Efficient Latent Space Representation Learning

Shimon Smith, Himanshu Arora, Raksha Harish, Tavleen Sahota

INTRODUCTION

An autoencoder consists of an encoder, a bottleneck layer, and a decoder. The layer between the encoder and decoder is known as the latent space. This latent space contains an encoded, compressed representation of an image, which must be used by the decoder to reconstruct the original image. In order to perform this reconstruction, the network must effectively learn the most relevant features present in the latent space. This is known as "Efficient latent space representation learning".

Variational Autoencoders (VAE) are generative models that encode data in latent space as regularized distributions and then decode it to reconstruct the data. Besides having applications in data compression, vision-based robotic navigation, and noisy/corrupted data reconstruction, VAEs have been successfully trained to encode RGB image data and reconstruct the image from it. However, there is little research in their ability to encode depth information.

Therefore the aim of this project is to develop a VAE model that learns depth information in addition to the RGB data and successfully reconstructs RGB-D images. We start by using RGB-D inputs to train a model to reconstruct RGB-D outputs. A number of models were implemented to reconstruct RGB-D images, and these models have been compared to find the most optimal architecture for RGB-D reconstruction.

DATSETS :

1. MNIST Dataset of handwritten digits (labeled dataset) – 60,000 training samples and 10,000 test samples of RGB images
2. NYU Depth Dataset V2 – consists of a variety of indoor scenes recorded by both RGB and Depth cameras with 464 scenes, 407,024 unlabeled frames and 1449 processed pairs of RGB and depth images.

COMPARATIVE MODELS

The following models were implemented initially to test the quality of RGB reconstruction on MNIST Dataset :

1. Fully Connected (FC) VAE – consists of Encoder and Decoder with a sequential set of fully connected neural net layers.
2. Conditional VAE (C-VAE) – This is an extension of the Vanilla VAE, where the RGB image reconstruction is conditioned on the latent vector space to control the generation of data by the decoder based on the labeling used in MNIST dataset.
3. VGG-16 based VAE – In this model, a VGG block containing 16 convolutional layers is defined and used by the encoder to improve the encoder's performance for compressing the original image.

The above 3 models performed poorly when implemented for RGB-D image reconstruction in NYU Dataset as they failed to capture the details of the scenes.

PROPOSED MODEL

To train our network we maximize the ELBO. We use the formulation in terms of KL divergence and a reconstruction term.

$$\text{ELBO} = \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log p(x|z)\right] - \text{KL}\left(q_\phi(z|x)\,||\,p(z)\right)$$

Both our prior, $p(z)$, and posterior approximation, $q_\phi(z|x)$, are Gaussian. As in Kingma et al. we can integrate analytically to get,

$$\text{ELBO} = \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log p(x|z)\right] + \frac{1}{2}\sum_{j=1}^{J}\left(1+\log(\sigma_j^2) - \mu_j^2 - \sigma_j^2\right)$$

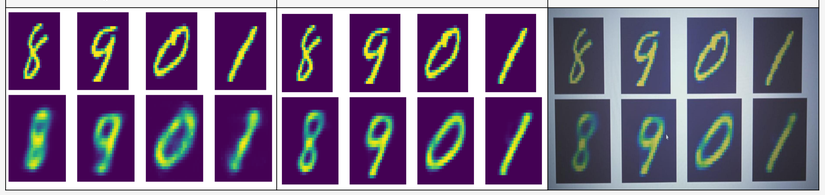
Where J is the latent dimension and μ, σ are the variational parameters ϕ. For the reconstruction term we take MSE which is equivalent to assuming a Gaussian likelihood. The ELBO we maximize is,

$$\text{ELBO} = ||x-\hat{x}||_2 + \frac{1}{2}\sum_{j=1}^{J}\left(1+\log(\sigma_j^2) - \mu_j^2 - \sigma_j^2\right)$$

Where $\hat{x}$ is the stochastically sampled reconstruction of the input x

RGB to RGB Reconstruction - MNIST Dataset

Figure: Reconstructed images on MNIST dataset using various models at epoch = 20.



RGB-D to RGB-D Reconstruction - NYU Dataset

Figure: Reconstructed RGB-D images for fully connected model (above) and Conditional VAE model (below) for the Labeled NYU Dataset at epoch = 100.

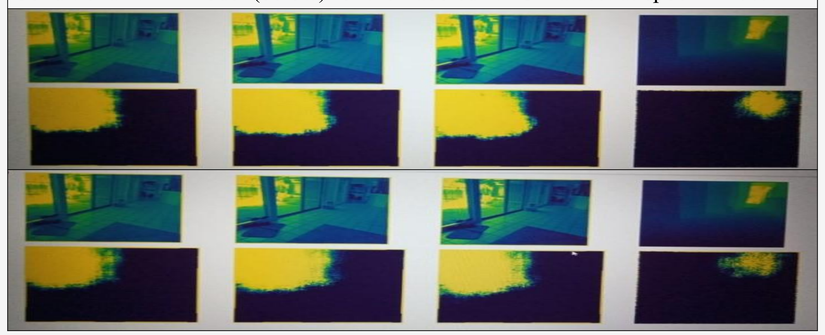


Figure: Reconstructed RGB-D images for VGG model for the Labeled NYU Dataset at epoch = 100.

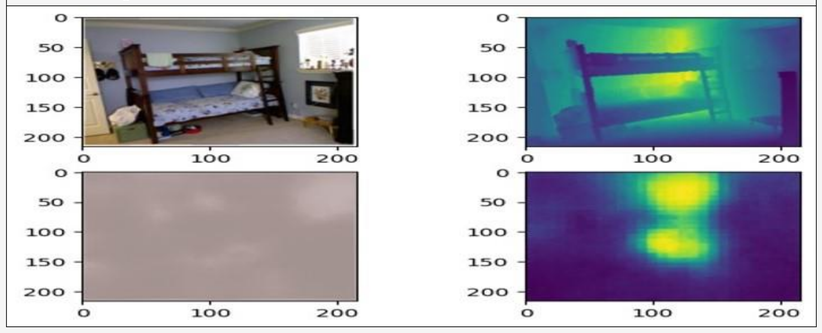
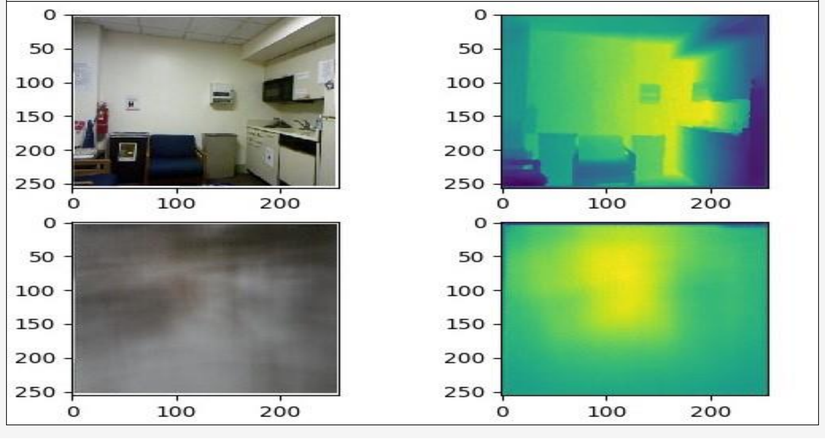


Figure: RGB-D reconstruction image for proposed model for NYU Dataset at epoch = 500.



ARCHITECTURE

We aim to create an architecture capable of learning the diverse scenes in a larger subset of the NYU Dataset.

To build the proposed model, we used the below architecture:

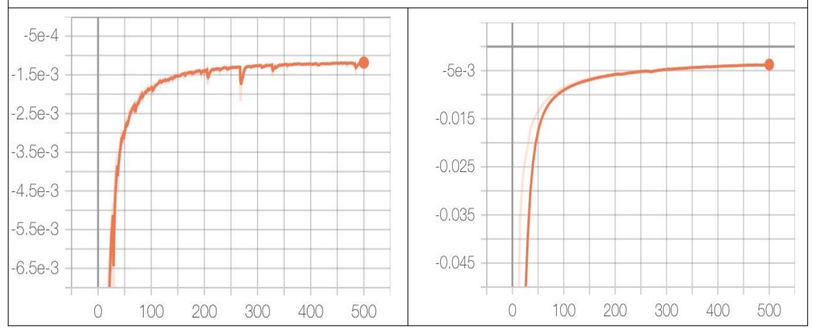
- Encoder: Pretrained ResNet34 with the first two blocks of convolutions frozen
- Decoder: The decoder is a sequence of residual blocks and transpose convolutions followed by Leaky ReLU activations. The final activation function is tanh since we use MSE as our reconstruction loss.

RESULTS

Table 1: Comparative Mean Sq Error of each model

MNIST dataset			NYU Dataset	
Fully Connected VAE	Conditional VAE	VGG	VGG	Proposed Model
13.31	9.96	8.97	161.72	1086231.16

Figure: Train loss (left) and test loss (right) graphs for NYU Dataset



For NYU Dataset, the quality of reconstruction observed, was very poor for the Fully Connected VAE and the Conditional VAE models as they could not effectively encode depth into the RGB image sequence. The deeper convolutional architectures we tried preform better but still fail to capture the high variance of indoor scenes.

DISCUSSION AND FUTURE ENHANCEMENTS

- We implemented various VAE architectures and showed our results on two datasets - MNIST and NYU Depth. Most works so far have shown results on the MNIST or CelebA dataset and it is challenging to get good results on a complex scene dataset such as NYU Depth.
- In future, we hope to improve our results on NYU Depth and make reconstruction more sharp. We also plan to reconstruct depth information just from RGB data during test time.

REFERENCES

1. Punarjay Chakravarty, Praveen Narayanan, and Tom Roussel. GEN-SLAM: Generative Modeling for Monocular Simultaneous Localization and Mapping. In The *IEEE 2019 International Conference on Robotics and Automation (ICRA),* May 2019.
2. Jing Zang, Deng-Ping Fan, et.al. UC-Net: Uncertainty Inspired RGB-D Saliency Detection via Conditional Variational Autoencoders. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* April 2020.
3. Yunchen Pu, Zhe Gan, et.al. Variational Autoencoder for Deep Learning of Images, Labels and Captions. In The 30th Conference on Neural Information Processing Systems (NIPS), 2016.
4. https://pyro.ai/examples/vae.html
5. MNIST Dataset - http://yann.lecun.com/exdb/mnist/
6. NYU Dataset - https://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html