Deep Learning Based Mineral Dust Detection and Feature Selection

CyberTraining: Big Data + High-Performance Computing + Atmospheric Sciences

Ping Hou[1], Peng Wu[2],
Research assistant: Pei Guo[3], Faculty mentor: Aryya Gangopadhyay[3]

[1]School for Environment and Sustainability, University of Michigan
[2]Department of Hydrology and Atmospheric Sciences, University of Arizona
[3]Department of Information Systems, UMBC

Technical Report HPCF–2019–14, hpcf.umbc.edu > Publications

Abstract

Dust storm affects human health and the environment. In this study, we develop deep learning models to identify dust from cloud and surface using MODIS observations and CALIPSO data. We also identified the best subset of channels for dust detection by a shuffling procedure and a genetic algorithm. Results show the important features determined by the two methods are very similar. And the genetic algorithm selected a subset of features that have a comparable performance with the model with all features, proving the effectiveness of the genetic algorithm. The chosen subset will reduce future data collection efforts for dust detection.

1 Introduction

Dust storm occurrence is getting higher at the background of global climate change, especially in the arid and semi-arid regions. Dust, originated from the soil, acts as an effective type of aerosol that affects human health and the environment. Dust storms reduce visibility and cause dangers for high way traffic. For people with respiratory conditions like asthma, chronic obstructive airways disease (COAD) or emphysema, even small increases in dust concentration can make their symptoms worse (https://healthywa.wa.gov.au/Articles/F_I/Health-effects-of-dust).

After been blowing to the air, dust can absorb and scatter solar radiation and warm the surrounding air and reduce the sun's radiation that reaches the surface, imposing a shortwave cooling effect. On the other hand, dust absorbs longwave radiation and re-emits to the surface, imposing a warming effect to the surface. Dust particles can also act as cloud condensation nuclei (CCN) or ice nuclei (IN) in cloud formation processes and alter cloud life time and radiative effect by changing cloud droplet number concentration and size (Twomey, 1974; Albrecht, 1989). Due to the complex effect of dust on radiation, the uncertainty of future climate projection due to dust uplift and loading is one of the largest one in the Intergovernmental Panel on Climate Change (IPCC) report.

With the rapid development of satellite remote sensing, various methods have been proposed to utilize multi-channel observations to detect and retrieve dust information. The Moderate Resolution Imaging Spectroradiometer (MODIS) was a widely used passive sensor in dust detection. The retrieval methods include (a) Normalized Difference Dust Index (NDDI) [5]. However, NDDI is only appropriate in detecting dust storms when a dust-free image from a nearby time period is

available; (b) Brightness Temperature Difference (BTD) (band 31 − band32); (c) BTD (band 20 − band 31). The BTD methods are simple and efficient in detecting dust. However, any pixel that have BTD exceed the threshold difference will be classified as dust pixel and this can mis-identify land pixels as dust; (d) Reflective Solar Band (RSB). However, the RSB method requires significant amount of dust-free pixels in determining the threshold and this method has the same problem of mis-identifying land pixels as dust as in the BTD methods.

Starting 2011, as a replacement to MODIS, the Visible Infrared Imaging Radiometer Suite (VIIRS) on board Suomi National Polar-orbiting Partnership (NPP) spacecraft and Aqua satellite was launched. The VIRRS sensor has 16 moderate resolution (750 m) channels, while MODIS has 36 channels. Retrieval methods have been proposed to identify dust aerosols from the moderate resolution channels and some of them were adopted from the MODIS methods. For example, the deep blue algorithm uses absorbing aerosol index (AAI) and dust smoke discrimination index (DSDI) to identify dust pixels from clear sky. The AAI and DSDI depend on the reflectance ratios from 0.41, 0.44, and 2.2 μm channels. The NDDI and RSB methods can also be used in detecting dust aerosol. These two methods use the differences between 12 and 11 μm, the difference between 12 and 8.4 μm, and set different thresholds for identifying dust, smoke or clear sky.

The physical-based methods discussed above, however, highly depend on the thresholds to differentiate dust and dust-free pixels. Also, the detection accuracy, according to the Team3 project report [8] from 2018 CyberTraining, is 40-50% when compared with collocated CALIPSO dust index. Different from the machine learning methods from last year's project, we develop and test a deep learning model for dust detection in this project. Besides simply classifying dust and dust-free pixels, we also add cloud information in the learning procedure. Dust affect climate through cloud by changing cloud lifetime, thus it is imperative to study dust effect on cloud using synthetic dust and cloud observations. The true dust and cloud information is from collocated CALIPSO level-2 data and it is used in training and verifying the deep learning model.

The rest of the report is organized as follow: Section 2 introduces the data sets and deep learning model used in this project, Section 3 presents the results of cloud and dust identification from the deep learning model, followed by conclusions in Section 4.

2 Data and Methodology

2.1 Data

The Visible Infrared Imaging Radiometer Suite (VIIRS) is one of the key instruments onboard the Suomi National Polar-Orbiting Partnership (Suomi NPP) spacecraft, which was successfully launched on October 28, 2011. Viirs has 22 channels, 16 of which are moderate resolution bands (M-bands) and have a spatial resolution of 750 m at the Nadir. The other six channels are made up of five imaging resolution bands (I-bands), which have a spatial resolution of 375 m at the nadir, and one day/night panchromatic band with a spatial resolution of 750 m.The 22 channels cover wavelengths from 0.41 to 12.5 μm and can provide data records for clouds, aerosol, sea surface temperature, snow and ice, vegetation and fire.

The satellite, sun, and target relative positions are important factors affecting the amount of

radiation received by the satellite sensor. Though it is hard to quantify the 3D radiative effect, the geometric information cannot be ignored in classifying pixel categories. In this project, in addition to the radiometric channels, four geometric parameters from VIIRS: view and solar zenith angels, and view and solar azimuth angels, are used in training the deep learning model.

The dust and cloud information from CALIPSO satellite was used to train and test the deep learning model. CALIPSO was launched in 2006 as part of A-Train and onboard CALIPSO there is an active remote sensor, lidar, available that can provide reliable cloud and dust aerosol index.

Four categories were classified in this project with (1) dust with no cloud, (2) cloud with no dust, (3) dust with cloud, and (4) any other. From global perspective, cloud occurrence is much higher than dust, this will likely to cause unbalanced samples in the training and testing data sets. To avoid this, we chose the same number of samples for each category. For test purpose and to speed up the code, we selected total 10,000 samples in March 2012 and the samples are equally distributed among the four categories.

2.2 Deep learning

A deep leaning model simulates the way biological nervous systems (e.g., human brain) process information [4]. A deep learning model composes multiple layers of neurons. The first layer contains input predictors (in this case, satellite measurement at each channel, plus geometric information), and the last layer contains output responses (in this case, the classification of the images). Between the input layer and the output layer are one or more hidden layers interconnected with each other by hidden neurons. It is most common to have a reasonably large number of hidden neurons and train them with regularization. Choice of the number of hidden layers is guided by background knowledge and experimentation. Each layer extracts features of the input for regression or classification. Use of multiple hidden layers allows construction of hierarchical features at different levels of resolution [2]. In this study, we determine the number of hidden layers and the number of hidden neurons by trial and error, and use L2 regularization to suppress the large weights and result in a model that is more stable and less like to overfit the training data. The activation function and the learning algorithm in deep learning models are also selected by trial and error.

2.3 Feature selection

Another goal of this study is to find the important input features for dust detection. We use two approaches to select the most important features in the deep learning model: shuffling procedure and genetic algorithm. Shuffling procedure gives an importance order of the features, while genetic algorithm selects a subset of the features that can give the optimal prediction performance.

2.3.1 Shuffling procedure

The procedure is first to get a benchmark test accuracy by training the model once and then predict multiple times while randomizing each variable in the test set (Figure 2.1). The difference of the benchmark test accuracy and the test accuracy after permuting the variable, meaning with

and without the help of this variable, is used as an importance measure (i.e., permutation importance). If the accuracy after randomizing a variable is lower than the benchmark test accuracy, it is an important variable. On the other hand, if nothing changes or the accuracy is higher than the benchmark, it is a useless variable. We randomize 50 times and get an average test accuracy for each variable and compare with the benchmark test accuracy.

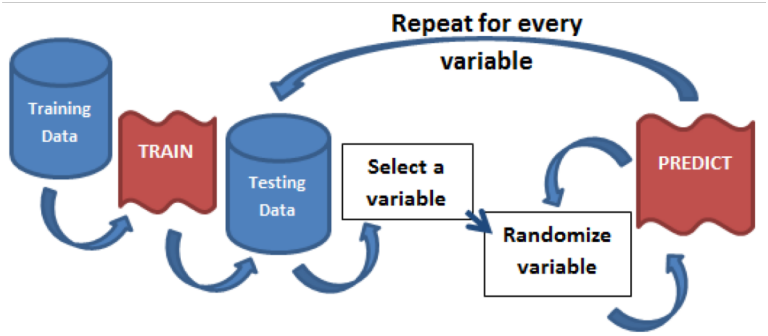

Figure 2.1: Shuffling procedure.

2.3.2 Genetic algorithm

Genetic algorithm is a directed random search technique that simulates the natural selection and evolution process [3]. Because it can be directly integrated to existing simulations and models, genetic algorithm has been widely used for many optimization problems which have a large number of parameters and their analytical solutions are hard to derive [6]. Rationally, genetic algorithm has also been used to optimize deep learning models [1, 7]. Here, we use a genetic algorithm to select a subset of features in our deep learning model. A sequence indicating whether a features is selected or not is defined as a genome. We use the genetic algorithm to guide the feature selection procedure in the following steps (Figure 2.2):

1. Initialization: we create a certain number of deep learning models with randomly generated genomes to be the population of the first generation;

2. Fitness evaluation: we train each model in the population and evaluate its performance on the test set using classification accuracy;

3. Selection: we rank all models in the population by accuracy and keep 20% of the top-ranked models to become part of the next generation to breed children. we also randomly keep 10% of the rest of the models. This helps find potentially successful combinations between worse-performers and top-performers, and also helps avoiding stuck in local maximum.

4. Crossover: crossover is the combination process from two members of a population to generate one or more children. Besides the top 20% models and the randomly kept 10% non-top models, to keep our population of 30 models, 21 children are generated for breeding in each generation.

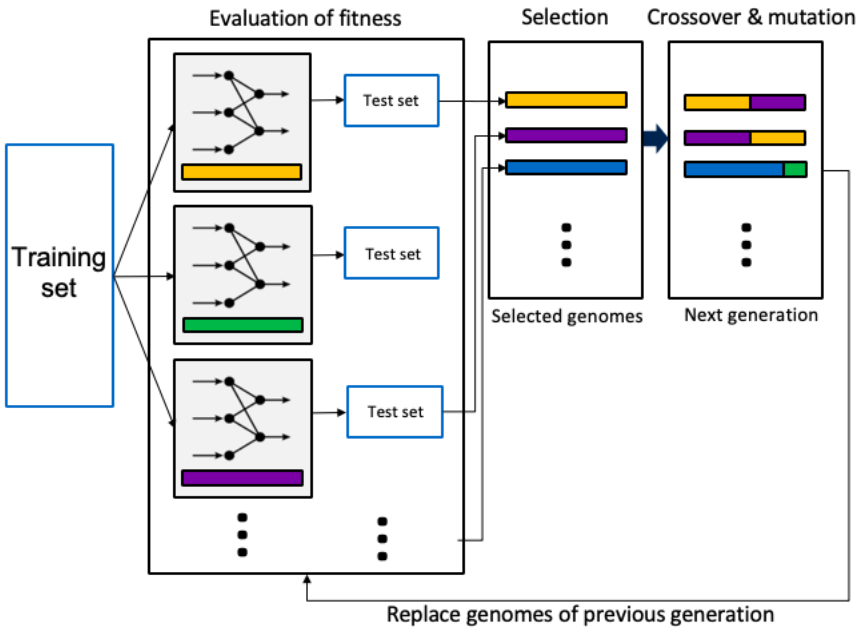

Figure 2.2: Genetic algorithm guided feature selection.

5. Mutation: we randomly mutate some of the genomes on some of the kept models

6. Genome replacement: genomes of the previous generations are replaced using the genomes after crossover and mutation.

7. Step 2 to step 6 are repeated for multiple generations until the model performance converges, i.e.,the test accuracy will not get any better. The best performed genome in the final generation is the selected best deep learning model, and best performed genome indicates the selected subset of the features.

3 Results

3.1 Deep learning model

Our final developed deep learning model has 5 hidden layers with 512 neruons in each layer. The activation function in each layer is 'relu', which is found to achieve better results than other functions. And we use "adam" to train the model. Figure 3.1 shows the training and validation loss on the test set. The classification accuracy on the test set is 71.1%.

3.2 Feature selection by shuffling procedure

Figure 3.2 shows the test accuracy when shuffling each feature in the model. The red horizontal line indicates the benchmark test accuracy (68.3%) when no variable is shuffled. Note the accuracy is

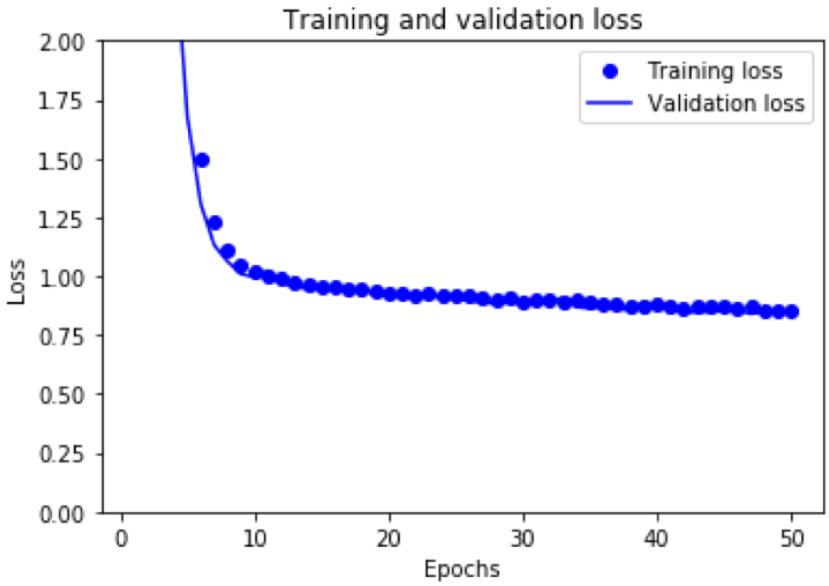

Figure 3.1: Training and validation loss on the test set.

different from 71.1% in section 3.1. This is due to the randomness in the training process. Neural networks and genetic algorithm are both stochastic, which means they make use of randomness (e.g., random weights initialized in the deep learning model, population random generated in the genetic algorithm) and therefore each time can produce different results. Each box in figure 3.2 indicates the distribution of the test accuracy of each variable being shuffled 50 times. Overall, the average test accuracy of all the variables are less than the benchmark test accuracy, which means all the variables contribute to the model to a certain degree. But some are more important than others. The most important features are the geometric information of solar zenith angle (18), view zenith angle (20), and solar azimuth angle (17), followed by channels 15 and 16 (11 and 12 μm) which are consistent with the NDDI and RSB methods from physical retrievas. Because dust particles are non-sphere particles, so their reflectance are different from different view angles, so are the emittance.

3.3 Feature selection by genetic algorithm

Table 3.1 shows the selected features by genetic algorithm and their performance. We test scenarios with different population size and different number of generations. The results show that, in general, when we have a larger population size and more generations, the genetic algorithm is able to find a better solution. When we have a population size as 64 and 8 generations, the best test accuracy can achieve 71.5%, which is even better than using all the features (71.1%). This proves that genetic algorithm is able to find a subset of features that can generate comparable results with all available features.

For the selected features in Table 3.1, the geometric information are important regardless of the population size and number of generations. This is consistent with the result from the shuffling pro-

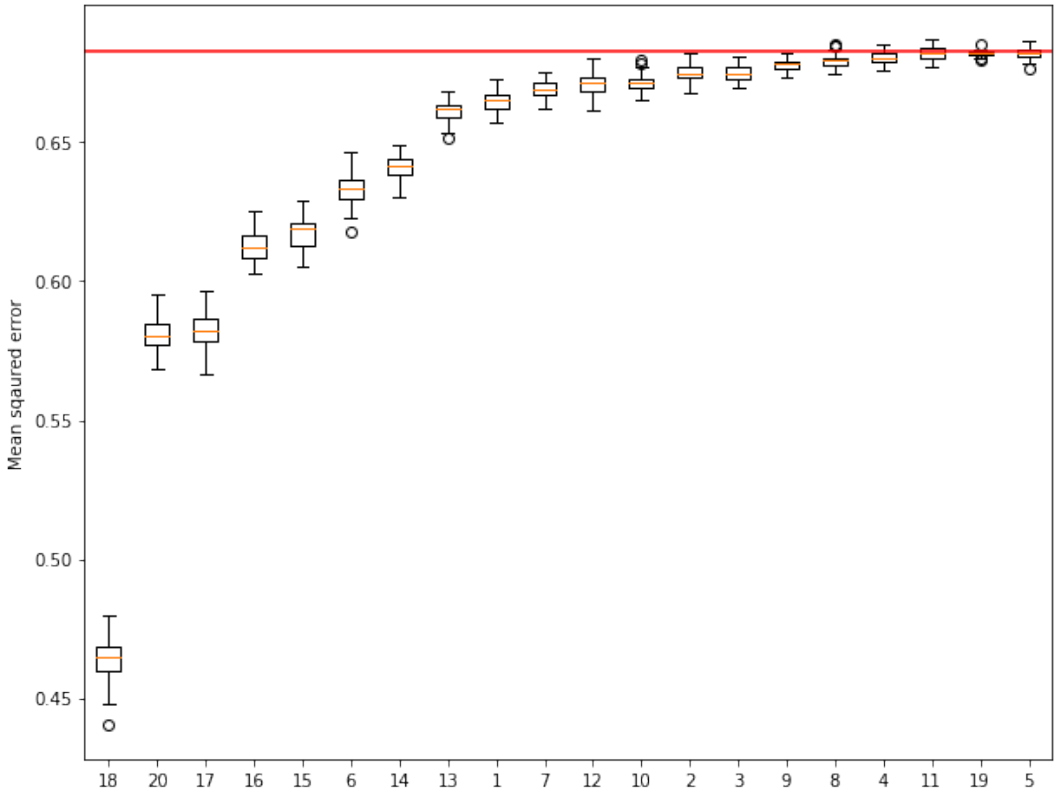

Figure 3.2: Training and validation loss on the test set.

cedure and is physically reasonable due to the non-spherical radiative properties of dust particles. The geometric matrix, however, are not adequately characterized in the physical based classification methods. Results from 2018 CyberTraining Team 3 [8] show that machine learning methods that include geometric angles have higher prediction accuracy than the physical based methods. The physical methods are not tested in this study, but the deep learning method is expected to have higher prediction accuracy.

The channels used in physical based are all selected as important features in Table 3.1, suggesting the deep learning model is capable of classifying dust and cloud in a physically feasible way.

Note that the prediction accuracy from our deep learning model is lower than what were presented in 2018 CyberTraining Team 3 report [8]. This is in part due to different satellite sensors in the two studies. The VIIRS, which has 16 radiometric channels, are used in this study while MODIS, which has 36 channels was used in [8]. Also, cloudy pixels are also classified in this study rather than only classifying dust pixels in [8]. Besides, due to the data size, the training and testing data sets were from the same day in [8], this will likely increase the prediction accuracy because of the similar dust and atmospheric properties. In this study, the training and testing data sets are randomly chosen and are not necessarily from the same day or over a specified region. To further improve and test our deep learning model, however, a larger data set will be required.

Table 3.1: Features selected by genetic algorithm

Population size	Number of generations	Selected features	Best test accuracy
8	4	2, 3, 6, 8, 9, 11, 12, 14, 15, 17, 18, 19	67.8%
16	4	2, 4, 8, 11, 13, 14, 15, 17,18, 19, 20	68.1%
32	4	1, 2, 3, 4, 5, 9, 10, 12, 13, 14, 16, 17, 18, 19, 20	70.3%
32	8	1, 3, 6, 8, 9, 10, 11, 17, 18, 19, 20	70.1%
64	8	1, 5, 7, 9, 12, 15, 16, 17, 18, 19, 20	71.5%

4 Conclusions

In this study, a deep learning model was trained and used to classify dust and cloud using VIIRS data. The deep learning model achieved a benchmark prediction accuracy of 71%. Through a careful tuning of population size and number of generations, the model is able to predict with comparable accuracy using a subset of the variables. The selected subsets consist of the channels used in the physical based classification methods. The geometric angles are always important in the subsets. A natural extension of this study is to implement the geometric angles into physical based methods. To further improve and test the deep learning model, more data points are needed. To facilitate the computation, more CPU nodes or GPU will likely needed.

Please see the codes for this project at https://github.com/big-data-lab-umbc/cybertraining/tree/master/year-2-projects/team-4

Acknowledgments

This work is supported by the grant CyberTraining: DSE: Cross-Training of Researchers in Computing, Applied Mathematics and Atmospheric Sciences using Advanced Cyberinfrastructure Resources from the National Science Foundation (grant no. OAC–1730250). The hardware in the UMBC High Performance Computing Facility (HPCF) is supported by the U.S. National Science Foundation through the MRI program (grant nos. CNS–0821258, CNS–1228778, and OAC–1726023) and the SCREMS program (grant no. DMS–0821311), with additional substantial support from the University of Maryland, Baltimore County (UMBC). See hpcf.umbc.edu for more information on HPCF and the projects using its resources.

References

[1] PG Benardos and G-C Vosniakos. Optimizing feedforward artificial neural network architecture. *Engineering Applications of Artificial Intelligence*, 20(3):365–382, 2007.

[2] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.

[3] John Henry Holland et al. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press, 1992.

[4] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[5] S. D. Miller. A consolidated technique for enhancing desert dust storms with MODIS. *Geophysical Research Letters*, 30(20), oct 2003.

[6] Duc Pham and Dervis Karaboga. *Intelligent optimisation techniques: genetic algorithms, tabu search, simulated annealing and neural networks*. Springer Science & Business Media, 2012.

[7] Marylyn D Ritchie, Bill C White, Joel S Parker, Lance W Hahn, and Jason H Moore. Optimizationof neural network architecture using genetic programming improvesdetection and modeling of gene-gene interactions in studies of humandiseases. *BMC bioinformatics*, 4(1):28, 2003.

[8] Peichang Shi, Song Qianqian, Janita Patwardhan, Zhibo Zhang, and Jianwu Wang. Mineral dust detection using satellite data. *Technical Report HPCF-2018-13, UMBC High Performance Computing Facility*, pages 0–11, 2018.