Climate Change Prediction-to-Action Pipeline

Aarushi Dubey
University of Maryland [XXX]
College Park, MD

Abstract

Climate change is having current and future effects on our environment and fellow species. Material change through policy is needed to reduce the impacts we are already feeling. In order to ease the learning curve policy-makers and the public may face in implementing the correct actions to combat climate change, I have created a design/implementation of a pipeline of deep learning applications for "climate change prediction to action." Specifically, the pipeline will (1) simulate climate change impacts and then (2) provide policy recommendations to improve these future impacts. This pipeline can help provide research-based actionable items on climate change remedies for audiences not as climate-acclimated. Utilizes RNNs and GPT-based reccomender systems to do so.

- - - - - -

1 The Introduction

Climate change is a slow-burn ecological collapse that is currently impacting and developing society WHO [2021]. We can see the impacts of climate change right now, with various geographic areas going through changes such as sea ice reduction, carbon dioxide increase WHO [2021]. The food supply, available water, health conditions, and natural + man-made environments for humans and other species are being detrimentally affected WHO [2021]. Due to the irreversability of previous actions done by humans such as industrial development, we cannot put a full stop onto the ongoing impacts of climate change and by proxy global warming NASA. However, we can try to reduce the severity of the impacts somehow in the long run. One way to improve climate change outcomes is through making policy changes that can create material effects on society.

Climate change is a topic that seems very broad and daunting to many, but actually is a field where a plenitude of research is conducted on and many recommendations are given for Klingelhöfer et al. [2020]. Rather, the issues is that information on climate change and its remedies may not be the most easily palatable or understandable for the general public who is not specialized in meteorology. I think that using deep learning to aid in this field can help take advantage of the immense efforts being put in by climate scientists to increase awareness on climate change impacts for the future, and can also help in communicating the risks of not making accurate / beneficial policy changes to both the government and the general public in a way that is palatable for users.

2 The Approach

For my final project, I created a theoretical pipeline of deep learning applications that can be used to provide aid in simulating the impacts of doing policy (non)choices regarding climate change for a hypothetical region of the world, and then use those generated hypotheses alongside external health data to get recommendations on what the 'next best steps' would be. The pipeline's implementation consists of the following two parts:

Preprint. Under review.

2.1 Hypothesizing the Impacts of Climate Change

2.1.1 Data

The first part of the pipeline is a predictive deep learning model that would use data from the following dataset: World Bank Climate Change Dataset WorldBank [2017]

This dataset consists of attributes such as "annual precipitation change," "annual temperature change," "annual cool days/cold nights," "annual hot days/warm nights" - which can be classified as factual attributes as these can be influenced by social and political changes, but cannot be controlled specifically. There are also attributes such as "child malnutrition, underweight (percent of under age 5)", "access to electricity (percent of total population)", "number of physicians per 100 people", "nurses and midwives (per 1,000 people)", "investment in transport, telecoms, energy, water w/ private participation ($)", "CO2 emissions per units of GDP (kg/$1,000 of 2005 PPP $)", "geographic region" and more, which we can refer to as political attributes as these attributes can be changed directly by policy and social changes. The dataset has groups of factual and political attributes for 200+ countries/geographic regions varying over 10 to 15 years each. This dataset is available for public use on dataworld.com, provided by World Bank.

Political Factors Used	
AG.YLD.CREL.KG	Cereal yield (kg per hectare)
BX.KLT.DINV.WD.GD.ZS	Foreign direct investment, net inflows (% of GDP)
EG.USE.COMM.GD.PP.KD	Energy use per units of GDP (kg oil eq./$1,000 of 2005 PPP $)
EG.USE.PCAP.KG.OE	Energy use per capita (kilograms of oil equivalent)
EN.URB.MCTY.TL.ZS	Population in urban agglomerations >1million (%)
ER.LND.PTLD.ZS	Nationally terrestrial protected areas (% of total land area)
NY.GNP.PCAP.CD	GNI per capita (Atlas $)
SE.ENR.PRSC.FM.ZS	Ratio of girls to boys in primary & secondary school (%)
SE.PRM.CMPT.ZS	Primary completion rate, total (% of relevant age group)
SH.DYN.MORT	Under-five mortality rate (per 1,000)
SH.MED.PHYS.ZS	Physicians (per 1,000 people)
SP.POP.GROW	Population growth (annual %)
SP.URB.GROW	Urban population growth (annual %)

Factual Factors Used	
EN.ATM.CO2E.KT	CO2 emissions, total (KtCO2)
EN.ATM.CO2E.PC	CO2 emissions per capita (metric tons)
EN.ATM.CO2E.PP.GD.KD	CO2 emissions per units of GDP (kg/$1,000 of 2005 PPP $)
EN.CLC.GHGR.MT.CE	GHG net emissions/removals by LUCF (MtCO2e)
EN.CLC.PCAT.C	Projected annual temperature change (2045-2065, Celsius)
EN.CLC.PCCC	Projected change in annual cool days/cold nights
EN.CLC.PCHW	Projected change in annual hot days/warm nights
EN.CLC.PCPT.MM	Projected annual precipitation change (2045-2065, mm)

2.1.2 Implementation Details

The model curated for the first part of the pipeline intakes sets of political attributes, and uses them to predict what factual attributes would result for 1 year, 5 years, 10 years from now. For this part of the pipeline, I implemented a Recurrent Neural Network (RNN). I chose to implement a recurrent neural network because RNN's are best suited for sequential or time-dependent data, and the data provided by the dataset consisted of various political & factual attributes being measured for various geographic areas as yearly measurements from 1990 to 2011. Specifically, I implemented a LSTM RNN - a Long Short-Term Memory Network due to the advantage that LSTM's have in holding onto previous outputs for a longer time in memory. However, before implementing the LSTM, I needed to manipulate the dataset to create a dataloader that pulled out a sequence of previous features to learn from, not just a single previous value. After developing my SequentialDataset, I developed and trained an LSTM from scratch on the dataset I developed using PyTorch with three linear layers, which also utilized a MSE Loss function and an Adam Optimizer.

2.2 Recommending Based on Resulting Hypotheses

The second part of the pipeline intakes the resulting factual attribute situations from part 1 of the pipeline, joins health statistics for the given geographic region of the world to the previous output, and then provides recommendations on what policies should be centered on to mitigate health issues that may impact the particular population affected by such climate effects - all based on an outlook on what the health statistics of the affected population (described by the output of part 1 of the pipeline) will look like in the future.

2.2.1 Data & Implementation Details

To build the the first part of the recommender system that will use an outlook on health statistics for the given population, I needed to aggregate myself a dataset that has both health statistics for populations around the world, and has climate statistics for the same populations. Formally, the dataset needed to consist of environmental public health indicators, which are defined by the US Centers for Disease Control and Protection as indicators that "provide information about a population's health status with respect to environmental factors." CDC I compiled the following datasets together:

- information on air quality measures in the United States of America regarding ozone concentration over various regions within the US: Air Quality Measures
- toxic release inventory data: Toxic Release Inventory
- chronic disease indicators data Chronic Disease Indicator Data

The reccomender system would then, ideally, do a sentiment analysis first on a snapshot of some chunk of the internet / a dataset that holds existing policies, formally proposed policies, online word-of-mouth proposals, and scientific papers regarding what to do when it comes to dealing with climate change *for a population with the factual attributes and potential health statistics given*.

This second part of the recommender system needs its own dataset that consists of labeled policy documents that have some type of labels according to the following coding scheme: (a) what goals the policy targets and (b) what actions are proposed to solve such goals. To do this, I downloaded 50 policy documents from the US Congress website (congress.gov) that were under the 'Environmental Protection' policy subject area. I then began to apply a bag-of-words computational approach onto each policy to pull out the top 5 most common words from each document. Then, I began to utilize Princeton's WordNet, which is a "a lexical database of over 100,000 word senses with definitions and hierarchical relationships" Ahn [2017] and Stanford's CoreNLP toolkit to pull out the hypernyms for each of the top 5 words. Hypernyms provide overarching parent terms of words, which allowed me to get a very generalized label for each document that gave some blended sense of what the policy goals and policy actions are.

After developing the fully labeled policy dataset utilizing WordNet, I would then need to develop a ChatGPT-like model or a collaborate filtering recommender system - which would take in the predicted climate statistics output from part 1 of the pipeline augmented with environmental health statistics of the area - and then use the labeled policy dataset to advise what actions should be taken by providing linkage to the best policy legislation available from the dataset that addresses the concern.

3 The Experiments & Results

The final LSTM network I developed utilized a learning rate of 0.00003, 16 hidden units, a batch size of 8, a sequence size of 5, forecast lead of 4 years, and 300 epochs. By the end of the training of the model, the LSTM had a training loss of change from 1.06 to 0.56, and a validation loss change from 7.67 to 5.89. In order to finalize on the hyperparameters, I did a grid search on combinations of hidden units sizes in the options of [8, 16, 24], batch sizes of [5, 6, 7, 8], and sequence sizes of [3, 4, 5]. An example of the model forecast on one of the outputted factual attributes (GHG net emissions/removals by LUCF (MtCO2e)) can be seen below:

year	EN.CLC.GHGR.MT.CE_lead4	Model forecast
1990-01-01	-850.050344	-784.814453
1991-01-01	-795.375724	-788.264771
1992-01-01	-679.822703	-796.090027
1993-01-01	-727.705738	-795.813599
1994-01-01	-601.993296	-795.782471
1995-01-01	-503.702132	-805.790222
1996-01-01	-540.259927	-815.408325
1997-01-01	-635.810488	-824.282349
1998-01-01	-830.615927	-842.290283
1999-01-01	-980.170627	-852.791260
2000-01-01	-1033.640402	-866.119141
2001-01-01	-1027.881749	-880.424072
2002-01-01	-1014.528899	-889.278748
2003-01-01	-1013.383977	-894.310974
2003-01-01	-1013.383977	-894.387390
2004-01-01	-1007.251509	-903.211548
2005-01-01	-990.061776	-910.844788
2006-01-01	0.000000	-911.607361
2007-01-01	0.000000	-910.797729

4 The Related Work

The following paper proposes a reccomender dataset which includes 'greenness' on decisions given and recommended to encourage more sustainable choices in people. Nudging Towards Sustainable Choices via Recommender Systems Kalisvaart [2023]

The following paper compares NLP approaches to labeling policies using NLP models to generate governmental datasets for political research, which was extremely insightful in helping me create my own environmental policy datasets for the reccomender system. Comparing NLP Methods for Identifying Policy Decisions in Government Documents Ahn [2017]

The following paper does a systematic review of current deep learning and machine learning models applied in the climate change field: Deep Learning and Machine Learning in Hydrological Processes Climate Change and Earth Systems a Systematic Review. Ardabili et al. [2020]

5 The Conclusion & Discussion

An issue I ran into early on was that the majority of easily available public datasets that I found that pertained to environmental health focused on the US, so I had to narrow down my entire pipeline to be specifically about simulating and recommending on climate change impacts in the USA, rather than globally. Therefore, for part one, I had to only choose USA specific climate change statistics, which reduced my original 13000+ data items to about 25 data items to train and validate my model on. Unfortunately, this extremely small dataset caused my model to overfit on my training dataset a bit and not do as well on my validation data, as indicated by the respective losses mentioned above. For the second part of the pipeline, because I had to manually create multiple datasets through various NLP methods, the majority of my time was dedicated to data creation and augmentation. Therefore, I wasn't able to physically develop the recommender system / GPT model, but I did try to plan out the theoretical frameworks for them before going into the dataset creations.

6 The Future Work

In the future, I would like to find more global datasets on environmental indicators so that I can utilize a larger amount of the World Bank dataset for part one of the pipeline (predicting climate change impacts for a geographical region). I would also like to add GPS coordinates for various regions on the dataset, so that way the model can estimate climate change impacts not just for countries already in the database, but also for geographical areas that may not have a label but can be identified through coordinates. For part two of the pipeline, I would like to be able to finish curating the datasets for the labelled policies with not just one overarching hypernym label, but with two labels identifying the policy action and the policy goal separately. For that, I would most likely need to replace my bag of words approach for feature extraction to one more structured. Then, I would like to implement the recommender system using the datasets I curated. Kordik [2018]

References

WHO. Climate change and health, 2021. URL https://www.who.int/news-room/fact-sheets/detail/climate-change-and-health.

NASA. Is it too late to prevent climate change? URL https://climate.nasa.gov/faq/16/is-it-too-late-to-prevent-climate-change/#:~:text=While%20the%20effects%20of%20human,otherwise%20persist%20for%20essentially%20forever.

Doris Klingelhöfer, Ruth Müller, Markus Braun, Dörthe Brüggmann, and David A. Groneberg. Climate change: Does international research fulfill global demands and necessities? *Environmental Sciences Europe*, 32(1), Oct 2020. doi: https://doi.org/10.1186/s12302-020-00419-1. URL https://enveurope.springeropen.com/articles/10.1186/s12302-020-00419-1.

WorldBank. Climate change data, 2017. URL https://data.world/worldbank/climate-change-data.

CDC. Environmental public health indicator. URL https://ephtracking.cdc.gov/indicatorPages?selectedContentAreaAbbreviation=22selectedIndicatorId=146selectedMeasureId=.

Natalie Ahn. Comparing nlp methods for identifying policy decisions in government documents*, 2017. URL https://natalieahn.github.io/Ahn_PINet.pdf.

Raoul Kalisvaart. Nudging towards sustainable choices via recommender systems, 2023. URL https://repository.tudelft.nl/islandora/object/uuid:dede59cc-5cd8-4420-b6ba-f6002f965b7c.

Sina Ardabili, Amir Mosavi, Majid Dehghani, and Annamária R. Várkonyi-Kóczy. Deep learning and machine learning in hydrological processes climate change and earth systems a systematic review. In Annamária R. Várkonyi-Kóczy, editor, *Engineering for Sustainable Future*, pages 52–62, Cham, 2020. Springer International Publishing. ISBN 978-3-030-36841-8.

Pavel Kordik. Machine learning for recommender systems — part 1 (algorithms, evaluation and cold start), 2018. URL https://medium.com/recombee-blog/machine-learning-for-recommender-systems-part-1-algorithms-evaluation-and-cold-start-6f696683