DEGENERATION IN VAE: IN THE LIGHT OF FISHER INFORMATION LOSS

Huangjie Zheng, Jiangchao Yao & Ya Zhang
Cooperative Medianet Innovation Center
Shanghai Jiao Tong University
Shanghai, 200240, China
{zhj865265,sunarker,ya_zhang}@sjtu.edu.cn

Ivor W. Tsang
Centre for Artificial Intelligence
University of Technology Sydney
Sydney, Australia
{ivor.tsang}@uts.edu.au

1 INTRODUCTION

Variational Autoencoder (VAE) (Kingma & Welling, 2013) is one representative generative model to combine variational inference with deep learning, and many variants of VAE are proposed to improve the effectiveness in both latent representation learning and generation (Burda et al., 2015; Tomczak & Welling, 2017; Rezende & Mohamed, 2015; Kingma et al., 2016; Sønderby et al., 2016; Gregor et al., 2015; Germain et al., 2015; Gulrajani et al., 2016). As previous works have made VAE more flexible, deeper architecture is needed for the increasing complexity of raw data and sophisticated models, whose impacts on VAE instead are still rarely explored.

Intuitively, deeper feed-forward networks are more powerful in hierarchical feature learning (Zeiler & Fergus, 2014) and in function approximation (Zhao et al., 2017), which should benefit the latent representation and generation in VAE. However, this conjecture is not exactly in accord with our observation: Compared to VAE in shallow architecture, 1) a deeper decoder leads VAE to produce more high-quality generation, while the latent representation becomes abstruse and harms downstream work like classification; 2) a deeper encoder helps the latent representation learn more global information to benefit classification, while the generation turns to be more blurry; 3) when encoder and decoder both go deeper, VAE performs worse in both latent representation and generation.

In this paper, we propose a Fisher Information (Brunel & Nadal, 1998) measure to quantify the information loss layer by layer. With such measure, we demonstrate that the information loss generally exists in the encoder and decoder, leading to the previous three types of degeneration. In addition, we demonstrate that skip connections could serve as a complementary information flow to help mitigate the degeneration in VAE models. Thus a variant named SCVAE, *i.e.*, VAE with skip connections, is proposed to preserve information when encoder and decoder go deeper. Finally, we conduct a series of experiments on widely used MNIST dataset (Lecun et al., 1998). Comprehensive results indicate that SCVAE performs well in information preservation, and mitigates degeneration to ensure a promising performance in latent representation learning and generation at same time.

2 FISHER INFORMATION LOSS ANALYSIS

The Fisher Information reflects how accurate an approximation of distribution is (Brunel & Nadal, 1998). Since encoding and decoding networks are supposed to approximate their corresponding distribution $q_\phi(z|x)$ and $p_\theta(x|z)$, we can evaluate the approximation with Fishier Information. We formally show how Fisher Information varies among layers with the following proposition:

Proposition 1. *Suppose a feed-forward network F with L layers, parameterized by $\Phi = \{\phi_1, \phi_2, \ldots, \phi_L\}$: $F = f_L \circ f_{L-1} \circ \cdots \circ f_1(X)$. When it is well optimized, for the Fisher Information in layer l and $l+1$, we have:*

$$\mathcal{I}_F(\phi_{l+1}) = \mathcal{I}_F(\phi_l) \left(\frac{\nabla_{\phi_l} F}{\nabla_{\phi_{l+1}} F} \right)^2 \tag{1}$$

where $\nabla_{\phi_l} F$ and $\nabla_{\phi_{l+1}} F$ signify the gradient in the lth and $(l+1)^{th}$ hidden layer.

Many works such as Saxe et al. (2013) have reported that the gradient tends to get smaller as we move backward through the hidden layers, which indicates that deeper layers tend to obtain less informa-

tion according to Equation 1. Therefore, when either encoder or decoder goes deeper, increasing depth leads to more information loss. In this way, VAE hardly maintains information sufficient to provide learning dynamic (Orhan & Pitkow, 2017), which causes previous degeneration.

Recall that skip connections are extra connections between nodes in different layers of a neural network, which have shown great promise in dealing with vanishing gradient problem in deep neural networks(He et al., 2016; Orhan & Pitkow, 2017). In this paper, we have the following proposition to show that skip connections are a possible way to alleviate the information loss.

Proposition 2. *Suppose the lth $(l \geq 1)$ hidden layer parameterized by ϕ_l receive information $f(h_{l-1})$ from the former layer h_{l-1}. Modeling with Fishier Information, when connected with skip connections, this layer shall receive more information compared with non-skip architecture:*

$$\mathcal{I}_{(f_l(h_{l-1}),c(h_{l-k}))}(\phi_l) > \mathcal{I}_{f_l(h_{l-1})}(\phi_l)$$

where the skip connection passes information $c(h_{l-k})$ from layer h_{l-k} by skipping k layers.

In this way, the skip connection can be regarded as a complementary information flow between layers. Following this idea, we propose a VAE model equipped with skip connections, named SCVAE. We make SCVAE skip one or more layers for information preservation. Skip connections as a simple method to preserve information flow, do not increase computation complexity and is compatible with deep neural networks. Most importantly, in the experiments, we observe that SCVAE mitigates the degeneration and achieves better performance in both latent representation and generation.

3 Experimental Results

The experiments are conducted on the MNIST dataset (Lecun et al., 1998) that consists of 10 categories of 28×28 hand-written digits. We follow the standard split 50,000/10,000/10,000 to partition the dataset as the training, validation and test parts. Fishier Information in network hidden layer in the following experiments is computed as suggested in Desjardins et al. (2015).

Plain VAE and SCVAE are implemented using MLPs with layers of 500 parameters. The shallow VAE model is of depth 1 hidden layer. When we make the model's encoder (resp. decoder) deeper, we note as q++ (resp. p++). Otherwise encoder has the same depth with decoder. We also use a model SCVAE-L which only modifies SCVAE to contain only one long-skipping-distance connection in encoder to demonstrate the effect of long-skipping-distance connection.

3.1 Information Preservation

In the first experiment, we respectively make VAE and SCVAE deeper to investigate the impact of depth on information amount. Figure 1 presents how information amount varies w.r.t. model depth. When VAE makes either encoder or decoder go deeper, the average information amount keeps a decreasing tendency. Different from plain VAE, SCVAE could generally maintain the information amount to the same level. Although these two models have similar amount of information in shallow case, as the model goes deeper, SCVAE remains more information amount than plain VAE.

In the next experiment, we respectively compute the average Fisher Information in encoder and decoder. Shown in Figure 2, deep VAE maintains little information amount. When encoder goes deeper, information amount mainly decays in decoder; while information amount decays more in encoder when decoder goes deeper. This indicates that information loss in encoder or decoder reflects corresponding degeneration types. The advantages of skip connection in information preservation is shown in SCVAE models, where SCVAE maintains the closest mean information amount as the shallow model does, as well as the ratio of information amount in encoder and in decoder. SCVAE-L leverages the long-skipping-distance connection and augment the information amount in encoder, but the capacity of preservation is finite due to lack of skip connection in intermediate layers.

3.2 Degeneration mitigation

To present how degeneration affects in representation learning and generation, we respectively evaluate these tasks with classification accuracy and negative log-likelihood (NLL). A simple SVM (Support Vector Machine) is trained and test with the learned latent representation in classification.

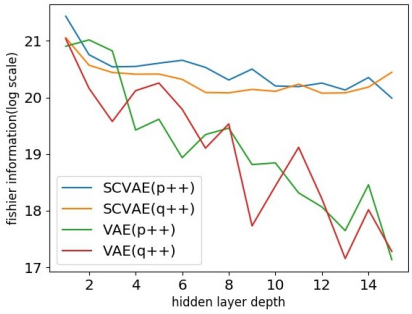

Figure 1: Mean Fisher Information w.r.t depth in VAE models and SCVAE models.

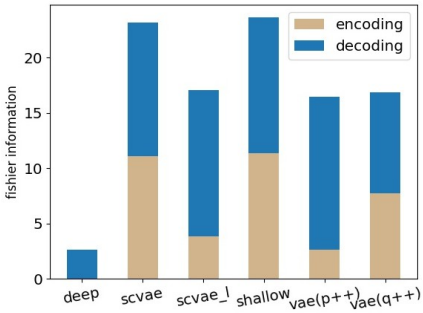

Figure 2: Comparison of mean Fisher Information in different models.

Table 1 suggests that unilaterally extending encoder or decoder's depth results in an improvement in a specific task, though at same time the performance in another task becomes worse: VAE(q++) achieves a better classification result, but sacrifices the NLL performance and vice versa for VAE(p++). Nevertheless, models with skip connections (SCVAE, SCVAE-L) achieve well performance in both tasks. Especially, SCVAE outperforms in both two tasks than other models.

Table 1: Test negative log-likelihood (NLL) and classification accuracy of VAE and SCVAE models on MNIST

Model	NLL	Acc
VAE(1L)	-87.89	0.8421
VAE(11L)	-206.09	0.1135
VAE(q++)	-91.13	0.9352
VAE(p++)	-81.59	0.7120
SCVAE	**-80.19**	**0.9588**
SCVAE-L	-84.02	0.9216

In Figure 3, we present qualitative results to have an intuitive understanding. As we analyzed in previous part, models suffer from degeneration due to the lack of information in encoder or decoder. When degenerated in encoder (VAE, VAE(p++)), the latent representation degenerates layer by layer in encoder; when degenerated in decoder (VAE, VAE(q++)), reconstructions are not only blurry but also contain incorrect digits. When free from degeneration, SCVAE benefits from deep architecture to produce more clear reconstructions and to learn a comprehensible representation. Recall that SCVAE-L contains less information in intermediate layers (Figure 2), we notice that intermediate layers show an abstruse presentation, implying degeneration in these layers.

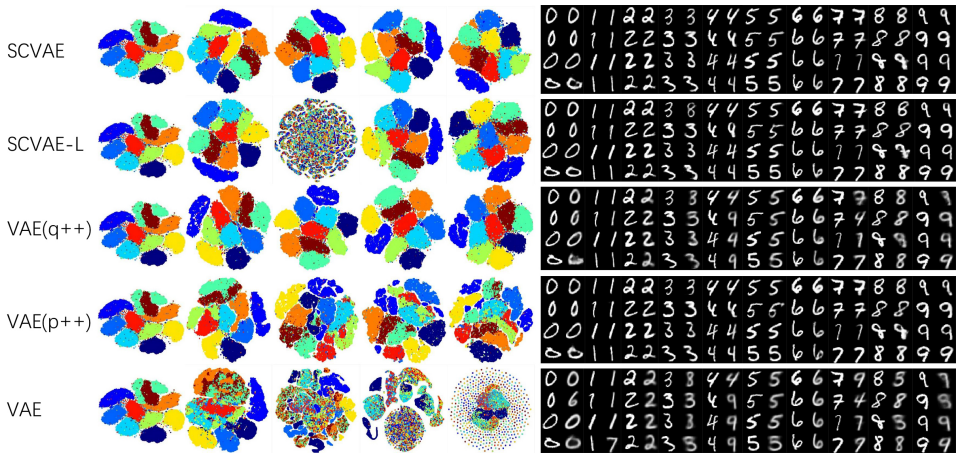

Figure 3: Left: with t-sne (Maaten & Hinton, 2008), representation visualization of raw data, first layer output, intermediate layer output, last layer output of the encoder, and latent space (from left to right). **Right**: ground truth (odd columns) and reconstruction (even columns).

REFERENCES

Nicolas Brunel and Jean-Pierre Nadal. Mutual information, fisher information, and population coding. *Neural computation*, 10(7):1731–1757, 1998.

Yuri Burda, Roger B. Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *CoRR*, abs/1509.00519, 2015. URL http://arxiv.org/abs/1509.00519.

Guillaume Desjardins, Karen Simonyan, Razvan Pascanu, and Koray Kavukcuoglu. Natural neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, pp. 2071–2079, Cambridge, MA, USA, 2015. MIT Press. URL http://dl.acm.org/citation.cfm?id=2969442.2969471.

Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. MADE: masked autoencoder for distribution estimation. *CoRR*, abs/1502.03509, 2015. URL http://arxiv.org/abs/1502.03509.

Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*, 2015.

Ishaan Gulrajani, Kundan Kumar, Faruk Ahmed, Adrien Ali Taiga, Francesco Visin, David Vázquez, and Aaron C. Courville. Pixelvae: A latent variable model for natural images. *CoRR*, abs/1611.05013, 2016. URL http://arxiv.org/abs/1611.05013.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013. URL http://dblp.uni-trier.de/db/journals/corr/corr1312.html#KingmaW13.

Diederik P Kingma, Tim Salimans, and Max Welling. Improving variational inference with inverse autoregressive flow. *arXiv preprint arXiv:1606.04934*, 2016.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov 1998. ISSN 0018-9219. doi: 10.1109/5.726791.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

A Emin Orhan and Xaq Pitkow. Skip connections eliminate singularities. *arXiv preprint arXiv:1701.09175*, 2017.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015.

Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. In *Advances in Neural Information Processing Systems*, pp. 3738–3746, 2016.

Jakub M Tomczak and Max Welling. Vae with a vampprior. *arXiv preprint arXiv:1705.07120*, 2017.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer, 2014.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. Learning hierarchical features from generative models. *arXiv preprint arXiv:1702.08396*, 2017.

A PROOF OF PROPOSITION

Proof of Proposition 1. Since F is twice differentiable w.r.t. ϕ, by definition, the Fishier Information in layer l can be written as:

$$\mathcal{I}_F(\phi_l) = -\mathbb{E}\left[\frac{\partial^2 logF}{\partial \phi_l^2}\right]$$

For layer $l+1$, we have:

$$\begin{aligned}
&\mathcal{I}_F(\phi_{l+1}) \\
&= -\mathbb{E}\left[\frac{\partial^2 logF}{\partial \phi_{l+1}^2}\right] \\
&= -\mathbb{E}\left[\frac{\partial}{\partial \phi_{l+1}}\left(\frac{\partial logF}{\partial \phi_l}\cdot\frac{\partial \phi_l}{\partial \phi_{l+1}}\right)\right] \\
&= -\mathbb{E}\left[\frac{\partial^2 logF}{\partial \phi_l \partial \phi_{l+1}}\cdot\frac{\partial \phi_l}{\partial \phi_{l+1}} + \frac{\partial^2 \phi_l}{\partial \phi_{l+1}^2}\cdot\frac{\partial logF}{\partial \phi_l}\right] \\
&= -\mathbb{E}\left[\frac{\partial^2 logF}{\partial \phi_l^2}\right]\left(\frac{\partial \phi_l}{\partial \phi_{l+1}}\right)^2 - \mathbb{E}\left[\frac{1}{F}\cdot\frac{\partial F}{\partial \phi_l}\right]\cdot\frac{\partial^2 \phi_l}{\partial \phi_{l+1}^2}
\end{aligned}$$

Additionally, between epoch $\{t, t+1\}$ in gradient descend optimization, $\phi_l^{(t+1)} = \phi_l^{(t)} + \lambda \cdot \nabla_{\phi_l} F$, where λ is the learning rate. Then we have $\partial \phi_l = \lambda \cdot \nabla_{\phi_l} F$. When it is well optimized, $\frac{\partial F}{\partial \phi_l} \simeq 0$, we can remove the corresponding expection term and have

$$\mathcal{I}_F(\phi_{l+1}) = \mathcal{I}_F\left(\phi_l\right)\left(\frac{\nabla_{\phi_l} F}{\nabla_{\phi_{l+1}} F}\right)^2.$$

□

It is interesting to notice the difference between a typical feed-forward neural network and VAE in perspective of information loss. Actually, the information loss widely exists in deep neural networks, but it is often ignored. Deep networks is powerful and usually applied in learning hierarchical features (Zeiler & Fergus, 2014). In fact, learning hierarchical features is a process that features become compact and discard superfluous information. In VAE, this process and its inversion compose the entire learning process. Therefore, the information loss is not tolerable for VAE's learning, which implies that VAE cannot **simply** go deeper.

Proof of Proposition 2. When one layer is equipped with a skip connection, this layer receives information from both former layer and the skip connection, thus the received information can be modeled as Fisher Information measure on a jointly distributed random variables $(f_l(h_{l_1}), c(h_{l-k}))$. According to Fishier Information's decomposition rule and its positivity, we have:

$$\begin{aligned}
\mathcal{I}_{(f_l(h_{l-1}), c(h_{l-k}))}\left(\phi_l\right) &= \mathcal{I}_{f_l(h_{l-1})}\left(\phi_l\right) + \mathcal{I}_{c(h_{l-k})|f_l(h_{l-1})}\left(\phi_l\right) \\
&> \mathcal{I}_{f_l(h_{l-1})}\left(\phi_l\right)
\end{aligned}$$

□