Variational Autoencoder: A Deep Generative model

Ayush Raina
IISc Bangalore
ayushraina@iisc.ac.in

Arnav Bhatt
IISc Bangalore
arnavbhatt@iisc.ac.in

Anushka Dassi
IISc Bangalore
anushkadassi@iisc.ac.in

Abstract

Autoencoders are deep **generative** models. They are widely used for tasks such as image generation, data compression, denoising and capturing the most important features of the data. Autoencoders contains encoder and decoder networks. Variational Autoencoders introduce **probabilistic modelling** into encoding process in which we learn the latent distribution, which enables generation of data which are similar to training data.

1 Introduction

1.1 Autoencoder

This is a neural network which consists of two parts: **Encoder(ϕ),Decoder(θ)** (where ϕ, θ are the parameters)and bottleneck layer between these two,whose dimension is much less than input layer. Encoder network learns the latent representation of the input data and decoder network learns to reconstruct the input data from the latent representation. The loss function is defined as the pixelwise difference between input and output data.

1.2 Aim

We want output of the network as input itself. The main challenge is that input has to pass from a bottleneck layer whose dimension is much less than input layer. We can say that the network learns to compress the data to hidden state.

1.3 Objective Function

In this case we are dealing with images and we want our output image to be as similar as input image. Since our image passes through encoder and decoder network we can set up the loss function as follows:

$$min \ \sum_{i=1}^{m} || \ x^{(i)} - Decoder_\theta(Encoder_\phi(x^{(i)})) \ ||^2$$

2 Theorems and Results

2.1 Expectation Maximization(EM) Algorithm

EM algorithm is used to perform the maximum likelihood estimation in the presence of latent variables. This is a two step process consisting of E-step and M-step. Suppose ψ are the parameters we want to learn and $z^{(i)}$ are the latent variables. Then we perform these two steps iteratively until convergence:
E-step: For all i, set $Q_i(z^{(i)}) = p(z^{(i)}|x^{(i)};\psi)$, M-step: $\psi^{(new)}$ = $argmax_\psi \sum_{i=1}^{m} ELBO(x^{(i)};Q_i,\psi)$, where **ELBO** is called evidence lower bound. We cannot use EM algorithm here because in E-step, we assumed that the posterior distribution $p(z^{(i)}|x^{(i)};\psi)$ can be calculated, but sometimes it may be extremely complex to calculate this distribution, for example suppose $z \sim N(0, I_{k\times k})$ and $x|z \sim NeuralNetwork(z^{(i)};\theta)$, then it is almost impossible to calculate the posterior distribution $p(z^{(i)}|x^{(i)};\psi)$ because of the complexity of the neural network.
So we need to find a way to approximate this distribution. There are many ways to approximate this distribution, like **Gibbs sampling, Variational Inference**, etc. In this paper we will discuss about the variational inference.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

2.2 Variational Inference

We know that $log(P(x)) = ELBO(x;Q) + D_{KL}(Q||P_{z|x})$, where $ELBO(x;Q) = E_{z\sim Q}[log(\frac{log(P(x,z)}{Q(x)}]$ and $D_{KL}(Q||P_{z|x})$ is the **KL Divergence** between the Q and the true posterior distribution $P_{z|x}$. We want Q to be as close as possible to the true posterior distribution $P_{z|x}$, which mean $D_{KL}(Q||P_{z|x}) \longrightarrow 0$. So if we maximise the ELBO with respect to Q, then we are minimising the KL Divergence between Q and $P_{z|x}$, because the $log(P(x))$ is constant. So we can say that $P_{z|x} \sim argmax_{q\sim Q}ELBO(x;q)$, where we will assume all $q's$ are in the same family of distributions(Q) called as **Variational family**. These are distributions over vectors of dimension k. In our case we will also assume that Q is a family of Gaussian distributions.

2.3 Generation with Autoencoders

Till now we have seen that autoencoders take input and maps it to **hidden** representation(latent representation) and decoder reconstructs the input back from the hidden representation. After the training, we can remove the encoder part and feed random latent representation to the decoder to generate output. But we may not get meaningful outputs because latent representations lies in very small **subspace** of the input space.
To generate meaningful outputs we need to feed the latent variables which is similar to the training data. In other words we want latent variables to be sampled from $P(z|X)$.But autoencoders learn hidden representation z but not a distribution $P(z|X)$. Same is in the case of decoder every time we feed the same latent representation we get the same output, which means both encoder and decoder are deterministic in this case. Now we will look over Variational Autoencoders which have same structure but they learn the distribution over the latent variables.

3 Variational Autoencoders

We are interested in $P(z|X)$ which corresponds to learning hidden representation. Variational Autoencoders use encoders to learn this distribution by method of variational inference. We will assume that our variational family Q is Gaussian and encoder network will give us the mean and variance of this distribution. Hence $Q = N(\mu(X), \Sigma(X))$ where $\mu(X), \Sigma(X) = Encoder_{\phi}(X)$. Note that $\mu(X), \Sigma(X)$ are functions of input X. Once we learned this distribution we can sample from this distribution and feed this sample to the decoder to generate the output. Here is the visual representation of the model:

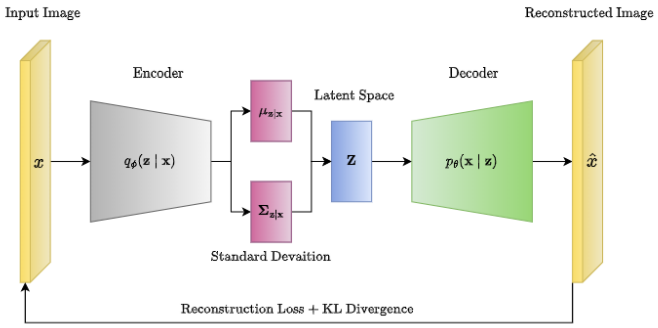

Figure 1: Variational Autoencoder

4 Experimental Results and Conclusions

Reconstructions on MNIST and Animal Face dataset

Here are the reconstructed outputs of autoencoder on 5 randomly choosen images. In first case we have used only **dense** layers and in second,third case we have used **convolutional** layers.Animal faces are reconstructed in Grayscale.

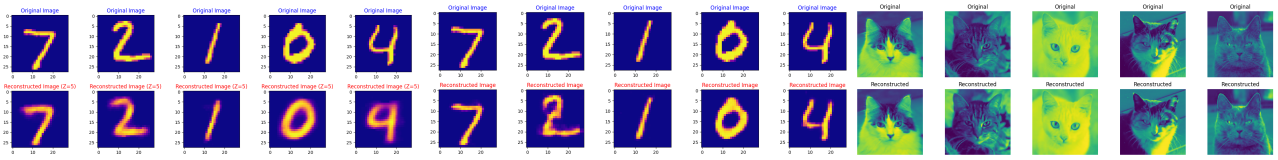

Figure 2: Reconstructions of Autoencoder without and with CNN

Experimental Results on Animal Face Dataset

After many experiments we observed that autoencoder was able to reconstruct the images well but was not able to generate new images due to reason that we have discussed earlier, where as in case of variational autoencoder, it was not able to reconstruct the images as well as autoencoder but was able to generate new images better than autoencoder.

Reconstructions and Generations by VAE on partial Animal Face Dataset

Here are the results when we feed separate images of cats. We have resized all images to 128×128 and trained for 40 epochs.

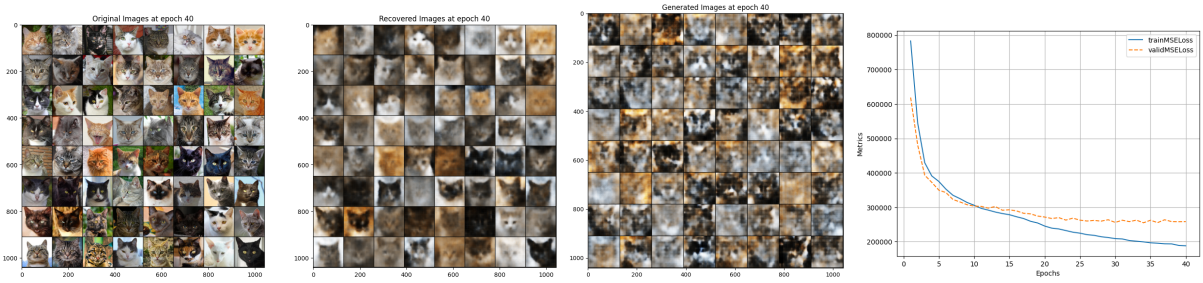

Figure 3: Reconstructions and Generations of Cats by Variational Autoencoder

Here are the results when we feed separate images of wild animals, this time for 70 epochs.

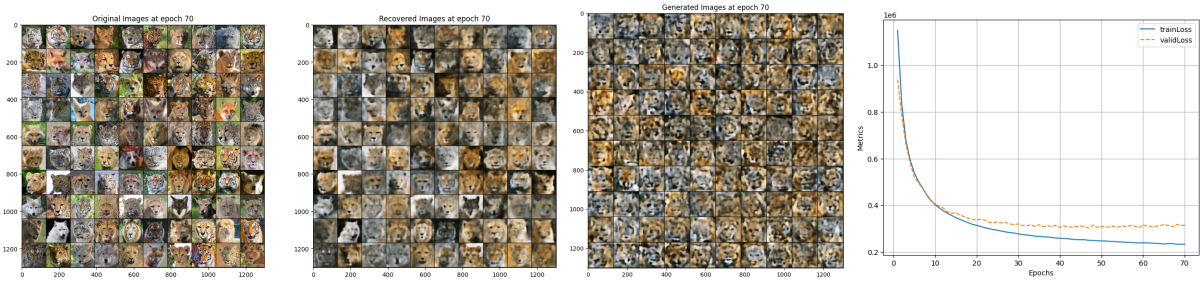

Figure 4: Reconstructions and Generations of Wild Animals by Variational Autoencoder

Reconstructions and Generations by VAE on full Animal Face Dataset

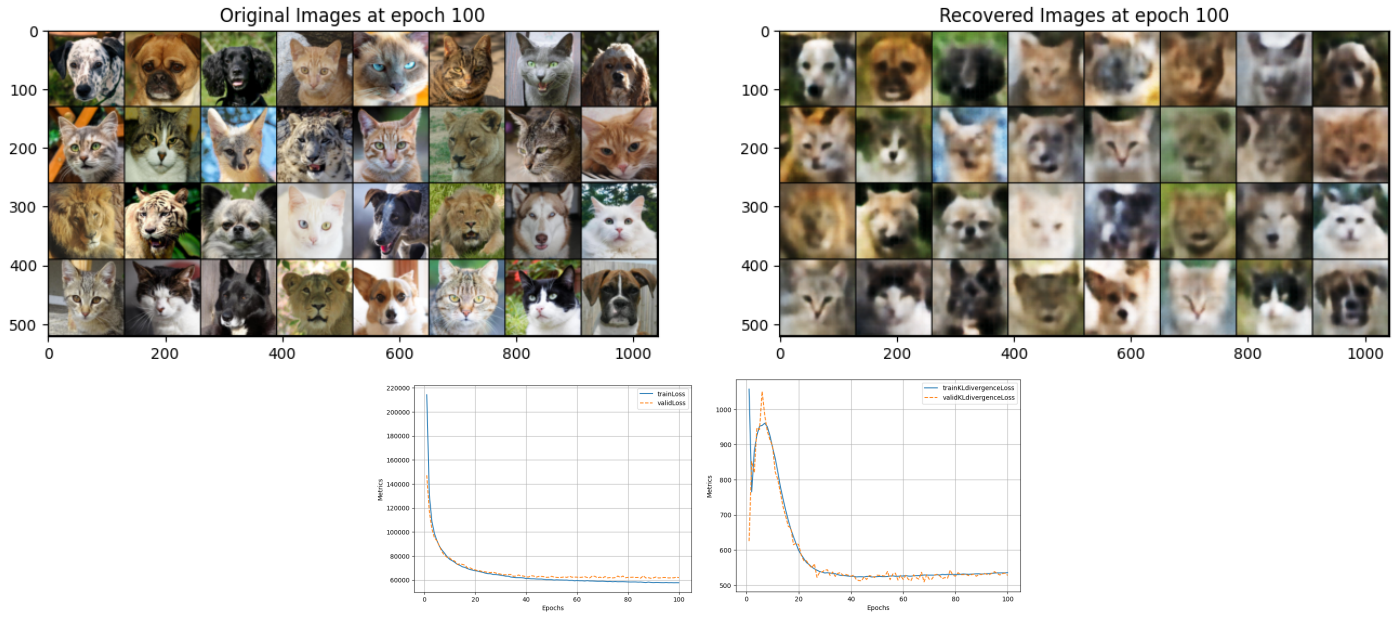

Figure 5: Reconstructions of Animal Faces by Variational Autoencoder

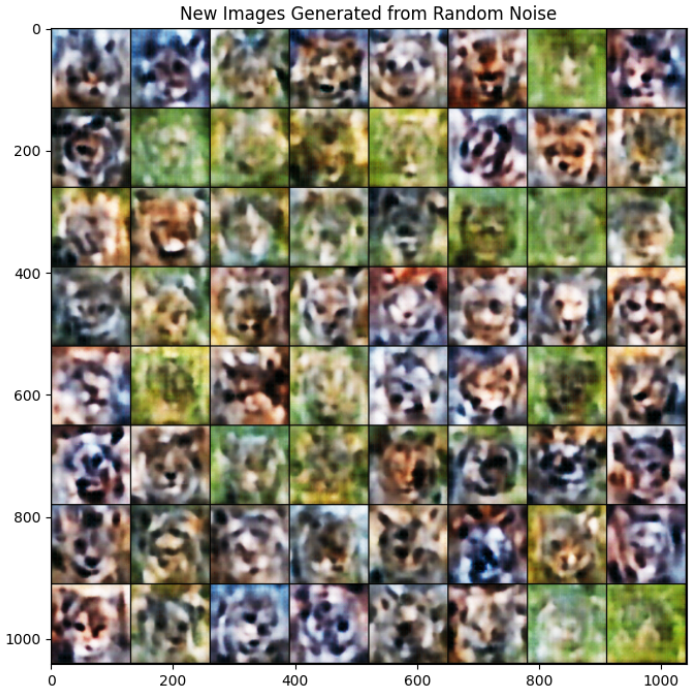

Figure 6: Generations of Animal Faces by VAE

```
ayushraina@administrator-B760M-DS3H-DDR4:~/Desktop/4th Semester/ML Assignm
ent/Term Paper/Codes/Final$ python3 -m pytorch_fid generateNew/ Reals/
100%|                                                  | 40/40 [00:07<00:00,  5.4
8it/s]
100%|
                                                       | 40/40 [00:08<00:00,  4.74it/s]
FID:  269.9804641747991
```

Figure 7: FID Score between real and generated images

5 Key Takeaways

Here are the main experimental observations:
1. Autoencoders can reconstruct the images pretty well than variational but they cannot generate better images.
2. During our experiments we have seen importance of learning rate, if not choosen properly loss may explode to infinity.
3. We also saw the importance of batch size, as it should be choosen precisely so that GPU memory does not get overloaded.We also saw that in VAE why it is important to minimize both KL Loss and MSE loss together.
4. We also designed deep neural networks with multiple layers and observed that deeper networks are able to learn better representations but they take more time to train.
5. We also saw how important is the KL weight in the loss function, higher the Kl more better is the generation but it may lead to poor reconstructions.

6 References

1. ELBO Surgery: Yet Another Way to Carve Up the Variational Evidence Lower Bound, `https://approximateinference.org/2016/accepted/HoffmanJohnson2016.pdf`
2. CS229 Lecture Notes for VAE, `https://cs229.stanford.edu/summer2019/cs229-notes8.pdf`
3. Tutorial on Variational Autoencoders, `https://arxiv.org/pdf/1606.05908.pdf`