Marcus Klausen
Daniela Simonelli
COMP 451

Sunday, December 15th, 2024

Final Report: Prediction of Visible Auroras

Source Code: https://github.com/MarcusFriisKlausen/COMP451-AuroraProject

1. Introduction

1.1 Background & Motivation

Interactions between solar activity and the Earth's magnetic field is a key area of study in space weather research. Solar flares are intense bursts of radiation released from the sun, which cause magnetic storms, that can cause auroras to form in the ionosphere.

Not only is the prediction of auroras of scientific interest but also practical interest. On one hand, we wish to predict auroras, so people know when they can expect to experience the beautiful phenomenon, but on the other, auroras have negative implications on technology, as they directly impact systems such as HF radio and satellite navigation.

Auroras are inherently hard to predict, as they depend on solar activity, so this project intends to implement a robust machine learning model, which makes use of various solar flare parameters such as intensity and duration, to better predict auroras.

1.2 Objectives

Some objectives for the project are to have accurate predictions on the visibility of an aurora. A procedure of testing a set number of dates with their given data on visible auroras would be able to measure such accuracy along with data on auroras on correlating dates. We would like to create a model that is robust so that it can accurately predict the visibility of an aurora no matter the season or solar cycle. Our model should not be overly dependent on specific data trends such as solar cycles and seasons, but instead it should ensure consistency. To do this, we will try to make sure our testing data has an equal number of dates ranging throughout at least a whole year in order to capture all seasons. However, since a solar cycle generally lasts around a decade, it may be a lot harder to do, so to be able to at least predict throughout half the cycle of a solar flare could also be effective since the data would still be taken from different solar trends.

2. Literature review

We want to create a machine learning model for predicting if aurorae will be visible on given days, and thus we must determine the relevant data features that cause aurorae to form. Generally, it is believed that bright aurorae form in the ionosphere due to Birkland currents. Birkland currents are electrical currents flowing from one of the Magnetosphere's poles down into the Ionosphere and are driven by the solar wind (Schield et al., 1969). Thus, important data for our model is gonna be data regarding solar wind and data regarding the conditions of the Ionosphere and Magnetosphere. Trivially we will of course need data indicating whether aurorae have been visible given solar, Ionosphere, and Magnetosphere data on a given day, as the model must predict visibility of aurorae on a given day.

Natras et al. (2021) mentions that the most important factor in space weather prediction is geomagnetic storms as well as VTEC (vertical total electron content). They also mention the importance of the temporal resolution of the data.

For solar data we can look to a study done by Natras et al. (2023) where they train a machine learning model to predict space a weather, which is highly connected to what our model should do. Here they use the following data points from NASA's OMNIWeb data service: "sunspot number, F10.7 solar radio flux, solar wind plasma speed, interplanetary magnetic field Bz index, geomagnetic field (GMF) Dst index, GMF Kp index, auroral electrojet (AE) index" (Natras et al. 2023). Here they also specify the longitude and latitude points for the VTEC data, along with the specific features derived from the VTEC, being the exponential moving average of VTEC over the last 30 and 4 days and first and second VTEC derivatives. As they are trying to forecast specific VTEC values in the Ionosphere, and we just wish to predict visible aurorae, we might not need a dataset as complicated, but that is to be tested later in the report.

In addition, Nature posted a report on “an automated detection system using deep learning” (Nanjo et al. 2022). This system uses pictures of citizen scientists that capture images of auroras in Norway, Tromsø to classify diverse types of aurorae.

Some of the interesting aspects they considered are the use of a wide range of digital cameras, the 11-year solar cycles, the rise of occurrence of aurorae in autumn and spring (Nanjo et al. 2022). They also compare multiple methods and their accuracies. Marcov models, k-nearest

neighbors and SVMs were discussed as having been used for automated classification of auroral images in the past (Nanjo et al. 2022). However, they were deemed slightly less efficient compared to newer models due to their dependency to manually select features. They decided to therefore use deep neural networks, specifically a model called ResNet-50 which reached an accuracy score of 92% (Nanjo et al. 2022). However, their work differs from us as they predict if an aurora was present and its type from images, while we will work with confirmed data on the aurora to predict the likeliness of future sightings.

In the topic of models, one of the models used by Natras et al., in a similar project, was an ensemble of convolutional neural networks with some least squares influence for the geomagnetic Dst index (which measures magnetic activity) prediction (2023). This, along with a class-balanced loss function, helped address the imbalance between storm and non-storm cases and deliver probabilistic Dst predictions (Natras et al. 2023). On the other hand, an artificial neural network was also deemed helpful to predict thermosphere density which can greatly affect the visibility of aurorae (Natras et al. 2023). They also implemented a Bayesian neural network, which they describe as similar to an artificial neural network with the added detail of using probabilities of weights instead of the actual weights. It was used to predict the geomagnetically induced currents (Natras et al. 2023).

3. Methodology

3.1 Data

The dataset our project is based on is first and foremost the Aurorasaurus Real-Time Citizen Science Aurora Data Version v1.0 (Kosar et al. 2018) dataset. It contains verified citizen sightings of auroras from all of 2015 and 2016. We combined this with the Kaggle dataset "Viewing Solar Flares" based on a study from Milligan et al. (2017). Processing these datasets, we ended up with a dataset with dates ranging from January 1, 2015, to December 31, 2016, with average solar flare duration, when the flares peaked on average from when they started, the flare count for the given date, and lastly a binary indicator for if at least one aurora was sighted that day.

Mirroring Natras et al. (2023), we used NASA's OMNIWeb data service to get more complex features. To our dataset we added the daily averages of parameters: sunspot number,

F10.7 solar radio flux, solar wind plasma speed, interplanetary magnetic field Bz index, geomagnetic field (GMF) Dst index, GMF Kp index and auroral electrojet (AE) index.

All this was achieved by creating, concatenating and manipulating pandas dataframes in a Python Jupyter Notebook which can be found as "/data_preprocessed/dataprocessing.ipynb" in our project source code.

3.2 Model & Training

For the model we chose to build and train a deep neural network for binary classification. Using Keras for the Tensorflow package, the best structure we found was 3 dense layers with ReLU activation into an output layer using sigmoid activation. The dense layers have 128, 64, and 32 neurons and between each of those layers a dropout of 50% is applied.

We randomly split the dataset into 80% for training and 20% for testing. The 80%-part was then split into another 80/20-split where the ladder is used as a validation split. The model was trained over 50 epochs using Adam, a variant of stochastic gradient descent, as optimizer and binary cross entropy as loss function.

4. Empirical Evaluation

In the below table we present our model's performance on each of the sets:

	Training	Validation	Test
Accuracy	0.823	0.786	0.798
Loss	0.466	0.497	

5. Discussion & Future Improvements

While an accuracy of ~80% is fine, we do believe it is possible to train an even better model for binary classification of aurora visibility. First of all, we must acknowledge that the sample size of our dataset is relatively small. With only 730 data points, including the data points for the test and validation sets, we suspect that the model has not been able to pick up on the most intricate patterns in the data. The reason for not collecting a bigger set of data is simply that we were unable to find any more publicly available aurora sighting data. Another reason our model most likely has not picked up on these patterns is, that we were not able to incorporate

daily VTEC values into our dataset, which from our literature review, we found were very vital in predicting space weather, which directly impacts the formation of auroras. These problems are reflected in the loss values, which we were not able to get lower than ~0.46-0.5 for our training and validation sets.

Lastly, it is important to point out that there might be aurorae visible on days not verified in the dataset, which might mean that the model doesn't pick up on certain patterns in the data, which it otherwise might have.

6. Conclusion

To conclude, we were able to train deep neural network on a dataset based on aurora sightings, solar flare data, and various averages on different solar parameters from NASA's OMNIweb achieving a test accuracy of 79.8%. This accuracy, while being fine, could along with the loss values for the training and validation sets most likely be improved in the future by collecting a larger dataset, which on top of that should include daily VTEC values.

7. Self-assessment

Overall, we're happy with how the model turned out given the size of our dataset, though it would have been nice to be able to include some VTEC features to see how much, if any, it would have improved the model. We did find VTEC data from NASA, but we found it very late, and the data was quite split up and formatted in a strange way, so it would have taken more time than we had left to properly process, so it could be ready for use. Given our small dataset, it might also have been interesting to train another model like Random Forest and compare its performance to the Deep Neural Network.

Bibliography

Nanjo, S; Nozawa, S.; Yamamoto, M.; Kawabata, T.; Magnar, J G.; Tsuda, T. T.; Hosokawa, K; "An automated auroral detection system using deep learning: real-time operation in Tromsø, Norway", https://www.nature.com/articles/s41598-022-11686-8, 2022

Natras, R.; Schmidt, M.; "Machine Learning Model Development for Space Weather Forecasting in the Ionosphere", 2021

Natras, R.; Soja, B; Schmidt, M.; "Uncertainty Quantification for Machine Learning-Based Ionosphere and Space Weather Forecasting: Ensemble, Bayesian Neural Network, and Quantile Gradient Boosting", https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2023SW003483, 2023

Schield, M. A.; Freeman, J. W.; Dessler, A. J.; "A source for field-aligned currents at auroral latitudes", 1969

Kosar, Burcu C.; MacDonald, Elizabeth A.; Case, Nathan A.; Heavner, Matthew "Aurorasaurus Real-Time Citizen Science Aurora Data" (V1.0) [Dataset], https://zenodo.org/records/1255196, 2018

MichaelKirk (Kaggle user), "Viewing Solar Flares" (V1.0) [Dataset], https://www.kaggle.com/datasets/heliodata/instruments-solarflares, 2017

Milligan, Ryan O; Ireland, Jack; "On the Performance of Multi-Instrument Solar Flare Observations During Solar Cycle 24", https://arxiv.org/abs/1703.04412, 2017