Journal of Korea Multimedia Society Vol. 24, No. 1, January 2021(pp. 13-20)
https://doi.org/10.9717/kmms.2020.24.1.013

WGAN의 성능개선을 위한 효과적인 정칙항 제안

한 희 일†

Proposing Effective Regularization Terms for Improvement of WGAN

Hee Il Hahn†

ABSTRACT

A Wasserstein GAN(WGAN), optimum in terms of minimizing Wasserstein distance, still suffers from inconsistent convergence or unexpected output due to inherent learning instability. It is widely known some kinds of restriction on the discriminative function should be considered to solve such problems, which implies the importance of Lipschitz continuity. Unfortunately, there are few known methods to satisfactorily maintain the Lipschitz continuity of the discriminative function. In this paper we propose techniques to stably maintain the Lipschitz continuity of the discriminative function by adding effective regularization terms to the objective function, which limit the magnitude of the gradient vectors of the discriminator to one or less. Extensive experiments are conducted to evaluate the performance of the proposed techniques, which shows the single-sided penalty improves convergence compared with the gradient penalty at the early learning process, while the proposed additional penalty increases inception scores by 0.18 after 100,000 number of learning.

Key words: Deep Learning, Generative Model, Lipschitz Continuity, Training Stability, Wasserstein Distance, Wasserstein GAN, Regularization Terms

1. 서 론

GAN(Generative Adversarial Network)은 게임 이론에 기반한 생성모델로서, 학습을 통하여 원하는 데이터 표본을 위한 생성모델 구축을 목표로 한다 [1]. 이전의 생성모델로 널리 알려진 가변 자동 인코더(variational autoencoder)가 로그우도(log likelihood)를 최대화하도록 학습하는 과정이 요구되는 반면에 [2], GAN은 개념적으로 완전히 다른 모델이다. 즉, GAN은 생성기(generator)와 분별기(discriminator)를 서로 대립시켜 학습하는 모델이다. 분별기는 원하는 데이터 그룹에 속하는 데이터와 생성기가 출력한 데이터를 정확히 분별하도록 학습시키는 반면에 생성기는 분류기가 전혀 구별할 수 없도록 학습시킨다.

생성기의 확률분포 P_g를 학습하기 위해서는 우선 생성기의 입력신호로 이용될 랜덤잡음의 확률분포 $P_z(z)$가 정의되어야 한다. 생성기는 함수 $g(z;\theta)$로 나타내는데, 여기서 g는 미분가능하고 매개변수 θ를 갖는 딥러닝 알고리즘으로 구현된다. 랜덤변수 $z \sim P_z(z)$를 적절히 표본화하여 이를 생성기 g에 입력시킴으로써 원하는 데이터 분포 P_r를 갖는 출력을 생성하도

※ Corresponding Author : Hee Il Hahn, Address: (17035) 81, Oedae-ro, Mohyeon-eup, Cheoin-gu, Yongin-si, Gyeonggi-do, Republic of Korea, TEL : +82-31-330-4356, FAX : +82-31-333-4256, E-mail : hihahn@hufs.ac.kr
Receipt date : Nov. 27, 2020, Revision date : Dec. 24, 2020
Approval date : Jan. 4, 2021

† Dept. of Information & Communications Eng., College of Engineering, Hankuk University of Foreign Studies
※ This research was supported by the Research Program 2020 funded by Hankuk University of Foreign Studies

록 학습한다. 반면에 또 다른 딥러닝으로 구현되는 분별기 $d(x;w)$는 학습데이터와 생성기 출력을 입력받으면 각각 1과 0을 출력하도록 학습된다. 이와 같이 생성기와 분별기는 모두 딥러닝으로 구현되는 비선형 함수이고 학습데이터의 확률분포는 두 함수 내에 근사적으로 내재된다. 문제는 매개변수가 수렴함에 따라 해당되는 확률분포도 함께 수렴하는가에 있다. GAN은 JS 발산(Jensen-Shannon divergence)을 최소화시키는 방향으로 학습된다[1]. 그런데 학습 초기에는 대부분 P_r과 P_g가 측정되는 영역이 서로 겹치지 않아서(disjoint) P_g가 P_r로 수렴하기 위한 정보를 얻을 수 없기 때문에 분별기 학습 시 학습 안정성을 기대하기 어려운 문제를 갖고 있다. 이를 해결하기 위해서는 유연하면서도 수렴에 초점을 맞춘 다른 메트릭이 필요한데, 이의 대안으로 Wasserstein 거리가 이용된다. 이 거리를 이용한 GAN을 WGAN(Wasserstein GAN)이라고 부른다[3]. WGAN은 분별기와 유사한 역할을 담당하는 내부함수(critic)가 립쉬츠 연속(Lipschitz continuous)이어야 안정적인 수렴이 보장되는 제약조건이 있다. 이 조건이 P_r과 P_g의 측정영역이 겹치지 않아도 P_g가 P_r로 수렴하기 위한 정보를 제공하는 역할을 수행하므로 보다 안정적인 수렴을 얻을 수 있는 장점이 있다.

WGAN의 성능을 개선하기 위해서는 분별기 함수의 립쉬츠 연속을 안정적으로 유지시키는 것이 관건인데, 이를 해결하기 위한 많은 연구가 진행되고 있다. 본 논문에서는 이론적 분석을 통하여 이 문제를 해결하기 위한 기법을 제안하고 실험을 통하여 기존의 대표적인 알고리즘과 그 성능을 비교분석한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련연구를 리뷰한다. 3장에서는 WGAN에 대한 이론적인 내용과 분별기의 립쉬츠 연속의 효과를 알아보고 제안 알고리즘의 이론적 특성을 설명한다. 4장에서는 기존의 대표적인 WGAN 알고리즘들과 제안 알고리즘들의 수렴속도와 출력이미지의 화질 등에 대하여 실험결과를 기반으로 비교분석한다. 마지막으로 5장에서는 결론을 맺고 향후 연구 진행방향에 대하여 논의한다.

2. 관련 연구

GAN은 다음과 같이 주어진 목적함수를 이용하여 분별기와 생성기를 학습시킨다.

$$L(d_w, g_\theta) = E_{x \sim P_d}[\log d_w(x)] + E_{x \sim P_g}[\log(1 - d_w(x))] \quad (1)$$

GAN의 학습은 두 단계의 학습이 반복되는 과정으로 구성된다. 첫 번째 단계에서는 이전 과정에서 획득된 생성기를 이용하여 $L(d_w, g_\theta)$가 최댓값을 갖도록 분별기를 학습시킨다. 두 번째 단계에서는 첫 번째 단계에서 얻어진 분별기를 이용하여 $L(d_w, g_\theta)$가 최솟값을 갖도록 생성기를 확습시킨다.

최적의 분별기 $d^*(x)$가 구해지면 이에 따라 식 (1)을 최소화시키는 생성기 g를 구한다. 그런데 실제로 학습해보면, 분별기가 잘 학습됨에 따라 생성기는 오히려 성능이 저하되는 불안정한 현상이 많이 노출된다. 분별기가 생성기보다 먼저 최적으로 학습되어 있는 상태에서 손실함수를 최소화시키는 학습을 진행하면 P_r과 P_g 간의 JS 발산을 최소화시키는 작업으로 귀결된다. 이것은 결과적으로 분별기가 포화단계로 수렴하면서 미분크기가 사라지는(vanishing gradient) 문제를 야기시킨다[4]. Radford, *et al.* [5]는 DMLP(Deep Multi-Layer Perceptron) 대신 CNN (Convolutional Neural Network)을 사용한 DCGAN (Deep Convolutional GAN)을 제안하여 이미지 생성 등의 응용에 특히 성능이 향상되는 이득은 있으나 안정적인 학습을 어렵게 하는 GAN 고유의 문제점을 해결하지는 못하고 있다. 이외에도 GAN의 학습 불안정성을 규명하고 개선하려는 연구는 다양한 방법으로 발표되었으나[6,7,8,9], 근본적인 해결책을 제시하지 못하고 있다.

생성기 출력의 확률분포 P_g가 학습을 통하여 매개변수를 최적화시킴으로써 데이터집합의 확률분포 P_r에 안정적으로 수렴하여야 한다. 이를 위해서는 두 분포 P_r과 P_g 간의 거리개념인 손실함수가 매개변수에 연속이어야 한다. 그런데 GAN에서 채택된 JS 발산은 수렴의 연속성을 만족시키지 못하기 때문에 이의 안정적 학습이 어렵고 성능을 저해하는 원인으로 작용한다[3]. 이에 반해서 Wasserstein 거리는 P_r과 P_g의 측정영역이 서로 겹치지 않아도(disjoint) 이들 간의 거리를 측정할 수 있고 또한 연속이어서 수렴을 판정하는데 매우 유연한 장점이 있다. Arjovsky, et al.[3]은 Wasserstein 거리를 채택한 WGAN을 제안한다. WGAN은 GAN의 치명적 약점인 모드 드로핑

(mode dropping)을 줄이고 분별기와 생성기에 대한 보다 안정적인 학습을 가능하게 하는 것으로 알려져 있다. 그런데 문제는 생성기 함수 g가 립쉬츠 연속이어야 Wasserstein 거리가 연속임을 보장할 수 있다. 이를 위하여 Arjovsky, *et al.*[3] 은 딥러닝으로 분별기를 학습하는 모든 과정에서 가중치를 고정된 임계값으로 클리핑하는 기법을 도입한다.

이 방식은 가중치를 강제로 클리핑함으로써 분별기가 지나치게 간략화된 함수로 수렴하고 학습 데이터 분포의 고차원 모멘트를 왜곡하는 것으로 알려져 있다. Gulrajani, *et al.*[10]는 함수의 미분크기와 1 간의 거리를 정칙(regularization) 항으로 이용함으로써 1-립쉬츠 연속함수를 구하고자 한다. Gulrajani의 알고리즘은 최적의 분별기 f^*가 미분 가능하고 x와 y가 최적의 결합 분포 π^*에서 샘플링되었다는 가정하에 유도된다. 하지만 주변 분포 P_r과 P_g로부터 독립적으로 샘플링된 것이 π^*의 정의구역에서 벗어난 곳에서 $x \sim P_g$와 $y \sim P_r$를 샘플링할 가능성이 높고, f^*가 미분 가능일 필요도 없다. 립쉬츠 연속인 함수는 옹골공간(compact space)에서 균일 미분가능 함수로 근사화시킬 수 있다. 그렇더라도 미분가능하지 않는 영역에서도 경사벡터(gradient)의 크기를 1로 제한하는 것은 매우 강한 제약이다. 이를 고려하여 미분크기가 1 보다 클 때에만 양의 값을 부과하는 정칙 항으로 변경하면 수렴속도와 생성기의 성능이 개선됨을 보여주는 연구결과도 발표된다[11,12]. 본 논문에서는 함수의 1-립쉬츠 연속을 보다 안정적으로 보장하는 기법을 이용하여 WGAN의 성능을 보다 향상시키는 방법을 제안하고 실험으로 성능을 확인한다.

3. 정칙 항을 통한 WGAN의 성능개선

3.1 WGAN

Wasserstein 거리함수 $W(P_r, P_g)$는 Dobrushin이 1970년에 제안한 것으로 다음과 같이 정의된다[13].

$$W(P_r, P_g) = inf_{\gamma \in \pi(P_r, P_g)} \mathbb{E}_{(x,y)\sim\gamma}[d(x(w), y(w))] \quad (2)$$

여기서 w는 모집단 Ω의 표본이고, $\pi(P_r, P_g)$는 모든 결합 확률분포(joint distribution) $\gamma(x,y)$의 집합으로서 $P_r = \int \gamma(x,y)dy$ 이고 $P_g = \int \gamma(x,y)dx$ 이다. 즉, $W(P_r, P_g)$는 $\pi(P_r, P_g)$ 중에서 $d(x,y)$의 기댓값을 최소로 추정한 값이다. 샘플링을 반복할수록 $\gamma(x,y)$는 최적의 연결(optimal coupling)로 수렴한다. 이 때, x와 y의 연결이 어떻게 이루어지는지에 따라 $d(x,y)$의 분포가 달라진다. 이와 같이 Wasserstein 거리는 여러 가지 $\gamma(x,y)$ 중에서 $d(x,y)$의 기댓값이 가장 작게 나오는 확률분포를 취한다.

KL(Kullback-Leibler) 발산이나 JS 발산은 P_r과 P_g가 서로 겹치지 않는(disjoint) 경우에서는 불연속인데 반해서 Wasserstein 거리는 연속이어서 수렴을 판정하는데 매우 유연한 장점이 있다. 즉, 분포 수렴(converge in distribution)한다[14]. $d(x,y)$를 유클리드 거리 $\|x-y\|$로 지정하고 P_r과 P_g의 확률측도(probability measure)를 각각 μ와 ν로 정의하면 식 (2)는 Kantorovich-Rubinstein 이중성 [15]에 의해 다음과 같이 유도될 수 있다.

$$W(P_r, P_g) = sup_{\|f\|_L \le 1}\{\mathbb{E}_{y\sim\mu}[f(y)] - \mathbb{E}_{x\sim\nu}[f(x)]\} \quad (3)$$

여기서 $\|f\|_L \le 1$는 함수 f가 1-립쉬츠 연속임을 의미한다. 1-립쉬츠 연속함수란 정의구역 내의 모든 x, y에 대하여 다음 식을 만족시키는 함수를 의미한다.

$$|f(y)-f(x)| \le \|y-x\| \quad (4)$$

GAN에서 분별기는 입력신호가 실제 데이터인지 아니면 생성된 데이터인지를 판별하도록 학습되는 대신 함수 $f(x)$는 Wasserstein 거리를 계산하는데 도움이 되는 함수로 학습된다는 점에서 GAN에서의 분별기와 그 역할이 기본적으로 다르고 "critic"이라고 부르지만, 편의상 본 논문에서는 분별기라고 부른다.

3.2 립쉬츠 연속 조건의 효과

P_r과 P_g의 측정영역(support)은 학습 초기에는 일반적으로 서로 겹치지 않는다. GAN에서 생성기는 그 출력을 업데이트시키기 위하여 경사벡터 $\nabla_x d_w(x)$를 이용하지만 $\nabla_x d_w(x)$는 실제로 P_r의 측정영역이 어디에 있는지 알지 못한다. 즉, 식 (1)에서 생성기가 분별기로부터 얻는 경사벡터는 다음과 같이 구할 수 있다.

$$\nabla_x L(d_w, g_\theta) = E_{x\sim P_g}\left[\nabla_{d_w}\log(1-d_w(x)) \bullet \nabla_x d_w(x)\right] \quad (5)$$

우측의 첫 번째 항은 log 함수의 변화량을 경사벡터의 크기에 영향을 주므로 본 논문에서는 큰 의미를 두지 않는다. 두 번째 항은 생성기가 출력한 $x \sim P_g$에

서 최적화하기 위해 생성기가 향해야 할 방향을 제시한다. 그런데 P_r과 P_g의 측정영역이 서로 겹치지 않을 때에는 $\nabla_x d_w(x)$가 P_r에 대한 유의미한 정보를 포함하지 않기 때문에 P_g가 P_r로 수렴하는 것을 보장하지 못한다. 이것이 GAN을 학습시키기 어려운 주원인 중의 하나이다. 반면에, 식 (3)을 만족하려면 모든 $x \sim P_g$와 $y \sim P_r$에 대하여 $f(y) > f(x)$이고 이를 식 (4)와 결합하면 다음과 같은 제약조건을 얻을 수 있다.

$$\frac{f(y)-f(x)}{\| y-x \|} \leq 1 \tag{6}$$

이와 같은 립쉬츠 연속 제약조건은 P_r과 P_g가 분리되어 있더라도 $f(x)$를 통하여 P_g와 P_r을 연결시켜 주는 역할을 한다. 즉, 생성기가 출력한 신호 $x \sim P_g$에서 충분히 수렴된 최적의 f인 f^*의 경사벡터 $\nabla_x f^*(x)$는 $y \sim P_r$을 향하게 되어 매 학습 단계마다 P_g가 P_r로 접근할 수 있도록 유도하는 역할을 한다.

3.3 립쉬츠 연속을 위한 기존의 구현방안

WGAN은 식 (6)을 만족시키는 1-립쉬츠 연속함수 f를 분별기로 이용하는데, 이는 신경망을 통하여 구현할 수 있다. Arjovsky, et al [3]은 함수 f가 1-립쉬츠임을 유지하기 위하여 f의 가중치를 다음과 같이 일방적으로 클리핑시킨다.

$$w = \begin{cases} c & w \geq c \\ -c & w \leq -c \end{cases} \tag{7}$$

이 방식(WGAN으로 명명한다.)은 가중치를 강제로 클리핑함으로써 함수 f의 미분크기가 매우 작은 값을 갖거나 매우 큰 값을 갖는 등의 이유로 안정적인 학습이 어렵고 수렴하더라도 지나치게 간략화된 함수로 수렴할 뿐만 아니라 학습 데이터 분포의 고차원 모멘트를 왜곡하여 최적화 과정을 어렵게 한다. 이러한 현상은 가중치 크기 제한과 목적함수 간의 상호작용에 기인하는 것으로 알려져 있다. 이러한 문제점을 보완하기 위하여 Gulrajani, *et al.*[10]는 식 (3)에 주어진 함수에 부호를 변경하여 손실함수로 변환한 다음 경사 페널티(gradient penalty)를 가하는 정칙(regularization) 항을 다음과 같이 추가함으로써 1-립쉬츠 함수를 구하는 WGAN-GP 알고리즘을 제안한다.

$$L_{gp} = \mathbb{E}_{x \sim P_g}[f(x)] - \mathbb{E}_{y \sim P_r}[f(y)] + \lambda \mathbb{E}_{v \sim P_v}\left[(\| \nabla_v f(v) \| - 1)^2\right] \tag{8}$$

위 식에서 $v = \alpha y + (1-\alpha)x$, $0 \leq \alpha \leq 1$는 x와 y를 연결하는 직선 상에서 랜덤하게 선택한 점이고 P_v는 그러한 v의 확률분포이다. 우식의 앞 두 항은 식 (3)의 부호를 변경한 것이고 세 번째 항은 f가 1-립쉬츠 연속을 유지하도록 부가된 정칙 항이다. 이 방식은, 미분가능한 함수는 립쉬츠 연속임을 이용하여 미분크기가 1을 갖도록 제한함으로써 f가 1-립쉬츠 연속이 되도록 유도한다.

하지만 이 기법은 학습 표본과 생성기가 만든 표본이 어떤 결합분포에서 추출되어야 하는데 실제로는 각 주변 분포(marginal distribution)에서 독립적으로 추출된다. Petzka [12]에 의하면 x와 y가 최적의 결합분포 π^*에서 샘플링되면 $v = \alpha y + (1-\alpha)x$, $0 \leq \alpha \leq 1$에서 $|f^*(y) - f^*(v)| = \| y - v \|$를 만족하지만 x와 y가 주변 분포에서 독립적으로 샘플링되면 위 식은 $|f^*(y) - f^*(v)| \leq \| y - v \|$으로 수정되어야 한다. 식 (8)에 주어진 Gulrajani의 정칙항 $\mathbb{E}_{u \sim P_v}\left[(\| \nabla_v f(v) \| - 1)^2\right]$ 대신에 다음과 같은 단면 페널티(single-sided penalty)를 가하는 정칙 항을 이용할 때(WGAN_SP라고 명명한다.) 수렴속도와 성능이 개선됨을 확인할 수 있다[11,12].

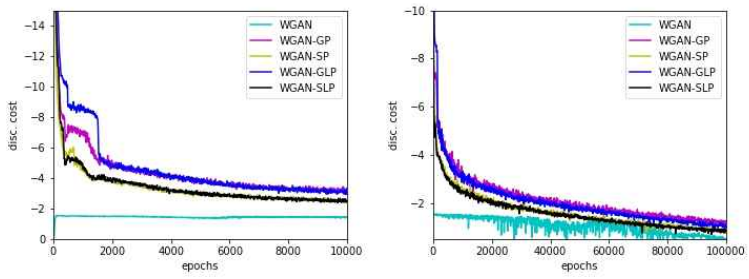

Fig. 1. The negative training and validation losses of WGAN, WGAN_GP and the proposed models when CIFAR-10 dataset is used. (a) The training losses (b) The validation losses.

Fig. 2. CIFAR-10 samples generated by WGAN, WGAN-GP and the proposed models at the epochs of 1,000, 2,000, 3,000, 4,000 and 5,000. (a) WGAN (b) WGAN-GP (c) WGAN-SP (d) WGAN-GLP (e) WGAN-SLP

$$L_{sp} = \mathbb{E}_{x \sim P_g}[f(x)] - \mathbb{E}_{y \sim P_r}[f(y)] + \lambda \mathbb{E}_{v \sim P_v}\left(\| \nabla_v f(v) \| - 1 \right)^2 u\left(\| \nabla_v f(v) \| - 1 \right) \quad (9)$$

식 (9)에서 $u(x)$는 단위 계단함수를 의미한다. 이와 같이 분별기 $f(x)$에 1-립쉬츠 연속 제약 조건을 강화시켜 주면 WGAN의 성능 향상을 기대할 수 있다.

3.4 제안 알고리즘

식 (8)에 제시된 정칙 항은 f의 미분크기가 1이 아니면 이에 비례한 페널티를 부과함으로써 f의 미분크기가 1을 유지하도록 유도하는 역할을 수행한다. f의 미분크기가 1을 유지하면 f가 1-립쉬츠 연속이기는 하지만, 1 이하의 값을 가져도 1-립쉬츠 연속이므로 식 (9)를 정칙항으로 이용함으로써 어느 정도의 성능향상을 얻을 수 있다. 본 논문에서는 보다 안정된 립쉬츠 연속을 만족시키기 위하여 식 (6)의 조건에서 벗어날 때 페널티를 가하는, 다음과 같은 립쉬츠 페널티를 추가하는 방식을 제안한다.

$$\sum \left\{ \max\left(0, \frac{|f(y) - f(x)|}{\| y - x \|} - 1\right) \right\}^2 \quad y \sim P_r,\ x \sim P_g \quad (10)$$

즉, WGAN-GP에 식 (10)에 주어진 립쉬츠 페널티를 추가한 알고리즘을 WGAN-GLP로 부르고 다음과 같은 손실함수를 갖는다.

$$L_{glp} = \mathbb{E}_{x \sim P_g}[f(x)] - \mathbb{E}_{y \sim P_r}[f(y)] + \lambda \mathbb{E}_{v \sim P_v}\left[\left(\| \nabla_v f(v) \| - 1 \right)^2 \right] + \mu \sum \left\{ \max\left(0, \frac{|f(y) - f(x)|}{\| y - x \|} - 1\right) \right\}^2 \quad (11)$$

그리고 WGAN-SP에 립쉬츠 페널티를 추가시킨 WGAN-SLP의 손실함수는 다음과 같이 주어진다.

$$L_{slp} = \mathbb{E}_{x \sim P_g}[f(x)] - \mathbb{E}_{y \sim P_r}[f(y)] + \lambda \mathbb{E}_{v \sim P_v}\left[\left(\| \nabla_v f(v) \| - 1 \right)^2 u\left(\| \nabla_v f(v) \| - 1 \right) \right] + \mu \sum \left\{ \max\left(0, \frac{|f(y) - f(x)|}{\| y - x \|} - 1\right) \right\}^2 \quad (12)$$

학습 초기에는 생성기가 출력한 x가 실제 데이터 y와 매우 다르므로 이들 간의 거리가 커서 식 (10)의 립쉬츠 페널티는 거의 0에 가까운 값을 갖는다. 따라서 립쉬츠 페널티만 단독으로 이용하면 학습이 잘 이루어지지 않는다. 하지만, WGAN-GP나 WGAN-SP와 결합하면 수렴이 진행됨에 따라 x와 y 간의 거리가 작아지면서 립쉬츠 페널티는 유의미한 정보를 갖게 되어 성능향상을 기대할 수 있다.

4. 실험 및 토론

본 논문의 모든 실험은 최대한 동일한 조건에서 WGAN, WGAN-GP 등과 제안모델들을 비교하기 위하여, 각 알고리즘들을 학습시키는데 있어서 생성기와 분별기의 구조, 학습률 등의 매개변수(hyper-parameter) 등은 해당 논문에서 제시한 최적의 결과를 그대로 채택한다. 제안 알고리즘의 성능을 확인하기 위하여 학습데이터로 CIFAR-10[16]을 이용한다. CIFAR-10 데이터 집합은 60,000개의 32×32 화소 크기의 컬러 이미지로 구성되어 있고 비행기, 자동차, 새, 고양이 등의 10개 부류 당 각각 6,000개의 이미지를 포함한다. 본 논문에서 제안한 알고리즘과 WGAN, WGAN-GP 등의 알고리즘의 성능을 비교하기 위하여 1,000개의 이미지를 평가 이미지로 이용한다.

Fig. 1은 각 알고리즘들에 대하여 학습 횟수에 따른 손실함수의 변화와 학습데이터에 속하지 않은

Fig. 4. CIFAR-10 samples generated by WGAN, WGAN-GP and the proposed models after 100,000 epochs of training. (a) WGAN (b) WGAN-GP (c) WGAN-SP (d) WGAN-GLP (e) WGAN-SLP.

1,000개의 랜덤으로 추출한 이미지에 대한 손실함수(validation loss)를 그래프로 보여준다.

이 그래프를 보면, 모든 알고리즘이 기본적으로 안정적으로 수렴하고 있으며, 특히 WGAN은 가중치를 강제로 클리핑하기 때문에 손실함수가 시작점부터 최적 값에서 크게 벗어나 있지 않다. 이 알고리즘은 학습횟수가 증가해도 별 감소효과도 보이지 않아 손실함수의 변화로 다른 알고리즘들과 비교하는 것은 큰 의미가 없다. 다른 알고리즘들에 대해서는 학습회수가 증가함에 따라 WGAN-SP와 WGAN-SLP가 가장 빠르고 WGAN-GLP, WGAN-GP의 순으로 수렴하는 것을 볼 수 있다. 이러한 사실은 단면 페널티가 수렴속도의 관점에서 경사 페널티보다 성능이 우수함을 보여준다. 이러한 내용이 생성된 이미지의 화질에 어떻게 반영되는지 확인하기 위하여 5개의 고정된 학습회수에서 각 모델이 생성한 이미지들의 예를 Fig. 2에 제시한다. 이 그림을 보면 학습초기에는 WGAN-GP보다 WGAN이 오히려 수렴속도가 더 빠르고 2,500회 이상의 학습 후에 역전되는 현상이 나타난다. 이에 비하여 WGAN-SP, WGAN-GLP 및 WGAN-SLP 등의 제안 알고리즘은 처음부터 훨씬 빠른 속도로 수렴함을 알 수 있다. Fig. 1의 결과와 비교하면 정량적인 손실함수의 변화와 이에 따른 화질의 변화는 반드시 일치하지는 않는다는 것을 알 수 있다. 이는 MNIST 같은 다른 데이터집합으로 학습해도 이와 유사한 결과를 보여준다.

생성된 화질의 객관적 지표로 인셉션 스코어(inception score)[7]가 널리 이용되는데, Fig. 3과 Table 1은 학습회수에 따른 각 알고리즘의 인셉션 스코어를 각각 그래프와 정량적 수치로 보여준다. 그래프를 통해 알 수 있는 바와 같이, WGAN을 제외한 알고리

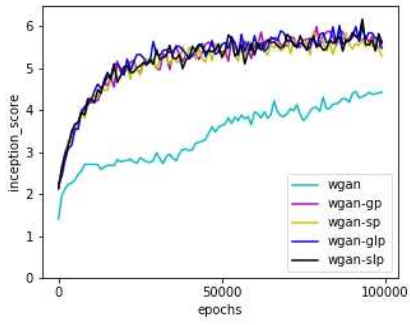

Fig. 3. CIFAR-10 Inception scores over generator iterations.

Table 1. Inception scores for CIFAR-10.

Method	Score
WGAN	4.4236
WGAN-GP	5.6518
WGAN-SP	5.6727
WGAN-GLP	5.8231
WGAN-SLP	5.8393

즘들은 우열을 구별할 수 없을 정도의 지표를 보여주고 있으나 Fig. 2의 결과와 함께 분석해 보면, 수렴속도나 화질 면에서 근소한 차이지만 WGAN-GLP와 WGAN-SLP가 보다 우수한 결과를 보여 주고 있다.

Fig. 4는 CIFAR-10 데이터집합에 대하여 100,000회의 학습과정 후에 각 모델의 생성기가 출력시킨 이미지를 제시한다. 각 모델 모두 안정적으로 수렴하고 있고, 각 생성기가 발생시킨 이미지의 화질은 그 차이가 눈에 띌 정도로 크지는 않으나 제안 알고리즘인 WGAN-SLP가 생성한 이미지가 보다 선명하다고 볼 수 있다.

이상의 실험결과로부터 제안 알고리즘은 WGAN이나 WGAN-GP보다 학습의 수렴속도와 생성기가 발생시킨 이미지의 화질 등의 관점에서 성능이 향상됨을 확인할 수 있다. 이를 토대로 분석해보면 모델의 분별기 함수가 립쉬츠 연속을 보다 안정적으로 유지할수록 수렴속도와 더불어 생성 이미지의 화질 향상을 얻을 수 있을 것으로 기대된다.

5. 결 론

본 논문에서는 분별기 함수가 립쉬츠 연속을 안정적으로 유지시키기 위한 알고리즘을 제안하고 다양한 실험을 통하여 수렴속도와 출력 이미지의 화질을 비교분석함으로써 그 성능을 확인하였다. 분별기 함수의 미분크기를 1로 유지하기 위한 강한 제약의 정칙 항인 경사 페널티를 추가하여 1-립쉬츠 연속을 보장하고자 하는 기법이 현재까지 널리 이용되고 있다. 하지만 실험에 의하면, 이보다는 미분 크기가 1 이상일 때에만 적용하는 단면 페널티 기법이 학습횟수가 2,000 이하인 학습초기에 특히 손실함수가 반 이하로 떨어져 이에 해당되는 만큼의 수렴속도가 개선됨을 보여주고 있다. 이와 함께 WGAN-GLP와 WGAN-SLP에서와 같이 알고리즘이 어느 정도 수렴하였을 때 부가적인 페널티를 추가하는 방식을 채택하면 인셉션 스코어가 0.18 정도 더 높아져 화질개선 효과를 얻을 수 있다.

향후에는 학습데이터의 분포를 보다 조밀하게 표본화하는 기법을 개선하여 립쉬츠 연속을 보다 안정적으로 유지하기 위한 연구를 계속 진행할 계획이다.

REFERENCE

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al., "Generative Adversarial Nets," In *Advances in Neural Information Procesing Systems* 27, pp. 2672-2680, 2014.

[2] D.P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," In *Proceedings of the International Conference on Learning Representations(ICLR)*, 2014.

[3] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," *arXiv:1701.07875v3 [stat.ML]*, Dec. 2017.

[4] M. Arjovsky and L. Bottou, "Towards Principled Methods for Training Generative Adversarial Networks," *arXiv:1701.04862v1 [stat, ML]*, Jan. 2017.

[5] A. Radford, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," *arXiv:1511.06434v2 [cs.LG]*, Jan. 2016.

[6] I. Toolstikhin, S. Gelly, O. Bousquet, C. Simon-Gabriel, and B. Scholkopf, "Adagan: Boosting Generative Models," *arXiv e-prints, arXiv:1701.02386*, 2017.

[7] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved Techniques for Training GANs," In *Advances in Neural Information Procesing Systems* 29, pp. 2234-2242, 2016.

[8] J. Adler and S. Luntz, "Banach Wasserstein GAN," *arXiv:1806.06621v2 [cs.CV]*, Jan. 2019.

[9] G. Qi, "Loss-Sensitive Generative Adversarial Networks on Lipschitz Densities," *arXiv:1701.06264v6 [cs.CV]*, Mar. 2018.

[10] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved Training of Wasserstein GANs," *arXiv e-prints, arXiv:1704.00028v3 [cs.LG]*, Dec. 2017.

[11] H.I. Hahn, "Improving the Performance of WGAN Using Stabilization of Lipschitz Continuity of the Discriminator," *Journal of The Institute of Electronics and Information*

Engineers, Vol. 52, No. 2, Feb. 2019.

[12] H. Petzka, A. Fischer, and D. Lukovnicov, "On the Regularization of Wasserstein GANs," *arXiv:1709.08894v2 [stat.ML]*, Sep. 2017.

[13] R.L. Dobrushin, "Prescribing a System of Random Variables by Conditional Distributions," *Theory of Probability and its Applications*, No. 15, pp. 458-486, 1970.

[14] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. Third Ed. McGrow-Hill, 1991.

[15] C. Villani, *Optimal Transport: Old and New*, Vol. 338, Springer Science & Business Media, 2008.

[16] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," *Technical Report TR-2009*, University of Toronto, 2009.

한 희 일

1980년 3월~1984년 2월 서울대학교 제어계측 공학과 공학사

1984년 3월~1986년 2월 서울대학교 제어계측 공학과 공학석사

1992년 8월~1995년 12월 University of Arizona 전기및 컴퓨터 공학과 공학박사

1987년 1월~1998년 3월 한국전자통신연구원, 선임연구원

1998년 3월~현재 한국외국어대학교 정보통신공학과 교수

관심분야 : 신호처리, 영상처리, 컴퓨터비전, 딥러닝, 미분기하 및 토폴로지