Capstone Project Proposal 5 November 2019

Machine Learning Engineer Nanodegree

Deep Autoregressive Recurrent Models for Flood Prediction

Prashidha Kharel, PhD

Background

Flood is an extreme event in which the water overflows from a stream or river that can result in loss of properties and even lives if it happens near built environment. Heavy rainfalls over a long period of time can cause the river discharge levels to increase, resulting in flood. Various magnitudes of floods can recur in various frequency of time. For example 1 year flood (i.e. the highest discharge in a typical year) may not cause overflow, but once every 10 or 50 years, we can expect a high level of flood that can have devastating consequences to our society. Hence if we can predict such flood events in advance, we can save lives and property by conducting early evacuation or other countermeasures.

In this project, I will use historical rainfall and river discharge data in order to build a deep autoregressive recurrent model (DeepAR) in Sage Maker (Salinas et al. 2019) to predict the flood events in advance.

Problem Statement

The goal is to create a deep autoregressive recurrent model (DeepAR) in Sage Maker that can predict future river flow discharge level in a given river. It will investigate three possible models.

- A deep neural network model to predict hourly discharge using single rainfall input.
- A deep neural network model to predict downstream hourly discharge using multiple rainfall and upstream discharge data as input (upto 7 hours into the future).
- A deep neural network model to predict downstream daily discharge using multiple rainfall, sun exposure, temperature and upstream discharge data as input (upto 7 days into the future).

Datasets and Inputs

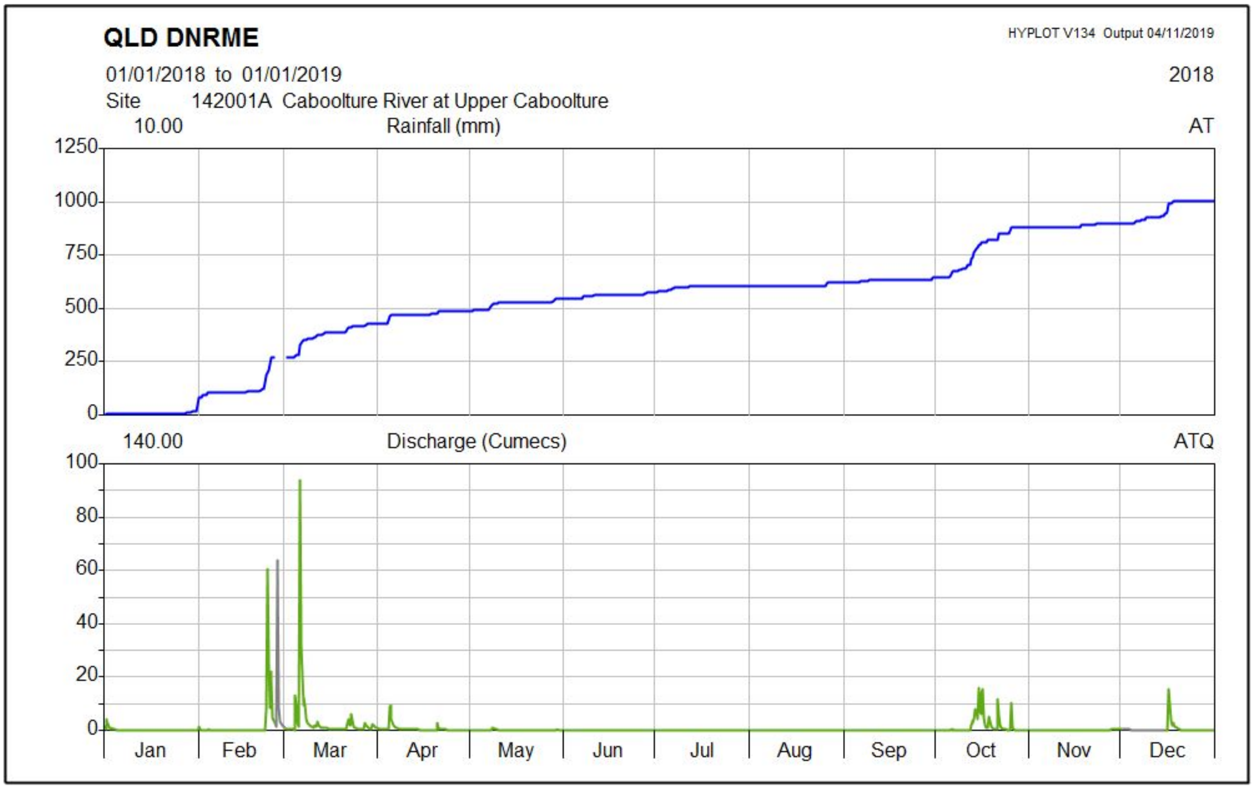
The hydrological data available at Queensland Water Monitoring Information Portal[1] will be used. The selected catchment is. Hourly as well as daily water flow data is available at various stations. Rainfall data is available in some of the stations. I shall utilise all the upstream discharges and rainfall data to predict the discharge downstream. The data are time series with hourly or daily discharge and rainfalls.

For hourly discharge prediction, the idea is to use rainfall and upstream flow data to predict the flow downstream.

For daily discharge prediction, I'll look into other Meteorological Data like rainfall, temperature, solar exposure, etc in the surrounding area as input.

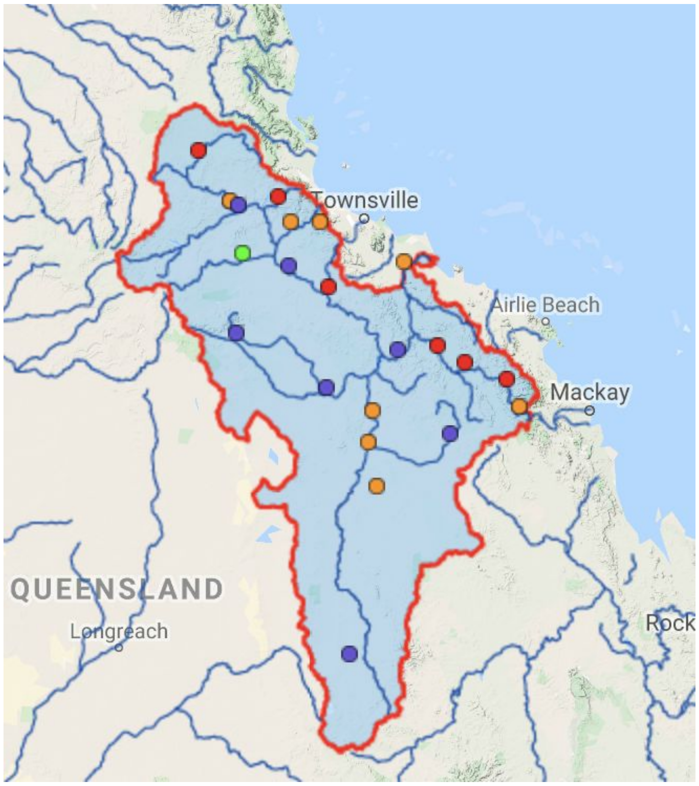
Since RORB only uses one rainfall data as input, in order to compare our deep learning model with the RORB, we'll be using Caboolture River at Upper Caboolture. It consists of hourly rainfall and discharge data from 1995 to 2019. Here is an example of rainfall and discharge data for 2018.

[1] Queensland Water Monitoring Information Portal. URL: https://water-monitoring.information.qld.gov.au/



The first model will be composed of an hourly discharge as the target time series and a single rainfall time series as input variable. This model will be used for comparision with RORB's predictions.

For the second model, we will look into Burdekin basin which consists of a number of rainfall and discharge stations shown below. A total of 17 hourly rainfall and discharge stations were selected (as per period of availability of data) as input for the model for data ranging from 1995 to 2019. This model will consist of one target time series discharge at downstream and 16 feature series. Hence a total of 20 years worth of training data will be used. This turns out to be about 175,000 hours of data.



For the final daily discharge model, the data will be converted to daily rainfall and dischage. Furthermore, daily solar exposure, temperature and vapour pressure data from nearby stations will be used as training data. Hence, this model will consist of one target time series at the downstream and 27 feature series. Hence a total of 20 years worth of training data will be used. This turns out to be about 7300 days of training data.

Solution Statement

Since we are working with time series data, the DeepAR model will be used to create the model time series prediction model. The target input of the model will be the discharge at the downstream of the catchment. Other input features will include upstream stations flow data and rainfall data. In a river catchment almost all of the rain and discharge will ultimately reach the downstream end. Hence I'll use all available rainfall and upstream discharge data as input for our model to predict the discharge at the downstream end of the catchment.

Benchmark Model

RORB[2] model is generally employed for calculating design flood discharges. It uses many assumptions and is manually calibrated to one flooding event. This will be used as benchmark model for comparison purpose.

Evaluation Metric

Root Mean Squared Error (RMSE) will be used to evaluate the models prediction. RMSE will be defined for each prediction time intervals.

$$\mathrm{RSME}(\Delta t) = \frac{1}{N}\sum_{t}(y_{t+\Delta t} - \hat{y}_{t+\Delta t})^2$$

Here,

$\hat{y}_{t+\Delta t}$ is a predicted discharge Δt time into the future, using input data up to t.

$y_{t+\Delta t}$ is the actual discharge Δt time into the future.

The overall RSME is then defined as the average for all interval predictions. I aim to do at least 7 intervals worth of prediction. RSME will be preferred because the prediction for higher flood level is of more important than small flood levels.

Project Design

The project will be divided into the following stages.

Data preprocessing

All required data will be downloaded and then processed into the required format. This includes hourly as well as daily discharge data. Any other meteorological data will also be looked out for, for example wind, sunshine, cloudcover etc.

Deciding which features to use for both hourly and daily models. Extracting extra features from available data that could help in the discharge prediction.

---

[2] RORB software for runoff routing.URL: https://www.monash.edu/engineering/departments/civil/research/themes/water/rorb

All the time series data will be combined into a single data frame based on the timestamp. Missing data will be handled by replacing them with a moving average. All the data will be rescaled by dividing by the maximum value.

The data will then be divided into training, validation and test set. The last couple of years worth of data will be used for test and the preceding couple of years as validation, while the rest of the data as training data. Validation set will be used to do hyperparameter tuning.

Model Training and Tuning

This involves testing various model types and tuning of the hyperparameters. The aim will be to reduce the loss for the validation set. Some of the hyperparameters include prediction length, drop out rates, learning rates and number of layers in the model. The DeepAR implementation in SageMaker will be used for the training product.

Comparison with Benchmark

The final model will be deployed and tested with the test data. An RORB model will also be calibrated and tested on the first model. Then the performance of the two models will be compared. For the second and third model, I will experiment with various combinations of input features, e.g. using solar exposure vs not using solar exposure as input, etc. The outcome on the difference in performance will be discussed.

Concluding Remarks

Any strengths and weaknesses of the model will be discussed with any recommendations for future flood prediction modeling. Recommendation on how this model can be used in other catchments and its use will also be discussed.

References

Salinas, David, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. 2019. “DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks.” *International Journal of Forecasting*. https://doi.org/10.1016/j.ijforecast.2019.07.001.