ON THE CLOUD DETECTION FROM BACKSCATTERED IMAGES GENERATED FROM A LIDAR-BASED CEILOMETER: CURRENT STATE AND OPPORTUNITIES

Supplementary Material

Data was collected with a Lufft CHM 15k ceilometer that leverages Light Detection and Ranging (LiDAR) technology. This system emits short light pulses from a solid-state laser microchip into the atmosphere, where they interact with aerosols, droplets, and other air molecules in the atmosphere. The reflected portion of light, called backscatter, contains valuable information that are processed by the device. The time-of-flight of laser pulses is measured to calculate the distance of the scattering event, and the height profile of the reflected signal is analysed to derive the backscatter intensity β-raw. The attenuated backscatter coefficient β-att is then calculated using a calibration constant.

From this data, various useful parameters can be determined, such as the height of clouds and aerosol layers. The detection system is based on a photon counting process.
This measuring device is located near San Giovanni La Punta (Catania, Italy), precisely at the coordinates [37° 34′ 43.997" N, 15° 6′ 11.002" E]. From the ceilometer data, images expressing the instrument's detection in a spatial and temporal sense were constructed (backscatter profiles). Specifically, we plot the time on the x-axis and the height of the measured particles (contained in the backscatter coefficient) on the y-axis. The colour in the backscatter profiles indicates particle intensity, with intense blue signifying an absence of particulates and red indicating a high concentration. Numerically, the scale ranges from 0 to $5 \cdot 10^{-6}$.

We generated backscatter profiles for each day of observation and we further divided them for every hour, resulting in 24 profiles per day. These profiles were then labelled using a *Weather Research and Forecasting* (WRF) model. For each pressure level, the presence or absence of clouds was determined from the output of the WRF model, providing precise and reliable ground-truth labels for each backscatter profile.

The dataset is organised in two folders: train and test set, respectively. Each of these folders contain positive (True) and Negative (False) samples, as summarised below:

- Train
    - True
    - False
- Test
    - True
    - False

We generate a total of 1,568 images of size 150 × 1,000, and we used 70% as training set and 30% as test set. In our benchmark, the training set was subdivided into training and validation set with a 70%-30% proportion.

The dataset is made public and available for download at the following link: https://zenodo.org/records/10616434
The DOI of the dataset is `10.5281/zenodo.10616434`

This dataset was used to train the following state-of-the-art pretrained deep learning architectures: VGG-16, ResNet50, InceptionV3, EfficientNet, and ViT by employing the Pytorch framework. We tested several hyper-parameters and two optimisers: Stochastic Gradient Descent (SGD) and Adam. In particular, the hyper-parameters used with SGD and Adam are these:

- SGD:
  1- lr=1e-3, momentum=0.9, weight_decay=1e-4
  2- lr=1e-4, momentum=0.8, weight_decay=1e-5
  3- lr=1e-5, momentum=0.7, weight_decay=1e-6
  4- lr=1e-6, momentum=0.9, weight_decay=1e-6
  5- lr=1e-2, momentum=0.9, weight_decay=1e-3
  6- lr=1e-2, momentum=0.8, weight_decay=1e-2
  7- lr=1e-2, momentum=0.8, weight_decay=1e-4
- Adam:
  1- lr=1e-3, weight_decay=1e-4
  2- lr=1e-4, weight_decay=1e-5
  3- lr=1e-5, weight_decay=1e-6
  4- lr=1e-6, weight_decay=1e-6
  5- lr=1e-2, weight_decay=1e-3
  6- lr=1e-2, weight_decay=1e-2
  7- lr=1e-2, weight_decay=1e-4

We carried out a total of 70 experiments, which are all available at the following GitHub repository: https://github.com/alessiochisari/CeilometerDatasetBenchmark

The authors will maintain the dataset on the platform used for a minimum time period of 10 years, during which time updates to the dataset are planned to increase the number of samples within the dataset.

The authors declare that this work does not infringe upon the rights of third parties, including but not limited to intellectual property rights, privacy rights, and other legal rights. Furthermore, the authors confirm that the data used in this study is subject to an appropriate license that permits its inclusion and usage in this scientific publication. In the event of any issues arising related to copyright or data license, the authors commit to fully cooperating to resolve the matter in accordance with applicable laws and editorial policies.