EGU22-4988, updated on 16 Aug 2022
https://doi.org/10.5194/egusphere-egu22-4988
EGU General Assembly 2022

Deep learning cluster techniques for large aerosol datasets

David Topping[1], Ian Crawford[1], Martin Gallagher[1], Maxamillian Moss[1], Man Nin Chan[2], Hing Bun martin Lee[2], Sinan Xing[2], Tsin Hung Ng[2], and Amos Tai[2]
[1]University of Manchester, Centre for Atmospheric Science, Simon building, Manchester, United Kingdom of Great Britain – England, Scotland, Wales (david.topping@manchester.ac.uk)
[2]Faculty of Science, The Chinese University of Hong Kong

The impacts that aerosol particles have on the climate, air-quality and thus human health, are linked to their evolving chemical and physical characteristics. We know that many processes taking place in/on atmospheric aerosol particles are accompanied by changes in the particles' morphology (size and shape). Likewise, particles of primary origin [e.g. desert dust, volcanic ash, soot, pollen] can have widely varying morphological features that should nonetheless offer significant information to aid detection and classification. There have been a number of developments in instrument capability that capture spectral signatures and images of individual particles. This includes the development of instruments designed to identify biological particles through a combination of fluorescent profiles, fluorescence lifetime decay and scattering images. Whilst the development of experimental frameworks continue, the marriage of developing compact systems and associated data analytics is lacking. There has been a rapid proliferation of data science methodologies, now wrapped up in commonly accessible open source environments. However, the tuning of appropriate hyperparameters, and the choice of architecture can be difficult. Likewise, where the challenge is to include multivariate datasets into one routine, such as chemical signatures and scattering information, it is difficult to know how best to combine them.

In this presentation we present an evaluation of an approach that uses a combination of convolutional neural network autoencoders to 1) lead to cluster strategies for long term ambient datasets, no matter the size and 2) combines sell-supervised learning with small amounts of laboratory data to arrive at classification routines. In each case the architecture is automatically tuned using a hyperband approach and various data augmentation strategies are applied. The initial focus is on data derived from the PLAIR Rapid-E instrument , though developments are applicable to a number of measurement techniques.