Public datasets

Signal-background classification with Parametric Neural Networks

- Dataset: https://zenodo.org/record/6453048
- Agenda: https://agenda.infn.it/event/34607/
- Contact: Luca Anzalone

HEPMASS-IMB is a benchmark dataset for *signal-background classification* in High-Energy Physics (HEP), derived from HEPMASS (Baldi et al.) by imbalancing it two times: on the class labels, as well as on the mass labels.

- It has 27 feature columns (named from *f0* to *f26*), and a 28-th mass feature (named *mass*).
- The 27 features are *already normalized* to have approximately zero-mean and unitary variance.
- The mass feature has five unique values: *500*, *750*, *1000*, *1250*, and *1500*.
- There are two class labels: 1 (signal), and 0 (background).
- The dataset describes the decay of an hypothetical particle: Xtt¯X>tt¯W+bWb¯.

Further details about the original dataset are available here, whereas a description of our modifications is presented in our paper.

NOTE:

- The files provided here represent only the *training-set*, since it's what is diverse compared to the original HEPMASS.
- The label column has been renamed from "# label" to "type".
- There are two new columns: *name*, and *weight*.

NPLM (New Physics Learning Machine)

- Dataset: https://zenodo.org/record/4442665
- Agenda: https://agenda.infn.it/event/34963/
- Contact: Gaia Grosso

Archive of synthetic data used for the studies presented in arXiv:1912.12155

MLaaS4HEP for the Higgs boson ML challenge

- Dataset: https://www.kaggle.com/competitions/higgs-boson/data, http://opendata.cern.ch/record/328
- Agenda: https://agenda.infn.it/event/35136/
- Contact: Luca Giommi

This dataset comes from the Higgs boson ML challenge, a competition held in 2014, organized by a group of ATLAS physicists and data scientists, and hosted by the Kaggle platform. For the full description of the dataset see the official documentation.

Cell counting with cell-ResUnet

- Dataset: http://amsacta.unibo.it/6706
- Agenda: https://agenda.infn.it/event/34695/
- Contact: Luca Clissa

By releasing this dataset, we aim at providing a new testbed for computer vision techniques using Deep Learning. The main peculiarity is the shift from the domain of "natural images" proper of common benchmark dataset to biological imaging. We anticipate that the advantages of doing so could be two-fold: i) fostering research in biomedical-related fields - for which popular pre-trained models perform typically poorly - and ii) promoting methodological research in deep learning by addressing peculiar requirements of these images. Possible applications include but are not limited to semantic segmentation, object detection and object counting. The data consist of 283 high-resolution pictures (1600x1200 pixels) of mice brain slices acquired through a fluorescence microscope. The final goal is to individuate and count neurons highlighted in the pictures by means of a marker, so to assess the result of a biological experiment. The corresponding ground-truth labels were generated through a hybrid approach involving semi-automatic and manual semantic segmentation. The result consists of black (0) and white (255) images having pixel-level annotations of where the stained neurons are located. For more information, please refer to Morelli, R. et al., 2021. Automating cell counting in fluorescent microscopy through deep learning with c-ResUnet. Scientific reports, (in press). https://doi.org/10.1038/s41598-021-01929-5. The collection of original images was supported by funding from the University of Bologna (RFO 2018) and the European Space Agency (Research agreement collaboration 4000123556).

Fast classifier-based goodness of fit test for online data quality monitoring

- Dataset: https://zenodo.org/record/7128223
- Agenda: https://agenda.infn.it/event/35822/
- Contact: Marco Letizia

Dataset for DQM for Drift Tube Chambers. The dataset include a reference sample and smaller data samples characterized by anomalous effects. Plots for data visualization are provided.