CS 583 — Deep Learning December 24, 2024

Wildfire Danger Advisories using Long Short-Term Networks

Lecturer: Zhaozhuo Xu *By: Matthew Kearney (CWID:20006175)*

I would first like to give a big thank you to Dr. Karen Short at the Missoula Fire Lab for helping me get deeper insight into the field of wildfire data and advisory operations, as well as Copernicus by the EU, the supplier of the modeled data.

1 Introduction

The state of the climate is at an all-time high. According to a study of studies from Earth science actors like NASA, NOAA, and more, global surface temperatures have reliably risen to our current climate levels of 1.2 degrees Celsius above the stable pre-industrial levels of the 1850s and before. Climatic factors such as dryness, heat, and terrain typical to the American Southeast make it highly susceptible to natural wildfires.

The Center for Climate and Energy Solutions verifies that wildfires depend on several factors, primarily temperature, soil moisture, organic matter, and other potential fuel. Humans cause 4 out of 5 wildfires, yet climate research projects a 30 percent increase in area burned by lightning-ignited wildfires in the southeastern US by 2060. NOAA estimates that the repair cost against wildfires in 2017 and 2018 alone in just California and Alaska exceeded 40 billion dollars, through the loss of property, infrastructure, and much more.

The essence of the model is to understand the conditions prone to wildfires and to deliver real-time advisories. Initially, the idea was influenced by Smokey the Bear-type metrics. I wanted to be able to deliver wildfire advisories of (yes y, hold a wildfire warning/advisory or no n, negligibly low wildfire probability/risk) in real-time. After getting lost in the pursuit of data, I ended up reaching a phone conversation with Dr. Karen Short, a research ecologist at Missoula Fire Sciences Laboratory, who told me more about indices (probabilities) used in fire advisory modeling.

”Ignition Component”, the target column I ended up finding, is a wildfire danger index that denotes the probability of a firebrand starting a fire that needs suppressive action. The model also has the capacity for the real-time generalization of wildfire advisory but has changed from a classification task to a regression task, based on data availability and research purposes.

As explained by the National Wildfire Coordinating Group, ”An IC of 100 means that every firebrand will cause an actionable fire if it contacts a receptive fuel. Likewise, an IC of zero would mean that no firebrand would cause an actionable fire under those conditions. Note the emphasis is on action.”. In terms of public safety, this model should lead to some insights into the relationship between temperature and wind when it comes to seasonal wildfire danger.

Further research and contributions about the model should be evaluated as it has a high potential to improve public safety by performing advisory leading to the lives saved and assets protected.

2 Data Collection and Processing

2.1 Data Source

By choosing observed data, I can be more certain of the quality of the data and the outcome of the model. Copernicus, the European Union's central climate agency, collects, processes, and releases climate and emergency management data consistently to their databases, meaning it would be possible to perform real-time advisory to local areas. Under a free student research license agreement, I was able to acquire 23 months (699 days) of viable data for the project, from January 1 to November 2023.

While I could not get all of the variables for the project that I had initially imagined, notably humidity and total precipitation, we have a strong suite of simple features and a well-rounded target column from which my model will learn. Future projects in environmental studies should seek quality data and deeper knowledge of the problem context to realize ideal variables for modeling this problem nature. We should also look into understanding the advantageous ML/Deep Learning paradigms to find optimal results, where model selection should seek the best models for different problems. In this problem, I decided to learn the implementation of a Long Short-Term Memory network (LSTM), which has the benefits of other RNN's while improving on 1) the vanishing gradients problem and 2) long-term dependency problems of time-series data.

After domain research on (populated) wildfire-prone areas in the nation, I decided to narrow the location for this project to Yosemite National Park, and most specifically a forested coordinate (38N, 120W) east of a small town named Deadwood, California. In the area, there are dozens of towns and other rural communities well within 10 miles, such as Cherry Valley Campground, Buchanan, Mather, Buck Meadows, and Pine Mountain Lake. The objective is to provide wildfire risk modeling for these areas and give some insight into further-reaching areas including Yosemite National Park and Yosemite Valley which are due 20 miles Southeast.

This was not only a convenient place to pull data (at discrete integer coordinates), but also centered in the forest inside of a populated hub. Based on generalization, we will be able to say that wildfire analytics produced from the model are similar in the general area. Further research should understand the factors involved in making an 'effective radius' (i.e, around 38N, 120W) that generalizes well across the area. For our purposes, these communities and the greater Yosemite Valley and National Parks are in consideration, but a precise distance can not be given at this time.

Our model target, the Ignition Component, measures the probability that a firebrand will start that needs human intervention. I was able to find this data from January 1st 2023 to November 30th 2024. This target column was found in the Copernicus Emergency Management Service (CEMS). Feature data, involving temperature and wind, is derived from Copernicus' Climate Data Store Era 5 Reanalysis from January 1st 2023 to December 15 2024. Please see the references, and 'DATA COP.py' to view these exact sources and processes. The particular feature and target data are as follows:

- Skin Temperature (Earth surface temperature)
- Soil Temperature Level 1 (Temperature of the topsoil)

- U-Component Wind 10M (Wind in the East-West 10 meters above surface)
- V-Component Wind 10M (Wind in the North-South 10 meters above surface)
- Ignition component (target, probability a fire will start that needs suppression)

2.2 Data Processing

- I used the Pandas library to load the CSV ('data.csv') as derived from the Data Collection part. Here we have climate and wildfire probability data ready to be explored. Various data manipulation, analysis, and visualization tasks were used to experiment with the model.
- We found zero examples of empty observations upon searching for null or otherwise inconsistent data, and we certify that we are ready for further exploratory analysis to understand the form of the data. Deeper visualizations and analytics are available in the file 'VISUALIZE.ipynb'
- The nature of our data, formed from real observations and calculations, should not need extra processing to remove outliers, as we should assume for model simplicity and place credibility in the operations at Copernicus, our sensors, and earth science knowledge.
- The following libraries are used in the Data Analytics and Processing: pandas, matplotlib, numpy, scikit-learn, and torch.
- Pytorch's 'LSTM' is used to develop this model.

Model Preparation

- While we have ideal data to leverage LSTM's for wildfire predictions, LSTM's are understood to perform best from normalized data. I use MinMaxScaler() by scikit-learn. This helps to not allow the model to 'overfit' onto one variable over another by scaling all data onto a level playing field.
- Feature engineering was considered in a couple of ways for this project.
- Wind components in the u and v-directions ('u10' and 'v10'), to use the distance equation and determine the magnitude of wind. Since the test data follows separating wind in the directions, it was kept this way for model development. Using this project idea in real-time advisory could benefit from divisive feature engineering such that we don't need split directions, just the total magnitude based on data availability.
- Temperature for both the soil and surface could be future engineered as well. It may be interesting to provide another 'difference' column, that is ['skt' - 'stl1'] to understand how the soil and surface compare precisely influences the model. Does it answer a line of questions like "did it rain recently?", which lead to model improvements.
- We are dealing with continuous data for all pieces of our model, and all tensors are created from dtype = torch.float32. It should be noted that this idea of using a continuous probability calculation for wildfire outbreaks has fertile ground to convert into encoded labels, which would themselves communicate the danger of wildfire.

- We could convert danger ranges from their numerical values, which people don't learn anything from, into 'fire danger rating' schemes such as ['low', 'medium', 'high', 'extreme']. Using this model with the fundamental idea as "Smokey the Bear", the ideas displayed could be generalized into real-time and provide real-time advisory with alignment from development hyperparameter tuning. For my purposes, we use these probabilities such that we can develop higher-quality metrics and learning visualizations towards the end. For our purposes, we will stick to the numerical data we have now as it leads to further precision on the computation side and can leverage better evaluation metrics and visuals in our testing.

3 Model Development

3.1 Machine Learning Model Considered

With just under two years of data, I decided to split 75 percent (the first 3/4 of the data from January 1 2023) into the training set and the last 25 percent) the past 6 (or so) months into the testing set. We have a theoretical 1.5 fire seasons in the training set, and we could test how the model has learned for predicting the rest of that season. Successful results are seen from this idea, as in our learning visualization that shows predictions trending along the true target values.

While multiple deep learning algorithms were considered for the model, including RNNs, LSTMs, GRUs, and CNN-RNN hybrids, it was realized that an LSTM would be an interesting and relevant model to develop.

3.2 Chosen Model Architecture

LSTM's are personalized for serial data, and as we will see, using a new hyperparamater of 'frames' or data windows to moderate generalizations and learning for particular tasks and data frames. This architecture allows valid information to flow to the next steps while sifting the data that does not lead to model improvement. The model development was very fast in this small data/problem size and happened in a matter of seconds. The LSTM paradigm is also known for solving two large problems in ML: vanishing gradients and long-term range dependencies, which were appealing in time for choosing a model. It was found that we were able to successfully tune the model. By the end, we are able to see low error, from metrics such as MSE, MAE, RMSE, and R-squared.

3.3 Hyperparameter Tuning

There are many components of a LSTM that we can leverage during hyperparameter tuning. These include:

- 'window size', size of a given frame or group of observations
- 'learning rate'
- 'hidden size', size of a hidden layer
- 'num layers', number of LSTM's

We are able to explore segues for model improvements through hyperparameter tuning, as is discussed in the below results section.

4 Results Analysis

4.1 Exploratory Data Analysis

Through exploratory analysis (Figures.pdf and VISUALIZATION.ipynb) we have an improved understanding of the domain insight of wildfire data. We can see some interesting results. Using 699 days of observations in Yosemite Valley, here are some of the conclusions made:

- We can see sinusoidal waves, as expected, forming over the time domain. This makes sense, temperatures rise in the summer and fall in the winter. These temperature plots, for 'Skin Temperature' and 'Soil Temperature Level 1' appear smooth and moderated.

- Reflecting on the relationship between wildfire risk and temperature, we can see that the 'ignition component' also forms a sinusoidal wave through the intense wildfire seasons. The main difference lies in the greater variance between wildfire probability over time and magnitude. This makes the plotting for the ignition component appear sharp, rigid, and chaotic.

- Based on this, wildfire risks seem to shoot up and down with some factors that are not modeled (i.e., not wind and not temperature), as wildfire probabilities vary more day-by-day compared to the temperature features, which are moderated more definitively by seasonality.

- Wind seems relatively unrelated to our ignition component target, while we can note that north-to-south winds seem to pick up in certain seasons, there is no exuberant evidence that wind helps the model greatly.

- Ignition component, our target variable, makes fire season waves around the average of '20'. When there is little-to-zero risk we see '0' and during intense fire seasons, the target reaches up to 50. This suggests that wildfire seasons are extreme when they are extreme, but never seem to fully dissipate over the year making them a threat year-long around Yosemite.

- Soil seems to both absorb and regulate heat more than the surface. The soil temperature averages 284.3K and the surface temperature at 281.4. Beyond this, surface temperature has a variance of 7.3, whereas surface temperature has a standard variance of 6.7.

4.2 Correlation Analysis Results

The correlation summary of the features, available in 'VISUALIZE.ipynb' in my dataset are interpreted as:

- The U-Component and V-Component of wind has a weak correlation with the target variable (icnfdr).

- The temperate (soil temperature and surface temperature show a positive correlation with the target variable (icnfdr).

Based on the correlation results, it can be concluded that factors including temperature have a notable influence on the risk of wildfires. Future research should be done to extend the variables involved in this analysis to optimize results.

4.3 Results of Machine Learning Modeling

4.3.1 Hyperparameter Tuning

Through hyperparameter tuning, the loss of our model decreased from 0.033 to 0.0286. In particular, 'windows' are a powerful hyperparameter related to LSTM networks that may provide a strong hypothetical basis with our wildfire model by 'playing with' the models' view of the data. Window size in an LSTM network addresses how strong of 'glasses' to put on the data when training.

Using a window size of one to three days, the LSTM network may not be able to understand seasonal trends. However, on the scale of seasons (3 months, 90 days), we lose some of the focus in the model, where it is not able to pick up on finer details based on weekly and sub-seasonal trends.

It was found that with a window size within the range of 7 to 60 [days/observations] or a time frame of one week to two months, we see similar results in the model. With a window size of 7, we reach a loss of 0.033 after 200 epochs and convergence. With a window size of 60, we reach a loss of 0.034. The trend remains similar through these evaluations (14 and 30). Based on this analysis, I chose a window size of 7 to continue tuning.

It was found that a smaller learning rate leads to better results than that of high magnitude. Initially, I used a learning rate of 0.0001, but it was found that a learning rate of 0.005 or 0.001 leads to model improvement. We also keep the number of layers as one, as this serial data and form may be a simple problem to solve with just one. I saw some model disengagement after increasing the number of layers, as we don't need to add increasing complexity to this smaller problem type.

4.3.2 Evaluation Metrics

Based on evaluation metrics on a testing set of 175 observations (about half of a year, and a half of a fire season), I found the model with a Mean Square Error (MSE) of around 0.05, Mean Absolute Error (MAE) of around 0.2, Root Mean Square Error (RMSE) of 0.2, and R-squared score of 0.4250.

4.3.3 Learning Visualization

For further understanding, it was advantageous to model the predictions from our model against the true values under our testing and evaluation set. While I was aware of what the test data set looks like ad With the understanding that we have data from June to November of 2024, I expected to see some fall in our predictions but otherwise had no preconceived notions.

As seen in Figure 2) of the accompanying file 'Figures', our model can discover the overall trends of particular seasons, while not reaching the fine precision of reality. For the cases of my interests, I was largely interested in exploring seasonal predictions. One great success of the model is that we can see the predictions plot along the true ignition component values. Although they don't reach the same magnitudes of variance, whether high or low, we can get some pretty precise insight into weekly trends from our model, and how the overall recent past conditions seem to fare through the future. Using a longer evaluation set can and should be explored down the line as needed to prove the model concept and validity in fine detail.

5 Conclusions

Notably, our model learns the seasonal and sub-seasonal trends involved in wildfire seasons. While not having the granularity to determine how every day or few days should be exactly considered, we have effectively modeled temperatures and even wind rates into the probability of wildfires requiring suppressive action through the end of the 2024 fire season. We have modeled the capacity for an LSTM network to capture greater trends of wildfire probability, suggesting a paradigm that could be helpful to further emergency advisory research like for flash floods and tornadoes.

Some variables to include down the line are humidity, precipitation, and even lightning frequency while looking into future projections of the climate. Future research to improve upon ideas displayed here should focus on acquiring sufficient features and target(s) to ascertain the more granular detail of capturing wildfire risks, as well as a greater volume of data of perhaps 3-5 years minimum, to begin capturing and validating greater seasonal trends from a climactic scope (it would also be interesting to test the model on a couple full seasons). Feature engineering and strong domain knowledge will help for improved model development leading to improved real-time advisory outcomes.

References

1. Ecmwf. ERA5 Hourly Data on Single Levels from 1940 to Present, 18 Dec. 2024, cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels?tab=overview.

2. Ecmwf. Fire Danger Indices Historical Data from the Copernicus Emergency Management Service, 18 Dec. 2024, ewds.climate.copernicus.eu/datasets/cems-fire-historical-v1?tab=download.

3. “Fire Danger: NFDRS System Inputs and Outputs.” NWCG, www.nwcg.gov/publications/pms437/fire-danger/nfdrs-system-inputs-and-outputs

4. Kearney, Matthew, and Karen Short (Dr.). “Wildfire Modeling and Data Inquiry.” 26 Nov. 2024.

5. “LSTM.” LSTM - PyTorch 2.5 Documentation, pytorch.org/docs/stable/generated/torch.nn.LSTM.html. Accessed 18 Dec. 2024.

6. Tam, Adrian. “LSTM for Time Series Prediction in Pytorch.” MachineLearningMastery.Com, 7 Apr. 2023, machinelearningmastery.com/lstm-for-time-series-prediction-in-pytorch/.

7. “What Is LSTM - Long Short Term Memory?” GeeksforGeeks, GeeksforGeeks, 10 June 2024, www.geeksforgeeks.org/deep-learning-introduction-to-long-short-term-memory/.

8. “Wildfire Climate Connection.” National Oceanic and Atmospheric Administration, www.noaa.gov/noaa-wildfire/wildfire-climate-connection. Accessed 18 Dec. 2024.

9. “Wildfires and Climate Change.” Center for Climate and Energy Solutions, 14 July 2023, www.c2es.org/content/wildfires-and-climate-change/.