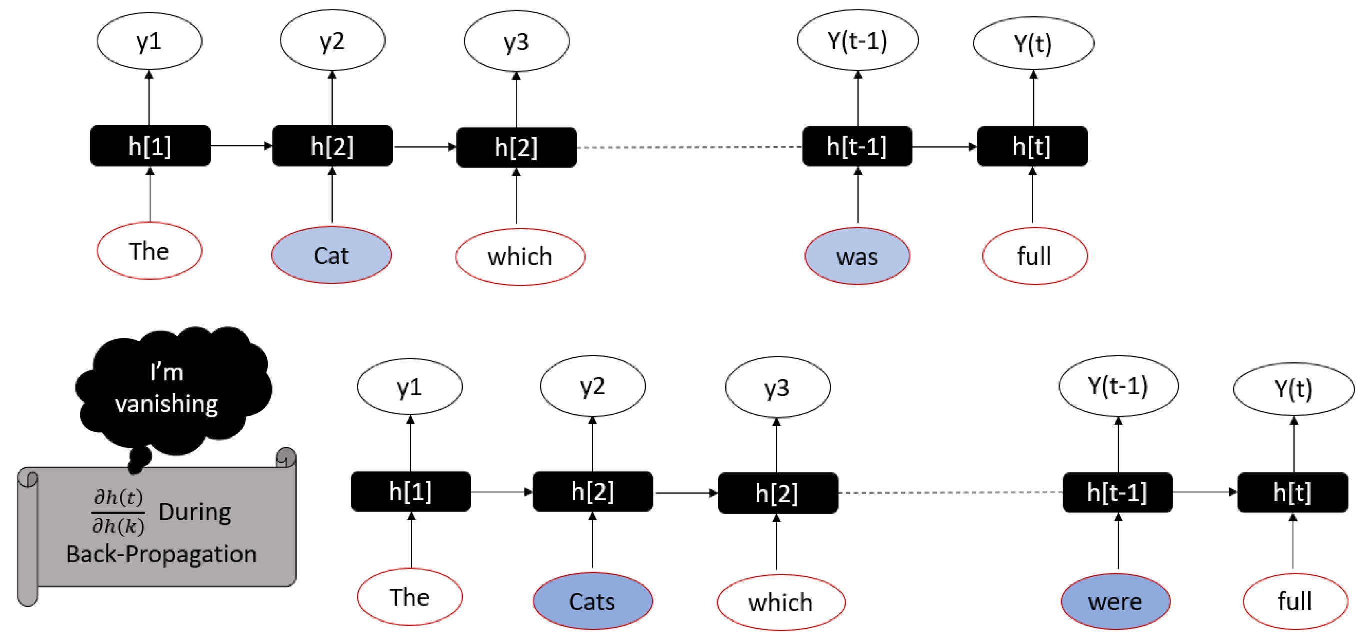
LSTM: Long Short Term Memory

Aviral Verma

CS460 : Machine Learning

National Institute of Science Education and Research, Bhubaneshwar

Problem of Vanishing Gradient



A standard recurrent neural network (RNN) could face the vanishing gradient problem during the training when long temporal sequential input data is provided.

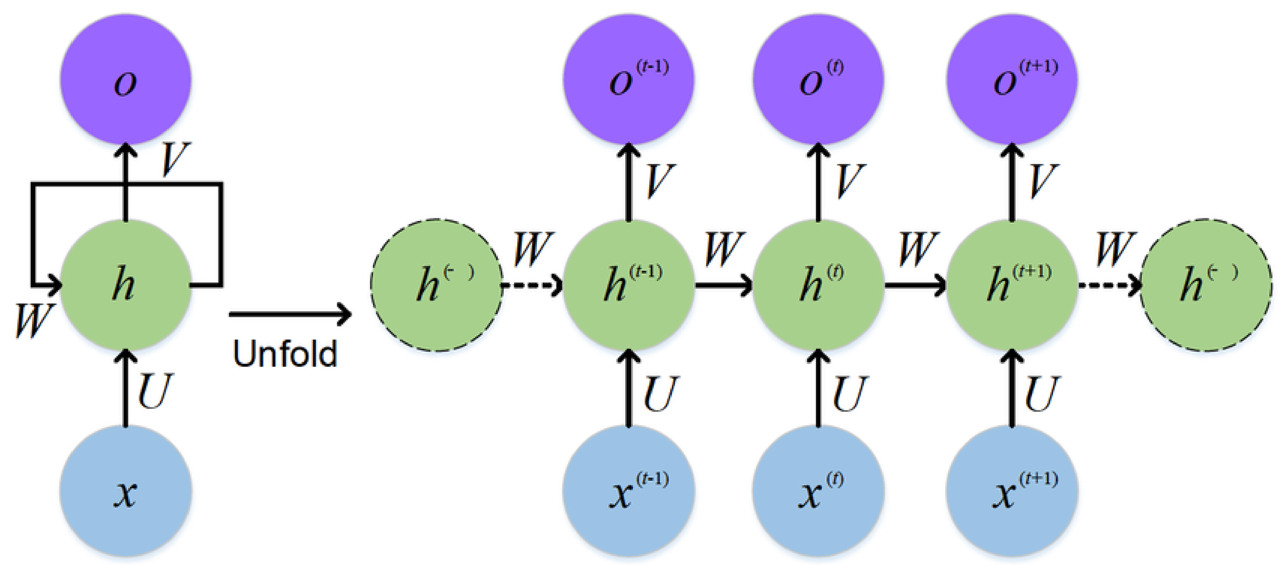


Figure 1. Recurrent Neural Network

An RNN model is usually trained by Back Propagation Through Time (BPTT). To update the weights, we find the gradient with respect to each parameter. Consider L is the loss function. During the training we backpropagate the gradients over time to update the weights.

$$\frac{\partial L^{(t)}}{\partial W} = \frac{\partial L^{(t)}}{\partial O^t} \cdot \frac{\partial O^t}{\partial h^{(t)}} \cdot \left(\Sigma_{k=1}^{t} \frac{\partial h^{(t)}}{\partial h^{(k)}} \cdot \frac{\partial h^{(k)}}{\partial W} \right)$$

Where,

$$\frac{\partial h^{(t)}}{\partial h^{(k)}} = \prod_{i=k+1}^{t} \frac{\partial h^{(i)}}{\partial h^{(i-1)}}$$

Overcoming the Vanishing Gradient Problem with LSTM

LSTMs are a special kind of RNN, capable of learning long-term dependencies. The architecture of LSTM includes specialised mechanisms that allow it to store and retrieve information over long sequences.

Understanding LSTM Architecture

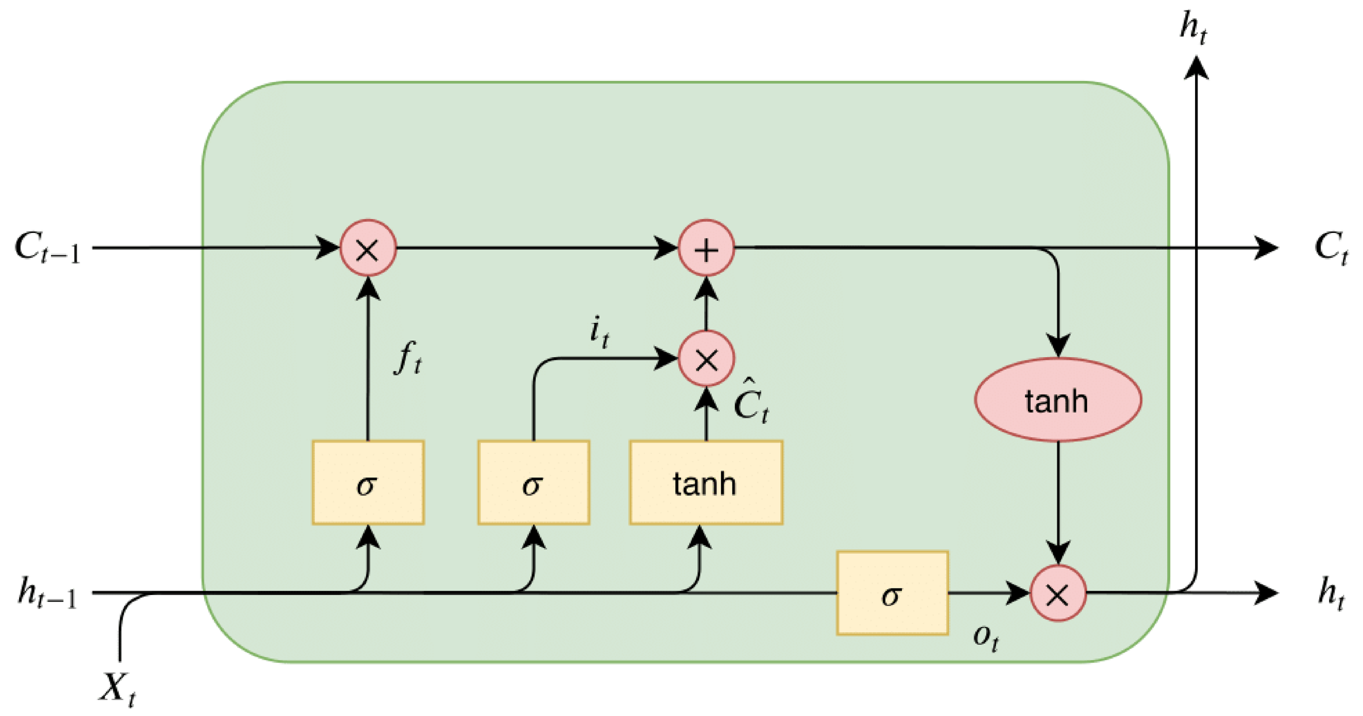


Figure 2. LSTM Cell Architecture

- **Forget Gate (f_t) :** Determines which information from the previous cell state C_{t-1} should be discarded or forgotten.
- **Learn Gate ($\hat{C}_t$) :** Captures new information from the current input X_t and the previous hidden state h_{t-1}.
- **Input Gate (i_t) :** Regulates the flow of new information into the cell state C_t.
- **Remember Gate (C_t) :** Allows LSTM to "forget" irrelevant information and retain the necessary information from current input and previous cell state.
- **Output (O_t) :** Decides which information from the cell state should be output as the hidden state, controlling the extent to which the cell state influences the network's predictions.

Operations of LSTM

Forget Gate: $C_{t-1} \odot f_t$

$$f_t = \sigma(W_f * h_{t-1} + W_f * X_t)$$

Learn Gate: $\hat{C}_t \odot i_t$

$$\hat{C}_t = tanh(W_n * h_{t-1} + W_n * X_t)$$
$$i_t = \sigma(W_i * h_{t-1} + W_i * X_t)$$

Remember Gate:

$$C_t = [(C_{t-1} \odot f_t) \oplus (\hat{C}_t \odot i_t)]$$

Output Gate: h_t = $tanh(C_t) \odot O_t$

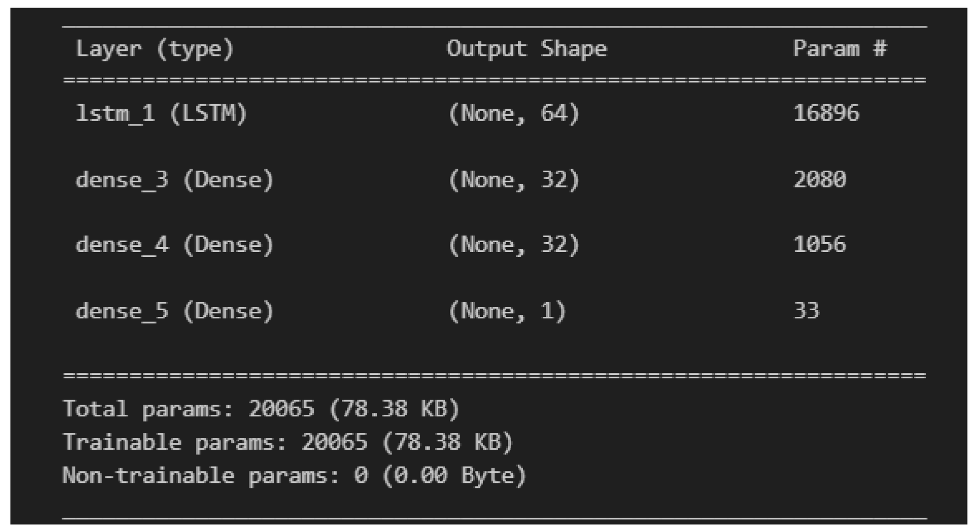
$$O_t = \sigma(W_v * h_{t-1} + W_v * X_t)$$

Application of LSTM

Predicting the Trend of a Stock with the help of LSTM Neural Network

DATASET: We will be using the last 4 years of Microsoft Corporation (MSFT) stock data

Model Details



Layer (type)	Output Shape	Param #
lstm_1 (LSTM)	(None, 64)	16896
dense_3 (Dense)	(None, 32)	2080
dense_4 (Dense)	(None, 32)	1056
dense_5 (Dense)	(None, 1)	33

Total params: 20065 (78.38 KB)
Trainable params: 20065 (78.38 KB)
Non-trainable params: 0 (0.00 Byte)

Figure 3. Model details

Result

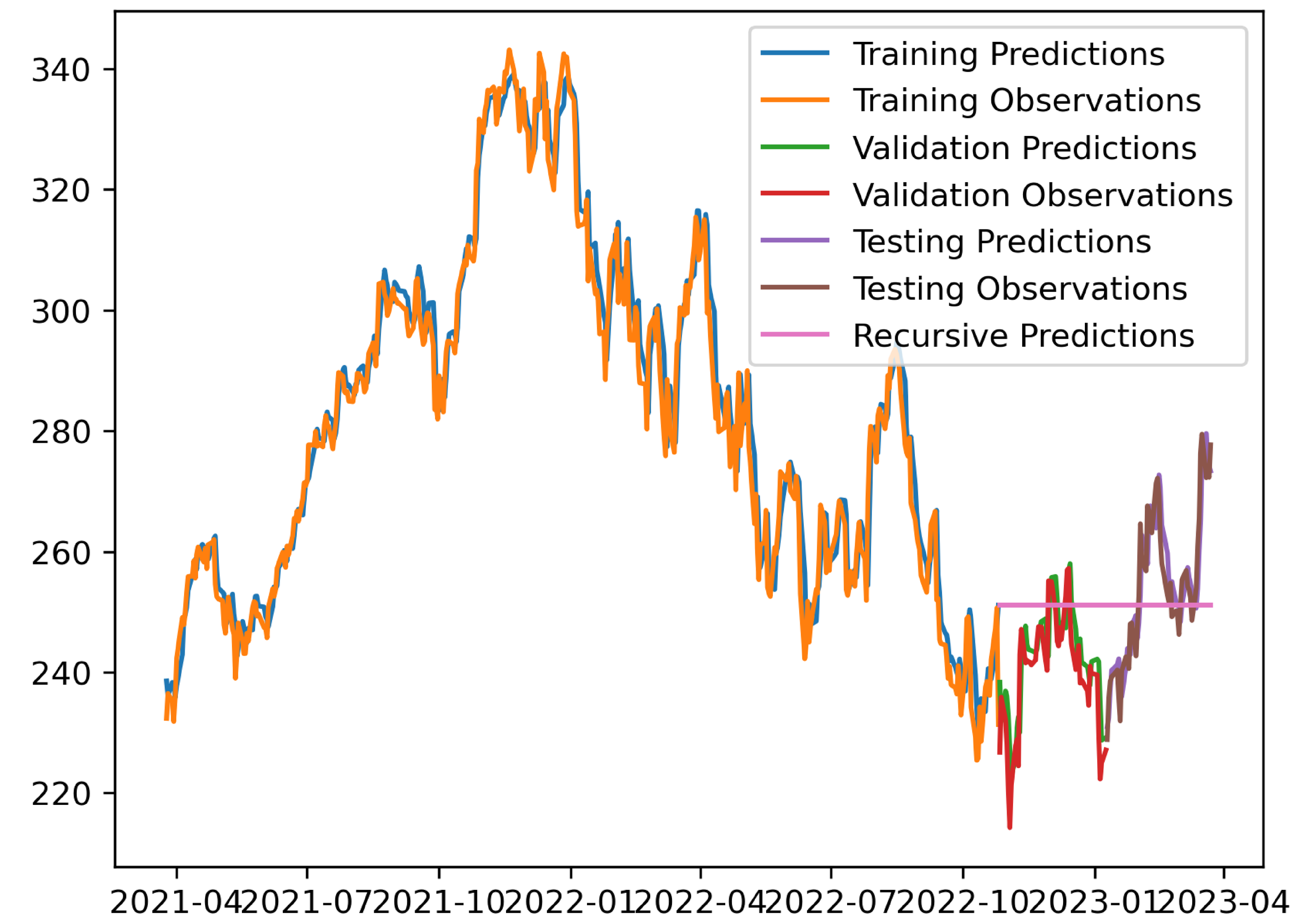


Figure 4. Prediction of Data and comparison with Recursive model

References

1. Olah, C. (2015) 'Understanding LSTM Networks', colah's blog, 27 August.

aviral.verma@niser.ac.in