Final Project Report

-Spencer Cain, Trina Dutta

We have chosen the paper “A self-attention based message passing neural network for predicting molecular lipophilicity and aqueous solubility”. The paper leverages the message passing framework to learn on graph representations. However, the graph representations passed through the message passing framework seem redundant and possibly ineffective. For instance, the original SAMPN model depends on atom-to-bond, bond-to-atom, atom-to-atom, and bond-to-bond relationships, where atom-to-bond and bond-to-atom map a bond index to the atom index of the starting atom. Each bond receives a directed edge to and from the atoms connected by the bond. We intend to improve this implementation in two ways. Rather than using attention weights placed on atoms following the MPN framework and the diluted graph representation, we employ graph attention layers that process the graph in the form of edge lists and a DGL graph as the graph representation. Our idea is to verify if this new representation of chemical compounds in graphical representation improves the prediction of lipophilicity and aqueous solubility.

Description of the new model:

The new model contains two submodels. The first submodel operates on graph data. This submodel contains two graph attention layers followed by an average pooling layer, which condenses the graph representation to a tensor with 139 dimensions (originally representing the one hot vector of the atomic number). The second submodel then classifies the tensor using the same feed forward network employed in the original SAMPN model.

Result on the new model:

We initially trained and tested the new model on the same datasets that were provided with SAMPN. The new model does not perform better than SAMPN on both datasets. Further, since the new model depends on uniquely sized graphs and cannot be batched, the metrics R^2 and Pearson correlation could not be computed. The datasets are split into 80%, 10%, and 10% for train, valid, and test datasets. Further, all models are trained on 50 epochs. The following results are from the test split.

Lipophilicity	SAMPN	New Model (lr = 5e-4)
RMSE	0.2979	0.5380
MSE	0.0894	0.5013

MAE	0.2188	0.5380
R^2	0.9211	-
Pearson	0.9636	-

Solubility	SAMPN	New Model (lr = 5e-4)
RMSE	0.7650	0.5362
MSE	0.5932	0.6376
MAE	0.6428	0.5362
R^2	0.2540	-
Pearson	0.7287	-

Further information about the model can be found in the README.md file provided with the code.