Maria Alder
Climate.UAR
5/2/2024

Unified machine learning tasks and datasets for climate change mitigation and adaptation

1. Introduction

Climate change as caused by the anthropogenic emission of greenhouse gasses is disrupting the equilibrium of the planet. This has prompted the urgent need to develop mitigation and adaptation strategies that reduce greenhouse gas emissions and prepare us for a changing world. Artificial Intelligence (AI) and Machine Learning (ML) are powerful computational tools that can enhance technologies and shape policies related to climate change mitigation and adaptation.

With regard to mitigation strategies, we will need to make sweeping changes to electricity systems, transportation, buildings, and land use. These changes require making existing systems more efficient in order to reduce energy consumption and emissions. AI and ML models can leverage the vast amounts of data produced by these systems to make their operation more efficient. For adaptation, building resilience and managing disasters will be key. AI and ML are essential tools for forecasting extreme weather events and identifying areas that may be vulnerable to disasters. Furthermore, these computational methods can improve weather and climate models which are used to understand long-term atmospheric changes.

ML tasks that are related to climate change typically deal with spatio-temporal data and systems bound by physical laws which can be challenging to represent with present-day ML models. These challenges stem from the need for large datasets so that the models can learn physical relationships. Oftentimes, datasets are not large enough or are too specific to a certain domain that it doesn't allow the model to generalize and adequately derive those relationships. This prompts the need to better understand and quantify specific characteristics of datasets that contribute to the underperformance of ML models for tasks related to climate change.

Currently, we do not understand the properties that are prevalent in climate change-related tasks and datasets. For instance, characteristics like data imbalance and distribution can affect the ability of a model trained on that data to generalize. Due to the spatio-temporal nature of datasets in the climate change domain, certain characteristics may be more prevalent. Furthermore, there exists a gap in our understanding of what attributes of a task and its associated data leads to poor performance for particular types of ML models. We are particularly interested in large-scale multi-tasking ML models which are models that can be applied to tasks in multiple domains while still maintaining performance or even outperforming more specialized models [1]. We hypothesize that multi-tasking models perform similarly on tasks with datasets that share common characteristics even if they are from different domains.

As such, we attempt to bridge this gap by performing a comprehensive analysis of tasks and datasets related to climate change. This can facilitate the design of new ML solutions in climate change domains, and potentially lead to advancements in research on more general ML methods [2, 3]. Furthermore, it can inform the development and improvement of useful datasets by highlighting what characteristics they might lack.

Previous work has explored potential applications of ML and AI for climate change-related tasks. Rolnick et al. created a comprehensive overview of ML applications for tackling climate change including a significant number of applications associated with enhancing renewable energy [4]. Mosavi et al. and Donti et al., review numerous applications of ML for facilitating the development and operation of sustainable energy [5, 6]. Nguyen et al. developed the first foundation model for ML tasks in weather and climate modeling [7], suggesting the potential for a multi-tasking type of model within this domain. Koh et al. introduce a collection of 10 datasets, some of which are associated with climate change-related tasks, and assess how the distribution of the data can affect default model performance [8].

This paper begins by introducing 2 climate change-related tasks. These tasks are an extension of work previously done on this project by Aryandoust et al. [9] which introduced and evaluated 6 tasks associated with 17 total datasets. In order to evaluate these tasks and datasets, we present a method for converting the datasets into a unified format. This conversion to a single representation allows for simpler comparison between data of different modalities and for a smaller set of multi-tasking ML models to be used on the data. Then, using this unified format, various scores are calculated for each dataset. We then compare the scores between datasets to determine similarities and differences between the tasks that they are used for and potential ML model types that are suitable for tasks with particular scores. We plan to release a public repository with the datasets and code that performs the calculation of the scores.

2. Tasks and Datasets

Previous work on this project by Aryandoust et al. [9] has evaluated 6 different tasks consisting of a total of 17 datasets. These tasks include predicting the electric load profile of buildings, active power generation of wind farms, average travel time of a car for a certain path, atmospheric radiative transfer between layers of the atmosphere, and structure of catalyst-adsorbate pairs for hydrogen electrolysis and fuel cells, as well as the annotation of policy directives and regulations. Building upon these tasks, this work will evaluate 2 additional tasks and datasets including capacity expansion planning for power grids and solar panel identification from aerial imagery. We add these datasets in order to cover additional machine learning paradigms that previous work has not explored. In particular, capacity expansion planning involves estimating a solution to an optimization problem while solar panel identification is image classification.

Capacity Expansion Planning. Given the hourly fuel mix, hourly interface flow, hourly real time price, hourly generation for thermal generators larger than 25W, daily nuclear capacity factor, and monthly hydro generation data for the New York State grid, we want to determine viable types of generators to add to the grid. As demand for renewable energy grows, it will be crucial to understand the optimal locations to place new generators so that they meet the demand of consumers while keeping the grid balanced.

Solar Panel Identification. Given an aerial image of a PV installation, we want to create segmentation masks that highlight the solar panels. As small-scale PV installations become more common, the identification and mapping of these installations will be valuable for power system operators responsible for balancing the grid and who currently have limited knowledge of their distribution and generation. Previously, Wang et al. [10] developed a model that accurately identifies solar panels and drew conclusions from the model about the adoption of solar in low-income communities. The dataset consists of 13000 installations with ground truth segmentation masks which we refer to as *solar-13000*.

3. Unified Representation

In order to create a single representation for the data, we leverage the fact that each dataset consists of data points that vary across time, space, or both. This allows us to create a unified spatio-temporal representation. To do so, we divide each datapoint's features into those that are time-variant $\mathbf{x_t}$ meaning that they are constant across space and only change across time, space-variant $\mathbf{x_s}$ meaning that they are constant across time and only change across space, and space-time-variant $\mathbf{x_{s,t}}$ meaning that they change across both time and space. By splitting the features like this, we are able to make comparisons between equivalent categories between datasets. This is useful for climate change-related tasks which are heavily dependent upon spatio-temporal data specifically bounded by physical laws.

4. Dataset Scores

After converting the datasets into the unified representation, four scores will be calculated for each dataset: sample imbalances (SImb-score), spatio-temporal out of distribution (STood-score), input-output (IO-score), and outlier.

SImb-score. The sample imbalances score quantifies selection bias by measuring the size of data imbalances and sample biases. It is calculated using the average of the JensenShannon divergence (JSD) between the distribution of each feature and the uniform distribution.

STood-score. The spatio-temporal out of distribution score quantifies the distribution shift of data points for the unified data format across time and space. This provides an estimate of the epistemic uncertainty associated with the dataset. It is calculated using the average JSD between the distribution of features in our training data and our validation or testing data.

IO-score. The input-output score quantifies the sensitivity of the labels with respect to changes in single features. This provides an estimate of Aleatoric uncertainty. It is calculated by taking the mean incremental ratio between all feature-label paris.

Outlier-score. The outlier score identifies the presence of subgroups and edge cases in the data which can help quantify evaluation bias in the dataset. It is calculated using the distribution of outlier values determined by the Tukey’s fences method.

7. References

[1] Sanh, V. & et al. Multitask prompted training enables zero-shot task generalization. ICLR (2021).

[2] Donti, P. L. Bridging Deep Learning and Electric Power Systems. Carnegie Mellon University (2022).

[3] Aryandoust, A. Artificial Intelligence for the renewable energy transition. ETH Zurich (2023).

[4] Rolnick, D. & et al. Tackling Climate Change with Machine Learning. ACM Comput. Surv. 55 (2), 42 (2022).

[5] Mosavi, A. & et al. Machine Learning for Sustainable Energy Systems. Energies 12 (7), 1301 (2019).

[6] Donti, P. L. & Kolter, J. Z. Machine Learning for Sustainable Energy Systems. Annual Review of Environment and Resources 46, 719-747 (2021).

[7] Nguyen, T., Brandstetter, J. Kapoor, A., Gupta, J. K. & Grover, A. AutoML for Climate Change: A Call to Action. Preprint at https://arxiv.org/abs/2301.10343 (2023).

[8] Koh, P. W. & et al. WILDS: A Benchmark of in-the-Wild Distribution Shifts. In 38th Intern. Conf. Mach. Learn. (2021).

[9] Aryandoust, A., Rigoni, T., Di Stefano, F. & Patt, A. Unified machine learning tasks and datasets for enhancing renewable energy. arXiv preprint arXiv:2311.06876 (2023).

[10] Wang, Z. & et al. DeepSolar++: Understanding residential solar adoption trajectories with computer vision and technology diffusion models, Joule, Volume 6, Issue 11, (2022).