Using Dynamic Neural Networks for Server Load Prediction

Pavlo Pukach[1], Volodymyr Hladun[2]

Lviv Polytechnic National University, Department of Applied mathematics

[1]pavlopukach@gmail.com, [2]v_hladun@yahoo.com

Abstract. In this paper, the approach of using neural networks for making time series predictions of strongly nonlinear data is used. A brief examination of neural networks usage for time series predictions is given, as well as the definition and schematics of LSTM blocks. The comparison between the time delay values and the prediction accuracy is given. It is shown that certain values of time delay can greatly increase the prediction accuracy.

Keywords: neural networks, long short-term memory, time series prediction;

1 Introduction

Nowadays, there is a considerable amount of web applications running all over the world. In a classical web app architecture, a physical or virtual machine, called a node, is required for running the application. The node's characteristics, such as CPU and RAM amount directly affect the performance of the application – namely the amount of requests it is able to serve concurrently. Most applications require a single node for their deployment, but when the amount of its users gets significantly higher, a single node is not enough for optimal application performance.

Modern top-tier web apps rely greatly on scaling. They run on a cluster of nodes, with each node containing a single instance of the application, behind a load balancer tool, which decides what requests should be served by particular nodes depending on their actual workload. The issue with this setup is that the number of requests (the application or server load) varies over time. Thus, for effectively serving requests to end users, the system has to increase or decrease the number of nodes depending on user activity.

The goal of this paper is to use dynamic neural networks (DNN) for predicting server load. Accurate predictions can ensure that the node count is close to optimal, so there is no over-usage of computing time. This can greatly reduce application hosting costs, especially with cloud providers like Amazon EC2, which give a possibility to request nodes on-demand.

COLINS'2018, Volume II: Workshop. Lviv, Ukraine, June 25-27, 2018, ISSN 2523-4013
http://colins.in.ua, online

1.1 Neural Networks

Neural Networks approach is usually involved in time series predictions in which traditional prediction may not be able to capture the non-linear pattern in data[1].

Neural Networks can be classified into two categories: static and dynamic. Static (feedforward) networks have no feedback elements and contain no time delay. In another words, the output is calculated directly from the input through the connections. In dynamic network, the output depends not only on the current input, but also on the previous inputs to the network or estimated output of the network [2]. In this paper, we will use LSTM networks.

Long short-term memory (LSTM) units (or blocks) are a building unit for layers of a recurrent neural network (RNN). A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. Each of the three gates can be thought of as a "conventional" artificial neuron, as in a multi-layer (or feedforward) neural network: that is, they compute an activation (using an activation function) of a weighted sum. Intuitively, they can be thought as regulators of the flow of values that goes through the connections of the LSTM. There are connections between these gates and the cell.

The **time delay** (TD) is the number of previous network outputs that are taken into account for prediction of the next value.

2 Server load prediction

For the purposes of predicting server load based on time series, we conducted several experiments, using LSTM neural networks. The sample dataset that was used in this work, is derived from the request count logs on a AWS EC2 container. The numbers represent the web application request count per 5 minutes. The size of the dataset is 2016 entries, representing roughly one week of application request count logging.

Several LSTM neural networks were developed, with different values of time delay *d* – 1, 25 and 30.

We have trained the three networks on 90% of the sample dataset, and made test predictions on the rest of it. To compare the results, the RMSE (root mean square error) metric is used:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2}$$

3 Conclusion

Comparing the results, it is shown that the prediction accuracy varies depending on the time delay. For the first neural network, it is seen on the graph (Figure 1), that the network does not take into account the peaks of server load. The second and the third networks have greater values of time delay, and their predictions (Figure 2, Figure 3) are way more accurate. It is also shown that all of the networks do not predict the two high peaks at and *t*=159. These peaks might represent a DOS-attack or some other abnormal cases, and such anomalous behavior is extremely hard to predict. However, the overall prediction results tend to get the more accurate the more the value of the time delay is. It follows from here, that dynamic NN's with certain values of time delay can make better predictions of strongly nonlinear data, such as server load.

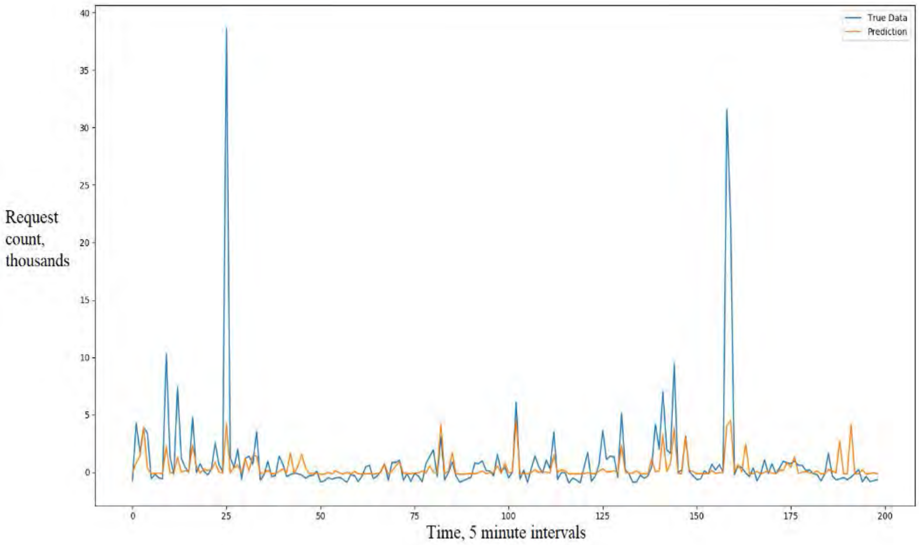

Fig. 1. Load prediction (*d*=1, *RMSE*=3.60)

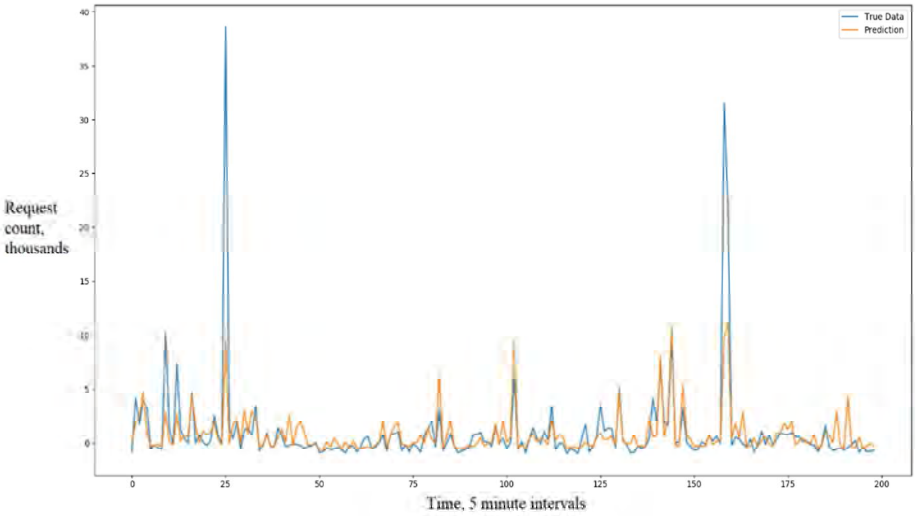

Fig. 2. Load prediction (*d*=25, *RMSE*=3.02)

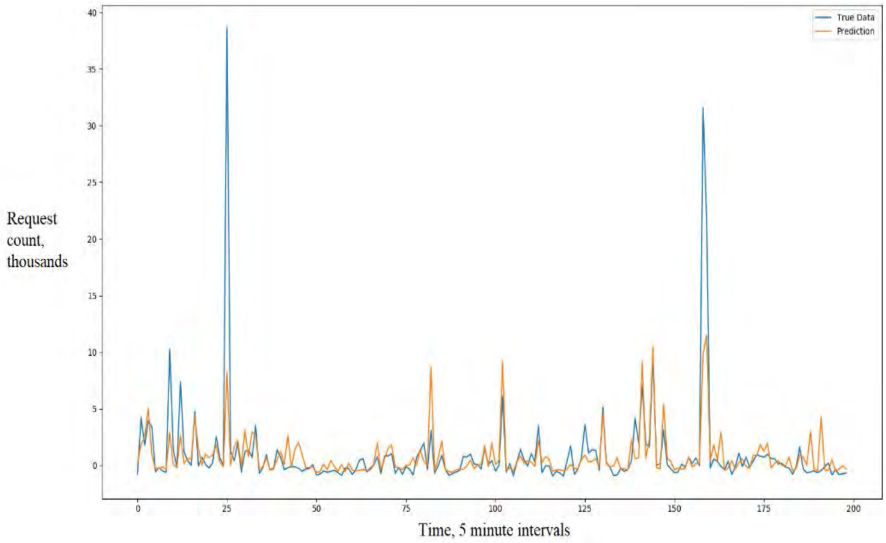

Fig. 3. Load prediction (*d*=30, *RMSE*=2.95)

References

1. M. Beale, M. Hagan, and H.Demuth, “Matlab neural network toolbox user’s guide,” The Math Works Inc., 2010, http://www.mathworks.com/help/pdfdoc/nnet/nnet ug.pdf.
2. C. A. Mitrea, C. K. M. Lee, and Z. Wu, “A comparison between neural networks and traditional forecasting methods: A case study,” in International Journal of Engineering Business Management, vol. 1, no. 2 2009, pp. 19–24.