MSSE277b: Machine Learning and Optimization for Chemical Problems

Project Assignment

Assigned March 17 and Due May 6

For the final project we will develop a supervised learning ANN model applied to the ANI-1 data set. We will do a check-in once per week to see steady progress with appropriate entries of dates in the jupyter notebooks on what has been accomplished. I.e. this is not a project assignment that should be finished the night before. This will be part of the assessment.

The finals project has the following expectations for assessment:

Part I: An individual jupyter notebook should be maintained during the course of the project. There will be 3 progress check-ins, each 20% of the grade. You'll submit your notebook to Gradescope.

(1) (April 14) Data preparation. Show that you're able to load the data, process them into model input format, and split the data into train, validation and test set with batching.

(2) (April 21) Network construction and workflow development. At this point you should have a working code that can train the network, demonstrated on small subset of the data.

(3) (April 28) Regularization strategies and hyperparameter tuning. Use more data to train the network. Play with the architecture, hyperparameters and the regularization strateges. Show your work and defend the final choice of your model.

Part II: (Due May 8) Submit a jupyter notebook on final results. Train your model with all data (at least upto 5 heavy atoms) and compare your results to what's reported in the paper. The expectations are:

1. The notebook should be concise and readable.
    a. There should be explanatory text for each major step and your code should be well-commented.
    b. Please suppress long printings that may last for pages, and delete unnecessary print outs, such as those for debugging purpose
2. A conclusion should be given in the end. For instance what have you achieved, how good is the model compare to other methods or literature values, what else can you improve etc.
3. You are expected to show the learning curves for any deep learning training. Please present this as a learning curve plot instead of printing out all loss values.