Using TensorFlow to Predict Connection Table Information within Chemical Structures

Brodie Schroeder[1], Ryan Richard[2], Theresa Windus[2]
[1]Kennesaw State University, [2]Iowa State University

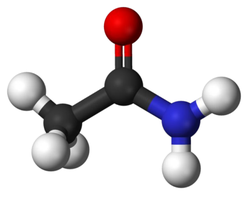
Abstract

Using the popular machine learning library TensorFlow, an open source project developed by Google Brain, we explore some of the possible use cases it has within the computational chemistry domain. The goal of this machine learning task is to develop a program, written in Python, that is able to parse large quantities of information from a file, build a training dataset based on this information, and define a sufficient machine learning algorithm that predicts the connection tables for a given chemical structure with a high degree of accuracy. The training data that will be parsed from the openly available files hosted on the PubChem.gov website which contain information that detail thousands of chemical structures. We will use this information to perform supervised training on our machine learning model to try and "teach" the computer how to accurately predict connection information.

Purpose

Obtaining this connection information is useful in that it will allow researchers to avoid calculating this information manually which can be very time consuming.

Sample



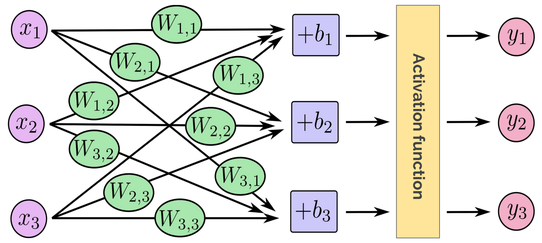
```
 1.9050   -0.7932   0.8766 C
 1.9050   -2.1232   1.3211 C
 0.7531   -0.1282   6.3411 C
 0.7531   -2.7882   0.5322 C
-0.3987   -0.7932   6.3221 C
-0.3987   -2.1232   4.7832 C
-4.2212    6.2138   3.2312 O
```

```
6.0,  0.0,  4.312,  6.223, 7.321,  3.221,  9.023,
6.0,  1.542,  0.0,  4.222, 8.231,  6.321,  1.999,
6.0,  2.221,  5.012,  0.0, 4.223,  6.723,  7.232,
8.0,  7.010,  3.011,  7.221, 0.0,  5.434,  7.777,
6.0,  4.312,  3.221,  3.563, 7.212,  0.0,  6.521,
8.0,  2.333,  5.321,  6.872, 6.454,  8.991,  0.0,
```

```
2 1 1 0 0 0 0
3 1 2 0 0 0 0
4 2 2 0 0 0 0
5 3 1 0 0 0 0
6 4 1 0 0 0 0
6 5 2 0 0 0 0
```

```
0.0,  1.0,  1.0, 0.0,  0.0,  0.0,
1.0,  0.0,  0.0, 1.0,  0.0,  0.0,
1.0,  0.0,  0.0, 0.0,  1.0,  0.0,
0.0,  1.0,  0.0, 0.0,  0.0,  1.0,
0.0,  0.0,  1.0, 0.0,  0.0,  1.0,
0.0,  0.0,  0.0, 1.0,  1.0,  0.0,
```

Methods



We use a training set of approximately 10,000 chemical structure and bond table pairs to train a multilayer neural network. This is a supervised learning technique.

Our artificial neural network (ANN) consisted of three layers. Between the layers we use the rectifier, or ReLU, function.

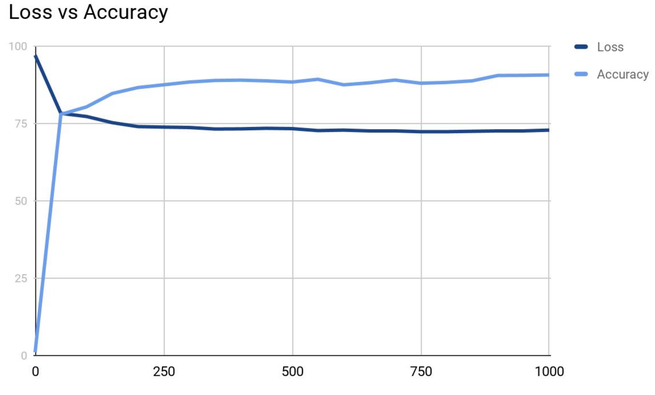
$f(x) = x^+ = max\ (o, x)$

On the output layer we use a sigmoid function for activation.

$S(x) = e^x / (e^x + 1)$

Results

Using this multilayer neural network we were able to achieve and accuracy of approximately **0.8867** or **88%** This is still lower than we would like for machine learning purposes and can be improved upon at a later date. We trained this model with about 10,000 target classes examples.



Discussion

There is still a lot to be explored in the application of machine learning to this problem. As you can see from our results, after the 250th training step the model does not improve as a result of more training steps. This is a possible scenario of overtraining and this still needs to be explored.

Additional improvements may also include adding more layers to the neural network and implementing some form of filtering or optimizing the weight selection.

We may also look into other types of neural networks such as a recurrent neural network (RNN) which can be better suited to the type of classification task we are attempting to perform.

Conclusions

We have discovered through this project that predicting the bonding tables of basic chemical structures does indeed seem to be possible through the use of machine learning. While the accuracy is fairly low, we now know that further optimization and the use of more advanced machine learning techniques will likely improve the accuracy of the model.

Acknowledgements

This work was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internships Program (SULI).