MLD Advance Project Phase 1

Topic: Weather Classification using ML

By: 1. Vishwashree V Karhadkar [Student ID: 202307962]

2. Gagandeep Singh [Student ID: 202303876]

OVERVIEW

Kaggle: https://www.kaggle.com/datasets/jehanbhathena/weather-dataset

We have fetched a weather-related dataset consisting of weather images into various folders. The folders consist of a few values like dew, rain, rainbow frost, lighting, snow, etc. It is a compatible dataset for supervised learning. All the images are sorted in a particular order, so each folder has a particular weather condition.

The dataset is fetched from Kaggle but its main source comes from the Harvard dataset[1]. The link to Kaggle is given above, whereas the actual Harvard source link with its description is given in the References section. There is a research article published using the same dataset in the Advanced Earth and Space Science publication which is explained in the Publication section[2].

Data set variables

The Dataset consists of a total of **6,862** images of various weather conditions collectively added in **11** different folders which are named according to weather conditions and have those specific weather images in them. All images in this dataset are in **JPG** format type. All images are in **colored** (RGB) format. The size of images ranges from 3-5 kb to 3-5 MB. The total size of the dataset is **636.73 MB**. These provide a diverse set of visual data to analyze helping us to train our ML models and help predict weather which is the primary objective of our advanced project.

The Listed weather conditions folder names with the count of images within it are as follows:

Dew: 698 files
Fog/Smog: 851 files
Frost: 475 files
Glaze: 639 files
Hail: 591 files
Lightning: 377 files

Rain: 526 files
Rainbow: 232 files
Rime: 1,160 files
Sandstorm: 692 files
Snow: 621 files

As far as the tabular data is concerned, our dataset doesn't contain any tabular data information associated with it.

Goal and Questions to Predict

Our primary objective and hypothesis in this advanced project would be to predict and classify different weather conditions based on the image by utilizing feature sample images available in the dataset and using our open-source software df-analyze for the prediction. We will be doing **multiclass classification** for this project. Also, We hypothesize that by using our open source software, running it multiple times with feeding different images, re-running it, or maybe with different backgrounds and/or with interfering objects; we could find an interesting comparison with the prediction accuracy percentage given by the df-analyze and comparing those results with the published paper prediction analysis to understand more and assist with the development of a more accurate ML Program and technology for the dataset using our application.

References and Publications

[1] Dataset Actual Source: Xiao, Haixia. 2021. Weather phenomenon database (WEAPD), Version 1.0. Harvard Dataverse, V1. doi:10.7910/DVN/M8JQCR. URL-[https://doi.org/10.7910/DVN/M8JQCR]

The Actual source of this dataset has its name as “Weather Phenomenon Database (WEAPD)”, which was created by Haixia Xiao in 2021 and is designed to help classify various weather Phenomena. The keywords as they described are Earth and Environmental Sciences. The dataset is officially published on the Harvard Dataverse and can be accessed using the link given above. It has been given to use in the public domain and is seen on the dataset page and on the website [1].

[2]Xiao, Haixia, Zhang, Feng, Shen, Zhongping, Wu, Kun, and Zhang, Jinglin. 2021. *Classification of Weather Phenomenons from Images by Using Deep Convolutional Neural Network.* Earth and Space Science. Published 07 April 2021. doi:10.1029/2020EA001604.URL-[https://doi.org/10.1029/2020EA001604]

The above-published paper has an approach to classify weather phenomena classification using a deep neural network (CNN) which is called METeCNN. The authors have used the exact same dataset from Harvard which we are using for our advanced level project. In their findings, MeteCNN has achieved a classification accuracy of **92.68%** which is outstanding compared to traditional models and various mainstream models available such as VGG16, ResNet34, and EfficientNet-B7. The model has the ability to succeed because of its ability to effectively use and analyze lean training on weather-related data and features and avoid the errors that are generally seen in human observations. The paper also states that the model has also seen challenges within a few of the weather types which is due to image stability and complexity of the dataset and few feature images. The paper suggests a future scope where, as they have seen interference objects and having complex backgrounds in images affects the accuracy of the model, they suggest working on the interference objects and the backgrounds present on the images of the dataset to achieve more accuracy[2].