Atmos. Chem. Phys. Discuss.,
https://doi.org/10.5194/acp-2017-1189-RC1, 2018

Atmospheric Chemistry and Physics Discussions

Interactive comment on "Identification of new particle formation events with deep learning" *by* Jorma Joutsensaari et al.

Anonymous Referee #1

Received and published: 19 February 2018

Summary

This paper presents the use AlexNet, a specific class of deep learning method for automatize new particle formation identification. The study is interesting and important to substitute the current manual based NPF days classification to ease the study of atmospheric sciences.

Response overview

This article explains well the importance of applying data-driven method to replace human-based classification. In general, the materials, such as the used data, and the method are presented well. The obtained results are also adequate and interesting. The plots are clear and made well.

However, the reviewers would like to give some constructive comments, mentioning in the following paragraphs.

First, the article claimed that in the abstract and throughout the paper (section 1: lines 29-30) that this is the first time that their method classified successfully NPF events. Of course this statement is not true. In fact, Junninen et al. (2007), Kulmala et al. (2012) and later Zaidan et al. (2017) also succeed classifying this automatically. Although they were using different methods and aiming slightly for different NPF classes, but the primary aim is the same: automatic NPF event classification. Zaidan et al. (2017) also obtained the accuracy about 84% in their recent study, which they may considered also a success. Therefore, the word "first time" here undermines the previous contributions. Instead, you should mention how different this method and/or how this method is better compared to the previous methods.

References:

- Junninen, Heikki, et al. "An Algorithm for Automatic Classification of Two-dimensional Aerosol Data." Nucleation and Atmospheric Aerosols. Springer, Dordrecht, 2007. 957-961.
- Kulmala, Markku, Tuukka Petäjä, Tuomo Nieminen, Mikko Sipilä, Hanna E Manninen, Katrianne Lehtipalo, Miikka Dal Maso, et al. 2012. "Measurement of the nucleation of atmospheric aerosol particles." Nature protocols 7 (9): 1651–1667.
- Zaidan, M.A, V. Haapasilta, R. Relan, H. Junninen, P.P. Aalto, F.F. Canova, L.

Laurson, and A.S. Foster. Neural network classifier on time series features for predicting atmospheric particle formation days. In The 20th International Conference on Nucleation and Atmospheric Aerosols, 2017

In the lines 21-22 (section 1), the article said that “both of these studies were able to construct models to predict the probability of NPF occurrence with reasonable accuracy.” This statement is not quite true. In fact, the paper by Hyvonen et al (2005) used data mining, such as ML classifier, to find relevant atmospheric variables to NPF. It is not to construct models to predict the probability of NPF. This statement should be revised.

Reference;

- Hyvönen, S., Junninen, H., Laakso, L., Dal Maso, M., Grönholm, T., Bonn, B., Keronen, P., Aalto, P., Hiltunen, V., Pohja, T., Launiainen, S., Hari, P., Mannila, H., and Kulmala, M.: A look at aerosol formation using data mining techniques, Atmos. Chem. Phys., 5, 3345-3356, 10.5194/acp-5-3345-2005, 2005.

The introduction section seems redundant and too long. There is a mixed reviews between the important of aerosol study, the use of data driven method and the algorithm review that is used in atmospheric study. This section should be narrowed down, by focusing only motivation of aerosol study and the use of data mining algorithms, such as CNN, in the field of atmospheric sciences and related discipline. General algorithm review can be pointed out to a specific reference sources.

In section 2 , lines 7-9, is it possible to get the latest statistics? The figure is 11 years old, this may be interesting to present the latest one, because the analyzed dataset was started from March 2017 (section 2: line 23).

Also, I may misunderstood your statement in section 2, lines: 23-25. Could you please clarify how the total days was 5534 from 24 March 2012 until 16 May 2017? (since in a year, there is only 365/366 days).

The explanation of class categorizes are clear and help readers understand.

For section 2.3, the first paragraph seems containing a lot of new details about the properties of CNN. This needs more clarification, for example,what are kernels, RELU and sofmax functions. Please also include other relevant mathematical details in the section 2.3? For example, in addition to Figure 2, it is good to include the mathematical representation of CNN.

Any justification why you used standard procedure and options? (line 17, section 2.3)

Section 2.4, line 13, what happened to the first reference?

Another part of your data pre-processing is to uniform the pixel size of the images, could you explain how this has been done? Is it an automatic process?

What is the value to estimate class 0? because human might put some ambiguous days into this class., which they are not sure if the days are event and non-event. This class is very subjective, the CNN learning in this subjective class may confuse the model and bias the results.

What do you think if you filter out the bad data before you feed this into CNN learning and analysis? In this case, the CNN learning can be simplified by reducing the number of classes.

There is no method that is perfect. Describe the weakness of your method in the conclusion part?

Interactive comment on Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2017-1189, 2018.