# [Supplementary] Variational Laplace Autoencoders

**Yookoon Park** [1] **Chris Dongjoo Kim** [1] **Gunhee Kim** [1]

## 1. Image Samples

Figure 1 and 2 illustrates examples of reconstruction and generation samples made by the VLAE on MNIST (LeCun et al., 1998), Fashion MNIST (Xiao et al., 2017), Omniglot (Lake et al., 2013), SVHN (Wang et al., 2011) and CIFAR 10 (Krizhevsky & Hinton, 2009). Overall, the reconstructions are sharp but generated samples tend to be blurry, especially when the data is complex (e.g. CIFAR10). We expect using convolutional architectures to be helpful for improving image generation qualities.

## 2. Experimental Details

We optimize using ADAM (Kingma & Ba, 2015) with learning rate 0.0005. Other parameters of the optimizer is set to default values. We experiment with $T = 1, 2, 4, 8$ where $T$ is the number of iterative updates for VLAE and SA-VAE, or the number of flow transformations for VAE+HF. We set the batch size to 128. All models are trained up to 2000 epochs at maximum and evaluated using the checkpoint that gives the best validation performance.

We set $\alpha_t = 0.5/(t+1)$ as decay for the VLAE update. For SA-VAE, the variational parameter $\boldsymbol{\lambda}_t$ is updated $T$ times using SGD: $\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t + \alpha \frac{\partial}{\partial \boldsymbol{\phi}} \mathcal{L}_{\boldsymbol{\theta}}(\mathbf{x}; \boldsymbol{\lambda}_t)$ with $\alpha = 0.0005$. The value of $\alpha$ is determined using a grid search among $\{1.0, 0.1, 0.001, 0.0005, 0.0001\}$ on the small network. We estimate the gradient using a single sample of $\mathbf{z}$ and apply the gradient norm clipping to avoid divergence of SA-VAE.

In our experiments, we find that the bigger models are susceptible to the parameter initialization, and their latent variables are prone to collapse if not properly initialized. We also observe VLAE and VAE+IAF are relatively robust to hyperparameter settings compared to other models. In order to prevent latent variable collapse, we use He's initialization (He et al., 2015) to preserve variance of backward propagation with gain of $2^{1/3}$ to account for the network structure that consists of two ReLU layer and one linear layer. In this way, the variance of gradients is preserved in initial phase of training. Furthermore, the data is mean-normalized and scaled so that the reconstruction loss at initial state is approximately 1. These changes successfully prevent latent variable collapse and significantly improve overall performance of the models.

This finding hints that it is crucial to preserve gradient variance throughout the networks. Note that the gradient signal to the encoder comes from two sources: (1) KL divergence term $D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))$ (2) Reconstruction term $\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[\ln p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})]$. While the gradient from the KL divergence term is directly fed into the encoder, the gradient from the reconstruction term - which is essential for preventing the latent variable collapse - have to propagate backwards through the decoder to reach the encoder. We hypothesize that if the networks are not initialized properly, the reconstruction gradient is overwhelmed by the KL divergence gradient which drives the approximate posterior $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ to collapse to the prior $p(\mathbf{z})$ in the initial stage of training.

## 3. Conjugate Gradient Method

To measure the performance of the Conjugate Gradient (CG) method as an alternative to the update equations of the main draft, we implement the VLAE+CG model where the mode update equation is replaced with the CG ascent step. We use nonlinear Conjugate Gradient of Polak-Ribiére method. For more details on nonlinear Conjugate Gradient methods, we refer readers to (Shewchuk, 1994; Dai, 2010). With $T = 4$, we find that VLAE+CG yields about 25% speed-up compared to the VLAE with minor performance loss ($\sim$ 1%) on MNIST.

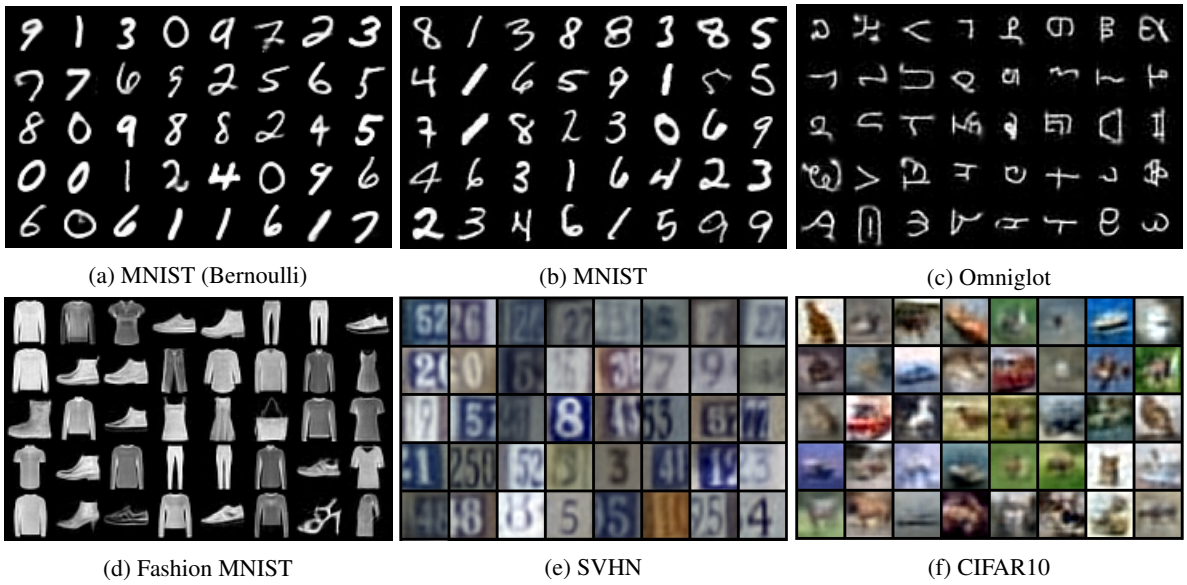
## References

Dai, Y. Nonlinear conjugate gradient methods. *Wiley Encyclopedia of Operations Research and Management Science*, 2010.

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015.

Kim, Y., Wiseman, S., Millter, A. C., Sontag, D., and Rush, A. M. Semi-amortized variational autoencoders. In *ICML*, 2018.

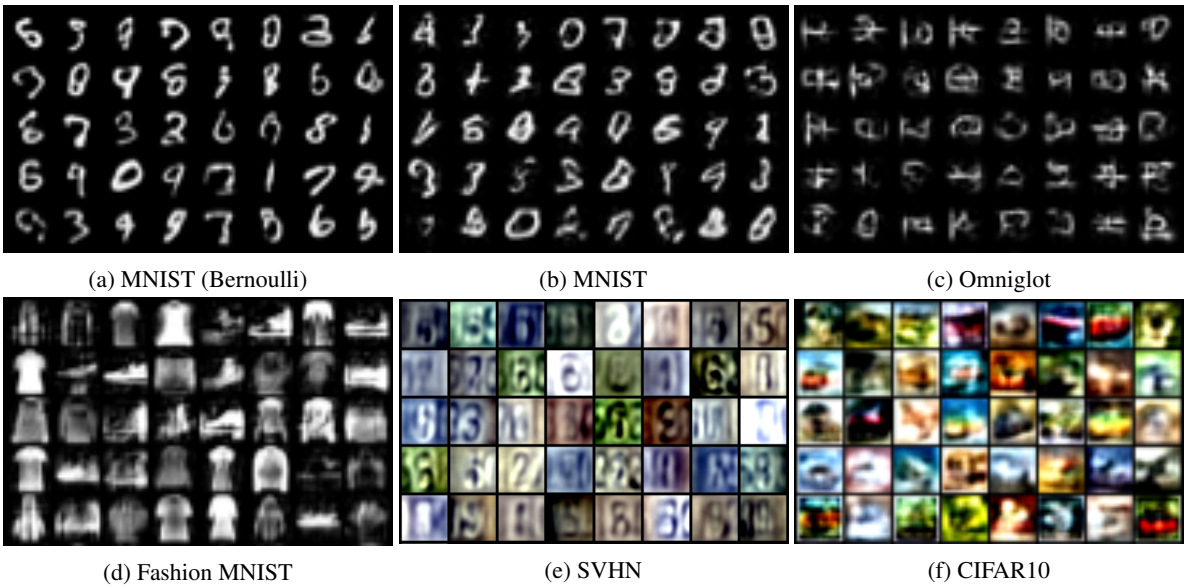
Kingma, D. and Ba, J. Adam: A Method for Stochastic Optimization. In *ICLR*, 2015.



(a) MNIST (Bernoulli) (b) MNIST (c) Omniglot

(d) Fashion MNIST (e) SVHN (f) CIFAR10

*Figure 1.* Examples of sample reconstructions by VLAE. Gaussian output distribution is used unless otherwise stated.



(a) MNIST (Bernoulli) (b) MNIST (c) Omniglot

(d) Fashion MNIST (e) SVHN (f) CIFAR10

*Figure 2.* Examples of output samples generated by VLAE. Gaussian output distribution is used unless otherwise stated.


Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. Improved variational inference with inverse autoregressive flow. In *NeurIPS*, 2016.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. Technical report, Computer Science Department, University of Toronto, 2009.

Lake, B. M., Salakhutdinov, R. R., and Tenenbaum, J. One-shot learning by inverting a compositinoal causal process. In *NeurIPS*, 2013.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient based learning applied to document recognition. In *IEEE*, 1998.

Shewchuk, J. R. An introduction to the conjugate gradient method without the agonizing pain, 1994.

Tomczak, J. M. and Welling, M. Improving variational auto-encoders using householder flow. In *NeurIPS Workshop on Bayesian Deep Learning*, 2016.

Wang, Y. N. T., Coates, A., Bissacco, A., Wu, B., and Ng,

A. Y. Reading digits in natural images with unsupervised feature learning. In *NeurIPS*, 2011.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. In *arXiv*, 2017.