277B: Machine Learning Algorithms

Homework assignment #8: Recurrent NN and LSTMs

Assigned Mar. 19 and Due Apr. 1

1. **RNN applied to SMILES string generation.** (50pt) Using the SMILES string from the ANI dataset with up to 6 heavy atoms, build a RNN generative model that can generate new smiles string with given initial character.
   (a) (15pt) Process the SMILES strings from ANI dataset by adding a starting character ("SOS") at the beginning and an ending character ("EOS") at the end. Look over the dataset and define the vocabulary, use one hot encoding to encode your smiles strings.
   (b) (25pt) Build a vanilla RNN model with 1 recurrent layer with hidden size 32. Use Adam optimizer with learning rate 1e-3, batch size 128, L2 regularization with lambda=1e-5 to train the model for 500 epochs. Then use the provided code to grow 1000 string character by character starting from the starting character using model prediction until it reaches the ending character. Use the provided "validate" function, how many valid and unique SMILES strings do you get?
   (c) (10pt) Build a LSTM model with 1 recurrent layer with hidden size 32 and the same hyperparameter setting. Again, generate 1000 SMILES strings. How many valid and unique strings do you get this time?