A WORLD ON FIRE

CONFRONTING THE GROWING CHALLENGE OF WILDFIRE DETECTION AND SUPPRESSION

A World on Fire was compiled in December 2021 by Benjamin Feuer, Jinyang Xue, Subei Han, Dennis Pang and Yuvraj Raina

ABSTRACT

This report aims to succinctly document the existence of a novel and growing threat in the United States – the rise of uncontrolled megafires driven primarily by climate change. It lays out the reasons it is reasonable to believe that the frequency and intensity of fires is increasing, and why systemic climate change is the likely culprit for these changes. Finally, it addresses how recent advances in distributed technology can potentially lead to successful and timely interventions by professional firefighting teams.

Below is a summary of the three major points we cover --

- I. Recent evidence derived from data released by US Government agencies leads us to conclude that wildfires pose an increasing threat to the lives and property of United States citizens.
- II. This threat is driven by changes at the level of society rather than the individual, and as such, we must seek systemic, structural solutions to the problem.
- III. Distributed deep learning models for image classification as a potential method of improving wildfire response times by emergency teams and thereby reducing wildfire spread.

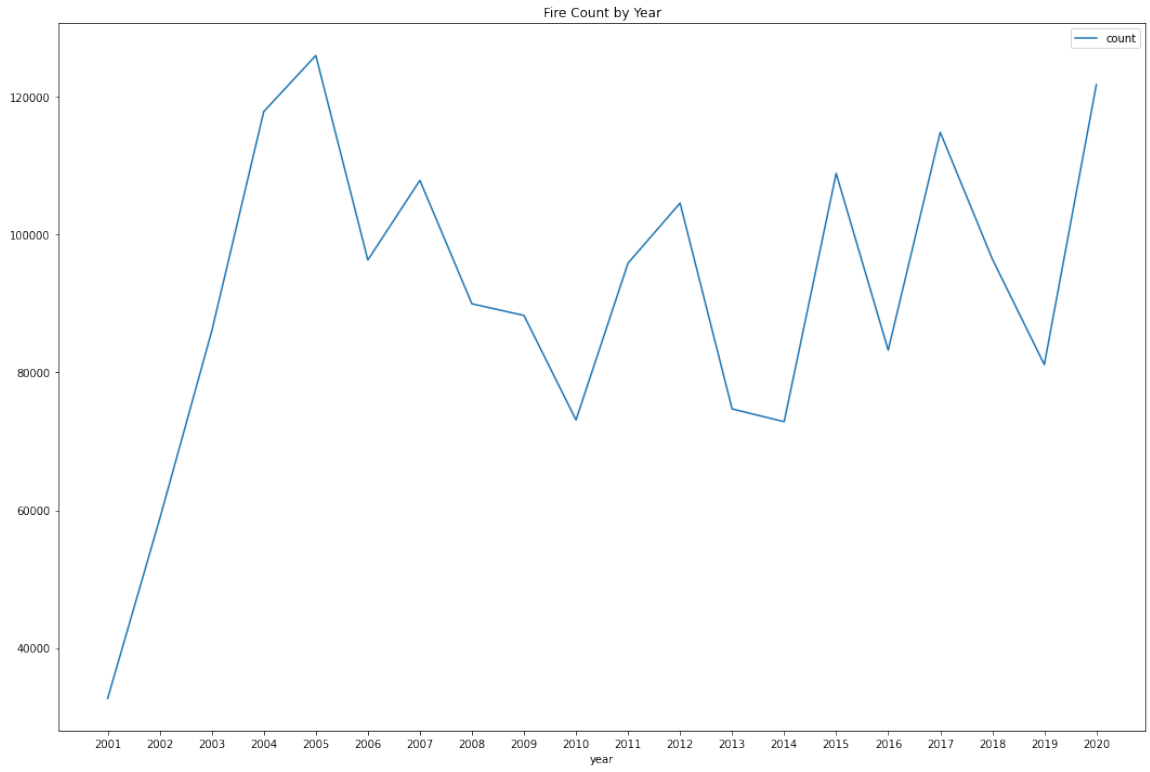
Afterwards, we offer a brief conclusion and a list of resources for further exploration.

SECTION I: Recognizing the Increasing Toll of Wildfires

In order to justify increased action on wildfire prevention, we must first address the underlying implications of our investigation.

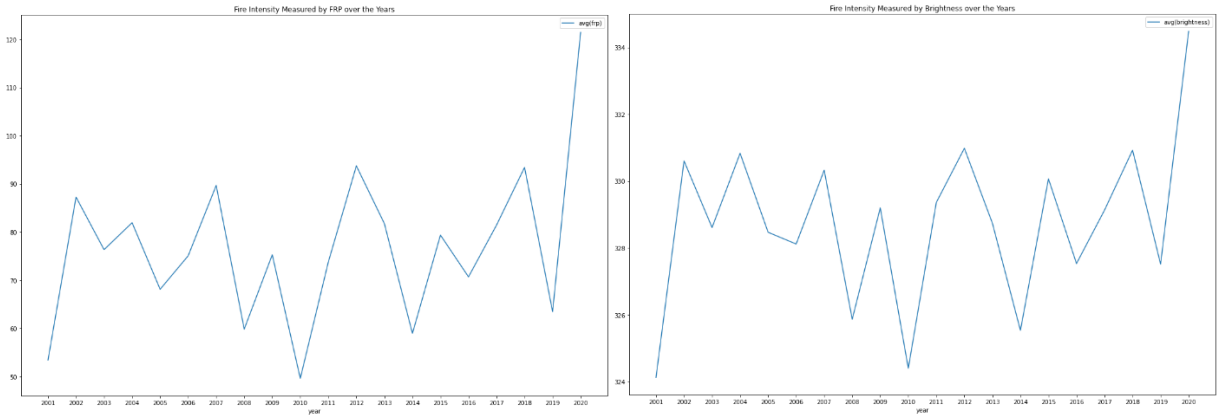
Do wildfires pose an increasing threat to the lives and property of United States citizens?

One possible measure of wildfire threat is a raw frequency count of wildfires per year, like the one we see here, aggregated from NASA's FIRMS dataset. (2021 is omitted throughout this report because for that year, the data remain incomplete as of this writing.)



By this metric, wildfire frequency in 2020 was the 2nd-highest out of the entire 20-year record, and 60% of the 5 peak years occurred between 2015 and 2020, which represents only 30% of the dataset.

The data on the intensity of these wildfires tells a similar story. We see that the average fire intensity, as measured by sub-pixel fire radiative power retrievals (**FRP**)[Csiszar et al., 2014] and apparent brightness, both reached new records in 2020.



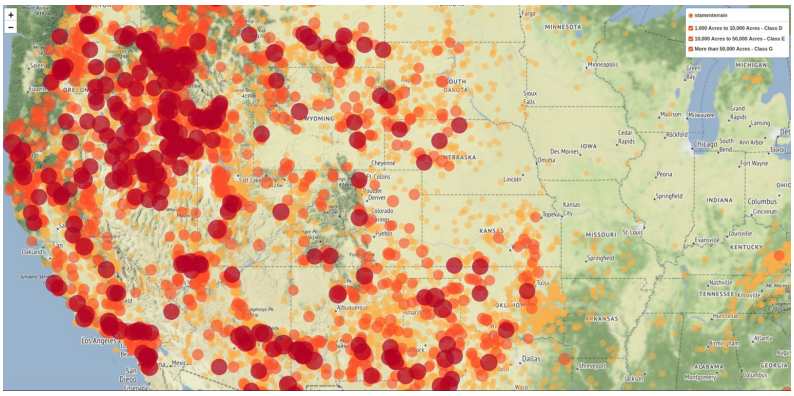
By all three of these measures, we can see that wildfires are increasing in frequency and intensity. It is logical to conclude that more frequent and more intense wildfires will cause greater damage to life, property and ecosystem and will require greater resources to combat, and indeed the evidence bears this out -- federal wildfire suppression costs in the United States have spiked from an annual average of about $425 million from 1985 to 1999 to $1.6 billion from 2000 to 2019, according to data from NIFC. In 2017 alone, damage from wildfires across the US exceeded a staggering $18 billion. Other parts of the world are being hit hard, too. This past summer, Spain suffered the worst wildfires that it's seen in 20 years, while thousands of fires burned in the Amazon Rainforest, an increase of over 80 percent compared to the same time period last year. *This vast and growing crisis demands our attention.*

SECTION II: Understanding Key Wildfire Causes

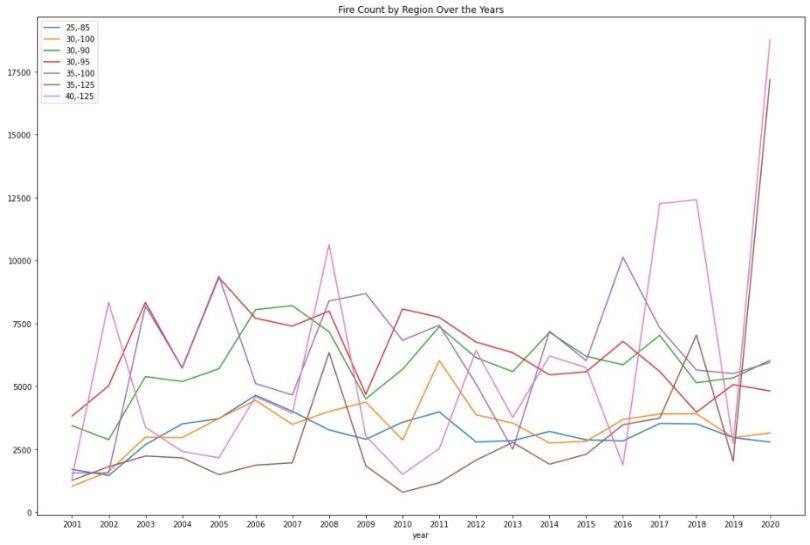
In Section I, we showed that wildfires have indeed grown more frequent and more intense. However, in order to consider how best to combat the issue, we must consider another aspect of the problem.

Can we draw any meaningful conclusions about the causes and locations of these wildfires, and can we use this information to better evaluate potential solutions?

To answer this question, we may begin by noting that this increase is not evenly distributed among all states. Historically, the drier states west of the Mississippi river have accounted for nearly all of the major wildfires by acreage burned, and this continues to be true, as this NCWG dataset shows.



Drought-ridden California, in particular, accounts for a disproportionate share of the increase in 2020, as we can see below in this FIRMS dataset graph of fire frequency by region (with regions represented as (lat, long) coordinate pairs).



Indeed, if we consider the raw increase in number of reported fires year-over-year, California consistently shows up as a hotspot.

georegion	year	count	pyCount	yearlyChange	yearlyInc
40,-125	2020	18763	2716	16047	5.908321060382916
35,-125	2020	17184	2016	15168	7.523809523809524
40,-125	2017	12254	1870	10384	5.552941176470588
45,-120	2015	12636	3557	9079	2.552431824571268
40,-125	2002	8337	1332	7005	5.259009009009009
40,-125	2008	10617	3930	6687	1.701526717557252
35,-100	2003	8184	1565	6619	4.229392971246006
30,-85	2011	12669	7603	5066	0.6663159279231882
35,-100	2014	7188	2497	4691	1.8786543852623148
35,-125	2008	6335	1960	4375	2.232142857142857
45,-120	2017	6872	2506	4366	1.742218675179569
35,-100	2016	10128	6014	4114	0.6840705021616229
40,-125	2012	6426	2527	3899	1.5429362880886426
35,-100	2008	8390	4656	3734	0.8019759450171822
35,-100	2005	9380	5740	3640	0.6341463414634146
30,-95	2005	9319	5722	3597	0.6286263544215309
30,-95	2010	8070	4682	3388	0.7236223835967536
30,-85	2017	11027	7657	3370	0.4401201514953637
30,-95	2003	8334	5020	3314	0.6601593625498008
35,-125	2018	7036	3736	3300	0.8832976445396146

only showing top 20 rows

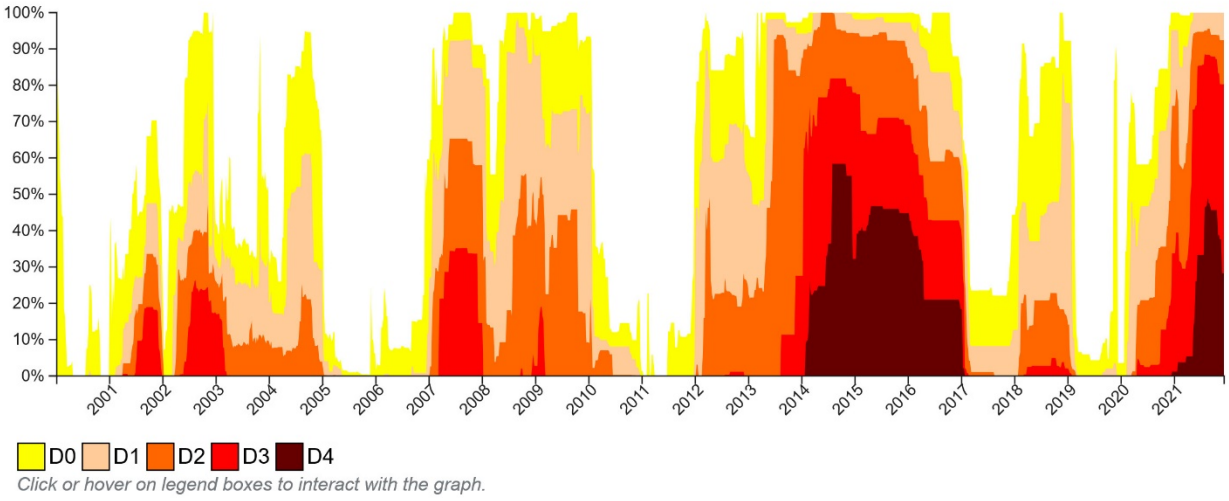
Using data from NOAA's NIDIS Drought monitor, we can see that the recent droughts in California correlate closely with peak fire years recorded in FIRMS.

2000 - Present (Weekly)

The U.S. Drought Monitor (USDM) is a national map released every Thursday, showing parts of the U.S. that are in drought. The USDM relies on drought experts to synthesize the best available data and work with local observers to interpret the information. The USDM also incorporates ground truthing and information about how drought is affecting people, via a network of more than 450 observers across the country, including state climatologists, National Weather Service staff, Extension agents, and hydrologists. Learn more.

Time Period (Years): to

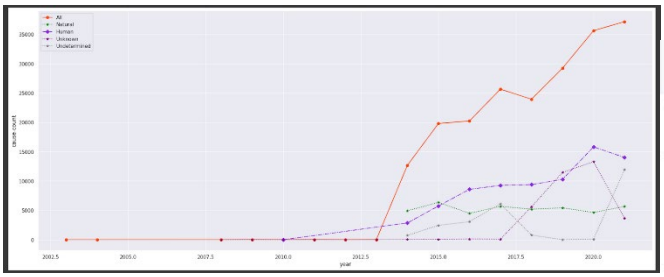
Latest Available Data:2021-12-07



These data suggest that an increase in drought conditions is strongly correlated with an increase in wildfire frequency and intensity, and that ongoing conditions leading to extreme shifts in climate may also be driving the surge in wildfires we have documented above.

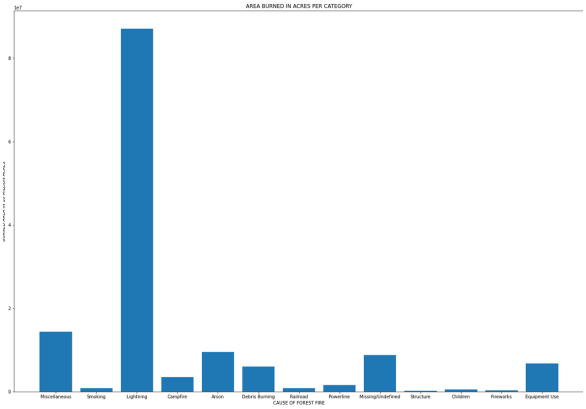
We can further confirm this by considering how regional data on wildfire causes, sourced from NIFC WFIGS Wildfire Locations data, influence the problem.

By raw count, human-initiated wildfires exceed that of natural wildfires. Furthermore, human-initiated wildfires are increasing in frequency, whereas the frequency of natural wildfires has remained flat.



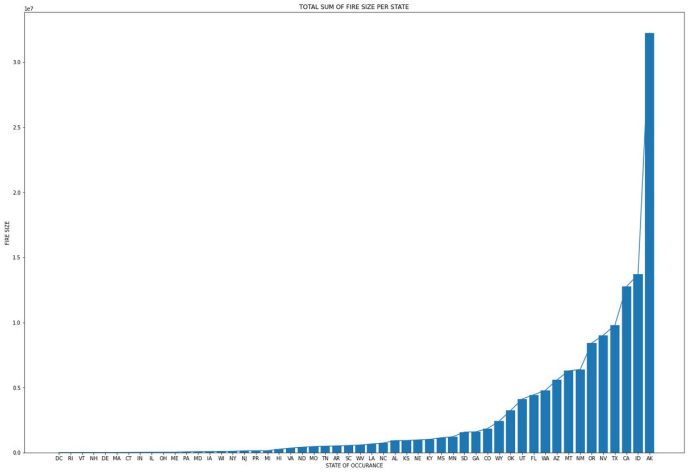
```
filterred_df.groupBy("FireCause").count().show()
✓ 1.7s
+-------------+-----+
|   FireCause|count|
+-------------+-----+
|     Unknown|60840|
|Undetermined|25199|
|       Human|76055|
|     Natural|42423|
+-------------+-----+
```

This might seem to contradict our earlier conclusions about drought being a major driver of increasing wildfire intensity. However, when we look at area burned in this [NCWG dataset](), we can see that the story is in fact quite different. Lightning alone accounts for more burned area than all other causes combined, with room to spare.



The explanation for this is well understood -- fires started by human beings tend to occur in populated areas, and are therefore more likely to be detected and contained before they grow into uncontained megafires. Fires caused by lightning (and to a lesser extent, malfunctioning electrical equipment) tend to happen in remote and inaccessible areas and are more difficult to detect and combat effectively.

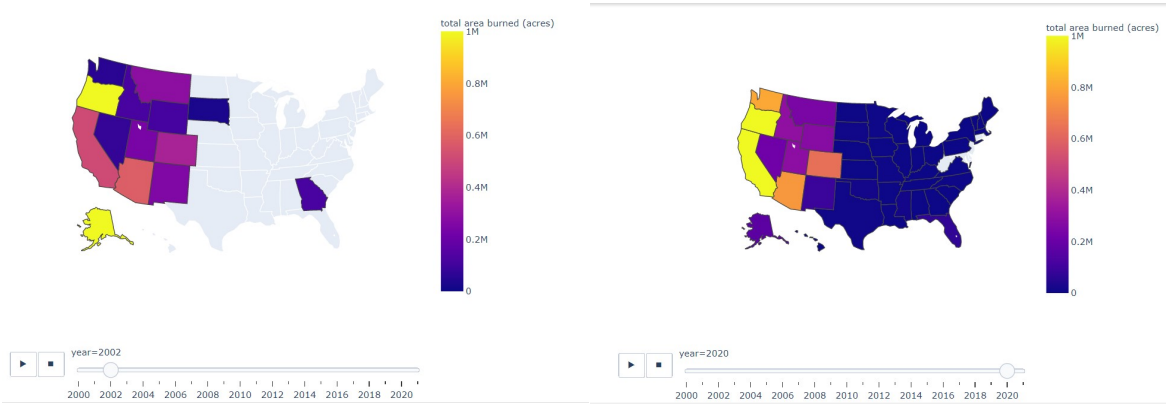
This fact accounts for the following graph. Alaska is the bottom third of states for wildfire frequency. However, in terms of acreage burned, it exceeds the bottom twenty states in the list combined. Alaska's low population and rugged terrain make it a very difficult environment for firefighters.



Although large uncontrolled forest fires have always been a historical fact of life, that does not mean that the size and frequency of wildfires we are seeing today is somehow natural. Indeed, it is likely that the side-effects of man-made climate change, such as the droughts we documented above, are driving a vicious cycle in the environment which itself contributes to further climate change. [Consider that from 1950 until 2009, forest fires in Alaska have released CO2 equal to half of all carbon emissions from the european union.]() Uncontrolled wildfire, natural or not, poses a significant climate threat when ignored.

Alaska's northerly latitude has previously protected it from large-scale wildfires – simply put, where there's ice, there isn't fire. However, ice coverage has dropped precipitously in Alaska since the 1970s, and at least 50% of this loss is caused by greenhouse gas emissions.

Furthermore, there is reason to believe that as Earth's climate continues to change, so may the regional character of fire distribution in the United States. According to NIFC data, in 2021, midwestern states have had fire counts similar to those found in West Coast states in 2014 and 2015. Even more alarmingly, as we can see in the maps below from 2002 (left) and 2020 (right), the majority of wildfire burned area is no longer a Western state problem; in 2020, all but a handful of states saw enough annual burned area to register on the chart.



Overall, the data lead us to conclude that the size and scale of lightning-caused wildfires is the major driver of the overall increase in wildfire intensity and acreage burned, which is in turn driven by changes in Earth's climate. We further conclude that the scope of the problem is broadening to include areas which were previously protected, which is likely to continue the vicious cycle of the destruction of forests, which serve as natural carbon sinks, and the corresponding release of CO_2 into the atmosphere.

The unhappy fact is that the challenges represented by these findings are predominantly the result of long-term, collective policies and incentives rather than individual choices. While it remains vitally important for vacationers to extinguish their campfires and cigarettes fully before leaving a forest, it will not and cannot make Alaskan mountain ranges easier to traverse, nor can it stop the polar ice caps from vanishing.

As such, we contend that systemic solutions are demanded. Governor Gavin Newsom of California recognized this fact with his recent dedication of funds to wildfire suppression. While national and state policy and budgeting decisions are well beyond the scope of this report, we can address a different aspect of the challenge – the role of deep learning models in early detection of wildfires.

SECTION III: Taking Action on Wildfires

Wildfires play an important role in natural systems, and their elimination is neither practical nor desirable. However, containment of megafires could achieve the best of both worlds, preventing unnatural surges while allowing healthy fires to burn under close monitoring. Unfortunately, it is impossible to contain a fire if you don't know where it is.

For this reason, it is a well-documented fact that early detection, partnered with prevention, must form the cornerstones of any successful wildfire management policy. Even small delays in detection can be very costly. Megafires can move at speeds of nearly 15 miles per hour under favorable winds.

At its peak, the Camp Fire is estimated to have burned an area greater than one football field (roughly 1.32 acres) in a single second.

Pano AI, a San Francisco startup, is one of a number of private and public ventures that are currently attempting to use deep learning algorithms to detect wildfires quickly. While Pano and U. Nevada's ALERT focus on aerial and ground-level detection, respectively, placing high-definition panoramic cameras and satellite transmitters everywhere from telephone poles to remote mountaintops, other systems, like U.C. Berkeley's FUEGO and UCSD's WiFire, have chosen to focus on detecting wildfires from satellite imagery. Still others have proposed using fleets of drones to sweep high-risk areas.

One thing that all of these solutions have in common is their reliance on state of the art deep learning models to distinguish images which contain fires from those which do not. These methods can be extremely fast and accurate, particularly when they combine image classification and semantic segmentation models with traditional algorithmic approaches. Research is ongoing in the area, but results are promising, and data on active fires are currently collected in real time from NASA and NOAA satellites, among others.

How can distributed technology help combat wildfires?

Our investigation focused on the potential of distributed systems to improve the speed and accuracy of image classifiers on fire data, since improvements in these areas could decrease the cost and improve the utility of all fire detection systems.

Our methodology was as follows;

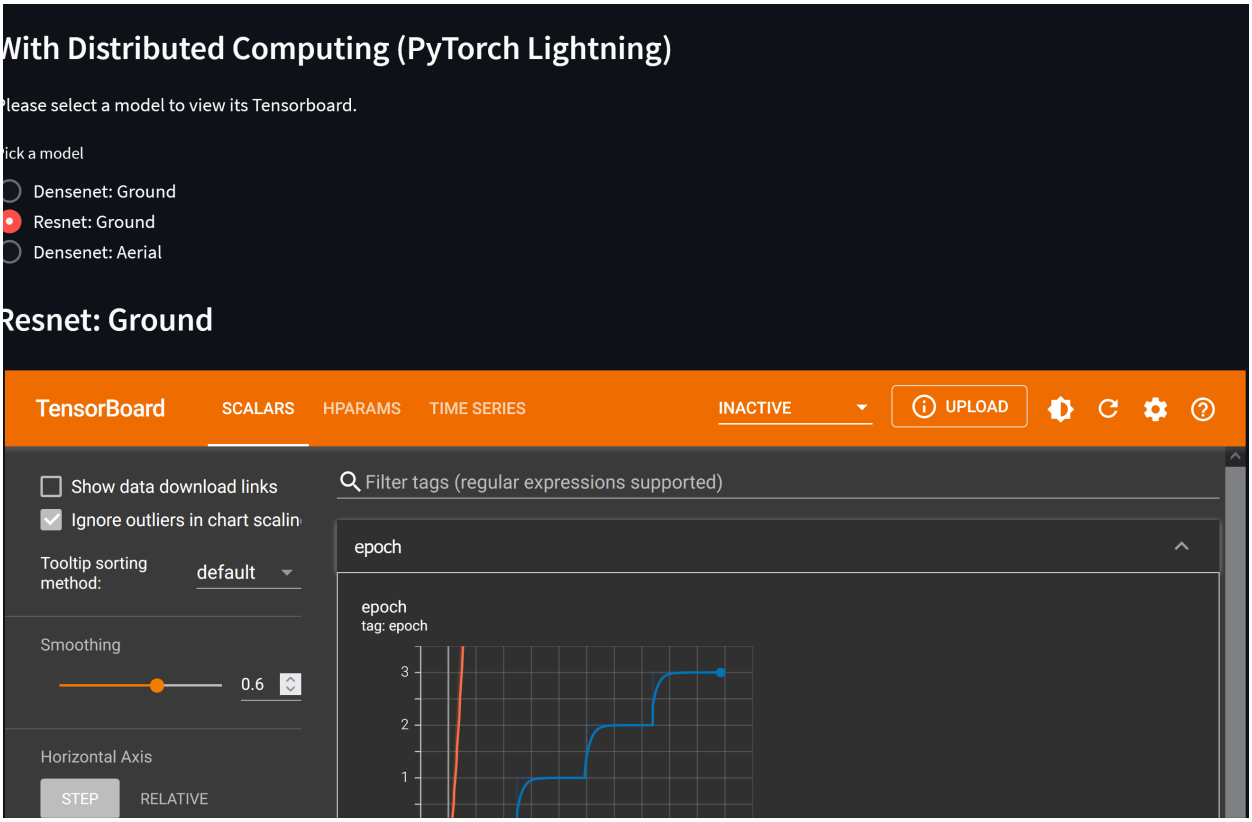
1. We collected and cleaned three image classification datasets of fire / non-fire images. The datasets were chosen to simulate a range of natural environments in which a remote sensor could be placed; at ground level, in the air, and on a satellite.
2. We designed and implemented a control group of image classifiers using convention ML and DL frameworks (XGBoost and PyTorch, respectively)
3. We re-implemented our models in distributed frameworks. Wherever possible, we replicated the hyperparameters and architecture of the model. We did, however, adapt the data ingest process to account for the unique demands of distributed systems, switching from image directories to parquet files in situations where they were supported.
4. We logged the results of our experiments to a remote database and compared them.

Our key findings are summarized below. For an in-depth account of our code, datasets, methods and experimental results, please refer to our Github repository and our interactive dashboard.

- With proper data cleaning and methodology, modern DL and ML models are capable of achieving classification accuracy greater than 90% on all three types of data, without the benefit of metadata or additional algorithmic techniques.
- On average, the traditional ML boosted tree classifier was generally less accurate than deep CNNs – however, the difference was often not large, and careful cleaning and preparation of data it was possible to eliminate much of the difference
- The boosted tree classifier was, on average, between 2x and 3x faster than the deep CNNs, on both training and inference.
- When multispectral data was available (as in the case of the Landsat-8 GeoTIFF files), training the classifier on all 10 bands was prohibitively slow and led to worse results. The choice of bands was pivotal in achieving peak accuracy, particularly the boosted tree classifiers, which are not location-invariant. Specifically, the use of SWIR data (bands 6 and 7 for Landsat-8) improved accuracy dramatically.
- Google Colab Pro is a poor environment for distributed computing experiments because of its lack of support for virtual environments, unpredictable hardware assignments and frequent timeouts. Future experiments should be conducted on a different cloud service or on the HPC.
- Dask-ml and RAPIDS are considerably less mature than PySpark, and remain rough around the edges. For instance, RAPIDS cannot be installed in a non-conda environment, and Dask and RAPIDS both lack direct support for ANY deep learning framework, although it is possible to distribute certain aspects of a PyTorch model manually using Dask's joblib (we did not attempt this). However, their potential for distributed learning is unparalleled because RAPIDS allows the offloading of massive datasets from the GPU to RAM without disrupting training – this means that the size of a RAPIDS dataset is (in theory) bounded only by the hardware it is running on.

- PyTorch Lightning includes built-in SLURM support and highlights the powerful flexibility of the underlying PyTorch architecture – switching from a fully local to a fully distributed environment in Lightning involves changing just one line of code, unlike in Dask/RAPIDS, where it requires a complete refactor.
- Even in the completely sub-optimal distributed environment (one node, one GPU, remote training) under which these experiments were conducted, distributed models were either as fast or faster than their non-distributed counterparts (20% faster, on average). We attribute this difference to more efficient data processing.
- Accuracy varied wildly between distributed and non-distributed models. We are still trying to account for the reasons for these differences – however, we suspect it has to do with the different frameworks either processing the data differently or running with different default hyperparameters.

SAMPLE: TENSORBOARD RESULTS



FUTURE WORK

Now that we have designed and tested the models, we hope to re-run our experiments on NYU's HPC cluster, where we can get more realistic results on running time and accuracy and iron out the accuracy fluctuations we detected in our initial round of experiments.

Section IV: The Human Element

Although our findings are extremely compelling, the true costs of a megafire simply cannot be reckoned in analytics alone. Sometimes, it is the little data that is truly revealing.

Consider 2021's Dixie Fire in California. The scope of this fire is difficult to imagine – at its peak, it was larger than the state of Rhode Island.

Yet, at ground level, its message is all too easy to comprehend.

Source: NY Times

This year, it was the homes, stores and services Plumas County that were razed to the ground.

Next year, it could be ours.

SECTION V: Appendix

CITATIONS

The citations for this report are included in-line via hyperlinks.

All visualizations, analytics and models generated in this report were produced with code written by the report team, unless otherwise noted, either in this report or in our Github repository notebooks, where our code can be viewed. Pre-trained model weights for certain classifiers are available on request.

An interactive summary of our key findings can be explored via our interactive dashboard.

OUR TEAM

Ben Feuer served as the project lead, designed and performed all the tests in Part III, created the interactive dashboard, created and maintained the repository, and wrote the final report.

All other team members produced supplemental project reports, which are available in the repository.

Subei Han lead the FIRMS dataset and created all of the analytics used in Part I, as well as portions of Part II.

Dennis Pang and **Jinyang Xue** co-lead the investigation into the NIFC data and created portions of the analytics used in Part II.

Yuvraj Raina lead the investigation into the NCWG dataset and created portions of the analytics used in Part II.

TECHNOLOGIES USED

All code was written and all models were trained on Google Colab Pro and written in Jupyter notebooks.

We used PySpark environment for the majority of our analytics. The techniques we used included aggregation, pivoting and window functions. When necessary, we used miniconda within Colab as an environment manager.

Our graphs and charts were produced using Matplotlib, Seaborn, Altair, and Plotly Express.

We used Fiona and RasterIO to process geospatial data and images.

Our map visualizations were generated using Folium and Plotly Express.

Our base classifier models were written in PyTorch and XGBoost.

We used PyArrow to generate parquet files for the distributed classifiers.

Our distributed classifiers were written in PyTorch Lightning, Dask-ml and RAPIDS.

MongoDB Atlas and Tensorboard were used to log the results of the classifiers.

Our interactive dashboard was created using streamlit.

LICENSE