About WGAN

Gaspard Ulysse Fragnière

August 2022

1 Introduction

Arjovsky et al. (2017) introduce a new type of Generative Adversial Network, namely the Wasserstein GAN (WGAN). The claim is that WGAN improves the stability in learning and get rid of typical problem of the traditional GAN approach such as Mode Collapse (TODO: look into that).

More specifically, Arjovsky et al. (2017) provides the following insights:

TODO: Here needs to state the problem that the Earth Mover distance (distance between probability) try to solve.

- Analyses how the Earth-Mover (EM) distance, also known as Wasserstein distance behaves compared to other distance between probability distribution (e.g. Kullback-Leibler distance)
- Define a GAN that minimizes the an approximation of the EM distance, namely the WGAN.
- Show that unlike traditional GANs, WGANs do not need to maintain a balance when training the discriminator and generator. Indeed in regular GAN approach, it was crucial to avoid the discriminator to become too good before the generator, since this would prevent the generator to learn any distribution.

2 The Earth-Mover or Wasserstein distance

TODO: Understand and explain loss function structure + Check if the loss function used is the same as the one used in the original Wasserstein paper.

The goal of WGAN, remains the same as GAN, namely approximation of of the probability distribution P_r of some data by distribution P_θ. Typically a family of distribution $(P_\theta)_{\theta \in \mathbb{R}^d}$. For this reason, it is necessary to have metrics to quantify distance between two probability distributions P_r and P_m (e.g. Kullback-Leibler divergence, Jensen-Shannon divergence, ...). In WGAN, the distance used is the Earth-Mover (EM) distance or Wasserstein-1, defined as

$$W(\mathbb{P}_r, \mathbb{P}_m) = \inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_m)} \mathbb{E}_{(x,y)\sim\gamma}[||x-y||] \tag{1}$$

Where $\Pi(\mathbb{P}_r, \mathbb{P}_m)$ is the set of all joint distribution $\gamma(x, y)$ whose marginals are respectively $\mathbb{P}_r$ and $\mathbb{P}_m$. Informally, $\gamma(x, y)$ shows how much "mass" must be carried to trasnform $\mathbb{P}_r$ into $\mathbb{P}_m$. The EM distance is then "the cost" of the optimal "transport".

Arjovsky et al. (2017) shows that this distance converges for some simple probablity disitrbution, where other common distances (e.g. Kullback-Leibler divergence, Jensen-Shannon divergence) do not. Moreover, in this paper it is shwon that the EM distances has nice properties (**TODO**: elaborate ???) conmpared to other distances.

Unfortunately the EM distance is intractable, due to the infinum part in its equation, but the Kantorovich-Rubinstein duality tells us that

$$W(\mathbb{P}_r, \mathbb{P}_\theta) = \sup_{||f||_L \leq 1} \mathbb{E}_{x \sim \mathbb{P}_r}[f(x)] - \mathbb{E}_{x \sim \mathbb{P}_\theta}[f(x)] \tag{2}$$

Where the supremum is over 1-Lipschitz function. It is important to note that if we replace $||f||_L \leq 1$ by $||f||_L \leq K$, i.e. consider also the K-Lipschitz function, then we obtain $K \cdot W(\mathbb{P}_r, \mathbb{P}_\theta)$. Hence, for a family of functions $\{f_w\}_{w \in \mathcal{W}}$ (all functions being K-Lipschitz), we can consider solving the optimization problem:

$$\max_{w \in (W)} \mathbb{E}_{x \sim \mathbb{P}_r}[f_w(x)] - \mathbb{E}_{z \sim p(z)}[f_w(g_\theta(z))] \tag{3}$$

Note: we can approximate the solution of the above mentionned problem with a Neural Network with weights $\mathcal{W}$. $\mathcal{W}$ need to be compact to assure that the function f_w are K-lpischitz. Therefore in order to have $\mathcal{W}$ being compact, we can simply clip the weights in a small box (e.g. $[-0.01, 0.01]^l$)

3 Changes between GAN and WGAN

Implementation of a WGAN requires a few changes from implementation of a regular GAN, i.e.

- Use a linear activation function in the output layer of the critic model (instead of sigmoid).
- Use Wasserstein loss to train the critic and generator models that promote larger difference between scores for real and generated images.
- Constrain critic model weights to a limited range after each mini batch update (e.g. [-0.01,0.01]). As seen above, this is to ensure, that the function estimated in the for approximting the Wasserstein distance are K-lipschitz, a necessary condition.
- Update the critic model more times than the generator each iteration (e.g. 5). Contrary as in a GAN, this is not a issue. Indeed, switching to the EM distance, allow for more stability when training the two networks in the WGAN. Moreover, the fact that the EM distance is continuous and differentiable means that we should train the critic until optimality.
- Use the RMSProp version of gradient descent with small learning rate and no momentum (e.g. 0.00005). (quote: "... we report that WGAN training becomes unstable at times when one uses a momentum based optimizer such as Adam [...] We therefore switched to RMSProp ...")

4 Implementation of EM distance or Wasserstein-1:

TODO: How to implement a Wasserstein loss:

- **Goal:** increase the gap between the scores for real and generated images

- **Critic Loss**: difference between average critic score on real images and average critic score on fake images
- **Generator Loss** the negation of the average critic score on fake images
-

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.