EECS 6322 Week 8, Paper 2
NVAE: A Deep Hierarchical Variational Autoencoder

Shivani Sheth (Student #: 218011783)
shivs29@yorku.ca

I. Summary and contributions

The paper proposes a deep hierarchical VAE architecture to generate high images using batch normalization and depth-wise separable convolutions. It parameterised the residual connections of the Normal distributions and stabilizes its training using spectral regularizations. The architecture is first to generate high quality images as large as 256×256 pixels using VAEs. Further, it also produces strong results as compared to the other non-autoregressive likelihood-based models on datasets such as, CIFAR-10, CelebA HQ, CelebA 64, MNIST, and FFHQ.

The neural network architectures for VAEs have been long overlooked and more work on the statistical domain of the VAEs have come forward. However, the neural network architectures for VAEs are found to be suitable for different tasks in deep learning for a variety of reasons. Firstly, the VAE models retain as much information as possible in the latent vectors as they aim to maximize the mutual information between the input data and the latent vectors. This is in contrast to other classification models that discard information from the input data. Secondly, over-parameterizing may lead to the reduction in the amortization gap which might help the encoder to build better models, whereas the decoder part of the VAE model may actually harm from over-parameterizing. The VAEs also have large receptive fields that enable them to capture the long range dependencies. Hence, the authors propose VAEs as a useful model for the image generation tasks.

The architecture proposed by the authors make use of depth-wise convolutions that significantly increase the receptive field of the network, while restricting the increase in the parameters. The architecture also includes Batch Normalization, Spectral Regularization, and residual parameterization of the approximate posterior parameters to increase the stability of the VAE model. The architecture utilizes approximate posterior distribution or encoder in the model instead of the true posterior since the true posterior is intractable. The approximate posterior is given by $q(z|x) = \prod_l \ q(z_l|z_{<l}, x)$ *which are represented by factorial Normal distributions. The architecture follows a top-down approach that generates the parameters of each conditional probabilities and at each level a sampling is done from each group that combines with the deterministic feature maps and is passed onto the next group. A bottom-up deterministic network is also used to infer variables from* $q(z|x)$ *and extract representation from input* x*. The representations extracted in the top-down model are reused in the generation process in order to avoid extra computational costs.*

The proposed model uses a hierarchical multi-scale VAE model to capture the long range dependencies of the input data. It initially generates a small spatially arranged latent variable as z_1 *and gradually doubles the spatial dimensions by sampling group-by-group from the hierarchy. This multi-scale approach enables NVAE to capture global long-range correlations at the top of the hierarchy and local fine-grained dependencies at the lower groups. The long-range dependencies of the model can also be increased by increasing the receptive field of the networks, which can be done by increasing the kernel sizes in the convolutional path. A (1 × 1) regular convolution is used as a bottleneck to reduce the parameter size and computational costs. The architecture also uses a modified version of the Batch Normalization where the momentum parameter of BN is modified to catch up faster with the batch statistics and an additional regularization is applied to ensure that the mismatches are not amplified by the BN.*

Additionally, the model uses Swish Activations, Squeeze and Excitation (SE), and residual cells with depthwise convolutions that are similar to MobileNetV2 but have two additional BN layers at the beginning and the end of the cell and additionally use Swish activation function and SE. The residual connections are also used in the bottom-up models to increase its performance. To reduce the memory requirement, the model is defined in mixed-precision using the NVIDIA APEX library, which enables the model to reduce the GPU memory by 40%. To reduce memory, an additional gradient checkpointing method is also used which computes the Batch Norm in the backward pass. To improve the KL optimization and stabilize the training, Residual Normal Distributions and Spectral Regularization (SR) methods are proposed and used by the authors, which help in optimizing the hierarchical structure of the model.

The model was experimented on several datasets such as the MNIST, CIFAR-10, ImageNet, CelebA, CelebA HQ, and FFHQ, where the NVAE model outperforms previous non-autoregressive models IAF-VAE [4] and BIVA [36] on most datasets and reduces the gap with autoregressive models. The model also achieves very competitive performance compared to the autoregressive models. It is noted that the model without

flows outperforms many existing generative models by itself and hence it indicates that the network architecture is an important component in VAEs and the Normal distributions in the encoder can compensate for some of the statistical challenges. In the ablation studies, each component of the model is studied independently as well as in relation with the other components of the model. These components include Normalization and Activation Functions, Residual Cells, Residual Normal Distributions, The Effect of SR and SE, Sampling Speed, Posterior Collapse, and Reconstruction. It is found that a combination of all these components are required for stabilizing the model and hence contributes to the competitive efficiency of VAEs in the image generation domain.

II. Strengths

The paper proposes a novel NVAE architecture which is a hierarchical structure of a VAE with some modified components. This is a novel architecture that is able to outperform many of the state of the art models in the image generation domain. The main strengths of the model include its ability to capture long range dependencies, lower memory and computational requirements by shared latent vectors, and residual parameterization of Normal distributions in the encoder along with spectral regularization for stabilizing the training of very deep models. It is also one of the first models to produce large high-quality images and is trained without changing the objective function of VAEs.

III. Weaknesses

The model produces an enhanced version of the reconstructed pictures such as human faces with an increase in smoothness which can be seen as a little less real as compared to the results produced by other models. It is also competitive to the autoregressive models but is not yet able to outperform them on datasets such as ImageNet.

IV. Correctness

The claims and empirical methodology given by the authors of the paper are supplemented by strong theoretical grounding and comparisons with the current state of the art methods that prove that they are correct. The model is also tested on several datasets and it is found that the modifications and the hierarchical structure of the model helps the model gain better efficiency as compared to its counterparts.

V. Clarity

The paper is well written in general with sufficient comparison tables and illustrations of the model architectures wherever required. The concepts have also been explained crisply along the paper, while elaborating them in the appendix that enables the reader to understand well.

VI. Relation to prior work

Prior work in the field of VAEs neural network architectures include VQ-VAE-2, IAF-VAEs, BIVA, DRAW, and Conv DRAW models. The VQ-VAE-2 model utilizes PixelCNN in its prior for latent variables up to 128×128 dimensions, whereas the NVAE utilizes an unconditional decoder in the data space. The VQ-VAE's objective also does not correspond to a lower bound on data log-likelihood unlike the NVAEs. Although the NVAEs are related to the second IAF-VAEs models, they differ from those models in terms of parameterizations, neural network implementation, and scaling up the images. The NVAEs also differ from the third BIVA models since the BIVA models are trained on images as large as 64×64 px and use neural networks similar to IAF-VAEs. Finally, the NVAEs differ from DRAW, and Conv DRAW models since they use RNNs or recurrent neural networks to model hierarchical dependencies, while the NVAEs have a different regularization, parameterization and a different neural network architecture as compared to the previous models.

VII. Reproducibility

Enough details are provided by the authors of the paper in terms of github links and the theoretical structure of the model to implement the major details of the work. The architectural complexities have also been explained in the paper which make it easier to implement the model and compare it with the previous state of the art results.

VIII. Additional comments

An extended study with image datasets in different domains related to animals, birds, etc. would have helped in better understanding of the various strengths of the model.