Reproduced paper: Tighter Variational Bounds are Not Necessarily Better

Amine M'Charrak, Vít Růžička, Sangyun Shin, Madhu Vankadari

Department of Computer Science, University of Oxford

{amine.mcharrak,vit.ruzicka,sangyun.shin,madhu.vankadari}@cs.ox.ac.uk

UNIVERSITY OF OXFORD

Introduction

1. Performing efficient approximate inference and learning for directed probabilistic models continuous latent variables and intractable posterior distributions is a challenging problem.
2. Recently, this problem is solved using a generative model which pairs a top-down generative network with a bottom-up recognition network trained to maximize a variational lower bound (ELBO) on the intractable model evidence.
3. It is often assumed that using tighter ELBOs is universally beneficial, at least whenever this does not in turn lead to higher variance gradient estimates
4. In [1] this implicit assumption is questioned by demonstrating that, although using a tighter ELBO is typically beneficial to gradient updates of the generative network, it can be detrimental to updates of the reconstruction network. Illustrated below:

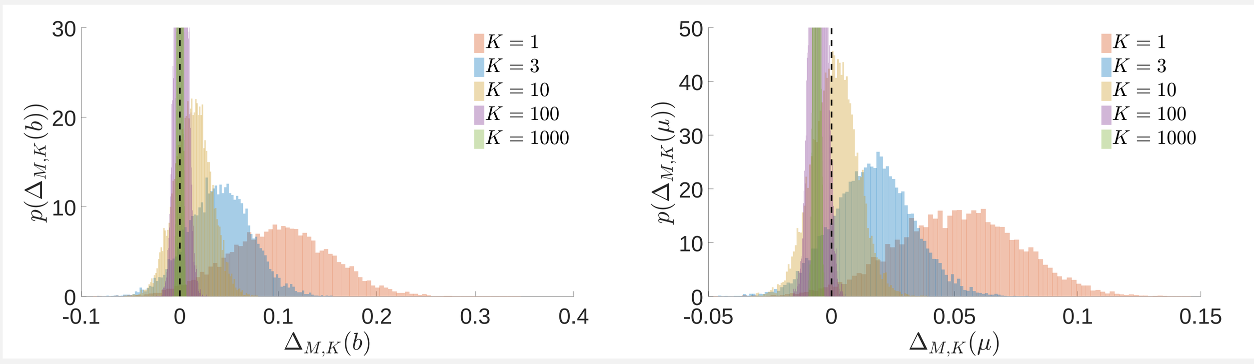

Figure 1:Illustration from [1] which shows the histograms of gradient estimates for the encoder (left) and decoder (right) networks.

Our contributions

1. Reproduction of the original paper and proposal of CIWAE with β as a learnable parameter.
2. Analysis of increasing the number of parameters in the models.
3. Extended evaluation on the Omniglot dataset and exploring the generalization ability across the two datasets.
4. Released an open-source Github repository: `https://github.com/madhubabuv/TightIWAE/`.

Theory foundation and variational autoencoder (VAE) methods

The VAE is a probabilistic model for density estimation and representation learning in continuous latent variable models $p_\theta(x|z)$ with posterior distribution $p_\theta(z|x)$. The distribution of interest arises by marginalization over the latent variables z:

$$p(x) = \int p_\theta(x|z)p_\theta(z)\mathrm{d}z \qquad (1)$$

Theory foundation and variational autoencoder (VAE) methods

To avoid the marginalization costs in Equation (1) we optimize an evidence lower-bound by introducing an inference network $q_\phi(z|x)$ instead, with learnable parameters ϕ. This allows us to cast the intractable inference problem into an optimization by learning $q_\phi(z|x)$ instead of the intractable posterior distribution $p_\theta(z|x)$. The ELBO

$$\log p(x) \geq \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p_\theta(x|z)p_\theta(z)}{q_\phi(z|x)}\right] =: \mathcal{L}_{VAE}(= ELBO) \qquad (2)$$

is maximized by minimizing the gap between the constant log likelihood of the data and the ELBO, quantified by $D_{KL}(q_\phi(z|x)\,||\,p_\theta(z|x))$.

Variational autoencoder methods: MIWAE, CIWAE, and PIWAE The three proposed algorithms are variations of the importance weighted autoencoder (IWAE), which performs a Monte Carlo estimation based on K samples for the argument inside the logarithm of Equation (2).

- MIWAE estimates the IWAE objective using M samples, instead of only 1 as in IWAE, to estimate the gradients of the objective.
- CIWAE is a binary convex combination of the VAE and IWAE objectives weighted by the parameter β. This bound is looser than the IWAE but tighter than VAE bound.
- PIWAE considers the SNR issues arising in the inference network of the IWAE due to increased K. Thus, $q_\phi(z|x)$ is optimized using the MIWAE bound, while $p_\theta(x|z)$ is optimized with the tighter IWAE bound to ensure improved density estimation.

Reproduced results

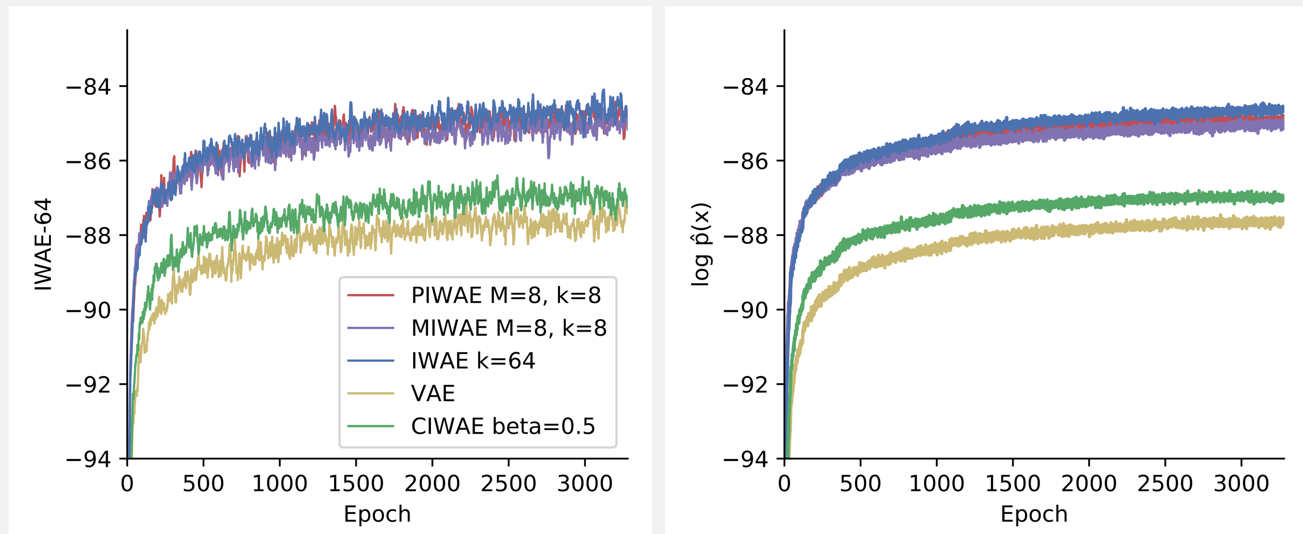

Figure 2:The metrics of IWAE$_{64}$ and logp̂(x) of re-implemented models from [2, 1].

Table 1:Reproduction of models proposed by [1] trained on MNIST for 3280 epochs.

Metric	IWAE	PIWAE$_{(8,8)}$	MIWAE$_{(8,8)}$	CIWAE$_{\beta=0.5}$	VAE
IWAE-64	-84.64	-84.72	-84.98	-87.05	-87.26
logp̂(x)	-84.64	-84.90	-85.04	-87.00	-87.66
-KL(Q‖P)	0.00	0.19	0.06	-0.05	0.40

Extended results

Table 2:Extension of the experiments over a larger tested model. Trained on MNIST for 3280 epochs.

Bigger model	IWAE	PIWAE$_{(8,8)}$	MIWAE$_{(8,8)}$	CIWAE$_{\beta=0.5}$	VAE
IWAE-64	-83.85	-83.86	-83.47	-84.92	-85.33
logp̂(x))	-83.92	-83.84	-83.66	-84.98	-85.31
-KL(Q‖P)	0.06	-0.02	0.19	0.06	-0.02

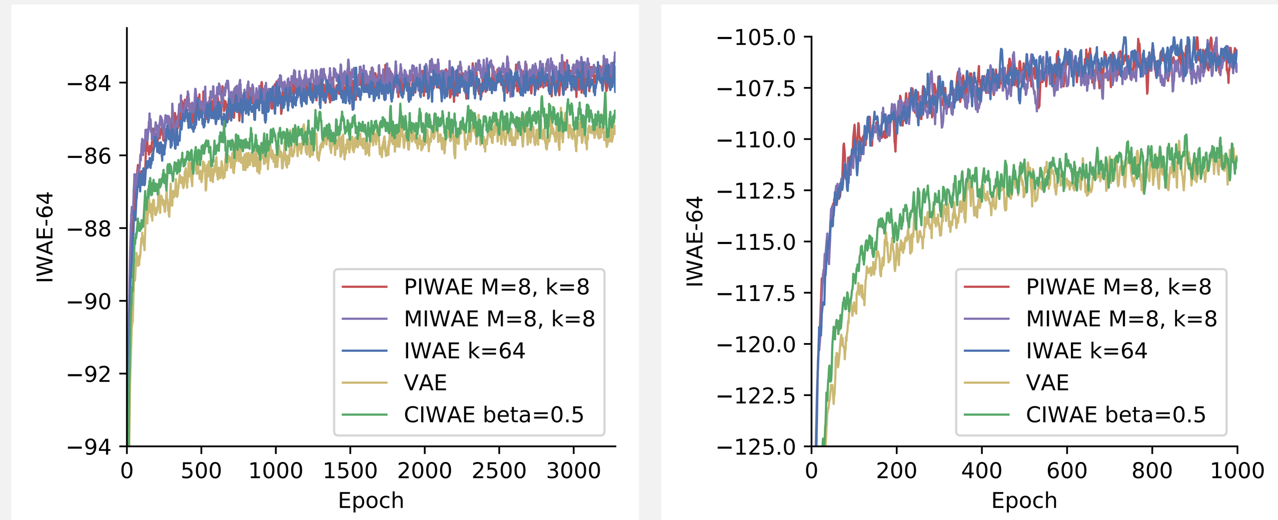

Figure 3:Extended evaluation of newly explored model variants: Left: Training process of the larger model (aproximately twice the number of parameters). Right: Training process on the Omniglot dataset (with the regular sized model).

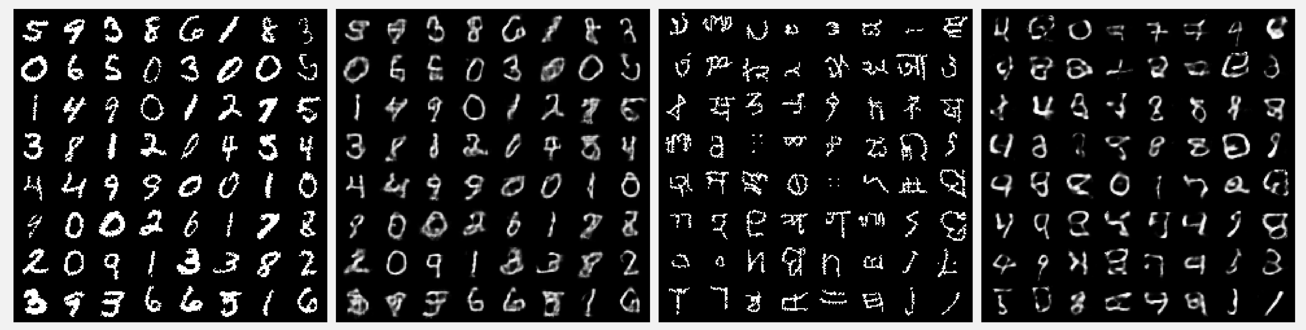

Figure 4:Generalization ability measured by reconstruction quality. Left to right: original MNIST input, reconstruction with MIWAE$_{(8,8)}$ trained on Omniglot, original Omniglot input, reconstruction with MIWAE$_{(8,8)}$ trained on MNIST.

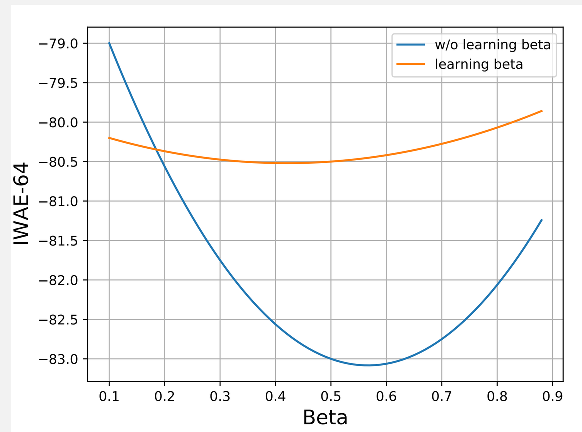

Figure 5:Evaluation of the proposed CIWAE model with a learnable parameter β. The average score when learning β is -80.34±0.49, whereas the average score is -81.99±2.98 without learning β. Our method is more independent to the initial setting of the β parameter.

References

[1] Tom Rainforth, Adam Kosiorek, Tuan Anh Le, Chris Maddison, Maximilian Igl, Frank Wood, and Yee Whye Teh. Tighter variational bounds are not necessarily better. In *International Conference on Machine Learning*, pages 4277–4285. PMLR, 2018.

[2] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.