OPTIMAL TRANSPORT GAN

Alokendu Mazumder

SPECTRUM Lab, Dept. of EE, IISc, Bangalore (PhD, 1^{st} year)

ABSTRACT

The integral loss function of GAN being intractable, is replaced by point estimates or approximated. Here we propose a new metric to compute distance between data and generator distribution, called mini-batch energy distance.

***Index Terms*—** Generative models, Loss function, metric

1. INTRODUCTION

To improve stability of GAN and get rid of other problems like mode collapse, WGAN [1] was proposed. As Wasserstein distance is intractable (in WGAN), OTGAN [4] was proposed which used a smooth version of Wasserstein distance over mini-batches of real data and generated data.

2. TECHNICAL DETAILS

OTGAN leverages the concept mini-batches used in deep learning. It minimizes the mini-batch energy distance between mini-batches of data and generator distribution.

2.1. Part 1: Mini-batch Sinkhorn Distance

The mini-batch version of sinkhorn distance [2] can be written as:

$$W_c(X,Y) = \underset{M \in m}{\operatorname{argmin}} Tr[MC^T] \tag{1}$$

Here, **M** is a matrix of (k,k) size corresponds to set of all joint distributions with marginals as data distribution and generator distribution. k is the size of mini-batch. **C** is the transport cost matrix of the cost function (here **cross-entropy** is used) used with same dimension, where elements of **C**, c_{ij} is the cost of transporting $\mathbf{x}_i$ in mini-batch **X** to $\mathbf{y}_j$ in mini-batch **Y**. All row and columns of **C** sums to one and has sufficient entropy.

2.2. Part 2: Cramer's GAN

Equation (1), being tractable allows a very stable training for GAN. But the gradients of above equation, are not unbiased estimator of real sinkhorn distance, hence to counter this, Cramer's distance [3] (Energy distance) was introduced:

$$D_{ED} = \sqrt{2E[||\mathbf{x}-\mathbf{y}||] - E[||\mathbf{x}-\mathbf{x}'||] - E[||\mathbf{y}-\mathbf{y}'||]} \tag{2}$$

Email: alokendum@iisc.ac.in, SR No: 04-03-00-10-12-21-1-20134

Where **x**, **x**' and **y**, **y**' are independent samples from real data distribution and generator distribution respectively.

2.3. OT-GAN

Equation (2) is not a valid metric over probability distribution, but this distance is indeed valid over mini-batches, called the mini-batch energy distance, defined as:

$$D_{MED} = \sqrt{2E[W_c(\mathbf{X},\mathbf{Y})] - E[W_c(\mathbf{X},\mathbf{X}')] - E[W_c(\mathbf{Y},\mathbf{Y}')]} \tag{3}$$

Where **X**, **X**' and **Y**, **Y**' are independent samples from real data distribution and generator distribution respectively. Hence, unlike equation (2), equation (3) is a valid metric by leveraging the energy distance in generic form. Hence, it makes OT-GAN highly stable.

3. RESULTS

The model is trained over MNIST dataset with 60,000 samples. Adam is used with 10^{-3} learning rate. It is run for 7 epochs. Below are few generated images: Due to limitation

Fig. 1. Results after 5 epoch (L) & after 2 epoch (R)

of hardware, I was not able to run it for long and hence results are not very visually attractive. Even in real paper, authors have used 8 GPU's and trained OTGAN for several days.

4. CONTRIBUTION

The codes are re-implemented by myself. In original paper, cosine similarity was used as cost function, here I have used cross-entropy.

5. RESOURCES

The paper can be found **here**.

6. REFERENCES

[1] M. Arjovsky, S. Chintala and L. Bottou "Wasserstein GAN," International conference on machine learning, 2017.

[2] A. Genevay, G. Peyre and M. Cuturi "Learning generative models with sinkhorn divergence," In proceedings of the 21^{st} International conference on artificial intelligence and statistics, 2018.

[3] M. G. Bellemare, I. Danihelka, W. Dabney, S. Mohamed, B. Lakshminarayanan, S. Hoyer and R. Munos "The cramer distance as a solution to biased wasserstein gradients," In arXiv 1705.10743, 2017.

[4] T. Salimans, H. Zhang, A. Radford and D. Metaxas "Improving GANs Using Optimal Transport," In International Conference on Learning Representations, 2018.