Multi-Stage Variational Auto-Encoders for Coarse-to-Fine Image Generation

Lei Cai * Hongyang Gao † Shuiwang Ji ‡

Abstract

Variational auto-encoder (VAE) is a powerful unsupervised learning framework for image generation. One drawback of VAE is that it generates blurry images due to its Gaussianity assumption and thus ℓ_2 loss. To allow the generation of high quality images by VAE, we increase the capacity of decoder network by employing residual blocks and skip connections, which also enable efficient optimization. To overcome the limitation of ℓ_2 loss, we propose to generate images in a multi-stage manner from coarse to fine. In the simplest case, the proposed multi-stage VAE divides the decoder into two components in which the second component generates refined images based on the course images generated by the first component. Since the second component is independent of the VAE model, it can employ other loss functions beyond the ℓ_2 loss and different model architectures. The proposed framework can be easily generalized to contain more than two components. Experiment results on the MNIST and CelebA datasets demonstrate that the proposed multi-stage VAE can generate sharper images as compared to those from the original VAE.

1 Introduction

In recent years, progress in deep learning has promoted the development of generative models[1, 2, 3, 4, 5] that are able to capture the distributions of high-dimensional dataset and generate new samples. Variational auto-encoder (VAE)[6] is a powerful unsupervised learning framework for deep generative modeling. In VAE, the input data is encoded into latent variables before they are reconstructed by the decoder network. The VAE learns the transformation parameters by optimizing a variational lower bound of the true likelihood. The lower bound consists of two components. The first component is the Kullback-Leibler (KL) divergence between the approximate posterior and a prior distribution, which is commonly a normal distribution. The second component is the reconstruction loss given a latent variable. The VAE assumes that the output follows a normal distribution given the latent variable, thereby leading to an ℓ_2 loss in the objective function. It has been shown that the ℓ_2 loss leads to blurry images when the data are drawn from multi-modal distributions.

To make the VAE generate high quality images, some approaches have been proposed to improve the decoder network [8, 9, 10]. Since the decoder network is usually implemented with convolutional neural networks (CNNs) [11], we can increase the network depth to improve the capacity of decoder networks as in [12, 13, 14]. However, deeper networks can be difficult to optimize. Therefore, we employ the deep residual blocks, which are easy to optimize, to increase the capacity of decoder. By employing residual blocks in the decoder network, the VAE can generate high quality images. However, it still suffers from the effect of ℓ_2 loss and thus generates blurry images.

In this work, we propose a multi-stage VAE framework to generate high quality images. The key idea of multi-stage VAE is to generate images from coarse to fine. One challenge is that, since the decoder network is trained end-to-end, it is difficult to control the decoder network and make it generate images from coarse to fine. A simple solution is to train two models separately in which the first model generates a coarse image and the second model refines the coarse image. A drawback of this simple approach is that it reduces the efficiency of the model and involves more computational costs. To obtain fine images efficiently, we propose to employ an ℓ_2 loss in the middle of the decoder network, thus requiring coarse images to be generated in an intermediate stage of the decoder network. The remaining parts of the encoder network can be considered as a model that takes coarse images as inputs and generates refined versions of them as outputs. Indeed, the second network can be considered as a super-resolution network. Following this interpretation, we can employ any loss functions to refine the images in the super-resolution network[15], thereby overcoming the effect of ℓ_2 loss. In this way, we can generate images from coarse to fine and alleviate the effect of ℓ_2 loss without introducing extra parameters. Experimental results on the MNIST and CelebA datasets demonstrate that the proposed multi-stage VAE can capture more details and generate sharper images than the original VAE. Some sample results are given in Figure 1.

2 Multi-Stage Variational Auto-Encoder

2.1 Variational Auto-Encoder Variational auto-encoder (VAE) [16] is a generative model that is able to capture

*School of Electrical Engineering and Computer Science at Washington State University. Email: lei.cai@wsu.edu

†Department of Computer Science & Engineering at Texas A&M University. Email: hongyang.gao@tamu.edu

‡Department of Computer Science & Engineering at Texas A&M University. Email: sji@tamu.edu

Copyright © 2019 by SIAM
Unauthorized reproduction of this article is prohibited

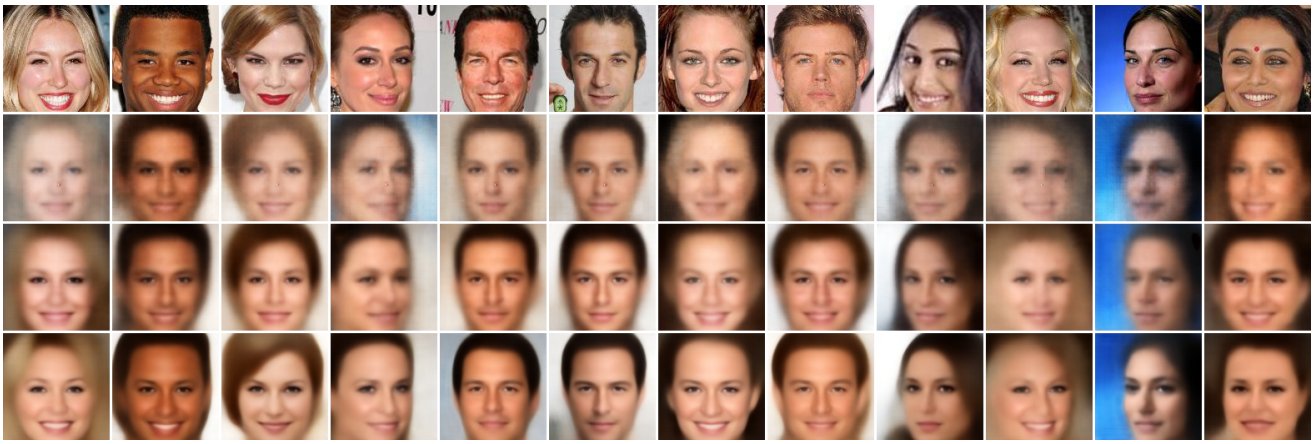

Figure 1: Comparison of reconstructed images from the CelebA dataset. The first row is the input images in the CelebA training set. The second row is the reconstructed images generated by the original VAE. The third and fourth rows are the results of deep residual VAE and multi-stage VAE, respectively.

the probability distribution over high-dimensional datasets. For image generation tasks, given a dataset $X = \{x^{(i)}\}_{i=1}^{N}$, we wish to learn a distribution function that can capture the dependencies among pixels. To tackle this problem, we can train a distribution model $p_{\theta_1}(x)$, parameterized by θ_1, to approximate the data distribution and optimize the model by maximizing the log likelihood as follows:

(2.1)
$$\log p_{\theta_1}(X) = \log p_{\theta_1}(x^{(1)}, \ldots, x^{(N)}) = \sum_{i=1}^{N} \log p_{\theta_1}(x^{(i)}).$$

However, probability distributions in high-dimensional space are very difficult to model. Thus, a low-dimensional latent variable z is usually introduced. It has been shown in [16] that the latent variable models can be optimized efficiently by maximizing a variational lower bound on the likelihood function as

$$\begin{aligned} \log p_{\theta_1}(x) &\geq \mathbb{E}_{q_\phi(z|x)}[\log p_{\theta_1}(x|z)] - D_{KL}[q_\phi(z|x)|p_{\theta_1}(z)] \\ &= -\mathcal{L}_{VAE}, \end{aligned} \tag{2.2}$$

where $\mathcal{L}_{VAE}$ is the loss function we need to minimize in VAE, and $q_\phi(z|x)$ is an approximate representation of the intractable $p_{\theta_1}(z|x)$ parameterized by q_ϕ. The output distribution in the first term is often Gaussian as:

(2.3)
$$p_{\theta_1}(x|z) = \mathcal{N}(x; f_{\theta_1}(z), \sigma^2 I) = C \times \exp\left(-\frac{(x - f_{\theta_1}(z))^2}{2\sigma^2}\right)$$

where C is a constant, and $f_{\theta_1}(\cdot)$ is computed by CNNs [12]. Therefore, the log likelihood can be expressed as:

$$\begin{aligned} \log p_{\theta_1}(X|z) &= \sum_{i=1}^{N} \log C \times \exp\left(-\frac{(x^{(i)} - f_{\theta_1}(z^{(i)}))^2}{2\sigma^2}\right), \\ &= N \times C - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (x^{(i)} - f_{\theta_1}(z^{(i)}))^2, \end{aligned} \tag{2.4}$$

where $N \times C$ is a constant that is irrelevant to $f_{\theta_1}(\cdot)$ and can be ignored in optimization. The first term in $\mathcal{L}_{VAE}$ is a ℓ_2 loss between x and $f_{\theta_1}(z)$. The second term corresponds to the Kullback-Leibler (KL) divergence between $q_\phi(z|x)$ and $p_{\theta_1}(z)$. VAE assumes that $q_\phi(z|x) = \mathcal{N}(z; \mu_\phi(x), \sum_\phi(x))$ and $p_\theta(z) = \mathcal{N}(z; 0, I)$. $\mu_\phi(x)$ and $\sum_\phi(x)$ are also implemented by CNNs. The second term in $\mathcal{L}_{VAE}$ can be considered as a prior regularization. Therefore, the loss function of VAE can be written as

$$\mathcal{L}_{VAE} = \mathcal{L}_{\ell_2} + \mathcal{L}_{prior}, \tag{2.5}$$

where

$$\begin{aligned} \mathcal{L}_{\ell_2} &= -\mathbb{E}_{q_\phi(z|x)}[\log p_{\theta_1}(x|z)] \\ &= \frac{1}{2\sigma^2} \sum_{i=1}^{N} (x^{(i)} - f_{\theta_1}(z^{(i)}))^2, \end{aligned} \tag{2.6}$$

$$\mathcal{L}_{prior} = D_{KL}[q_\phi(z|x)|p_{\theta_1}(z)]. \tag{2.7}$$

2.2 Deep Residual Variational Auto-Encoder VAE has shown promising results in image generation tasks[17, 18, 19]. However, the images generated by VAE are blurry. This is caused by the ℓ_2 loss, which is based on the assumption that the data follow a single Gaussian distribution. When samples in dataset have multi-modal distribution, VAE cannot generate images with sharp edges and fine details. In

Copyright © 2019 by SIAM
Unauthorized reproduction of this article is prohibited

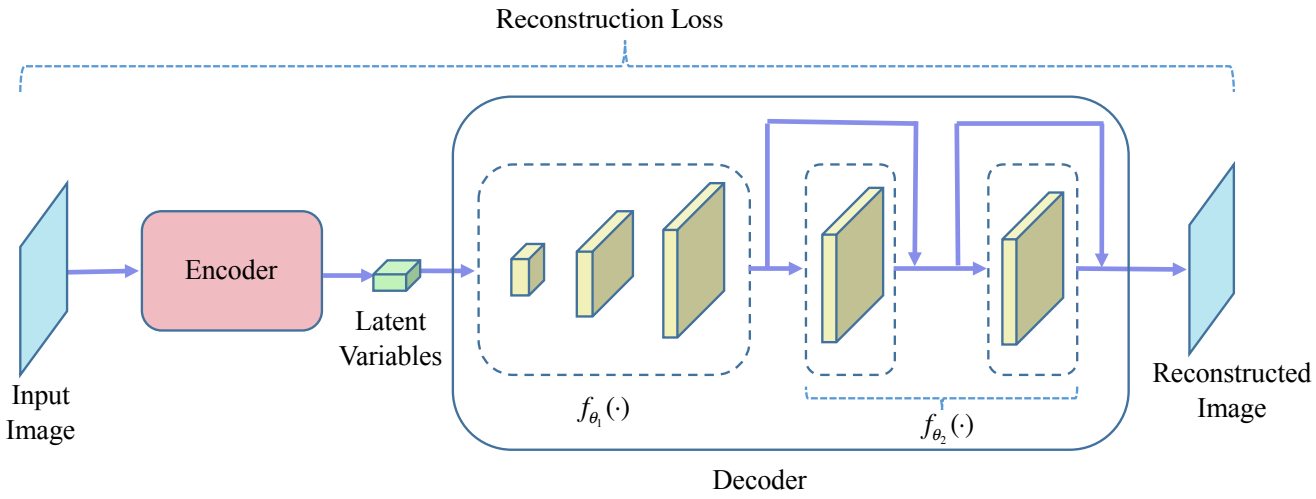

Figure 2: The network architecture of deep residual VAE. In this model, the encoder takes images as input and generate latent variables. The latent variables are fed into decoder network to recover the original spatial information. To make the decoder generate better image, we concatenate the original decoder network $f_{\theta_1}(\cdot)$ with the residual block network $f_{\theta_2}(\cdot)$ to increase the capacity of model.

VAE, images are generated by $f_{\theta_1}(\cdot)$. It is possible to generate better images by using more complex model for $f_{\theta_1}(\cdot)$. One solution is to employ the autoregressive model [20][21] for decoder function $f_{\theta_1}(\cdot)$. In the autoregressive model, each pixel is conditioned on previously generated pixels. The autoregressive model increases the dependency between pixels and generates images with fine details. However, since it must generate images pixel by pixel, the prediction procedure of autoregressive model is much slower compared with other generative models such as VAE.

Since the decoder of VAE is implemented with CNNs, a direct way to generate better images is to employ deeper networks, resulting in increased capacity of the decoder model [13]. The difficulty that deep neural networks facing is the degradation problem. As the network depth increases, the performance of deep networks initially improves and then degrades rapidly. Although deep neural network models with higher capacity usually yield better performance, it is also challenging to optimize them. To efficiently train deep neural networks, the batch normalization method is proposed in [22] by reducing internal covariate shift. Another solution is the residual learning framework proposed in [23], which employs the residual blocks and skip connection to back-propagate the gradients more efficiently in the network. The introduction of skip connection and residual block makes the optimization of deep neural networks more efficient. It is possible to employ deeper neural works on complex tasks. The residual learning framework has already been successfully applied to image recognition, object detection, and image super-resolution. To increase the capacity of decoder in VAE and optimize the model efficiently, we concatenate the original VAE decoder with several residual blocks. The architecture of deep residual VAE is illustrated in Figure 2. Given the original decoder $f_{\theta_1}(z)$, the deeper decoder networks can be denoted as $f_\theta = f_{\theta_2}(f_{\theta_1}(z))$, where $f_{\theta_2}(\cdot)$ corresponds to the residual network. Compared with the original VAE decoder, the deeper decoder networks can capture more details. The loss function of deep residual VAE can be written as:

$$\mathcal{L}_{RSVAE} = \mathcal{L}_{\ell_2} + \mathcal{L}_{prior}, \tag{2.8}$$

where

$$\begin{aligned} \mathcal{L}_{\ell_2} &= -\mathbb{E}_{q_\phi(z|x)}[\log p_{\theta_1}(x|z)] \\ &= \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x^{(i)} - f_{\theta_2}(f_{\theta_1}(z^{(i)})))^2, \qquad (2.9) \\ \mathcal{L}_{prior} &= D_{KL}[q_\phi(z|x)|p_{\theta_1}(z)]. \qquad (2.10) \end{aligned}$$

2.3 Multi-Stage Variational Auto-Encoder Experiment results in Section 3 show that deep residual VAE can capture more details than the original VAE by adding residual blocks to the decoder network. But the performance of deep residual VAE saturates rapidly as more residual blocks are added. As the depth of decoder network increases, the quality of generated images improves with smaller and smaller margins. This saturation effect is not a surprise as the network still employs ℓ_2 loss and thus generates blurry images. On the other hand, it is natural to use a step-by-step procedure to generate high-quality images[24]. Specifically, in image generation, we can generate a coarse image with

Copyright © 2019 by SIAM
Unauthorized reproduction of this article is prohibited

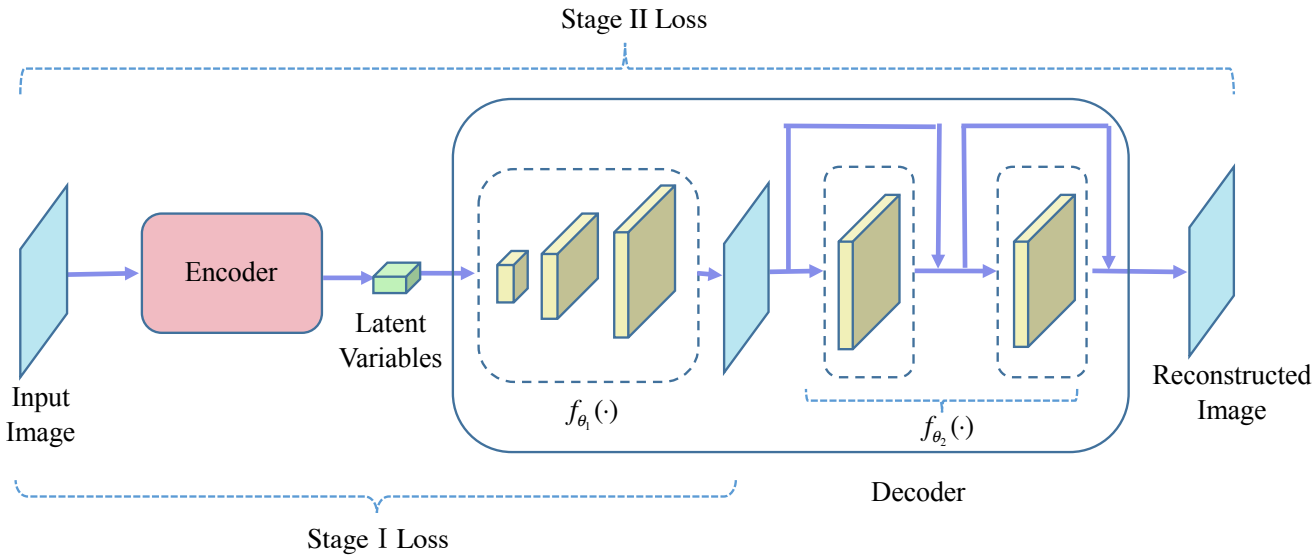

Figure 3: The network architecture of multi-stage VAE based on the deep residual VAE. In the first stage, the sub-network $f_{\theta_1}(\cdot)$ generates a coarse image $f_{\theta_1}(z)$. In the second stage, the coarse image $f_{\theta_1}(z)$ is fed into the model $f_{\theta_2}(\cdot)$ to produce a fine image $f_{\theta_2}(f_{\theta_1}(z))$.

rough shape and basic colors first and then refine the coarse image to a high quality one. In VAE, the decoder network is trained end-to-end. Thus we cannot control the process of image generation. To make the decoder network generate images step-by-step, we need to divide the decoder network into two components, where the first component generates a coarse image, and the second component refines it to a high quality one. To achieve this, we propose to add a loss function at some location in the decoder network and enforce the network to generate images at that location.

Here we use two stage deep VAE to illustrate how this idea works. Since in the first stage we only need to generate a coarse image, it is possible for the original VAE to accomplish this using the decode function $f_{\theta_1}(\cdot)$. Then we need to build a model to refine the coarse images. When we require the sub-network $f_{\theta_1}(\cdot)$ in the decoder of deep residual VAE to generate a coarse image, the input of $f_{\theta_2}(\cdot)$ is not some arbitrary intermediate feature maps but a coarse image. In this way, the sub-network $f_{\theta_2}(\cdot)$ acts as a model to refine the coarse images generated from $f_{\theta_1}(z)$. The architecture of the proposed multi-stage VAE is illustrated in Figure 3. The loss of the multi-stage VAE can be written as:

$$\begin{aligned}\mathcal{L}_{MSVAE} = &-\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] + D_{KL}[q_\phi(z|x)|p_\theta(z)] \\ &+ \mathcal{L}_{rf}(x, f_{\theta_2}(f_{\theta_1}(z))). \end{aligned} \tag{2.11}$$

Compared with deep residual VAE, multi-stage VAE has two cost functions in the decoder network. The cost function of the first stage corresponds to $-\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)]$ in the original VAE, and it is used to generate coarse images. The cost function of the second stage corresponds to the third term in Equation 2.11, and it is used to refine the coarse images. In multi-stage VAE framework, the second network is independent of the VAE model. Therefore, we can employ loss function on $\mathcal{L}_{rf}(x, f_{\theta_2}(f_{\theta_1}(z)))$. It also overcomes the effect of ℓ_2 loss under the assumption that data have a single Gaussian distribution. By employing different loss functions, the second model can recover more detailed information from blurry images. The $\mathcal{L}_{MSVAE}$ can be written as:

$$\mathcal{L}_{MSVAE} = \mathcal{L}_{\ell_2} + \mathcal{L}_{prior} + \mathcal{L}_{rf}(x, f_{\theta_2}(f_{\theta_1}(z))), \tag{2.12}$$

where

$$\begin{aligned} \mathcal{L}_{l_2} &= -\mathbb{E}_{q_\phi(z|x)}[\log p_{\theta_1}(x|z)] \\ &= \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x^{(i)} - f_{\theta_1}(z^{(i)}))^2, \end{aligned} \tag{2.13}$$

$$\mathcal{L}_{prior} = D_{KL}[q_\phi(z|x)|p_{\theta_1}(z)]. \tag{2.14}$$

In addition, generating higher resolution images (e.g., 128×128) is challenging for generative models. In multi-stage VAE, the coarse images generated in the first stage provide additional information and subsequently enables the multi-stage VAE to generate high-resolution images. The idea of tackling complex tasks in a multi-stage manner is also employed by Stack GAN [24]. Stack GAN employs two separate models to generate low-resolution images and high-resolution images, respectively. The two models are trained separately. However, our model divides the decoder network

Copyright © 2019 by SIAM
Unauthorized reproduction of this article is prohibited

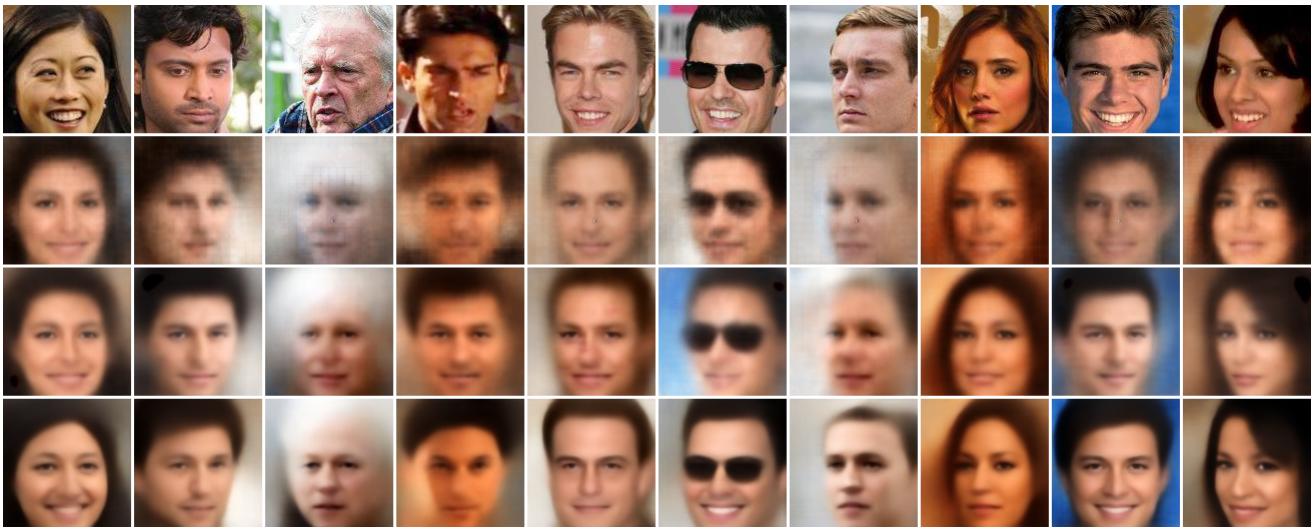

Figure 4: Comparison of reconstructed images from the CelebA dataset. The first row is the input images in the CelebA training set. The second row is the reconstructed images generated by the original VAE. The third and fourth rows are the results of deep residual VAE and multi-stage VAE respectively.

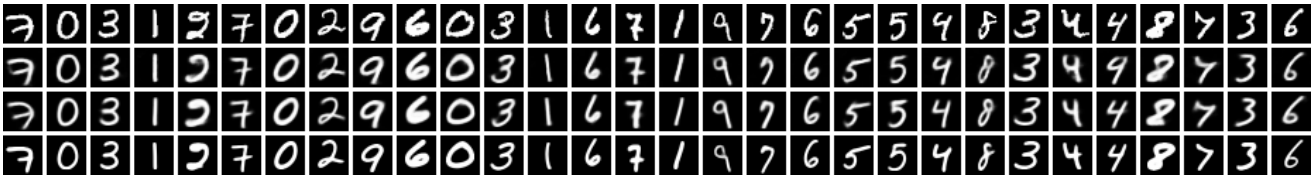

Figure 5: Comparison of reconstructed images from the MNIST dataset. The first row is the input images from the MNIST training set. The second row is the reconstructed images generated by the original VAE. The third and fourth rows are the results of deep residual VAE and multi-stage VAE, respectively.

into two components with different loss functions, and both networks are trained jointly.

2.4 Connections with Super-Resolution We employ residual networks in the second stage of our multi-stage VAE to refine the coarse images generated in the first stage. The key idea of the second model is similar to the super-resolution residual net (SRResNet) [25]. In SRResNet, a low-resolution image is fed into a network composed of residual blocks and up-sampling layers. Then an image with high resolution is generated by SRResNet.

In multi-stage VAE, we employ a pixel-wise loss function to recover the details between low-resolution images and high-resolution images. Minimizing the pixel-wise loss encourages the model to generate the average of plausible solutions, thus leading to poor perceptual quality [17, 26]. A plausible loss function applied in image super-resolution tasks is the combination of Euclidean distances in feature space and adversarial loss. In fact, our multi-stage VAE framework can work with any plausible super-resolution model by replacing the loss function in $\mathcal{L}_{rf}$ and the model architecture of $f_{\theta_2}(\cdot)$.

3 Experiments

In this section, we evaluate the deep residual VAE and multi-stage VAE[1] on the MNIST and CelebA datasets. Since the evaluation methods cannot guarantee the performance of generative models, we compare the quality of generated images with the original VAE [27]. In addition, we employ the structural similarity (SSIM) [28] to measure the similarity between generated images and real images. Results show that the proposed multi-stage VAE generates higher-resolution images as compared to those generated by the original VAE and deep residual VAE.

3.1 Settings CelebA [29] is a large scale face dataset that contains $202,599$ face images. The size of each face image is 178×218. Most prior VAE work using this dataset crops the images to 64×64. In order to demonstrate the performance of our multi-stage VAE in generating high-

[1] https://github.com/divelab/msvae/

Copyright © 2019 by SIAM
Unauthorized reproduction of this article is prohibited

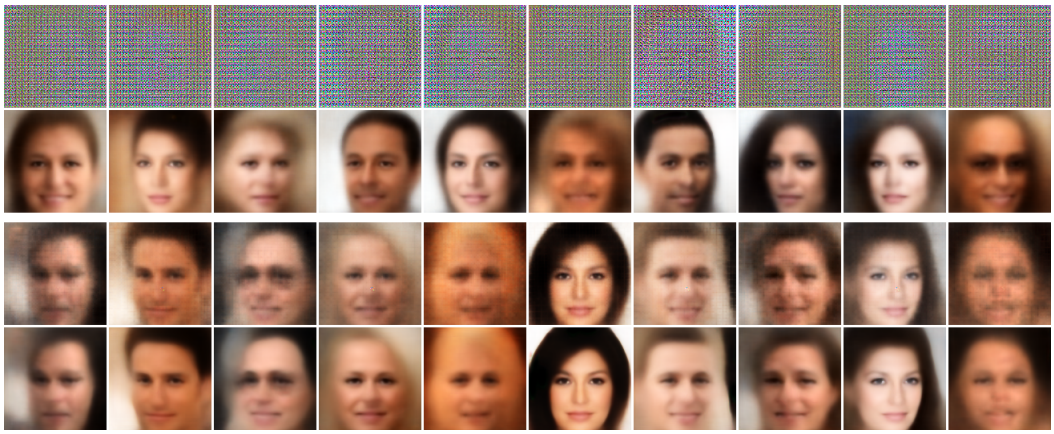

Figure 6: Illustration of decoder outputs on the CelebA dataset. The first and third rows are the output of $f_{\theta_1}(\cdot)$ in deep residual VAE and multi-stage VAE, respectively. The second and fourth rows are the outputs of $f_{\theta_2}(\cdot)$ in deep residual VAE and multi-stage VAE, respectively.

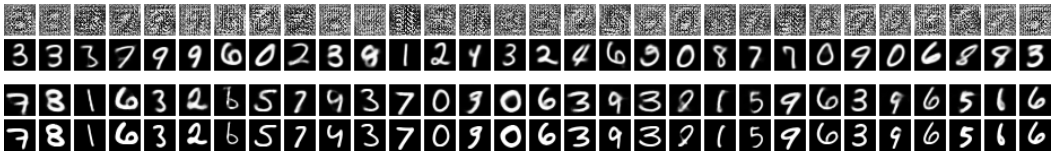

Figure 7: Illustration of decoder output on the MNIST dataset. The first and third rows are the outputs of $f_{\theta_1}(\cdot)$ in deep residual VAE and multi-stage VAE, respectively. The second and fourth rows are the outputs of $f_{\theta_2}(\cdot)$ in deep residual VAE and multi-stage VAE, respectively.

resolution images, we crop the image to 128×128. We train three models for 200000 iterations. with batch size of 32 and a learning rate of 2e-4. The encoder model of VAE consists of four layers. Each layer consists of a convolution layer with stride 1 followed by a convolution layer with stride 2. The latent variable size of VAE is 512 for the CelebA dataset. The decoder network consists of four deconvolution layers. To generate images with high quality, five residual blocks are employed in the decoder network. The ℓ_1 loss is used in the objective function of the second network.

The MNIST is a handwritten digits dataset where the size of each image is 28×28. We train three models on the training set of $60,000$ images. Each model is trained for $100,000$ iterations with a batch size of 256 and a learning rate of 1e-3. The encoder model of VAE consists of three convolution layers with a stride of 2. The latent variable size of VAE is set to 128. The decoder reconstructs the image from the latent variable with three deconvolution layers. To increase the complexity of decoder network, we concatenate the original VAE decoder with five residual blocks in the deep residual network. Each residual block consists of two convolution layers followed by a batch normalization layer. In multi-stage VAE, we add an ℓ_2 loss function at the location of the output in the original VAE. The residual network is employed to refine the coarse images generated in the first stage. To overcome the blurry effect of ℓ_2 loss, we employ ℓ_1 loss in the objective function of the second network.

Table 1: Comparison of SSIM by three models

Model	CelebA	MNIST
VAE	0.646	0.834
Residual-VAE	0.648	0.839
MSVAE	0.690	0.836

3.2 Quantitative Results and Analysis The goal of our proposed method is to improve the quality of generated images and encourage generated images to be close to the original images. Therefore, we employ the structural similarity (SSIM) to measure the similarity between generated images and real images as an evaluation of our model. We compute SSIM on the dataset, and the results are shown in Table 1. We can observe from the results that the images generated by residual VAE are closer to the original images than those of VAE. We employ multi-stage loss in our model without adding extra parameters. The experimental results show that

Copyright © 2019 by SIAM
Unauthorized reproduction of this article is prohibited

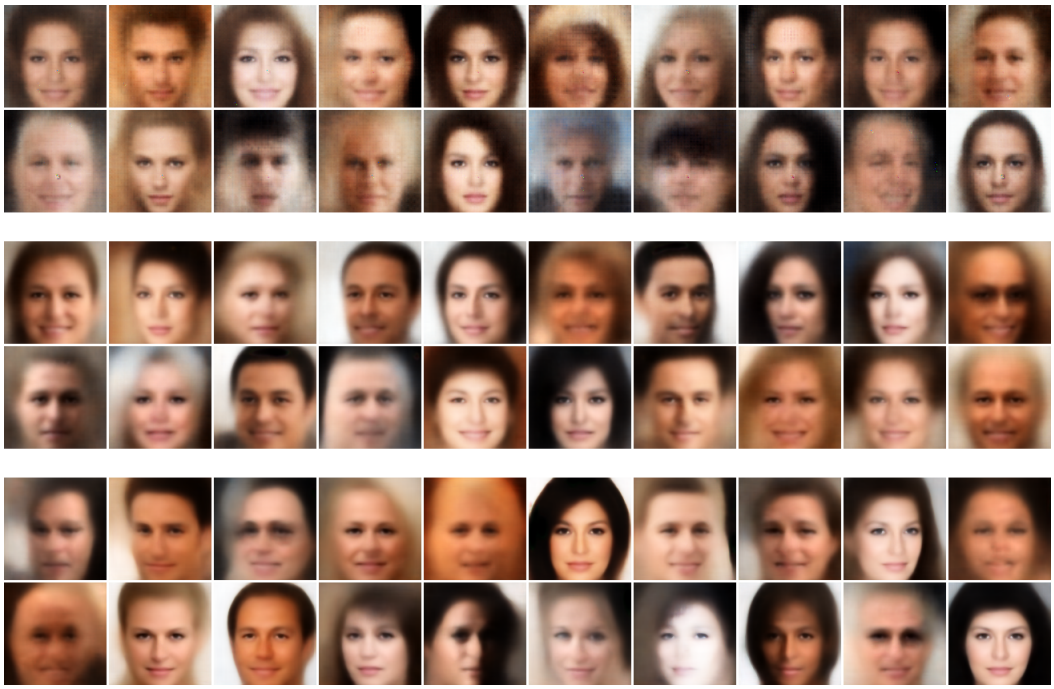

Figure 8: Sample images generated by different models when trained on the CelebA dataset. The three parts correspond to standard VAE, residual VAE and multi-stage VAE.

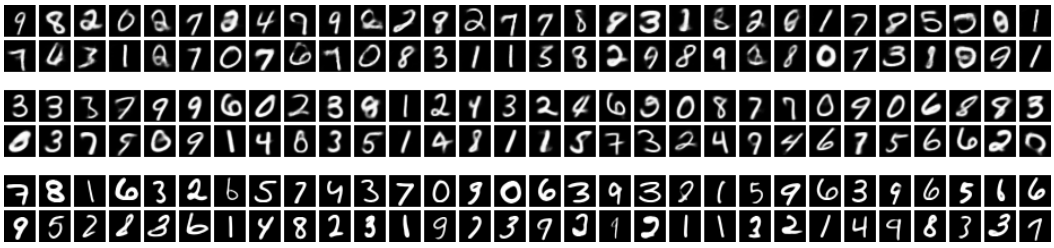

Figure 9: Sample images generated by different models when trained on the MNIST dataset. The three parts correspond to standard VAE, residual VAE and multi-stage VAE.

our proposed model achieves better performance than VAE and residual VAE.

3.3 Qualitative Results and Analysis Figures 4 and 5 provide some reconstructed images by different models. We can see that the deep residual VAE can capture more details than the original VAE by employing more complex decoder network. However, the images generated by deep residual VAE are still blurry due to the effect of ℓ_2 loss. We also observe that the effect of ℓ_2 loss is largely overcome by employing the multi-stage loss. The blurry region becomes clearer through the multi-stage refine process. These results demonstrate that the proposed multi-stage VAE goes beyond the bottleneck of increasing the capacity of decoder network, thereby effectively overcoming the blurry effect caused by the ℓ_2 loss.

Figures 6 and 7 provide some reconstructed images and intermediate outputs of $f_{\theta_1}(\cdot)$ by the deep residual VAE and multi-stage VAE. We can see that at the intermediate location in the decoder network of multi-stage VAE, a blurry image is generated, and it is fed into the residual networks. Through the refined operation of the second network, an image with high quality is generated. Since the whole decoder network of deep residual VAE only contains a single loss function, the generation process suffers from the effect of ℓ_2 loss. Therefore, the images generated by deep residual VAE are still blurry.

Figures 8 and 9 provide some sample images generated by the original VAE, deep residual VAE, and multi-stage VAE when the models are trained on the CelebA and MNIST datasets. We can see that the images generated by the multi-stage VAE have higher resolution than those generated by

Copyright © 2019 by SIAM
Unauthorized reproduction of this article is prohibited

other two methods. Also the images generated by the deep residual VAE are clearer than those generated by the original VAE. These results demonstrates that the proposed multi-stage VAE is effective in generating high resolution images.

4 Conclusion and Future Work

In this work, we propose a multi-stage VAE that can generate higher quality images than the original VAE. The original VAE always generated blurry images due to the effect of ℓ_2 loss. To generate high quality images, we propose to improve the decoder capacity by increasing the network depth and employing residual blocks and skip connection. Although the deep residual VAE can capture more details, it still suffers from the effect of ℓ_2 loss and generates blurry images. To overcome the limitation of ℓ_2 loss, we propose to generate images from coarse to fine. To achieve this goal, we require the decoder network to generate a coarse image by employing a ℓ_2 loss function in the first stage. The subsequent stage in the decoder network acts as a super-resolution network that takes a blurry image as input and generates a high quality image. Since the super-resolution network is independent of the VAE model, it can employ other loss functions to overcome the the effect of ℓ_2 loss, thereby generating high quality images. Experimental results on the MNIST and CelebA datasets show that the proposed multi-stage VAE can overcome the effect of ℓ_2 loss and generate high quality images.

One interpretation of our proposed framework is that, the network in the second stage can be considered as a super-resolution module. Following this interpretation, we plan to use other model architectures and loss functions commonly used for super-resolution, such as the adversarial loss [25]. As has been mentioned, the proposed multi-stage framework can be generalized to more than two components. We plan to explore more stages in the future[16].

Acknowledgments

This work was supported in part by National Science Foundation grants DBI-1661289 and CHE-1738305.

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[2] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," in *9th ISCA Speech Synthesis Workshop*, pp. 125–125.

[3] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real nvp," *arXiv preprint arXiv:1605.08803*, 2016.

[4] G. E. Hinton and T. J. Sejnowski, "Learning and releaming in boltzmann machines," *Parallel Distrilmted Processing*, vol. 1, 1986.

[5] Y. Bengio, E. Laufer, G. Alain, and J. Yosinski, "Deep generative stochastic networks trainable by backprop," in *International Conference on Machine Learning*, 2014, pp. 226–234.

[6] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *Proceedings of The 31st International Conference on Machine Learning*, 2014, pp. 1278–1286.

[7] C. Doersch, "Tutorial on variational autoencoders," *arXiv preprint arXiv:1606.05908*, 2016.

[8] I. Gulrajani, K. Kumar, F. Ahmed, A. A. Taiga, F. Visin, D. Vazquez, and A. Courville, "Pixelvae: A latent variable model for natural images," *arXiv preprint arXiv:1611.05013*, 2016.

[9] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling, "Improving variational autoencoders with inverse autoregressive flow," in *Advances In Neural Information Processing Systems*, 2016, pp. 4736–4744.

[10] X. Chen, D. P. Kingma, T. Salimans, Y. Duan, P. Dhariwal, J. Schulman, I. Sutskever, and P. Abbeel, "Variational lossy autoencoder," *arXiv preprint arXiv:1611.02731*, 2016.

[11] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[14] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[15] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1646–1654.

[16] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[17] A. Dosovitskiy and T. Brox, "Generating images with perceptual similarity metrics based on deep networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 658–666.

[18] S. Zhao, J. Song, and S. Ermon, "Towards deeper understanding of variational autoencoding models," *arXiv preprint arXiv:1702.08658*, 2017.

[19] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *Proceedings of The 33rd International Conference on Machine Learning*, 2016, pp. 1558–1566.

[20] A. v. d. Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," *arXiv preprint arXiv:1601.06759*, 2016.

Copyright © 2019 by SIAM
Unauthorized reproduction of this article is prohibited

[21] A. van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves *et al.*, "Conditional image generation with pixelcnn decoders," in *Advances in Neural Information Processing Systems*, 2016, pp. 4790–4798.

[22] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[24] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," *arXiv preprint arXiv:1612.03242*, 2016.

[25] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," *arXiv preprint arXiv:1609.04802*, 2016.

[26] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," *arXiv preprint arXiv:1511.05440*, 2015.

[27] L. Theis, A. v. d. Oord, and M. Bethge, "A note on the evaluation of generative models," *arXiv preprint arXiv:1511.01844*, 2015.

[28] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[29] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

Copyright © 2019 by SIAM
Unauthorized reproduction of this article is prohibited