Autoencoder

Industrial AI Lab.

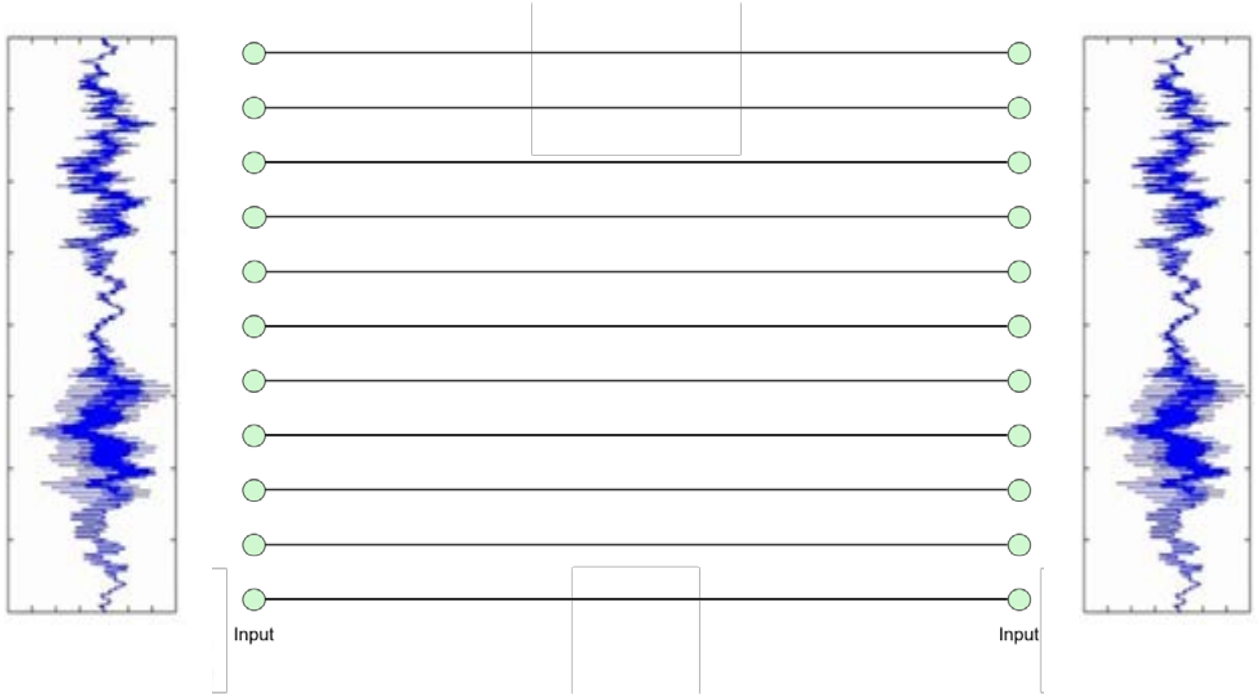
Prof. Seungchul Lee

Autoencoders

- It is like ‘deep learning version’ of unsupervised learning
- Definition
  - An autoencoder is a neural network that is trained to attempt to copy its input to its output
  - The network consists of two parts: an encoder and a decoder that produce a reconstruction
- Encoder and Decoder
  - Encoder function : $z = f(x)$
  - Decoder function : $x = g(z)$
  - We learn to set $g\big(f(x)\big) = x$

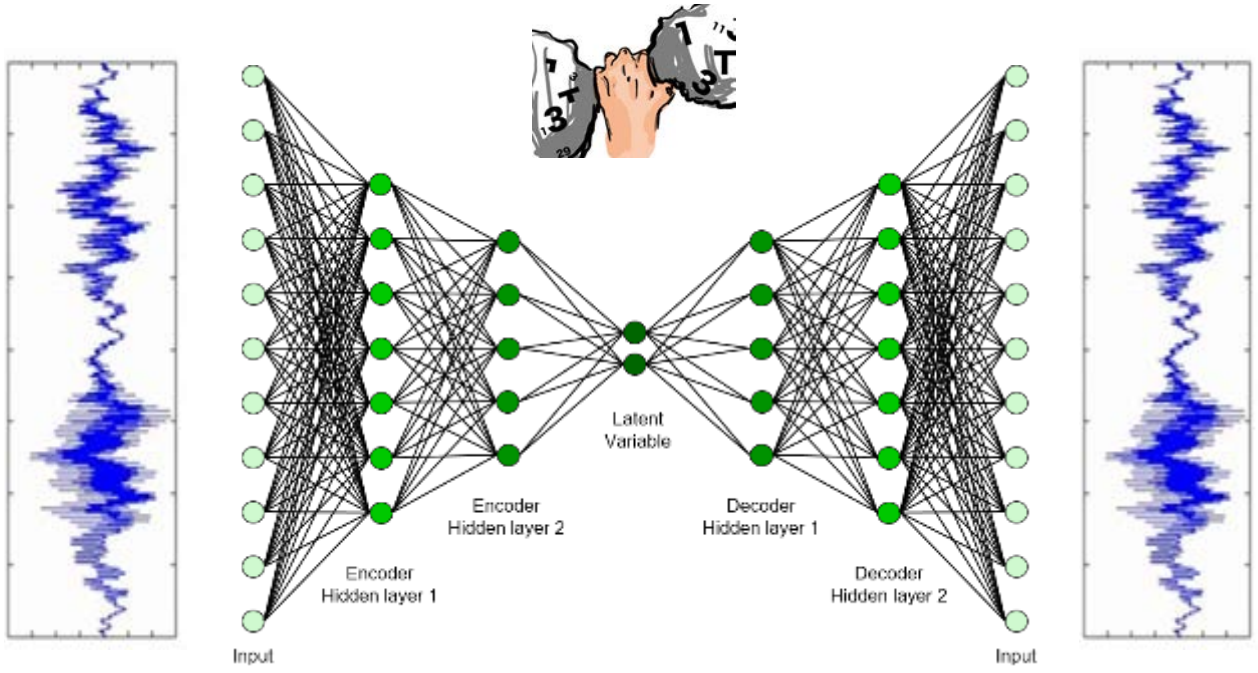
Autoencoder

- Dimension reduction
- Recover the input data



Autoencoder

- Dimension reduction
- Recover the input data
  - Learns an encoding of the inputs so as to recover the original input from the encodings as well as possible

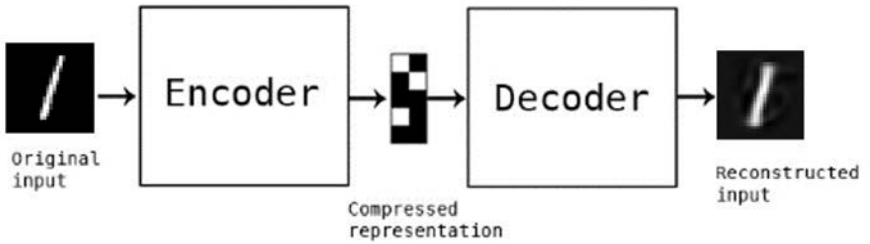


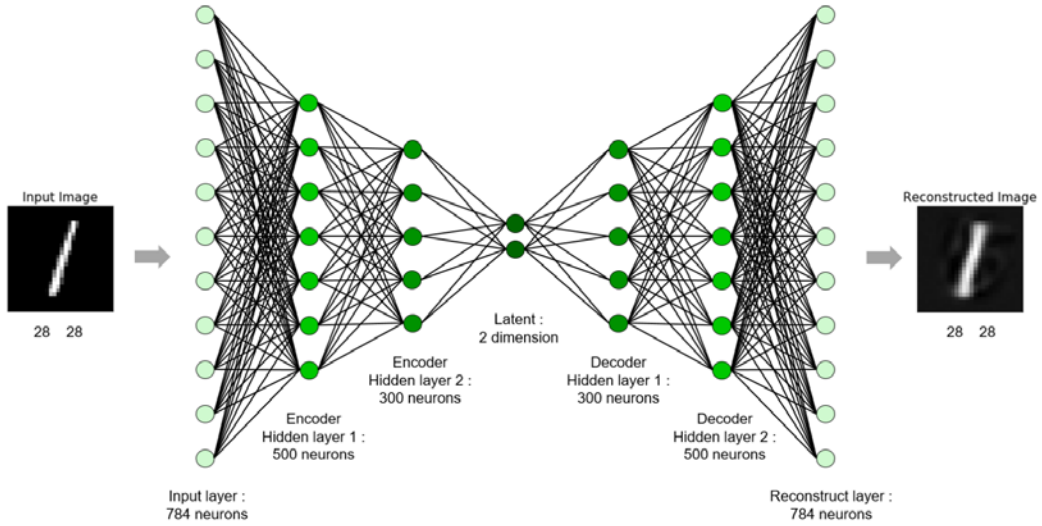
Original space Latent space

Autoencoder with MNIST

Autoencoder with TensorFlow

- MNIST example
- Use only (1, 5, 6) digits to visualize in 2-D



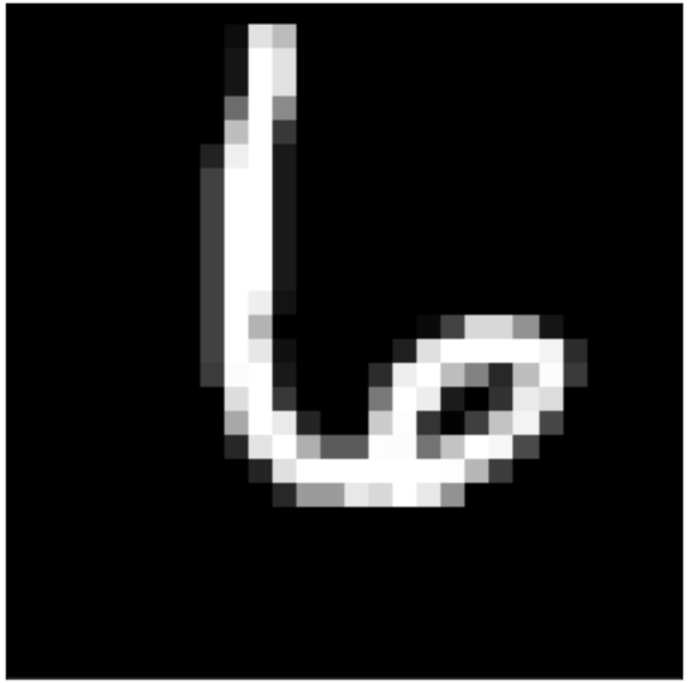


$$\frac{1}{m}\sum_{i=1}^{m}(t_i - y_i)^2$$

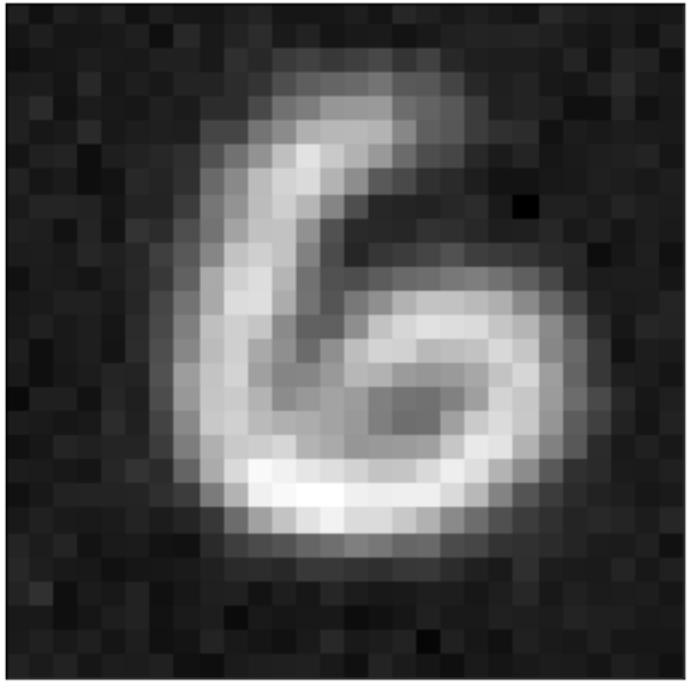
Test or Evaluation

```
test_x, _ = test_batch_maker(1)
x_reconst = sess.run(reconst, feed_dict = {x: test_x})
```

Imput Image



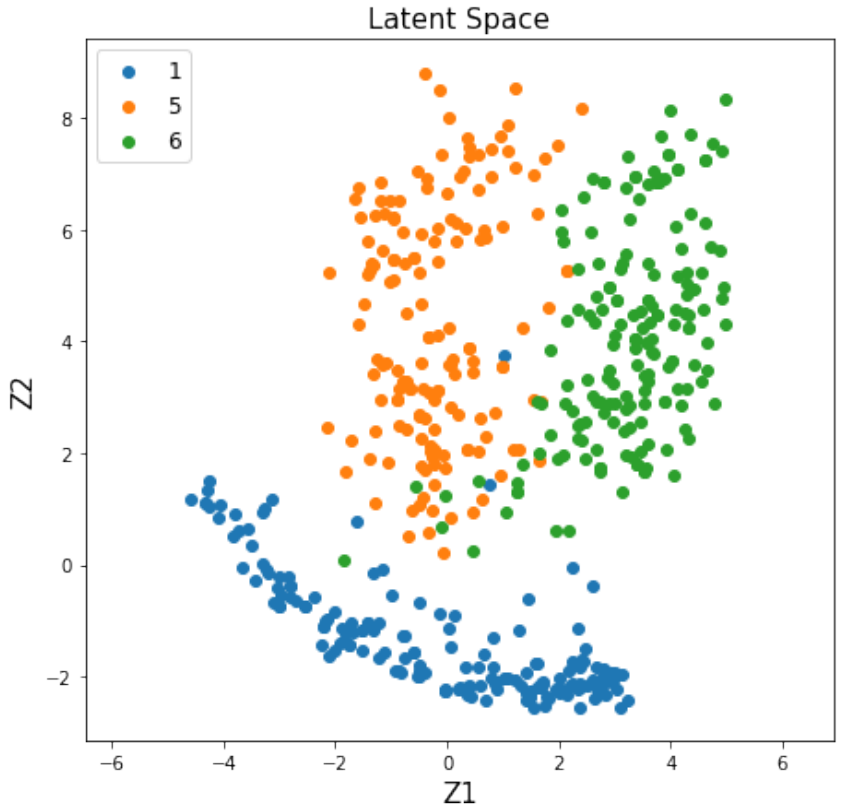
Reconstructed Image



Distribution in Latent Space

- Make a projection of 784-dim image onto 2-dim latent space

```
test_x, test_y = test_batch_maker(500)
test_y = np.argmax(test_y, axis = 1)
test_latent = sess.run(latent, feed_dict = {x: test_x})
```



Generative Capabilities