Using Variational AutoEncoders for Image Completion

Bo Liu, Yunhan Zhao, Cuiqing Li

Department of Computer Science of the Johns Hopkins University

Introduction

Variational Autoencoder (VAE) is a type of interesting generative model for unsupervised learning of complicated distributions.

Unlike conventional autoencoder, counting on the expressiveness of its neural network encoder and decoder, VAE not only learns a compact hidden representation **z** for the input **X**, but also forces **z** to follow a simple distribution (e.g. a normal distribution) so that **X** can be easily sampled according to **P(X | z)**.

VAEs are widely used in computer vision. We found the generativity of VAE is particularly interesting and hypothesized that we can take advantage of its generativity to do image completion task. **The goal of the Project** then is to let VAE learn the generation process of a given type of images, then try to recover the occluded part of images by selecting the most promising reconstructed one. We developed 4 variations of VAEs and test their performance on occluded images from MNIST data.

Related Work

Variational Autoencoders were simultaneously discovered by Kingma et al and Rezende et al in 2014. Only in a few years, variational autoencoders become one of the most popular generative models used in computer vision.

According to Dr. Carl Doersch's tutorial on VAEs, we know that the model have already shown promise in generating many kinds of complicated data, including handwritten digits (Tim Salimans, et al 2015), faces (Tejas D Kulkarni, et al, 2015), house numbers (Diederik P Kingma, et al, 2014), CIFAR images (Karol Gregor, et al, 2015) and some other interesting real-world applications

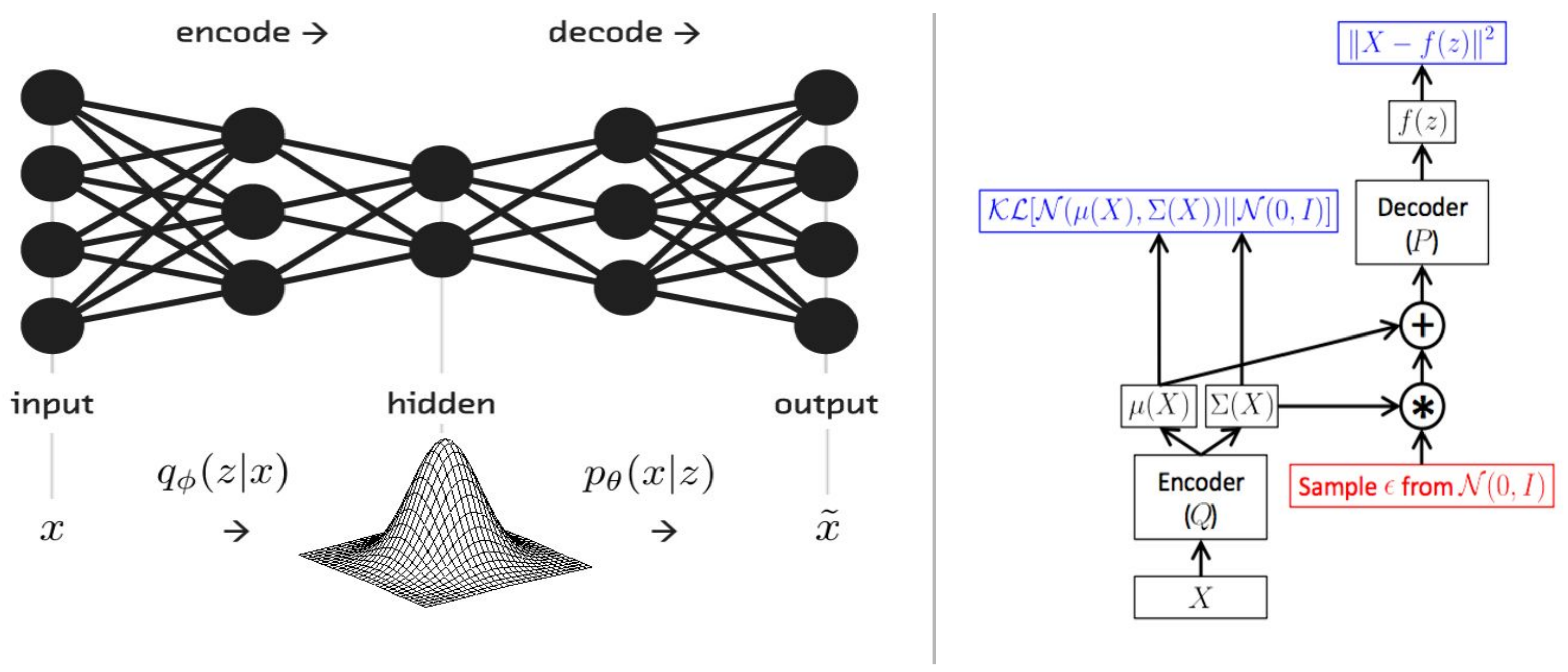
Approach

As mentioned above, the original idea of this project derived from the generativity of VAE. By viewing tons of images of a type, we hypothesized that the model could be clever enough to capture the generation logic of the original images and thus being able to reconstruct the occluded part. The basic approach is then to randomly sample a certain number of latent representation **z,** pass them into the decoder to get the reconstructed images, and then choose the one that offers the lowest reconstruction loss.

However, this naive approach did not perform well on testing time. This is due to the fact that though VAE successfully learned the generation logic, we still have no control of the generation process and it becomes hard to sample the "correct" **z** from the distribution. Therefore, we come up with several alternatives. An easy modification is to pass only occluded images as input, but compare the reconstructed images with the gold images so that the model will be able to learn how to complete the image.

Unfortunately, this is still not good enough. Therefore we shift our focus to a variation called **conditional variational autoencoder**. By conditioning on additional variable **c**, the decoder now models the distribution **P(X | z, c).** A natural approach then is to condition the image on its category. In the MNIST case, this is just the number the image represents. Although this provides us with some significant improvement on test data, we still think conditioning on the number category is not good enough. Then we came up with the idea that we should rather condition on the incomplete image directly, resulting in our last model which rendered the best performance.

Model



There are 4 types of VAE we implemented for this project.

- Vanilla Variational AutoEncoder (VAE)
- VAE with incomplete images as inputs
- Conditional VAE with labels as condition
- Conditional VAE with incomplete images as condition

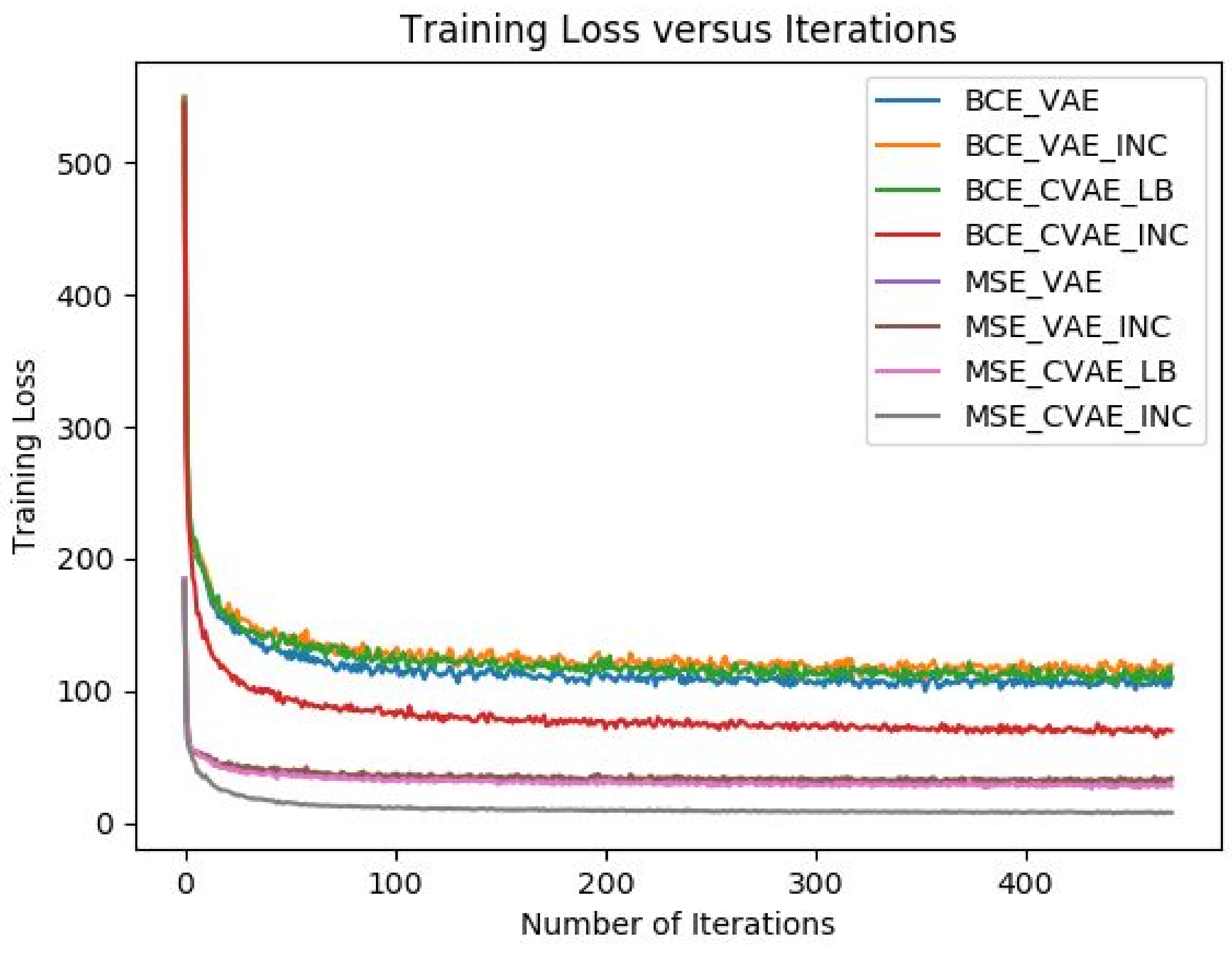
We will briefly elaborate on the last model and the others are in a similar fashion. For a Conditional VAE with occluded image as its condition, we are essentially optimize the following objective:

$$\log P(X|c) - D_{KL}[Q(z|X,c)\|P(z|X,c)] = E[\log P(X|z,c)] - D_{KL}[Q(z|X,c)\|P(z|c)]$$

which consists of a reconstruction loss (either BCE loss or MSE loss) and a variational inference KL divergence loss.

Result and Analysis

We experimented the 4 types of models on 2 different loss functions, namely the Binary Cross Entropy (BCE) loss and the Mean Squared Error (MSE) loss. As in the following plot, the training losses of each model keep decreasing. Particularly, we noticed that the two CVAE models with MSE loss rendered both the lowest loss and the "best" reconstructed images. An example of test sample (7) is indicated in the table. As one can tell, the vanilla VAE and VAE using incomplete images may not be able to even correctly predict the original number based on the content. In addition, we observed that the CVAE model with labels as condition worked better on MSE loss. But the CVAE model with incomplete image as condition outperformed all models above.



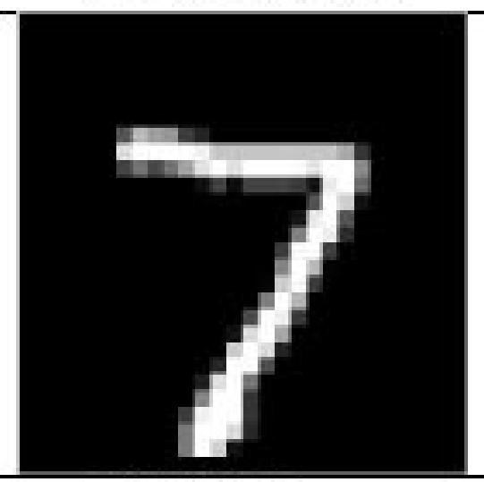
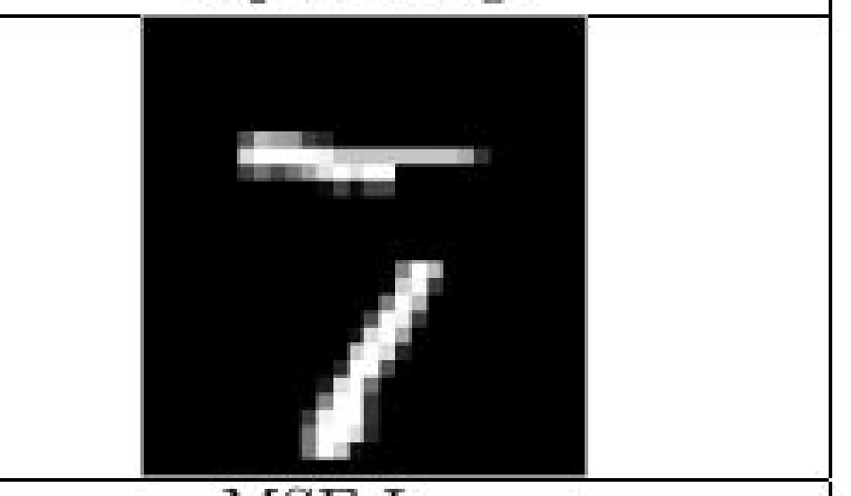
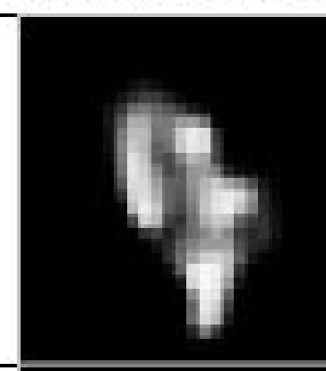
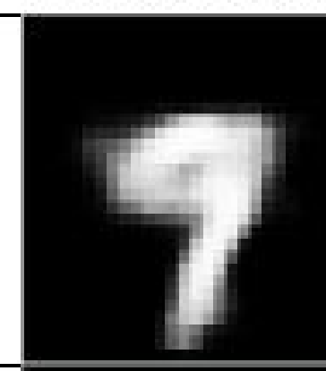
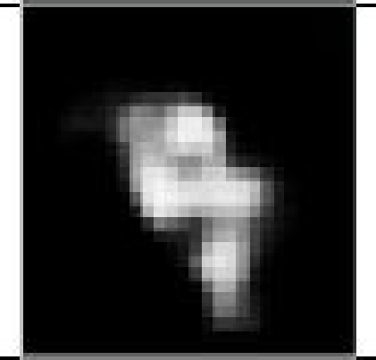
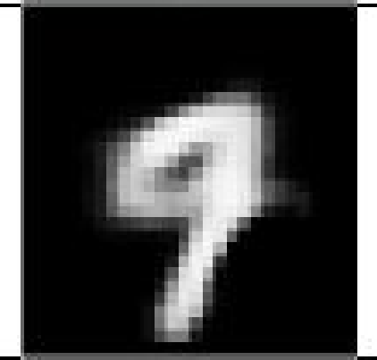
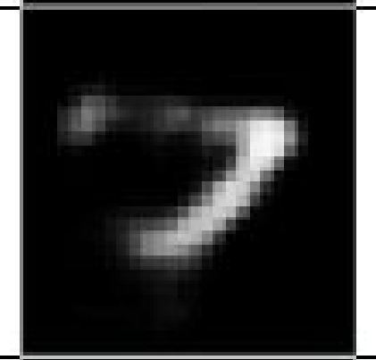
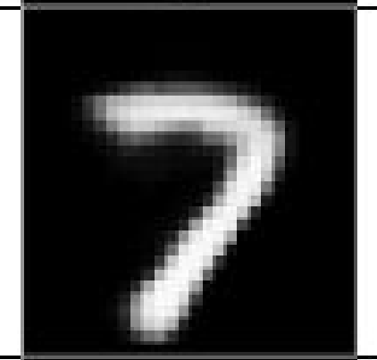
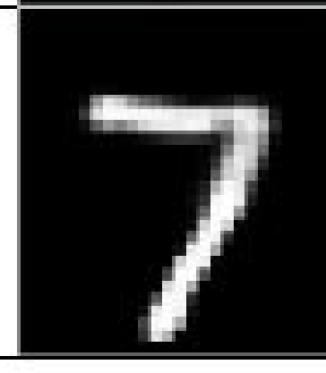
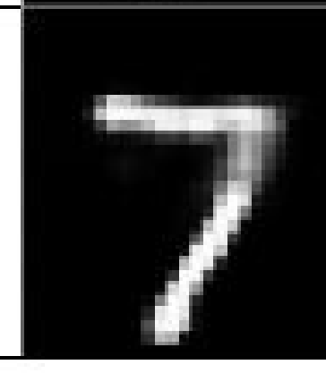
	Gold Image		Input Image	
				
Method	BCE Loss		MSE Loss	
	Recon Image	Avg Test Loss	Recon Image	Avg Test Loss
VAE		389.39		67.36
VAE-INC		339.48		65.12
CVAE-LB		250.29		49.32
CVAE-INC		50.34		2.93

Table 1: Project Demo

Conclusion and Future work

Although VAE model is famous for its ability to model complicated distribution in a generative fashion, we found that it is hard to sample the correct latent variable for the specific image we would like to generate. On the other hand, by conditioning on the rest of the image, the model quickly learned how to recover the occluded part and the test loss dropped to single digit number on a MSE measure. Some **future work** will include measuring the reconstruction loss between the non-occluded parts of both the original and the reconstructed images, so that the model will be able to learn the generation logic without complete inputs given.