1)^2] + \mathbb{E}_{\hat{x}\sim p_g}[(D(\hat{x})^2],$$
$$L_G^{LSGAN} = \mathbb{E}_{\hat{x}\sim p_g}[(D(\hat{x})-1)^2].$$

2.3.4. WGAN

GAN의 생성망은 JSD를 최소화하는 학습을 한다. 그런데 JSD는 두 개의 분포가 완전하게 일치하지 않으면 항상 $\log 2$를 출력하므로 기울기가 0이 되고 이로 인해 미분이 불가능하게 되므로 학습이 진행되지 않는 문제를 가지고 있다. JSD 대신 두 개의 분포를 일치시키기 위해 필요한 질량의 최소 운반비용을 의미하는 Wasserstein 거리를 이용하는 Wasserstein GAN (WGAN)을 Arjovsky 등 (2017)이 제안하여 좋은 성과를 보였다. 가치함수는 아래와 같이 정의된다.

$$L_D^{WGAN} = \mathbb{E}_{x\sim p_d}[D(x)] - \mathbb{E}_{\hat{x}\sim p_g}[D(\hat{x})],$$
$$L_G^{WGAN} = \mathbb{E}_{\hat{x}\sim p_g}[D(\hat{x})].$$

2.3.5. WGAN-GP

WGAN에서는 가중치 제약의 임계값이 크면 수렴하는데 많은 시간이 소요되며 반대로 임계값이 적으면 기울기 소실 문제가 발생하여 최적화에 어려움이 뒤따랐다. Gulrajani 등 (2017)은 입력에 대한 판별망 출력의 기울기를 직접적으로 제한하는 제약 조건을 적용하는 WGAN-GP (Wasserstein GAN - gradient penalty)으로 이를 해결하였다. 생성망의 가치함수는 WGAN과 동일하며 판별망은 WGAN의 가치함수에 기울기 벌칙 (penalty)이 더해진 형태로 아래와 같다.

$$L_D^{WGAN-GP} = L_D^{WGAN} + \lambda \mathbb{E}_{\hat{x}\sim p_g}[(||D(\alpha x + (1-\alpha\hat{x}))||_2 - 1)^2],$$
$$L_G^{WGAN-GP} = L_G^{WGAN}$$

여기에서 λ는 벌칙의 강도를 조절하는 모수이고 α는 자료를 조절하는 모수다.

2.3.6. DRAGAN

학습 과정에서 실제 자료 주변에서 판별망 함수의 기울기가 아주 크면 생성망을 단일 출력으로 이끌어 모드붕괴가 발생한다. Kodali 등 (2017)은 판별망을 정규화하여 데이터 공간에서 기울기를 제한하는 deep regret analytic GAN (DRAGAN)을 제안하여 학습을 빠르게 하고 안정성을 향상하였다. 가치함수는 GAN의 판별망에 기울기 벌칙을 포함하는 아래 식으로 표현한다.

$$L_D^{DRAGAN} = L_D^{GAN} + \lambda \mathbb{E}_{\hat{x}\sim p_d + N(0,c)}[(|| \bigtriangledown D(\hat{x})||_2 - 1)^2],$$
$$L_G^{DRAGAN} = L_G^{GAN},$$

여기에서 λ는 벌칙의 강도를 조절하는 모수이다.

2.3.7. BEGAN

Berthelot 등 (2017)이 제안한 boundary equilibrium GAN (BEGAN)은 판별망 구조를 오토인코더로 제안한 energy-based GAN (Zhao, 2016)과 WGAN을 응용한 모형으로 실제 자료와 생성자료의 차이를 오토인코더 손실 분포의 Wasserstein 거리를 이용하여 비교하는데 가치함수는 다음과 같다.

$$L_D^{BEGAN} = \mathbb{E}_{x\sim p_d}[||x - AE(x)||_1] - k_t \mathbb{E}_{\hat{x}\sim p_g}[||\hat{x} - AE(\hat{x})||_1],$$
$$L_G^{BEGAN} = \mathbb{E}_{\hat{x}\sim p_g}[||\hat{x} - AE(\hat{x})||_1],$$

여기에서 AE는 오토인코더이고 $k_t \in [0, 1]$는 생성자료에 중점을 주는 정도를 제어하기 위한 모수로 초기값은 0을 가진다.

3. 분석 및 결과

3.1. 자료

분석에 사용된 자료는 딥러닝 평가에 전형적으로 사용되는 다음의 4개 자료다. Mnist 자료는 0부터 9까지의 숫자로 구성된 손글씨 이미지자료로 훈련이미지 60,000개, 시험이미지 10,000개으로 구성되어 있으며 각 이미지는 28×28 크기의 회색조로 이루어져 있다. Fashion-Mnist는 10개의 패션 클래스로 구성된 자료인데 훈련이미지 60,000개와 시험이미지 10,000개로 구성되어 있으며 각 이미지는 2828 크기의 회색조로 이루어져 있다. Cifar-10 자료는 총 10개의 클래스로 이루어진 이미지자료로 각 이미지는 32×32 크기의 RGB로 이루어져 있고 훈련이미지 50,000개와 시험이미지 10,000개로 구성되어있다. CelebA자료는 유명 인사 얼굴에 관한 이미지자료이다. 10,177명에 대한 202,599장의 이미지가 있으며 각 이미지는 178×218 크기의 RGB로 이루어져 있으며 40개의 이진형 특성 주석을 가진다.

3.2. 측정

3.2.1. 인셉션 점수

인셉션모형 (Inception model; Szegedy 등, 2016)은 1,000개의 클래스와 120만개의 이미지로 구성된 ImageNet (Deng 등, 2009) 자료를 사전학습 (pre-trained)한 CNN모형으로 어떤 이미지가 입력되면 각 1,000개의 클래스에 속할 확률 벡터를 출력한다. 이는 전이학습 (transfer learning)과 미세조정

(fine tunning)에 널리 사용된다. 생성자료를 인셉션모형에 입력한 결과를 이용하여 다음과 같이 IS를 계산하여 생성모형을 평가하는데 큰 값이 좋은 품질을 의미한다 (Barratt과 Sharma, 2018).

$$IS(G) = \exp(\mathbb{E}_{\hat{x}\sim p_g} KL(p(y|\hat{x})||p(y))),$$

여기에서 $p(y|\hat{x})$는 조건부 클래스 분포 (conditional class distribution)이고 $p(y)$는 주변 클래스 분포 (marginal class distribution)이다. IS는 1이상 1,000 이하의 값을 가질 수 있지만 대개는 2 근방의 값을 가진다 (Barratt과 Sharma, 2018).

3.2.2. Fréchet 인셉션 거리

Heusel 등 (2017)이 개발한 FID는 두 정규 분포의 차이를 측정한 것으로 IS가 실제 자료의 분포를 사용하지 않는 단점을 보완하기 위해 제안되었으며 아래와 같이 계산되는데 작은 값이 좋은 품질을 의미한다.

$$FID = ||m - m_w||_2^2 + Tr(C + C_w - 2(CC_w)^{1/2}),$$

여기에서 (m, C)와 (m_w, c_w)는 생성자료 분포와 실제 자료의 평균과 공분산이다. FID는 두 분포의 평균과 분산의 차이를 계산하므로 큰 값을 가질 수 있다.

3.3. 분석 방법

잠재공간의 적절한 차원 크기를 추정하기 위하여 1, 2, 3, 10, 50, 100 차원, 즉 $z \sim U^d(-1, 1)$, $d = 1, 2, 3, 10, 50, 100$을 고려하였고, 각 차원별로 GAN-MM, GAN-NS, LSGAN, WGAN, WGAN-GP, DRAGAN 그리고 BEGAN을 학습하였다. 모든 GAN의 학습조건이 동일할 수는 없겠지만 비슷한 환경을 조성하도록 노력하였고 필요한 초모수 (hyperparemeter)는 격자탐색 (grid search)을 이용하여 선택하였다. GAN의 학습이 끝나면 생성자료의 IS와 FID 값을 측정하였는데, 이는 계산시간의 한계를 고려하여 3번 반복하였다.

모든 실험은 구글에서 만든 머신러닝 라이브러리 Tensorflow를 사용하였고 OS는 Ubuntu 16.04.4 LTS, CPU는 Intel Xeon Bronze 3106, GPU는 Nvidia Geforce Titan XP 그리고 RAM은 64GB를 사용하였다.

3.4. 분석 결과

Table 3.1과 Table 3.2는 각 차원별, GAN별 평균 FID와 IS를 나타낸다. 각 모형에서 가장 작은 FID와 큰 IS는 굵게 표시하였고, 각 자료에서 가장 작은 FID와 큰 IS를 가지는 모형도 굵게 표시하였다. Figure 3.1과 Figure 3.2은 실험에서 나온 각 값을 그림으로 나타내었는데 (a)는 Mnist (b)는 Fashion-Mnist (c)는 Cifar-10 그리고 (d)는 CelebA를 나타낸다.

Figure 3.3은 Mnist, Fashion-Mnist를 LSGAN모형이 그리고 Cifar-10, CelebA를 GAN_NS가 생성한 자료를 잠재공간의 크기 (a)는 1차원 (b)는 2차원 (c)는 3차원 (d)는 10차원 (e)는 50차원 (f)는 100차원에 따라 나타낸다.

Table 3.1 Average FID for various data, dimensions and GANs

Data	Model	1	2	3	10	50	100
Mnist	GAN-MM	316.36	33.56	17.10	**4.96**	5.46	5.27
	GAN-NS	186.43	29.60	14.21	4.93	4.93	**4.90**
	LSGAN	166.79	26.67	13.96	**4.79**	4.99	4.98
	WGAN	142.57	65.88	44.43	25.77	27.25	**24.35**
	WGAN-GP	176.28	92.95	45.21	18.94	20.94	**17.67**
	DRAGAN	221.04	44.64	22.68	8.46	**6.94**	8.36
	BEGAN	136.32	49.19	25.35	23.11	**18.79**	19.70
Fashion	GAN-MM	366.01	65.25	38.10	17.27	**16.31**	17.59
-Mnist	GAN-NS	232.19	68.33	37.46	15.94	15.87	**15.80**
	LSGAN	311.81	71.72	39.33	16.28	**15.28**	15.53
	WGAN	177.42	95.01	75.55	**61.83**	67.35	66.73
	WGAN-GP	197.58	93.35	54.78	36.21	**28.88**	30.12
	DRAGAN	255.11	87.64	42.76	23.49	**22.91**	27.28
	BEGAN	164.90	71.91	41.00	36.95	35.61	**33.25**
Cifar-10	GAN-MM	477.09	220.43	161.25	96.46	87.21	**85.29**
	GAN-NS	320.36	197.49	150.07	88.45	91.08	**83.20**
	LSGAN	326.35	205.97	160.87	86.03	**85.81**	89.41
	WGAN	270.80	178.49	157.73	**130.27**	148.65	144.60
	WGAN-GP	360.58	318.33	226.03	207.80	**179.99**	192.14
	DRAGAN	332.29	222.51	182.40	117.88	**101.53**	102.44
	BEGAN	361.71	236.29	186.26	117.88	152.20	**117.06**
CelebA	GAN-MM	459.99	319.79	414.02	**60.96**	72.69	88.90
	GAN-NS	376.76	160.36	94.71	51.21	**46.23**	53.26
	LSGAN	430.22	299.40	112.83	57.68	61.40	**54.17**
	WGAN	319.16	160.55	113.88	77.91	**76.51**	78.49
	WGAN-GP	442.79	379.45	323.68	212.01	118.12	**103.28**
	DRAGAN	436.35	168.98	167.05	72.52	**62.81**	63.33
	BEGAN	290.72	190.59	119.51	**67.23**	69.93	68.37

3.4.1. FID vs. IS

Table 3.2에서 IS는 Cifar-10을 제외한 자료에서는 1, 2, 3 차원에서도 가장 큰 값을 가지는 것으로 나타나는데 특히 CelebA에서 그런 현상이 많이 나타난다. 이것은 Figure 3.2에서도 확인 할 수 있는데 Cifar-10을 제외한 자료의 IS가 차원에 따라 증가하는 추세를 확인할 수 없다. 그렇지만 FID값은 1, 2, 3 차원에서 가장 작은 값을 가지는 경우가 없는 것으로 Table 3.1에서 나타난다. 그리고 모든 자료에서 FID 값은 10차원 까지는 그 값이 감소하는 것을 Figure 3.2에서도 확인 할 수 있다. 이는 Figure 3.3에서 1, 2, 3차원으로 생성된 이미지의 품질을 고려하면 IS보다는 FID가 생성자료의 품질을 측정하는 측도로 더 우수한 것으로 판단된다.

3.4.2. 잠재공간 크기

대부분의 모형과 자료에서 10차원까지는 FID 값이 급격히 감소하지만 10차원 이상에서는 큰 변화가 없는 것을 Table 3.1과 Figure 3.1에서 알 수 있다. 10차원에서 최소값을 갖는 경우도 많이 있고 또한 50차원과 100차원에서 갖는 값이 10차원의 값보다 유의하게 작은 경우는 많지 않음을 알 수 있는데 이것은 Figure 3.1 에서도 확인할 수 있다. 그리고 Figure 3.3과 Figure 3.4에서 1차원에서는 미완성된 이미지를 보이지만 Mnist와 Fashion-Mnist와 같은 회색조이미지는 2차원에서도 좋은 결과를 보임을 알 수 있다. 그러나 Figure 3.5와 Figure 3.6에서 좀 더 복잡한 칼라이미지인 Cifar-10과 CelebA에서는 10차원 미만차원에서 생성된 이미지는 온전하지 않지만 10차원 이상에서는 실제 자료와 유사한 자료를

Table 3.2 Average Inseption score for various data, dimensions and GANs

Data	Model	1	2	3	10	50	100
Mnist	**GAN-MM**	1.15	2.17	2.17	**2.19**	2.18	**2.19**
	GAN-NS	1.76	**2.22**	2.18	2.19	2.19	2.19
	LSGAN	1.82	2.15	**2.29**	2.19	2.17	2.18
	WGAN	1.92	2.14	2.15	2.31	2.23	2.23
	WGAN-GP	1.92	**2.27**	2.22	2.19	2.20	2.22
	DRAGAN	1.68	2.14	**2.25**	2.19	2.19	2.17
	BEGAN	1.90	**2.26**	2.17	2.21	**2.26**	**2.26**
Fashion	GAN-MM	1.50	4.02	4.06	**4.43**	4.33	4.35
-Mnist	GAN-NS	2.46	4.11	4.31	**4.43**	4.37	4.40
	LSGAN	1.36	4.04	**4.39**	4.32	4.33	4.34
	WGAN	3.19	3.61	**3.91**	3.89	3.79	3.76
	WGAN-GP	2.74	3.65	4.03	4.20	**4.28**	4.24
	DRAGAN	2.32	3.72	4.04	**4.42**	4.02	4.05
	BEGAN	2.82	3.82	**4.26**	4.25	4.02	4.05
Cifar-10	GAN-MM	1.01	2.46	3.15	4.57	5.12	**5.29**
	GAN-NS	2.13	2.81	3.28	4.92	4.67	**5.19**
	LSGAN	1.72	2.85	3.05	4.70	**5.04**	4.83
	WGAN	2.34	2.95	3.09	**3.10**	2.69	2.85
	WGAN-GP	1.67	1.89	2.18	2.21	2.38	**2.49**
	DRAGAN	1.90	2.67	2.80	3.79	**4.28**	4.16
	BEGAN	1.61	2.40	2.92	**4.32**	3.62	4.26
CelebA	GAN-MM	1.58	1.52	1.01	2.13	2.20	**2.43**
	GAN-NS	**2.58**	2.02	2.15	2.37	2.19	2.26
	LSGAN	2.05	2.24	2.02	2.23	**2.9**	2.24
	WGAN	**2.77**	2.25	2.39	2.27	2.24	2.26
	WGAN-GP	1.72	2.47	2.55	**3.03**	2.65	2.45
	DRAGAN	2.22	2.00	**2.58**	2.21	2.21	2.24
	BEGAN	1.47	**2.35**	2.12	2.14	2.22	2.22

생성하는 것으로 확인된다.

많은 연구에서 전형적으로 잠재공간의 크기를 100차원으로 사용하지만 본 연구에서는 10차원도 충분한 것으로 확인하였다. 이는 차원을 줄이면 계산에 필요한 공간과 시간을 절약할 수 있으므로 의미있는 결론으로 평가된다.

3.4.3. GAN 비교

Lucic 등 (2017)에서도 다룬 것처럼 실험 환경과 조건을 동일하게 하는 것이 어렵기 때문에 어떤 GAN이 가장 우수한지에 대한 논란은 계속된다. 본 연구의 FID를 기준으로 평가하면 Mnist와 Fashion-Mnist에서는 LSGAN, Cifar-10과 CelebA에서는 GAN-NS가 가장 작은 값을 가지는 좋은 모형으로 평가된다.

성능이 안정된 10차원 이후에서 모형의 성능을 비교한 결과 GAN-NS와 LSGAN은 회색조이미지와 컬러이미지 모두에서 안정적으로 우수한 성능을 보인 반면 많은 주목을 받았던 WGAN은 회색조이미지에서 낮은 성능을 보였고 컬러이미지에서는 WGAN-GP가 낮은 성능을 보였다. 이상의 결과를 종합하여 알고리즘에 대한 이해가 쉽고 성능이 우수한 것으로 평가되는 LSGAN의 사용을 권장한다.

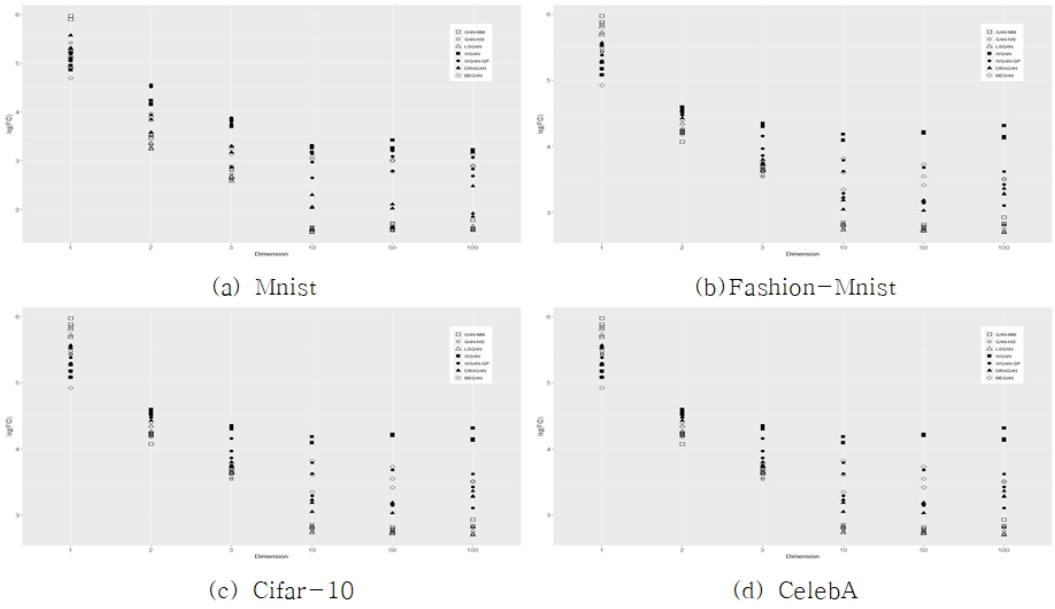

Figure 3.1 Scatter plots of log(FID) versus dimension with Mnist (a) Fashion-Mnist (b) Cifar-10 (c) CelebA (d)

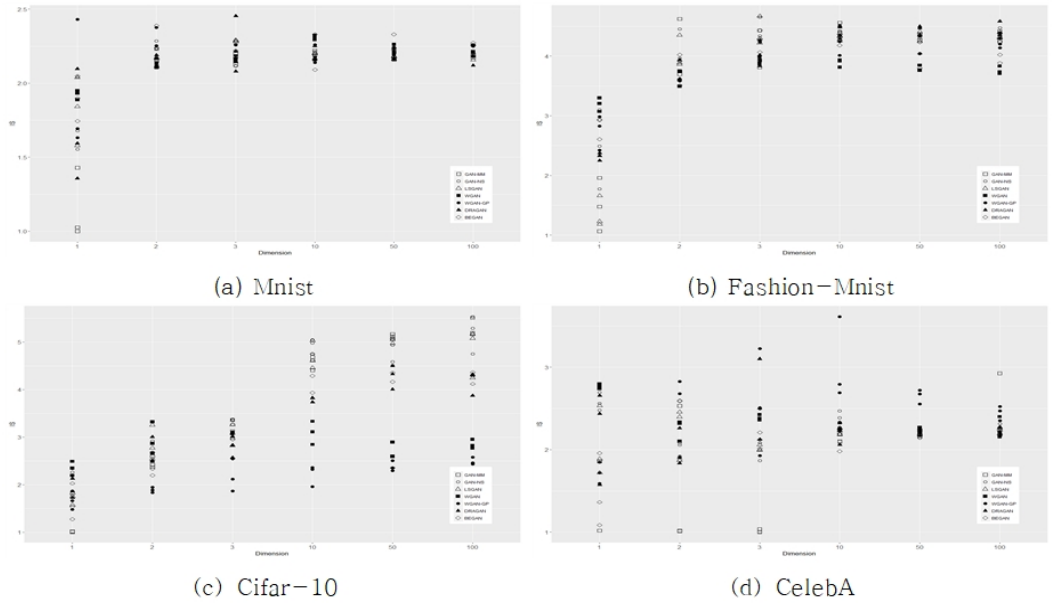

Figure 3.2 Scatter plots of Inseption score versus dimension with Mnist (a) Fashion-Mnist (b) Cifar-10 (c) CelebA (d)

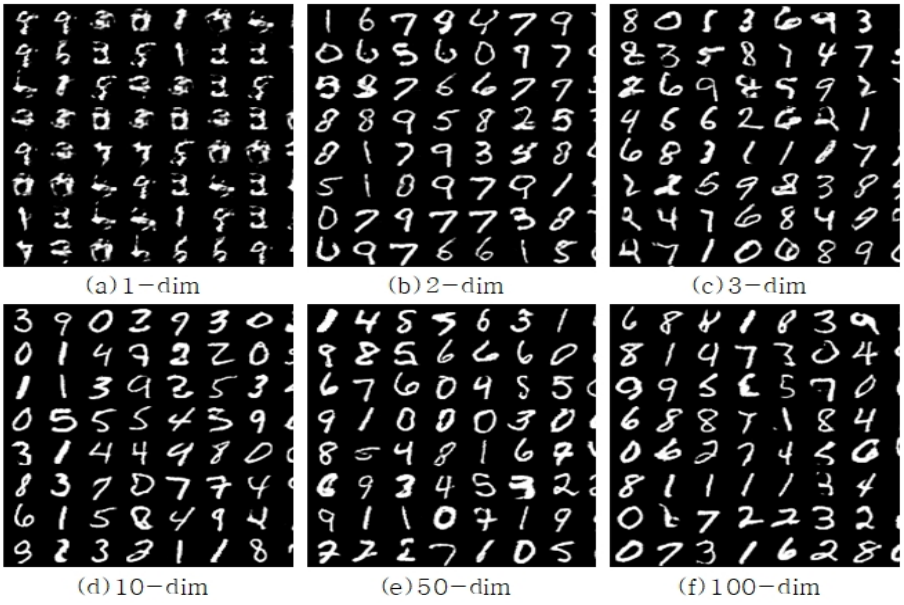

Figure 3.3 Generated images of Mnist with LSGAN for 1, 2, 3, 10, 50 and 100 dimensional latent space

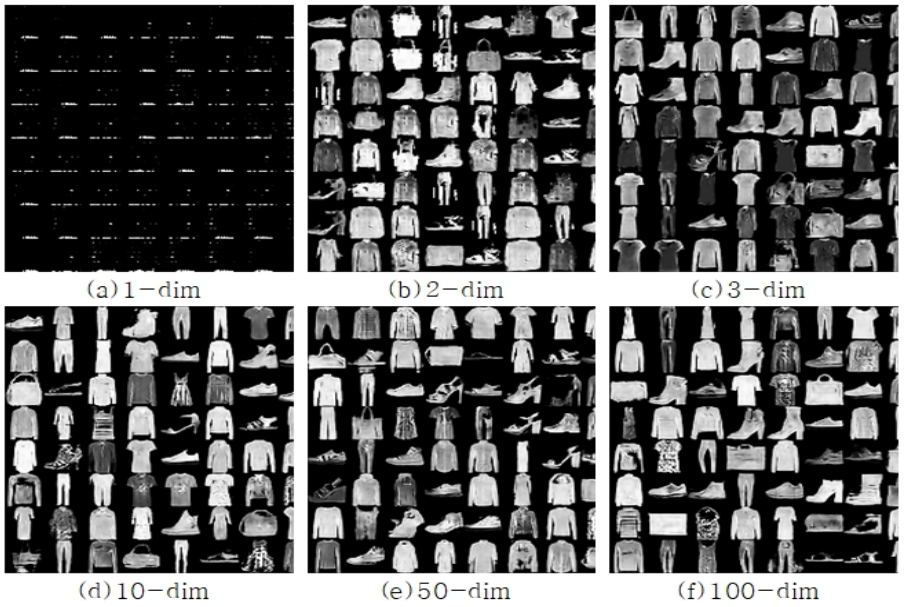

Figure 3.4 Generated images of Fashion-Mnist with LSGAN for 1, 2, 3, 10, 50 and 100 dimensional latent space

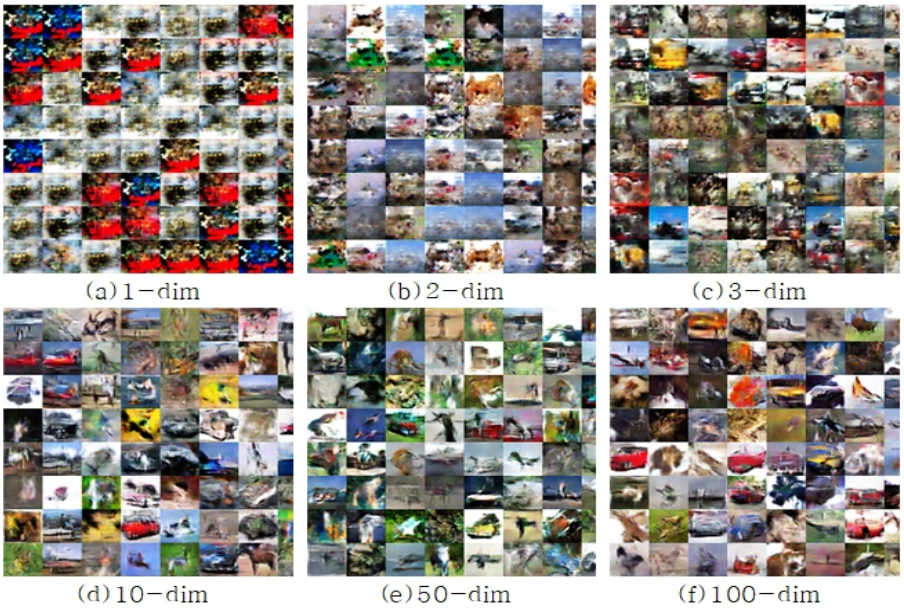

Figure 3.5 Generated images of Cifar-10 with GAN-NS for 1, 2, 3, 10, 50 and 100 dimensional latent space

Figure 3.6 Generated images of CelebA with GAN-NS for 1, 2, 3, 10, 50 and 100 dimensional latent space

4. 결론 및 제언

본 연구에서는 GAN을 비롯해 성능이 우수하다고 평가되는 GAN-NS, BEGAN, DRAGAN, LSGAN, WGAN, WGAN-GP를 정리 및 요약하고 성능을 평가하였다. 그리고 GAN의 입력으로 이용되는 잠재공간의 차원크기를 1, 2, 3, 10, 50, 100으로 나누어 적절한 차원크기를 추정하였고 생성 자료를 통해 FID와 IS의 적절성도 평가 하였다.

학습 자료는 회색조이미지로 Mnist와 Fashion-Mnist를 사용하였고 컬러이미지로 Cifar-10, CelebA을 사용하여 3번의 반복실험을 실시하였다.

잠재공간의 적절한 차원크기를 추정하는 연구에서는 1, 2, 3, 10차원까지는 성능이 향상되는 모습을 보였고 10, 50, 100차원에서는 성능을 유지하여 10차원에서도 전형적인 100차원과 차이가 없는 것을 확인 할 수 있었다. 그러나 WGAN-GP의 경우 다른 모형들과 달리 CelebA자료에서 10, 50, 100차원에서도 지속적인 성능향상이 있었다.

성능이 안정된 10차원 이후에서 모형의 성능을 비교한 결과 GAN-NS와 LSGAN은 회색조이미지와 컬러이미지 모두에서 안정적으로 우수한 성능을 보인 반면 WGAN은 회색조이미지에서 낮은 성능을 보였고 컬러이미지에서는 WGAN-GP가 낮은 성능을 보였다. 성능을 평가하는 측도인 IS는 일부 자료에서는 1차원에서도 성능이 높은 것으로 평가하는 등 FID 보다 부족한 모습을 보였다.

신경망의 구조를 비롯한 학습률, 배치정규화 (batch normaization) 유무, 드롭아웃 (dropout)비율 등의 초모수를 다양하게 이용하여 비교하는 추후 연구과제를 제언하는 바이다. 그리고 더 많은 환경에서 객관적으로 두 측도를 비교하는 연구도 요구된다.

References

Arjovsky, M., Chintala, S. and Bottou, L. (2017). Wasserstein GAN. *arXiv preprint arXiv:1701.07875*.

Barratt, S. and Sharma, R. (2018). A note on the Inception score. *arXiv preprint arXiv:1801.01973*.

Berthelot, D., Schumm, T. and Metz, L. (2017). Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *In 2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248-255.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014). Generative adversarial nets. *In Advances in Neural Information Processing Systems*, 2672-2680.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V. and Courville, A. C. (2017). Improved training of Wasserstein GANs. *In Advances in Neural Information Processing Systems*, 5769-5779.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Klambauer, G. and Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a Nash equilibrium. *arXiv preprint arXiv:1706.08500*.

Kodali, N., Hays, J., Abernethy, J. and Kira, Z. (2017). On convergence and stability of GANs. *arXiv preprint arXiv:1705.07215*.

Lucic, M., Kurach, K., Michalski, M., Gelly, S. and Bousquet, O. (2017). Are GANs created equal? A large-scale study. *arXiv preprint arXiv:1711.10337]/*.

Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z. and Smolley, S. P. (2017). Least squares generative adversarial networks. *In 2017 IEEE International Conference on Computer Vision (ICCV)*, 2813-2821.

Radford, A., Metz, L. and Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A. (2015). Going deeper with convolutions. *In 2015 IEEE Conference on Computer Vision and Pattern Recognition*, 1-9.

Zhao, J., Mathieu, M. and LeCun, Y. (2016). Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*.

Journal of the Korean Data & Information Science Society
2018, **29**(5), 1155–1167

http://dx.doi.org/10.7465/jkdi.2018.29.5.1155
한국데이터정보과학회지

A study on the performance of generative adversarial networks[†]

Yeongjae Lee[1] · Kyungha Seok[2]

[12]Department of Statistics, Inje University

Received 5 July 2018, revised 10 August 2018, accepted 20 August 2018

Abstract

Generative Adversarial Networks (GAN) is one of the most popular models in generative deep learning models. Many derivatives have been published and researches have been conducted in various fields. In this study, we review the derivatives of GAN and compare them. We determine the proper dimension of the latent space and compare the metrics Fréchet Inception distance (FID) and Inception score (IS) which are used for evaluating generated data. The experiments show that GAN-NS and LSGAN works well and FID is superior to IS. And the 10 dimensional latent spaces yield good results, which is not much different from the result of typical 100 dimensions.

Keywords: Deep learning, Fréchet Inception distance, generative adversarial networks, Inception score, latent space.

[†] This work was supported by the 2017 Inje University research grant.
[1] Master of Science, Department of Statistics, Inje University, Gyungnam 50834, Korea.
[2] Corresponding Author: Professor, Institute of Statistical Information, Department of Statistics, Inje University, Gyungnam 50834, Korea. E-mail: statskh@inje.ac.kr