MSSE277b: Machine Learning
Homework assignment #11 (extra credit):
Recurrent Neural Network, LSTM and VAE
Assigned Apr. 17 and Due Apr. 24

1. **LSTM applied to SMILES string generation.** (10 pt) Using the SMILES string from the ANI dataset with upto 6 heavy atoms, build a LSTM generative model that can generate new smiles string with given initial character.
   (a) (3pt) Process the smiles strings from ANI dataset by adding a starting character at the beginning and an ending character at the end. Look over the dataset and define the vocabulary, use one hot encoding to encode your smiles strings.
   (b) (7pt) Build a LSTM model with 1 recurrent layer. Starting with the starting character and grow a string character by character using model prediction until it reaches a ending character. Look at the string you grown, is it a valid SMILES string?

2. **Variational Autoencoder(VAE) applied to MNIST dataset.**(10 pt)
   Train an VAE model for the MNIST dataset. The encoder and decoder of the VAE model are convolutional neural networks. Encoder have 4 convolutional layers, each with 4, 8, 16, 32 channels, kernal size of 4x4, padding of 1 and stride of 2. and the decoder is the reverse of that. In the bottleneck region, the encoder output is flattened and mapped to two latent vector μ and σ each represented with 32 hidden neurons by two separate linear layers. Then the latent state z with 32 hidden neurons is formulated by applying reparameterization with addition of noise ε, which is then passed to decoder. Use binary cross entropy plus KL divergence as your loss function. Train this model with the MNIST dataset and use the provided reconstruction code to show that your model is able to reproduce the images.