VARIATIONAL COMPOSITE AUTOENCODERS

Jiangchao Yao[†,‡], **Ivor W. Tsang**[‡], **Ya Zhang**[†]
Cooperative Medianet Innovation Center[†], Center for Artificial Intelligence[‡]
Shanghai Jiao Tong University[†], University of Technology Sydney[‡]
Sunarker@sjtu.edu.cn, Ivor.Tsang@uts.edu.au, ya_zhang@sjtu.edu.cn

ABSTRACT

Learning in the latent variable model is challenging in the presence of the complex data structure or the intractable latent variable. Previous variational autoencoders can be low effective due to the straightforward encoder-decoder structure. In this paper, we propose a variational composite autoencoder to sidestep this issue by amortizing on top of the hierarchical latent variable model. The experimental results confirm the advantages of the proposed method.

1 INTRODUCTION

The latent variable model in Figure 1(a) has been widely explored for the representation learning. One line of research has concentrated on a broad range of data structures. For example, Kipf & Welling (2016) proposed a variational graph autoencoder to consider the graph-structured data. Chen et al. (2017) introduced a variational lossy autoencoder to learn the controllable representation of the image. Kusner et al. (2017) proposed a grammar variational autoencoder to directly learn from a parse tree described by the grammar. Another line of research pays attention to the optimization difficulty on the intractable latent variable. For instance, Mnih & Gregor (2014) proposed a general score function method with control variates to optimize the model with continuous or discrete latent variables. Maddison et al. (2017) introduced a *Concrete distribution*, the continuous relaxation of the discrete distribution, to approximately reparameterize the discrete latent variable. Nalisnick & Smyth (2017) extended variational autoencoders to perform posterior inference for the latent variable of Stick-Breaking processes.

Although the promising performance has been achieved, previous variational autoencoders in Figure 1(c) can be low effective due to the straightforward encoder-decoder structure. In the presence of the very sophisticated data or the unstable optimization on the latent variable, it is challenging with only resort to the implementation of the encoder and the decoder. This problem motivates us to explore a more expressive structure in the representation learning, which can further improve the encoding efficiency and ease the optimization burden.

In this paper, we propose a variational composite autoencoder (VCAE) to sidestep this issue by amortizing on top of the hierarchical latent variable model. As shown in Figure 1(b), a surrogate variable is introduced between the latent variable and the data to intermediate their gap. Specifically, in terms of the complex data structure, the representation learning is implemented hierarchically, i.e., the raw data is first projected into a moderate surrogate space, on top of which a latent variable is then learned. For the optimization difficulty on the latent variable, VCAE provides a possible framework to unify previous approaches, score function methods and reparameterization. Experimental results demonstrate the superior performance of VCAE over state-of-the-arts in these two aspects.

2 VARIATIONAL COMPOSITE AUTOENCODERS

The key idea is illustrated in Figure 1. Compared with the one-off encoding and decoding process in variational autoencoder (VAE) (Kingma & Welling, 2014), we decouple the representation learning into two sub-procedures by introducing a surrogate variable. Then the learning difficulty either on the sophisticated data structure or the unstable optimization can be alleviated in this surrogate space.

In VCAE, we use two inference networks $q_{\theta_1}(s|x)$ and $q_{\theta_2}(z|x)$ to represent the encoding processes from x to s and z respectively. The generative processes from z to s and s to x are modeled with

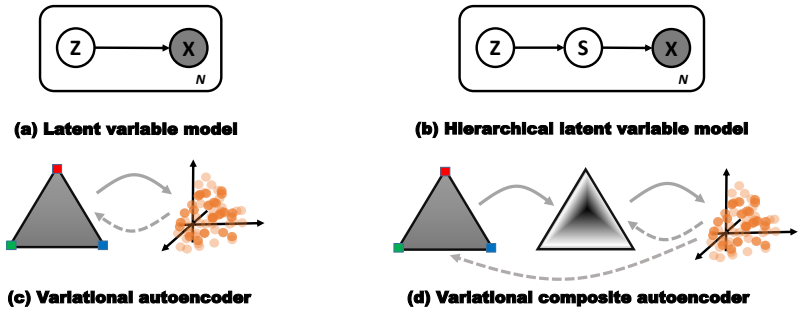

Figure 1: The idea of our work. (a) and (b) are the graphical representations of two models. z represents the latent variable, s is the surrogate variable and x is the observation. (c) and (d) illustrate the variational autoencoder and our variational composite autoencoder. We introduce a surrogate variable to sidestep the learning difficulty between the latent variable and the observation.

two networks $p_\psi(s|z)$ and $p_\phi(x|s)$. The graphical representation of our hierarchical latent variable model is illustrated in Figure 1(b). According to the encoding and decoding processes in Figure 1(d), we can deduce a variational lower bound as follows,

$$\ln p(x) \geq \mathbb{E}_{q_{\theta_1}}\left[\ln p(x|s)\right] - \mathbb{E}_{q_{\theta_2}}\left[\mathrm{D}_{KL}(q_{\theta_1}(s|x)||p_\psi(s|z))\right] - \mathrm{D}_{KL}(q_{\theta_2}(z|x)||p(z)). \quad (1)$$

Previous stochastic optimization on the variational Monte Carlo objective can be seamlessly applied to optimize this bound, which makes VCAE general to many applications. For example, when both z and s are reparameterizable, the objective can be directly optimized by the stochastic backpropagation (Rezende et al., 2014). When z is the discrete variable and s is chosen to be reparameterizable, score function methods can be used jointly with reparameterization for optimization. In this paper, we will focus on the comparison between VAE and VCAE in the representation learning, and the optimization effectiveness of VCAE compared with previous state-of-the-arts.

3 Evaluation

We evaluate the effectiveness of VCAE using two benchmark tasks, generative modeling and structured prediction, with the statically binarized MNIST dataset Salakhutdinov & Murray (2008). The standard partition 50000/10000/10000 is used to split the MNIST dataset into the training, validation and test sets. All sub-modules in VCAE are parameterized with the sigmoid belief network. Specifically, 200 stochastic units interleaved by one linear layer (denoted as **Linear**) or two layers of 200 tanh units (denoted as **Nonlinear**) are used for two encoding channels $q_{\theta_1}(s|x)$, $q_{\theta_2}(z|x)$ and one decoding channel $p_\phi(x|s)$. One sigmoid layer is fixed for $p_\psi(s|z)$. All models are randomly initialized and optimized by the ADAM optimizer Kingma & Ba (2015) with a learning rate of 3e-4. During the test, we compute the importance-sampled estimate of the log-likelihood objective.

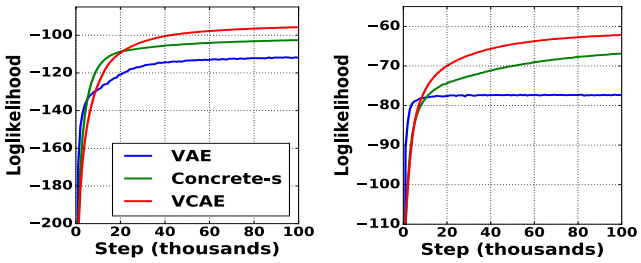

Figure 2: Comparison among VAE, Concrete-s and VCAE on generative modeling (left) and structured prediction (right).

Methods	Generative modeling	Structured prediction
VAE	-112.8	-77.1
Concrete-s	-105.3	-66.3
VCAE	**-95.7**	**-62.1**

Table 1: The importance-weighted estimate of the log-likelihood objective on two tasks in the test dataset.

In the first experiment, we use VCAE to directly compare with VAE (Kingma & Welling, 2014). In the hierarchical latent variable model, the Gaussian variable and the reparameterizable Concrete

Methods		NVIL	REBAR	Concrete-z	Concrete-s	VCAE
Generative modeling	Linear	-108.3	-107.6	-107.5	-105.3	**-97.9**
	Nonlinear	-100.9	-102.0	-101.2	-100.6	**-95.0**
Structured prediction	Linear	-67.9	-66.4	-66.1	-66.3	**-64.4**
	Nonlinear	-63.8	-63.5	-67.0	-63.3	**-62.9**

Table 2: The importance-weighted estimate of the log-likelihood objective in the test dataset.

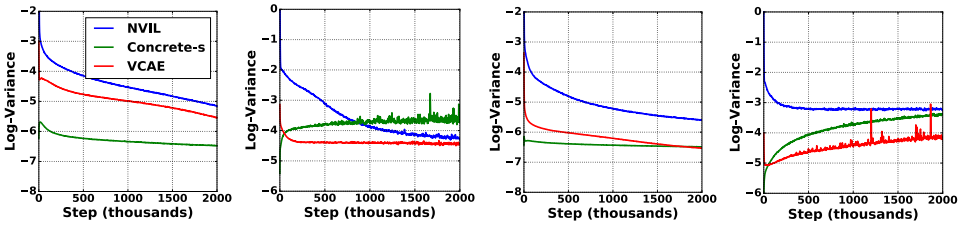

Figure 3: Gradient variance comparison among NVIL, Concrete-s and VCAE (Linear: the 1st and 3rd figures, Nonlinear: the 2nd and 4th figures) on generative modeling (left two) and structured prediction (right two).

variable (Maddison et al., 2017) are respectively chosen as the latent variable and the surrogate variable. Thus our VCAE will includes the Gaussian reparameterization (VAE) and the Concrete reparameterization (Concrete-s) parts. To comprehensively understand our method, we also use Concrete-s as a baseline in the experiments. Figure 2 and Table 1 present the performance of three methods in the Linear case. As can be seen, VCAE outperforms two baselines in the training procedure, and achieves -95.7 and -62.1 on the two tasks respectively, which is significantly better than VAE. Besides, as an important component in our method, Concrete-s also shows promising performance, but still underperforms VCAE. This indicates our hierarchical representation learning structure is more effective than those in previous VAEs.

In the second experiment, we use VCAE to optimize the challenging discrete variable model. Similarly, the Concrete variable is set as the surrogate variable. Then we unify the score function method, NVIL (Mnih & Gregor, 2014), and the reparamterization method, Concrete-s, into VCAE by applying them on $q_{\theta_2}(z|x)$ and $q_{\theta_2}(s|x)$ respectively. The resulting method is to compare with the state-of-the-art approaches and the final performance is summarized in Table 2. For the generative modeling task, the importance-weighted estimate of the log-likelihood objective for VCAE (Linear) is -97.9, while that of the best unbiased estimators REBAR (Linear) (Tucker et al., 2017) and the biased reparameterization method Concrete-z[1] are -107.6 and -107.5, respectively. For the structured prediction task, VCAE outperforms all unbiased estimators and the biased reparameterization method (Concrete-z) even in the case of continuous variable (Concrete-s). This confirm the advantages of VCAE in the optimization. To present a fine-grained performance analysis of VCAE, we further compare VCAE with NVIL and Concrete-s by tracing their gradient variance in the training phase, as the lower gradient variance the faster convergence is (Johnson & Zhang, 2013). As shown in Figure 3, in the Linear case, the gradient variance of VCAE is between that of NVIL and Concrete-s, and in the Nonlinear case, VCAE not only outperforms NVIL, but also surpasses Concrete-s. This indicates on the basis of the score function method in Eq. (1), the variance is further reduced due to the reparameterizable Concrete variable, and the improvement is even better than both of them.

4 Conclusion

In this paper, we propose a variational composite autoencoder to improve the effectiveness of representation learning. Compared with previous VAEs, VCAE enjoys the advantages of amortizing on top of the hierarchical latent variable model. By introducing a surrogate variable to learn the intermediate space, both learning burden and optimization difficulty can be alleviated. Experimental results in these two aspects demonstrate VCAE outperforms the state-of-the-art methods.

[1]In the test phase, the binarized Concrete samples are used to evaluate the lower bounds.

REFERENCES

Xi Chen, Diederik P Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational lossy autoencoder. *ICLR*, 2017.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, 2013.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *ICLR*, 2014.

Thomas N Kipf and Max Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016.

Matt J Kusner, Brooks Paige, and José Miguel Hernández-Lobato. Grammar variational autoencoder. *Proceedings of the 34th International Coference on International Conference on Machine Learning*, 2017.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *ICLR*, 2017.

Andriy Mnih and Karol Gregor. Neural variational inference and learning in belief networks. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

Eric Nalisnick and Padhraic Smyth. Stick-breaking variational autoencoders. *ICLR*, 2017.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *Proceedings of the 31st International Coference on International Conference on Machine Learning*, 2014.

Ruslan Salakhutdinov and Iain Murray. On the quantitative analysis of deep belief networks. In *Proceedings of the 25th international conference on Machine learning*, 2008.

George Tucker, Andriy Mnih, Chris J. Maddison, and Jascha Sohl-Dickstein. Rebar: Low-variance, unbiased gradient estimates for discrete latent variable models. *Advances in Neural Information Processing Systems*, 2017.