TAMING VAES

BAYESIAN METHODS OF MACHINE LEARNING 2018

A PREPRINT

Anna Kuzina **Denis Prokopenko** **Valentina Shumovskaia**

October 26, 2018

1 Problem

Deep variational auto-encoders (VAEs) have made substantial progress in the last few years on modelling complex data such as images, speech synthesis, molecular discovery, robotics and 3d-scene understanding. The most popular method for training these models is through stochastic amortized variational approximations, which use a variational posterior to construct the evidence lower-bound (ELBO) objective function.

Our main goal is to implement an approach proposed by Danilo J. Rezende and Fabio Viola [1]. The main question is: whether this approach is really better than [2] or [3]? We did experiments with VAE, β-VAE, GECO algorithm, compared them on MNIST and CIFAR-10 data sets. The next goal is to try to improve this approach by combining this method and [4], [5]. Such a model was implemented and compared with the others on MNIST data set.

1.1 VAE and β-VAE

We solve the following optimization problem:

$$L(\theta, \phi, x^i) = -\beta D_{KL}(q_\phi(z|x^{(i)})||p(z)) + \mathbb{E}_{q_\phi(z|x^{(i)})}(\log p_\theta(x^{(i)}|z)), \tag{1}$$

where x is a data, z is a latent space, $p(z)$ is a latent distribution. The specific case $\beta = 1$ is an initial VAE model, q_ϕ is an encoder, p_θ is a decoder.

1.2 GECO

Danilo J. Rezende at al. in [1] offered to take a look at this problem from another point of view, which helps to treat β as a trainable parameter. The idea is problem of ELBO maximization can be reformulated as a conditional minimization problem: find the minimal possible KL-divergence for a given value of a reconstruction loss, with the following Lagrangian:

$$KL(q(z|x)\|p(z)) + \lambda^T \mathbb{E}_{q(z|x)}(Re(x, g(z)) \to min \tag{2}$$

Where reconstruction loss can be computed as:

$$\|x - g(x)\|^2 - \kappa^2$$

As a result, λ is adjusted interactively. At the beginning, when reconstruction is poor, λ increases from iteration to iteration. But as soon as reconstruction becomes equal to predefined tolerance level, λ stats to fall, resulting it increased weight of KL-divergence in the loss. The run of the algorithm is illustrated on 1.

1.3 IWAE

Yuri Burda et al. [4] introduced the improvement of variational autoencodes. Instead of (1) we consider the following variational k-sample importance weighting estimate of the log-likelihood with reparametrization trick:

$$\mathcal{L}_k(x) = \mathbb{E}_{\mathbf{h}_1,\ldots,\mathbf{h}_k \sim q(\mathbf{h}|\mathbf{x})}\left[\log \frac{1}{k}\sum_{i=1}^{k} \frac{p(x, h_i)}{q(h_i|x)}\right] \tag{3}$$

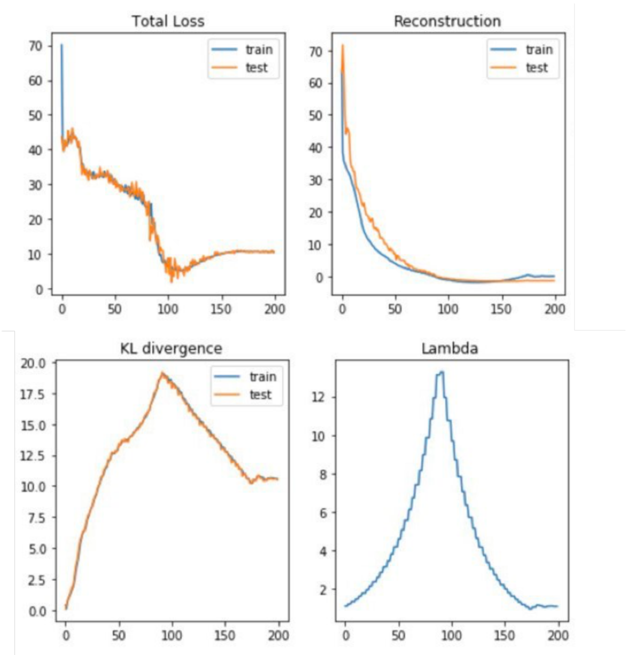

Figure 1: Training of GECO Algorithm on MNIST

When $k = 1$, k-sample importance weighting estimate of the log-likelihood becomes usual VAE estimate log-likelihood. According to the Theorem 1 from [4]:

$$\log p(x) \geq \mathcal{L}_{k+1} \geq \mathcal{L}_k. \tag{4}$$

As a result such estimation increases flexibility to model complex posteriors, which do not fit the VAE modeling assumptions. It may be shown that IWAEs learn richer latent space representations than VAEs, leading to improved test log-likelihood on density estimation benchmarks.

Training procedure is similar to the original, however there appears importance weighting due to differentiation:

$$\nabla_\theta \mathcal{L}_k(\mathbf{x}) = \nabla_\theta \mathbb{E}_{\mathbf{h}_1,\ldots,\mathbf{h}_k \sim q(\mathbf{h}|\mathbf{x})} \log \frac{1}{k} \sum_{i=1}^{k} w_i = \nabla_\theta \mathbb{E}_{\epsilon_1,\ldots,\epsilon_k} \log \frac{1}{k} \sum_{i=1}^{k} w_i(\mathbf{x}, \mathbf{h}(\mathbf{x}, \epsilon_i, \theta)) =$$

$$= \mathbb{E}_{\epsilon_1,\ldots,\epsilon_k} \nabla_\theta \log \frac{1}{k} \sum_{i=1}^{k} w(\mathbf{x}, \mathbf{h}(\mathbf{x}, \epsilon_i, \theta)) = \mathbb{E}_{\epsilon_1,\ldots,\epsilon_k} \sum_{i=1}^{k} \tilde{w}_i \nabla_\theta \log w(\mathbf{x}, \mathbf{h}(\mathbf{x}, \epsilon_i, \theta))$$

$$\tilde{w}_i = \frac{w_i}{\sum_{j=1}^{k} w_j}$$

2 Discussion

2.1 GECO

The idea of the discussed algorithm is quite logical. But when it came to realization, a few problems appeared. At the beginning of the training reconstruction loss it too high, resulting in fast increase of λ. When finally reconstruction converge to some reasonable value, value of λ is already too big and KL diverges converges to a higher values that expected. This peculiarity was not discussed in the paper, but we noticed, that the graphs which they present, actually begin from 800-1000th iteration. So, we assumed, that it would be better to fix λ at some low level for the beginning of the training in order to get proper reconstruction.

2.2 IWAE

Now, move on to IWAE implementation. The GECO algorithm is assumed to be used. However, the initial IWAE approach cannot be applied straightforward in that case. Hence, the IWAE estimation should be splitted into two parts:

Model	Marginal KL	Reconstruction loss
VAE	1.5813	55.8484
β-VAE	10.5938	15.1434
GECO	10.4975	14.6269
GECO + IWAE	1.3640	12.9214

Table 1: MNIST data set

Model	Marginal KL	Reconstruction loss
VAE	31.9662	77.6570
β-VAE	59.4998	57.9653
GECO	383.08495	11.4413

Table 2: CIFAR-10 data set

reconstruction and KL divergence. Chris Cremer et al. introduced the way to split using q_{IW}.

$$q_{IW}(z|x) = \frac{p, z}{\frac{1}{k}\sum_{j=1}^{k} \frac{p(x,z_j}{q(z_j|x)}}$$

Then the considered Lagrangian will have the following form:

$$\mathcal{L}oss_{IW} = \lambda \mathbb{E}_z \log p(x|z) + KL(q_{IW}(z|x)||p(z)) = (\lambda - 1)\mathbb{E}_z \log p(x|z) + \mathcal{L}_k \tag{5}$$

As a result the network may be trained by decreasing (5), which consists of reconstruction error and IWAE log-likelihood.

3 Experiments

3.1 MNIST

All in all, the VAE, $\beta-$VAE, VAE and IWAE trained by GECO were tested on MNIST dataset. The model of variational autoencoder consisted of fully connected encoder: $784 \to 512 \to 256 \to 200$ and decoder: $200 \to 256 \to 512 \to 784$. Tolerance was equal 4, batch size was equal 2000. Initial $\lambda_{VAE} = 1$ with step 100,for β-VAE we fixed $\beta = 0.5$ $\lambda_{IWAE} = 10000$ with step $= 100$. In order to compute proper loss function in case of IWAE, the $k = 5$ samples were drawn from latent space for each original picture.

The GECO algorithm showed better results for VAE and IWAE both compared to standard VAE and $\beta-$ VAE, see fig. 2 and table 1. The generation case showed other useful result. The GECO algorithm applied to VAE outperform the IWAE with GECO and showed the best results, while simple VAEs failed.

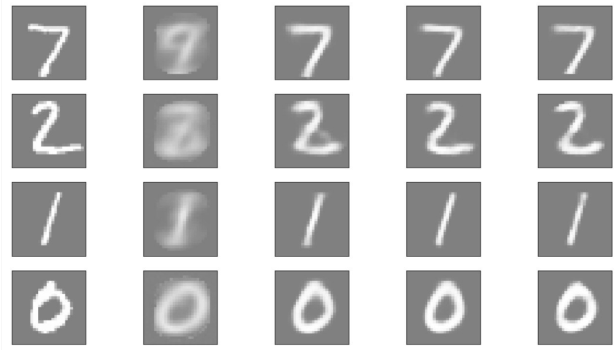

Figure 2: Ground truth and reconstruction results of VAE, β-VAE with $\beta = 0.5$, GECO, GECO + IWAE models trained on MNIST (from left to right).

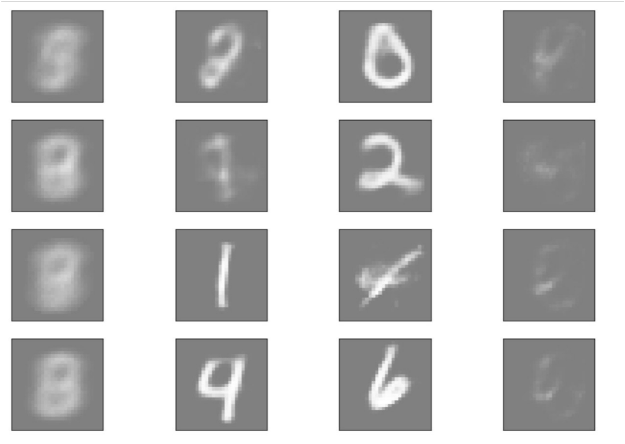

Figure 3: Generation results of VAE, β-VAE with $\beta = 0.5$, GECO, GECO + IWAE models trained on MNIST (from left to right).

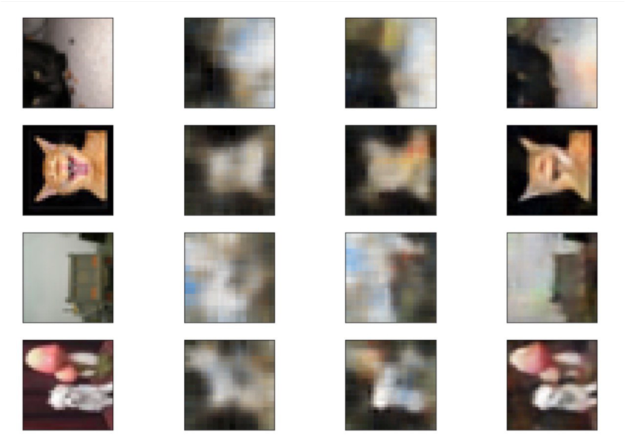

Figure 4: Ground truth and reconstruction results of VAE, β-VAE with $\beta = 0.5$, GECO, GECO + IWAE models trained on CIFAR-10 (from left to right).

3.2 CIFAR10

For the CIFAR10 dataset, simple fully-connected model did not perform well, therefore, we use ConvolutionalDRAW model instead. It consisted of one LSTM modules, where linear layers were substituted by convolutions.

To estimate VAE, β-VAE with $\beta = 0.5$ and GECO-VAE performance, we used hidden dimension of 256 and batch size 300. For GECO we've chosen tolerance of 0.6, which produce nice reconstruction. Initial value of $\lambda = 1$ was adjusted every 100 iterations.

We can see from table 2 and figure 4, that quality of reconstruction achieved by GECO is much better than the one for simple VAE. But at the same time, posterior density of the latent variables is too far from the prior. It is highly likely, that this is a result of the problem, discussed in the previous part: λ grows too fast at the begging, resulting in the extremely complex posterior density, which cannot be simplified further.

References

[1] Rezende Danilo and Viola Fabio. Taming vaes. In *arXiv:1810.00597v1*, 2018.

[2] Kingma Diederik and Welling Max. Auto-encoding variational bayes. *arXiv:1312.6114*, 2013.

[3] Higgins Irina, Matthey Loic, Pal Arka, Burgess Christopher, Glorot Xavier, Botvinick Matthew, Mohamed Shakir, and Lerchner Alexander. β-vae: Learning basic visual concepts with a constrained variational framework. *ICLR 2017*, 2017.

[4] Burda Yuri, Grosse Roger, and Salakhutdinov Ruslan. Importance weighted autoencoders. In *arXiv:1509.00519v4*, 2016.

[5] Cremer Chris, Morris Quaid, and Duvenaud David. Reinterpreting importance-weighted autoencoders. *ICLR 2017*, 2017.