Paper Review on Web Traffic Prediction of Wikipedia Pages

This paper presents a forecasting model to predict the web traffic. The prediction of web traffic can help the site owner in many ways, including better understanding of the user behaviour and determine an efficient strategy for load-balancing for future loads. This has made forecasting an active area of research. The model in this paper focuses on the temporal observations that emerge from analysis and forecast.

The model assumes that the medians considered over a window of exponentially varying sizes, and their median may also affect the forecast and includes them in the features of the input to the model. The model aims at bringing out the relationship between the magnitude of the features used for prediction model, its overall performance and the optimal number of features required while maintaining high accuracy prediction results.

The data, Google's Web Traffic Time series forecasting, was obtained from Kaggle. It was then cleaned since the missing values were replaced with 0. The model was rebuilt from the RNN seq2seq model, using the encoder decoder architecture where cuDNN GRU was used to encode and TensorFlow GRUBlockCell as decoder with the output of the decoder passed on to the next step until the end of the sequence.

Rather than considering the entire time period, the model considers divided windows of time. The median in each time window is added as the new feature. A rolling window was used to provide more weightage to more recent data. The window started off with a smaller window size of 6 and followed the Fibonacci sequence to increase the size of the window. This rolling window was used to find the seasonality and trend.

The new model performed better than the previous model which is evident from the SMAPE: improved 0.351 to 0.349 for 804 days and 0.354 to 0.351 for 740 days in the training set. The model however, considers only 4 pages and does not incorporate the relations among the pages. In our point of view, the model may be sensitive external factors like spikes in traffic which could have occured during a DDoS attack.