école normale supérieure paris–saclay

Optimal Transport applied to Generative Adversarial Networks

Deep Learning Project

DÉMARRE Camille - NAÏT S. Thiziri

Master 2 Student

MVA

ENS Paris-Saclay

January 24, 2021

école normale supérieure paris–saclay CentraleSupélec École des Ponts ENSAE ParisTech ENSTA MINES PARIS ParisTech l'X École Polytechnique TELECOM ParisTech Université Paris Descartes DAUPHINE UNIVERSITÉ PARIS

Outline

1. Introduction
2. Principle of the Generative Adversarial Network GAN
3. Wasserstein GAN
4. Difference between GAN and WGAN
5. Implementation
6. Some results
7. Conclusion

Introduction

- **Problem:** Lack of data in some fields (medicine, genetics, ...) for training

  $\longrightarrow$ Neuroimaging studies in 2017 and 2018 with median sample size = 12
- **Solution:** Generation of new data with same distribution

Principle of GAN (1)

Adversarial : generator and discriminator **competing** with each other

- **Generator G** turns noise *z* into an imitation of the data to fool the discriminator
- **Discriminator D** tries to identify real instances from fake generated samples

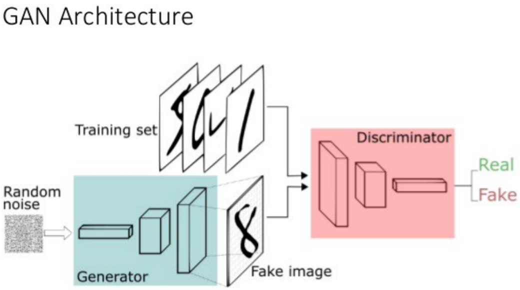


Figure: Principle of GAN

Principle of GAN (2)

The objective is dual :

$$\min_G \max_D \mathbb{E}_{x\sim p} \log D(x) - \mathbb{E}_{z\sim\mathcal{N}(0,\mathcal{I})} \log(1 - D(G(z))$$

which is approximated by :

$$\min_G \max_D \frac{1}{|\mathcal{T}|} \sum_{x\in\mathcal{T}} \log D(x) + \frac{1}{N_z} \sum_{z\sim\mathcal{N}(0,\mathcal{I})} \log(1 - D(G(z)))$$

$\longrightarrow$ Roughly equals to the **Jensen-Shannon divergence**

Wasserstein GAN

- **Different metric :** 1 - Wasserstein distance between two probability measures α, β

$$\mathcal{W}_1(\alpha, \beta) := \min_{\pi \in \Pi(\alpha,\beta)} \mathbb{E}_{(x,y)\sim\pi}(||x - y||)$$

Kantorovich-Rubinstein duality gives :

$$\mathcal{W}_1(\alpha, \beta) = \sup_{||f||_L \leqslant 1} \mathbb{E}_{x\sim\alpha}[f(x)] - \mathbb{E}_{x\sim\beta}[f(x)] \quad (1)$$

Wasserstein GAN (2)

The objective is dual :

$$\min_{\theta} \mathcal{W}_1(p, p_\theta) = \min_{\theta} \max_{||f||_L \leqslant 1} \mathbb{E}_{x\sim p}(f(x)) - \mathbb{E}_{y\sim p_\theta}(f(y))$$

which is approximated by :

$$\min_{\theta} \max_{w s.t. ||f_w||_L \leqslant 1} \frac{1}{|\mathcal{T}|} \sum_{x \in \mathcal{T}} f_w(x) - \frac{1}{N_z} \sum_{z \sim \mathcal{N}(0,\mathcal{I})} f_w(g_\theta(z))$$

f is called the critic

Difference between GAN and WGAN

- GAN : Jensen-Shannon divergence

  $\longrightarrow$ Loss discontinuity

  $\longrightarrow$ Training instability

- WGAN : Wasserstein distance

  $\longrightarrow$ Loss smoothness

  $\longrightarrow$ Improved Training stability

WGAN - Lipschitz constraint

Problem: How to enforce $||f_w||_L \leqslant 1$?

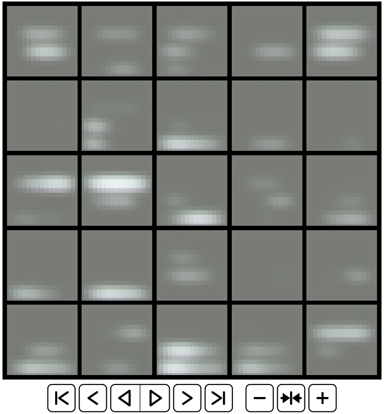
- **Solution :** Weight clipping

  $\longrightarrow$ Naive brute force

- **Improved solution :** Gradient Penalty term in the objective :

$$\lambda_{gp} \frac{1}{T} \sum_{x \in T} (||\nabla f_w(x)|| - 1)^2$$

Embedded Animation



Implementation

Algorithm 1 WGAN Pseudo Code Algorithm

for *iter* to N_{epochs} **do**
  for k_{critic} steps **do**
    Sample minibatch of m samples $\{z^{(1)}, \ldots, z^{(m)}\}$;
    Sample minibatch of m samples $\{x^{(1)}, \ldots, x^{(m)}\}$;
    Update parameters w for the critic f_w by stochastic gradient ascent
    $\nabla_\omega = \frac{\partial}{\partial w}\left[\frac{1}{m}\sum_i \left(f_w(x^{(i)}) - f_w(g_\theta(z^{(i)}))\right)\right];$
    $w \leftarrow \text{clip}(w, -c, c);$
  end for
  Sample minibatch of m samples $\{z^{(1)}, \ldots, z^{(m)}\}$;
  Update parameters θ for the generator g_θ by stochastic gradient descent
  $\nabla_\theta = \frac{\partial}{\partial \theta}\left[\frac{1}{m}\sum_i f_w(x^i) - \frac{1}{m}\sum_i f_w(g_\theta(z^i))\right];$
end for

Some results

MNIST dataset

Standard GAN

WGAN

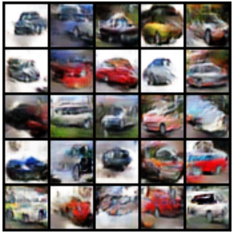
WGAN-GP

Some results

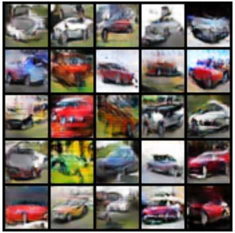
CIFAR dataset

Standard GAN

WGAN



WGAN-GP



Conclusion

- GANs: Promising results but instability, mode collapse
- WGANs : Introduction of Optimal Transport with ML
- To be studied: Sample quality metric, hyper parameter tuning

école normale supérieure paris-saclay

References

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu,D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks

[2] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan

[3] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved training of wasserstein gans

Thank You

for Listening.