Wasserstein GAN
MAS480 Final Report

20200130 Yujun Kim

June 2023

1 Introduction

GAN은 사실과 유사한 Distribution을 얻기 위해 사용하는 기계학습의 framework이다. 이미지 생성을 예시로 들자. 여러 사람의 얼굴을 찍은 사진 데이터가 있을 때, 이 데이터는 어떤 분포로 부터 얻었다고 가정할 수 있다. 그 분포를 알게 되었을 때, 새로운 데이터도 그 분포에서 sampling을 하여 얻게 된다면 실제 사람 얼굴과 유사한 사진을 얻을 수 있을 것이다. 그렇다면 여러 개의 training data를 바탕으로 분포를 어떻게 학습하는 것일까.

각 데이터들이 $x_i, i = 1, \cdots m$가 분포 parametric distribution P_θ를 따른다고 했을 때, sample한 데이터가 x_i들이 될 확률의 log를 취한 값은 다음과 같다.

$$\frac{1}{m}\sum_{i=1}^{m} \log P_\theta(x_i)$$

이 값을 최대화 하는 θ에 의해 만들어진 분포가 원래 분포와 같다고 생각하는 것이 가장 합리적이다. 만약 true distribution이 P_r이라고 한다면, 위 값을 키우는 것과 KL divergence $KL(P_r||P_\theta)$를 줄이는 것이 m이 커질때 asymtotic하게 같아진다.

P_r을 직접 찾는 것 대신 prior distribution $p(z)$를 가지는 random variable Z를 이용하여, $g_\theta(Z)$가 데이터들을 생성한다고 생각하자. 이 논문에서는 두 분포 간의 차이를 측정하는 다양한 기준 $\rho(P_\theta, P_r)$를 비교하고, 왜 EM distance를 사용하는 것이 유리한지, 그리고 EM distance를 줄이기 위한 Algorithm을 어떻게 짜야 하는지를 이야기 하고 있다. 그리고 마지막으로 실험적으로 새로운 알고리즘이 어떤 이점들을 가지는지 증명하고있다.

2 Backgrounds

두 distribution간의 차이를 측정하는 기준 ρ로 Total variation, KL divergence, JS divergence, , EM(Wasserstein) distance 등이 있다. 논문의 첫번째 theorem은 $W(P_r, P_\theta)$가 theta에 대해 continuous everywhere, differentiable almost everywhere이라는 것이 이야기해주고, 따라서 gradient based algorithm을 사용해서 W를 줄이는 θ를 찾을 수 있다. 논문의 두번째 theorem은 EM distance W와 다른 distance/divergence간의 비교를 할 수 있는 정리다. 이 정리는 KL divergence의 0으로의 convergence가 JS divergence의 0으로의 convergence를 imply하고, JS divergence의 0으로의 convergence가 EM distance의 0으로의 convergence를 imply함을 이야기한다.

그렇다면, real distribution과의 EM distance를 줄이는 방향으로 parameter θ를 조정해 나가는 알고리즘을 생각해 볼 수 있을 것이다. EM distance의 0으로의 convergence가 다른 조건들보다 더 강력하기 때문에 이렇게 얻은 parameter θ는 더 좋은 성질들을 가질 것으로 기대해 볼 수 있다. 한편, 무조건 수렴이 어려운 distance 혹은 divergence를 사용하는 것은 practical한 측면에서서 좋지 않다.

예를들어 zero-one distance(혹은 discrete metric)를 생각해보자. 이 distance는 두 input 분포가 같으면 0 아니면 1을 거리로 준다. 이 distance d에서 $d(P_r, P_{\theta_n})$가 0으로 수렴하기 위해서는 θ_n이

$$\theta_1, \theta_2, \cdots, \theta_n, r, r, \cdots$$

와 같은 수열이 되어야 한다. 즉 eventually real distribution으로 stationary해지는 수열만이 zero-one distance에서 converge하게 된다. 우리의 목표는 r에 근접한 θ를 찾는 것이기 때문에 당연히 이런 수열을 찾을 수 있으면 좋을 것이다. 하지만, zero-one distance의 정보만으로는 gradient를 취해 θ를 어떻게 update해 나가야 할지 정할 수 없다. 반면 Wasserstein distance는 θ에 대한 미분 가능성을 첫번째 정리에서 이야기 해주기 때문에 gradient based algorithm으로 θ를 찾을 수 있다.

3 Wasserstein GAN

이제 이런 이론적 배경을 바탕으로 EM distance의 gradient를 어떻게 계산할 것이며, Iterative algorithm이 어떻게 될지를 알아보자. 우선 EM distance의 정의는 infimum을 포함하고 있어 gradient를 바로 계산하는 것이 까다롭다. 이 문제를 해결하기 위해서, 우선 Kantorovich-Rubinstein duality를 이용해서

$$W(P_r, P_\theta) = \inf_{\gamma \in \Pi(P_2, P_\theta)} \mathbf{E}_{(x,y)\sim\gamma}[||x - y||]$$

로 정의된 EM distance를

$$W(P_r, P_\theta) = \sup_{||f||_L \leq 1} \mathbf{E}_{x\sim P_r}[f(x)] - \mathbf{E}_{x\sim P_\theta}[f(x)] \tag{1}$$

로 바꿔준다. 여기서 supremum은 1-Lipschitz한 모든 function f에 대해 취해졌다. 실제 알고리즘 적으로 계산을 할 때 모든 1-Lipschitz한 function에 대해 supremum을 취하는 것은 불가능하기 때문에, f를 w로 parametrized된 function f_w이라고 생각하고 $w \in \mathcal{W}$내에서 w값을 잘 찾아 supremum을 attain하는 f_w를 찾고자 한다. 여기서 $\mathcal{W}$는 어떤 compact set으로 정해 $w \in \mathcal{W}$일 때 f_w가 $K - Lipschitz$인 상수 K가 존재하도록 한다. 1-Lipschitz조건이 $K - Lipschitz$로 바뀌게 되면 supremum을 취하려고 했던 값이 상수배 차이가 날 뿐이다. 다만, $\mathcal{F}_\mathcal{W} := \{f_w\}_{w\in\mathcal{W}}$이 $K - Lipschitz$인 function을 모두 포함하는 것은 아닌데, 이 논문에서는 $\mathcal{F}_\mathcal{W}$에서 (1)의 supremum이 attain됨을 가정하고 넘어간다.

이렇게 Practical하게는 f를 w로 parametrized되었다고 생각하고 w를 찾지만, 이론적으로도 supremum을 attain하는 어떤 f가 존재함을 세번째 정리가 말해준다. 다만 이 f는 $\mathcal{F}_\mathcal{W}$에 포함되지 않을 수 있다. 3번째 정리는 gradient based iterative algorithm을 제시하기위한 핵심적인 정리로, supremum을 attain하는 f(이 f를 critic이라고 한다)를 찾으면, 그 f를 이용해서 EM distance의 gradient를 다음의 공식으로 쉽게 계산할 수 있음을 말해준다.

$$\nabla_\theta W(P_r, P_\theta) = -\mathbf{E}_{z\sim p(z)}[\nabla_\theta f(g_\theta(z))]$$

이 식은 EM distance의 gradient를 supremum없이 계산할 수 있도록 해준다. WGAN의 algorithm은 매 θ를 update하기 위한 iteration마다 critic w에 대한 f_w를 찾고, 이를 이용해 θ를 gradient descent방향으로 update해주게 된다. 공식에서 평균이 들어간 부분들은 평균의 non-biased estimator인 sample mean을 이용해 계산한다.

4 Conclusion and Interpretation

GAN의 고질적인 문제는 학습이 완전히 망가지는 mode collapsing이 일어나곤 한다는 것이다. WGAN은 기존 GAN과 다르게 Variable 수를 줄여도 실험적으로 mode collpasing이 일어나지 않는 것을 확인하였다. 이 논문에서 인상깊었던 점은 infimum으로 정의된 함수의 gradient를 계산하기 위해 사용된 수학적 테크닉(duality, theorem3) 및 이론적 배경을 바탕으로 한 Algorithm의 구현이었다. Expectation의 근사로 sample mean을 사용하고, discriminator혹은 critic을 parametrized된 함수로 근사하는 등의 작업이 신선했다.