Wasserstein Auto-Encoders

Anastasia Rogachevskaya
Faculty of Mathematics
Higher School of Economics
avrogachevskaya@edu.hse.ru

Mark Garnitskiy
Faculty of Computer Science
Higher School of Economics
markmitt@yandex.ru

Abstract

The report covers the implementation of the Wasserstein Auto-Encoder proposed by Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Scholkopf. It is a new model to generate data distribution and can be considered as an modified extension of the Variational Auto-Encoder. The main difference is the shift from f-divergences used in the regularizer to another family of divergences between probability distributions known as optimal transport distances. In this work, we present our implementation of proposed model and try to repeat experiments from the original paper.

1 Introduction

In last years, the Variational autoencoder became a powerful tool for generative modeling due to its fundamental property that separates them from standard autoencoders: latent spaces of encoded vectors are continuous, allowing random sampling and interpolation. However, there are certain problems with quality of reconstructed data since a model is known to generate blurry images. In this case, generative adversarial networks (GANs) are able to produce more impressive samples with high visual quality but still stay harder to train and suffer from 'mode collapse' problem.
Authors propose new approach from the optimal transport point of view which leads to minimize $W_c(P_X, P_G)$ for any cost function c, the true data distribution P_X and a latent variable model P_G. It should be mentioned that the objective function of WAEs is consisted of two parts: c - reconstruction cost and a regularizer $D_Z(P_Z, Q_Z)$ to penalize for the difference between two probability distributions. In this paper two types of a regularizer are discovered: one is based on adversarial training and the other uses the maximum mean discrepancy. Finally, there were presented experimental results of the WAE's implementation with two different regularizers (WAE-GAN and WAE-MMD) on MNIST and CelebA datasets.

2 Method description

In the original paper the theory of optimal transport is used to construct a different regularizer: one that matches the prior with the aggregated posterior – the average posterior over the training data. Authors define a latent variable model P_G: a sample Z is taken from the prior distribution P_Z and is mapped to the image $X = G(Z)$ where $G : \mathcal{Z} \to \mathcal{X}$. The goal is to find model distribution P_G which is similar to data distribution P_X. Note that only generative models with a determistical map G are considered. This generative model can written as the following equation for densities:

$$p_G(x) = \int_{\mathcal{Z}} p_G(x|z)p_z(z)dz$$

In order to define a new family of divergences we should formulate the optimal transport problem (according to Kantorovich):

$$W_c(P_X, P_G) := \inf_{\Gamma \in \mathcal{P}(X \sim P_X, Y \sim P_G)} \mathbb{E}_{(X,Y)\sim\Gamma}[c(X, Y)],$$

where $c(x, y)$ is any measurable cost function and $\mathcal{P}(X \sim P_X, Y \sim P_G)$ is a set of all joint distributions of measures (X, Y) such that their marginals are equal to P_X and P_Y respectively. We want to find the joint that gives the minimum cost and it leads to the Wasserstein distance is in itself an optimization problem. Note that according to the formulation $W_c(P_X, P_G)$ is needed to be optimized over all couplings between two random variables X and Y from space $\mathcal{X}$ that X distributed according to P_X and P_G respectively. However, authors used the following theorem to avoid this:

Theorem 1 *(Bousquet et al. (2017)) For P_G as defined above with deterministic $P_G(X|Z)$ and any function $G : \mathcal{Z} \to \mathcal{X}$*

$$\inf_{\Gamma \in \mathcal{P}} \mathbb{E}_{(X,Y)\sim\Gamma}[c(X, Y)] = \inf_{Q:Q_Z=P_Z} \mathbb{E}_{P_X}\mathbb{E}_{Q(Z|X)}[c(X, G(Z))],$$

where Q_Z is the marginal distribution of Z when $X \sim P_X$ and $Z \sim Q(Z|X)$.

Due to this fact it is possible to optimize over probabilistic encoders $Q(Z|X)$ from $\mathcal{Q}$. In other words, authors ask the marginal distribution of encoded images (the aggregated posterior) to match the prior. It allows to construct the objective function for WAEs:

$$D_{WAE}(P_X, P_G) := \inf_{Q(Z|X)\in\mathcal{Q}} \mathbb{E}_{P_X}\mathbb{E}_{Q(Z|X)}[c(X, G(Z))] + \lambda \cdot D_Z(Q_Z, P_Z)$$

While the first term serves as the reconstruction loss of the autoencoder, the last term is the regularizer of the objective with added hyperparameter $\lambda > 0$ to penalize the difference between probability distributions Q_Z and P_Z. As it was mentioned above, there were proposed two different algorithms to define D_Z.

2.1 WAE-GAN

According to this approach, Jensen-Shannon divergence (denoted by D_{JS}) is used as D_Z. Besides, authors introduced a discriminator D to differ samples from the prior P_Z and ones from encoding of real data points Q_Z. The WAE-GAN model optimizes the same adversarial objective as the regular GAN, except that the objective is defined on the latent space instead of the actual input/output space:

$$D_{JS}(P_Z, Q_Z) = \sup_D \mathbb{E}_{Z\sim P_Z}[logD(Z)] + \mathbb{E}_{Z^*\sim Q_Z} log[1 - D(Z^*)]$$

2.2 WAE-MMD

This divergence is based on the maximum mean discrepancy:

$$MMD_k(P_Z, Q_Z) = \| \int_{\mathcal{Z}} k(z, \cdot) dP_Z(z) - \int_{\mathcal{Z}} k(z, \cdot) dQ_Z(z) \|_{\mathcal{H}_k},$$

where $k : \mathcal{Z} \times \mathcal{Z} \to \mathcal{R}$ is a positive-definite reproducing kernel defined on the latent space and $\| \cdot \|$ is a measure (distance) in a reproducing kernel Hilbert space denoted by $\mathcal{H}_k$.

ALGORITHM 1 Wasserstein Auto-Encoder with GAN-based penalty (WAE-GAN).

Require: Regularization coefficient $\lambda > 0$.
Initialize the parameters of the encoder Q_ϕ, decoder G_θ, and latent discriminator D_γ.
while (ϕ, θ) not converged **do**
Sample $\{x_1, \dots, x_n\}$ from the training set
Sample $\{z_1, \dots, z_n\}$ from the prior P_Z
Sample $\tilde{z}_i$ from $Q_\phi(Z|x_i)$ for $i = 1, \dots, n$
Update D_γ by ascending:

$$\frac{\lambda}{n} \sum_{i=1}^{n} \log D_\gamma(z_i) + \log\big(1 - D_\gamma(\tilde{z}_i)\big)$$

Update Q_ϕ and G_θ by descending:

$$\frac{1}{n} \sum_{i=1}^{n} c\big(x_i, G_\theta(\tilde{z}_i)\big) - \lambda \cdot \log D_\gamma(\tilde{z}_i)$$

end while

ALGORITHM 2 Wasserstein Auto-Encoder with MMD-based penalty (WAE-MMD).

Require: Regularization coefficient $\lambda > 0$, characteristic positive-definite kernel k.
Initialize the parameters of the encoder Q_ϕ, decoder G_θ, and latent discriminator D_γ.
while (ϕ, θ) not converged **do**
Sample $\{x_1, \dots, x_n\}$ from the training set
Sample $\{z_1, \dots, z_n\}$ from the prior P_Z
Sample $\tilde{z}_i$ from $Q_\phi(Z|x_i)$ for $i = 1, \dots, n$
Update Q_ϕ and G_θ by descending:

$$\frac{1}{n} \sum_{i=1}^{n} c\big(x_i, G_\theta(\tilde{z}_i)\big) + \frac{\lambda}{n(n-1)} \sum_{\ell \neq j} k(z_\ell, z_j) + \frac{\lambda}{n(n-1)} \sum_{\ell \neq j} k(\tilde{z}_\ell, \tilde{z}_j) - \frac{2\lambda}{n^2} \sum_{\ell, j} k(z_\ell, \tilde{z}_j)$$

end while

3 Experiments

3.1 MNIST

The MNIST dataset consists of 70k black-whine images of numbers. We run our experiments exactly in the same way it was proposed in the original paper. Authors trained WAE-MMD and WAE-GAN using $c(x, y) = \| x - y \|^2$ as a cost function, the Gaussian prior P_Z and Euclidean latent space $\mathcal{Z} \in \mathcal{R}^{d_z}$ with $d_z = 8$. As it was mentioned in the original paper, Adam with $\beta_1 = 0.5$ was used for optimization. The architecture of a model was built in this way: convolutional deep neural network architectures for encoder mapping and decoder similar to the DCGAN with batch normalization. Convolutional layers have filters of 4×4 size, also the encoder has convolutions with strides are equal to 2 and SAME padding. In particular, for WAE-GAN we used a composition of fully connected layers with $ReLu$ layers as a discriminator D. WAE-MMD was trained with the inverse multiquadratics kernel $k(x, y) = C/(C + \| x - y \|^2)$. Besides, λ for D_{WAE} was set to 10 according to the paper. We should notice that there are no results to provide the comparison of WAE-MMD and WAE-GAN on MNIST in the original paper.

Encoder architecture:

$$\begin{aligned} x \in \mathcal{R}^{28 \times 28} &\to \mathrm{Conv}_{128} \to \mathrm{BN} \to \mathrm{ReLU} \\ &\to \mathrm{Conv}_{256} \to \mathrm{BN} \to \mathrm{ReLU} \\ &\to \mathrm{Conv}_{512} \to \mathrm{BN} \to \mathrm{ReLU} \\ &\to \mathrm{Conv}_{1024} \to \mathrm{BN} \to \mathrm{ReLU} \to \mathrm{FC}_8 \end{aligned}$$

Decoder architecture:

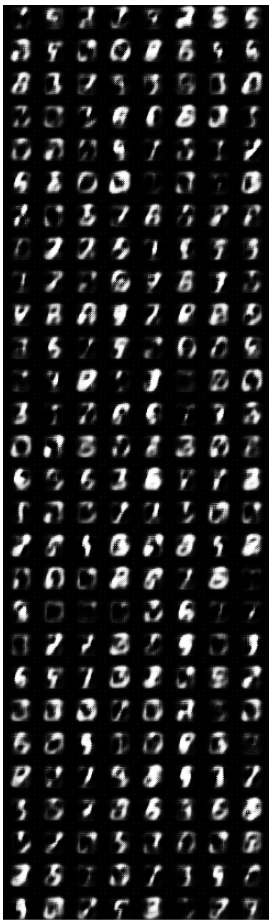
$$\begin{aligned} z \in \mathcal{R}^8 &\to \mathrm{FC}_{7 \times 7 \times 1024} \\ &\to \mathrm{FSConv}_{512} \to \mathrm{BN} \to \mathrm{ReLU} \\ &\to \mathrm{FSConv}_{256} \to \mathrm{BN} \to \mathrm{ReLU} \to \mathrm{FSConv}_1 \end{aligned}$$

Adversary architecture for WAE-GAN:

$$\begin{aligned} z \in \mathcal{R}^8 &\to \mathrm{FC}_{512} \to \mathrm{ReLU} \\ &\to \mathrm{FC}_{512} \to \mathrm{ReLU} \\ &\to \mathrm{FC}_{512} \to \mathrm{ReLU} \\ &\to \mathrm{FC}_{512} \to \mathrm{ReLU} \to \mathrm{FC}_1 \end{aligned}$$

Currently we train[1] only WAE-MMD for 50 epochs since we have several difficulties with WAE-GAN. We set a learning rate for Adam optimizer in encoder-decoder to 0.001. We did not manage to implement WAE-GAN but we got several MNIST images (above one is original image and below one refers to produced by the decoder of WAE-MMD implemention. As we can see, image is very different from the original.

(a) Original



(b) After decoder

References

Ilya Tolstikhin, Sylvain Gelly, Olivier Bousquet, & Bernhard Scholkopf (2018) Wasserstein Auto-Encoders, *In International Conference on Learning Representations, 2018*, URL `https://openreview.net/pdf?id=HkL7n1-0b`.

O. Bousquet, S. Gelly, I. Tolstikhin, C. J. Simon-Gabriel, and B. Scholkopf. From optimal transport to generative modeling: the VEGAN cookbook, 2017.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN, 2017.

[1]The code is available at `https://github.com/179mark/WAE/`