Machine Learning of Water Molecule Energies

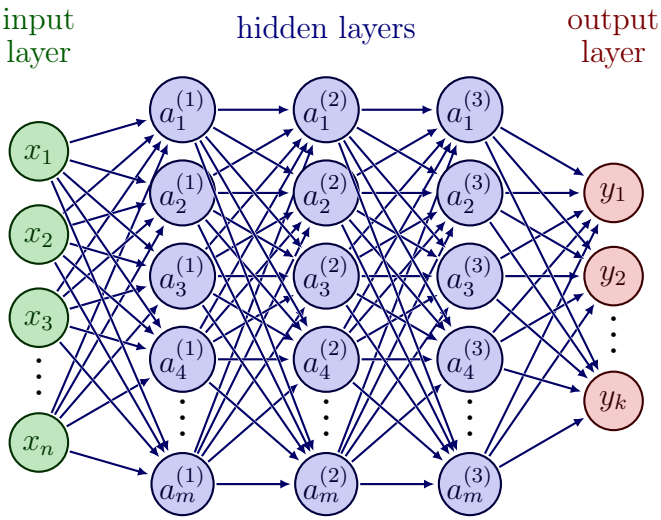
Conor Ryan

Introduction

The focus of this report is on the application of machine learning methods to the prediction of the energies of water molecules in various geometries. The geometries under study consider the stretching of distances between the atoms, as well as translations and rotations of the entire molecule.
The accuracy and performance of the machine learning model is judged by analysing the convergence of the model over a number of epochs, how it learns from training sets of different sizes and how its energy predictions compare to the values given in the test set.

Method

The machine learning model used in this work is a neural network. A diagrammatic representation is shown below.



Firstly, the x, y, z-coordinates of the atoms in each molecule are converted into distances between each of the atoms, r_{HO}, r_{OH} and r_{HH}. This is done because the energy of the H_2O molecule depends on the distances between its constituent atoms, not their specific x, y, z-coordinates. The predictions of the model will then be more generalisable as molecular geometries which are identical up to a rotation or translation in space will result in identically accurate energy values.
With the atomic distances as the features and the energy values as the labels, the data is then normalised. The data can be normalised using different methods and the method chosen in this case is scaling the data into the range $[0, 1]$,

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

Scaling the data into this range puts it on a similar scale, allowing the model to train faster and not give very high initial loss values due to the initially random weights and biases. This is an effective normalisation method given that the data has few outliers and is distributed in a relatively uniform way.

The three normalised atomic distances for each molecule are then input into the neural network and the model is trained to optimize the weights and biases in order to accurately predict the training energies. Initially the data is shuffled and throughout the training 20% of the training data is used as a validation set to check that the model is performing as desired.
Once the model is satisfactorily trained, a test set which is pre-processed in the same way as the training set, is given to the model to make predictions on so its performance can be tested on completely new unseen data.

Unrotated H_2O Molecule

The first training set contains H_2O geometries where only the atomic distances are changed with no overall rotations or translations of the molecule.
Through varying the model and training parameters, such as the number of layers, neurons in each layer, activation functions, regularization and number of epochs, and checking the accuracy of the energy predictions on the test set, the final neural network architecture was found to consist of an input layer of 3 neurons(for the atomic distances), hidden layers with 8, 12 and 8 neurons and an output layer of a single neuron(for the energy). The number of neurons in each layer were chosen to allow for sufficient complexity and flexibility in the model, without making it overly complicated and difficult to train. Each of the hidden layers use the *relu* activation function and both kernel and bias regularization. The layers with 8 neurons use $L2$ regularization with a penalty parameter of 10^{-4} and the 12 neuron layer uses $L1$ regularization with the same penalty parameter. The regularization is introduced to limit overfitting in the model. *relu* is chosen as the activation function as the model trained faster using this choice and the loss values of the model at the start of training were smaller than for other choices of activation function. The input data is also flattened as dense neural network layers expect flattened data. The weights and biases of the model were optimised using the *adam* optimizer and its performance measured using the mean square error(MSE) as the loss function.
The following results were produced using this neural network model by training over 500 epochs.

(a) Change in model loss over its training

(b) Training and test loss for different training set sizes

(c) Predicted H_2O energies compared with test energies

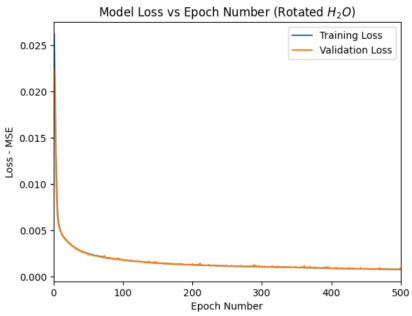
Figure 1: Results Produced for the Unrotated Dataset

Figure 1a shows that the model loss converges to a final smallest value which is 7.151×10^{-4}. By converging to a stable final value this plot shows that the training is performed over a large enough number of epochs. The test loss and validation loss being effectively identical indicate that there is no major underfitting or overfitting occurring in the model.
In Figure 1b both the test and training loss are seen to decrease relatively smoothly for increasing training set sizes. The smoother behaviour in the results is found from training the model 5 times at each training set size and taking an average of the losses. Intuitively, this is expected for the test loss, as the model will generalise better to unseen data when it has more data to learn from. However, this is not what is logically expected for the training loss. The training loss should be smallest for small training sets when it becomes easy for the model to overfit the data and achieve very small losses for the training set and high losses for the test set. At larger training set sizes the training loss should increase as it becomes more difficult for the model to overfit for a larger dataset.

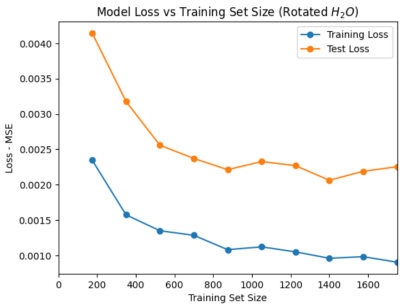
The energy predictions seen in Figure 1c show the model can make relatively accurate predictions on new test data, with an MSE of 0.005675305351629409. For a perfect model all predicted energy points will lie on the red line, which is seen not to be the case. These results indicate that the final model is a good but perhaps not optimal choice.

Rotated H_2O Molecule

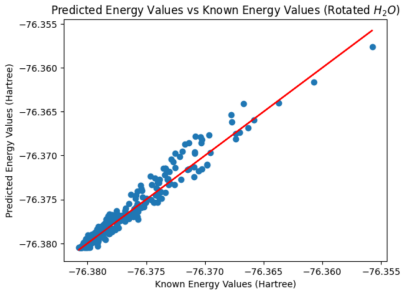
The next training set contains H_2O geometries which include rotations and translations, meaning it is more general than the previous training data.
Preparing the data in the same way described in the Method section, the final model is found to be the same as for the unrotated dataset. This is due to the fact that the original x, y, z-coordinate data is converted to atomic distances. This conversion then results in both datasets being very similar since atomic distances are invariant under rotations and translations.
The following results were produced after training the model over 500 epochs.



(a) Change in model loss over its training



(b) Training and test loss for different training set sizes



(c) Predicted H_2O energies compared with test energies

Figure 2: Results Produced for the Rotated Dataset

Each of the results produced in Figure 2 are almost identical to those produced in Figure 1 as a consequence of training very similar datasets on the same neural network architecture.
Minor differences occur in the precise loss values in the results between both datasets. The final loss reached by the model over 500 epochs for the rotated dataset is 9.1141×10^{-4}, which is of the same order as before but a slightly different value. This occurs due to the slightly different values in the new dataset and also the statistical nature of training a machine learning model, which won't always reach the exact same loss value after every training run.
The predictions on the test data produce energy values with a similar accuracy to Figure 1c, just with a slightly different MSE of 0.005600178421942583.

Conclusions

It has been shown that it is possible to train neural networks to predict the energies of H_2O molecules in various geometries to a sufficiently high accuracy. These high accuracy results were produced by training the model on datasets of molecular geometries with stretched bond distances, rotations of atomic positions and translations of the atoms in space.
Comparing the predicted energy values with the known values in the test set it was found that the results are not perfect and allow for the possibility of improving the model to make more accurate predictions.
Using the outlined method the choice of rotated or unrotated data has no effect on the model results or performance due to the atomic distances and energy values contained in both datasets becoming very similar once the data has been pre-processed to be input into the model.