**Final Report**

**Design of RNN**

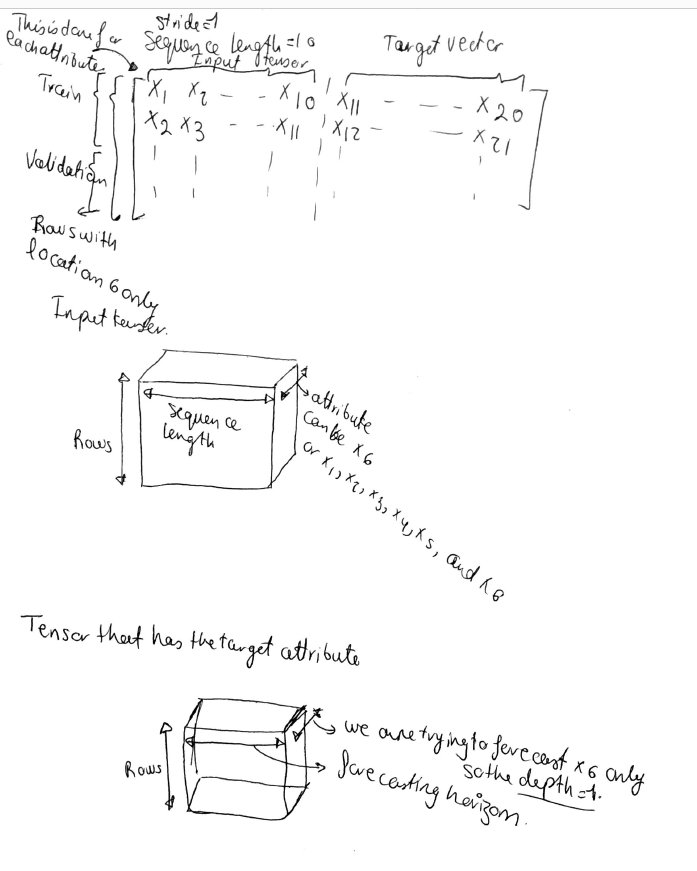
We relied on a simple design of RNN to forecast which outputs 20 values at the last time step. The network is trained with an input sequence followed by the next 20 sequential values from the time series as the target output. The output layer dimension is 30. The design of the tensor which will be input to the RNN is as follows.

* Rows correspond to the data instances
* Columns correspond to the time steps (days in our case)
* Depth corresponds to the attributes measured (which is going to be ‘x6’ if we have a ‘single’ variable option within the python code option and 'x1', 'x2', 'x3', 'x4', 'x5', 'x6' if we use the ‘multiple’ variables option within the python code.

We started with a RNN model that uses data from only the target attribute (‘x6’) and the target location (‘location 6’). Then, we tried different parameters to increase the complexity of the RNN network such as increasing the number of layers and the hidden size. Also, we changed the batch size and the number of epochs (the number of epochs can also control the overfitting or the underfitting). For these different options we plotted, the loss curve which contains the average mean absolute error for the training data and the validation data.

**Preprocessing of Data**

We scaled the data using the Min Max Scaler. Also, we prepared the tensor which will be the input as below. Also, the below tensor which has the target attribute has the below design. We chose only the rows that have location 6.



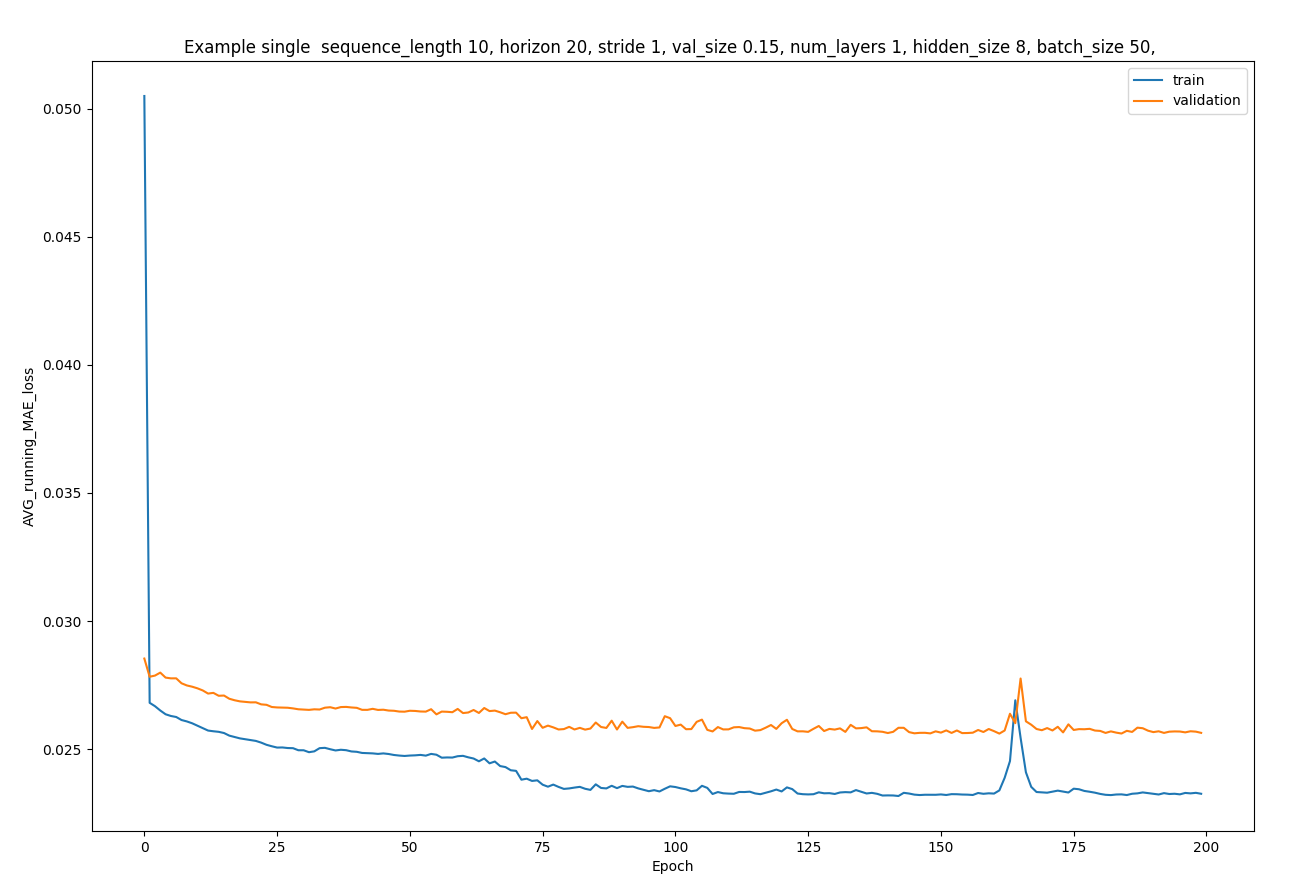
**Methods**

We used an RNN model with a design mentioned in Design of RNN section. Parameters are discussed in the next section.

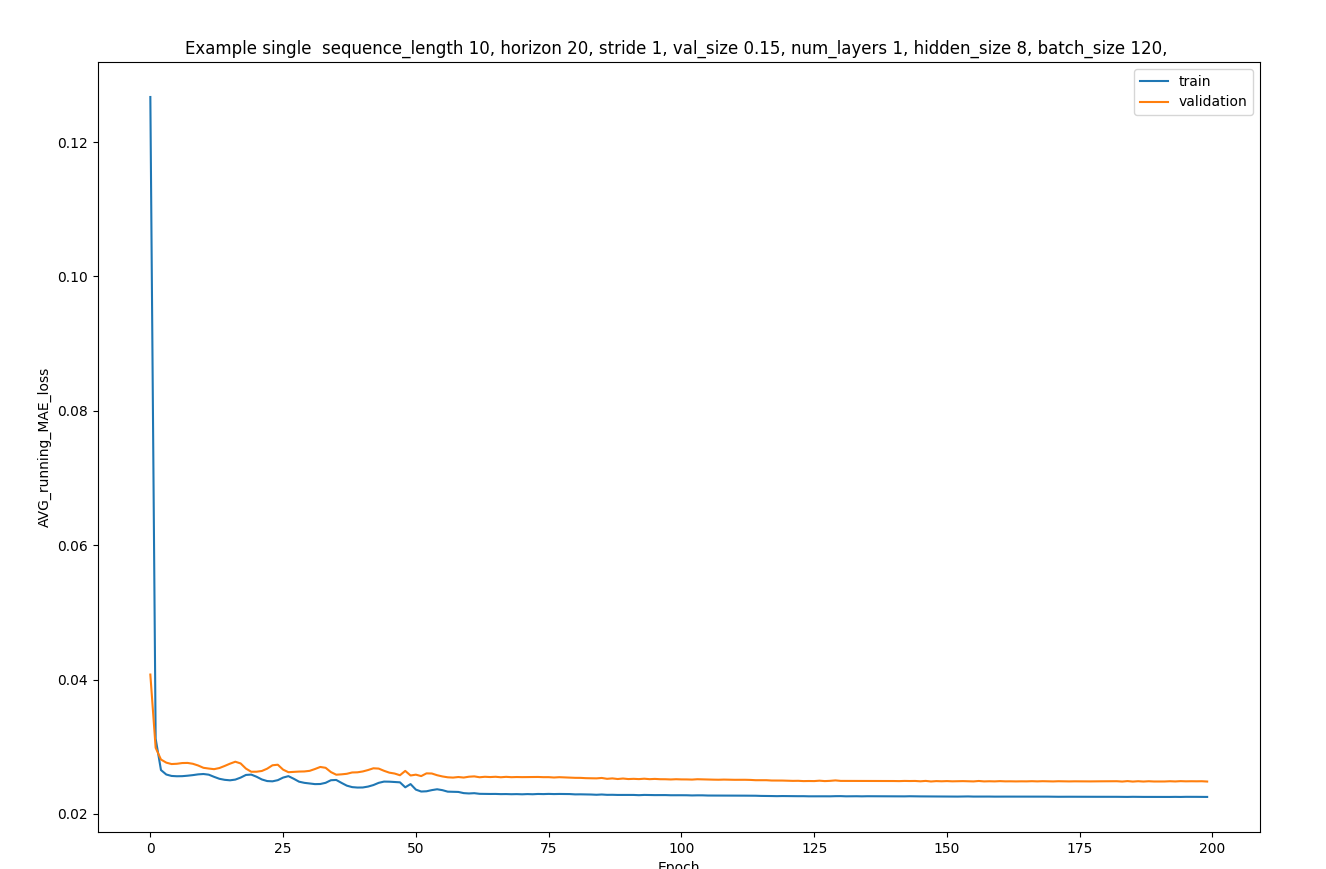
**Evaluation**

We mainly relied on loss curve which has the average Mean Absolute Error for the training and the validation error. Below are some parameters that we tried.

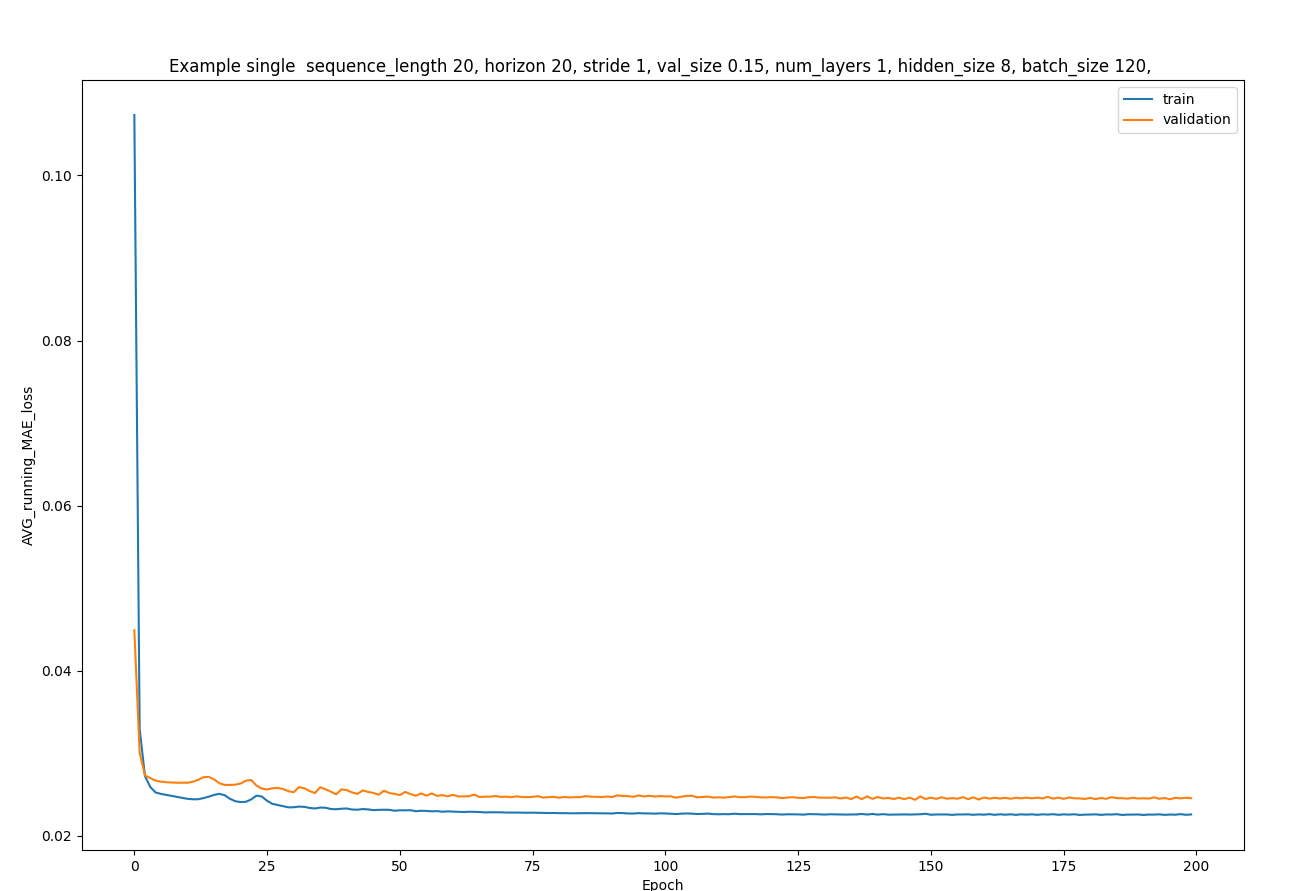
1. Single Variable. Using only the rows with location 6. Hidden Size=8, Number of Layers =1, Batch Size= 50, and Epochs=200. The minimum validation error is at 0.02509 which is at epoch 161. Below is the loss curve.



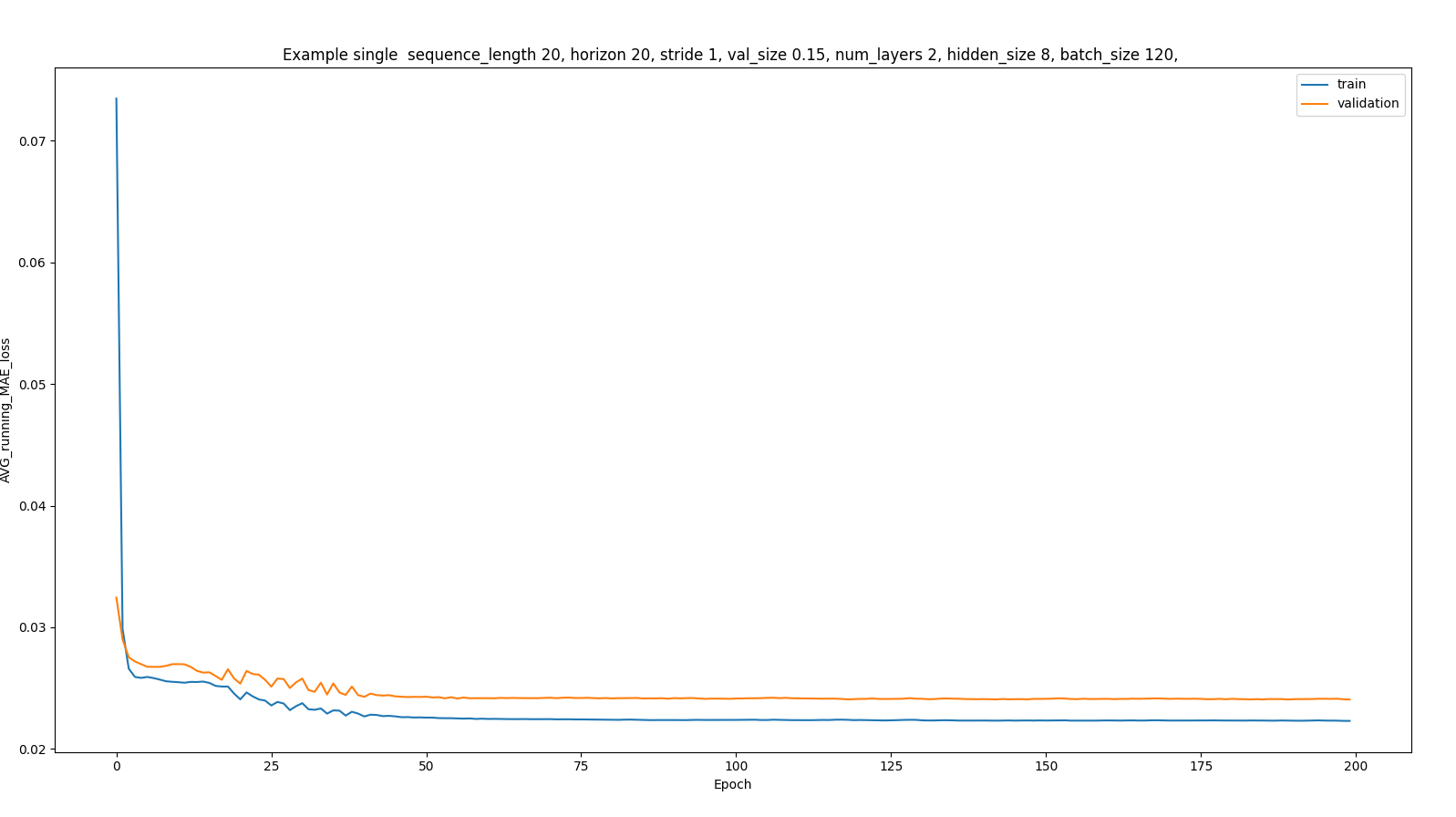
1. We can increase the batch size and see if it is going to make a difference. So, we increased the batch size to 120, while keeping the other parameters the same. The minimum validation error is 0.024824 which is a minor improvement in error. Also, we can see less fluctuations in the error which indicates that there is less variance which might indicate less overfitting.



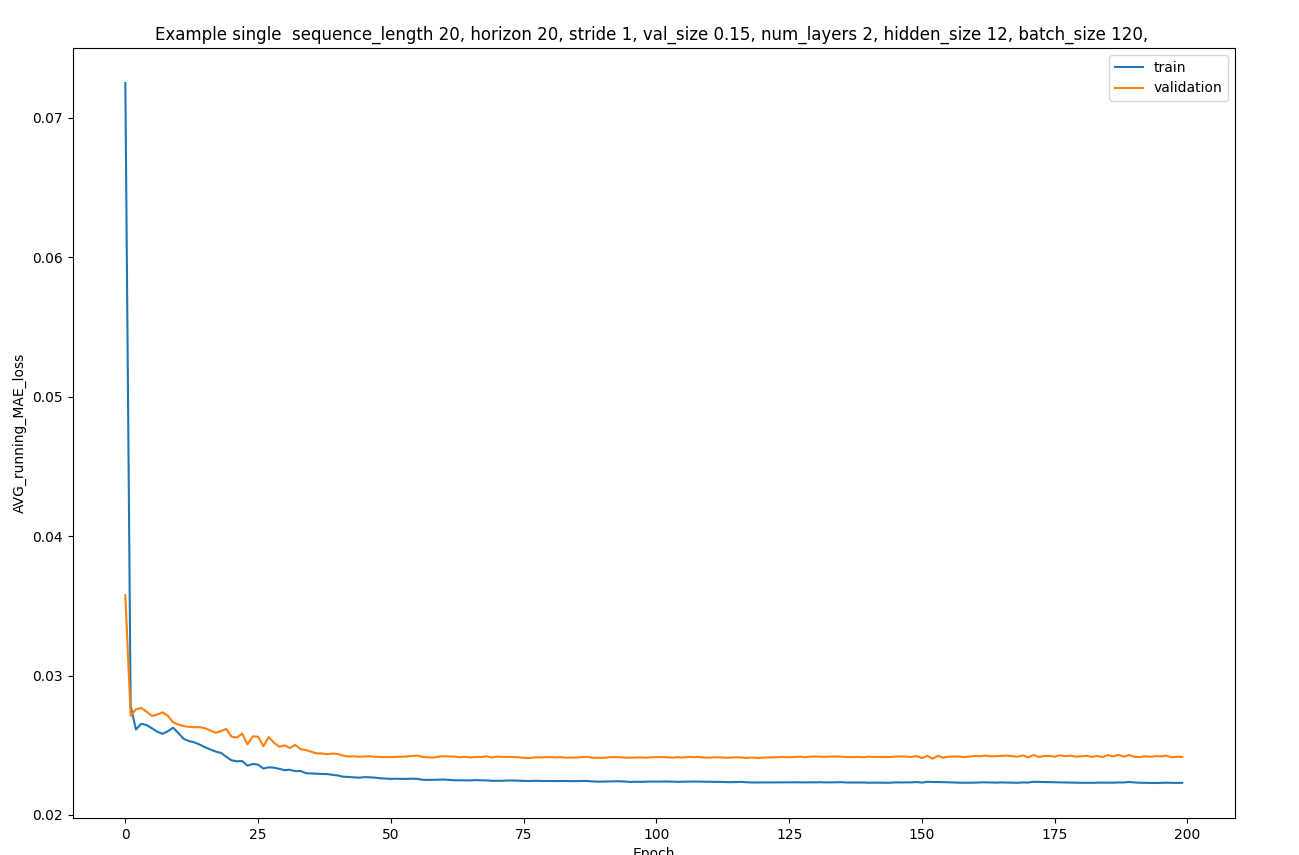
1. Also, we can increase the sequence length instead of 10. So, we increase the sequence length to 20, while keeping the other parameters the same.. The minimum validation error is 0.024325 which is a minor imporvement on the previous trial.



1. We can increase the complexity of the RNN. Two ways to do that is either by increasing the number of layers or by increasing the hidden size. We increased the number of layers to 2. The minimum validation error is 0.023994 which is considered an improvement.



1. We can further increase complexity by increasing the size of hidden units. The hidden size is increased to 12. The minimum validation error is 0.024032 which is not considered an improvement.



1. We can include different attributes other ‘x6’ in the input tensor described in the preprocessing of data. To pick which attributes to choose we calculate the cross correlation between the variables and we choose the ones that has high correlation with ‘x6’.

The correlation is calculated for rows with location 6 only.

Correlation between x1 and x6= 0.4579422828022259

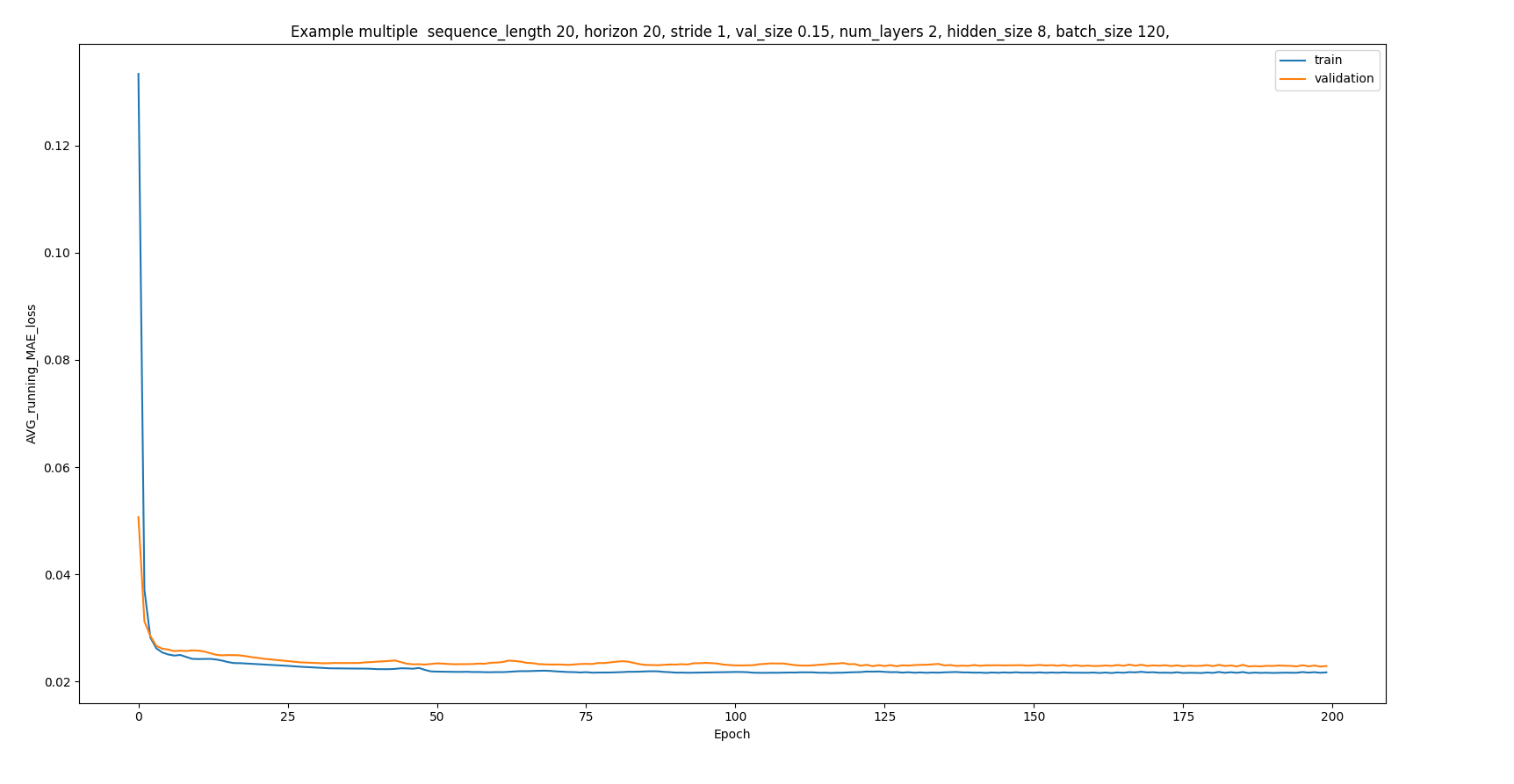
Correlation between x2 and x6=0.03202997014885508

Correlation between x3 and x6=0.47334917730620185

Correlation between x4 and x6=0.3532249338846274

Correlation between x5 and x6=0.4914658456551111

We add variables x1,x3,x4, x5, and x6



The minimum validation error is 0.022755 which is considered an improvement. The optimum number of epochs is 199.

**Final Model**

The final model will have the following parameters:

Sequence Length= 20

Stride =1

Number of layers =3

Batch Size =120

Number of Epochs=200

The variables used 'x1','x3','x4', 'x5', 'x6'

This model is chosen since it has the lowest validation data error (MAE).