Aidan LaBella
DATA/EEPS 1720
Project Milestone 3
4/18/2024

Forecasting Air Quality Metrics With Evolved Recurrent Neural Networks

1 Introduction

The main question my project attempts to answer is "can we **forecast** air quality data when sensor data is unreliable and/or not present"? This project, at a high level, is to train a model that is effective at predicting air pollution given certain parameters. I also propose using a toolkit for neural architecture search while training the model. The applications of a model that can predict air quality vary from being able to estimate the pollutants in the air based on other parameters, to being able to accurately forecast when the pollution will reach its high/low levels, much like a weather forecast.

Recurrent Neural Networks, or RNNs are an effective way to regress over a dataset. Since RNNs are able to output a set of parameters for a variable number of time-steps, this makes them ideally suited for time series forecasting. Air quality data takes the form of hourly readings from sensors that measure different contents of elements in the air. As such, RNNs can be said to be well-suited for this type of time series forecasting. However, one of the drawbacks that RNNs bring is a computational overhead due to their need to be "unrolled" through time. Because of this, creating smaller recurrent neural networks has been a priority in the DL field but finding the right architecture has been a challenge. Using the Evolutionary eXploration of Augmenting Memory Models (EXAMM)[4], we can "evolve" recurrent neural networks to (hopefully) be the "smallest" yet best performing model for a given task. Not only does this reduce the complexity of the model, it can reduce the computational cost which in turn makes the model more environmentally friendly. We will show that our evolved networks can achieve the same levels of performance as a traditional RNN in PyTorch or Tensorflow.

2 Related Work

Previous applications of the EXAMM[4] toolkit with neural architecture search (NAS) include making predictions for coal power plants[1]. We seek to use this work as a foundation and proof-of-concept for our task of forecasting air quality data.

2.1 The EXAMM toolkit

The EXAMM toolkit was developed by the Distributed Data Science Systems lab (D2S2) at Rochester Institute of Technology, before the "explosion" of modern neural network frameworks like PyTorch and Tensorflow. EXAMM is written entirely in C++ and utilizes cpu-bound MPI and multithreaded compute methods in lieu of GPUs. EXAMM also focuses on feed-forward and recurrent neural networks for its evolutionary processes. This is due to the fact that neuro-evolution for CNNs is a bit trickier to conduct due to the intensive computational resources required for training and evaluation. One added benefit of a NAS apparoach is its ability to identify the necessary input columns for a successful prediction. In turn, this will allow us to answer **RQ1**.

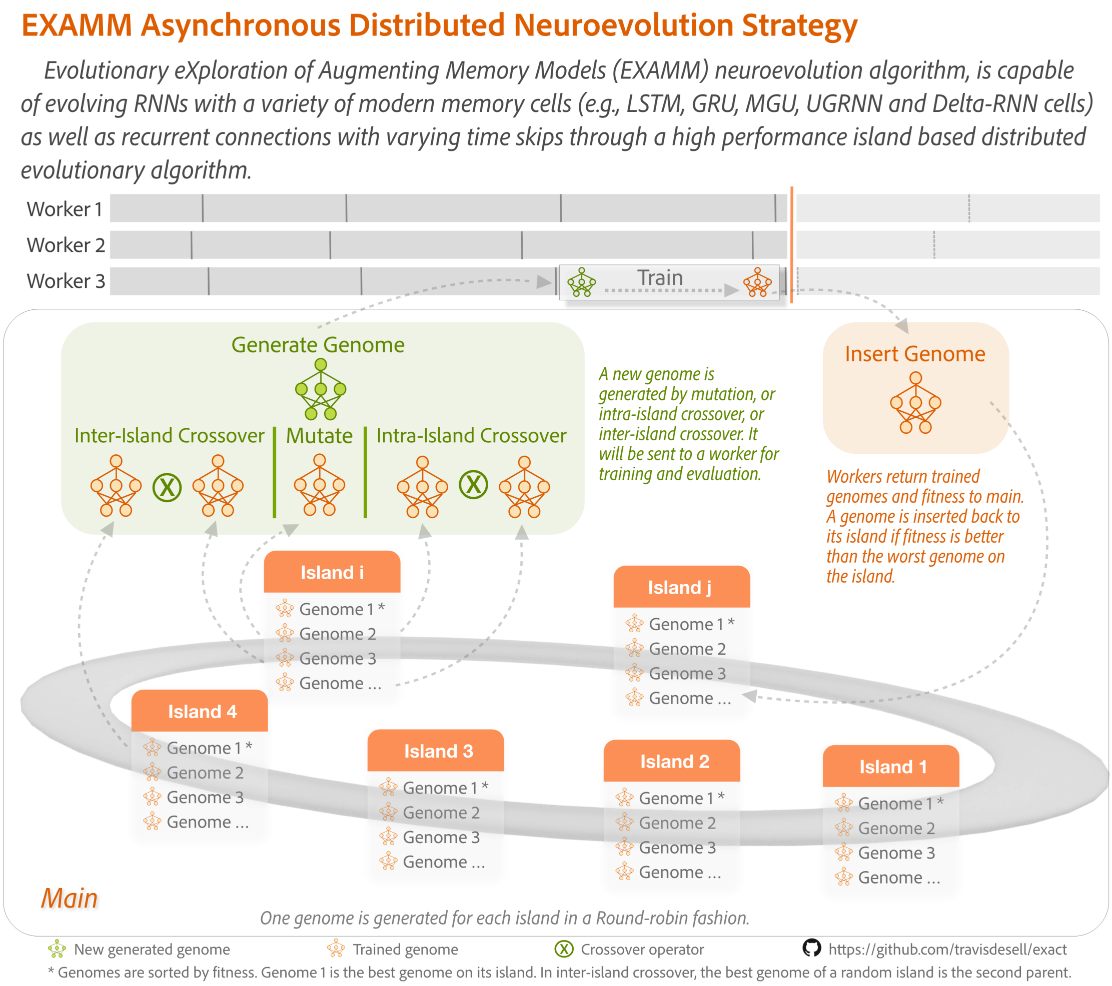

Figure 1: EXAMM's evolutionary process [3]

3 Dataset(s)

For this project, multiple datasets were explored including the UCI Air Quality repository and the NASA MERRA-2 dataset.

3.1 UCI Air Quality

It is important to first note that this work can be classified as a multivariate time series classification forecasting task. The data that will be used to train the models will be multivariate in form, with one or more columns potentially missing or corrupted to simulate unreliable data. To train a well-performing recurrent neural network, we must first find a well-curated dataset that has enough sensor readings to learn from. The UC Irvine Air Quality dataset [5] uses 9358 instances of hourly averaged responses from an array of 5 sensors within an Italian city from March 2004 to February 2005, approximately 1 year in time.

3.2 MERRA-2

The MERRA-2 [2] dataset was chosen as a more complete and representative dataset with regard to US air quality. For the purposes of this project, we will look at the following columns:

Parameter	**Description**
AIRDENS	Air density, in kilograms per cubic meter (kg/m^3)
SO4	Sulfate concentration in micrograms per cubic meter ($\mu g/m^3$)
SO2	Sulfur dioxide measured in parts per billion (ppb)
RH	Relative humidity as a percentage
PS	Atmospheric pressure, in hectopascals (hPa) or millibars (mbar)
H	Hydrogen concentration, in parts per million (ppm)
O3	Ozone concentration, in parts per billion (ppb)
T	Temperature of the air, in Kelvin (K)
U	Horizontal wind speed, meters per second (m/s)
V	Vertical wind speed, in meters per second (m/s)
CO	Carbon monoxide concentration in parts per million (ppm)

Table 1: Glossary of MERRA-2 Parameters

Parameter Forecasting To have a meaningful measure of the air quality at a given point in time, we will narrow our focus to the following particulate air measures. Our main focus will be on CO levels but for mutivatiate output layers we will look at additional parameter levels.

1. CO or Carbon Monoxide
2. O3 or Ozone Concentration
3. SO_4 or Sulfate Concentration

4. SO_2 or Sulfur Dioxide Concentration

4 Exploratory Analysis

For a "proof of concept", the entire dataset was used to "evolve" network(s) and evaluate their performance. In NAS, evolution refers to the process of augmenting the model's architecture through random mutations, followed by gradient descent for adequate evaluation. In the experiments conducted, we attempt to impute the ground truth readings based on date, time, temperature and other sensor readings. We evolve the networks for 2000 genomes (series of mutations) with 5 epochs of gradient descent for evaluation. The preliminary MSE values of well-performing genomes (models) for univariate ouputs has been found to be in the ballpark of $[.02, .04]$, an acceptable range for any deep learning model. For multivariate outputs, the loss was higher (≥ 0.2). Thus, I am confident that an evolved RNN will be well-suited to make predictions for this task.

5 Methodology

The methodology of this work will focus around the following research questions:

RQ1: What input columns can accurately estimate the levels of pollutants such as Carbon Monoxide, CO?

RQ2: Can multiple prediction columns exist in the output layer with minimal performance impact? I.e. can we predict multiple columns at once?

RQ3: How do our **forecasted** values compare with actual sensor readings?

RQ4: What is the inference time of the best performing model? How does it compare to a standard Jordan, Elman and standard RNNs?
Since one of the goals of this project is to be able to forecast air quality parameters

5.1 Multivatiate Output Layer

5.2 Univariate Output Layer

6 Evaluation

To evaluate both **RQ1** and **RQ2**, both mean average error (MAE) and mean squared error (MSE) will be used. Plots nd distributions will also be created to showcase the models performance for *each* timestep in the series.
We then make two plots, one with real versus predicted and the absoulte error between then (per time-step). This aims to answer **RQ1** & **RQ3**.

7 Acknowledgments

I want to thank my B.S. advisor, Dr. Travis Desell and my former lab-mates, Josh Karns and Zimeng Lyu at RIT for their help and support with the EXAMM code base!

References

[1] AbdElRahman ElSaid, Steven Benson, Shuchita Patwardhan, David Stadem, and Travis Desell. Evolving Recurrent Neural Networks for Time Series Data Prediction of Coal Plant Parameters. In Paul Kaufmann and Pedro A. Castillo, editors, *Applications of Evolutionary Computation*, volume 11454, pages 488–503. Springer International Publishing, Cham, 2019. Series Title: Lecture Notes in Computer Science.

[2] Ronald Gelaro, Will McCarty, Max J Suárez, Ricardo Todling, Andrea Molod, Lawrence Takacs, Cynthia A Randles, Anton Darmenov, Michael G Bosilovich, Rolf Reichle, et al. The modern-era retrospective analysis for research and applications, version 2 (merra-2). *Journal of climate*, 30(14):5419–5454, 2017.

[3] Zimeng Lyu, Alexander Ororbia, and Travis Desell. Online evolutionary neural architecture search for multivariate non-stationary time series forecasting. *Applied Soft Computing*, 145:110522, 2023.

[4] Alexander Ororbia, AbdElRahman ElSaid, and Travis Desell. Investigating recurrent neural network memory structures using neuro-evolution. In *Proceedings of the Genetic and Evolutionary Computation Conference*, GECCO '19, pages 446–455, New York, NY, USA, July 2019. Association for Computing Machinery.

[5] Saverio Vito. Air Quality. UCI Machine Learning Repository, 2016. DOI: https://doi.org/10.24432/C59K5F.