A review of the Wasserstein Auto-Encoders paper

Yves Greatti

December 30, 2018

1 Overview

As mentioned in the OpenReview comments of ICLR 18, one could argue that the paper from Tolstikhin et al. [3], is not being innovative enough. However, it will be unfair as it is a solid paper. The authors show a good familiarity with the literature regarding auto-encoders, GAN and Optimal Transport. They carefully compare their method with the approaches made in each domain. Variational auto-encoders (VAEs) tend to generate blurry images, and generative adversarial networks (GANs) are notoriously hard to train, and suffer from "mode collapse" problem or limit cycle (oscillation around the equilibrium). The paper proposes a new generative model which has close connections to these existing algorithms and addresses their shortcomings. The authors of the paper give an explanation why the samples generated by their model, WAE, are more realistic:

1. In the VAE, $Q(Z|X)$ matches the prior $p(z)$ for each point by averaging , the VAE collapses most dimensions in the latent representation which can yield blurry images.

2. By forcing the constraint $Q_Z(Z) = \mathbb{E}_{X\sim P_X}\left[Q(Z|X)\right] = P_Z(z)$ (see Figure 1), WAE is not averaging and covers the whole set of data points $X \in \mathcal{X}$.

The major contribution of the paper is to appeal to the same idea of a variational lower bound, the marginal log-likehood $E_{P_X}[\log p_G(X)]$, the theoretical framework employed by VAEs, but in the context of optimal transport (OT) (P_X is the data distribution and P_G the model). The model P_G is a two-step procedure, first Z is sampled on a latent space $\mathcal{Z}$ from a distribution $Q_Z(Z) = \mathbb{E}_{X\sim P_X}\left[Q(Z|X)\right]$ and then Z is mapped deterministically to $G(z)$ with $P_G(X|Z=z) = \delta_{G(z)}$. The result is a density of the form:

$$p_G(x) := \int_{\mathcal{Z}} p_G(x|z)p_z(z)dz, \quad \forall x \in \mathcal{X}$$

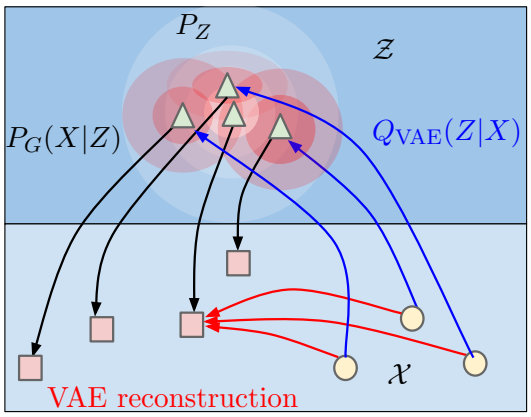
To be able to optimize over deterministic encoders $Q(Z|X)$ instead of optimizing over all couplings between X and Y, the authors of the paper prove and use the theorem:

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

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

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

(a) VAE



(b) WAE

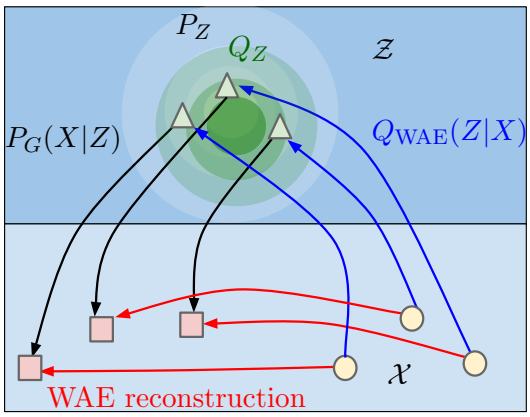


Figure 1

This leads to the WAE objective function:

$$\min_{Q(Z|X)\in\mathcal{Q}} \mathbb{E}_{P_X}\mathbb{E}_{Q(Z|X)}\big[c\big(X, G(Z)\big)\big] + \lambda \cdot \mathcal{D}_Z(Q_Z, P_Z) \tag{1}$$

where $\mathcal{Q}$ is any nonparametric set of encoders, $Q(Z|X) = \int Q(Z|X)P_X(X)dX$, $\mathcal{D}_Z$ is an arbitrary divergence measure between two distributions Q_Z and P_Z, and $\lambda > 0$ is a regularization parameter.

The first reconstruction term ensures that the encoded images can be reconstructed by the decoder. The second regularization term forces the posterior Q_Z to match the prior distribution P_Z for all data points x. Two algorithms are proposed using two different penalties for the term $\mathcal{D}_Z(Q_Z, P_Z)$, one using Jensen penalty equivalent to the one used by Goodfellow et al. [1] , and the other one is *the maximum mean discrepancy* (MMD) allowing to use Kernel methods which is more appropriate for multi-modal distributions.

2 Latent Dimensionality and Random Encoders and Critique

It seems that for the GAN-based penalty, $\mathcal{D}_Z(Q_Z, P_Z)$, the loss is based on the Jensen-Shannon divergence and could be the Earth-Mover distance (W_1, the 1-Wasserstein distance), which provides more stable gradients when the probability distributions, Q_Z and P_Z, have largely disjoint low-dimensional supports. Another restriction made by the authors of the paper is to consider, for $Q(Z|X)$, deterministic encoders.

This restriction is addressed in a second paper from Rubenstein et al. [2], in which experiments are made using random encoders, which map an image $x \in \mathcal{X}$ to a distribution $Q(Z|X = x)$ over the latent space: $Q(Z|X = x) = \delta_{\varphi(x)}$. When there is a large mismatch between the dimensionality of the latent space $d_{\mathcal{Z}}$ and the intrinsic dimensionality of the data distribution $d_{\mathcal{I}}$, i.e. $d_{\mathcal{I}} \ll d_{\mathcal{Z}}$, the deterministic encoder leaves large holes in the latent space on which the decoder is never trained which results in generated samples of poor quality or wrong proportion of generated images (a model trained on MNIST dataset will generate too few samples of some numbers, e.g. for example, not enough 3s and too many 7s).

Usually random encoders do not exhibit such behavior, by filling additional dimensions with noise, making Q_Z accurately matching P_Z preserving the quality of the generated samples. However, the authors observe that random encoders can suffer from a variance collapse mode which

they cannot really explain but in practice they propose to eliminate it by adding, to the loss function, an L_1 regularization term on the log-variances of $Q(Z|X = x)$:

$$\frac{\lambda_p}{N} \sum_{n=1}^{N} \sum_{i=1}^{d_{\mathcal{Z}}} \left| \log \left(\sigma_i^2(x_n) \right) \right|^p$$

where N is the size of the mini-batch, $d_{\mathcal{Z}}$ size of the latent space, and $\lambda_p \geq 0$ is a regularization parameter.

3 Conclusion

After mentioning few limitations of the WAE, other directions of research could be investigated such as:

- Regarding the regularization term, usage of other divergences from the f-divergence class.
- Reduction of number of hyper-parameters (random-encoder WAEs automatically adapting to $d_{\mathcal{Z}}$ thus without the need of an extra hyperparameter)
- A stronger theory unifying the best of GANs and VAEs.

References

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014. URL `http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf`.

[2] Paul K. Rubenstein, Bernhard Schoelkopf, and Ilya Tolstikhin. Wasserstein auto-encoders: Latent dimensionality and random encoders, 2018. URL `https://openreview.net/forum?id=r157GIJvz`.

[3] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=HkL7n1-0b`.