Towards Consistent Variational Auto-Encoding (Student Abstract)

Yijing Liu,[*1] **Shuyu Lin,**[*2] **Ronald Clark**[3]
[2]University of Oxford
Email: slin@robots.ox.ac.uk
Post: Balliol College, Oxford, OX1 3BJ, UK

Abstract

Variational autoencoders (VAEs) have been a successful approach to learning meaningful representations of data in an unsupervised manner. However, suboptimal representations are often learned because the approximate inference model fails to match the true posterior of the generative model, i.e. an inconsistency exists between the learnt inference and generative models. In this paper, we introduce a novel consistency loss that directly requires the encoding of the reconstructed data point to match the encoding of the original data, leading to better representations. Through experiments on MNIST and Fashion MNIST, we demonstrate the existence of the inconsistency in VAE learning and that our method can effectively reduce such inconsistency.

Introduction

Variational autoencoders (VAEs, (Kingma and Welling 2013)) are a popular generative model $p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})$ that aims to summarise the regularities of a given dataset $\mathcal{D}_N = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$ in a low dimensional latent space $\mathbf{z}$. Utilizing variational inference, VAEs introduce an approximate inference model $q_\phi(\mathbf{z}|\boldsymbol{x}_i)$ and is able to replace the often intractable data likelihood objective $\int p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})\mathrm{d}\mathbf{z}$ in the generative models by an alternate evidence lower bound (ELBO) objective. The optimal ELBO will only be reached when the approximate inference model matches the true inference model, i.e. $q_\phi(\mathbf{z}|\boldsymbol{x}_i) = p_\theta(\mathbf{z}|\boldsymbol{x}_i) = \frac{p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{\int p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})\mathrm{d}\mathbf{z}}$, and this indicates that the best explanation of the given dataset in the target latent space will only be obtained if a consistent auto-encoding process can be found.

However, various works have shown that VAEs often deliver a sub-optimal inference model that does not match with the generative model completely, leading to inaccurate representation of the observed data in the latent space and worse generation quality (Cremer, Li, and Duvenaud 2018). Many works have been proposed to indirectly mitigate such effects by using either a more expressive inference model $q_\phi(\mathbf{z}|\boldsymbol{x}_i)$ (Rezende and Mohamed 2015; Kingma, Salimans, and Welling 2016; Ranganath, Tran, and Blei 2016) or a more flexible model for the prior $p(z)$ (Dilokthanakul et al. 2016; Tomczak and Welling 2018).

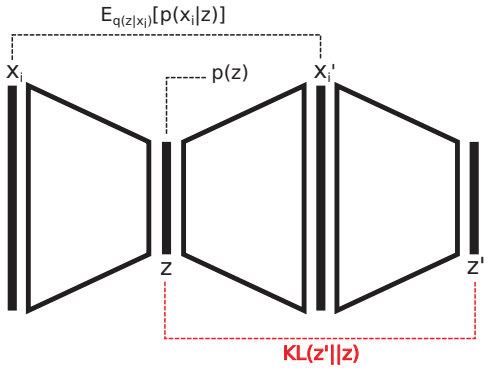


Figure 1: Our proposed encoder-decoder matching loss.

Different from the aforementioned works, we directly measure the degree of inconsistency between the inference and the generative models through the drift in the latent space if the decoded data points are encoded again using the learnt inference model, as shown in Figure 1. We then propose to reduce the inconsistency by minimizing such drifts together with the VAE ELBO objective. Our solution introduces no additional parameters to the original VAE algorithm. Further, our proposed objective is in fact a lower bound to the original ELBO loss and the optimum is obtained when the ELBO approaches the data log likelihood, so the original VAE's learning objective is unmodified. Through experiments, we demonstrate that our method leads to much more consistent inference and generative models.

Our Proposal

The ELBO loss $\mathcal{L}(\mathbf{x};\theta,\phi)$ that VAE models use for learning consists of the following two terms:

$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big[\log p_\theta(\mathbf{x}|\mathbf{z})\big] - D_{\mathrm{KL}}\big[q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})\big], \quad (1)$$

which are indicated in Figure 1 by the two black dashed lines to require the reconstructed data point $\boldsymbol{x}_i'$ to preserve key information of the original data point $\boldsymbol{x}_i$ and the encoding distribution $q_\phi(\mathbf{z}|\boldsymbol{x}_i)$ to remain close to a prior distribution

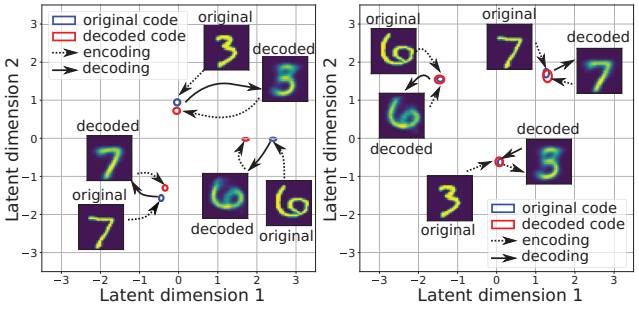
$p(\mathbf{z})$. In addition to these two losses, we introduce another consistency loss. To derive it, we need to encode the reconstructed data point $\boldsymbol{x}'_i$ again using the inference model and obtain a second encoding distribution $q_\phi(\mathbf{z}'|\boldsymbol{x}'_i)$. If the inference model matches the generative model perfectly, then the two encoding distributions would be the same. On the other hand, if inconsistency exists, then $q_\phi(\mathbf{z}'|\boldsymbol{x}'_i)$ will drift away from $q_\phi(\mathbf{z}|\boldsymbol{x}_i)$. Therefore, the distance between $q_\phi(\mathbf{z}|\boldsymbol{x}_i)$ and $q_\phi(\mathbf{z}'|\boldsymbol{x}'_i)$, such as measured by Kullback-Leibler (KL) divergence (i.e. $D_{\mathrm{KL}}(\mathbf{z}'||\mathbf{z})$), can be a good indication of the level of inconsistency and should be minimized. This gives us our modified learning objective as:

$$\mathcal{L}(\mathbf{x};\theta,\phi) - \alpha \cdot D_{\mathrm{KL}}(\mathbf{z}'||\mathbf{z}), \tag{2}$$

where $\alpha \geq 0$ is an adjustable weight that controls the strength of the consistency constraint. As $D_{\mathrm{KL}}(\mathbf{z}'||\mathbf{z})$ is non-negative, our objective is a lower bound to the original ELBO objective and the bound is tight when $D_{\mathrm{KL}}(\mathbf{z}'||\mathbf{z}) = 0$, i.e. when the inference and the generative models perfectly match. Therefore, our learning objective does not alter the original VAE's learning objective.

Results

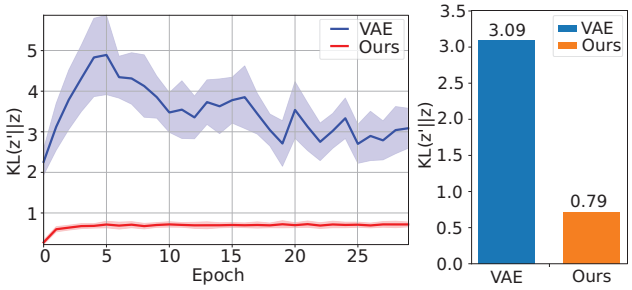
To visualise the effect of our method in removing the inconsistency between the inference and the generative models, we show the encodings of 3 MNIST digits $q_\phi(\mathbf{z}|\boldsymbol{x}_i)$, the reconstructed digits and the encodings of the reconstruction $q_\phi(\mathbf{z}'|\boldsymbol{x}'_i)$ for both VAE and our models in Figure 2. A clear drift occurs between the two encodings (red and blue circles) as a result of the VAE learning, indicating the optimised inference model has not been matched with the generative model. In contrast, learning under our loss given in Equation (2), we are able to remove the drift and hence obtain a consistent pair of inference and generative models, indicating to reach the optimal learning outcome, i.e. ELBO approaches the data log likelihood.



(a) VAE. (b) Ours.

Figure 2: 3 pairs of encodings of original and reconstructed images are ploted in the 2D latent space for VAE and our models. A drift occurs in VAE encodings indicating inconsistency between the inference and generative models, whereas our method successfully removes such a drift.

We also quantitatively evaluate the inconsistency between the inference and the generative models by the $D_{\mathrm{KL}}(\mathbf{z}'||\mathbf{z})$ between the two encodings $q_\phi(\mathbf{z}|\boldsymbol{x}_i)$ and $q_\phi(\mathbf{z}'|\boldsymbol{x}'_i)$ for both VAE and our methods during and at the end of training on a test set of 10k images. In Figure 3, the representation learnt under VAE algorithm constantly suffers from the inconsistency between the two models, whereas our method manages to control the inconsistency at very low level throughout the training. As a result, the inconsistency has been reduced 3 times using our method at the end of training.



(a) During training. (b) After training.

Figure 3: $D_{\mathrm{KL}}(\mathbf{z}'||\mathbf{z})$ during and after training on the held-out test set for MNIST dataset (lower is better).

Conclusion and Future Work

We aim to solve the suboptimality issue in VAE learning due to the inconsistency between the inference and the generative models. By introducing a novel consistency loss that directly requires the encoding of the reconstructed data point to match the encoding of the original data, we effectively coax the inference and the generative to be an inverse function of each other and, hence, approach the optimal solution to VAE objective. We notice the weight α on the consistency loss has an impact on the learning result and we would like to investigate how to determine the optimal α in the future.

References

Cremer, C.; Li, X.; and Duvenaud, D. 2018. Inference suboptimality in variational autoencoders.

Dilokthanakul, N.; Mediano, P. A. M.; Garnelo, M.; Lee, M. C. H.; Salimbeni, H.; Arulkumaran, K.; and Shanahan, M. 2016. Deep unsupervised clustering with gaussian mixture variational autoencoders. *CoRR* abs/1611.02648.

Kingma, D. P., and Welling, M. 2013. Auto-encoding variational bayes. *CoRR* abs/1312.6114.

Kingma, D. P.; Salimans, T.; and Welling, M. 2016. Improving variational inference with inverse autoregressive flow.

Ranganath, R.; Tran, D.; and Blei, D. M. 2016. Hierarchical variational models. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016*, 324–333.

Rezende, D. J., and Mohamed, S. 2015. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*.

Tomczak, J. M., and Welling, M. 2018. VAE with a vampprior. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2018*, 1214–1223.